DATA SCIENTIST : DU REVE A LA REALITE

Découvrez ce que c'est réellement la Data Science et les actions à suivre pour devenir Data scientist

Mr. Mint

Machine Learning made easy

YOUNES BENZAKI

Table des matières

Pourquoi ce livre ?

Introduction

La *Data Science* (ou la science de données), en voilà un mot fantaisiste. Pourtant, c'est un terme qui fait le buzz ces dernières années. Tellement de buzz que la *Harvard Business Review* l'a qualifié du job le plus sexy du 21ème siècle.

L'une des conséquences de cette hypermédiatisation de la *Data Science*, est l'engouement sans précédent pour le poste de *Data Scientist.* On le décrit comme le magicien qui fait parler les données muettes de l'entreprise. Ou encore celui qui trouve des pépites d'informations là où personne n'aurait songé à en trouver.

L'engouement pour ce poste a été principalement perçu aux Etats-Unis. Notamment, après le rôle très probant qu'a joué la *Data Science* pour mettre le 44ème président des Etats-Unis, *Barack Obama*, au bureau ovale.

Persuadés de l'importance de cette science, les européens se sont, eux aussi, faits entrainés dans cette sphère. Ainsi l'engouement est devenu quasi-mondial.

Qui dit engouement, dit chaos. Tout le monde hurle qu'il connaît l'ultime vérité sur la *Data Science*, comment la pratiquer et comment la transmettre. Résultat des courses, le prochain venu au domaine se retrouve perdu entre ce que disent les experts, ce qui est médiatisé dans la presse écrite et audiovisuelle, et ce que racontent certains charlatans !

Face à cette (triste) réalité, le nouvel apprenti, motivé, se retrouve perdu face à ce chaos et ce bouilli culturel/scientifique. Par conséquent, soit il laisse tomber ce domaine porteur, soit il s'arme de patience et de courage. Pour les courageux, ils se baseront sur des ressources en ligne et parfois sur des *MOOC* assez généralistes. Ou bien, ils se pencheront sur des bouquins théoriques trop poussés dédiés aux chercheurs scientifiques.

Finalement, les apprentis se retrouvent frustrés. D'une part de la complexité de cette discipline. D'autre part du manque de recul et de retour d'expériences concrets.

Du fait de la jeunesse de la *Data Science*, le chemin d'apprentissage de cette dernière n'est pas encore bien clair et défini. De mon humble point de vue, j'estime que comprendre le périmètre d'un domaine que l'on cherche à étudier est la première et la plus importante étape pour un apprentissage efficace et agréable. Surtout quand il s'agit d'une discipline si novatrice et dynamique telle que la *Data Science*.

Pour cela, il faut prendre de la hauteur, et répondre à des questions existentielles comme : que cherche à résoudre la *Data Science* ? pourquoi résoudre ces problématiques ? comment les résoudre...etc.

Lors de ce livret, il ne s'agit de rentrer dans les détails techniques inhérents à la *Data Science*. Mais plutôt de comprendre "*l'esprit"* de ce domaine. Ainsi, vous aurez assez de recul pour comprendre ce qu'on cherche à faire et par conséquent, mieux prioriser ce qu'il faut apprendre en premier.

Le but de cet ebook est de vous aider à y voir plus clair. Démystifier le vrai du pur charlatanisme. Par conséquent, vous verrez ce que c'est la *Data Science*, ce que c'est un *Data Scientist*, et ce qu'il faut pour en devenir un.

Les formules mathématiques et des *snippet* de code informatique n'auront pas leur place dans ce livret. Il s'agit surtout de prendre de la hauteur et de voir comment apprendre à apprendre, savoir nommer chaque chose par son nom et finalement avoir une méthodologie et une vision claire et efficace de l'apprentissage de ce domaine.

A l'issue de la lecture de ce livret, vous aurez une idée claire de ce que c'est la *Data Science* et, ce qu'est un *Data Scientist*. Vous saurez également comment s'y prendre pour devenir un professionnel des sciences de données. Ainsi, grâce à cette clairvoyance du chemin d'apprentissage, vous ne vous perdrez pas en chemin lors de votre progression.


LA DATA
SCIENCE
EST
L'ELDORADO
DE DEMAIN !

La Data Science : La nouvelle ruée vers l'or

La Data Science : pourquoi un tel buzz ?

Data Scientist est un métier récent aux contours assez flous. De très nombreux débats ont lieu sur l'éventail des compétences qui le caractérisent. Ce flou rend difficile de statuer sur le rôle de ce métier nouveau au sein des entreprises.

Bien que certains considèrent un *Data Scientist* comme "le mouton à cinq pattes", la science de données permet de répondre à des vraies problématiques. En effet, la *Data Science* est un domaine permettant de tirer des **informations utiles** depuis des gisements de données.

Au vu du potentiel que permet la *Data Science* aux entreprises, ce domaine connaît un réel engouement. En effet, cette science permet de meilleures décisions stratégiques pour une entreprise. Ceci est possible parce qu'elle permet de trouver des informations utiles pour les décideurs.

Une information utile différera, selon le contexte, d'une entreprise à une autre. En effet, une entreprise opérant dans le e-commerce n'aura pas les mêmes problématiques à résoudre qu'une institution bancaire.

Les décideurs d'une entreprise e-commerce seront, éventuellement, plus intéressés de savoir si les produits qu'ils mettent en avant dans leurs boutiques en ligne impactent le prix moyen du panier d'un client.

Par ailleurs, dans le cas d'une banque, elle sera notamment plus intéressée de détecter des transactions et flux financiers frauduleux. Ceci en se basant sur le comportement du client et ses habitudes financières.

En remarquant le potentiel et les problématiques que peut résoudre la *Data Science*, les entreprises (startups, multinationales...) à l'échelle mondiale se sont mises à étudier ce domaine. Notamment, en montant des laboratoires *R&D* pour construire des *Data Product.* Ces dernières s'adaptent à l'utilisateur en lui proposant des contenus et des produits adéquats à ses goûts. Cette adaptation et contextualisation de l'offre permet un marketing de proximité résultant à des taux de conversion plus élevés.

