# DECRYP'TECH
# LA DATA SCIENCE
## PAR CAROLE, CONSULTANTE IT

Weeneo

***La Data Science ou Science des Données, c'est l'analyse et la présentation de données dans le but de produire de la connaissance.***

## CHIFFRES CLES

- 1,7 Mo c'est le nombre de données générées par une personne chaque seconde dans le monde (Baromètre Data Never Sleeps, Domo).
- 701% d'augmentation du volume de données entre 2016 et 2018 en France (rapport Global Data Protection Index, Dell EMC).
- 189 milliards de dollars de revenus produits par le secteur dans le monde (cabinet IDC).
- 10,6 % de croissance du marché de la Data Science entre 2020 ($206.95 billion) et 2021 ($231.43 billion).
- 13,2 % de croissance du marché de la Data Science prévue sur les 10 prochaines années atteignant en 2028 $549.73 billion (CAGR).
- $46 000 à $90 000 c'est le salaire annuel d'un Data Analyst junior, $96 000 pour un Data Analyst plus expérimenté et $105 000 à partir de 10 ans d'expérience (Glassdoor).

## EXEMPLES DE DOMAINES D'APPLICATION

**Marketing** ciblé/ Fidélisation client

**Santé :** classification d'affection bénigne ou maligne

Prédictions de la **météo**

**Industrie :** prédictions pour diminuer les risques

Monitoring de systèmes complexes

Prédictions du comportement d'un utilisateur sur un site web

Spotify créé et propose des playlists par rapport aux prédictions

**Sport :** prédictions des performances individuelles

Collecte massive des données : Google, Amazon

RAPIDE HISTORIQUE

## *Années 1930*

Existence de certaines méthodes d'analyse de données restreintes par le manque de données lié à la capacité mémoire limité et par les limites des machines en termes de puissance et de temps de calcul.

## *1992*

*A*pparition du terme data science lors du 2e colloque franco-japonais de statistique tenu à l'Université Montpellier II. Les participants ont reconnu l'émergence d'une nouvelle discipline au cœur de laquelle se trouvent des données de toutes origines, tailles, types et structures. Cette activité doit s'appuyer sur des concepts et des principes reconnus de la statistique et de l'analyse des données tout en exploitant pleinement la puissance croissante des outils informatiques.

## *2001*

William Cleveland reprenait essentiellement les mêmes idées dans un article programmatique paru en 2001 "Data Science : An Action Plan for Expanding the Technical Areas of the Field of Statistics" qui précise les contours de cette discipline émergente. Cette discipline est issue de l'apparition et du développement des bases de données et d'Internet. Elle a reçu beaucoup d'attention dernièrement grâce à l'intérêt grandissant pour les "données massives " ou Big Data.

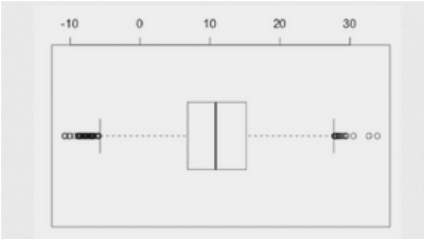
## *Fin des années 2000*

*L*'explosion de la puissance disponible associée à la baisse des coûts de stockage de données a permis l'application concrète de ces théories dormantes. Les données devenant dans le même temps accessibles massivement à distance, l'accélération a été fulgurante.

# DEBUTER EN DATA SCIENCE

## DES BASES EN STATISTIQUES

- Les statistiques : moyenne, médiane, variance, écart-type et intervalles de confiance ;
- Les représentations graphiques : histogramme, box plot ou boîte à moustache (pour étudier la distribution), diagramme en bâtons, diagramme de dispersion ;

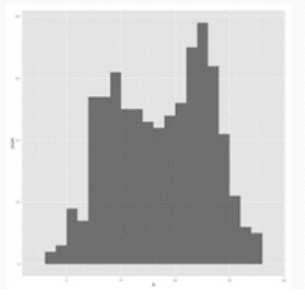



## EXPLORER; FOUILLER ET PREMIERES ANALYSES

- La technique du **“hack”**, c'est-à-dire de la bidouille ;

- Pour débuter, pourquoi pas commencer par ouvrir et explorer vos données dans un tableur, Microsoft Excel ou Google Spreadsheet, réaliser des tableaux pour résumer les données (compter les valeurs, donner la répartition, croiser les données et les informations) et des graphiques pour les visualiser sont des actions de base de tout Analyst et Data Scientist.

- La fouille des données requiert un bon aperçu de l'ensemble des données. A l'échelle d'une entreprise, cela nécessite de se familiariser avec des données provenant de divers services et qui sont historisées depuis plusieurs années. Bien entendu, cela a ses limites, on ne fera pas du Big Data avec un tableur.



