Mémoire de fin d'études

Big Data

Comment le Big Data peut influencer notre vie quotidienne ?

Valentin MEYNE Architecture des Logiciels
Ismaïl NGUYEN Architecture des Logiciels

Maître de Mémoire : Gael COAT

ESGI
école supérieure de génie informatique

5A 2016/2017

RÉSEAU GES
GRANDES ÉCOLES SPÉCIALISÉES

REMERCIEMENTS

Nous adressons nos remerciements aux personnes qui nous ont aidé dans la réalisation de ce mémoire.

En premier lieu, nous tenons à remercier M. COAT, professeur à l'ESGI. En tant que Directeur de mémoire, il nous a guidé dans notre travail et nous a aidé à trouver des solutions pour avancer.

Nous remercions aussi nos camarades de classes, ainsi que tous nos professeurs, qui nous ont aidé et soutenu durant toute ces années au sein de l'école.

TABLE DES MATIÈRES

RÉSUMÉ

Vouloir comprendre le passé et prédire l'avenir. C'est un rêve désormais à la portée de l'humanité.

Pouvoir incommensurable de prédire les faits et gestes d'un individu. En se basant sur ses goûts, ses préférences, son vécu, son corps et ses actions, ces données permettent de le confondre entre mille. Ces données nous les produisons tous, chaque instant. Elle sont moissonnées sur les sites internets et endroits sur Terre que nous visitons. Néanmoins, la source principale à ce jour est constituée de nos objets connectés.
Ne vous croyez pas épargné si vous n'êtes pas féru de nouvelles technologies car nous en possédons tous au moins un aujourd'hui. L'objet connecté auquel nous faisons allusion est notre smartphone.

Cette quantité titanesque de données, tellement grande, met à l'épreuve les géants de l'industrie informatique qui collectent ces données pour les transformer en information pertinente.

Elle sont appelées **Big Data**.

Avec un pouvoir aussi grand entre les mains d'entreprises les plus puissantes du monde nous pouvons à juste titre nous poser la question suivante :

Comment le Big Data peut influencer notre vie quotidienne ?

C'est là que nous intervenons avec ce présent document. Nous, deux étudiants tombés dans la technologie étant petits, essayons d'apporter une réponse, en recherchant, regroupant, analysant divers articles de presses, des conférences, des livres, d'experts ou de penseurs.

Ce qui en ressort c'est que cette avancée technologique a un impact global. En effet, elle commence à provoquer une mutation de grand nombre de secteurs.
De l'optimisation de la production, à l'anticipation des besoins du consommateur, toute la chaîne de grande distribution semble en bénéficier.
La sécurité du citoyen se voit dotée d'un potentiel d'amélioration énorme. Des déplacements, aux accidents ménager en passant par la santé, l'analyse de chaque situation va permettre de sécuriser notre quotidien à des niveaux inatteignables aujourd'hui.

Notre quotidien va donc se voir grandement sécurisé et notre consommation être optimisée et personnalisée aux dépens de notre autonomie et notre liberté.

En effet, nous pouvons voir notamment en politique, qu'il y a une recherche d'obtention de méthodes pour s'octroyer les faveurs des électeurs. Adapter un discours en fonction des résultats de l'analyse de ce que les gens veulent est déjà une méthode existante. Aussi, les gouvernements sont tentés de se servir des informations collectées sur chacun pour détecter les éléments problématiques de leur population. Définition des plus floues sur le long terme car ce qui reste l'exception antiterroriste aujourd'hui peut devenir la norme de demain.
Chaque entreprise peut être tentée, afin de rester compétitive, de jouer sur les prédisposition d'addiction, connues ou inconnues, de son client et de le garder captif. Ainsi, notre vie privée pourrait se voir dévoilée sans possibilité de lutter contre des entitées n'ayant pas les mêmes intérêts que nous.
Tous ces points se résument en un mot : "Contrôler". En effet, le plus grand danger est que, connaissant l'avenir et ses leviers, un acteur l'influence pour son intérêt.

Nous conclurons que le Big Data et les changements qu'il implique sont aussi grand, voir plus, que la découverte de l'atome en son temps et que malgré son potentiel dévastateur, le Big Data peut améliorer la vie quotidienne de chacun. Seul l'Histoire nous dira avec certitude quels usages en seront fait sur le long terme mais il influencera grandement notre vie quotidienne.

ABSTRACT

Wanting to understand the past and foresee the future. It is from now on a dream within the reach of Humanity.

Immeasurable power to foresee the comings and goings of an individual. Basing on this individual tastes, preferences, experiences, body and acts, this data allows to identify him. We all produce them, at any time. It is harvested on websites and places on Earth we visited. Nevertheless, the main source to this day is our connected objects.
Don't believe you are spared if you are not passionate about new technologies because today we all possess at least one. It is our smartphone.

This huge amount of Data, so big that it challenges the giants of IT industry that collect it to obtain relevant informations.

It is called **Big Data**.

With this huge power in the hands of the most powerful companies in the world we can ask ourselves :

How Big Data can influence our everyday life ?

This is where we come up with this document. We, two students who growth with computers, try to bring an answer by searching, grouping, analysing numerous press articles, conferences, books, from experts or thinkers.

It appears that this technological progress have a global impact. Indeed, many field have started to mutate.
From optimisation of production to anticipating consumer needs, the large retailers seems to benefit from it.
The citizen security can be improved too. It can be trips, domestic accidents or health, by analysing each situation, the everyday life will be more secure than we could imagine.
Our everyday life will be more secure and our consommation will be optimized at the expense of our independence and freedom.

Actually, in politics, we can see there is a research for methods to obtain electors favours. Adapting a speech according to the results of the analyse of what people want is already an existing method. Also, governments are tempted to use the

informations collected on everyone to detect the ones that can make trouble. The definition of someone who make trouble in the eyes of a government can change in the long run. An exception created today for anti-terrorism can become the norm tomorrow.
Each company can be tempted, to stay competitive, to take advantage of consumer addiction predispositions, known or not by the client and keep him captive. In this way, our private life could be revealed without any possibility to fight back against entities that don't have the same goal as our.
All this points sum up in one word : "Control". In fact, the greatest danger is to, knowing the future and his levers, a player influences it in his own interest.

We will conclude by saying that Big Data and the changements it involves are greater than the discovery of atom in its time and despite its devastating potential, Big Data can improve the everyday life of each of us. Only History will tell us with certainty which use of it will be done in the long run but it will influence our everyday life.

MOTS CLÉS

Données en masse

Objets connectés

Intelligence Artificielle

Données liées

Apprentissage machine

Prédiction

Algorithme

Analytique

KEYWORDS

Big Data

Internet of Things

Artificial Intelligence

Linked datas

Machine Learning

Prediction

Algorithm

Analytics

INTRODUCTION

Google, Amazon, Facebook, Apple, Microsoft, ces entreprises sont souvent regroupées sous l'acronyme "GAFAM". Elles sont les leaders du marché des nouvelles technologies informatiques. Malgré la différence dans les produits et services qu'elles proposent, toutes ont commencé à s'intéresser au cours des dernières années à, puis ont intégré au coeur de leur business model, la collecte massive de nos données. Données que nous produisons tous, à chaque instant, par l'usage de nos matériels informatiques. Smartphones, tablettes, objets connectés, ordinateurs, automobiles, logements connectés (domotique), tous ces appareils sont des extensions de nous-mêmes et sont concernés par ce moissonnage.

Les données amassées sont tellement nombreuses que leur traitement est devenu impossible par les systèmes de gestion de base de données classiques. Ces entreprises ont dû innover en inventant, ou en promouvant l'invention, d'outils informatiques permettant le traitement et l'exploitation de ces données massives. Ce sont ces données qui sont communément appelées Big Data.

Au vu de l'Histoire, notamment l'histoire militaire à l'échelle mondiale et l'histoire du marketing, il est admis que tout type de données recoupées entre-elles peuvent faire apparaître des patterns et informations utiles à celui qui les génère ou à la personne, la communauté, l'entreprise ou l'Etat à qui ces données vont être diffusées ou vendues. Informations sur les comportements de vie privée, d'utilisation, d'achat, de communication, les habitudes, préférences et intentions de chacun ou d'un groupe de personnes.

L'optique et la promesse de l'avènement d'une technologie offrant un tel pouvoir a suscité depuis 2012, l'engouement des grandes entreprises informatiques et des dirigeants des pays développés pour l'ère du numérique.
Les appareils informatiques alimentent désormais une technologie de puissance comparable à la découverte de la fusion nucléaire. A chaque nouvelle technologie, de nouvelles pratiques émergent. La réticence à l'idée de les adopter et explorer les possibilités réside dans la connaissance que l'Homme a et dans sa capabilité à se servir d'une technologie comme d'une arme ou comme méthode servant à esclavager.

Une grande partie, si ce n'est la totalité, des oeuvres d'anticipation et de science-fiction qui abordent le thème de la surveillance ou d'une totale connexion à un Internet dépeignent un futur dystopique et concentrent leurs récits sur les dérives et effets néfastes sur l'Homme.
Elles forgent notre imaginaire mais aussi notre façon de visualiser les dégâts futurs sur le monde avant même d'avoir la possibilité de mettre une technologie en application.
Parmis elles nous pourrions citer, et recommander, *1984* de George Orwell, décrivant la dérive de la surveillance de masse par un Etat totalitaire, *Black Mirror* de Charlie Brooker, explorant les faiblesses humaines exacerbées par les technologies de l'information actuelles et futures, *Ghost in the Shell* de Masamune Shirow, projetant dans un monde cyberpunk, notre relation avec le monde du numérique, dont les frontières deviennent de plus en plus floues, ou encore *Person Of Interest* de Jonathan Nolan où les premières intelligences artificielles fortes sont créées et développent à l'image de leurs créateurs, leur emprise sur la population des Etats-Unis du Patriot Act.

Comme nous, rédacteurs de ce document, sommes pragmatiques et avant tout ne voulons pas nous morfondre dans des idées pessimistes, nous pensons qu'il existe une utilisation utile à l'Homme, comme les centrales nucléaires pour l'exploitation de la fusion, et qu'il ne faut donc pas stopper le progrès à cause de nos peurs, mais apprendre à les maîtriser ainsi que cette nouvelle technologie.
Nous nous plaçons donc dans le camp des optimistes, des visionnaires. Nous tenons à préciser que nous vous laissons seuls juges de l'honnêteté de notre position sur le sujet.

Nous vous proposons donc avec ce document, de découvrir si les craintes à propos de cette technologie qu'est le Big Data sont fondées, s'il y aura un incident digne d'un apocalypse nucléaire, pour revenir sur l'analogie avec la fusion nucléaire, ou s'il y aura une nouvelle forme d'esclavage moderne. Au contraire, il est possible que notre société pourra être optimisée. Enfin, elle atteindra la perfection d'une utopie ou le consommateur sera complètement satisfait, que les entrepreneurs et investisseurs soient sur d'avoir un commerce pérenne, que l'Etat pourra garantir une sécurité optimale et que les citoyens soient libres, s'épanouissent et respirent le bon air frais.

Plusieurs questions se posent donc auxquelles nous allons y apporter nos réponses, en essayant d'être aussi exacte que le permet notre méthode et nos moyens. Notre méthode de recherche consiste à récolter des informations dans les publications, émissions et conférences d'experts et de scientifiques sur la période s'étendant du début de l'engouement pour le Big Data, c'est à dire 2012, à aujourd'hui. Nous

procédons à la vérification des sources et de la crédibilité à accorder à leurs auteurs. Nous croisons les informations afin d'en faire ressortir les plus fortes tendances ou le consensus sur ces questions à propos du Big Data et les confronter à notre point de vue.

- Qu'en est-il du Big Data à l'heure actuelle ?
- Quels secteurs d'activités sont ou seront liés au Big Data ?
- L'intrusion dans notre vie privée au quotidien est-elle si néfaste ?
- Quel est le positionnement des gouvernements ?
- Quels sont les bénéfices d'une surveillance constante par des entreprises privées ?

Toutes ces questions nous mèneront à notre problématique :

- Comment peut-on améliorer la vie du consommateur ?

Nous allons dans la première partie de ce document parler de la technique. Cette première partie regroupe donc un état de l'art que nous allons établir avant de voir les connexions et interaction avec des autres domaines de l'informatique. Ces domaines, comme l'Internet of Things (IoT) et la Domotique, ont eux aussi suscité un fort engouement au cours de la dernière décennie.

La deuxième partie de ce document abordera les impacts du Big Data. Différents périmètres seront abordés tel que les bénéfices au quotidien, le respect de la vie privée, les réactions psychiques sur la population, la législation, l'environnement et le commerce.

A travers ce plan, nous pourrons au fur et à mesure apporter notre réponse aux questions évoquées ci-dessus. Ces réponses vont servir de pièces pour compléter le puzzle qu'est la problématique, afin de pouvoir conclure sur celle-ci.

I - TECHNIQUE

1.1- L'état de l'art du Big Data

1.1.1 - L'escalade de la puissance

Le Big Data veut littératéralement dire grande donnée, son exploitation est impossible par l'homme car si nous nous comparons avec une machine, la quantité de données dépasse notre capacité cérébrale estimée entre quelques téraoctets à quelques pétaoctets selon les études, la vitesse d'une machine pour des opérations logiques nous est inaccessible, entre 12 000 et 80 000 pensées par jour estimées pour un humain contre 170 x 10^{12} opérations basiques telle une addition (FLOPS) pour un supercalculateur en 2016 (Deep Learning System de NVidia).

