TELECOM ParisTech

TELECOM Evolution

Une école de l'IMT

DATA SCIENTISTS !

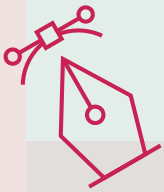
Les métiers qui façonnent les transitions vers demain

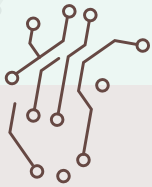
UPWARD DATA

AIRBUS

Sommaire

Fiches Métiers





Fiches pratiques

Avant-propos

Chère lectrice, cher lecteur,

Si vous lisez ces lignes, c'est sans doute parce que la science des données ou le big data ont éveillé votre curiosité. Quoi de plus naturel, tant ce vaste champ d'investigation scientifique a aujourd'hui envahi notre quotidien. Mon souhait est que, au fil de ces pages, vous passiez de l'interrogation à la vocation !

Tous les témoins de ce livre, qu'ils soient enseignants-chercheurs, anciens élèves, professionnels de la donnée au sein de grandes entreprises ou de start-up, vous le diront : la data science est avant tout une passion. Par bonheur, elle offre de nombreux métiers enthousiasmants, au cœur des enjeux de la transition numérique.

Télécom ParisTech, première grande école française d'ingénieurs dans le domaine du numérique, a demandé à une trentaine de professionnels de donner leur vision, leurs impressions de parcours et leurs conseils. Au-delà de la culture de la donnée qu'ils partagent tous, leur forte proximité avec les métiers de l'entreprise montre combien leur rôle est ancré dans la réalité économique.

Je voudrais à présent remercier tous les contributeurs de ce livre, ainsi que l'ensemble des équipes de l'école qui ont réussi à créer une dynamique exceptionnelle associant la recherche, la formation et le monde de l'entreprise. Dans le domaine du big data, l'interaction entre ces trois pôles est en effet particulièrement éloquente, comme vous allez le découvrir.

Cette année 2017 apportera son lot d'interrogations et de nouveaux défis. Le règlement général sur la protection des données personnelles (RGPD) incite déjà les entreprises à pouvoir « rendre des comptes ». Les réussites brillantes, comme les échecs, de start-up dont le modèle économique repose sur les données, ne cessent d'alimenter des débats passionnants sur l'ubérisation et l'intelligence artificielle.

Que vous vous destiniez à devenir data scientist, envisagiez une thèse de doctorat ou désiriez créer votre start-up, je suis persuadé que ce livre saura guider vos choix et vous ouvrir de nouveaux horizons. Et si vous êtes amenés à recruter ou, tout simplement, à travailler avec des professionnels de la donnée dans votre entreprise, il vous éclairera certainement sur ces métiers dont on parle tant.

Yves Poilane
Directeur de Télécom ParisTech

L'ère numérique guidée par la donnée

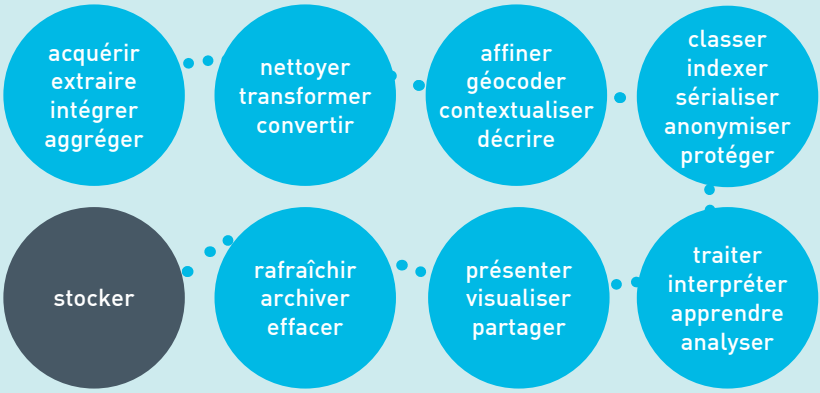
S'il existe un domaine qui a provoqué, puis accéléré, la transition numérique en cours, ce n'est pas tant le déploiement des réseaux ou la convergence des médias au début des années 2000, mais bien l'abondance de la donnée, produite par notre utilisation effrénée des réseaux et des objets connectés. Depuis 2012, année où le « *big data* » a été choisi comme terme tendance dans le monde du numérique, la situation a fortement évolué. L'utilisation de telles *données massives* n'était cependant pas inconnue auparavant des directions des systèmes d'information, et c'est la soudaine production de nouvelles données de toute nature, la mise à disposition en *open data* de nombreuses sources de données externes et la facilité avec lesquelles il était possible de les manipuler, les assembler, et effectuer des calculs avec, qui a provoqué la création de services innovants. La transition numérique de la société a pu commencer sérieusement, guidée par les sciences de la donnée. Le terme *data-driven economy* faisait son apparition.

Nature des données

Il est habituel de présenter les données massives à travers quelques-unes de leurs caractéristiques fondamentales, qui commencent toutes par la lettre V. Les trois principales et immédiates sont leur *volume*, la *vitesse* à laquelle elles sont produites, captées, consommées, et leur *variété*, les données à traiter étant souvent non structurées, composées de textes, d'images, de suites de chiffres... Leur *variabilité* caractérise leur propension à changer de format ou de structure au cours du temps, par l'ajout de nouveaux champs ou par l'intégration de données similaires améliorant leur diversité.

Les données en soi sont des *faits bruts*, et pour créer de l'information, puis de la connaissance, elles doivent être interprétées. L'intérêt des données réside alors dans leur *valeur* intrinsèque qui émerge des traitements qu'on leur fait subir, des simples statistiques à l'apprentissage machine (*machine learning*) plus évolué.

De l'acquisition à l'utilisation, les données subissent de nombreux traitements



Déverrouiller la valeur tapie au cœur des données ne sert cependant pas à grand-chose si leur mise en contexte n'est pas respectée et si ce qu'elles ont à dire n'est pas bien mis en lumière. *Valoriser*, rendre *visibles* et permettre de *visualiser* les données sont une même démarche qui consiste à rendre les données compréhensibles, interprétables, utiles, partageables, vivantes. Reste une condition : que ces données et leurs conséquences soient conformes à la vérité et aux faits. Ce souci de *véracité* des données est un marqueur fort à préserver. Ajoutons que les données ne sont *jamais neutres* : de nouvelles combinaisons de données peuvent créer de nouvelles connaissances aux conséquences difficilement prévisibles.

Culture de la donnée

Les *data scientists* – un terme large qui se décline en une multitude de métiers, évolutifs – sont les femmes et les hommes qui naviguent dans les données au quotidien. Le volume de ces dernières n'est pas toujours leur caractéristique majeure. Les *open data*, ces informations publiques librement accessibles et réutilisables, désignant également des données d'acteurs privés qu'ils libèrent dans les mêmes conditions, sont un type de données vecteur d'innovations et d'enrichissement des données que chaque acteur produit. Le caractère *personnel* de certaines données est une qualité qui nécessite une attention particulière et une éthique de tous les instants. La place déterminante qu'elles ont prise dans nos sociétés rend le développement de la culture des données –comprendre leur rôle et se donner les moyens d'agir avec– essentiel à double titre : l'examen critique de leur place au cœur de nos échanges et de nos décisions, et le renouvellement de nos façons d'agir et d'interagir. Moteur de la transition numérique et des autres transitions en cours, porteuse d'enjeux juridiques et sociétaux, non bridée par la technologie pour l'instant, la donnée est à la fois une affaire de technicité et d'humains, et les data scientists en sont les artisans et le liant.

Une stratégie de la donnée, en France et en Europe

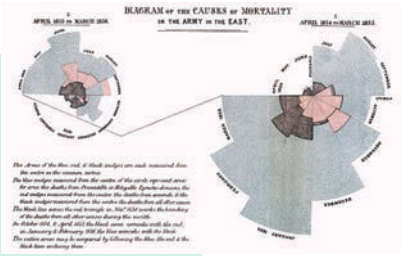
Lancé en France en 2013, le Plan Big Data a affirmé la donnée comme une des priorités nationales, avec l'objectif de devenir une référence mondiale, et l'ambition de créer 10 000 emplois directs d'ici 2019, notamment via la formation des data scientists. En 2015 le programme Industrie du futur a défini trois priorités liées au big data : l'économie des données, les objets intelligents et la confiance numérique. Les données sont également l'une des grandes orientations de la stratégie de recherche France-Europe 2020. Issue du Plan d'investissements d'avenir, la première plate-forme big data sponsorisée par l'État français, Teralab, dotant l'innovation, la recherche et l'enseignement d'importantes capacités de traitement big data, a reçu le label « Silver i-Spaces » par la *Big Data Value Association* fin 2016, label détenu par seulement trois plate-formes en Europe. Teralab est pilotée par l'IMT et le GENES, en partenariat avec l'INSEE.

Dans le monde entier, tous les domaines d'activité sont touchés par l'économie de la donnée, chacun à des stades différents. Des femmes et des hommes data scientists racontent à présent leur métier passionnant, dans les secteurs historiques que sont la santé et les transports.

L'enjeu des données de santé

Quand en 2008 Google révèle qu'il semble capable de prédire les lieux d'épidémie de la grippe deux semaines avant les centres nationaux de suivi des maladies, à partir de corrélations via les requêtes faites sur son moteur de recherche, une prise de conscience sur le tour que pourrait prendre les flux massifs de données appliquées à la santé s'opère. En exploitant des jeux de données nouveaux, provenant des comportements de la multitude, il semble soudain envisageable de prédire le futur, dans une certaine mesure. Avec les objets connectés, c'est également l'accès à des données relatives au bien-être (sommeil, alimentation, gestion des moments de sport) qui devient possible. Enfin, les progrès dans l'analyse d'images et d'autres signaux d'origine médicale, et le recours à des systèmes d'intelligence artificielle, permettent de mieux détecter certaines pathologies.

Les données de santé ne sont pas des données comme les autres. Elles font l'objet d'une définition officielle dans le Règlement européen 2016/679 : il s'agit d'une donnée médicale ou relative aux déterminants généraux de la santé, se rapportant à l'état de santé d'une personne, qui révèle des informations sur sa santé physique ou mentale passée, présente ou future, y compris des informations relatives à son enregistrement pour la prestation de services de santé, ou obtenues lors de tests ou d'examens, et d'autres types d'informations de santé quelle qu'en soit la source. Ces questions règlementaires et le grand nombre d'organismes de validation ou de contrôle existant, les questions de sécurité et de protection de ces données, les questions d'éthique associées, font de la santé un secteur où la variété d'acteurs intervenants doit gérer les données avec soin.



Pionnière de l'utilisation de données de santé, Florence Nigthindale dirige une équipe d'infirmières pendant la guerre de Crimée et constate dans les hôpitaux des conditions horribles de mauvaise hygiène, de manque de ressources et de désorganisation du personnel et des dossiers médicaux, notamment lors des transferts de malades. Elle recueille des données pour les analyser, et à son départ en 1856, les conditions dans les hôpitaux se seront considérablement améliorées, avec des taux de mortalité tombant de 42% à 2%. L'analyse de ces données avait démontré les vraies causes de la mortalité – les conditions de vie insalubres et non pas le manque de nourriture. Elles montraient également que les hôpitaux civils étaient concernés pour les mêmes raisons. Pour convaincre la reine Victoria et le parlement d'enclencher des réformes, Florence Nigthindale crée une représentation graphique nouvelle, un diagramme polaire dit *coxcomb*, dont les multiples dimensions véhiculent et démontrent mieux les faits.

Objectif Data Driven 2020

Yoann Janvier
Lead Data Scientist
IPSEN

Yoann Janvier est Lead Data Scientist chez Ipsen (Euronext: IPN; ADR: IPSEY), groupe pharmaceutique international de spécialité. Le département auquel il est rattaché est dirigé par un vice-président big data & analytics et est composé d'un directeur des données, d'un chef de projet big data et d'un business analyst. Yoann quant à lui est en charge du Data Lab avec une petite équipe de data scientists externes. Il pilote et réalise des projets exploratoires avec l'objectif de création de valeur à partir de la donnée. Les méthodologies agile et *test and learn* sont appliquées tout au long de la chaîne : collecte et exploration des données, recherche de caractéristiques pour alimenter les algorithmes d'apprentissage machine, data-visualisations…

« Un mix entre data scientist et manager »

Les données proviennent du système d'information, de la R&D, du web, ou sont acquises en externe. Leur volumétrie reste faible mais de futurs projets sur les données cliniques et génétiques, et des projets de maintenance prédictive avec des flux de données de capteurs, vont significativement accroître les volumes. L'exploitation de ces données très variées, structurées ou non (bases de données, articles scientifiques, tweets…) nécessite une intelligence algorithmique importante. Enfin, l'un des objectifs poursuivis est de révéler des informations cachées dans les données (segmentation marketing par exemple) et de faire de la prédiction (par exemple sur l'épidémie de gastro-entérite).

En 2017 un projet big data d'industrialisation a été lancé avec pour objectif de passer à l'échelle en terme de valeur apportée et de volume d'utilisateurs cibles. La nouvelle plate-forme rendra plus robuste la collecte des données, les développements exploratoires et l'exposition des données. Ce type de plate-forme aidera à mettre en œuvre des projets plus complexes, en particulier avec le traitement de données en temps réel. « *Cette Data Factory est un guichet unique pour toutes les initiatives exploratoires et pour manipuler des données quotidiennement. Les données y sont exposées et accessibles via des API, internes et externes. Par analogie, notre Data Factory est une bibliothèque où chaque nouveau service est un nouveau livre.* » Ipsen évalue également d'autres technologies innovantes tout au long du parcours de soins du patient : remontée de données d'essais cliniques avec les objets connectés, gestion de la douleur ou programme de rééducation par la réalité augmentée, avec des systèmes d'intelligence artificielle.

linkedin.com/in/yjanvier/fr
@yoannjanvier

Retrouvez Yoann Janvier page 8

Explorer les données de transport

Avec les données de santé, les données de transport sont parmi les premières à avoir été largement utilisées pour créer des preuves de concept de services innovants à l'ère numérique. À l'occasion de nombreux hackathons, et en utilisant des ensembles de données de plus en plus ouvertes, les amoureux de la donnée de transport et de la donnée géographique ont imaginé des services d'aide à la multi-modalité, de suivi et de visualisation de trajets, ou encore de covoiturage. Côté matériel, les véhicules autonomes et les drones ouvrent de nouvelles perspectives. Enfin, c'est de ce secteur qu'est né le terme « *ubérisation* ».

Le secteur du transport est un domaine où l'on commence à avoir suffisamment de recul, et s'il reste encore beaucoup à explorer, de nombreux chantiers sont passés en phase industrielle, ce qui offre un panorama assez large des projets de data scientists.

Améliorer la qualité des produits

Data scientist pour le véhicule connecté et expert en analyse de données, Alain Abramatic est un ancien élève de Télécom ParisTech qui est entré chez PSA Peugeot Citroën en 1989 après dix années passées chez Schlumberger. Cet ingénieur, attiré par l'analyse de données dès sa formation initiale, a été manager et expert dans plusieurs services du constructeur automobile, avant que son parcours ne l'amène à manipuler des données de transport sur l'ensemble de la chaîne de transformation de la data.

Les informations traitées chez PSA sont de nature très variée : données structurées collectées dans les véhicules (avec l'accord du client), bases de données internes (fabrication, garantie, diagnostic), verbatims associés, données issues de la relation clientèle… Un exemple parmi d'autres : une étude sur la consommation de carburant des véhicules.

« *Mon travail consiste à valoriser les données des véhicules connectés en développant de nouveaux services à valeur ajoutée pour l'ensemble de nos clients, particuliers ou professionnels, et de nos partenaires.* » Ceci passe par le croisement des informations et la mise en œuvre de techniques d'apprentissage machine. Dans ce domaine, et à condition d'avoir les bons capteurs, l'internet des objets est nécessaire pour obtenir de nouvelles données. Il doit cependant apporter une réelle plus-value, et l'utilisation d'intelligences artificielles est indispensable pour cela. Le risque de rejet d'un nouvel usage est une boussole pour le data scientist.

Ce travail se fait dans le respect des exigences réglementaires de chaque pays où les données sont collectées. La maîtrise des volumes et la qualité, l'optimisation des traitements, la pertinence des résultats obtenus sont fondamentaux.

Accélérer la démarche orientée données

Angélique Bidault-Verliac
Directrice Datascience & Connaissance Client
Voyages-sncf.com

www.linkedin.com/in/angeliquebidault
@Bidaultverliac

Les projets autour des big data au sein du marketing chez Voyages-sncf.com ont débuté en 2013. Ils concernaient à l'origine les données de navigation des clients, qui provenaient du site web et de l'application mobile. L'objectif était de les aider à trouver ce qu'ils cherchaient et à revenir vers eux avec des suggestions de destination. En quelques années, les data scientists ont beaucoup appris, notamment sur le soin à apporter dans le nettoyage préalable des données et le traitement des cas particuliers, pour éviter que la chaîne de données ne se rompe en production. « *Ces constats ont incité les équipes à s'organiser différement entre data scientists et unités métier, afin de rendre plus fluides leurs relations.* » Depuis les premiers prototypes de services en 2013, l'accent est donc mis sur l'organisation et l'industrialisation. Les équipes travaillent de telle manière qu'en cas d'absence d'un des leurs, chacun connaissant le projet de l'autre peut continuer l'activité globale.

« Le marketing vit sa transformation grâce à la donnée »

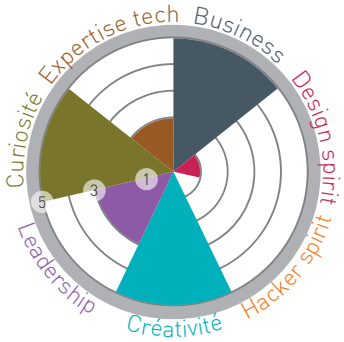
La période actuelle, très motivante, consiste à accélérer cette démarche *data driven*. Le maître-mot est de démocratiser la donnée. Un programme est mis en place pour rendre autonomes les personnes qui ont besoin de données de comportement clients anonymisés dans l'entreprise. Le premier chantier consiste à former ces personnes aux agrégats clients (segmentation RFM, flag clients acquis...) et le second, mené en parallèle, à mettre à disposition d'un plus grand nombre toutes les analyses statistiques réalisées. « *L'objectif est de libérer les data scientists et les data miners, et placer tout ce qui relève du reporting directement dans les équipes opérationnelles.* » Cela change le périmètre d'action des commerciaux, accédant à présent dans leur quotidien à toutes les études data en interne et qui, sensibilisés à l'importance des données, peuvent devenir sponsors de futurs produits.

Un autre enjeu important est celui de la veille technologique. « *De nombreux outils d'apprentissage machine sortent régulièrement. Les data scientists doivent s'emparer de ces outils, et avec leur appui un membre de l'équipe spécialisé en innovation orchestre cette veille.* »

Retrouvez Angélique Bidault-Verliac page 9

Data & marketing

Grands consommateurs de données à des fins d'analyse et de segmentations de plus en plus fines, les métiers du marketing ont été dans les premiers à s'emparer des big data pour y chercher de nouvelles sources de connaissance des clients. Les profils sachant allier la maîtrise de la donnée et celle du marketing sont fortement recherchés par les annonceurs, les agences et les équipes marketing au sein des entreprises. Leur rôle est de recueillir, analyser et mettre en perspective les données issues des parcours des clients, tant sur le web et les mobiles, qu'en offline et cross-canal. Ces *digital analysts* effectuent du reporting et doivent être capables de chercher à comprendre les comportements des clients et d'être force de propositions et de recommandations pour les évolutions des sites web et des applications. Ils sont également aujourd'hui amenés à utiliser les données en provenance ou à destination des agents conversationnels et chatbots.



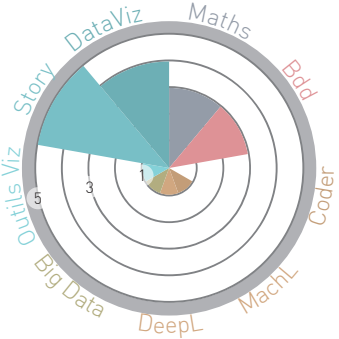
Connaître pour anticiper

Profil Formation généraliste commerciale

Le parcours idéal commence en agence ou chez un acteur du web à fort trafic, l'évolution vers des responsabilités de plus en plus large se faisant rapidement

Compétences **Outils :** solutions logicielles de suivi de trafic

Qualités : parfaite connaissance des outils digitaux, des évolutions des usages et des pratiques des consommateurs ; capacité d'interprétation, d'analyse stratégique et de synthèse ; sensibilité à l'ergonomie et à l'optimisation des sites et applications



Chez Voyages-sncf.com, la donnée ne relève pas uniquement du marketing. On trouve également celle des équipes commerciales, des équipes produits, les données transactionnelles, les achats (voir ci-contre), les comptes client... Cependant, compte-tenu du métier principal de Voyages-sncf.com, qui est l'expert de la distribution du train et de la destination France, filiale du Groupe SNCF, la donnée marketing reste un bien central de l'entreprise.

Le travail sur les destinations est un bon exemple de ce que l'analyse des données a apporté. « *Nous avons collecté des informations de navigation qui se rapportent notamment à la destination recherchée par l'internaute. Mes équipes ont créé des algorithmes de recommandation pour identifier les destinations auxquelles vous pouvez être sensible. Ainsi, chaque newsletter est personnalisée. Nous avons 13 millions de visiteurs uniques, ce qui fait que nous avons suffisamment de données à disposition.* » explique Angélique Bidault-Verliac. Ces données sont couplées à des informations météorologiques et à une bibliothèque d'images –et leurs attributs– de destination autour d'une ville.

1 million d'appels par mois

760 millions de recherches par an sur voyages-sncf.com

En utilisant des techniques d'A/B testing, les newsletters personnalisées sur la destination ont permis une augmentation de 20% de volume d'affaires par e-mail et une hausse de 25% du taux de clics.

Une organisation en Feature Team

« *À une époque nous avions une équipe de six data scientists et six data ingénieurs sur un chantier d'optimisation d'achats media, en plus d'une dizaine d'autres projets. Cette organisation n'était pas optimale car les personnes travaillaient de manière isolée. Pendant une semaine, un data scientist et une personne du marketing relationnel se sont retrouvés pour lister leurs besoins en ressources sur ce chantier, et voir quel type d'équipe proposer.* »

Le résultat a été la proposition d'un chef de projet marketing relationnel, ne connaissant pas nécessairement les data sciences, trois data scientists, un ingénieur data et un développeur. Se voyant à présent tous au quotidien, les data scientists appréhendent bien mieux les enjeux et besoins du marketing. « *Nous déclinons ce schéma des* feature teams *sur d'autres domaines, comme l'acquisition et les cibles prioritaires. C'est la création d'équipes auto-organisées et multidisciplinaires.* »

Des équipes en évolution

Le management des équipes s'en est trouvé transformé. Chacune d'entre elles gère à présent sa propre feuille de route, et Angélique Bidault-Verliac s'assure que les projets avancent sans difficulté. Cette autonomie gagnée lui a donné plus de temps pour s'impliquer sur les méthodes et pour effectuer un coaching adapté aux besoins et aux rythmes des différentes personnes. « *Des communautés de data scientists sont créées pour qu'ils puissent échanger, qu'ils voient de nouvelles perspectives et ne s'ennuient jamais. Il y a des instances de partage régulier, dont il faut trouver le bon rythme. L'évolution des équipes et celle des carrières sont liées.* »

Des métiers en évolution

Si les équipes de data scientists sont aujourd'hui en évolution, c'est également parce que les métiers de data scientist ont subi leur propre évolution, grâce à l'expérience des premières années de data science dans les entreprises et dans les organismes d'enseignement supérieur et de recherche.

Ardent créateur de liens entre l'enseignement, la recherche et les entreprises, Stephan Clémençon est professeur à Télécom ParisTech depuis 2007, qu'il a rejoint pour y développer la recherche et l'enseignement du machine learning. Responsable de la première chaire de recherche en big data (*Machine Learning for Big Data*, page 57), il est également en charge du programme de Mastère Spécialisé® « Big Data », et a conçu le Certificat d'études spécialisées « Data Scientist ».

« *Auparavant* », rappelle-t-il, « *les entreprises effectuaient des recrutements big data sans discerner s'il s'agissait d'infrastructures big data, de manipuler des technologies telles que Casandra, MongoDB, des bases de données, des graphes, faire du machine learning ou de la data visualisation, ou encore avoir quelqu'un qui mette en place un lac de données –les technologies et leurs termes évoluent également… Tous étaient recrutés sans distinction, sous la même étiquette big data.* » Maintenant les entreprises commencent à identifier véritablement leurs besoins : « *s'agit-il d'infrastructures, d'apprentissage machine ou d'exploitation des données ? Sous le vocable data scientist aujourd'hui, nous cherchons des profils avec une dimension machine learning et analyse statistique des données, souvent appelé analytics.* »

Selon la maturité des entreprises en matière de culture de la donnée, qu'il convient de jauger, et bien sûr du type de données qu'elles utilisent, les profils de poste peuvent encore présenter sous un même vocable des réalités différentes, et les collaborations des uns avec les autres en découlent. Chez Voyages-sncf.com par exemple, les data scientists ont un socle mathématique très poussé, tandis que les ingénieurs big data sont plus orientés informatique, s'occupant de la collecte de la donnée et créant effectivement le code. Certes, les data scientists codent (en Python), mais pour passer en mode production, et industrialiser, leur code est revu par les ingénieurs big data, avec une réflexion sur le cadencement pour que le système global soit robuste.

Les passerelles existent entre les métiers, ou vers ces métiers, et c'est le rôle des responsables d'équipes et des chief data officers (voir ce métier pages 40–41) de faire progresser les personnes selon leur appétence, et d'identifier celles qui aiment la donnée pour les former vers ces métiers, en partant de leur connaissance antérieure de l'entreprise.

Une explosion des méthodes

Les data scientists en poste doivent également s'adapter à la multiplication des nouvelles solutions technologiques.

