Erwan Le Pennec

Professeur à l'Ecole polytechnique, Université Paris Saclay

Erwan Le Pennec, 40 ans. Je suis professeur associé au département de mathématiques appliquées de l'École polytechnique et j'effectue mes recherches en traitement du signal et en statistique au CMAP. Je suis le porteur de la chaire Data Scientist créée en octobre 2014 par l'École polytechnique, Keyrus, Orange et Thales et portée par la fondation de l'X.

Pascal Massart

Pascal Massart, 58 ans, Professeur à l'Université Paris-Sud depuis 1990. Je me suis intéressé très tôt aux connexions entre le calcul des probabilités dans les espaces de Banach et la statistique avec un goût particulier pour les résultats et inégalités non asymptotiques. Ce goût ne s'est jamais démenti avec un souci de plus en plus affirmé au fil des ans de confronter la théorie à la pratique, ce qui m'a conduit à m'investir dans l'élaboration de critères de choix de modèle entièrement dirigés par les données dont le fondement théorique repose sur la théorie de la concentration de la mesure. Nommé membre de l'IUF en 2010, je me suis investi depuis plusieurs années dans la construction d'un master de mathématiques à l'échelle de l'université Paris Saclay qui a vécu sa première rentrée en 2015.

Le « Big data » et les mathématiques

En l'espace de quelques années, le « big data » a envahi l'espace économique, scientifique et même médiatique. Big data par ci, big data par là, ce terme aiguise les appétits des uns et agace les autres, mais derrière l'effet de « buzz » et les inévitables mythes et fantasmes, il faut bien comprendre que se cachent à la fois une véritable réalité économique et une révolution scientifique.

Un déluge de données

Afin d'appréhender ce phénomène, il convient de partir d'un fait qui n'échappe à personne : les « données » sont aujourd'hui partout. Elles ont évidemment toujours existé au sein des entreprises aussi bien que dans les laboratoires scientifiques. L'élément nouveau n'est pas tant l'augmentation de leur volume que leur exploitations dans des contextes inattendus. Elles ne sont plus seulement générées volontairement en vue d'une exploitation économique bien ciblée ou issues d'une expérimentation scientifique destinée à valider une hypothèse bien identifiée. Récoltées par des capteurs multiples parfois même à l'échelle des individus via des objets connectés qui font partie de notre quotidien, elles ont à l'instar de la créature de Frankenstein une forte tendance à échapper à leur créateurs pour vivre une vie qui leur est propre.

C'est dans l'idée même qu'on puisse les exploiter à d'autres fins que celles pour lesquelles elles ont été conçues que se situe réellement le déclic révolutionnaire du « big data ». Le bras armé de cette idée réside évidemment dans les formidables progrès technologiques accomplis au cours de ces dernières années : explosion de la capacité de stockage des données et facilité de captation, de circulation et d'accès via internet. Parallèlement, ces progrès concernent également les capacités du calcul, exécuté à présent sur des machines de moins en moins coûteuses équipées de processeurs de plus en plus puissants qu'on peut qui plus est faire coopérer grâce au calcul distribué.

Les données, celles dont on dispose et celles qu'on pourrait songer à acquérir avec pour ambition de mieux comprendre celles qu'on possède déjà, ressemblent à un immense gisement dont l'œil peine à embrasser les contours. L'envie de les utiliser comme une ressource pour prendre ou évaluer des décisions ou bien pour comprendre un phénomène scientifique devient dès lors naturelle et le rapport favorable coût/puissance du calcul rend possible cette aspiration.

Une nouvelle matière première

Pour les entreprises, le « big data » constitue une mine d'or, pour les chercheurs c'est un levier puissant de la recherche scientifique. De nombreux produits emblématiques sont ainsi nés des données dans les dernières décennies : les moteurs de recherche des pages constituant le web, les publicités en lignes de nos traces de navigation, la médecine personnalisée à partir des profils médicaux, l'optimisation dans les villes des réseaux de capteurs intelligents ou encore, en science, les thérapies géniques à partir des profils génomiques ou la mise en évidence du boson de Higgs à partir des mesures du Cern. Les données sont perçues comme une source potentielle de richesse permettant d'optimiser le développement et la fabrication de produits existants, d'en construire des nouveaux ou encore

de proposer des nouveaux usages. Dans le même temps, l'utilisation de données toujours plus personnelles de manière parfois assez peu contrôlée est ressentie comme un sujet d'inquiétude.

La Science des données

La Statistique est un domaine scientifique multiforme. Au contact direct de l'environnement socio-économique et des autres disciplines scientifiques via les « données », elle propose un corpus de méthodes d'analyse et d'interprétation fondées sur des modèles mathématiques. Ce statut particulier l'expose à des mutations parfois brutales, principalement liées à sa très grande porosité vis à vis de l'extérieur.

