La « vague de fond » du BigData

Entretien avec Arnaud LAROCHE

Co-fondateur et dirigeant de Bluestone, société de conseil en Datascience

Ce qui caractérise l'ère du « BigData », ce n'est pas seulement la taille des fichiers de données, ou la vitesse à laquelle on doit les traiter : c'est surtout la place qu'on leur attribue dans les processus industriels et de services. Désormais, c'est « la donnée » qui pilote des applications et des décisions. Les cas d'usage se multiplient, du micro-ciblage de publicités au pilotage de réseaux d'énergie ou de transport. Des secteurs comme les télécommunications, l'assurance, la banque sont directement concernés ; et l'information économique et sociale publique sera elle aussi touchée. Ce vaste mouvement donne à réfléchir : quelle est la place des analyses statistiques traditionnelles par rapport aux algorithmes « data-driven » ? Comment donner une maîtrise au citoyen-consommateur sur ses propres informations ? Comment concevoir et mieux maîtriser les implications sociales de la généralisation des algorithmes ?

Statistique & Société : Pouvez vous présenter votre société « Bluestone » ?

Arnaud Laroche : C'est une société que nous avons créée en 1998 avec quelques amis issus de l'ENSAE[1] pour proposer un service d'exploitation de la donnée aux entreprises désireuses de mieux comprendre leur environnement, pour prévoir et agir. Avec le temps, et les changements technologiques, nous avons évolué dans notre orientation sur la façon de traiter les données, nous sommes passés d'une approche statistique « classique » fondée sur les modèles à une approche qui fait place aux démarches nouvelles comme le « machine learning » où l'on laisse davantage parler les données. Mais il s'agit toujours d'utiliser les données pour répondre aux grands enjeux de l'entreprise. Aujourd'hui Bluestone compte 120 « ,data-scientists ».

S&S : Quelle définition donnez-vous du « BigData » ?

AL : Le BigData, c'est une vague de fond qui résulte de quatre évolutions, ou révolutions. La première, c'est la digitalisation de notre monde, élément déclencheur : nous sommes désormais entourés de capteurs, nous laissons partout des traces informatiques. Ensuite, il y a la révolution technologique : le remplacement de gros « supercalculateurs » par une myriade de petites machines travaillant en parallèle, dont le nombre peut être augmenté ou réduit en fonction des besoins. Cette révolution technologique divise les investissements requis pour entrer dans ce domaine et les met à portée de petites sociétés innovantes comme la nôtre. En troisième lieu, vient le progrès de la science des algorithmes capables d'opérer sur de très

1. Ecole Nationale de la Statistique et de l'Administration Économique.

gros volumes de données de façon plus exploratoire, comme ceux qui relèvent du « machine learning », etc. Et enfin, last but not least, depuis quelques années, et surtout depuis un an ou deux, les dirigeants d'entreprises ont changé d'attitude vis-à-vis des données. C'est une sorte de révolution culturelle : autrefois, les données de l'entreprise étaient vues comme des sous-produits des activités de gestion, analysées par des équipes de « data-mining » dont l'influence dans l'entreprise était réduite ; aujourd'hui, on construit des applications, des services, des processus qui sont conduits par les données. Le « buzz médiatique » est à la fois cause et conséquence de cette sensibilisation du « management ».

S&S : A partir de quelle taille des données est-on dans le domaine du « BigData » ? Peut-on esquisser un ordre de grandeur ?

AL : Je ne m'y essaierais pas. Bien avant qu'on parle de « BigData », on traitait dans certains domaines (astrophysique, génomique) de grandes bases de données. Mais c'était avec des architectures informatiques centralisées. L'émergence d'outils et de technologies différentes, notamment « Open source » autour de l'écosystème « Hadoop », permet de faire plus de choses à moindre coût, tant du point de vue des quantités de données que du point de vue du temps : on peut désormais réagir en continu à des données évolutives. Et sur cette nouvelle base technique s'est développé un changement « culturel » de la relation aux données : on met les données à la racine des efforts de l'entreprise pour traiter des enjeux économiques, sociétaux, etc. C'est cet ensemble technologique et culturel qui caractérise le « BigData ».

S&S : Avec cette définition, qui fait vraiment du BigData en France ? Il y a beaucoup d'entreprises qui en sont là ?

