Cyril de Sousa Cardoso
Emmanuelle Galou
Aurore Kervella
Patrick Kwok

DATA POWER

Comprenez et exploitez la valeur de la donnée

Éditions EYROLLES

La *data* est un enjeu de pouvoir : l'organisation qui la maîtrise et la développe accède à une source de création de valeur majeure, en même temps qu'à une aide précieuse à la prise de décision. Mais comment mettre en place une stratégie dans des entreprises, organisations publiques ou associations qui, souvent, n'ont que peu de culture *data* ?

À la fois introduction et guide de mise en œuvre, cet ouvrage permet de faire les premiers pas :

- **Comprendre ce qu'est la *data*** et les différents modes de traitement de la donnée.
- **Anticiper l'impact de la *data* pour vous :** marketing, industrie, médias, finance, médecine, territoire, politique, etc.
- **Utiliser la *data* dans votre business :** explorer, apprendre, modéliser, prédire.
- **Décoder les enjeux et les perspectives de la *data* :** IA, IoT, questions écologiques, enjeux de propriété intellectuelle, etc.

Rédigé par quatre auteurs experts, *Data power* donne une vision large et concrète des enjeux et des applications de la donnée aujourd'hui.

Cyril de Sousa Cardoso est entrepreneur dans l'univers du conseil et des start-up. Il est l'auteur de plusieurs ouvrages sur le thème de l'innovation et expert en conduite de projets technologiques.

Emmanuelle Galou est experte en exploitation de la *data*. Elle a travaillé au sein de nombreuses start-up, de Deezer à Captain Contrat en passant par LeKiosk (Cafeyn).

Aurore Kervella dirige une équipe de *data scientists* chez Heetch. Elle a été *marketing analytics manager* chez Uber à San Francisco, après avoir mis son expertise *data* au service de Parfums Christian Dior et Sephora.

Patrick Kwok est statisticien dans le secteur public. Il a travaillé à l'Insee, au ministère de l'Écologie et à la Banque de France.

www.editions-eyrolles.com

Data power

Éditions Eyrolles
61, bd Saint-Germain
75240 Paris Cedex 05
www.editions-eyrolles.com

ISBN : 978-2-212-57389-3

Cyril de Sousa Cardoso
Emmanuelle Galou
Aurore Kervella
Patrick Kwok

Data power

Comprenez et exploitez la valeur de la donnée

Éditions EYROLLES

Sommaire

Les auteurs

Cyril de Sousa Cardoso est entrepreneur dans l'univers du conseil et des start-up et président du mouvement Innovation Commando. Il est l'auteur de plusieurs ouvrages sur le thème de l'innovation et expert en conduite de projets technologiques.

Emmanuelle Galou est experte en utilisation de la *data*. Elle a travaillé au sein de nombreuses start-up, de Deezer à Captain Contrat en passant par LeKiosk (Cafeyn), où elle a développé des stratégies à fort impact pour acquérir, convertir et retenir les clients.

Aurore Kervella dirige une équipe de *data scientists* chez Heetch. Elle s'est tournée vers le monde de la tech en tant que *marketing analytics manager* chez Uber à San Francisco, après avoir mis son expertise *data* au service du marketing et de la relation client pour Parfums Christian Dior et Sephora.

Patrick Kwok est statisticien dans le secteur public. Il a travaillé dans les domaines de la *data* au ministère de l'Écologie et à la Banque de France.

Remerciements

Nous tenons à remercier l'ensemble des experts qui ont accompagné l'écriture de cet ouvrage et dont les interviews ponctuent ce livre. Nous remercions également Lauriane Madec, Joëlle Galou-Chartier et Vincent Lehérissé pour leur soutien constant lors de l'écriture de cet ouvrage et leur relecture attentive.

Merci à Maël d'avoir participé silencieusement à nos rencontres.

Nous remercions l'ENSAI de nous avoir insufflé la passion de la *data* et de nous avoir permis de nous rencontrer.

Introduction

Alors que la donnée – ou *data* – est une source de connaissance pour la science, elle est aussi une source de valeur économique, sociale ou encore politique et son utilisation à travers les technologies du numérique est aujourd'hui exponentielle. Au cœur d'enjeux stratégiques pour nos entreprises et nos organisations publiques, la donnée numérique soulève des problématiques sociales et politiques. La donnée et son utilisation dessinent le monde dans lequel nous vivrons demain.

