1 • VINCENT BLONDEL : « Nous étudions de nouveaux objets scientifiques »

Entretien L'accroissement rapide du volume des données numériques enregistrées promet une compréhension inédite des comportements sociaux. Mais il nécessite de nouvelles méthodes d'analyse.

Vincent Blondel
est professeur à l'université catholique de Louvain, en Belgique, et professeur invité au Massachusetts Institute of Technology, aux États-Unis.

LA RECHERCHE : Comment définissez-vous les Big Data?

VINCENT BLONDEL : Le premier critère est le volume, sous-entendu par le mot « big ». Le domaine des Big Data s'intéresse à des ensembles de données digitales qui, de par leur taille, ne peuvent être traitées avec des méthodes traditionnelles ; en fonction des applications, ce peut-être de l'ordre du gigabit*, du térabit* ou plus encore. Ensuite, ce volume ne cesse de croître à grande vitesse. On estime que le volume de données stockées dans le monde double tous les quatre ans. On a ainsi stocké plus de données depuis 2010 qu'on ne l'avait fait depuis les débuts de l'humanité ! Le troisième critère, c'est la grande diversité des données auxquelles on s'intéresse. Ce peut être la consommation d'électricité dans tous les quartiers de France à tout moment, les 30 milliards de « j'aime » journaliers sur Facebook ou les 5 000 photographies déposées chaque minute sur le site de partage Flickr [1]. Enfin, on s'attache à la « véracité » : les données recueillies sont souvent bruitées et imprécises et doivent être traitées pour en extraire de l'information utile.

*LE BIT est une unité de mesure en informatique désignant la quantité élémentaire d'information ; 1 gigabit = 10^9 bits 1 térabit = 10^{12} bits.

L'essentiel

- LES BIG DATA sont caractérisées par le volume, la vitesse d'accumulation, la variété et la véracité des données numériques.
- IL FAUT METTRE au point des méthodes de traitement dont le nombre d'opérations n'augmente pas trop vite avec la quantité de données.
- CE DOMAINE fournit aussi de nouveau outils pour faire de la science, notamment des sciences sociales.

En quoi ces traitements diffèrent-ils de ceux que l'on réalise déjà en informatique ?

V.B. Ce qui nous a fait entrer dans l'ère des Big Data, c'est l'explosion des capacités de stockage. Un petit disque dur de l'épaisseur d'un livre suffit par exemple à stocker les informations sur les communications téléphoniques belges d'une année. Et cela pour un prix très modeste. Il devient alors possible de s'interroger sur la façon de traiter ces données afin d'en tirer des informations utiles : cette capacité de stockage crée de nouveaux objets à étudier, et il nous faut imaginer comment le faire. Même si nous arrêtions aujourd'hui de recueillir des données, nous aurions besoin de plusieurs années de travail pour comprendre comment analyser tout ce que nous avons déjà enregistré. Mais les données continuent d'arriver, toujours plus vite !

Donc vous recherchez de nouvelles méthodes d'analyse ?

V.B. Exactement. Prenons l'exemple d'un réseau dans lequel des entités sont connectées les unes aux autres. Un problème classique et très général consiste à rechercher des « communautés » : des zones plus densément connectées que d'autres. C'est un problème bien défini mathématiquement, et nous avions depuis longtemps des méthodes pour le résoudre. Mais elles n'étaient pas assez efficaces : il aurait fallu des années pour traiter les énormes réseaux d'aujourd'hui, formés par les utilisateurs de Facebook, qui sont 1 milliard, ou les pages web reliées par des hyperliens, que l'on compte par dizaines de milliards. Désormais, de nouvelles méthodes permettent de résoudre rapidement ces problèmes à l'aide d'un simple ordinateur de bureau. L'efficacité, c'est-à-dire la vitesse de traitement, est aussi l'un des obstacles à surmonter quand il s'agit de détecter des corrélations dans des ensembles très grands ou d'identifier des événements anormaux dans des séries.


© VINCENT BLONDEL

Comment rendre ces méthodes plus efficaces ?

