SOMMAIRE

INTRODUCTION GENERALE : LES PROJETS DE SCIENCE CITOYENNE COMME ORGANISATION SCIENTIFIQUE EPHEMERE.

> *Foule (n.f.) : grand nombre de personnes rassemblées de manière désorganisée ou indisciplinée (traduit à partir du Oxford Living Dictionaries)*

Depuis les années 2000, les organisations font face à l'émergence d'une culture de la collaboration massive que ce soit dans les entreprises pour la conception de projets de R&D, en informatique pour développer de nouveaux outils, pour financer des projets ou même dans des nouvelles formes de collaboration comme Wikipédia que l'on peut voir apparaître sur internet (Chesbrough, 2006; Hippel & Krogh, 2003; Howe, 2006; Nielsen, 2011; Surowiecki, 2004; Tapscott & Williams, 2006). Cette tendance est à l'inverse d'une perception négative tenace du fonctionnement d'une foule. Elle est issue de nombreuses expérimentations en psychologie réalisées durant le 20è siècle décrivent comment l'individu et les décisions qu'il prend se retrouvent altérées par la psychologie de la foule (Gilovich, et al., 2012; Le Bon, 1908). En effet, une littérature émergente depuis les années 2000 propose une nouvelle approche de la collaboration entre les individus, où l'action collective d'une foule a plus de valeur que la somme des actions individuelles (Lüttgens, Pollok, Antons, & Piller, 2014; Smith, 2009; Surowiecki, 2004). Des études mettent en évidence un ensemble de situations dans lesquelles la perception et la résolution de problèmes sont plus efficaces quand elles sont réalisées par la foule que par des individus seuls (Mollick & Nanda, 2016; Poetz & Schreier, 2012). Pour que la foule atteigne de tels résultats, elle doit présenter ce que Surowiecki (2004) définit comme une forme de « sagesse » qui lui assure une diversité, une indépendance et une décentralisation.

Fort de cette approche, plusieurs organisations scientifiques ont cherché à utiliser la foule au sein de leurs processus en déléguant des tâches qui sont habituellement réalisées en interne (Brokaw, 2011; Buecheler, Sieg, Füchslin, & Pfeifer, 2010; Cooper, Dickinson, Phillips, & Bonney. R., 2007; Franzoni & Sauermann, 2014; Theisz, 2017). Ces organisations « ouvrent » une partie de leur processus à des groupes d'individus non limité en taille et non identifiés par une quelconque expertise ou un relation établie avec l'organisation. Plusieurs exemples dans la littérature montrent l'efficacité de l'ouverture à la foule pour des situations simples, mais également pour des résoudre des problèmes complexes (Franzoni & Sauermann, 2014; Nielsen, 2011). Cette nouvelle forme d'organisation contemporain de la science pose cependant des questions fondamentales : Comment pouvons-nous identifier et rendre compte de ce nouveau phénomène ? Dans quelles conditions et jusqu'où est-il possible d'ouvrir le processus ? Peut-on mettre en place des pratiques

de gestion qui permettent de s'assurer de sa performance ? Si oui, est-il possible de s'assurer de leur efficacité non pas sur un projet isolé mais comme une pratique continue dans le temps ?

Cette thèse se propose d'étudier et d'analyser le phénomène d'ouverture du processus scientifique dans un contexte contemporain. Nous allons voir qu'il y a une opportunité à interroger cette forme d'organisation en lien avec une transformation contemporaine du rapport des sciences avec leurs données. Nous allons préciser dans l'introduction générale le contexte de cette recherche, les motivations théoriques et empiriques, les principales questions de recherche ainsi que les grandes lignes de la méthodologie permettant de répondre aux différentes questions de recherche. Nous présentons enfin, le synopsis du manuscrit.

1. LES SCIENCES CITOYENNES COMME NOUVELLE FORME D'ORGANISATION SCIENTIFIQUE EPHEMERE

1.1. EMERGENCE D'UN NOUVELLE FORME D'ORGANISATION DE LA SCIENCE

En 2007, Kevin Schawinski, alors doctorant en astronomie dans le laboratoire de Chris Lintott à l'Université d'Oxford, cherchait à étudier les galaxies elliptiques qui avaient formé les étoiles les plus récentes. Ses premières hypothèses étaient basées sur un échantillon limité de galaxies que Schawinski avait codé manuellement, mais plus de données étaient nécessaires pour les vérifier. Le groupe de chercheurs auquel il appartient a alors eu l'idée d'utiliser les 930 000 images de galaxies lointaines que le Sloan Digtal Sky Survey (SDSS) avait mis à disposition quelques mois plus tôt. Les images étaient brutes et devaient être codées (type de galaxie, caractéristiques) pour être exploitables. Or les ordinateurs ne sont pas particulièrement bons pour la détection automatique d'images et, la grande quantité d'images disponibles ne pouvait être traité rapidement par une poignée de scientifiques. Les chercheurs ont donc eu l'idée de créer une plateforme en ligne, appelée *Galaxy Zoo*, dans laquelle n'importe quel volontaire pouvait s'inscrire afin d'aider les scientifiques à classer ces images. La tâche était accessible à tous sans compétence préalable requise et le volontaire était aidé grâce à un tutoriel ainsi qu'un ensemble d'exemples disponibles sur la plateforme. Chaque volontaire devait visualiser les images puis coder six propriétés différentes des objets astronomiques. La participation est devenue rapidement virale, et sept mois après le lancement du projet environ 900 000 galaxies furent codées par plus de 250 000 volontaires uniques. Afin de réduire la probabilité d'un codage incorrect, les galaxies furent codées plusieurs fois par différents volontaires (environ 50 codages par image), pour un total d'environ 50 millions de classifications. A titre de comparaison, ces 50 millions de classifications auraient requis plus de 83 années à plein temps pour un scientifique seul. Les données récoltées par Galaxy Zoo ont permis à l'équipe de chercheurs de mener à bien l'étude initialement prévue, mais ont également été bénéfiques pour d'autres astronomes et la découverte de nouveaux objets astronomiques (Cardamone et al., 2009).

Ce projet est caractéristique d'une tendance forte et des expériences de ce type se sont significativement multipliées ces 10 dernières années (Franzoni & Sauermann, 2014; Houllier, 2016; Wiggins & Crowston, 2011). *Foldit*, par exemple, est un projet collaboratif incluant plusieurs centaines de milliers de participants, et dont l'objectif est d'améliorer les connaissances scientifiques en biologie sur le repliement des protéines. La plateforme *DREAM* challenge met en relation des biologistes avec des centaines de spécialistes de l'analyse de données pour résoudre des problèmes de biologie computationnelle. *Polymath* est un blog qui implique des mathématiciens hautement reconnus avec des amateurs afin de résoudre collectivement des problèmes qui ont longtemps échappé aux approches traditionnelles des mathématiques. Ce phénomène, loin d'être isolé, apparait dans un grand nombre de disciplines scientifiques : microbiologie, médecine, écologie, astronomie, neuroscience, physique, histoire, mathématique pour n'en citer que quelques-unes (Franzoni & Sauermann, 2014; Haklay, 2015; Houllier, 2016; Nielsen, 2011; Wiggins & Crowston, 2011). En Août 2018, le site SciStarter[1] répertoriait plus de 2700 projets de recherche faisant participer des inconnus dans le processus scientifique. Définis entre autres par le terme de **science citoyenne** (Franzoni & Sauermann, 2014; Hand, 2010; Houllier, 2016; Wiggins & Crowston, 2011), ces projets tirent parti de l'effort fourni par une foule de contributeurs variés.

Les projets de science citoyenne intéressent de plus en plus la communauté scientifique, mais également des structures publiques, des agences de financement qui cherchent à évaluer ses avantages et ses défis potentiels considérables. Par exemple, des revues ont publié des numéros spéciaux autour de la science citoyenne (R. Bonney et al., 2014), et le sujet a été abordé dans le cadre des sciences de gestion dans des médias tels que le *Sloan Management Review* (Brokaw, 2011). Plusieurs rapports publics[2] font également état de la multiplication de ce type de projet et démontrent de l'intérêt des sphères publiques à s'approprier la question des sciences citoyennes (e.g. Haklay, 2015; Houllier, 2016). Enfin, il existe une réelle volonté de la part de certains Etats d'investir dans le développement de projets de science citoyenne : le gouvernement canadien a lancé un appel à projets « Protection des Grands Lacs » pour lequel il financera tout projet basé sur un principe de science citoyenne[3] ; le gouvernement australien fournit des fonds à toute organisation souhaitant développer des projets scientifiques qui font participer le public[4] ; les sciences citoyennes sont également l'une des cinq orientations stratégiques du nouveau

[1] https://scistarter.com/about

[2] Voir également entre autres le rapport du COMETS du CNRS en 2015 *Les science citoyennes* ; *Paper on Citizen Science* de la Commission européenne en 2013 et 2014 ; ainsi que des rapports menés dans des disciplines spécifiques comme l'astronomie ou la biodiversité

[3] https://www.canada.ca/fr/environnement-changement-climatique/services/protection-grands-lacs/financement/accroitre-mobilisation-public-grace-science-citoyenne.html

[4] https://www.business.gov.au/assistance/inspiring-australia-science-engagement/citizen-science-grants

programme de travail 2018-2020 «Une science avec et pour la société» (*SwafS*) dans Horizon 2020 de la Commission Européenne[5].

L'appel à la contribution d'un public non-scientifique au sein du processus scientifique n'est pas aussi récent que l'on pourrait le croire. En 1900 par exemple, l'américain Frank Chapman, fondateur du magazine Bird-Lore, proposa aux citoyens de compter les oiseaux à Noël. Cette année-là, 27 observateurs prirent part au comptage à 25 endroits aux États-Unis et au Canada. Le succès de l'initiative poussa les organisateurs à réitérer le processus chaque année à la période de Noël et depuis, les dénombrements ont intégré un nombre croissant d'observateurs. En 2009, les données collectées grâce à ce programme auront permis la rédaction de 350 publications scientifiques autour de la biodiversité (Cohn, 2008; Houllier, 2016). A une époque où les moyens de collecter les données étaient plus réduits, ce type d'organisation offrait deux avantages aux scientifiques : la facilité d'accès à des terrains parfois répartis géographiquement, l'intégration de groupes de personnes concernées (par ex. pour l'installation d'un site nucléaire). Des scientifiques en biologie, écologie, agronomie, ou astronomie ont pu tirer parti de l'implication de volontaires et d'amateurs dans leur domaine pour collaborer avec eux et augmenter leurs ressources expérimentales (Bonney et al., 2009; Callon, Lascoumes, & Barthe, 2001).

Or, des éléments diffèrent entre les cas historiques et les initiatives contemporaines qui justifient de son intérêt grandissant : la multiplication et la variété des cas contemporains, le nombre de participants dans certains projets qui peut atteindre parfois plusieurs centaines de milliers de personnes, le type d'activité ouverte à la participation du public. Les exemples historiques concernent généralement des tâches qui ne nécessitent pas de compétences particulières de la part des participants : compter le nombre d'oiseaux dans une zone géographique délimitée, calculer des valeurs numériques à partir de formules mathématiques (Poitou, 1982) déterrer des fossiles d'animaux (Torrens, 1995). Au contraire, les projets contemporains s'appuient sur différents niveaux d'apport par les participants : certains collectent des données, d'autres les analysent, certains prêtent leurs ordinateurs d'autres fournissent un financement (Nielsen, 2011). Les scientifiques tirent également parti des efforts et des connaissances fournis par une base de contributeurs importante et diversifiée, élargissant potentiellement l'éventail des problèmes scientifiques pouvant être traités à un coût mieux maîtrisé, en tout les cas plus abordable. Un nombre croissant de publications revendiquent une utilisation avec succès du public pour résoudre des problèmes scientifiques parfois complexes (e.g. Azencott et al., 2017; Ewing et al., 2015; Griffith et al., 2017; Saez-Rodriguez et al., 2016; Warby et al., 2014). Le cas de Foldit en est un bon exemple. Les scientifiques cherchaient depuis plus de 10 ans à modéliser la structure tri-dimensionnelle d'une protéine du virus du SIDA. Le concours a été lancé, 200 000 personnes ont participé, et trois ont fait une proposition pertinente. Cela aura pris moins de trois semaines, et le modèle conçu par ces trois participants est maintenant utilisé par les biologistes pour étudier les différents sites disponibles sur la molécule de la protéine (Khatib et al., 2010).

[5] http://www.sisnetwork.eu/media/sisnet/Policy_brief_Citizen_Science_SiSnet.pdf

1.2. LA SCIENCE CITOYENNE COMME ORGANISATION SCIENTIFIQUE EPHEMERE

Actuellement, la littérature continue de cumuler les initiatives des projets de science citoyenne et de présenter des exemples, généralement des « *success stories* ». Bien que certains scientifiques cherchent à rationaliser le processus (Franzoni & Sauermann, 2014; Haklay, 2015; Houllier, 2016; Wiggins & Crowston, 2011), les exemples d'application restent pour l'heure plutôt considérés comme des évènements ponctuels (Brokaw, 2011; Nielsen, 2011). La mobilisation d'un public au sein du processus de production de connaissance est encore largement transparente du point de vue des institutions scientifiques traditionnelles. Si les participants sont cités comme auteurs dans les publications (e.g. Cranshaw & Kittur, 2011; Khatib et al., 2010), leur place au sein du modèle de production de connaissances n'est pas considérée, et ils échappent pour le moment à un système normé et reconnu. Le manque d'analyse actuel pointe qu'il y a besoin d'étendre, si ce n'est de concevoir un cadre conceptuel adapté.

Nous proposons d'envisager les effets de la science citoyenne dans le domaine de la gestion des connaissances et des organisations scientifiques. Dans ce cadre d'analyse, les sciences citoyennes sont considérées comme une nouvelle forme **d'organisation scientifique éphémère.** Par éphémère, nous mettons en avant deux caractéristiques des projets: leur limite dans le temps et dans l'espace, et l'intégration d'acteurs éphémères. En effet, alors que le processus scientifique se tient généralement dans des structures dédiées avec une organisation, des normes et un cadre structuré, les organisations scientifiques éphémères sont souvent hors des murs de ces institutions et limitées dans le temps. Ensuite, les acteurs d'une organisation scientifique éphémère ne participent que pour le projet et n'ont pas d'intérêts directs à ce que le projet aboutisse. Par intérêt, nous entendons un intérêt pour leur carrière (reconnaissance par les pairs), une obligation contractuelle, ou encore un intérêt personnel (concerné par les résultats du processus de recherche).

Une difficulté immédiate dans ce type d'organisation éphémère est le faible niveau de contrôle que les organisateurs peuvent avoir avec les participants. Beaucoup d'éléments leurs sont inconnus : les compétences et les expertises des participants, le nombre de participants, mais également le niveau d'engagement au sein du projet. Comment organiser un tel dispositif? Serait-il possible de mobiliser le public de manière régulière et répétitive ? Y'a-t-il des processus scientifiques où une contribution des citoyens est plus adaptée que d'autres ? Une organisation scientifique qui souhaiterait se lancer dans ce type de projet devrait avoir une vision plus claire sur ces aspects.

Une analyse plus approfondie des divers avantages et défis associés à ces projets peut, en principe, nous permettre d'évaluer dans quel contexte la science citoyenne peut être une forme d'organisation scientifique adaptée. Par ailleurs, de telles études pourraient définir quels types de problèmes scientifiques peuvent s'apprêter plus à l'usage de ce dispositif. Cependant, la littérature scientifique sur cette question relève plutôt de la science de gestion et de l'organisation. Dans

notre thèse, nous nous intéressons à la manière d'organiser les projets de science citoyenne pour gérer la performance de ces projets.

2. LA SCIENCE CITOYENNE COMME PHENOMENE SCIENTIFIQUE

Une partie de la littérature sur le sujet a considéré l'ouverture du processus scientifique à des non scientifiques comme une caractéristique fondamentale des projets de science citoyenne. Ces auteurs analysent la vague d'émergence des projets de science citoyenne comme un mouvement de ré-interrogation de la pratique de la science, de son organisation et de sa gouvernance et militent pour une ouverture de l'activité de recherche (Fecher & Friesike, 2014; Nosek et al., 2015). Dans un article intitulé « *Promoting an open research culture* » publié dans le magazine *Science* en 2015, Nosek et ses collègues critiquent de façon univoque un réel manque de transparence, d'ouverture et de reproductibilité dans la science, alors que ces caractéristiques sont généralement considérées par les chercheurs comme fondamentales et propres à la culture scientifique.

Les projets de science citoyenne s'inscrivent dans ce courant comme un moyen d'ouvrir le processus de production de connaissance scientifique à la société et aux citoyens. Les citoyens semblent eux-mêmes demandeurs de contribuer plus à la production scientifique (Reed et al., 2013). Le succès de l'ouverture à la foule est souvent présenté comme étant la conséquence d'une « sagesse de la foule » (Surowiecki, 2004) et la supériorité d'une foule par rapport aux individus. Dans cette perspective, les sciences citoyennes sont vues comme un phénomène sociologique et sociétal dont le principe est d'intégrer la culture de l'ouverture au sein des organisations existantes. Cette caractérisation des projets par leur ouverture aux citoyens peut être caractérisée comme une approche militante et engagée de la production scientifique dont les objectifs seraient d'accéder à une meilleure démocratisation de la science vis-à-vis de la société. En revanche, ce courant n'explique pas si on peut et comment ce dispositif procède de la méthode scientifique de manière récurrente et maîtrisée. Au contraire, une approche gestionnaire des projets de science citoyenne pourrait nous permettre de mieux comprendre comment les organisations scientifiques peuvent tirer parti du public dans leur activité scientifique.

2.1. LA SCIENCE CITOYENNE ET LA GESTION TRADITIONNELLE DE LA SCIENCE

Notre projet vise à proposer la particularité des modes de gestion et d'organisation des projets de science citoyenne. Dans cet objectif, le modèle traditionnel de gestion de la science peut sembler le plus naturel au premier abord. La collaboration entre scientifiques et acteurs non reconnus comme scientifiques existe déjà dans beaucoup de structures dédiées à la production scientifique. En fait, le scientifique travaille le plus souvent au sein d'un petit groupe en interaction avec des non scientifiques comme des stagiaires, des laborantins ou des techniciens à qui il délègue une partie de son activité (Hagstrom, 1964; Pelz, 1960). Ces derniers, bien que baignés dans la culture

scientifique (Cole, 2011), ne possèdent pas généralement les connaissances spécifiques aux scientifiques mais développent à la place d'autres types de connaissances qui prennent d'autres formes (Polanyi, 1966). Certains techniciens de laboratoires seraient plus compétents que les scientifiques à suivre des protocoles, ou des techniciens en radiologies seraient meilleurs à lire les radios. La collaboration entre acteurs ayant des rôles différents dans la production de connaissance demande alors une organisation et une coordination entre les individus (Allen, Lee, & Tushman, 1980; Chompalov, Genuth, & Shrum, 2002; Elias, Cavana, & Jackson, 2002; König et al., 2013; Liberatore & Titus, 1983; Stokols et al., 2005).

Les scientifiques sont également amenés dans certains cas à travailler avec des non scientifiques hors des systèmes institutionnels pour accéder à une expertise particulière ou alimenter la réflexion sur un projet de recherche (Houllier, 2016; Israel et al., 1998). Ces acteurs sont intégrés au processus non pas en fonction de leurs capacités à répondre au problème scientifique, mais par rapport à l'histoire singulière et leur implication qui les lient à la question scientifique (Callon et al., 2001). Les sociologues soulignent l'apparition de modèles d'organisation démocratiques qui diminuent l'impact de la place dominante du sachant, le scientifique, vis-à-vis des profanes, et forment des jeux de négociation entre les différents acteurs (Callon et al., 2001).

Si les formes de collaboration entre scientifiques et non scientifiques sont variées (Chompalov et al., 2002), savoir laquelle convient pour organiser les projets de science citoyenne n'est pas évident étant donné les spécificités des projets de science citoyenne. D'abord, il n'existe aucune relation contractuelle qui lie les volontaires aux organisateurs du projet ou à l'institution pour obliger à intégrer le projet. La participation dépend de leur motivation qui ne peut être fondée sur les objectifs traditionnels des employés d'une institution comme conserver son emploi, faire carrière dans l'institution ou obtenir la reconnaissance de ses pairs (Merton & Storer, 1973). Ensuite, les organisateurs ne connaissent pas à l'avance les compétences ni le savoir-faire des citoyens de la science, et ne peuvent donc prévoir si le projet sera correctement exécuté, surtout lorsque celui-ci nécessite des compétences spécifiques. De plus, il n'y a pas de limite à la participation dans les projets de science citoyenne (Franzoni & Sauermann, 2014). Les scientifiques ne peuvent donc savoir à l'avance combien de personnes vont contribuer. Savoir qui et combien de personnes vont contribuer à un projet dans une organisation est souvent essentiel pour pouvoir estimer le temps que peut prendre le projet, ainsi que la capacité à mener le projet à bien. Dès lors que ces conditions deviennent incertaines, cela limite fortement la capacité à organiser le travail à réaliser. Enfin, les projets de science citoyenne sont toujours des organisations limitées dans le temps. Contrairement à une organisation scientifique traditionnelle basée sur des programmes de recherche longs, où du moins dont la durée est souvent incertaine (par exemple les incertitudes sur la durée d'une thèse de doctorat (Delamont & Atkinson, 2001)), les projets de science citoyenne sont relativement courts, de quelques jours à quelques mois (Rotman et al., 2014).

2.2. MECANISMES DE REUTILISATION DURANT LE PROJET VIA LE CROWDSOURCING

La littérature actuelle sur l'organisation et la gestion de projets de sciences citoyennes s'accorde plutôt à traiter ce phénomène en lien avec le crowdsourcing (e.g. Buecheler, Sieg, Füchslin, & Pfeifer, 2010; Erickson, 2011; Franzoni & Sauermann, 2014; See et al., 2016; Wiggins & Crowston, 2011). En effet, ce modèle est particulièrement en vogue dans la recherche et un nombre croissant de publications dans des revues prestigieuses comme *Nature* ou *Science* en revendiquent une utilisation avec succès pour résoudre leurs problèmes (Azencott et al., 2017; Ewing et al., 2015; Griffith et al., 2017; Saez-Rodriguez et al., 2016; Warby et al., 2014). De plus, plusieurs similitudes entre crowdsourcing et science citoyenne justifient ce rapprochement (Theisz, 2017).

Le crowdsourcing est souvent défini comme le fait d'intégrer un ensemble non limité de personnes non sélectionnées dans un processus afin d'exécuter une tâche sans aucun engagement pour accomplir cette tâche (Afuah, 2018; Afuah & Tucci, 2012; Howe, 2006). L'exécution de la tâche ne se limite pas à une personne en particulier ou à un groupe de personnes pré-identifiées, mais est ouverte à tous. Le crowdsourcing est construit sur un modèle organisationnel de type « seeker-solver » (Jeppesen & Lakhani, 2010; Sieg, Wallin, & von Krogh, 2010) ou d'appel au marché (Terwiesch & Xu, 2008). Les seekers formulent un problème bien délimité qu'ils diffusent ensuite à un ensemble indéfini de solvers sous la forme d'un appel ouvert. Dans cette littérature les projets de science citoyenne sont analysés par un jeu économique dans lequel la performance est étudiée au travers de l'effort fourni par les participants, des incitations à la participation, ainsi que de l'incertitude à trouver une solution à un problème (Terwiesch & Xu, 2008).

La littérature sur le crowdsourcing et plus généralement sur les organisations incluant des communautés de participants s'intéresse de plus en plus à la question de la collaboration au sein du processus (e.g. Riedl & Woolley, 2017; Tucci, Afuah, & Viscusi, 2018; West & Sims, 2018; Zhao & Zhu, 2014). En effet, la collaboration au sein d'une communauté est un contexte idéal pour partager et réutiliser des informations par opposition à un système classique de collaboration qui est souvent contraint à un ensemble de règles et de rôles à respecter (Anderson & Dron, 2014). Cette réutilisation a été mise en œuvre dans de nombreuses situations comme dans les logiciels open source (Hippel & Krogh, 2003; Murray & O'Mahony, 2007; von Krogh, Spaeth, & Lakhani, 2003), Wikipedia (Kittur, Kraut, & Kraut, 2008), mais également dans les projets de science citoyenne comme le projet Foldit en biologie computationnelle (Nielsen, 2011), ainsi que le projet Polymath pour la construction de preuves mathématiques (Gowers, 2009). Les études sur la collaboration par réutilisation se sont généralement intéressées aux formes de contrat qui régissent la réutilisation entre les parties prenantes (Boudreau & Lakhani, 2015; Pearce, 2016), aux moyens pour favoriser la coordination (Faraj & Majchrzak, 2011; Lakhani & Von Hippel, 2003; Levine et al., 2014; Wasko et al., 2017), ou encore à l'impact de la diversité chez participants (Cronin & Weingart, 2007; Dahlin et al., 2005; Phillips et al., 2004). Par ailleurs, des travaux menés au sein du Centre de Gestion Scientifique des Mines Paristech montrent l'émergence de

nouveaux modes de crowdsourcing avec des communautés de concepteurs seeker-solvers capables d'explorer et de développer collectivement de nouvelles idées (Kokshagina, Gillier, Cogez, Le Masson, & Weil, 2014).

En revanche, les mécanismes de collaboration par réutilisation qui peuvent exister au sein d'une foule, notamment dans un cadre généralement bien maîtrisé comme celui du crowdsourcing, manquent d'un modèle clair de performance de la production. Les organisateurs de projets de crowdsourcing devraient pouvoir évaluer quel mode d'organisation mettre en œuvre suivant le type de résultat qu'ils souhaitent obtenir. Par ailleurs, les modèles d'organisation faisant appel à une foule sont souvent basés sur une approche statique, où le seeker n'intervient qu'au début et à la fin du processus. Quels seraient les effets d'une intervention répétée au sein du processus ?

2.3. PILOTER LA PERFORMANCE DANS LA REPETITION DES PROJETS DE SCIENCE CITOYENNE

Penser les projets de science citoyenne comme une organisation scientifique éphémère pose la question de leur intégration au sein des structures traditionnelles (Brabham, 2008; Gallaugher & Ransbotham, 2010; Schlagwein & Bjorn-Andersen, 2018; Wales, 2005). Les projets incluant des foules sont généralement présentés comme un moyen pour les organisations d'accéder à des compétences qu'ils ne possèdent pas en interne (Afuah & Tucci, 2012). Ce qui est produit par ces projets n'est donc pas facilement intégrable à cause de la distance possible avec les compétences internes de l'organisation. Les recherches managériales qui en découlent s'intéressent ainsi à la façon dont les organisations vont effectivement utiliser la production, mettant en place des pratiques managériales adaptées pour maximiser l'efficacité des projets.[6]

En revanche, il n'existe pas à notre connaissance d'études qui se soient penchées sur la question de la répétition de ces projets au sein d'une organisation, notamment la question du transfert d'informations de projets à projets. En effet, les recherches semblent se fixer sur le côté éphémère et donc incompatible avec la répétition. L'aspect éphémère tend à considérer que la nature ponctuelle et non récurrente des activités de projet laisse peu de place à un apprentissage systématique (Hobday, 2002) ou à une répétition systématique (Gann & Salter, 1998, 2000) comparé à des activités à volume élevé (c'est-à-dire où le processus produit le même élément en grand nombre). Or, c'est dans la répétition que nous pensons trouver des redondances dans la façon de gérer ces projets, et éviter de les regarder comme une anomalie.

Un des concepts correspondant à cette question de la répétition est la notion de *capacité dynamique*. Les capacités dynamiques désignent « les processus d'innovation stratégiques utilisés pour adapter, intégrer et reconfigurer les compétences, ressources et routines internes et externes d'une entreprise en réponse à des conditions changeantes et instables » (Eisenhardt & Graebner,

[6] On peut citer les études sur l' *« absorptive capacity »*, c'est-à-dire la capacité à reconnaître la valeur de ce qui est produit et de l'intégrer (e.g. King & Lakhani, 2012; Spithoven et al., 2009).

2007; Winter, 2003; Zollo & Winter, 2003). Ce concept a été appliqué dans la littérature de gestion de projet pour identifier comment les entreprises déploient plusieurs projets pour des clients existants et lancent des projets innovants pour développer de nouvelles technologies et créer de nouveaux marchés (Brady & Davies, 2004; Davies & Brady, 2000; Ethiraj, Kale, Krishnan, & Singh, 2005; Gann & Salter, 2000).

Des travaux remarquables sur cette question dans le cadre de projets temporaires ont été réalisés par les chercheurs Brady et Davis (2000 ; 2004). Ils ont étudié la question de la répétition et de l'apprentissage organisationnel dans la conception de produits et systèmes complexes (Brady & Davies, 2004; Brady, Marshall, Prencipe, & Tell, 2002; Davies & Brady, 2000). Ils définissent le concept de « *project capabilities* » comme principe d'apprentissage organisationnel pour réduire le risque de répéter les mêmes erreurs d'un projet à un autre (Brady & Davies, 2004). Le cas des projets de science citoyenne et de manière plus générale du crowdsourcing peut bénéficier d'un tel cadre théorique pour analyser la transmission d'information de projet à projet. Cependant, les spécificités des projets incluant une foule nécessitent d'étudier comment ce cadre peut s'adapter au cadre proposé par Brady et Davies. En effet, les projets de science citoyenne amènent à plusieurs difficultés pour assurer l'apprentissage : réduction de la collaboration spontanée entre les participants (Levine & Prietula, 2012; Levine et al., 2014), risque d'un manque de transmission d'un projet à l'autre par les participants clés du projet précédent. Comment dans ce cas assurer la transmission d'informations et de ce qui est produit entre les épisodes successifs afin d'améliorer de manière continue et dynamique la performance des projets de science citoyenne ?

2.4. LE RAPPORT AUX DONNEES COMME CADRE IMPENSE POUR OBSERVER LE PHENOMENE

Etudier les projets de science citoyenne comme organisation éphémère demande, entre autres, de caractériser au préalable les activités ou les tâches déléguées aux volontaires. La plupart du temps, une typologie de ces projets est établie en analysant les cas déjà existants (Rick Bonney, Ballard, et al., 2009; Haklay, 2015; Houllier, 2016; Wiggins & Crowston, 2011). Ainsi, les chercheurs distinguent-ils souvent les tâches simples comme collecter les données, des tâches plus intellectuelles et plus incertaines comme résoudre des problèmes (Franzoni & Sauermann, 2014; Haklay, 2015). Cependant, cette approche peut s'avérer restrictive car elle ne reflète pas tout l'ampleur du phénomène : elle est en effet limitée au contexte analysé par le chercheur relativement à son domaine de validité (une discipline spécifique, une zone géographique par exemple). Au contraire, analyser les projets de science citoyenne en fonction du contexte dans lequel ce type d'organisation a émergé semble plus approprié pour entrevoir les enjeux de cette ouverture de la science au public.

Certains auteurs suggèrent que cette démocratisation de la pratique a été largement favorisée par l'apparition de nouvelles technologies de communication et la baisse des coûts de fabrication des

outils de mesure (R. Bonney et al., 2014; Pearce, 2016). Restreindre l'émergence d'un phénomène à une opportunité technique et technologique nous semble cependant insuffisant. Au contraire, nous suspectons que l'apparition des projets de science citoyenne répond à un besoin dans les organisations scientifiques traditionnelles. En regardant un grand échantillon de cas reportés dans les différentes publications et dans les sites spécialisés (voir **annexe 2**), nous constatons qu'il y a une opportunité d'un point de vue des sciences de gestion à analyser les projets de science citoyenne sous l'angle **des données**. En effet, les différentes études menées sur les modes d'ouverture de la science ont constaté implicitement ce lien entre ouverture et données (Franzoni & Sauermann, 2014; Haklay, 2015; Houllier, 2016; Wiggins & Crowston, 2011), sans pour autant l'exploiter. Dans des disciplines comme la biologie, l'environnement, ou le développement durable les chercheurs analysent les projets de science citoyenne selon une grille d'analyse basée sur les activités autour des données : la collecte, la labellisation, l'analyse, ou encore la construction de modèles prédictifs à partir de bases de données (R. Bonney et al., 2014; Wiggins & Crowston, 2011). Par ailleurs, les cas d'utilisation des projets de science citoyenne autour des données se multiplient dans différentes disciplines[7].

De l'autre côté, un ensemble de chercheurs considère que notre époque se situe à l'aune d'une transformation profonde dans la science due à l'évolution de plusieurs caractéristiques des données : leur accessibilité, les méthodes et les techniques pour analyser les données, le lien entre les données et la formulation des hypothèses scientifiques (B. C. Anderson, 2008; Gray, 2009; Kitchin, 2014; Miller, 2010; Shmueli, 2011). Cette transformation ne concerne pas une seule discipline isolée mais impacterait un grand nombre de champ de recherche tel que les sciences environnementales, les sciences de la terre, la santé, mais également la biologie ou l'astronomie (Gray, 2009; Swan, 2013).

L'avalanche de bases de données massives pousse le processus scientifique vers ses limites en terme de capacité de production pour traiter, stocker et coder ces données[8] (Gligorov, 2015; Laney, 2001), est susceptible d'avoir d'autres conséquences sur les méthodes d'analyse des données, mais également sur le modèle de raisonnement scientifique (Kitchin, 2014; Shmueli, 2011). Par ailleurs, certains chercheurs suggèrent que l'avalanche de données massives est susceptible de modifier le processus de génération des hypothèses scientifiques dans un paradigme appelé « *data-driven science* », et donc d'impacter le processus de production de connaissances (Agrawal & Choudhary, 2016; Dubois, Hájek, & Prade, 2000; Kelling et al., 2009; Kitchin, 2014).

---

[7] Le projet Zooniverse, une des principales plateformes regroupant des projets de science citoyenne, fournit un bon proxy pour se rendre compte de ce phénomène. Cette plateforme s'est donnée pour mission *« d'utiliser la « sagesse de la foule » afin de produire des données exploitables et de bonne qualité »*[7]. Des projets de science citoyenne basés sur les données ont émergé dans une grande variété de disciplines comme la biologie (45 projets), la climatologie (9 projets), l'histoire (12 projets), la linguistique (9 projets), la médecine (7 projets), les sciences naturelles (48 projets), la physique (12 projets), les sciences sociales (9 projets), l'astronomie (17 projets)

[8] Le cas de Galaxy Zoo en est un bon exemple. Les scientifiques n'avaient pas les moyens techniques et humains internes pour coder les 900 000 images de galaxies dans un temps convenable.

En passant d'un environnement pauvre en données à une avalanche de données accessibles (Baraniuk, 2011; Miller, 2010), le « *bottleneck* » ou goulot d'étranglement qui conditionne l'amélioration de la production de nouvelles hypothèses change de nature (Sivasubramanian et al., 2003) : on passe d'une contrainte où ce sont les données qui conditionnent l'amélioration de la production de connaissance vers une contrainte où ce sont les capacités de traitement et de formulation de nouvelles hypothèses. Dans ce cadre, il est légitime de se demander si l'utilisation des citoyens de la science ne peut pas s'étendre à la génération des hypothèses pour répondre à ce manque de ressources.

L'ouverture de cette phase à des citoyens de la science est cependant problématique, car comment gérer ou organiser des acteurs sans rétribution ni connaissance *ex* ante de leurs compétences est une question difficile. Alors que multiplier le nombre d'acteurs pour formuler des hypothèses augmente la probabilité d'obtenir des hypothèses intéressantes, la question de la réutilisation durant les projets et entre les projets reste en suspens.

Enfin, la science data-driven interroge sur le rôle du scientifique au sein du processus. Le scientifique est traditionnellement défini comme le garant d'un savoir spécifique à son domaine qui lui permet de choisir les directions de recherche à prendre et de formuler les hypothèses adéquates : c'est lui en effet qui recherche la reconnaissance de ses pairs pour pouvoir obtenir des ressources supplémentaires pour avancer dans sa recherche. Or, nous avons vu que dans la science data-driven, la génération des hypothèses est principalement portée par l'existence des bases de données dès le début du processus scientifique. Ainsi, la place de la connaissance scientifique qui va de pair avec le rôle du scientifique n'est plus aussi déterminante. Durant le 20è siècle, plusieurs tentatives ont été menées pour automatiser le processus de découverte scientifique, remettant en cause place du scientifique (R. D. King et al., 2009; Kulkarni & Simon, 1988; Sen, 2010). En revanche, ces réflexions ont été appliquées dans des environnements spécifiques sans que les résultats soient répliqués dans d'autres contextes. Quel serait le rôle du scientifique dans un processus data-driven où ce sont les citoyens de la science qui génèrent les hypothèses ? Plus généralement, que faut-il pour générer des hypothèses ?

Notre étude des organisations scientifiques éphémères instanciées dans les projets de science citoyenne s'articule donc autour d'un contexte de transformation du rapport du scientifique aux données : comment s'intègrent les formes d'organisation scientifique éphémère dans un contexte de transformation du rapport aux données ? Dans notre thèse, nous nous intéressons uniquement aux disciplines scientifiques qui sont susceptibles d'être impactées par ce changement de paradigme data-driven : ce sont notamment les disciplines qui s'intéressent aux phénomènes naturels comme la physique, la biologie, l'astronomie. Nous intégrons également certaines disciplines en sciences sociales où les chercheurs utilisent des méthodes quantitatives comme la psychologie ou l'économie. Nous excluons par ailleurs toutes les sciences formelles comme la logique, les mathématiques, ou l'informatique théorique. Ainsi quand nous parlerons de

« science », de « scientifique » ou de « discipline scientifique », nous nous référerons exclusivement à ce périmètre de recherche.

Au long de l'itinéraire de recherche, cette problématique se décompose en trois questions de recherche.

3. QUESTIONS DE RECHERCHE

3.1. LES DONNEES COMME FACTEUR D'OUVERTURE DE LA SCIENCE

Le premier élément de questionnement concerne le lien que nous suspectons entre données scientifiques et multiplication des projets de science citoyenne. Le cadre scientifique où positionner les projets de science citoyenne reste encore aujourd'hui impensé. D'un côté, certains épistémologues suggèrent que l'avalanche de bases de données massives ainsi que l'arrivée d'outils techniques pour les analyser (machine learning, intelligence artificielle) mèneraient à l'apparition d'un nouveau paradigme du raisonnement scientifique, appelé « data-driven » (Gray, 2009; Kelling et al., 2009; Kitchin & McArdle, 2016) : dans ce paradigme les données massives sont existantes avant la formulation de l'hypothèse scientifique et servent à sa formulation. De l'autre côté, nous avons remarqué que la plupart des projets de science citoyenne visent à produire des données scientifiques, en particulier des bases de données massives. Le changement de paradigme identifié est néanmoins parcellaire car les épistémologues ne considèrent implicitement qu'un acteur, le scientifique, alors que nous cherchons à étudier l'ouverture de la science à d'autres types d'acteurs à savoir le public. En effet, l'épistémologie s'intéresse essentiellement à ce qu'est une connaissance scientifique : elle étudie les postulats, les méthodes et les conclusions au sein d'une discipline pour en vérifier la portée scientifique et la valeur de vérité que l'on peut attribuer (Matthias, 2018). Le paradigme scientifique au contraire ne se limite pas au seul raisonnement et englobe des éléments comme les méthodes, les pratiques, ainsi que l'organisation scientifique (acteurs, lieux, outils de gestions,…) (Kuhn & Ian, 1962).

Pour être étudié, ce lien entre données et projets de science citoyenne peut être élargi à la question plus générale du rapport entre l'observation scientifique et l'ouverture du processus scientifique. Les chercheurs qui se sont intéressés à ce lien à partir de l'analyse de cas contemporains ou historiques apportent un éclairage intéressant à cette problématique. En effet, il est courant dans les grandes organisations scientifiques contemporaines que les activités comme la collecte ou l'analyse de données par exemple soit réalisée presque exclusivement par des assistants et des laborantins (Hagstrom, 1964). De même, il est rare de croiser des scientifiques qui construisent eux-mêmes leurs instruments de mesure (Price & Tukey, 1963). Pourtant, la division du travail scientifique, sa bureaucratisation et la multiplicité des acteurs dans le processus n'ont pas toujours été comme tel. En fait, les chercheurs estiment la diffusion de cette forme d'organisation aux alentours du 19è siècle (Shapin, 2008; Walsh & Lee, 2015). Auparavant, la pratique scientifique était plutôt hétérogène et la répartition du travail entre les acteurs de la science n'était

pas aussi claire (e.g. Shapin, 1989). De plus, si les savants étaient assistés pour des activités simples comme l'observation, ils conservaient autant que possible les activités stratégiques de l'activité scientifique, comme la validité des instruments, l'analyse des données, la formulation des hypothèses ou la construction de modèles théoriques. Ces transformations concernent généralement des activités autour des données : la construction des instruments scientifiques, la collecte des données, l'analyse des données. Cette évolution dans l'organisation suggère que l'activité scientifique a subi des transformations importantes dans son organisation.

L'évolution de la science et de son organisation a donné lieu notamment à des études en sociologie (Bagla-Gökalp, 1996; Jouvenet, 2009; Latour & Woolgar, 1988; Shinn & Ragouet, 2000; Vinck, 2006) et en histoire des sciences (Berger, 1999; Licoppe, 1996; Matthews, 2016; Schaffer, 1988; Shapin, 1989). Deux périodes nous semblent intéressantes à étudier : l'introduction entre le 17è et le 19è siècle des instruments scientifiques, ainsi que l'apparition de la mesure stochastique de phénomènes naturels dans plusieurs disciplines scientifiques entre le 19 et le 20è siècle. Ces disciplines n'apportent cependant qu'une vision parcellaire de comment les données peuvent impacter le paradigme du processus scientifique : les sociologues étudient les liens entre méthode et pratique mais supposent que le raisonnement reste inchangé ; les historiens analysent des cas de l'histoire des sciences mais ne les confrontent pas au paradigme contemporain. Or, dans notre problématique l'objectif est de construire, s'il existe, un processus causal entre données et ouverture de la science qui nous permette de mieux comprendre le phénomène des sciences citoyennes. D'où notre première question de recherche :

> Q1 : En quoi les données constituent-elles un facteur pour l'ouverture de la science à des citoyens et affecte la place du scientifique ?

3.2. MODELISATION DE LA DELEGATION DE L'ACTIVITE SCIENTIFIQUE

Ensuite, si le lien entre la transformation du rapport aux données et l'ouverture du processus scientifique est avéré, l'organisation de la science va changer de paradigme et ouvrir l'activité scientifique. Pour comprendre et repenser cette organisation, nous avons besoin d'un modèle pour représenter l'activité scientifique et sa délégation notamment par l'ouverture au public. Ce modèle n'a pas pour vocation à représenter toute la science, mais doit nous aider à penser la délégation des activités scientifiques dans un cadre bien précis et restreint pour les disciplines touchées par la transformation des données.

Le modèle de délégation de l'activité scientifique que nous cherchons à élaborer doit prendre en compte à la fois toutes les activités qui sont déjà déléguées dans les projets de science citoyenne et dans les exemples que nous avons pu rencontrer dans l'histoire, mais il doit intégrer les activités qui sont susceptibles d'être déléguées dans le cadre du paradigme « data-driven ». En particulier,

les épistémologues considèrent que le passage vers une science data-driven va impacter le processus de génération des hypothèses. Dans le paradigme scientifique traditionnel, défini comme « knowledge-driven » (Kitchin, 2014), les théories et les connaissances existantes sont utilisées pour orienter le processus de découverte des connaissances. La façon dont les données sont générées ou utilisées est guidée par la formulation d'hypothèses, appuyées par les connaissances et des expériences théoriques et pratiques. Les scientifiques développent des stratégies de génération des données : elles ne sont pas générées par tous les moyens possibles, mais avec un choix méthodologique de générer certaines données et pas d'autres. Le traitement et l'analyse des données est également bien pensé et l'attention des scientifiques se concentre en général sur la voie la plus probable, mettant de côté les hypothèses triviales ou absurdes (Miller, 2010). Dans le nouveau paradigme « data-driven », la construction du résultat scientifique et la génération des hypothèses s'appuie sur l'existence de bases de données massives déjà existantes. Au lieu de construire l'hypothèse uniquement à partir de réflexions intellectuelles et de connaissances sur un sujet scientifique, l'hypothèse se construit en explorant les bases de données déjà existantes. Ainsi, l'hypothèse peut-elle être formulée en même temps que celle-ci est vérifiée par l'analyse de données.

Est-ce que les représentations de l'activité scientifique issues de la littérature sont suffisantes pour modéliser les tâches déléguées dans le paradigme data-driven ? Une littérature en lien avec notre problématique est celle des modèles informatiques de l'activité scientifique. Ces modèles mettent en avant l'existence de deux espaces à explorer lors de la génération des hypothèses (R. D. King et al., 2009; Klahr & Dunbar, 1988; Klahr & Simon, 1999; Kulkarni & Simon, 1988; Schunn & Klahr, 1992). D'un côté, la formulation d'une hypothèse scientifique répond à des raisonnements purement intellectuels qui reprennent en partie les notions d'inférences (abduction, induction, déduction). Ensuite, un autre espace est apparenté à tout activité expérimentale dans laquelle le scientifique agit sur les objets du monde. Cependant, cette représentation est généralement étudiée dans des conditions spécifiques où les espaces sont bornés et connus dès le démarrage de l'exploration. Or, dès que la modélisation n'est pas sur un ordinateur mais analyse un processus humain, il arrive que l'espace sur lequel opère le scientifique soit spontanément « expansionné » (Hatchuel, 2001). De nombreuses découvertes ou constructions comme l'ADN, les nombres imaginaires, ou les quarks, ne peuvent être simplement considérées comme une combinaison de toutes les connaissances préalablement existantes dans la connaissance scientifique et supposent à un moment une extension par ajout d'une nouvelle connaissance.

Ensuite, une fois qu'un modèle des activités qui s'ouvrent dans le cadre des projets de science citoyenne est établi, quels sont les principes de gestion associés pour piloter la performance et la productivité du processus ? Cela nous amène à notre deuxième question de recherche :

> Q2 : Quel modèle élaborer pour se rendre compte de l'impact que provoquent les données sur l'ouverture du raisonnement scientifique?

3.3. GERER LA PERFORMANCE D'UNE ORGANISATION SCIENTIFIQUE EPHEMERE

La réponse à notre première question de recherche va nous apporter des éléments de réponse sur le lien qui existe entre science citoyenne et avalanche des données scientifiques. Nous verrons que la transformation du rapport aux données dans la science a toujours poussé à une ouverture du processus de production scientifique, mais a remis en cause à chaque fois le rôle et les activités des scientifiques au sein du processus. Etant donné que les projets de science citoyenne sont susceptibles d'être appliqués à une activité stratégique comme la génération des hypothèses, le rôle du scientifique est encore une fois remis en cause.

Notre deuxième question de recherche nous apporte ensuite un cadre d'analyse pour étudier les problématiques organisationnelles et de gestion autour des projets de science citoyenne. Cette question nous permet d'étendre le modèle de résolution de problème implicite du crowdsourcing pour y intégrer la notion de tâches couplées inventives. Nous montrons ensuite que la gestion de la productivité des projets de science citoyenne doit prendre en compte la perte de production, à la fois durant et entre les épisodes.

Enfin, la littérature sur la gestion des projets n'est pas suffisante pour prendre en compte l'organisation d'une production scientifique citoyenne. La première limite est que les mécanismes de réutilisation qui peuvent exister au sein d'une foule, notamment dans un cadre généralement bien maîtrisé comme celui du crowdsourcing, ne sont pas très clair. Or, la collaboration entre acteurs demande une forme « d'orchestration » de la part du promoteur du projet pour s'assurer que chaque participant puisse s'exprimer en harmonie avec les autres (Renault, 2014). Quels sont les effets de la réutilisation de ce qui est produit par la foule sur la performance ? Quel est le rôle de la diversité dans le cadre d'une collaboration ?

Une deuxième limite est que le crowdsourcing étudie les projets de science citoyenne comme des épisodes ponctuels et indépendants et non leur continuité dans le temps. Par ailleurs, nous avons vu que si la notion de « project capabilities » correspondait à cette question de transfert d'informations entre les épisodes et d'amélioration dynamique des projets (Brady & Davies, 2004), elle n'a pas été étudiée dans un cadre où les participations au projet sont éphémères. Cela pose plusieurs questions organisationnelles. Comment s'assurer que ce qui a été appris durant un épisode puisse être réintégré dans les épisodes suivants ? Quel pourrait être le rôle du scientifique dans ce contexte sachant que c'est lui qui est habituellement le garant de la connaissance et de la génération des hypothèses ? La performance du public dans le cas où cette activité est déléguée via un projet de science citoyenne est une question qui n'a pas été abordée dans la littérature. Cela nous amène à notre troisième question de recherche :

> Q3 : Comment gérer la performance d'une organisation scientifique éphémère?

Pour résumer, ce travail se construit sur trois principales questions de recherche.

Question de recherche 1 : En quoi les données constituent-elles un facteur pour l'ouverture de la science à des citoyens et affecte la place du scientifique?

Question de recherche 2 : Quel modèle élaborer pour se rendre compte de l'impact que provoquent les données sur l'ouverture du raisonnement scientifique?

Question de recherche 3 : Comment gérer la performance d'une organisation scientifique éphémère?

4. APPROCHE METHODOLOGIQUE

Chaque question de recherche que nous avons identifiée s'appuie sur des littératures différentes pour lesquelles les méthodologies varient. La première question de recherche se base sur une littérature en histoire et en sociologie ; la deuxième question de recherche interroge les modèles de logique de découverte scientifique ; la troisième question de recherche mobilise une littérature en gestion et organisation spécifique. Ainsi, nous avons utilisé diverses méthodes, comprenant des études de cas empiriques, la modélisation formelle, ainsi qu'une étude longitudinale historique pour construire le modèle d'action nécessaire aux projets de science citoyenne selon quatre dimensions : définition des objets, contexte et performance, les méthodes et les processus, les acteurs et l'organisation. La recherche a été conduite dans le cadre d'un contrat de recherche affilié au Centre de Gestion Scientifiques à l'école des Mines ParisTech pendant la période 2015-2019.

1) Valider l'impact des données sur l'organisation de la science : l'étude historique de la science et de son rapport aux données

Pour mieux comprendre le contexte organisationnel contemporain et valider les liens supposés entre projets de science citoyenne et science « data-driven », notre recherche suit une méthode historico-comparative du phénomène (Kieser, 1994). Cette approche a pour objectif de produire un modèle causal entre la transformation du rapport aux données et l'ouverture de la science qui nous permet de justifier le lien supposé dans notre contexte contemporain. Notre étude se concentrera notamment sur l'apparition de nouveaux acteurs, la redéfinition du rôle des acteurs existants (celui notamment du scientifique), et proposera une ébauche de formalisme des différentes étapes du processus scientifique qui ont été ouvertes à de nouveaux acteurs. Nous

analyserons les deux périodes historiques de la science que nous avons identifiées dans les études historiques : l'introduction des instruments scientifiques entre le 17è et le 19è siècle ; l'apparition d'une approche stochastique dans les mesures de phénomènes naturels entre le 19è et le 20è siècle.

2) Modélisation formelle des projets de science citoyenne et de leur efficacité

Pour poursuivre l'exploration des projets d'ouverture de la science via les projets de science citoyenne, un modèle formel ainsi que les méthodes de gestion et les logiques de performance peuvent être établis. La méthodologie comprend : une étude et une critique des modèles informatiques pour représenter le raisonnement scientifique ; l'élaboration d'un modèle basé sur la notion de constructivisme imaginatif développée par Akin Kazakçi (Kazakçi; 2013, 2014). Pour étendre l'ouverture des projets de science citoyenne à la génération des hypothèses basées sur les données, la notion de constructivisme imaginatif permet d'interpréter de manière précise les modèles informatiques développés par des chercheurs pour représenter la génération d'hypothèses, et donc de proposer une interprétation de cette activité.

Ensuite, nous nous sommes appuyés sur des exemples de la littérature (historiques et contemporains) ainsi que sur nos contraintes de systématisation et d'efficacité des projets pour déterminer les métriques liées à la performance des projets en fonction du type de tâche.

3) Deux études de cas pour identifier une figure de gestion adéquate et des structures organisationnelles pour le pilotage des projets de science citoyenne

D'un point de vue méthodologique, nous adopterons dans notre approche une posture épistémologique de « recherche collaborative » (Shani et al., 2008) ou « recherche-intervention » (David, Hatchuel, & Laufer, 2012) menée par des chercheurs et des praticiens afin de créer des connaissances concrètes pour l'organisation et de nouveaux modèles théoriques pour la recherche en sciences de gestion (David & Hatchuel, 2008). Notre étude porte sur deux cas empiriques.

Les cas étudiés dans cette thèse porteront sur deux typologies de projets de science citoyenne autour des problématiques de collaboration et de répétition des épisodes de science citoyenne. Nous étudierons d'abord le cas d'un outil de gestion, le RAMP (pour Rapid Analytics and Model Prototyping), développé par le Centre de Data Science de Paris-Saclay. Cet outil propose de développer des projets basées sur des problématiques et des bases de données fournies par des scientifiques de disciplines variées (économie, biologie, physique des particules,…). Chaque problème est formalisé comme un problème d'optimisation d'algorithme d'analyse de données et soumis à une foule de participants. Cette plateforme est un cas presque unique pour étudier de manière fine avec des métriques quantifiées les effets de la réutilisation dans le cas d'un projet ouvert à la foule

Nous analyserons également un cas unique de délégation de la génération d'hypothèses basées sur les données via Epidemium, un programme de recherche collaboratif basé sur l'épidémiologie, qui

s'est déroulé entre novembre 2015 et mars 2018. Ce projet est le seul cas à notre connaissance où la génération d'hypothèses basées sur les données est ouverte à une foule de participants.

5. ITINERAIRE ET SYNOPSIS DE LA RECHERCHE :

5.1. PARTIE 1 : UN PHENOMENE NOUVEAU ANALYSE DANS UN CADRE IMPENSE

Le point de départ de notre recherche est l'observation d'un phénomène organisationnel nouveau dans la science : les projets scientifiques qui ouvrent le processus de production de connaissance à une foule de contributeurs. Comme nous l'avons souligné, l'intérêt suscité par ce phénomène provient de la multiplication des cas rencontrés dans un nombre grandissant de disciplines scientifiques, de la capacité de plusieurs de ces projets à résoudre des problèmes que les scientifiques n'arrivent pas à résoudre seuls ainsi que du décalage que cette forme organisationnelle opère avec le modèle traditionnel de production de connaissance.

Dans un premier temps, nous cherchons à dessiner les caractéristiques de ces projets. A partir d'une revue de littérature nous caractérisons les projets de science citoyenne, d'abord par rapport aux initiatives d'intégration de publics dans la science (Houllier, 2016), ensuite sur les dimensions d'ouverture qui forment la spécificité de ces projets (Franzoni & Sauermann, 2014). Ainsi ce que nous appelons science citoyenne peut être défini comme des projets de science participative incluant une foule de participants (c'est-à-dire un nombre non défini et non limité de citoyens) pour une tâche à réaliser dans un temps limité.

Les « *success stories* » des projets de science citoyenne suggèrent plusieurs bénéfices en terme de production (Nielsen, 2011; Sauermann & Franzoni, 2014). Comparés au modèle traditionnel de production de connaissance scientifique, les projets de science citoyenne apportent des atouts indéniables en proposant un système de production plus avantageux dans certaines conditions par rapport aux critères classiques de production (coût-qualité-délai). Malgré cela, la littérature existante a surtout considéré la science citoyenne comme un phénomène ponctuel capable d'augmenter de manière éphémère les ressources dans le processus scientifique. En revanche, peu d'attention a été porté sur la question de la performance en terme de productivité dès lors que ces projets sont utilisés de manière récurrente. Nous proposons d'étudier les projets de science citoyenne comme une nouvelle forme **d'organisation scientifique éphémère.**

Nous montrons que le modèle traditionnel de performance de la science n'est pas adapté pour étudier ce mode d'organisation. A la place, la littérature a privilégié le crowdsourcing comme modèle dominant pour analyser les problématiques relatives à la gestion d'une foule de participants (Lichten et al., 2018; Sauermann & Franzoni, 2014; See et al., 2016). Nous montrons que le crowdsourcing a également des limites pour notre étude des projets de science citoyenne. D'abord, il n'existe pas de modèle clair de la performance dès lors que les participants peuvent

réutiliser ce qui a été produit par les autres durant le processus. Ensuite, peu de littérature sur le crowdsourcing ne s'est intéressée à la question de la transmission d'un projet à un autre pour limiter les pertes de production. De plus, les modèles se sont surtout concentrés sur la question de transmission entre le projet de crowdsourcing et l'organisation qui pilote le projet, et pas sur la transmission de projet à projet. Nous proposons à la place de nous appuyer sur le modèle de « project capabilities » développé par Brady et Davies (2004) pour étudier cette organisation.

Dans le deuxième chapitre, nous nous intéressons à la transformation du rapport aux données suggérée par un certain nombre d'épistémologues avec l'arrivée de bases de données massives dans le processus scientifique. Nous montrons que cette transformation est susceptible d'avoir des impacts organisationnels sur le processus de production de connaissance. Notamment, nous montrons l'émergence d'une dichotomie entre deux modes organisationnels de la science. D'un côté le modèle traditionnel, dit « knowledge-driven », qui s'est construit dans un environnement pauvre en données (Baraniuk, 2011; Miller, 2010),. Les mesures de la réalité étaient difficiles, coûteuses et lourdes à obtenir, à stocker et à manipuler. De l'autre côté, l'émergence d'une science dite « data-driven » où les coûts de capture, de stockage et de manipulation des données numériques ont fortement baissé, et les technologies de communication et d'information sont maintenant largement déployées dans les organisations scientifiques (Miller, 2010). Nous suggérons alors que les projets de science citoyenne sont susceptibles d'apporter une réponse organisationnelle à cette transformation, mais que celle-ci nécessite d'être gérée.

5.2. PARTIE 2 - CONSTRUCTION D'UN MODELE THEORIQUE : LA CAPITALISATION COMME SUPPORT D'ORGANISATION DES PROJETS DE SCIENCE CITOYENNE

Avant d'explorer les problématiques managériales des projets de science citoyenne, nous avons besoin d'éprouver notre hypothèse sur le lien entre les données et l'ouverture de la science. La construction d'une réponse à cette hypothèse au chapitre 4 suit une méthode historico-comparative du phénomène (Kieser, 1994). Cette approche a pour objectif de produire un modèle causal entre transformation du rapport aux données et ouverture de la science qui nous permette de justifier le lien supposé dans notre contexte contemporain. Nous nous basons sur deux périodes dans l'histoire de la science : l'introduction entre le 17è et le 19è siècle des instruments scientifiques ; l'apparition de la mesure stochastique de phénomènes naturels dans plusieurs disciplines scientifiques entre le 19 et le 20è siècle. Les observations de ces périodes historiques nous permettront de construire un schéma causal où l'évolution du rapport aux données scientifiques implique une complexification du processus. Les scientifiques, garants de la méthodologie ne peuvent plus assmuer tout le processus de production de connaissance par manque de ressources ou de savoir-faire et doivent déléguer une partie de l'activité stratégique. Historiquement, les scientifiques ont fait preuve de volonté de concentrer leurs activités sur les parties les plus essentielles – construction théorique, génération des hypothèses – et de ne déléguer que des activités dans lesquelles leur valeur ajoutée est minime par rapport aux autres

acteurs. Ainsi nous retraçons la lente délégation de la construction des instruments scientifiques aux artisans et les conflits qui y sont associés. Ensuite, nous montrons comment l'approche stochastique a poussé à intégrer en grand nombre des acteurs pour répondre aux nouveaux besoins en ressources.

Cette analyse historique nous permettra également d'engager un travail de modélisation du raisonnement scientifique à partir des cas de délégation historiques couplés aux exemples contemporains. Nous établirons un modèle basé sur la notion de tâche pour caractériser les différentes activités déléguées au cours de l'histoire : les tâches élémentaires, les tâches de type recette et les tâches de type résolution de problèmes. Nous montrerons que jusqu'à présent, le scientifique a toujours conservé les activités les plus sensibles sans déléguer des tâches comme la génération des hypothèses scientifiques.

Dans le chapitre 5, nous reprenons le travail de modélisation de manière plus formelle. Nous nous interrogeons notamment sur la capacité du crowdsourcing et du modèle de résolution de problèmes à intégrer les activités scientifiques qui peuvent être déléguées aux citoyens de la science. A partir d'une analyse des modèles informatiques du raisonnement scientifique et du modèle du « dual space search » (Klahr & Dunbar, 1988), nous montrons que la génération d'hypothèses ne rentre pas dans le modèle de base. En effet, la génération d'hypothèses est implicitement associée à l'exploration de deux espaces extensibles - un espace pour formuler l'hypothèse et un espace pour réaliser des actions physiques (analyser les données, faire des expérimentations) – tandis que le modèle traditionnel qui découle des simulations informatiques ne contient qu'un seul espace. Nous proposons d'analyser la science comme un régime particulier de conception (Hatchuel et al., 2013) et nous étendons alors notre modèle sur un formalisme à deux espaces, celui du constructivisme imaginatif (Kazakçi, 2014). Ce modèle, en s'inspirant du constructivisme de Brouwer (Kazakçi, 2013) se base sur le principe que tout objet n'existe que s'il existe une méthode qui permet de le construire. Le constructivisme imaginatif suggère que l'on construise en même temps le quoi (la définition de l'objet) et le comment (une méthode pour construire cet objet). Contrairement au modèle de résolution de problèmes, l'objet n'est pas défini conceptuellement avant de réaliser le processus : il se construit en même temps qu'on construit la méthode.

Nous élaborons à partir de ce modèle la notion de **tâche couplée inventive**, qui consiste à formuler un problème (ou état désiré) en même temps que la construction de son plan d'action, et par laquelle nous pourrons analyser la génération des hypothèses. Ce modèle étendu nous permet de construire un modèle de la performance. Nous montrons que l'intérêt majeur des projets de science citoyenne est l'externalisation d'une partie des coûts d'exécution des tâches aux citoyens de la science, réduisant fortement le coût global de la tâche. De plus, les tâches peuvent être parallélisées et donc diviser le temps d'exécution de tâches similaires par le nombre de participants. En revanche, l'ouverture au public des tâches réduit la fiabilité du processus d'exécution de la tâche : en effet, les scientifiques n'ont pas moyen de connaître à l'avance les

compétences des participants ni leur propension à fournir un effort suffisant pour résoudre la tâche. Pour pallier à cette baisse de fiabilité, nous montrons que les managers peuvent mettre en place un système de redondance des tâches.

De manière générale, la question de la répétition de la tâche génère de nouvelles problématiques de performance qui n'existent pas lorsque la tâche est réalisée une seule fois. La perte de production par exemple est un problème courant que l'on retrouve dans tous les types de tâches (e.g. Leifert & Woodward, 2013; Shankar, Mittal, Rabinowitz, Baveja, & Acharia, 2013). Un laborantin chimiste est susceptible de faire de mauvaises manipulations et d'obtenir des produits chimiques non désirables ; un data scientist [9] qui doit résoudre un problème produira potentiellement des lignes de code inutiles avant d'arriver à la solution ; un scientifique va tester plusieurs hypothèses avant d'en trouver une qui soit intéressante et vérifiable. Lorsque les tâches ne sont pas répétées, les pertes de production sont difficilement maîtrisables. Or, dès que la tâche est répétée, le volume de production potentiellement perdu peut devenir important et nécessite donc de mettre en place des modèles de gestion. De manière générale, il est possible d'intégrer de nouveaux critères de performance liés à la répétition que l'on étudie suivant deux angles : la notion de perte d'une tâche à l'autre ainsi que les pertes durant la tâche.

Nous proposons alors un nouveau critère dans les organisations scientifiques éphémères pour piloter les pertes lorsque les tâches sont répétées : la **capitalisation**, c'est-à-dire la capacité de réintégrer dans une tâche toute production, retour d'expérience, outils, ou autres éléments produit antérieurement. La capitalisation se distingue de la notion d'apprentissage dans le sens où il n'est pas possible de compter sur le fait que les citoyens de la science apprennent. En effet, l'important turn-over chez les participants aux projets de science citoyenne implique que la plupart des apprentissages individuels sont perdus et ne peuvent pas être gérés. Au contraire, la capitalisation suppose que tout ce qui a été produit est formalisé et potentiellement réutilisable. Le principe de capitalisation durant la tâche existe déjà dans la gestion de tâches ne nécessitant pas d'exploration (tâches élémentaires et recette dans notre modèle). Dans l'exemple de Galaxy Zoo, la réduction des pertes de la production par les participants a été gérée par l'agrégation des productions individuelles des participants. Afin de réduire la probabilité d'un codage incorrect, les galaxies furent codées plusieurs fois par différents volontaires (environ 50 codages par image) puis agrégées grâce à des outils statistiques pour diminuer le risque d'erreurs. En revanche, la question de la capitalisation durant les tâches de type résolution de problèmes et tâches couplées inventives n'a pas été traité. De même, comment éviter les pertes entre deux tâches successives est une problématique qui reste en suspens.

---

[9] Une définition classiquement utilisée pour les data scientist: "*Data scientists are the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information*". (Davenport & Patil, 2012)

Nous proposons alors deux formes de gestion de la capitalisation qui répondent aux notions de pertes que nous avons identifiées pour les tâches de types résolution de problèmes et couplée inventive. La **capitalisation croisée,** dont le principe est de limiter les pertes durant la tâche en laissant la possibilité pour les participants de réutiliser ce qui est produit durant la tâche. Nous proposons une deuxième forme que nous appelons **capitalisations séquentielle** pour piloter la perte entre des tâches successives (e.g. deux projets de science citoyenne distincts). Dans un premier temps, les participants explorent les espaces en proposant des solutions. A chaque solution proposée, ils améliorent leur connaissance de l'espace et peuvent capitaliser sur cette connaissance pour soumettre des solutions de meilleure qualité. Une fois la tâche terminée, il est possible de cartographier l'ensemble de l'exploration réalisée par chacun des participants. Cette cartographie permet de capitaliser sur ce qui a été produit durant la tâche et de le réutiliser dans la tâche successive.

5.3. PARTIE 3 : IMPLICATIONS MANAGERIALES : COMMENT GERER LES PROJETS DE SCIENCE CITOYENNE ?

La troisième partie de la recherche consiste à replonger le modèle et ses développements théoriques dans le contexte empirique. L'enjeu principal de cette partie est de comprendre à partir de ces cas empiriques les principes en terme de gestion et d'organisation. En effet, la simple rencontre entre des scientifiques et des citoyens de la science autour d'une activité scientifique n'est pas un dispositif autosuffisant en terme d'organisation. Elle nécessite un certain nombre d'instruments de gestion et de conditions de gouvernance qui doivent assurer l'efficacité de la collaboration. Autrement dit, nous avons besoin de comprendre comment le pilotage de la capitalisation s'intègre dans les projets de science citoyenne, les outils de gestion et les figures d'acteurs adaptés pour supporter cette gestion.

Si la mise en place de projets de science citoyenne implique effectivement l'existence de pertes importantes, la gestion de la capitalisation est un enjeu important dans un contexte de standardisation de ce mode d'organisation de la science. Nous avons vu dans la partie précédente que la capitalisation durant les tâches de type élémentaire ou recette peut être pilotée par un modèle de gestion par agrégation. Ce dernier a d'ailleurs été utilisé avec succès dans le cas de Galaxy Zoo. Les cas empiriques que nous analysons dans cette partie serviront à étudier deux situations d'utilisation de la foule que la littérature a négligé : d'abord nous analysons la capitalisation croisée dans le cas d'une tâche de type résolution de problème. Ensuite nous étudions la capitalisation séquentielle dans le cas d'une tâche couplée inventive, à savoir la génération d'hypothèses basées sur les données. Il est à noter que le cas expérimental de tâches couplées inventives que nous analysons est un cas unique à notre connaissance.

Premièrement, nous analysons dans le chapitre 7 les effets de la réutilisation de la production des autres participants durant une tâche de type résolution de problème. En effet, un des éléments caractéristiques des projets de science citoyenne est l'ouverture des résultats intermédiaires

durant l'exécution de la tâche (Franzoni & Sauermann, 2014). De plus, une publication récente montre que laisser les participants réutiliser la production des autres augmente la qualité de la meilleure solution proposée par rapport à un mode compétitif classique (Boudreau & Lakhani, 2015). Cette analyse porte sur un instrument de gestion, le RAMP, qui sert de plateforme virtuelle pour des tâches de type résolution de problèmes dans le domaine des sciences de données. Nous montrons que laisser la possibilité de réutiliser les productions des autres participants durant la tâche fait converger rapidement vers quelques type de solutions (que nous appelons *plateaux de fixation*) et donc diminue la diversité. Cette baisse de la diversité a un impact négatif sur l'efficacité globale (Afuah & Tucci, 2012). Nous constatons en revanche que l'intégration d'une phase « fermée » en amont où les participants ne peuvent pas réutiliser la production des autres (mode compétitif) réduit cet effet. Cela permet d'augmenter à la fois la diversité (durant la phase fermée), la qualité des solutions (durant la phase ouverte), tout en limitant les pertes durant la tâche. Ce processus de réutilisation de la production durant la tâche peut être assimilé à un processus de capitalisation croisée, où la réutilisation de la production par les participants réduit le nombre de solutions non exploitées, et donc les pertes.

Notre deuxième étude dans les chapitres 8 et 9 porte sur Epidemium, un programme de recherche collaborative basé sur l'épidémiologie. Nous analysons deux projets consécutifs mis en place par le programme. En s'appuyant sur le modèle de performance que nous avons développé dans les chapitres 5 et 6, ces projets sont modélisés comme une tâche couplée inventive. Leur analyse permet d'identifier le dispositif organisationnel mis en place par le programme pour capitaliser sur ce qui a été produit durant et entre les tâches grâce à la notion de capitalisation séquentielle développée dans la partie 2. Nous montrons que la capitalisation séquentielle présente un paradoxe : elle nécessite d'évaluer la valeur de ce qui est produit pour capitaliser dessus tout en n'ayant pas de fonction de valeur *ex ante*. Pour pouvoir capitaliser sur ce qui a été produit, il est nécessaire de construire une fonction de valeur *ad hoc*. Celle-ci va permettre à chaque fin de projet d'évaluer la valeur de chaque production et les zones des espaces explorées. Cette évaluation permet de construire une cartographie des zones de l'espace qui ont été explorées par les participants. Cette représentation de la production sert d'outil d'aide à la décision pour décider des futures zones des espaces à explorer. Nous montrons également que cette fonction de valeur peut être étendue entre deux tâches successives par l'introduction d'une nouvelle dimension. Nous suggérons ainsi que la réalisation d'une tâche couplée inventive ne peut pas être pensée comme un projet isolé, mais plutôt comme une succession de projets éphémères.

Ces avancées dans la compréhension des projets de science citoyenne font apparaître un modèle de gestion des projets de science citoyenne. Ce modèle comporte six étapes : 1) Définir la problématique, 2) animer la foule, 3) coordonner la production, 4) évaluer et réintégrer les résultats dans le processus scientifique, 5) piloter le transfert d'informations entre projets, 6) Systématiser l'apprentissage entre projets. La réalisation des étapes nécessite la présence de figures d'acteurs capables de gérer cette logique et de conduire à un pilotage de la productivité. S'il existe, quelles compétences indispensables doit-il avoir pour supporter l'efficacité des projets ? Où

se trouve cette figure managériale dans l'organisation ? Notre étude nous a conduit à définir que l'efficacité des projets de science citoyenne au sein dans les organisations scientifiques pourrait être assurée avec l'aide d'une figure comme « **le gestionnaire de foules inventives**».

Nous pouvons le rapprocher de la figure du manager de projets ou du manager de portefeuille de projets. Cet acteur n'est pas nécessairement expert dans la discipline scientifique, mais il doit être capable de définir la problématique soulevée par l'organisation et la transcrire sous la forme d'un projet de science citoyenne : il doit isoler la tâche du reste du processus, formuler la problématique, mais également définir et fournir les outils et les dispositifs nécessaires à mettre en place dans le cadre du projet, ainsi que de décider les formes de collaboration entre les participants (compétitif, collaboratif, hybride), et de maîtriser les pertes de production durant l'exécution de la tâche. Il est également responsable d'animer la communauté durant le projet et de gérer les systèmes de motivation. Ensuite, dans le cas où le projet s'étale sur plusieurs épisodes, il doit être capable de cartographier l'exploration des espaces réalisée par les participants basée sur une fonction de valeur, tout en étant capable d'identifier l'apparition de nouvelles dimensions de la valeur en fonction de la production. Enfin, il est garant de ce qui est produit par les participants et donc de réintégrer la ce qui est produit dans le processus scientifique.

En outre, cette figure managériale partage certaines des étapes avec le scientifique. En effet, la cartographie de la valeur suffit à piloter l'exploration et ne nécessite pas la connaissance théorique caractéristique du rôle du scientifique. En fait, le scientifique a essentiellement un rôle de garant final et d'évaluateur de l'intérêt scientifique avéré du projet. De la même manière que nous avions observé une remise en cause de la place du scientifique dans notre analyse historique, la fonction du scientifique au sein du processus de production de connaissance évolue : il est demandé au scientifique d'être capable d'identifier la valeur scientifique des hypothèses formulées par d'autres acteurs. C'est le rôle notamment des comités scientifiques que nous avons pu voir dans le programme Epidemium.

Enfin, nous suggérons qu'il y a plusieurs avantages à ce que les projets de science citoyenne soient gérés par des structures intermédiaires. Dans cette forme d'organisation, le gestionnaire des foules inventives n'est pas un employé de l'organisation scientifique mais un des acteurs de la structure intermédiaire. La relation entre scientifique et gestionnaire des foules inventives peut s'apparenter à une relation entre un client et un fournisseur : le scientifique attend des résultats de la part de la structure intermédiaire tandis que le gestionnaire des foules inventives assure la performance de la délégation à la foule.

6. SYNTHESE DE LA THESE

Parties de la thèse	I. Cadre d'analyse de l'ouverture de la science	II. Modélisation du processus : la capitalisation comme élément central de l'efficacité	III. Implications managériales
Question de recherche	**Q1 : En quoi les données constituent-elles un facteur pour l'ouverture de la science à des citoyens et affecte la place du scientifique?**	**Q2 : Quel modèle élaborer pour se rendre compte de l'impact que provoquent les données sur l'ouverture du raisonnement scientifique?**	**Q3 : Comment gérer la performance d'une organisation scientifique éphémère?**
Objectif	Etablir le lien entre transformation du rapport aux données et réorganisation de la science	Construire un modèle de performance pour les activités déléguées de science citoyenne	Pilotage et instrumentation gestionnaire de la performance
Cadre théorique	Littérature en histoire des sciences (Licoppe, 1996; Schaffer, 1988)	Modèle formel de tâche et de résolution de problèmes (Simon & Newell, 1971) Dual-space search (Klahr & Dunbar, 1988) Constructivisme imaginatif (Kazakçi, 2013, 2014)	Crowdsourcing (Afuah & Tucci, 2012), collaboration dans les communautés (Boudreau & Lakhani, 2015), « *project capabilities* » (Brady & Davies, 2000)
Matériau et méthodes	Méthode historico-comparative (Kieser, 1994)	Littérature et cas existants	Recherche-intervention (David et al., 2012): RAMP et Epidemium
Principaux résultats	Lien de causalité entre transformation du rapport aux données et ouverture du processus scientifique Interrogation sur le rôle du scientifique	Modèle formel de tâches déléguées à la foule, notion de tâche couplée inventive pour la génération des hypothèses Modèle de performance basé sur la « capitalisation » pour réduire les pertes	Collaboration pilotée par la possibilité de réutiliser ce qui a été produit Répétition entre les tâches : Cartographie de l'exploration des espaces par évaluation de la valeur des projets Apparition d'une nouvelle figure d'acteur, le gestionnaire des foules inventives

PARTIE 1 – ETUDIER LA PERFORMANCE DES PROJETS DE SCIENCE CITOYENNE DANS UN PROCESSUS « DATA-DRIVEN »

CHAPITRE 1 - LES LIMITES DES MODELES DE GESTION POUR ETUDIER L'OUVERTURE PAR LES PROJETS DE « SCIENCE CITOYENNE »

RESUME DU CHAPITRE 1

Dans ce chapitre, nous cherchons à dessiner les caractéristiques de ces projets. A partir d'une revue de littérature nous caractérisons les projets de science citoyenne, d'abord par rapport aux initiatives d'intégration de publics dans la science (Houllier, 2016), ensuite sur les dimensions d'ouverture qui forment la spécificité de ces projets (Franzoni & Sauermann, 2014). Nous proposons ensuite d'observer les projets de science citoyenne comme une forme d'organisation scientifique éphémère : c'est une organisation sous forme de projets limités dans le temps et où les acteurs du projet n'ont pas d'intérêts directs à ce que le projet aboutisse. Par intérêt, nous entendons un intérêt pour leur carrière (reconnaissance par les pairs), une obligation contractuelle, ou encore un intérêt personnel (concerné par les résultats du processus de recherche).

Nous montrons que le modèle traditionnel de performance de la science n'est pas adapté pour étudier ce mode d'organisation. A la place, la littérature a privilégié le crowdsourcing comme modèle dominant pour analyser les problématiques relatives à la gestion d'une foule de participants (Lichten et al., 2018; Sauermann & Franzoni, 2014; See et al., 2016). Nous montrons cependant que le crowdsourcing a également des limites pour notre étude des projets de science citoyenne. D'abord, il n'existe pas de modèle clair de la performance dès lors que les participants peuvent réutiliser ce qui a été produit par les autres durant le processus. Ensuite, peu de littérature sur le crowdsourcing ne s'est intéressée à la question de la transmission d'un projet à un autre pour limiter les pertes de production. Les modèles organisationnels se sont surtout concentrés sur la question de transmission entre le projet de crowdsourcing et l'organisation qui pilote le projet, et pas sur la transmission de projet à projet. Nous proposons à la place de nous appuyer sur le modèle de « project capabilities » développé par Brady et Davies (2004) pour analyser le transfert d'informations entre projets.

Le but de ce chapitre est de révéler les défis en terme de gestion de la performance que les projets de science citoyenne génèrent et les limites de la littérature pour les étudier. La première section précise comment la littérature distingue les projets de science citoyenne des organisations habituellement rencontrées dans la science. Elle propose ensuite de les définir comme une forme d'organisation scientifique éphémère. Ceux-ci présentent un certain nombre d'opportunité pour la production scientifique mais également des difficultés de gestion. La section deux montre que le modèle de gestion traditionnel utilisé dans la science n'est pas adapté pour étudier ce mode d'organisation. La section 3 présente le modèle du crowdsourcing qui est le modèle dominant pour étudier les projets de science citoyenne. Nous montrons qu'il est également limité pour modéliser la performance des projets quand les participants ont la possibilité de réutiliser ce qui est produit. La section 4 étudie la question de l'amélioration continue entre les projets en se basant sur la notion de « project capabilities », et montre qu'elle aussi ne permet pas de traiter cette question avec une foule de participants.

1. LES PROJETS DE SCIENCE CITOYENNE COMME ORGANISATION SCIENTIFIQUE EPHEMERE

Les premiers efforts pour étudier l'ouverture du processus scientifique à des non scientifiques ont été menés directement par les chercheurs des disciplines concernées, afin de rendre compte de ce qu'ils observaient eux-mêmes comme transformation dans l'organisation de leurs recherches. Ces études sont issues de la littérature en Bioscience, en Biodiversité et en Géographie où les chercheurs font face à une forte augmentation du nombre de projets scientifiques avec de grands groupes de volontaires (Bonney, Cooper, et al., 2009; Cohn, 2008; Silvertown, 2009; Wiggins & Crowston, 2011). Historiquement les projets sont souvent associés à un besoin de main d'œuvre supplémentaire pour les scientifiques afin d'aller collecter des données scientifiques sur le terrain parfois réparties géographiquement dans plusieurs zones différentes. L'expérience montre qu'avec un design adapté et suivant les circonstances, l'ouverture à la foule peut générer des données de haute qualité aboutissant à des résultats scientifiques fiables et valables (Trumbull, Bonney, Bascom, & Cabral, 2000). Si une grande partie des projets dans ces disciplines se limite à la collecte de données, les participants peuvent également participer à d'autres étapes du processus scientifique, comme l'analyse de données voire la co-construction des projets de recherche (Cooper, Dickinson, Phillips, & Bonney. R., 2007; Wiggins & Crowston, 2011; Wilderman, 2007). Cette collaboration varie généralement en fonction de la capacité des participants à être « concernés de manière directe par un problème et mobilisés par leur volonté de mieux connaître des phénomènes qui les concernent, ou d'agir sur leurs conditions propres ou sur leurs environnements proches ou lointains. » (Arnstein, 1969; Houllier, 2016). Depuis ces premières études sur l'intégration de volontaires dans le processus scientifique, plusieurs chercheurs ont proposé de caractériser les projets de science citoyenne de manière plus générale et transdisciplinaire. Deux analyses remarquables permettent de mieux situer ces projets dans le champ des modes de production scientifique : par rapport aux sciences participatives, c'est-à-dire

les projets scientifiques qui font intervenir des citoyens dans le processus scientifique (Houllier, 2016) ; par rapport aux modes d'ouverture du processus (Franzoni & Sauermann, 2014).

1.1. Science participative et « science citoyenne » : caractérisation des projets d'ouverture de la science

Dans un rapport récent, François Houllier, Président-directeur général de l'Institut national de la Recherche Agronomique (INRA) a réalisé un état des lieux des sciences participatives en France et dans le monde, c'est-à-dire des modes d'organisation entre acteurs de la science et membres de la société. Il distingue trois types de projets participatifs (Houllier, 2016) : les projets de *science citoyenne* où les participants sont intégrés sans connaître leurs compétences au préalable ; les projets de type *community-based research* où les participants au projet interviennent activement à différentes activités de recherche dans l'objectif d'améliorer leurs conditions d'existence en produisant des connaissances actionnables ; les *projets de recherche participative*, où la collaboration se met en place entre chercheurs et groupes de citoyens ou de professionnels pour produire des connaissances actionnables et résoudre des problèmes liés à la vie sociale. Plusieurs caractéristiques distinguent les projets de science citoyenne des autres formes de collaboration. Alors que dans les projets de type community-based research et les projets de recherche participative les groupes de réflexion collaborent entre eux en fonction de leurs compétences et de leurs spécificités, la relation entre scientifiques professionnels et citoyens est rarement égalitaire dans les projets de science citoyenne. Les citoyens apparaissent souvent non pas en tant que partenaires du processus, mais plutôt comme une main-d'œuvre accessible révélant une organisation verticale du travail (Sauermann & Franzoni, 2014). De plus, alors que chaque participant dans un projet participatif est clairement identifié en fonction de son apport au processus scientifique et où tous les acteurs voient un intérêt dans l'aboutissement du processus, dans les projets de science citoyenne les scientifiques sont les seuls acteurs du projet définis par leur compétence et leur intérêt à voir l'aboutissement du projet.

Bien que Houllier (2016) restreigne la science citoyenne à la collecte et l'analyse de données, ce mode d'organisation est celui qui correspond le mieux au phénomène que nous avons identifié. En effet, il représente la seule forme de collaboration entre des scientifiques et participants qui ne sont pas définis par leurs compétences ou par leur intérêt dans le projet. Au contraire les projets de type community-based research ou les recherches participatives correspondent à des traditions anciennes de collaboration entre scientifiques et des acteurs qui sont leurs sujets d'étude ou fortement affiliées à leurs problématiques de recherche. Dans les projets de science citoyenne, l'intérêt des participants pour le sujet scientifique peut paraître comme une source de motivation à la participation mais n'est pas une condition pour participer au projet. Par la suite, les acteurs de la foule seront indifféremment appelés « citoyens de la science », « citoyens », « participants » ou « volontaires ».

	Les sciences citoyennes	La community-based research	Les recherches participatives
Objectif	Contribution des citoyens au processus scientifique à l'initiative d'un projet proposé par les scientifiques	Collaboration pour diagnostiquer et résoudre des problèmes qui affectent les participants	Collaboration pour résoudre des problèmes sociétaux
Histoire	Très longue tradition de la participation des amateurs à la production des sciences naturalistes	Tradition longue aux États-Unis, en santé publique, au Canada, en relation avec les communautés indigènes	Tradition longue dans le domaine de la recherche pour le développement. Différentes approches influencées par des traditions intellectuelles différentes (Kurt Lewin, Paolo Freire, Chambers, etc.)
Caractéristiques des participants	Participants constituant une foule indéterminée	Groupes concernés par des problèmes qui les affectent	Groupes de citoyens ou professionnels sollicités pour résoudre des problèmes liés à la vie sociale
Organisation	Position ascendante des scientifiques (organisation verticale)	Collaboration entre chercheurs et groupes	Collaboration entre chercheurs et groupes

Tableau 1. Typologie des projets participatifs issu de (Houllier, 2016)

1.2. TYPOLOGIE DE L'OUVERTURE DU PROCESSUS SCIENTIFIQUE : DEFINIR LA « SCIENCE CITOYENNE » A PARTIR DES DIMENSIONS DE L'OUVERTURE

La notion d'ouverture est souvent utilisée pour définir les projets de science citoyenne (Franzoni & Sauermann, 2014; Lee et al., 2014; Wiggins & Crowston, 2011). Cette notion d'ouverture existe déjà pour caractériser l'organisation scientifique, mais avec une définition différente (David, 2007; Nosek et al., 2015). Elle fait référence dans ce sens à l'accessibilité à la connaissance scientifique par les communautés de chercheurs et par la société au sens large et serait née en Europe au 17è siècle avec l'apparition de la revue académique (David, 2007). La production scientifique était auparavant souvent financée par un système de patronage où ce qui était produit par le chercheur appartenait exclusivement à ceux qui avaient fourni le financement. Avec l'augmentation de l'activité scientifique, la demande d'accès à la connaissance avait atteint un tel point qu'il a fallu que des groupes scientifiques de plus en plus grands, dispersés et spécialisés partagent leurs connaissances et leurs ressources pour pouvoir faire collectivement leur travail.

Dans les projets de science citoyenne, l'ouverture ne concerne pas uniquement la question de l'accessibilité à la production, mais se définit également comme une ouverture du processus scientifique à un nombre non limité de participants. Dans une analyse réalisée sur un grand échantillon d'exemples, Franzoni et Sauermann (2014) identifient deux caractéristiques clés qui différencient les projets de science citoyenne des autres régimes d'ouverture du processus

scientifique. La première est que la participation aux projets est ouverte à un grand nombre de contributeurs potentiels qui généralement ne se connaissent pas entre eux et ne connaissent pas les organisateurs. Une deuxième caractéristique est que les projets de science citoyenne ouvrent une partie des résultats intermédiaires utilisés dans la production de connaissances tels que les bases de données ou les méthodes de résolution de problèmes (**Figure 1**). Il distinguent ainsi quatre régimes de production scientifique en fonction de leur ouverture : la science traditionnelle dite « mertonienne », les tournois d'innovation, la science traditionnelle avec ouverture des données et des publications, et la science citoyenne (ou *« crowd science »*).

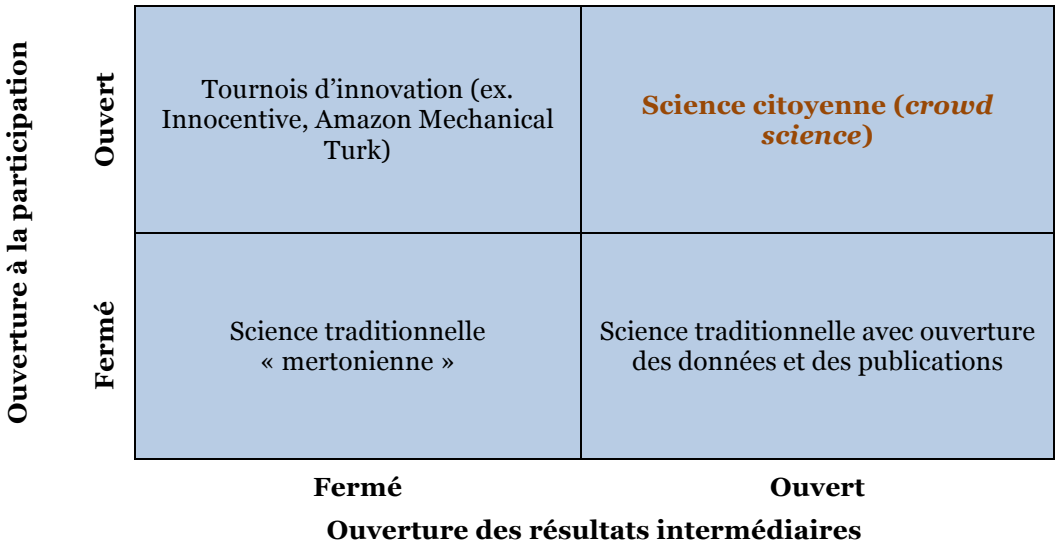


Figure 1. Régime de production de connaissance suivant les degrés d'ouverture de Franzoni et Sauermann (2014)

Les sciences citoyennes se distinguent le plus du quadrant inférieur gauche du schéma, qui se réfère à un régime de science traditionnelle et synthétise la manière dont la science a été réalisée au cours du siècle dernier. L'activité scientifique y est basée sur un modèle où le scientifique évolue dans un environnement constitué d'un ensemble de méthodes, au moyen desquelles la connaissance est produite et certifiée, mais également d'un ensemble de valeurs culturelles et de mœurs non codifiées qui gouvernent l'activité scientifique (Merton, 1942, 1957). Elles garantissent que la production scientifique soit le fruit d'une évaluation impersonnelle, mais également la diffusion des découvertes comme bien public, la recherche non partisane de la vérité ainsi qu'une critique rationnelle qui ne découlent non pas d'une moralité des scientifiques mais d'un dispositif d'incitation basé sur la reconnaissance de la valeur d'une contribution par les pairs (Merton, 1957; Whitley, 2000). Dans ce mode, la notion d'ouverture fait traditionnellement écho au fait que les résultats finaux de l'activité scientifique doivent être librement accessibles à tous (David, 2007). L'accessibilité permet aux scientifiques de produire de nouvelles connaissances qui se basent souvent sur les connaissances déjà produites (Merton & Storer, 1973; Murray & O'Mahony, 2007). Cependant, alors que la science peut être considérée ouverte dans ce sens, elle est largement fermée dans le cadre des deux dimensions présentées dans le modèle. Au contraire, la

fermeture permet aux scientifiques d'assurer une reconnaissance par les pairs ou être les premiers à publier de nouveaux résultats.

Le cadran en bas à droite regroupe les cas où seules les productions de résultats intermédiaires sont divulguées au public. De nos jours, cette question de l'ouverture des résultats intermédiaires est au cœur d'un débat scientifique, où certains chercheurs suggèrent que la production scientifique devrait être plus facilement accessible (Boulton, Rawlins, Vallance, & Walport, 2011; Fecher & Friesike, 2014; Franzoni & Sauermann, 2014; Hand, 2010; Irwin, 2006; Molloy, 2011; Nielsen, 2011). Ces chercheurs considèrent que l'ouverture doit être renforcée sur l'ensemble de l'activité scientifique, sans se limiter à la seule accessibilité des résultats finaux. L'ouverture dans ce cadre est souvent défendue comme un moyen pour résoudre des problématiques au sein de l'activité scientifique, comme les problèmes des délais de publications (Björk & Solomon, 2013) ou les difficultés à reproduire les expérimentations en ayant uniquement accès aux résultats finaux (Ioannidis, 2016; Nosek et al., 2015; West, 2016).

Le quadrant en haut à gauche présente les projets où la participation est ouverte alors que l'ouverture des résultats intermédiaires n'existe pas. Ce mode d'organisation inclut des processus comme Amazon Mechanical Turk où les individus sont payés pour effectuer des tâches telles que la collecte de données sur des sites Web (Kittur, Chi, & Suh, 2008), mais également des plateformes d'innovation utilisées pour résoudre des problèmes scientifiques (Terwiesch & Xu, 2008). Dans ces tournois d'innovation, les organisateurs affichent un problème en ligne et offrent des récompenses monétaires pour les meilleures solutions. Une fois qu'un gagnant est déterminé, il transfère généralement le droit de propriété de la solution au chercheur en échange d'un prix (Jeppesen & Lakhani, 2010). Il n'y a pas de collaboration dans ces tournois et les participants travaillent indépendamment les uns des autres, contrairement à la plupart des projets de science citoyenne. Il est à noter que la collaboration entre participants n'est pas systématique dans les projets de science citoyenne : par exemple le fonctionnement des projets de la plateforme DREAM (**Annexe 2**) s'apparente plus aux plateformes d'innovation où chaque participant travaille indépendamment des autres dans une compétition afin de proposer la meilleure solution.

Les typologies proposées de Houllier (2016) ainsi que de Franzoni et Sauermann (2014) permettent de dégager un ensemble de critères spécifiques aux projets de science citoyenne qui les distinguent à la fois des projets de science participative, mais également des initiatives d'ouverture du processus scientifique: leur ouverture non limitée aux participants, une organisation verticale entre les équipes de scientifiques et les volontaires non scientifiques qui participent au projet, et de manière générale l'ouverture des résultats intermédiaires et notamment la possibilité de collaborer en réutilisant la production des autres. Quels sont les effets de ces spécificités organisationnelles sur les capacités de production de connaissance scientifique ? Nous proposons ci-dessous un cadre pour étudier cette question.

1.3. LES PROJETS DE SCIENCE CITOYENNE COMME ORGANISATION SCIENTIFIQUE EPHEMERE

1.3.1. Avantages des projets de science citoyenne

Les « *success stories* » des projets de science citoyenne suggèrent plusieurs bénéfices en terme de production (Nielsen, 2011; Sauermann & Franzoni, 2014) sur des questions de coût, de temps et de performance. Premièrement, des avantages en terme de coût. Les participants sont généralement mus par des motivations intrinsèques comme le désir de contribuer à la science, plutôt que de rechercher des compensations financières (Domroese & Johnson, 2016; Raddick et al., 2013). Ainsi la plupart des projets de science citoyenne ne comportent pas de rémunération ou de système de prix, diminuant d'autant le coût de réalisation d'une tâche par rapport à un processus plus traditionnel. Ensuite, la délégation des tâches à un nombre important de participants permet, lorsque c'est possible, de lancer un grand nombre de sous-tâches en parallèle et de réduire le temps d'exécution. Le projet Galaxy Zoo a permis de coder plus de 50 millions d'images de galaxies en quelques mois, alors que cela aurait pris 83 ans à un scientifique seul. Enfin, l'intégration de nouveaux acteurs apporte des bénéfices en terme de performance et de réalisation d'une tâche. Pour des tâches ne demandant pas de compétences de haut niveau comme collecter des données, les citoyens de la science sont au moins aussi bons que les scientifiques (van der Velde et al., 2016). Par ailleurs, les sciences citoyennes permettent également de mener à bien des tâches complexes que les scientifiques ne savent pas résoudre eux-mêmes. A partir du moment où les scientifiques diffusent un problème à un grand nombre de contributeurs potentiels, ils augmentent la probabilité d'accéder à des compétences et des connaissances spécifiques nécessaires pour réaliser la tâche, et qui ne sont pas forcément reliées à la connaissance scientifique (Afuah & Tucci, 2012; Jeppesen & Lakhani, 2010; A. King & Lakhani, 2013). Le cas de *Foldit* est un bon exemple. Les scientifiques cherchaient depuis plus de 10 ans à modéliser la structure tri-dimensionnelle d'une protéine du virus du SIDA. Le concours a été lancé, 200 000 personnes ont participé, trois ont fait une proposition pertinente. Cela aura pris moins de trois semaines, le modèle est maintenant utilisé par les biologistes pour étudier les différents sites disponibles sur la molécule (Khatib et al., 2010).

Comparé au modèle traditionnel de production de connaissance scientifique, les projets de science citoyenne apportent des atouts indéniables en proposant un système de production plus avantageux dans certaines conditions par rapport aux critères classiques de production (coût-qualité-délai). Malgré cela, la littérature existante a surtout considéré la science citoyenne comme un phénomène ponctuel capable d'augmenter de manière éphémère les ressources dans le processus scientifique. En revanche, peu d'attention a été porté sur la question de la performance en terme de productivité dès lors que ces projets sont utilisés de manière récurrente (Franzoni & Sauermann, 2014). Pourtant, certains projets ont commencé à systématiser leur utilisation. Par exemple, la campagne réalisée par Galaxy Zoo pour coder les images de galaxies a donné lieu à une nouvelle version améliorée appelée Galaxy Zoo 2 et entrée en service en février 2009. Par la

suite, le principe de codage de données par la foule s'est étendu via la création du portail de science citoyenne Zooniverse : ici les internautes ne sont pas uniquement invités à classer les galaxies, mais peuvent participer à une multitude de projets de science citoyenne, aussi bien en astronomie qu'en biologie, en climatologie, en histoire, en littérature ou encore en médecine[1]. A ce jour, le portail compte 96 projets actifs de science citoyenne. Par ailleurs, d'autres projets comme le DREAM challenges ou RAMP cherchent également à systématiser les projets de science citoyenne en traitant des problèmes variés issus de collaborations avec différents scientifiques. Comment ces projets peuvent s'inscrire dans la durée et assurer leurs utilisateurs de leur efficacité ?

A notre connaissance, il n'existe pas de littérature qui se soit intéressée de manière spécifique à la question de la systématisation des projets de science citoyenne et de leur performance dans le long terme. Etant donné leur intérêt certain, est-il possible de piloter et de contrôler *ex ante* qui permettent de résoudre des problèmes que les scientifiques ne peuvent pas résoudre eux-mêmes ? Grâce à leurs faibles coûts et à leur capacité à fédérer des ressources importantes en très peu de temps, les projets de science citoyenne peuvent fournir une solution performante à de multiples situations rencontrées par les scientifiques. Ainsi, est-il possible de déterminer à l'avance les conditions d'utilisations de ces projets en fonction du besoin des scientifiques ? Quel modèle de performance permet de gérer l'efficacité de ce type de projet ? Nous proposons d'étudier les projets de science citoyenne comme une nouvelle forme **d'organisation scientifique éphémère.** Par éphémère, nous mettons en avant deux caractéristiques des projets de science citoyenne qui ont été négligés dans les études de ces projets : leur limite dans le temps et dans l'espace, et l'intégration d'acteurs éphémères. En effet, alors que le processus scientifique se tient généralement dans des structures dédiées avec une organisation, des normes et un cadre structuré, les organisations scientifiques éphémères sont souvent hors des murs de ces institutions et sont limitées dans le temps. Ensuite, les acteurs d'une organisation scientifique éphémère ne participent que pour le projet et n'ont pas d'intérêts directs à ce que le projet aboutisse. Par intérêt, nous entendons un intérêt pour leur carrière (reconnaissance par les pairs), une obligation contractuelle, ou encore un intérêt personnel (concerné par les résultats du processus de recherche).

1.3.2. Efficacité et perte dans les projets de science citoyenne

Une difficulté immédiate dans ce type d'organisation éphémère est le faible niveau de contrôle que les organisations peuvent avoir avec les participants. Beaucoup d'éléments leurs sont inconnus : les compétences et les expertises des participants, le nombre de participants, mais également le niveau d'engagement des participants au sein du projet. Les projets de science citoyenne font intervenir de nouveaux acteurs au sein du processus scientifique sans aucune relation contractuelle *ex ante* entre eux et les scientifiques (Franzoni & Sauermann, 2014). Dans de nombreux projets, la majorité des participants ne font que des contributions petites et peu

[1] https://www.zooniverse.org/projects?page=1&status=live

fréquentes, s'arrêtant souvent rapidement après leur adhésion. À titre d'illustration sur les 16 400 participants au projet *Old Weather* [2] (projet pour aider les scientifiques à récupérer les observations météorologiques arctiques et mondiales enregistrées dans les journaux de bord depuis le milieu du XIXe siècle) en septembre 2012, seul un petit nombre de personnes ont fait un très grand nombre de contributions, tandis que la plupart des participants ont très peu contribué (Franzoni & Sauermann, 2014). Comme John Timmer le note également[3], la plupart des gens qui participent à un projet de science citoyenne ne reviennent jamais : les projets les plus intéressants ont 60% de leurs utilisateurs qui ne reviennent pas après leur première visite, ce taux pouvant aller jusqu'à 83% dans les cas analysés. Comment dans ce cas peut-on assurer *ex ante* la performance de ce type d'organisation ?

La littérature a déjà étudié ce type de communauté dans d'autres contextes : communautés en ligne (Janzik & Herstatt, 2008), projets open source (Hippel & Krogh, 2003), tournois d'innovations ouverts (Terwiesch & Xu, 2008), crowdsourcing (Afuah & Tucci, 2012). Certains chercheurs qui étudient la performance dans un mode compétitif (c'est-à-dire sans collaboration entre les participants) préconisent un système de performance basé sur les statistiques : plus il y a de participants à un projet, plus il y a de chance que le projet soit une réussite (Afuah & Tucci, 2012; A. King & Lakhani, 2013; Terwiesch & Xu, 2008). Par ailleurs, d'autres chercheurs se sont intéressés aux organisations où les participants collaborent entre eux en s'appuyant sur les autres et en réutilisant ce qui est produit : idées, solutions, avis, écrits,... La foule est un contexte idéal pour partager des informations par opposition à un système classique de collaboration qui est souvent contraint à un ensemble de règles et de rôles à respecter (Anderson & Dron, 2014). Cette réutilisation a été mise en œuvre dans de nombreuses situations comme dans les logiciels open source (Hippel & Krogh, 2003; Murray & O'Mahony, 2007; von Krogh, Spaeth, & Lakhani, 2003), Wikipedia (Kittur, Kraut, & Kraut, 2008), mais également dans les projets de science citoyenne comme le projet Foldit en biologie computationnelle (Nielsen, 2011), ainsi que le projet Polymath pour la construction de preuves mathématiques (Gowers, 2009). Les chercheurs suggèrent que cette réutilisation a des impacts positifs sur la performance du processus (Boudreau & Lakhani, 2015). En revanche, ces effets ont encore peu été étudiés et manquent d'un modèle comme il en existe pour les modes compétitifs (Afuah & Tucci, 2012 ; Terwiesch & Xu, 2008).

Ensuite, la question de la performance des projets de science citoyenne comme système de production doit être considéré dans la continuité. En effet, analyser les projets de science citoyenne comme une organisation scientifique implique de les inscrire non pas comme une forme unique et isolée, mais de s'intéresser à la succession des épisodes. Dans cette perspective, les organisateurs doivent être capables d'assimiler ce qui a été produit durant un épisode pour améliorer les épisodes suivants. En effet, les connaissances produites par les citoyens de la science sont rarement entièrement utilisées par les scientifiques universitaires (Wildschut, 2017). Cela

[2] https://www.oldweather.org/

[3] https://arstechnica.com/science/2015/01/most-participants-in-citizen-science-projects-give-up-almost-immediately/

engendre des pertes importantes de production entre les épisodes et les organisateurs sont susceptibles de ne pas prendre en compte tout ce qui a été produit ou appris durant les épisodes précédents, augmentant le risque de refaire des erreurs ou de reproduire les mêmes éléments. Quel mode de pilotage peut permettre d'assurer cette amélioration continue dans les projets de science citoyenne ?

Dans les sections de ce chapitre, nous analysons comment la littérature en gestion et en organisation de la science et de la foule peut nous aider à répondre à ces questions. Nous montrerons notamment que ces aspects sont encore peu étudiés par les chercheurs dans notre contexte de projets de science citoyenne.

2. MODELE DE PERFORMANCE DE LA PRODUCTION SCIENTIFIQUE TRADITIONNELLE : TROIS ECHELLES DE COORDINATION

Notre projet vise à identifier la particularité des modes de gestion et d'organisation des projets de science citoyenne. Dans cet objectif, le modèle traditionnel de gestion de la science peut sembler le plus naturel au premier abord. L'intégration d'une foule au sein du processus scientifique peut être regardée comme la collaboration des scientifiques avec des volontaires au sein des structures traditionnelles dédiées à la science que sont les laboratoires, observatoires, ou encore les départements de R&D. Peut-on dans ce contexte étudier les projets de science citoyenne dans le modèle de performance existant dans la gestion traditionnelle de la science, notamment sur les questions de collaboration et de continuité ? Nous proposons dans cette section d'étudier les modèles de performance de la production de connaissance scientifique via une triple échelle suivant le type d'acteurs. Le premier niveau va traiter de la performance à l'échelle individuelle (le scientifique). Le deuxième niveau présentera les modes la collaboration au sein d'une structure scientifique. Le troisième niveau étendra les modes de collaboration entre une structure scientifique et des acteurs externes.

2.1. PREMIER NIVEAU : LE SCIENTIFIQUE GUIDE PAR LA RECONNAISSANCE PAR LES PAIRS

Dans le modèle scientifique traditionnel, le scientifique évolue dans un environnement constitué d'un ensemble de méthodes au moyen desquelles la connaissance est produite et certifiée, mais également d'un ensemble de valeurs culturelles et de mœurs non codifiées qui gouvernent l'activité scientifique. Ces valeurs sont transmises par les scientifiques au travers de l'enseignement et de l'exemple (Merton, 1942; Merton & Storer, 1973). L'ouverture, la transparence, mais également l'universalisme, le désintéressement et le scepticisme défendus par les scientifiques assurent la recherche d'une vérité non sectatrice par le chercheur, une évaluation impersonnelle et une critique rationnelle par la communauté scientifique, ainsi que la diffusion libre des découvertes. Ces valeurs ne découlent pas d'une potentielle moralité du chercheur sur le travail qu'il accomplit. En fait, dans certains cas les chercheurs n'hésitent pas à falsifier leurs

productions (Fanelli, 2009). Elles résultent plutôt d'un système d'incitation instillé au sein de la communauté scientifique : la reconnaissance par les pairs de la valeur de la contribution constitue à la fois une gratification pour le chercheur, et en même temps s'avère être l'objectif final à toute production qui garantit l'avancée générale de la connaissance.

Bien que les scientifiques puissent également se soucier d'autres objectifs, cette reconnaissance par les pairs, parfois accompagnée de récompenses comme des titres honorifiques dans les instituts ou des prix (prix Nobel, médaille Fields) pour les plus prestigieux, est essentielle dans leur activité car elle peut générer des avantages indirects : travaux cités par d'autres, sécurité de l'emploi, ressources pour la recherche, financements, augmentation du nombre d'étudiants,... (Latour & Woolgar, 1988; Sauermann & Roach, 2012). Ce système incite les scientifiques à déployer des efforts importants pour être les premiers à divulguer les résultats de recherche avant les autres. En effet, la plus grande partie de la reconnaissance reviendra à la personne qui découvre et publie de nouvelles connaissances. Il y a donc parfois un jeu de concurrence entre les scientifiques où chacun cherche à se munir d'un avantage par rapport aux autres. Ce désir de reconnaissance par les pairs devient l'enjeu majeur qui motive les chercheurs à diffuser leurs résultats auprès de la communauté scientifique et dont ils attendent en retour la gratitude (Hagstrom, 1965). Pour les quelques scientifiques qui obtiennent un prix Nobel, l'impact et la pertinence de leur recherche sont incontestables. Pour le reste cependant, l'évaluation des performances se fait le plus souvent en comptant le nombre de fois où les publications ont été citées par d'autres scientifiques (Egghe, 2006; Hirsch, 2005). Cet indicateur joue souvent un rôle crucial dans la détermination des subventions accordées, la manière dont les candidats à un poste sont classés et même le sort des institutions scientifiques.

Ce système de récompense décourage les scientifiques à aider les autres scientifiques : en fait il explique en partie pourquoi les résultats intermédiaires de la recherche comme les données, ou les stratégies de résolution sont souvent tenues secrètes (Haeussler & Sauermann, 2012; Nosek et al., 2015). Certains chercheurs développent cependant des stratégies de collaboration temporaires qui s'avèrent souvent plus payantes que des tactiques individuelles. L'association à de nombreux auteurs pour publier comme co-auteur d'un papier permet aux scientifiques d'augmenter leur nombre de citations (Yan & Ding, 2009), de jouir d'une plus grande notoriété (Li, Liao, & Yen, 2013). Par ailleurs d'autres études ont cherché à étudier le rôle que jouent les réseaux de collaborations au sein des communautés scientifiques par rapport à la productivité (Balconi, Breschi, & Lissoni, 2004). Par exemple, des recherches ont montré que les réseaux sociaux étaient inextricablement liés à la création de connaissances en influençant à la fois les processus de recherche et de recombinaison (McFadyen et al., 2004; Thompson & Hanley, 2017).

2.2. DEUXIEME NIVEAU : COLLABORATION AU SEIN D'UNE STRUCTURE SCIENTIFIQUE

2.2.1. Organisation de l'activité scientifique dans les structures dédiées

Le travail quotidien du scientifique est rarement une activité isolée, mais repose plutôt sur des échanges permanents, de la négociation et de la persuasion (Latour & Woolgar, 1988). En fait, le produit de la recherche résulte surtout d'un travail d'équipe (Hagstrom, 1964). Le scientifique, qui est généralement rattaché à un lieu dédié comme un laboratoire ou une académie, doit associer la gestion de son évolution individuelle à sa participation aux programmes de recherche de ce lieu. En contrepartie, l'institution l'allège de ses tâches administratives, souvent gérées dans les organisations de recherches publiques par un administrateur de recherche dont le rôle est d'être au service des scientifiques tout en étant le garant des règles administratives (Kaplan, 1959).

La collaboration au sein d'une structure scientifique est souvent considérée comme bénéfique pour la production scientifique, notamment parce que « les groupes sont capables de résoudre les problèmes plus vite et mieux que les individus » (Hagstrom, 1964). Le scientifique travaille en interaction avec des acteurs à qui il délègue une partie de son activité comme des stagiaires, des laborantins, des techniciens ou encore des doctorants (Hagstrom, 1964; Pelz, 1960), et collabore également avec d'autres scientifiques pour produire de la connaissance (Beaver & Rosen, 1979). Cette collaboration entre les acteurs d'une même structure scientifique peut se faire de façons variées dans les organisations scientifiques en fonction de la place que prend le leadership et la répartition du travail. Chompalov, Genuth et Shrum (2002) ont étudié 53 collaborations scientifiques interinstitutionnelles dans les sciences expérimentales afin d'identifier la façon dont les collaborations sont organisées. A partir de leur analyse, ils proposent quatre modèles de collaboration en fonction des connaissances et des compétences des acteurs: bureaucratique, sans leadership, non spécialisé et participative.

1) *Les collaborations bureaucratiques* sont basées sur des règles et des règlements écrits qui gouvernent l'activité, une autorité hiérarchique avec des chefs de projet nommés officiellement, une formalisation des responsabilités ainsi qu'une division du travail. Les scientifiques sont régulièrement soumis à des évaluations externes approfondies, ainsi qu'à la formation de comités. Dans ce modèle, le scientifique est une ressource qui apporte une solution au chef de projet qui en a fait la demande. Le scientifique se doit de produire des connaissances dans son champ d'expertise conformément aux exigences. Ce modèle d'organisation bureaucratique est particulièrement présent dans les départements de R&D entreprises où un contexte économique beaucoup plus tendu implique une forme accrue de compétitivité (Weil, 1999).

