Les *big data*

OBJET

Internet, les réseaux sociaux, les capteurs de toutes sortes installés dans les appareils, les vidéos, les films, les textes et les graphiques produisent une masse impressionnante de données. En 2015, 29 000 gigaoctets d'information ont été produits dans le monde à la seconde, soit 914,5 exaoctets ($914{,}5*10^{18}$ octets) pour l'année[1]. Et ce volume continue de croître à un rythme effréné. On prévoit que, en 2020, il atteindra 35 zettaoctets ($35*10^{21}$ octets). Le couple formé par cette quantité d'informations et les technologies qui la produisent et la traitent est appelé *big data;* données massives ou mégadonnées en français.

Les *big data* sont caractérisées par le volume (grande quantité), la vélocité (vitesse de collecte, de traitement, d'analyse et d'utilisation) et la variété, soit des données structurées (prix, dates, températures, poids, données boursières, etc.) et non structurées (vidéos, images, données audio, textes, etc.).

L'avènement des données massives donne lieu à de nouvelles statistiques et exige de nouveaux outils de traitement et d'analyse que les chercheurs sont à mettre au point. Révolution industrielle, or noir du XXIe siècle ou encore mystification entretenue par les entreprises qui y investissent des sommes faramineuses sont quelques termes utilisés pour caractériser les *big data* dans la littérature. Celle-ci fait néanmoins consensus sur deux défis majeurs du phénomène : le traitement des données et la protection de la vie privée.

Les cinq lectures qui suivent (et les cinq autres pour aller un peu plus loin) permettront au lecteur intéressé de prendre la mesure des *big data*, de leur potentiel et de leurs limites. Évidemment, l'innovation étant récente et évolutive, la littérature qui la couvre n'est qu'à ses débuts. Il importe donc de continuer à suivre les développements qui ne manqueront pas de survenir dans ce domaine au cours des prochaines années.

LES CINQ LECTURES POUR COMPRENDRE

1/ **Warin, Thierry, et Nathalie de Marcellis-Warin, « Un état des lieux sur les données massives », *Rapport Bourgogne,* CIRANO, juin 2014.**

Ce texte présente les *big data* comme un futur instrument de mesure des activités des pays et des entreprises. Le rapport, écrit en juin 2014, explique ce qui fait des données massives une nouvelle révolution industrielle. Et comme les révolutions industrielles précédentes, celle des *big data* s'accompagne aussi d'innovations radicales de procédés que l'on commence à peine à découvrir. Les auteurs définissent deux caractéristiques de ces données : la territorialité qui leur confère une citoyenneté et l'exigence de nouveaux outils de traitement[2].

[1] On prévoit dépasser le zettaoctects ou 10^{21} en 2016.
[2] Les outils les plus utilisés actuellement sont les architectures Hadoop et MapReduce ainsi que le langage de programmation R.

À l'aide d'exemples variés empruntés aux entreprises, à la bourse, à la santé, à la politique et à l'éducation, le rapport illustre comment les données massives affectent les stratégies et les opérations des entreprises, les procédés d'affaires, les institutions gouvernementales et les choix des individus.

Par ailleurs, les auteurs expliquent que la citoyenneté qu'acquièrent les données massives comporte des enjeux juridique, moral, politique ou économique. L'enjeu juridique est relié au droit de propriété intellectuelle des données ou le territoire des sièges sociaux ou des filiales des entreprises qui les collectent. L'enjeu moral a trait aux risques d'atteinte à la vie privée de ceux sur qui portent les données. En ce sens, les données ont une personnalité morale distincte de celle de l'entité qui les génère. L'enjeu politique consiste en la perte de contrôle des entités qui collectent ou gèrent les données, notamment dans le cas des gouvernements. L'enjeu économique consiste en l'avantage concurrentiel que les données pourraient procurer aux entités qui auront la capacité de les collecter et de les analyser pour en faire ressortir des éléments stratégiques.

Selon les auteurs, les données massives apparaissent aujourd'hui comme incontournables et risquent de devenir une base d'analyse de la compétitivité, du niveau de concurrence et de la croissance d'un pays et de ses entreprises.

2 / Cukier, Kenneth Neil, et Viktor Mayer-Schoenberger, « The rise of big data, how it's changing the way we think about the world », *Foreign Affairs*, 3 avril 2013.