« *Dans le machine learning existe un cycle vertueux entre les applications et la théorie»,* poursuit Stephan Clémençon. *« Les méthodes sont requises par les applications et les solutions sont apportées par les praticiens et ensuite revisitées et améliorées par les mathématiciens. Attention, car les effets d'annonce ne reflètent pas toujours la réalité scientifique. Apprentissage par renforcement, apprentissage sur des séries temporelles ne peuvent pas se traiter de façon simple et naïve, sans un fort background en mathématique sous peine de réinventer la roue. En formation, j'alerte les futurs data scientists opérationnels sur le fait que la discipline va bouger et que c'est important de ne pas apprendre les techniques de l'état de l'art comme étant figées. Le panorama des méthodes aura largement changé dans quelques années simplement car la technologie change et le type de données auquel on est confronté, avec l'internet des objets en particulier, n'est pas le même que celui avec lequel on a pu faire la reconnaissance de formes il y a quelques décennies.* »

Et les liens tissés entre chercheurs et entreprises profitent également aux élèves.

Se former en proximité avec les entreprises et avec la recherche

Le dynamisme de la filière big data découle d'une recherche pluridisciplinaire stratégique et unique en Europe. Cette filière s'articule autour de nombreuses thématiques : graph-mining et exploration des réseaux sociaux, ranking et filtrage collaboratif, détection d'attaques et d'anomalies, mathématiques financières, maintenance prédictive, ciblage marketing, indexation et recherche de documents multimédia, reconnaissance de sons et d'images… Elle interroge aussi les aspects juridiques, économiques, politiques et philosophiques en relation avec l'utilisation des données personnelles.

À Télécom ParisTech, la proximité avec les entreprises s'incarne par le choix des intervenants dans les formations big data : « *Stéphane Gentric (Research Unit Manager dans le groupe Morpho (Safran)) peut donner des cours sur le deep learning, et notamment sur l'outil TensorFlow qu'il pratique et que l'on ne retrouve pas dans un laboratoire par exemple.* » Les intervenants sont soit issus du monde professionnel, avec une forte compétence opérationnelle et proches des besoins des entreprises, soit des enseignants chercheurs qui offrent une vision à long terme de leurs disciplines, et continuent à faire progresser le socle de connaissances qu'ils transmettent. L'enseignement effectué est aussi une initiation à la recherche et offre la capacité aux élèves de lire des articles de recherche et de continuer à progresser avec les bases dont ils disposeront. « *Aujourd'hui, on va être amené à traiter des flux de données, des données hétérogènes échantillonnées de différentes façons, sous des contraintes de mémoire, de traiter du quasi temps réel, qui sont très différentes de ce qu'on a connu. Ces méthodes sont en cours d'élaboration, et c'est pour cela que la recherche est également totalement indissociable de la formation.* »

Cette vision à long terme du domaine renforce les thématiques d'actualité, traitées via des projets fil rouge et des séminaires, qui sont autant d'opportunités pour les étudiants de rencontrer les entreprises et d'être au plus près de leurs besoins.

Des chaires et des liens

Télécom ParisTech, en partenariat avec des entreprises, et avec le soutien de la Fondation Télécom, anime trois chaires de recherche et d'enseignement big data, au sein de l'Institut Mines-Télécom.

Les chaires *Valeurs et Politiques des Informations Personnelles* et *Machine Learning for Big Data* sont présentées en pages 34 et 57 respectivement.

Talel Abdessalem, responsable de la chaire *Big Data & Market Insights*, explique l'intérêt d'une chaire pour les entreprises : « *Dans une chaire existe une certaine flexibilité pour l'équipe de recherche, car la nature même des chaires, le mécénat, fait qu'il n'y a pas d'obligation de transfert. Les comités de pilotage et les comités opérationnels orientent le travail de l'équipe de recherche vers des sujets concrets pour les entreprises. Il y a une proximité avec les équipes de l'entreprise, qui suivent le déroulement de la recherche, voient les résultats qui se dessinent, et apprennent des choses, y compris si rien n'aboutit directement. La phase de transfert, c'est-à-dire concrètement produire quelque chose qui est transférable pour l'entreprise, peut toutefois arriver par la suite.* » Et comme pour les liens entre data scientists et métiers de l'entreprise, les chercheurs de la chaire voient comment leur problématique de recherche sera utile aux personnes avec qui ils discutent dans les entreprises. « *Les financements permettent également de soutenir une activité de recherche fondamentale, dont la face applicative, l'utilité, n'est pas forcément visible tout de suite et le sera à long terme. Les entreprises voient un double intérêt à la chaire : une utilité directe, et un moyen de financer des recherches fondamentales sur le laboratoire, sur la recherche en France et sur le développement de la science en général, dont elles pourront profiter également.* »

Chaire Big Data & Market Insights

Créée en 2014 avec le soutien de la Fondation Télécom et financée jusqu'à fin 2016 par quatre entreprises partenaires, Deloitte, Groupe BPCE, Groupe Yves Rocher et SNCF, la Chaire « Big Data & Market Insights », portée par la professeur Talel Abdessalem, regroupe des chercheurs spécialisés dans la gestion et la fouille de données massives, l'extraction de connaissances à partir du web et l'analyse de réseaux sociaux.

Cette chaire est partie du constat que de plus en plus d'entreprises disposent de masses de données relatives aux consommateurs, hétérogènes, évolutives, provenant de multiples sources internes ou externes – capteurs, réseaux sociaux, transactions en ligne, traces laissées sur le web… Leur prise en compte efficace dans les processus métiers représente dès lors autant de défis technologiques.

Devenir professionnel de la data

Télécom ParisTech est l'une des premières écoles d'ingénieurs à s'être investie dans le big data avec la création d'un Mastère Spécialisé® en septembre 2013. Cette formation pluridisciplinaire débouche sur un savoir-faire opérationnel et prépare à l'ensemble des métiers dans le domaine de la science des données. Elle couvre aussi bien les aspects techniques que les aspects transverses. Son programme évolue chaque année, selon les retours d'expérience des diplômés et les recommandations des entreprises. *Voir l'enquête insertion pages* 16 17.

L'école délivre également, via Télécom Evolution, un Certificat d'Études Spécialisées « Data Scientist », destiné aux professionnels souhaitant accroître leurs compétences. Très opérationnelle, la formation permet la maîtrise des techniques de gestion et d'analyse des big data et des principaux algorithmes de machine learning. Son objectif est l'obtention d'un savoir-faire technique, à l'interface des mathématiques appliquées et de l'informatique. *Voir l'enquête insertion pages* 18 19.

Co-responsable du Master 2 « Data Sciences » de l'Université Paris-Saclay, co-habilité avec l'École polytechnique et Télécom ParisTech, l'ENSAE ParisTech et l'Université Paris Sud, Florence d'Alché-Buc est à l'origine, et principale organisatrice, de la première « *Junior Conference on Data Science* » qui s'est déroulée en septembre 2016.

« *Cette première édition de la Junior Conference sur la science des données s'adressait aux étudiantes et étudiants de première année de thèse, de master et de deuxième ou troisième année d'école d'ingénieur à l'Université Paris-Saclay. Cet événement leur a fourni l'occasion de présenter les travaux issus de leurs années d'études, et de développer leur sens critique à travers une véritable conférence. Parmi les invités se trouvaient Hervé Jégou, chercheur en intelligence artificielle chez Facebook, Patrick Valduriez de l'Inria, Gabor Lugosi de l'Université Pompeu Fabra (Barcelone) et Isabelle Guyon, ingénieure en machine learning.* »

Florence d'Alché-Buc est également co-responsable de la nouvelle chaire d'enseignement en science des données. Cette deuxième chaire d'enseignement de Télécom ParisTech, conclue avec Bearing Point, vise à développer la formation en sciences des données pour les élèves ingénieurs et étudiants en Mastère Spécialisé® Big Data. La chaire étudiera particulièrement l'impact du big data sur la stratégie et la transformation du modèle économique des grandes entreprises.

MOOC Fondamentaux pour le Big Data

Premier contact avec les Sciences de la donnée

Premier contact ou reprise de contact avec le monde des données, les MOOC offrent une solution flexible, accessible et compatible avec une activité professionnelle, permettant d'apprendre à son rythme. C'est également l'occasion de discuter avec une grande variété d'autres apprenants, et une solution idéale pour pouvoir situer son appétance à aller plus loin dans un domaine d'actvité.

Télécom Evolution propose sur 7 semaines le MOOC « Fondamentaux pour le Big Data ». S'adressant à un public ayant des bases en mathématiques et en algorithmique (niveau L2 validé), il permet un rafraîchissement de ces connaissances pour suivre des formations en data sciences et en big data. Les compétences visées constituent un préalable indispensable dans les domaines de l'analyse, de l'algèbre, des probabilités, des statistiques et des bases de données.

Le MOOC, qui se termine par un quizz validant les acquis, se compose de sept parties : programmation Python, limites des bases de données relationnelles, algèbre, analyse, probabilités, statistiques et d'un exemple de classifieur, le perceptron. Six heures de vidéo ont également été produites en appui des cours.

Ce MOOC peut être suivi en préparation du Mastère Spécialisé® « Big data : Gestion et analyse des données massives », du Certificat d'Études Spécialisées « Data Scientist » et de la formation courte « Data Science : Introduction au Machine Learning », proposés par Télécom ParisTech et Télécom Evolution.

Le MOOC « Fondamentaux pour le Big Data » est classé **6e**
sur la plate-forme France Université Numérique.

Des formations proposées par Télécom Evolution

Télécom Evolution est la marque de formation continue des 3 grandes écoles d'ingénieur : IMT Atlantique, Télécom ParisTech et Télécom SudParis. Spécialisé dans le domaine du numérique, il conçoit et produit des solutions de formation innovantes. La valeur des formations, certifiantes, en inter-entreprises ou élaborées sur mesure, vient de l'attention portée aux besoins réels des entreprises, avec qui Télécom Evolution travaille en étroite collaboration.

www.telecom-evolution.fr/fr/domaines/big-data

Ons Jelassi est enseignante – chercheuse à Télécom ParisTech et coordinatrice du MOOC « Fondamentaux pour le Big Data ». Ses travaux de recherche portent sur le passage à l'échelle des algorithmes d'apprentissage et de prédiction sur les données massives. Elle rappelle que ce MOOC a été conçu au départ pour préparer les personnes désireuses de suivre des formations big data plus poussées. « *Le sujet des données impactant tous les secteurs d'activité, nous faisions face à une très forte demande pour suivre nos formations diplômantes. Il était nécessaire de construire un parcours préalable pour leur permettre de se préparer et de valider le niveau de connaissances nécessaires.* »

La première session attire un peu plus de 7400 personnes, et 293 décrochent leur attestation de réussite, « *ce qui nous mettait dans la bonne moyenne des MOOC où l'on observe un taux de complétion de 5%. C'était même un résultat plutôt élevé, compte-tenu du niveau technique à atteindre sur 6 semaines à l'époque, avec en plus la nécessité de valider un projet.* » Avec la sixième session qui s'est terminée en avril 2017, ce sont environ 33400 personnes qui ont débuté ce MOOC, et près de 1750 qui ont obtenu l'attestation de réussite. À chaque session une communauté se construit sur les forums de discussion (156 fils différents lors de la dernière session), facilitant l'entraide et la constitution des premiers réseaux pour les futurs data scientists, les échanges se faisant ou non en présence des formateurs. Ces derniers sont la force de ce MOOC, car « *Télécom ParisTech réunit un écosystème riche, avec des chercheurs en droit, en sciences économiques et sociales, dans les réseaux, dans l'informatique et dans les mathématiques appliqués, et dans des domaines connexes comme l'Internet des objets, les transports intelligents ou la cybersécurité.* »

Télécom Evolution propose également dix stages de 1 à 5 jours permettant un focus sur des compétences précises. Les deux premiers offrent un panorama plus général et s'adressent à un public de non spécialistes.

Big data : premiers succès et retours d'expérience *(tout public)* // Big data : enjeux stratégiques et défis technologiques *(tout public)* // Big data : panorama des infrastructures et architectures distribuées // Data science dans le Cloud : big data, statistiques et machine learning // Data science : introduction au machine learning // Introduction à la sécurité du big data // Visualisation d'information (InfoVis) // Moteurs de recommandation et extraction de données du Web // Text-Mining // Introduction à R

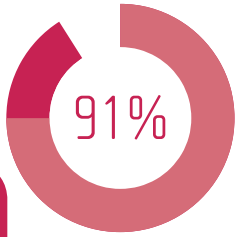
Mastère Spécialisé® Big Data

ENQUÊTE INSERTION

Accélérateur de carrière numérique

96,8%

des étudiants et des diplômés sont satisfaits de leur formation

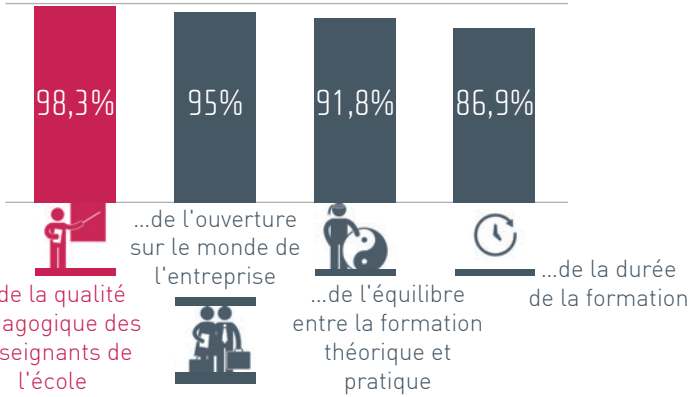


ont trouvé un emploi facilement ou très facilement

et 75% avant même la fin de leur stage de thèse professionnelle

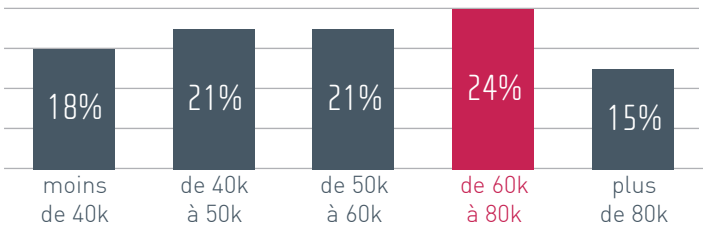
Ce qu'ils apprécient dans la formation

Ils sont plutôt satisfaits ou très satisfaits...



Les entreprises qui recrutent nos étudiants

des entreprises qui les emploient considèrent que le big data est une priorité



Salaire : + de 50k€ pour 60% des diplômés

Salaire brut primes comprises

Résultats issus d'une enquête conduite par Télécom ParisTech auprès des quatre premières pr

* Recueil réalisé en ligne du 21 septembre au 1er novembre 2016. 111 répondants sur un total d

Un Mastère Spécialisé® est une formation professionnalisante qui s'adresse à un public de diplômés Bac +5 (ingénieurs et masters) en poursuite d'études ou en reconversion.

Impact positif de la formation

92% intérêt du poste

78% niveau de responsabilité

78% rémunération

Bénéfices de la formation

93,3% accroît les chances de trouver un emploi dans le Big Data

90% formation valorisée dans le monde du travail

91,5% permet de créer des contacts professionnels

93% permet aux salariés en reconversion de se réorienter professionnellement

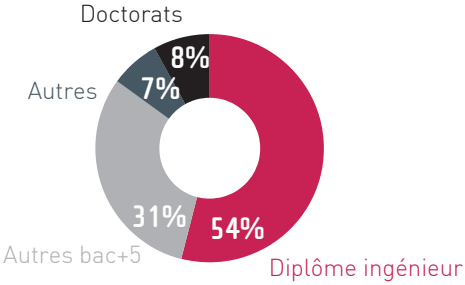
Les deux grands profils types d'étudiants

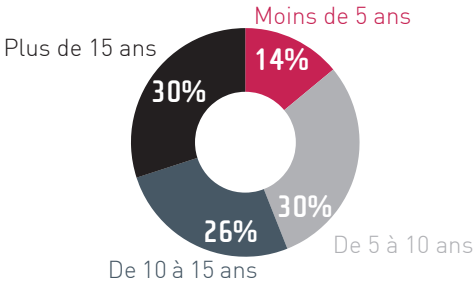
40% en poursuite d'études

Diplôme à l'entrée dans la formation

60% en reprise d'études

Expérience professionnelle à l'entrée dans la formation





1. Acquérir une double compétence
2. Trouver un emploi ou en changer
3. Formation professionnalisante
4. Améliorer sa rémunération

1. Trouver un emploi ou en changer
2. Acquérir une double compétence
3. Formation professionnalisante
4. Améliorer sa rémunération

omotions du Mastère Spécialisé ® Big Data
e 152 étudiants et diplômés

www.telecom-paristech.fr/ms-bigdata

Certificat d'Études Spécialisées

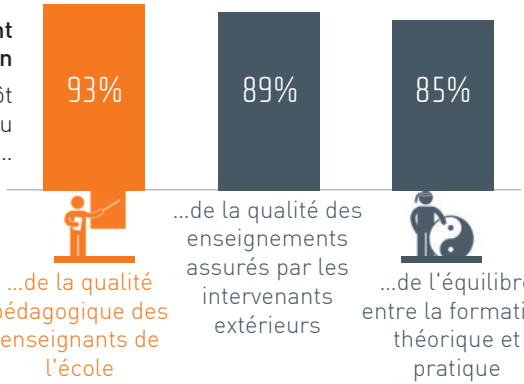
Un cursus certifiant de haut niveau

89%

des répondants sont satisfaits de leur formation

Ce qu'ils apprécient dans la formation

Ils sont plutôt satisfaits ou très satisfaits...

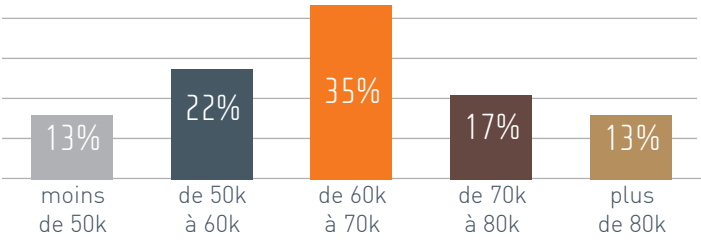


Impact positif de la formation

- 85% intérêt du poste
- 82% niveau de responsabilité
- 52% rémunération

Principales motivations

1. 54% Suivre une formation professionnalisante
2. 44% Changer d'emploi
3. 39% Acquérir une double compétence



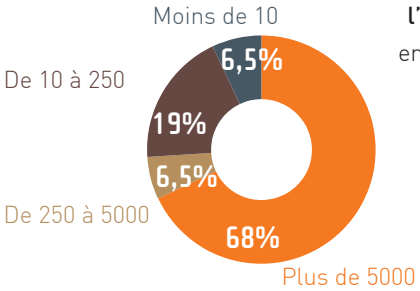
Salaire : + de 60k€ pour 65% des diplômés

Salaire brut primes comprises

Enquête conduite en novembre 2016 auprès des 3 premières promotions du CES « Data Scientist » 33 répondants sur 45 interrogés soit 73% de répondants

Data Scientist

Parcours avant le CES



Taille de l'entreprise
en nombre de salariés

Formation initiale

Diplôme d'ingénieur **40%**
Master **28%**
Doctorat **19%**
Autre **13%**

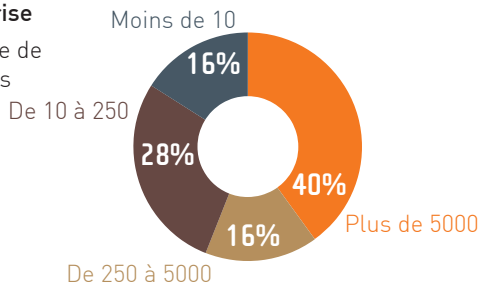
Années d'expérience

Plus de 16 ans **28%**
10 à 15 ans **41%**
5 à 9 ans **25%**
Moins de 5 ans **6%**

Fonctions occupées

Manager
Ingénieur d'étude
Chef de projet
Ingénieur R&D

Évolution professionnelle



Secteurs d'activité

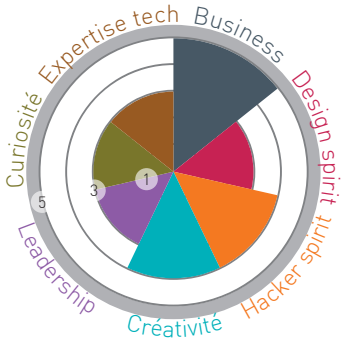
28% Activités informatiques et services d'information
28% Activités financières et d'assurance
8% Industrie
8% Énergie
28% Autres secteurs : édition, audiovisuel, recherche, santé, administration...

Data Scientist
Data Analyst
Architecte SI big data
Ingénieur big data
Chef d'entreprise
Manager data science
Consultant
Chef de projet

La formation se répartit en 12 sessions de deux jours sur 10 mois, ce qui permet le maintien d'une activité professionnelle. Chaque session présente cours, travaux dirigés et travaux pratiques ponctués par le témoignage d'un professionnel. Trois grands domaines sont abordés : les données, l'apprentissage statistique et l'informatique distribuée. Le CES se conclut par un projet personnel sur 3 mois.

Data analyst

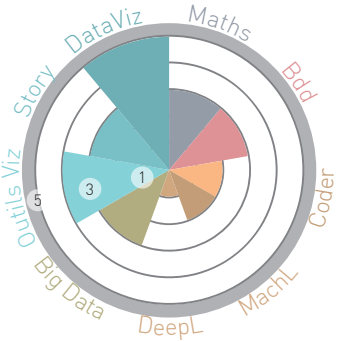
Les data analysts examinent les données d'une unique source et travaillent sur des données déjà connues. Leur boîte à outils statistiques et informatiques leur permet d'organiser, synthétiser et traduire les informations utiles aux organisations pour orienter les prises de position des acteurs décisionnels. Ils agissent en aval de la chaîne de traitement de la donnée tout en collaborant avec les data scientists sur les dimensions technico-scientifiques. Ils explorent et exploitent, extraient et analysent les données en définissant des indicateurs clefs de performance (KPI) pertinents. Ils sont amenés à vulgariser et à restituer de manière pertinente et sous un format exploitable les résultats aux décideurs, notamment au travers de data visualisations. Avec les profils d'expert en data visualisation et ingénieurs big data, ils sont une des composantes du métier de data scientist et peuvent évoluer vers celui-ci.



Les spécialistes d'une source de données

Profil Formation type école de commerce ou école d'ingénieur

Bonne connaissance des outils analytiques. Goût pour les chiffres, sensibilité aux enjeux business. Ces profils peuvent également être des consultants freelance.



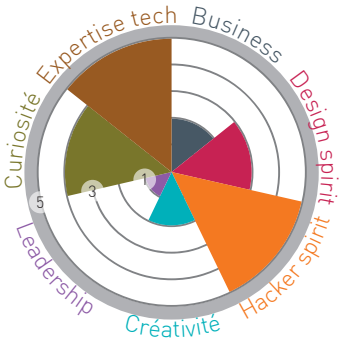
Compétences **Outils :** Excel VBA, SQL, R, Python, outils de visualisation de données (type Tableau)...

Qualités : Capacité d'analyse, aptitude pour le travail en équipe, communication, curiosité intellectuelle...

Ingénieur big data

Les ingénieurs big data s'occupent de la maintenance au quotidien des bases de données et des frameworks big data. Ils développent, entretiennent, testent et évaluent des solutions big data, pour assurer que les infrastructures techniques tiennent la charge sous la masse de données exploitée par les data scientists. Experts en data warehousing, ce sont également eux qui font migrer les bases de données et les frameworks des entreprises vers les évolutions les plus récentes. Ce métier est un bon point d'entrée dans le monde de la donnée pour des profils issus d'écoles d'ingénieur très poussés en informatique, avec des capacités d'évolution vers le métier plus large de data scientist, dont il est une des composantes.

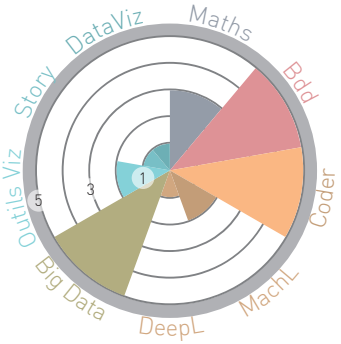
Tenir les promesses du big data grâce à la technique



Profil

Formation IT

Ces profils peuvent également être ou devenir des consultants freelance.



Compétences

Outils : frameworks Big Data (Hadoop, spark…) et bases de données NoSQL (MongoDb, ElasticSearch, Cassandra…), Python, Java, Scala…

Qualités : Curiosité intellectuelle, rigueur, méthode, adaptabilité, anglais technique, communication orale et écrite, travail en équipe…

La data peut-elle tout résoudre ?

« *La data est le nouveau pétrole !* » Si l'expression est souvent prononcée, elle est assez trompeuse, relève Henri Verdier, ancien directeur d'Etalab, le service du premier ministre chargé de l'ouverture des données publiques, et aujourd'hui administrateur général des données, dans un billet de 2013. Contrairement au pétrole, les données ne sont pas des ressources rares, les transformer ne les détruit pas, les utiliser peut même leur faire prendre de la valeur. Elles n'ont pas intérêt à être stockées en attente d'un acquéreur, elles n'ont pas de valeur tant qu'elles restent brutes, isolées, inanimées, non agissantes. « *Elles sont le substrat dans lequel il faut apprendre à se mouvoir et, plus qu'une matière première ou une énergie, le code au cœur du réseau et le flux sur lequel se greffent les autres applications.* »

Ces gisements soudains de valeur suscitant un grand engouement, le recours excessif à la donnée pour résoudre tous les problèmes a parfois atteint des limites et pu décevoir. Certains se sont retrouvés noyés sous un déluge de données qu'ils ne savaient plus contenir, d'autres ont oublié la nécessité d'arrêter l'exploration à temps, et celle de sortir des produits industrialisables sur la base d'un volume et d'une variété maîtrisée des données. Le recours systématique aux données et aux chiffres risque de pousser à une confiance aveugle, or il faut aller voir les métiers pour comprendre leur contexte, il faut écouter les clients qui remontent des ressentis sortant de l'ordinaire. Tout ne peut pas non plus être fait avec les données collectées des clients, comme par exemple les utiliser pour leur vendre des services en plus qu'ils n'avaient pas sollicités. Collecter et conserver les traces numériques – déplacements, conversations, accès internet – des personnes sans leur consentement, ou de manière disproportionnée, n'est pas la bonne voie pour concevoir des services rencontrant l'approbation et la confiance des utilisateurs.

Le rythme effréné des évolutions technologiques, et l'ensemble des combinaisons possibles de ces outils ouvertes aux data scientists, peuvent également constituer un frein pour les utiliser de manière pertinente et adaptée. Ces technologies et les usages qu'elles offrent, enthousiasmants, restent en constante évolution, et il faut savoir garder raison et modestie dans leur application.

