COURS DE DATA MINING

Stéphane TUFFERY

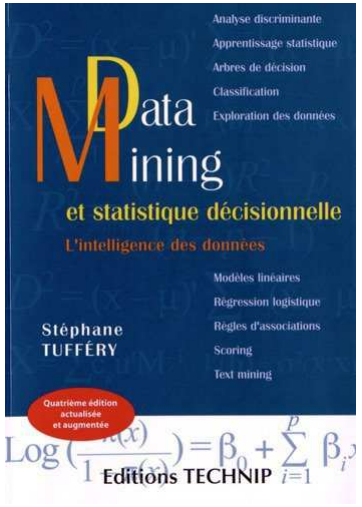
Université Rennes 1

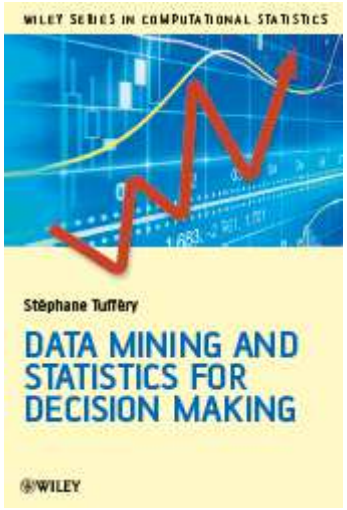
Master 2 Ingénierie économique et financière

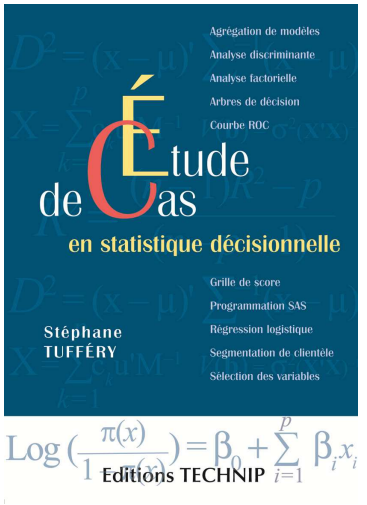
7 février 2014

Présentation de l'intervenant

- Responsable de l'équipe statistique dans un groupe bancaire français
- Enseigne à l'ENSAI et à l'Université Catholique de l'Ouest (Angers)
- Docteur en Mathématiques
- Auteur de :
  - *Data Mining et Statistique Décisionnelle*, Éditions Technip, 2005, 4e édition 2012, préface de Gilbert Saporta
  - *Data Mining and Statistics for Decision Making*, Éditions Wiley, mars 2011
  - *Étude de cas en Statistique Décisionnelle*, Éditions Technip, 2009
  - *Computational Actuarial Science with R* (ouvrage collectif), Éditions Chapman & Hall, 2014







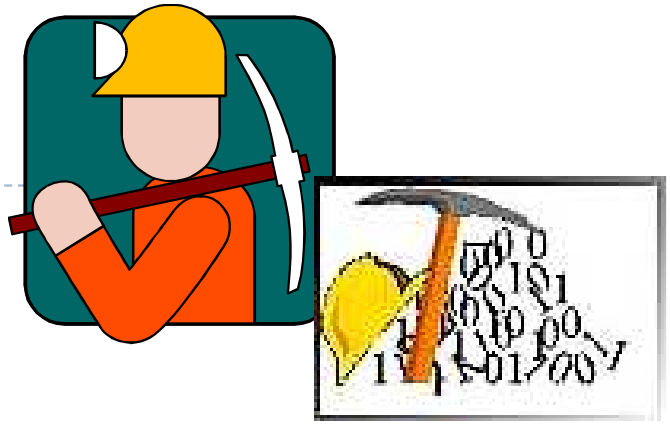
Plan

- Qu'est-ce que le data mining ?
- Qu'est-ce que le Big Data ?
- À quoi sert le data mining ?
- À quoi sert le Big Data ?
- La réforme de Bâle et le ratio de solvabilité
- L'élaboration d'un modèle de scoring

- La sélection des variables
- La modélisation
- Quelques principes du data mining
- L'agrégation de modèles
- Méthodes pour le Big Data
- La détection des règles d'association
- Conclusion

Qu'est-ce que le data mining ?

La fouille de données

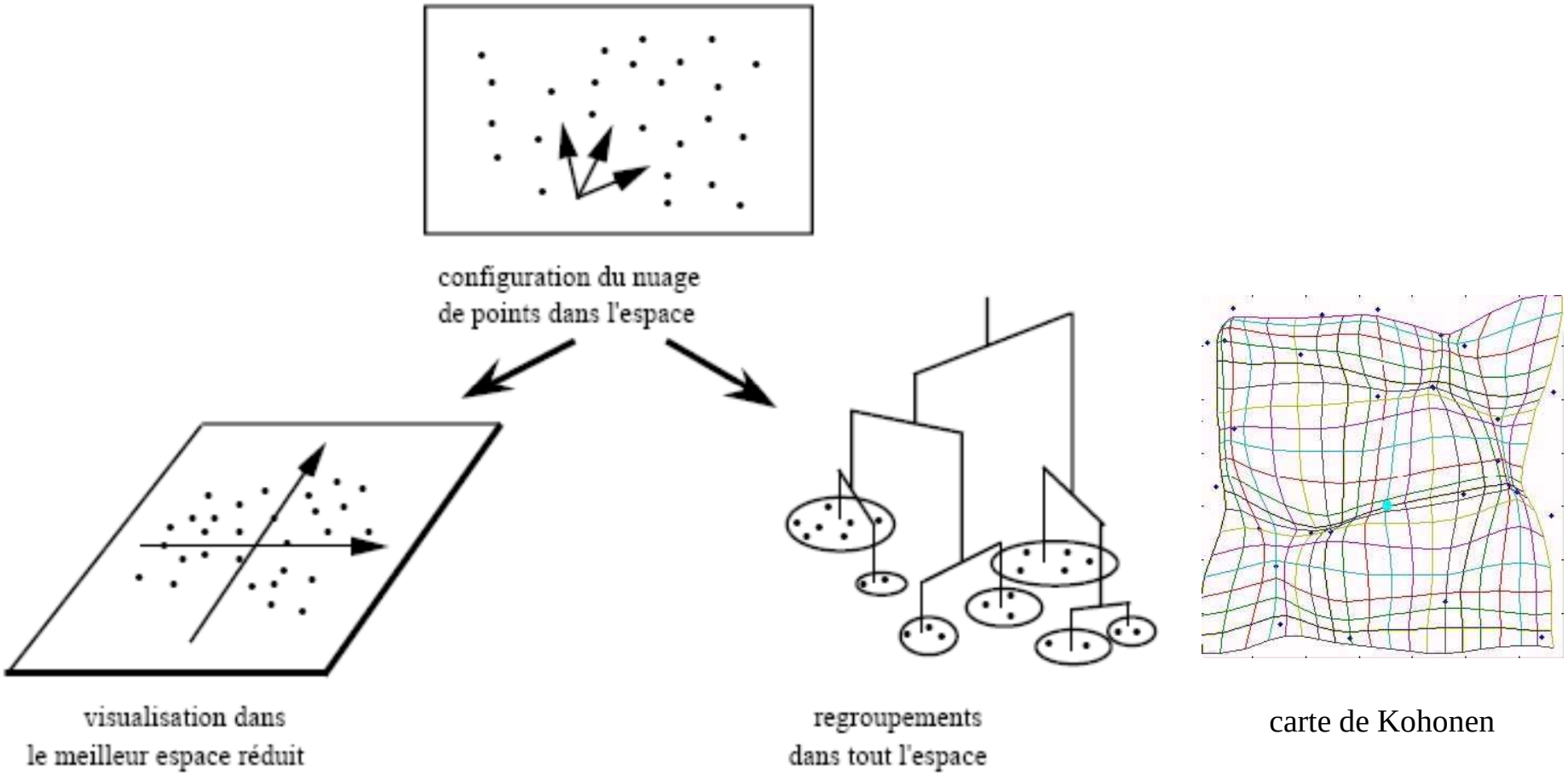


- Le **data mining** est l'ensemble des :
  - méthodes scientifiques
  - … destinées à l'exploration et l'analyse
  - … de (souvent) grandes bases de données informatiques
  - … en vue de détecter dans ces données des profils-type, des comportements récurrents, des règles, des liens, des tendances inconnues (non fixées *a priori*), des structures particulières restituant de façon concise l'essentiel de l'information utile
  - … pour l'aide à la décision
- On parle d'extraire l'information de la donnée
- Selon le MIT, c'est l'une des 10 technologies émergentes qui « changeront le monde » au XXIe siècle

Les 2 types de méthodes de data mining

- Les méthodes descriptives (recherche de « patterns ») :
  - visent à **mettre en évidence des informations présentes** mais cachées par le volume des données (c'est le cas des *segmentations* de clientèle et des *recherches d'associations* de produits sur les tickets de caisse)
  - réduisent, résument, synthétisent les données
  - il n'y a pas de variable à expliquer
- Les méthodes prédictives (modélisation) :
  - visent à **extrapoler de nouvelles informations** à partir des informations présentes (c'est le cas du *scoring*)
  - expliquent les données
  - il y a une variable à expliquer

Les 2 principales familles de méthodes descriptives



Source : Lebart-Morineau-Piron, *Statistique exploratoire multidimensionnelle*, page 10

Qu'est-ce que la classification ?

- Regrouper des objets en groupes, ou classes, ou familles, ou segments, ou *clusters*, de sorte que :
  - 2 objets d'un même groupe se ressemblent le plus possible
  - 2 objets de groupes distincts diffèrent le plus possible
  - le nombre des groupes est parfois fixé
  - les groupes ne sont pas prédéfinis mais déterminés au cours de l'opération
- Méthode descriptive :
  - pas de variable à expliquer privilégiée
  - décrire de façon simple une réalité complexe en la résumant
- Utilisation en marketing, médecine, sciences humaines…
  - segmentation de clientèle marketing
- Les objets à classer sont :
  - des individus
  - des variables
  - les deux à la fois (biclustering)

Complexité du problème !

- Le nombre de partitions (classes non recouvrantes) de *n* objets est le nombre de Bell : $B_n = \frac{1}{e}\sum_{k=1}^{\infty}\frac{k^n}{k!}$
- Exemple : pour *n* = 4 objets, on a B_n = 15, avec
  - 1 partition à 1 classe (abcd)
  - 7 partitions à 2 classes (ab,cd), (ac,bd), (ad,bc), (a,bcd), (b,acd), (c,bad), (d,abc)
  - 6 partitions à 3 classes (a,b,cd), (a,c,bd), (a,d,bc), (b,c,ad), (b,d,ac), (c,d,ab)
  - 1 partition à 4 classes (a,b,c,d)
- Exemple : pour *n* = 30 objets, on a $B_{30} = 8{,}47.10^{23}$
- $B_n > \exp(n) \Rightarrow$ Nécessité de définir des critères de bonne classification et d'avoir des algorithmes performants

Classement et prédiction

- Ce sont des méthodes prédictives
  - on parle aussi d'apprentissage supervisé (réseaux de neurones)
- **Classement** : la variable à expliquer (ou « cible », « réponse », « dépendante ») est *qualitative*
  - on parle aussi de **classification** (en anglais) ou **discrimination**
- **Prédiction** : la variable à expliquer est *quantitative*
  - on parle aussi de **régression**
  - exemple : le prix d'un appartement (en fonction de sa superficie, de l'étage et du quartier)
- **Scoring** : classement appliqué à une problématique d'entreprise (variable à expliquer souvent binaire)
  - chaque individu est affecté à une classe (« risqué » ou « non risqué », par exemple) en fonction de ses caractéristiques

Quelques types de scores

- Score d'appétence
  - prédire l'achat d'un produit ou service
- Score de (comportement) risque
  - prédire les impayés ou la fraude
- Score de pré-acceptation
  - croisement des deux précédents
- Score d'octroi (ou d'acceptation)
  - prédire en temps réel les impayés
- Score d'attrition
  - prédire le départ du client vers un concurrent
- Et aussi :
  - En médecine : diagnostic (bonne santé : oui / non) en fonction du dossier du patient et des analyses médicales
  - Courriels : spam (oui / non) en fonction des caractéristiques du message (fréquence des mots…)

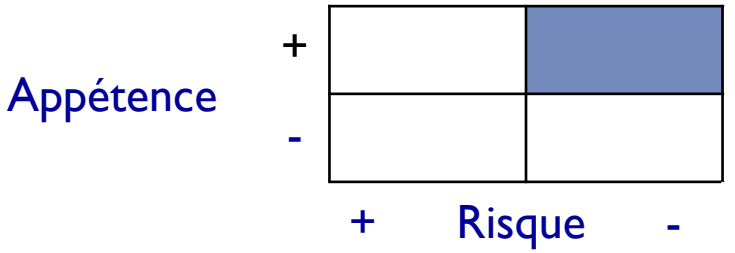


Tableau des méthodes descriptives

type	famille	sous-famille	méthode
méthodes descriptives	modèles géométriques	analyse factorielle (projection sur un espace de dimension inférieure)	analyse en composantes principales ACP (variables continues)
			analyse factorielle des correspondances AFC (2 variables qualitatives)
			analyse des correspondances multiples ACM (+ de 2 var. qualitatives)
		analyse typologique (regroupement en classes homogènes)	méthodes de partitionnement (centres mobiles, *k*-means, nuées dynamiques)
			méthodes hiérarchiques (ascendantes, descendantes)
		analyse typologique + réduction dimens.	classification neuronale (cartes de Kohonen)
	modèles combinatoires		classification relationnelle (variables qualitatives)
	modèles à base de règles logiques	détection de liens	détection d'associations

En grisé : méthodes « classiques »

Tableau des méthodes prédictives

type	famille	sous-famille	méthode
méthodes prédictives	modèles à base de règles logiques	arbres de décision	arbres de décision (variable à expliquer continue ou qualitative)
	modèles à base de fonctions mathématiques	réseaux de neurones	réseaux à apprentissage supervisé : perceptron multicouches, réseau à fonction radiale de base
		modèles paramétriques ou semi-paramétriques	régression linéaire, ANOVA, MANOVA, ANCOVA, MANCOVA, modèle linéaire général GLM, régression PLS, SVR (variable à expliquer continue)
			analyse discriminante linéaire, régression logistique, régression logistique PLS, SVM (variable à expliquer qualitative)
			modèle log-linéaire, régression de Poisson (variable à expliquer discrète = comptage)
			modèle linéaire généralisé, modèle additif généralisé (variable à expliquer continue, discrète ou qualitative)
	prédiction sans modèle		*k*-plus proches voisins (*k*-NN)

En grisé : méthodes « classiques »

Statistique inférentielle et data mining

- Statistique (avant 1950) :
  - quelques centaines d'individus
  - quelques variables recueillies avec un protocole spécial (échantillonnage, plan d'expérience…)
  - fortes hypothèses sur les lois statistiques suivies (linéarité, normalité, homoscédasticité)
  - le modèle prime sur la donnée : il est issu de la théorie et confronté aux données
  - utilisation en laboratoire
- Analyse des données (1960-1980) :
  - quelques dizaines de milliers d'individus
  - quelques dizaines de variables
  - construction des tableaux « Individus x Variables »
  - importance du calcul et de la représentation visuelle
- Data mining (depuis 1990) :
  - plusieurs millions d'individus
  - plusieurs centaines de variables
  - certaines variables non numériques
  - données recueillies avant l'étude, et souvent à d'autres fins
  - données imparfaites, avec des erreurs de saisie, des valeurs manquantes…
  - pour l'aide à la décision
  - nécessité de calculs rapides, parfois en temps réel
  - on ne recherche pas toujours l'optimum théorique, mais le plus compréhensible pour des non statisticiens
  - faibles hypothèses sur les lois statistiques suivies
  - la donnée prime sur le modèle : le modèle est issu des données et on en tire éventuellement des éléments théoriques
  - utilisation en entreprise

Qu'est-ce que le Big Data ?

L'explosion de la production de données

- Données signalétiques et sociodémographiques
- Données de comportement (utilisation du téléphone, de la carte bancaire, du véhicule…)
- Données CRM (contact avec un service client, fidélisation…)
- Données externes provenant des mégabases de données privées ou des administrations (Open Data)
- Informations remontées par les capteurs industriels, routiers, climatiques, puces RFID, NFC, objets connectés (caméras, compteurs électriques, appareils médicaux, voitures…)
- Géolocalisation par GPS ou adresse IP
- Données de tracking sur Internet (sites visités, mots-clés recherchés...)
- Contenu partagé sur Internet (blogs, photos, vidéos…)
- Opinions exprimées dans les réseaux sociaux (sur une entreprise, une marque, un produit, un service…)

Caractérisation des Big Data : les 3 « V »

- Volume
  - L'ordre de grandeur est le pétaoctet (10^{15} octets)
  - L'accroissement du volume vient de l'augmentation :
    - du nombre d'individus observés (plus nombreux ou à un niveau plus fin)
    - de la fréquence d'observation et d'enregistrement des données (mensuel -> quotidien, voire horaire)
    - du nombre de caractéristiques observées
  - Cet accroissement vient aussi de l'observation de données nouvelles, provenant notamment d'Internet : pages indexées, recherches effectuées, éventuellement avec des données de géolocalisation
  - Cet aspect est peut-être le plus visible et le plus spectaculaire, mais il n'est pas le plus nouveau (grande distribution, banque, téléphonie manipulent de grands volumes de données)

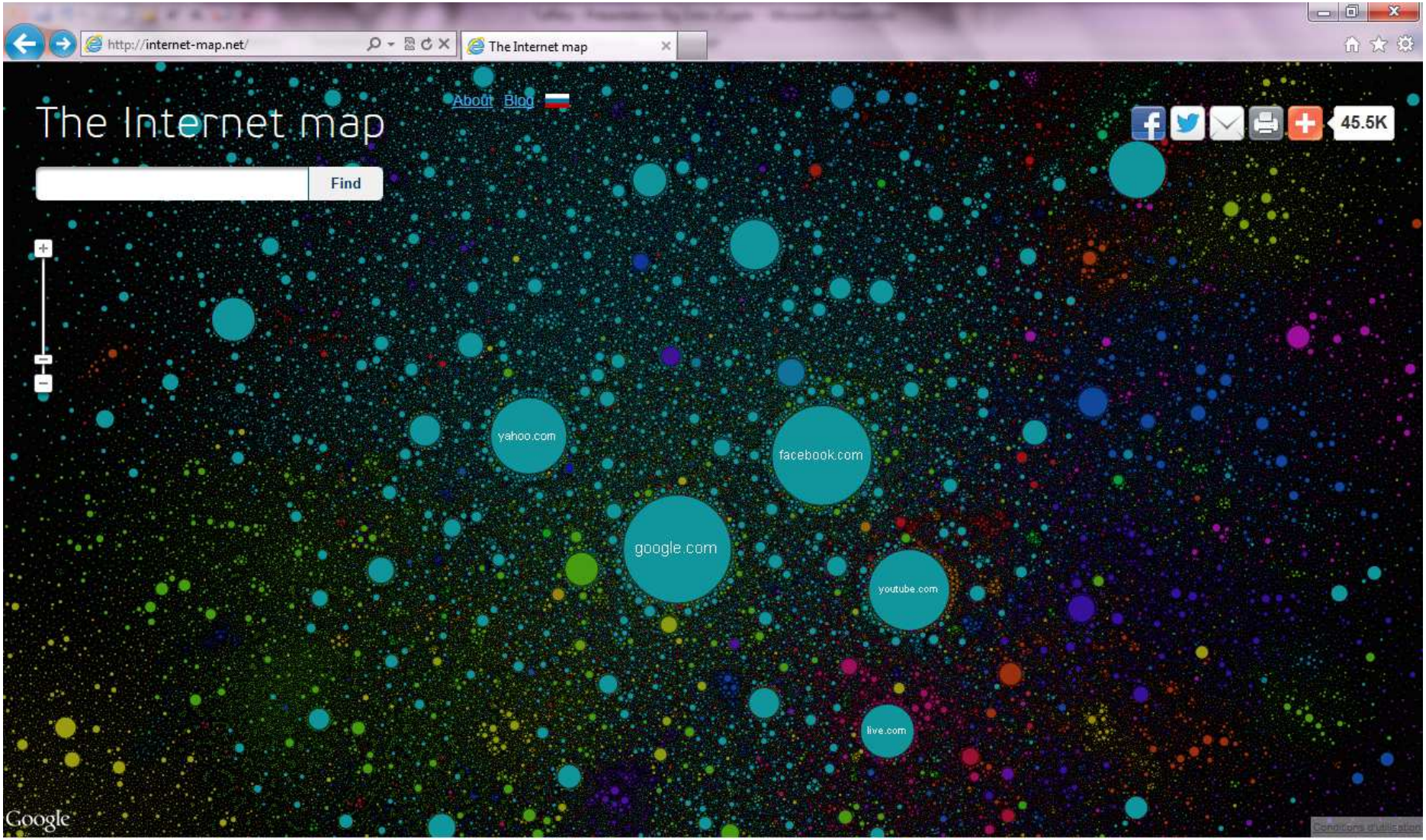
Caractérisation des Big Data : les 3 « V »

- Variété
  - Ces données sont de natures et de formes très diverses : numériques, logs web, textes (Word, PDF, courriels, SMS…), sons, images, données fonctionnelles…
  - Cette variété rend difficile l'utilisation des bases de données usuelles et requiert une variété de méthodes (text mining, web mining…)

Caractérisation des Big Data : les 3 « V »

- Vitesse, ou Vélocité
  - Vélocité des données qui proviennent de sources où elles sont mises à jour rapidement, parfois en temps réel
  - Vitesse des traitements à mettre en œuvre sur ces données
    - La décision du client sur Internet se fait vite car il suffit d'un clic pour changer de site, aussi faut-il instantanément lui faire la meilleure offre commerciale
    - La détection de la fraude par carte bancaire doit bien sûr aussi être instantanée
  - Dans certains cas, vitesse de mise à jour des modèles, et pas seulement vitesse de leur application

Le Big Data d'Internet



Quelques exemples d'utilisations de ces données 1/2

- Transports : fixation dynamique du prix des billets d'avion, amélioration du trafic routier par géolocalisation, recherche de la station-service la plus proche, des places libres de stationnement, facturation dans les zones payantes grâce à la lecture et l'OCR des plaques d'immatriculation…
- Marketing : la géolocalisation permet l'envoi d'une promotion ou d'un coupon sur votre smartphone quand vous passez à proximité d'un commerce, d'une alerte quand vous passez à côté d'une librairie contenant un ouvrage consulté la veille sur Internet, l'analyse des préférences, des recommandations, éventuellement en lien avec les données de vente, permet de mieux cibler les consommateurs
- Grande distribution : analyse des tickets de caisse et croisement avec les données du programme de fidélité
- Ressources humaines : analyse des CV enrichie par la détection des liens noués par le candidat sur les réseaux sociaux
- Scientifiques : météorologie, génomique, épidémiologie, imagerie médicale, astronomie, physique nucléaire…

Quelques exemples d'utilisations de ces données 2/2

- Yield (ou revenue) management :
  - intéresse les activités avec des capacités disponibles limitées (transport, hôtellerie, espaces publicitaires, tourisme…)
  - détermine en temps réel les quantités appropriées à mettre en vente, au prix approprié, de façon à optimiser le profit généré par la vente
  - né dans les années 1980 dans le transport aérien
- Informatique : surveillance des machines et réseaux, et détection de dysfonctionnements ou d'incidents sécuritaires
- Sécurité : vidéo-surveillance, renseignement
- Enseignement : analyse des réseaux sociaux pour connaître la popularité des enseignements et la satisfaction des élèves

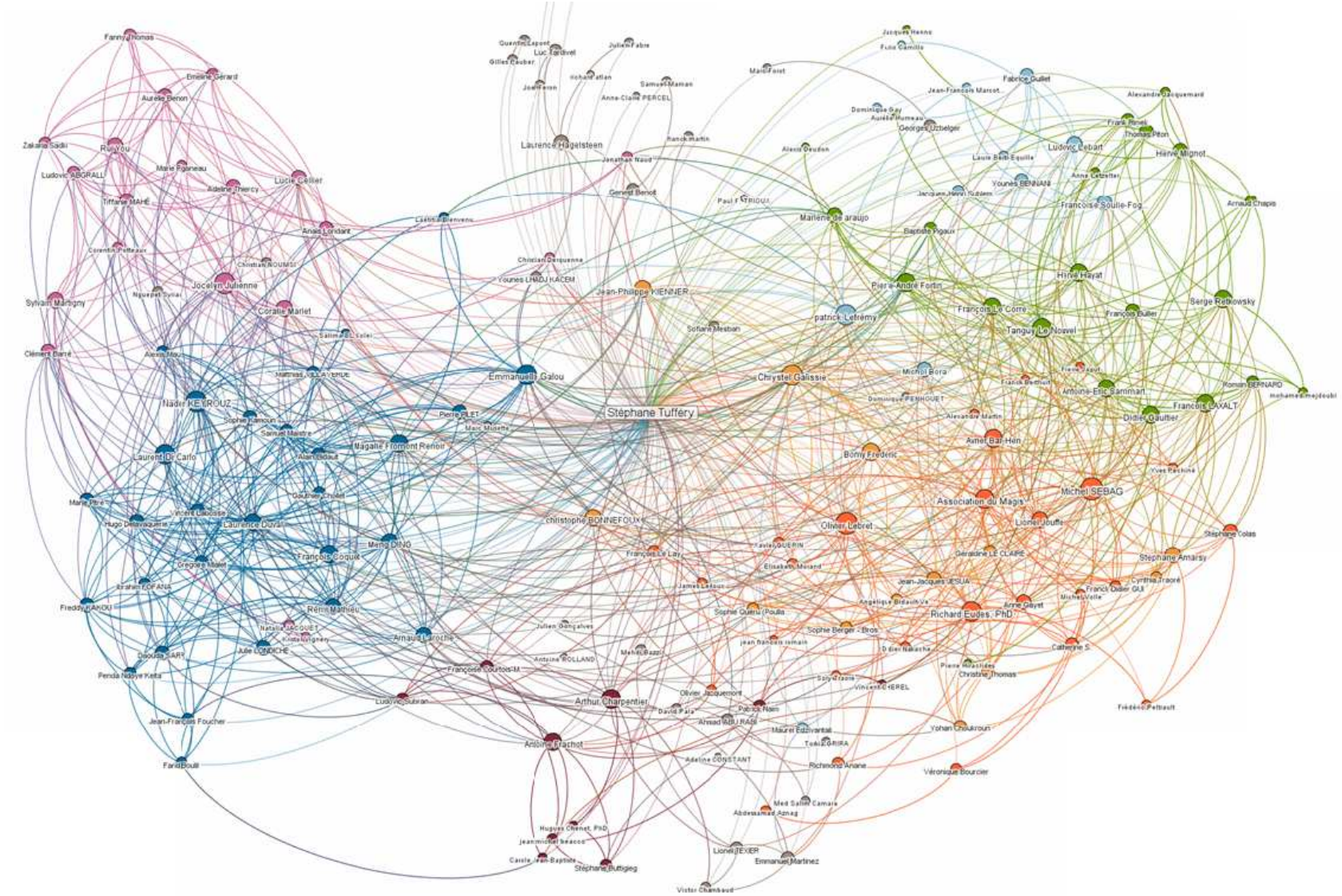
Les réseaux sociaux 1/3

- Un réseau social est un ensemble d'acteurs (individus, groupes ou organisations) reliés par des liens sociaux (familiaux, amicaux professionnels)
- On le représente sous la forme d'un graphe dont les acteurs sont les sommets et les liens sont les arêtes
- On peut étudier le graphe, son nombre de sommets, d'arêtes, sa densité, son diamètre, ses éléments centraux (avec le plus de liens)
- Dans la recherche sur Internet, on peut aussi s'intéresser à des réseaux de sites et regrouper les sites par similarité
- Les sites de commerce en ligne identifient des groupes d'acheteurs en ligne pour formuler des conseils d'achat

Les réseaux sociaux 2/3

- Exemple de LinkedIn InMaps : http://inmaps.linkedinlabs.com/network
- Le graphe représente les connexions des contacts avec une personne et leurs connexions entre eux
- Les connexions de LinkedIn sont utilisées (au 1er et 2e niveau), mais aussi les invitations de connexions, les adresses e-mail, les numéros de téléphone, les messages, les groupes de discussion, le profil (formation, qualifications, postes, recommandations…)
- Les couleurs distinguent les différents groupes détectés (collègues, camarades de promotion, participants à un projet…)
- Des packages graphiques pour les réseaux sociaux existent aussi dans R

Les réseaux sociaux 3/3



A quoi sert le data mining ?

