Hors-Série N°2 - Avril 2015

FSMP
Fondation Sciences Mathématiques de Paris

MathsInfos

LE BIG DATA

*Tous les propos de ce hors-série ont été recueillis auprès de **Bertrand Michel** et d'**Agathe Guilloux**, enseignants-chercheurs à l'Université Pierre et Marie Curie, membres du **LSTA**, et de **Fabrice Rossi**, enseignant-chercheur à l'Université Paris 1 Panthéon-Sorbonne et membre de l'équipe de recherche **SAMM**.*

L'essor du Big Data

Collecter, stocker et traiter des données en très grand nombre n'a rien d'un phénomène nouveau. Des entreprises comme MasterCard et Orange, ou encore des centres de recherche tel le CERN dans le canton de Genève, recueillent et utilisent depuis longtemps des données massives et possèdent à cet effet des équipements aux capacités gigantesques. Ce qui est nouveau, en revanche, c'est l'augmentation considérable du nombre d'organismes ayant recours à des données massives. Désormais, celles-ci ne sont plus l'apanage de quelques structures géantes, mais concernent également des entreprises de moindre taille. L'essor du *cloud computing*, dans les années 2000, explique en partie ce phénomène. En effet, de nombreuses petites entreprises du web ont profité du fonctionnement commun de serveurs informatiques à distance pour accéder à une puissance de stockage et de calcul jusqu'alors inédite pour elles. Elles ont pu collecter des données de manière plus efficace : par exemple, via les jeux et les applications sur le réseau social Facebook, qui permettent de toucher un public conséquent. Grâce au *cloud*, ces petites entreprises du web sont donc facilement rentrées en contact avec un très grand nombre de clients et ont pu récolter des données de l'ordre du téraoctet.

Les mathématiques sollicitées

La récente médiatisation du Big Data révèle les possibilités qu'offre l'analyse des mégadonnées. Pour beaucoup d'entreprises, c'est le déclic : il est désormais possible d'extraire des renseignements de toutes les informations qu'elles accumulent depuis des années. Certaines d'entre elles n'hésitent pas alors à solliciter l'aide de laboratoires mathématiques, tels le LSTA et l'équipe SAMM, pour résoudre les problèmes complexes que pose le traitement de leurs données. Les entreprises du web ne sont pas les seules à se tourner vers les mathématiciens ; de manière générale, les demandes du secteur privé ont augmenté.
Ces sollicitations nouvelles amènent les mathématiciens à s'emparer des problématiques propres au Big Data. Ils s'attellent par exemple aux questions relatives à la protection de la vie privée, devenues incontournables (voir *Comment rendre les données anonymes* au verso). Dorénavant, mathématiques et Big Data sont intrinsèquement liés.

Le Laboratoire de Statistique Théorique et Appliquée (LSTA, Université Pierre et Marie Curie) et l'équipe Statistique, Analyse et Modélisation Multidisciplinaire (SAMM, Université Paris 1 Panthéon-Sorbonne) ont récemment rejoint la Fondation Sciences Mathématiques de Paris. Fortes au total d'une quarantaine d'enseignants-chercheurs et d'autant de doctorants, ces deux unités représentent un potentiel important pour la statistique parisienne.
Soucieux d'inscrire leurs activités dans les thématiques les plus en pointe, les deux groupes s'adaptent aux bouleversements induits par l'analyse statistique des mégadonnées (en anglais, *Big Data*). Ce terme générique fait référence au volume et à la diversité des données, désormais accessibles grâce au développement à grande échelle de l'informatique et du traitement numérique de l'information.
Les mégadonnées se singularisent par leur taille (un avion de ligne peut enregistrer jusqu'à 1 teraoctet - 10^{12} octets - de données par vol), leur variété (nombres, images, textes, etc.) et bien souvent leur flux en continu. Elles nécessitent de repenser certains fondamentaux de la statistique tout en proposant de nouvelles idées. Cette évolution est rendue possible par l'avènement de nouvelles technologies informatiques de stockage et de calcul, comme par exemple les architectures MapReduce et Hadoop. Egalement poussé par une très forte demande en matière d'emplois, ce mouvement s'accompagne d'un besoin de formation important de spécialistes du traitement des mégadonnées : les *data scientists*.
L'adaptation de nos laboratoires à ce nouveau paradigme s'effectue donc non seulement au niveau de la recherche et de l'enseignement, mais également au niveau des liens avec les autres organismes de recherche et les entreprises - les propositions de collaboration du secteur privé sont considérables. L'équipe SAMM et le LSTA sont prêts à relever le défi !