Il est cependant indéniable que la data science a tenu de belles promesses. Comme le rappelle Stephan Clémençon, professeur à Télécom ParisTech, « *personne ne se lancerait aujourd'hui dans le ciblage commercial sans un outil de scoring ou de recommandation. C'est une activité reconnue qui pèse plus de 20% de chiffre d'affaires en plus. Dans le milieu industriel, tout le monde utilise le terme de maintenance prédictive et l'identifie comme un catalogue d'outils qui permettraient de mieux gérer des infrastructures très complexes. C'est le cas dans le domaine du transport, aérien en particulier, et dans les grands réseaux d'énergie. Comme on vend un service plutôt qu'un produit, il s'agit de pouvoir le maintenir à un certain niveau et de faire des économies en le maintenant mieux.* »

S'éduquer à la donnée

Pour le data scientist Yoann Janvier, le principal frein à l'innovation en science des données n'est pas vraiment la technologie : « *c'est l'accès à la data, qui, pour des raisons réglementaires et organisationnelles, pose parfois problème. Ce nouveau métier est également très mystérieux dans les entreprises dites classiques : il faut passer du temps à expliquer, éduquer, développer une culture de la data. Sans relais dans les organisations, le data scientist ne peut pas grand chose.* » En effet, si se former à la donnée fait partie du quotidien des data scientists, s'éduquer à la donnée est notre affaire à tous. D'autant que les réglementations en lien avec la donnée sont nombreuses. Citons par exemple le décret de juillet 2016 sur l'obligation d'ouverture et de mise à disposition des données de production et de consommation d'énergie des opérateurs, l'obligation règlementaire européenne faite aux assureurs et aux banquiers de rendre compte des risques de solvabilité, ou encore le règlement général sur la protection des données personnelles (RGPD, voir page 35).

Gilles Babinet, représentant la France auprès de la Commission européenne pour les enjeux du numérique, et auteur d'ouvrages sur le big data et sur la transformation digitale, y voit là « *un vrai sujet d'éducation des citoyens. Je ne cesse d'être confronté à des fantasmes et à de la méconnaissance à ce sujet. Comprendre par exemple les bases du RGPD est loin d'être superflu. C'est un texte fondateur et il ouvre des enjeux de choix aussi à l'échelon du citoyen. Cette éducation vise également le personnel politique.* »

La France joue un rôle moteur pour améliorer la culture de la donnée sur tout le territoire et en Europe, par son engagement ancien et soutenu dans le développement des *open data* – nouvelle licence parue fin avril 2017 –, à travers certains articles de la loi pour une République numérique d'octobre 2016, comme l'ouverture en avril 2017 du service public de la donnée de référence (base adresse nationale, plan cadastral informatisé, répertoire des entreprises et des établissements…), ou la tenue régulière de hackathons à portée nationale. « *Sur les enjeux de régulation* », poursuit Gilles Babinet, « *la France a largement œuvré pour participer à l'élaboration du RGPD. Même si le texte final a fait l'objet de nombreux compromis, visant notamment à le ramener à sa dimension principielle, il ne fait que peu de doute que certains éléments, comme par exemple les pénalités imposables aux contrevenants, ont été influencés par la position française.* »

Les obligations règlementaires et les possibilités offertes par des technologies à l'évolution rapide sont à l'origine de nouvelles fonctions et de nouveaux métiers. Garant d'équilibres précieux, un déontologue de la donnée s'assurera que l'utilisation de données par une entreprise ne nuit pas à sa réputation et que sa « *stratégie donnée* » reste conforme à ses objectifs généraux. L'éducation à la donnée passe aussi par l'étude et l'impact de nouvelles combinaisons technologiques. L'immutabilité et la transparence inhérentes à certaines blockchains, par exemple, pose des questions en termes de sécurité et de protection des données personnelles.

Maîtriser les algorithmes

Les données ne sont rien sans les algorithmes qui permettent de les manipuler, les transformer, les classer. Tri à bulles, tri cocktail, tri par tas, tri *quicksort*… l'élève ingénieur perçoit rapidement que de nombreux algorithmes existent pour trier des données, que leur efficacité dépend du type de ces données, et que leur choix a des implications sur la mémoire utilisée et l'énergie consommée. Cette famille d'algorithmes de tri est une bonne introduction à la complexité des algorithmes en général et aux questions techniques à se poser en les créant. Et même pour un objectif simple comme le tri, visualiser leur fonctionnement est souvent nécessaire pour en saisir toutes les subtilités et y trouver des pistes d'optimisation.

Le grand public s'est quant à lui peu à peu familiarisé avec l'algorithme du moteur de recherche de Google – sans en connaître nécessairement les détails – ou avec celui de la *timeline* de Facebook. Il a appris que « *tout le monde n'est pas égal devant l'algo* », et que selon son historique, sa provenance géographique, et d'autres types de données souvent personnelles, les algorithmes pouvaient présenter une vision du monde adaptée à celui qui l'utilisait. Avec les algorithmes de recommandation qui influent nos décisions et nos comportements, ils peuvent même finir par façonner le monde à notre insu.

Présenté comme une recette de cuisine, avec ses données-ingrédients, illustré par du code informatique abscons ou s'écoulant comme dans Matrix, l'algorithme est tour à tour inoffensif ou inquiétant.

Un *algorithme* est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre un problème ou d'obtenir un résultat.

Si cette définition est valable pour les algorithmes simples comme ceux du tri, elle laisse de côté les conséquences des algorithmes plus généraux, larges et complexes, en particulier ceux qui aident à prendre des décisions. L'évaluation des impacts d'un algorithme est donc tout aussi essentiel que leur vérification – ils font effectivement ce pour quoi ils ont été créés – ou leur sûreté – ils ne sont pas détournables.

L'État a montré l'exemple en publiant en mars 2017 un décret d'application de la loi pour une République numérique, relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique. Dans un monde où nous serions « *tous demain algorithmés* », selon la formule d'une conférence proposée par le Secrétariat Général pour la Modernisation de l'Action Publique, l'État redonne à chacun la possibilité de connaître et comprendre les décisions administratives prises à son encontre. Que ce soit en matière de fiscalité, ou d'admission post-bac, les applications sont déjà en cours. C'est un modèle à suivre pour les acteurs privés également, et dans tous les cas un point de vigilance pour les utilisateurs, qui doivent améliorer leur culture algorithmique, tout autant que leur culture des données.

Design, Données & Algorithmes

Qu'est-ce que concevoir de bons algorithmes ? Qu'est-ce que faire un traitement respectueux des données ? Actuellement préoccupation exclusive de l'ingénieur, la modélisation d'algorithmes pourrait bien bénéficier de l'approche et de la sensibilité des designers.

Maguelonne Chandesris, data scientist, responsable de la thématique « *Data, Mobilités et Territoires* » chez SNCF Innovation & Recherche, est docteur en mathématiques et diplômée du Mastère Spécialisé ENSCI « Innovation By Design ». Elle y a mené une recherche sur les enjeux du design dans la forme des décisions algorithmiques, que l'on retrouve sous forme d'un essai dans le n°4 de la revue Sciences du Design. Illustrant son propos d'exemples puisés dans le domaine des transports – l'usage de la boîte de vitesse automatique, les évolutions encore en cours de l'automatisation de la circulation aux carrefours et les algorithmes de calcul d'itinéraire –, elle rappelle que l'automatisation des décisions n'implique pas toujours son adoption, qu'elle modifie l'environnement et les comportements, et qu'elle peut brider les désirs des utilisateurs au profit de tiers non nécessairement connus.

Pour prendre des décisions sur des cas précis, deux manières de raisonner sont possibles : soit l'établissement d'un modèle permettant de décider par approche inductive et déductive, soit le raisonnement par analogie ou différence, selon une approche transductive. Le choix du modèle pour la première porte en soi une vision du monde, et n'est donc pas anodin. La seconde méthode, qu'on retrouve dans les algorithmes de recommandation, est moins rigoureuse mais plus simple à mettre en œuvre. Dans les deux cas le code informatique façonne le monde et établit une gouvernance algorithmée.

L'automatisation des décisions pose des questions à enjeux forts. La sécurité et la fiabilité des systèmes automatiques doivent être garanties. Du concepteur à l'utilisateur, en passant par le constructeur et le propriétaire, les responsabilités doivent pouvoir être établies. La délégation de décisions aux machines et la capacité de ces dernières à prendre des décisions éthiques doivent être étudiées. Les désirs et l'imagination ne doivent pas être bridés par des algorithmes véloces et trop friands de notre attention.

L'acceptabilité des algorithmes est en jeu. Pour cela l'humain doit trouver une vraie place dans le couplage qui le lie de plus en plus aux systèmes numériques. L'introduction d'empathie dans ces systèmes, d'une dose d'imprévu, et la possibilité de ressentir et jouer avec l'algorithme, sont des pistes que le design peut ouvrir.

الخوارزمي, *du nom du mathématicien du IX^e siècle Al-Khwârizmî, également géographe, astrologue et astronome, considéré comme le père de l'algèbre, et qui a proposé une première classification des algorithmes connus à l'époque, dont celui d'Euclide.*

Boîte à outils d'algos

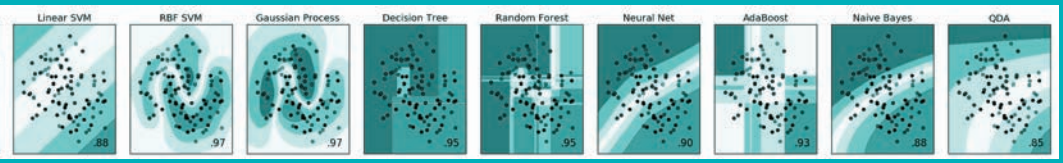
Choisir un algorithme ou une combinaison d'algorithmes qui permettent de résoudre un problème de classification est un art et un savoir-faire qui peuvent s'acquérir en partant d'exemples très visuels.
Télécom ParisTech participe au développement de Scikit-learn, un ensemble d'outils en Python très prisé pour l'analyse et la fouille de données.

Un outil de la communauté

Le développement de Scikit-learn a commencé en 2006 lors d'un « *Google Summer of Code* », grâce au travail de David Cournapeau. La communauté scientifique Python avait alors besoin de briques logicielles sur certains thèmes, comme le *machine learning* ou le traitement d'image, et c'est ainsi qu'est né Scikit-learn, le *kit scientifique* pour faire du *machine learning*. Entre 2006 et 2010 un autre français, Mathieu Brucher, a repris le projet et l'a fait vivre pendant sa thèse. Alexandre Gramfort, maître de conférences à Télécom ParisTech pendant 5 ans, se souvient. Il était en post-doc à l'INRIA Saclay en janvier 2010, et « *avec Gaël Varoquaux, Bertrand Thirion, Vincent Michel, de l'INRIA, ainsi qu'Olivier Grisel et Fabian Pedregosa, des personnes qui avaient pas mal d'expérience logicielle dans le monde de Python, nous nous sommes enfermés dans une salle à Saclay et nous avons commencé à collecter tous les bouts de code qu'on possédait pour faire du machine learning, à les assembler dans une bibliothèque, puis à commencer une documentation, écrire des tests, faire de l'intégration continue. La plate-forme GitHub n'existait pas encore à l'époque et tout était mis sur Sourceforge. On a communiqué dessus assez vite.* »

Scikit-learn a été créé par les personnes faisant du machine learning pour des problèmes très appliqués. « *Nous avions besoin de briques logicielles ultra solides car l'enjeu n'était pas de faire le logiciel, mais de répondre à un travail applicatif. Nous avions la vision d'un logiciel qui soit simple à utiliser, très pragmatique, qui ne réponde pas forcément à tous les cas d'usage possibles, mais qui permette de faire bien et simplement, de façon efficace, 80 ou 90% d'entre eux.* » Et cet état d'esprit a perduré.

« *Depuis janvier 2010, plusieurs autres personnes ont été financées pour continuer le développement de Scikit-learn.* » Arrivé à Télécom ParisTech en septembre 2012, Alexandre Gramfort lui consacre une fraction de son temps. Depuis 2016 et dans le cadre d'un contrat industriel dont il est responsable, deux ingénieurs y travaillent avec lui à plein temps. Des missions doctorales ont été financées par le *Center for Data Science* de l'Université Paris-Saclay, et des doctorants de l'équipe de statistique du département Image Données Signal à Télécom ParisTech s'impliquent également. On compte aujourd'hui plus de 600 contributeurs dans le monde entier, et la première publication scientifique rédigée en 2011 a été citée plus de 5000 fois.



Scikit-learn dans l'enseignement, et à Télécom ParisTech

À Télécom ParisTech, tous les enseignements en *machine learning* aujourd'hui sont faits en Python. « *Je suis un peu responsable de ce mouvement vers Python* », explique Alexandre Gramfort, qui poursuit : « *la plupart des enseignements utilisent Scikit-learn, et quand ce n'est pas le cas, c'est soit parce que les outils ne sont pas encore disponibles dans Scikit-learn, soit parce que l'on souhaite que les élèves manipulent la donnée par eux-mêmes, plutôt que d'utiliser des briques toutes faites. Les enseignants-chercheurs de Télécom ParisTech maîtrisent Scikit-learn, l'utilisent en TP, et s'en servent en cours pour faire des démonstrations. C'est également utilisé massivement dans toutes les grandes écoles de Paris, et dans plusieurs universités, y compris à l'international. C'est vraiment un ensemble d'outils très répandu qu'il est bon de connaître.* »

L'enseignement par des spécialistes n'est pas la seule raison qui fait de Télécom ParisTech un centre de gravité pour Scikit-learn. On y croise des *core developpers*, qui ont les droits de validation et qui garantissent la bonne évolution de l'outil. « *Le Center for Data Science dont je m'occupe avec d'autres personnes à l'échelle de l'Université Paris-Saclay pousse à l'utilisation de la data science à travers les disciplines scientifiques expérimentales. C'est un lieu où les gens qui savent traiter et modéliser des données vont pouvoir rencontrer ceux qui ont des données et qui ont des problèmes de la vie réelle.* » Scikit-learn fait partie de la boîte à outils de tout data scientist désireux d'être correctement équipé pour explorer les nombreux pans du monde des données.

scikit-learn.org

Qui utilise Scikit-learn aujourd'hui ?

200 000 utilisateurs viennent chaque mois du monde entier télécharger les outils et les mises à jour, inspecter la documentation. « *Sans nécessairement le dire officiellement, de nombreuses entreprises nous utilisent, des grandes comme de plus petites.* » Les logiciels et bibliothèques *open source* sont souvent cités par les entreprises, et les data scientists sont invités à contribuer à leur développement, et faire de la veille sur ces outils.

Rendre les données visibles

Des premières visualisations de données par Florence Nightingale (page 4) et de la représentation de la campagne de Russie 1812-1813 par Charles Joseph Minard, aux travaux du statisticien Jacques Bertin et sa sémiologie graphique, puis aux documents auto-expliquants du designer Bret Victor en 2011, savoir communiquer des données visuellement pour en révéler les ressorts et les expliquer à ses interlocuteurs est une discipline à part entière.

La *visualisation de données (data visualisation, dataviz)* désigne un ensemble de techniques et d'outils permettant de mieux comprendre et analyser les données, en les traitant sous une forme visuelle, interactive et graphique.

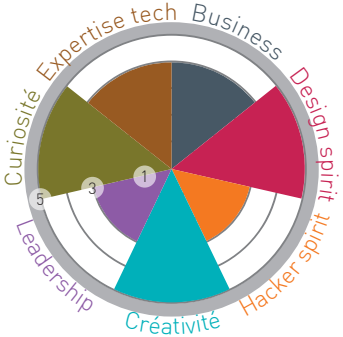
Un premier objectif est de rendre les données plus intelligibles, plus immédiatement accessibles. Il s'agit d'attirer l'attention du lecteur sur les faits saillants, par un choix judicieux d'organisation spatiale, de liaison des éléments, de couleurs, de formes, de typographies. Une dataviz réussie fera rentrer progressivement l'utilisateur dans l'histoire des données, chaque choix visuel lui en offrant une nouvelle clé de compréhension. Deuxième objectif, la visualisation exploratoire permet à l'analyste de naviguer à travers des ensembles de données multidimensionnels et complexes, les diverses représentations graphiques à sa disposition lui permettant de saisir les grands phénomènes et les motifs cachés. « *L'un des grands enjeux est de construire des systèmes qui permettent à l'utilisateur et au système de travailler ensemble, en tirant profit de ce que chacun fait bien* », précise James Eagan, maître de conférences à Télécom ParisTech en interactions homme-machine. « *Il y a des éléments sur la représentation qui peuvent faire changer l'interprétation. La visualisation peut être aussi un outil pour aider l'utilisateur à explorer ses données. Ainsi, si on ne sait pas quel est le modèle, si on ne sait pas ce qu'on cherche, la visualisation est un outil qui aide à explorer et comprendre quand on ne sait pas encore ce qu'on veut chercher !* »

La visualisation de données ne s'exprime pas que sur des supports fixes et figés. Manipuler la dynamique d'un ensemble de données dans le temps, régler des paramètres pour focaliser sur tel ou tel aspect, travailler la donnée de manière interactive, poussent l'imagination un cran plus loin. Hans Rosling, statisticien décédé début 2017, a donné ses lettres de noblesses à cet art. Enfin, aidés par les techniques de réalité augmentée, et pilotés par des intelligences artificielles, il devient possible de plonger au cœur des données. « *La visualisation de données est une discipline complète* », explique James Eagan, « *l'union des statistiques, du design, de l'informatique, complétées par les systèmes d'intelligence artificielle, la fouille de données et le machine learning.* » Pour Charles Miglietti, cofondateur de la start-up dataviz Toucan Toco (voir page 53), la data visualisation est tout simplement « *le dernier kilomètre de la donnée.* »

Expert Data visualisation

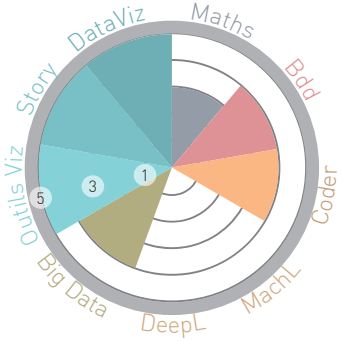
Le profil d'expert en data visualisation présente deux facettes : il peut être utilisateur expert d'outils de data visualisation, et effectuer du reporting et du story telling sur les données, ou bien développeur front-end et back-end créant des applications de data visualisation – en intranet, sur le web, sur des applications mobiles et sur papier. Grâce à leur travail sur les interfaces, les experts en data visualisation permettent également aux équipes opérationnelles d'y voir plus clair dans leurs données et d'identifier des pistes d'analyse qui n'apparaîtraient pas de manière aussi évidente à leur simple lecture. Ce métier demande beaucoup de travail et une grande culture pour choisir les visualisations les plus pertinentes et les moins susceptibles d'apporter des biais dans l'interprétation des données. Ces profils peuvent évoluer à terme vers le métier de data scientist au sens large, dont il est un des rôles.

Le data storytelling est le futur du reporting



Profils & outils

Les *profils data* maîtrisent les outils de visualisation comme Tableau, Qlik, Microsoft Power BI, Excel...



Pour les *développeurs web front end* : Javascript, D3.js, Angular.js, Django, html, CSS... les *développeurs front end & back end* qui développent les applications de data viz de A à Z, plus les technologies *front end* de type Node.js, Java...

Compétences

Qualités : curiosité intellectuelle, créativité, sensibilité au design, autonomie, aptitude pour le travail en équipe, rigueur...

De man of science à scientist

L'origine du mot « *scientist* », que l'on traduit le plus souvent en français par « scientifique », remonte au début du XIX[e] siècle, au Royaume-Uni. Mary Somerville, une chercheuse écossaise, y avait rédigé des livres savants réunissant des domaines jusqu'alors disparates des mathématiques, de l'astronomie, de la géologie, de la chimie et de la physique. Son travail était si clair et si fluide que ses textes formaient le cœur du premier programme scientifique de l'Université de Cambridge.

En 1834, son traité *On the Connexion of the Physical Sciences* impressionna fortement ses pairs. William Whewell, professeur à Cambridge, écrivit un article élogieux sur Mary Somerville, utilisant pour elle le terme « *scientist* ». Il avait proposé ce mot l'année précédente à la *British Association for the Advancement of Science*, à la fois parce que la désignation valant jusqu'alors, « *man of science* », était de plus en plus controversée, et parce que les travaux des scientifiques devenaient de plus en plus interdisciplinaires. Mary Somerville n'était en effet pas simplement mathématicienne, astronome ou chimiste… Elle avait identifié les liens entre ces différentes disciplines et leurs méthodes d'exploration scientifique, avait su les articuler et communiquer cette vision à ses contemporains de manière élégante. Pour Whewell, il n'était pas suffisant de créer un mot qui ne fasse plus allusion au genre. Il fallait également traduire cette capacité à synthétiser des champs scientifiques alors distincts, une idée rendue possible par ce qu'il décrit comme l'« *illumination particulière de l'esprit féminin* ». « *Scientist* » est créé à partir de *Science* et *Artist*. Il reconnaît en Mary Somerville et en d'autres scientifiques de l'époque le génie créateur de l'artiste capable d'établir et de faire apparaître des liens invisibles pour le commun des mortels, et les traduire de belle manière, poétique parfois, pour toucher durablement le lecteur.

Raconter les découvertes de la science et les rendre moins absconses pour les non initiés, en les communiquant via des images et des émotions ressenties, est un art qui mérite d'être reconnu et exercé.

Mary Somerville fréquentait le cercle intellectuel *The Analyticals*, un groupe de scientifiques qui souhaitait réformer et professionaliser la science britannique, auquel participait Charles Babbage, l'inventeur de la machine analytique. Elle lui présenta Ada Lovelace dont elle était la mentor. Cette dernière créa le premier algorithme destiné à être exécuté par une machine, et contribua à jeter les bases de notre informatique moderne.

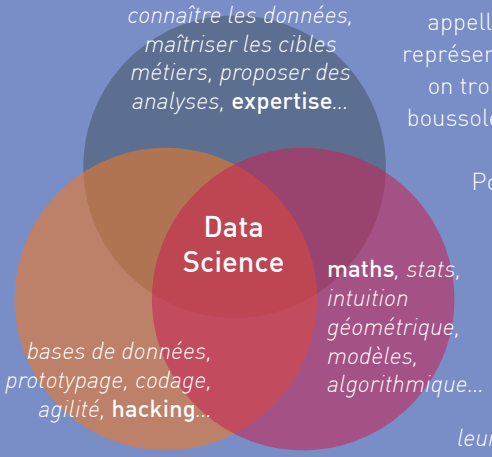
Combinant maîtrise des techniques et des technologies, vision transdisciplinaire et art de raconter et d'enseigner, tout en étant impliquée dans les grands mouvements de son temps, Mary Somerville, la première *scientist*, devrait être une source d'inspiration pour chaque data scientist.

L'hybridation des compétences, une évolution de la data science

Si à l'époque de Mary Somerville il était naturel d'évoluer dans plusieurs disciplines à la fois, et de se nourrir des unes et des autres, la cybernétique naissante au sortir de la deuxième guerre mondiale contribue à privilégier des approches mono-disciplinaires, qui ne seront pas les plus fécondes. Dans le même temps, les hommes prennent de plus en plus de place dans les métiers de l'informatique, au point qu'une majorité de femmes aujourd'hui pensent que coder n'est pas de leur ressort. Cette perte de diversité de points de vue et des expériences s'est ressentie fortement dans toute l'aventure de l'intelligence artificielle, et commence à être prise au sérieux, notamment suite à l'observation que les robots algorithmiques et mécaniques reproduisent les biais cognitifs de ceux qui les concoivent.

Transdisciplinaires par nécessité

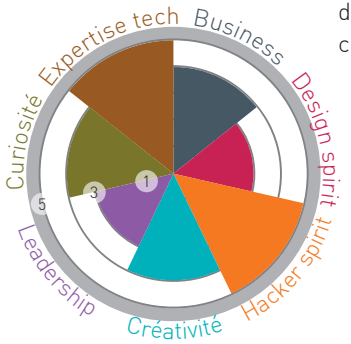
Les métiers de la data science puisent dans de nombreuses disciplines. Celle ou celui qui combine des compétences de data analyst et d'ingénieur big data, avec une conscience des besoins métiers et des fondements marketing, ainsi qu'une appétence à rendre les données visibles et compréhensibles, cette personne aux talents hybrides est celle qu'on appelle data scientist. Cette polyvalence est souvent représentée par le diagramme de Venn ci-contre, dont on trouve plusieurs variantes et qui reste une bonne boussole pour se situer dans cette galaxie de métiers.



Pour Yoann Janvier, ajouter ces diverses cordes à son arc est une chance pour les futurs data scientists : « *D'anciens dataminers par exemple vont devoir acquérir l'agilité pour aller puiser dans l'open source et pour faire de la data visualisation. Ailleurs dans les équipes, des juniors un peu trop geeks doivent travailler leur communication et leur organisation.* » Les métiers de la data science contiennent en eux leurs principes d'évolution.

Data Scientist

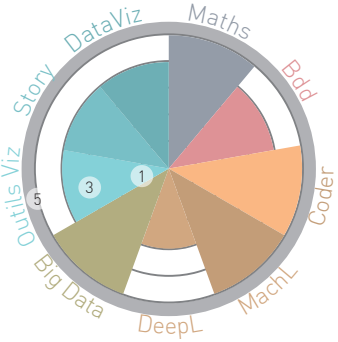
Être data scientist, c'est être au cœur de la valorisation des données, c'est comprendre les enjeux et les problématiques stratégiques de l'entreprise et mettre en place des algorithmes qui y répondent. Les data scientists vont à la rencontre des métiers pour en définir les besoins, identifient les indicateurs et données pertinentes, et les analysent à l'aide d'algorithmes qu'ils ont conçus. Ils interviennent à toutes les étapes de la chaîne de données : définition du problème, collecte des données, nettoyage, mise en place des modèles et création des algorithmes. Ils doivent ensuite savoir présenter et prioriser les résultats pour les rendre exploitables par les décideurs. Il leur faut donc d'excellentes capacités de communication car ils sont au contact des opérationnels métiers, des profils plus techniques et des décideurs, et doivent adapter leur discours à chacun. C'est particulièrement vrai dans les grandes structures en cours de transition, où ils vont être sollicités pour faire de la vulgarisation et expliquer au reste de l'entreprise la démarche qu'ils ont adoptée et les conclusions qu'ils en ont tirées.



