DATA SCIENCE : UNE FORMATION INTERNATIONALE DE NIVEAU MASTER EN SCIENCE DES DONNÉES

Massih-Reza AMINI[1], Jean-Baptiste DURAND[2], Olivier GAUDOIN[3], Éric GAUSSIER[4] et Anatoli IOUDITSKI[5]

TITLE

Data Science: an international training program at master level

RÉSUMÉ

Nous présentons la formation internationale de niveau master 2 en *Data Science* de l'Université Grenoble Alpes et de Grenoble INP : spécificités et positionnement de la formation, fonctionnement et aspects historiques de sa création, programme de cours détaillé et perspectives d'évolution.

Mots-clés : *formation internationale, master 2, data science.*

ABSTRACT

We present the international training program in Data Science at master 2 level. This program is supported by both Grenoble Alpes University and Grenoble INP. In this article, we elaborate on the specific features of the program, its strategic position, operating and historical features, the detailed contents of courses and perspectives of evolution.

Keywords: *international program, master 2, data science.*

1 Positionnement et fonctionnement de la formation

1.1 Positionnement

La spécificité de la science des données par rapport à l'informatique et à la statistique semble faire en grande partie consensus (Besse et Laurent, 2015 ; Overton, 2015). Il s'agit de mettre en œuvre un faisceau de compétences multidisciplinaires à l'interface de l'informatique (systèmes et bases de données répartis, programmation), la statistique, l'apprentissage automatique, l'optimisation et la connaissance d'un champ d'application. Cependant des divergences d'appréciations demeurent quant à l'importance donnée à chacune de ces compétences. Par exemple Overton (2015) met en avant la démarche d'investigation face à une question informelle traduite en hypothèse, et la nécessité de savoir justifier et traduire de manière informelle les méthodes de résolution plutôt que de savoir démontrer leurs propriétés théoriques.

La formation *Data Science*[6] présentée dans cet article reprend les ingrédients ci-dessus, suivant un programme détaillé en Section 2. Il vise à répondre à une demande de formation

[1] Université Grenoble Alpes, Laboratoire d'Informatique de Grenoble (LIG), Massih-Reza.Amini@imag.fr
[2] Grenoble INP Ensimag, Laboratoire Jean Kuntzmann (LJK), Jean-Baptiste.Durand@imag.fr
[3] Grenoble INP Ensimag, Laboratoire Jean Kuntzmann (LJK), Olivier.Gaudoin@imag.fr
[4] Université Grenoble Alpes, Laboratoire d'Informatique de Grenoble (LIG), Eric.Gaussier@imag.fr
[5] Université Grenoble Alpes, Laboratoire Jean Kuntzmann (LJK), Anatoli.Iouditski@imag.fr
CS40700 - 38058 Grenoble cedex 9, France

croissante de diplômés de niveau master, en science des données. Cette formation se veut adaptée, d'une part, à l'intégration des diplômés en entreprise (entreprises de services en science des données ou scientifique des données dans une entreprise dont ce n'est pas l'activité principale). Elle doit également permettre la poursuite d'études doctorales en informatique, statistique, apprentissage automatique ou à l'interface de ces disciplines et d'un domaine d'application. À ce titre, c'est à la fois une formation professionnelle et de recherche. Tous les enseignements doivent donc fournir les compétences nécessaires à ces deux objectifs. Ainsi, l'accent est mis sur la problématique d'adaptation de modèles, méthodes et algorithmes existants au contexte des données distribuées voire massives, plutôt que sur d'autres problématiques de recherche relevant principalement de la méthodologie statistique décrits dans Fan *et al.* (2013), comme l'accumulation du bruit en grande dimension, l'apparition des corrélations empiriques fortuites ou le phénomène d'endogénéité.

Data Science sera, à partir de septembre 2016, une option des parcours de masters 2 internationaux en mathématiques appliquées (MSIAM) et en informatique (MoSIG) de l'Université Grenoble Alpes et Grenoble INP. Le recrutement se fera à la fois sur les filières d'informatique et de mathématiques appliquées, au niveau M1.