2) *Les collaborations sans leadership* ont des responsables administratifs mais pas de responsables pour les activités à proprement scientifiques. Les responsables administratifs sollicitent les contributions des scientifiques et les mettent en charge de projets spécifiques. Ces collaborations ont également des règles formelles et réglementaires pour la participation, ainsi qu'un conseil d'administration.

3) Dans *la collaboration non spécialisée*, la gestion est également hiérarchique mais avec moins de formalisation et de différenciation des rôles et des responsabilités. Plusieurs équipes effectuent des tâches similaires (analyser des données par des modèles standard, labelliser les données). Dans ces collaborations, le leadership scientifique sert à établir et maintenir des normes et les tâches administratives sont partagées entre les membres.

4) *Les collaborations participatives* fonctionnent sans leader scientifique ou administratif. Il n'existe pas de règlement mais plutôt un fonctionnement tacite. Les acteurs publient les résultats collectivement. Ce modèle organisationnel est particulièrement présent en physique des particules, où l'activité scientifique est dispersée entre plusieurs laboratoires dans le monde entier.

Ces modèles d'organisation et de gestion se caractérisent notamment par une distinction entre l'activité scientifique, basée sur des compétences de disciplines spécifiques, et le travail administratif, supporté le plus souvent par des acteurs spécifiques. On voit apparaître également le rôle de manager de projet comme autorité hiérarchique dont le rôle est de piloter le bon déroulement des projets de recherche scientifique (Allen, Lee, & Tushman, 1980; Elias, Cavana, & Jackson, 2002; Liberatore & Titus, 1983). Cette autorité peut avoir différents niveaux d'implication dans l'activité scientifique : dans les processus de *collaboration participative*, la conception et la réalisation du projet est sujet à un jeu de négociation entre les acteurs sans leadership, et le rôle du manager se limite à la gestion de budget et du planning, sans interférer sur les décisions prises au niveau scientifique. Dans des processus plus bureaucratiques, c'est le manager qui décide des directions scientifiques à prendre. Dans ce cas, le scientifique est vu comme une main-d'œuvre qui doit répondre aux besoins du chef de projet en fonction de son domaine d'expertise. Le manager de projets est responsable du bon déroulement du projet, en réunissant notamment les expertises nécessaires à l'avancement de celui-ci. Ces types de projets intègrent également une fonction de gestion des ressources humaines, responsables du recrutement du personnel mais également de leur évolution de carrière au sein du lieu dédié (Allen & Katz, 1986; Cabanes, 2017; Louvel, 2011).

2.2.2. Performance dans les modes de collaboration

En même temps que la grande variété dans les modes d'organisations au sein des structures scientifiques, il y a dans le système académique des variations extrêmes entre les instituts scientifiques en terme de productivité et d'impact sur les recherches ultérieures (Lotka, 1926). Certains chercheurs considèrent que les organisations et les équipes de projet au sein des organisations qui intègrent plus efficacement leurs diverses compétences seront plus performants (Grant, 1996). Les théories des organisations distinguent entre autres quatre activités de coordination qui aide à mieux utiliser les expertises au sein des organisations scientifiques et mènent à de meilleurs résultats (Cummings & Kiesler, 2007).

Une première activité de collaboration implique de diviser le travail à réaliser afin d'attribuer la responsabilité des tâches à exécuter en fonction des expertises et des spécialistes appropriés. Un projet dans lequel la gestion de la coordination est attribuée à plusieurs individus différents ou groupes permet de limiter la dépendance excessive entre les individus et de réduire les problèmes de communication (Weick, 1979). Ce mode de collaboration a également des effets positifs sur la productivité en réduisant les coût de communication directe entre les acteurs (Porac et al., 2004). Une deuxième activité de collaboration concerne le partage de ressources: cela peut être un site web, une base de données partagée ou encore des outils d'analyses ou des logiciels informatiques. Ce recours à des ressources communes permet à chacune des parties prenantes de pouvoir accéder à des ressources qu'elles ne possèdent pas (Powell, Koput, & Smith-Doerr, 1996), mais également de maintenir les valeurs collaboratives tel que le bien-être des partenaires de la collaboration, la distribution équitable des récompenses ainsi que la qualité des relations (Appley & Winder, 1977; Jap, 2003). En effet, une approche de la collaboration en terme d'équité indique que les personnes estiment qu'un résultat est juste lorsque le rapport entre les ressources et les productions propres est égal au ratio ressources / production de celui des autres (Adams, 1966; Walster, Walster, & Berscheid, 1978). Une troisième activité de collaboration, la plus courante est la communication directe entre les participants par le biais de réunions et de discussions spontanées. Une communication plus fréquente est associée à une plus grande confiance ainsi qu'au respect et à des normes participatives (Cummings & Kiesler, 2007). Enfin une dernière activité de collaboration consiste à apprendre et à transférer des informations et des connaissances pour des effets potentiellement synergiques. Le partage des connaissances scientifiques est essentiel au progrès scientifique, à tel point que la norme mertonienne de partage inconditionnel des connaissances est considérée comme l'une des caractéristiques déterminantes de la vie universitaire (Merton, 1973). Les publications sont divulguées dans le domaine public (revues académiques, bases de données sur papier de travail), après quoi d'autres peuvent réutiliser leur contenu et leurs idées en les citant dans leurs propres publications.

2.3. TROISIEME NIVEAU : COLLABORATION INTER-STRUCTURES ET AVEC DES ACTEURS HORS DE LA STRUCTURE SCIENTIFIQUE

2.3.1. Collaboration multi-institutionnelle

Enfin, la recherche scientifique est de plus en plus distribuée et les collaborations scientifiques impliquent de nombreuses institutions (Corley et al., 2006). Ces collaborations massives entre scientifiques sont souvent associées à la vague relative à la multiplication des projets de grande envergure, ou « Big Science », apparus peu après la Seconde Guerre mondiale où les scientifiques ont collaboré pour optimiser les coûts d'équipements scientifique onéreux et former des spécialistes (de Solla Price, 1963). Depuis l'avènement des réseaux informatiques, ce type de collaboration se multiplie et permet aux scientifiques de différentes régions géographiques de pouvoir collaborer ensemble et partager des ressources communes (Kouzes et al., 1996). Par ailleurs, ces collaborations entre scientifiques ont des effets positifs sur la qualité de la production issue de la collaboration mais également sur les stratégies de recherches individuelles des scientifiques (He et al., 2009).

Dès lors que plusieurs universités ou centres de recherche sont impliqués dans un projet commun, la complexité de la coordination entre les différents acteurs augmente (Hagstrom, 1964). Par exemple, les conversations spontanées et informelles issues d'un contexte social partagé sont réduites par la distance entre les acteurs (Kiesler & Cummings, 2002). Comparés à des projets d'une seule université, les projets avec des chercheurs de différentes universités auront probablement plus de difficulté à créer un terrain d'entente (Clark & Brennan, 2004), faciliter une prise de conscience de ce que font les autres (Weisband, 2002) et s'adapter rapidement aux surprises (Olson & Olson, 2000). Les améliorations technologiques offrent des possibilités de collaborer de nouvelles manières, mais ne facilitent pas forcément la gestion de cette collaboration et la coordination. Dans des études de projets de recherche où les acteurs sont répartis géographiquement dans différents instituts, les chercheurs ont montré un retard dans les projets (Herbsleb et al., 2004), des rivalités entre institutions (Armstrong & Cole, 2004) ainsi que des échecs en matière de partage d'informations et de communication efficace (Hinds & Mortensen, 2005).

Dans une étude sur 491 collaborations de recherche scientifique, Cummings et Kiesler (2007) montrent que plus il y a de parties prenantes impliquées dans un projet inter-institutions, plus il sera difficile de gérer la collaboration et la coordination entre les acteurs du projet et que cela aura un impact sur la performance de la collaboration scientifique. Au contraire, des moyens de coordonner comme la division des responsabilités entre les acteurs par exemple sont des facteur essentiel pour estimer la qualité et le type de production à la sortie (Cummings & Kiesler, 2007).

2.3.2. Collaboration dans les sciences participatives

Dans la plupart des modes de collaboration que nous avons présenté, les acteurs du processus de production de la connaissance sont identifiés par leurs compétences et leurs activités plus ou moins prédéfinies au travers de relations contractuelles. Cependant, dans certains cas, les acteurs des lieux dédiés ne sont pas les seuls concernés par les problématiques scientifiques (Bonney, Cooper, et al., 2009; Houllier, 2016; Israel, Schulz, Parker, & Becker, 1998). Parfois définis comme « *community-based research* » ou recherche participative (Houllier, 2016), ces projets sont principalement différenciés par la participation et l'influence de personnes non académiques dans le processus de production de connaissance. Des projets sociétaux sur des maladies graves ou des décisions territoriales qui dépassent le cadre d'une seule discipline font interagir à la fois des spécialistes de différentes disciplines scientifiques (Callon, Lascoumes, & Barthe, 2001), mais également de nouveaux acteurs, souvent volontaires, concernés par le problème. Ces volontaires sont intégrés dans le processus à cause de leur implication vis-à-vis de la question proposée par les chercheurs. Ils sont alors intégrés au processus non pas en fonction de leurs capacités à répondre au problème scientifique, mais par rapport à l'histoire singulière qui les lie à la question scientifique. Dans une étude sociologique, Callon, Lascousmes et Barthe (Callon et al., 2001) présentent différents exemples mettant en jeu des problématique scientifique dans des contextes sociaux (implémentation d'un site nucléaire, stratégie de recherche pour une maladie). Ils mettent en évidence des problématiques de gouvernance qui poussent à l'émergence de modèles de coopérations entre experts et groupes concernés "profanes". Les chercheurs soulignent l'apparition de modèles d'organisation démocratiques qui diminuent l'impact de la place dominante du sachant, le scientifique, vis-à-vis des profanes, et forment des jeux de négociation entre les différents acteurs. Un rapport technique, axé sur l'éducation, analyse également les niveaux de participation de volontaires dans le cadre de projets de recherche liés aux biosciences (Bonney, Ballard, et al., 2009). Dans ces projets, les participants volontaires apportent aux scientifiques des connaissances spécifiques, telle que la connaissance du terrain ou des connaissances techniques. Leur intégration dans le processus de production de connaissance fait suite à un entraînement spécifique aux méthodologies scientifiques liées à l'activité qu'ils vont mener.

2.4. UN MODELE DE PERFORMANCE DE LA PRODUCTION SCIENTIFIQUE BASE SUR UNE FIABILITE DES ACTEURS

Nous avons vu que les modèles de gestion et d'organisation du processus scientifique sont protéiformes. Ils diffèrent en fonction de la taille de la structure étudiée, mais également suivant la discipline concernée et le type d'acteurs impliqués. Dans tous ces modèles, le processus de production de connaissances scientifiques fait intervenir différents acteurs : des scientifiques de disciplines différentes et des managers de projet, mais aussi des non-scientifiques, comme des techniciens, des laborantins ou des ingénieurs de recherche, des administrateurs ainsi que des non académiciens (entreprises, financeurs, volontaires). Tous ces acteurs se regroupent

généralement au sein d'une structure dédiée à la production de connaissances scientifiques où tous ont un rôle à jouer plus ou moins bien déterminé. Lorsque les structures sont de petite taille, c'est-à-dire que le nombre d'acteurs est faible (jusqu'à quelques dizaines), la répartition du travail se fait souvent de manière informelle et est supportée la plupart du temps par les scientifiques. La gestion est essentiellement basée sur la communication entre les parties prenantes. Les scientifiques pilotent eux-mêmes leur activité basée sur un système d'incitation global à produire de la connaissance pour recevoir en retour la reconnaissance de leurs pairs, des financements, des ressources,... Les non scientifiques sont reliés contractuellement à la structure et leurs motivations sont d'abord de faire avancer la science, mais également sur le désir de conserver leur emploi, voire d'obtenir des évolutions de carrière.

Dans le cas où les collaborations réunissent un grand nombre d'acteurs, les activités assignées à chaque acteur sont généralement plus formalisées et répondent à un modèle de collaboration bureaucratique, où chaque acteur contribue à la production suivant ses connaissances et ses compétences. On peut voir apparaître des managers de projets qui dirigent à la fois le travail des non scientifiques (laborantins, techniciens) mais également celui des scientifiques. Si le travail des non scientifiques répond toujours à un modèle bureaucratique, l'activité des scientifiques peut elle être organisée via un système de collaboration participative où la conception et la réalisation du projet est sujet à un jeu de négociation entre les acteurs sans leadership, et le rôle du manager se limite à la gestion de budget et du planning, sans interférer sur les décisions prises au niveau scientifique. Enfin dans les projets de type community-based research, les scientifiques collaborent avec des non académiciens durant le processus (entreprises, volontaires). La répartition du travail entre ces acteurs répond plutôt à un jeu de négociation en fonction des compétences et des connaissances de chacun.

Processus de production	Acteurs	Modèles de gestion	Modèle organisationnel
Scientifique seul	Scientifique	Evolution de carrière	Incitatif (reconnaissance par les pairs)
Laboratoires de petite taille	Scientifique, Laborantin, Doctorant, Stagiaire	Coordination, Répartition des tâches	Communication
Grands laboratoires, projets multi-institutions	Scientifiques de plusieurs disciplines, Laborantin, Doctorant, Stagiaire	Management de projet, Evolution de carrière, Répartition des tâches	Emergence de coordinations démocratiques, Distinction des rôles administratif, scientifique et de gestion de projets
Science participative	Scientifiques, non scientifiques	Intégration des non scientifique	Consultatif, Négociation, Pré-apprentissage

Tableau 2. Modèles organisationnels de la science suivant les processus de production.

Les modèles de performance de production scientifique reposent en grande partie sur une fiabilité des acteurs au sein du processus. Cette fiabilité peut se traduire par la connaissance *ex ante* des compétences des participants, ainsi que leur motivation à aller au bout du projet : objectif de reconnaissance par les pairs pour les scientifiques, relations contractuelles, implications personnelles dans le projet. L'ouverture du processus scientifique à la foule fait intervenir un nouveau type d'acteur, les « citoyens de la science » dont le manque de fiabilité ne permet pas d'appliquer les mêmes modèles de performance que ceux habituellement développés dans la science.

1) *Non-connaissance ex ante des compétences*. Le citoyen de la science n'est pas défini par ses compétences ou ses connaissances utiles dans le cadre de l'activité scientifique. Le scientifique qui fait appel aux citoyens de la science n'a pas connaissance des compétences des acteurs avant que ceux-ci aient réalisé la tâche. Cette caractéristique est unique vis-à-vis des acteurs traditionnellement inclus dans le processus scientifique.
2) *Pas de limite dans le nombre de participants*. Une des caractéristiques propres aux projets de science citoyenne est de ne pas limiter le nombre de participants au projet et donc de ne pas savoir combien de personnes participeront, rendant d'autant plus complexe d'anticiper l'organisation.
3) *Pas de relation contractuelle*. Alors que dans un processus de production de connaissance scientifique le travail est généralement réparti entre les différents acteurs suivant des relations contractuelles *ex ante*, il n'existe aucune obligation du citoyen de la science vis-à-vis du scientifique de réaliser la tâche. Il y a alors un risque dans l'ouverture du processus que la tâche ne soit pas réalisée ou du moins mal exécuté.
4) *Pas de recherche d'évolution de carrière*. Les citoyens de la science ne recherchent pas une carrière au sein des structures scientifiques : alors que le recrutement d'un chercheur, d'un laborantin ou d'un technicien est un investissement de long terme pour un laboratoire, la relation avec les citoyens de la science est beaucoup plus courte et diminue les difficultés d'engagement pour les directeurs de recherche. En contrepartie, l'ouverture de la science pose d'autres questions sur la gestion de la communauté pour pérenniser l'effort fourni par les citoyens.
5) *Pas de recherche de reconnaissance par les pairs*. Les citoyens de la science ne cherchent pas une reconnaissance par la communauté scientifique pour leur travail, et leurs actions sont régies par d'autres formes d'incitations.

Bien que des projets scientifiques comme les collaborations multi-institutions ainsi que les projets incluant des volontaires ou profanes se rapprochent de la situation que l'on retrouve dans les projets de science citoyenne, la spécificité des acteurs – les citoyens de la science – rend inconsistant les modèles traditionnels pour étudier les contraintes spécifiques de gestion relatives à ce type d'organisation. Au contraire, la littérature a préféré s'appuyer sur le modèle du crowdsourcing pour étudier ces projets incluant des foules.

3. *LE CROWDSOURCING COMME MODELE DE PERFORMANCE DOMINANT POUR ETUDIER L'OUVERTURE A LA FOULE*

Dans la littérature en gestion des projets de sciences citoyennes, la plupart des acteurs s'accordent à traiter ce phénomène en lien avec le *crowdsourcing* (Buecheler et al., 2010; Erickson, 2011; Franzoni & Sauermann, 2014; Haklay, 2015; Lichten et al., 2018; Rotman et al., 2012; See et al., 2016; Swan et al., 2010; Wiggins & Crowston, 2011). On observe en effet un nombre croissant de publications où les scientifiques revendiquent une utilisation avec succès du processus de crowdsourcing pour résoudre leurs problèmes (Azencott et al., 2017; Ewing et al., 2015; Griffith et al., 2017; Saez-Rodriguez et al., 2016; Warby et al., 2014). Suivant une première approche, le crowdsourcing est le fait d'intégrer un ensemble non limité de personnes non sélectionnées dans un processus afin d'exécuter une tâche sans aucun engagement pour accomplir cette tâche (Afuah, 2018; Afuah & Tucci, 2012; Howe, 2006). L'exécution de la tâche ne se limite donc pas à une personne en particulier ou à un groupe de personnes pré-identifiées, mais est ouverte à tous. Le terme a donné lieu à beaucoup de définitions dans la littérature (Estellés-Arolas & González-Ladrón-De-Guevara, 2012), qui peuvent mener à des interprétations différentes du phénomène (Hossain & Kauranen, 2015), cependant la définition la plus utilisée pour parler du crowdsourcing est sans doute celle proposée par Howe (2006) :

> *«Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call »*

Le crowdsourcing peut être compétitif, collectif ou collaboratif (Afuah, 2018). Dans le modèle compétitif, chaque participant génère seul une solution au problème. Ainsi, plus il y a de participants, plus il y aura de solutions proposées. La plupart des tournois d'analyse de données comme *Kaggle*, *TopCoder* ou *DREAM* sont basés sur ce modèle. Dans le mode collectif, chaque participant soumet uniquement une partie de la solution au problème, et ces composants sont ensuite agrégés pour obtenir la solution finale. C'est le cas de Galaxy Zoo où les participants codent les galaxies individuellement pour que les résultats soient ensuite compilés. Enfin le modèle collaboratif est un hybride des deux. Par exemple, dans le cas de Polymath les différents volontaires interagissent au travers d'un blog où chacun, indépendamment de ses compétences dans le domaine, peut proposer une piste ou une idée ou rebondir sur les idées des autres.

Le crowdsourcing est construit sur un modèle organisationnel de type « seeker-solver » (Jeppesen & Lakhani, 2010; Sieg, Wallin, & von Krogh, 2010) ou d'appel au marché (Terwiesch & Xu, 2008). Il comporte quatre étapes majeures (Afuah & Tucci, 2012). Tout d'abord, les seekers qui sont généralement les responsables d'une organisation formulent un problème bien délimité qu'ils diffusent ensuite à un ensemble indéfini de personnes sous la forme d'un appel ouvert. Deuxièmement, les participants qui souhaitent résoudre le problème choisissent de le faire sans qu'ils soient spécifiquement assignés à cette tâche par l'organisation. Chaque solver travaille

ensuite de façon autonome sans avoir la garantie qu'il puisse fournir la solution au seeker. Troisièmement, lorsque le crowdsourcing est en mode compétitif, les solutions générées sont évaluées, souvent par une métrique, pour déterminer un gagnant. Dans certains cas, les participants peuvent avoir accès à un tableau de classement qui leur permet de connaître leur position en fonction des solutions des autres. Dans les modes collectifs ou collaboratifs, les différentes solutions soumises sont agrégées pour obtenir la solution. Enfin la dernière étape du crowdsourcing concerne l'implémentation de la solution au sein du processus scientifique (Lüttgens et al., 2014; Piezunka & Dahlander, 2015).

3.1. APPLICATIONS DU CROWDSOURCING

Le crowdsourcing est une nouvelle manière d'organiser une partie d'un processus sans le confier à des personnes en interne ou à des personnes connues pour leurs compétences, mais plutôt en le déléguant à une foule. Différents types d'applications ont été étudiés dans la littérature pour lesquelles on retrouve des similarités dans les projets de science citoyenne : réaliser des micro-tâches, résoudre des problèmes, générer de nouvelles idées.

Certains chercheurs considèrent le crowdsourcing comme un moyen pour vendre et acheter de la main-d'œuvre à la demande pour réaliser des micro-tâches rémunérées ou non sur des temps courts (Kittur, Chi, et al., 2008). Ces tâches ne nécessitent en général pas de compétences particulières, et peuvent être réalisées par n'importe qui. Elles sont conçues à partir d'une tâche plus importante, comme la traduction d'un rapport ou d'un livre, qui est décomposée en un ensemble de micro-tâches indépendantes (Olsen & Carmel, 2013). Ce type d'utilisation du crowdsourcing met en général en jeu trois types d'acteurs : en plus du seeker et des solvers, un opérateur de plateforme s'assure de la transmission des tâches et de leur réalisation (Hossain & Kauranen, 2015). Le seeker formule son besoin qu'il traduit par un ensemble de micro-tâches et rémunère les solvers lorsque les tâches sont réalisées. L'opérateur de plateformes fournit quant à lui une infrastructure, souvent une plateforme web, qui sert de médiateur entre le seeker et les solvers. Les solvers eux, accèdent à la plateforme dans laquelle sont catégorisées un ensemble de micro-tâches à exécuter. Certains projets de science citoyenne présentent des similitudes avec le crowdsourcing pour les micro-tâches (Haklay, 2015). Dans le projet Galaxy Zoo par exemple, le codage des images de galaxies est divisé en un ensemble de micro-tâches indépendantes (labelliser une image) qui est ensuite facilement transmissible aux citoyens de la science. Pour rappel, le projet a permis de coder plus de 50 millions d'images de galaxies en quelques mois (Mao, Kamar, & Horvitz, 2013). Les mécanismes de délégation de la tâche et les avantages en terme de réduction de temps coïncident avec modèle développé du crowdsourcing de micro-tâches.

D'autres chercheurs voient le crowdsourcing comme une nouvelle forme d'organisation capable de résoudre des problèmes (Afuah & Tucci, 2012; Brabham, 2008b; Terwiesch & Xu, 2008). Un tel problème pourrait être de trouver des données compatibles avec une hypothèse scientifique,

trouver un algorithme capable d'analyser les données scientifiques pour répondre à une question de recherche, détecter des formes spécifiques sur des images, diagnostiquer une maladie rare. Si un manager connaît quelqu'un qui peut résoudre son problème, il peut assigner ce problème sous forme d'une tâche à cette personne, à l'intérieur ou à l'extérieur de l'organisation, sous forme d'un contrat ou d'un engagement (Malone, Laubacher, & Dellarocas, 2010). Dans le cas contraire, il peut utiliser le crowdsourcing, c'est-à-dire faire sous-traiter le problème sous la forme d'un appel ouvert sans que cette tâche soit affectée spécifiquement à quelqu'un que le manager pense susceptible de résoudre la tâche. Dans ce cas, il n'existe pas de contrat *ex ante* ou engagement tenu pour accomplir la tâche (Afuah, 2018). Chaque participant développe sa propre stratégie individuelle ou collective pour résoudre le problème, en fonction de ses connaissances et de ses compétences. Le seeker n'a pas besoin d'évaluer les compétences nécessaires pour résoudre le problème : ce qui peut être un problème insoluble pour lui s'avèrera potentiellement être résolvable pour un des seekers (Afuah & Tucci, 2012). Ainsi, l'augmentation du nombre de participants pour résoudre le problème va augmenter le nombre de voies explorées pour trouver la solution, et donc améliorer la probabilité de trouver une solution (Afuah & Tucci, 2012; A. King & Lakhani, 2013; Terwiesch & Xu, 2008). La tâche que les participants doivent résoudre doit cependant respecter un ensemble de critères afin que le processus de crowdsourcing soit efficace (Afuah & Tucci, 2012) : le problème doit être facile à délimiter et à isoler du reste du processus ; les connaissances requises pour résoudre le problème doivent être cherchées hors de l'organisation ; plus il y a de savoir-faire dans la foule évaluation, plus il y a de chances que le problème soit résolu ; la solution doit pouvoir être évaluée par un grand nombre de personnes.

Enfin, un autre courant de recherche considère que le crowdsourcing apporte un moyen pour utiliser la créativité de la foule afin de proposer de nouvelles idées (Poetz & Schreier, 2012; Schemmann, Herrmann, Chappin, & Heimeriks, 2016). Alors que l'innovation est traditionnellement supportée par des acteurs internes (Schulze & Hoegl, 2008), beaucoup d'organisations externalisent le processus d'idéation par le biais du crowdsourcing dans l'espoir de pouvoir récolter de nouvelles idées. En effet, un certain nombre d'études suggèrent que l'externalisation du processus d'idéation permet d'améliorer la performance des idées générées en termes de quantité, mais également en termes de nouveauté et d'originalité (Poetz & Schreier, 2012b; Schweitzer et al., 2012). Poetz and Shreier (2012) ont démontré notamment que les foules peuvent surpasser les professionnels dans la génération d'idées de nouveaux produits. Le seeker peut tirer parti du fait d'avoir accès à une large population de solvers qui lui permet d'augmenter la diversité des solutions proposées (Surowiecki, 2004; Terwiesch & Xu, 2008).

3.2. MODELE DE PERFORMANCE DES PROJETS BASES SUR UNE FOULE

Le modèle traditionnel de performance du crowdsourcing et plus généralement des projets incluant une foule est souvent défini comme la combinaison de l'effort fourni par les participants durant le processus, de leur capacité à pouvoir résoudre le problème ainsi que du niveau

d'incertitude lié à la résolution du problème (Terwiesch & Xu, 2008). La recherche s'est donc concentrée sur ces aspects. Un premier champ de recherche porte sur la maximisation de l'effort fourni par les participants. En effet, comme les solvers ne sont pas liés aux seekers par une relation contractuelle leur participation dépend essentiellement de leur motivation. Les seekers peuvent alors mettre en place des incitations financières pour récompenser les meilleures solutions et fournir une motivation aux participants (DiPalantino & Vojnovic, 2009; Zheng, Li, & Hou, 2011). Le deuxième élément concerne la compétence des participants à trouver une solution au problème. Les chercheurs ont montré que lorsque l'incertitude à résoudre le problème est importante, les seekers devraient baisser les barrières à l'entrée pour maximiser le nombre de solutions proposées (Terwiesch & Xu, 2008). En effet la recherche d'une solution est souvent sujette à de « mauvais chemins », de « l'expérimentation », de la « sérendipité » et de « l'incertitude » (Abernathy & Rosenbloom, 1969; Boudreau, Lakhani, & Lacetera, 2008; Loch, Terwiesch, & Thomke, 2001). La découverte d'une solution à un problème dépend fortement des compétences et du savoir-faire de la personne qui va résoudre le problème (Afuah & Tucci, 2012). Ainsi l'augmentation du nombre de participants permet de multiplier le nombre de points d'entrée et donc augmenter la probabilité de trouver la meilleure solution (Afuah & Tucci, 2012; Boudreau et al., 2011; King & Lakhani, 2013).

Pour que le modèle du crowdsourcing soit efficace, les organisateurs doivent respecter un certain nombre de critères **(Tableau 3)** : sur la structure du problème, la structure des connaissances, les caractéristiques de la foule, évaluation de la solution (Afuah & Tucci, 2012).

Variable	Caractéristiques	Description
Problème	Facilité à isoler et à transmettre Modularité	Le problème est facile à délimiter et isoler du reste du processus, mais également à transmettre (peu d'informations tacites). Le problème sera d'autant plus facilement délégable à la foule s'il est modularisable.
Connaissance	Distance effective Caractère tacite et complexité	Les connaissances requises pour résoudre le problème ne font pas partie des connaissances au sein de l'organisation. De plus l'organisation devrait intégrer des connaissances tacites et complexes.
Foule	Présence de savoir-faire Motivation	Plus il y a de savoir-faire pour la résolution de problèmes dans la foule, plus il y a de chances que le problème soit résolu. De plus, plus les participants sont motivés, plus il y aura de chance qu'ils résolvent le problème.
Solution	Evaluation Evaluateur	La solution doit pouvoir être évaluée par un grand nombre de personnes

Tableau 3. Variables et caractéristiques pour l'utilisation du crowdsourcing dans le cadre de la résolution de problème (inspiré de Afuah & Tucci 2012)

Ce modèle de performance ignore les possibles collaborations entre les participants durant les projets, notamment la possibilité de réutiliser la production des autres participants. En effet, il suppose implicitement que chacun des participants travaille indépendamment les uns des autres, sans aucune interaction. Réutiliser ce que les autres ont produit peut avoir des effets positifs et négatifs sur la performance : d'un côté il peut pousser à une forme de production cumulée de

connaissance et donc d'améliorer la performance globale (Murray & O'Mahony, 2007). De l'autre, la réutilisation peut avoir des effets négatifs sur la diversité des solutions et donc diminuer l'effet statistique du modèle traditionnel (Boudreau & Lakhani, 2015). Etant donnée l'importance de la collaboration et de la réutilisation dans les projets de science citoyenne, nous avons besoin de mieux comprendre les mécanismes mis en jeu dans ce contexte.

3.2.1. Processus de réutilisation des produits intermédiaires dans un système sans contractualisation

La question de la réutilisation de produits intermédiaires fait souvent intervenir les conditions contractuelles. Dans leur étude, Boudreau et Lakhani (2015), comparent différentes formes de politiques de divulgation des produits intermédiaires dans des formes de gouvernance variées : système de brevet, science académique, projets open source entre autres. Dans les systèmes de brevet, les brevets sont protégés par un système de licence permettant de régir la réutilisation (Kitch, 1977). Une fois que le système de protection du brevet est expiré, celui-ci tombe dans le domaine public et peut donc être réutilisé librement. Dans le système académique de la science, les publications sont divulguées dans le domaine public (revues académiques, bases de données sur papier de travail), à la suite de quoi d'autres scientifiques peuvent réutiliser le contenu et les idées en citant leurs sources (Stephan, 1996). Les productions résultantes d'un processus open source sont généralement associées à un système de licence libre ou licence commune qui fournit les réglementations pour définir les conditions de réutilisation et de droit d'accès à des tiers (Rosen, 2004).

Chacun de ces systèmes est associé à des règles et des normes de gouvernance qui définissent les conditions dans lesquelles les produits intermédiaires ou finaux peuvent être réutilisés. Ces règles de gouvernance permettent de définir qui a le droit de voir ce qui a été produit, mais également les conditions sous lesquelles ce qui a été produit peut ensuite être réutilisé. La mise en place de ces règles nécessite cependant d'avoir un contrôle (propriété intellectuelle, relation contractuelle, système juridique) sur les personnes susceptibles de réutiliser ce qui a été produit. Or, dans le cas des projets de science citoyenne, les participants ne sont généralement pas reliés entre eux par l'existence d'un contrat défini *ex ante* et définissant les conditions de réutilisation. De plus, il est difficile pour la personne en charge de piloter le projet d'avoir un contrôle sur tous les moyens de communication utilisables par les participants. Au contraire, les projets de science citoyenne ont plutôt tendance à favoriser la collaboration entre les participants et donc maximiser la réutilisation des produits intermédiaires. La problématique n'est donc pas de définir quelles sont les règles de réutilisation des produits intermédiaires mais plutôt de favoriser cette réutilisation. Cela est essentiel dans un environnement où les individus ou les groupes d'une foule agissent de manière autonome pour développer une solution plutôt que de développer les idées des autres (Madsen et al., 2012).

3.2.2. Diversité, coordination et réutilisation

Ensuite, si on considère que les conditions sont suffisantes pour favoriser la réutilisation de produits intermédiaires, comment cette réutilisation se met effectivement en place et quelles sont les conditions pour qu'elle soit efficace ? La qualité d'une collaboration se base principalement sur deux aspects importants (Riedl & Woolley, 2017) : la coordination entre les membres du projet (Faraj & Majchrzak, 2011; Fayard & DeSanctis, 2008; Lakhani & Von Hippel, 2003; Moon & Sproull, 2008) et la diversité dans les informations partagées (Mello & Rentsch, 2015).

Favoriser la coordination

La recherche sur les communautés en ligne notamment a mis en évidence un certain nombre de facteurs expliquant le succès de certaines communautés en ligne à surmonter les obstacles à la collaboration (Faraj & Majchrzak, 2011; Lakhani & Von Hippel, 2003; Levine et al., 2014; Wasko et al., 2017). Pour maintenir la collaboration dans de telles communautés, il est important de disposer d'animateurs et d'experts mais également de leaders capables de stimuler la conversation (Gray & Tatar, 2004; Wasko et al., 2017), ainsi que de développer des pratiques de communication qui reflètent leur légitimité et leur autorité (Galegher et al., 1998). Il est également important de mettre en places différentes formes de pratiques hétérogènes pour créer des opportunités variées afin de partager l'information (Fayard & DeSanctis, 2008). D'autre part, plusieurs études ont également souligné l'importance du feedback et de pratiques discursives spécifiques pour favoriser la collaboration (Moon & Sproull, 2008; Wooten & Ulrich, 2017). Enfin, les chercheurs ont mis en avant l'importance de la coordination des efforts des participants dans le temps comme prédicteur fort de la performance dans les projets, même quand des métriques comme les systèmes d'incitation ou les compétences des participants sont contrôlées (Riedl & Woolley, 2017).

Diversité dans la production

Un deuxième aspect lié à la performance de la collaboration et de la réutilisation de la production des autres est la diversité des informations. D'un côté, cette diversité est généralement reconnue dans la littérature comme essentielle pour la performance dans une équipe (Cronin & Weingart, 2007; Dahlin et al., 2005; Phillips et al., 2004). Les chercheurs en organisation ont démontré qu'une exposition accrue d'équipes de travail à diverses informations pouvait améliorer les performances, en particulier pour les tâches qui demandaient de la créativité (Jackson & Bantel, 1989; McLeod, Lobel, & Cox, 2007). La diversité peut également fournir un plus grand nombre d'idées au sein d'un groupe de travail (Milliken et al., 1997), ainsi que la non redondance de ces idées (De Dreu & West, 2001).

D'un autre côté, lorsque les projets cherchent à résoudre un problème complexe et incertain, le processus requiert de chercher une solution en fonction des propres stocks initiaux de connaissances et de conviction de chaque participant (Rosenkopf & Almeida, 2003). L'expérimentation par essais-erreurs fournit ensuite des informations en retour et des

informations utiles s'accumulant dans le stock initial de connaissances. Cette recherche d'une solution est souvent sujette à de « mauvais chemins », de « l'expérimentation », de la « sérendipité » et de « l'incertitude » (Abernathy & Rosenbloom, 1969; Boudreau et al., 2008; Loch et al., 2001). Ainsi toute information supplémentaire, comme des résultats intermédiaires, sont susceptibles d'influencer les trajectoires individuelles de chaque participant qui va chercher à réduire son incertitude. A partir du moment où les participants peuvent réutiliser les solutions des autres participants, l'exploration menée par les participants résulte d'un processus beaucoup plus coordonné, où les participants ont tendance à se concentrer sur les solutions qui ont le plus marché (Boudreau & Lakhani, 2015). En contrepartie, ils diminuent le degré d'expérimentation dans la recherche d'une solution et donc la diversité des solutions formulées. Malgré cette baisse de la diversité, l'étude menée par Boudreau et Lakhani (2015) sur des projets de crowdsourcing a montré que les projets collaboratifs (les participants peuvent réutiliser la production des autres) étaient plus performants que les projets compétitifs.

Ces résultats suggèrent que la diversité n'est pas suffisante pour interpréter la performance dans le cas de la réutilisation. Peut-on modéliser et comprendre les mécanismes dans ce type de collaboration ? Quelle est la place de la diversité dans ces projets ? Nous allons tenter de répondre à ces questions dans notre thèse.

4. PERFORMANCE DANS LA REPETITION DES PROJETS DE SCIENCE CITOYENNE

Une petite partie de la littérature analyse la place du crowdsourcing du point de vue de l'organisation. Certains articles décrivent le crowdsourcing comme un nouveau modèle économique d'entreprise. Cela passe par la co-création d'encyclopédies via Wikipedia (Wales, 2005) ou à la conception de t-shirt avec la plateforme Threadless (Brabham, 2008). D'autres études analysent comment des organisations déjà existantes adoptent le principe de crowdsourcing au sein de leur modèle existant : par exemple pour améliorer le service client (Gallaugher & Ransbotham, 2010), pour intégrer les utilisateurs dans le design de produit (Schlagwein & Bjorn-Andersen, 2018), pour aider à collecter de l'or (Tapscott & Williams, 2007). Cette utilisation du crowdsourcing dans un contexte organisationnel pourrait nous être utile à mieux comprendre comment les projets de science citoyenne peuvent piloter le processus d'amélioration continue entre les épisodes.

Certains se sont intéressés au crowdsourcing au sein de l'organisation suivant le principe d' « *absorptive capacity* » pour expliquer le principe d'innovation ouverte (e.g. King & Lakhani, 2012; Spithoven et al., 2009). L'absorptive capacity peut être définie comme la capacité d'une entreprise à reconnaître la valeur d'une nouvelle information, l'assimiler, et l'appliquer en vue de fins commerciales. L'idée sous-jacente pour le crowdsourcing est que l'absorptive capacity d'une

organisation est renforcée par l'utilisation du crowdsourcing. Cependant, cette approche s'intéresse au transfert entre ce qui est produit par le crowdsourcing vers une organisation, tandis que nous sommes plutôt intéressés par un transfert de projet à projet. Dans ce cas, la notion de transfert ne passe pas nécessairement par l'organisation (comme nous le verrons dans notre cas d'étude). Hormis ce rapprochement avec la notion d' « absorptive capacity », il y a un manque important de recherches pour théoriser le crowdsourcing vis-à-vis des théories organisationnelles (Geiger et al., 2011; Majchrzak & Malhotra, 2013).

4.1. LES « PROJECT CAPABILITIES » POUR GERER LA TRANSMISSION ENTRE LES PROJETS

Un des concepts organisationnels correspondant à la question de la gestion de la transmission entre projets est la notion de *capacité dynamique*. Les capacités dynamiques désignent les processus d'innovation stratégiques utilisés pour adapter, intégrer et reconfigurer les compétences, ressources et routines internes et externes d'une entreprise en réponse à des conditions changeantes et instables (Eisenhardt & Graebner, 2007; Winter, 2003; Zollo & Winter, 2003). Ce concept a été appliqué dans la littérature de gestion de projet pour identifier comment les entreprises déploient plusieurs projets pour des clients existants et lancent des projets innovants pour développer de nouvelles technologies et créer de nouveaux marchés (Brady & Davies, 2004; Davies & Brady, 2000; Ethiraj, Kale, Krishnan, & Singh, 2005; Gann & Salter, 2000).

Des travaux remarquables sur cette question dans le cadre de projets temporaires ont été réalisés par les chercheurs Brady et Davis (2000 ; 2004). Ils ont étudié la question de la répétition et de l'apprentissage organisationnel dans la conception de produits et systèmes complexes en faibles volumes vis-à-vis des exigences clients (Brady & Davies, 2004; Brady, Marshall, Prencipe, & Tell, 2002; Davies & Brady, 2000). Ils proposent une approche dans la lignée du modèle de *resource-based view* et de la construction de nouvelles aptitudes dans l'entreprise (Chandler, 1990; Penrose, 1959). Dans ce modèle, l'entreprise possède un ensemble de ressources spécialisées et de connaissances utiles pour adresser des offres à certains marchés. En plus de développer des connaissances uniques qui permettent à l'entreprise de se différencier de la concurrence, comme des « *core competencies* » (Hamel & Prahalad, 1994) ou des « *distinctive capabilities* » (Richardson, 1972), les entreprises dépendent en même temps de leur capacité interne au changement : des capacités stratégiques pour piloter les opérations internes et adapter les stratégies à un changement d'environnement ; des capacités fonctionnelles organisées en département ou en silo pour produire des produits ou des services à fort volume (Chandler, 1990). Davies et Brady (2000) introduisent le concept supplémentaire de « *project capabilities* », qui fait référence aux activités principales des entreprises qui conçoivent et produisent des produits et des systèmes complexes en faibles volumes, en fonction des besoins spécifiques des clients. Celles-ci sont nécessaires pour s'engager avec leurs clients dans des activités pré-offre stratégiques ; préparer des propositions ou, si l'entreprise est impliquée dans un partenariat stratégique avec le

client, présenter des «offres» ; gérer les activités de cycle de vie impliquées dans la mise en œuvre du projet, le transfert au client et l'assistance continue.

Brady et Davies (2004) proposent que l'organisation de projets au sein d'une organisation intègre un processus d'apprentissage organisationnel pour réduire le risque de répéter les mêmes erreurs d'un projet à un autre. En effet, les projets lancés par les organisations ont souvent des similarités impliquant des pratiques organisationnelles répétables et prédictibles. L'apprentissage organisationnel peut être réalisé durant trois phases relatives au projet :

> Au cours de la phase 1, une nouvelle organisation de projet est établie à la pointe d'une organisation afin d'explorer les opportunités stratégiques permettant d'accéder à de nouvelles bases technologiques ou de marché, ou de s'adapter à un environnement de marché en mutation. Les membres du projet utilisent leurs connaissances existantes pour les guider dans leurs actions, tandis que l'expérience acquise est partagée par le groupe impliqué dans le projet. Dans cette phase d'apprentissage «intra-projet» (Keegan & Turner, 2001), l'expérience acquise est partagée par le groupe impliqué dans le projet. Etant donné l'aspect souvent exploratoire de ces projets, les participants doivent être prêts à s'éloigner de leurs routines organisationnelles pour inventer de nouvelles règles et méthodes de travail plus efficaces (Ayas & Zeniuk, 2001)
>
> Au cours de la phase 2, l'apprentissage d'un projet à l'autre prédomine alors que l'on tente de saisir et de transférer l'expérience et les connaissances des participants au projet séminal vers les équipes de projet suivantes qui peuvent en tirer parti. Dans ce processus d'apprentissage entre projets (Keegan & Turner, 2001), des mécanismes d'apprentissage formels sont développés pour saisir l'apprentissage des projets, le codifier et le rendre accessible aux autres équipes de projet. Les membres clés d'un projet précédent peuvent être réintroduits dans les projets futurs.
>
> Une fois qu'un nombre suffisant de nouveaux types de projets ont été initiés par l'organisation, la phase 3 permet de développer un apprentissage entre les projets. Des efforts sont entrepris par les organisations pour systématiser l'apprentissage et la transmission des connaissances accumulées vers l'ensemble de la division responsable des projets. Les entreprises peuvent également avoir besoin de créer des cellules spécialisées pour prendre en charge un nombre croissant de projets. Cela permet de s'assurer que les connaissances acquises restent effectivement dans la mémoire de l'organisation (Brady & Davies, 2004).

4.2. PILOTER LA TRANSMISSION D'INFORMATION ENTRE DES PROJETS INCLUANT UNE FOULE

Le contexte d'étude de Brady et Davis s'adresse à des organisations qui systématisent l'utilisation de projets pour améliorer leur compétitivité au sein d'un marché. Les projets sont développés pour répondre à des problématiques complexes de développement de nouveaux produits ou de nouveaux services. S'ensuivent pour l'organisation la construction de nouvelles connaissances et de compétences spécifiques qui lui permet de se distinguer de la concurrence et de la création de nouvelles routines organisationnelles associées à la fois aux nouvelles connaissances et à cette organisation par projet. Or, ces études ne se sont pas intéressées à cette question dans le cadre d'organisations scientifiques éphémères, c'est-à-dire dans des organisations où les participants au projet n'ont pas d'intérêts directs à ce que le projet aboutisse, et donc à faciliter la transmission de l'information. Ils n'ont aucune obligation contractuelle qui puisse assurer la transmission codifiée de tout ce qu'ils ont produit et appris. En fait, il n'y a pas de certitude que les projets vont se terminer durant la phase projet définie *ex ante*. On ne peut pas compter sur les citoyens de la science comme ressource pérenne comme on pourrait le faire sur des acteurs qui sont embauchés spécifiquement pour finir le projet.

Cette spécificité des projets de science citoyenne amène à plusieurs difficultés pour assurer l'apprentissage dans les différentes phases proposées par Brady et Davies. D'abord dans la phase 1, la collaboration entre les participants pour qu'ils puissent partager leurs apprentissages n'est pas évidente. En effet, bien qu'une collaboration spontanée s'observe au sein des communautés en ligne, les profils des participants sont variés et seuls une petite partie, appelée « coopérateurs », facilitent la transmission d'information au sein de la communauté (S.S. Levine & Prietula, 2012; Sheen S Levine et al., 2014). Ainsi une grande partie de la production et de l'apprentissage risque d'être perdue. Ensuite, la phase 2 implique généralement les participants des anciens projets dans les projets successifs pour favoriser la transmission (Brady & Davies, 2004). Or, cette transmission est fragilisée dans les projets de science citoyenne, car les participants ne reviennent que rarement dans les projets suivants et que ces derniers n'ont pas forcément des profils de coopérateurs. Ensuite, quand un projet se termine, les équipes en charge du projet ont rarement le temps ou la motivation de faire un retour d'expérience et de documenter quelles sont les connaissances transférables à recycler pour de nouveaux projets (Coombs & Hull, 1998). Faciliter cet apprentissage demanderait de mettre en place un ensemble d'outils et de logiques de gestion *ex ante* pour s'assurer que tout ce qui est produit suit un schéma et une logique prévisible. Or, la mise en place de ce type de mécanisme est rapidement limitée car elle se confronte à la création de connaissances qui se fait souvent en tant que sous-produit involontaire de l'activité de projet (DeFillippi, Jones, & Arthur, 2001). Comment dans ce cas supporter la transmission d'informations et de ce qui est produit entre les épisodes successifs afin d'améliorer de manière continue et dynamique la performance des projets de science citoyenne ? Notre travail de thèse cherchera à apporter des éléments de réponse à cette question.

CHAPITRE 2 – LE CONTEXTE DES DONNEES COMME CADRE D'ETUDE : L'EFFET DE L'AVALANCHE DE DONNEES SUR LE PROCESSUS SCIENTIFIQUE

RESUME DU CHAPITRE 2

Dans ce chapitre, nous nous intéressons à la transformation du rapport aux données suggérée par un certain nombre d'épistémologues avec l'arrivée de bases de données massives dans le processus scientifique. Nous montrons que cette transformation est susceptible d'avoir des impacts organisationnels sur le processus de production de connaissance, notamment sur le processus de génération des hypothèses scientifiques. Nous montrons l'émergence d'une dichotomie entre deux modes organisationnels de la science. D'un côté le modèle traditionnel, dit « knowledge-driven », qui s'est construit dans un environnement pauvre en données (Baraniuk, 2011; Miller, 2010),. Les mesures de la réalité étaient difficiles, coûteuses et lourdes à obtenir, à stocker et à manipuler. De l'autre côté, l'émergence d'une science dite « data-driven » où les coûts de capture, de stockage et de manipulation des données numériques ont fortement baissé, et les technologies de communication et d'information sont maintenant largement déployées dans les organisations scientifiques (Miller, 2010). Nous suggérons alors que les projets de science citoyenne sont susceptibles d'apporter une réponse organisationnelle à cette transformation, mais que celle-ci nécessite d'être gérée.

La littérature sur les sciences citoyennes, les nombreux exemples de projets (voir l'échantillon en **annexe** 2 par exemple) ainsi que l'avalanche de données dans de nombreuses disciplines scientifiques suggèrent qu'il y a une opportunité d'un point de vue des sciences de gestion à analyser les sciences citoyennes en rapport avec les données. D'un côté, les différentes études menées sur les modes d'ouverture de la science ont constaté implicitement ce lien entre ouverture et données (Franzoni & Sauermann, 2014; Haklay, 2015; Houllier, 2016; Wiggins & Crowston, 2011), sans pour autant l'exploiter. Dans des disciplines comme la biologie l'environnement, ou le développement durable les chercheurs analysent les projets de science citoyenne selon une grille d'analyse basée sur les activités autour des données : la collecte, la labellisation, l'analyse, ou encore la construction de modèles prédictifs à partir de bases de données (Bonney et al., 2014; Haklay, 2015; Wiggins & Crowston, 2011). Par ailleurs, les cas d'utilisation des projets de science citoyenne autour des données se multiplient. Le projet Zooniverse, une des principales plateformes regroupant des projets de science citoyenne, permet de se rendre compte en partie du phénomène. Cette plateforme s'est donnée pour mission *« d'utiliser la « sagesse de la foule » afin de produire des données exploitables et de bonne qualité »* [1]. Des projets de science citoyenne basés sur les données ont émergé dans une grande variété de disciplines comme la biologie (45 projets), la climatologie (9 projets), l'histoire (12 projets), la linguistique (9 projets), la médecine (7 projets), les sciences naturelles (48 projets), la physique (12 projets), les sciences sociales (9 projets), l'astronomie (17 projets)[2].

De l'autre côté, un ensemble de chercheurs considère que notre époque se situe à l'aune d'une transformation profonde dans la science due à l'évolution de plusieurs caractéristiques des données : leur accessibilité, les méthodes et les techniques pour analyser les données, le lien entre les données et la formulation des hypothèses scientifiques (Anderson, 2008; Gray, 2009; Kitchin, 2014; Miller, 2010; Shmueli, 2011). Cette transformation ne concerne pas une seule discipline isolée mais impacterait un grand nombre de champ de recherche tel que les sciences environnementales, les sciences de la terre, la santé, mais également la biologie ou l'astronomie (e.g. Gray, 2009; Swan, 2013).

L'avalanche de bases de données massives pousse le processus scientifique vers ses limites en terme de capacité de production pour traiter, stocker et coder ces données[3] (Laney, 2001). Plusieurs structures scientifiques sont obligées de repenser leurs systèmes de stockage et de traitement des données pour répondre à un besoin grandissant (Hey, Tansley, & Tolle, 2009). Par exemple, les équipes travaillant sur le plus grand collisionneur de particules, le Large Hadron Collider (LHC), prévoient que la quantité de collision générée sera multipliée par 100 d'ici 2020 alors que les technologies de traitement des données existantes ne sont pas adaptées (Gligorov, 2015). Bien que leur système de stockage, de filtrage et de traitement des données soit souvent

[1] https://www.zooniverse.org/about

[2] (site visité en Mars 2019)

[3] Le cas de Galaxy Zoo en est un bon exemple. Les scientifiques n'avaient pas les moyens techniques et humains internes pour coder les 900 000 images de galaxies dans un temps convenable.

reconnu pour son efficacité unique au monde, l'ensemble du système est susceptible d'être saturé et de ne pas pouvoir accueillir en l'état cette augmentation du nombre de données disponibles. Une des idées soumises par un des experts en analyse de données serait de pouvoir systématiser une analyse en temps réel des données collectées afin de ne conserver que le produit de cette analyse et de supprimer les données restantes (Gligorov, 2015). L'application de ce principe transformerait profondément l'organisation du LHC, et pose des questions sur la réplicabilité des expériences tandis qu'actuellement toutes les collisions sont conservées en trois exemplaires différents pour éviter de perdre la trace des données brutes de l'expérience.

Par ailleurs, comme nous allons le voir dans ce chapitre, l'avalanche des données massives est susceptible d'avoir d'autres conséquences sur les méthodes d'analyse des données, mais également sur le modèle de raisonnement scientifique (Kitchin, 2014; Shmueli, 2011). Nous verrons notamment que l'avalanche de données massives est susceptible de modifier le processus de génération des hypothèses scientifiques dans un paradigme appelé « *data-driven science* », et donc d'impacter le processus de production de connaissances. Nous suggérerons que les sciences citoyennes peuvent apporter une réponse en terme de capacité de production à cette transformation, mais qui demande de penser à un mode de gestion adapté.

1. LA TRANSFORMATION PAR LES DONNEES COMME CADRE IMPENSE POUR L'ETUDE DE L'OUVERTURE

1.1. L'AVALANCHE CONTEMPORAINE DES DONNEES DISPONIBLES

Au cours des dernières décennies, la science est passée d'un contexte limité en données à une abondance des données disponibles (Miller, 2010). Cette transformation touche un large panel de disciplines scientifiques comme les sciences de la terre, la médecine et les domaines de la santé, la physique des particules, la génétique, mais également plusieurs disciplines des sciences sociales (George, Haas, & Pentland, 2014; Gligorov, 2015; González-Bailõn, 2013; Gray, 2009; Raghupathi & Raghupathi, 2014). L'émergence de ces données représente une opportunité pour les scientifiques afin d'étudier des phénomènes jusqu'alors inaccessible que de trop faibles échantillons ne permettaient pas d'envisager. Dans le domaine de la santé par exemple, de multiples avantages apportés par l'émergence de ces bases de données massives sont suggérés : détecter des maladies à des stades précoces ; gérer la santé des individus de manière plus rapide et efficace ; prédire ou estimer certains évènements comme des complications, des taux de remplissage des hôpitaux, ou les progressions de la maladie (Raghupathi & Raghupathi, 2014).

Ces bases de données, souvent regroupées sous le terme « Big Data », ne sont pas nécessairement compatibles avec les méthodes scientifiques traditionnelles et engendrent un ensemble de défis organisationnels pour les scientifiques (Fan, Han, & Liu, 2014; Gandomi & Haider, 2015; Kitchin

& McArdle, 2016; Labrinidis & Jagadish, 2012; Lazer, Kennedy, King, & Vespignani, 2014; Russom, 2011). En effet pendant la majeure partie de son histoire, la science a opéré dans un environnement pauvre en données. Les mesures de la réalité étaient difficiles, coûteuses et lourdes à obtenir, à stocker et à manipuler. Par conséquent, une grande partie de l'appareil scientifique a été conçue pour tirer des informations de rares observations. Une problématique récurrente liée au Big data est celle de la taille des jeux de données disponibles (Kitchin & McArdle, 2016; Laney, 2001). A la fin des années 1990, la difficulté majeure rencontrée par les organisations était de pouvoir stocker des volumes de données pour lesquelles elles n'étaient pas équipées techniquement parlant. Durant les années 2000 et 2010, les coûts de capture, de stockage et de manipulation des données numériques ont fortement baissé, et les technologies de communication et d'information pour récupérer les jeux de données générées se sont largement déployées dans les organisations scientifiques (Miller, 2010). Les structures scientifiques se sont donc dotées d'une organisation capable de stocker et de traiter ces données.

Aujourd'hui, les opportunités liées à la baisse des barrières techniques pour stocker et traiter les bases de données favorisent la collecte et le codage de bases de données massives dans les organisations scientifiques. Cependant, d'autres problématiques de gestion liées à l'avalanche de données restent encore largement d'actualité, notamment l'intégration des méthodes d'analyse des bases de données massives ainsi que la génération des hypothèses scientifiques.

1.2. NOUVELLES METHODES D'ANALYSES DES DONNEES ISSUES DE L'INTELLIGENCE ARTIFICIELLE

En parallèle de cette avalanche de données, plusieurs chercheurs font état de l'apparition de nouvelles méthodes pour analyser ces données et construire des modèles scientifiques (Boyd & Crawford, 2011; Dhar, 2013; Kitchin, 2014; Lin & Lucas, 2013). Ces algorithmes, souvent regroupés sous le terme Intelligence Artificielle (IA), ont l'avantage d'être particulièrement efficaces lorsque les dimensions de bases de données explorées sont très grandes (de l'ordre de plusieurs milliers voire plusieurs millions de variables). Une IA peut être définie comme un système qui perçoit son environnement et réalise des actions afin de maximiser ses chances d'atteindre un but préalablement défini (Poole, Mackworth, & Goebel, 1998). Bien que le principe de l'IA existe depuis les années 1950, ce sont les récentes avancées considérables dans ce domaine qui ont entrouvert la possibilité de les utiliser de façon large dans le cadre de la méthode scientifique (Mjolsness & DeCoste, 2001).

1.2.1. Pluralité d'application des nouvelles méthodes d'analyses des données

Si les algorithmes d'IA sont de plus en plus utilisés pour des données scientifiques, ils ont également un intérêt pour d'autres types de données rencontrées au sein du processus. Selon Ilias Tagkopoulos, informaticien à l'Université de Californie à Davis, tout problème d'optimisation complexe pour lequel une amélioration est bien définie peut faire l'objet d'une automatisation, que ce soit pour que l'algorithme dicte quelle expérience doit être faite pour maximiser le gain d'information, prédire l'évolution des bactéries dans un environnement hospitalier changeant, ou encore concevoir de meilleurs snacks. « Si les machines peuvent remplacer les humains dans certaines tâches scientifiques, il y a des chances que de nombreux scientifiques les adoptent » affirme Bohannon (2017). Selon lui, un ensemble de tâches dans le processus scientifiques sont fastidieuses et pourraient être mieux exécutées par un ordinateur plutôt que par un humain. L'avalanche de données n'impacte pas uniquement les données à proprement scientifique, mais peut également renseigner sur le fonctionnement du laboratoire de recherche, la bibliographie d'un sujet de recherche, ou bien encore peut améliorer le fonctionnement de machines utilisées pour les expérimentations. En fait, les possibilités ne sont limitées que par la capacité d'intégrer de nouvelles données. John Bohannon propose un aperçu des différentes utilisations possibles de l'intelligence artificielle dans la recherche en biologie. Il montre que les algorithmes ne se limitent pas à l'interprétation des données scientifiques mais peuvent être intégré à toutes les étapes du processus scientifique : l'exploration de la littérature scientifique, la conception des expérimentations, la réalisation des expérimentations, l'interprétation des données, et l'écriture du papier scientifique. Chaque élément de la séquence se voit associé des outils d'amélioration basés sur les méthodes de l'intelligence artificielle et de l'analyse de données massives **(tableau 4)**. Par exemple, Citeomatic est un modèle de machine learning qui lit les papiers scientifiques et prédit quelles sont les citations manquantes. Cet algorithme peut également servir d'outil utile pour réaliser une revue de littérature à tout moment du processus d'écriture.

Les algorithmes d'IA offrent des outils efficaces pour l'analyse des bases de données massives, parfois bien supérieures aux méthodes traditionnelles. Dans un article publié dans *Nature* par exemple, des scientifiques ont utilisé un algorithme d'IA pour connaître les spécificités des séquences d'ADN et d'ARN (Alipanahi et al., 2015). Ils ont montré que l'utilisation de cet algorithme permet de réduire considérablement le nombre de paramètres du modèle par rapport à un algorithme classique et obtiennent par la même des résultats bien supérieurs que les méthodes de pointe utilisées actuellement. Si les algorithmes d'IA surpassent sur certains points les méthodes plus traditionnelles, ils peuvent en revanche être moins performants dans leur capacité à expliquer un phénomène. En effet, les variables dans les algorithmes d'IA sont souvent couplées de façon complexe, réduites ou transformées et ne permettent que difficilement d'expliquer les relations existantes entre elles (Snoek, Larochelle, & Adams, 2012). Ces modèles sous forme de « boîtes noires » réduisent la capacité explicative de l'intelligence artificielle, pourtant indispensable pour formuler des modèles théoriques scientifiques. En fait, certains

chercheurs suggèrent que leur mise en place implique un changement important dans l'analyse de données, passant d'algorithmes traditionnels à forte capacité explicative à des algorithmes d'IA qui présentent plutôt de fortes capacités prédictives (Shmueli, 2011).

Activité	Description	Outils d'IA
1. Explorer la littérature scientifique	Trouver les papiers les plus pertinents parmi des millions, détecter des nouveaux sujets émergents	**Semantic scholar** Un moteur de recherche qui extrait les bons mots d'un papier, mais aussi les graphiques et les citations influentes **Iris.AI** Un outil pour explorer les papiers scientifiques via le concept qui les relient
2. Concevoir des expérimentations	Trouver le bon équilibre entre l'exploration de nouveauté et l'exploitation de phénomènes connus	**Zymergen** Entreprise avec une AI qui pistes des milliers de variables dans les génomes de microbe
3. Réaliser les expérimentations	Conserver les pistes de milliers de tubes, molécules et cellules, minimisant l'imprécision et les erreurs	**Transcriptic, Emerald Cloud Lab** Laboratoires robotiques en cloud pour faire des expériences automatiquement
4. Interpréter les données	Donner un sens aux résultats variés qui viennent des expérimentations	**Nutonian** Une plateforme software qui intègre beaucoup de données et ressort une théorie mathématique qui explique les variations dans les données
5. Ecrire un papier scientifique	L'écriture d'un papier peut être aidée par un software	**Citeomatic** Un outil en ligne qui lit les papiers et prédit quelles citations sont manquantes

Tableau 4. Bohannon 2017, Nature

1.2.2. Des algorithmes d'intelligence artificielle à fort pouvoir prédictif

Si les algorithmes d'IA présentent des avantages indéniables sur leurs capacités à traiter de larges bases de données, ils posent en mêmes temps des questions sur leur capacité à s'intégrer comme méthode d'analyse dans une méthode scientifique bien établie. D'un côté, les techniques d'analyse de données scientifiques traditionnelles ont été conçues pour extraire des informations à partir de jeux de données pauvres, statiques, propres, échantillonnés scientifiquement ayant été générées et analysées avec une question spécifique en tête (Miller, 2010). L'analyse est réalisée au travers d'outils statistiques dont le modèle de base est proche de l'inférence inductive : les données obtenues sont analysées et leur résultat produit un verdict qui induit une généralisation du résultat obtenu sur la base de données. Ces analyses sont souvent réalisées au travers de méthodes statistiques pour tester la causalité d'une théorie avec une approche fréquentiste (Shmueli, 2011). A partir de modèles théoriques, deux hypothèses de causalité se confrontent pour réaliser le test statistique : une hypothèse dite nulle H0 et une hypothèse alternative H1, conçue comme la négation logique de H0. Les scientifiques choisissent ensuite un seuil afin de valider ou de rejeter l'hypothèse H0. Cette analyse mène les scientifiques à des conclusions statistiques qui leurs

permettront de statuer sur la valeur de vérité de l'hypothèse et donc de produire une nouvelle connaissance. Dans cette approche, c'est l'aspect explicatif de la modélisation qui prime sur le reste. L'objectif de l'analyse est de comprendre l'essence d'un phénomène et d'en apporter une représentation simplifiée et compréhensible des effets de causalité entre les différentes variables. Le modèle doit donc être suffisamment clair pour être interprétable et donc être suffisamment explicatif.

Avec les algorithmes d'IA basés sur le Big data, l'analyse des données est souvent confrontée à « une abondance, une exhaustivité et une variété, un dynamisme, du désordre et de l'incertitude, des relations fortes et le principe que ce qui est généré n'est pas régi par une question particulière ou provient d'une autre activité » (Kitchin, 2014). Les liens de causalité entre les variables ne sont pas toujours explicites et le modèle théorique ne permet pas aussi facilement de se représenter *ex ante* une idée précise des relations. Les modèles statistiques classiques ne sont pas toujours adaptés lorsque les données ne sont pas collectées dans un but précis, que celles-ci présentent une forte hétérogénéité ou qu'elles soient massives. Les techniques d'intelligence artificielle (IA) proposent une réponse à ce problème en élaborant des relations complexes entre des bases de données massives ou provenant de sources très hétérogènes (Han & Kamber, 2012; Hastie, Tibshirani, & Friedman, 2009; Shmueli, 2011). Alors que les modèles traditionnels sont d'abord modélisés par le scientifique qui va ensuite les tester sur les données, les algorithmes d'IA testent eux-mêmes des modèles et apprennent des résultats de façon autonome afin de faire évoluer le système. Ces méthodes ont l'avantage d'être particulièrement efficaces lorsque les dimensions de bases de données explorées sont très grandes.

1.2.3. Elaboration d'un modèle d'analyse basé sur un processus d'essai-erreur

Si les modèles statistiques traditionnels créent des liens de causalité entre les variables facilement représentables, les méthodes issues de l'IA peuvent être utilisées juste en ayant en tête quelques possibles corrélations entre ces variables. La recherche du meilleur modèle pour analyser un jeu de données repose alors sur un processus d'essai-erreur où des centaines d'algorithmes différents peuvent être appliqués sur un même jeu de données (Siegel, 2013). Les algorithmes d'IA élargissent les moyens mis à disposition des scientifiques tout en complexifiant cette partie du travail. Le processus d'optimisation pour chercher le meilleur algorithme peut s'avérer relativement coûteux en ressources pour les scientifiques qui peuvent passer beaucoup de temps à explorer une piste sans savoir si celle-ci est une bonne voie. En effet, ce processus dépend principalement du talent et des connaissances d'une poignée de scientifiques qui ne sont pas nécessairement les plus à-même à trouver le meilleur algorithme (Afuah & Tucci, 2012; Dhar, 2013; Provost & Fawcett, 2013).

Au lieu de chercher à optimiser eux-mêmes leurs modèles, certains scientifiques externalisent ce processus par le biais de projets de science citoyenne. Le projet « Dialogue for Reverse

Engineering Assessments and Methods » (DREAM) par exemple est une plateforme lancée en 2006 spécialisée dans les challenges issus de la biologie computationnelle (voir **annexe 1**). Les biologistes considéraient que leurs modèles étaient élaborés sur de trop petites bases de données et pouvaient difficilement être inférés à un large panel de situations non encore testées expérimentalement. Au lieu de réaliser un ensemble de modèles chacun validé suivant quelques mesures expérimentales, le groupe de travail a proposé l'idée de combiner les différentes informations contenues individuellement dans chaque modèle afin de pouvoir faire des prédictions plus précises. Ils ont donc mis en place différents challenges sous forme de projets de type science citoyenne pour demander à des volontaires de réfléchir à leurs problématiques. Au cours des différents challenges, ils ont pu obtenir des algorithmes bien meilleurs que les algorithmes habituellement utilisés et ont publié dans des revues prestigieuses comme *Nature Biotechnology* (4 papiers), *Nature Review Genetics* ou *Science*.

Ce type d'organisation permet de réduire le temps d'exploration et d'augmenter la probabilité d'atteindre la meilleure solution au problème (Afuah & Tucci, 2012; A. King & Lakhani, 2013; Terwiesch & Xu, 2008). Pour se représenter l'intérêt de déléguer un problème à une foule de personnes plutôt qu'à le résoudre seul, illustrons le principe d'essai-erreur par le modèle classique informatique de type *hill climbing*. Le principe est d'implémenter un algorithme itératif qui, à partir d'une solution donnée, va chercher à trouver la meilleure solution en proposant des changements incrémentaux à la solution. Cette approche considérée comme faisant partie de la famille des méthodes de recherche locale présente ses limites si la solution initialement choisie n'est pas située proche de la solution optimale. En effet, la solution peut se retrouver localement optimale (meilleure que tous ses voisins proches) sans pour autant être la meilleure solution globale. La résolution de problèmes par une foule via un modèle compétitif permet d'augmenter la probabilité de trouver la solution optimale. Ce n'est plus un agent seul mais un ensemble d'agents qui vont chercher à résoudre individuellement le problème. La recherche de la solution dans l'espace des actions consiste alors en la multiplication du nombre de points d'entrée dans l'espace des actions, chaque point d'entrée provenant d'un participant. La recherche de la solution optimale n'est plus une recherche distante pour un agent mais une multitude de recherches locales pour l'ensemble des agents (Afuah & Tucci, 2012).

2. LA FORMULATION DES HYPOTHESES SCIENTIFIQUES : DE LA SCIENCE « KNOWLEDGE-DRIVEN » A LA SCIENCE « DATA-DRIVEN »

2.1. POSITIONNEMENT EPISTEMOLOGIQUE DE LA TRANSFORMATION PAR LES DONNEES

Les transformations sur l'accessibilité aux bases de données ainsi que l'apparition de nouvelles techniques d'analyses mènent à des interrogations sur l'impact épistémologique de ces nouveaux objets dans la science. En fait, avoir accès à toutes ces données et aux algorithmes pour les étudier pourrait faire penser que l'on peut se passer des modèles théoriques scientifiques (Anderson, 2008). Le travail du scientifique consisterait dans ce cas uniquement à explorer de manière opportuniste les bases de données qui sont à sa disposition jusqu'à ce qu'il découvre une « heureuse » relation entre des variables. Trois courants de pensée existent dans la littérature scientifique ayant des avis divergents sur le sujet. Premièrement, un courant voit l'avalanche de données comme associé à un nouveau paradigme dans la gestion des données scientifiques. Un deuxième courant y voit le retour à un empirisme baconien, où la science serait faite sans théorie. Enfin, un troisième courant suggère que l'apparition du Big Data dans le processus scientifique va surtout impacter le processus de génération des hypothèses scientifiques, passant d'une science dite « data-driven » vers une science dite « knowledge-driven ». Nous passons en revue ces différentes approches dans la littérature.

2.1.1. Le quatrième paradigme selon Jim Gray

Une des visions les plus répandues de la transformation du rapport aux données est sans doute celle que propose Jim Gray, un chercheur reconnu pour ses contributions majeures dans le domaine de l'informatique. Selon lui le processus scientifique pourrait être amené à ce qu'il appelle un quatrième paradigme de la science, aussi connu sous le nom de « data-intensive science » (Hey et al., 2009). Pour expliquer cela, Gray propose une relecture historique de l'évolution de la science par rapport aux capacités des organisations à stocker l'activité scientifique. Selon lui, chaque paradigme que la science a connu est associé à une contrainte supplémentaire pour les scientifiques dans leur capacité à gérer le flux de données générées. Lors du premier paradigme entre le 17è et le 18è siècle, les organisations scientifiques faisaient face à un défi majeur : plus les expériences se complexifiaient et plus les bases de données devenaient plus importantes. Par conséquence il devenait de plus en plus difficile de citer l'entièreté de ces bases de données dans les publications scientifiques. Pour éviter de perdre ces données, les organisations ont mis en place des systèmes d'archives pour à la fois déposer les données et les rendre facilement disponibles. Cependant, le nombre de scientifiques à continuer à croître en même temps que le nombre de publications. Dès lors que la littérature scientifique est devenue trop importante en terme de volume, des outils et des pratiques ont du être développés pour gérer le changement d'échelle. On a ainsi vu apparaître des revues spécialisées, des systèmes de citation et d'index, des revues bibliographiques. Le troisième paradigme apparaît selon lui au milieu du

20è siècle avec la croissance des technologies computationnelles et des techniques de simulation numérique. Non seulement les données continuaient à croître, mais les résultats des simulations et des expériences devenaient des ensembles de données volumineux et complexes qui ne pouvaient être résumés, car les technologies de calcul étaient complexes, longues et parfois hasardeuses. Il a fallu plusieurs dizaines d'années pour que les ordinateurs soient suffisamment fiables pour les calculs à grande échelle, et réduire le volume de ce qu'il fallait conserver.

Le quatrième paradigme, dans lequel nous sommes actuellement entré, est caractérisé par le déploiement des technologies numériques qui permet de multiplier les modes de publications numériques, et qui à terme pourrait transformer le modèle de publication traditionnel des revues scientifiques électroniques. Selon Gray, de nouveaux éléments seront intégrés aux publications scientifiques qui pourront accueillir par exemple des vidéos, mais également l'ensemble des données utilisées, ainsi que les méthodes d'analyse. À cette ère de l'informatique « data-intensive », les chercheurs utilisent les publications scientifiques de deux manières. Suivant une approche traditionnelle, ils se contentent de lire les documents comme ils le font depuis des siècles, mais avec des outils informatiques qui leur permettent une grande facilité, précision et flexibilité. D'un autre point de vue, les chercheurs peuvent étudier les publications scientifiques dans leur ensemble, en tant que corpus de textes et ensemble de données interconnectées. Par le biais de nouveaux outils informatiques, ils pourront à partir de ces bases de données identifier les papiers d'intérêt, suggérer des hypothèses, ou produire directement de nouvelles données ou résultats.

Paradigme	Nature	Forme
Premier	Science expérimentale	Empirisme, description de phénomènes naturels
Deuxième	Science théorique	Modélisation et généralisation
Troisième	Science computationnelle	Simulation de phénomènes complexes
Quatrième	Science exploratoire	Data-intensive, exploration statistique et data mining

Tableau 5. Les quatre paradigmes de la science selon Gray (Hey et al., 2009)

2.1.2. Un retour à l'empirisme

La transformation du rapport aux données décrite par Jim Gray fait plutôt écho pour certains chercheurs au retour à une époque où la science était marquée par l'empirisme (Dyche, 2012; Siegel, 2013). Le chef de file de ce courant est sans doute Chris Anderson, auteur de plusieurs ouvrages sur l'économie de l'internet et de la gratuité. Dans un article volontairement provocateur, il suggère que l'accessibilité à de larges volumes de données associée à de nouvelles techniques d'analyses permettront aux données de pouvoir s'exprimer sans avoir besoin de se baser sur des modèles théoriques ni de formuler des hypothèses (Anderson, 2008). Selon lui, les

corrélations que l'on peut découvrir au sein des bases de données massives sont suffisantes pour construire de la connaissance sur des phénomènes :

> *There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.*

Les exemples utilisés pour illustrer son propos sont principalement issus du monde de l'entreprise et du marketing : son article part d'ailleurs du constat que Google n'a pas besoin de comprendre les préférences de chaque utilisateur, il a simplement besoin de les connaître. Or, un certain nombre de chercheurs critiquent cette vision car elle est justement trop peu centrée sur les contraintes de la méthodologie scientifique. Plusieurs arguments vont en défaveur de l'idée de l'existence d'une science basée sur les données totalement privée de théorie (Crawford, 2013; Kitchin, 2014) : une base de données ne pourra jamais être exhaustive pour rendre compte d'un phénomène ; les bases de données ne sont jamais collectées de nulle part, sans idée en tête ; l'idée que les données peuvent parler d'elles-mêmes supposent que n'importe qui pourrait les interpréter sans connaître le contexte de l'étude ou avoir un minimum de connaissance sur le sujet. Depuis les années 2010, les partisans d'un retour à un empirisme pur se font plus rares dans la littérature, et cette vision est progressivement mise en retrait.

2.1.3. De la science « knowledge-driven » à la science « data-driven »

Enfin, un ensemble de chercheurs considère que le paradigme de la data-intensive science va amener à une science qui utilisera de façon hybride les inférences d'abduction, d'induction et de déduction dans la compréhension d'un phénomène (Blaikie & Priest, 2019; Kelling et al., 2009; Kitchin, 2014; Sarker et al., 2018). Selon eux, le quatrième paradigme de la science ferait émerger une science où les hypothèses sont formulées à partir des bases de données existantes au lieu de provenir uniquement de la théorie. Sans prétendre que le processus scientifique peut être privé de théorie comme le suggèrent les partisans du retour à l'empirisme, ces scientifiques suggèrent que la génération d'hypothèses se baserait à la fois sur les données disponibles ainsi que les connaissances scientifiques. Ils esquissent alors une dichotomie entre une forme classique de production de connaissance, appelée « *knowledge-driven* », et une forme émergente qui se profile avec l'avalanche du Big data dans la science, appelée « *data-driven* ». Dans notre thèse, nous partageons cette vision de la transformation de la science par les données.

2.1. PLACE DES DONNEES DANS LA SCIENCE DATA-DRIVEN

Nous proposons d'illustrer ces deux modes de raisonnement scientifique par des exemples. Deux exemples de processus *knowledge-driven* proviennent de synthèses réalisés dans la littérature à partir des 'cahiers de laboratoire' des scientifiques : la synthèse de l'urée par Hans Krebs (Kulkarni & Simon, 1988) et la découverte du vaccin contre le choléra des poules par Pasteur

(Cadeddu, 1985). Les exemples du processus *data-driven* sont plus récents et proviennent de sources diverses. Le premier est l'exemple du programme astronomique Sloan Digital Sky Survey dont l'objectif est de cartographier une partie du ciel en identifiant l'ensemble des objets célestes (Nielsen, 2011). Un deuxième exemple est une initiative réalisée par des chercheurs d'un centre psychiatrique qui ont cherché à valoriser des bases de données déjà existantes (Menger et al., 2016). Le troisième exemple est celui de l'émergence d'une nouvelle science, appelée science des villes.

Deux exemples de processus knowledge-driven

EXEMPLE 1 : LA DECOUVERTE DE LA SYNTHESE DE L'UREE PAR HANS KREBS

La découverte de la synthèse de l'urée par Hans Krebs est un de ces récits authentiques d'une découverte scientifique qui a pu nous parvenir grâce à ces cahiers de laboratoire (pp. 445-446, Antonsson & Cagan, 2001). Le processus suivi par Krebs commence par un problème : découvrir les réactions en chaîne qui synthétisent la sécrétion de l'urée dans le foie. A l'époque, les biochimistes connaissaient déjà quelques réactions dans la chaîne mais aucun n'avait le processus global. Ce qui différenciait Krebs de ses contemporains résidait dans son protocole expérimental : au lieu de réaliser des expériences sur l'organe dans son ensemble, il utilisait des fines sections de tissus de celui-ci ce qui lui permettait de gagner un temps précieux. Il réalisa plusieurs expériences dans lesquelles il appliquait différentes substances sur ces tissus afin de mesurer le taux d'urée. Ce n'est qu'après de nombreuses tentatives et tâtonnements qu'il finit par obtenir les fortes concentrations d'urée qu'il espérait atteindre. Grâce à la littérature en chimie, il fût à même de proposer une plausible réaction qui combinaient l'ensemble des éléments qu'il avait mis en jeu dans son protocole expérimental.

EXEMPLE 2 : PASTEUR ET LE CHOLERA DES POULES.

Le cas de la découverte par Pasteur d'un vaccin pour la maladie prénommée le « choléra des poules » est un autre récit historique basé sur des cahiers de laboratoire du scientifique. Ces études menées par Pasteur et leurs résultats sont considérées comme décisifs dans la consolidation de la théorie sur la vaccination. Des travaux réalisés à l'époque par ses prédécesseurs suggèrent que lorsqu'un sujet contractait une maladie infectieuse sous sa forme la plus bénigne, cela préservait l'organisme de sa forme la plus grave (Cadeddu, 1985). Durant l'année 1879, Pasteur décide d'effectuer des recherches sur un cas de maladie infectieuse appelée choléra des poules. Pour ce faire, il met en culture la bactérie suivant différents protocoles pour ensuite les injecter aux poules et observer les effets sur leur mortalité. Par le fait d'un heureux hasard, Pasteur est obligé de s'absenter quelques mois de son laboratoire, laissant les cultures microbiennes se développer pendant des temps plus longs qu'à l'accoutumée. Après avoir inoculé ces cultures aux poules, il se rend compte que plusieurs d'entre elles contractent la maladie mais n'en meurent pas. A la suite, Pasteur et son assistant décident de procéder à une deuxième inoculation avec des cultures extrêmement virulentes et ils se rendent compte que les poules y résistent fort bien. Pasteur ne conclut pas directement qu'il a véritablement découvert un moyen de vaccination efficace, et il faudra encore mener plusieurs expérimentations pour atténuer au mieux selon lui les circonstances spécifiques de sa découverte.

Dans la science présentée comme knowledge-driven, le processus démarre à partir d'une hypothèse H : celle-ci est formulée à partir des connaissances existantes sur le problème, la détection d'anomalies à partir de cas expérimentaux ou l'utilisation d'analogies (Strevens, 2006). Ces connaissances s'appuient sur des théories, des modèles qui représentent l'ensemble de la connaissance accumulée sur le phénomène, mais également des observations réalisées sur l'objet d'étude. Le tout constitue un ensemble de prédicats sur un phénomène existant. Dans le cas où les prédicats sont déjà tous vérifiés et que de leur existence peut être formulée une hypothèse, alors celle-ci devient une cause probable pour le fait observé qui a une validité scientifique et qui justifie de son étude approfondie. A partir de l'hypothèse, le scientifique conçoit un processus

expérimental qui lui permettra de collecter les données scientifiques nécessaires à l'étude de la validité ou non de l'hypothèse. Comme l'illustrent les exemples, rares sont les expérimentations qui permettent de statuer directement sur la validité de l'hypothèse. Au contraire, le travail du scientifique est fait de tâtonnement, de reformulation et de réitération successives et multiples (Popper, 1959). Ainsi le scientifique modifiera probablement plusieurs fois son hypothèse au cours de ces expérimentations. En fait, chaque expérience réalisée génère un nouveau savoir que le scientifique intègre pour modifier son idée de base.

Trois exemples de processus data-driven

EXEMPLE 1 : LE PROGRAMME SLOAN DIGITAL SKY SURVEY (SDSS) ET LES GALAXIES PETIT POIS.

Le SDSS est un programme de relevé des objets célestes ayant pour but de cartographier 25% du ciel afin d'enregistrer des images relatives à plus de 100 millions d'objets célestes. Depuis l'année 2000, toutes les nuits, le télescope produit environ 200 Go de données, soit l'équivalent d'un petit disque dur d'ordinateur personnel. Une fois les images et leurs spectres collectées et triées, le SDSS rend disponible l'ensemble des bases de données au public. Ainsi, une image en couleur de n'importe quelle région du ciel couverte par une diffusion de données SDSS peut être obtenue simplement en fournissant ses coordonnées. Les travaux du SDSS constituent une des bases de données les plus citées dans l'histoire de l'astronomie : à ce jour, plus de 7 700 publications utilisent ou citent leurs données[4]. La taille de la base de données élaborée offre une opportunité unique de faire des découvertes scientifiques inattendues ou originales. La découverte des galaxies 'petits pois' en fait partie. En juillet 2007, un groupe de discussions d'amateurs intéressés par les bases de données du SDSS s'amuse de l'existence de plusieurs images représentant des objets célestes verts (Nielsen, 2011). Cette anecdote, au départ plutôt humoristique, est reprise plus tard par des scientifiques qui se rendent compte qu'ils sont face à un nouveau groupe de galaxies. Par une recherche plus systématique mise en place, le groupe de contributeurs a réussi à identifier au total 251 galaxies petit pois parmi le million d'objets existants dans les bases de données. Un an plus tard, en novembre 2009, les auteurs C. Cardamone, Kevin Schawinski, M. Sarzi, S. Bamford, N. Bennert, C. Urry, Chris Lintott, W. Keel et neuf autres personnes publient un article dans le MNRAS intitulé « *Galaxy Zoo Green Peas: Discovery of A Class of Compact Extremely Star-Forming Galaxies* ». Par la suite, ces travaux ont donné lieu à une série de publication sur les galaxies petits pois et ont fait émerger un nouvel objet d'étude dans le contexte astronomique (Amorín, Pérez-Montero, & Vílchez, 2010; Hawley, 2012; Izotov, Guseva, & Thuan, 2011; Pilyugin et al., 2012).

EXEMPLE 2 : LES DONNEES SUR LA PRATIQUE DE SANTE MENTALE.

Le département de psychiatrie du Centre Médical Universitaire d'Utrecht (UMCU) aux Pays-Bas est un centre de santé spécialisé dans les troubles affectifs sévères, les troubles psychotiques, les troubles du développement et les soins urgents. Comme beaucoup d'autres établissements de santé, plusieurs sources de données peuvent fournir des informations pertinentes aux médecins, telles que le dossier patient électronique (RPE), des mesures de laboratoire, des données d'imagerie et d'autres bases de données contenant des informations pertinentes. En externe, des données de recensement, des données géographiques et des données recueillies par d'autres institutions de soins contiennent également des informations potentiellement pertinentes. Afin de valoriser ces bases de données peu exploitées dans leur ensemble, l'UMCU a collaboré avec des techniciens spécialistes de l'analyse de données pour chercher à produire de la connaissance scientifique (Menger et al., 2016). Les techniciens se sont d'abord entretenus avec les médecins pour définir des thématiques de recherche en rapport avec les données disponibles. Ces données ont ensuite été triées et catégorisées suivant leur source, le type de données et leur structure. Les techniciens ont ensuite organisé des ateliers de travail avec des groupes de spécialistes de la santé : durant ces séances, les techniciens mettaient à disposition leurs compétences ainsi que des outils d'analyses et de visualisation tandis que les médecins orientaient l'exploration de par leurs découvertes progressives. En trois mois, 19 ateliers ont été organisées avec deux ou trois experts de domaine à chaque fois et au moins un expert technique. A la fin de ces ateliers, 29 hypothèses scientifiques ont pu être identifiées, toutes étant basées sur les données existantes. Sur ces 29 hypothèses, 24 n'avaient pas été imaginées ou suggérées par les spécialistes de santé durant les premiers entretiens préalables. A titre d'exemple, les médecins se sont rendus compte qu'ils existait une corrélation positive entre la saison durant laquelle est faite l'admission des patients et la longueur de cette admission. Sans fournir nécessairement de résultats scientifiques immédiatement valides, cette approche a permis de fournir aux médecins des pistes d'exploration pour de

[4] https://www.sdss.org/science/

futures recherches.

EXEMPLE 3 : L'EMERGENCE D'UNE SCIENCE DE LA VILLE.

Comme le souligne Marc Barthélémy, expert en physique statistique des systèmes complexes, nous sommes dans une période où la compréhension et la modélisation de la structure et de l'évolution des villes est plus importante que jamais[5]. La conception de ces modèles est cependant confrontée au manque d'une fondation théorique solide sur la science des villes capable d'expliquer et de représenter les transformations qui s'opèrent. L'approche à adopter pour développer cette science des villes semble donc être vouée à se construire d'abord sur des considérations empiriques. Si historiquement l'approche empirique est généralement basée sur un contexte où les bases de données sont rares et donc en faible quantité, il y a selon lui une forte opportunité à utiliser les nouvelles sources de données disponibles à portée de main. « Les nouvelles sources de données, telles que le GPS ou la téléphonie mobile en particulier, permettent une photographie instantanée et inédite de l'activité d'une ville et de sa structure. » Grâce à ces données, les scientifiques peuvent déterminer la position de chaque utilisateur et mesurer la densité à l'intérieur des villes. Ils ont ainsi pu modéliser des centres d'activité au sein des villes, et montrer que le nombre de ces centres varie suivant une relation sous-linéaire simple avec la population P. Or jusqu'à présent, aucun modèle théorique d'économie spatiale ainsi que des simulations agent-centré ne peuvent expliquer cette relation. Grâce à leurs relations, les scientifiques ont pu élaborer un modèle simple qui met en relation ces centres d'activités avec leur trafic routier, et ont ainsi pu montrer qu'une ville basée sur la voiture individuelle n'est pas durable.

Contrairement à la science knowledge-driven , l'approche data-driven utilise de façon hybride les inférences d'abduction, d'induction et de déduction dans la compréhension d'un phénomène (Kitchin, 2014). Plutôt que de découvrir des potentielles incohérences à partir d'un ensemble de savoirs scientifiques construits, les hypothèses sont formulées en prenant en compte les bases de données massives déjà disponibles dès le début du raisonnement. Le processus de génération des hypothèses n'est pas ancré dans les connaissances déjà existantes, mais plutôt « guidé » par celles-ci pour identifier les hypothèses susceptibles d'être examinées et testées. Dans le cas du centre de santé psychiatrique par exemple, les scientifiques vont structurer l'exploration en définissant des domaines d'intérêt basé sur les bases de données disponibles. Une des différences avec la science knowledge-driven est sur la façon de collecter les données. Dans les méthodes traditionnelles, les données sont collectées dans un but qui sert l'hypothèse initiale : elles arrivent donc après la formulation de l'hypothèse. Dans la science data-driven, les données sont déjà présentes en abondance avant de démarrer le processus. A noter que ces données ne sont pas générées par tous les moyens possibles et utilisables, mais généralement plutôt par des stratégies bien pensées et appuyées sur un corpus de savoirs existants. Pourtant, nous avons vu au travers des exemples que la science data-driven n'utilise pas nécessairement les données avec le même but que lorsqu'elles ont été collectées. Au contraire, les scientifiques sont plutôt dans une démarche opportuniste, où ils considèrent que la multiplication de sources hétérogènes de données permet d'augmenter les probabilités de trouver des hypothèses et des résultats intéressants. Au lieu de tester de façon automatique si chaque relation identifiée est véridique, l'attention se concentre sur les relations qui sont les plus probables ou au moins les plus valables scientifiquement parlant. Cette exploration opportuniste de larges bases de données limite la capacité des scientifiques à formuler clairement la problématique à laquelle ils souhaitent

[5] http://www.cea.fr/multimedia/Documents/publications/clefs-cea/CLEFS64-BIGDATA.pdf

répondre. Au lieu que ce soit la théorie qui formule l'hypothèse de base, les données aident le scientifique à formuler des hypothèses probables. La théorie n'est plus au cœur du processus de génération des hypothèses mais un appui sur lequel le scientifique se repose pour se guider dans l'exploration des données.

La dichotomie entre la science data-driven et la science knowledge-driven remet en cause la place de la donnée au cœur du processus scientifique. Alors que celle-ci était traditionnellement la résultante des expériences menées par les scientifiques ou des observations préalables, elle devient un élément omniprésent du début à la fin du processus. Elle ne sert plus uniquement à valider les hypothèses ou les questionnements du scientifique ; elle est en même temps l'outil qui va les aider à construire leurs problématiques. C'est à partir de la donnée que la science data-driven construit ses questionnements et expose ses hypothèses de manière opportuniste. D'une ressource rare, la donnée est devenue abondante et constitue une matière première pour la génération des hypothèses scientifiques. Quel est l'impact de ce nouveau mode de raisonnement « data-driven » sur la gestion de la production scientifique ? Nous présentons ci-dessous comment l'intégration des sciences citoyennes permet de répondre à une partie des contraintes de production, tout en ajoutant de nouvelles contraintes. Nous suggérons également que le modèle traditionnel de problem solving implicitement utilisé pour le crowdsourcing est insuffisant pour étudier la génération d'hypothèses data-driven et sa délégation à une foule.

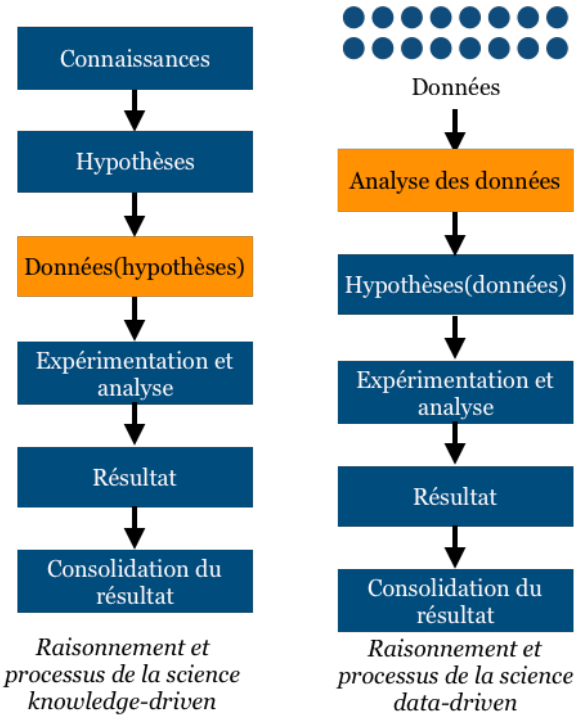
2.2. IMPACT DE LA SCIENCE DATA-DRIVEN SUR LE MODELE DE PERFORMANCE DE PRODUCTION DE CONNAISSANCES

Cette transformation du raisonnement scientifique impacte l'organisation du processus de production scientifique. Dans le modèle knowledge-driven, les modélisations du processus scientifique séparent généralement l'activité cognitive comme la génération ou le choix des données des activités expérimentales - collecte de données, réalisation des expériences entre autres (e.g. King et al., 2009; Klahr & Dunbar, 1988; Klahr & Simon, 1999; Kulkarni & Simon, 1988). Les données sont collectées en fonction des hypothèses scientifiques préalablement formulées, puis les hypothèses sont raffinées suivant les découvertes menées au cours du processus. Ainsi, il est possible de répartir les activités entre acteurs : les scientifiques par exemple vont s'occuper de générer des hypothèses tandis que d'autres acteurs (laborantins, assistants, citoyens par exemple) sont en charge de l'expérimentation et de la collecte des données. La productivité dans ce processus de production de connaissance est limitée par les systèmes de capteurs utilisés par les scientifiques pour collecter les données (Baraniuk, 2011; Miller, 2010).

Dans le processus data-driven, les données sont en abondance dès le début du processus. Ainsi, les contraintes des scientifiques ne sont plus sur leur capacité à collecter des données ni à les stocker (sauf dans certains cas extrêmes, voir l'exemple du LHC), mais dans leur capacité à analyser ces données pour formuler des hypothèses. En passant d'un environnement pauvre en données à une avalanche de données accessibles (Miller, 2010), le « *bottleneck* » ou goulot

d'étranglement qui conditionne l'amélioration de la production de nouvelles hypothèses change de nature (Sivasubramanian, Selladurai, & Gunasekaran, 2003) : on passe d'une contrainte liée où ce sont les données qui conditionnent l'amélioration de la production de connaissance vers une contrainte où ce sont les capacités de traitement et de formulation de nouvelles hypothèses.

Dans ce cadre, il est légitime de se demander si l'utilisation des citoyens de la science ne peut pas s'étendre à la génération des hypothèses pour répondre à ce manque de ressources. Nous avons vu que l'un des intérêts majeurs des sciences citoyennes est de fournir en grande quantité des acteurs capables de travailler sur des activités scientifiques. Ce système permet d'amplifier la capacité de production de manière momentanée le temps de l'ouverture du processus, mais également de réduire fortement les coûts. En effet, les projets de science citoyenne rémunèrent rarement le travail réalisé par les citoyens de la science. Au contraire, embaucher le même nombre de scientifiques pour multiplier le nombre d'explorations et d'analyses des données pour générer des hypothèses augmenterait d'autant le coût de la production.



Figures 2 et 3. Comparaison entre les deux modes de production de connaissance scientifique (inspiré de Shmueli, 2010). Les cases en orange symbolisent les goulots d'étranglement (« bottleneck »).

L'ouverture de cette phase à des citoyens de la science est cependant problématique, car comment gérer ou organiser des acteurs sans rétribution ni connaissance *ex* ante de leurs compétences est une question difficile. Si la délégation de la tâche à la foule réduit les coûts et les délais en terme d'exécution de la tâche, elle réduit en même temps la maîtrise de la qualité de ce qui est produit. Contrairement à une organisation classique où la tâche est déléguée en interne à un individu dont les compétences sont reconnues, ici les organisateurs ne maîtrisent ni les compétences des

participants, ni leur motivation à réaliser la tâche. Pour améliorer la fiabilité du système, plusieurs stratégies peuvent être mises en œuvre par les organisateurs : mise en place d'une redondance des éléments du système, amélioration de la fiabilité des sous-éléments qui constituent le système (mais potentiellement avec des coûts plus élevés), nouvelle conception du système, ou encore faire une combinaison de tout ce qui précède (E. Fyffe, W. Hines, & Kee Lee, 1968). Etant donné qu'il n'est pas possible de contrôler de façon sûre les entrées dans le processus, la redondance est la méthode qui semble être la plus appropriée pour réduire les risques de ne pas formuler de nouvelles hypothèses (chaque participant formule une ou plusieurs hypothèses).

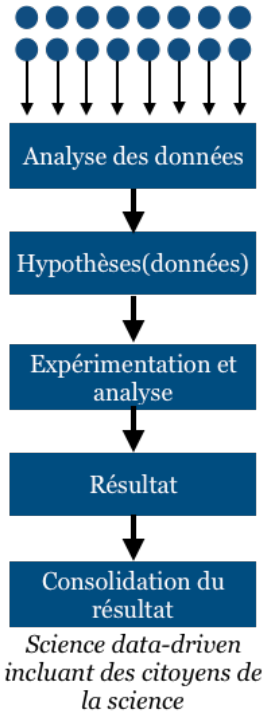


Figure 4. Ouvrir la génération des hypothèses data-driven à la foule.

La mise en place d'une redondance dans le processus de génération des hypothèses par la foule fait réapparaître les problématiques que nous avions identifiées pour des activités classiquement déléguées aux citoyens de la science dans le cas où on considère la science citoyenne comme une organisation scientifique éphémère. En effet, comment gérer la collaboration au sein des projets ainsi que la transmission d'informations entre les projets reste une question centrale dans le cas de la génération des hypothèses data-driven.

2.3. La generation d'hypotheses par les citoyens de la science

La science data-driven interroge également sur la place du scientifique au sein du processus. Le scientifique est traditionnellement défini comme le garant d'un savoir spécifique à son domaine qui lui permet de choisir les directions de recherche à prendre et de formuler les hypothèses adéquates : c'est lui en effet qui recherche la reconnaissance de ses pairs pour pouvoir obtenir des ressources supplémentaires pour avancer dans sa recherche. Or, nous avons vu que dans la science data-driven, la génération des hypothèses est principalement portée par l'existence des bases de données dès le début du processus scientifique. Ainsi, la place de la connaissance scientifique qui va de pair avec le rôle du scientifique n'est plus aussi déterminante. Dans l'exemple du SDSS, ce sont des volontaires non scientifiques qui ont découvert l'existence des galaxies pois verts, sans avoir de connaissances préalables sur le sujet. Les scientifiques ne sont intervenus que dans un deuxième temps, pour constater que le sujet n'avait jamais été traité dans la littérature. Jusqu'où est-il possible de remplacer le scientifique n'est pas une question récente. Durant le 20è siècle, plusieurs tentatives ont été menées pour automatiser le processus de découverte scientifique, remettant en cause place du scientifique (R. D. King et al., 2009; Kulkarni & Simon, 1988; Sen, 2010). En revanche, ces réflexions ont été appliquées dans des environnements spécifiques sans que les résultats soient répliqués dans d'autres contextes. Quelle serait la place du scientifique dans un processus data-driven où ce sont les citoyens de la science qui génèrent les hypothèses ?

2.4. Modele pour etudier la formulation d'hypotheses data-driven

Alors que les études des projets de science citoyenne se concentrent généralement sur des tâches comme la collecte, la codage ou l'analyse de données, les scientifiques ne considèrent pas le cas où ce sont les citoyens de la science qui vont générer les hypothèses directement à partir des bases de données disponibles. Pour analyser ce processus, la première idée serait d'utiliser le modèle de raisonnement implicite du crowdsourcing - le *problem solving* (Afuah & Tucci, 2012; Brabham, 2008). Le principe du *problem solving* suivant Herbert Simon consiste en l'utilisation de méthodes existantes ou *ad hoc* afin de trouver une solution à un problème défini. Ce processus est généralement modélisé comme l'exploration - procédure qui mène à la solution la plus satisfaisante, sans nécessairement être la plus optimale (Simon, 1955) - dans un espace de solutions afin d'atteindre le résultat final espéré (Simon & Newell, 1971).

A partir des années 1980, Simon chercha à analyser le processus de découverte scientifique grâce au principe du problem solving. Avec son collègue Kulkarni, il simula par exemple à l'aide d'un algorithme informatique KEKADA le raisonnement du chimiste Hans Krebs en matière de recherche sur la production d'urée (Kulkarni & Simon, 1988). Bien que leurs études aient grandement aidé à la compréhension du processus scientifique, des recherches ultérieures laissent

à penser que le modèle de résolution de problèmes à un espace n'est pas suffisant pour étudier la logique de découverte scientifique (Klahr & Simon, 1999). Par exemple dans KEKADA, le programme informatique sépare l'activité de formulation des hypothèses de l'activité expérimentale. En effet, alors que la génération des hypothèses est supportée par le programme informatique, la partie expérimentale reste réalisée par le scientifique qui ensuite implémente les résultats obtenus dans le programme[6]. Les scientifiques suggèrent que le nombre correct d'espaces à prendre en compte pour représenter le processus scientifique n'est pas fixe et dépend de la nature du contexte de découverte (Langley et al., 1987) : ainsi les modèles peuvent se baser sur deux (Klahr & Dunbar, 1988), trois (Thagard, 1998) voire quatre espaces (Schunn & Klahr, 1992) en fonction de ce qu'ils ont besoin de représenter. Si le nombre d'espaces à considérer n'est pas une donnée fixe, en revanche les chercheurs sont généralement d'accord avec la nécessité de prendre en compte deux espaces au moins : un espace pour formuler les hypothèses scientifiques et un espace pour expérimenter, collecter et analyser les données (Klahr & Dunbar, 1988).

Si le modèle classique du problem solving à un espace n'est pas suffisant pour représenter la logique de découverte scientifique, quel modèle pourrait correspondre qui nous permettrait d'étudier sa délégation via des projets de science citoyenne ? Alors que la plupart des projets de science citoyenne peuvent être interprétés comme de la résolution de problème (Franzoni & Sauermann, 2014), la génération d'hypothèses basées sur les données ne semble pas convenir à ce modèle : en effet, elle demande à priori de travailler à la fois à la construction de l'hypothèse scientifique tout en analysant les bases de données existantes. Nous avons besoin de construire un modèle de cette activité pour pouvoir l'étudier.

[6] Cette partie est étudiée plus en détail dans le chapitre 5 pour la construction d'un modèle de tâches déléguées aux citoyens de la science.

3. QUESTIONS DE RECHERCHE

Jusqu'à présent, peu d'études ont considéré les projets de science citoyenne comme un moyen d'améliorer la production de connaissance scientifique de manière répétée et intégrée dans l'organisation scientifique (Franzoni & Sauermann, 2014). Pourtant, les différents cas étudiés dans la littérature montrent que ce type d'organisation apporte des avantages importants pour les organisations scientifiques. Par notre analyse dans la littérature et le rapprochement des sciences citoyennes avec l'émergence d'un processus « data-driven », nous suggérons les différentes questions dans la littérature qui méritent d'être étudiées.

D'abord, nous avons formulé l'hypothèse qu'il existait un lien entre transformation du rapport aux données scientifiques et la multiplication de l'utilisation des projets de science citoyenne. Cependant, notre hypothèse se base uniquement sur notre observation du nombre de projets en rapport avec les données et de la transformation suggérée par les épistémologues d'une science knowledge-driven vers une science data-driven. De plus, les chercheurs ont plutôt tendance à justifier les projets de science citoyenne au travers d'une démocratisation des outils de communication. Cela constitue le canevas de notre première question de recherche : en quoi les données constituent-elles un facteur pour l'ouverture de la science à des citoyens et comment cette ouverture affecte le rapport du scientifique à la science?

Deuxièmement, le modèle de résolution de problèmes à la base du crowdsourcing ne semble pas suffisant pour étudier toutes les activités potentiellement délégables aux citoyens de la science. En effet, la génération d'hypothèses basées sur les données ne peut être réduite à l'exploration d'un seul espace et demande à la fois de générer les hypothèses et d'analyser les bases de données existantes. Il nous faut un modèle des activités déléguées pour étudier leur organisation et leur gestion de la productivité. D'où notre deuxième question : Quel modèle élaborer pour se rendre compte de l'impact que provoquent les données sur le processus de raisonnement scientifique?

Finalement, nous avons vu que la mise en place des projets de science citoyenne posait des questions en terme de gestion de la productivité. Alors que la fiabilité peut être gérée par un système de redondance de l'activité, d'autres critères nécessitent d'introduire des modes de gouvernance notamment les pertes de productivité durant le projet et entre les épisodes. Par ailleurs, la délégation de la génération des hypothèses aux citoyens de la science pose des questions sur la place du scientifique dans ce processus et la tâche qu'il lui est assigné. D'où notre troisième question de recherche : Comment gérer la performance d'une organisation scientifique éphémère?

Question de recherche 1 : En quoi les données constituent-elles un facteur pour l'ouverture de la science à des citoyens et comment cette ouverture affecte le rapport du scientifique à la science?

Question de recherche 2 : Quel modèle élaborer pour se rendre compte de l'impact que provoquent les données sur l'ouverture du raisonnement scientifique?

Question de recherche 3 : Comment gérer la performance d'une organisation scientifique éphémère?

CHAPITRE 3 – APPROCHE METHODOLOGIQUE ET PRESENTATION DU MATERIEL DE RECHERCHE

Ce chapitre a pour objectif de décrire l'approche méthodologique ainsi que l'itinéraire de recherche adopté pendant cette thèse. La recherche a été conduite dans le cadre d'un contrat doctoral au sein du Centre de Gestion Scientifique de l'école Mines ParisTech entre 2015 et 2019. En tant que chercheur en gestion, l'utilisation d'une méthode de recherche est souvent la conséquence d'un choix épistémologique de la part du chercheur. Cette épistémologie permet de cadrer le retour critique que l'on porte sur notre objet de recherche et sur la connaissance en elle-même que celui peut apporter afin *« de décrire, de comprendre, de prédire ou d'expliquer des phénomènes liées aux organisations »* (Ben Aissa, 2001). De manière générale, deux épistémologies sont présentes dans les disciplines de sciences sociales : l'approche positive et l'approche constructiviste. L'approche positiviste a longuement été prédominante comme épistémologie dans les sciences sociales suivant l'influence des travaux d'Auguste Comte pour qui le « mot positif désigne le réel » (Le Moigne, 1995). Dans cette représentation, le réel est régi par un ensemble de lois préexistantes dont le rôle de la science est d'en découvrir l'existence. L'objet de recherche est indépendant du chercheur qui a permis d'arriver à son élaboration. Cette approche implique cependant un certain nombre de principes issu de la logique aristotélicienne comme la notion d'identité, de non contradiction ou de tiers exclus qui ne peuvent être facilement soutenus dans le contexte de la gestion (David, 1999). A la place, les études de cas en science de gestion se basent plutôt sur une approche constructiviste qui considère qu'« un objet existe si on est capable de le construire, d'en exhiber un exemplaire ou de le calculer explicitement » (Largeaut, 1993). Le chercheur n'est plus indépendant de la construction de l'objet qu'il étudie, mais fait partie intégrante de ce processus. David (1999) propose une typologie des différentes démarches de recherche lorsque celles-ci sont basées sur une approche constructiviste en croisant deux critères pour les différencier : l'objectif du chercheur suivant qu'il produit une construction mentale ou concrète de la réalité, et la démarche que celui-ci met en œuvre en fonction de s'il part d'une observation des faits ou d'un projet de transformation ou d'une situation idéalisée.

		Objectif	
		Construction mentale de la réalité	Construction concrète de la réalité
Démarche	Partir de l'observation des faits	**Observation, participante ou non** Elaborer un modèle de fonctionnement du système étudié	**Recherche-action, étude clinique** Aider à transformer le système à partir de sa propre réflexion sur lui-même
	Partir d'un projet de transformation ou d'une situation idéalisée	**Conception de modèles de gestion** Elaborer des outils de gestion potentiels, des modèles possibles de fonctionnement	**Recherche-intervention** Aider à transformer le système à partir d'un projet concret de transformation plus ou moins complètement défini

Dans notre approche, nous partageons plutôt une vision proche de la recherche-action ou de la recherche-intervention selon laquelle la recherche en gestion n'est pas simplement une recherche sur l'action, mais plutôt une recherche dans l'action, « une recherche transformative où le chercheur, participant à la vie de l'organisation, conçoit, met en œuvre, analyse, communique, diffuse les résultats obtenus tant à l'intérieur de l'organisation auprès des praticiens, qu'à l'extérieur en direction des milieux académiques » (Lallé, 2004). Au lieu de se poser la question du « comment » à partir d'un objectif bien défini, le chercheur part de ses études de cas pour se poser à la fois la question du « comment » et du « pourquoi » (Yin, 2003). Le processus de recherche basé sur les études de cas constitue une stratégie de recherche globale, reposant sur de multiples sources de données, les données devant converger tout en bénéficiant du développement au préalable de propositions théoriques pour guider la collecte et l'analyse de données.

1. ITINERAIRE ET CADRE METHODOLOGIQUE GENERAL DE LA THESE

Dans notre étude, le projet de recherche initial se base sur le questionnement suivant : **quelles sont les logiques de gestion à mettre en place pour s'assurer de l'efficacité systématique des projets de science citoyenne dans le cadre d'un processus data-driven ?** Ce projet de recherche est issu d'observations manifestes dans des cas empiriques observés par le chercheur ou au travers d'exemples issus de la littérature mais dont l'interprétation n'est pas immédiate. Une variété d'approches a été mobilisée pour définir un modèle de gestion adapté aux projets de science citoyenne. Par la suite, chaque méthode est présentée en lien avec la question de recherche identifiée.

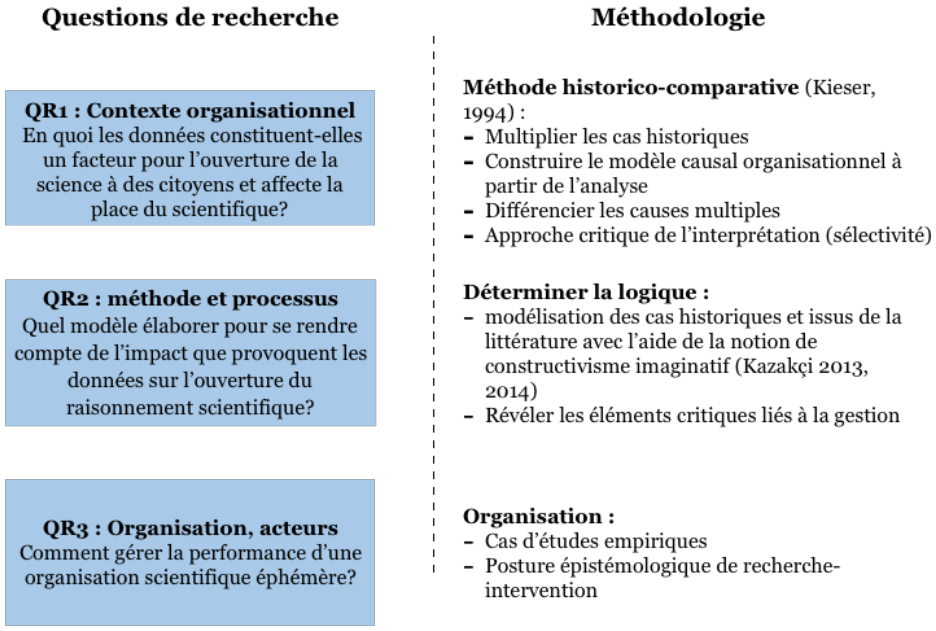


Figure 5. Méthodologie de recherche.

