BIG DATA, L'ACCÉLÉRATEUR D'INNOVATION

Livre blanc de l'institut G9+

En partenariat avec

renaissance numérique
le think tank de la société numérique

INTRODUCTION

 Date de parution: décembre 2014

VICE-PRÉSIDENT
DE L'INSTITUT G9+

ADMINISTRATEUR
DE RENAISSANCE NUMÉRIQUE

LUC BRETONES

Nous avons choisi en 2013 d'analyser le potentiel du marché des objets connectés et ses dynamiques internationales en publiant, sur la base d'entretiens avec les meilleurs spécialistes, notre livre blanc «les nouveaux eldorados de l'économie connectée» et en lui dédiant avec succès notre rencontre annuelle.

2014 est donc naturellement l'année du Big Data prédictif

pour l'Institut G9+ qui vient d'organiser au premier semestre la principale conférence sur le sujet en France sous le titre « ils font parler les données pour accélérer l'innovation ».

Nous voyons en effet, dans l'explosion des données générées par les objets connectés et les activités humaines, l'émergence ultra rapide d'un nouveau paradigme, celui de la « mise en donnée » de nos vies et des potentiels d'analyse de corrélations relatives.

Plus précisément, la multiplication des objets connectés va accélérer l'émergence de gisements de données personnelles pour de nombreux acteurs économiques dans tous les secteurs (opérateurs télécom, banques, assurances, industriels, distributeurs, transporteurs...).

La rencontre des données issues de ces objets connectés, qu'elles proviennent de grands groupes ou d'autres acteurs, peut permettre de constituer des collections inédites de type Big Data, dont le volume, la précision, la richesse et la portée seront à la source d'énormément d'analyses poussées, d'opportunités de croisement et de corrélations par l'intermédiaire de services et d'applications qui sauront « révéler » des informations de plus haut niveau. Dans le même temps, les données générées par les particuliers et les entreprises sont désormais quasi exclusivement numériques et en croissance volumique exponentielle. Souvenons-nous qu'en 2007, déjà, seulement 7 % des données restaient au format analogique, or les données numériques font

plus que doubler tous les 14 mois. Axelle Lemaire, secrétaire d'Etat au numérique, n'y voit-elle pas le pétrole du XXIème siècle ? Certains préconisent même de les intégrer systématiquement au bilan des entreprises. Et pour cause, la valeur de l'économie globale, après s'être développée et concentrée massivement dans le logiciel, qui, comme le décrit si justement Marc Andreessen[1] , « dévore le monde », tend à prendre un nouveau virage et une forme encore plus élaborée, celle de l'analyse mais surtout de la possession des données du monde.

Ce mouvement vers les acteurs qui contrôlent les données, au-delà des experts capables de les analyser, prépare des bouleversements majeurs dans la chaîne de valeur économique mondiale et dans les business modèles des entreprises.

Sommes-nous à l'orée d'une émergence oligopolistique de géants mondiaux de la donnée, ou au contraire de nouvelles sociétés agiles, ultra expertes de ce domaine et bénéficiant de la « taille sans la masse »[2] ? Dans un monde ainsi mis en données, les mathématiques, les statistiques et la programmation deviendront-elles les nouvelles langues vivantes, aussi incontournables que fondamentales ? Que devons-nous attendre des Etats en matière d'orientation de l'éducation d'une part et d'investissement en infrastructure de l'autre ?

Et au-delà des opportunités, quels sont les risques engendrés par ce nouveau paradigme, sur notre vie privée bien sûr, mais également sur notre libre arbitre, notre choix individuel, face à une dictature potentielle de la prévision.

Il semble que le principe de précaution appliqué au Big Data porterait un coup d'arrêt au potentiel important de ce nouveau marché. Pour autant, il conviendra de définir rapidement les conditions d'utilisations secondaires innovantes des données collectées ou accédées. De même, l'anonymisation parfaite des données étant impossible à partir d'un certain volume[3], et la mise à jour de tout ou partie des graphes sociaux à portée d'algorithme, ces conditions devront assurer aux individus et organisations un strict respect et les moyens de leur protection.

Comme l'humanité a su le faire avec les précédentes révolutions technologiques, je suis convaincu que l'usage du Big Data sera régulé ; ce n'est qu'une question de temps et d'apprentissage .Quant au déterminisme potentiellement extrême induit par la mise en données du monde, gardons à l'esprit que le génie humain ne dépend pas du Big Data, pas plus que l'invention de la voiture n'a fait l'objet d'une demande prévisible des cavaliers, ou ceux de l'Ipad d'un besoin exprimé par les utilisateurs d'ordinateurs personnels.

Dans ce monde de données mises à nu en temps réel, je pense que les dimensions humaines de discernement, d'expérience et de

(1) Entre 2006 et 2014, le classement Financial Times 500 par secteur économique mentionne une progression de +116% des services logiciels et informatiques à 1 744 928,4 millions de dollars contre une progression de seulement +7% pour le secteur télécom fixe et mobile.
(2) Selon l'expression du professeur Brynjolfsson (MIT – Digital Business)
(3) Paul Ohm, professeur, Université du Colorado à Boulder

créativité, seront encore plus cruciales, encore plus différenciantes.

Et comme le note Kenneth Cukier dans son ouvrage Big data, la révolution des données est en marche, le monde présent du Big Data nous paraîtra sous peu aussi dépassé que les quatre kilo octets de mémoire vive de l'ordinateur de bord d'Apollo 11.

 Date de parution: décembre 2014

FONDATEUR
DE 1000MERCIS-NUMBERLY

ADMINISTRATEUR
DE RENAISSANCE NUMÉRIQUE

THIBAUT MUNIER

Si le Big Data représente une avancée technologique généralement peu contestée, ses possibilités d'utilisation cristallisent trop souvent les doutes et les peurs d'une large partie de la population.

Big Data : une triple opportunité à ne pas laisser passer

Si le Big Data représente une avancée technologique généralement peu contestée, ses possibilités d'utilisation cristallisent trop souvent les doutes et les peurs d'une large partie de la population. La complexité du sujet et la variété des domaines impactés conduisent parfois à faire des amalgames hâtifs et dangereux, ainsi qu'on a pu le voir après les révélations de Snowden sur les systèmes de surveillance massive.

A l'instar de nombreuses innovations technologiques, le Big Data peut certes donner lieu à des dérives liberticides qui doivent évidemment être identifiées, régulées et si possible éradiquées. Mais il paraît tout aussi fondamental de ne pas se contenter de cette vision restrictive et de considérer avec au moins autant d'attention les **immenses opportunités que cette révolution contient en germe.**

Le Big Data doit avant toute chose être considéré comme une révolution technologique dans la capacité de collecte, de stockage et d'exploitation des données. Apparu sur la côte Ouest des Etats Unis à la suite du développement massif des usages digitaux[4], le Big Data constitue aujourd'hui **une triple opportunité** pour les consommateurs, les entreprises et la croissance de notre pays.

Une opportunité pour les internautes et les consommateurs

Grâce à ces nouvelles capacités de stockage et de traitement des données, les consommateurs ont progressivement pu découvrir des services nouveaux, qu'ils ont par-

(4) En février 2001 Doug Laney, analyste au Meta Group, publie une note de recherche intitulée "3D Data Management: Controlling Data Volume, Velocity, and Variety." 10 ans plus tard les 3 Vs sont les 3 dimensions généralement utilisées pour définir le Big Data bien que le concept n'apparaisse pas dans l'article de Laney.

fois plébiscités, entraînant des besoins plus grands encore et souvent une nouvelle grappe d'innovations.

À titre d'exemple de ces nouveaux usages directement issus de l'essor du Big Data, on pourrait citer les moteurs de comparaison de prix qui nécessitent bien sûr d'immenses capacités de traitement de données en temps réel et qui permettent chaque mois à des millions d'internautes d'acheter mieux et moins cher dans de multiples secteurs.

Une autre demande forte des consommateurs qui a pu commencer à être adressée grâce aux technologies du Big Data concerne la communication directe Marques-Consommateurs.

Devant l'afflux de messages souvent non ciblés et sans intérêt pour leurs destinataires, les internautes ont joué de leurs « contre-pouvoirs digitaux » : plaintes, désinscriptions, non réactions, réclamant avec force une communication plus pertinente de la part des marques dont ils sont par ailleurs des clients exigeants et souvent fidèles.

Grâce aux possibilités offertes notamment par les bannières publicitaires achetées aux enchères en temps réel (Real Time Bidding) il devient aujourd'hui possible pour les marques de reconnaître leurs clients sur les différents terminaux de connexion qu'ils utilisent, et ainsi de les solliciter beaucoup moins fréquemment mais à bon escient, d'être globalement plus intelligentes et moins envahissantes.

C'est par le développement d'algorithmes sophistiqués que chaque marque peut espérer donner du sens aux données collectées et simplifier la vie de ses clients en limitant le nombre de messages et en recréant de la valeur lors de chaque interaction. Par une communication et des services beaucoup plus pertinents, il s'agit en quelque sorte pour les entreprises de **rendre à chaque consommateur la valeur des données confiées.**

Une opportunité pour les entreprises

Grâce au Big Data les entreprises ont également devant elles des opportunités formidables pour revoir leur chaîne de valeur et transformer leurs points de vente.

Avec les produits connectés, il devient en effet envisageable pour une marque de capter de façon automatique et anonyme une quantité importante d'informations sur l'utilisation de chaque produit pour en améliorer la qualité, la durée de vie et en cas de panne (par exemple pour une voiture) pour établir le diagnostic et définir la réparation nécessaire.

Le Big Data permet enfin à beaucoup d'entreprises d'envisager une transformation de leurs points de vente et du rôle de leurs vendeurs. Equipé d'une tablette un vendeur pourra par exemple accéder à l'historique d'activités de ses clients ou à des recommandations personnalisées et ainsi compléter son propre jugement en face à face sur un point de vente afin d'ap-

porter un meilleur conseil dans le cadre d'une relation enrichie et d'un métier totalement réinventé.

Connaître et comprendre un consommateur n'empêche d'ailleurs pas la surprise et l'inattendu : à tout algorithme de recommandation peut être intégré une dimension de sérendipité, d'exploration ou de hasard pour éviter un systématisme rapidement inefficace.

Une opportunité pour la croissance et l'emploi dans notre pays

Du fait de la qualité de ses structures d'enseignement et de recherche en mathématiques appliquées, notre pays possède tous les atouts pour être aux premiers rangs dans la formation des « Data Scientists » et dans la création d'entreprises et de nouveaux usages qui en découleront. C'est dans l'environnement actuel une opportunité rare en termes d'emploi et de croissance et il ne serait pas concevable de la négliger.

En conclusion, le Big Data offre aujourd'hui un vaste champ d'applications possible, et demeurent aux prémices de leur développement[5]. Ces technologies et leurs applications méritent bien qu'on les observe sans naïveté ni a priori, d'un œil critique et avec discernement, mais de façon d'abord positive et entrepreneuriale avec ambition et l'envie de construire.

Après tout, **le Big Data ne sera que ce que nous en ferons.** Il constitue aujourd'hui un champ unique d'opportunités et elles méritent une grande attention si nous voulons peser dans ce débat global qui est déjà ouvert. De nombreuses questions relatives aux données, à leur collecte et leur utilisation notamment par la robotique vont en effet se poser dans des domaines aussi variés que la protection des données personnelles, la santé ou la place de l'homme dans la société.

Etudiants, entrepreneurs, décideurs, ou chercheurs doivent tous ensemble participer à cette construction et à ce débat qui ne peuvent qu'avancer de pair. Avec une double exigence : être rapide car la concurrence est mondiale et pragmatique car c'est en faisant qu'on comprend les enjeux.

(5) L'Association française des éditeurs de logiciels (Afdel) a ainsi estimé que la création de valeur liée au Big Data pourrait atteindre en France 2.8 milliards d'euro et 10 000 emplois directs d'ici cinq ans. Le cabinet américain Gartner estime de son côté que le secteur Big Data créera 4,4 millions d'emplois dans le monde d'ici à 2015, dont 1,9 million aux États- Unis.

PRÉSIDENTE
INSTITUT G9+

ADMINISTRATRICE
INGÉNIEURS ET SCIENTIFIQUES DE FRANCE

DIRIGEANTE IT
TRANSITION

La gestion des données était jusque là réservée à des domaines d'expertise, spécialistes chacun de son métier.

VALENTINE FERRÉOL

Big Data : un levier supplémentaire pour imaginer, construire, s'inspirer

Dimension industrielle

La gestion des données était jusque là réservée à des domaines d'expertise, spécialistes chacun de son métier. Qu'ils soient techniques : stockage, sauvegarde, archivage au sein de datawarehouse ; ou fonctionnels : mathématiciens, traders, exploitants de centrale électrique, contrôleurs aériens ou encore les services publics, la santé, la culture etc… La performance - performance de la chaine de valeur de nos entreprises et de l'économie de demain - réside dans le bon fonctionnement des réseaux qui coopèrent à l'élaboration des produits ou services. Le Big Data est un gisement colossal de gain en productivité si les données utilisées correspondent à une facette de la réalité que nous cherchons à étudier.

En faisant parler les données, en leur donnant du relief chaque acteur a potentiellement accès à une meilleure compréhension du contexte de sa filière, son entreprise, de son métier, peut en percevoir les évolutions (service ou produit). Cette mise en perspective génère également une sorte d'Empathie avec tous les acteurs de la chaine de valeur qu'ils soient collaborateurs, fournisseurs, partenaires, clients. Cette coopération « enrichie » est le facteur-clé de succès de notre économie.

Dimension sociétale

Les traces que nous laissons volontairement ou involontairement, directement ou indirectement, de part nos comportements, nos actions, les

objets que nous utilisons, nos propos, notre appartenances à telle ou telle communauté sont autant de data utilisées pour des études sociodémographiques, socioéconomiques, sociologiques. Ces data sont elles aussi utilisées grâce à des algorithmes très sophistiqués qui ont pour objectif de prédire nos comportements aussi bien individuellement que par « catégorie » ou groupe d'individus.

Et qu'en est-il donc lorsque la signification, la portée, le contenu associés à ces traces changent de sens ? Car tous les codes évoluent, se cassent et se reconstruisent à une vitesse folle : langues, langages, codes informatiques, codes culturels, les stéréotypes, le fonctionnement « en tribu » de communautés qui se font et se défont. Quelle valeur et quelle validité dans le temps peuvent avoir les prédictions ainsi constituées ? A chaque modélisation et chaque algorithme sont associées des hypothèses qu'il convient de (re)préciser, des paramètres intrinsèquement évolutifs qu'il convient de réajuster, qui mettent en scène des données collectées dans un contexte très spécifique.

Il convient donc aussi de mettre à l'épreuve de manière continue les comportements ainsi modélisés. Cela revient à modéliser de manière dynamique l'évolution des algorithmes pour tendre à les rendre intelligents. Pour autant…nous vivons dans un monde réel. Augmenté ? Souvent. Connecté ? De plus en plus. Mais dans un monde qui est toujours bel et bien réel. Nous sommes des hommes et des femmes bel et bien vivants, avec notre quotidien, nos émotions, nos valeurs et nos rêves.

Que portent donc ces fameuses Data, devenues Big ? Imaginer, concevoir, modéliser, implémenter… et aussi observer. Observer pour (re) trouver le sens et s'ouvrir vers de nouvelles inspirations. Prendre en compte l'ampleur de cette période que nous traversons, certes une période de crise permanente mais aussi fabuleuse car porteuse de tant d'avenirs potentiels. L'ampleur de cette révolution que nous sommes en train de vivre et dans laquelle aujourd'hui nous avons la possibilité et l'ambition de redevenir partie prenante.

Alors, je dis « oui » aux big data ! Hackons ensemble, et que ce soit pour le meilleur !

PARTIE I

LE BIG DATA :

POURQUOI PARLER DE RÉVOLUTION ?

Depuis des années, les mathématiciens élaborent des modèles mathématiques pour faire parler des jeux de données. Cela commence par un simple modèle statistique, basé sur un jeu de quelques informations, à un modèle prédictif élaboré, basé sur des milliards de données, permettant de prévoir demain quelle région du monde sera la plus touchée par une maladie ou comment réguler le trafic pour éviter les pics de pollution.

Si le traitement de données massives existe depuis déjà des dizaines d'années, notamment dans les pratiques de marketing ciblé utilisées par toutes les grandes entreprises depuis leur fichier clients, pourquoi le terme de révolution est-il alors tant employé aujourd'hui ? Le Big Data représente-t-il un vrai tournant, et pour quels acteurs ? S'agirait-il d'une révolution mathématique, technologique, politique et sociale ?

Pour Henri Verdier, Administrateur général des Données en France, la révolution de la donnée que nous traversons est le troisième acte de la révolution numérique[6]. Cette dernière a débuté dans les années 1980 avec la révolution informatique et l'augmentation fantastique de la puissance de calcul des ordinateurs, puis, à partir des années 1990, la révolution Internet qui mit en réseau les ordinateurs et, avec l'avènement du web 2.0, les humains du monde entier. La révolution de la donnée s'est faite jour avec l'intensification de nos pratiques en ligne et la massification des capteurs, à commencer par nos téléphones mobiles. Outre la technologie mise en place, l'aspect révolutionnaire du Big Data repose dans la multitude d'applications possibles, qui touche tous les pans de notre société. Les océans de données disponibles sont au centre des choix stratégiques des organisations, alimentent le débat public (vie privée notamment) et modifient les comportements des individus (santé/bien-être, goûts culturels, vie sociale...).

Cette première partie a pour ambition de définir les facteurs qui font que le Big Data peut être considéré comme une révolution aujourd'hui. Poser le postulat de cette révolution par la donnée et son traitement exige un travail de définition et de compréhension du concept de Big Data, souvent négligé par des discours marketing peu enclins à s'attarder sur cette question. Quelle définition pour le Big Data ? Quelles sont ses implications tangibles ? Qui en sont les acteurs ?

C'est ce changement de paradigme qui nous permet de parler de révolution dans son sens le plus strict : un bouleversement violent dans notre perception du monde.

« La valeur de l'informatique était de créer des outils pour manipuler les données puis dans la création des process qui manipulent ces outils. Maintenant, on se rend compte que la valeur se trouve dans la donnée elle-même ».

Gaëlle Recourcé, Directrice Scientifique, Evercontact.

A BIG DATA : D'UNE DÉFINITION STATIQUE À UN PROCÉDÉ

> Demandez à n'importe quel chief data officer de définir Big Data et il va se mettre à regarder ses chaussures. En réalité, il y a de forte chance pour que vous obteniez autant de définitions différentes que le nombre de personnes auxquelles vous poserez la question
>
> MIT Review[7]

Au cours des dernières années, définir le terme "Big Data" s'est révélé être un exercice périlleux. Quel est le critère de définition premier : le volume de données traitées ? Le logiciel de traitement de la donnée ? La nature des traitements qui leurs sont appliqués ?

I. LA TYPOLOGIE DES 3 V

Dans le maquis des définitions, les 3V se distinguent comme le plus petit dénominateur commun. Apparue en 2001, elle est le fruit des analyses de Doug Laney, employé de Gartner, dans son rapport « 3D Data Management: Controlling Data Volume, Velocity, and Variety » [8]. Omniprésente dans la littérature sur le Big Data, elle identifie trois critères définitionnels : le volume, la vitesse et la variété des jeux de données.

(7) Big Data Gets Persona, MIT Review, Octobre 2013
(8) Cabinet Gartner, Janvier 2012, http://blogues.gartner.com/doug-laney/files/2012/01ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Volume

Le volume de données traité est considéré comme le premier critère pour qu'un ensemble de données relève du Big Data. Pourtant, ce premier V est le moins opérant et le plus variable en fonction du secteur et de l'organisation concernés : où faut-il placer le curseur ? Peut-on parler de seuil au dessus duquel on entrerait dans le monde du Big Data ? Pour Florian Douetteau, fondateur de Dataiku, ce critère confine au non-sens : « Jongler entre péta et téra octet, après tout, il n'y a qu'un facteur mille entre les deux ... ! ». de manipuler de larges volumes de données («Volume»), possiblement disparates («Variété»), nécessitant ou non d'être manipulées en temps réel («Vélocité»). Selon le besoin, on va privilégier tel ou tel module de notre plate-forme, pour optimiser le traitement des données.

Ainsi, la vélocité est cruciale quand il s'agit de scorer un visiteur lors de son parcours sur site, pour nourrir les plate-formes de ciblage publicitaires ; a contrario, on privilégiera la capacité à traiter en fort volume des données variées, quand il s'agit d'analyser à froid la valeur «lifetime» d'un client, ou de croiser les parcours digitaux avec la base CRM d'un client pour affiner le risque d'attrition de sa base client ».[9] De plus, la problématique du seuil se pose aussi pour ce deuxième critère.

Vitesse (Velocity)

Ce critère de vitesse renvoie à la faculté de traiter les jeux de données en un temps record, voire, le plus souvent, en temps réel. Cela permet de créer des services directement fondés sur les interactions présentes. Pourtant, parmi les organisations qui traitent du Big Data, toutes n'offrent pas cette instantanéité ou n'en ont tout simplement pas besoin : « Nous disposons d'une plate-forme de mining propriétaire qui nous permet

Variété

La variété des données traitées est un enjeu singulier au Big Data et incarne par là un critère important de leur définition. La diversité des sources et des formats des jeux de données représente un véritable défi technologique. À titre d'exemple, le CRM – customer relationship management, gestion de la relation client – d'une entreprise peut contenir des données issues des réseaux sociaux, des cartes de fidélité physiques ain-

(9) Entretien avec Arnaud Massonnie, Co-fondateur et Directeur Général de l'agence fifty-five

si que de l'interaction en magasin. Agréger ces données pour les traiter ensemble est la première difficulté que rencontrent entreprises et organisations et souligne l'enjeu primordial de l'interopérabilité des données. La suggestion du cabinet NewVantage Partners de remplacer le terme Big Data par Mashup Data est à cet égard très significatif[10].

Valeur et Véracité ?

En outre, il est fréquent de voir la définition des 3 V complétée par un 4ème voire un 5ème V, pour véracité, qui recouvre la précision et l'exactitude des données, et valeur, portant son attention sur la capacité intrinsèque de la donnée de créer de la valeur.

II. BIG DATA : UN CONCEPT LARGE AUX DIFFÉRENTES ACCEPTIONS

« La terminologie et les frontières du Big Data sont floues parce que ce concept connaît des champs d'applications très différente » - Romain Lacombe, Chargé de l'innovation et du développement de la mission Etalab

Santé, sport, ressources humaines, transports urbains : comme mode opératoire, le Big Data recouvre une multitude de réalités - ce qui, pour certains acteurs, rend leur définition complexe. Dans son rapport en mai 2011, McKinsey écrivait : « Il est important de noter que la définition peut varier par secteur, en fonction de quels types de logiciels sont disponibles et de la taille des jeux de données dans telle ou telle industrie ».

Cette difficulté à définir ce qu'est le Big Data provient de la diversité des acteurs qui se sont emparés de cette expression. Chaque définition est ainsi colorée différemment en fonction d'objectifs et d'intérêts[11]. Ainsi, il n'est pas surprenant de constater que la définition d'Oracle met l'accent sur l'infrastructure qui constitue le cœur de son activité : « Le Big Data est le résultat de l'exploitation d'une base de données traditionnelle, enrichie par des données non structurées. »

De la même manière, Intel fonde la sienne sur son expérience avec ses clients : « Les opportunités offertes par le Big Data sont issues des organisations générant environ 300 terabytes de données par semaine. Le type le plus répandu de données analysées de cette façon sont les transactions commerciales, suivies des documents, emails, données capteur, blogues et medias sociaux. » Microsoft, quant à lui, insiste sur le besoin en puissance de calcul : "Big Data est le terme de plus en plus employé pour décrire le processus qui applique la puissance informatique : machine learning et intelligence artificielle à un jeu massif et souvent très complexe d'informations ».

(10) Big Data Executive Survey, 2013, Cabinet NVP, http://newvantage.com/wp-content/uploads/2012/12/NVP-Big-Data-Survey-Themes-Trends.pdf
(11) Définitions collectées dans Undefined By Data: A Survey of Big Data Definitions, Jonathan Stuart Ward and Adam Barker, School of Computer Science at University of St Andrews, UK, Octobre 2013, p.1

III. LE BIG DATA : LA DÉFINITION PAR LE PROCÉDÉ

Les deux points précédents illustrent la difficulté à définir le Big Data comme un fait statique. Pour mieux appréhender la notion, il conviendrait de distinguer ce qui est nouveau – nombre de données et nouvelles opportunités technologiques – de ce qui ne l'est pas : son principe de fonctionnement.

Un fonctionnement traditionnel en trois temps

On peut définir le Big Data comme un processus de traitement de la donnée qui comporterait trois étapes : collection, agrégation et analyse. Ce n'est qu'à travers ces trois actions que des ensembles de données, si vastes et véloces soient-ils, deviennent du Big Data.

La collection des données

Construire une base de données nécessite de récolter une multitudes d'informations générées tant par la navigation en ligne (du clic au surlignage d'un texte), les objets connectés de notre quotidien, les organisations publiques ou privées qui libèrent des jeux de données (Open Data), etc.

Agrégation

L'objectif est de préparer une base de données opérationnelles à partir de données initialement hétérogènes et non exploitables telles quelles. Cette étape est essentielle car elle conditionne le travail d'analyse : seules des données nettoyées et cohérentes peuvent délivrer du sens. L'agrégation de données provenant de sources différentes constitue le défi majeur.

Analyse

À ce stade, les données sont interopérables entre elles et prêtes à être analysées. Les applications Big Data varient naturellement d'un secteur et d'un acteur à l'autre. On peut distinguer trois utilisations majeures[12] :

Détecter et optimiser : L'afflux et le croisement de données en temps réel permettent une compréhension fine de l'environnement. La prise de décision est facilitée et les activités peuvent être pilotées plus efficacement.

Tracer et cibler : La granularité des données analysées autorise la découverte et le suivi à un niveau très fin, par exemple l'individu dans le cadre d'une population d'un pays.

Prévoir et prédire : Les vastes données disponibles sur un phénomène ou une population permettent de construire des modèles prédictifs. Leurs capacités sont puissantes mais présentent des limites dans l'anticipation de phénomènes nouveaux. Ce fonctionnement s'inscrit dans les pas du datawarehousing – une technique vieille de plus de trente ans (cf encadré).

(12) Institut de l'Entreprise, Faire entrer la France dans la 3ème Révolution Industrielle, Mai 2014, p.19

NOUVELLES DONNÉES, ANCIENNES TECHNIQUES ?