Suite aux retours encourageants des géants du web et de différentes entreprises. Les entreprises investissent de plus en plus dans la *Data Science*. Cela se concrétise par des recrutements de *Data Scientist* bien qu'ils restent assez rares encore.

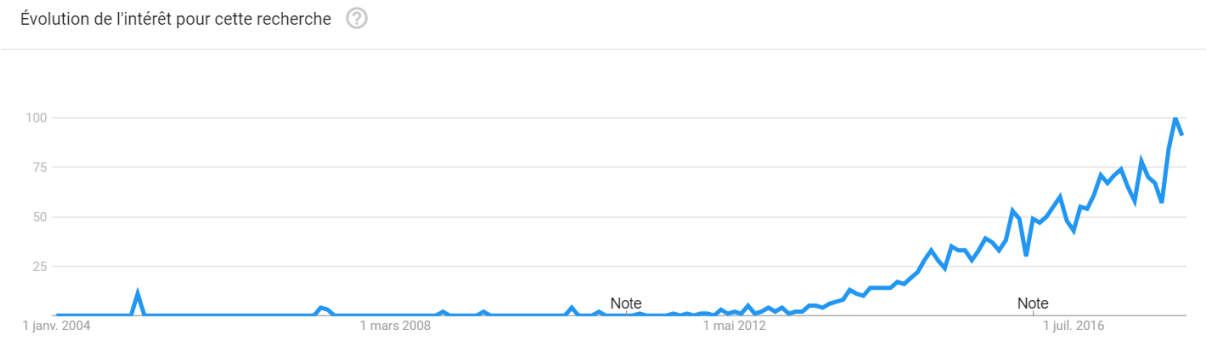

Figure 1 : Evolution de l'intérêt et l'usage du mot Data Scientist depuis 2004

La figure ci-dessus illustre l'intérêt mondial pour ce domaine et ce métier. Tout le monde croit en l'existence d'un réel besoin et s'accorde à dire que lors des dix prochaines années, le profil de *Data Scientist* sera très recherché. Un peu à l'image des traders financiers de *Wall Street* lors des années 1980 et 1990.

Quel intérêt à apprendre la Data Science

Les *Data Product* qui sont entre nos mains de nos jours puisent leurs innovations de la *Data Science* et du *Machine Learning*. Toutefois, ces applications restent assez peu nombreuses encore. Pour ainsi dire, nous vivons les prémisses d'une révolution des technologies de l'intelligence artificielle et de la *Data Science*.

Par ailleurs, la quête de l'humain à produire des assistances numériques, automatisées et intelligentes, poussera ce dernier à innover en intelligence artificielle et en *Data Science*.

En effet, certains métiers à basse qualification disparaîtront pour laisser place à des programmes intelligents (pensez à *Uber* qui essaie de se passer de ses chauffeurs au profit des voitures autonomes). Même les professions "pointues" comme la médecine, les métiers de droit (avocats…) ne seront épargnés. En ce qui concerne la médecine, certains professionnels de la santé expérimentent des programmes intelligents pour l'assistance au diagnostic et aux procédés chirurgicaux. Ceci ne relève pas de la science-fiction car il est déjà arrivé que des algorithmes de *Machine Learning* ont su déceler des pathologies cardiaques qui ont été manquées par des cardiologues qualifiés.

Bien que l'engouement sur la *Data Science* subsiste, des profils qualifiés et maîtrisant les tenants et les aboutissants de ce domaine se font encore rares. Certainement à cause de la complexité pluridisciplinaire de la *Data Science* et la jeunesse de cette dernière. En effet, le nombre moyen d'années d'expériences professionnelles des *Data Scientist* à travers le monde est d'environ 4 ans. Ce sont surtout des profils qui ont fait un changement de cap dans leurs carrières et ont bondi le pas pour se reconvertir en des *Data Scientist*. Or, les gens qui ont osé relever ce pari, ont certainement fait le bon choix. Du fait de la rareté du profil, ils se forgent un nom, une expérience nouvelle, tout en jouissant d'un beau package de rémunération (certainement) !

Votre intérêt à apprendre la *Data Science* est de rattraper cette vague pour ne pas dire ce tsunami technologique. Un tsunami de robotisation et d'applications intelligentes. Le virage stratégique vers l'intelligence artificielle sera pris tôt ou tard par toutes les entreprises à travers le globe. A la lumière de ce que présage ce paysage technologique, capitaliser sur la *Data Science* vous permettra de vous positionner sur un métier d'avenir. Un métier qui sera au centre des stratégies des entreprises.

Apprendre la *Data Science* et le *Machine Learning* c'est aussi, être acteur de cette innovation tout en capitalisant sur un métier d'avenir. En réalité, la *Data Science* est beaucoup plus qu'un simple buzz médiatique. C'est un domaine qui est là, dans notre vie et qui perdurera. On la voit maintenant comme un luxe. Demain, elle deviendra une nécessité de la société moderne.

Pour ma part, je perçois la *Data Science* et le *Machine Learning* comme un train de l'innovation dont les sonnettes alarmant de son départ retentissent. J'ai pris le pari de monter dans ce train et je vous invite via ce livret de le prendre avec moi.

Le prendriez-vous ... ?

Applications concrètes de l'utilisation de la Data Science

Maintenant qu'on comprend macroscopiquement le potentiel de la *Data Science*, on assimile mieux pourquoi le buzz est si prononcé et mondial. En effet, une réelle opportunité se présente. Nous vivons sans doute les prémisses d'une nouvelle révolution, technologique et mondiale cette fois-ci : celle de l'intelligence artificielle et de la *Data Science*.

Pour se convaincre de la réalité de l'existence et le bien fondé des applications de la *Data Science* dans le monde, voici quelques exemples de produits, issues de la science de données, que nous utilisons quotidiennement sans nous en rendre compte. Ces exemples ne donnent qu'un mince aperçu de ce que permet la *Data Science*. Il ne s'agit certainement pas d'une liste exhaustive.

Reconnaissance de photos

Si vous êtes un détenteur d'un smartphone *IPhone*, vous connaissez peut-être l'application *People*. Cette dernière arrive à faire des catégorisations des images par lieu et personnes se trouvant dans une photo.