Jean-Marc Bardet (directeur de l'équipe SAMM) et Gérard Biau (directeur du LSTA)

L'équipe SAMM et le LSTA ont respectivement rejoint la Fondation Sciences Mathématiques de Paris les 1er mars 2014 et 1er janvier 2015.

Statistique, Analyse, Modélisation Multidisciplinaire
EA 4543
Université Paris 1 Panthéon-Sorbonne

Les domaines de recherche présents au sein du SAMM couvrent de nombreux champs des mathématiques appliquées (analyse fonctionnelle appliquée, apprentissage statistique, contrôle optimal, équations d'évolution, probabilités et statistique), et quelques thématiques en informatique (graphes, automates cellulaires).
Axes de recherche :
(a) Apprentissage Statistique et Réseaux
(b) Equations d'évolution
(c) Statistique

Laboratoire de Statistique Théorique et Appliquée
EA 3124
Université Pierre et Marie Curie

Comme l'indique le nom du laboratoire, la recherche effectuée en son sein couvre un très large éventail de la statistique, tant dans ses composantes théoriques que dans les applications. Elle est structurée en cinq thématiques majeures. Celles-ci regroupent, sans que ceci soit pour autant limitatif, les axes de recherche les plus actifs du LSTA :
(a) Apprentissage statistique et grande dimension
(b) Biostatistique
(c) Statistique mathématique
(d) Actuariat
(e) Statistiques industrielles

Hors-Série N°2 - Avril 2015

FSMP
Fondation Sciences Mathématiques de Paris

Le Big Data à la croisée des disciplines

D'emblée, le champ disciplinaire du Big Data semble difficile à circonscrire, tant le terme est aujourd'hui employé. Les mathématiciens du LSTA et de l'équipe SAMM reconnaissent volontiers qu'il s'agit là d'une tâche délicate. Les outils mathématiques qu'ils emploient relèvent en effet de divers domaines : statistique, apprentissage statistique (une discipline à cheval entre l'informatique et la statistique), optimisation convexe, probabilités, apprentissage automatique... Cette difficulté à définir ce que recouvre précisément le Big Data, aux yeux des mathématiciens et des informaticiens, se manifeste notamment lors de la mise en place de formations spécialisées : le choix des enseignements n'a alors rien d'une évidence. Déjà, reconnaître quelles données relèvent ou non du Big Data n'est pas si simple en soi. Les mathématiciens s'accordent toutefois à dire que le Big Data concerne des données de taille supérieure à mille téraoctets. Pour les traiter, il faut les répartir efficacement sur des centaines de cœurs de calcul, chacun d'entre eux étant capable à l'heure actuelle de traiter jusqu'à six milliards d'opérations à la seconde.

Les modèles et les outils mathématiques en jeu

Face aux mégadonnées, ce sont souvent aux techniques de l'***apprentissage automatique*** que les mathématiciens ont recours. L'objectif est alors d'établir un modèle mathématique qui soit le plus pertinent possible. Mais pour trouver ce modèle, il faut tout d'abord qu'il y ait une question préalable à laquelle tenter de répondre. Par exemple, face à un fichier contenant des millions de données du secteur automobile, les mathématiciens de l'équipe SAMM cherchent à déterminer quel est le meilleur prix possible pour chaque voiture d'occasion. Pour ce faire, ils s'attellent donc à créer un modèle, capable de donner ce prix.
Les outils mathématiques qu'ils utilisent relèvent de l'optimisation, des statistiques et de l'analyse. L'optimisation intervient pour déterminer le meilleur temps de calcul pour le modèle, afin que celui-ci donne le plus rapidement possible la réponse attendue. Les statistiques permettent, elles, d'estimer la qualité du modèle : savoir si celui-ci répond bien à la question posée. Enfin, l'analyse mathématique indique si le type de modèle retenu est le plus efficace. Etablir le modèle le plus pertinent, c'est alors trouver un équilibre entre les contraintes établies par l'optimisation, les statistiques et l'analyse.