Cet ouvrage a pour objectif de donner au plus grand nombre les clés de compréhension de cette ère de la donnée dans laquelle nous vivons déjà, de la démystifier et de la rendre accessible. Il vise à mettre en perspective l'histoire de la donnée et de ses technologies pour mieux en comprendre les applications modernes, mais également les enjeux actuels et futurs. Ce livre s'adresse aux citoyens qui doivent s'éclairer sur le sujet, aux décisionnaires et à tous ceux qui souhaitent mieux comprendre la *data*, quel que soit leur domaine d'action, car au XXI^e^ siècle, la donnée est déjà partout.

Partie 1

Qu'est-ce que la *data* ?

Chapitre 1

L'histoire de la donnée

Un monde de données

9 h 15 min 00 s. 48° 54' 12.2" de latitude N et 2° 16' 10.8" de longitude E. Anissa passe à la caisse. En présentant sa carte de fidélité, elle provoque la remontée des informations de son ticket de caisse à un système marketing centralisé qui va mettre à jour son score et détecter en croisant son âge et son historique d'achats qu'elle vient d'avoir un bébé avec une probabilité de 98 %. Le système identifie qu'un email présentant une offre de réduction pour des couches aura plus de chances de provoquer un acte d'achat par Anissa s'il est envoyé dimanche à 19 h 50.

9 h 15 min 30 s. 46° 19' 04.8" de latitude N et 0° 25' 11.5" de longitude O. Au volant de leur véhicule, Fabien et Joëlle prennent une sortie d'autoroute. Leur passage devant le compteur radar qu'ils n'ont pas vu comptabilise leur véhicule parmi tous ceux passés par cette route depuis 6 heures. L'information va être exploitée par la direction générale des services tech-

niques de la ville pour optimiser la séquence des travaux de réfection de la voirie à venir, afin de perturber le moins possible le trafic. Devant eux, alors qu'il arrive au feu, Étienne freine brusquement, il n'avait pas anticipé l'arrêt du véhicule le précédant. L'ordinateur de bord de son véhicule électrique a enregistré ce freinage brusque et va faire remonter ces données accompagnées des informations fournies par le radar au système central du constructeur. Ces données vont être exploitées automatiquement pour perfectionner le système de conduite autonome des futurs véhicules de la marque qui ne produiront pas le type d'erreur de conduite réalisée par Étienne.

9 h 16 min 00 s. 43° 41' 49.4" de latitude N et 7° 16' 14.0" de longitude E. Maxime et Marie traversent la grande place en promenant leur fille en poussette. Au même moment, une des caméras placées en hauteur détecte leurs visages et les compare instantanément à une banque de données de personnes recherchées. Le système de traitement de l'image, en cours de tests, estime avec une probabilité supérieure à 95 % que ces individus n'en font pas partie. Simultanément, le smartphone de Maxime se connecte à une nouvelle antenne relais indiquant qu'il rentre sur la zone du front de mer. Cette connexion le comptabilise parmi les touristes fréquentant la zone depuis le début de la semaine. Information qui sera exploitée par l'agence de développement régionale pour mettre en lumière l'attractivité du quartier pour les nouveaux commerçants.

9 h 16 min 30 s. 48° 22' 57.5" de latitude N 4° 30' 55.6" de longitude O. Devant son écran, Vincent prend connaissance des opérations bancaires potentiellement frauduleuses détectées par le système central sur les comptes de ses clients et doit confirmer leur validité. Au sein de la même banque, Erwan s'apprête à recevoir Mélanie et Aurélien pour une demande de prêt immobilier. Exploitant l'historique de leurs comptes, leurs statuts professionnels et contractuels, leur situation fami-

liale, leurs âges, l'adresse et la typologie du bien qu'ils souhaitent acquérir, le système central confirme l'octroi de ce crédit estimant avec une probabilité supérieure à 97 % le bon remboursement futur de ce crédit.

9 h 17 min 00 s. 45° 44' 48.6" de latitude N et 4° 49' 34.2" de longitude E. Dans leur laboratoire, Magda et Claire prennent connaissance des statistiques issues de la dernière expérimentation médicale. La nouvelle molécule testée impacte significativement l'état de santé des patients tests. Ces résultats vont permettre à l'entreprise pharmaceutique qui a commandé l'étude de finaliser le dossier de mise sur le marché de son nouveau médicament.

