



GILLES
BABINET



Big Data,
penser l'homme
et le monde autrement

Préface d'Erik Orsenna
de l'Académie française

LE PASSEUR
— EDITEUR —

BIG DATA, PENSER L'HOMME ET LE MONDE AUTREMENT

Issues de nos réseaux sociaux, de notre téléphone mobile, de notre passage en caisse, de l'utilisation d'une carte de transport, les données sont partout. Gilles Babinet montre que les techniques de *Big Data* peuvent être utilisées pour développer des outils marketing perfectionnés, améliorer la prise en charge médicale, permettre aux villes de mieux gérer leurs ressources ou encore allier productivité agricole et développement durable. Les données pourraient bien être l'une des énergies motrices de la nouvelle révolution globale en cours.

Pourtant, de nombreuses questions se posent. Sommes-nous sur le point d'entrer dans une société semblable à celle que décrit George Orwell dans *1984*? Quels facteurs permettraient à une société démocratique et respectueuse des libertés individuelles de bénéficier de cette nouvelle technologie ? Gilles Babinet fait toute la lumière sur ces questions en rappelant que ce ne sont pas les technologies qui façonnent l'humanité, mais bien l'usage que l'on choisit collectivement d'en faire.



D. P.

Gilles Babinet a créé de nombreuses sociétés dans divers secteurs. Premier président du Conseil national du numérique, puis « Digital Champion », il représente la France auprès de la Commission européenne pour les enjeux du numérique. Il est l'auteur de *L'Ère numérique, un nouvel âge de l'humanité* (Le Passeur Éditeur, 2014).

ISBN 978-2-35890-260-8



9 782368 902608

Diffusion Solédis
Code Sotis S560141

20,50 € TTC

Table

PRÉFACE.....	9
PRÉAMBULE.....	11
INTRODUCTION.....	19

PREMIÈRE PARTIE

Big Data, genèse et évolution

1. Les origines.....	29
<i>Big what ? (29) – Un peu de mathématiques (35) – La règle des 3 V (40) – Révéler (43) – Prédire (46) – Réagir (52) – Manipuler des données : enfer et paradis (55)</i>	
2. Vivre mieux et en meilleure santé.....	57
<i>La fin du système de santé, tel que nous le connaissons (57) – Pas de nouvelle stratégie de santé (58) – Un océan d'opportunités (63) – Épidémiologie et A/B Testing (67) – Des capteurs et des données (70) – La marginalisation du service de santé publique ? (72) – Open data, épidémiologie et médicaments (73) – Centraliser les données de santé ? (75) – Le blue button et les data personnelles (78)</i>	
3. Agriculture, environnement et complexité.....	85
<i>Des champs de données (85) – L'exemple du Qatar en agriculture (92) – Environnement : le Big Data pour sauver le monde ? (94)</i>	

4. Des villes, et des fluides qui les animent..... 102
Big Data, ville et sécurité : compatible avec les libertés publiques ? (106) – *Des poubelles et des déchets urbains* (109) – *Un peu d'open data* (111) – *À propos des démocraties rationnelles* (113) – *L'eau* (115) – *Smart Grid : au-delà du buzzword, une nouvelle approche* (118) – *Big Data et consommation* (122)

DEUXIÈME PARTIE

Entrer dans l'univers des data

1. Culture ou technologie ?..... 129
L'ère de la plateforme (129) – *Étape 1. Comprendre et faire comprendre le potentiel des données, au sein de leur organisation* (132) – *Étape 2. Identifier des domaines d'application prioritaires en interne et au travers de partenariats* (133) – *Étape 3. Jouer avec les données* (135) – *Étape 4. Administrer et faciliter l'usage de données partagées* (137) – *Étape 5. Construire une offre compréhensible et attractive pour l'écosystème numérique ainsi que pour les utilisateurs finaux* (138) – *Étape 6. Mobiliser l'écosystème numérique pour développer les services thématiques* (140) – *Former, recruter et gagner en agilité digitale* (141) – *Identifier et nommer un chief digital officer (CDO)* (142) – *Data scientist : vraiment le métier le plus cool du monde ?* (144) – *Data miner, data analyst, data manager* (145) – *Le chief privacy officer : gadget ou nécessité ?* (146) – *Devenir data-compatible* (148)
2. Data, Big Data et... marketing..... 149
Agences publicitaires et agences créatives (149) – *Marques et annonceurs* (151) – *Le panel comme outil de compréhension du consommateur* (153) – *Le plan médias* (156) – *Communautarismes numériques* (159) – « *Marketing as a Big Data platform* » (161) – *Big data et marketing : pourquoi c'est évident* (165)

3. L'impact économique des données.....	168
<i>Servification (168) – Les gains d'opportunité (175) – La fin de la croissance et des infrastructures ? (177) – Effet de mode ou révolution ? (181)</i>	