Cet article va plus loin que le précédent en montrant comment les *big data* modifient la façon dont l'humanité perçoit, traite et analyse le monde qui l'entoure. Pour les auteurs, deux prémisses sont à la base de la montée des *big data*. D'abord, l'idée que l'on peut apprendre des données massives des choses que ne peuvent révéler les relatifs petits échantillons traditionnels. Ensuite, l'idée que l'on peut mettre en données à peu près n'importe quel aspect du monde.

Les petits échantillons traditionnels procurent une plus grande précision, mais ne permettent pas d'étudier les groupes qui les constituent. Ils supposent aussi la détermination à l'avance des données à collecter et la façon de les traiter. Les données massives n'ont pas ces deux limites. En revanche, en étant souvent parasitées (par des données plus ou moins pertinentes), dynamiques, hétérogènes et interreliées, elles exigent que l'on sacrifie un peu de précision et que l'on s'appuie davantage sur la corrélation entre les événements plutôt que sur la causalité. Cet arbitrage supplante les fluctuations individuelles des données et permet de découvrir des modèles de connaissance cachés. Plusieurs utilisations satisfaisantes sont illustrées dans le texte.

À titre d'exemple, Google a utilisé les 50 millions de termes de recherche les plus utilisés sur son moteur de recherche aux États-Unis entre 2003 et 2008 et les a comparés avec les données historiques de l'apparition de la grippe compilées par les Centres de contrôle et de prévention des maladies (CCPM) à travers le pays. Après plus de un demi-milliard d'itérations, Google a pu identifier 45 termes (tels que maux de tête ou congestion nasale) qui sont fortement corrélés avec l'éclosion de la grippe. Avec plus de un milliard de recherches par jour sur Google, il aurait été impossible pour quiconque de deviner les termes qui seraient les meilleurs prédicteurs en vue de les tester. Actuellement, les CCPM se basent sur la fréquentation des hôpitaux et des cliniques pour déclarer les épidémies grippales, mais ils ont besoin des données d'une semaine ou de deux semaines avant d'en venir à une telle conclusion. Le système de Google en revanche est presque en temps réel.

Concernant la mise en données des aspects du monde, l'article donne des exemples de situations qui s'y prêtent. Il donne aussi les applications qui peuvent découler de l'opération de mise en donnée.

Évidemment, ces utilisations pratiques et intéressantes, ont plusieurs coûts que ne manquent pas de relever les auteurs. Au nombre de ceux-ci : la menace à la protection de la vie privée et l'accentuation de l'asymétrie d'information entre certains gouvernements et leur population. De plus, l'utilisation de ces données rend obsolète la méthode scientifique traditionnelle.

3 / Nora, Dominique, Nicole Pénicaud et Claude Soula, « Révolution *big data*, gros *business* », *Nouvel Observateur*, 17 avril 2014, p. 66-70.

Les auteurs de l'article s'intéressent davantage à l'aspect commercial des *big data*. Ils définissent le phénomène comme la résultante de la multiplication des sources de données et de la croissance exponentielle de ces dernières. Ainsi, 90 % de toutes les données disponibles sur la planète en 2014 ont été créées au cours des deux années précédentes. Et pour illustrer leur propos, ils soutiennent que, à elle seule, la voiture sans chauffeur de Google contient davantage d'électronique que la totalité du programme Apollo.

C'est aussi le pays de Google et du programme Apollo, soit les États-Unis, qui domine dans la course à l'exploitation des données massives. Né et mis en place dans ce pays par les grands agrégateurs de données tels Yahoo, Amazon et Google, le phénomène s'y développe sans être menacé par la concurrence internationale. Ainsi, en 2014, sur une échelle de 0 à 5, les Américains se situaient entre 4 et 5 alors que leurs principaux rivaux, les Européens se situaient entre 1 et 2.

Comme dans le précédent article, les auteurs voient dans la révolution *big data*, deux défis majeurs. Le premier est la menace à la vie privée. Le second, économique celui-là, réside dans la pénurie de main-d'œuvre qualifiée pour exploiter les mégadonnées. En effet, en 2010-2011, 80 % des emplois disponibles dans le secteur de l'exploitation des données massives aux États-Unis n'ont pu être pourvus.

4 / Oillion, Étienne, et Julien Boelart, « Au-delà des *big data*, les sciences sociales et la multiplication des données numériques », *Sociologie*, nº 3, vol. 6, 2015 p. 295-310.

