L'émergence d'une nouvelle filière de formation : « data scientists »

Serge Abiteboul (INRIA, CNNum), coordinateur
François Bancilhon (Data publica)
François Bourdoncle (Dassault systèmes)
Stephan Clemencon (Telecom ParisTech)
Colin de la Higuera (U. Nantes, SIF)
Gilbert Saporta (CNAM)
Francoise Soulie-Fogelman (Kxen)

François Bourdoncle et Paul Hermelin ont été nommés « chefs de file » de la filière Big Data française. Leur mission est d'identifier les leviers que l'Etat peut actionner pour favoriser l'émergence d'un marché du Big Data exportateur net et créateur d'emplois en France. C'est donc un objectif économique et industriel, même si la recherche a, à l'évidence, son rôle à jouer dans la filière. L'un des axes importants identifiés pour tous les secteurs industriels qu'ils ont souhaité cibler est l'existence, en nombre, de besoins de profils de « data scientists ».

Ils nous ont chargés de réfléchir sur ce sujet. Le but est d'écrire quelques pages pour fin novembre 2013. La réflexion pourra évidemment être approfondie par d'autres après.

Sujet

La data science[1] (la « science des données ») et ses liens avec le big data (parfois traduit par la « datamasse ») ont été beaucoup discutés. Voir par exemple la définition de Wikipedia. Nous ne reviendrons pas ici sur cette définition ni sur le fait qu'il est nécessaire de développer en France une filière de formation de « data scientists ». La France disposant d'excellents enseignements en mathématiques, statistiques et informatique, on devrait pouvoir rapidement former d'excellents data scientists. La question qui est traitée ici est : « comment s'y prendre ? ».

Le terme « data scientist » est pris ici au sens large pour ce qui est des compétences, comme scientifique et ingénieur, peut-être technicien, et son domaine d'application. Il recouvre des besoins : des entreprises (ingénieurs, techniciens) et des institutions de recherche (scientifiques)

Une mesure qualitative/quantitative des besoins est compliquée et serait nécessaire. Lors de rencontres pour préparer ce rapport, nous avons beaucoup entendu :

- C'est du buzz et il n'y a pas de besoin spécifique (typiquement dans des entreprises classiques).
- On va avoir besoin pour rester dans la course de data scientists « haut de gamme » combinant des compétences gestion de données, machine learning et business/management (typiquement aussi dans des entreprises classiques).
- Le besoin explose, et on a déjà du mal à trouver des spécialistes (typiquement des entreprises plus Web).

[1] Davenport et Patil, « Data Scientist: The Sexiest Job of the 21st Century », Harvard Business Review, 2012

- On est au début de la vague et on fonce vers la pénurie que connaissent déjà d'autres pays plus avancés comme les Etats-Unis.

Les formations de data scientists croulent sous les demandes. Mais elles sont trop peu nombreuses et récentes pour pouvoir mesurer une demande extrêmement évolutive.

Méthode suivie

La plus ouverte et collaborative possible ; n'hésitez pas à faire circuler des versions de ce texte et à envoyer vos commentaires/critiques à mailto://serge.abiteboul@inria.fr.

Recommandations

Que faut-il enseigner ?

L'histoire[2] du terme « data scientist » remonte aux années 60 mais le consensus aujourd'hui est de définir le data scientist à l'intersection de trois domaines d'expertise[3] : (i) Informatique, (ii) Statistique et mathématiques, et (iii) Connaissances métier.

Liste indicative de cours pour illustrer :

- Analyse du problème et conception.
- Acquisition de données non structurées : trouver, choisir, représentation, crawling, scrapping, ETL.
- Manipulation de données : nettoyage (cleansing), stockage (NoSQL, bases colonnes, bases graphes), analyse de qualité, intégration, extraction des métadonnées, distribution des données.
- Analyse de données : statistique, apprentissage automatique (machine learning), data et text mining), données temps réel (time series, flux d'évènements), données imprécises/incertaines.
- Utilisation de réseaux sociaux et moteurs de recherche.
- Passage à l'échelle, parallélisme (Hadoop, map/reduce), cloud computing, architecture matérielle/logicielle.
- Visualisation de données massives (data viz), explication des résultats.
- Exploitation des données, décisionnel (BI).
- Aspects légaux, éthiques et sociaux, confidentialité, anonymisation, sécurité.
- **Séminaire de cas d'usage** présentant des expériences de sciences de données.