TROISIÈME PARTIE

Nous et les machines

1. Surveillance, machines et transparence	185
<i>Le Big Data, interprété par la NSA (185) – Contre-pouvoirs ? (193) – Liberté, égalité, data : écoutes et régulation des écoutes (197)</i>	
2. Le code des données	202
<i>Politique publique, prévision et modèle de société (203) – Observer la puissance de la pensée collective (209) – Plateformes : nouveaux agents économiques et forces de lobbying (212) – La régulation : affaire de spécialistes ou confiée aux citoyens ? (221) – Les données peuvent-elles mentir ? (227) – Le code des données (229)</i>	
OUVERTURE : La nécessité de choisir.....	231
Lexique	245
Remerciements	249

Préface

BIG Data *ou* Big Brother ? Big Data *et* Big Brother ? Il y a quelque temps, la science-fiction nous faisait frémir en annonçant un âge où les ordinateurs domineraient les humains. Aujourd'hui, la menace se précise. Les « données » ont-elles pris le contrôle de nos vies ?

Désormais, nous sommes suivis, pistés, démasqués, mis en catégories, enregistrés. À chacune de nos innombrables connexions quotidiennes, nous dévoilons un peu plus de notre intimité. On sait tout de nous, de nos préférences, de nos espoirs, de nos petites manies inavouables...

Qu'en est-il de nos droits, à commencer par deux d'entre eux, le droit à l'ombre et le droit à l'oubli ? Comment vivre sous ce projecteur perpétuel ? Que fait-on de toutes ces informations qui nous ont été volées ? Et qui les rassemble, et qui décide, un beau jour, une sale nuit, de les utiliser ?

Gilles Babinet est une personne active. À 47 ans, il a déjà créé neuf entreprises, dont CaptainDash, spécialisée, justement, dans l'analyse de ces « données ». Nicolas Sarkozy le nomme au Conseil national du numérique, chargé d'informer les autorités publiques sur les enjeux de cette révolution. Il en sera le premier président. Aujourd'hui, il est devenu Digital Champion,

autrement dit le représentant de la France auprès de la Commission européenne pour toutes ces affaires.

Dans un ouvrage récent et passionnant, *L'Ère numérique, un nouvel âge de l'humanité*, Gilles Babinet nous avait expliqué les changements apportés par les nouvelles technologies dans cinq domaines clefs : la connaissance, l'éducation, la santé, la production et l'État. Effarés, nous prenions conscience des bouleversements à l'œuvre. Ravis, nous découvriions des potentiels inimaginables. Furieux, nous constations les retards de notre pays. Inquiets, nous nous demandions quelle sorte de société était en train d'advenir : celle décrite par George Orwell dans *1984*, une sorte d'irrespirable tyrannie ? Ou un nouvel humanisme, plus riche en libertés, ouvert sur de joyeux potentiels ?

C'est pour apporter quelques éléments de réponse à cette question que Gilles Babinet a écrit le livre que vous allez lire.

Il faut dire que ce passionné a déclaré un jour que la Cnil, Commission nationale informatique et libertés, est un « ennemi de la nation ». Cette Cnil n'avait-elle pas pour mission, ô combien noble et utile, de protéger le citoyen ? C'est bien beau la modernité, mais si l'on doit y perdre son âme... Cette affirmation, pour le moins choquante, devait, au plus vite, être expliquée. Et justifiée. À vous la parole, Gilles Babinet.

Vous l'avez bien compris, il y va du type de monde dans lequel nous allons vivre. Ne nous y trompons pas, l'espèce humaine va devoir affronter *en même temps* trois transitions : climatique, biologique et numérique. Comment en sortirons-nous ? *Augmentés* ou *machinisés* ?

ERIK ORSENNA,
de l'Académie française

Préambule

MON téléphone se mit à sonner – longuement – avant que je sorte finalement de mon sommeil. Mes yeux accrochèrent le cadran numérique d'une horloge : il était 3 heures du matin. Comme souvent lorsque je voyageais à l'étranger, j'avais manifestement oublié d'actionner la fonction veille pour éviter les appels intempestifs. Peu à peu, il me revint à l'esprit que j'étais à Las Vegas, pour un séminaire dont l'un des sujets était justement les data. Trois heures du matin, cela voulait dire midi à Paris. Je laissai passer l'appel et tentai de me rendormir lorsque le téléphone sonna de nouveau. J'étendis le bras et trouvai à tâtons mon smartphone : le numéro n'était pas identifié mais je décrochai malgré tout. C'était un journaliste qui tenait à évoquer ma « déclaration à l'égard de la Cnil^{*1} ». J'avoue que sur l'instant, je fus incapable de comprendre de quoi il s'agissait : je bredouillai qu'il serait préférable de me rappeler d'ici quelques heures puis raccrochai, dubitatif et vaguement inquiet : quelque chose d'important était-il en train de se passer sans que je comprenne de

1. Les mots suivis d'un astérisque sont répertoriés et explicités dans le lexique, p. 245.

quoi il s'agissait ? J'escomptai pourtant me rendormir et essayai, dans un état semi-éveillé, de mettre mon téléphone en mode avion. C'est alors que je réalisai que de nombreux SMS m'étaient parvenus au cours des dernières heures : deux attirèrent particulièrement mon attention. L'un de Jean-Christophe, un de mes proches amis : « Une nouvelle croisade ? – *Good Luck*. » Et l'autre, de Fleur Pellerin, alors ministre en charge du Numérique : « Tu as pétié les plombs ? » Manifestement, quelque chose d'anormal venait de se produire.