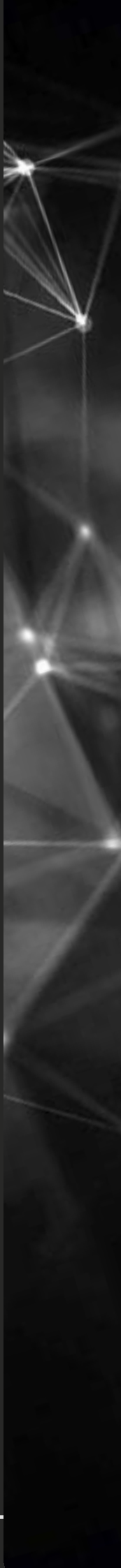
Qu'est ce que le datawarehouse ?

Un datawarehouse (ou entrepôt de données) est un serveur informatique dans lequel est centralisé un volume important de données consolidées à partir des différentes sources de renseignements d'une entreprise. L'organisation des données est conçue pour que les personnes intéressées aient accès rapidement et sous forme synthétique à l'information stratégique dont elles ont besoin pour la prise de décision.

Le datawarehouse s'est développé tout au long des années 1980 simultanément à l'essor de l'informatique dans le monde de l'entreprise. La principale différence entre le datawarehouse et les Big Data ne reside pas dans le fonctionnement mais plutôt dans le volume et la complexité des données traitées. Le Big Data renvoie ainsi aux jeux de données dont le volume dépasse les capacités de collecte du datawarehousing.

On peut même remonter l'origine du Big Data encore plus en arrière. En effet, si les progrès technologiques dans le stockage et le traitement des données ont permis l'émergence du Big Data, les analyses qui en sont déduites restent en partie fondées sur des techniques statistiques anciennes. Pour Christophe Benavent, chercheur en marketing à Paris-10 et membre de notre panel, « une partie Big Data n'est qu'une industrialisation du théorème de Bayes » (ndlr: théorème fondateur de la statistique formulé en 1761).

Il y a confusion entre les différentes étapes du traitement des ensembles Big Data : « Les pratiques corrélatives et prédictives sur les bases de données sont employées depuis plusieurs décennies voire plusieurs siècles. Ce qui change réellement, c'est le volume de données traitées et comment elles sont agrégées. », Samuel Goëta, doctorant à Télécom ParisTech - Sociologie de la production et de la libération de données publiques.



BIG DATA = BIG BANG OU BIG BLUFF?

« Le Big Data, c'est avant tout le marketing qui nous dit : il y a de la valeur à exploiter les données disponibles ».

Arnaud Massonnie, Co-fondateur et Directeur Général de l'agence fifty-five

L'innovation du Big Data est difficile à cerner. Son fonctionnement s'inscrit dans celui de techniques plus anciennes. Il est également difficile de délimiter une définition précise du Big Data. On peut alors se demander si cette révolution ne serait qu'en réalité un phénomène marketing qui comme une bulle retomberait dans peu de temps. Les entreprises sont de plus en plus nombreuses à saisir l'intérêt d'analyser les données clients. Mais cette prise de conscience consiste-elle en soi à une révolution inédite ? Il semblerait plutôt que cet actuel état d'esprit des départements marketing dérive de la nécessité des entreprises de créer de l'attraction autour de projets nouveaux comme le souligne Arnaud Massonnie, « Le marketing s'est emparé du sujet de l'exploitation des données et a réinventé des choses existantes pour « vendre » l'innovation, « la rupture ». In fine, derrière le terme Big Data, il s'agit essentiellement de savoir valoriser et explorer son patrimoine data, au service de l'expérience client ou de la performance opérationnelle. ».

Pour un certain nombre de penseurs du numérique, le pouvoir transformateur du Big Data est une idéologie ou un phénomène de mode. Sans nier la réalité des chiffres, ils adoptent une posture critique qui fournit une base théorique à l'emballement médiatique autour du Big Data, et un contrepoint intéressant dans la littérature foisonnante sur le sujet

Clyde Thompson de Wired, décrit son ouvrage Smarter Than You Think l'influence du «biais de la nouveauté» dans l'appréhension des technologies innovantes. Il explique que les contemporains de l'apparition d'une technologie tendent toujours à perdre le recul nécessaire pour juger le potentiel d'une technologie. Rien de surprenant donc à ce que les analystes rivalisent de milliards de dollars pour estimer le poids du Big Data.

Où se cache la révolution du Big Data ? Sceptiques ou non, le constat de l'entrée dans l'ère de la donnée massive est unanime. Il convient alors d'identifier et de comprendre les leviers de cette transition vers une société où de plus en plus de faits deviennent des informations à valoriser dans des bases de données. Deux facteurs convergent : d'une part, nos comportements et notre environnement produisent plus de données que jamais, et d'autre part, nous disposons de la technologie nécessaire pour stocker et analyser ces océans de données.

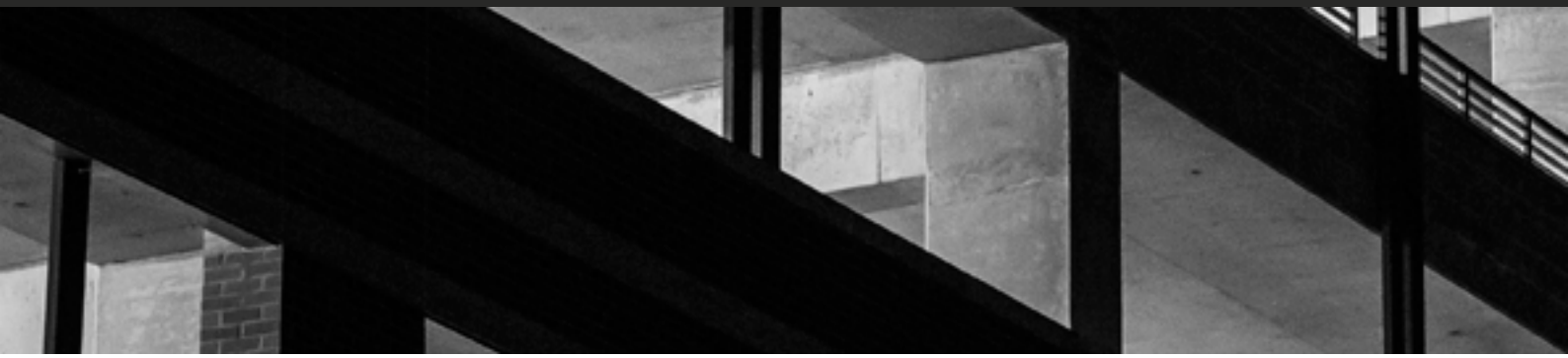
B

BIG DATA :
EN QUOI EST-IL RÉVOLUTIONNAIRE ?

“

Les progrès technologiques ont réduit le coût de créer, capturer, analyser et stocker de l'information au sixième de ce qu'il était.

Rapport Podesta [14]



I. LA RÉVOLUTION PAR LA TECHNIQUE

Les progrès techniques et la baisse des prix associée dans la gestion de la donnée sont les premiers facteurs d'émergence du Big Data. Ces progrès concernent à la fois les logiciels de traitement de données et l'architecture informatique nécessaire à son transit et à son stockage.

« Le tera data existe déjà depuis très longtemps car nous avons toujours stocké les données. Ce qui fait un projet «Big Data», c'est la technologie que l'on utilise. Avec ces technologies, ce qui change, c'est la puissance et la rapidité du calcul qui nous permet d'être davantage « time to market » et de capter de façon plus automatique les comportements clients. »

Ekbel Bouzgarrou, Chief Technologie Officier, Air France KLM

(14) Rapport Big Data: seizing opportunities, preserving values, Executive Office of the President, Mai 2014 - http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Une architecture agile : le « cloud computing »

Avant la popularisation de l'informatique dans les nuages, les données étaient rattachées à l'entrepôt de données (datawarehouse) dans lequel elles étaient stockées. Ainsi, au sein d'une entreprise ou d'une organisation, chaque département possédait son propre silo de données et il était nécessaire de relier physiquement les entrepôts de données pour les mutualiser. Aujourd'hui, le « cloud computing » stocke les données en ligne et les rend ainsi accessibles sans contrainte de lieu ni de temps.

Cette dématérialisation du stockage de données constitue la première couche technologique indispensable pour le traitement Big Data. Un tiers des données seront stockées dans le cloud d'ici à 2020[15], selon Stéphane Grumbach de l'INRIA.

Pour que les données donnent lieu à des opportunités innovantes, il est nécessaire de disposer d'importantes capacités de calcul. Or, ces moyens sont principalement disponibles dans de grands data centers. Le cloud computing permet de dépasser cette difficulté en « louant » puissance de calcul et espace de stockage. En effet, peu d'entreprises et d'organisations possèdent l'infrastructure nécessaire pour traiter du Big Data.

Le cloud computing permet deux grandes innovations. Tout d'abord, une démocratisation du Big Data. Cette innovation devient accessible à des PME ou aux acteurs peu familiers du traitement de données. Mais surtout, le cloud permet l'analyse de ces données en temps réel.

L'écosystème Hadoop : clef d'entrée dans le monde du Big Data

Pour Samuel Goëta, doctorant à Télécom ParisTech, « Avec le NoSQL, Hadoop est le point de départ technologique Big Data ». Hadoop a été créé en 2004 par Doug Cutting qui voulait agrandir la taille de l'index de son moteur Open Source Nutch. Le terme ne désigne pas un logiciel particulier mais un environnement technologique dont le but est de réaliser des traitements sur des volumes massifs de données.

Son fonctionnement se base sur le principe des grilles de calcul : répartir l'exécution d'un traitement sur des grappes de serveurs c'est-à-dire plusieurs ordinateurs indépendants. La grande innovation de Hadoop réside dans cette distribution de l'information. Les architectures plus traditionnelles adossent le traitement de données à une grappe unique.

L'étude de l'institut IDC[16] souligne que l'écrasante majorité (98 %) des entreprises portant des projets Big Data ont recourt à Hadoop. Néanmoins, le prix pour la migration de ses bases de données sur Hadoop reste un frein : 45 % des entreprises interrogées ont dû dépenser entre $100.000 et $500.000 et 30 % d'entre elles, plus de $500.000. Troquer une architecture basée sur un entrepôt de données pour un projet Hadoop

(15) Stéphane Grumbach, Big Data, the global imbalance, octobre 2012 ; www.fr.slideshare.net/slidesharefing/lift12fr-stephane-grumbach
(16) http://www.redhat.com/rhecm/rest-rhecm/jcr/repository/collaboration/sites%20content/live/redhat/web-cabinet/static-files/library-assets/Trends%20in%20enterprise%20Hadoop%20deployments

représente donc un coût elevé. Néanmoins, cette dernière technologie est en moyenne cinq fois moins chère qu'un datawarehouse classique. Ce chiffre comprenant le matériel, le logiciel et le déploiement de l'infrastructure. Sans compter qu'une plateforme Big Data stocke environ cinq fois plus d'informations qu'un datawarehouse traditionnel. Aux données de ventes, sont en effet associées toutes les informations relatives aux comportements des clients en magasin, sur le web ou les réseaux sociaux, etc.

CHIFFRE CLÉ :

15 minutes : c'est le temps que met désormais Suravenir Assurances, du Crédit Mutuel, pour simuler les sommes à provisionner sur trente ans pour ses quelques deux millions d'emprunteurs, grâce aux technologies Hadoop. Hier, il fallait 24 heures pour ce même calcul.

Map reduce : l'architecture analytique

Hadoop est composé d'une architecture de développement dédiée aux calculs parallèles et distribués, nommée MapReduce. Modèle de programmation, elle permet la manipulation des données en très grande quantité, distribuées sur le cluster de nœuds de serveurs qui composent l'architecture de la solution Big Data déployée. C'est ainsi que des données non structurées peuvent faire l'objet d'un traitement analytique et que cette découpe en blocs accélère le traitement, jusqu'à se rapprocher du temps réel.

En fin de processus de l'analyse du Big Data, grâce à Map Reduce, l'analyse des résultats prend la forme de tableaux de bord, de reporting ou de graphiques qui reflètent les interactions ou les corrélations entre les données. L'interprétation de ces sorties passe alors par l'adoption d'un raisonnement prédictif : c'est là le changement majeur opéré par les technologies Big Data.

Une structuration spécifique des bases de données : le NoSQL

Le NoSQL (Not only SQL) est un type de systèmes de gestion de base de données (SGBD). Leur fonction est de manier un grand volume de données et une plus grande échelle (habilité d'un produit à répondre à une mutation d'ordre de grandeur de la demande).

Leur grande innovation est de pouvoir contenir des données hétérogènes. En effet, le NoSQL se distingue des SGBD relationnelles (SGBDR) qui sont construits pour stocker des données normalisées :

(17) http://www.zdnet.fr/actualites/quelle-est-l-activite-sur-internet-en-1-minute-39763269.htm
(18) http://www.liberation.fr/economie/2013/11/03/15-milliards-d-objets-connectes-et-moi- emoi_944254
(19) http://www.lesnumeriques.com/video-poids-lourd-reseau-n9201.html

les champs et les relations entre les tables respectent le même modèle. Le NoSQL est majoritairement utilisé par les sites à grand trafic ou par des réseaux sociaux comme Facebook ou Twitter. Apparu à la fin des années 2000 aux Etats-Unis, le NoSQL a perfectionné les analyses en temps réel, les statistiques et les capacités de stockage. Ce type de base de données permet de soutenir la volumétrie très importante du Big Data.

II. LA MISE EN DONNÉE DU MONDE

« Le Big Data est né de l'explosion de l'information disponible »

Gaëlle Recourcé,
Directrice scientifique, Evercontact

Au delà d'un volume gigantesque, c'est la diversité des sources de données qui donne au Big Data toute son ampleur. Deux leviers principaux soutiennent cette croissance de la production de données : l'effacement de la frontière entre comportements online et offline et la mise à disposition des données publiques. On identifie aujourd'hui quatre grands facteurs responsables de l'explosion de la production de données par nos comportements connectés.

1) Les réseaux sociaux

A chaque minute écoulée, on compte sur internet au niveau mondial : 98 000 tweets, 695 000 mises à jour de statuts et onze millions de messages instantanés sur Facebook. Ce dernier s'occupe également de la gestion de 50 milliards de photos[17].

2) Les objets connectés

Selon la Commission européenne, un Européen dispose en moyenne de deux objets connectés en 2012. En 2015, il en disposera sept. En 2020, il y aurait entre 30 et 80 milliards de nouveaux objets connectés dans le monde[18].

3) Les technologies mobiles

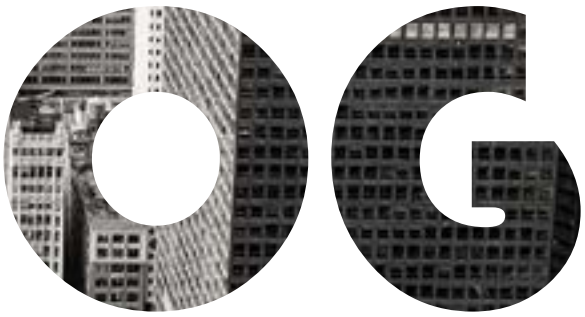
On considère qu'un smartphone génère environ 60 gigabytes chaque année. Si on multiplie ce chiffre par le nombre de smartphones dans le monde soit environ un milliard, on obtient une production de données par an de 56 exabytes soit la totalité de la bande passante consommée en 2013, dans le monde[19]. Le terme Big Data prend alors tout son sens. En 2018, les prévisions estiment qu'il y aura 3,3 milliards de smartphones dans le monde[20].

4) Les comportements numériques scrutés, analysés et stockés

A chaque minute écoulée, on compte sur Internet 700 000 recherches Google, 12 000 annonces sur Craigslist, 600 nouvelles vidéos Youtube et 1 500 articles de blogues [21]. Selon IDC, on comptera en 2016 dans le monde plus de deux milliards d'ordinateurs connectés à Internet[22].

(20) http://www.lefigaro.fr/flash-eco/2013/03/08/97002-20130308FILWWW00351-33-milliards-de-smartphones-en-2018.php
(21) ibid référence 17
(22) http://pro.01net.com/editorial/562702/pres-de-deux-milliards-dordinateurs-connectes-dici-2016/

PDG
D'IMAGE & DIALOGUE GROUP



OLIVIER GUÉRIN

Une des application du Big Data consiste à recueillir et analyser en temps réel des milliers de données diffusée sur Internet

Les outils Big Data marquent-ils la fin des sondages d'opinion classiques ?

Pour comprendre l'opinion, pourquoi aller interroger des personnes, effectuer des enquêtes longues et coûteuses, parfois biaisées par la forme de l'enquête alors qu'il suffit de simplement récolter et analyser les milliers d'avis publiés spontanément et gratuitement sur le web 2.0 ?

Cette analyse de l'opinion sur Internet offre de nombreuses opportunités tant pour la communication d'une organisation ou d'une marque, que pour la sphère civile ou journalistique.

Aujourd'hui, il y a deux manières de procéder pour recueillir et analyser les opinions.

- D'un côté les analyses quantitatives et qualitatives des contenus publiés sur le web à partir du « Text Mining » qui va permettre d'analyser la volumétrie, les thématiques, la tonalité et les sentiments exprimés au sujet d'une organisation, d'une marque, d'une personnalité ou d'un produit.

- De l'autre côté, la technique, souvent très pertinente dans la mise en œuvre de stratégies d'influence, celle du « Link-Mining » qui va permettre d'identifier (ou non) une communauté d'intérêt sur un sujet, de cartographier cette communauté pour mieux comprendre les différentes opinions exprimées, les « suiveurs », « contributeurs » et « influenceurs » de cette communauté.

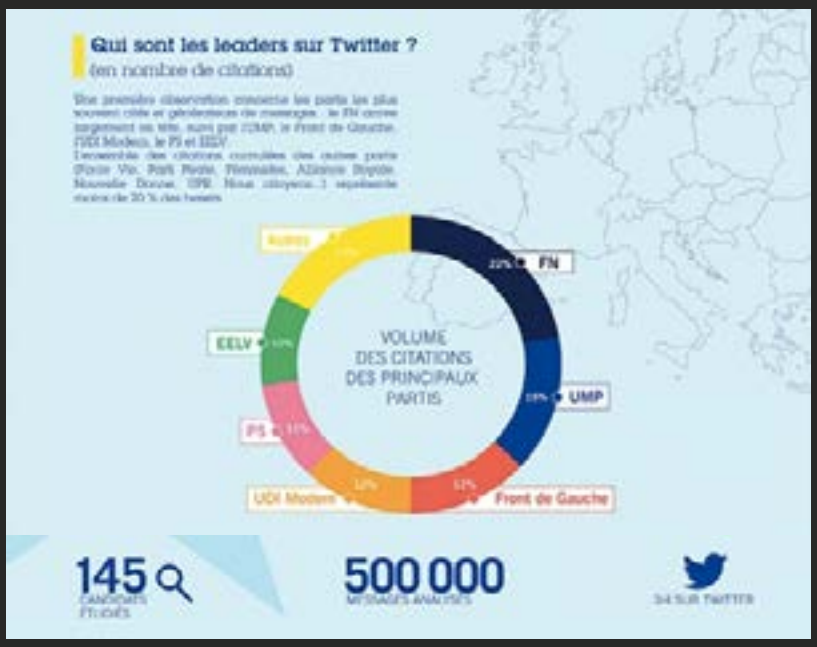
Attention, de très nombreux logiciels dits « d'e-réputation » se vantent d'opérer de telles analyses, mais souvent aux travers de méthodologies ou d'algorithmes peu fiables.

 Date de parution: décembre 2014

Avec l'analyse de cette masse de données à haute valeur ajoutée, nous pouvons aller plus loin, dans la mesure où ces données sont traitées en temps réel. En effet, il est possible, au sein de l'énorme bruit généré par les milliers de conversations, de détecter les signaux faibles, c'est-à-dire, l'information qui va prendre de l'importance rapidement voire même générer du buzz et parfois, des crises. C'est, bien entendu, un moyen de mesurer la notoriété ou l'influence d'une entité sur internet mais surtout au sein d'un environnement, d'une communauté et auprès d'influenceurs.

Ce lien permanent hyper personnalisé et en temps réel avec l'opinion est en train de bouleverser les stratégies et l'organisation des organisations en les obligeant à revoir leurs modèles, leurs produits et leurs services.

Enfin, un cran plus loin, nous pouvons entreprendre d'extrapoler voire de faire du « prédictif » basé sur ces millions de données. Avec le think tank Renaissance Numérique, image & dialogue group a mené de telles analyses lors des dernières élections européennes et municipales. En avril-mai 2014, en recueillant les propos de 145 candidats aux Européennes, il apparaissait déjà que le Front National remportait un écho très important. Le découpage géographique de nos données permettait également de prédire, plusieurs semaines avant les élections, le très fort impact rencontré par ce parti dans certaines régions.



L'effacement de la frontière entre comportements online et offline

Tout d'abord, parce que les individus connectent de plus en plus leur quotidien et génèrent par là même de plus en plus de données facilement exploitables. Ainsi, chacune de nos actions en ligne, du clic au temps parcouru sur une page, des images ou commentaires postés sur les réseaux sociaux, produit une multitude de données.

De plus, les objets connectés et les capteurs intelligents font exploser les compteurs en transmettant un flux permanent de données. Les voitures, l'électroménager domestique, les vêtements et le mobilier urbain deviendront des sources inextinguibles de données. Pour Jean-Luc Errant, directeur de CityzenSciences, « d'ici à 2020, les objets connectées seront le principal adjuvant du Big Data »

En 2013, dans son rapport annuel[23], la société Ericsson annonçait que « le nombre d'abonnements avec un smartphone était de 1,1 milliard fin 2012 et [...] qu'il atteindra 3,3 milliards d'ici à la fin 2018 ». Que ce soit l'envoi d'un message, l'utilisation d'une application, une recherche sur Internet, un coup de fil, un email, une photo ou une vidéo partagée ou téléchargée… le smartphone génère et stocke une masse de données très importante qui peut avoir un intérêt pour une quantité infinie de services : gestion des flux automobiles, offres commerciales spécialisées… Une utilisation bénéfique est son usage par les autorités médicales en cas de pandémie. En Afrique, des scientifiques utilisent ces données pour déterminer l'origine des foyers du paludisme et la localisation des individus malades. La finalité est alors d'optimiser la logistique et la distribution des traitements[24].

(23) Ericsson, rapport annuel 2012 : « Bringing the networked society to life »
(24) http://m.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/

LE TRAITEMENT INFORMATIQUE DE LA LANGUE ET DE L'IMAGE

La structuration de l'information vise à transformer le texte en langage informatique. Les linguistes-informaticiens développent une grammaire de détection qui structure l'information textuelle pour la rendre compréhensible par une machine. L'entreprise Evercontact, par exemple, analyse les signatures des mails pour en extraire automatiquement des fiches de contact. De la même manière, d'un mail, d'un tweet mais aussi d'une image ou photo postée sur un réseau social, il est possible d'extraire une donnée quantifiable traduisant une émotion, un sentiment ou une satisfaction.

« IL FAUT CRÉER DE L'INTELLIGENCE À PARTIR DES OCTETS QUI CIRCULENT »

GAËLLE RECOURCÉ,
DIRECTRICE SCIENTIFIQUE DE EVERCONTACT

Cette discipline, où la France possède une filière d'excellence, participe à la croissance exponentielle de la production de données. Elle permet l'élaboration d'une couche de services intelligents où la donnée créée est mise au service de l'utilisateur.

Le web sémantique (ou langage naturel ou encore le web de données) apparaît aujourd'hui comme la nouvelle quête du Saint Graal des chercheurs en informatique. Il s'oppose au web actuel dit syntaxique.

Cette évolution consiste à rendre le web intelligent c'est-à-dire que les informations ne seront plus seulement stockées mais également comprises par les ordinateurs. Le web sémantique permettrait ainsi

d'agréger différentes données par exemple pour une image : la localisation, la date, l'identité des personnes y figurant, son auteur…

Les possibilités de recherches seraient bien plus nombreuses. Le web de données transformerait également d'autres aspects : recherche intelligente, classement documentaire, e-commerce…

Dans son article « The Prose of the Machines » [25] Will Oremus s'intéresse à l'émergence des robots journalistes – le terme de systèmes-journalistes est plus adéquat. Yahoo ou le site de vente en ligne de voiture Edmunds utilisent ces logiciels pour écrire respectivement des articles sur les résultats de football américain et pour des descriptions d'automobiles.

Ces systèmes ne remplaceront pas les journalistes de sitôt. Le cerveau humain semble – pour l'instant – irremplaçable pour l'écriture d'articles de fond. Ces systèmes ne parviennent pas non plus à adopter un ton humoristique. Ce qui sépare ces systèmes des journalistes n'est pas la qualité d'écriture des machines, c'est la qualité des données. Plus les données sont riches et diversifiées, plus les systèmes seront capables de fournir une analyse fine.

Ils présentent de nombreux intérêts : ils peuvent rédiger de courts articles sur des brèves pour un coût proche de zéro, une fois le système acheté et installé. Le principal logiciel d'écriture automatique Automated Insights a rédigé plus de 300 millions textes en 2013 à un rythme de 9.5 textes par seconde. L'objectif du groupe pour 2014 est de tripler ce chiffre.

(25) Publié le 14 juillet 2014 sur Slate.com - http://www.slate.com/articles/technology/technology/2014/07/automated_insights_to_write_ap_earnings_reports_why_robots_can_t_take_journalists.html

La libération des données

L'Open Data constitue une autre source de gisement de données. Cette dynamique de libération des données publiques est portée par de nombreuses administrations. Malgré de nombreux freins techniques et politiques, ce mouvement représente une opportunité pour obtenir de nouvelles données. Cependant, le volume de ces jeux de données publiques reste faible relativement aux autres sources de données décrites précédemment. Pour Samuel Goëta, doctorant à Télécom ParisTech, ***« ce n'est pas son volume que l'Open data apporte au Big Data mais la fiabilité : les données publiques deviennent des données de référence »***.

