Université Constantine 2
جامعة قسنطينة 2

Technologies for Data Science and Artificial Intelligence

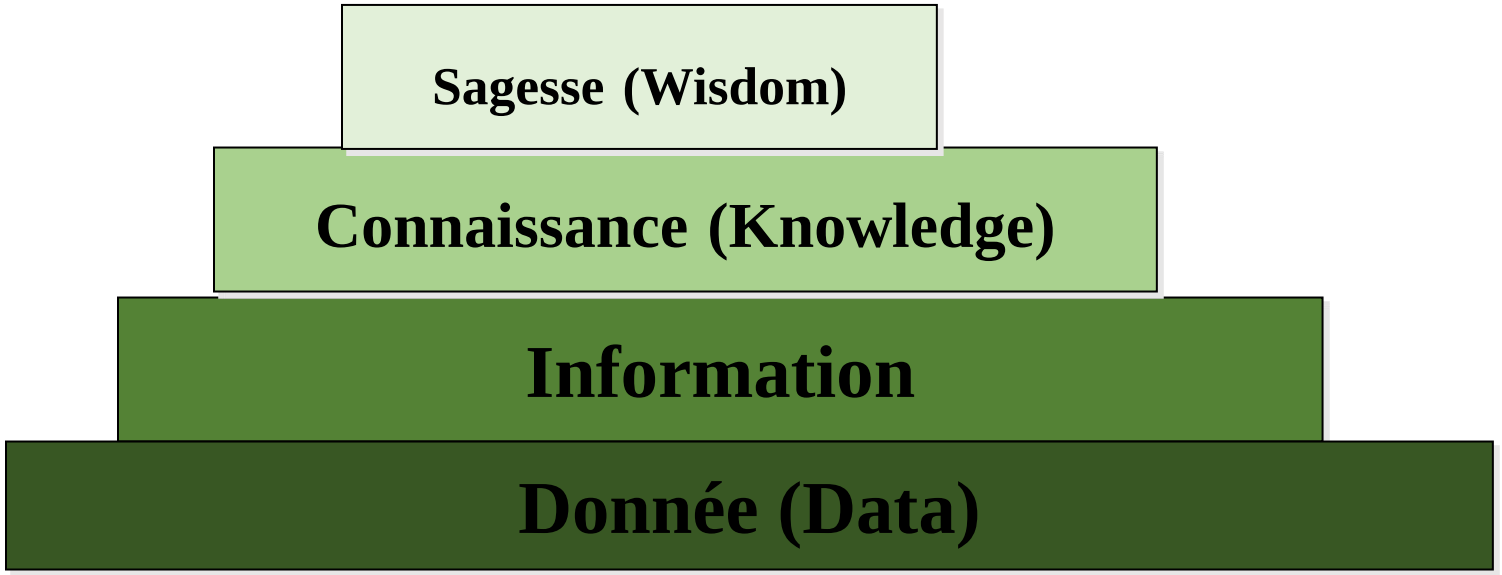
Chapitre 1: Introduction à la science de données

Dr. S.ZERABI
Faculté des NTIC
Soumeya.zerabi@univ-constantine2.dz

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
Nouvelles technologies	IFA	Master 1	SDIA

Introduction



Pyramide de la sagesse (D.I.K.W)

Définitions

- Donnée (Data)

«le résultat d'une observation faite sur une population ou sur un échantillon» *(Dodge, 2004).*

Ex. « Benali », « Mohamed », 25

- Information (Information)

Donnée + sens

Ex. (Nom: Benali), (Prénom: Mohamed), (Salaire: 25).

- Connaissance (Knowledge)

Information + règles

Ex. Benali Mohamed a un âge supérieur à 18 ans.

- Sagesse (Wisdom)

Connaissance + expertise

Ex. Benali Mohamed est majeur.

- jeu de données (Dataset)

C'est une collection de données.

Définitions

➢ Donnée structurée

Une donnée structurée décrit une propriété (e.g., nom, adresse, Numéro de carte de crédit) d'une entité (e.g., client, produit) selon un modèle fixé.

Exemple

- Données stockées dans des feuilles (e.g., Fichier Excel).
- Enregistrements stockés dans les tables d'une base de données relationnelle.

Définitions

➢ Donnée semi-structurée

Une donnée semi-structurée possède une structure où les entités et leurs propriétés peuvent être facilement distinguées, MAIS l'organisation de la structure n'est pas rigoureuse comme celle de la table de la base de données.