Je me levai, pressentant que je ne pourrais plus dormir. Je remis mon téléphone en mode normal. Presque immédiatement, un nouvel appel se fit entendre. C'était un journaliste de 01Net, un site Web spécialisé. Il me parlait d'une interview et de mes déclarations. Je pris peu à peu conscience de ce qui arrivait : un journaliste avait titré le papier issu d'une interview que j'avais donnée pour *L'Usine nouvelle* avec une affirmation que j'avais lâchée sans trop y faire attention : « La Cnil* est un ennemi de la nation. » Évidemment, ça faisait tache ; et surtout ça avait l'air d'en fâcher plus d'un.

J'ouvris mon ordinateur : des messages arrivaient presque en continu. Un autre site Web, Numerama, avait déjà fait un article pour commenter mes déclarations. Ça n'était pas tendre. Les commentaires des internautes non plus. Rares étaient ceux qui comprenaient que l'on puisse s'en prendre à la Cnil* ; cela donnait le sentiment que j'avais commis un crime de lèse-majesté. Puis, je lus l'article de mon interview dans *L'Usine nouvelle*.

Je réalisai que rien de ce qui était écrit n'était faux, mais que la journaliste avait surtout retenu mes commentaires « d'ambiance », qui émaillaient le fond de notre échange,

et qu'elle les avait articulés d'une manière qui faisait ressortir un esprit général très agressif à l'égard de la Cnil*.

Depuis quelques mois je travaillais avec l'Institut Montaigne à la rédaction d'un rapport sur les facteurs d'agilité numérique des nations. Nous avons auditionné beaucoup de monde, des grandes entreprises du numérique, d'autres traditionnelles, des directions d'administrations publiques, ainsi que la Cnil*. Avec cette dernière, cela s'était plutôt mal passé. Je me souviens encore avoir reproché à son secrétaire général d'avoir fait en sorte que son institution défende des positions que je jugeais archaïques et, surtout, de nature à amplifier encore l'importance du principe de précaution, à présent gravé dans notre Constitution. Ce n'était d'ailleurs pas la première fois que je m'accrochais avec la Cnil*. Quelques semaines auparavant, lors du traditionnel déjeuner des ambassadeurs, Fleur Pellerin m'avait demandé d'intervenir pour parler du numérique en général. Je me souviens que j'étais alors assis à table aux côtés des ambassadeurs de Chine, d'Allemagne, des États-Unis, de Fleur Pellerin et... d'Isabelle Falque-Pierrotin, la présidente de la Cnil*. On nous avait tous deux demandé de résumer en quelques minutes en quoi le digital était sur le point de bouleverser le monde.

Il était convenu que j'interviendrais juste après Isabelle. Je l'avais patiemment écoutée, non sans noter combien j'étais en désaccord avec son approche. Lorsque ce fut mon tour, je ne pus me retenir de lâcher quelques observations sur le fait que son institution était l'une des causes qui expliquaient le retard français. Par exemple, en ce qui concernait le plan de numérisation du système de santé, dans la mesure où elle avait émis des recommandations alambiquées, compliquant d'autant

les développements du système d'information¹. Fleur était restée impassible, mais je la connaissais suffisamment pour savoir qu'elle avait modérément apprécié. Elle comptait sur moi pour la soutenir, mais maintenant qu'elle était ministre, elle avait probablement le sentiment que je ne faisais que la mettre en danger. Ce fut d'ailleurs l'une des dernières fois où j'intervins à ses côtés.

Mais à présent que cet article incendiaire était paru – le contenu était à la même enseigne que le titre –, on me demandait de m'expliquer. En quoi la Cnil*, dont la mission est avant tout de protéger les citoyens des abus des entreprises et des administrations publiques, pourrait-elle être condamnable ?

J'avais bien quelques idées. Cela tenait à peu près la route, mais ce n'était ni clair ni structuré. J'ai annulé ma participation à deux réunions sans grande importance prévues dans la matinée et passais deux heures à lire ce que je pouvais trouver sur la gestion des données personnelles. C'était passionnant et surtout cela ébranlait mes certitudes. Finalement, que savais-je vraiment de ces enjeux ? Est-ce que je pouvais me permettre de remettre en cause le travail d'une institution – la Cnil* – qui emploie des centaines de personnes et qui existe depuis plusieurs décennies, sans accident notoire ? D'autres questions également émergeaient. Pourquoi les données semblaient-elles faire si peur ? Et pourquoi ceux qui semblaient les craindre le plus étaient-ils souvent eux-mêmes des utilisateurs fréquents des services d'Apple, Google, Amazon ou Facebook, pour ne citer qu'eux ?