Le data mining dans la banque

- Naissance du score de risque en 1941 (David Durand)
- Multiples techniques appliquées à la banque de détail et la banque d'entreprise
- Surtout la banque de particuliers :
  - grand nombre de dossiers
  - dossiers relativement standards
  - montants unitaires modérés
- Essor dû à :
  - développement des nouvelles technologies
  - nouvelles attentes de qualité de service des clients
  - pression mondiale pour une plus grande rentabilité
  - surtout : ratio de solvabilité Bâle 2

Brève histoire du credit scoring

- 1936 : analyse discriminante de Fisher
- 1941 : utilisation par David Durand pour modéliser le risque de défaut d'un emprunteur à partir de quelques caractéristiques telles que son âge et son sexe
- Après la 2e guerre mondiale : intérêt des entreprises confrontées à une pénurie d'analystes de crédit
- 1958 : développement des ordinateurs et premier système de credit scoring de Fair Isaac
- 1968 : Z-score d'Altman, fonction discriminante de 5 ratios financiers, capable de prévoir à un an la défaillance d'une entreprise, avec une fiabilité d'environ 94 %
- 1998 : premiers travaux sur le ratio de solvabilité Bâle 2

Le data mining dans l'assurance de risque

- Des produits obligatoires (automobile, habitation) :
  - soit prendre un client à un concurrent
  - soit faire monter en gamme un client que l'on détient déjà
- D'où les sujets dominants :
  - attrition
  - ventes croisées (*cross-selling*)
  - montées en gamme (*up-selling*)
- Besoin de décisionnel dû à :
  - concurrence des nouveaux entrants (bancassurance)
  - bases clients des assureurs traditionnels mal organisées :
    - compartimentées par agent général
    - ou structurées par contrat et non par client

Le data mining dans la téléphonie

- Deux événements :
  - fin du monopole de France Télécom dans la téléphonie fixe
  - arrivée à saturation du marché de la téléphonie mobile
- D'où les sujets dominants dans la téléphonie :
  - score d'attrition (*churn* = changement d'opérateur)
  - optimisation des campagnes marketing
    - et aussi le *text mining* (pour analyser les lettres de réclamation)
- Problème du *churn* :
  - coût d'acquisition moyen en téléphonie mobile : 250 euros
  - plus d'un million d'utilisateurs changent chaque d'année d'opérateur en France
  - les lois facilitant le changement d'opérateur
  - la portabilité du numéro facilite le churn

Le data mining dans le commerce

- Vente Par Correspondance
  - utilise depuis longtemps des scores d'appétence
  - pour optimiser ses ciblages et en réduire les coûts
  - des centaines de millions de documents envoyés par an
- e-commerce
  - personnalisation des pages du site web de l'entreprise, en fonction du profil de chaque internaute
  - optimisation de la navigation sur un site web
- Grande distribution
  - analyse du ticket de caisse
  - détermination des meilleures implantations (géomarketing)

Autres exemples

- De l'infiniment petit (génomique) à l'infiniment grand (astrophysique pour le classement en étoile ou galaxie)
- Du plus quotidien (reconnaissance de l'écriture manuscrite sur les enveloppes) au moins quotidien (aide au pilotage aéronautique)
- Du plus ouvert (e-commerce) au plus sécuritaire (détection de la fraude dans la téléphonie mobile ou les cartes bancaires)
- Du plus industriel (contrôle qualité pour la recherche des facteurs expliquant les défauts de la production) au plus théorique (sciences humaines, biologie…)
- Du plus alimentaire (agronomie et agroalimentaire) au plus divertissant (prévisions d'audience TV)

A quoi sert le Big Data ?

Le Big Data dans le marketing

- L'analyse des réseaux sociaux, des forums et des moteurs de recherche permet de découvrir les centres d'intérêt et les préférences des internautes, et donc leur comportement possible face à une proposition de produit ou de service
- C'est particulièrement utile pour les entreprises qui font du B to B to C, ont des contacts avec des distributeurs et non leurs clients finaux, sur lesquels elles ont peu d'informations directes
- L'analyse des réseaux sociaux n'est pas seulement utile à la vente et elle peut aider à la conception de nouveaux produits, par l'analyse de la perception positive ou négative de certaines caractéristiques des produits, et la comparaison avec la concurrence
- Des packages R existent pour traiter les données de Twitter et Facebook

Le Big Data dans la finance

- Risque boursier
  - Une étude parue dans *Nature* (2013) démontre une corrélation entre les mots clés saisis sur Google et l'évolution des cours de bourse. Avant une chute des indices boursiers, les investisseurs sont préoccupés et recherchent sur Internet des informations les aidant à décider de conserver ou vendre leurs titres.
- Risque financier
  - Ce que l'on dit d'une entreprise, son image chez ses partenaires, les analystes financiers ou le grand public, sa réputation, son image en termes de qualité, d'innovation, de respect social et environnemental… ces éléments peuvent concourir à sa santé financière à moyen/long terme et peuvent être intégrés dans les analyses
- Risque de fraude
  - Les données de géolocalisation des détenteurs de smartphones peuvent être comparées aux informations relatives au terminal de paiement pour s'assurer qu'elles sont cohérentes

Le Big Data dans l'assurance

- Aviva a mis au point une application pour smartphone (Aviva Drive) qui analyse le style de conduite des conducteurs afin de leur proposer des tarifs appropriés (http://www.aviva.co.uk/drive/)
- Un projet similaire avait été imaginé en 2006 mais abandonné en 2008 en raison de la difficulté d'installer des « boîtes noires » dans les véhicules
- Cette application analyse pendant 300 km le nombre de kilomètres parcourus, le temps, le type de route…
- Un changement radical de comportement pourra faire suspecter une fraude
- Des capteurs sur la voiture pourraient même signaler des risques de panne, indiquant au conducteur la conduite à tenir et le garage le plus proche

Le Big Data dans l'industrie

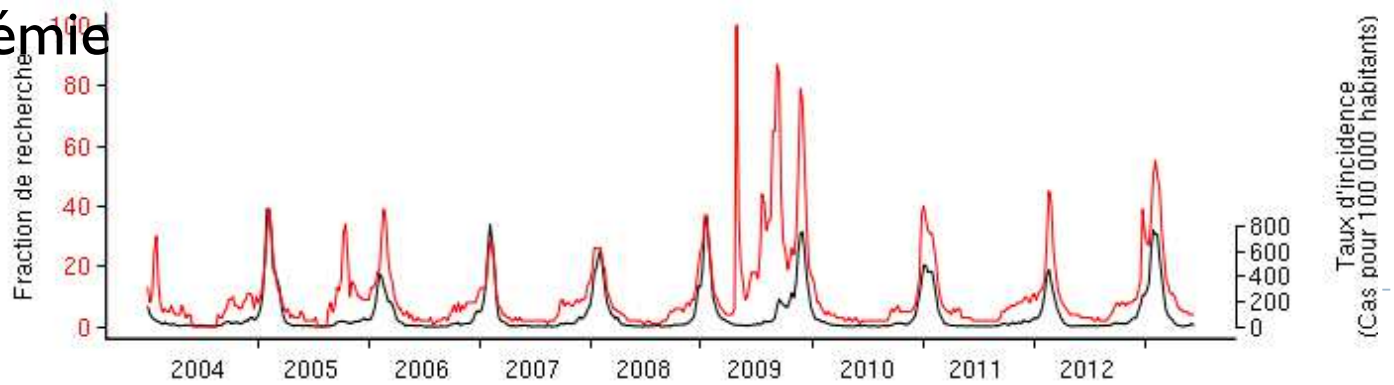
- Les nombreux capteurs (température, pression, vibration, usure…) placés sur les composants de l'appareil productif permettent de remonter en temps réel et à distance de nombreuses informations qui, analysées et modélisées, peuvent fournir une probabilité de défaillance, de rupture d'une pièce, et permettre un arbitrage entre :
  - Des opérations de maintenance inutilement lourdes et fréquentes, entraînant des dépenses inutiles
  - Des opérations de maintenance insuffisantes et laissant se produire des défaillances coûteuses, voire dangereuses
- Optimisation de la chaîne d'approvisionnement (supply chain)
- Prédiction en temps réel de la consommation électrique, mais aussi des dysfonctionnements, et facturation plus économique et plus rapide, grâce aux compteurs connectés (Linky)

Le Big Data dans la santé 1/2

- Diagnostic médical à distance : détection de risques de crise cardiaque
- Des applications pour smartphones savent analyser les données transmises par des capteurs (rythme cardiaque, pression sanguine…)
- Monitoring des grands prématurés : analyse en temps réel des données fournies par des capteurs placés sur les bébés
- Génomique :
  - Liens entre données génomiques et apparition d'une maladie ou réponse à un traitement
  - Génomique des populations

Le Big Data dans la santé 2/2

- En analysant les mots clés sur son moteur de recherche, Google a pu établir une corrélation entre certaines requêtes et l'apparition d'une épidémie de grippe. Cette corrélation a été corroborée par les organismes de veille sanitaire et a fait l'objet d'une publication dans *Nature* (2009).
  - Voir : http://www.google.org/flutrends/intl/en_us/about/how.html et http://websenti.u707.jussieu.fr/sentiweb/?page=google
- Cet exemple illustre le V de la vitesse, avec des mises à jour de données quotidiennes et non hebdomadaires comme dans les suivis traditionnels : permet une détection plus rapide de l'épidémie



Le Big Data dans la statistique publique

- L'analyse des messages Twitter aux Pays-Bas a montré une corrélation entre les sentiments exprimés et l'indice public de confiance des ménages
- Twitter a succédé aux médias classiques dans les analyses classiques en sciences humaines sur les discours, l'opinion…
- Les journalistes de Bloomberg intègrent aussi les données de Twitter
- D'autres données peuvent aussi être utiles : tickets de caisse et calcul du taux d'inflation, sites de recherche d'emploi et estimation du taux de chômage…
- Ces exemples illustrent l'apport possible des analyses privées de Big Data à la statistique publique, avec des indicateurs équivalents mais calculés bien plus rapidement et peut-être, du moins à terme, à moindre coût

La réforme de Bâle et le ratio de solvabilité

Les principaux types de risques financiers

- Crédit : risque que l'emprunteur ne rembourse pas sa dette à l'échéance fixée
  - Nombreuses méthodes statistiques développées depuis 1941, surtout des modèles binaires dont la variable à expliquer est le défaut de remboursement
  - Passage de Bâle I à Bâle II d'une approche forfaitaire à une approche de rating
- Marché : risque que la valeur d'un actif (d'une dette) détenu(e) par une institution financière varie en raison de l'évolution des prix sur les marchés financiers
  - Modèles économétriques
- Opérationnel : risque de pertes directes ou indirectes résultant d'une inadéquation ou d'une défaillance attribuable à des procédures, des personnes, des systèmes internes ou à des événements extérieurs
  - Introduit dans la réforme du ratio de solvabilité Bâle II
  - Inclut le risque juridique mais exclut le risque stratégique
  - Méthodes probabilistes et à dire d'expert

Autres types de risques financiers

- De liquidité : risque de ne pouvoir vendre un actif suffisamment rapidement pour éviter une perte par rapport au prix qu'on aurait dû obtenir
  - Bâle III demande aux banques de détenir un stock d'actifs sans risque et facilement négociables (cash, titres d'État…) lui permettant de résister pendant 30 jours à une crise de liquidité
- De réputation : risque résultant d'une perception négative de la part des clients, des contreparties, des actionnaires, des investisseurs ou des régulateurs qui peut affecter défavorablement la capacité d'une banque à maintenir ou engager des relations d'affaires et la continuité de l'accès aux sources de financement
- De taux : risque de déséquilibre entre les taux des emplois et les taux des ressources
- De change : risque lié aux activités en devise
- Stratégique
- …

Le ratio de solvabilité Bâle I

- La solvabilité d'une banque est sa capacité à rembourser ses dettes
- 1988 : instauration d'un « ratio Cooke » visant à :
  - Renforcer la solidité et la stabilité du système bancaire international
  - Promouvoir des conditions d'égalité de concurrence entre les banques à vocation internationale
- Ce ratio de 8% est le rapport entre les encours pondérés et le montant des fonds propres de la banque
  - Ratio de 4% pour les fonds propres Tier 1
- Les crédits sont pondérés selon la catégorie d'actifs considérée (0% pour les Souverains, 20% pour les Banques, 50% pour l'immobilier hypothécaire, 100% pour le reste) mais non selon la qualité de la signature
- Au risque de crédit est ajouté le risque de marché en 1996

Le ratio de solvabilité Bâle II

- 2004 : accords Bâle II
- Trois piliers
  - Pilier 1 : exigences minimales en fonds propres
  - Pilier 2 : couverture des risques non pris en compte dans le pilier 1
  - Pilier 3 : transparence et la discipline de marché.
- Pilier 1 : instauration d'un nouveau « ratio Mc Donough »
  - toujours égal à 8%
  - mais diversifie les risques pris en compte (en incluant les risques opérationnels)
  - et affine la méthode de pondération des risques, notamment en autorisant l'utilisation de systèmes (« notations internes ») de classification des emprunteurs à partir des probabilités de défaillance prédites dans les différents types de portefeuille de la banque : souverains, banques, entreprises, banque de détail (particuliers et professionnels), titres, titrisation et autres

L'accord Bâle III

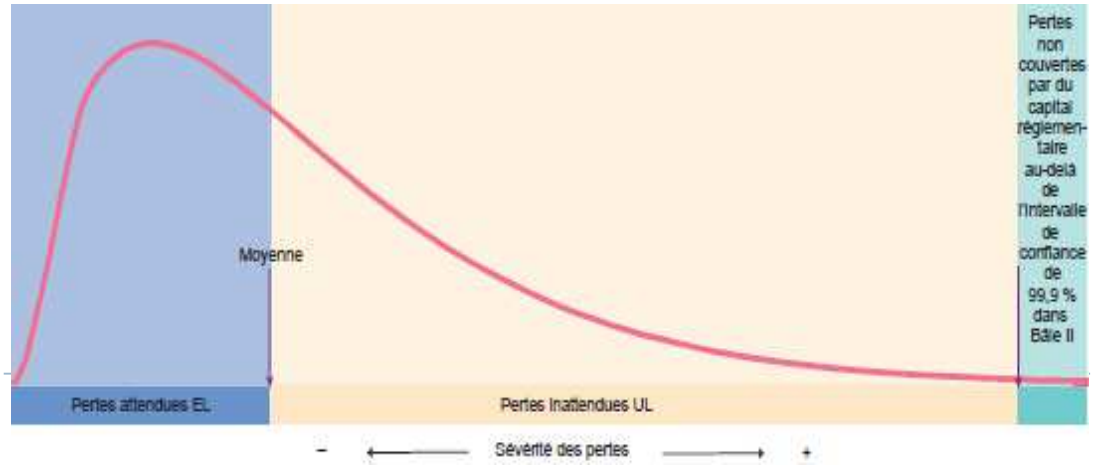
- 2010 : recommandations Bâle III
- Liquidité :
  - Instauration d'un ratio de liquidité LCR (Liquidity Coverage Ratio) à 30 jours et d'un ratio NSFR (Net Stable Funding Ratio) à un an
- Fonds propres :
  - Renforcement de la qualité et du niveau des fonds propres
  - Mise en place d'un coussin de conservation alimenté dans les périodes favorables
  - Surcharge systémique pour les établissements les plus importants
  - Instauration d'un ratio d'effet de levier (ratio « fonds propres / total des actifs non pondérés ») > 3%

Le risque de crédit

- Les modèles de scoring permettent d'attribuer une probabilité de défaut (PD) de paiement à toute entité notée, sur un horizon donné
- La perte encourue par la banque dépend de deux autres facteurs :
  - EAD (Exposure At Default) : montant du crédit exposé si l'emprunteur passe en défaut (encours bilan + CCF x encours hors-bilan)
  - CCF (Credit Conversion Factor) : part de l'encours hors-bilan qui sera utilisée par l'emprunteur au moment du défaut
  - LGD (Loss Given Default) : taux de perte (y compris frais de recouvrement) subi par la banque (après activation des éventuelles garanties) en cas de défaut de l'emprunteur
- Un établissement bancaire peut avoir une approche Bâle II :
  - Standard (application de pondérations forfaitaires à l'encours exposé)
  - Interne « fondation » (IRBF) : estimation par l'établissement de la PD, le CCF et la LGD étant forfaitaires
  - Interne « avancée » (IRBA) : estimation par l'établissement de tous les paramètres
- Utilisation possible pour le calcul de l'exigence en fonds propres sous réserve d'une validation indépendante par l'autorité de tutelle

Pertes attendues et inattendues

- Pertes attendues (EL : expected losses)
  - Pertes annuelles moyennes : EAD x PD x LGD
  - Doivent être couvertes par les provisions et éventuellement par des fonds propres
- Pertes inattendues (UL : unexpected losses)
  - VaR = pertes annuelles si élevées qu'elles ne sont possibles qu'une fois sur 1000 : EAD x f(PD) x LGD
  - UL = VaR - EL
  - Doivent être couvertes par les fonds propres réglementaires



Calcul de l'exigence en fonds propres

- Actifs pondérés : RWA (risk weighted assets)
  - 12,5 x EAD x (f(PD) - PD) x LGD pour le risque de crédit
- Exigence en fonds propres (couvrir les pertes inattendues)
  - EFP = 8% (RWA + 12,5 x capital risqué au titre du risque de marché + 12,5 x capital risqué au titre du risque opérationnel)
- Rappel : Exigence en fonds propres Bâle I
  - EFP = 8% x Actifs pondérés Cooke
  - Actifs pondérés Cooke = encours crédit x pondération

Nature du risque	Pondération
Souverain	0 %
Banques	20 %
Immobilier	50 %
Autres crédits	100 %

Bâle II : pondérations en méthode standard

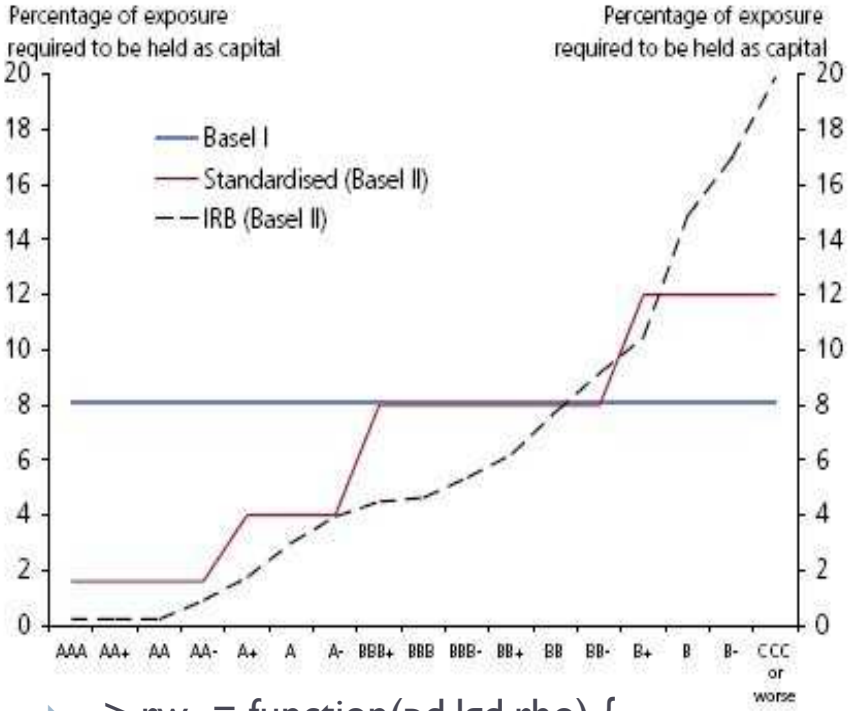
- Même méthode que Bâle I avec une pondération des expositions fixée par le texte et affinée :

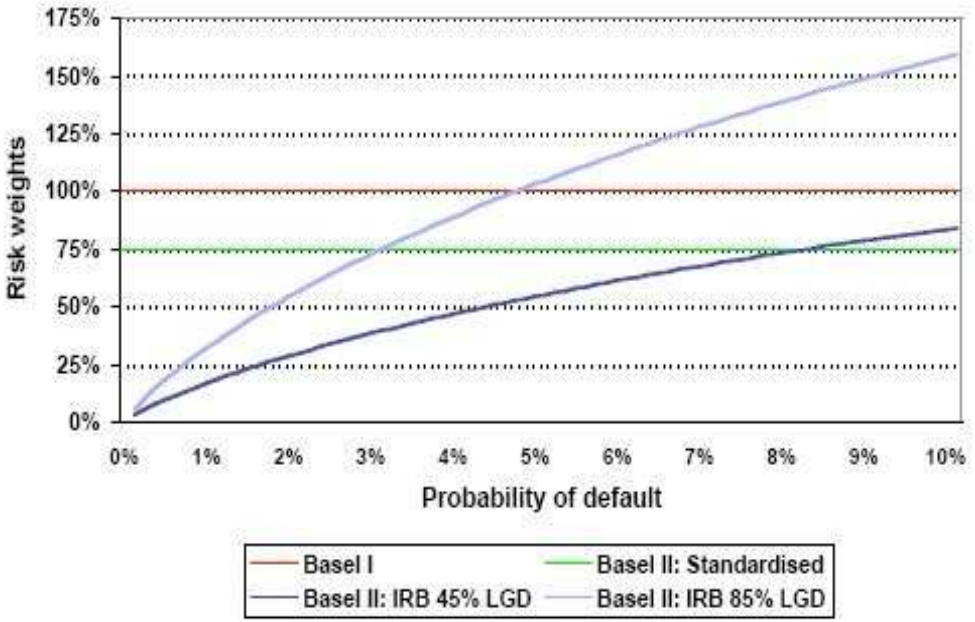
	Notations externes						
Contreparties	**AAA à AA-**	**A+ à A-**	**BBB+ à BBB-**	**BB+ à B-**	**Inférieur à B -**	**Non noté**	
Souverains	0 %	20 %	50 %	100%	150%	100%	
Banques	20%	50%	50%	100%	150%	50%	
Entreprises	20%	50%	100%	jusqu'à BB- : 100%	<BB - : 150%	100%	
Retail Immobilier						35 %	
Retail Autres						75 %	

Pondération des risques de crédit

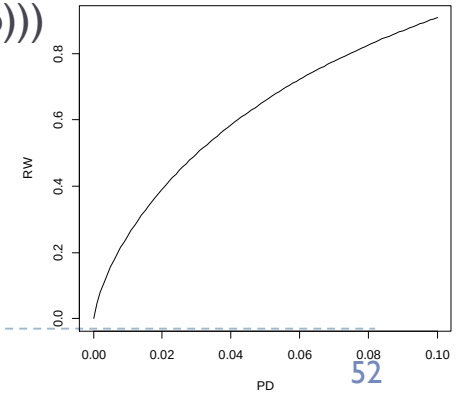
- Pondération des risques
  - RW = 12,5 x (f(PD) - PD) x LGD pour le risque de crédit
  - EFP = 8% x RW x EAD
- Cette pondération RW est à comparer :
  - Au tableau précédent pour la méthode standard Bâle II (par exemple, 75% pour le Retail Autre)
  - Aux valeurs pour Bâle I : 100 % Autres Crédits, 50% Immobilier…
- Exemple de calcul en R : RW d'un crédit habitat avec une PD = 3% et une LGD à 20%
  - > ead <- 100
  - > pd <- 0.03
  - > lgd <- 0.2
  - > rho <- 0.15 # immobilier mortgage
  - > fpd <- pnorm(qnorm(pd)*sqrt(1/(1-rho)) + qnorm(0.999)*sqrt(rho/(1-rho)))
  - > (rw <- 12.5*lgd*(fpd-pd))
  - [1] 0.4977229
- C'est quasiment le RW = 50% de Bâle I et plus que le 35% Standard

Comparaison des approches





- > rw = function(pd,lgd,rho) {
- + fpd <- pnorm(qnorm(pd)*sqrt(1/(1-rho)) + qnorm(0.999)*sqrt(rho/(1-rho)))
- + rw <- 12.5*lgd*(fpd-pd)
- + return(rw)
- + }
- > rpd <- seq(0,0.1,by=.001)
- > plot(rpd,rw(rpd,lgd,rho),type="l",xlab="PD",ylab="RW")



Calcul des actifs pondérés

Corporate exposure[10]

Correlation (R) = $0.12 \times (1 - EXP(-50 \times PD)) / (1 - EXP(-50)) + 0.24 \times [1 - (1 - EXP(-50 \times PD))/(1 - EXP(-50))]$

Maturity adjustment (b) = $(0.11852 - 0.05478 \times \ln(PD))^2$

Capital requirement[68] (K) = $[LGD \times N[(1 - R)^{-0.5} \times G(PD) + (R / (1 - R))^{0.5} \times G(0.999)] - PD \times LGD] \times (1 - 1.5 \times b)^{-1} \times (1 + (M - 2.5) \times b)$

Risk-weighted assets (RWA) = $K \times 12.5 \times EAD$

Corporate exposure adjusted for SME[11]

Correlation (R) = $0.12 \times (1 - EXP(-50 \times PD)) / (1 - EXP(-50)) + 0.24 \times [1 - (1 - EXP(-50 \times PD))/(1 - EXP(-50))] - 0.04 \times (1 - (S-5)/45)$

Residential mortgage exposure[12]

Correlation (R) = 0.15

Capital requirement (K) = $LGD \times N[(1 - R)^{-0.5} \times G(PD) + (R / (1 - R))^{0.5} \times G(0.999)] - PD \times LGD$

Risk-weighted assets = $K \times 12.5 \times EAD$

Qualifying revolving retail exposure[13] **(credit card product)**

Correlation (R) = 0.04

Capital requirement (K) = $LGD \times N[(1 - R)^{-0.5} \times G(PD) + (R / (1 - R))^{0.5} \times G(0.999)] - PD \times LGD$

