La data science : quésaco ?

Cédric Bohnert

23/10/2021

Notes de lecture/cours issues des ressources suivantes :

- The Data Science Manual - Steven S. Skiena
- The Data Science Specialization - Jeff Leek and the Data Science Specialization Team

Qu'est-ce que la data science ?

Une très large définition serait de stipuler que faire de la data science, c'est répondre à une question en utilisant des données.

La data science est une science interdisciplinaire qui croise la statistique, la science informatique et la mathématique.

On y rencontre des tâches comme nettoyer, formater, visualiser, modéliser des données.

De nombreuses compétences techniques et de connaissances conceptuelles se rassemblent dans ce métier.

La data science est apparue du fait de l'énorme quantité de données actuellement disponible et celle-ci est générée quotidiennement à une vitesse proportionnelle à cette quantité.

Bien entendu, cette dernière affirmation fausse stricto sensu, c'est-à-dire au sens mathématique du terme.

Cependant on pourrait s'intéresser de manière rigoureuse à l'estimation du volume de données disponible dans le monde.

Un travail a été réalisé dans ce sens par Maurice de Kunder sur l'estimation du nombre de pages indexées dans le World Wide Web :

- www.worldwidewebsize.com
- Estimating search engine index size variability: a 9-year longitudinal study

On pourrait tenter de comprendre ce travail tout en réalisant un TD ou TP pour y extraire un succint aperçu de la méthodologie de l'étude scientifique.

Dans data science, il y a le mot science !

Un data scientist efficace est quelqu'un qui pense et réfléchit comme un scientifique :

- Un scientifique cherche et fouille des données pour répondre à ses interrogations.
- Un scientifique se soucie de la signification des données et du sens de ses réponses.
- Un scientifique se soucie des biais et erreurs dans les données qui affectent la robustesse de ses réponses.
- Un scientifique est n'est pas tant concerné par la précision de ses réponses que par leur signification.

Un aspirant data scientist se doit donc d'apprendre à penser comme un vrai scientifique.

Il a pour rôle de produire de la connaissance à partir des données.

La data science est une science propulsée par les données et un bon data scientist est concerné par les deux questions suivantes :

- Un problème étant donnée, quelles données sont disponible pour aider à le résoudre ?
- Quels problèmes intéressants peut-on résoudre avec un jeu de données spécifique ?

Poser d'intéressantes questions à partir de données !

Les data scientists se posent toujours des questions et s'intéressent à beaucoup de choses.

- Quels trucs pouvons-nous apprendre d'un jeu de données?
- Que voulons-nous apprendre à propos du monde?
- Qu'est-ce que cela signifie pour nous de savoir ce truc?

Devenir data scientist implique d'apprendre à poser de bonnes questions concernant des données.

Pour cela, la clé est de ratisser large : les réponses aux grosses questions se cachent souvent dans des jeux de données très spécifiques.

Exemples de jeu de données pour se poser des questions :

Les quatre sections suivantes présentent brièvement des jeux de données avec lesquelson peut s'entrainer à formuler des questions intéressantes.

L'encyclopédie américaine de baseball : www.baseball-reference.com

Le baseball est un sport très riche en statistiques et permet de nourrir un esprit quantitatif curieux de plein de questions intéressantes.

Par exemple, voici 5 questions autour des finances que l'on peut tenter de répondre avec les données de ce sport :

- Quel est le joueur le plus cher du MLB ?
- Combien coûte en moyenne une MLB ?
- Quel est le meilleur joueur le moins cher de l'histoire du MLB ?
- Quel est le pire joueur le plus cher de l'histoire du MLB ?
- Combien va coûter la prochaine compétitions ?

A vous de jouer ! Formuler 5 questions autour du baseball auxquelles on pourrait répondre grâce à ce jeu de données.

Deux éléments sont essentiels lorsque l'on questionne un jeu de données :

- Les métadonnées sont souvent aussi importantes que les données elles-même.
- La proximité des données, où l'on substitue les données que l'on a à celles que l'on aimerait vraiment.

Un bon data scientist est pragmatique et regarde ce qu'il a au lieu de ce qu'il rêverait d'avoir.