Figure 2 : l'application People de l'IPhone permettant un regroupement de photos par personnes présentes dans la photo

Le smartphone va apprendre les informations contenues dans les photos pour les regrouper par similarité. Il apprendra davantage si vous lui dites le nom des personnes dans les photos. De sorte que, quand vous cherchez dans votre galerie de photos par personne, il saura les retrouver.

L'intérêt d'une telle application est de donner une sémantique aux photos. Ainsi au lieu d'avoir une galerie de photos statistiques, l'application *Photos* "*comprendra*" les informations contenues dans les images. Par ailleurs, l'utilisateur pourra faire des recherches intelligentes pour retrouver les photos où une personne y est présente.

Publicité ciblée

Le roi de la publicité ciblée est certainement *Google*. D'ailleurs, la publicité ciblée compose une grande partie des revenus du géant de *Mountain View*.

En se basant sur les requêtes de recherche, les produits Google que vous utilisez (Android, Google Agenda, Gmail, Maps, Picasa…), le géant du web a un dossier de profilage complet sur chaque utilisateur (même s'ils ne l'avouent pas). Les données que nous laissons (souvent à notre insu) suite à notre activité en ligne, sont récoltées et analysées. Grâce à la *Data Science* et aux techniques de *Machine Learning*, *Google* saura quelle publicité aura une meilleur réception et acceptation par l'internaute.

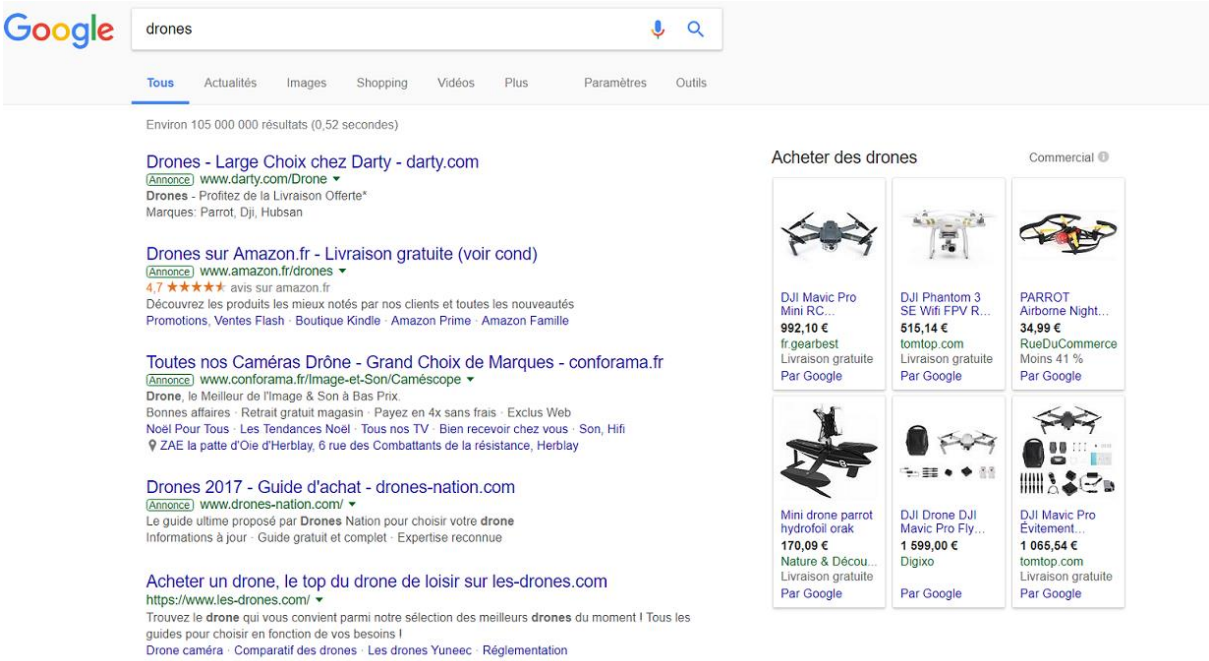

Figure 3 : publicité ciblée proposée par google suite à la recherche du mot clé "Drones"

Proposant ce ciblage de publicité précis, les annonceurs de pub paient au géant du web des sommes astronomiques pour faire de publicité de leurs produits. A titre d'information, certains mots clés de recherche (*search keywords)* coûtent jusqu'à 20€ par clic de l'internaute !

Dans la même lignée, Facebook agit de la même sorte sur son réseau social mondialement connu (environ 2 milliards d'abonnées). Imaginez à quel point la publicité ciblée est lucrative !

Recommandation de produits

Amazon utilise la *Data Science* pour proposer des articles pertinents pour ses clients. Ceci en se basant sur leurs historiques de navigation, d'achat sur le site ainsi que les données des autres clients.

Figure 4 : Amazon proposant d'acheter deux produits se vendant fréquemment ensemble (système de recommandation)

L'utilisation de la *Data Science* pour produire son système de recommandation de produits a permis à Amazon une nette augmentation de ses ventes et par conséquent une augmentation de son chiffre d'affaire.

Conduite Autonome

La conduite autonome fait rêver plus d'une personne. Ces dernières années, de multiples prototypes fonctionnels ont vu le jour. Certains constructeurs, notamment *Tesla Motors* commercialise d'ores et déjà des voitures électriques avec une conduite autonome.

Figure 5 : Une voiture du constructeur Tesla Motors en mode conduite autonome

Grâce à des techniques de *Deep Learning* (un sous domaine du *Machine Learning*), *Tesla* a su mettre en œuvre des voitures fiables. L'adoption à échelle mondiale n'est qu'une question de temps. Par ailleurs, grâce à la fiabilité qu'offre les voitures autonomes, des horizons meilleurs sont à espérer pour la sécurité routière.


DATA SIENCE

La Data Science : Loin du Mythe

Il est difficile de donner une définition exacte au terme *Data Science*. Cette difficulté est certainement due à la jeunesse de cette science et son chevauchement avec plusieurs disciplines.

En raison de sa nature interdisciplinaire, la *Data Science* se positionne à l'intersection des besoins métiers, le traitement de différents types de données et le développement de modèles algorithmiques. Le but étant de trouver des informations utiles en se basant sur des données et des algorithmes prédictifs pour un contexte métier particulier.