V.B. Il faut que le nombre d'opérations à réaliser, donc le temps nécessaire, n'augmente pas trop vite quand le volume des données s'amplifie [fig. 1]. Pour la détection de communautés, par exemple, ce nombre d'opérations croissait comme le carré de la quantité de données : pour un réseau 10 fois plus gros, il fallait 100 fois plus de temps. Par exemple, imaginons qu'une heure de calcul suffise pour mener une analyse sur les communications téléphoniques d'une seule journée à Bruxelles. Pour traiter les communications de toute la Belgique, cela prendra 100 heures. Et pour les communications de toute l'Europe, il faudra environ 250 000 heures, soit plus de vingt-huit ans. Ce n'est pas possible. Nous devons donc trouver des méthodes dont le temps de calcul croît moins vite avec la taille des données. Linéairement par exemple : le temps augmente seulement proportionnellement à la quantité de données. C'est le minimum si l'on veut lire toutes les données.

Peut-on néanmoins faire mieux ?

V.B. Oui, nous savons aujourd'hui analyser un ensemble de données sans les consulter toutes, en donnant néanmoins des garanties sur la fiabilité de la réponse. Voilà une problématique scientifique récente et typiquement Big Data. Par exemple, il y a quotidiennement des milliards de transactions avec des cartes de crédit. Un algorithme qui n'en analyse que 10 ou 100 millions pourra tout de même indiquer qu'aucune carte n'a eu un parcours correspondant à une usurpation d'identité. La réponse ne sera pas garantie à 100 %, parce que des comportements anormaux pourraient exister dans les données qui n'ont pas été analysées. Mais la probabilité qu'elle soit vraie sera quantifiée rigoureusement.

Cela ressemble à des sondages ?

V.B. En quelque sorte, mais ce type de méthode permet de répondre à des questions plus complexes que celles posées lors de sondages d'opinion. Par exemple, l'évolution au cours du temps des communautés qui structurent un réseau, ou la détermination qu'une entité a eu un parcours différent des autres. La théorie nous permet de déterminer la distribution de probabilités suivant laquelle il faut choisir les données à analyser pour optimiser la précision de la réponse. Elle nous permet aussi de donner des bornes mathématiques pour l'écart entre cette réponse et celle que l'on aurait obtenue si l'on avait examiné toutes les données. Bien entendu, tout cela repose sur des hypothèses en lien avec la structure de l'ensemble de données et dépend du problème particulier que l'on souhaite résoudre.

Deux auteurs ont affirmé récemment que les Big Data sont porteuses d'une révolution scientifique comparable à celle entraînée par l'invention du microscope [2]. Qu'en pensez-vous?

V.B. Les Big Data permettent effectivement de faire de la science de façon totalement >>>

« Nous étudions de nouveaux objets scientifiques »

Entretien avec Vincent Blondel

nouvelle, notamment pour l'étude de phénomènes sociaux. Par exemple, avec Samuel Martin et Corentin Vande Kerckhove, dans mon laboratoire, nous travaillons en psychologie sociale sur les dynamiques d'opinion : comment, dans un groupe, des personnes qui doivent faire un choix s'influencent-elles mutuellement ? Des modèles mathématiques ont été proposés, mais il faut les tester. Autrefois, nous aurions mené les expériences avec quelques dizaines de personnes. Aujourd'hui, grâce au « Turc mécanique » de la société Amazon, nous pouvons très simplement recruter plusieurs milliers de participants qui réaliseront l'expérience de chez eux, en échange d'une somme modique [3]. Nos résultats auront une autre portée !

Un autre exemple est lié au développement des « cours en ligne ouverts et massifs », les MOOC, selon l'acronyme anglais. Des universités proposent des cours en accès gratuit sur Internet. À l'université catholique de Louvain, nous travaillons en partenariat avec la plateforme internationale edX, fondée par l'université Harvard et le Massachusetts Institute of Technology, aux États-Unis, dont l'interface enregistre l'ensemble du parcours de formation de l'étudiant : à quels moments il se connecte, combien de temps il reste sur chaque page, son taux de réussite aux tests qui lui sont régulièrement proposés, éventuellement les questions qu'il pose au sein des forums mis en place, etc. Cela permet de faire des observations et des expérimentations pédagogiques à une échelle inaccessible jusqu'ici. En corrélant le comportement des étudiants avec leurs résultats et leur progression, on pourra comprendre les processus d'apprentissage mieux qu'on ne l'a jamais fait, déterminer si certains sont plus efficaces que d'autres et offrir un parcours personnalisé.