Risk-weighted assets = $K \times 12.5 \times EAD$

S =

Min(Max(SalesTurnover),5),50

[10] Function is taken from paragraph 272

[11] Function is taken from paragraph 273

[12] Function is taken from paragraph 328

[13] Function is taken from paragraph 329

In Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework (BCBS) (November 2005 Revision)

Risques opérationnels : matrice « lignes de métier x types de risque »

	EL1 - Fraude interne	EL2 - Fraude externe	EL3 - Emploi et sécurité	EL4 - Pratiques commerciales	EL5- Dommages aux actifs corporels	EL6 - Dysfonctionnements des systèmes	EL7- Exécution des processus	TOTAL
BL1 - Opérations financières								
BL2 - Opérations de marché								
BL3 - Banque de Détail								
BL4 -Banque commerciale								
BL5 - Paiements et Règlements								
BL6 - Traitement des Titres								
BL7 - Gestion d'Actifs								
BL8 - Courtage								
TOTAL								

Traitement des risques opérationnels

- Certains risques ne sont « que » potentiels : leur probabilité de survenance est très faible mais leur gravité très grande : ce sont les risques de gravité, pour lesquels on élabore des scénarii avec les experts (on obtient des expositions, des gravités et des probabilités de survenance conditionnées par des facteurs appelés KRI)
- D'autres risques sont plus fréquents mais leur gravité plus faible : ce sont les risques de fréquence, pour lesquels on recherche des ajustements sur les historiques de pertes unitaires de lois théoriques pour la survenance (loi de Poisson) et la gravité (loi log-normale, de Weibull…)
- Puis simulations très nombreuses de sinistres selon les paramètres établis, calcul de pertes unitaires puis cumulées sur 1 an, et obtention de la moyenne (= EL) et du quantile à 99,9 % (= VaR = UL + EL) de la perte cumulée, c'est-à-dire de la perte pouvant survenir 1 fois sur 1000, soit 1 fois tous les 1000 ans

Approches des risques opérationnels

- Contrairement au risque de crédit, pas de relation simple entre les pertes attendues et inattendues : ces dernières peuvent « exploser », surtout si la survenance d'un risque est liée à la survenance d'un autre risque
- Comme pour le risque de crédit, trois niveaux d'approche pour l'exigence en fonds propres :
  - De base : un pourcentage du PNB moyen des 3 années précédentes (fixé à 15%)
  - Standard : identique à l'approche de base, mais pourcentages différenciés (entre 12% et 18%) par lignes de métiers (voir matrice précédente)
  - Avancée (AMA) : l'établissement détermine lui-même ses besoins en fonds propres par des modèles internes

Politique du risque et gestion dans la banque

- Politique du risque dans la banque
  - Enjeux majeurs en termes financiers, opérationnels, réglementaires et d'image
    - Sous le contrôle permanent de l'ACPR : Autorité de Contrôle Prudentiel et de Résolution
  - Intrication de ces enjeux
- Appréhension dans l'entreprise
  - Nécessaire appropriation de la politique du risque par tous les acteurs
  - Est évaluée par l'autorité de tutelle
  - C'est à la fois une condition et une conséquence de la qualité des outils de maîtrise de risque : un cercle vertueux à mettre en place
- Les modèles de risque doivent être :
  - Bien conçus
  - Bien mis en œuvre
  - Bien suivis (outils à mettre en place – tableaux de bord – procédures)
  - Bien appropriés
- Gouvernance du risque
  - Les outils ne suffisent pas : il faut gouverner leur utilisation, leur suivi et leur évolution

L'élaboration d'un modèle de scoring

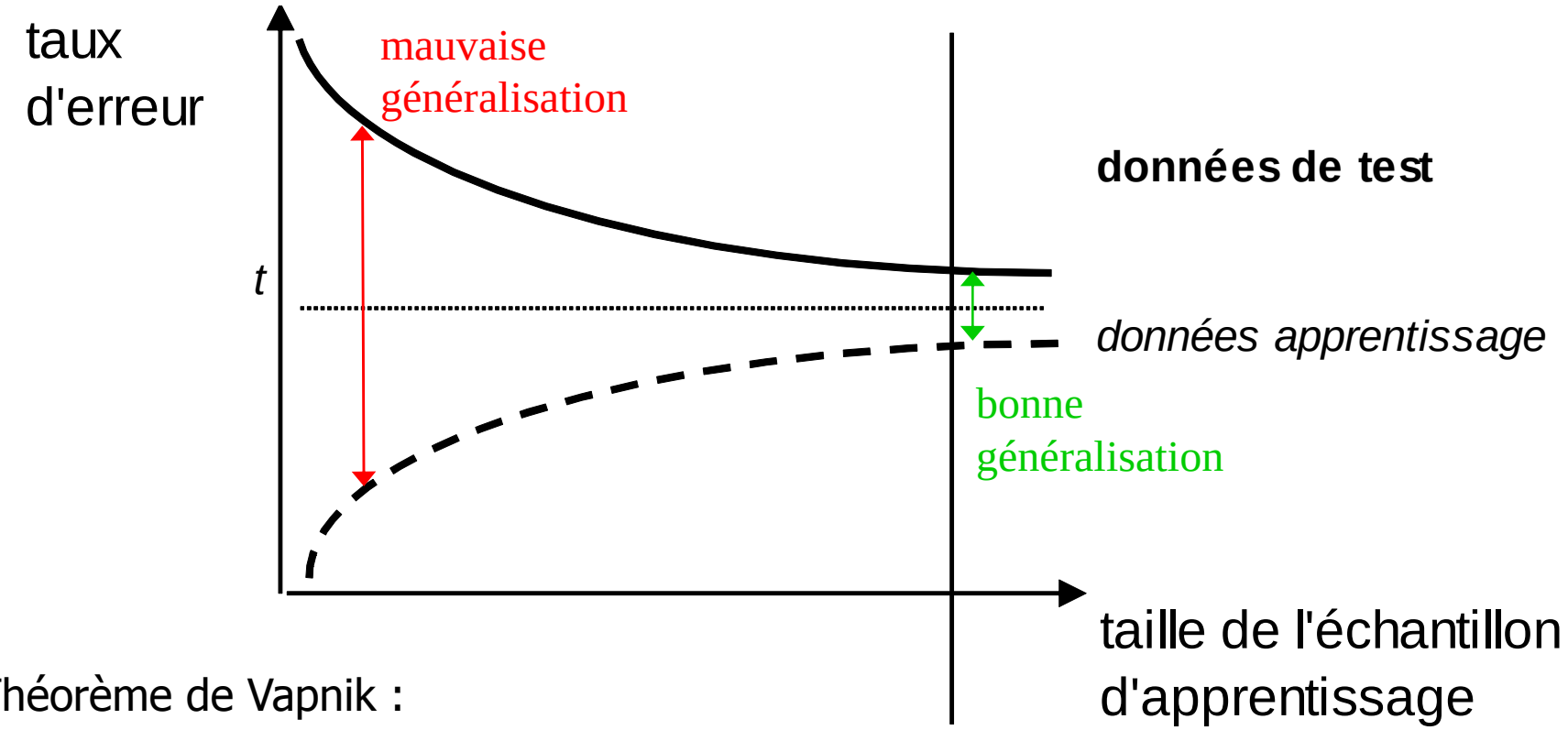
Définition de la variable à expliquer

- En médecine : définition souvent naturelle
  - un patient a ou non une tumeur (et encore faut-il distinguer les différents stades d'une tumeur)
- Dans la banque : qu'est-ce qu'un client non risqué ?
  - aucun impayé, 1 impayé, *n* impayés mais dette apurée ?
- Dans certains modèles, on définit une « zone indéterminée » non modélisée :
  - 1 impayé $\Rightarrow$ variable à expliquer non définie
  - aucun impayé $\Rightarrow$ variable à expliquer = 0
  - $\geq$ 2 impayés $\Rightarrow$ variable à expliquer = 1 ($\geq$ 3 impayés pour Bâle 2)
- Définition parfois encore plus problématique en attrition
  - dans la banque, contrairement à la téléphonie ou l'assurance, on peut partir brutalement ou progressivement

Biais de sélection

- En risque : certaines demandes sont refusées et on ne peut donc pas mesurer la variable à expliquer
  - certaines populations ont été exclues de la modélisation et on leur applique pourtant le modèle
  - il existe des méthodes « d'inférence des refusés », mais dont aucune n'est totalement satisfaisante
  - et parfois aucune trace n'est conservée des demandes refusées !
- En appétence : certaines populations n'ont jamais été ciblées et on ne leur a pas proposé le produit
  - si on les modélise, elles seront présentes dans l'échantillon des « mauvais » (clients sans appétence) peut-être à tort
  - contrairement au cas précédent, on peut mesurer la variable à expliquer car il y a des souscriptions spontanées
  - envisager de limiter le périmètre aux clients ciblés
- Fraude à la carte bancaire : certaines transactions ont été rejetées et on ne sait pas toujours si elles étaient frauduleuses

Taille de l'échantillon



Théorème de Vapnik :

$$R < R_{emp} + \sqrt{\frac{h\,(\log(2n/h)+1) - \log(\alpha/4)}{n}}$$

Représentativité de l'échantillon d'étude

- Hypothèse fondamentale :
  - l'échantillon d'étude est représentatif de la population à laquelle sera appliqué le modèle
- N'implique pas un échantillonnage aléatoire simple :
  - événement à prédire rare $\Rightarrow$ stratification non proportionnelle de l'échantillon sur la variable à expliquer
  - parfois : 50 % de positifs et 50 % de négatifs
  - nécessaire quand on utilise CART pour modéliser 3 % de positifs, sinon CART prédit que personne n'est positif $\Rightarrow$ excellent taux d'erreur = 3 % !
  - change la constante du logit de la régression logistique
  - intéressant en cas d'hétéroscédasticité dans une analyse discriminante linéaire

Inventaire des données utiles

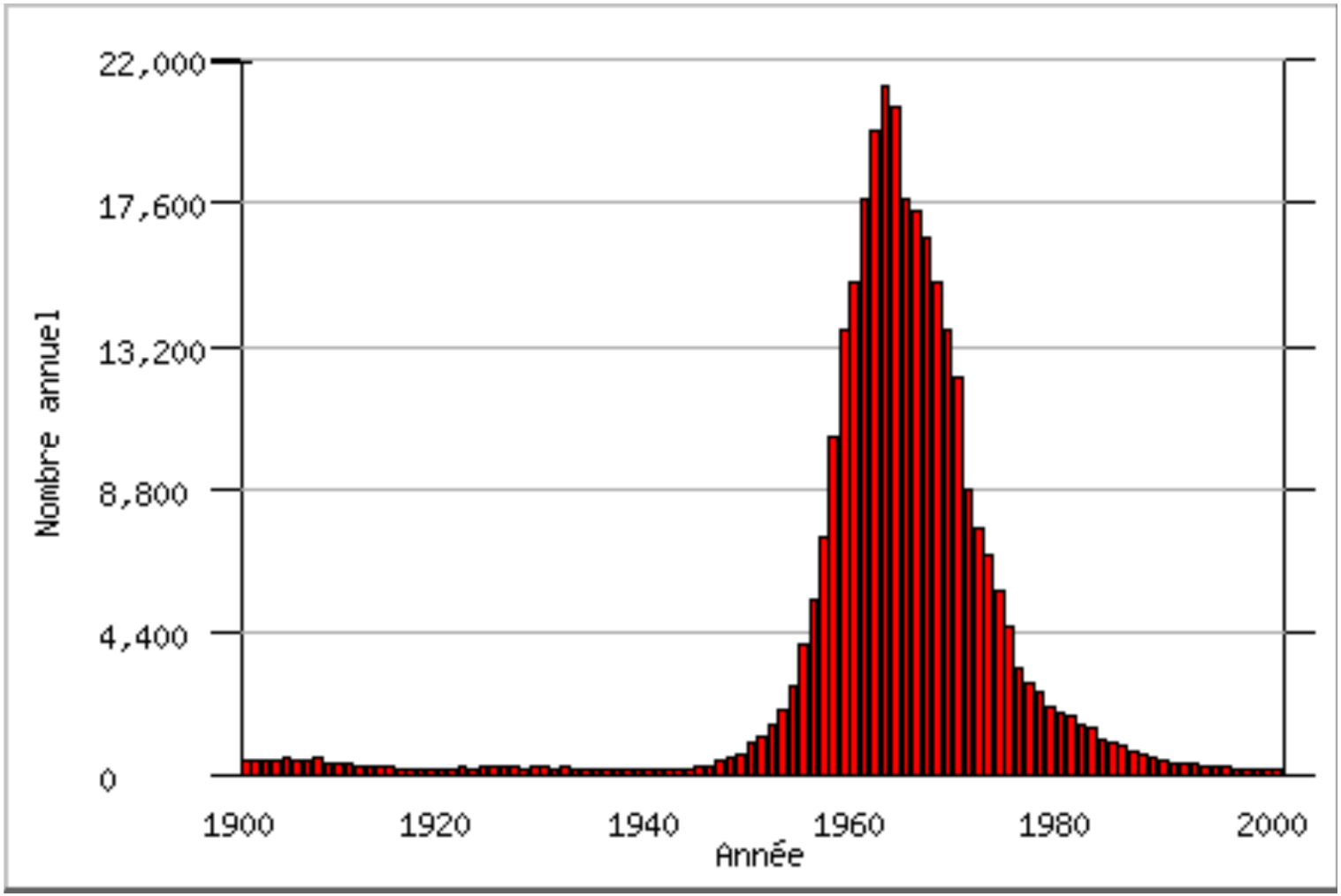
- Recenser avec les spécialistes métier et les informaticiens, les données utiles :
  - accessibles raisonnablement (pas sur microfilms !)
  - fiables
  - suffisamment à jour
  - historisées, si besoin est
  - légalement utilisables
- Il y a les données :
  - du système d'information (SI) de l'entreprise
  - stockées dans l'entreprise, hors du SI (fichiers Excel...)
  - achetées ou récupérées à l'extérieur de l'entreprise
  - provenant d'Internet et des réseaux sociaux
  - calculées à partir des données précédentes (indicateurs, ratios, évolutions au cours du temps)

Quand on manque de données

- Enquêtes auprès d'échantillons de clients
  - en les incitant à répondre à des questionnaires en leur proposant des cadeaux
- Utilisation des mégabases de données (Acxiom, Wegener Direct Marketing)
- « Scoring prénom »
- Utilisation de données géodémographiques (type d'habitat en fonction de l'adresse)
  - données moins précises que des données nominatives
  - mais disponibles pour des prospects
- Recours à des modèles standards préétablis par des sociétés spécialisées (ex : scores génériques)
  - quand on a des données actuelles mais peu d'historique

Scoring prénom

Pascal



Données géodémographiques

- Données économiques
  - nombre d'entreprises, population active, chômage, commerces et services de proximité, habitudes de consommation…
- Données sociodémographiques
  - population, richesse, âge et nombre d'enfants moyens, structures familiales, niveau socioprofessionnel…
- Données résidentielles
  - ancienneté, type et confort des logements, proportion de locataires et propriétaires…
- Données concurrentielles
  - implantation de l'entreprise, implantation de ses concurrents, parts de marché, taux de pénétration…
- Type d'habitat (classification sur les données précédentes) : beaux quartiers, classe moyenne, classe ouvrière, centre ville et quartiers commerçants...

Construction de la base d'analyse

n° client	variable cible : acheteur (O/N)	âge	PCS	situation famille	nb achats	montant achats	…	variable explicative *m*	échantillon
1	O	58	cadre	marié	2	40	…	…	apprentissage
2	N	27	ouvrier	célibataire	3	30	…	…	test
…	…	…		…	…	…	…	…	…
…	…	…		…	…	…	…	…	…
k	O	46	technicien	célibataire	3	75	…	…	test
…	…	…		…	…	…	…	…	…
…	…	…		…	…	…	…	…	…
1000	N	32	employé	marié	1	50	…	…	apprentissage
…	…	…		…	…	…		…	…

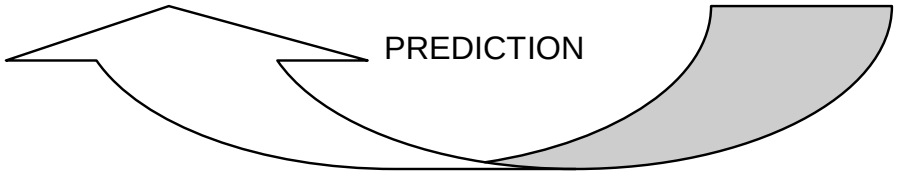
au moins 1000 cas

variable à expliquer observée année n

variables explicatives observées année n-1

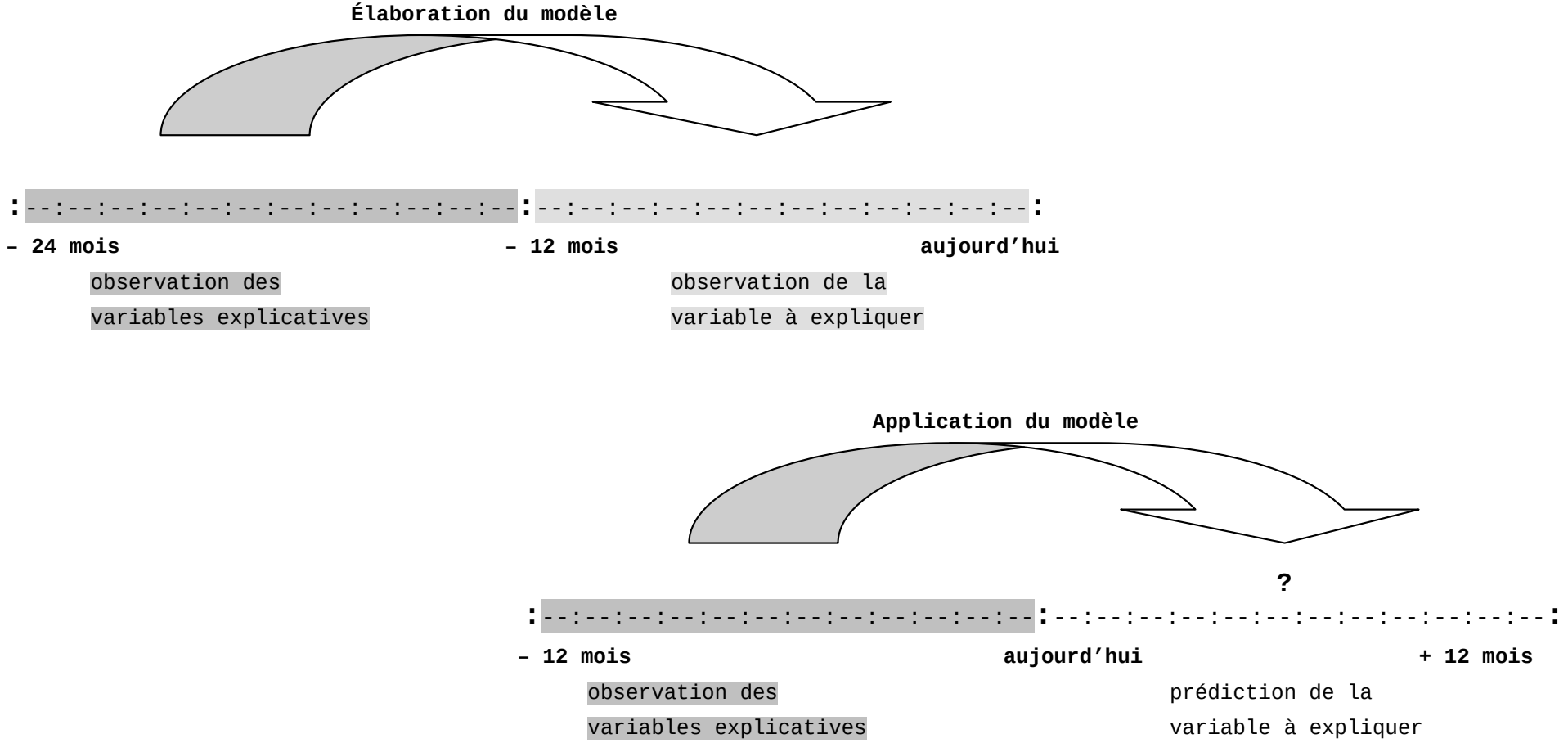
répartition aléatoire des clients entre les 2 échantillons

O : au moins 500 clients ciblés dans l'année *n* et acheteurs
N : au moins 500 clients ciblés dans l'année *n* et non acheteurs



f

Sélection des périodes d'observation



Le **modèle** sera par exemple une fonction *f* telle que :

Probabilité(variable cible = x) = *f*(variables explicatives)

Pré-segmentation

- Segmentation (classification) de la population :
  - en groupes forcément distincts selon les données disponibles (clients / prospects) : homogénéité du point de vue des variables explicatives
  - ou en groupes statistiquement pertinents vis-à-vis des objectifs de l'étude : homogénéité du point de vue de la variable à expliquer
  - ou selon certaines caractéristiques sociodémographiques (âge, profession…) si elles correspondent à des règles métiers (offres marketing spécifiques)
- Autres caractéristiques recherchées :
  - Simplicité de la segmentation (pas trop de règles)
  - Nombre limité de segments et stabilité des segments
  - Tailles des segments généralement du même ordre de grandeur

Intérêt de segmenter : le paradoxe de Simpson

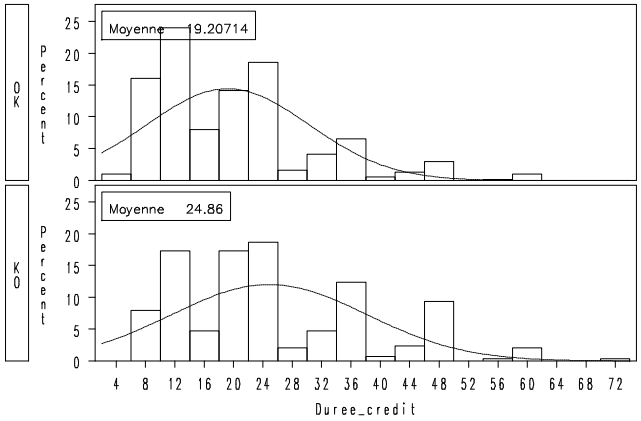
Hommes				
	sans achat	*avec achat*	*TOTAL*	*taux d'achat*
courriel	950	50	1 000	5,00%
téléphone	475	25	500	5,00%
TOTAL	1 425	75	1 500	5,00%
Femmes				
	sans achat	*avec achat*	*TOTAL*	*taux d'achat*
courriel	450	50	500	10,00%
téléphone	900	100	1 000	10,00%
TOTAL	1 350	150	1 500	10,00%
Tous clients				
	sans achat	*avec achat*	*TOTAL*	*taux d'achat*
courriel	1 400	100	1 500	6,67%
téléphone	1 375	125	1 500	8,33%
TOTAL	2 775	225	3 000	7,50%

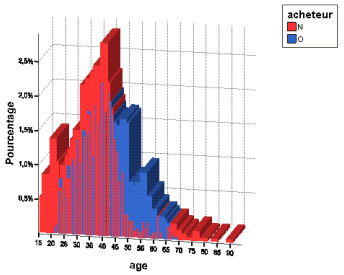
Paradoxe de Simpson : explication

- Dans le dernier exemple :
  - les hommes ne répondent pas mieux au téléphone qu'au courriel
  - de même pour les femmes
  - et pourtant, le téléphone semble avoir globalement un meilleur taux d'achat
- Explication :
  - un individu pris au hasard ne répond pas mieux au téléphone
  - mais les femmes achètent plus et on a privilégié le téléphone pour les contacter
    - liaison entre les variables « sexe » et « canal de vente »
- Autre exemple publié dans le *Wall-Street Journal* du 2/12/2009 :
  - le taux de chômage est globalement plus faible en octobre 2009 (10,2 %) qu'en novembre 1982 (10,8 %)
  - et pourtant, ce taux de chômage est plus élevé en 2009 à la fois pour les diplômés et pour les non-diplômés !
  - l'explication est l'existence d'une liaison entre l'année et le niveau d'étude : le niveau moyen d'étude est plus élevé en 2009, et le taux de chômage est plus faible chez ceux dont le niveau d'étude est plus élevé

Analyse exploratoire des données 1/2

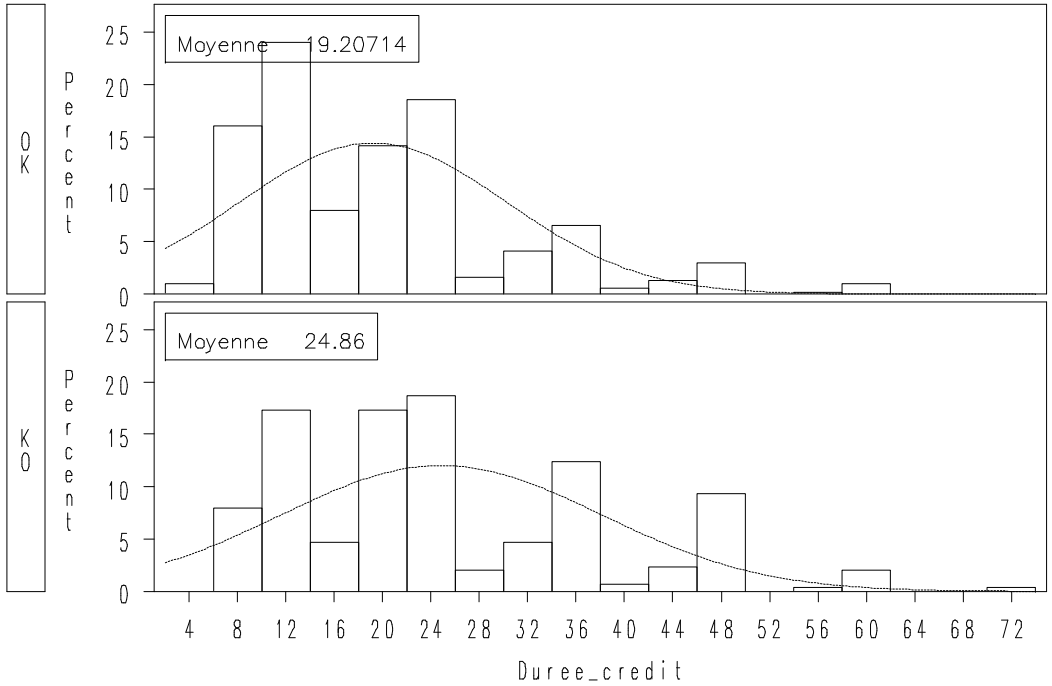
- Explorer la distribution des variables
- Vérifier la fiabilité des variables
  - valeurs incohérentes ou manquantes
    - suppression ou imputation ou isolement
  - valeurs extrêmes
    - voir si valeurs aberrantes à éliminer
  - certaines variables sont fiables mais trompeuses
    - le profil de souscripteurs peut être faussé par une campagne commerciale ciblée récente
- Variables continues
  - détecter la non-monotonie ou la non-linéarité justifiant la discrétisation
  - tester la normalité des variables (surtout si petits effectifs) et les transformer pour augmenter la normalité
  - éventuellement discrétiser : découper la variable en tranches en fonction de la variable à expliquer
    - et isoler les valeurs manquantes ou aberrantes