1.1. IDENTIFIER LES CAUSALITES ENTRE TRANSFORMATION DU RAPPORT AUX DONNEES ET ORGANISATION SCIENTIFIQUE

Pour mieux comprendre le contexte organisationnel contemporain et valider les liens supposés entre projets de science citoyenne et science data-driven, notre recherche suit une méthode historico-comparative du phénomène (Kieser, 1994). Cette approche a pour objectif de produire un modèle causal entre transformation du rapport aux données et ouverture de la science qui nous permette de justifier le lien supposé dans notre contexte contemporain. Notre étude se concentrera notamment sur l'apparition de nouveaux acteurs, la redéfinition du rôle des acteurs existants (celui notamment du scientifique), et proposera une ébauche de formalisme des différentes étapes du processus scientifique qui ont été ouvertes à de nouveaux acteurs. Au moins quatre raisons selon Kieser justifient l'intérêt de l'analyse historique pour étudier des phénomènes organisationnels contemporains : 1) L'analyse d'un comportement organisationnel ne peut être séparé d'un effet culturel inclus dans la dimension culturelle. 2) Recontextualiser des problèmes d'organisation contemporains à des situations similaires dans le passé pour éviter des effets idéologiques et des tendances actuelles « à la mode ». 3) Les analyses historiques nous apprennent à interpréter les structures organisationnelles existantes non pas telles que déterminées par les lois, mais comme le résultat de décisions prises dans le cadre d'opportunités de choix passés, certaines intentionnellement et d'autres implicitement. Des possibilités de choix qui n'avaient pas été utilisées à l'avantage des acteurs impliqués peuvent éventuellement se présenter à nouveau ou être restaurées d'une manière ou d'une autre. Les analyses historiques peuvent nous préparer à mieux identifier et à mieux utiliser les opportunités de choix. 4) En confrontant les théories du changement organisationnel aux évolutions historiques, ces théories peuvent être soumises à un test plus radical que celui qu'elles doivent passer lorsqu'elles sont simplement confrontées à des données sur les changements à court terme. En revanche, les analyses historiques au sein des organisations présentent un certain nombre de faiblesses qui peuvent être limitées en établissant un ensemble de bonnes pratiques méthodologiques.

D'abord, une étude restreinte à un seul cas type historique peut mener à des biais interprétatifs. Au contraire, les comparaisons avec d'autres situations similaires permettent d'augmenter la visibilité d'une structure en la contrastant avec une autre. Ainsi, dans notre étude nous allons analyser deux périodes distinctes dans lesquelles nous espérons retrouver des phénomènes similaires. D'abord l'introduction et l'émergence des instruments scientifiques entre le 17 et le 19è siècle au sein de la pratique scientifique. Nous nous appuierons sur les travaux de Christian Licoppe qui a analysé cette période à la fois en France et en Angleterre afin de réduire l'effet culturel possible de notre interprétation. Nous multiplierons les exemples durant la période pour justifier l'existence d'une tendance paradigmatique dans la science ou du moins dans un certain nombre de disciplines scientifiques. Nous étudierons également une deuxième période entre le 19 et le 20è siècle comme l'introduction de la stochastique dans l'analyse des phénomènes naturels dans certaines disciplines scientifiques. Le contraste entre ces deux périodes ponctuées de multiples exemples permettra de réduire ces biais.

Ensuite, Kieser propose une voie stratégique pour étudier les régularités causales que l'on peut extraire dans l'analyse historique. Au lieu de guider l'analyse sur une hypothèse préconçue issue d'une modélisation et d'une réflexion théorique antérieure unique, la construction de la relation causale est constamment modifiée et générée dans un dialogue constant avec les données historiques. Le chercheur n'essaie pas de prouver un modèle existant; au lieu de cela, il s'engage à générer des schémas de causalité capables d'expliquer les développements historiques. Ainsi dans notre approche méthodologique, nous suggérons un lien de causalité entre transformation du rapport aux données et réorganisation de la science sans imposer le schéma causal que nous découvrirons durant notre analyse historique.

Un autre problème est que les événements historiques ont toujours des causes multiples qui ne doivent pas nécessairement s'exclure mutuellement mais peuvent être complémentaires. Ainsi si plusieurs facteurs ont été identifiés en tant que causes possibles d'un événement historique, le chercheur doit alors préciser si une seule cause aurait suffi à en provoquer l'apparition.

Enfin, l'auteur suggère que le contenu historique est inépuisable et donc mène inévitablement à une sélection de la part du chercheur qui doit être justifiée. Dans notre cas, le choix des deux périodes que nous étudions n'est pas anodin. Ils représentent l'étendue de ce que l'on définit comme l'histoire de la science moderne. D'un point de vue organisationnel, ils cherchent à recréer un pont entre un mode de production de la connaissance où on passe du scientifique seul ou accompagné de quelques assistants à notre époque moderne constituée d'acteurs hétérogènes qui ne peuvent être définis comme scientifique.

1.2. Modelisation des projets de science citoyenne : dans quelles conditions les projets sont efficaces ?

Alors que les projets de science citoyenne s'intègrent dans un contexte de transformation du rapport aux données et de systématisation de son utilisation, il devient nécessaire de préciser les conditions qui permettent de s'assurer de son efficacité. Pour identifier quelles sont les caractéristiques nécessaires à la bonne gestion de ce type de projets, nous avons pour objectif de modéliser les différents types de projets de science citoyenne que nous pouvons rencontrer et d'analyser dans quelles conditions ce modèle est suffisant dans le processus data-driven.

Durant notre revue de littérature, nous avons montré que le modèle dominant pour étudier les projets de science citoyenne est celui du crowdsourcing. Or, celui-ci se base essentiellement sur l'augmentation de la diversité et n'est pas suffisant dans le cas où les participants peuvent interagir entre eux ainsi que pour la génération des hypothèses basées sur les données. Le but est donc de construire un modèle qui prenne en compte la systématisation des tâches afin de déterminer les critères qui permettent d'assurer l'efficacité des projets. Notre modélisation démarre initialement sur les exemples que nous avons rencontrés dans notre analyse historique ainsi que sur les cas contemporains de projets de science citoyenne. Nous nous appuyons sur les

théories sous-jacentes du modèle du crowdsourcing pour construire un modèle de tâche basé sur le principe de résolution de problème (Simon & Newell, 1971). Ensuite, nous analysons un ensemble de cas de modélisation informatique de la génération des hypothèses scientifiques pour montrer que le modèle de résolution de problème n'est pas suffisant pour interpréter cette tâche (e.g. King et al., 2009; Kulkarni & Simon, 1988). Nous proposons alors d'étendre le modèle initial en nous basant sur une théorie issue de la conception appelée « constructivisme imaginatif » (Kazakçi; 2013, 2014).

A partir de ce modèle formel, nous tentons de relier la systématisation des projets dans une organisation et les critères qui vont établir le lien avec son efficacité. Il est important de souligner que ce modèle n'a pas pour vocation à être une représentation robuste et exhaustive des projets de science citoyenne, mais plutôt de fournir les caractéristiques nécessaires pour s'assurer de l'efficacité des projets dans le cas où ils sont systématisés.

1.3. DEUX CAS D'ETUDE D'ORGANISATION POUR IDENTIFIER UNE FIGURE MANAGERIALE ET DES LOGIQUES ORGANISATIONNELLES

La troisième étape de notre recherche consiste à réinterroger notre modèle et ses implications théoriques dans un contexte empirique. L'enjeu principal est de comprendre quels sont les principes qui permettent de gérer les tâches qui n'apparaissent pas dans les modèles de gestion traditionnels. Nous adopterons dans notre approche une posture épistémologique de « recherche collaborative » (Shani et al., 2008) ou « recherche-intervention » (David, Hatchuel, & Laufer, 2012) menée par des chercheurs et des praticiens afin de créer des connaissances concrètes pour l'organisation et de nouveaux modèles théoriques pour la recherche en sciences de gestion (David & Hatchuel, 2008). Nous avons choisi d'inscrire notre recherche expérimentale dans deux contextes particuliers.

Nous étudierons d'abord le cas d'un outil de gestion, le RAMP (pour Rapid Analytics and Model Prototyping), développé par le Centre de Data Science de Paris-Saclay. Cet outil propose de développer des projets basés sur des problématiques et des bases de données fournies par des scientifiques de disciplines variées (économie, biologie, physique des particules,…). Chaque problème est formalisé comme un problème d'optimisation d'algorithme d'analyse de données et soumis à une foule de participants. Le RAMP est utilisé comme plateforme de compétition et de collaboration où les spécialistes des données travaillent sur un problème pour des délais relativement courts (généralement un ou deux jours). Cela peut être considéré comme une forme de Hackathon avec la principale différence que l'objectif est d'optimiser une métrique claire. L'intérêt de la plateforme est qu'elle a été conçue également comme outil d'observation du travail des data scientists. Des mesures sont réalisées durant les challenges pour lesquelles il est possible d'extraire des informations quantitatives sur les trajectoires des participants. En collaboration

avec l'équipe organisatrice de la plateforme, nous avons mené un travail d'analyse des données à partir d'outils statistiques ainsi que de méthodes de visualisation. Cette analyse a permis de mettre en avant les caractéristiques critiques sur l'efficacité de la plateforme et donc a aidé à aiguiller dans les choix organisationnels mis en œuvre.

Le deuxième terrain expérimental est Epidemium, un programme de recherche collaboratif basé sur l'épidémiologie, qui s'est déroulé entre novembre 2015 et mars 2018. Epidemium, financé en partie par les Laboratoires Roche, a pour mission de rassembler une communauté autour de de bases de données massives afin d'explorer ces bases de données pour générer des hypothèses scientifiques. L'étude expérimentale a été menée en partie avec Olga Kokshagina, une collègue chercheuse ayant travaillé au sein du laboratoire du CGS. Elle est basée à la fois sur un ensemble de documents, une participation active pour le déroulement et l'organisation de certains évènements, une communication avec les organisateurs et les participants, et la recherche de mise en place d'outils pour piloter l'efficacité du projet.

2. CONTEXTE DES TERRAINS DE RECHERCHE DANS LEURS DOMAINES DE SCIENCE

L'analyse et les principaux résultats organisationnels de la thèse sont basés sur les deux terrains que nous avons présentés ci-dessus. Ils présentent plusieurs caractéristiques qui justifient de leur intérêt méthodologique pour notre analyse. D'abord, la plateforme RAMP nous permet d'avoir accès à des données uniques pour étudier le comportement de participants durant un projet de type science citoyenne. En plus d'un accès privilégié aux résultats et aux solutions de tous les participants, les organisateurs ont intégré des métriques qui permettent d'analyser le déroulement de la collaboration entre les participants et d'estimer son efficacité. Ce type de métriques est unique dans les projets de type challenge. Un autre élément est que la philosophie gestionnaire portée par la plateforme suppose l'émergence de l'importance des méthodes d'analyse des données regroupées sous le terme de « sciences de données ». Nous présentons dans cette section le contexte du RAMP et de sa construction.

Ensuite, le programme Epidemium est un projet unique d'ouverture de la génération d'hypothèses basées sur les données à une foule de participants. Ce programme est caractéristique d'un besoin grandissant dans de plus en plus d'organisations scientifiques qui cherchent à valoriser les larges bases de données qui leurs sont disponibles. Pourtant, il n'existe pas de méthodologie pour construire des hypothèses scientifiques à partir de bases de données massives. Nous montrons dans cette section que notre étude se place dans un cadre plus large d'épidémiologie populaire qui inclut les acteurs de la société dans le processus scientifique. Ce processus est amplifié dans le cadre d'une explosion des données de santé qui pousse les scientifiques à chercher des solutions pour générer de la valeur à partir de ces bases de données.

2.1. CADRE D'ETUDE POUR LE RAMP : LES « SCIENCES DES DONNEES »

Des quantités massives de données sont produites dans l'environnement scientifique actuel (Miller, 2010). Au-delà des problèmes d'infrastructure et d'ingénierie, l'extraction d'informations à partir de données est devenue un défi majeur pour les universités et exige de repousser les limites des techniques d'analyse actuelles et de développer des avancées radicales. Cet impératif a donné lieu à la notion de « science des données » dans les milieux universitaires (Agarwal & Dhar, 2014; Davenport & Patil, 2012). La science des données peut être définie au sens large comme la conception de méthodes automatisées pour analyser des données massives et complexes afin d'extraire des informations utiles. La science des données et le Big data sont étroitement liés mais ne concernent pas les mêmes problématiques. Alors que le Big data couvre un large éventail de thèmes sur la capture, le transfert, le stockage, la recherche, le partage sécurisé, l'archivage et l'analyse de données massives, la science des données se concentre sur les aspects algorithmiques et mathématiques de l'extraction de nouvelles connaissances à partir de données. En tant que telle, la science des données se situe à la croisée de l'informatique, des mathématiques appliquées et des statistiques. Le débat académique sur la notion de science des données s'accompagne d'initiatives institutionnelles à travers le monde. Par exemple, après l'annonce de l'Initiative nationale de R&D sur les données massives de la Maison Blanche en 2012, les agences de financement nationales (NSF, NIH et DARPA) et les universités ont mené des actions à grande échelle pour promouvoir la science des données et la recherche sur le Big data. Les cas suivants méritent d'être mentionnés. La Research Data Alliance (RDA) est créée pour accélérer l'innovation axée sur les données dans le monde entier grâce au partage et à l'échange de données de recherche. L'Université de New York a inauguré son Centre for Data Science. L'Université de Washington a fondé son institut eScience. Berkeley a lancé son Institute for Data Science. Les fondations Moore et Sloan ont annoncé une initiative interinstitutionnelle de 37,8 millions de dollars sur cinq ans pour soutenir les trois instituts précédents. Harvard a créé le Harvard Data Science Initiative pour accélérer la collaboration entre la recherche et l'enseignement. En Europe, l'Université d'Amsterdam a annoncé la création de son centre de recherche Data Science. L'Université d'Edimbourg a lancé son Centre de formation doctorale en science des données. L'Imperial College London s'est associé à l'université de Zhejiang pour lancer une collaboration en science des données. Depuis 2016, l'école Polytechnique en France a créé l'Initiative Data Science qui regroupe une équipe de chercheurs et d'enseignants afin de fournir des réponses concrètes à des questions qui émanent des milieux industriels.

Ces initiatives sur les sciences de données sont généralement conçues et organisées de manière à trouver et recruter temporairement un chercheur spécialiste en intelligence artificielle et en analyse de données, dans l'espoir que des progrès significatifs seront réalisés durant cette période. En effet, un principe largement partagé dans les communautés scientifiques considère que le manque de spécialistes des données constitue un des problèmes majeurs des projets basés sur les données complexes (Davenport & Patil, 2012). Cependant, les retours d'expérience sur les projets

sur les données tendent à montrer que cette vision est limitée, et que le problème essentiel est plutôt organisationnel (Kazakçi, 2015; Kégl et al., 2018). En général la pratique du processus scientifique s'organise dans un système où les chercheurs sont incités à faire carrière par le biais de publications qu'ils produisent et la recherche d'une reconnaissance par les pairs (Merton, 1957; Merton & Storer, 1973). Cette conception de la science néglige le fait qu'une grande partie de la science se fait aujourd'hui dans des communautés dispersées aux sein de différentes organisations de R&D (von Zedtwitz, Gassmann, & Boutellier, 2004). La principale limite de ces processus scientifiques basés sur les données est cette « incapacité de la communauté scientifique dans son ensemble à comparer rapidement et à évaluer la valeur scientifique d'un jeu de données et de sa configuration analytique » (Kégl et al., 2018). Entre les années 1990 et les années 2000 la recherche en intelligence artificielle était construite autour de jeux de données de référence qui permettaient une comparaison entre les différents algorithmes. Cependant la situation a grandement évolué et les scientifiques sont maintenant dotés de plusieurs algorithmes très performants sur des critères de performance établis. Ainsi le problème n'est plus d'obtenir la performance d'un algorithme suivant une configuration donnée, mais plutôt de se concentrer sur la configuration en elle-même (Kégl et al., 2018).

Depuis une dizaine d'années, la communauté en intelligence artificielle et machine learning a réalisé des efforts importants pour standardiser les algorithmes de modèles prédictifs par le biais de plateformes tel que Scikit-learn (Vanderplas, J. et al., 2011) ou Keras. Ces travaux ont permis de grandement faciliter la réutilisation des modèles existants et de distinguer la phase d'expérimentation de la phase d'optimisation dans la construction de modèles prédictifs. En effet, séparer la conception des processus de la partie optimisation garantit que l'optimisation ne démarre que lorsque l'expérience est entièrement spécifiée. La recherche en intelligence artificielle et en machine learning s'est organisée sous forme de défis appelés « data challenges », dans lesquelles des communautés ou des personnes indépendantes acceptent de travailler sur des problèmes scientifiques majeurs afin d'optimiser une métrique bien définie. La configuration standard des data challenges est un concours de compétition pur - les participants opèrent en parallèle sans communication et le ou les gagnants obtiennent une récompense. La principale hypothèse de ces challenges est qu'un grand nombre de participants augmenteront les chances de trouver des solutions exceptionnelles. La littérature montre que, lorsqu'une forte incitation est présente (Boudreau, Lacetera, & Lakhani, 2011), cette configuration est susceptible de stimuler les efforts fournis par les participants et d'augmenter la qualité globale des solutions (Afuah & Tucci, 2012). D'un autre côté, cette configuration présente l'inconvénient de ne pas capitaliser pleinement sur la production entière de la foule, puisque seules les solutions gagnantes sont divulguées à la fin (Kazakçi, 2015). Des idées potentiellement bonnes qui ne donnent pas de succès immédiat sont perdues (Boudreau & Lakhani, 2015). C'est un handicap important pour les problèmes nécessitant une collaboration étroite entre les domaines (par exemple, la physique) et les spécialistes de l'analyse de données. Ce constat a été l'un des principaux moteurs du Center for Data Science (CDS) de l'université Paris-Saclay, une des initiatives de data science, qui a souhaité

fournir une plateforme plus flexible où les configurations collaboratives et compétitives étaient possibles. C'est dans ce cadre que la plateforme RAMP a été créée.

En conclusion, le besoin grandissant de compétences spécifiques pour appliquer les méthodes de sciences de données couplées à une standardisation de ces méthodes ont grandement facilité et poussé les scientifiques à exploiter des ressources externes notamment sous la forme de projets d'analyse des données. La performance et de l'efficacité de la répétition de ces projets est donc une question fondamentale.

2.2. L'EPIDEMIOLOGIE COMME CADRE D'ETUDE POUR LE PROGRAMME EPIDEMIUM

2.2.1. Avalanches de données dans le domaine médical

Les scientifiques du domaine médical et les acteurs du domaine de l'épidémiologie font face à une explosion du nombre de données accessibles et de leur variété (Chiolero, 2013). Alors qu'une grande partie des données relatives à la santé tel que la tenue de registres administratifs, les données relatives aux patients (notes et ordonnances écrites du médecin, imagerie médicale, laboratoire, pharmacie, assurance et autres données administratives) étaient généralement stockées sous forme de papier, la tendance actuelle est à la numérisation en grandes bases de données (Raghupathi & Raghupathi, 2014). Par ailleurs, des instituts internationaux compilent des bases de données massives sur des sujets relatifs à la santé et les rendent accessibles au grand public. A titre d'exemple, l'Organisation Mondiale de la Santé a mis en place depuis 2012 une base de données de veille sanitaire accessible aux habitants des 194 Etats membres comprenant plus de 1000 indicateurs. Les grandes bases de données que constitue l'ensemble des publications scientifiques sont maintenant stockées numériquement et peuvent être plus facilement analysées et traitées afin de faire émerger des problématiques scientifiques (Bohannon, 2017; Gray, 2009; Sybrandt, Shtutman, & Safro, 2017). De la même manière, l'analyse des résultats d'essais cliniques prend également une toute autre dimension à l'ère des données massives (DerSimonian & Laird, 2015). Enfin, les données issues des plateformes de réseaux sociaux tel que Twitter ou Facebook peuvent potentiellement être une source d'information pour la pharmacovigilance (Bian, Topaloglu, & Yu, 2012).

Les rapports indiquent que les données du seul système de santé américain ont atteint, en 2011, 150 exaoctets et atteindront bientôt l'échelle des zettaoctets (1021 gigaoctets) et, peu après, le yottabyte (1024 gigaoctets) (Cottle & Hoover, 2013). Pour les spécialistes du Big Data, il existe, parmi cette vaste quantité de données, une opportunité. En découvrant les associations entre les variables que constituent les bases de données, l'analyse du Big Data a le potentiel d'améliorer les soins, de sauver des vies et de réduire les coûts en augmentant le rôle de la médecine préventive.

2.2.2. L'épidémiologie entre médecine, statistique et science populaire

2.2.2.1. Avalanche de données en épidémiologie

L'épidémiologie a un rôle fondamental à jouer dans cette transformation. Cette discipline impliquant à la fois les médicaments et les statistiques étudie principalement les facteurs de risque associés à l'incidence ou à la mortalité des maladies. Depuis les années 1950, les études épidémiologiques ont utilisé des méthodes statistiques qui leur permettent d'extrapoler les résultats obtenus sur des échantillons à des populations beaucoup plus importantes. Cette approche a conduit à l'émergence de nombreuses études sur les facteurs de risque comportementaux tels que l'exposition à l'alcool, le tabagisme ou la nutrition. Les biais statistiques dans l'échantillonnage affectent cependant l'extrapolation des phénomènes locaux et plusieurs études ont mis en évidence que les résultats sont parfois contradictoires sur des facteurs de risque similaires : par exemple, un aliment peut prévenir et favoriser le cancer selon différentes études (Schoenfeld & Ioannidis, 2013). L'émergence récente de bases de données massives sur l'incidence et la mortalité des maladies est considérée dans la communauté de l'épidémiologie comme une opportunité pour des études épidémiologiques susceptibles de réduire les problèmes actuels et les limites des approches existantes, et de récentes études montrent comment les méthodes de machine learning peuvent être utilisées en épidémiologie (Szymczak et al., 2009).

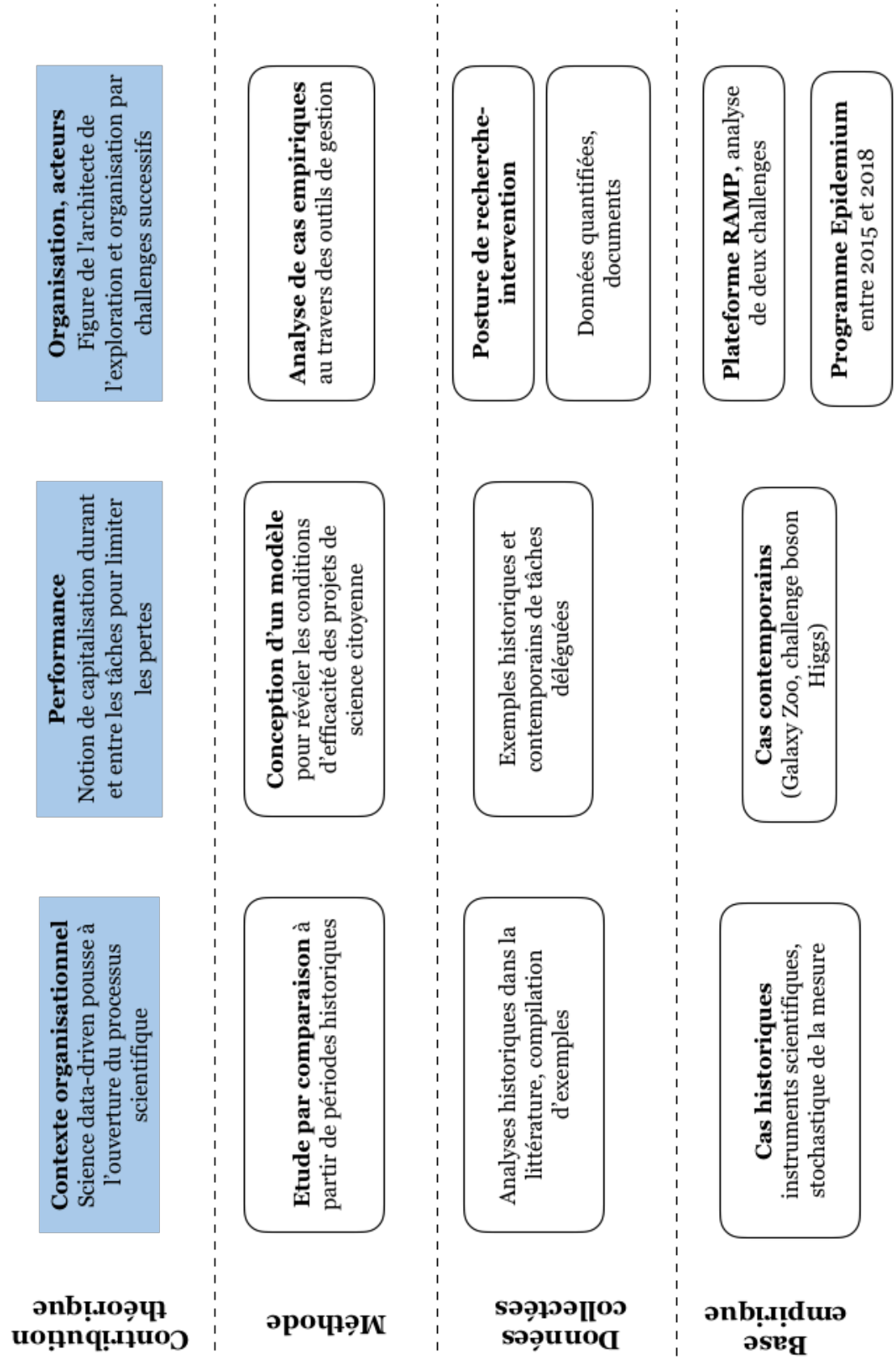
2.2.2.2. Une tradition récente de « l'épidémiologie populaire »

La recherche épidémiologique est généralement réalisée en laboratoire par des épidémiologistes ou des spécialistes de la santé comme des oncologues ou des cliniciens. Cependant, une forme d'épidémiologie fondée sur une large participation du public telle que rencontrée dans les sciences citoyenne fait l'objet d'un courant de recherche depuis la fin des années 1980 sous le terme « épidémiologie populaire ». L'épidémiologie populaire désigne le processus par lequel des citoyens collectent eux-mêmes des données et mobilisent des connaissances scientifiques pour comprendre la distribution et les causes d'une maladie (Barthe, 2013). Cette notion a été créée par le sociologue Phil Brown pour qualifier le travail d'enquête réalisé par les riverains d'un site contaminé afin d'établir l'origine des leucémies infantiles qui frappaient leur communauté (Brown, 1987, 1992). Par la suite, la notion a été mobilisée dans un certain nombre de champs scientifiques relatifs à la santé environnementale (Barthe, 2013). Si cette forme de pratique scientifique intéresse les chercheurs en sciences sociales, elle crée également un certain nombre de débats entre les professionnels de la santé pour reconnaître ce type d'enquête comme viable d'un point de vue scientifique. Ainsi, certains chercheurs militent pour trouver des moyens d'intégrer cette pratique en créant de nouveaux protocoles de recherche, tandis que d'autres chercheurs y voient l'apparition d'une pratique dangereuse car pouvant intégrer au débat public des résultats sans valeur scientifique.

Notre étude se place dans cette tendance d'une épidémiologie populaire, à la différence que les données n'ont pas été collectées par les participants mais par un ensemble d'acteurs hétérogènes,

puis ont été choisies et compilées par des équipes de scientifiques. De plus, la validation d'un nouveau résultat scientifique ne provient pas uniquement du travail des participants, mais est mesuré, évalué et reconnu comme tel en accord avec des spécialistes de la santé et des méthodes épidémiologiques. La tâche que nous analysons ici, à savoir la formulation d'hypothèses basées sur les données, n'est pas synonyme de ce qui pourrait être considéré comme un retour aux méthodes épidémiologiques de la première heure qui consistaient à proposer des analyses descriptives des objets étudiés sans se baser sur des méthodes de statistiques inférentielles (Schwartz in Lechopier, 2010). En effet notre démarche n'est pas purement inductive mais cherche au contraire à relier des effets à des causes. Par le fait, notre étude porte sur la formulation des hypothèses scientifiques et la vérification de ces hypothèses grâce aux données qui sont disponibles. La validation ensuite de cette construction comme résultat scientifique ou non dépendra ensuite de la fiabilité des données collectées ainsi que de la pertinence des méthodes employées. Suivant la pertinence des résultats obtenus, l'hypothèse pourra faire l'objet dans un cadre ultérieur d'une étude plus approfondie grâce à la collecte de données spécifiques.

3. SYNTHESE DE L'ITINERAIRE DE RECHERCHE ET DES METHODES CHOISIES



PARTIE 2 – ELABORATION D'UN CADRE D'ANALYSE DES PROJETS DE SCIENCE CITOYENNE

CHAPITRE 4 – REPARTITION DES ACTIVITES SCIENTIFIQUES ENTRE ACTEURS ET REMPLACEMENT DU SCIENTIFIQUE DANS LE PROCESSUS : APPROCHE HISTORIQUE

RESUME DU CHAPITRE 4

Ce chapitre a pour objectif de produire un modèle causal entre transformation du rapport aux données et ouverture de la science qui nous permette de justifier le lien supposé dans notre contexte contemporain. Nous nous basons sur deux périodes dans l'histoire de la science notamment pour éviter des biais interprétatifs: l'introduction entre le 17è et le 19è siècle des instruments scientifiques ; l'apparition de la mesure stochastique de phénomènes naturels dans plusieurs disciplines scientifiques entre le 19 et le 20è siècle. Les observations de ces périodes historiques nous permettront de construire un schéma causal où l'évolution du rapport aux données scientifiques implique une complexification du processus. Les scientifiques, garants de la méthodologie ne peuvent plus supporter tout le processus par manque de ressources ou de savoir-faire et doivent déléguer une partie de l'activité stratégique. Historiquement, les scientifiques ont fait preuve de volonté de concentrer leurs activités sur les parties les plus essentielles – construction théorique, génération des hypothèses – et de ne déléguer que des activités dans lesquelles leur valeur ajoutée est minime par rapport aux autres acteurs. Ainsi nous retraçons la lente délégation de la construction des instruments scientifiques aux artisans et les conflits qui y sont associés. Ensuite, nous montrons comment l'approche stochastique dans la mesure a poussé à intégrer en grand nombre des acteurs pour répondre aux nouveaux besoins en ressources. Le paradigme data-driven et l'organisation de projets de science citoyenne présentent plusieurs similitudes avec le modèle causal : organisation sous forme de division du travail, manque de ressources et de compétences, transformation du rapport aux données, interrogation sur la place du scientifique dans le processus de production. Cette analyse historique donne toute la légitimité de s'interroger alors sur la place du scientifique dans notre contexte et sur une possible évolution.

L'analyse historique nous permet également d'engager un travail de modélisation du raisonnement scientifique à partir des cas de délégation historiques couplés aux exemples contemporains. Nous établirons un modèle basé sur la notion de tâche pour caractériser les différentes activités déléguées au cours de l'histoire : les tâches élémentaires, les tâches de type recette et les tâches de type résolution de problèmes.

Nous avons suggéré dans la première partie de cette thèse que les transformations liées aux données peuvent avoir un impact sur l'organisation du processus scientifique. En particulier, un ensemble de chercheurs contemporains considèrent que nous vivons à notre époque une transformation profonde des caractéristiques des données, assimilable à un changement de paradigme de la science (Dubois et al., 2000; Kelling et al., 2009; Kitchin, 2014; Menger et al., 2016). Cette transformation va amener à une dichotomie entre deux sciences : une science traditionnelle dite « knowledge-driven » où les scientifiques génèrent des hypothèses à partir des connaissances existantes puis collectent des données pour vérifier la véracité de leur hypothèse ; une science « data-driven » où les hypothèses sont formulées à partir de bases de données existantes et le processus n'est pas ancré dans les connaissances déjà existantes, mais plutôt « guidées » par celles-ci pour identifier les hypothèses susceptibles d'être examinées et testées.

L'émergence de cette pratique de la science pousse à interroger sur l'organisation scientifique et la place du scientifique au sein du processus de production de connaissance. D'un côté, le scientifique est garant de la connaissance qui lui permet de formuler les hypothèses scientifiques. La relation entre les autres acteurs au sein du processus scientifique est généralement une position dominante menant à une relation hiérarchique (Chompalov, Genuth, & Shrum, 2002; Walsh & Lee, 2015) et une division forte du travail entre les individus suivant leurs compétences. Les non scientifiques comme les techniciens ou les assistants par exemple, bien que baignés dans la culture scientifique (Cole, 2011), ne possèdent pas généralement les connaissances spécifiques aux scientifiques mais développent à la place d'autres types de connaissances qui prennent d'autres formes (Polanyi, 1966) : certains techniciens de laboratoires seraient plus compétents que les scientifiques à suivre des protocoles, ou des techniciens en radiologies seraient meilleurs à lire les radios.

D'un autre côté, la génération des hypothèses dans le paradigme « data-driven » se fait principalement de manière opportuniste à partir de bases de données déjà disponibles dans le processus (Menger et al., 2016). La place de la connaissance scientifique dans la science data-driven n'est donc plus clairement identifiée comme elle l'est dans la science knowledge-driven (Kitchin, 2014). De plus, la formulation des hypothèses demande également de posséder des compétences en analyse et en interprétation de bases de données massives. Enfin, nous avons vu que l'avalanche des bases de données massives déplaçait le goulot d'étranglement du processus de production de connaissance vers l'analyse des données pour générer des hypothèses. Il est de ce fait légitime de se demander si le processus de génération des hypothèses scientifiques basée sur les données pourrait effectivement être délégué au public.

Ce n'est pas la première fois dans l'histoire que le scientifique se retrouve à déléguer une tâche relative aux données à des non-scientifiques. Il est courant dans les grandes organisations scientifiques que la collecte de données par exemple soit réalisée par des assistants et des laborantins. De même, il est rare maintenant de croiser des scientifiques qui construisent eux-mêmes leurs instruments de mesure. Quelles sont les activités liées aux données que le

scientifique a progressivement déléguées et comment se sont-elles organisées ? Quels sont les liens entre cette délégation et le rapport que les scientifiques avaient avec les données ? Comment cela a-t-il remis en cause et remet-il en cause le rôle du scientifique en fonction de ce qui est délégué ? Avoir une réponse à ces questions nous permettrait de mieux appréhender les problématiques de gestion contemporaines que nous rencontrons dans le cadre du paradigme data-driven et des sciences citoyennes.

Pour traiter le sujet, nous réalisons une étude longitudinale historique de l'évolution du rapport aux données dans la science. Nous nous intéressons dans ce chapitre à deux moments dans l'histoire de la science où les scientifiques ont changé leur rapport aux données: l'introduction entre le 17è et le 19è siècle des instruments scientifiques ; l'introduction de la mesure stochastique de phénomènes naturels dans plusieurs disciplines scientifiques entre le 19 et le 20è siècle. Nous montrerons que les données et leurs transformations ont poussé durant ces périodes à une réorganisation de la science, et ont mené les scientifiques à intégrer de nouveaux acteurs. Nous nous pencherons également sur la façon dont cette ouverture du processus scientifique a redessiné la place des scientifiques et des nouveaux acteurs au sein du processus scientifique et la façon dont les différentes tâches ont été réparties. Nous tracerons les limites implicites que les scientifiques ont établies entre les tâches du processus scientifique qui ont été déléguées et celles qui ne le sont pas, et dans quelles conditions.

1. ETUDIER L'HISTOIRE POUR DETERMINER LES LIMITES DE L'OUVERTURE DE LA SCIENCE

Des historiens des sciences ainsi que des sociologues ont apporté des éclairages fondamentaux sur l'évolution de la science (e.g. David, 2007; Kuhn & Ian, 1962; Licoppe, 1996; Schaffer, 1988; Shapin, 2008). Notre intention dans ce chapitre est de nous appuyer sur ces études pour souligner quelques moments clés dans l'histoire des sciences où le processus scientifique, initialement porté par un ensemble plus ou moins établi d'acteurs identifiés, a multiplié et organisé l'ouverture à d'autres acteurs non scientifiques. En nous intéressant aux transformations du rapport aux données, nous cherchons à comprendre les impacts qu'ils ont pu avoir sur l'organisation de la science et sur les relations entre les acteurs impliqués dans le processus. Nous nous appuierons sur les études menées par ces historiens ainsi que sur des documents d'époque, généralement des récits originaux de savants qui ont narré les différentes expériences qu'ils ont menées.

On pourrait nous reprocher une représentation unidirectionnelle et transdisciplinaire de la science. Cela n'est pas notre propos. Chaque discipline et chaque communauté scientifique développent leurs propres courants de pensée et méthodologies. Nous supposons en revanche que certaines périodes de l'histoire ont été marquées par des phénomènes qui ont touché un grand nombre d'acteurs et de disciplines à un moment donné. Dans cette analyse, nous nous contentons

d'éclairer ces moments dans l'histoire des sciences au travers d'exemples pour comprendre quelles sont les tensions qui ont mené les scientifiques à ouvrir et déléguer certaines de leurs activités.

L'évolution paradigmatique des sciences dans leur rapport aux données a déjà été le sujet d'intérêt de Jim Gray dans le livre *The Fourth Paradigm* publié en 2009 (Hey, Tansley, & Tolle, 2009). Selon lui, nous sommes arrivés à ce qu'il nomme le quatrième paradigme, symbolisé par l'émergence du big data et de la science dite data-driven, les trois autres paradigmes étant : la science expérimentale et empirique post-baconienne, la science théorique et la modélisation, et enfin la science computationnelle et la simulation de phénomènes complexes. Il montre dans son analyse comment les scientifiques ont construits les outils adaptés pour compiler et gérer des bases de données de plus en plus importantes. Son approche permet de mieux caractériser la transformation actuelle d'un point de vue épistémologique, cependant elle néglige les acteurs qui interviennent dans le processus scientifique et ne permet pas de rendre compte des leviers qui poussent à l'ouverture du processus.

Nous proposons dans notre analyse d'étudier deux périodes liées à la transformation du rapport aux données. Une première période est caractérisée par l'introduction des instruments scientifiques de manière généralisée dans les pays occidentaux. Des instruments comme le microscope, le télescope ou le baromètre ont permis aux scientifiques d'observer et de mesurer des phénomènes naturels qui étaient inaccessibles jusqu'alors. Nous allons montrer comment ces appareils ont été introduits puis généralisés dans les communautés de savants entre le 17 et le 19è siècle. Nous nous intéresserons particulièrement à l'impact de cette transformation sur l'organisation du processus scientifique. La deuxième période que nous analysons est celle du passage à une approche stochastique dans l'analyse des données scientifiques. Comme nous le montrerons, le 19è et le 20è siècle ont été le témoin de l'introduction de la notion de hasard et d'incertitude dans l'étude des phénomènes naturels. Nous montrerons au travers d'exemples comment certains scientifiques ont intégré dans leurs modèles cette vision stochastique et comment cela a impacté l'organisation scientifique.

2. DU 17E AU 19E SIECLE : REDEFINIR LE ROLE DU SCIENTIFIQUE FACE AUX FABRICANTS D'INSTRUMENTS SCIENTIFIQUES

Dans cette section, nous analysons comment l'introduction et la démocratisation des instruments scientifiques de mesure a poussé les scientifiques à changer d'organisation et à intégrer de nouveaux acteurs dans le processus scientifique. Cette période a été modélisée suivant trois régimes distincts par l'historien Christian Licoppe (1996). D'abord un régime de curiosité au 17è siècle avec la recherche de faits extraordinaires dans des représentations publiques. Les expérimentations n'avaient pas pour objectif de construire de nouvelles connaissances scientifiques mais plutôt de convaincre un public majoritairement adeptes du syllogisme. Ensuite un régime de l'utilité où les différentes disciplines se sont moins intéressées à émerveiller un public qu'à stabiliser les expérimentations et entrevoir la possibilité de reproduire les expériences. Enfin, un régime de l'exactitude caractérisé par une recherche de standardisation de l'activité scientifique. Notre analyse portera sur l'évolution des rôles des acteurs gravitant autour du processus scientifique et des relations entre ces acteurs.

2.1. L'INSTRUMENT SCIENTIFIQUE COMME OUTIL AU 17E SIECLE POUR PRODUIRE DES FAITS EXTRAORDINAIRES

Le 17è siècle est l'âge d'or des premiers appareils d'optique : « microscope», « télescope » ou encore « baroscope » permettent d'observer des phénomènes qui restaient inaccessibles aux sens humains. Ces instruments sont construits par les physiciens ou les astronomes eux-mêmes ou encore par des artisans mais sous l'étroite surveillance des premiers. L'instrument scientifique représente à cette époque[1] le premier mouvement d'opposition fort des intellectuels de l'époque contre les savoirs secrets des alchimistes, mais également contre la rhétorique syllogistique provenant des écrits grecs anciens et considérés jusqu'alors comme source première de vérité. Il y a en effet une scission à l'époque entre partisans d'un raisonnement syllogistique où la production de connaissance est la conséquence unique et directe d'un pur exercice de pensée, et l'émergence d'une méthode empirique qui prône la connaissance basée sur l'expérience. Une petite partie de savants développent de coûteux instruments pour réaliser devant des témoins des expérimentations qui suscitent la curiosité de leur public. Par la construction de faits extraordinaires, ces démonstrations tendent à opposer la vision traditionnelle de l'*experientia* (la

[1] On peut remonter à l'antiquité grecque pour voir l'utilisation d'instruments scientifiques (Daumas, 1950; Gaulon, 1997). Cependant, à part des esprits curieux en avance sur leur temps, comme Roger Bacon ou Pierre de Maricourt, « il faudra attendre la chute de Constantinople (1453) et le reflux vers l'Europe de l'ouest des érudits byzantins avec leurs originaux des oeuvres grecques, pour déclencher la Renaissance scientifique du XVIe siècle. » (Gaulon, 1997). Le 16è verra une résurgence de l'instrument scientifique au travers de figures comme Francis Bacon ou Tycho Brahe.

recension des lieux communs) propre à l'épistémologie grecque et encense le régime de l'*experimentum*, la mise à l'épreuve artificielle. Après la gloire d'un mode de raisonnement essentiellement porté sur le syllogisme, ces philosophes portent dans leurs démonstrations la valeur d'une rigueur de l'expérience, où l'on fait attention à distinguer fortement les faits de leurs interprétations. Le but à ce moment n'est pas de démontrer mais de convaincre un maximum de personnes que l'expérience locale menée avec les instruments et dans les circonstances particulières peut être généralisée.

Progressivement, les nombreuses représentations extraordinaires réalisées devant les aristocrates et les puissants ainsi que le succès international des travaux du physicien Isaac Newton favorisent la diffusion du processus expérimental et de l'utilisation des instruments scientifiques. Newton se présente comme fervent défenseur de la méthode expérimentale et rejette unilatéralement le syllogisme des grecs anciens. Il devient la figure de proue d'un mouvement de démocratisation de la science expérimentale et un exemple à suivre pour tous les savants qui suivront. La célèbre phrase qu'il écrivit dans son livre Principes édité en 1702 gouvernera la philosophie du siècle à venir : « Hypotheses non fingo » (Je ne feins pas d'hypothèses) (Cariou, 2009).

Figure 6. Illustration de Galilée proposant une expérience publique au doge de Venise (peinture de Giuseppe Bertini, 1858)

Dans ce régime de faits extraordinaires, seuls les scientifiques sont représentés dans le processus. Ce sont eux qui conçoivent et fabriquent leurs instruments scientifiques, parfois aidé par des artisans ; c'est encore eux qui font les expérimentations et collectent des données. Bien que les récits historiques aient démontré l'existence de « techniciens invisibles » (Shapin, 1989), leur rôle est généralement peu décrit dans les récits de découverte : certains non-scientifiques sont de simples assistants qui aident à faire tourner l'instrument scientifique tandis que des assistants de

l'ombre participent activement à l'élaboration du résultat scientifique ou de la fabrication des instruments.[2]

2.2. STANDARDISER LE TRAVAIL DU FABRICANT D'INSTRUMENT

2.2.1. Problèmes de reproductibilité des expériences au 18è siècle

Après ce premier régime de l'expérimentation instauré dans les communautés de savants, un deuxième régime va se constituer dès le début du 18è siècle que Licoppe qualifie d'«utile ». L'objectif à partir de ce moment est de réussir à stabiliser les résultats des expériences afin de les diffuser dans des cercles plus étendus que les aristocrates et les seuls expérimentalistes. Les expériences menées avaient été jusqu'alors plutôt qualitative et ayant pour objectif de proposer des démonstrations extraordinaires, mais dont la reproductibilité est difficile à mener. Les causes en sont multiples. D'abord une trop faible connaissance théorique des phénomènes mène parfois à des interprétations trompeuses de la part des philosophes naturels, qui fournissent des protocoles erronés à leurs pairs. Ensuite, les instruments scientifiques sont généralement des pièces uniques fabriquées par les scientifiques eux-mêmes ou avec l'aide de quelques artisans. Une différence de résultats dans les expériences peut alors être interprétée à tort comme étant la cause d'un instrument de mauvaise qualité. La chose se traduit par une volonté des philosophes naturels de rationaliser les moyens de mesures utilisés et permettre que le résultat d'une expérience soit le moins possible remis en cause. La qualité de l'instrument fabriqué devient une fonction essentielle pour réussir à produire collectivement une connaissance fiable. L'instrument scientifique n'est plus l'objet d'une curiosité de la part des philosophes mais devient instrument dans le sens où celui-ci se doit d'être suffisamment fiable pour étudier les phénomènes naturels.

L'EXPERIENCE DU BAROMETRE LUMINEUX

Le cas du baromètre lumineux vers la fin du 17è siècle est un bon exemple des difficultés de reproductibilité que peuvent rencontrer les scientifiques. C'est en transportant un baromètre à mercure dans un voyage de nuit que l'abbé Picard fait une découverte surprenante: à chaque mouvement brusque du métal provenant des mouvements de la voiture, le mercure contenu dans le baromètre devient légèrement lumineux. Ce phénomène intéressa un de ses amis philosophes, Philippe de la Hire, qui tenta au début sans succès de reproduire le phénomène que jusqu'à présent seul Picard avait observé. De la Hire ira jusqu'à démonter et remonter à plusieurs reprises l'ancien baromètre de l'abbé Picard pour parvenir en avril 1694 à reproduire lors d'une expérience une lumière assez vive. Un autre scientifique, l'astronome Cassini, observera également que son baromètre émet une lumière, mais d'une manière légèrement différente que celle du baromètre de de la Hire.

La luminosité du baromètre reste une curiosité jusqu'à ce que le mathématicien Jean Bernoulli soit au courant de ce phénomène et cherche à le reproduire avec son propre matériel. Il y parvient, tandis que les tentatives menées par les membres de l'Académie des Sciences à la même époque étaient restées jusqu'alors infructueuses. Pour expliquer l'échec de ses pairs, Bernoulli suppose la présence d'une pellicule de saleté qui apparaît au frottement du mercure avec l'air et qui vient obstruer les pores du mercure et donc empêche la luminosité de celui-ci. Son raisonnement l'amène à concevoir deux épreuves expérimentales pour éviter cette contamination : une première où il aspire le mercure avec sa bouche et recrache les premières gouttes

[2] Un contre-exemple notable de cette relation entre assistants et scientifiques est la collaboration mise en œuvre par Tycho Brahe au sein de l'observatoire d'Uraniborg qu'il aura fait construire au 16è siècle (Christianson, 2000). Cas unique d'organisation de cette taille à l'époque, Tycho Brahe était entouré d'une famille d'assistants qui le soutenaient dans les expérimentations et les observations du ciel.

souillées, une seconde où il utilise la machine à vide afin de bénéficier d'un environnement avec une quantité faible d'air pour limiter la contamination. Bernoulli envoie une lettre aux académiciens en leur suggérant de se baser sur ses méthodes. Cependant, les académiciens réussissent l'expérience sans recourir à la méthode de Bernoulli. A la place, ils suggèrent que c'est la qualité du mercure de Bernoulli qui explique l'apparition d'une pellicule de saleté.

Bernoulli rétorque les arguments des académiciens et soutient que s'il avait été là lors des expériences ordinaires, il aurait mis en avant des spécificités propres aux conditions d'expérimentations. Il envoie alors son propre mercure aux académiciens et leurs propose de réitérer l'expérience en se basant sur trois règles : partir d'une fiole très sèche, ne pas remuer le mercure avant de le vider, et réaliser un très bon vide dans la fiole. Les académiciens qui tenteront de reproduire l'expérience échoueront une seconde fois. Ils en tirent alors leurs propres conclusions, à savoir que la pellicule de saleté à enlever suggérée par Bernoulli n'est pas nécessaire. A la place, ils suggèrent que c'est le contact du mercure avec d'autres substances qui le dote d'une certaine disposition à être lumineux. Ils soutiennent cet argument en précisant qu'ils ont rendu des baromètres lumineux juste après avoir nettoyé le mercure à la chaux.

Les discussions continueront entre Bernoulli et les membres de l'académie, chacun étant convaincu que son modèle est le bon et que les difficultés rencontrées chez l'autre proviennent de spécificités liées la qualité de l'instrument utilisé ou sur le protocole mis en place. En fait, la construction de la conjoncture supposée de Bernoulli ressemble plus à un bricolage de sa part. Il cherche à faire correspondre les faits observés à une prétendue explication plus générale : à chaque itération où le système se voit mis en défaut par de nouvelles expérimentations, notamment les expériences faites par les membres de l'académie, Bernoulli propose une parade en produisant de nouvelles règles ou indique des précautions supplémentaires à suivre et qui sont nécessaires pour reproduire le phénomène. Au lieu de remettre en cause son système, Bernoulli préfère critiquer la qualité des expérimentateurs et la fiabilité de leurs dispositifs. La véracité des expériences ne peut malheureusement compter que sur la présence de spectateurs ou de témoins qui pourront attester de ce qu'ils auront vu, ainsi que par la précision du compte-rendu de l'expérience menée. C'est en cela que le phénomène de luminosité du baromètre échoue à être considéré comme autre chose qu'une curiosité, car les expérimentateurs n'arrivent pas à se mettre d'accord sur un protocole général et reproductible, attestant de la réalité de celui-ci. Il émerge le besoin de rendre l'expérience indépendante du contexte dans laquelle elle a été organisée, ainsi que de celui qui réalise l'expérience. L'expérience n'est plus un phénomène d'intérêt mais doit également être reproductible : elle n'a plus rien de mondain, et les mesures devront être prises dans des lieux et des horaires prédéfinis et des conditions de mesure similaires.

Dès lors que le régime de l'utilité s'impose face au régime de curiosité, un nouveau public d'ingénieurs et d'artisans se densifie au sein des activités d'expérimentations (Morus, 2016). Les fabricants d'instruments y jouent un rôle crucial dans la révolution scientifique de l'époque (Daumas, 1953). Maurice Daumas, chimiste et historien français suggérait que «les scientifiques du XVIIe siècle, dont la plupart étaient également des artisans, n'auraient pas pu créer leur appareil sans la collaboration de l'artisan professionnel». Si les philosophes naturels étaient généralement les concepteurs de ces nouveaux instruments, il fallait également un travail régulier de l'artisan pour reproduire et améliorer l'objet afin d'en faire un instrument à usage quotidien. L'argument porté par Daumas est que les constructeurs d'instruments ne pouvaient pas être simplement considérés comme des acteurs mettant en œuvre les concepts développés par les philosophes, mais que leurs compétences artisanales étaient un élément essentiel à l'accomplissement d'un modèle de la science que défendaient les partisans de l'expérimentation. Il existerait selon lui un ensemble de connaissances tacites propres aux artisans que les savants n'avaient pas et qui ont aidé à faire prospérer la recherche.

Au cours du 18è siècle, les instruments servant à l'astronomie tel que les astrolables et les télescopes sont de plus en plus populaires. Des compétences spécifiques comme la fabrication du verre pour les lentilles sont très recherchées et la taille des verres est un sujet de vif intérêt pour

plusieurs savants de l'époque. Bien que cela ne soit pas leur activité principale, plusieurs s'employèrent à son étude tel que Descartes, Huygens, Leuwenhoeck ou Leibniz dans l'espoir d'améliorer la qualité souvent médiocre des instruments d'optique de l'époque (Daumas, 1950). Les savants étaient souvent fascinés par la capacité des artisans à fabriquer des instruments de grande qualité.

Certains de ces savants cherchent à rationnaliser la fabrication des instruments en étudiant les méthodes des artisans. Ce travail d'appropriation des savoir-faire artisanaux se fait en injectant les nouvelles lois de la physique dans les pratiques professionnelles. Pour eux, le travail fait par les artisans pourrait être grandement amélioré s'il était mieux encadré. Le comte de Buffon par exemple réalise et diffuse des tables révisées sur la résistance du bois. Il critique ouvertement le travail des artisans : « Le principal usage du Bois dans les Bâtiments et dans les constructions de toute espèce est de supporter des fardeaux ; la pratique des ouvriers qui l'employent, n'est fondée que sur des expériences, à la vérité souvent réitérées, mais toujours assez grossières ; ils ne connaissent que très imparfaitement la force et la résistance des matériaux [3] ». Pourtant, l'extraction des connaissances tacites des artisans est complexe. En effet, la communication est difficile à cause d'une forte différence de statut social : les artisans étaient des hommes de métier plutôt que des gentlemen à une époque où la philosophie naturelle était de plus en plus associée à la bourgeoisie. De plus, les artisans cherchent à conserver leur savoir-faire dans le secret pour conserver un ascendant financier et un avantage concurrentiel. Ce point est illustré par les relations qu'entretenait Christiaan Huygens avec les artisans spécialisés dans la fabrication des lentilles pour les télescopes et les microscopes.

LA FABRICATION DES LENTILLES PAR HUYGENS

Lorsque, en 1685, l'astronome Huygens entreprend la rédaction d'un traité intitulé *De Telescopiis et Microscopiis*, il se retrouve vite limité par son manque de compétences sur le travail du verre[4]. Grand amateur du métier de l'artisanat, il cherchait à comprendre et maîtriser la fabrication de la lentille. Cependant, comme il l'explique, la qualité du verre sur lequel il travaillait de même que ses capacités étaient insuffisantes pour obtenir de bons résultats. De plus, l'artisanat était un monde entouré de secret et les maîtres lunettiers avec lesquels il était en contact et qui savaient construire de grand verre avaient « chacun leur manières et méthodes qu'ils ne veulent pas que d'autres sachent ». La fabrication des instruments scientifiques s'avérait en effet être un commerce lucratif que les artisans préféraient garder secret pour le monopole. Les hommes qui fabriquaient ces instruments étaient des artisans qualifiés qui avaient généralement intégré le métier par le biais de l'apprentissage et de l'appartenance à une guilde. Huygens dans ses recherches a rendu visite à plusieurs reprises au fabricant de télescopes Philippe-Claude Lebas lors à Paris. Lebas semblait avoir trouvé une méthode de polissage supérieure à ce qui existait à l'époque. Malgré tous ses efforts, il n'a jamais été en mesure de découvrir exactement le fonctionnement de cette nouvelle méthode. À la mort de Lebas, Huygens tenta même de persuader la veuve de Lebas de bien vouloir révéler la méthode, mais en vain, car elle aussi protégeait le secret de son mari (Louwman, 2004). Malgré ces

[3] BUFFON, « Expériences sur la force du bois (premier mémoire) », *op. cit.,* p. 453

[4] Christiaan Huygens, Oeuvres complètes. Tome XXI. Cosmologie (ed. J.A. Vollgraff). Martinus Nijhoff, Den Haag 1944

barrières, Huygens s'évertua à observer et à comprendre le travail des artisans avec lesquels il avait l'occasion de travailler ou d'échanger, ce qui lui permis d'extraire des procédures précieuses pour son ouvrage. Il n'hésite pas dans ses écrits à admettre que quelques-unes des techniques de fabrication proviennent directement des ouvriers avec lesquels il travaille : « En polissant [le verre] commence a reluire par tout également, et c'est un des avantages des grandes formes, a ce que dit l'ouvrier, car dans les petites tousjours les bords demeurent un peu moins polis » ; ou encore après avoir vu que Lebas mettait sous la forme, pour l'empêcher de plier elle-même, « quelques ronds de carton qui obéissent tant soit peu quand on travaille le verre », qu'il serait peut-être une bonne idée de placer une plaque de plâtre pour le fortifier.

2.2.2. Le tournant du 19è siècle : définir le statut du scientifique face aux fabricants d'instruments

Au cours du 19è siècle, la fabrication d'instruments devient de plus en plus centralisée dans les grandes villes telles que Londres ou Paris. Les expérimentations de grande taille confient de plus en plus certains de leurs processus à des grandes entreprises avec des compétences bien spécifiques (Sorrenson, 2013). Face à ces transformations, les fabricants d'instruments cherchent à défendre leur savoir-faire spécifique vis-à-vis d'un simple technicien. Le tournant du 19è siècle est rude pour ces derniers : c'est une période où les savants luttent ouvertement pour se distinguer du rôle de fabricant d'instrument et où l'artisanat devenait une voie moins prometteuse pour devenir membre d'institutions prestigieuses. Ceci est bien illustré dans le conflit qui a opposé le philosophe naturel Charles Babbage et son technicien Joseph Clément (Morus, 2016; Schaffer, 1994).

> Babbage avait fait appel à Clément dans les années 1820 pour concevoir des composants du moteur sur une machine capable de calculer des tables de navigation et d'astronomie. Après une collaboration fructueuse, les deux hommes se sont retrouvés face à un dilemme. Ils n'étaient pas d'accord pour savoir à qui appartenait quoi dans ce qui avait été produit. Clément estimait que les outils qu'il avait conçus et fabriqués pour produire les composants du moteur lui appartenait, comme il était coutume de faire à l'époque. De plus, il souhaitait pouvoir reproduire le moteur de calcul sans la permission de Babbage. De son côté, Babbage proposait de nationaliser le moteur, les outils et les conceptions, ce qu'il estimait être la solution généralement suivie dans une grande partie de l'industrie. Au-delà des différends de propriété sur le moteur et les outils nécessaires à sa fabrication, les ingénieurs et artisans voyaient dans cette démarche un défi à leurs droits et à leur savoir-faire. A l'opposé, Babbage considérait que l'apport des artisans provenaient de connaissances tacites et donc toute production ne pouvait être considérée comme appartenant formellement à quelqu'un.

Il y a dans ce différend le symptôme de deux visions qui se confrontent : d'un côté celle de l'artisan qui cherche à tout prix à valoriser son métier et un savoir-faire manuel ; de l'autre côté, des philosophes qui veulent séparer dans le processus scientifique tout ce qui relève d'un travail de la pensée au dépens du savoir-faire manuel. A partir du début du 19è siècle les savants dressent un compte-rendu de ce qui relève de la philosophie naturelle (étymologie ancienne du scientifique) et de ce qui n'en relève pas. Dans son livre *Preliminary Discourse of the Study of Natural Philosophy* publié en 1845, John Herschel défend un point de vue dans lequel la

philosophie naturelle est ouverte et transparente, tandis que l'artisanat est fermé et secret. Selon lui, c'est la notion d'accessibilité qui est déterminante pour distinguer ce qui est scientifique de ce qui ne l'est pas. La science doit se « dépouiller, autant que possible, de difficultés artificielles et dépourvu de toute technicité telle qu'elle tend à la situer à la lumière d'un métier et d'un mystère, inaccessible sans une sorte d'apprentissage [5]». L'artisanat est relégué au second plan, comme un moyen de faire de la science mais qui ne peut se confondre avec la noblesse de la philosophie naturelle. La fabrication d'instruments scientifiques est devenue essentielle à la création de nouvelles connaissances, mais les fabricants d'instruments ne sont pas considérés comme des hommes de science.

Progressivement, la demande de fabrication d'instruments scientifique devient tellement importante que celle-ci se voit déléguée à des industries composées d'ingénieurs et d'artisans, où les processus de standardisation permettront de vendre à bas coût et à beaucoup de savants des instruments de bien meilleure facture. L'artisan solitaire ne peut plus répondre aux exigences de fabrication et à la standardisation nécessaire pour une reproductibilité des phénomènes observés. De plus, les outils se complexifient et demandent d'intégrer de nouvelles techniques dans les instruments que les artisans ne connaissent pas forcément.

2.3. UNE REORGANISATION DU PROCESSUS SCIENTIFIQUE ET UNE REDEFINITION DU ROLE DU SCIENTIFIQUE

L'introduction de l'instrument fait apparaître une relation entre organisation de la science et changement du rapport aux observations. D'un côté, l'introduction des instruments de mesure tels que le télescope, le microscope ou le baromètre donnent accès à de nouvelles données sur des phénomènes naturels. Progressivement, les simples sens du scientifique ne sont plus suffisants pour pratiquer la recherche savante, et l'instrument s'impose comme une nécessité pour progresser dans la compréhension du monde. L'effet de communauté qui joue également un rôle dans cette démocratisation, et notamment le besoin de reproductibilité pousse les savants à se munir des dernières techniques et technologies. De l'autre côté, l'émergence de ces instruments dans le quotidien du scientifique fait apparaître les failles d'une organisation où le savant travaille seul, ou accompagné de quelques assistants sans besoin de coordonner véritablement la collaboration. En effet, la conception et la fabrication des instruments scientifiques demande au scientifique d'acquérir des connaissances techniques et artisanales ainsi que de consacrer un temps généralement long à leur conception. Il faut à la fois penser l'instrument, le construire, mais également concevoir et réaliser les expérimentations avec cet instrument. Ces nouvelles activités poussent les scientifiques à déléguer une partie à d'autres acteurs. Cela donne lieu à des

[5] Herschel, John. 1831. *Preliminary Discourse on the Study of Natural Philosophy*. London: Longman, Reese, Orme, Browne, Greene and Taylor. p. 70.

collaborations entre scientifiques et artisans qui eux étaient porteurs d'un savoir-faire manuel pour construire les instruments.

Cette collaboration s'est progressivement confrontée à des tensions : d'un côté des artisans souhaitaient valoriser leur art et conserver un ascendant sur leur technique en gardant leur connaissances secrètes ; de l'autre côté, une communauté de savants qui considéraient que la connaissance devait être accessible à tous. Les scientifiques de l'époque se sentaient menacés par les artisans et leur savoir-faire. En effet, si les artisans savaient eux construire les instruments, et pouvaient même parfois accéder au rang de savant, que restait-il aux scientifiques qui faisait leur spécificité ? La communauté scientifique mit en place au 19è siècle deux stratégies pour asseoir une forme de supériorité de leur activité face aux artisans et aux techniciens. La première était de démystifier le travail des artisans : les scientifiques ont cherché à appliquer systématiquement leurs méthodes d'analyse sur le métier de l'artisanat. Leur objectif était à la fois de comprendre, de rationaliser et d'améliorer les pratiques. Progressivement, la fabrication des instruments scientifiques est devenue industrielle, portée la plupart du temps par des entreprises situées dans les grandes villes. La deuxième stratégie a été de définir le rôle du scientifique par opposition au métier de l'artisanat. Les scientifiques ont délaissé la plupart des activités relatives à la fabrication des instruments scientifiques, et se sont positionnés comme les garants de la conception de ces instruments et de la construction de la connaissance scientifique. Finalement, ce conflit a poussé à une forme de division du travail entre acteurs qui a amené à rationaliser la fabrication des instruments scientifiques et mieux redéfinir le rôle du savant dans le processus de production de connaissance. Notons que nous ne réduisons pas la réorganisation du processus scientifique à l'introduction de l'instrument scientifique. Ce que l'on peut définir comme la période moderne ou la révolution industrielle du 19è siècle avec les théories d'Adam Smith ont joué un grand rôle réflexif dans l'organisation du processus scientifique. Nous mettons cependant en avant comment l'introduction et la multiplication de nouveaux acteurs dans le processus scientifique a mené à des conflits organisationnels qui ont trouvé une solution dans un mode d'organisation rationalisé et dans une définition du rôle des acteurs.

3. DU 19E SIECLE A NOS JOURS : OUVERTURE DES DISCIPLINES SCIENTIFIQUES AUX LABORANTINS ET AUX STATISTICIENS

La fin de cette période coïncide avec le début de notre deuxième analyse. Dans cette deuxième section, nous allons montrer comment l'idée d'une mesure stochastique s'est propagée dans plusieurs disciplines scientifiques entre le 19 et le 20è siècle. Au travers de l'histoire de l'observatoire de Greenwich au 19è siècle, que nous utiliserons comme fil rouge, nous montrerons comment l'introduction de certaines disciplines à la stochastique a poussé les scientifiques à agrandir le nombre d'acteurs et à une deuxième vague de rationalisation du processus de production de connaissance.

3.1. D'UN MODELE DE LA SCIENCE DETERMINISTE A UNE SCIENCE STOCHASTIQUE

Dès le 19è siècle, un courant d'intellectuels et de scientifiques s'interroge sur la méthodologie communément partagée par la plupart des savants. En effet, une majorité de scientifiques adopte une approche déterministe, qui suppose que toutes les observations soient liées à une cause. Mathias Girel compare cette épistémologie à une « chaîne qui relie chaque maillon [...] et qui articule des individus, des évènements ou des atomes et qui symbolisent le déterminisme causal »[6]. Le travail du scientifique est de prévoir le phénomène en bout de chaîne causale en fonction uniquement des données initiales. Ce principe était cohérent avec plusieurs grandes découvertes scientifiques, comme les principes fondamentaux de la dynamique développés par Newton. Ses équations permettaient de connaître avec précision la position et la vitesse d'un corps à n'importe quel temps donné, uniquement grâce aux conditions initiales. L'idée de l'existence de lois mathématiques permettant de relier une cause à un effet était alors profondément ancrée chez la plupart des savants de l'époque. Les scientifiques considéraient que, ayant pris leurs mesures eux-mêmes, ils avaient confiance en leurs résultats. Ils étudiaient alors les choses jusqu'à ce qu'ils soient sûrs de la vérité.

Or si l'approche déterministe était performante dans certains cas, elle était beaucoup plus difficile à implémenter dans des situations où les phénomènes naturels étaient sujets à des variabilités incontrôlables. Trop simpliste, la représentation sous forme de chaîne causale a fait l'objet de plusieurs critiques. Dans des disciplines scientifiques comme l'astronomie, la physique ou la biologie, certains savants étaient face à des problèmes d'indétermination dans leurs mesures. Charles Boltzmann montra que le principe d'irréversibilité d'un système physique à l'équilibre macroscopique pouvait être représenté comme l'évolution indéterminée de particules identiques à l'échelle microscopique. Jacques Hadamard, mathématicien et physicien de la fin du 19è siècle, travaillait lui sur la stabilité du système solaire et des interactions entre les corps célestes. Il montra que dans certaines situations, une légère variation dans les conditions initiales des corps a un impact énorme dans leur trajectoire. Or, dans les problèmes astronomiques les conditions initiales ne sont jamais connues qu'avec une certaine erreur. « Si petite soit-elle, cette erreur pourrait amener une perturbation totale et absolue dans le résultat recherché » (Hadamard *in* Shaposhnikova Tatyana, 2005). Enfin, Charles Darwin fut sans doute celui qui à l'époque a le plus profondément transformé la vision de ses contemporains. Ses travaux sur l'évolution et la sélection naturelle repoussaient le paradigme religieux, profondément déterministe, et instillait dans l'évolution des espèces la notion novatrice de hasard.

Progressivement le « darwinisme » devient, comme l'a pu être Newton à son époque, un courant de pensée, un modèle à suivre pour repenser la science et la construction de la connaissance. Ce que certains savants vont déceler chez Darwin et les autres, c'est l'aspect contingent des objets d'observation de la science : au lieu de penser la science de façon entièrement déterministe, ils y

[6] http://savoirs.ens.fr/expose.php?id=2205

entrevoient une approche stochastique. Charles Sanders Pierce, un des défenseurs de cette approche darwinienne, soulignera la réussite de l'application de l'approche stochastique sur le comportement des gaz dans son livre *Essay on Classification* (Peirce *in* Becquemont, 2016) :

> « Mr Darwin s'est proposé d'appliquer la méthode statistique à la biologie. La même chose s'est faite dans une branche très différente de la science, la théorie des gaz. Bien qu'incapables de dire ce que serait les mouvements d'une molécule particulière de gaz d'après une hypothèse certaine concernant la constitution de cette classe de corps, Clausius et Maxwell réussirent cependant, par l'application de la doctrine des probabilités, à prédire qu'à la longue telle ou telle proportion de molécules, dans des circonstances données, acquerrait telle ou telle vélocité : que, chaque seconde, se produirait tel ou tel nombre de collisions, et, à partir de ces propositions, ils réussirent à déduire certaines propriétés des gaz, spécialement dans leur rapport de chaleur. »

L'approche stochastique intègre dans la mesure des phénomènes naturels la notion d'aléatoire, de hasard et propose en même temps une interprétation de cette incertitude par des lois mathématiques.

3.2. VERS UNE MESURE REPETEE ET STANDARDISEE : LE CAS DE L'EQUATION PERSONNELLE DES ASTRONOMES

Dès lors que la mesure est considérée comme sujette à du hasard, l'observation seule n'est plus suffisante pour faire des conclusions viables. L'expérimentation se doit au contraire d'être répétée, contrôlée et suffisamment standardisée pour confronter les différentes mesures entre elles. Les travaux menés à l'observatoire de Greenwich au 19è siècle sur la variabilité de la mesure astronomique illustrent bien comment les scientifiques ont intégré cet aspect dans leur activité quotidienne (Schaffer, 1988).

A l'époque, les astronomes avaient l'habitude de mesurer la position d'une planète ou d'une étoile au moment où celle-ci croisait le méridien. On considérait que la qualité d'une mesure dépendait principalement des compétences de la personne qui la réalisait. Ainsi, quand en 1796 l'astronome royal de l'observatoire de Greenwich Nevil Maskelyne se rendit compte que les mesures faites par son assistant différaient en général d'une demie à une seconde avec ses propres mesures, il le jugea trop imprécis et décida de le congédier. Plus tard, au cours des années 1810 et 1820, l'astronome allemand Friedrich Bessel reprit les travaux de Maskelyne et de son assistant. Il avait reçu comme objectif de réduire la taille des catalogues d'observation de Greenwich. Remarquant cette différence entre les mesures de Maskelyne et de son assistant, il s'intéressa à la cause de ce phénomène. Pour lui, les différences de mesure résultaient potentiellement d'un phénomène propre à l'observateur et non pas d'une « erreur » de mesure. Personne ne savait en effet à cette époque d'où provenaient ces différences individuelles. Profitant des fréquentes visites des astronomes à l'observatoire, il entreprit de mettre en place un programme systématique pour

étudier les différences relatives entre les observateurs astronomiques qui venaient au laboratoire. En se basant sur la mesure d'une étoile « artificielle » pour laquelle la vitesse de transition pouvait être standardisée, Bessel et son successeur George Airy enregistrèrent les vitesses d'observation des astronomes afin de calculer leur « équation personnelle », c'est-à-dire un facteur correctif moyen à appliquer systématiquement lors des mesures effectuées. Ainsi, la qualité de la mesure de chaque observateur pouvait être contrôlée grâce à son équation personnelle préalablement déterminée. Les variations de mesure entre chaque personne n'étaient plus vues comme une erreur due à la compétence des personnes, mais bien un phénomène inhérent à leur personnalité.

Au-delà de l'amélioration de la qualité des mesures effectuées, l'équation personnelle a apporté aux scientifiques et aux dirigeants de l'observatoire un moyen de standardiser le travail des observateurs comme pourrait l'être n'importe quel instrument de mesure. L'observateur devenait un composant de l'instrument de mesure qu'il convenait de calibrer. Cette transformation s'est accompagnée d'un processus de réorganisation sociale. L'observatoire fonctionnait de manière comparable à une usine : les observateurs étaient relégués à la base d'une hiérarchie où leurs mesures étaient inspectées par la direction. L'observation était mécanisée et les observateurs transformés en surveillants de machines. Le scientifique n'était plus celui qui observait car il était plus intéressant de lui attribuer des tâches qui correspondent à ses compétences. En fait, l'évaluation de l'équation personnelle permettait d'être plus confiant dans les mesures qui étaient réalisées par les non scientifiques. Le scientifique s'est mué en gestionnaire qui pilotait le travail de ses assistants. De la même manière que les scientifiques ont distingué le rôle du philosophe naturel de celui du fabricant d'instrument, la répartition des rôles entre scientifiques et assistants se précisait. L'observateur se contentait de la partie la plus mécanique de la tâche pour libérer le scientifique de toute erreur possible. Dans son livre sur la vie de William Rowan Hamilton, Graves explique comment le statut du scientifique a évolué dans l'observatoire de Greenwich :

> « It is not necessary for a man to observe, himself; he may render… most important services to Science by his calculations, and make his assistants observe for him. Schumacher [at Altona] observes very little himself, but is very accurate in superintending his assistants…. Bessel would be a first rate Professor of Astronomy, even though he never put his eye to a telescope. »[7]

[7] Graves, R. P. (1975). *Life of Sir William Rowan Hamilton* (Vol. 2). Arno Press.

3.3. Integration d'outils statistiques dans la pratique scientifique : la multiplication de la division du travail

L'équation personnelle des astronomes témoigne également d'un manque d'outils mathématiques à l'époque pour analyser ce phénomène. En effet, les astronomes de Greenwich se sont rapidement rendus compte de la difficulté à obtenir une équation personnelle suffisamment précise pour chaque observateur. Des mesures répétées avec un même observateur aboutissaient généralement à des résultats différents et ils n'avaient pas d'outils pour maîtriser cet aspect aléatoire des résultats. Partant de ce constat Peirce s'intéressa dans son livre *On the theory of Errors and Observations* publié en 1870 à cette question. Selon lui, l'équation en elle-même pouvait être sujet à une évolution de la part de l'observateur. Il souhaitait voir s'il était possible d'améliorer l'équation personnelle d'un observateur par l'entraînement. Il paya un jeune garçon inexpérimenté qui, pendant un mois, pressa plus de 500 fois par jour sur un bouton. Peirce analysa la différence temporelle entre le signal initial et le moment où le garçon appuyait effectivement sur le bouton. Au lieu d'étudier chaque pression individuellement, il présenta les résultats quotidiennement sous la forme d'une gaussienne. Grâce à cette forme, il pu mettre en évidence que le jeune garçon avait amélioré son équation personnelle entre le début et la fin de l'expérience, passant de 1/7è de retard à 1/80è en un mois.

Ce n'est pas tant le résultat de Peirce qui est intéressant ici, mais aussi la méthode qu'il emploie. L'aspect stochastique est présent dans son expérience à plusieurs endroits ; d'abord Peirce est conscient que le jeune va améliorer sa technique à chaque itération, et qu'on ne peut considérer son équation personnelle comme la simple moyenne sur l'ensemble de ces essais. Il prend en compte un nouvel aspect dans l'analyse, à savoir l'évolution de l'objet étudié durant l'analyse de celui-ci. Ensuite, il étudie les essais non pas de façon isolée mais s'intéresse à leur distribution quotidienne et à la variation de l'écart-type de la distribution. Il s'approprie ainsi un certain nombre de concepts développés durant ce siècle par quelques mathématiciens. Il reprend la forme de distribution développée par Carl Friedrich Gauss quelques décennies plus tôt, ainsi que des outils développés par son contemporain Francis Galton comme l'écart-type, la corrélation ainsi que le principe de régression.

En effet, avant cette période il y avait parfois un refus catégorique des scientifiques de combiner les données entre elles (Stigler, 1986). Les astronomes calculaient des moyennes simples de mesures presque parfaitement répliquées, mais l'idée de pouvoir augmenter la précision en combinant des mesures effectuées dans différentes conditions tardait à venir. Ils craignaient que des erreurs commises dans une observation en contaminent d'autres, qu'elles se multiplient au lieu de se compenser. Pour qu'une inférence statistique fondée sur les probabilités soit combinée, il était indispensable que les mesures soient effectuées dans des conditions considérées comme identiques ou ne différant que de la manière qui pourrait être prise en compte dans l'analyse.

Par la suite, ce processus d'appropriation des outils mathématiques pour étudier les phénomènes stochastiques se généralisa (Stigler, 1986). Entre le 19è et le 20è siècle, les statistiques sont au cœur d'un développement à la fois vertical et horizontal dans la méthode scientifique. Vertical dans la compréhension du rôle des probabilités comme une analogie des jeux de chance et des modèles de probabilités pour les mesures, menant à l'introduction des probabilités inverses et au début de l'inférence statistique. Horizontal car plusieurs disciplines adoptent ces outils statistiques : la médecine, la psychologie, l'astronomie, la biologie, la thermodynamique ainsi que la physique des particules pour n'en citer que quelques-uns (Berger, 1999; Matthews, 2016; Schaffer, 1988). La théorie des probabilités et les statistiques initiales sont systématisées au 19ème siècle et les chercheurs en sciences sociales utilisent un raisonnement statistique et des modèles de probabilités pour faire progresser les nouvelles sciences de la psychologie et de la sociologie expérimentales, ainsi que des physiciens en thermodynamique et en mécanique statistique. Le développement du raisonnement statistique devient étroitement associé au développement de la logique inductive et de la méthode scientifique, préoccupations qui éloignent les statisticiens du domaine plus étroit de la statistique mathématique.

La statistique va offrir une source d'outils à une époque où de nombreux domaines scientifiques faisaient l'objet d'une quantification croissante et où les problèmes de collecte, d'analyse et d'interprétation des données atteignaient un seuil critique de complexité. La maîtrise de ces outils étant parfois complexe, les scientifiques travaillent parfois avec les statisticiens pour développer ensemble des modèles adaptés à leurs objets d'observation.[8]

LES PROFESSIONNELS DE SANTE DANS LA MEDECINE

Les besoins de scientificité de la médecine apparaissent dès la fin du 19è siècle avec : « la conscience de la nécessité d'un langage qui passerait de l'à peu près qualitatif à l'univers de la précision chiffrée ; Exemple du précédent de la chimie qui, avec Lavoisier et autour de lui, acquiert un statut de science plus rigoureux qu'antérieurement, des chimistes préoccupés de physiologie (Lavoisier) ou des chimistes-docteurs en médecine (Chaptal) servent de transition ; Exemple simultané de l'application des statistiques aux réalités économiques et sociales, et notamment aux réalités de l'hygiène publique, considérée alors comme une discipline médicale » (Comiti, 1976). Des scientifiques comme Chaptal, Bernoulli, Villermé développent les méthodes statistiques pour étudier les populations de malades, et démocratisent la quantification de la mesure par opposition à une approche individuelle de la médecine.

Au milieu du 19è siècle, des tests de laboratoire sont introduits pour détecter la tuberculose, le choléra, la typhoïde et la diphtérie (Berger, 1999). Les médecins commencent à étudier le pouls, la pression artérielle, la température corporelle et d'autres indicateurs physiologiques et l'utilisation de mesures précises devient la norme. Les tests oculaires standardisés, les tables de poids et de taille et les tests de QI faisaient partie d'un mouvement visant à identifier les normes de la physiologie et du comportement humains. La multiplication des tests, les besoins grandissant en instruments de mesure, ainsi que l'émergence d'entreprise pharmaceutique complexifient les liens entre les travaux des médecins et une certaine emprise des entreprises privées. Les médecins cherchent alors à conserver à tout prix leur autonomie et empêcher des tiers de réaliser des bénéfices. Soutenus par les gouvernements, ils ont mis en place une infrastructure médicale leur permettant de déléguer à d'autres professionnels de la santé un travail répétitif et chronophage. Pour conserver leur autonomie, les médecins avaient besoin d'assistants techniques dans les hôpitaux et les laboratoires sans être employés par ces établissements. Les professionnels paramédicaux ont commencé à émerger au cours des années 1900-1930. Ces assistants techniques devaient être suffisamment compétents pour travailler en l'absence des médecins, sans toutefois menacer leur autorité. Cette politique

[8] Nous pouvons citer la collaboration entre le biologiste Pierre Charles Alexandre Louis avec le mathématicien Karl Pearson qui à eux deux développèrent la biostatistique.

s'est notamment développée : (1) en encourageant une sorte de professionnalisme responsable parmi les cadres supérieurs des personnels de santé subalternes, et (2) en employant des femmes dans ces rôles auxiliaires qui ne contesteraient pas l'autorité du médecin.

3.4. UNE NOUVELLE ORGANISATION AUTOUR DE L'EXPERIMENTATION STOCHASTIQUE

L'approche stochastique des mesures de phénomènes naturels s'est progressivement intégrée dans un ensemble de disciplines scientifiques qui ne se suffisaient pas d'une approche déterministe. Conséquemment, la mesure des phénomènes devient un processus qui doit être répété un nombre suffisant de fois afin d'en tirer des conclusions scientifiques. En effet, les outils mathématiques qui permettent de dégager des lois à des phénomènes hasardeux comme les courbes de Gauss, les mesures d'écart-type ou la méthode des moindres carrés nécessitent de travailler avec des échantillons de mesure suffisamment grands pour convenir aux conventions[9]. L'activité autour de la collecte de données se multiplie et demande des ressources beaucoup plus importantes pour effectuer les récoltes des bases de données. Ces activités étant trop chronophages pour les équipes de scientifiques, l'organisation scientifique voit apparaître de nouveaux acteurs spécifiquement dédiés à l'expérimentation et à la collecte des données : les professionnels de santé dans la médecine, les laborantins dans les laboratoires de biologie ou de chimie, les observateurs dans le laboratoire de Greenwich. Alors que les assistants et les techniciens étaient auparavant des « hommes à tout faire » pour les scientifiques qui parfois avaient un rôle essentiel dans la conception et la réussite des expériences scientifiques (Shapin, 1989), ici l'activité est clairement divisée entre scientifiques et non scientifiques. Par ailleurs, les scientifiques favorisent la collaboration avec les statisticiens : tandis que ces derniers peuvent fournir des modèles statistiques appropriés aux scientifiques, les statisticiens profitent de l'existence des problématiques scientifiques pour affiner la théorie statistique.

[9] Fischer, un illustre statisticien, fournira aux scientifiques les outils parmi les plus utilisés en science au début du 20è siècle : le test de l'hypothèse nulle ainsi que la notion de p-valeur. La p-valeur correspond à la probabilité d'obtenir avec un modèle, une différence au moins égale à celle observée; et si cette p-valeur est inférieure à une limite de référence choisie dans le respect de certaines conventions arbitraires, alors on considère que la différence observée est significative. Cette standardisation des valeurs seuil dans l'inférence statistique pousse les scientifiques à collecter des échantillons suffisamment grands pour obtenir des résultats statistiques satisfaisants (Peto et al., 1976)

4. OUVERTURE DU PROCESSUS SCIENTIFIQUE DANS LE CADRE DE LA SCIENCE DATA-DRIVEN

4.1. SIMILARITES ENTRE LES EXEMPLES HISTORIQUES ET LES SCIENCES CITOYENNES

Ces deux périodes historiques que nous avons présentées peuvent être interprétées suivant un même schéma causal qui pousse à la réorganisation du processus de production de connaissance (**figure 7**), et que nous retrouvons également dans notre situation contemporaine avec l'avalanche de données.

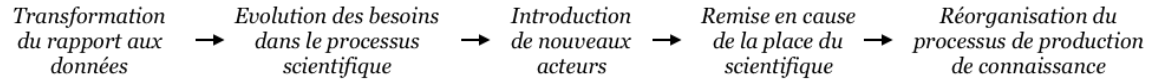


Figure 7. Modèle d'intégration de la transformation du rapport aux données par une réorganisation du processus de production de connaissance.

Premièrement, les trois périodes analysées sont basées sur une transformation du rapport des scientifiques à l'observation des phénomènes naturels. Avec l'utilisation généralisée de l'instrument scientifique, le scientifique a accès à de nouveaux observables : il peut voir l'infiniment grand comme l'infiniment petit, mais également mesurer des phénomènes qui lui étaient jusqu'à présent inaccessibles – la pression de l'air, la température, les ondes électromagnétiques pour n'en citer que quelques-uns. Le domaine d'observation de la nature s'en trouve changé, et intègre nouvelles dimensions possibles de la mesure. Durant la deuxième période historique, c'est l'interprétation des données qui est transformée. Alors que les modèles d'analyse étaient généralement basés sur une approche déterministe, un ensemble de scientifiques proposent un changement de paradigme pour étudier les phénomènes et incluent la notion d'aléatoire dans la mesure. Enfin, dans le modèle contemporain de la science, les épistémologues suspectent que l'avalanche de données accessibles est susceptible de faire émerger une science dite data-driven, où les hypothèses du processus scientifique se construisent directement à partir des bases de données disponibles.

Ensuite, de nouveaux besoins émergent dans le processus de production de connaissance qui poussent les scientifiques à intégrer de nouveaux acteurs. A chaque période, la transformation du rapport aux données fait apparaître de nouvelles activités (conception et fabrication d'instruments scientifiques, génération d'hypothèses basées sur les données) ou fait évoluer les activités existantes (répétition des expérimentations, analyse de bases de données massives). Ces transformations nécessitent à la fois de nouvelles ressources dans le processus de production de connaissance ainsi que des compétences non nécessairement présentes chez les scientifiques. Par exemple, la fabrication d'instruments scientifiques demande une connaissance artisanale sur des matériaux complexes à travailler comme des lentilles pour les télescopes et les microscopes. De la

même manière, l'analyse des données massives dans le paradigme data-driven est souvent liée à des méthodes statistiques nouvelles issues de l'intelligence artificielle que les scientifiques ne maîtrisent pas forcément. Ces nouveaux besoins poussent les scientifiques à collaborer avec de nouveaux acteurs au sein du processus scientifique pour supporter une partie de ces tâches. Dans les exemples historiques, nous avons vu que cela avait mené à l'intégration de fabricants d'instruments, de statisticiens mais également de laborantins, de techniciens ou bien de professionnels de la santé.

Dans certains cas, l'intégration de ces nouveaux acteurs perturbe le rôle du scientifique dans le processus de production de connaissance. Par exemple, la collaboration entre artisans et scientifiques a mené à une redéfinition du rôle du savant pour mieux se distinguer des fabricants d'instruments. L'apparition d'une science data-driven est susceptible de créer un conflit de même nature si la génération d'hypothèses est déléguée. En effet, la génération des hypothèses est une des activités fondamentales du scientifique. Le scientifique définit les hypothèses importantes et les théories dans son domaine de spécialité. C'est encore lui qui conçoit les outils et les instruments qui lui seront utiles pour ces expérimentations. Le système de récompense développé récompense ensuite sa capacité à proposer des hypothèses originales et la qualité de sa production scientifique lui servira à obtenir plus de financements et de ressources pour mener à bien ces recherches. Le rôle du scientifique s'est défini au cours de l'histoire sur cette loi fondamentale forte et implicite qui distingue les tâches de l'intellect des autres tâches (Shapin, 1989).

Enfin, dans les deux épisodes historiques l'apparition de nouveaux acteurs dans le processus scientifique a poussé à une réorganisation et à mettre en place un modèle de gestion adapté. Dès le 18è siècle, les scientifiques rationalisent la production scientifique et standardisent les outils et les instruments utilisés. Ils étudient les pratiques des fabricants d'instrument afin d'en extraire une méthode et donc de faciliter la répartition du travail et la délégation de tâches aux non scientifiques (David, 2007; Shapin, 1989). Le scientifique est en charge de la conception de l'instrument tandis que l'artisan le fabrique. Ce modèle de division du travail est également appliqué au sein des lieux dédiés à la production scientifique. Lorsque George Airy prend ses fonctions pour diriger l'observatoire royal de Greenwich, il met en place une hiérarchie stricte et une division du travail dans lesquelles des calculs de routine et de répétition sont effectués par des cadres d'employés avant d'être transmis pour vérification. Cette bureaucratisation de la science, initialement instaurée dans les universités allemandes, se développera ensuite largement dans les pays occidentaux, notamment en France, en Angleterre, en Italie et aux Etats-Unis (Whitley, 2000).

4.1.1. Analogie avec le contexte contemporain « data-driven »

Qu'est-ce que l'étude historique nous permet d'apprendre sur l'apparition des sciences citoyennes dans le contexte data-driven ? Les projets de science citoyenne semblent apporter une réponse organisationnelle adaptée à l'avalanche de données dans notre contexte contemporain avec un

modèle organisationnel qui diffère peu de ce qu'on retrouve dans les épisodes historiques. En effet, la science citoyenne propose une collaboration sous forme de division du travail entre scientifiques et non-scientifique basée sur un modèle seeker-solver : le seeker conçoit et organise la tâche tandis que le solver l'exécute. En revanche, le mode de collaboration entre les non-scientifiques reste pour l'heure inconnu, tout comme le modèle de performance.

Ensuite, de la même manière qu'elle a été remise en question dans l'histoire, la place du scientifique et la réorganisation du rôle entre les acteurs du processus de production de connaissance semble être réinterrogée. Les analyses historiques nous montrent que ce n'est pas la première fois que les scientifiques sont confrontés à cette question. Généralement, le processus scientifique a toujours distingué ceux qui conçoivent l'activité (ce sont les tâches les moins répétitives) des personnes qui exécutent pour utiliser de la manière la plus efficace les connaissances et les compétences de chacun. Ainsi les scientifiques se sont-ils toujours attribués les tâches qui relèvent de la conception plutôt que celle relatives à l'exécution. Pourtant, dans le cas où la génération des hypothèses n'est pas réalisée par le scientifique mais par une foule de citoyens, cette hypothèse est remise en cause. Quelle serait la place du scientifique dans cette autre forme de répartition du travail ?

	Introduction des instruments scientifiques	**Stochastique dans les phénomènes naturels**	**Science data-driven**
Transformation du rapport aux données	Extension du domaine d'observation des phénomènes naturels	Interprétation stochastique des observations	Les données comme base pour la génération des hypothèses
Evolution des besoins dans le processus scientifique	Conception et fabrication des instruments	Répétition des observations et analyse	Collecte et analyse de bases de données massives, générer des hypothèses basées sur les données
Introduction d'acteurs	Fabricants d'instruments, artisans	Observateurs, laborantins, professionnels de santé, Statisticiens	Citoyens de la science (public**), nouveaux acteurs pour gérer la foule ?**
Remise en cause de la place du scientifique	Place face à l'artisan	Le scientifique qui ne fait pas ses expérimentations	**Quelle place pour le scientifique ?**
Réorganisation du processus de production de connaissance	Division du travail	Organisation bureaucratique	Modèle seeker-solver

Tableau 6. Similitudes entre les épisodes historiques et l'approche data-driven.

4.2. CONSTRUCTION D'UN MODELE TACHE DELEGUEE AUX CITOYENS DE LA SCIENCE

Au travers de notre analyse historique succincte sur l'évolution du rapport aux données dans la science, nous avons rencontré un certain nombre de cas et d'exemples historiques où certaines activités du processus de production de connaissance ont été déléguées à des non-scientifiques. Ces exemples sont précieux pour modéliser les différentes activités qui peuvent être ouvertes par le biais des sciences citoyennes. En effet, dans la suite de notre thèse, nous ne cherchons pas à proposer un modèle général de la découverte scientifique. Par ailleurs les recherches qui ont été menées sur le sujet ont conclu que le modèle dépendait largement du contexte dans lequel il était appliqué et qu'il n'était pas possible de construire un modèle universel (Klahr & Simon, 1999). Au contraire, nous cherchons à construire un modèle des activités déléguées dans le processus de production scientifique. A partir des exemples historiques et d'exemples contemporains de projets de science citoyenne, nous proposons un modèle de trois tâches : la tâche élémentaire, la recette et la résolution de problèmes. Dans le chapitre suivant, nous étudierons de manière formelle ces modèles de tâche et nous chercherons à établir si ils sont suffisants pour prendre en compte les tâches déléguées dans les projets de science citoyenne, notamment la génération d'hypothèses basées sur les données. Ce modèle formel nous permettra d'établir les principes généraux de performance inhérents à chacune des tâches et de voir comment cette performance peut être gérée.

4.2.1. La tâche élémentaire

La tâche élémentaire est définie comme l'application d'une unique action pour arriver à l'objectif désiré. Cette action est entièrement connue et souvent décrite au travers d'un protocole standardisé. L'objectif est de maîtriser au maximum l'action produite afin de contrôler sa productivité : on connaît l'objectif à atteindre ainsi que l'action qui va mener à cet objectif. La principale difficulté dans la tâche élémentaire réside dans la capacité à obtenir des résultats similaires lorsque celle-ci est répétée plusieurs fois. En effet, plusieurs éléments peuvent varier entre les tâches : l'objet initial sur lequel est appliqué la tâche n'est jamais exactement le même, et l'action à mener peut également être exécutée de différentes manières.

LA COLLECTE DE DONNEES DANS LE PROCESSUS SCIENTIFIQUE : LE COMPTAGE DES SCINTILLEMENTS DANS LES EXPERIENCES DE RUTHERFORD

Les premières expériences de radioactivité dépendaient d'observations difficiles et répétitives effectuées sur des cadres d'observateurs formés à une tâche spécifique. Lorsque Rutherford entreprit ses recherches sur la radioactivité parallèlement à ses activités militaires en 1917, il chercha à explorer les effets des particules α énergétiques provenant de substances radioactives naturelles sur divers atomes, dans le but d'en savoir plus sur les forces autour des noyaux et leurs autres caractéristiques. Les expériences consistaient à compter les scintillations produites par des particules énergétiques lorsqu'elles frappaient un écran de sulfure de zinc préparé avec soin. Or, la vue et le tempérament de Rutherford ne convenaient pas parfaitement au métier de comptage de scintillation, et Rutherford se tourna vers son assistant William Kay pour cette tâche. En effet Kay avait montré à plusieurs reprises sa grande capacité et sa grande précision pour ce type de tâches. De son aveu, la tâche était lourde en raison de la difficulté d'obtenir des résultats précis, cohérents, reproductibles et fiables. Ainsi, en 1917 et 1918, Kay passa des heures assis dans une pièce sombre, muni

d'un puissant microscope, comptant les minuscules éclairs de lumière produits par les expériences de Rutherford. C'est ce travail qui a conduit Rutherford à la conclusion que les collisions avec des particules α énergétiques pourraient éjecter des noyaux d'hydrogène - des protons, comme il les baptiserait plus tard - des noyaux d'azote. Cela suggérait les idées de constitution nucléaire et de structure nucléaire: les origines conceptuelles de la physique nucléaire moderne (Hughes, 2008). Plus tard, Johannes Wilhelm Geiger travaillera sur la conception de méthodes électriques afin de compter ces pulsations.

4.2.2. La recette

La recette peut être vue comme une succession de tâches élémentaires. La recette peut être assimilée à une séquence d'actions à réaliser dans un ordre précis. Les remarques que nous avons sur les de tâches élémentaires s'appliquent donc au cas de la recette. Il faut également inclure que la séquence d'actions doit être réalisée dans le bon ordre. Le cas du projet Galaxy Zoo présenté en annexe 1 est un exemple typique de délégation à la foule pour ce type de tâche. Pour rappel, la plateforme Galaxy Zoo est un projet ayant regroupé plus de 250 000 volontaires qui ont aidé au codage d'images astronomiques issues de télescopes, et qui ont contribué à la découverte de nouvelles classes de galaxie. La participation est devenue rapidement virale, et sept mois après le lancement du projet environ 900 000 galaxies furent codées. A titre de comparaison, 50 millions de classifications auraient requis plus de 83 années à plein temps pour un scientifique seul. Afin de réduire la probabilité d'un codage incorrect et donc de mauvaise qualité, les galaxies furent codées plusieurs fois par différents volontaires, pour un total d'environ 50 millions de classifications.

LES TABLES DE CALCUL DE PRONY – UN EXEMPLE DE RECETTE

Au début du 19è siècle, Prony s'était engagé à concevoir des tables logarithmiques et trigonométriques qui constituaient à l'époque « le monument de calcul le plus vaste qui eut jamais été exécuté, ou même conçu » (Laboulaye & Babbage, 2016). Les logarithmes devaient être calculés pour les nombres allant de 1 à 200 000, et Prony se rendit vite compte que même associé à une équipe de mathématicien, il lui aurait fallu plus d'une vie pour accomplir la tâche. En se basant sur le modèle de division sociale du travail développé par son contemporain Adam Smith, il eut l'idée de mettre en place une organisation du travail à trois niveaux. Le premier niveau, constitué de cinq ou six mathématiciens était chargé de trouver la meilleure méthode pour effectuer le calcul numérique tandis que le deuxième niveau devait convertir les formules en opérations numériques simples et faciles à contrôler. Enfin, le troisième niveau, était constituée de 80 personnes qui calculait les opérations conçues par le deuxième niveau. Si les membres du premier et du deuxième niveau étaient des mathématiciens de métier, les neuf dixièmes de membres du troisième niveau ne connaissaient que les règles élémentaires des mathématiques : l'addition et la soustraction.

4.2.3. La résolution de problèmes

La résolution de problème est une tâche où l'objectif est clairement désigné, mais la solution n'est pas connue (Simon & Newell, 1971). Par exemple, la construction d'un nouvel instrument scientifique fait appel à un ensemble de connaissances tacites qui ne sont pas partagées par les scientifiques. Lorsque ces derniers conçoivent l'instrument, une partie de la méthode pour arriver à la construction finale reste encore à déterminer par les artisans. C'est le cas par exemple dans la

collaboration entre Huygens et le spécialiste de la fabrication de lentilles, ou encore entre Charles Babbage et Joseph Clement.

OPTIMISATION DE L'ANALYSE DE DONNEES DANS LE PROCESSUS SCIENTIFIQUE : DATA CHALLENGE POUR LA DETECTION DU BOSON DE HIGGS

La validation d'un résultat expérimental dans le modèle de la physique des particules correspond principalement à deux étapes majeures : la collision de particules à très forte énergie dans un accélérateur à particules qui génère de nouvelles particules durant des temps infimes, puis l'analyse des signaux détectés par les différents détecteurs de particules installés aux alentours de la collision. Les signaux détectés représentent un vecteur d'environ cent mille dimensions à partir duquel l'énergie et la direction dans l'espace de chaque particule sont estimées. Une fois l'ensemble des évènements reconstitués, des algorithmes de sélection sont utilisés pour filtrer les évènements considérés comme inintéressants puis conserver les autres évènements via un système de stockage. Une analyse hors ligne de ces évènements a ensuite pour objectif de trouver dans l'espace une région dans laquelle il existe un excès significatif d'évènements par rapport à ce que les processus connus en arrière-plan peuvent expliquer. Une fois la région fixée, un test statistique de comptage est appliqué pour déterminer si l'observation est significative.

Dans ce système, le choix de la région est déterminant. Si l'objectif associé est facilement quantifiable et correspond à trouver une région de signal indépendante du signal de fond, la découverte d'un modèle algorithmique qui réponde parfaitement à l'objectif est très difficile. Les physiciens ont récemment mis en place une campagne d'optimisation pour la détection de ces régions à partir d'un data challenge en 2014 (tournoi basé sur l'analyse de données) appelé Higgs Boson ML (Adam-Bourdarios, Cowan, Germain, & Guyon, 2015). Le challenge se basait sur les bases de données récoltées en 2012 et utilisées pour analyser l'existence du boson de Higgs. Le challenge a été implémenté sur la plateforme en ligne Kaggle et était ouvert à toute personne intéressée. Les participants n'étaient pas nécessairement spécialistes de la physique des particules mais étaient capables, à partir d'une base de données ainsi que d'un problème bien formulé, d'explorer l'espace des solutions possibles et de proposer la solution qui leur semblait la meilleure. Plusieurs participants ont proposé des solutions plus performantes que les méthodes traditionnellement utilisées et le meilleur modèle a permis une amélioration de l'objectif de 19%, passant de 3.2 sigma à 3.8 sigma.

CHAPITRE 5 – MODELE FORMEL DES ACTIVITES DELEGUEES DU PROCESSUS DE DECOUVERTE SCIENTIFIQUE : NOTION DE « TACHE COUPLEE INVENTIVE » ET CRITERES DE PERFORMANCE

RESUME DU CHAPITRE 5

Dans ce chapitre, nous réinterprétons les tâches – élémentaire, recette, résolution de problèmes - présentées dans le chapitre historique (chapitre 4) au travers du modèle général de la résolution de problèmes (Simon, 1966). Ce modèle est basé sur la notion d'états ainsi que d'actions qui permettent de passer d'un état à un autre. D'un point de vue de la performance, ce modèle permet d'illustrer des problématiques de coût liées à la réalisation d'une tâche, notamment les coûts d'exploration et d'exécution des actions. Bien que ce modèle général ait été critiqué sur plusieurs aspects, il offre un formalisme simple pour interpréter l'activité humaine et adapté pour notre utilisation.

Nous montrons en revanche que la génération d'hypothèses ne rentre pas dans le modèle. En effet, la génération d'hypothèses est implicitement associée à l'exploration de deux espaces extensibles : un espace pour formuler l'hypothèse et un espace pour réaliser des actions physiques (analyser les données, faire des expérimentations). Le modèle classique de résolution de problèmes n'est donc pas suffisant pour modéliser la génération des hypothèses car il se contente d'explorer un seul espace. Par ailleurs, les travaux menés par Klahr et Danbar (1988) pour construire un modèle de la logique de découverte scientifique basé sur un double espace ne prennent pas en compte l'expansion possible des espaces.

Nous proposons alors de baser la génération des hypothèses sur un formalisme à deux espaces extensibles, celui du constructivisme imaginatif (Kazakçi, 2014) et issu des travaux plus généraux sur les liens entre science et conception (Hatchuel et al., 2013). Ce modèle se base sur le principe que tout objet n'existe que s'il existe une méthode capable de le construire. Le constructivisme imaginatif suggère que l'on construise en même temps le quoi (la définition de l'objet) et le comment (une méthode pour construire cet objet). Contrairement au modèle de résolution de problèmes, l'objet n'est pas défini conceptuellement avant de réaliser le processus : il se construit en même temps qu'on construit la méthode. De ce modèle, nous élaborons la notion de tâche couplée inventive, qui consiste à formuler un problème (ou état désiré) en même temps que la construction de son plan d'action.

Enfin, nous nous intéressons à la question de la performance dans le cas où les tâches sont répétées. Si cette répétition est classiquement intégrée pour des tâches élémentaires ou recettes, elle est beaucoup moins étudiée dans le cas de tâches de type résolution de problème et tâche couplée inventive. Nous suggérons que cette question de performance est fortement reliée à la notion de capitalisation : capitalisation entre les tâches répétées et capitalisation durant la tâche. Nous verrons dans le chapitre suivant à quel point cette notion de capitalisation dans les tâches répétées est importante dans la performance des tâches ouvertes à la foule.

Nous avons caractérisé dans les chapitres précédents la nature du processus scientifique basé sur les données. Nous avons montré que plusieurs activités liées au processus pouvaient se rapporter à des tâches : tâche élémentaire, recette, résolution de problèmes. Dans ce chapitre, nous poursuivons cet effort de modélisation en reprenant de façon plus formelle les tâches que nous avons définies. Nous montrons également que ce modèle est insuffisant pour étudier la performance dans la génération des hypothèses. En effet, il se base généralement sur l'exploration d'un seul espace, alors que la génération des hypothèses nécessite de penser à la fois à la formulation de l'hypothèse en même qu'au processus pour analyser les données qui mènent à la construction de cette hypothèse. Nous proposons d'étendre le modèle en introduisant la notion de **« tâche couplée inventive ».**

1. PRESENTATION DU MODELE GENERAL

1.1. REPRESENTATION PAR DES ETATS ET DES ACTIONS

Nous allons supposer un ensemble $E = \{e_1, e_2, ...\}$ qui représente l'ensemble des états possibles du monde. Dans ce modèle, il est toujours possible de transiter de n'importe quel état e_i du monde vers un état e_j par l'application d'une action $a_{i,j}$. Pour simplifier l'explication, nous utiliserons un indice unique pour énumérer les actions. L'ensemble des actions sera alors noté $A = \{a_1, a_2, ...\}$. Il est à noter qu'une action peut être applicable ou non dans un état donné. La structure $< E, A >$ possède une structure de graphe orienté. Dans ce graphe, un état e_i peut être relié à un état e_j (mais pas nécessairement) suivant l'application d'une action. Réciproquement, l'état e_j peut être relié au même état e_i par une autre action.

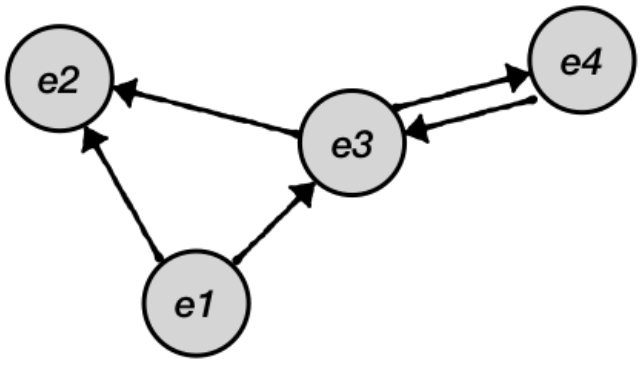


Figure 8. Illustration de la structure du graphe < E, A>

L'exécution d'une tâche est représentée comme la transition d'un état initial $e_{initial}$ vers un nouvel état e_{final} désiré. Tout problème peut être modélisé par un triplet $(e_{initial}, e_{final}, test(e_j, e_{final}))$ tel que :

$e_{initial}$ représente l'état du monde initial tel qu'il peut être observé ;
e_{final} représente l'état du monde espéré que l'on cherche à atteindre ;

test(e_j, e_{final}) est une fonction qui permet de vérifier si tout état e_j est égal à l'état e_{final} recherché.

Le but de celui qui résout le problème va être de trouver la séquence d'action a_i , a_{i+1}, …, a_j tel que chaque a_k de la séquence soit applicable à l'état du monde qui résulte de l'application de l'action précédente a_{k-1}. En effet, nous avons vu que certains états ne sont pas reliés directement les uns aux autres par une action et la transition nécessite de passer par des états intermédiaires. Dans notre illustration (**figure 8**), le passage de l'état *e4* à l'état *e2* nécessite de transiter par l'état intermédiaire *e3*. L'exécution d'une tâche comme une séquence d'actions peut être vue différemment suivant le contexte d'application : cela peut être une suite d'instructions informatique, un algorithme, une fonction, une méthode ou bien un plan d'action.

ILLUSTRATION AVEC LE LABYRINTHE

Prenons le cas classique d'un labyrinthe dont l'objectif est de trouver la sortie. Les états ici sont représentés par la position discrète sur le tableau de la **figure 9**. Par exemple le point de départ, symbolisé par le point gris, est à la position (3 ;2) tandis que la sortie en orange est au point (8 ;8). Il existe quatre actions élémentaires qui peuvent être appliquées pour changer l'état du monde : monter (ce qui correspond au mouvement monter (0 ;1) ; descendre (0 ;-1) ; aller à gauche (-1 ;0) ; aller à droite (1 ;0)). A chaque action, l'exécutant teste sa position par rapport à la sortie grâce à la fonction test. Si les deux sont différentes, alors l'exécutant appliquera une nouvelle séquence d'action jusqu'à arriver à la position finale. Nous avons représenté sur la figure une séquence d'actions qui permet de passer de l'état initial (la position (3 ;2)) à la position finale représenté par la porte de sortie. Cette séquence n'est pas la seule possible entre l'état initial et l'état final.

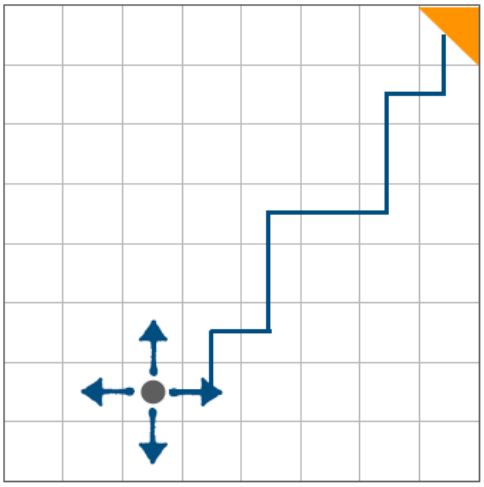


Figure 9. Exemple du labyrinthe

1.2. LE COUT DE L'EXPLORATION VS LE COUT D'UNE SOLUTION

Dans notre exemple simple du labyrinthe, l'espace de tous les états possibles est représentable entièrement. Or, ce n'est pas toujours le cas dans un processus de résolution de problèmes, et l'exécutant doit souvent explorer plusieurs combinaisons d'actions avant d'en trouver une qui mène à l'état désiré. En fait, plus la taille de l'espace d'actions augmente, plus il y a de combinaisons possibles à générer et donc à tester. En terme pratique, cela engendre des coûts liés à l'exploration. Par exemple, s'il s'agit d'une résolution algorithmique, le nombre d'opérations possibles à séquencer peut rapidement devenir très grand et donc augmenter d'autant le nombre de combinaisons d'opérations possibles. En parallèle de ces coûts liés à l'exploration, il existe également le coût de la mise en œuvre d'une solution (ou séquence d'actions). En effet, les actions à appliquer peuvent avoir des coûts associés. Cela veut dire que non seulement l'exécutant doit trouver une solution, mais il doit également faire en sorte que le coût de cette solution soit le minimum possible ou qu'il reste en dessous d'un certain budget. Cette fonction de coût peut simplement être le temps passé à réaliser une séquence d'action, ou bien être multicritères : par exemple une combinaison entre le coût monétaire et le temps passé. Remarquons que lorsque la fonction de coût est connue, elle est utilisée pour guider l'exploration : souvent, le processus d'exploration va privilégier les directions qui vont baisser le coût global.

LE COUT DANS LE CALCUL DE CORRELATION

Pour illustrer ces notions de coûts, prenons l'exemple du calcul de l'écart type de la variable X par un ordinateur. Un estimateur du coefficient de corrélation est donné par :

$$\widehat{\sigma_X} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \quad \text{avec} \quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \ ;$$

Pour effectuer le calcul numérique, nous proposons deux séquences d'actions différentes. La première consiste à calculer chacun des termes de l'équation indépendamment les uns des autres, c'est-à-dire le carré de la différence pour chaque terme entre lui-même et la moyenne $\bar{x}$. Pour chaque terme, l'ordinateur devra calculer à chaque itération cette moyenne et la supprimer de sa mémoire vive. Ainsi la moyenne sera calculée N fois, N étant le nombre de termes pour la variable X. La deuxième méthode consiste à calculer d'abord la moyenne, la conserver dans la mémoire vive de l'ordinateur, puis calculer la différence entre chaque terme avec la moyenne sans avoir besoin de recalculer la moyenne à chaque fois. La différence majeure entre les deux méthodes peut se quantifier en terme de temps de calcul sur l'ordinateur : il y a N-1 actions supplémentaires dans la première méthode, sans que cela modifie le résultat final obtenu. En fait, l'ordinateur calculera à chaque fois la moyenne des termes, au lieu de ne la calculer qu'une seule et unique fois.

1.3. REMARQUES SUR LE MODELE

1.3.1. Représentation des états

La première remarque de fond à propos de ce formalisme concerne simplement la notion « d'état ». En fait, dire qu'on peut avoir des descriptions univoques et claires d'un « état » de la nature du monde est une hypothèse fondamentale. Or, elle laisse en suspens un certain nombre de questions : qu'est-ce qu'un état? Peut-on le représenter ou peut-il même être connu? Cela ne dépendrait-il pas des capacités d'une entité cognitive? Ces questions sont difficiles à trancher sur le plan philosophique, épistémologique et cognitif et font débat depuis l'introduction de ce modèle, dont les limites ont été discutées sous plus d'un angle (e.g. Dummett, 2000; Harman, 2002; Shapiro, 1996; Simon, 1973). En pratique, un état est souvent présenté comme un ensemble de relations $R_1, R_2, \ldots \in R$ qui sont vraies dans cet état. L'ensemble R des relations entre les objets suppose l'existence d'un langage qui permet de décrire le monde, par exemple en utilisant la logique de premier ordre. Ces relations décrivent les états et les relations qui s'appliquent à un domaine d'objet D. Une relation peut concerner un objet seul ou un groupe d'objets. Par exemple l'objet « pomme » est associé au prédicat « rouge » qui lui attribue une couleur, tandis que cette même pomme est également reliée à une « table » par la relation « sur » qui permet de dire que « la pomme est sur la table ». Par la suite, nous supposerons l'existence d'un langage ou d'une ontologie capable de décrire les différents objets que l'on observe suivant l'état décrit.

