Académie des technologies

Big Data :
un changement de paradigme peut en cacher un autre

Opportunités et menaces liées à l'émergence de nouveaux écosystèmes

Rapport de l'Académie des technologies
de la Commission TIC
Rapporteur : Yves CASEAU

Imprimé en France
ISBN : 978-2-7598-1780-1

RÉSUMÉ

Le domaine du *Big Data*[1] représente une vraie révolution informatique, qui s'exprime dans des dimensions multiples, depuis la technologie jusqu'aux applications en passant par les pratiques. Il devient possible d'analyser un ensemble très important de « traces numériques » et de connaître les intentions des clients des entreprises avec une précision inégalée. La manipulation des données issues des smartphones et objets connectés ouvre des opportunités de nouveaux services, tandis que ces nouvelles méthodes permettent une réduction importante des coûts des systèmes d'information.

La maitrise du *Big Data* est un enjeu majeur de compétitivité pour les entreprises. Ce rapport va s'intéresser à l'impact du *Big Data* sur les entreprises, même si de nombreux autres domaines sont touchés, par exemple les sciences ou le monde politique et citoyen. Maîtriser ces méthodes permet aux entreprises une nouvelle proximité dans la relation avec ses clients. Il devient fondamental de

[1] Nous employons le terme anglais *Big Data* parce qu'il est le plus populaire et celui utilisé au sein des entreprises. Le terme français dont l'usage est recommandé est « megadonnées », qui a été publié au journal officiel le 22 août 2014.

comprendre ces nouveaux outils pour faire face à la compétition mondiale des entreprises « stars » de l'Internet.

Le *Big Data* est une rupture dans l'analyse des données et l'utilisation des méthodes statistiques pour les entreprises, fondée sur une approche systémique et des cycles réactifs courts. La démarche classique qui sépare une phase d'extraction de connaissances de la phase d'application – pour du ciblage marketing par exemple – est remplacée par une boucle itérative dans laquelle les motifs détectés sont tout de suite confrontés à la mise en situation et jugés sur leur efficacité opérationnelle.

Le *Big Data* représente également une nouvelle façon de programmer, de façon massivement parallèle et centrée sur les données. Le *Big Data* n'est pas simplement une collection d'outils, c'est également une autre façon de concevoir les algorithmes. Cette différence vient de la distribution des traitements sur des milliers voire des dizaines de milliers de machines, des exigences de performance liées aux très gros volumes traités et du besoin de mettre les algorithmes au point par apprentissage.

l'Académie des technologies estime que le *Big Data* est un enjeu majeur pour les pouvoirs publics et les entreprises françaises. Ces changements de paradigmes méritent une prise de conscience et un fort accompagnement en termes de formation. La technologie et la pratique y jouent des rôles essentiels ; il s'agit de développer de nouvelles façons de travailler, dont le champ d'action est extrêmement large et n'est pas restreint à l'exploration de nouvelles opportunités.

SOMMAIRE

INTRODUCTION

Le terme de *Big Data* est apparu il y a une dizaine d'années pour désigner le traitement informatique d'un volume de données dépassant les capacités de traitement utilisées jusqu'alors en utilisant la solution classique d'un système de gestion de base de données relationnelle (SGBDR). Wikipédia cite un rapport du META Group / Gartner de 2001 qui fait la première mention des « 3V » : volume, vélocité et variété. Ces « trois V » décrivent les trois dimensions qui posent des défis aux bases de données relationnelles « classiques ». Par construction, le volume qui justifie l'appellation de *Big Data* prend en compte les aspects technologiques et varie donc au cours du temps ; en 2012, ce seuil correspondait à plusieurs dizaines de téraoctets.

Puisque le *Big Data* a été présenté dès son introduction comme un domaine technologique en rupture, il a fait l'objet de beaucoup d'attentions, en particulier des fournisseurs de technologie. Les possibilités émergeant de ces nouveaux outils ont souvent été exagérées, créant comme pour chaque innovation significative un débat sur la radicalité de ce changement. La commission TIC de l'Académie des technologies s'est penchée sur ce thème en 2012 et a conduit un cycle d'interviews

et de réflexion[2] pour approfondir la question qui aurait pu s'intituler « le *Big Data* est-il une révolution ou une évolution ? ». La décision de produire un rapport qui résume ces investigations peut surprendre, tant le sujet est devenu médiatique et très bien couvert. En particulier, le rapport de prospective 2030 « Un principe et sept ambitions pour l'innovation » [1], présenté par la commission « Innovation 2030 » présidée par Anne Lauvergeon en 2013, fait la part belle au *Big Data* comme un des sujets d'avenir. Le présent rapport s'inscrit dans la continuité de ces recommandations, mais il s'intéresse à d'autres aspects plus disruptifs qui ne sont pas forcément mis en valeur dans les documents récemment publiés sur le *Big Data*. Là où la plupart des documents cités dans la bibliographie, et en particulier le rapport précité ainsi que celui qui vient d'être remis au Président Obama [2], s'intéressent à la capacité d'innovation (nouveaux services et nouvelles opportunités pour des nouveaux métiers), l'Académie des technologies s'est plutôt concentrée sur la capacité de disruption (faire différemment et plus efficacement les services actuels des métiers d'aujourd'hui). Nous allons nous placer dans la continuité des propos de François Bourdoncle [3] – un des experts audités pendant nos séances de travail – et de Nicolas Colin (co-auteur de « L'âge de la multitude » [4] et d'une célèbre chronique sur Internet intitulée « Les Barbares Attaquent »).

Parmi les ouvrages de référence, on peut citer l'excellent livre de Viktor Mayer-Schönberger et Kenneth Cukier *Big Data – A Revolution That Will Transform How We Live, Work and Think* [5] paru en 2013. Nous allons emprunter de nombreux exemples à ce livre qui choisit clairement l'option de la révolution par rapport à l'évolution. Au cours de notre développement nous allons identifier cinq « révolutions » possibles :

- la révolution de la connaissance client, beaucoup plus précise et utilisable en temps réel, qui permet de mieux vendre au client des produits et services existants, en s'appuyant sur les souhaits perçus et inférés, ce que Doc Searl appelle l'économie de l'intention [6] ;
- la révolution de la vie numérique et de l'internet des objets, qui crée un flux de données formant les traces numériques de nos existences et qui permet d'inventer des nouveaux services ;

[2] La liste des intervenants est citée à la fin de ce document dans la section « remerciements »

- la révolution technologique associée, celle de la baisse spectaculaire des coûts unitaires, qui permet de reconstruire avec de nouveaux outils des systèmes d'information beaucoup moins chers, créant un potentiel de disruption pour ceux qui maîtrisent ces outils (les « barbares » mentionnés par Nicolas Colin ou François Bourdoncle) ;
- la révolution dans la démarche intellectuelle et, en particulier, les méthodes de programmation, fondée sur le parallélisme massif et le rôle central de la donnée, ce que Henri Verdier nous a décrit comme « Data is the new code » ;
- la révolution dans l'analyse des données, que nous pourrions qualifier de « statistiques sans modèles », dans laquelle l'analyse de la causalité est remplacée par l'utilisation en boucle asservie de corrélations détectées par les méthodes de *Big Data*. Cette approche est illustrée par la célèbre citation de Chris Anderson, l'éditeur en chef de la revue Wired, en 2008 : « le déluge de données rend la méthode scientifique obsolète ». Ce sera l'objet de la section 3.3.

L'objectif principal de ce rapport est de sensibiliser le lecteur sur les trois derniers sujets de cette liste. Nous aborderons brièvement les deux premiers. Le thème de l'amélioration de la connaissance client est central dans le *Big Data* et celui de l'internet des objets et des nouvelles opportunités de service est très bien mis en valeur par le rapport de la commission Anne Lauvergeon. Nous souhaitons attirer l'attention des pouvoirs publics et des dirigeants des grandes entreprises sur le potentiel important de disruption des méthodes et outils du *Big Data*. Notre objectif est de contribuer à une prise de conscience qui débouche sur des efforts massifs de formation et de développement de compétences pratiques (*cf.* la note de la commission TIC sur l'enseignement de l'informatique [7]), dans l'enseignement supérieur, tout comme dans les entreprises.

Ce document est organisé comme suit. La section 2 présente le *Big Data* comme une suite logique de l'évolution de l'informatique, dans un contexte d'explosion du volume de données disponible, et marquée par un ensemble de ruptures technologiques. La section 3 prend le contrepied de cette vision évolutionniste et s'intéresse aux ruptures de paradigmes, en particulier pour approfondir les « révolutions » que nous avons listées dans cette introduction. La dernière section propose une analyse des risques et des recommandations

pour les lecteurs de ce rapport, tournées vers le renforcement de la compétitivité de la France. Nous insistons sur les menaces pour les grandes entreprises et sur la nécessité de favoriser l'émergence d'un écosystème de pratique, de formation et d'innovation.

LA GENÈSE DU « *BIG DATA* »

Le terme de *Big Data* s'inscrit dans le contexte de l'omniprésence de l'informatique. D'un certain côté, il s'agit de la suite logique de l'évolution de l'informatique, qui traite depuis sa naissance de la manipulation de données. L'évolution, de type exponentiel, des capacités de stockage et de traitement que nous constatons depuis 50 ans conduit maintenant à traiter des très grands volumes de données. Pourtant, il existe également des ruptures technologiques qui accompagnent cette évolution.

2.1 Le contexte : l'explosion des données

Tous les experts que nous avons audités convergent sur ce point : la genèse du *Big Data* repose sur la baisse spectaculaire des coûts unitaires de stockage et de l'explosion de la densité d'intégration, qui ont rendu possible le fait de conserver des données d'usage qui seraient restées temporaires il y a quelques années. Matthew Komorowski (*cf.* figure 1) a calculé que la quantité de stockage sur disque accessible pour un prix donné avait doublé tous les 14 mois depuis 30 ans. Ceci

fait que le coût du Téraoctet est passé de 1 million de dollars en 1995 à moins de 40 $ en 2013. Pour fixer un ordre de grandeur, l'entrepôt de données de Bouygues Télécom en 2000 représentait une dizaine de To. Cela signifie que l'on peut mettre aujourd'hui sur un produit grand public de quelques centaines d'euros toutes les connaissances clients du système d'information d'il y a dix ans pour un million de clients.

