insy2s

Découverte des métiers de la Data

Le parcours

Découvrir différents métiers et leurs composantes au travers des présentations et de la pratique personnelle

Découverte des métiers de la Data

Sommaire :

- L’évolution du stockage de données
- L’évolution de la collecte de données
- Le big data
- Les différents métiers de la Data
- La data Science

Evolution du stockage de données

- 1956 :

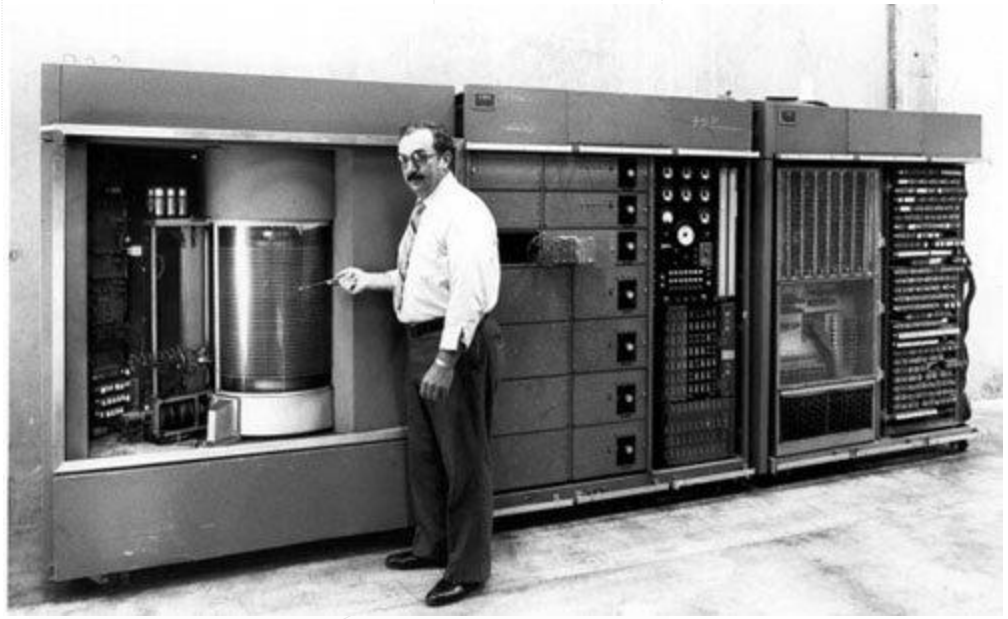
1er DISQUE DUR :

IBM RAMAC 305

Capacité : 5 Mo

Poids : environ 1 Tonne

Tarif : 50 000$ soit 10 000 $/Mo



Evolution du stockage de données

- Début des années 80 :

Support fixe :

DISQUE DUR

Capacité : 5 Mo

Format : 5,25 pouces

Tarif : 1500$ soit 300 $/Mo

Support Amovible :

DISQUETTE 3,5 pouces

Capacité : 1,44 Mo

Format adapté à une poche de chemise

Inventé par Sony, démocratisé par Apple sur ses Macintosh.

Pour la plupart des ordinateurs c'est le seul outil de stockage.

Evolution du stockage de données

- Fin des années 80 :

Support fixe :

DISQUE DUR

Capacité max : 1 Go

À cette époque, la capacité moyenne des disques durs est de 10 Mo.

Support Amovible :

CD

Capacité : 500 à 700 Mo

Inventé par Sony et Philips.

D'abord utilisé pour la musique, il servira plus tard à stocker des données via la démocratisation des lecteurs-graveurs.

Evolution du stockage de données

- Fin des Années 90 :

Support fixe :

DISQUE DUR

Capacité max : 25 Go en 1998

Le standard pour les PC de bureau est de 2 Go

Support Amovible :

DVD

Capacité : 4,7 Go

Inventé par Sony et Philips

Evolution du stockage de données

- Années 2000 :

Support fixe :

DISQUE DUR

Capacité max : 500 Go (2005), 3 To (2010)

Vers 2002, les disques durs de 40 Go sont courants pour des PC de bureau.