La solution si nous ne pouvons nous augmenter serait donc d'accroître le nombre d'humains. Notre capacité à travailler en parallèle et donc nous répartir en équipe tout en restant coordonné est limitée même sur des tâches simples.
Sans évoquer le sentiment d'implication qu'il est recommandé d'avoir afin de réaliser un travail efficace. Notion de sentiments que n'a pas une machine.
Nous citerons la méthode Agile, même si elle ne fait pas l'unanimité dans le monde de la gestion de projet, qui conseille de limiter le nombre de personne dans une équipe à 7.

Le traitement de données nécessite donc assez rapidement de disposer d'outils d'un niveau technologique assez poussé; Outils permettant entre autres de stocker les données, les transférer à l'unité de calcul qui doit avoir la puissance de calcul nécessaire. Parcourons brièvement l'histoire du premier usage des données, c'est à dire la statistique avec un parallèle des outils informatiques.

Dans les années 1950, la statistique était faite avec quelques centaines d'individus et quelques variables, recueillis dans un laboratoire selon un protocole strict pour une étude scientifique.

Niveau informatique, les machines étaient encore dans les prémices de ce que l'on appellerait ordinateur.
Le vocabulaire de l'époque était "cartes perforées", "lampes à vide", "bande magnétiques".
Le nombre de mots en mémoire était alors dans les milliers.

La première miniaturisation permettait de le faire tenir dans une pièce et non dans un immeuble.

Figure 1 : Photographie de l'IBM 650, premier ordinateur produit en série (1951)

Et le prix d'une de ces machines, comme l'IBM 650, valait environ 150 000 dollars avec un coût de location de 3 000 dollars par mois, limitant le nombre d'entreprises et de laboratoires pouvant en faire usage. Croiser les données à la main restait plus rentable.

Dans les année 1960-1980, l'analyse des données se faisait avec quelques dizaines de milliers d'individus et quelques dizaines de variables, recueillis de façon rigoureuse pour une enquête précise.

Au cours de ces deux décennies, après les transistors, les premiers circuits imprimés furent inventés et diverses techniques pour concentrer un maximum de ces premiers sur les seconds. On observe alors une forte progression de la puissance de calcul, suivant les lois de Moore de 1973 à 2004.

Dans les années 1980, les premiers systèmes d'informations rentables apparaissent et sont assez puissants pour être exploités pour de la statistique.

Entre 1980 et 2000, le Data Mining se fait avec plusieurs millions d'individus et plusieurs centaines de variables, recueillis dans le système d'information des entreprises pour de l'aide à la décision.

Date	**Gamme**	**Vitesse de calcul / Vitesse processeur**	**Capacité de stockage**	**RAM maximum**	**Prix ($)**
1953	IBM 650	Addition : 1,63 ms Multiplication : 12,96 ms Division : 16,90 ms	2000 mots	N/A	150 000
1980	HP Model 85	613 kHz	210 kB	32 ko	3 250
1994	Power Macintosh 6100	6 MHz	500 MB	72 Mo	1 700
2004	HP Pavilion	2.1 GHz	120 GB	2 Go	1 500
2014	Asus ROG	4.5 GHz	2 TB	128 Go	1 700

Figure 2 : Tableau comparatif de gammes d'ordinateurs de différentes décennies

L'évolution s'accélère à partir des années 2010, le Big Data arrive avec plusieurs centaines de millions d'individus et plusieurs milliers de variables, de tous types, recueillis dans les entreprises, les systèmes, Internet, pour de l'aide à la décision, de nouveaux services.

Or, depuis l'année 2004, le problème de dissipation thermique fait que la puissance d'un processeur n'augmente plus suivant la loi de Moore. Pourtant les besoins sont grands, dans l'industrie, notamment la CAO (Conception Assistée par Ordinateur), du décisionnel, des sites à plusieurs centaines de visiteurs par jours. Afin de continuer de progresser en puissance de calcul et de s'adapter à ces besoins toujours plus grands, la solution mise en place a été de mettre en parallèle plusieurs processeurs, puis à leur tour de mettre des serveurs en parallèle.

1.1.2 - Quelques estimations

- 1,8 zettaoctets : données stockées par toute l'humanité jusqu'en 2011
- 40 zettaoctets en 2020
- Moins d'1% de ces données analysées
- Moins de 20% sont protégées

Sources : Digital Universe study of International Data Corporation, décembre 2012

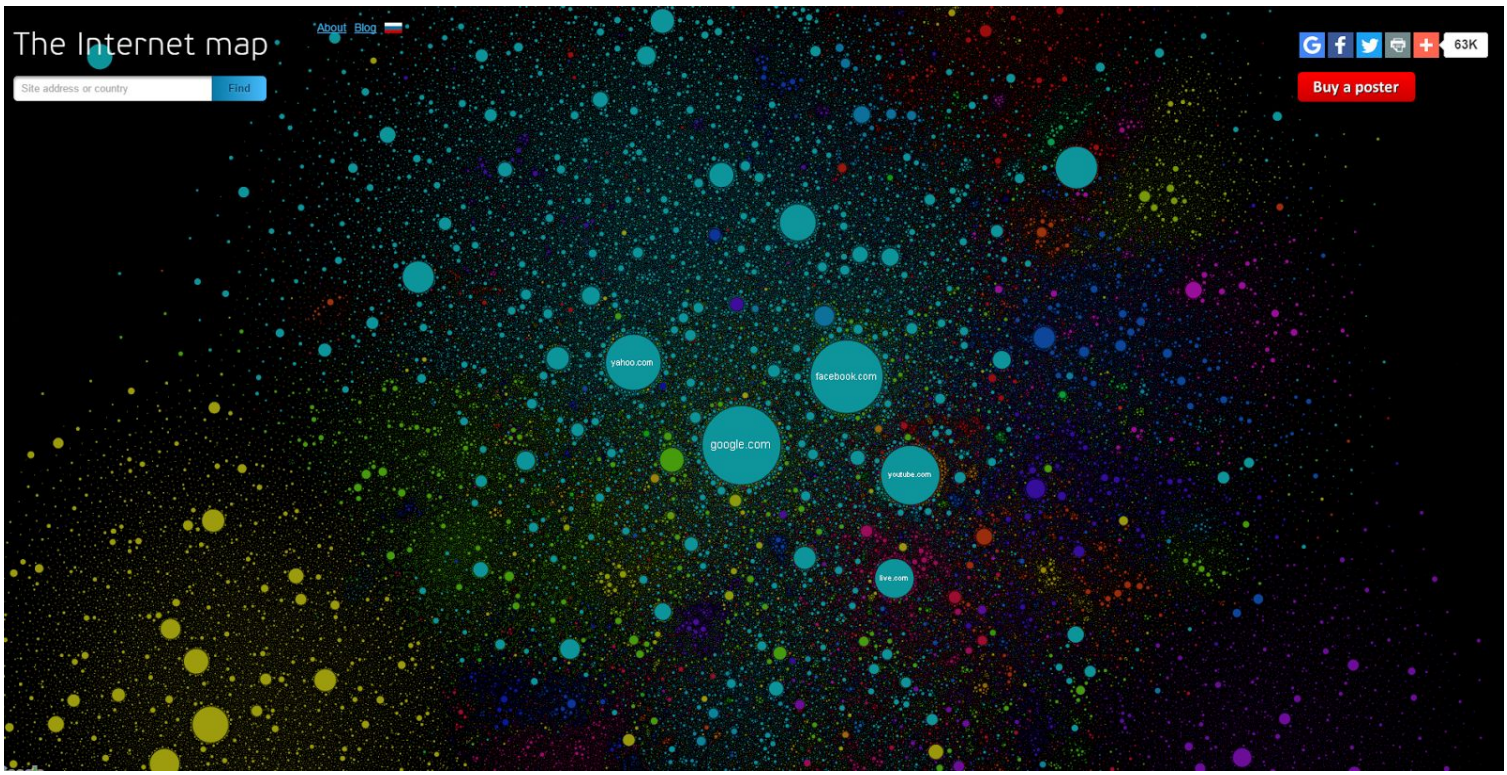

Figure 3 : Cartographie du Big Data d'Internet

1.1.3 - Caractérisation des données Big Data : les 3 "V"

1.1.3.1 - Volume

L'ordre de grandeur de ces volumes est le pétaoctet (10 puissance 15 octets).
L'augmentation du volume vient de la hausse :

- du nombre de personnes observées (plus nombreux ou à un niveau plus fin)
- de la cadence d'analyse et de stockage des informations
- du nombre d'attributs observées

Cet augmentation vient aussi de l'observation de récentes données, arrivant surtout du Web : pages référencées, recherches effectuées, hypothétiquement avec des informations de géolocalisation.
Cette particularité est éventuellement la plus visible et la plus incroyable, mais elle n'est pas la plus nouvelle (grande distribution, banque, téléphonie).

1.1.3.2 - Variété

Ces informations sont de natures très diverses : numériques, logs internet, textes, sons, images, données fonctionnelles...
Cette disparité rend compliqué l'usage des bases de données usuelles et demande une disparité/variété de méthodes (Text Mining, Web Mining...)

1.1.3.3 - Vitesse

Ces informations parviennent de sources où elles sont mises à jour dans les plus brefs délais, parfois en temps réel, et doivent la plupart du temps être traitées aussi rapidement.
Le choix du consommateur sur le Web se fait vite car il suffit d'un clic pour changer de site, aussi faut-il immédiatement lui faire la meilleure proposition commerciale.
La détection de la fraude par carte bancaire doit évidemment aussi se faire de manière instantanée.
Dans certains cas, ce n'est pas seulement l'application du modèle, mais sa mise à jour qui se fait en temps réel ou du moins très fréquemment.

1.1.4 Les outils et architectures

Pour exploiter ces données respectant les 3V évoqués plus tôt, il faut d'une part pouvoir les stocker, tâche que le cloud accomplit parfaitement. On trouve comme solution cloud Amazon avec Amazon Web Services, Microsoft avec Azure et Google avec Google Cloud. D'autre part, il est nécessaire d'avoir une technologie de traitement. Chaque technologie de traitement possède et est principalement distinguée par un type d'architecture et de modèle de données.

Trois types d'architectures existent. La différence réside entre la séparation ou non des processeurs, de la mémoire vive et des disques durs.

La première est la machine à mémoire partagée, *shared memory computer* en anglais, sur la même machine plusieurs processeurs partagent la même mémoire vive et les mêmes disques durs et un unique système d'exploitation. Cette architecture est en général plus coûteuse à faire évoluer et convient pour les application à écritures intensives.

La seconde est le cluster avec disques partagés, *shared disk cluster* en anglais, comme son nom l'indique chaque processeur à sa propre mémoire vive et son propre système d'exploitation et seul les disques sont partagés. Cette architecture est utilisé pour des applications à lectures/écritures intensives. Ce type de cluster s'adapte rapidement aux charges de travail, permet une haute disponibilité mais est plus coûteux que le cluster sans partage à faire évoluer.

La troisième et dernière est le cluster sans partage, *shared nothing cluster* en anglais. Tout est séparé, cela permet une évolution relativement peu coûteuse car le seul lien est la connexion réseau et le matériel utilisé est celui de serveurs classiques donc moins coûteux. Utilisé pour les application nécessitant des lectures intensives. Le principal défaut est la complexité d'appliquer des mise à jour entre tous les serveurs.

Le but d'avoir ces différentes architectures, est d'offrir plusieurs possibilités pour un concepteur de solution Big Data et choisir celle qui réponds le mieux au problème. Ce principe ne se limite pas qu'au Big Data.

Afin de gagner en vitesse, les bases de données pour le Big Data laissent de côté le modèle relationnel classique, et utilisent le NoSQL (not only SQL). Différents modèles de données existent, parmis les plus répandus nous pouvons citer :

- Modèle clé-valeur
- Modèle relation
- Modèle graphe
- Modèle document
- Modèle colonnes

Autre point clé dans le choix d'une base de données pour le Big Data est la position de la technologie par rapport au théorème CAP. Le théorème énoncé par Brewer en 2002 déclare que : "*it is impossible for a distributed computer system to simultaneously provide more than two out of three of the following guarantees:*

Consistency, Availability and Partition tolerance". Il est impossible pour un système distribué de garantir simultanément plus de 2 des 3 points suivants : Cohérence, Disponibilité et Tolérance au partitionnement :

- **Cohérence** (ou consistance des informations) : l'ensemble des nœuds du système voient précisément les mêmes données au même moment.

- **Disponibilité** : assurance que la majorité des requêtes obtiennent une réponse.

- **Tolérance au partitionnement** : aucun incident, différent de la coupure du réseau, ne doit empêcher le système d'apporter une réponse convenablement ou qu'en cas de morcellement en sous-réseaux, chacun doit parvenir à fonctionner de façon autonome.

Le diagramme suivant répartit les différentes bases de données (Mysql et NoSQL) en fonction des 2 points qu'elles garantissent. La couleur permet de savoir quel modèle de données elles utilisent.