1.3.2. Complexité et capacité de calcul

Sur le plan informatique, les limites du modèle sont d'un autre ordre et se réfèrent à la fois à la complexité du problème étudié ainsi qu'à la capacité à bien exécuter les actions. Quand le problème concerne un ensemble restreint d'objets et de relations simples, il est possible de se représenter l'état du monde tel qu'il est. On suppose, par exemple, que l'espace d'état est fixe et connu au départ. Cela signifie que l'on a un ensemble dès le départ, tel qu'on l'a décrit ici, soit, on connaît une fonction de transition $T : EE$ qui indique quels sont les états accessibles à partir de l'état initial $e_{initial}$. Dans tous les cas, ce formalisme fait l'hypothèse qu'on connait les actions possibles et leurs conditions d'application, ce qui permet la transition d'un état à l'autre. Il est alors possible de tester et donc de connaître toutes les combinaisons d'actions possibles pour atteindre l'état final espéré. Dès que la taille du problème atteint quelques centaines de variables, le modèle devient rapidement trop complexe pour que toutes les combinaisons puissent être connues et testées. Pire, même dans le cas où celui-ci connaîtrait toutes les conséquences de ces actions, il pourrait très bien faire des erreurs dans l'application de certaines de ces actions.

1.3.3. Appropriation du modèle dans notre contexte

Ce formalisme, bien connu a été largement utilisé dans diverses littératures aussi bien pour étudier le raisonnement cognitif et le processus de conception que pour construire la logique informatique. Il permet en effet d'avoir une représentation simple et intuitive d'un processus de

raisonnement et de proposer un cadre et une méthode pour résoudre un problème. Nous avons également vu que sa simplicité a également été fréquemment critiquée, remettant en cause la légitimité de son utilisation dans un certain nombre de situations. Dans le cadre de notre étude, nous allons élaborer notre interprétation des tâches déléguées à la foule au travers de ce formalisme. Nous ne le présentons ni comme un modèle cognitif ni dans le but d'une résolution informatique, mais plutôt comme un modèle qui nous permet de cerner et discuter la question de la performance d'une foule. Nous allons reprendre l'ensemble des tâches que nous avons déjà identifiées dans le chapitre précédent, à savoir tâche élémentaire, recette et résolution de problèmes et les réinterpréter dans le cadre du formalisme traditionnel. Nous montrerons que ce formalisme permet d'identifier les critères de performance associés à chacune des tâches. Ensuite nous verrons quels sont ces limites pour prendre interpréter le processus de formulation d'hypothèses basées sur les données et nous proposerons une extension de ce formalisme.

2. TACHE ELEMENTAIRE, RECETTE, RESOLUTION DE PROBLEME ET PERFORMANCE

2.1. LA TACHE ELEMENTAIRE

Dans le chapitre précédent, l'exemple de l'expérience de Rutherford a servi à illustrer la notion de tâche élémentaire. En reprenant le formalisme que nous avons présenté plus haut, la notion de tâche élémentaire peut être modélisée comme l'application d'une unique action a à partir de l'état du monde initial $e_{initial}$. Cette action a pour effet d'atteindre un nouvel état désiré e_{final} situé au voisinage direct de l'état initial selon le graphe orienté. L'espace E des états est réduit au couple $\{e_{initial}, e_{final}\}$ et l'espace des actions se limite à la seule action $a : e_{initial} \rightarrow e_{final}$. Cette action est supposée ne pas pouvoir être facilement divisible en une séquence de sous actions menant à des états intermédiaires. Dans l'exemple de Rutherford, l'action consiste à détecter l'éclair de lumière formé sur l'écran de sulfure de zinc durant l'expérience. L'état du monde initial est formalisé par l'unique objet d'intérêt, à savoir l'écran de sulfure de zinc, ainsi que par la propriété « éclair de lumière » sur l'objet. A noter que si la tâche élémentaire n'est réduite qu'à une seule action elle peut en revanche nécessiter une grande dextérité et pouvant mener à des erreurs de la part de l'exécutant. Pour nous en persuader, nous illustrons un cas de tâche élémentaire dans le domaine de la cuisine.

Analogie avec la cuisine : La découpe du poisson par un maître sushi

Dans la cuisine japonaise, la préparation d'un sushi répond à un cérémonial complexe que seuls quelques initiés peuvent pratiquer en théorie. En effet, le sushi représente un élément essentiel de la gastronomie locale, pour laquelle la conception de celui-ci représente plus un art de vivre qu'un simple geste technique. Pour obtenir le statut de maître-sushi (appelé *sushiya*), pas moins de dix années de formation sont nécessaires. La découpe de fines tranches de poissons crus découle d'un entraînement très long et d'une connaissance totale des poissons et de la façon dont ceux-ci

doivent être découpés. Nous pouvons supposer qu'il existe un geste de découpe optimal du poisson qui permet d'évaluer le travail du maître-sushi sur un ensemble de critères : la finesse de la tranche, la forme de la découpe dans le filet, l'aspect visuel de la tranche.

CRITERES DE PERFORMANCE DE LA TACHE ELEMENTAIRE

Les tâches élémentaires sont généralement des tâches répétitives, où la même tâche va être exécutée plusieurs fois successivement. Pourtant, si l'action est globalement identique à chaque itération, les états du monde initiaux $e_{initial}$ et finaux e_{final} ne sont jamais exactement semblables. Un même geste répété sur des états du monde différents générera potentiellement des résultats légèrement différents. Ce ne sont pas N tranches de poissons que le maître-sushi obtient, mais N fois une tranche de poisson légèrement différente dont chacune est différente de toutes les autres tranches de poisson.

D'abord, il y a toujours plus ou moins de précision dans l'application de l'action. La fatigue du cuisinier qui découpe le poisson peut influer au cours du temps sur la qualité de la découpe de ses sushis. Ensuite, d'autres éléments peuvent avoir un effet sur l'application de l'action comme la qualité des ingrédients (deux filets de poisson ne seront jamais parfaitement identiques) ou les outils utilisés (quel type de couteau). L'analyse de la performance ne se mesure généralement pas sur une tâche exécutée seule, mais plutôt sur l'ensemble de tâches similaires exécutées. Cette performance correspond triptyque traditionnel qualité-coût-délai. Dans le cas du maître sushi, la qualité du geste peut se baser sur un ensemble de critères : la finesse de la tranche, la forme de la découpe dans le filet, l'aspect visuel de la tranche. On peut également prendre en compte le temps passé à découper chaque tranche de sushi ainsi que le coût (usure de couteau après un ensemble de N découpes ? Nombre de pièce de sushis défectueuses ?)

2.2. LE MODELE DE LA RECETTE

La tâche de type recette est définie comme l'application d'une succession d'actions $\{a_1, a_2, \dots, a_n\}$ suivant un ordre précis. Cette séquence a pour effet d'atteindre un nouvel état désiré e_{final} en passant par autant d'états intermédiaires qu'il existe d'actions dans la séquence à exécuter. L'espace *E* est constitué de l'ensemble des états à atteindre $\{e_{initial}, e_1, e_2, \dots, e_{n-1}, e_{final}\}$. Il est à noter que l'ordre d'exécution des actions est important car l'état du monde qui résulte de l'action a_i doit être compatible avec l'action a_{i+1} qui suit. Dans l'exemple de Galaxy Zoo, l'état du monde initial correspond à la photo de la galaxie non encore codée et l'état final correspond à la photo finale codée. Chaque action correspond à une étape prédéterminée par les organisateurs et fournit une information qui impacte les choix qui vont suivre. Si par exemple l'exécutant code que la galaxie est lisse et sans « traînées », c'est-à-dire sans spirales, le logiciel lui demandera à l'étape suivante de préciser si celle-ci a une forme ronde ou en cigare.

La recette est une tâche qui peut être décrite via un protocole suivant la succession d'actions à mener. Les ambiguïtés sont en général minimisées pour que la tâche puisse être réalisée avec le minimum d'improvisation de la part de l'exécutant. Dans le cadre du processus scientifique, une tâche de type recette peut correspondre au travail expérimental réalisé par le laborantin. Chaque étape de l'expérience est définie au préalable suivant les directives imposées par le scientifique en charge de la recherche et constitue le protocole expérimental. Le laborantin ou le technicien de laboratoire exécute ensuite les différentes actions et répertorie dans un carnet toutes ses observations et les résultats.

CRITERES DE PERFORMANCE DE LA RECETTE

Les conclusions que nous avons établies sur les cas de tâches élémentaires s'appliquent au cas de la recette. La différence principale entre la tâche élémentaire et la tâche de type recette réside en l'existence pour cette dernière des états intermédiaires. Il faut alors évaluer si la séquence d'actions est bien réalisée dans le bon ordre, et si les états intermédiaires permettent d'appliquer l'action suivante. Lorsque nous préparons une omelette, une poêle légèrement trop chaude pourra faire subir une action irréversible sur l'état du monde (brûler la partie inférieure de l'omelette) et ne permettra pas d'obtenir le résultat final espéré. Une mesure de la performance peut être réalisée à chaque état intermédiaire afin d'éventuellement d'appliquer des mesures correctives.

2.3. LA RESOLUTION DE PROBLEMES

La résolution de problèmes est décrite comme la recherche d'une solution à partir d'un état e_{final} fixé au départ. Contrairement aux tâches élémentaires et aux tâches de type recette, la séquence d'actions n'est pas connue à l'avance ou bien n'est pas compatible avec les objectifs définis (budget, temps d'exécution, simplicité). La résolution de problèmes consiste justement en l'exploration de l'espace des actions pour trouver une séquence qui va permettre de résoudre le problème. L'exécutant choisit et applique une séquence $\{a_i, a_{i+1}, \dots, a_{i+k}\}$ d'actions puis exécute la fonction *test* pour évaluer si le problème est résolu. Tant que la séquence d'action ne permet pas d'avoir une solution satisfaisante, il en essaie une autre. A la manière d'une exploration dans un labyrinthe où on doit trouver la sortie, les exécutants explorent différentes séquences d'actions jusqu'à obtenir une solution. Une fois une solution trouvée celle-ci peut être réappliquée plusieurs fois : nous revenons à une tâche de type recette.

L'état final désiré e_{final} peut ne pas être entièrement connu à l'avance dans la résolution de problème. En fait le problème à résoudre est défini par un couple $< R, D >$ d'objets et de relations entre ces objets à atteindre. Plusieurs états peuvent inclure cette condition au sein d'une structure d'objets et de relations plus grande $< R', D' > \subset < R, D >$. Le deuxième point spécifique à la résolution de problèmes est que l'espace des actions possibles $\{a_1, a_2, \dots, a_n\}$ est considéré

comme prédéterminé en amont de la tâche. Cela signifie qu'il est possible de représenter entièrement l'espace des actions possibles que l'on peut appliquer à partir de notre état de départ et des états intermédiaires (Simon, 1973). La taille de cet espace varie énormément suivant le type de problème rencontré. Plus celui-ci est grand, plus le processus d'exploration peut potentiellement mener à tester un grand nombre de tests de séquences différentes. Dans l'exemple du challenge du boson de Higgs que nous avons présenté dans le chapitre 4, le problème d'optimisation défini par les organisateurs est représenté par un unique critère de performance à améliorer : le degré de confiance en la détection d'une anomalie. L'espace des actions possibles est très grand : celui-ci contient au moins l'ensemble des algorithmes applicables à une base de données ainsi que toutes les combinaisons d'algorithmes, plus toutes leurs variantes. En effet, tout algorithme ou méthode informatique peut être considéré comme une action car elle transforme les états de la machine.

Exemple en cuisine : le plat trop salé

Nous illustrons la tâche de type résolution de problème par une situation simple en cuisine. Imaginons le cas d'un commis en cuisine à qui on a demandé de faire une assiette de pâtes. Avant de dresser l'assiette, le chef en cuisine goûte le plat et se rend compte qu'il est trop salé. Il demande alors au commis de faire en sorte de modifier l'assaisonnement de l'assiette pour diminuer l'effet salé du plat. Le commis peut alors tester plusieurs séquences d'actions avec les ingrédients et les ustensiles qu'il a en sa possession dans sa cuisine : utiliser du jus de citron pour atténuer l'effet salé, laver les pâtes à grande eau, intégrer plus d'ingrédients dans la préparation pour répartir la quantité de sel, mettre temporairement une pomme de terre crue dans le plat pour absorber l'excès de sel. L'état final auquel il parviendra dépend du choix de la séquence qu'il va mettre en œuvre. Par exemple, s'il augmente la quantité d'ingrédients, il y aura plus de quantité du plat que s'il avait utilisé une pomme de terre crue.

CRITERES DE PERFORMANCE DANS LA RESOLUTION DE PROBLEMES

Contrairement aux tâches de type élémentaire et recette, le problème majeur de la résolution de problèmes est de trouver la bonne séquence d'action ou la bonne recette. La performance du processus est liée à la fois au processus d'exploration de l'espace des actions et à la solution choisie. En effet, nous avons signalé dans la section 1.2 que l'exploration de l'espace des actions engendre un coût qui augmente d'autant que la taille de l'espace est importante. Chaque tentative infructueuse d'application d'une séquence d'action implique des coûts en terme d'exécution ainsi qu'un coût en terme de temps. Il peut donc être intéressant de chercher une méthode afin de réduire le nombre d'itération dans le processus d'exploration, et donc d'augmenter la *vitesse* pour trouver une bonne solution.

Ensuite, les critères de *qualité-coût-délai* s'appliquent dès lors qu'un plan d'action ou une recette est défini. D'un point de vue du coût, nous avons constaté dans la première section que l'exécution d'une séquence d'action représentait un coût que l'exécutant devrait chercher à minimiser. Ainsi, en plus de chercher une solution permette de résoudre le problème, celle-ci doit correspondre à

un ensemble de contraintes budgétaires. Ensuite, la notion de qualité peut être interprétée de deux manières. D'abord comme la qualité de l'exécution du plan d'action comme cela a été précisé pour la tâche élémentaire et recette. Ensuite, la qualité peut être associée à un critère associé au problème. Dans le cas du challenge sur le boson de Higgs, il est toujours possible d'améliorer l'unique critère de performance, en proposant une solution qui a un meilleur score. Enfin, le délai est le temps que prend la recette à être appliquée.

Afin de spécifier les contraintes de gestion, supposons un ensemble de critères à respecter que l'on peut représenter par une valeur $v_{désirée}$. Ces critères peuvent être liées au budget alloué, à des contraintes temporelles pour exécuter la tâche, ou à un niveau de qualité désirée. A chaque fois que le participant soumet un plan d'action, l'organisateur évalue l'état final obtenu e_{final} avec l'état désiré $e_{désiré}$ grâce à la fonction *test()*. Le résultat obtenu peut être associé à un point de la fonction de valeur $v(e_{final})$ pour l'état final e_{final}. Si $v(e_{final}) \geq v_{désirée}$, le processus s'arrête. Dans le cas contraire, le processus continue.

La performance de l'exploration concerne également la vitesse à laquelle le problème est résolu. Cette vitesse dépend en grande partie de la position de départ (i.e. l'état initial $e_{initial}$) choisie par le participant. Nous pouvons l'illustrer par une méthode d'exploration bien connue en informatique qui est celle du processus d'escalade (hill climbing). Le principe est d'implémenter un algorithme itératif qui, à partir d'une solution donnée, va chercher à trouver la meilleure solution en proposant des changements incrémentaux à la solution. Dans ce cas, l'opérateur ne proposera que des solutions qui sont au voisinage de la solution qu'il a précédemment soumise. Par exemple s'il propose initialement un plan d'action $\{a_1, a_2, \dots, a_n\}$ qui mène à l'état e_1, la solution suivante sera par exemple de la forme $\{a'_1, a_2, \dots, a_n\}$, où l'action a_1 a été remplacée par l'action a'_1. Ces séquences d'action mènent respectivement aux états e_1 et e_2 auxquelles sont associées les valeurs $v(e_1)$ et $v(e_2)$. Ces deux valeurs fournissent un gradient qui aide à diriger l'exploration. Si $v(e_2) > v(e_1)$, alors l'opérateur considère qu'il est dans la bonne direction et continue à proposer des solutions du même type (par exemple proposer une autre action à la place de a'_1). Au contraire, si $v(e_2) < v(e_1)$, l'opérateur proposera d'autres modifications (par exemple changer d'autres actions de la séquence initiale). L'opérateur s'arrête au moment où il atteint une valeur maximale sur la fonction de valeur, c'est-à-dire que toute transformation incrémentale est associée à une valeur *v()* plus faible.

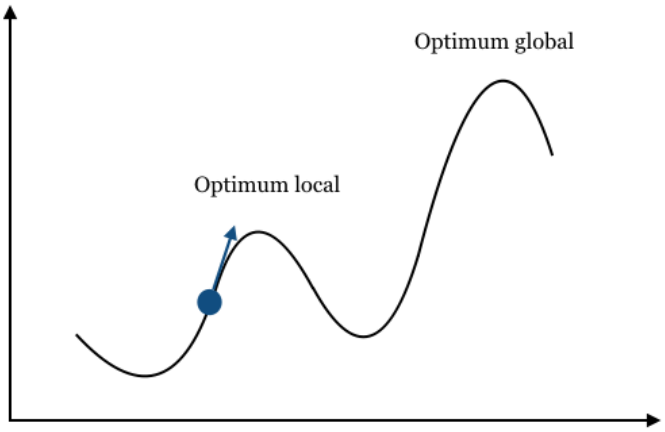


Figure 10. Illustration de la fonction de valeur avec un optimum local différent de l'optimum global.

Cette approche présente ses limites si la solution initialement choisie n'est pas située proche de la solution optimale. En effet, la solution peut se retrouver localement optimale (meilleure que tous ses voisins proches) sans pour autant être la meilleure solution globale, c'est-à-dire celle avec la valeur *v()* la plus élevée. Ainsi si la solution de départ est loin de la solution optimale, il sera tenté d'atteindre la meilleure solution de son voisinage (optimum local) sans savoir que celle-ci n'est pas la meilleure solution (**Figure 10**).

La difficulté réside dans la difficulté de connaître de façon exhaustive tous les points de la fonction de valeur *v()*. Dès que le nombre d'actions possibles est très grand, cela prendrait beaucoup trop de temps au seul participant d'explorer toutes les séquences d'actions possibles. Plusieurs études ont montré que lorsque l'incertitude à résoudre le problème est importante, la recherche d'une solution est souvent sujette à de « mauvais chemins », de « l'expérimentation », de la « sérendipité » et de « l'incertitude » (Abernathy & Rosenbloom, 1969; Boudreau, Lakhani, & Lacetera, 2008; Loch, Terwiesch, & Thomke, 2001). Cela amène alors à augmenter les pertes de production.

2.4. INCONSISTANCE DES MODELES POUR LA GENERATION D'HYPOTHESES SCIENTIFIQUES BASEES SUR LES DONNEES

Nous avons vu au travers des différents exemples du chapitre 4 que les modèles de tâche présentés sont applicables à plusieurs situations dans le processus de découverte scientifique. Ce modèle formel permet d'interpréter notamment certaines tâches déléguées dans le cadre des sciences citoyennes comme la collecte et le codage de données, l'application du protocole

expérimental ou l'analyse de données. Mais est-il suffisant pour interpréter la génération des hypothèses scientifiques ? Pour répondre à cette question, nous allons présenter comment la génération d'hypothèses a été modélisée par les informaticiens.

2.4.1. Méthodes informatiques d'automatisation de la génération d'hypothèses : le dual-space search

Un certain de nombre de travaux remarquables ont été réalisés dans la modélisation informatique du processus de découverte scientifique. Les programmes développés ont généralement pour objectif d'automatiser tout ou partie du processus afin de faciliter la génération des hypothèses (Kalinichenko, Kovalev, Kovaleva, & Malkov, 2015). Dans leurs modèles, la génération d'hypothèses est décrite comme faisant partie d'une activité de résolution de problèmes plus générale, et donc modélisable avec le modèle traditionnel. Nous avons identifié trois niveaux d'automatisation du processus de découverte et de génération d'hypothèse scientifique. Un premier niveau dans lequel les hypothèses scientifiques sont intégrées dans le programme qui automatise l'exploration de l'espace des hypothèses. Dans un deuxième niveau, le programme formule les hypothèses scientifiques automatiquement à partir des informations fournies par les scientifiques. Enfin, le troisième niveau propose un système entièrement automatique, un robot, qui génère les hypothèses scientifiques et réalise lui-même les expérimentations.

2.4.1.1. Exploration de l'espace des hypothèses

Les programmes d'exploration sont généralement développés dans un domaine scientifique particulier tels que la biologie (Callahan, Dumontier, & Shah, 2011; Racunas et al., 2004; Soldatova & Rzhetsky, 2011) ou la neuromédecine (Gao et al., 2006). Le principe est d'accumuler dans un programme informatique les savoirs connus du phénomène d'intérêt, ainsi que les différentes théories et hypothèses existantes dans le domaine. Cette programmation facilite l'exploration au sein de l'espace des hypothèses et peut être associé à un algorithme qui automatise le processus pour valider ou infirmer les hypothèses formulées par les scientifiques vis-à-vis des savoirs connus.

Dans le projet HyBrow (Hypothesis Space Browser) (Racunas et al., 2004) par exemple, les hypothèses sont représentées sous la forme d'un ensemble de phrases. Les concepteurs de l'outil ont développé une ontologie des termes et des concepts du domaine biologique. On retrouve ainsi des « agents» (gènes, protéines,…) qui sont reliés entre eux par des « opérateurs » (active, est similaire à, transporte,…). Ces phrases sont issues des théories et des modèles scientifiques issus de la discipline. Un ensemble d'axiomes est défini en tant que règles qui modélisent des faits biologiques connus ainsi que des données expérimentales et qui constituent la base de connaissance. Cette base peut infirmer ou valider certaines phrases contenues dans l'espace des hypothèses, laissant les autres phrases candidates à une découverte. Ainsi, à mesure que davantage de données expérimentales sont obtenues et que les règles associées sont identifiées, les phrases de l'espace des hypothèses deviennent des faits qui sont valides ou invalides. Dans le

cas où une phrase est contredite, les règles à l'origine des problèmes doivent être identifiées et éliminées de la théorie substantielle aux hypothèses. En fait, le programme HyBrow permet de vérifier les théories scientifiques existantes par l'analyse du statut logique des hypothèses que l'on peut en déduire. Le logiciel met en avant les incohérences avec les données expérimentales afin de pouvoir modifier le modèle.

2.4.1.2. Formalisation d'hypothèses scientifiques

D'autres programmes sont développés pour créer de nouvelles hypothèses scientifiques à partir des hypothèses et théories déjà existantes. Ils se basent sur un langage séquentiel au travers d'équations, de formules mathématiques ou de langage logique afin de pouvoir représenter les théories existantes comme des séquences d'actions (Asgharbeygi et al., 2006; Kulkarni & Simon, 1988; Tran et al., 2005). Les scientifiques implémentent ensuite un ensemble de règles et d'heuristiques à respecter pour que le programme puisse générer de nouvelles hypothèses à partir des hypothèses existantes. Une fois ces éléments mis en place, les scientifiques intègrent dans le programme de nouvelles observations non prises en compte dans les théories et les hypothèses existantes. Le programme génère alors à partir de toutes ces entrées de nouvelles hypothèses qui intègrent les nouvelles observations.

Asgharbeygi et ses collègues présentent par exemple un algorithme général de découverte scientifique dont le principe est d'améliorer les modèles scientifiques existants (Asgharbeygi et al., 2006). Chaque modèle est représenté de manière formelle sous la forme d'une équation ou d'une formule mathématique. La révision du modèle nécessite que l'utilisateur spécifie quatre entrées dans le programme : un modèle initial qui code les croyances sur les processus les plus susceptibles d'être impliqués; un ensemble de contraintes représentant les modifications acceptables du modèle initial et spécifiant quels processus initiaux doivent être corrigés, supprimés ou dont leurs paramètres doivent être modifiés; un ensemble de processus génériques pouvant être ajoutés au modèle initial; les observations auxquelles le modèle révisé devrait s'adapter. Le modèle initial constitue la meilleure estimation par l'utilisateur des processus présents dans le système, alors que les modifications autorisées indiquent ses zones d'incertitude. Ces éléments fournissent une heuristique qui guide la recherche vers des parties de l'espace compatibles avec la connaissance du domaine.

Le programme fonctionne ensuite en deux étapes principales. La première consiste à rechercher toutes les structures de modèles compatibles avec les contraintes spécifiées. Le système génère d'abord toutes les instanciations des processus génériques recommandés par l'utilisateur qui satisfont aux contraintes imposées aux types de variables; ceux-ci deviennent des candidats à l'ajout au modèle. Ensuite, le programme effectue une recherche dans l'espace des structures de modèle, en utilisant le modèle initial comme état de départ. Le principe est de s'intéresser d'abord aux modèles les plus proches du modèle initial pour s'assurer de la cohérence globale. Le résultat est un ensemble de modèles révisés qui cherchent à expliquer les relations entre les variables.

Une autre série de travaux remarquables sur ce sujet provient de modèles informatiques développés par Herbert Simon et ses collègues (Antonsson & Cagan, 2001; Kulkarni & Simon, 1988; Lindsay et al., 1993). KEKADA, par exemple, est un algorithme développé par Kulkarni et Simon en analysant le processus du biologiste Hans Krebs en matière de recherche sur la production d'urée (Kulkarni & Simon, 1988). Cet algorithme a permis d'automatiser certaines tâches du processus scientifique. Dans son principe, KEKADA prévoit une séquence d'expériences afin de produire des observations pouvant être utilisées pour formuler des théories descriptives et explicatives d'un ensemble de phénomènes. Le processus de formulation d'hypothèses dans KEKADA est présenté comme un processus en deux étapes: la génération de nouvelles hypothèses puis le choix de l'hypothèse selon les heuristiques issues d'un processus décisionnel.

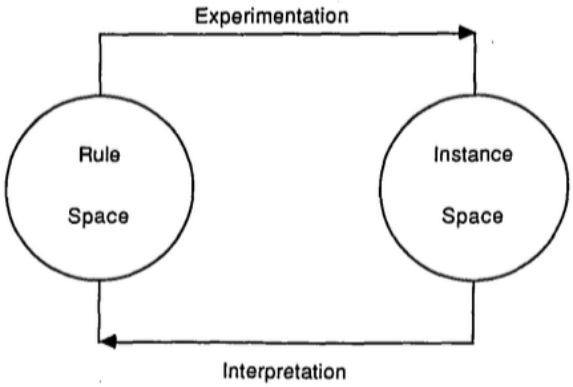


Figure 11. Modèle à double espace (Kulkarni & Simon 1988)

Le compte-rendu des recherches de Krebs montre un processus de génération d'hypothèse itératif. L'arrivée à la découverte finale peut être divisée en trois grands segments. Krebs a commencé par tester l'efficacité de divers acides aminés dans la production d'urée, avec des résultats généralement négatifs. Lors de l'expérience avec l'ornithine (l'un des acides aminés les moins courants) et l'ammoniac, de grandes quantités inattendues d'urée ont été produites. Il s'est ensuite concentré sur l'effet ornithine. Krebs a ensuite suivi une stratégie standard: si un composé donné exerce une action particulière, vérifier si les dérivés de ce composé ont une action similaire. Ainsi, il a effectué des tests sur certains dérivés de l'ornithine et des substances similaires à l'ornithine. Mais aucune de ces substances n'avait d'effet comparable à celui de l'ornithine. Enfin, il s'est intéressé au chemin de réaction qui mène de l'ornithine à la production de l'urée. Il s'est rendu compte que la réaction (connue) de l'arginine, par laquelle l'arginine est convertie en ornithine et en urée, pourrait être liée à l'effet ornithine. En concluant à partir des données quantitatives que l'ornithine ne pouvait être qu'un catalyseur, il en a déduit que l'ornithine avec de l'ammoniac produit de l'arginine, laquelle produit à son tour de l'urée et de l'ornithine.

Dans le modèle de Kulkarni, le processus de découverte scientifique est modélisé suivant un espace de règles et un espace d'instance. La formulation des hypothèses scientifiques est réalisé au sein de l'espace des règles piloté par un ensemble d'heuristiques automatisées dans le programme

KEKADA sous forme de modules : un générateur et un sélectionneur de problèmes, un générateur et un sélectionneur d'hypothèses, un modificateur d'hypothèses et un modificateur de niveau de confiance d'une hypothèse sur la base de nouvelles informations. L'espace d'instance intègre l'ensemble des résultats expérimentaux obtenus par les scientifiques eux-mêmes et implémentés au sein du programme.

2.4.1.3. Automatisation de la découverte scientifique

Un groupe de chercheurs a développé à l'université d'Aberystwyth un « robot scientifique » pour le domaine de la génomique où le processus de découverte scientifique est entièrement automatisé (King et al., 2009). Adam, le premier prototype de ce type de robot, est capable d'exécuter de manière entièrement automatique des cycles d'expérimentation et de découverte scientifiques: formulation d'hypothèses, sélection d'expériences pour tester ces hypothèses, exécution d'expériences avec un système robotique, analyse des résultats et interprétation. Ce système est conçu comme une boucle fermée dans laquelle les résultats obtenus sont utilisés pour en tirer des enseignements en réintroduisant les connaissances obtenues dans les modèles expérimentaux **(figure 12)**.

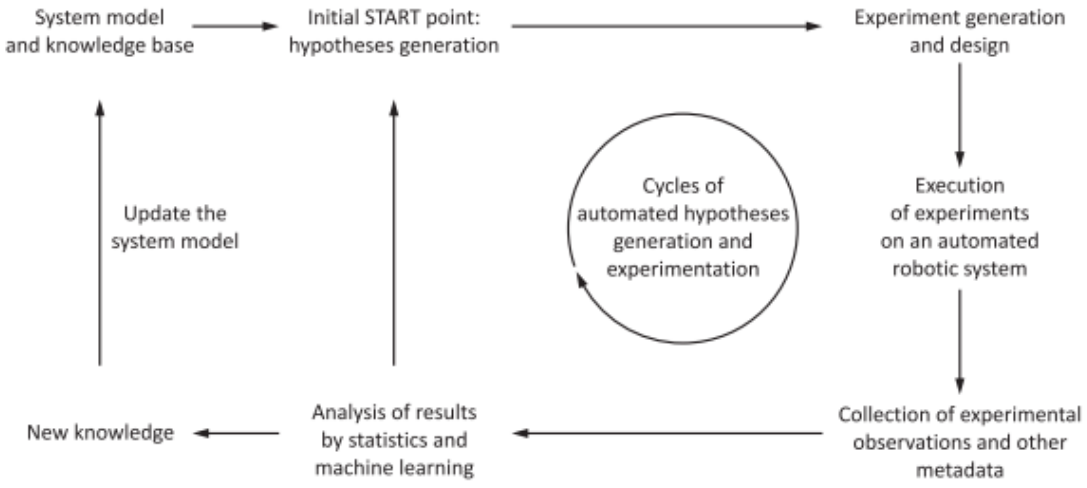


Figure 12. Cycle en boucle fermée du robot scientifique (Kalinichenko et al., 2015)

Dans le système Adam, la formulation automatisée d'hypothèses est basée sur un ensemble de composantes clés du programme : représentation calculable par la machine des connaissances du domaine; les nouvelles hypothèses sont produites par inférence abductive ou inductive; sélection des hypothèses; processus de déduction des conséquences expérimentales des hypothèses. Pour formaliser les hypothèses, le robot se base sur une ontologie associée aux différentes étapes de recherche : essai, étude, cycle d'étude et de reproduction, stratégie de conception, disposition des plaques, résultats réels attendus.

Adam a été conçu pour mener des expériences de croissance microbienne afin d'étudier la génomique fonctionnelle de la levure *Saccharomyces cerevisiae*, en particulier pour identifier les

gènes codant pour «des enzymes localement orphelines». Des enzymes sont dites orphelines quand on ne connaît pas le gène, ou la séquence de ce gène, responsable de leur création et de leur fonction. Adam utilise un modèle logique complet du métabolisme de la levure, associé à une base de données bioinformatique et des techniques classiques de recherche d'homologie bioinformatique permettant de faire l'hypothèse de gènes candidats probables pouvant coder les enzymes localement orphelines. Ce processus de génération d'hypothèses est abductif.

Deux types d'hypothèses sont générés par le programme. Le premier niveau relie une enzyme orpheline, représentée par son numéro de classe d'enzyme (E.C.), à un gène (ORF) qui la code potentiellement. Cette relation est exprimée sous la forme d'un système à deux positions, le premier argument étant l'ORF et le second l'E.C. Le deuxième niveau d'hypothèse implique l'association entre une souche spécifique, référencée via le nom de son ORF manquant, et un composé chimique qui devrait affecter la croissance de la souche, s'il est ajouté comme nutriment à son environnement. Ce niveau d'hypothèse est dérivé du premier par déduction logique en utilisant un modèle spécifique du métabolisme de la levure.

Adam conçoit ensuite les tests expérimentaux nécessaires pour tester l'exécution de ces hypothèses sur le système robotique. Ces expériences sont basées sur une conception à deux facteurs qui comparent plusieurs réplicats des souches avec et sans métabolites par rapport à des souches de contrôle de type sauvage avec et sans métabolites. Pour sélectionner les expériences, Adam prend en compte le coût variable des expériences et les différentes probabilités d'hypothèses. Adam choisit ses expériences pour minimiser le coût attendu d'éliminer toutes les hypothèses sauf une.

2.4.2. Génération des hypothèses dans un double espace

Dans tous ces programmes informatiques, la génération des hypothèses scientifiques est présentée comme un processus itératif alternant des expérimentations. Par la mise en place d'heuristiques, les scientifiques génèrent et choisissent une hypothèse dans l'espace des hypothèses. Ensuite, ils conçoivent et réalisent un certain nombre d'expérimentations pour évaluer la validité de cette hypothèse. Ce processus apporte de nouvelles observations qui nourrissent les heuristiques et donc permettent de modifier l'hypothèse de base pour qu'elle corresponde aux nouvelles observations. Le processus s'arrête au moment où le scientifique est satisfait des résultats qu'il obtient, c'est-à-dire au moment où il a une hypothèse viable et qu'il est capable de statuer sur la véracité de celle-ci.

Toute la partie expérimentale reste peu décrite par les modèles. En effet, ceux-ci se concentrent surtout sur la génération de l'hypothèse plutôt que sur sa vérification et ne décrivent pas cette partie du processus. Dans le cas du programme KEKADA, la partie expérimentale n'est même pas décrite par le programme. En fait, c'est le scientifique qui réalise les expérimentations et qui ensuite implémente les résultats dans le programme. Seul dans le cas du robot Adam la partie expérimentale du processus de découverte scientifique est automatisée et donc décrite. Celle-ci

inclut la conception de l'expérimentation, la réalisation de l'expérimentation et la collecte de données expérimentales, l'analyse des résultats expérimentaux par des méthodes statistiques, l'extraction de nouvelles connaissances, l'intégration de ces connaissances dans la base de savoirs déjà existants. L'espace des actions possibles est borné : il correspond à toutes les actions possibles que le robot peut effectuer.

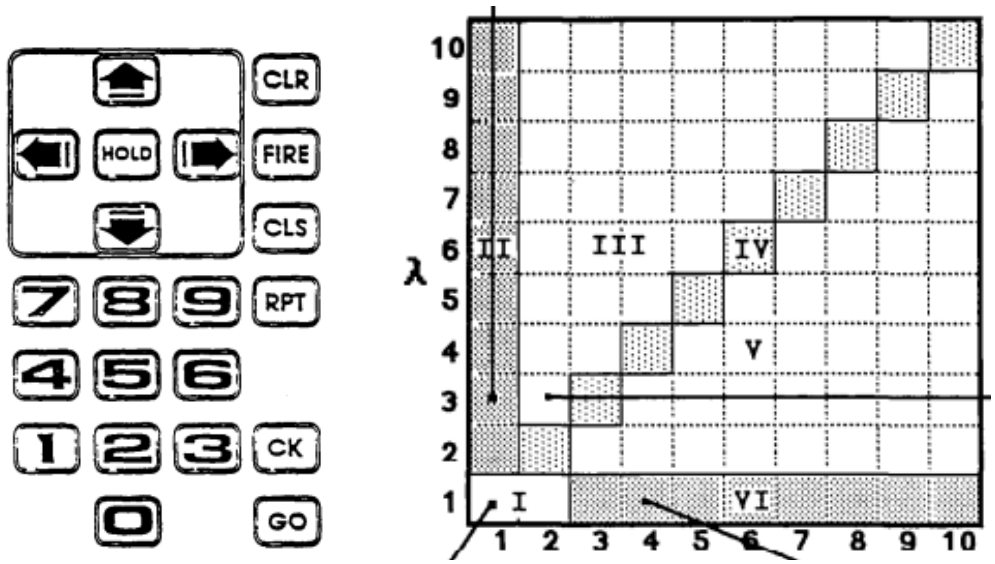
La génération des hypothèses fait donc appel à deux processus extrêmement différents. D'un côté, la formulation d'une hypothèse scientifique répond à des raisonnements purement intellectuels. Ce mécanisme fait généralement appel à des modes d'inférence, tel que la déduction, l'induction ou l'abduction. Ensuite, l'expérimentation demande d'agir sur les objets du monde.

De ces deux activités découle naturellement le besoin de travailler sur deux espaces différents : un premier espace appelé « espace des hypothèses » dans lequel se construisent les hypothèses scientifiques, et un second espace pour l'expérimentation que l'on appelle « espace des plans d'action ». L'espace des hypothèses est basé sur l'ensemble des données disponibles et utilisées pour formuler des hypothèses. L'espace des plans d'action comprend toutes les actions possibles que le scientifique peut mettre en œuvre : entre autres les expérimentations, l'analyse de données, la collecte de données. Bien sûr, le modèle n'est qu'une représentation partielle du processus de découverte scientifique et il est tout à fait possible de rajouter d'autres espaces (Klahr & Simon, 1999). Par exemple, nous ne prenons pas en compte dans ces modèles la personnalité du scientifique, les ressources auxquelles il a accès ou bien les instruments scientifiques utilisés. Cependant, nous considérons que la représentation à deux espaces permet de se focaliser sur les points essentiels de la formulation des hypothèses basées sur les données et donc suffit à l'analyse de nos terrains de recherche.

2.4.3. Le modèle du dual-space search

La génération des hypothèses basées sur les données ne peut donc se réduire au modèle de résolution de problèmes dans un seul espace. En fait, elle demande d'explorer deux espaces à la fois : l'espace des hypothèses et l'espace des algorithmes pour analyser les données. Un modèle à deux espaces a déjà été proposé dans la lignée des travaux menés par Kulkarni et Simon (1988) et le programme KEKADA. Ce modèle, appelé dual-space search, a été présenté par (Klahr & Dunbar, 1988) comme logique de découverte scientifique. Pour le construire, les chercheurs ont réalisé deux expérimentations avec respectivement 10 et 20 participants afin d'identifier les stratégies de construction et de validation d'hypothèses scientifiques. Ils se sont particulièrement intéressés à la performance de ces stratégies suivant que l'exploration part de l'espace des hypothèses ou de l'espace des expérimentations. L'expérimentation menée consiste en un robot commandé par ordinateur (appelé « BigTrak ») qui est programmé en un langage de type LOGO. Il s'agit d'un véhicule à six roues et alimenté par batterie, mesurant environ 30 cm de long, 20 cm de large et 15 cm de hauteur. L'interaction se fait via un clavier en haut de l'appareil (**figure 13**)

tandis que l'espace d'expérimentation peut être représenté comme un espace physique borné de 10x10 cases dans lequel le robot évolue.



Figures 13. Clavier de contrôle qui représente l'espace des hypothèses (à gauche) et espace d'expérimentation (à droite)

Chaque touche du clavier de contrôle correspond à une commande pré-enregistrée par les chercheurs. Le robot n'effectue les actions qu'une fois la touche « Go » activée, permettant de réaliser des combinaisons d'actions. Dans l'expérimentation, les chercheurs ont expliqué aux participants le rôle de toutes les touches, exceptée la touche « RPT ». Le but était d'analyser le processus mené par les participants afin de déterminer comment fonctionnait la touche « RPT ». Les résultats de l'expérimentation montrent que la découverte de la fonction de la touche « RPT » a demandé aux participants d'explorer les deux espaces en même temps : d'un côté ils effectuent des combinaisons de touche sur la manette tandis que l'espace d'expérimentation permet d'observer les effets de ces combinaisons.

Dans ce modèle dual, les auteurs font l'hypothèse que les deux espaces dans lesquels l'exploration est réalisée sont entièrement représentables *ex ante*. En effet, toutes les touches à partir desquelles les participants peuvent faire des combinaisons sont déjà existantes, de même que l'espace d'expérimentation est borné. Ainsi dès qu'un « point » (une hypothèse ou une case de l'espace d'expérimentation) est en dehors des espaces, il est non pertinent. Par exemple, les participants n'auraient pas pu intégrer une nouvelle touche supplémentaire pour tester la fonction « RPT », ni agrandir l'espace d'expérimentation en intégrant une nouvelle case.

De manière générale, le problème des points ou des zones qui ne rentrent pas dans les espaces préalablement déterminés ne sont pas intégrés par les informaticiens, car il n'existe pas de mécanisme qui permette d'y arriver (générer un point qui est en dehors de l'espace initial). Or, dès que la modélisation n'est pas sur un ordinateur mais analyse un processus humain, il arrive que

l'espace sur lequel opère soit spontanément « expansionné » (Hatchuel, 2001). De nombreuses découvertes ou constructions comme l'ADN, les nombres imaginaires, ou les quarks, ne peuvent être simplement considérées comme une combinaison de tous les objets préalablement existants dans la connaissance scientifique et supposent à un moment une extension par ajout d'un nouvel objet. On peut donc supposer qu'il y a un opérateur de disjonction qui permette de générer la description d'un point qui serait en dehors des espaces connus (Hatchuel & Weil, 2009a). Cet opérateur, même si on ne sait pas "comment" il opère, est utile, notamment dans le contexte scientifique où la qualité d'une hypothèse est mesurée en partie par son originalité. Ce principe d'expansion des espaces n'est proposé ni dans le modèle initial de résolution de problème de Simon, ni dans son extension par (Klahr & Dunbar, 1988). Or, nous avons donc besoin de construire un modèle à double espace qui inclut le principe d'expansion pour intégrer la génération des hypothèses.

3. RESOLUTION DE PROBLEMES VS FORMULATION DE PROBLEMES : LA NOTION DE « TACHE COUPLEE INVENTIVE »

3.1. DEFINIR L'ACTIVITE SCIENTIFIQUE COMME UN ACTE PARTICULIER DE CONCEPTION

Nous avons suggéré dans la section précédente que les travaux menés dans les années 1980-1990 par Simon, Khlar et Dunbar entre autres, sont limités pour modéliser l'activité scientifique et la formulation des hypothèses. En effet, ces modèles ne prennent pas en compte l'expansion possible des espaces, c'est-à-dire l'intégration de nouveaux éléments. Leurs études se basent essentiellement sur le principe de problem solving développé qui est limité dès lors que le problème n'est pas bien posé (Simon, 1973). Ces approches modélisent le processus comme un acte de décision, c'est-à-dire de choisir une solution parmi l'ensemble des solutions disponible, généralement la plus optimale par rapport à un problème défini.

Depuis ces travaux, des avancées importantes dans la recherche en design ont permis une nouvelle interprétation de l'activité scientifique comme un processus de conception particulier (Hatchuel et al., 2013). En effet les deux processus, l'acte de concevoir et l'activité scientifique, présentent un ensemble de similitudes qui justifient de s'appuyer sur les théories de la conception pour chercher à modéliser le processus cognitif relatif à l'activité scientifique. Nous suivons ici la logique de la théorie de la conception C-K pour formaliser la fonction générative de la conception (Hatchuel & Weil, 2003, 2009). D'un côté, la conception vise à définir et à réaliser un objet qui n'existe pas déjà ou qui ne pourrait pas être obtenu par une déduction d'objets et de connaissances existants (Hatchuel et al., 2011). De manière similaire à un scientifique qui établit une hypothèse scientifique pour laquelle il ne connaît pas le statut logique, le concepteur est face à un objet inconnu défini par un ensemble de propriétés qui lui sont désirables mais qui ne sont pas encore observables dans l'ensemble des objets existants. Suivant cette représentation, l'objet inconnu n'existera uniquement que si le processus de conception est une réussite. En effet,

l'existence de cet objet n'est pas déductible à partir de l'ensemble des connaissances existantes. Ainsi, la conception du nouvel objet passe nécessairement par la construction de nouvelles connaissances nécessaires pour prouver l'existence de l'objet désiré et de ses propriétés. Ces connaissances ne se substituent pas aux connaissances existantes, mais viennent en complément : il y a donc une expansion de l'ensemble des connaissances.

D'un autre côté, la méthode scientifique se base sur une logique de modélisation. Ainsi, il est largement admis que la science est un processus qui produit des connaissances en utilisant à la fois des observations et des modèles (principalement mathématiques, mais pas uniquement). Les discussions scientifiques portent essentiellement sur la cohérence, la validité, la testabilité des modèles et, surtout, sur la manière dont les modèles peuvent s'adapter aux observations existantes ou provoquées de manière expérimentale. Dans une célèbre lettre écrite en 1933 à son ami Maurice Solovine, Albert Einstein propose une représentation de l'acte de modélisation (Einstein, 1993, p. 138-139) :

> *« (1) Les Es (expériences immédiates) sont nos données.*
> *(2) Les axiomes sur lesquels nous tirons nos conclusions sont indiqués par A. Psychologiquement, les As dépendent des Es. Mais il n'existe pas de voie logique menant de l'Es à l'As, mais seulement une connexion intuitive (psychologique), qui est toujours «en train de se retourner».*
> *(3) Logiquement, les instructions spécifiques S, S ', S' 'sont déduites de A; ces déclarations peuvent prétendre à l'exactitude.*
> *(4) Les As sont connectés aux Es (vérification par l'expérience). Un examen plus approfondi montre que cette procédure appartient également à la sphère extralogique (sphère intuitive), car la relation entre les notions montrées en S et les expériences immédiates n'est pas de nature logique.*
> *Mais la relation entre Ss et Es est (pragmatiquement) beaucoup moins certaine que celle entre As et Es. (Prenez la notion 'chien' et les expériences immédiates correspondantes.) Si une telle relation ne pouvait pas être établie avec un degré de certitude élevé (même si elle pouvait échapper à la logique), un mécanisme logique n'aurait aucune valeur dans la « compréhension de la réalité » (exemple: théologie). Tout cela se résume à la relation éternellement problématique entre le monde des idées et ce qui peut être expérimenté (expériences immédiates des sens). »*

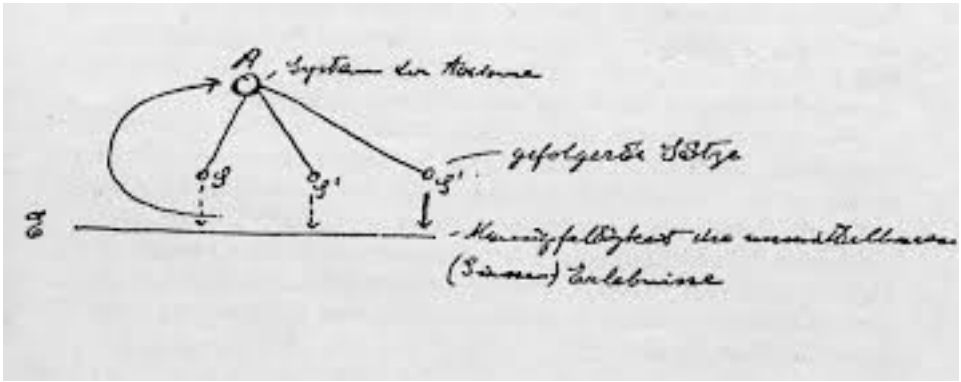


Figure 14. Dessin d'Einstein sur le mystère de l'épistémologie.

Einstein synthétise le raisonnement scientifique comme une suite d'expériences E1, E2,… et un système d'axiomes A. A partir d'un certain moment, les expériences E ne peuvent rendre compte par elles-mêmes d'un phénomène et il est nécessaire de remonter en abstraction de créer un nouvel objet, un A inconnu, tel que A rende compte de la suite des expériences Au final, l'acte de modélisation est très similaire au processus de conception, et peut être assimilé à un cas particulier de conception (Hatchuel et al., 2013). La différence majeure entre les deux apparaît dans les hypothèses: dans la modélisation, l'inconnu est considéré comme une «réalité externe» pouvant être observée; dans la conception, il est souhaitable de créer une entité inconnue.

Etudier l'activité scientifique comme un cas particulier de conception permet notamment d'intégrer l'expansion possible des espaces à explorer, indispensable pour notre étude. En revanche, cette approche est trop générale pour décrire le fait qu'un individu produit des solutions et en même temps formule des hypothèses. Dans notre cas, l'individu cherche à la fois le modèle de connaissance qui va permettre d'inférer sur les expériences, mais doit également sélectionner dans l'espace des expériences celles qui ont un facteur commun lié au modèle. Nous proposons de nous baser sur le modèle du constructivisme imaginatif développé par Akin Kazakçi (2013 ; 2014).

3.2. LA NOTION DE TACHE COUPLEE INVENTIVE SUIVANT LE MODELE DU CONSTRUCTIVISME IMAGINATIF

3.2.1. Constructivisme imaginatif comme une recherche à double espace extensible

Kazakci (2014) propose dans la littérature en design un modèle sur deux espaces avec expansion qu'il nomme constructivisme imaginatif. Selon lui, il ne suffit pas d'imaginer de nouvelles solutions, il faut aussi être capable de mettre au point une méthode qui permettrait de construire la solution imaginée. Il s'agit d'un double processus constructiviste, dans lequel la construction de la définition d'objet interagit avec la construction de la méthode qui permettrait de mettre en œuvre cet objet. Cette représentation suppose que l'objet que l'on cherche à construire n'existe pas encore au moment où on cherche à le construire. Dans le constructivisme imaginatif, on construit à la fois le « quoi » de l'objet (on définit l'objet) en même temps que l'on construit le « comment » - un moyen réel de construire effectivement cet objet.

Il existe une multitude de situations dans laquelle les acteurs de la science observent ou définissent de nouveaux objets ou de nouvelles propriétés sans pour autant développer une méthode capable de construire ces objets avec les propriétés recherchées. Il est tout à fait légitime que ces acteurs puissent manipuler ces objets pour en déduire certaines propriétés, sans connaître de méthode pour pouvoir construire ces objets. En fait, de récentes études expérimentales (Edelman, 2011; Eris, 2006) ont montré que dans le processus de conception les acteurs pensent et agissent différemment quand ils pensent *à quoi* l'objet devrait ressembler et à *comment* l'objet peut être construit (Kazakçi, 2014). A titre d'exemple, l'hypothèse de l'existence boson de Higgs,

formulée dans les années 1960, est le résultat d'un processus purement intellectuel pour lequel les scientifiques de l'époque n'avaient pas accès ou connaissance d'une méthode qui permette de vérifier effectivement l'existence de cet objet. Il a en effet fallu attendre plusieurs dizaines d'années pour qu'en 2012, une expérimentation rigoureuse menée par le CERN en 2012 s'assure de la véracité de l'existence du Boson de Higgs.

Dans notre analyse, nous avons vu que la génération d'hypothèses basées sur les données demande de prendre en compte les deux espaces à la fois, afin de formuler l'hypothèse et de trouver le bon algorithme pour analyser les données. En fait, la construction d'une hypothèse sans prendre en compte les bases de données existantes n'aurait pas de sens car elle aurait peu de chances de pouvoir être vérifié. Aussi le modèle du constructivisme imaginatif nous semble adapté à notre étude.

Les fondements du constructivisme imaginatif proviennent de l'étude d'une discipline scientifique particulière : les mathématiques. Kazakçi (2013) se base sur la notion d'intuitionnisme développée par le mathématicien Brouwer, qui est l'une des approches constructivistes majeures de la mathématique. L'intuitionnisme conçoit les mathématiques comme une libre création de l'esprit humain où tous les objets qu'elle manipule doivent pouvoir être construits. Cette représentation du raisonnement se base sur plusieurs propriétés fondamentales. Premièrement, elle explique l'activité mathématique en tant que processus de raisonnement exécuté dans le temps. Deuxièmement, elle met l'accent sur la constructibilité des objets plutôt que sur la vérité de leur existence. Troisièmement, elle reconnaît le caractère incomplet de la connaissance et la possibilité de construire de nouveaux objets. Quatrièmement, la construction d'objets sans précédent et imprévisible est prise en compte par une notion de créativité de la mathématicienne: ses choix libres. De manière plus générale, les mathématiques constructivistes défendent l'argument selon lequel un objet ne peut être défini et exister que s'il peut être construit explicitement par une méthode. Une méthode est un ensemble d'opérations ordonnées qui transforme une entrée en une sortie. Dans notre modélisation, une méthode est une séquence d'actions ordonnées qui permet de passer d'un état initial à un état final. La dynamique constructiviste imaginative permet ainsi de révéler un double constructivisme dans les processus de conception en laissant un libre choix à celui qui construit l'objet.

Dans le modèle de constructivisme imaginatif, Kazakçi suppose que la construction de l'objet se situe dans deux espaces extensibles. Cette propriété n'existe pas dans le modèle de résolution de problèmes où la structure de l'espace explorée est entièrement connue avant de résoudre le problème. Il est selon lui toujours possible de pouvoir intégrer une nouvelle propriété à l'objet qui n'existe pas dans la représentation de l'espace des objets que l'on pourrait avoir à un temps t. Ainsi l'espace des objets se construit en même temps que les objets lui-même. Aussi, les actions mises en œuvre pour construire cet objet ne sont pas limitées à un nombre fini d'action A = $\{a_1, a_2, \ldots, a_n\}$. En fait, selon lui il est toujours possible de proposer une nouvelle action qui n'a pas encore été proposée.

Pour comprendre comment pourrait être mis en place le principe de constructivisme imaginatif, nous l'illustrons par l'exemple du labyrinthe. La situation initiale diffère légèrement du labyrinthe que nous avons présenté pour le modèle traditionnel. L'objectif est toujours de trouver un plan d'action pour sortir du labyrinthe, sauf que nous ne savons pas où est la sortie. De plus, les espaces explorés sont extensibles. Ainsi, il est toujours possible de rajouter une nouvelle case au labyrinthe que nous connaissons. Pareillement, il est toujours possible de trouver une nouvelle sortie autre que celle que nous aurions potentiellement trouvé. Chaque plan d'action construit pour sortir du labyrinthe fournit des informations à la fois sur le chemin suivi et sur l'existence ou non d'une sortie tout au long du chemin. En fait, l'exploration par l'application des actions permet en même temps de statuer sur l'existence et la position des sorties du labyrinthe.

3.2.2. Le modèle de « tâche couplée inventive »

Le modèle de constructivisme imaginatif nous permet de définir une nouvelle tâche pour étudier la génération d'hypothèses basées sur les données. Cette tâche que nous définissons par **« tâche couplée inventive** » se déroule dans un double espace extensible. Elle consiste à construire un état final désiré e_{final} ainsi que le plan d'action $\{a_1, a_2, \dots, a_n\}$ permettant d'atteindre cet état final. Un premier espace est constitué de tous les plans d'actions possibles à partir de l'état initial du monde $e_{initial}$. Contrairement à une tâche couplée classique, la tâche couplée inventive implique le couplage de deux tâches hétérogènes dont la résolution ne peut être réduite à des activités de type élémentaire ou recette et demande un certain niveau d'exploration et d'inventivité de la part de celui qui va résoudre la tâche.

La différence principale avec le modèle du dual-space search est que les espaces sont infiniment extensibles, c'est-à-dire qu'il est toujours possible de créer de nouvelles actions qui ne sont pas encore dans l'espace. Le deuxième espace est constitué de l'ensemble des états possibles du monde. Contrairement à la résolution de problème, les caractéristiques requises de l'état final ne sont pas prédéterminées et la tâche consiste à définir cet état final en même temps que de trouver le plan d'action pour atteindre cet état. L'exploration réalisée dans les deux espaces est donc « couplée ».

Certaines tâches du processus scientifique peuvent être perçues comme une tâche couplée inventive tel que la formulation des hypothèses basées sur les données ou la fabrication des instruments scientifiques. En effet, nous avons vu dans les différents exemples historiques que la fabrication des instruments scientifiques pouvait impliquer l'existence d'une connaissance tacite de la part du fabricant d'instrument que le concepteur de l'instrument ne maîtrise pas nécessairement. Ainsi, la fabrication finale de l'instrument implique un certain nombre de réajustement, voire de modification des propriétés mêmes de l'instrument scientifique final.

Nous pouvons remarquer que la tâche de type résolution de problèmes est un cas particulier de tâche couplée inventive. Cette tâche peut être représentée dans les deux espaces en mêmes temps.

En effet, nous avons vu que le processus de search se réalise dans un espace de plan d'action. L'objectif est de trouver une séquence d'action qui va permettre d'atteindre l'état final désiré. Or, l'état final peut lui aussi être représenté dans un espace. Celui-ci appartient en effet à l'espace contenant l'ensemble des états possibles. Simplement, dans le cas de la résolution de problème, cet état final désiré est considéré comme fixe : il n'est en effet pas possible de changer l'état final durant le processus de résolution de problème.

3.3. LES SITUATIONS SIMILAIRES A LA TACHE COUPLEE INVENTIVE DANS LA LITTERATURE

Le rapprochement entre recherche d'une solution et construction de cette solution a déjà fait l'objet d'étude dans d'autres domaines que le design et que la logique de découverte scientifique. Ici, nous présentons quelques études qui présentent des similarités et peuvent souligner des critères utiles à l'étude de la tâche couplée inventive : une approche managériale avec le principe de *« need-solution pair »* (von Hippel & von Krogh, 2015) ; les méthodes pour déboguer les softwares informatiques (Raymond, 1999) ; la notion de sérendipité.

Récemment, von Hippel et von Krogh (2015) ont proposé de définir théoriquement une nouvelle façon pour les organisations de résoudre des problèmes. Habituellement, la résolution de problèmes dans les organisations se conçoit comme le processus qui va mener, à partir d'un problème formulé, vers une solution. Ce principe est largement hérité des travaux de Simon et la distinction qu'il propose entre un problème bien formulé et un problème mal formulé. Dans leurs travaux, von Hippel et von Krogh s'intéressent à l'existence de situations où la recherche d'une solution se fait alors même que le problème n'existe pas : le problème est en fait élaboré en même temps que sa solution. Cette approche, qu'il définissent suivant le terme de « *need-solution pair »* est particulièrement intéressante pour les organisations car elle permet de limiter le coût de formulation du problème en un problème bien défini, souvent très élevé.

Pour représenter conceptuellement la recherche d'une paire de solution et de problème, ils proposent de représenter deux « champs » ou « landscapes » : un champ représentant les besoins ou les problèmes et un champ comprenant toutes les solutions. Chacun de ces champs présente des creux et des bosses : plus un point dans un des champs est élevé, plus ce point est satisfaisant ou désirable. Ainsi, le point le plus élevé dans le champ des problèmes représente le problème pour lequel une solution serait très importante pour l'entreprise. Au contraire, un point dans un creux représenterait un problème peu désirable. Leur représentation met en avant le fait que se limiter à de la résolution de problèmes reviendrait à chercher à faire correspondre deux points, un dans chacun des champ. Or, cette restriction néglige totalement la richesse des champs et des multiples correspondances possibles entre les deux champs. Limiter la résolution à un unique problème revient alors à ignorer les autres problèmes possibles que l'on peut rencontrer et donc la possibilité de trouver de répondre à d'autres problèmes.

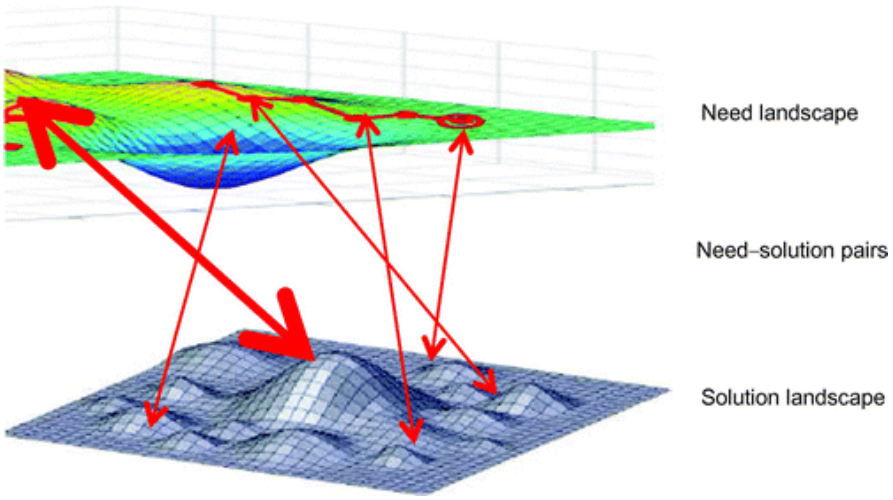


Figure 15. Champs de solutions et de problèmes reliées par des paires de « need-solutions » (issu de von Krogh & von Hippel, 2015)

Le champ de réflexion ouvert par cette notion est encore très récent et ne bénéficie pour l'instant que de peu d'analyses expérimentales. Par ailleurs, les chercheurs n'ont pas encore exploré la structure des champs qu'il propose, notamment la question de leur expansion et de la gestion de cette expansion. En revanche, ils suggèrent de manière théorique de mettre en pratique durant le processus de résolution de problèmes deux étapes pour favoriser l'apparition des « need-solution pairs ». D'abord, ils proposent que lors de la résolution de problèmes, les managers proposent des problèmes les plus larges possibles, afin de favoriser la possible découverte de nouveaux problèmes. Ensuite, ils suggèrent de commencer la résolution de problèmes avec un problème formulé, puis de l'ajuster de manière itérative au fur et à mesure que la résolution de problèmes progresse afin d'augmenter les chances de découvrir des paires de solutions et de problèmes. Dans chaque cycle, un point du paysage de la solution (c'est-à-dire une solution) est testé par rapport au point souhaité sur le paysage des besoins en termes de viabilité. Si l'ajustement n'est pas bon, la formulation du problème est modifiée et la résolution recommence. Le cycle d'essais et d'erreurs se poursuit jusqu'à trouver ou créer un appariement acceptable besoin-solution (Marples 1961 ; Simon & Simon 1962 ; von Hippel et Tyre 1996 ; Thomke 2003 ; Hsieh et al. 2007 ; Nelson 2008). Nous verrons par la suite que ces étapes ont été appliquées dans notre cas expérimental (voir chapitre 8 et 9) mais qu'elles nécessitent néanmoins la mise en place de pratiques managériales pour s'assurer de leur performance.

La notion de tâche couplée inventive peut également se retrouver dans le cas du débogage de logiciels informatiques. Dans le logiciel, découvrir et réparer des erreurs de code ou des bogues peut être très coûteux (Brooks, 1979). Il existe toujours des anomalies dans les codes des logiciels construits, cependant les informaticiens ne savent pas quels sont ces bugs et par conséquence ne peuvent pas anticiper la solution à un problème qui n'aurait pas été identifié par avance. Cependant, la même tâche peut être considérablement réduite en coûts, mais également plus rapide et plus efficace lorsqu'elle est ouverte à une large communauté d'utilisateurs de logiciels

(Raymond, 1999) : «Chaque [utilisateur] aborde la tâche de caractérisation des bogues avec un ensemble de perception légèrement différent et une boîte à outils analytique, avec un angle différent pour le problème. … Donc, ajouter plus de bêta-testeurs… augmente la probabilité que la boîte à outils de quelqu'un corresponde au problème de manière à ce que le bogue soit superficiel pour cette personne » (Raymond 1999, p. 43–44). Dans son analyse, Raymond va dans le sens de notre suggestion sur la délégation des tâches couplées inventives à des foules de participants. En effet, chaque participant va se concentrer sur une partie différente des espaces des hypothèses et des plans d'action pour pouvoir construire à la fois l'hypothèse et l'analyse de données associée.

Enfin, la tâche couplée inventive peut être rapprochée de la notion de sérendipité c'est-à-dire au principe de découverte hasardeuse. La sérendipité est souvent considérée comme un élément fondamental de la logique de découverte scientifique. Merton l'a d'ailleurs défini comme une méthode scientifique à part entière (Merton et Barber 2004). Parfois, un événement fortuit lors du processus scientifique peut déclencher une nouvelle approche d'un problème déjà formulé. En effet lorsque les scientifiques sont suffisamment surpris de manière positive, ils réorientent souvent la recherche de solutions voir restructurent entièrement leur problème (Foster & Ford 2003).

Les recherches sur la sérendipité ont cherché à améliorer la pratique. Certains chercheurs se sont demandés comment on pouvait augmenter artificiellement le nombre de découvertes hasardeuses, mais également la capacité à détecter et reconnaître ces découvertes (van Andel 1994 ; Denrell et al. 2003). Des études ont notamment montré l'importance d'améliorer les compétences des utilisateurs (Beale 2007, p. 421), ainsi que de créer des lieux d'échanges riches et variés (Bathelt et al. 2004) pour accroître le nombre de fois où les phénomènes de sérendipité peuvent être observés. En revanche, la question de la sérendipité pose des questions dans le cas où l'activité est gérée par une foule. En effet, dans ce cas on se retrouve face à une double incertitude : incertitude sur les évènements surprenants et incertitude sur les compétences des participants. La littérature sur la sérendipité n'a pas traité ce problème spécifique.

3.4. LE SEARCH COMME METHODE GENERIQUE POUR CONSTRUIRE DES PLANS D'ACTION

Comment explorer les espaces dans le cas d'une résolution de problème ? Plusieurs méthodes existent dans la littérature pour construire des plans d'action à partir d'une problématique. Une des méthodes les plus connues est le « search ». Lorsque l'espace de l'ensemble des solutions à considérer est énorme, voire potentiellement infini, des procédures de "search" permettent de trouver des solutions aux problèmes.

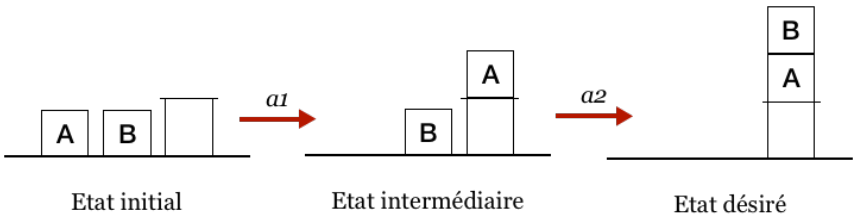


Figure 16. Un diagramme de blockworld. L'état initial est transformé en état désiré par l'application des actions a1 puis a2 (Kazakçi, 2014)

Pour mieux comprendre quels types d'objets sont élaborés par le search, nous illustrons le processus par un exemple bien connu de planification de tâches, le modèle des blockworlds **(figure 16)** (Kazakçi, 2014; Russell & Norvig, 2007). Le blockworld est un monde abstrait et fermé pour expérimenter des systèmes d'IA ou pour l'étude des capacités de raisonnement humain en psychologie cognitive (Cross, 2001). Il consiste en une surface plane, un ensemble de blocs et un robot (ou un bras robotique) capable de déplacer les blocs en appliquant certaines actions (par exemple, *ramasser()* ou *empiler()*). Les états sont des descriptions du monde au moyen de prédicats (par exemple *sur*(A, B) ou *surTable*(A)). À partir d'une configuration initiale (ou d'un état), le programme doit rechercher une combinaison d'actions pour atteindre une configuration finale souhaitée, appelée état désiré. Souvent, il n'y a pas de solution triviale pour la recherche combinatoire et le programme peut rester bloqué (par exemple, dans l'exemple ci-dessus, si B est placé au-dessus de la table avant A), auquel cas il est nécessaire de revenir en arrière. La recherche recommence à partir d'une configuration précédemment explorée et susceptible de mener à une solution. Le programme s'arrête si les ressources de calcul disponibles sont dépassées ou si une séquence d'actions menant à l'état désiré a été trouvée.

Est-ce que ce processus peut servir à étudier la génération des hypothèses ? Dans ce contexte, l'état désiré est interprété comme étant l'hypothèse et le scientifique doit réaliser un certain nombre d'actions (i.e. faire des expérimentations, analyser les bases de données) pour atteindre cet état désiré. Le processus de search crée deux choses. Tout d'abord, un plan d'actions - avec un ensemble d'entrées et de sorties clairement défini - est créé, ce qui correspond à une méthode pour construire l'état final, voire même à une preuve de son existence. La recherche crée un ensemble d'instructions (en tant qu'algorithme) qui transforme l'état d'entrée $e_{entrée}$ en une sortie

e_{sortie}. Dans le cas du processus scientifique, cela correspondrait à l'exécution d'un algorithme pour analyser une base de données disponible. Deuxièmement, l'application du plan d'action permet d'atteindre un état de sortie e_{sortie}. Dans le cas de la génération d'hypothèses basée sur les données, cela correspondrait par exemple à un score de corrélation entre plusieurs variables. L'état de sortie peut donc être différent de l'état désiré. L'hypothèse peut par exemple supposer l'existence d'un lien de causalité entre deux variables tandis que l'analyse de ces variables au travers d'un algorithme ne fournit qu'un score de corrélation.

Peut-on considérer qu'il y a eu création d'une hypothèse ? En fait, la conception de l'hypothèse est déjà terminée avant le début du processus de recherche. Il représente l'état désiré donné au programme en tant que paramètre. Le processus de search ne crée que la méthode qui permette de construire ou de prouver l'hypothèse, ainsi que l'objet de sortie qui est dans le cas de l'analyse de données un score de corrélation entre un ensemble de variables. Il existe potentiellement plusieurs de ces objets de sortie qui correspondent à différentes combinaisons d'actions possibles développés par le processus de search. Ces éléments de sortie ne peuvent être définis comme une hypothèse qu'au moment où on peut définir une hypothèse différente que celle définie au départ, et de toutes celles déjà existantes. Pour cela, il est nécessaire de connaître toutes les hypothèses qui existent déjà et de pouvoir comparer la nouvelle hypothèse à celle potentiellement créée.

À moins que le processus de search ne puisse modifier la définition des types d'objets en cours d'exécution, il n'est pas possible d'imaginer ces nouvelles hypothèses, et donc de nouveaux états désirés. Dans l'exemple du blockword, la recherche s'applique uniquement à une dimension de la définition de l'objet, réduisant ainsi le processus à une recherche mono-espace. En fait, le processus de search correspond soit à la description de l'objet d'intérêt soit à la construction de la méthode pour le construire, mais pas les deux en même temps. Une autre difficulté de ce type de processus de recherche est son incapacité à créer de nouveaux objectifs. Cela suppose non pas la capacité de formuler des états désirés, mais également des états désirés qui ne peuvent être atteints par l'ensemble des actions en cours. Dans un tel cas, de nouvelles actions doivent être apprises par le programme. Il doit donc être capable de créer de nouvelles hypothèses sans précédent tout en acquérant de nouveaux types d'action.

La méthode de search n'est donc pas adaptée pour explorer le double espace dans le cadre de la génération des hypothèses, et de manière plus générale pour les tâches couplées inventives. Pour modéliser le processus de génération des hypothèses il nous faut pouvoir prendre en compte les cas où on est amené à reformuler un problème.

3.5. STRUCTURE ET NATURE DES ESPACES DANS LES TACHES COUPLEES INVENTIVES

3.5.1. De la résolution de problèmes à la formulation de problème

Nous avons vu que la résolution de problème et la méthode de search ne tiennent pas compte des cas où on pourra être amené à reformuler un problème. Ils se contentent de chercher un plan d'action étant donné un problème déjà défini. Pourtant, en partant du modèle de base, il est possible de définir la dynamique qui mène à la formulation d'un problème. Nous définissons ainsi la formulation de problème comme la génération d'un état final désiré e_{final} étant donné E, l'ensemble comprenant tous les états. C'est donc une opération de $H : E \rightarrow F$ où l'état désiré final est inclus dans l'espace *F*. Il existe deux cas de figure:

- L'état généré comme objectif final est déjà contenu dans l'ensemble d'états *E* de départ. Dans ce cas, l'espace *F* se confond avec l'espace *E* (*F* = *E*).
- L'état généré est une extension de l'ensemble d'états *E* de départ, c'est-à-dire que e_{final} n'appartient à l'espace des états connus E. Dans ce cas $F = \{e_{final}\} \cup E$.

L'espace d'état *E* est donc partiel et extensible. D'abord partiel car on ne connaît pas tous les états, en revanche on peut les découvrir. Ensuite extensible car on peut toujours rajouter un état. Supposons par exemple dans le cas du blockworld que l'espace d'état n'est constitué que de l'état initial $e_{0,0}$, l'état intermédiaire $e_{A,0}$ et l'état final $e_{A,B}$ que nous avons illustrés sur la figure. Nous avions supposés qu'il était possible que l'on soit dans la situation où le bloc A est placé au-dessus du bloc B. Dans ce cas l'espace des états $E = \{e_{0,0}, e_{A,0}, e_{A,B}\}$ peut être agrémenté d'un nouvel état $e_{A,\acute{B}}$ non encore décrit tel que $E' = \{e_{A,\acute{B}}\} \cup E$. Notons que le terme $\acute{B}$ définit le bloc B tel que celui –ci a subi une rotation de 45° vers la droite.

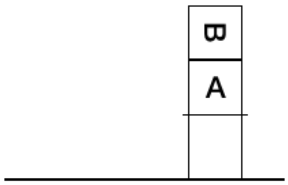


Figure 17. Nouvel état dans le blockworld. Le bloc B est tourné à 45° vers la droite.

Nous définissons également la fonction de valeur *v()* qui pour tout état dans l'espace *E* associe une valeur. Cette valeur peut correspondre à la désirabilité d'atteindre un état plutôt qu'un autre. Dans le cas du blockworld, l'état final désiré $e_{A,B}$ a plus de valeur que l'état initial $e_{0,0}$ ou l'état intermédiaire $e_{A,B}$ car il est plus désirable. La valeur $v(e_{A,\acute{B}})$ du nouvel état n'est pas connue ou prise en compte avant son existence. Cela montre que la fonction de valeur *v()* est partielle et est

définie seulement dans *E*, au moment de la recherche. Lorsqu'on va transiter de *E* vers *F*, il va falloir aussi étendre *v()* dans *F* pour estimer la valeur du nouvel états introduit. Nous reviendrons sur ce sujet dans l'analyse de nos terrains.

3.5.2. Structure de l'espace des hypothèses

Dans l'espace des hypothèses, une hypothèse est structurée comme une phrase, c'est-à-dire que l'ensemble des mots mis bout à bout doit avoir un sens, un contenu transmis par le message. Plusieurs caractéristiques la distinguent pourtant d'une phrase quelconque. Premièrement, le message transmis n'a pas de statut logique, c'est-à-dire qu'on ne peut pas statuer si l'hypothèse a valeur de vérité à partir de l'ensemble des savoirs existants au moment de la formulation. Par exemple, la phrase « Il pleut » n'est pas une hypothèse car le sens est uniquement descriptif et la proposition a déjà un statut logique connu (je sais qu'il pleut à ce moment précis) contrairement à la phrase « il va pleuvoir demain ». Deuxièmement, pour être d'un intérêt scientifique, on doit pouvoir déduire de l'hypothèse un certain nombre de conséquences vraies ou fausses. Lorsque les belges François Eglert et Robert Brout, et l'Ecossais Peter Higgs formulent en 1964 l'existence d'une particule appelée boson de Higgs, cette hypothèse a pour objectif de proposer une explication au fait que certaines particules dans l'univers ont une masse et d'autres non. La confirmation ou l'infirmation de l'hypothèse a des conséquences sur les savoirs scientifiques existants, notamment dans ce cas le fait d'apporter des preuves solides de la validité du modèle standard sur les particules. Enfin si le statut logique de l'hypothèse n'est pas connu au moment de sa formulation, il doit être vérifiable par la suite. L'existence de l'objet construit par le scientifique est par définition observable mais non encore observé.

Le langage utilisé dans l'espace des hypothèses a une ontologie propre à chaque domaine scientifique qui construit au fur et à mesure de ses découvertes des mots, synthétisant les familles concepts scientifiques ou objets qui lui sont sous-jacent. Par exemple, des astronomes vont plutôt parler de « planètes » ou « d'étoiles » tandis que les spécialistes en physique des particules s'intéresseront aux « quarks », aux « photons » ou aux « neutrinos ». De plus, le scientifique peut inventer un nouvel objet si les concepts déjà existants ne suffisent pas pour décrire le phénomène d'intérêt, ce qui arrive régulièrement dans la science : ADN, neutrino, électron, cellule pour n'en citer que quelques-uns. Ces objets *D* sont ensuite combinés grâce à des mots servant à relier les concepts et former le sens de la phrase. Dans l'hypothèse « La pomme est sur la table », les mots « pomme » et « table » sont les objets définis pour caractériser ce qui est observé dans le monde sensible. Les mots « est » et « sur » quant à eux créent la relation entre les termes « pomme » et « table », et formulent ainsi la supposition, en plus des éléments de langage essentiels à la formulation d'une phrase (les articles « le » et « la » dans notre cas). Les relations *R* font partie des connecteurs logiques souvent utilisés dans le domaine de la linguistique et de l'informatique. Il en existe une très grande quantité que l'on regroupe la plupart du temps suivant leurs rôle : additif (et, de plus, puis), causal (car, pour, étant donné que), comparaison (comme, identique à, de même que, semblable à) pour n'en citer que quelques-uns. Nous pouvons ainsi décrire la

structure d'une hypothèse à partir de deux catégories de mots : les familles de concepts scientifiques ou objets *D* et les relations *R* entre ces objets. L'espace des hypothèses est structuré tel que chaque hypothèse puisse être écrite suivant le modèle :

$$\mathcal{L}_{\mathcal{H}} = \textbf{relations } \boldsymbol{R}(\textbf{objets } \boldsymbol{D} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n}\})$$

Il faut noter que la notion de famille de concepts est ambiguë. En effet, une famille de concepts peut comporter plusieurs autres familles de concepts. Par exemple, le concept de « particule » en physique contient la famille de concept « quark » qui elle-même contient les familles « quark up », « quark down », « quark bottom » ou encore « quark charm ». Cette relation d'inclusion n'est pas suffisante pour différencier les hypothèses entre elles. En effet, un physicien peut très bien formuler une hypothèse sur les quarks en général, de même qu'il pourra avoir une hypothèse spécifique aux quarks ups. La précision nécessaire à l'existence d'une hypothèse dépend du contexte scientifique.

3.5.3. Extension et recherche dans l'espace d'action, suite à la formulation d'un nouvel état final

En même temps que de nouveaux problèmes sont formulés, le processus de tâche couplée inventive suppose la construction d'un plan d'actions permettant d'atteindre l'état désiré. Or, chercher à résoudre un nouveau problème peut ne pas toujours être possible à partir des actions connues. Dans ce cas, une partie importante de la recherche serait de pouvoir générer de nouvelles actions a_j qui n'appartiennent pas à *A*, l'ensemble d'actions connu au départ. L'espace d'état *A* est donc également partiel et extensible comme l'espace *E* des actions.

Reprenons le cas du blockworld avec comme état désiré le nouvel état $e_{A,\acute{B}}$. Dans cet état, le bloc B est posé sur le bloc A et le bloc B est tourné de 45° vers la droite par rapport à sa position initiale. Les seules actions possibles que le robot peut exécuter sont de *ramasser* un bloc et *empiler* ce bloc sur un autre bloc. Or, aucune de ces actions ne permet d'avoir un bloc tourné à 45° vers la droite. Il est donc nécessaire d'intégrer une nouvelle action. Il y a donc extension de l'espace des actions qui inclut la nouvelle action nécessaire pour atteindre l'état désiré. A noter qu'il est possible de créer de nouveaux états en intégrant cette action dans le plan d'action.

3.5.4. Structure de l'espace des plans d'action

L'espace des plans d'action possède une structure similaire à l'espace des hypothèses. De la même manière que l'hypothèse, un plan d'actions ou une séquence d'actions est une combinaison d'actions définies par le scientifique suivant un certain ordre d'exécution. Pour comprendre la nature d'un plan d'action, il faut imaginer la forme un commis qui cherche à faire une omelette. Ce dernier choisit la recette qu'il applique puis ensuite exécute cette dernière : « cassez les œufs », puis « les battre afin de mélanger le jaune et le blanc », « faire chauffer une poêle à feu moyen », « mettre le mélange dans la poêle »… Chacune des étapes représente une action que le commis va

mener. Toutes ces actions appartiennent à un espace structuré comme un espace « descriptif », et se génèrent à travers un langage $\mathcal{L}_{\mathcal{A}}$ propre à la discipline. Les deux langages $\mathcal{L}_{\mathcal{A}} + \mathcal{L}_{\mathcal{H}}$constituent ainsi le vocabulaire propre à une discipline scientifique.

Les termes employés correspondent au champ lexical propre au domaine. Dans la recette de l'omelette, on distinguera ainsi deux champs lexicaux distincts. D'abord les objets sur lesquels vont être appliqués la transformation : ici ce sont les mots « œuf », « poêle », « blanc » et « jaune ». Tous ces termes représentent des objets appartenant au monde réel et caractérisés par leur état. Ces objets appartiennent tous à une catégorie d'objets semblable. Quand on parle d'un œuf dans la recette, on ne précise pas sur quel œuf on va agir et on suppose que l'action menée transformera tous les œufs d'une façon similaire. Mais on peut également parler d'un objet précis en particulier : par exemple un algorithme construit pour une base de données précise suivra un plan d'actions spécifique à ces données et non nécessairement applicable à n'importe quelle données semblables. On ne parle donc pas ici de la catégorie d'objets *D* mais de l'objet précis *D(x)*, où x est un des objets appartenant à une catégorie plus générale. Ensuite les verbes correspondent à l'action A à mener : « casser », « faire chauffer », « battre ». Nous pouvons ainsi décrire la structure de l'espace des plans d'action à partir de deux catégories de mots : les objets scientifiques et les actions à mener sur ces objets.

$$\mathcal{L}_{\mathcal{A}} = \boldsymbol{actions\ A}(\boldsymbol{objets\ scientifiques\ D} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n}\})$$

Notons que l'ordre dans lequel les actions vont être exécutées importe beaucoup. Dans notre exemple de l'omelette, le fait de mettre les œufs dans la poêle sans les avoir cassé au préalable aura un effet différent et ne donnera probablement pas d'omelette à la fin !

4. PERFORMANCE ET CAPITALISATION DANS LES TACHES

4.1. PERFORMANCE DANS LES TACHES COUPLEES INVENTIVES

La question de la performance de l'exploration de l'espace des hypothèses n'est pas abordée par les modèles que nous avons présentés. De la même manière, Kazakçi n'a pas pris en compte cette notion de performance dans son modèle de constructivisme imaginatif (Kazakçi, 2013; 2014). Pourtant, tout processus de search dans un espace peut être associé à des questions de coûts d'exploration et de coût de la solution et donc de l'efficacité du processus. Que devient alors la question de la performance dans le cas d'une tâche ? Nous avons montré au début de ce chapitre que ces questions de coût, de qualité et de délai étaient présentes dans le cadre de l'exploration de l'espace des plans d'action. De manière similaire, l'exploration dans l'espace des hypothèses devrait être associé à une deuxième logique de performance.

Toute tâche couplée inventive peut donc être associée à une double logique de performance dans chacun des espaces explorés. D'un côté, il y a les critères de performance classiques que nous pouvons retrouver dans toute exploration d'un espace de plan d'action. Nous avons vu dans le cas de la résolution de problèmes que l'exploration de cet espace peut être associée à des coûts d'exploration ainsi que des coûts d'exécution. Nous pouvons également intégrer les notions de qualité pour définir la qualité de l'exécution du plan d'action, mais également la qualité de la solution obtenue évaluée par la fonction *test*(). Dans le cas de la tâche couplée inventive, le test est réalisé vis-à-vis d'un état final désiré construit pendant la tâche. Cependant, contrairement à la résolution de problème, l'état final désiré n'est pas fixe et peut évoluer durant la tâche. Enfin, la notion de délai reste identique au cas de la résolution de problème.

De l'autre côté, qu'en est-il pour l'exploration de l'espace des hypothèses ou des états ? Similairement à l'espace des plans d'action, la recherche d'un état engendre des coûts en terme d'exploration. En effet, les exemples d'automatisation de la génération d'hypothèses montrent que les coûts relatifs à l'exploration de l'espace des hypothèses dépendent du nombre d'étapes à suivre pour sélectionner ou construire un nouvel état désiré. Lorsque ces étapes sont réalisées par un ordinateur, il est alors facile de compter le nombre d'étapes suivies par celui-ci pour atteindre un nouvel état désiré. Ensuite, la notion de qualité est plus difficile à évaluer car elle ne repose pas sur une fonction *test()* existante avant la tâche. En fait, la qualité peut faire intervenir un ensemble de critères *ad hoc* qui peuvent être mis en place pour la tâche : par exemple la capacité à trouver une séquence d'actions pour construire l'état final désiré, le nombre d'actions minimum pour passer d'un état existant à l'état final désiré, ou combien de parties de l'espace des états ont été exploré. Remarquons qu'une partie des critères de qualité présentés se basent directement sur l'exploration dans l'espace des plans d'action. Enfin la notion de délai peut facilement être interprété dans l'espace des états comme étant le temps passé pour construire un nouvel état désiré.

Nous proposons de construire des critères de performance pour évaluer l'exploration dans le cas de la génération des hypothèses basées sur les données. Ces critères seront utiles pour analyser le cas d'étude que nous avons présenté dans le chapitre 3 et sur lequel nous reviendrons dans les chapitres 8 et 9. Nous définissons des métriques pour l'exploration dans l'espace des hypothèses et dans l'espace des plans d'action :

Espace des hypothèses	• Nombre d'hypothèses produites • Total de données explorées sur l'ensemble des données disponibles • Cohérence avec la littérature • Nombre d'hypothèses vérifiées par des méthodes statistiques
Espace des plans d'action	• Robustesse de la méthode de vérification de l'hypothèse

4.2. TACHE REPETEE ET CAPITALISATION

Nous avons vu au début du chapitre que la notion de performance pour les tâches élémentaires et de type recette implique de considérer la répétition de la tâche. De manière générale, la question de la répétition de la tâche génère de nouvelles problématiques de performance qui n'existent pas lorsque la tâche est réalisée seule. La perte de production par exemple est un problème courant dans l'organisation scientifique que l'on retrouve dans tous les types de tâches. Un laborantin chimiste est susceptible de faire de mauvaises manipulations et d'obtenir des produits chimiques non désirables ; un data scientist qui doit résoudre un problème produira potentiellement des lignes de code inutiles avant d'arriver à la solution ; un scientifique va tester plusieurs hypothèses avant d'en trouver une qui soit intéressante et vérifiable. Lorsque les tâches ne sont pas répétées, les pertes de production sont faibles ou difficilement maîtrisables. Or, dès que la tâche est répétée, le volume de production potentiellement perdu peut devenir important et nécessite donc de mettre en place des modèles de gestion. De manière générale, il est possible d'intégrer de nouveaux critères de performance liés à la répétition que l'on étudie suivant deux angles : la performance d'une tâche à l'autre et la performance sur l'ensemble des tâches. Ces critères de performance vont nous permettre d'étudier les problématiques que nous avions identifiées dans la partie 1 concernant la performance des projets de science citoyenne. Nous reviendrons plus en détail sur le lien entre performance et science citoyenne dans le chapitre 6.

4.2.1. Capitalisation entre les tâches répétées

Dans les tâches élémentaires et de type recette, il n'y a pas d'incertitude dans l'exécution de la tâche : la performance ne se mesure qu'à partir de l'état final désiré, du coût et du délai de l'exécution. Dès lors qu'il y a répétition de la tâche, on peut s'intéresser aux variations d'exécution entre les tâches, et donc évaluer la performance dans la répétition. Prenons l'exemple que nous avions cité dans le chapitre précédent autour de l'équation personnelle de l'astronome. Dans son expérimentation, Peirce avait montré que le jeune assistant qu'il avait embauché pour répéter plusieurs fois la mesure de la position de l'astre avait grandement amélioré son temps moyen de détection entre le premier et le dernier jour. Cet effet d'amélioration n'avait pas été observé par les astronomes du laboratoire de Greenwich. En effet dans leur cas, chaque mesure de l'équation personnelle était indépendante car réalisée par un seul acteur. Il y a un processus de capitalisation de l'assistant : entre chacune des tâches, celui-ci interroge et interprète la manière qu'il a d'exécuter sa mesure. A chaque itération, il modifie légèrement son action, améliore certains aspects dans l'exécution des tâches afin de se rapprocher au mieux de l'objectif optimal, à savoir mesurer le passage de l'astre sans qu'il n'y ait de décalage de mesure. Cette capitalisation a plusieurs effets sur la performance de la répétition des tâches. D'abord elle améliore la qualité de la production entre chaque tâche. Ensuite, elle réduit les pertes, c'est-à-dire qu'elle tire parti de ce qui a déjà été fait et évite de faire les mêmes erreurs deux fois.

La **capitalisation** peut être définie comme une accumulation de tout ce qui peut être compris de ses expériences passées (ici les tâches déjà exécutées) afin d'en tirer profit pour les tâches futures. A noter que nous distinguons la notion d'apprentissage de la notion de capitalisation. En effet il nous semble plus judicieux de parler de capitalisation plutôt que d'apprentissage dans un contexte de performance. D'abord, l'apprentissage est difficilement mesurable. Comment être certain qu'un certain élément a bien été appris par une personne ? Cela nécessite de pouvoir évaluer le raisonnement cognitif mis en œuvre par celui-ci et donc de pouvoir valider par un quelconque moyen l'évolution de ce raisonnement. Au contraire notre approche se considère comportementaliste dans le sens où on s'intéresse plutôt au comportement extérieur de l'individu qu'à son raisonnement cognitif. De plus, l'apprentissage est un processus personnel et donc dépendant du sujet qui exécute la tâche. Une démission, un changement d'équipe, implique alors une perte nette de tout ce qui a été appris auparavant. Au contraire, la capitalisation peut être étudiée indépendamment de l'humain et la capitalisation devient indépendante des acteurs.

La notion de capitalisation a également un effet sur la performance dans les tâches couplées inventives (dont la résolution de problèmes) (Sitruk & Kazakçi, 2019). Dans ces tâches, les individus explorent des espaces (espace des plans d'action, espace des états) pour trouver des séquences d'action et des états désirés. A la fin d'une tâche, certaines zones de l'espace auront été explorées tandis que d'autres restent des zones inconnues. Lorsque la tâche est répétée, la connaissance des zones de l'espace déjà explorées permet de limiter la redondance de l'exploration (explorer plusieurs fois la même zone de l'espace) et donc d'augmenter les chances de trouver un état désiré intéressant. En effet, le risque de ne pas générer d'états finaux intéressants est potentiellement élevé dans les tâches couplée inventives. Dans le cas de la génération d'hypothèses basées sur les données par exemple, l'espace descriptif du langage est défini par les bases de données accumulées sur la problématique. Imaginons le cas d'une base de données élémentaire constituée de 3 variables $\{v_1, v_2, v_3\}$ et une seule relation R possible entre ces caractéristiques. Dans ce cas, il est possible de formuler $\binom{3}{2} + \binom{3}{3} = 4$ combinaisons possibles et donc autant d'hypothèses à tester. Ce nombre augmente exponentiellement à mesure que le nombre de variables augmente. Pour une base de données de 10 variables et une seule relation R, le nombre d'hypothèses formulables est de $\sum_{k=2}^{10}\binom{10}{k} = 2^{10} - 1 - 10 = 1013$, et de l'ordre de 10^{301} pour une base de données comprenant 1 000 variables différentes. Lorsque le scientifique est amené à manipuler des données massives, le nombre de variables peut atteindre facilement plusieurs milliers voire plusieurs millions d'éléments à combiner : par exemple si la base de données est constituée d'images de qualité standard 640x480 pixels, le nombre de variables est égal au nombre de pixels soit 307 200. Pire, l'espace des données étant extensible, il est toujours possible d'ajouter une nouvelle caractéristique à la base de données déjà existante. La probabilité de formuler une hypothèse intéressante est donc très faible dans le cas où le processus de génération est basé sur un processus aléatoire et sur des bases de données massives. Nous reviendrons sur ce point dans notre étude du programme Epidemium.

4.2.2. Performance sur l'ensemble des tâches répétées

La question de la capitalisation existe également durant la tâche. Par exemple, lorsque l'on cherche à étudier si un dé est pipé, il faut le lancer un nombre suffisant de fois pour avoir une estimation statistique satisfaisante. Ainsi la performance sera associée non pas seulement au lancer unique, mais à la répétition du lancer. Chaque lancer est considéré comme indépendant des autres : il n'y a donc pas de capitalisation entre chaque lancer pour modifier l'action du lancer. C'est l'analyse statistique de l'ensemble des lancers qui peut être considérée comme une capitalisation. Au lieu d'évaluer la production de chaque tâche individuellement, la capitalisation est agrégative : elle propose un résultat basé sur une analyse de l'ensemble des lancers. Nous pouvons retrouver ce principe de **capitalisation par agrégation** dans les tâches élémentaires et de type recette. Chaque tâche réalisée de façon indépendante peut être étudiée par des outils statistiques pour étudier la performance globale. Ce système de capitalisation est classiquement mis en œuvre dans les usines pour évaluer le niveau de production d'une chaîne de fabrication par exemple. Qu'en est-il de la capitalisation durant la tâche dans le cas de la résolution de problèmes et de la tâche couplée inventive ? Prenons l'exemple d'un robot qui explore un labyrinthe pour trouver la sortie. La capitalisation représente la mémoire du robot. Le robot va proposer une première séquence d'action pour passer de l'état initial à la sortie. Si la séquence n'est pas une solution au problème, alors il en teste une autre. Lorsque le robot doit choisir une nouvelle séquence d'action, il utilise la mémoire de tout ce qu'il a déjà exploré. S'il n'avait pas de mémoire, l'exploration recommencerait de zéro à chaque fois et le robot augmenterait le risque de répéter les mêmes actions qui ne mènent pas à la solution. Nous reviendrons plus en détail sur ce point dans le chapitre suivant.

CHAPITRE 6 – GESTION DE LA PRODUCTIVITE D'UNE FOULE : PERFORMANCE ET RISQUE DE PERTES DURANT ET ENTRE LES TACHES

RESUME DU CHAPITRE 6

Dans ce chapitre, nous utilisons le modèle formel de tâches développé dans le chapitre 5 pour étudier la performance de l'ouverture de ces tâches dans le cadre des projets de science citoyenne. Les projets de science citoyenne sont organisés sous une forme seeker-solver, où les seekers (les organisateurs) conçoivent et formulent la tâche et les solvers (les participants ou les citoyens de la science) exécutent ou résolvent la tâche. Nous montrons que l'intérêt majeur des projets de science citoyenne est l'externalisation d'une partie des coûts d'exécution des tâches aux citoyens de la science, réduisant fortement le coût global de la tâche. De plus, les tâches peuvent être parallélisées et donc diviser le temps d'exécution de tâches similaires par le nombre de participants.

En revanche, l'ouverture à la foule des tâches réduit la fiabilité du processus d'exécution de la tâche : en effet, les organisateurs n'ont pas moyen de connaître à l'avance les compétences des participants ni leur propension à fournir un effort suffisant pour résoudre la tâche. Pour pallier à cette baisse de fiabilité, les organisateurs peuvent mettre en place un système de redondance des tâches (répéter plusieurs fois la même tâche). Cette redondance des tâches peut mener cependant à des pertes importantes de productivité, pouvant parfois atteindre jusqu'à 100% de perte (par exemple dans les projets compétitifs où seule la meilleure solution est conservée).

Nous proposons de gérer ces pertes par l'introduction d'un processus de capitalisation. Dans le cas des tâches de type élémentaire ou recette, la capitalisation peut se faire par agrégation statistique, dont son fonctionnement est illustré avec le projet Galaxy Zoo. Dans le cas des tâches exploratoires comme la résolution de problèmes ou les tâches couplées inventives, les pertes peuvent être gérées par deux types de capitalisation : la capitalisation croisée, où les participants capitalisent durant la tâche sur ce que les autres participants ont produit. La capitalisation séquentielle : dans un premier temps, les participants explorent les espaces en soumettant plusieurs fois des solutions. A chaque solution soumise, ils améliorent leur connaissance de l'espace et peuvent capitaliser sur cette connaissance pour soumettre des solutions de meilleure qualité. Une fois la tâche terminée, il est possible de cartographier l'ensemble de l'exploration réalisée par chacun des participants. Cette cartographie permet de capitaliser sur ce qui a été produit durant la tâche et de le réutiliser dans la tâche successive.

Nous avons montré dans le chapitre 5 que les tâches scientifiques ouvertes à la foule peuvent être interprétées suivant un modèle à quatre niveaux : la tâche élémentaire, la recette, la résolution de problèmes et la tâche couplée inventive. Nous avons suggéré que la notion de performance dans ce modèle relève à la fois des questions classiques de coût-qualité-délai relatives à l'exécution des tâches ainsi qu'à l'exploration des espaces, mais intègre également d'autres critères dès lors que la tache est répétée. Comment s'organise dans ce modèle la répartition du travail entre les différents acteurs des projets de science citoyenne ? Pour répondre à cette question, nous avons besoin de modéliser les acteurs de la tâche.

Les projets de science citoyenne font intervenir essentiellement deux types d'acteurs dans le processus, un seeker et un solver (Sieg et al., 2010). D'un côté il y a le seeker ou « organisateur » O dont le rôle est de concevoir, déléguer et piloter l'exécution de la tâche. C'est lui qui est en charge de gérer les critères de performance. De l'autre côté il y a les solvers ou « participants », qui vont exécuter la tâche. Nous définissons l'espace $P(n) = \{P_1, P_2, \dots, P_n\}$ comme l'ensemble des participants avec n le nombre de participants. Toute personne est définie comme participant à la tâche dès lors qu'il soumet une proposition ou qu'il fait partie d'une équipe qui a soumis une proposition. Dans le contexte des projets de science citoyenne, P est un espace extensible. En effet les projets de science citoyenne ne limitent pas généralement le nombre participant et il peut toujours y avoir un nouveau participant. Le nombre n n'est donc généralement pas connu à l'avance par le seeker. Dans l'exemple du labyrinthe, le seeker est celui qui va définir le contexte (le labyrinthe), fournir les règles de déplacement (se déplacer sur les cases adjacentes et non en diagonale), définir l'objectif de la tâche (trouver la sortie), mais également définir un budget et des contraintes de réalisation. Les solvers soumettent des plans d'action au seeker, c'est-à-dire des chemins dans le labyrinthe.

Notre travail s'inscrit dans les recherches menées sur les modes de coopération entre concepteurs dans le cadre de partenariats d'exploration qui utilisent la théorie de la conception comme cadre d'analyse. Les chercheurs ont exploré différentes situations expérimentales de collaboration entre institutions dans le cas où l'incertitude est très élevée. Une forme de coopération consiste à identifier au sein de chaque partenaire des zones d'intérêts communes d'exploration : à partir d'une approche dite matching-building (Gillier et al., 2010), les concepteurs évoluent par étape afin de converger autour d'un concept sur lesquel ils trouvent un intérêt commun. Une deuxième forme de coopération concerne les situations où différents concepteurs doivent collaborer dans des projets à très forte incertitude (inconnu marché et inconnu technologique). La stratégie proposée consiste à développer un « inconnu commun » pour gérer le risque (Kokshagina et al., 2012). Cet inconnu commun est défini comme le point de rencontre des connaissances désirables par l'ensemble des participants ou concepteurs : dès lors que cet inconnu existe, chacun d'entre eux peut générer à partir une valeur propre grâce à ses connaissances et ses expertises spécifiques. Une troisième forme de coopération consiste à faire interagir des acteurs dont les domaines d'expertises sont très éloignés. En forçant la collaboration par interactions répétées, les concepteurs peuvent entrevoir des concepts ou des idées, pourtant banales dans le domaine

initial, et qui apportent un regard neuf et désirable dans leur champ d'expertise (Salgueiredo & Hatchuel, 2016).

Dans ces exemples, l'attention est portée sur la qualité de la coopération entre acteurs, et les chercheurs ne s'intéressent pas à l'efficacité de la coopération et à la productivité. Or, de par la structure des projets de science citoyenne sous forme de seeker-solver et de challenges, la question de la performance est omniprésente : la question n'est pas tant de s'intéresser à la qualité de la coopération mais plutôt de l'impact de cette coopération sur la qualité de la production. Par ailleurs, la variété des expertises et la non connaissance *ex ante* de celles-ci par les seekers ne permettent pas de mettre en place les stratégies proposées présentées et demande de nouvelles investigations.

Dans ce chapitre, nous étudions la question de la performance dans ce modèle classique seeker-solver pour les projets de science citoyenne. Nous présentons dans la première section la gestion de la performance durant le projet. Dans la deuxième section, nous nous intéressons à la gestion des pertes entre les projets.

1. PERFORMANCE DURANT LES PROJETS DE SCIENCE CITOYENNE : CAPITALISATION PAR AGREGATION ET CAPITALISATION CROISEE

1.1. SPECIFICITE DES PROJETS DE SCIENCE CITOYENNE : DELEGATION DES COUTS VS FIABILITE DU SYSTEME

Dans le modèle seeker-solver, la performance de la tâche est gérée par l'organisateur (le seeker). C'est lui qui conçoit la tâche, établit le cahier des charges à respecter dans la réalisation de cette tâche et évalue à la fin l'ensemble du processus. Nous avons vu que les critères de performance utilisés peuvent être ramenées généralement à un triptyque qualité-coût-délai. Le coût de la tâche peut être représenté comme la somme d'un coût de conception et un coût d'exécution (soit $coût = coût_{conception} + coût_{exécution}$). Le coût de conception comprend l'ensemble des éléments nécessaires à la tâche et à son exécution. Ce sont notamment les ressources humaines nécessaires à la conception de la tâche, les outils associés à l'exécution de la tâche (plateforme virtuelle pour les challenges, instrument de mesure scientifique), la communication pour rendre visible la tâche et intégrer les citoyens de la science, le coût de transfert des solutions des participants vers les organisateurs. Le coût d'exécution lui comprend l'ensemble des moyens nécessaires à l'exécution de la tâche. Ce sont notamment le temps de calcul alloué à un serveur ou un ordinateur, les ressources humaines nécessaires pour exécuter la tâche ainsi que toute forme de récompense associée à l'exécution de la tâche.

Nous avons vu dans la partie 1 qu'un des avantages principaux des projets de science citoyenne est l'externalisation d'une partie des coûts d'exécution de la tâche. En effet, l'ouverture de la tâche permet au moins de déléguer les coûts relatifs aux ressources humaines nécessaires pour exécuter

la tâche. Dans le cas de Galaxy Zoo par exemple, la délégation a permis aux organisateurs d'économiser l'équivalent de 83 années de salaires pour une personne travaillant à temps plein sur le projet (Franzoni & Sauermann, 2014) - soit l'équivalent de 1.5 millions € pour une personne payée au smic. Il faut bien sûr ôter à ce chiffre les surcoûts de conception comme la création de la plateforme virtuelle, toutefois les gains en terme de coûts restent importants. D'autres projets exploitent ce gain pour réduire le temps de calcul de leurs serveurs en externalisant le calcul sur les ordinateurs individuels des citoyens de la science (voir les projets *LHC@home* et *Foldit@home*).

Un autre avantage des projets de science citoyenne est de pouvoir réduire le temps d'exécution d'un ensemble de tâches similaires. En multipliant le nombre de composants du système (les participants), l'ouverture des tâches permet d'exécuter ou de résoudre plusieurs tâches en parallèle. Le temps d'exécution de toutes les tâches est donc globalement divisé par n, n étant le nombre de participants au projet. En reprenant l'exemple de Galaxy Zoo, les participants ont codé les 900 000 images de galaxies en quelques mois, au lieu de 83 ans pour un scientifique seul.

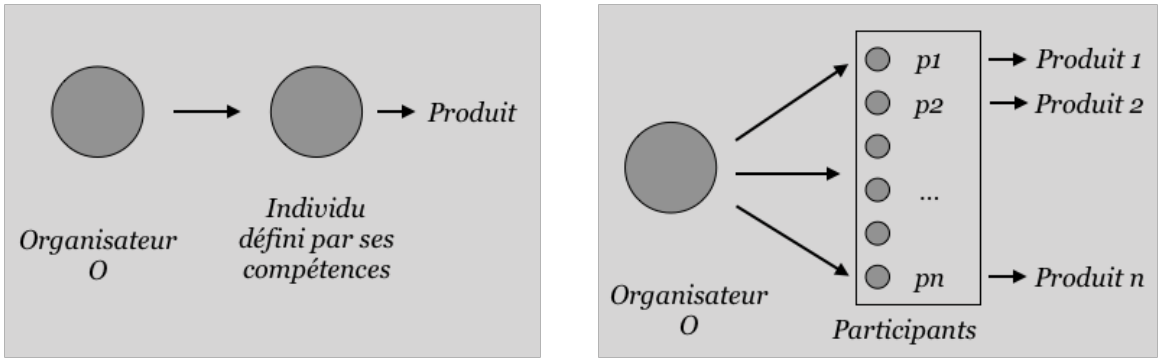


Figure 18. Système 1 (à gauche) : tous les composants du système sont maîtrisés, assurant sa fiabilité. Système 2 (à droite) : pas de maîtrise des caractéristiques des composants du système (les participants).

Si la délégation de la tâche à la foule réduit les coûts et les délais en terme d'exécution de la tâche, elle réduit en même temps la maîtrise de la qualité de ce qui est produit. Contrairement à une organisation classique où la tâche est déléguée en interne à un individu dont les compétences sont reconnues, ici les organisateurs ne connaissent ni les compétences des participants, ni leur motivation à réaliser la tâche. On peut se représenter l'organisation comme un système avec les caractéristiques de la tâche à l'entrée et le produit final de la tâche à la sortie. Chaque acteur peut être représenté comme un composant de ce système dont la fiabilité a des conséquences sur la performance globale du système (**figures 17 et 18**). On passe d'un premier système 1 fiable (mais plus coûteux et plus long) où tous les composants sont maîtrisés à un système 2 moins coûteux mais moins fiable où une partie des caractéristiques des composants du système ne peut être maîtrisés.

1.2. AMELIORER LA FIABILITE PAR REDONDANCE

Pour améliorer la fiabilité d'un système de production, plusieurs stratégies peuvent être mises en œuvre par les organisateurs : mise en place d'une redondance des éléments du système, amélioration de la fiabilité des sous-éléments qui constituent le système (mais potentiellement avec des coûts plus élevés), nouvelle conception du système, ou encore faire une combinaison de tout ce qui précède (E. Fyffe, W. Hines, & Kee Lee, 1968). Les projets de science citoyenne présentent toutefois des contraintes qui limitent les champs d'action pour améliorer sa fiabilité. Premièrement, il n'est pas possible de concevoir un nouveau système. Cela reviendrait à changer la foule par un autre composant du système, et donc à passer à une autre forme d'organisation. Deuxièmement, la maîtrise des composants du système est délicate à mettre en œuvre. D'un côté elle peut être gérée en augmentant les barrières à l'entrée pour participer au projet, ce qui est contraire au modèle d'ouverture promulgué par les partisans des sciences citoyennes (Franzoni & Sauermann, 2014). Une autre stratégie consisterait à cibler la communication sur le projet dans des environnements maîtrisés (des universités, des associations de spécialistes). Cette approche peut augmenter la fiabilité globale du système mais limite le nombre de participants possibles.

Les organisateurs peuvent également mettre en place une redondance dans les composants du système. La redondance consiste à disposer plusieurs composants identiques dans un système qui ont les mêmes fonctions. Elle permet d'améliorer la fiabilité globale du système de deux manières :

- elle réduit le risque de panne : si un des composants tombe en panne, le système peut passer par un autre composant.
- Elle augmente les performances du système. Le système peut réaliser la même action plusieurs fois par des composants différents. Un système d'évaluation peut être mis en place ensuite pour choisir le produit qui est de la meilleure qualité.

Le principe de redondance est courant dans des systèmes où les exigences en terme de sécurité sont élevées. Dans les centrales nucléaires de production électrique par exemple, la conception des systèmes de sûreté des réacteurs est réalisée conformément à des calculs statistiques pour minimiser les risques de pannes ou d'incident nucléaire. C'est le cas notamment des groupes électrogènes de secours à moteur diesel. En fonctionnement normal, le réacteur nucléaire est alimenté électriquement par des sources électriques externes. En cas de perte de ces sources, des groupes électrogènes de secours sont utilisés pour alimenter en électricité et permettre le fonctionnement des systèmes de sauvegarde qui seraient mis en œuvre en cas d'accident. Tous ces systèmes de sûreté sont doublés ou triplés, voire quadruplés tel que si l'un d'eux ne fonctionne pas, le système en réserve s'y substitue.

Bien que la redondance améliore fortement la fiabilité d'un système, celle-ci est souvent accompagnée de coûts importants. Dans les systèmes de sûreté dans les centrales nucléaires, le

seul coût partagé dans la construction des groupes électrogènes est celui de la conception du système. En revanche, la construction et l'entretien de ces systèmes sont indépendants et le coût de la redondance est environ égal au coût d'un système multiplié par le nombre de système $co\hat{u}t_{redondance} \approx n\ x\ co\hat{u}t_{composant}$.

Dans les projets de science citoyenne en revanche, nous avons vu que le coût peut être en partie externalisé, donc non supporté par les organisateurs. En fait, il est possible de mettre en place un système de redondance qui s'appuie sur le fait que le nombre de participants est très grand, donc au-delà de ce qu'il est nécessaire pour réaliser la tâche. Au lieu de déléguer une tâche unique à chaque participant, les organisateurs peuvent mettre en place une redondance dans le système pour faire exécuter plusieurs fois la même tâche à des participants différents. Ainsi, le coût de la redondance n'augmente pas ou très peu : il est dépendant du coût d'exécution associé à chaque participant.

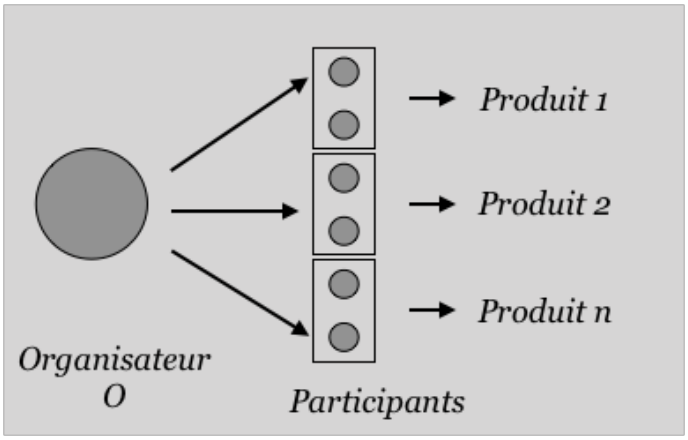


Figure 19. Principe de redondance dans les projets de science citoyenne pour augmenter la fiabilité du système.