AL : Aujourd'hui, on met cette estampille partout, y compris sur des applications traditionnelles. Mais il y a des secteurs où déjà de vrais projets « BigData » au sens où je viens de le définir sont en place : télécommunications, assurances, un peu dans les banques. Dans beaucoup d'autres secteurs, on passe actuellement d'une phase d'observation à une phase d'industrialisation, et la demande de compétences sur les technologies et les outils BigData est en pleine explosion.

S&S : Pouvez-vous donner des exemples d'applications ?

AL : Les usages les plus connus relèvent de la personnalisation de la « relation-client ». Il s'agit d'utiliser les données internes de l'entreprise et des données web pour mieux gérer la relation client. On peut citer comme exemple le microciblage en temps réel du client potentiel réalisé en France par Critéo[2] pour offrir aux annonceurs une publicité sur internet à la performance personnalisée déclinable à l'échelle mondiale. Dans la même veine Netflix, récemment introduit en France a fondé son modèle économique sur des algorithmes qui recommandent à ses clients des films suceptibles de les intéresser. Les films ainsi proposés à l'abonné le sont sur la base d'algorithmes et de critères calculés sur des grandes masses. Le choix ainsi adapté au client semble pousser à la découverte, mais il présente également par construction le risque d'enfermer dans une mono culture. Les banques qui ont une relation client plutôt basée sur l'offre produit travaillent sur une plus grande personnalisation de l'offre en s'appuyant sur des données clients internes (transactions) croisées avec des données externes sur les moments de vie (recherches des clients sur le web via les cookies, parcours de recherches sur le web).
Ce sont les applications les plus médiatisées aujourd'hui. Mais il y en a beaucoup d'autres, moins visibles par le grand public.
On assiste depuis un an en France à une montée en puissance sur la « maintenance prédictive », en particulier dans l'industrie. En effet, d'une part l'utilisation intensive de capteurs se généralise dans les process de fabrication pour mieux contrôler les paramètres techniques (température

2. http://www.criteo.com/fr/what-we-do/technology/

pression…) et donc le pilotage, d'autre part les données sont stockées et analysées plus facilement et plus rapidement. Ces évolutions permettent de mettre en place des systèmes de détection des anomalies des systèmes basés sur les signaux faibles, qui ne peuvent pas être perçus dans des contrôles qualité de type échantillonnage. Ils permettent de détecter les problèmes plus en amont et de gagner en efficacité et en rapidité par rapport aux systèmes antérieurs basés sur les seuls experts des métiers concernés. On trouve de tels systèmes en France pour les forages de Total. General Electric[3, 4] propose une offre packagée aux entreprises industrielles. Les capteurs embarqués dans les avions permettent d'identifier les dysfonctionnements plus en amont pour mieux programmer les opérations de maintenance. L'afflux de données permet de mieux comprendre les enchainements temporels et de cerner plus précisément le lien entre les réparations effectuées et la résolution des dysfonctionnements constatés. Bien sûr, cette analyse ne peut se faire sans l'apport de l'expertise propre au métier. On pourrait trouver d'autres exemples de telles innovations dans la gestion des réseaux de transport, d'eau, d'électricité.

Les techniques d'analyses sont également mobilisées pour optimiser les chaines logistiques « supply chain management »: dimensionnement des entrepôts, optimisation des tournées de livraison, des capacités d'un parc de transport. Citons comme exemple l'initiative récente de Chronopost visant à diminuer ses délais de livraison[5, 6].
Les récurrences dans la configuration des données peuvent également être utilisées pour faire émerger des suspicions de fraude et cibler les contrôles (douanes, carte bleue, assurance chômage).
Les données comme celles de la téléphonie mobile, ou celles qui viennent de la gestion des réseaux de transport peuvent être mobilisées pour réaliser des analyses fines des flux afin de dimensionner les infrastructures, d'optimiser les transports. Elles apportent une amélioration par rapport aux enquêtes auprès des usagers qui sont coûteuses et ne peuvent être réalisées que parcimonieusement. Enfin, les compteurs électriques intelligents devraient permettre à terme de faire des diagnostics et des recommandations à distance.

S&S : Les bonnes conclusions, les bonnes décisions vont elles découler naturellement de la seule analyse des données ?

