

## Big data

### Description

Expression anglophone désignant l'exploitation des masses de données numériques, quelle que soit leur nature (textes, chiffres, photos, sons, vidéos, graphiques, clics, liens hypertextes, pages vues...), transmises grâce à Internet. Leur volume ne cesse d'augmenter avec le développement des usages numériques. La multiplication des terminaux connectés, la popularité des réseaux sociaux et l'informatique en nuage (*cloud computing*, voir *REM* n°9, p.43) en ont accéléré la croissance depuis 2010. Les ordinateurs, les téléphones portables et les tablettes sont autant d'émetteurs de données numériques en provenance des particuliers, des entreprises, des institutions ou des Etats, auxquels s'ajoutent les objets connectés (ampoules, réfrigérateurs, compteurs électriques, voitures, jouets, plantes vertes, pour ne citer que des applications domestiques), dont le nombre devrait atteindre 50 milliards en 2020 selon Cisco, ainsi que les caméras de surveillance, les sondes météo et autres télescopes. De grandes capacités de stockage, des algorithmes et une puissance de calcul qui augmente de façon exponentielle constituent les ingrédients nécessaires à l'émergence de ce nouveau défi consistant à produire de la connaissance à partir de l'analyse du volume incommensurable de données générées par les techniques de l'information et de la communication.

Le champ d'application du *big data* est infini, des sciences dures aux sciences sociales, de l'étude des phénomènes climatiques ou astronomiques à celle des mouvements sociaux ou des politiques économiques, avec pour finalité, toujours, de mieux comprendre le monde. Plus prosaïquement, derrière le concept de *big data* se cache aussi une activité lucrative liée à Internet, celle de prédire les comportements des consommateurs en particulier et de débusquer ceux des citoyens en général. Le marché du *big data* est estimé à 50 milliards de dollars dans cinq ans. Sur les 4,4 millions d'emplois créés dans le secteur de l'informatique d'ici à 2015, dont un peu moins de la moitié aux Etats-Unis, grâce en partie au *big data*, deux tiers d'entre eux pourraient ne pas trouver preneur faute de qualifications adéquates. Doté de compétences en informatique, en statistiques et en marketing, le *data scientist* est un profil déjà recherché par un tiers des entreprises aux Etats-Unis.

En 2011, la somme des données enregistrées sur Internet atteint 1,8 zettaoctet (Zo), soit 1 800 milliards de gigaoctets. Chaque minute circulent sur le réseau mondial 2,3 millions de requêtes sur Google, 1,7 millionde contenus partagés sur Facebook, 300 000 tweets, 200 millions de courriels, 72 heures de vidéos mises en ligne, 13 400 nouveaux sites web, 12 000 applications téléchargées sur l'App Store, etc. En 2013, 5exaoctets de données sont engendrées en dix minutes seulement contre deux jours en 2011. Avec plus d'untiers de la population mondiale connectée aux réseaux, soit plus de 2 milliards d'internautes, la sommed'informations produites en 48 heures est égale à celle produite depuis le début de l'humanité. En 2020, elleatteindra 40 zettaoctets, soit 40 000 milliards de gigaoctets.

L'exploitation de ce flux grossissant de données stockées dans d'immenses fermes de serveurs à travers le monde constitue un enjeu majeur pour la connaissance et la compréhension des activités humaines, aussi bien d'ordre économique, social que politique. Le téléchargement de films ou de musiques, les échanges de messages ou de photos sur les réseaux sociaux, la géolocalisation, la navigation sur le Web, ainsi que les requêtes adressées aux moteurs de recherche constituent des données dont l'analyse se révèle riche d'enseignements pour de nombreux secteurs d'activités scientifiques ou commerciales. Les capacités de stockage informatique supérieures au téraoctet (To = 1 000 Go) et les puissances de calcul utilisées communément par les géants d'Internet comme Google dépassant le pétaoctet (1 000 To) rendent d'ores et déjà possible le brassage d'un nombre infini de données mises en ligne afin d'en tirer des renseignements utiles aux stratégies des entreprises. Installées dans le monde entier, les fermes de serveurs détenues par le groupe Amazon hébergent plus de 1 000 milliards de documents, et plus de 500 millions de particuliers à travers le monde avaient recours à un service d'informatique en nuage en 2012.

