# *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.

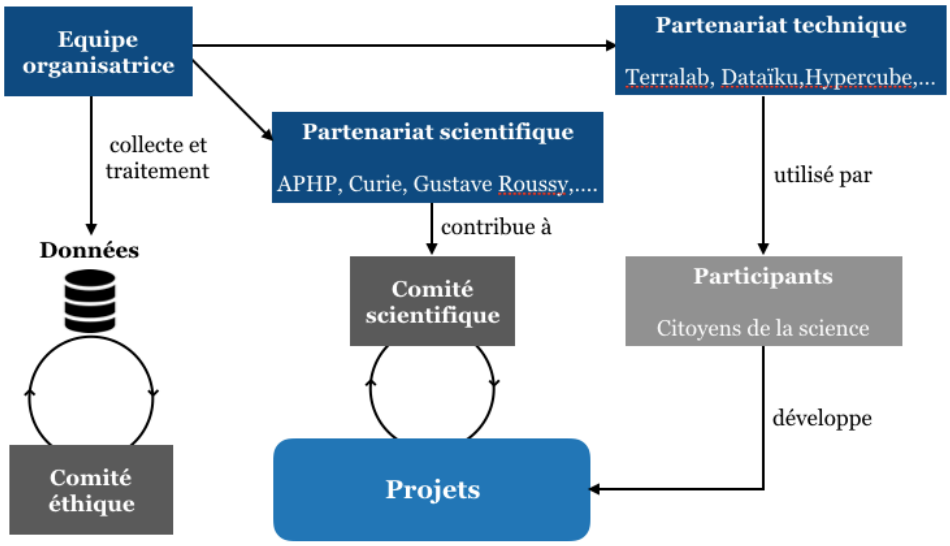


**Figure 47. Organisation d'Epidemium.**

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 $\tau$ 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,…)

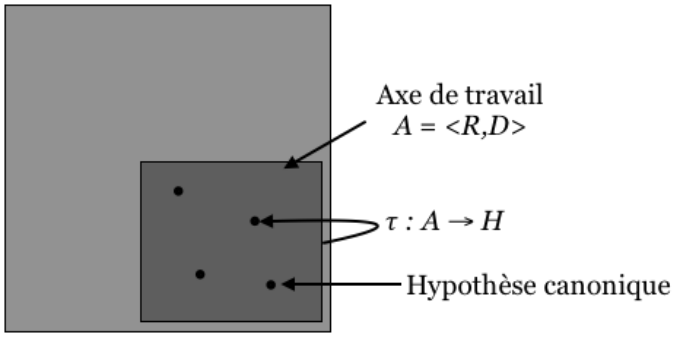


**Figure 48. Espace des hypothèses et fonction de transformation de l'axe de travail vers l'hypothèse scientifique**

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.

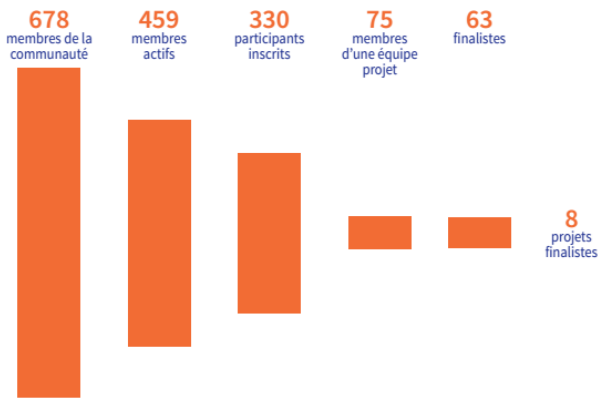


**Figure 49. Taux de conversion des membres de la communauté en projets actifs.**

### *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.

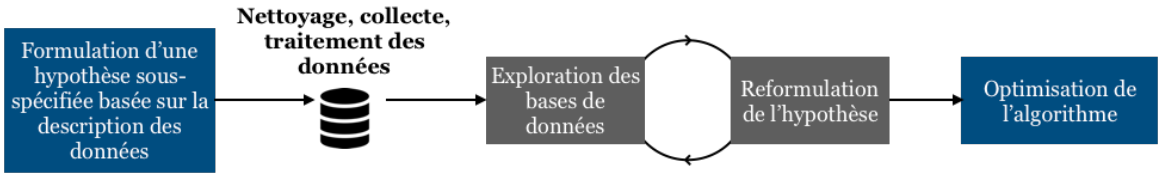


**Figure 50. Processus d'exploration durant le premier challenge**

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.