Grâce à la *Data Science*, on peut générer de la valeur ajoutée de deux façons différentes :

- Trouver des patterns (*Insights*) dans les données
- Le développement de *Data Product*

Trouver des patterns dans les données

Cet aspect de la *Data Science* se concentre sur l'étude des amas de données qu'une entreprise a collecté auparavant dans le cadre de son activité. Le but étant de comprendre ces données, et les étudier en détail pour en tirer des informations utiles : des tendances, des relations cachées entre les données...etc.

Le but étant d'aider les décideurs à prendre des décisions stratégiques plus aguerries en se basant sur données factuelles. Utilisée à bon escient, la *Data Science* permettra aux entreprises de prendre des décisions plus novatrices grâce aux relations "cachées" (*insights)* qu'aura découvert le *Data Scientist*.

Certaines entreprises ont su innover grâce à la *Data Science*, en voici quelques exemples :

- Amazon se base sur les temps de réponse de ses sites pour savoir si la vitesse d'affichage des pages de ces dernières impacte le prix du panier moyen de l'internaute
- Netflix utilise les données de visionnage des films de ses abonnées pour comprendre l'intérêt et les goûts de chaque abonné. Par la suite, Netflix utilise ces patterns pour décider quels genres de séries elle produira dans ses studios.

Pour le cas d'Amazon, penser à une relation de causalité entre le temps de réponse du site e-commerce et le prix du panier moyen ne semble pas intuitif et facilement décelable. La *Data*

Science, a permis à Amazon de détecter ce genre de relations “cachées”. Ainsi, les décideurs du site peuvent agir en conséquence à la lumière de ces informations.

Développement de Data Product

Un *Data Product* est un logiciel qui se base sur des données comme entrée, et génère un résultat. Le résultat généré est calculé en se basant sur un modèle prédictif que le *Data Scientist* aura construit auparavant.

Ces programmes sont le "cœur" permettant de personnaliser une autre application aux goûts de l'utilisateur. Par exemple, le site Amazon, se base sur un programme de recommandations de produits (le *Data Product*) pour suggérer des produits à l'utilisateur du site. Ces suggestions différeront en fonction des utilisateurs.

Voici d'autres exemples de *Data Product* :

- Le filtre anti-spam de service de messagerie *Gmail*, lit vos e-mails pour les catégoriser en Spam ou non.
- *YouTube* propose des vidéos similaires à ce que vous avez visualiser auparavant et le type de contenus qui vous intéressent.
- Le programme de traduction en direct d'une communication *Skype* entre deux interlocuteurs parlant deux langues différentes est aussi un *Data Product*

Les *Data Product* ont une particularité intéressante. Ils **s'améliorent avec le temps**. En effet, plus vous vous en servez, mieux ils vous "comprendront". Le filtre *Anti-Spam* de *Gmail* en est un bon exemple. Si vous indiquez à votre messagerie Google qu'un mail est un spam alors qu'il l'a laissé passer, la messagerie catégorisera en Spam les prochains mails le ressemblant. Cette amélioration à l'utilisation s'appelle *l'Online Training*.

Et le Machine Learning dans tout ça ?

Souvent, le terme "*Machine Learning*" est évoqué quand on parle de *Data Science*. Cela montre à quel point ils sont liés.

Le *Machine Learning* est un ensemble de méthodes algorithmiques permettant, entre autres, de produire des modèles prédictifs, ou détecter des patterns dans les données.

Algorithmes de prédictions

Dans ce type de problèmes, un algorithme va apprendre les caractéristiques (*features)* des données d'entraînement (*Training Set)* pour produire un modèle prédictif. On retrouve deux classes de prédictions : la **régression**, et la **classification**. Un algorithme de régression pourra, par exemple, apprendre à prédire le prix d'une maison en fonction de ses caractéristiques. Par ailleurs, un algorithme de classification nous détectera si une transaction bancaire est frauduleuse ou non.

La régression tout comme la classification fait partie de l'*apprentissage supervisé (Supervised*

Learning). La régression permettra de prédire des valeurs continues (prix d'une maison, poids d'une personne...etc.). Quant à la classification, elle permet de déduire une valeur discrète (une classe) comme Spam/non Spam, tumeur maligne/bénigne etc...

Découverte de Patterns

Dans ce type de problématique, on retrouve les algorithmes *d'apprentissage non supervisés (Unsupervised Learning).* Ces algorithmes servent à déceler des patterns dans les données. Parmi ces patterns : **rassembler les éléments par similarité**. Ainsi, les algorithmes de *clustering* vont rassembler des items similaires. Notamment, regrouper des photos par contenu, regrouper des morceaux de musique par genre musical...etc.

Les algorithmes de clustering sont utiles dans plusieurs situations. Notamment, la segmentation de la clientèle. Ainsi, le *clustering* détectera, par exemple, des clients "similaires". Cette similarité est déduite du jeu de données qu'on donne à l'algorithme. Cela peut être un regroupement par âge, par sexe, par habitudes de consommation etc…

Un autre exemple d'application concrète du clustering est le service *Google News.* Si vous utilisez ce service, vous remarquerez qu'il agrège des articles de presse par thématique (sport, technologie, politique etc…). Pour y parvenir, le service de Google se basera, entre autres, sur le contenu des articles de presse pour en déduire la catégorie.

Il existe d'autres algorithmes d'apprentissage non supervisés comme Analyse à Composantes Principales (ACP), les *modèles cachés de Markov*, ainsi que des méthodes qui sont à mi-chemin entre l'apprentissage supervisé et l'apprentissage non supervisé. Notamment le *filtrage collaboratif (Collaborative Filtering)* et la *détection d'anomalie* (*Anomaly detection*). Pour garder cet ebook concis, Nous ne les développerons pas davantage ici.

Le *Machine Learning* est prépondérant en *Data Science*. Apprendre ce domaine **est primordial** pour tout *Data Scientist* qui se respecte. Ce dernier aura à sa charge de choisir les bons algorithmes en fonction de son contexte et des différentes contraintes.

DEVENIR DATA SCIENTIST

Le Data Scientist : le nouvel alchimiste !