Exemple

documents XML, JSON, HTML.

Définitions

- Donnée non structurée

Une donnée non structurée décrit une entité qui ne possède pas une structure à cause de ses propriétés qui ne peuvent pas être distinguées les unes des autres.

Exemple

Un fichier txt.

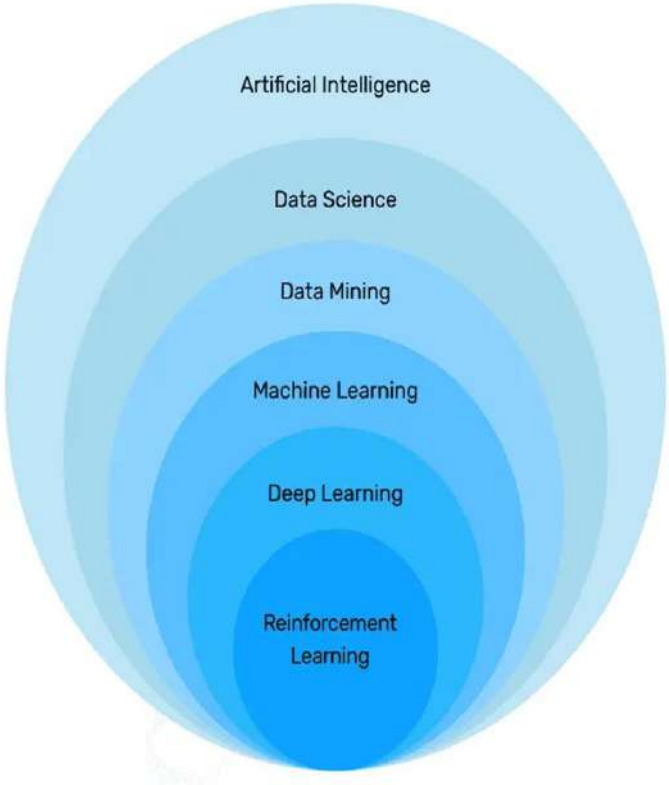
Science de données

La « **Science des Données** » est l'ensemble des techniques et outils permettant de **collecter**, **nettoyer**, **organiser**, **explorer**, **modéliser**, **visualiser** les données.

L'objectif est d'en extraire des **informations pertinentes** permettant de prendre les bonnes décisions (estimation, prévision, classification, …).

La science de données est un domaine multidisciplinaire, elle fait intervenir les disciplines suivantes:

- ✓ **Intelligence artificielle.**
- ✓ **Big Data.**
- ✓ **Les mathématiques** (Statistiques, Algèbre linéaire, probabilité, analyse, etc).
- ✓ **Programmation** (Python, R, etc).



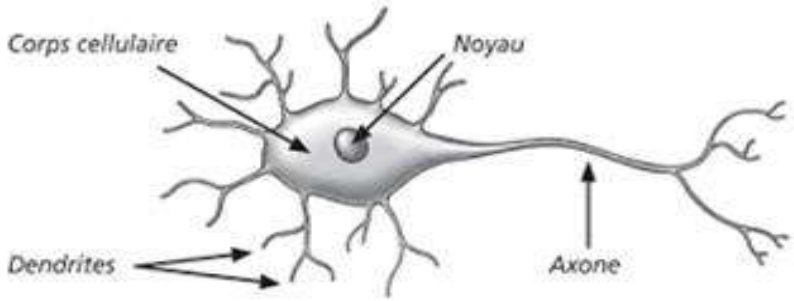
Scientifique de données (Data Scientist)

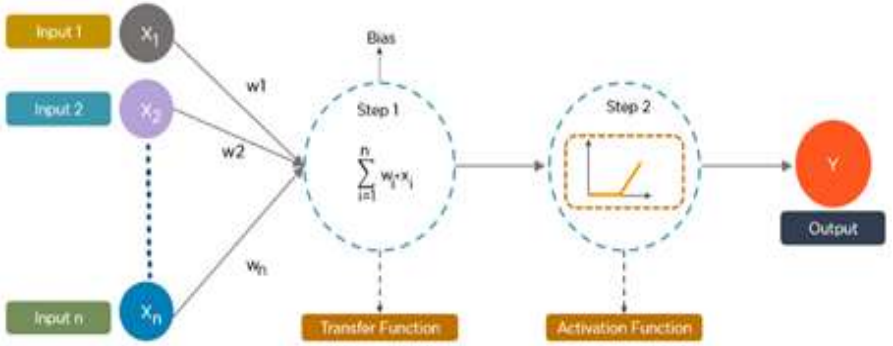
Répondre aux questions suivantes:

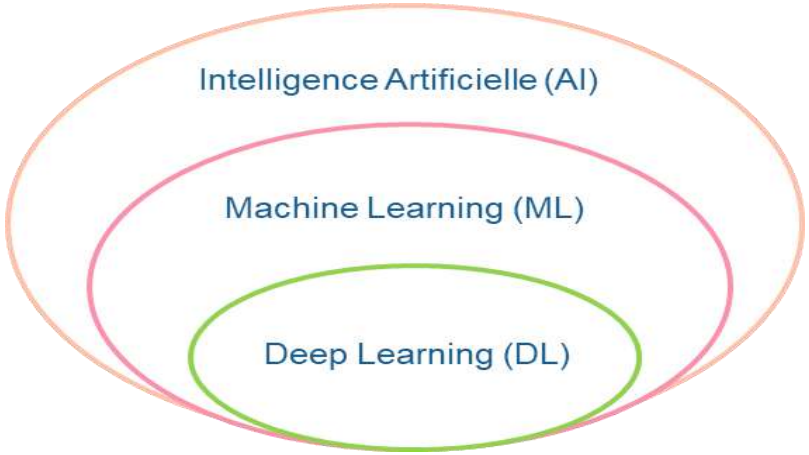
- ✓ Que s'est-il passé?
- ✓ Pourquoi cela s'est-il passé?
- ✓ Que va-t-il se passer?
- ✓ Que peut-on faire avec ces résultats?

Intelligence Artificielle

- L'IA est l'ensemble de techniques et théories qui cherchent à développer des modèles capables de simuler le comportement humain afin d'effectuer des tâches complexes.
- L'IA a été inspiré du neurone biologique.
- Premier Neurone artificiel en 1943.