AL : On voit renaître aujourd'hui la vieille controverse entre les analyses « conduites par les données » (« data-driven ») et les analyses reposant sur des modèles. La statistique traditionnelle adopte la démarche hypothético-déductive, qui utilise des modèles, teste des hypothèses, cherche à comprendre ; le data mining cherche des corrélations sans hypothèses préalables, et cherche à prévoir. Vieille controverse : elle était déjà vive en France lors du renouveau de l'analyse factorielle vers 1970. Aujourd'hui certains de mes confrères disent : « avec le BigData, plus besoin d'être intelligent », « du moment que ça marche, c'est bon ». Cela me semble totalement à l'opposé de la réalité. Certes, il y a des contextes dans lesquels l'efficacité à court terme prime, sans qu'on ait besoin de savoir « pourquoi ça marche » : pour faire en temps réel les meilleures propositions commerciales, pour détecter le plus vite possible les pannes, on peut concevoir des algorithmes « data-driven ». Mais en même temps, s'engager dans une telle voie demande un surcroît d'intelligence, car il faut apprendre à contrôler ces algorithmes pour en maitriser la pérennité. Fonctionnent-ils convenablement ? Fonctionnent-ils de façon stable ou dérivent-ils dans la durée ? Pour le savoir, il faut les soumettre à un véritable « monitoring » dans lequel on doit s'interroger sur l'interprétation des phénomènes, sur la valeur structurelle des modèles, et faire appel à des techniques d'expériences contrôlées comme les sondages. Pour moi, les deux écoles devraient se répondre plutôt que de s'opposer.

3. http://www.ge-ip.com/ii/industrial-internet
4. http://www.ge-ip.com/products/rtoi/c564
5. http://www.decideo.fr/Chronopost-peaufine-ses-delais-de-livraison-avec-le-Data-Science-Studio-de-Dataiku_a7501.html
6. https://evenement.inter.laposte.fr/labpostal/images/conference/data_daitaku.pdf