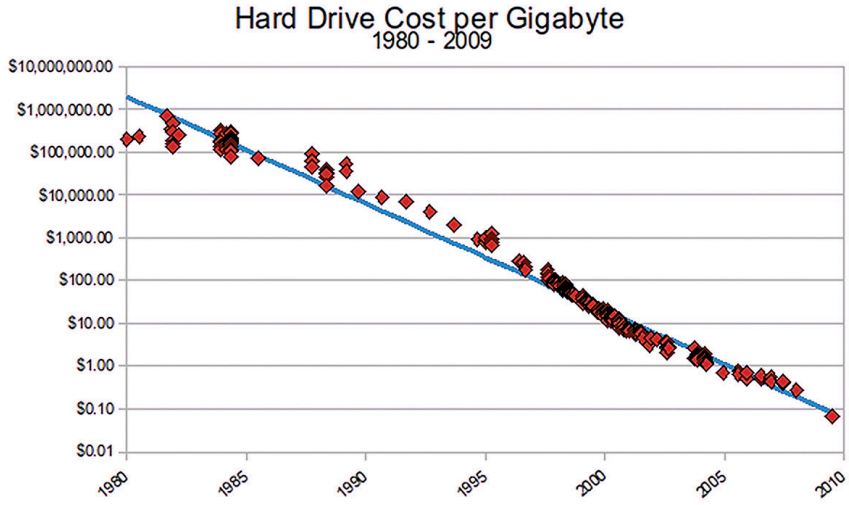

Figure 1 : évolution des coûts de stockage sur disque (Matthew Komorowski)

Cette capacité à stocker un volume sans cesse croissant de données a rencontré une explosion des sources. La numérisation de l'ensemble des activités humaines [8] conduit à la prolifération des « traces numériques ». Toutes nos actions, que ce soient des transactions (achats, réservations, requêtes) ou une simple lecture de contenus numériques ou encore notre navigation sur Internet, produisent des « logs », des représentations numériques dans les systèmes d'information, qu'il est aujourd'hui réaliste de conserver à cause de la baisse des coûts. Cette explosion des traces est amplifiée par la prolifération des smartphones (6 milliards dans le monde en 2014), qui sont autant de balises numériques de nos vies quotidiennes. Une des manifestations les plus spectaculaire de cette « vie numérique » est le développement massif des réseaux sociaux (Facebook avec plus de 1,3 milliard d'utilisateurs, WhatsApp avec 500 millions ou encore Twitter avec 300 millions). Ces réseaux permettent aux internautes de fournir de façon

volontaire (?) des traces numériques de ce qu'ils font, ce qu'ils pensent et avec qui ils interagissent. Cette tendance n'est pas prête de changer ; au contraire, elle devrait s'amplifier avec l'explosion des « senseurs », ces capteurs connectés qui sont omniprésents, que ce soit sur nos smartphones ou sur les fameux « objets connectés » de l'Internet des objets (IOT : *Internet of Things* ou IOE : *Internet of Everything*). Tim O'Reilly emploie l'expression de *Web Squared* (web au carré) pour représenter cette double connexion du monde réel et du monde virtuel : les objets physiques deviennent « intelligents » en étant connectés au Web, tandis que les programmes du « cloud » ont accès aux « sens du monde physique » à travers la myriade de capteurs (image, température, son, pression, etc.) de ces objets connectés.

L'explosion des données s'accompagne de l'apparition de données dites « non-structurées », qu'il s'agisse de textes, de photos, de sons ou de vidéos. L'explosion du trafic sur Internet est fortement liée au développement de la vidéo, sous toutes ses formes (de la distraction à la communication, de YouTube à Facebook). Ces nouvelles formes de données non structurées posent des défis en termes de collecte, de stockage, d'indexation, de recherche et de manipulation pour les systèmes d'information classiques.

Même si la majorité des traitements *Big Data* consiste à effectuer des opérations simples sur des très grands volumes de données, il faut pouvoir le faire très rapidement. Critéo, une entreprise française qui est le leader mondial de son domaine et un des centres d'excellence du *Big Data*, est un parfait exemple de cette maîtrise de la vélocité puisque son métier consiste à faire du *retargeting*, c'est-à-dire choisir en temps réel les meilleures publicités en fonction de l'historique de l'intérêt d'un internaute perçu au travers de sa navigation.

2.2. Les technologies pivots

Les technologies qui soutiennent l'essor du *Big Data* se sont développées depuis une grosse dizaine d'années pour faire face aux trois enjeux représentés par les 3V :

- volume : comment stocker des volumes qui s'expriment aujourd'hui en pétaoctets (10^{15}) et demain en exaoctets (10^{18}) ?

- variété : comment stocker des données non-structurées, c'est-à-dire de façon brute (et souvent compressée, comme pour la vidéo) sans perdre les capacités d'indexation et de recherche ?
- vélocité : une grande partie des « traces numériques » représentent des informations volatiles, qui n'ont d'intérêt que si elles peuvent être utilisées rapidement.

La première réponse à ces défis est d'utiliser un stockage massivement distribué (plusieurs milliers, voire dizaines de milliers de serveurs) des données, en évitant toute forme de déplacement pendant le traitement. Le traitement se fait également de façon distribuée, au plus proche de la donnée. La technologie la plus emblématique de répartition de données distribuées est *Hadoop*, dont nous allons reparler et qui est presque synonyme de *Big Data* en ce qui concerne le déploiement de ces dernières années. *Hadoop* a popularisé le concept de *MapReduce*, consistant à découper un très gros volume de données en petites unités et à effectuer une très grande partie du traitement de façon distribuée sur chaque « petit » serveur (la partie *map*), tandis que le travail d'agrégation (*reduce*) est limité le plus possible. Tous les traitements ne prêtent pas à cette décomposition, mais on peut rendre à Google l'honneur d'avoir montré très tôt (sur *Google Search*) et sur de nombreux exemples (en particulier *Gmail*) la puissance de cette approche massivement distribuée.

La distribution des traitements s'appuie sur la disponibilité de moyens de calculs distribués (*grids*, *cloud*) identiques à des coûts qui baissent également de façon exponentielle. Pour illustrer ce point, citons l'anecdote rapportée par François Bancilhon, de Data Publica, pour une expérience d'analyse de données qui s'appuyait sur le graphe complet du Web (un graphe représentant 70 To), il a été possible de faire un calcul distribué mobilisant jusqu'à 2 000 cœurs de calcul pour un coût total de 300 € chez Amazon. Notons également que l'évolution exponentielle de la puissance concerne dans des proportions variées le calcul, le stockage sur disque, mais également la capacité de mémoire vive et la vitesse de transfert des réseaux, ce qui est nécessaire pour le fonctionnement de ces architectures distribuées.

Une autre rupture technologique est liée au traitement des données non-structurée : l'apparition des bases de données *no-SQL*, dont *Appache Cassandra* ou *MongoDB* sont des exemples emblématiques. SQL est le langage

de requête structuré commun à l'ensemble des bases de données relationnelles. L'approche *no-SQL* consiste à ne plus placer des données sous une forme structurée de tables, mais d'accepter un stockage plus « primitif », convenant mieux aux contenus tels que la vidéo ou aux « traces numériques » qui sont encapsulées dans les langages informatiques de structuration et de transfert de données tels que XML ou JSON.

Pour finir ce tour d'horizon synthétique, le besoin de traitement en temps réel a produit de nombreux outils informatiques autour de concepts tels que le *flow computing* (traitement sur des flots de données sans stockage) ou de *complex event processing* (CEP), pour fournir des solutions de traitement de très gros volumes de données en quasi temps-réel. Dans tous les cas, il s'agit d'approches distribuées qui s'adaptent aux architectures massivement parallèles du *Big Data*. Par exemple, la plate-forme *Storm* est une évolution de *Hadoop* qui est mieux adaptée au traitement d'évènements et qui a été rendue populaire par son adoption par Twitter.

2.3 Ingénierie système et écosystème logiciel

Marko Erman, de Thales, a insisté sur la vision complète « système » qui est nécessaire pour réussir un traitement de *Big Data*, ce qui est illustré de façon sommaire par la figure 2. Il faut maîtriser et intégrer des compétences « hardware » (pour construire et opérer des réseaux de ressources distribuées), des compétences « systèmes informatiques » (en particulier, une bonne compréhension du système d'exploitation et de ses mécanismes de distribution), des compétences logicielles applicatives (pour maîtriser les solutions techniques telles que *Hadoop*, *Storm* ou *Elastic Search*, pour ne citer que quelques exemples) et des compétences algorithmiques (nous allons y revenir, les algorithmes du *Big Data* sont forcément spécifiques car ils doivent avoir des complexités sous-linéaires).

Il est naturel de faire remonter l'origine du *Big Data* à Google, qui a développé toute une approche du calcul parallèle pour mettre en œuvre son moteur de recherche. Google a développé ses propres serveurs (en mettant en œuvre le principe du *commodity computing*, consistant à construire des nœuds de calcul à très faible coût en partant des processeurs des PC grand-public et en jouant sur

la redondance), sa propre version de système d'exploitation (en partant de LINUX et en écrivant ses propres algorithmes de distribution et d'ordonnancement), son propre mécanisme de gestion de fichiers et de stockage (donnant naissance à *BigTable*, qui est l'ancêtre de *Hadoop*) et sa première réalisation d'une plate-forme *MapReduce*. *Hadoop* est née comme une initiative *open source*, en partie supportée par Yahoo (à l'origine de *Hadoop* en 2005), pour reproduire cet environnement de calcul distribué construit par Google. *Hadoop*, parce qu'il n'impose pas de structure de données (c'est une forme de système de fichiers, utilisé avec d'autres outils, tels que *Pig* ou *Hive*, pour construire des traitements) supporte de très larges volumes de données (petaoctets) hétérogènes.

Le logiciel libre et *open source* joue un rôle fondamental dans l'approche *Big Data*. *Hadoop* n'est qu'un des maillons d'une chaîne logicielle qui intègre de multiples outils, allant de la configuration à la visualisation en passant par la supervision. Nos nombreux experts ont insisté sur l'importance de l'*open source*, qui dépasse le cadre de ce rapport sur le *Big Data*. Dans un mouvement de balancier historique, l'utilisation de code en *open source* permet de réutiliser des fragments de code par insertion, dans une « approche boîte blanche », qui est très différente de l'approche par composants « boîte noire » qui était l'état de l'art il y a 20 ans. Le code *open source* est développé de façon émergente, avec une distribution du travail de test et de validation, qui est une réponse systémique efficace à la complexité et la variété des sujets à traiter.

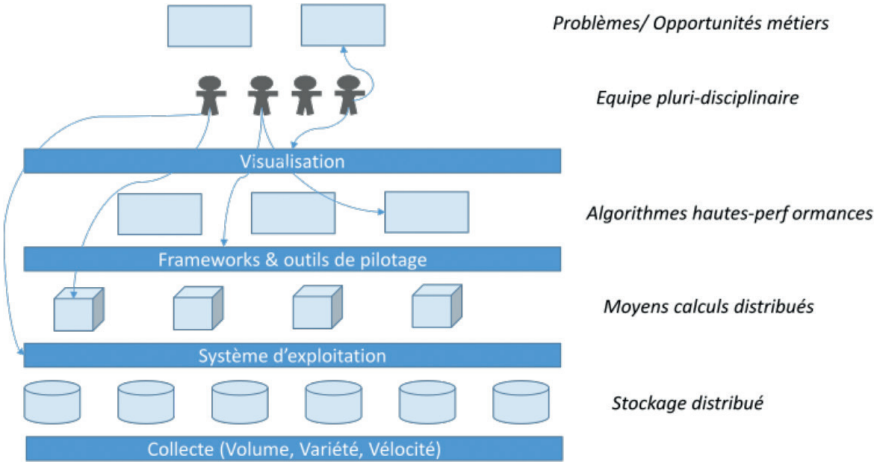

Figure 2 : Le *Big Data* en tant que système

Nous allons voir dans la section suivante qu'il y a plus que des ruptures technologiques dans le *Big Data*, mais il ne faut pas sous-estimer la technologie, en particulier la maîtrise des outils. De par la prédominance des outils *open source*, c'est une véritable culture technique qu'il faut s'approprier pour pouvoir exploiter ce « nouveau pétrole que représente les données », pour reprendre une métaphore abondamment employée. Pour conclure, notons que si les textes écrits restent une bonne source pour prendre du recul sur le sujet *Big Data*, les meilleures sources pour comprendre la technologie sont les vidéos (par exemple sur *YouTube*, *cf.* [22]), des exposés techniques proposés par les architectes qui ont construit des grands systèmes opérationnels (tels que Netflix, Spotify, etc.) et des *MOOCs* (cours en lignes – *cf.* glossaire).

TROIS RUPTURES DE PARADIGME

Même si la massification des données traitées dans les systèmes informatiques s'inscrit dans la continuité de la discipline, il est apparu au travers des entretiens de la commission TIC que l'on pouvait distinguer « trois ruptures de paradigme », c'est-à-dire des changements dans la façon de concevoir et d'opérer des systèmes informatiques lorsque ceux-ci traitent des très grands volumes de données et utilisent les technologies que nous venons de décrire.

