Ambassade de France à Washington
Service pour la Science et la Technologie
4101 Reservoir Road NW, Washington, DC 20007
Tél. : +1 202 944 6246
Mail : info@france-science.org
URL : http://france-science.org

Domaine : Mathématique, Informatique, Santé, Physique, Politique Scientifique
Document : Rapport d'Ambassade
Titre : Recherche, Innovation et Formations sur le Big Data aux Etats-Unis
Auteur(s) : Sébastien Fischman
Date : Septembre 2013
Contact : Marc Daumas (attache-ntics@ambascience-usa.org)

Mots-clés :	Politique scientifique, Recherche et Développement
Résumé :	L'évolution récente des technologies (Internet, smart phones, appareils photos, caméras, satellites...) a fait exploser la quantité de données créées et stockées. Chaque jour, 540 millions de SMS sont envoyés dans le monde, 143 milliards de courriels sont échangés, 40 000 giga-octets de données sont produites au LHC (Large Hadron Collider), 400 millions de tweets sont postés, 104 000 heures de vidéos sont ajoutées sur Youtube, etc. Le Big Data va certainement trouver des applications dans tous les domaines imaginables. La médecine, l'astrophysique, les sciences sociales ou la recherche scientifique en général voient s'ouvrir les portes d'un univers différent. Cependant, au delà de toutes les promesses du Big Data, cette révolution amène également de nombreuses interrogations. Il est primordial de ne pas tomber dans ce que certains appellent déjà « la dictature des données ». Faire des prédictions ou prendre des décisions en suivant aveuglément des schémas que l'on entrevoit dans un amas de données non pertinentes est une grossière erreur. Former des « data scientists » compétents et capables de tirer profit du Big Data tout en sensibilisant l'ensemble de la société, y compris nos responsables politiques, à ce changement profond et à ses conséquences : voici le nouveau défi du Big Data.

NB : Retrouvez toutes nos publications sur : http://france-science.org/Bulletins-Electroniques-Etats-Unis.html

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Recherche, Innovation et Formations sur le Big Data aux Etats-Unis

Ambassade de France à Washington

Sébastien Fischman
sebastien.fischman@ens-cachan.fr

Travaux effectuées de mars à juillet 2013

Table des matières

Remerciements

Je tiens ici à remercier toutes les personnes qui m'ont aidé pendant la rédaction de ce rapport et sans qui il n'aurait pu voir le jour.
Tout d'abord merci à Kirk Borne, professeur d'astrophysique et d'informatique à George Mason University et à Jonathan Epstein, bio-informaticien au NIH (National Institute of Health) pour m'avoir accordé un peu de leur temps pour répondre à mes questions et m'éclairer sur les problématiques des chercheurs.
Merci à Marc Daumas qui a supervisé depuis le début l'avancement de ce rapport.
Merci également à Annick Suzor-Weiner pour son accueil chaleureux au sein de la Mission Scientifique et Technologique.
Pour finir, je tiens à remercier mes collègues Marie Imbs, Thomas Debacker et Frédéric Lohier pour leurs précieuses relectures et corrections qui ont très certainement contribué à rendre ce rapport meilleur.

Introduction

Le monde du numérique a récemment basculé dans une nouvelle ère : celle du Big Data. En effet, l'évolution récente des technologies (Internet, smart phones, appareils photos, caméras, satellites...) a fait exploser la quantité de données créées et stockées. Tous les deux jours [1], nous générons autant de données que l'ensemble de l'Humanité depuis le début de son existence jusqu'en 2003, 90% des données existantes ont été générées durant ces deux dernières années. Chaque jour [9], 540 millions de SMS sont envoyés dans le monde, 143 milliards de courriels sont échangés, 40 000 giga-octets de données sont produites au LHC (Large Hadron Collider), 400 millions de tweets sont postés sur Twitter, 104 000 heures de vidéos sont ajoutées sur Youtube... Et la quantité de données que nous générons double tous les deux ans [2]. Et pourtant seulement 0,05% des données digitales que l'on crée sont analysées [25][26].

La question qui se pose alors est la suivante : comment pouvons-nous utiliser intelligemment cette immense masse de données et ne pas laisser disparaître toutes ces informations, submergés par leur nombre ? C'est le défi que propose de relever le Big Data. Ce terme de «Big Data», très à la mode, est souvent utilisé à tort et à travers. En effet, les promesses de cette nouvelle approche scientifique sont nombreuses [4] ce qui explique un tel engouement. La notion de Big Data étant récente il n'existe pas de définition bien arrêtée. Le terme aurait été inventé par le cabinet d'étude Gartner en 2008. Le NIST (National Institute of Standards and Technology) vient de lancer le 19 juin 2013 un programme ouvert pour s'accorder sur une définition précise afin de faciliter les échanges marchands et contractuels. Nous entendons ici par «Big Data» toute méthode (stockage, distribution, analyse...) ayant pour but d'extraire des informations et mettant en jeu un gros volume de données (de l'ordre du peta-octet, 10^{15} octets) ou un gros flux de données et dont l'étude nécessite des techniques avancées. L'une des difficultés principales, au delà de la taille des bases de données et de la contrainte temporelle (résoudre un problème quasiment en temps réel est souvent l'une des exigences du Big Data), est l'extraction d'informations au sein de données non structurées (photos, vidéos, commentaires écrits, courriels...) que l'ordinateur ne peut comprendre tout seul. Comme le dit Kenneth Cukier auteur de «Big Data : A revolution that will transform how we live, work, and think.» [4], on fait du Big Data dès lors qu'on tire profit de beaucoup de données et que l'extraction d'informations n'aurait pu être possible avec moins de données. Parvenir à gérer les données stockées et à en obtenir des informations pertinentes est aujourd'hui devenu un enjeu majeur dans tous les domaines. Aux Etats-Unis, l'administration du président Obama a annoncé le 29 Mars 2012, la création d'un fond de 200 millions de dollars pour investir dans la R&D liée au Big Data sous le nom de «Big Data Research and Development Initiative» [3]. En France, Fleur Pellerin (Ministre déléguée aux PME, à l'Innovation et à l'Économie numérique), Arnaud Montebourg (Ministre du Redressement productif) et Louis Gallois (commissaire général à l'investissement) ont annoncé, le 5 avril 2013, une aide de 11,5 millions d'euros dans 7 projets «Big Data» dans le cadre des Investissements d'Avenir [4].

1. C'est ce qu'a affirmé Eric Schmidt, ancien PDG de Google, lors d'une conférence aux Etats-Unis en 2010.
2. Ce phénomène peut être vu comme l'un des aspects de la loi de Moore.
3. Voir le communiqué officiel `http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf`.
4. Voir l'appel à projet Big Data `http://investissement-avenir.gouvernement.fr/content/big-data`.

Ce rapport a pour objectif d'expliquer, dans un premier temps, les idées et principes fondamentaux à la base du Big Data. Il sera porté, tout au long du rapport, une attention plus particulière aux aspects mathématiques que l'on pourra rencontrer. L'objectif étant de comprendre le rôle essentiel que joue la mathématique dans cette révolution numérique.

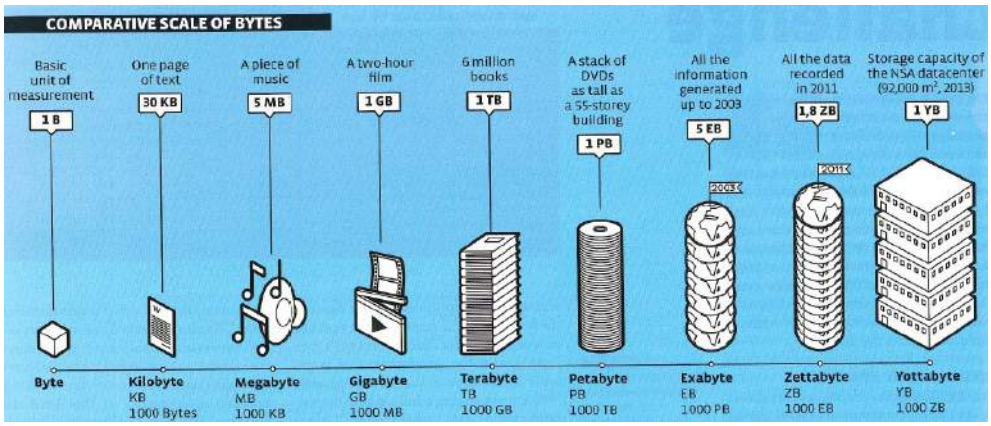
Cependant, pour faciliter la lecture de ce rapport, nous avons décidé de changer de police pour les parties «techniques» qui fourniront aux seuls lecteurs intéressés de plus amples informations.

Nous parlerons d'abord succinctement du Map Reduce (section 1 p.9), méthode informatique fondamentale pour le Big Data et de calcul parallèle. Puis nous nous arrêterons plus longuement sur le cas de Google (section 2 p.11), pionnier et leader dans le Big Data.
Nous ferons ensuite un tour d'horizon des très nombreux domaines d'application du Big Data. Nous verrons qu'il jouera un rôle central dans les années à venir, aussi bien dans le domaine industriel ou marketing (section 6 p.27) qu'aux niveaux de la recherche scientifique (section 5 p.21) et des politiques publiques (section 7 p.32). Nous mettrons également en avant un échantillon d'entreprises américaines présentes sur le marché du Big Data, promis à un avenir florissant. Un rapport de l'institut McKinsey [20] affirme en effet qu'il manquait en 2011 aux Etats-Unis entre 140 000 et 190 000 «Data Analysts» hautement qualifiés et 1,5 millions de managers/analystes capables de relever les défis du Big Data.
Pour finir, nous verrons quels types de formations universitaires sont proposés aux Etats-Unis pour devenir "data analyst" et quelles mathématiques sont nécessaires pour évoluer dans ce nouvel eldorado (partie III p.37). Il est en effet indispensable pour un «data scientist» de posséder, parmi d'autres compétences, des bases solides en mathématiques.

Première partie

Idées et principes fondamentaux du Big Data

Notre capacité à générer et stocker des données est en augmentation exponentielle (double tous les 18 mois). Alors que depuis toujours le manque d'informations était le principal frein à la compréhension du monde qui nous entoure, c'est aujourd'hui la gestion des données qui pose problème. L'évolution extrêmement rapide des technologies contribue également au flou entourant la définition du Big Data. Ce qui semble être une grosse quantité de données aujourd'hui ne le sera plus d'ici quelques années. En 1990, la taille des disques durs sur le marché se comptait en méga-octets, en 2000 en giga-octets et aujourd'hui en téra-octets. Martin Hilbert, professeur à USC (University of Southern California), estime qu'en 2013, 1,2 zetta-octets (1,2 10^{21} octets) sont stockés à travers le monde [12]. Bien que le terme Big Data soit récent, le problème de la gestion de grandes quantités de données (par rapport aux capacités du moment) se pose depuis de nombreuses années. La nouveauté provient de l'explosion du nombre de capteurs en tout genre (téléphones, appareils photos, satellites, télescopes mais aussi les courriels et les messages sur les réseaux sociaux) qui permettent de «photographier» le réel, ce qui donne aux données collectées un intérêt accru. C'est une révolution conjoncturelle : nous disposons désormais d'énormément de données. Comment les utiliser ?



Ordres de grandeurs pour les échelles de données (extrait de [9]).

1 MapReduce, l'outil informatique omniprésent

Le MapReduce [8] est essentiellement une approche informatique du traitement des données qui consiste à faire fonctionner plusieurs ordinateurs parallèlement sur un même problème [35]. Elle ne fait appel à aucune considération mathématique particulière. C'est une méthode du type «Divide and conquer» ou «Diviser pour régner», ce qui n'est pas récent en informatique [1]. Cependant, c'est un outil central du Big Data et certainement l'un des plus largement utilisés. C'est pourquoi nous en expliquons brièvement le principe.

Le MapReduce repose en fait, comme son nom l'indique, sur deux fonctions : Map et Reduce.

La fonction Map : L'ensemble des données à traiter est reçu par un ordinateur principal («Master node» en anglais) qui va fractionner les données et les redistribuer à d'autres ordinateurs («node» ou «shrink node») qui peuvent éventuellement eux-mêmes redistribuer à d'autres nœuds, créant ainsi une structure arborescente. Cette structure fait souvent intervenir des centaines d'ordinateurs (ce qu'on appelle un «cluster») et peut aller au delà de 100 000 machines connectées. Le supercalculateur IBM Sequoia, élu machine la plus rapide du monde en 2012 fonctionne avec 98000 ordinateurs connectés les uns aux autres. Il est aujourd'hui 3ème au classement [2] Top500 de juin 2013. Chacun des ordinateurs mis en jeu va alors traiter les données qu'il détient en parallèle des autres ordinateurs et de manière indépendante. Ceci permet de traiter un grand nombre de données simultanément et plus rapidement.

La fonction Reduce : Lorsqu'un ordinateur du cluster a fini de traiter ses données il remonte le résultat à l'ordinateur principal. La fonction Reduce consiste donc à la collecte et au traitement, par l'ordinateur principal, des données intermédiaires. Le résultat final est ensuite rendu par l'ordinateur principal.