Le cas du projet Galaxy Zoo (**annexe 1**) est un exemple typique d'utilisation de la redondance. Pour rappel, la plateforme Galaxy Zoo est un projet ayant regroupé plus de 250 000 volontaires qui ont aidé au codage d'images astronomiques issues de télescopes, et qui ont contribué à la découverte de nouvelles classes de galaxie. La participation est devenue rapidement virale, et sept mois après le lancement du projet environ 900 000 galaxies furent codées. A titre de comparaison, 50 millions de classifications auraient requis plus de 83 années à plein temps pour un scientifique seul. Afin de réduire la probabilité d'un codage incorrect et donc de mauvaise qualité, les galaxies furent codées plusieurs fois par différents volontaires, pour un total d'environ 50 millions de classifications, soit environ 55 codages différents par image (**tableau 7**). Une fois les galaxies codées, les organisateurs ont analysé les codages groupés des participants sur chaque image grâce à des outils statistiques et ont choisi un codage pour une image de galaxie.

Gestion de Galaxy Zoo

Tâche à exécuter	Coder manuellement 900 000 images de galaxies
Division en sous-tâches	Coder individuellement chaque image de galaxie suivant une séquence d'actions prédéterminée
Séquence d'états pour chaque sous-tâche	Classification des galaxies prédéfinie via un arbre décisionnel: - lisse, caractéristiques ou disque, étoile ou artefact - complètement rond, entre les deux, en forme de cigare, - ...
Actions à réaliser	Evaluer visuellement les caractéristiques de la galaxie sur chaque image
Gestion de la qualité de l'exécution	Limiter la probabilité d'un codage incorrect
Redondance	Codage de 50 millions d'images au lieu de 900 000 uniques et évaluation par un modèle statistique

Tableau 7. Gestion de la délégation d'une tâche de type recette : le cas de Galaxy Zoo.

1.3. REPETITION, PERTE ET CAPITALISATION

Dans l'exemple de Galaxy Zoo, la redondance est gérée par un processus de **capitalisation par agrégation**. Ce modèle de capitalisation est celui que nous avions présenté dans le chapitre précédent la notion de capitalisation appliqué durant la tâche. En fait, nous pouvons rapprocher de manière plus générale le principe de redondance à la notion de tâche répétée. Créer une redondance dans l'exécution d'une tâche revient exactement à répéter plusieurs fois la même tâche au sein d'un système. Les questions de performance que nous avons présentées dans le cas où la tâche est répétée s'appliquent dès lors qu'il y a redondance dans le système de production de la tâche.

Baisse de fiabilité du système → *Redondance* → *Répétition des tâches* → *Perte* → *Capitalisation*

Nous avions vu que la répétition des tâches peut mener à des variations de performance entre les tâches et durant la tâche, comme des problèmes de pertes. Dans le cas de la délégation à la foule, ces pertes peuvent être très importantes. En effet, la fiabilité du système dans les tâches ouvertes à la foule repose souvent sur une approche statistique (concordance statistique sur des tâches répétées dans le cas de la recette, augmentation de la probabilité de trouver une bonne réponse pour la résolution de problèmes, multiplication des explorations dans le cas des tâches couplées inventives) et une grande partie de ce qui est produit par les participants n'est pas utilisé. Dans le cadre de la résolution de problème par exemple, chaque participant travaille individuellement et seule la meilleure solution est conservée. Toutes les autres productions peuvent être vues comme un ensemble de mauvais chemins qui n'ont pas servi à l'élaboration de la solution finale : le taux

de perte est proche de 100% ($\frac{n-1}{n}$ solutions non utilisées). D'un point de vue économique, une telle perte est un non sens pour la productivité.

1.4. PERTE DURANT LES TACHES AVEC EXPLORATION DE L'ESPACE : NOTION DE CAPITALISATION CROISEE

1.4.1. Redondance et perte dans la résolution de problèmes

Dans le cas de la résolution de problèmes, le mode d'organisation par redondance permet de multiplier les positions de départ dans l'exploration des solutions, et donc d'augmenter la probabilité d'être proche de la solution optimale (Afuah & Tucci, 2012; King & Lakhani, 2013). Chaque participant va explorer une partie de l'espace des plans d'action créant une multitude de recherches locales (Afuah & Tucci, 2012). L'organisation de la tâche est donc basée sur *l'augmentation de la diversité des solutions proposées*. Ce type d'organisation a fait l'objet de plusieurs études dans le cas de la résolution de problèmes et plusieurs chercheurs ont suggéré son intérêt pour obtenir des solutions de meilleure qualité (voir e.g. Amatriain, 2012; Boudreau, Lacetera, & Lakhani, 2011; Brabham, 2008; Cavallo, Street, York, & Haven, 2012; Poetz & Schreier, 2012). En plus de fournir des solutions plus performantes qu'une exploration individuelle (Poetz & Schreier, 2012), ces projets sont peu coûteux car les organisateurs ne rémunèrent généralement que le vainqueur.

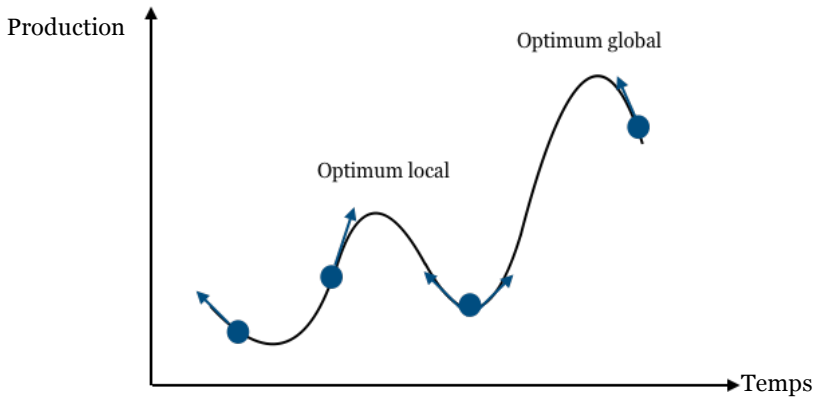


Figure 20. Multiplication des points d'entrée et des chemins d'expérimentation dans l'espace des plans d'action.

En revanche, l'organisation compétitive pour la résolution de problèmes présente l'inconvénient majeur de générer une perte importante de productivité. En effet, seule la séquence d'actions du vainqueur est exploitée, tandis que les autres séquences ne sont généralement jamais divulguées aux organisateurs. En terme de rentabilité, le ratio de perte de production est proche de 100% lorsque le nombre de participants est grand. Pour limiter cette perte, des économistes et des gestionnaires suggèrent de nouveaux modèles d'organisation incluant une coopération entre les participants (Afuah, 2018; Benkert & Letina, 2016; Boudreau & Lakhani, 2015). Au lieu que

chaque participant travaille individuellement, les participants peuvent avoir accès à tout ou partie de ce qui a été produit par les autres. Ce système est compatible avec les projets de science citoyenne, où les organisateurs ouvrent les résultats intermédiaires durant le processus. Chaque citoyen de la science a accès aux séquences d'action que les autres participants ont produit, et peut les utiliser sans contrainte particulière. Contrairement au modèle compétitif, l'approche collaborative de la résolution de problèmes diminue le risque de perte d'une « bonne idée » qui peut être réutilisée par les autres participants. Ce modèle d'organisation a été montré comme donnant des solutions de meilleure qualité que le processus compétitif dans les mêmes circonstances (Boudreau & Lakhani, 2015). Or, si l'ouverture des résultats intermédiaires augmente les performances du processus les mécanismes d'interactions entre les participants sont peu clairs. En effet, l'étude réalisée par Boudreau et Lakhani montre que si la meilleure solution est de meilleure qualité, il y a moins de diversité dans les solutions, contredisant le modèle compétitif. Alors que les modèles de gestion de la résolution de problèmes par la foule se basent généralement sur un modèle input/output où les organisateurs n'interagissent pas durant le processus, l'ouverture à la collaboration force à repenser ce paradigme et à concevoir un modèle de gestion qui prenne en compte les interactions entre les participants durant le processus.

1.4.2. Perte durant les tâches couplées inventives

Les études de performance menées sur l'ouverture des tâches de type résolution ne s'appliquent pas dans le cas des tâches couplées inventives. En effet, ces études suggèrent que la performance de l'ouverture à la foule nécessite l'existence d'un problème bien formulé, c'est-à-dire qu'il existe un état désirable formulé. Cependant, nous pouvons étendre les notions de perte à l'exploration de l'espace des états et celui des plans d'action. Dans le cas d'une tâche couplée inventive compétitive, chaque participant explore de manière individuelle les deux espaces. Cette exploration permet d'augmenter la diversité des zones des espaces explorées. En effet, nous avions vu dans le chapitre précédent que le nombre d'hypothèses possibles croît exponentiellement avec le nombre de variables à combiner. Dans le cas où la tâche est déléguée à une foule, la grande taille de l'espace des hypothèses diminue la probabilité que deux participants formulent la même hypothèse. En revanche plus il y a de participants à explorer les espaces, plus il y a de chances que certaines zones explorées des espaces aient des parties en commun. Par exemple, deux participants travaillant sur des hypothèses différentes mais dont les relations ainsi qu'une partie des objets sont identiques. Or, cette zone commune pourrait être délaissée par un des participants qui n'en voit pas l'usage tandis que l'autre y aurait un intérêt dans son exploration. L'exploration menée par le premier participant serait dans ce cas une pure perte. Ces risques de perte durant la tâche sont moins importantes que pour la résolution de problèmes car ils dépendent de multiples facteurs : existence d'une zone commune d'exploration, variation dans les intérêts sur cette zone commune. C'est pourquoi notre étude se concentrera essentiellement sur les tâches de type résolution de problèmes.

1.4.3. Limiter les pertes par capitalisation croisée

Alors que la résolution compétitive de problèmes est basée sur la diversité des solutions proposées, l'organisation collaborative cherche à diminuer les pertes en améliorant la *capitalisation* sur ce qui est produit par les participants. Pour améliorer la performance globale du processus de résolution de problèmes, la gestion de la capitalisation est donc un élément central que la littérature n'a que peu étudié dans le cadre d'une résolution collective. En effet, la réutilisation entre les participants, que nous définissons comme **capitalisation croisée**, permet notamment de réduire le risque que plusieurs participants explorent plusieurs fois la même zone de l'espace.

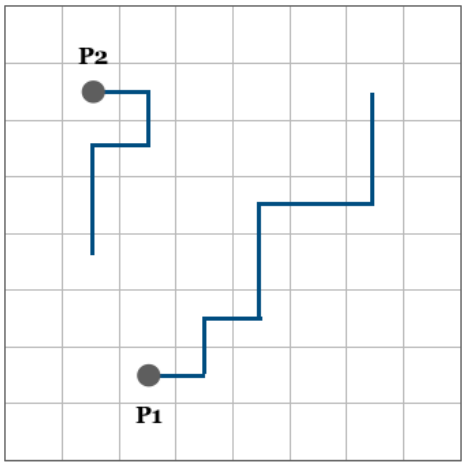


Figure 21. Illustration des parcours de deux participants dans le labyrinthe.

Nous pouvons illustrer la capitalisation croisée au travers de l'exemple du labyrinthe. Supposons deux participants *P1* et *P2* qui soumettent chacun une séquence d'action *A1* et *A2* respectivement. Supposons également que les participants n'ont pas trouvé de séquence qui mène à la sortie et doivent proposer une nouvelle solution. Dans le cas où la résolution est compétitive, chaque participant ne sait pas ce que l'autre a soumis. Chacun ne connaît donc que son état initial, son état final et les états intermédiaires de la séquence qu'il a soumise. La mémoire des états possibles dans le labyrinthe est individuelle. Il y a donc un risque que le participant *P1* soumette une nouvelle séquence dont certains états se confondent avec la séquence d'action *A2*. Dans notre exemple, il a 6 chances sur 64 que cela se produise. De la même manière, *P2* a 10 chances sur 64 de proposer un état qui aura déjà été exploré par *P1*. Dans le cas où la résolution est collaborative, chaque participant connaît la séquence d'action soumise par l'autre et voit toutes les zones de l'espace qui ont été explorées. La probabilité de proposer un état déjà exploré tombe à 0. De plus les deux participants connaissent chacun 16 états sur 64 de l'espace et donc ont moins d'états à tester.

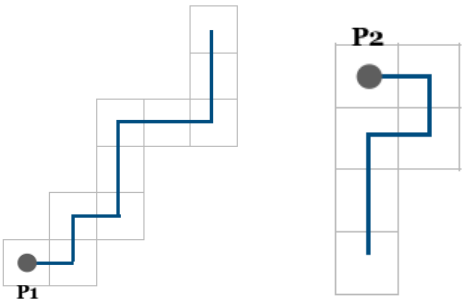


Figure 22. Dans la résolution compétitive, chaque participant ne peut capitaliser que sur sa séquence d'action.

Quel comportement est adopté par les participants lors de la capitalisation croisée ? Est-ce qu'il réutilisent une partie des séquences d'actions soumises par l'autre participant ou bien préfèrent-ils explorer des zones de l'espace encore inconnues ? Nous allons étudier cette question dans le chapitre suivant en analysant le cas du RAMP, une plateforme de data challenge. Nous montrerons que les participants ont un comportement globalement identique qui crée des plateaux de fixation autour de certaines zones de l'espace. Nous montrerons également que la capitalisation croisée peut être améliorée si les organisateurs mettent en place des challenges compétitifs puis collaboratifs.

2. *Performance entre les taches : capitalisation sequentielle*

2.1. Limiter les pertes entre les taches de type elementaire ou recette

Nous avons vu les problématiques de perte lorsque plusieurs tâches identiques sont réalisées en même temps. La redondance peut également être organisée de manière séquentielle, c'est-à-dire sans recoupement entre les tâches identiques. Dans le cas où ce sont les mêmes acteurs qui participent aux projets (du moins en partie), la transmission d'information et la gestion des pertes se fait en partie de manière tacite par des échanges entre les participants (Brady & Davies, 2004). Cependant, dès lors que les tâches sont déléguées à une foule, le risque de pertes est décuplé. En effet, les participants restent rarement sur plus d'un projet, limitant la transmission tacite d'information. Comment alors gérer cette perte dans les projets ?

Dans les tâches de type élémentaire ou recette, la partie principale de l'apprentissage concerne ce que les organisateurs peuvent apprendre de l'utilisation de citoyens dans le processus, des compétences des participants et des types d'actions qu'ils peuvent leur demander d'exécuter (Sauermann & Franzoni, 2014). Le retour d'expérience fait par les organisateurs permet de définir les contours de la méthode utilisée et de mieux cibler pour les prochains projets ce qu'il est possible de demander à la foule et sous quelle forme la tâche doit être présentée. A titre

d'exemple, le premier projet Galaxy Zoo a permis de démontrer aux organisateurs l'utilité de la méthode pour la production de catalogues à grande échelle et de découvertes fortuites d'objets individuels (Thomas et al., 2013). Depuis lors, cette méthode a été étendue au-delà des morphologies de galaxies pour inclure l'identification de la supernova (Smith et al., 2011), la découverte d'exoplanètes (Fischer et al., 2012) et un recensement des bulles associées à la formation d'étoiles dans la Voie Lactée (Simpson et al., 2012), ainsi que divers problèmes liés aux «données volumineuses» en dehors de l'astronomie avec la plateforme Zooniverse. Alors que le projet original Galaxy Zoo identifiait les galaxies comme étant des types précoces, des types tardifs ou des fusions, le projet Galaxy Zoo 2 s'est attelé à mesurer des caractéristiques morphologiques plus fines comme les barres, les renflements et les formes des disques à bords, ainsi que la quantification des forces relatives des renflements galactiques et des bras spiraux.

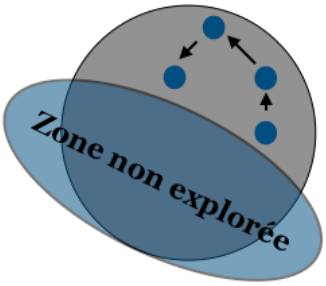
2.2. PERTE ENTRE LES TACHES AVEC EXPLORATION DE L'ESPACE : LE PRINCIPE DE CAPITALISATION SEQUENTIELLE

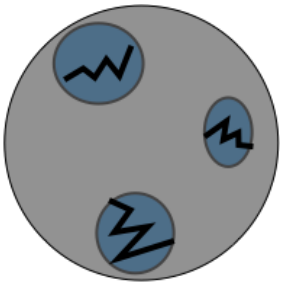
La transmission d'informations entre les projets est plus complexe dans le cadre de tâches incluant l'exploration d'un ou plusieurs espaces. En effet la recherche d'une solution est souvent sujette à de « mauvais chemins », de « l'expérimentation », de la « sérendipité » et de « l'incertitude » (Abernathy & Rosenbloom, 1969; Boudreau, Lakhani, & Lacetera, 2008; Loch, Terwiesch, & Thomke, 2001). La découverte d'une solution à un problème se confronte alors à la création de connaissances qui se fait souvent en tant que sous-produit involontaire de l'activité de projet (DeFillippi, Jones, & Arthur, 2001).

Supposons par exemple que la résolution d'un problème soit donnée successivement à différents participants ou groupes de participants. Durant la première résolution, chaque participant va explorer une partie de l'espace en proposant un ensemble de séquences d'action pour résoudre le problème. Nous pouvons représenter cette exploration par une succession de petits chemins dans l'espace (**figure 23**). Cette exploration va former une zone de l'espace qui aura été explorée par le participant. Si plusieurs participants explorent l'espace en même temps, nous pouvons représenter les explorations individuelles avec une cartographie de l'espace. Cela permet de distinguer les zones de l'espace explorées des zones de l'espace non explorées, ainsi que des distances possibles entre les zones de l'espace explorées (**figure 24**). Cette cartographie fournit des informations pour les tâches successivement répétées : elle représente la mémoire de ce qui a déjà été explorée dans l'espace. Sans cette mémoire, les participants des tâches suivantes ne pourraient pas s'appuyer sur les explorations déjà faites, et donc recommenceraient l'exploration comme si rien n'avait été fait auparavant. Or, cette cartographie permettrait de savoir si certaines zones de l'espace ont été très explorées tandis que d'autres ont été négligées, et potentiellement réduire les coûts d'explorations des participants.

L'existence de deux espaces à explorer dans les tâches couplées inventives augmente d'autant plus les choix dans l'exploration et donc le risque de partir sur de mauvais chemins. Nous avons

montré dans le chapitre 5 par exemple que la génération d'hypothèses basée sur les données est associée à un risque très fort de générer des hypothèses sans intérêt scientifique.





Figures 23 et 24. A gauche : exploration individuelle de l'espace. A droite : cartographie des explorations individuelles.

Réduire les pertes en utilisant la production des tâches précédentes fait apparaître un processus de **capitalisation séquentielle**. Dans un premier temps, les participants explorent les espaces en soumettant plusieurs fois des solutions. A chaque solution soumise, ils améliorent leur connaissance de l'espace et peuvent capitaliser sur cette connaissance pour soumettre des solutions de meilleure qualité. Une fois la tâche terminée, il est possible de cartographier l'ensemble de l'exploration réalisée par chacun des participants. Cette cartographie permet de capitaliser sur ce qui a été produit durant la tâche et de le réutiliser dans la tâche successive.

Ce principe de capitalisation séquentielle permet de gérer la production entre les tâches et de réduire les pertes de production en utilisant successivement ce qui est produit. Cependant, si le principe de capitaliser sur ce qui a été produit semble assez intuitif dans un processus de performance, les moyens pour gérer cette capitalisation séquentielle doivent être éclaircis. Dans le cas d'une tâche de type résolution de problèmes, la production durant la tâche peut être résumée à la séquence d'action qui a le meilleur score à la fonction *test()*. Celle-ci peut par exemple servir de base pour la répétition successive de la tâche. Ce processus de capitalisation séquentielle est moins clair dans le cas d'une tâche couplée inventive. Qui est en charge de la capitalisation dans le processus de capitalisation séquentielle ? Quels sont les moyens à mettre en œuvre pour gérer cette capitalisation ? Nous tenterons de répondre à ces questions au travers de notre analyse du cas Epidemium dans les chapitres 8 et 9.

3. *SYNTHESE DES SITUATIONS DE GESTION DE L'OUVERTURE*

Le modèle que nous avons construit dans les chapitres 5 et 6 propose de distinguer quatre type de tâches déléguées par l'ouverture du processus scientifique à la foule : tâche de type élémentaire ou recette, résolution de problème, tâche couplée inventive. Chacune de ces tâches est associé à des activités différentes (exécution d'une séquence d'action, exploration d'un ou deux espaces) et demande de mettre en place des moyens de gestion différents. Nous avons montré dans ce chapitre comment la capitalisation a été gérée pour des tâches de type élémentaire ou recette au travers de l'exemple de Galaxy Zoo. Par la suite, nous allons tenter de montrer comment cette capitalisation est gérée dans les projets de science citoyenne sur des tâches de type résolution de problèmes et tâche couplée inventive.

	Tâche élémentaire ou recette	**Résolution de problèmes**	**Tâche couplée inventive**
Type d'espace dans le modèle	Un espace constitué de l'ensemble des plans d'action	Un espace constitué de l'ensemble des plans d'action	Deux espaces : - Un espace constitué de l'ensemble des plans d'action - Un espace constitué des hypothèses
Activité	Exécution d'une action ou d'un plan d'action prédéterminé (état du monde espéré entièrement connu)	Exploration de l'espace des plans d'action grâce guidée par l'existence d'une fonction test et d'un état final désiré	Exploration des deux espaces suivant le libre choix de l'explorateur pour construire le quoi et le comment des hypothèses
Collaboration	Organisation hiérarchique seeker-solver, division du travail entre les participants	Organisation hiérarchique seeker-solver, coordination par réutilisation de la production durant la tâche	Organisation hiérarchique seeker-solver,
Risque de perte dans la répétition des tâches	Distinguer les bonnes exécutions des mauvaises, ne pas exploiter au maximum la foule	Perte des « bonnes idées » dans l'exploration de l'espace des plans d'action durant la tâche (risque élevé), risque de recommencer l'exploration de zéro (perte entre les tâches)	Zones communes d'exploration dans les espaces (risque faible), risque élevé de produire des éléments non intéressants
Capitalisation durant la tâche	Capitalisation par agrégation	**Capitalisation croisée**	**Capitalisation croisée**
Capitalisation entre les tâches	Retour d'expérience	**Capitalisation séquentielle**	**Capitalisation séquentielle**
Organisation de la capitalisation	Production par les participants, Capitalisation par les organisateurs	Production et capitalisation croisée par les participants Capitalisation séquentielle gérée par les organisateurs	Capitalisation séquentielle gérée par les organisateurs ?

Tableau 8. Synthèse des tâches au sein du processus scientifique et des types de capitalisation.

PARTIE 3 – ANALYSE ET EXPERIMENTATION DE DISPOSITIFS ORGANISATIONNELS

CHAPITRE 7 – PILOTAGE DE LA PERFORMANCE DURANT LES PROJETS DE SCIENCE CITOYENNE AVEC EXPLORATION D'ESPACES : FONCTIONNEMENT ET IMPACT DE LA « CAPITALISATION CROISEE »

RESUME DU CHAPITRE 7

Dans ce chapitre, nous analysons l'impact de la **capitalisation croisée** - réutilisation des résultats intermédiaires durant la résolution d'une tâche incertaine - sur la perte de production ainsi que les métriques classiques de performance (qualité, coût, délai). La perte de production dans les tâches ouvertes est très peu étudiée dans la littérature. Or, dans le cas de la résolution de problèmes compétitif traditionnel, seule la meilleure solution est utilisée ce qui crée des pertes de production de l'ordre de 100%.

Notre terrain d'étude est la plateforme RAMP qui sert à mettre en place des data challenges. Les organisateurs du RAMP ont effectué une série de tests pour évaluer lequel des deux modes (compétitif ou collaboratif – solutions accessibles) était le plus performant en terme de qualité de la meilleure solution. Leurs résultats ont corroboré les conclusions du papier de Boudreau et Lakhani, à savoir que le mode collaboratif obtient de meilleures performances que le mode compétitif. Par la suite, les organisateurs du RAMP ont eu besoin de mettre en place des challenges hybrides (fermé puis ouvert) pour faciliter la notation d'étudiants. Ils ont constaté que les performances étaient généralement meilleures que pour les challenges uniquement ouverts. Nous avons étudié deux challenges hybrides pour comprendre les mécanismes qui impactent la qualité des solutions lorsqu'il y a capitalisation croisée.

Notre analyse montre que les modes d'exploration diffèrent entre la phase fermée et la phase ouverte. Dans la phase fermée, chaque participant explore de façon individuelle l'espace des solutions, augmentant ainsi la diversité des solutions proposées et donc la probabilité de trouver une bonne solution. Dans la phase ouverte, les participants réutilisent les soumissions existantes et « convergent » rapidement vers un type de solution unique. Cela permet aux participants de réduire le coût d'exploration et l'incertitude liée au problème. La réutilisation des soumissions réduit ainsi la perte de production. Nous montrons ensuite que lors d'un challenge hybride, la diversité en amont de la phase de capitalisation a un impact positif sur la qualité de la meilleure solution. En effet, une plus grande diversité dans les solutions augmente la probabilité de trouver des « bonnes idées ».

La notion de perte est très peu présente dans la littérature qui étudie les tournois et la délégation de tâche à une foule. En fait, dès que la résolution de la tâche est incertaine, le modèle d'ouverture se base essentiellement sur l'opportunité statistique d'accéder à un grand nombre de solutions et donc d'augmenter la probabilité d'avoir une bonne solution (e.g. Afuah & Tucci, 2012; Terwiesch & Xu, 2008). En ce sens, les modèles ont tendance à résumer la performance au travers du score de la meilleure solution. D'un point de vue statistique, l'objectif est de maximiser la valeur extrême et les modèles ne s'intéressent pas à la valeur moyenne de toutes les solutions.

Cette approche présente ses limites dès lors que l'on cherche à minimiser les pertes en capitalisant sur ce qui est produit durant la tâche. Nous proposons d'étudier la notion de « **capitalisation croisée** » que nous avons développé dans le chapitre 6 : dans ce cas, les solutions sont accessibles à tous les participants durant le processus (l'ouverture des résultats intermédiaires dans les projets de science citoyenne). Ainsi, toutes les solutions proposées peuvent jouer un rôle dans la construction de la meilleure solution possible. Les organisateurs doivent ainsi s'intéresser à la fois à la valeur de la solution maximale mais également aux autres solutions. Pour proposer des stratégies de pilotage adaptées à la capitalisation croisée, deux questions se posent : Quel processus d'exploration est mis en place par les participants quand les solutions sont accessibles durant la tâche ? Quel est l'impact de la capitalisation croisée dans la performance du processus ? L'unique étude expérimentale comparée dans la littérature a montré que les participants réutilisaient spontanément les solutions des autres durant le tournoi pour proposer de nouvelles solutions (Boudreau & Lakhani, 2015). Les participants regardent les solutions et s'inspirent de celles-ci pour proposer de nouvelles solutions. Cependant, les stratégies d'exploration adoptées par les participants ne sont pas clairement établies et souffrent d'un manque de formalisme. De plus, les auteurs se sont intéressés à contraster le mode compétitif (pas de capitalisation) avec le mode collaboratif (capitalisation possible), sans étudier la gestion de la performance dans le cadre collaboratif. Ce sont ces questions auxquelles nous allons tenter de répondre dans ce chapitre.

Notre terrain d'étude est la plateforme RAMP, une plateforme pour mettre en place des tournois de data challenge. Les problèmes ouverts à la foule peuvent être assimilés à des tâches de type résolution de problèmes tel que nous les avons modélisés dans le chapitre 5. La plateforme peut mettre en place des challenges compétitifs (fermé) ou collaboratifs (ouvert). Dans le cas compétitif, les codes soumis sont fermés aux autres participants; les participants ne peuvent voir que leur propre code ainsi que le score de chacun. En mode collaboratif, tous les participants ont accès à tous les codes soumis, et ils sont encouragés à les regarder et à réutiliser les solutions. Les organisateurs du RAMP ont effectué une série de tests pour évaluer lequel des deux modes était le plus performant en terme de qualité de la meilleure solution. Leurs résultats ont corroboré les conclusions du papier de Boudreau et Lakhani, à savoir que le mode collaboratif obtient de meilleures performances que le mode compétitif. Par la suite, les organisateurs du RAMP ont eu besoin de mettre en place des challenges hybrides (fermé puis ouvert) pour faciliter la notation d'étudiants. Ils ont constaté que les performances étaient généralement meilleures dans les challenges hybrides que pour les challenges uniquement ouverts. Ce résultat est en contradiction

avec l'analyse faite par Boudreau et Lakhani : eux avaient observé que le mode hybride donnait des résultats intermédiaires (entre le mode compétitif et le mode collaboratif). Quelles sont les conditions qui expliquent cette différence entre leur résultat et les multiples constats empiriques dans le cas du RAMP ?

Nous analysons le cas de deux challenges hybride afin de comprendre le processus d'exploration dans la phase fermée, la phase ouverte et le lien entre les deux. Notre étude montre que durant la phase ouverte les participants ont tendance à réutiliser spontanément les soumissions avec les meilleurs scores suivant trois types de comportement : copie totale de la soumission, combinaison de plusieurs parties de séquences d'action, transformation par intégration de nouvelles actions. La réutilisation des soumissions permet aux participants de réduire l'incertitude lors de l'exploration et de faire baisser leurs coûts d'exploration. Cela crée un effet global de convergence des soumissions vers des un voisinage de quelques solutions types (généralement celles qui ont le meilleur score). Enfin, nous montrons que dans les challenges hybrides la phase fermée augmente la diversité des solutions soumises ce qui permet de « nourrir » la phase ouverte. Plus il y a de solutions différentes en début de la phase ouverte, plus les participants peuvent combiner différentes séquences d'action pour construire une solution optimale.

1. INITIATIVES EMERGENTES POUR REUNIR SCIENTIFIQUES ET EXPERTS EN ANALYSE DE DONNEES : LE CAS DU RAMP

1.1. UNE INITIATIVE POUR LA RESOLUTION COLLABORATIVE DE PROBLEMES LIES A L'ANALYSE DE DONNEES: LE CENTER FOR DATA SCIENCE ET L'OUTIL RAMP

Le Center for Data Science (CDS) de l'université Paris-Saclay est un projet de Lidex et l'une des premières initiatives de science des données en France. Son objectif consiste à développer des méthodes et des outils permettant d'analyser de grandes quantités de données et d'en extraire des informations utiles pour la physique, la biologie, la médecine, la chimie, l'environnement et les sciences humaines. Cette initiative est multidisciplinaire; elle nécessite des recherches sur les méthodologies analytiques (statistiques, processus d'apprentissage automatique, extraction de connaissances, visualisation de données), ainsi que sur la conception de logiciels. Elle regroupe 50 laboratoires d'informatique et plus de 300 chercheurs, couvrant une grande variété d'expertises. Ce centre se concentre principalement sur les applications de données scientifiques telles que les données ATLAS provenant des expériences du CERN et les projets CASD (données économiques et financières françaises / européennes). Le RAMP (pour *Rapid Analytics and Model Prototyping*) est une plateforme de data challenge développée en tant qu'outil interne au Center for Data Science (CDS) de l'université Paris-Saclay pour répondre au déséquilibre spécifique entre l'offre et la demande de projets scientifiques. La conception de cette plateforme rassemble depuis 2014 des chercheurs de l'équipe Théories et méthodes de conception pour l'innovation du Centre

de Gestion Scientifique (CGS) des Mines ParisTech et des spécialistes de l'analyse de données du CDS. Aujourd'hui, RAMP a deux types d'utilisation primaire. Premièrement, il est utilisé comme plateforme de compétition et de collaboration où les spécialistes des données travaillent sur un problème pour des délais relativement courts (généralement un ou deux jours). Cela peut être considéré comme une forme de Hackathon avec la principale différence que l'objectif est d'optimiser une métrique claire. Ce format a été mis en place dans plusieurs endroits comme l'Université Paris-Saclay mais également à l'École d'économie de Paris, au Musée national d'histoire naturelle, sur la plateforme Epidemium ou au Centre national de recherche atmosphérique aux États-Unis. Deuxièmement, RAMP est utilisé en tant que support à l'éducation dans divers cours de science des données. Il a été utilisé à ce jour dans 6 établissements pour 8 cours de science des données et avec la participation d'environ 400 étudiants.

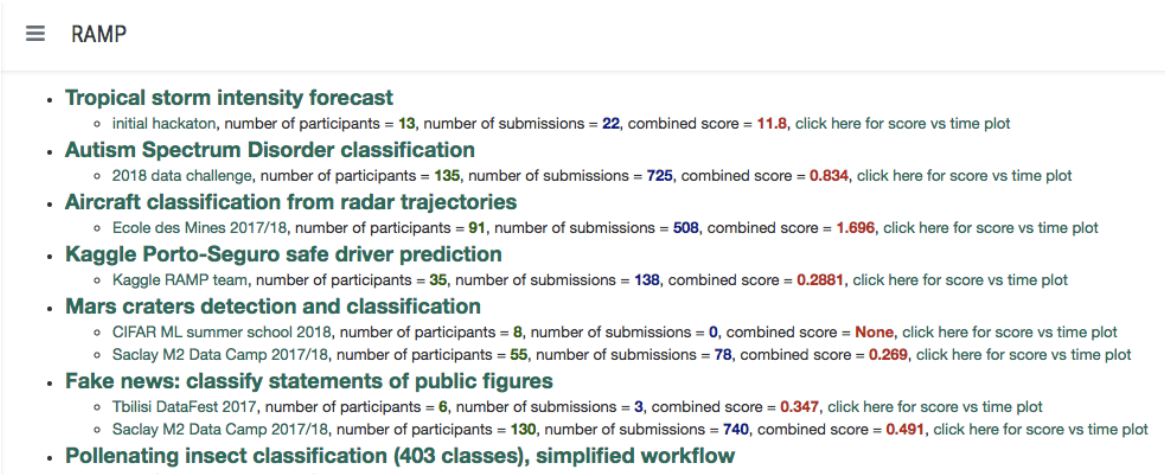
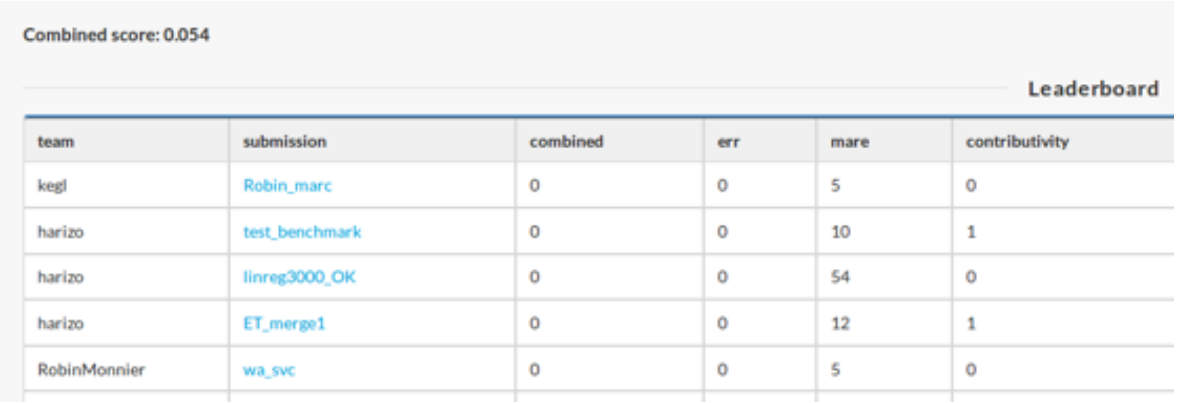


Figure 25. Extrait de la liste des projets réalisés par le RAMP[1]

Le RAMP est utilisé dans le contexte opérationnel suivant. À l'instar d'un data challenge, un scientifique fournit un problème et un ensemble de données à la plateforme. De manière générale, le problème doit souvent être remanié pour pouvoir formuler une hypothèse compatible avec les méthodes d'analyse ainsi qu'une métrique d'évaluation. De plus, les données doivent également être traitées, nettoyées et les données manquantes doivent être comblées pour être exploitable. Cette activité est généralement réalisée par une équipe distincte des participants qui participent au RAMP en soumettant des modèles. Le processus peut prendre de deux semaines à six mois et donne lieu à la conception d'un système de récompenses (des prix, un système de notation,...) en fonction des résultats ainsi qu'un kit de démarrage, généralement un bloc-notes (environnement informatique interactif dans lequel il est possible de combiner du code, du texte enrichi, des mathématiques, des tracés et du contenu multimédia) qui présente le problème de la science du domaine, décrit les données et présente une première solution servant de base.

[1] https://ramp.studio/problems

Le problème est ensuite configuré à l'aide du logiciel RAMP et un événement RAMP est organisé qui est dans la plupart dont la participation est généralement ouverte. Au cours du RAMP, les participants soumettent leurs solutions sous forme de code qui sont ensuite traitées sur un serveur principal distant. Cette étape crée les modèles de prédiction correspondants pour chaque solution soumise. L'expérience montre que les soumissions uniquement des prédictions réalisées par le code sont inutiles pour les scientifiques du domaine, car ils ne fournissent pas de modèles reproductibles et de code prêt pour la production. Les scores sont automatiquement calculés pendant le traitement et affichés sur un classement (**Figure**). Un challenge peut être compétitif (fermé) ou collaboratif (ouvert). Dans le cas compétitif, les codes soumis sont fermés aux autres participants; les participants ne peuvent voir que leur propre code ainsi que le score de chacun. En mode collaboratif, tous les participants ont accès à tous les codes soumis, et ils sont encouragés à les regarder et à réutiliser les solutions. Suivant l'expérience, le mode collaboratif accélère le processus de développement puisque les bonnes idées se répandent rapidement: en effet les challenges sous forme compétitive sont lents, les participants n'échangent pas leurs idées et gaspillent du temps précieux (Kégl et al., 2018). La configuration collaborative a été principalement utilisée dans des événements d'une journée où différents ensembles de problèmes tels que le défi HiggsML (physique des particules), la prédiction de la mortalité (soins de santé), la classification des étoiles variables (astrophysique), la prédiction des insectes (écologie), et des simulations basées sur des agents (macroéconomie) ont été utilisées. Chacun de ces événements a entraîné une amélioration significative par rapport aux solutions de base. Les organisateurs ayant accès à tout le code, le résultat de la journée est un prototype entièrement fonctionnel, réutilisable et prêt à être déployé.


Combined score: 0.054

Leaderboard

team	submission	combined	err	mare	contributivity
kegl	Robin_marc	0	0	5	0
harizo	test_benchmark	0	0	10	1
harizo	linreg3000_OK	0	0	54	0
harizo	ET_merge1	0	0	12	1
RobinMonnier	wa_svc	0	0	5	0

Figure 26. Exemple de classement pour le challenge Drug Spectra

Par la pratique, les organisateurs ont développé un principe organisationnel clé. En utilisant les RAMP dans la formation en science des données, les challenges ont dû être en partie fermés (mode compétitif) pour pouvoir noter les étudiants. La répétition de ce type de challenge a montré que l'exploration est plus performante en termes d'amélioration du score que le modèle purement collaboratif. À partir de là, les organisateurs ont systématisé la construction des challenges hybride: une phase compétitive dans laquelle les participants ont uniquement accès aux scores du

modèle, suivi d'une phase de collaboration au cours de laquelle, en outre, le code source des modèles peut être dévoilé.

1.2. *Principe de fonctionnement du RAMP*

Dans cette section sont résumées les principales composantes du RAMP. Tous les kits RAMP des expériences répertoriées sur le site de soumission sont disponibles ouvertement sur la plateforme GitHub[2]. Dans la **figure 27** sont repris les éléments principaux qui constituent le RAMP.

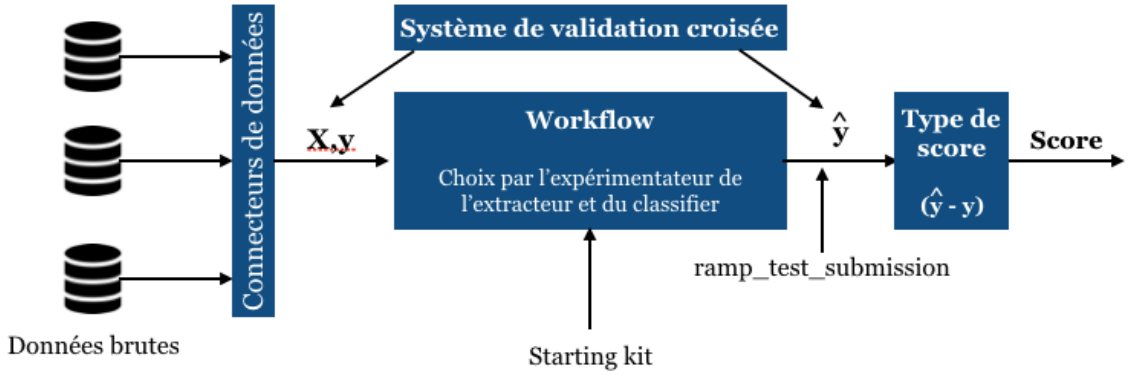


Figure 27. Processus du RAMP

1.2.1. Le « bundle »

Le « bundle » représente l'architecture globale du problème soumis aux expérimentateurs. Grâce à plusieurs itérations, le modèle type de bundle de la plateforme RAMP a été affiné et développé autour de quatre composantes qui définissent l'expérimentation ou le problème prédictif : le *workflow*, le type de score, les connecteurs de données et la validation croisée. Tous ces éléments peuvent être entièrement personnalisable. La principale conséquence de cette modularisation a été la réorganisation du travail: elle a séparé la mise en place des expériences de la résolution et de l'optimisation des modèles. Nous présentons en détail chacun de ces composants du bundle.

1. Le **workflow** décrit les éléments de calcul algorithmique par lequel les données vont passer pour générer un vecteur de prévisions. Un workflow élémentaire peut contenir un seul élément comme un classificateur, un régresseur ou un détecteur mais la plupart du temps celui-ci est modélisé par au moins un extracteur des caractéristiques des bases de données ainsi qu'un prédicteur. Une fois le workflow configuré, les expérimentateurs vont optimiser et soumettre un seul fichier de code pour chaque élément de workflow défini.
2. Les **types de score** (ou métriques) définissent la façon dont les prédictions des soumissions sont évaluées. Plusieurs types de score peuvent être nécessaires pour un seul projet. De manière générale, la prédiction réalisée par le biais de l'algorithme fournit un

[2] https://github.com/ramp-kits

vecteur de prédiction $\hat{y}$ à partir des bases de données test. Cette valeur est ensuite comparée avec la valeur de vérité y des éléments de la base par la fonction $test\ (\hat{y}, y) = \widehat{(y - y)}$ ce qui génère un score.

3. Les **connecteurs de données** sont des fonctions qui à partir d'un fichier de données brutes renvoient un objet X manipulable par les algorithmes et une valeur y qui correspond à la valeur recherchée (le score).
4. La **validation croisée** définit la manière par laquelle le modèle algorithmique va s'entraîner ainsi que la façon dont il va être évalué. Son choix est déterminant pour éviter tout problème de sur-apprentissage dans les algorithmes (surinterprétation d'un algorithme qui « colle » trop aux données et ne peut pas être généralisé). Il existe plusieurs systèmes existants déjà prêts à l'emploi et disponibles dans les bibliothèques de Scikit-Learn. De manière générale, le modèle algorithmique va choisir comment à partir de la base de données existantes les données vont être réparties entre celles qui vont servir à entraîner le modèle (train set) et les données qui serviront à évaluer le modèle (test set). L'algorithme de prédiction s'entraîne et apprend à partir de la base de données X qui est visible pour lui, puis sa capacité de prédiction est évaluée sur le test set Y.

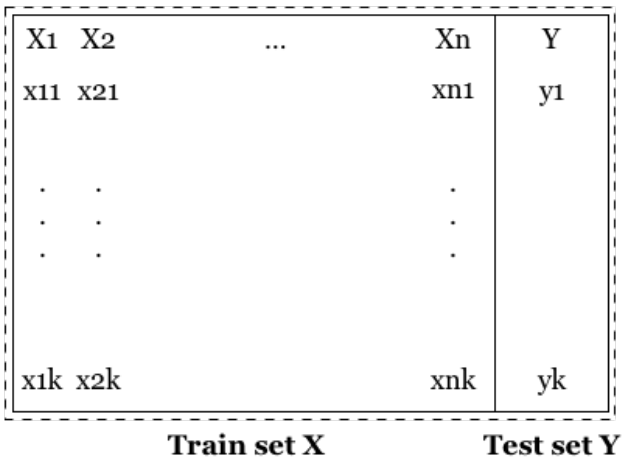


Figure 28. Illustration d'une base de données divisée en train set X et test set Y.

1.2.2. Le « starting kit »

Le « starting kit » est un exemple simple fourni aux expérimentateurs en début de challenge. Celui-ci est conçu par un des organisateurs du RAMP en amont du challenge dans les mêmes conditions que celles fournies aux participants. Il a un double rôle. Tout d'abord, il réduit les barrières à l'entrée pour les participants en leur permettant de commencer par une solution de base valide au lieu de perdre du temps à partir d'une feuille blanche. Deuxièmement, il sert au concepteur expérimental pour effectuer les tests afin de valider le choix de son bundle ainsi que la validation croisée.

1.2.3. Le script « ramp_test_submission »

Une fois que le bundle et le starting kit sont définis et que les données sont en place, un seul script appelé ramp_test_submission est utilisé pour exécuter l'expérience. Ce script a plusieurs rôles:

1. Il permet aux expérimentateurs de réaliser des tests rapides.
2. Il permet aux concepteurs des expérimentations de tester l'expérience eux-mêmes.
3. Il permet aux organisateurs de challenges d'externaliser le développement d'une nouvelle expérience en vérifiant rapidement la conformité ainsi que de s'assurer que le design expérimental est bien respecté.

1.3. LE SETTING DU RAMP COMME DE LA RESOLUTION DE PROBLEMES DANS UN ESPACE EXTENSIBLE

L'architecture générale du RAMP peut être considérée à plusieurs égards comme étant compatible avec le modèle de résolution de problèmes que nous avons présenté dans le chapitre 5. Un problème conçu dans RAMP est généralement de la forme $Y = f(X)$ où Y est l'état final désiré e_{final} caractérisé les objets contenus dans le test set Y ; X représente la base de données d'entraînement de l'algorithme ; f est le code soumis par les participants. Une fois soumis, l'algorithme est évalué grâce à la fonction ramp_test_submission, équivalent de la fonction $test()$ dans notre modèle. Cette fonction évalue la distance entre le vecteur de prédiction $\hat{y} = f(X)$ obtenu par le modèle et l'état final désiré y tel que $test\ (\hat{y}, y) = (\hat{y} - y)$. A noter que d'autres métriques de performance peuvent être intégrées comme la vitesse d'exécution de l'algorithme ou encore le besoin en ressources de calcul pour construire l'algorithme. Une fois le code analysé, celui-ci obtient un score de prédiction, c'est-à-dire un point sur la fonction de valeur.

Chaque solution proposée est sous la forme d'un code informatique, ce qui correspond à une séquence d'instruction pour l'ordinateur qu'il exécute suivant un ordre précis. L'espace contenant tous les codes informatiques peut être représentée comme un espace extensible dans lequel chaque vecteur de la base est associé à une famille de méthodes informatiques (**figure 29**). L'espace contient entre autres les familles de méthodes traditionnelles issues de l'intelligence artificielle utilisées dans les algorithmes de prédiction tel que XGBoost, Random Forest, k-NN, réseaux neuronaux, ainsi que toutes leurs variantes : modification des paramètres, variations à l'intérieur d'une famille de méthodes. Ce sont ces méthodes qui sont généralement privilégiées pour résoudre des problèmes de prédiction. L'exploration se concentre principalement dans cette partie de l'espace, où les participants explorent les variations possibles des familles de méthode.

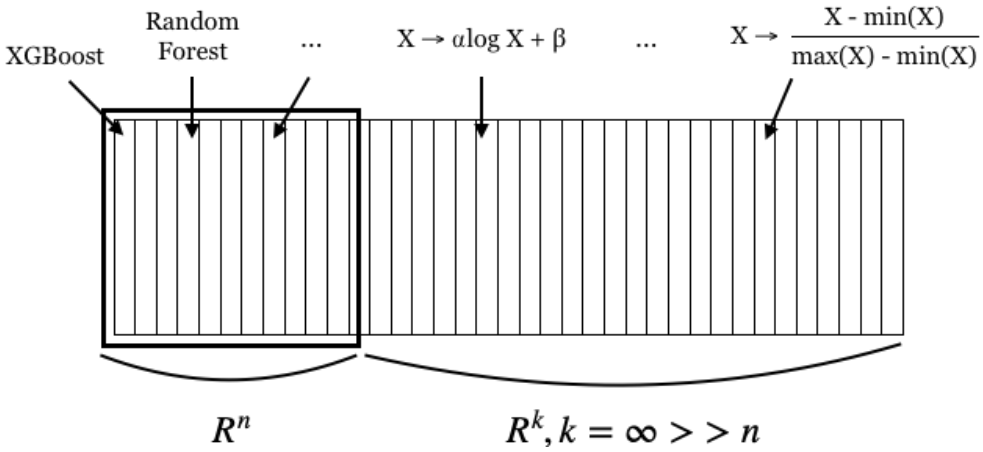


Figure 29. Représentation de l'espace du code informatique.

L'espace du code informatique contient également toutes les autres instructions possibles que l'on peut appliquer sur des données. On retrouve par exemple des familles de méthodes mathématiques basiques comme l'application de la fonction *log*, la transformation linéaire ou encore la normalisation d'un vecteur objet. Cette portion de l'espace est elle de dimension infinie (ou très grande) et extensible (on peut toujours trouver une nouvelle famille d'actions à appliquer sur les données). De manière générale, les participants explorent moins cette partie de l'espace. En effet, sa taille est tellement grande qu'elle contient potentiellement plus d'instructions inutiles qu'intéressantes pour la résolution du problème, impliquant un risque élevé de d'augmenter le coût de l'exploration et de perdre du temps.

Figure 30. Quelques images d'abeilles issues de SPIPOLL.

Pourtant, le retour d'expérience du RAMP a montré l'importance de l'exploration dans cette partie de l'espace. Un des exemples les plus frappants est celui du challenge SPIPOLL[3]. L'objectif du

[3] https://github.com/ramp-kits/pollenating_insects_3_simplified

challenge était de classer les images d'insectes du projet de crowdsourcing SPIPOLL du Muséum d'histoire naturelle de Paris, dans le but d'étudier quantitativement les insectes pollinisateurs en France. Lors d'une des premières éditions de ce challenge, la base de données était constituée de 20 000 images d'insectes pour 18 types d'abeilles. En seulement une journée d'activités, les participants sont passés du score de 0.31 pour le starting kit à un score de prédiction de 0.70. En analysant le code, les organisateurs ont observé que ce qui avait le plus impacté le score n'était pas le choix du modèle prédictif, mais deux lignes de codes ajoutées par un participant. Ces lignes consistaient à couper les bords de l'image pour n'analyser que son centre. En effet, il avait remarqué que les photographes centraient tout le temps l'abeille sur l'image. Ainsi les informations du bord des images avaient tendance à parasiter le modèle et donc générer de moins bons résultats. Cette simple action a permis de démultiplier les performances des soumissions qui ont suivies et qui se sont inspirées de cette idée.

1.4. METHODOLOGIE ET METRIQUES D'ANALYSE

Dans la méthodologie employée par Boudreau et Lakhani (2015) pour leur étude, chaque solution a été analysée par des experts en construisant des catégories de méthodes employées. Les experts ont ainsi identifié 56 combinaisons ou approches de solution distinctes à partir d'une base de 10 familles de méthodes. Chaque code soumis est donc une combinaison de ces 10 méthodes. A noter qu'il existe des variations entre chaque approche dans la mise en œuvre particulière et la qualité des solutions, ce qui explique l'existence de scores très différents avec un même type d'approche.

Ce modèle combinatoire utilisé pour analyser les différentes solutions restreint l'exploration à une partie de l'espace du code dans un espace de dimension 10. Or, une grande partie de la valeur ajoutée de l'ouverture à la foule provient de l'exploration de la partie infinie de l'espace comme l'illustre le cas de SPIPOLL. Dans notre analyse, nous avons besoin de rendre compte de cette exploration afin de mieux percevoir les variations de l'exploration entre la phase fermée et la phase ouverte. Chaque solution est en effet une combinaison de l'exploration des familles de méthodes classiques et de l'extension de l'espace par introduction de nouvelles idées.

Au lieu d'analyser directement le code, nous nous intéressons au comportement de celui-ci, c'est-à-dire à sa capacité prédictive. Nous utiliserons pour cela un outil informatique, le t-SNE, qui permet de comparer entre eux les comportements de différents codes et de les représenter sur un espace à deux dimensions (Van Der Maaten & Hinton, 2008). Cet outil permet ainsi d'évaluer la « distance » entre deux codes en terme de comportement. En plus de donner des indications sur le fait que les participants aient réutilisé ou non une partie du code des autres, l'outil permet d'évaluer à quel point la réutilisation a été proche. Nous couplerons cette analyse avec une étude du comportement des participants durant la phase ouverte, c'est-à-dire la façon d'utiliser les soumissions des autres participants. Nous analyserons deux indicateurs : d'abord le clic sur une soumission, enregistré automatiquement par le programme RAMP. Ensuite, le crédit alloué à chaque soumission : l'organisateur du RAMP a demandé aux participants d'indiquer les crédits

qu'ils allouent aux soumissions précédentes. À titre d'exemple, un crédit de 30% à sa soumission montre que le participant estime qu'il a été influencé à 70% par le modèle qu'elle attribue. Nous proposerons enfin une explication à la différence de résultats sur les challenges hybrides entre les cas du RAMP et le cas étudié par Boudreau et Lakhani.

Elemental techniques used in solutions.

Method	Description
1	**Filtering by unmapped alignment scone (Hamming distance)**: Compare the query string against strings from sets A, B and/or C, trying various possible offsets.
2	**Filtering by comparing frequencies of hashed chunks:** For both the query string and strings from A, B and/or C. move a sliding window across the string and make a frequency table of the chunks that appear in the window, optionally after hashing the chunks. Select the best match(es) between the frequency table obtained from the query and those from the carpus.
3	**Dynamic programming:** Compute the actual Levanshtein distance between a portion of the query and strings from sets A, B and/or C.
4	**Dynamic programming extended to more than one section (A, E, C) at once:** Extend the dynamic programming Levenshtein distance computation to find the optimal edit distance between (a portion of) the query and all possible A+B, B+C or A+B+C combinations.
5	**Bit optimizations:** Use bitwise arithmetic to operate or multiple characters at a time.
6	**SEE optimizations:** Use Streaming SIMB Extensions (a CPU instruction set enabling single-instruction multiple-data (SIMD) parallelization) to process up to 16 characters or strings at once.
7	**Refinement of choices after finding initial solution:** As a post-processing step, hold two of the three selections fixed and reoptimize the third.
8	**Fast approximation of edit distance in well-matched regions:** Use restricted dynamic programming. Hamming distance or variants thereof to speed up the cumputation.
9	**Precomputation of statistics on the string corpus:** Perform offline analysis of the provided sets A, B and C, and use the data obtained for decision making in the algorithm.
10	**Explicitly prefer shorter B strings:** In heuristic approaches, give bonuses to shorter strings from set B (which empirically have greater likelihood of producing high scores).

Figure 31. Liste des 10 méthodes employées dans le challenge de Boudreau et Lakhani (2015)

2. CAS D'ETUDES : LE DRUG SPECTRA ET LE HEP CHALLENGE

Nous analysons le déroulement de deux challenges hybrides. Ce type de challenge nous permet d'analyser séparément les phases fermées et les phases ouvertes et d'étudier le comportement sur chacune de ces phases. De plus, cela permet d'étudier l'influence de la phase fermée sur la phase ouverte. Enfin, l'étude de deux challenges hybrides dans deux conditions distinctes (durée et nombre de participants variable, problème différent) permet de clarifier les phénomènes propres à chaque contexte et les conclusions que l'on peut induire de notre étude de cas. Le premier challenge Drug Spectra s'intéresse à la prédiction des caractéristiques des médicaments avant injection pour réduire les risques d'erreur. Un deuxième challenge, le High Energy Physics (HEP), cherche à détecter les potentielles erreurs qui interviennent lors de la collecte de données sur les collisions de particules dans le cadre de la physique des particules. Nous présentons tout d'abord quelques indicateurs statistiques clés du déroulement des challenges incluant le nombre de participants, de soumissions, la moyenne des scores, la distribution du score, l'évolution du score au cours du temps.

2.1. PRESENTATION DES CHALLENGES

Drug Spectra challenge Pour éviter les médicaments inappropriés, la réglementation française impose la vérification des médicaments anticancéreux avant leur administration à un patient. Le but de ce challenge est de vérifier que les emballages tels que les sacs, les diffuseurs ou les seringues contiennent les bons agents chimiothérapiques avec la bonne proportion. L'ensemble des données utilisées est basé sur un total de 360 spectres mesurés pour 4 types d'agents chimiothérapeutiques utilisant la spectroscopie Raman (Butler et al., 2016). De plus amples informations sur l'organisation, les données et les objectifs scientifiques et l'historique du challenge sont disponibles sur le site Web de RAMP. Le starting kit a été construit par les administrateurs du système RAMP en collaboration avec des scientifiques du domaine de la chimie analytique. Le starting kit combine deux *pipelines* (séquences de procédé et d'analyse des données) correspondant à deux problèmes de prédiction distincts. Le premier problème est un problème de classification où les molécules sont prédites avec un spectre. Le second problème est un problème de régression où la concentration de la molécule est prédite à partir de son spectre. Les deux pipelines consistent en un module d'extraction de caractéristiques suivi d'un module de prédiction.

HEP Challenge Une tâche courante en physique des hautes énergies consiste à identifier les événements ou les ensembles de données qui diffèrent de ceux rencontrés normalement. Par exemple, deux essais enregistrés le même jour avec des conditions d'accélérateur et de détecteur identiques ne doivent pas être distingués statistiquement. Si tel est le cas, un effet systématique inattendu doit être présent pour agir sur chaque événement ou sous-ensemble des événements, ce qui entraîne une anomalie. Plusieurs raisons peuvent être à l'origine de telles anomalies: par exemple, le logiciel d'acquisition ou de reconstruction de données peut être mal configuré ou une

sous-composante du détecteur peut présenter des dysfonctionnements. À l'inverse, un ensemble de données par ailleurs considérées comme normales peut contenir des événements individuels inhabituels. Ces anomalies ponctuelles peuvent survenir suite à un problème avec le détecteur, à l'acquisition de données, à une reconstruction qui ne se produit que dans de rares circonstances. Un modèle prédictif capable d'analyser automatiquement les jeux de données entrants, de détecter toute caractéristique anormale et d'alerter un être humain pour qu'il puisse effectuer une investigation détaillée serait utile. Ce type de problème peut être considéré comme un problème de classification. Une version modifiée du jeu de données provenant du challenge HiggsML (Adam-Bourdarios et al., 2015; Kazakçi, 2015b) a été utilisée dans le challenge. Le starting kit contenait un pipeline initial comprenant un extracteur de fonctionnalités et un classificateur.

2.2. LA CONFIGURATION DU CHALLENGE: PASSER DE FERME A OUVERT

Dans les deux challenges, les participants ont reçu au début un starting kit. Ce kit a fourni à tous les participants des explications détaillées sur le problème, les données et le pipeline initial. Les deux challenges comprenaient deux phases. Dans la phase fermée, les participants ont soumis leur code au système mais n'ont pas accès au code des autres participants. Ils pouvaient néanmoins voir leurs scores et leurs positions dans le classement. Dans la phase ouverte, toutes les soumissions ont été ouvertes à tous les participants (y compris les nouvelles soumissions pendant la phase ouverte). Par conséquent, tout participant peut visualiser toutes les soumissions précédentes de tous les participants. Cela donne aux participants la possibilité d'examiner et de réutiliser le code des autres participants. Au total, 186 étudiants ont participé aux challenges (voir les **Figures** pour des statistiques récapitulatives). Tous étaient des étudiants en sciences de données de niveau master ayant des connaissances en mathématiques ou en informatique. Les participants ont été fortement incités car ils ont reçu des notes en fonction de leur performance en phase fermée et de leur activité en phase ouverte. Tous les modèles ont été formés par le système pour calculer leurs scores de prédiction. Dans le système, un algorithme supplémentaire est utilisé pour combiner certains des modèles afin de créer des prédicteurs encore meilleures; il s'agit d'une nouvelle utilisation appelée « assembling » dans la littérature en machine learning (Zhang & Ma, 2012). Les modèles à combiner sont sélectionnés en fonction de l'amélioration qu'ils apportent à au modèle d'ensemble. Cela apporte un score supplémentaire à chaque soumission, appelé contribution et affiché comme une colonne supplémentaire dans le classement. Afin de faciliter l'analyse des processus et de fournir des informations supplémentaires dans la phase ouverte, les participants ont été incités à créditer (en pourcentage) les modèles précédents utilisés suivant l'influence qu'ils ont eu sur leur soumission. Pendant le challenge, la plateforme enregistre également plusieurs événements clés et des informations générales sur les participants et ses actions sur la plateforme. Ces événements incluent un clic sur un lien particulier ou le chargement d'une page Web. Tous ces éléments sont associés à un utilisateur et à un horodatage. Toute action que l'utilisateur peut effectuer au-delà du système RAMP, telle que la navigation Web ou le codage dans différents éditeurs, n'est pas suivie.

3. COMPORTEMENT DES PARTICIPANTS DURANT LES PHASES FERMEES ET OUVERTES

3.1. DES COMPORTEMENTS INDIVIDUELS DIFFERENTS ENTRE LE MODE FERME ET LE MODE OUVERT

La **tableau 9** donne un aperçu des phases fermées et ouvertes des deux défis en observant l'activité exercée ainsi que les évolutions de score. Le déroulement de chaque phase est présenté d'un point de vue global mais également en analysant les participants qui ont au moins une soumission avec un score entre 90 et 100% du score maximum durant la phase. Pour une lecture plus facile, le score est normalisé sur une échelle de 100 entre le score du kit de départ et le meilleur score global du challenge. L'analyse de la participation montre que plus de participants ont été plus actifs pendant la phase fermée (pourtant plus courte dans les deux cas) que la phase ouverte et ont soumis plus de solutions (**tableau 9**). Cela n'est pas surprenant puisque 80% des notes proviennent de la performance individuelle pendant la phase fermée. Les statistiques sur l'activité des participants montrent également une baisse de l'activité moyenne par participant durant la phase ouverte. Cela est d'autant plus flagrant que la phase ouverte prend en compte le clic sur la soumission des autres. De manière générale, la tendance dans les deux challenges est à une baisse globale de la participation durant la phase ouverte (moins d'activité, moins de participants, moins de soumissions).

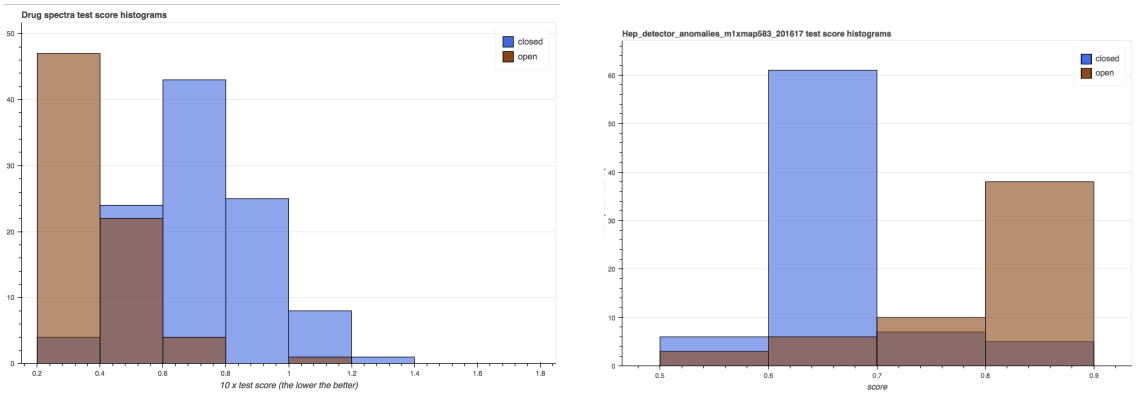
	Drug Spectra 2 challenge Durée (du 13/12/16 au 03/01/17)				**HEP challenge** Durée (du 02/01/17 au 17/01/17)			
	Phase fermée 4 jours		Open Phase 15 jours		Phase fermée 7 jours		Phase ouverte 8 jours	
	global	top 10% score*	global	top 10% score*	global	top 10% score*	global	top 10% score*
Participation								
Nombre de participants actifs	107	8	75	47	79	8	56	26
No soumissions	316	25	151	116	148	22	97	56
Soumissions par participant	3,0	3,1	2,0	2,5	1,9	2,9	1,7	2,2
Activités								
Nombre d'événements enregistrés	9573	1291	6295	4175	6142	1509	4816	2553
Activité moyenne par participant	89,5	161,4	83,9	88,8	77,7	188,6	86,0	96,2
Performance								
Max	94,9	94,9	100	100	79,7	79,7	100	100
Moyenne	47,9 (E.T. 22,6)	75,6 (E.T. 17,0)	84,7 (E.T. 24,1)	93,6 (E.T. 2,2)	40,3 (E.T. 15,9)	62,3 (E.T. 18,2)	77,3 (E.T. 25,3)	89,6 (E.T. 17,8)
Médiane	49,7	88,7	90	93,4	39	73,4	83,9	97,5

* Les participants avec au moins une soumission avec un score supérieur à 0,9 * (best_score_phase) au cours de la phase

Tableau 9. Analyse statistique des challenges Drug Spectra et HEP

En terme de performance, la plupart des améliorations du score maximum ont été réalisés pendant la phase fermée, néanmoins la phase ouverte a permis des progrès significatifs, notamment au regard du nombre de soumissions. A noter que pour le challenge HEP, la phase ouverte a permis d'améliorer le score maximum de plus de 20%. Nous observons deux comportements différents en terme de score entre la phase fermée et la phase ouverte. Durant les phases fermées, les scores sont bas par rapport à la tête de classement (moyenne et médiane basses par rapport au score max). Le nombre de participants dans le top 10% est faible dans les deux challenges. Durant la phase ouverte en revanche, on constate une amélioration considérable du score moyen : les scores des soumissions sont plus proches de la meilleure soumission et le nombre de participants dans le top 10% est 3 à 6 fois plus important. Manifestement, l'accès aux solutions durant la phase ouverte a eu un impact sur les soumissions des participants.

Les **figures 32 et 33** montrent la répartition du score suivant la phase du challenge. L'histogramme des scores dans la phase fermée (bleue) est approximativement gaussien, alors qu'il devient fortement asymétrique dans la phase ouverte (marron), indiquant que le gros des participants «rattrape» le meilleur score des meilleurs experts. Dans les deux cas, les participants bénéficient de l'ouverture en termes d'amélioration du score. Lorsque le challenge est ouvert, chaque participant peut réutiliser des éléments entiers ou partiels des soumissions précédentes et l'inclure dans sa prochaine soumission. Cependant, les notes seules ne permettent pas de distinguer les soumissions qui copient des soumissions antérieures de celles qui l'ont transformée, soit en combinant différentes soumissions, soit en modifiant certaines parties. L'amélioration du score après l'ouverture laisse supposer qu'au moins une partie des soumissions sur la phase ouverte développe de nouvelles idées. Pour mieux comprendre cet impact, nous regardons l'évolution temporelle du score entre la phase fermée et la phase ouverte.



Figures 32 et 33. Distribution du score suivant la phase fermée (bleu) ou ouverte (marron) – Drug Spectra (à gauche) et HEP (à droite) – Le score est en abscisse et le nombre de soumissions en ordonnées.

3.1.1. Evolution du score

Les **figures 34, 35, 36** présentent l'évolution des scores au cours des phases des challenges. La ligne rouge correspond à l'évolution du meilleur score global et les lignes pointillées associée aux points bleus représentent les progressions individuelles pour chaque participant.

3.1.1.1. Drug Spectra

Durant la phase fermée, un grand nombre de solutions ont été soumises (les points bleus) sans amélioration significative du score moyen entre le début et la fin. L'exploration est assez erratique et la répartition en terme de score est importante. La soumission étant autorisée toutes les 12 heures, les points ont tendance à se regrouper verticalement et cinq cycles de soumissions sont visibles. Le score de la meilleure soumission augmente régulièrement durant le challenge, particulièrement durant la dernière journée où le score a été battu 4 fois. Durant la phase ouverte (**figure 35**), on note immédiatement la baisse importante du nombre de soumissions. Celles-ci ont également tendance à être réparties dans le temps. Une salve de soumissions est visible juste après l'ouverture qui mène à une amélioration immédiate du score. De manière générale, les soumissions (points bleus) sont plus proches de la ligne rouge que dans la phase fermée.

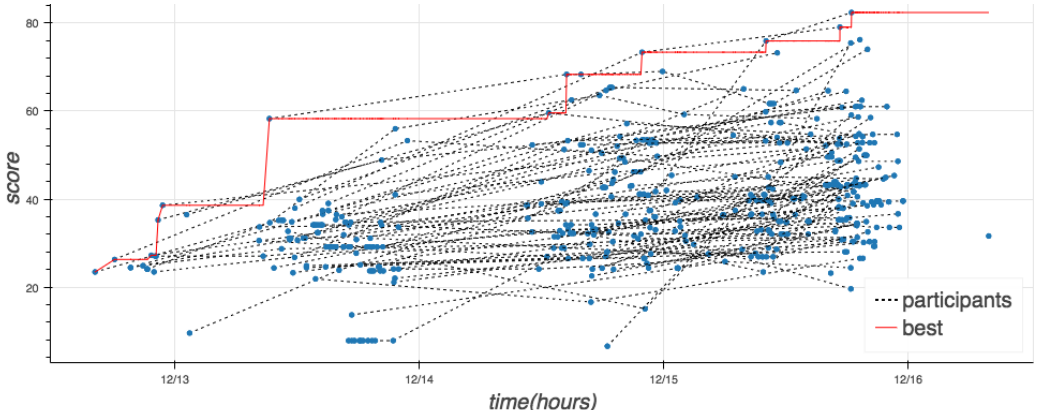


Figure 34. Drug Spectra - Evolution des scores durant la phase fermée

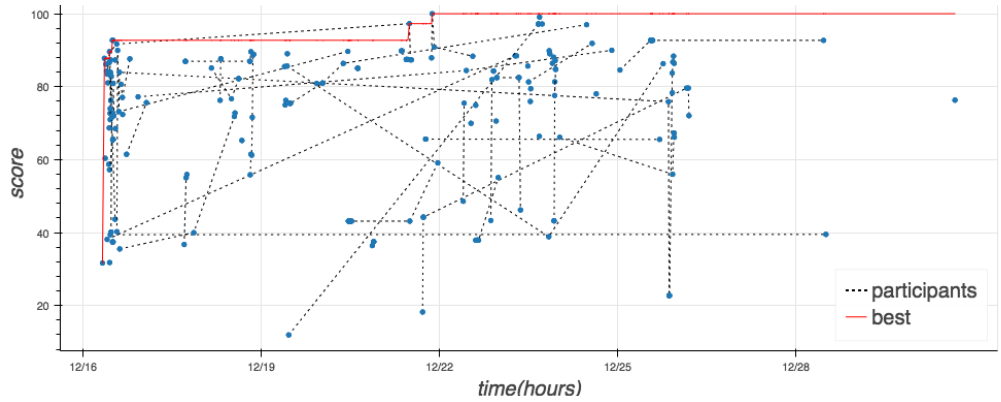


Figure 35. Drug Spectra - Evolution des scores durant la phase ouverte

3.1.1.2. HEP challenge

La **figure 36** montre l'évolution du score au cours des phases du challenge HEP (le moment de l'ouverture est représenté par la ligne verticale). Contrairement au challenge Drug Spectra, les soumissions sont moins erratiques en phase fermée. Un seuil à environ 73% sépare les soumissions en trois groupes : sous le seuil, au niveau du seuil, au-dessus du seuil. Les scores sous le seuil sont distribués également au-dessous du seuil. Il y a ensuite un "saut" en terme de score d'au moins 10 points pour les soumissions au-dessus du seuil. Après l'ouverture, une salve de soumissions atteint le niveau précédemment établi par les meilleures soumissions en phase fermée. La diminution du nombre de soumissions pendant la phase ouverte est moins visible que pour le challenge Drug Spectra.

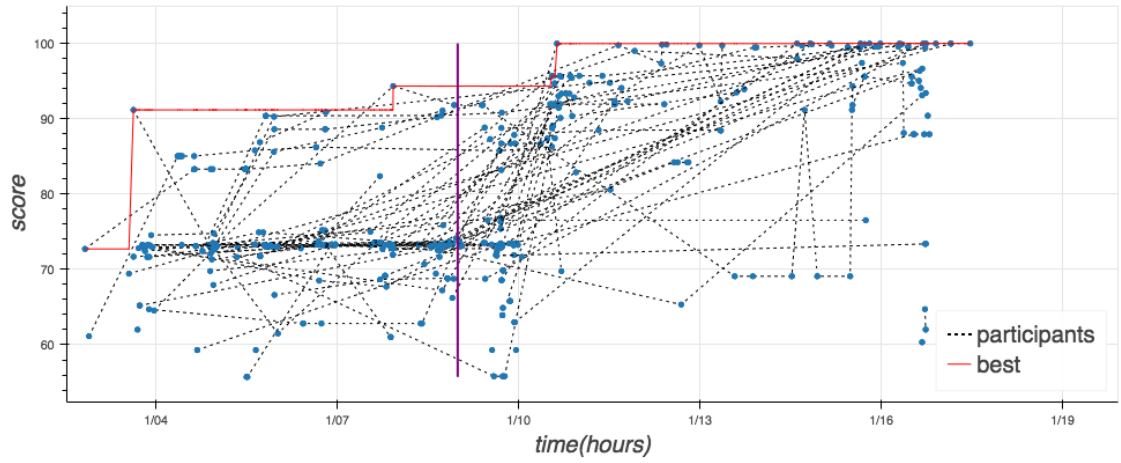


Figure 36. HEP - Evolution des scores durant la phase fermée et la phase ouverte

3.1.2. Phases fermées et ouvertes

Dans les phases fermées, les chemins suivis par les participants sont dans les deux cas erratiques et moins réguliers que dans la phase ouverte. Les scores ne semblent pas suivre une progression cohérente mais font plutôt l'objet d'une logique stochastique sans comportement global commun. La distribution des scores montre une répartition sous forme de gaussienne avec une majeure partie des soumissions ayant eu un score moyen et peu de scores très élevés. Chaque participant explore individuellement l'espace du code informatique et suit une trajectoire individuelle, de l'ordre de 2 à 3 soumissions par participant. Cependant on n'observe pas de comportement global cohérent, comme l'amélioration significative de la moyenne ou de la médiane du score entre le début et la fin de la phase.

Dans les phases ouvertes, on observe une amélioration nette du score moyen qui se vérifie dans les courbes de répartition du score. Les soumissions sont de meilleure qualité en terme de score avec une valeur moyenne proche du meilleur score obtenu. Contrairement à la phase fermée, les

scores des soumissions sont plus concentrés autour d'une valeur. Dans le même temps, on observe une baisse de la participation, avec une diminution d'environ 25% du nombre de participants actifs. Cela s'explique en partie dans notre cas parce que la note des étudiants se mesure principalement durant la phase fermée.

3.2. *Les meilleures soumissions comme standard pour la réutilisation*

A quel point les soumissions précédentes ont un impact sur un participant lorsque celui-ci propose une nouvelle solution durant la phase ouverte ? Nous analysons les liens entre une soumission et les soumissions précédentes grâce à deux indicateurs : le crédit alloué par le participant aux soumissions précédentes et le fait que le participant ait cliqué sur une soumission. Les **figures 37, 38** illustrent les liens entre les soumissions au cours de la présentation précédente et le crédit accordé à une soumission. Les soumissions sont rangées dans l'ordre chronologique en abscisse. Un arc bleu signifie que le participant a consulté une soumission antérieure. Un arc rouge signifie que le dernier participant a crédité une soumission antérieure. L'épaisseur de l'arc rouge indique le crédit alloué par le participant. Les graphes nous permettent d'observer de manière qualitative et globale les comportements des participants durant la phase ouverte et la phase fermée. Nous analysons également le code des 12 participants du top 10% dans le challenge HEP ainsi que toutes les soumissions sur lesquelles ils ont cliqué pour comprendre quel comportement a été adopté.

3.2.1. *Drug Spectra challenge*

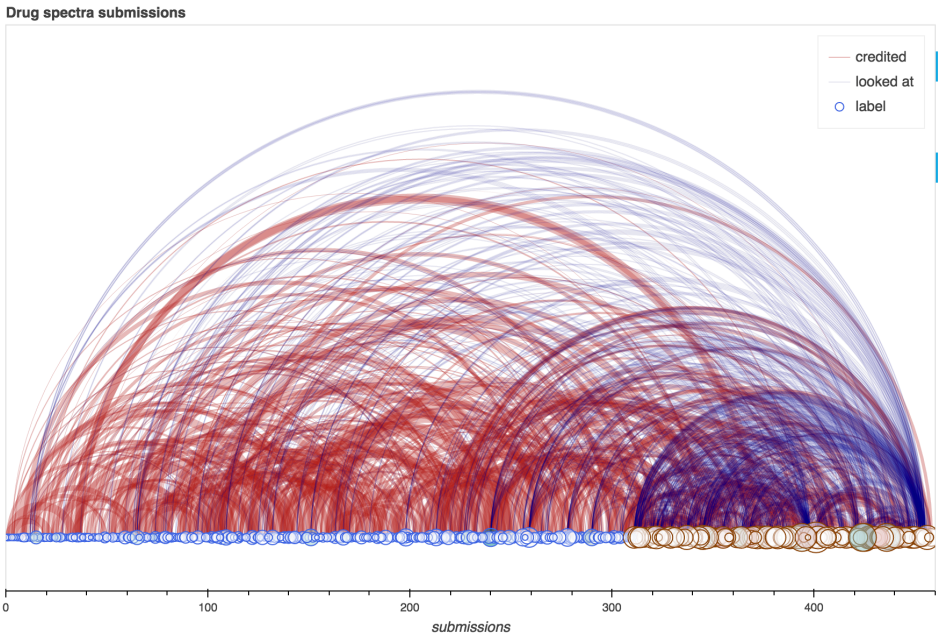


Figure 37. Drug Spectra - Liens entre les soumissions au cours de la phase précédente et le crédit accordé à une soumission

La grande concentration des arcs bleus à la fin du graphe montre que la plupart des solutions regardées ont été produites durant la phase ouverte (le nombre d'arcs bleus reliant la phase fermée à la phase ouverte est très faible). Les arcs rouges durant la phase fermée indiquent simplement que les participants se sont inspirés de leurs propres soumissions : en effet, ils ne peuvent pas voir la soumission des autres durant cette phase. A noter qu'il n'existe pas dans le graphe de façon claire de « nœuds », c'est-à-dire une ou plusieurs soumissions qui concentrent un grand nombre d'arcs de cercles (bleus ou rouges). Cela signifie que les participants ne se sont pas concentrés exclusivement sur une seule soumission de référence. En fait, les arcs sont généralement courts en distance, ce qui laisse supposer que seules les soumissions les plus récentes sont regardées et créditées par les participants. L'étude quantitative des indicateurs de « crédit » et de « clic sur soumission » le confirme.

Les soumissions les plus regardées par les participants sont la 311, 337, 345, 259, 368. Elles présentent plusieurs particularités : elles ont quasiment toutes été soumises durant la phase ouverte (sauf la n° 259) ; elles ont les meilleurs scores du classement ; elles font partie des soumissions les plus créditées (entre 6 et 18 fois par soumission). En conséquence, la réutilisation durant la phase ouverte concerne majoritairement les soumissions ayant le meilleur score et les plus récentes. Dès qu'une soumission a un meilleur score que celles précédente, elle devient la nouvelle référence temporaire pour les participants. En fait dans le cadre de la phase ouverte, les participants font évoluer leurs stratégies individuelles en fonction d'un ensemble restreint de soumissions qui vont « nourrir » leur exploration. A noter que la soumission n° 259 est celle qui a le score le plus élevé durant la phase fermée et a été par la suite la soumission de référence : c'est elle qui est le plus regardé par les participants dès le début de la phase d'ouverture (avant d'être remplacée par la soumission 311 qui a un meilleur score).

3.2.2. HEP Challenge

La **figure 38** montre une densité globale plus faible d'arcs bleus et rouges par rapport au challenge Drug Spectra. Cela provient d'un plus faible nombre de soumissions enregistrées ainsi que d'activités durant la phase ouverte. Une deuxième différence par rapport à Drug Spectra est la présence de nœuds d'arcs bleus et rouges, c'est-à-dire l'existence de soumissions avec un nombre significatif de crédits et de clic par rapport à la moyenne des autres. La soumission 165 par exemple, qui a réalisé le meilleur score juste après l'ouverture compile la plus grande somme de crédits alloués (créditée 24 fois). Contrairement au challenge Drug Spectra, il existe plusieurs de ces nœuds issus de la phase fermée. En fait, il semble que la progression générale dans le HEP challenge est moins linéaire que dans le cas de Drug Spectra. Les participants ne réutilisent pas nécessairement les solutions les plus récentes (4 solutions sur les 5 les plus créditées sont issues de la phase fermée – n° 104, 66, 137, 124) : cela s'observe notamment par une plus grande densité d'arcs de grande taille.

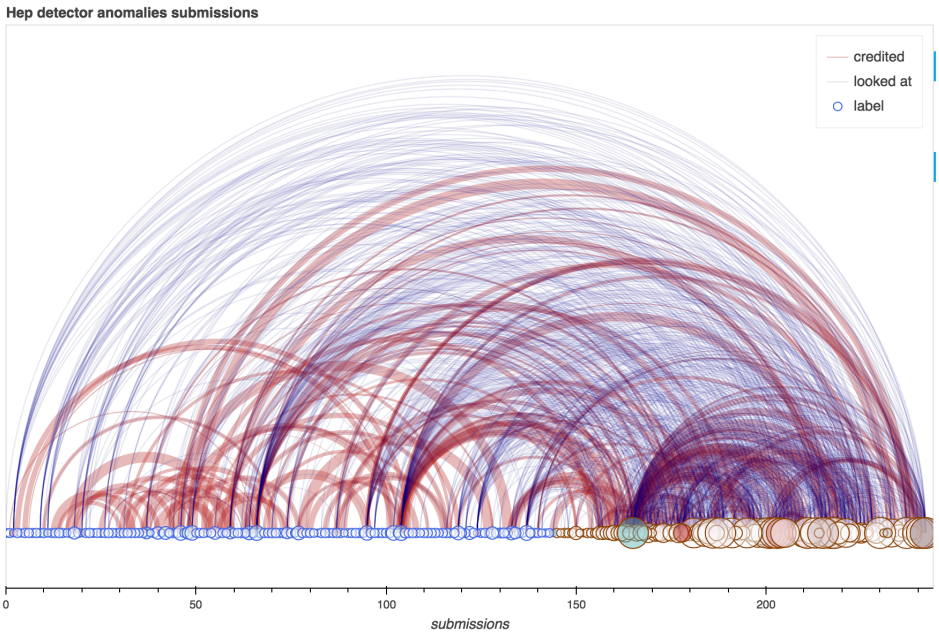


Figure 38. Drug Spectra - Liens entre les soumissions au cours de la phase précédente et le crédit accordé aux soumissions

L'analyse des crédits et des clics sur les soumissions montre une différence de stratégie d'exploration entre les deux challenges. Comme nous l'avions suggérée, cette différence peut s'expliquer par différents facteurs liés au problème ou aux participants. En revanche, les deux challenges sont similaires en deux points. D'abord, les soumissions examinées par les participants sont principalement celles ayant le meilleur score au cours du processus. Ensuite, l'activité durant la phase fermée a une influence sur les stratégies d'exploration durant la phase ouverte. En effet, dès l'ouverture les participants regardent les meilleures soumissions de la phase fermée et les réutilisent pour proposer des solutions dans la phase ouverte. Certaines soumissions deviennent des références ou des standards pour les participants étant donné leur score. Chaque soumission ultérieure est généralement associé à un clic sur cette soumission ainsi qu'un crédit alloué par les participants.

3.3. TYPE DE REUTILISATION DU CODE EN PHASE OUVERTE : LE CAS DE HEP CHALLENGE

Afin de comprendre comment les participants réutilisent les soumissions, nous avons analysé avec l'aide d'un data scientist de l'équipe du RAMP le code des soumissions examinées par les participants qui ont eu les meilleurs scores finaux durant le HEP challenge. Nous avons établi les liens entre les différentes soumissions de chaque participant durant la phase fermée et la phase ouverte et les soumissions créditées et regardées.

```python
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.preprocessing import Imputer
from sklearn.pipeline import Pipeline
from sklearn.base import BaseEstimator

class Classifier(BaseEstimator):
    def __init__(self):
        self.clf = Pipeline([('imputer', Imputer(strategy='most_frequent')),
        ('rf', AdaBoostClassifier(base_estimator=RandomForestClassifier(max_depth=5,
n_estimators=100),
                                  n_estimators=20))])

    def fit(self, X, y):
        self.clf.fit(X, y)

    def predict(self, X):
        return self.clf.predict(X)

    def predict_proba(self, X):
        return self.clf.predict_proba(X)

```

Figure 39. Exemple de code soumis par un participant dans la plateforme RAMP.

Le code analysé est une succession de lignes écrites par le participant dans le programme suivant un langage de programmation (Python). Chaque ligne correspond à une action que le programme exécute dans l'ordre **(figure 39)**. Une fois toutes les lignes exécutées, le programme fournit le modèle de prédiction. La comparaison entre les codes est facilitée par une structure imposée par les organisateurs. Dans l'analyse nous avons distingué les situations où les lignes de code diffèrent dans leur écriture tout en ayant un effet strictement identique. En effet, une même action peut être écrite de plusieurs manières différentes.

Sur les 12 codes des soumissions des top 10% pendant la phase, nous avons observé trois comportements différents des participants: la *copie* d'un code à l'identique ; la *combinaison* de différents codes pour proposer une nouvelle solution ; la *transformation* d'un code existant. La transformation peut être comprise comme l'ajout de certaines fonctions dans le code ou le choix d'en utiliser d'autres, ou encore de modifier des paramètres. La plupart des participants (9 sur 12) proposent de nouvelles soumissions, soit en combinant différents éléments de codes soit en transformant le code existant. Nous constatons qu'une grande partie des soumissions des participants durant la phase ouverte provient de la combinaison de soumissions précédentes associé à une variété de transformations : modification des paramètres des méthodes, changement de l'ordre d'exécution. Les résultats sont similaires à ceux observés dans l'étude menée par Boudreau et Lakhani et la plupart des bonnes soumissions sont des combinaisons de familles de méthodes déjà proposées dans des soumissions précédentes.

4. PROCESSUS D'EXPLORATION ET IMPACT DE LA PHASE FERMEE SUR LA PHASE OUVERTE

Nous avons clarifié le processus de réutilisation durant la phase ouverte. Cependant, l'impact de la phase fermée sur la phase ouverte reste peu clair. Dans cette section, nous analysons ce lien entre les phases afin de comprendre pourquoi les challenges hybrides sont plus performants que les challenges uniquement ouverts. Pour traiter cette question, nous nous intéressons au comportement du code en se basant sur le vecteur de prédiction du code $\hat{y}$. Cette analyse permet de distinguer les codes suivant s'ils se comportent de façons similaires ou différentes sur la base de leur prédiction. En effet, deux codes peuvent avoir des scores très proches (résultat à la fonction *test()*), mais pour lesquels leur efficacité se mesure sur des éléments différents de la base de données.

Nous utilisons un algorithme appelé t-SNE qui permet d'étudier le comportement des codes dans un espace à 2 dimensions (Van Der Maaten & Hinton, 2008). Afin de comprendre le fonctionnement de l'algorithme, il faut se représenter le vecteur de prédiction $\hat{y}$ obtenu par le code comme un vecteur colonne dont chaque variable est la prédiction d'un objet d'une base de données. Pour illustrer, si l'objet à prédire est la présence ou non d'un chat sur une image, chaque variable de l'algorithme correspond à la probabilité pour une image que celle-ci comporte effectivement un chat. Le principe du t-SNE est de projeter ce vecteur de prédiction dans un espace à deux dimensions, appelé espace de comportement du code, qui peut ensuite être facilement représenté graphiquement. L'algorithme applique une fonction de projection $P : H \rightarrow R^2$, où H est un espace à n dimensions (n étant le nombre de prédictions, d'images dans le cas du chat) vers un espace à 2 dimensions. Dans cette représentation, plus les points sont éloignés les uns des autres, plus leur façon de prédire est différente (et réciproquement). Nous pouvons ainsi déterminer facilement si les codes ont un comportement différent les uns des autres ou au contraire similaire.

Dans la représentation graphique les points bleus sont les soumissions de la phase fermée et les points oranges celles de la phase ouverte. La taille de chaque point est proportionnelle à son score (plus le point est gros, plus le score est important).

4.1. APPARITION DE PLATEAUX DE FIXATION DANS LA PHASE OUVERTE

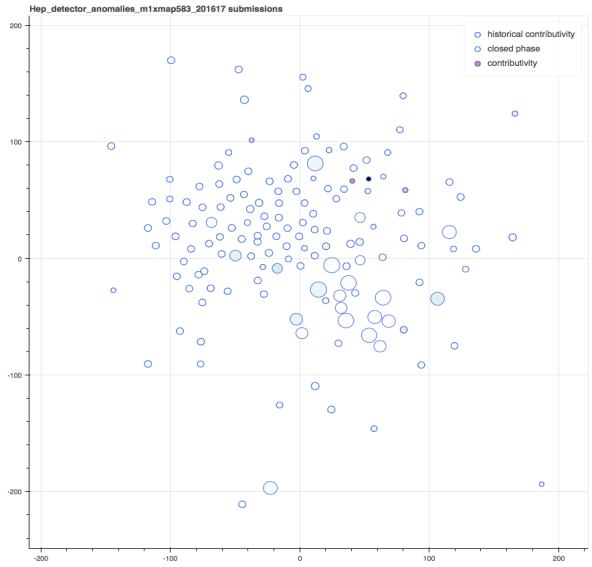


Figure 40. HEP - Espace du comportement des codes à la fin de la phase fermée

La répartition des points dans la **figure 40** montre que les algorithmes soumis ont des comportements hétérogènes en terme de prévisions sur le même ensemble de données. Cette large répartition met en avant une *diversité* de l'exploration par l'ensemble des participants qui n'était pas visible lors de l'analyse du score. Chaque code a sa spécificité propre qui le distingue des autres codes. La diversité des soumissions provient du fait que chaque participant travaille de façon individuelle durant la phase fermée. Le processus d'exploration n'est donc pas influencé par les propositions des autres. La recherche de la solution peut être considérée comme une *recherche locale*, c'est-à-dire qu'elle dépend entièrement des caractéristiques du participant (Afuah & Tucci, 2012). A noter que deux groupes de points se distinguent en fonction de leur taille (leur score). Alors que la majorité des soumissions ont des scores bas représentés par les petits points, une douzaine de soumissions ont un score similaire plus élevé. Ce constat est compatible avec l'effet de seuil que nous avions identifié dans l'étude de l'évolution du score. La grande partie des gros points est concentrée dans une zone de l'espace, ce qui indique que ces soumissions ont des approches algorithmiques similaires avec le même ensemble de données. Les points isolés quant à eux peuvent être interprétés comme des solutions originales.

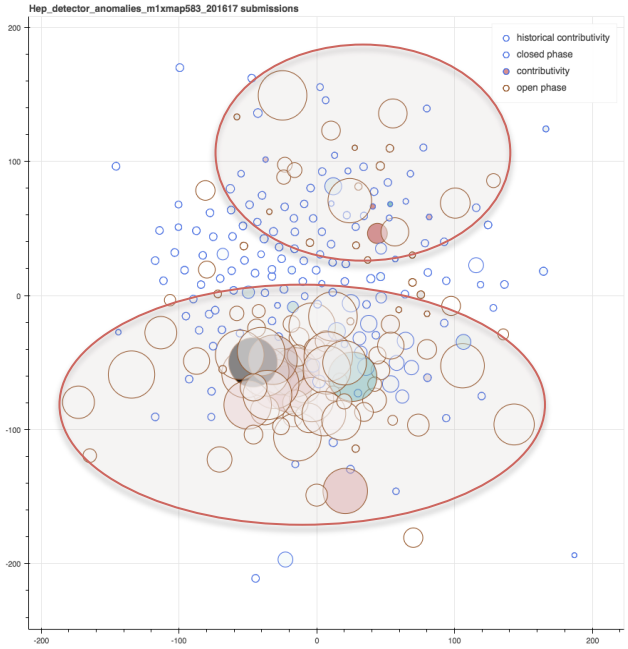


Figure 41. HEP. Espace du comportement à la fin de la phase ouverte

Dans la **figure 41,** la majeure partie des soumissions de la phase ouverte est concentrée autour d'une zone spécifique. Ces soumissions ont toutes des scores globalement élevés caractérisés par la taille des points. A noter que certains points avec de hauts scores sont isolés, laissant suggérer des approches originales de la part des participants. La concentration des points montre une similarité dans leur comportement : contrairement à la phase fermée où la diversité est importante, la phase ouverte semble faire *converger* les solutions vers des zones de l'espace que l'on définit comme des **plateaux de fixation**. Ce constat est confirmé par les observations réalisées par Boudreau et Lakhani ainsi que par le challenge Drug Spectra (**figures 42, 43**). Les auteurs suggèrent que cette convergence s'explique par une économie en terme de coût pour les participants et permet d'atténuer les incertitudes dans le processus d'exploration, créant des incitations à reproduire ce qui a déjà fonctionné. L'ouverture des résultats intermédiaires incite les participants à réutiliser des solutions qui ont de bons scores et limiter les expérimentations indépendantes.

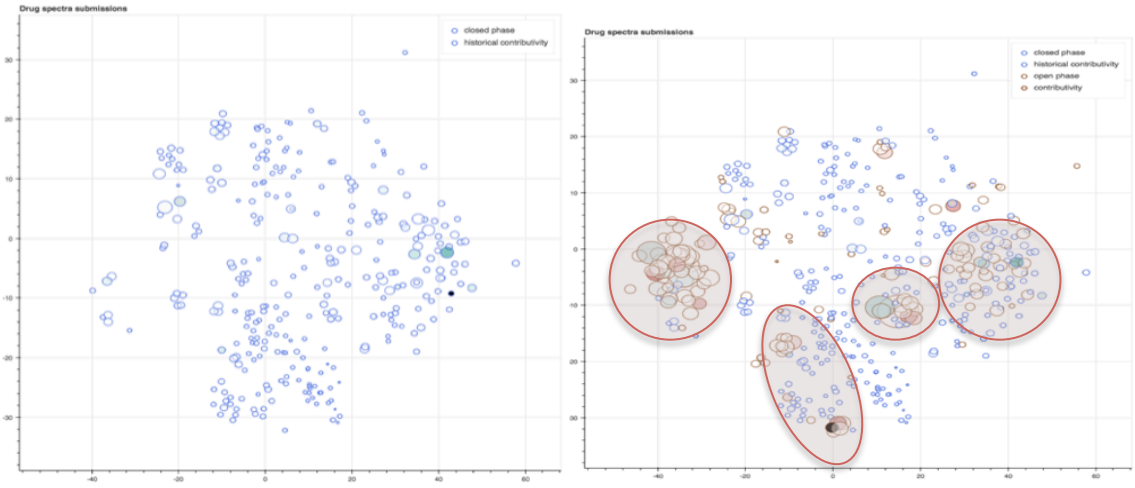


Figure 42 et 43. Emergence de plateaux de fixation dans le challenge Drug Spectra

4.2. Les limites du processus d'exploration en phase ouverte

Le mode ouvert fait apparaître un type d'exploration où seules certaines solutions sont privilégiées. Or, ce type de comportement peut mener à une exploration très localisée de l'espace des solutions et donc augmente les risques de passer à côté de meilleures solutions. Nous illustrons ce phénomène au travers de la fonction de valeur des solutions de l'espace. Dans cette représentation, la fonction de valeur est une fonction continue dans laquelle les solutions voisines en terme de code ont des scores similaires. Supposons que plus la valeur est basse, meilleur est le score de la soumission. Au fur et à mesure du challenge, plusieurs solutions sont proposées par les participants (représentées par les points bleus sur la courbe). Le code de chaque soumission est visible par les participants, ainsi que son score (visible sur le tableau de classement). Dès qu'une solution obtient un bon score, elle devient une référence pour les autres participants. Ainsi la majeure partie des soumissions ultérieures s'inspire de celle-ci, car cela réduit leur coût d'exploration ainsi que l'incertitude quant à la réussite d'une solution. Les participants vont donc explorer dans le voisinage de la meilleure solution, créant un plateau de fixation qui mène à un optimum local de la fonction de valeur. Or cette concentration de l'effort fourni autour du plateau de fixation réduit la probabilité de proposer une solution autre, et donc de tomber au voisinage de l'optimum global.

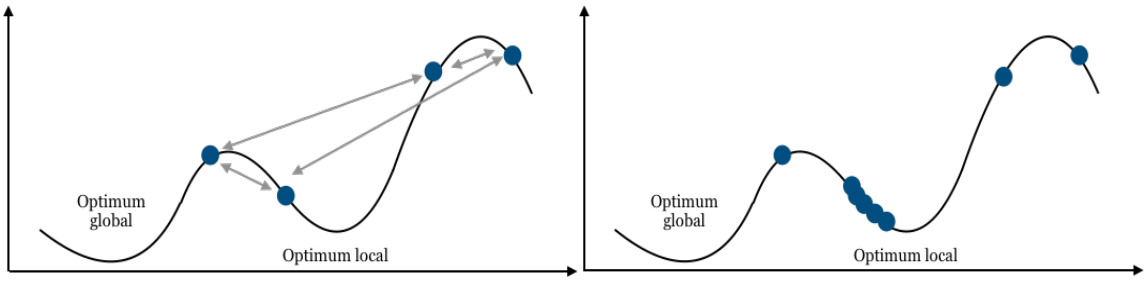


Figure 42 et 43. Début de l'exploration (à gauche) et création d'un plateau de fixation (à droite)

Dans les différentes expérimentations menées dans le cadre du RAMP ainsi que dans le conclusions de Boudreau et Lakhani, on constate pourtant que de manière générale ce mode est plus performant que le mode fermé. Cette comparaison est le reflet de la performance de deux modèles d'exploration de l'espace. D'un côté, le mode fermé est basé exclusivement sur l'augmentation de la diversité en multipliant le nombre de points d'entrée dans l'espace des solutions. Chaque participant explore localement l'espace des solutions indépendamment des autres participants, augmentant la probabilité de tomber sur une bonne solution. De l'autre le mode ouvert concentre l'exploration sur quelques zones de l'espace et cherche à optimiser localement quelques solutions de référence. Au lieu que la recherche locale autour d'une bonne solution ne soit faite que par un participant, elle est déléguée à un grand nombre de participants qui explorent de façon répétée le voisinage de la solution, augmentant la probabilité de trouver la solution optimale dans le voisinage.

4.3. IMPACT DE LA PHASE FERMEE DANS LES CHALLENGES HYBRIDES

Si les conclusions sur le RAMP et sur les travaux de Boudreau et Lakhani sont cohérentes pour les challenges fermés et ouverts, elles divergent sur la performance des challenges hybrides. La plateforme RAMP a eu l'occasion de prouver à plusieurs reprises que les challenges hybrides étaient plus performants que les challenges uniquement ouverts, contrairement aux conclusions des deux auteurs. Comment expliquer cette différence de l'impact de la phase ouverte sur la phase fermée ? Durant la phase fermée, les participants explorent de façon indépendante un espace extensible. Chaque participant débute l'exploration à l'aveugle : il n'a pas moyen de savoir quel sera l'impact des actions qu'il va coder. Il n'a donc pas de moyens pour réduire le niveau d'incertitude de son exploration, et est plus susceptible de prendre des risques. Les participants sont donc plus susceptibles de travailler sur la partie extensible de l'espace et donc de détecter des « bonnes idées » comme nous l'avons présenté plus haut avec l'exemple de SPIPOLL.

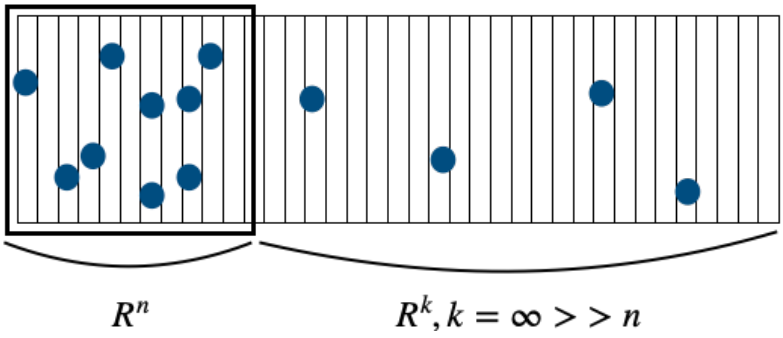


Figure 44. Exploration de l'espace en phase fermée : extension de l'espace

Ainsi, au moment de l'ouverture, les participants ont accès à plus de variétés dans les solutions qui peuvent être réutilisées. Ils augmentent les chances de converger vers un plateau de fixation autour d'un optimum meilleur que s'il n'y avait eu que la phase ouverte. En fait, la phase fermée « nourrit » la phase ouverte de solutions avec de bons scores et susceptibles de contenir des bonnes idées. Rapidement, les solutions convergent vers le plateau de fixation et réduisent la probabilité de créer une bonne idée.

Dans l'étude de cas de Boudreau et Lakhani, l'exploration de l'espace semble être limitée à 10 familles de méthodes et leurs variantes. Ainsi, la phase fermée ne permet pas une extension de l'espace comme nous l'avons montré dans le cas de SPIPOLL. De plus, nous pouvons constater que le nombre de participants actifs durant le challenge est faible par rapport au nombre de participants (43 pour 244), soit moins de deux fois le nombre de participants durant les challenges du RAMP. Ainsi, la probabilité de proposer des recherches locales durant la phase fermée avec de bonnes idées est plus faible. Malheureusement, les auteurs ne s'attardent pas sur ce cas qu'ils considèrent comme intermédiaire aux deux autres et il nous manque une analyse plus fine de l'exploration réalisée sur la partie extensive de l'espace.

4.3.1. Illustration avec le parcours du gagnant du challenge HEP

Pour illustrer l'importance de la phase fermée précédant la phase ouverte, nous prenons l'exemple du participant *P* qui a gagné le challenge HEP. La **figure 45** montre le parcours (en orange) suivi par *P* dès sa première soumission (trois au total). Le participant soumet deux solutions durant la phase fermée et une solution dans la phase ouverte (la meilleure du challenge en terme de score). Quatre soumissions sont créditées par le participant pour cette soumission (encerclées en violet) : sa propre soumission (la deuxième) ainsi que trois autres soumissions ayant eu un bon score durant la phase fermée. En analysant les différences entre les codes crédités et sa soumission, nous avons remarqué que sa meilleure soumission est une combinaison des quatre soumissions précédentes avec une modification des paramètres. *P* a choisi d'utiliser le classificateur de sa précédente soumission (combinaison d'un algorithme XGB et DMatrix) avec de nouveaux paramètres. C'est en effet le seul qui utilise ce classificateur parmi toutes les soumissions

créditées. *P* a également utilisé une combinaison des trois autres soumissions pour construire un algorithme de transformation des données.

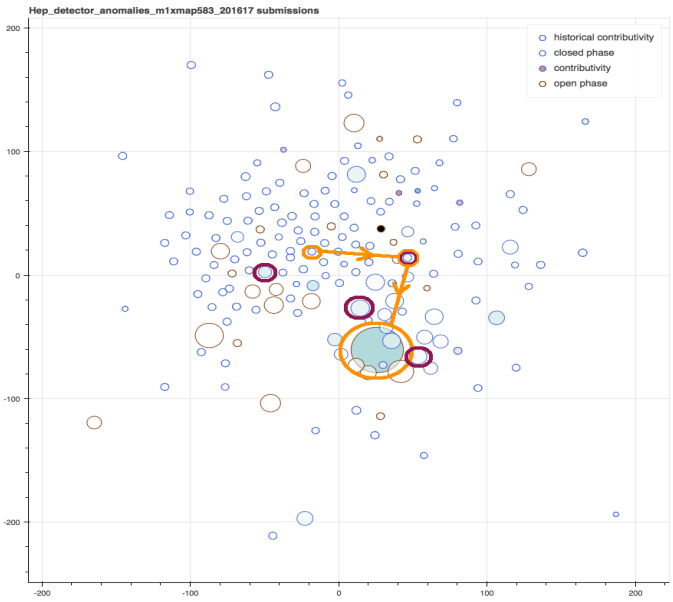


Figure 45. Trajectoires des soumissions du participant *P*

5. GERER LA CAPITALISATION DANS LA RESOLUTION DE PROBLEMES

5.1. EFFET DE LA CAPITALISATION CROISEE SUR LES METRIQUES DE PERFORMANCE

Nous avons montré dans les chapitres précédents qu'un défaut du crowdsourcing est la hausse de la perte de production (proche de 100% - une seule solution utilisée sur toute la production). Nous avons proposé de réduire ces pertes en capitalisant sur ce qui est produit durant la tâche par un système de capitalisation croisée. Dans un premier cas, les participants peuvent voir et réutiliser la soumission des participants durant tout le challenge. Notre étude ainsi que celle de Boudreau et Lakhani montrent que ce mode donne de meilleurs résultats que le mode traditionnel (fermé). Au lieu que chaque participant évolue individuellement, l'effort d'une grande partie des participants est concentré sur quelques zones de l'espace, créant des plateaux de fixation. Pour un même coût de production et pour des délais similaires, nous montrons que le mode ouvert permet d'accéder à de meilleurs résultats.

		Modèle traditionnel	Modèle fermé (crowdsourcing ou appel au marché)	Modèle ouvert	Modèle hybride
Fiabilité	**Fiabilité du système**	Elevée (experts ou groupe d'experts identifiés)	Incertaine (Foule)		
	Gestion de la fiabilité	Niveau de qualification des experts	Redondance		
Coût	**Coût d'installation**	Conception du problème	Conception du problème, starting kit, plateforme, récompense financière		
	Coût de production	Salaires, machines,…	Serveurs (très faible)		
Délai (cadence)		Temps de l'exploration	Constitution de la communauté + temps de l'exploration		
Qualité		Dépend de l'expert et du problème (recherche locale)	Equivalent (Mollick & Nanda, 2016) ou meilleur que les experts (Afuah & Tucci, 2012)	Plus performant que le mode fermé (Boudreau & Lakhani, 2015)	Plus performant que le mode fermé (Boudreau & Lakhani, 2015) et généralement ouvert (RAMP)
Perte de production durant la tâche		Faible	(n-1)/n, (n nombre de participants)	Capitalisation croisée (≈96%*n pour HEP)	Capitalisation croisée (≈98%*n pour HEP)

Tableau 10. Synthèse des métriques de performance suivant les modes de délégation.

Nous avons également souligné que le mode ouvert présentait des limites en terme d'exploration de l'espace. En effet, la convergence vers des plateaux de fixation limite l'exploration de participants dans d'autres zones de l'espace et est susceptible de fixer les participants dans un extrema local de la fonction de valeur. Notre étude de cas sur le RAMP a montré que la mise en place d'un mode fermé avant le mode ouvert permet de multiplier la diversité de l'espace exploré et donc augmente les chances de trouver des zones de l'espace où la fonction de valeur est élevée. D'un point de vue de la qualité des résultats soumis, nous observons que les meilleures soumissions ont de meilleurs résultats que le mode fermé. De plus, l'étude répétée et empirique du RAMP a suggéré que les résultats obtenus étaient meilleurs que dans le mode ouvert.

La capitalisation croisée permet également de réduire la perte de production durant la tâche. Dans le cas du challenge HEP, quatre soumissions ont été utilisées par le participant *P* pour élaborer la solution avec le meilleur score. La capitalisation entre la phase fermée et la phase ouverte est donc de $\frac{4}{148+97} \sim \frac{1}{61}$, soit 98% de perte. Pour comparaison, la construction de la meilleure solution uniquement durant la phase fermée est de $\frac{1}{148}$ soit 96% de perte. Le taux de perte a donc diminué entre la phase fermée et la phase ouverte d'un facteur 2,5. Nous pourrions rétorquer qu'il reste 244 solutions non utilisées pour produire le résultat final. En fait, une grande partie de ces soumissions n'est pas forcément réutilisable : certaines ne sont que des copies d'autres

soumissions ; d'autres n'ont pas forcément de « bonnes idées » à extraires ; d'autres encore sont des variantes moins performantes de la meilleure soumission (effet de convergence dans la phase ouverte). De plus il est difficile d'estimer a priori si une solution intègre une bonne idée.

5.2. LIMITES DE LA CAPITALISATION SEQUENTIELLE DANS LES TACHES DE TYPE RESOLUTION DE PROBLEMES

Contrairement aux autres plateformes de data challenge comme Kaggle ou TopCoder, le RAMP a la particularité de répéter plusieurs fois les mêmes challenges, principalement dans un but éducatif et d'observation. Chaque challenge génère de nouveaux codes avec de meilleurs scores que le starting kit. Capitaliser sur ces codes permettrait d'améliorer au fur et à mesure les starting kits afin que chaque itération se base sur la dernière meilleure solution. Dans le cas de SPIPOLL par exemple, ajouter les lignes de code pour couper les bords de l'image dans le starting kit permettrait d'éviter qu'à chaque itération les participants doivent recommencer l'exploration. Or, il y a eu très peu de capitalisation sur le code durant les RAMP. Les solutions proposées durant chacun des challenges ne sont pas devenues les nouveaux starting kit et chaque challenge a été répété indépendamment des autres.

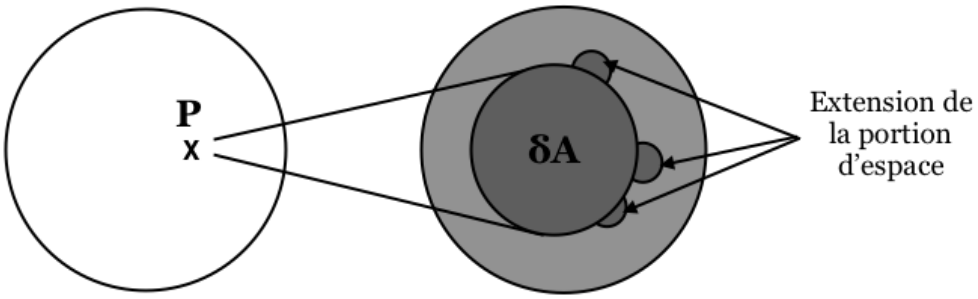


Figure 46. La définition du problème contraint la zone de l'espace d'action à explorer.