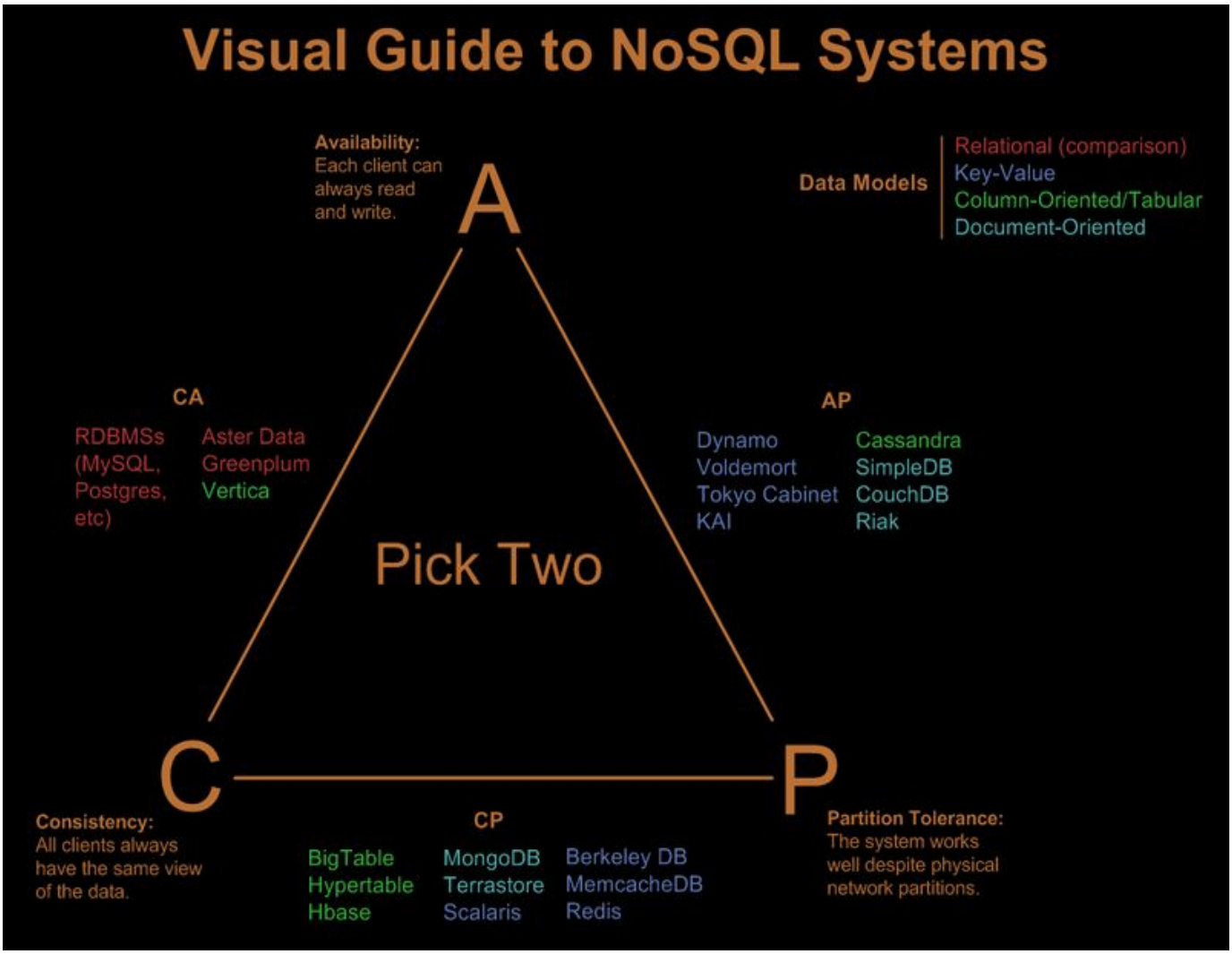

Figure 4 : Guide Visuel des systèmes NoSQL - Nathan Hurst

Parmis les principaux éditeurs de technologies Big Data ont retrouve sans surprise Google avec BigTable, Amazon avec Dynamo et SimpleDB mais aussi de l'Open Source avec Hadoop, Neo4J, Cassandra d'Apache entre autres.

Actuellement la grande majorité des outils pour le Big Data est de type NoSQL, mais de nouvelles technologies émergent comme le NewSQL permettant de combiner les propriété ACID (atomicité, cohérence, isolation et durabilité) des bases de données relationnelles et la même puissance de traitement que les bases de données NoSQL.

1.1.5 - Autres enjeux informatiques

Les défis du Big Data ne sont pas seulement de puissance de traitement et de type de stockage (cloud).

La sécurité des données est cruciale quand autant de données personnelles sont stockées, elles doivent être à l'abri de tout vol, perte ou détérioration. La protection contre le piratage et les cyberattaques doit être renforcée.
La sécurité est très importante pour des raisons financières (ces données ont de la valeur), et de protection des données personnelles, donc de confiance.
La qualité est toujours capitale et d'autant plus sensible avec les Big Data que :

- de nombreuses données proviennent de sources externes et hétérogènes.
- certaines données, typiquement celles collectées sur Internet et les réseaux sociaux, comportent une grande part de bruit.

La QoD (Quality of Data, en français qualité des données) est un autre des enjeux cruciaux du Big Data et de tout système d'information en général. Le problème prend une toute autre échelle avec le Big Data cependant les même préconisations sont faites que pour les bases de données classiques, c'est-à-dire, procéder à une validation rigoureuse des données, les standardiser, détecter les erreurs, avoir un moyen de faire du rattrapage, pouvoir les enrichir et les résoudre.

1.1.6 - Le Big Data ne se résume pas à Internet

Les entreprises ont dans leurs systèmes d'informations des quantités de données inexploitées.
Elles peuvent explorer d'autres méthodes pour collecter de la donnée avant d'avoir à requérir à celles de Facebook, Twitter ou de Google.
Parmis les sources de données existantes, elles peuvent exploiter les données de logs de leurs sites Web (Web Mining).
Elles peuvent également analyser les courriers et courriels (de réclamation, dans le cas d'une entreprise d'e-commerce) de leurs clients et aller explorer les forums de discussion (Text Mining).
Enfin, elles peuvent consommer les données d'API publiques et d'Open Data (Moissonnage).

Un des cas où énormément de données sont générées mais pas ou peu exploitées est La Poste. En effet, La Poste achemine chaque année des milliards de plis, dont elle pourrait analyser l'adresse du destinataire, le lieu d'envoi, la date, l'affranchissement, le type (journal, facture, publicité...), le poids, la taille dans le cas d'un colis. Autant de données pouvant servir à enrichir des bases de données géo-démographiques.

Que l'informatique soit la base de leur business ou non, beaucoup d'entreprises négligent par priorisation de faibles coûts l'introduction d'outils d'introspection dans leurs systèmes d'information.
Pour la grande majorité d'entres-elles, le recueil de données se traduit par un fichier client, un fichier fournisseur, un fichier catalogue de produit et d'entrepôt pour le commerce et purement informatique les fichiers de logs, ces derniers uniquement pour surveiller les incidents applicatifs ou serveurs. Parfois les dates de connexion des clients sont aussi enregistrées en base.

On constate que les systèmes d'informations sont victimes d'une rationalisation de leur utilisation dans le seul but de satisfaire les besoins métiers, sans prendre en compte que n'importe quelle action, informatique ou non, génère de la donnée et qu'elle peut être fortement valorisée. Ces données doivent ensuite servir de boussole dans l'évolution d'un business model ou l'organisation des processus de l'entreprise ou des services publics.

Une prise de conscience a eu lieu de la part des géants de l'informatiques, des départements et communes françaises dans le cadre des lois de l'Open Data et de certaines start-ups spécialisées. Pour le reste des entreprises, leurs systèmes d'informations n'ont pas encore ou que très peu intégré le Big Data.

Le temps et l'argent sont, nous dirons toujours, les raisons indirectes de la non adoption d'une nouvelle façon de voir l'informatique.

D'une part, nous avons l'existant applicatif qui n'inclut pas l'émission de données, il faut donc de la main d'oeuvre pour modifier ou remplacer les applications. Demandant des fonds et du temps, sans impact bénéfique direct sur le chiffre d'affaire. Et le coût est d'autant plus grand qu'il y a d'application héritées. Sans compter la révision nécessaire des contrats de maintenance.

De l'autre, nous avons le matériel, dimensionné pour des besoins précis, capable de remplir sa fonction pour plusieurs années.

Au final, comme toute technologie, elle se démocratise avec le temps. Temps nécessaire pour que le changement ne soit plus une option de confort et que la nouvelle technologie y soit intégrée à moindre coût et que les experts la maîtrisent.
La solution pour accélérer l'adoption du Big Data et de produire des données à faible coût pourrait être l'Internet des Objets (IoT, ou Internet of Things en anglais).

1.2 - La possible fusion avec les mondes IoT et Domotique connectés

Le monde dans lequel nous vivons aujourd'hui est très différent de ce qu'il était il y a cinq ou dix ans. L'analyse de données enrichie et les avancées technologiques innovantes ont transformé notre façon de penser et ont permis aux entreprises de réimaginer comment elle se connectent avec le consommateur final.
L'Internet des Objets (IoT) a évolué grâce au Big Data, permettant aux entreprises de fournir un service prescriptif personnalisé, disponible sur n'importe quel appareil, n'importe où et n'importe quand à un particulier.
Ce nouveau monde numérique connecté offre de nombreuses possibilités pour les entreprises de fournir des services enrichis, des soins de santé, des produits et finalement, améliorer la durabilité pour un environnement plus vert.

Comme l'opposition qu'il y a eu dans le début des années 2000, entre téléphone et ordinateur pour savoir ce que serait un smartphone. Il y a opposition entre appareils classiques non optimisés car pouvant exécuter n'importe quelle tâche comme une montre connectée et un appareil ne servant qu'à accomplir une tâche particulière dont son architecture est optimisée comme un bracelet connecté.

Les données personnelles volent partout tout autour de nous, les heures de sommeil accumulées, les calories perdues jusqu'au bureau, les détails de nos virées en vélo, notre position GPS, jusqu'à notre carte grise. Une mine de renseignement, toutes ces données une fois récupérées et bien archivées, c'est ce que l'on appel le Big Data.
Ces informations, nous en sommes consciemment les principaux émetteurs.
Elles documentent de surcroît la moindre activité et ce n'est qu'un début.
Au cours de l'année 2015, le Big Data fut alimenté grâce à environ 15 milliards d'appareils connectés à Internet et ce chiffre pourrait atteindre jusqu'à 50 milliards d'ici 2020.
IBM, numéro un du domaine, a investi 24 milliards de billets verts depuis une décennie uniquement sur l'analyse des données.
Quel est dès lors l'enjeu de cette boulimie de données ?

Prenons l'exemple des maisons intelligentes, des logements pourvues de détecteurs qui analysent notre permanent :
Allstate, une compagnie d'assurance américaine, offre déjà un quart de remises à la totalité des propriétaires qui en équiperaient leur habitation.
Autre domaine, les automobiles, des boîtiers installés dans les automobiles permettent aux automobilistes de jouer avec leurs propres statistiques, mais l'assureur les analyse pour affiner les bonus et malus, et donc les contrats types.
Dans l'hexagone, 70% des conducteurs seraient prêts à équiper leur voiture de capteurs en contrepartie de réduction sur leurs contrats.
La santé, Axa propose à certains de ses assurés français un bijou/bracelet connecté (rythme cardiaque, calories brûlées, oxygène dans le sang) qui stocke tout un flot d'informations. De fait, l'assureur récompense l'assuré s'il repère un mode de vie sain, notamment jusqu'à 100 euros si l'assuré marche plus de 10 000 pas par jour. Parce qu'un assuré en bonne forme, c'est un compagnie d'assurance rentable.

1.2.1 - L'explosion de la production de données

Le Big Data a vu le jour pour faire face à l'explosion des données. Inventé par les géants du Web, comme Amazon ou Google, ces solutions sont destinées à offrir un accès en temps réel à des bases de données géantes.

L'horizon de l'Internet of Things et du Big Data est celui d'un monde encore plus densément connecté. Il relie les individus, les données et les objets dans un environnement numérique dorénavant global.
L'Organisation des Nations-Unies (ONU) indique que plus de données ont été créées au cours de l'année 2011 que dans toute l'histoire de l'humanité et, certaines sources indiquent qu'entre 30 à 212 milliards d'objets pourraient être connectés avant la fin de cette décennie. Cette liaison, qualifiée d'ubiquitaire, soulève à présent tant d'inquiétudes que de promesses d'opportunités sociétales et économiques.

Les objets connectés, ainsi que le Big Data représentent à eux seuls un imposant relais du développement économique d'après une multitude d'études. Ils ouvrent la capacité de connecter les individus ou les objets de façon plus pertinente, de fournir au bon moment la bonne information au bon destinataire, mais également d'en faire ressortir les informations clés à la prise de décision.