Avantages de la méthode : Cette méthode permet de traiter des quantités de données très massives sans disposer de machines extrêmement puissantes, la tâche étant répartie sur de nombreux ordinateurs de puissance moyenne (on parle de «commodity computing» ou COTS). La technologie est donc plus accessible et moins coûteuse. Il est par exemple possible de connecter plusieurs ordinateurs de particuliers en réseau même si une connection LAN efficace (Local Area Network) est indispensable. Sur le marché, la société Sablacore [3], implantée depuis 11 ans dans la location de services informatiques, propose de louer des calculateurs pour $\$0,20$ par cœur et par heure, à la demande. Pour se faire une idée, le plus grand supercalculateur du monde [4], le Tianhe-2 (MilkyWay-2) est chinois et possède 3 120 000 cœurs.
Un autre avantage de cette distribution du travail est d'éviter qu'une panne paralyse entièrement le système lors d'un gros traitement. Chaque ordinateur étant indépendant des autres,

1. Les langages LISP ont été inventés en 1958 et faisaient déjà appel à ce paradigme.
2. `http://www.top500.org/lists/2013/06/`.
3. Voir `http://www.sabalcore.com/`.
4. Classement au Top500 en juin 2013, `http://www.top500.org/lists/2013/06/`.

si un problème survient tout le système n'est pas interrompu. En pratique, chaque ordinateur envoie périodiquement un signal au poste principal qui, s'il ne reçoit pas de signal, considère l'ordinateur comme déficient et envoie les données traitées par cet ordinateur (s'il en existe une copie) à un autre qui le remplace.

Google fut l'un des premiers à utiliser le MapReduce en conjonction avec son Google File System, avec le succès que l'on connait. Aujourd'hui, le consensus se fait autour du logiciel open source Hadoop MapReduce. Toutes les grosses entreprises (Google, IBM, Microsoft...) utilisent Hadoop comme base dans leurs logiciels intégrés.

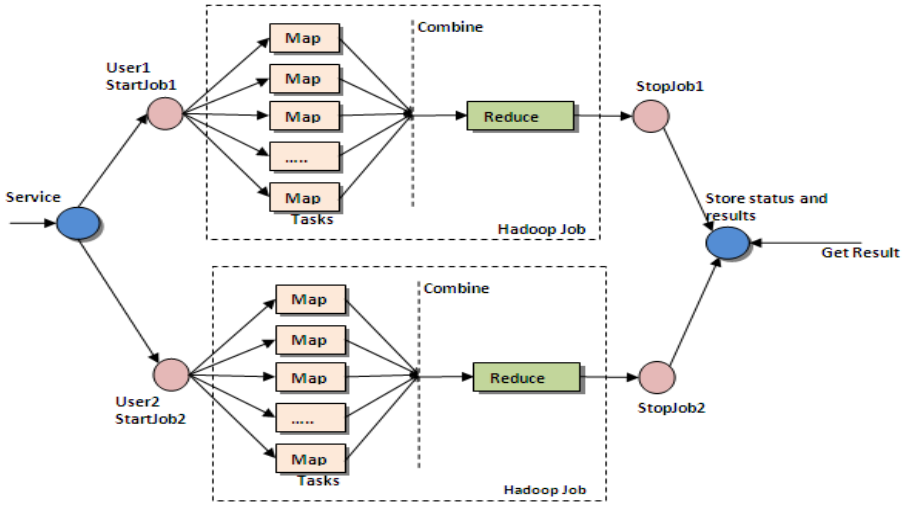


FIG. 1.1 – Schéma du fonctionnement de Hadoop MapReduce

FIG. 1.2 – Logo de Hadoop

2 Comment fonctionne l'algorithme PageRank de Google

L'algorithme PageRank, dont le nom rappelle son inventeur Larry Page, créateur et actuel PDG de Google, est à la base du moteur de recherche de Google. Il permet de classer les pages Web de manière pertinente en fonction de chaque requête. Son efficacité que tout le monde connaît a inspiré bien d'autres applications Big Data que nous détaillerons plus loin. C'est pourquoi nous expliquons ici le fonctionnement de PageRank en tant que moteur de recherche. Mais c'est surtout parce qu'il illustre parfaitement le fait que le Big Data est plus une révolution conceptuelle que technologique, une nouvelle manière d'utiliser les données.

2.1 Explorer le Web et stocker les informations

Google parcourt en permanence le Web grâce à des «logiciels araignées» («spiders» ou «crawlers» en anglais) qui se promènent de liens en liens pour enregistrer les pages Web. On compterait 30 000 milliards de pages individuelles [1]. Il faut également avoir en tête la constante augmentation du nombre de données : en une journée 9 000 nouveaux articles paraissent sur Wikipedia, 400 millions de tweets sont postés sur Twitter, 104 000 heures de vidéos sont ajoutées sur Youtube [9]... Le travail quotidien d'exploration et de stockage est donc indispensable. Google dispose ainsi d'une base de données gigantesque, et lorsque qu'une demande est faite sur le moteur de recherche c'est vers cette base de données que se tourne Google et non vers Internet directement.

FIG. 2.1 – Un data center de Google dans l'Iowa (image Google).

2.2 Trouver les pages correspondantes

Lorsqu'un utilisateur entre un mot ou un groupe de mots dans le moteur de recherche, Google parcourt l'ensemble de ses données pour trouver toutes les pages qui sont susceptibles d'être pertinentes et les proposer à l'utilisateur. Toute page contenant l'un des mots de la recherche

1. Chiffres avancés par Google Inside Search, «How Search Works» en Mars 2013.

ou même un synonyme est considérée comme potentiellement intéressante. Mais alors comment parcourir plus de 100 peta-bits (10^{17} bits) de données [2] en moins d'une seconde ? Pour son algorithme qui demeure secret, Google utilise du MapReduce, principe fondateur du Big Data (section 1 p.9) mais aussi des techniques mathématiques particulièrement efficaces. Nous allons maintenant détailler son fonctionnement.

2.3 Classer les pages de manière pertinente

Une fois le premier écrémage fait (retenir les pages contenant un mot de la requête ou un synonyme), il est primordial de classer les pages restantes de façon pertinente. Tapez «algorithme» sur Google France vous obtenez 5 240 000 résultats [3]. Si ces derniers n'étaient pas classés le moteur de recherche n'aurait aucun intérêt.

La question se pose alors de savoir comment classer les sites ?

— Classez les par ordre chronologique de création et après quelques heures un site n'apparaitra plus, noyé sous la masse des nouveaux sites.

— Classez les par nombre de liens auxquels le site renvoie ou bien par nombre de liens qui envoient sur ce site ne fonctionne pas non plus. En effet, il sera alors très facile d'obtenir un site bien référencé en fabriquant des pages vides ou totalement inintéressantes renvoyant au site. Ou inversement, mettre des liens inutiles du site en question vers d'autres sites inintéressants pour le référencer.

Ainsi la méthode choisie par Google est d'attribuer une note à chaque page Web. Cela est fait pour l'intégralité de sa base de données. La notation tient compte de nombreux paramètres comme la confiance dans le site, la confiance en son auteur, les liens extérieurs et intérieurs, le nombre de fois où apparaissent les mots de la recherche, leurs emplacements (dans le titre, dans l'URL, dans le texte)... La recette exacte de cette notation est gardée secrète et varie selon les moteurs de recherche. C'est d'ailleurs une source de conflit : Google a accusé, en février 2011, Bing (Microsoft) de copier leurs résultats, preuves à l'appui [4] selon Matt Cutts ingénieur anti-webspam chez Google.

2.4 Comment utiliser les notes attribuées ?

Contrairement à ce qui pourrait sembler logique *a priori*, Google n'utilise pas directement cette notation globale pour classer les pages. L'algorithme fait appel à la théorie des graphes, aux marches aléatoires et aux chaines de Markov.

2.4.1 Représentation par graphe

Le problème est mis sous la forme d'un graphe. Supposons que N pages ont été retenues parce qu'elles contenaient une partie des mots de la requête. Chaque page constitue un nœud. Les nœuds sont reliés entre eux par des flèches qui correspondent aux liens Internet pointant vers d'autres pages.
Prenons un exemple très simple avec $N = 3$. Appelons A, B et C nos trois pages et supposons

2. Toujours selon Google.
3. En juillet 2013.
4. Voir `http://googleblog.blogspot.com/2011/02/microsofts-bing-uses-google-search.html`.

que A possède un lien qui pointe vers B, B et C ayant chacune un lien l'une vers l'autre, alors le graphe obtenu sera du type : $A \longrightarrow B \longleftrightarrow C$

2.4.2 La marche aléatoire

L'idée[5] est à présent de simuler un utilisateur Internet qui se promènerait sur les pages. De clic en clic, il se déplace aléatoirement en suivant les liens qui lui sont proposés. C'est ici qu'intervient la notation des pages. L'utilisateur choisit aléatoirement le lien qu'il va suivre mais pas de façon équiprobable : il aura davantage tendance à suivre un lien qui mène à une page bien notée. On pondère ainsi la probabilité de suivre chaque lien pour respecter les notes des différentes pages accessibles. Pour éviter que l'utilisateur reste bloqué sur une page qui ne contient aucun lien, on offre également systématiquement une probabilité ϵ (faible) d'aller sur n'importe quelle autre page.

2.4.3 Un peu de mathématiques

Ce chapitre explique l'idée mathématique qui sous-tend l'algorithme de Google. Il est ici intéressant de voir à quel point des problèmes pratiques sont résolus par des mathématiques théoriques. C'est également un bon exemple de comment deux approches mathématiques différentes (l'analyse et la topologie d'une part et les probabilités et les chaines de Markov d'autre part) se complètent pour obtenir une solution. Les parties techniques (écrites dans cette police) peuvent être sautées en première lecture. **Pour gagner du temps, nous vous invitons à vous rendre directement page 17.**

La marche aléatoire décrite précédemment définit en fait une chaîne de Markov (voir section 2.4.5 p.15). On peut alors représenter notre système de N pages par un matrice $A = (a_{i,j})_{1 \leq i,j \leq N}$ où $a_{i,j}$ est la probabilité qu'en étant sur la page i on aille directement sur la page j. Cette matrice est appelée matrice de transition.

Ici notre matrice de transition est du type $A = (1-\epsilon)P + \epsilon Q$ où P est la matrice de transition liée à la marche aléatoire correspondant aux notes et aux liens réellement existants et $Q = \begin{pmatrix} 1/N & \dots & 1/N \\ \vdots & \dots & \vdots \\ 1/N & \dots & 1/N \end{pmatrix}$ représente le fait d'avoir une faible probabilité de retourner sur n'importe quelle page à tout moment. Google utiliserait $\epsilon = 0{,}15$, ce qui correspond à imaginer qu'un utilisateur suit 7 liens en moyenne avant de recommencer une nouvelle recherche ($100/7 \approx 0{,}15$).
Voici la matrice de transition obtenue pour l'exemple précédent du graphe très simple $A \longrightarrow B \longleftrightarrow C$ (on considère ici que les trois pages possèdent la même notation) :

$$A = 0,85 \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + 0,15 \begin{pmatrix} 0,33 & 0,33 & 0,33 \\ 0,33 & 0,33 & 0,33 \\ 0,33 & 0,33 & 0,33 \end{pmatrix} = \begin{pmatrix} 0,10 & 0,05 & 0,05 \\ 0,90 & 0,05 & 0,90 \\ 0,10 & 0,90 & 0,05 \end{pmatrix}$$

L'idée est de calculer les préférences de cet utilisateur imaginaire. Pour cela on cherche ce qu'on appelle une distribution stationnaire, on va donc chercher des points fixes.

5. La plupart des idées décrites ici sont inspirées soit de la "vignette" de Christiane Rousseau «Comment Google fonctionne : chaînes de Markov et valeurs propres» soit du papier de Michael Eisermann http ://www-fourier.ujf-grenoble.fr/ eiserm/Enseignement/google.pdf. Cependant les étapes de la preuve et les preuves en elles-mêmes sont personnelles.

2.4.4 Théorèmes du point fixe

On fait ici appel à des théorèmes d'existence, utilisant de la topologie classique.

Théorème du point fixe de Picard : Soit (E, d) un espace métrique complet, $f : E \to E$ une application contractante, ie $\exists C < 1$ tel que

$$\forall x, y \in E, d(f(x), f(y)) < Cd(x, y)$$

Alors f admet un unique point fixe et $\forall x_0 \in E$ la suite définie par $\forall n \in \mathbb{N}$, $x_{n+1} = f(x_n)$ converge vers ce point fixe.

preuve : On considère la suite $(x_n)_{n\in\mathbb{N}}$ avec x_0 quelconque et $x_{n+1} = f(x_n)$. Montrons que cette suite converge, il suffit de montrer qu'elle est une suite de Cauchy.
Soit $p, q \in \mathbb{N}$,

$$|x_{p+1} - x_p| < C|x_p - x_{p-1}| < \ldots < C^p|x_1 - x_0|$$

Ainsi

$$\begin{aligned} |x_{p+q+1} - x_{q+1}| &\leq |x_{p+q+1} - x_{p+q}| + \ldots + |x_{q+2} - x_{q+1}| \\ &< C^{p+q}|x_1 - x_0| + \cdots + C^q|x_1 - x_0| \\ &< C^q \sum_{k=0}^{p} C^k|x_1 - x_0| \end{aligned}$$

Or

$$\lim_{q\to\infty} C^q = 0 \text{ et } \sum_{k=0}^{\infty} C^k = \frac{1}{1-C}$$

La suite est donc bien de Cauchy. Par passage à la limite et par continuité de f on est sûr que cette limite est un point fixe.
Il reste à montrer l'unicité. Nous allons raisonner par l'absurde.
Soit x et y deux points fixes distincts, on a alors :

$$|f(x) - f(y)| = |x - y| < C|x - y|$$

Ceci est impossible. D'ou l'unicité.
Ceci achève la preuve.