Depuis que vous avez commencé à lire ce chapitre, à travers le monde, environ 16,5 millions de mégots de cigarettes ont été jetés par terre, 588 tonnes de poissons ont été pêchées et 360 requins tués pour leur aileron, 9 000 arbres ont été coupés. Aux États-Unis, 300 oiseaux sont morts en heurtant un pare-brise, 3 840 en s'écrasant contre un gratte-ciel et 14 000 ont été tués par des chats. En France, 1 440 yaourts encore consommables ont été jetés. Le principal glacier du pôle Sud, Pine Island, a perdu 305 000 m^3 cubes d'eau, 288 tonnes de sable ont été extraites des plages, tous les iPhones de la planète ont rejeté 9 tonnes de CO_2[1]. Durant ces deux minutes de lecture, dans le monde, environ 646 personnes sont mortes et 1 538 bébés sont nés. 125 millions de litres d'eau ont été consommés par l'humanité. 286 personnes sont devenues obèses et 122 sont mortes de faim. 854 voitures et 1 641 vélos ont été produits. 2 751 ordinateurs et 25 017 téléphones mobiles ont été vendus. 1 milliard d'emails a été envoyé[2].

Que ces chiffres vous impressionnent, vous effraient ou vous interrogent, bienvenue dans votre monde, un monde de données, plus ou moins fantasmées, invisibles et pourtant omniprésentes.

Données, informations et connaissances

Qu'est-ce qu'une donnée ? Une donnée, ou *data*, est un élément brut qui décrit de manière élémentaire une réalité, soit issue de l'observation, soit d'une mesure. Cette définition souligne le caractère « brut » d'une donnée, pas encore transformée en information. En effet, une donnée mise en contexte et interprétée s'enrichit d'une valeur ajoutée qui la transforme alors en information. Cette information devient à son tour connaissance si elle est comprise et utilisée pour aboutir à une décision ou une action. La donnée est donc l'élément de base du raisonnement humain, lui permettant de décider et d'agir.

Différents types de données

Une donnée peut prendre une multitude de formes : séquence de nombres, de lettres, de sons, d'images... En informatique, il est possible de distinguer les données structurées des données non structurées.

Les données structurées possèdent une structure préalable, c'est-à-dire qu'elle se définissent par différents champs dans une base de données ou par différentes balises dans un code informatique ou sur une page Web. Une donnée structurée est facilement interprétable par un programme informatique.

Les données non structurées se définissent par opposition comme des données sans structure préalable identifiée. Les contenus d'une série de textes, d'images, de pistes audio ou encore de vidéos peuvent être considérés comme des données non structurées.

On parlera de **données semi-structurées** quand une partie de la donnée peut être codifiée. Par exemple, une lettre écrite contient une adresse destinataire, le nom de l'expéditeur, un objet...

Une donnée peut également être qualifiée selon sa typologie. On distingue ainsi :

Les données quantitatives : ces données se réfèrent à des données numériques, des chiffres ou des données qui peuvent être converties en chiffres. Parmi les données quantitatives, on distingue les ***données discrètes***, ce sont des données dénombrables*, c'est-à-dire qui peuvent être comptées (souvent le comptage d'un nombre d'éléments, par exemple le nombre de voitures sur un parking), des ***données continues***, qui mesurent souvent des quantités ou des distances, qui sont par définition indénombrables (par exemple la quantité d'eau dans une bouteille).

Les données qualitatives : par opposition aux données quantitatives, ces données se réfèrent à la qualité d'un objet non quantifiable ou mesurable (la couleur d'une voiture).

On peut également définir les **données catégorielles** qui permettent de classer un élément dans des catégories qui peuvent se référer à des mesures quantitatives (par exemple classer les parkings en deux catégories : ceux de moins de 100 places et ceux de plus de 100 places) ou qualitatives (classer les voitures selon leur motorisation, essence ou diesel). Enfin, ces données catégorielles peuvent être **ordonnées** (classer les parkings selon leur nombre de places ou leur ancienneté).