Des projets « les yeux dans les yeux de données en vraie grandeur » complémenteront ces cours en faisant intervenir tous les aspects de la science des données, depuis l'acquisition jusqu'au plan d'exploitation. Ces projets doivent encourager **la créativité** des élèves et s'appuyer, en particulier dans le cadre de la formation continue, sur leurs compétences existantes, leurs intérêts, leur domaine d'expertise. Ces formations s'appuieront sur des outils logiciels qui pourront être présentés par des industriels du domaine et utilisés pour les expérimentations.

Une telle formation est par nature pluridisciplinaire et il est indispensable de ne pas occulter une de ses dimensions techniques ou sa dimension métier :

---

[2] A very short history of data science, www.forbes.com.
[3] Diagramme de Venn de la data science, drewconway.com.

- **Complétude.** Nous parlerons ici de formation « complète » pour les formations tenant compte de ces aspects dans leur totalité. Si nous insistons sur ce point, c'est qu'aujourd'hui, par effet de mode, on fait parfois le marketing de formations informatiques ou de formations de statistiques comme formation de data scientists. Une telle approche est positive en ce qu'elle permet de faire bouger les lignes, il ne faut pas que cela perdure. La formation doit être à terme pluridisciplinaire. Selon les organismes formateurs, on pourra avoir cependant sans doute, à terme, des formations « à majeure » informatique, statistique ou métier.
- **Différentiation tardive.** La formation doit offrir un tronc commun le plus généraliste possible. La raison est simple : il est simplement impossible d'offrir des formations de data scientists spécialisées à chacun des métiers concernés. La formation de data scientists devra par contre offrir des spécialisations métiers en fin de cursus peut-être par le biais d'options, certainement à l'occasion de projets avec des interactions fortes avec les entreprises.
- **Illustration.** Un exemple très particulier permettra d'illustrer ces notions, celui de « journaliste de données ». On a de tout temps réalisé la difficulté de faire comprendre les statistiques aux décideurs et au grand public[4]. Ce qui a changé, c'est que l'utilisation de l'informatique s'est généralisée, et une personne, le journaliste de données, est capable de trouver des données, de les analyser et de communiquer, d'être le lien entre les données et les décideurs, le grand public. C'est bien là tout le spectre d'un data scientist. On peut imaginer une option journalisme de données dans une formation de data scientists et évidemment, comme nous l'encouragerons plus loin, l'introduction de cours de data science dans des écoles de journalismes.

Grands domaines :

Aéronautique & Automobile	*Manufacturing*	*Utilities*
Agriculture	*Trading financier*	*Publicité*
Assurance	*Santé*	*E-commerce*
Banque	*Télécommunications*	*Environnement*
Distribution	*Tourisme*	*Qantified self (= santé)*
Energie	*Transport & Smart Cities*	*Media & entertainment*
Luxe & beauté		

Quelques pistes d'action pour une réaction rapide

1. **Formations en ligne.** Il faut développer, dans ce domaine, les Moocs (Massive on-line open courses) ou autres formation à distance en insistant sur *la qualité* et *la complétude*. De tels cours peuvent être déployés très vite. Un but serait notamment de rendre disponible assez vite des cours qui servent de référence dans le monde francophone.
   Pour y arriver, on pourra s'appuyer sur la plateforme Fun du MESR ; il faudra affecter de vraies ressources : temps de préparation du cours, TP, plateforme technique (matériel, logiciel et jeux de données), support technique. Par exemple, l'ordre de grandeur pour produire un vrai Mooc est au minimum l'équivalent d'un professeur temps plein sur un an.

[4] Y a-t-il des compétences "Data" spécifiques ?, fing.tumblr.com.

On s'appuiera sur des cours existants de qualité pour développer une offre de cours dématérialisés.
Ce type de formation pourra servir à la formation continue dans les entreprises et à la formation des enseignants eux-mêmes.