L'évolution drastique du concept même de « données » décrit plus haut place cette discipline face à l'une des mutations les plus importantes de son histoire au travers de l'émergence de ce qu'il convient d'appeler désormais la Science des données (*Data Science* dans le monde anglophone) qui constitue au fond une re-fondation de la Statistique actant la montée en puissance du rôle du numérique dans la gestion des données, en amont aussi bien qu'en aval. Positionnée à la frontière entre Informatique et Mathématique, la Science des Données veut penser l'exploitation des données comme un tout : de l'acquisition au produit, en passant par le stockage, le traitement, l'analyse ou encore la visualisation.

Data scientist : un métier aux multiples visages

Cette révolution dans la perception et l'usage des données engendre un besoin de disposer au sein de chaque entreprise de compétences dans le domaine de la Science des données. Le nouveau héros dépositaire de ce savoir serait le « *data scientist* ». Celui-ci maîtriserait les trois clés de la Science des données : il connaitrait les méthodes avec leurs fondements mathématiques, il saurait les implémenter et il disposerait d'une connaissance approfondie du domaine d'application envisagé. En pratique, ce héros mythique n'existe pas et la maîtrise de toutes ces compétences n'est possible qu'à travers des équipes mélangeant des profils variés. Un « *data scientist* » est de façon plus réaliste un expert dans l'un de ces champs disposant de connaissances plus larges lui permettant d'être à l'interface entre plusieurs d'entre eux. On voit donc fleurir actuellement à divers endroits de par le monde des formations de niveau master dédiées à la Science des Données avec une inflexion plus ou moins forte vers un ancrage méthodologique ou vers la gestion des données. C'est vrai aussi en France évidemment où la très forte demande de recrutement de jeunes scientifiques dans ce secteur offre des opportunités d'emploi exceptionnelles aux futurs diplômés.

Et les mathématiques alors ?

Les thématiques statistiques liées à l'analyse des données de grande dimension se sont développées de façon très vivace durant ces dix dernières années. Classer des données en grande dimension, analyser les données massives et possiblement hétérogènes du« *big data* », bâtir des prévisions à partir de données fonctionnelles, analyser des données structurées en grands réseaux : voici autant de défis auxquels les statisticiens sont désormais confrontés. Dans le même temps, l'enracinement mathématique de la statistique reste plus fructueux et plus varié que jamais. Ceci est non seulement vrai au travers du lien historique avec le calcul des probabilités, mais aussi via la théorie de l'approximation et l'optimisation, ces deux derniers domaines étant eux aussi exposés à la révolution de la grande dimension. Ainsi les méthodes dites de « *compressed sensing* » possèdent-elles une résonance à la fois en théorie de l'approximation en grande dimension et en théorie statistique du signal, tout en reposant fondamentalement sur des propriétés spectrales fines de grandes matrices aléatoires. Dans la même veine, les solutions efficaces pour sélectionner au sein d'un grand nombre de variables les plus influentes d'entre elles, reposent sur des critères d'optimisation dont le bon fonctionnement est intimement lié à des questions de géométrie convexe.

Que ce soit pour la conception, l'expérimentation ou l'application des méthodes statistiques, la réflexion mathématique est aujourd'hui indissociable de la réflexion sur les algorithmes permettant son expression et sa mise en application. Si l'analyse des données en grande dimension a grandement stimulé les recherches des mathématiciens pour faire progresser la méthodologie statistique durant ces dix dernières années, les mathématiques du « big data » vont continuer à s'écrire. La validité des informations extraites des données renvoie à la question cruciale de la reproductibilité des découvertes scientifiques effectuées à partir de l'analyse de données massives. La nécessité de préserver la vie privée tout en analysant des données de plus en plus personnelles pose des questions méthodologiques liées à la théorie de l'information. Ces questions constituent des exemples parmi d'autres de problématiques donnant lieu à une intense activité de recherche à l'heure actuelle, s'appuyant de manière indispensable sur la puissance inductive des mathématiques.

Une nouvelle ère s'est donc ouverte qui voit les données, de tout temps matière première de la réflexion du statisticien, devenir aujourd'hui une matière première tout court possédant à la fois une valeur économique et scientifique. Cette ère est celle de la Science des données qui stimule et stimulera (nous en sommes intimement convaincus) des développements mathématiques passionnants.

Petite bibliographie

- Doing Data Science: Straight talk from the frontline, R. Schutt and C. O'Neil - O'Reilly
- Data science : fondamentaux et études de cas Machine learning avec Python et R, Éric Biernat / Michel Lutz - Eyrolles

Version du 18 juin 2016