# DEBUTER EN DATA SCIENCE

## LE SQL POUR TOUS

- Le travail sur base de données permet de s'affranchir de nombreuses limites de Excel. On ne travaille généralement plus sur un poste mais sur un serveur dédié (voire même un ensemble) et cela ouvre la possibilité au traitement de gros volumes de données, de façon simultanée et reproductible ;
- Il existe un langage normalisé servant à exploiter des bases de données : le SQL (Structured Query Language). Grâce au SQL, il est possible d'interroger les bases de données afin de retrouver des données, de les croiser et de créer des statistiques sur ces mêmes données ;
- Le langage SQL est très utilisé dans tous les domaines, pas seulement en informatique. Biologistes, médecins, économétriciens, marketeurs et toutes les professions qui ont besoin de travailler sur des données statistiques s'y forment. Il offre l'intérêt de manier d'importants volumes de données, de faire des agrégations ou des calculs sur les données sans s'occuper de l'implémentation technique. Il n'y a pas besoin de coder des boucles, de trouver un algorithme performant par exemple.

## CONNAISSANCE DES ALGORITHMES DE MACHINE LEARNING

- Après l'exploration et la préparation vient la **phase de modélisation et d'apprentissage** sur les données. La vraie valeur ajoutée sur vos données est obtenue sur cette phase.
- Pour débuter, il est bien d'avoir des notions des modèles statistiques et algorithmes d'apprentissage automatique (**Machine Learning** en anglais).

- 3 techniques de modélisation ou d'apprentissage statistique :

  **La régression** qui permet d'expliquer une variable numérique en fonction d'autres variables. A noter qu'il existe plusieurs types de régressions (linéaire et non linéaire, simple et multiple) ;

  **La classification** qui range des éléments dans des classes. Très souvent, la classification est binaire (c'est-à-dire qu'il y a seulement deux classes, par exemple « oui » ou « non »). Il existe plusieurs algorithmes de classification dont l'apprentissage par arbres de décision.

  **Le clustering** qui consiste à diviser un ensemble d'éléments en groupes homogènes. Un algorithme assez connu est celui des k-means

Des disques de couleurs pour visualiser les clusters, réalisés avec RAW.

# DEBUTER EN DATA SCIENCE

## LA VISUALISATION DES DONNEES

- La visualisation que l'on appelle parfois data-viz intervient tout le long d'un projet sur la donnée afin d'accompagner la compréhension. Construire des graphiques ou autres visualisations aide à appréhender les données et les chiffres.

  - Surtout, la data-viz permet sur les phases finales de restituer, expliquer et mettre en valeur le travail de l'Analyst ou du Data Scientist. Cela participe à la retransmission de l'information.

  - Là, les outils accessibles facilement sont plutôt nombreux. En plus des solutions logicielles (citons seulement Excel et Tableau), de nombreux sites en lignes permettent des créations en quelques minutes. Un exemple parmi tant d'autres : RAW.

  - Par contre, la création de visualisations animées est plus compliquée. Il faudra probablement faire appel à du JavaScript et à la librairie D3.js.

## OUTILS INCONTOURNABLES DE LA DATA SCIENCE AUJOURD'HUI

- Le langage R spécialisé dans l'analyse statistique

- Le langage Python et ses librairies associées (Pandas, Scikit-learn...)

- Pour travailler sur de gros volumes de données, Hadoop est devenu un standard. Il est indispensable de maitriser le framework dans un cadre professionnel. Avec Hive, il est possible de manipuler des données sur Hadoop via des requêtes SQL.

# LE DATA SCIENTIST

- 2 profils se distinguent quand on parle de Data Science :
  Le Data Scientist qui créé des modèles et le Data Analyst qui les exploite.

- Un Data Scientist a pour objectif de produire des méthodes de tri et d'analyse de données de masse et de sources plus ou moins complexes ou disjointes de données, afin d'en extraire des informations utiles ou potentiellement utiles.

- Les activités d'un Data Scientist :
  - La fouille de données
  - Le nettoyage/formatage des données (data wrangling)
  - Le traitement des données
  - Traitements classiques (fonctions mathématiques)
  - Traitements d'apprentissage automatique
  - Visualisation de données
  - L'exploitation des résultats
  - Tableaux de bords et outils d'aides à la décisions (pouvant être notamment intégrés sur des sites web)
  - Publications de résultats de recherche (interne à l'entreprise, ou publics)
  - Le data scientist est donc souvent appelé à manipuler les statistiques, le traitement du signal. Il s'intéresse donc à la classification, au nettoyage, à l'exploration et à l'analyse de bases de données plus ou moins interopérables.

## PODCAST

Le portrait de Yannis,
consultant Data Science chez WEENEO

*Sources* :
Courrierinternational.com,, Empirik.fr, , Jereze.com, , Python.org, , Wikipedia, Jereze.com