Un *Data Scientist* est quelqu'un qui travaille dans la *Data Science*. Il a pour mission d'aider les organisations à tirer profit de leurs données. C'est un profil qui portera plusieurs casquettes afin de mener à bien sa fonction.

Le *Data Scientist* se doit de maitriser les différents aspects de l'élaboration de modèles prédictifs et de création de *Data Product*. Pour y parvenir, il se positionnera à l'intersection des trois domaines suivants :

- Expertise métier
- Expertise mathématique
- Expertise informatique

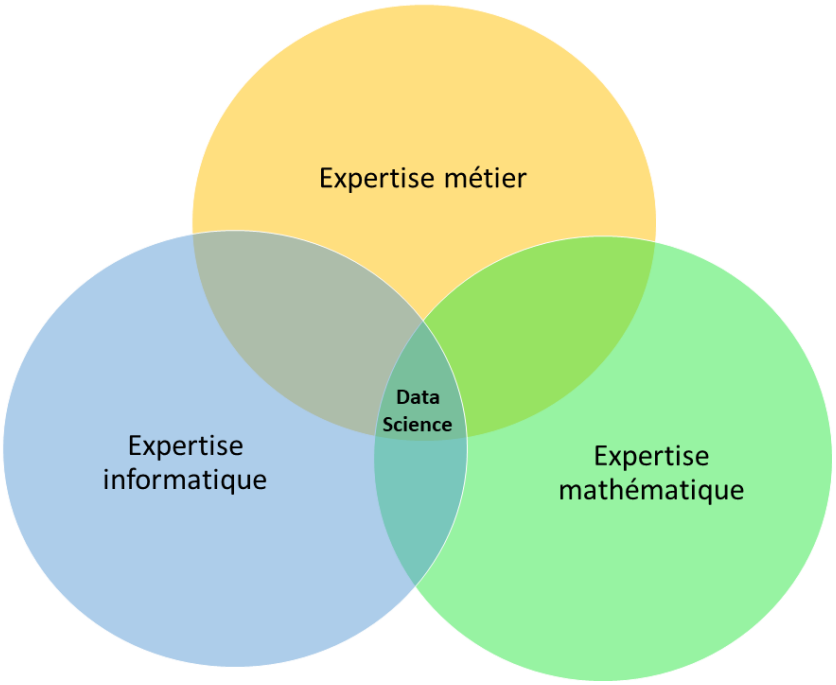

Figure 6 : Intersection des disciplines formant un besoin de Data Science

Expertise métier

Pour produire des *Data Product* ou des modèles prédictifs, un *Data Scientist* doit **comprendre l'enjeu métier** de l'entreprise. Ceci lui est primordial pour comprendre dans quelle direction cette dernière cherche à s'orienter.

Comprendre la finalité métier permettra au *Data Scientist* de se poser les bonnes questions et définir les pistes les plus prometteuses à explorer pour mener à bien sa mission. Par ailleurs, comprendre la finalité métier n'est pas suffisant en soi. En effet un *Data Scientist* **travaille conjointement avec les équipes métiers** pour comprendre leurs problématiques. Cette collaboration lui servira pour définir les bons indicateurs, et KPI (*Key Performance Indicator*) sur lesquels il se basera pour qualifier son travail et l'améliorer. Pour reprendre l'exemple de l'institution bancaire qui souhaite détecter des transactions bancaires frauduleuses, un bon KPI pourra être "le nombre de transactions frauduleuses détectées). Si à l'issue du travail du *Data Scientist*, le nombre de détection de fraude augmente, l'équipe métier saura quantifier cette amélioration.

Cette compréhension des règles et contraintes métier permettra au *Data Scientist* de juxtaposer sa compréhension avec ce que révèlent les données. Ainsi, il pourra contextualiser ses découvertes et donner un sens aux observations et aux modèles qu'il produit lors de ses explorations.

Toujours est-il que, grâce à sa compréhension du métier, le *Data Scientist* saura **communiquer dans un discours compréhensible** par les gens du métier, les trouvailles et découvertes qu'il obtient lors de ses explorations de données. Par conséquent, il pourra expliquer les modèles prédictifs qu'il produit tout en cachant la complexité technique sous-jacente à leur élaboration.

Finalement, la vraie valeur ajoutée d'un *Data Scientist* ne réside pas dans ses compétences techniques, mais plutôt dans sa capacité savante de traduire en des mots simples et compréhensibles par les décideurs, ce que cherchent à dire les données.

Expertise mathématique

L'exploration (*Data Mining*) des données et la construction de *Data Product* requiert de manipuler et quantifier les données. Ces dernières, au cœur de la *Data Science*, viennent dans différentes formes, textures et corrélations. Egalement, elles peuvent être modélisées dans des formulations mathématiques. Ces dernières synthétisent le comportement des données et permettent de faire des extrapolations et des déductions pour construire des modèles prédictifs.

Les modèles prédictifs construits à l'aide des algorithmes de *Machine Learning* se basent sur des calculs et formules mathématiques. Souvent, la majorité des gens ne retiennent, **à tort**, que les statistiques comme branche des mathématiques applicables à la *Data Science*. L'algèbre linéaire, les statistiques (classiques et bayésiennes), les probabilités, et les calculs de dérivés sont tout aussi importants pour avoir un profil complet et à l'aise avec les rouages algorithmiques et mathématiques du *Machine Learning*.

Toutefois, il ne s'agit pas de devenir un mathématicien or paire pour faire de la *Data Science*. Il faut surtout être à l'aise avec les mathématiques pour comprendre les fondements théoriques

et les limitations de chaque algorithme de *Machine Learning*. Ainsi, vous saurez choisir les bons outils en fonction des situations et ce que vous souhaitez accomplir.

Expertise informatique

A l'inverse de l'expertise métier et mathématique qui relève du théorique, le volet informatique revêt le côté pratique de la *Data Science*.

Un *Data Scientist* traite souvent de larges volumes de données, qui peuvent provenir de différentes sources et dans différents formats. Pour cette raison, il doit être capable de coder des prototypes dans un langage de programmation.

