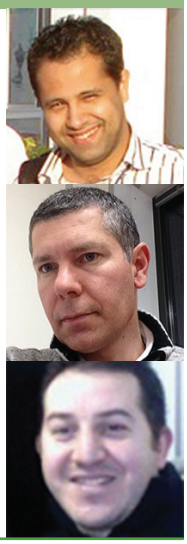
Les trois défis du *Big Data*[1]
Éléments de réflexion



Khalid BENABDESLEM[2]
Christophe BIERNACKI[3]
Mustapha LEBBAH[4]
Groupe « Data mining et apprentissage » de la SFdS

Genèse informatique du Big Data

Le phénomène qu'on appelle aujourd'hui communément le « *Big Data* », ou données massives, appelle une vision globale, c'est-à-dire non limitée aux seuls aspects informatiques par exemple, même si ce phénomène tire essentiellement son origine dans l'accroissement des moyens informatiques et numériques à un coût toujours plus réduit. En effet, le coût de stockage par Mo est passé de 700$ en 1981 à 1$ en 1994 puis à 0.01$ en 2013[5] (prix divisé par 70 000 en une trentaine d'années) tandis que l'on trouve maintenant des disques durs de l'ordre de 8 To à comparer aux 1.02 Go de 1982[6] (capacité multipliée par 8 000 sur la même période) et une vitesse de traitement pour l'ordinateur le plus performant du moment passant d'un gigaFLOPS (le FLOP correspond à *FLoating-point Operations Per Second*) en 1985 à plus de 33 petaFLOPS en 2013[7] (vitesse multipliée par 33 millions).

Il faut bien avoir conscience qu'aucun domaine n'échappe à cette accumulation de données numériques, ce qui justifie pleinement l'intérêt d'un aperçu global. On peut citer une liste bien longue mais qui donne l'ampleur sociétale du phénomène : commerce et affaires (système d'information d'entreprise, banques, transactions commerciales, systèmes de réservation…), gouvernements et organisations (lois, réglementations, standardisations, infrastructures…), loisirs (musique, vidéo, jeux, réseaux sociaux…), sciences fondamentales (astronomie, physique et énergie, génome…), santé (dossier médical, bases de données du système de sécurité sociale…), environnement (climat, développement durable, pollution, alimentation…), humanités et sciences sociales (numérisation du savoir, littérature, histoire, art, architectures, données archéologiques…). Toute la société converge ainsi vers un monde numérique, au point qu'en 2007 plus de 94% de l'information stockée l'était sous forme numérique (les 6% restants sous forme analogique), à comparer à seulement 1% en 1986 (voir la figure 1). En outre, cette

1. Compte rendu de la journée thématique du 13 mars 2015 organisée par la SFdS. Le descriptif de la journée est disponible ici : http://www.sfds.asso.fr/393-Big-Data. On trouvera aussi un lien vers les présentations de cette journée sous forme de pdf et de vidéos (la journée avait été retransmise en direct sur le web).
2. Université Lyon 1, CNRS UMR 5205 LIRIS
3. Université Lille 1, CNRS UMR 8524 Painlevé, Inria
4. Université Paris 13, CNRS UMR 7030 LIPN
5. http://www.capital.fr/enquetes/documents/la-folle-evolution-du-stockage-informatique-953110
6. http://fr.wikipedia.org/wiki/Disque_dur#.C3.89volution_en_termes_de_prix_ou_de_capacit.C3.A9
7. http://fr.wikipedia.org/wiki/FLOPS

quantité d'information stockée dépasse maintenant les 280 Eo (exaoctet), contre 0.02 Eo en 1986 (14 000 fois plus).

