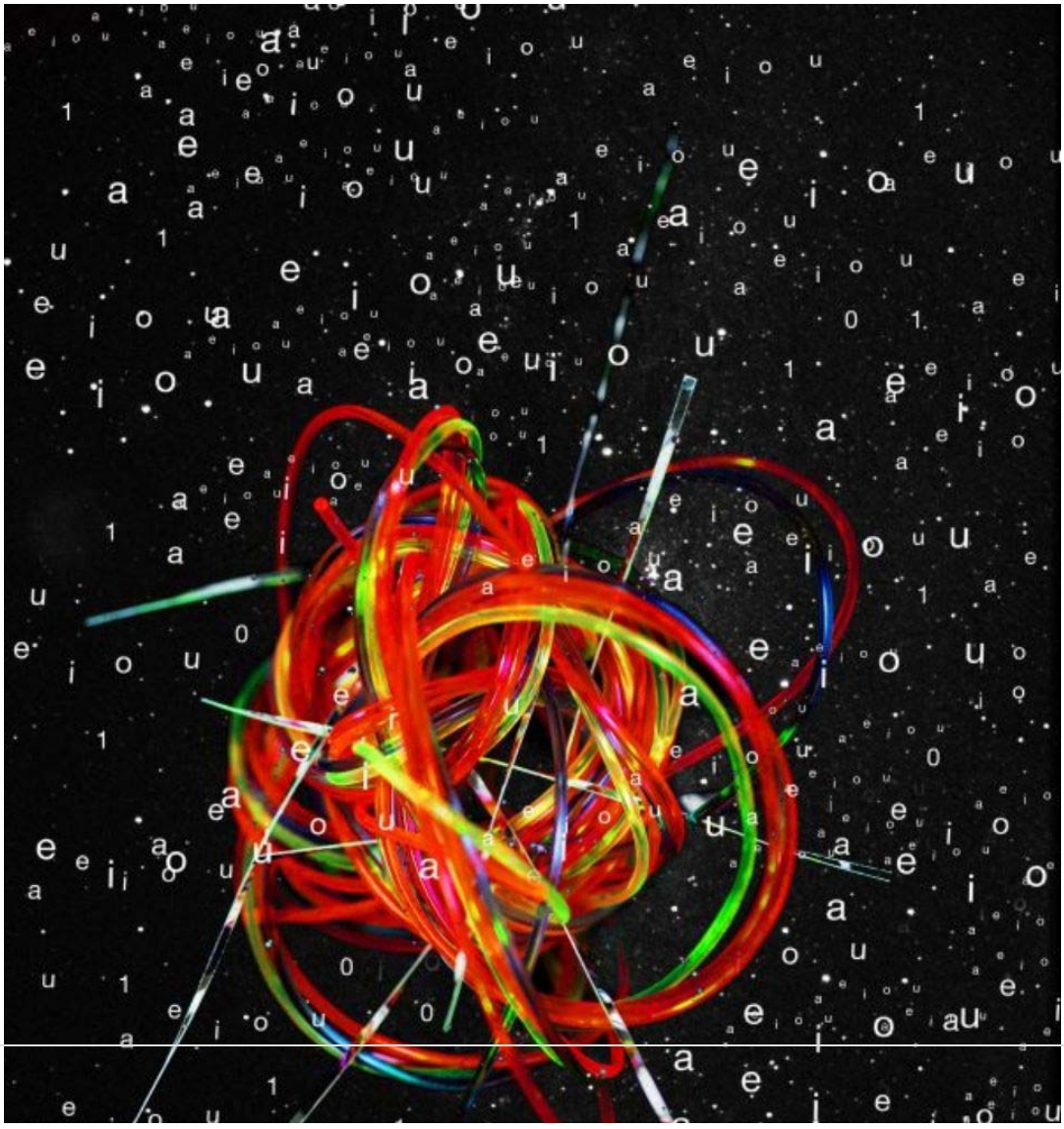
ENQUÊTE

Données le vertige

Par Gabriel Siméon(http://www.liberation.fr/auteur/13388-gabriel-simeon) — 3 décembre 2012 à 19:01



Données le vertige

L'humanité produit autant d'informations en deux jours qu'elle ne l'a fait en deux millions d'années. L'avenir appartient à ceux qui sauront utiliser cette profusion.

Le big data a de l'avenir

Des flots d'octets, un océan de données, un déluge de connaissances... A mesure qu'Internet tisse sa toile, le volume d'informations numérisées n'en finit plus d'exploser. D'ici huit ans, cette masse vertigineuse de «datas» sera 50 fois supérieure à ce qu'elle est aujourd'hui, prédit le cabinet d'études IDC. Et il faudra dix fois plus de serveurs informatiques pour espérer gérer cette déferlante. Pas par crainte d'être submergés, mais plutôt pour être en mesure de retrouver, d'extraire et d'exploiter cette nouvelle manne.