Il ne s'agit de maîtriser un langage de programmation de ses moindres détails, mais plutôt d'avoir un bon *background technique* pour se débrouiller et produire un code opérationnel. Ainsi, les *Data Scientist* sont généralement familiers avec des langages de programmation comme *SQL*, *Python* ou encore *R*.

Par ailleurs, être familier avec des langages comme *Java* et des infrastructures *Big Data* comme *Hadoop* est un grand plus. En effet, une telle familiarisation permettra au *Data Scientist* de "chapoter" et superviser le travail des *Data Engineers.* Ces derniers se chargeront de traduire le prototype fait par le Data Scientist (souvent écrit en *R* ou *Python*) dans un langage de programmation et une infrastructure plus robuste (comme *Java* ou *Scala* sur une pile *Hadoop* et *Spark*).

Les *Soft Skills* d'un Data Scientist

En outre des *Hard Skills* qu'on vient de présenter jusqu'ici que tout *Data Scientist* se doit de maîtriser, d'autres viennent pour compléter son profil. Il s'agit de compétences non techniques qui lui serviront dans son travail quotidien de *Data Scientist*.

Esprit curieux

Fabriquer un modèle prédictif ou un *Data Product* est un processus heuristique et itératif : On s'améliore au fil des itérations tout en explorant les différentes pistes.

Le *Data Scientist* est quelqu'un de curieux. Il innove en explorant les différentes facettes du problème qu'il résout. Il n'a pas peur de sortir des sentiers battus pour trouver des solutions innovantes.

Par ailleurs, de par la nature des problèmes que traite la *Data Science*, un *Data Scientist* sait qu'il n'y a pas de solution finale à un problème. Son esprit curieux le poussera toujours à aller plus loin dans la quête de l'amélioration du *Data Product* ou modèle prédictif qu'il construit.

Un bon communicant

La communication est un *soft skill* important pour tout *Data Scientist*. De par son profil multi-casquettes, il travaille avec différents profils et départements de l'entreprise. De ce fait, il doit être capable d'adapter son discours en fonction de son interlocuteur.

Pour mieux imager l'importance de la communication pour un *Data Scientist*, imaginons cette mise en situation : Un *Data Scientist* souhaite construire un outil de prédiction des prix de biens immobiliers. Pour un tel projet, plusieurs acteurs seront concernés :

- Les gens du métier : Le *Data Scientist* se rapprochera des agents immobiliers pour comprendre les critères permettant d'estimer le prix d'un bien.
- L'administrateur de base de données (DBA) : Il se rapprochera du *DBA* pour faire l'inventaire de la base de données et recenser les données qui seront utiles à la résolution du problème.
- Les designers HTML : Par ce qu'il s'agit d'un produit professionnel, Ce dernier lui faut une belle interface graphique, simple et intuitive. Par conséquent, il communiquera avec les designers pour qu'ils construisent les écrans applicatifs nécessaires.
- Les décideurs : à la fin du prototype, le *Data Scientist* expliquera par une démonstration l'efficacité de son application

Dans cette mise en situation, on recense plusieurs types d'interlocuteurs. Ainsi, le *Data Scientist* "parlera" plus sur les caractéristiques importantes d'une maison avec les agents immobiliers. Et parlera plus base de données et SQL avec le DBA. La même logique s'applique pour la communication avec les designers et les décideurs.

Ainsi, cette capacité à vulgariser en des mots simples, ce que cherche à dire le *Data Scientist* tout en abstrayant la complexité technique sous-jacente, permettra une bonne dynamique de travail. Par conséquent, ce n'est que la qualité du *Data Product* qui en bénéficiera.

En plus de ses aptitudes de communication, le *Data Scientist* mène souvent un travail de **présentation** de ses idées. Ainsi, il pourra capter l'intérêt des décideurs et de tous les acteurs du projet de *Data Science*.

Les supports de présentation que choisira le *Data Scientist* devront lui permettre de véhiculer clairement ses idées tout en cachant la complexité technique inhérente à son domaine. Par conséquent, le support de présentation doit être fonction de l'interlocuteur, sa fonction dans l'entreprise et son *background* (métier, décideur, DBA...etc.).

Esprit d'analyse et d'abstraction

Un *Data Scientist* est quelqu'un avec une capacité d 'abstraction prononcée. Cette capabilité lui permet de démystifier un problème métier et le traduire sous forme de problématique de *Data Science*.

Souvent, les équipes métiers ou les décideurs ne sauront pas décrire clairement ce qu'ils cherchent à accomplir via un *Data Product*. Le génie du *Data Scientist* se manifeste quand il

analyse leurs expressions de besoins pour en déduire une opportunité de *Data Product* ainsi que les KPI nécessaires à son élaboration.

Finalement, en couplant ses compétences de communication et celles d'analyse, le *Data Scientist* saura mettre en des mots simples ce que les données d'entreprises suggèrent. En effet, les données d'une entreprise sont un grand *asset* qui n'attend qu'à être utilisé. Le *Data Scientist* pointera l'entreprise et l'équipe dirigeante vers les bonnes direction stratégiques… Avec des données à l'appui !

Aiguisez vos talents de Data Scientist

Maintenant, qu'on vient de définir ce que cherche à résoudre la *Data Science*, et ce que c'est un *Data Scientist*. Regardant à présent comment en devenir un. L'idée est de savoir ce qui est important à maîtriser pour optimiser ses efforts d'apprentissage.

Il ne s'agit pas d'être un expert en tout !

Comme on vient de le voir jusqu'ici, la Data Science recouvre plusieurs facettes et domaines. Notamment, le *Data Mining*, le *Machine Learning*, les maths (statistiques, algèbre linéaire...), la programmation...etc.

Devenir expert dans chacun de ces domaines prendra un temps considérable. Bien que ce ne soit pas le but. Rappelez-vous qu'un *Data Scientist* est un ninja ou un bricoleur qui trouve des parades aux problèmes qu'il rencontre afin d'arriver à son but : faire parler les données ou construire un *Data Product*.

Apprendre la *Data Science* ne s'agit pas de connaître tous les arcanes d'un langage de programmation comme *Python*, ou de maîtriser toutes subtilités mathématiques derrière un théorème. A l'inverse d'un mathématicien ou d'un statisticien qui se veut rigoureux dans ses modèles mathématiques, le *Data Scientist* prendra de la hauteur pour extraire de l'information depuis les données. Quitte à faire des concessions et être moins rigoureux mathématiquement parlant.