2. **Curriculum et label « data science »**. On définira précisément un curriculum général de formation de data scientists ainsi que de formations spécialisées dans des domaines spécialisés comme le marketing, les sciences, le journalisme. On contrôlera la qualité des formations de data scientists par exemple par le biais d'un label.
3. **Formations professionalisantes**. On soutiendra leur développement en insistant sur *la qualité* et *la complétude* :
   a. Stages courts de quelques jours (Fac, CNAM, écoles, entreprises de formation)
   b. Formations diplômantes (universités, CNAM, IUT, écoles)
4. **Faire évoluer les formations actuelles**
   a. Encourager les formations statistiques à inclure plus d'informatique et les formations informatiques à inclure plus de statistique.
   b. Encourager à introduire une vraie exposition aux aspects business
   c. Mettre à leur disposition des intervenants universitaires ou industriels pour combler les manques.
5. **Réseau.** On associera les entreprises à la formation de data scientists par le biais d'un réseau avec une communauté de partage de ressources (outils, données, formation) et d'expériences. Ce réseau pourra en particulier proposer des projets d'expérimentation aux élèves. Il pourra aller jusqu'à la mise en place de collaborations avec mise à disposition de personnel.
   Ce réseau pourrait se faire dans le cadre du *Centre de ressource technologique* proposé dans le cadre de la Filière Big data.
6. **Nouveaux centres d'éducation.** Envisager des coopérations avec des écoles comme Simplon.co ou 42 et des expériences nouvelles d'éducation des data scientists.

Pour structurer des formations à plus long terme

1. **Agir au niveau européen** : Voir Annexe B.
2. **Evaluation** : Réaliser une évaluation des besoins tant qualitatifs que quantitatifs.
3. **Les formations métiers.** Développer la formation de data scientists dans les formations métier : Ecoles d'ingénieurs, école de management et de commerce, école de journalisme, école d'administration.
4. **Le droit des données.** Développer dans les études juridiques la place des droits du numérique, et en particulier des droits des données.
5. **Matériel éducatif** : développer une gamme de matériels éducatifs ouverts et de qualité dans ce domaine, comme des livres, des collections de données tests, des Moocs.
6. **Assouplissement des structures éducatives publiques**.
   a. Il y a de véritables volontés d'enseignants de développer de nouveaux enseignements autour de la data science. Ces volontés sont freinées par des procédures – il faut aujourd'hui plus de 4 ans pour lancer un nouveau master. Il faudrait introduire plus de souplesse/rapidité en générale ou au minimum dans ce domaine où tout évolue très vite.
   b. La formation continue publique fonctionne plutôt bien en ce qui concerne la formation diplômante. La formation à la demande (stages intensifs)

rentre mal dans le cadre rigide du publique. Il faudrait introduire plus de souplesse dans l'utilisation du personnel enseignant ainsi que des prestataires extérieurs.

c. Il faudrait faciliter les échanges entre entreprises et centres de formation, par exemple, en simplifiant les conventions de stages trop lourdes pour des stages très courts, ou en installant des instruments de mobilités pour que des enseignants puissent passer du temps dans des entreprises, et dans l'autre sens aussi.

7. **Bac+3.** Envisager un parcours de licence bi-disciplinaire Mathématiques-Informatique orienté autour des données et permettant une formation équilibrée qui pourrait d'une part satisfaire les besoins des entreprises à ce niveau-là et servir de socle à des Masters en Sciences de Données.

Annexe A : European cooperation

This comes right in an objective of the European community: Information and Communication Technologies in Horizon-2020 (ICT 15)

Goal

To contribute to capacity-building by designing and coordinating a network of European skills centres for big data analytics technologies and business development. The network is expected to identify knowledge/skills gaps in the European industrial landscape and produce effective learning curricula and documentation to train large numbers of European data analysts and business developers, capable of (co)operating across national borders on the basis of a common vision and methodology.

Outcome

Availability of deployable educational material for data scientists and data workers and thousands of European data professionals trained in state-of-the-art data analytics technologies and capable of (co)operating in cross-border, cross-lingual and cross-sector European data supply chains.

A group is being formed to submit a proposal with Volker Markl from Germany. For now, Serge Abiteboul and Stefan Clémençon are representing France.

Annexe B : l'existant en France

Un article récent du Monde de l'éducation parle du sujet[5].

Telecom-Paris

http://www.telecom-paristech.fr/formation-continue/masteres-specialises/big-data.html Le Mastère Spécialisé « Big Data : gestion et analyse des données massives » qui a démarré cette année est ce qui nous paraît le plus dans l'esprit du programme complet précisé précédemment. Il nous semble être la bonne base pour former des « data scientist ».

INP-Grenoble

A Grenoble l'Ensimag et la GEM (Grenoble Ecole de Management) vont ouvrir l'an prochain un Mastère spécialisé « Data scientists ». Semble tout à fait dans l'esprit du programme complet. Il sera composé de 5 mois de cours et 10 mois de mission en entreprise.

Il existe déjà un master mais plutôt MathAppli avec juste une option « data scientists ».

L'ENSAE ParisTech

L'ENSAE ouvre en octobre 2013 une nouvelle filière de 3ème année de son cycle ingénieur : la voie Data Science. Cette filière permettra, entre autres, d'acquérir les compétences attendues pour les postes de Data Scientist et Chief Data Officer qui émergent avec le développement des Big Data.