Dans le cas qui nous intéresse ici, on a $E = \left\{ X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \in \mathbb{R}^N, x_i \geq 0, \sum_{i=0}^{N} x_i = 1 \right\}$ l'espace des vecteurs de probabilités de transitions, $f = X \in E \longrightarrow AX \in E$ et $d(X, Y) = \sum_{i=0}^{N} |x_i - y_i|$.
Même si on ne peut pas affirmer *a priori* que notre fonction est bien contractante (et donc utilisé ce théorème) on sait que : $d(f(X), f(Y)) \leq d(X, Y)$. Ce qui semble assez proche [6]. On

6. On prend ici un parti différent de celui de Michael Eisermann qui facilite les choses en considérant $f(X) = (1-\epsilon)PX + \epsilon \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, ce qui assure à f d'être contractante mais ne semble pas approprié à la vision «Markovienne» du problème.

imagine également que la méthode d'itération pourrait être efficace en pratique. On peut néanmoins prouver l'existence d'une distribution stationnaire grâce au théorème suivant.

Théorème :
Soit (E, d) un espace métrique complet, convexe et borné, $f : E \to E$ telle que

$$\forall x, y \in E, d(f(x), f(y)) \leq d(x, y)$$

Alors f admet un point fixe et $\forall x_0 \in E$ la suite définie par $\forall n \in \mathbb{N}$, $x_{n+1} = f(x_n)$ converge vers ce point fixe.

preuve : On définit ici $\forall n \in \mathbb{N}, \forall x, f_n(x) = (1 - \frac{1}{n})f(x) + \frac{x}{n}$
$g_n : x \longrightarrow (1 - \frac{1}{n})f(x)$ est $1 - \frac{1}{n}$ contractante donc admet un unique point fixe x_n.
Montrons à présent que (f_n) converge uniformément vers f :
Soit $x \in E$, $d(f_n(x), f(x)) \leq \frac{M}{n} + \frac{M}{n}$ où M est une borne de E.
D'où la convergence uniforme.
La suite (x_n) étant bornée elle admet une sous suite convergente, appelons x sa limite.
Donc en passant à la limite dans l'équation $f_n(x_n) = (1 - \frac{1}{n})f(x_n) + \frac{x_n}{n}$ on obtient $f(x) = x$
D'où l'existence du point fixe pour f .

Le choix de E et d nous assure le caractère convexe et borné. Nous avons donc bien une preuve d'existence mais pas d'unicité. Il est par exemple clair que l'identité convient et ne possède que des points fixes.

2.4.5 Les chaînes de Markov

Pour parvenir à une preuve d'existence et d'unicité exacte, il faut faire appel aux probabilités et aux chaînes de Markov.

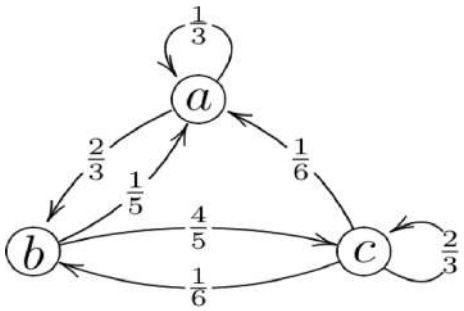


FIG. 2.2 – Graphe d'une chaîne de Markov

Théorème : Une chaîne de Markov dont la matrice de représentation est irréductible admet une distribution invariante si et seulement si tous ses états sont récurrents positifs. De plus cette distribution invariante est unique.

La présence du terme en Q nous assure, puisqu'on travaille ici avec une chaîne finie, à la fois le caractère irréductible de A et le fait que les états sont récurrents positifs.
Le théorème nous assure donc l'existence et l'unicité d'une distribution stationnaire. C'est cette distribution qui est utilisée pour classer les pages, celle possédant le plus grand temps de présence sera la première et ainsi de suite.

3 Méthodes pratiques de calcul

Non seulement la théorie nous assure de l'existence et de l'unicité d'une distribution stationnaire, mais elle nous offre aussi un moyen de la trouver. En effet, il suffit de promener l'internaute imaginaire suffisamment longtemps pour qu'il s'approche de la distribution stationnaire.
En pratique cela revient à calculer successivement les $(A^k X_0)_{0 \leq k \leq p_\epsilon}$ où X_0 est une distribution initiale quelconque et p_ϵ un entier bien choisi (assez grand pour fournir une convergence et assez petit pour permettre un temps de calcul raisonnable).

3.1 Un algorithme naïf

Voici un exemple d'une programmation très simple en MATLAB qui est un langage de programmation très utilisé pour les problèmes de calcul numérique.

```
function [X] = rank(A, precision_voulue)
N = size(A, 2);
X_prec = zeros(N,1);
X_prec(1)=1;
X= zeros(N,1);
X(2)=1;
precision_voulue=1/1000;
while(norm(X-X_prec, 2) > precision_voulue)
X_prec=X;
X = A * X;
end
```

On obtient ainsi une distribution proche de la distribution stationnaire (en norme euclidienne). Une implémentation aussi simple est bien sûr sans commune mesure avec les algorithmes optimisés utilisés par Google.

3.2 L'algorithme de Google

Google utilise évidemment des méthodes bien plus complexes pour la mise en place de son algorithme PageRank. L'algorithme exact n'étant pas public, nous faisons ici des suppositions. Cependant il est clair qu'une méthode itérative classique ne peut être envisagée

à cause des grandes tailles des matrices.
En effet, les algorithmes connus aujourd'hui sur la décomposition en valeurs singulières [1] (SVD) fonctionne en $O(n^3)$ (lire «grand o de n puissance 3) où n est le nombre de nœuds dans le réseau. C'est-à-dire que les valeurs pour n sont de l'ordre de 10 millions, il faudrait une puissance de calcul de 10^{21} flops (FLoating-point Operations Per Second) pour mener le calcul aussi rapidement, ce qui n'est pas faisable aujourd'hui. En effet, l'ordinateur le plus rapide du monde en juin 2013, le Tianhe-2 (MilkyWay-2), monte seulement à 54,9 peta-flops, soit 54,9 10^{15} flops. Il faut également prendre en compte que Google reçoit plus de 2 300 000 requêtes par minute.
Il faut faire appel à des méthodes dites «de Monte-Carlo» pour obtenir des algorithmes rapides et puissants. On obtient alors des algorithmes qui convergent «avec grandes probabilités» vers la distribution stationnaire. Il est possible de faire tourner PageRank en $O(\sqrt{log(n)}/\epsilon)$ étapes, où ϵ est la probabilité (définie précédemment) d'entamer une nouvelle recherche. L'idée, décrite et démontrée dans [28], est la suivante.
On fait simultanément n marches aléatoires (1 par nœud). On part de chaque nœud du graphe et on fait K pas d'une marche aléatoire : avec une probabilité ϵ on termine la marche aléatoire directement là où on est, avec une probabilité $1-\epsilon$ on se déplace aléatoirement vers un autre nœud en suivant les liens existants. Pour voir que cette marche s'arrête on peut la voir comme un problème de population. Au départ il y a n individus (un sur chaque nœud). A chaque nouvelle génération, un individu a 1 descendant avec probabilité $1-\epsilon$ (qui vit dans un autre nœud aléatoirement choisi) ou 0 descendant avec probabilité ϵ. Ici le taux de natalité est strictement inférieur à 1 donc la population est sûre de disparaître (avec probabilité 1).
Chaque nœud ν compte le nombre de visiteurs ζ_ν qu'il a accueilli pendant ces K tours et calcul son PageRank $\widetilde{\pi}_\nu = \frac{\zeta_\nu \epsilon}{nK}$. La durée de vie moyenne d'un individu est $1/\epsilon$ et $\frac{nK}{\epsilon}$ représente le nombre moyen de visites sur l'ensemble des nœuds durant K tours. $\widetilde{\pi}_\nu$ est ainsi une estimation de la distribution stationnaire.

4 Le Big Data ne s'arrête pas là pour Google.

Google ne se contente pas d'enregistrer toutes les pages Web existantes et de leur attribuer une note pour permettre au moteur de recherche de fonctionner. Chaque requête faite par un utilisateur est enregistrée : les mots de la recherche, le résultat sur lequel l'utilisateur a cliqué, la date, l'heure, la localisation de l'ordinateur... Toutes ces données, qui étaient d'abord stockées sans dessein particulier, sont aujourd'hui une mine d'or pour Google qui dispose de toutes les archives de leur moteur de recherche depuis plus de 10 ans. Google possède très certainement l'une des plus grandes bases de données au monde. Voici un exemple impressionnant et très représentatif du potentiel «Big Data» qu'elle représente.

1. En utilisant la méthode des «Householder reflections» de Golub & Christian Reinsch qui fonctionnent en $4mn^2 - 4n^3/3$, avec mxn la taille de la matrice à diagonaliser, cf Wikipedia «Singular Value Decomposition» et [1].

4.1 Google Flu Trend

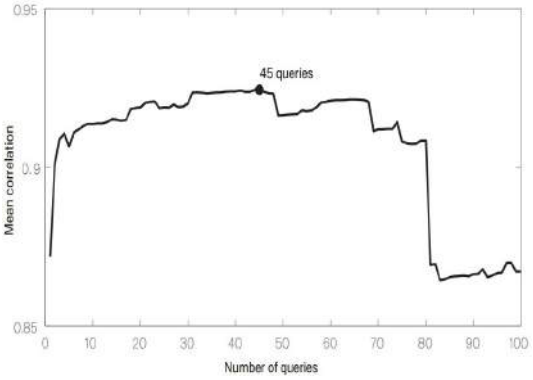
En 2009, peu avant l'apparition du virus H1N1 dont la propagation préoccupa la terre entière, les chercheurs de Google eurent l'idée d'essayer de suivre les propagations d'épidémies [10] grâce à leur moteur de recherche. Ils se sont d'abord intéressés à la grippe, qui touche chaque année plusieurs millions de personnes et qui cause la mort de plus de 250 000 personnes [1]. Pour cela ils ont simplement décidé de regarder de plus près les requêtes effectuées sur le moteur de recherche. Il paraît en effet évident que les personnes malades ont plus tendance à se préoccuper des symptômes grippaux sur la toile que les gens en bonne santé. On estime à 90 millions le nombre d'Américains se renseignant sur Internet pour des raisons médicales chaque année. Cependant, l'état des connaissances sociologiques sur le comportement des internautes ne permettait pas de fabriquer un modèle mathématique *ex nihilo*. C'est là qu'intervient la vision Big Data.
Même si les chercheurs ne savaient pas quelles requêtes, ni combien d'entre elles seraient corrélées à l'épidémie de grippe, ils étaient persuadés qu'un lien existait. L'ordinateur serait capable de le trouver. Ils sont donc partis d'un modèle mathématique très simple :

$$log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 log\left(\frac{Q}{1-Q}\right) + \epsilon$$

où P est la proportion de visites médicales liées à la grippe, Q la proportion de requêtes liées à la grippe sur Google, β_0, β_1 des termes multiplicatifs et ϵ un terme d'erreur.
Pour obtenir les valeurs de P semaine par semaine, Google s'est appuyé sur les archives des CDC (Centers for Disease Control and Prevention) qui sont les organisme chargés du suivi des épidémies aux Etats-Unis. Il fallait également définir proprement Q. Quels mots-clés sont liés à la grippe ? Combien en prendre ? Google a simplement fait tourner son modèle sur les 50 millions de mots-clés les plus fréquemment utilisés aux Etats-Unis durant la période d'enregistrement disponible dans leur base de données soit de 2003 à 2008. Ce qui correspond à plusieurs centaines de milliards de requêtes. Ils ont ainsi obtenu les 100 mots-clés qui rendaient le modèle Google le plus proche de celui des archives du CDC. Puis ils ont essayé de voir combien de ces 100 requêtes il fallait prendre en compte pour obtenir les meilleures prévisions possibles. Parmi ces 100 requêtes certaines n'étaient d'ailleurs pas nécessairement en rapport avec la grippe. La requête «high school basketball» y figurait par exemple, puisqu'il s'avère que la saison de basket universitaire coïncide aux Etats-Unis avec celle de la grippe... Ils ont finalement retenus 45 de ces requêtes.

Au final, Google a testé plus de 450 millions de modèles différents. Ils ont effectué des régressions linéaires et transformations de Fisher pour observer la correspondance du modèle avec les résultats du CDC. Ils ont pour cela utilisé leur puissance de calcul parallèle mise à disposition par leurs data centers. Le modèle Google s'avère très efficace et dispose d'un gros avantage sur les modèles du CDC. En effet, le modèle Google est actualisé jour par jour et potentiellement pourrait l'être heure par heure, tandis que les modèles des CDC basés (entre autres) sur le nombre de visites chez le médecin ont systématiquement 1 à 2 semaines de retard sur les prévisions Google. L'élève a dépassé le maître. Les résultats sont disponibles gratuitement sur Google Flu Trend.