Bien avant que nous nous mettions, en tant qu'Homo sapiens, à récolter et utiliser nos propres données pour produire de l'information et des connaissances, l'évolution naturelle a fait de nous, comme de tous les êtres vivants, de formidables centres de traitement de données. Nos sens (la vue, l'ouïe, l'odorat, le goût, le toucher) sont de véritables capteurs qui nourrissent en

* La réciproque n'est pas vraie, par exemple, l'ensemble des nombres rationnels est dénombrable mais pas discret.

nous des algorithmes* biochimiques transformant ces données en informations et en connaissances.

Ce que l'on appelle, par exemple, notre instinct de survie, se fonde sur un système biochimique complexe de traitement de données. Si vous vous retrouvez, lors d'une manifestation collective, au milieu d'un mouvement de foule, que vous ne pouvez ni voir ni savoir ce qui se passe, vos sens vont vous fournir un grand nombre de données et votre corps va en conséquence prendre des décisions favorisant votre survie. Votre vue va observer des personnes courir rapidement et des visages probablement marqués par la peur, votre ouïe va très certainement entendre des cris et tous vos sens vont vous indiquer que les personnes fuient le foyer probable du danger. Avant même que vous en ayez eu conscience, vous aurez commencé à courir dans la même direction. Vos algorithmes biochimiques et votre cerveau, façonnés à travers des centaines de milliers d'années d'évolution pour favoriser votre survie et votre capacité à vous reproduire, auront capté des données pour en extraire deux informations : la présence d'un danger imminent et le sens dans lequel fuir. L'information est immédiatement transformée en action grâce à l'augmentation de votre rythme cardiaque permettant notamment au sang d'affluer en quantité vers vos jambes, sollicitées pour courir.

La vérité, c'est les autres

Le psychologue Robert Cialdini décrit ce système de pilotage automatique comme celui de « la preuve sociale ». Façonné par l'évolution naturelle, notre système biochimique est tel qu'il semble considérer qu'en règle générale, nous faisons moins d'erreurs en agissant « conformément aux indications que nous fournit le groupe social qu'en allant à son encontre.

* Voir le lexique.

[...] Nous avons apparemment tendance à supposer que si beaucoup de gens font la même chose, c'est qu'ils savent quelque chose que nous ne savons pas. Surtout quand nous sommes incertains, nous sommes disposés à témoigner une confiance excessive à la science collective de la foule. »

Les systèmes de recommandation d'hôtels ou de restaurants s'appuient sur ce vieil algorithme biochimique. Cependant, comme le souligne Robert Cialdini dans son ouvrage *Influence et manipulation*[3], ce même système automatique et naturel de traitement de la donnée peut également se tromper lors, par exemple, des mouvements de foules qui peuvent se produire à la suite d'une détonation malencontreuse ne provenant pas d'une source dangereuse, par exemple un pétard festif.

Histoire de la « donnée »

Si nous sommes, en tant qu'êtres vivants, des centres naturellement performants pour traiter certains types de données, avec l'émergence de l'intelligence humaine est apparue en nous une obsession, celle de vouloir connaître et comprendre notre monde, en le qualifiant et en le quantifiant pour le résumer et le saisir. L'espèce humaine a pour cela développé, tout au long de son histoire, des outils de plus en plus performants. Dès la préhistoire, l'apparition de la peinture avec des fresques murales dans certaines grottes peut être envisagée comme une volonté de mémoriser et de stocker de l'information. Cette volonté s'accélérant il y a plus de 5 000 ans avec l'invention de l'écriture. Le besoin d'organiser et d'administrer les premières grandes cités humaines a probablement contribué de manière majeure à l'émergence de cette invention, qui était donc, avant tout, comptable. Les premiers documents écrits connus, découverts à Uruk, sont d'ailleurs des inventaires de biens[4].

Cette volonté d'accumuler et d'utiliser de la donnée semble d'ailleurs universelle[5]. En Chine, en 2238 avant J.-C., l'empereur

chinois Yao ordonne le recensement des productions agricoles. En Égypte, en 1700 avant J.-C., le pharaon Amasis organise le recensement de la population, tout comme le roi indien Chandragupta au IVe siècle avant notre ère. Des réflexions statistiques apparaissent en Grèce dès le VIIe siècle avant J.-C. et même les Incas organisaient des statistiques de récoltes. Le réflexe d'accumulation de données est donc ancien, se développant tout au long de l'histoire avec l'organisation d'inventaires, de registres et de recensements pour organiser les villes, les armées ou encore les impôts.

