

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.

Kiloctet (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 ?).

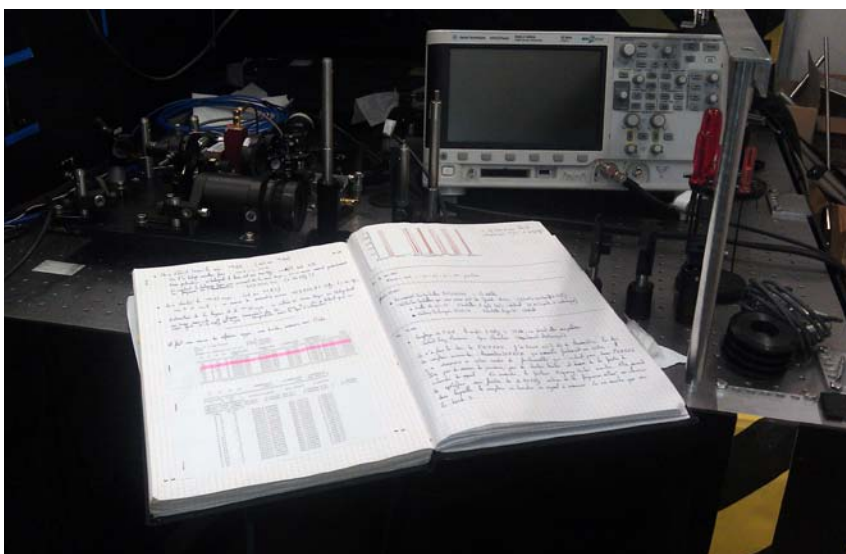
informatique aussi bien des données observées que de la théorie elle-même, qui doit être exprimée sous une forme prédictive, c'est-à-dire algorithmique, pour qu'elle puisse être automatiquement confrontée aux données observées. Et l'utilisation des observations pour imaginer des théories demande un traitement 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 devenant 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 : $F = m d^2 \vec{x} / dt^2$. 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.

La grammaire d'une langue, par exemple, contient sans doute des milliers de règles dont certaines ne portent que sur des expressions très particulières. En français, on peut dire «je viendrai lundi 2 septembre» ou «je viendrai le lundi 2 septembre»; on peut aussi dire «je viendrai demain», mais pas «je viendrai le demain». Il semble ainsi y avoir des règles de grammaire locales propres aux dates, qui diffèrent des règles de grammaire relatives aux adverbes de temps. Et si nous mettons bout à bout toutes ces règles de grammaire locales propres aux dates, aux noms de pays et autres toponymes (ainsi, on va «en Russie», mais «au Brésil» et «à Madagascar», et on prend le bateau pour aller «à Noirmoutier», mais «à l'île de Ré» et non «à Ré»), on obtiendra probablement des connaissances qui ne peuvent pas s'exprimer, comme la deuxième loi de Newton ou les équations de Maxwell régissant les champs électromagnétiques, en quelques dizaines de caractères (*voir l'encadré ci-contre*).

Le rôle des scientifiques est appelé à changer

En fait, cette idée est légèrement antérieure au déluge de données. Elle est sans doute apparue quand nous avons pris conscience que les lois qui régissent l'évolution de l'atmosphère ou le fonctionnement d'une cellule, par exemple, ne pouvaient pas s'exprimer en quelques lignes.

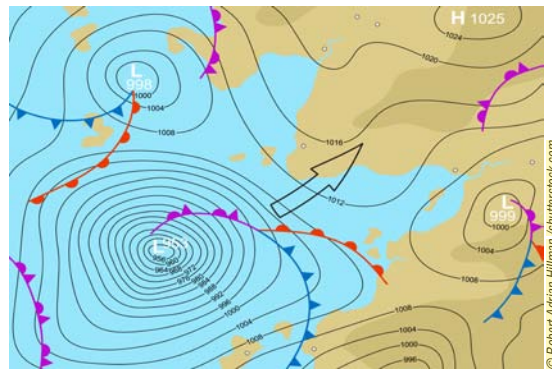
Lorsqu'une théorie s'exprime par des énoncés ou des algorithmes qui font des centaines de pages, le lien entre prévision et explication se distend. Ainsi, nous avons pris l'habitude d'accepter que les prévisions météorologiques indiquent le temps du lendemain avec une probabilité de succès assez élevée, mais qu'elles ne nous expliquent pas pour quelles raisons ce lendemain sera ensoleillé ou pluvieux. Les lois de Newton, au contraire, nous permettent non seulement de prédire que les trajectoires des planètes sont des ellipses, mais aussi de comprendre pourquoi.

