RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY

Faculty of Technology
Common core of computer engineer

COURS: FONDEMENT DES SCIENCES DE DONNÉES

Préparé par :
Bilal Dendani

bilal.dendani@univ-annaba.dz

SPÉCIALITÉ: SCIENCES DE DONNÉES

علم البيانات

2024-2025

Objectifs :

- Fournir une base solide en sciences des données, en mettant l'accent sur *les outils mathématiques essentiels*, notamment *l'algèbre linéaire*.

- Préparer les étudiants à aborder des cours plus avancés en analyse des données et en machine learning en leur donnant les théoriques et pratiques nécessaires.

- Prérequis : notions de base en mathématiques, statistiques et programmation.

Contenu du cours

Chapitre 1. Introduction aux sciences de données

- Qu'est-ce qu'une science de données?
- Origines et enjeux de la science des données
- Facettes et types de données
- Comment fonctionne la science des données?
- Cas d'usage et domaines d'application
- L'écosystème du big data et la science des données

Chapitre 2. Le processus de science des données

- Rôles et responsabilités dans un projet de science des données
- Présentation du cycle de vie d'un projet de science des données
- Étape 1 : Définir les objectifs de recherche et créer une charte de projet
- Étape 2 : Récupération des données
- Étape 3 : Nettoyer, intégrer et transformer les données
- Étape 4 : Analyse exploratoire des données
- Étape 5 : Construire les modèles
- Étape 6 : Présentation des résultats et création d'applications au-dessus d'eux

Contenu du cours

Chapitre 3 : Outils et technologies utilisés en Data Science

- Les outils de stockage de données
- Les outils de préparation de données
- Les outils de visualisation de données
- Les outils IDE notebooks
- Les plateformes complètes de Data science

Chapitre 4 : Principes de Base de l'Algèbre Linéaire

- Vecteurs et Espaces Vectoriels :
  - Définition et opérations sur les vecteurs.
  - Espaces vectoriels et sous-espaces.
- Matrices :
  - Définition, types de matrices, opérations sur les matrices (addition, multiplication, inversion).
  - Matrices spéciales (matrices diagonales, orthogonales, identités).
- Systèmes d'Équations Linéaires :
  - Résolution des systèmes linéaires (méthode de Gauss-Jordan, décomposition LU)

Contenu du cours

Chapitre 5 : Modèles Linéaires

- Régression Linéaire Simple :
  - Modèle de régression linéaire simple, estimation des paramètres.
  - Interprétation des résultats, erreurs et diagnostics.
- Régression Linéaire Multiple :
  - Extension aux modèles à plusieurs variables explicatives.
  - Sélection de modèles et régularisation (Ridge, Lasso).

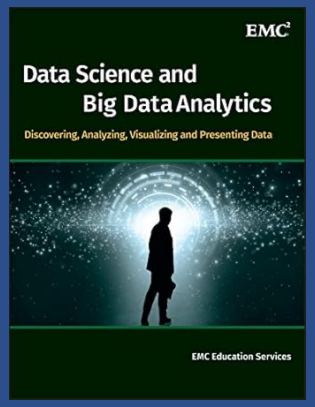
Chapitre 6 : Algèbre Linéaire Avancée

- Valeurs Propres et Vecteurs Propres :
  - Calcul et interprétation des valeurs propres et des vecteurs propres.
  - Applications en réduction de dimensionnalité (analyse en composantes principales).
- Décomposition en Valeurs Singulières (SVD) :
  - Théorie et calcul de la décomposition en valeurs singulières.
  - Applications pratiques (compression d'images, filtrage collaboratif).

Chapitre 7 : Méthodes Numériques en Sciences des Données

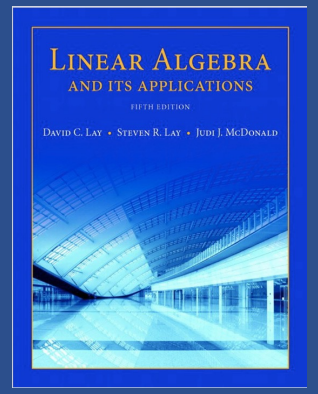
- Techniques d'Optimisation :
  - Introduction aux méthodes d'optimisation (gradient descent, Newton-Raphson).
  - Applications en apprentissage machine.
- Algorithmes Numériques :
  - Résolution numérique des systèmes d'équations.
  - Techniques de recherche de racines et d'intégration numérique.

