Le Mastère Spécialisé Big Data de Télécom ParisTech

Stéphan Clémençon [1]

[1] *Institut Mines Télécom - LTCI UMR Télécom ParisTech/CNRS No. 5141, stephan.clemencon@telecom-paristech.fr*

Résumé. Les espoirs, comme les craintes, suscitées par le Big Data, la perspective d'usages matrisés de megadonnées désormais perçues comme un levier de progrès et d'innovation dans de nombreux secteurs invitent les équipes académiques à définir de nouveaux programmes de formation, interdisciplinaires, associant technique (mathématiques et informatique) et réflexion stratégique (aspects légaux, création de valeur économique, cas d'usage) en collaboration étroite avec l'Industrie et les Services.

Mots-clés. Megadonnées, Interdisciplinarité, Interaction industrie/académie, Science des Données, Apprentissage Statistique

Abstract. Hopes, and fears as well, aroused by Big Data, possible mastered uses of data masses, considered now as a source of growth and innovation in a wide variety of economic and societal sectors should encourage academics to define novel interdisciplinary higher education programs, mixing technique (maths and computer science) and humanities (law, value creation), in closed collaboration with professionals from the sector of Industrie and Services.

Keywords. Big Data, Interdisciplinarity, Interaction Industry/Académy, Data-science, Machine-learning

Big Data, cette expression qui fait florès depuis quelques temps, désigne tout autant la collection de briques technologiques promues par les géants de l'informatique tels Google, Amazon ou Facebook permettant de gérer et d'analyser en temps quasi-réel des données massives de formats divers (par exemple nombres, texte, graphes, images/vidéos, signaux audio) en réponse au challenge des 3V (pour Volume, Vélocité et Variét́), qu'un véritable phénomène sociétal. La formation progressive d'un consensus sur le fait que les données, combinées à l'essor des sciences et technologies de l'information, pourraient jouer dans un futur proche un rle décisif dans la transformation de presque tous les champs de l'activité humaine : commerce, sécurité, santé, science, communication, bâtiments, transports, énergie,... En effet, un nombre croissant d'acteurs de la vie économique ou scientifique partagent aujourd'hui la conviction que les flux de données qui nous sont de plus en plus facilement accessibles du fait des progrès technologiques récents et de l'ubiquité des capteurs dans la société moderne (par exemple systèmes embarqués, téléphone mobile, internet) sont susceptibles de permettre l'optimisation de nombreux processus, tels que la détection de fraude par exemple, ainsi que la création de nouveaux services, à

la personnalisation accrue, à l'instar des moteurs de recommandation proposés par les sites d'e-commerce sur le web. Après avoir passé en revue quelques faits témoignant de l'ampleur du phénomène, du rôle joué par les sciences et techniques de l'information dans de nombreux secteurs et de la nécessité de transformer de nombreux cursus universitaires en y intégrant les potentialités de la science des masses de données (data-science), nous présenterons l'ambition et l'esprit des formations proposées par Télécom ParisTech dans le domaine Big Data, en interaction avec le monde de l'entreprise.

Ubiquité. Dans les secteurs de la défense et de la sécurité, l'actualité récente nous rappelle que les images satellitaires, les données biométriques, les informations générées par le trafic internet, les systèmes de télécommunications et les réseaux sociaux sont d'ores et déjà intensément exploitées par les services de renseignement. En contrepartie, la cybersécurité, la protection de l'information numérique stratégique, devient un enjeu majeur, non seulement dans le domaine de la défense mais aussi dans bien d'autres secteurs d'activité. Dans le domaine de la prévention de la criminalité, les masses de données décrivant les conditions dans lesquelles les crimes et délits se produisent nourrissent aujourd'hui des applications logicielles prédictives ayant permis de réduire le taux de criminalité dans certaines régions du monde avec succès, comme certains romans d'anticipation le promettaient. La santé se voit elle aussi largement impactée par le phénomène Big Data, la médecine se transformant peu à peu en une science de l'information. L'analyse des requtes sur les moteurs de recherche permet d'élaborer des applications comme Google flu pour le suivi en temps réel des épidémies de maladies infectieuses transmissibles, parfois plus performantes que les modèles statistiques épidémiologiques ajustés à partir des données remontées par de coteux réseaux sentinelle . Les outils de bioinformatique accélérant considérablement le séquençage d'un génome, les technologies analytiques (spectrométrie de masse, RMN) permettant d'observer le métabolisme des patients sont sans doute les prémisses d'une médecine individualisée. Avec les progrès en matière de miniaturisation, l'utilisation de systèmes embarqués se voit également généralisée, ouvrant la voie à l'élaboration d'une maintenance prédictive permettant d'assurer un niveau de service optimal dans de nombreux domaines tels que les transports (par exemple aéronautiques, routiers, ferrés), la distribution de l'énergie ou la gestion de bâtiments intelligents . Le commerce fut indéniablement l'un des premiers secteurs bouleversé par la révolution Big Data. Si la relation client est gérée au moyen de techniques quantitatives depuis plusieurs décennies, l'instantanéité des données aujourd'hui disponibles, leur volume et leur diversité l'ont fait récemment considérablement évoluer. Les techniques de filtrage collaboratif permettent en particulier d'ajuster véritablement l'offre commerciale au profil du client, dans le cadre de l'e-commerce tout particulièrement. Ces applications fondées sur des méthodes statistiques parfois très élaborées sont largement utilisées pour la recommandation de produits sur les portails internet commerciaux ou de liens publicitaires sur le web (retargeting) et sont amenées prochainement à tre généralisées dans bien d'autres domaines, pour l'élaboration de polices d'assurance ou de produits financiers personnalisés, voir [1].