1. Chiffres avancés dans [10].



Search Query Topic	Top 45 Queries N	Top 45 Queries Weighted	Next 55 Queries N	Next 55 Queries Weighted
Influenza Complication	11	18.15	5	3.40
Cold/Flu Remedy	8	5.05	6	5.03
General Influenza Symptoms	5	2.60	1	0.07
Term for Influenza	4	3.74	6	0.30
Specific Influenza Symptom	4	2.54	6	3.74
Symptoms of an Influenza Complication	4	2.21	2	0.92
Antibiotic Medication	3	6.23	3	3.17
General Influenza Remedies	2	0.18	1	0.32
Symptoms of a Related Disease	2	1.66	2	0.77
Antiviral Medication	1	0.39	1	0.74
Related Disease	1	6.66	3	3.77
Unrelated to Influenza	0	0.00	19	28.37
	45	**49.40**	**55**	**50.60**

FIG. 4.1 – Graphe de la qualité du modèle en fonction du nombre de mots-clés retenus (A gauche). Nature des 100 mots-clés retenus (A droite, source des images [10]).

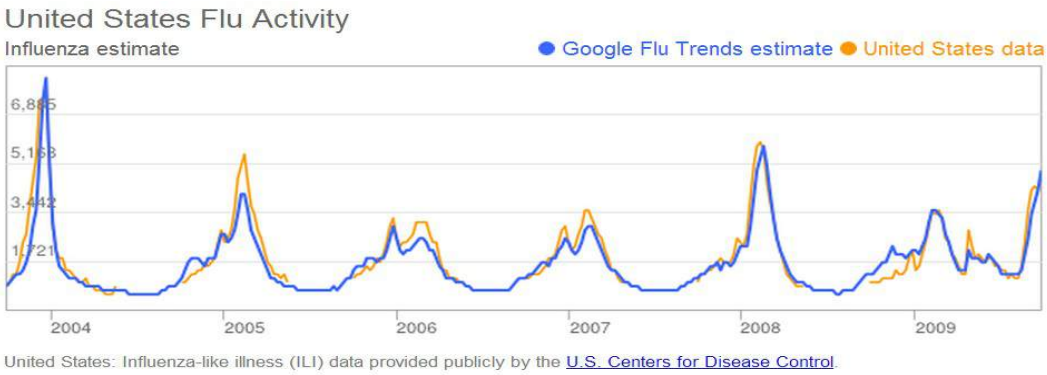


FIG. 4.2 – Résultats comparés du modèle Google et de celui des CDC (source : Nature).

Deuxième partie

Domaines d'application

Le déluge de nouvelles données, dont 85% d'entre elles sont non structurées (Photos, vidéos, textes de messages Facebook, dans Tweeter, de SMS etc...), est une mine d'informations potentiellement utiles dans tous les domaines. Que ce soit pour la recherche scientifique, l'industrie, le marketing ou simplement l'optimisation de services en tout genre. L'exploitation des bases de données pour en extraire des informations utiles (ce qu'on appelle le «data mining») peut s'avérer extrêmement prolifique. Nous allons ici donner quelques exemples représentatifs d'utilisation du Big Data.

5 Le quatrième paradigme de la science

Bien que les techniques de traitement de l'information utilisées pour le Big Data ne soient pas fondamentalement révolutionnaires (le calcul parallèle existe depuis les années 50 et les outils mathématiques telles que les chaînes de Markov datent du début du XXème siècle), la nouvelle approche scientifique rendue possible par tant de données est à l'origine de ce qu'on appelle le quatrième paradigme de la science suite à la parution du livre à succès de Tony Hey [11].

— Le premier paradigme est celui de l'empirisme, né il y a quelques millénaires, la science étant à l'origine essentiellement basée sur l'observation des phénomènes naturels.
— Le second paradigme, apparu plus tard, est lié au développement de ce qu'on appelle les sciences «dures» où la théorisation permet l'explication des phénomènes observables.
— Depuis quelques décennies, l'avènement des ordinateurs a amené le troisième paradigme qui consiste à simuler des expériences sur ordinateur pour valider ou réfuter des théories.
— L'ère du Big Data est celle du quatrième paradigme. Nous disposons actuellement de tellement de données disponibles qu'il est possible de laisser l'ordinateur faire des découvertes par lui-même [16], simplement en trouvant des liens statistiques au sein de milliards de données. Le chercheur n'a aujourd'hui plus nécessairement besoin de théoriser pour faire une découverte. Nous allons ici donner quelques exemples.

5.1 La recherche biomédicale

5.1.1 Trouver automatiquement des liens insoupçonnés : le projet brainSCANr

Face à la montagne d'articles existants dans la bibliothèque PubMed [1] (plus de 22 millions de publications accessibles actuellement et 40 000 à 50 000 de plus chaque mois), Jessica et Bradley Voytek, respectivement informaticienne et neurologue à l'université de San Francisco (UCSF), ont décidé de faire de la recherche autrement. En 2010, ils lancent le projet brainSCANr [2] (Brain Systems, Connections, Associations, and Network Relationships) [34] qui consiste à faire tourner un logiciel sur 3,5 millions de résumés d'articles pour trouver automatiquement des relations entre différentes maladies et différentes parties du cerveau.
L'idée est de choisir quelques mots-clés (pour leur premier essai 124 régions du cerveau, 291 fonctions cognitives et 47 maladies) et de fabriquer un graphe qui associe les mots entre eux. On distingue alors deux types d'associations entre deux termes. Les associations fortes lient un terme à un autre lorsqu'ils apparaissent simultanément dans une même publication [3]. Les associations faibles, et ce sont les plus intéressantes car ce sont celles que seul l'ordinateur permet de trouver, lient deux termes lorsqu'ils sont tous les deux plus de 1000 fois fortement liés à un mot commun.
Jessica et Bradley Voytek ont ainsi trouvé une conneion forte entre «sérotonine» et «migraine»

1. PubMed est une bibliothèque publique d'articles médicaux en ligne qui dépend du NCBI (National Center for Biological Information).
2. Voir `www.brainscanr.com`.
3. Le parti pris est que deux termes associés à une même publication sont bien «interconnectés» et non pas «antagonistes».

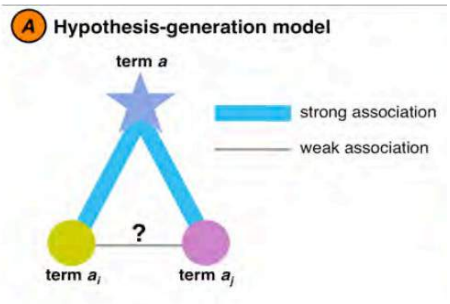


FIG. 5.1 – Exemple de connections fortes et faibles (source de l'image : [34]).

(2943 articles en commun) ainsi qu'entre «sérotonine» et «striatium» (4782 articles) alors que «striatium» et «migraine» n'apparaissent que dans 16 articles en même temps. L'idée est ainsi de mieux orienter les recherches futures pour voir si ce lien entre migraine et striatium n'aurait pas échappé aux chercheurs en neurosciences jusque là ...

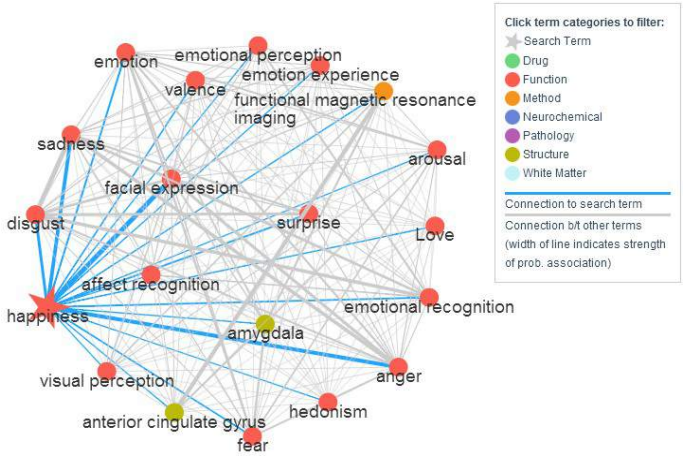


FIG. 5.2 – Graphe représentant les différentes connections trouvées (source de l'image : [34]).

5.1.2 Applications dans la génomique

Ce chapitre a pu être écrit grâce aux savants éclairages de Jonathan Epstein, bio-informaticien au NIH (National Institute of Health) et anciennement au NCBI (National Center for Biotechnology Information), que nous tenons ici à remercier chaleureusement.

Il a fallu 11 ans (1990-2001) et 3 milliards de dollars pour réaliser le premier séquençage ADN humain complet (ou presque [4]), soit 3 milliards de nucléotides. Aujourd'hui il est possible de séquencer 100 génomes (soit 600 Go de données) en dix jours et ce pour moins de 1000 euros par génome. Certaines entreprises comme Ion Torrent [5] promettent même prochainement

4. Voir Projet génome humain, Wikipedia `http://fr.wikipedia.org/wiki/Projet_g%C3%A9nome_humain`.
5. Voir `www.iontorrent.com`.

un séquençage complet en 15 minutes et pour 900 euros. Dans ce contexte, les chercheurs et biologistes disposent d'une base de données toujours plus massive. Par exemple, la base de données GenBank possède, en libre accès, une bibliothèque de plus de 150 milliards de bases de nucléotides contenues dans 162 millions de séquences. Il est donc indispensable de trouver des méthodes pour en extraire des connaissances [2].

Le logiciel BLAST

Pour tirer profit de ces immenses bases de données, les bio-informaticiens utilisent le logiciel BLAST (Basic Local Alignement Search Tool) [18] qui s'apparente à un moteur de recherche pour biologistes. Ce logiciel est géré par le NCBI dans son «cluster» personnel. Après avoir isolé une séquence de nucléotides ou d'acides aminés qui lui paraît intéressante, le chercheur entre cette séquence dans la recherche BLAST. Le logiciel compare cette séquence à la base de donnée choisie et renvoie les séquences les plus proches. Ce qui aide à percer les mystères de la génétique.

Le fonctionnement de BLAST

BLAST n'est pas un algorithme d'alignements parfaits, le temps de calcul nécessaire pour rechercher des alignements exacts [6] étant trop long en pratique [3] si l'on essaie de repérer les éventuelles mutations. L'idée est donc de rechercher des alignements locaux. L'utilisateur met en entrée une séquence à comparer avec l'intégralité d'une base de données choisie (typiquement GenBank). Le logiciel divise la séquence en petits mots (usuellement en mot de 3 lettres pour une séquence de protéines et de 11 pour une séquence ADN) et calcule un score d'alignement à l'aide d'une matrice dite de similarité (cf.figure ci-après).

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

A L Q F V C G D
A I Q Y V C P D

4+2+5+3+4+9+(-2)+6
=31

FIG. 5.3 – Exemple de calcul de score d'alignement entre deux séquences (ici de 7 lettres) à l'aide de la matrice BLOSUM62

Chaque matrice de score est basée sur des prévisions statistiques différentes. Certaines attribuent les scores selon des probabilités liées aux mutations les plus courantes, d'autres se basent seulement sur la proportion globale connue de chaque paire. La théorie n'est pas encore suffisamment avancée pour trancher objectivement, l'approche est plutôt empirique. La plus utilisée est la matrice BLOSSUM62 qui est la 62ième itération de la matrice BLOSSUM, chaque itération représentant un degré d'évolution.

6. On peut citer en exemple de tels programmes l'algorithme de Smith-Waterman qui s'effectue en $O(MN)$ où M et N sont les tailles des deux séquences à comparer.

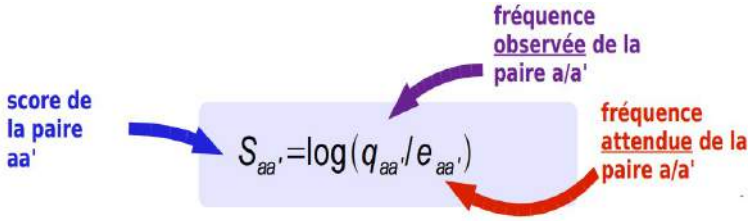


FIG. 5.4 – Calcul du score d'un couple aa'

Pour chaque mot issu de la requête initiale, on calcule le score de similarité avec tous les mots possibles de la même taille (pour un mot de 3 lettres on a 20x20x20=8000 possibilités car il 20 acides aminés chez l'Homme) et on conserve seulement ceux dont le score est au-dessus d'un seuil T choisi à l'avance. Le logiciel cherche ensuite un alignement parfait des mots restants (au dessus du seuil) avec la base de données. S'il en trouve un, il cherche des High-scoring Segment pairs (HSP) qui sont donc des séquences de fort alignement. Pour cela il étend la recherche d'alignement (vers la gauche et vers la droite) en partant du point d'exact alignement aussi longtemps que le score d'alignement augmente.