Les trois piliers de l'ère de la *data*

Cependant, l'accumulation de données ne suffit pas, seule, à expliquer l'apparition de l'ère de la *data* dans laquelle nous vivons aujourd'hui. Ce que nous définissons comme « *data* » ou « *data science* » repose en effet sur trois piliers.

Définition : la* data science *consiste à accumuler des données (premier pilier) pour en extraire de l'information (deuxième pilier) de manière automatique ou semi-automatique en exploitant des outils informatiques (troisième pilier).

Le premier pilier, l'accumulation de données, va donc se développer avec ce réflexe comptable et administratif ancien pour l'être humain. En revanche, l'extraction d'information de ces données va prendre son essor avec les sciences mathématiques et statistiques couplées à l'automatisation des traitements se développant avec l'algorithmique et l'informatique.

La volonté d'analyser et d'extraire de l'information des données dans une visée de prédiction apparaît au XVIIe siècle avec

des initiatives telles que celle de John Graunt qui cherche à prévoir à Londres l'apparition de la peste par l'analyse des registres de mortalité[6]. L'exploration de données va peu à peu se formaliser du XVIII^e^ au XX^e^ siècle avec Bayes, Laplace, Legendre ou encore Fisher[7]. Cependant, tant que les sciences statistiques s'appuient sur des calculs manuels, leurs applications restent limitées.

Le monde entre véritablement dans l'ère de la *data* dans la deuxième moitié du XX^e^ siècle avec l'apparition et la généralisation de l'informatique qui vient compléter l'accumulation de données et les sciences statistiques. Ce troisième pilier est lui-même fruit d'une longue histoire qui débute avec la formalisation des premiers algorithmes. Un algorithme est une suite d'opérations ou d'instructions à réaliser pour résoudre un problème. Les premiers algorithmes de résolution d'équations apparaissent chez les Babyloniens[8] dès le III^e^ millénaire avant J.-C. L'algorithmique se développe ensuite avec Euclide★, Archimède★★ et le mathématicien persan Al-Khwârizmî qui lui donnera son nom latinisé *Algoritmi*. C'est cependant Alan Turing qui va, en 1936, donner la première définition précise du concept d'algorithme, ouvrant l'ère des sciences informatiques et, en rencontrant les sciences statistiques et des données, l'ère de la « *data* ».

La formalisation, les méthodes et les outils de la *data science* vont aux XX^e^ et XXI^e^ siècles se développer de manière accélérée. Les premières méthodes de segmentation, de classification, d'arbres de décision vont se conjuguer à l'apparition de nouveaux types

★ Euclide, mathématicien de la Grèce antique, va définir dans le livre VII des *Éléments* un algorithme permettant de déterminer le plus grand commun diviseur de deux entiers sans connaître leur factorisation.

★★ Archimède va être le premier à définir un algorithme de calcul de la constante Π.

d'algorithmes, à l'augmentation de la puissance calculatoire informatique et à l'accroissement exponentiel de la quantité de données captées et stockées[9].

Avec l'avènement de l'informatique grand public, des cartes à puce, d'Internet, du GPS, de la téléphonie mobile, des réseaux de télécommunication ou encore des objets connectés, les êtres humains et leurs environnements sont de plus en plus nombreux à être hyper connectés, pourvoyeurs et utilisateurs de données à travers leur géolocalisation ou encore les réseaux sociaux. Les sources de données se multiplient en même temps que se développent les capacités de les stocker et de les analyser pour en extraire rapidement informations et connaissances. Les calculs et algorithmes se développent dans le *cloud*, les données, l'information et les instructions circulant à travers des réseaux de plus en plus rapidement, à la vitesse de la 4G et maintenant de la 5G. La *data* est la nouvelle électricité. Si sa circulation est invisible à nos yeux, elle est pourtant partout.