À la croisée de trois domaines : mathématiques, IT, business

Profil Formation mathématiques-statistiques ou informatique

L'orientation à dominante IT ou à dominante maths dépend des problématiques de l'entreprise. Sensibilité aux enjeux business, notamment dans les secteurs comme le marketing, le web, la publicité...



Compétences **Méthodes :** Analyse de données, modélisation, machine learning... **Outils :** Python, R , Java, C, C++, Matlab, écosystème Hadoop (Hadoop, Hbase, Hive, Pig, Mapreduce...), Spark... **Qualités :** curiosité intellectuelle, capacité d'apprentissage, rigueur, aptitude pour le travail en équipe, communication...

Des data scientists au quotidien

Après avoir obtenu le Mastère Spécialisé® Big Data de Télécom ParisTech, **Kim Pellegrin** a intégré une direction des systèmes d'information chez Dassault Systèmes. « *Ma mission consiste à améliorer les outils de traitement de l'information en apportant des solutions innovantes de machine learning. Cela concerne les moteurs de recherche, les moteurs de recommandation et les outils de support métiers. J'interviens en apportant une expertise et en réalisant des prototypes, l'industrialisation des solutions retenues se faisant avec les équipes de développement. À la différence de l'approche déterministe du développement classique, la data science s'inscrit dans une méthode expérimentale et itérative. Il faut réaliser des prototypes pour pouvoir évaluer la faisabilité et la valeur du service final. Il est donc important de savoir expliquer aux métiers ces nouveaux modes opératoires. Être data scientist nécessite une expérience technique et des connaissances théoriques solides, de prendre du temps pour faire de la veille, d'identifier des points sur lesquels on est moins à l'aise. C'est un cheminement de plusieurs années, pour un métier passionnant dont l'apprentissage nécessite un investissement personnel important, mais c'est un parcours qui est tout à fait réalisable pour un esprit curieux et toujours avide d'apprendre. C'est un métier d'explorateur qui amène à découvrir et apprendre en permanence.* »

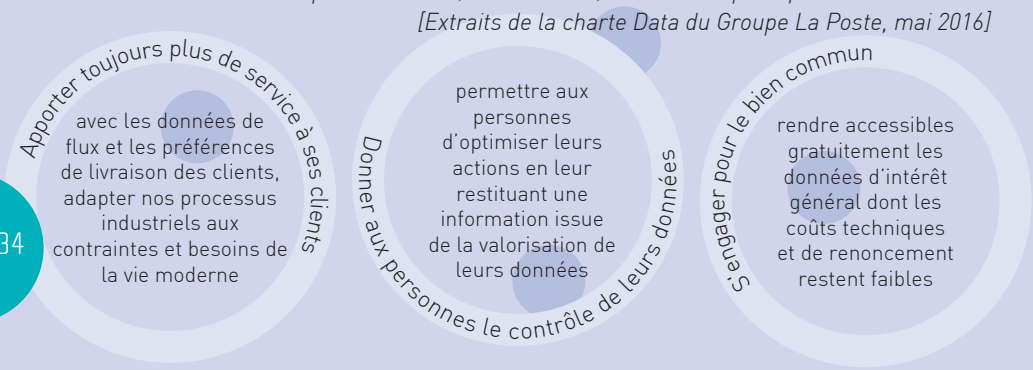
Pierre Achaichia, responsable de dix data scientists chez Enedis (lire page 51) estime qu'« *au-delà des connaissances purement techniques en modélisation, la méthodologie et les capacités de communication/vulgarisation sont les compétences clés. Un profil idéal de data scientist répond au moins aux critères suivants : il a été confronté à des problèmes complexes ayant challengé sa créativité ; il a une expérience dans le développement d'un projet informatique et a réalisé des travaux faisant appel à de la modélisation statistique ; il peut expliquer clairement son métier et vulgariser les résultats d'une étude à une personne non initiée.* »

Pour **Alain Abramatic** chez PSA Groupe (lire page 6), à la fois data analyst et data scientist, « *le data scientist vient en complément et en support des business developpers pour la construction de services à forte valeur ajoutée. Notre rôle est d'échanger avec le métier sur ses besoins d'information ou proposer des opportunités, imaginer comment les données disponibles peuvent permettre d'y répondre, proposer des axes d'analyse, construire un prototype pour valider le concept et sa valeur, engager un projet de réalisation.* »

Yoann Janvier dirige une équipe de data scientists chez Ipsen (lire page 5). Il livre quelques règles utiles pour être à l'aise dans ce poste multi-forme. « *Comprendre comment fonctionne l'environnement autour de soi. Prendre de la hauteur vis-à-vis des solutions suggérées et garder un focus sur les attentes business. Mettre en place de la délégation et structurer l'équipe ou le data lab pour éliminer les tâches répétitives, routinières et sans valeur ajoutée. Se concentrer, justement, sur les activités à valeur ajoutée. Obtenir rapidement des premiers succès pour convaincre de la valeur de la piste explorée, et pouvoir communiquer. Travailler pour cela en mode agile et éviter absolument le mode tunnel. Garder du temps pour l'auto-formation et la veille, ce sont des conditions de survie à long terme du data scientist. Enfin, éduquer l'environnement sur ce que la data science peut ou ne peut pas apporter.* »

Déployer une charte data

Au Groupe La Poste, les activités de services courrier, colis, de la distribution, de la banque, de l'assurance, de gestion des bâtiments et véhicules, des télécommunications et du numérique s'appuient de plus en plus sur des données. Quelles que soient leurs caractéristiques – données de gestion, données industrielles, données confiées par les clients particuliers et entreprises, données de partenaires, données d'identification, données de transaction, données d'interaction aux guichets, avec les postiers, au service client, sur les automates, sur les sites web et applications mobiles et demain avec des objets connectés… –, toutes sont nécessaires à l'efficacité des activités postales. Pour conjuguer confiance et traitement des données, le Groupe La Poste, tiers de confiance, a formalisé ses engagements à toutes ses parties prenantes, citoyens, clients particuliers, professionnels, associations et entreprises clientes, fournisseurs, collectivités publiques et actionnaires.
[Extraits de la charte Data du Groupe La Poste, mai 2016]



Chaire Valeurs et Politiques des Informations Personnelles

Coordonnée par Claire Levallois-Barth, maître de conférences en droit à Télécom ParisTech, la chaire Valeurs et Politiques des Informations Personnelles réunit une équipe pluridisciplinaire de chercheurs de Télécom ParisTech, Télécom SudParis et Télécom École de Management. Elle traite des aspects juridiques, techniques, économiques et philosophiques qui concernent la collecte, l'utilisation et le partage des informations personnelles ainsi que leurs conséquences sociétales.

La Chaire bénéficie du mécénat de l'Imprimerie Nationale, de BNP Paribas, d'Orange, de LVMH, de Dassault Systèmes et d'un partenariat conclu avec la CNIL et la DINSIC. Après les Identités numériques, son deuxième Cahier de recherche (octobre 2017), s'intéresse aux marques et labels de confiance.

www.informations-personnelles.org

Protéger les données personnelles

Données de santé, données de localisation, bulletins de salaire dématérialisés, historique des sites que l'on visite, photographies de famille… toutes ces données personnelles que l'on confie aux serveurs circulent et peuvent être analysées et exploitées à des fins utiles, ou plus préoccupantes. Pour autant, elle ne peuvent pas être réutilisées de n'importe quelle manière. Elles sont en effet soumises à des règles juridiques strictes, qui visent à assurer la confiance des utilisateurs et donc le développement de nouveaux services. L'ancienne directive européenne 95/46/EC, transposée en France via la Loi Informatique et Libertés, sera abrogée et remplacée le 25 mai 2018 par un règlement, le Règlement Général sur la Protection des Données. Celui-ci s'inscrit dans la continuité de la directive de 1995, tout en ajoutant de nouvelles obligations.

« *Il est important pour les data scientists de repérer dans le cadre d'un traitement big data les données juridiquement qualifiées de personnelles, pour ensuite se conformer aux principes qui s'appliquent* », souligne la maître de conférence en droit Claire Levallois-Barth. « *Il y a d'une part les principes clés déjà connus (finalités, qualité et durée de conservation des données, mesures de sécurité et de confidentialité, droits de la personne concernée…) qu'il faut reprendre, et d'autres part les nouveaux principes (protection des mineurs, droit à la portabilité des données, Privacy by design…) qui doivent absolument être respectés. L'objectif est de renforcer l'effectivité du droit fondamental à la protection des données personnelles, et donc la confiance du citoyen européen dans les nouvelles technologies.* »

En particulier, un nouveau principe est introduit, le principe de responsabilité (*accountability*, que l'on peut traduire par obligation de rendre des comptes). « *À tout moment le responsable de traitement des données doit être en mesure de démontrer qu'il remplit ses obligations légales, notamment qu'il gère les risques d'atteinte aux données personnelles, et a mis en place les outils pour en garantir la protection effective.* » Ces éléments de démonstration, constitutifs de la confiance, peuvent prendre la forme d'une politique de protection des données, d'un code de conduite ou d'un mécanisme de certification approuvés. Un des grands chantiers actuels pour les data scientists – et notamment le ou la Délégué.e à la protection des données, voir page 49 - est donc de se préparer à ce règlement. « *Cette nécessité impose notamment de développer une véritable culture Informatique et Libertés et une approche transversale. Il s'agit en effet de prendre en compte la protection des données dès la conception d'un service ou d'un produit, d'identifier les risques associés aux opérations de traitement et de prendre les mesures nécessaires à leur prévention.* »

Les banques & assurances bougent

Banques et assurances sont de grandes maisons, pour la plupart anciennes, ayant une très large clientèle qui effectue de nombreuses transactions journalières. La variété et le volume des données disponibles y sont significatives, même si chaque donnée est encore aujourd'hui relativement petite et structurée. Ces données, auparavant exploitées avec des outils statistiques classiques, le sont aujourd'hui avec des techniques d'analyse qui ont changé la manière de travailler dans plusieurs types d'activité.

En matière de prédiction, on sait mieux aujourd'hui détecter les fraudes, par l'analyse de comportements atypiques, et identifier les profils client susceptibles d'être insatisfaits et de résilier leur compte. Mieux concevoir la tarification des services, notamment dans le domaine des assurances, ce qui est le métier historique des actuaires, accélérer la gestion des sinistres et celle des accès à un prêt, bénéficient également d'une meilleure connaissance et utilisation des données disponibles. Sans oublier l'optimisation des placements monétaires et la réduction des risques inhérents par une répartition optimale des portefeuilles suggérée par l'analyse des données financières.

Un secteur poussé par la réglementation

La connaissance et l'analyse des tickets de caisse des clients pourraient également ouvrir la porte à de nouveaux services : aide à la maîtrise du budget, suggestion d'achats, adaptation des primes d'assurances… mais ces données personnelles des clients ne doivent pas être manipulées sans précautions, et sans suivre le Règlement général européen sur la protection des données personnelles.

« *Concernant la prédiction du surendettement,* » explique Talel Abdessalem, enseignant-chercheur à Télécom ParisTech, « *les aspects réglementaires poussent les entreprises à faire appel au big data. La charte d'inclusion bancaire et de prévention du surendettement oblige les banques à essayer de comprendre si leurs clients sont en train d'aller vers du surendettement, et d'essayer de mettre en place les mécanismes pour les aider à en sortir.* » Les banques sont obligées d'avoir un tableau de bord sur leur activité. Elles doivent produire très rapidement certaines données et mettre en place les infrastructures techniques permettant cette réactivité. « *On demande aux entreprises de plus en plus d'agilité, et pour gagner en agilité il faut mettre en place en interne des mécanismes de remontée, de traitement et d'analyse de données quasiment en temps réel qui font appel à des technologies big data.* »

Un secteur FinTech à la pointe

Manipuler la donnée dans le secteur bancassurance n'est pas un travail austère. Pour imaginer les services innovants de la banque de demain et se distinguer de ceux de la concurrence, les données sont étudiées finement à la recherche de signaux faibles, et les technologies d'apprentissage machine, de réalité augmentée, les objets connectés ou encore les cryptomonnaies font partie du quotidien.

La donnée connectée

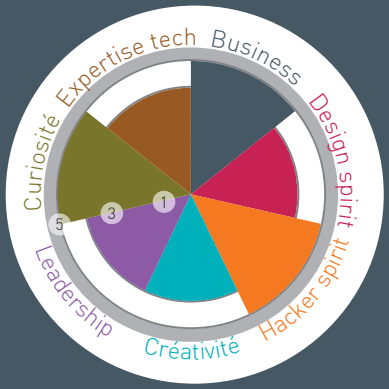
Société Générale
146 000 salariés
50 000 serveurs
30 Po de données stockées

« *Les données sont au cœur de la recherche d'innovations de rupture* » s'enthousiasme Emmanuel Bavière, responsable Centre d'Innovation chez Société Générale, « *Avec mon équipe de cinq personnes, nous jouons le rôle de catalyseur de la banque, pour tester rapidement des idées à travers des preuves de concept et répondre aux besoins des entreprises numériques et culturelles tels que le service client 2.0, l'amélioration de la carte de crédit, le paiement mobile, l'utilisation de robots, le monde 3D, les surfaces tactiles, l'intégration collaborative, la réalité augmentée, l'internet des objets, ou encore la blockchain…* »

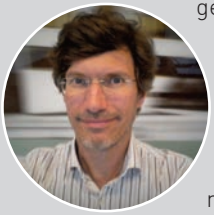
« *On expérimente, on essaye, on se plante rapidement s'il le faut, et surtout on apprend* » poursuit Emmanuel Bavière, dont le rôle de déchiffreur consiste à trier les idées en amont et à communiquer les résultats en aval. Savoir partager et diffuser ces expérimentations est en effet essentiel, et tant l'utilisation et la production de données en *open data* que les principes d'innovation ouverte sont des moteurs des data scientists de son équipe.

« La donnée moteur de l'innovation »

La banque s'est ainsi posé la question de l'utilisation d'un robot comme Pepper. Trois cas ont été explorés : lors de forums étudiants et forum professionnels ; dans l'utilisation interne chez Société Générale ; dans une agence bancaire. « *Dans les forums, par exemple, les étudiants apportent leur CV, et soudain il y a trop de monde. Le robot capte les personnes pour les aider, repère les personnes qui ont juste besoin d'une URL, ou renvoie vers le bon interlocuteur.* » Développé par des étudiants, ce système sera versé sur GitHub, et des données associées seront placées en *open data*. La même approche est retenue pour les agents conversationnels, pour lesquels l'interaction doit changer en fonction des comportements de chacun : « *On arrive à mieux valoriser les résultats en les ouvrant.* »



Retrouvez Emmanuel Bavière page 72



Au sein de la Banque de financement et d'investissement de la Société Générale, Marc van Oudheusden est *Head of Data Science*, responsable d'une équipe agile et homogène de data scientists sans rôles spécialisés, provenant du monde des stats, des maths appliquées ou de l'informatique, et travaillant en symbiose avec les process innovation de la banque. Il est également sponsor de l'équipe big data de la banque d'investissement, dont il précise les projets à traiter. Les sujets, nombreux, sont non triviaux.

La détection d'anomalies, par exemple, ne concerne pas que la fraude. Les équipes doivent certifier et valider des chiffres, détecter des saisies incorrectes, des données manquantes. Sur des classes de données de marché, la structure de la donnée est une courbe de taux, non uni-dimensionnelle. C'est un objet complexe sur lequel une recherche d'anomalies mal conduite pourrait entraîner de nombreuses fausses alertes. Il y a tout un travail de labellisation subtil à effectuer.

Un sujet récurrent est celui de l'automatisation du classement de données, pour détecter tel type d'événements sur tel périmètre d'observation, et fournir de l'aide à la décision pour conforter les clients dans leurs choix. La navigation de ces derniers sur les outils digitaux bénéficie du reste aujourd'hui de recommandations algorithmiques personnalisées, selon leurs demandes de cotations antérieures et leurs centres d'intérêt.

Le langage naturel fait également partie des projets de data science. Classer automatiquement des documents, zoomer sur les endroits à traiter sont des fonctionnalités très appréciées par le back office en charge de revoir des contrats.

Une organisation décentralisée

La Société Générale a une logique de data scientists répartis au plus près des métiers : banque de détail, banque d'investissement, assurances. Des relais Chief Data Officer ont été nommés dans ces lignes métiers, ainsi qu'à la direction financière, à la direction des risques, dans les grandes implantations…

Pour autant ces data scientists ne sont pas isolés. Organisés en communautés, ils s'appuient sur les services informatiques, où s'opèrent un nombre limité de datalakes. Ils se retrouvent pour échanger sur les pratiques et les méthodes.

Plusieurs centaines de personnes travaillent aujourd'hui autour de la donnée, et ce secteur recrute : des data scientists, des ingénieurs data, des data architectes, des data quality managers.

Ces derniers profils viennent de la maîtrise d'ouvrage stratégique, en charge de piloter la qualité de la donnée sur un périmètre. Ils analysent la qualité, et organisent des actions de remédiation, des campagnes qualité, ou bien des modifications sur le système d'information.

Créer les conditions pour développer les usages de la donnée

Dans une entreprise où le développement de l'usage de la donnée se fait partout, il faut un *Chief Data Officer*. Chez Société Générale, Emmanuelle Payan y occupe cette fonction. Rattachée à la direction des ressources et de l'innovation, son rôle de facilitatrice consiste à créer les conditions dans lesquelles le Groupe développe l'usage de ses données, dans le respect des principes de sécurité et de protection des données qu'ils se sont définis.

« *Nous avons désormais accès à des technologies plus performantes et des volumes de données plus massifs. Cela nous permet d'améliorer la personnalisation des services à nos clients, notre efficacité opérationnelle, être encore plus performant en matière de gestion de risques. En tant qu'acteur bancaire, nous souhaitons être un acteur responsable du traitement des données de nos clients. Cela peut nous amener à nous poser des questions de réaliser ou pas certains traitements.* » Sa mission est de développer la culture de la donnée, de créer la gouvernance (rôles et responsabilités), et de définir avec les métiers leurs principes d'action sur les données, ainsi que leur protection et leur mise en qualité… Dans son champ d'action également, les programmes réglementaires comme le RGPD (cf. p35).

Au plus près des métiers

S'appuyant sur les lignes métier pour les déploiements, Emmanuelle Payan les aide à définir la stratégie de la banque data. « *On veut être tiers de confiance : comment mettre cela réellement au cœur de nos projets ? Quel niveau d'information donner aux clients sur l'utilisation de la donnée ? Comment faire en sorte qu'un grand nombre de personnes aient une vraie culture de la donnée ?* » À longueur de déplacements, de conférences internes et de modules d'e-learning, sa mission est finalement simple: « *que chaque collaborateur dans le groupe devienne data fluent !* »

Travailler dans une banque

L'essentiel pour un jeune data scientist est d'être attaché à quelqu'un des métiers, qui a une vraie connaissance de son métier, voit la valeur qu'il tire des données, et le placera en face de vrais sujets.

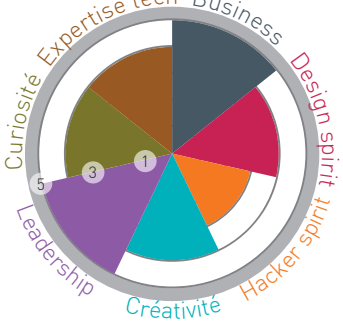
Un acteur bancaire est une industrie qui par son histoire traite beaucoup de données, beaucoup plus que dans une start-up, avec lesquelles on peut faire beaucoup plus de choses. C'est très intéressant pour les jeunes data scientists, car les cas d'usage sont assez variés, et les datasets sont passionnants.

Compte tenu des enjeux de recrutement, celui de garder des collaborateurs experts, et d'accompagner l'évolution des personnes qui ont des compétences statistiques, la banque a mis les moyens afin de créer les conditions pour s'épanouir dans un environnement agile. « *Cette attractivité passe par le travail en réseau et en communautés, les incitations à faire et organiser des hackathons et la tenue de Techweeks. Il y a toujours la possibilité au datalab et au Centre de compétences big data de travailler sur des tests, et de s'essayer à de nouvelles technologies.* »

Chief Data Officer

Le Chief Data Officer est le directeur des données de l'entreprise. C'est un cadre dirigeant qui participe au pilotage de la stratégie globale de l'entreprise, met en place les moyens et les équipes – notamment les data labs – pour qu'elle soit *data driven*. Ses missions sont également de s'assurer de la bonne collecte des données et de la transmission des informations les plus pertinentes pour la prise de décision. La gouvernance des données, leur contrôle, leur sécurité, leur niveau de confidentialité, la définition de leurs propriétaires, sont également rattachés à ce poste.

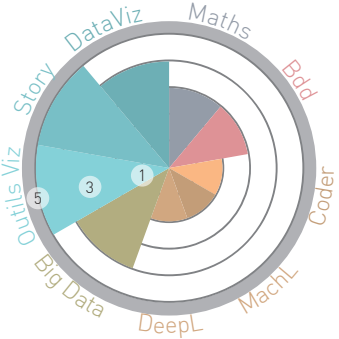
Présent principalement dans les grandes structures manipulant une grande variété de données et confrontées à de nombreuses réglementations, le Chief Data Officer est proche du comité exécutif, sans en faire nécessairement partie. Leurs échanges sont cependant réguliers, pour insuffler la culture des données dans le rythme de l'entreprise, comme ils le sont avec les directions métiers pour lesquelles il apporte les solutions data. Il est parfois rattaché au Chief Digital Officer, dont la mission plus globale est de piloter la stratégie de transformation numérique de l'entreprise. Ce poste va prendre à terme une plus grande importance dans les organisations.



Une fonction multiple, qui se crée et se transforme

Profil Formation type école d'ingénieur ou école de commerce

Excellente connaissance des technologies du big data, expertise métier. Ce métier étant assez récent, il est également possible d'y parvenir lors d'une évolution de carrière.



Compétences **Qualités :** capacité à convaincre, travail en équipe, solide expérience dans les domaines du management, de l'informatique et du marketing, connaissances des règlementations internationales et de leurs évolutions

Les missions des CDO

Par leurs postes, dont ils sont souvent les premiers titulaires, les Chief Data Officer sont confrontés à de très vastes ensembles de données provenant de tous les métiers et services de l'entreprise. À l'aide de leurs équipes de data scientists et des équipes métier, ils placent les données au cœur de la stratégie et de la culture de l'entreprise.

Organiser & manager

Fabrice Otañao, le Chief Data Officer du groupe Accor –92 pays, 4000 hôtels, 1400 tableaux de bord, des prévisions à 405 jours–, a eu carte blanche pour réorganiser des équipes auparavant dispersées (finances, distribution, IT, base de donnée référentiel, optimisation des prix, data scientists), soit 90 personnes en France et 700 dans le monde. Dans un secteur où la chaîne de distribution est fortement disruptée par des acteurs *pure player*, la donnée est centrale pour réussir la transformation digitale : « *Ma mission est triple :* think, *anticiper la vision business de demain ;* build, *construire les outils big data ;* run, *les faire tourner en continu et s'assurer que les optimisations trouvées passent à l'échelle industrielle.* »

« Il n'y a pas de fiche de poste définie à l'avance »

Évangéliser

C'est sans doute la mission centrale, précisent-ils tous, aller convaincre les clients internes : « *Nous avons des tableaux de bord, on peut à présent travailler mieux, qu'est-ce que vous pouvez utiliser de ces données que nous vous présentons ?* » Pour Michel Lutz, Group Data Officer chez Total, un des objectifs sur cette partie de la mission est la mise en production de modèles d'apprentissage statistique et l'amélioration des interfaces utilisateur donnant accès à la donnée et aux modèles.

Rendre le groupe data driven

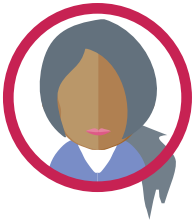
« *Que chaque collaborateur devienne data fluent* » comme le dit Emmanuelle Payan chez Société Générale, est en effet le graal. L'action du CDO est de diffuser la culture de la donnée dans toutes les équipes, de créer et de développer des réseaux, de manier les bonnes alchimies pour que des équipes et des données émergent les bonnes idées. Chez Accor, pour augmenter les revenus, on doit être plus prédictif, et pour cela élargir les capacités data du groupe. Des réseaux de *business analysts* passés auparavant chez Fabrice Otañao sont mis en place.

Des conseils pour ce poste ?

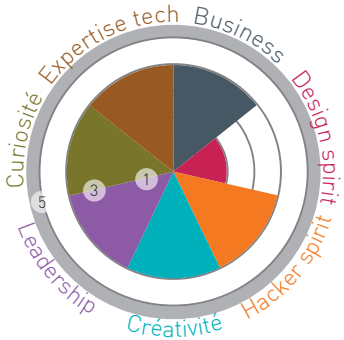
La donnée est agile, c'est une valeur qui va se diffuser dans les entreprises, qu'il faut aider dans cette transition culturelle. Pour Stéphane Ternot, EDF, il faut « *mettre en œuvre le collaboratif, savoir partager différents points de vue, et la valeur de la donnée augmentera. C'est pour cela qu'on monte des plateaux multi-compétences autour d'un problème.* » Pour Michel Lutz, « *il ne faut pas avoir peur de l'inconnu !* », et Stéphane Ternot ajoute : « *On peut se tromper, on fait avec naïveté et bienveillance.* »

« Dans nos équipes, la diversité des profils, c'est la clé »

Head of Data



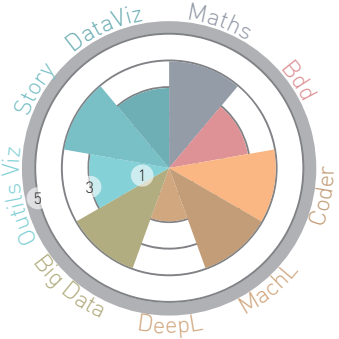
Fort de son expérience à la fois en sciences de données et en management, le ou la Head of Data est en charge de l'équipe data science. Data scientist d'origine, ces chefs d'équipe échangent avec les métiers de l'entreprise, pour en comprendre leurs modes de fonctionnement et identifier des cas d'usage, et ce qu'une meilleure utilisation des données pourrait leur apporter. Ils mettent en place des processus d'idéation qui permettent de faire émerger de nouvelles idées, et de repérer les données pertinentes à traiter ou collecter. En lien avec les direction métiers, ils travaillent alors à des solutions avec leur équipe de data scientists, en mode exploratoire. Ils ne développent plus nécessairement au jour le jour, et ont plus un rôle d'identification de problématiques à résoudre, souvent en binôme avec quelqu'un des métiers pour en acquérir la culture. Ils sont également en charge de trouver les sponsors en interne pour financer leurs travaux exploratoires, organisent des hackathons internes et mettent en place des data labs.