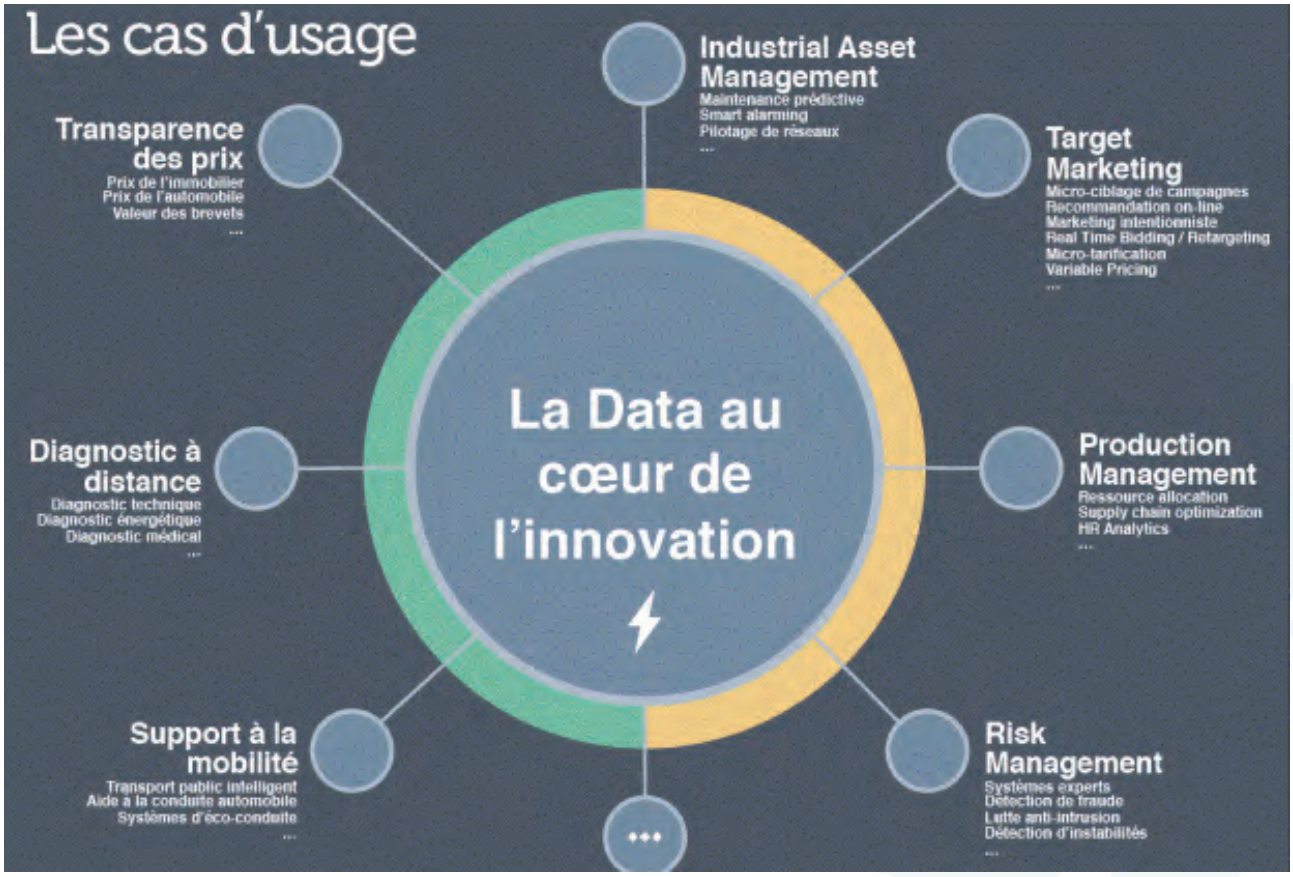

Figure 1 : Sept cas d'usage des « BigData »

S&S : Faut-il pour caractériser le BigData faire des distinctions entre les différents types de données ? Les données structurées, les « traces » qu'on laisse sur Internet, les données textuelles, etc. ?

AL : Il y a bien entendu de grandes différences du point de vue des techniques de traitement des données. Par exemple, lorsqu'on a affaire à des données issues du recueil de formulaires administratifs, les matrices « observations-variables » sont bien remplies, il y a en général peu de valeurs manquantes ; alors que, lorsqu'on utilise les traces laissées sur Internet par les consommateurs pour concevoir un système de recommandation de produits, la matrice « individus x produits » est très « creuse », et cela appelle des techniques de traitement particulières. En matière de traitement des textes, c'est pareil : le « text mining » existe depuis longtemps, mais le traitement des textes spontanés, récupérés sur des forums par exemple, pose des difficultés nouvelles. Il y a un foisonnement de recherches là-dessus pour mettre au point des algorithmes adaptés. Cela dit, ces différences ne me semblent pas être au cœur de la caractérisation du phénomène « BigData ».

S&S : : La statistique publique est-elle menacée par l'émergence du « BigData » ?

AL : Qu'est-ce qu'on attend de la statistique publique ? Qu'elle produise des chiffres sûrs, selon des méthodologies éprouvées, en respectant des principes clairs. Personne ne va s'amuser à lui faire de l'ombre sur ce terrain. La contrepartie est un certain manque d'agilité. Si des initiatives issues du BigData peuvent lui porter tort, c'est dans un domaine bien particulier, celui de la création d'indicateurs économiques avancés à partir de données captées « dans la vraie vie » à

partir de données qui n'ont pas fait l'objet d'un plan de recueil préalable. Un exemple, dans le domaine de l'immobilier, des prix des logements. La statistique publique s'appuie sur les bases des notaires et pour diverses raisons ne peut pas avoir moins de deux mois de retard par rapport à l'évènement. Un indice comme celui de « Meilleurs agents », fondé sur la captation de offres d'agences immobilières partenaires, paraît beaucoup plus tôt, et les 2/3 des articles de la presse spécialisée le citent. En ce cas précis, ce n'est pas du BigData ; mais cela pourrait le devenir. Selon moi, la statistique publique aurait tort de balayer d'un revers de la main ce genre de démarche en disant simplement « ce n'est pas propre » : elle devrait plutôt chercher à innover en traitant ce type de problème – utiliser des données tout venant pour calculer des indicateurs avancés fiables – avec le regard des statisticiens publics. L'exemple de « Google Flu » qui a prédit à tort une épidémie de grippe à New-York il y a quelques années, avec des conséquences fâcheuses pour l'action publique, montre qu'il y a là un réel besoin.

S&S : Venons-en maintenant aux risques des BigData pour les droits des individus. Comment les caractérisez-vous ?