3.1. Nouveaux services et nouveaux métiers autour des données

La richesse des informations disponibles, puisque nous sommes capables de stocker les traces numériques, permet un niveau de *profiling* exceptionnel, qui représente une rupture par rapport à ce qui était possible jusqu'alors [2]. Dans le livre « *Big Data* » [5], les auteurs donnent de nombreux exemples spectaculaires, tirés par exemples de Target ou de Walmart aux États-Unis (deux acteurs de la grande distribution). Dans un exemple devenu célèbre, Target a été capable de

détecter qu'une jeune fille était tombée enceinte, à partir de l'analyse de son ticket de caisse, alors que ses parents n'étaient pas encore au courant. C'est d'ailleurs cette capacité de « prévision » qui pose des questions de respect de la vie privée, sur lesquelles nous reviendrons dans la section 4.1. La thèse fondamentale de *l'économie de l'intention* est qu'il est maintenant possible de proposer au consommateur le produit ou le service dont il a envie, au bon moment, en capturant son intention à partir des traces numériques (ses navigations, ce qu'il écrit sur les réseaux sociaux, les personnes avec qui il interagit, etc.). Cette nouvelle forme de marketing est très bien décrite dans le livre *Le Marketing Synchronisé* de Marco Tinelli [11]. Il faut noter que, contrairement à d'autres services personnalisés que nous allons évoquer plus loin, le *profiling* et la recommandation personnalisée peuvent s'appuyer sur des agrégats liés aux produits (ce que font Amazon ou Netflix pour leurs algorithmes de recommandation, ou ce que fait Target sur l'analyse des tickets de caisse [5]), ce qui ne pose pas de la même façon la question du respect de la vie privée. Il y a bien deux questions distinctes : celle de la divulgation d'informations collectées (par exemple la géolocalisation) et celle de la nature des inférences qu'on a le droit de faire.

Cette richesse de « *profiling* » est la première source de valeur qui fait penser au « pétrole », puisque chaque entreprise qui dispose de riches données d'usage peut en extraire des informations qui ont une valeur plus large. L'exemple le plus évident est celui des opérateurs téléphoniques, qui disposent de vecteurs de déplacements (dans temps et dans l'espace), croisés avec des données sociodémographiques. Même s'il faut respecter des processus de regroupement et d'anonymisation (*cf.* Section 4.1), cela permet, comme l'a fait Telefonica en créant Telefonica Systems [5], de créer de nouvelles sources de diversification des revenus. Le *Big Data* crée des opportunités de type « B2B2C » [3]. Pour prendre un exemple concret et classique, une entreprise de distribution peut choisir la localisation de ses magasins en utilisant des nuages de points « déplacement » fournis par un opérateur téléphonique [12]. De façon duale, le *Big Data* donne des

[3] dans laquelle une entreprise – le premier B – expose des services de données au moyen d'API – Application Programming Interfaces, le cœur de ce que l'on appelle une approche de plate-forme [10] – qui permettent à une autre entreprise – le second B – de mieux servir ses propres clients (C) parce qu'elle les connaît mieux grâce aux services de la première entreprise.

très bons résultats pour détecter des comportements « anormaux » et est utilisé avec succès par les entreprises contre la fraude [5].

À côté de l'exploitation de données « anonymisées » et collectives, se trouve le second « gisement » de l'exploitation des données individuelles. Il est possible, pour simplifier, de distinguer deux modèles : celui de l'exploitation de données « peu sensibles » avec un échange « gagnant-gagnant » (le client accepte que ses données d'usage soient exploitées en échange d'un service gratuit) et celui de l'exploitation de données sensibles, pour lesquelles le client exige la non-divulgation et a, par conséquent, besoin d'une relation de confiance avec l'entreprise. Le premier modèle est celui des Google, Facebook et autres réseaux sociaux ou celui des acteurs B2B, tels que Critéo, qui exploitent les « traces de navigations » (que l'on nomme « cookies », et qui sont collectées par les navigateurs). Les enquêtes d'opinion montrent que les internautes sont très lucides sur ce type d'échange de valeur : ils comprennent que la mise à disposition gratuite d'un service tel que Gmail ou Facebook est obtenu en contrepartie d'un ciblage publicitaire renforcé. Une étude commissionnée par Orange a montré que l'internaute évalue à environ 60 € par an la valeur qu'il souhaite recevoir (de son opérateur téléphonique par exemple) en échange de l'utilisation de ses données d'usage à des fins commerciale. L'exemple d'O_2 au Royaume-Uni est significatif : il propose un programme qui permet de gagner des récompenses, telles que des places de concerts, en échange de l'utilisation commerciale des données de navigation mobile. Cette offre, en *OPT-IN* (acceptation explicite des clients), a trouvé un succès considérable (plusieurs millions de clients, avec un taux de satisfaction remarquable). On retrouve les trois principes clés de ce premier modèle : le choix consenti, la transparence (plus ou moins relative) sur ce qui est collecté et le partage de la valeur. Ici aussi, il faut faire une différence entre le fait de partager une donnée ou une inférence faite sur ces données. Par exemple, les données de réseaux sociaux (les « amis » de Facebook, les « contacts » téléphoniques…) peuvent servir à établir des inférences commerciales, en vertu de l'homophilie (« qui se ressemble s'assemble ») qui veut que les comportements des individus ressemblent, statistiquement, à ceux de leurs amis. Les opérateurs téléphoniques, tels que Vodaphone ou Telefonica, ont déjà démontré expérimentalement la puissance que donne la connaissance du réseau social d'un individu pour prévoir son comportement.

Le second modèle – celui de la confiance et du consentement explicite – est par exemple celui des objets connectés tels que la balance Withings, mais

également celui de la banque ou de l'assurance : l'entreprise collecte des données qui sont personnelles et qui n'ont pas vocation à être partagées en B2B. Le *Big Data* permet à l'entreprise qui collecte d'offrir des nouveaux services en exploitant le passé (la capacité d'analyse des données accumulées – on trouve par exemple le champ du *quantified self*, qui permet à l'internaute de mieux se connaître au travers un ensemble de mesure et d'outils de visualisation) et la prévision à court terme. Le service *Google Now* est un bon exemple de *Big Data* appliqué à l'assistance à la personne, fondé sur la capacité à prévoir l'environnement de l'utilisateur dans les heures qui viennent. Le *Big Data* appliqué aux données personnelles va se développer fortement avec le déploiement des objets connectés et des senseurs, comme cela a été expliqué en introduction. Nous renvoyons le lecteur au rapport de la commission Anne Lauvergeon [1], ou aux ouvrages de la bibliographie (en particulier [5], [8], [9], [13]). Même si cela pose des problèmes de respect de la vie privée que nous allons évoquer dans la section suivante, deux choses sont claires : ce mouvement est en marche de façon irréversible [2] [13][4], et il représente une formidable opportunité de création de valeur pour nos entreprises [1] [2] [5] [13][5].

Nous avons fait référence aux « 3V » qui sont traditionnellement associés au *Big Data* sans évoquer les variantes, tels que l'ajout d'un « V » pour « value », parce qu'il nous semble implicite (effectivement, le traitement de grands volumes de données n'est pas un but en soi, il a pour objectif de créer de la valeur pour l'entreprise qui le pratique, mais c'est le propre de tout traitement d'un système d'information) ou l'ajout d'un « V » pour véracité. Ce dernier pourrait sembler logique, puisque un des principes du *Big Data* est de combiner de multiples sources d'information, dont une bonne partie est externe, obtenue grâce aux API d'autres organisations qui jouent un rôle de plate-forme, tels que les réseaux

[4] Nous citons [13] parce que le côté irréversible est explicitement mentionné, mais aussi [2] parce que le rapport au Président Obama montre la fierté des États-Unis face à l'écosystème de données et de ciblage publicitaire, considéré, à juste titre, comme un avantage stratégique et compétitif.

[5] Dans ce même rapport [2], on trouve "When wrestling with the vexing issues *big data* raises in the public sector, it can be easy to lose sight of the tremendous opportunities these technologies offer to improve public services, grow the economy, and improve the health and safety of our communities".

sociaux ou les répertoires d'information publiques. Mais toutes les statistiques publiées sur la qualité des données dans les systèmes d'information (*cf.* [14]) montrent que les données inexactes sont courantes dans la très grande majorité des systèmes d'information, et il n'existe aucune étude sérieuse qui montrerait que l'agrégation avec des données externes est un facteur de dégradation de la qualité des données. Autrement dit, traiter les incertitudes et les erreurs dans les données est une problématique classique de tout système d'information, pas une spécificité du *Big Data*[6].

3.2. Une autre façon d'utiliser des statistiques

Un des débats central de l'approche *Big Data* est celui sur la différence entre corrélation et causalité (qui fait l'objet d'un chapitre important dans [5]). Les méthodes numériques du *Big Data* sont dans leur grande majorité des analyses de corrélation (en particulier avec un usage important de la régression logistique) ou des méthodes d'apprentissage (une utilisation simplifiée – à cause de la grande taille – des réseaux bayésiens). Une critique fréquente (par exemple, lire le rapport de Deutsche Bank Research [13]) est faite autour de la double constatation de l'absence de modèle et du fait que corrélation n'est pas causalité. C'est ici qu'intervient le deuxième changement de paradigme : l'approche *Big Data* ne s'inscrit pas dans un cadre statique (une analyse pour prévoir puis agir), mais dans un cadre dynamique d'ajustement continu (*cf.* la référence au « Marketing Synchronisé » [11] d'où l'on retient cette citation « La stratégie, la création et l'exécution sont devenue une seule et même chose, avec un seul arbitre : le feedback consommateur qu'est la performance mesurée »). Ceci signifie que les corrélations qui sont obtenues par « fouille de données » n'ont pas d'intérêt en soi, elles servent à agir, dans un contexte de mesure instantanée de l'effet de l'action. Les algorithmes efficaces de *Big Data* (qu'on pense à Google pour *AdSense* ou à Critéo) sont placés au cœur d'une boucle détection - action - mesure - réaction,

[6] De façon un peu générale, puisqu'il est vrai que le grand volume de données donne accès à des méthodes propres de détection d'erreur ou de réduction d'incertitude.

ce qui rend le besoin d'un modèle de causalité beaucoup moins convaincant. La mesure de corrélation s'inscrit au sein d'une boucle de contrôle asservie et c'est le résultat financier, l'augmentation du chiffre d'affaires, qui est le juge de paix de l'efficacité de l'algorithme.

Autrement dit, l'approche de fouille des données dans le cadre du *Big Data* est pour l'essentiel un travail expérimental, placé dans le domaine du contrôle dynamique. L'objectif n'est pas de produire de la connaissance client (i.e. comprendre ce que veut le client), mais de produire un processus adaptatif qui conduit à la meilleure satisfaction du client. Le travail de corrélation (par exemple, sur les tickets de caisse, lorsque que l'on cherche les corrélations entre différents types d'achat) produit des pépites de corrélation, qui peuvent être interprétées ou non comme des connaissances client. Si on le fait, on court le double risque de se tromper dans l'instant (mauvaise analyse du cycle de causalité qui peut être très complexe) ou à terme (on vient de détecter une corrélation du moment qui est liée à un contexte qui ne va pas durer). Ce qui rend la méthode efficace, c'est qu'on associe à cette corrélation une action dont on mesure l'efficacité, telle que la proposition de vente (si vous avez acheté un objet, on vous propose un autre dont les ventes sont corrélées ou on crée des promotions dynamiques). Si on reprend l'exemple célèbre de Walmart [5] qui constate la corrélation entre les ventes de paquets de couches et de packs de bières, personne ne s'intéresse à la compréhension causale de cette constatation, mais Walmart se contente de placer les deux à proximité et de constater si ce déplacement augmente les ventes. Ce qui reste complexe dans le monde physique devient un jeu d'enfant dans le monde virtuel, ce qui est le principe du marketing synchronisé [11]. Cette notion de double boucle dynamique n'est pas nouvelle dans le domaine de la connaissance client [14] – on parle de *Real-time analytics* dans le monde des systèmes d'information – mais les méthodes du *Big Data* renouvellent complètement ce qu'il est possible de faire.