Dans la banque, la finance ou l'assurance, la *data* structure nombre de processus décisionnels, de l'octroi de crédits à la détection de fraudes en passant par des décisions d'investissement. Elle est exploitée pour assurer de la surveillance, programmer des maintenances et détecter des incidents dans l'industrie, l'informatique, l'aéronautique ou encore l'automobile. La donnée est utilisée pour développer l'efficacité des réseaux de transport, de distribution, de télécommunication ou encore pour développer la productivité de l'agriculture. Les scientifiques, de l'économiste au météorologue, en passant par les chercheurs en génomique, en épidémiologie, en imagerie médicale, en astronomie ou encore en physique nucléaire l'exploitent. La donnée façonne notre environnement en développant notre compréhension complexe des mécanismes écologiques, mais aussi en influençant la ville et l'urbanisme à travers les domaines de la ville intelligente (*smart*

city) ou encore du réseau électrique intelligent (*smart grid*). La *data science* influence les médias et les pratiques journalistiques. Elle est devenue le cœur de métier du marketing, de la publicité ou encore du e-commerce.

★★

Références bibliographiques du chapitre 1

1. Cette statistique et toutes celles qui précèdent sont calculées à partir des données fournies dans l'ouvrage de Mathieu Vidard, *Le Carnet scientifique*, Grasset, 2016.

2. Les 10 statistiques qui précèdent cette note ont été calculées grâce au site Worldometers : https://www. worldometers.info/fr/

3. Robert Cialdini, *Influence et manipulation : comprendre et maîtriser les mécanismes et les techniques de persuasion*, First Éditions, 2004.

4. Jean-Christophe Messina et Cyril de Sousa Cardoso, *L'Art de l'Innovation*, Éditions Eyrolles, 2017.

5. Jean-Claude Oriol, *Éléments d'histoire de la statistique*, HAL, archives-ouvertes.fr, 23 mars 2010.

6. Jean-Marc Rohrbasser, « John Graunt et les bulletins de Londres : une statistique de la mortalité au XVII^e siècle », *Dix-septième siècle*, 2009/2 (n°243), pp. 345-368 : https://www. cairn.info/revue-dix-septieme-siecle-2009-2-page-345.htm

7. Jean-Paul Benzécri, « Histoire et préhistoire de l'analyse des données », *Les Cahiers de l'analyse des données*, 1977.

8. Donald Knuth, *Éléments pour une histoire de l'informatique*, Éditions Eyrolles, 2011.

9. « Pendant la seule année 2011, le volume de l'information qui a été numérisée dans le monde a atteint 10^{21} octets. En 2013, ce volume a été 4,4 fois supérieur. À ce rythme, en 2020, l'humanité

devrait stocker 44 zettaoctets, soit 44 000 milliards de gigaoctets de données » dans ses *data centers*, ordinateurs, tablettes, smartphones et autres objets connectés. Mathieu Vidard, *Le Carnet scientifique*, Grasset, 2016.

Chapitre 2

Data science : les différents modes de traitement de la donnée

L'ÉMERGENCE DE LA *DATA SCIENCE*

La *data* est un concept à la fois global et flou qui regroupe les données, leurs modes de traitements et les valeurs qu'on en extrait. S'il est difficile de donner une date de naissance à cette discipline, le terme de *data science* semble émerger entre les années 1990[1] et 2000[2]. Cette science s'affirme rapidement comme une nouvelle discipline qui cherche à résoudre des problèmes par l'exploration de données. Les opportunités des sciences de la donnée se sont développées au fil de l'enregistrement et du stockage d'informations et de connaissances de plus en plus nombreuses et complexes, à travers notamment le développement de l'informatique grand public, d'Internet, des réseaux de télécommunication ou encore de la téléphonie

mobile, certains parlant d'un « big bang de l'information stockée[3] ». Couplée au développement de la puissance de calcul des outils informatiques, la *data science* s'est peu à peu enrichie de nouvelles disciplines, de la théorie de l'information, au traitement du signal et de l'image, en passant par la compression, l'apprentissage automatique ou la visualisation de données.

Le terrain de jeu du *data scientist* se développe avec l'innovation dans les sciences des nanotechnologies, des biotechnologies, des technologies de l'information et des sciences cognitives (ces quatre domaines étant regroupés sous le sigle de NBIC). L'intérêt pour la discipline bénéficiant de l'intérêt grandissant pour le *big data* et l'intelligence artificielle.