Références

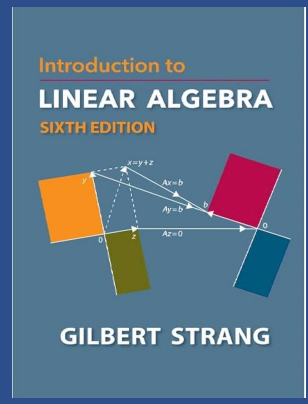


Dietrich, D., “Data science & big data analytics: discovering, analyzing, visualizing and presenting data”, Wiley, 2015.

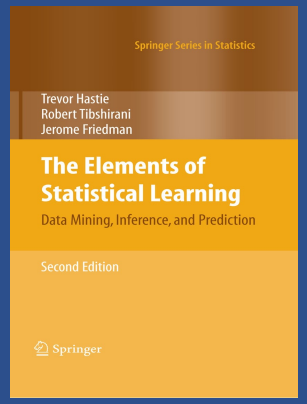
Lutz, M., Biernat, E., “Data Science: fondamentaux et études de cas: Machine Learning avec Python et R”, Editions Eyrolles, 2015.



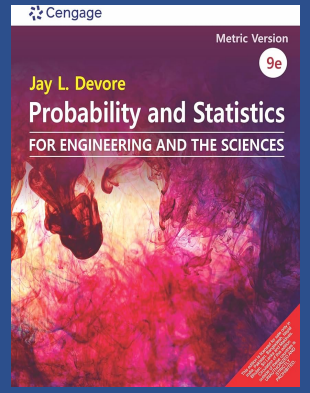
«Linear Algebra and Its Applications» by David & Steven Lay, and Judi McDonald



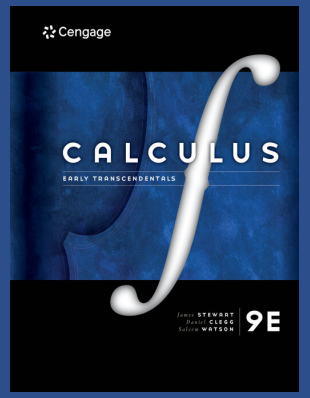
"Introduction to Linear Algebra" de Gilbert Strang.



"The Elements of Statistical Learning: Data Mining, Inference, and Prediction" de Trevor Hastie, Robert Tibshirani, et Jerome Friedman.



"Probability and Statistics for Engineering and the Sciences" de Jay L. Devore.



"Calculus: Early Transcendentals" de James Stewart.

Chapitre 1

Introduction aux sciences de donnés

Préparé par :
Dr. Bilal Dendani

جـامعة بـاجـي مختـار - عنـابـة
BADJI MOKHTAR - ANNABA UNIVERSITY

Chapitre 1 : Introduction aux sciences de données

- Qu'est-ce que les données ?
- C’est quoi la science des données ?
- Quelques définitions de domaines liés au science de données
  - Qu’est ce que le Big data ?
  - Qu’est ce que l’intelligence artificielle, l’apprentissage automatique et l’apprentissage approfondie ?
- Les origines scientifiques de la science de données
- Quelques historiques liés à l’emergence de la science de données.
- Enjeux de la science de données
- Facettes et types de données (structurés, non structurés et semi-structurés)
- Comment fonctionne la science de données ?
- Cas d'usage et domaines d'application
- Eosystème de Big data et de la science de données.

Qu'est-ce que les données ?

- Aujourd'hui les données sont omniprésent , nous sommes constamment entourés de données. Le texte que vous êtes en train de lire est une donnée. La liste des numéros de téléphone de vos amis dans votre smartphone est une donnée, tout comme l'heure affichée sur votre montre.
- Cependant, les données sont devenues beaucoup plus importantes avec la création des ordinateurs et l'émergence de l'internet.
- Les données représentent des éléments bruts ou des faits non traités, y compris des nombres et des symboles, des textes et des images.
- Lorsqu'ils sont collectés et observés sans interprétation, ces éléments restent des données simples et non organisées.