Guillaume, fondateur de LMP, startup en stratégie électorale qui développe des modèles prédictifs, souligne que les données fournies par l'INSEE sont un carburant indispensable à son activité : ***« ÉtaLab a fait un travail remarquable pour rendre accessibles à tous des milliers de jeux de données ».***

La libération des données est un levier de croissance. En rendant accessibles certaines informations, des entrepreneurs peuvent y identifier une offre pas encore présente sur le marché. Edouard Schlumberger après un échec à l'examen du permis de conduire, décide de se réinscrire dans une nouvelle auto-école. Il contacte alors les préfectures de police pour accéder aux taux de réussite des différentes agences. Il essuie de nombreux refus alors que ces données doivent être publiques. Il saisit alors la Commission d'accès aux documents administratifs (CADA) pour obtenir enfin les informations qu'il recherchait. E.Schlumberger décide alors en 2013 de créer vroomvroom.fr, un site qui recense le taux de réussite de l'ensemble des auto-écoles françaises et qui – contre rémunération – développe la visibilité de certaines d'entre elles sur le web. L'entrepreneur déclare: ***« L'Open Data, c'est un produit d'appel pour donner envie aux gens de venir nous voir. Monétiser la data seule, ça me parait compliqué. Ce qu'il faut c'est monétiser la résolution d'une problématique. La data, c'est un levier parmi d'autres »*** [26].

Les débouchés de la libération des données ne sont pas uniquement économiques. Elles peuvent également être politiques. La victoire d'Obama en 2008 s'explique – en partie – par le choix innovant de son équipe de campagne de recruter de jeunes talents spécialisés dans la collecte et l'analyse de données. L'équipe démocrate utilisait la technique du data-crunching, en français le « croqué de données ».

En s'appuyant sur des systèmes spécialisés dans le calcul (algorithmes) de haute vitesse capables d'analyser un grand volume de données, le bureau de campagne d'Obama parvenait à identifier précisément les attentes de l'électorat. En effet, le croisement d'informations diverses comme l'âge de l'électeur, son origine ethnique, sa structure fami-

(26) http://lentreprise.lexpress.fr/open-data-liberer-les-donnees-mais-pour-quoi-faire_1534854.html#DdYzEouTCiE7Arm8.99

liale, sa catégorie sociale… permet de dresser le modèle de l'électeur moyen du secteur étudié. Obama pouvait ainsi moduler et personnaliser son discours et répondre au mieux aux attentes des électeurs. Cette technique a également permis un meilleur ciblage dans l'organisation de la récolte de fonds ou dans l'identification des électeurs indécis. Par ailleurs, dans sa politique même, en tant que Président des Etats-Unis, Barack Obama a cherché à donner une vraie impulsion au mouvement de la libération des données par les administrations. En 2009, il a demandé aux organismes fédéraux de publier un maximum de données possibles et créé data.gov.

« Ce site est passé de 47 ensembles de données en 2009 à près de 450 000 provenant de 172 organismes au moment de son 3ème anniversaire en juillet 2012. »
expliquent Viktor Mayer-Schönberger et Kenneth Cukier[27]. Deux startups américaines illustrent les dynamiques de marché et de service qu'engendre l'Open data : « OPower », qui utilise des données énergétiques et climatiques pour aider des familles à réduire leur facture d'électricité et de gaz, et « iTriage » qui aide les Américains à choisir des professionnels de santé correspondant à leurs besoins près de chez eux.

III. LE PRINCIPAL DÉFI TECHNIQUE : L'INTEROPÉRABILITÉ

Les données sont là. Néanmoins, elles n'ont pas été stockées de la même manière. Depuis les années 1980, il existe de nombreux outils de stockage de données dont les infrastructures ne sont pas identiques. Elles ne s'articulent pas entre elles : on parle alors d'absence d'interopérabilité. Un des intérêts du Big Data est le croisement de données. Il serait par exemple intéressant de regarder la relation entre le nombre d'accidents de la route et l'usage des transports publics dans un secteur donné pour observer s'il existe un lien de causalité entre les deux éléments.

Si ces deux jeux de données disposent d'infrastructures différentes, il est impossible d'étudier cette relation en temps réel. Le grand défi à relever pour que la révolution du Big Data réponde à ses promesses est de trouver une architecture interopérable à travers notamment l'adoption de normes communes. L'Organisation internationale de la normalisation (ISO) et plus particulièrement le Comité Technique Commun sur les technologies de l'information (JTC1) est en train de dresser un état des lieux afin d'avancer des solutions sur cette question. La publication de leurs travaux n'a pas encore de date prévue.

L'enjeu de l'interopérabilité ne dépend pas uniquement d'une question de normes. Une des innovations du Big Data est de pouvoir croiser un très grand nombre de jeux de données provenant de bases éclatées. Le problème de l'agrégation et de l'indexation se pose alors.

(27) Big Data: A Revolution That Will Transform How We Live, Work & Think, Viktor Mayer-Schönberger et Kenneth Cukier, mars 2013

PARTIE II

L'ALGORITHME :

CHEF
D'ORCHESTRE
DE LA RÉVOLUTION
BIG DATA

Un des aspects de la révolution Big Data, on l'a vu, repose sur des technologies plus puissantes et accessibles et de l'explosion du nombre de données disponibles. Mais un autre moteur indispensable à cette nouvelle donne est la puissance de formules mathématiques permettant de faire parler les données : les algorithmes.

Au départ une simple formule statistique, les algorithmes permettent aujourd'hui, à partir d'un traitement de données conséquent, d'établir des modèles corrélatifs qui prévoient et préviennent des éléments futurs.

Ainsi, au coeur du Big Data se trouve les algorithmes : tels des chefs d'orchestre, ils mettent en musique des jeux de données massifs. Ils ordonnent, trient, hiérarchisent les gigantesques bases de données, et les rendent intelligibles via un modèle de corrélation ou de prédiction. Pour l'utilisateur, ce sont eux qui transforment des océans de données en des services personnalisés en temps réel.

A COMMENT CONSTRUIT-ON UN **ALGORITHME ?**

“

De gigantesques ramifications dans lesquelles se succèdent des décisions binaires suivant une suite de règles pré-établies.

Christophe Steiner

I. QU’EST CE QU’UN ALGORITHME ?

Formule mathématique, un algorithme désigne initialement la suite de calculs nécessaires pour effectuer une opération complexe. Aujourd’hui l’omniprésence du calcul informatique dans nos vies quotidiennes a élargi cette définition à “une suite d’instructions et de processus requis pour réaliser une tâche”, explique Dominique Cardon, sociologue au sein du département SENSE des Orange Labs, et professeur associé à l’Université de Marne la Vallée-Paris Est.

Christophe Steiner, auteur de Automate This: How Algorithms Came to Rule Our World (non traduit en français) définit les algorithmes comme “des gigantesques ramifications dans lesquelles se succèdent des décisions binaires suivant une suite de règles pré-établies.”

Aujourd'hui, les algorithmes de recherche, de recommandation ou de suggestion structurent notre manière de naviguer sur Internet et la nature même du réseau. Appliqués à une autre échelle, comme celle de la ville, les algorithmes permettent de réguler la circulation des transports en commun.

II. COMMENT ÉLABORE-T-ON UN ALGORITHME ?

Un algorithme trouve donc sa définition et sa formule dans sa finalité. Selon qu'il recommande, ordonne ou déduit, il sera construit différemment.

Construire un algorithme de recommandation

Pour Thibaut Munier, fondateur de 1000mercis-numberly, Administrateur de Renaissance Numérique, un algorithme de recommandation comme celui d'Amazon, qui conseille sur le choix d'un livre en fonction des choix précédents du consommateur, est composé de trois types de calcul distincts qui correspondent à trois questions différentes. Il est étonnant de constater à quel point ces questions relèvent du bon sens humain plus que du savoir scientifique :

Similarité

Quels sont les ouvrages qui abordent une thématique ou un genre similaire à l'ouvrage choisi ?

Complémentarité

Quels sont les ouvrages qui complètent l'ouvrage choisi ?

Diversité

Au sein de cette thématique, quels sont les ouvrages les plus éloignés de l'ouvrage choisi ? Pour fournir la liste de recommandations finales, ces trois questions fondamentales sont pondérées par les informations disponibles sur l'utilisateur (âge, localisation, habitude de lecture, notations d'autres ouvrages).

Algorithme de prédiction

« La puissance et la qualité d'un algorithme dérivent directement de la qualité et de la quantité de données que nous pouvons collecter » Rand Hindi, fondateur de Snips.
Guillaume Liegey, fondateur du cabinet LMP, souligne que l'élaboration de modèles prédictifs se fait en deux étapes :

Identifier les variables et rassembler les données pertinentes.

Celles-ci sont de natures différentes : données publiques fournies par l'INSEE sur les chiffres du chômage, les données électorales passées fournies par le ministère de l'Intérieur et les données politiques publiques ou récoltées sur le terrain (popularité du gouvernement et notoriété du candidat). Il est ensuite nécessaire de nettoyer ces données : colmater les trous, corriger les erreurs et assurer leur interopérabilité.

Affecter les pondérations.

Selon les analyses escomptées, toutes les données croisées dans une même base ne recouvrent pas le même intérêt, d'où la nécessité de les pondérer. À ce stade, l'équipe du cabinet LMP utilise des modèles de régression pour estimer les pondérations de chaque variable à l'aide de logiciels comme Stata ou MathLab. Les données de l'élection précédente sont rentrées dans ce nouvel algorithme et comparées aux résultats connus : les pondérations sont ensuite modifiées jusqu'à ce que les prédictions de l'algorithme correspondent aux résultats réels.

III. L'ALGORITHME AUTONOME GRÂCE AU MACHINE -LEARNING ?

" Aujourd'hui, pour produire un algorithme intéressant, les technologies de machine-leaning doivent être au cœur de son fonctionnement "

Rand Hindi, fondateur de Snips.

Le machine learning est l'innovation mathématique qui permet, une fois encore, de parler d'une véritable révolution par le Big Data. L'apprentissage automatique, ou machine-learning, est la discipline de l'intelligence artificielle qui vise à développer la capacité des machines et des logiciels à apprendre de leurs résultats.

Les algorithmes utilisés pour développer ces systèmes permettent à un système d'adapter ses comportements et réponses de façon autonome, en fonction d'une base de données empiriques.

Pour reprendre l'exemple précédent des campagnes électorales, on parle de machine-learning dans le cas où l'algorithme rectifie tout seul les pondérations des données en fonction du résultat obtenu à l'élection précédente, et rectifie sa formule pour ne pas répéter les inexactitudes repérées dans l'élection suivante. En d'autres termes, l'algorithme apprend et se corrige de façon autonome.

L'apprentissage automatique entre donc pleinement dans les stratégies d'analyse prédictives, puisqu'il considère que les corrélations entre les jeux de données suffisent pour prévoir les nouveaux modèles à appliquer.


BIG DATA

SENIOR DATA SCIENTIST
CHEZ PARKEON

En école d'ingénieur, on apprend aux étudiants les fondements de la théorie de l'information. Rapidement, l'élève connaît les trois niveaux (données, information et connaissance) ainsi que la transition entre ces concepts.

MEHDI CHOUITEN

Machine Learning et valorisation des données

De manière très basique, une information peut être vue comme l'interprétation d'une ou plusieurs données. La connaissance peut être vue comme l'interprétation d'une ou plusieurs informations. Par exemple : Pierre et Paul ont obtenu 9 et 8 respectivement à l'examen de Machine Learning = données 9 > 8 = information Pierre est meilleur que Paul en Machine Learning = connaissance.

Par ce petit exemple, on comprend aisément que les données en elles-mêmes sont d'une utilité très limitée. Leur intérêt réside essentiellement dans l'exploitation que l'on en fait. Le parallèle peut être fait avec la matière première utilisée pour la fabrication d'un objet à forte valeur ajoutée technologique. La valeur d'un smartphone par exemple représente plusieurs milliers de fois celle du plastique et des métaux utilisés pour sa fabrication.

En dehors du stockage et de l'accessibilité des données, la forte valeur créée par le Big Data réside dans l'interprétation et l'exploitation de ces données.

Une exploitation statistique de ces données est souvent faite pour analyser des situations, des comportements d'utilisateurs, des paramètres qui impactent les données et, le cas échéant, essayer d'en déduire des "règles business".

Outre l'exploitation classique offline de ces données, les algorithmes de Machine Learning permettent d'incorporer l'exploitation des données de manière dynamique au système qui permettra d'une part de prédire des situations futures. Et, dans un second temps adaptera automatiquement son fonctionnement à ce qu'il "apprend" non seulement des données à disposition ainsi que des "règles business" établies manuellement par des experts du métier.

Le fonctionnement typique d'un système d'apprentissage se déroule en plusieurs étapes. L'objectif est de

construire en premier lieu un modèle basé sur des données connues et validées. Ce modèle sert à comprendre quel est l'impact des différentes données sur un objectif déterminé (Etape 1 de la figure ci-dessous). Par exemple : pour un client de site de e-commerce, comment l'âge, le genre, le nombre d'amis inscrits sur le site, et le pays de résidence affectent son panier d'achat moyen. Une fois le modèle constitué, il peut être exploité pour prédire le panier d'achat moyen d'un nouveau client (Etape 2 de la figure ci-dessous).

Enfin, en fonction d'objectifs à atteindre et connaissant la manière dont les données influent sur ces objectifs (règles business), nous pouvons décider des actions à mener.

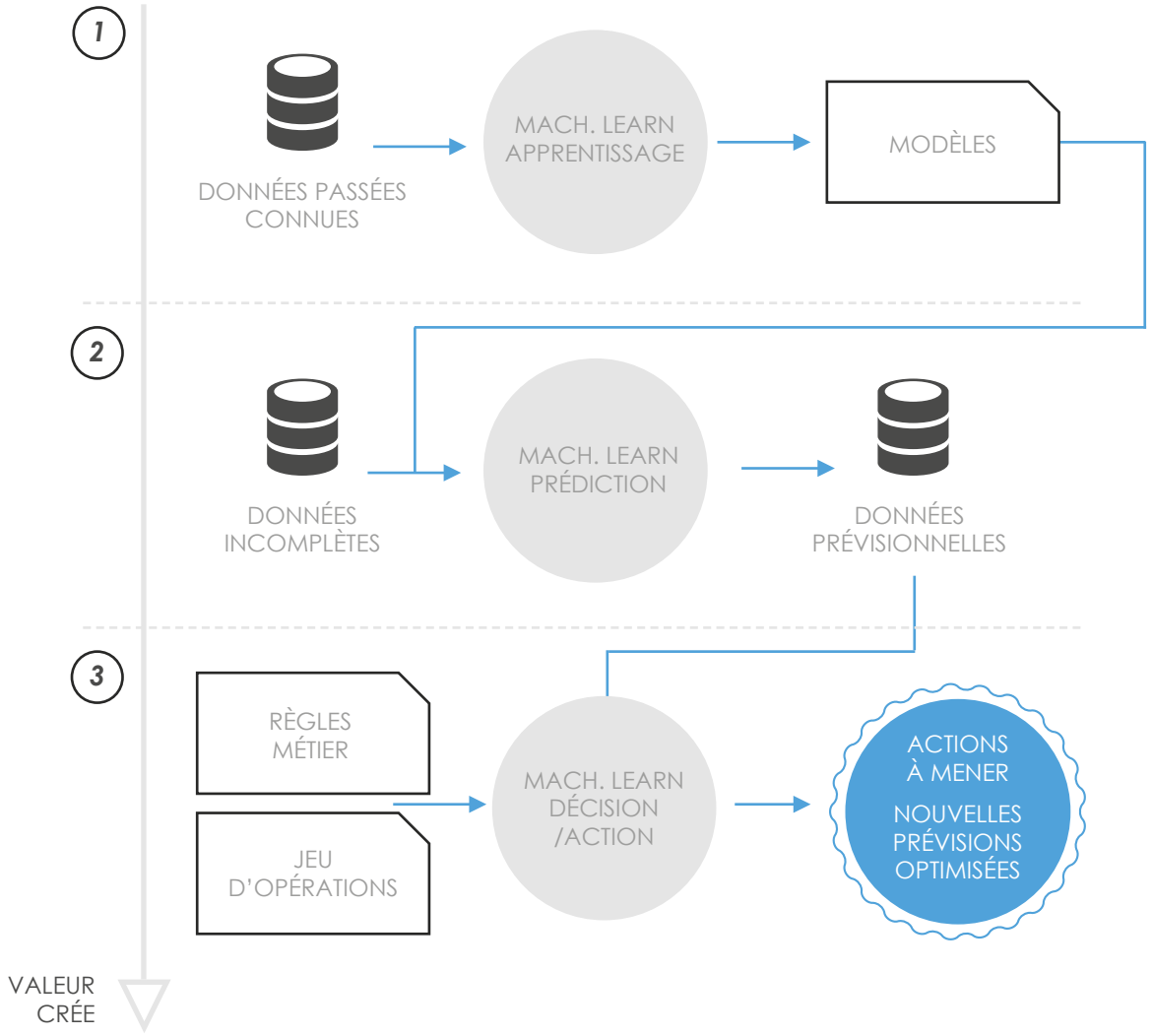
Dans l'exemple précédent, nous pouvons par exemple décider de créer un système de parrainage si nous remarquons que le nombre d'amis inscrits sur le site affecte le panier d'achat moyen. Selon le cas, certaines de ces décisions peuvent être semi-automatisées en mettant à disposition d'un algorithme, un jeu d'opérations possibles associées à des objectifs / contraintes (règles business).

A titre d'exemple, pour un géant du commerce en ligne, des exemples de règles business peuvent être :

- Maximiser le revenu par client dans les pays où la part de marché est supérieure ou égale à 10 %
- Maximiser les parrainages pour les pays où la part de marché est inférieure à 10 %
- Quand les surfaces de stockage sont remplies à 80%, minimiser pour chaque produit la variable (temps de stockage * surface occupée)

Figure 1. Processus typique d'apprentissage, d'exploitation des données et création de valeur

Dans un système disposant de suffisamment de données, il est imaginable de se projeter dans une solution ne contenant que l'objectif à atteindre et où les règles métier sont elles-mêmes déduites par le système. Aujourd'hui, ce sont ces règles qui nécessitent le plus de temps et de ressources humaines (data analysts notamment). Les algorithmes et concepts mathématiques utilisés pour la prédiction sont très similaires d'une tâche à l'autre. C'est notamment la définition des objectifs et des contraintes qui s'avère être la tâche la plus complexe.



Pour les chercheurs en Machine Learning, le Saint Graal serait de disposer - à la manière des tissus cérébraux humains - d'algorithmes génériques qui s'adaptent à chaque tâche et démontrent une capacité à identifier d'elles-mêmes les features les plus pertinentes pour la réussite d'un objectif.

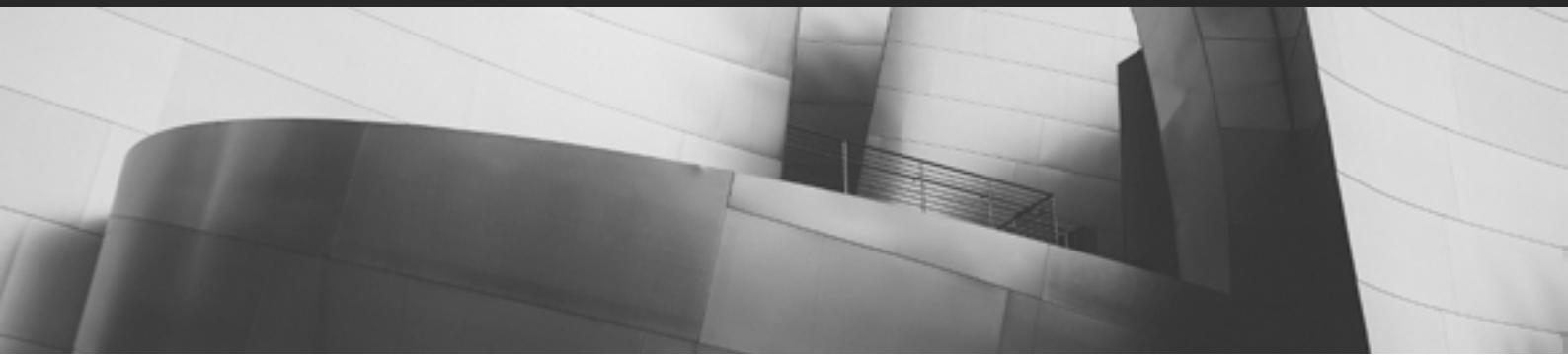
En effet, un cerveau humain est capable d'apprendre en utilisant les mêmes tissus, à distinguer une moto d'un vélo (tâche et features visuelles) aussi bien qu'il peut distinguer une voix d'un bruit tiers (tâche et features sonores).

Les chercheurs poursuivant cet objectif appellent cela le "Deep Learning". Au-delà du Buzzword, les applications sont infinies. Bien que le deep learning n'en soit qu'à ses balbutiements, il est déjà appliqué par Google (notamment dans la classification d'images par le contenu et non plus uniquement en se basant sur les mots clés présents sur la page).

B

VERS « L'ALGORITHMISATION » **DU MONDE ?**

> Si vous étudiez un système d'informations sans tenir compte de sa structure, ses réseaux et ses composantes, vous passez à côté de dimensions essentielles : qui relèvent de l'esthétique, la justice et l'innovation
>
> Susan Leighn



L'agencement de notre fil Facebook, les recommandations d'achats Amazon ou bien le Page Rank Google sont autant d'exemples quotidiens qui mettent en lumière la place des algorithmes pour sélectionner l'information à laquelle nous avons accès (dans un modèle corrélatif) et ordonner l'ordre des choses à venir (dans un modèle prédictif).

La puissance de ces formules mathématiques invite à s'interroger sur le périmètre à leur accorder dans l'agencement du monde par l'humain. Si un algorithme peut gérer mathématiquement les flux des transports urbains, faire baisser la criminalité et la pollution, alors quelle est la place du maire dans la ville intelligente ?

(28) http://peerproduction.net/issues/issue-1/peer-reviewed-papers/caring-about-the-plumbing/
(29) Bruno Latour, La Vie de laboratoire : la Production des faits scientifiques), 1988
(30) http://www.framablogue.org/index.php/post/2010/05/22/code-is-law-lessig
(31) Dominique Cardon, revue Réseaux, Politiques des algorithmes, numéro 177, http://www.cairn.info/revue-reseaux-2013-1-page-9.htm#no2

I. L'ALGORITHME : UNE CONSTRUCTION HUMAINE ET POLITIQUE

« Si vous étudiez un système d'informations sans tenir compte de sa structure, ses réseaux et ses composantes, vous passez à côté de dimensions essentielles : qui relèvent de l'esthétique, la justice et l'innovation » [28] - Susan Leigh

Avant même la naissance d'Internet, Bruno Latour affirmait que la structuration d'un système d'information était «de la politique par d'autres moyens» [29]. En 2001, Lawrence Lessing, dans son célèbre article de « Code is Law », insistait sur la puissance régulatrice du code « dans la manière dont nous vivons le cyberespace » [30].

Aujourd'hui, cette question de la construction humaine et politique de l'algorithme se pose avec plus d'intensité parce que les algorithmes ont pénétré de nombreux domaines de notre vie quotidienne et structurent notre accès à l'information.

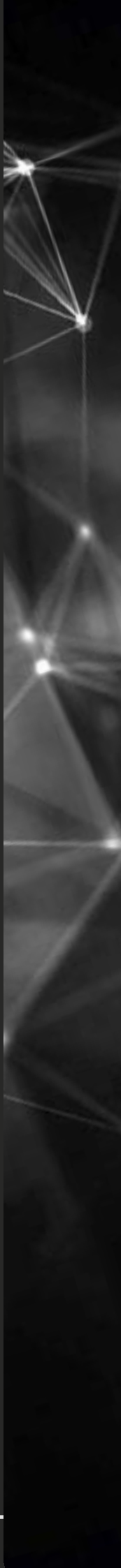
D'un côté, l'approche algorithmique est une nécessité pour rendre intelligible la masse d'informations disponibles, de l'autre, utilisée à mauvais escient, elle peut orienter complètement la connaissance et donc la décision d'un individu. En effet, les algorithmes décident de ce qui est pertinent ou non pour l'utilisateur. De fait, ils déterminent dans le cas d'un moteur de recherche par exemple, ce qui doit être vu et ce qui doit rester caché, ou dans le cas d'un algorithme prédictif, ce qui doit advenir ou non. Pour Dominique Cardon, ce pouvoir est éminemment politique : ***« En décidant de ce qui doit être vu, ils encouragent ou découragent la confrontation et la discussion, participent à la construction de l'agenda public et sélectionnent les bons interlocuteurs »*** [31].

Pour Ted Striphas, auteur The Late Age of Print, la personnalisation des contenus proposés sur Internet grâce aux algorithmes de recommandation, représente un changement décisif dans la culture occidentale. Pour lui, la massification des algorithmes dans la programmation culturelle tend à créer une « culture algorithmique ». C'est bien l'essence même des activités culturelles qui est remise en question : ***« le choix et la hiérarchisation des hommes, des œuvres et des idées »*** [33].

Sur ce point, l'expérience de Netflix est éclairante : la plateforme de diffusion a segmenté son public en 76 897 micro-genres cinématographiques, tels que « films d'action et d'aventure violents et à suspens des années 1980 » ou encore « comédies romantiques à propos de mariages ayant remportées des Oscars » [34].

Plus encore, sa série à succès, House of Cards, a été produite en fonction des données des expériences culturelles de ses utilisateurs[35]. Le scénario et le casting d'House of Cards sont une compilation des préférences de ses utilisateurs analysée sous le crible du Big Data : la série est une reprise d'une série britannique à succès des années 1990 produite par la BBC. Les données de Netflix ont pu lier les goûts de ses utilisateurs pour ce drame politique avec un intérêt pour les films réalisés par David Fincher ou ceux dans lesquels joue l'acteur Kevin Spacey.

(32) https://medium.com/futurists-views/algorithmic-culture-culture-now-has-two-audiences-people-and-machines-2bdaa404f643
(33) Ibid
(34) http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/
(35) http://rebellionlab.com/is-big-data-the-future-starting-point-of-creation/



ÉTHIQUE DE LA DÉCISION À L'ÈRE DE L'ALGORITHME : UN ROBOT A-T-IL LE DROIT DE VIE OU DE MORT ?

Dans son ouvrage Théorie du Drone[36], Grégoire Chamayou donne un exemple extrême du pouvoir des algorithmes. Ce sont des algorithmes qui ont déterminé les cibles des drones américains à la frontière du Pakistan et de l'Afghanistan en scannant les communications des habitants et en évaluant ainsi leur inclinaison à perpétrer des actions terroristes. La liste de ces cibles est in fine ratifiée par la Maison-Blanche.

Dans la zone démilitarisée qui sépare la Corée du Sud de sa voisine du Nord, des robots sur roues de la société Samsung Techwin sont capables de détecter par infrarouges la présence d'êtres humains. Ils sont pour le moment actionnés par des soldats mais disposent d'une technique de tir automatique s'activant à la détection de la chaleur.

