Module 1
Introduction à la science des données

Sommaire

Dernière mise à jour le 19 septembre 2018

Introduction

La science des données est une discipline, qui n'est certes pas nouvelle, mais qui a pris une grande ampleur dans les dernières années. Comme son nom l'indique, la science des données ne concerne pas un domaine en particulier, mais elle s'intéresse plutôt à tous les aspects liés aux données notamment la collecte, le stockage, l'analyse, le transfert, le nettoyage, le filtrage, l'anonymisation, le cryptage, la visualisation, etc. Bien que les données et leur traitement existaient bien avant l'émergence du concept de science des données, le volume très important de données disponibles présentement (données massives - *big data*) et leur démocratisation, ainsi que l'augmentation des puissances de calcul nous amènent vers une nouvelle ère d'analyse des données qui n'était certainement pas possible à l'époque. Dans ce qui suit, nous allons présenter les différents aspects liés à la science des données et les caractéristiques de ceux et celles qui font de la science des données communément appelés scientifiques de données (*data scientists*). Nous allons d'abord définir le concept de la science des données, sa relation avec les données massives, ses perspectives ainsi que les compétences nécessaires pour réussir la fonction de scientifique de données.

1.1 Définitions

Dans cette section, nous allons définir le concept de la science des données. Il n'existe pas une définition claire et communément adoptée par la communauté scientifique et industrielle. Il existe quelques définitions qui correspondent plus à la réalité de la science des données indépendamment du domaine d'application.

Définition 1. Science des données
Selon l'université Johns Hopkins[1], la science des données « **c'est l'art de répondre à des questions d'importance avec des données** ».

Définition 2. Science des données
Selon INVESTOPEDIA[2], la science des données « **c'est un domaine des données massives qui cherche à fournir des informations significatives à partir de grandes quantités de données complexes** ».

Définition 3. Science des données
Selon les auteurs Cathy O'Neil et Rachel Schutt dans Doing Data Science[3], la science des données « **est l'ingénierie civile des données. Ses acolytes possèdent une connaissance pratique des outils, associée à une compréhension théorique de ce qui est possible** ».

Ces définitions résument des points importants qui requièrent d'être soigneusement analysés par celui qui fait la science des données. En même temps, ces points représentent, entre autres, les tâches d'un scientifique de données.

— Connaissance des méthodes et outils pour le traitement et l'analyse des données.
— Identification d'une question pertinente.
— Assemblage des données pour répondre à la question.
— Développement des modèles statistiques.

— Communication des résultats d'une manière compréhensible et utilisable.

La science des données est une discipline qui fait appel à plusieurs domaines à savoir les mathématiques, les statistiques et l'informatique comme le montre la figure suivante.