1.2 Historique et fonctionnement

En septembre 2013, le master 2 recherche de mathématiques appliquées de Grenoble est devenu un master international dont les cours sont entièrement en anglais et s'appelle maintenant *Master of Science in Industrial and Applied Mathematics* (MSIAM[7]). L'objectif général de ce master est de former des experts de haut niveau pour les métiers de la recherche et de l'ingénierie en mathématiques appliquées dans un large spectre de domaines où la demande du monde socio-économique est forte. Au moment de la création de MSIAM, une option *Data Science* a été ajoutée aux trois spécialisations existantes de la deuxième année de master, *Modelling and Scientific Computing*, *Geometry Image and Computer-Aided Design*, et *Statistics*.

Par ailleurs, le master recherche en informatique de Grenoble est devenu international depuis plus longtemps et se nomme *Master of Science in Informatics at Grenoble* (MoSIG[8]). Ces masters sont depuis toujours co-habilités entre l'Université Grenoble Alpes (c'est-à-dire avant 2016 l'Université Joseph Fourier) et l'Institut Polytechnique de Grenoble (Grenoble INP), et co-gérés par l'UFR Informatique, Mathématiques et Mathématiques Appliquées de Grenoble (IM2AG) de l'Université Grenoble Alpes (UGA) et l'École Nationale Supérieure d'Informatique et de Mathématiques Appliquées (Ensimag) de Grenoble INP. Les enseignants des deux établissements interviennent à part égale dans la formation et apportent les compétences complémentaires d'une université scientifique et d'une école d'ingénieurs. Cette étroite symbiose permet de mutualiser des cours entre les différents masters, ainsi qu'entre les masters et les filières d'ingénierie de l'Ensimag. Aussi, à sa création, l'option *Data Science*, incluse dans le master 2 MSIAM, a mutualisé des cours avec le master 2 MoSIG et avec la troisième année de l'Ensimag dans sa filière Modélisation Mathématique, Image et Simulation.

[6] formation dispensée en anglais, voir Section 1.2.
[7] http://msiam.imag.fr
[8] http://mosig.imag.fr

Après trois ans d'existence, une nouvelle étape sera franchie à la rentrée 2016, puisque l'option *Data Science* sera proposée conjointement dans les masters MSIAM et MoSIG. En effet, la science des données s'appuie par nature à la fois sur les mathématiques et sur l'informatique. Or une des caractéristiques phares de Grenoble est l'imbrication forte entre ces deux domaines, illustrée notamment par l'existence de l'Ensimag. Par conséquent, la formation en *Data Science* propose une véritable double compétence en mathématiques appliquées et informatique. Le contenu de la formation a été revu pour renforcer cet aspect, et sera présenté en Section 2.

Le recrutement de l'option *Data Science* se fait en France et dans le monde entier. Le recrutement international des dernières promotions s'est entièrement fait hors de l'Union Européenne : Russie, Inde, Iran et Colombie. Les élèves ingénieurs de l'Ensimag peuvent valider leur diplôme d'ingénieur avec ce parcours. En 2014-2015, *Data Science* comportait six étudiant(e)s. En 2015-2016, nous avons actuellement 13 étudiant(e)s en *Data Science*.

Grenoble INP propose également un autre type de formation sur les mégadonnées, un mastère spécialisé *Big Data*[9], en collaboration avec Grenoble Ecole de Management (GEM). De niveau bac +5, cette formation est accessible à des diplômés de niveau master (écoles d'ingénieurs, écoles de management ou master universitaire) ou à des personnes en formation continue. Elle est composée de 5 mois de cours et 10 mois de mission en entreprise. Là où le parcours *Data Science* propose une formation purement scientifique, le mastère *Big Data* propose une formation à l'articulation entre les aspects scientifiques, techniques, managériaux, business et stratégiques, où la statistique n'a qu'une place marginale.

2 Programme de cours

La formation est axée autour de cours scientifiques en informatique et en mathématiques appliquées, de nombreux enseignements étant multidisciplinaires. Cette formation est dispensée par des enseignants-chercheurs à l'université (UGA et Grenoble INP) ou par des chercheurs au CNRS ou à Inria. Le programme est complété par des enseignements d'ouverture vers des champs d'application, de langues et par des projets. Comme pré-requis, il demande des connaissances fondamentales en probabilités, statistique, algèbre linéaire, analyse, optimisation et programmation.