Support Amovible :

CompactFlash

Capacité : 1 Mo au début puis jusqu'à 512 Go

Inventé par SanDisk et Toshiba
Utilisé pour les téléphones, appareils photo, Caméscope, consoles de jeux et lecteurs MP3

En 1991, 1 Go de mémoire Flash coûtait 45 000 $

Evolution du stockage de données

Années 2010 :

Support fixe :

DISQUE DUR

Capacité max : 3 To (2010), 10 To (2015)

En 2010, le standard pour les PC de bureau est de 1 To (à partir de 0,1 €/Go) et de 500 Go pour les PC portables.

Support Amovible :

Mémoire Flash Nand, Carte SD & Clé USB

Capacité : 32Go à 1 To aujourd'hui

Evolution du stockage de données

- Depuis 2015 :

Support fixe :

DISQUE DUR SSD

Jusqu'à 30 To (Samsung PM1643)

250 Go à 2 To pour les modèles les plus courants en 2019

Le Cloud :

1.000 milliards d'octets (1 To)

pour quelques euros par mois

Démocratisé par Amazon qui souhaitait rentabiliser en dehors des périodes de forte affluence, ses immenses serveurs informatiques qui étaient sous-utilisés.

Evolution du stockage de données

- Depuis 2015 :

Support fixe :

DISQUE DUR SSD

Jusqu'à 30 To (Samsung PM1643)

250 Go à 2 To pour les modèles les plus courants en 2019

Le Cloud :

1.000 milliards d'octets (1 To)

pour quelques euros par mois

Démocratisé par Amazon qui souhaitait rentabiliser en dehors des périodes de forte affluence, ses immenses serveurs informatiques qui étaient sous-utilisés.

Le déluge de données

- Aujourd'hui :

Tous les appareils et leurs utilisateurs génèrent des données en masse…

Le déluge de données

- Depuis 2015 :

3,8 Zetta Octets de données en 2015

L' équivalent d'une pile de Blu-Ray qui ferait 7 fois le tour de la terre !

1 Zo
= mille milliards de milliards d'octets
1 000 000 000 000 000 000 000 octets

Le déluge de données

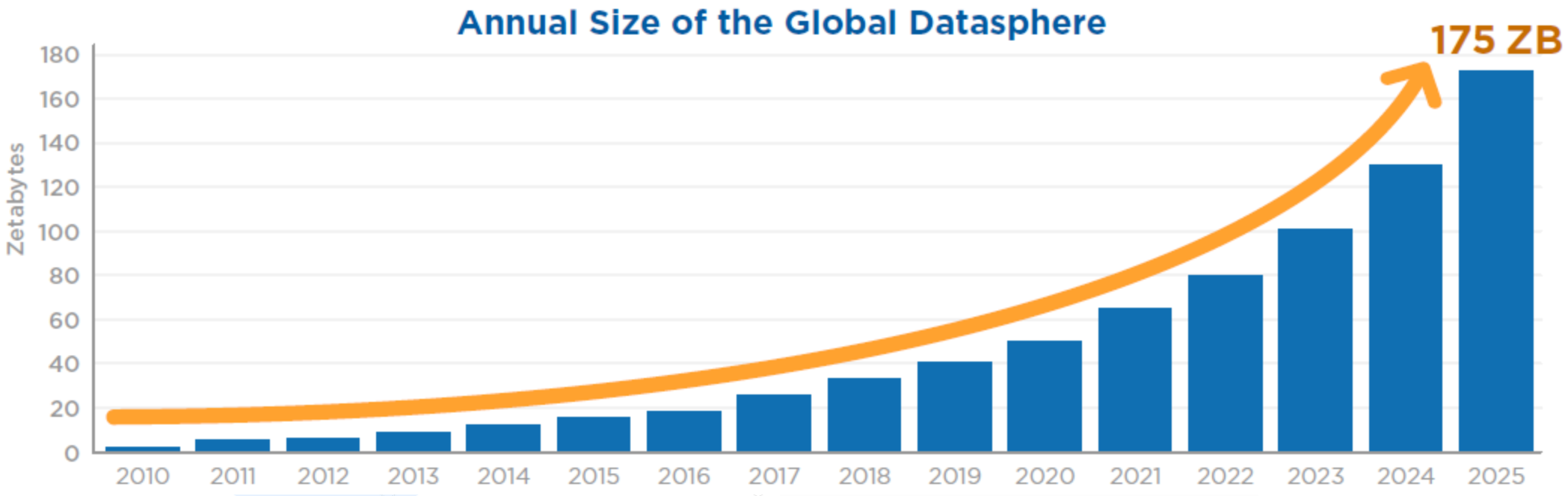
- Depuis 2015 :