Cet article s'intéresse à l'utilisation des données massives en sciences sociales. Le rapport souligne que les *big data* soulèvent un certain nombre de questions qui n'ont pas encore eu de réponse satisfaisante. Quelques-unes de ces questions sont reliées aux difficultés techniques inhérentes à la masse et à l'hétérogénéité des mégadonnées, aux enjeux de mesure et de traitement, à l'éthique dans l'utilisation des données, à leur capacité à améliorer la recherche et la connaissance dans les sciences sociales.

Aux yeux des auteurs, l'absence de réponse satisfaisante à ces questions est attribuable à deux raisons. Premièrement, l'intérêt des entreprises ayant massivement investi dans les mégadonnées à entretenir l'idée qu'une révolution est en cours sans plus de démonstration. Deuxièmement, le flou qui entoure la définition même des données massives. Selon eux, même la caractérisation par les « 3v » (volume, vélocité et variété) est à la fois trop générale et trop restrictive pour s'intégrer à la pratique scientifique.

Les auteurs trouvent le terme « données massives » trop englobant. Ils lui préfèrent celui de données numériques qu'il importe de distinguer en fonction de leur source respective. L'appellation *big data* tend,

selon eux, à confondre la multiplication des bases de données avec leur taille qui peut être modeste pour certaines.

Ils terminent l'article en évoquant que les *big data* ne sont que la face visible d'un phénomène plus large, soit la multiplication exponentielle des données numériques. Le principal défi de l'heure réside dans l'évaluation des sources de ces données numériques en fonction de leur intérêt propre afin de juger leur pertinence pour les sciences sociales, entre autres.

5 / Varian, Hal, et Martin Flemming, « Using big data, the urgency of now », *The Economist*, *the world in 2016*, p. 108-109.

Dans cet article, les auteurs affirment que les économistes travaillent souvent avec des données gouvernementales inexactes et non à jour, bien que soigneusement documentées et traitées. Cela vient du délai plus ou moins long entre la mesure des indicateurs économiques et leur publication et les corrections qui leur sont apportées par la suite.

Ils voient donc dans l'avènement des données massives qui sont collectées et traitées en temps réel une opportunité de corriger cette lacune. Pour eux, la combinaison des données publiques (bien construites, mais à faible fréquence) avec celles du secteur privé (à très haute fréquence, mais plus ou moins précises) pourrait produire des statistiques plus exactes et plus à jour que les données officielles actuelles.

À cet égard, des entreprises et des chercheurs intègrent déjà les données instantanées de diverses sources et les mettent sur le marché. C'est le cas d'Intuit[3], de Zillow[4], d'IBM Commerce[5] et du MIT Billion Prices Project[6]. Pour sa part, Google possède des données sur l'embauche qui permettent d'estimer le taux de chômage courant.

On peut donc noter que les auteurs de cet article rejoignent Kennet Neil Cukier et Viktor Mayer-Schoenberger (voir le deuxième article) en ce qu'ils sont prêts à accepter la relative inexactitude des *big data* pour permettre aux gouvernements et aux banques centrales de produire plus rapidement des indicateurs économiques plus à jour.

[3] Intuit est une entreprise américaine de développement de logiciel, notamment d'impôt, destiné aux PME et aux particuliers. Elle offre un indice d'emploi des PME.
[4] Zillow est une entreprise d'annonce immobilière fondée en 2006 et basée à Seattle.
[5] IBM Commerce compile les tendances des ventes au détail sur une base quotidienne.
[6] Le MIT Billion prices Project est une initiative du MIT qui agrège sur une base quotidienne les prix collectés en ligne auprès de plus de 300 détaillants dans plus de 70 pays. Ils couvrent quelque 5 millions de produits.

ET CINQ AUTRES LECTURES (POUR ALLER PLUS LOIN)

1/ Ouellet, Maxime, André Mondoux, Marc Ménard, Maude Bonenfant et Fabien Richert, « *Big data*, gouvernance et surveillance », *Cahier du CRISIS*, 2014-1, 61 p.

2/ Gallet, Mathieu, *La France et l'Europe face à l'enjeu du* big data , Éditions Choiseul, 2014/2, nº 69, p. 7-23.

3/ Berto, John Carlo, Ursula Gorham et autres, « Big data, open government and e-government : issues, policies and rcommendations », *Information policy`* , nº 19, 2014, 15 p.

4/ Reimsbach-Kounatze, Christian, « The proliferation of big data and implications for official statistic and statistic agencies », *OECD digital economy papers* nº 45, 2015.

5/ Australian government, « The Australian public service big data strategy », Department of finance and deregulation, 2013.

Préparé par Samuel Houngué, septembre 2016.