« Et si vous deveniez “ data scientist ” ? »

Comme l'illustrent les trois précédents articles de ce cahier, les entreprises ont de plus en plus besoin de *data scientists*. Mais comment se former à cette nouvelle profession ? Le point sur les différentes voies pour y parvenir…

«*Métier le plus sexy du XXIe siècle*». C'est en ces termes que la prestigieuse Harvard Business Review qualifie le job de *data scientist*! Un métier en plein boom consistant à extraire de nouvelles connaissances à partir de données (*data*), grâce à des techniques issues des mathématiques appliquées, de la statistique et de l'informatique. En quelques années à peine, les formations se sont multipliées. Alors, comment s'y retrouver?

Les formations de qualité remplissent généralement trois critères: elles sont souvent d'un niveau master, adossées à des laboratoires ou des chaires… et à un écosystème industriel. L'objectif étant de pouvoir ainsi adapter les enseignements au rythme des évolutions de la recherche et de l'industrie. On trouve de tels masters dans les universités et les grandes écoles, le plus réputé restant le MVA (Mathématiques, Vision, Apprentissage) de l'ENS Paris-Saclay… qui a fêté ses vingt ans en 2017! Certains sont plutôt à dominante «maths», d'autres plus à dominante informatique.

Trois casquettes

Car dans la pratique, le métier de *data scientist* en regroupe plusieurs. Il y a bien sûr le concepteur d'algorithmes, nécessitant une solide formation générale en maths, en informatique scientifique et en modélisation. Mais il y a aussi l'«intégrateur» ou «développeur informatique» qui lui met en œuvre des algorithmes et des méthodes d'apprentissage, et doit gérer problèmes opérationnels et informatiques. Sans oublier le «spécialiste métier» ou «utilisateur»: véritable expert dans la manipulation de ces outils, et capable de comprendre le contexte d'utilisation (médical, marketing, transports…).

CEPE, École Polytechnique, TelecomParisTech, université Paris-Descartes… de nombreux organismes proposent aussi des formations continues qui se sont fortement développées ces dernières années, notamment dans le domaine du *big data*. Enfin, même s'ils ne remplacent pas les formations classiques, les cours en ligne ouverts à tous (Mooc) peuvent aussi se révéler très utiles.

«Les data scientists *devront de plus en plus se spécialiser.»*
NICOLAS VAYATIS
Responsable du master MVA de l'ENS Paris-Saclay

« L'enjeu est à la fois de faire évoluer nos statisticiens en interne vers ces nouveaux métiers et de recruter en externe. »
OLIVIER BAES
Responsable du Datalab de la MAIF

« Nous avons créé avec l'École des ponts et chaussées la première Chaire en France regroupant optimisation et apprentissage. »
ISABEL GOMEZ GARCIA DE SORIA
Directrice de la Recherche Opérationnelle chez Air France

«Nous initions des concours en sciences des données depuis 2014.»
MAGUELONNE CHANDESRIS
***Data scientist* chez SNCF Innovation & Recherche**

En chiffres

- La France aura besoin de **130 000** *data scientists* d'ici à 2020
- La *data science* pourrait créer **3 millions d'emplois** en Europe
- Un *data scientist* gagne en moyenne de **45 000 à 55 000 euros** par an

DATA CHALLENGES: DE VRAIS DÉFIS!

Aujourd'hui, de nombreuses formations intègrent des «data challenges», compétitions dans lesquelles des équipes d'étudiants data scientists s'affrontent pour résoudre des problèmes très concrets: prévoir la consommation électrique d'une zone, la fréquentation d'une gare, l'intérêt de consommateurs pour tel ou tel produit, améliorer des algorithmes de recommandations d'achat en ligne, etc. Souvent proposés par des entreprises, y compris les grandes (Air France, SNCF…), ces *data challenges* permettent aussi à certains étudiants de se faire repérer pour décrocher un stage… voire un job!

L'ESSENTIEL

- Les données, de diverses natures et toujours plus abondantes, sont interrogées *via* des requêtes ou analysées pour y dénicher des corrélations, des indicateurs...
- Les applications nécessitent en amont des algorithmes qui optimisent la collecte, la gestion et le prétraitement des données.
- Ces traitements sont facilités par divers outils, tels que l'algèbre relationnelle, le calcul distribué, le modèle MapReduce...
- Au-delà du calcul proprement dit, la gestion des données passe aussi par leur indexation et leur protection.

LES AUTEURS

MOKRANE BOUZEGHOUB est professeur à l'université de Versailles et directeur adjoint scientifique de l'INS2I, au CNRS.

ANNE DOUCET est professeure à l'université Pierre-et-Marie-Curie, déléguée scientifique à l'international de l'INS2I, au CNRS.