Mathématiques en Mouvement

Chaque année, la **Fondation Sciences Mathématiques de Paris** organise une journée afin d'illustrer la prodigieuse diversité de la recherche mathématique. Plusieurs chercheurs y proposent de courts exposés, accessibles aux étudiants. Pour cette 7ème édition, qui se déroulera le **samedi 6 juin 2015** à l'Institut Henri Poincaré, c'est de Big Data dont il sera question. Programme et inscription (gratuite mais obligatoire) sur **www.sciencesmaths-paris.fr**.

Dans la presse

Le Big Data a les honneurs de la presse, parfois au prix de quelques confusions. Ainsi, l'analyse statistique des mégadonnées permet certes d'appréhender le comportement de tel ou tel groupe d'individus dans une situation bien définie (par exemple, savoir si les personnes qui ont acheté tel livre achèteront tel autre), mais elle ne s'avère plus guère pertinente à l'échelle individuelle, quand il s'agit de prévoir le comportement d'une personne en particulier.

L'aventurier du Big Data

Retrouvez sur **www.sciencesmaths-paris.fr** l'interview vidéo et le portrait de **David Bessis**, mathématicien et créateur de la startup tinyclues, spécialisée dans le Big Data.

Les algorithmes du Big Data


© agsandrew - Fotolia.com

Les données du Big Data proviennent de contextes très variés : elles sont par exemple issues du secteur médical, du web ou encore de nos téléphones portables (certaines applications enregistrent nos données de géolocalisation). Même si les idées et les méthodes sous-jacentes sont souvent similaires, chaque type de données nécessite des algorithmes spécifiques, le plus souvent proposés par la communauté informatique. Statisticiens et informaticiens travaillent alors sur les mêmes objets, mais leurs motivations ne sont pas strictement identiques. Pour schématiser, les informaticiens cherchent avant tout l'efficacité prédictive et algorithmique tandis que les mathématiciens sont souvent obligés de travailler sur des algorithmes simplifiés, mais dont l'analyse mathématique reste possible. Ces deux approches sont à l'évidence complémentaires ; un rapprochement toujours plus fort entre elles, en France, permet d'avancer dans le domaine de la science des données. Les algorithmes du *deep learning* illustrent cette situation. Ces dernières années, des améliorations significatives de ces algorithmes ont été proposées par la communauté du *machine learning*. Cependant, la complexité de ces méthodes rend encore difficile leur analyse mathématique.

Comment rendre les données anonymes

Les mathématiciens ont désormais une meilleure compréhension des algorithmes qu'ils développent, et peuvent envisager une répartition plus pertinente des mégadonnées sur plusieurs cœurs de calcul. De nombreux enjeux se profilent toutefois. L'un d'entre eux consiste à définir des stratégies efficaces d'anonymisation des données. C'est à cette tâche que s'attelle notamment le mathématicien **Michael I. Jordan (Université de Californie, Berkeley), Chaire Senior FSMP 2012**. Dans l'article *Privacy Aware Learning* dont il est co-auteur, il cherche à modifier des données personnelles - par exemple, en intervertissant les caractéristiques « homme » et « femme » -, de manière à ce qu'il ne soit plus possible d'identifier une personne en particulier à partir des éléments la concernant. Toute la difficulté réside dans le fait que cette perturbation doit remplir cet objectif d'anonymisation sans pour autant fausser l'exploitation des données.

Directeur de la publication : Jean Dolbeault
Responsable de la rédaction : Alice JACQUET - Contact : jacquet@fsmp.fr
Fondation Sciences Mathématiques de Paris - www.sciencesmaths-paris.fr
IHP, 11 rue Pierre et Marie Curie 75231 Paris Cedex 05
Tél. : +33 (0) 1 44 27 68 03 - Fax : +33 (0) 1 44 27 68 04

Fondateurs
Partenaires scientifiques
LabEx SMP
DIM RDM-ÎdF