La base de données internet des films (IMDb) : www.imdb.com

C'est un autre exemple très riche en données pour exercer ses talents de data scientist.

Par exemple, on peut se poser les 5 questions suivantes autour des acteurs et de leur famille :

- Quelle est la plus grande famille d'acteurs du cinéma ?
- Quels sont les jumeaux les plus populaires du cinéma ?
- Quel est l'acteur mort prématurément le plus populaire de l'histoire du cinéma ?
- Quel est le film qui réunit le plus d'acteurs d'une même famille ?
- Quelle est la famille de l'histoire du cinéma la plus détestée par le public ?

A vous de jouer ! Formuler 5 questions autour du cinéma auxquelles on pourrait répondre grâce à ce jeu de données.

L'outil Google Ngrams : Google a entrepris de digitaliser l'ensemble des livres publiés depuis l'invention de l'impression par Gutemberg. Un corpus de 20%, c'est-à-dire près de 30 millions d'ouvrages, a ainsi été digitalisé depuis le début de ce projet.

Google met à disposition un outil permettant de fournir la fréquence à laquelle apparaissent de courtes phrases dans le corpus :

Google Ngrams

On peut ainsi comparer l'évolution temporelle d'expressions et de termes dans les sociétés. C'est potentiellement un fantastique télescope pour observer dans le passé de l'humanité.

Voici 5 exemples de questions que l'on peut être amener à se poser grâce à cet outil :

- Quel est le sentiment qui perdurent le plus dans l'histoire des écrits ?
- A quelle période parle-t-on le plus de guerre et de paix ?
- Comment évolue la notion de nourriture dans l'histoire ?
- A partir de quand une expression se modifie-t-elle dans l'histoire ?
- La notion de politique a-t-elle bonne réputation ?

A vous de jouer ! Formuler 5 questions auxquelles on pourrait répondre grâce à cette outil.

Les enregistrements des taxi de New York : Les taxi new yorkais enregistrent toutes les données de leurs courses grâce à un GPS et un terminal de paiement. Ces données sont centralisées et rendues accessibles à l'agence publique 'New York Taxi and Limousine Commission'.

Voici 5 exemples de questions que l'on peut se poser avec ce jeu de données :

- Quel secteur géographique de New York enregistre le plus de courses ?
- Peut-on prédire les pics de demande de courses afin d'optimiser le nombre de taxi nécessaires ?
- Quel est le budget de déplacement d'un new yorkais vivant dans un quartier spécifique ?
- Peut-on prédire avec ces données les problèmes de congestion des rues new yorkaises ?
- Quelle est la part de l'activité des taxi dans l'économie du transport à New York ?

A vous de jouer ! Formuler 5 questions auxquelles on pourrait répondre grâce à ce jeu de données new yorkais.

Les données ont des propriétés !

Pour une meilleure compréhension de son travail, un data scientist se doit d'apprécier les propriétés d'un jeu de données.

Tout d'abord, revenons à trois principales caractéristiques du phénomène de disponibilité des données générées grâce aux avancées technologiques.

C'est le phénomène de mode que l'on appelle couramment Big Data :

- Le volume : de très gros jeux de données de l'ordre du teraoctet sont déjà disponibles.
- La vélocité : les données sont collectées et générées plus vite que jamais dans le passé.
- La variété : les données sont disponibles sous des formats très divers (textes, nombres, images, sons, vidéos, etc)

Ces trois caractéristiques font que la data science émerge et se développe parallèlement à la technologie.

A vous de jouer ! Si 300 heures de vidéos sont uploadées sur YouTube par minute, pendant combien d'années devrions nous visionner le volume de vidéos uploadé sur une période d'un an ?

Voici une taxonomie des propriétés de la donnée :

Données structurées versus données non structurées : On parle de données structurées lorsqu'elles se présentent sous la forme d'une matrice où les lignes représentent des objets (ou enregistrements) distincts et les colonnes des propriétés de ces objets.

On parle de données non structurées lorsqu'elles ne se trouvent pas sous cette forme de tableau.

Les images et liens de Wikipédia sont des données non structurées alors qu'un tableau Excel est un parfait exemple de données structurées.

Données quantitatives versus données catégorielles :