Des cours de remise à niveau sont proposés en début d'année scolaire. Ils sont plutôt à destination des nouveaux arrivants à Grenoble, par opposition à ceux qui arrivent de première année de master MSIAM ou de l'Ensimag, qui sont déjà familiarisés à ces notions. Ces cours ne sont pas comptabilisés dans le total des ECTS et ne donnent pas lieu à une évaluation comptant dans l'attribution du diplôme. Ils peuvent éventuellement permettre d'évaluer le niveau des participants en programmation.

En tant que parcours de master professionnel et recherche, le parcours MSIAM doit comprendre des unités d'enseignements (UE) permettant aux étudiants d'être évalués sur des compétences valorisables soit plutôt en entreprise, soit plutôt en laboratoire. À ce titre, les UE scientifiques ci-dessous comprennent au choix soit environ neuf heures additionnelles de travaux pratiques destinés de manière privilégiée au parcours professionnel, soit des travaux

[9] http://ensimag.grenoble-inp.fr/formation/big-data-analyse-management-et-valorisation-responsable-584683.kjsp?RH=ENSIMAG_FR

de recherche orientés vers la poursuite d'études doctorales. Les UE ci-dessous sont proposées au semestre 1. Le semestre 2 est dédié au stage en entreprise ou laboratoire de recherche.

2.1 Remise à niveau en analyse numérique matricielle

Cette remise à niveau vise à introduire les concepts en analyse matricielle requis dans le reste du programme : dans un premier temps, les notions de matrices, valeurs propres, conditionnement et applications linéaires. Puis les aspects numériques des méthodes de résolution sont abordés : décompositions matricielles, résolution d'équations linéaires, calculs d'éléments propres (3H de cours et 3H de travaux dirigés).

Des travaux pratiques (3H) sont réalisés dans le but de fournir une vision concrète du coût des opérations élémentaires et des applications du calcul matriciel (telles que la régression linéaire).

2.2 Remise à niveau en optimisation numérique

Cette remise à niveau vise à introduire les concepts en optimisation requis dans le reste du programme : dans un premier temps, une introduction à l'optimisation apporte les définitions fondamentales, telles que les ensembles et fonctions convexes, en s'appuyant sur des exemples. Puis les algorithmes pour l'optimisation sans contraintes sont présentés : méthode de gradient, de Newton et quasi-Newton (3H de cours et 3H de travaux dirigés).

Des travaux pratiques (3H) sont réalisés pour aborder l'optimisation dans le cadre de la régression logistique, et afin de comparer différents algorithmes.

2.3 Outils pour le développement logiciel

Il s'agit une UE optionnelle de 3 ECTS comportant 9H de cours et 30H de travaux pratiques. Elle sert de remise à niveau en informatique au besoin, et vise à fournir les compétences nécessaires à la programmation appliquée au calcul scientifique, avec l'utilisation de méthodes, langages et bibliothèques adaptées. Il s'agit principalement de développement en C++, gestion de projets et évaluation des performances et ressources (outils cmake, subversion, qtcreator, gdb, gprof et valgrind). L'algèbre linéaire est abordée via la bibliothèque Eigen, les interfaces utilisateurs via Qt, le traitement de données via XML, et le prototypage et l'interfaçage via Python.

2.4 Optimisation convexe et distribuée

Cette unité d'enseignement fortement orientée autour d'un projet est décrite en détail dans la section 3 « Stage et projet ».

2.5 Calcul hautes performances et systèmes distribués

Ce cours de 6 ECTS vise à fournir les concepts nécessaires à l'utilisation de systèmes distribués et à la mise en œuvre de méthodes de calcul sur ces systèmes. Il est composé des volets « Calcul hautes performances » et « Introduction aux systèmes distribués ». Le calcul hautes performances (3 ECTS, 18H) est dédié à l'acquisition de compétences en calcul parallèle avec une forte orientation vers le calcul scientifique. Les enseignements de systèmes distribués (3 ECTS, 18H) introduisent les principes fondamentaux des systèmes distribués,

notamment sur le plan algorithmique. Ils en présentent les caractéristiques essentielles, ainsi qu'un certain nombre d'algorithmes utilisés dans ces systèmes : exclusion mutuelle, *snapshot*, diffusion causalement ordonnée, diffusion totalement ordonnée, consensus, etc.

2.6 Gestion des données à grande échelle

La gestion de données à grande échelle (3 ECTS, 18H) vise à fournir les compétences en bases de données réparties, réseaux pair à pair, virtualisation, calcul distribué et NoSQL, avec également une sensibilisation à des méthodes de fouille de données.