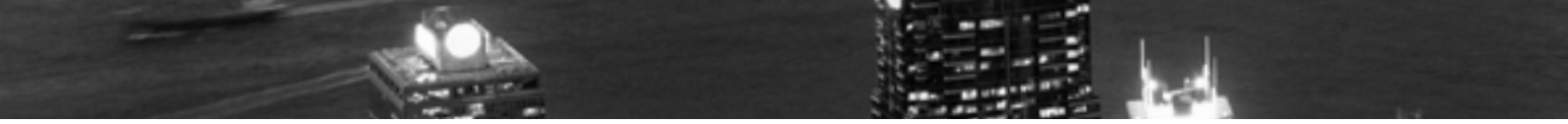
Ainsi, si les drones-tueurs ne sont pas encore des armes de guerre effectives, il convient de s'interroger sur la limite d'autonomie décisionnelle à donner à un robot, surtout quand il peut décider de la vie ou de la mort d'un être humain. C'est la question qu'a soulevé l'ONU en avril 2014 dans le cadre de la Convention sur certaines armes classiques (CCA). D'un côté l'on avance les arguments de réduction des budgets de défense ou de sécurité des soldats, de l'autre on affirme qu'un robot dénué de compassion et d'empathie ne peut avoir droit de vie sur quelqu'un.

« On peut définir comme robot-tueur tout système qui a l'autonomie d'interprétation d'une situation, d'analyse du risque et de prise de décision. Entre son capteur de danger et l'action réalisée, il existe tout une chaîne qui repose sur une sorte d'intelligence artificielle "

Emmanuel Remy
Spécialiste des questions de défense[37]

Les robots-tueurs présentent un cas pratique qui interroge les limites juridiques, philosophiques et éthiques que posent les algorithmes gérés de façon autonome.

(36) Grégoire Chamayou,Théorie du Drone, La Fabrique, 2013
(37 http://www.france24.com/fr/20140514-robots-tueurs-sont-plus-a-craindre-le-cyberespace-ailleurs-armee-drone-ethique/



II. CONNAÎTRE ET PRÉDIRE PAR L'ALGORITHME

Adossés au Big Data, les algorithmes représentent une avancée extraordinaire pour la recherche. Ils permettent d'établir des corrélations qui seraient restées invisibles avec une base de données plus réduite. Ces corrélations peuvent être la base de mise en place de modèles prédictifs.

Ce saut quantitatif a bouleversé par exemple la recherche génétique. La détection d'un des gènes responsables de la schizophrénie était impossible en analysant seulement 3 500 malades mais quand les chercheurs ont pu faire fonctionner les algorithmes avec 35 000 cas, la détection a été très rapide : « il y a un point d'inflexion à partir duquel tout change »[38].

Derrière cette réflexion, se trouve l'idée que le quantitatif, à un certain niveau, modifie le qualitatif. Ce bond est similaire au passage de la physiologie à la biochimie : en changeant d'échelle, de nouveaux phénomènes se font jour et des nouvelles techniques d'interventions apparaissent.

« Avec le Big Data, il s'agit du quoi, et non du pourquoi. Il n'est pas toujours nécessaire de connaître la cause d'un phénomène ; laissons plutôt les données parler elles-mêmes ! » [39]

Ainsi, quand la recherche de la causalité est un échec, il peut être pertinent de se fier au modèle corrélatif. La médecine bénéficie, par exemple, d'une compréhension très fine des mécanismes causaux derrière le diabète, pourtant, elle est incapable de prédire avec précisions les hyperglycémies ou hypoglycémies.

III. NOTRE FUTUR RÉDUIT À UNE FORMULE MATHÉMATIQUE ?

« Le futur n'est pas une déclaration du passé » Benjamin Sarda, Directeur Marketing chez Orange Healthcare

Toutefois, il convient de se demander si toute corrélation peut servir de base à une vérité scientifique. Dans le sens où le modèle déductif de la corrélation répond à la règle de la falsifiabilité de Popper, certainement : mais peut-on baser sur ces corrélations nos hypothèses futures ?

« La donnée brute est un mythe, elle est toujours construite, elle ne vient pas de la nature mais de l'instrument qui la mesure »

Christophe Benavent, chercheur en marketing à Paris-10

Pour la recherche, le Big Data représente un gisement fantastique d'informations. Néanmoins, celles-ci ne peuvent se convertir directement et automatiquement en connaissance. L'utilisation massive de ces données est plus complexe qu'il n'y paraît. En effet, toute donnée comporte une part d'arbitraire, qu'elle provienne de l'instrument de mesure ou de l'organisation qui la collecte ; selon

(38) Manolis Kellis, "Importance of Access to Large Populations," Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice, Cambridge, MA, March 3, 2014,
(39) Big Data: A Revolution That Will Transform How We Live, Work & Think, Viktor Mayer-Schönberger et Kenneth Cukier, mars 2013

le mot de Bruno Latour, « il n'y a pas de données, il y a des obtenus ».

Pour Thomas Lefèvre, médecin de santé publique, ingénieur Mines-Télécom, docteur en sciences, chercheur associé à l'IRIS (CNRS/INSERM/EHESS/Paris 13), deux théories statistiques remettent en question la puissance prédictive des algorithmes.

Les attracteurs étranges :

Dans les années 1960, le météorologue E. Lorenz démontre en trois équations simples l'existence d'attracteurs dit «étranges». Autrement dit, que « certains systèmes sont intrinsèquement sujets à des variations imprédictibles à moyen terme au niveau individuel, c'est-à-dire que deux sujets initialement infiniment proches et semblables peuvent évoluer complètement différemment bien que pris globalement, le système auquel ils appartiennent présente un comportement bien délimité dans l'espace ». Si les algorithmes prédictifs sont performants pour un ensemble d'individus, ils sont incapables de prédire avec certitude ou précision les évolutions individuelles. Cela a des implications très fortes, dans le domaine de la santé notamment.

La malédiction de la dimension :

Richard Bellman, mathématicien américain a mis au point ce concept relativement jeune et encore peu diffusé dans le monde universitaire. Cette théorie démontre que pour des systèmes présentant de nombreuses variables (possiblement, dès une vingtaine de variables), les analyses classiques vont inexorablement tendre vers un résultat moyen et deviendront aveugles aux spécificités de chaque objet. L'outil statistique ne sera plus capable de différencier deux individus : pour une population d'individus définis par de nombreux traits (leurs gènes, leur taille, âge, sexe, etc.), même si beaucoup présenteront des caractéristiques très différentes, l'outil statistique va les assimiler...

En plus de ces limites statistiques, les algorithmes prédictifs, parce qu'ils sont uniquement basés sur des données antérieures à ce qu'ils essayent de prédire, ne sont pas capables d'anticiper des variations dans le futur.

Un algorithme prédictif est extrêmement puissant pour prolonger la courbe mais est aveugle pour anticiper l'innovation. C'est tout le sens du trait d'humour de C&WS, « Si Henry Ford avait demandé à des algorithmes Big Data ce que les clients désiraient, ils lui auraient répondu 'des chevaux plus rapides' »

« La question de fond est celle de la finalité : est-ce que vous voulez comprendre ou est-ce que vous voulez prédire ? »

Thomas Lefèvre, médecin de santé publique, ingénieur Mines-Télécom, docteur en sciencesshercheur associé à l'IRIS (CNRS/INSERM/EHESS/Paris 13)

C

PENSER LA GOUVERNANCE **DES ALGORITHMES**

> Les progrès de l'ingénierie algorithmique, les possibilités d'automatisation qu'elle ouvre (…) nous obligent à construire dès maintenant un corpus d'analyse et de réflexion qui pourra seul nous laisser en situation de comprendre les enjeux de cette deuxième vague d'externalisation : l'externalisation de nos stratégies décisionnelles, émotionnelles, affectives.
>
> Olivier Ertzscheid

Olivier Ertzscheid[40] : maître de conférences en Sciences de l'information et de la Communication à l'Université de Nantes

I. L'ALGORITHME : « HUMAIN, TROP HUMAIN » ?

Parce qu'ils constituent un prisme de lecture et de compréhension du réel de plus en plus présents, les algorithmes et les données doivent faire l'objet de règles de gouvernance réfléchies. Plusieurs exemples mettent en lumière comment une utilisation malintentionnée ou malencontreuse des technologies Big Data peut transformer un algorithme en une machine à discriminer, systémique et silencieuse.

(40) http://ecrans.liberation.fr/ecrans/2014/05/12/bienvenue-dans-le-world-wide-orwell_1015427 (41) http://europepmc.org/articles/PMC2545288/pdf/bmj00275-0003.pdf
(41) http://europepmc.org/articles/PMC2545288/pdf/bmj00275-0003.pdf
(42) http://knowledge.wharton.upenn.edu/article/the-social-credit-score-separating-the-data-from-the-noise/

 Date de parution: décembre 2014

Le risque de l'erreur humaine

Afin d'éliminer le biais humain et de limiter le poids administratif dans son processus d'admission, l'université de médecine St Georges en Angleterre a mis en place en 1988 un modèle algorithmique de sélection des étudiants[41]. Durant les années qui suivirent, le nombre d'étudiants féminins et d'origines étrangères chuta sévèrement, jusqu'à ce que deux professeurs de l'université découvrent la présence de biais discriminants dans la composition de l'algorithme.

En réalité, l'algorithme se basait sur les anciennes données d'admissions de l'université, à une époque où les femmes et les étudiants issus de communautés étrangères étaient minoritaires. L'algorithme a transposé cette inégalité passée et refusait des candidatures. L'université fut condamnée par la justice britannique et coopéra activement pour réparer sa faute.

Il est intéressant alors de constater, d'une part que l'algorithme est bien le fruit d'un cerveau humain, puisqu'il vient même à en reproduire les failles et les limites ; et d'autre part que les formules ne peuvent exister en pleine autonomie, mais nécessitent toujours un contrôle et une gouvernance humaine.

Les algorithmes rendent invisibles des pratiques discriminantes

En avril 2014, le Wall-Street Journal[42] a révélé que des compagnies de crédits américaines utilisaient des données issues des réseaux sociaux pour construire les algorithmes qui déterminent l'accès au crédit et les taux pratiqués. Ces algorithmes complètent le credit score officiel de leurs clients[43]. Deux exemples :

• La startup Neo Finance qui analyse la qualité des connections Linkedin d'un individu pour estimer les revenus futurs et la stabilité de l'emploi de son client ;

• Lenddo, basée à Honk Kong, puise dans les données Facebook et Twitter pour dresser un profil social de chaque client.

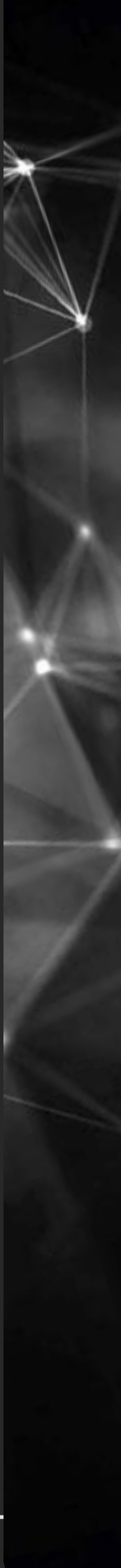
Les pondérations affectées à chaque variable restent inconnues et ne permettent pas de préjuger de l'utilisation qui est faite de ces données. Cependant, l'utilisation de tels algorithmes ouvre la porte à des pratiques discriminantes et intensifie les inégalités.

En effet, ces pratiques permettent aux individus qui disposent des ressources d'améliorer leur accès au crédit en dynamisant artificiellement leurs profils en ligne. N'oublions pas qu'il suffit de quelques dollars pour acheter des followers sur Twitter. Ainsi, des algorithmes discriminants se superposent et amplifient les inégalités existantes[44].

« Quand on réduit une personne à une somme de statistiques et de probabilités, on la transforme en une caricature culturelle qui en dit plus sur les maux de notre société que sur les valeurs et comportement réels de cette personne. » [45]

explique Cécilia Rabess dans un article de The Bold Italic.

(43) Aux Etats-Unis, un credit score est affecté chaque titulaire d'un compte bancaire en fonction de ses revenus et ses mouvements financiers. Il est standardisé et est partagé par toutes les institutions financières. Cette pratique est strictement encadrée par le Equal Credit Opportunity Act.
(44) Gandy, Oscar (2010). "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems," Ethics and Information Technology 12, no. 1, 29–42.
(45) http://www.thebolditalic.com/articles/4502-can-big-data-be-racist



LA NOUVELLE FRACTURE NUMÉRIQUE : CELLE DE LA DONNÉE ?

Aujourd'hui, beaucoup des outils Big Data sont calibrés pour un habitant de Manhattan – qui va générer de larges quantités d'informations. Mais parmi les individus connectés et dont les données sont collectées, nombre d'entre eux génèrent une quantité trop faible de données pour entrer dans le périmètre d'analyse des entreprises qui utilisent les techniques Big Data pour formuler leurs offres et leurs produits. C'est ainsi qu'après l'accessibilité et le haut débit, une nouvelle facette de la fracture numérique se construit : celle du Big Data.

Pour Jonas Lerman, membre du Minister of State américain, il ne s'agit pas simplement de passer à côté de promotions, mais bel et bien d'être pénalisé économiquement et exclu de la vie politique :

« Les magasins n'ouvriront peut être pas dans leurs quartiers, jugés moins attractifs pour les entreprises, tuant dans l'œuf des possibilités d'emploi (…) et ne seront plus dans le périmètre d'intérêt des partis politiques, qui est une condition d'une citoyenneté pleine".

Jonas Lerman suggère que les acteurs publics fournissent des garanties à ces « Big Data's marginalized groups » afin qu'ils ne soient pas exclus de la vie démocratique. C'est paradoxalement aux Etats-Unis, où la protection de la vie privée en ligne est plus faible qu'en Europe, que le débat sur cette nouvelle forme de fracture numérique est le plus vif [46].

(46) Pour aller plus loin dans ce débat : Jonas Lerman, «Big Data and Its Exclusions», Stanford Law Review, septembre 2013, http://www.stanfordlawreview.org/online/privacy-and-big-data/big-data-and-its-exclusions

En se gardant de généraliser les deux exemples précédents, les risques inhérents à « l'algorithmisation du monde » rendent nécessaire l'élaboration de mécanisme de contrôle. Ceux-ci renforceraient la confiance des individus dans le Big Data et serait, à terme, bénéfiques pour l'ensemble des acteurs.

II. TROIS SCÉNARIOS POUR RÉGULER LE BIG DATA

« Notre incapacité à décrire et comprendre l'infrastructure technologique réduit notre portée critique, nous laissant à la fois impuissants et assez souvent vulnérables. L'infrastructure ne doit pas être fantôme. » [47]

- Julian Oliver, membre du collectif artistique berlinois Weise 7 qui a imaginé des Hommes en gris : des hommes qui captent et récoltent les données qu'échangent nos ordinateurs avec les routeurs des hotspots Wi-Fi que nous utilisons, récompensé en 2010 à Ars Electronica.

Une évolution vers davantage de transparence est la condition préalable à la mise en place d'une régulation. Se pose ensuite la question de l'instance de contrôle.

Une exigence : la transparence

L'application du principe de transparence par une ou plusieurs entités de contrôle et régulation aux algorithmes impliquerait que les données utilisées et les calculs effectués soient accessibles afin de voir si les pratiques mises en place sont respectueuses des enjeux de vie privée et d'éthique.

Le premier frein à cette démarche réside dans la complexité technique : la composition d'un algorithme requiert des compétences très élevées en mathématiques et en statistiques pour être décryptée. De plus, l'immense majorité des algorithmes est la propriété des entreprises qui les utilisent ; elle est donc de fait protégée par les lois nationales et internationales de propriété intellectuelle, ce qui complexifie la tâche du régulateur.

« Les algorithmes sont des secrets bien gardés et rendre publiques leurs recettes poserait des problèmes de concurrence et de manipulation »
- Governing Algorithms : a provocation piece[48]

Quel régulateur ? Trois scenarios

« Quand un déluge d'informations financières a dû être géré au début du XXème siècle, sont apparus les comptables et les auditeurs. »

Viktor Mayer-Schönberger
et Kenneth Cukier [49]

Le contrôle des algorithmes à l'aune de la législation en vigueur demande une expertise technique semblable à celle d'un Data Scientist et implique

(47) http://www.internetactu.net/2014/02/26/les-algorithmes-sont-ils-notre-nouvelle-culture/
(48) http://governingalgorithms.org/resources/provocation-piece/
(49) Big Data: A Revolution That Will Transform How We Live, Work & Think, Viktor Mayer-Schönberger et Kenneth Cukier, p.219

la création d'une nouvelle catégorie d'experts. Celle-ci serait strictement encadrée et pourrait agir en interne et en externe des entreprises.

Ces algorithmists[50] répondraient à une demande du marché pour anticiper et éviter les problèmes évoqués plus haut et répondre au besoin de plus de transparence et de sécurité des utilisateurs. Comme dans des domaines aussi variés que la médecine et le droit, les pratiques seraient encadrées par une règlementation et un code déontologique stricts.

a) L'hypothèse d'une régulation par le haut, où les experts-contrôleurs seraient employés par une institution publique, est pertinente pour l'audit des algorithmes à l'œuvre dans les organisations publiques. Ils pourraient s'appliquer de la même manière que des contrôles administratifs ou de sécurité. Cette instance conseillerait les agences de l'État sur les meilleures utilisations possibles des algorithmes.

b) Ces missions de contrôles pourraient également être effectuées par des entreprises agrémentées, à la manière des cabinets de comptabilité ou d'audit. Ces organisations seraient certifiées par une institution de référence, qui pourrait être la CNIL ou une autre institution publique ou ministère.

c) Les entreprises utilisatrices des algorithmes pourraient elles-mêmes assurer le contrôle en interne. À la manière des médiateurs en place dans des grands médias, elles assuraient aux utilisateurs une utilisation juste de leurs données et préserveraient la confiance avec les utilisateurs.

Ces régulateurs posent alors à la puissance publique un nouveau défi : celle d'identifier les compétences nécessaires et de les recruter au juste prix du marché.

L'intensification et complexification du trajet de l'information doivent faire l'objet d'une régulation adaptée. Celle-ci doit prendre en compte le rôle clef que jouent les algorithmes et l'élaboration d'une forme de régulation centrée sur la vérification par des tiers certifiés semble pouvoir permettre de fluidifier le marché tout en préservant la confiance des utilisateurs.

(50) Big Data: A Revolution That Will Transform How We Live, Work & Think, Viktor Mayer-Schönberger et Kenneth Cukier, p.219

PARTIE III

LA RÉVOLUTION INDUSTRIELLE DU BIG DATA :

UN LEVIER DE CROISSANCE DANS DE NOMBREUX SECTEURS

A LE BIG DATA, MOTEUR DE CROISSANCE ET DE MUTATIONS

> Le Big Data est la révolution technologique qui est le nerf de la guerre d'une révolution industrielle en cours
>
> François Bourdoncle

François Bourdoncle, fondateur et CEO de FB & Cie, rapporteur du plan Big Data pour le Ministère du Redressement productif

Le Big Data est l'écho d'une dynamique transversale à tous les secteurs de l'économie qui fait de la donnée la source de valeur principale. Dans ce nouveau paradigme où la donnée devient matière première, les économies traditionnelles doivent questionner leur modèle économique.

À l'instar de l'électricité au tournant XIXème siècle, le Big Data est le déclencheur d'une nouvelle révolution industrielle. François Bourdoncle identifie quatre marqueurs de cette révolution :

I. PREMIER MARQUEUR - L'HYBRIDATION DES MÉTIERS

Issus des deux premières révolutions numériques, les géants de l'industrie numérique possèdent d'importantes réserves de liquidités et une flexibilité organisationnelle qui leur permettent de conquérir de nouveaux marchés bien au-delà de leur activité traditionnelle. Parce que le numérique a pénétré toutes les facettes de notre quotidien, le cloisonnement entre les marchés devient de plus en plus labile. Les entreprises capables de donner du sens à une chaîne de données éparses sont avantagées. C'est par exemple la stratégie de Google qui investit autant dans la domotique que dans la santé ou l'automobile afin de relier toutes ces activités dans un même ensemble et chaîne de valeur.

II. DEUXIÈME MARQUEUR - ÉVOLUTION DES INDUSTRIES TRADITIONNELLES VERS DES BUSINESS-MODEL SOUS FORME DE SERVICE

Conséquence de ces nouveaux entrants sur les marchés traditionnels, les entreprises vont devoir recentrer leur modèle économique autour de l'exploitation de la donnée et sur le service personnalisé qui en découle, plutôt que sur un produit uniforme. Autolib' est l'exemple phare de cette « servicisation » de l'industrie automobile. C'est donc la connexion numérique directe avec le client qui est essentielle pour comprendre les usages et in-fine vendre le service. La connaissance précise des comportements permet de minimiser la prise de risque et devient un avantage compétitif décisif.

III. TROISIÈME MARQUEUR - DES BUSINESS-MODEL QUI SE RAPPROCHENT DE CEUX DES STARTUPS

Les modèles économiques classiques des startups, qui consistent à dégager un très grand volume de liquidités pour l'investir très rapidement sur un nouveau marché, migrent vers l'industrie lourde. La levée de fonds d'un milliard de dollars d'Uber pour s'emparer du marché de la logistique urbaine est significative. Cela est rendu possible par les capitaux auxquels ont accès les fonds d'investissement américains après la croissance phénoménale de l'économie numérique à partir des années 1990.

IV. QUATRIÈME MARQUEUR - LE MODÈLE « FULL-STACK STARTUP »

Ce dernier marqueur correspond à l'évolution des entreprises vers une maîtrise totale de la production. Le meilleur exemple est la décision de Netflix, originellement distributeur de contenu, de produire ses propres séries pour ne plus dépendre d'Hollywood. Là encore, cette évolution implique une compréhension fine des usages et un rapport direct avec le client. Sous l'influence de ces quatre marqueurs, tous les pans de notre économie, toutes les strates de notre société seront contraints d'opérer une mutation profonde pour mettre la donnée au centre de leur organisation. Cette partie identifie la transition numérique par le Big Data de plusieurs secteurs de notre économie traditionnelle.

B LE BIG DATA, UNE RÉVOLUTION QUI TRANSFORME TOUS LES SECTEURS DE NOTRE ÉCONOMIE

> Le Big Data peut en particulier aider à réduire les pertes et le gaspillage au niveau du transport et de la distribution des produits agricoles.

LE BIG DATA ET L'AGRICULTURE

CHIFFRES CLÉS :

20 milliards de dollars : ce sont les bénéfices supplémentaires obtenus par Monsanto grâce à ses technologies Big Data en 2013

10 000 : c'est le nombre d'exploitants français qui utilisent les techniques de l'agriculture de précision

La technologie Big Data intéresse de plus en plus les industries agricoles comme en témoigne le rachat de Climate Corp, entreprise d'analyse des données, par Monsanto. Alors que la population mondiale va dépasser les 9 milliards d'individus d'ici 2050 et que les besoins alimentaires grandissent, le Big Data esquisse une des solutions pour améliorer et optimiser la production agricole mondiale.

Diminuer les risques inhérents à la culture du sol

Avec la baisse des coûts des capteurs connectés, il devient de plus en plus attrayant pour les exploitants agricoles de se procurer des systèmes d'analyse et de prévision des aléas climatiques. Données météorologiques, pollinisation, qualités des sols ou de l'air (température, humidité…), les analyses agricoles gagnent en précision ce qui impacte directement les rendements agricoles. Monsanto estime ainsi à 20 milliards de dollars les bénéfices supplémentaires obtenus grâce à cette « agriculture de précision ».

La société américaine Farm Intelligence travaille par exemple dans le Minnesota avec les producteurs de maïs et de soja pour aider à identifier les signes avant-coureurs de pucerons ou de maladies des plantations.

Par ailleurs, des images aériennes des exploitations prises depuis des satellites ou des drones donnent des informations cruciales sur la croissance des plantes et peuvent, couplées avec des données météorologiques, établir des modèles prédictifs analysant les qualités des cultures, les besoins en eau et engrais, et ce jusqu'à 48 heures à l'avance. En France, déjà 10 000 exploitants utiliseraient les techniques de l'agriculture de précision selon l'Institut national de Recherche en Informatique et Automatique. Pour ce qui est de l'élevage, les applications du Big Data peuvent aller du suivi des animaux, à la détection anticipée d'infections (par exemple des infections mammaires à la couleur du lait) jusqu'à l'adaptation de l'alimentation.

Encourager une agriculture plus respectueuse de l'environnement

La transition vers une agriculture connectée permet de gérer des systèmes d'irrigation intelligents, capables de s'activer automatiquement en fonction des données de précipitations ou de sécheresse du sol.

La société Libellium fournit des capteurs connectés à des vignerons espagnols qui grâce à eux ont amélioré significativement la productivité de leurs exploitations : la productivité des vignobles a augmenté de 15 % et les pesticides ont été réduits de 20 %.

Le Big Data peut en particulier aider à réduire les pertes et les gaspillages au niveau du transport et de la distribution des produits agricoles. Au Brésil par exemple, de nombreuses routes vétustes peuvent être rapidement impraticables pour les camions de transport à la suite de fortes pluies. Les données météorologiques et les cartes des réseaux routiers permettent alors en temps réel de changer les itinéraires et d'améliorer les réseaux de distribution, en minimisant les pertes.

LE BIG DATA ET L'ASSURANCE

CHIFFRES CLÉS :

800 millions d'euros : C'est la somme investie par Axa en 2014, sur trois ans, dans des projets digitaux au niveau mondial

67% : c'est le nombre d'acheteurs d'assurance qui, aux Etats-Unis, ont obtenu leur tarif en ligne

L'assurance, dont le modèle économique est basé sur la gestion du risque et donc la connaissance des individus et les analyses statistiques, est logiquement un des secteurs les plus impactés par l'essor du Big Data.

En effet, l'hyperconnectivité des individus et ainsi la récolte de données massives permettent une connaissance très précise des modes de vie de chacun : l'hygiène de vie de l'individu peut être calculée grâce aux applications quantified self, la qualité de sa conduite est limpide si la voiture est connectée ou géolocalisée, ou encore, la gestion du foyer est transparente si l'accès est donné aux compteurs intelligents d'eau ou d'électricité. Avec cette nouvelle volumétrie de données, c'est la matière première de l'assureur qui évolue en profondeur. Si l'assurance accède à ces données, il lui sera alors facile de faire évoluer ses produits, ses garanties et ses méthodes de gestion de risques pour envisager une offre extrêmement personnalisée en fonction du profil de l'assuré.