Différentes sources de données :

- Données sociodémographiques et signalétiques
- Données comportementales (utilisation du téléphone, du véhicule, de la carte bancaire...)
- Données des systèmes de gestion de la relation client (CRM ou Client Relation Management en anglais) (fidélisation, carte de fidélité, contact avec un service client...)
- Données externes provenant des administrations (Open Data) ou des mégabases de données provenant du secteur privé
- Données remontées par les capteurs routiers, climatiques, industriels, puces RFID, tags NFC, objets connectés (voitures, domotique, compteurs électriques...)
- Données de localisation par GPS ou par adresse IP
- Données de tracking sur le Web (mots-clés recherchés, sites visités...)
- Contenu partagé sur les réseaux sociaux (photos, vidéos, blogs...)
- Opinions et avis exprimées à travers les blogs et les réseaux sociaux

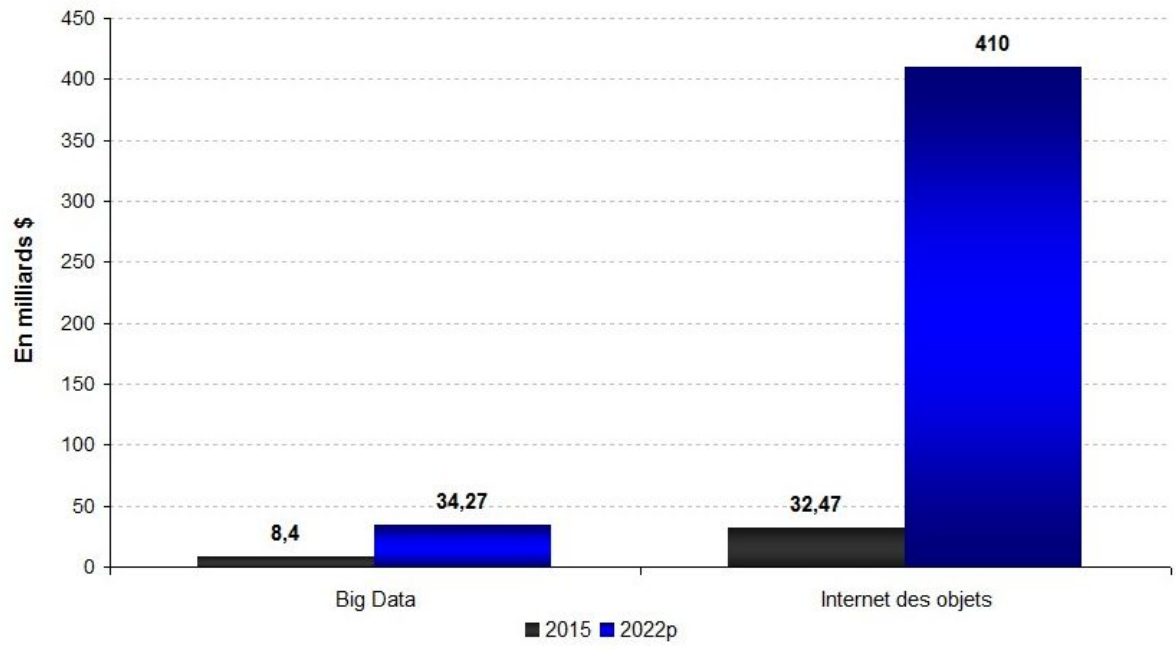

Figure 5 : Marché mondial du Big Data et de l'Internet des Objets en santé (2015 et 2022)

p = projections
Sources : ReseachAndMarket, janvier 2016, Market Research Future, juillet 2016, MarketsAndMarkets, octobre 2015, BI Intelligence Estimates, 2015, Grand View Research, mai 2016.

Autre secteur, en pleine croissance, et ayant connu un très fort engouement depuis le début des années 2010, le domaine de la recherche opérationnelle et de l'Intelligence Artificielle pourrait s'avérer être la clé de l'exploitation à son maximum de toutes ces données.

1.3 - Le traitement des données par l'Intelligence Artificielle

Les données seules ne sont pas importantes. Ce qui compte est comment la donnée est gérée, analysée et utilisée.

Mais venons-en à la partie la plus intéressante qui est la partie exploitation intelligente des données : le Machine Learning.
Le Machine Learning consiste en un ensemble de modèles qui permettent d'apprendre à partir de données sans que les règles inférées par les données soient explicitement programmées.
En fait dans les média de vulgarisation, lorsqu'on parle de Big Data, Smart Data, ... se cache derrière du Machine Learning.
Pour donner quelques exemples d'apprentissage supervisé, une branche du Machine Learning, nous citerons par exemple : la régression linéaire, la régression logistique, les classifieurs naïfs Bayésien, le Support Vector Machine et les réseaux neuronaux, ...

Le Machine Learning permet de prédire le future proche grâce à de l'analytics sur données et des scénarios évolutifs. Le Machine Learning consiste à programmer par l'exemple.

Exemples d'outils utilisant le Machine Learning :

- Churn analysis
- Ad targeting
- Forecasting
- Fraud detection
- Anomaly detection
- Image detection & classification
- Equipment monitoring
- Spam filtering (un des premiers algorithme de Machine Learning)

Puis pour bien exploiter le résultat final, ou bien plus en amont, pour mieux comprendre les données de manière à en dériver les "features" qui elles-mêmes vont servir à alimenter le Machine Learning, un Data Viz sera nécessaire pour permettre la visualisation des données.

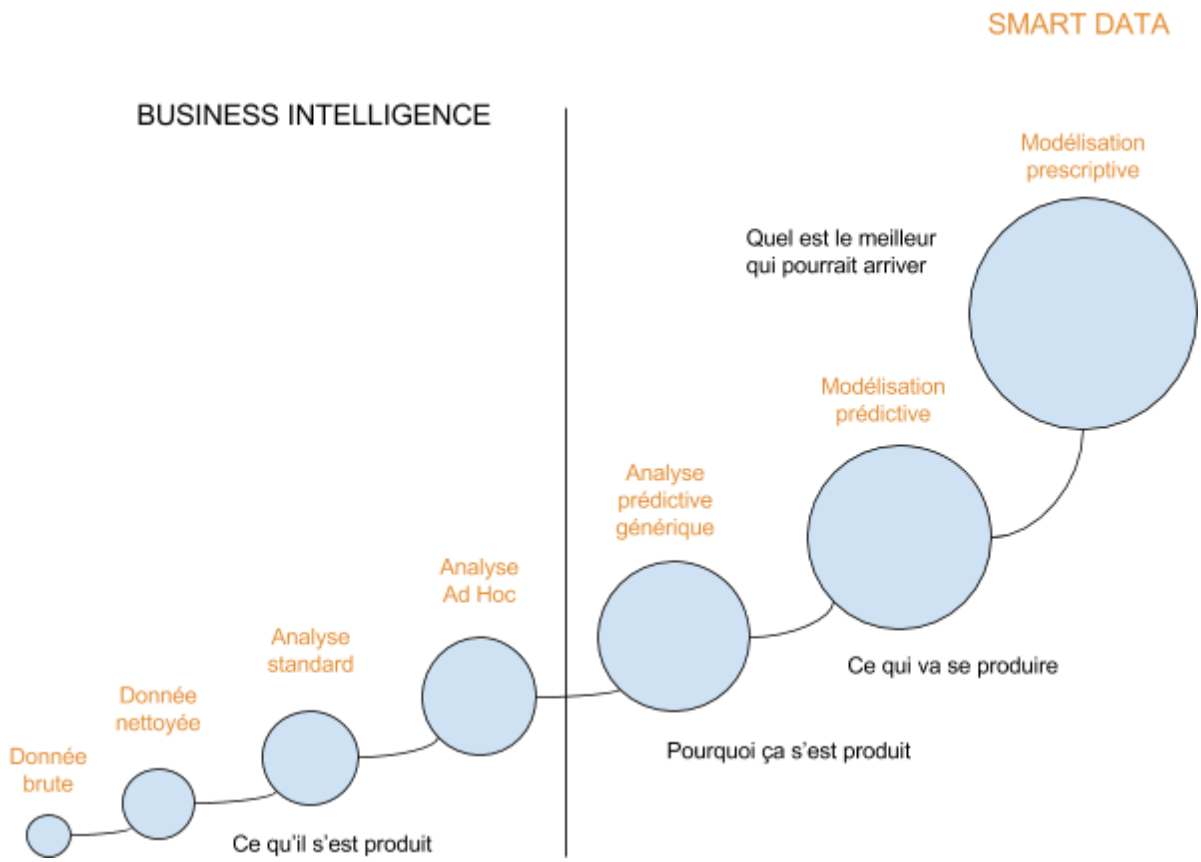

Figure 6 : Processus du Big Data Analytics

Nous allons maintenant voir - sans être exhaustif - quelques domaines où le Big Data Analytics apporte une forte valeur métier : des nouveaux produits, des nouveaux services, des avantages concurrentiels ...
Derrière le terme Big Data Analytics se cachent deux concepts différents. Tout d'abord la partie Big Data pure correspondant aux traitements et au stockage massivement parallèle avec scalabilité et tolérance aux fautes. Et d'autre part la partie exploitation intelligence de ces données que l'on pourrait qualifier de Smart Data avec derrière en particulier le Machine Learning.

Nous allons dans un premier temps aborder les domaines transverses de l'OCR (Reconnaissance Optique de Caractères, ou Optical Character Recognition en anglais), de la reconnaissance du visage et de la parole, dans lesquels le Machine Learning est particulièrement efficient. La reconnaissance optique des caractères - l'OCR - est depuis longtemps l'apanage des réseaux neuronaux et fonctionne très bien.
Depuis quelques années, la reconnaissance faciale a fait beaucoup de progrès avec les approches d'apprentissage profond : le Deep Learning.

Par exemple, Deep Face, le nouveau système de reconnaissance faciale de Facebook obtient des performances similaires à l'être humain avec une fiabilité de l'ordre de 97% et avec une proportion de faux positifs très faible.
La reconnaissance automatique de la parole a également fait beaucoup de progrès et fait désormais partie du quotidien. Il suffit par exemple de citer Siri, Cortana ou Ok Google.

Un autre domaine de prédilection du Big Data Analytics est l'optimisation des campagnes marketing et du processus de conversion. Pour ce faire, le Big Data Analytics permet une meilleure compréhension des besoins et des comportements clients, permet une meilleure segmentation de la clientèle, permet d'identifier les clients susceptibles d'acheter des services de plus haut niveau, et permet de cibler les campagnes marketing et les publicités.

Un autre domaine intéressant du Big Data Analytics est l'aide à la prise de décision. Des systèmes comme Watson permettent d'aider à prendre des décisions, voire de prendre des décisions dans des domaines aussi variés que la finance et la médecine en répondant à des questions formulées en langage naturel d'un humain.

La sécurité est un autre domaine où le Big Data Analytics apporte de la valeur. C'est par exemple le cas à Paypal où le Deep Learning a été mis en oeuvre avec succès pour la détection des suspicions de fraude. Le Machine Learning est également utilisé pour détecter les suspicions d'attaque en particulier en analysant les logs. Certaines polices aux États-Unis et en particulier à Los Angeles utilisent le Big Data pour prédire quand et où des crimes ou des cambriolages peuvent survenir. Ce n'est pas encore Minority Report - et heureusement - mais cela a quand même entraîné une réduction de 33% de cambriolages et de 21% de crimes.

Le Big Data Analytics va également permettre d'optimiser les processus industriels et les services, comme l'optimisation de la logistique et de la Supply Chain et l'optimisation de la rentabilité de la production agricole en conseillant les fermiers sur les dates de plantation optimales sur la base de capteurs, de conditions météo et de rentabilité passée.

Le Big Data Analytics est également utile dans le monde de la santé, il permet de trouver de nouvelles molécules, de prédire des maladies sur la base des analyses ADN. Il permet également de prédire les chances de survie des patients dans le cas des cancers, de la transplantation d'organes en utilisant des réseaux Bayésiens.
Mais on peut également envisager d'autres débouchés innovants comme par exemple l'identification qu'un patient a du diabète à partir de photographies de la rétine.

Le Big Data Analytics a également envahie le monde du sport. Il y a de plus en plus d'athlètes ou d'équipes - en foot par exemple - qui optimisent leurs performances en analysant des données.
Les transactions à cadence élevée sont l'exécution à haute vitesse de transactions financières réalisées par des calculs informatiques qui ressemblent à du "big data" et qui ont la faculté d'exécuter des manoeuvres sur les marchés financiers en quelques microsecondes.
La recherche quant à elle, produit et consomme un volume considérable de données. Le LHC au CERN par exemple produit annuellement 30 petabytes de données. C'est également le cas de l'astrophysique et des simulations climatiques.
Dans le monde des Smarts Cities, le Big Data va permettre d'optimiser la consommation énergétique et de réguler le trafic.
Même dans le monde de la politique, le Big Data est utilisé ! Il est de notoriété publique que l'ancien président Obama a utilisé le Big Data dans sa compagne afin de cibler le démarchage des personnes les plus susceptibles de changer d'avis.

De nombreuses applications du Machine Learning impliquent des interactions humaines. L'humain peut apporter sa contribution à un algorithme d'apprentissage, y compris l'apport sous la forme de tags, d'événement, de corrections, de classements ou d'évaluations. Et ils pourront ainsi s'attendre à des résultats de l'algorithme sous la forme de rétroaction, de prédictions ou d'événements.
Bien que l'humain soit une partie intégrante du processus d'apprentissage, les systèmes d'apprentissage traditionnels utilisés dans ces contextes sont agnostiques au fait que les entrées/sorties proviennent de/pour des humains.
Cependant, une communauté croissante de chercheurs au croisement entre le Machine Learning et l'interaction homme-machine font de l'interaction avec l'humain le coeur du développement des systèmes d'apprentissage automatique. Ces systèmes sont appelées Interactive Machine Learning, qui est défini comme étant du Machine Learning avec une dimension humaine dans la boucle d'apprentissage, tout en observant les résultats de l'apprentissage et en fournissant les informations destinées à améliorer les résultats d'apprentissage.
Leur méthode d'apprentissage comprend l'application de principes d'Interaction Design aux systèmes de Machine Learning, en utilisant des tests sur des sujets humains afin d'évaluer les systèmes de Machine Learning et d'en découler de nouvelles méthodes.

Figure 7 : Le Big Data dans la santé : réalités et perspectives en France

Au final, le Big Data peut concerner tous les secteurs d'activités, son adoption est pour l'heure loin d'être généralisé mais tend à le devenir dans un futur proche. Les entreprises du Web commencent à intégrer le commerce de données dans leur business model. Un fort lien avec le monde domotique et Internet of Things existe car il permet de produire encore plus de données sur les individus et leurs environnements à bas coût. Lien plus fort que pour l'Intelligence Artificielle, qui n'est pas être encore une technologie rodée. En tout cas il est clair que pour nous, l'intelligence artificielle aura un rôle déterminant dans l'exploitation des données.

II - IMPACTS

2 500 000 000 000 000 000 octets de données (2,5 exaoctets) par jour, voici la quantité de données actuellement générées. Le point tournant de la mi-décennie que nous vivons actuellement est une capacité sans précédent de traiter ces données, de générer des idées. Ces prévisions, recueillies à partir de nouvelles techniques de traitement de données, peuvent avoir un incidence sur la façon dont nous faisons des emplettes, comment nous trouvons des emplois, des rencontres, et même comment les médecins diagnostiquent des maladies.