Créateurs de valeur, en lien avec les métiers de l'entreprise

Profil Formation mathématiques-statistiques ou informatique. L'orientation à dominante IT ou à dominante mathématiques dépend des enjeux de l'entreprise, de ses problématiques données, et de la taille des équipes.

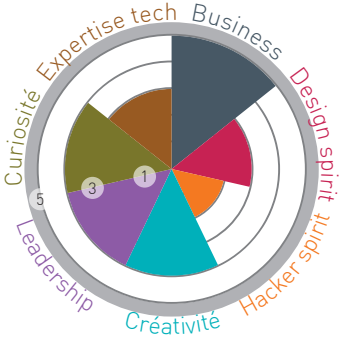
Compétences **Qualités :** bon relationnel, capacités de vulgarisation, aptitude pour le travail en équipe, rigueur, aptitudes managériales **Méthodes :** analyse de données, modélisation, machine learning **Outils :** Python, R, Java, Matlab, Spark, écosystème Hadoop (Hadoop, Hbase, Hive, Pig, Mapreduce…)



Chef de projet data

Grâce à leur connaissance des enjeux et des problématiques liés au big data, mais également des enjeux business sur un secteur ou une entreprise, les chefs de projet data gèrent les projets data de l'entreprise au quotidien. Ils font la liaison entre les profils IT et les profils plus opérationnels. Rattachés au responsable de l'équipe données (Head of Data), ils gèrent plus spécifiquement un aspect de la stratégie données de l'entreprise, comme la gestion et l'enrichissement du socle de données, ou bien le déploiement d'outils big data. Dans leur version « Data manager », ils sont également les garants de la conformité et de la bonne organisation des données dans les systèmes d'information : à la fois les données référentielles – liées aux catalogues fournisseurs, clients, articles, etc. – et les métadonnées structurantes – liées aux normes et aux règlements.

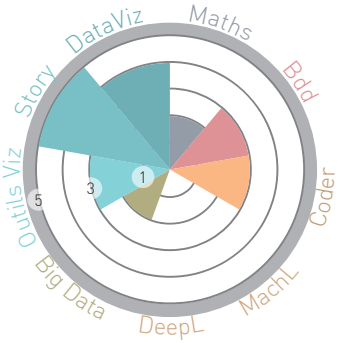
Opérationnels de la donnée et des projets orientés données



Profil

Formation type école d'ingénieur ou école de commerce

Selon les entreprises, le profil recherché sera plus ou moins technique. Très bonne connaissance métier, très bonne connaissance des enjeux data, forte orientation business

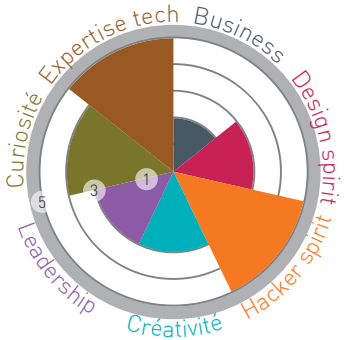


Compétences

Qualités : Capacités de communication et bon relationnel, aptitude pour le travail en équipe, leadership, autonomie, qualités d'écoute, d'analyse et de synthèse... **Outils :** Python, R, Java, Matlab, Spark, écosystème Hadoop (Hadoop, Hbase, Hive, Pig, Mapreduce...)

Architecte Big Data

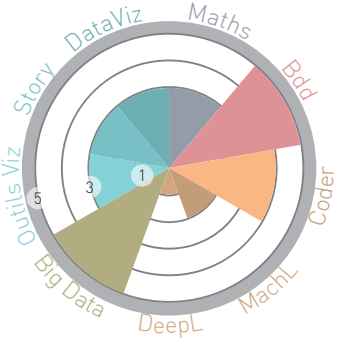
Les architectes big data interviennent le plus en amont de l'organisation du traitement de la donnée, en lien avec les équipes informatiques et les managers de la donnée. Leur rôle est de mettre en place toute l'infrastructure technique nécessaire à la collecte et au traitement de gros volumes de données. Ils élaborent des schémas de systèmes de gestion de données qui facilitent l'acquisition et la circulation des données, qu'ils affinent et surveillent ensuite en permanence. Ils développent également l'inventaire des données et les modèles de données. Grâce à leur vision d'ensemble des technologies big data, ils assurent la cohérence de la structure des bases de données et celle des frameworks, afin qu'ils soient en phase avec les besoins de l'organisation et restent adaptés aux enjeux de l'entreprise.



Organisateurs et responsables de la vision d'ensemble des données

Profil Formation IT

Vision d'ensemble de l'écosystème technique, forte expertise technique. Bonne compréhension des enjeux métiers et des problématiques de data science. Ces profils peuvent également être des consultants freelance.



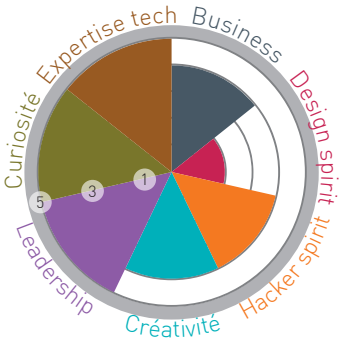
Compétences **Outils :** base de données NoSQL, écosystème Hadoop, Spark…

Qualités : Curiosité intellectuelle, autonomie, communication…

Chief Technology Officer

Rattachés à la direction générale dans les entreprises, et éléments clé dans les start-up, les Chief Technology Officers sont des responsables de haut niveau expérimentés, qui ont pour rôle d'insuffler la culture de l'innovation dans leurs organisations. Ils ont une vision globale du système d'information de l'entreprise et sont un levier essentiel pour le développement de la compétitivité de leur entreprise. Ils repèrent et testent, ou développent avec les équipes de la direction des systèmes d'information, les outils et les solutions technologiques innovantes, en pilotent leur mise en œuvre et leurs évolutions. Ils s'impliquent dans la R&D et font le lien avec les start-up. Comme le métier de chef de projet informatique, il ne s'agit pas a priori d'un métier directement orienté data, mais être expert des technologies big data devient de plus en plus indispensable pour les CTO.

Apporteurs d'innovation

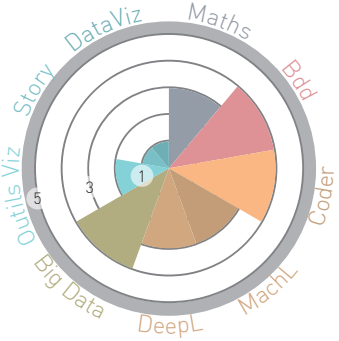


Profil

Formation école d'ingénieur

Spécialistes des usages des nouvelles technologies, à la pointe de l'innovation et de la recherche. Une expérience significative dans la direction de projets informatiques est nécessaire.

Compétences



Expertise en gestion de grands projets et de projets transverses, appétence pour la R&D, compréhension des enjeux juridiques de l'utilisation des données...

Qualités : leadership, curiosité, expertise technique, goût des défis, organisation, rigueur

Des données sous haute protection

Vol de données d'entreprise ou de données personnelles collectées et stockées par les acteurs de la nouvelle économie, piratage et détournement d'objets connectés, corruption de données en amont pour tromper les algorithmes, ces quelques cas dont l'actualité se fait l'écho régulièrement montrent à quel point la sécurisation des données est une préoccupation de tous les instants.

Deux métiers de la data science sont directement liés à la sécurité et à la protection des données : l'expert sécurité, qui intervient principalement dans les couches informatiques et télécoms, et doit anticiper les risques augmentés ou créés par l'utilisation des big data, et le *Data protection officer*, un métier actuellement en pleine définition, qui doit prendre en compte l'évolution des contraintes réglementaires, et mettre en place la nécessaire sensibilisation des collaborateurs.

La sécurisation des données s'effectue à tous les stades de la chaîne de traitement de la donnée, et relève de chacun des V du big data. Leur volume et leur vélocité, leur traitement en temps réel, nécessitent des infrastructures de stockage et de calcul de haut technicité, qui ne se situent pas nécessairement dans les locaux de l'entreprise et impliquent donc également leur transport. C'est autant d'angles d'attaque, qui peuvent aller du déni de service au vol de données. La variété des données, et la nécessité de les croiser pour en trouver les caractéristiques utiles, est la source de risque de désanonymisation de ces données, ou de faiblesses dans la protection intellectuelle. Des données corrompues, des données de mauvaise qualité, ont également un impact sur la véracité des données traitées, dont l'utilisation conséquente peut porter atteinte à la réputation de l'entreprise et à la confiance que ses partenaires mettent en elle.

Se former à la sécurité

La sécurité est l'affaire de tous, et comme celle des données doit s'effectuer sur l'ensemble de la chaîne de traitement, les data scientists sont concernés dans toutes les déclinaisons de leur métier.

Compte-tenu des enjeux, et des techniques d'attaque qui ne cessent d'être imaginées, les diverses formations big data de Télécom ParisTech et de Télécom Evolution disposent toutes d'un volet sécurité. « *En matière de cybersécurité* », explique Ons Jelassi, en charge du domaine big data à Telecom Evolution, « *on collecte des informations sur les attaques qui ont déjà eu lieu, on a à disposition des données de vulnérabilité, des données de menaces sur les systèmes de sécurité en place… Plus on a d'informations de ce type et plus les actions de contre-mesure vont être efficaces.* » Et pour sensibiliser les élèves, les formations font appel à des intervenants du terrain, pour qui la sécurité est le quotidien : « *Parmi eux, un commandant de l'armée de terre vient faire des cours à nos étudiants sur l'utilisation du machine learning dans la cyberdéfense.* »

Assurer à l'ère des big data

L'arrivée des big data a fortement changé le métier des assurances, dont la principale composante est l'analyse et l'évaluation du risque, et donc de savoir faire un pari sur l'avenir. La collecte de données de plus en plus proches des assurés, via par exemple des objets connectés, l'accès à des données météorologiques plus précises et à des données géospatiales enrichies (voir page 74), les systèmes prédictifs de plus en plus performants et les analyses comportementales effectuées sur de grandes cohortes d'assurés, ont ouvert la voie à des produits d'assurances personnalisés et avancés, avec parfois le risque d'aller trop loin, ou d'être soupçonné par les clients d'en savoir trop sur eux.

Concernant les systèmes prédictifs, le professeur Talel Abdessalem, Télécom ParisTech, observe : « *Il y a la prédiction du risque, et puis il y a la détection d'anomalies et la maintenance prédictive, c'est-à-dire comment prédire qu'un équipement, un logiciel, un système risque de tomber en panne ou d'entrer dans une phase de fonctionnement anormal, pour pouvoir résoudre le problème à l'avance. Ceci est également encouragé par l'internet des objets, le développement des capteurs, qui fournissent plus de données sur le fonctionnement des équipements.* » Ce qui s'applique au matériel pourrait bien s'appliquer aux humains également, le métier d'assureur évoluant pour devenir coach de l'assuré et l'aider à se comporter de telle manière à éviter de s'approcher des situations de risque, qu'il s'agisse de conduite automobile ou de gestion de sa santé. Le suivi au plus près, et notamment en matière de géolocalisation, exige cependant le respect de règles déontologiques strictes. Et les assureurs doivent compter avec les grands acteurs de l'internet, qui en savent à présent également beaucoup sur leurs clients, et pourraient venir bouleverser leur relation.

Executive MBA « Manager data scientist des métiers de l'assurance »

Comme pour les métiers liés à la sécurité et à la protection des données, les métiers de data scientists liés au risque et aux assurances peuvent bénéficier de formations spécialisées.

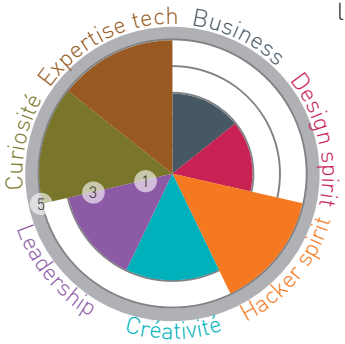
Conçu en partenariat avec l'École Polytechnique d'Assurances, l'Executive MBA « *Manager data scientist des métiers de l'assurance* » s'adresse à des ingénieurs, techniciens, chefs de projet, informaticiens, statisticiens, mathématiciens souhaitant développer leurs compétences et encadrer des équipes spécialisées dans le big data appliqué à l'assurance.

La formation compte 67 jours en présentiel sur 13 mois à raison d'une semaine par mois. Le programme repose sur 4 piliers : technique, à la fois en data science et sur les métiers de l'assurance, stratégique, éthique et management. La validation de la partie data sciences du MBA donne lieu à la délivrance par Télécom ParisTech du Certificat d'Études Spécialisées « Data Scientist ».

www.epassurances.fr/emba

Expert sécurité

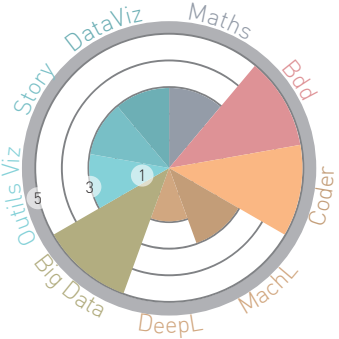
Dans un contexte global où les cyberattaques se déploient à grande échelle, les experts sécurité (ou cybersécurité) sont les maîtres d'œuvre de la politique de sécurité d'une entreprise. Ils évaluent le niveau de vulnérabilité des systèmes d'information et des systèmes de gestion de données, que ce soit lors de l'acquisition des données, leur transport, leur traitement ou leur stockage, rédigent les politiques et les standards de sécurité, préparent les solutions pour les sécuriser et administrent les droits d'accès au réseau et aux données. Ils doivent également mettre en échec les tentatives d'intrusion ou de déni d'accès à ces systèmes. Ils ont un rôle de prévention et de remontée de risques en amont, de détection et de lutte en direct, et d'explication et de réparation en aval pour assurer la continuité de l'activité. Ils effectuent une veille technologique poussée pour anticiper les futures défaillances ou attaques, et peuvent se transformer en attaquant pour chercher et repérer les failles de leurs systèmes afin de mieux les contrer.



Protéger les données des cyber attaques

Profil Ingénieur ou Master en informatique, en télécommunications & réseaux, en systèmes d'information, avec une spécialité ou une dominante sécurité

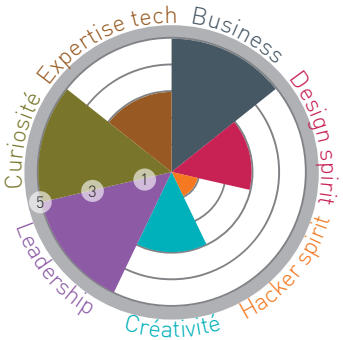
Compétences Organisation, rigueur, communication et pédagogie



Data protection officer

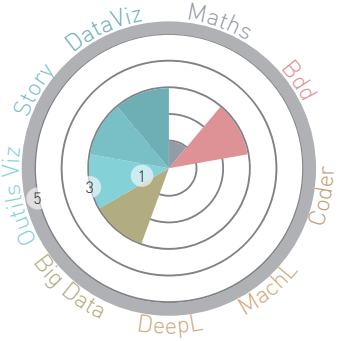
Pour piloter la gouvernance des données personnelles, et en vertu des règlements européens, il est recommandé et parfois obligatoire, selon la taille des organismes, de désigner une personne ayant une mission d'information, de conseil et de contrôle en interne : le ou la délégué.e à la protection des données. En France ce rôle prendra la suite de celui, bien connu, du correspondant informatique et libertés. C'est une première étape essentielle pour se préparer à l'arrivée du Règlement général sur la protection des données personnelles. Ce métier comprend un volet essentiel de sensibilisation des collaborateurs aux implications juridiques de l'utilisation des données. Son défi est de se tenir au courant de tous les projets lancés autour des données, pour pouvoir y apporter des préconisations suffisamment en amont.

Protéger les données confiées, dans le respect des règlements



Profil

Évolution du correspondant informatique et libertés ou nouveau poste créé, les profils sont plutôt issus d'écoles de commerce ou juristes, avec plusieurs années d'expérience dans ces domaines.



Compétences

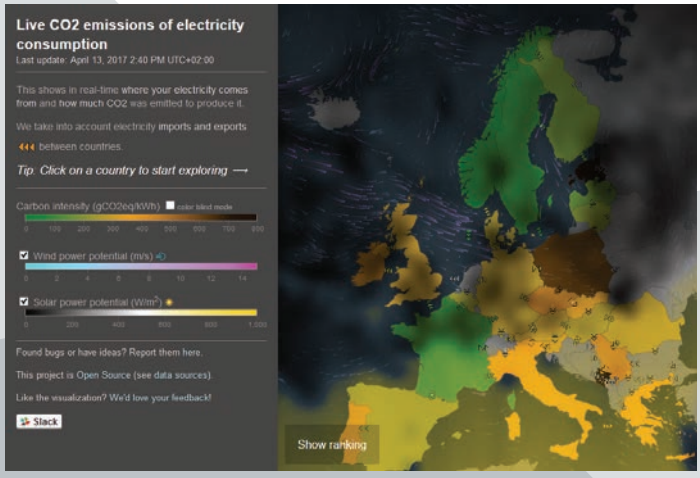
Culture générale, culture internationale, connaissance et suivi des règlements et des normes, connaissances juridiques, capacités de communication et de conviction

Libérer l'énergie des données

En ce début de XXIe siècle, la transition énergétique, un des volets de la transition écologique, passe par une prise de conscience de la crise énergétique et la possibilité de choix éclairés. Les données permettent d'observer et de prédire simultanément et à toutes les échelles les impacts de nos décisions énergétiques, de nos usages et de nos gestes du quotidien.

La vision du producteur

Stéphane Ternot est un data scientist du secteur de l'énergie. Entre 2013 et 2015 il fonde le datalab de GRDF, dont l'un des projets permettra à ce distributeur de mieux cibler, en mode prédictif, des ouvrages potentiellement défaillants sur sa chaîne de communication gazière. Il rejoint ensuite EDF où il développe une approche similaire au sein de la direction optimisation amont-aval trading, une entité très opérationnelle de 400 personnes. « *Les données que nous traitons sont des séries temporelles pour l'ensemble des moyens de production d'EDF, nucléaire, thermique à flamme, hydraulique… Le travail de cette direction consiste à élaborer et optimiser le programme de fonctionnement de ces moyens de production, du pluri-annuel jusqu'à H-1, à un pas de 30mn. Les angles d'analyse sont multiples : temporels, échéances, fiabilité des données, types de production, valeurs du prix de marché.* » Dans un premier temps un cluster Hadoop est déployé, où les données sont envoyées en masse avant d'être mobilisées. Une plate-forme logicielle est ensuite utilisée pour sortir des indicateurs de pilotage de ces données. C'est alors une collaboration entre métiers et SI, et non plus la seule DSI qui teste les jeux de données. « *En 2016, l'objectif devient de partager les données au plus grand nombre. Il s'agit de faire de la donnée sur les usages de la donnée. Qui utilise quoi pour faire quoi ? Qu'est-ce qui apporte le plus de valeur ?* » Cet outil attire l'attention d'autres personnes travaillant avec la R&D. Data scientists et consommateurs de données se retrouvent autour d'un wikipedia de la donnée, qui facilite la recherche et indique quelle donnée est utilisée par quel projet. Il devient possible de faire des études très rapides et d'être très réactif.



Objectif: quantifier

Un mix énergétique ne se construit pas de manière isolée. Les pays échangent entre eux l'énergie qu'ils produisent à partir de sources diverses. Cette carte, fondée sur des données publiques temps réel, permet d'en visualiser les effets.

www.electricitymap.org

Pour le producteur d'énergie, l'objectif est de mieux comprendre un portefeuille client qui évolue. En effet, les changements de comportements, les possibilités d'autoconsommation, font que la baisse globale de la consommation individuelle n'est pas uniforme, rendant plus complexe encore la prévision des moyens à assurer à court et long termes. « *Il s'agira de confirmer ou infirmer des choses cachées, d'identifier des singularités signifiantes qu'on n'aurait pas vues autrement, de faire des requêtes sur des données bizarres, en méthode agile et rapide.* » Tout est souvent question d'échelle. À grosse maille géographique s'expriment des moyennes, alors qu'une observation plus fine à cheval sur plusieurs zones fera émerger des dissimilarités invisibles. Disposer de ces grands volumes de données sous forme graphique est par ailleurs essentiel.

La vision du distributeur

Collecter les données et faciliter les changements de comportement se fait aussi chez l'habitant. C'est dans cette optique qu'Enedis, qui exploite 95 % du réseau de distribution électrique français, est engagé sur un grand projet de remplacement des compteurs par des compteurs communicants, avec la possibilité pour le client de suivre sa consommation sur un site Internet, mieux la comprendre et agir pour la maîtriser.

Objectif: smart grids

Pour accompagner la transition énergétique, la croissance de la production d'énergies renouvelables et les nouveaux usages comme les véhicules électriques, le réseau de distribution d'électricité évolue vers un réseau dit intelligent. Une directive européenne a fixé l'objectif de déployer des compteurs communicants dans 80 % des foyers européens d'ici 2020. Enedis envisage d'en remplacer 90 % dans 35 millions de foyers en France d'ici 2021.

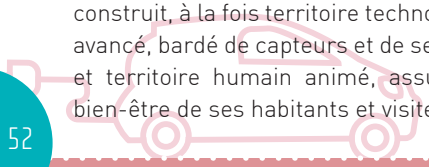
Pierre Achaichia est responsable de l'activité data science (dix data scientists), rattaché à Pierre Gotelaere, manager de l'activité data & analytics (70 personnes environ, mixant compétences IT, data analysts pour la qualité des données, et les data scientists), pour le système Linky. « *Nos données proviennent du matériel Linky et de la chaîne communicante Linky. Certaines sont structurées, d'autres non. On trouve des données tabulaires classiques, des données techniques, des données orientées graphe et des séries temporelles. Ce sont par exemple les données du déploiement, comme la date de changement de compteur, l'entreprise de pose… Dans tous les cas, la confidentialité et l'anonymisation de ces données sont centrales.* » Toutes ces données permettent à Enedis d'automatiser les processus de supervision du système Linky, par exemple pour faire de la maintenance prédictive, et développer de nouveaux services pour les métiers du distributeur, comme la détection des incidents réseau en temps réel.

Se trouver à la fois au coeur d'un projet de déploiement d'objets connectés de grande ampleur, et d'un enjeu de transition de portée mondiale, voilà le type de chantiers de longue haleine qui permet de voir toutes les facettes de la data science. « *Au delà de la technique et de sa maîtrise, l'état d'esprit, la curiosité, la volonté de travailler en équipe et de s'ouvrir et partager sont primordiales pour réussir de tels projets de data science* », soulignent les deux équipiers d'un même élan.

La ville, terrain de jeux de données

Les 35 millions de compteurs Linky à terme sont autant de capteurs pour renseigner sur l'état du réseau basse tension, et préparer ce réseau à l'intégration en masse des énergies renouvelables et des véhicules électriques. Pour les collectivités locales, les aménageurs du territoire et la puissance publique, se profile une meilleure connaissance des flux d'énergie et un patrimoine d'équipements suivi avec plus de précision.

Le *smart grid* est un des cœurs battants de la *smart city*, cette ville sensible qui se construit, à la fois territoire technologique avancé, bardé de capteurs et de senseurs, et territoire humain animé, assurant le bien-être de ses habitants et visiteurs.



D'autres ensembles de capteurs et jeux de données sont également présents, utilisés ou produits par les collectivités, les entreprises –en délégation de service public ou non– ou les citoyens. En partageant ces données, ces acteurs permettent l'émergence de nouveaux services, parfois issus du croisement improbable de données. Lors de hackathons, des équipes hybrides réunissant toutes les compétences de la data science, du développement à la visualisation, de la connaissance métier à celle des usages, de la fouille à l'exploitation agile, créent de nouvelles synergies. Ces rencontres, qui ne se limitent d'ailleurs pas qu'aux enceintes de la ville, et se font dans les territoires ruraux également, ou à cheval entre territoires, sont

La data science, des statistiques à l'apprentissage machine

Pour Stephan Clémençon, professeur à Télécom ParisTech et responsable de la chaire *Machine Learning for Big Data* (voir page 57), l'exemple de la ville connectée est typique d'un déploiement à plusieurs étages de la data science.

« *Quand on parle de ville connectée et de transports intelligents, cela sous-entend qu'on collecte des données sur les usages et le fonctionnement des infrastructures. Ces données sont d'un volume considérable, elles sont collectées en continu et il faut pouvoir les stocker. Elles le sont souvent de manière assez brute, avec énormément de formats différents. Seules les infrastructures et technologies big data permettent de stocker ces données et d'y effectuer des requêtes de manière optimale ensuite.*

Dans un premier temps, leur traitement peut être basique et ne demande pas un bagage avancé de data science. Cela peut prendre la forme de simples statistiques descriptives (nombre de voyageurs), mais c'est généralement plus complexe, car l'information est massive. Il faut rapidement des outils de data visualisation pour la rendre intelligible et permettre de monitorer et analyser le fonctionnement de l'infrastructure. Or, on attend souvent un peu plus que tout cela : calculer des prédictions pour anticiper les pics d'affluence par exemple. On souhaite que ces données permettent d'optimiser un certain nombre de décisions. C'est là que le machine learning intervient. »

l'occasion pour des data scientists aguerris comme pour de futurs professionnels d'échanger, de partager des points de vue, d'aller sur le terrain en situation et de tester de nouvelles techniques et approches. L'ouverture des données est un appel à l'ouverture des esprits.

Avec le développement des usages, certains de ces services vont plus loin que leur visée initiale. Waze ne promet plus seulement aux conducteurs de trouver un trajet dégagé. L'application filiale d'Alphabet aide aussi telle ville à réguler son trafic, telle autre à alimenter ses données prédictives, elle équipe tel constructeur de capteurs, permet à tel assureur d'ajuster ses primes en fonction des pratiques de conduite. Sous l'impulsion de la dissémination et de la gestion partagée de ces données, et avec l'apparition de data services privés, la frontière se brouille entre le champ d'intervention des acteurs publics et privés sur un territoire de données qui ne sont pas encore un bien commun.