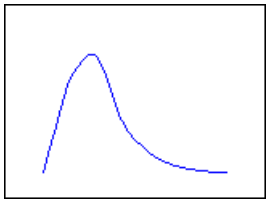
Examen de la distribution des variables

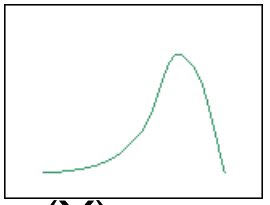
- La durée du crédit présente des pics prévisibles à 12, 24, 36, 48 et 60 mois
- On constate assez nettement la plus forte proportion de crédits plus longs parmi ceux qui ont des impayés
- Pas de valeur manquante ou aberrante



Normalisation : transformations

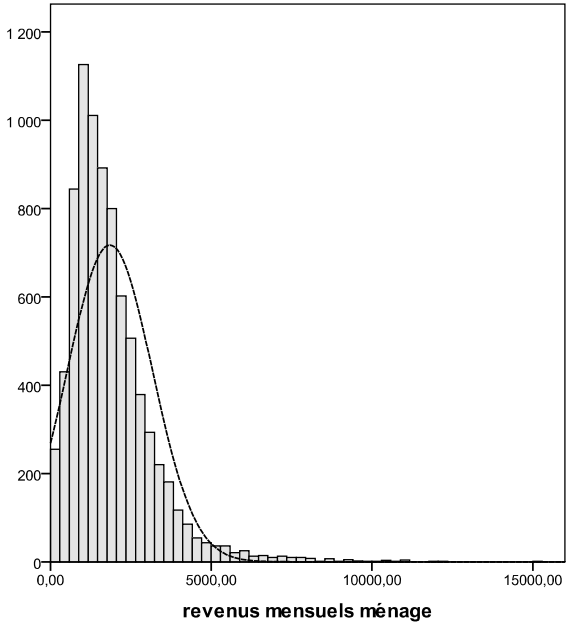
- Log (V)
  - transformation la plus courante pour corriger un coefficient d'asymétrie > 0
  - Si V ≥ 0, on prend Log (1 + V)
- Racine carrée (V) si coefficient d'asymétrie > 0
- -1/V ou $-1/V^2$ si coefficient d'asymétrie > 0
- V^2 ou V^3 si coefficient d'asymétrie < 0
- Arc sinus (racine carrée de V/100)
  - si V est un pourcentage compris entre 0 et 100
- La transformation de Box-Cox ($f(X) = (X^{\lambda}-1)/\lambda$ si $\lambda \neq 0$, et $f(X) = \log(X)$ sinon) recouvre un ensemble de transformations possibles, selon la valeur du paramètre λ déterminée par maximisation de la vraisemblance (en écrivant la densité d'une loi normale), et est implémentée dans plusieurs logiciels, dont R (fonction `boxplot` du package `MASS`)

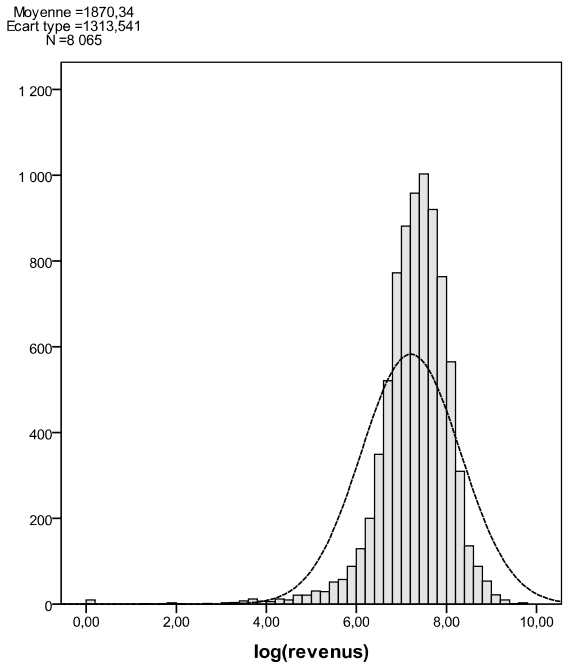


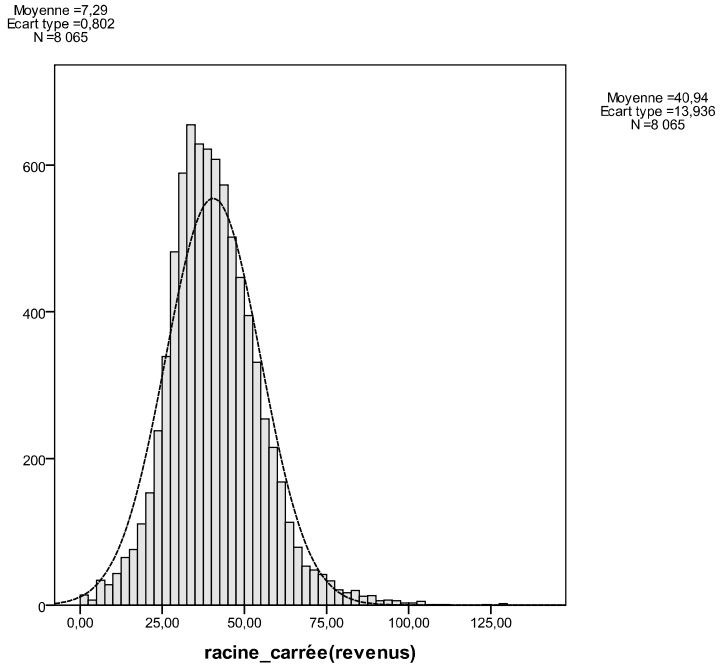


Transformation	exp(V)	V^3	V^2	V	√V	log(V)	-1/V	$-1/V^2$
Correction	asymétrie à gauche			pas de correction	asymétrie à droite			
Effet	fort ← moyen				moyen → fort			

Normalisation : un exemple







Revenus :

Asymétrie = 2,38

Aplatissement = 11,72

Log(1+revenus) :

Asymétrie = - 2,03

Aplatissement = 12,03

Racine(revenus) :

Asymétrie = 0,64

Aplatissement = 1,76

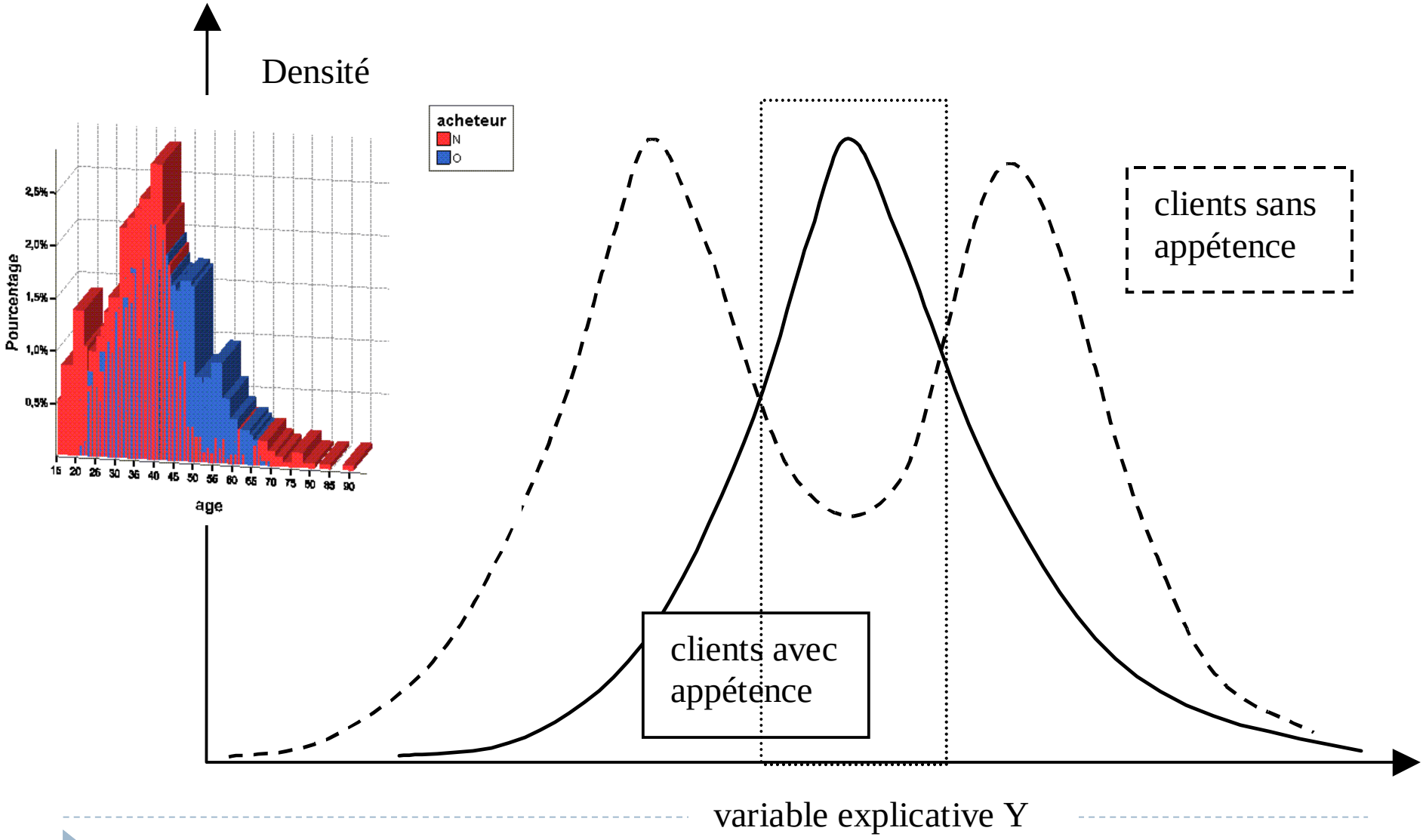
La racine carrée normalise ici mieux que le logarithme

(Loi normale : asymétrie = aplatissement (– 3) = 0)

Utilité de la normalisation

- Une des hypothèses de l'analyse discriminante linéaire :
  - multinormalité de X/G_i et égalité des matrices de covariances
- N'est en pratique jamais satisfaite
- Mais on constate une amélioration des performances de l'analyse discriminante lorsque l'on s'en rapproche :
  - en neutralisant les « outliers » (individus hors norme)
  - en normalisant les variables explicatives susceptibles d'entrer dans le modèle
- Moralité : mieux vaut connaître les contraintes théoriques pour se rapprocher des conditions optimales

Discrétisation en tranches naturelles



Pourquoi discrétiser ?

- Appréhender des liaisons non linéaires (de degré >1), voire non monotones, entre les variables continues et la variable à expliquer
  - par une analyse des correspondances multiples, une régression logistique ou une analyse discriminante DISQUAL
- Neutraliser les valeurs extrêmes (« outliers »)
  - qui sont dans la 1ère et la dernière tranches
- Gérer les valeurs manquantes (imputation toujours délicate)
  - rassemblées dans une tranche spécifique ou regroupée avec une autre
- Gérer les ratios dont le numérateur et le dénominateur peuvent être tous deux > 0 ou < 0
  - EBE / capital économique (rentabilité économique), résultat net / capitaux propres (rentabilité financière ou ROE)
- Améliorer parfois le pouvoir prédictif
- Faciliter la lisibilité du modèle (grille de score)

Exemple de discrétisation

- On commence par découper la variable explicative en déciles, et à regarder à quelle valeur correspond chaque décile
- Par exemple , le 2e décile est 25 ans

Analysis Variable : Age			
Rang pour la variable Age	N Obs	Minimum	Maximum
0	105	19.0000000	23.0000000
1	85	24.0000000	25.0000000
2	101	26.0000000	27.0000000
3	120	28.0000000	30.0000000
4	105	31.0000000	33.0000000
5	72	34.0000000	35.0000000
6	113	36.0000000	39.0000000
7	98	40.0000000	44.0000000
8	105	45.0000000	52.0000000
9	96	53.0000000	75.0000000

Exemple de discrétisation

- Le tableau de contingence montre que les deux premiers déciles de l'âge correspondent à un taux d'impayés nettement supérieur à celui des autres déciles. Il y a donc un seuil à 25 ans
- Aucun autre seuil ne se distingue nettement, les taux d'impayés fluctuant ensuite entre 20 % et un peu plus de 30 %
- Le découpage de l'âge en deux tranches est donc décidé

Table de dAge par Cible			
dAge(Rang pour la variable Age)	Cible		
FREQUENCE Pourcentage Pct en ligne	1	2	Total
0	63 6.30 60.00	42 4.20 40.00	105 10.50
1	47 4.70 55.29	38 3.80 44.71	85 8.50
2	74 7.40 73.27	27 2.70 26.73	101 10.10
3	79 7.90 65.83	41 4.10 34.17	120 12.00
4	72 7.20 68.57	33 3.30 31.43	105 10.50
5	55 5.50 76.39	17 1.70 23.61	72 7.20
6	89 8.90 78.76	24 2.40 21.24	113 11.30
7	70 7.00 71.43	28 2.80 28.57	98 9.80
8	84 8.40 80.00	21 2.10 20.00	105 10.50
9	67 6.70 69.79	29 2.90 30.21	96 9.60
Total	700 70.00	300 30.00	1000 100.00

Analyse exploratoire des données 2/2

- Variables qualitatives ou discrètes
  - regrouper certaines modalités aux effectifs trop petits
  - représenter les modalités dans une analyse des correspondances multiples
- Créer des indicateurs pertinents d'après les données brutes
  - prendre l'avis des spécialistes du secteur étudié
  - création d'indicateurs pertinents (maxima, moyennes, présence/absence…)
  - utiliser des ratios plutôt que des variables absolues (exemple : plafond ligne de crédit + part utilisée $\Rightarrow$ taux d'utilisation du crédit)
  - calcul d'évolutions temporelles de variables
  - création de durées, d'anciennetés à partir de dates
  - croisement de variables, interactions
  - utilisation de coordonnées factorielles
- Détecter les liaisons entre variables
  - entre variables explicatives et à expliquer (bon)
  - entre variables explicatives entre elles (colinéarité à éviter dans certaines méthodes)

Exemple de regroupement de modalités

- Regroupement de « < 100 » et « [100-500 euros[dont les taux d'impayés sont proches (35,99% et 33,01%)
- Regroupement de « [500-1000 euros[» et « >= 1000 euros » : leurs taux d'impayés sont moins proches mais la 2e modalité est trop petite pour rester seule
- On pourrait même regrouper ces deux modalités avec « Pas d'épargne »

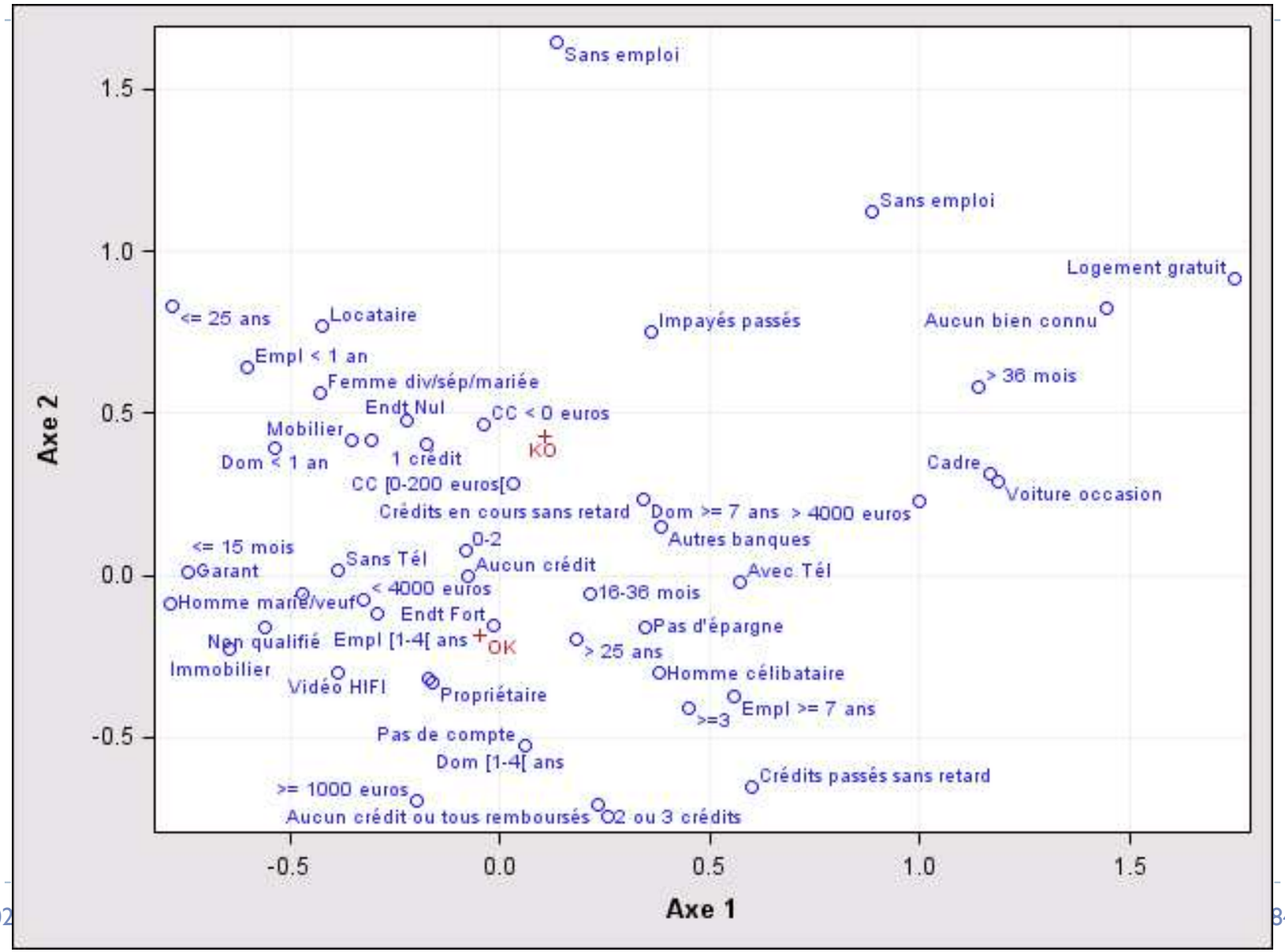
Table de Epargne par Cible			
Epargne	**Cible**		
FREQUENCE **Pourcentage** **Pct en ligne**	**OK**	**KO**	**Total**
Pas d'épargne	151 15.10 82.51	32 3.20 17.49	183 18.30
< 100	386 38.60 64.01	217 21.70 35.99	603 60.30
[100-500 euros[	69 6.90 66.99	34 3.40 33.01	103 10.30
[500-1000 euros[	52 5.20 82.54	11 1.10 17.46	63 6.30
>= 1000 euros	42 4.20 87.50	6 0.60 12.50	48 4.80
Total	700 70.00	300 30.00	1000 100.00

Autre exemple de regroupement de modalités

- Le regroupement des modalités « Locataire » et « Logement gratuit » est évident
- Elles sont associées à des taux d'impayés proches et élevés (39,11% et 40,74%)
- Les propriétaires sont moins risqués, surtout s'ils ont fini leur emprunt, mais pas seulement dans ce cas, car ils sont généralement plus attentifs que la moyenne au bon remboursement de leur emprunt

Table de Statut_domicile par Cible			
Statut_domicile	Cible		
FREQUENCE Pourcentage Pct en ligne	OK	KO	Total
Locataire	109 10.90 60.89	70 7.00 39.11	179 17.90
Propriétaire	527 52.70 73.91	186 18.60 26.09	713 71.30
Logement gratuit	64 6.40 59.26	44 4.40 40.74	108 10.80
Total	700 70.00	300 30.00	1000 100.00

Exploration avec une ACM



Traitement des valeurs manquantes

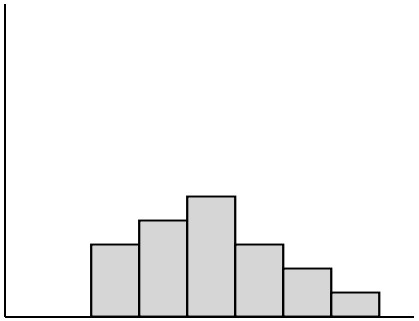
- D'abord vérifier que les valeurs manquantes ne proviennent pas :
  - d'un problème technique dans la constitution de la base
  - d'individus qui ne devraient pas se trouver dans la base
- Sinon, plusieurs solutions sont envisageables selon les cas :
  - supprimer les observations (si elles sont peu nombreuses ou si le non renseignement de la variable est grave et peut laisser suspecter d'autres anomalies dans l'observation)
  - ne pas utiliser la variable concernée (surtout si elle est peu discriminante) ou la remplacer par une variable proche mais sans valeur manquante
    - mieux vaut supprimer une variable *a priori* peu utile, mais qui est souvent non renseignée et conduirait à exclure de nombreuses observations de la modélisation
  - traiter la valeur manquante comme une valeur à part entière
  - imputation : remplacer la valeur manquante par une valeur par défaut ou déduite des valeurs des autres variables
  - remplacer les valeurs manquantes grâce à une source externe (rarement possible)
- Mais aucune solution n'est idéale ☹

Imputation des valeurs manquantes

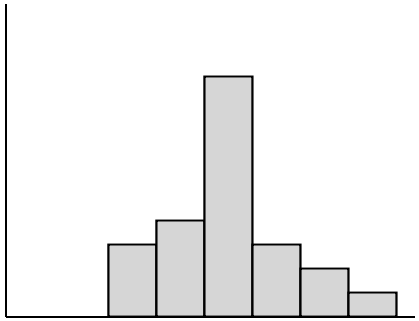
- Imputation statistique
  - par le mode, la moyenne ou la médiane
  - par une régression ou un arbre de décision
  - imputation
    - simple (minore la variabilité et les intervalles de confiance des paramètres estimés)
    - ou multiple (remplacer chaque valeur manquante par *n* valeurs, par exemple *n* = 5, puis faire les analyses sur les *n* tables et combiner les résultats pour obtenir les paramètres avec leurs écart-types

L'imputation n'est jamais neutre

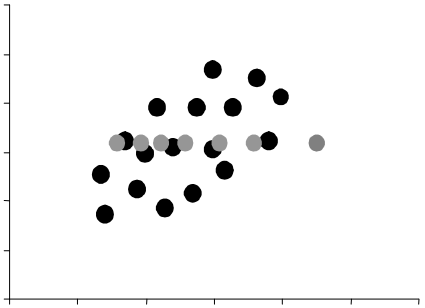
- Surtout si les données ne sont pas manquantes au hasard
- Déformation des variances et des corrélations



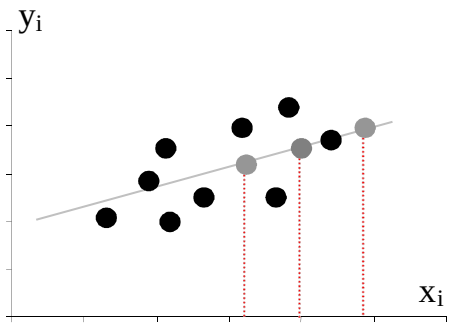
avant imputation

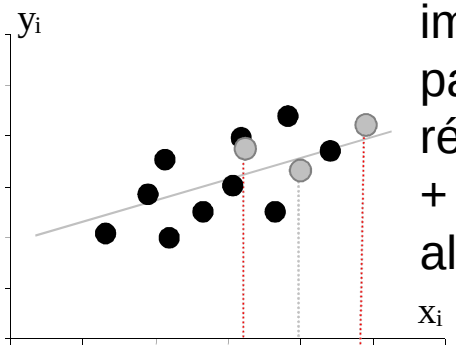


après imputation par la moyenne



imputation par ← moyenne ou régression →





imputation par régression + résidu aléatoire

source : J.-P. Nakache – A. Gueguen, RSA 2005

Schéma des valeurs manquantes

- Exemple de sortie produite par la procédure MI de SAS

Caractéristiques des données manquantes								
						Moyennes de groupes		
Groupe	Var1	Var2	Var3	Fréq	Pourcentage	Var1	Var2	Var3
1	X	X	X	6557	80.79	12.217310	0.245615	3.102462
2	X	.	X	3	0.04	0	.	0.166667
3	.	X	X	1108	13.65	.	-0.075471	0.595276
4	.	X	.	353	4.35	.	0.160265	.
5	.	.	X	91	1.12	.	.	0.000916
6	O	O	O	4	0.05	.	.	.