2.1- Impact d'une augmentation du Big Data dans notre vie quotidienne

Ces dernières années, les entreprises ont commencé à investir massivement dans de grandes données en recueillant des informations sur les consommateurs par le biais du Web, des centres d'appels et des interactions en personne.
Parmi les avantages de cette nouvelle richesse d'information, on compte l'augmentation des inscriptions, des ventes, du retour des investissements et de la satisfaction des consommateurs.

Une partie du succès du Big Data est due à sa capacité à identifier les principaux facteurs démographiques dans les consommateurs qui sont des facteurs majeurs de la conduite du comportement des consommateurs.
Au-delà de l'information démographique de base, comme l'âge, le revenu et la profession, les entreprises ont également été en mesure d'identifier les croyances des consommateurs, les attitudes et les sentiments envers les produits, et comment ces changements se font au fil du temps.

En dessinant une image plus claire des consommateurs grâce au Big Data, les entreprises ont été en mesure de répondre au mieux aux besoins des différents consommateurs.
Des études et des enquêtes récentes menées par Forbes Insights et Rocket Fuel ont réussi à mettre en évidence de nombreuses autres façons dont le Big Data affecte le comportement des consommateurs.
85% des agences américaines et des dirigeants de marques disent que le Big Data a conduit plus de la moitié des initiatives de marketing visant à accroître les connaissances sur les comportements des consommateurs.

Avant 1993, la manière dont les gens trouvaient du travail, la gestion de la bourse, la rencontre amoureuse étaient totalement différentes.
Aujourd'hui, grâce au Big Data, on peut se demander, comment gérer notre santé ? comment communiquer avec mes amis ? comment rester en contact avec mes amis ? comment investir mon argent ? comment trouver de nouveaux clients pour mon entreprise ?
Globalement, notre vie au quotidien est touchée par cette révolution technologique. Entre 1970 et 1993 si nous utilisions toutes les données commerciales que nous pouvions trouver pour résoudre n'importe quel type de problème dans la vie personnelle, nous devrions traiter environ un milliard de gigabytes de données.

La naissance d'Internet a considérablement accéléré la façon dont les données sont recueillies, stockées, et disponibles pour l'utilisation. Aujourd'hui, il y a plus de cinq milliards de gigabytes de données par jour qui sont structurées et disponibles à l'utilisation. Il faut savoir que ces cinq milliards de gigabytes de données par jour sont construits à partir de 25 milliards de gigabytes de données recueillies et stockées tous les jours, car la technologie que nous possédons actuellement n'est capable de structurer seulement environ cinq milliards de gigabytes de données.

Nous savons tous que posséder une quantité importante de données, sans en avoir la capacité à la traiter est sans trop d'utilité.
Une autre chose remarquable sur cet écosystème est qu'à l'instant à laquelle la donnée existe, les technologies arrivent à la traiter, la structurer, la nettoyer, (c'est-à-dire de la rendre propre pour la consommation dans les systèmes décisionnels) et donc d'en extraire des décisions et idées à partir de cette dernière.

De cette manière, la façon dont nos vies vont changer est qu'aujourd'hui nous sommes amenés à nous demander "quelles décisions dois-je prendre ?", "de quelles données existantes pourrais-je me servir ?". Sous peu de temps, les décisions que nous devrons prendre nous serons apportées par ces technologies. Nous n'avons plus besoin de savoir exactement ce que nous cherchons.
Prenons deux exemples concrets à l'heure actuelle, les deux principaux instituts de lutte contre le cancer aux Etats-Unis ont rassemblé toutes les données numériques (environ quinze années de données électroniques sur le traitement du cancer sur chaque patient individuellement). Il s'agit, ici de dizaines de millions de cas.
Les meilleurs cancérologues dans ces meilleurs centres, lorsqu'ils sont faces à un nouveau cas de cancer, disent qu'ils peuvent se souvenir de six à huit cas similaires, tandis que le système est capable de mémoriser dix mille cas similaires.
De ce fait, dans ces laboratoires expérimentaux, les meilleurs cancérologue au monde, demandent au système ce que serait le meilleur traitement pour ce patient, car pour les dix mille semblables le système en connaît leur résultat.

De nos jours, la technologie peut faire le travail de recherche d'un cancérologue, et en mieux dans 85% du temps. Donc nous pouvons imaginer ce type de système dans les centres de lutte contre le cancer dans le monde, où les cancérologue n'auraient probablement pas de cas semblables.

Ceci sera notre nouvelle façon de penser, lorsque vous voudrez aller skier dans les Alpes, vous n'aurez plus besoin de regarder les trains et hôtels, mais plutôt dire au système "Je veux aller skier dans les Alpes" et le système se chargera de planifier l'itinéraire pour vous en tenant en compte l'expérience de tous les autres utilisateurs ayant effectué le même voyage que vous.

2.1.1 - Quelques exemples d'utilisations de ces données

2.1.1.1 - Transports

Détermination dynamique du prix des vols, recherche de la station-service la moins éloignée, des places libres de parking (Parkopedia), optimisation du trafic routier par géolocalisation (Waze), facturation dans les zones payantes à travers la reconnaissance optique de caractères (Optical Character Recognition) des plaques d'immatriculation...

2.1.1.2 - Marketing

La géolocalisation permet de recevoir des bons de réductions et/ou promotions sur notre smartphone au moment où nous passons à proximité d'un commerce, d'une alerte au moment où nous passons à côté d'une librairie possédant un ouvrage consulté en ligne la veille, l'analyse des préférences, des recommandations, peut-être en lien avec les données de vente, offre la possibilité de mieux prendre pour cible les consommateurs.

2.1.1.3 - Grande distribution

Croisement des tickets de caisses avec les données du programme de fidélité.

2.1.1.4 - Ressources humaines

Analyse des Curriculum Vitae enrichie par le repérage des liens noués par le candidat sur les réseaux sociaux.

2.1.1.5 - Scientifiques

Imagerie médicale, épidémiologie, génomique, astronomie, physique nucléaire, météorologie.

2.1.1.6 - Yield management

Intéresse les activités avec des capacités disponibles limitées (espaces publicitaires, tourisme, hôtellerie, transport...).
Détermine en temps réel les quantités appropriées à commercialiser, au tarif approprié, de manière à maximiser le bénéfice produit par la commercialisation.
Né dans les années 1980 dans le transport aérien.

2.1.1.7 - Informatique

Surveillance des machines et infrastructures, et repérage de dysfonctionnements ou d'incidents sécuritaires.

2.1.1.8 - Sécurité

Renseignement, vidéo-surveillance.

2.1.1.9 - Enseignement

Analyse des réseaux sociaux afin de connaître la satisfaction des élèves vis-à-vis de l'éducation mais aussi la popularité des enseignements auprès de ces derniers (MyGES).

2.1.2 - Le Big Data pour les conducteurs

Le Big Data a déjà fait un pas dans le monde de l'automobile et les données seront cruciales pour les Intelligences Artificielles entre autres. Pour preuve, le MIT (Massachusetts Institute of Technology) nous demande déjà sur le site http://moralmachine.mit.edu/hl/fr d'évaluer les situations dans lesquels un sacrifice humain devrait être fait plutôt qu'un autre sacrifice. Mais avant de devoir faire un choix aussi lourd, le Big Data a d'autres applications comme l'aide à la conduite et l'aide communautaire.

2.1.2.1 - Aide à la conduite automobile

Aide à la conduite automobile avec les systèmes :

- de guidage intelligents (localisation du véhicule, état du trafic et contraintes horaires du conducteur)
- d'aide à la conduite économique (en intégrant en plus la consommation du véhicule dans ses différentes phases de roulement : arrêt, accélération, croisière, freinage)
- d'aide à la conduite se basant sur les informations communiquées par les autres véhicules, telles que la présence d'une côte ou d'une descente permettant d'anticiper une augmentation ou une diminution du régime moteur (à l'étude dans certaines entreprises de transport routier)

2.1.2.2 - Passage d'intérêts privés à des intérêts collectifs

La collecte d'informations fournies en temps réel par les smartphones des conducteurs permet aussi de fournir des renseignements à la communauté sur le trafic routier, fluidifiant le trafic et pouvant même éviter des accidents (Waze, Google Maps...).
Aussi, des capteurs météorologiques sur les smartphones embarqués peuvent fournir des informations plus fines que celles de la météorologie nationale.

2.1.3 - Le Big Data pour les citoyens

Pendant un an (février 2012 - 2013) la ville de Toulouse a constitué une base de plus de 1,6 millions de documents provenant de blogs, de forums de discussion, de Facebook, Twitter et de divers médias issus de la presse nationale ou régionale.

L'analyse de ces documents a permis à la ville de Toulouse de détecter les attentes et les préoccupations de ses habitants, dans le but de mieux cibler ses investissements (implantation du tramway), ses travaux de rénovation, ses nouveaux services ou sa communication.
Il y a cependant un risque de dérive vers une surveillance de l'opinion et des réseaux personnels des citoyens.
Les réseaux sociaux peuvent aussi être scrutés pour anticiper les résultats d'élections.

2.1.4 - Statistiques publiques

En faisant l'analyse des mots clés sur son moteur de recherche, la firme de Mountain View a pu définir une corrélation entre plusieurs requêtes et l'avènement d'une épidémie de grippe. D'après une publication dans Nature (2009), cette corrélation a été confirmée par les organismes de la veille et sécurité sanitaire (VSS).

Une enquête très récente également parue dans Nature (2013) explique une corrélation entre les mots clés saisis sur Google et la progression des cours de bourse. Avant une baisse des indices boursiers, les actionnaires sont préoccupés et recherchent sur sur le Web des données leur permettant de choisir entre conserver ou vendre leurs titres.

2.1.4 - Aide à la prise de décision

Le projet Open Food System est un exemple d'aide à la prise de décision et espère bien révolutionner la cuisine domestique.
Le but de ces recherches d'ampleur est d'aider les clients à mieux manger et de façon équilibrée.
Le principe de cette technologie repose sur un moteur de recommandation de recettes intelligent. Concernant le produit final, il se matérialisera sous l'apparence d'un appareil de cuisson connecté. Avec du Machine Learning, l'appareil s'adaptera à ce qu'il apprend par le biais de l'utilisateur afin d'être capable de lui recommander des recettes en prenant en compte ses contraintes nutritionnelles, ses envies et ses goûts. Notamment, si l'appareil note que la personne se prépare toujours des mets peu complexes, il pourra lui proposer des mets un peu plus élaborés (difficulté croissante).
Le but recherché est d'offrir à toute personne chargée de la préparation du souper une offre de produits et services subsidiaires comportant en simultané de l'équipement ménager et des contenus (outils d'aide, applications, recettes...).

Au final, la cuisine serait capable de ressembler à un espace intelligent où les équipements électroménagers principaux (table de cuisson, lave vaisselle, réfrigérateur...) seront interconnectés entre eux. Au travers de ce moyen, la cuisine sera bientôt un nouveau terrain de jeu assurant une expérience utilisateur sans précédent...

Quelques autres exemples:

- Les données institutionnelles et publiques sont de plus en plus utilisées pour améliorer l'efficience fonctionnelle des villes : Cisco a annoncé début décembre 2015 son association avec la Startup Streetline pour s'attaquer à la gestion des places de parking à San Francisco en temps réel.
- La numérisation de la médecine serait susceptible à terme de simplifier le diagnostic des médecins et le traitement des patients, tout en améliorant les coûts.

2.2 - Impact sur le public d'une surveillance constante. Besoin de légiférer ?

Les moyens de communications modernes ont toujours été au coeur d'une tension permanente. Mais le problème avec l'exception, c'est lorsqu'elle en devient la règle.

Pour les services de renseignement, surveiller ce qui transite est un pari majeur.
Pour tous, la préservation du caractère privé de nos discussions l'est tout autant. Une tension est toujours plus importante au fil du temps. Entre 1994 et 2014, la quantité d'abonnés en téléphonie mobile dans l'hexagone est passé de 280 000 à 78,4 millions.

Il y a 4 ans, 196 milliards de SMS ont été envoyé en France, 6% supplémentaire que l'année précédente.
Additionné avec les données échangées l'année suivante par plus de 54 millions de cybernautes sur le sol Français, nous allons finir par obtenir un trésor d'informations.
Du coup, tout change : Il y a quelques années, afin de savoir si nous nous trouvions au sein d'une pièce, il était dès lors nécessaire d'y trouver nos empreintes puis d'y ajuster au moins 12 points identiques avec l'une de celles hébergées dans les bases de données policières.

À présent, à partir de quatre connexions sur quatre antennes téléphonique différentes, suffisent pour identifier une personne avec une précision à 95%.

Alors dans ce rapport de force entre hyper-communication et désir de surveillance, l'équilibre tient en une simple question : Qui surveille les surveillants ?
Et là où la loi devrait assurer cet équilibre, une notion vient régulièrement la mettre en danger : l'exception.

Au lendemain des attaques du 11 septembre 2011, le congrès américain vote le Patriot Act, une loi dite d'exception, censée aider temporairement à lutter contre cet ennemi invisible qu'est le terrorisme. 15 ans plus tard cette loi est toujours en vigueur et l'exception est devenue la règle. Au passage, l'article 215 a notamment permit de récupérer les données de communications de l'ensemble des abonnés de l'opérateur téléphonique Verizon. Face à la menace, pas de temps à perdre avec les principes démocratiques.

Il y a 3 ans, la cours de justice de l'Union Européenne invalide une directive sur la conservation des données personnelles qui précédemment était adoptée en 2006, suite aux actes terroristes de Madrid et de Londres. Trop de données stockées trop longtemps sans même que l'utilisateur soit informé en cas d'utilisation.
Une attaque du droit fondamental à la vie privée selon la justice Européenne.