  - **60% de croissance par an** des volumes d'informations
  - **1 Boeing produit 20 To / heure de données**
  - **250 milliards d'e-mail envoyés chaque jour** (80% de spam)
  - **72h de vidéo déposées chaque minute** sur Youtube

Le déluge de données

- Les prévisions d'ici 2025 :

Figure 1 – Annual Size of the Global Datasphere



Le déluge de données

- Le problème n'est plus le stockage

80% de l'information est **non-structurée**

95% de l'information est **non-exploitée**

L’ère du BIG DATA

L'objectif principal du Big Data est de réussir à faire apparaître des enseignements et des connexions entre de gros volumes de données de nature hétérogène qui seraient impossible à obtenir avec les méthodes classiques d'analyse des données.

Métiers du BIG DATA

- Data Analyst
- Data Scientist
- Data Engineer

Métiers du BIG DATA

Data Analyst :

Le Data Analyst a pour mission d'**exploiter et interpréter les données pour en dégager des observations business utiles**. Ainsi, les rapports fournis permettent d'orienter les prises de décision du Management et améliorer les performances et les stratégies Marketing.

Le Data Analyst créé, administre et modélise la base de données et s'assure d'une mise à jour régulière pour en faciliter l'exploitation par les équipes métiers. Il est souvent amené à travailler avec plusieurs équipes et les résultats fournis vont avoir un impact sur la croissance de l'entreprise. Après analyse des données qu'il mettra ensuite à la disposition des équipes **Marketing, Finance, Communication** et de la Direction.

En savoir plus :

https://www.michaelpage.fr/advice/metiers/digital-marketing-communication/fiche-métier-data-analyst

Métiers du BIG DATA

Data Scientist

Le data scientist développe des algorithmes d'apprentissage automatique selon les besoins des équipes métiers. Ses compétences en statistiques lui permettent de construire des modèles de machine learning et ses connaissances en informatique l'aident à anticiper leur mise en production. En amont de ces deux missions, il/elle est également en charge de structurer et d'analyser les données qu'il/elle utilise.

Plus d'info :

https://www.apec.fr/tous-nos-metiers/informatique/data-scientist.html

https://www.michaelpage.fr/advice/metiers/systemes-dinformation/fiche-métier-data-scientist

Métiers du BIG DATA

Data Engineer
Le Data Engineer est un ingénieur. Son rôle est donc **de concevoir et de fabriquer.** Ces missions sont :

- collecter des données brutes en provenance de multiples sources. Puis conçoit et gère les bases de données de l'organisation.

- **- mettre en place un pipeline** permettant d'automatiser les différentes étapes de l'acquisition de données, de l'extraction au stockage. Dans un second temps, le Data Engineer " nettoie " les données et les transforme. L'objectif est qu'elles soient prêtes à être analysées par les Data Scientists.

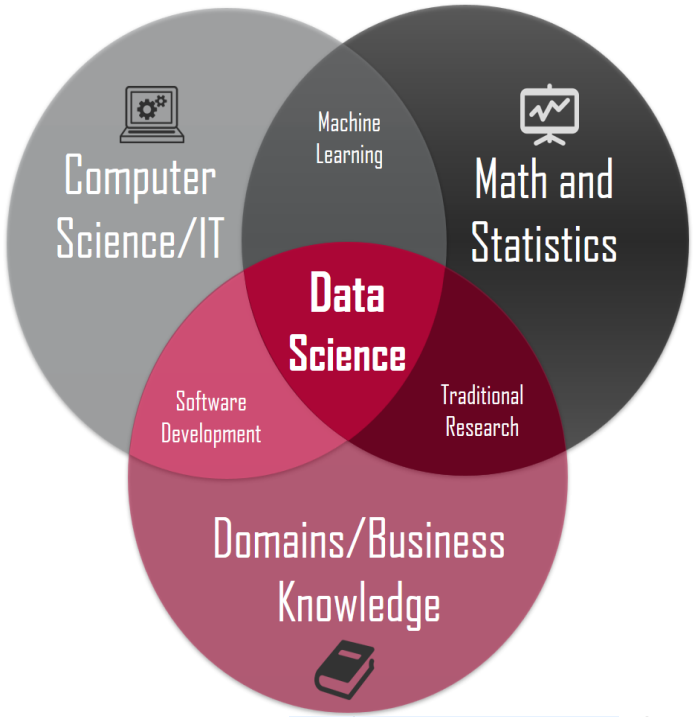
Le Data Engineer doit aussi créer des outils et algorithmes permettant aux Data Scientists, et éventuellement à d'autres employés ou cadres de l'organisation, **d'accéder facilement aux données** dont ils ont besoin.

