

Et les futurs historiens rendent jaloux ceux d'aujourd'hui, car ils disposeront de documents bien plus nombreux et complets...

Cette médaille a son revers : il n'est pas si facile de vivre en sachant que chacun de ses gestes laisse potentiellement une trace à la postérité – c'est vivre avec un *Big Brother* hypermnésique. Aussi est-il probable que se développera un véritable droit à l'oubli, afin de permettre à chacun de décider des traces qu'il souhaite effacer ou transmettre. Ces préoccupations sont à l'origine des recherches des informaticiens qui développent des techniques pour protéger la vie privée, par exemple celle des personnes enregistrées dans des bases de données (voir l'article de T. Allard, B. Nguyen et Ph. Pucheral dans ce numéro).

Vers de nouvelles institutions politiques ?

D'autres éléments de la culture humaine transformés par cette capacité nouvelle à collecter, stocker, transmettre et transformer de grandes quantités d'information sont les institutions politiques. Notre démocratie représentative a été inventée aux XVII^e et XVIII^e siècles, à une époque où le débit de l'information que chaque citoyen pouvait communiquer à la collectivité – et, de façon symétrique, la collectivité à chaque citoyen – était très faible.

De fait, en participant tous les cinq ans à l'élection d'un chef d'État parmi 16 candidats, chaque citoyen communique à la collectivité une information de quatre bits. Le débit d'information est donc de 0,8 bit par an – soit 2×10^{-8} bit par seconde. On peut ajouter l'information communiquée lors des élections législatives, locales, etc., l'ordre de grandeur reste le même : quelques bits par an.

Aussi, jusqu'au XX^e siècle, de nombreuses stratégies – tracts, pétitions, manifestations, etc. – ont été inventées pour accroître cette quantité. Ces formes de militantisme ont vieilli et sont aujourd'hui remplacées par d'autres : sites Web, pétitions en ligne, etc. La faible quantité d'information échangée a également pour conséquence que, lors d'une unique élection, nous décidons à la fois du devenir de l'armée, de la police, de la justice, de la diplomatie, de l'éducation, de la recherche, de la santé, des arts, du logement, de l'économie, de l'agriculture, du travail, etc. Un chef d'État, comme



© JN-Design - Dimedro68/shutterstock.com

3. LORSQU'UN CITOYEN VOTE, il communique à la collectivité, via son bulletin, quelques bits d'information à peine. La technologie actuelle permet pourtant de transmettre beaucoup plus d'information par des sites Web, des blogs, etc. À quand des institutions politiques adaptées, qui profiteraient de ces capacités techniques ?

■ BIBLIOGRAPHIE

Dossier « Universités en ligne », *Pour la Science* n° 431, pp. 33-49, septembre 2013.

M. Serres, *Petite Poucette*, Le Pommier, 2012.

E. Hoog, *Mémoire année zéro*, Seuil, 2009.

un roi de jadis, est en effet en charge de toutes ces questions.

L'augmentation du débit de l'information que chaque citoyen peut communiquer à la collectivité et que la collectivité peut communiquer à chaque citoyen permet d'imaginer d'autres formes de démocraties, par exemple des systèmes électifs où plusieurs représentants, élus indépendamment les uns des autres, gèrent chacune de ces questions. La séparation des pouvoirs, jusqu'ici limitée aux trois pouvoirs législatif, exécutif et judiciaire, sera peut-être bientôt rendue possible à une toute autre échelle. Cette augmentation de la quantité d'information que chaque citoyen peut échanger avec la collectivité est sans doute l'une des causes du vieillissement rapide de nos institutions, qui ne sont plus en phase avec les techniques.

De l'information pour nourrir l'économie

Une autre transformation est celle de nos économies, dont l'information est, selon une métaphore déjà un peu usée, le nouveau pétrole. Par exemple, chacun de nos achats en ligne, ainsi que dans un magasin dont nous détenons une carte de fidélité, permet aux commerçants de construire notre profil de consommateur et de nous proposer des produits adaptés à nos besoins et nos désirs. Mais ces constructions de profils individuels ne sont qu'une partie de l'énorme iceberg que représentent les données en économie.

Il est ainsi devenu possible de suivre en temps réel la progression d'une épidémie de grippe simplement en comptant le nombre de requêtes dans un moteur de recherche qui portent sur un mot lié à la grippe – alors que les signalements par les médecins aux autorités de santé ne permettent de suivre la progression d'une épidémie qu'avec plusieurs jours de retard. Ce type d'information est beaucoup plus utile à l'industrie pharmaceutique, qui essaie d'anticiper nos besoins en médicaments, que de construire le profil de tel ou tel consommateur.

De même, les achats de pétrole ou de gaz sont conditionnés à la fois par les prévisions météorologiques à moyen terme et par l'analyse du comportement collectif des consommateurs en réaction à une vague de froid ou à un hiver clément. Sans de telles analyses, il deviendra rapi-

dement impossible pour une entreprise de se maintenir sur le marché du gaz ou des médicaments. Et cela explique, parmi d'autres raisons, que certains pays, telle la Chine, aient décidé de développer leurs propres moteurs de recherche et de limiter l'accès aux moteurs de recherche étrangers, pour ne pas laisser fuir cette information, comme nous le faisons en Europe.