Si nous ne pouvons pas expliquer le temps du lendemain, c'est précisément parce que la situation météorologique n'est pas le résultat d'une cause unique, comme le mouvement d'une planète est le résultat de l'attraction d'une étoile, mais de milliers de causes particulières desquelles il est difficile de faire émerger une explication globale.

DES THÉORIES SCIENTIFIQUES QUI S'ALLONGENT ?

Les phénomènes électriques et magnétiques, telles des étincelles (*ci-dessous à gauche*), sont bien décrits à partir de la théorie électromagnétique élaborée par James Clerk Maxwell vers 1870. Cette théorie s'exprime sous forme compacte par quatre équations vectorielles (les «équations de Maxwell») mettant en jeu le champ électrique $\vec{E}$ et le champ magnétique $\vec{B}$. En revanche, les prévisions météorologiques (*ci-dessous à droite*) résultent de modélisations théoriques et de calculs qu'il est impossible d'exprimer sous forme courte. Avec les données massives, les situations de ce type seront probablement de plus en plus fréquentes.

$$\begin{aligned}\nabla \cdot \vec{E} &= \frac{\rho}{\epsilon_0} \\ \nabla \cdot \vec{B} &= 0 \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t} \\ \nabla \times \vec{B} &= \mu_0 \vec{J} + \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t}\end{aligned}$$



Comme les données, ces théories sont trop complexes pour être utilisées «à la main»: seul un ordinateur peut les exploiter pour faire des prédictions. En extrapolant – à nouveau, peut-être dangereusement – les évolutions de la forme des théories scientifiques, nous pouvons imaginer que l'élaboration des prédictions sera un jour entièrement déléguée aux ordinateurs, et que les scientifiques auront alors pour rôle de concevoir des algorithmes permettant aux ordinateurs de calculer des prédictions à partir d'une théorie.

Il est cependant essentiel de garder en tête que, comme les autres théories scientifiques, ces théories restent des hypothèses, toujours à la merci d'une réfutation par les observations, et que leur pertinence ne tient qu'à leur capacité à résister à ces réfutations potentielles. Le risque est grand de prendre pour des oracles des prédictions produites par des ordinateurs à partir de théories trop complexes pour être comprises. Par exemple, contrairement à ce que l'on entend parfois, ce n'est pas le fait que tous les modèles climatiques s'accordent dans leurs prédictions qui importe, mais le fait que les prédictions passées aient été en accord, ou non, avec les observations.

Comme souvent, ces transformations de la technique et de la science en annoncent d'autres, dans divers autres domaines de notre culture. Un premier changement,

conséquence de cette capacité nouvelle de collecter, stocker, transformer et transmettre de grandes quantités d'information, est celui de notre rapport à la mémoire et au passé.

La plupart des quelque 100 milliards de femmes et d'hommes qui ont vécu avant nous n'ont laissé aucune trace de leur passage sur Terre. Par exemple, malgré l'écriture et l'imprimerie, peu d'entre nous peuvent citer le nom et le prénom de leurs 16 arrière-arrière-grands-parents. Et les plus anciennes personnes dont nous connaissons le nom, sans doute quelques souverains de Mésopotamie ou d'Égypte, remontent à un passé relativement récent. De l'histoire de l'humanité durant les centaines de milliers d'années qui les ont précédés, depuis l'apparition des premiers *Homo sapiens*, nous n'avons presque aucune trace.

L'humanité devient hypermnésique

Les femmes et les hommes des siècles prochains auront une vision bien différente de leur passé, car les capacités actuelles de stockage de l'information rendent l'humanité hypermnésique. Ainsi, nos descendants connaîtront mieux que nous leur arbre généalogique; et au lieu d'un nom laissé, pour toute trace, dans un registre d'état civil, ils pourront conserver sur chacun de leurs lointains ancêtres un grand nombre d'informations.