Big data

La *data science* est une discipline qui s'intéresse à tous les ensembles de données, quels que soient leurs origines, leurs volumes ou encore leurs structures (structurées, non structurées, semi-structurées…). C'est cependant le besoin de traiter ces ensembles de plus en plus massifs qui ont fait de la discipline un domaine stratégique pour les entreprises, les organisations ou encore les scientifiques. Pour bien comprendre et définir ce qu'est le *big data*, il est nécessaire de ne pas s'arrêter uniquement à l'idée de traitement de bases de données massives. Il est indispensable de comprendre que les technologies de l'information n'ont pas simplement développé le volume des données stockées, mais qu'elles ont provoqué une explosion de la quantité de données au-delà de notre capacité cognitive à le saisir, poussant également à leur limite les puissances de calcul et de stockage de nos outils informatiques « classiques ».

Le *big data* marque le passage à une autre échelle, où les ordres de grandeur explosent★.

Dans l'univers de la finance, par exemple, le *trading* haute fréquence cherche à générer des ordres d'achat et de vente à une fréquence relevant de la nanoseconde, insaisissable pour l'esprit humain. Des prises de décision avec une telle vélocité nécessitent que les systèmes captent d'immenses volumes de données, les traitent et en extraient des connaissances pour prendre des décisions à la même fréquence.

À cette nouvelle échelle, trois enjeux permettent de vraiment définir la problématique du *big data* et sont symbolisés par la règle des 3V★★ : Volume, Vélocité et Variété. Le cabinet de conseil et de recherche Gartner définit le *big data* comme « des bases de données volumineuses, créées ou traitées à grande vitesse et/ou dont les données sont très variées[4] ». Le *big data* (ou « mégadonnées ») a aussi depuis le 22 août 2014 une définition au *Journal officiel* : « données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés[5] ».

Derrière ce problème de « taille » se trouvent plusieurs défis : les capacités de stockage et d'accès, mais aussi celles d'un traitement et d'une analyse rapides et efficaces, en faisant appel aux technologies de l'information. Car, toujours selon Gartner : « ces bases de données *big data* nécessitent des formes de traitement de l'information novatrices et rentables qui permettent d'automatiser les processus pour développer des savoirs ou

---

★ Pour bien saisir à quelle vitesse le stockage de données explose, il suffit de noter qu'en 2010, environ 1,2 zettaoctet supplémentaire de données a été stocké, puis 1,8 zettaoctet en 2011 et 2,8 zettaoctets en 2012. En 2020, 40 zettaoctets de données seront stockés.

★★ Rejoints rapidement par la Véracité et la Valeur des données, menant le compte à 5V pour décrire le phénomène *big data*.

prendre des décisions* ». L'échelle *big data* impose une complexité des traitements due au volume des données traitées, mais aussi à leur variété (diverses dans leur format et leurs sources) et à la fréquence (vélocité) à laquelle il faut à la fois les générer, les capturer ou encore les partager. Le *big data* ne part pas de modélisations préexistantes pour extraire de l'information des données (ce que fait l'informatique décisionnelle par exemple), il est capable de trouver des modèles directement dans les données, sans *a priori*, en utilisant des techniques statistiques sur des volumes inédits.

Uber, le service de voiture de transport avec chauffeur (VTC) né en 2009, utilise une quantité et une variété importante de données pour exercer son activité. La société analyse non seulement les réseaux de transport public des villes où son service est disponible, mais aussi chaque trajet. Cette analyse lui permet d'anticiper la demande en période de pénurie et surtout de répartir efficacement ses ressources : en utilisant le service, les chauffeurs ajustent en temps réel le prix de l'offre à la demande. Uber a repoussé à une nouvelle échelle les principes de tarification dynamique** utilisés couramment par les hôtels et les compagnies aériennes : les données sont volumineuses (couverture géographique et utilisation de son application de plus en plus importantes), de types variés (coordonnées géographiques, données textuelles converties en localisation) et son service doit répondre en quelques secondes à chaque sollicitation imposant une cadence élevée à ses infrastructures. La précision de cette

---

* C'est également ce que l'on appelle le « broyage de données » (ou *big analytics*) dans le *big data*.

** La tarification dynamique consiste à ajuster les prix aux variations de la demande. C'est une stratégie exploitée dans le domaine du *yield management* qui vise à gérer les tarifs d'un service en fonction des disponibilités pour optimiser le « remplissage » et donc le chiffre d'affaires. Domaine exploité par exemple dans l'hôtellerie ou le transport ferroviaire ou aérien.