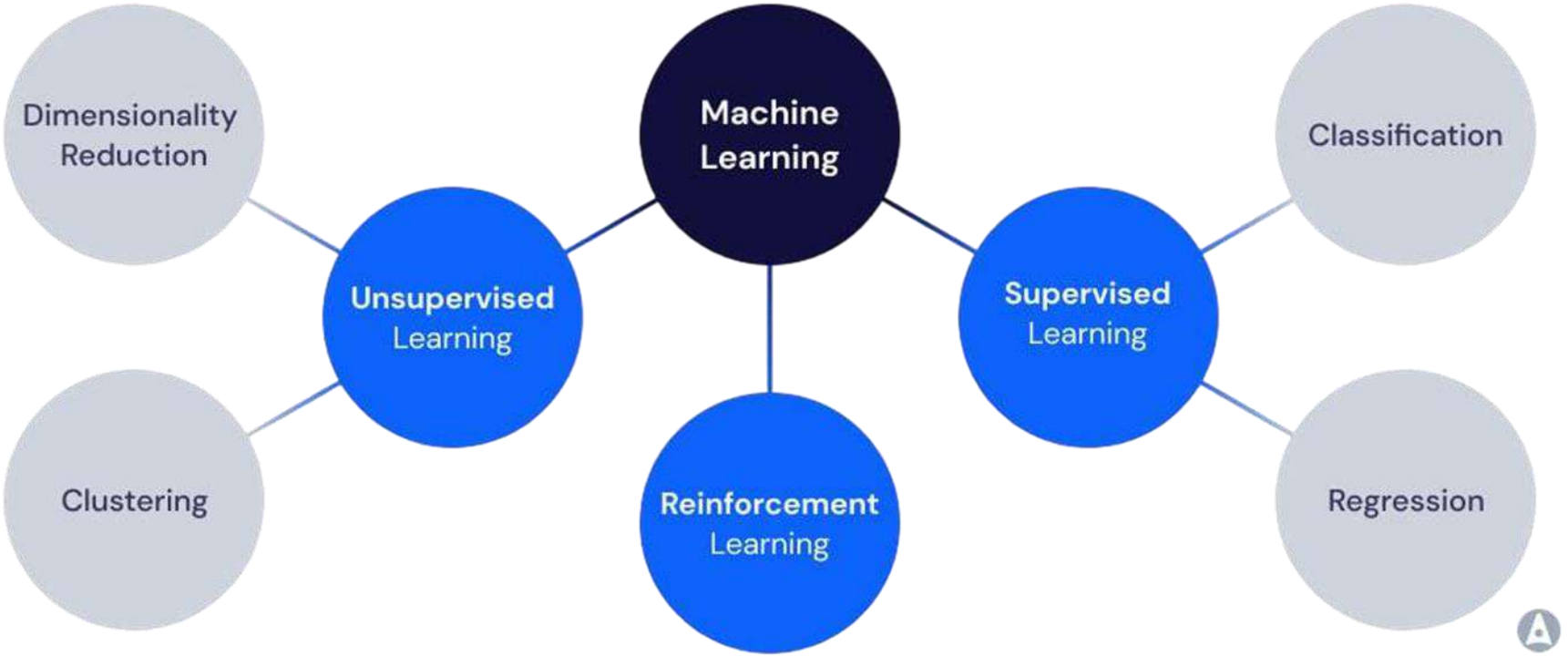






Machine learning

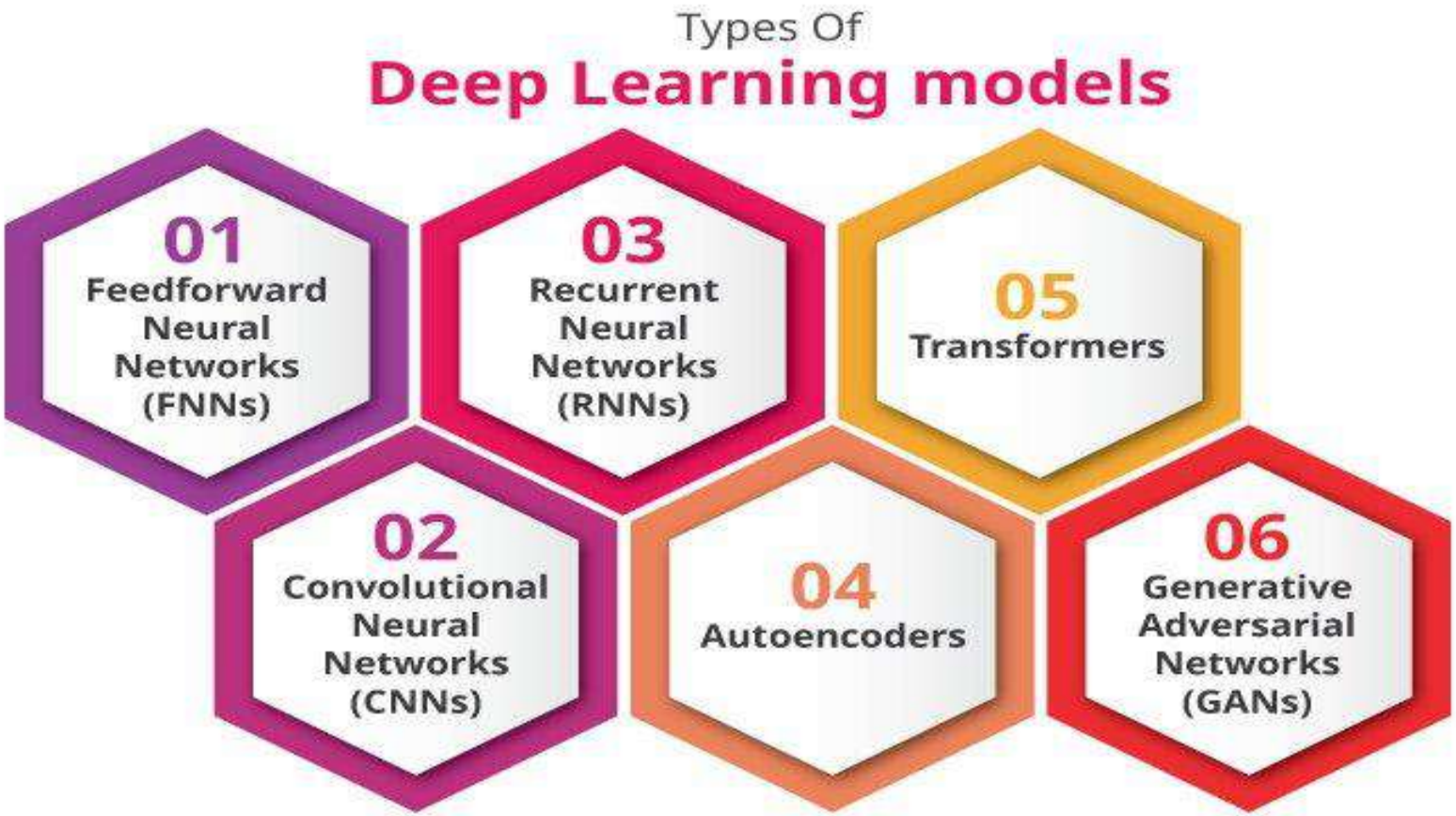
Machine learning (ou Apprentissage automatique) est un champ d'étude qui fournit aux ordinateurs la capacité d'apprendre sans avoir été programmés explicitement. *(Arthur Samuel, 1959) .*