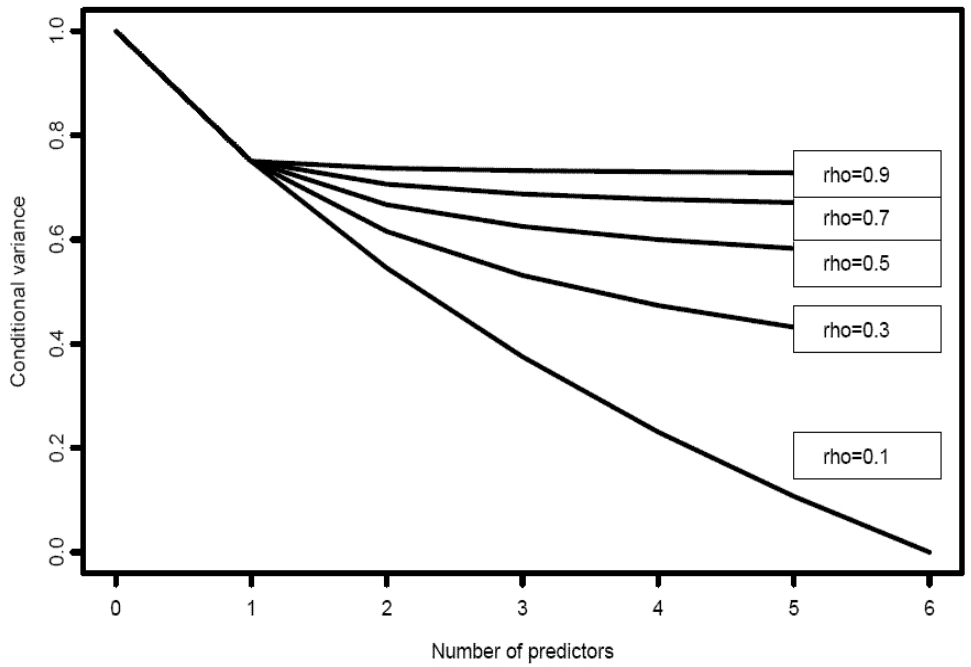
Le problème de la qualité des données : trois niveaux

- Données non correctes (manquantes ou aberrantes)
  - Pas toujours faciles à détecter
    - 0 est-il 0 ou manquant ? 9999..999 est-il manquant ou aberrant ?
  - S'agit-il d'une erreur ou d'un individu hors norme ?
  - Les données manquantes ou extrêmes sont plus faciles à détecter que les autres erreurs, qui ne se voient souvent que par croisement des données entre elles
  - Comment corriger en apprentissage / en application ?
- Données correctes mais non cohérentes
  - Venant du rapprochement de données correctes isolément MAIS
    - mesurées à des dates différentes
    - ou sur des échelles différentes
    - ou issues de règles de calcul différentes
- Données correctes et cohérentes mais trompeuses
  - Par exemple, en appétence, le profil des souscripteurs peut être faussé par une campagne commerciale ciblée récente

La sélection des variables

Importance de la sélection des variables

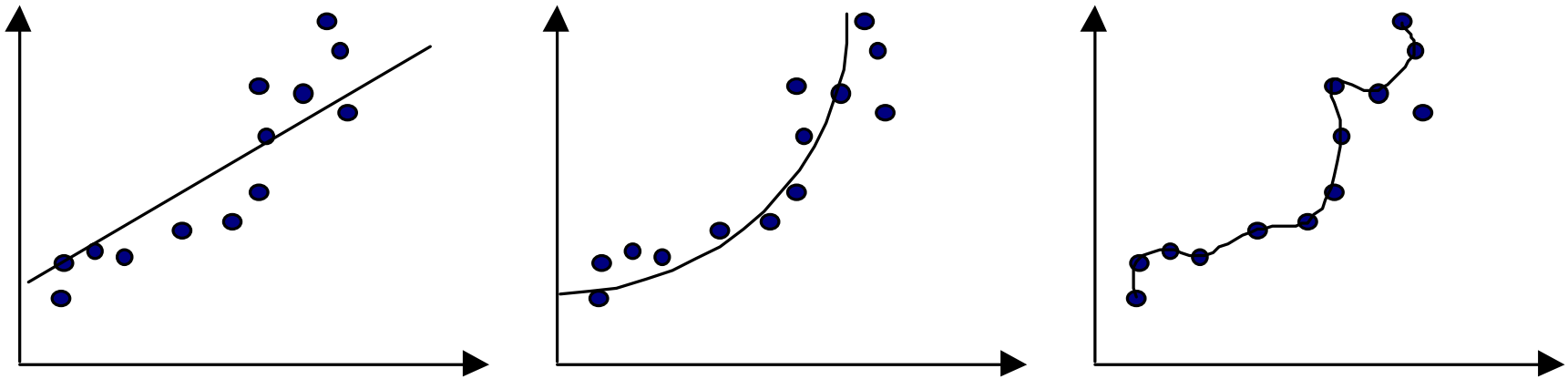
- Exemple de David Hand (2005) : régression avec un coefficient de corrélation linéaire 0,5 entre chaque prédicteur (variable explicative) et la variable à expliquer, et un coefficient de corrélation ρ entre chaque prédicteur
  - Les courbes représentent 1-R^2 (proportion de la somme des carrés non expliquée) en fonction du nombre de prédicteurs



Limiter le nombre de variables sélectionnées

- En présence de colinéarité entre les prédicteurs, l'apport marginal de chaque prédicteur décroît très vite
- Et pourtant, ici chaque prédicteur est supposé avoir la même liaison avec la variable à expliquer, ce qui n'est pas le cas dans une sélection pas à pas réelle où la liaison décroît !
- Conclusion :
  - Éviter au maximum la colinéarité des prédicteurs
  - Limiter le nombre de prédicteurs : souvent moins de 10
  - Alternative : la régression PLS ou régularisée (ridge…)
- Remarque :
  - Dans une procédure pas à pas, le 1er prédicteur peut occulter un autre prédicteur plus intéressant

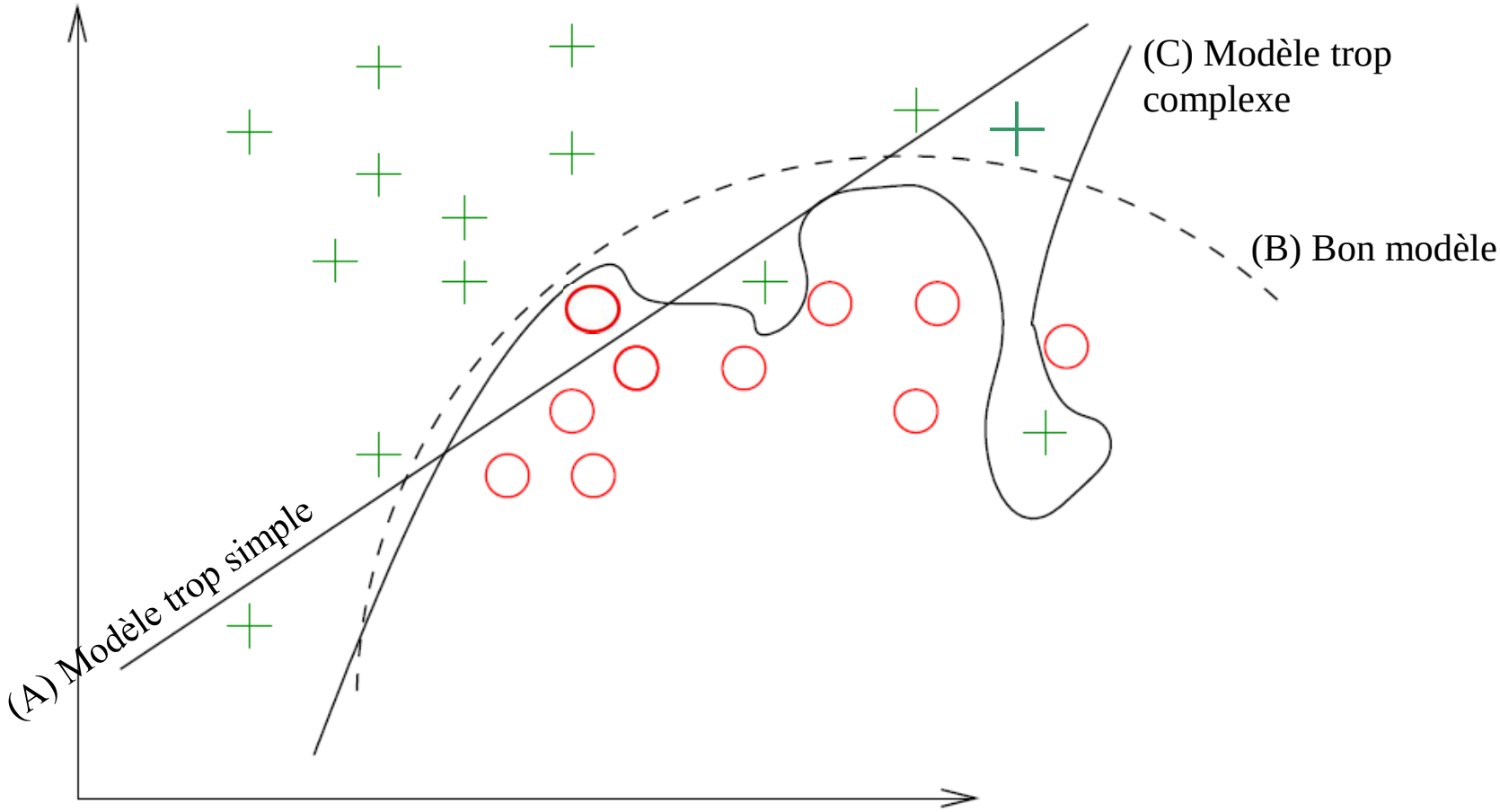
Sur-apprentissage en régression



(A) Modèle trop simple (B) Bon modèle (C) Modèle trop complexe

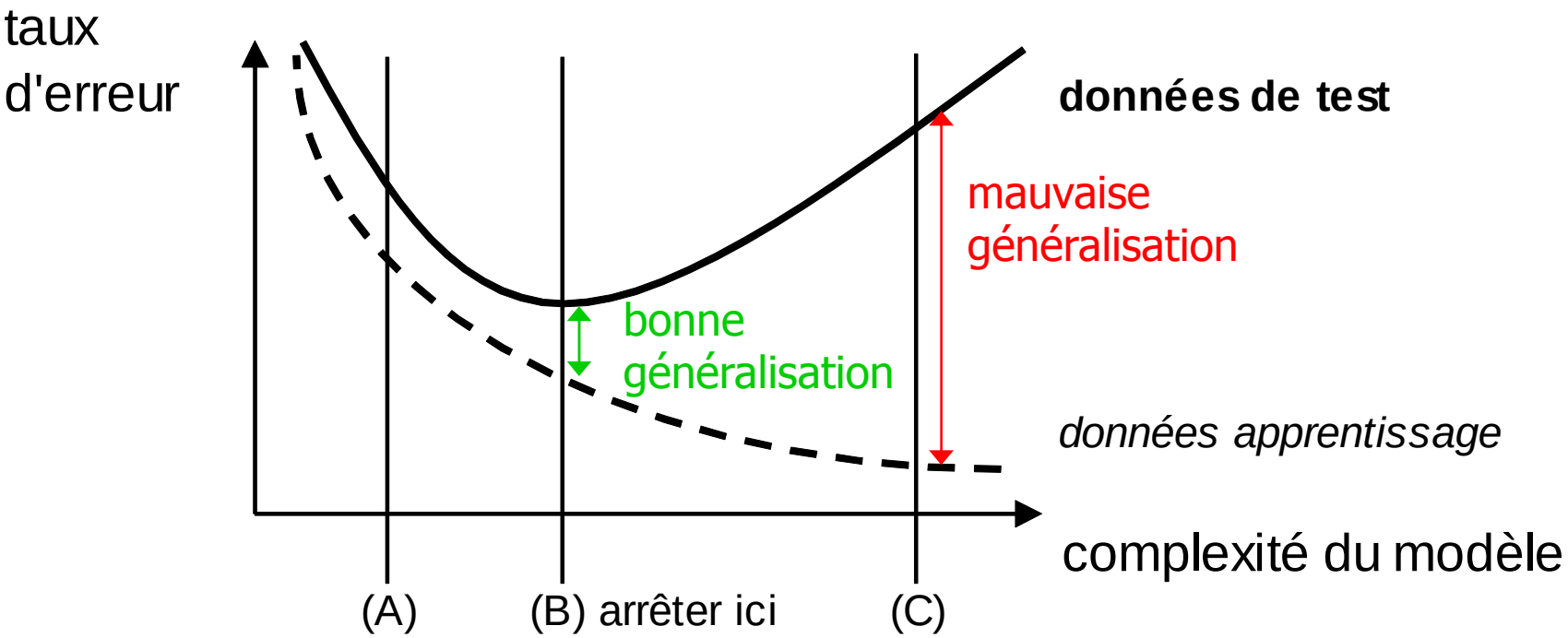
- Un modèle trop poussé dans la phase d'apprentissage :
  - épouse toutes les fluctuations de l'échantillon d'apprentissage,
  - détecte ainsi de fausses liaisons,
  - et les applique à tort sur d'autres échantillons
- On parle de sur-apprentissage ou sur-ajustement

Sur-apprentissage en classement



Source : Olivier Bousquet

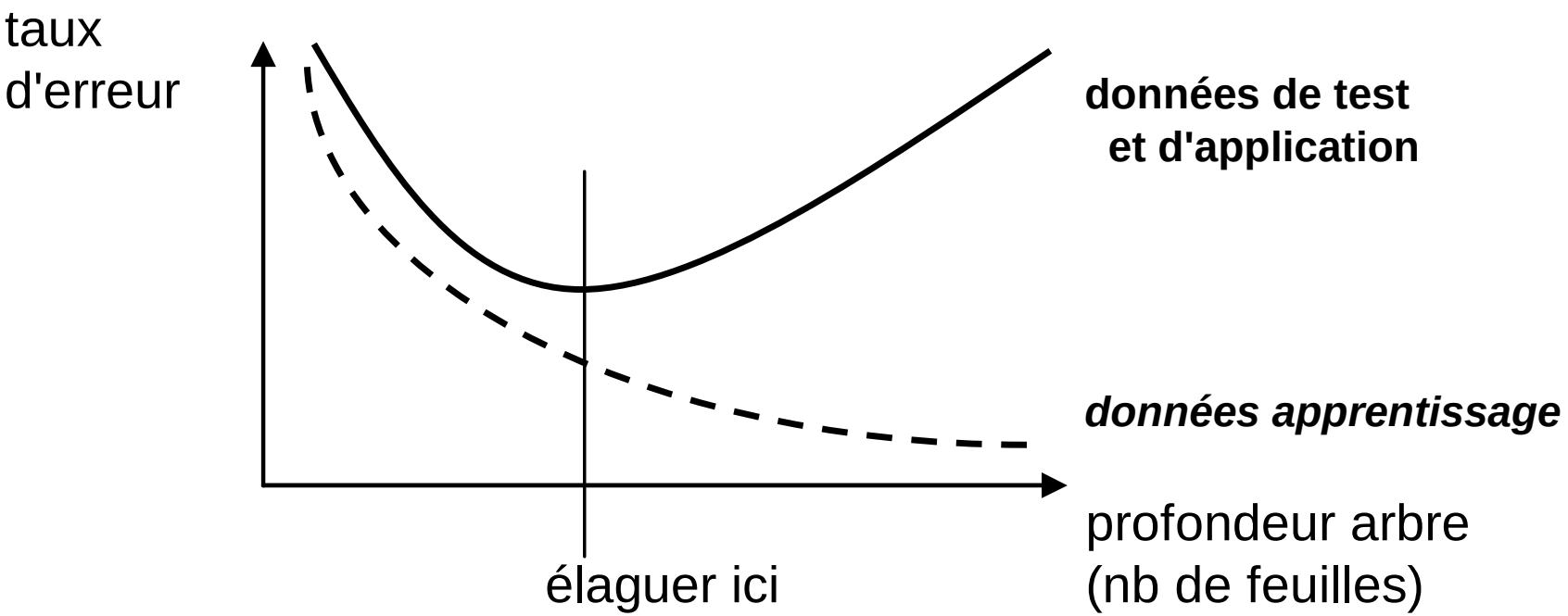
Taux d'erreur en fonction de la complexité du modèle



Théorème de Vapnik :

$$R < R_{emp} + \sqrt{\frac{h\,(\log(2n/h)+1) - \log(\alpha/4)}{n}}$$

Élagage d'un arbre de décision



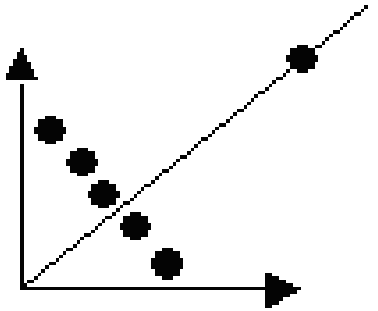
- Un bon arbre doit être élagué pour éviter la remontée du taux d'erreur due au sur-apprentissage
- Dans l'exemple précédent, il faut élaguer les feuilles 9 et 10

Sélection des variables explicatives

- En présence de corrélation linéaire entre les prédicteurs, l'apport marginal de chaque prédicteur décroît très vite
- Il peut même altérer le modèle (inversions de signes des paramètres) et réduire son pouvoir prédictif
- On doit effectuer des tests statistiques de liaison
- On peut préférer un prédicteur moins lié à la variable à expliquer s'il est moins corrélé aux autres prédicteurs
- On peut travailler sur les coordonnées factorielles
- Il est plus facile de limiter le nombre de prédicteurs si la population est homogène
- **Et même s'ils sont peu corrélés, les prédicteurs doivent être suffisamment peu nombreux (ou bornés comme dans la régression pénalisée) pour éviter d'avoir un modèle trop complexe et du sur-ajustement**

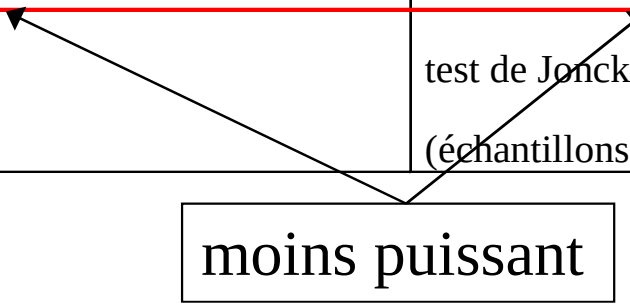
Rappel sur les tests

- Tests paramétriques
  - supposent que les variables suivent une loi particulière (normalité, homoscédasticité)
  - ex : test de Student, ANOVA
- Tests non-paramétriques
  - ne supposent pas que les variables suivent une loi particulière
  - se fondent souvent sur les rangs des valeurs des variables plutôt que sur les valeurs elles-mêmes
  - peu sensibles aux valeurs aberrantes
  - ex : test de Wilcoxon-Mann-Whitney, test de Kruskal-Wallis
- Exemple du *r* de Pearson et du ρ de Spearman :
  - $r > \rho \Rightarrow$ présence de valeurs extrêmes ?
  - $\rho > r \Rightarrow$ liaison non linéaire non détectée par Pearson ?
    - ex : x = 1, 2, 3… et y = e^1, e^2, e^3…



Liaison entre une variable continue et une variable de classe

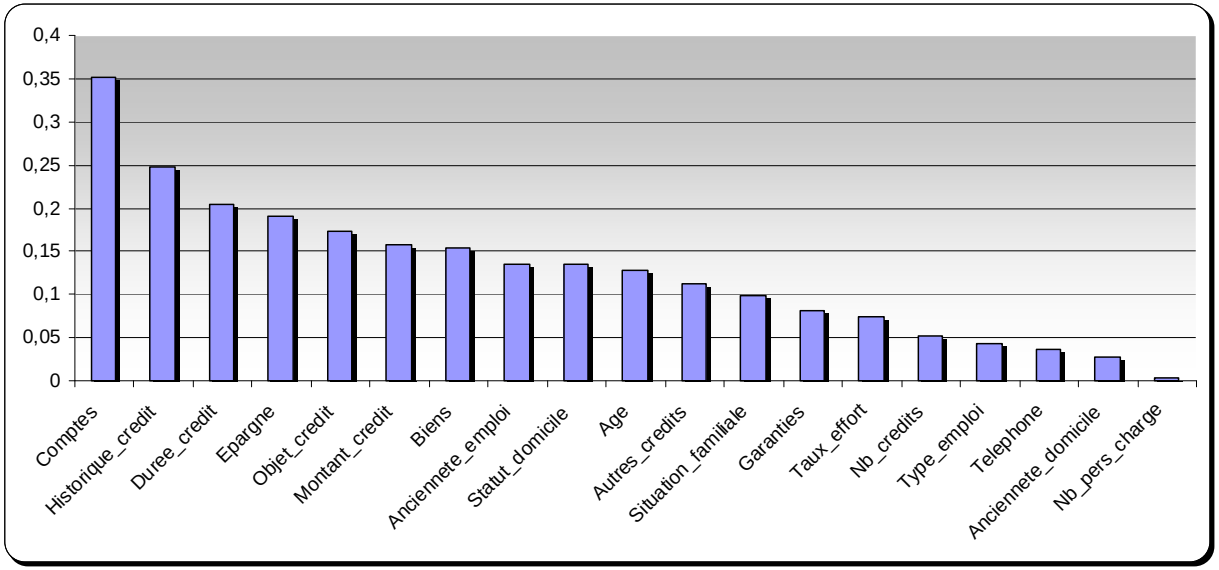
lois suivies	**2 échantillons**	**3 échantillons et plus (***)**
normalité – homoscédasticité (*)	test T de Student	ANOVA
normalité – hétéroscédasticité	test T de Welch	Welch - ANOVA
non normalité – hétéroscédasticité (**)	Wilcoxon – Mann – Whitney	Kruskal – Wallis
non normalité – hétéroscédasticité (**)	test de la médiane	test de la médiane
non normalité – hétéroscédasticité (**)		test de Jonckheere-Terpstra (échantillons ordonnés)



(*) Ces tests supportent mieux la non-normalité que l'hétéroscédasticité.
(**) Ces tests travaillant sur les rangs et non sur les valeurs elles-mêmes, ils sont plus robustes et s'appliquent également à des variables ordinales
(***) ne pas comparer toutes les paires par des tests T ⇒ on détecte à tort des différences significatives (au seuil de 95 % : dans 27 % des cas pour 4 moyennes égales)

Exemple de liste des variables

- Liste des variables par liaison décroissante avec la variable à expliquer
- Ici les variables sont qualitatives et la liaison mesurée par le V de Cramer



Obs	V_Cramer	Variable
1	0.35174	Comptes
2	0.24838	Historique_credit
3	0.20499	Duree_credit
4	0.19000	Epargne
5	0.17354	Objet_credit
6	0.15809	Montant_credit
7	0.15401	Biens
8	0.13553	Anciennete_emploi
9	0.13491	Statut_domicile
10	0.12794	Age
11	0.11331	Autres_credits
12	0.09801	Situation_familiale
13	0.08152	Garanties
14	0.07401	Taux_effort
15	0.05168	Nb_credits
16	0.04342	Type_emploi
17	0.03647	Telephone
18	0.02737	Anciennete_domicile
19	0.00301	Nb_pers_charge

Pourquoi le V de Cramer ?

	Classe 1	Classe 2	Ensemble
Effectifs observés :			
A	55	45	100
B	20	30	50
Total	75	75	150
Effectifs attendus si la variable est indépendante de la classe :			
A	50	50	100
B	25	25	50
Total	75	75	150
Probabilité du χ^2 = 0,08326454			
V de Cramer = 0,14142136			

	Classe 1	Classe 2	Ensemble
Effectifs observés :			
A	550	450	1000
B	200	300	500
Total	750	750	1500
Effectifs attendus si la variable est indépendante de la classe :			
A	500	500	1000
B	250	250	500
Total	750	750	1500
Probabilité du χ^2 = $4{,}3205.10^{-8}$			
V de Cramer = 0,14142136			

- Quand la taille de la population augmente, le moindre écart finit par devenir significatif aux seuils usuels

Le V de Cramer

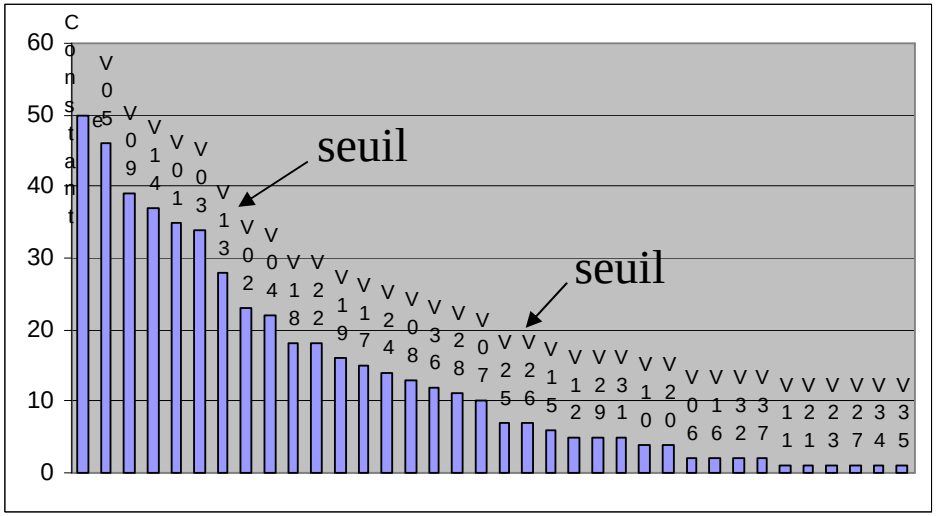
- V de Cramer = $\sqrt{\frac{\chi^2}{\chi^2_{max}}}$
  - mesure directement l'intensité de la liaison de 2 variables qualitatives, sans avoir recours à une table du χ^2
  - indépendamment du nombre de modalités et de l'effectif
  - en intégrant l'effectif et le nombre de degrés de liberté, par l'intermédiaire de χ^2_{max}
  - χ^2_{max} = effectif x [min (nb lignes, nb colonnes) – 1]
  - V compris entre 0 (liaison nulle) et 1 (liaison parfaite)

Sélection des variables : bootstrap

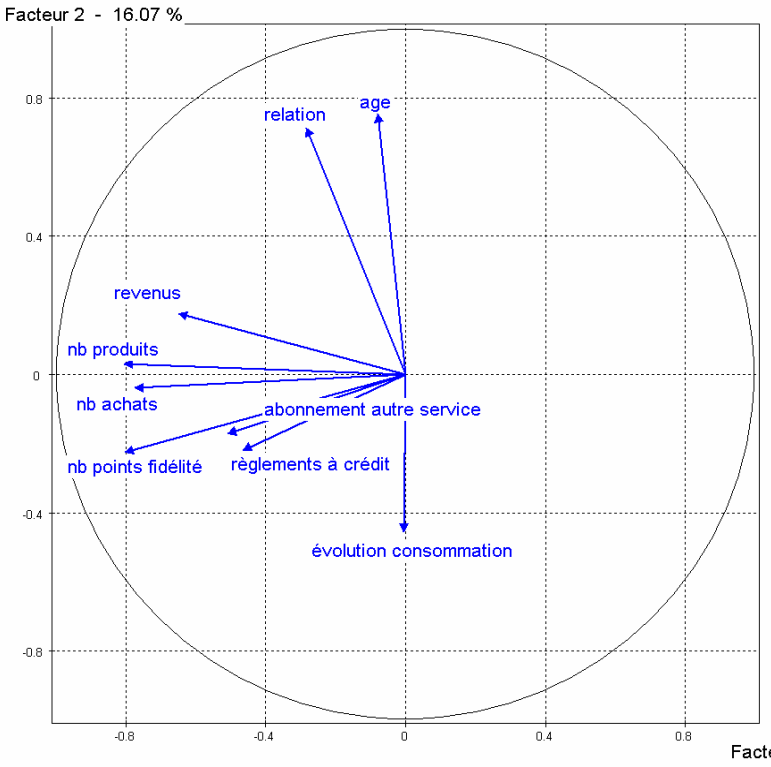
On effectue une régression logistique stepwise sur chacun des échantillons bootstrap

Variable	Nb occurrences		Variable	Nb occurrences
Constante	50		V25	7
V05	46		V26	7
V09	39		V15	6
V14	37		V12	5
V01	35		V29	5
V03	34		V31	5
V13	28		V10	4
V02	23		V20	4
V04	22		V06	2
V18	18		V16	2
V22	18		V32	2
V19	16		V37	2
V17	15		V11	1
V24	14		V21	1
V08	13		V23	1
V36	12		V27	1
V28	11		V34	1
V07	10		V35	1

Bootstrap : B tirages aléatoires avec remise de *n* individus parmi *n* et sélection de variables sur chacun des B échantillons bootstrap



Sélection des variables : classification à l'aide d'une ACP avec rotation



Cluster	Variable	R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio	Variable Label
Cluster 1	nbpoints	0.6546	0.0011	0.3458	nb points fidélité
	nbproduits	0.6189	0.0183	0.3882	nb produits
	nbachats	0.5950	0.0007	0.4053	nb achats
	revenus	0.4551	0.0234	0.5580	revenus du client
	abonnement	0.2537	0.0042	0.7495	abonnement autre service
	utilcredit	0.2312	0.0002	0.7689	règlements à crédit
Cluster 2	age	0.6033	0.0000	0.3967	âge
	relation	0.6461	0.0336	0.3662	relation (ancienneté client)
	evolconsom	0.2151	0.0027	0.7870	évolution consommation

```
PROC VARCLUS DATA=fichier_client;
VAR age relation nbpoints nbproduits nbachats revenus abonnement evolconsom
utilcredit;
RUN;
```

La modélisation

Méthodes inductives : 4 étapes

- Apprentissage : **construction du modèle** sur un 1er échantillon pour lequel on connaît la valeur de la variable à expliquer
- Test : **vérification du modèle** sur un 2d échantillon pour lequel on connaît la valeur de la variable à expliquer, que l'on compare à la valeur prédite par le modèle
  - si le résultat du test est insuffisant (d'après la *matrice de confusion* ou la courbe *ROC*), on recommence l'apprentissage
- **Validation du modèle** sur un 3e échantillon, éventuellement « out of time », pour avoir une idée du taux d'erreur non biaisé du modèle
- **Application du modèle** à l'ensemble de la population

valeur prédite → valeur réelle ↓	A	B	TOTAL
A	1800	200	
B	300	1700	
TOTAL			4000

Quelques méthodes classiques de scoring

- Analyse discriminante linéaire
  - Résultat explicite $P(Y/ X_1, \dots, X_p)$ sous forme d'une formule
  - Requiert des X_i continues et des lois X_i/Y multinormales et homoscédastiques (attention aux individus hors norme)
  - Optimale si les hypothèses sont remplies
- Régression logistique
  - Sans hypothèse sur les lois X_i/Y, X_i peut être discret, nécessaire absence de colinéarité entre les X_i
  - Méthode très souvent performante
  - Méthode la plus utilisée en scoring
- Arbres de décision
  - Règles complètement explicites
  - Traitent les données hétérogènes, éventuellement manquantes, sans hypothèses de distribution
  - Détection d'interactions et de phénomènes non linéaires
  - Mais moindre robustesse

Grille de score

- Passage de coefficients (« Estimation ») à des pondérations dont la somme est comprise entre 0 et 100

Analyse des estimations de la vraisemblance maximum						
Paramètre		**DF**	**Estimation**	**Erreur std**	**Khi 2 de Wald**	**Pr > Khi 2**
Intercept		1	-3.1995	0.3967	65.0626	<.0001
Comptes	**CC >= 200 euros**	1	1.0772	0.4254	6.4109	0.0113
Comptes	**CC < 0 euros**	1	2.0129	0.2730	54.3578	<.0001
Comptes	**CC [0-200 euros[**	1	1.5001	0.2690	31.1067	<.0001
Comptes	**Pas de compte**	0	0	.	.	.
Historique_credit	**Crédits en impayé**	1	1.0794	0.3710	8.4629	0.0036
Historique_credit	**Crédits sans retard**	1	0.4519	0.2385	3.5888	0.0582
Historique_credit	**Jamais aucun crédit**	0	0	.	.	
Duree_credit	**> 36 mois**	1	1.4424	0.3479	17.1937	<.0001
Duree_credit	**16-36 mois**	1	1.0232	0.2197	21.6955	<.0001
Duree_credit	**<= 15 mois**	0	0	.	.	.
Age	**<= 25 ans**	1	0.6288	0.2454	6.5675	0.0104
Age	**> 25 ans**	0	0	.	.	.
Epargne	**< 500 euros**	1	0.6415	0.2366	7.3501	0.0067
Epargne	**pas épargne ou > 500 euros**	0	0	.	.	.
Garanties	**Avec garant**	1	-1.7210	0.5598	9.4522	0.0021
Garanties	**Sans garant**	0	0	.	.	.
Autres_credits	**Aucun crédit extérieur**	1	-0.5359	0.2439	4.8276	0.0280
Autres_credits	**Crédits extérieurs**	0	0	.	.	.