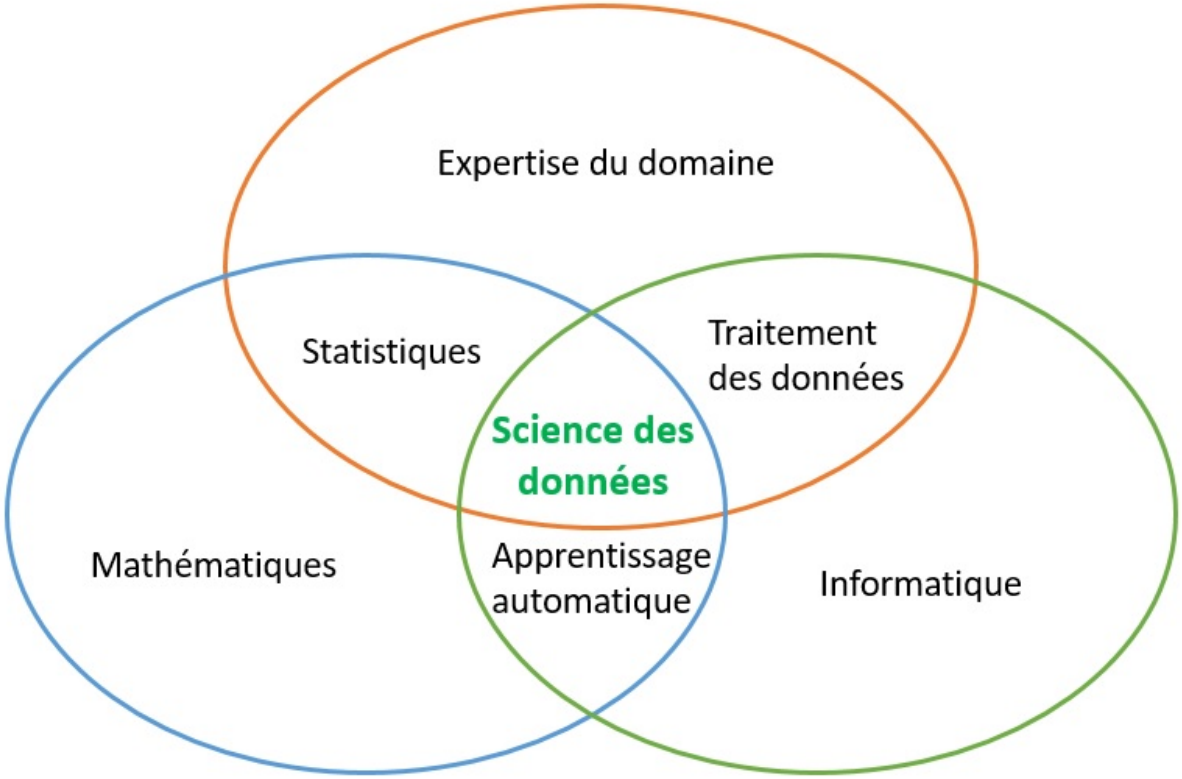

Figure 1.1: Positionnement de la science des données par rapport aux autres domaines. Traduction de l'image tirée de la source : Palmer, Shelly. Data Science for the C-Suite. New York : Digital Living Press, 2015. Print.

La science des données n'est pas un synonyme de l'apprentissage automatique. La science des données utilise plutôt l'apprentissage machine comme une étape dans un processus global d'analyse des données.

1.2 Science des données et les données massives

La science des données a été, dans les derniers temps, souvent liée aux données massives. Bien que la science des données existait avant l'apparition du terme données massives, le besoin de traiter des données massives et le développement rapide des infrastructures et outils de traitement des données massives font en sorte que la science des données émerge de plus en plus dans pratiquement tous les domaines. La science des données est liée au données peu importe leur taille ou provenance.

Définition 4. Données massives
Selon le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) de la France[4], les données massives (Big Data) « **désigne à la fois la production de données massives et le développement de technologies capables de les traiter afin d'en extraire des corrélations ou du sens** ».

Selon le CEA toujours, « c'est dans les années 1990 que le terme Big Data (données massives) prend sa signification actuelle d'un défi technologique à relever pour analyser de grands ensembles de données, d'abord scientifiques, mais de plus en plus souvent collectés au quotidien par divers moyens techniques ».
Le term Big Data (données massives) est lié à trois caractéristiques fondamentales souvent appelées VVV ou 3V (Volume, Variété et Vélocité) comme le montre la figure suivante.

— Volume : décrit la quantité de données générées.

Figure 1.2: Les trois caractéristiques des données massives. Image tirée du IBM InfoSphere streams (https ://www.ibm.com/developerworks/library/bd-streamsintro/index.html).

— Variété : décrit la diversité des types de données provenant de sources multiples.

— Vélocité : décrit la fréquence à laquelle les données sont générées, capturées, parta-

gées et traitées.

Les trois caractéristiques des données massives engendrent, donc, beaucoup de défis techniques et computationnels pour les scientifiques de données afin d'analyser ces données et d'en extraire des informations pertinentes en temps réel.

1.3 Scientifique de données

Il est important de parler des caractéristiques d'un scientifique de données, les compétences requises, et les différentes tâches qu'il devrait accomplir. Nous avons vu, dans les sections précédentes, le concept de la science des données et les différents domaines auxquels il fait appel comme les mathématiques, les statistiques et l'informatique. Il n'existait pas, il y a quelques années, des formations académiques pour la discipline de la science des données pour la formation des scientifiques de données. Cependant, récemment beaucoup d'universités et instituts à travers le monde ont lancé des programmes relatifs à la science des données et les données massives.

1.3.1 Compétences

Selon un article publié par le magazine Le Big Data en France [5], le métier de scientifique de données a été élu en janvier 2017, dans le site de recherche d'emploi Glassdoor, en première position de son top 25 des meilleurs métiers du monde. Cela démontre entre autres la démocratisation de la science des données et le besoin accru d'experts pour ce domaine.
Selon le même article, bien que le métier de scientifique de données et sans doute passionnant, il s'agit également d'un métier à haute responsabilité qui nécessitent des prédisposions naturelles et une éducation de haut niveau. L'article définit 13 compétences

5. Voici les 13 compétences nécessaires pour devenir Data Scientist, https ://www.lebigdata.fr/13-competences-necessaires-devenir-data-scientist

nécessaires pour exercer le métier de scientifique de données. Ces compétences sont résumées dans les points suivants :