Une conséquence de cette approche est que le risque d'erreur est important si l'on sort la détection des corrélations de ce contexte dynamique et que l'on souhaite en faire une méthode prédictive. Les différents experts que nous avons audités, ainsi que les conférenciers de *Frontiers of Engineering 2013* – lors de la session sur le *Big Data*, en particulier Jeff Hammerbacher de Cloudera et Thomas Hofmann de Google – ont insisté sur le fait que le danger naturel du *Big Data* est de produire des « informations » que l'on ne comprend pas, et qu'il est donc

facile de mal utiliser. Cela rejoint un phénomène bien connu des statisticiens (les *spurious correlations*), qui veut que si l'on analyse un nuage de points très riche avec un très grand nombre de variables, on trouve statistiquement un grand nombre de corrélations sans véritable signification. C'est pour cela que l'approche *Big Data* doit s'inscrire dans une « culture système complète » et ne se réduit pas à de la fouille de données. Un très bel exemple de mise en garde est fourni par *Google Flu Trends* (GFT). En analysant les recherches faites sur le moteur de Google autour des mots liés à la grippe, les chercheurs de Google ont constaté qu'ils pouvaient prédire avec une certaine efficacité la propagation des épidémies de grippe. Certains ont tout de suite crié victoire sur la puissance du *Big Data*[7]. Il se trouve que l'analyse plus scientifique et plus critique [15] [16] a montré que cette méthode de prévision était médiocre, avec de nombreux cas importants d'erreur, et que d'autres méthodes plus simples étaient plus robustes. L'article publié sur le blog de Harvard [15] ne remet pas en cause l'approche GFT : « *The initial vision regarding GFT – that producing a more accurate picture of the current prevalence of contagious diseases might allow for life-saving interventions – is fundamentally correct, and all analyses suggest that there is indeed valuable signals to be extracted* ». En revanche, il met en garde contre « l'orgueil du *Big Data* » (ce que nos experts avaient souligné – le fait de travailler en circuit fermé n'est pas un gage de sens) et le besoin de transparence, de la capacité à répliquer et à étudier de façon comparative ces algorithmes[8].

Ce changement de paradigme explique pourquoi il y a souvent une incompréhension sur ce qu'est un *data scientist*. Une plaisanterie qui circule sur le Net veut « qu'il s'agisse d'un informaticien qui connaît plus de statistique que les autres informaticiens, ou d'un statisticien qui connaît plus d'informatique que

7 Non sans raison, puisque Netflix exploite les informations de perte de service qui circulent sur les réseaux sociaux dont Twitter pour améliorer la supervision de son réseau, obtenant de la sorte des meilleures performances que ce qu'il pouvait faire avec des outils classiques de supervision.

8 Ce dernier point est du pur bon sens, mais l'exemple de l'analyse « faussement épidémiologique » qui a fait prédire la fin de Facebook en 2017, et qui a été largement reprise par la presse sans la moindre vérification de l'algorithme utilisé [17], montre que la question de l'expertise et de la certification des algorithmes d'analyse et de prévision est importante.

les autres statisticiens ». Il y a trois raisons de penser que la première formulation est la plus proche de la réalité. Premièrement, il faut maîtriser les outils et la culture *open source* associée, ce que nous avons décrit dans la section précédente. La réussite dans une science expérimentale passe par la maîtrise de la technologie. Deuxièmement, les algorithmes sont simples, même si leur mise en œuvre est complexe. Lors de la conférence *Frontiers of Engineering* précitée, les experts ont déclaré : « *simple methods working on very large data sets outperform complex algorithms working on smaller sets* ». Nous reviendrons sur les complexités de l'algorithmique sous-linéaire, mais les concepts statistiques utilisés sont assez simples. La majorité des programmes les plus célèbres, de Google Adsense ou de Critéo, sont fondés sur des régressions logistiques (selon Thomas Hofmann « Google Add Analytics is the largest machine learning system in the world ») … « avec des millions de paramètres ». Troisièmement, la mise en œuvre d'une boucle de contrôle dynamique est bien plus large que l'aspect de « fouille de données » et demande une certaine compétence en algorithmique et en programmation.

Une dernière conséquence de ce changement de paradigme est que les méthodes de développement sont forcément de type « agile » [18]. Les méthodes agiles prônent un développement incrémental par cycle court, autour d'équipes pluridisciplinaires qui travaillent de façon synchronisée sur le même objectif. Ces nouvelles méthodes de travail sont au cœur de la culture des *Géants du Web*, pour reprendre le titre du livre collectif d'OCTO [19]. Le cabinet de conseil OCTO a étudié les traits communs aux grandes entreprises du Web (Amazon, Google, Facebook, etc.) et a produit un ouvrage passionnant qui montre les différences avec les méthodes classiques de développement de systèmes d'information. Nous y reviendrons dans la section suivante, mais notons tout de suite que le mode de développement agile et collaboratif est un point central. Dans le cas du *Big Data* (dont les géants du Web sont des champions), la structure d'équipe agile permet de mélanger les compétences informatiques, statistiques (et algorithmiques) et métier. Parce que le *Big Data* se pratique sur les bases de données complètes (au lieu d'utiliser un échantillon, ce qui fait dire aux auteurs de [5] qu'une caractéristique du *Big Data* est de prendre la base complète comme source), et puisqu'il faut inclure le travail de détection dans une boucle opérationnelle, ce travail exige une forte collaboration entre les métiers du développement et de la production informatique. En conséquence, on retrouve également une forte présence des

principes et de la culture DevOps [20], un prolongement des méthodes agiles qui unifie le travail des équipes de développement et de production pour implémenter un cycle de déploiement continu.

La transformation induite par le *Big Data* sur notre façon de penser et résoudre les problèmes est plus large que ce qui vient d'être évoqué ici. La combinaison du *Big Data*, de la modélisation et de la simulation permet de donner un nouveau sens au terme « fouille de données », qui va beaucoup plus loin que les méthodes statistiques traditionnelles (à titre d'exemple, c'est le positionnement de la startup TheCosmoCompany). Le livre *The Fourth Paradigm – Data-Intensive Scientific Discovery* [21] propose un tour d'horizon passionnant sur l'impact du *Big Data* sur un très grand nombre de domaines scientifiques et sur l'évolution propre de la « méthode scientifique » qui donne naissance à la *eScience* pour reprendre les termes de Jim Gray. De façon plus générale, comme cela a été dit en introduction, le champ d'application des méthodes de *Big Data* est plus vaste que le domaine de l'entreprise. Qu'il s'agisse des sciences fondamentale – par exemple en biologie ou en génétique – ou dans le secteur public – la gestion des transports multi-modaux ou des villes – les méthodes de *Big Data* et d'apprentissage ont de très vastes domaines d'application.

3.3. Une autre façon de programmer : « *Data is the new code* »

Henri Verdier a introduit le thème *Data is the new code* en nous expliquant que les équipes de Google, lorsqu'elles s'intéressent à une startup ou une autre entreprise pour une acquisition, évaluent la valeur de cette entreprise à partir du volume et de la qualité des données collectées, bien plus qu'elles ne s'intéressent au code développé. La valeur des données dépend à la fois de la difficulté qu'il y aurait à les acquérir (coût de collecte) et de leur valeur estimée d'usage. Le code est perçu comme un artefact lié aux données, destiné à évoluer et facile à remplacer : les données restent (et s'accumulent), le code passe.

Ce slogan provocateur *Data is the new code* signifie que dans ce monde du *Big Data*, le code est conceptuellement moins important que la donnée sur laquelle il s'applique. Il est moins important parce qu'il change constamment, parce qu'il

est composé d'algorithmes simples (les seuls que l'on sache exécuter sur des petaoctets de données) et parce qu'il est le résultat d'une boucle d'apprentissage (autrement dit, et pour caricaturer, Google dit aux entreprises : « donnez-moi vos données et je saurai reconstruire le code » … d'aucuns ajouteraient « mieux que vous »).

Cette nouvelle façon de programmer ne s'applique pas qu'aux nouveaux problèmes et nouvelles opportunités. Elle permet également de reconstruire des systèmes « plus classiques » en utilisant la combinaison du *commodity computing*, de la programmation massivement parallèle et les outils *open source* de distribution de données. L'ensemble permet de gagner un, voire deux, ordres de grandeur sur les coûts, un chiffre qui nous a été confirmé par EDF qui a réalisé un pilote de ré-implémentation de son *datawarehouse* (entrepôt de données) avec des technologies *Big Data* en *open source*. Dans le livre précédemment cité [5], on trouve l'exemple intéressant de VISA qui a utilisé les technologies *Big Data* pour ré-implémenter un traitement et qui a constaté des écarts spectaculaires en performance (il est passé d'un mois de temps de calcul à 13 minutes pour traiter 73 milliards de transaction dans une opération de segmentation) et en coût.

Il est possible de caractériser cette « nouvelle façon de programmer » autour d'un volume très important de données. Premièrement, c'est une programmation massivement parallèle. La clé de distribution des données devient la clé de distribution des traitements (plus le volume est important, plus c'est nécessaire), donc l'architecture des données devient la colonne vertébrale de l'architecture logicielle. Deuxièmement, les contraintes de performance obligent à rechercher des algorithmes « sous-linéaires », dont le temps de calcul croît moins vite que la taille des données qu'ils manipulent. Marko Erman a insisté sur l'importance de cette algorithmique sous-linéaire et sur le fait que les entreprises devaient la maîtriser, alors qu'elle ne « s'achète pas sur étagère » (pour l'instant du moins). Paolo Boldi lors de *Frontiers of engineering 2013* a fait un exposé fort intéressant sur le calcul du diamètre du graphe de Facebook[9], dans lequel il utilise des compteurs « Hyperloglog », un très bel exemple d'algorithmique sous-linéaire. Troisièmement, ces algorithmes sont auto-adaptatifs et partiellement produits par

[9] Chaque individu sur la planète est en moyenne relié par une chaine de 4.7 amis à tout autre individu.

apprentissage à partir des données. L'apprentissage (*machine learning*) devient une compétence fondamentale de la programmation lorsqu'on travaille sur des volumes très importants de données [22].

Lors de cette même conférence FoE 2013, Thomas Hoffman a tenu des propos similaires à ceux d'Henri Verdier : "*Big data is getting at the core of computer science*". Autrement dit, les problèmes auxquels s'intéressent les informaticiens aujourd'hui (intelligence artificielle, robotique, traitement du langage naturel, apprentissage adaptatif de comportement, etc.) nécessitent tous les trois caractéristiques que nous avons évoquées : besoins de calcul massifs donc fortement parallèles (ce qu'on appelle le HPC : *High Performance Computing*[10]), très grand corpus de données, auto-adaptation des algorithmes par apprentissage. La conséquence logique est que le « style de programmation *Big Data* deviendra progressivement le mode dominant de programmation ». Le programme « *Human Brain* » de l'Europe dirigé par l'EPFL est un exemple emblématique de cette approche [23]. Un autre exemple est fourni par les progrès constants en traduction automatique, correction grammaticale et compréhension du langage naturel, qui s'appuient sur l'utilisation de corpus de données sans cesse croissants. Notons également que cette façon de programmer, qui demande de développer une nouvelle forme d'intuition, s'appuie sur de nouveaux outils, en particulier de visualisation des données. Jeff Hammerbacher a prédit lors de *FoE 2013* que la prochaine décennie serait marquée par l'explosion des techniques d'exploration et visualisation, tandis que les outils de stockage et gestion semblent avoir atteint une certaine maturité.