Une des raisons à cette non capitalisation est liée à la structure des challenges. Ces derniers sont construits comme des tâches de type résolution de problème définies par les scientifiques. Cela signifie que la tâche consiste à explorer un espace, ici celui du code informatique, en fonction du problème établi : un état final désiré e_{final}, une base de données ainsi qu'une fonction $test()$ pour évaluer la qualité des soumissions. Bien que la résolution du problème se situe dans un espace extensible, la plupart des explorations sont réalisées dans une portion finie de l'espace des actions (dans notre cas les familles de méthodes traditionnelles du machine learning). A noter que les participants peuvent explorer d'autres dimensions et étendre ponctuellement la taille de la portion de l'espace cependant plus les solutions proposées se rapprochent du score optimal, plus il devient difficile et coûteux de trouver des nouvelles familles de méthodes pour améliorer le score. En fait, la définition du problème spécifie la valeur attendue des solutions et borne l'espace d'action à explorer. L'exploration de l'espace des actions serait rapidement limité par

l'amélioration continue du starting kit et rendrait les challenges ultérieurs peu éducatif, complexe et peu incitatifs.

CHAPITRE 8 – PILOTAGE DE LA PERFORMANCE DES PROJETS DE SCIENCE CITOYENNE REPETES : LES DISPOSITIFS DE GESTION DE LA « CAPITALISATION SEQUENTIELLE »

RESUME DU CHAPITRE 8

Conformément à notre méthodologie de recherche, ce chapitre étudie Epidemium, un programme de recherche collaboratif basé sur l'épidémiologie, qui s'est déroulé entre novembre 2015 et mars 2018. Le programme a mis en place deux challenges successifs pour les participants, construits sur une base de données ouverte et massive (plus de 21 000 jeux de données). En s'appuyant sur le modèle formel que nous avons développé dans les chapitres 5 et 6, nous analysons Epidemium comme un projet de science citoyenne de délégation d'une tâche couplée, en l'occurrence la génération d'hypothèses scientifiques à partir de bases de données. Nous étudions le dispositif organisationnel mis en place par le programme ainsi que les stratégies d'exploration des participants pour construire les hypothèses et les vérifier. Nous nous intéressons particulièrement à la performance du programme ainsi qu'au système de capitalisation durant et entre la tâche.

Notre analyse montre que les stratégies d'exploration se basent sur des aller-retour entre l'espace des hypothèses et l'espace des plans d'action (ici le code informatique) qui sont parasités par la faibles qualité des données ouvertes. La difficulté de cette exploration dans deux espaces ainsi que la limite de temps du challenge limite la productivité des participants et les contraint à restituer non pas des résultats scientifiques finaux, mais des prototypes ou des produits intermédiaires. Nous définissons ces éléments comme des stepping stones, c'est-à-dire des étapes intermédiaires dans le processus d'exploration : ce sont des hypothèses inabouties (axe de travail), des algorithmes à améliorer, des bases de données nettoyées, ou des outils (ou prototypes) d'aide à l'exploration dans chacun des espaces. La capitalisation sur ces stepping stones permet alors de définir les prochaines actions à mener dans les challenges suivants et réduire les coûts d'exploration.

Dans le cas d'Epidemium, cette capitalisation a été « sauvage », c'est-à-dire sans gestion, où toutes les parties prenantes (financeurs, participants, organisateurs) ont capitalisé sur ce qu'ils ont appris durant le challenge. Ce type de capitalisation est limité pour plusieurs raisons : il ne propose pas de critères de comparaison entre les productions ; il ne prend pas en compte toute la production ; il est géré des acteurs potentiellement éphémères (les participants) et donc peu fiables. Au final, une grande partie de ce qui est produit durant le premier challenge a été perdu. Nous proposons pour limiter ces pertes d'intégrer des critères de valeur pour évaluer la production. Au lieu de recommencer chaque exploration de zéro, l'évaluation de la valeur de chaque stepping stone permet de cartographier les zones explorées dans les espaces et de décider des stratégies à adopter pour les prochains challenges.

Dans le chapitre suivant, nous analysons l'impact de la capitalisation sauvage sur le déroulement du challenge 2. Nous analysons également si notre cadre d'étude de la valeur est suffisant dans le cadre d'une organisation en challenges successifs.

Après avoir étudié la notion de capitalisation dans le cas d'une tâche de type résolution de problèmes, ce chapitre a pour ambition de déterminer les moyens à mettre en œuvre pour gérer la capitalisation séquentielle dans le cas des tâches exploratoires. Comme nous l'avions déjà évoqué dans notre méthodologie de recherche, notre analyse porte pour cette question sur l'étude d'un programme de recherche ouvert Epidemium, destinée à la recherche scientifique et consacrée à la compréhension et à l'épidémiologie du cancer. En effet, les scientifiques du domaine médical et les acteurs du domaine de l'épidémiologie font face à une explosion du nombre de données accessibles et de leur variété (Chiolero, 2013). Face à cette accumulation de ressources disponibles, de plus en plus de structures liées à la santé cherchent des moyens de valoriser les bases de données en cherchant à produire des outils ou des résultats de recherche scientifique. Epidemium nous permet d'étudier une des premières initiatives de ce genre dans le cadre de l'épidémiologie du cancer. Contrairement à d'autres domaines scientifiques de la santé qui demandent un haut niveau de compétences pour concevoir de nouvelles hypothèses scientifiques, l'épidémiologie est une discipline à la croisée des spécialistes de la santé et des études statistiques qui a déjà fait l'objet de collaboration entre scientifiques et un public via l'épidémiologie populaire.

Dans le programme Epidemium, une équipe d'organisateurs délègue la génération d'hypothèses scientifiques ainsi que la conception d'algorithmes pour valider ces hypothèses à la foule. Ce projet est un cas unique dans la littérature que nous avons tenté dans ce chapitre et le suivant d'étudier. Nous avons modélisé le programme Epidemium comme la délégation d'une tâche couplée inventive à une foule et nous nous sommes intéressé à la performance de cette délégation, et notamment au processus de capitalisation mis en œuvre. Nous avons également retracé les processus d'exploration des participants au sein des deux espaces. Ce dispositif remet en cause la place même du scientifique au sein de processus de production scientifique et interroge sur les rôles de chacun des acteurs impliqués.

1. ELABORATION DU PROGRAMME EPIDEMIUM : ORGANISATION, FINANCEMENT

Le projet Epidemium est une initiative issue d'une collaboration entre deux acteurs, Olivier de Fresnoye et Mehdi Benchoufi, et d'une volonté d'introduire la dimension d'ouverture de la science dans une discipline aussi normée que la santé et le médical. Afin de réunir des acteurs médicaux et des experts en analyse de données pour l'exploration, les deux initiateurs du projet ont collaboré avec deux structures : les laboratoires Roche ainsi qu'un nouveau type de laboratoire de recherche en santé, l'association La Paillasse, un laboratoire de recherche ouvert à tous.

Les laboratoires Roche, une des plus importantes entreprises pharmaceutiques en terme de chiffre d'affaires, ont été intéressés par le projet car ils souhaitent évaluer comment l'analyse de

données massives et de sources hétérogènes pourrait être un catalyseur pour une nouvelle médecine plus préventive et personnalisée[1]. Alors que Roche est coutumier de la mise en place d'enquêtes épidémiologiques (voir par exemple l'étude ObEpi en 2012[2]), les experts internes sont confrontés à une double contrainte par rapport aux méthodes statistiques conventionnelles. Premièrement, les équipes de laboratoire ne sont pas expertes des méthodes d'analyses basées sur l'intelligence artificielle. Ces méthodes sont relativement récentes, et les spécialistes n'ont pas reçu de formations spécifiques. Deuxièmement, la méthode scientifique à mettre en œuvre diffère des méthodes statistiques conventionnelles. Alors que la collecte de données est directement liée à une hypothèse prédéterminée, l'épidémiologie data-driven cherche à interroger une base de données déjà collectée avant que l'hypothèse ne soit définie.

Ensuite, la Paillasse est d'abord un lieu physique qui recycle des instruments scientifiques dont les laboratoires se débarrassent pour un deuxième usage. Ce lieu permet à toute personne intéressée de mener des actions afin d'amorcer ou d'accélérer des projets scientifiques, entrepreneuriaux ou artistiques[3]. En plus de ses activités, La Paillasse regroupe un ensemble d'acteurs, souvent passionnés par la recherche et qui militent pour une science plus ouverte. Cette collaboration avec un laboratoire communautaire et ouvert n'est pas évidente pour un grand groupe pharmaceutique comme Roche. Au-delà des différences organisationnelles, cette collaboration met en avant des postures idéologiques sur la science très éloignées et parfois contradictoires. Pourtant, le projet a suscité un fort engouement au sein des laboratoires Roche avec la participation de plus d'une cinquantaine d'employés : un groupe projet dédié de 10 personnes, 24 ambassadeurs, et plus de 20 collaborateurs impliqués[4].

Les laboratoires Roche fournissent un financement, une expertise et ouvrent une partie de leurs données, tandis que la Paillasse fait profiter de sa culture de la science ouverte, l'accès à une communauté de scientifiques sensibles à l'ouverture des sciences ainsi que des locaux. Pour assurer l'unité et une forme d'indépendance, la collaboration entre Roche, La Paillasse et les initiateurs du projet ont créé ensemble en 2015 le projet Epidemium[5], conçu comme une structure destinée à la recherche scientifique et consacrée à la compréhension et à l'épidémiologie du cancer. Le but du projet est de réunir l'ensemble des données rendues disponibles relatives aux potentiels facteurs de risque liés au cancer et de développer des projets à visée scientifique à partir de leur exploration et de leur exploitation. Les participants sont incités à explorer les bases de données pour construire des hypothèses scientifiques et des méthodes d'évaluation de ces hypothèses. Le principe d'Epidemium suit les caractéristiques propres aux projets de science ouverte. D'abord une ouverture des résultats intermédiaires, en rendant disponibles des bases de données scientifiques ainsi qu'un ensemble d'outils pour faciliter leur analyse. Ensuite, une

[1] http://www.roche.fr/innovation-recherche-medicale/big-data-sante.html

[2] http://www.roche.fr/content/dam/roche_france/fr_FR/doc/obepi_2012.pdf

[3] https://lapaillasse.org/

[4] https://medium.com/epidemium/lengagement-de-roche-35f11a777419

[5] http://www.epidemium.cc/

ouverture à tous les participants. En effet, les barrières à l'entrée sont très faibles, et toute personne intéressée peut participer au projet. C'est cette structure, son organisation et son activité que nous avons étudiées dans notre thèse.

1.1. COMPILATION DE DONNEES EPIDEMIOLOGIQUES

La première étape d'Epidemium a consisté à collecter toutes les données ouvertes disponibles relatives à l'épidémiologie du cancer et les traiter pour les rendre plus facilement exploitables. Les données collectées ont été globalement divisées en deux types : les données sur la mortalité et l 'incidence du cancer, et les données sur les facteurs de risque potentiels. Un ensemble de données a été compilé sur l'incidence et la mortalité du cancer à partir des bases de données disponibles sur les sites de l'OCDE et de l'Organisation Mondiale de la Santé[6]. L'OCDE fournit des jeux de données ouverts sur la mortalité du cancer sur la période 1960-2012 en fonction du pays, du type de cancer et du sexe. L'Organisation Mondiale de la Santé fournit des jeux de données à la fois sur la mortalité et sur l'incidence du cancer. Les jeux de données sur la mortalité du cancer concerne la période 1950-2012 et ceux sur la mortalité du cancer la période 1953-2007. Ils sont classés par pays, type de cancer, tranche d'âge et par sexe. Un jeu de données spécifique sur l'incidence des cancers en France sur une période 2009-2012 est également disponible avec un classement par région et par type de cancer.

Les organisateurs ont également collecté des bases de données sur des facteurs de risque potentiels pouvant être corrélés à la mortalité et à l'incidence du cancer. La base de données la plus complète disponible concerne les jeux de données concernant la répartition des infections sexuellement transmissibles (VIH, Syphilis, tuberculose, hépatites,…) en particulier aux Etats-Unis. Ces bases de données concernent globalement la période 1990-2011 mais celle-ci varie suivant la maladie. La plateforme Epidemium a également collecté un ensemble de jeux de données sur des informations très générales, et dont l'exploration peut potentiellement mener à la découverte de facteurs de risque, ou au moins au classement des facteurs de risque suivant leur ordre d'importance. Ces données sont cataloguées suivant des thématiques :

- Démographique : âge, population, taux de suicide, taux de mortalité et de fécondité, nombre d'enfants par femme,…
- Environnemental : émission de CO2 et de GES, pourcentage de terres agricoles, biomasse en forêt,… (issues notamment du site de la FAO, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture)
- Travail : emploi et condition de travail, revenus, chômage, temps de travail, scolarisation,…
- Economique : croissance, PIB par habitant, score démocratique,…

[6] http://wiki.epidemium.cc/wiki/Donn%C3%A9es

- Comportement : consommation de tabac et d'alcool, utilisation de charbon, consommation téléphonique,...
- Santé : maladies, moyens de contraception,...

Afin d'étendre la portée des études possibles, l'équipe organisatrice d'Epidemium a mis à la disposition des participants des publications en épidémiologie issues de la littérature scientifique médicale ou d'essais cliniques. Plusieurs jeux de données ont été intégrés, y compris des essais cliniques rassemblés sur la plateforme de l'OMS, ClinicalTrials.gov, la demande de données d'étude clinique et la base de données complète des publications PubMed Open Access et des publications sur PubMed. Enfin, les laboratoires Roche ont mis à disposition un ensemble de d'études réalisées par le laboratoire. Ces données permettent de réaliser des études méta-épidémiologiques ou scientométriques, c'est-à-dire d'analyser les résultats scientifiques afin d'en tirer des conclusions ou de trouver des hypothèses non encore élucidées par la littérature. Au total, Epidemium a compilé plus de **21 000 jeux de données** relatifs à l'épidémiologie du cancer accessibles à tous les participants et libres de droits et d'utilisation. Ces données servent de base pour les participants dans la génération des hypothèses scientifiques.

1.2. ORGANISATION DU PROGRAMME EPIDEMIUM

Epidemium est dirigé par une équipe de 6 personnes, des experts en science ouverte et en gestion communautaire ainsi qu'un chef de clinique en épidémiologie. Tout autour de cette équipe centrale que constitue le cœur d'Epidemium, plusieurs structures ont été développées correspondant chacune à des fonctions précises au cœur du projet (voir **figure 47**). L'équipe organisatrice est en lien très étroit avec l'Assistance Publique des Hôpitaux de Paris (APHP) et donc avec des acteurs de la santé potentiellement intéressés par la démarche d'ouverture. En effet, un nombre certes restreint mais grandissant d'acteurs au sein du domaine médical militent pour une plus grande ouverture de la science et se retrouvent dans les valeurs défendues par le projet Epidemium. Différents partenariats sont initiés avec des spécialistes issus d'institutions publiques reconnues dans le domaine du cancer tel que l'Institut Curie, Gustave Roussy ainsi que le centre de recherche sur le cancer CLARA à Lyon. Cette communauté de spécialiste est indispensable pour asseoir une certaine crédibilité mais également pour fournir l'expertise scientifique essentielle à la bonne réussite des projets.

L'équipe organisatrice a également constitué deux comités de spécialistes pour évaluer de façon indépendante les projets. Le premier est un comité scientifique dont l'objectif est de « garantir la qualité des outils et connaissances mis à disposition des participants au Challenge, définir une grille de critères d'évaluation des projets, valider d'éventuelles publications, identifier les applications des projets au terme du challenge, enfin, il doit participer à la rédaction d'un cadre méthodologique. »[7]. En pratique son rôle est de s'assurer que les méthodologies employées par les participants durant le processus sont cohérentes vis-à-vis des exigences scientifiques. Ce comité

[7] http://wiki.epidemium.cc/wiki/Comité_scientifique

intervient trois fois durant la réalisation des projets par les participants. Une première fois au début pour s'accorder sur les critères d'observation des projets des participants. Une deuxième fois au milieu de projet pour évaluer l'avancement des projets et permettre un retour d'experts. Enfin une troisième fois à la fin pour évaluer le projet. Le comité est constitué de 9 personnes avec en proportion équivalente des experts en oncologie et des spécialistes de l'analyse de données.

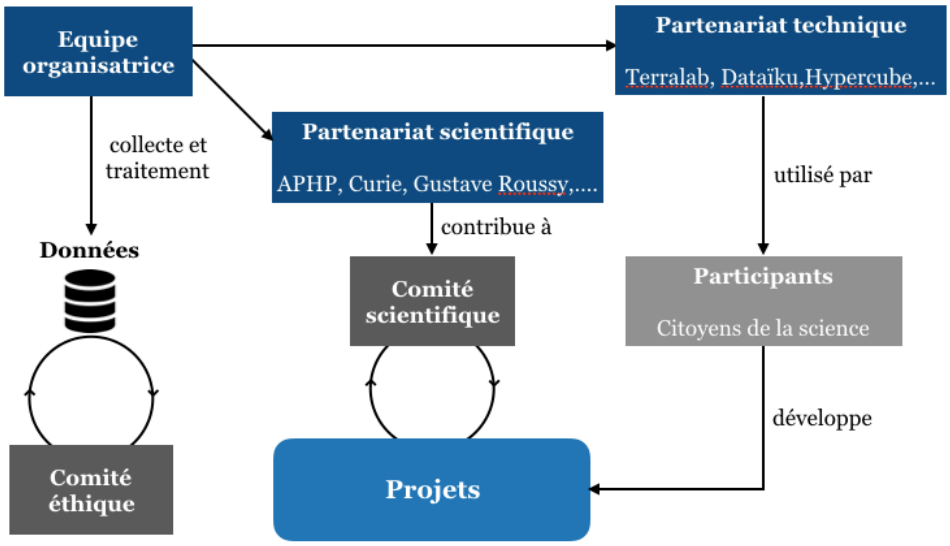


Figure 47. Organisation d'Epidemium.

Le deuxième est un comité éthique, fondamental dans toute démarche de santé publique[8]. De nombreux enjeux concernent les données médicales comme la privatisation, le consentement à l'usage des données personnelles ou les risques de sur-diagnostics qui doivent être pilotés d'une point de vue éthique. La garantie de l'anonymisation est d'autant plus complexe à gérer avec des jeux de données ouverts. Le comité a en charge principalement d'évaluer si les pratiques mises en œuvre relatives aux données utilisées respectent les règles relatives à toute analyse médicale. Les membres du comité ont ainsi mis en place une charte éthique, qui précise le bon comportement à avoir par rapport aux données rendues accessibles. Ils sont également garants de la conformité éthique des projets vis-à-vis de cette charte. Le comité est constitué de 10 membres d'origines variées : chercheurs en Big Data, directrice d'association de patients, avocat, ou encore mathématicien (en l'occurrence Cédric Villani, médaille Fields 2010 en mathématiques).

1.3. Dispositifs et outils de gestion au sein du programme Epidemium

Les organisateurs du programme Epidemium ont mis en place deux challenges successifs, appelés Challenge4Cancer, entre 2015 et 2018 durant lesquels les citoyens ont été incités à participer librement sur des thématiques qui avaient été préalablement définies. Lors de ces challenges, les

[8] http://wiki.epidemium.cc/wiki/Comité_d%27%éthique

participants se sont généralement regroupés en équipe pour travailler sur une période de six mois pour formuler des hypothèses de recherche à partir des bases de données rendues disponibles par Epidemium. Notre analyse porte sur l'étude de ces challenges, afin d'en inférer les bons principes et d'en déduire un modèle de gestion approprié. Ce modèle a pour ambition de fournir un ensemble de critères et de règles dans le cadre d'une systématisation de l'ouverture des tâches couplées inventives que nous avons suggérée dans notre revue de littérature.

Nous présentons dans cette section le programme Epidemium tel qu'il a été conçu avant le premier challenge. Cette présentation de l'organisation d'Epidemium et de sa gouvernance est réalisée au prisme des outils et des dispositifs de gestion qui ont été mis en œuvre. Au-delà de leur fonctionnalité technique, les outils de gestion peuvent être un moyen d'analyser un système de gestion. En plus d'avoir un rôle de médiation dans l'organisation, les outils de gestion constituent une forme privilégiée d'intervention pour construire de nouvelles capacités d'action, mais également un formidable appareil pour observer les transformations dans les organisations (Aggeri & Labatut, 2010).

1.3.1. Challenge4Cancer

Les organisateurs d'Epidemium ont lancé un challenge, baptisé Challenge4Cancer, basé sur les jeux de données collectées. L'objectif déclaré du challenge est double: identifier les hypothèses pertinentes à partir des bases de données disponibles et développer des méthodes pour tester ces hypothèses sur la base. Le premier challenge s'est déroulé sur 6 mois entre le 5 novembre 2015 et le 6 mai 2016. Au total, 678 contributeurs ont participé. Les organisateurs ont proposé aux participants de répondre à l'une des quatre thématiques suivantes :

Thématique	Modélisation
T1 : Comprendre la répartition du cancer dans le temps et dans l'espace	$\mathcal{T}_1$ = *Répartition({type de cancer}, {zone géographique}, {temporalité})*
T2 : Facteurs de risque et facteurs de protection du cancer	$\mathcal{T}_2$ = *Relations({type de cancer}, {facteurs de risques})*
T3 : Méta-épidémiologie: comprendre le cancer à partir de la littérature scientifique médicale	$\mathcal{T}_3 = \mathcal{H}$
T4 : Changements environnementaux et cancer	$\mathcal{T}_4$ = *Relations({type de cancer}, {facteurs de risques = changements environnementaux})*

Tableau 11. Thématiques proposées par les organisateurs d'Epidemium pour le premier challenge

Ces thématiques sont délibérément sous-spécifiées pour laisser place à diverses hypothèses de recherche. Elles ne peuvent être considérées en soi comme des hypothèses : plutôt, elles suggèrent au participants de se concentrer sur certains concepts. Ainsi les thèmes 2 et 4 traitent

exclusivement des liens entre l'apparition de « type de cancer » et les « facteurs de risque » associés tandis que le thème 1 cherche à étudier l'impact de la « répartition » des cancers en fonction d'une « zone géographique » et de la « temporalité ». Enfin le thème 3 ne spécifie pas véritablement de famille de concepts, il propose plutôt aux participants d'axer leur exploration sur un type de données, à savoir la littérature scientifique.

Chaque participant ou équipe choisit l'une des thématiques et définit un problème à résoudre à partir des données relatives à la thématique. Les équipes de projet sont également invitées à remplir une page wiki lors de l'exécution du projet. Au terme du challenge, les comités éthiques et scientifiques évaluent les projets. Trois projets lauréats reçoivent un prix: 5 000 € pour le premier et 2 000 € pour le deuxième et le troisième.

1.3.2. Mise à disposition d'outils techniques

En plus de développer une légitimité scientifique et de rassembler des individus, le projet Epidemium a tissé des partenariats techniques avec des entreprises pour rendre accessible des outils d'analyse et de traitement des données. Plusieurs entreprises ont donné accès durant la durée du challenge à des technologies d'analyse Big data permettant d'explorer les interactions entre une variable à prédire et les variables explicatives d'un jeu de données complexe (Hypercube, Dataïku). Ces outils sont également déployés dans le cadre d'analyse de données massives en entreprise et peuvent servir de support pour des personnes non spécialistes de l'analyse de données. En parallèle, Epidemium a développé un partenariat avec le Center for Data Science de Paris Saclay afin d'utiliser la plateforme de data challenge RAMP (Rapid Analytics and Model Prototyping). Contrairement aux plateformes classiques de data challenge tel que Topcoder ou Kaggle, le RAMP est une plateforme collaborative durant lequel les modèle soumis par les participants peuvent être regardés et utilisés par les autres. Les équipes de projet ont été également sollicité pour renseigner une page wiki sur le déroulé du projet, de la question à l'avancement final, ainsi que d'utiliser la plateforme Github. Suivant le règlement, les participants placent leur contribution sous une licence de leur choix respectant les conditions d'ouverture de l'Open Source Initiative (disponible sur http://opensource.org/osd) en fonction de leur contribution. D'autres outils gratuits ont été déployés pour la gestion de projet, la conservation des éléments produits ainsi que pour la communication entre les participants (Slack, Q&A, Wiki et GitHub). La gamme d'outils proposée par Epidemium permet de baisser les barrières à l'entrée pour les participants en proposant des outils avec des niveaux de compétences variées et donc adaptée à chaque profil de participant.

1.3.3. Synthèse des outils de gestion

Chaque élément présenté dans cette section peut être interprété comme un dispositif ou d'un outil de gestion répondant à une fonction spécifique dans le cadre d'Epidemium. La représentation de tous ces outils permet d'avoir une trace observable, une photo du modèle de gestion pensé et mis en place par les organisateurs d'Epidemium au début du challenge (**tableau 12**).

Dispositif/outil de gestion	Fonctions	Dispositif/outil de gestion	Fonctions
Gestion de l'exploration		***Outils de gestion de projet***	
Epidemium	Organisation globale	Slack, Q&A	Outils de communication
Challenge4Cancer	Guide pour l'exploration des espaces	Basecamp, Drive	Outils de gestion de projets
Jeux de données	Base de connaissances pour la formulation des hypothèses	Dataïku, Hypercube	Outils d'analyse big data
Comité scientifique	Contrôle de la qualité des productions du Challenge4Cancer	Teralab	Cluster big data pour stockage
Comité éthique	Vérification de la conformité éthique des projets	RAMP	Outil de développement de modèle big data
Gestion de la communauté		***Outils de capitalisation***	
Epidemium	Fabriquer une communauté	Wiki	Outil de partage de l'avancement
Challenge4Cancer	Motivation à l'entrée, favoriser la collaboration	GitHub	Stocker les connaissances (nouvelles base de données, modèles)
Meet-ups (+100)	Rencontre, partage de connaissance, maintien de la communauté, recruter des talents		
RAMP	Outil d'incitation pour la communauté de data scientists		

Tableau 12. Dispositifs et outils de gestion développés dans le programme Epidemium

Ce modèle est évolutif en fonction du temps et des résultats des challenges et nous reviendrons dessus pour illustrer les transformations au sein du programme à travers l'évolution des fonctions de gestion auquel Epidemium doit répondre. Cela nous permettra de mettre en avant les fonctions qui auront été considérées comme inutiles par les organisateurs, ainsi que celles manquantes. Nous avons regroupé les fonctionnalités suivant quatre catégories :

- la gestion de l'exploration : elle représente l'ensemble des fonctionnalités dont l'objectif est de fournir un cadre à l'exploration dans la résolution de la tâche
- la gestion de la communauté : ces outils permettent de construire, consolider, et gérer la communauté Epidemium
- la gestion de projets : ce sont les outils mis en place pour les participants afin de les aider à piloter les projets
- la gestion de la capitalisation : outils pour capitaliser sur la production durant les challenges

1.4. CRITERES DE VALIDITE DES HYPOTHESES SCIENTIFIQUE EN EPIDEMIOLOGIE DU CANCER : HYPOTHESE ET « AXE DE TRAVAIL »

Dans le programme Epidemium, les compétences des participants ne sont pas connues *ex ante*. Ils ne sont donc pas nécessairement des experts en épidémiologie du cancer. Or pour comprendre et évaluer leurs résultats, il est nécessaire de définir ce qui est considéré comme une hypothèse valide dans le domaine de l'épidémiologie du cancer. Il n'est ici pas question de juger la valeur de l'hypothèse formulée, mais bien de sa validité vis-à-vis d'une communauté de scientifiques. Pour établir ce paradigme, nous nous sommes basés sur un ensemble de publications scientifiques dans la littérature concernant l'épidémiologie du cancer. Celles-ci peuvent être perçues comme des explorations scientifiques dont le résultat et la méthodologie ont été validées par la communauté d'expert, et donc considérés comme valides.

Bien qu'il y ait un grand nombre d'études possibles en épidémiologie du cancer, les types de publications que l'on retrouve dans les journaux de la discipline sont au final d'un nombre assez restreint. Nous avons pu reconstituer une typologie relativement exhaustive des publications que l'on peut retrouver dans les journaux spécialisés en épidémiologie. Nous avons étudié les publications de l'année 2018 relatives à l'épidémiologie du cancer dans les revues *International Journal of Epidemiology*, *Cancer Causes & Control*, *Cancer Epidemiology Biomarkers & Prevention*, *Epidemiologic Reviews*, *European Journal of Epidemiology* et *Cancer Epidemiology* afin de proposer une typologie présentée dans le (**tableau 13**). Cette typologie constient six types d'études : la revue critique, les modèles descriptifs et tendances, l'étude des facteurs de risque, le dépistage et la prévention, la survie, la méthodologie. Chaque type est illustré d'exemples issus du volume 55 du journal *Cancer Epidemiology* d'Août 2018.

A partir de cette typologie, nous cherchons à établir un ou des critères qui permettent d'évaluer si une hypothèse formulée durant le challenge correspond aux standards que nous trouvons habituellement dans la discipline. Ce résultat a pour objectif de filtrer rapidement les hypothèses valides des hypothèses non valides. Une typologie trop restrictive serait potentiellement préjudiciable au critère d'originalité que doit respecter toute nouvelle hypothèse. C'est pourquoi nous ne proposons non pas des règles strictes de formulation mais plutôt un système d'évaluation simple et potentiellement évolutif qui permettra d'analyser les hypothèses générées par le programme Epidemium. L'analyse que nous avons menée se base exclusivement sur le modèle que nous avons construit au chapitre 5. Nous étudions les hypothèses suivant un espace du langage que nous cherchons à expliciter et qui correspond à la discipline étudiée.

Type d'étude	Objectif	Exemple
Revue critique	Analyse méta-épidémiologique sur des études réalisées et publiées	*Review of methodological challenges in comparing the effectiveness of neoadjuvant chemotherapy versus primary debulking surgery for advanced ovarian cancer in the United States* (Cole et al., 2018)
Modèles descriptifs et tendances	Etude de la mortalité ou l'incidence (souvent temporelle) d'un type de cancer suivant la population, la zone géographique,…	*Lung cancer incidence trends in Uruguay 1990–2014: An age-period-cohort analysis* (Alonso et al., 2018)
Etude des facteurs de risque	Etudier les facteurs de risque associé à un cancer ou à un type de cancer	*Benzene exposure at workplace and risk of colorectal cancer in four Nordic countries* (Talibov et al., 2018)
Dépistage et prévention	Etudier les effets de la prévention sur les risques de développer un cancer	*Avoidable colorectal cancer cases in Denmark – The impact of red and processed meat* (Lourenço et al., 2018)
Survie	Etude de la survie des patients atteints d'un cancer	*Mortality of patients examined at a diagnostic centre: A matched cohort study* (Næser et al., 2018)
Méthodologie	Méthodes relatives aux études épidémiologiques	*Childhood cancer registration in New Zealand: A registry collaboration to assess and improve data quality* (Ballantine et al., 2018)

Tableau 13. Typologie des publications scientifiques en épidémiologie du cancer (issu du journal *Cancer Epidemiology* Volume 55)

Notre typologie permet de mettre en avant deux spécificités dans les publications en épidémiologie du cancer. D'une part, la plupart des hypothèses sont élaborées sur l'étude d'un type de cancer précis et non d'une agrégation de plusieurs cancers. De la même manière, la zone géographique est souvent précisée, que ce soit une région, un pays ou plus rarement à l'échelle d'un continent. La formulation d'hypothèses semble suivre une logique où la granulométrie doit être la plus fine possible. Cela n'est pas étonnant. En effet, avoir un résultat basé sur une hypothèse trop vague ne permettrait pas de tirer des conclusions intéressantes en terme de santé publique. Pire il pourrait également passer à côté de certaines spécificités. Nous tenons cependant à préciser qu'il n'existe pas de règles absolues pour ce critère et une hypothèse pourra très bien être valide sans préciser un type de cancer. Pour autant, il est raisonnable d'affirmer que les hypothèses sont généralement plus proches de ce type de granulométrie. D'autre part, les concepts utilisés pour constituer l'hypothèse sont en nombre restreint et peuvent être facilement énumérés : ce sont par exemple le type de population, la temporalité de l'étude, ou encore le type de facteur de risque pris en compte. Ainsi nous considérons que pour être valide une hypothèse en épidémiologie du cancer doit au moins être basée sur **un seul type de cancer (sauf cas particulier) et un seul impact (mortalité, survie,…).** Elle devra également respecter une forme du type :

$\mathcal{H}$ = *impact{mortalité, survie, incidence, efficacité,…}(type de cancer, facteur de risque, zone géographique, temporalité, profil des patients, type de traitement,…)*

A titre d'exemple, le papier D'Alonso et ses collègues (2018) peut être modélisé comme $\mathcal{H}$ = *incidence({type de cancer = cancer du poumon}, {zone géographique = Uruguay}, {temporalité = 1990-2014}, {profil = âge, genre})*. Toute hypothèse qui ne respecte pas ces critères est appelée « **axe de travail** ». Nous rappelons que l'espace des hypothèses est extensible et que l'ensemble des catégories est également non fini, simplement cela permet de fournir un cadre initial pour interpréter la validité d'une hypothèse.

Les thématiques proposées par Epidemium peuvent clairement être assimilées à des axes de travail : en effet, elles réduisent la taille de l'espace des hypothèses, mais ne précisent pas quel type de cancer ni quel type de relation entre les variables.

2. *LE PROGRAMME EPIDEMIUM COMME LA RESOLUTION D'UNE TACHE COUPLEE*

Le programme Epidemium peut être assimilé à la délégation d'une tâche couplée inventive telle que nous l'avons définie dans le chapitre 5. Nous pouvons représenter l'exploration des participants au travers de deux espaces. Un premier espace est constitué de tout le code informatique. Dans celui-ci, les participants cherchent les corrélations existantes entre les variables (les colonnes) issues des bases de données collectées. Par exemple, l'étude des facteurs de risque consiste à chercher les relations entre les cancers existants Y et les facteurs de risques X tel que $Y = f(X)$ avec f la relation de corrélation entre les variables. Les participants construisent des séquences d'actions afin de connaître ces relations et la validité statistique de celle-ci. A noter que dans le cas où le problème existerait *ex ante* (i.e. on a déterminé à l'avance l'état final désiré Y, le type de données X ainsi qu'une fonction *test()*), la tâche peut être réduite à de la résolution de problèmes, semblable à ce que nous avons étudié dans le cas du RAMP.

Les participants explorent également un deuxième espace pour construire les hypothèses scientifiques. Celui-ci est constitué de l'ensemble des hypothèses que l'on peut établir à partir des variables issues des 21 000 bases de données. Dans le challenge, chaque thématique développée par Epidemium constitue une partition de cet espace. Passer d'une thématique A à une hypothèse scientifique H revient à appliquer une fonction de transformation τ tel que $\tau : A \rightarrow H$. Pour que l'hypothèse soit valide, cette fonction doit réaliser au moins deux actions :

- Associer à la notion de *cancer* une sous-famille de concept correspondant à un type de cancer

- Formuler une relation *R* entre les familles de concepts (mortalité, survie, incidence, efficacité,…)

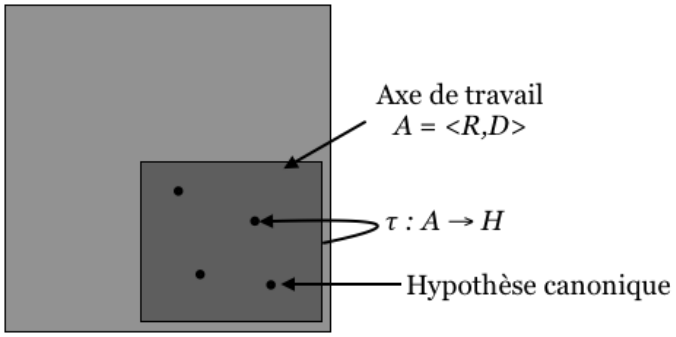


Figure 48. Espace des hypothèses et fonction de transformation de l'axe de travail vers l'hypothèse scientifique

L'exploration de l'espace des hypothèses représente un réel défi à cause de la taille des bases de données. Comme nous l'avons vu, les variables dans la base de données concernent des domaines très larges comme l'environnement, les conditions de travail, la santé, des statistiques sur les cancers, ou des résultats cliniques. Même si chaque jeu de données ne possédait qu'une seule variable unique, le nombre de combinaisons possibles à partir des bases de données serait de 21000^2 = 44 100 000 000. Ce nombre atteint 21000^3 = 9,261 x 10^15 quand la relation est faite sur trois indicateurs (cancer, facteur de risque et région du monde par exemple). Dans le cas d'une exploration aléatoire, la probabilité de tomber sur une hypothèse intéressante est extrêmement faible. Si l'exploration était faite par des scientifiques, leur connaissance sur le sujet permettrait de réduire de façon importante le nombre d'hypothèses inutiles ou non pertinentes. Cependant, ils seraient également rapidement limités en ressource (temps d'exploration par rapport au nombre de combinaisons possibles) ainsi que dans leurs compétences en analyse de données. De plus, les données provenant de sources hétérogène, ils n'ont pas une vision synthétique et globale des différentes bases de données qui leur permet de savoir exactement où aller explorer. Au contraire la délégation de l'exploration de ces bases de données par le biais des sciences citoyennes permet de profiter de la capacité des participants à utiliser les algorithmes d'analyse de données ainsi que de la multiplicité des explorations relative au nombre de participants.

3. EXPLORATION ET PRODUCTION DURANT LE PREMIER CHALLENGE4CANCER

3.1. CAPACITES DES ORGANISATEURS A FEDERER UNE COMMUNAUTE

Une des difficultés dans la réalisation d'un projet ouvert tel que les sciences citoyennes est de pouvoir construire la communauté de participants et recruter les talents essentiels à sa réussite.

Cette tâche demande aux organisateurs de multiplier les moyens de communication dans des environnements propices et de savoir transformer un simple intérêt en un réel engagement dans le projet. Epidemium a réalisé pour le premier challenge 115 présentations dans un large éventail d'organisations externes considérées comme des partenaires potentiels ou des centres où des talents peuvent être recrutés pour le challenge. Dans une communauté de 678 membres, dont 331 participants inscrits au challenge (54% de scientifiques de données, 28% d'informaticiens et 18% de professionnels de la santé ou de chercheurs médicaux), 75 personnes ont participé à l'un des 16 projets, avec 63 finalistes pour 8 projets sélectionnés par les comités. Ces présentations ont été un vecteur de motivation mis en place par les organisateurs d'Epidemium afin d'encourager les équipes à collaborer entre elles en incluant dans l'évaluation finale le niveau de coopération du projet pendant le challenge et en favorisant les échanges entre les participants. Une rencontre hebdomadaire organisée dans les locaux de La Paillasse a facilité l'intégration de nouveaux contributeurs dans les projets et la rencontre physique des différents participants pour des collaborations potentielles.

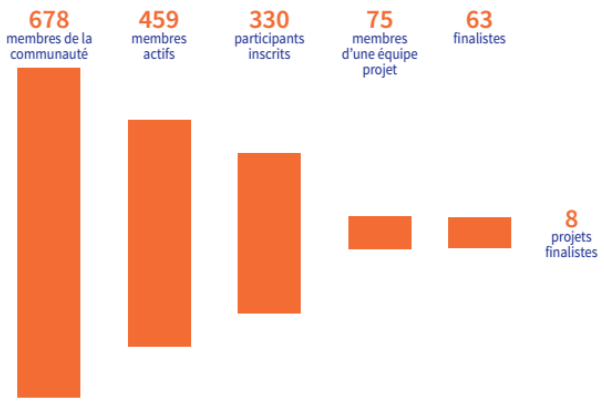


Figure 49. Taux de conversion des membres de la communauté en projets actifs.

3.2. BILAN GLOBAL DU PREMIER CHALLENGE

Parmi les 678 volontaires inscrits au sein de la communauté, 16 équipes se sont constituées pour proposer un projet correspondant aux thématiques du challenge. Au final, seuls 8 sont arrivés jusqu'au bout en soumettant leur production aux comités à la fin du challenge, pour un total de 63 volontaires, soit un taux de transformation de 9% (nombre de participants actifs par rapport au nombre total d'inscrits). Malgré ce faible taux, il est cependant remarquable que 63 volontaires se soient investis fortement dans un projet ouvert jusqu'au bout. Cet investissement permet d'atteindre un niveau de ressources humaines que les scientifiques seuls auraient difficilement pu atteindre avec les mêmes investissements et le même budget (dix fois plus que l'équipe organisatrice d'Epidemium). Les 8 projets diffèrent dans leur approche et dans le nombre de participants (de deux ou trois personnes à plusieurs dizaines). Une première catégorie de projet a cherché à construire des modèles causals ou prédictifs entre différents facteurs pour tester certaines hypothèses (*Baseline*, *Approche Prédictive et Risque de Cancer - APRC*). Une deuxième

catégorie d'équipes a développé des outils de visualisation des données afin de faciliter la formulation des hypothèses (*Viz4Cancer*, *CancerViz*) ou explorer la littérature scientifique (*OncoBase*, *BD4Cancer*, *Venn*). Enfin, un projet unique a proposé d'utiliser les données pour développer un outil pédagogique afin de sensibiliser aux facteurs de risque des cancers (*ELSE*).

Thème 1 - Comprendre la répartition du cancer dans le temps et dans l'espace		
Viz4Cancer	Site internet informatif qui permet de représenter graphiquement d'une part l'évolution de différents types de cancer en France et d'autre part la variation de différents facteurs socio-environnementaux comme les polluants atmosphériques ou les dépenses net en tabac.	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
CancerViz	Outil de data visualisation facilitant la phase d'acquisition des données	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
Thème 2 - Facteurs de risques et Facteurs protecteurs du cancer		
Baseline	Prévoir l'incidence / la mortalité / la survie au cancer en utilisant des facteurs de risque provenant de sources de données ouvertes (avec une portée mondiale et une granularité régionale)	$\mathcal{A}$ = *{impact =* ***incidence, mortalité, survie****} ({type de cancer}, {zone géographique =* ***granularité régionale****})*
Approches prédictives et risque de cancer	Mieux identifier les facteurs de risque du cancer dont certains font l'objet de travaux et de recherche comme les radiofréquences, les pesticides ou les nanoparticules.	$\mathcal{A}$ = *impact({type de cancer}, {facteurs de risque =* ***environnementaux****})*
Thème 3 - Meta-épidemiologie : comprendre le cancer à partir de la littérature scientifique médicale		
OncoBase	Produire une base de données unifiée d'articles de la littérature scientifique pour la communauté Epidemium, mêlant des données issues de sources diverses et variées, et permettant de construire des analyses statistiques solides	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
BD4Cancer	Identifier des événements liés à l'usage des médicaments anti-cancers, dont les effets secondaires (Drug Side Effects; DSE) et les effets indésirables (Adverse Drug Reaction; ADR) à partir de la littérature scientifique.	$\mathcal{A}$ = *{impact =* ***effets indésirables****}({type de cancer}, {facteurs de risque =* ***médicaments anti-cancers****})*
Thème 4 - Changements environnementaux et cancer		
ELSE	Jeu de data visualisation basé sur les données pour sensibiliser aux facteurs de risque (outil pédagogique sur les facteurs de risque basé sur les données réelles).	Développement d'outils dans l'espace du code informatique
Venn	Construire une procédure d'extraction et d'analyse de données textuelles issues des méta-données des papiers de recherche afin de mettre en lumière les liens entre cancer et pollution de l'air mis en évidence par la recherche.	$\mathcal{A}$ = *impact({type de cancer}, {facteurs de risque =* ***pollution de l'air****})*

Note : les familles de concepts spécifiées par les projets sont marquées en gras

Tableau 14. Présentation des projets du challenge 1 d'Epidemium.

Dans l'ensemble les projets ont été élaborés à partir d'axes de travail et non sur des hypothèses considérées comme valide par la discipline (**tableau 14**). En effet, il est difficile pour les participants de formuler une hypothèse sans connaître au préalable les données disponibles. Au contraire, il semble plus judicieux de commencer l'exploration de manière un peu large pour pouvoir ensuite construire des hypothèses valides. A noter également que sur les 8 projets, seuls quatre cherchent à produire de la connaissance scientifique. Les participants ont la liberté de choisir ce qu'ils vont produire, et ne sont pas restreints par des objectifs clairs. Cela fait naître dans cette forme d'organisation ouverte à la fois un risque et une opportunité pour les organisateurs de voir émerger des projets inattendus.

Les six mois d'activité durant le challenge ont été foisonnantes et ont mobilisé un grand nombre de participants. Pourtant, la plupart des projets ne sont restés qu'à l'état de prototype et bien que quelques hypothèses aient pu être formulées, aucune d'entre elles n'a pu être vérifiées à partir des bases de données. Le principal facteur a été l'incapacité des groupes à terminer à temps. Ceux qui ont soumis leur projet final n'ont pas réussi à atteindre les objectifs qu'ils avaient initialement fixés et ont dû présenter des prototypes ou des versions simplifiées de leur projet initial. Ces modifications à la volée ont plusieurs sources selon les participants : difficultés techniques dans la recherche d'un modèle d'apprentissage automatique efficace, problèmes de qualité des données et impossibilité d'explorer efficacement le grand nombre d'ensembles de données.

L'exploration des espaces par les participants n'a pas été un processus linéaire. Au contraire, elle résulte de nombreux aller-retour entre l'espace des hypothèses et l'espace du code informatique pour élaborer une hypothèse compatible avec les bases de données. Nous allons voir quelles sont les stratégies qui ont été mises en œuvre par les participants pour produire des hypothèses scientifiques et construire des algorithmes.

3.3. La confrontation des projets aux donnees disponibles : les trajectoires d'exploration des participants

3.3.1. Le traitement des bases de données

Chaque projet a débuté avec la formulation d'un axe de travail. Celui-ci se construit à partir des thématiques proposées par Epidemium ainsi qu'avec le descriptif fourni des bases de données disponibles. En effet, chaque thème est associé par les organisateurs à un ensemble de bases de données jugées utiles pour la problématique. A partir des hypothèses, les participants cherchent ensuite à analyser les données disponibles pour établir une potentielle corrélation.

Dès le début de leur exploration, les participants ont à une mauvaise qualité des bases de données fournies par Epidemium : la qualité d'une base de données peut être définie comme le contraste entre la description de la base de données (le titre des colonnes et des lignes) et le contenu réel. C'est un problème récurrent en analyse de données notamment dans le cas de données ouvertes, car les bases de données sont rarement complètes et la valeur des cellules ne correspond pas nécessairement aux domaines de variation prévue dans leurs descriptions (colonnes et lignes). Bien que l'équipe organisatrice d'Epidemium a largement contribué à rendre homogène les bases de données provenant de sources hétérogènes, parmi les 21 000 bases de données proposées initialement, peu de données étaient de bonne qualité et donc facilement exploitable. Cela a poussé les équipes projets à improviser, développer des stratégies afin de créer de nouvelles bases de données en s'associant pour construire des données plus adaptées ou en allant chercher des bases de données non existantes dans le projet Epidemium.

C'est notamment le cas sur une partie des bases de données sur la mortalité et l'incidence du cancer. Le projet BD4Cancer s'est associé au projet Baseline afin de créer une nouvelle base de données, appelée EpidemiumDB. La collecte des données a été effectuée selon un processus standardisé conçu par les équipes projet : chaque personnes qui souhaitait contribuer pouvait s'intégrer au processus de collecte en choisissant une région du monde (pays, région d'un pays) et fournissait les données associées à chacune des régions sur les taux d'incidence et de mortalité de chaque cancer, ainsi que les informations sur les facteurs de risques connus. La particularité est que le processus ne s'est pas limité à la seule communauté d'Epidemium, et des contributeurs hors du programme sont intervenus sur la tâche pour constituer la base de données. Au total de plus de 200 participants ont contribué à collecter les données. Alors que les projets BD4Cancer et Baseline sont constitués d'un nombre relativement restreint de participants, ils ont pu mettre en place un processus pour permettre d'accéder à une communauté beaucoup plus large de contributeurs éphémères sur une tâche indépendamment du cadre d'Epidemium. Contrairement aux autres projets d'Epidemium, ces tâches ne répondent pas à une demande directe des organisateurs, mais sont pilotées entièrement par des équipes projets.

D'autres projets ont ajouté de nouvelles bases de données dans le programme Epidemium suivant leurs besoins propres. Par exemple, le projet BD4Cancer développait un outil de veille sanitaire en analysant sur les réseaux sociaux (Twitter) l'apparition d'effets secondaires quant à l'utilisation de médicaments anti-cancer. L'équipe projet a donc constitué deux nouvelles bases de données sur la liste des médicaments anti-cancer ainsi que la liste des effets secondaires associés, rendues ensuite librement accessible à la communauté Epidemium.

3.3.2. Modélisation de l'analyse des bases de données par les participants

3.3.2.1. Premières analyses statistiques sur les bases de données : le cas de Baseline

Une fois les bases de données collectées, nettoyées et traitées, les participants ont exploré l'espace du code informatique. La première exploration réalisée par l'équipe Baseline a consisté à analyser la base de données EpidemiumDB à partir de modèles statistiques basiques (de type régression multi-linéaires) afin de détecter de possibles corrélations entre facteurs de risques et incidence et mortalité du cancer. Leur étude a permis de mettre en avant des corrélations intéressantes pour l'épidémiologie. Ils ont notamment constaté qu'aux Etats-Unis, les populations afro américaines ont une plus grande incidence et mortalité du cancer de la prostate que les autres groupes ethniques. Cette première exploration a permis de formuler cette première hypothèse :

> $\mathcal{H}_1(\mathcal{L})$ = *{impact =* ***incidence, mortalité****}({type de cancer =* ***pancréas****}, {facteurs de risques =* ***changements environnementaux****}, {zone géographique =* ***Etats-Unis****}, {profil des patients =* ***population noire africaine****})*

Cette hypothèse permet de préciser un certain nombre de critères par rapport à l'axe de travail initialement formulé : le type de cancer (cancer du pancréas), le type d'impact observé (incidence, mortalité), la catégorie de facteurs de risque (environnementaux), la zone géographique (Etats-Unis) ainsi que le type de population (noire africaine). En plus de restreindre l'hypothèse de base en une nouvelle hypothèse potentiellement valide pour la communauté scientifique, l'équipe projet développe une méthode statistique pour construire le niveau de vérité de cette hypothèse. Il y a à la fois la construction du quoi (quelle hypothèse) et du comment (construire sa valeur de vérité). Ces premières analyses statistiques sont encourageantes mais pas suffisantes pour conclure d'une validité scientifique et nécessitent de nouvelles investigations avec des données différentes ou d'autres algorithmes. De plus l'hypothèse obtenue est déjà connue par la littérature en épidémiologie. En effet, des études montrent déjà que le risque de pancréatite est de 2 à 3 fois plus élevé chez la population noire que chez la population blanche (Yadav & Lowenfels, 2013). Pour autant, la découverte de l'hypothèse uniquement avec les données ouvertes disponibles constitue une preuve qu'il est possible d'utiliser les bases de données ouvertes pour découvrir des résultats scientifiques.

3.3.2.2. Deuxième exploration : la mise en place de deux tournois d'analyses des données

Suite à ces premiers résultats encourageants, les projets Baseline et BD4Cancer ont organisé un data challenge avec l'outil RAMP avec la base de données EpidemiumDB. L'objectif était d'explorer s'il était possible d'extraire des corrélations intéressantes entre les colonnes des bases de données à partir des algorithmes de machine learning. Les participants étaient incités à développer un modèle algorithmique sous la forme $Y = f(X)$, où Y est le taux de mortalité d'un certain cancer et X les possibles facteurs de risque associés. Plutôt que de se baser sur une

hypothèse canonique, les équipes projets ont préféré avoir une approche large, et les participants étaient libres de choisir leurs variables de corrélation : le choix d'un ou plusieurs cancers, la zone géographique, ou encore le type de facteurs de risque.

Le data challenge a réuni plus de 40 participants. Si beaucoup des participants provenaient des équipes projets Epidemium, certains n'appartenaient même pas au programme. Une telle démarche a permis aux équipes projets de profiter d'un grand nombre de spécialistes pour chercher à analyser les données, sans se limiter à leur entourage proche. En une après-midi, les participants ont développé environ 40 algorithmes dont plusieurs assez performants en terme de capacité de prédiction. Pourtant, ces algorithmes n'ont pas pu pour la plupart être exploité car les modèles sous-jacents ne permettaient pas d'en déduire une hypothèse. En effet, contrairement à la première analyse qui a permis de mettre en relation les populations noires africaines face au risque de cancer du pancréas, il n'était pas possible d'extraire de façon claire les liens entre facteurs de risques *X* et taux de mortalité *Y*. En fait, un seul algorithme de type « GLM aggregate » permettait d'établir des corrélations intéressantes : celui-ci estimait le poids de chaque facteur de risque en fonction de son effet sur la mortalité du cancer. Les facteurs de risque identifiés par l'algorithme étaient connus de la littérature : le chômage de longue durée, la consommation d'alcool ou encore le cholestérol. Bien que cette redécouverte de résultats grâce à de nouvelles méthodes est encourageante pour l'utilisation du machine learning dans le cadre de l'épidémiologie, aucune corrélation statistique n'a été découverte et le processus a été jugé globalement décevant par les équipes projets.

Un deuxième RAMP a été organisé par la suite. Cette fois, les équipes projets ont introduit la variable âge corrélée avec les risques de cancer, ce qui rendait la base de données beaucoup plus volumineuse et riche. De plus, les risques d'incidence du cancer ont également été intégrés en plus de la mortalité. La problématique était de modéliser les risques de mortalité par cancer digestif (intestin, côlon, rectum et anus, foie, vésicule biliaire) en fonction de l'âge et d'autres types de cancer. Forts du constat du premier RAMP, les équipes projets ont cherché à préciser la question de recherche initiale en s'intéressant à des problèmes moins agrégés (un seul type de cancer dans une seule région du monde sur une période courte et dans une couche assez étroite de la population par exemple) pour pouvoir fournir des modèles interprétables. Au total, 40 modèles ont été soumis par plus de 30 participants (soit une moyenne d'environ un modèle par participant) durant une après-midi. Sans apporter de réponse précise, plusieurs points ont été soulignés par cette deuxième analyse notamment :

- A2($\mathcal{L}$) : L'âge est de loin la variable la plus importante, suggérant que les cancers digestifs sont clairement associés au vieillissement avant tout (avant le tabagisme, soleil, alcool, etc.); des recherches biomédicales associant le vieillissement et le cancer pourraient donc apporter des solutions

- A3($\mathcal{L}$) : L'origine ethnique semble être une réalité médicale pour certains cancers, qu'il est nécessaire d'inclure dans certaines recherches. Par exemple, il semble important de rechercher des facteurs de risque communs entre Afro-Américains et Carabéens en Afrique.

Ces observations, bien que peu originales dans la littérature, permettent de restreindre l'exploration sur des concepts qui semblent adaptés aux bases de données disponibles.

L'équipe projet APRC a également profité de l'organisation du RAMP2 pour tester quelques algorithmes sur les données de la FAO et sur l'incidence du cancer du pancréas. Leurs premières analyses font ressortir que la variable la plus discriminante pour expliquer le cancer du pancréas, parmi les variables agro-environnementales, est la consommation d'énergie dans les secteurs de l'agriculture et de la foresterie en pourcentage du total de la consommation d'énergie. Autrement dit, l'analyse a permis de formuler l'hypothèse canonique suivante :

> $\mathcal{H}_2(\mathcal{L})$ = *{impact =* ***incidence****}({type de cancer =* ***pancréas****}, {facteurs de risques =* ***consommation d'énergie dans le secteur agricole et forestier****}, {zone géographique =* ***régions du monde****})*

L'exploration menée par l'équipe APRC a permis d'appliquer une fonction de transformation dans l'espace des hypothèses permettant de préciser plusieurs familles de concepts à partir de leur axe de travail initial : le type d'impact, le type de cancer, le choix d'un facteur de risque environnemental et une analyse basée sur des régions du monde.

3.3.3. Trajectoire des équipes projet durant le challenge : processus de reformulation

Tout au long du challenge les participants explorent à la fois l'espace des hypothèses et l'espace du code informatique pour produire des hypothèses et des algorithmes compatibles avec les données existantes et obtenir des résultats scientifiques. Pour réduire le risque de formuler une hypothèse non valide ou sans valeur, les participants commencent généralement avec une idée plus ou moins précise de ce qu'ils cherchent dans les données, même si cette idée est mal formulée. Ils peuvent alors par exemple s'intéresser à l'impact d'un type de facteur de risque sans préciser au départ quel type de cancer sera étudié. C'est seulement durant le processus d'exploration que l'hypothèse sera raffinée (application de la fonction de transformation). Ainsi, au lieu de démarrer l'exploration par une hypothèse valide vis-à-vis de la discipline, ils peuvent commencer à travailler sur une hypothèse de type « axe de travail », où la problématique n'est pas clairement formulée. Les stratégies d'exploration au sein des équipes projet sont sensiblement similaires et répondent à une procédure générale (voir **figure 50**) :

- *Formulation d'un axe de travail* : l'hypothèse initiale est volontairement sous-spécifiée pour limiter le risque de formuler une hypothèse non compatible avec les données disponibles. Cette hypothèse est incluse parmi les thèmes proposés par les organisateurs et se base sur la description des données fournie par les organisateurs. Les participants extraient une partie des variables ou de familles de variables existantes dans les bases de données pour formuler l'axe de travail.
- *Nettoyage, traitement et collecte de données* : Le nettoyage et le traitement des données est un élément qui n'avait pas été anticipé par les équipes projet, et qui a constitué une part importante de leur investissement dans leur participation au programme. De même, les organisateurs n'ont pas semblé mesurer l'importance de cette activité.
- *Exploration des bases de données* (**figure 51**) : Une fois les hypothèses choisies et les données nettoyées, les participants explorent les bases de données en appliquant des algorithmes d'analyses des données. L'application de séquences d'action sur les jeux de données permet de produire des relations entre différentes variables.
- *Reformulation de l'hypothèse* (**figure 52**): Les participants reformulent leur hypothèse initiale en fonction des relations entre les variables initiales et de nouvelles variables. Cette nouvelle hypothèse est plus restrictive et en même temps plus compatible avec les bases de données disponibles. Elle devient un nouveau point de départ pour explorer les bases de données.
- *Processus d'optimisation* : une fois que l'hypothèse découverte est valide d'un point de vue de la communauté scientifique, les équipes projets cherchent à optimiser le modèle algorithmique utilisé pour analyser les bases de données. Aucune équipe projet n'a abouti à cette étape durant le premier challenge, probablement par manque de temps.

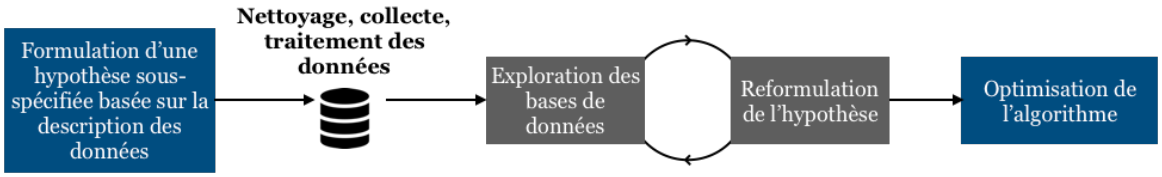


Figure 50. Processus d'exploration durant le premier challenge

Nous pouvons illustrer ce processus grâce au projet Baseline. Initialement, les chefs de projet souhaitaient prédire l'incidence, la mortalité et la survie du cancer en utilisant des facteurs de risque provenant de sources de données ouvertes (avec une portée mondiale et une granularité régionale). Rapidement, ils se sont rendus compte que la qualité des données était insuffisante pour les analyser et ont remplacé la base de données existante sur l'incidence et la mortalité du cancer par une nouvelle base de données, EpidemiumDB, de meilleure qualité. Cette base de données a ensuite été utilisé dans plusieurs challenges RAMP pour déterminer des relations entre des variables de la base. Cela a permis de reformuler l'axe de travail initial en de nouvelles hypothèses.

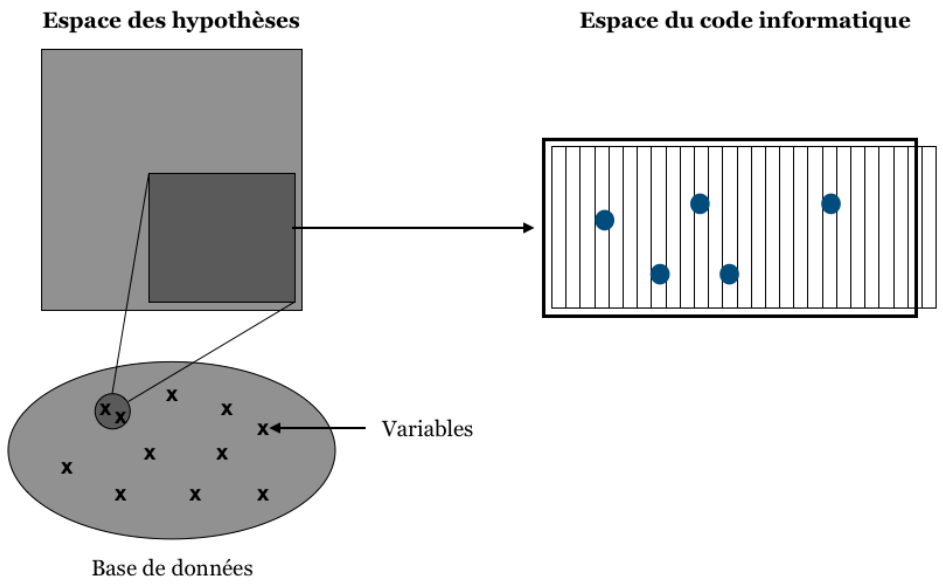


Figure 51. Formulation d'un axe de travail et exploration des bases de données

Le carré plus sombre dans l'espace des hypothèses modélise l'axe de travail initialement choisi par les participants. Cet axe de travail est peu restrictif, et couvre une large partie de l'espace d'action à explorer.

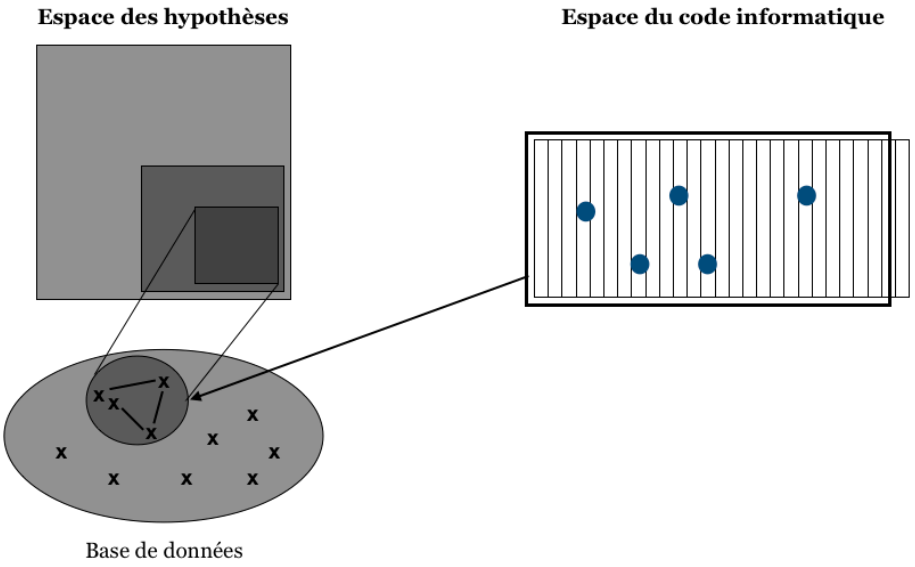


Figure 52. Reformulation de l'hypothèse à partir des relations qui ont été apprises sur les bases de données.

Les résultats statistiques obtenus par exploration de l'espace du code informatique produisent des relations entre des variables de la base de données. Ces relations étendent le nombre de variables prises en compte dans l'exploration et donc transforment l'axe de travail initial en nouvelle hypothèse.

3.4. Evaluer la production par les participants : l'accumulation de « stepping stones »

Quel est le bilan de l'exploration menée par les participants ? L'évaluation globale de la production du challenge est en demi-teinte quand il est observé uniquement sous le spectre de la production de résultat scientifique (production d'hypothèses scientifiques valides et intéressantes, élaboration d'algorithmes pour analyser les données). Cependant, cette nous allons voir que cette représentation est limitée pour rendre compte des efforts et des éléments produits durant le challenge. En effet, malgré les difficultés des participants à finaliser leurs projets, nous verrons que les participants ont produit non pas des résultats scientifiques valides mais ce que l'on définit comme des **stepping stones**, c'est-à-dire des résultats intermédiaires capables d'aider les scientifiques et les participants d'Epidemium à l'exploration de l'espace des hypothèses et du code informatique.

3.4.1. Evaluation par les organisateurs

L'ensemble de la production des projets représente un panel riche et hétérogène et engendre une difficulté pour les organisateurs et pour les comités pour développer une grille d'évaluation afin de déterminer les gagnants. En effet, le challenge avait été initié sans avoir de critère au départ pour déterminer ce qui avait de la valeur. Bien que les organisateurs aient supposé que cette variété serait bénéfique pour explorer l'espace, ils se sont vite rendus compte qu'il n'y avait pas de moyen facile d'évaluer des projets de nature très différente. Pour déterminer les gagnants, ils ont adopté une méthode *ad hoc* pour comparer les projets entre eux à partir d'une liste de critères **(voir annexe pour la grille complète)** :

- La clarté du projet et la pertinence de l'approche proposée,
- L'originalité du projet,
- Les méthodes de travail (travail collaboratif et complémentarité, appropriation du technologies et outils mis à disposition),
- Les résultats et conclusions (caractère innovant et travail accompli, compréhension et clarté des résultats),
- L'impact sur la santé des patients (pertinence médicale scientifique, utilisation et appropriation par la communauté médicale) et perspectives (vision à long terme, durée de vie estimée du projet).

À l'aide de ces critères, trois projets lauréats ont été choisis (dans l'ordre décroissant) : Baseline, CancerViz et ELSE. Les projets Baseline et BD4Cancer ont été déclarés comme étant les plus fédérateurs dans la communauté et les plus soutenus par des experts reconnus dans le domaine médical et l'analyse des données, ce qui leur a permis de s'étendre largement au-delà de la communauté Epidemium. Le projet CancerViz a été notamment récompensé par la qualité du projet et son approche pour la visualisation des indicateurs du cancer dans le temps. C'était le seul

notamment à proposer une approche multi-variable pour représenter l'évolution du cancer dans le temps. Enfin le projet ELSE a été retenu comme étant le plus original. L'équipe du projet ELSE a démarré le challenge seulement quelques semaines avant la fin et n'ont eu le temps de mettre en place qu'un prototype de leur idée d'outil. Pourtant, l'idée a largement séduit le comité et ils ont pu obtenir la troisième place du challenge.

Cette grille d'analyse est très éloignée de ce que l'on pourrait attendre d'un processus visant à produire de la connaissance scientifique. Au lieu de valoriser les hypothèses qui ont été produites, il semble que la grille d'analyse ait été conçue pour valoriser le travail fourni par les participants ainsi que l'originalité dans les approches proposées. Quels sont les résultats des explorations menées au sein des deux espaces n'est pas clair. Nous proposons tout d'abord d'évaluer la production à partir des métriques de performance que nous avons développé dans le chapitre 6. Nous analysons cette productivité en évaluant l'exploration sur l'espace des hypothèses (production d'hypothèses, qualité des données) et sur l'espace des actions (vérification des hypothèses, fonction de valeur).

Objet d'analyse	Métrique	Productivité
Espace des hypothèses	Analyse des axes de travail et des hypothèses canoniques produites	• Nombre d'axes de travail et d'hypothèses canoniques produites • Conversion en hypothèse canonique • Explorer toutes les données disponibles
	Qualité des ressources et extension de l'espace des hypothèses	• Qualité des données • Couvrir toute la littérature
Espace des actions	Test statistique	• Nombre d'hypothèses vérifiées – Avancement • Cohérence avec la littérature
Valeur des projets	Valeur scientifique	• Typologie de la valeur des productions

Tableau 15. Métriques de productivité durant le challenge

3.4.2. Evaluation des hypothèses et des bases de données

Au total, 3 projets finalistes sur 8 ont proposé des hypothèses scientifiques pour un total de cinq hypothèses de type axes de travail et deux hypothèses canoniques. Ce ratio est faible par rapport au nombre de participants (7 hypothèses sur 75 participants soit environ 9%). De plus la capacité à convertir un axe de travail en hypothèse canonique est mince (2 hypothèses canoniques sur 3 axes de travail). Enfin si plusieurs études ont été réalisées sur les hypothèses formulées, aucune n'a fourni un résultat scientifique fiable. En fait, bien que la plupart des hypothèses soient présentées comme cohérentes par rapport à la littérature, elles ne permettent pas de conclure à un résultat scientifique. Pourtant le premier challenge a été considéré comme une réussite par les

partenaires d'Epidemium comme Roche, l'Institut Curie et l'Institut Gustave Roussy. Les organisateurs et les partenaires ont plutôt eu tendance à juger l'engouement autour du projet Epidemium que de regarder la productivité. En effet, la difficulté majeure pour les projets de science citoyenne d'une telle envergure est de pouvoir rassembler une grande communauté autour d'un projet commun. Il est vrai que de ce point de vue, le processus a été réussi.

Projets	Hypothèse	Analyse par le langage	Bases de données
Hypothèses de type axe de travail			
Baseline	Etude des nouveaux facteurs de risque par le biais des méthodes de machine learning	$\mathcal{A}1$ = *{impact =* ***incidence, mortalité, survie****} ({type de cancer}, {zone géographique =* ***granularité régionale****})*	EpidemiumDB (incidence et mortalité des cancers par région du monde)
Baseline	Analyse de la mortalité dans les cancers en fonction du vieillissement de la population	$\mathcal{A}2$ = *{impact =* ***mortalité****} ({type de cancer}, {facteurs de risque =* ***vieillissement population****}, {zone géographique =* ***granularité régionale****})*	EpidemiumDB (incidence et mortalité des cancers par région du monde) + âge
Baseline	Analyse de l'impact de l'origine ethnique des patients dans l'incidence et la mortalité des cancers	$\mathcal{A}3$ = *{impact =* ***incidence, mortalité****} ({type de cancer}, {facteurs de risque =* ***origine ethnique****}, {zone géographique =* ***granularité régionale****})*	EpidemiumDB (incidence et mortalité des cancers par région du monde)
APRC	Identifier l'impact de plusieurs facteurs environnementaux sur le cancer	$\mathcal{A}4$ = *impact({type de cancer}, {facteurs de risque =* ***environnementaux****})*	EpidemiumDB et données FAO
BD4Cancer	Étude des effets indésirables des médicaments sur les patients grâce aux données des réseaux sociaux (système en temps réel) et aux essais cliniques (mécanismes génétiques)	$\mathcal{A}5$ = *{impact =* ***effets indésirables****}({type de cancer}, {facteurs de risque =* ***médicaments anti-cancers****})*	Liste médicaments, effets secondaires, liste des cancers
Hypothèses canoniques			
Baseline	Les risques d'être sujets au cancer du pancréas sont plus élevés pour la population noire africaine des Etats-Unis	$\mathcal{H}1(\mathcal{L})$ = *{impact =* ***incidence, mortalité****}({type de cancer =* ***pancréas****}, {facteurs de risques =* ***changements environnementaux****}, {zone géographique =* ***Etats-Unis****}, {profil des patients =* ***population noire africaine****})*	EpidemiumDB (incidence et mortalité des cancers par région du monde)
APRC	La consommation d'énergie dans le secteur agricole et forestier est liée à un facteur de risque d'apparition du cancer du pancréas	$\mathcal{H}2(\mathcal{L})$ = *{impact =* ***incidence****}({type de cancer =* ***pancréas****}, {facteurs de risques =* ***consommation d'énergie dans le secteur agricole et forestier****}, {zone géographique =* ***régions du monde****})*	EpidemiumDB et données FAO

Tableau 16. Liste des hypothèses formulées durant le challenge 1.

Une des difficultés principales pour les participants a été de travailler avec des bases de données de mauvaise qualité. En effet, une grande partie du travail réalisé par les équipes a consisté au nettoyage des données ainsi qu'à la collecte de bases de données supplémentaires. Il est ambitieux de dire que face à une telle quantité de bases de données (de l'ordre de 21 000) les organisateurs auraient pu nettoyer l'ensemble des bases de données pour les rendre exploitables. De ce point de vue, la stratégie qui consiste à laisser les participants s'occuper de cette tâche semble plus judicieuse. En effet, les participants ne vont pas chercher à nettoyer toutes les bases de données mais à priori uniquement celles dont ils ont besoin en fonction de l'hypothèse qu'ils auront initialement formulé. En revanche, le temps passé à nettoyer les bases de données limite nécessairement la productivité des équipes et donc la production finale du challenge. Les organisateurs doivent donc prendre en compte l'existence de cette tâche et décider s'ils doivent la traiter eux-mêmes ou la déléguer aux participants.

Plusieurs des équipes ont donc passé une grande partie de leur temps à travailler sur les bases de données afin de partir sur des bases de meilleure qualité. Dans le challenge, la collaboration entre les projets Baseline et BD4Cancer a permis de constituer une base de données EpidemiumDB plus précise que celle existante auparavant sur l'incidence et la mortalité des cancers en fonction des régions du monde. Le projet Oncobase a également produit une base de données unifiée pour la communauté Epidemium, mêlant des données issues de sources diverses et variées, et permettant de construire des analyses statistiques solides. L'équipe projet s'est concentrée sur la constitution d'un algorithme permettant d'automatiser des procédures visant à faciliter l'exploitation de données ouvertes. La plus grande part de leur activité a consisté à structurer ces bases de données, notamment en travaillant sur la forme des fichiers (nettoyage, tri, classement, regroupement). Enfin le projet Venn a développé le prototype d'un algorithme capable de croiser les articles scientifiques sur la recherche contre le cancer afin de mettre en lumière les liens entre cancer et pollution de l'air mis en évidence par la recherche. La base de données de départ est constituée de plus de deux millions d'articles scientifiques de la base PubMed dans lesquels l'équipe a extrait les informations qui lui étaient nécessaires : le titre, la date de publication, les mots-clés associés, le journal où l'article a été publié, le laboratoire d'origine du premier auteur. Une sélection des articles a été faite grâce à la construction d'un lexique des termes entre cancer et pollution de l'air à partir d'un monographe d'un centre de recherche sur le sujet[9]. Les articles ont ensuite été catégorisés suivant les mots clés pour y accéder facilement à partir d'un moteur de recherche.

3.4.3. Evaluation de la production dans l'espace du code informatique

L'exploration de l'espace du code informatique durant les challenges RAMP 1 et 2 a permis de construire plus de 80 modèles informatiques pour analyser les données. Deux de ces algorithmes ont amené à formuler des hypothèses scientifiques potentiellement valides en épidémiologie

[9] http://publications.iarc.fr/Book-And-Report-Series/Iarc-Scientific-Publications/Air-Pollution-And-Cancer-2013

($\mathcal{H}_1(\mathcal{L})$ et $\mathcal{H}_2(\mathcal{L})$), tandis qu'une autre partie d'entre eux a servi à construire deux nouveaux axes de travail ($\mathcal{A}_2(\mathcal{L})$ et $\mathcal{A}_3(\mathcal{L})$). Cependant, aucun des modèles statistiques proposés ne permet d'avoir des résultats compatibles avec les exigences scientifiques. En effet, la corrélation possible mesurée entre les différentes variables des bases de données n'est pas suffisante pour statuer d'une causalité épidémiologique. Améliorer les algorithmes existants demande de pousser l'explorer du code informatique et de vérifier la qualité des données utilisées.

3.4.4. Conception d'outils pour explorer les espaces

Certaines équipes ont préféré se concentrer sur la fabrication d'outils permettant de faciliter l'exploration des données existantes. Si ces équipes n'ont pas travaillé directement à la production d'hypothèses scientifiques, leurs outils peuvent ouvrir des pistes intéressantes pour les travaux futurs d'Epidemium et constituent des étapes intermédiaires potentielles à la génération d'hypothèses. Par exemple, le projet Venn a pour objectif de développer un outil pour faire émerger des corrélations entre cancer et facteurs de risque à partir des connaissances issues de la littérature. Les projets Viz4Cancer et CancerViz ont également développé des outils pour visualiser l'évolution du cancer en fonction de différents facteurs. La représentation des bases de données est souvent essentielle pour les professionnels du métier afin d'explorer facilement une base de données et formuler des hypothèses de corrélations entre différents facteurs, notamment dans le cadre de l'épidémiologie (Shelly, By, & Birnbaum, 1996). Le projet Viz4Cancer a développé un premier prototype d'un site web pour la visualisation de données. En plus de la visualisation, ils ont construit un modèle prédictif simple pour estimer les tendances des courbes représentées sur les prochaines années à venir. Bien qu'ayant démarré quelques jours avant la date de rendu, le projet CancerViz est allé plus loin en proposant le prototype d'un logiciel capable de proposer des visualisations interactives multicritères pour faciliter le choix par les scientifiques des données qui ont un intérêt scientifique potentiel. Bien que ces projets soient à l'état de prototype et nécessitent encore un temps de développement, ils ouvrent la voie à des outils capables d'aider les scientifiques et les participants d'Epidemium à l'exploration de l'espace des hypothèses et d'aider à la formulation.

3.4.5. Une production d'états intermédiaires : les « stepping stones »

Au final, aucune équipe n'a abouti à un résultat final satisfaisant, c'est-à-dire la génération d'hypothèses scientifiques dont le statut logique a été validée par les données disponibles. Tous les projets ont terminé sur des prototypes d'outils d'aide à l'exploration, des hypothèses non valides scientifiquement (axes de travail) ou des hypothèses scientifiques qui n'ont pas pu être vérifiées par une méthode algorithmique. Cela n'est pas étonnant : nous avons vu que la plus grande difficulté des équipes a été de pouvoir terminer les projets à temps (Sitruk & Kazakçi, 2018). De plus, les équipes ont du intégrer le nettoyage et la collecte de nouvelles bases de données dans le temps qui leur était imparti. Aussi, évaluer les projets uniquement sur leur capacité à aboutir à

des résultats finaux ne semble pas pertinent pour prendre en compte tout ce qui a été produit durant le challenge.

Pourtant, il serait inapproprié de qualifier le challenge d'échec. Alors que les axes de travail formulés au début étaient assez flous, l'exploration de l'espace des hypothèses a permis de construire de nouveaux axes de travail ou hypothèses qui sont plus compatibles avec les bases de données. De la même manière, les algorithmes développés par les participants fournissent des bases intéressantes pour continuer l'exploration de nouveaux algorithmes basés sur les données existantes. En fait, toutes les productions durant le challenge (hypothèses, algorithmes, outils, bases de données,...) peuvent être vues comme des **stepping stones**, c'est-à-dire des étapes intermédiaires dans le processus d'exploration.

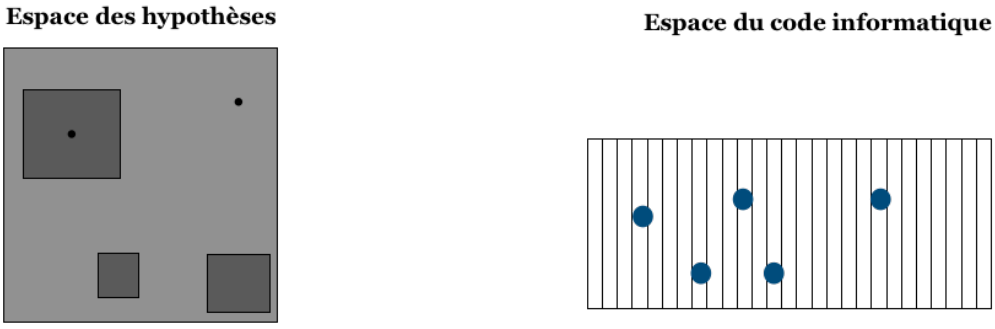


Figure 53. Illustration de l'exploration des espaces à la fin du challenge.

Entre le début et la fin du challenge, plusieurs parties des espaces des hypothèses et du code informatique a été explorée, améliorant la compréhension de certaines zones des ces espaces. Au lieu de démarrer les prochaines explorations par des thématiques *ad hoc*, les participants des prochains challenges peuvent s'appuyer sur ces stepping stones qui peuvent devenir de nouveaux points d'entrée. Cela permet de capitaliser sur ce qui est déjà connu et donc potentiellement de réduire les coûts d'exploration des prochains participants.

4. Organisation et dispositifs de gestion au sein d'Epidemium

Quels ont été les processus organisationnels mis en œuvre durant le challenge 1 ? Comment les organisateurs ont capitalisé sur la production ? Dans cette section, nous établissons un bilan critique de la gestion du processus d'ouverture que l'on a pu observer dans le cadre d'Epidemium. Nous montrerons que la capitalisation entre les tâches est un élément essentiel pour augmenter les chances de construire des résultats scientifiques et que celle-ci nécessite d'être gérée.

4.1. UNE GRANDE LIBERTE ORGANISATIONNELLE : EMERGENCE DE « SOUS-COMMUNAUTES EPHEMERES »

Il n'y a pas de supervision directe de l'exploration par les organisateurs d'Epidemium durant le challenge. En fait, les organisateurs et les comités ont plutôt un rôle d'évaluateur des projets et n'ont pas autorité à imposer les orientations stratégiques. Cette grande liberté dans le choix des projets se retrouve également dans les thématiques formulées en amont par les organisateurs. Cela est particulièrement flagrant dans le projet *ELSE* qui concerne un aspect pédagogique de l'épidémiologie, bien éloigné des thématiques proposées par Epidemium au départ. Par ailleurs, les participants ont également la possibilité de s'organiser comme ils le souhaitent et les équipes projet organisent elles-mêmes la répartition des activités entre les différents participants. Dans tous les cas rencontrés émergent ainsi un porteur de projet qui va structurer le travail à l'intérieur des équipes. Grâce aux multiples interactions durant le processus, les porteurs de projet sollicitent différents partenaires au sein de la communauté Epidemium pour les intégrer à leurs projets. L'opportunité d'accéder à une ressource d'acteurs avec des compétences diverses est à la fois un moteur pour les organisateurs d'Epidemium mais également pour les porteurs de projet eux-mêmes. Chaque membre est généralement intégré aux projets en fonction des compétences qu'il possède : analyse des données, informatique, spécialiste en épidémiologie ou en santé publique.

Les organisateurs d'Epidemium poussent également à la collaboration entre les équipes projet en valorisant la collaboration (cf. grille d'analyse des projets) ainsi qu'en organisant des rencontres hebdomadaires des membres de la communauté dans les locaux de la Paillasse. Cela a mené à deux situations de collaboration entre des équipes projet durant le challenge. Dans le premier cas, les équipes Baseline et BD4Cancer avaient identifié que les bases de données qu'ils souhaitaient utiliser étaient de mauvaise qualité et ont décidé de reconstituer une nouvelle base de données de meilleure qualité. Le travail de collecte consistait à trouver un certain nombre d'informations relatives au cancer, à sa mortalité et son incidence pour chaque pays. La tâche à exécuter pouvait donc être réduite à une tâche de type recette. Les porteurs de projet ont conçu une grille standardisée qui pouvait être déployée pour chaque région du monde. Cumulée, la collecte des données pour chaque région du monde aurait pris beaucoup de temps aux équipes projet et auraient limité leur exploration durant le challenge. Les porteurs de projets des équipes ont donc décidé de déléguer la tâche. S'appuyant sur le modèle d'Epidemium, ils ont ouvert la collecte de données à toute personne intéressée à contribuer. La portée du projet est allée au-delà du programme Epidemium puisque des personnes hors de la communauté Epidemium ont participé. Dans le deuxième cas, le projet Baseline cherchait à analyser les données qu'ils avaient collecté en partant de leur hypothèse de départ. Au lieu de travailler uniquement avec les membres du projet Baseline, ils ont décidé d'utiliser la plateforme RAMP pour mettre en place deux data challenges basés sur les données EpidemiumDB. Ce processus leur a permis de générer 40 algorithmes différents en une après-midi développé par 40 participants, dont certains ne faisaient pas partie de la communauté Epidemium.

Dans les deux cas, les porteurs de projet isolent une tâche bien spécifiée (élémentaire, recette ou résolution de problèmes) qu'ils délèguent à la foule. Contrairement aux projets Epidemium, ces tâches ne répondent pas à une demande directe des organisateurs, mais sont pilotées entièrement par les équipes projets. Cette organisation est unique dans le cadre des projets ouverts. En effet, de manière générale les projets d'ouverture sont conçus pour que les participants travaillent chacun de façon indépendante des autres participants. Dans le cas d'Epidemium, les thématiques proposées aux participants sont larges, et les équipes projet sont susceptibles de partager des tâches similaires, favorisant la collaboration. Les participants ne sont plus uniquement des exécutants, mais deviennent eux-mêmes organisateurs de projets collaboratifs. Via un système de communication indépendante du projet chapeau, ils constituent une autre communauté pour résoudre la tâche, que l'on appelle « **sous-communauté éphémère** ». La création de ces sous-communautés permet aux porteurs de projet de réaliser des tâches dans le temps imparti par le challenge qu'ils n'auraient probablement pas eu la possibilité de faire avec leurs propres ressources. Ces sous-communautés présentent plusieurs caractéristiques :

- *Indépendance* : la délégation de la tâche ne dépend pas du projet chapeau (dans ce cas Epidemium) et donc n'est pas limitée à la communauté qu'il constitue
- *Extension* : aucune limite de taille dans le nombre de participants
- *Ephémère* : la communauté créée est non pérenne dans le temps et se dissout dès que la tâche est terminée
- *Tâche bien déterminée* : la tâche déléguée à la communauté est bien définie et peut être réductible à une tâche de type élémentaire, recette ou résolution de problèmes
- *Ressources limitées* : les organisateurs de la sous-communauté n'ont pas suffisamment de ressources pour réaliser la tâche

Cette organisation du collectif laisse beaucoup de libertés aux participants : c'est en effet une des valeurs souvent revendiquée dans les projets de science citoyenne. Cependant, nous avons vu qu'il existait un risque important de se perdre dans l'exploration des deux espaces et de ne pas produire de choses intéressantes. Au-delà des valeurs morales recherchées dans ce type de projet, cette tâche demande une gouvernance de la performance et de la fiabilité du système. La question de la gestion de la capitalisation se pose. Est-ce qu'il y a de la capitalisation durant la tâche ? De plus, nous avons vu que les challenges produisaient principalement des stepping stones, sans que les organisateurs n'aient introduit d'outils de gestion pour capitaliser sur cette production entre les tâches. Comment s'est organisé cette organisation *ad hoc* ? Quels moyens peuvent être mis en œuvre pour améliorer cette capitalisation ?

4.2. FAIBLE CAPITALISATION DURANT LA TACHE

Contrairement au RAMP, les participants travaillent individuellement ou en équipe et interagissent peu durant le processus. Chaque équipe commence son exploration par la formulation d'un axe de travail qui lui est propre. La probabilité que les zones de l'espace définies

par l'axe de travail se recoupent est faible, et donc il y a peu de chances que les participants puissent capitaliser sur les codes informatiques des autres ou les hypothèses générées. En revanche, nous avons vu que les équipes projets partagent des besoins communs sur les bases de données ou sur l'utilisation d'outils pour faciliter l'exploration. Par exemple, la base de données EpidemiumDB a été utilisée par plusieurs équipes projets qui n'avaient pas participé à sa construction, comme l'équipe *APRC*.

4.3. De la capitalisation « sauvage » a la mise en place d'un outil de gestion de la valeur

Epidemium a mis en place un deuxième challenge le 6 juin 2017 (fin mars 2018) suite au premier Challenge4Cancer. Les acteurs du premier challenge (organisateurs, instituts partenaires, participants) ont établi auparavant un bilan de ce qu'ils avaient appris pour déterminer les directions futures à prendre.

4.3.1. Capitalisation par les organisateurs d'Epidemium

Les organisateurs d'Epidemium ont cherché à améliorer le fonctionnement du challenge sur plusieurs points. Premièrement, ils ont constaté un manque de transdisciplinarité au sein des participants. En effet, seul un projet (Baseline) intégrait toutes les compétences médicales nécessaires (cancérologie, épidémiologie, santé publique) et pouvait garantir de la validité des projets vis-à-vis de la littérature et de la méthode scientifique. Cela a limité la qualité des projets d'un point de vue de la cohérence scientifique et la plupart des études menées ont été sur des hypothèses agrégées d'où il aurait été difficile d'en extraire des connaissances scientifiques. L'équipe organisatrice ainsi que les laboratoires Roche ont alors décidé de mettre en place une communauté de scientifiques experts du domaine, et facilement accessibles tout au long du processus par les participants. Cette communauté peut être consultée régulièrement et servir de support pour les non experts afin de les guider dans le processus scientifique. Deuxièmement, il y a eu beaucoup de reformulation entre les projets de départ et les rendus finaux. Bien que ce processus d'exploration soit inévitable, il est coûteux en temps et en ressource. Afin de maximiser la productivité et réduire les pertes, les thématiques du deuxième challenge ont été axées sur des problématiques plus adaptées vis-à-vis des expertises dominantes des participants, principalement des informaticiens ou des spécialistes de l'analyse de données. Deux thématiques ont donc été formulées pour le challenge 2 :

- Construire une visualisation de données de l'incidence des cancers en exposant les facteurs épidémiologiques associés à leur dynamique;
- Développer un outil prédictif pour la progression du cancer dans le temps et dans l'espace, en fonction des facteurs connus ou supposés qui déterminent son évolution.

Enfin, le projet Epidemium a suscité beaucoup d'enthousiasme lors du premier challenge et de nombreuses écoles d'ingénieurs françaises renommées, telles que Centrale-Supélec et Polytechnique, étaient intéressées par l'utilisation du challenge comme plateforme pour des projets étudiants. Les organisateurs ont donc mis en place une thématique supplémentaire pour les étudiants en machine learning afin de *prédire la mortalité par cancer dans les pays en voie de développement*. Les jeux de données ont été simplifiés et adaptés pour qu'ils correspondent aux compétences des étudiants.

4.3.2. Capitalisation par les instituts partenaires, les financeurs et les participants

Les laboratoires Roche et les instituts partenaires ont également tiré quelques enseignements du premier challenge. Similairement aux organisateurs, ils ont constaté que les projets étaient de manière générale peu cohérents avec la littérature scientifique. Ainsi, en plus de la communauté de scientifiques experts mise en place, les laboratoires ont demandé à ce que les projets soient systématiquement associés à la littérature scientifique afin de pouvoir juger de leur validité. En fin du challenge 2 les équipes projets ont dû présenter une question de recherche basée sur une revue de littérature scientifique, puis détailler un protocole de recherche. En fait, plutôt que de se baser sur la première méthode d'évaluation *ad hoc*, la grille d'analyse des projets a été très fortement inspirée des méthodes utilisées pour évaluer un résultat scientifique. Deuxièmement, si les participant ont eu des difficultés à structurer scientifiquement les projets, les instituts ont constaté le très fort engouement autour d'Epidemium et les nombreuses contributions potentielles. Durant le deuxième challenge, les représentants des instituts ont été plus actifs dans les projets. De plus de nouveaux partenaires sont apparus comme le centre de recherche CLARA à Lyon. Si le premier challenge était très localisé géographiquement, les instituts et les organisateurs ont cherché à internationaliser la démarche, en traduisant le challenge en anglais et en élargissant les réseaux de communication.

Enfin, certains participants du premier challenge ont été force de proposition pour mettre en place de nouveaux projets durant le deuxième challenge, et ont pris en considération la mauvaise qualité des données qu'ils avaient pu rencontrer lors du challenge 1. Un des projets issus de l'équipe *APRC* a notamment travaillé à la réalisation d'un algorithme permettant de faciliter le nettoyage et le remplissage des bases de données.

	Problèmes soulevés dans le Challenge 1	Solutions intégrées dans le Challenge 2
Equipe organisatrice Epidemium	Manque de transdisciplinarité (très peu de médecins)	Création d'une communauté de scientifiques au service des projets
	Majoritairement des spécialistes de l'analyse de données	Recentrage sur des thématiques techniques, thème 3 adapté aux étudiants
	Coût de reformulation des hypothèses élevé	Thématiques plus précises et plus ciblées
Instituts partenaires et financeurs	Manque de valeur scientifique (hypothèses agrégées)	Compatibilité avec la littérature
	Capacité à créer des ressources éphémères à volonté	Notoriété auprès d'instituts scientifiques (nouvelles collaborations)
Participants	Données de mauvaise qualité	Projet d'algorithme pour nettoyer automatiquement les données

Tableau 17. Synthèse de la capitalisation entre les deux challenges

4.3.3. De la capitalisation « sauvage » à la capitalisation via un outil de pilotage de la valeur

Chaque partie prenante du programme Epidemium a cherché à améliorer le processus à partir de ce qu'elles ont constaté durant le premier challenge. Pourtant, plusieurs limites à cette organisation montre la nécessité de mettre en place un système de gestion adapté. Premièrement la capitalisation se base essentiellement sur une vision tronquée du résultat du premier challenge, où les parties prenantes voient le premier challenge comme une réussite d'un point de vue de la communication et de l'investissement mais un échec en terme de productivité. Ils ne prennent pas ou très peu en compte ce que nous avons défini comme des stepping stones ce qui mène à une perte potentiellement importante de la productivité. Deuxièmement, la capitalisation a été répartie entre tous les acteurs du projet : organisateurs, financeurs, participants. Or, il semble délicat de déléguer une partie de la capitalisation aux participants, ces derniers n'étant contractuellement pas impliqués dans le programme Epidemium. En effet, dans de nombreux projets de sciences citoyennes, la majorité des participants ne font que des contributions petites et peu fréquentes, s'arrêtant souvent rapidement après leur intégration dans le projet (Franzoni & Sauermann, 2014). Enfin, la capitalisation mise en œuvre dans le programme Epidemium ne formalise pas ce qui a été produit durant le challenge, augmentant le risque de perte.

Bien que la plupart des projets ne soient rendus qu'à l'état de prototype ou stepping stones, une cartographie de la production permet d'évaluer quels sont les projets qui méritent un approfondissement et les projets dont la valeur n'est pas explicite ou peu intéressante pour mériter qu'on s'y attarde. Ce problème n'existe pas dans le cas de tâche de type résolution de problèmes. En effet, dans ce cadre toute production est évaluée au travers d'une fonction de valeur préalablement déterminée. Cette fonction est essentielle afin de déterminer si l'exploration va dans un sens où la valeur s'améliore. Dans le cas d'Epidemium, l'exploration des espaces est réalisée au travers d'une succession de challenges où les organisateurs sont les garants des

directions à prendre dans le projet. Nous proposons d'élaborer des critères de valeur en collaboration avec les organisateurs. Chaque stepping stone peut être évalué via ce système de valeur qui servira ensuite de support d'aide à la décision afin de savoir si l'exploration doit être poursuivie dans les zones de l'espace déjà explorées.

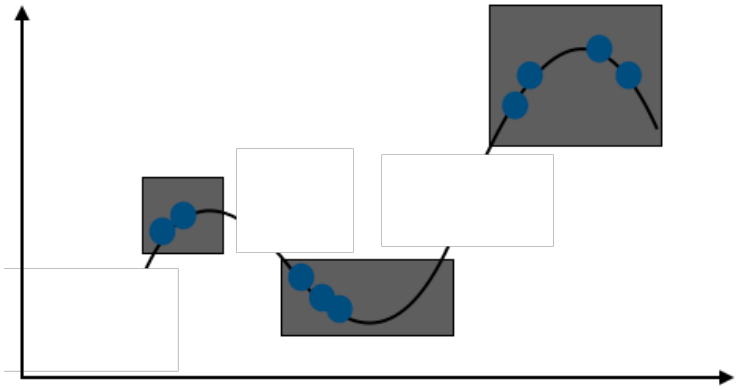


Figure 54. Les zones explorées par les participants (en gris) donnent des indications sur la fonction de valeur.

Pour élaborer cette grille, nous avons collaboré avec les organisateurs et le comité scientifique. A noter que celle-ci a été construite après le démarrage du deuxième challenge et n'a pas été utilisé pour capitaliser entre le premier et le deuxième challenge. Plusieurs éléments ont été pris en compte pour élaborer cette grille :

- *Valeur scientifique* : celle-ci correspond à l'intérêt scientifique du potentiel résultat obtenu et de l'hypothèse générée dans le domaine de l'épidémiologie du cancer.
- *Compatibilité avec les données existantes*
- *Politique publique* : certains projets peuvent avoir un faible impact en terme de résultat scientifique mais peuvent être utiles en terme de politique publique comme instrument scientifique d'aide à la décision
- *Interaction entre acteurs potentiels* : nous avons également évalué si les produits générés durant le challenge avaient le potentiel d'améliorer la communication entre les épidémiologistes et les autres acteurs de l'analyse Big data : spécialistes des données, citoyens de la science, patients, sociologues et d'autres.
- *Originalité* : nous avons intégré la notion d'originalité pour justifier l'intérêt de projets n'ayant aucune valeur scientifique mais apportant une vision originale des bases de données existantes. C'est le cas notamment du projet ELSE dont l'objectif était de construire un outil afin de sensibiliser une personne aux facteurs de risque du cancer en fonction de son style de vie et de son environnement.

Pour chaque case, les organisateurs ont été invités à donner un niveau de valeur en fonction d'une échelle que nous leur avions fourni. L'échelle va de « 0 » à « xxx » ; 0 est associé à une valeur non

explicite ; et xxx un projet pour lequel la valeur est avérée. La valeur « - » est utilisée lorsque les participants ont démontré le manque de valeur.

	Science	Compatibilité données	Politique publique	Interaction acteurs potentiels	Originalité	Total
H1	xx	x	x	xx	0	40%
H2	xx	xx	x	x	0	40%
A1	xx	xx	x	x	0	40%
A2	xx	xx	x	x	0	40%
A3	xx	xx	x	xx	0	47%
A4	xx	xx	x	x	0	40%
A5	x	0	x	x	0	20%
Outil de sensibilisation (*ELSE*)	0	0	x	xx	xx	33%
Méta-épidémiologie (*VENN*)	xx	xx	x	x	x	47%
Méta-épidémiologie (*Oncobase*)	xx	xx	x	x	x	47%
Visualisation de données (*CancerViz*)	x	xx	0	xx	x	40%
Visualisation de données (*Viz4Cancer*)	0	x	0	x	0	13%
Base de données *Oncobase*	x	xx	0	0	0	20%
Base de données EpidemiumDB	x	xx	0	0	0	20%

Tableau 18. Analyse de la valeur par projet du challenge 1 Epidemium.

Bien que la plupart des projets ne soient rendus qu'à l'état de prototype, cette représentation permet d'évaluer quels sont les projets qui méritent un approfondissement et les projets dont la valeur n'est pas explicite ou peu intéressante pour mériter qu'on s'y attarde. Nous avons proposé dans la dernière colonne un exemple d'agrégation de la valeur de chaque projet (moyenne non pondérée de chaque colonne). Avec cette représentation, les projets ayant la plus grande valeur potentielle sont ceux de Baseline et les deux projets de méta-épidémiologie, *Oncabase* et *Venn*. Notons que les résultats diffèrent de la grille d'évaluation utilisée par Epidemium.

CHAPITRE 9 – GESTION DES TACHES COUPLEES INVENTIVES PAR PROJETS SUCCESSIFS PAR EXTENSION DES CRITERES DE PERFORMANCE

RESUME DU CHAPITRE 9

Ce chapitre a pour ambition d'analyser les effets de la capitalisation sur le challenge 2 du programme Epidemium. Dans un premier temps, nous analyserons le déroulement du challenge 2 et nous présenterons les résultats obtenus. Nous montrerons que les bonnes performances du challenge 2 sont principalement dues à la capitalisation menée par les participants eux-mêmes. Nous mettrons ensuite en avant un effet de double extension qui ne pouvait être observé sur un seul challenge. D'abord une extension dans la taille des espaces explorés avec l'apparition de projets qui ne proviennent pas des bases de données disponibles. Ensuite une extension des fonctions de valeur via des projets qui ne peuvent être évalué uniquement avec les critères de valeurs qui ont été créés durant le challenge 1. Nous suggérons que ces phénomènes d'extension doivent être gérés durant le processus de capitalisation.

Nous avons vu dans le chapitre 8 que la capitalisation séquentielle dans le cas de tâches couplées inventives impliquait de piloter la transmission entre les projets de l'avancement de l'exploration des espaces. Dans ce chapitre, nous étudions les effets de la capitalisation qui a été mise en œuvre par les organisateurs d'Epidemium sur la performance du deuxième challenge, et nous tirons des conclusions sur les moyens à mettre en œuvre pour piloter la capitalisation séquentielle.

1. BILAN GLOBAL DU DEUXIEME CHALLENGE

1.1. UNE PLUS GRANDE COMMUNAUTE AVEC MOINS DE PARTICIPATION

Entre la fin du premier et du deuxième challenge la communauté Epidemium, définie comme le nombre de personnes inscrites aux meetups et aux newsletter, a doublé passant de 600 à 1200 inscrits. En parallèle, le nombre de participants actifs a diminué sur le deuxième challenge : 22 volontaires ont participé sur les deux premières thématiques tandis que le thème 3 a regroupé 32 étudiants issus des écoles Centrale Paris, Polytechnique et ESIEA. Cela représente un total de 54 membres (dont 32 étudiants), soit un taux de transformation de 4.5% (2% sans les étudiants). Ce taux est beaucoup plus faible que durant le premier challenge, où il y a eu environ 10% de la communauté ayant été active. Les raisons sont multiples. Les participants évoquent des barrières à l'entrée plus élevée que durant le premier challenge rendant la participation plus complexe. En effet, les organisateurs ont imposé la cohérence du projet avec la littérature scientifique. De ce fait, les conditions pour soumettre une proposition ont demandé un travail plus structuré et plus approfondi pour les participants. Deuxièmement, certains participants ont reconnu qu'il y avait une double difficulté d'acculturation pour intégrer le programme Epidemium : d'abord pour comprendre le monde relatif au domaine médical et scientifique, d'autant plus renforcée par l'obligation de fournir un résultat scientifique. Ensuite la mise en pratique d'une culture de l'ouverture et de la collaboration qu'on ne retrouve pas traditionnellement dans les projets scientifiques. Enfin, le projet Epidemium ne bénéficie plus de l'effet de nouveauté.

1.2. PRESENTATION DES PROJETS

13 équipes ont proposé un projet, dont 10 sont arrivés jusqu'au bout en soumettant leur projet au comité à la fin du challenge. Sur les 10 projets, 5 sont des projets d'étudiants relatifs à la thématique 3. Les 10 projets diffèrent dans leur approche suivant les thématiques auxquelles ils sont raccrochés (**tableau 19**). La thématique 1 ne contient qu'un seul projet, *IDEA*, dont l'objectif est de développer un outil de visualisation des données basé sur un algorithme de machine learning. Le principe est qu'à chaque fois que l'outil sera utilisé, la visualisation qu'il aura choisie (histogramme, carte, courbes,…) sera notée par l'utilisateur en fonction de plusieurs paramètres comme le type d'utilisateur ou le type de base de données. Ainsi plus l'outil est utilisé, plus il « apprend » de ses utilisateurs afin de fournir le type de visualisation le plus adapté. Ce projet est réalisé par des employés de l'entreprise CONIX, qui avaient déjà participé au premier

challenge dans le cadre du projet ELSE. Une deuxième catégorie d'équipes issus de la thématique 2 a cherché à développer des outils algorithmiques pour prédire la progression du cancer dans le temps et dans l'espace en fonction du régime alimentaire (*Cancer Diet*), pour étudier l'impact des essais cliniques (*AROUND*) ou encore pour aider les autorités publiques à prendre des décisions en terme de santé publique (*Locapred*). Une troisième catégorie concerne les projets étudiants sur la prédiction de la mortalité des cancers dans les pays en voie de développement (*Prévenir pour mieux guérir, Cancerinfl, Osy3a, Oma, Octopus*). Contrairement aux thématiques proposées durant le challenge 1, la thématique 3 se concentre sur un nombre de variables limitées et cherche à multiplier les explorations dans un espace restreint de l'espace des hypothèses. Pour rappel, la thématique 3 est de *prédire la mortalité par cancer dans les pays en voie de développement*. Dans cette thématique, les organisateurs précisent le type d'impact {mortalité}, demandent à ne travailler que sur un seul type de cancer, et réduisent l'exploration aux {pays en voie de développement}.

Enfin, un projet unique est issu d'une association appelée *Cancer Au Travail (CAT)*, qui s'intéresse à un champ encore peu exploré par la littérature scientifique : l'impact du modèle social d'un pays, sa vitalité démocratique, ses conditions de travail ou son modèle de production et leur impact sur la survivance du citoyen ? Cette notion de survivance fait émerger un concept qui n'existe pas dans la littérature médicale ni dans les données Epidemium, et qui fait référence à l'étude des patients guéris du cancer. Habituellement, la littérature médicale s'intéresse à cette population quand les scientifiques cherchent à quantifier les taux de survie au cancer et étudier l'efficacité d'un traitement. Ici, le projet CAT propose d'étudier le patient dans son environnement social, et l'impact que cet environnement peut avoir sur sa guérison ou sa rechute. Cela se modélise dans l'espace des hypothèses comme une extension de l'espace par l'intégration d'une nouvelle relation (**figure 55**), définie comme {survivance}. L'exploration ici n'est donc plus limitée aux seules données disponibles ou existantes, mais peut s'étendre par l'apparition de nouveaux mots.

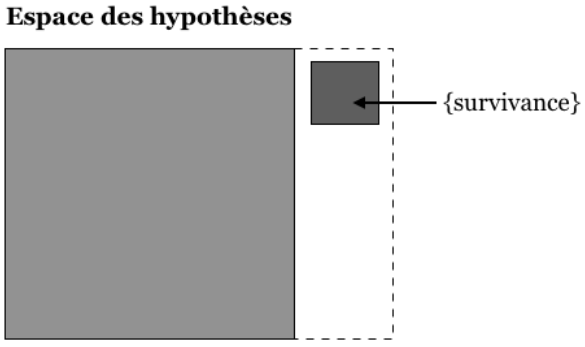


Figure 55. Extension de l'espace des hypothèses par le projet CAT et la notion de « survivance ».

Thème 1 - Construire une visualisation de données de l'incidence des cancers en exposant les facteurs épidémiologiques associés à leur dynamique		
IDEA	Développer un outil de visualisation des données sur le cancer qui va apprendre la bonne représentation en fonction des bases de données, du type d'utilisateur,... pour proposer la solution la plus adaptée	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
Thème 2 - Développer un outil prédictif pour la progression du cancer dans le temps et dans l'espace, en fonction des facteurs connus ou supposés qui déterminent son évolution.		
Cancer Au Travail	Déterminer le niveau de survivance du cancer vis-à-vis des facteurs de risques sociétaux (type de sécurité sociale, travail,...)	$\mathcal{A}$ = *{impact =* ***survivance****} ({type de cancer}, {facteurs de risque =* ***sociétaux****})*
Locapred	Construire un outil prédictif afin d'aider les autorités à prendre des décisions en terme de santé publique	Développement d'outils pour faciliter l'exploration de l'espace du code informatique
Cancer Diet	Développer un algorithme pour prédire le taux de mortalité du cancer en fonction du régime alimentaire	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer}, {facteurs de risque =* ***régime alimentaire****})*
AROUND	Déterminer l'impact des essais cliniques sur l'incidence et la mortalité des cancers	$\mathcal{A}$ = *{impact =* ***incidence, mortalité****} ({type de cancer}, {facteurs de risque =* ***essais cliniques****})*
Thème 3 – Prédire dans le temps et dans l'espace la mortalité des cancers dans les pays en voie de développement		
Prévenir pour mieux guérir	Prédire les taux de mortalité des cancers digestifs dans des pays dont les régimes alimentaires et les conditions environnementales sont différentes : la France et le Brésil	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestifs****}, {zone géographique =* ***France, Brésil****})*
Cancerinfl	Etude de l'évolution de la mortalité des cancers gynécologiques dans les pays d'Asie en voie de développement	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***gynécologiques****}, {zone géographique =* ***pays d'Asie en développement****})*
Osy3A	Prédire l'évolution de la mortalité des cancers digestifs dans les pays en voie de développement	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestif****}, {zone géographique =* ***pays en développement****})*
Oma	Etude de l'évolution de la mortalité du cancer de l'estomac dans les pays en voie de développement	$\mathcal{H}$ = *{impact =* ***mortalité****} ({type de cancer =* ***estomac****}, {zone géographique =* ***pays en développement****})*
Octopus	Prédire l'évolution de la mortalité des cancers colorectaux dans les pays en voie de développement	$\mathcal{A}$ = *{impact =* ***mortalité****} ({type de cancer =* ***colorectaux****}, {zone géographique =* ***pays en développement****})*

Note : les catégories spécifiées par les projets sont marquées en gras

Tableau 19. Présentation des projets du challenge 2 d'Epidemium.

1.3. BILAN DU DEUXIEME CHALLENGE4CANCER

Bien que la communauté Epidemium ait doublée entre le premier et le deuxième challenge, le nombre de participants actifs (hormis les équipes d'étudiants) a fortement diminué. La hausse de la barrière à l'entrée exigée par les organisateurs (revue de littérature et validité scientifique) a certainement contribué à la baisse des incitations des participants. En contrepartie, les projets soumis à la fin du challenge sont de bien meilleure qualité : dans l'ensemble, tous les outils développés par les participants sont beaucoup plus aboutis, et certains d'entre eux ont mené à une publication dans des revues scientifiques. De même, les hypothèses formulées se rapprochent beaucoup plus des exigences de validité d'hypothèses scientifiques en épidémiologie.

Notre analyse du challenge dans ce chapitre montrera que les organisateurs ont mieux répondu à certains besoins en terme de gestion durant le challenge, notamment en mettant à disposition des participants une équipe de scientifiques spécialistes en épidémiologie et capables d'aider et d'aiguiller la production. De plus, nous montrerons que la réussite du challenge est également dû à une capitalisation réalisée par les organisateurs entre le challenge 1 et le challenge 2 sur les pistes intéressantes à explorer à partir des données disponibles. En effet, la seule présence des bases de données ne permet pas aux organisateurs d'anticiper quelles sont les bonnes explorations à mener et donc de formuler des thématiques pertinentes. Au contraire, le processus d'exploration pour la tâche couplée inventive demande une gestion par la mise en place de challenges successifs. C'est un processus de **capitalisation séquentielle** : d'abord les participants explorent les bases de données durant les challenges afin de reformuler et de préciser quelles sont les bonnes hypothèses à explorer. Dans un deuxième temps, les organisateurs capitalisent sur la production globale entre les challenges afin d'orienter les explorations futures.