Variable	Modalité	Nb points
Age	> 25 ans	0
Age	≤ 25 ans	8
Autres_credits	Aucun crédit extérieur	0
Autres_credits	Crédits extérieurs	7
Comptes	Pas de compte	0
Comptes	CC ≥ 200 euros	13
Comptes	CC [0-200 euros[	19
Comptes	CC < 0 euros	25
Duree_credit	≤ 15 mois	0
Duree_credit	16-36 mois	13
Duree_credit	> 36 mois	18
Epargne	pas épargne ou > 500 euros	0
Epargne	< 500 euros	8
Garanties	Avec garant	0
Garanties	Sans garant	21
Historique_credit	Jamais aucun crédit	0
Historique_credit	Crédits sans retard	6
Historique_credit	Crédits en impayé	13

Exemples de notations

- Note d'un jeune de moins de 25 ans, qui demande pour la première fois un crédit dans l'établissement et qui n'en a pas ailleurs, sans impayé, avec un compte dont le solde moyen est légèrement positif (mais < 200 €), avec un peu d'épargne (< 500 €), sans garant, qui demande un crédit sur 36 mois :
  - 8 + 0 + 19 + 13 + 8 + 21 + 0 = 69 points
- Note d'un demandeur de plus de 25 ans, avec des crédits à la concurrence, sans impayé, avec un compte dont le solde moyen est > 200 €, avec plus de 500 € d'épargne, sans garant, qui demande un crédit sur 12 mois :
  - 0 + 7 + 13 + 0 + 0 + 21 + 0 = 41 points
- On constate la facilité de l'implémentation et du calcul du score

Découpage de la note de score

- On peut calculer les déciles du nombre de points et leurs taux d'impayés correspondants :

Analysis Variable : nbpoints			
Rang pour la variable nbpoints	N Obs	Minimum	Maximum
0	104	6.0000000	29.0000000
1	95	33.0000000	37.0000000
2	107	39.0000000	42.0000000
3	120	43.0000000	48.0000000
4	98	49.0000000	54.0000000
5	93	55.0000000	60.0000000
6	81	61.0000000	65.0000000
7	104	66.0000000	69.0000000
8	92	70.0000000	74.0000000
9	106	75.0000000	95.0000000

Seuils de taux

Table de dnbpoints par Cible			
dnbpoints(Rang pour la variable nbpoints)	Cible		
FREQUENCE Pct en ligne	OK	KO	Total
0	99 95.19	5 4.81	104
1	89 93.68	6 6.32	95
2	100 93.46	7 6.54	107
3	101 84.17	19 15.83	120
4	71 72.45	27 27.55	98
5	60 64.52	33 35.48	93
6	48 59.26	33 40.74	81
7	60 57.69	44 42.31	104
8	38 41.30	54 58.70	92
9	34 32.08	72 67.92	106
Total	700	300	1000

Taux d'impayés par tranches de score

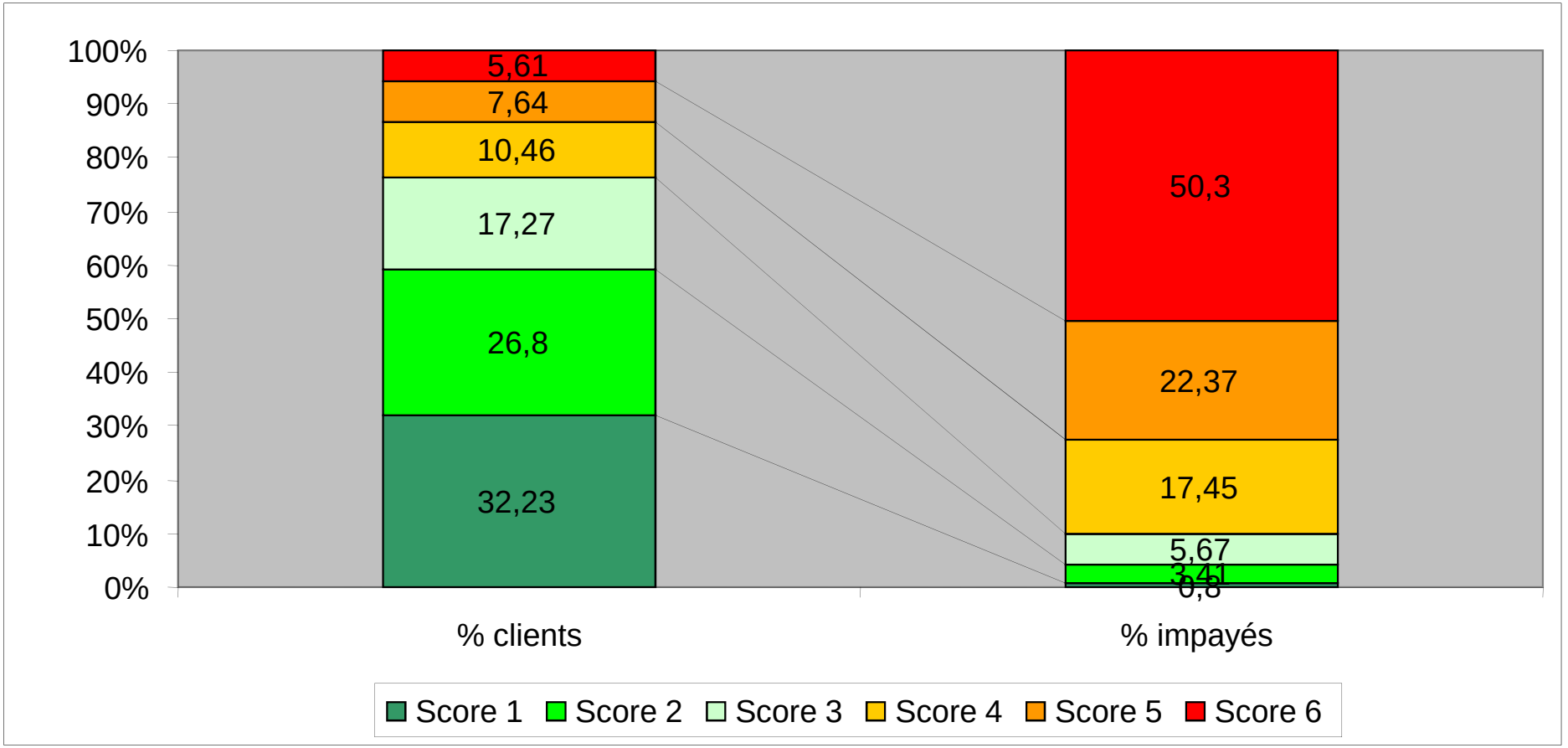
Table de nbpoints par Cible			
nbpoints	**Cible**		
FREQUENCE **Pourcentage** **Pct en ligne**	**OK**	**KO**	**Total**
risque faible [0 , 48] points	389 38.90 91.31	37 3.70 8.69	426 42.60
risque moyen [49 , 69] points	239 23.90 63.56	137 13.70 36.44	376 37.60
risque fort ≥ 70 points	72 7.20 36.36	126 12.60 63.64	198 19.80
Total	700 70.00	300 30.00	1000 100.00

- Tranche de risque faible :
  - 8,69% d'impayés
  - octroi du crédit avec un minimum de formalités
- Tranche de risque moyen :
  - 36,44% d'impayés
  - octroi du crédit selon la procédure standard
- Tranche de risque élevé :
  - 63,64% d'impayés
  - octroi du crédit interdit sauf par l'échelon hiérarchique supérieur (directeur d'agence)

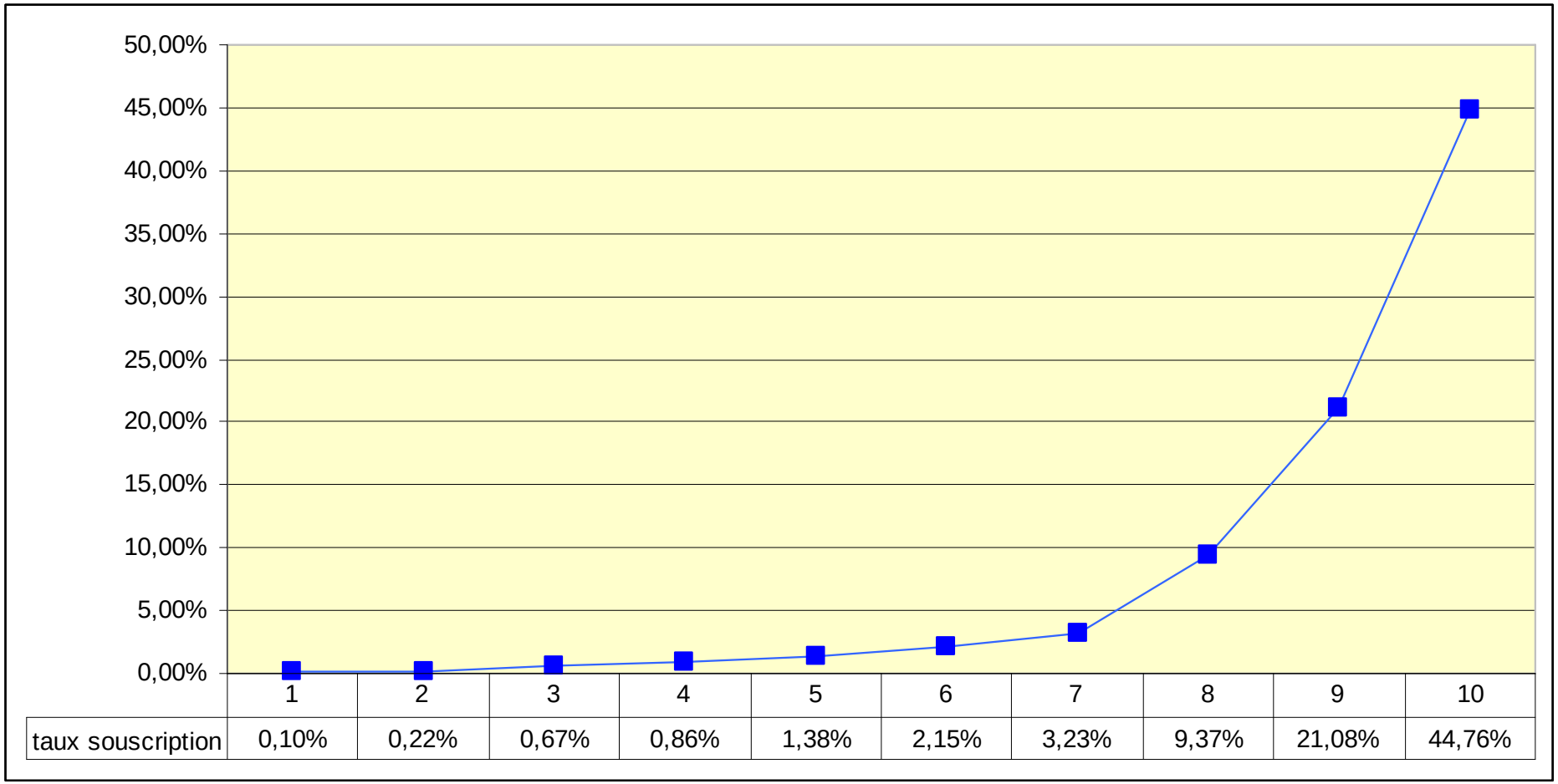
Reprenons nos exemples

- Demandeur de moins de 25 ans, qui demande pour la première fois un crédit dans l'établissement et qui n'en a pas ailleurs, sans impayé, avec un compte dont le solde moyen est légèrement positif (mais < 200 €), avec un peu d'épargne (< 500 €), sans garant, qui demande un crédit sur 36 mois :
  - 69 points ⇒ risque moyen
  - On est à la limite du risque élevé et cette limite aurait été franchie avec un crédit sur plus de 36 mois
- Demandeur de plus de 25 ans, avec des crédits à la concurrence, sans impayé, avec un compte dont le solde moyen est > 200 €, avec plus de 500 € d'épargne, sans garant, qui demande un crédit sur 12 mois :
  - 41 points ⇒ risque faible

Exemple de prédiction des impayés à 12 mois



Les résultats du modèle retenu (autre exemple)



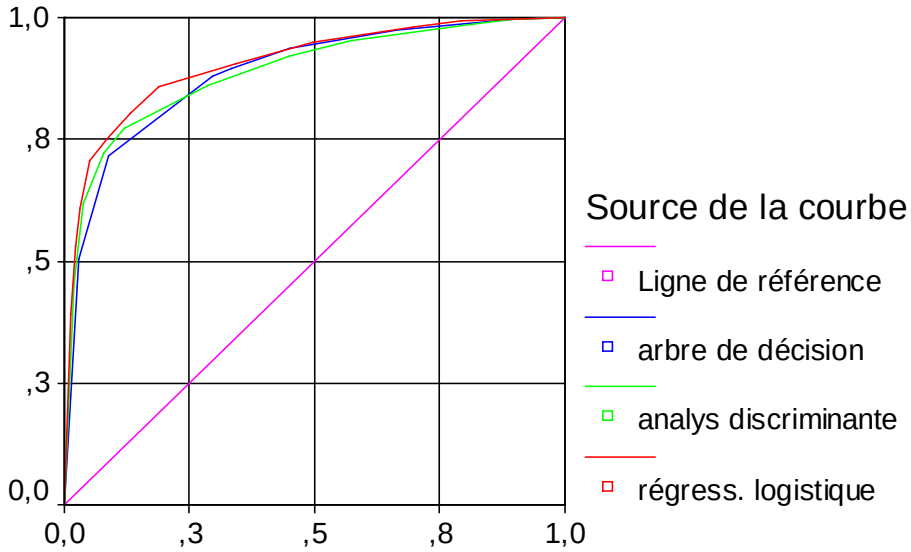
	1	2	3	4	5	6	7	8	9	10
taux souscription	0,10%	0,22%	0,67%	0,86%	1,38%	2,15%	3,23%	9,37%	21,08%	44,76%

- Observer l'évolution « exponentielle » du taux de souscription

Sensibilité et spécificité

- Pour un score devant discriminer un groupe A (les positifs; ex : les risqués) par rapport à un autre groupe B (les négatifs ; ex : les non risqués), on définit 2 fonctions du seuil de séparation s du score :
  - sensibilité = $\alpha(s)$ = Prob(score $\geq$ s / A) = probabilité de bien détecter un positif
  - spécificité = $\beta(s)$ = Prob(score < s / B) = probabilité de bien détecter un négatif
- Pour un modèle, on cherche s qui maximise $\alpha(s)$ tout en minimisant les faux positifs 1 - $\beta(s)$ = Prob(score $\geq$ s / B)
  - faux positifs : négatifs considérés comme positifs à cause du score
- Le meilleur modèle : permet de détecter le plus possible de vrais positifs avec le moins possible de faux positifs

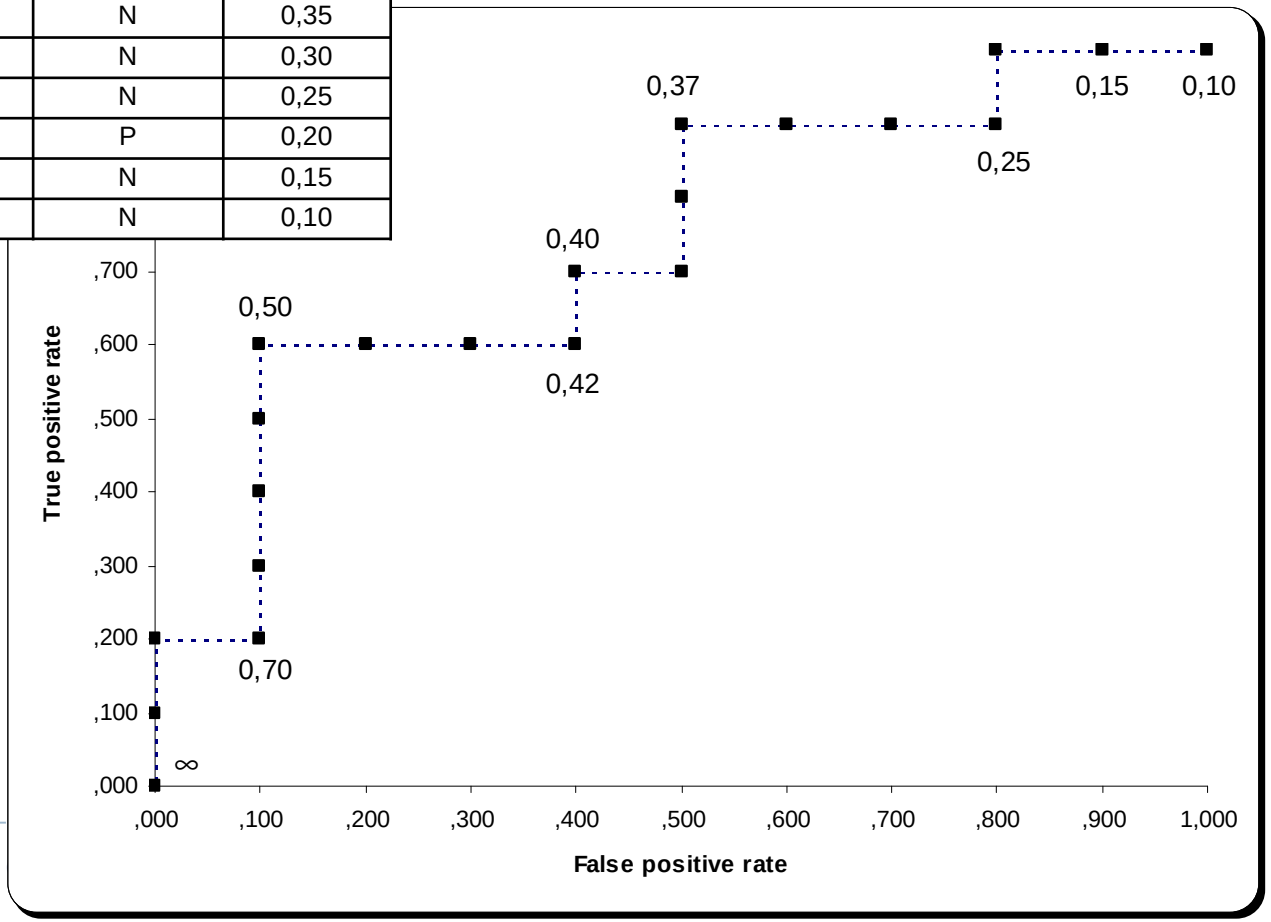
Courbe ROC



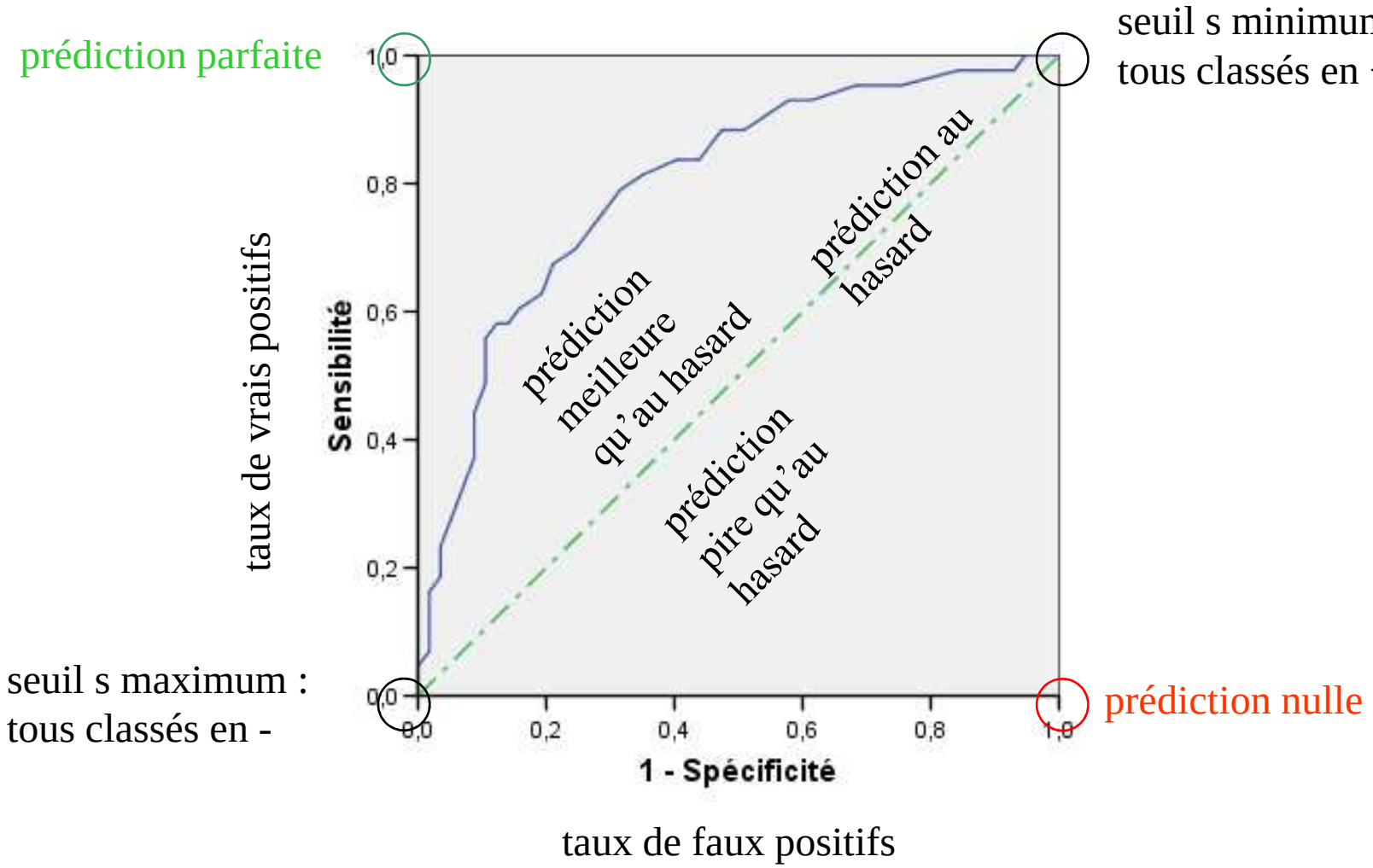
- La courbe ROC
  - sur l'axe Y : sensibilité = $\alpha(s)$
  - sur l'axe X : 1 - spécificité = 1 - $\beta(s)$
  - proportion y de vrais positifs en fonction de la proportion x de faux positifs, lorsque l'on fait varier le seuil s du score
- Aire AUC sous la courbe ROC = probabilité que score(x) > score(y), si x est tiré au hasard dans le groupe A (à prédire) et y dans le groupe B
  - 1ère méthode d'estimation : par la méthode des trapèzes
  - 2e méthode d'estimation : par les paires concordantes
  - 3e méthode équivalente : par le test de Mann-Whitney
- Le modèle est d'autant meilleur que l'AUC s'approche de 1
- AUC = 0,5 $\Rightarrow$ modèle pas meilleur qu'une notation aléatoire

Exemple de courbe ROC

#	Classe	Score	#	Classe	Score
1	P	0,90	11	P	0,40
2	P	0,80	12	N	0,39
3	N	0,70	13	P	0,38
4	P	0,65	14	P	0,37
5	P	0,60	15	N	0,35
6	P	0,55	16	N	0,30
7	P	0,50	17	N	0,25
8	N	0,45	18	P	0,20
9	N	0,44	19	N	0,15
10	N	0,42	20	N	0,10