2.7 Apprentissage automatique

Ce cours de 6 ECTS présente les concepts fondamentaux de la théorie de l'apprentissage supervisé (3 ECTS, 18H). Nous exposons, entre autre, la notion de consistance du principe de la minimisation du risque empirique selon lequel la plupart des algorithmes en apprentissage supervisé ont été développés.

L'étude de cette consistance nous mènera à l'exposé du second principe fondamental en apprentissage qui est la minimisation du risque structurel, ouvrant le champ au développement de nouveaux modèles en apprentissage automatique. En particulier, nous présentons cinq modèles de classification classiques qui ont été les précurseurs dans le développement d'algorithmes d'apprentissage plus complexes, notamment les machines à noyaux. (3 ECTS, 18H). Ces algorithmes seront mis en œuvre à travers un challenge d'analyse de données impliquant un travail d'implémentation conséquent (réalisé en dehors des heures encadrées).

Notre souci est de proposer un exposé cohérent reliant la théorie aux algorithmes présentés dans ces cours.

2.8 Fouille de données

Ce cours de 6 ECTS vise à fournir des compétences, à la fois en terme de programmation et de modélisation, en fouille de données et *clustering* dans des bases de données. Une première partie est dédiée aux modèles génératifs et leur application au *clustering* et à la classification supervisée (3 ECTS, 18H). On y aborde les modèles probabilistes graphiques, les modèles à variables latentes (dont les mélanges), l'analyse en composantes principales, l'estimation de densité et les modèles de Markov cachés. C'est l'occasion de traiter des données de nature vectorielle, séquentielle ou graphique, les problèmes d'hétérogénéité des données, de discuter des avantages et inconvénients des modèles génératifs, et de présenter des principes généraux de modélisation statistique.

La deuxième partie (3 ECTS, 18H) se concentre sur les algorithmes importants (et les modèles associés) de la fouille de données. Nous passons en revue en particulier les algorithmes de recherche de motif fréquents, de calcul du *PageRank*, de Monte Carlo et Monte Carlo par chaînes de Markov, d'échantillonnage de Gibbs, de factorisation matricielle et de *clustering* non probabilistes (k-moyennes et k-moyennes généralisées, *clustering* spectral et partitionnement de graphes). Les applications associées sont nombreuses et couvrent plusieurs champs, depuis l'analyse de données (de différentes types : textes, images

ou réseaux sociaux par exemple) à la recherche d'information, en passant par les systèmes de recommandation.

2.9 Ouverture vers un champ d'application

Une composante importante de la science des données est la formalisation de problèmes mal spécifiés dans des domaines d'application divers. Le module de spécialisation met en œuvre de la modélisation et de l'analyse de données réelles dans un champ d'application particulier. Ce module consiste au choix de deux UE de 3 ECTS, en principe parmi celles des autres options des parcours MSIAM et MoSIG – par exemple Visualisation de l'information, Biologie computationnelle, Modèles stochastiques pour les neurosciences, Méthodes de Monte Carlo pour l'ingénierie financière, Recherche d'information, Traitement d'images…

2.10 Langues

Les étudiants non-francophones suivent des cours de « Français Langue Étrangère » (FLE). Les étudiants francophones suivent des cours d'anglais. Ces derniers ne sont pas regroupés par niveau mais ont à choisir parmi quatre thèmes possibles (pratique de l'anglais dans quatre contextes professionnels ou culturels différents). Le cours de langue compte pour 3 ECTS.

3 Stage et projets

Les enseignements ci-dessus seront complétés par un projet de mise en œuvre de problématiques en science des données au premier semestre en optimisation convexe et distribuée, par le choix d'un cours optionnel à visée applicative (paragraphe 2.9) et par un challenge d'analyse de données en apprentissage automatique (projet non encadré nécessitant des implémentations conséquentes). Un stage en entreprise ou en laboratoire de recherche aura lieu au second semestre.

Les étudiants suivant l'option « pro » du parcours *Data Science* auront des travaux pratiques renforcés (TP de 3H à 9H) dans les enseignements suivants : calcul hautes performances et systèmes distribués, apprentissage automatique, fouille de données, optimisation convexe et distribuée. Il s'agira de mettre en œuvre des problématiques de fouille de données, apprentissage et calcul distribué sur des données libres. Ces travaux donneront lieu à une évaluation spécifique sur la base d'un rapport écrit complémenté, pour certaines unités d'enseignement, par une soutenance. Les étudiants suivant l'option « recherche » seront évalués à l'aide d'un travail alternatif aux TP basé sur la lecture d'articles de recherche. Un examen écrit commun aux deux options complètera cette évaluation différenciée.

