PAR
ALEXEI GRINBAUM
(Direction de la recherche fondamentale)

Alexei Grinbaum est physicien et philosophe. Il travaille au Laboratoire de recherche sur les sciences de la matière (Institut de recherches sur les lois fondamentales de l'Univers du CEA).

DÉFINITION

Big Data : de quoi parle-t-on ?

C'est dans les années 1990 que le terme Big Data prend sa signification actuelle : un défi technologique à relever pour analyser de grands ensembles de données, d'abord scientifiques, mais de plus en plus souvent collectés au quotidien par divers moyens techniques. Big Data désigne à la fois la production de données massives et le développement de technologies capables de les traiter afin d'en extraire des corrélations ou du sens. Définition en sept étapes…

OCTET

Unité de mesure de la quantité de données pouvant être produites ou stockées.
Un kilooctet (Ko) correspond à mille octets (quelques Ko, c'est le poids d'un simple fichier texte), un mégaoctet (Mo) à un million d'octets (un CD-Rom fait 650 Mo), un gigaoctet (Go) à un milliard d'octets (la taille d'une clef USB varie usuellement de 1 à 8 Go, certaines allant jusqu'à 128 voire 256 Go) et un teraoctet (To) à mille milliards d'octets, soit la capacité de stockage d'un disque dur performant.

ALGORITHME

Description, traduisible sous forme d'un programme dans un langage informatique, d'une suite finie d'étapes à exécuter pour obtenir, à partir de données en entrée, des données en sortie en vue d'un objectif prédéterminé.

CALCUL HAUTE PERFORMANCE

Représenter virtuellement des objets, des phénomènes ou des systèmes particulièrement complexes nécessite d'utiliser des calculateurs extrêmement puissants (les supercalculateurs). Aujourd'hui, les plus performants sont capables de réaliser plusieurs millions de milliards d'opérations à la seconde (petaflop/s).
D'où le terme de calcul haute performance (ou HPC pour High Performance Computing) qui désigne également, par extension, la science développée autour de ces équipements (matériels, logiciels etc.).

VOLUME

Qui dit données massives dit volumes allant du kilooctet au petaoctet, dépassant toute capacité de traitement rapide par le cerveau humain.

VÉLOCITÉ

Fréquence à laquelle les données sont générées, traitées et mises en réseau. Cette fréquence étant de plus en plus élevée, il est très souvent nécessaire d'employer les ressources du calcul haute performance (extreme computing). Climatologues [voir page 27], astrophysiciens [voir page 32] comme spécialistes en génomique [voir page 33] en sont de fervents utilisateurs.

VARIÉTÉ

Les données peuvent être textuelles, visuelles ou sonores, scientifiques ou provenant de la vie courante, structurées ou non. D'où la nécessité de les analyser automatiquement par des algorithmes pour en extraire des corrélations et des connaissances (data mining) et, quelquefois, de les représenter sous forme visuelle (data visualisation).

CORRÉLATION

L'analyse de données permet de dégager des corrélations souvent insoupçonnées et instructives (data analytics). Cependant, l'existence de corrélations ne signifie pas la réalité des liens de cause à effet entre leurs référents. Et une corrélation n'équivaut pas une signification ou une connaissance. La tension fondamentale entre une science fondée sur la causalité et une analyse qui s'appuie sur les corrélations est au centre des débats épistémologiques actuels [voir page 4].

BIAIS

Certaines données peuvent contenir des biais ou être discriminatoires. Leur traitement automatique transmettra ces biais aux conclusions qui en seront tirées. L'éthique du Big Data cherche à en éviter les conséquences néfastes en préconisant des procédures de contrôle et vérification des données.

Rapport Stratégie France IA : www.enseignementsup-recherche.gouv.fr/cid114739/rapport-strategie-france-i.a.-pour-le-developpement-des-technologies-d-intelligence-artificielle.html

TRAÇABILITÉ

Il doit être possible de suivre les actions d'un système qui apprend en analysant les données (machine learning) par la mise à disposition d'un journal suffisamment détaillé. C'est même essentiel pour déterminer les responsabilités et fonder, le cas échéant, un recours juridique.

Initiative IEEE : www.standards.ieee.org/develop/indconn/ec/autonomous_systems.html

EXPLICABILITÉ

Dans certains cas, le machine learning inventera et utilisera des repères ou des concepts qui lui sont propres, et dont l'humain ne comprendrait pas nécessairement la signification. Le compromis entre la performance de l'apprentissage et l'explicabilité doit être apprécié en fonction de l'usage.

Travaux de la Cerna : www.cerna-ethics-allistene.org