Il y a vingt ans, nous stockions encore nos fichiers sur des disques durs de quelques mégaoctets (1 Mo équivaut à 1 000 000 d'octets, soit 10^6 octets, 1 octet valant 8 bits ; le bit est l'unité de base en informatique, à savoir un 0 ou un 1). Aujourd'hui, la capacité des outils de stockage a dépassé le téraoctet (To, soit 10^{12} octets) et il n'est plus rare pour les entreprises et les organismes de recherche de manipuler des volumes supérieurs au pétaoctet (Po, soit 10^{15} octets). Les

nouveaux usages suivent : une sauvegarde de vos films sur un disque dur externe ? Une photo partagée sur les réseaux sociaux ou une géolocalisation depuis votre smartphone ? Ce sont autant de données qui viennent s'ajouter à la masse enregistrée sur les ordinateurs et les serveurs du monde entier. Même la façon de les interroger devient information : notre historique de navigation sur le Web, nos recherches sur Google...

Les chiffres donnent le tournis : chaque minute, environ 350 000 tweets, 15 millions de SMS et 200 millions de mails sont envoyés au niveau mondial ; pendant le même laps de temps, des dizaines d'heures de vidéos sont mises en ligne sur YouTube, des centaines de milliers de nouveaux fichiers sont archivés sur les serveurs de Facebook. L'ancien PDG de Google, Eric Schmidt, estimait en 2010 que nous produisions tous les deux jours environ 5 exaoctets (Eo, soit 10^{18} octets) d'informations... soit autant *«qu'entre le début de la culture humaine et 2003»* ! Selon l'institut IDC, 1,8 zettaoctet de données (Zo, 10^{21} octets) a été créé en 2011. *«L'information disponible à la surface de notre planète en 2020 devrait tourner autour des 40 Zo... Mais ces estimations sont rendues fausses d'année en année par les nouveaux usages»*, précise Jean-Yves Pronier, directeur marketing du gestionnaire de données EMC.