1. La Cnil s'est notamment opposée à ce que le Nir (numéro d'identification des usagers du système de santé) soit l'identification unique, rendant difficile la gestion des dossiers par identifiant électronique (www.cnil.fr/fileadmin/documents/approfondir/dossier/NIR/Rapport%20NIR.pdf).

Je réalisais que, derrière la technologie, il y avait également des visions du monde. Celle de l'Europe, qui avait, il y a cent trente ans, vu l'innovation comme un projet sans concession, pour finalement l'aborder avec beaucoup plus de circonspection à l'issue de deux guerres mondiales. Celle des États-Unis, qui l'avaient utilisée pour s'affirmer peu à peu au cours du xx^e siècle et aussi celle de l'Asie, sans doute plus pragmatique encore à son égard.

Dès mon retour de Las Vegas, plusieurs médias me demandèrent de préciser mon propos. Lors de la première interview, sur la radio et télévision BFM, je me suis rendu compte qu'il n'était pas si simple de faire une réponse unique, qui embrassait tous les propos et toutes les situations. Bien entendu, il y avait le principal : je n'avais rien contre la régulation à l'égard du numérique et des données, mais elle ne devait pas renforcer le nihilisme ambiant que l'on observe trop souvent en France. La Cnil* caractérise – encore aujourd'hui – à mon sens cette crispation française à l'égard du futur et du progrès, qui nous a fait inscrire le principe de précaution dans notre Constitution. Le propre de l'homme est d'essayer et essayer signifie aussi échouer. Si nous cherchons à réguler le progrès, nous sortons *de facto* de la route du futur. L'enjeu n'est pas d'avoir une seule forme de progrès, mais bien d'en avoir plusieurs et elles ne sont pas toutes nécessairement technologiques. Mais toutes sont des tentatives avec des risques inhérents. Et si nous devions attendre qu'une armée de juristes ait préalablement « débeugué » tous les types de risques – ceux relevant d'une probabilité notoire et d'autres, imaginaires – il est probable que nous y perdions une partie de notre âme, celle d'une nation innovante et tournée vers l'avenir.

Tout cela, j'étais capable de l'énoncer à peu près clairement. En revanche, il m'apparut crûment que je n'avais pas nécessairement de position arrêtée sur beaucoup de sujets. À moins d'être manichéen, il devenait difficile de continuer à débattre au sujet des données.

En réalité, je ne cessais de me confronter à de nombreuses questions : Google a-t-il vraiment le moyen d'imposer une situation de monopole à l'égard des données ? En a-t-il le projet ? Et, si oui, comment réguler ? Est-ce que la souveraineté territoriale a encore un sens dans un monde largement virtualisé et ubiquitaire ? Les États ont-ils vocation à devenir des dictatures numériques ? Y a-t-il un risque de type *Minority Report*¹ que l'on nous condamne avant que l'on ait commis un délit ?

Nombreux, parmi nous, ont l'intuition qu'il existe un sens commun qui rend évidentes les réponses à l'égard de chacune de ces questions. Certains, par exemple, aimeraient couper Google en morceaux ou le nationaliser. D'autres pensent que les États vont rapidement jouer à Big Brother², simplement parce que les technologies le permettront et qu'il faut donc, dès à présent, les en empêcher. D'autres encore craignent une forme de lobotomisation de l'esprit du fait de l'automatisme des machines. D'autres enfin aimeraient ériger des barrières numériques autour de l'Europe, pour que seules les données qui sont à l'intérieur puissent circuler libre-

1. *Minority Report* (2002) est un film de science-fiction de Steven Spielberg, adapté d'une nouvelle de Phillip K. Dick. L'action se déroule en 2054 et des êtres humains mutants, doués de prescience, prédisent les crimes à venir.

2. Big Brother : nom d'un personnage du roman de science-fiction de George Orwell, 1984. Big Brother est devenu l'incarnation de l'État policier et de la perte des droits individuels.

ment et qu'un contrôle strict ait lieu sur le peu d'activité des grandes entreprises californiennes que l'on aurait continué à tolérer.

Tout cela semblait fort envisageable et je ne cherchais pas à le nier. Simplement, il me semblait impossible ne serait-ce que d'essayer de répondre honnêtement à ces questions sans en avoir fait le tour ou, tout au moins, sans avoir réellement essayé de comprendre de quels types d'enjeux elles relevaient.

À force d'en discuter autour de moi, plusieurs personnes m'ont invité à écrire un livre à ce sujet. J'étais alors en pleine rédaction de *L'Ère numérique, un nouvel âge de l'humanité*, un essai que j'ai finalement publié en 2014. Je ne me voyais pas me remettre aussi rapidement à écrire un autre ouvrage. Je m'aperçus pourtant que j'avais amassé tant de notes et disposais de tant de notions, qui me semblaient finalement passionnantes à expliciter, que le projet du livre s'est imposé tout seul. C'est celui que vous tenez entre vos mains.