Cette motivation explique pourquoi les assurances réfléchissent toutes aujourd'hui aux moyens de mettre en place une collecte d'informations massives sur le mode de vie de leurs assurés. Avec les objets connectés et les applications santé, fini les longues fiches de renseignements et les questionnaires à remplir par l'assuré ! On peut alors imaginer de nouvelles offres, aux tarifs presque personnalisés, pour assurer nos risques quotidiens, amenées à évoluer en temps réel, en fonction de la vie quotidienne de chacun.

Au-delà de la tarification au plus proche des risques, le Big Data offre la possibilité d'effectuer une lutte contre la fraude à l'assurance redoutablement efficace en identifiant de manière automatique les comportements anormaux.

Les assureurs auto ont été les premiers à explorer les opportunités Big Data avec des formules « Pay as you drive ». Les assureurs américains Progressive et Allstate viennent ainsi de lancer des offres où le calcul de la prime prend en compte non seulement le nombre de kilomètres parcourus mais aussi une évaluation du comportement de l'assuré à travers des données comme l'heure à laquelle il prend la route, le nombre de freins

brusques, le nombre d'accélérations rapides et la vitesse. Restituées sur un espace privé en ligne, ces données une fois analysées donnent lieu à des tarifs minorés ou majorés.

Les défis à relever pour faire entrer l'assurance dans l'ère Big Data

La collecte de la donnée

Puisque celle-ci constitue la matière première du marché de l'assurance, cette dernière doit s'atteler à nouer des partenariats avec des entrepreneurs des objets connectés ou applications mobiles pour collecter les données à la source : à l'instar du partenariat Withings / Axa noué en 2014 qui propose d'équiper gratuitement leur client de bracelets connectés.

Certification de la donnée

Si les sources de données sont multiples à l'ère du Big Data, établir leur traçabilité est de plus en plus complexe. Aussi, le cheminement de la donnée, son changement de statut, peut se révéler être un vrai casse-tête pour l'actuaire chargé de vérifier et certifier la donnée. Pour Omptimind Winter[51], l'actuaire de demain «pourrait devenir le correspondant du régulateur sur les questions de conformité dans le recueil et l'usage du Big Data ».

Protection de la vie privée

Dans ce secteur particulièrement, le déploiement du Big Data doit être encadré par le régulateur à l'instar de la CNIL qui aujourd'hui regarde d'un œil attentif les pratiques des applications Quantified self ou des boîtiers connectés aux voitures donnant lieu à des offres Pay-as-you-Drive. On peut très vite imaginer, par exemple, les dérives inégalitaires d'un système de santé où les assurances se fondent sur l'analyse des données personnelles pour finaliser les tarifs de prise en charge.

L'assurance nouvel acteur pour la prévention

Ce pouvoir de collecte et d'analyse des données, et les opportunités qu'il fait naître, invite naturellement les assureurs, à même de mieux comprendre les risques encourus à partir d'un comportement type, à devenir des acteurs de la prévention. Ainsi, l'ère du Big Data pour l'assureur rime-t-elle avec l'émergence de nouvelles responsabilités ?

« L'assurance doit prendre conscience d'elle-même comme d'un agrégateur et d'un gestionnaire de données. L'assurance transforme des données en services de protection. Le service de l'assurance consiste en effet à isoler dans la masse des données disponibles celles qui ont un caractère prédictif et peuvent servir à organiser des services de protection contre les conséquences patrimoniales d'événements futurs. »

François Ewald, Professeur honoraire au CNAM et International Research Fellow de la Law School de l'Université du Connecticut[52]

(51) Optimind Winter, Dossier technique d'information «Big Data», Octobre 2013, http://www.optimindwinter.com/wp-content/themes/optimind/upload_dbem/2013/10/201310_Dossier_technique_Optimind_Winter_Big_Data.pdf
(52) Entretien avec François Ewald, « Big Data et assurance », Institut Montparnasse, http://www.institut-montparnasse.fr/big-data-et-assurance/

LE BIG DATA ET LA CULTURE

CHIFFRES CLÉS :

Le magazine Forbes a estimé à 0,03 dollar la valeur d'un goût individuel exprimé sur un lien culturel 53

Aujourd'hui, 52 % des commentaires sur Facebook portent sur les programmes diffusés à la télévision 54

Sur Netflix, 75 % des programmes consommés le sont grâce au système de recommandation. Près de 800 ingénieurs travaillent, au sein de l'entreprise, à l'élaboration et l'amélioration de ces algorithmes de recommandation.

La culture comprend deux dimensions. La première relève de l'intime ; nos pratiques culturelles dévoilent nos goûts, nos hobbies, nos aspirations… notre identité, en somme. La culture renvoie également à des pratiques sociales et communautaires. La donnée personnelle culturelle possède ainsi une valeur particulière : « La donnée personnelle culturelle renferme des informations contextuelles fortes et permet de qualifier de façon assez précise le pouvoir d'achat de l'être numérique mais aussi de prévoir son comportement » [55].

Aux Etats-Unis, des chercheurs de l'université Stony Brook (New York) ont développé un algorithme capable de prédire avec 84 % le succès d'un livre. Le principe ? Le programme se base sur l'analyse d'autres romans qui ont été choisis pour leur succès littéraires (récompenses/critiques). A l'instar des logiciels anti-plagiat, le système étudie le degré de similarité entre la base d'étude et l'œuvre en question.

Selon le programme, les éléments qui font d'un livre un succès sont le choix des prépositions, noms, pronoms, déterminants et adjectifs (à l'inverse, les mauvais livres utiliseraient plus de verbes et d'adverbes qui renvoient à des mots d'actualités, des clichés, des lieux communs). Les bons livres aborderaient plus le vocabulaire de la réflexion que celui de l'action…[56]

Le Big Data au service d'une meilleure diffusion pour une grande interaction avec le public

Les acteurs de l'industrie culturelle ont un double-défi à résoudre : instaurer et assurer une relation privilégiée avec ses clients. Le Big Data leur permet d'atteindre cet objectif. En scrutant et en analysant les réseaux sociaux – principalement – l'industrie culturelle est en mesure d'observer quelles sont les attentes du moment, mais aussi de les anticiper.

Les données personnelles culturelles permettent également de prolonger l'expérience culturelle et la relation entre acteurs et usagers cultu-

(53) http://www.strategies.fr/etudes-tendances/tendances/224438W/le-big-data-au-service-de-la-culture.html
(54) Comportements culturels et données personnelles au cœur du Big data – EY & Forum d'Avignon, 2013 : p.12
(55) Ibid
(56) http://substance.etsmtl.ca/un-algorithme-pour-predire-le-succes-litteraire-la-maniere-de-triz/

rels. Après une expérience culturelle que ce soit un spectacle ou la visite d'une exposition, la collecte et le traitement des informations relatives à l'évènement donnent lieu à la création de communautés web ou de services complémentaires.

L'utilisation du Big Data dans le tourisme : l'exemple à suivre.

En 2012, le Comité Régional de Tourisme Côte d'Azur et Orange ont quantifié et modélisé les déplacements des touristes dans la région. En utilisant les données de ses clients notamment et en les croisant avec les informations géographiques de l'I.G.N, Orange est parvenu à produire des analyses quant aux déplacements des touristes, le temps passé, les lieux les plus visités, nombre de nuitées… La finalité de l'opération était d'optimiser l'expérience touristique : emplacement des structures d'hébergement, de restauration mais aussi de s'adapter aux coutumes nationales des visiteurs[57].

Cette initiative peut être reprise pour d'autres zones touristiques en France. Son principe peut également être appliqué à une échelle plus modeste. Un musée pourrait analyser de la sorte les données émises par ses visiteurs afin d'améliorer sa logistique d'organisation (estimation en temps réel de l'attente pour l'achat des tickets) d'optimiser le parcours de l'exposition (rendre plus accessibles les œuvres qui plaisent le plus) ou l'emplacement de ses services annexes (boutiques, restaurants).

Le Big Data : nouveaux gains pour l'industrie ?

L'écosystème de la culture voit ses sources de financement tarir à cause du contexte économique difficile. Largement dépendante des deniers publics par le passé, la culture doit trouver de nouvelles sources de financement. Le rapport EY & Forum d'Avignon 2013 présente une nouvelle piste de réflexion intéressante :

« Un projet culturel pourrait demain valoriser, au moment de son financement, sa capacité à générer des données pour le distributeur, au même titre qu'il peut générer des ventes. Les plans de financement de projets cinématographiques ou discographiques pourraient, par exemple, intégrer la valorisation des données nouvelles collectées : un producteur exécutif céderait à un coproducteur le droit d'administrer la communauté de l'œuvre créée, et les revenus publicitaires éventuellement générés. »

(57) http://reseauculture21.fr/wp-content/uploads/2014/07/EtudeATELIER_FA_2013.pdf

UNIVERSITÉ PARIS DAUPHINE

VICE PRÉSIDENT
DE RENAISSANCE NUMÉRIQUE

Le commerce électronique, qu'il soit mobile, desktop ou sur tablette, génère quantité de données qui sont la base des web analytics que tout e-commerçant se doit de suivre avec attention.

HENRI ISAAC

Le commerce électronique, qu'il soit mobile, desktop ou sur tablette, génère quantité de données

qui sont la base des web analytics que tout e-commerçant se doit de suivre avec attention. Si les volumétries conséquentes de données ont longtemps été l'apanage des principaux sites d'e-commerce, le développement constant de ce secteur amène de nombreux sites à gérer des volumes croissants de données liées au trafic, à la navigation, à l'achat, à la relation client.

L'arrivée des technologies Big Data change radicalement la donne dans ce secteur et ce sur plusieurs problématiques propres au commerce électronique : la conception des interfaces marchandes, la recommandation et la personnalisation, le pricing, la gestion du catalogue.

LE BIG DATA POUR OPTIMISER LES INTERFACES MARCHANDES

Afin d'améliorer les interfaces de sites marchands, de très nombreuses sociétés proposent d'utiliser des tests A/B qui visent à exposer deux groupes de clients à deux pages différentes afin de déterminer la plus efficace en terme de souscription ou de vente ou de tout autre problématique d'ergonomie. Si cette méthodologie présente un intérêt avéré, elle présente en revanche des limites dès lors que l'on introduit simultanément plusieurs changements sur une page.

Si les méthodes du Big Data sont depuis longtemps utilisées dans la recommandation de produits et d'offres (cf. § suivant), elles investissent désormais le champ de la conception d'interfaces en analysant simultanément en temps réel des milliers - voir des millions- de parcours de navigation en y appliquant des analyses statistiques afin de détermi-

ner l'interface la plus performante. La société Content-Square[58] est très illustrative des méthodes du Big Data appliquées à la conception d'interface. Un des principaux apports des méthodes Big Data est leur capacité à fournir des éléments de réponse dans des délais fortement réduits (quelques jours versus plusieurs mois).

L'adaptation de l'offre d'un site et les algorithmes de recommandation

Un des enjeux du e-merchandising est de fournir une assistance à la vente sans vendeur. Une méthode pour y parvenir consiste à détecter un client et à adapter l'offre de produits au profil, à la navigation. Longtemps apanage de sites marchands aux ressources importantes, les algorithmes de personnalisation et de recommandation deviennent accessibles avec des offres SaaS nombreuses (Sparkow, Tynyclues, Nosto, Ezako, NuuKik, Target2Sell, PlanetWorld, etc.).

Les moteurs de recommandation s'appuient tous sur de l'apprentissage artificiel (machine learning) afin d'apprendre des comportements des internautes. La disponibilité d'Apache Mahout (https://mahout.apache.org), logiciel open-source de machine learning issu du projet Hadoop, va encore accélérer le déploiement du Big Data sur de nombreux sites marchands.

La gestion dynamique des prix (Dynamic Pricing)

Boomerang Commerce[59] permet aux e-commerçants d'ajuster leurs prix en temps réel en fonction de ceux d'Amazon et d'autres e-commerçants. Le logiciel parcours les sites de la concurrence et analyse les prix d'un produit donné. Il peut ensuite ajuster le prix automatiquement, à la hausse ou à la baisse. Il peut aussi faire des suggestions au lieu d'un ajustement automatique, et par exemple proposer d'augmenter un prix par rapport à la concurrence, afin d'accroître des marges dans une catégorie de produits. Ainsi, les e-commerçants ont la possibilité d'automatiser leurs prix en fonction de nombreux facteurs, comme leur stock, les changements de prix de la concurrence, l'heure du jour ou la météo. Boomerang Commerce offre aussi de tester différentes stratégies de prix via un indice de perception des prix pour un produit donné. Il permet aussi d'optimiser les prix en fonction des canaux de distribution utilisés.

Le Big Data est utilisé pour évaluer l'impact d'un changement de prix sur le chiffre d'affaires, et ainsi aider chaque entreprise à établir la meilleure stratégie de prix en fonction de ses objectifs. Les volumes des catalogues (SKUs) et le nombre de concurrents à surveiller en temps réel nécessitent de recourir à des technologies de Big Data pour le stockage des données et des algorithmes d'apprentissage (machine learning) qui reposent ici sur la théorie des jeux[60].

(58) http://www.content-square.fr/
(59) http://www.boomerangcommerce.com
(60) http://www.ecommercebytes.com/cab/abn/y14/m07/i18/s02

Performance des catalogues e-commerce

Les catalogues des e-commerçants peuvent comporter de très nombreuses références de produits générant des bases de données de plusieurs milliers, centaines de milliers, voire plusieurs millions de produits dans le cas des marketplaces[61].

Les bases de données des progiciels e-commerce doivent donc stocker des fiches-produits avec des données très hétérogènes – un livre ne se représente pas avec les mêmes attributs qu'un aspirateur. En outre les caractéristiques des produits peuvent évoluer dans le temps. L'une des techniques classiques employée pour répondre à cette problématique dans une base de données relationnelle est de proposer un modèle dit Entity-Attribute-Value[62].

Le principe de cette modélisation est de séparer les données fixes du produit de ses attributs qui sont stockés dans des tables spécifiques (5 tables au total dans le progiciel E-commerce Magento, par exemple).

Ce modèle présente un avantage important par rapport au stockage à plat des données lorsqu'il s'agit d'opérer la mise à jour du modèle de stockage des produits puisqu'il n'est pas nécessaire de modifier la structure des tables de stockage. Cette opération est en effet très difficile à opérer dès lors que le volume contenu dans une table devient important. Cela rend très clairement le passage à un modèle à plat difficile à envisager. En contrepartie, le modèle EAV présente un coût important pour certaines opérations basiques. Ainsi pour obtenir un produit ou une liste de produits, il est nécessaire de procéder à des opérations de jointure assez lourdes pour récupérer l'ensemble des attributs du produit. Sur un petit catalogue, c'est tout à fait acceptable. Mais dès lors qu'il s'agit de manipuler une base de plusieurs millions de produits, le coût de ces opérations devient vite prohibitif. Sur une base de 5 millions de produits avec une moyenne de 20 attributs produits, on effectuerait des opérations de jointure sur une centaine de millions de lignes.

Dès lors, le recours à une base NoSQL documentaire comme MongoDB est une solution idéale pour améliorer les performances. La SSLL Smile a ainsi réalisé une adaptation du progiciel e-commerce Magento en y intégrant MongoDB[63]. Cette solution de base de donnée NoSQL permet de gérer des catalogues de très grande envergure avec des performances d'affichage (temps de réponse) et de recherche optimisées[64] .

(61) A titre d'exemple, un site come Pêcheur.com gère un catalogue de plus de 154 000 produits, Amazon.fr possède à l'été 2014 plus de 119 millions de références et Amazon.com 253 millions. Source : Export.com
(62) voir une présentation pédagogique de ce modèle : http://www.magentix.fr/divers/modele-eav-magento-database.html
(63) disponible sur GitHub https://github.com/Smile-SA/mongogento
(64) http://www.ecommerce-performances.com/

LE BIG DATA ET LA FINANCE

CHIFFRES CLÉS :

98 % : c'est le pourcentage de baisse du coût du stockage pour un gigabit de data financière. Ainsi, une entreprise opérant plus de 20 millions d'opérations quotidiennes fait passer le coût de stockage de 17$ à 21 cents par gigabit avec une architecture Hadoop 65 .

Sur les marchés européens et américains, sept transactions financières sur dix sont automatisées : le Big Data, par le truchement du Trading à Haute Fréquence, sont au cœur des organismes financiers.

« Les entreprises financières sont parmi les premières à avoir compris que la donnée était la nouvelle source de valeur»-Stéphane Buttigieg, Directeur général adjoint Institut Louis Bachelier

Dès les premiers pas de l'informatique dans les années 1980, le monde de la finance a tenté de maîtriser les nouveaux flux d'informations numériques. Ce qu'on appelait alors le Business Intelligence répondait aux mêmes problématiques que le Big Data. La différence fondamentale réside dans le volume alors traité.

Les banques, les sociétés d'assurances et les entreprises financières sont les premières à avoir embauché en masse des profils type Data-Scientists, notamment à travers la formation d'excellence Polytechnique – ENSAE qui est devenue la référence mondiale. Aujourd'hui, le secteur bancaire est le premier client des entreprises informatiques qui fournissent l'infrastructure de gestion du Big Data.

La vente ou l'achat automatisés d'actifs boursiers en l'espace de quelques nanosecondes est une pratique en place à travers les places boursières du monde entier. Pour certains, elle favorise la spéculation à outrance et est le reflet d'une finance déconnectée des enjeux de l'économie réelle, alors que pour d'autres, le Trading Haute Fréquence est un moyen efficace de dégager les liquidités nécessaires au marché.

Un algorithme ne peut pas fonctionner seul : il doit être régulièrement contrôlé, corrigé et réorienté ! Pour autant, le rôle de l'humain reste primordial et la machine ne sera jamais que l'écho de ses choix, comme le souligne Stéphane Buttigieg « Automatisées ou pas, les décisions prises par la machine sont toujours le reflet d'une intelligence humaine ».

Les spéculations sur les matières premières au début des années 1990 ou encore la crise des SubPrimes de 2008 sont le fruit de décisions humaines et leurs mécanismes ne sont pas liés à la généralisation du Trading à haute fréquence. En 2011, le piratage du compte Twitter de l'Associated Press par

(65) http://inside-bigdata.com/2014/10/13/adopting-big-data-finance/

des hackers syriens et la diffusion d'un prétendu attentat contre la Maison-Blanche a berné les algorithmes de Wall Street : en l'espace de quelques instants, le marché s'est effondré et a nécessité l'intervention humaine pour revenir à la normale.

LE BIG DATA ET LA GESTION DES RESSOURCES HUMAINES

CHIFFRES CLÉS :

22 % : augmentation de la performance des centres d'appels Xerox qui ont « automatisé » leur recrutement

4 millions : le nombre de profils de développeurs analysés et classés par l'algorithme de Gild

Après l'Organisation Scientifique du Travail de Taylor, le Big Data est la nouvelle révolution des techniques de travail et d'optimisation de la chaîne de production.

Plus de productivité grâce aux données

« Nous assistons à une Révolution de la mesure, et cette révolution va transformer l'économie de l'organisation et l'économie personnelle » - Erik Brynjolfsson, directeur du Centre des affaires numériques à la Sloan School of Management du MIT.

Le Big Data provoque un changement d'échelle dans les études du comportement des travailleurs : de la fréquence des emails envoyés au moindre clic de souris, l'intégralité de l'activité de milliers de travailleurs peut être analysée et mise au service de l'efficacité de l'entreprise. Ces données nouvelles établissent des modèles corrélatifs qui identifient les variables explicatives de la performance des employés.

Bank of America a équipé 900 de ses employés de badges développés pour étudier leurs mouvements et interactions afin de comprendre la façon dont ils travaillent. Résultat : une productivité augmentée de 10 %[66]. La société Citizen invite ses employés à renseigner leur régime alimentaire, leurs activités sportives et leur temps de sommeil afin de déterminer comment augmenter leur productivité au travail.

L'utilisation de nouvelles technologies peut toutefois se heurter à des barrières juridiques en France. Légalement, la surveillance des salariés répond à un cadre légal strict. Si elles peuvent être placées dans un couloir, les caméras de vidéosurveillance par exemple ne peuvent servir à espionner un employé. Un dispositif tel que mis en place par Bank of America serait sanctionné par la CNIL en France.

L'automatisation du recrutement : une nouvelle méritocratie ?

(66) http://internetactu.blogue.lemonde.fr/2013/05/03/le-recrutement-et-la-productivite-a-lheure-des-big-data/

« Nous allons bientôt assister à la prolifération des systèmes de recrutement automatique qui feront automatiquement correspondre les candidats aux emplois. Imaginez qu'au lieu de recevoir des recommandations de films de Netflix vous receviez des propositions d'emploi de Monster ou LinkedIn - et que ces emplois soient effectivement bons pour vous. » – Tomas Chamaro-Premuzic, contributeur à l'Harvard Business Review

"I no longer look at somebody's CV to determine if we will interview them or not," - Teri Morse, responsable des ressources humaines à Xerox Services Xerox, leader du marché des imprimantes, a confié aux algorithmes de la startup spécialisée Evolv le remplacement de 22 000 opérateurs pour ses centrales d'appels[67]. Afin de prédire quels seront les employés les plus fidèles et les plus performants, Evolv a croisé les résultats de tests de personnalités avec les données fournies par Xerox sur les comportements de ses employés en central d'appel. Xerox bénéficie alors d'un portrait de l'employé idéal et peut automatiser sa décision en fonction de l'écart du candidat par rapport à cet idéal type. Les résultats contredisent les idées reçues : par exemple, une expérience préalable dans un centre d'appel ne conditionne par nécessairement une performance plus haute.

Cependant, l'automatisation du recrutement exclut les candidats qui ne se trouvent pas dans le périmètre des outils scannés par le recruteur : un candidat qui ne dispose pas de profil Linkedin est-il une moins bonne recrue que celui qui l'alimente ?

LE BIG DATA ET L'ÉCOSYSTÈME SPORTIF

CHIFFRES CLÉS :

25 par seconde : c'est le nombre d'informations qu'émettaient les joueurs de la Mannschaft équipés de matériel connecté pendant la Coupe du monde 2014

L'émergence du Big Data dans le monde du sport constitue une étape clef dans le dépassement des limites et des exploits sportifs. Le Big Data permet d'affiner avec précision les mouvements, les tactiques et les prouesses des joueurs sur le terrain, confortant la tendance contemporaine à un « culte de la performance » comme l'observe le sociologue Alain Erhenberg. Les sportifs, médias et publics cherchent toujours plus à quantifier, chiffrer et objectiver les performances sportives, les analyser sous le crible de la puissance des algorithmes et in fine tirer de nouvelles stratégies permettant de poursuivre l'effort vers le dépassement de soi.

Les joueurs, générateurs de données

Selon Philippe Gargov, le Big Data représente les troisième et quatrième générations de statistiques utilisées dans le monde du sport :

La géolocalisation : ces vastes

(67) http://www.ft.com/intl/cms/s/2/e3561cd0-dd11-11e3-8546-00144feabdc0.html#ixzz374JVEd7M

plages de données servent alors à analyser en détails les mouvements des joueurs sur le terrain et affiner les tactiques de déplacements. Dès 2012, le club du Paris Saint-Germain s'est ainsi doté de brassards GPS équipés sur ses joueurs lors des entraînements pour capter leurs déplacements et leurs efforts.

Les données physiologiques : une quatrième génération de statistiques fournies par le Big Data s'intéresse aux données de santé des joueurs à travers des capteurs physiologiques.

L'équipe nationale d'Allemagne a largement eu recourt au Big Data pour la préparation de la Coupe du monde 2014 grâce aux logiciels de traitement de données de la société allemande SAP capable de récolter plus de 25 informations à la seconde. D'un côté, des capteurs biométriques posés sur les joueurs permettent de collecter des données physiques telles que le rythme cardiaque, les accélérations et décélérations ou les distances parcourues, de l'autre des caméras fournissent des données vidéo capturées sous plusieurs angles détaillant les trajectoires des joueurs sur le terrain.

L'Internet des objets s'empare du sport

En plus de connecter les joueurs, ce sont tous les objets sportifs que l'on connecte : du ballon de basket-ball augmentés 94fifty pour Nike au ballon de football Smart Ball pour Adidas, afin de transmettre en instantané à des supports connectés des données sur la précision, la puissance ou l'angle des tirs d'une frappe et d'accumuler leurs historiques. À l'avenir, l'ensemble de ces dispositifs high-tech pourrait être utilisé en temps réel pour permettre aux entraîïeurs de changer de tactiques en cours de match, ainsi qu'être élargi à de nombreux autres sports.

Quantifier, chiffrer, rationnaliser les prouesses sportives : quelle place pour l'humain ?

« Il y a vingt ans, on ne pesait ni les chevaux ni les cavaliers avant une course hippique. Désormais, on pèse les chevaux, on regarde leur poids, on les mesure, on collecte un certain nombre d'informations qui ont un rôle crucial sur les paris sportifs. Le Big Data permet de démultiplier les sources d'informations, ce qui permet d'affiner le suivi de performances des uns et des autres et in fine les paris sportifs. Le Big Data, c'est le perfectionnement de l'information. » Jean-Luc Errant, Directeur de Cityzen Sciences

Néanmoins, « le Big Data ne remplacera pas l'humain » selon Jean-Luc Errant. Les analyses algorithmiques ne permettent pas tant de prédire avec exactitude les résultats sportifs - car malgré l'immensité des données la part d'incertitude reste grande - que d'améliorer la compréhension des performances sportives et surtout de prévenir des situations à risque dans une démarche orientée vers le bien-être des joueurs.

DIRECTEUR
AFFAIRES JURIDIQUES
ET AFFAIRES PUBLIQUES

MICROSOFT

Dans un monde de plus en plus interconnecté, à l'heure où les flux de données s'accroissent de façon exponentielle

MARC MOSSÉ

LE BIG DATA ET LA VILLE

CHIFFRES CLÉS :

Près de 50 % de la population mondiale vit aujourd'hui dans une zone urbaine

25 % : c'est le pourcentage de réduction de la consommation électrique de la ville de Seattle grâce à l'analyse prédictive et l'optimisation des équipements électriques contrôlés par des logiciels

Le marché des villes intelligentes devrait atteindre 39 milliards de dollars en 2016, contre 10 milliards en 2010 selon ABI Research.