3.1 Optimisation convexe et distribuée

Cette unité d'enseignement de 27H (3 ECTS) comporte d'une part des cours et travaux dirigés (12H) et d'autre part un projet d'application (15H).

Le cours introduit des notions avancées d'optimisation convexe, dont des éléments d'analyse convexe (dualité, opérateurs proximaux). L'accent est mis sur l'identification des difficultés d'un problème d'optimisation avec des illustrations en apprentissage supervisé

(problèmes de classification et de régression) et en recherche opérationnelle (méthodes de décomposition de matrices). Puis les algorithmes d'optimisation convexe sont abordés (gradient, gradient proximal, gradient conditionnel, *alternating direction method of multipliers*).

Les principes du calcul distribué sont ensuite présentés : architectures de calcul, schéma *MapReduce*, bibliothèque MPI, cadriciel Spark. Enfin, ces principes sont revisités dans le cadre de problèmes d'optimisation : algorithmes d'optimisation distribuée, algorithmes stochastiques et méthodes d'optimisation asynchrones.

Le projet vise à appliquer des méthodes générales de calcul parallèle, puis plus spécifiquement des méthodes d'optimisation distribuée. Elles sont mises en œuvre pour l'élaboration d'un système de recommandation, avec également une application à la régression logistique parcimonieuse en grande dimension.

3.2 Stage de second semestre

Suivant l'orientation préférentielle des étudiants, un stage de second semestre a lieu en entreprise ou en laboratoire de recherche (27 ECTS). Le stage doit durer de 4 à 6 mois, et le plus souvent dure environ 22 semaines. C'est en définitive le choix de réaliser le stage en entreprise ou en laboratoire de recherche qui donne sa coloration à la formation : pro ou recherche. Le projet est également validé comme projet de fin d'étude ingénieur pour les étudiants Ensimag.

4 Perspectives d'évolution

Si les effectifs de l'option *Data Science* venaient à augmenter d'une dizaine d'étudiants dans les prochaines années, on pourra envisager d'en faire un parcours de master à part entière.

Par ailleurs, comme évoqué en introduction, les compétences attendues en entreprise pour les scientifiques des données (voir Wikipédia : Science des Données [4] et Wikipedia : *Data Science* [5]) sont sujets à débats et variations d'appréciation. Toutes les possibilités ne se retrouvent pas forcément dans les enseignements de cours de tronc commun proposés, et notamment : nettoyage et préparation des données, échantillonnage, visualisation des données (qui est un enseignement optionnel), programmation web, aspects juridiques, confidentialité et sécurité des données. On devra donc veiller aux retours des entreprises et laboratoires sur l'aspect critique ou non de ces compétences, et quant à la nécessité les développer. Ces retours seront collectés notamment dans le cadre des conseils de perfectionnement des masters MSIAM et MoSIG, mis en place à partir de la rentrée 2016, et qui incluront des représentants d'entreprises et de laboratoires de recherche. La valorisation des données en entreprise avec ses aspects managériaux et stratégiques ne sera a priori pas abordée, dans la mesure où le mastère spécialisé Big Data de Grenoble INP et GEM traite de ces dimensions de la science des données.

Références

[1] Besse, P. et Laurent, B. (2015), De Statisticien à Data Scientist, *<Hal-01205336v2>* http://hal.archives-ouvertes.fr/hal-01205336v2.

[2] Fan, J., Han, F. et Liu, H. (2014). Challenges of big data analysis. *National science review*, **1**(2), 293-314.

[3] Overton, J. (2015). What to look for in a data scientist, http://www.oreilly.com/ideas/what-to-look-for-in-a-data-scientist

[4] Wikipédia, l'encyclopédie libre. (19 nov 2015, 12:23 UTC). *Science des données*. <http://fr.wikipedia.org/w/index.php?title=Science_des_données&oldid=120625264>.

[5] Wikipedia, The Free Encyclopedia. (13 déc 2015, 05:23 UTC). *Data science*. <https://en.wikipedia.org/w/index.php?title=Data_science&oldid=695017731>