Cela ne fait pas de l'approche *Data is the new code* une panacée [24]. Le principe de la distribution massive des données a ses propres contraintes et se heurte à des difficultés fondamentales de l'architecture de données, connues sous le nom de « théorème de CAP » [25] ou des « *snapshot algorithms in distributed computing* » [14]. Dit simplement, il n'est pas possible d'obtenir à la fois la consistance des données, leur disponibilité et la résilience à la défaillance d'une

[10] La référence au HPC est discutable. Il y a bien un thème commun, celui de faire des calculs complexes grâce au parallélisme massif, mais les architectures de la plupart des systèmes HPC sont très différentes du « *commodity computing* » qu'utilise traditionnellement le *Big Data*. À l'inverse, tous les programmes HPC sont traversés par des problématiques et des expériences *Big Data* (par exemple pour la recherche pharmaceutique).

partie des ressources. Les bases de données du monde *Big Data* choisissent une forme affaiblie de consistance ou de disponibilité. La conséquence logique est qu'il reste des domaines (liés aux transactions – les exigences ACID[11] – et aux garanties de très faible latence) pour lesquels des architectures « plus classiques » restent appropriées.

[11] ACID est un acronyme qui résume les quatre exigences fondamentales d'un système de base de données transactionnelle : Atomicité, Consistance, Isolation et Durabilité.

ÉCOSYSTÈME DU *BIG DATA* : MENACES ET OPPORTUNITÉS POUR LA FRANCE

Cette dernière section va approfondir quelques questions difficiles identifiées pendant les discussions avec nos intervenants. Nous employons le mot d'écosystème pour souligner la nature systémique et le nombre important de parties prenantes d'une démarche *Big Data*. En premier lieu se pose la question des données, qui est la matière première de cette chaîne de création de valeur (d'où la métaphore du « pétrole »). Nous allons revenir sur les tensions entre le respect de la vie privée et la capacité à innover et découvrir. Nous avons souligné dans la section précédente que le *Big Data* dépasse de beaucoup le champ du *Business Analytics* et s'invite dans tous les métiers de l'entreprise comme une nouvelle façon de programmer des systèmes. Nous allons donc insister, comme cela a été dit en introduction, sur le risque disruptif que la non-maîtrise de ces techniques pose aux entreprises. Pour terminer, nous résumerons les préconisations « systémiques » que nous avons pu collecter pendant nos entretiens, une contribution modeste et partielle à l'effort de réflexion sur la compétitivité de la France.

4.1. Écosystème des données et respect de la vie privée

Nous avons identifié deux tensions qui rendent la question du contrôle des données difficiles. La CNIL s'appuie sur une approche « traitement » qui associe, fort justement, une donnée au processus qui la traite. On pourrait parler d'une approche « boîte noire » : la donnée est encapsulée dans un contexte qui détermine sa finalité (ce qu'on va faire avec la donnée) et son emploi (le processus de traitement). C'est l'ensemble qui doit être validé auprès de la CNIL dès qu'une entreprise utilise des données personnelles. La première tension, fort bien expliquée par François Bourdoncle [3], est que l'approche *Big Data* dissocie la finalité du stockage : on collecte des données avant de savoir ce à quoi elles pourront servir. Ce qui est souhaitable pour déployer les méthodes du *Big Data* est une approche « boîte blanche » dans laquelle la donnée collectée pour un premier processus devient accessible pour un second. Le droit au respect de la vie privée (*privacy rights* en anglais) stipule que chaque individu a le droit de décider quelles informations ne doivent pas être divulguées à d'autres personnes (y compris aux employés d'une entreprise qui lui rend un service). Paul Ohm, dans un article célèbre [26], constate que « *privacy and (data) utility are at war* ». Cette deuxième tension exprime que les objectifs du citoyen et de l'écosystème industriel sont divergents. S'il s'agit de données nominatives, le respect de la vie privée pousse à minimiser leur collecte et restreindre leur échange. S'il s'agit de données anonymisées, les contraintes à mettre en œuvre pour garantir une « vraie anonymisation » produisent des jeux de données tellement agrégés (nous allons y revenir) qu'ils ne sont plus capables de servir les ambitions décrite dans la section 3.1. Il y a un parallèle évident entre ces deux tensions que nous pourrions reformuler comme suit :

- une entreprise peut-elle utiliser des données collectées (légitimement) pour un premier traitement dans un deuxième traitement de finalité différente ?
- une entreprise peut-elle créer de la valeur à partir de données personnelles en fournissant un service à un tiers tout en respectant la vie privée du client d'origine ?

Il est difficile de formuler des recommandations ; pourtant trois idées semblent se dégager qui peuvent servir à réguler ces tensions[12] :

- le client ou utilisateur doit décider de façon explicite – que l'on appelle *OPT-IN* – quelles données personnelles échappent au cadre juridique par défaut, qui garantit sa vie privée en encapsulant toute donnée dans un processus (légitimé et figé dans sa finalité). C'est le double principe de l'*OPT-IN* (choix explicite du client d'accepter que des données personnelles soient collectées pour des usages ultérieurs) et du « gagnant-gagnant » (le client étant conscient que cet usage ultérieur va produire de la valeur, il veut en recevoir une partie) ;
- même dans le cadre d'un usage ultérieur, le client souhaite comprendre et maîtriser la finalité de cet usage. Même si cette contrainte pourrait être assouplie par rapport au fonctionnement actuel (*cf.* 1ère tension), ce principe de la finalité – fondateur pour la CNIL – demeure pertinent. En particulier, il pousse à distinguer deux modes : celui dans lequel l'entreprise rend des services, mais conserve ses données (par exemple, identification de prospects, ciblage publicitaire) et celui dans lequel elle revend ses propres données. Il semble très difficile de rassurer un individu sur le respect de la finalité dans le deuxième cas. Par exemple, je peux accepter qu'une entreprise collecte ma géolocalisation parce que j'ai confiance dans cette entreprise et parce que son service m'intéresse (de façon volontaire, en *OPT-IN*), mais cela ne signifie nullement que ces informations puissent être divulguées ;
- le choix consenti (*OPT-IN*) doit s'accompagner du droit à l'oubli, du droit de regard (savoir quelle données sont conservées à un instant donné) et d'un « droit de contexte » qui reste à inventer et qui caractériserait un domaine de finalité, pour l'entreprise qui collecte (*cf.* le point précédent, qui fait écho au *Customer Privacy Bill of Right* [2]).

[12] Le rapport [2] mentionne le Customer Privacy Bill of Rights, une proposition de la Maison Blanche datant de février 2012, sur laquelle cette section est relativement alignée, en particulier en ce qui concerne le respect du « contexte ». Les sept droits du Customer Privacy Bill of Rights sont : le contrôle individuel, la transparence, le respect du contexte, la sécurité, l'accès et la précision, la restriction de la collecte et la responsabilisation de ceux qui collectent.

Le sujet de l'anonymisation des données est un sujet complexe, précisément à cause de la puissance du *Big Data* ! Les outils que nous avons évoqués permettent, à partir de données anonymes et d'informations nominatives collectées sur les nombreuses sources publiques (celles dont nous avons parlé dans les sections précédentes, en particulier les réseaux sociaux), de retrouver les individus anonymisés (« désanonymisation »). Les exemples abondent. Ainsi, des études du MIT ont montré qu'il était facile de retrouver des individus à partir de leur déplacements (qui permettent d'identifier le domicile et le lieu de travail) ; autre exemple célèbre la dé-anonymisation de profils de consommation de films fournis par Netflix. L'article précédemment cité *Broken promises of Privacy : Responding to the Surprising Failure of Anonymization* [26] fait un tour d'horizon de cette question. La solution, qui est promue par la CNIL, consiste à combiner anonymisation et agrégation, pour que les données ne soient plus ré-attribuables à un individu (quasi) unique. Fournir un processus d'anonymisation certifié est un des objectifs prioritaires du plan *Big Data* [3] [13]. Pour décupler la puissance d'innovation, il faut passer à un modèle d'*open innovation* dans lequel ce ne sont pas les mêmes qui collectent et qui valorisent les volumes de données. C'est le schéma qui permet à des petits acteurs innovants – dont les startup ou les laboratoires de recherche – de travailler avec des grandes entreprises. Les contraintes de respect de la vie privée font que ce schéma suppose de travailler, dans l'immense majorité des cas, sur des volumes de données anonymisées. Les contraintes réglementaires et les risques font que l'écosystème *Big Data* a besoin d'un processus de certification qui garantisse aux grandes entreprises qu'elles peuvent échanger avec d'autres acteurs innovants des données anonymes sans prendre de risque sur la vie privée de leurs clients.

Pour les données qui ne sont pas anonymisées, il est intéressant de les séparer en trois catégories : les données de profil (données liées au client qui sont les mêmes pour toutes les entreprises, par exemple l'adresse), les données d'usage (collectées par l'entreprise et propres à son activité) et les données calculées (inférées par les systèmes d'information). De par leur nature, les données de profil se prêtent bien

[13] François Bourdoncle est, avec Paul Hermelin, un des co-auteurs du plan *Big Data* demandé par Arnaud Montebourg. Son approche est que même l'anonymisation doit s'inscrire dans un processus industriel complet et dans une finalité et qu'il est quasi-impossible de faire une « anonymisation complètement générique ».

à l'approche « boîte blanche », ce qui rend leur échange (une fois l'*OPT-IN* validé) acceptable entre des entreprises partenaires (c'est d'ailleurs la base des services d'identification / authentification des géants du net). Les données d'usage représentent l'enjeu le plus essentiel, puisqu'elles rassemblent les données de navigation, de communication et les multiples traces des objets connectés. Ce sont ces données qui nécessitent la définition d'un « domaine de finalité » même dans le cadre d'une collecte consentie en *OPT-IN*, et qui ne se prêtent pas à la revente ou au partage. La raison en est simple : la notion de finalité est abstraite et très complexe à circonscrire. Pour le client, elle s'identifie à l'entreprise ou à la marque, de façon couplée à la confiance qu'a le client qui accepte le contrat de l'*OPT-IN*. Ce contrat est implicitement lié à l'entreprise et il serait très difficile de produire les « clauses juridiques » qui garantissent un respect de la vie privée tout en autorisant le déplacement d'une entreprise à une autre. La troisième catégorie concerne les données que l'entreprise produit (à partir d'algorithmes *Big Data* et de fouille de données par exemple). Ici, le principe de la transparence sur la finalité se heurte au problème de la propriété intellectuelle de l'entreprise et de la complexité des traitements. Ces données doivent donc rester dans le « domaine de la vie privée », soumises aux contraintes actuelles de la CNIL (mode « boîte noire » dans lequel le traitement est validé et figé, la finalité est validée par une expertise et n'a pas vocation à être formalisée et exposée). Cette proposition peut être résumée par le tableau suivant.