Multidisciplinarité. Les mégadonnées , associées à des technologies informatiques et mathématiques en plein essor, sont donc largement perçues comme un moyen de créer de la valeur et de faire avancer la recherche scientifique. L'ère Big Data recèle également de nombreux dangers, au delà des problèmes éventuels posés par un assujettissement de certaines activités critiques à des systèmes d'information et des infrastructures pouvant tomber en panne. En l'absence d'une planification expérimentale préalable et d'un cadre d'analyse statistique rigoureux, l'exploitation des masses de données récoltées à la volée peut en effet conduire à des résultats fortement biaisés, erronés. L'accessibilité accrue aux données personnelles bouscule le concept de vie privée qui prédominait jusqu'alors. Le Big Data appelle ainsi indéniablement de nouveaux profils sur le marché de l'emploi : des hommes et des femmes disposant de compétences techniques pointues dans les domaines de l'informatique et des mathématiques appliquées pour gérer et analyser l'information bien sr, de connaissances métier également, une capacité à anticiper les services et usages rendus possibles par les mégadonnées dans divers secteurs d'activité mais aussi une compréhension des aspects juridiques relatifs à la collecte, au stockage et à l'exploitation des données personnelles. Relever les défis du Big Data requiert de concevoir et proposer de nouvelles formations, permettant aux futurs cadres et décideurs d'acquérir des connaissances générales relatives aux technologies à mettre en ?uvre pour réaliser l'acquisition et l'exploitation des données, à ce que leur traitement statistique rend possible aujourd'hui, ainsi qu'aux dangers afférents. De ce point de vue, l'un des plus grands challenges est l'élaboration d'un curriculum s'affranchissant du carcan disciplinaire et articulant sciences de l'information, business et droit afin de former au métier de data scientist . C'est précisément ce que cherche à réaliser le MS Big Data de Télécom ParisTech.

Data-science. L'analyse statistique des données dans le but d'élaborer des outils d'aide à la décision, baptisée parfois fouille de données (data-mining) ou encore intelligence économique (business intelligence) dans un contexte métier , n'est pas une activité nouvelle. La gestion des risques financiers ou le contrle de qualité dans l'industrie par exemple mobilisent depuis bien longtemps le travail de statisticiens. De telles tâches convoquent des connaissances et des techniques issues de différentes branches des mathématiques appliquées (probabilités et statistique, optimisation, analyse et calcul numériques, traitement du signal entre autres) ainsi que l'utilisation de solutions informatiques très encadrées. Mais il y a encore peu de temps, les données étaient collectées via de coteux plans d'expérience ou des sondages. Elles nécessitaient, du fait de leur rareté, un prétraitement considérable reposant sur l'expertise humaine avant de permettre l'élaboration de modèles statistiques prédictifs. La complexité des données disponibles aujourd'hui (leur très grande dimension, leur volumétrie explosive nécessitant un stockage distribué, leur nature très variable : texte, nombres, graphes, images/vidéos, etc.) et la nécessité de les traiter en temps réel de façon automatique sont à l'origine de l'engouement actuel pour le machine-learning, l'apprentissage statistique. Cette discipline se situe à l'interface des mathématiques appliquées et de l'informatique et a pour but de produire des algorithmes permettant d'apprendre automatiquement des données les représentations ou les modèles