1. Formation : selon l'article, 88% des scientifiques de données sont diplômés au minimum d'un master[6], et 46% d'entre eux possèdent un doctorat. Cette formation semble nécessaire pour développer le niveau de connaissance suffisant pour permettre l'exercice de ce métier.
2. Connaissance des statistiques : le domaine de statistiques est l'un des domaines de base de la science des données, et la connaissance des statistiques représente un élément de base pour qu'un scientifique de données remplisse avec succès les mandats qui lui sont confiés. La plupart des scientifiques de données ont une formation dans le domaine des mathématiques et des statistiques (32%), et (19%) ont une formation en informatique, et (16%) ont une formation en sciences de l'ingénieur.
3. Maîtrise des outils analytiques : la maîtrise des outils d'analyse des données est importante pour pouvoir traiter les données. Ces outils incluent entre autres R, SAS, Matlab, etc. Pour la science des données, les préférences convergent vers R. Les auteurs de l'article[7] préfèrent le langage informatique R qui est normé pour l'analyse et l'exploration des données. Le langage R est utilisé dans ce cours SCI 1016 mais aussi dans les autres cours du programme court en science des données.
4. Langage de programmation : le scientifique de données devrait maîtriser au moins un langage de programmation pour faciliter la mise en production. Le langage le plus couramment utilisé est Python, mais d'autres langages peuvent être utilisés également comme Java, C++, etc.
5. Apprentissage automatique : la connaissance de quelques méthodes d'apprentissage automatique est un réel atout pour un scientifique de données. Les méthodes d'apprentissage automatiques incluent entre autres les arbres de décision, les K plus proches voisins, K moyennes, les méthodes Bayésiennes et les méthodes d'ensembles comme les forêts aléatoires.

6. Un master est équivalent à une maîtrise au Québec.
7. https ://www.lebigdata.fr/13-competences-necessaires-devenir-data-scientist

6. L'algèbre linéaire : beaucoup de calculs scientifiques et de méthodes d'apprentissage automatiques sont basés sur l'algèbre linéaire. Bien que l'algèbre linéaire est implémentée dans la plus part des langages de programmation, il est très utile d'avoir des connaissances de cette branche des mathématiques.
7. Hadoop[8] : la connaissance de la plateforme Hadoop est essentielle et est le plus souvent requise, même si certaines entreprises ne l'exigent pas. La connaissance des outils de traitement des données massives (Spark, Pig, Hive, etc.) est souvent un atout et représente un argument supplémentaire en vue d'un recrutement.
8. Connaissance du SQL : le SQL est le langage d'interrogation des bases de données le plus utilisé. Bien que les nouvelles générations des bases de données non-structurées exigent d'autres langages comme NoSQL, le SQL reste le moyen le plus utilisé pour formuler et exécuter des requêtes.
9. Données non-structurées : les données qui proviennent des réseaux sociaux, du Web en général incluant des textes, videos et audios sont des données non-structurées qui nécessitent des méthodes particulières pour pouvoir les traiter et analyser. Le scientifique de données devrait avoir des connaissances lui permettant de traiter les données comportant des imperfections telles que des valeurs manquantes ou des chaînes de format incohérentes.
10. Ingénierie logicielle : le scientifique de données devrait avoir des compétences en ingénierie logicielle qui lui permettent de prendre en charge le développement d'un produit de bout en bout, c'est-à-dire, de la phase de définition des besoins jusqu'à la mise en production du produit.
11. Curiosité intellectuelle : le scientifique de données devrait être créatif et poserait ses propores questions pertinentes, et non seulement répondre aux questions qui lui sont posées.
12. L'esprit d'un entrepreneur : il est important pour le scientifique de données de comprendre le mode de l'entreprise pour qui il travaille et le business auquel il est

8. http://hadoop.apache.org/

affilié en particulier. Cela permettra de comprendre les problèmes à résoudre et les différentes possibilités que les données peuvent offrir.

13. Communication : le sens de la communication est primordial pour un scientifique de données. Cela lui permettra de communiquer clairement ses idées, ses résultats, et ses découvertes techniques aux autres employés de l'entreprise dans les différents domaines tel que le marketing, les ventes,

La figure ci-dessous montre les différents domaines d'intérêt et les compétences requises pour un scientifique de données. Figure tirée de [9].