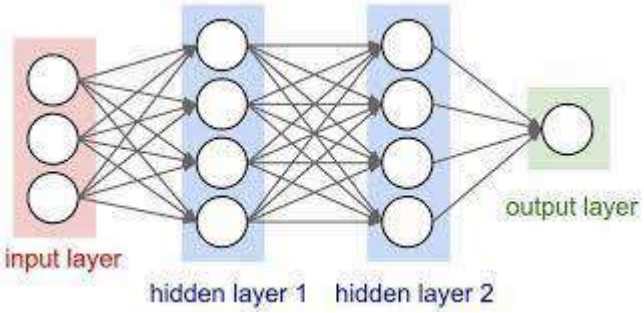
Deep learning

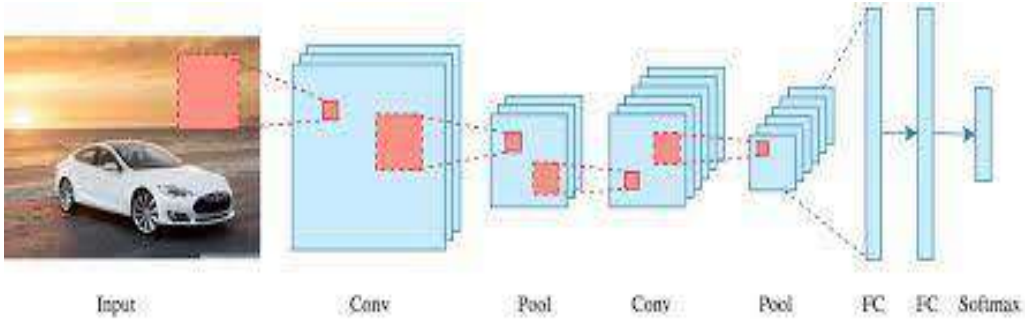
Deep learning (ou Apprentissage profond) est un sous domaine du machine learning, basé sur les réseaux de neurones artificielles (*ANN*).

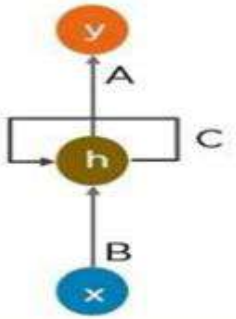
Il est appelé "profond" car il utilise des réseaux de neurones artificiels avec plusieurs couches de traitement des données.

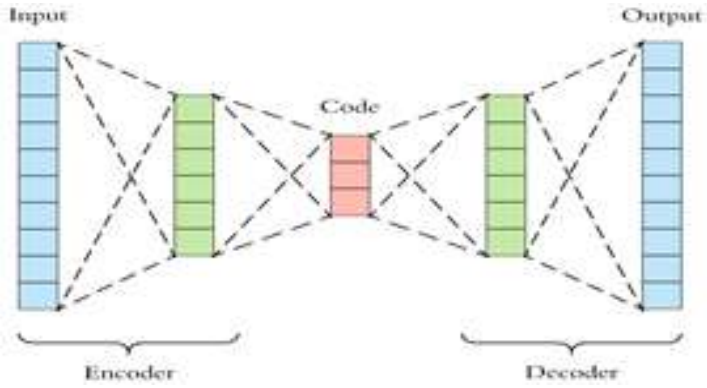

Types Of
Deep Learning models
01
Feedforward Neural Networks (FNNs)
03
Recurrent Neural Networks (RNNs)
05
Transformers
02
Convolutional Neural Networks (CNNs)
04
Autoencoders
06
Generative Adversarial Networks (GANs)

