

L'explosion mémorielle change la donne

On peut aujourd'hui facilement stocker d'énormes quantités d'information. Cette capacité nouvelle n'est pas qu'un changement quantitatif : elle annonce des transformations profondes dans la société et ses activités.

Gilles Dowek

Au cours des dernières années, les capacités de production et de stockage de l'information ont énormément augmenté, menant à un « déluge de données » (voir l'article de S. Abiteboul et P. Senellart dans ce numéro). Cette croissance s'accompagne de nombreux défis scientifiques et techniques : il faut notamment améliorer les performances des matériels et des logiciels, garantir la sécurité et la pérennité des données, concevoir des méthodes pour les exploiter, etc. Ces questions scientifiques et techniques sont intimement liées à celle de l'impact de cette augmentation sur les individus et les sociétés.

Le déluge de données est d'abord un changement de nature quantitative. Nous pouvions naguère collecter, stocker, transmettre et transformer de petits volumes d'information ; nous pouvons désormais faire de même avec des volumes d'information plus grands. Toutefois, comme souvent, un changement quantitatif annonce un changement qualitatif. La lunette de Galilée n'était qu'un œil qui permettait de voir des objets plus petits

L'ESSENTIEL

- Un disque de quelques centimètres, qui coûte moins de 100 euros, peut contenir aujourd'hui l'équivalent d'une bibliothèque nationale.
- De telles capacités de stockage, par les possibilités qu'elles offrent, bouleversent l'activité scientifique et nécessitent des développements informatiques.
- Elles modifient aussi notre rapport au passé et ouvrent des perspectives dans de nombreux domaines tels que la politique, l'économie ou l'éducation.

© Roman Gordeev, Zaslavsky Vectors/Shutterstock.com

et plus lointains ; mais la découverte de ces objets a révolutionné l'astronomie et, au-delà, la physique. De façon comparable, l'augmentation des capacités des mémoires numériques transforme le champ des problèmes scientifiques, les méthodes de la recherche et la forme même de nos connaissances.

La portée de ces changements ne concerne pas uniquement la science : notre rapport au temps, nos institutions politiques, nos économies, nos systèmes éducatifs, etc. se trouvent bouleversés et confrontés à de nouveaux défis.

Pour mieux comprendre ces transformations, commençons par donner quelques ordres de grandeur. Les quantités d'information s'expriment, comme toute autre grandeur, dans une certaine unité. L'unité fondamentale est le bit, qui correspond à un choix binaire (0 ou 1), mais on utilise souvent une unité dérivée, l'octet, égal à huit bits et qui correspond donc à un choix parmi $2^8 = 256$ possibilités.

Un octet est, *grosso modo*, la quantité d'information contenue dans une lettre de notre alphabet qui, en comptant les



1. DE PETITS DISQUES DE SAUVEGARDE ayant une capacité de un téraoctet sont désormais courants et sont vendus pour quelques dizaines d'euros. Or un téraoctet correspond à peu près à l'information contenue dans une bibliothèque d'un million de livres de 1 000 pages chacun !

majuscules, les minuscules, les lettres accentuées, les chiffres, les ligatures, les signes de ponctuation, etc., en contient un peu plus d'une centaine.

Un kilooctet, soit 1 000 octets, est à peu près la quantité d'information contenue dans une page de livre. Un mégaoctet, soit 1 000 kilooctets, correspond à l'information contenue dans un livre d'environ 1 000 pages. Un gigaoctet, égal à 1 000 mégaoctets, est *grosso modo* la quantité d'information contenue dans une bibliothèque personnelle de 1 000 volumes. Un téraoctet, soit 1 000 gigaoctets, est approximativement la quantité d'information contenue dans la collection de livres et imprimés de la Bibliothèque nationale de France, qui contient quelques millions de volumes. Le multiple suivant est le pétaoctet, qui vaut 1 000 téraoctets. Certains suggèrent qu'un pétaoctet correspond à la quantité d'information contenue dans l'ensemble des cultures humaines, même si cette quantité est difficile à évaluer.

