PRES@JE.COM
Droit, Economie, Justice
Rédaction : Jacques Barraux et Albert Merlin
FÉVRIER et MARS 2015 - N°25

Comment utiliser le Big Data

par Michel Volle, *co-président de l'institut de l'iconomie*

La plupart des articles et commentaires autour du Big Data s'appuient sur une conception erronée de ce que sont les données et de la façon dont on peut les utiliser, constate Michel Volle. Il est vrai que l'internet apporte des moyens éditoriaux puissants aux institutions qui produisent des statistiques. Il faut bien sûr être conscient des possibilités et des dangers nouveaux que cela comporte.

Les commentateurs manient avec trop peu de précautions les bombes sémantiques que sont les mots « donnée » et « information »[1]. Des expressions comme « numérisation de tout », « société de l'information », « masse de données », « une ressource peu différente des matières premières comme le charbon ou le minerai de fer » sont en effet trompeuses : incitant à considérer les données selon leur volumétrie, elles engagent l'intuition sur la pente de la « théorie de l'information » de Shannon, qui assimile l'information qu'apporte un message au logarithme de sa longueur après compression. Shannon disait « meaning doesn't matter », affirmation dont l'énergie impressionne mais qui masque une absurdité.

Voici ce qu'enseigne la pratique de la statistique[2] :

1. Les « données » sont en fait des *observations sélectives* : elles ne sont pas « données » par la nature mais définies a *priori* par un observateur afin que leur *mesure* puisse être ensuite « donnée » à l'ordinateur.

2. L'« information » donne au cerveau de celui qui la reçoit une « forme intérieure » qui lui confère une capacité d'action : Gilbert Simondon a conçu une théorie qui s'appuie sur ce constat et elle éclaire le Big Data mieux que ne le fait celle de Shannon. La capacité d'action ne peut cependant se dégager que si les données sont *interprétées*, ce qui suppose de concevoir un lien de *causalité* entre les concepts dont la mesure a été observée.

3. L'analyse des données la plus pointue ne fait cependant que constater des corrélations. Il faut posséder une bonne maîtrise de la *théorie* du domaine observé pour pouvoir passer de la corrélation à la causalité : la corrélation n'est qu'un un *indice*, au sens qu'a ce mot dans une enquête policière, et il faut savoir l'interpréter.

Quelques mots sur ce dernier point : la théorie, c'est le trésor des interprétations antérieures, condensé sous la forme de liens de causalité entre les concepts - trésor qu'il faut souhaiter exempt du dogmatisme, du pédantisme et de l'étroitesse qui sont pour une théorie autant de maladies.

Celui qui ignore la théorie tombera fatalement, comme cela m'est arrivé, dans quelqu'une des naïvetés que les théoriciens ont depuis longtemps appris à éviter. Cependant la tentation est forte : Jean-Paul Benzécri, pionnier de l'analyse des données, a prétendu que celle-ci révélait « le pur diamant de la véridique nature » et les économètres, dont la discipline se rattache pourtant à la théorie, commettent souvent par précipitation la même erreur.

Les auteurs d'un livre à succès ont érigé cette erreur en principe de la démarche, ce qui indique que le

risque est élevé[3] : « *move away from the age-old search for causality. As humans we have been conditionned to look for causes, even though searching causality is often difficult and may lead us down the wrong paths. In a big data world, by contrast, we won't have to be fixed on causality; instead we can discover patterns and correlations in the data that offer us novel and invaluable inisghts.* »

L'expérience des services de renseignement montre cependant que l'interprétation (qu'ils appellent « analyse ») importe *beaucoup plus* que la collecte : mieux vaut collecter peu de données bien choisies, et que l'on sache interpréter, plutôt que de se laisser écraser par une collecte massive.

Tout observer, c'est en effet ne rien comprendre car l'intellect est *nécessairement sélectif : dans la complexité du monde, chacun doit choisir à chaque instant de voir ce qui importe pour son action et donc de ne pas voir le reste. L'intellect est submergé s'il ne trie pas parmi les signaux qui sollicitent la perception : un conducteur qui se laisse distraire par les détails du paysage est dangereux.*

Il est donc périlleux de situer la valeur ajoutée dans les seuls stockage et traitement informatiques des données, et si l'on commet cette erreur le Big Data n'apportera que de la confusion. Si l'on sait par contre s'y prendre pour interpréter les corrélations en s'appuyant sur les acquis de la théorie du domaine observé, le Big Data constitue une ressource, et donc un enjeu.

[1] *Par exemple Stéphane Grumbach et Stéphane Frénot, « Les données, puissance du futur », Le Monde, 7 janvier 2013.*

[2] *Michel Volle, Le métier de statisticien, Economica, 1984.*

[3] *Viktor Mayer-Schonberger et Kenneth Niel Cukier, Big Data, John Murray, 2013*

Télécharger le PDF de l'article

<< Retour au sommaire

PRES@JE.COM

Une publication de l'Institut PRESAJE
(Prospective, Recherche et Etudes Sociétales Appliquées à la Justice et à l'Economie)
30 rue Claude Lorrain 75016 Paris
Tél. 01 46 51 12 21 - E-mail : contact@presaje.com - www.presaje.com
Directeur de la publication : Michel Rouger

Pour ne plus recevoir d'e-mails de la part de Presaje, cliquez ici

>> CONSULTER LES PRECEDENTS NUMEROS