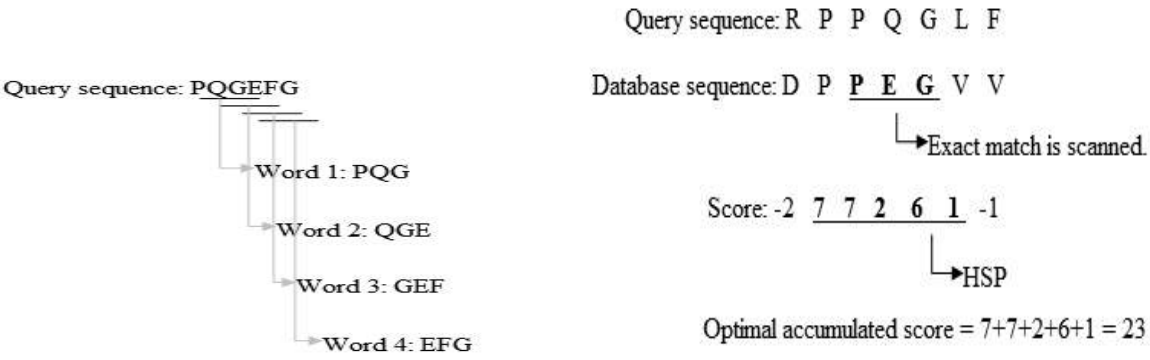


FIG. 5.5 – Découpage d'une séquence en petits mots (à gauche), processus d'extension de l'alignement (à droite)

Toute la difficulté est de savoir quels processus (aléatoires ?) régissent les mutations ou substitutions qui peuvent apparaître dans une séquence. La méthode utilisée de nos jours fait appel au modèle de Markov caché ou HMM (Hidden Markov Model [23]). Ce concept mathématique permet lorsqu'on observe une suite d'évènements, sans savoir par quel mécanisme ils ont été générés, de reconstruire, à partir des observations, le mécanisme le plus probable ayant généré la suite d'évènements. On peut ainsi essayer de trouver dans quelles proportions interviennent mutations, délétions, insertions, erreurs de transcription (c'est à dire tous les états cachés) etc... C'est une extension de BLAST nommée PSI-BLAST qui utilise cette méthode. On a également recours à des méthodes de décomposition en valeurs singulières [1].
Une chaîne de Markov cachée est une chaîne de Markov dont certains états, ainsi que leurs transitions associées, sont inconnus. Typiquement en génomique les observations, sont les séquences ADN. Les états cachés représentent les différents types de mutations génétiques. A partir des observations on reconstruit la chaîne de Markov la plus probable ayant engendré toutes ces observations.

La puissance du parallélisme (utilisation d'ordinateurs en parallèle) est également en train d'être mise à disposition des biologistes. Des chercheurs d'Oxford [13] ont mis en place des algorithmes de calculs de modèles de Markov optimaux. Pour une séquence de taille T et un nombre d'états N avec T très grand devant N et en mettant en parallèle $KT/2N^2$ machines l'algorithme a une complexité en $O(Nlog(T))$. K (qui est une constante) et N (le nombre d'observations et donc le nombre de lettres) sont de l'ordre de la centaine, alors que T est de l'ordre du million. Ainsi les supercalculateurs permettent des traitements ultra-rapides.

Il existe également une version «MapReduce» de BLAST nommée BlastReduce [29] par son inventeur Michael Schatz. En connectant seulement 24 machines en réseau, grâce notamment à Hadoop MapReduce, on atteint des performances jusqu'à 250 fois plus rapide qu'avec le BLAST traditionnel. En utilisant un "cluster" plus gros, le gain de temps est prodigieux (on ne compte plus en jours mais en minutes). La progression superlinéaire est due également à une augmentation de la mémoire disponible et pas seulement à l'augmentation de la puissance de calcul.
L'utilisation d'arbre des suffixes [33], qui consiste à stocker les données sous une forme arborescente particulière (c'est un arbre dont les branches présentent tous les suffixes que l'on peut créer à partir d'un mot), a également permis des recherches de séquences bien plus rapides dans des grosses bases de données. En effet, une fois la base de données mise sous forme d'arbre des suffixes (cette opération [7] s'effectue en $O(nlog(n))$ opérations où n est le nombre de lettres dans la base de données), la recherche d'une séquence S de m lettres s'effectue en $O(m)$ et ce indépendamment de la taille n de la base de données ! Ceci représente un gain de temps extraordinaire par rapport aux méthodes antérieures.

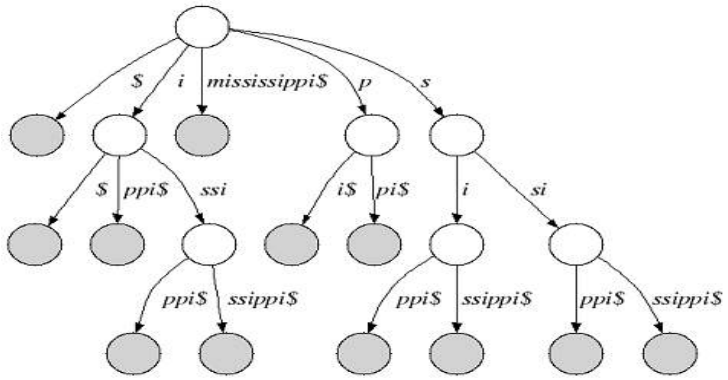


FIG. 5.6 – Arbre des suffixes du mot mississippi

Il existe également d'autres logiciels que BLAST qui traitent de sujets similaires. On peut par exemple citer le logiciel GRAIL (Gene Relationships Among Implicated Loci) [24] qui s'appuie également sur la base de données PubMed.

7. cf l'algorithme d'Ukkonen [33].

5.2 La recherche en astrophysique

5.2.1 Le programme LSST

Nous tenons ici à remercier Kirk Borne, professeur d'astrophysique et d'informatique à George Mason University à Fairfax en Virginie, qui nous a accordé un précieux entretien pour expliquer les interrogations des chercheurs à ce sujet.

Le projet LSST (Large Synoptic Survey Telescope) est l'un des grands défis du Big Data. Ce centre d'observation, dont la mise en marche est prévue à l'horizon 2020, scrutera le ciel avec un appareil photo numérique de 3 200 Méga-pixels (ce qui en fait le plus puissant appareil photo numérique au monde). Chaque nuit, 30 Tera-bits de données seront créés et devront être analysés. L'enjeu ici est que l'ordinateur détecte tout seul des points de la galaxie qui pourraient intéresser les chercheurs. Chaque partie du ciel photographiée est ainsi mise en équation sous forme de vecteur. On y indique par exemple la taille apparente de l'astre, sa luminosité, la longueur d'onde émise, la vitesse de déplacement, la direction etc... En cas de fluctuation soudaine l'ordinateur doit émettre une alerte pour pouvoir permettre une observation approfondie. Une gestion en temps réel d'une aussi volumineuse quantité de données serait aujourd'hui impossible, les chercheurs sont donc lancés dans une course contre la montre. L'énorme base de données qui sera créée sera également un terrain d'exploitation pour améliorer les performances dans le domaine du Big Data.
Au delà des améliorations technologiques que l'on peut légitimement espérer d'ici 2020, c'est en faisant preuve d'ingéniosité que les chercheurs diminuent le nombre des données. La principale technique consiste à réduire la dimension du problème en trouvant des corrélations mathématiques entre les différents paramètres physiques. Un des exemples proposé par Kirk Borne est celui du plan fondamental des galaxies elliptiques. On sait relier la brillance de surface moyenne $< I >$, le rayon effectif R_{eff} et la dispersion des vitesses centrales σ_0 d'une galaxie elliptique par l'équation :

$$log(R_{\text{eff}}) = 0,34\frac{< I >}{\mu_B} + 1,4\sigma_0$$

Ainsi il n'est pas nécessaire de conserver ni de calculer ces trois grandeurs mais seulement deux d'entre elles. Sur un échantillon de 100 milliards de points, ce gain de temps et d'espace n'est pas du tout négligeable ! Les chercheurs utilisent également les données pour étayer la théorie et trouver de nouvelles formules qu'ils testent notamment avec la méthode des moindres carrés, qui permet de trouver la fonction qui correspond le mieux aux observations au sens des moindres carrés (on minimise la somme des carrés des écarts observés).
Une des pistes explorées par Kirk Borne est justement de transposer les méthodes de la recherche en génétique à l'astrophysique. Chaque point de l'espace, représenté par un vecteur contenant tout une série d'informations (luminosité, couleur, vitesse apparente de déplacements etc...) peut en effet être vu comme une séquence codante, semblable à un gène. Ainsi, si l'on parvient à corréler des caractères phénotypiques à des gènes pourquoi ne pas pouvoir corréler des phénomènes physiques aux astres ?

5.2.2 Le programme SKA

Le projet SKA (Square Kilometer Array) est encore plus ambitieux que celui du LSST. Il prévoit la construction du plus grand radiotélescope au monde, qui s'étendra sur près d'un kilomètre carré et comptera plusieurs milliers d'antennes. Le radiotélescope générera 100 peta-octets

de données chaque jour. Pour pouvoir traiter toutes ces données, le plus grand supercalculateur jamais construit devrait voir le jour, sa puissance de calcul devrait atteindre 100 peta-flops. Les défis à relever sont les mêmes que pour le programme LSST.

FIG. 5.7 – Vue d'artiste du projet SKA

6 Des applications innovantes

C'est ici un aspect différent du Big Data que nous abordons, et c'est probablement celui qui transformera le plus notre quotidien dans les futures années. Le Big Data permet une recherche d'information qui nous était jusqu'ici inaccessible. Transformer des données brutes en informations pratiques est aujourd'hui devenu un vrai business : c'est ce qui explique en partie cette ferveur autour du Big Data. Voici quelques exemples parmi bien d'autres, le champ des possibles étant quasiment infini.

6.1 Prédire l'avenir grâce aux données GPS : Far Out

Que serez-vous en train de faire dans 158 jours à 14h30 ?
La réponse à cette question, à laquelle peu de gens se risqueraient à répondre avec certitude, pourrait bien nous parvenir... d'un ordinateur. En effet, deux chercheurs de Microsoft travaillent sérieusement sur une manière de répondre à ce genre de questions [27]. Leur approche est simple : ils ont suivi 307 personnes et 396 véhicules (bus, taxis, voitures personnelles...) grâce à des balises GPS pendant de longues périodes (jusqu'à 1247 jours). En enregistrant les données heure par heure, jours après jours et après un travail mathématique que nous décrirons plus bas, les deux chercheurs sont capables de prédire avec de bonnes chances de réussite à quel endroit vous serez effectivement dans 158 jours à 14h30. En espérant améliorer encore le modèle, nos deux chercheurs en viennent même à imaginer des annonces commerciales du type : «Vous avez besoin d'aller chez le coiffeur ? Ca tombe bien ! Dans 4 jours vous serez à 100m d'un coiffeur qui vous proposera 10% de réduction !». Une autre idée est de proposer une application qui, pour un groupe d'amis choisis, définirait le meilleur endroit et le meilleur moment pour se retrouver autour d'un verre. Il est également possible d'imaginer une meilleure gestion du trafic routier au sein d'une agglomération par exemple.

Comment cela fonctionne ?

L'utilisation des chaînes de Markov (qu'elles soient classiques ou cachées) ne permet pas selon les auteurs de prédire correctement dans un futur lointain. En effet, le critère de dépendance uniquement entre l'instant t et $t+1$ des chaînes de Markov ne rend pas bien compte d'une évolution à long terme. La méthode mathématique utilisée [27] fait appel à la décomposition en série de Fourier, à l'analyse en composantes principales (ACP) et à la décomposition en valeurs singulières. L'analyse en composantes principales étant une méthode mathématique de réduction de la dimension. C'est une méthode pour trouver des axes privilégiés qui contiennent les informations les plus pertinentes et représentatives en leur sein. En regardant seulement l'information contenue sur ces axes principaux, on gagne énormément de temps tout en perdant un minimum d'informations.

Nous exposons ici seulement la méthode discrète d'un point de vue spatial, la méthode continue étant très similaire [27]. La surface du globe est d'abord découpée en cellules triangulaires équilatérales de 400m de côté. Les données GPS sont ensuite mises sous formes de vecteurs $(X_k)_{k\in\mathbb{N}}$, X_k codant le jour k. Seules les 10 cellules les plus fréquentées sont retenues pour le modèle ainsi qu'une onzième qui correspond à une cellule «ailleurs» regroupant toutes les autres cellules. Il s'avère qu'empiriquement 10 cellules suffisent à balayer la grande majorité de la vie courante d'un sédentaire : maison, lieu de travail, restaurant préféré etc... Le modèle fonctionne heure par heure et prend en compte le jour de la semaine ainsi que le fait d'être un jour férié ou non. L'ajout de plus de données pertinentes (comme par exemple la saison, la météo ou pourquoi pas les anniversaires ou même les jours de paie) sont facilement envisageables et ne présentent aucun défi théorique, il suffit d'être capable de récolter plus d'informations.

Ainsi $X_k = (c_{1,1},\dots,c_{11,1},c_{1,2},\dots,c_{11,2},\dots,c_{1,24},\dots,c_{11,24},\dots,j_1,\dots,j_7,f) \in \mathbb{R}^{272}$ est un vecteur de taille 272 et $c_{i,j}$ est la proportion de temps passé dans la cellule i pendant la j-ième heure du jour d'observation k, les $(j_q)_{q\in 1\dots 7}$ et f étant des booléens indiquant le jour de la semaine et si ce jour est férié ou non.
On obtient ainsi une matrice $D \in \mathcal{M}_{272,d}$ à 272 lignes et d colonnes où d est le nombre de jours durant lesquels une personne a été localisée par GPS. Chaque colonne de cette matrice représentant les différents déplacements lors d'une journée.
D'un point de vue statistique on détient les observations :
$Pr(C = c|T = t, J = j)$ qui est la probabilité de se trouver dans la cellule c à un instant t au jour particulier j. C,T et J étant des variables aléatoires représentant respectivement la cellule dans laquelle on se trouve, l'heure et le jour (par exemple un jeudi férié ensoleillé).
On veut à présent calculer les composantes principales de notre matrice pour obtenir ce qu'on appellera des «jours propres» (eigendays en anglais). Ceci correspond en fait à calculer les vecteurs propres de la matrice de variance-covariance $V = \frac{1}{d}DD^T \in \mathcal{M}_{272}$. D'un point de vue pratique, il est plus rapide de calculer la décomposition en valeur singulière directement à partir de la matrice rectangulaire D plutôt que de diagonaliser V.
De tous ces jours propres seuls les 10 premiers (c'est à dire les 10 plus grands en terme de valeur propre associée) sont considérés comme utiles. On réduit ainsi la totalité du temps d'étude (d jours) à 10 jours propres qui sont représentatifs du train de vie typique de notre sujet d'étude.