L'ENSAE a aussi ouvert un master spécialisé aligné sur cette troisième année.

La formation est surtout orientée statistiques plus que gestion de données massives.

L'ENSAI

L'ENSAI a aussi ouvert une filière « data science et big data ». Plus informatique que l'ENSAE.

Paris 6

Paris 6 va ouvrir un master de data science « Données, Apprentissage, et Connaissances », succédant au Master actuel IAD, orienté plutôt informatique et data mining, mais qui devrait évoluer.

CNAM

Le CNAM a des formations prévues pour la rentrée 2014 : des certificats maison sur un an (certificat de spécialisation, mi informatique, mi statistique, pour ingénieurs en poste ; certificat de compétences, niveau licence, visant les besoins de gestion de data centers), évolution d'un parcours du Master Statistique vers le « Big Data » incluant des modules d'informatique

Le CNAM travaille par ailleurs sur une offre de MOOC, dans la prolongation de leur expérience de formation à distance. Mais pour l'instant, rien n'est prévu sur la data science.

---

[5] Sophy Caulier, La montée en puissance de la Datamasse, Le Monde 10.12.2013.

HEC

HEC a créé une filière Big Data et business analytics en partenariat avec IBM. Il s'agit de sensibiliser de futurs responsables business et manageurs aux possibilités de l'utilise de l'analyse de données massives.

Université Paris Saclay

Dans sa spécialité M2 « Information, Apprentissage, Cognition », l'Université Paris Saclay offre des enseignements liés à la Data Science, regroupant des thématiques autour de la gestion de larges volumes de données hétérogènes, de l'apprentissage, et de connaissances. https://www.dep-informatique.u-psud.fr/formation/lmd/M2R_IAC
A partir de la rentrée 2015, l'Université Paris Sud participera - avec d'autres établissements du périmètre Paris Saclay - à deux parcours M2 autour des Big Data et Data Science. Notamment, l'un de ces parcours (Data & Knowledge) inclut la majorité des enseignements mentionnés dans le programme précisé précédemment.

Plateforme de l'Institut Mines-Télécom et Genes

La plateforme d'expérimentation sur les Big Data offrira un grand corpus de données et intègrera des technologies matérielles, logicielles. Accès de l'OpenData ou des données privées en mode sécurisé. Exploitation en mode SAAS/PAAS par un catalogue de services et applicatifs. La plate-forme est une initiative de l'Institut Mines Télécom à laquelle s'est associée le Genes. Son utilisation n'est pas réservée aux seuls chercheurs de ces institutions ; elle se destine également à des chercheurs de tout organisme et d'entreprises (PME et grands groupes) en dehors de toute exploitation commerciale

Associations professionnelles

- http://data-tuesday.com/ (Open data, data viz, big data)
- http://www.alliancebigdata.com/ (un portail Web)
- http://www.kdnuggets.com/ (« Data Mining Community's Top Resource for Data Mining and Analytics Software, Jobs, Consulting, Courses, and more»)
- http://ww.gfii.fr/ (GFII - Les acteurs du marché de l'information et de la connaiss ance)

Annexe C : Some courses in the world

Une étude publiée en Angleterre en novembre 2013 (http://www.e-skills.com/research/research-publications/big-data-analytics/#November%20report) indique avoir identifié 31 000 emplois de spécialistes big data et 383 000 d'utilisateurs de big data. En janvier 2013 l'étude avait identifié 3 790 offres d'emploi en UK pour le 3ème trimestre 2012, dont 42% pour des postes de développeurs, 10% des architectes, 8% des analystes et 6% des administrateurs, avec une croissance de 912% sur les 5 dernières années.

Aux Etats-Unis, la société http://www.wantedanalytics.com indique que 88 000 offres d'emploi en septembre 2013 réclamaient des compétences « big data ». Les formations devraient donc viser à couvrir des besoins à la fois de spécialistes (analystes, administrateurs, architectes et développeurs) et d'utilisateurs ; ce qui recoupe les trois domaines indiqués précédemment (informatique, statistiques et expertise métier). Une étude sur les besoins en outils Big data dans la région de Washington DC http://bigdatastudio.com/2013/10/20/3707/ fait apparaître des besoins informatiques standard, à côté d'autres très spécifiques (Hadoop, MapReduce) : le besoin en développeurs est très significatif.