	Données Profils clients	Données d'usage	Données calculées
Collecte & droit de regard usager	- Collecte explicite - importation d'autres sources sous *OPT-IN* - droit de regard (accès complet aux données) / oubli	- *OPT-IN* dès que le stockage excède la durée du traitement opérationnel lié au processus métier - Droit de savoir/ oubli	Contrôle externe (type CNIL), hors du droit de regard, dans un contexte de processus métier bien défini.
Usage interne *Big Data*	OK	OK sous contraintes de finalité et *OPT-IN*	Sous contrôle externe (lié au processus métier)
Revente en B2B	OK sous *OPT-IN*	Données anonymisées	non

La situation actuelle du statut des données personnelles est moins claire que ce que nous venons d'écrire dans cette section. Il existe une « zone grise » de données d'usage considérées comme « non sensibles » (telles que les données issue de la navigation, par exemple les « cookies »), qui ne bénéficient pas du mécanisme d'*OPT-IN*, et pour laquelle l'approche préconisée est plutôt l'*OPT-OUT /do not track* [2] (dans le meilleur des cas, par exemple les messages qui préviennent de l'utilisation des cookies qui sont apparus très récemment). Une des questions ouvertes est de savoir s'il est légitime d'enrichir des données de profils avec des « scores » calculés en fonction de ces données « d'usage peu sensibles » (une activité de « data broker » telle que décrite dans le rapport [2]). Par ailleurs, notons également que l'*OPT-IN* n'est pas suffisant pour rendre légitime la collecte de données personnelles. Pour éviter les risques de discrimination, par exemple sur des minorités, le législateur est conduit à exclure des types de données (par exemple, la religion ou l'ethnicité) de la collecte et à en protéger d'autres (par exemple, les données médicales). Toutefois, cette problématique n'est pas propre au *Big Data* et sort donc du cadre de ce rapport.

Il faudrait également souligner le rôle fondamental de la géopolitique des données. La maîtrise des données depuis le niveau de l'État à celui de l'individu en passant par celui des entreprises est la seule possibilité, pour un État, d'être maître de sa souveraineté, comme, pour un citoyen, de sa vie privée ou encore, pour une entreprise, de sa stratégie. L'asymétrie du contrôle des informations crée un pouvoir, que les États-Unis se sont accaparé de façon majoritaire. Cette domination américaine est d'autant plus frappante qu'elle se concentre sur un petit nombre d'acteurs dont la majorité est privée. Seule la Chine semble avoir construit une politique engagée de maîtrise de sa souveraineté sur ses propres données[14]. Cette dimension géopolitique complique le débat sur le respect de la vie privée puisqu'une législation plus contraignante dans un pays peut se transformer en risque de perte de compétitivité économique par rapport à d'autres pays qui seraient plus tolérants en termes de collecte de données.

[14] Ce sujet dépasse le cadre de ce rapport. Pour prendre connaissance avec le sujet, on peut lire la page de Stéphane Grumbach de l'INRIA : http://who.rocq.inria.fr/Stephane.Grumbach/bigdata.html

4.2. « Attaque des Barbares » : le risque de disruption ordinaire

Un des messages les plus importants que nous avons reçu des différents experts est que le *Big Data* touche tous les domaines d'activité et tous les métiers de l'entreprise. Le rapport du CIGREF, *Big Data : la vision des grandes entreprises* [27] énonce que « la finalité du *Big Data* est d'améliorer l'efficacité des prises de décision et rendre l'ensemble de la chaîne de valeur plus efficiente ». La première partie de l'affirmation concerne le « *business analytics* » et recouvre un consensus partagé. Mais c'est la deuxième partie de l'affirmation qui est essentielle : le *Big Data* remet en cause la façon dont nous opérons les « métiers ordinaires » parce qu'il permet de faire différemment et moins cher. Les différents scénarii d' « attaque des Barbares » [15] dont nous avons parlé en introduction en faisant référence à Nicolas Colin et François Bourdoncle s'appuient sur deux principes. Le premier consiste à reprendre un processus métier existant et de l'exécuter « de façon numérique » en profitant de cette nouvelle approche informatique que nous avons évoquée dans la section 3.3. Cette première révolution est une double révolution d'efficacité (dont la rapidité) et de baisse des coûts. Cela nous a été répété plusieurs fois pendant nos entretiens : la première cause de disruption est l'effondrement des coûts. Le second principe consiste à atteindre par une ré-implémentation « complètement numérique » une bien meilleure satisfaction client. Cette meilleure satisfaction est la double conséquence d'une interaction plus riche et d'une capacité à apprendre du client, à co-développer les produits et les services (c'est un axiome fondamental de l'économie numérique, le pivot du best-seller *The Lean Startup* [28], mais également développé abondamment dans les ouvrages cités en référence, tels que [4] [8] [9] [10] [11] [19]). Ceci permet de recomposer la chaîne de valeur et d'inventer des nouveaux modèles d'affaires. Les « Barbares » maîtrisent mieux la relation client car ils comprennent mieux les besoins et les envies de leurs clients.

[15] Le terme de « barbares » renvoie à l'invasion de l'empire Romain, et n'est nullement péjoratif dans ce contexte. Il évoque le choc de culture (informatique) et l'efficacité liée à l'agilité.

- Nous avons déjà mentionné le livre *Les Géants du Web* d'OCTO [19] qui est particulièrement pertinent pour comprendre comment les entreprises doivent évoluer pour ne pas se faire prendre de court sur leur domaine existant d'activité. Sans surprise, on retrouve trois dimensions qui font écho au reste de ce document :
- il faut acquérir des compétences techniques pratiques autour des outils et des méthodes de cette « nouvelle façon » de faire de l'informatique. La maîtrise des outils *open source*, des méthodes et outils de programmation distribuée, des ressources de calcul massivement parallèle[16] est indispensable pour lutter à armes égales ;
- il faut développer une culture de la mesure et pratiquer des boucles d'amélioration continue sur des cycles courts, mais répétés de nombreuses fois. Le livre d'OCTO insiste sur cette pratique de la mesure, qui est la marque des approches expérimentales, parce qu'elle remet en cause une tradition plus conceptuelle et cartésienne de la culture française ;
- les méthodes de travail et les méthodes de développement doivent évoluer pour donner plus d'autonomie, de pouvoir et de reconnaissance à ceux qui font (par rapport à ceux qui conçoivent et ceux qui dirigent). Le livre d'Octo [19] tout comme la culture DevOps [20] insistent à la fois sur l'importance de la collaboration (essentiel dans une équipe de *Big Data* qui exige une collaboration pluridisciplinaire) et de la reconnaissance du talent technique (en particulier du talent de développeur). Ce n'est pas à proprement parler spécifique au *Big Data* [18], mais toutes les entreprises leaders du monde numérique ont développé des cultures qui valorisent les compétences techniques, ainsi que les processus collaboratifs – dont le développement *open source* – pour les enrichir et les entretenir de façon continue.

[16] Cette omniprésence du parallélisme va s'accentuer dans la décennie à venir : multi-cœurs, multi-processeurs, multi-lames.

4.3. Écosystème de création de valeur

Nous allons terminer ce document en évoquant un certain nombre de pistes qui peuvent favoriser le développement d'un écosystème de partenaires autour du *Big Data* en France. Puisque le point de départ est la donnée, il faut encourager toutes les initiatives qui mettent à disposition de cet écosystème (universités, laboratoires, startup et entreprises) des corpus de données anonymisées de grande taille. La démarche Etalab qu'anime Henri Verdier est un exemple à poursuivre et à répliquer[17]. La caisse nationale d'asssurance maladie, par exemple, pourrait fournir des données, soigneusement anonymisées et agrégées, de trajectoires de parcours de santé à un ensemble de partenaires, depuis les fabricants de médicaments jusqu'à la communauté scientifique. Pour aller plus loin, la France doit se poser la question de l'opportunité de la création d'un identifiant numérique unique. La mise en place d'un identifiant numérique unique est un choix de société. Les nombreux pays qui l'on fait bénéficient de coûts d'intégration plus faibles et d'une capacité plus grande à proposer des services numériques. Un tel changement en France passe par un débat consultatif large et un changement législatif en profondeur, ce qui pourrait éventuellement faire l'objet d'un référendum.

Pour développer les collaborations entre la recherche publique et l'industrie, il faut faire émerger un protocole d'expérimentation, permettant de confier un jeu de données sensibles (anonymisées, mais sans garantie d'un processus irréversible) à un partenaire de recherche, sous une double contrainte de non-divulgation et de non-désanonymisation. Autrement dit, il doit exister des règles strictes pour les échanges entre entreprises (*cf.* Section 4.1) et une forme dérogatoire pour les expérimentations. Les pouvoirs publics ont un rôle à jouer pour définir les « règles du jeu » et les sanctions si celles-ci ne sont pas respectées. Ne pas créer d'espace d'expérimentation conduit à appliquer un « principe de précaution » qui stérilisera une partie du potentiel de ces technologies de *Big Data* et placera la France dans une position d'infériorité. L'exemple classique de Netflix que nous avons évoqué, qui crée des concours d'algorithmes autour de ses propres données

[17] Dans la lignée de l'approche data.gov du gouvernement américain, bien mise en valeur dans le rapport [2].

pour améliorer constamment ses méthodes de recommandation, est un modèle à suivre, une fois fourni le cadre d'expérimentation qui régule ce que les partenaires académiques peuvent faire.

La plupart des experts audités nous ont confirmé ce qu'on peut lire partout : nous manquons crucialement de ressources humaines disposant de compétences en *Big Data*. Il est difficile de recruter des jeunes ingénieurs ayant la familiarité requise avec les outils de développement *open source* que nous avons mentionnés et il est encore plus difficile de trouver des ingénieurs ayant quelques années d'expérience, parce que les entreprises mondiales se les arrachent. Il est donc important que les pouvoirs publics, mais aussi les grandes entreprises, favorisent le développement de cycles de formation dans le domaine *Big Data*. Ce besoin de formation se décline dans des multiples domaines [27] : technologique, algorithmique, juridique, systémique, métier. Il faut donc intégrer les formations *Big Data* dans de nombreux cursus telles que les écoles de marketing et commerce ou les écoles de management. Néanmoins, il nous semble que ce dont les entreprises ont besoin pour pouvoir construire leur propre démarche d'apprentissage et d'expérimentation, est, en premier lieu, de jeunes ingénieurs formés aux technologies de la programmation distribuée, de la manipulation de très grands volumes de données et de la programmation système avec des outils *open source*. Comme nous l'avons souligné dans un rapport consacré à l'enseignement de l'informatique [7], il s'agit d'un besoin de compétences techniques qui s'acquièrent par la pratique. Il faut donc favoriser la création de centres de ressources technologiques, tels que les préconise le rapport de la commission Anne Lauvergeon [1]. Il faut également mettre en place des formations en ligne de type MOOC accessibles aussi bien en entreprise que pour l'ensemble du grand-public.

CONCLUSIONS

Ce rapport s'inscrit dans la continuité d'autres rapports [1] [3] [13] [27] et d'ouvrages publiés récemment [4] [5] [8] [9] qui font de *Big Data* une véritable opportunité de croissance et de création de valeur pour l'économie numérique en général et pour la France en particulier. Nous souscrivons entièrement à la conjonction favorable des approches *Big Data* et de l'Internet des objets. Ce que nous avons voulu mettre en exergue dans ce document peut se résumer en trois idées :

- le *Big Data*, c'est une nouvelle façon de faire de l'informatique massivement parallèle, à une époque où les ressources de calculs seront de plus en plus parallèles[18]. Il est donc essentiel d'acquérir ces nouvelles compétences, à la fois techniques et culturelles (savoir profiter de l'écosystème de l'*open source*) ;
- le *Big Data*, c'est une autre façon de programmer des systèmes, en boucle fermée et de façon adaptative, en incluant ses clients ou utilisateurs dans

[18] Puisque la course à la fréquence d'horloge est terminée, la « loi de Moore » ne survit que par un recours au parallélisme.

cette boucle. C'est avant tout une démarche expérimentale – même si elle laisse la place aux développements conceptuels et théoriques – qui est favorisée par un changement de culture de travail, agile et collaborative. Les méthodes d'apprentissage ne se limitent pas à la fouille de données, c'est toute l'entreprise qui pratique le « machine learning » ;

- le *Big Data*, c'est un changement de paradigme qui mérite un accompagnement national des pouvoirs publics et une prise de conscience des grandes entreprises. Le besoin évident en formation doit être couvert par la mise à disposition de centres de ressources informatiques et de corpus de données, pour que les étudiants et les ingénieurs puissent développer par la pratique ces compétences essentielles pour le 21e siècle.