## L'échelle des octets

Octet (1 o)	Unité de base (série de huit chiffres 0 ou 1)
Kilooctet (1 Ko = 1 000 o)	Une page de texte = 30 Ko
Mégaoctet (1 Mo = 1 000 Ko)	Un morceau de musique = 5 Mo
Gigaoctet (1 Go = 1 000 Mo)	Un film de deux heures = 1 Go
Téraoctet (1 To = 1 000 Go)	Six millions de livres = 1 To
Pétaoctet (1 Po = 1 000 To)	Une pile de DVD de la hauteur de la tour Montparnasse = 1 Po
Exaoctet (1 Eo = 1 000 Po)	Toutes les informations produites depuis les premières mesures jusqu'en 2003 = 5 Eo
Zettaoctet (1 Zo = 1 000 Eo)	La totalité des données enregistrées en 2011 = 1,8 Zo
Yottaoctet (1 Yo = 1 000 Zo)	La capacité du nouveau data center de la NSA prévu en 2013 = 1 Yo

Source : CNRS, d'après *Les Echos*, 10 décembre 2012.

Le *big data* constitue un nouveau marché de l'information sur lequel opèrent des experts en algorithmes, ainsi que des entreprises qui se sont faites comme spécialité de conférer du sens au traitement de milliards de fichiers numériques. Selon ces professionnels des *datas*, l'exploitation des bases de données internet, assortie de l'*open data* promu par les pouvoirs publics (data.gov depuis 2009 aux Etats-Unis et data.gouv.fr

---

depuis fin 2011 en France), est un facteur majeur d'innovation, contribuant à améliorer la productivité des entreprises. Selon le cabinet d'études Gartner, *le big data* pourrait améliorer de 20 % leur performance d'ici à 2015, autour de 6 % estime le MIT, ou encore permettre aux administrations européennes d'économiser 250 milliards d'euros par an selon McKinsey qui évalue le gain de productivité à 0,7 % dans le secteur de la santé, 0,5 % dans l'administration et de 0,5 % à 1 % dans le commerce. Au-delà des applications marketing permettant un profilage de plus en plus précis des consommateurs, le *big data* est déjà utile à de nombreux secteurs comme la santé, la sécurité ou les transports. Il est déjà possible d'utiliser les requêtes effectuées à partir de Google pour suivre la progression d'une épidémie de grippe ou encore de prévoir les embouteillages routiers en se basant sur les données fournies par les GPS des automobilistes. La police de Memphis dans l'Etat du Tennessee, aux Etats-Unis, se sert d'un programme développé par IBM qui analyse les interactions entre les jours de paye, la population par quartier et le calendrier des manifestations sportives, afin de déterminer les zones potentiellement à risque en termes de criminalité. En septembre 2013, la National Security Agency (NSA), chargée du renseignement informatique aux Etats-Unis, inaugurerait le plus grand centre de traitement de données au monde, capable de traiter plus d'un yottaoctet ( $Y_0 = 1\,000\,Z_0$ ) de données, dans le but d'archiver toutes les communications à l'échelle de la planète.

Au rythme où vont les choses, les bases de données classiques construites sur l'indexation de textes ou de chiffres pour répondre à des questions prédéterminées appartiendront bientôt à l'histoire de l'informatique. Des modèles de programmation informatique permettent désormais d'assurer le traitement de très grandes quantités de données (vidéos, sons, images, voix...) réparties sur des grappes de serveurs (*clusters*), tel MapReduce de Google ou Hadoop, projet open source développé par la Fondation Apache et utilisé par la plupart des grands groupes internet, notamment Facebook, Yahoo!, Twitter et Microsoft. Afin de vérifier sa e-réputation, le groupe Disney a passé au crible tous les blogs, forums et réseaux sociaux de la Toile avec Hadoop. Louant des capacités de mémoire et de calcul auprès des spécialistes de l'informatique en nuage, comme IBM ou Amazon, de nombreuses start-up prestataires de services spécialisés dans l'analyse des données naissent sur le marché du *big data*. Mesagraph propose aux chaînes de télévision des études d'audience d'un genre nouveau, basées sur les échanges effectués sur Twitter, restituant sous forme de graphiques le comportement des téléspectateurs sur le réseau social lié à la diffusion d'un programme. L'algorithme inventé par Tinyclues sert à déterminer le sexe et l'âge d'un internaute à partir de son adresse électronique, avec une marge d'erreur ne dépassant pas les 15 %. MFG Labs scrute les sites de partage de photos afin de déterminer les lieux de vacances privilégiés des touristes. Semicast surveille l'image des marques à partir des conversations en temps réel sur Twitter.