Plus d'info : https://datascientest.com/data-engineer-tout-savoir
https://www.data-transitionnumerique.com/fiche-metier-data-engineer/

Data Science : Qu’est-ce que c’est ?

La définition de la data science peut être résumée à l’extraction et transformation de données brutes et inorganisées en information traitée et exploitable par les entreprises.

Data Science



- La data science est une discipline scientifique visant à établir des modèles statistiques et mathématiques afin d'analyser et extraire de gros volumes de données en vue de leur exploitation
- Multidisciplinaire, la science des données fait appel notamment aux mathématiques, aux statistiques, à l'analyse de données, à la programmation informatique…

Science des données

- La data science a pour but d'extraire des données exploitables à partir d'un gros volume de données brutes
- L'objectif principal est de dégager à partir de ces données, des tendances et des prédictions
- Avec l'essor du big data, la data science se développe de plus en plus, dans tous les secteurs d'activités

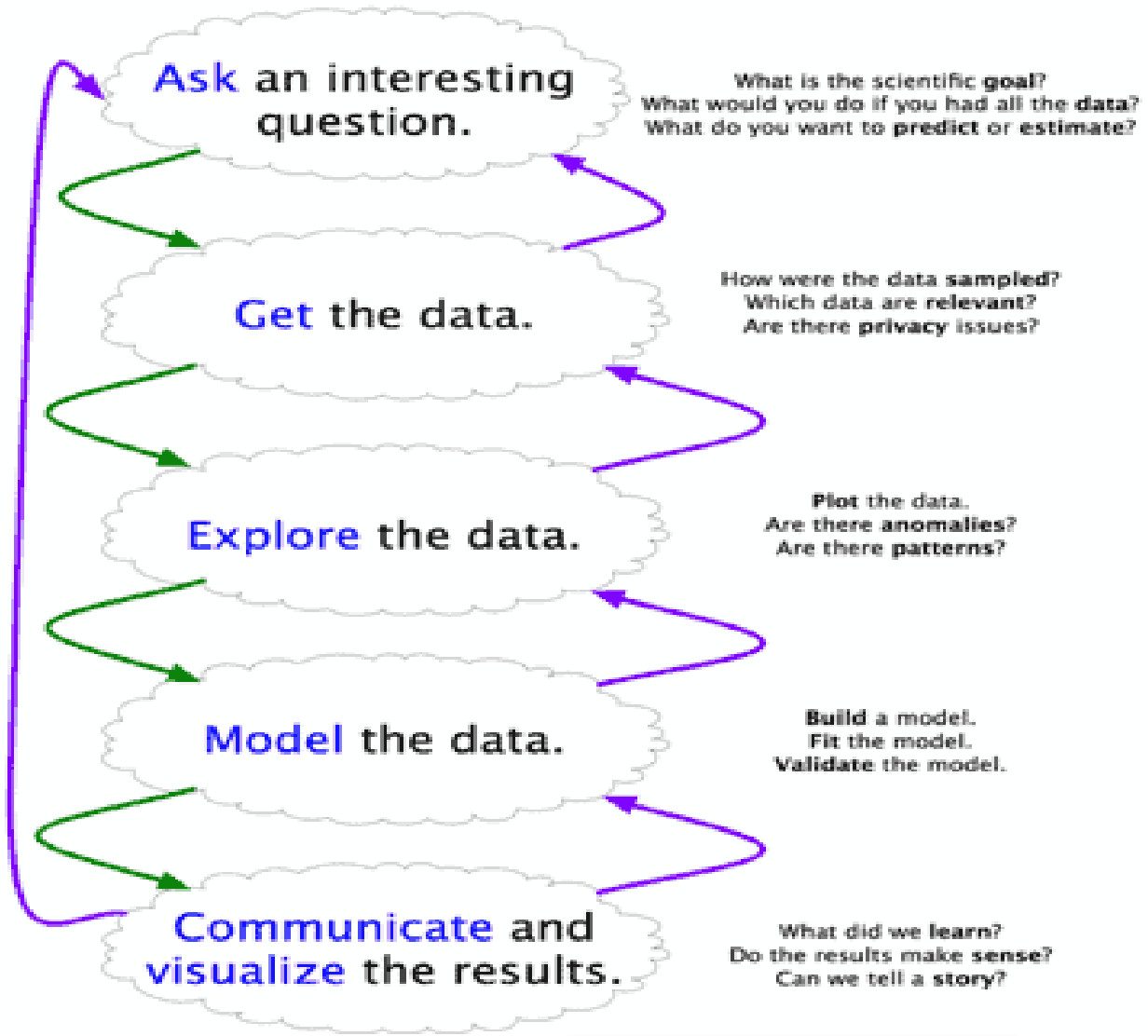
Science des données

- Dans la vente, la data science va être utilisée comme un outil d'aide à la décision afin de dégager les futures attentes des consommateurs par exemple. Sa finalité sera essentiellement marketing
- Dans l'environnement, la data science va permettre de modéliser des phénomènes climatiques et d'établir des projections d'impacts liées à ces phénomènes