Des formations supérieures pour tous ?

Un autre domaine qui subit l'influence du déluge de données est celui de l'éducation. Depuis quelque temps déjà, les enseignants savent que toutes les connaissances qu'ils transmettent à leurs élèves sont déjà disponibles sur le Web. Le rôle des enseignants n'est plus d'être une source d'information, la bouche d'où sourd la connaissance, mais de donner un sens à cette information, d'aider leurs élèves à se l'approprier et de leur en montrer la cohérence et l'utilité.

Le déluge de données bouleverse le secteur de l'éducation plus radicalement encore, en transformant l'enseignement

à distance. Nous prenons conscience rétrospectivement du caractère élitiste des universités du XX^e siècle, qui s'adressaient uniquement à ceux qui pouvaient être présents quand la connaissance sortait de la bouche de l'enseignant. Pour les autres, qui travaillent en plus d'étudier, ou qui vivent dans un village ou un bidonville loin des centres universitaires, rien n'était possible. L'augmentation des capacités de stockage de l'information, en permettant l'enregistrement des cours en vidéo et leur accès de manière souple et depuis le monde entier, contribuera sans doute à réaliser ce vieux rêve d'une éducation pour tous.

Nous comprenons plus ou moins l'impact de ces possibilités nouvelles de collecter, stocker, transformer et transmettre de l'information à grande échelle sur tel ou tel élément de notre culture – la science, la politique, l'économie, l'enseignement, etc. Mais nous avons encore du mal à penser globalement ces transformations. C'est le défi qui s'offre aujourd'hui à nous. ■

LE DÉFI EST DE PENSER GLOBALEMENT

les transformations induites par ces possibilités nouvelles de collecter, stocker, transformer et transmettre de l'information.

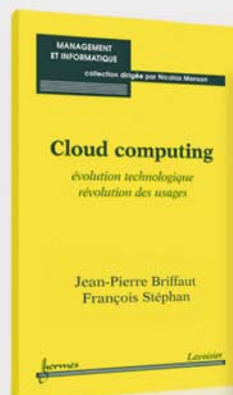
Technologies en mutation... Bouleversement des usages

Disponibles chez
votre libraire habituel ou sur
www.lavoisier.fr



ISBN 9782743014247
316 p. - 2012 - 59 €

ISBN 9782746239128
261 p. - 55 €



ISBN 9782746245112
270 p. - 2013 - 65 €



ISBN 9782746245204
203 p. - 49 €



Lavoisier
Éditeur scientifique, technique et médical



Un déluge de données

Serge Abiteboul et Pierre Senellart

Dans de nombreux domaines, scientifiques ou non, les données s'accumulent en masse.

Les gérer et les exploiter est le défi posé à l'informatique du *Big Data*.

Au CERN, près de Genève, le collisionneur LHC (*Large Hadron Collider*) est équipé d'énormes détecteurs capables d'enregistrer les traces de dizaines à centaines de millions de collisions proton-proton par seconde. Évaluons grossièrement le volume de données que cela représente, en faisant des hypothèses basses. L'information relative aux produits de chaque collision est représentée à l'aide de quatre octets (soit $(2^8)^4 = 256^4 = 4,3$ milliards de possibilités), l'accélérateur fonctionne dix heures par jour en moyenne, et 100 millions de collisions sont enregistrées chaque seconde. On calcule facilement qu'au bout d'un an, le LHC produit ainsi 5×10^{15} octets, soit cinq pétaoctets d'information (une évaluation plus précise fournit des valeurs supérieures, voir l'article de F. Malek dans ce numéro). Il faut 5 000 disques durs du commerce pour stocker une telle masse de données.

Cependant, ce n'est pas le stockage qui est la difficulté principale, mais l'exploitation : extraire de l'information utile – de la connaissance – de cette masse de données dépasse largement les capacités humaines.

Pour relever de tels défis, une nouvelle science est apparue, la science des données. Il s'agit d'extraire des contenus intéressants à partir de masses de données gigantesques, changeantes, hétérogènes, incertaines, et ce à l'aide d'algorithmes efficaces. L'objet de cet article est de décrire les principaux enjeux de cette science des données et les difficultés qu'elle doit surmonter.

L'exemple du LHC par lequel nous avons commencé est extrême en termes de volumes d'octets, mais de nombreux autres domaines scientifiques et industriels doivent faire face à une avalanche similaire de données. En recherche médicale, par exemple, l'un des problèmes a longtemps été l'absence de données en nombre suffisant ; aujourd'hui, c'est plutôt l'inverse : les chercheurs sont confrontés à la difficulté de produire des connaissances utiles à partir des myriades de données accumulées par les différents essais cliniques, les services hospitaliers, etc. Citons la base de données MIMIC-II (*Multiparameter Intelligent Monitoring in Intensive Care*), constituée aux États-Unis par une équipe de chercheurs du