Pour les sciences sociales, le *big data* renouvelle les méthodes d'enquête, permettant de procéder par induction, mettant en lumière des causalités à partir de l'analyse des corrélations existantes entre des masses de données. Ignorant l'individu, le *big data* dresse des profils statistiques à son insu. Selon les chercheurs, le risque existe alors de privilégier les corrélations créées de toute pièce par les algorithmes au détriment du travail par vérification d'hypothèses préalablement définies.

Au fur et à mesure que nos vies se numérisent, se pose la question de la protection des données privées contre le piratage, dont furent victimes, en 2011, plus de 100 millions de titulaires d'un compte Sony

---

(PlayStation Network, Sony Online Entertainment et Qriocity). Mais se pose aussi la question de la commercialisation légale des données privées. La loi française LOPPSI 2 sur la sécurité intérieure autorise le ministère de l'intérieur à revendre à des tiers les informations personnelles des détenteurs d'une carte grise obtenue depuis août 2011, notamment à des fins de prospection commerciale. Les règles applicables au respect de la vie privée, à la protection des libertés individuelles et à l'éthique en matière d'exploitation des données personnelles, ainsi que les normes de sécurité contre le piratage ou la perte, se trouvent régulièrement dépassées par des techniques nées avec Internet en perpétuelle évolution et imprévisibles. En outre, les données archivées dans des fermes de serveurs réparties dans le monde entier posent le problème crucial de la « *souveraineté numérique* », notamment au regard du *Patriot Act* américain. Cette loi antiterroriste, votée à la suite des attentats du 11 septembre 2001, autorise les autorités américaines à accéder aux données numériques européennes hébergées par des infrastructures appartenant à des entreprises américaines, même si celles-ci sont installées en Europe. Face à la domination des entreprises américaines sur le marché de l'informatique en nuage, les Etats européens, notamment l'Allemagne, le Royaume-Uni et la France (projet Andromède), ont entrepris de construire leur nuage souverain pour héberger leurs données nationales (administrations, défense, nucléaire...).

Les données liées aux traces laissées par les internautes sont le nouvel or noir de la société de l'information, et la valorisation de cette matière première numérique est principalement l'affaire des groupes américains, Google, Apple, Facebook ou Amazon, qui règnent sur Internet. Après un an et demi d'enquête, la Federal Trade Commission a d'ailleurs annoncé, le 3 janvier 2013, qu'elle abandonnait les poursuites pour abus de position dominante à l'encontre de Google, accusé de privilégier ses propres services aux dépens de ceux de ses concurrents dans les réponses apportées aux requêtes des internautes sur son moteur de recherche. Les conclusions de l'enquête similaire en cours menée par les services de la concurrence de la Commission européenne seront-elles moins clémentes, au regard des amendes substantielles infligées antérieurement aux géants Microsoft et Intel ? En janvier 2013, une mission lancée par le gouvernement français propose de soumettre les géants américains d'Internet à une taxe fiscale portant sur la collecte, la gestion et l'exploitation commerciale de données personnelles des internautes localisés en France, en espérant les inciter ainsi à plus de transparence dans leurs pratiques (voir *supra*).

« *Désormais, nous devons faire confiance aux machines. Notre histoire appartient à Google, Facebook et Apple. Sans les armées de serveurs de ces nouveaux maîtres du monde numérisé, que restera-t-il de notre mémoire individuelle, collective, de nos rêves et réalisations dans dix ans, un siècle, un millénaire ? Rien.* » écrit Jean-Christophe Féraud dans *Libération* du 3 décembre 2012.

## Categorie

1. A retenir
2. Repères & tendances

## date créée

21 décembre 2012

## Auteur

françoise