Dans un monde de plus en plus interconnecté, à l'heure où les flux de données s'accroissent de façon exponentielle et où les capacités analytiques des machines s'enrichissent continuellement, le Big data représente la prochaine vague technologique qui impactera durablement et positivement les services rendus par les collectivités et renforcera le potentiel des agents publics et des citoyens. Adossée à la puissance du Cloud computing liée à la diffusion des objets connectés et aux réseaux sociaux, la révolution du Big data - que prolongent les potentialités du Machine Learning - constitue une opportunité afin de changer les choses.

Les données sont une ressource précieuse, un véritable actif.

Toutes les villes du monde sont submergées de données, mais ne savent pas toujours comment les utiliser de façon pertinente.

Les villes ont besoin de solutions qui permettent aux données de circuler au sein d'une infrastructure, intégrant des capteurs, des compteurs, des canaux de médias sociaux et des marchés de données publiques collectant des informations cruciales, mais aussi des systèmes de back-end où les données peuvent être transformées en informations et en ressources que la population et les machines savent exploiter.

Pour y parvenir durablement, il importe que la confiance, la sécurité et la protection soient au coeur de la collecte et du traitement des données.

Aujourd'hui, nous sommes à la fois des témoins et des acteurs privilégiés de ces grandes évolutions qui aident les métropoles à répondre aux attentes de leurs citoyens et de leurs agents. En mettant l'humain et ses droits fondamentaux au centre de leurs priorités et en s'appuyant sur des partenaires stratégiques, les villes renforcent leur rôle de moteurs de l'innovation et du progrès.

La ville intelligente passe par la connexion des agents municipaux

La modernisation numérique des villes doit s'appuyer sur l'innovation en privilégiant l'individu et les usages et en laissant le soin au secteur public, aux entreprises et aux citoyens de bâtir l'avenir de leurs villes. Privilégier l'individu signifie mobiliser toutes les idées, toutes les énergies et toute l'expertise des habitants de la ville pour créer une cité plus démocratique, plus durable et plus efficace. C'est aussi confier aux agents municipaux des appareils et des applications de qualité professionnelle en leur donnant la possibilité de rester connectés via un appareil mobile avec leur bureau et leurs collègues, quel que soit l'endroit où ils se trouvent, afin que le service apporté aux citoyens ne soit pas interrompu dès qu'ils quittent leur lieu de travail.

Prenant en compte les usages des agents, le projet CityNext a mis au point des offres sur mesure qui autorisent et facilitent le paramétrage et l'utilisation de divers types d'appareils, qu'ils fonctionnent sous des systèmes d'exploitation Windows, Android, ou Apple. Les agents ont des idées bien précises quant aux appareils qu'ils souhaitent utiliser et nous pensons que l'interopérabilité offre plus de flexibilité et de confort de travail.

L'Autorité portuaire de Hambourg (HPA) gère le plus grand port d'Allemagne. Elle souhaitait tirer parti des appareils nomades de ses agents pour accroître la mobilité de ses collaborateurs. La HPA a collaboré avec Microsoft et son partenaire, Blue Communications Software, pour adopter une solution Office 365 ProPlus, basée sur le Cloud, afin de permettre à ses collaborateurs d'être productifs quelque soit l'appareil utilisé ou l'endroit où ils se trouvent dans le port. Les problèmes de compatibilité ont ainsi été résolus et le risque d'interruption limité conduisant à une réduction de 75 % du temps consacré par les administrateurs à la gestion du déploiement. Autant de temps disponible pour un meilleur service.

Les Big Data pour valoriser le potentiel humain de la ville

De nombreux projets d'innovation urbaine de grande ampleur ont pour principal objet de rendre les infrastructures « intelligentes » en y intégrant des capteurs et en accroissant les capacités des réseaux. Bien que cet élément soit essentiel, se limiter aux infrastructures engendre le risque de passer à côté de l'énorme potentiel humain qu'offre la ville. Les villes de la prochaine génération comptent sur les personnes au sein de l'Etat et des collectivités, dans les entreprises et les citoyens pour bâtir – via l'innovation – une cité durable dans toutes ses sphères : économique, environnementale et sociale.

Les technologies Big Data peuvent aider les villes à relever des défis de plus en plus pressants. Après des années de collaboration avec des maires du monde entier, Microsoft a identifié plus de 40 domaines d'applications répartis dans huit secteurs critiques : énergie et eau ; bâtiments, infrastructures et planification ; transports ; sécurité publique et justice ; tourisme, loisir et culture ; éducation ; santé et services sociaux ; administration publique.

L'une des applications concrètes du Big data s'illustre en matière de consommation d'énergie avec les réseaux intelligents (Smart grids) et l'analyse prédictive. La ville de Seattle s'est associée à Microsoft, Accenture, un fournisseur d'électricité local et une structure à but non lucratif, pour créer un programme de bâtiment intelligent qui rend possible une réduction de la consommation électrique de 25% grâce à l'analyse prédictive et l'optimisation des équipements électriques contrôlés par des logiciels.

Pas de ville intelligente sans Open Data

La ville intelligente se déploie également grâce à l'ouverture des données publiques et la création d'un écosystème de développeurs imaginatifs et créateurs d'applications mobiles. L'exemple de l'entreprise gérant les transports du Grand Manchester (Transport for Greater Manchester) en témoigne : Transport for Greater Manchester utilise la plateforme Windows Azure, pour héberger des données publiques recueillies. Il est maintenant possible de connaître en temps réel la localisation des transports en communs mais également le nombre de places disponibles sur les itinéraires les plus utilisées[68].

(68) http://www.microsoft.com/casestudies/Case_Study_Detail.aspx?casestudyid=710000003034

UNIVERSITÉ PARIS DAUPHINE

VICE PRÉSIDENT
DE RENAISSANCE NUMÉRIQUE

Le marketing fait d'ores et déjà l'objet d'une révolution profonde grâce à la donnée. L'arrivée des données en volume et en temps réel conduit à d'importantes transformations des outils, des méthodes et des compétences nécessaires pour analyser et comprendre les comportements d'un prospect, d'un client.

HENRI ISAAC

LE BIG DATA ET LE MARKETING

De nombreuses méthodologies, au cœur du marketing sont questionnées : les études, la notion même de campagne. Au-delà de la fonction marketing elle-même, c'est un nouveau paradigme de pilotage de l'entreprise qui est en jeu.

Le consommateur de plus en plus décrypté

D'abord, l'omniprésence des capteurs générateurs de données, la « Révolution des capteurs » [69] selon les termes de Christophe Benavent, démultiplie les possibilités de connaissance du consommateur.

• L'opinion et l'attitude sont identifiées au travers de l'analyse automatique des sentiments et de l'analyse usages d'un service (temps, intensité, fréquences d'utilisation : les variables à étudier sont infinies).

• La micro-localisation : des capteurs type iBeacon localisent un client au sein d'un centre commercial pour lui proposer les promotions les plus adaptées à son parcours, au mètre près.

• Les dispositifs de reconnaissance faciale disposés dans les affiches publicitaires présentes dans le métro parisien.

Ciblage comportemental et re-ciblage (retargeting)

Le ciblage comportemental (Behavioral Targeting), désigne l'ensemble des technologies et des outils qui permettent d'afficher des publicités, des contenus éditoriaux en adéquation avec le comportement d'un internaute. Cette technique publicitaire consiste à employer des élé-

(69) http://www.butter-cake.com/big-data-christophe-benavent-de-letude-a-laction-en-marketing/

ments comportementaux, comme l'historique des pages visitées, les recherches effectuées sur les sites, les produits mis en panier et/ou achetés en ligne, le clic sur bannière publicitaire, pour déterminer avec précision les centres d'intérêt d'un internaute ou d'un mobinaute. La construction de ces profils, leur analyse et leur commercialisation nécessitent des technologies Big Data. Le ciblage comportemental est désormais très largement utilisé par les annonceurs. Il est désormais mieux compris avec le développement du retargeting, dont l'entreprise française Critéo est le leader mondial.

Du Real Time Biding (RTB) à l'achat programmatique

L'explosion des inventaires publicitaires en ligne a conduit à des volumes d'espaces invendus conséquents qui ont finalement été vendus aux enchères par les éditeurs. Progressivement, les techniques de d'achat-vente d'espace ont évolué. Les éditeurs ont construits des plateformes de ventes (Sell-Side Platform, SSP) où les agences peuvent acheter des audiences en temps réel.

Les annonceurs ont, quant à eux, construit des plateformes d'achat (Demand-Side Platform, DSP). Les espaces sont commercialisés aux enchères (plusieurs annonceurs peuvent s'affronter pour acheter un profil d'internaute) et tout se déroule dans un temps qui est inférieur à la seconde et de l'ordre de la milliseconde. De chaque côté, les acteurs ajoutent des données à des Data Management Platform (DMP) afin, d'augmenter la valeur de ces inventaires pour les éditeurs en qualifiant leurs audiences, et de cibler plus précisément les internautes du côté des annonceurs. Ces techniques d'achat en temps réel sont appelées Real Time Biding (RTB). Cet écosystème publicitaire en ligne repose fondamentalement sur des technologies Big Data par les volumes de données traitées, les algorithmes mobilisés et les compétences nécessaires pour bâtir de telles méthodes.

Alors que les techniques du RTB ont historiquement été utilisées pour acheter/vendre des inventaires publicitaires excédentaires, l'achat programmatique est une généralisation et automatisation des achats médias à tous les inventaires publicitaires (y compris les Private MarketPlaces, PMP)[70]. La croissance de l'achat programmatique est forte en France (hausse de 125 % en glissement annuel 2012/13) et 22 % au premier semestre de 2014 [71]. La conception des campagnes et l'achat d'espace reposent donc désormais de plus en plus sur des compétences Big Data, tant pour les éditeurs et les annonceurs.

CRM, DMP et gestion de campagnes

Le développement de l'achat programmatique et des technologies de reciblage (re-marketing) fait évoluer les frontières traditionnelles entre les métiers du marketing. La capacité d'identifier les profils des internautes permet également, lorsque l'on lie la

(70) voir IAB Europe, AppNexus and WARC, (2014), « Why and How Programmatic is Emerging as key to Real-Time Marketing Success », June
(71) Observatoire de l'e-Pub SRI et PwC
(72) http://www.orange-business.com/fr/big-data-analytics
(73) http://www.visitprovence.org/agence_flux_vision_tourisme.asp
(74) Voir par exemple les données de Google sur le sujet : http://www.thinkwithgoogle.com/tools/customer-journey-to-online-purchase.html

base de données client enrichie (par exemple par les données des réseaux sociaux), à la plateforme DMP (Data Management Platform) nécessaire aux campagnes digitales, de personnaliser les messages, leur contenu, leur nature. Outre l'efficacité accrue des campagnes et l'optimisation des budgets, ces transformations liées à l'utilisation des technologies Big Data modifient les métiers de la relation client qui se rapprochent des métiers du media planning.

Le Big Data au service de la conception et de l'innovation produit

La possibilité d'accéder à de nouvelles données massives en temps réel constitue une rupture forte dans la façon d'aborder les études, la conception et l'adaptation des offres et services. A cet égard, l'offre Flux Vision[72] d'Orange Business Service constitue un exemple intéressant de ces transformations dans la conception des offres. Cette offre permet à toute société d'accéder en temps réel aux données de déplacement des utilisateurs du réseau mobile Orange.

L'Office du Tourisme des Bouches-du-Rhône utilise ces outils pour analyser en temps réel les flux touristiques dans le département[73]. Il obtient ainsi en temps réel des données sur les événements, les lieux, les flux de déplacement, la durée des séjours, les lieux visités. Les données anonymisées identifient plusieurs catégories de touristes : les locaux, les excursionnistes, les étrangers (grâce aux données de roaming). On perçoit bien au travers de cet exemple le bouleversement potentiel que le Big Data apporte dans le champ du marketing.

Continuous commerce

Ce que le Big Data contribue certainement le plus à transformer c'est la notion même de campagne et donc la façon d'exécuter une stratégie marketing. Les processus de décisions d'achat des clients se sont complexifiés[74] (réseaux sociaux, App mobile, magasin, TV, tablette, ordinateur, catalogue, affichage, radio, presse, etc.) du fait d'une information disponible abondante et accessible pour le consommateur[75].

De nouveau vocable apparus dans le champ du marketing illustrent bien cette complexité croissante : pre-marketing[76] et re-marketing[77] ne sont que les phases plus complexes et denses d'un processus continu et temps réel que l'entreprise doit analyser, suivre et sur lequel agir. Certains, comme Ogilvy parlent de continuous commerce[78]. La nécessaire maîtrise de cette complexité et du temps réel ne font que renforcer l'utilisation des outils Big Data.

De nouvelles organisations et compétences nécessaires

L'arrivée des méthodologies Big Data bouscule quelque peu les métiers historiques du marketing. L'outillage croissant des décisions marketing, le pilotage des actions et de leur budget nécessitent de nombreuses

(75) voir les données de Google par pays et secteurs disponibles sur le processus de décision d'achat en ligne : http://www.thinkwithgoogle.com/tools/customer-journey-to-online-purchase.html.
(76) Voir par exemple le cas dans l'automobile http://www.largus.fr/actualite-automobile/le-marketing-est-mort-vive-le-pre-marketing-5132379.html
(77) http://www.thinkwithgoogle.com/products/remarketing.html
(78) http://continuouscommerce.ogilvydo.com

nouvelles compétences[79]. Au côté du Directeur Marketing (Chief Marketing Officer, CMO), on voit apparaître des Chief Data Officer, voir des Chief Digital Officer. Si l'enjeu du digital pour les entreprises n'est pas fonctionnel, il n'en demeure pas moins que la fonction marketing est en première ligne dans cette phase rapide de transformation. Si de nouveaux métiers au sein de la fonction marketing émergent (data scientist, data analyst, data visualizer), ce sont la plupart des métiers historiques qui évoluent profondément avec le digital (études, media planning, RP, etc.)

Ce qui est en jeu, c'est la maîtrise des outils digitaux, des méthodologies et de la culture de cet environnement mouvant. Nombreux sont ceux qui appellent à une nouvelle relation entre la Direction Marketing et la Direction des Systèmes d'information (DSI)[80]. D'autres appellent à un directeur technique au sein de la direction marketing[81]. Ce débat concerne d'ailleurs tout autant les agences qui conseillent ou exécutent les décisions marketing. Elles font face à des enjeux tout aussi complexes : comment attirer des data scientists ? Comment faire évoluer et préserver la culture créative à l'ère de la mathématisation des décisions et des campagnes ?

Au-delà de l'organisation de la direction marketing ou des agences, c'est le renouvellement constant de ces compétences qui est le véritable enjeu.

LE BIG DATA ET LA SÉCURITÉ PUBLIQUE

CHIFFRES CLÉS :

20 % : c'est le nombre de crimes qui aurait été commis en moins à Santa Cruz grâce à l'équipement des équipes de police de technologies prédictives fondées sur le Big Data

Une surveillance à grande échelle, rempart contre le terrorisme ?

La protection et la défense des citoyens sont des missions régaliennes pour lesquelles le Big Data permet une efficacité accrue. Toutefois, c'est dans ce domaine que la tentation Orwelienne est la plus présente, à l'image du scandale mondial déclenché par les révélations d'Edward Snowden.

Les documents fournis par cet ancien consultant informatique travaillant pour la National Security Agency (NSA) ont levé le voile sur certaines de ses pratiques contraires au respect de la vie privée. Dans un contexte de risque terroriste accru, l'utilisation du Big Data ne peut se réduire à une caricature où les données deviendraient l'instrument d'une police politique.

(79) http://www.journaldunet.com/solutions/analytics/metier-big-data-data-scientist.shtml
(80) http://www.accenture.com/us-en/Pages/insight-cmo-cio-alignment-digital-summary.aspx
(81) Brinker, Scott, McLellan, Laura (2014), « The Rise of the Chief Marketing Technologist », Harvard Business Review. Jul/Aug, Vol. 92, Issue 7/8, pp. 82-85

Surveiller pour ne plus punir ?

La capacité du Big Data à tracer, cibler et suivre un individu permet de renforcer le contrôle des populations, notamment dans le cadre de menace terroriste. La coopération avec des entreprises génératrices de données, comme des fournisseurs d'accès internet ou des opérateurs téléphoniques, autorise un niveau de profilage très élevé.

Par ailleurs, les capacités prédictives du Big Data ouvrent la porte à une lutte contre le crime a priori, à l'instar du monde décrit par Philip K. Dick dans Minority Report. La collecte et le traitement des données permettraient de cartographier de façon très fine les zones les plus à risque et, grâce aux technologies de Machine Learning, de prévoir, peut-être, le prochain crime ou délit.

Un groupe de chercheurs de l'UCLA, mené par le professeur Jeff Brantingham, a analysé 13 millions de crimes. Avec l'aide du mathématicien George Mohler de l'université de Santa Clara, ils ont appliqué sur ce corpus les algorithmes prédictifs dérivés de ceux annonçant les répliques d'un tremblement de terre. Le logiciel, exploité par la police de Los Angeles, est maintenant capable de définir une zone de quelques centaines de m2 où un crime devrait se produire dans les 12 heures.

À long terme, le risque éthique est de glisser vers « une pénalisation des intentions ».

Gestion des risques et sécurité publique

La sécurité publique, c'est aussi la gestion des risques quotidiens des citoyens : accidents de la route, scandales sanitaires, etc. Dans ces domaines là également, les analyses prédictives permises par les technologies Big Data peuvent être des leviers d'efficacité redoutables.

Ellis-Car est une startup qui, grâce à une solution permettant de connecter les flottes automobiles, souhaite prédire les accidents de la route. Un module embarqué sous le volant du véhicule permet de récupérer un certain nombre de données sur l'état de la voiture mais aussi sur le comportement du conducteur (vitesse, accéléromètre, données GPS etc).

Toutes ces informations sont ensuite stockées par les serveurs de la startup qui les conjugue à toutes les données ouvertes liées à la météo et à la circulation par exemple. À partir de là, un nouvel algorithme permettrait de réaliser des prédictions sur les risques d'accident. Le créateur de cette startup, Rand Hindi, auditionné pour ce présent livre blanc, a été désigné jeune innovateur français de l'année lors du concours organisé par la MIT Technology Review en avril dernier.

BUREAU DE L'INSTITUT G9+

ASSOCIÉE TÉLÉCOMS DIGITAL ET MÉDIAS CHEZ SIA PARTNERS

ISABELLE DENERVAUD

L'immortalité serait-elle à portée de main ? C'est ce que la croissance fulgurante des technologies NBIC (Nanotechnologies, Biotechnologies, Intelligence Artificielle et Sciences Cognitives) dans le secteur de la santé laisserait imaginer à terme.

LE Big data et la quête de l'immortalité

En effet, la recherche dans ce secteur pourrait à terme faire des miracles grâce à la croissance exponentielle des données issues des objets connectés, de la génomique ou de la biologie moléculaire. L'annonce récente par Google d'un projet de recherche de diagnostic de maladies comme le cancer basé sur l'utilisation de nanoparticules artificielles en constitue un exemple frappant. L'assaut pour dépasser la mort est donc officiellement lancé mais jusqu'où ira-t-on pour prolonger la vie ?

Un champ des possibles inspirationnel...

Dans la santé, le séquencement du génome dont le coût devrait passer de 1000$ à 100$ d'ici à 2020 [1] , permet déjà de détecter certaines maladies génétiques ou prédispositions à des maladies. Cet examen interdit en France permet également de fournir un traitement personnalisé aux patients selon leur patrimoine génétique. Sergei Brin, co-fondateur de Google, a publié en 2008 l'analyse de son ADN et sa sa forte probabilité de développer la maladie de Parkinson… Il a changé ses habitudes de vie pour minorer cette éventualité.

Demain, la constitution et l'exploitation de bases de données sur le génome pourraient ouvrir à la voie à une recherche à grande échelle sur les maladies génétiques, la régénération des organes grâce aux cellules souches ou encore la greffe d'organes artificiels. L'immortalité ne serait donc plus qu'à quelques pas si on imagine remplacer les organes défaillants par de nouveaux artificiels et chaque jour plus endurants. En France, Carmat a déjà ré-

[1] Le Monde, 2014

alisé deux greffes de cœur artificiel depuis le début de 2014, et a démontré que malgré le décès du premier patient deux mois après l'opération en mars, le concept d'une telle prothèse est bien validé. Ainsi, un deuxième patient a bénéficié de cette greffe de cœur artificiel et d'autres laboratoires et chercheurs s'intéressent d'ores et déjà au développement d'autres organes artificiels, comme le foie, les reins ou encore les poumons, qui pourraient un jour remplacer le don d'organe.

...et sans limites éthiques ?

Si les perspectives d'allongement de l'espérance de vie se profilent déjà, de nombreux points éthiques demeurent en suspens, comme celui de l'eugénisme induit par les technologies NBIC et le Big Data. La détection prénatale de maladies génétiques, telles que la trisomie 21, est déjà possible aujourd'hui grâce au séquençage de l'ADN présent dans le sang de la mère.

Pour les spécialistes du domaine comme Alexandre Laurent, ce n'est que la première étape du tri des embryons [2] : demain, ira-t-on jusqu'à choisir les "bons" embryons selon les gènes qu'ils comportent ? Cette possibilité est déjà à l'étude en Chine où le Beijing Genomics Institute étudie le patrimoine génétique de 2200 personnes avec un QI supérieur à 160 pour identifier les gènes de l'intelligence. La sélection et la modification des gènes d'ici quelques années ne semblent plus une utopie. La protection et la commercialisation des données de santé, dont le patrimoine génétique, restent également sans réponse aujourd'hui. Les données de bien-être sont quant à elles déjà utilisées pour ajuster au mieux les prix des contrats d'assurance selon le comportement des clients, comme chez Axa où des réductions sont activées en fonction du nombre de pas réalisés par jour. La prédictibilité personnalisée des risques pour affiner les tarifs peut être à double tranchant pour le financement de la santé, notamment pour les mutuelles, où les cotisations variabilisées des membres pourraient provoquer un dangereux déséquilibre...

Vers un meilleur des mondes ?

Le débat autour du progrès technique et scientifique apparaît aujourd'hui encore plus qu'hier un incontournable. Il a été récurrent dans l'histoire, comme l'a illustrée la longue période de transition précédant la diffusion des idées humanistes au XVIIIe siècle. Un nouvel équilibre est à rechercher entre idéal, valeurs, science et progrès, dans un monde en accélération continue : une transparence sur la collecte, l'utilisation et la commercialisation des données personnelles est attendue. Avec le vieillissement de la population mondiale et son impact sur les dépenses de santé, en particulier dans les pays développés, les Etats ont tout intérêt à initier le débat et mobiliser les citoyens et entreprises pour non seulement traiter la question économique, mais aussi délimiter les terrains

[2] Usbek & Rica, 2014

de jeux des expérimentations pour garantir un niveau de confiance.

En une décennie à peine, les NBIC et le Big Data se sont imposés dans le domaine de la santé comme potentiel ultime remède aux maux médicaux de l'humanité. En l'absence de cadre légal et éthique défini, géants du web, acteurs de la pharmaceutique et start-up spécialisées n'ont pas attendu pour investir dans la recherche et l'expérimentation de solutions pour allonger l'espérance de vie humaine. Cependant, un nouvel équilibre reste encore à construire au vu des interrogations économiques, politiques et philosophiques soulevées, entre liberté et déterminisme individuel, afin de dessiner ensemble les esquisses d'un « meilleur des mondes possibles », différent de celui décrit par Aldous Huxley dans son roman en 1930, ou de celui de Candide dans le conte philosophique de Voltaire.

FUTURS USAGES DES **OBJETS CONNECTÉS** ET **BIG DATA** ?

L'offre d'objets connectés est très en avance sur les usages. Le flot de données grandissant d'objets connectés soutient la croissance du Big Data qui, à son tour, facilite l'explosion des usages.

Yannick Lacoste,
CEO - beConect.com

Jean-François Vermont
Chaiman - beConect.com

Pour sa croissance, le Big Data attend beaucoup de l'internet des objets, que l'on nomme – et c'est peu dire – le web 4.0. Prévoir l'évolution du Big Data passe donc, en partie du moins, par la connaissance de ce que sera cet inter- net des objets que l'on voit comme la prochaine grande révolution du web.
Quelle projection, quelle anticipation, quelle esquisse peut-on faire de l'évolution de ce secteur « prometteur » ? Nous avons choisi d'analyser ce marché par la dynamique d'évolution des usages, qui est, selon nous, le meilleur moyen d'y parvenir.

Aujourd'hui, le marché et les usages des objets connectés peuvent se ranger en plusieurs catégories distinctes.

Grand public

L'usage auprès du grand public se propage en cercles concentriques à partir de besoins initiaux simples pour la maison connectée, le fitness (et les fameux wearables) et les loisirs, pensons notamment aux drones.

Au sein de la maison, le monitoring (et plus largement la sécurité) ainsi que la gestion intelligente de l'éclairage sont les deux principaux points d'entrée en terme d'usage. Attention ! On ne dira plus « domotique » pour ces nouveaux usages – terme renvoyant à un contrôle plutôt décentré de l'usager – mais bien « maison connectée » ou « intelligente ».

Les deux logiques de Big Data sont d'ailleurs très différentes : on passe d'un contrôle unique, le plus souvent par le biais d'un spécialiste, à un contrôle multiple directement par les usagers et les systèmes de traitement intelligent nodaux. On touche ensuite aux besoins plus évolués qui nécessitent des transformations parfois plus coûteuses des objets traditionnels : chauffage, climatisation, électroménagers.

La santé est un autre usage majeur qui, en raison de sa forte réglementation, se développe d'abord par des usages simples : tensiomètre, pèse-personne, brosse à dents… Ce qui ne l'empêche pas d'entrer, lentement mais sûrement, dans les hôpitaux et dans les pratiques médicales en général. En 2012 déjà, plus d'un médecin sur deux ayant un smartphone utilisait une application santé.*

Finalement, un ensemble d'usages divers apparaît, allant de la localisation d'objets, perdus ou volés, au microscope connecté. À ce titre, il est intéressant de noter toute la richesse créative en cours qui, parfois et contre toute attente, amène à un usage massif et quasi instantané de l'objet à peine créé, alors qu'ils était inconnu auparavant. Par exemple, une valise connectée vient tout juste de récolter 1,2 M$ via un financement participatif.

Professionnels

Il s'agit ici traditionnellement de l'internet industriel, ou du « machine to machine », qui repose sur des processus de production qui s'appuient sur des composants communiquant entre eux et avec les acteurs de l'écosysteme de l'entreprise : fournisseurs, clients pour la personnalisation des commandes, ressources humaines dédiées à l'exécution de la production, logisticiens pour l'évacuation de la production, etc. Cet usage est en profonde mutation.