Interprétation de la courbe ROC



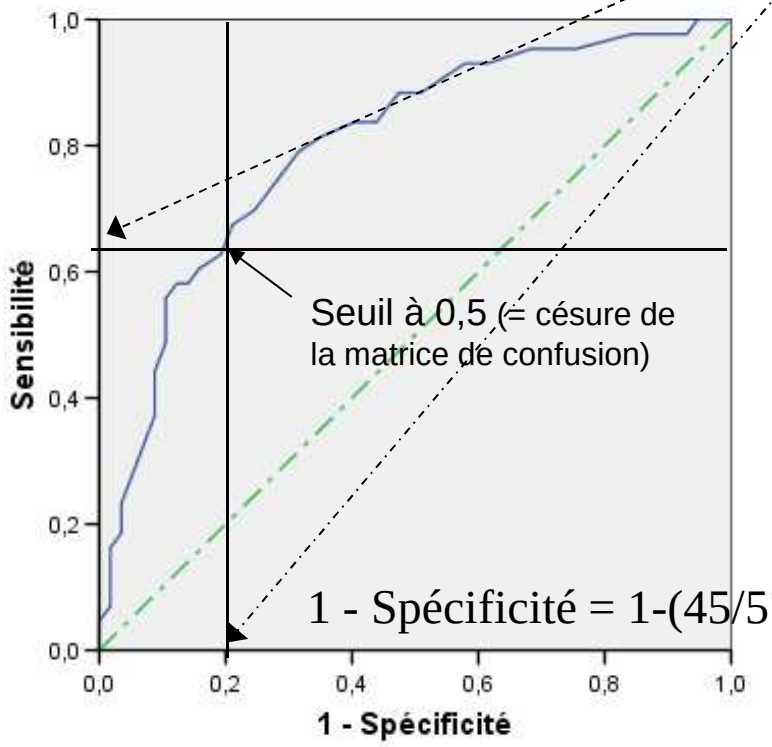
Matrice de confusion et courbe ROC

Tableau de classement[a]

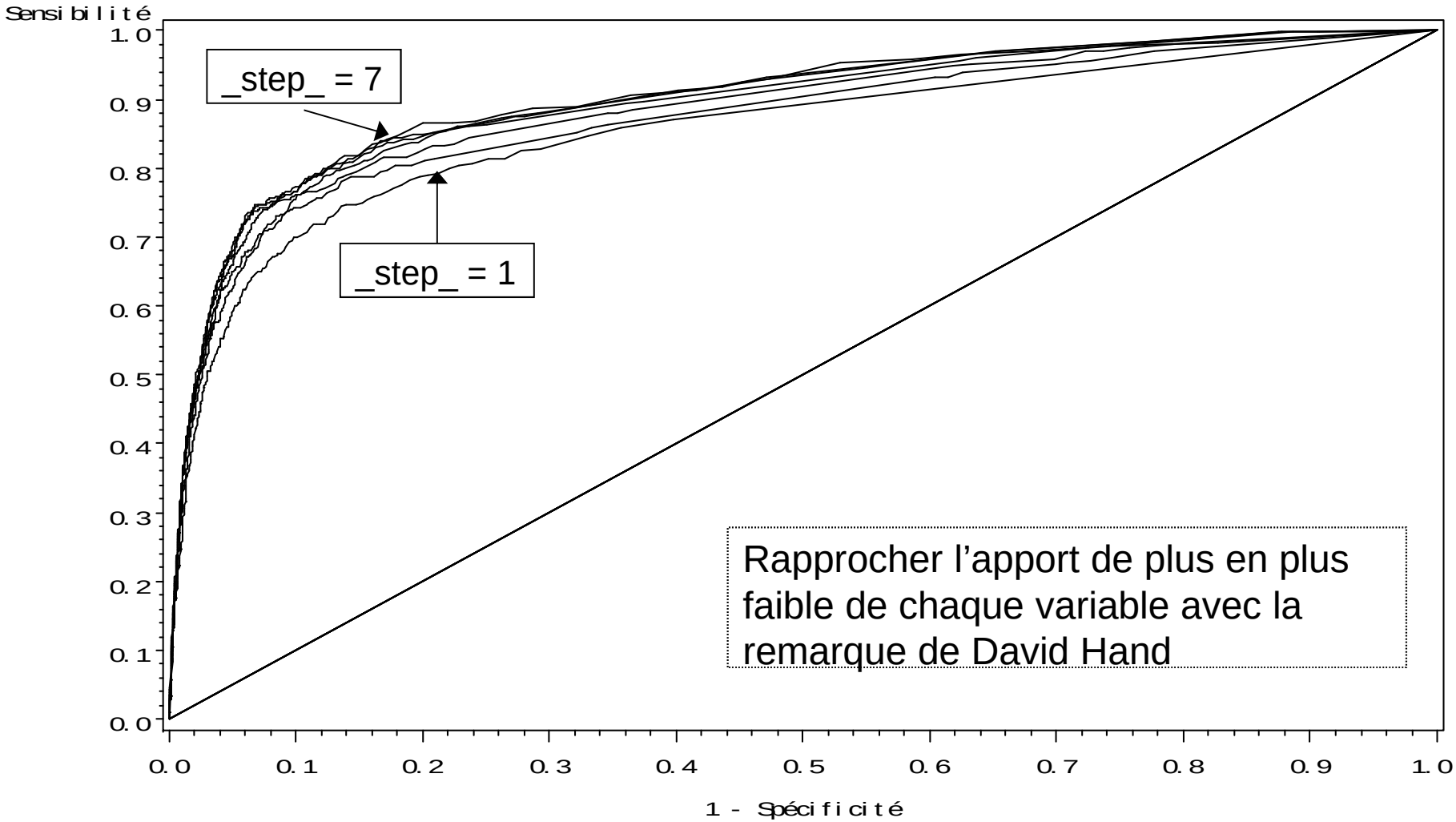
Observé		Prévu		
		CHD		Pourcentage correct
		0	1	
CHD	0	45	12	78,9
	1	16	27	62,8
Pourcentage global				72,0

a. La valeur de césure est ,500

Sensibilité = 27/43 = 0,63



Courbes ROC avec entrée progressive des variables du modèle



Quelques principes du data mining

Les 8 principes de base de la modélisation

- La préparation des données est la phase la plus longue, peut-être la plus laborieuse mais la plus importante
- Il faut un nombre suffisant d'observations pour en inférer un modèle
- Validation sur un échantillon de test distinct de celui d'apprentissage (ou validation croisée)
- Arbitrage entre la précision d'un modèle et sa robustesse (« dilemme biais – variance »)
- Limiter le nombre de variables explicatives et surtout éviter leur colinéarité
- Perdre parfois de l'information pour en gagner
  - découpage des variables continues en classes
- On modélise mieux des populations homogènes
  - intérêt d'une classification préalable à la modélisation
- La performance d'un modèle dépend souvent plus de la qualité des données et du type de problème que de la méthode

Qualités attendues d'une technique prédictive 1/2

- La précision
  - le taux d'erreur doit être le plus bas possible, et l'aire sous la courbe ROC la plus proche possible de 1
- La robustesse
  - être le moins sensible possible aux fluctuations aléatoires de certaines variables et aux valeurs manquantes
  - ne pas dépendre de l'échantillon d'apprentissage utilisé et bien se généraliser à d'autres échantillons
- La concision
  - les règles du modèle doivent être les plus simples et les moins nombreuses possible

Qualités attendues d'une technique prédictive 2/2

- Des résultats explicites
  - les règles du modèle doivent être accessibles et compréhensibles
- La diversité des types de données manipulées
  - toutes les méthodes ne sont pas aptes à traiter les données qualitatives, discrètes, continues et… manquantes
- La rapidité de calcul du modèle
  - un apprentissage trop long limite le nombre d'essais possibles
- Les possibilités de paramétrage
  - dans un classement, il est parfois intéressant de pouvoir pondérer les erreurs de classement, pour signifier, par exemple, qu'il est plus grave de classer un patient malade en « non-malade » que l'inverse

Choix d'une méthode : nature des données

explicatives ➔ **⬇ à expliquer**	**1 quantitative (covariable)**	**n quantitatives (covariables)**	**1 qualitative (facteur)**	**n qualitatives (facteurs)**	**mélange**
1 quantitative	rég. linéaire simple, régression robuste, arbres de décision	rég. linéaire multiple, rég. robuste, PLS, arbres, réseaux de neurones	ANOVA, arbres de décision	ANOVA, arbres de décision, réseaux de neurones	ANCOVA, arbres de décision, réseaux de neurones
***n* quantitatives (représentent des quantités ≠)**	régression PLS2	régression PLS2, réseaux de neurones	MANOVA	MANOVA, réseaux de neurones	MANCOVA, réseaux de neurones
1 qualitative nominale ou binaire	ADL, régression logistique, arbres de décision	ADL, rég. logistique, reg. logistique PLS, arbres, réseaux de neurones, SVM	régression logistique, DISQUAL, arbres	régression logistique, DISQUAL, arbres, réseaux de neurones	régression logistique, arbres, réseaux de neurones
1 discrète (comptage)	modèle linéaire généralisé (régression de Poisson, modèle log-linéaire)				
1 quantitative asymétrique	modèle linéaire généralisé (régressions gamma et log-normale)				
1 qualitative ordinale	régression logistique ordinale (au moins 3 niveaux)				
***n* quantitatives ou qualitatives**	modèle à mesures répétées (les n variables représentent des mesures répétées d'une même quantité)				

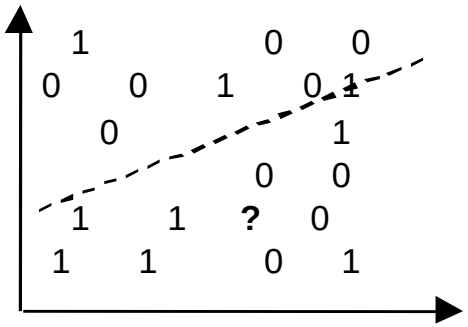
Choix d'une méthode : précision, robustesse, concision, lisibilité

- Précision : privilégier la régression linéaire, l'analyse discriminante linéaire, DISQUAL et la régression logistique, et parfois les SVM et les réseaux de neurones en prenant garde au sur-apprentissage (ne pas avoir trop de neurones dans la ou les couches cachées)
- Robustesse : éviter les arbres de décision et se méfier des réseaux de neurones, préférer une régression robuste à une régression linéaire par les moindres carrés
- Concision : privilégier la régression linéaire, l'analyse discriminante et la régression logistique, ainsi que les arbres sans trop de feuilles
- Lisibilité : préférer les arbres de décision et prohiber les réseaux de neurones. La régression logistique, DISQUAL, l'analyse discriminante linéaire et la régression linéaire fournissent aussi des modèles faciles à interpréter

Choix d'une méthode : autres critères

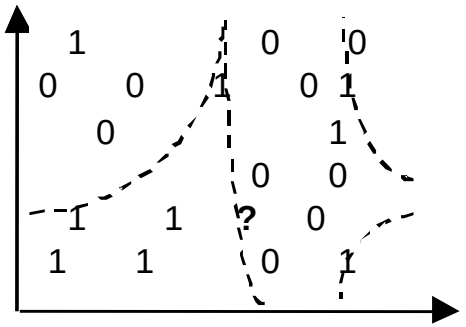
- Peu de données : éviter les arbres de décision et les réseaux de neurones
- Données avec des valeurs manquantes : essayer de recourir à un arbre, à une régression PLS, ou à une régression logistique en codant les valeurs manquantes comme une classe particulière
- Les valeurs extrêmes de variables continues n'affectent pas les arbres de décision, ni la régression logistique et DISQUAL quand les variables continues sont découpées en classes et les extrêmes placés dans 1 ou 2 classes
- Variables explicatives très nombreuses ou très corrélées : arbres de décision (pour limiter le nombre de variables du modèle), régression régularisée ou PLS (pour conserver le maximum de variables dans le modèle)
- Mauvaise compréhension de la structure des données : réseaux de neurones (sinon exploiter la compréhension des données par d'autres types de modèles)

Choix d'une méthode : topographie des classes à discriminer



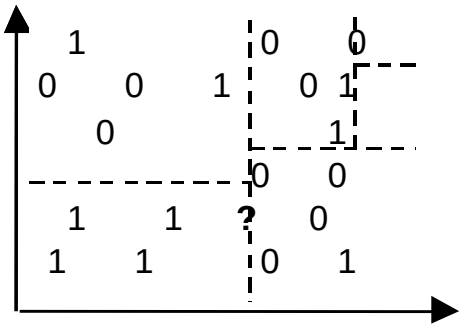
? est classé en "1"

Analyse discriminante



? est classé en "0"

Réseau de neurones



? est classé en "0"

Arbre de décision

- Toutes les méthodes de classement découpent l'espace des variables en régions, dont chacune est associée à une des classes à discriminer
- La forme de ces régions dépend de la méthode employée

Influence des données et méthodes

- Pour un jeu de données fixé, les écarts entre les performances de différents modèles sont souvent faibles
  - exemple de Gilbert Saporta sur des données d'assurance automobile (on mesure l'aire sous la courbe ROC) :
    - régression logistique : 0,933
    - régression PLS : 0,933
    - analyse discriminante DISQUAL : 0,934
    - analyse discriminante barycentrique : 0,935
  - le choix de la méthode est parfois affaire d'école
- Les performances d'un modèle dépendent :
  - un peu de la technique de modélisation employée
  - beaucoup plus des données !
- D'où l'importance de la phase préliminaire d'exploration et d'analyse des données
  - Collecter des données pertinentes nouvelles (ex : sémiométriques)

L'agrégation de modèles

Fonction de perte et risque d'un modèle

- L'erreur de prédiction d'un modèle se mesure par une fonction de perte :
  - y continue $\Rightarrow L(y,f(x)) = (y - f(x))^2$
  - $y = -1/+1 \Rightarrow L(y,f(x)) = \frac{1}{2}|y - f(x)|$
- Risque (ou risque réel) = espérance de la fonction de perte sur l'ensemble des valeurs possibles des données (x,y)
  - comme on ne connaît pas la loi de probabilité conjointe de x et y, on ne peut qu'estimer le risque
  - l'estimation la plus courante est le risque empirique
    $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ ou $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}|y_i - f(x_i)|$
  - on retrouve le taux d'erreur pour $y = -1/+1$ (n = effectif)
- Dans le cas quadratique, le risque se décompose en :
  - Biais²(modèle) + Variance(modèle)
  - (différence entre espérance de la prédiction f(x) et valeur moyenne de y)² + variance de la prédiction

Dilemme Biais-Variance

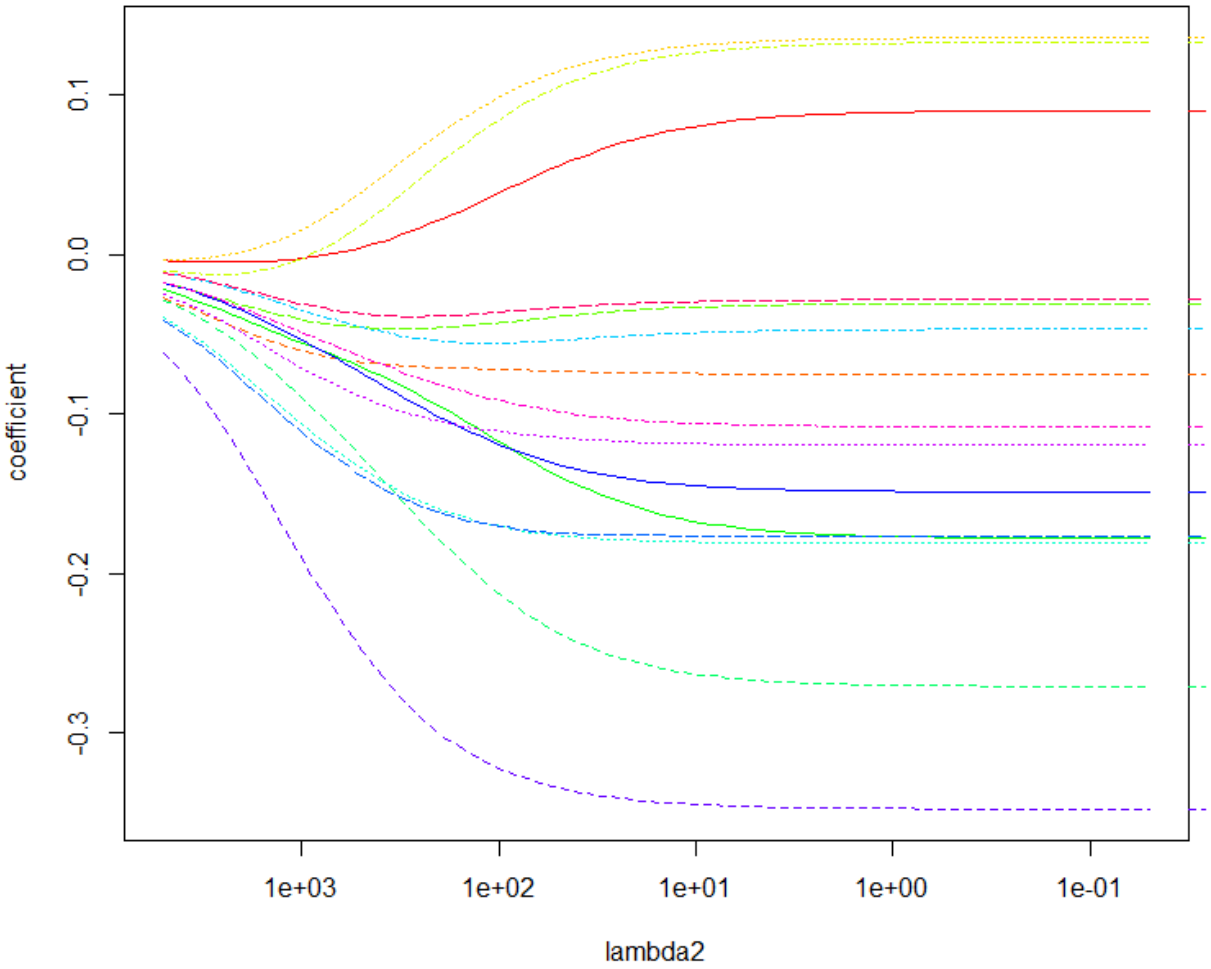
- Plus un modèle est complexe, plus son biais diminue mais plus sa variance augmente
- Nous devons trouver le bon réglage (trade-off) entre biais et variance, entre ajustement aux données d'apprentissage (biais) et capacité de généralisation (variance)
- Dans quelques cas simples, la complexité d'un modèle est égale au nombre p de paramètres
- Dans certaines situations, on ne peut pas diminuer le nombre de paramètres car les utilisateurs veulent voire apparaître simultanément des critères même s'ils sont fortement corrélés
  - Médecine, avec des mesures physiologiques, des résultats d'analyses
  - Banque, avec des critères qualitatifs saisis sur les entreprises
- Cette complexité peut être diminuée par l'introduction de bornes $||\beta|| \leq C$ dans la recherche des coefficients d'un modèle de régression (les observations étant dans une sphère de rayon R)
  - complexité ≤ min [partie entière $(R^2.C^2)$,p] + 1

Solutions de réduction de complexité

- La régression avec pénalisation L^d ($d \geq 0$)
  - Minimiser -2.log-vraisemblance (ou Σ des carrés) + $\lambda\Sigma|\beta_i|^d$, $d \geq 0$
  - $\Leftrightarrow$ minimiser -2.log-vraisemblance avec la contrainte $\lambda\Sigma|\beta_i|^d \leq C$
  - $d \leq 1$: sélection de prédicteurs (AIC, BIC si $d = 0$, Lasso si $d = 1$)
  - $d > 1$: rétrécissements de coefficients (Ridge si $d = 2$)
- La régression ridge (ou logistique ridge) est la plus répandue
  - Elle réduit les coefficients dans toutes les directions, surtout celles à faible variance (le coefficient de la ridge sur la 1ère composante principale diminue moins que le coefficient sur la 2e composante, etc.)
- La complexité peut aussi être réduite par la régression PLS
  - Avec une seule composante : les signes des coefficients sont égaux aux signes des corrélations entre prédicteurs et variable réponse
  - La régression PLS réduit les coefficients dans les directions à faible variance, mais peut provoquer une hausse trop grande dans les directions à forte variance $\Rightarrow$ l'erreur de prédiction de la PLS est souvent un peu supérieure

Ridge plot

- Évolution des coefficients en fonction de la pénalisation



Introduction aux méthodes d'agrégation

- Nous avons vu que la complexité d'un modèle doit être maîtrisée pour lui assurer une faible somme « biais² + variance » et donc une bonne généralisation
- La complexité d'un modèle peut être diminuée par :
  - La diminution du nombre de prédicteurs
  - L'introduction de bornes sur les coefficients de régression de ces prédicteurs
  - L'augmentation de la marge des SVM
- Nous allons voir une autre approche avec les méthodes d'agrégation (synonyme : méthodes d'ensemble) qui consistent à agréger les prédictions de plusieurs modèles de même type, d'une façon qui permette de réduire la variance et éventuellement le biais du modèle agrégé

Principe des méthodes d'agrégation

- La moyenne de B variables aléatoires i.i.d. de variance σ^2, a une variance σ^2/B
- L'espérance de cette moyenne de variables aléatoires est égale à l'espérance de chaque variable
- Si les variables sont identiquement distribuées mais dépendantes, avec une corrélation positive ρ, la variance de la moyenne est $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$
- Cette formule peut être appliquée à la fonction de prédiction définie par chaque modèle dans l'agrégation. Si ces fonctions sont fortement corrélées, l'agrégation réduira peu la variance, même si B est grand

Bagging 1/2

- En moyennant des classifieurs (par exemple des arbres) sur B échantillons bootstrap, on obtient un classifieur :
  - dont le biais n'a pas diminué
  - dont la variance a diminué d'autant plus que la corrélation entre les classifieurs est faible
- Les *n* modèles sont agrégés :
  - par un vote ou une moyenne des probabilités P(Y=1|X) quand on sait calculer cette moyenne (classement)
  - par une moyenne des estimations (régression)
- C'est le bagging : Bootstrap AGGregatING, Breiman, 1996
- La procédure de vote appliquée à des arbres de faible qualité peut conduire à un résultat pire lors de l'agrégation
  - Supposons que $Y = 1$ pour tout x et que chaque classifieur prédise 1 avec la probabilité 0,4 et 0 avec la probabilité 0,6. L'erreur de classement de chaque classifieur vaudra 0,6 mais l'agrégation par vote donnera un classifieur dont l'erreur vaudra 1.

Bagging 2/2

- Le classifieur de base est le même à chaque itération : arbre de décision, réseau de neurones…
- La corrélation entre les classifieurs est diminuée par :
  - le mécanisme de bootstrap
  - l'augmentation de la complexité
- ⇒ Le bagging s'applique mieux aux classifieurs à faible biais et variance élevée ⇒ particulièrement les arbres de décision
- ⇒ La stratégie d'élagage est simple : préférer le bagging sur des arbres profonds
- ☹ Bagging inefficace sur un classifieur fort, dont les différents modèles seront trop corrélés pour réduire la variance
- R : packages *ipred, randomForest*

Forêts aléatoires 1/3

- Le bagging manque d'efficacité quand les modèles sont trop corrélés $\Rightarrow$ on veut donc les décorréler
- Introduction d'une 2e randomisation : sur les individus (bagging) mais aussi sur les prédicteurs, en ajoutant à chaque scission un tirage aléatoire d'un sous-ensemble de taille *q* (constante) parmi l'ensemble des *p* prédicteurs (forêts aléatoires, Breiman, 2001)
- Plus la corrélation baisse (elle peut atteindre $\rho = 0{,}05$) plus la variance du modèle agrégé diminue : $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$
- Évite de voir apparaître trop souvent les mêmes variables les plus discriminantes
- Chaque arbre élémentaire est moins performant mais l'agrégation conduit à un modèle agrégé plus performant : l'augmentation du biais est plus que compensée par la diminution de la variance
- R : packages *randomForest* (sur arbre CART) et *party* (sur arbre Ctree)

Forêts aléatoires 2/3

- Diminuer le nombre q de prédicteurs augmente le biais mais diminue la corrélation entre les modèles et la variance du modèle final $\Rightarrow$ q permet de régler le trade-off entre biais et variance
- Pour le classement, Breiman suggère un sous-ensemble de $q = \sqrt{p}$ variables ou [log(p)+1] ou 1
- Mais il ne faut pas que ce nombre q soit trop faible si une forte proportion de variables sont peu discriminantes
- Les forêts aléatoires commencent à être très efficaces lorsque la probabilité de sélectionner un prédicteur discriminant est > 0,5. Cette probabilité est donnée par la loi hypergéométrique.
- Si 6 variables discriminantes sont mélangées à 30 non discriminantes, la probabilité de tirer au moins une variable discriminante parmi 6 est :
  - > cumsum(dhyper(1:6, 6, 30, 6))
  - [1] 0.4389771 0.6500237 0.6917119 0.6950619 0.6951543 **0.6951548**

Forêts aléatoires 3/3

- Le nombre q de prédicteurs est le seul paramètre réellement à régler
- Le nombre d'itérations est moins sensible à régler, et on a intérêt à le choisir assez élevé
  - Le nombre de modèles à agréger devrait croître avec le nombre de prédicteurs
  - À noter la convergence des performances atteinte avec un nombre de modèles agrégés parfois très inférieur au nombre de combinaisons de p variables parmi n ($n!/p!(n–p)!$), c'est-à-dire bien avant que toutes les combinaisons possibles de variables soient apparues.
  - Les forêts aléatoires résistent bien au sur-apprentissage (contrairement aux réseaux de neurones et au boosting) même quand le nombre de modèles agrégés est grand

Similarités entre forêts aléatoires et régression pénalisée ridge

- Le rétrécissement λ des coefficients dans la régression ridge ⇔ sélection de $q < p$ prédicteurs aléatoirement parmi les p prédicteurs
- Augmenter λ ou réduire q :
  - augmente le biais, puisque la solution est cherchée dans un sous-espace fixé par la contrainte
  - réduit la variance, de façon à compenser la hausse du biais
- Autre analogie : tous les prédicteurs peuvent apparaître dans le modèle
  - par rétrécissement de leurs coefficients dans la régression ridge
  - ou par sélection au hasard dans les forêts aléatoires
  - $\Rightarrow$ le travail de sélection des variables est simplifié !
- ☺ Pouvoir prédictif élevé !