- Dans les transports, la data science va être utilisée pour adapter la fréquence des transports au flux de voyageurs à certaines heures
- Dans la banque et l'assurance, la data science va être utilisée pour analyser la perte clients par exemple ou pour proposer des produits en fonction des besoins clients

Science des données

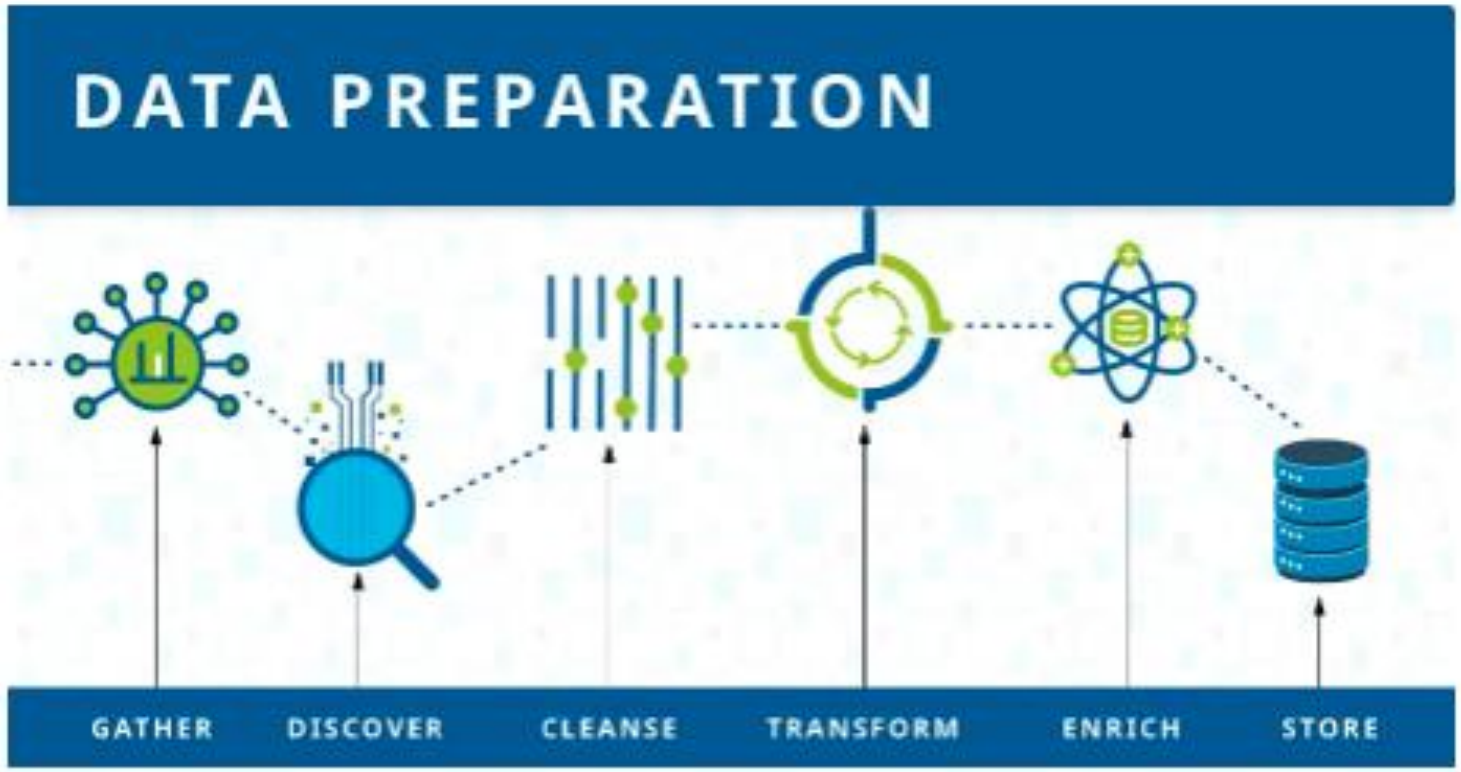
- Sur les réseaux sociaux, la data science va permettre d'organiser des campagnes publicitaires ciblées, de proposer des recommandations de vidéos ou films à regarder
- Dans l'industrie, la data science va être utilisée pour générer de l'innovation, pour faire de la maintenance prédictive, pour analyser des images et vidéos, pour faire du text mining... Cela va leur permettre d'améliorer des produits déjà existants ou de proposer de nouveaux produits adaptés aux attentes des clients
- Dans la logistique, les entreprises vont utiliser la data science pour améliorer leurs délais de livraison, réduire les coûts, analyser le trafic ou les conditions météorologiques
- Dans l'automobile, la data science est de plus en plus utilisée grâce au développement des véhicules autonomes. Le déploiement de l'intelligence artificielle va permettre d'analyser une grosse quantité de données permettant de rendre autonome une voiture au niveau du freinage d'urgence par exemple ou encore du stationnement automatique