L'écosystème, jadis centré sur les professionnels, devient désormais multicentrique. Le client en est un épicentre évident, par exemple dans l'automobile, mais aussi différentes communautés jadis indépendantes le deviennent. Pensons à une communauté d'utilisateurs de voitures qui partageraient des informations entre eux, ou encore une

(82) http://www.euractiv.fr/sections/innovation-entreprises/la-commission-europeenne-met-la-protection-des-donnees-en-haut-de
*http://vidalfrance.com/presse/premier-barometre-sur-les-medecins-utilisateurs-dun-smartphone/

interaction entre le concessionnaire et son client, entre un utilisateur et des stations-essence ou centres de services. Globalement, on considérera les fonctionnalités des objets connectés comme une résultante de la juxtaposition de plusieurs couches :

- les possibilités et caractéristiques des capteurs aptes à collecter des données mesurables : vitesse, accélération, composition chimique, température… ;

- les modalités de transmission des in- formations collectées : RFID, Blue-Too- th, WiFi, 4G, satellite data… ;

- les méthodes de captation, de stockage et d'analyse des données, cette dernière étant particulièrement importante ;

- les fonctionnalités support aux services rendus à l'utilisateur, qui peuvent être par exemple un retour d'information.

Chaque couche de cette pile constitutive des objets connectés fait l'objet d'intenses efforts de recherche, de développement et d'amélioration.

Les capteurs notamment se mniaturisent et se diversifient de façon ingénieuse dans leur aptitude à collecter des données. Aussi, les convertisseurs des variations de valeurs physiques des capteurs en data se perfectionnent et sont de plus en plus économes en énergie.

Une illustration impressionnante de

cette nouvelle génération de capteurs est leur implantation dans une lentille oculaire souple afin de corriger la vue. Un véritable exploit, fruit de quinze années de recherche.

Les méthodes de captation des données échangées se diversifient également et se miniaturisent (bornes relais de captation), ou s'appuient, lorsqu'elles le peuvent, sur des appareils existants, comme les téléphones mobiles.

Un grand nombre d'acteurs agissent pour transformer les objets privés et publics en véritable bornes de collecte. Dans le domaine du mobilier urbain, Citelium transforme par exemple ses réverbères en antenne 4G, en borne Wifi, en support de caméra ou en borne de recharge.

Parallèlement, les outils de traitement du flux de données, de stockage et d'analyse développent chaque jour leur puissance (Cloud, Big Data). Les fonctionnalités ouvertes aux utilisateurs se développent sans limite, et dépassent le plus souvent l'imaginaire dles utilisateurs potentiels.

À quoi peut servir une brosse à dents connectée ? La réponse donnée par

certains concepteurs : à créer une animation familiale parents-enfants par un concours de celui qui se lave le mieux les dents avec des points à gagner, des goodies, des paris, et la possibilité de partager les résultats avec le dentiste. Nous laissons le lecteur apprécier la puissance de l'imaginaire des créateurs de ces objets.

Le nombre d'entrepreneurs, d'inventeurs, de découvreurs, de développeurs et de chercheurs mobilisés croît de façon exponentielle. L'offre prend certes beaucoup d'avance par rapport aux usages, mais cet usage elle le crée !

Doit-on s'attendre à un effet soufflé, à une bulle qui va se dégonfler et décevoir les prévisions optimistes de développement du Big Data ?

Comme l'exprimait Jacques Attali le 25 novembre 2014 lors de la conférence du G9+ sur le thème « Internet va-t-il tuer le capitalisme ? », les révolutions à venir seraient plu- tôt du côté des biotechnologies et des nanotechnologies que de l'internet des objets et du Big Data.

Une réponse est à chercher dans la dynamique de création de ce marché, qui en est d'ailleurs déjà un. Aujourd'hui se met en place un mécanisme de création d'usage fondé sur un web 4.0, web symbiotique pour reprendre l'expression de Joël de Rosnay dans lequel le réel et le virtuel se rejoignent dans un continuum de perception et d'action.

Il ne s'agit pas d'une mécanique d'open innovation, ni de création collaborative dans laquelle créateurs et consommateurs exercent une coresponsabilité de création. Il s'agit d'une mécanique « communautaire » dans laquelle des initiateurs s'investissent dans la création tous azimuts d'objets connectés aux fonctions les plus rares, les plus innovantes, en apparence futiles pour certaines d'entre elles, dans lesquelles l'initiateur recherche un vote, un assentiment des utilisateurs à venir par tout type de mécanisme social, comme par exemple la possibilité de précommander des produits en cours de développement.

Le ressort de la création de la demande repose sur l'appel aux besoins, aux désirs et aux fantasmes des premiers utilisateurs à vouloir non seulement un objet, mais un objet et une communauté d'appartenance, Dans un premier temps, la communauté des utilisateurs de ces nouveaux objets complètement hybrides relient nolens volens les utilisateurs entre eux, avec les organisations qui les mettent en place.

Pour maximiser leur chance de succès, les initiateurs sont prêts à faire « pivoter » leur modèle et prendre en compte les remarques des utilisateurs et le résultat des observations, dans une approche web 4.0 entièrement en ligne avec l'aspect « social » des objets et la numérisation des activités en mode réalité augmentée. Si l'on résume, l'alchimie qui se met en place est fondée sur une hyperstimulation des initiateurs, l'envie,

ou la nécessité, de servir une communauté aux quatre étages de la pile constitutive des objets connectés. Cette alchimie repose également sur un gisement de latences et d'attentes de consommateurs, lesquels souhaitent vivre une existence « augmentée » et une forte intégration sociale dans leur communauté.

Ce modèle est extrêmement dynamique, générateur d'inventions, d'usages et de marchés. En se limitant à ce niveau d'observation, nous pourrions en conclure que le Big Data sera fortement propulsé par tous les flux de données à capter, à stocker, à analyser, à rediriger et à sécuriser. Toutefois, des freins puissants peuvent venir casser la croissance du marché des objets connectés. Ces freins viennent du fait que les objets connectés touchent à l'intime et au personnel. Les données collectées peuvent être utilisées pour enfermer l'utilisateur dans une relation conditionnée et « obligée », au sens ancien du terme qui évoque une nouvelle allégeance à venir.

Il est probable que certains acteurs économiques et institutionnels collecteront des données dans le but premier de développer des stratégies d'influence et de contrôle, et ce en marge des attentes des futurs consommateurs. Pour le comprendre, prenons une analogie avec le reciblage publicitaire sur internet (le retargeting). Lorsqu'un internaute navigue sur un site,

plusieurs acteurs suivent sa navigation et ses cookies, principalement le responsable du site et le publicitaire avec lequel il a passé un accord. En temps réel, la navigation est décortiquée et les cookies de tous les sites consultés sont pris en compte, dégageant ainsi son historique d'utilisation. Les données clients des sites consultés sont rapprochés à des modèles d'influence qui déterminent, grâce au traitement que permet le Big Data, des corrélations que l'on n'imaginait pas il y a quelques années et qui déterminent les messages et les publicités que vous allez recevoir pendant et après votre consultation, soit par email, soit lors d'une navigation ultérieure.

Les résultats sont là, le modèle d'influence est établi, et cela fonctionne : les internautes sont influencés et opèrent des transactions en conformité avec les modèles d'influences mis en œuvre.

Maintenant, projetons-nous dans le domaine des objets connectés.

Sans régulation, ni autolimitation, ni moyen de résistance, des informations bien plus personnelles et intimes qu'une navigation internet seront collectées, disséquées et mis dans des modèles d'influence, voire de contrôle. Pensons simplement à la mesure du rythme cardiaque : que diriez-vous de recevoir une publicité pour l'achat d'un défibrillateur alors qu'un stress récent vous aurait conduit à avoir une mesure anormale ?... Ou encore plus près de nous, les failles de sécurité actuelles sur l'espoir d'un eldorado et sur des caméras de surveillance révélées

par des sites comme insecam.com.

Avec la forte probabilité que certains acteurs économiques et institutionnels soient malveillants et cherchent à considérer les consommateurs comme une ressource à exploiter et à contrôler à leur profit, et non pas au profit de leurs clients et usagers, un écart grandissant risque de se former entre les attentes des consommateurs et les fournisseurs d'objets connectés.

De fortes tensions sont à attendre dans ce scénario, et il est tout à fait imaginable que se produisent quelques scandales médiatiquement mis en scène dans les pays démocratiques dénonçant des acteurs exploitant sans scrupule des données intimes et créant un rapport d'influence ressemblant à de l'abus de faiblesse. Il est alors aussi imaginable que le marché des objets connectés passe de l'enthousiasme le plus débridé à une plus grande méfiance.

L'avenir du marché des objets connectés reposera donc sur le développement de mécanismes de confiance, à titre d'exemples :

- la propriété des données accordée à l'utilisateur et l'interdiction faite aux acteurs d'exploiter ces données sans un consentement explicite, dont la forme reste d'ailleurs à imaginer ;

- la possibilité de se mettre en mode « maison », par analogie au mode « avion » des téléphones mobiles, afin de se déconnecter à tout moment des remontées vers les acteurs économiques et institutionnels, ou de fonctionner dans un mode « dégradé » pour certains objets connectés que l'utilisateur jugerait pertinent.

Une conclusion provisoire est que le Big Data est promis à un bel avenir pour les dix prochaines années, et ce à travers l'explosion des usages des objets connectés qui seraient conçus et exploités dans une logique de respect et de bienveillance envers leurs utilisateurs, ce qui – il faut le dire – est encore loin d'être une évidence.

Avec les objets connectés, les fournisseurs de solution touchent à l'intime, et donc à une matière sensible, voire potentiellement explosive. Les acteurs économiques, institutionnels et étatiques, devront reconnaître que les données personnelles ne leur appartiennent pas, et que des mécanismes de contrôle par des autorités techniques, par des lois et des juges, sont indispensables à leur propre réussite.

Nous recommandons au lecteur intéressé par le futur du Big Data et des objets connectés de se faire son avis tout simplement en achetant des objets connectés, en les utilisant, en les observant et en s'intéressant au débat sur la vie privée en lien avec le Big Data. Pour ceux qui souhaiteraient prolonger le débat, vous pouvez joindre les auteurs dans la rubrique contact de beconect.com. Ils seront ravis d'échanger sur les thèmes abordés, dont les enjeux sociétaux sont extrêmement puissants.

D QUELS SONT LES ENJEUX JURIDIQUES DE CETTE **RÉVOLUTION** ?

Un État transparent sur son fonctionnement et protecteur des libertés personnelles : voilà les deux conditions d'une démocratie à l'ère du Big Data

Romain Lacombe,
Chargé de l'innovation et du développement de la mission Etalab

D'une part, la présentation des différentes applications du Big Data présentent un potentiel économique, social et politique énorme. Ces progrès technologiques viennent questionner de nouveaux enjeux éthiques. Ils concernent la vie privée des individus, la rationalisation des choix et la place de l'humain dans les processus de décisions, la confiance accordée à la technologie, la capacité de réguler des formules mathématiques, etc.

« La demande pour plus de transparence est une modification structurelle face à laquelle les instances de régulations et les entreprises n'ont d'autres choix que de s'adapter »

Yves-Alexandre de Montjoye, chercheur associé au MIT

 Date de parution: décembre 2014

Pour la puissance publique, la révolution des données provoquée par le Big Data bouleverse le cadre législatif et réglementaire en place, structuré en France autour la loi Informatique et Libertés de 1978. Le législateur est confronté à une double problématique : comment continuer à assurer la protection des données privées à l'ère du Big Data sans bloquer l'immense potentiel d'innovation qu'il propose ?

En phase de concertation, le régulateur réfléchit à de nouveaux modèles de régulation, constatant un certain essoufflement de la loi actuelle basée sur le principe de la notification et du consentement. Le régulateur est en attente du General Data Regulation Plan européen à l'ordre du jour de l'actuelle mandature européenne qui a jusqu'en 2015 pour achever ce texte[82]. Cette réflexion est partagée par l'administration américaine qui jusqu'ici a adopté, comme de coutume, une approche de self-regulation.

« Le problème majeur est que la loi de 1978 est structurée autour de la collecte des données et de la finalité de cette collecte : si vous ne collectez des données que pour X motif, vous ne pourrez pas les utiliser pour un motif Y, Y étant jugée incompatible avec X. » François Bourdoncle, Président de FB&Cie, co-fondateur d'Exalead, et co-rapporteur du plan Big Data pour le Ministère de l'Economie

Big Data : l'enjeu de la rgulation est l'utilisation de la donnée, non sa collecte

La régulation actuelle des données, centrée sur la modalité de sa collecte et non sur sa finalité, interdit les croisements de jeux de données de différentes natures et leurs utilisations par les entreprises. Pourtant, la donnée est l'adjuvant essentiel de la révolution industrielle en cours et les conditions de son utilisation font maintenant partie des conditions de croissance des entreprises.

« Le discours alarmiste sur les données personnelles nourrit une défiance qui est un frein pour le développement de l'économie numérique de la France »

Thibaut Munier, Fondateur de 1000mercis-numberly, Administrateur de Renaissance Numérique.

Écrite en 1978 en réaction au projet SAFARI qui visait à croiser les fichiers nominatifs de l'administration, la Loi Informatique et Libertés doit opérer un changement de fond si l'on veut permettre aux entreprises d'exploiter le potentiel économique de leur base de données. À l'heure actuelle, les autorisations de croisement des données sont délivrées par la CNIL au cas par cas, en fonction de chaque entreprise. Ce processus ralentit la pénétration du Big Data dans les entreprises françaises.

(82) http://www.euractiv.fr/sections/innovation-entreprises/la-commission-europeenne-met-la-protection-des-donnees-en-haut-de

LA FIN DE L'ANONYMISATION DES DONNÉES = LA FIN DE LA VIE PRIVÉE ?

Avec la multitude de données collectées sur une personne et ses différentes activités, il sera toujours possible aujourd'hui de retrouver l'origine et donc l'identité d'une donnée, en la croisant avec les autres informations contenues dans d'autres jeux de données. Des études récentes montrent les limites techniques de l'anonymisation comme protection efficace de la vie privée.

- En 2006, AOL avait ouvert les données de recherche de ses utilisateurs pour qu'elles puissent faire l'objet de recherches : les historiques de recherche sur trois mois de 658 000 utilisateurs ont été publiques. En théorie, les données avaient été anonymisées et les utilisateurs n'étaient identifiés que par un numéro. Pourtant, l'ensemble des requêtes d'un internaute peut s'avérer suffisant pour identifier un individu. Ainsi, le New York Times est parvenu à identifier une utilisatrice du New Jersey sur la base des ses recherches pour acquérir une nouvelle maison.

- Dans son étude « Unique dans la foule », l'équipe du MIT du professeur Sandy Pentland[83] a démontré que dans la base de données anonymisées d'un opérateur de téléphone d'un million et demi de personnes, il suffisait de quatre éléments spatio-temporels pour identifier 95 % des participants de l'expérience. Ces éléments peuvent être par exemple un statut sur Facebook avec la mention d'un lieu, mais aussi l'usage d'une borne libre d'accès Wifi. En d'autres termes, la prévisibilité de nos déplacements quotidiens nous rend identifiables malgré l'anonymisation des métadonnées (date et heure de l'appel, récepteur et émetteur de l'appel).

Bien que la puissance de calcul déjoue les mécanismes d'anonymisation, cela ne veut pas dire que toute vie privée est impossible à l'ère de la collecte des données et que l'on peut lire dans un jeu de données comme dans un livre. C'est donc sur cet aspect, l'individu derrière une donnée que s'est concentré le régulateur jusqu'ici.

LA CNIL DONNE ACCÈS AUX DONNÉES DE SOINS DES ASSURÉS AXA DANS UN CADRE EXPÉRIMENTAL

Remis en 2003, le rapport Barbusiaux préconisait d'autoriser, sous conditions, les complémentaires santé d'avoir accès aux données de santé de leurs assurés. Ainsi, la CNIL a permis à Axa, en 2010, dans un cadre expérimental, d'obtenir les données de soins en pharmacie de ses assurés santé en préservant le secret médical. L'expérimentation a été menée auprès d'un panel de 41 000 assurés dans dix départements de France.

« Le but final est de proposer aux assurés des garanties pouvant mieux correspondre à leurs besoins (remboursement optique en fonction du défaut visuel, remboursement des médicaments non pris en charge par la Sécurité sociale, etc) »,

explique Axa dans un communiqué.

Ici, par la multiplication, le croisement et l'analyse de données, la voie vers une médecine plus personnalisée, plus performante et moins coûteuse a été ouverte.

(83) Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3, Article number: 1376, Mars 2013 - http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

QUELQUES PISTES ACTUELLEMENT EN COURS D'EXAMEN EN EUROPE ET DANS LE MONDE POUR FAIRE ÉVOLUER LE CADRE JURIDIQUE QUI ENCADRE LES DONNÉES PERSONNELLES

Privacy by Default

« We live in a Track-Me world, one from which opting out is, as a practical matter, often not possible. »

Lauren E.Willis, Université de Berkeley

La Privacy by Default consiste à paramétrer par défaut les plus hautes options de protection des données personnelles dans les produits et services numériques. Elle est fondée sur trois constats : (1) le paramétrage initial proposé n'est pas modifié par l'utilisateur, (2) l'utilisateur est favorable à une meilleure protection de ses données personnelles, (3) les entreprises devront être plus transparentes pour convaincre l'utilisateur d'ouvrir ses options de confidentialité.

En somme, le concept de Privacy by Default considère que l'utilisateur n'est pas suffisamment informé et compétent pour être le seul responsable de la protection de sa vie privée. En effet, en 2013, 63 % des utilisateurs Facebook aux États-Unis n'ont jamais essayé de régler leurs options de confidentialité. Celles-ci doivent donc être garanties dans le paramétrage même des plateformes en ligne. L'avantage de la Privacy by Default réside dans le fait que ce modèle de régulation systémique ne souffre pas du volume ou de la variété des données collectées

La Privacy by Default est au coeur de la politique européenne de régulation des données personnelles. L'ancienne vice-présidente de la Commission Européenne, Viviane Reding, en a fait le 3ème pilier du General Data Regulation Plan, aux côtés de la transparence et du droit à l'oubli.

Privacy by Design

Né durant les années 90 aux États-Unis, le concept de Privacy by Design consiste à mettre la protection des données privées au coeur de la conception même du produit : celles-ci sont protégées a priori par le design du produit ou service et non plus par un contrôle a posteriori. Son implantation dans l'architecture même du produit ou service permet d'apporter une réponse globale à la protection des données personnelles, adaptée au Big Data. C'est le modèle en place, par exemple, dans la gestion des données traitées par les caméras de surveillance aux États-Unis.

Présente dans les textes européens, au coeur des réflexions de la CNIL, la Privacy by Design implique de lourds investissements et manque d'applications concrètes de la part des entreprises.

Le modèle émergent de protection par certifications

« C'est le processus qui détermine la finalité des données qui importe: pourquoi les croise-t-on ? Comment ? Aussi, ce que l'on doit réguler et juger c'est la légitimité des traitements qui sont faits par le croisement des données en fonction de la finalité du processus lui-même et non la finalité de la collecte»

François Bourdoncle, Président de FB&Cie, co-fondateur d'Exalead, et co-rapporteur du plan Big Data pour le Ministère de l'Economie.

La réflexion autour de ce nouveau modèle de régulation est encore jeune. Il été mis en avant par John Podesta dans son rapport pour la Maison Blanche et par le rapport Big Data remis par François Bourdoncle et Paul Hermelin, PDG de Cap Gemini France, au Ministère du Redressement productif en 2014. Il fait écho à l'inadéquation entre le cadre « Notice & Consent» et le contexte Big Data où la collection apparaît incontrôlable.

L'idée sous-jacente est de ne pas «couper le robinet» des données à la base mais bien de contrôler leur usage responsable a posteriori. Cette régulation sectorielle s'oppose à une tradition européenne de législation avançant par grands textes fondateurs, comme c'est le cas en ce moment avec le General Data Protection Regulation actuellement en cours de rédaction par la Commission.

La restitution de leurs données aux individus : les projets VRM dans le monde

Le principe d'un projet VRM, pour Vendor Relationship Management, est de restituer aux individus toutes les informations qu'ils délivrent par leur comportement. Le VRM ne suffit pas à constituer un cadre juridique structurant pour le Big Data mais cela peut être un levier vers plus d'autonomie et de liberté pour les citoyens. Les initiatives de « Self Data » tentent de mettre à mal l'ambiguïté autour de la notion de données personnelles, à la fois perçues comme une manne par les entreprises et comme un danger pour les opinions publiques.

Renaud Francou, porteur du projet MesInfos pour La FING, indique ainsi que 78 % des consommateurs ne font pas confiance aux entreprises pour l'exploitation de leurs données personnelles : l'asymétrie entre entreprises et consommateurs dans le domaine de la récolte et de la gestion des données personnelles engendre un délitement de la confiance de ces derniers et la montée d'un désir de plus en plus fort de contrôle et de maîtrise de ses données.

En France, c'est La FING qui, depuis novembre 2013, mène ce type d'expérience avec le projet Mes Infos. L'expérimentation a ainsi réuni pendant six mois 300 individus volontaires clients d'au moins deux des huit entreprises partenaires qui ont accepté de participer à ce retour de data, parmi lesquelles Axa,

le Crédit coopératif, la Banque postale, les Mousquetaires, Orange, la Société générale, Google et Solocal. Une plateforme sécurisée de cloud personnel a été mise en ligne sur laquelle les quelques 300 testeurs pouvaient avoir accès à l'ensemble de leurs données telles que leurs relevés de comptes bancaires, leurs historiques d'achats, leurs données de géolocalisation ou encore de communications.

Dans le même temps la FING a lancé en partenariat avec des développeurs et des écoles un concours de création d'applications et de services capables de réutiliser de façon innovante les données mises en jeu.

Facilitation du quotidien, classements, alertes, self-coaching, mise en relation, bons de réduction, la cérémonie a ainsi été l'occasion de présenter les quelques trente concepts et la dizaine de prototypes élaborés pour l'occasion, à l'image de l'application « Moi » qui propose de fournir chaque mois dans une démarche de « quantitative self » un relevé de l'ensemble des activités de l'utilisateur, comme le nombre de kilomètres parcourus ou l'évolution des achats au supermarché.

Une équipe de sociologues a accompagné l'expérience et mené une série d'enquêtes quantitatives et qualitatives pour rendre compte du ressenti des 300 testeurs. Les résultats ont montré la confirmation du phénomène du « privacy paradox » : un niveau de préoccupation élevé pour les questions de protection de la vie privée dans le panel des expérimentateurs mais qui ne se concrétise pas directement par une utilisation plus précautionneuse des services en ligne proposés.

L'initiative de la FING et la mouvance de « Self Data » repose sur des projets similaires menés :

- aux Etats-Unis, avec le projet Blue Button qui permet, en un clic, de télécharger ses données dans les secteurs de l'énergie, de la santé ou de la formation ;

- au Royaume-Uni qui a mis en place avec le soutien des pouvoirs publics le projet MiData : les entreprises participantes s'engagent à rendre aux individus les données personnelles et transactionnelles les concernant, dans un format lisible.

PARTIE IV

LA FRANCE À L'HEURE DU **BIG DATA**

Les enjeux et dynamiques qui traversent la révolution du Big Data exigent des décideurs politiques et économiques qu'ils saisissent le phénomène et s'attèlent à favoriser son avènement en France.

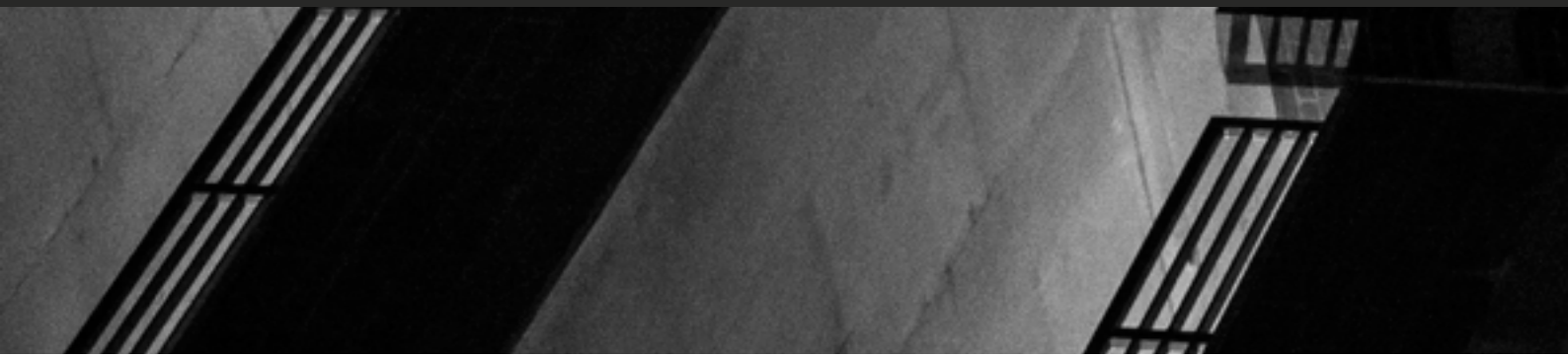
Depuis plusieurs années, les gouvernements successifs montrent leur intérêt pour le Big Data et le levier économique qu'il représente. Les politiques publiques concernant le Big Data se trouvent à la confluence de deux problématiques :

- L'équation entre vie privée et compétitivité. La France ne doit pas laisser échapper de potentiels leviers de croissance, tout en agissant dans un cadre légal protecteur des libertés individuelles.

- Adapter l'économie traditionnelle aux modèles économiques basés sur l'analyse de la donnée. De nombreux pans de l'économie française peuvent être bouleversés par les acteurs qui maîtrisent la donnée et qui ne craignent pas de remettre en cause les équilibres économiques traditionnels. Les grandes entreprises nationales voient déjà naître un nouveau type de concurrence face auquel elles peinent à innover.

A L'ÉTAT, UTILISATEUR EXEMPLAIRE DES TECHNOLOGIES BIG DATA

> À titre d'exemple, pour faire tomber les peurs et parce que ces outils sont redoutablement efficaces, la puissance publique pourrait être le premier acteur à adopter en masse les technologies Big Data pour l'élaboration et l'évaluation de ses politiques publiques.



La révolution de la donnée constitue pour l'État une opportunité de dynamiser les services publics, la lutte contre le chômage ou la gestion des hôpitaux. Une bonne maîtrise du Big Data permet une meilleure connaissance et un meilleur suivi des citoyens et offre également, dans un contexte de réduction budgétaire, la possibilité d'optimiser l'allocation des ressources.