Passé par la mission Etalab (voir page 22) Romain Lacombe fait partie de ces jeunes entrepreneurs pionniers qui mobilisent sur des enjeux de société les données publiques et ouvertes, ainsi que celles que chacun de nous peut produire, en s'appuyant sur le développement d'objets connectés et d'applications de visualisation. Début 2017, sa start-up Plume Labs a ainsi dévoilé Flow, un capteur personnel mobile permettant de mesurer le degré de pollution auquel chacun s'expose chez soi, à l'extérieur ou dans les transports. Il y avait le *quantified self* pour mesurer les données de son corps, il y a maintenant le *quantified environement*.

Entrepreneur de la dataviz

Rendre la ville intelligible à travers des data visualisations est une nécessité. Ces outils permettent également aux aménageurs de l'espace public de mieux analyser la performance des dispositifs qu'ils gèrent. Intégrant huit sources de données distinctes –patrimoine, qualité, audience, cibles socio-démo et géo-comportementales, points d'intérêt et points de vente, profils consommateurs, bases d'images et de textes–, la start-up Toucan Toco a créé en 2016 une application permettant à 250 utilisateurs commerciaux et marketing au sein des équipes JCDecaux d'accéder à ces données en mobilité.

Son fondateur, Charles Miglietti, est un data entrepreneur et un pédagogue de la donnée. « *J'ai développé et codé le produit de data visualisation de Toucan Toco, qui est née en 2014. Notre solution est un outil de reporting visant à simplifier la compréhension de la donnée par les néophytes en entreprise. Aujourd'hui, mon activité se partage entre le commercial et le conseil en dataviz, le management, et le développement et l'architecture de notre solution.* »

Intervenant dans les formations big data de Télécom ParisTech, Charles Miglietti incarne avec passion deux aspects fondamentaux de la data science : entreprendre pour aller explorer le monde des données et en tirer les ressources d'avenir, et savoir donner la parole à ces données pour qu'elles expriment toutes leurs saveurs.

De l'open data à l'open innovation

Avant de pouvoir faire le story-telling des données de ses clients, les connecteurs de Toucan Toco agrégent les données venant de leurs systèmes d'information et de leurs bases de données. Des données externes à leurs applications, disponibles en open data, « *comme les données macro-économiques ou la météo qui peuvent influer sur la performance des activités de nos clients,* » précise le data entrepreneur Charles Miglietti, sont également traitées, enrichissant la base initiale.

Ces données ouvertes et les nombreux outils en logiciel libre sont une des raisons qui permettent de se lancer rapidement dans un projet entrepreneurial, y compris pendant ses études. « *L'investissement est relativement modéré,* » relève Stephan Clémençon, professeur à Télécom ParisTech, « *une idée de services, élaborée à partir d'algorithmes, et des compétences intellectuelles suffisent. Les services de Cloud permettent d'exploiter des infrastructures qu'on ne possède pas et qu'on va simplement intégrer dans le business model de la start-up.* » Le professeur Talel Abdessalem ajoute : « *Les personnes qui se lancent dans des start-up big data sont motivées par les success stories. Certaines travaillent sur des outils open source, cherchant à développer un outil qui devienne une référence, utilisé par le plus grand nombre de personnes, et cherchent ensuite un business model. Certaines le sont aussi par le développement d'outils ayant un impact.* »

Producteurs et consommateurs de données à grande échelle, les services de l'État et les grandes entreprises sont friands de rencontres avec les étudiants en data sciences et les start-up, prenant des formes diverses comme des hackathons, des challenges ponctuels ou à plus long terme. L'État a ainsi lancé fin 2016 le programme « Entrepreneur d'intérêt général », recrutant 10 personnes sur 10 mois pour résoudre, grâce aux données, des défis d'intérêt général au sein d'administrations pionnières.

De son côté, SNCF met à disposition nombre de ses données en open data sur un site dédié et animé. Celles et ceux qui souhaitent prototyper, coder ensemble et accéder à des données de transport très riches, comme les horaires planifiés et temps réel, les équipements et services en gare, la régularité des trains, l'accessibilité des gares, sont régulièrement conviés à relever des défis dans des événements. Certains projets fil rouge 2017 du Mastère Spécialisé Big Data de Télécom ParisTech ont d'ailleurs bénéficié de cet accélérateur d'innovation au service de la mobilité que représente cette initiative.

Télécom ParisTech héberge une spin-off, Score4Biz, qui s'appuie sur des technologies propriétaires brevetées développées par des équipes de statisticiens et de traitement big data de l'École. Son objectif : permettre aux entreprises d'améliorer la performance de leur activité, en répondant à des questions business à fort enjeu, par application de technologies d'analyse big data aux données des entreprises.

Des start-up de la donnée

L'incubateur de Télécom ParisTech, ParisTech Entrepreneurs, a accueilli en quinze ans plus de 400 projets innovants du numérique. Beaucoup d'entre eux utilisent les technologies du big data. Dans l'analyse et le traitement des données : Invenis, Lefty, Datapred, Linkurious, DCbrain, Predictice. Dans le marketing et le e-commerce : Botfuel, Adomik, Vigicolis, Catalisio, Beyable. Dans l'industrie et les transports : iDMog, Safety Line. Dans la santé : Dreamquark. Dans l'innovation et la hight-tech : Stim, L2 Technologies, Sevenhugs. Dans les services Internet : Ownpage, FocusMatic. Zoom sur quatre de ces start-up :

botfuel

Botfuel développe et commercialise une plate-forme de développement de « chatbots » à destination des entreprises qui cherchent à transposer un service existant sous forme conversationnelle. Cette plate-forme cible les besoins qui ne sont pas bien couverts par les plate-formes destinées à une plus large audience : gestion des conversations complexes, montée en charge, testabilité, internationalisation et confidentialité des données.

Predictice

Predictice propose un outil d'aide à la décision destiné aux professionnels du droit. Son objectif est d'accroître la transparence, la prédictibilité et la performance de la Justice et de ses acteurs. Son algorithme de justice prédictive permet de calculer facilement les chances de succès d'un litige, le montant des indemnités ou d'identifier les éléments les plus influents. La solution permet également d'accéder à la jurisprudence et aux textes de loi via une barre de recherche en langage naturel.

DCbrain

DCbrain propose une solution d'intelligence artificielle dédiée aux problématiques des gestionnaires de réseaux complexes (eau, gaz, électricité, vapeur, logistique). Ces acteurs ont les mêmes enjeux : comment équilibrer et optimiser les réseaux. Les outils classiques de gestion de réseau ne permettent pas cela : ces outils ont d'abord été créés comme système d'alarming et de planning. DCbrain permet de visualiser l'ensemble des flux s'écoulant dans un réseau, d'identifier et de prédire des anomalies et de modéliser des évolutions de réseau.

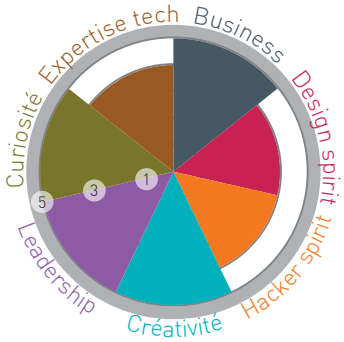
Invenis

Invenis propose un logiciel d'analyse de données issu de technologies big data et d'intelligence artificielle, simple, performant. À destination des analystes et des équipes métiers, Invenis offre la possibilité d'exploiter tout le potentiel de ses données au service d'une meilleure performance business, en toute autonomie. Grâce à des algorithmes de machine learning, les analyses prédictives deviennent accessibles à tous.

www.paristech-entrepreneurs.fr

Data entrepreneur

La possibilité d'accéder à des bases de données en open data de plus en plus nombreuses, et à des capacités de stockage, de calcul et d'apprentissage machine en ligne à des coûts réduits, est une véritable aubaine et une opportunité pour créer de nouvelles activités et chercher à capter la valeur de ces données. Grâce à ces technologies et données facilement accessibles, il est aujourd'hui possible de lancer des activités innovantes avec une petite équipe de passionnés, avec un investissement de départ réduit, et d'avoir une réelle capacité à passer à l'échelle et devenir un acteur international en quelques années. Les entrepreneurs qui se lancent sur ce territoire défrichent de nouvelles pratiques et de nouveaux usages au cœur des différentes transitions, numériques bien sûr, mais également écologiques, énergétiques et sociales…



Explorateurs et défricheurs d'un nouveau monde

Profils Tous profils.

L'équipe d'une start-up data réunira tous ces types de profil. Les *profils business* sont en charge des business plans et business models (positionnement sur le marché), la stratégie commerciale, la stratégie marketing / communication. Les *profils techniques* sont en charge du développement produit et responsables techniques.

Compétences Savoir-être, dynamisme, ambition, culture générale, goût du risque, facultés d'adaptation, imagination et curiosité

Vers la transition cognitive

Transition énergétique, transition écologique, transition sociale, toutes reposent à un degré ou à un autre sur la transition numérique. Cette dernière n'est pourtant que l'arbre qui cache une forêt bien plus vaste et profonde : celle de la nécessaire transition cognitive, le passage de tous nos objets, nos usages, nos activités, notre quotidien et notre société vers des états où les systèmes d'intelligence artificielle augmentent nos capacités intellectuelles individuelles et collectives.

Tous les pans de la société sont en train de se *cognitiser*, comme il y a un siècle ils se sont électrifiés, et donner un nom à ce phénomène c'est affirmer qu'il faut s'y préparer comme on s'est préparé et adapté au numérique. Cette transition ne remplace pas pour autant la nécessaire transition numérique ; elle en est à la fois l'étape suivante et le niveau supérieur. Elle ne se limite pas aux seuls systèmes d'intelligence artificielle et aux techniques d'apprentissage machine ; elle rappelle qu'il faut prendre en compte toutes les intelligences humaines, tant individuelles que collectives, dans leur diversité, dans leur provenance et leur mobilité.

En englobant toutes ces intelligences – et on peut y ajouter la reconnaissance des spécificités de la cognition animale –, là où la transition numérique actuelle est plus un substrat technique et socio-technique, la transition cognitive serait un moteur de l'évolution humaine. Et dans ce mouvement d'ampleur, les data scientists doivent prendre en compte, comme le rappelle Alain Abramatic (lire page 6), les « *réticences face à ces transitions qui nécessitent une charge cognitive trop importante et qui apparaissent comme très intrusives. Il y a un réel effort à faire pour inventer des services qui apportent une véritable aide et qui soient faciles à comprendre.* »

Chaire Machine Learning for Big Data

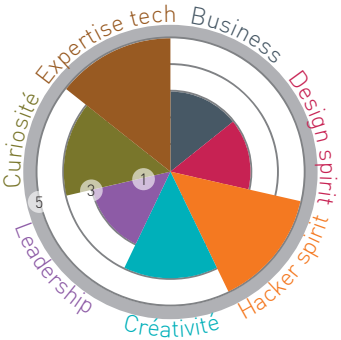
Créée en septembre 2013 avec le soutien de la Fondation Télécom et financée par quatre entreprises partenaires, Safran, PSA Groupe, Criteo et BNP Paribas, la Chaire « Machine Learning for Big Data » portée par le professeur Stephan Clémençon vise à produire une recherche méthodologique répondant au défi de l'analyse statistique des données massives et d'animer la formation dans ce domaine à Télécom ParisTech.

La variété, la volumétrie et les dimensions des données disponibles rendent en effet inopérantes les méthodes statistiques traditionnelles. Les data scientists spécialisés en apprentissage machine élaborent et étudient les algorithmes permettant à des machines d'apprendre automatiquement à partir des données, à effectuer des tâches de façon performante, et assister les humains dans leurs utilisations de la donnée.

machinelearningforbigdata.telecom-paristech.fr

Machine learning specialist

Ces data scientists se spécialisent sur les techniques et algorithmes d'apprentissage automatique sur les données et les systèmes d'intelligence artificielle. Ils développent des algorithmes avancés pour étendre les capacités des outils de traitement big data déjà en production. Ils utilisent ces techniques pour trouver des caractéristiques nouvelles dans les ensembles de données, pour créer de nouveaux modèles de données, pour développer et évaluer des modèles prédictifs. Ils élaborent et testent des hypothèses, puis analysent et interprètent les résultats. Ils produisent des solutions à partir d'une analyse exploratoire d'ensembles de données multi-dimensionnels et complexes. Compte-tenu des avancées actuelles dans ces domaines, ils effectuent une veille continue et collaborent avec les équipes de R&D.

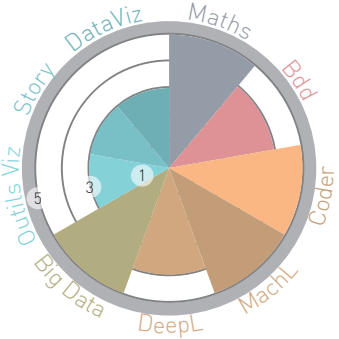


Explorateurs de données complexes et chercheurs de nouvelles pistes

Profil Masters ou Thèses en informatique, sciences cognitives, statistiques, mathématiques...

Compétences Esprit d'initiative, esprit pratique

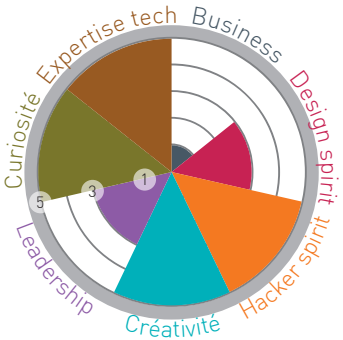
Méthodes : Deep Learning, Random Forests, Modèles de Markov cachés, SVM, Regression, Séries temporelles, Traitement du signal...



Chercheur en data sciences

Faire de la recherche en data sciences peut signifier plusieurs choses. Il y a tout d'abord les chercheurs qui font progresser les techniques et technologies au cœur des différentes disciplines qui composent les data sciences. Il y a ensuite celles et ceux qui utilisent les techniques de la data science pour faire avancer les recherches dans leur propre discipline, selon les nouveaux paradigmes scientifiques pour lesquels la donnée, puis depuis peu les systèmes d'intelligence artificielle, sont leurs nouveaux instruments. Il y a enfin des chercheurs pour lesquels l'évolution de la data science est un champ de recherche en soi : designers, juristes, sociologues, économistes... Les avancées de la data science se faisant à grandes enjambées et à l'échelle mondiale, ces chercheurs publient en open source et en open data les logiciels et données qu'ils développent, et leurs articles scientifiques suivent souvent les principes de l'open science.

Inventeurs du futur de la data science

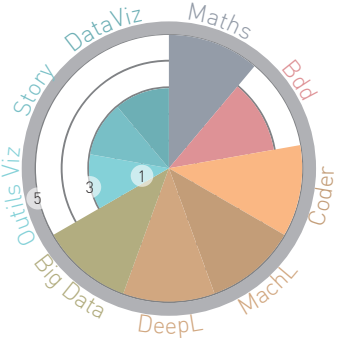


Profil

Masters ou Thèses en informatique, sciences cognitives, statistiques, mathématiques...

Compétences

Ouverture d'esprit, curiosité, transdisciplinarité



La Recherche en Data Sciences

Les données traitées au sein de la chaire Big Data & Market Insights (voir page 12) proviennent de multiples sources, comme des capteurs, des réseaux sociaux, ou des traces laissées en ligne. Entretien avec son responsable, Talel Abdessalem, professeur à Télécom ParisTech, sur les grands sujets abordés par son équipe.

« L'intelligence artificielle devient une thématique de recherche de plus en plus présente, qui intéresse directement les entreprises. Le succès des applications de deep learning, sur des données de type texte ou image, inspire les chercheurs qui essayent d'adapter et d'étendre les travaux à d'autres types de données et également d'améliorer les techniques existantes.

L'Internet des Objets sera probablement dans les prochaines années la source de données la plus importante. Quand on a des volumes de données de grande ampleur, ce que l'on souhaite c'est pouvoir réagir et faire des détections d'anomalie à la volée. C'est un sujet de recherche essentiel sur lequel nous travaillons : l'apprentissage sur des flux de données.

Avec l'analyse de grands graphes, de nouveaux outils émergent qui améliorent les capacités de traitements et d'analyse de données. On utilise des graphes comme modèles de données pour permettre ou faciliter certaines analyses. Il s'agit de graphes qui sont inférés à partir de transactions, d'échanges... Les enjeux sont nombreux, par exemple sur les graphes probabilistes qui permettent de manipuler des données associées à des valeurs d'incertitude, et pour lesquels on cherche à analyser, traiter, extraire de l'information. Se posent également des problèmes d'échelle, quand il s'agit de trouver ces informations dans de grands graphes sans tester tous les motifs possibles. Et dans tous les cas nous visons des qualités de résultat les meilleures possibles et l'impératif de pouvoir répondre dans des temps raisonnables.

Prenons l'exemple de graphes inférés à partir de la mobilité des personnes, dans le cadre d'une application de transport, dans une grande ville comme Paris. On peut, à partir de données Twitter

ou de réseaux sociaux, voir où les gens postent des informations. En récupèrant ces traces-là, on obtient des graphes. On ne les garde pas sous forme de données brutes, on transforme ces données pour coupler ces mobilités en trajets, avec des contraintes comme : le trajet doit être fait dans un certain laps de temps. Ceci donne des fragments de graphes, en très grand nombre, et on essaie d'en sortir des motifs (patterns) *de mobilité pour dire que le lundi entre 8h et 10h, la population parisienne bouge selon tel modèle, etc. Cela donne des informations intéressantes pour gérer l'offre de transport et la multimodalité par exemple.*

Un autre exemple est celui de la recommandation dans un cadre de réservation touristique. On part d'un graphe de similarités, entre les produits qu'on recommande ou entre les utilisateurs. Les graphes de points d'intérêts –tel monument ayant été visité, quelle prochaine visite serait intéressante à proposer– sont très complexes et très étendus, avec des pondérations, des probabilités... Dans une ville, il y a énormément de points d'intérêt. Ces graphes de similarité sont construits à partir des profils des historiques de visite des utilisateurs. Traiter et extraire de l'information à partir de ces graphes reste toujours un défi, et cela se traduit en termes de recherche. Une compétition s'est établie entre les différentes équipes des laboratoires de recherche, chacune essayant de proposer de nouvelles techniques qui améliorent le domaine.»

La plupart des problèmes sont encore ouverts...

Quelles perspectives pour la recherche en big data à moyen et long terme ?

Le big data est un ensemble de défis. Ceux-ci ont été à l'origine portés par des volumes de données importants, auxquels les premières entreprises à avoir eu accès étaient les entreprises à l'échelle du Web. Les volumes de données vont continuer à augmenter et les besoins d'analyse et de traitement vont perdurer quoiqu'il arrive.

On peut affirmer que le big data est un ensemble de technologies qui continue de se développer, comme par exemple dans le cadre du traitement de données à large échelle, où Spark est en train de supplanter Hadoop. La technologie va évoluer également autour des techniques statistiques, on voit l'émergence du deep learning de plus en plus présent dans les travaux, et de nouvelles techniques de machine learning qui se développent.

Ce qui va changer, c'est que petit à petit les technologies big data vont être au service de nouvelles applications comme on le voit avec l'Internet des objets. Il est maintenant acquis que de nouveaux problèmes big data ne cessent de se présenter et qu'il ne s'agit pas d'une mode. Le développement de ces outils ne va pas s'arrêter et le big data est de plus en plus au service d'autres challenges.

Questions de recherche

Repousser les limites du machine learning

Le machine learning est une classe d'algorithmes dont la capacité à pouvoir analyser des données sans faire d'hypothèses préalables, sans chercher de modèles, a fait grand bruit ces dernières années, parfois avec raison, parfois par abus.

Si le « *machine learning permet de proposer des services à haute valeur ajoutée avec un avantage concurrentiel* », comme le souligne Alain Abramatic (PSA Groupe), il n'est pas obligatoirement la panacée, rappelle Alexandre Gramfort (Télécom ParisTech), « *de nombreux problèmes de data peuvent être résolus avant de faire du machine learning, avec de la visualisation, en posant les bonnes questions.* » Yoann Janvier (IPSEN) confirme que ces outils sont employés de manière mesurée : « *Le machine learning supervisé n'est pas utilisé pour tous les projets, car il manque bien souvent des données étiquetées. Le machine learning non supervisé (par exemple le clustering) ne donne pas toujours des résultats probants. De plus, l'identification de cas d'usages avec du machine learning nécessite bien souvent un travail approfondi avec des interlocuteurs métiers qui soient très mûrs sur ces sujets, ce qui est encore rare.* » La recherche reste donc très active, notamment sur « *tout ce qui est statistical machine learning où on construit un système prédictif, où les exemples doivent prendre du sens* », précise Alexandre Gramfort. « *Il y a des choses qui restent compliquées à faire, tout ce qui relève de l'apprentissage non supervisé notamment.* »

Un autre enjeu, moins connu , est celui des données de test, quand elles sont faites de données simulées, « *notamment quand il s'agit de travailler sur des domaines où les données sont classifiées* », explique Arnaud Cauchy (Airbus Defense & Space, page 73) « *car elles ne sont pas assez crédibles et les systèmes d'intelligence artificielle qui les traitent ne trouvent rien à corréler. Les relations sémantiques sous-jacentes n'ont pas été bien traduites, et le machine learning ne sort pas de vrais problèmes. C'est là un vrai problème théorique.* »

Visualisation et Interaction homme-machine

James Eagan, maître de conférences à Télécom ParisTech, développe une recherche dans le domaine de l'Interaction homme-machine, à l'intersection de la visualisation d'information et la programmation par l'utilisateur final. « *Cela donne des outils qui peuvent énormément aider avec le machine learning. L'union des deux aide l'utilisateur à explorer, comprendre, découvrir, et les data scientists à faire des analyses, construire des modèles qui peuvent avoir un pouvoir explicatif, analytique, à l'échelle de ce qu'on ne pouvait pas traiter manuellement avant.* » Il identifie trois défis actuels et futurs de la data visualisation : développer la *visual analytics*, pour un meilleur mariage de la visualisation et des outils automatiques ; savoir inciter les utilisateurs à interagir avec les données ; démocratiser la visualisation en encourageant les personnes à manipuler leurs données personnelles et réapprendre à les interpréter graphiquement.

Data sciences et cognition

Le machine learning s'inspire en partie des recherches effectuées sur le cerveau et, dans un mouvement d'enrichissement circulaire, les avancées en data sciences et en visualisation permettent de mieux explorer les connaissances issues des sciences cognitives. Alexandre Gramfort (voir aussi pages 26-27 ses contributions sur l'outil Scikit-learn) effectue ainsi sa recherche dans le domaine de la neuro imagerie fonctionnelle, c'est-à-dire la compréhension du cerveau en fonctionnement, grâce notamment au logiciel open-source MNE, spécialisé en traitement du signal des électroencéphalogrammes et des magnétoencéphalogrammes. « *Ce sont des technologies qui créent beaucoup de données : 40 minutes de magnéto encéphalographie produisent 10 Go de données. Mes travaux de recherche aident les personnes qui collectent ces données tous les jours à avoir les meilleurs outils statistiques, les algorithmes pour les traiter, et mieux comprendre le cerveau.* »

Titulaire d'une bourse « *Starting grant* » par le *European Research Council* (ERC), visant à récompenser des travaux de jeunes chercheurs et les encourager à construire leur équipe, le chercheur recrute des data scientists, sur des expertises allant de l'extraction de données au développement de nouveaux outils de traitement. « *C'est très applicatif. Un des défis est de traiter des données actuellement non exploitées car trop sujettes aux signaux parasites, qui peuvent venir des capteurs comme du cerveau lui-même. Les médecins ont également besoin de faire de l'exportation non supervisée, pour faire de la visualisation, pour faire des systèmes prédictifs, extraire automatiquement des bio marqueurs…* »

Des processus de recherche transformés

Autre apport possible des data sciences dans la recherche, leur nature multidisciplinaire facilite l'ouverture des équipes de recherche vers de nouvelles disciplines. Maguelonne Chandesris est aujourd'hui responsable de la thématique « *Data, Mobilités et Territoires* » chez SNCF Innovation & Recherche. Son rôle est de faire émerger, co-construire et porter la vision partagée d'un programme pluri-annuel de recherche, organiser et effectuer une veille dans le domaine et mettre en place des partenariats. Motivée par l'idée que « *les données permettent d'imaginer les nouvelles manières de (ne plus) faire* », elle a travaillé sur une démarche d'« *algorithmes à dess(e)ins* », qui serviraient à la fois la représentation graphique des données et l'intention donnée à voir. C'est l'intégration d'une designer dans une équipe auparavant constituée de statisticiens qui en a été le déclencheur.

Les données étudiées étaient celles des requêtes d'itinéraires, venant de l'application et du site Transilien, des données moins habituelles par rapport aux données de transport effectué. L'équipe de Maguelonne Chandesris cherchait à définir quelle valeur ces données pouvaient avoir, et comment produire du nouveau et s'approprier de manière collective ces données complexes et abstraites. L'adoption d'une démarche d'innovation par le design a offert à l'équipe de data scientists effectuant ces travaux, dont les résultats étaient incertains et qui restaient exploratoires, plus de liberté pour aborder le sujet et choisir le rendu final. De nouvelles méthodes de travail ont été adoptées.

Une thèse en machine learning

Claire Vernade est doctorante en dernière année de thèse, et une des premières thésardes de la chaire Machine Learning for Big Data. **Qu'est-ce qui a incité cette diplômée de Télécom ParisTech à poursuivre ses études dans le domaine de la recherche en machine learning ?**

« Cela a été en fait un hasard total pour moi d'atterrir dans le domaine du machine learning. J'ai toujours été passionnée par l'idée que des décisions complexes puissent être éclairées par des algorithmes, particulièrement quand ces décisions vont impliquer beaucoup de contraintes qui demandent de prendre en compte beaucoup de variables. J'ai d'abord commencé par étudier la recherche opérationnelle à Télécom ParisTech. Ce sont des méthodes plus théoriques fondées sur des modèles mathématiques, des problèmes qui sont ensuite traités d'un point de vue plus algorithmique. J'ai pris un cours de machine learning en fin d'année et ça a été la révélation.