Science des données



Science des données

- Préparation des données – Workflow général



Préparation des données

1. Collecte de données

Le processus de préparation des données commence par la recherche des données les plus utiles. Ces données peuvent provenir d'un catalogue existant ou être collectées à partir des sites web (des techniques de web scrapping peuvent être utilisées).

Science des données

Préparation des données

2. Découvrir et évaluer les données

Lorsque les données ont été collectées, il est important de découvrir les différents datasets. Cette étape permet de mieux connaître les données et de déterminer les traitements à leur appliquer avant qu'elles deviennent exploitables dans un contexte particulier.

Science des données

Préparation des données

3. Nettoyer et valider les données

Le nettoyage des données est l'étape la plus longue du processus de préparation des données. Lors du nettoyage, les tâches importantes sont notamment les suivantes :

- ✓ Supprimer les données superflues et les valeurs aberrantes
- ✓ Imputer les valeurs manquantes
- ✓ Adapter les données à une structure standard
- ✓ Masquer les données privées ou sensibles

Lorsque les données ont été nettoyées, elles doivent être validées, à savoir déterminer si des erreurs se sont produites dans le processus de préparation des données jusqu'à ce point (il peut arriver qu'une erreur apparaisse pendant cette étape, et il est alors nécessaire de la corriger avant de poursuivre).

Science des données

Préparation des données

4. Transformer et enrichir les données

Transformer les données consiste à mettre à jour les données de manière à obtenir un résultat clairement défini ou à rendre les données plus faciles à comprendre.

Enrichir les données consiste à ajouter ou agréger des données de manière à dégager des connaissances approfondies.

Science des données

Préparation des données

5. Stocker les données

Lorsque la préparation des données est terminée, celles-ci doivent être stockées avant leur traitement et analyse.

Science des données

Analyse et traitement des données

1. DataViz

Les visualisations aident à percevoir des motifs. Même quand le volume des données est très important, des tendances peuvent être perçues de façon rapide et simple.

La visualisation facilite la transmission des informations de façon universelle et facilite le partage d'idées avec les autres

C'est un moyen rapide d'obtenir des informations à travers l'exploration visuelle, des rapports fiables et un partage d'informations aisé

Science des données

Analyse et traitement des données

2. Statistiques

Dans un projet data science, la statistique permet :

de trouver les relations qui peuvent exister entre les caractéristiques des données (corrélation, dépendance, etc.)

pour normaliser et la mise à l'échelle des données

pour identifier la nature de distribution des données

…

Science des données

Etablir des modèles d'apprentissage automatique

L'objectif des modèles de machine learning est de comprendre le comportement du phénomène que l'on souhaite étudier (marché immobilier, analyse de sentiment, reconnaissance d'image, etc) en se basant sur des données générées par ce phénomène et le répliquer sur de nouvelles observations.