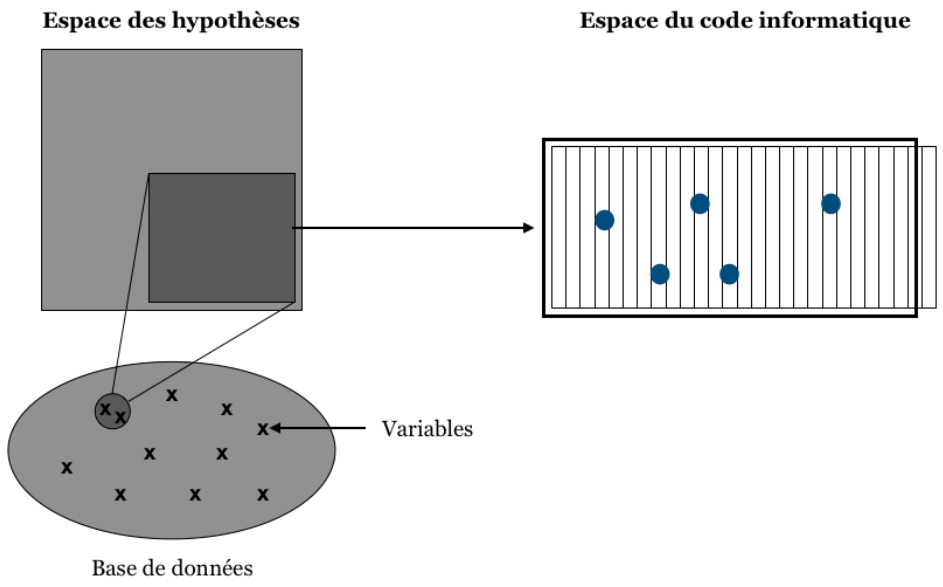


**Figure 51. Formulation d'un axe de travail et exploration des bases de données**

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.

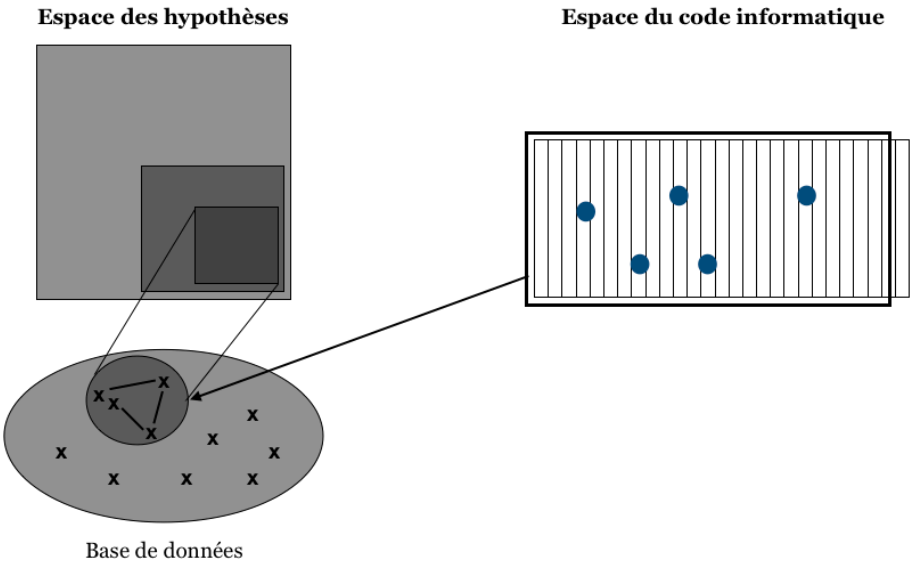


**Figure 52. Reformulation de l'hypothèse à partir des relations qui ont été apprises sur les bases de données.**

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<br>• Conversion en hypothèse canonique<br>• Explorer toutes les données disponibles |
| | Qualité des ressources et extension de l'espace des hypothèses | • Qualité des données<br>• Couvrir toute la littérature |
| **Espace des actions** | Test statistique | • Nombre d'hypothèses vérifiées – Avancement<br>• 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.

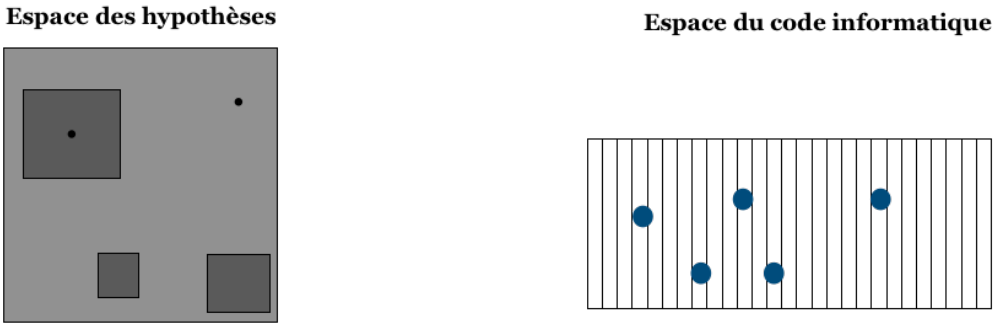


**Figure 53. Illustration de l'exploration des espaces à la fin du challenge.**

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.

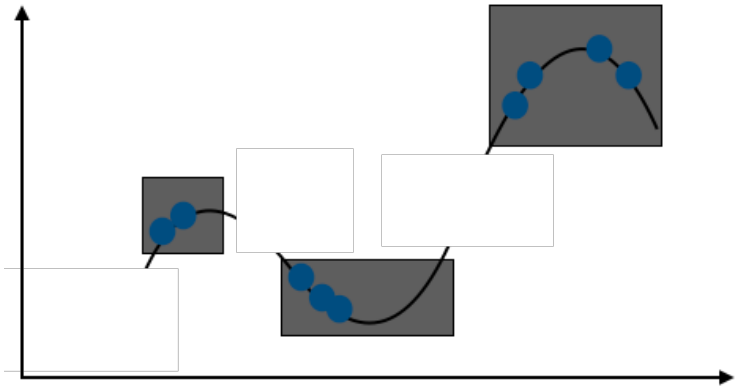

**Figure 54. Les zones explorées par les participants (en gris) donnent des indications sur la fonction de valeur.**

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.