Alors que vous êtes sur le point d'en commencer la lecture, je ne peux que vous recommander une chose : prenez quelques minutes pour lire le lexique qui se trouve à la fin de l'ouvrage (p. 245). J'ai sincèrement cherché à éviter d'être jargonnant. Mais j'ai dû faire quelques entailles à ce vœu. Aussi, il me semble que si vous aviez à l'esprit les quelques mots compliqués qui sont évoqués à longueur de pages, cela vous rendrait la lecture sans doute plus fluide et accessible et vous permettrait de comprendre, sans avoir de bagage technique particulier, cette notion si particulière qu'est le Big Data.

Introduction

IL est vain de chercher à retenir le progrès : pas un exemple dans l'histoire ne nous permet de croire qu'une technique accessible ait pu être durablement cantonnée. Qu'il s'agisse de la soie chinoise ou du métier à filer anglais, les exemples sont nombreux des techniques que les États ou les inventeurs ont cherché à préserver pour leurs intérêts propres, sans finalement y parvenir. C'est pourquoi l'avènement des technologies liées aux données ne représente désormais plus une hypothèse que l'on peut infléchir, mais bien une certitude à laquelle nous devons nous préparer. Le choix que nous pouvons encore faire consiste, en revanche, en celui d'un modèle d'émergence de ces techniques qui soit le plus harmonieux pour nos sociétés. Il nous faut donc comprendre, le mieux possible, ce que permettent les données, ce que sont leurs opportunités, leurs dangers ; et cela, le plus vite possible. Car l'avènement de cette révolution est proche. Cela va vite, de plus en plus vite.

Ainsi, l'humanité accroît sa production de données de façon étourdissante : chaque minute, environ 350 000 tweets, 15 millions de SMS, 200 millions de mails sont envoyés dans le monde, tandis que des dizaines

PREMIÈRE PARTIE

Big Data, genèse et évolution

Les origines

En 1941, l'*Oxford English Dictionary* évoqua la multiplication des données sous l'expression « *information explosion* ». C'était une première. L'information était si précieuse que personne n'avait songé au fait qu'il pourrait y en avoir trop. Trois ans suffirent pour qu'un universitaire américain, Fremont Rider, évoque le phénomène d'« explosion des données ». Il estimait que la taille des bibliothèques américaines doublait tous les seize ans et que, en se fondant sur ce taux de croissance, celle de Yale contiendrait environ deux cents millions de volumes, nécessitant plus de six mille employés pour garantir un accès convenable aux ouvrages. Il fut suivi dans les années 1950 par Derek Price, qui publia un article mettant en lumière cette évolution exponentielle : le nombre d'articles universitaires, de son point de vue, serait amené à être multiplié par dix tous les siècles. Il s'agissait là de signes avant-coureurs d'un phénomène dont personne n'avait encore idée de l'ampleur.

Big what ?

Il est assez difficile de retracer précisément l'histoire de l'expression « Big Data ». Le premier à l'avoir

Vivre mieux et en meilleure santé

La fin du système de santé, tel que nous le connaissons

C'est plus qu'une hypothèse : il suffit de se rendre dans un service d'urgence pour en faire le constat. La qualité de notre médecine ne cesse de se dégrader. De déremboursement en déremboursement, nombreux sont ceux qui ne parviennent plus à se soigner convenablement. En conséquence, ils se rendent aux urgences, où seul le ticket modérateur est dû, pour des services qui restent malgré tout d'assez bonne qualité.

Il faut l'admettre : la santé dans l'ensemble des pays riches (ceux de l'OCDE) ne se porte pas au mieux. Plusieurs études pointent du doigt le fait que dans certaines catégories sociales, on observe une diminution de la durée de vie en bonne santé¹; une première depuis l'émergence de la médecine moderne. De nombreux signes montrent que le système est à bout de souffle. Malgré des budgets de plus en plus importants, la qualité ne suit pas. Au sein de l'OCDE, on estime que de 10 à 17 % des diagnostics médicaux sont tout simplement erronés, avec tous les risques que cela induit pour les patients².

1. Voir les travaux de l'Ined et de l'OMS sur : www.ined.fr et : www.oecd-ilibrary.org.

2. « Diagnostic Errors are the most common type of Medical Mistake », *Time*, avril 2013.

Agriculture, environnement et complexité

Des champs de données

Parmi les acteurs économiques, les agriculteurs sont, depuis longtemps, en pointe en matière d'utilisation des technologies. Qu'il s'agisse de la mécanisation, de l'emploi d'intrants, de celle d'automates de traite ou de distribution alimentaire, ou encore de l'informatisation, l'adoption a été précoce et souvent très massive. Cette aisance avec la technologie s'exprime notamment par le concept d'« agriculture de précision », apparu aux États-Unis au milieu des années 1990. L'idée initiale a consisté à permettre à des tracteurs dotés de GPS et de systèmes de pilotage automatique de se frayer un chemin entre les rangs de maïs pour déposer les quantités adéquates d'engrais et éviter de passer deux fois au même endroit. Au-delà, il s'agissait également d'effectuer des mesures en grand nombre pour ajuster les systèmes d'irrigation et les volumes d'engrais et autres pesticides nécessaires à chaque parcelle de terre. Et cela fonctionnait. L'American Farm Association évoque des rendements accrus de 10 à 20 % lorsque ces techniques sont mises en œuvre. À tel point que, aux États-Unis, il n'est plus envisageable