Le traitement statistique/probabilistique de ces mêmes problèmes de décisions –ou disons, des problèmes avec des objectifs proches– m'a tout de suite séduite. J'ai rapidement pris la décision de faire le Master MVA (Mathématiques / Vision / Apprentissage de l'ENS Cachan), une formation que j'ai adorée car j'y ai appris beaucoup de choses très vite. Le master commence fin septembre et les cours se terminent fin mars, ce qui laisse six mois pour traiter une grande variété de problèmes en machine learning. En décembre j'ai compris que ça ne serait pas suffisant, que cela ne me permettrait pas de rentrer dans le fond des choses. Comme j'avais particulièrement aimé le cours d'apprentissage par renforcement, j'ai commencé à me renseigner sur les possibilités de stages et de thèses dans ce domaine. Je n'ai pas eu à chercher longtemps car justement Stephan Clémençon était en train de boucler la Chaire Machine Learning for Big Data et m'a proposé d'être parmi les premiers doctorants. C'était une opportunité à prendre, les planètes s'alignaient soudain parfaitement, et je n'ai donc pas hésité longtemps.

Une voie inconnue et un monde incroyable

Au moment où j'ai pris cette décision, je n'avais absolument aucune idée où me mènerait ce choix de parcours. Je ne savais pas trop ce que c'était de faire une thèse, je n'avais pas vraiment envisagé cette option avant le MVA donc je m'engageais dans une voie un peu inconnue. Et trois ans plus tard, je pense que c'était en effet une excellente décision. J'ai découvert un monde incroyable, celui de la Recherche académique, et de son pendant industriel qui tend à se développer dans le domaine du machine learning. Ce monde est juste fantastique. Il est fait de gens qui passent leur vie à se poser des questions souvent très complexes et non-triviales et à chercher des solutions à ces problèmes tout en se remettant en question en permanence. Cela m'apporte

beaucoup personnellement et je suis maintenant bien plus sereine quant à mon avenir et à ma carrière. Je ne sais pas exactement ce que je ferai, mais je suis sûre que ce sera avec des gens intelligents et que je continuerai à apprendre.

Pourquoi le machine learning est-il un domaine d'avenir ?

Lorsque je suis arrivée en thèse, j'ai eu le privilège de pouvoir assister à la conférence NIPS grâce à la Chaire. Il y avait 2000 personnes à Montréal, près du double de l'année précédente. Deux ans plus tard, nous étions plus de 6000 à Barcelone. Le monde a pris le chemin de l'intelligence artificielle, les gens commencent à faire confiance aux machines et à l'idée que les machines puissent contrôler certaines choses dans nos vies, ou bien nous orienter. Un exemple frappant est l'engouement pour les voitures Tesla. La fonction auto-pilot de ces voitures est particulièrement recherchée par les clients, et presque tous les constructeurs maintenant essayent de créer un modèle équivalent. Cela veut dire que nous, les chercheurs dans ce domaine, allons devoir être à la hauteur du défi technologique que cela représente, d'être capable de guider une voiture transportant des personnes dans un milieu très complexe. L'objectif est encore assez loin d'être atteint, malgré les avancées rapides.

Beaucoup d'autres signes tendent à montrer que le machine learning va faire partie de nos vies de plus en plus. J'écris ces lignes dans la Sillicon Valley actuellement, et ici tout le monde se déplace en Uber –qui est contrôlé par de nombreux algorithmes d'optimisation–, tout le monde se fait livrer toutes sortes de choses et on parle de livraisons autonomes (sans conducteur ou via des drones) pour bientôt. Ici, personne ne serait choqué de recevoir son repas le soir, livré par une voiture sans chauffeur... Je ne sais pas ce que l'on doit en penser, mais il est certain que la technologie pour faire tout cela est loin d'être prête. Ce sont donc autant de défis technologiques et sociétaux pour l'avenir du machine learning.

Des défis pour la société

Enfin, dans un domaine bien plus crucial, le machine learning va probablement jouer un rôle important dans la gestion de nos ressources énergétiques. Récemment, Google a économisé des millions en électricité en laissant un algorithme contrôler sa consommation. Ce type d'applications va se généraliser et donc participer à diminuer drastiquement le gaspillage de l'énergie.

Quelle est aujourd'hui la place d'un chercheur en data science dans notre société ?

Un de nos rôles est de rendre possible ce à quoi la société dans laquelle nous vivons aspire. Je pense que les chercheurs sont là pour se poser des questions que personne ne se pose, pour soulever des problèmes et proposer des solutions. Tout le monde devrait se poser des questions mais pour accéder à certaines questions, on doit avoir un bagage de compétences techniques qui permettent d'appréhender la complexité des problèmes en jeu. Le rôle des chercheurs est plus de poser des questions pertinentes que de proposer des solutions optimales. Ma courte expérience m'a montré que c'est en réalité la partie la plus compliquée du travail de recherche. »

Les nouveaux paradigmes scientifiques

Le plus puissant télescope du monde, le SKA (*Square Kilometre Array*), actuellement en cours de construction, commencera à acquérir des données en 2020, avec les technologies big data en cours de développement. Les astronomes estiment qu'il collectera l'équivalent de 35 000 DVDs par seconde, soit l'ensemble du web actuel chaque jour ! Cette arrivée massive de données, dans tous les domaines, change profondément le visage de la science.

En 2007, Jim Gray, un chercheur en base de données réputé, propose le nom de *eScience* pour désigner une nouvelle méthode scientifique dans laquelle « *IT meets scientists* ». Cette proposition entérine le fait qu'il existe des scientifiques qui ne regardent presque plus directement dans leurs instruments (télescopes, microscopes, accélérateurs de particules…) de nouveaux éléments. Ils examinent les données capturées par ces instruments et qui n'ont pas encore été étudiées, et celles créées par les simulations, toutes de plus en plus massives. On parle d'*exploration de données*, et il s'agit du *quatrième paradigme scientifique*.

Ce quatrième paradigme se construit sur la base des trois précédents, et les complète. La Science s'est développée en effet tout d'abord à partir d'une méthodologie empirique fondée sur l'observation et l'étude des phénomènes observables (premier paradigme). Les pratiques scientifiques deviennent par la suite de plus en plus théoriques, utilisant des modèles et faisant appel aux abstractions et à la généralisation. C'est ainsi qu'apparaissent les lois de la gravitation, celles de l'électromagnétisme, traduites en formules mathématiques. L'arrivée des ordinateurs au XXe siècle marque une troisième évolution, celle où la programmation devient l'outil de travail et d'expression des chercheurs, qui utilisent les machines pour modéliser les phénomènes complexes.

Exploration de données spatiales

Parmi les figures pionnières de la future data science, deux étudiants ont dans les année 60 ouvert la voie de l'exploration spatiale grâce à la manipulation de données. On l'a oublié, mais il y a seulement 50 ans l'envoi de sondes vers d'autres planètes était jugé impossible, en raison d'une simple question d'énergie nécessaire pour s'affranchir à la fois de l'attraction terrestre et de celle du soleil. Un des fondements mathématiques à résoudre était le problème des 3 corps, c'est-à-dire celui des équations du mouvement de Newton (deuxième paradigme) de corps interagissant gravitationnellement, connaissant leurs masses ainsi que leurs positions et vitesses initiales. Le problème à trois corps –celui d'un engin spatial qui part d'un corps céleste pour un atteindre un autre– possède une solution analytique exacte, découverte en 1909,

qui se présente sous la forme d'une série infinie convergeant très lentement, hélas inutile en pratique pour faire des prédictions en un temps raisonnable.

En 1961, un étudiant en mathématiques, stagiaire au *Jet Propulsion Lab*, Michael Minovitch, utilise le temps de calcul de l'ordinateur le plus rapide de l'époque, l'IBM 7090, pour résoudre le problème (troisième paradigme). Il approche si bien de la solution qu'on l'autorise à utiliser des données plus précises sur les positions des planètes. Son modèle s'en trouve conforté ; Michael Minovitch a résolu le problème des 3 corps.

L'histoire ne s'arrête pas là. En 1965, un autre stagiaire d'été, Gary Flandro, s'intéresse aux données de son prédécesseur, dans l'idée d'explorer les planètes extérieures (quatrième paradigme). Il reporte les données sur des graphes (visualisation de données) sans savoir ce qui l'attend. Sur un de ces graphes, les lignes représentant les positions des planètes externes se recoupent, signifiant qu'une fenêtre de tir existe pour pouvoir les explorer toutes d'un coup. Et la fenêtre de tir est 1977, permettant de visiter 4 planètes externes en 12 ans. Or la prochaine fenêtre de ce type est… 176 ans plus tard. le programme Voyager venait de naître.

Vers un cinquième paradigme

Cette science « *data-intensive* » consiste en trois principales activités : l'acquisition de donnée (capture), la « *curation* », puis l'analyse. Les données ici traitées proviennent à la fois des instruments et des simulations. Elles sont mises à disposition en open data pour toujours à des fins d'analyse continuelle, car on ne sait jamais grâce à qui, et dans quelles circonstances, ces données vont être à l'origine de nouvelles découvertes ou inventions. C'est la sérendipité qui a animé Gary Flandro, et la transdisciplinarité qui a sans doute été la chance de Michael Minovitch, lui qui a examiné un vieux problème scientifique d'un œil nouveau.

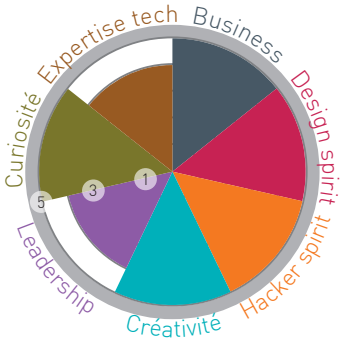
L'arrivée des systèmes d'intelligence artificielle (IA) dans les équipes scientifiques est en train de dessiner un nouveau paradigme. En 2015, une équipe australienne a utilisé une IA pour refaire l'expérience du condensat de Bose-Einstein, un état particulier de la matière qui a été prédit en 1925 par Albert Einstein (paradigme 2), et réalisée en 1995, valant le prix Nobel de physique en 2001 à l'équipe. L'IA de 2015 devait s'occuper notamment du paramétrage des lasers impliqués dans l'expérience. Moins d'une heure a été nécessaire pour refaire l'expérience à partir des conditions de départ, au grand étonnement de l'équipe scientifique, d'autant plus surprise que l'IA avait fait des choix techniques auxquels aucun humain n'avait pensé avant, et qui pourraient ouvrir de nouvelles pistes d'investigation.

Les IA utilisées dans ce 5e paradigme et les chercheurs en data science peuvent donc faire progresser la science en général d'au moins trois manières : en accélérant la recherche de conditions expérimentales et de protocoles scientifiques optimaux ; en proposant des méthodes auxquelles les humains ne pensent pas ; en lançant les scientifiques sur de nouvelles pistes d'exploration, grâce à ces nouvelles méthodes et aux idées trouvées en cherchant à les comprendre, stimulant ainsi leur créativité.

Consultant Data & Analytics

Les consulants data aident les organisations à définir et mettre en place leur stratégie data. Ils ont comme interlocuteurs les opérationnels métiers et font le lien avec l'ensemble des personnes agissant sur les données, des ingénieurs big data aux chief data officers. Ils participent à la création de la culture de la donnée dans les entreprises.

Les data scientists, les data analysts, les ingénieurs big data et les architectes big data peuvent devenir des consultants, dès lors qu'ils possèdent les qualités de bonne communication et bon relationnel, les capacités de synthèse et de vulgarisation, et l'appétence pour la diversité des métiers de leurs clients.



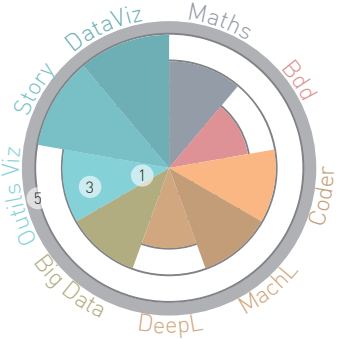
le lien entre une compréhension métier pointue et la manipulation technique des données et des algorithmes

Profil Formation école d'ingénieur ou de commerce

Compétences **Qualités :** créativité, curiosité

Savoir raconter ce que les données ont à dire

Connaissance étendue des outils du marché, compréhension des enjeux métier



Des données et des humains

linkedin.com/in/denisoblin/fr
@OblinDenis

Denis Oblin
Consultant Data
associé Memorandum

Leur vision très large des écosystèmes de la donnée permet aux consultants data d'apporter leur assistance et savoir-faire, aussi bien aux équipes de data scientists établies qu'aux entreprises n'ayant pas ces compétences en interne. Le quotidien de Denis Oblin est ainsi très varié : « *Animations de comités de direction, ateliers d'émergence de besoin, découverte métier, codage… les formes d'intervention auprès de nos clients (grands comptes, PME, start-up…) traduisent la diversité des situations qu'ils rencontrent, et ce qu'ils peuvent attendre des données.* » Tour à tour data scientist puis expert métier, il vise à apporter une réponse opérationnelle. Il alterne pour cela la discussion avec la data : « *coder, tous les jours, tester de nouvelles approches techniques* », et avec les humains que la donnée concerne : « *tout est dans la compréhension du métier* ». L'essentiel, et la différence, ne se joue en effet pas tant dans les packages algorithmiques utilisés, que dans la compréhension métier qui a été insufflée dans la préparation des données en amont et lors du dépouillement en aval.

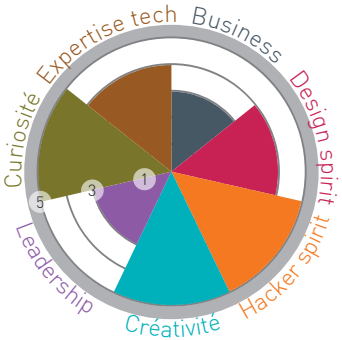
« La data n'est qu'une matière première, pas une fin en soi »

Denis Oblin identifie les problèmes sur lesquels la donnée peut, ou ne peut pas, aider, puis manipule ces données. « *En 3-4 cycles de production agiles de 15 jours à 3 semaines, le client progresse dans la formulation de son objectif au même rythme que je progresse dans la réponse.* » Certains cherchent de la prédiction, d'autres veulent un diagnostic opérationnel. En d'autres termes : veut on prédire l'avenir ou le changer ?

Être consultant data permet également de diffuser la culture de la donnée hors des domaines les plus courants. Denis Oblin cite cette entreprise de location de matériel professionnel disposant de plusieurs sites de distribution, recevant 15 000 devis par an, concrétisés aux deux tiers. « *En travaillant ces données, nous les avons considérablement enrichies jusqu'à une centaine de caractéristiques par devis. Nous avons construit un score, mis en production, qui annonce avec 90% de performance si le devis va être gagné ou perdu.* » Cette entreprise sans site web profite à plein de la transition numérique. « *Il n'y a pas de petite donnée* », conclut cet artisan de la donnée, qui continue à inventer chaque jour son métier avec son associé et son équipe.

Data journalist

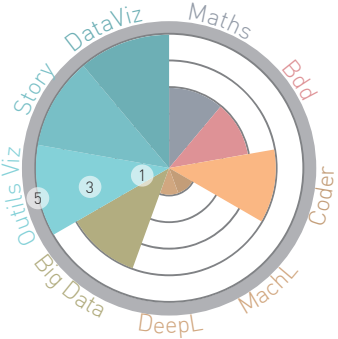
Les data journalists sont des journalistes qui collectent, nettoient, vérifient, croisent, classent et analysent des données massives pour y trouver des informations nouvelles ou pour les présenter à leur lectorat de manière compréhensible, accessible et interactive. Ils ne rédigent pas nécessairement d'articles, mais proposent une visualisation de l'information avec les mêmes outils graphiques que les experts en data visualisation. Ils travaillent le plus souvent avec les développeurs web et graphistes de leur média, se concentrant sur la mise en lumière des faits et la mise en scène des histoires qu'ils doivent conter.



Un journalisme d'enquête sur le terrain des données

Profil École de journalisme

Compétences Curiosité, rigueur, grande culture générale



Un journalisme qui redonne du sens

« *Qui est le prince mouillé de Bel Air ?* » La question est posée fin 2015 en Californie, après la découverte dans les données publiques du département de l'eau et de l'énergie de Los Angeles qu'une personne résidant dans le quartier chic de Bel Air avait consommé près de 45 millions de litres d'eau en un an, dans une région souffrant de fortes sécheresses, et en pleine restriction de consommation d'eau. Le fournisseur d'eau ne souhaitant pas donner le nom de ce grand consommateur, des citoyens engagés dans la chasse aux gaspilleurs rendent publique l'affaire, et un journaliste commence son enquête dans le LA Times.

Confronté à l'impossibilité de faire le tour de ce quartier huppé, et au refus de divulgation des noms, il va se tourner vers les données accessibles publiquement et les faire parler. En utilisant des images satellites publiques et des techniques d'imagerie spécialisées en reconnaissance des végétaux – car tous ne consomment pas l'eau de la même manière, comme il l'apprendra dans une base de données scientifiques –, et en croisant le tout avec les cartographies publiques – bâtiments, adresses… – de manière à relier les parcelles à leur propriétaire, le journaliste va réussir à cerner sept noms potentiels. Il publiera fin 2016 sa méthode, ses sources et ses astuces.

Renouveler la manière de faire du journalisme

Si l'utilisation de données statistiques pour appuyer une enquête a régulièrement été utilisée par la presse, le « journalisme de données », poussé par le double mouvement de l'abondance de données et d'information devenues inextricables, et de la défiance grandissante dans les médias traditionnels, a subi un développement important il y a une dizaine d'année, au point que la plupart des rédactions aujourd'hui disposent de leurs data journalistes. Ceux-ci ont pour mission de plonger dans de larges bases de données pour y collecter, nettoyer, vérifier, trier, analyser et croiser les données, puis visualiser et scénariser ce que ces données ont à raconter. Certains se spécialisent dans le *fact-checking*, qui consiste à contrôler la véracité des propos ou des chiffres avancés par tel ou tel, tandis que d'autres explorent ces masses de données comme un terrain d'enquête d'un nouveau genre. Tous sont animés par le souci de la donnée proprement utilisée pour expliquer la complexité du monde.

Leurs compétences sont celles des data scientists classiques, et leurs outils et méthodes, très largement publics, ont fait l'objet de formations ouvertes en ligne qui sont un premier pas accessible et ludique pour de futurs explorateurs de la donnée. Leur expertise du *storytelling* et de la mise en visualisation des informations en font également une source d'inspiration précieuse pour des data scientists ayant des données compliquées à mettre en musique. Les data journalistes, comme certains artistes qui créent des œuvres à partir de données, sont des data scientists spécialisés. Peut-être les premiers d'une liste de nouveaux métiers à inventer.

Naviguer en données complexes

Si expliquer la complexité du monde et redonner du sens aux faits est le quotidien des data journalistes, découvrir des corrélations utiles dans de grands corpus de données oblige bien souvent les data scientists à jouer un rôle d'enquêteurs. En attirant leur attention sur telle ou telle donnée et en les étiquetant de meta-informations, la combinaison de systèmes d'intelligence artificielle et de réalité virtuelle et augmentée leur facilite la tâche. Ces systèmes allègent la charge cognitive nécessaire pour évoluer dans des ensembles de données à la structure difficile à appréhender, ou pour naviguer simultanément dans des flux d'informations de différentes natures.

C'est particulièrement vrai en situation de stress. Pour Emmanuel Bavière chez Société Générale (voir page 37), le cas des traders ou celui des pilotes de chasse est à ce titre symptomatique. La navigation dans leur environnement de données se fait la plupart du temps à vue d'aigle, et arrivent des moments de vérité où il est nécessaire de leur apporter la bonne information, correctement packagée, en provenance directe des données massives. Il faut alors disposer d'interfaces beaucoup plus fluides et légères que le classique clavier/ souris. Ces nouvelles interfaces utilisent par exemple le *eye tracking* pour repérer l'intérêt soudain de l'utilisateur sur une partie du champ d'informations et en améliorer subtilement sa compréhension, ou encore les agents conversationnels pour ne pas mobiliser les yeux. Les data scientists doivent concevoir ces interfaces modernes qui augmentent les sens et aident le cerveau à naviguer dans la donnée complexe.

Système autonome de commande, de contrôle, de communication et de gestion de mission, Fightacs est un système d'information flexible conçu par Airbus pour les pilotes d'hélicoptères et de chasseurs. Connecté aux radios embarquées et aux systèmes de navigation de l'aéronef hôte, il permet à ses utilisateurs l'accès à des données complexes comme la carte de navigation 3D, les menaces électroniques et la situation tactique. Il a été conçu pour transformer à faible coût n'importe quel avion ou hélicopère d'ancienne génération en une plate-forme adaptée à l'exécution des missions modernes.

Faire voler les data

Un groupe de 136 000 salariés organisé en trois divisions : Commercial Aircraft, Helicopters, Defence and Space.

Il y a 23 000 capteurs dans un A320

Opérant dans des domaines sensibles pour une grande variété de clients, eux-mêmes s'adressant à une multitude de clients, Airbus est une entreprise de taille mondiale confrontée à des disruptions majeures sur son volume d'affaires principal. Du terabyte de données enregistrées par vol d'A350 à l'indexation des images satellites en temps réel, de la complexité d'un avion lors de sa conception, puis de son exploitation, à l'évolution des besoins et usages de ses passagers, la donnée est au cœur de la transition numérique d'Airbus. Il n'est dès lors pas étonnant de retrouver dans les équipes de data scientists des personnes également impliquées dans la transformation digitale de l'entreprise.

« Les données au cœur de la transformation digitale de l'entreprise »

Arnaud Cauchy est l'un d'entre eux. *Digital transformation officer* depuis 2015, travaillant sur des projets à l'échelle du groupe, il devient responsable en 2017 d'un département d'*engineering for digital services*, dans lequel lui et sa petite équipe agile de data scientists de niveau international imaginent des services nouveaux, associés à de nouveaux modèles d'affaire, qui puissent se réaliser effectivement. Comme partout en effet, vendre des produits seuls ne suffit plus : il faut leur adjoindre des services. Bardé de capteurs, le produit devient *smart product* auquel sont attachés des *smart services*.

Deux cas sont envisagés. Le premier consiste, à partir d'un produit existant, à y collecter des données opérationnelles et à prolonger son cycle de vie. C'est l'exemple de la maintenance prédictive, qui évite les pannes soudaines non anticipées. Appliquée aux satellites, les données télémétriques analysées permettent ainsi de détecter des signes de faiblesse, les pallier, et assurer la continuité de mission. Le deuxième cas est celui où le modèle écomique ne fonctionne plus et où le produit est en train de disparaître. Il est alors remplacé par du service dématérialisé, par une offre naissante. Avec l'accord des clients, des données anonymisées sont collectées. Riches de situations très différentes et croisées avec des données externes, elles permettent aux data scientists d'imaginer de nouveaux services, dont certains correspondront parfois à de nouveaux clients.

@DrMattEvans

Matt Evans est *Digital Transformation Leader* au sein du groupe Airbus. Il y dirige le développement et la mise en œuvre d'une stratégie de données complète pour l'ensemble du groupe, comprenant l'intégration de données et les plateformes, les technologies et compétences analytiques, la gestion des données et leur gouvernance, ainsi que l'interaction avec les clients et les fournisseurs pour leur assurer l'accès aux données.

La majorité de sa journée consiste à rencontrer les personnes du terrain (processus, fabrication, assistance & services…) pour comprendre leurs métiers, leurs objectifs, et pouvoir identifier les liens entre les données et ces objectifs. « *Il faut avoir la capacité de coder, de créer des visualisations, pour guider le client final via un produit minimum viable* », précise-t-il, l'industrialisation du projet arrivant dans un deuxième temps. Avec son équipe de 25 data scientists, ce *Chief Data Officer* a ainsi déployé sur les lignes d'assemblage des A350 (49 avions livrés en 2016) un projet dont il est particulièrement fier. Sur un tel avion où tous les problèmes de conception n'ont jamais été vus auparavant, la recherche rapide de solutions, dans un contexte de haute qualité, est un souci constant. Partant de l'idée que « *certes tel problème est nouveau, mais des choses comparables ont peut-être eu lieu un jour ailleurs* », l'équipe de data scientists a créé une interface utilisateur facile d'emploi, permettant de faire des associations d'idées entre problèmes – certains étant parfois décrits sur 40 pages, à croiser avec des données structurées des bases des fabricants –, de fournir des recommandations et de proposer des avis. Cette application qui a littéralement élargi la vision des superviseurs (5 ou 6 personnes sous leur responsabilité) a été plebiscitée, et est utilisée par plus de 1 000 personnes.

Donnée géospatiale enrichie

À travers Data Management Solutions, Airbus Defence and Space propose une gamme de produits et de services qui permet à ses clients d'accéder facilement, de gérer et de diffuser différents types de données géospatiales.

En croisant des images satellitaires de Spot ou Pléiades avec des données externes comme les prévisions météorologiques, l'analyse biochimique de sols et des données sur les pesticides, il est possible de conseiller les agriculteurs qui souhaitent utiliser ces derniers avec parcimonie. Les assureurs en cas de tempête peuvent recevoir plus que des images avant/après, et disposer de cartes des dommages et d'une estimation des coûts. Dans les pays ne faisant pas de recensement, des modèles urbains et des modèles culturels permettent de prédire le nombre de personnes par toit.

Tous ces acteurs n'ont pas besoin d'être des spécialistes du pixels. Ils veulent des services d'aide à la décision, et la production de données enrichies les leur apporte. Les data scientists férus de cartographie et de croisements de données originaux trouvent chez Airbus Defence and Space de quoi nourrir leurs passions.

Données externes mobilisées

Même avec près de 10 000 aéronefs civils Airbus volant dans le monde, le volume de données qui en provient ne suffit pas pour créer le dossier de toute la vie d'un avion. Ces données doivent être croisées avec les données opérationnelles et de maintenance, et celles issues des usages des compagnies aériennes. Cette meilleure connaissance de l'utilisation des avions facilite la création des modèles de maintenance prédictive et est partagée avec les compagnies. Concernant l'amélioration de l'expérience des passagers, l'anonymisation des données personnelles collectées est faite avec une grande vigilance.

Savoir maîtriser le cloud

Anne Chanié, responsable offres futurs segments sol d'observation – nouvelles technologies chez Airbus Defence and Space, participe à l'élaboration de solutions de SI Big Data faisant venir les utilisateurs à la donnée, en leur fournissant une plate-forme où sont proposés des services à forte valeur ajoutée, ainsi que les outils nécessaires à leur développement (API ouvertes, services de cloud computing). Il s'agit de véritables systèmes d'exploitation de la donnée – enrichie en amont, indexée en temps réel à la sortie des gros flux satellitaires – fournissant de l'information élaborée. Dans ce cadre, les infrastructures *cloud* telles celles de Google ou Amazon sont vite devenues incontournables. Les data scientists doivent savoir maîtriser les outils très spécifiques du cloud, ainsi que les outils du machine learning tels TensorFlow ou Caffe. La puissance de calcul de ces infrastructures facilite notamment des techniques de « *croissance d'algorithmes* », pour lesquels divers paramétrages peuvent être testés dans des temps raisonnables, et des solutions converger en quelques heures. Il s'agit également de stocker la donnée de manière à en optimiser l'accès, et de concevoir des solutions techniques et algorithmiques qui soient capables de passer à l'échelle.