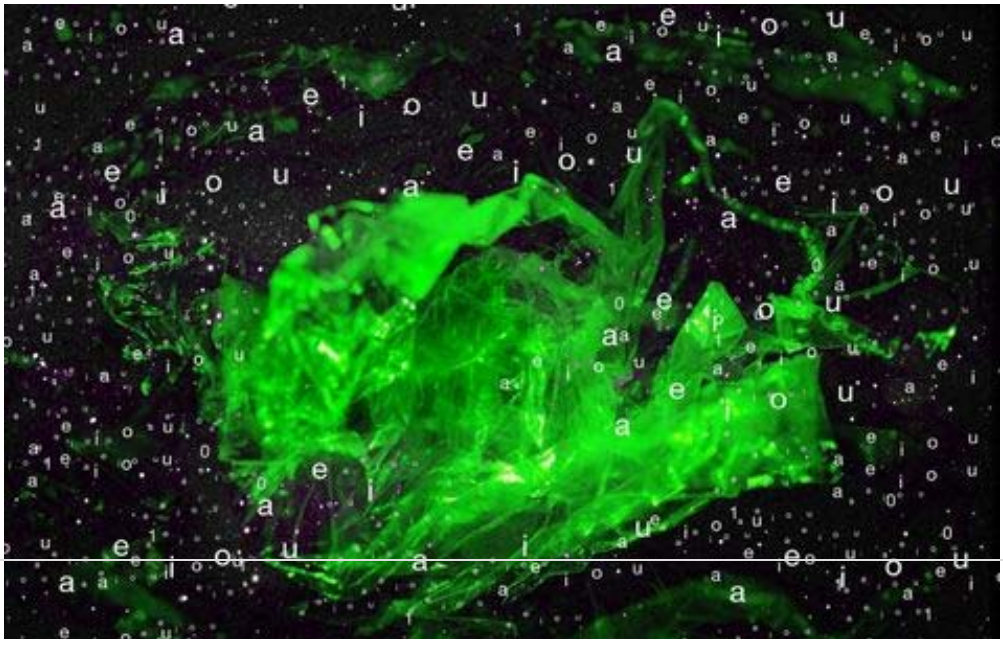


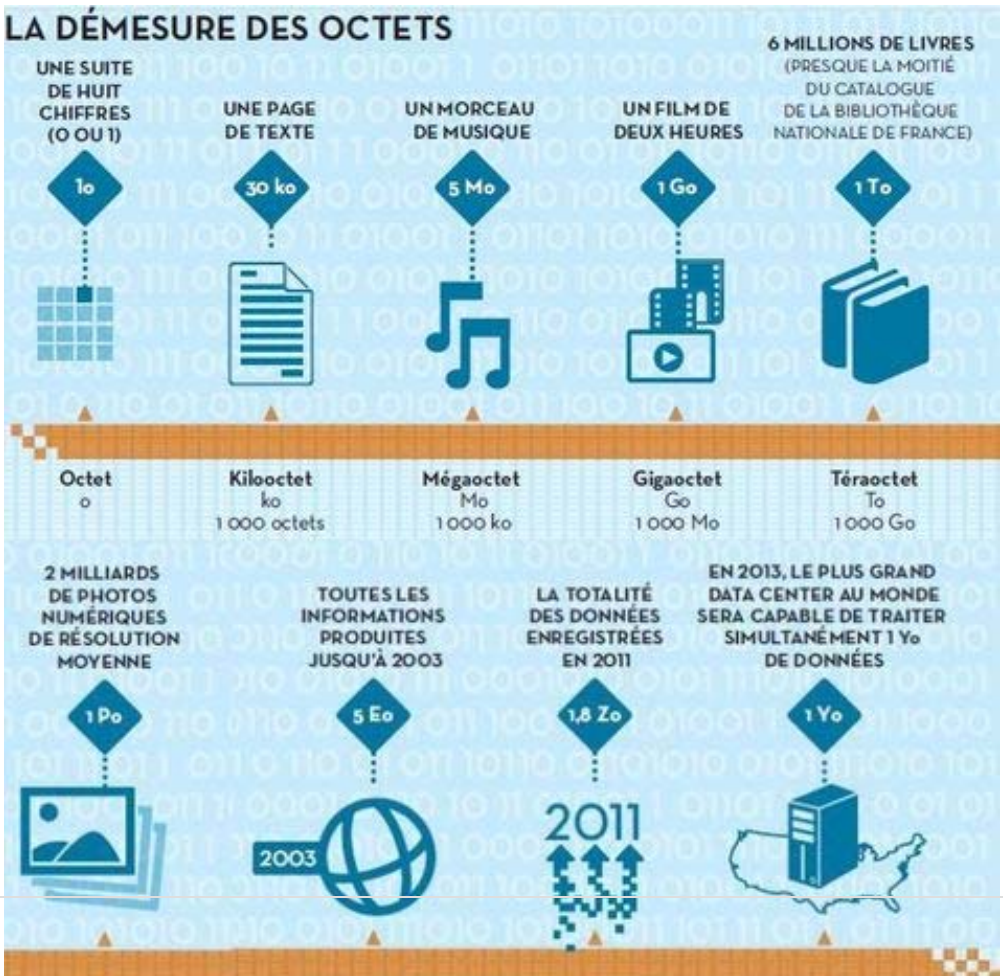
Photo: Emmanuel Pierrot.Vu pour Libération

«Capteurs». *«L'essentiel du volume d'informations généré aujourd'hui l'est encore par des humains,* note Bernard Benhamou, délégué aux usages d'Internet auprès du ministère de l'Enseignement supérieur et de la Recherche. *Mais, dans les prochaines années, il sera produit par des capteurs.»* Caméras de surveillance, sondes météo, cartes bancaires et autres télescopes géants constituent déjà des mines d'informations considérables pour les secteurs concernés. Mises en réseau ou rendues publiques, elles profitent désormais à bien d'autres domaines. *«La nouveauté, c'est la capacité à croiser toutes les données en provenance des capteurs, du Web et de l'open data* [les informations mises à disposition par les pouvoirs publics, ndlr]», explique Serge Abiteboul, de l'Institut national de recherche en informatique et en automatique (Inria). *«C'est bien d'avoir des données, mais c'est mieux de les faire parler. Et, pour cela, les technologies traditionnelles ne suffisent plus»,* souligne Jean-Yves Pronier.

C'est là qu'intervient une nouvelle discipline : le «big data». Il consiste à analyser ces immenses bases de données en faisant tourner des algorithmes qui vont traquer le plus infime lien entre chacun des éléments stockés, puis à livrer les informations en quelques dixièmes de seconde, pour peu que la capacité de calcul des ordinateurs impliqués dans l'opération soit suffisante. Rien de bien nouveau pour Google, habitué à jongler quotidiennement avec des pétaoctets de données pour les besoins de son moteur de recherche. Mais le géant du Web a entraîné dans son sillage nombre de grands groupes désireux de faire émerger les connaissances cachées dans leurs milliards de fichiers texte.

par les autres. Pas étonnant que de nombreuses start-up se soient créées autour de l'analyse des big datas.