Des villes, et des fluides qui les animent

DEUX MILLE DOUZE aura marqué l'histoire de l'humanité pour être la première année où plus de 50 % de la population mondiale vit désormais dans les villes, soit environ 3,6 milliards de personnes. En Asie, en Amérique du Sud et maintenant également en Afrique, ces villes dépassent fréquemment les 10 millions d'habitants, et cette tendance à croître n'est pas près de s'arrêter. Selon l'Onu, d'ici 2040, ce ne seront plus 50 %, mais bien les deux tiers de l'humanité qui vivront dans des villes. Ainsi une mégapole comme Lagos croîtrait au rythme insoutenable de 12 % par an. Inutile de souligner combien de défis ce développement pose aux autorités qui en ont la charge. Dans des villes comme Djakarta, Dhaka, Manille et d'autres encore, l'on estime qu'un pourcentage important d'habitants sont sujets à des maladies chroniques. L'exposition à des niveaux de pollution élevés, la difficulté d'accéder à une eau de bonne qualité, à un système d'évacuation domestique des eaux usées, au système de santé, toutes causes généralement liées à un faible niveau de vie sont autant de défis de premier plan pour les maires et autres administrateurs de ces gigantesques cités.

DEUXIÈME PARTIE

Entrer dans l'univers des data

Culture ou technologie ?

L'ère de la plateforme

Nombreux sont les chefs d'entreprises, les acteurs institutionnels, politiques, les simples citoyens qui ne réalisent pas ce que représente le potentiel des données. Et parmi les chefs de grandes entreprises, beaucoup nient même qu'il existe un réel potentiel. Lorsque émergea l'univers du Web dans les années 1990, on entendait à peu près la même rengaine de leur part : « Ça n'est pas pour moi, ça n'est pas une nécessité immédiate. » Les entreprises ne percevaient pas la valeur économique de l'Internet qui, au mieux, leur apparaissait comme un nouveau format de diffusion de leurs brochures publicitaires.

Quand, au cours de l'été 2014, CaptainDash a commencé l'analyse de l'agilité numérique du Cac 40 avec le magazine *Enjeux Les Échos*, il est apparu très clairement que le Big Data effrayait les directeurs des systèmes d'information, tant il s'inscrivait en rupture par rapport aux processus traditionnels. Car, contrairement à une idée reçue, le Big Data ne représente pas qu'un enjeu technologique : introduire le Big Data au sein d'une entreprise revient à privilégier les processus transversaux au détriment de l'organisation en silos, si propice

Data, Big Data et... marketing

C'EST peu de dire que le marketing représente l'un des domaines de prédilection du Big Data, mais aussi l'un de ceux qui suscitent le plus d'inquiétudes quant aux abus possibles. La simple innovation qui a consisté à permettre aux publicités d'un site d'e-commerce que l'on aurait visité de se répandre sur d'autres sites que l'on visiterait ultérieurement – le *retargeting* – a fait l'objet de vives critiques et continue à faire polémique. Il ne s'agit pourtant encore que d'une mise en œuvre très limitée des données, et plutôt d'une astucieuse gestion des *adserver*, ces plateformes qui distribuent les bannières et autres formats publicitaires sur Internet, mais elle n'en a pas moins fait polémique.

Agences publicitaires et agences créatives

Un livre ne suffirait pas à raconter dans le détail toute l'histoire du développement de la publicité, donc du marketing. Cette histoire est d'ailleurs intrinsèquement liée à l'expansion de la société de consommation et de l'ère industrielle. Le marketing symbolise on ne peut mieux le monde occidental : alors que celui-ci est mû

L'impact économique des données

Servification

Et finalement, pourquoi les data et le Big Data ne seraient-ils rien de plus qu'une nouvelle technique ? Pourquoi ne seraient-ils pas, au-delà des gains de productivité évidents qu'ils procurent, des forces de nature à changer les règles du jeu ?

En réalité, les règles du jeu ne seront désormais plus les mêmes. Et une entreprise qui ne les aurait pas comprises n'aurait à terme guère de chance de subsister. Le Big Data n'est en effet pas une technique que l'on apprivoise pour l'intégrer dans le processus de production. Il s'agit plutôt d'un changement de paradigme majeur. Un fabricant de matelas peut utiliser le Big Data pour essayer d'optimiser son processus de production. Prévoir un accroissement de la demande, détecter le moment où il importe d'effectuer des opérations de maintenance sur son outil de production... en cela oui, le Big Data peut être assimilé à un outil, un moyen d'accroître la productivité.