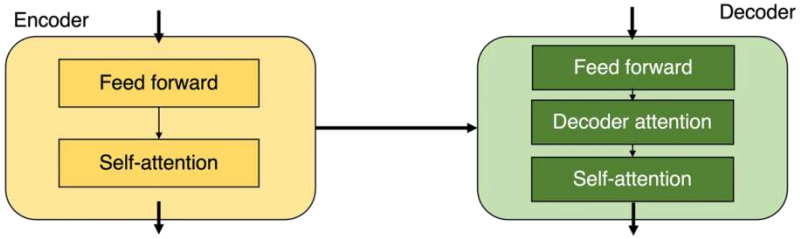
Intelligence artificielle

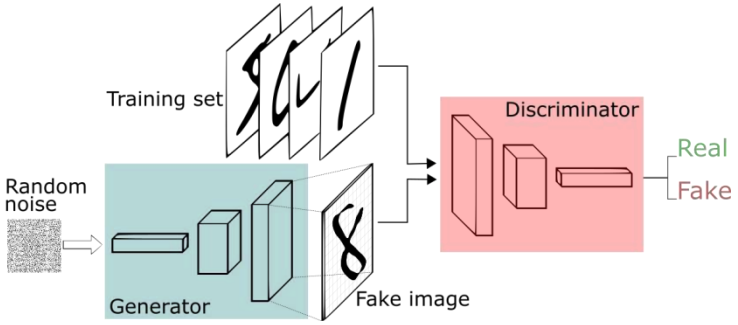












Introduction au Big Data

Emails

Objets connectés

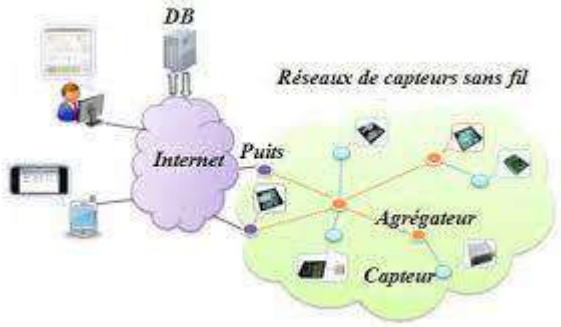
Smart phones et tablettes

E-commerce

Diverses sources de données

Instruments scientifiques

Réseaux sociaux et medias



Réseaux de Capteurs

Caméras de surveillance

Introduction au Big Data

- L'humanité génère de plus en plus de données
- Entre 2011 et 2013, le volume mondial de données a été multiplié par 9.

Préfixes multiplicatifs

signe	préfixe	facteur	exemple représentatif
k	kilo	10^3	une page de texte
M	méga	10^6	vitesse de transfert par seconde
G	giga	10^9	DVD, clé USB
T	téra	10^{12}	disque dur
P	**péta**	10^{15}	
E	**exa**	10^{18}	FaceBook, Amazon
Z	**zetta**	10^{21}	internet tout entier depuis 2010