L'agence de renseignement américaine (NSA) qui scrute les réseaux, se permet de remonter jusqu'à 3 degrés de relation à partir d'un individu potentiellement suspect : cela signifie qu'en fonction des gens qui sont en contact, avec les gens qui sont en contact, avec les gens qui sont en contact, avec nous, nous sommes potentiellement louches.

A partir d'une cible possédant environ une quarantaine de contacts dans le carnet d'adresse de son smartphone, 2,5 millions d'êtres humains peuvent ainsi devenir suspects. En France, trois mois à partir des actes terroristes survenus dans la capitale, le futur amendement sur le renseignement a comme objectif de moderniser les ressources des services de renseignement, là aussi afin de se battre contre le terrorisme.
L'un des articles du texte prévoit l'installation de boîtes noires pour surveiller le trafic sur Internet plus précisément pour détecter des comportements suspects, mais pour identifier un comportement anormal, donc suspect, il faut surveiller les communications privées d'une majorité considérée comme non-suspectes, c'est-à-dire "monsieur et madame tout le monde".

D'après le projet de loi, des algorithmes placés dans cette boîte noire aideront à cette surveillance, la question n'est donc plus de surveiller les surveillants, mais de surveiller des algorithmes écrits pour des surveillants. La transparence devient opaque.

Et sans garde fou solide, la normalité risque vite de devenir tout ce qui n'est pas anormal, donc ce que les algorithmes veulent bien laisser hors de leur radars.

Étant donnée qu'il y a aussi du risque en politique, le Big Data doit arriver à aider des élections.
Les équipes de l'ex-président américain ne s'y sont pas trompées.
Les fonds recueillis durant la campagne démocrate de 2012 ont battu des records avec plus d'un milliard de dollars engrangés, sachant qu'environ 70% étaients issus de donations en lignes, laissant des données personnelles avec lesquelles Barack Obama a plus cibler au mieux les appels aux dons.

De façon similaire, ils ont pu étudier l'attitude d'électeurs clés, les indécis.
En Ohio notamment, des milliers de données de vote furent injectées dans un algorithme qui a su reproduire 66 000 fois l'élection suivant différents scénarios, en simulant leur comportement, les équipes de l'ex-président Obama ont pu en tirer les arguments qui potentiellement sauraient les faire basculer dans leur camps.

2.2.1 - Protection des données personnelles

La segmentation ne profile pas les utilisateurs en tant que tels, mais plutôt les utilisateurs selon une variété de caractéristiques.
La personnalisation par segmentation s'est révélée efficace, mais non sans une certaine controverse.

Le site de voyage américain Orbitz a ébranlé les utilisateurs de Mac quand il a été découvert que les prix des hôtels qui leurs sont proposés sont plus élevés que pour les utilisateurs de PC. Les utilisateurs Mac dépensent en effet en moyenne 30% de plus dans la location d'un hôtel sur le site.

Encore une fois, c'est une méthode très efficace qui permet aux entreprises de diriger des publicités personnalisées vers les groupes concernés.
Cependant, cette méthode s'est également révélée un peu controversée.
Orbitz a irrité beaucoup de ses utilisateurs quand le journal de Wall Street a rapporté que le site montrait aux utilisateurs de Mac des hôtels à prix plus élevés que les utilisateurs de PC.

Leur PDG a répondu à la controverse en soulignant que les utilisateurs de Mac sont 40% plus susceptibles de réserver des hôtels de quatre ou cinq étoiles.
Leur but était simplement d'afficher les produits les plus pertinents pour chaque consommateur.

L'efficacité de la publicité ciblée par l'utilisation de profils d'utilisateurs ou la segmentation du marché est indéniable.
Par exemple, on peut considérer que les annonces graphiques réorientées sont 10 fois plus susceptibles de provoquer un utilisateur à rechercher des produits que les autres formes de publicité.
Mais malgré l'attrait de la publicité ciblée, de nombreux consommateurs craignent que leurs données puissent être utilisées de manière nocive.
En effet, les trois quarts des utilisateurs mobiles considèrent la publicité ciblée comme une atteinte à la vie privée.
De plus, 9 consommateurs sur 10 souhaiteraient que leurs navigateurs disposent d'une option «ne pas suivre» pour éviter les marketeurs.

N'est pas légalement ni encore moins éthiquement possible, tout ce qui l'est techniquement; et l'utilisation à outrance des données personnelles se révélerait contre-productive pour l'entreprise.
Même si beaucoup de données analysées sont anonymisées, les législations sur la protection des données personnelles, devront probablement évoluer pour prendre en compte l'existence de ces immenses quantités de données personnelles disponible "librement" sur Internet, et dont le titulaire n'a sans doute même pas toujours conscience.

L'anonymisation n'offre pas une protection totale, car chaque internaute, même non identifié, peut être associé à un profil très précis de comportement voire de personnalité.

Il faut aussi éviter les erreurs qui pourraient se produire lors de rapprochement de sources de données, entre lesquelles n'existent pas de clés universelles.

2.2.2 - Big Brother is watching us

Globalement, on peut rajouter que les réseaux sociaux, qui sont sans détour liés à la communication et au marketing, influent sur l'ensemble des types de vente des entreprises. Ces réseaux sont une vitrine virtuelle incontestable pour les biens et services et sont très fréquemment aussi présents dans la publicité actuelle, notamment la publicité en ligne. La position de leader dans la sollicitation d'entreprise pour des annonces publicitaires est occupée par Facebook.

Comme toute révolution industrielle, elle fait apparaître des risques.
Tant que la donnée est monétisable il y a risque.

Une enquête datant de novembre 2016 dans la presse américaine, sur le Wall Street Journal, fait part de l'enthousiasme du gouvernement chinois pour le traitement de données massives. La république populaire de Chine est pour l'instant occupée à tester, dans trois grandes régions dont Shanghaï pour ensuite l'étendre à toute la nation avant la fin de cette décennie, un système de notation qui repose sur le Big Data, le premier système centralisé de scoring personnel.

Il s'agit d'un scoring social basé sur trois types de données, d'abord les données classiques pour un scoring financier comme les remboursements de prêts, les incidents de paiement, les retards sur les paiements d'impôts auxquels s'ajoutent d'autres données classiques du Big Data commercial comme les préférences sur internet, la qualité de l'information postée ou encore la nature des messages que les gens postent. Mais en plus s'ajoute à tout cela ce qui concerne les sens civiques, c'est-à-dire les amendes, les activités de volontariat, le parcours scolaire, notamment si la personne a triché à des examens, le sens de la famille et notamment la façon dont une personne s'occupe de ses parents.

En fonction du ranking de chacun, les prix d'assurances, l'accès au crédit, l'accès au monde du travail, au emplois gouvernementaux en particulier ou encore l'accès des enfants aux écoles en sera affecté. Pour mettre ce système en place, des grandes firmes chinoises, tel que Alibaba, sont prêtes à partager leurs données.

Données privées gérées sur : https://www.google.com/settings/dashboard

- Historique et manière de consommer une vidéo Youtube
- Courriels, contacts, agenda, documents
- Historique des positions, historique Internet
- Applications smartphone

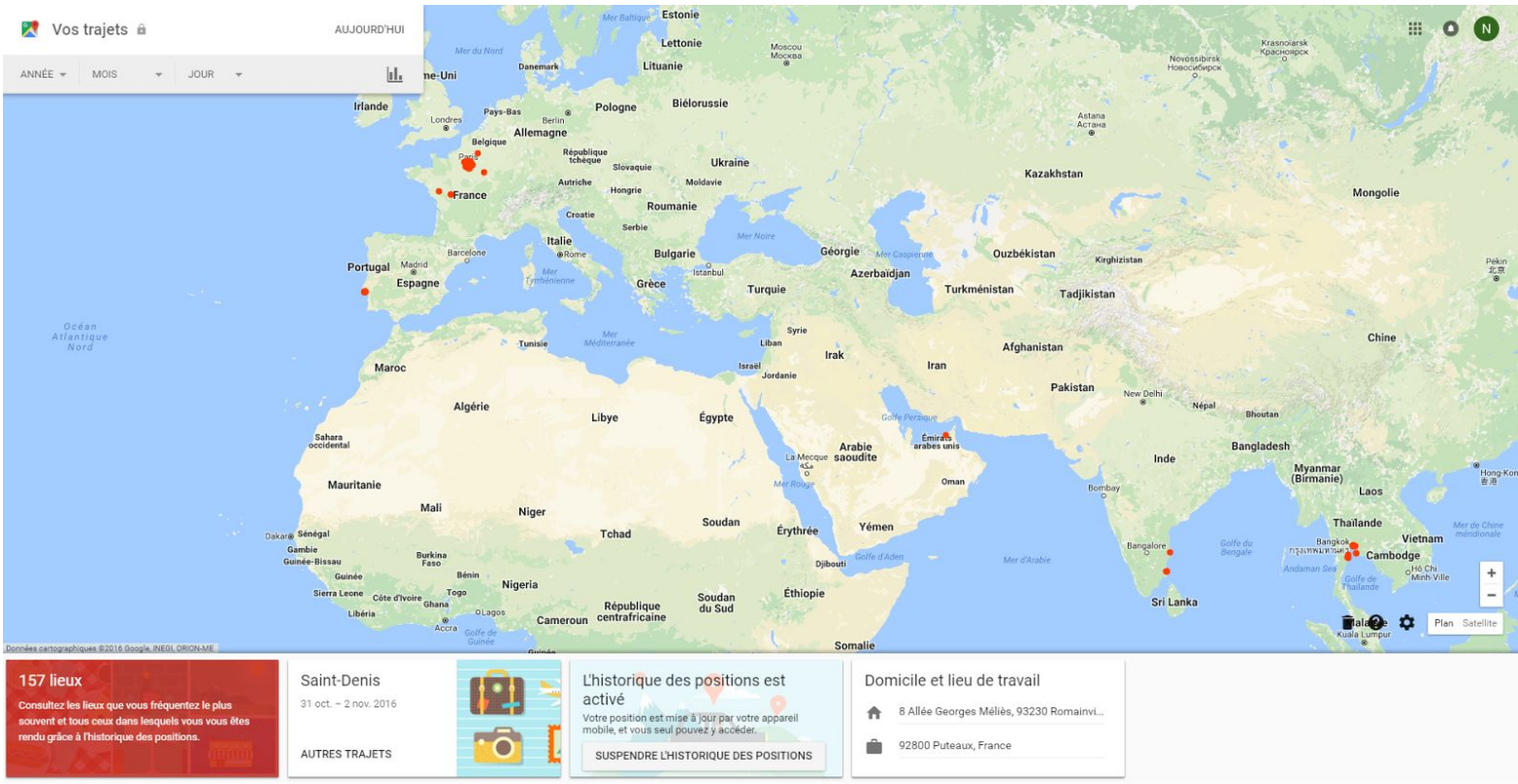

Figure 8 : Historique des positions Google

2.3 - Enjeux environnementaux

Certaines entreprises traitaient déjà les données de masse bien avant qu'apparaisse la transformation numérique; ERDF en est une qui peaufine depuis des années la modélisation de la consommation électrique de 35 millions de clients. C'est un atout colossale pour l'électricien qui souhaite aujourd'hui tirer le meilleur de la révolution digitale.

ERDF a considéré que dans leur secteur, l'énergie, la révolution numérique était une formidable opportunité pour utiliser le digital comme un levier à la fois de transformation technologique mais aussi de transformation culturelle. ERDF met à disposition en open data un certain nombre de données qui paraissent pertinentes pour un grand nombre d'acteurs, par exemple ils mettent à disposition sur leur site internet la production éolienne sur leur réseau. Ceci permet donc aux internautes d'aller consulter leur consommation comme par exemple constater un pic de consommation en février 2012 car la France a connu un pic de grand froid le 8 février 2012.

Ces flux de données permettent une meilleure transparence vis-à-vis des consommateurs mais aussi des producteurs, mais les données permettent aussi d'optimiser les performances de l'entreprise par exemple en améliorant l'entretien du réseau.
Des algorithmes sont développés à partir de ces données pour aider à anticiper les pannes.

Le Big Data permet aussi d'évaluer les fuites énergétiques sur un réseau, d'identifier les systèmes énergétiques ou thermiques concurrents et de mieux les gérer. Ainsi, on note que les sociétés qui utilisent le Big Data pour gérer leur management énergétique rapportent une baisse d'approximativement 20% de leurs frais énergétiques. C'est autant en énergie économisée et dès lors des impacts environnementaux évités.

2.4 - L'adaptation des commerçants et entreprises à ce nouvel outil

Le Big Data a un impact sur tous les aspects de la culture d'entreprise.
Les entreprises doivent évoluer de leurs méthodes et pratiques traditionnelles pour pouvoir utiliser le Big Data dans le but d'améliorer leur organisation.

2.4.1 - L'emploi

Le 12 septembre 2013, le gouvernement français lance 34 plans industriels, dans lequel fait partie le Big Data.

En effet, le gouvernement a besoin de "scientifiques de la donnée" (data scientist) qui connaissent les technologies informatiques, les enjeux métiers, et les méthodes statistiques.
Des centaines de milliers d'emplois de data scientists sont annoncés dans le monde.

L'absence de data scientists se laisse sentir dans le monde. On peut l'imputer à la prise de conscience récente de la capacité recelée par les données et à une mise en valeur toujours insuffisante du data scientist en entreprise.

Les premières formations spécialisées sont apparues en 2013 avec un master de Big Data à l'Université de New York et un mastère Big Data de de l'école Telecom-Paris Tech, peu de temps avant celui de l'ENSAI.