Mesagraph fournit ainsi à Canal + une modélisation de son audience à partir des conversations sur Twitter. *«Les téléspectateurs font souvent autre chose pendant qu'ils regardent une émission : ils vérifient les informations diffusées, commentent sur les réseaux sociaux... Et nous arrivons à dire combien tweetent en regardant le Grand Journal, puis zappent sur Secret Story»* , affirme Sébastien Lefebvre, patron de Mesagraph. Comment ? Grâce à une application *«qui collecte les tweets qui nous intéressent, ceux contenant le nom d'une émission ou un hashtag spécifique, puis qui crée des métadonnées décrivant ces tweets,* poursuit l'informaticien. *Une fois analysées, ces informations sont ensuite restituées via une API»,* à savoir une interface qui rend lisible de manière graphique les résultats du traitement informatique (nuage de mots-clés, camembert, etc.).



Épidémie. Santé, sécurité, consommation, transports, sciences, marketing... Les domaines d'application semblent sans limite. *«Les assurances pourront bientôt vous verser des primes en fonction de votre style de conduite, grâce à des sortes de boîtes noires installées dans votre voiture qui enregistreront la moindre information. C'est déjà le cas aux Etats-Unis»,* illustre Jean-Yves Pronier. Le logiciel HealthMap, qui traite en temps réel des données en provenance, entre autres, de l'Organisation mondiale de la santé (OMS), de Google News et bientôt de Twitter pour dresser une carte planétaire des foyers de maladies, a permis de suivre l'évolution d'une épidémie de choléra en Haïti avec près de deux semaines d'avance sur les observations des autorités de santé.

Aux Etats-Unis, un programme développé par IBM est utilisé par la police de Memphis (Tennessee) pour prédire les «zones chaudes» et réduire la criminalité, grâce au croisement de données aussi diverses que les jours de paie, le type de populations par quartier et les rencontres sportives. A Singapour, on sait désormais pourquoi il faut se battre pour trouver un taxi quand il pleut. Une étude menée en 2012 a croisé les données GPS de 16 000 taxis avec les relevés météo et montré que les chauffeurs s'arrêtent de rouler dès les premières gouttes de peur d'être impliqués dans un accident et de devoir payer un malus d'assurance élevé. Le cabinet d'études Gartner estime que les entreprises qui auront intégré toutes les dimensions du big data d'ici à 2015 seront plus performantes de 20% par rapport à leurs concurrentes. Pour des chercheurs du MIT (Massachusetts Institute of Technology, à Boston), ce serait plutôt entre 5% et 6%. Les administrations publiques européennes y gagneraient aussi en efficacité, à en croire un rapport de

des économies pouvant être réalisées.

La course à l'équipement informatique bat donc son plein. En France, un appel à projet doté d'une enveloppe de 25 millions d'euros a été lancé pour développer des technologies d'exploitation de ces très gros volumes de données. Aux Etats-Unis, on voit plus grand encore. Après avoir alloué 200 millions de dollars (155 millions d'euros) à la recherche dans ce domaine en mars, le pays inaugurera en septembre 2013 le plus grand centre de traitement de données au monde. Un centre d'espionnage, à vrai dire : capable d'analyser simultanément plus d'un yottaoctet d'informations (Yo, 10^{24} octets), il aura pour mission d'intercepter, de déchiffrer et de stocker la totalité des communications mondiales !

Qui dit big data dit-il forcément Big Brother ? Le piratage de 24 millions de comptes Sony en 2011 - contenant notamment les informations bancaires des utilisateurs - ou l'affaire des «target coupons» - un Américain a découvert la grossesse de sa fille en voyant la teneur des publicités hyperciblées, envoyées par les commerçants sur la base de l'examen de ses tickets de caisse - obligent à se poser la question de la sécurité et de la confidentialité des informations.

Surtout que la moitié seulement des données nécessitant une protection en bénéficie réellement, selon IDC. Saviez-vous, par exemple, que le ministère de l'Intérieur commercialise les données personnelles de ceux qui ont immatriculé leur véhicule après août 2011 ? Et si, à l'avenir, l'obtention d'un crédit bancaire dépend d'un examen préalable de votre profil numérique, comment éviter les dérapages ?

«Compagnon». L'optimisme semble pourtant de mise pour Jean-Yves Pronier : *«Cela va naturellement bénéficier à la société et aux entreprises. Le big data sera un compagnon de tous les jours pour chacun d'entre nous.»* Vision idyllique

peines. *«Au-delà de 2020, il va sans doute falloir trouver de nouvelles techniques de stockage et des algorithmes encore plus performants*, observe Christine Collet, chercheuse spécialisée en base de données au Laboratoire d'informatique de Grenoble (LIG). *Sans la donnée, on ne peut rien faire. C'est une vraie matière première. Et celle qui aura été transformée vaudra cher.»*Alors, ira-t-on jusqu'à taxer ces informations à forte valeur ajoutée pour renflouer les finances publiques ? Pour cette chercheuse, *«c'est une question qu'on peut se poser»*.

Photos: Emmanuel Pierrot(http://www.emmanuelpierrot.com).Vu pour Libération

Gabriel Siméon (http://www.liberation.fr/auteur/13388-gabriel-simeon)