Mais que penser de sa capacité à remettre en cause la vocation même de ce fabricant de matelas ? Quelle magnitude sismique affecter à une technique qui ferait

TROISIÈME PARTIE

Nous et les machines

Surveillance, machines et transparence

Le Big Data interprété par la NSA

À voir la mine fermée qu'il affichait lors de ses dernières apparitions publiques, nul doute que pour le général Keith Alexander, les auditions à répétition devant le Congrès commençaient à devenir pesantes. En dépit de sa parfaite maîtrise du sujet, malgré ses médailles, malgré le soutien d'une large majorité de l'administration américaine, il ne parvenait plus à dissimuler une certaine nervosité, à quelques semaines de la retraite, d'avoir à répondre à ce qu'il percevait de plus en plus comme des procès en inquisition. Cela faisait neuf mois que rien n'allait plus ; depuis juin 2013 exactement. Un obscur consultant du nom d'Edward Snowden avait alors commencé à faire fuir de grandes quantités d'informations *top secret* via des journaux comme *The Washington Post* ou *The Guardian*. Peu à peu l'ampleur du désastre apparaissait : ce consultant en savait suffisamment pour détailler le fonctionnement de nombreux programmes ultrasecrets de la NSA (National Security Agency), l'agence dont le général Alexander avait la charge. Le pire était que nul ne savait où il allait s'arrêter.

Le code des données

NE nous méprenons pas : ces enjeux numériques vont s'imprégner dans notre quotidien dans une mesure qui se situe au-delà même de ce que nous pensons généralement. Certes, les pouvoirs conférés par la technologie à la société civile vont s'accroître très significativement et avec eux, également, des risques de toutes sortes. Pour autant, les opportunités individuelles et collectives issues du digital et plus encore du Big Data sont immenses. Il ne s'agit donc pas d'être naïf : il est probablement raisonnable d'admettre que services secrets et écoutes sont un mal nécessaire. Mais ce n'est pas en mettant en place des dispositifs qui sont proches des lois d'exception¹ que nous allons recréer un sentiment de confiance de la part des citoyens. En revanche, faire en sorte que le monde politique et l'appareil administratif acquièrent les codes numériques, comprennent

1. La loi dite « article 20 » du code de programmation militaire (cadre de l'article L241-1) légalise les procédures d'interception. Lors du débat qui a précédé son adoption, les parlementaires ont eu beau jeu d'expliquer qu'ils avaient limité sa portée aux métadonnées. Probablement savaient-ils déjà que le dispositif technologique de la DGSE était justement conçu pour n'intercepter en général que ces données, et non le contenu principal des messages. Or, contrairement à une idée reçue, les métadonnées en disent beaucoup plus long sur nous que nous ne le pensons.

OUVERTURE

La nécessité de choisir

EN avril 2014, Stephen Hawking, probablement l'un des scientifiques vivants les plus brillants, prit la parole pour dénoncer les risques cachés, intrinsèques de l'intelligence artificielle¹. Beaucoup n'y virent que les élucubrations d'un homme maintenant âgé et évidemment diminué par une terrible infirmité. Pourtant, Hawking n'est pas un réactionnaire. Il croit profondément au progrès et a souvent émis des hypothèses qui démontrent sa capacité d'envisager un monde renouvelé, sensiblement différent de celui que nous connaissons, notamment du fait de possibles avancées scientifiques ou techniques. Hawking « voit » littéralement des univers que le commun des mortels ne peut pas conceptualiser. C'est d'ailleurs en cela que son point de vue est intéressant. Mais sa vision, tout aussi perçante qu'elle soit, ne représente néanmoins que l'un des futurs possibles.

L'humanité dans son ensemble est confrontée aujourd'hui à des choix probablement les plus importants de son histoire, et ce de façon nette et brutale. Nous allons devoir gérer la transition énergétique de la planète, et répondre aux défis environnementaux que poseront ses 9 ou 10 milliards d'habitants. Ces défis sont connus et

1. « Stephen Hawking is terrified of artificial intelligence », *Huffington Post*, 5 mai 2014.

Lexique

Api : acronyme d'*Applications Programming Interface*. Prise de courant numérique, l'Api est une interface de programmation qui permet à de multiples services numériques de se « brancher » sur une application pour échanger des données.

Une Api est généralement ouverte et proposée par le propriétaire du programme. Elle peut être normalisée par des organismes comme l'Iso ou l'IUT.

Cluster : grappe de serveurs ou « ferme de données », structure générique des applications distribuées en Big Data.

Cnil : acronyme de Commission nationale de l'informatique et des libertés. La Cnil représente une autorité administrative indépendante chargée de veiller à ce que l'informatique soit au service du citoyen et qu'elle ne porte atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques (source Wikipédia).

CRM : acronyme de *Customer Relationship Management*. Il s'agit des progiciels qui permettent de traiter directement avec le client, que ce soit au niveau de la vente, du marketing ou des services annexes, et que l'on regroupe souvent sous le terme de *front-office*, par opposition aux outils de *back-office* que sont les progiciels de gestion intégrés ou ERP (source Wikipédia).