2.4.2 - Impact sur la chaîne logistique (Supply Chain Management)

- Optimisation des livraisons :
  L'analyse Big Data ouvre le chemin au monitoring et à l'amélioration en temps réel des livraisons. Au moyen des informations de localisation provenant des appareils télématiques des automobiles et à des informations tierces comme la circulation automobile, on est en mesure de choisir de re-router des livraisons de façon dynamique et de les calibrer jusqu'au niveau de la ruelle avec un impact positif sur les frais de transport.

- Gestion des approvisionnements :
  Avec des hangars, des fabriques et des centres de distribution éparpillés, la planification des infrastructures logistiques défie les moyens d'optimisation classiques. Le Big Data doit soutenir au façonnage de l'infrastructure avec en prime plus de vitesse, des tarifs de fret et d'entreposage diminués ainsi qu'une baisse de l'empreinte écologique.

- Amélioration des services clients.

- Elaboration de scénarios prédictifs (*eg.* Livraison prédictive d'Amazon) :
  Amazon a déposé un brevet au mois de décembre 2013 sur la livraison prédictive de produits. Le mastodonte américain du commerce en ligne désire diminuer le temps de livraison, les délais pouvant décourager certains achats sur la toile, et que l'article n'attende plus sagement son investisseur dans son centre de distribution mais soit déjà dans le hub de l'expressiste censé livrer le consommateur, bien avant que ce dernier n'ait eu le temps d'acheter l'article.
  Cette nouvelle approche, qui a fait le buzz au mois de janvier sur le Web, se base sur le Big Data : commandes, recherches d'articles sur le site, liste de souhaits, contenus des paniers, retours consommateurs et également combien de temps le client a passé sa souris sur un article.

- Supply Chain Management basée sur la demande :
  La management de la Supply Chain repose essentiellement sur les estimations des ventes, la plupart du temps imprécises face à une demande volatile et à des portefeuilles de produits complexes.

L'analyse combiné de récentes sources d'informations externes (réseaux sociaux, environnement, fournisseurs, senseurs) permet de mieux répondre et plus vite aux changement de demande. A la clé une amélioration des performances et une diminution des inventaires.

2.4.3 - Utilisation marketing des réseaux sociaux

De nombreux rapports récents des médias ont affirmé que le Big Data va bientôt remplacer les études de marché.
En effet, les géants des médias sociaux comme Facebook, Twitter et Yelp sont en train de changer leurs méthodes de recherches.

La croissance régulière des médias sociaux a généré des quantités massives de données sur les attitudes, les comportements et les préférences des consommateurs.

L'analyse de moteurs de recherches, des forums, et des réseaux sociaux permet de connaître les centres d'intérêts et les préférences des utilisateurs par conséquent leur comportement envisageable devant une proposition de service ou de produit.

Les entreprises se tournent de plus en plus vers cette grande nouvelle source de données comme un nouveau débouché pour l'étude de marché.
Dans le cas des sociétés qui font du B to B to C, cela s'avère particulièrement utile, car elles ont des contacts seulement avec des distributeurs et n'ont pas les moyens d'obtenir des informations directes sur leurs clients finaux.

L'analyse des réseaux sociaux n'est pas seulement utile dans le commerce et elle peut aider à la conception de produits novateurs, par l'analyse du sentiment favorable ou non de plusieurs caractéristiques des produits.

Mais alors que les abondantes données des médias sociaux reflètent les tendances actuelles des consommateurs, elles ne permettent souvent pas aux spécialistes du marketing de faire des prévisions significatives. Ceci est nécessaire à la compréhension de la psychologie du consommateur, qui est généralement masquée par les médias sociaux.

Beaucoup d'entreprises se détournent des efforts coûteux de l'étude de marché et s'appuient uniquement sur le Big Data pour trouver le bon chemin.

L'échec récent du mouvement autour de la TV 3D illustre parfaitement l'inconvénient potentiel de la dépendance excessive au Big Data et de l'accent mis sur la recherche fondamentale.
Après le succès écrasant du film Avatar, les téléviseurs 3D ont triomphé comme la prochaine révolution dans la télévision.

Dans le secteur de l'assurance ou du crédit, la recherche des profils d'internautes allant sur les sites de ventes de voiture, permet d'essayer de détecter dans sa propre clientèle ceux qui ont ce profil et seraient susceptibles de souscrire prochainement un contrat d'assurance ou de prêt.

En parlant de crédit, la société Lenddo, présente à Hong-Kong, scrute les flux Facebook et Twitter afin d'établir si ses clients méritent un prêt ou non.
En Allemagne, la société Kreditech, en fait de même, elle analyse pas moins de 8 000 paramètres depuis notre localisation, jusqu'à nos achats en lignes.
Ou encore Zest Finance, qui évalue le risque lié à ses clients à partir de dizaines de milliers de flux différents.

Les entreprises doivent garder un oeil attentif sur les médias sociaux, mais ne peuvent pas trop en dépendre.

2.4.4 - Le marketing digital

L'une des avancées les plus importantes facilitée par les bases de données croissantes est la capacité des entreprises à profiler les consommateurs.
Bien que la pratique ait conduit à de nombreuses préoccupations éthiques et juridiques, elle est devenue une méthode indispensable pour que les entreprises rendent leurs campagnes de marketing plus efficaces.
Une façon dont les entreprises recueillent des données pour créer de tels profils de ses utilisateurs est de promouvoir les connexions sociales.

Plutôt que d'exiger que les utilisateurs s'inscrivent à un site Web, il est souvent plus avantageux de leur permettre d'utiliser leurs connexions aux médias sociaux.
En connectant les utilisateurs à leurs profils sociaux, les entreprises peuvent immédiatement récupérer beaucoup d'informations sur un utilisateur.

Ces données démographiques peuvent ensuite être utilisées pour créer une distribution de courrier électronique plus pertinente pour chaque utilisateur en fournissant des messages de marque ciblés.
Semblable au profilage d'utilisateur, la segmentation est une autre stratégie de marketing populaire enracinée dans l'analyse Big Data.
Alors que le profilage des utilisateurs cherche à identifier chaque consommateur avec un large éventail de caractéristiques, la segmentation du marché est une approche moins ambiguë.

Au lieu de cibler les consommateurs individuels, le but de la segmentation est de définir un ensemble de caractéristiques dans une population et des groupes d'utilisateurs associés en eux.

Des plates-formes virtuelles automatisées (Ad Exchange) mettent en relation directe les acheteurs et vendeurs de publicité sur Internet. Quand un internaute est dirigé vers un site commerçant par une bannière, il ne se contente pas seulement de diriger l'internaute vers le site commerçant, mais enregistre des informations à l'aide d'un cookie déposé sur l'ordinateur de l'internaute.

L'agence connaît ainsi le profil de l'internaute, et les navigations qu'il a effectuées sur le Web. Elle sait aussi que les bannières n'ont aucun effet quand elles sont placées sur certains sites, et elle sait où les placer de façon optimale pour la conversion (achat, téléchargement...).
Elle sait même repérer les enchaînements de sites visités qui mènent le mieux à la conversion souhaitée. Lorsqu'elle repère qu'un internaute est en train d'exécuter cet enchaînement, elle peut acheter en temps réel une bannière publicitaire qui sera placée sous les yeux de l'internaute.
L'exploitation des cookies permet donc d'afficher en temps réel au moment où un internaute arrive sur un site, un contenu correspondant à son profil.

2.4.5 - Le Big Data dans la santé et l'industrie

Les innombrables capteurs (pression, température, usure, vibration...) positionnés sur les composants des appareils permettent une remontée instantanée d'une multitude d'informations qui, évaluées et modélisées, peuvent fournir une probabilité de problème, de rupture d'une pièce et de permettre une décision entre :

- des manoeuvres d'entretien inutilement lourdes et courantes, provoquant des frais non nécessaires.

- des manoeuvres d'entretien insuffisantes et laissant arriver des défaillances coûteuses, ou scabreuses.

Prédiction en temps réel de la consommation d'énergie, et également des problèmes, ainsi qu'une facturation plus rapide et économique, au moyen de compteurs intelligents (Linky).
On voit d'évidentes utilisations dans le domaine de la santé avec des capteurs sur un patient.

2.4.6 - Le Big Data pour l'assurance automobile

Aviva a créé une application mobile (Aviva Drive) qui analyse le style de conduite des automobilistes pour leur suggérer des tarifs appropriés.

Un projet équivalent avait été pensé en 2006 mais laissé à l'abandon deux années plus tard du fait de la complexité d'installer des boîtes permettant l'enregistrement des données dans les véhicules (boîtes noires).
Cette application scrute durant une centaine de kilomètres de nombreuses informations tel que le temps écoulé, le type de route mais surtout le nombre de kilomètres parcourus.

Un radical changement comportemental sera capable de faire suspecter une fraude.
Des applications plus classiques fournissent bien sûr des services tels que la localisation du dépanneur le plus proche, les numéros d'urgence...
Des capteurs sur le véhicule seraient également susceptible de signaler des risques de panne, indiquant à l'utilisateur le garage le plus proche ainsi que la conduite à tenir.

2.4.7 - Apports du Big Data aux entreprises

- Augmentation de parts d'audience et de marché, de la fidélisation :
  Augmentation des ventes induites par des recommandations.

- Aide à la prise de décision :
  Prévision des évolutions boursières et macro-économiques.

- Optimisation des processus :
  Exemple : les prévisions des besoins de l'entreprise et du prix des matières premières (en fonction de différents critères météorologiques, économiques...) dans le temps permettent d'acheter les bonnes quantités au bon moment.

- Mise en place de services novateurs :
  C'est le client lui-même qui bénéficie d'une aide à la décision, que ce soit par exemple pour le téléchargement de films ou de musiques (système de recommandation) ou pour la conduite automobile.

III - CONCLUSION ET OUVERTURE

La façon dont notre vie a été changée, est qu'aujourd'hui nous nous demandons : Quelle décision dois-je prendre ? De quelles données présentes pourrais-je me servir ? A quel point doit-je violer ma vie privée pour bénéficier de leurs services ?

Ceci a un impact fort sur la pensée.
Se demander "Qu'y a-t-il là ?" vous fera perdre loin dans l'espace.
Se demander "Qu'est-ce que j'aurais voulu avoir ?" vous ramènera au départ.

Nous n'avons plus besoin de savoir exactement ce que nous cherchons.

3.1 - Influence du Big Data sur sa propre évolution

En voulant, par leur propre volonté ou en étant contraint de suivre l'instinct humain de limitation des risques, les acteurs du Big Data finissent par créer les nouvelles normes, la nouvelle morale mondiale, déterminer ce qui est bon, et ce qui est mal, de mettre une étiquette sur chaque chose qui compose notre monde, que ce soit concret ou abstrait. Décisions prises entre deux conflits d'intérêts. En d'autres termes, les acteurs du Big Data sont les prophètes de cette divinité du 21ème siècle.

Par instinct de préservation, l'Homme essaie depuis toujours de prévoir l'avenir, par des méthodes scientifiques ou non, de la météo à l'astrologie en passant par les tests de personnalités. Combiné à la volonté d'efficacité de notre système économique, la question qui se pose est donc la suivante :

Comment manier les nouveaux algorithmes pour prévoir l'imprévisible et augmenter la part des bénéfices ?

Les possibilités ont aussi bien été comprises du côté des consommateurs.
A l'instar de Google Search, qui est devenu vital pour la majorité d'entre nous car annihile notre frustration de ne pas savoir, et ceci, en quelques secondes et à tout moment, la prédiction commence à être présente sur nos terminaux.

Plus seulement des réponses à des problèmes passés mais des réponses à des problèmes dans un futur plus ou moins proche. La phrase précédente est proche d'un slogan publicitaire et est dorénavant à la portée de nos téléphones et de nos

ordinateurs portables. Des applications sur ces derniers reliés à de puissants moteurs de calcul pourraient prédire et faciliter notre quotidien. Voilà la solution.

Les promesses à son propos, comme toute campagne marketing ou phénomène ayant eu un engouement fort récemment, peuvent décevoir et même si le Big Data véhicule quelques illusions et surtout redécouvre ce que beaucoup savaient, même s'il entre dans l'engouement pour le Big Data une part de buzz soigneusement entretenue, nous espérons avoir montré tout ce que le Big Data peut apporter dans nos vies sous la forme de nouveaux services et les questions qu'il pose sur la protection des données personnelles et de la vie privée.

Nous pensons qu'il serait dommage que l'Europe et la France restent à l'écart de cette vague de fond, qu'ils s'écartent de cette possibilité sans précédent de progrès.

En effet, nous affirmons que notre pays et l'alliance de nations dans lesquels nous vivons doivent développer leurs formations de data scientists, susciter de grands acteurs dans le stockage et le traitement de données, dans les services numériques, dans les moteurs de recherche, réseaux sociaux, système de blog, dans le cloud, détenant des parts de marché importantes.

On peut attendre de l'économie de l'information qu'elle apporte sa contribution à l'économie globale sans pour autant dépendre ou être en opposition totale aux géants Américains, comme Google.

Avant de conclure avec le rôle de l'intelligence artificielle, nous allons vous présenter le scénario suivant :

Alice vit à Paris. Une nouvelle boutique d'équipements sportifs vient d'ouvrir dans sa rue. Étant une sportive aguerrie, Alice s'y rend pour découvrir ce nouveau commerce qui, pense-t-elle, pourrait l'intéresser.

Une fois à l'intérieur de la boutique, elle se rend compte après un long tour dans les rayons relatifs à ses pratiques sportives que rien ne l'intéresse et rebrousse chemin.

Ce à quoi Alice n'a pas fait attention, c'est que la boutique indiquait sur sa porte d'entrée que les visiteurs, par leur entrée dans le magasin, autorisent la collecte et le traitement des données par le groupe possédant cette boutique.