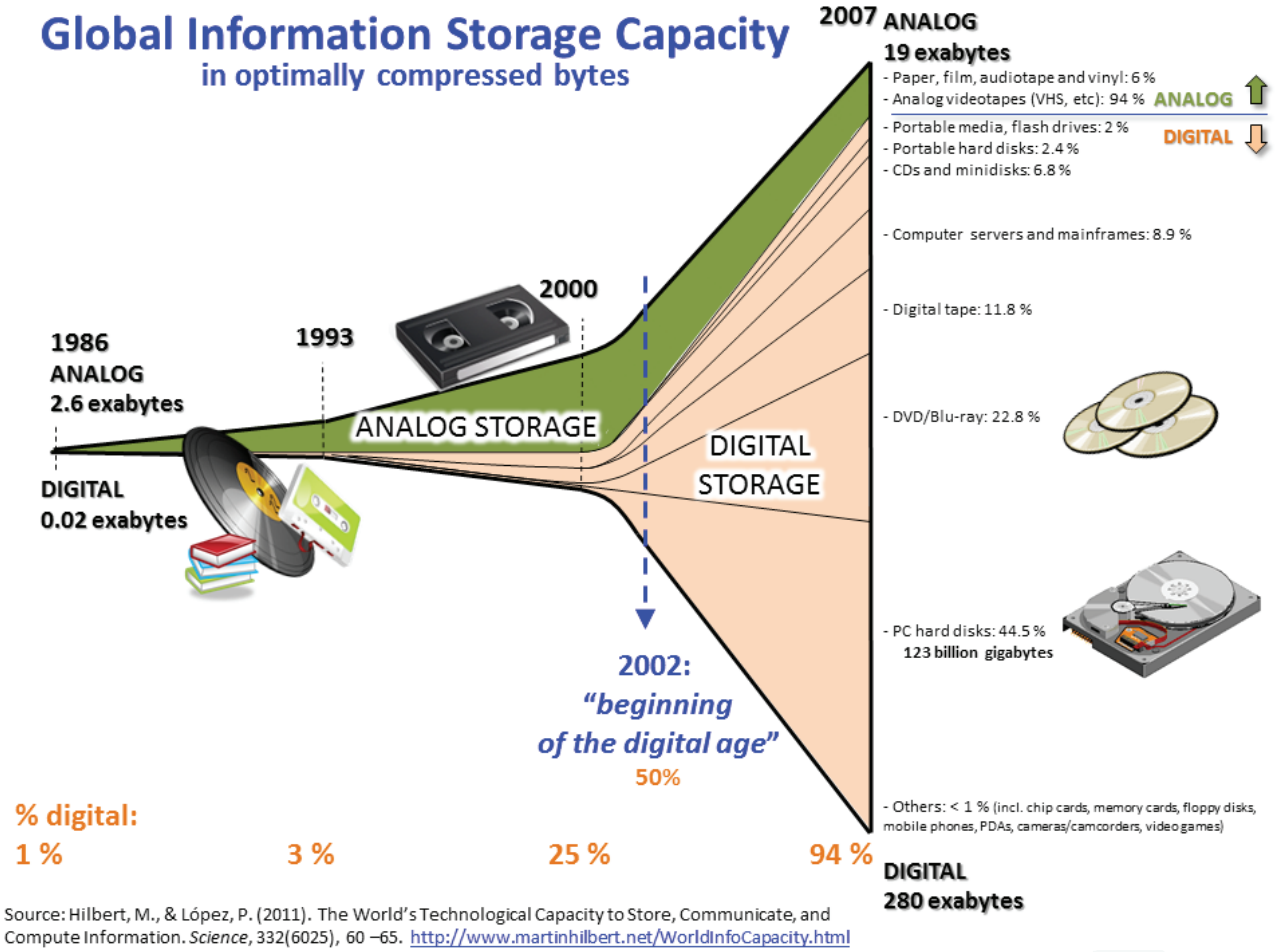


Figure 1. Evolution de la capacité de stockage numérique

Mais cette nouvelle société du « tout numérique » est-elle suffisamment mature pour s'acclimater en profondeur ? Cette avalanche de données pose en effet des grands défis, incontournables, qui ne sont pas totalement résolus à ce jour. Ils sont de trois types : (i) Le stockage et la préservation à long terme et sa pérennisation, (ii) la gestion et l'analyse adéquate en un temps raisonnable et (iii) l'impact sociétal et économique. Ils englobent, sous une autre typologie, les habituels « 5V » qui caractérisent généralement le *Big Data* : Volume, Vitesse, Variété, Véracité et Valeur. La journée thématique du 13 mars 2015, intitulée « *Big Data*: une vision globale: Gestion, Analyse, Éthique et Logiciels », a permis d'aborder l'ensemble de ces aspects par des acteurs spécialisés. Cette journée a permis de rendre compte que ce mouvement de « *Big Data* », qui est très profond, ne peut pas se limiter uniquement aux cinq caractéristiques ci-dessus. C'est aussi une opportunité pour un dialogue interdisciplinaire inédit, suscitant de vraies questions éthiques et juridiques.

Trois défis pour le *Big Data*

Défi du stockage

Comme discuté plus amont, les données massives proviennent essentiellement des facilités d'acquisition et de stockage des données. Cependant, le volume est encore appelé à croître de façon très rapide, par exemple en astrophysique avec le projet Gaïa (2013) ou le projet Euclid

(2021) qui prévoit d'atteindre de l'ordre de 50 Go par jour d'acquisition de nouvelles données. De façon connexe, il ne sert à rien de stocker ces informations si la performance des accès (transfert typiquement) n'est pas garantie pour un futur traitement par exemple, ou encore si leur disponibilité n'est pas assurée. La question de leur protection et de leur préservation à long terme, pour les générations futures, est également posée.

Défi de l'analyse

La facilité de stockage massif conduit inévitablement à peu de sélection *a priori* sur les données à acquérir. Cela peut être vu comme une véritable chance, permettant de garder toute latitude sur les futurs usages potentiels et qui ne sont en fait pas toujours totalement définis au moment même de leur acquisition.

En particulier, de nombreuses questions qui étaient considérées comme hors de portée auparavant deviendront accessibles, avec à la clé une plus-value potentielle importante (avantage compétitif par exemple).

Il faut cependant garder à l'esprit que des données plus massives ne sont pas toujours de meilleures données. Cela dépend si elles sont ou non bruitées, et si elles sont représentatives de ce qui est recherché. En sus, lorsque le nombre de variables croît, le nombre de corrélations erronées croît également. La partie analyse devra prendre en considération ces aspects essentiels.