Voir

- http://datascience101.wordpress.com/2012/04/09/colleges-with-data-science-degrees/ et
- http://www.kdnuggets.com/education/index.html

pour une liste complète.

Quelques pays semblent en avance : US (California), Israel, Singapoure. Par exemple, Israël est depuis des années leader en fouille de données et machine learning. Ils ont de nombreux projets en data science avec des compagnies multinationales qui installent des équipes de R&D en Israël dans ces domaines et des startups. Nous n'avons pas eu le temps d'obtenir d'information de pays comme la chine ou la Corée.

Moocs dans Coursera

- Introduction to data science: B. Howe – U. Washington (https://www.coursera.org/course/datasci)
- Machine learning: A. Ng – U. Stanford (https://www.coursera.org/course/ml)
- Neural Networks for Machine Learning: G. Hinton – U. Toronto (https://www.coursera.org/course/neuralnets)
- Social network Analysis: L. Adamic – U. Michigan (https://www.coursera.org/course/sna)

IBM http://bigdatauniversity.com/ (around Big data and data analysis)

Singapore undergraduate

The Bachelor of Science (Business Analytics) degree program is an inter-disciplinary undergraduate degree program offered by the School of Computing with participation from the Business School, Faculty of Engineering, Faculty of Science, and Faculty of Arts

and Social Sciences. This is a four-year direct honours program which offers a common two-year broad-based inter-disciplinary curriculum where all students will read modules in Mathematics, Statistics, Economics, Accounting, Marketing, Decision Science, Industrial and Systems Engineering, Computer Science and Information Systems.

http://www.comp.nus.edu.sg/undergraduates/ug-bsc-ba_prospective.html

Singapore master

http://msba.nus.edu/

Singapore Industry: Big Data analytics and data scientists training center opens

1. Dell, Intel Corporation, and Revolution Analytics have set up a Big Data Innovation Center in Singapore. The new center brings together expertise across all three organizations to provide extensive training programs, proof-of-concept capabilities and solution development support on big data and predictive analytic innovations for Asia market.
2. As part of efforts to provide more receptacles for "expert" level skills training, Singapore's Infocomm Development Authority (IDA) has announced two Centers of Attachment (COA) in partnership with EMC and Microsoft respectively.

U.C. Berkeley Master of Information and Data Science

2 years fully on line

Caltech

Learning from data. Yaser Abu-Mostafa. http://work.caltech.edu/lectures.html

Cornell Tech

Joint Cornell-Technion campus in NYC. Data management/analysis is one focus.

New York University Masters

The new NYU Center for Data Science is offering a 2 year MS in Data Science requiring a strong background in math, computer science and statistics.
http://datascience.nyu.edu/.

Israel

Master program at Ben Gurion University in “Data Science and Business intelligence” ; mostly academic.

Interesting courses at Tel Aviv : http://www.cohenwang.com/edith/bigdataclass2013/

Filière d’un Master Erasmus Mundus

Master Erasmus Mundus IT4BI « Information Technologies for Business Intelligence », http://it4bi.univ-tours.fr/. Erasmus Mundus est un programme européen de coopération et de mobilité dans le domaine de l’enseignement supérieur (http://eacea.ec.europa.eu/index_en.php). Ce programme de 2 ans a pour objectif de former des informaticiens pour comprendre et développer des stratégies décisionnelles. Délivré par ULB en Belgique, Univ. de Tours (Patrick Marcel), Central Paris, Politècnica de Catalunya, la TU Berliun. Volker Markl à TU Berlin a une filière Big data.

Annexe D : Remerciements

Nous remercions les personnes suivantes pour des discussions et leurs commentaires sur des versions de ce rapport. Leurs contributions n'entraînent pas nécessairement leurs adhésions à toutes ses recommandations.

Marie-Aude Aufore (Central Paris)
Francis Bach (ENS, Inria)
Christine Ballagué (Telecom Paris, CNNum)
Noureddine Belkhatir (IUT Grenoble)
Marc Chemin (Cap Gemini)
Vassilis Christophides (Technicolor)
Jean-Louis Constanza (Criteo)
Matthieu Cornec (Cdiscount)
Yves Denneulin (IMAG)
Antoine Frachau (Directeur Génes)
Patrick Gallinari (Paris 6)
Pascal Guitton (Inria)
Radu Ispas (Keyrus)
Daniel Kaplan (Fing, CNNum)
Nicolas Kayser-Bril (Data journalist)
Brigitte Plateau (Ensimag)
Julien Pouget (ENSAE)
Pierre Senellart (Telecom Paris)