En enregistrant et analysant le parcours des visiteurs, plusieurs informations peuvent être déduites par l'équipe en charge. La boutique peut donc savoir à propos d'Alice de ce qu'elle aime, ce qu'elle n'aime pas, ce genre d'informations pourra

établir un profil et servir aux prédictions d'achats des visiteurs de la boutique en se basant sur les préférences de chacun.

De ce fait, Pierre étant en charge du cas d'Alice et recoupant les données de ses moindres faits et gestes effectués lors de sa visite au magasin détermine qu'elle a quarante ans, qu'elle a deux enfants, qu'elle a tendance à aimer les jeans slim, qu'elle n'est pas sédentaire et même qu'elle s'entraîne à des triathlons, et consomme des barres énergétiques chaque week-end. Alice boit des boissons sucrées et travaille dans un environnement à haut niveau de stress.

Pierre, au vu de son expérience et du nombre de profils qu'il a auparavant traité, a remarqué qu'une quarantaine de profils de femmes similaires à celui d'Alice avaient un risque de cancer des ovaires sensiblement plus élevé que la moyenne des femmes de son âge.

Maintenant que le récit est arrivé son terme et en se basant sur celui-ci, nous pouvons nous poser la question suivante : Est-t-il légitime de faire savoir à Alice qu'elle a un grand risque d'avoir un cancer des ovaires ?

D'un côté, le poids moral est lourd car ne pas la prévenir pourrait l'empêcher de bénéficier des soins nécessaires et laisser le cancer atteindre un stade de non retour. Une des pires conséquences serait de priver le mari de sa femme et les enfants de leur mère.

De l'autre, les données devraient avoir un usage uniquement restreint au secteur d'activité, ici les équipements sportifs, pour éviter toutes les dérives imaginables ou non. De la plus évidente comme la manipulation du client pour le convaincre d'acheter en se basant sur ses faiblesses personnelles qu'il peut ne pas connaître, à la plus loufoque comme des hypocondriaques qui passeraient leur journée dans les magasins afin de bénéficier de cette analyse complète et gratuite.

Que doit faire Pierre ? Doit-il alerter ses responsables et/ou Alice ? Qu'est-ce qui est légal ?

Même si nous avons pris un cas extrême et qu'il entrera possiblement dans les cas exceptionnels, ce type de questions juridiques, éthiques et morales que nous essayons actuellement de traiter, sont encore aujourd'hui sans réponses.

Un cas, heureusement plus joyeux, est déjà arrivé. Il est donc important d'être prêt pour les prochains. Un couple se voyait proposer des produits pour bébé dans le commerce où ils allaient à l'habitué. En effet, l'algorithme traitant les données avait identifié le profil type via leur changement dans leurs choix alimentaires. Se voir

proposer des produits pour bébé sans raisons apparentes mena la femme du couple à faire un test de grossesse. Il se révéla positif.

Donc nous allons vous proposer un plan qui permettra de comprendre la prise de décision axée sur les données et de résoudre avec espoir certaines de ces questions morales et éthiques.

Ce plan se divise en cinq composantes qu'il est important de retenir. Ces cinq composantes sont la donnée, l'information, la connaissance, la compréhension et la sagesse.

Commençons par la donnée, nous entendons souvent ce mot et beaucoup de références envers ce mot. En latin "*data*" signifie "*une chose qui m'a été donnée et que j'ai interprété*".
Donc une donnée peut être des lettres, des numéros, des images, toutes ces choses que nous retrouvons sur Internet, toutes ces choses connectées grâce à nos ordinateurs. Par exemple 010184 est une donnée, cela peut être une date de naissance, ça peut être un code confidentiel ou même le montant d'un compte bancaire. Mais cette suite de chiffre n'est rien sans son contexte.
Et voilà que le Big Data n'est qu'un ensemble de symboles qui vont ensuite être traités par des machines.

Par ailleurs lorsqu'on prend toutes les données sur Twitter, tous les films numériques qui étaient créés, toutes les données partagées à travers Internet, dans les mails et autres, cette quantité correspond à plus de données qu'il n'est possible de stocker et de traiter par une simple machine traditionnelle.

De ce fait, en plaçant la précédente suite de chiffres (010184) dans un certain contexte, on peut comprendre qu'il s'agit d'informations, d'une date, 01 janvier 1984 et cette donnée prend donc toute son utilité. Ce récent mouvement, appelé "Données liées", est un important mouvement tendance qui court sur les scènes de l'Internet. Ce que nous faisons, c'est que nous donnons aux machines des moyens de donner du sens aux données. Nous prenons en réalité une partie de données sans connexion apparente et nous créons des liens.

Une fois que nous commençons à relier les données entre elles et que nous leur donnons un sens, nous pouvons commencer à poser des questions très intéressantes et les organiser, nous pouvons interroger ces données et nous pouvons même dire que la vraie connaissance qu'on peut en tirer vient de l'organisation de l'information. Une fois que nous avons identifié certains modèles et tendances dans ces informations, nous pouvons commencer à apprendre des choses intéressantes de ces données liées.

De cette manière, il est possible de savoir que lorsqu'il commence à faire beau, il y a une forte chance que la vente de glaces croît.
Ainsi une grande quantité de recherches est faite dans le monde de l'informatique pour identifier des modèles et des prédictions dans les données en se servant des outils de visualisations et de Machine Learning. Toutefois, avoir la connaissance d'une donnée ne signifie pas la comprendre et ceci est la quatrième composante du plan précédemment énoncé. C'est alors que nous savons que lorsqu'il fait beau, il y a plus de chance de vente de glaces et que nous pouvons ainsi utiliser notre intelligence pour comprendre pourquoi c'est le cas, parce que lorsqu'un être humain a chaud, il a besoin d'être refroidi.

Cependant, les ordinateurs trouvent qu'il est très difficile d'interpréter de telles informations, c'est pourquoi d'importantes recherches sont effectuées dans le but d'apprendre aux ordinateurs à comprendre les données et d'en tirer des décisions pertinentes.

La dernière composante de ce plan est la sagesse, c'est-à-dire l'utilisation de la compréhension, il s'agit là de mettre en pratique notre compréhension afin de prendre de meilleures décisions.

Ces cinq composantes commencent à être rassemblées dans le milieu universitaire et nous commençons à construire des systèmes qui permettent d'agréger des données, de "miner" (exploiter), de trouver des "patterns" (modèles) et qui essaient de prendre des décisions basées sur certains de ces modèles.
Nous pouvons dire que nous nous rapprochons de l'Intelligence Artificielle, nous allons examiner toutes les données que nous rassemblons tous ensemble et nous construisons des modèles et nous exécutons sur ces modèles des traitements pour en tirer des décisions.

Une fois que nous apportons ces cinq composantes ensemble dans un ordinateur, on peut commencer à parler d'Intelligence Artificielle car nous avons un comportement adaptatif; le mot "adaptatif" est essentiel dans cette définition car c'est ce que nous, l'Homme, faisons très bien, tandis que la machine ne le fait pas correctement. Peut-être que dans les prochaines décennies nous aurons des ordinateurs qui pourront apprendre et s'adapter automatiquement.

Maintenant que les cinq composantes du précédent plan sont acquises et en partant du principe que plus nous avons de données et meilleures les décisions prises seront, nous pouvons nous demander "*Qu'en serait-il si toutes les données étaient ouvertes ?*", si nos relevés bancaires étaient à découvert, si nos registres de santé étaient accessibles par tous, et absolument toutes nos données étaient complètement transparentes.

Nous pouvons imaginer que dans un monde avec plus de données transparentes et plus de connaissances publiques, nous nous comporterions mieux envers nous-mêmes. Nous n'en connaissons pas la réponse certaine mais peut-être que des études permettront d'observer le contenu de la pensée de chacun afin de savoir à quoi pense chaque personne, pense-t-il à une pomme ou à une maison.
Cet accès à la pensée de chaque individu pourrait permettre de créer d'encore meilleures prédictions.

Par contre qu'en serait-il si un jour nous créons des algorithmes qui portent des préjudices ? Qui serait responsable de contenus racistes proposés par des algorithmes ? Car des algorithmes pourront utiliser des critères de couleur de peau, d'ethnie, voire même de sexe afin de déterminer si une personne a le droit à une assurance par exemple. De plus, il est très difficile d'aller chercher à l'intérieur d'algorithmes et de comprendre comment ces inférences ont pu être faites. Donc la réponse à l'auteur de ces inférences reste à ce jour inconnue.

La prise de décision est jusqu'à aujourd'hui difficile, que ce soit pour un humain ou pour une machine. Mais aujourd'hui le Big Data est utilisé pour nous aider à faire un choix entre plusieurs produits qui pourraient nous intéresser.
A un moment donné, nous allons commencer à avoir des ordinateurs qui prendront des décisions de plus en plus complexes et finalement nous allons intégrer des systèmes d'Intelligence Artificielle dans des machines, des robots. Ces robots serviront à aider les humains à répondre à des questions complexes telles que "*Comment mettre la fin à la pauvreté dans le monde ?*", "*Comment mettre fin aux cancers ?*". Mais une réponse toute simple permet de résoudre ces deux questions, qui est l'extinction de l'espèce humaine, voilà ce qu'un système d'Intelligence Artificielle pourrait interpréter.

Donc nous devons être très méfiants vis-à-vis du pouvoir que nous pourrons accorder aux systèmes de prises de décisions car bien évidemment une machine ne peut contextualiser certaines choses de la même manière que l'Homme le fait avec toutes les contraintes qui s'imposent à lui comme le fait de ne pas détruire l'humanité toute entière pour éviter la pauvreté ou le cancer.

Il faut admettre qu'il est vrai de penser que ce qui est de l'ordre de la science fiction aujourd'hui, sera de la science factuelle dans les prochaines décennies et que le Big Data et l'Intelligence Artificielle sont parmis les choses les plus révolutionnaires qui puissent arriver grâce à l'existence d'Internet.

Pour finir, nous rappelons la célèbre citation (de comics) suivante :
"*Un grand pouvoir implique de grandes responsabilités*" - Oncle Ben

BIBLIOGRAPHIE

Livres

Marc DUGAIN et Christophe Labbé, L'Homme Nu, la nouvelle dictature du numérique, Editions Robert Laffont, 2016.

Telit Communication, Internet of Things - Executive Handbook 2016, 2016.

Diagramme

Technology Hype Cycle - Gartner Inc.

Articles

Big data santé : et le respect de la vie privée ?
12 septembre 2016 - Nathalie Devillier
https://www.contrepoints.org/2016/09/12/265561-big-data-sante-vie-privee

Big data : Axa fait un pas de plus vers l'assurance connectée - juin 2014 - Samuel Vasquez
http://www.jechange.fr/assurance/mutuelle-sante/news/big-data-axa-assurance-connectee-11-06-2014-3298

Lower Your Car Insurance Bill, at the Price of Some Privacy - Ron Lieber - AUG. 15, 2014
http://www.nytimes.com/2014/08/16/your-money/auto-insurance/tracking-gadgets-could-lower-your-car-insurance-at-the-price-of-some-privacy.html?_r=2

Time - Exclusive: Obama's 2012 Digital Fundraising Outperformed 2008 - Michael Scherer
http://swampland.time.com/2012/11/15/exclusive-obamas-2012-digital-fundraising-outperformed-2008/

Inside the Secret World of the Data Crunchers Who Helped Obama Win - Michael Scherer
Nov. 07, 2012
http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/

Leo Mirani, London's bike-share program unwittingly revealed its cyclists' movements for the world to see, April 16, 2014.
http://qz.com/199209/londons-bike-share-program-unwittingly-revealed-its-cyclists-movements-for-the-world-to-see/

"Un Français sur deux se dit prêt à confier ses données personnelles à son assureur pour obtenir la meilleure offre en assurance de biens et responsabilités (habitation, automobile...)"
Communiqué de presse - PWC France - 2014

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, Unique in the Crowd: The privacy bounds of human mobility, 25 March 2013.
http://www.nature.com/articles/srep01376

Théorème CAP : https://en.wikipedia.org/wiki/CAP_theorem

Emissions

Les enjeux du big data - FUTUREMAG - ARTE
https://www.youtube.com/watch?v=ye-DsD_EHKk

La tronche en biais - Liberté du numérique, invité développeur de Framasoft
https://www.youtube.com/watch?v=WjBf_G1UH4Y

Conférences

Susan Etlinger, What do we do with all this big data ?, TED at IBM, septembre 2014

Kenneth Cukier, Big data is better data, TED Salon Berlin, juin 2014

Joel Selanikio, The surprising seeds of a big-data revolution in healthcare, TEDxAustin, 2013.

Ben Wellington, How we found the worst place to park in New York City using big data, TEDx New York, novembre 2014.

Charlie Stryker, Big Data will impact every part of your life, TED Fulton Street, septembre 2014.

Daniel Hulme, Big Data and dangerous ideas, TEDx UCL, janvier 2015.

Sudha Ram, Creating a smarter world with Big Data, TEDx Tucson, janvier 2014.

Jake Porway, Big Data in the service of humanity, TEDx Montreal, juin 2012.

Royal Institution of Great Britain, Information, Evolution, and intelligent Design With Daniel Dennett, Londres, mai 2015
https://www.youtube.com/watch?v=AZX6awZq5Z0

Photographies et Diagrammes

Théorème CAP
http://blog.nahurst.com/visual-guide-to-nosql-systems

IBM 650
https://upload.wikimedia.org/wikipedia/commons/thumb/c/c7/IBM_650_at_Texas_A%26M.jpg/280px-IBM_650_at_Texas_A%26M.jpg

Cartographie Internet
http://internet-map.net/

Historique Google
http://www.google.com/settings/dashboard/