Les problèmes mathématiques posés par la recherche de corrélations dans les données provenant des avions restent nombreux. Airbus collabore avec des universités et des start-up qui proposent de nouvelles manières d'analyser les données, élaborent des algorithmes efficaces et créent des outils de visualisation adaptés aux besoins des data scientists.

Au sein de l'IRT Saint Exupéry, à travers l'accélérateur Airbus BizLab, lors de hackathons, les occasions sont multiples de croiser des talents externes avec ceux d'Airbus, sur des problèmes toujours passionnants et complexes qui nécessitent la somme de toutes les énergies et des compétences des divers métiers de la datascience. Et travailler dans le paysage des données d'Airbus ne signifie pas n'être que mathématicien. Les data scientists ici doivent avoir l'envie de croiser des technologies avec des sources ouvertes qui peuvent être assez éloignées, et se demander, avec un état d'esprit geek, ce que cette combinaison de données internes et externes jamais faite encore pourrait bien produire. Jeunes recrutés, plus anciens en reconversion interne, start-up et chercheurs, toutes et tous reflètent bien la multitude de parcours et d'aptitudes possibles pour être data scientist.

Un secteur en tension

« *Recommandations algorithmiques, prévisions commerciales, trading algorithmique, détection d'anomalie, traitement du langage naturel… appliquez des modèles d'apprentissage machine sur des projets du monde réel !* », le recrutement de data scientists prend de plus en plus des airs de mobilisation générale, tant la demande augmente. C'est ce que constate Talel Abdessalem : « *Le big data s'installe comme une thématique importante pour les entreprises et pour les laboratoires de recherche. Il y a aussi un besoin d'expertise, le marché des data scientists ne se tarit pas, on l'observe par le nombre d'inscriptions à nos formations. La demande des personnes à se former sur le big data a explosé et continue d'augmenter, et la demande de recrutement sur le big data des entreprises augmente dans le même temps. Et les changements fréquents de postes, le turn over, est un indicateur du manque de data scientists sur le marché.* »

Une discipline en mouvement

Quels conseils Stephan Clémençon donnerait-il à quelqu'un souhaitant s'engager dans un cursus de data science ? « *Tout dépend s'il a envie de faire de la recherche ou de traiter rapidement des applications. Pour la recherche, je suggère d'avoir un socle de connaissances fondamentales le plus large possible plutôt que de se spécialiser trop vite. Le domaine bouge énormément et a besoin d'idées nouvelles. Il faut des connaissances très générales pour pouvoir ensuite traiter des problèmes spécifiques, mais de façon entièrement nouvelle. Je travaille par exemple sur des problématiques d'analyse de données de préférence – celles que nous exprimons à travers une poignée de films ou sur quelques objets d'un catalogue considérable proposant des millions de produits – qui s'exprime avec des types d'objets mathématiques tout à fait nouveaux, alors que ce problème est ancien. Dans les travaux que j'ai développés avec mes étudiants et mes collègues, nous avons besoin de topologie algébrique et de domaines qui n'étaient pas traditionnellement évoqués dans le machine learning. À quelqu'un voulant faire de la recherche, je conseillerai également d'apprendre l'informatique pour être capable de se confronter aux contraintes des applications modernes.* »

« *Il faut absolument un background mathématique* », renchérit la directrice data science Angélique Bidault-Verliac chez Voyages-sncf.com. Ons Jelassi invite ainsi les data scientists postulants à « *vérifier leur appétence pour les aspects purement mathématiques, et leur intérêt pour l'informatique distribuée, car il faut être à l'aise sur les deux.* » Fabrice Otañao, Chief Data Officer du groupe Accor, demande à ses data scientists « *de faire des modèles industrialisables, qui passent à l'échelle, et d'être opérationnel.* ». Et pour quelqu'un voulant être plus opérationnel, Stephan Clémençon suggérera « *de ne pas non plus négliger les enseignements théoriques qui lui permettront de continuer à progresser. La discipline est en mouvement. Je lui conseillerai d'être l'incarnation de cette interdisciplinarité entre informatique, mathématiques appliquées et usage.* »

Prouver ses compétences

Les recrutements de data scientists sont réguliers au groupe Accor. « *Il y a deux chemins* », précise Fabrice Otañao, « *la technicité, et j'ai trois lead data scientists qui challengeront les postulants ; et les bonnes personnalités capables d'interagir avec les métiers.* » Chez Toucan Toco, le *test & learn* fait partie des valeurs et des critères de recrutement chez les développeurs. « *Nous testons l'ensemble de notre code et avons des environnements sandbox dans lesquels nous essayons nos nouvelles offres* », explique Charles Miglietti. « *Il est important de pouvoir prouver ses compétences techniques lors d'exercices.* » Consultant data, Denis Oblin conseille de « *continuer à suivre des MOOC et fréquenter des meet-ups.* » Il faut également fréquenter des plate-formes comme Kaggle, où des entreprises proposent des problèmes et récompensent les data scientists ayant obtenu les meilleurs performances.

Aimer être sur le terrain

Denis Oblin a la possibilité de voir de nombreux cas d'usages dans des secteurs d'activité très différents. « *Il est très important de savoir raconter des choses qui parleront aux gens du métier. On trouve parfois des résultats non attendus ou non demandés par le client. C'est parce qu'on a partagé avec tous les opérationnels, été à côté d'eux, multiplié les interactions, affiché le maximum de data visualisation pour permettre des commentaires spontanés, et été sur place, là où la donnée est produite.* » Ce constat, à savoir qu'il est essentiel d'installer le data scientist en immersion chez ses clients, internes ou externes, se vérifie dans les entreprises, qui sont de plus en plus nombreuses à intégrer leurs data scientists directement auprès des opérationnels.

Anne Chanié, Airbus Defence and Space, prolonge cette nécessaire proximité du terrain par celle de la donnée traitée : « *Dans mon domaine, la connaissance et la compréhension de la donnée sont essentielles. On ne met pas des données dans un pot en attendant que l'information sorte toute seule. L'expertise de la donnée initiale est une compétence à avoir si l'on veut en être un bon interprète.* »

Montrer sa motivation

Comment convaincre un recruteur de ses compétences et de sa motivation ? Kim Pellegrin, Dassault Systèmes : « *Le meilleur moyen est de présenter des réalisations probantes, les mettre à disposition sur un dépôt Git par exemple. Réviser avant l'entretien le code et les réalisations qu'on souhaite présenter de façon à ne pas redécouvrir le sujet face au recruteur et avoir une présentation fluide. Faire un travail continu de veille technologique pour se tenir au courant des avancées.* »

« *Dans son CV, il faut faire preuve d'humilité et de curiosité, ne pas empiler les références aux technologies, ne pas survendre et être capable de mettre en œuvre pendant l'entretien les compétences annoncées* », poursuit Yoann Janvier, IPSEN. « *Je fais systématiquement passer des tests techniques pour m'assurer que les compétences sont au rendez-vous et il y a pas mal de surprises !* »

Faire son CV de data scientist

Comme dans tout CV, il est essentiel de faire prendre connaissance d'un parcours académique et professionnel en un seul coup d'œil, et de laisser une bonne impression parfois en quelques secondes. Toutes les sections et informations doivent être immédiatement identifiables. Maïté Allain, consultante en recrutement et responsable d'Upward Data : « *Nous préférons les CV simples et classiques, en une page, avec pour chaque expérience – datée – mise en avant, une description rapide de l'entreprise (titre de la fonction exercée, nom de l'entité à l'intérieur de l'entreprise, domaine d'activité), les problématiques data qui ont été rencontrées, les outils et les méthodes qui ont été utilisées, et les résultats obtenus. Des données chiffrées et des termes techniques, tant qu'ils restent compréhensibles par le recruteur, peuvent être appréciés.*

Dans la catégorie des compétences, mettre, de manière classique, les connaissances techniques en rapport avec le domaine d'activité, les connaissances informatiques, et les niveaux dans les langues pratiquées, justifiés éventuellement par des certifications ou des séjours à l'étranger. La catégorie formation regroupe les cursus suivis et les diplômes obtenus pendant et après la formation initiale, notamment les certificats obtenus à la suite d'un MOOC, ou sur une technologie particulière, et les formations certifiantes.

La science des données étant fortement évolutive, les data scientists doivent continuellement se former, et c'est valable également pour les plus jeunes. Les projets développés en marge de sa formation et de son emploi seront très appréciés, ainsi que la participation à des hackathons, à des compétitions de type Kaggle. »

Quatre types de CV de data scientists passent devant les yeux de la recruteuse, chacun avec leurs spécificités. Les CV des jeunes data scientists, ceux qui ont une forte connaissance métier et souhaitent se réorienter, ceux qui ont utilisé ou développé des algorithmes et souhaitent également s'orienter vers les métiers de la données, et les architectes big data qui sont souvent à part. Ces derniers fournissent en effet des CV plus longs, avec une page de synthèse, puis plusieurs pages pouvant développer leurs expériences.

« *Les CV doivent rester classiques et complets, car nous n'avons pas le temps d'aller voir les blogs des candidats, par exemple. En revanche, ces derniers peuvent être mentionnés, c'est toujours un plus. Et pour les experts en data visualisation, le CV peut servir de support pour en démontrer la maîtrise.* » Dernier point, le CV sur Linkedin, sur lequel on peut détailler ses expériences. « *En plus des expériences, leur durée, la maîtrise des outils, il permet de mettre en avant les compétences recommandées par les autres. C'est également l'occasion de rencontrer d'autres data scientists, d'échanger dans des forums, et de faire sa veille.* »

Se réorienter vers la donnée

Pour Matt Evans, groupe Airbus, la reconversion est un vrai sujet. «*De nombreuses personnes sont désireuses de se reconvertir dans les métiers de la donnée. Or la formation essentielle reste tout de même les mathématiques et les statistiques. Et faire ce choix de reconversion induit un vrai changement de métier. Ces personnes, même celles en début de leur carrière vers 35 ans, doivent bien peser tous les tenants et les aboutissants.*» C'est là que réside tout l'intérêt de suivre un Certificat d'études spécialisées – et de préparer celui-ci en participant à un MOOC – qui permettra à chacun de faire le bilan de ses connaissances et de s'assurer qu'il mesure bien le chemin à parcourir pour devenir data scientist.

Ons Jelassi observe que les profils des personnes voulant suivre le CES «Data Scientist» de Télécom ParisTech sont de plus en plus solides : «*Parmi les dossiers que je reçois, les étudiants ont déjà fait des formations par e-learning avant de venir nous voir, ils se forment à la programmation Python par eux-même, et viennent tout de même candidater. Ils ont l'habitude d'utiliser les outils d'apprentissage de type MOOC... Le label CES est cependant très utile pour légitimer les connaissances. La certification est aujourd'hui encore très importante auprès des étudiants et des entreprises. Quand je regarde un dossier de candidature avec les MOOC machine learning ou massive data mining de Stanford, cela me parle et cela veut dire que la personne s'est intéressée au sujet et l'a creusé. Mais par rapport à tout ce qui est pratique et professionnalisant, la formation en présentiel est un plus dans une procédure d'embauche.*»

En donnant sa définition du data scientist, Pierre Gotelaere d'Enedis veut faire vibrer celles et ceux qui s'apprêtent à rejoindre le monde des données en cours de carrière, et qui disposent d'une maturité plus grande : «*Les compétences mathématiques et statistiques restent bien sûr primordiales. Cependant, le data scientist doit avoir selon moi quatre cordes supplémentaires à son arc. D'abord, il lui faut avoir une capacité à vulgariser et communiquer pour emporter l'adhésion autour de ses travaux ; puis une sensibilité business (recherche de valeur), car ces profils sont coûteux pour l'entreprise ; un vernis système d'information pour faciliter le lien avec les équipes IT en charge de l'industrialisation des travaux d'études ; et enfin, savoir travailler en mode projet pour avancer avec différents interlocuteurs dans le cadre d'un projet industriel.*»

Les fiches pratiques et les fiches métiers de ce livre blanc ont été rédigées en partenariat avec Upward Data, cabinet de recrutement spécialisé dans les métiers de la data science et du big data. Résolument pro-candidats, Updward Data accompagne les postulants sur le long terme, sur un marché en constante évolution, les guidant vers des entreprises vers lesquelles ils n'avaient pas nécessairement pensé se tourner, et aidant ces dernières à dénicher des profils encore rares.

Grand groupe ou start-up ?

« *Les innovations vont se multiplier dans les années qui viennent, notamment par le biais des start-up* », observe déjà Yoann Janvier dans le domaine de la health tech. « *Bon nombre d'entreprises vont s'appuyer sur ces start-up pour accélérer leurs innovations et cela va contribuer à faire davantage évoluer le métier de data scientist.* » Alors, où débuter sa carrière, dans un grand groupe qui offrira des perspectives d'évolution ou dans une start-up réputée plus agile ?

Pour Maïté Allain, Upward Data, « *les avantages du grand groupe seront le nom sur son CV, l'apprentissage du monde de l'entreprise – organisation, services, métiers – et de ses codes, une très bonne sécurité de l'emploi couplée à des avantages salariaux et sociaux généralement plus intéressants que dans les start-up.* » Attention à bien identifier au préalable le niveau de maturité vis-à-vis des données, certains services en interne pouvant être encore réticents à ouvrir leurs données. Certaines entreprises de taille respectable ont cependant gardé, ou acquis, un esprit start-up. « *Ce sont des pépites que tout le monde cherche car elles allient sécurité de l'emploi et agilité.* » Les start-up réelles, celles qui cherchent encore leur modèle économique, apportent des défis d'un autre ordre. « *Rejoindre ce type de start-up à effectifs réduits, c'est participer à une vraie aventure entrepreneuriale avec tous les risques que cela comprend. Il ne faut pas avoir d'aversion au risque. Les data scientists sont très vite responsabilisés et doivent être très autonomes et accepter un rôle très souvent ultra-polyvalent. En arrivant au tout début de l'aventure, on intervient vite sur les décisions stratégiques de l'entreprise.* » Et il existe des secteurs, comme par exemple les RH, où beaucoup reste à inventer avec la donnée, que ce soit dans les start-up ou dans les entreprises établies.

Des échanges réguliers

« *Comme les applications et les projets de data science sont de plus en plus portés par des start-up, les entreprises maintiennent une veille très active sur ce qui est peut être développé à l'extérieur* », remarque Yoann Janvier. Les échanges, et les passerelles, sont continus entre les deux types d'organisation, pour le plus grand intérêt des data scientists qui, par nature, aiment les parcours hybrides.

La chaire Big Data & Market Insights organisait en octobre 2016 sa journée annuelle, consacrée cette année à la place des start-up dans la stratégie big data des entreprises. Talel Abdessalem y a noté que « *beaucoup d'entreprises étaient plus intéressées par les personnes qui travaillent dans une start-up que par leur modèle économique. La start-up agit comme un catalyseur de personnes hyper dynamiques et efficaces, qui veulent vraiment produire des choses. Ce sont des éléments essentiels pour les grandes entreprises, qui recherchent des compétences. Les start-up sont aussi des laboratoires démontrant qu'une telle technologie est possible. Passer sous une grande entreprise, c'est l'industrialisation.* » Une trajectoire de plus à prendre en compte.

Se former en continu

« *La datascience est un art nouveau, amené à changer. Ce qu'il faut préparer, c'est une formation personnelle continue aux nouvelles technologies à venir* », invite Charles Miglietti, fondateur de Toucan Toco. « *Il y a la possibilité de se former par soi-même car la communauté est ouverte et fonctionne beaucoup par data challenges* », renchérit Stephan Clémençon. Il existe aussi une particularité que la data science partage avec le monde de la recherche. « *On s'attend à ce que la recherche soit reproductible, et c'est pareil dans le monde des données : si un algorithme fonctionne bien, il faut en faire la preuve, et les codes de ces algorithmes sont souvent disponibles en open source, comme un certain nombre de jeu de données de référence. Tout cela est livré à la communauté. C'est ce qui lui permet d'en faire usage et d'appréhender ces outils. Du matériel est donc disponible et la communauté attend que les usagers fassent un retour de ces outils et les documentent.* »

Une expérience qui se cultive

Ce qu'on demande à un data scientist, à partir de données, d'un problème parfois mal formulé, éventuellement avec l'aide d'une personne du métier, c'est « *cuisiner les données jusqu'à ce qu'elles se confessent et qu'elles arrivent à faire ce qu'on leur avait demandé* », sourit Alexandre Gramfort. « *Pour y arriver, il faut pas mal d'expérience, il faut s'être battu pendant des jours sur du traitement de données, comprendre quelles sont les formes des algorithmes, en comprendre la complexité et les subtilités, quand est-ce qu'ils fonctionnent ou pas…* »

« *Aujourd'hui, quand on me donne un problème de machine learning* », poursuit le chercheur, « *pour en avoir vu un certain nombre, je sais déjà la première chose à tester avant d'autres et cet avis est en grande partie fondé sur l'expérience. Il faut certes avoir été à l'école pour comprendre les algorithmes qui existent, leur fonctionnement intime, mais il y a beaucoup de choses qui viennent aussi par la pratique.* »

C'est en ayant traité cent problèmes de machine learning que, finalement, on trouve les dénominateurs communs, les algorithmes qui ont le plus de chance de marcher, et qu'on peut écarter ceux qui seraient une très bonne idée, mais malheureusement inadaptés car le volume de données est trop important.

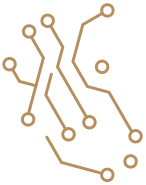
Relever sans cesse des défis

Les data scientists ont la chance et l'obligation de se remettre sans cesse en question et de se défier. Cette caractéristique de leur métier est également ce qui leur permet de se faire repérer, et de faire évoluer leur carrière. Pour Alexandre Gramfort, Kaggle fournit une excellente formation. « *Le but de la plateforme Kaggle est de réunir les bonnes personnes qui apprennent vite, de façon à pouvoir les identifier et leur proposer d'évoluer, de faire d'autres mission… Les data scientists sont nombreux à aller sur Kaggle pour se former, parce qu'il y a des forums, parce qu'on y apprend en faisant, parce que finalement, la science des données est une science bien plus expérimentale que ce qu'on peut imaginer.* »

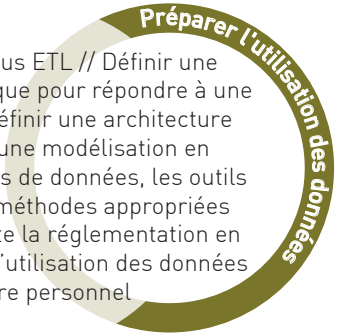
Retrouvez toutes nos formations courtes en page 15

Le quotidien des data scientists

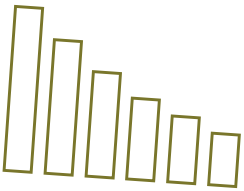
Mise en place / maintenance de plates-formes de données // Mise en œuvre de modèles / algorithmes en production // Planification de grands projets logiciels ou de systèmes de données // Mettre en place les outils technologiques et la stratégie adaptée pour sécuriser les données de l'entreprise



Développer / utiliser du hardware // Développer / utiliser des produits qui dépendent de l'analyse de données en temps réel // Développer des tableaux de bord // Collaborer sur des projets de code // Développement de prototypes de modèles // Développer / utiliser des logiciels d'analyse de données



Faire des processus ETL // Définir une modélisation statistique pour répondre à une problématique // Définir une architecture de traitement et une modélisation en déterminant les types de données, les outils statistiques et les méthodes appropriées // Prendre en compte la réglementation en vigueur concernant l'utilisation des données à caractère personnel



Identifier les besoins en données de l'entreprise

Identifier les besoins et la problématique des directions métiers // Comprendre et mobiliser les connaissances métiers spécifiques // Identifier des cas d'usage à résoudre avec l'analyse de données // Proposer des axes de gestion et d'analyse de grandes masses de données // Définir et mettre en œuvre un projet transverse dans l'entreprise // Prendre en compte au quotidien le modèle économique et la stratégie

Répondre à des questions / exploiter les données

Utilisation de tableaux de bord pour prendre des décisions // Extraction de caractéristiques // Analyse des données pour répondre à des questions // Analyse de base des données exploratoires // Utiliser des outils d'analyse et de gestion de bases de données de types variés dans de grands volumes, en faisant preuve de réactivité et d'adaptation afin de surmonter les obstacles rencontrés

Communiquer & partager

Communiquer avec des personnes extérieures à l'entreprise // Création de visualisations // Communiquer les résultats aux décideurs // Organiser et synthétiser les résultats d'une analyse sous une ou des formes adaptées au besoin (rapport, graphique…) et exploitable

Travail d'équipe

Enseigner / former d'autres personnes // Organiser et guider des projets d'équipe

Source :

« Tâches quotidiennes des data scientists », selon le *2016 Datascience salary Survey*, par O'Reilly (983 répondants, 45 pays + 45 États des USA)

Les compétences des data scientists

Le métier de data scientist recouvre des réalités très différentes, et le terme lui-même de data scientist (voir fiche métier page 32) peut décrire, selon les entreprises, des profils plutôt orientés mathématique ou plutôt orientés informatique. Les métiers présentés dans ce livret sont tous en évolution, nécessitent un ensemble de qualités et de compétences qui leur sont transverses, et se complètent pour constituer des équipes de data scientists. Pour chaque métier, nous proposons un profil type, articulé autour des compétences orientés data et des compétences générales ci-dessous. Ces profils peuvent servir de boussole (radars ci-contre à titre d'exemples) pour choisir son premier métier, ou évoluer d'un métier à un autre.

Crédits

Conception, rédaction, mise en page et suivi de réalisation

Aymeric Poulain Maubant, Nereÿs www.nereys.fr

Suivi de projet et contenus additionnels

Stéphane Menegaldo, Télécom ParisTech

Contenus complémentaires sur les fiches métiers et les fiches pratiques

Maïté Allain, Upward Data

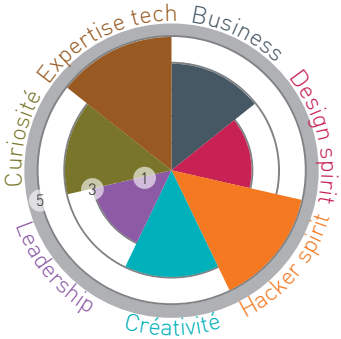
Crédits photos : Xavier Granet, Fonds Télécom ParisTech, droits réservés

Crédits pictos : ArtFeelsGood, Noun Project (Barracuda, Gan Khoon Lay, Hea Poh Lin, Gabriele Malaspina, Samy Menai, Tinashe Mugayi, Anusha Narvekar, Sergey Shmidt)

Merci à l'ensemble des professionnels qui ont apporté leurs témoignages ainsi qu'aux relecteurs de cet ouvrage.

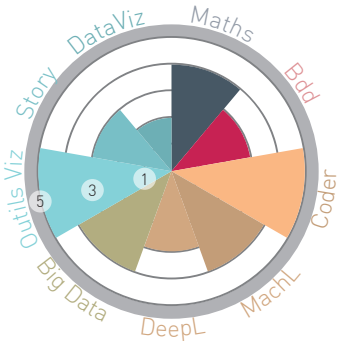
Compétences générales

Passion / Curiosité intellectuelle / Créativité – Des évolutions technologiques constantes, qui demandent une veille permanente // Capacités d'apprentissage // Rigueur et précision // Capacités de communication – Être capable de communiquer clairement auprès des opérationnels qui ne sont pas des profils techniques // Culture du partage pour échanger, se tenir informé, monter en compétence et faire monter en compétences // Autonomie // Sensibilité aux enjeux business // Ouverture d'esprit – Comprendre les problématiques métiers et opérationnelles // Esprit d'équipe // Leadership



Compétences sur la donnée en particulier

En science des données : Modélisation statistique ; Machine learning : arbres de décision, régression logistique, traitement automatique du langage naturel, deep learning… // *En informatique :* Langages statistiques : R, Python ; Langages de programmation compilée : C#, C++, Java… ; Bases de données SQL et NoSQL ; Frameworks : écosystème big data Hadoop, Spark… // *Communication :* Outils de data visualisation : Tableau, Qlikview… Storytelling : capacité à raconter les données, Dataviz : capacité à les mettre sous forme graphique



Data Scientists ! Mais qu'est-ce donc ? et pourquoi tant d'enthousiasme ? Si le livre que vous tenez entre les mains aurait aussi pu s'appeler « Toutes et tous data scientists ! », c'est parce que les « data », les informations de toutes natures qui circulent notamment sur Internet, ont pris une place tellement centrale dans notre quotidien que chacun et chacune d'entre nous est concerné. Et le point d'exclamation vient souligner l'incroyable effervescence du domaine, l'engouement passionné des professionnels qui se lancent dans l'exploration de ce nouveau monde.

Collectées par les objets qui nous entourent, générées en flux continu par les ordinateurs des multinationales, ou directement mises en ligne par des millions d'internautes, les données sont au cœur de presque tous les processus de notre vie moderne : achat en ligne, détection de cyber attaques, cartes de fidélité, prévision d'épidémies, suggestions de voyages, gestion de l'énergie, reconnaissance de visages et demain : voiture autonome, médecine personnalisée, intelligence artificielle… !

Ce livre est une véritable plongée dans les métiers de la donnée, depuis ses fondements historiques jusqu'à ses perspectives d'un avenir qu'on touche déjà du doigt. Il s'adresse aux étudiants et futurs étudiants en science des données, aux spécialistes du domaine, aux entreprises qui souhaitent les recruter, aux professionnels de la formation et de l'orientation, aux curieux, en bref à tous ceux qui veulent découvrir et mieux comprend cet univers foisonnant qui nous offre des possibilités infinies.

Retrouvez cet ouvrage en version numérique sur : www.telecom-paristech.fr/datascientists

www.telecom-paristech.fr - www.telecom-evolution.fr

Institut Mines-Télécom

université PARIS-SACLAY

INSTITUT CARNOT Télécom & Société numérique