Le même raisonnement pourrait s'appliquer à la mise en place de modèles avec le langage *Python*, ou *R*. Le *Data Scientist* se soucierait moins de la rigueur des bonnes pratiques de programmation et se concentrera plus sur la mise en place et l'optimisation de son modèle de *Machine Learning*. La robustesse du code du *Data Product* sera délaissée aux *Data Engineers* qui réécriront le code *Python* ou *R* du *Data Scientist* dans des infrastructures plus robustes comme *Java* et *Spark* en utilisant les bonnes pratiques de programmation.

Le Data Mining pour donner un sens aux données

L'essence même de la *Data Science* est la *Data* (les données). Comprendre ces dernières est fondamental pour tout projet de *Data Science*. Ainsi, il est important de connaître les techniques de base d'exploration de données.

Parmi ces techniques, on retrouve les **statistiques descriptives, l'analyse univariée et multivariée** des données. L'analyse univariée et les statistiques descriptives nous donnent une vision sur comment les *features* de notre jeu de données se comportent unitairement et comment elles se répartissent dans l'espace de valeurs. Cette vision est possible grâce à des calculs de moyenne, variance, écart-type etc…

Quant à l'analyse multivariée, elle permet de comprendre les corrélations entre les différentes features. Des outils mathématiques comme la *Covariance* et la *Corrélation* permettent de voir le comportement des features entre-elles.

La visualisation des données (Dataviz)

La visualisation des données est un moyen de représenter ces dernières de façon graphique et visuelle. Bien que les statistiques et l'analyse exploratoire des données soient utiles, elles ne sont, cependant, pas suffisantes.

Visualiser les données permet de mieux comprendre ces dernières. En effet, le cerveau humain a une plus grande facilitée à comprendre des concepts par des images. Exploiter cette faculté naturelle permettra une compréhension plus accrue des données.

Il est important pour tout *Data Scientist* de **connaître les techniques de base de visualisation** de données. Notamment les *histogrammes, les Scatter Plots, box plots*… etc. Il ne s'agit pas de savoir les dessiner, mais de plutôt de savoir quand il faut les utiliser, quelles sont leurs limites et surtout comment les interpréter. L'important est qu'à l'issu d'une étape de *Data Visualisation*, vous ayez des pistes et des hypothèses sur votre jeu de données que vous venez de visualiser.

Un Data Scientist est aussi un matheux

La modélisation des modèles prédictifs de *Machine Learning* se base sur des calculs mathématiques. Entre autres, sur le calcul matriciel, le calcul vectoriel, l'analyse, la théorie de probabilités, les statistiques… etc.

Bien entendu, il ne s'agit pas d'être *Henri Poincaré* ni autre mathématicien hors pair. Mais plutôt d'être à l'aise avec concepts mathématiques utiles à la *Data Science* et ne pas être allergique à des formulations mathématiques rigoureuses.

Personnellement, lors de mon apprentissage de la *Data Science* et du *Machine Learning*, je me suis pris du temps pour me rafraîchir la mémoire sur certaines notions mathématiques. Notamment, les statistiques, la théorie de probabilités (les lois de distributions, les variables aléatoires …) ainsi qu'un peu de calcul matriciel.

Mon conseil est que si vous êtes comme moi, quelqu'un qui aime comprendre la théorie derrière les concepts, quitte à mettre un peu plus de temps dans votre apprentissage, attardez-vous sur ces concepts mathématiques. C'est un bon investissement que vous faites qui vous armera mieux pour votre apprentissage du *Machine Learning*.

A l'inverse, si vous êtes du genre à apprendre par la pratique et que vous n'aimez pas trop la théorie. Laissez de côté les formulations mathématiques. Vous y reviendrez quand le besoin se fera ressentir.

Accédez aux coulisses du Machine Learning

Le *Machine Learning* représente un grand pavé de la *Data Science*. Grâce à ce dernier, les *Data Scientist* construisent des modèles prédictifs qui apprennent à partir des données.

Vous pouvez imaginer le *Machine Learning* comme la boîte à outil du *Data Scientist*. Elle contient plusieurs outils qu'il faudra maîtriser et savoir utiliser à bon escient et dans la bonne situation. Ainsi, ces outils sont les algorithmes de *Machine Learning*. Il en existe une multitude et chacun de ces algorithmes à ses propres caractéristiques, ses prérequis d'utilisation, ses limitations...etc.

Généralement, ces algorithmes sont déjà codés dans des librairies dédiées au *Machine Learning*, comme *Scikit Learn* de *Python*, ou encore *Spark MLlib* de *Java*. Par conséquent, vous n'aurez pas à les coder par vous-même, il suffira juste de les utiliser.

Toutefois, ce qui vous permettra de vous démarquer des autres, est la maîtrise et la compréhension du fonctionnement interne de ces algorithmes. Ainsi, ces librairies ne vous sembleront pas comme des boîtes noires au fonctionnement magique.

Comprendre le fonctionnement interne d'un algorithme de *Machine Learning*, revient à comprendre son procédé d'apprentissage depuis les données (comment il calcule le modèle prédictif), ce qu'il demande comme prérequis d'entrée (le format des données, leur distribution

statistique etc…). Sans oublier, qu'il faut être capable d'expliquer le modèle qu'il vous aura calculé.

Comprendre ces notions sur chaque algorithme vous permettra, premièrement, de choisir le bon algorithme en fonction du problème que vous cherchez à résoudre. Et deuxièmement, de comprendre le résultat produit par cet algorithme et comment procéder pour optimiser et améliorer les performances du modèle prédictif produit.

Finalement, **les algorithmes du *Machine Learning* sont nombreux**, connaître les subtilités de chacun de ces algorithmes prendra du temps. Ce sont des choses qui viennent par la pratique et avec le temps. C'est surtout dans la phase d'optimisation d'un modèle prédictif, qu'il faudra voir en détail l'algorithme pour mieux aligner vos données sur ses prérequis et propriétés algorithmiques.

Votre apprentissage...dans la pratique !