les plus performants. Cette approche couple la formulation statistique des problèmes prédictifs avec des principes algorithmiques puissants et intègre dans la mise en oeuvre les contraintes opérationnelles, relatives à l'accès aux données par exemple. Fort du succès de logiciels fondés sur de tels algorithmes pour la reconnaissance vocale ou de caractères manuscrits par exemple dès le début des années 90, le machine-learning remplace progressivement la statistique traditionnelle dans de nombreux domaines. Le data-scientist n'est donc pas seulement un statisticien mais un expert capable d'articuler des compétences en mathématiques appliquées et en informatique tout à la fois, de façon à appréhender la chane de traitement des données dans toute sa globalité : de l'étape d'acquisition des données à la solution analytique, en passant par les étapes de stockage et de représentation/visualisation. L'innovation technologique semble exiger de plus en plus souvent cette polyvalence et le temps des organisations "en silo" où, par exemple, les services informatiques de l'entreprise transmettait un fichier "plat" au département en charge de la modélisation puis se voyait renvoyer un modèle statistique très parcimonieux encapsulé dans une structure spécifique et à recoder entièrement pour la mise en production semble désormais révolu à l'ère Big Data.

Le Programme du Mastère Spécialisé "Big Data". La formation dispensée en vue de l'obtention du diplôme de MS Big Data délivré par Télécom ParisTech se veut à la fois complète et progressive. La quasi-ubiquité des problématiques Big Data s'accompagne naturellement d'une grande variété des secteurs d'activité concernés et d'une inévitable hétérogénéité des profils des candidats au mastère spécialisé. Trois types de profil ont émergé des deux premières sessions de recrutement : de jeunes diplômés (ingénieurs et/ou diplômés de master en informatique, télécommunications ou maths appliquées), des salariés du domaine de l'IT ayant besoin de monter en compétences et enfin des candidats à une reconversion professionnelle dans un secteur très porteur. Le programme s'articule autour de trois champs disciplinaires et mobilise des enseignants-chercheurs issus de plusieurs départements: le département de traitement de l'image et du signal, le département informatique et réseaux et le département de sciences économiques et sociales. Plus précisément, les enseignements sont répartis sur une dizaine de cours, délivrant des connaissances approfondies sur les thèmes suivants et proposant de nombreuses séances de TP.

Gestion des données. Le but est la maîtrise des systèmes de gestion de données hétérogènes, massives et peu structurées. Partant de concepts et techniques élémentaires relatifs au modèle relationnel et au langage SQL, l'enseignement aborde les notions essentielles de stockage (distribué), d'indexation, d'évaluation/optimisation/répartition de requtes et détaille de nombreuses briques technologiques amenées à devenir d'éventuels standards (par exemple Cassandra, MongoDB , ElasticSearch).

Données du Web. L'objectif de l'enseignement est de permettre la compréhension et l'utilisation des technologies du Web : des techniques de base (par exemple HTML, CSS, JavaScript, PHP), aux méthodes permettant d'exploiter automatiquement les données du Web. Les technologies d'informatique décisionnelle, telles que celles mises en ?uvre par

les moteurs de recherche et de recommandation, seront couvertes en détail : modélisation (Web sémantique, graphe du Web), extraction (wrappers), indexation (langage naturel), calcul à grande échelle (MapReduce).

Machine-Learning. Le programme aborde de nombreux aspects de ce domaine, à l'interface des mathématiques et de l'informatique, dédié à l'élaboration, l'analyse et la mise en oeuvre d'algorithmes permettant à une "machine" d'extraire des informations à partir de données, afin d'accomplir automatiquement des tâches de prédiction, d'aide à la décision ou de représentation efficace des données (indexation, compression). La volonté d'automatisation se conçoit généralement dans des situations o les données à disposition sont massives à tel point que les méthodes statistiques classiques, reposant en partie sur l'expertise et le prétraitement humains, s'avèrent impossibles à mettre en ?uvre et/ou inefficaces. Le machine-learning constitue en effet un véritable corpus de méthodes algorithmiques, pouvant s'adapter à des données de nature différente, sur lequel repose de nombreux systèmes décisionnels. L'enseignement proposé couvre les concepts et techniques essentiels en apprentissage supervisé et non supervisé (théorie de Vapnik, support vector machines, méthodes d'agrégation, modèles graphiques) ainsi que les avancées récentes réalisées dans le domaine, motivées par les problématiques du Big Data : apprentissage distribué, optimisation stochastique et apprentissage par renforcement, apprentissage multitâche, graph-mining, ranking.