COMPOSITION DU GROUPE DE TRAVAIL *BIG DATA*

Membres de l'Académie

Maurice Bellanger
Yves Caseau
Marko Erman
Hervé Gallaire
Erol Gelenbe
Michel Frybourg
Jacques Lukasik
Pierre Perrier
Alain Pouyat
Bruno Revellin-Falcoz
Gérard Roucairol
Gérard Sabah
Christian Saguez
Éric Spitz

Personnalités extérieures à l'Académie

Alain Brénac
Laurent Gouzènes
Claude Kirchner
Jacques Serris
Hélène Serveille

REMERCIEMENTS

La commission TIC remercie vivement les différents experts que nous avons invités à participer à nos débats :

Claude Kirchner, *INRIA*
George Hébrail, *EDF*
Ludovic Cinquin, *OCTO*
François Bourdoncle, *Dassault Systèmes*
François Bancilhon, *Data Publica*
Henri Verdier, *Etalab*
Yan Georget, *Critéo*
Marko Erman, *Thales*
Gilles Babinet, *Champion numérique auprès de l'Europe*
Pascal Buffard, *CIGREF & AXA*
Judicael Phan & Stéphane Grégoire, *CNIL*
Philippe Dewost, *Caisse des dépots*

S'ils ne peuvent être tenus responsables de nos inexactitudes ou erreurs, ils ont grandement contribué à notre éducation sur le sujet du *Big Data*. L'Académie a également profité de sa participation à l'organisation de la conférence *Frontiers of*

Engineering 2013, avec la NAE américaine, dont le *Big Data* était un des thèmes. En particulier, ce rapport a été influencé par les présentations des orateurs suivants :

Sergei Vassilvitskii, *Google*
Jeff Hammerbacher, *Cloudera*
Thomas Hoffman, *Google*
Paulo Boldi, *Université de Milan*

GLOSSAIRE

Big Data : Megadonnées – le domaine de la collecte, du stockage et du traitement de très large volumes de données.

Business Analytics : Méthode d'analyse des données clients pour mieux les connaître et augmenter la performance métier.

CEP (Complex Event Processing) : Technologies de traitement des événements du système d'information, le plus souvent sous forme de règles.

Commodity computing : Utiliser des ressources distribuées accessibles à tous (commodity), pour réaliser des tâches de calcul massives.

Cookies : Éléments d'information échangés et stockés sur les navigateurs, permettant de tracer l'usage des internautes sur les sites Web.

Datawarehouse : Entrepôt de données, système de stockage massif des données de l'entreprise.

HADOOP : Système logiciel ouvert permettant le stockage et le traitement distribué de larges volumes de données.

IoT (Internet of Things) : Internet des objets

JSON (JavaScript Object Notation) : Format d'échange de données, populaire dans le monde du Web.

Machine learning : Méthode informatique d'apprentissage à partir des données.

MapReduce : Méthode d'analyse de données pour le *Big Data*, qui consiste à décomposer, effectuer des traitements de façon distribuée puis assembler le résultat final.

MOOCS (Massively Open On-line Courses) : Formations en ligne ouvertes à tous (FLOT).

Open source : Logiciel libre, dont le code source est accessible.

Profiling : Méthode d'analyse des données dont le but est la segmentation fine (déterminer les « profils ») des utilisateurs.

Quantified Self : Une approche liée aux objets connectés qui promeut l'utilisation et le partage de la mesure de son activité sous toutes ses formes.

Real-time analytics : Méthodes d'analyse des données comportementales d'un client qui visite un site pour pouvoir adapter le contenu du site en temps reel.

Retargeting : Technique permettant de déduire l'intérêt d'un internaute qui visite un premier site pour personaliser les annonces qu'il reçoit sur un second site.

SQL (Structured Query Language) : Langage de traitement de requêtes de bases de données, dont la standardisation a marqué le succès des bases de données relationnelles.

Web Squared : Expression due à Tim O'Reilly qui signifie l'interpénétration du monde réel et du monde virtuel, avec enrichissement réciproque.

XML (eXtensible Markup Language) : Format d'échange de données utilisé massivement dans le monde des systèmes d'information.

RÉFÉRENCES

[1] Commission Anne Lauvergeon. *Un principe et sept ambitions pour l'innovation.* 2013.

[2] John Podesta & al. *Big Data : Seizing Opportunities, preserving values*. Executive Office of the President, May 2014.

[3] François Bourdoncle. "Peut-on créer un écosystème français du *Big Data* ? ", *Le Journal de l'École de Paris* n°108, Juillet/Août 2014.

[4] Henri Verdier, Nicolas Colin. *L'Age de la Multitude*. Armand Colin 2012.

[5] Viktor Mayer-Schönberger, Kenneth Cukier. *Big Data – A Revolution That Will Transform How We Live, Work and Think*. John Murray, 2013.

[6] Doc Searls. The Intention Economy. *Harvard Business Review Press*, 2012.

[7] Rapport de la commission TIC. *Le rôle de la technologie et de la pratique dans l'enseignement de l'informatique*. Communication à l'Académie des technologies, 2014.

[8] Gilles Babinet. *L'ère numérique, un nouvel âge de l'humanité : Cinq mutations qui vont bouleverser votre vie*. Le Passeur, 2014.

[9] Jean-Pierre Corniou, SIA conseil. *Le choc numérique*. Nuvis, 2013

[10] Phil Simon. *The Age of The Platform – How Amazon, Apple, Facebook and Google have redefined business*. Motion Publishing, 2011.

[11] Marco Tinelli. *Le Marketing Synchronisé.* Eyrolles, 2012.

[12] IBM Global Business Services, Analytics : *Real-world use of big data in telecommunications – How innovative communication service providers are extracting value from uncertain data*. IBM Institute for Business Value, Avril 2013.

[13] Thomas Dapp. *Big Data* – The untamed force *Deutsche Bank Research*, May 5, 2014.

[14] Yves Caseau. *Urbanisation, SOA et BPM*. Dunod, 2011 (4e edition).

[15] David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani. *The Parable of Google Flu : Traps in Big Data Analysis*, http://blogs.iq.harvard.edu/netgov/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf

[16] Tim Harford. *Big data :* are we making a big mistake ?, *Financial Times*, March 28th, 2014.

[17] Juliette Garside. Facebook will lose 80 % of users by 2017, say Princeton researchers. *The Guardian*, January 22nd, 2014.

[18] Jurgen Appelo. *Management 3.0 – Leading Agile Developpers, Developping Agile Leaders*. Addison-Wesley, 2010.

[19] Octo Technology. *Les Géants du Web : Culture – Pratiques - Architecture*. Octo2012.

[20] Paul Swartout. *Continuous Delivery and DevOps : A Quickstart Guide*. Packt Publishing, 2012.

[21] Tony Hey, Stewart Tansley, Kristin Tolle (eds). *The Fourth Paradigm – Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[22] Max Lin. *Machine Learning on Big Data – Lessons Learned from Google Projects*. http://fr.slideshare.net/npinto/harvard-cs264-09-machine-learning-on-big-data-lessons-learned-from-google-projects-max-lin-google-research

[23] *Human Bain Project*. European Commission. https://www.humanbrainproject.eu/

[24] Michael Kopp. *Top Performance Problems discussed at the Hadoop and Cassandra Summits, July 17, 2013*. http://apmblog.compuware.com/2013/07/17/top-performance-problems-discussed-at-the-hadoop-and-cassandra-summits/

[25] Eddy Satterly. *Big Data Architecture Patterns*. https://www.youtube.com/watch?v=-N9i-YXoQBE

[26] Paul Ohm. Broken Promises of Privacy : Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, Vol. 57, p. 1701, 2010

[27] CIGREF, *Big Data : La vision des grandes entreprises*, 2013
[28] Eric Ries. *The Lean Startup*. Crown Business, 2011.

PUBLICATIONS DE L'ACADÉMIE

Les travaux de l'Académie des technologies sont l'objet de publications réparties en quatre collections[1] :

- Les rapports de l'Académie : ce sont des textes rédigés par un groupe de l'Académie dans le cadre du programme décidé par l'Académie et suivi par le Comité des travaux. Ces textes sont soumis au Comité de la qualité, votés par l'Assemblée, puis rendus publics. On trouve dans la même collection les avis de l'Académie, également votés en Assemblée, et dont le conseil académique a décidé de la publication sous forme d'ouvrage papier. Cette collection est sous couverture bleue.

[1] - Les ouvrages de l'Académie des technologies publiés entre 2008 et 2012 peuvent être commandés aux Éditions Le Manuscrit (http://www.manuscrit.com). La plupart existent tant sous forme matérielle que sous forme électronique.
- Les titres publiés à partir de janvier 2013 sont disponibles en librairie et sous forme de ebook payant sur le site de EDP sciences (http://laboutique.edpsciences.fr/). À échéance de six mois ils sont téléchargeables directement et gratuitement sur le site de l'Académie.
- Les publications plus anciennes n'ont pas fait l'objet d'une diffusion commerciale, elles sont consultables et téléchargeables sur le site public de l'Académie www.academie-technologies.fr, dans la rubrique « Publications ». De plus, l'Académie dispose encore pour certaines d'entre elles d'exemplaires imprimés.

- Les communications à l'Académie sont rédigées par un ou plusieurs Académiciens. Elles sont soumises au Comité de la qualité et débattues en Assemblée. Non soumises à son vote elles n'engagent pas l'Académie. Elles sont rendues publiques comme telles, sur décision du Conseil académique. Cette collection est publiée sous couverture rouge.
- Les « Dix questions à… et dix questions sur… » : un auteur spécialiste d'un sujet est sélectionné par le Comité des travaux et propose dix à quinze pages au maximum, sous forme de réponses à dix questions qu'il a élaborées lui-même ou après discussion avec un journaliste de ses connaissances ou des collègues (Dix questions à…). Ce type de document peut aussi être rédigé sur un thème défini par l'Académie par un académicien ou un groupe d'académiciens (Dix questions sur…). Dans les deux cas ces textes sont écrits de manière à être accessibles à un public non-spécialisé. Cette collection est publiée sous une couverture verte.
- Les grandes aventures technologiques françaises : témoignages d'un membre de l'Académie ayant contribué à l'histoire industrielle. Cette collection est publiée sous couverture jaune.
- Par ailleurs, concernant les Avis, l'Académie des technologies est amenée, comme cela est spécifié dans ses missions, à remettre des Avis suite à la saisine d'une collectivité publique ou par auto saisine en réaction à l'actualité. Lorsqu'un avis ne fait pas l'objet d'une publication matérielle, il est, après accord de l'organisme demandeur, mis en ligne sur le site public de l'Académie.
- Enfin, l'Académie participe aussi à des co-études avec ses partenaires, notamment les Académies des sciences, de médecine, d'agriculture, de pharmacie…

Tous les documents émis par l'Académie des technologies depuis sa création sont répertoriés sur le site www.academie-technologies.fr. La plupart sont peuvent être consultés sur ce site et ils sont pour beaucoup téléchargeables.