Le but est de prédire ce que fera notre cobaye au jour j au moment t. Pour cela on ne s'intéresse dans un premier temps qu'aux dernières composantes des vecteurs qui codent le type de jours (un lundi pluvieux et travaillé pour être moins optimiste que dans le dernier exemple).

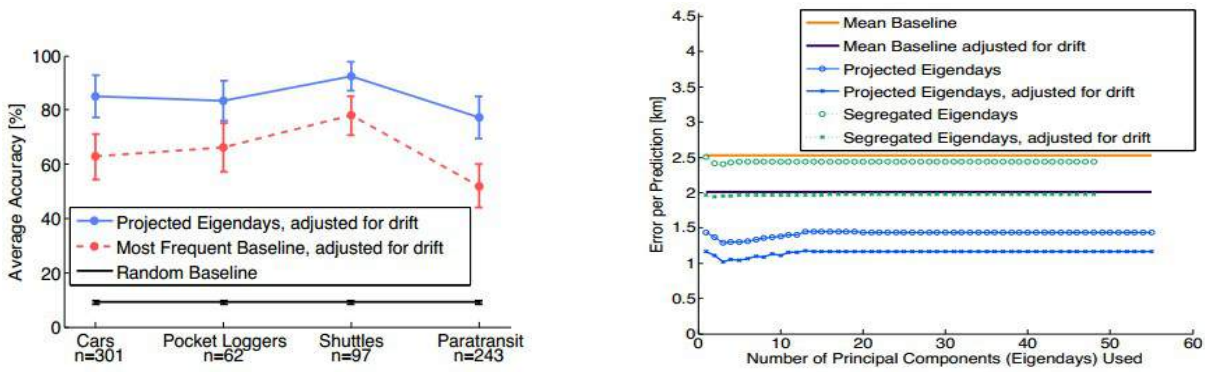


FIG. 6.1 – Résultats des prédictions pour différentes méthodes de prédiction, en bleu la méthode décrite précédemment.

On projette donc le vecteur j codant les paramètres jours, sur le sous espace engendré par les 10 jours propres mais uniquement par rapport aux dernières composantes (toujours celles qui codent le type de jour). On obtient ainsi un ensemble de poids $(\omega_i)_{i\in\{1,\dots,10\}}$ correspondant à chaque jour propre : $j = \sum_{i=1}^{10} \omega_i \mathcal{P}_i$, $\mathcal{P}_i$ étant la fin du jour propre i codant le type de jour. On pondère ensuite les poids pour que leur somme fasse 1 puis on calcule les probabilités d'être à l'instant t_q dans chacune des cellules. Les taux de réussite des prédictions varient selon les modèles mais sont dans l'ensemble prometteurs.

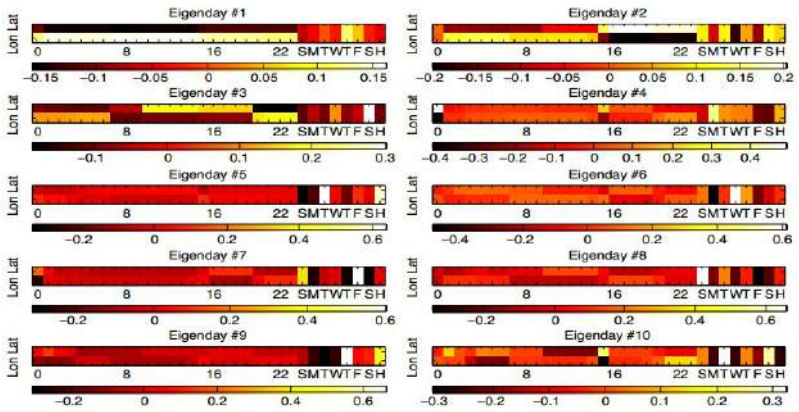


FIG. 6.2 – Représentation des 10 jours propres d'un individu

Ce type de prédiction n'intéresse pas uniquement les publicitaires mais aussi les enquêteurs de police. Il est en effet possible de localiser un téléphone portable et ainsi de prévoir des interpellations ou faciliter des surveillances. Aux Etats-Unis 87% des adultes peuvent être géolocalisés grâce à leur téléphone portables [32]. Ce genre de pratiques suscite évidemment de nombreuses craintes quant à la protection de la vie privée des citoyens. Nous reviendrons sur ce point plus tard.

6.2 Améliorer la sécurité aérienne

La reconnaissance de motifs similaires au sein de plusieurs séquences s'avère utile en génomique comme nous l'avons vu précédemment (section 5.1.2 p.22). Mais le problème inverse, c'est-à-dire trouver les motifs qui diffèrent au sein d'un ensemble de séquences, est également intéressant.
L'idée de Tom Oates [22], chercheur au département de sciences informatiques de l'Université de Massachusetts, est de comparer les enregistrements des boîtes noires d'avions ayant eu un accident avec celles de vols n'ayant pas rencontré de problèmes. En effet, la plupart des informations entre deux vols seront similaires : décollage, montée en altitude de croisière, manœuvres classiques ... Ce qui est intéressant pour comprendre un crash aérien est ce qui ne ressemble pas aux autres vols. Analyser un très grand nombre d'enregistrement pourrait permettre de progresser dans la compréhension et la détection de pannes et incidents majeurs.
La méthode utilisée pour détecter les motifs «distinctifs» s'appuie sur des méthodes statistiques classiques comme le partitionnement des données par la méthode des k-moyennes (K-means clustering) ou des tests de Student sur des modèles gaussiens. Les données se présentent sous forme de séries temporelles auxquelles il faut parfois appliquer des déformations temporelles dynamiques pour pouvoir travailler avec différents types de séries (en cas par exemple d'échelles de temps différentes). Chaque élément de la série étant un vecteur codant par exemple l'altitude de l'avion, sa vitesse, la température extérieure, la météo, la jauge de carburant, l'état des différents réservoirs en tout genre, l'état de chacun des moteurs... Le partitionnement par la méthode des k-moyennes consiste à diviser l'ensemble des enregistrements en k groupes distincts. L'un pouvant être les vols qui se sont bien passés, un autre celui des crashs ou encore des vols ayant présenté un problème moteur etc... C'est ensuite une simple classification barycentrique.
On espère ainsi être capable de prévenir des accidents en reconnaissant en temps réel une série déviante, annonciatrice d'un problème plus ou moins imminent.

6.3 Prévenir des complications chez les bébés prématurés

Fig. 6.3 – Publicité IBM pour la détection des infections chez les prématurés

Une équipe de chercheurs d'IBM travaille à appliquer les méthodes de modélisations propres au Big Data pour trouver des schémas permettant la détection plus rapide d'infections chez les prématurés [21]. En couveuse un bébé prématuré est constamment connecté à de nombreuses machines qui collectent sa température, son rythme cardiaque, son rythme respiratoire, son taux de globule rouge etc... Les électrocardiogrammes peuvent enregistrer jusqu'à 1000 valeurs

par seconde, soit 86,4 millions par jour. Chaque patient crée plus de 39 Go de données par mois. Ces données étaient habituellement utilisées pour un contrôle en temps réel par le personnel soignant qui était prévenu en cas de résultats inquiétants. Cependant l'énorme quantité de données produites en continue par l'ensemble de ces capteurs n'était pas utilisée pour de la prédiction. Le travail de Carolyn MacGregor, docteur en informatique à l'Université Ontario Institute of Technology, a été de trouver des schémas de détection d'infections grâce à l'ensemble des données ayant été collectées par le passé. Le programme qui en a découlé a d'ailleurs été appelé Artemis en référence à la déesse grecque qui accompagne les enfants vers l'âge adulte. Ce sont ici des algorithmes de machine learning [14] qui sont utilisés. C'est-à-dire des algorithmes de classification statistiques qui évoluent de manière autonome au fur et à mesure que de nouvelles données sont traitées. Plus le nombre de données utilisées est grand, plus l'algorithme est efficace. Les résultats [1] de ces recherches ont permis de détecter certaines infections graves jusqu'à 24h plus tôt qu'avec les méthodes traditionnelles.
Cependant les hôpitaux ne disposent en général pas de moyens suffisants pour enregistrer toutes les données émises par les capteurs médicaux. C'est pourquoi IBM utilise le logiciel InfoSphere Streams [2] [30] qui permet la capture et l'analyse des données en temps réel sans jamais enregistrer sur un disque dur le flux de données. Le logiciel Artemis Cloud permet aussi de stocker les données à distance pour permettre aux données d'être réutilisées pour de prochaines recherches.

1. Voici le lien d'une vidéo explicative produite par IBM `http://www.youtube.com/watch?v=WNccLBzR_I4`
2. Voir sur le site d'IBM `http://www-03.ibm.com/software/products/us/en/infosphere-streams/`.

7 Le Big Data et le gouvernement américain

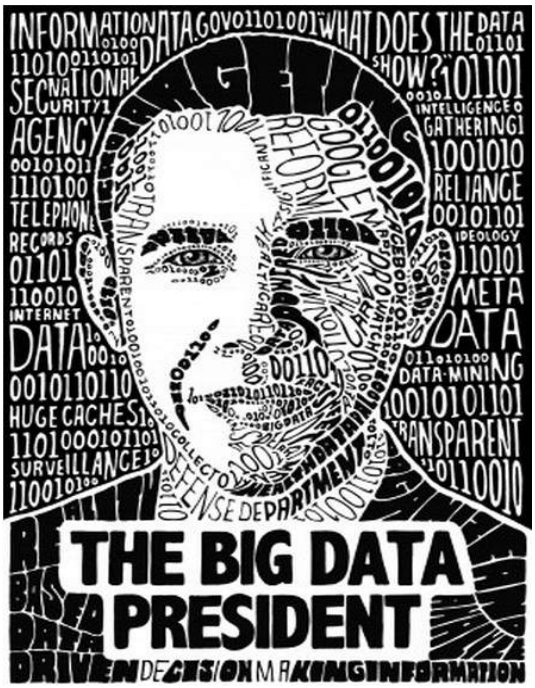
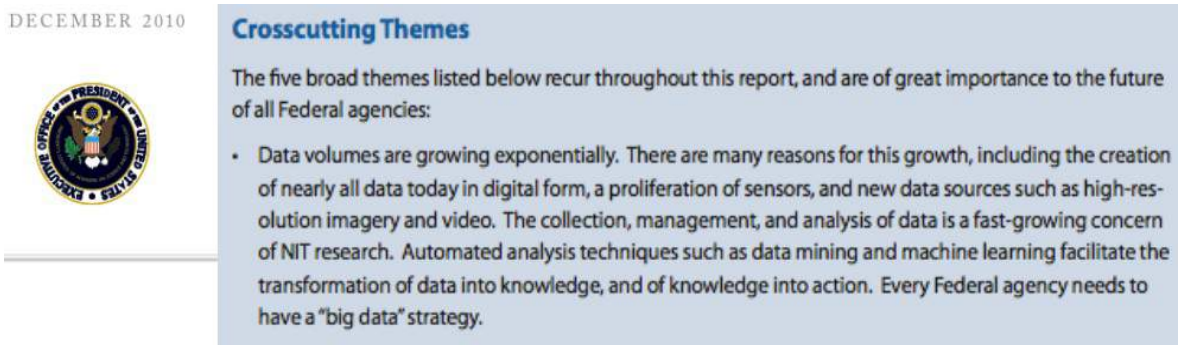


FIG. 7.1 – «The Big Data President». Dessin de Sarah A.King pour le Washington Post.

Le gouvernement américain a officiellement pris conscience de l'importance du Big Data lorsque l'administration Obama a annoncé en Mars 2012 la création d'un fond de 200 millions de dollars pour la recherche sur le Big Data («Big Data Research and Development Initiative»). On peut lire dans un rapport destiné à la Maison Blanche en décembre 2010 : «Toutes les agences fédérales ont besoin d'avoir une stratégie Big Data.». Le Président Obama semble en effet particulièrement réceptif aux enjeux du Big Data. Tellement réceptif que le Washington Post va jusqu'à parler du «Président Big Data» [1].

Les agences fédérales disposent en effet d'énormes quantités de données provenant de différentes sources à travers tout le pays. La question du stockage et de l'analyse de ces données est aujourd'hui l'une des priorités aux Etats-Unis. C'était d'ailleurs le sujet d'une audition au Congrès américain, le 24 avril 2013, durant laquelle les professeurs David McQueeney, vice-président de la stratégie technologique chez IBM Research, Michael Rappa, directeur du «Institute for Advanced Analytics» et Farnam Jahanian, directeur du département CISE (Computer and Information Science and Engineering) à la NSF (National Science Fondation), répondaient

1. «The Big Data President», Washington Post, `http://goo.gl/0BD4b`.