DPI : acronyme de *Deep Packet Inspection*. Technologie permettant de reconstituer des messages à partir de l'interception des *packets* IP qui transitent dans un câble de télécommunication sous-marin par exemple.

En informatique, la *Deep Packet Inspection* (en français Inspection des paquets en profondeur) est, pour un équipement d'infrastructure de réseau, l'analyse du contenu (au-delà de l'en-tête) d'un paquet réseau (paquet IP le plus souvent) de façon à en tirer des statistiques, à filtrer ceux-ci ou à détecter des intrusions, du spam ou tout autre contenu prédéfini. Le DPI* peut servir notamment à la censure sur Internet ou dans le cadre de dispositifs de protection de la propriété intellectuelle (source Wikipédia).

ETL : acronyme d'*Extract-Transform and Load*. Représente une passerelle faite sur mesure entre deux systèmes d'information.

Elle est fondée sur des connecteurs (*extract*) servant à exporter ou à importer les données dans les applications (par exemple des connecteurs Oracle ou SAP...), des transformateurs (*transform*) qui reformatent les données (agrégations, filtres, conversions...), et des mises en correspondance (*load*). Les solutions d'ETL sont apparues dès les années 1970 pour faciliter la conversion régulière de données entre applications dans le monde bancaire et financier. Les ETL sont souvent le cauchemar des directeurs informatiques pour leur complexité. Elles s'opposent aux Apis qui visent à normaliser les échanges là où les ETL font des traitements spécifiques.

Gafa : l'acronyme de Google, Apple, Facebook et Amazon, quatre grandes firmes américaines emblématiques de ce qu'est l'économie numérique avec son développement très rapide, et dominant chacune leurs marchés.

Hadoop : à la base, Hadoop représente un *framework* conçu en *open source* et permettant de réaliser des traite-

ments sur des volumes de données massifs, de l'ordre de plusieurs pétaoctets (soit plusieurs milliers de téraoctets). Aujourd'hui, il s'agit davantage d'une définition générique d'outils de Big Data *open source*, compatibles entre eux.

Hana : acronyme de *High Performance Analytic Appliance*. C'est une technologie *in memory* de traitement en mémoire de masse, propriétaire, développée par SAP AG. Hana fonctionne en mode massivement parallèle, exploitant ainsi au maximum de processeurs multicœurs et permettant l'exécution particulièrement rapide des requêtes.

In memory : les systèmes de Big Data « en mémoire » travaillent sur des données stockées dans de la mémoire vive (ou flash) pour accélérer le traitement de requêtes nécessitant de faire de nombreux appels de données. Actuellement beaucoup plus coûteux à exploiter que les systèmes en disques traditionnels, ils n'en représentent pas moins une avancée significative par la vitesse de traitements des algorithmes complexes qu'ils permettent.

Learning machine : système qui pratique l'analyse de situations à partir de données et qui est capable de commencer une action en fonction des typologies de données. Par exemple, prévenir un usager d'un dysfonctionnement lorsque l'infrastructure de son opérateur est tombée en panne.

MapReduce : architecture de développement informatique, inventée par Google, dans laquelle sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses, typiquement supérieures en taille à 1 téraoctet.

Les termes *map* et *reduce* ainsi que les concepts sous-jacents sont empruntés aux modèles de programmation. MapReduce permet de manipuler de grandes quantités de données en organisant leur distribution dans un *cluster** de

machines afin d'y être traitées. Ce modèle connaît un grand succès auprès de sociétés possédant d'importantes quantités de données à traiter, comme Amazon ou Facebook.

NoSQL : en informatique, NoSQL (*Not only SQL* en anglais) désigne des systèmes de gestion de base de données (SGBD) qui ne sont pas fondés sur l'architecture classique des bases relationnelles. L'unité logique n'y est plus la table, et les données ne sont en général pas manipulées avec le système générique SQL. Les systèmes NoSQL sont généralement rudimentaires en termes de fonctionnalités, mais permettent une grande agilité dans le traitement des données. Ils se sont progressivement imposés avec l'explosion de données qu'ont constatée les grands acteurs de l'Internet qui ne parvenaient plus à faire fonctionner leurs services avec des systèmes relationnels traditionnels.

Open data : une « donnée ouverte » est une donnée numérique d'origine publique ou privée, accessible à tous. Elle peut être notamment produite par une collectivité, un service public ou une entreprise. Elle est diffusée de manière structurée selon une méthodologie et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière.

SaaS : acronyme de *Software as a Service*. Service logiciel auquel on accède en ligne et que l'on paye en fonction de l'utilisation (par mois, par volume, etc.).

Spark : dispositif *open source* de type Hadoop* qui s'affranchit des architectures de type MapReduce* en faisant ses traitements directement dans de la mémoire vive (*in memory*). Accélération ainsi ses traitements de façon importante, allant jusqu'à 100 fois plus vite que les systèmes Hadoop traditionnels, Spark est un programme prioritaire d'Apache Foundation.