Différents exemples illustrent la puissance du Big Data au service de l'intérêt général :

- La prédiction dans le domaine de la santé.

En 2009, une université canadienne a développé une plateforme d'analyse en temps réel des flux de l'Hôpital des Enfants Malades de Toronto. L'établissement est parvenu à prévoir ainsi l'apparition d'infections nosocomiales 24h avant les premiers symptômes[84].

(84) IHTT, Transforming Health Care Through Big Data, 2013, p.8

Le ciblage dans la lutte contre la fraude.

Avec cent millions d'euros de fraudes détectées en 2009, Pôle Emploi compte sur un meilleur traçage des utilisateurs pour limiter la fraude. La Cour des Comptes[85] préconise le croisement des données avec la Sécurité Sociale mais également des acteurs privés comme les banques et les opérateurs téléphoniques. Il serait par exemple possible de détecter si un bénéficiaire réside à l'étranger alors que cela est interdit. Des systèmes similaires ont été mis en place pour lutter contre la fraude et l'évasion fiscale dans plusieurs pays. (cf encadré page suivante)

Une meilleure gestion des villes.

La population mondiale réside maintenant en majorité dans des zones urbaines. La part des urbains au sein de la population mondiale atteindra 70 % en 2050, soit 6 milliards d'individus. Le Big Data apporte de nombreuses réponses aux défis engendrés par cette urbanisation en pleine explosion. Les agents municipaux doivent en effet assurer la qualité des services publics tout en veillant à une bonne gestion financière : le contexte actuel valorisant le précepte du "faire plus avec moins". A Issy-les-Moulineaux, la municipalité a décidé de travailler avec dix entreprises pour développer IssyGrid, le premier réseau de quartier intelligent en France. Parmi différentes innovations, ce réseau a permis d'optimiser la gestion de l'eau et le traitement des eaux usées grâce à des systèmes d'analyse de consommation, réduisant de 10 à 20 % la consommation et la facture énergétique.

La prédiction en matière de crimes et délits.

Dans sa nouvelle Minority Report, Philip K.Dick dépeint un monde où les crimes sont devenus impossibles grâce à trois mutants qui ont le pouvoir de prédire l'avenir. La réalité dépasse la fiction. Le logiciel PredPol – predictive policing – fonctionne sur un algorithme dessiné par un mathématicien, un anthropologue et un criminologue. En agrégeant des données aussi diverses que la composition démographique d'un quartier ou l'historique des infractions passées par exemple, les autorités policières peuvent distinguer les zones où les prochaines infractions sont les plus probables.

De cette manière, les forces de l'ordre peuvent dépêcher des hommes sur des zones à risques et empêcher une infraction de se produire. Les premiers tests du logiciel PredPol ont été réalisés dès 2011 par la police de Santa Cruz, en Californie. Différentes villes ont répété l'expérience : New York, Los Angeles… Dans la mégalopole californienne, PredPol a été utilisé entre novembre 2011 et mai 2012. Les crimes et infractions ont diminué de 13 % alors que dans le reste de l'Etat – qui n'a pas déployé cette technologie – ce chiffre a augmenté de 0,4 %.

(85) www.ccomptes.fr/content/download/.../2_6_Pole_emploi_tome_II.pdf

LE BIG DATA : NOUVEL ARME CONTRE LA FRAUDE À L'ASSURANCE MALADIE.

L'EXEMPLE AMÉRICAIN DU FRAUD PREVENTION SYSTEM

En moyenne le montant de la fraude à l'assurance santé équivaut à presque 7 % de la totalité des dépenses de santé d'un pays soit pour 2014 un coût mondial équivalant à 376 milliard d'euros.

La situation est particulièrement inquiétante aux Etats-Unis. La fraude à l'assurance santé représente entre 3 et 10 % du total des dépenses de santé soit entre 75 et 250 milliards de dollars par an. Alors que le vieillissement de la population s'accélère et que le nombre de maladies chroniques augmente, les autorités ont décidé d'agir en conséquence.

En juin 2011, le Ministère de la santé américain a déployé le Fraud Prevention System (FPS). Cette technologie fonctionne selon des technologies du Big Data. Il collecte et agrège des données. Puis un protocole d'analyse fondé sur des algorithmes examine au fur et à mesure les demandes de remboursement présentées. Ces demandes sont notées en fonction du risque de fraude. Si une demande semble présenter de forts risques de fraude, les autorités sont alertées avant de procéder au remboursement afin de vérifier l'authenticité du document.

Le contrôle de données fonctionne sur quatre types d'algorithmes :

- Rules-based models : filtrent les demandes de remboursement. Ils identifient par exemple les factures qui portent un numéro d'identification Medicare volé ou utilisé de manière anormale.

- Anomaly models : détectent les comportements anormaux en les comparant à des comportements de référence. Par exemple, un fournisseur de soins de santé facturant bien plus de services de soins que 99 % des fournisseurs analogues en une seule journée sera identifié.

- Predictive models : évaluent des comportements à l'aune de cas précédemment identifiés comme frauduleux.

- Network models : analysent des liens associés entre différents acteurs. Par exemple, les services d'un fournisseur lié ayant un comportement frauduleux seront identifiés comme frauduleux grâce à l'analyse de localisation.

Un retour sur investissement avantageux

L'investissement pour construire et mettre en place le FPS fût relativement lourd : environ 41 millions de dollars. Mais le retour sur investissement est très intéressant. En effet, le système préventif a permis à Medicare d'économiser 210 millions de dollars. Ainsi pour un dollar dépensé, cinq dollars ont été économisés.

Informations et chiffres issus du livre blanc « D'un système de santé curatif à un modèle préventif grâce aux outils numériques », Renaissance Numérique, Septembre 2014

L'État pourrait ainsi, en utilisant les technologies Big Data, être valeur d'exemple pour encourager d'une part le marché français du Big Data à se développer, et d'autre part encourager les grandes entreprises traditionnelles à s'engager dans le secteur du Big Data.

B

L'ÉCOSYSTÈME FRANÇAIS : DE VRAIS ATOUTS POUR DEVENIR LEADER EUROPÉEN DU BIG DATA

> De nombreuses startups et agences spécialisées sont nées de cette nécessité de maîtriser la donnée pour les entreprises et organisations.

Un écosystème en trois strates

De nombreuses startups et agences spécialisées sont nées de cette nécessité de maîtriser la donnée pour les entreprises et organisations. Cet écosystème florissant se compose de trois couches distinctes :

La production de la donnée.

Il s'agit des startups qui participent à la production et collecte des données en fabriquant des capteurs, à l'instar du tee-shirt connecté produit par CityzenSciences, ou en rendant accessibles des données publiques, comme, par exemple, Kel Quartier qui dessine le portrait-robot d'une zone urbaine : revenu moyen des habitants, taux d'insécurité ou densité du tissu commercial.

Les outils de traitement et d'analyse de la donnée.

Ce sont les entreprises qui proposent aux grands groupes des solutions technologiques et des conseils pour mieux maîtriser la donnée. À cheval entre l'organisation d'une agence et d'une startup, elles développent des outils en interne qu'elles associent à ceux existants comme Hadoop. En France, 1000mercis-numberly et Fifty-Five font figure de leader du marché qui connaît un taux de croissance formidable. Fondées plus récemment, des entreprises comme Dataiku, Captain Dash et Squids Solution font également parties déjà des acteurs de ce marché dit de « l'analytics ».

Les applications qui exploitent la donnée pour proposer de nouveaux services.

Cette dernière strate d'entreprises met en action les données disponibles pour concevoir des applications innovantes. Ces données peuvent être publiques, comme l'application Transilien développé par l'entreprise Snips et qui exploitent les données fournies par le STIF, ou bien privées.

C'est par exemple le cas de Criteo qui utilise les données fournies par ses clients pour fournir une solution de re-ciblage publicitaire à travers un puissant algorithme. Si IDC estime que le poids des technologies et services liés à l'analyse et à l'exploitation des données en grande quantité en temps réel atteindra 16,9 milliards au niveau mondial en 2015, en France, il est estimé à seulement 387 millions d'euros en 2013. Notons toutefois, que la hausse du secteur est estimée à 40 %. Si la hausse prévue est donc déterminante, reste que le marché français, qui dispose pourtant de tout un écosystème français prêt à développer des projets Big Data, reste frileux.

LES ETATS-UNIS : PASSAGE OBLIGATOIRE POUR LES ENTREPRISES FRANÇAISES DE BIG DATA ?

« Le savoir-faire technique, la taille et la maturité du marché américain restent supérieurs au marché français » Thibaut Munier, Fondateur de 1000mercis-numberly, Administrateur de Renaissance Numérique.

Selon Transparency Market Research[86] qui évalue les chiffres du marché du Big Data dans le monde, l'Amérique du Nord capte aujourd'hui, à elle seule, près de 55 % du marché mondial. Sur ce marché, les entreprises américaines que sont HP, Teradata, Opera Solution, Mu Sigma and Splunk Inc détenaient, en 2012, 60 % du marché.

Ainsi, pour les startups spécialisées dans la mise en place de projets Big Data, démarcher en France n'est pas aisé.
« Nous avons de belles réussites ici, mais en règle générale les grands groupes français restent trop frileux pour confier leurs jeux de données à une startup. Au-delà du risque, ils n'identifient pas encore clairement le retour sur investissement direct du passage à une approche data-driven. Aux Etats-Unis, le marché est plus mature et nous ne rencontrons pas ce type de barrière » explique Marine Romezin, Communications Manager chez Squid Solutions, qui vient d'ouvrir un bureau à San Francisco.

(86) Transparency Market Research, Big Data Market - Global Scenario, Trends, Industry Analysis, Size, Share and Forecast, 2012 - 2018, http://www.transparencymarketresearch.com/big-data-market.html
(87) On peut noter les rapprochements autour de cursus spécialisés Big Data entre Grenoble Ecole de Management et l'EMSI, entre l'EPSI et l'IDRAC, HEC et Télecom Paris Tech.

VALORISER LE SAVOIR-FAIRE FRANÇAIS POUR MAÎTRISER LE BIG DATA

L'éducation supérieure française et la recherche sont les deux leviers pour la maîtrise technique des flots de données ; condition sine qua non à l'activation du Big Data. Elles sont traversées par une problématique commune : approfondir l'interdisciplinarité pour répondre aux défis techniques du Big Data

Le nouveau besoin en experts opérationnels s'accroît fortement et les formations proposées par les universités scientifiques et les écoles d'ingénieurs sont fortement valorisées. Sans qu'aucun chiffre ne fasse autorité sur le sujet, on peut raisonnablement estimer que vingt-mille à trente-mille nouveaux professionnels seront nécessaires chaque année pour répondre aux besoins des entreprises et des organisations françaises, structurer et valoriser leurs données et automatiser leurs services.

La conduite de projets Big Data demande plusieurs compétences répondant à des formations distinctes :

- bagage technique, fourni en France par les écoles d'ingénieurs ou les facultés de mathématiques et de statistiques ;
- une compréhension des enjeux commerciaux, financiers et managériaux ;
- la gestion de projets Big Data qui va de la phase de collecte auprès des différents acteurs pertinents, à la visualisation et la compréhension des analyses fournies par les technologies Big Data.

Cette hybridation des profils demande aux instituts d'éducation supérieure de se recomposer, à l'image de l'inflation des doubles formations écoles d'ingénieurs – écoles de commerces [87]. Étant donnée la diversité des métiers du Big Data et des compétences requises, tous les degrés de l'université sont concernés, des formations technologiques et spécialisées aux masters et doctorats.

« Le leader de demain ne sera ni ingénieur, ni manager : ce dont nous avons besoin aujourd'hui, c'est de caractères hybrides, capables de manier les données mais également d'aller chercher, de trouver les bases de données intéressantes à compiler, etc. » Nicolas Glady, Professeur Associé, Titulaire de la Chaire Accenture Strategic Business Analytics, ESSEC

On peut noter les rapprochements autour de cursus spécialisés Big Data entre Grenoble Ecole de Management et l'EMSI, entre l'EPSI et l'IDRAC, HEC et Télecom Paris Tech.

QU'EST CE QU'UN DATA SCIENTIST ?

" Un Data Scientist c'est plus qu'un statisticien avec un Mac ! " Ce trait d'humour de Florian Douetteau – fondateur de Dataiku – révèle toute l'ambiguïté du métier de Data Scientist, à la fois statisticien, ingénieur et chef de projet.

• Une solide formation en statistiques et en mathématiques est nécessaire pour pouvoir décrypter les données, formuler des intuitions et in fine transformer la masse d'informations en intelligence au service d'une organisation.

• L'efficacité d'un Data Scientist provient également de sa capacité à se plonger dans des bases de données pour les nettoyer, les rendre opérationnelles et construire des modèles prédictifs. Rand Hindi, fondateur de Snips, déplore le manque d'expérience pratique des étudiants français dans l'élaboration de ce genre de modèle : « la majorité des étudiants en mathématiques anglais ont été amené à construire des modèles durant leurs études, notamment dans le cadre de cours de finance quantitative : c'est un vrai manque des étudiants français ».

• Transformer les méthodes de travail et de prise de décisions à l'aune des connaissances obtenues grâce au Big Data est la dernière facette du métier de Data-Scientist. Elle requiert des compétences en management et en business pour parvenir à mettre le Big Data aux services des équipes de l'entreprise.

En octobre 2012, la Harvard Business Review affirmait que Data Scientist était le métier «le plus sexy du XXIème siècle» et, face à la pénurie d'individus qualifiés, prévoyait une future guerre des talents. Le cabinet Gartner prévoit la création de quatre millions et demi d'emploi pour répondre aux besoins du Big Data dans le monde d'ici à 2015.

Les métiers de la donnée requièrent des compétences spécifiques, à la croisée des mathématiques, de la statistique, de l'informatique et du management. Face à cette hybridation des compétences, les écoles d'ingénieurs et les universités[88] ont adapté leurs cursus pour proposer des formations spécialement dédiées au Big Data. Les entreprises s'arrachent ces étudiants extrêmement qualifiés et les salaires à la sortie d'écoles grimpent rapidement[89].

Parallèlement, on assiste à une mi-

(88) L'École Polytechnique, ENSAE, les Écoles Centrales, ParisTech et les facultés d'Orsay et de Jussieu on régulièrement été cité par notre panel
(89) Étude de l'entreprise américainde Kforce et accessible sur http://www.lemondeinformatique.fr/actualites/lire-les-salaires-des-8-competences-les-plus-recherchees-en-big-data-56610.html
(90) https://www.gov.uk/government/news/73-million-to-improve-access-to-data-and-drive-innovation

gration des employés de la finance quantitative, notamment des étudiant issus du cursus X – ENSAE, vers les sociétés technologiques. Cela est dû d'une part, à la baisse d'attractivité de la finance et, d'autre part, à l'imaginaire positif du monde de la startup qui, par ricochet, valorise les métiers de l'informatique. De plus, les salaires des sociétés technologiques tendent à s'aligner avec ceux de l'industrie financière et constituent une incitation supplémentaire.

« L'excellence des écoles d'ingénieurs françaises et des formations universitaires en mathématiques et statistiques forment chaque année des milliers d'étudiants très compétents » Florian Douetteau, fondateur de Daïtaku .

Cependant, le recrutement à l'étranger, notamment dans les Ivy League américaines et en Angleterre, reste une option pour beaucoup d'employeurs français. Pour Rand Hindi, fondateur de Snips, « un étudiant sortant de Stanford sera bien plus compétent opérationnellement qu'un étudiant de l'ENS ou de l'X ».

Soutenir la création d'un centre de recherche interdisciplinaire sur la donnée

En parallèle de la formation, la recherche académique autour du Big Data doit être un levier d'innovation pour les organisations. L'exemple des centres de recherche américains, comme le MIT cité à de nombreuses reprises dans ce livre blanc, souligne qu'ils jouent aussi bien un rôle clef dans l'innovation technologique que dans les débats sur la régulation. Ils forment et attirent les talents, nouent des partenariats avec des entreprises nationales et conseillent l'État sur les politiques publiques.

Les exemples anglo-saxons montrent la marche à suivre :

- En février 2014, l'Angleterre a investi 98 millions d'euros dans quatre centres de recherche qui interrogent le rôle de la donnée dans les problématiques santé, urbanisme, énergie et culture [90].
- La Maison-Blanche a lancé, en Novembre 2013, un plan d'investissement de 200 millions d'euros pour la recherche en Big Data pour les grandes entreprises et les universités, notamment dans le domaine de la santé [91].

Pour canaliser l'expertise française, la création d'un centre français, voire européen, de recherche sur la donnée permettrait d'allier recherche fondamentale en statistiques et en mathématiques et travaillerait à des applications dans tous les domaines de l'action publique. Par exemple, à l'instar de l'OpenPDS développé par une équipe du MIT, des solutions techniques pour protéger la vie privée de manière structurelle (Privacy by Design) pourraient émerger d'un tel institut.

(91) http://www.whitehouse.gov/sites/default/files/microsites/ostp/Data2Action%20Press%20Release.pdf

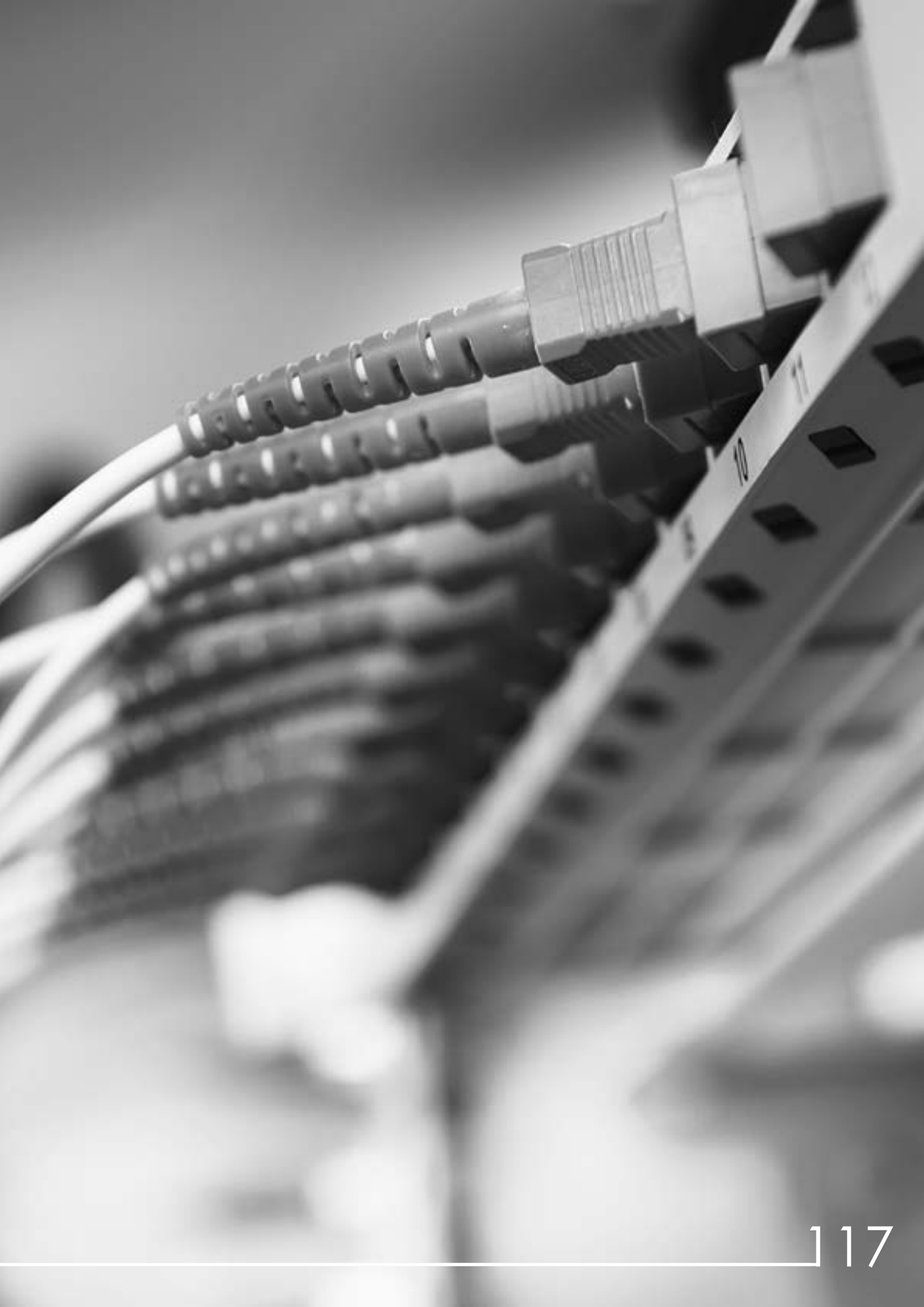
C ÊTRE EN TÊTE DE LA RÉFLEXION SUR LA NOUVELLE RÉGULATION À L'ÈRE DE LA DONNÉE

De nombreuses startups et agences spécialisées sont nées de cette nécessité de maîtriser la donnée pour les entreprises et organisations.

Parce qu'il n'appartient ni à la technique, ni aux intérêts économiques des entreprises de décider de l'avenir de la société, le législateur français et européen doit permettre à nos économies de tirer pleinement profit des promesses du numérique, sans avoir à abandonner un haut niveau de protection de la vie privée.

Penser la règlementation des risques algorithmiques

L'appareil législatif peut s'avérer trop lourd comparé à une régulation par cas ou par secteurs, dans le souci toujours de préserver les opportunités économiques du Big Data. Les algorithmes à l'œuvre dans le Big Data ont une influence politique, culturelle et scientifique de plus en plus importante. Ils sont décisifs pour la personnalisation des contenus et services proposés à l'utilisateur mais peuvent recéler des biais discriminants. L'opacité qui entoure leur composition interdit une prise en compte des risques inhérents à leur massification.



CONCLUSION

SIX PROPOSITIONS DU G9+ POUR FAIRE DE LA FRANCE UN ACTEUR MOTEUR DE LA RÉVOLUTION BIG DATA

PROPOSITION 1 : Déployer sur 3 ans des programmes test d'utilisation de technologies Big Data dans certains secteurs des politiques publiques pour dégager des économies directes : par exemple dans le cadre de la lutte contre la fraude à l'assurance maladie, ou dans la gestion de certaines politiques publiques de santé. Les acteurs publics doivent alors obtenir des dérogations de la CNIL. De telles initiatives dynamiseraient tout l'écosystème Big Data, en promouvant la coopération entre les startups expertes de ces technologies et les grands groupes détenteurs de données complémentaires.

PROPOSITION 2 : Une loi sur l'Open data pour contraindre les administrations stratégiques à ouvrir leurs données concernant les événements et statistiques qui touchent directement à « la vie, la santé et le patrimoine des personnes ». À l'instar de l'Estonie, contraindre par la loi les organisations publiques à ouvrir leurs données à les diffuser sur la plateforme data.gov.fr, le portail national des données publiques en France. Inscrire dans cette même loi, la gratuité des données : Aujourd'hui l'article 15 de la loi de Juillet 1978 (mise à jour en 2003) postule que les données publiques peuvent avoir un prix[92]. Cette facilité financière d'accès aux données dynamiserait l'écosystème de startup et interdit sa captation par un groupe d'entreprises.

PROPOSITION 3 : Développer une offre de formation couvrant l'intégralité de la chaîne de métiers reliés au Big Data.

PROPOSITION 4 : Valoriser l'expertise française en mathématiques, statistiques et télécommunications et parvenir à les hybrider autour de projets et centres de recherches communs.

PROPOSITION 5 : Faire émerger le débat de l'évolution de la régulation Big Data au sein du gouvernement, des Parlements français et européens et des CNIL européennes : l'éthique de la décision à l'ère des algorithmes ou encore la régulation par le traitement et le processus de croisement de la donnée sont des enjeux dont les pouvoirs publics et la société civile doivent se saisir. Evoluer vers une régulation unifiée pour l'Europe permettant aux acteurs européens innovants de bénéficier d'un marché continental.

PROPOSITION 6 : Réfléchir à la possibilité d'audit des algorithmes par un régulateur certifié sur la protection de la vie personnelle à l'ère du Big Data : pour les entreprises comme pour les acteurs publics. Cela permet une forme d'une régulation qui se focalise sur la manière dont les données sont utilisées et non comment elles sont collectées.

(92) "La réutilisation d'informations publiques peut donner lieu au versement de redevances"
Art. 15, Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal, http://legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000339241

LISTE DES PERSONNES AUDITIONNÉES POUR LE LIVRE BLANC :

Christophe Benavent
Chercheur en marketing à Paris-10

François Bourdoncle
Président de FB&Cie, co-fondateur d'Exalead, co-rapporteur du plan Big Data pour le Ministère de l'Economie

Ekbel Bouzgarrou
Chief Technologie Officier
Air France KLM

Stéphane Buttigieg
Institut Louis Bachelier,

Mehdi Chouiten
Data Scientist senior chez Parkeon

Yves-Alexandre De Montjoye
Doctorant au MIT, laboratoire de dynamique humaine du Media Lab

Florian Douetteau
Fondateur de Dataiku

Jean-Luc Errant
Fondateur de la société Cityzen Sciences-Cityzen Data

Nicolas Glady
Professeur Associé
Titulaire de la Chaire Accenture Strategic Business Analytics

Samuel Goëta,
Doctorant à Télécom ParisTech

Olivier Guérin,
Pdg d'image & dialogue group
Adhérent de Renaissance Numérique

Rand Hindi
Fondateur de Snips

Romain Lacombe
Chargé de l'innovation et du développement de la mission Etalab.

Thomas Lefèvre
Médecin de santé publique
Ingénieur Mines-Télécom
Docteur en sciences
Chercheur associé à l'IRIS (CNRS/INSERM/EHESS/Paris 13)

Guillaume Liegey
Fondateur de Liegey-Muller-Pons

Arnaud Massonie
Co-fondateur et Directeur Général de l'agence fifty-five

Thibaut Munier
Administrateur de Renaissance Numérique

Gaëlle Recourcé
Directrice scientifique,
Evercontact

Marine Romezin
Communications Manager chez Squid Solutions

EQUIPE DE RÉDACTION DU LIVRE BLANC :

Luc Bretones
Vice président
Institut G9+

Henri Isaac
Vice président de Renaissance Numérique

Jean-François Vermont
Trésorier Institut G9+

Basile Michardiere
Chargé de mission
Renaissance Numérique

Camille Vaziaga
Déléguée générale
Renaissance Numérique

Pierre Balas
Chargé de mission
Renaissance Numérique

 Date de parution: décembre 2014