Bien entendu, dans cette échelle – lettre (octet), page (kilooctet), livre (mégaoc-tet), bibliothèque personnelle (gigaoc-

tet), bibliothèque nationale (téraoctet), ensemble des cultures (pétaoctet) –, ce sont les ordres de grandeur et non les détails qui comptent. Par exemple, les images et les vidéos sont plus denses en information que les textes – une heure de vidéo contient ainsi une dizaine de gigaoctets, alors qu'une heure de lecture (quelques dizaines de pages) contient quelques dizaines de kilooctets. Et la collection de livres et imprimés de la Bibliothèque nationale de France comporte en réalité 14 millions de volumes, et non un million.

Cependant, cette échelle aide à comprendre les transformations que nous sommes en train de vivre, depuis qu'un disque permettant de stocker un téraoctet ne coûte que quelques dizaines d'euros. Ces changements concernent toutes nos activités qui consistent à collecter, stocker, transmettre et transformer de l'information... c'est-à-dire une grande partie de nos activités !

Commençons par les sciences, qui ont toujours comporté une part de collecte d'information. Par exemple, l'observation

des satellites de Jupiter par Galilée ou les observations recueillies par Darwin lors de son voyage à bord du *Beagle* ont joué un rôle essentiel dans la construction de leurs théories. Mais la quantité d'information collectée par les scientifiques d'aujourd'hui est d'un tout autre ordre de grandeur.

Un impact sur les sciences

Le résultat du séquençage d'un génome humain, qui comporte 3,4 milliards de paires de bases, équivaut ainsi à un texte dont la taille est de l'ordre d'un gigaoc-tet, soit la taille d'une bibliothèque de 1 000 volumes (*voir l'article de G. Perrière dans ce numéro*). Quant au LHC, le grand collisionneur de particules du CERN, il produit *via* ses détecteurs de l'ordre de 15 pétaoctets de données chaque année, soit plusieurs milliers de fois la taille de la collection de livres et imprimés de la Bibliothèque nationale de France (*voir l'article de F. Malek dans ce numéro*). Et les astronomes collectent des volumes encore

plus astronomiques de données (voir l'article de C. Reylé dans ce numéro).

Parce qu'ils demandent de brasser de telles quantités d'information, des projets tels que le séquençage du génome ou la recherche du boson de Higgs n'étaient pas à la portée des scientifiques avant la fin du XX^e siècle. Ce sont donc des domaines de recherche entiers qui s'ouvrent, depuis peu, devant nous.

Si cette transformation touche la nature des problèmes que les scientifiques peuvent se poser et tenter de résoudre, elle a également un impact sur la démarche scientifique et la forme même de nos connaissances.

La démarche d'abord : naguère, les scientifiques notaient leurs observations dans des « cahiers de manip » (voir la figure 2) et s'inspiraient de ces données pour imaginer des théories – c'est ce que l'on appelle l'induction – ou pour confronter les théories imaginées aux observations, afin de réfuter celles qui n'étaient pas en adéquation avec ces dernières. L'importance relative de l'induction et de la réfutation est, aujourd'hui encore, une question débattue. Quoi qu'il en soit, avec des « cahiers de manip » dont la taille se mesure en gigaoctets ou en téraoctets, il est devenu impossible tant de confronter « à la main » les théories aux observations que de s'inspirer de ces observations pour imaginer des théories.

La confrontation des théories aux observations demande un traitement

■ L'AUTEUR



Gilles DOWEK est chercheur en informatique à l'Inria (anciennement nommé Institut national de recherche en informatique et automatique), au sein de l'équipe *Deducteam*.

L'échelle des octets

Octet : unité égale à huit bits, pouvant coder $2^8 = 256$ caractères différents.

Kilooctet (ko) : 1 000 octets (environ une page d'un livre).

Mégaoctet (Mo) : Un million d'octets (un livre de 1 000 pages environ).

Gigaoctet (Go) : Un milliard d'octets (une bibliothèque de un millier de livres).