AL : Tout d'abord, il faut savoir mettre en regard les nouveaux services offerts et les dangers réellement encourus. Le dévoilement de données personnelles à des tiers est beaucoup plus fréquent qu'il y a vingt ans : on délivre de l'information sur soi à beaucoup de gens, sans savoir toujours qui ils sont, où ils sont, et sans maîtriser ce qu'ils peuvent en faire et avec qui ils peuvent la partager. On le fait généralement en échange de services qui ne sont accessibles que si on a dit « oui » ! Et l'on désire obtenir ces services. Aussi, il serait vain de s'opposer à une telle déferlante. Je suis sceptique sur la capacité de résister, et je trouve qu'ériger la protection des données personnelles en un principe sacro-saint est une démarche vaine. Mais on peut s'efforcer de rendre de la maîtrise au citoyen-consommateur.

S&S : Comment ?

AL : J'ai un point de vue libéral, qui repose sur l'idéal d'un contrat informé entre l'individu qui veut avoir accès à de plus grands services et les fournisseurs qui ont besoin des données des individus pour développer ces services. Un tel contrat suppose qu'on donne à chaque individu un moyen simple de savoir quelle information sur lui-même il livre, à qui, et pour quoi faire. En particulier, il doit pouvoir contrôler les utilisations en cascade de ses données : à qui seront-elles transférées, ou vendues, et pour quels usages. Actuellement on est perdu : personne ne sait ce qu'il a lui-même autorisé. Il y a un manque de conscience de l'information qu'on laisse, et d'éducation sur les enjeux que cela comporte. Techniquement, un meilleur contrôle est possible : se développent actuellement des outils de « VRM[7] » permettant aux clients d'avoir accès à l'information détenue par les entreprises sur eux-mêmes, et leur donnant une certaine maîtrise sur ces contenus, par un renversement de la logique de « CRM[8] » dans laquelle les fournisseurs de services « managent » leurs clients.

S&S : N'est-il pas impossible de définir à l'avance tous les usages possibles des données ?

AL : On pourrait définir des catégories d'usages, que l'individu pourrait autoriser ou non, selon qu'il souhaiterait ou non accéder à des services plus étendus. Dès à présent on peut acheter de la donnée « Twitter », et il existe des modalités prévues pour cela dans Facebook. Il y a un modèle économique qui se construit autour de la réutilisation des données.

7. « Vendor relationship management » voir http://data-tuesday.com/2013/10/22/decouvrez-les-presentations-de-la-data-tuesday-vrm-8-octobre-2013/
8. « Customer relationship management »

S&S : L'émergence du BigData est-elle porteuse d'autres risques, cette fois au niveau de la société toute entière ?

AL : Se pose la question d'une éventuelle « sur-mathématisation du monde », c'est-à-dire du nombre de plus en plus grand des décisions prises par des machines. Qu'il s'agisse de finance, de décisions concernant des personnes, des interactions sociales, l'invasion des algorithmes n'est pas un mythe : une société vient même de faire entrer un robot dans son conseil d'administration ! Stephen Hawking, cosmologiste, alerte sur les dangers de l'intelligence artificielle et sur une possible perte de contrôle de l'homme sur la machine.

S&S : Vraiment ?

AL : Je pense à l'exemple du trading à haute fréquence : à un certain degré, les effets d'ensemble deviennent incontrôlables. Les modèles de scoring peuvent écarter des pans entiers de la population. La publicité sur Internet devient un marché financier, dans lequel les produits visibles sont déterminés par des algorithmes conçus pour montrer à chacun ce qui est censé lui convenir : cela entraîne un « normage de la société » où tout le monde est informé exclusivement selon sa place à l'intérieur d'une segmentation a priori. Encore une fois c'est une lame de fond, un mouvement qui se fait de toute façon ; mais un mouvement qu'il est souhaitable de contrôler par des démarches qui, elles, ne peuvent pas relever de l'algorithmique. Il faut y réfléchir, et mêler à cette réflexion des gens qui ne soient ni dans l'adhésion complète, par exemple du fait de leur implication professionnelle, ni dans l'opposition systématique qui conduit à ne penser qu'en termes de réglementation. De toutes façons, le BigData change notre monde : il faut y faire face hors des lobbies et des logiques doctrinaires.