Finalement, ne renoncez-vous pas à établir des lois scientifiques explicatives au profit de simples corrélations, que seul l'ordinateur maîtrise ?

V.B. Si un algorithme peut vous dire avant votre médecin, et sans que l'on comprenne totalement pourquoi, que vous avez une probabilité élevée d'avoir un cancer, je ne vois pas pourquoi on s'en priverait. Ensuite, les Big Data touchent les sciences sociales dans lesquelles les chaînes causales d'explication sont moins claires qu'en physique ou en biologie. Enfin, les analyses de Big Data sont des outils qui ne se substituent pas à la compréhension des scientifiques : elles attirent l'attention sur des corrélations détectées afin que ces derniers recherchent ensuite des explications causales. Bien entendu, pour les entreprises qui s'intéressent seulement aux applications, pour mieux vendre leurs produits par exemple, les modèles explicatifs ne sont pas nécessaires. En science par contre, les Big Data peuvent bien être vues comme un outil, à l'image d'un microscope, pour faire progresser la connaissance

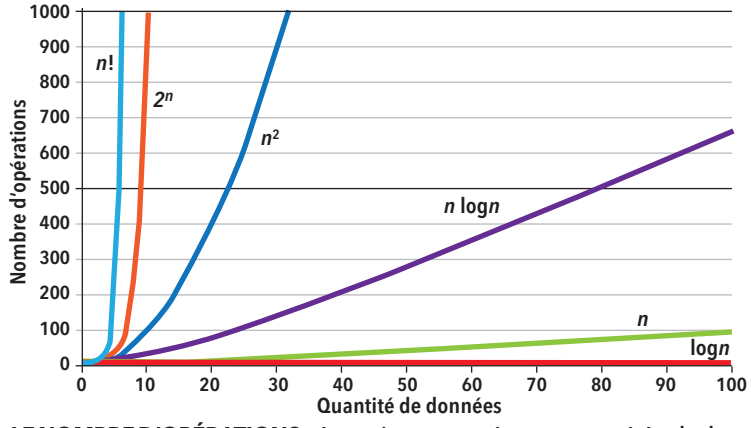
Y a-t-il une question sur laquelle les Big Data rencontrent des difficultés sérieuses aujourd'hui ?

V.B. La protection de la vie privée. Les Big Data promettent des bénéfices énormes pour la société, en faisant progresser la médecine personnalisée, la prédiction de la propagation de virus ou les modèles de croissance économique. Mais comme ces données sont souvent issues des comportements de chacun d'entre nous, il y a des risques d'intrusion, ce qui suscite des craintes. Il est de la responsabilité des scientifiques de contribuer à ces problématiques et d'aider les citoyens et les législateurs qui s'interrogent sur les limites à mettre, comme le font en ce moment ceux de l'Union européenne. Trouver la juste mesure ne doit pas être seulement du ressort de juristes ou de techniciens mais bien de toute la société. Ceux qui élaborent des modèles d'utilisation des données doivent aussi montrer scientifiquement l'intérêt de le faire, les difficultés qui se présentent lorsqu'on veut rendre des données anonymes et quantifier les dangers auxquels on s'expose en partageant des données.

Propos recueillis par Luc Allemand

[1] www.flickr.fr
[2] V. Mayer-Schöneberger et K. Cukier, *Big Data*, John Murray, 2013.
[3] www.mturk.com

Fig.1 La complexité des algorithmes



LE NOMBRE D'OPÉRATIONS nécessaires pour traiter une quantité *n* de données ne doit pas augmenter trop vite avec *n*. Des algorithmes acceptables pour peu de données (avec une variation en n^2 par exemple) prennent trop de temps dans le domaine des Big Data. Des méthodes d'échantillonnage, où toutes les données ne sont pas lues, permettent une variation plus faible que *n*.