Traiter les BIG DATA avec raffinement

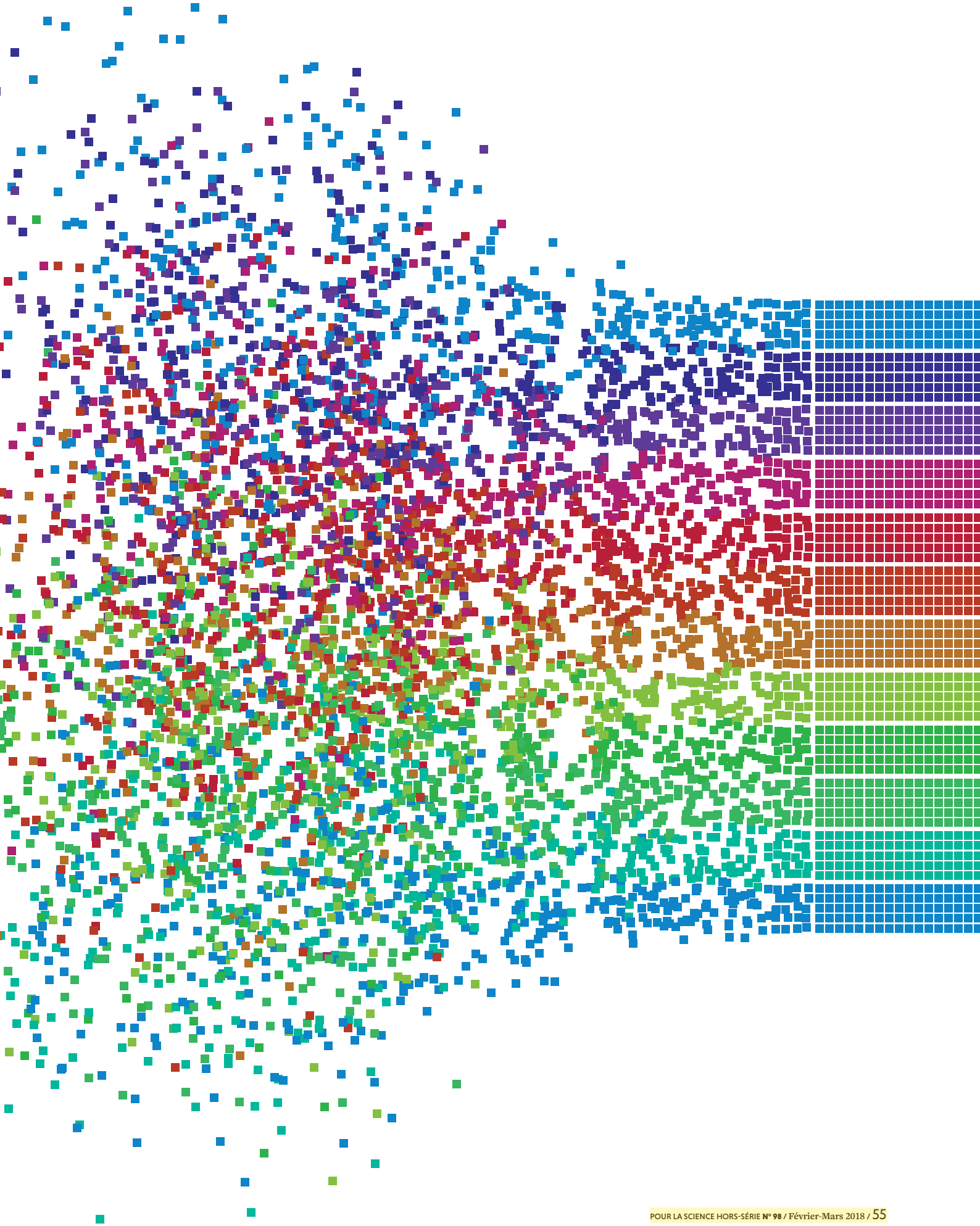
La gestion des *big data*, de la collecte jusqu'à l'analyse, requiert de nombreux algorithmes. Tous sont confrontés à des problèmes de volume, d'hétérogénéité et de qualité des données... Comment relever le défi ?

Au sortir d'un puits, le pétrole est un mélange complexe de nombreux hydrocarbures, d'autres composés et d'impuretés. Pour être utile, ce cocktail doit être raffiné de façon à séparer les produits exploitables: bitume, fioul, kérosène, gazole, essence... De même, les données réunies sous le terme de *big data* subissent un long processus de raffinage, de sélection et de transport avant de pouvoir livrer une quelconque information, c'est-à-dire d'être interrogées ou analysées. En quoi consistent ces traitements?

Lorsqu'on évoque les *big data*, on pense aussitôt à l'analyse de données plus qu'à leur exploration par des requêtes, comme si le volume croissant des données, en atteignant un certain niveau, rendait caduque leur interrogation directe. En réalité, ce sont deux approches complémentaires, toutes deux contraintes par les limites technologiques dues au traitement massif des données. Qui plus est, l'analyse au sens large nécessite souvent une exploration et un prétraitement individualisé des données. La chaîne de valorisation des *big data* est longue et semée d'embûches, la partie analyse ne représentant qu'un petit maillon de l'ensemble en bout de chaîne.