9. The Role of a Data Scientist in 2016 : https ://www.infoq.com/articles/role-of-a-data-scientist-in-2016

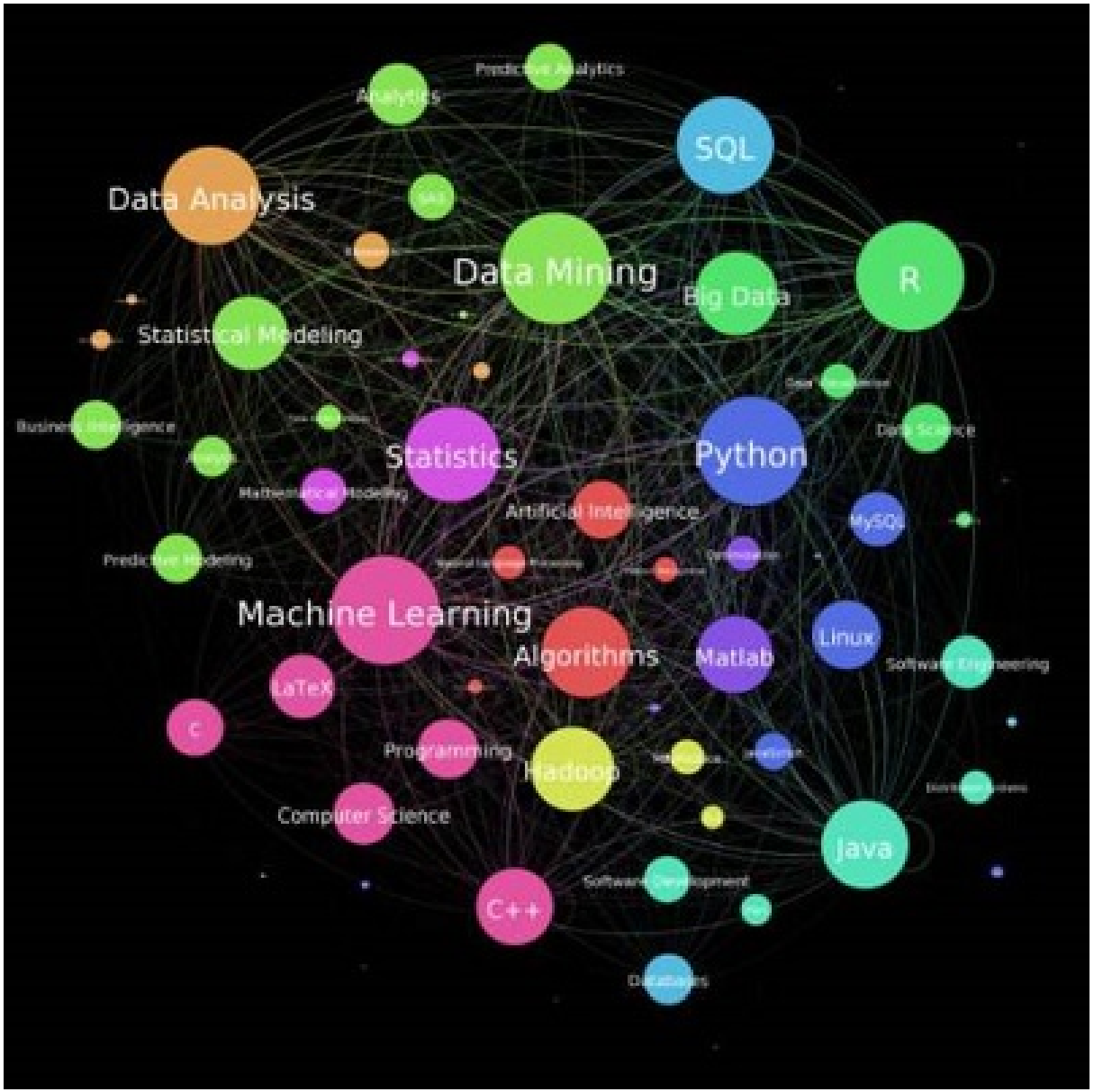

Figure 1.3: Les différents domaines d'intérêt et les compétences requises pour un scientifique de données.

1.3.2 Rôle d'un scientifique de données

Nous avons vu, dans la section précédente, les compétences requises pour un scientifique de données. Dans cette section, nous allons présenter son rôle plus particulièrement dans les entreprises.
Le rôle d'un scientifique de données est de produire des méthodes (automatisées, autant que possible) de tri et d'analyse de données massives et de sources plus ou moins complexes ou disjointes de données, afin d'en extraire des informations utiles ou potentiellement utiles.
Le rôle du scientifique de données ne se limite pas à analyser les données et voir ce qui se passait dans le passé, mais plutôt il va plus loin que ça en regardant qu'est-ce qui va se passer dans le futur en effectuant des analyses prédictives. À l'aide des méthodes statistiques avancées et la modélisation complexe des données, le scientifique de données devra être capable d'extraire des patrons cachés dans les données et d'effectuer des prédictions.

1.4 Perspectives

L'émergence des données massives et la démocratisation de l'intelligence artificielle, font en sorte que le domaine de la science des données occupera une place assez large autant sur le plan académique et scientifique que sur le plan industriel. La compagnie Gartner Inc a publié un rapport [10] indiquant que plus de 40% des tâches effectuées par un scientifique de données seront automatisées d'ici 2020.

10. https://www.gartner.com/newsroom/id/3570917