La présentation de ce deuxième challenge reprendra celle que nous avons réalisée pour le premier. Nous présenterons d'abord le processus d'exploration mené par les équipes. Nous nous intéresserons particulièrement à un nouveau processus d'exploration mené par l'équipe CAT. Nous analyserons ensuite la production globale du challenge au travers des métriques que nous avons présenté dans le chapitre 8. Enfin, nous proposerons une analyse critique de la capitalisation durant le challenge.

2. *EXPLORATION DES ESPACES ET EVALUATION DE LA PRODUCTION*

2.1. LE PROJET CAT COMME EXTENSION DE L'ESPACE DES HYPOTHESES

De manière générale, les projets orientés vers la formulation d'hypothèses scientifiques basées sur les données ont suivi le même processus d'exploration que celui présenté lors du premier challenge. Après avoir formulé leur hypothèse de départ à partir des données disponibles, les équipes ont confronté l'hypothèse aux données disponibles et ont procédé à une reformulation cohérente avec les données et les analyses fournies. Si les projets ont suivi peu ou prou le même processus que durant le premier challenge, le projet CAT nécessite une analyse distincte de par sa spécificité. En effet, contrairement aux méthodes déployées par les autres participants, le projet est initié non pas à partir des bases de données disponibles mais par une interrogation que les porteurs de projet ont sur le cancer et la notion de survivance. Les premiers travaux ont consisté à explorer à la fois les bases de données disponibles et la littérature pour situer le projet. Le projet CAT a notamment été soutenu et accompagné par la communauté de scientifiques mises à disposition par les organisateurs d'Epidemium et les laboratoires Roche. Suite à leur première exploration des données, un premier constat apparaît : la catégorie de publication "survie et modèle social" n'existe pas dans les bases étudiées et le sujet est principalement traité dans la littérature sous l'angle des cancers professionnels et les questions du travail sont abordées dans le cadre de la psycho-oncologie sous l'angle des "traces psychique" ou "la fabrique psychique du cancer".

En conséquence, les bases de données récoltées ne sont pas suffisantes pour rendre compte du phénomène. De plus, le projet CAT réalise rapidement que les quelques bases de données qui pourraient être utiles sont de mauvaise qualité. Grâce aux meetups, l'équipe projet se met en relation avec le projet Locapred qu'il convainc de nettoyer les bases de données OIT (Organisation Internationale du Travail). Cette organisation fournit un ensemble de données relatives au travail en fonction des pays. Cette collaboration permet d'aligner les objectifs des deux projets : d'un côté, le projet CAT va obtenir des bases de données fiables pour son hypothèse, tandis que le projet Locapred va utiliser les bases de données OIT comme matière première pour consolider son algorithme. Ces bases de données ont permis au projet CAT de constituer 5 familles de métriques pour décrire le modèle social en lien avec les incidences des cancers (tous les types de cancers) :

- Indicateurs du collectif public (conventions collectives, pourcentage de population avec ou sans couverture sociale, accessibilité au soin et aux soignants, dépenses de santé)
- Dynamisme social (niveau de productivité, de formation et de revenus, taux de chômage et de pauvreté)
- Production dominante/type d'activité (prédominance agricole / industrielle /tertiaire, protection sociale, médecine du travail)

- Conditions de travail (horaires, travail des enfants, économie déclarée ou non,…)
- Indicateurs socio-démographiques (disparité hommes/femmes, présence des seniors au travail,…)

Ainsi, au lieu d'adapter leur hypothèse aux bases de données existantes, le projet CAT a choisi la stratégie opposée. L'équipe de participants a cherché à définir les différents éléments de langage de son hypothèse ({travail}, {survivance}) au travers de variables déjà existantes dans les bases de données. On comprend qu'il y a projection de la question de recherche sur les données déjà existantes (**figure 56**). Chaque terme utilisé dans son hypothèse est retranscrit au travers d'un ensemble de termes déjà existants dans les bases de données. La majeure partie du travail consiste à retranscrire de la manière la plus fidèle possible les concepts développés au début du projet.

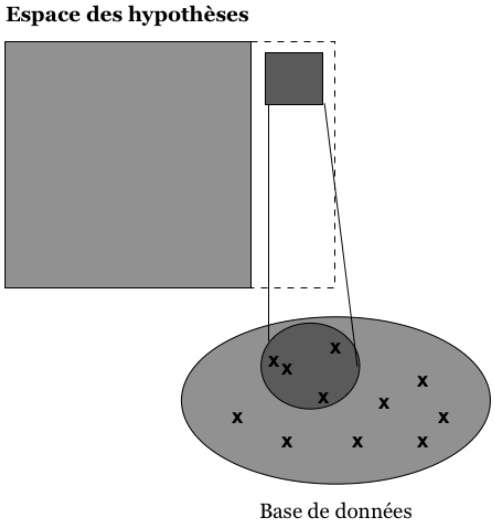


Figure 56. Projection de l'hypothèse de recherche sur les bases de données existantes.

Suite à ces premières analyses, le projet CAT est entré en relation avec l'institut Curie et l'institut Gustave Roussy pour définir des stratégies possibles afin de continuer le projet. En effet, ces instituts ont considéré comme intéressants l'approche de CAT, et étaient prêts à aller chercher plus loin.

Le processus d'exploration suivi par le projet Cancer Au Travail diffère des processus suivis par les autres équipes :

- *Formulation d'une hypothèse de travail indépendamment des données* : l'hypothèse initiale est formulée à partir des connaissances existantes en épidémiologie et sur les observations réalisées. En ce sens , la formulation correspond à une approche knowledge-driven

- *Projection de l'hypothèse sur les bases de données existantes* : L'équipe sélectionne des catégories issues des bases de données susceptibles de correspondre aux catégories de l'hypothèse de travail. Il y a projection de l'hypothèse sur les bases de données
- *Nettoyage, traitement et collecte de données* : Une fois les hypothèses choisies et les données déterminées, les participants nettoient les bases de données
- *Exploration des bases de données* : L'équipe développe un algorithme pour analyser les données et vérifier la valeur de vérité de l'hypothèse.
- *Processus d'optimisation* : une fois que l'hypothèse découverte est valide d'un point de vue de la communauté scientifique, les équipes projets cherchent à optimiser le modèle algorithmique utilisé pour analyser les bases de données. Aucune équipe projet n'a abouti à cette étape durant le premier challenge, probablement par manque de temps.

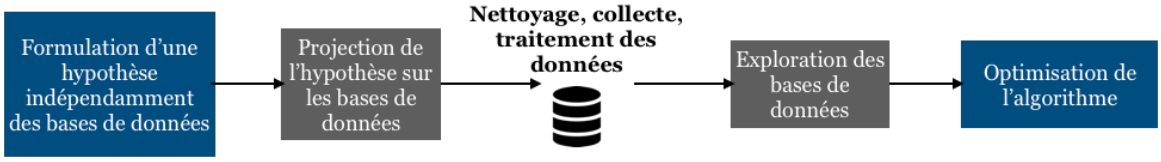


Figure 57. Processus d'exploration durant le premier challenge

2.2. EVALUATION FINALE DES PROJETS PAR LES COMITES

Les projets ont été évalués en deux temps par le comité scientifique : les organisateurs ont séparé l'évaluation des projets de la thématique 3 qui a été conçue spécifiquement pour des projets étudiants. De plus, les projets des thématiques 1 et 2 se sont déroulés entre Juin 2017 et Janvier 2018 alors que les projets de la thématique 3 se sont déroulés entre Novembre 2017 et Mars 2018. A l'aide de critères similaires à ceux utilisés dans le challenge 1, trois projets lauréats ont été choisis pour les thématiques 1 et 2 (dans l'ordre décroissant) : Locapred, IDEA et Cancer Diet. Bien que leurs approches soient originales d'un point de vue scientifique, les projets CAT et AROUND n'ont pas réussi à fournir à la fin du challenge des rendus suffisamment avancés par rapport aux autres projets. Cependant, ils ont été récompensés par leur côté éthique et leur originalité dans l'approche envisagée. Contrairement au premier challenge, le classement s'est avéré plus sélectif et les organisateurs ont souhaité mettre en avant les projets les plus aboutis.

Concernant la thématique 3, les projets ont été évalués indépendamment dans le cadre scolaire, et ont également reçu une évaluation par le comité scientifique. Les trois projets lauréats sont dans l'ordre décroissant : Osy3A, Octopus, Oma. Mis à part le projet vainqueur Osy3A, le résultat est en demi-teinte selon le comité. En effet, les étudiants semblaient ne pas avoir pris conscience de l'importance que le projet soit cohérent avec le cadre de recherche habituel dans la littérature en

épidémiologie. La plupart des projets ont proposé des modèles prédictifs agrégés, réduisant l'intérêt scientifique et en terme de politique publique pour l'exploitation des résultats.

2.3. ANALYSE DE LA PRODUCTION ET EXTENSION DE LA FONCTION DE VALEUR

Nous avons de notre côté évalué la production au travers des mêmes métriques que celles utilisées lors du challenge 1.

2.3.1. Formulation et vérification des hypothèses

Malgré un plus faible nombre de projets que dans le premier challenge, le nombre d'hypothèses générées durant le challenge est important, avec 12 hypothèses dont 8 axes de travail et 4 hypothèses canoniques. Ce résultat vient principalement du grand nombre d'hypothèses formulées par les étudiants via la problématique du cancer dans les pays en voie de développement. Pour autant ces hypothèses sont essentiellement des axes de travail qui étudient plusieurs cancers en même temps, et il n'a pas été possible d'extraire d'hypothèses scientifiquement intéressantes de leurs travaux. Ce problème avait déjà été identifié durant le premier challenge comme l'illusion des données. En effet, les communications entre étudiants et communauté scientifique ont été assez rare, et les scientifiques n'ont pas pu orienter les travaux afin d'éviter ce type d'exploration peu fécond.

Alors qu'aucune hypothèse formulée par les étudiants n'a été transformée en hypothèse canonique vérifiée, leurs travaux ont permis en revanche de mieux comprendre le fonctionnement des algorithmes de machine learning dans la recherche de corrélation entre facteurs de risque et incidence ou mortalité d'une pathologie. La plupart des projets étudiants ont permis d'extraire au travers de leurs modèles prédictifs une liste des facteurs de risque les plus discriminants pour une pathologie en fonction de son « poids » dans le modèle algorithmique. A partir de ces listes, les scientifiques pourraient se concentrer sur des facteurs de risque ou des facteurs protecteurs qui semblent jouer un rôle fondamental dans le développement de la pathologie. A titre d'exemple le projet Locapred a testé son modèle prédictif sur l'incidence du cancer du colon par genre et groupe d'âge. A partir de leur analyse, ils ont pu extraire les quatre facteurs les plus importants pour estimer l'incidence du cancer du colon : le taux de pollution de l'air à 2.5PPM, la consommation de cigarette, la consommation domestique de fruits, le taux de chômage. Sans apporter de résultat scientifique probant, cette analyse permet de formuler des hypothèses canoniques sur des facteurs de risques et protecteurs soupçonnés pour le cancer du colon. De plus, leur analyse met en avant la valeur potentielle des bases de données utilisées pour tester l'hypothèse.

Le projet CAT n'a pas pu fournir d'hypothèses canoniques à la fin du challenge. Le projet est en effet dans une phase très amont et leurs travaux ont permis pour l'instant d'identifier quelques

variables potentiellement utiles pour étudier quantitativement leur hypothèse. Enfin, les projets Cancer Diet et AROUND ont réalisé des explorations intéressantes mais ne possédaient pas les compétences suffisantes en terme d'analyse de données pour obtenir des résultats convaincants d'un point de vue scientifique. Il est à noter cependant que l'exploration du projet AROUND a permis de formuler une hypothèse canonique originale et potentiellement viable d'un point de vue scientifique.

Projets	**Hypothèse**	**Analyse par le langage**	**Bases de données**
Cancer Au Travail	Déterminer le niveau de survivance du cancer vis-à-vis des facteurs de risques sociétaux (type de sécurité sociale, travail,…)	$\mathcal{A}1$ = *{impact =* ***survivance****} ({type de cancer}, {facteurs de risque =* ***sociétaux****}*	Bases de données ILO
Cancer Diet	Développer un algorithme pour prédire le taux de mortalité du cancer en fonction du régime alimentaire	$\mathcal{A}2$ = *{impact =* ***mortalité****} ({type de cancer}, {facteurs de risque =* ***régime alimentaire****})*	Données FAO
AROUND	Déterminer l'impact des essais cliniques sur l'incidence et la mortalité des cancers	$\mathcal{A}3$ = *{impact =* ***incidence, mortalité****} ({type de cancer}, {facteurs de risque =* ***essais cliniques****})*	Essais cliniques (clinicaltrials.org), données INSEE (mortalité cancer en France)
AROUND	Déterminer l'impact des essais cliniques sur la mortalité des cancers des poumons, du pancréas, du sein et colorectal dans 5 régions françaises	$\mathcal{A}4$ = *{impact =* ***mortalité****} ({type de cancer =* ***poumon, pancréas, sein colorectal****}, {facteurs de risque =* ***essais cliniques****}, {zone géographique =* ***5 régions françaises****})*	Essais cliniques (clinicaltrials.org), données INSEE (mortalité cancer en France)
Prévenir pour mieux guérir	Prédire les taux de mortalité des cancers digestifs dans des pays dont les régimes alimentaires et les conditions environnementales sont différentes : la France et le Brésil	$\mathcal{A}5$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestifs****}, ({facteur de risque =* ***régime alimentaire****}, {zone géographique =* ***France, Brésil****})*	Données mortalité, FAO, World Bank data
Cancerinfl	Etude de l'évolution de la mortalité des cancers gynécologiques dans les pays d'Asie en voie de développement	$\mathcal{A}6$ = *{impact =* ***mortalité****} ({type de cancer =* ***gynécologiques****}, {zone géographique =* ***pays d'Asie en développement****})*	World Bank data, FAO, ILOstat
Osy3A	Prédire l'évolution de la mortalité des cancers digestifs dans les pays en voie de développement	$\mathcal{A}7$ = *{impact =* ***mortalité****} ({type de cancer =* ***digestif****}, {zone géographique =* ***pays en développement****})*	WHO, Wolrd Bank, FAO
Octopus	Prédire l'évolution de la mortalité des cancers colorectaux dans les pays en voie de développement	$\mathcal{A}8$ = *{impact =* ***mortalité****} ({type de cancer =* ***colorectaux****}, {zone géographique =* ***pays en développement****})*	WHO, Wolrd Bank, FAO
Locapred	Les particules 2.5PPM dans l'air sont un facteur de risque de l'incidence dans le cadre des cancers du colon	$\mathcal{H}1$ = *{impact =* ***incidence****} ({type de cancer =* ***colon****}, {FdR =* ***polluation air 2.5PPM****}, {profil =* ***féminin, âge 20-29 ans****})*	Données mortalité, FAO, World Bank data

Locapred	La consommation de fruit est un facteur de présention de l'incidence dans le cadre des cancers du colon	$\mathcal{H}1$ = *{impact =* ***incidence****} ({type de cancer =* ***colon****}, {FdR =* ***consommation de fruits****}, {profil =* ***féminin, âge 20-29 ans****})*	Données mortalité, FAO, World Bank data
Oma	Etude de l'évolution de la mortalité du cancer de l'estomac dans les pays en voie de développement	$\mathcal{H}3$ = *{impact =* ***mortalité****} ({type de cancer =* ***estomac****}, {zone géographique =* ***pays en développement****})*	WHO, Wolrd Bank, FAO

Tableau 20. Récapitulatif des hypothèses formulées durant le challenge 2

2.3.2. Elaboration d'outils pour l'exploration dans les espaces

Contrairement au premier challenge, les projets d'outils sont plus aboutis. Les deux équipes (IDEA et Locapred) ayant choisi de construire des outils pour l'exploration des données sont partiellement issues d'équipes ayant participé au premier challenge, et elles avaient anticipé certains problèmes et points durs à éviter. Le premier challenge leur a permis un premier contact avec les données qui les a mené à identifier les points faibles de l'exploration. Le projet IDEA a permis de fournir à la fin du challenge à un prototype fonctionnel de son logiciel de visualisation des données. Les scientifiques y voient un réel potentiel pour la recherche en épidémiologie c'est pourquoi l'équipe a été récompensée en étant en seconde place. Depuis la fin du challenge, l'équipe IDEA multiplie les tests de l'algorithme dans des cercles médicaux afin d'améliorer la qualité de l'outil. En fonction des résultats obtenus, l'équipe IDEA a pour objectif de soumettre une publication scientifique sur leur méthode originale.

Le projet Locapred, grand gagnant du challenge 2, a quant à lui développé un outil servant à nettoyer les bases de données ouvertes contenant des éléments manquants ou aberrants. En effet, une des conclusions majeures de la fin du premier challenge était la difficulté d'explorer les bases de données étant donné leur qualité souvent médiocre. Cet outil est utilisable clé en main pour des personnes non expertes, mais son code est également librement disponible pour des spécialistes de l'analyse de données afin de modifier tout ou partie des paramètres choisis. Les premières bases de données constituées par le logiciel ont intéressé une doctorante qui travaille en Angleterre sur des méthodes de nettoyage des bases de données. Leur collaboration a permis de publier un papier dans une conférence scientifique sur le machine learning (Chelly Dagdia et al., 2018). De plus, les membres du projet continuent de travailler sur le logiciel dont la partie sur la partie données manquantes va faire l'objet de deux publications : une dans le domaine de l'épidémiologie sur les questions des données ouvertes, et une technique sur la question des méthodes de remplissage des données.

2.3.3. Extension de la fonction de valeur

Le challenge 2 a enregistré une baisse importante de la participation (en dehors des projets étudiants) par rapport au challenge 1. Cependant, les projets menés ont été beaucoup plus aboutis et ont mené pour certains à des publications scientifiques à la fois dans le domaine de l'épidémiologie mais également dans le domaine du machine learning. De plus, les différents instituts partenaires du programme Epidemium ont engagé des collaborations avec au moins trois des projets finaux (CAT, Locapred, IDEA). Ainsi, contrairement au premier challenge, certaines équipes continuent leur projet hors du cadre d'Epidemium. Si les résultats sont assez positifs pour les projets des thèmes 1 et 2, le résultat est plus en demi-teinte pour les projets étudiants. En effet, les organisateurs n'ont pas cherché à tirer suffisamment parti de la présence d'un nombre important de participants pour piloter l'exploration autour des questions du cancer dans les pays en voie de développement. Ainsi, les projets des étudiants n'ont pas permis d'apporter un nouveau regard suffisamment pertinent pour ce sujet de recherche.

Certains projets ont également ouverts des voies intéressantes dans l'exploration de l'espace des hypothèses. D'abord le projet CAT a intégré un nouveau concept dans l'espace des hypothèses, créant une extension de l'espace et la possibilité d'interroger les bases de données sous un nouvel angle. Ensuite, le projet AROUND a cherché à explorer si la mise en place d'essais cliniques dans les hôpitaux avaient un impact sur la mortalité du cancer en France. La combinaison de ces deux termes {essais cliniques} et {mortalité} du cancer est une approche originale selon les scientifiques et suffisamment intéressante pour être traitée de façon isolée afin de proposer des indicateurs dans le cadre de la méta-épidémiologie et des études réalisées. Enfin, le projet *Locapred* a permis de montrer que les bases de données ouvertes utilisées permettaient d'interroger des facteurs de risques connus ou suspectés par la littérature et donc de pousser à investiguer dans ces directions.

Il est intéressant de voir que la valeur scientifique du projet *Locapred* est évaluée comme haute. Pourtant, d'un point de vue de l'épidémiologie du cancer, celui-ci n'offre pas de résultats pertinents. En fait, sa valeur scientifique a été montrée dans le domaine du machine learning et des techniques de data mining. Il y a extension de la fonction de valeur initialement construite : nous reviendrons sur ce point dans la dernière section.

	Science	Compatibilité données	Politique publique	Interaction acteurs potentiels	Originalité	Total
H1	xx	x	x	o	o	**27%**
H2	xx	x	x	o	o	**27%**
H3	o	o	o	o	o	**0%**
A1	xx	o	xx	x	xx	**47%**
A2	x	x	x	o	o	**20%**
A3	x	x	x	o	x	**27%**
A4	xxx	x	x	xx	o	**47%**
A5	o	o	o	o	o	**0%**
A6	o	o	o	o	o	**0%**
A7	o	o	o	o	o	**0%**
A8	o	o	o	o	o	**0%**
Outil de data visualisation (*IDEA*)	o	xxx	-	xx	xx	**47%**
Outil de nettoyage des données (*Locapred*)	xxx	xx	x	x	x	**53%**

Tableau 21. Analyse de la valeur par projet du challenge 2 Epidemium.

3. *Effet de la capitalisation sequentielle : extension de l'espace des hypotheses et de la fonction de valeur*

3.1. La reussite du challenge 2 portee en partie par la capitalisation des participants

Le principal changement en terme de gestion opéré entre le challenge 1 et le challenge 2 est un renforcement des contraintes en terme de résultats : en plus des thématiques suggérées, les organisateurs ont imposé aux participants de situer leurs activités vis-à-vis de la littérature scientifique. En contrepartie, ils ont mis à disposition des équipes un ensemble de spécialistes du domaine constitués d'épidémiologistes, d'oncologues ou d'acteurs de la santé publique. Si la valeur des résultats produits est dans certains projets meilleurs que dans le challenge 1, il serait précipité de dire que c'est uniquement grâce à l'action des organisateurs et des financeurs. En effet, plusieurs éléments nous portent à croire que ce n'est pas le cas.

Premièrement, nous avons vu que les rendus des étudiants ont été très décevants. En effet, les résultats obtenus par les projets étudiants sont souvent basés sur plusieurs cancers en même

temps, et ne permettent pas d'obtenir le niveau de validité exigé par la discipline. Bien que ces derniers avaient accès à un ensemble de spécialiste du domaine pour s'interroger sur leur approche, peu d'entre eux en ont effectivement tiré parti, et leur travail a été plutôt le résultat d'un vase clos. Or, les équipes étaient constituées uniquement d'étudiants spécialistes en analyse de données et n'avaient pas de compétences en épidémiologie ou de façon plus large dans les questions de santé. Au final, les organisateurs ont appris peu de choses sur la thématique 3 et les projets étudiants ont plutôt servi de vecteur de communication afin de diffuser l'existence et l'intérêt du programme Epidemium dans les cercles universitaires. Pourtant, le problème de la validité des hypothèses formulées avait déjà été identifié lors du premier challenge. Nous avions mis en avant que les équipes n'avaient pas suffisamment pris en compte les conditions de validité relative à la discipline. L'accessibilité à des spécialistes du domaine ne semble pas suffisante pour éliminer ce problème, et les organisateurs devraient annoncer explicitement ces conditions de validité en même temps que les thématiques.

Deuxièmement, les projets avec les résultats les plus aboutis (*IDEA* et *Locapred*) et qui ont été récompensé par Epidemium proviennent d'équipes qui étaient déjà présentes au premier challenge. Lors de nos discussions avec les membres de ces équipes projet, nous avons compris que leur expérience lors du premier challenge avait été capitale pour eux dans le déroulement de leur deuxième intervention. Les choix qu'ils ont réalisés, autant dans la formulation de leur problématique que dans leur rapport aux données, ont été la conséquence de l'apprentissage qu'ils avaient eux-mêmes fait de leur expérience passée. Comme nous l'avions précisé pour le premier challenge, la capitalisation sauvage entre les deux challenges n'a pas été suffisante pour rendre compte des problématiques gestionnaires mais également de la capitalisation sur la production. Nous avons montré qu'une partie de cette capitalisation avait été portée par les équipes des participants. Ainsi, le résultat du challenge aurait été tout autre si les équipes n'avaient pas reconduit leur participation, incluant le bagage d'un apprentissage tacite.

Au final, si le deuxième challenge a eu de meilleurs résultats que le premier challenge, c'est principalement grâce aux équipes qui ont elles-mêmes capitalisé sur ce qu'elles avaient appris durant le premier challenge. Or, il est nécessaire que cet apprentissage soit géré par les organisateurs.

3.2. EXTENSION DE L'ESPACE DES HYPOTHESES

Le deuxième challenge a mis en évidence l'émergence de projets comme CAT qui construisent des hypothèses indépendamment des bases de données. Au lieu de chercher à produire une hypothèse qui colle le mieux aux données disponibles, les membres de l'équipe CAT ont réalisé le processus inverse : ils ont considéré que les données existantes étaient potentiellement suffisantes pour traiter un problème que les participants avaient identifié au préalable. Ce phénomène est caractéristique d'un fort imaginaire sur la capacité des données massives. Durant notre analyse de cas, nous avons pu constater à de multiples reprises que l'existence même de bases de données

massives dont le contenu est incertain est souvent associée à une croyance sur ce qu'il est possible de créer à partir de ces données. Cela peut mener les citoyens de la science à formuler une hypothèse : le nombre de variables à l'intérieur des bases de données massives est tellement grand qu'il est statistiquement possible que une hypothèse formulée de manière *ad hoc* puisse être vérifiée par ces données. Les participants sont alors sujets à une « appétence » qui les poussent à exploiter les données comme une ressource pour étudier un sujet.

Si cette appétence ne peut être maîtrisée dans un processus totalement ouvert, elle ne doit pas être négligée par les organisateurs car elle crée une opportunité de faire émerger des questions intéressantes que l'on peut tester à partir des bases de données existantes. Nous avons vu que, contrairement à l'approche classique qui consiste à reformuler les hypothèses de départ pour qu'elles correspondent aux bases de données, la stratégie ici consiste à trouver des variables suffisantes pour expliquer l'hypothèse préalablement formulée. Dans le cas du projet CAT, la notion de « survivance » n'existe pas dans les bases de données existantes, ainsi que dans la littérature en épidémiologie du cancer. En fait, le terme permet de faire une extension de l'espace des hypothèses en intégrant une nouvelle relation entre les familles de concept existantes.

3.3. EXTENSION DE LA FONCTION DE VALEUR

Enfin, nous avons montré que si la fonction de valeur est nécessaire pour piloter l'exploration des espaces dans les tâches déléguées à la foule, elle ne peut pas être considérée comme étant inconditionnellement figée. En effet, la fonction de valeur créée durant le premier challenge n'a pas été suffisante pour rendre compte de ce qui a pu être produit durant le challenge 2. Le projet *Locapred* a par exemple débouché sur la publication de papiers scientifiques dans le domaine du machine learning, bien éloigné du domaine de l'épidémiologie du cancer et donc de la définition que l'on avait proposé de la « valeur » scientifique. Tout se passe comme si l'exploration réalisée par l'équipe *Locapred* avait étendu les critères de valeurs qui avaient été préalablement établi par les organisateurs. Nous pouvons représenter la fonction de valeur comme un ensemble de critères $\{c_1, c_2, \dots, c_n\}$ avec n fixé (5 dans notre cas durant le challenge 1). Dans la **figure 58** que nous avons utilisée pour notre étude, les valeurs sont représentées dans un espace à deux dimensions, où les valeurs fluctuent en fonction de la position dans l'espace explorée.

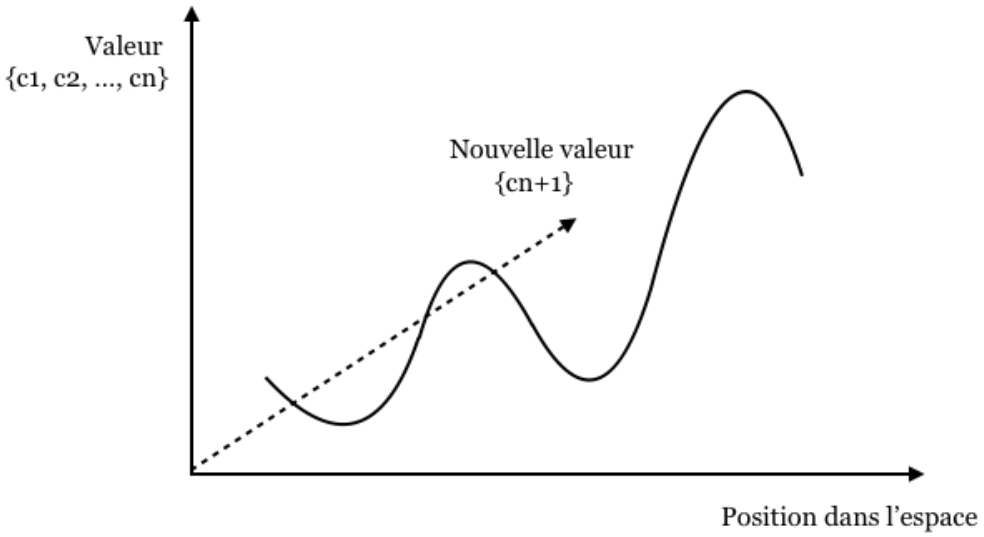


Figure 58. Nouvelle valeur $\{c_{n+1}\}$ intégrée dans les critères définis préalablement

Le projet *Locapred* permet une extension de cette fonction de valeur en intégrant la notion de « valeur scientifique dans le domaine du machine learning ». Cela représente une extension des critères de valeurs établis préalablement en ajoutant une dimension $\{c_{n+1}\}$. Sans cette extension, l'évaluation du projet serait perçue comme incompatible avec les objectifs du programme et donc écartée du processus de capitalisation séquentielle. Ainsi, en plus de gérer la capitalisation par la valeur des projets, il y a nécessité de gérer l'extension de cette valeur.

CHAPITRE 10 – IMPLICATIONS MANAGERIALES : ORGANISATION ET APPARITION DE LA FIGURE DE « GESTIONNAIRE DES FOULES INVENTIVES »

RESUME DU CHAPITRE 10

Dans ce chapitre, nous faisons apparaître un modèle général de gestion des projets de science citoyenne. Ce modèle comporte six étapes : 1) Définir la problématique, 2) animer la foule, 3) coordonner la production, 4) évaluer et réintégrer les résultats dans le processus scientifique, 5) piloter le transfert d'informations entre projets, 6) Systématiser l'apprentissage entre projets. La réalisation des étapes nécessite la présence de figures d'acteurs capables de gérer cette logique et de conduire à un pilotage de la productivité. Notre étude nous conduit à définir que l'efficacité des projets de science citoyenne au sein dans les organisations scientifiques pourrait être assurée avec l'aide d'une figure comme « le gestionnaire de foules inventives ».

Nous pouvons le rapprocher de la figure du manager de projets ou du manager de portefeuille de projets. Cet acteur n'est pas nécessairement expert dans la discipline scientifique, mais il doit être capable de définir la problématique soulevée par l'organisation et la transcrire sous la forme d'un projet de science citoyenne : il doit isoler la tâche du reste du processus, formuler la problématique, mais également définir et fournir les outils et les dispositifs nécessaires à mettre en place dans le cadre du projet, ainsi que de décider les formes de collaboration entre les participants (compétitif, collaboratif, hybride), et de maîtriser les pertes de production durant l'exécution de la tâche. Il est également responsable d'animer la communauté durant le projet et de gérer les systèmes de motivation. Ensuite, dans le cas où le projet s'étale sur plusieurs épisodes, il doit être capable de cartographier l'exploration des espaces réalisée par les participants basée sur une fonction de valeur, tout en étant capable d'identifier l'apparition de nouvelles dimensions de la valeur en fonction de la production. Enfin, il est garant de ce qui est produit par les participants et donc de réintégrer la ce qui est produit dans le processus scientifique.

En outre, cette figure managériale partage certaines des étapes avec le scientifique. En effet, la cartographie de la valeur suffit à piloter l'exploration et ne nécessite pas la connaissance théorique caractéristique du rôle du scientifique. En fait, le scientifique a essentiellement un rôle de garant final et d'évaluateur de l'intérêt scientifique avéré du projet. De la même manière que nous avions observé une remise en cause de la place du scientifique dans notre analyse historique, la fonction du scientifique au sein du processus de production de connaissance évolue : il est demandé au scientifique d'être capable d'identifier la valeur scientifique des hypothèses formulées par d'autres acteurs. C'est le rôle notamment des comités scientifiques que nous avons pu voir dans le programme Epidemium.

Enfin, nous suggérons qu'il y a plusieurs avantages à ce que les projets de science citoyenne soient gérés par des structures intermédiaires. Dans cette forme d'organisation, le gestionnaire des foules inventives n'est pas un employé de l'organisation scientifique mais un des acteurs de la structure intermédiaire. La relation entre scientifique et gestionnaire des foules inventives peut s'apparenter à une relation entre un client et un fournisseur : le scientifique attend des résultats de la part de la structure intermédiaire tandis que le gestionnaire des foules inventives assure la performance de la délégation à la foule.

1. STRUCTURE ORGANISATIONNELLE ET ROLE MANAGERIAL POUR LES PROJETS DE SCIENCE CITOYENNE

Comme indiqué dans la partie 1, cette thèse détermine les moyens de gestion à mettre en œuvre dans les organisations scientifiques éphémères, à savoir les projets de science citoyenne, pour piloter la performance des projets et s'assurer de minimiser les pertes durant le processus.

Les résultats de la thèse révèlent que la gestion des projets de science citoyenne nécessite de s'intéresser à un nouveau type de tâche non encore pris en compte dans les modèles de science citoyenne, la tâche couplée, pour étudier la génération des hypothèses basées sur les données. Ce modèle de tâche est basé sur l'exploration couplée de deux espaces en même temps. Nous construisons donc un modèle basée sur quatre types de tâches qui peuvent être extraites du processus scientifiques pour être déléguées à la foule : tâche élémentaire, recette, résolution de problèmes, tâche couplée. En suivant ce modèle, nous montrons que la gestion de la performance est relativement bien traitée dans les tâches de type élémentaire et recette, notamment en favorisant une capitalisation par agrégation de la production durant le projet et la mise en place de retours d'expériences entre les projets. Cependant, la résolution de problèmes et les tâches couplées inventives demande une exploration des espaces pour trouver des solutions pour laquelle la stratégie n'est pas claire. Nous montrons que la question de la capitalisation durant et entre la tâche est critique pour s'assurer de l'efficacité de la délégation des projets à une foule pour ce type de tâche. Afin d'étudier ces formes de capitalisation, nous introduisons deux notions, la capitalisation croisée et la capitalisation séquentielle, dont l'objectif est de maximiser la réutilisation de ce qui est produit. Par la suite, nous étudions les effets de la capitalisation croisée sur la performance de la tâche et son lien avec la diversité. Nous étudions également la notion de capitalisation séquentielle pour les tâches couplées inventives. Les résultats révèlent plusieurs choses : d'abord un modèle de performance lorsqu'il y a capitalisation croisée durant le processus. Nous montrons notamment l'effet de la diversité dans les solutions sur la performance des projets. Ensuite, nous montrons que les tâches couplées inventives ne peuvent que difficilement se résoudre en un projet mais nécessitent de mettre en place une succession de projet dont les transitions sont gérées par capitalisation séquentielle.

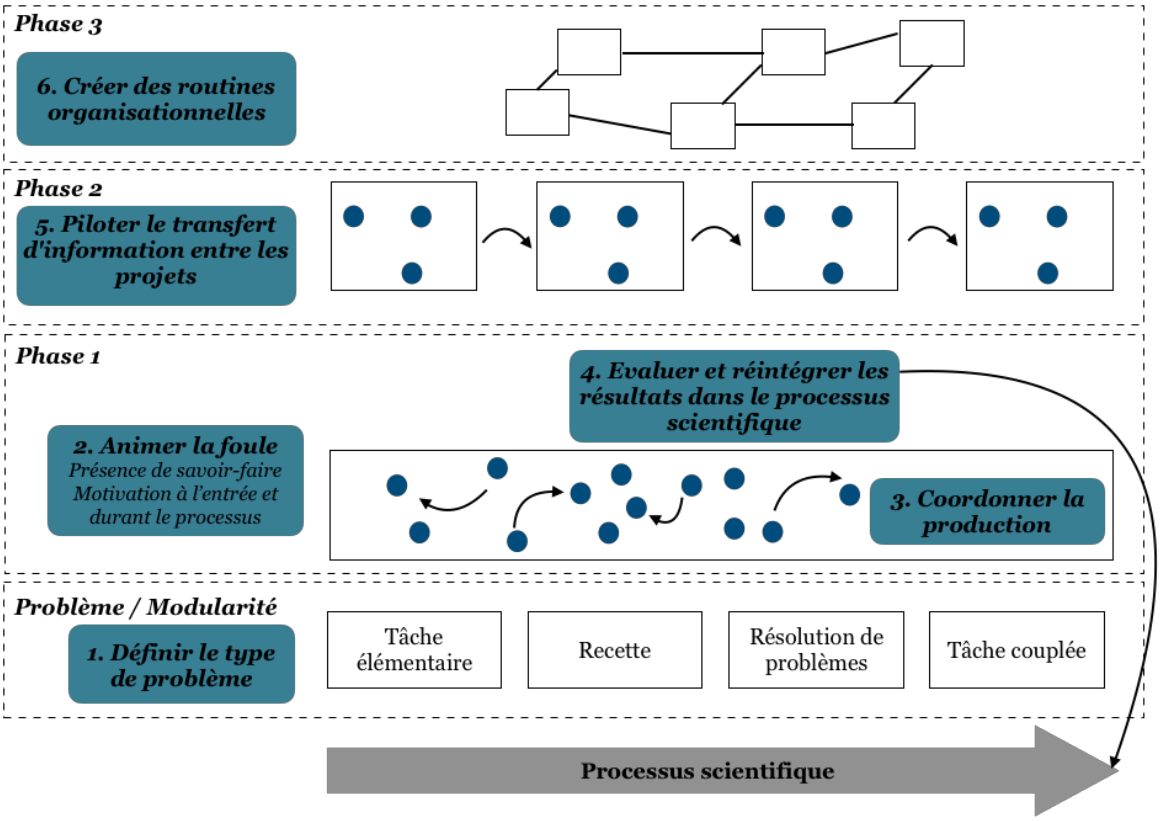


Figure 59. Modèle de performance des projets de science citoyenne.

A partir du modèle de gestion proposé par Afuah et Tucci (2012) pour les projets de crowdsourcing, de l'étude des project capabilities développées par Brady et Davies (2004) et de l'étude que nous avons mené dans la thèse, nous proposons un modèle de gestion pour les projets de science citoyenne (**figure 59**). Ce modèle comporte six étapes :

1. **Définir le type de problèmes**. Les organisateurs doivent identifier les tâches au sein du processus scientifique qui pourraient bénéficier des avantages d'une délégation à la foule. Deux cas peuvent se présenter : dans le premier, le problème est facile à délimiter et isoler du reste du processus, mais également à transmettre (peu d'informations tacites). Le problème sera d'autant plus facilement délégable à la foule s'il est modularisable (Afuah & Tucci, 2012). Dans le cas d'une tâche couplée inventivecependant, le problème ne réunit pas ces conditions. Les organisateurs doivent s'assurer que l'ensemble des informations nécessaires soient transmis et suffisamment claires : les bases de données nettoyées, outils d'analyse et de gestion,... Enfin les organisateurs estiment les coûts d'installation (plateforme web, communication,...) et la répartition des coûts de production.
2. **Animer la foule**. Les organisateurs doivent faire coïncider autant que possible les profils des participants avec le type de compétences recherchées. Par exemple, pour résoudre un problème d'analyse de données il sera préférable de communiquer dans des sphères sociales incluant des spécialistes de l'analyse de données. En revanche, la

définition des compétences nécessaires se confrontent également aux limites de compétences de l'organisation, pour laquelle la solution est potentiellement liée à des compétences distantes des siennes (Afuah et Tucci, 2012). Par ailleurs, les organisateurs doivent s'assurer d'une bonne communication pour inciter les participants, mais également doivent avoir des capacités d'animation durant le projet pour favoriser le temps passé par les participants sur la résolution, et donc l'effort.

3. **Coordonner la production**. Le mode de coordination et de collaboration entre les participants est défini suivant le type de tâche déléguée et les conditions contractuelles qui sont définies par les organisateurs. La fiabilité de la production est gérée par une redondance des tâches, c'est-à-dire une répétition de la même tâche pour augmenter statistiquement la chance d'avoir une bonne solution. Dans le cas des tâches élémentaires ou recette, les productions sont ensuite agrégées pour choisir par des moyens statistiques (moyenne, valeur la plus élévée/basse, médiane,…) afin d'obtenir le résultat final (capitalisation par agrégation). Dans les tâches nécessitant une exploration d'espaces, il est préférable dans la mesure du possible (conditions contractuelles) de favoriser la réutilisation de la production durant le processus (capitalisation croisée). Les organisateurs devront favoriser également la diversité des solutions proposées en proposant par exemple en amont une phase fermée (compétitive).
4. **Evaluer et réintégrer les résultats dans le processus scientifique**. Les organisateurs doivent vérifier si ce qui a été produit est conforme aux critères définis au départ avant de réintégrer la production dans le processus scientifique. Lorsque c'est possible, il est préférable de mettre en place des métriques d'évaluation *ex ante* qui permettent de contrôler facilement ce qui est produit et de choisir les productions. Dans certains cas en revanche (Galaxy Zoo par exemple), évaluer la production demanderait presque le même effort que la tâche en elle-même. Les organisateurs peuvent en ce cas mettre en place d'autres méthodes pour faciliter l'évaluation : déléguer l'évaluation à un grand nombre de personnes (Afuah & Tucci, 2012), mettre en place une redondance.
5. **Piloter le transfert d'information entre les projets**. Durant cette phase, on cherche à transférer ce qui a été appris durant un projet vers les projets suivants. Cette tâche est particulièrement critique dans les projets de tâche couplée inventiveoù la résolution de la tâche se fait par projets successifs. Les organisateurs doivent mettre en place un système de capitalisation séquentielle où la problématique de chaque projet évolue en fonction des productions des projets précédents. Ce système demande à mettre en place un système d'évaluation potentiellement extensible et réévalué en fonction de ce qui est produit.
6. **Systématiser l'apprentissage entre projets**. Une fois qu'un nombre suffisant de nouveaux types de projets ont été initiés par l'organisation, cette phase permet de développer un apprentissage entre les projets. Des efforts sont entrepris par les organisations pour systématiser l'apprentissage et la transmission des connaissances

accumulées vers l'ensemble de la division responsable des projets. Les organisations peuvent également avoir besoin de créer des cellules spécialisées pour prendre en charge un nombre croissant de projets. Cela permet de s'assurer que les connaissances acquises restent effectivement dans la mémoire de l'organisation (Brady & Davies, 2004).

Dans le **tableau 22,** nous reprenons quelques cas de projets de science citoyenne afin d'illustrer comment ces étapes sont gérées : un projet de tâche de type recette (Galaxy Zoo), trois projets de type résolution de problèmes (RAMP, Foldit, DREAM challenges), et un projet de tâche couplée inventive(Epidemium). Chaque projet ne traite pas nécessairement toutes les étapes. Par ailleurs, ce comparatif illustre bien comment les organisations coordonnent ou non le travail de la foule, mais également mettent en place des moyens organisationnels pour piloter le transfert d'informations entre projets.

La réalisation des étapes nécessite la présence de figures capables de gérer cette logique et de conduire à un pilotage de la productivité. Cependant quels sont ces acteurs ? Quelle est leur place dans l'organisation ? Ces acteurs doivent être capable de : 1) Définir la problématique, 2) animer la foule, 3) coordonner la production, 4) évaluer et réintégrer les résultats dans le processus scientifique, 5) piloter le transfert d'informations entre projets, 6) Systématiser l'apprentissage entre projets. **Est-ce que cet acteur existe ?** Peut-on le trouver au sein d'une organisation ?

Comme nous l'avions signalé dans les chapitres 2 et 4, la mise en place des projets de science citoyenne peut potentiellement impacter le rôle du scientifique dans ce processus. D'un côté, le scientifique n'est pas nécessairement le plus qualifié pour répondre à toutes les activités nécessaires dans le processus, notamment dans la gestion d'une foule et dans sa coordination. Par ailleurs, la tendance historique de la position du scientifique dans le processus de production de connaissance est de concentrer son action sur les tâches critiques pour lesquelles sa valeur ajoutée est la plus importante. D'un autre côté, le scientifique est le garant de la connaissance dans le cadre de sa discipline ou de sa spécialité. C'est lui qui définit les besoins nécessaires pour s'assurer que le processus scientifiques possède les moyens suffisants pour être mené à bien. C'est également grâce à son aval que l'on peut assurer que ce qui est produit durant le processus est bien validé scientifiquement.

Comment se répartissent les activités entre le scientifique et l'autre acteur gestionnaire ? Pour étudier cette question nous reprenons les cas expérimentaux, RAMP et Epidemium.

Tableau 22. Illustration des étapes du processus par des exemples de projets de science citoyenne.

Type de projets	RAMP	Epidemium	Foldit	Galaxy Zoo	DREAM challenge
Produit final	Code informatique (analyse de données)	Hypothèses et méthode d'analyse	Méthode repliement protéine, recette	Images annotées des galaxies	Code du meilleur score
1. Définir la tâche	Résolution de problèmes - problème scientifique, base de données, site web, métriques de performance - Calcul par serveurs	Tâche couplée - Base de données, logiciels, espace physique, suivi de la production	Résolution de problèmes - site gamifié, problèmes définis - calcul local	Recette - Site web, tutoriel, serveurs, images	Résolution de problème - problème scientifique, base de données, site web, métriques de performance
2. Animer la foule	Communauté constitué de data scientists ou d'étudiants en data science Projets courts (moins d'1 mois, souvent quelques jours)	Communication dans des endroits ciblés (instituts de recherche, écoles,…) Plusieurs outils de communication (meetups, chats, forums, wikis, réunions,…)	Plateforme gamifiée, forums de discussion	Plateforme gamifiée	Grosses récompenses, communication dans des revues prestigieuses, reconnaissance professionnelle
3. Coordonner la production	Capitalisation croisée directe (accès à la production durant le tournoi)	Capitalisation croisée favorisée (projets communs – e.g. EpidemiumDB)	Capitalisation croisée traçable (sur le site)	Capitalisation par agrégation	Système compétitif
4. Evaluer et réintégrer les résultats	Evaluation du score (calcul simple) Transfert du code aux scientifiques	Evaluation par métriques *ad hoc* Contacts avec les instituts intéressés (Curie, Gustave Roussy)	Evaluation directe du score, intégration dans le logiciel Rosetta	Concordance statistique (agrégation de 50 tâches identiques)	Evaluation directe du score,
5. Piloter le transfert d'informations entre les projets	Faible – pas de transfert des solutions entre les épisodes	Capitalisation séquentielle : extension continue de la fonction de valeur	Délégation de plusieurs problèmes similaires	Retour d'expériences	Amélioration continue au fur et à mesure des challenges
6. Systématiser l'apprentissage entre projets	Amélioration des puissance de calcul des algorithmes (serveur Amazon)	Pas de transfert visible	Modèle répliqué en biologie computationnelle	Plateforme Zooniverse	Routines organisationnelles (unique chef de projet, montant de la récompense,…)

2. LE ROLE DU GESTIONNAIRE DE FOULES INVENTIVES

Dans les projets Epidemium et RAMP, nous avons eu l'occasion de rencontrer des figures en charge d'une ou plusieurs des étapes que nous avons identifiées. Au travers de ces deux cas, nous cherchons à mettre en évidence comment ces acteurs se sont emparés de ces activités et la façon dont ils ont piloté la performance du processus.

2.1. RETOUR SUR LES CAS ETUDIES : PRESENTATION DES MANAGERS RENCONTRES

2.1.1. Cas 1 : le projet RAMP

Comme nous l'avions présenté dans le chapitre 3, le RAMP est une plateforme de data challenges à l'initiative du Center for Data Science de Paris Saclay. L'initiateur et le gestionnaire principal du projet RAMP est Balasz Kégl, pour lequel l'intérêt principal de la plateforme est d'avoir un support permettant de réunir à la fois des scientifiques en recherche de modèles algorithmiques et des data scientists et informaticiens capables de fournir des réponses adaptées. Chaque projet du RAMP a démarré avec un problème plus ou moins bien formulé autour de l'analyse de bases de données et provenant de scientifiques de diverses disciplines (biologie, physique,...). Pendant plusieurs mois, les scientifiques et les équipes du RAMP (Balasz Kégl ainsi que des spécialistes en machine learning) travaillent avec les scientifiques afin de proposer une formulation du problème qui puisse répondre aux contraintes d'un data challenge : problème nécessitant très peu de connaissance dans la discipline scientifique, objectif formulé sous la forme d'une ou plusieurs métriques de performance, bases de données nettoyées, séparation entre la base de données d'entraînement et la base test. Notons que les membres de l'équipe ne sont pas experts de la discipline, mais experts en machine learning. Par ailleurs, les équipes du RAMP élaborent également une baseline, c'est-à-dire un premier modèle d'algorithme non optimisé mais qui permet de proposer une première solution efficace. Une fois le problème défini et posé, l'équipe organisatrice définit la durée du data challenge, les éventuelles récompenses mais également si le challenge est ouvert, fermé ou hybride. Ils sollicitent ensuite des communautés de data scientists susceptibles d'être intéressés par le projet et lancent le challenge.

Durant le challenge, les organisateurs observent l'évolution des scores, et poussent les participants à soumettre le plus possible de solutions. Dans certains cas, le data challenge se passe dans des lieux physiques au travers lesquels les organisateurs ont la possibilité d'interagir facilement avec les participants. Ils peuvent également choisir de prolonger ou réduire la phase fermée dans les challenges hybrides. Une fois le challenge terminé, la meilleure solution soumise est ensuite proposée au scientifique afin de l'intégrer à son processus. A noter que les mêmes problèmes sont réalisés plusieurs fois, pour pouvoir faire des observations dans des contextes

différents mais également comme support pédagogique dans les écoles. Cependant, les organisateurs ne font pas évoluer le baseline entre les différents épisodes. En effet, les problèmes rencontrent un plateau de performance qui rend difficile l'amélioration cumulée notamment dans un contexte pédagogique. Il y a donc des pertes importantes entre les projets dans le principe où la réutilisation entre les projets permettrait d'améliorer le modèle global.

Si les organisateurs n'ont pas mis en place un système de transfert entre les projets, ils tirent cependant des enseignements généraux sur l'organisation des projets et systématisent des solutions. Par exemple, les organisateurs ont amélioré le système de calcul des solutions soumises par les participants en intégrant des serveurs externes (Amazon) pour répondre au besoin des participants d'avoir un retour rapide sur leurs solutions.

2.1.2. Cas 2 : le programme Epidemium

Le programme Epidemium a pour objectif de tirer au mieux parti de bases de données massives ouvertes autour de la question de l'épidémiologie du cancer. Ce programme est piloté par deux acteurs : Olivier de Fresnoye, spécialisé dans la communication scientifique et l'animation de communautés, et Mehdi Benchoufi, professeur assistant au centre d'épidémiologie clinique au sein de l'APHP. Les organisateurs se sont entourés de plusieurs structures et acteurs capables de répondre à des besoins spécifiques : financeurs, scientifiques, entreprises, entre autres. Ainsi, les organisateurs créent un comité scientifique et éthique constitué d'experts des disciplines (machine learning et épidémiologie) pour évaluer les projets. Ils rendent également disponibles aux organisateurs un ensemble d'outils techniques et de gestion de projets pour faciliter leur travail sur les données. Enfin, une équipe dédiée a été utilisée pour compiler, traiter et nettoyer les bases de données disponibles sur le sujet. Les bases de données ont ensuite permis de développer quatre thématiques de recherche, qui serviront de base pour l'exploration durant le challenge.

Une fois l'organisation mise en place, les organisateurs ont été amenés à diffuser le projet au sein de structures existantes et de communautés de pratiques. Leurs liens avec la structure La Paillasse, les laboratoires Roche ainsi que la communauté RAMP a permis de fédérer un certain nombre de participants potentiellement intéressés par le sujet. Par ailleurs, les organisateurs sont entrés en contact avec plusieurs instituts de recherche (Gustave Roussy, Curie, CLARA) qui ont permis d'élargir le nombre de participants potentiels. Enfin, les organisateurs ont mis en place un grand nombre de présentations dans différents cercles (étudiants, hackathons, environnement de startups) afin d'élargir et de maximiser le nombre d'entrées et de participants. Une fois les participants intégrés au projet, les organisateurs ont également mis en place des meetups hebdomadaires dans les locaux de la Paillasse pour favoriser les rencontres et maintenir l'effort et la communication entre les personnes. Ces échanges ont participé à une collaboration entre les participants et une réutilisation de la production (bases de données nettoyées, outils développés,...). Par ailleurs, des personnes étaient en charge d'une communication régulière sur

les réseaux sociaux. Enfin, une évaluation de mi-parcours a été faite pour les projets intéressés avec la participation du comité scientifique.

Une fois le premier challenge terminé, le comité scientifique en collaboration avec les organisateurs ont créé une grille de valeur *ad hoc* pour analyser et évaluer les projets. Pour autant, l'avancement des projets ne permettait pas d'en tirer en l'état actuel des résultats probants qui puisse avoir de la valeur pour une organisation existante ou bien auprès de la communauté scientifique. Les organisateurs ont donc lancé un deuxième projet, en partie basé sur les expériences du projet passé, mais dont une grande partie de la production a été délaissée. Au final, la plupart des apprentissages ont été transmis de manière tacite par la participation de volontaires qui étaient déjà présents lors du premier challenge.

2.1.3. Enseignements tirés des deux cas

L'utilisation des citoyens de la science comme ressource est une opportunité pour améliorer les capacités de production de connaissances scientifique. Cependant, malgré les nombreux avantages en terme de coût et de temps passé, les projets de science citoyenne posent néanmoins des problématiques en terme de gestion au sein du projet mais également dans une logique d'utilisation en continu. Dans les deux cas que nous avons étudié dans la thèse, nous avons vu apparaître un nouveau type d'acteur, le « **gestionnaire de foules inventives** », qui partage avec les scientifiques les six étapes que nous avons montré préalablement : ainsi ils participent à la définition du problème, animent et coordonnent la communauté, aident à l'évaluation et à la réintégrations, mais également sont les acteurs indispensables pour transférer les informations entre les projets. Nous retrouvons en partie ce rôle de gestionnaire de foules inventives incarné par Olivier de Fresnoye et Balasz Kégl dans les deux cas que nous avons étudié.

2.2. LE ROLE DU « GESTIONNAIRE DES FOULES INVENTIVES »

Le gestionnaire de foules inventives peut être comparé au manager de projets ou de portefeuille de projets. A partir d'une problématique définie, il est en charge de mener à bien l'exécution du projet afin d'atteindre l'objectif annoncé. La différence principale entre le manager de projets et le gestionnaire de foules inventives est que la réalisation de la tâche se fait par l'introduction d'une foule dans le processus. Le gestionnaire doit donc être capable de piloter à la fois l'animation de la foule (aller chercher des participants potentiels et s'assurer qu'ils restent dans le processus), mais également doit s'assurer de la coordination du travail entre les participants. Par ailleurs, le gestionnaire de foules inventives est également responsable de capitaliser sur ce qui a été produit durant les projets et donc de transférer les informations de projet à projet afin de limiter les pertes. Nous avons vu que cette étape est particulièrement importante pour les tâches couplées inventives qui nécessitent de mettre en place des projets successifs et donc de gérer la transmissions entre les projets.

Dans les projets sur des tâches élémentaires, recette ou résolution de problème, le gestionnaire de foule n'est pas responsable de l'élaboration de la fonction de valeur associée à la production. Celle-ci est définie préalablement en accord avec le demandeur. Cependant, le gestionnaire est capable d'apprécier si la problématique est cohérente pour une délégation à une foule. De plus, il doit pouvoir juger si le problème est bien indépendant du reste du processus scientifique et ne demande pas de connaissances autres. Dans les tâches couplées inventives cependant, la fonction de valeur est définie *ex ante* avec le demandeur, mas elle est susceptible d'évoluer en fonction de ce qu'il est possible de découvrir lors du processus. Ainsi, le gestionnaire des foules inventives doit être capable de faire évoluer la fonction de valeur pour rendre compte au mieux de ce qui est produit. Sans être spécialiste du sujet ni avoir les compétences techniques, il doit être capable de cartographier tout ce qui a été produit durant le challenge en fonction de sa valeur afin de pouvoir prendre la décision des directions à prendre sur les challenges successifs.

Enfin, même si cette fonction est peu vue dans les exemples que nous avons présenté, le gestionnaire de foules inventives doit être capable de systématiser l'apprentissage entre projets. Une fois que le nombre de projets réalisé est suffisamment important, il doit pouvoir tirer des conclusions en terme d'organisation, d'amélioration de la productivité, afin de construire des règles de fonctionnement similaires à plusieurs projets en fonction de leurs caractéristiques.

L'apparition du gestionnaire de foules inventives dans notre étude s'aligne avec des travaux précédents autour des notions de « collège de l'inconnu » et « architecte de l'inconnu », qui caractérisent l'existence d'une personnalité managériale n'agissant pas pour son propre bénéfice mais désirant affecter les capacités de conceptions innovantes des autres acteurs (Agogué et al., 2017; Le Masson & Weil, 2014). Ils suggèrent que dans certaines situations dans lesquelles les technologies, les marchés et les acteurs impliqués sont inconnus, de nouveaux principes de gestion spécifiques pour l'intermédiation sont nécessaires. L'architecte de l'inconnu est un acteur qui exploite les idées et imaginaires d'autres acteurs d'un secteur pour améliorer les capacités de conception actuelles. Cette recherche montre comment un cluster ou une association, par exemple, peut agir au-delà d'un simple rôle de tierce partie pour approuver des activités exploratoires. En règle générale, un architecte de l'inconnu peut dévoiler des pistes d'innovation inexplorées que d'autres acteurs peuvent emprunter. Il peut également apporter des idées de l'extérieur du terrain pour favoriser les discussions collectives sur des questions non compétitives. Ce faisant, l'objectif principal de l'architecte de l'inconnu est de remettre en question les représentations cognitives établies.

2.3. COLLABORATION AVEC LES SCIENTIFIQUES

Nous avons vu dans les cas de science citoyenne que nous avons étudié que la figure de gestionnaire de foules inventives interagit régulièrement avec les scientifiques avec lesquels il collabore. C'est avec lui que se construit le cahier des charges du projet en amont de la délégation à la foule. Le projet est initié par une demande du scientifique vis-à-vis d'un manque de ressources, de compétences ou autres raisons nécessaires à la bonne exécution du processus scientifique. Le scientifique sollicite donc le gestionnaire des foules inventives pour réaliser son projet. Le gestionnaire des foules inventives doit alors identifier 1) le type de tâches que le scientifique cherche à déléguer, 2) si la délégation de la tâche est avantageuse à déléguer à une foule. Ensuite, le scientifique et le gestionnaire des foules inventives collaborent pour construire ensemble la problématique de recherche afin qu'elle soit indépendante du reste du processus scientifique (Afuah & Tucci, 2012), qu'elle soit associée à un nombre limité de métriques de performance. Ils définissent également le budget pour réaliser le projet ainsi que le délai annoncé. Le gestionnaire de la foule doit estimer également le nombre de participants nécessaires pour réaliser la tâche et définir un plan d'action pour communiquer sur le projet.

Une fois que le projet est terminé, le gestionnaire des foules inventives est en charge de synthétiser la production et de l'évaluer avec les métriques préalablement définies. Dans le cas des tâches couplées inventives, cette évaluation se fait en collaboration avec les scientifiques. Ces derniers sont en charge de définir si la production a une valeur au regard de la communauté scientifique et de l'avancée dans la production de connaissance. Le gestionnaire des foules inventives lui est responsable de définir l'existence d'autres formes de valeur et de déterminer les acteurs qui peuvent l'aider à déterminer cette valeur. Dans le projet Epidemium par exemple, les instituts de recherche ont montré un certain intérêt au projet CAT ainsi qu'aux outils développés par les participants. De la même manière, le gestionnaire des foules inventives doit être capable de trouver des acteurs potentiellement intéressés par ce qui est produit.

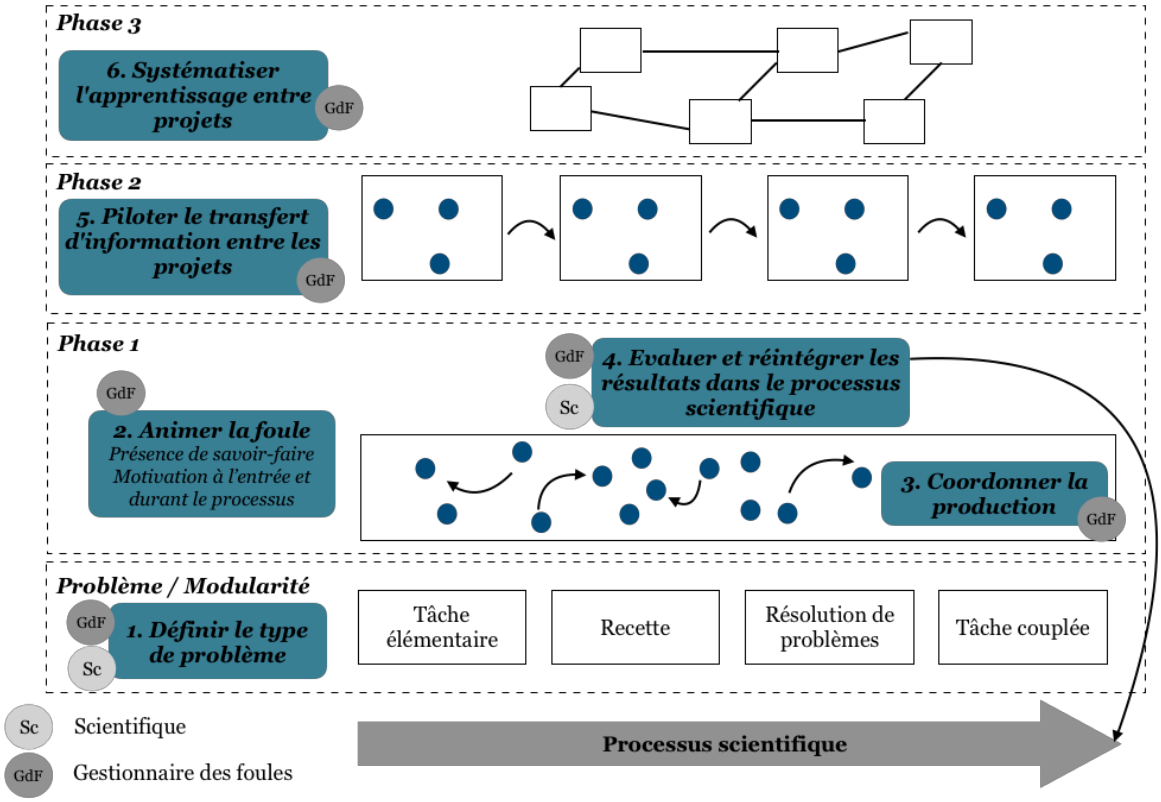


Figure 60. Répartition des activités entre le scientifique et le gestionnaire des foules inventives.

3. *LA PLACE DU GESTIONNAIRE DE FOULES INVENTIVES DANS LE PROCESSUS*

Dans les cas que nous avons étudié, les projets de science citoyenne sont généralement pilotés par une organisation indépendante des structures traditionnelle qui sert d'intermédiation entre l'organisation scientifique et l'accès à la foule. Pour exemples : le projet Galaxy Zoo, initialement à l'initiative d'un laboratoire d'astronomie, a ensuite étendu la collecte, le traitement et le codage des données à plusieurs disciplines via la plateforme Zooniverse ; la plateforme RAMP est issue d'une initiative du CDS et s'occupe des problèmes de data challenges de plusieurs disciplines scientifiques sans être rattachée à un domaine précis ; le programme DREAM challenges est spécialisé en biologie computationnelle et est utilisé par différents centres de recherche dans le domaine pour résoudre des problèmes liés aux modèles informatiques ; le programme Epidemium collabore avec plusieurs instituts sur les problématiques d'épidémiologie du cancer pour faire avancer la connaissance en épidémiologie du cancer, sans être rattaché à une structure.

Au lieu d'inclure une gestion spécifique au sein de l'organisation, les structures scientifiques font plutôt appel à des structures intermédiaires dont le rôle est de piloter le projet de science

citoyenne. Ce mode d'organisation présente plusieurs avantages, à la fois pour les structures traditionnelles et pour l'organisation intermédiaire. Premièrement, le nombre de projets de science citoyenne au sein d'une organisation scientifique n'est potentiellement pas élevé. De manière générale, les structures traditionnelles supportent la plupart de la production en interne et n'ont pas besoin de déléguer une grande partie de leur production. Les projets ayant plutôt tendance à être rares, le gestionnaire des foules inventives en charge des projets ne serait potentiellement pas occupé à temps plein. Par ailleurs, le faible nombre de projets ne permettrait pas suffisamment de systématiser des pratiques organisationnelles (phase 3). Au contraire, une organisation intermédiaire permet de multiplier les projets et de favoriser leur diversité.

Un deuxième avantage à l'utilisation de ces structures intermédiaires est la fidélisation des participants. En effet, les structures intermédiaires basése sur la foule ne peuvent fonctionner que si le nombre de participants aux projets est suffisamment important à chaque projet et les rendements sur les projets sont performants. Des plateformes comme Kaggle ou Innocentive développent des outils qui peuvent aider les participants à comprendre leur probabilité de gagner un concours particulier et à recommander des concours en fonction de leur inclination et de leurs performances passées (Chen et al., 2018). La répétition des projets augmente les chances pour la structure intermédiaire de créer des communautés actionnables de participants pour les projets. Par ailleurs, les projets de science citoyenne sont fortement associés à une culture de l'ouverture de la science de la part des participants. Une structure indépendante aura plus de chances de supporter cette image qu'un laboratoire scientifique traditionnel.

Enfin, les structures intermédiaires peuvent être pour les organisations privées de lancer des initiatives de données ouvertes sans révéler leurs stratégies et leurs problèmes de R&D à la concurrence. La structure intermédiaire, présentée ici sous le terme de « *boundary organization* » permet de contrôler la diffusion des problématiques scientifiques des organisations, quitte à développer des stratégies pour éviter que les participants ou les entreprises concurrent puissent remonter facilement jusqu'à l'entreprise demandeuse (Perkmann & Schildt, 2015). Ces structures permettent alors d'assurer des conditions contractuelles et le respect des stratégies internes dans un environnement qui favorise l'ouverture et la divulgation des informations.

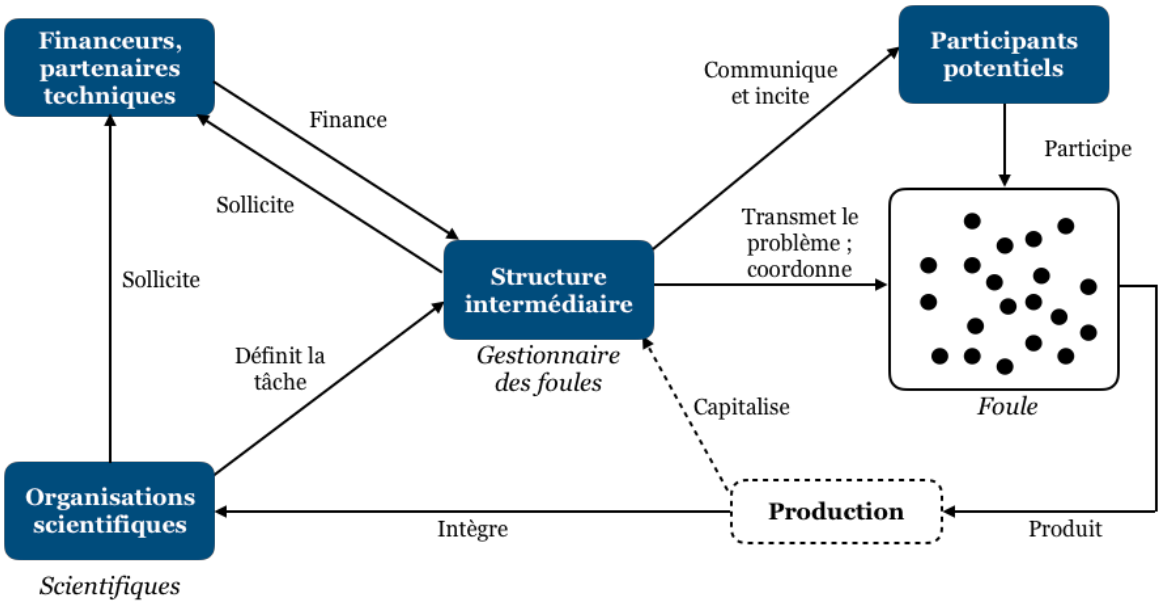


Figure 61. Modèle d'organisation via une structure intermédiaire entre les organisations scientifiques et la foule.

Dans cette forme d'organisation, le gestionnaire des foules inventives n'est pas un employé de l'organisation scientifique mais un des acteurs de la structure intermédiaire. La relation entre scientifique et gestionnaire des foules inventives peut s'apparenter à une relation entre un client et un fournisseur : le scientifique attend des résultats de la part de la structure intermédiaire tandis que le gestionnaire des foules inventives assure la performance de la délégation à la foule.

CONCLUSION GENERALE – SYNTHESE DES PRINCIPAUX RESULTATS ET PERSPECTIVES

Les principaux résultats développés dans les chapitres précédents amènent à distinguer les éléments qui permettent d'assurer la performance des projets de science citoyenne dans un contexte data-driven. Ce chapitre commence en proposant un positionnement de la thèse ainsi qu'une synthèse de ce qui a été appris. Nous présentons ensuite les implications théoriques et les implications managériales relatifs à ces résultats. Enfin, nous présentons les limites des résultats de la thèse et nous proposons quelques perspectives pour de futures recherches.

1. POSITIONNEMENT PRINCIPAL DE LA THESE : LES PROJETS DE SCIENCE CITOYENNE COMME ORGANISATION SCIENTIFIQUE EPHEMERE

La multiplication des projets de science citoyenne dans un nombre croissant de disciplines scientifiques suggère de s'intéresser à ce phénomène et aux caractéristiques de cette forme d'organisation. Dans ces projets, les organisations délèguent une partie de leur activité à des foules d'inconnus au lieu de gérer l'entièreté du processus scientifique en interne. Plusieurs exemples montrent que ces foules arrivent à être plus performantes à la fois en temps, en terme de coût et en qualité que si c'était l'organisation classique qui gérait le processus. Cependant, les études se contentent généralement pour le moment d'étudier ces projets comme des phénomènes ponctuels. C'est le point d'entrée de cette thèse : d'un côté les organisations scientifiques souhaitent profiter d'une ressource peu chère et efficace en utilisant des volontaires prêts à travailler pour des motivations autres que financières, tandis que d'un autre côté ces projets sont généralement étudiés comme des initiatives ponctuelles et non comme un processus systématique qui fasse partie intégrante du processus scientifique. Notre recherche s'articule donc sur le pilotage de la performance des projets de science citoyenne dans la durée et dans la répétition.

La principale difficulté de ce type d'organisation provient des participants : ni leur nombre, ni leurs compétences, ni le temps qu'ils vont rester à participer ne sont connus *ex ante* par les participants. Lorsque les activités qui leurs sont déléguées ne nécessitent pas de compétences spécifiques, gérer la performance du projet consiste essentiellement en la maximisation du nombre de participants pour multiplier l'effort fourni pour terminer le projet. Cependant, dès lors que l'aboutissement du projet est lié à de l'incertitude et que le résultat final n'est pas entièrement prévisible, la performance de la foule est plus difficile à établir et à modéliser. Comment est-il possible de s'assurer de la performance de ces projets dans le cas où il y a de l'incertitude durant le processus ? La gestion de la performance d'une foule pour la résolution de problèmes, notamment dans un cadre où les participants peuvent collaborer et réutiliser la production des autres nécessite un cadre théorique qui n'est pas clair dans la littérature. Par ailleurs, la littérature sur la gestion des foules ne s'est pas intéressée à la systématisation de ce processus et donc à la mise en place de moyens managériaux pour s'assurer de la transmission d'informations entre projets et assurer une amélioration continue de la performance globale. En effet, le phénomène est récent et les chercheurs se sont d'abord concentrés à en définir les contours.

Enfin, nous constatons qu'il y a une opportunité d'un point de vue des sciences de gestion à analyser les projets de science citoyenne sous l'angle **des données**. En effet, les différentes études menées sur les modes d'ouverture de la science ont constaté implicitement ce lien entre ouverture et données. Par ailleurs, les cas d'utilisation des projets de science citoyenne autour des données se multiplient dans différentes disciplines. De l'autre côté, un ensemble de chercheurs considère que notre époque se situe à l'aune d'une transformation profonde dans la science due à l'évolution de plusieurs caractéristiques des données : leur accessibilité, les méthodes et les techniques pour analyser les données, le lien entre les données et la formulation des hypothèses scientifiques. Notre époque contemporaine serait donc associée à l'émergence d'un nouveau paradigme de la science dite « data-driven » où la formulation des hypothèses serait faite directement à partir des bases de données. En passant d'un environnement pauvre en données à une avalanche de données accessibles, le goulot d'étranglement qui conditionne l'amélioration de la production de nouvelles hypothèses change de nature : on passe d'une contrainte liée où ce sont l'accessibilité aux données qui conditionne l'amélioration de la production de connaissance vers une contrainte où ce sont les capacités de traitement et de formulation de nouvelles hypothèses. Dans ce cadre, il est légitime de se demander si l'utilisation des citoyens de la science ne peut pas s'étendre à la génération des hypothèses pour répondre à cette transformation.

Le but de cette thèse est d'examiner les moyens pour piloter la performance des projets de science citoyenne dans un contexte data-driven en cherchant à répondre à ces questions : 1) En quoi les données constituent-elles un facteur pour l'ouverture de la science à des citoyens et affecte la place du scientifique? 2) Quel modèle élaborer pour se rendre compte de l'impact que provoquent les données sur l'ouverture du raisonnement scientifique? 3) Comment gérer la performance d'une organisation scientifique éphémère? Ces questions nous permettent de construire une logique générale des projets de science citoyenne et de son lien avec l'avalanche de bases de données massives dans les disciplines scientifiques.

La recherche consiste en partie en une analyse historique comparative, une étude théorique pour construire le modèle de performance et une analyse expérimentale de deux terrains. Nous étudions le cas d'un outil de gestion, le RAMP (pour Rapid Analytics and Model Prototyping), développé par le Centre de Data Science de Paris-Saclay. Cet outil propose de développer des projets basées sur des problématiques et des bases de données fournies par des scientifiques de disciplines variées (économie, biologie, physique des particules,…). Chaque problème est formalisé comme un problème d'optimisation d'algorithme d'analyse de données et soumis à une foule de participants. Le RAMP est utilisé comme plateforme de compétition et de collaboration où les spécialistes des données travaillent sur un problème pour des délais relativement courts (généralement un ou deux jours). Le deuxième terrain expérimental est Epidemium, un programme de recherche collaboratif basé sur l'épidémiologie, qui s'est déroulé entre novembre 2015 et mars 2018. Epidemium, financé en partie par les Laboratoires Roche, a pour mission de rassembler une communauté autour de de bases de données massives afin d'explorer ces bases de données pour générer des hypothèses scientifiques. L'étude expérimentale a été menée en partie

avec Olga Kokshagina, une collègue chercheuse ayant travaillé au sein du laboratoire du CGS. Elle est basée à la fois sur un ensemble de documents, une participation active pour le déroulement et l'organisation de certains évènements, une communication avec les organisateurs et les participants, et la recherche de mise en place d'outils pour piloter l'efficacité du projet.

2. PRINCIPAUX RESULTATS : MODELE MANAGERIAL ET ORGANISATIONNEL POUR LA PERFORMANCE DES PROJETS DE SCIENCE CITOYENNE DANS UN CONTEXTE DE SCIENCE DATA-DRIVEN

L'exploration des projets de science citoyenne comme une organisation scientifique éphémère nous a permis d'implémenter un nouveau cadre analytique pour l'analyse d'un nouveau phénomène organisationnel. Par organisation scientifique éphémère, nous mettons en avant deux caractéristiques des projets: leur limite dans le temps et dans l'espace, et l'intégration éphémère d'acteurs. En effet, alors que le processus scientifique se tient généralement dans des structures dédiées avec une organisation, des normes et un cadre structuré, les organisations scientifiques éphémères sont souvent hors des murs de ces institutions et limitées dans le temps. Ensuite, les acteurs d'une organisation scientifique éphémère ne participent que pour le projet et n'ont pas d'intérêts directs à ce que le projet aboutisse. Par intérêt, nous entendons un intérêt pour leur carrière (reconnaissance par les pairs), une obligation contractuelle, ou encore un intérêt personnel (concerné par les résultats du processus de recherche).

Les résultats principaux en lien avec les questions de recherche déterminent l'organisation et le pilotage des projets de science citoyenne dans un contexte data-driven. D'abord, Le résultat lié à la première question de recherche a permis de mettre en avant que le lien entre transformation du rapport aux données et évolution de l'organisation scientifique que nous avions suspecté est comparable à des situations historiques similaires. L'avalanche de bases de données massives met en défaut les capacités organisationnelles en terme de ressources dans le paradigme data-driven. Les scientifiques ne peuvent plus supporter seuls la génération d'hypothèses basées sur les données par manque de ressources et de compétences pour analyser les bases de données. Le paradigme data-driven pousse à l'apparition de nouveaux acteurs qui vont supporter une partie du processus scientifique, notamment la génération des hypothèses scientifiques. Or, l'histoire nous montre que la délégation d'activités à des non scientifiques peut mener à une redéfinition du rôle du scientifique dans le processus de production de connaissance. Cette étude historique ainsi que les exemples contemporains de projets de science citoyenne nous a enfin permis de construire un modèle d'activités déléguées à des non scientifiques basé sur trois types de tâches : les tâches élémentaires, les tâches de types recette et les tâches de résolution de problème.

Ce modèle est basé sur un modèle plus général de problem solving dans lequel les tâches sont représentées au sein d'un seul espace. Or, plusieurs études nous suggèrent que la logique de

découverte scientifique doit être représentée au moins sur deux espaces différents (e.g. Klahr & Dunbar, 1988 ; Kulkarni & Simon, 1988). Nous défendons alors le besoin d'étendre le modèle classique de résolution de problèmes à un espace vers un modèle à double espaces extensibles pour prendre en compte la délégation de la génération d'hypothèses à la foule : un espace comprenant l'ensemble des hypothèses formulables et un espace d'expérimentation dans lequel les acteurs explorent les différents codes pour analyser les bases de données massives. Nous élaborons à partir de ce modèle la notion de **tâche couplée**, qui consiste à formuler un problème (ou état désiré) en même temps que la construction de son plan d'action, et par laquelle nous pourrons analyser la génération des hypothèses.

Ce modèle étendu à quatre tâches nous permet d'étudier la performance des projets de science citoyenne. Nous montrons que l'intérêt majeur de ces projets est l'externalisation d'une partie des coûts d'exécution des tâches aux citoyens de la science, réduisant fortement le coût global de la tâche. De plus, les tâches peuvent être parallélisées et donc diviser le temps d'exécution de tâches similaires par le nombre de participants. Enfin, bien que le manque de visibilité sur les compétences des participants réduise la fiabilité globale du processus, celle-ci peut être compensée par un système de redondance (répétition des tâches). En revanche, nous voyons que la production à grand volume (plusieurs centaines voire milliers de participants) fait augmenter le risque de perte de production durant les tâches et entre les tâches.

Nous proposons alors un nouveau critère dans les organisations scientifiques éphémères pour piloter les pertes lorsque les tâches sont répétées : la **capitalisation**, c'est-à-dire la capacité de réintégrer dans une tâche toute production, retour d'expérience, outils, ou autres éléments produit antérieurement. La capitalisation se distingue de la notion d'apprentissage dans le sens où il n'est pas possible de compter sur le fait que les citoyens de la science apprennent. En effet, l'important turn-over chez les participants aux projets de science citoyenne implique que la plupart des apprentissages individuels sont perdus et ne peuvent pas être gérés. Au contraire, la capitalisation suppose que tout ce qui a été produit est formalisé et potentiellement réutilisable. Nous définissons deux formes de capitalisation qui répondent aux notions de pertes que nous avons identifiées pour les tâches de types résolution de problèmes et couplée inventive. La **capitalisation croisée,** dont le principe est limiter les pertes durant la tâche en laissant la possibilité pour les participants de réutiliser ce qui est produit durant la tâche. Nous proposons une deuxième forme que nous appelons **capitalisations séquentielle** pour piloter la perte entre des tâches successives (e.g. deux projets de science citoyenne distincts). Dans un premier temps, les participants explorent les espaces en proposant des solutions. A chaque solution proposée, ils améliorent leur connaissance de l'espace et peuvent capitaliser sur cette connaissance pour soumettre des solutions de meilleure qualité. Une fois la tâche terminée, il est possible de cartographier l'ensemble de l'exploration réalisée par chacun des participants. Cette cartographie permet de capitaliser sur ce qui a été produit durant la tâche et de le réutiliser dans la tâche successive.

Nous nous intéressons d'abord aux effets de la capitalisation croisée. Nous montrons que laisser les participants réutiliser la production des autres durant la tâche fait converger rapidement vers quelques types de solutions (que nous appelons *plateaux de fixation*) et donc diminue la diversité. Cette baisse de la diversité a un impact négatif sur l'efficacité globale (Afuah & Tucci, 2012). Nous constatons en revanche que l'intégration d'une phase « fermée » en amont où les participants ne peuvent pas réutiliser la production des autres (mode compétitif) réduit cet effet. Cela permet d'augmenter à la fois la diversité (durant la phase fermée), la qualité des solutions (durant la phase ouverte), tout en limitant les pertes durant la tâche.

Nous étudions ensuite le phénomène de capitalisation séquentielle dans le cas d'une tâche couplée inventive. Nous montrons que cette capitalisation présente un paradoxe : elle nécessite d'évaluer la valeur de ce qui est produit pour capitaliser dessus tout en n'ayant pas de fonction de valeur *ex ante*. Pour pouvoir capitaliser sur ce qui a été produit, il est nécessaire de construire une fonction de valeur *ad hoc*. Celle-ci va permettre à chaque fin de projet d'évaluer la valeur de chaque production et les zones des espaces explorées. Cette évaluation permet de construire une cartographie des zones de l'espace qui ont été explorées par les participants. Cette représentation de la production sert d'outil d'aide à la décision pour décider des futures zones des espaces à explorer. Nous montrons également que cette fonction de valeur peut être étendue entre deux tâches successives par l'introduction d'une nouvelle dimension. Nous suggérons ainsi que la réalisation d'une tâche couplée inventive ne peut pas être pensée comme un projet isolé, mais plutôt comme une succession de projets.

Finalement, notre étude nous a conduit à définir que l'efficacité des projets de science citoyenne au sein dans les organisations scientifiques pourrait être assurée avec l'aide d'une figure comme « **le gestionnaire des foules inventives** ». Nous pouvons le rapprocher de la figure du manager de projets ou du manager de portefeuille de projets. Cet acteur n'est pas nécessairement expert dans la discipline scientifique, mais il doit être capable de définir la problématique soulevée par l'organisation et la transcrire sous la forme d'un projet de science citoyenne : il doit isoler la tâche du reste du processus, formuler la problématique, mais également définir et fournir les outils et les dispositifs nécessaires à mettre en place dans le cadre du projet, ainsi que de décider les formes de collaboration entre les participants (compétitif, collaboratif, hybride), et de maîtriser les pertes de production durant l'exécution de la tâche. Il est également responsable d'animer la communauté durant le projet et de gérer les systèmes de motivation. Ensuite, dans le cas où le projet s'étale sur plusieurs épisodes, il doit être capable de cartographier l'exploration des espaces réalisée par les participants basée sur une fonction de valeur, tout en étant capable d'identifier l'apparition de nouvelles dimensions de la valeur en fonction de la production. Enfin, il est garant de ce qui est produit par les participants et donc de réintégrer la ce qui est produit dans le processus scientifique.

En outre, cette figure managériale partage certaines des activités avec le scientifique. En effet, la cartographie de la valeur suffit à piloter l'exploration et ne nécessite pas la connaissance théorique

caractéristique du rôle du scientifique. En fait, le scientifique a essentiellement un rôle de garant final et d'évaluateur de l'intérêt scientifique avéré du projet. De la même manière que nous avions observé une remise en cause de la place du scientifique dans notre analyse historique, la fonction du scientifique au sein du processus de production de connaissance évolue : il est demandé au scientifique d'être capable d'identifier la valeur scientifique des hypothèses formulées par d'autres acteurs. C'est le rôle notamment des comités scientifiques que nous avons pu voir dans le programme Epidemium.

3. IMPLICATIONS THEORIQUES : LES CHALLENGES SUCCESSIFS COMME UNE NOUVELLE FORME D'EXPLORATION ET DE RESOLUTION DE PROBLEMES POUR LES TACHES COUPLEES INVENTIVES

Dans cette section, nous discutons des implications théoriques de notre recherche. D'abord, ce travail étend le champ de recherche concernant l'organisation des projets impliquant une foule dans le modèle du crowdsourcing. Dans les approches classiques d'une activité collective menée par une foule, les recherches portent généralement sur un modèle compétitif où chaque participant travaille indépendamment des autres (Afuah & Tucci, 2012). Par ailleurs, les études qui ont étudiées la collaboration entre les participants ont généralement cherché les moyens de maximiser cette collaboration sans avoir de modèle pour comprendre l'impact de celle-ci sur la performance du projet. L'étude que nous avons mené avec le RAMP sur la possibilité de réutiliser les productions des autres a permis de montrer de manière fine quels étaient les effets de la réutilisation sur l'exploration de l'espace, mais également sur la performance globale du projet. Par ailleurs, notre étude montre que le seeker peut bénéficier d'une meilleure performance s'il agit durant le projet et non pas uniquement au début et à la fin de celui-ci. Cette approche rejoint des travaux théoriques faits par des chercheurs en économie qui suggèrent que le modèle classique est trop « statique » car il limite les interactions possibles entre les organisateurs et les participants. Selon eux un modèle plus « dynamique » permettrait d'adapter le tournoi au cours du temps suivant l'avancement des participants dans la résolution du problème et la diminution progressive de l'incertitude (Bimpikis, Ehsani, & Mostagir, 2015). Plusieurs leviers supplémentaires sont alors suggérés dans le modèle de performance tel que la mise en place de feedback (Wooten & Ulrich, 2017), de paliers intermédiaires (Benkert & Letina, 2016) ou encore l'ouverture des résultats intermédiaires durant le processus (Boudreau & Lakhani, 2015).

Deuxièmement, nous avons développé un cadre organisationnel pour le crowdsourcing afin de rendre compte de son impact au sein de l'organisation. En effet, la littérature étudie généralement les mécanismes du crowdsourcing et néglige une approche plus générale sur son impact dans le cas de sa systématisation. Dans la lignée des études proposant un cadre organisationnel pour les projets de crowdsourcing (e.g. Afuah & Tucci, 2012; King & Lakhani, 2012; Schlagwein & Bjorn-

Andersen, 2018), nous suggérons d'étudier les projets de science citoyenne et de manière plus générale le crowdsourcing à partir de la notion de « *project capabilities* » (Brady & Davies, 2004). Cette approche permet d'étendre la recherche aux notions de transmissions d'informations et de capitalisation de projet à projet, mais également sur la création de routines organisationnelles lorsque ces projets sont répétés un grand nombre de fois. Notre étude cherche notamment à limiter la perte de production dans la continuité et donc de mettre en place des systèmes pour maximiser la réutilisation de ce qui est produit. Nos terrains expérimentaux ne nous ont cependant pas permis d'explorer la mise en place de ces routines car elle nécessite des études à long terme avec un grand nombre de projets ouverts à la foule.

Troisièmement, notre étude a permis de mettre en avant des stratégies pour la gouvernance de projets soumis à la foule avec incertitude. Dans la plupart des projets avec une forte incertitude, il y a une tendance à privilégier des stratégies comme l'essai-erreur (Loch et al., 2008). Cependant, ces stratégies sont limitées dès lors que l'objectif n'est pas de trouver une solution à un problème, mais plutôt d'explorer l'ensemble des problèmes adressables dans un certain contexte et des solutions associées. En effet, les chercheurs mettent plutôt en avant le coût des essais et cherchent à limiter ces derniers pour réduire le coût total de construction d'une solution (Loch et al., 2001). Au contraire, dans notre étude de tâche couplée, l'intérêt dans l'exploration est de chercher non pas la meilleure hypothèse à construire à partir des bases de données existantes, mais plutôt d'organiser l'exhaustivité de l'exploration de toutes les bases de données pour maximiser le nombre d'hypothèses formulées et de méthodes algorithmiques pour tester ces hypothèses. Notre étude d'Epidemium a alors mis en avant une stratégie d'exploration par challenges successifs supporté par des foules de participants. Pour piloter l'avancement, nous suggérons que des acteurs soient en charge de cartographier l'exploration des espaces par les participants afin de mettre à jour, les explorations pour les challenges suivants.

Finalement, notre étude fournit des éléments pour mieux comprendre comment se met en place à long terme la gestion de la systématisation des projets de science citoyenne. Nous définissons le rôle de gestionnaire des foules inventives qui est responsable du pilotage du projet et de la transmission d'information entre projets. En outre, cette figure managériale partage certaines activités avec le scientifique. En effet, la cartographie de la valeur suffit à piloter l'exploration et ne nécessite pas la connaissance théorique caractéristique du rôle du scientifique. En fait, le scientifique a essentiellement un rôle de garant final et d'évaluateur de l'intérêt scientifique avéré du projet. De la même manière que nous avions observé une remise en cause de la place du scientifique dans notre analyse historique, la fonction du scientifique au sein du processus de production de connaissance évolue : il est demandé au scientifique d'être capable d'identifier la valeur scientifique des hypothèses formulées par d'autres acteurs.

4. IMPLICATIONS MANAGERIALES : ORGANISATION INTERMEDIAIRE POUR LA GESTION DES PROJETS DE SCIENCE CITOYENNE

En se basant sur l'analyse des plateformes RAMP et Epidemium ainsi que d'autres projets de science citoyenne (Galayx Zoo, DREAM, Kaggle), nous mettons en avant, nous suggérons qu'il y a plusieurs avantages à ce que les projets de science citoyenne soient gérés par des structures intermédiaires. Premièrement, le nombre de projets de science citoyenne au sein d'une organisation scientifique n'est potentiellement pas élevé. De manière générale, les structures traditionnelles supportent la plupart de la production en interne et n'ont pas besoin de déléguer une grande partie de leur production. Les projets ayant plutôt tendance à être rares, le gestionnaire des foules inventives en charge des projets ne serait potentiellement pas occupé à temps plein. Au contraire, une organisation intermédiaire permet de multiplier les projets et de favoriser leur diversité. Un deuxième avantage à l'utilisation de ces structures intermédiaires est la fidélisation des participants. En effet, les structures intermédiaires basées sur la foule ne peuvent fonctionner que si le nombre de participants aux projets est suffisamment important à chaque projet et les rendements sur les projets sont performants. Des plateformes comme *Kaggle* ou *Innocentive* développent des outils qui peuvent aider les participants à comprendre leur probabilité de gagner un concours particulier et à recommander des concours en fonction de leur inclination et de leurs performances passées (Chen et al., 2018). Enfin, les structures intermédiaires peuvent être pour les organisations privées de lancer des initiatives de données ouvertes sans révéler leurs stratégies et leurs problèmes de R&D à la concurrence (voir les travaux de Perkmann & Schildt, 2015).

Dans ces organisations intermédiaires, le gestionnaire des foules inventives n'est pas un employé de l'organisation scientifique mais un des acteurs de la structure intermédiaire. La relation entre scientifique et gestionnaire des foules inventives peut s'apparenter à une relation entre un client et un fournisseur : le scientifique attend des résultats de la part de la structure intermédiaire tandis que le gestionnaire des foules inventives assure la performance de la délégation à la foule. Cette structure a pour principales missions de :

- Co-définir la problématique avec les scientifiques,
- Animer la foule,
- Coordonner la production,
- Piloter le transfert d'informations entre projets,
- Systématiser l'apprentissage

5. LIMITES DU TRAVAIL DE RECHERCHE

Les limites de notre recherche portent principalement sur notre matériau de recherche et sur le choix de notre méthodologie de recherche. En optant pour une démarche analytique à partir de seulement deux études de cas (un projet de type résolution de problèmes et un projet de type tâche couplée), nous admettons une limite quant à la généralisation de nos résultats de recherche. Par conséquent, afin de renforcer la validité des résultats et des cadres d'analyse élaborés dans cette thèse, il nous paraîtrait intéressant de les tester sur d'autres terrains de recherche. Par exemple, il pourrait être judicieux d'étudier la gestion de tâches couplées inventives ouverte à une foule dans d'autres contextes que celui scientifique. En effet, nous voyons apparaître l'utilisation des citoyens dans des contextes variés, comme dans les exemples récents avec la mise en place de Référendum d'Initiative Citoyenne. Par ailleurs, il serait intéressant de mener des études comparatives sur la gestion de la génération d'hypothèses en comparant le processus lorsqu'il est porté par une foule ou par des individus en groupe mais dont le nombre est limité. Néanmoins, le choix des projets se justifie par plusieurs points. D'abord, le projet RAMP est un cas presque unique pour étudier de manière fine avec des métriques quantifiées les effets de la réutilisation dans le cas d'un projet ouvert à la foule. Ensuite, le cas d'Epidemium est également le seul cas à notre connaissance où la génération d'hypothèses basées sur les données est ouvert à une foule de participants. En outre, nous estimons que ces cas extrêmes constituent une base solide pour développer un modèle théorique qui paraît plausible en lui-même et pour lequel les cas servent à identifier les principaux paramètres (Sigglekow, 2007).

Sur le plan théorique, le modèle formel que nous avons développé permet de mettre en évidence des phénomènes qui avaient jusque-là été négligés par la littérature sur les sciences citoyennes et le crowdsourcing. Cependant, la notion de tâche couplée inventive pourrait être davantage analysée : y'a-t-il d'autres situations autres que la génération d'hypothèses dans lesquelles ce modèle de tâche peut s'appliquer ? Si oui, est-ce que son application engendre les mêmes effets que ceux que nous avons identifiés ?

6. PERSPECTIVES : EFFET A LONG TERME DES PROJETS DE SCIENCE CITOYENNE

Cette thèse peut servir comme point de démarrage pour étudier la considération à long terme des projets de science citoyenne dans le paysage scientifique, mais également de façon plus large l'intégration des projets incluant une foule dans les organisations. A travers nos investigations, nous avons formulés un certain nombre de questions sur le sujet et obtenu des résultats significatifs dont les résultats s'adressent de manière transverse à toute discipline scientifique qui souhaiterait intégrer cette forme d'organisation. Pour terminer, quelques futures directions de recherche sont présentées. Celles-ci sont principalement en lien avec l'impact à long-terme que peut générer l'utilisation systématique des projets faisant appel à une foule.

Le premier élément qui nous semble intéressant à explorer concerne la phase 3 présente dans le modèle de Brady et Davies (2000 ; 2004). Celle-ci interroge la façon dont l'organisation capitalise sur l'ensemble des projets mis en place pour créer des routines et des standardisations qui échappent à l'institutionnalisation car éphémères. Nous avons vu durant la thèse que certaines plateformes tel que Galaxy Zoo ou DREAM pourraient être des supports intéressants pour adresser ces questions.

Ensuite, un aspect négligé dans notre thèse est tout ce qui concerne le système de motivation et d'incitations des participants à intégrer un projet. Un certain nombre d'études ont été menées pour comprendre pourquoi les volontaires intégraient les projets de science citoyenne et les évolutions de leurs motivation durant le projet (Domroese & Johnson, 2016; Raddick et al., 2013; Rotman et al., 2012; Sauermann & Franzoni, 2014). Cependant, peu d'études se sont intéressées aux mécanismes à mettre en place pour maintenir la communauté en place et la capacité à renouveler cette communauté ainsi que les compétences nécessaires en fonction des problèmes adressés. D'autres travaux menés sur les communautés open-source (von Krogh, Spaeth, & Lakhani, 2003), de lead-users (Jeppesen & Frederiksen, 2006) ou encore des communautés d'apprentissage professionnels (Bolam et al., 2005) peuvent nous aider à concevoir des systèmes d'incitation adaptés. Par ailleurs, ces études devraient s'intéresser aux distinctions entre les motivations liées à l'activité en tant que telle, à la contribution à la science en général, ainsi qu'à l'intérêt spécifique à une discipline ou à un sujet en particulier.

Notre perspective sur le long terme des projets de science citoyenne fait également écho à des problématiques en sciences sociales, en sciences de gestion ainsi qu'en économie autour de la question du travail numérique. Des recherches mettent en avant par ces projets l'émergence d'un travail, souvent cognitif et dématérialisé, qui apparaît hors des lieux ordinaires. Ces activités peuvent en effet être assimilées à la définition d'un travail car elles sont « productrices de valeur, faisant l'objet d'un quelconque encadrement contractuel et soumises à des métriques de performance » (Casilli, 2015). Les travaux autour de cette nouvelle forme de travail s'appuient notamment sur la plateforme de crowdsourcing *Amazon Mechanical Turk*, où les participants sont payés quelques centimes pour réaliser des tâches simples : traduire une phrase, compter, vérifier l'exactitude d'une page Wikipedia (Kittur et al., 2008). Cette organisation du travail cognitif est associée à la notion de « capitalisme cognitif » (Scholz, 2013) : d'abord pensée sous des formes classiques de division du travail, notre travail pousse à étendre cette forme de capitalisme vers de nouvelles formes d'organisation, plus collaboratives avec moins de répartition établie du travail. Par ailleurs, dès que le résultat du projet n'est plus associé à un individu mais la résultante d'une synergie entre individus basée sur la réutilisation des productions antérieures, il devient difficile de maintenir une logique de rétribution classique entre les participants (Boudreau & Lakhani, 2015). Comment repenser dans ce cas de nouvelles formes de propriété intellectuelle ainsi que de règles et de normes pour protéger les contributeurs ? De plus, est-ce que ce système a encore un sens dans ce modèle de travail ?

Enfin, non examinées dans ce travail, les approches de conception sociales et psychologiques pourraient fournir de nouvelles perspectives sur le processus d'exploration dans les tâches de résolution de problèmes et les tâches couplées inventives. Nous avons vu que certains projets peuvent mener de façon inattendue à de nouvelles valeurs qui n'étaient pas prévisible au début du projet (par exemple le projet CAT ou le projet ELSE dans Epidemium). Est-il possible de piloter cette pensée innovante afin de la stimuler ou la réfréner suivant les objectifs définis par les organisateurs ? Des travaux menés par les équipes en psychologie au Centre de Gestion Scientifique de MINES ParisTech ont exploré cette question au travers de la notion d'effet de fixation (e.g. Agogué et al., 2014). Comment évoluent les effets de fixation de manière collective et individuelle dans le cas d'une tâche couplée inventive ?

REFERENCES CLES

Champ de recherche	*Références clés*
Sciences citoyennes	Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984. Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, *43*(1), 1–20. Houllier, F. (2016). Les sciences participatives en France: Etat des lieux, bonnes pratiques et recommandations. *Les Sciences Participatives En France(2016)*, 63.
1. Organisation des sciences citoyennes	- **Organisation et gestion du processus scientifique** Chompalov, I., Genuth, J., & Shrum, W. (2002). The organization of scientific collaborations. *Research Policy*, *31*(5), 749–767. - **Crowdsourcing et modèle d'organisation de la foule** Afuah, A., & Tucci, C. L. (2012). Crowdsourcing As A Solution To Distant Search DISTANT SEARCH, (February 2016) Boudreau, K. J., & Lakhani, K. R. (2015). "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. *Research Policy*, *44*(1), 4–19. Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine*, *14*(06), 1–5. - **Gestion de projets** King, A. A., & Lakhani, K. R. (2012). The Contingent Effect of Absorptive Capacity: An Open Innovation Analysis. *SSRN Electronic Journal*. Brady, T., & Davies, A. (2004). Building project capabilities: From exploratory to exploitative learning. *Organization Studies*, *25*(9), 1601–1621.
2. Transformation épistémologique dans un contexte du Big Data	- **Big Data** Laney, D. (2001). Application Delivery Strategies. *Meta Group*, (September). - **Algorithmes d'Intelligence Artificielle** Shmueli, G. (2011). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. - **Transformation épistémologique** Anderson, B. C. (2008). The End of Theory : The Data Deluge Makes the Scientific Method Obsolete The End of Theory : The Data Deluge Makes the Scientific Method Obsolete, 14–16. Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intesive Scientific Discovery. *E-Science and Information Management*, 1–1. Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848.

3. Histoire des sciences et transformation par les données	- **Introduction des instruments scientifique** Daumas, M. (1953). Les Instruments scientifiques aux XVIIe et XVIIIe siècles. *Revue Philosophique de La France Et de l'Etranger*, (146), 402–403. Licoppe, C. (1996). *La Formation de la pratique scientifique : le discours de l'expériènce en France et en Angleterre : 1630-1820*. La Découverte. Shapin, S. (1989). The Invisible Technician. *American Scientist*, *77*(6), 554–563. - **Stochastique dans le processus scientifique** Schaffer, S. (1988). Astronomers Mark Time: Discipline and the Personal Equation. *Science in Context*, *2*(1), 115–145. Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. *Technology and Culture* (Vol. 29). Belknap Press of Harvard University Press.
4. Modèles de l'activité scientifique	- **Modèles de la logique de découverte scientifique** Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, *12*(2), 139–175. King, R. D., Rowland, J., Aubrey, W., Liakata, M., Markham, M., Soldatova, L. N., … Pir, P. (2009). The robot scientist adam. *Computer*, *42*(8), 46–54. Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48. - **Modèles à double espace** Kazakçi, A. O. (2013). On the imaginative constructivist nature of design: A theoretical approach. *Research in Engineering Design*, *24*(2), 127–145. von Hippel, E., & von Krogh, G. (2015). Identifying Viable "Need–Solution Pairs": Problem Solving Without Problem Formulation. *Organization Science*, orsc.2015.1023.

REFERENCES

Abernathy, W. J., & Rosenbloom, R. S. (1969). Parallel Strategies in Development Projects. *Management Science*, *15*(10), B-486-B-505.

Adam-Bourdarios, C., Cowan, G., Germain, C., & Guyon, I. (2015). The Higgs boson machine learning challenge. *NIPS Workshop on High-Energy Physics and Machine Learning*, *42*, 19–55.

Afuah, A. (2018). Crowdsourcing: A Primer and Research Framework (Vol. 1). Oxford University Press.

Afuah, A., & Tucci, C. L. (2012). Crowdsourcing As A Solution To Distant Search DISTANT SEARCH, (February 2016).

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. Information Systems Research, 25(3), 443–448.

Aggeri, F., & Labatut, J. (2010). La gestion au prisme de ses instruments. Une approche généalogie des theories fondées sur les instruments de gestion. FInance Contrôle Stratégie, 13(3), 5–37.

Agogué, M., Comtet, G., Menudet, J.-F., Picard, R., & Le Masson, P. (2017). Managing innovative design within the health ecosystem : the Living Lab as an architect of the unknown. Management & Avenir Santé, N° 1(1), 17.

Agogué, M., Kazakçi, A. O., Hatchuel, A., Le Masson, P., Weil, B., Poirel, N., & Cassotti, M. (2014). The impact of type of examples on originality: Explaining fixation and stimulation effects. Journal of Creative Behavior, 48(1), 1–12.

Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. APL Materials, 4(5), 053208.

Allen, T. J. (1966). Studies of the problem-solving process in engineering design. IEEE Transactions on Engineering Management.

Allen, T. J., & Katz, R. (1986). The dual ladder: motivational solution or managerial delusion? R&D Management, 16(2), 185–197.

Allen, T. J., Lee, D. M. S., & Tushman, M. L. (1980). R&D Performance as a Function of Internal Communication, Project Management, and the Nature of the Work. IEEE Transactions On Engineering Management, 27(1), 2–12.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8), 831–838.

Alonso, R., Piñeros, M., Laversanne, M., Musetti, C., Garau, M., Barrios, E., & Bray, F. (2018). Lung cancer incidence trends in Uruguay 1990–2014: An age-period-cohort analysis. Cancer Epidemiology, 55, 17–22.

Amatriain, X. (2012). Netflix Recommendatons - Beyond the 5 Stars. ... Netflix. Com/2012/04/Netflixrecommendations-beyond-5 ..., (Part 2), 1–5.

Amorín, R. O., Pérez-Montero, E., & Vílchez, J. M. (2010). On the oxygen and nitrogen chemical abundances and the evolution of the "green pea" galaxies. Astrophysical Journal Letters, 715(2 PART 2), 128–132.

Anderson, B. C. (2008). The End of Theory : The Data Deluge Makes the Scientific Method Obsolete The End of Theory : The Data Deluge Makes the Scientific Method Obsolete, 14–16.

Anderson, T., & Dron, J. (2014). Teaching Crowds: Learning and Social Media. Teaching Crowds: Learning and Social Media.

Antonsson, E., & Cagan, J. (2001). Formal Engineering Design Synthesis. Cambridge University Press.

Appley, D. G., & Winder, A. E. (1977). An Evolving Definition of Collaboration and Some Implications for the World of Work. The Journal of Applied Behavioral Science, 13(3), 279–291.

Arnstein, S. R. (1969). A Ladder Of Citizen Participation. Journal of the American Planning Association, 35(4), 216–224.

Armstrong, D. J., & Cole, P. (2004). Managing distances and differences in geographically distributed work groups. In Diversity in work teams: Research paradigms for a changing workplace. (pp. 187–215). MIT Press.

Asgharbeygi, N., Langley, P., Bay, S., & Arrigo, K. (2006). Inductive revision of quantitative process models.

In Ecological Modelling (Vol. 194, pp. 70–79).

Ayas, K., & Zeniuk, N. (2001). Project-based Learning: Building Communities of Reflective Practitioners. Management Learning.

Azencott, C. A., Aittokallio, T., Roy, S., Norman, T., Friend, S., Stolovitzky, G., … Zinovyev, A. (2017, September 29). The inconvenience of data of convenience: Computational research beyond post-mortem analyses. Nature Methods.

Bagla-Gökalp, L. (1996). Le chercheur et son instrument: Changement des techniques de mesure et des pratiques scientifiques en mécanique des fluidesLe chercheur et son instrument: Changement des techniques de mesure et des pratiques scientifiques en mecanique des fluides. Revue Française de Sociologie, 37(4), 537.

Balconi, M., Breschi, S., & Lissoni, F. (2004). Networks of inventors and the role of academia: An exploration of Italian patent data. Research Policy, 33(1), 127–145.

Ballantine, K. R., Hanna, S., Macfarlane, S., Bradbeer, P., Teague, L., Hunter, S., … Skeen, J. (2018). Childhood cancer registration in New Zealand: A registry collaboration to asssess and improve data quality. Cancer Epidemiology, 55, 104–109.

Baraniuk, R. G. (2011, February 11). More is less: Signal processing and the data deluge. Science. http://doi.org/10.1126/science.1197448

Barthe, Y. (2013). Epidémiologie populaire. In Dictionnaire de la participation dictionnaire critique interdisciplinaire de la participation. GIS Participation du public, décision, démocratie participative.

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: Local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography.

Beale, R. (2007). Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. International Journal of Human Computer Studies.

Beaver, D. de B., & Rosen, R. (1979). Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. Scientometrics, 1(3), 231–245.

Becquemont, D. (1944-. . . .). (2016). Darwin, darwinisme, évolutionnisme. Editions Kimé.

Ben Aissa, H. (2001). Quelle méthodologie de recherche appropriée pour une construction de la recherche en gestion□? Xième Conférence de l'Association Internationale de Management Stratégique, 27.

Benkert, J.-M., & Letina, I. (2016). Designing Dynamic Research Contests. SSRN.

Berger, D. (1999). A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1—Ancient times through the 19th century. In Medical Laboratory Observer (Vol. 31(7), pp. 28–40).

Bian, J., Topaloglu, U., & Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In Proceedings of the 2012 international workshop on Smart health and wellbeing - SHB '12 (p. 25).

Bimpikis, K., Ehsani, S., & Mostagir, M. (2015). Designing Dynamic Contests. In Proceedings of the Sixteenth ACM Conference on Economics and Computation - EC '15 (pp. 281–282). New York, New York, USA: ACM Press.

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work. In Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work - GROUP '03 (p. 339).

Björk, B. C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. Journal of Informetrics, 7(4), 914–923.

Blaikie, N., & Priest, J. (2019). Designing Social Research The Logic of Anticipation. Polity Press.

Bohannon, J. (2017). The cyberscientist. Science, 357(6346), 18–21.

Bolam, R., Mcmahon, A., Stoll, L., Thomas, S., Wallace, M., Greenwood, A., … Smith, M. (2005). Creating and Sustaining Effective Professional Learning Communities. Department for education and skills Research Report No 637.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009). Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. A CAISE Inquiry Group Report.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. BioScience, 59(11), 977–984.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next Steps for Citizen Science. Science, 343(6178), 1436–1437.

Borgman, C. L. (2007). Scholarship in the digital age□: information, infrastructure, and the Internet. MIT Press.

Boudreau, K. J., Lacetera, N., & Lakhani, K. R. (2011). Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis. Management Science, 57(5), 843–863.

Boudreau, K. J., & Lakhani, K. R. (2015). "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. Research Policy, 44(1), 4–19.

Boudreau, K. J., Lakhani, K. R., & Lacetera, N. (2008). Parallel Search , Incentives and Problem Type□: Revisiting the Competition and Innovation Link. Technology, 1–41.

Boulton, G., Rawlins, M., Vallance, P., & Walport, M. (2011, May 14). Science as a public enterprise: The case for open data. The Lancet.

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. SSRN.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. Convergence, 14(1), 75–90.

Brady, T., & Davies, A. (2004). Building project capabilities: From exploratory to exploitative learning. Organization Studies, 25(9), 1601–1621.

Brady, T., Marshall, N., Prencipe, A., & Tell, F. (2002). Making Sense of Learning Landscapes in Project-Based Organisations. In European Conferenece on Organizational Knowledge, Learning and Capabilities. Athens.

Brokaw, L. (2011). Could 'Citizen Science' Be Better Than Academy Science?, 11–13.

Brooks, F. (1996). The mythical man-month: Essays on software engineering. IEEE Annals of the History of Computing.

Brown, P. (1987). Popular Epidemiology: Community Response to Toxic Waste-Induced Disease in Woburn, Massachusetts. Science, Technology, & Human Values, 124(3), 78–85.

Brown, P. (1992). Popular Epidemiology and Toxic Waste Contamination: Lay and Professional Ways of Knowing. Journal of Health and Social Behavior, 33(3), 267.

Brown, J. S., & Duguid, P. (2001). Knowledge and Organization: A Social-Practice Perspective. Organization Science, 12(2), 198–213.

Buecheler, T., Sieg, J. H., Füchslin, R. M., & Pfeifer, R. (2010). Crowdsourcing , Open Innovation and Collective Intelligence in the Scientific Method□: A Research Agenda and Operational Framework. In Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems (pp. 679–686). MIT Press.

Butler, H. J., Ashton, L., Bird, B., Cinque, G., Curtis, K., Dorney, J., … Martin, F. L. (2016). Using Raman spectroscopy to characterize biological materials. Nature Protocols, 11(4), 664–687.

Cabanes, B. (2017, June 20). Modéliser l'émergence de l'expertise et sa gouvernance dans les entreprises innovantes : des communautés aux sociétés proto-épistémiques d'experts. MINES ParisTech - PSL Research University.