Ce n'est pas tout……

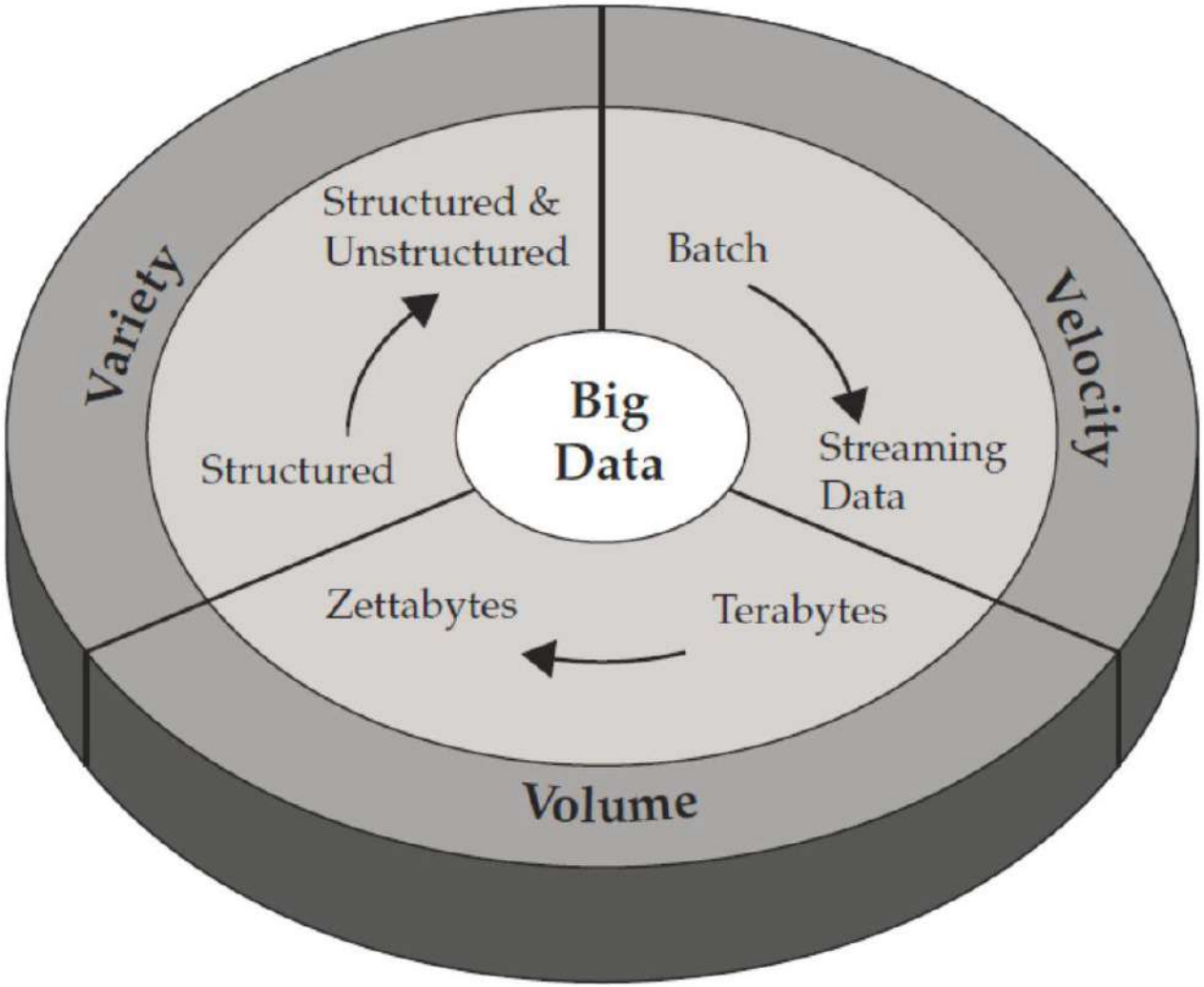
Définition du Big Data

Def 1. Le Big Data est une nouvelle génération de technologies et d'architectures conçues pour extraire de la valeur à partir d'un volume considérable de données très variées permettant leur capture et leur exploration à grande vitesse *(IDC)*.

Def 2. Le Big Data désigne des ensembles de données qui deviennent tellement volumineux qu'il devient difficile voire impossible à les manipuler avec des outils classiques de gestion de base de données ou de gestion de l'information.

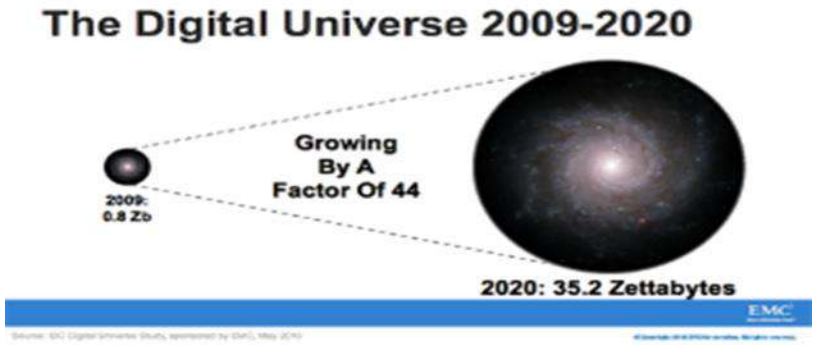
Caractéristiques des Big Data (les 3Vs)

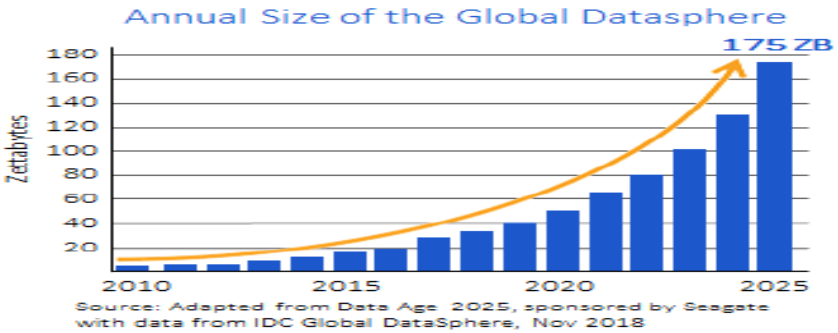
Les 3Vs des Big Data



Caractéristiques des Big Data (les 3Vs)

- Volume: La taille du dataset.