DECEMBER 2010

Crosscutting Themes

The five broad themes listed below recur throughout this report, and are of great importance to the future of all Federal agencies:

- Data volumes are growing exponentially. There are many reasons for this growth, including the creation of nearly all data today in digital form, a proliferation of sensors, and new data sources such as high-resolution imagery and video. The collection, management, and analysis of data is a fast-growing concern of NIT research. Automated analysis techniques such as data mining and machine learning facilitate the transformation of data into knowledge, and of knowledge into action. Every Federal agency needs to have a "big data" strategy.

FIG. 7.2 – Extrait du rapport de Décembre 2010 remis à la Maison Blanche par le comité de conseil du Président pour la Science et la Technologie.

aux questions des élus pour les sensibiliser à l'importance du Big Data. Dans un rapport [2] sur l'évolution globale du monde, rédigé par «The office of the director of National Intelligence» en décembre 2012, le Big Data est mentionné comme l'une des tendances qui influera le plus sur notre société d'ici à 2030. D'après une étude MeriTalk [3] «Smarter Uncle Sam : The Big Data forecast» le gouvernement fédéral américain estimerait que 14% d'économies pourraient être réalisées dans leur budget grâce au Big Data. Cette somme représente presque 500 milliards de dollars.

En période de coupes budgétaires liées à la Séquestration [4], les économies «Big Data» pourraient bien s'avérer capitales.

7.1 Lutte contre la fraude

Le service des impôts américains (IRS) a entrepris une collaboration avec les équipes d'IBM pour analyser leur immense base de données. L'IRS reçoit chaque année les déclarations d'impôts de l'ensemble des résidents américains. La collaboration avec IBM a pour but d'analyser les déclarations au regard de toutes les précédentes déclarations y compris celles que l'on a découvert frauduleuses.
Ainsi la recherche Big Data consiste ici à trouver des schémas typiques de fraudeurs. C'est encore des problèmes statistiques de classification que l'on se pose. On veut ranger les individus en seulement deux catégories (en utilisant une méthode du k-means avec $k = 2$) : les contribuables honnêtes et les fraudeurs. Au fil des ans et des nouveaux contrôles l'algorithme s'améliore car il dispose de plus de données et donc d'un modèle plus fiable. C'est de nouveau un problème d'apprentissage statistique (ou machine learning [14]). Après analyse, le logiciel propose lui-même les individus propices à un contrôle fiscal. Lors d'une conférence IBM le 21 mars 2013, Paul Seckar, partenaire associé aux relations publiques chez IBM, affirmait que grâce au logiciel mis en place par IBM le ratio personnes contrôlées/fraudeurs atteignait aujourd'hui $1,1$. Ce qui signifie que pour 11 personnes proposées par l'ordinateur comme étant potentiellement frauduleuses, 10 l'étaient effectivement. Grâce à un «simple» partitionnement statistique il a ainsi été possible d'augmenter sensiblement le rendement des inspecteurs des impôts.

2. Global trends 2030 : Alternative worlds.
3. `http://www.meritalk.com/home.php`.
4. Voir `http://en.wikipedia.org/wiki/Budget_sequestration_in_2013`.

Les services de lutte contre la fraude à la carte bancaire font également appel à des analyses poussées pour pouvoir détecter automatiquement et en temps réels les tentatives de fraudes.

7.2 La défense

Le département de la défense américaine (Department Of Defense) fait un «gros pari sur le Big Data». Dans un communiqué officiel de la Maison Blanche [5] datant du 29 mars 2012, il est annoncé que 250 millions de dollars seront alloués chaque année pour des projets Big Data au sein de l'administration américaine et notamment au sein des différents organismes de défense. Les deux objectifs affichés sont :

— Exploiter des grandes masses de données de façons nouvelles afin de créer des systèmes réellement autonomes capables de prendre des décisions par eux-mêmes.
— Améliorer l'état des connaissances en reconnaissance vidéo afin de venir en aide aux combattants sur le terrain ainsi que les technologies de «text mining» dans toutes les langues afin de faciliter le travail des analystes.

7.2.1 La sécurité intérieure

Au nom de la lutte anti-terroriste le gouvernement américain s'est lancé dans la construction du plus grand data center du monde. Situé dans l'Utah, sa capacité de stockage se comptera en yotta-octets (10^{24} octets). La nature exacte des données qui y seront stockées n'est pas publique. Les critiques quant aux violations de la vie privée se font d'ailleurs de plus en plus féroces depuis les révélations sur le programme PRISM [6]. La NSA (National Security Agency) possède en effet des accords avec les géants du net américains (Microsoft, Yahoo, Google, Facebook, YouTube, Skype, AOL, PalTalk ou encore Apple) pour accéder aux données personnelles des utilisateurs.

Les applications envisageables sont nombreuses. L'exemple du projet Far Out vu précédemment (section 6.1 p.27) donne une idée de ce qu'il est possible de faire simplement avec des méta-données. Parmi les programmes présentés officiellement, le programme ADAMS (Anomaly Detection At Multiple Scales) semble être du même genre. C'est un programme qui a pour but de repérer des changements suspects de comportements. L'idée étant de repérer un soldat en bonne santé mentale qui commence à perdre les pédales ou un «bon» citoyen qui se radicalise. De longues périodes d'enregistrement de données de types GPS, courriels, sites visités peuvent permettre de repérer de tels changements de comportement.

L'un des autres grands défis est d'améliorer les technologies de reconnaissance vidéo. Le programme «Mind's Eye» a pour objectif de permettre aux ordinateurs de comprendre tout seul les images vidéos. C'est-à-dire reconnaître des visages mais aussi des objets ou des comportements suspects (course-poursuites, bagarres ...) pour pouvoir rendre des résumés écrits automatisés.

Un autre programme nommé VIRAT (Video and Image Retreval and Analysis) se concentre

5. «Fact Sheet : Big Data Across the Federal Government» - `http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf`.
6. «U.S. confirmes that it gathers data overseas»- Charlie Savage, Edward Wyatt & Peter Baker - New York Times - 06/06/13 - `http://goo.gl/BuRMq`.

plus sur une aide à la reconnaissance aérienne en zone de combat : repérer les positions stratégiques au sol. L'un des objectifs de VIRAT est également de fabriquer une sorte de moteur de recherche au sein de grandes bibliothèques vidéos (vidéos prises par les avions ou les soldats).

Neuf projets au total sont à comptabiliser au sein de la DARPA (Defense Advanced Research Projects Agency). La plupart s'intéresse à la reconnaissance vidéo et textuelle ou plus généralement la gestion de grandes quantités de données non structurées. Deux autres projets du DHS (Department of Homeland Security) bénéficient également de soutiens financiers.

7.3 La santé

De nombreux projets ont pour but d'améliorer le système de santé. L'objectif principal est de s'orienter vers une médecine plus personnalisée, donc plus efficace et moins coûteuse.
Le ministère des anciens combattants (VA : Department of Veterans Affairs) a lancé un programme de centralisation et d'homogénéisation des données médicales des vétérans. Le CDW (Corporate Data Warehouse) [7] est un data center créé spécialement pour recevoir toutes les données provenant des différents services des vétérans. Le programme «Health Data Repository» est chargé de structurer toutes les données de manière homogène pour qu'elles puissent être traitées par le CDW. Deux programmes sont ainsi dédiés à la recherche pour la santé des vétérans :

- Le programme GenISIS (Genomic Information System for Integrated Science) enregistre les données génétiques des vétérans ainsi que leur dossier médical complet pour personnaliser les traitements.
- Le programme Million Veteran Program vient en aide au programme GenISIS en procédant au séquençage génétique des vétérans volontaires ainsi qu'à la description phénotypique des volontaires dans l'espoir de faire avancer la recherche médicale.

Dans un souci de réduction des coûts, le gouvernement américain veut centraliser les données des patients. Le CDC (Center for Disease Control & Prevention), dont nous avons parlé plus haut (section 4.1 p.18), ainsi que le CMS (Center for Medicare & Medicaid Services) sont deux centres subventionnés pour stocker et gérer les informations au sein du HHS (Health and Human Services). Le CMS a une architecture basée sur Hadoop et un évaluation sur l'utilisation des nouvelles technologies Big Data a été lancé. Certaines économies sont envisagées en automatisant les programmes d'éligibilité aux assurances et programmes d'aide Medicare et Medicaid.

Le NIH (National Institutes of Health) s'engage également dans la voie «Big Data». La lutte contre le cancer est l'une des priorités. Pour cela deux objectifs sont poursuivis. Le premier est d'améliorer la détection et le diagnostique de la maladie grâce aux techniques d'imageries médicales. Les archives du TCIA (The Cancer Imaging Archive) sont mises à la disposition des chercheurs afin qu'ils puissent améliorer les logiciels d'aide à la décision pour les médecins. Le projet TCGA (The Cancer Genome Atlas) consiste lui à collecter des séquences génétiques de patients atteints de cancer pour faire avancer la recherche sur ce sujet. D'ici 2014 le TCGA devrait détenir plusieurs péta-octets de données.

7. Voir `http://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm`.

7.4 La recherche

Depuis 2012 le gouvernement américain investit 60 millions de dollars par an pour la recherche Big Data dans les centres de recherche gouvernementaux.

7.4.1 La NASA

La NASA (National Aeronautics & Space Administration) en tant que centre de recherche aéronautique et aérospatial est un gros consommateur de Big Data. Plusieurs programmes mentionnés dans le rapport officiel «Fact Sheet : Big Data across the Federal Government» ont pour but d'utiliser les technologies Big Data pour analyser les données satellites. Le projet ESDIS (NASA's Earth Science Data and Information Systems) par exemple doit stocker puis analyser les images satellites prises depuis plus de 15 ans pour faire avancer l'état de la science à propos du changement climatique.
On peut également mentionné l'accord (Space Act Agreement) entre la NASA et l'entreprise «Cray, the supercomputer company», spécialisée dans la résolution de problème Big Data, qui a été signé pour accélérer l'analyse des données recueillies par la NASA.

7.4.2 L'énergie

Le DOE (Department Of Energy) n'est pas en reste puisqu'il compte 9 programmes liés au Big Data. Parmi eux, on compte un programme nommé «Mathematics for Analysis of Petascale Data» [15] et spécialement destiné au développement des méthodes mathématiques pour le Big Data. Un comité d'experts a été réuni dès 2008 pour mettre en avant les grands enjeux mathématiques. Il en ressort qu'une large palette de disciplines mathématiques est essentielle :

- les statistiques
- l'optimisation
- les calculs d'incertitudes
- l'apprentissage statistique ou machine learning
- l'analyse de graphes et de réseaux
- la topologie.

7.4.3 La NSF (National Science Foundation)

La NSF (National Science Foundation), l'agence fédérale pour la recherche fondamentale, mise aussi sur le Big Data. De nombreux programmes sont consacrés au Big Data et notamment des programmes de recherche en mathématiques fondamentales. Le département des sciences mathématiques de la NSF a récemment lancé un programme de soutien financier aux travaux mathématiques concernant le Big Data. La coopération au sein de la NSF entre le DMS (Division of Mathematical Sciences) et l'OCI (Office of Cyberinfrastructure) a donné naissance à un nouveau programme [8] qui conçoit le Big Data comme une nouvelle discipline scientifique à la frontière entre la mathématique, la statistique et l'informatique.
Parmi les 13 autres projets Big Data engagés au sein de la NSF on peut également mentionner le FRG (Focused Research Group, stochastic network models). C'est un projet de recherche sur les modèles stochastiques (c'est-à-dire aléatoire) pour développer et unifier un cadre théorique pour les méthodes de «text mining» qui doivent permettre une analyse automatique de texte par les ordinateurs.

8. Le programme s'intitule «The Computational and Data-enabled Science and Engineering in Mathematical and Statistical Sciences». Voir aussi [6].

Troisième partie

Quelles formations pour devenir «Data scientist» ?

Selon la Harvard Business Review, le métier de «data scientist» serait le métier le plus sexy du 21ème siècle [7]. Du fait de la très récente dynamique Big Data, beaucoup d'entreprises, parmi lesquelles on compte les géants américains Google, Microsoft, IBM, Facebook, Twitter, Linkedin mais également de très nombreuses start up, sont à la recherche de «data scientists». Il est toutefois difficile de caractériser et surtout de trouver les profils recherchés. Un «data scientist» doit en effet posséder de solides bases en informatiques et être capable de programmer. Il doit aussi avoir de nombreuses connaissances mathématiques. Mais il doit également avoir une vision globale sur de l'état de l'art ainsi qu'une curiosité pour lui permettre de prendre des initiatives afin d'aller chercher les informations utiles là où elles se trouvent. Cependant, les universités américaines réagissent avec inertie et les premiers Masters réellement orientés Big Data verront le jour à la rentrée 2013 dans quelques unes des plus grandes universités américaines comme Columbia University ou George Washington University.

8 Les connaissances requises

Les spécialistes actuels du Big Data ont des profils très variés et ont, au sein de leur entreprise, des postes différents. La dénomination de «data analyst» n'existait même pas il y a 5 ans, il règne ainsi un flou autour du profil des candidats recherchés. Il est difficile de définir le rôle et la place exacte du «data scientist» au sein de l'entreprise [17]. Même si il est de bon ton d'expliquer qu'un bon analyste ne se reconnait pas à ses capacités à coder mais plutôt à son ouverture d'esprit et à sa compréhension du monde qui l'entoure, on ne peut nier le caractère scientifique du métier, ou du moins les bases théoriques qu'il requiert.