Différences entre forêts aléatoires et régression pénalisée ridge

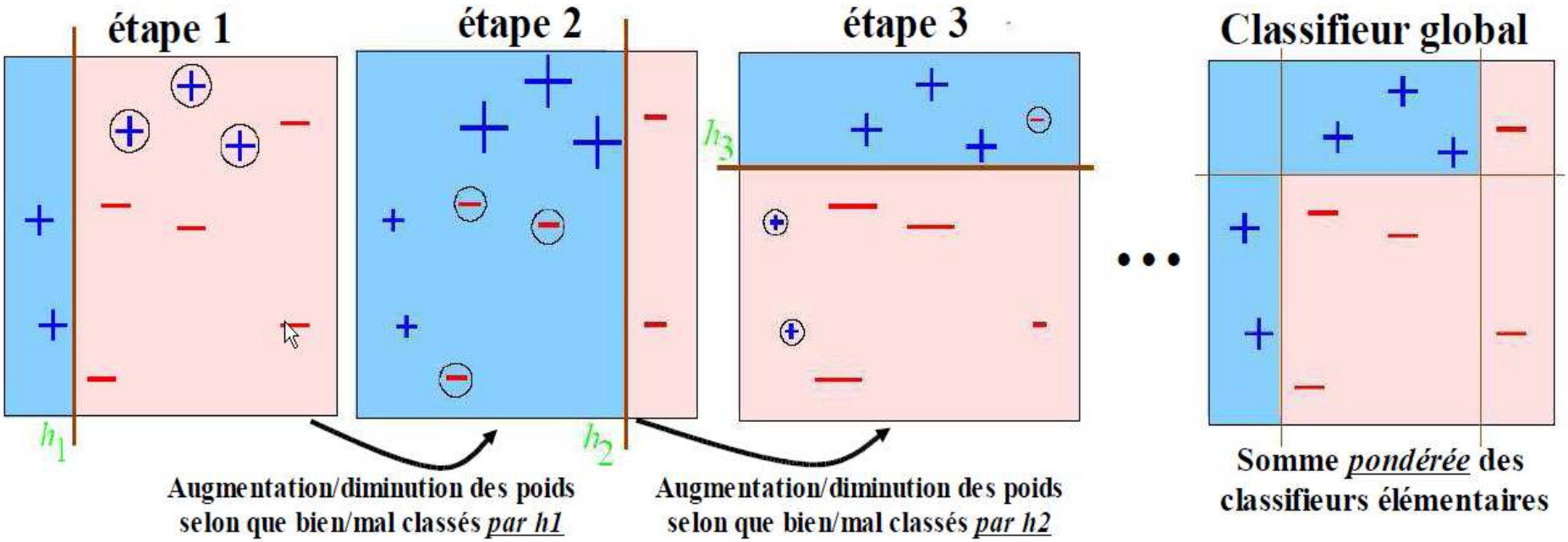
- Le paramètre de pénalisation λ permet un ajustement continu du biais-variance, alors que le nombre q est discret
- Il permet d'ajuster les coefficients à l'aide du ridge plot
  - en sorte que tous les coefficients aient un signe cohérent
  - voire que certains coefficients soit supérieur à un certain seuil fixé par les experts du domaine
- La régression pénalisée est déterministe
- Les calculs de la régression pénalisée sont plus rapides
  - mais les calculs des forêts aléatoires peuvent être parallélisés
- ☹ Manque de lisibilité d'un modèle de forêts aléatoires, qui détruit la structure d'arbre

Agrégation de modèles : le boosting

- BOOSTING, Freund et Schapire, 1996
- Algorithme adaptatif et généralement déterministe :
  - on travaille souvent sur toute la population
  - et à chaque itération, on augmente le poids des individus mal classés ou mal ajustés dans les itérations précédentes
  - à la fin, on agrège les modèles en les pondérant par leur qualité
- Diminue le biais et pas seulement la variance (grâce au mécanisme d'agrégation) mais peut être sujet au sur-ajustement
- Nombreux algorithmes : Discrete AdaBoost, Real AdaBoost, Gentle AdaBoost, LogitBoost, Arcing (Adaptative Resampling and Combining)…
- Performances pas toujours très différenciées sur des données réelles (voir plus loin l'arc-x4 de Breiman)
- R : packages *ada*, *gbm* et *mboost*

Illustration (Robert Schapire)

Extrait d'une conférence visible ici :
http://videolectures.net/mlss05us_schapire_b/



Algorithme Discrete AdaBoost

- 1) Initialiser les poids des N individus de l'échantillon d'apprentissage : $p_i = 1/N, i = 1, 2, \ldots, N$
- 2) Répéter pour m = 1 à M
  - ajuster le classifieur $f_m(x) \in \{-1,+1\}$ sur l'échantillon d'apprentissage pondéré par les poids p_i
  - calculer le taux d'erreur ε_m de $f_m(x)$ (tenant compte du poids de chaque observation mal classée) et calculer $\alpha_m = \ln((1-\varepsilon_m)/\varepsilon_m)$
  - on peut multiplier α_m par un paramètre de pénalisation $\lambda \geq 1$
  - si $\varepsilon_m < 0,5$, multiplier le poids p_i de chaque observation mal classée par $\exp(\alpha_m)$ (sinon : interrompre l'algorithme ou réinitialiser les poids) – le multiplicateur décroît avec le taux d'erreur
  - normaliser les poids p_i pour que leur somme soit 1
- 3) Le classifieur boosté est le signe de la somme $\sum_m \alpha_m f_m(x)$ (ou la valeur moyenne des $\alpha_m f_m(x)$)

Algorithme Arcing

- 1) Initialiser les poids des N individus de l'échantillon d'apprentissage : $p_i = 1/N, i = 1, 2, \ldots, N$
- 2) Répéter pour m = 1 à M
  - dans l'échantillon d'apprentissage, tirer avec remise N individus chacun selon la probabilité p_i
  - ajuster le classifieur $f_m(x) \in \{-1,+1\}$ sur l'échantillon ainsi tiré
  - sur l'échantillon d'apprentissage initial :
    - calculer le taux d'erreur ε_m pondéré des observations mal classées par $f_m(x)$ et calculer $\alpha_m = \ln((1-\varepsilon_m)/\varepsilon_m)$
    - si $\varepsilon_m < 0{,}5$, multiplier le poids p_i de chaque observation mal classée par $\exp(\alpha_m)$ pour $i = 1, 2, \ldots, N$ (sinon : interrompre l'algorithme ou réinitialiser les poids)
    - normaliser les poids p_i pour que leur somme soit 1
- 3) Le classifieur boosté est le signe de la somme $\sum_m \alpha_m f_m(x)$ (ou la valeur moyenne des $\alpha_m f_m(x)$)

Intérêt de l'algorithme Arcing

- L'arcing introduit un facteur aléatoire par un tirage avec remise et avec une probabilité de tirage plus importante pour les individus mal classés à l'itération précédente
  - contrairement au Discrete AdaBoost qui conserve chaque individu en modifiant son poids mais non sa probabilité d'être tiré
- Ce tirage aléatoire introduit une plus grande diversité dans les modèles obtenus et agrégés
- Variante arc-x4 de l'arcing
  - à chaque itération, le poids d'un individu est proportionnel à la somme de 1 et des puissances 4^e des nombres d'erreurs de classement des itérations précédentes
  - Breiman (Breiman, 1996) a choisi la puissance 4^e de façon empirique après avoir testé plusieurs valeurs
  - performances comparables à celle de l'algorithme standard
  - montre que l'efficacité d'un algorithme de boosting vient moins de son dispositif spécifique de pondération des observations que de son principe général de rééchantillonnage adaptatif

Algorithme Real AdaBoost

- 1) Initialiser les poids des N individus :
  - $p_i = 1/N, i = 1, 2, \ldots, N$
- 2) Répéter pour m = 1 à M
  - calculer la probabilité $p_m(x) = P(Y = 1|x)$ sur l'échantillon d'apprentissage pondéré par les poids p_i
  - calculer $f_m(x) = ½ \text{ Log}(p_m(x)/(1-p_m(x))$
  - multiplier le poids p_i de chaque observation (x_i,y_i) par $\exp(- \lambda. y_i.f_m(x_i))$ pour $i = 1, 2, \ldots, N$, où $\lambda \geq 1$ est un paramètre de pénalisation
  - normaliser les poids p_i pour que leur somme soit 1
- 3) Le classifieur boosté est le signe de la somme $\sum_m f_m(x)$

Comparaison des caractéristiques

BAGGING	FORÊTS ALÉATOIRES	BOOSTING
Le *bagging* est un mécanisme aléatoire	Idem bagging	Le *boosting* est un mécanisme adaptatif et généralement (sauf l'arcing) déterministe
À chaque itération, l'apprentissage se fait sur un échantillon bootstrap différent	Idem bagging	Généralement (sauf l'arcing), à chaque itération, l'apprentissage se fait sur l'échantillon initial complet
À chaque itération, l'apprentissage se fait sur l'ensemble des prédicteurs	À chaque itération, l'apprentissage se fait sur un sous-ensemble aléatoire de prédicteurs	À chaque itération, l'apprentissage se fait sur l'ensemble des prédicteurs
À chaque itération, le modèle produit doit être performant sur l'ensemble des observations	À chaque itération, le modèle produit doit aussi être performant sur l'ensemble des observations, mais l'est moins que le bagging, puisque tous les prédicteurs ne sont pas utilisés	À chaque itération, le modèle produit doit être performant sur certaines observations ; un modèle performant sur certains *outliers* sera moins performant sur les autres observations
Dans l'agrégation finale, tous les modèles ont le même poids	Idem bagging	Dans l'agrégation finale, les modèles sont généralement pondérés selon leur d'erreur

Comparaison des points forts / faibles

BAGGING	FORÊTS ALÉATOIRES	BOOSTING
Réduction de la variance par moyenne de modèles	Idem bagging, mais avec une plus grande réduction de la variance	Peut diminuer la variance et le biais du classifieur de base Mais la variance peut augmenter avec un classifieur de base stable
Perte de lisibilité sur des arbres de décision	Idem	Idem
Peu efficace sur les « stumps »	Efficace sur les « stumps »	Très efficace sur les « stumps »
Convergence plus rapide	Idem bagging	Convergence plus lente
Possibilité de paralléliser l'algorithme	Idem bagging	Algorithme séquentiel ne pouvant être parallélisé
Pas de sur-apprentissage : supérieur au boosting en présence de « bruit »	Idem bagging	Risque de sur-apprentissage si le nombre d'itérations est grand
Le bagging est le plus simple à mettre en œuvre mais est généralement moins discriminant que les forêts aléatoires et le boosting	Les forêts aléatoires sont toujours supérieures au bagging et assez souvent plus que le boosting (sauf si les prédicteurs discriminants sont très rares)	Le boosting est souvent plus efficace que le bagging, du moins sur les données non bruitées

Méthodes pour le Big Data

Méthodes pour le Big Data

- Les questions d'échantillonnage sont importantes, puisqu'elles peuvent permettre de diminuer le volume de données et d'inférer des conclusions générales à partir d'observations partielles
  - Mais la représentativité des échantillons est délicate à établir, avec des sources de données multiples, qui ne couvrent pas les mêmes populations et comportent un nombre important de valeurs manquantes
  - Il faut réussir à apparier les données et redresser les échantillons
- L'étude des matrices en grande dimension survient avec des matrices dont les lignes sont des clients et les colonnes des produits téléchargés, achetés ou recommandés
  - Il peut aussi s'agir de matrices représentant des relations entre individus ou entre institutions financières cotées (rendements journaliers croisés avec les rendements décalés) dans un contexte d'étude du risque systémique
- Un autre axe de recherche porte sur la visualisation des données en grande dimension

De nouvelles problématiques

- Les données fonctionnelles sont des données qui ne sont pas ponctuelles mais sont continues, comme des courbes ou des images
  - Ces données se sont multipliées avec les progrès technologiques qui permettent la collecte et le stockage d'observations de plus en plus fines, captant en continu les informations sur un objet étudié (météorologique, environnemental, médical, alimentaire…)
  - Au lieu de discriminer des individus au vu de quelques caractéristiques à des instants choisis, on n'a pas d'a priori sur le moment et la durée des différences entre deux courbes d'évolution
- Dans les problématiques liées au web, on ne recherche pas systématiquement des modèles robustes et lisibles, mais des modèles construits rapidement sur des micro-segments mouvants, afin de prédire les comportements ou les préférences d'un petit nombre d'internautes

p >> n

- La régression en grande dimension pose le problème classique de sélection des variables
- On rencontre aussi, par exemple en bio-statistique (séquençage de l'ADN) ou en chimiométrie (statistique appliquée aux données chimiques), des situations où le nombre de variables est supérieur, voire très supérieur, au nombre d'individus (on parle de tableaux plats), et où les méthodes classiques de régression ne s'appliquent pas et cèdent la place à des méthodes telles que la régression Lasso ou PLS
- Le nombre de variables étudiées peut varier entre 10^4 et 10^8, alors que le nombre d'observations est de quelques centaines

Nouvelles approches en machine learning

- Les méthodes de machine learning (agrégation de modèles, SVM, réseaux de neurones…) sont utilisées pour leur pouvoir prédictif élevé, dans des situations où la lisibilité du modèle n'est pas recherchée et où leur caractéristique « boîte noire » n'est pas un inconvénient
- Exemple d'une librairie en ligne, qui veut proposer des titres à ses clients. Dans ce problème, les variables (titres déjà achetés) sont excessivement nombreuses et créent des matrices creuses difficiles à modéliser. L'approche courante est de décomposer la clientèle en un très grand nombre de segments, éventuellement des milliers, recalculés en permanence par des techniques statistiques qui permettent de situer chaque client dans un petit segment de clients ayant des goûts proches. Ensuite, on lui propose les titres souvent acquis par les autres clients de son segment, que lui-même n'aurait pas encore acquis. Ces calculs sont refaits en permanence, sans recherche de segments et de modèles robustes et compréhensibles.

Les méthodes d'agrégation

- Les méthodes d'agrégation, ou méthodes d'ensemble, ainsi que le stacking, consistent à combiner entre elles des méthodes prédictives
- Dans le stacking, on combine différentes méthodes ; dans les méthodes d'agrégation, on applique un grand nombre de fois la même méthode
- Quand on agrège des modèles prédictifs, parfois simplement en faisant la moyenne de leurs prédictions, il vaut mieux agréger des modèles moins poussés, individuellement moins performants, pour obtenir un modèle final plus performant !
- Cela vient de ce que les modèles individuels plus poussés se ressemblent plus, et que le gain de leur agrégation est beaucoup moins grand
- On touche ici au besoin de puissance du Big Data, car ces méthodes peuvent être très gourmandes en temps de calcul

Remarque sur les méthodes appliquées au Big Data

- Toutes les méthodes utilisées pour le Big Data ne sont pas très récentes, et la plupart faisait du Big Data comme Monsieur Jourdain : « Par ma foi ! il y a plus de quarante ans que je dis de la prose sans que j'en susse rien, et je vous suis le plus obligé du monde de m'avoir appris cela. » Molière, *Le Bourgeois gentilhomme*, 1670
- À côté des méthodes classiques, on utilise des méthodes plus modernes (les méthodes « d'agrégation » par exemple) mais aussi des perfectionnements très récents de méthodes classiques (les méthodes « pénalisées », par exemple)
- C'est comme la musique dite classique, qui ne s'est pas arrêtée au XIXe siècle, et qui s'enrichit en permanence de nouvelles œuvres, certaines plus novatrices et originales que des œuvres de musique dite moderne (écoutons par exemple Henri Dutilleux)

Algorithme PageRank 1/2

- La structure d'une base de données classique permet d'en extraire des informations
- Mais le Web est immense et peu structuré
- La recherche par mots-clés ne permet pas de limiter suffisamment le nombre de réponses
- D'où la recherche d'un algorithme pour trier les réponses selon leur pertinence ⇒ algorithme PageRank de Google (cofondateur Larry Page)
- Principe : classement des pages Web selon leur popularité sur le Web, donc selon le nombre de lien pointant sur elles
- Un lien d'une page A vers une page B augmente le PageRank de B
  - l'augmentation du PageRank de la page B est d'autant plus importante que le PageRank de la page A est élevé
  - l'augmentation du PageRank de la page B est d'autant plus importante que la page A fait peu de liens

Algorithme PageRank 2/2

- Soient A_1, A_2, ..., A_n les pages pointant vers une page B, $PR(A_k)$ le PageRank de A_k, $N(A_k)$ le nombre de liens sortants présents sur la page A_k, et d un facteur compris entre 0 et 1, souvent fixé à 0,85
- $PR(B) = (1-d) + \{ d \times [PR(A_1)/N(A_1) + ... + PR(A_n)/N(A_n)] \}$
  - Si aucune page ne pointe vers B, alors $PR(B) = 1-d$
  - $PR(A_k)$ = contribution de la page A_k à l'ensemble des autres pages
- Le PageRank dépend des liens et non des clics
- L'algorithme PageRank est inspiré par le système de référence des publications universitaires dans lequel la valeur d'une publication est déterminée par le nombre de citations que cette publication reçoit
- Référence : Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The pagerank citation ranking: bringing order to the web, *Technical report*, Stanford Digital Library Technologies Project

La détection des règles d'associations

Les recherches d'associations

- Rechercher les **associations** consiste à rechercher les règles du type :
  - « Si pour un individu, la variable A = x_A, la variable B = x_B, etc, alors, dans 80% des cas, la variable Z = x_Z, cette configuration se rencontrant pour 20 % des individus »
- La valeur de 80% est appelée **indice de confiance** et la valeur de 20% est appelée **indice de support**
- Par exemple, dans l'ensemble de transactions ci-contre :
  - l'indice de confiance de « B $\Rightarrow$ E » = 3/4
  - l'indice de support de « B $\Rightarrow$ E » = 3/5

T26	**A**	**B**	**C**	**D**	**E**
T1245	**B**	**C**	**E**	**F**	
T156	**B**	**E**			
T2356	**A**	**B**	**D**		
T145	**C**	**D**			

Les associations : définitions

- Une règle est donc une expression de la forme :
  - Si *Condition* alors *Résultat*
- Synonymes :
  - Condition = Antécédent
  - Résultat = Conséquent
- Les éléments d'une règle $\{A = x_A, B = x_B, ...\} \Rightarrow \{Z = x_Z\}$ sont les *items*
- Exemple :
  - Si *riz* et *vin blanc*, alors *poisson*
- L'indice de support est la probabilité :
  - Prob (*condition* et *résultat*)
- L'indice de confiance est la probabilité :
  - Prob (*condition* et *résultat*) / Prob (*condition*)

Intérêt d'une règle d'association

- Dans l'exemple précédent, on a :
  - indice de confiance de l'association C $\Rightarrow$ B est 2/3
  - indice de support = 2/5
- Or, Prob (B) = 0,8
  - B est présent dans presque tous les tickets de caisse
- Cette probabilité est supérieure à l'indice de confiance de C $\Rightarrow$ B, ce qui fait que l'on ne gagne rien à utiliser la règle C $\Rightarrow$ B pour prédire B
- Si l'on suppose aléatoirement qu'un ticket de caisse contient B, on n'a qu'1 chance sur 5 de se tromper, contre 1 chance sur 3 en appliquant la règle C $\Rightarrow$ B

Lift d'une règle : mesure son intérêt

- L'amélioration apportée par une règle, par rapport à une réponse au hasard est appelée « lift » et vaut :
  - lift (règle) = confiance (règle) / Prob (*résultat*)
  - = Prob (*condition* et *résultat*) / [Prob (*condition*) x Prob (*résultat*)]
- Quand le lift est < 1, la règle n'apporte rien
  - car Prob (*résultat*) > indice de confiance (règle)
- Exemples :
  - lift (C $\Rightarrow$ B) = 5/6 (règle inutile)
  - lift (B $\Rightarrow$ E) = 5/4 (règle utile)

Lift de la règle inverse

- Il faut noter que si le lift de la règle
  - Si *Condition* alors *Résultat*
- est < 1, alors le lift de la règle inverse, c.a.d. de :
  - Si *Condition* alors *NON Résultat*
- est > 1, puisque :
  - confiance (règle inverse) = 1 - confiance (règle)
- et
  - Prob (*NON résultat*) = 1 - Prob (*résultat*)
  - d'où Prob (*NON résultat*) < confiance (règle inverse)
- Si une règle n'est pas utile, on peut donc essayer la règle inverse… en espérant que cette dernière soit intéressante en termes de métier ou de marketing

Algorithme *Apriori*

- C'est l'algorithme le plus répandu (Agrawal et al.)
- Il fonctionne en deux étapes :
  - il commence par rechercher les sous-ensembles d'items ayant une probabilité d'apparition (support) supérieure à un certain seuil *s*
    - 1e passe : élimination des items moins fréquents que *s*
    - 2e passe : constitution des combinaisons de deux items parmi les précédents, et élimination des combinaisons moins fréquentes que *s*
    - etc : les ensembles fréquents de taille *n* qui nous intéressent sont ceux provenant d'ensembles de taille $n - 1$ eux-mêmes fréquents
  - puis il tente de décomposer chaque sous-ensemble sous une forme {Condition $\cup$ Résultat} telle que le quotient « Prob (Condition et Résultat) / Prob (Condition) » (indice de confiance), soit supérieur à un certain seuil
    - difficulté : pour chaque sous-ensemble d'items E à *n* éléments, il y a $2^{n-1} - 1$ règles de la forme $A \Rightarrow \{E - A\}$
    - optimisation d'*Apriori* pour l'identification des règles à conserver

Mise en œuvre

- En pratique, les règles demeurent très nombreuses, et la plupart des logiciels permettent de stocker ces règles dans un fichier, dans lequel il est possible de filtrer les règles *Condition* $\Rightarrow$ *Résultat* en deçà d'un certain indice de support, et de les trier selon leur support, leur confiance ou leur lift
- On est généralement plus sévère sur le seuil de confiance que de support, surtout si l'on recherche des règles rares, et un exemple courant de filtre sera 75 % pour la confiance et 5 % pour le support (et bien sûr 1 pour le lift)
- Même avec ces filtres, le nombre de règles peut vite atteindre plusieurs millions pour seulement quelques centaines d'items et quelques milliers d'observations
- Certains logiciels permettent d'ajouter un filtre sur le contenu des règles, pour ne conserver que celles qui contiennent un item donné dans leur résultat ou leurs conditions
- Les logiciels permettent aussi de fixer une limite à la taille des règles : on dépasse rarement 10 items

Taxinomie : définition

- Les produits peuvent être définies avec un niveau plus ou moins fin de détail
- On peut par exemple considérer :
  - les produits d'épargne bancaire, financière…
  - parmi les produits d'épargne bancaire, les comptes de chèques, les livrets…
  - parmi les livrets, les livrets A, les Codevi, les LEP…

- La **taxinomie** des produits est l'ensemble de ces niveaux

Taxinomie : utilisation

- Le **niveau le plus fin** permet d'entreprendre des actions commerciales plus précises
  - Mais travailler au niveau le plus fin multiplie les règles, parmi lesquelles un grand nombre n'auront qu'un faible support et seront peut-être éliminées
- Travailler au **niveau le plus général** permet d'obtenir des règles plus fortes

> Les 2 points de vue ont leurs avantages et leurs inconvénients

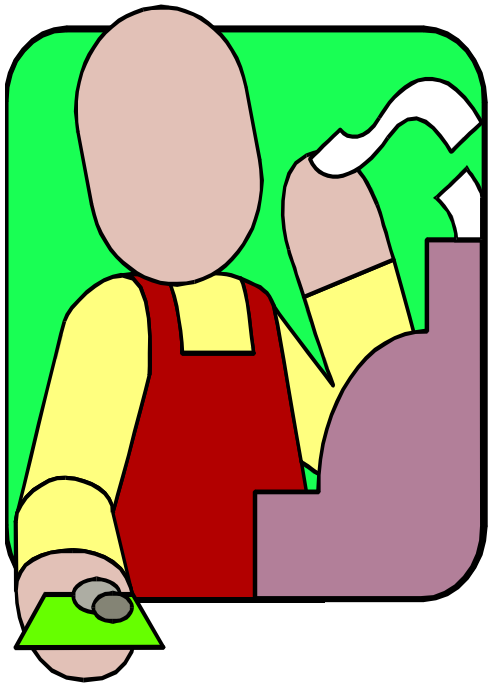
> Il faut adapter le niveau de généralité à chaque produit, en fonction notamment de sa rareté

Taxinomie : intérêt

- Les articles les plus rares et les plus chers (exemple : micro-informatique ou HIFI dans un grand magasin) seront codifiés au niveau le plus fin
- Les articles les plus courants (exemple : produits alimentaires) seront codifiés à un niveau plus général
- On regroupera par exemple tous les yaourts, fromages blancs, flancs… en « produits laitiers », tout en distinguant un téléviseur d'un magnétoscope ou d'un caméscope
- L'intérêt de cette façon de procéder est d'obtenir des règles plus pertinentes, dans lesquelles les articles les plus courants ne dissimulent pas, par leur fréquence, les articles les moins courants

L'analyse du ticket de caisse

- Cette technique est très utilisée dans la *grande distribution* :
  - d'où les termes d'analyse du **ticket de caisse** ou du **panier de la ménagère** (market basket analysis) pour désigner la recherche d'associations
- Autres usages :
  - associations d'options retenues dans les produits packagés (banque, téléphonie, assurance…)
  - web mining (analyse de la navigation sur un site internet)
- Difficultés :
  - volumes de données importants
  - trouver des règles intéressantes noyées parmi les règles triviales ou non utilisables



Utilisation de variables supplémentaires

- En ajoutant des variables temporelles (jour et heure de la transaction), on pourra rechercher l'ensemble des événements qui débouchent sur l'acquisition d'un nouveau produit, sur le départ du client…
- En ajoutant le nom du fabricant, on pourra détecter des phénomènes d'attachement à une marque
- Autres variables supplémentaires :
  - canal de distribution
  - mode de paiement
  - …
- Le développement des cartes de fidélité permet de croiser les achats avec de nombreuses autres données : âge, adresse…

Conclusion

Perspectives professionnelles

- Finance
  - Réglementations Bâle II (et Bâle III)
  - Évolution des marchés boursiers
- Marketing
  - Dont marketing direct et sur le web
  - Étude des préférences et des comportements des consommateurs
  - Revenue management
- Assurance (scoring et actuariat)
- Industrie
  - Contrôle qualité
- Industrie pharmaceutique, santé
  - Tests cliniques, pharmacovigilance, épidémiologie
- Médecine
  - Analyses de survie, causes, prévention et traitement des maladies
- Environnement et Météorologie
  - Études sur le climat, la pollution
- Recherche scientifique
- …

Le Big Data et l'emploi

- Le Big Data fait partie des 34 plans industriels lancés par le gouvernement français le 12 septembre 2013
- Le Big Data a besoin de « data scientists » qui connaissent :
  - les enjeux métiers (marketing, risque, production…)
  - les technologies informatiques (architecture, algorithmes, logiciels)
  - les méthodes de statistique et de machine learning
- Des centaines de milliers d'emplois de data scientists annoncés dans le monde
- Le manque de data scientists se fait sentir dans tous les pays. On peut l'imputer à la prise de conscience récente du potentiel recelé par les données, et à une valorisation encore insuffisante du data scientist en entreprise.
- Premières formations spécialisées en 2013 aux USA et en France

Quelques liens

- Site de la Société Française de Statistique : www.sfds.asso.fr
- Site de Gilbert Saporta (contenu riche, avec de nombreux cours) : http://cedric.cnam.fr/~saporta/
- Site de Philippe Besse (très complet sur les statistiques et le data mining) : www.math.univ-toulouse.fr/~besse/
- Site du livre *The Elements of Statistical Learning* de Hastie, Tibshirani et Friedman : http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Un livre complémentaire : http://www-bcf.usc.edu/~gareth/ISL/index.html
- StatNotes Online Textbook (statistiques) : www2.chass.ncsu.edu/garson/pa765/statnote.htm
- Statistique avec R : http://zoonek2.free.fr/UNIX/48_R/all.html
- Données réelles : http://www.umass.edu/statdata/statdata/index.htm
- Site d'Olivier Decourt (spécialiste de SAS) : www.od-datamining.com/
- Blog d'Arthur Charpentier : http://freakonometrics.blog.free.fr/