Seront aussi stockées des données hétérogènes (structurées, non-structurées) ou encore des données incomplètes ou incertaines pour lesquelles des traitements spécifiques sont nécessaires. A ce sujet d'ailleurs, des traitements spécifiques sont déjà requis pour les données plus standard, le volume des données posant déjà en lui-même des difficultés théoriques et pratiques inconnues jusqu'alors, même si certaines vieilles méthodes restent efficaces. Ainsi, les simples tests statistiques[8] deviennent inopérants pour de grandes tailles d'échantillons. On peut citer aussi la difficulté d'analyses multidimensionnelles sur des grands ensembles de données, problème parfois appelé « fléau de la dimension ». Au-delà de l'extraction des connaissances se pose aussi leur interprétation, la visualisation étant jusqu'à présent un outil extrêmement puissant mais qui risque de devenir inopérant par effet de saturation graphique tout simplement. En sus, l'analyse en temps réel de flux continus de données émanant de différentes sources pose elle aussi des difficultés spécifiques. Toutes ces questions impliquent la mise au point de nouvelles statistiques pour le *Big Data*, nécessitant de revoir par exemple des calculs de base comme les tests statistiques et les corrélations[9]. Par rebond, de nouveaux profils de statisticiens pour les mettre en œuvre devront être définis.

Ces outils d'analyse méthodologiques ne peuvent bien entendu être dissociés des outils informatiques et d'écosystèmes dédiés au *Big Data* comme NoSQL, Hadoop, MapReduce ou encore Spark.

Défi sociétal et économique

Le phénomène de *Big Data* ne concerne pas que l'informaticien ou le statisticien. La protection de la vie privée, le droit à l'oubli, les droits de propriété, les droits d'exploitation, le coût énergétique du stockage ou du transfert sont autant de questions touchant le plus grand nombre. D'un point de vue économique, la définition du rôle pris par ces données est aussi posée : matière première ? produits dérivés ? ou capital ? Avec à la clé leur valorisation économique. La fin des monopoles durables est également possible, la donnée permettant de contourner les traditionnelles barrières à l'entrée que représentent par exemple la détention exclusive de

7. Raftery, A.E. Bayesian model selection in social research. Sociological methodology 25, 111-164, 1995.
8. Meyer-Schönberger, V. & Cukier, K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan, 2013.

ressources technologiques ou encore certaines connaissances autrefois monopolisées par les pouvoirs publics. Certes aujourd'hui l'innovation et la découverte de nouveaux procédés avec l'analyse des données créent pour l'entreprise qui en est l'auteur une situation provisoire de monopole mais cette situation est amenée à disparaître rapidement, au rythme accéléré des innovations concurrentes[10]. D'un point de vue sociétal, le statut des données entre propriété privée, domaine public et objet commercial reste souvent flou. Les communautés auto-régulées et le développement transnational vont aussi bouger les lignes actuelles par la grande fluidité de l'information. La gestion fine des déplacements des individus, leur profilage et leur ciblage, le travail sur des recensements plus que des sondages soulèvent des questions de protection des libertés individuelles. D'un point de vue juridique et fiscal, le rôle des États et de leurs instances officielles de surveillance, rôle exercé en France par la Commission Nationale Informatique et Libertés (CNIL), peut changer avec des pertes de ressources fiscales et une pertinence amoindrie des normes juridiques. Ce ne sont pas nécessairement les principes de la loi Informatique et Libertés qu'il faut remettre en cause mais c'est assurément les outils de la régulation qu'il faut adapter. C'est ainsi dans un état d'esprit pragmatique, ouvert et soucieux d'accompagner l'innovation que s'inscrit dorénavant la CNIL[11]. D'un point de vue sécuritaire, on ne peut éviter de penser à une société de surveillance ou de contrôle, symbolisée par Prism et la NSA.

Vers une science des données

La disponibilité de très grandes masses de données et les capacités computationnelles de les traiter de manière efficace sont en train de modifier la manière dont nous faisons de la science. L'informaticien et le statisticien ont pris conscience de l'émergence d'une science des données (*data science*), caractérisée par une collecte massive et variée de données associée à des méthodes de traitements pour en extraire des connaissances nouvelles. A la clé de ce changement de paradigme scientifique se profile aussi un bouleversement de l'enseignement, pour former non seulement les futurs acteurs de cette discipline mais également les citoyens de ce nouveau monde connecté.

Le groupe « DMA » de la SFdS

Le groupe Data Mining et Apprentissage (DMA) de la SFdS vise à tisser des liens étroits entre l'informatique et la statistique, à favoriser les interactions entre théorie, méthodologie et applications, à travailler au renforcement des collaborations entre les membres de la SFdS et ceux d'autres associations savantes connexes. A ce titre, il regroupe des membres de ces différents horizons : informaticiens, statisticiens, praticiens et théoriciens.

10. http://www.creg.ac-versailles.fr/IMG/pdf/La_concurrence_imparfaite.pdf
11. http://www.cnil.fr/linstitution/actualite/article/article/enjeux-2015-2-la-protection-des-donnees-cle-de-voute-de-linnovation/