Il est évidemment nécessaire pour faire «data analyst» d'être familier avec l'informatique et de savoir coder. L'une des difficultés est de concevoir des programmes qui vont utiliser les données là où elles sont, sans les faire venir à soi (car déplacer des très grandes quantités de données est trop coûteux). Mais nous nous concentrons ici sur les aptitudes mathématiques qui peuvent être mises en jeu dans le Big Data. Nous avons déjà vu, en témoigne l'index, que de nombreux pans de la mathématique sont impliqués dans le développement du Big Data. Dés juin 2008, un comité d'expert américain avait été réuni par le DOE (Department of Energy) pour clarifier les enjeux mathématiques du Big Data. Leur conclusion peut être lue dans [15]. Il y a plusieurs défis mathématiques qui se présentent.

- **La très haute dimension.**
  Les programmes informatiques/mathématiques (machine learning, classification, détection d'anomalies...) d'aujourd'hui ne fonctionnent pas suffisamment rapidement ou présentent des problèmes de robustesse lorsqu'on travaille avec de très grandes dimensions (plusieurs centaines de milliers). Comment repérer des schémas de petites dimensions au milieu de données de très grande dimension ?
- **La modélisation.**
  Il ne faut pas perdre de vue que les modèles utilisés doivent être prouvés mathématiquement. Les modèles doivent également être fiable lorsque les données sont de mauvaise qualité. C'est-à-dire soit à cause d'un fort bruit soit à cause de données manquantes soit par un déséquilibre dans les différentes catégories de données présentes. Il faut également tenir compte qu'avec un très grand nombre d'observations, l'apparition de phénomènes «rares» devient non négligeable.
- **Le calcul d'incertitude.**
  Il est primordial de pouvoir calculer l'incertitude d'un résultat et de porter un regard critique sur le travail effectué. Il est également important de minimiser le taux de «faux-positif» dans un processus de classification.

Chaque problème met en jeu plusieurs disciplines mathématiques différentes, les principales étant :

- Les statistiques (K-means clustering, transformation de Fisher, méthodes de Monte-Carlo, modèles gaussiens, tests de Student...)
- L'optimisation (analyse)
- Les calculs d'incertitude

— Le machine learning (apprentissage statistique)
— L'analyse de réseaux et de graphes (algèbre, chaînes de Markov...)
— La topologie
— Le calcul stochastique (modèles de Markov, modèles aléatoires...)

9 Les formations proposées

Il n'existe aujourd'hui aucun diplômé «Big Data». Les premières formations académiques Big Data ne commenceront qu'à la rentré 2013. Les premiers diplômés n'arriveront pas avant 2014. Certaines entreprises, conscientes de la nécessité de posséder un «data scientist» dans leur équipe, proposent déjà depuis quelques années des formations avancées. A l'embauche, les jeunes étudiants en thèses sont familiarisés par ces formations aux enjeux et techniques du Big Data.

L'une des méthodes à la mode pour trouver les meilleurs «data scientists» du moment est d'organiser des compétitions ouvertes à tous. L'entreprise Kaggle propose librement sur Internet des problèmes «Big Data» que lui soumettent des entreprises partenaires. Chaque problème est accompagné d'une récompense financière [1]. En juillet 2013, on peut voir un problème d'Amazon qui offre $5000 à qui trouvera la meilleure solution pour gérer les droits d'accès sur différentes serveurs des nouveaux employés. La société californienne Belkin, qui fabrique des composantes électroniques, offre elle $25 000 pour fabriquer un programme qui analyse en temps réel les consommations d'énergie de chaque employé afin de proposer des économies personnalisées. En plus des prix offerts, les participants aux compétitions Kaggle obtiennent des points à chaque participation. Ces points servent à mettre à jour le classement Kaggle [2] qui met en avant les meilleurs «data scientists» du moment.

Le fort besoin en «data scientists» a finalement décidé les grandes universités américaines à proposer des parcours spécialisés Big Data. La célèbre université Columbia à New York a créé au sein de l'IDSE (Institute for Data Sciences and Engineering) un nouveau diplôme Big Data appelé «Certification of Professional achievement in Data Sciences». Ce nouveau programme qui ouvrira accueillera des étudiants pour la première fois à l'automne 2013 proposera 4 matières principales [3].

— Probabilités et Statistiques. Ce cours a pour but de fournir aux étudiants les outils mathématiques suffisant pour comprendre et naviguer sereinement dans le monde du Big Data.
— Algorithmes pour Big Data. Les étudiants apprennent ici comment utiliser les outils informatiques indispensables à un «data analyst».
— Machine learning. Un cours spécifique est reservé à l'apprentissage statistique étant donné

1. Voir les compétitions ouvertes sur `http://www.kaggle.com/competitions`.
2. Voir le classement sur `http://www.kaggle.com/users`.
3. Une vidéo explicative est disponible sur le site Internet de l'université `http://idse.columbia.edu/certification-professional-achievement-data-sciences`.

l'importance qu'elle joue dans le Big Data.

— Exploration des données. Ce cours s'intéresse à l'analyse mais surtout à la visualisation des données. Comment rendre un résultat visible parmi toutes les données traitées.

Stanford University délivre uniquement via des cours sur Internet depuis 2012 un diplôme orienté Big Data dénommé «Mining Massive Data Sets Graduate Certificate» [4]. George Washington University lance également à l'automne prochaine un nouveau Master Big Data «Master of Science in Business Analytics» qui a pour but de former des «data scientists» prêt à travailler en entreprises. Au total plus d'une vingtaine d'universités américaines [5] lancent l'année prochaine des Masters Big Data.

Les formations Big Data se doivent d'être d'un haut niveau théorique en mathématiques et en informatiques. La spécialisation Big Data au niveau Master 1 (M1), après une licence 3 (L3) en mathématiques ou en informatique ou une école d'ingénieur, semble la plus appropriée. Certains problèmes Big Data feront également, à n'en pas douter, l'objet de nombreuses thèses.

4. Pour en savoir plus, consulter `http://scpd.stanford.edu/public/category/courseCategoryCertificateProfile.do?method=load&certificateId=10555807`.

5. Pour plus d'informations voir `http://www.informationweek.com/big-data/slideshows/big-data-analytics/big-data-analytics-masters-degrees-20/240145673`.

Conclusion

La révolution Big Data est en marche et va certainement trouver des applications dans tous les domaines imaginables. Ce nouveau paradigme promet des avancées majeures dans notre compréhension du monde. La médecine, l'astrophysique, les sciences sociales ou la recherche scientifique en général voient s'ouvrir les portes d'un univers différent. Ce changement d'échelle de données n'est pas seulement quantitatif. Il nous permet d'accéder à de nouvelles connaissances qui nous étaient auparavant inaccessibles. Pour reprendre l'analogie de Viktor Mayer-Schönberger, l'un des auteurs de [4], les lois de la physique qui prédominent à l'échelle microscopique diffèrent de celles qui importent à l'échelle macroscopique [6]. Le Big Data est un bouleversement de même nature.

Cependant, au delà de toutes les promesses du Big Data, cette révolution amène également de nombreuses interrogations. De nouveaux types d'abus ne sont-ils pas à craindre ? Comment encadrer cette nouvelle technologie ? Les révélations sur le programme PRISM sont une première alerte sur les dangers du Big Data. Les données informatiques personnelles à disposition de certains grands groupes (comme Google, Apple, Facebook pour ne citer qu'eux) et les usages qu'ils peuvent en faire dépassent de très loin l'imagination des consommateurs. Comment savoir à quelles fins nos données sont utilisées ? Peut-on accepter une surveillance de l'Etat à travers nos données personnelles [19] ?

Il est également primordial de ne pas tomber dans ce que certains appellent déjà «La dictature des données» [5]. Le Big Data doit rester une science qui apporte des solutions concrètes et sûres. Faire des prédictions ou prendre des décisions en suivant aveuglément des schémas que l'on semble apercevoir dans un tas de données non pertinentes est une grossière erreur [31]. Former des «data scientists» compétents et capables de tirer profit du Big Data tout en sensibilisant l'ensemble de la société, y compris nos responsables politiques, à ce changement profond et à ses conséquences : voici le nouveau défi du Big Data.

6. Voir l'interview de Viktor Mayer-Schönberger par Dominique Leglu, directrice de la rédaction de Sciences et Avenir : `http://sciencesetavenir.nouvelobs.com/decryptage/20130708.OBS8487/les-big-data-vont-revolutionner-nos-vies-notre-travail-et-notre-pensee.html`.

Bibliographie

[1] Orly Alter, Patrick O. Brown et David Botstein : Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Stanford University*, Jan 2013.

[2] Mikuail Bautin : Identification of most frequent subsequences. *Computational biology project*, Dec 2006.

[3] Alexander Chan : Analysis of pariwise sequence alignment algorithm complexities : Needleman-Wunsch, Smith-Waterman, FASTA, BLAST and gapped BLAST.

[4] Kenneth Cukier et Victor Mayer-Schönberger : *Big Data : A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt, 2013.

[5] Kenneth Cukier et Viktor Mayer-Schönberger : The dictatorship of data. *MIT Technology Review*, May 2013.

[6] James Curry et al. : Proposal and award policies and procedures guide : grant proposal guide. *National Science Foundation*, Oct 2012.

[7] Thomas Davenport et D.J. Patil : Data scientist : the sexiest job of the 21st century. *Harvard Business Review*, Oct 2012.

[8] Dean et Ghemawat : Mapreduce : simplified data processing on large clusters, 2004.

[9] Fabrice Demarthon, Denis Delbecq et Gregory Flechet : The Big Data revolution. *CNRS international magazine*, 28:20–27, Jan 2013.

[10] Jeremy Ginsbert, Matthew H.Mohebbi et al. : Detecting influenza epidemics using search engine query data. *Nature*, 457, Feb 2009.

[11] Tony Hey, Stewart Tansley et Kristin Tolle : *The Fourth Paradigm : Data-Intensive Scientific Discovery.* 2009.

[12] Martin Hilbert : How much information is there in the "information society" ? *Significance*, 9:8–12, Aug 2012.

[13] Chris Holmes, Peter Kecskemethy et Chris Gamble : Methods for big data in medical genomics : Parallel Hidden Markov Models in Population Genetics, Jan 2013.

[14] David Jensen et Jennifer Neville : Correlation and sampling in data mining. *University of Massachussetts.*

[15] Philip Kegelmeyer et al. : Mathematics for analysis of petascale data. *Report on a Department of Energy Workshop*, Jun 2008.

[16] F. Lassage, R. Iconikoff, A. Debroise, M. Grousson, J. Michaux et M.Fontez : Google, le nouvel Einstein. *Science & Vie*, pages 46–63, Jul 2012.

[17] Jessica Leber : In a data deluge, companies seek to fill a new role. *MIT Technology Review*, May 2013.

[18] Tom Madden : The BLAST sequence analysis tool. *The NCBI Handbook*, Aug 2003.

[19] Jean-Marc Manach : La DGSE a le "droit" d'espionner ton Wi-Fi, ton GSM et ton GPS aussi. *Le Monde*, Jul 2013.

[20] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh et Angela Hung Byers : Big data : the next frontier for innovation, competition and productivity. *McKinsey Global Institute*, May 2011.

[21] Carolyn McGregor : Big data in neonatal intensive care. *Computer*, pages 54–58, Jun 2013.

[22] Tim Oates : Identifying distinctive subsequences in multivariate time series by clustering.

[23] Valery A. Petrushin : Hidden markov models : fundamentals and applications. *Online Symposium for Electronics Engineer*, 2000.

[24] Soumya Raychaudhuri *et al.* : Identifying relationships among genomic disease regions : predicting genes at pathogenic SNP associations and rare deletions. *Public Library of Science*, 2009.

[25] Antonio Regalado : The data made me do it. *MIT Technology Review*, May 2013.

[26] Antonio Regalado, Patrick Tucker *et al.* : Big data gets personal. *MIT Technology Review*, May 2013.

[27] Adam Sadilek et John Krumm : Far Out : Predicting Long-Term Human Mobility. *AAAI Conference on Artificial Intelligence*, 2012.

[28] Atish Das Sarma, Anisur Rahaman Molla, Gopal Pandurangan et Eli Upfal : Fast Distributed PageRank Computation. *Lecture Notes in Computer Science*, 7730:11–26, 2013.

[29] Michael C. Schatz : Blastreduce : High Performance Short Read Mapping with Map Reduce, 2009.

[30] Tom Simonite : Watson goes to work in the hospital. *MIT Technology Review*, Apr 2013.

[31] Nassim Nicholas Taleb : *Fooled by randomness : the hidden role of chance in life and in the markets*. Texer, 2004.

[32] Patrick Tucker : Has Big Data made anonymity impossible ? *MIT Technology Review*, May 2013.

[33] Esko Ukkonen : On-line construction of suffix trees. *Algorithmica*.

[34] Jessica Voytek et Bradley Voytek : Automated cognome construction and semi-automated hypothesis generation. *Journal of Neuroscience Methods*, 2010.

[35] Zhenghua Xue et al. : Compression-aware I/O performance analysis for big data clustering, 2012.

Index