

- > 1 000 kilooctets, correspond à l'information contenue dans un livre d'environ 1 000 pages. Un gigaoctet, égal à 1 000 mégaoctets, est approximativement la quantité d'information contenue dans une bibliothèque personnelle de 1 000 volumes. Un téraoctet, soit 1 000 gigaoctets, est environ 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 (mégaoctet), bibliothèque personnelle (gigaoctet), 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!

VERS L'ORDINATEUR THÉORICIEN

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.

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 gigaoctet, soit la taille d'une bibliothèque de 1 000 volumes. Quant au LHC, le grand collisionneur de particules du CERN, il produit *via* ses détecteurs plusieurs 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. Et les

astronomes collectent des volumes de données encore plus... astronomiques.

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, à 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 page ci-contre) 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 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 devenant alors de concevoir des algorithmes qui permettraient de construire des théories à partir des observations.

L'échelle des octets

OCTET: unité égale à huit bits, codant $2^8 = 256$ caractères.

KILOOCTET (Ko): 10^3 octets (environ une page d'un livre).

MÉGAOCTET (Mo): 10^6 octets (un livre de 1 000 pages environ).

GIGAOCTET (Go): 10^9 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?).