Un Apprentissage sur plusieurs fronts

Maintenant qu'on sait sur quels points il faut accentuer ses efforts d'apprentissage, il reste à définir la chronologie de ce dernier. Rappelez-vous qu'il faut attaquer aussi bien la théorie (les maths), que le côté pratique (exploration de données, programmation de modèles prédictifs) ...etc.

A mon avis, il ne faut pas attaquer votre apprentissage depuis un seul angle. Comprendre toutes les subtilités du calcul matriciel, puis plonger dans la théorie des algorithmes de *Machine Learning*, avant d'entamer l'apprentissage de *Python* est contre-productif. Cela parce que, chacun de ces sujets est vaste et vous mettrez trop de temps avant de coder votre premier modèle prédictif. Par conséquent, vous risquez de vous lasser rapidement.

Le mieux, est que vous entamiez votre apprentissage sur plusieurs fronts. Vous vous lasserez moins et vous construirez votre profil de *Data Scientist* de façon plus complète. Il est préférable de connaître un peu de *Dataviz*, un peu des maths, et de se débrouiller en *Python* ou *R* pour entamer vos premiers pas en *Data Science*. Vous allez éprouver des difficultés aux débuts, mais une fois vous les aurez surmontées, les solutions que vous aurez trouvées resteront à toujours dans votre esprit.

Un apprentissage par la pratique

C'est en forgeant qu'on devient forgeron. C'est cet état d'esprit du proverbe qu'il faut adopter. Vous aurez beau apprendre la théorie, la mise en pratique de la *Data Science* sera tout autre.

Comme nous l'avons vu, la *Data Science* est un domaine à l'intersection de plusieurs disciplines. Cet aspect multidimensionnel de la science de données, est dû aux étapes multiples nécessaires à la création d'un modèle prédictif ou d'un *Data Product*. Notamment, le

chargement, l'analyse et la visualisation des données, programmation d'un modèle de *Machine Learning* et l'optimisation de ce dernier.

Toutes ces facettes ne peuvent être assimilées qu'avec la pratique. Je ne parle pas de prendre des bouts de code, sur internet, tout prêt, tout fait et les analyser. Je parle de coder manuellement vos propres modèles de *Machine Learning* et connaître par vous-même les difficultés inhérentes à un projet de *Data Science*. Ainsi vous vous forgerez votre propre profil de *Data Scientist*. Par ailleurs, lorsque vous construisez vos propres projets *Machine Learning*, vous pourrez les montrer lors des sessions de recrutements pour soutenir votre candidature.

Ne soyez pas découragé si vos premiers pas dans la *Data Science* se font timides. La traction au début est faible, mais plus vous avancez dans le domaine plus vous vous familiariserez et vos progrès seront plus notables.

Gérer la difficulté

En termes de difficulté, ne placez pas la barre trop haute, au risque de vous décourager, faute de résultats. Commencez par des jeux de données simples, comportant peu d'observations et un nombre faible de features. Concernant les algorithmes, commencez par les plus simples comme la *Régression linéaire ou la régression logistique.* Vous aurez largement le temps d'étudier des algorithmes plus complexes comme les réseaux de neurones, le *Random Forest,* et le *Deep Learning*.

Soyez organisé

La *Data Science* n'est pas que son aspect technique. C'est aussi l'aspect organisationnel. Prenez l'habitude de **décortiquer entièrement le problème** que vous étudiez. Ne vous vous arrêtez pas à coder un simple modèle prédictif. Même si vous apprenez seul, prenez le temps **de faire un bilan** rigoureux de chaque phase de votre projet. Faites comme si vous devez rendre votre travail à quelqu'un. Ce dernier doit être capable de comprendre votre démarche ainsi que votre raisonnement et apprécier à sa juste valeur vos conclusions. Cette rigueur à un double bénéfices :

- Vous aurez un bon support pour y revenir dans le futur pour vous rafraichir la mémoire en cas de besoin
- Cette démarche rigoureuse deviendra une habitude de travail qui vous sera utile dans le cadre professionnel

WHAT'S NEXT ?

Pour aller plus loin...

Un voyage de mille lieux commence toujours par un premier pas. A ce stade, vous avez déjà fait un grand pas dans le monde de la *Data Science*. Félicitations ! vous venez d'accéder au club très restreint des futurs *Data Scientist* !

J'espère que vous ai donné une vision plus claire via cet ebook sur la *Data Science* et comment entamer votre apprentissage de ce domaine. J'ai pris le soin de ne pas rentrer dans l'aspect technique de la *Data Science* pour que vous puissiez prendre de la hauteur et avoir une vision panoramique sur cette dernière. Cette vision nous aura permis, à travers les pages de ce livret, de définir un chemin d'apprentissage complet et cohérent de la *Data Science*.

Maintenant que vous ayez pris vos repères sur comment devenir *Data Scientist*, je vous invite à jeter un coup d'œil sur mon blog français https://mrmint.fr. Dans ce dernier, je rentre en détails dans les différents aspects techniques de la *Data Science* et le *Machine Learning*.

Si je dois vous donner un dernier conseil, mon conseil ultime en quelque sorte, c'est surtout de ne pas rester dans votre coin lors de votre apprentissage. La *Data Science* est un domaine en pleine ébullition. Il y a des groupes *Meetup*, des communautés et des discussions en ligne tenus par des gens tout aussi passionnés que vous. Rentrez en contact avec du monde, c'est votre apprentissage qui vous en remerciera.

Toutefois, la majorité de ces communautés sont en anglais. Pour, palier à ce vide linguistique, j'ai créé une page Facebook ainsi qu'une communauté sur Google+, toutes les deux dédiées aux francophones passionnés de *Data Science* et du *Machine Learning*. N'hésitez pas à y faire un tour pour partager vos expériences, nouer de nouvelles relations, et pour recevoir les dernières nouveautés !

Finalement, cet ebook touche à sa fin. J'espère que vous ayez pris autant de plaisir en le lisant que moi j'en aurai pris à sa rédaction. Toutefois, si vous avez des commentaires sur son contenu, des questions ou des idées d'amélioration, écrivez-moi sur cet e-mail : younes.benzaki@mrmint.fr ça sera mon plaisir d'entendre de vous et écouter vos idées.