Les algorithmes sous-jacents à la collecte des données, leur indexation, leur nettoyage, leur stockage, leur annotation... peu visibles de l'extérieur, constituent pourtant l'ossature du pipeline >

Les données disparates doivent rentrer dans le rang pour être exploitables. C'est le rôle des algorithmes.



qui alimente les applications de décision ou d'apprentissage. Qu'il s'agisse d'algorithmes simples ou complexes, déterministes ou prédictifs, opérant sur des données exactes ou bruitées, leur compréhension est indispensable pour entrevoir en quoi consiste l'exploitation des *big data*.

Occultés par les succès récents de l'apprentissage machine et autres programmes de jeux, ces algorithmes enfouis sont donnés pour acquis, avec un coût nul pour leur mise en œuvre et leur exécution. Nous verrons que ce n'est pas le cas, un travail permanent de recherche et d'ingénierie est nécessaire pour les adapter aux nouveaux contextes applicatifs, les étendre, les optimiser et en inventer de nouveaux.

Il y a autant de processus de gestion et de valorisation des données que de raisons de valoriser ces données. On y retrouve cependant de grandes activités génériques (*voir l'encadré page ci-contre*), sur chacune desquelles existent des verrous importants.

Pour réaliser ces grandes activités, il est important de spécifier les tâches constitutives en fonction des applications visées. La consultation intensive des données n'a pas les mêmes besoins qu'un traitement des flux de données en temps réel. Une application d'aide à la décision n'a pas les mêmes besoins qu'une application de prédiction du panier d'un consommateur ou d'aide au diagnostic d'une tumeur. La nature de l'application et des tâches algorithmiques qui en découlent doivent être mises en perspective avec le modèle de calcul le plus adapté.

DEUX MODÈLES DE CALCUL

La gestion des données n'est pas constituée d'algorithmes *ad hoc* réalisés au coup par coup, dans un langage donné, en fonction des applications. C'est une science qui a ses fondements théoriques, ses propres objets d'étude, ses propres abstractions et ses propres méthodes. Les technologies produites ont pour vocation de réduire l'effort nécessaire à la gestion des données, de contrôler l'accès à ces données et de faciliter l'évolution des systèmes sans réécriture de leur code. Pour simplifier, il y a deux modèles de calcul sur les grands volumes de données : celui dédié à la recherche d'information et celui adapté à l'analyse de données. Même si les deux approches partagent de nombreux concepts et techniques et peuvent se combiner, elles diffèrent d'un point de vue sémantique.

Dans le premier cas, le sens est porté par la requête de l'utilisateur ; si la requête est pertinente, les données délivrées seront bien conformes aux critères de recherche spécifiés. Par exemple, « rechercher les références de produits vendus en quantité supérieure à 200 dans chacun des magasins de la région parisienne durant la période de Noël », est une requête complexe, mais qui définit clairement le calcul à faire et les critères de sélection de données. Ainsi spécifié, le système de gestion de données se chargera de générer le code nécessaire à l'exécution de cette requête. D'autres requêtes sont plus approximatives, leurs critères étant spécifiés de façon lâche. C'est le cas des recherches sur le Web. Ainsi, « rechercher les magasins parisiens pratiquant les prix les moins chers pour le chocolat et le champagne durant la période de Noël » est une requête floue. Elle impose un traitement approché du critère « moins cher » obligeant à présenter les résultats dans un certain ordre.

Dans le second cas, la question du sens est plus complexe, car elle interroge à la fois la méthode utilisée pour l'analyse des données et l'interprétation par l'utilisateur des résultats produits. Pour une requête d'analyse de l'évolution des cours de bourse des entreprises du CAC40 ou de prédiction des ventes de smartphones par région et par période de l'année, les résultats ne sont pas interprétables sans connaissance du métier et des modèles de prédiction sous-jacents.

Ce problème peut devenir plus aigu dans des algorithmes encore moins transparents. Par exemple, dans un système de reconnaissance fondé sur l'apprentissage, la confusion entre un chat et un caniche fait sourire, mais celle d'un honnête citoyen avec un criminel peut être dramatique. Les algorithmes sous-jacents à ces systèmes posent des défis à la société.

Il y a autant de processus de valorisation des données que de raisons de valoriser ces données

D'un point de vue conceptuel, un modèle de calcul est défini par les structures de données qu'il reconnaît et les opérations qu'il offre sur ces structures de données. Les structures de données peuvent être des tables (ou matrices), des arbres, des graphes ou des séquences finies ou infinies. Les opérations sont associées à chaque structure : les opérations de tri, projection, produit et union sur