Dans la liste ci-dessous, les documents édités sous forme d'ouvrage imprimé commercialisé sont signalés par une astérisque. Les publications les plus récentes sont signalées sur le site des éditions. Toutes les publications existent aussi sous forme électronique au format pdf et pour les plus récentes au format ebook.

AVIS DE L'ACADÉMIE

1. Brevetabilité des inventions mises en œuvre par ordinateurs : avis au Premier ministre – juin 2001
2. Note complémentaire au premier avis transmis au Premier ministre – juin 2003
3. Quelles méthodologies doit-on mettre en œuvre pour définir les grandes orientations de la recherche française et comment, à partir de cette approche, donner plus de lisibilité à la politique engagée ? – décembre 2003
4. Les indicateurs pertinents permettant le suivi des flux de jeunes scientifiques et ingénieurs français vers d'autres pays, notamment les États-Unis – décembre 2003
5. Recenser les paramètres susceptibles de constituer une grille d'analyse commune à toutes les questions concernant l'énergie – décembre 2003
6. Commentaires sur le Livre Blanc sur les énergies – janvier 2004
7. Premières remarques à propos de la réflexion et de la concertation sur l'avenir de la recherche lancée par le ministère de la Recherche – mars 2004
8. Le système français de recherche et d'innovation (SFRI). Vue d'ensemble du système français de recherche et d'innovation – juin 2004
   - Annexe 1 – La gouvernance du système de recherche
   - Annexe 2 – Causes structurelles du déficit d'innovation technologique. Constat, analyse et proposition.
9. L'enseignement des technologies de l'école primaire aux lycées – septembre 2004
10. L'évaluation de la recherche – mars 2007
11. L'enseignement supérieur – juillet 2007
12. La structuration du CNRS – novembre 2008
13. La réforme du recrutement et de la formation des enseignants des lycées professionnels – Recommandation de l'Académie des technologies – avril 2009
14. La stratégie nationale de recherche et l'innovation (SNRI) – octobre 2009
15. Les crédits carbone – novembre 2009
16. Réduire l'exposition aux ondes des antennes-relais n'est pas justifié scientifiquement : mise au point de l'Académie nationale de médecine, de l'Académie des sciences et de l'Académie des technologies – décembre 2009
17. Les biotechnologies demain – juillet 2010

18. Les bons usages du Principe de précaution – octobre 2010
19. La validation de l'Acquis de l'expérience (VAE) – janvier 2012
20. Mise en œuvre de la directive des quotas pour la période 2013–2020 – mars 2011
21. Le devenir des IUT – mai 2011
22. Le financement des start-up de biotechnologies pharmaceutiques – septembre 2011
23. Recherche et innovation : Quelles politiques pour les régions ? – juillet 2012
24. La biologie de synthèse et les biotechnologies industrielles (blanches) – octobre 2012
25. Les produits chimiques dans notre environnement quotidien – octobre 2012
26. L'introduction de la technologie au lycée dans les filières d'enseignement général – décembre 2012
27. Évaluation de la recherche technologique publique – février 2013
28. L'usage de la langue anglaise dans l'enseignement supérieur – mai 2013
29. Les Académies d'agriculture, des sciences et des technologies demandent de restaurer la liberté de recherche sur les plantes génétiquement modifiées – mars 2014
30. La réglementation thermique 2012, la réglementation bâtiment responsable 2020 et le climat – novembre 2014
31. Les réseaux de chaleur – décembre 2014
32. Les enjeux stratégiques de la fabrication additive – juin 2015
33. Sur la loi relative à la "transition énergétique pour une croissante verte" – juin 2015

RAPPORTS DE L'ACADÉMIE

1. Analyse des cycles de vie – octobre 2002
2. Le gaz naturel – octobre 2002
3. Les nanotechnologies : enjeux et conditions de réussite d'un projet national de recherche – décembre 2002
4. Les progrès technologiques au sein des industries alimentaires – Impact sur la qualité des aliments / La filière lait – mai 2003
5. *Métrologie du futur – mai 2004
6. *Interaction Homme-Machine – octobre 2004

7. *Enquête sur les frontières de la simulation numérique – juin 2005
8. Progrès technologiques au sein des industries alimentaires – la filière laitière, rapport en commun avec l'Académie d'agriculture de France – 2006
9. *Le patient, les technologies et la médecine ambulatoire – avril 2008
10. *Le transport de marchandises – janvier 2009 (version anglaise au numéro 15)
11. *Efficacité énergétique dans l'habitat et les bâtiments – avril 2009 (version anglaise au numéro 17)
12. *L'enseignement professionnel – décembre 2010
13. *Vecteurs d'énergie – décembre 2011 (version anglaise au numéro 16)
14. *Le véhicule du futur – septembre 2012 (publication juin 2013)
15. *Freight systems (version anglaise du rapport 10 le transport de marchandises) – novembre 2012
16. *Energy vectors – novembre 2012 (vesion anglaise du numéro 13)
17. *Energy Efficiency in Buildings and Housing – novembre 2012 (version anglaise du numéro 11)
18. *Les grands systèmes socio-techniques / Large Socio-Technical Systems – ouvrage bilingue, juillet 2013
19. * Première contribution de l'Académie des technologies au débat national sur l'énergie / First contribution of the national academy of technologies of France to the national debate on the Future oF energies supply – ouvrage bilingue, juillet 2013
20. Renaissance de l'industrie : construire des écosystèmes compétitifs fondés sur la confiance et favorisant l'innovation - juillet 2014
21. Le Méthane : d'où vient-il et quel est son impact sur le climat ? – novembre 2014
22. Biologies blanches et biologie de synthèse – mai 2015
23. Impact des TIC sur la consommation d'Énergie à travers le monde (à paraître, 2015)

COMMUNICATIONS À L'ACADÉMIE

1. *Prospective sur l'énergie au XXI^e siècle, synthèse de la Commission énergie et environnement – avril 2004, MàJ décembre 2004
2. Rapports sectoriels dans le cadre de la Commission énergie et changement climatique :

- Les émissions humaines – août 2003
- Économies d'énergie dans l'habitat – août 2003
- Le changement climatique et la lutte contre l'effet de serre – août 2003
- Le cycle du carbone – août 2003
- Charbon, quel avenir ? – décembre 2003
- Gaz naturel – décembre 2003
- Facteur 4 sur les émissions de CO_2 – mars 2005
- Les filières nucléaires aujourd'hui et demain – mars 2005
- Énergie hydraulique et énergie éolienne – novembre 2005
- La séquestration du CO_2 – décembre 2005
- Que penser de l'épuisement des réserves pétrolières et de l'évolution du prix du brut ? – mars 2007

3. Pour une politique audacieuse de recherche, développement et d'innovation de la France – juillet 2004
4. *Les TIC : un enjeu économique et sociétal pour la France – juillet 2005
5. *Perspectives de l'énergie solaire en France – juillet 2008
6. *Des relations entre entreprise et recherche extérieure – octobre 2008
7. *Prospective sur l'énergie au XXIe siècle, synthèse de la Commission énergie et environnement, version française et anglaise, réactualisation – octobre 2008
8. *L'énergie hydro-électrique et l'énergie éolienne – janvier 2009
9. *Les Biocarburants – février 2010
10. *PME, technologies et développement – mars 2010.
11. *Biotechnologies et environnement – avril 2010
12. *Des bons usages du Principe de précaution – février 2011
13. L'exploration des réserves françaises d'hydrocarbures de roche mère (gaz et huile de schiste) – mai 2011
14. *Les ruptures technologiques et l'innovation – février 2012
15. *Risques liés aux nanoparticules manufacturées – février 2012
16. *Alimentation, innovation et consommateurs – juin 2012
17. Vers une technologie de la conscience – juin 2012
18. Les produits chimiques au quotidien – septembre 2012
19. Profiter des ruptures technologiques pour gagner en compétitivité et en capacité d'innovation – novembre 2012 (à paraître)
20. Dynamiser l'innovation par la recherche et la technologie – novembre 2012
21. La technologie, école d'intelligence innovante. Pour une introduction au lycée dans les filières de l'enseignement général – octobre 2012 (à paraître)
22. Renaissance de l'industrie : recueil d'analyses spécifiques – juillet 2014

23. Réflexions sur la robotique militaire – février 2015
24. Le rôle de la technologie et de la pratique dans l'enseignement de l'informatique (à paraître, 2015)

DIX QUESTIONS POSÉES À…

1. *Les déchets nucléaires – 10 questions posées à Robert Guillaumont – décembre 2004
2. *L'avenir du charbon – 10 questions posées à Gilbert Ruelle – janvier 2005
3. *L'hydrogène – 10 questions posées à Jean Dhers – janvier 2005
4. *Relations entre la technologie, la croissance et l'emploi – 10 questions à Jacques Lesourne – mars 2007
5. *Stockage de l'énergie électrique – 10 questions posées à Jean Dhers – décembre 2007
6. *L'éolien, une énergie du XXIe siècle – 10 questions posées à Gilbert Ruelle – octobre 2008
7. *La robotique – 10 questions posées à Philippe Coiffet, version franco-anglaise – septembre 2009
8. *L'intelligence artificielle – 10 questions posées à Gérard Sabah – septembre 2009
9. *La validation des acquis de l'expérience – 10 questions posées à Bernard Decomps – juillet 2012
10. Les OGM - 10 questions posées à Bernard Le Buanec - avril 2014

GRANDES AVENTURES TECHNOLOGIQUES

1. *Le Rilsan – par Pierre Castillon – octobre 2006
2. *Un siècle d'énergie nucléaire – par Michel Hug – novembre 2009

HORS COLLECTION

1. Actes de la journée en mémoire de Pierre Faurre et Jacques-Louis Lions, membres fondateurs de l'Académie des technologies, sur les thèmes de l'informatique et de l'automatique – 9 avril 2002 avec le concours du CNES

2. Actes de la séance sur "Les technologies spatiales aujourd'hui et demain" en hommage à Hubert Curien, membre fondateur de l'Académie des technologies – 15 septembre 2005
3. Libérer Prométhée – mai 2011

CO-ÉTUDES

1. Progrès technologiques au sein des industries alimentaires – La filière laitière. Rapport en commun avec l'Académie d'agriculture de France – mai 2004
2. Influence de l'évolution des technologies de production et de transformation des grains et des graines sur la qualité des aliments. Rapport commun avec l'Académie d'agriculture de France – février 2006
3. *Longévité de l'information numérique – Jean-Charles Hourcade, Franck Laloë et Erich Spitz. Rapport commun avec l'Académie des sciences – mars 2010, EDP Sciences
4. *Créativité et Innovation dans les territoires – Michel Godet, Jean-Michel Charpin, Yves Farge et François Guinot. Rapport commun du Conseil d'analyse économique, de la Datar et de l'Académie des technologies – août 2010 à la Documentation française
5. *Libérer l'innovation dans les territoires. Synthèse du Rapport commun du Conseil d'analyse économique, de la Datar et de l'Académie des technologies. Créativité et Innovation dans les territoires Édition de poche – septembre 2010 – réédition novembre 2010 à la Documentation française
6. *La Métallurgie, science et ingénierie – André Pineau et Yves Quéré. Rapport commun avec l'Académie des sciences (RST) – décembre 2010, EDP Sciences.
7. Les cahiers de la ville décarbonée en liaison avec le pôle de compétitivité Advancity
8. Le brevet, outil de l'innovation et de la valorisation – Son devenir dans une économie mondialisée – Actes du colloque organisé conjointement avec l'Académie des sciences le 5 juillet 2012 éditions Tec & doc – Lavoisier
9. Quel avenir pour les biocarburants aéronautiques ? – juillet 2015