Cadeddu, A. (1985). Pasteur et le choléra des poules: révision critique d'un récit historique. History and Philosophy of the Life Sciences, 7(1), 87–104.

Callahan, A., Dumontier, M., & Shah, N. H. (2011). HyQue: Evaluating hypotheses using Semantic Web technologies. Journal of Biomedical Semantics, 2(2), S3.

Callon, M., Lascoumes, P., & Barthe, Y. (2001). Agir dans un monde incertain: essai sur la démocratie technique. Seuil.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C. M., … Vandenberg, J. (2009). Galaxy Zoo Green Peas: Discovery of a class of compact extremely star-forming galaxies. Monthly Notices of the Royal Astronomical Society, 399(3), 1191–1205.

Cariou, J.-Y. (2009). Former l'esprit scientifique en privilégiant l'initiative des élèves dans une démarche s'appuyant sur l'épistémologie et l'histoire des sciences. Thèse.

Carlile, P. R. (2004). Transferring, Translating, and Transforming: An Integrative Framework for Managing Knowledge Across Boundaries. Organization Science, 15(5), 555–568.

Cavallo, R., Street, W., York, N., & Haven, N. (2012). Efficient Crowdsourcing Contests. Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), 8

Casilli, A. (2015). Digital Labor : travail, technologies et conflictualités.

Chandler, A. D. (1990). Scale and Scope: The Dynamics of Industrial Capitalism. Academy of Management Review.

Chelly Dagdia, Z., Zarges, C., Schannes, B., Micalef, M., Galiana, L., Rolland, B. B., … Benchoufi, M. (2018). Rough Set Theory as a Data Mining Technique: A Case Study in Epidemiology and Cancer Incidence Prediction. Dublin, Irland: European Conference on Machine Learning.

Chen, Y.-J., Dai, T., Korpeoglu, C. G., Körpeoğlu, E., Sahin, O., Tang, C. S., & Xiao, S. (2018). Innovative Online Platforms: Research Opportunities. SSRN.

Chesbrough, H. (2006). Open Innovation : A New Paradigm for Understanding Industrial Innovation.

Chiolero, A. (2013, November). Big data in epidemiology: Too big to fail? Epidemiology.

Chompalov, I., Genuth, J., & Shrum, W. (2002). The organization of scientific collaborations. Research Policy, 31(5), 749–767.

Christianson, J. R. (2000). On Tycho's Island: Tycho Brahe and His Assistants, 1570–1601. Cambridge U. Press.

Clark, H. H., & Brennan, S. E. (2004). Grounding in communication. In Perspectives on socially shared cognition. (pp. 127–149).

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly, 35(1), 128.

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? BioScience, 58(3), 192–197. http://doi.org/10.1641/B580303

Cole, S. A. (2011). Acculturating Forensic Science: What is 'Scientific Culture', and How Can Forensic Science Adopt It? SSRN (Vol. 38).

Cole, A. L., Austin, A. E., Hickson, R. P., Dixon, M. S., & Barber, E. L. (2018, August 1). Review of methodological challenges in comparing the effectiveness of neoadjuvant chemotherapy versus primary debulking surgery for advanced ovarian cancer in the United States. Cancer Epidemiology. Elsevier.

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? BioScience, 58(3), 192–197.

Comiti, V. (1976). Eléments historiques de l'utilisation de la méthode statistique en médecine, (2), 607–609.

Coombs, R., & Hull, R. (1998). "Knowledge management practices" and path-dependency in innovation. Research Policy.

Cooper, C. B., Dickinson, J., Phillips, T., & Bonney. R. (2007). Citizen science as a tool for conservation in residential ecosystems. Ecology and Society, 12(2), 1–9.

Corley, E. A., Boardman, P. C., & Bozeman, B. (2006). Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. Research Policy, 35(7), 975–993.

Cottle, M., & Hoover, W. (2013). Transforming Health Care Through Big Data, 1–24.

Cranshaw, J., & Kittur, A. (2011). The Polymath Project: Lessons from a Successful Online Collaboration in Mathematics. Proceedings of the 2011 Annual Conference on Computer Human Interaction, 1865–1874.

Crawford, K. (2013). The Hidden Biases in Big Data. HBR Blog Network, 9–10.

Cronin, M. A., & Weingart, L. R. (2007). REPRESENTATIONAL GAPS , INFORMATION PROCESSING , AND CONFLICT IN FUNCTIONALLY DIVERSE TEAMS, 32(3), 761–773.

Cross, N. (2001). Design Cognition: results from protocol and other empirical studies of design activity. In Design Knowing and Learning: Cognition in Design Education (pp. 79–103). Elsevier.

Cummings, J. N., & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. Research Policy, 36(10), 1620–1634.

Dahlin, K. B., Weingart, L. R., & Hinds, P. J. (2005). Team diversity and information use. Academy of Management Journal.

Daumas, M. (1950). Quelques fabricants d'instruments scientifiques anciens. Revue d'histoire Des Sciences, 3(4), 364–370.

Daumas, M. (1953). Les Instruments scientifiques aux XVIIe et XVIIIe siècles. Revue Philosophique de La France Et de l'Etranger, (146), 402–403.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: the sexiest job of the 21st century. Harvard Business Review, 90(10), 70–6, 128.

David, A. (1996). Structure et dynamique des innovations managériales. Cinquième Conférence de l'AIMS, (1), 1–29.

David, A. (1999). Logique, épistémologie et méthodologie en sciences de gestion. In Conférence de l'AIMS (pp. 1–23).

David, A., Hatchuel, A., & Laufer, R. (2012). Les nouvelles fondations des sciences de gestion : éléments d'épistémologie de la recherche en management. Mines ParisTech.

David, P. A. (2007). THE HISTORICAL ORIGINS OF ' OPEN SCIENCE ' An Essay on Patronage , Reputation and Common Agency Contracting in the Scientific Revolution.

Davies, A., & Brady, T. (2000). Organisational capabilities and learning in complex product systems: towards repeatable solutions. Research Policy, 29(7–8), 931–953.

De Dreu, C. K. W., & West, M. A. (2001). Minority dissent and team innovation: The importance of participation in decision making. Journal of Applied Psychology, 86(6), 1191–1201.

DeFillippi, R. J., Jones, C., & Arthur, M. B. (2001). Project-Based Learning as the Interplay of Career and Company Non-Financial Capital. Management Learning.

Delamont, S., & Atkinson, P. (2001). Doctoring Uncertainty : Mastering Craft Knowledge Doctoring Uncertainty : Mastering Craft Knowledge, 31(1), 87–107.

Denrell, J., Fang, C., & Winter, S. G. (2003). The economics of strategic opportunity. Strategic Management Journal. http://doi.org/10.1002/smj.341

DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. Contemporary Clinical Trials,

45, 139–145. http://doi.org/10.1016/j.cct.2015.09.002

DiPalantino, D., & Vojnovic, M. (2009). Crowdsourcing and all-pay auctions. In Proceedings of the tenth ACM conference on Electronic commerce - EC '09 (p. 119). New York, New York, USA: ACM Press. http://doi.org/10.1145/1566374.1566392

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73.

Domroese, M. C., & Johnson, E. A. (2016). Why watch bees? Motivations of citizen science volunteers in the Great Pollinator Project.

Dubois, D., Hájek, P., & Prade, H. (2000). Knowledge-Driven versus Data-Driven Logics. Journal of Logic, Language, and Information (Vol. 9).

Dummett, M. A. E. (2000). Elements of intuitionism. Oxford logic guides.

Dyche, J. (2012). Big data '"Eurekas!"' don't just happen.

E. Fyffe, D., W. Hines, W., & Kee Lee, N. (1968). System Reliability Allocation and a Computational Algorithm. IEEE Transactions on Reliability, R-17(2), 64–69.

Edelman, J. (2011). UNDERSTANDING RADICAL BREAKS: MEDIA AND BEHAVIOR IN SMALL TEAMS ENGAGED IN REDESIGN SCENARIOS. Stanford.

Egghe, L. (2006). Theory and practise of the g-index. Scientometrics, 69(1), 131–152.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. Academy of Management Journal, 50(1), 25–32.

Elias, A. A., Cavana, R. Y., & Jackson, L. S. (2002). Stakeholder analysis for R & D project management. R and D Management, 32(4), 301–310.

Erickson, T. (2011). Some Thoughts on a Framework for Crowdsourcing, 1–4.

Eris, O. (2006). Insisting on truth at the expense of conceptualization: Can engineering portfolios help? International Journal of Engineering Education, 22(3), 551–559.

Estellés-Arolas, E., & González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. Journal of Information Science, 38(2), 189–200.

Ethiraj, S. K., Kale, P., Krishnan, M. S., & Singh, J. V. (2005). Where do capabilities come from and how do they matter? A study in the software services industry. Strategic Management Journal.

Ewing, A. D., Houlahan, K. E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T. N., … Boutros, P. C. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature Methods, 12(7), 623–630.

Fan, J., Han, F., & Liu, H. (2014, June 1). Challenges of Big Data analysis. National Science Review. http://doi.org/10.1093/nsr/nwt032

Fanelli, D. (2009, May 29). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. (T. Tregenza, Ed.)PLoS ONE. Public Library of Science.

Faraj, S., & Majchrzak, A. (2011). Knowledge Collaboration in Online Communities, 22(5), 1224–1239.

Fayard, A.-L., & DeSanctis, G. (2008). Kiosks, Clubs and Neighborhoods: The Language Games of Online Forums. Journal of the Association for Information Systems, 9(10), 677–705.

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In Opening Science (pp. 17–47). Cham: Springer International Publishing.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. Research Policy, 43(1), 1–20.

Fischer, D. A., Schwamb, M. E., Schawinski, K., Lintott, C., Brewer, J., Giguere, M., … Zimmermann, V. (2012). Planet Hunters: The first two planet candidates identified by the public using the Kepler public archive data. Monthly Notices of the Royal Astronomical Society, 419(4), 2900–2911.

Foster, A., & Ford, N. (2003). Serendipity and information seeking: An empirical study. Journal of Documentation.

Galegher, J., Sproull, L., & Kiesler, S. (1998). Legitimacy, authority, and community in electronic support groups. Written Communication, 15(4), 493–530.

Gallaugher, J., & Ransbotham, S. (2010). Social media and customer dialog management at Starbucks. MIS Quarterly Executive.

Gao, Y., Kinoshita, J., Wu, E., Miller, E., Lee, R., Seaborne, A., … Clark, T. (2006). SWAN: A distributed knowledge infrastructure for Alzheimer disease research. Web Semantics, 4(3), 222–228.

Gaulon, P.-A. (1997). Les instruments scientifiques – Définition et historique. Bulletin de La Sabix, 18(18), 9–15.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.

Gann, D. M., & Salter, A. (1998). Learning and Innovation Management in Project-Based, Service-Enhanced Firms. International Journal of Innovation Management, 02(04), 431–454.

Gann, D. M., & Salter, A. J. (2000). Innovation in project-based, service-enhanced firms: the construction of

complex products and systems. Research Policy, 29(7–8), 955–972.

Geiger, D., Seedorf, S., Nickerson, R., & Schader, M. (2011). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In Proceedings of the 17th Americas Conference on Information Systems, AMCIS 2011 (pp. 1–11).

George, G., Haas, M. R., & Pentland, A. (2014, April 4). From the editors: Big data and management. Academy of Management Journal. Academy of Management Briarcliff Manor, NY.

Gibbons, C. M., Limoges, C., Nowotny, H., Scott, P., & Trow, M. (2010). The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies The new production of knowledge: The dynamics of science and research in contemporary societies, 155–167.

Gillier, T., Piat, G., Roussel, B., & Truchot, P. (2010). Managing innovation fields in a cross-industry exploratory partnership with C-K design theory. Journal of Product Innovation Management, 27(6), 883–896.

Gillier, T., & Sylvain, L. (2018). Experimenting in the Unknown: Lessons from The Manhattan Project. European Management Review.

Gligorov, V. V. (2015). Real-time data analysis at the LHC: present and future, 1–18.

González-Bailõn, S. (2013). Social science in the era of big data. In Policy and Internet (Vol. 5, pp. 147–160). Wiley-Blackwell.

Gowers, T. (2009). Massively collaborative mathematics, 461(October), 879–881.

Grant, R. M. (1996). TOWARD A KNOWLEDGE-BASED THEORY OF THE FIRM. Strategic Management Journal, 17, 109–122.

Gray, J. (2009). Fourth Paradigm: Data-Intensive Scientific Discovery - Microsoft Research.

Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., … Griffith, O. L. (2017, February 1). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nature Genetics.

Haeussler, C., & Sauermann, H. (2012). Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. Research Policy.

Hagstrom, W. O. (1964). Traditional and Modern Forms of Scientific Teamwork. Administrative Science Quarterly, 9(3), 241–263.

Hagstrom, W. O. (1965). The Scientific Community. Historisches Wörterbuch der Philosophie (Vol. 8).

Haklay, M. (2015). Citizen Science and Policy: A European Perspective. Common Labs. Case Study Series., 4, 76.

Hamel, G., & Prahalad, C. K. (1994). Competing for the future. Harvard Business School Press.

Harman, G. (2002). Internal critique: A logic is not a theory of reasoning and a theory of reasoning is not a logic. Studies in Logic and Practical Reasoning, 1(C), 171–186.

Han, J., & Kamber, M. (2012). Data mining□: concepts and techniques. Elsevier.

Hand, E. (2010, August 5). Citizen science: People power. Nature. Nature Publishing Group.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. New York, NY: Springer New York.

Hatchuel, A., & Weil, B. (1992). L'expert et le système□: gestion des savoirs et métamorphose des acteurs dans l'entreprise industrielle, suivi de quatre histoires de systèmes-experts. Paris: Economica.

Hatchuel, A. (2001). Towards design theory and expandable rationality: The unfinished programme of Herbert Simon. Journal of Management and Governance, 5(3–4), 260–273.

Hatchuel, A., & Weil, B. (2003). a New Approach of Innovative Design□: an Introduction To C-K Theory. In Iced 2003 (pp. 1–15).

Hatchuel, A., & Weil, B. (2009). C-K design theory: An advanced formulation. Research in Engineering Design, 19(4), 181–192.

Hatchuel, A., Le Masson, P., Reich, Y., & Weil, B. (2011). A Systematic Approach of Design Theories Using Generativeness and Robustness. In 11th International Conference on Engineering Design (pp. 1–12).

Hatchuel, A., Reich, Y., Le Masson, P., Weil, B., & Kazakçi, A. (2013). Beyond models and decisions: Situating design through generative functions. In DS 75-2: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol.2: Design Theory and Research Methodology, Seoul, Korea, 19-22.08.2013 (pp. 233–242).

Hawley, S. A. (2012). Abundances in "Green Pea" Star-forming Galaxies. Publications of the Astronomical Society of the Pacific, 124(911), 21–35.

He, Z. L., Geng, X. S., & Campbell-Hunt, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. Research Policy, 38(2), 306–317.

Herbsleb, J. D., Mockus, A., Finholt, T. A., & Grinter, R. E. (2004). Distance, dependencies, and delay in a global collaboration (pp. 319–328).

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intesive Scientific Discovery. E-Science and Information Management, 1–1.

Hinds, P. J., & Mortensen, M. (2005). Understanding Conflict in Geographically Distributed Teams: The Moderating Effects of Shared Identity, Shared Context, and Spontaneous Communication. Organization Science, 16(3), 290–307.

Hippel, E. Von, & Krogh, G. Von. (2003). open Source Software and the " Private-Collective " Innovation Model : Issues for Organization Science, (2), 209–224.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569–72.

Hobday, M. (2002). The project-based organisation: an ideal form for managing complex products and systems? Research Policy, 29(7–8), 871–893.

Hossain, M., & Kauranen, I. (2015). Crowdsourcing: a comprehensive literature review. Strategic Outsourcing: An International Journal, 8(1), 2–22.

Houllier, F. (2016). Les sciences participatives en France: Etat des lieux, bonnes pratiques et recommandations. Les Sciences Participatives En France(2016), 63.

Howe, J. (2006). The Rise of Crowdsourcing. Wired Magazine, 14(06), 1–5. http://doi.org/10.1086/599595

Hsieh, C., Nickerson, J. A., & Zenger, T. R. (2007). Opportunity discovery, problem solving and a theory of the entrepreneurial firm. Journal of Management Studies.

Hughes, J. (2008). William Kay, Samuel Devons and memories of practice in Rutherford's Manchester laboratory. Notes and Records of the Royal Society, 62(1), 97–121.

Ioannidis, J. P. A. (2016, February 1). Anticipating consequences of sharing raw data and code and of awarding badges for sharing. Journal of Clinical Epidemiology. Elsevier.

Irwin, A. (2006). The politics of talk: Coming to terms with the "new" scientific governance. Social Studies of Science, 36(2), 299–320.

Israel, B. A., Schulz, A. J., Parker, E. A., & Becker, A. B. (1998). REVIEW OF COMMUNITY-BASED RESEARCH: Assessing Partnership Approaches to Improve Public Health. Annual Review of Public Health, 19(1), 173–202.

Izotov, Y. I., Guseva, N. G., & Thuan, T. X. (2011). Green pea galaxies and cohorts: Luminous compact emission-line galaxies in the Sloan Digital Sky Survey. Astrophysical Journal, 728(2), 161.

Jackson, S. E., & Bantel, K. A. . (1989). Top management and innovations in banking□: Does the composition of the top team make a difference ? Strategic Management Journal, 10(S1), 107–124.

Jaime, A., Gardoni, M., Mosca, J., & Vinck, D. (2005). BASIC Lab: A software tool for supporting the production of knowledge in research organizations through the management of scientific concepts. Journal of Knowledge Management, 9(6), 53–66.

Janzik, L., & Herstatt, C. (2008). Innovation communities: Motivation and incentives for community members to contribute. Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology, ICMIT, 350–355.

Jeppesen, L. B., & Frederiksen, L. (2006). Why Do Users Contribute to Firm-Hosted User Communities? The Case of Computer-Controlled Music Instruments. Organization Science, 17(1), 45–63.

Jeppesen, L. B., & Lakhani, K. R. (2010). Marginality and Problem-Solving Effectiveness in Broadcast Search. Organization Science, 21(5), 1016–1033.

Jirotka, M., Lee, C. P., & Olson, G. M. (2013). Supporting scientific collaboration: Methods, tools and concepts. Computer Supported Cooperative Work: CSCW: An International Journal, 22(4–6), 667–715.

Jouvenet, M. (2009). La culture du «bricolage» instrumental et l'organisation du travail scientifique enquête dans un centre de recherche en nanosciences. Revue d'anthropologie Des Connaissances, 1, 2(2), 189.

Kalinichenko, L., Kovalev, D., Kovaleva, D., & Malkov, O. (2015). METHODS AND TOOLS FOR HYPOTHESIS-DRIVEN RESEARCH SUPPORT: A SURVEY *, 9.

Kaplan, N. (1959). The Role of the Research Administrator. Administrative Science Quarterly, 4(1), 20–42.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? Research Policy, 26(1), 1–18.

Kazakçi, A. O. (2013). On the imaginative constructivist nature of design: A theoretical approach. Research in Engineering Design, 24(2), 127–145.

Kazakçi, A. O. (2014). Conceptive artificial intelligence: Insights from design theory. In Proceedings of International Design Conference, DESIGN (Vol. 2014–Janua, pp. 33–48).

Kazakçi, A. O. (2015). Data science as a new frontier for design. In International Conference on Engineering Design (Ed.), (pp. 1–10). Milan, Italy.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. BioScience, 59(7), 613–620.

Keegan, A., & Turner, J. R. (2001). Quantity versus Quality in Project-based Learning Practices. Management Learning.

Kégl, B., Boucaud, A., Cherti, M., Kazakçı, A., Gramfort, A., Lemaitre, G., … Marini, C. (2018). The RAMP framework: from reproducibility to transparency in the design and optimization of scientific workflows. Stockholm: International Conference on Machine Learning.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. BioScience, 59(7), 613–620.

Khatib, F., Dimaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., … Baker, D. (2010). Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nature Structural and Molecular Biology, 18(10), 1175–1177.

Kieser, A. (1994). Why Organization Theory Needs Historical Analyses- Should Be And How This Performed. Organization Science, 5(4), 608–620.

Kiesler, S., & Cummings, J. N. (2002). What Do We Know about Proximity and Distance in Work Groups? A Legacy of Research 1 Sara Kiesler and Jonathon N. Cummings. Ditributed Work (Book), 76–109.

King, A. A., & Lakhani, K. R. (2012). The Contingent Effect of Absorptive Capacity: An Open Innovation Analysis. SSRN Electronic Journal.

King, A., & Lakhani, K. (2013). Using Open Innovation to Identify the Best Ideas. Sloanreview.Mit.Edu, 50(55121), 69–76.

King, R. D., Rowland, J., Aubrey, W., Liakata, M., Markham, M., Soldatova, L. N., … Pir, P. (2009). The robot scientist adam. Computer, 42(8), 46–54.

Kitch, E. W. (1977). The Nature and Function of the Patent System. The Journal of Law and Economics, 20(2), 265–290.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 205395171452848.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 205395171663113.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08 (p. 453). New York, New York, USA: ACM Press.

Kittur, A., Kraut, R. E., & Kraut, R. E. (2008). Harnessing the Wisdom of Crowds in Wikipedia□: Quality through Coordination Harnessing the Wisdom of Crowds in Wikipedia□:

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12(1), 1–48.

Klahr, D., & Simon, H. A. (1999). Studies of Scientific Creativity: Complimentary approaches and convergent findings. Psychological Bulletin, 125(5), 524–543.

Kokshagina, O., Le Masson, P., Weil, B., & Cogez, P. (2012). Risk Management strategies in a highly uncertain environment: understanding the role of common unknown. In 19th International Product Development Management Conference (pp. 1–26).

Kokshagina, O., Gillier, T., Cogez, P., Le Masson, P., & Weil, B. (2014). Towards a new form of ideas contests in high-tech environment: design community building. Academy of Management Proceedings, 2014(1), 16743.

König, B., Diehl, K., Tscherning, K., & Helming, K. (2013). A framework for structuring interdisciplinary research management, 42, 261–272.

Kouzes, R. T., Myers, J. D., & Wulf, W. A. (1996). Collaboratories: Doing science on the internet. Computer, 29(8), 40–46.

Kuhn, T. S., & Ian, H. (1962). The structure of scientific revolutions.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. Cognitive Science, 12(2), 139–175.

Laboulaye, C., & Babbage, C. (2016). Économie des machines et des manufactures. BnF-P.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032–2033.

Lakhani, K. R., & Von Hippel, E. (2003). How open source software works: "free" user-to-user assistance. Research Policy.

Lallé, B. (2004). Production de la connaissance et de l'action en sciences de gestion. Le statut expérimenté de «□chercheur-acteur. Revue Française de Gestion, 30(148), 45–65.

Laney, D. (2001). Application Delivery Strategies. Meta Group, (September).

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). Scientific discovery: computational explorations of the creative processes. MIT Press.

Largeaut, J. (1993). La logique. (PUF, Ed.). Paris: Que sais-je?

Latour, B., & Woolgar, S. (1988). La Vie de laboratoire. La production des faits scientifiques. La Découverte.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. Science.

Le Bon, G. (1908). The crowd: A study of the popular mind. Dover Publications.

Le Masson, P., & Weil, B. (2014). Réinventer l'entreprise□: la gestion collégiale des inconnus communs non appropriables. In L'entreprise, point aveugle du savoir.

Le Moigne, J. L. (1995). Les épistémologies constructivistes. PUF.

Lechopier, N. (2010). Recherche et non-recherche. Les valeurs à l'oeuvre dans l'évaluation des protocoles épidémiologiques. Revue d'Epidemiologie et de Sante Publique, 58(1), 41–48. http://doi.org/10.1016/j.respe.2009.09.006

Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., … Das, R. (2014). RNA design rules from a massive open laboratory. Proceedings of the National Academy of Sciences, 111(6), 2122–2127.

Leifert, C., & Woodward, S. (2013). Laboratory Contamination Management; the Requirement for Microbiological Quality Assurance. In Plant Cell, Tissue and Organ Culture (Vol. 52, pp. 237–244).

Levine, S. S., & Prietula, M. J. (2012). Open source, open innovation, open communities: What drives the performance of "open"? In Academy of Management 2012 Annual Meeting, AOM 2012.

Levine, S. S., Prietula, M. J., & Levine, S. S. (2014). Performance Open Collaboration for Innovation□: Principles and Performance, (February 2016).

Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. Research Policy, 42(9), 1515–1530. http://doi.org/10.1016/j.respol.2013.06.012

Liberatore, M. J., & Titus, G. J. (1983). The Practice of Management Science in R & D Project Management. Management Science, 29(8), 962–974.

Lichten, C., Ioppolo, R., D', C., Rebecca, A., Simmons, K., & Jones, M. M. (2018). Citizen science: crowdsourcing for research.

Licoppe, C. (1996). La Formation de la pratique scientifique□: le discours de l'expériènce en France et en Angleterre□: 1630-1820. La Découverte.

Lin, M., & Lucas, H. C. (2013). Too Big to Fail : Large Samples and the p -Value Problem. Information Systems Research, 7047(August 2016), 1–12.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). DENDRAL: A case study of the first expert system for scientific hypothesis formation. Artificial Intelligence, 61(2), 209–261.

Loch, C. H., Terwiesch, C., & Thomke, S. (2001). Parallel and Sequential Testing of Design Alternatives. Management Science, 47(5), 663–678.

Loch, C. H., Solt, M. E., & Bailey, E. M. (2008). Diagnosing unforeseeable uncertainty in a new venture. In Journal of Product Innovation Management (Vol. 25, pp. 28–46).

Lotka, A. J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16, 317–323.

Lourenço, S., Gunge, V. B., Andersson, T. M.-L., Andersen, C. L. E., Lund, A. S. Q., Køster, B., & Hansen, G. L. (2018). Avoidable colorectal cancer cases in Denmark – The impact of red and processed meat. Cancer Epidemiology, 55, 1–7.

Louvel, S. (2011). Des patrons aux managers□: Les laboratoires de la recherche publique depuis les années 1970. Presses universitaires de Rennes.

Louwman, P. (2004). Christiaan Huygens and his telescopes. In European Space Agency, (Special Publication) ESA SP (Vol. 1278, pp. 103–114).

Lundin, R. A., & Söderholm, A. (1995). A theory of the temporary organization. Scandinavian Journal of Management, 11(4), 437–455.

Lüttgens, D., Pollok, P., Antons, D., & Piller, F. (2014). Wisdom of the crowd and capabilities of a few: internal success factors of crowdsourcing for innovation. Journal of Business Economics, 84(3), 339–374.

MacLean, D., MacIntosh, R., & Grant, S. (2002). Mode 2 management research. British Journal of Management, 13(3), 189–207.

Madsen, T. L., Woolley, J., & Sarangee, K. (2012). Using Internet-based Collaboration Technologies for Innovation: Crowdsourcing vs. Expertsourcing. Academy of Management Proceedings, 2012(1), 14965.

Majchrzak, A., & Malhotra, A. (2013). Journal of Strategic Information Systems Towards an information systems perspective and research agenda on crowdsourcing for innovation. Journal of Strategic Information Systems, 22(4), 257–268.

Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The Collective Intelligence Genome THE LEADING. MIT Sloan Management Review, 51(51303), 21–31.

Mao, A., Kamar, E., & Horvitz, E. (2013). Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing. First Conference on Human Computation (HCOMP 2013), 103–111.

Marples, D. L. (1961). The Decisions of Engineering Design. IRE Transactions on Engineering Management.

Matthews, J. (2016). History of Biostatistics. Http://Journal.Emwa.Org/, 25(3), 8–11.

Matthias, S. (2018). Epistemology. In The Stanford Encyclopedia of Philosophy (Edward N.).

McFadyen, M. A., Jr, A. A. C., The, S., Journal, M., Oct, N., Mcfadyen, M. A. N. N., & Cannella, A. A. (2004). Social Capital and Knowledge Creation : Diminishing Returns of the Number and Strength of Exchange. Academy of Management Journal, 47(5), 735–746.

McLeod, P. L., Lobel, S. A., & Cox, T. H. (2007). Ethnic Diversity and Creativity in Small Groups. Small Group Research, 27(2), 248–264.

Mello, A. L., & Rentsch, J. R. (2015). Cognitive Diversity in Teams. Small Group Research, 46(6), 623–658.

Menger, V., Spruit, M., Hagoort, K., & Scheepers, F. (2016). Transitioning to a Data Driven Mental Health Practice: Collaborative Expert Sessions for Knowledge and Hypothesis Finding. Computational and Mathematical Methods in Medicine, 2016, 1–11.

Merton, R. K. (1942). Science and Technology in a Democratic Order. Journal of Legal and Political Sociology, 115–126.

Merton, R. K. (1957). Priorities in Scientific Discovery: A Chapter in the Sociology of Science. American Sociological Review, 22(6), 635.

Merton, R. K., & Storer, N. W. (1973). The Sociology of Science: Theoretical and Empirical Investigations. Contemporary Sociology, 5(5), 557.

Merton, R. K., & Barber, E. G. (2004). The travels and adventures of serendipity□: a study in sociological semantics and the sociology of science. Princeton University Press.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? Journal of Regional Science, 50(1), 181–201.

Milliken, F. J., Bartel, C. A., & Kurtzberg, T. R. (1997). Diversity and Creativity in Work Groups A Dynamic Perspective on. Creativity, 346.

Mjolsness, E., & DeCoste, D. (2001, September 14). Machine learning for science: State of the art and future prospects. Science. American Association for the Advancement of Science.

Mollick, E., & Nanda, R. (2016). Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. Management Science, 62(6), 1533–1553.

Molloy, J. C. (2011). The open knowledge foundation: Open data means better science. PLoS Biology, 9(12), e1001195.

Moon, J. Y., & Sproull, L. S. (2008). The role of feedback in managing the internet-based volunteer work force. Information Systems Research, 19(4), 494–515.

Morus, I. T. (2016). Invisible Technicians, Instrument Makers and Artisans. In B. V. Lightman (Ed.), A companion to the history of science (First, pp. 97–109). Oxford: Wiley.

Murray, F., & O'Mahony, S. (2007). Exploring the Foundations of Cumulative Innovation: Implications for Organization Science. Organization Science, 18(6), 1006–1021.

Næser, E., Møller, H., Fredberg, U., & Vedsted, P. (2018). Mortality of patients examined at a diagnostic centre: A matched cohort study. Cancer Epidemiology, 55, 130–135.

Nelson, R. R. (2008). Bounded rationality , cognitive maps , and trial and error learning, 67(December 2005), 78–89.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). The processes of creative thinking. In Contemporary approaches to creative thinking: A symposium held at the University of Colorado. (pp. 63–119). RAND Corporation.

Nielsen, M. (2011). Reinventing Discovery: The New Era of Networked Science. Portal□: Libraries and the Academy, 13(2), 214–216.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015, June 26). Promoting an open research culture. Science. NIH Public Access.

Olsen, T., & Carmel, E. (2013). The process of atomization of business tasks for crowdsourcing. Strategic Outsourcing: An International Journal, 6(3), 3–7.

Olson, G. M., & Olson, J. S. (2000). Distance matters. Human-Computer Interaction, 15(2), 139–178.

Partha, D., & David, P. A. (1994). Toward a new economics of science. Research Policy, 23(5), 487–521.

Pearce, J. M. (2016). Return on investment for open source scientific hardware development. Science and Public Policy, 43(2), 192–195.

Pelz, D. C. (1960). Interaction and Attitudes between Scientists and the Auxiliary Staff: II. Viewpoint of Scientists. Administrative Science Quarterly, 4(4), 410.

Perkmann, M., & Schildt, H. (2015). Open data partnerships between firms and universities□: The role of boundary organizations. Research Policy, 44(5), 1133–1143.

Penrose, E. (1959). The theory of the growth of the firm. New York: Oxford University Press.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V, … Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. British Journal of Cancer, 34(6), 585–612.

Phillips, K. W., Mannix, E. A., Neale, M. A., & Gruenfeld, D. H. (2004). Diverse groups and information

sharing: The effects of congruent ties. Journal of Experimental Social Psychology, 40(4), 497–510.

Piezunka, H., & Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. Academy of Management Journal, 58(3), 856–880.

Pilyugin, L. S., Vílchez, J. M., Mattsson, L., & Thuan, T. X. (2012). Abundance determination from global emission-line SDSS spectra: Exploring objects with high N/O ratios. Monthly Notices of the Royal Astronomical Society, 421(2), 1624–1634.

Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? Journal of Product Innovation Management, 29(2), 245–256.

Poitou, J. P. (1982). Prony et babbage: apercus sur l'histoire de la division du travail mental. History of European Ideas, 3(3), 295–302.

Polanyi, M. (1966). The tacit dimension.

Poole, D., Mackworth, A., & Goebel, R. (1998). Computational Intelligence: A Logical Approach. Book.

Popper, K. (1959). The logic of scientific discovery. Journal of the Franklin Institute, 268(3), 244.

Porac, J. F., Wade, J. B., Fischer, H. M., Brown, J., Kanfer, A., & Bowker, G. (2004). Human capital heterogeneity, collaborative relationships, and publication patterns in a multidisciplinary scientific alliance: a comparative case study of two scientific teams. Research Policy, 33(4), 661–678.

Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology. Administrative Science Quarterly, 41(1), 116.

Price, D. J. de S. (Derek J. de S., & Tukey, J. W. (John W. (1963). Little science, big science. Columbia Univ. Press.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1(1), 51–59.

Racunas, S. A., Shah, N. H., Albert, I., & Fedoroff, N. V. (2004). HyBrow: A prototype system for computer-aided hypothesis evaluation. In Bioinformatics (Vol. 20, pp. 1–8).

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Cardamone, C., Murray, P., … Vandenberg, J. (2013). Galaxy Zoo: Motivations of Citizen Scientists.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3.

Raymond, E. (1999). The cathedral and the bazaar. Knowledge, Technology & Policy, 12(3), 23–49.

Reed, J., Raddick, M. J., Lardner, A., & Carney, K. (2013). An exploratory factor analysis of motivations for participating in Zooniverse, a collection of virtual citizen science projects. In Proceedings of the Annual Hawaii International Conference on System Sciences (pp. 610–619).

Renault, S. (2014). Comment orchestrer la participation de la foule à une activité de crowdsourcing□? La taxonomie des 4 C. Systèmes d'information & Management, 19(1), 77.

Richardson, G. B. (1972). The Organisation of Industry. The Economic Journal.

Riedl, C., & Woolley, A. W. (2017). Teams vs. Crowds: A Field Test of the Relative Contribution of Incentives, Member Ability, and Emergent Collaboration to Crowd-Based Problem Solving Performance. Academy of Management Discoveries, 3(4), 382–403.

Rosen, L. (2004). Open source licensing: Software freedom and intellectual property law. Cited May.

Rosenkopf, L., & Almeida, P. (2003). Overcoming Local Search Through Alliances and Mobility. Management Science, 49(6), 751–766.

Rotman, D., Hammock, J., Preece, J., Hansen, D., Boston, C., Bowser, A., & He, Y. (2014). Motivations Affecting Initial and Long-Term Participation in Citizen Science Projects in Three Countries. In iConference 2014 Proceedings. iSchools.

Russell, S. J., & Norvig, P. (2007). Artificial intelligence a modern approach. (Pearson, Ed.).

Russom, P. (2011). Introduction to Big Data Analytics, 38.

Sarker, S., Chiang, R., & Abbasi, A. (2018). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. Journal of the Association for Information Systems, 17(2), I–XXXII.

Sauermann, H., & Franzoni, C. (2014). Crowd science user contribution patterns and their implications.

Schaffer, S. (1988). Astronomers Mark Time: Discipline and the Personal Equation. Science in Context, 2(1), 115–145.

Schaffer, S. (1994). Babbage's Intelligence: Calculating Engines and the Factory System. Critical Inquiry, 21(1), 203–227.

Scholz, T. (2013). Digital labor□: the Internet as playground and factory. Routledge.

Schunn, C. D., & Klahr, D. (1992). A 4-space model of scientific discovery. In Proceedings of the Fourteenth Annual Cognitive Science Society Conference (pp. 106–111).

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., … Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS International Journal of Geo-Information, 5(5), 55.

Sen, S. (2010, January 1). Automatic Effective Model Discovery. http://www.theses.fr. Rennes 1.

Shapin, S. (1989). The Invisible Technician. American Scientist, 77(6), 554–563.

Shapin, S. (2008). The scientific life: A moral history of a late modern vocation. Science Education.

Shapiro, S. (1996). The limits of logic : higher-order logic and the Löwenheim-Skolem theorem. The international research library of philosophy. Dartmouth.

Shaposhnikova Tatyana, T. G. M. V. (2005). Jacques Hadamard: Un mathématicien universel. EDP Sciences.

Shelly, M. A., By, E., & Birnbaum, D. (1996). Exploratory Data Analysis: Data Visualization or Torture? Statistics for Hospital Epidemiolog Exploratory Data Analysis: Data Visualization or Torture? Source: Infection Control and Hospital Epidemiology INFECTION CONTROL AND HOSPITAL EPIDEMIOLOGY, 17(605), 605–612.

Shinn, T. (1980). Division du savoir et spécificité organisationnelle□: Les laboratoires de recherche industrielle en France. Revue Française de Sociologie, 21(1), 3. http://doi.org/10.2307/3320898

Shmueli, G. (2011). To Explain or to Predict? Statistical Science, 25(3), 289–310.

Sieg, J. H., Wallin, M. W., & von Krogh, G. (2010). Managerial challenges in open innovation: A study of innovation intermediation in the chemical industry. R and D Management, 40(3), 281–291.

Siegel, E. (2013). Predictive analytics : the power to predict who will click, buy, lie, or die.

Sigglekow, N. J. (2007). PERSUASION WITH CASE STUDIES. Academy of Management Journal, 50(1), 20–24.

Silvertown, J. (2009). A new dawn for citizen science. Trends in Ecology and Evolution, 24(9), 467–471.

Simpson, R. J., Povich, M. S., Kendrew, S., Lintott, C. J., Bressert, E., Arvidsson, K., … Wolf-Chase, G. (2012). The Milky Way Project First Data Release: A bubblier Galactic disc. Monthly Notices of the Royal Astronomical Society, 424(4), 2442–2460.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. The Quarterly Journal of Economics, 69(1), 99–118.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. American Psychologist, 26(2), 145–159.

Simon, H. A. (1973). The structure of ill structred problems. Artificial Intelligence, (4), 181–202.

Sitruk, Y., & Kazakçi, A., (2019). Pilotage de la performance des projets de science citoyenne pour la génération d'hypothèses scientifiques data-driven : le principe de capitalisation séquentielle. Papier soumis pour publication.

Sitruk, Y., & Kazakçi, A. (2018). Crowd-based data-driven hypothesis generation from data and the organisation of participative scientific process. In Proceedings of International Design Conference, DESIGN (Vol. 4, pp. 1673–1684).

Sivasubramanian, R., Selladurai, V., & Gunasekaran, A. (2003). Utilization of bottleneck resources for profitability through a synchronized operation of marketing and manufacturing. Integrated Manufacturing Systems, 14(3), 238–246.

Smith, A. M., Lynn, S., Sullivan, M., Lintott, C. J., Nugent, P. E., Botyanszki, J., … Walters, R. (2011). Galaxy Zoo Supernovae. Monthly Notices of the Royal Astronomical Society, 412(2), 1309–1319.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. ArXiv E-Prints.

Soldatova, L. N., & Rzhetsky, A. (2011). Representation of research hypotheses. Journal of Biomedical Semantics, 2(2), S9.

Sonnenwald, D. H. (2007). Scientific Collaboration : A Synthesis of Challenges and Strategies. Annual Review of Information Science and Technology, 4(January 2006), 2–37.

Sorrenson, R. (2013). Perfect Mechanics: Instrument Makers at the Royal Society of London in the Eighteenth Century. (B. D. Press, Ed.)Isis.

Spithoven, A., Clarysse, B., & Knockaert, M. (2009). Building Absorptive Capacity to Organise Inbound Open Innovation in Low Tech Industries.

Stephan, P. E. (1996). The economics of science. In Handbook of the Economics of Innovation.

Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty before 1900. Technology and Culture (Vol. 29). Belknap Press of Harvard University Press.

Stokols, D., Harvey, R., Gress, J., Fuqua, J., & Phillips, K. (2005). In vivo studies of transdisciplinary scientific collaboration: Lessons learned and implications for active living research. In American Journal of Preventive Medicine (Vol. 28, pp. 202–213).

Strevens, M. (2006). Scientific explanation. In Encyclopedia of Philosophy. Stanford University.

Surowiecki. (2004). The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Choice Reviews Online, 42(03), 42-1645-42–1645.

Swan, M., Hathaway, K., Hogg, C., McCauley, R., & Vollrath, A. (2010). Citizen science genomics as a model

for crowdsourced preventive medicine research. J Participat Med, 2, e20.

Swan, M. (2013). The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. Big Data, 1(2), 85–99.

Saez-Rodriguez, J., Costello, J. C., Friend, S. H., Kellen, M. R., Mangravite, L., Meyer, P., … Stolovitzky, G. (2016, August 1). Crowdsourcing biomedical research: Leveraging communities as innovation engines. Nature Reviews Genetics.

Salgueiredo, C. F., & Hatchuel, A. (2016). Beyond analogy: A model of bioinspiration for creative design. Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM, 30(2), 159–170.

Sauermann, H., & Franzoni, C. (2014). Crowd science user contribution patterns and their implications.

Sauermann, H., & Roach, M. (2012). Science PhD career preferences: Levels, changes, and advisor encouragement. PLoS ONE, 7(5), e36307.

Schaffer, S. (1988). Astronomers Mark Time: Discipline and the Personal Equation. Science in Context, 2(1), 115–145.

Schlagwein, D., & Bjorn-Andersen, N. (2018). Organizational Learning with Crowdsourcing: The Revelatory Case of LEGO. Journal of the Association for Information Systems, 15(11), 754–778.

Schemmann, B., Herrmann, A. M., Chappin, M. M. H., & Heimeriks, G. J. (2016). Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development. Research Policy, 45(6), 1145–1154.

Schlagwein, D., & Bjorn-Andersen, N. (2018). Organizational Learning with Crowdsourcing: The Revelatory Case of LEGO. Journal of the Association for Information Systems, 15(11), 754–778.

Schoenfeld, J. D., & Ioannidis, J. P. (2013). Is everything we eat associated with cancer? A systematic cookbook review. The American Journal of Clinical Nutrition, 97(1), 127–134.

Schulze, A., & Hoegl, M. (2008). Organizational knowledge creation and the generation of new product ideas: A behavioral approach. Research Policy, 37(10), 1742–1750.

Schunn, C. D., & Klahr, D. (1992). A 4-space model of scientific discovery. In Proceedings of the Fourteenth Annual Cognitive Science Society Conference (pp. 106–111).

Schweitzer, F. M., Buchinger, W., Gassmann, O., & Obrist, M. (2012). Crowdsourcing: Leveraging Innovation through Online Idea Competitions. Research-Technology Management, 55(3), 32–38.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., … Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS International Journal of Geo-Information, 5(5), 55.

Sen, S. (2010, January 1). Automatic Effective Model Discovery. http://www.theses.fr. Rennes 1.

Shankar, R., Mittal, N., Rabinowitz, S., Baveja, A., & Acharia, S. (2013). A collaborative framework to minimise knowledge loss in new product development. International Journal of Production Research, 51(7), 2049–2059.

Shapin, S. (1989). The Invisible Technician. American Scientist, 77(6), 554–563.

Shapin, S. (2008). The scientific life: A moral history of a late modern vocation. Science Education.

Shinn, T., & Ragouet, P. (2000). Formes de division du travail scientifique et convergence intellectuelle . La recherche technico-instrumentale. Revue Française de Sociologie, 41(3), 447–473.

Shmueli, G. (2011). To Explain or to Predict? Statistical Science, 25(3), 289–310.

Sieg, J. H., Wallin, M. W., & von Krogh, G. (2010). Managerial challenges in open innovation: A study of innovation intermediation in the chemical industry. R and D Management, 40(3), 281–291.

Simon, H. A., & Simon, P. A. (1962). Trial and error search in solving difficult problems: Evidence from the game of chess. Behavioral Science, 7(4), 425–429.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. Synthese, 47(1), 1–27.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. American Psychologist, 26(2), 145–159.

Simon, H. A. (1973). The structure of ill structred problems. Artificial Intelligence, (4), 181–202.

Sivasubramanian, R., Selladurai, V., & Gunasekaran, A. (2003). Utilization of bottleneck resources for profitability through a synchronized operation of marketing and manufacturing. Integrated Manufacturing Systems, 14(3), 238–246.

Smith, K. (2009). The wisdom of crowds. Nature Reports Climate Change, (0908), 89–91.

Spithoven, A., Clarysse, B., & Knockaert, M. (2009). Building Absorptive Capacity to Organise Inbound Open Innovation in Low Tech Industries.

Stokols, D., Harvey, R., Gress, J., Fuqua, J., & Phillips, K. (2005). In vivo studies of transdisciplinary scientific collaboration: Lessons learned and implications for active living research. In American Journal of Preventive Medicine (Vol. 28, pp. 202–213).

Surowiecki. (2004). The wisdom of crowds: why the many are smarter than the few and how collective

wisdom shapes business, economies, societies and nations. Choice Reviews Online, 42(03), 42-1645-42–1645.

Swan, M. (2013). The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. Big Data, 1(2), 85–99. http://doi.org/10.1089/big.2012.0002

Sybrandt, J., Shtutman, M., & Safro, I. (2017). MOLIERE: Automatic Biomedical Hypothesis Generation System.

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., Kö Nig, I. R., Zhang, H., & Sun, Y. V. (2009). Machine Learning in Genome-Wide Association Studies. Genetic Epidemiology, 33, 51–57.

Talibov, M., Sormunen, J., Hansen, J., Kjaerheim, K., Martinsen, J. I., Sparen, P., … Pukkala, E. (2018). Benzene exposure at workplace and risk of colorectal cancer in four Nordic countries. Cancer Epidemiology, 55, 156–161.

Tapscott, D., & Williams, A. D. (2007). Innovation in the Age of Mass Collaboration. New York.

Tapscott, D., & Williams, A. D. (2006). Wikinomics: How mass collaboration changes everything. Portfolio.

Terwiesch, C., & Xu, Y. (2008). Innovation Contests, Open Innovation, and Multiagent Problem Solving. Management Science, 54(9), 1529–1543.

Thagard, P. (1998). Ulcers and bacteria I: Discovery and acceptance. Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences, 29(1), 107–136.

Theisz, K. (2017). Crowdsourcing and Citizen Science: Investigating Data Quality and Utility.

Thomas, D., Raddick, M. J., Willett, K. W., Skibba, R. A., Casteels, K. R. V., Kaviraj, S., … Edmondson, E. M. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 435(4), 2835–2860.

Thomke, S. H. (2003). Experimentation Matters: Unlocking the Potential of New Technologies for Innovation. Journal of Engineering and Technology Management.

Thompson, N., & Hanley, D. (2017). Science Is Shaped by Wikipedia: Evidence From a Randomized Control Trial. SSRN.

Tillas, A. (2012). Social Perception. In Encyclopedia of Philosophy and the Social Sciences (p. 621).

Torrens, H. (1995, September 5). Mary anning (1799–1847) of lyme; 'the greatest fossilist the world ever knew.' The British Journal for the History of Science, 28(3), 257–284.

Tran, N., Baral, C., Nagaraj, V. J., & Joshi, L. (2005). Knowledge-based integrative framework for hypothesis formation in biochemical networks. In Data Integration in the Life Sciences (pp. 121–136).

Trumbull, D. J., Bonney, R., Bascom, D., & Cabral, A. (2000). Thinking scientifically during participation in a citizen-science project. Science Education, 84(2), 265–275.

Tucci, C. L., Afuah, A., & Viscusi, G. (2018). Creating and capturing value through crowdsourcing.

Tuertscher, P. R., Garud, R., & Nordberg, M. (2008). The Emergence of Architecture: Coordination Across Boundaries at ATLAS, CERN. Academy of Management Annual Meeting Proceedings, 42.

van Andel, P. (1994). Anatomy of the unsought finding. Serendipity: Orgin, history, domains, traditions, appearances, patterns and programmability. British Journal for the Philosophy of Science.

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research (Vol. 9).

van der Velde, T., Davis, G., Perkins, G., Lawson, T. J., Wilcox, C., Lansdell, M., … Hardesty, B. D. (2016). Comparison of marine debris data collected by researchers and citizen scientists: Is citizen science data worth the effort? Biological Conservation, 208, 127–138.

Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. . (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Vinck, D. (2006). L'équipement du chercheur□: comme si la technique était déterminante. Ethnographiques.Org, (Numéro 9-février 2006).

von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. Research Policy, 32(7), 1217–1241.

von Hippel, E., & Tyre, M. J. (1996). How learning by doing is done: problem identification in novel process equipment. Research Policy.

von Hippel, E., & von Krogh, G. (2015). Identifying Viable "Need–Solution Pairs": Problem Solving Without Problem Formulation. Organization Science, orsc.2015.1023.

von Zedtwitz, M., Gassmann, O., & Boutellier, R. (2004). Organizing global R&D: challenges and dilemmas. Journal of International Management, 10(1), 21–49.

Wales, J. (2005). Internet encyclopaedias go head to head. Nature.

Walsh, J. P., & Lee, Y. N. (2015). The bureaucratization of science. Research Policy, 44(8), 1584–1600.

Walster, E., Walster, G. W., & Berscheid, E. (1978). Equity: Theory and Research. eweb:22046.

Warby, S. C., Wendt, S. L., Welinder, P., Munk, E. G. S., Carrillo, O., Sorensen, H. B. D., … Mignot, E. (2014).

Sleep-spindle detection: Crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nature Methods, 11(4), 385–392.

Wasko, M., Kudaravalli, S., & Faraj, S. (2017). Leading Collaboration in Online Communities. MIS Quarterly.

Weick, K. . (1979). The social psychology of organizing. Reading, Addison-Wesley (Vol. 2). Random House.

Weil, B. (1999, January 1). Conception collective, coordination et savoirs, les rationalisations de la conception automobile. Ingénierie et gestion. Paris, ENMP.

Weisband, S. (2002). Maintaining Awareness in Distributed Team Collaboration: Implications for Leadership and Performance. In Distributed Work (pp. 311–333). MIT Press.

West, J., & Sims, J. (2018). How Firms Leverage Crowds and Communities for Open Innovation. In Creating and Capturing Value through Crowdsourcing.

Whitley, R. (2000). The intellectual and social organization of the sciences. Oxford University Press.

Wiggins, A., & Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. In Proceedings of the Annual Hawaii International Conference on System Sciences (pp. 1–10). IEEE.

Wilderman, C. C. (2007). Citizen Science Toolkit Conference models of community science: design lessons from the field.

Wildschut, D. (2017). The need for citizen science in the transition to a sustainable peer-to-peer-society. Futures, 91, 46–52.

Winter, S. G. (2003). Understanding dynamic capabilities. Strategic Management Journal, 24(10 SPEC ISS.), 991–995.

Wooten, J. O., & Ulrich, K. T. (2017). Idea Generation and the Role of Feedback: Evidence from Field Experiments with Innovation Tournaments. Production and Operations Management, 26(1), 80–99.

Yadav, D., & Lowenfels, A. B. (2013, May). The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: a coauthorship network analysis. Journal of the American Society for Information Science and Technology, 60(10), 2107–2118.

Yin, R. K. (2003). Case Study Research: Design and Methods. Sage Publications.

Zhang, C., & Ma, Y. (2012). Ensemble machine learning□: methods and applications. Springer.

Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. Information Systems Frontiers, 16(3), 417–434.

Zheng, H., Li, D., & Hou, W. (2011). Task Design, Motivation, and Participation in Crowdsourcing Contests. International Journal of Electronic Commerce, 15(4), 57–88.

Zollo, M., & Winter, S. G. (2003). Deliberate Learning and the Evolution of Dynamic Capabilities. Organization Science, 13(3), 339–351.

ANNEXES

ANNEXE 1 – ECHANTILLON D'EXEMPLES D'OUVERTURE DU RAISONNEMENT SCIENTIFIQUE

Nom du projet	Domaine	Objectif
American Gut http://humanfoodproject.com/americangut/	Microbiologie	Etudier l'impact des microbes sur la santé en soumettant un échantillon de votre peau, de votre bouche ou de votre intestin.
Autoimmune Citizen Science https://joinaics.com/	Médecine	Utiliser les patients pour étudier les maladies auto-immunes
Bee Lab https://www.beelab.umn.edu/	Apiculture	Utiliser l'*Open Design* pour améliorer la pratique de l'apiculture dans un environnement imprévisible
Christmas Bird Count https://www.audubon.org/conservation/science/christmas-bird-count	Biologie	Collecter des données sur la migration des oiseaux à utiliser dans le domaine des sciences, en particulier la biologie de la conservation
Cybelle Méditerranée http://www.cybelle-planete.org/cybelle-mediterranee-2/le-programme/description.html	Biodiversité	Collecter des données sur la faune marine grâce aux plongeurs
Disk Detective https://www.diskdetective.org/	Astronomie	Trier les images de satellite pour déterminer les disques de débris poussiéreux
DREAM Challenge http://dreamchallenges.org/	Biologie computationnelle	Développer de nouvelles méthodes d'utilisation des données pour répondre aux questions fondamentales afin de mieux comprendre et améliorer les sciences biologiques et la santé humaine.
Epidemium http://www.epidemium.cc/	Médecine, santé publique	Explorer de nouvelles voies de recherche sur le cancer à partir de bases de données ouvertes
EteRNA http://eterna.cmu.edu/	Biochimie	Modifier la forme de la combinaison de bases d'ARN pour s'adapter à une structure pliée (jeu)
EyeWire https://www.eyewire.org/	Neuroscience	Reproduire à partir des données scientifiques une carte du cerveau (jeu)
Foldit https://fold.it/portal/	Biochimie	Modifier un modèle visuel 3D d'une protéine pour optimiser sa forme et son repliement (jeu)
Galaxy Zoo https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/	Astronomie	Inspecter et classifier les images de galaxies
Gravity Spy https://www.zooniverse.org/projects/zooniverse/gravity-spy	Physique	Améliorer la classification pour détecter les ondes gravitationnelles
Higgs Boson Machine Learning Challenge https://www.kaggle.com/c/higgs-boson	Physique des particules	Développer de nouvelles méthodes d'utilisation des données pour valider la détection du boson de Higgs
LHC@Home http://lhcathome.web.cern.ch/	Physique des particules	Utiliser le temps d'inactivité des ordinateurs pour aider les physiciens à comparer la théorie avec l'expérience, dans la recherche de nouvelles particules fondamentales
Milky Way Project https://www.zooniverse.org/projec	Astronomie	Identifier des bulles de vent stellaire dans la Voie

ts/povich/milky-way-project		Lactée
Mozak https://www.mozak.science/landing	Neuroscience	Construire un modèle des cellules du cerveau
Nutri-Net Santé https://www.etude-nutrinet-sante.fr/	Santé	Etudier les relations nutrition-santé par une étude de cohorte
Old Weather https://www.oldweather.org	Météorologie historique	Aidez les scientifiques à récupérer les observations météorologiques de l'Arctique et du monde enregistrées dans les journaux de bord depuis le milieu du XIXe siècle.
Polymath 1 à 14 https://polymathprojects.org/	Mathématique	Résoudre collectivement des problèmes qui ont longtemps échappé aux approches traditionnelles des mathématiques.
Planet Hunters https://www.planethunters.org/	Astronomie	Découvrir de nouvelles planètes à partir des données issues du vaisseau Kepler
Quantum Moves https://www.scienceathome.org/games/quantum-moves/	Physique quantique	Trouver des solutions à des problèmes qui sont des simulations d'opération logique dans un ordinateur quantique (jeu)
RAMP (Rapid Analytics and Model Prototyping) https://ramp.studio/	Divers	Développer de nouvelles méthodes d'utilisation des données pour répondre aux questions fondamentales dans différents domaines scientifiques
Scribes of the Cairo Geniza https://www.zooniverse.org/projects/judaicadh/scribes-of-the-cairo-geniza	Archéologie	Trier et transcrire les fragments de Cairo Geniza
SPIPOLL http://www.spipoll.org/	Biodiversité	Collecter des données quantitatives sur les insectes pollinisateurs et/ou floricoles sur l'ensemble de la France métropolitaine.
Stall Catchers https://stallcatchers.com/main	Médecine	Identifier les zones du cerveau avec des vaisseaux sanguins bouchés pour aider la recherche sur Alzheimer
Steller Watch https://www.zooniverse.org/projects/sweenkl/steller-watch	Biodiversité	Comprendre pourquoi l'otarie de Steller, une espèce en voie de disparition, continue de décliner dans les îles Aléoutiennes
The Plastic Tide https://www.zooniverse.org/projects/theplastictide/the-plastic-tide	Environnement	Entraîner l'algorithme de détection automatique de plastique sur les plages
Vigie-Nature Ecole https://www.vigienature-ecole.fr/	Biodiversité	Suivre la biodiversité ordinaire en partenariat avec des écoles
War Diary https://www.operationwardiary.org/	Histoire	Etudier les journaux intimes des soldats durant la première guerre mondiale

ANNEXE 2 : TROIS EXEMPLES D'OUVERTURE DU RAISONNEMENT SCIENTIFIQUE

Galaxy Zoo (issu de Nielsen 2011)

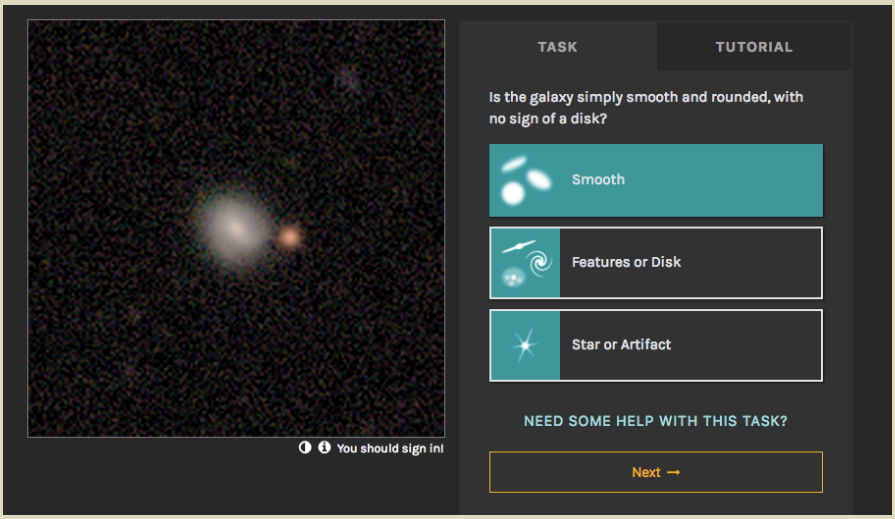


Figure . Interface d'utilisation de Galaxy Zoo

En 2007, Kevin Schawinski, alors doctorant en astronomie dans le laboratoire de Chris Lintott à l'Université d'Oxford, travaillait sur la compréhension de la formation des galaxies, et cherchait les galaxies elliptiques qui avaient formé les étoiles les plus récentes. Leurs premières idées étaient basées sur un échantillon limité de galaxies que Schawinski avait codé manuellement, cependant plus de données étaient nécessaires pour vraiment prouver leur point. Le groupe de chercheurs a alors eu l'idée d'utiliser 930 000 images de galaxies lointaines que le Sloan Digtal Sky Survey (SDSS) avait mis à disposition quelques mois plus tôt. Cependant, les ordinateurs ne sont pas particulièrement bons pour la détection d'images et, la grande quantité d'images ne pouvait être traité rapidement par quelques scientifiques. Les chercheurs ont donc créé la plateforme en ligne Galaxy Zoo dans laquelle les volontaires étaient incités à s'inscrire, lire un tutoriel, visualiser les images puis coder six propriétés différentes des objets astronomiques. La tâche était accessible tous sans compétence préalable requise et le volontaire était aidé avec un ensemble d'exemples. La participation est devenue rapidement virale, et sept mois après le lancement du projet environ 900 000 galaxies furent codées par plus de 250 000 volontaires uniques. Afin de réduire la probabilité d'un codage incorrect, les galaxies furent codées plusieurs fois par différents volontaires, pour un total d'environ 50 millions de classifications. A titre de comparaison, 50 millions de classifications auraient requis plus de 83 années à plein temps pour un scientifique seul. Les données récoltées par Galaxy Zoo ont permis à l'équipe de chercheurs de mener à bien l'étude initialement prévue.

Cependant, Galaxy Zoo a permis d'autres contributions, comme la découverte de nouveaux objets astronomiques, comme les « galaxies de pois verts » (Cardamone et al., 2009). Les données codées ont également été rendues publiques pour d'autres investigations et en septembre 2012, 141 articles scientifiques se basaient sur ces données. Après le succès du premier projet Galaxy Zoo, d'autres campagnes de codages de données furent lancées sur de nouvelles images astronomiques qui amenèrent à plus de 200 millions de classifications. Le succès de Galaxy Zoo a suscité de l'intérêt dans divers domaines scientifiques et a poussé l'équipe de chercheurs à une coopération avec d'autres institutions, la Citizen Science Alliance, pour exécuter un certain nombre de projets sur une plate-forme commune «The Zooniverse». Actuellement, la plateforme héberge 88 projets dans des domaines aussi variés que l'art, la biologie, le climat, ou encore l'histoire, la littérature et la médecine.

Dream Challenge project

Le projet « Dialogue for Reverse Engineering Assessments and Methods » (DREAM) a été officiellement lancé au printemps 2006, lorsqu'un groupe de biologistes, provenant à la fois de l'expérimentation et de la computation, s'est réuni à l'Académie des sciences de New York à Manhattan. Ces scientifiques partageaient une frustration commune sur les limites artificielles entre les sciences expérimentales et les sciences computationnelles dans la recherche biologique moderne. Suite aux discussions qu'ils ont mené durant la réunion, ils ont conclu que la recherche en biologie devrait concentrer ses efforts sur la compréhension des limites et des forces des multiples méthodes de rétro-ingénierie biologique. En effet, à cette époque, le domaine des réseaux biologiques d'ingénierie était en train de connaître une expansion considérable, ce qui engendrait beaucoup de confusion quant aux méthodes qui étaient vraiment utiles. Un problème clé est que ces méthodes computationnelles pouvaient, très rapidement, générer un grand nombre de prédictions - de quelques centaines à des centaines de milliers - dont la quasi-totalité n'étaient généralement pas testé expérimentalement. Pire, dans la plupart des cas, seul un très petit sous-échantillon de prédictions, habituellement trois ou un peu plus, forme la base d'une validation expérimentale du modèle, et sert comme critère valable pour la solidité de l'ensemble des prédictions.

Au lieu de réaliser un ensemble de modèles chacun validé suivant quelques mesures expérimentales, le groupe de travail a proposé l'idée de combiner les différentes informations contenues individuellement dans chaque modèle afin de pouvoir faire des prédictions plus précises. Cela a donné lieu à une première série de 5 challenges présentés à la communauté baptisés DREAM2 et basés sur les problématiques de réseau d'interactions moléculaires. Ces challenges ont été réalisés durant 3 mois entre Juillet et Octobre 2007. Au cours de ces trois mois, 109 équipes de 17 pays différents ont téléchargé les challenges, et 36 équipes ont soumis un total de 110 prédictions à l'évaluation. Même si les premiers résultats furent jugés décevants par les équipes organisatrices par rapport à leurs attentes, les challenges ont tout de même donné lieu à quelques publications intéressantes. Par exemple, le challenge BCL6 a permis d'obtenir un modèle pertinent qui n'existait pas dans la littérature et dont la publication fut acceptée dans la revue Cell. Les données collectées durant le challenge ont ensuite été mises à disposition de la communauté pour réaliser de futures analyses. Cette première expérience a surtout permis d'en apprendre beaucoup sur le fonctionnement communautaire des challenges. A partir de ce moment, DREAM a multiplié les challenges à la fréquence d'un groupe de challenges par an.

Ce travail répété a permis à l'équipe organisatrice d'améliorer le fonctionnement et la performance de leurs challenges au fur et à mesure et de gagner en productivité et crédibilité dans la communauté en biologie, regroupant à ce jour plus de 50 associations avec des insitutions partenaires. En 2013, DREAM Challenge s'est associé avec la SAGE Biotnetworks pour proposer un challenge sur le cancer du sein. Fort d'une nouvelle visibilité, le nombre de participants a explosé sur les challenges 8.5 et 9 en 2013 et 2014 pour un total de 1780 utilisateurs enregistrés, 159 équipes et 368 soumissions uniques. A partir de ce moment, DREAM a multiplié les publications, dont certaines dans des revues prestigieuses comme *Nature Biotechnology* (4 papiers), *Nature Review Genetics* ou *Science*, tout en améliorant continuellement le processus des challenges.

Foldit (issu de Franzoni & Sauermann, 2014)

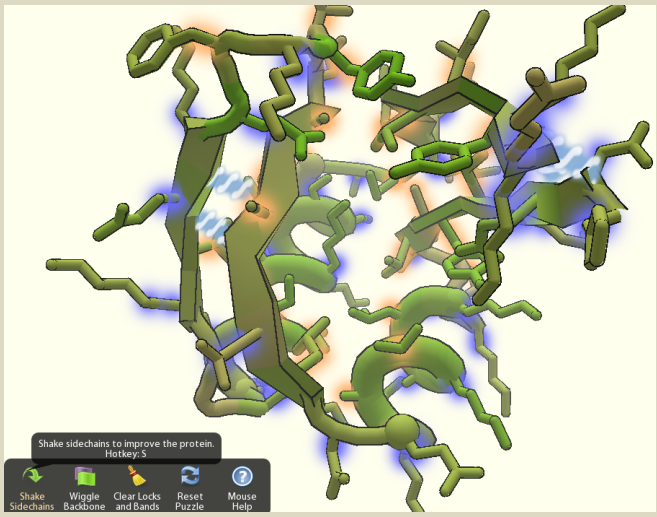


Figure . Interface d'utilisation de Foldit

Dans les années 1990, les scientifiques avaient développé des connaissances significatives sur la composition biochimique des protéines. Cependant, ils avaient une compréhension très limitée de la structure et des formes des protéines. Cette forme est importante car elle explique la façon dont les protéines fonctionnent et interagissent avec les cellules, les virus ou les protéines du corps humain. Par exemple, une protéine de forme appropriée pourrait bloquer la réplication d'un virus. Les méthodes conventionnelles pour déterminer les formes protéiques comme la cristallographie aux rayons X, la spectroscopie de résonance magnétique nucléaire et la microscopie électronique sont extrêmement coûteuses, coûtant jusqu'à 100 000 dollars pour une seule protéine, avec des millions de protéines à déterminer. En 2000, David Baker et son laboratoire de l'Université de Washington, à Seattle, ont reçu une subvention du Howard Hughes Medical Institute pour travailler sur la détermination de forme avec des algorithmes de calcul. Les chercheurs pensent qu'en principe, les protéines doivent se replier de sorte que leur forme utilise le minimum d'énergie pour éviter que la protéine ne se désintègre. Ainsi, les algorithmes de calcul devraient être capables de déterminer la forme d'une protéine en fonction des charges électriques de ses composants. Cependant, chaque partie d'une séquence protéique est composée d'atomes multiples et chaque atome a ses propres préférences pour se lier ou se différencier des autres atomes, résultant en un grand nombre de degrés de liberté dans une seule molécule, rendant les solutions informatiques extrêmement difficiles.

Baker et son laboratoire ont développé un algorithme appelé Rosetta qui combine des techniques déterministes et stochastiques pour calculer le niveau d'énergie de formes de protéines choisies au hasard à la recherche du meilleur résultat. Après plusieurs années d'améliorations, l'algorithme fonctionnait raisonnablement bien. A cause d'un besoin en calcul intensif, l'équipe a lancé à l'automne 2005 Rosetta@home, un système permettant aux bénévoles de mettre à leur disposition leur capacité de calcul de leurs ordinateurs personnels. Un élément qui peut sembler anodin, est que Baker et son équipe avait construit une interface visuelle qui montrait les protéines au fur et à mesure qu'elles se repliaient. Bien que les volontaires aient été sollicités pour contribuer uniquement à la puissance de calcul, en regardant les économiseurs d'écran de Rosetta, certains d'entre eux ont posté des commentaires suggérant de meilleures façons de plier les protéines que ce qu'ils ont vu faire l'ordinateur. Ces commentaires ont inspiré les étudiants diplômés du département d'informatique et de génie et les post-docs au laboratoire de Baker. Ils ont commencé à se demander si la capacité visuelle humaine pouvait compléter le pouvoir de

l'ordinateur dans les domaines où l'intelligence informatique était en train de se faire sentir. Ils ont développé un jeu en ligne appelé Foldit qui a permis aux joueurs de modéliser la structure des protéines avec le mouvement de la souris. Les joueurs peuvent inspecter une structure de protéine sous différents angles. Ils peuvent ensuite la déplacer, la faire pivoter ou retourner les branches de la chaîne à la recherche de meilleures structures. Le logiciel calcule automatiquement le niveau d'énergie des nouvelles configurations, montrant immédiatement une amélioration ou une détérioration. Un ensemble de fonctionnalité, spécifiques au repliement des protéines ont été mis en place pour faciliter l'exercice. Ces fonctionnalités, combinées à quelques tutoriels en ligne, ont permis aux gens de commencer à plier des protéines sans connaître pratiquement rien de la biochimie. Le plus intéressant, le logiciel a été conçu comme un jeu sur ordinateur et incluait un tableau de bord indiquant la performance des joueurs. En tant que tels, les joueurs ont eu la possibilité de grimper dans les classements et ils ont également pu créer des équipes et partager des stratégies pour rivaliser avec d'autres équipes.

Le jeu a été initialement lancé en mai 2008. En septembre de la même année, on comptait 50 000 utilisateurs. Les joueurs ont d'abord reçu des structures protéiques connues afin qu'ils puissent voir la solution désirée. Après quelques mois de pratique, plusieurs joueurs ont trouvé des formes très proches de la solution et dans plusieurs cas surpassé les meilleures structures conçues par Rosetta. De ces observations, il est devenu clair que l'intuition humaine était très utile car elle permettait aux joueurs de dépasser les pièges des optima locaux, ce qui créait des problèmes considérables pour les ordinateurs. Un an après le lancement, il y avait environ 200 000 joueurs Foldit actifs.

Dans les mois suivants, le développement de Rosetta et celui de Foldit se sont combinés. Certaines protéines où Rosetta échouait ont été distribuées aux joueurs de Foldit. En échange, les joueurs ont suggéré des outils automatiques supplémentaires qu'ils pensaient que l'ordinateur devrait leur fournir. Pendant ce temps, les joueurs ont mis en place des équipes avec des noms tels que "Another Hour Another Point" ou "Void Crushers" et utilisé des chats et des forums pour interagir. Certains joueurs ont également commencé à élaborer leurs propres «recettes», des stratégies encodées qui pourraient être comparées à celles créées en laboratoire. Par exemple, certaines stratégies consistaient à remuer une petite partie d'une protéine, plutôt que la totalité de la structure, d'autres étaient basées sur la fusion de protéines ou de protéines chaînes de baguage. Les stratégies complexes sont maintenant souvent intégrées dans des outils («recettes») qui peuvent ensuite être téléchargés et utilisés par d'autres3.

Les résultats étaient frappants. Une stratégie de joueur appelée "Bluefuse" a complètement dépassé Rosetta "Classic Relax", et a surpassé "Fast Relax", un morceau de code sur lequel les développeurs de Rosetta travaillaient depuis longtemps. Ces résultats ont été publiés dans PNAS et les acteurs de Foldit étaient co-auteurs sous un pseudonyme collectif (Khatib et al., 2010b). En décembre 2010, encouragés par ces résultats, Firas Khatib, Frank DiMaio dans le laboratoire de Baker et Seth Cooper dans l'informatique et l'ingénierie, parmi d'autres travaillant sur Foldit, pensaient que les joueurs étaient prêts à relever un vrai défi. En consultation avec un groupe d'expérimentateurs, ils ont choisi une protéase rétrovirale monomère, une protéine connue pour être critique pour le VIH de singe-virus dont la structure avait intrigué les cristallographes pendant plus d'une décennie. Trois joueurs de l'équipe "Contenders" Foldit sont arrivés à une solution assez détaillée de la structure des protéines en moins de trois semaines. En septembre 2012, les joueurs de Foldit étaient co-auteurs de quatre publications dans les meilleures revues.

Annexe 3 – Fiche d'evaluation des projets pour le challenge 1 d'Epidemium

PROJET VIZ4CANCER Note globale sur 50 :		Points /5	Commentaires
LE PROJET	Clarté et pertinence de l'approche envisagée		
	Originalité du projet		
LES MÉTHODES DE TRAVAIL	Travail collaboratif et complémentarité		
	Appropriation des technologies et outils mis à disposition		
LES RÉSULTATS ET CONCLUSIONS	Caractère innovant et travail accompli		
	Compréhension et clarté des résultats		
L'IMPACT SUR LA SANTÉ DES PATIENTS	Relevance médicale scientifique		
	Usage et appropriation par le milieu médical		
LES PERSPECTIVES	Vision à long terme		
	Durée de vie estimée du projet		

Verriez-vous le projet Viz4Cancer avoir...	(entourez la note que vous attribuez)
Le 1er prix	1 2 3 4 5 6 7 8 9 10
Le 2ème prix	1 2 3 4 5 6 7 8 9 10
Le 3ème prix	1 2 3 4 5 6 7 8 9 10
La mention Meilleur impact sur la santé des patients	1 2 3 4 5 6 7 8 9 10
La mention éthique	1 2 3 4 5 6 7 8 9 10
La mention Originalité du modèle	1 2 3 4 5 6 7 8 9 10
La mention Meilleure dataviz	1 2 3 4 5 6 7 8 9 10
La mention Meilleur wiki	1 2 3 4 5 6 7 8 9 10
La mention Inclusion et travail collaboratif	1 2 3 4 5 6 7 8 9 10

Annexe 4 – Crowd-based Hypothesis generation from data and the organisation of the participative scientific process (Design conference 2018, Sitruk & Kazakçi)

INTERNATIONAL DESIGN CONFERENCE - DESIGN 2018
Dubrovnik - Croatia, May 21-24, 2018.

CROWD-BASED HYPOTHESIS GENERATION FROM DATA AND THE ORGANISATION OF PARTICIPATIVE SCIENTIFIC PROCESS

[Authors will be inserted automatically]

Abstract

[Abstract will be inserted automatically]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[Do not delete or modify the abstract area]

[Keywords will be inserted automatically]
[---]

1. Introduction

During the past decade, the practice of science has been facing profound changes in its organization and its processes. The paper considers two major trends affecting the contemporary science - openness and data-drivenness, and how these trends affect, in turn, the generation of hypotheses in science. The ability to generate the *right* hypothesis is also the mark of a genuinely creative scientist, since a research hypothesis very often determines the *value* of the subsequent process and results.

Hypothesis generation is the least understood and most secluded activity in a scientific process. Despite the recent efforts to open up the scientific process (Franzoni & Sauermann 2013; Wiggins & Cruston 2011), hypothesis generation has remained confined to the laboratory, and most often, to the intellect of the lone researcher. Traditionally, openness in science is manifested by a group of scientists delegating the execution of a part of their scientific activity to the crowd. These activities include data collection and annotation, analysis of existing datasets or building models. One of the main reasons originating the open science movement is to momentarily increase resources available to the scientists with little or no cost. An often-cited case, Galaxy Zoo, that consisted in involving ordinary citizens to add some common sense, high-level semantic information to images of galaxies, enabled scientists to gather enough data for analysis in a few weeks, whereas it would have taken 83 years to collect that same amount of data for a single individual (Franzoni & Sauermann 2013). Several studies demonstrate that crowdsourcing scientific projects may foster innovation, by

harnessing the brainpower and imaginations of the many and leading to a larger variety and quality of solutions (King & Lakhani 2013; Afuah & Tucci 2012; Panchal 2015). Nevertheless, finding new and interesting hypotheses have not been a topic of open science literature so far.

In the meanwhile, the increasing availability of massive datasets seems to affect how hypotheses are being generated in science today. The emergence of Big Data is considered to ease the accessibility to data and the creation of science projects across disciplines and potentially by non-specialists (Laney 2001; Kitchin 2014; Boyd and Crawford 2012). Authors such as Kitchin (2014) and Gray (2009) consider that data-driven science is a fundamental paradigm shift that will transform the way humanity will produce scientific knowledge in the following way. In the previous paradigm, data collection follows the formulation of a research hypothesis (Prensky 2009). In other terms, the hypothesis drives the data collection process. In the new paradigm, researchers are no longer constrained by existing theories for generating research hypotheses - thanks to the advantages offered by the multiplicity of publicly available datasets (Kitchin 2014). Epistemologists define this shift of science as a move from a 'knowledge-driven' science to a 'data-driven' science (Kitchin 2014). Thus, big data will be considered at the very beginning of the scientific process and will play a fundamental role in generating hypotheses, and hence, the value of the scientific output.

Currently, open science and data-driven science literatures have no overlaps. While studies on open-science emphasize the creative potential of involving broader audiences into the scientific process, as far as the authors know, there are no studies or published evidence that the crowd can deploy this creative potential into the most challenging scientific activity where creativity is needed most: the generation of new and valuable hypotheses. On the other hand, while data-driven science literature argues that data will enable the generation of fundamentally different research hypotheses, it does not clarify whether a crowd of non-specialists can accomplish this particular activity.

The central questions we consider in this paper are thus motivated by the lack of intersection between these literatures: Is it possible that a crowd generates useful research hypotheses based on large amounts of data? We shall study this question based on case study of an open scientific community, called Epidemium, working on cancer research. Epidemium, sponsored by Roche Laboratories, is built with the mission to rally a community around publicly available 21,000 datasets related to the epidemiology of cancer. The 2 years research during which Epidemium organized a series of open challenges and built a community of participants demonstrate that hypotheses generation is indeed a non-trivial process for the crowd. First, the availability of data is not enough for generating hypotheses that are consistent with the available data. Difficulties faced by many participant teams points to a need to manage the exploration and the appropriation of data by the crowd for an effective hypothesis generation. Second, when working on a large number of disconnected datasets, it is not possible (nor, necessarily, desirable) to generate all potentially useful hypotheses in one pass. Organizers need to adopt a strategy of exploring the hypotheses space through successive challenges and to capitalize on the intermediary results to become able to help the community to develop better and better hypotheses.

The plan of the paper is as follows. First, we will review the notions of hypothesis generation in the literature and a lack for organizing it through a collective activity from existing data. We present then how crowdsourcing is used for well-defined problems in big data. Then, we introduce our case study in some detail. We will present a practical case of generating hypothesis by crowdsourcing in epidemiology of cancer. An analysis of the crowdsourcing process and its limits will be presented. Finally we will provide some insights for the design problems.

2. Theoretical background

2.1. Hypothesis generation: from individual problem-solving to collective design

Hypothesis generation process in science has been widely studied during the twentieth century by philosophers, epistemologists, logicians or designers (Douven 2017). A notable strand of work on this topic comes from computational models of scientific discovery from Herbert Simon and colleagues (Kulkarni & Simon 1988; Lindsay et al. 1993; Antonsson & Cagan 2005). In these models, hypothesis generation has clearly been described as a part of a more general problem-solving activity.

One example for such models is, KEKADA, which automated some tasks of the scientific process by analyzing Hans Krebs' process on urea production research. KEKADA plans sequence of experiments in order to produce observations that can be used to formulate descriptive and explanatory theories of a set of phenomena.
KEKADA conducts a double search process, in an instance space and a rule space. The possible experiments and experimental outcomes define the instance space (e.g. molecular substances), which is searched by performing experiments (run by an external user). The hypotheses and other higher-level descriptions, coupled with the confidences assigned to these, define the rule space.
Hypothesis generation process is conceptualized as a two-step process: the generation of new hypotheses based both on existing knowledge and facts from experiments, and the choice of the hypothesis according a set of rules from a decision-maker process. Each experiment generates new facts that are then evaluated through the set of rules and incorporated in the existing knowledge.
Other researchers conducted experiments to develop automated algorithm to specifically study hypothesis generation. DENDRAL for example is an algorithm to help organic chemists in identifying unknown organic molecules, by analyzing their mass spectra and using knowledge of chemistry (Lindsay et al. 1993). A subsystem incorporates specific knowledge of chemistry and mass spectrometry, accepts a mass spectrum and other experimental data from an unknown compound as input, and produces an ordered set of chemical structure descriptions hypothesized to explain the data.
The underlying problem-solving models to these problems, and to many similar research projects, are clearly based on well-defined problem domains. One might argue that the main difficulty in research, and particularly in hypothesis generation, is to be able to arrive to such a clear structuring of the space. Indeed, Simon himself acknowledges this point (Simon, Langley, & Bradshaw 1981). Hatchuel (2001), on the other hand, argues that what is usually called ill-structured problems are simply design projects, with clearly defined yet under-specified formulations due to a lack of understanding and clarity. As we shall see in our case study, both conditions are satisfied during the hypothesis generation based on a large number of data sets. This, in turn, implies that the whole process can be seen as a collective design activity, in this particular case carried out by a crowd.
Another fundamental shortfall of the above computational models is that hypothesis generation depends on the existing knowledge. As mentioned in the introduction, data-driven approach however is no longer constrained by the theories and existing knowledge in general and data instead forms the basis of the reasoning (Kitchin 2014). The process is fundamentally different since data provides much less insights on a phenomenon than knowledge (Davenport & Prusack 2005).
Finally, these models focus on the individual hypothesis generation, and thus ignores this activity can be accomplished by multiple actors. The major difference between the individual and the collective regimes is that the collective work requires organization. Literature on hypothesis generation is barren from that perspective - the question has not been considered, since historically, this was an activity that has never been opened to the crowd. The opening of this phase to the crowd is even more problematic, since, how to manage or organize is by definition a hard question: according to Oxford Living Dictionaries the crowd is defined as a large number of people gathered together in a *disorganized* or unruly way.

2.2. Crowdsourcing to outsource search process in the solution space

Many studies suggest that crowdsourcing process fosters innovation, by harnessing the brainpower and imaginations of many and leading to larger variety and quality of solutions (King & Lakhani 2013). Often implicitly crowdsourcing literature adopts Simon's original problem-solving metaphor. Afuah & Tucci (2012), for instance, explain that when a contributor conducts a *search* from its current position, he tends to focus on the alternatives around its *neighborhood*. Crowdsourcing thus can be seen as multiplying the number of contributors to *increase the number of local searches* and the probability of finding the right knowledge and contributor for solving the problem.
It comes thus no surprise that open innovation literature's main finding is that crowdsourcing is particularly efficient in a well-defined problem situation (Afuah & Tucci 2012). In particle physics for example, methods to detect Higgs boson from data were initially developed based on pre-simulated data that were inconsistent with real observations collected from the Large Hadron Collider (LHC).

Instead of internalization, physicists decided to outsource the search of a better model through the open challenge HiggsML (Bourdarios et al. 2014). The problem was designed so that no knowledge of particle physics was required to participate. As a result, more than 1700 teams have participated to the challenge, which was the biggest participation to a machine learning challenge at the time.
While crowdsourcing is an effective process for scientists to outsource the search process in the problem space, its use for hypothesis generation in a data-driven approach has not been considered. Indeed hypothesis generation from data can be considered as an ill-defined problem, where hypotheses are hard to structure or to evaluate since we do not know yet whether it leads to an original result nor it can be solved with the existing data. The outcome of a crowdsourcing process based on ill-structured problem is fuzzy since the organizer does not have a clear idea of what he is looking for.

3. Case study: a worldwide open medical project for cancer research

3.1. Method

Our research was conducted from November 2015 to November 2017 with Epidemium, an organization designed for scientific research dedicated to the understanding of epidemiology of cancer. We followed a collaborative management research (Shani et al. 2008), conducted by academics and practitioners in order to create actionable knowledge for the organization and new theoretical models in management research (David and Hatchuel 2008). Other written sources were solicited such as the wiki page of every project, the website and the white book of Epidemium. The purpose of this research was to investigate how an initiative based on Big Data should generate interesting hypotheses through a crowdsourcing process. From the perspective of Epidemium, the goal was to validate or invalidate whether a crowd may help research on epidemiology of cancer by renewing traditional research questions using the availability of disparate data sources. From our research perspective, our first aim was to gain better insight into when and how crowdsourcing maybe an interesting form of organization for generating research hypotheses. Second, we wanted to see what kind of theoretical frameworks would be needed in the crowdsourcing of a creative activity where, traditionally, it is thought that extensive knowledge and expertise is a necessary.

3.2. Big Data in the context of epidemiology for cancer

3.2.1. Epidemiology and Big Data

Epidemiology is a scientific discipline involving both medicine and statistics that studies risk factors associated with the incidence or mortality of diseases. Since the 1950s, epidemiological studies have used statistical methods that allow them to extrapolate results on samples to much larger populations. This approach have led to the emergence of numerous studies on behavioral risk factors such as exposure to alcohol, smoking or nutrition. Statistical biases in sampling, however, affect the extrapolation of local phenomena and several studies highlighted that the results are sometimes contradictory on similar risk factors (for example, a given ingredient in food can both prevent and cause cancer according to different studies; see Schoenfeld and Ioannadis, 2013). The recent emergence of massive databases on the incidence and mortality of diseases is seen in the Epidemiology community as an opportunity for epidemiological studies that could reduce the current problems and limits of existing approaches.

3.2.2. Epidemium as a structure to access knowledge communities

The notion of big data has engendered a wide appeal in health sector and several actors are seeking new opportunities. Roche Laboratories, a pharmaceutical company, wants to evaluate how big data analysis in epidemiology could be a catalyst for a new more preventive and personalized medicine. Although Roche have already participated in epidemiological surveys (e.g. ObEpi 2012 study), in-house experts face a double constraint compared to conventional statistical methods. First, the lab teams are not experts in the data science methods. Second, the scientific method to be implemented

differs from conventional statistical methods. While data collection is directly related to a predetermined hypothesis, Big Data Epidemiology seeks to query a database already collected before the hypothesis is defined. Moreover, data have not been collected by epidemiologists but rather heterogeneous institutes and thus restrict an overview given a particular objective. In order to bring together both medical players and data analysis experts, Roche initiated collaboration with a new kind of research laboratory, called La Paillasse, whose objective is to be a research institution open to all citizens. La Paillasse provides Roche laboratories with a culture of open science and access to a community of scientists sensitive to openness in science. To ensure unity and a form of independence, the two entities created the Epidemium project, which is intended as a structure designed for scientific research dedicated to the understanding and epidemiology of cancer.

3.2.3. Available data

The first step of Epidemium was to collect all available open data related to the epidemiology of cancer and prepare the data to make them easily exploitable. A core data set has been compiled on mortality and cancer incidence from the databases available on the OECD and World Health Organization sites. These datasets are extended over periods of about 60 years and specify the type of cancer, country or region, age group and period of death. Data sets on the risk factors related to Sexually Transmitted Infections (STIs), particularly in the United States, as well as a set of datasets on general information (demography, environment and agriculture, climatology, work and working conditions, economic indicators, potential or actual behavioral risk factors, general health data, cancer data) represents the predictor variables to be correlated with the main dataset. In order to extend the scope of possible studies, the Epidemium team has compiled information on all publications in epidemiology in the medical scientific literature. Several datasets have been integrated, including clinical trials gathered on the WHO platform, ClinicalTrials.gov, Clinical Study Data Request, and the full database of PubMed Open Access publications plus publications on PubMed. Finally, Roche laboratories have made available a dataset of studies carried out by the laboratory. In total, Epidemium made it possible to compile a set of about 21,000 datasets accessible to all participants and free of rights and use. In a medical setting, the projects as well as the data used must comply with an ethical framework. The guarantee of anonymization the data is indeed complex to manage with open datasets. An ethics committee has been set up to delimit Epidemium's framework and ethical charter.

3.3. Crowdsourcing contest to explore the data

Epidemium is led by a core team of 6 people, mainly experts in open science and community management. With the overall objective of fostering collaboration through a common scientific purpose in mind, the organizers of Epidemium decided to launch a crowdsourcing contest, named Challenge4Cancer, based on the collected datasets. The declared objective was twofold: identify relevant hypotheses from the available databases and develop methods to test those hypotheses based. Ability to identify missing knowledge and know-how, and identify relevant stakeholders forms the basis competency needed by the Epidemium core team to build the community. In order to promote the project, Epidemium made 115 presentations in a large variety of external organizations seen as potential partners or *hubs* where talents can be recruited for the challenge. Various partnerships with recognized institutions in the medical and scientific field (APHP, Institut Curie, Cancer research Cluster CLARA) as well as a set of technical partners that provide tools for management, storage and analysis of data (Teralab, Dataiku, Hypercube, Center for Data Science) were established. Existing data science communities, such as the RAMP data challenge platform were also involved in order to facilitate access to the existing pool of talents.

Challenge4Cancer took place over 6 months between November 5, 2015 and May 6, 2016. In total, 678 contributors participated. Epidemium defined four challenges from the available datasets:

- Understanding the distribution of cancer over time and space;
- Risk factors and protective factors of cancer;
- Meta-epidemiology: understanding cancer from medical scientific literature;
- Environmental changes and cancer.

The challenges are deliberately under-specified to allow room for a variety of research hypotheses. Note that these clear yet under-specified formulations are the primary source of ill-define problems. Any participant can further specify the problem she or he is trying to solve, thus engages in a design activity where both the objective and the solution should be specified (Hatchuel 2001, 2010). Each participant or team chooses one of the challenges and defines a problem to be solved from the data relating to the challenge. Epidemium designates a scientific committee whose objective is to accompany the contributors during the production process in order to control the compliance between the projects and the proposed challenges. In a community of 678 members including 331 participants registered in the tournament (54% of data scientists, 28% of computer scientists and 18% of health professionals or medical researchers), 75 people took part in one of the 16 projects, with 63 finalists for 8 projects selected by the committees. Epidemium encouraged teams to collaborate with each other by including in the final evaluation the level of cooperation of the project during the tournament and by fostering exchanges between the participants. A weekly meet up held in the premises of La Paillasse that makes easier to integrate new contributors in the projects and to physically meet the various participants for potential collaborations. Project teams are also asked to fill a wiki page on the project run. At the end of the challenge the ethical and scientific committees evaluate the projects. Three winning projects receives a prize: € 5,000 for the first and € 2,000 for the second and third.

4. An analysis of the crowdsourcing process for Challenge4Cancer

The first round of challenge organized by Epidemium was rich of insights. In this section, we will highlight some of observations and analyze them both from the perspective of participants and their hypotheses generation processes and from the perspective of the organizers of Epidemium and their learning in terms of the management of such a process.

4.1. The participants' processes: Specifying and reformulating hypotheses when confronted with data

The participants have encountered several difficulties in conducting their projects. The main factor that was associated with the high failure rate of the projects (only 16 project was submitted in the end out of 678) was the inability of the groups to terminate in time. Those who managed to submit a proposal did not manage to reach the objectives they initially fixed and needed to adapt their final submissions by providing prototypes or simplified versions of the original target. Several aspects of the problem have caused these on-the-fly modifications, such as technical difficulties in the search of an efficient machine learning model, data quality issues and the inability to explore efficiently the large number of data sets.

Only 16 teams proposed a project that fit into the objectives of the challenge. After further screening by the evaluation committee, 8 projects have reached the final of Challenge4Cancer where the number of participants varies across projects from one to several dozen. We can identify several categories. A first category strived to build causal or predictive models between various factors to test certain hypotheses (Baseline, Predictive approach and cancer risk). A second category of teams dealt with data visualization tools, in order to facilitate hypotheses formulation (Viz4Cancer, CancerViz) or to explore the scientific literature (OncoBase, BD4Cancer, Venn). Finally, a unique project proposed to use the data to raise awareness about cancer in a more targeted and data-driven way than the usual solutions (ELSE).

For the first category, we can cite the project 'The Predictive Approaches and Cancer Risk'. This project had to limit its exploration and reformulate their initial hypothesis during the contest. During the process they realized their initial ideas were too broad and difficult to test, so they needed to restrict the initial scope. Their research for a better-specified hypothesis was hindered though by the data quality problems they discovered progressively as they inspected different datasets more closely.

The Baseline project went through a similar cycle. Initially, project leaders wanted to predict cancer incidence, mortality and survival using risk factors from open data sources (with a global scope and a regional granularity). When they needed to program this question as a prediction problem, they realized the question was too broad to warrant a predictive modeling with the available data. The lack

of coherence and substance between the various datasets made it impossible to target a general prediction problem. One of the reformulated hypotheses was " the risk modeling of the data mortality for digestive cancers (gut, colon, rectum and anus, liver, gallbladder) as a function of age and other types of cancer".
While many groups discovered during the process the incompatibility of their initial target with what the available data or analysis tools can deliver, some other groups decided to add new datasets to the database made available by Epidemium. Indeed, data quality issues that were soon realized by several groups led BD4Cancer project to team up with Baseline project in order to create a new database, EpidemiumDB. The data collection was done according to a standardized process designed by the team leaders and the collection was divided between the contributors. Some groups have included this new datasets to continue their work.
As these examples illustrate, several groups needed to abandon or reformulate their original hypotheses once they have started to study the available data in some detail. Before this confrontation, many of the ideas that were put forward were beyond what the Epidemium data could provide. This can be seen as a form of undesired out-of-box phenomena.

4.2. Learning from the challenges: Structuring future work by developing new tools and mechanisms

As with the participants, the organizers have faced several difficulties. One such difficulty was to develop a sound method for evaluating the large variety of proposals. Although the organizers have assumed that this variety would be beneficial to map out the space, they soon realized there could be no easy way to evaluate projects that were very different in their nature. To determine the winners, they have adopted an ad hoc method to compare projects in a pairwise manner. This helped them to figure out which project stood out with what aspect. This gradually became a set of custom-made criteria that could be used for selection, derived directly from the analysis of each project:

- the project clarity and relevance of the proposed approach,
- the originality of the project,
- the working methods (Collaborative work and complementarity, Appropriation of the technologies and tools made available),
- the results and conclusions (Innovative character and work done, Understanding and clarity of the results),
- the impact patients' health (Scientific medical relevance, Use and appropriation by the medical community) and
- perspectives (Long-term vision, Estimated life of the project).

Using these criteria, the three winning projects were selected to be Baseline, CancerViz and ELSE. The Baseline and BD4Cancer projects were the most unifying in the community and the most supported by recognized experts in the medical field and data analysis, which allowed them to spread widely beyond the Epidemium community.
The second major difficulty that was identified concerns the time and effort needed by the participants to develop an understanding of the proposed datasets. Epidemium provided to the contributors both a file to download with the data and a synthesis of what is inside the data. For example, the general data set was presented through categories such as demography, environment and agriculture, work, behavior, or economics factor. Each category was then divided in subcategories: for example, behavior contained data on alcohol consumption, tobacco consumption, coal use, telephone consumption, death road accident. Initially, it was assumed that this was all that is needed. As we have seen, many participants have simply ignored the details of these datasets when generating their initial target - which proved costly since reformulation was needed. These reformulations were not deemed to be progress since, it was not about further specification of testable hypotheses, but rather, the grounding of initial vague ideas to what was accomplishable with the current time and resources.
Yet another important point affecting the process was the level of specifications of the challenge themes. As we have already pointed out, the challenge themes were under-specified. This led to the lack of ability from the contributors to propose advanced solutions during the challenge. However, this was also inevitable since the organizers themselves did not have very clear ideas about what the more

specific research targets could be. In a way, they needed a first round of challenge to promote a breadth-first search strategy, which would give them a better overview of the whole design space. This also implies that successive challenges are needed, where the scope would be reduced to more specific issues and there would be, in theory, more productive.
Finally, an important issue is how to capitalize, not on how to run future challenges, but, how to use the results to contribute to the scientific knowledge on the cancer research. During the process, some projects for example found initial results to contribute to the scientific knowledge by highlighting possible correlations between risk factors and the incidence of certain cancers. *Baseline* for example identified a possible correlation between Black African populations and the incidence of prostate cancer. 'Predictive Approaches and Cancer Risk' project focused its investigations on the incidence of pancreatic cancer. Initial analyses shown that the most discriminating variable for explaining pancreatic cancer, among agro-environmental variables, energy use in the agriculture and forestry sectors as a percentage of total consumption of energy. Currently, there is no scientific publication on this topic, and the result remains a local knowledge detained by the Epidemium community.

4.3. Setting up a second challenge: deepening and centring the search

After the first contest, Epidemium set up a second crowdsourcing in June 2017. Contributors in the first challenge had difficulties to submit a finished product, facing a gap between hypothesis formulation and existing database. Organizers wanted to limit such constraints by lowering the quantity of available data set and reducing the number of challenges proposed to two: constructing a Data-Visualization of the incidence of cancers by exposing the epidemiological factors associated with their dynamics; developing a predictive tool for the progression of cancer in time and space, depending on the known or supposed factors that determine its evolution. Moreover, objectives were readjusted, and organizers asked a final scientific publication to the teams to win. Epidemium generated a great deal of enthusiasm with the first contest and many well-known French engineering schools, such as Centrale-Supélec and Polytechnique, were interested for using the challenge as a platform for student projects. Epidemium therefore set up a student-related challenge on predicting cancer mortality in developing countries. Second Challenge4Cancer has been launched in November 2017 and should be finalized by March 2018.

5. A model of the crowdsourcing process for hypothesis generation

Our analysis of the previous section demonstrate that, in a scientific hypothesis generation context, crowdsourcing should be thought as an iterative activity where the organizers need to capitalize on the results of successive challenges to better learn both the value of emerging hypotheses and the ideal methods and tools for better managing the community in the successive events. In this section, we present a process model that describes how the process can be extended to manage this iterative process. This model was implicitly used by Epidemium although the internal steps were not considered as part of the crowdsourcing process.

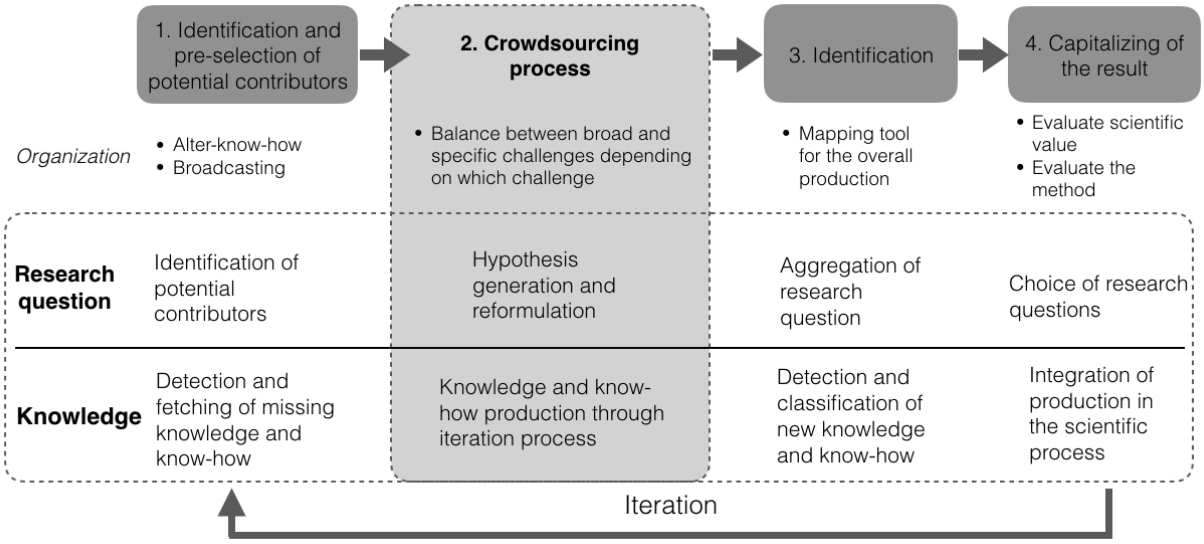


Figure 1. Model of the crowdsourcing process for hypothesis generation

5.1. Identification and pre-selection of potential contributors

As the Epidemium case demonstrate, it is important for the organizers to find the right participants that can bring the necessary expertise to the crowdsourcing process. This is an activity that should not be underestimated by the organizers, since it requires identifying missing knowledge and know-how, as well as identify where those resources can be found. The project leaders should *broadcast* the objectives of their initiatives through strategic events and leverage the intrinsic motivation of the targeted participants. We call this step the search for *alter-know-how*, the identification and fetching of missing know-how. One should be aware that every subsequent contest modify however the initial setup and modify also which kind of knowledge is needed inside the process. There is a need to systematize this action inside the iterative process while it is necessary every time a new contest is designed.

5.2. Crowdsourcing process

The Epidemium case demonstrates that attention needs to be paid to the level of specification of the challenge objectives. At this point, four different objectives compete. First, the objectives should be broad enough to allow room for the generation of a variety of hypotheses. This is particularly true if the organizer do not have a clear vision of the search space and the associated values of potential research questions. Second, the objectives might also need to be specific enough so that the cost of reformulation is not high and the participants remain a certain level of productivity. The organizers may need to reduce this time with tools to improve the appropriation process of the data by the crowd (particularly, in order to avoid the undesirable out-of-the-box effect seen in paragraph 4.1). We have seen in paragraph 4 that Epidemium committee already identified this as a critical part and integrated the generation of new tools as one of the two challenges in the second contest. As Escandon-Quintanilla (2017) suggest in ideation in engineering design processes, the way and the degree to which the participants are allowed to interact with the data has important consequences on the outcome of ideation. Third, the degree of specifications will depend on the current level of advancement in the previous episode of challenge. Epidemium has started with very broad topic which allowed to explore the space very broadly but they soon realized that the next challenge should be more specific and targeted, and possibly a third challenge can be run on a very specific subject that was generated during the second rounds and determined to be highly valuable from a scientific point of view. Fourth, the scientific value of the proposals should be monitored between successive cycles.

As the space is likely to be exponentially large on the number of datasets available, care must be taken for steering the exploration process in-between challenges.

5.3. Identification of the crowdsourcing outcome

During the first contest, challenges were too broad and gave not enough specific hypotheses generated by the crowd to be considered as a potential value for the scientific literature. However, some insights emerged providing tacit elements for the next challenge, such as potential correlations between variables or exploration of data analysis. These informations should be categorized to prove its worth and how it can be reintegrated in the next crowdsourcing. A mapping tool like CK theory for categorization might be useful as we first explore what can be done with the data (Hatchuel et al. 2011). The organizers should pay specific attention to what are the specific questions made by the crowd, and their level of specification. C-K provides interesting informations to evaluate the level of specification of a hypothesis and give some insights on the missing knowledge needed for a specific hypothesis. A first experimentation was made during the first challenge (Kokshagina & Sitruk 2017, see figure 2) and investigations should explore how this tool can be used to reintegrate the production of the previous contests.
Analyzing the overall production of the crowdsourcing process through tools for categorization should provide new metrics for evaluating the final contribution of every project.

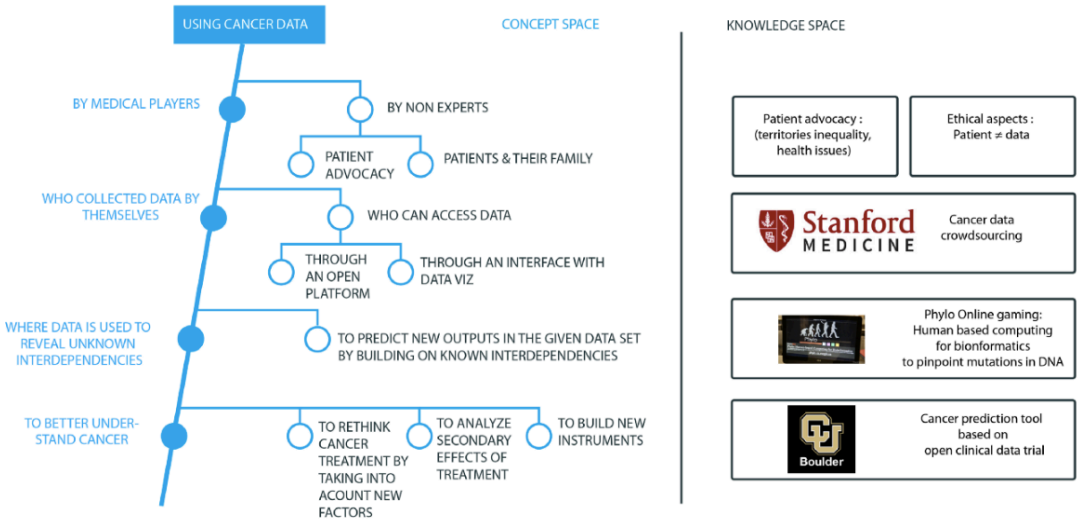


Figure 2. Extract from CK tool (Kokshagina & Sitruk 2017)

5.4. Capitalizing on the results

The organizing committee should be able to mobilize appropriate experts to evaluate the scientific value of those proposals. This will allow to set priorities to determine which challenge to be organized next. Result of step 3 should explicitly guide to the design of the new challenges. Organizers need also to take stock of the methods used and specify which element has been effective and others that need to be improved.

6. Discussion

This paper analyses the crowdsourcing method and identifies a lack of literature on the study of generating hypothesis using the crowd on data-driven projects. We conducted an in-depth analysis at Epidemium that highlights two organizational learning activities that need to be included in the crowdsourcing process: learning from the contributors and learning from the organizers. We propose a process that includes the two forms of learning identified. Further researches should be done to explore the applicability and performance of the proposed process in future Epidemium contests and other scientific contexts.

Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B., and Rousseau, D. (2014). The higgs boson machine learning challenge. In NIPS 2014 Workshop on High-energy Physics and Machine Learning, volume 42, page 37.

Afuah, A., & Tucci, C. L. (2012). Crowdsourcing as a solution to distant search. Academy of Management Review, 37(3), 355-375.

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. Wired, 23 June 2008. Available at: https://www.wired.com/2008/06/pb-theory/ (accessed 05 December 2017).

Antonsson, E. K., & Cagan, J. (Eds.). (2005). *Formal engineering design synthesis*. Cambridge University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

Callon, M. (2009). Acting in an uncertain world. MIT press.

David, A., Hatchuel, A., "From actionable knowledge to universal theory in management research", in: Shani, A.B. (Ed), Handbook of Collaborative Management Research, Sage Publications, Thousand Oaks, CA, 2008.

Douven, Igor, "Abduction", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/abduction/>.

Escandon-Quintanilla, M. L. (2017). Effects of data exploration and use of data mining tools to extract knowledge from databases (KDD) in early stages of the Engineering design process (EDP) (Doctoral dissertation, École de technologie supérieure).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. Research Policy, 43(1), 1-20.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., & Schader, M. (2011, August). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In AMCIS.

Hatchuel, A. (2001). Towards Design Theory and expandable rationality: The unfinished program of Herbert Simon. *Journal of management and governance*, *5*(3), 260-273.

Hatchuel, A., P. Le Masson, Y. Reich and B. Weil (2011). A systematic approach of design theories using generativeness and robustness. Proceedings of the 18th International Conference on Engineering Design (ICED11), Vol. 2: 87–97.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.

King, A., & Lakhani, K. R. (2013). Using open innovation to identify the best ideas. MIT Sloan management review, 55(1), 41.

Kokshagina O., Sitruk Y. Open Science: how to identify exploration axes in a transdisciplinary context? Medium, 17 October 2017. Available at : https://medium.com/epidemium/using-big-data-to-understand-cancer-epidemium-discover-how-c-k-method-can-be-used-to-identify-898120dbfac4 (accessed 5 December 2017)

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. Cognitive science, 12(2), 139-175.

Laney, D. (2001). 3D Data management: Controlling data volume, velocity and variety. Meta Group. http://blogs. gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Var iety.pdf. Accessed 16 Jan 2013.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, 61(2), 209-261.

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering applications of artificial intelligence*, *16*(4), 277-291.

Obépi-Roche, R. (2012). Enquête épidémiologique nationale sur le surpoids et l'obésité. *Paris: Inserm/TNS Healthcare/Roche*.

Panchal, J. H. (2015). Using Crowds in Engineering Design–Towards a Holistic Framework. In *2015 International Conference on Engineering Design, Design Society, Milan, Italy, July* (pp. 27-30).

Prensky, M. (2009). H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: journal of online education*, *5*(3), 1.

Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. Am J Clin Nutr 2013;97:127-34.

Shani, A. B., Mohrman, S. A., Pasmore, W. A., Stymme, B., Adler, N., "Handbook of Collaborative Management Research", Sage Publications, Thousand Oaks, CA, 2008.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese*, *47*(1), 1-27.

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Wiggins, A., & Crowston, K. (2011, January). From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii international conference on* (pp. 1-10). IEEE.

TABLES DES TABLEAUX ET DES FIGURES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

De plus en plus d'organisations scientifiques contemporaines intègrent dans leur processus des foules de participants assignés à des tâches variées, souvent appelés projets de science citoyenne. Ces foules sont une opportunité dans un contexte lié à une avalanche de données massives qui met les structures scientifiques faces à leurs limites en terme de ressources et en capacités. Mais ces nouvelles formes de coopération sont déstabilisées par leur nature même dès lors que les tâches déléguées à la foule demandent une certaine inventivité – résoudre des problèmes, formuler des hypothèses scientifiques - et que ces projets sont amenés à être répétés dans l'organisation. A partir de deux études expérimentales basées sur une modélisation originale, cette thèse étudie les mécanismes gestionnaires à mettre en place pour assurer la performance des projets délégués à la foule. Nous montrons que la performance est liée à la gestion de deux types de capitalisation : une capitalisation croisée (chaque participant peut réutiliser les travaux des autres participants) ; une capitalisation séquentielle (capitalisation par les participants puis par les organisateurs). Par ailleurs cette recherche met en avant la figure d'une nouvelle figure managériale pour supporter la capitalisation, le « gestionnaire des foules inventives », indispensable pour le succès des projets.

MOTS CLÉS

Crowdsourcing, Gestion de la performance, Tâche répétée, Science data-driven, Processus scientifique, Génération d'hypothèses, Théorie de la conception

ABSTRACT

A growing number of contemporary scientific organizations collaborate with crowds for diverse tasks of the scientific process. These collaborations are often designed as citizen science projects.The collaboration is an opportunity for scientific structures in a context of massive data deluge which lead organizations to face limits in terms of resources and capabilities. However, in such new forms of cooperation a major crisis is caused when tasks delegated to the crowd require a certain inventiveness - solving problems, formulating scientific hypotheses - and when these projects have to be repeated in the organization. From two experimental studies based on an original modeling, this thesis studies the management mechanisms needed to ensure the performance of projects delegated to the crowd. We show that the performance is linked to the management of two types of capitalization: a cross-capitalization (each participant can reuse the work of the other participants); a sequential capitalization (capitalization by the participants then by the organizers). In addition, this research highlights the figure of a new managerial figure to support the capitalization, the "manager of inventive crowds", essential for the success of the projects.

KEYWORDS

Crowdsourcing, Performance Management, Repeated task, Data-Driven Science, Scientific Process, Hypothesis Generation, Design Theory