Systèmes répartis. Le programme traite en particulier de l'architecture des systèmes répartis et de leurs fonctionnalités (par exemple processus, threads, communications, synchronisation, nommage, répartition des fichiers/données), des grandes tendances en matière de pair-à-pair, de cloud et d'informatique mobile. Il propose une étude détaillée des intergiciels (middleware), des briques technologiques et de l'algorithmique pour la construction de systèmes répartis.

Visualisation. Il s'agira d'enseigner les techniques récentes de visualisation permettent aux utilisateurs de logiciels de mieux comprendre l'information contenue dans les grandes masses de données, ainsi que les règles de décisions complexes fondées sur ces dernières, facilitant ainsi l'interaction entre système décisionnel et utilisateur final.

Sécurité. Le programme couvre à la fois les aspects techniques (sécurisation des OS, des bases de données, des sites web), organisationnels/méthodologiques (évaluation/certification de la sécurité des systèmes d'information) et juridiques (loi sur l'économie).

Au delà de la vision technique (informatique et mathématiques appliquées), la formation propose d'explorer les aspects sociétaux, juridiques (données personnelles, privacy) et économiques, du Big Data.

L'écosystème Big Data. Le Big Data, par son potentiel d'innovation multisectoriel, aura à son échelle un impact certain, forçant l'adaptation, permettant l'émergence ou poussant vers la sortie les acteurs selon leur position et leurs gènes business. Par un mélange de modèles et notions fondamentales et modèles, d'exemples réels et de témoignages de professionnels de cet écosystème, cet enseignement explorera comment les Big Data prennent appui sur l'environnement économique en place pour le modifier.

Données personnelles et économie de l'internet. Il s'agit d'aborder des éléments d'économie de protection de la vie privée, de la réputation et des asymétries d'information ainsi que de valorisation de données sur les moteurs de recherche et les réseaux socio-numériques. Ce cours propose également une étude prospective sur les scénarii possibles autour des données personnelles et des Big Data à moyen et long termes, tant le phénomène est susceptible de faire bouger les barrières légales.

Au delà des compétences reconnues de l'équipe académique de Télécom ParisTech, s'appuyant sur une activité de recherche très compétitive dans les domaines scientifiques afférents, la formation mobilisera des professionnels, issus de secteurs variés (par exemple internet, sécurité, défense, finance, e-commerce, consulting), de grands groupes, de PME innovantes ou de start-ups (celles de l'incubateur de Télécom ParisTech en particulier). Le comité de perfectionnement du MS Big Data compte en particulier des représentants de Thalès, du groupe Safran, de BNP Paribas, d'EADS, de Capgemini, de SAS, de Criteo, d'IBM et de Liligo. L'objectif de la formation étant de garantir l'acquisition d'un socle de connaissances théoriques mais aussi d'un savoir-faire opérationnel satisfaisant, il convient en effet de veiller à ce que les enseignements proposés intègrent les contraintes industrielles et soient étayés par des cas d'études correspondant aux enjeux réels du Big Data aujourd'hui. Les correspondants de nos partenaires industriels interviennent lors de séminaires, de séances de cours ou de travaux pratiques mais participent également à l'élaboration et à l'encadrement de projets fil rouge , réalisés en groupe de 4 à 5 élèves en parallèle avec les cours tout au long du cursus, autour d'une problématique industrielle. Ces projets permettent d'aborder en situation réelle de nombreuses facettes du Big Data : acquisition des données, stockage, solution analytique, visualisation, mise en SaaS, aspects légaux, modèle économique. La synergie entre Télécom ParisTech et le monde industriel s'incarne aussi à travers trois chaires de recherche étroitement liées au Big Data: "Machine-Learning for Big Data", "Market Insights and Big Data" et "Valeurs et Politiques des Informations Personnelles". Cette interaction très forte entre industrie et académie autour d'enjeux majeurs pour l'innovation assure la pertinence de cette formation.

Bibliographie

[1] S. Clémençon (2014), Les métiers du Big Data, Revue Documentaliste-Sciences de l'Information.
[2] Chaire Machine-Learning for Big Data: http://machinelearningforbigdata.telecom-paristech.fr/fr/