Téraoctet (To) : 10^{12} octets (une grande bibliothèque nationale).

Pétaoctet (Po) : 10^{15} octets (l'ensemble des cultures humaines ?).

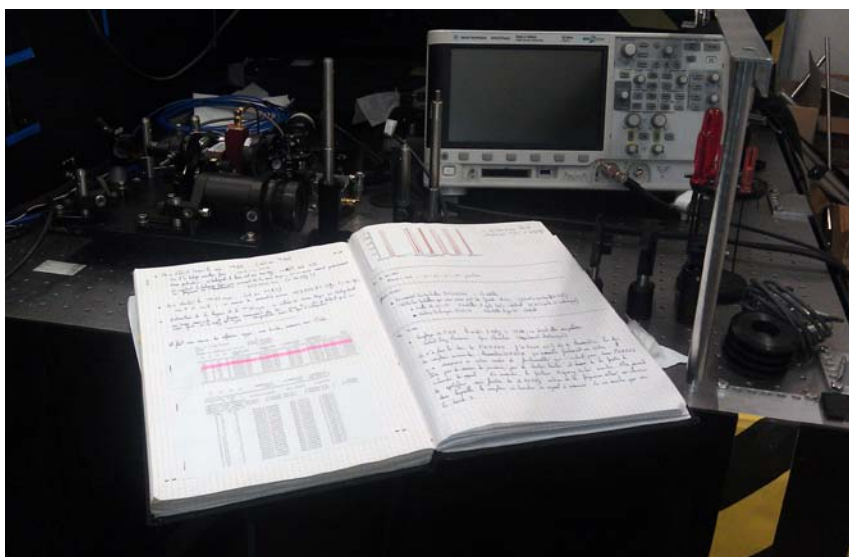
informatique aussi bien des données observées que de la théorie elle-même, qui doit être exprimée sous une forme prédictive, c'est-à-dire algorithmique, pour qu'elle puisse être automatiquement confrontée aux données observées. Et l'utilisation des observations pour imaginer des théories demande un traitement informatique plus difficile encore, appelé « apprentissage automatique ».

Une question typique en apprentissage automatique est de savoir si les lois de Kepler (qui décrivent les trajectoires des planètes autour du Soleil) peuvent être découvertes par un algorithme à partir des observations de Tycho Brahe, que Kepler a utilisées pour concevoir ses lois. Une autre est de savoir s'il est possible, à partir de grands corpus de textes, d'inférer automatiquement des règles de grammaire ou des règles de traduction d'une expression d'une langue en une autre.

En extrapolant – peut-être dangereusement – les évolutions actuelles de la méthode scientifique, nous allons peut-être vers un temps où les observations seront tellement volumineuses que la construction de théories à partir des observations sera une tâche laissée exclusivement aux ordinateurs, le rôle des scientifiques devant alors de concevoir des algorithmes permettant de construire des théories à partir des observations.

Une transformation de la connaissance ?

Cette transformation progressive de la démarche scientifique en induit une autre sur la connaissance scientifique elle-même. Jusqu'à la fin du XX^e siècle, les théories scientifiques étaient exprimées par des propositions de petite taille. Par exemple, la deuxième loi de Newton (selon laquelle la force s'exerçant sur un objet est égale au produit de sa masse par son accélération) s'exprime par une proposition d'une dizaine de caractères : $\vec{F} = m d^2 \vec{x} / dt^2$. Mais rien ne garantit que les théories automatiquement inférées à partir de gigantesques volumes d'observations soient petites. De fait, nous commençons à prendre conscience qu'il n'y a aucune raison *a priori* pour que les lois de la nature ou les règles de la grammaire d'une langue s'expriment par de courts énoncés.



2. UN CAHIER DE LABORATOIRE, où l'on consigne les résultats des expériences. De plus en plus, ces cahiers prennent une forme électronique. Certains d'entre eux contiennent des téraoctets de données, voire plus ; leur exploitation ne peut se faire qu'avec des moyens informatiques.