- Vélocité: La nécessité de traitement des données à leur arrivée.

- Variété: La nature hétérogène des données (structurées, semi-structurées, non structurées).

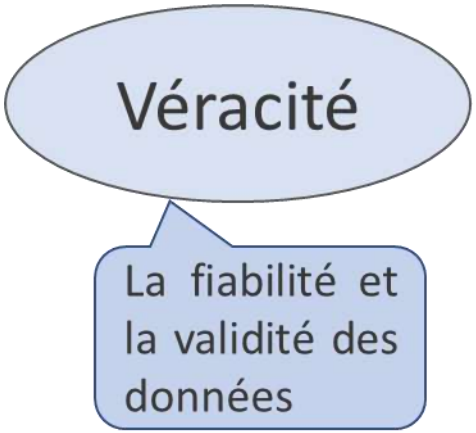
Exemple

Système d'analyse des Sentiments qui traite les tweets.

- Objectif : sentiment positif/négatif/neutre?
- Volume: Millions de tweets.
- Vélocité: flot constant de données (7,500 tweets/second).
- Variété: Textes, images et liens pages Web.

Caractéristiques des Big Data (les 3Vs)

Plus 2 autres Vs



Défis des Big Data

- Le Big Data ne peut pas être manipulé au niveau d'une seule machine, lorsque la complexité des applications dépassent la capacité de calculs.
- **Solution:** distribuer les calculs sur un cluster constitué de plusieurs machines
- Il y a deux défis principaux du Big Data:

Traitement + Stockage.

Défis des Big Data

- Traitement

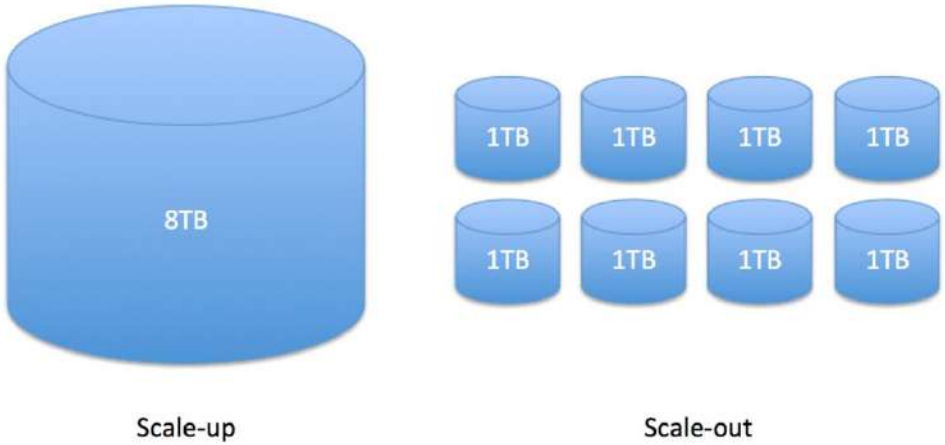
• Parallélisation du calcul sur les machines.

• Frameworks de traitement parallèle et distribué.

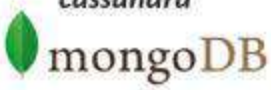
(e.g., Hadoop MapReduce, Spark, Storm) .

- Stockage

Bases de données (relationnelles/NoSQL) distribuées.



Outils de la science de données



Domaines d'application de la science de données

Reconnaissance vocale

Détection d'anomalies

Transport

- Optimisation des déplacements
- Contrôle de la circulation

TIC

Cyber sécurité

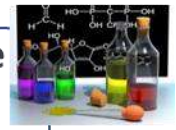
Marketing et vente en détail

- Gestion des promotions
- Connaissance des clients
- Optimisation des approvisionnements

Sécurité

- Détection de menaces et criminologie
- Gestion de catastrophes naturelles
- Fraude (détection/prévention)

Science et recherche



- Physique
- Chimie
- Environnement
- Science de la vie

Santé

- Alerte précoce d'épidémie
- Médecine à distance
- Génomique

Reconnaissance faciale

IOT

- 26.6 milliards d'objets connectés en 2020 ~ 41.6 milliards en 2025

Média et réseaux sociaux

Systèmes de recommandations