- Lorsque ces éléments sont analysés et mis en contexte, ils se transforment en quelque chose de plus significatif, en **information (d**onnée + sens).
- **Database (ensemble de données)** collection de données ou d'informations structurées, stockées et organisées.

https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README

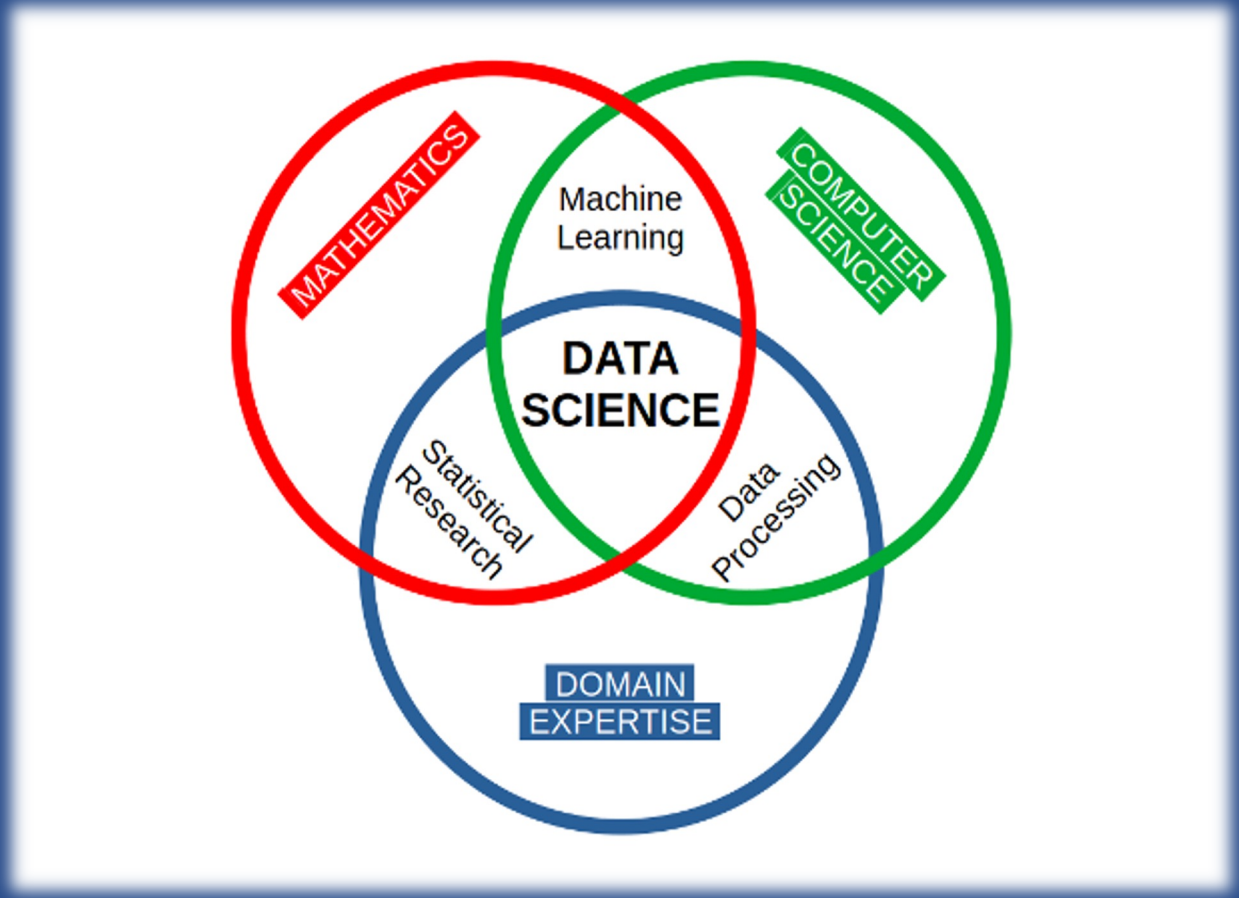
Qu'est-ce que la science des données ?

- La science des données est une discipline qui consiste à utiliser les données pour résoudre des problèmes complexes et prendre des décisions informées
- La science des données est un art qui transforme les données brutes en informations significatives (utiles).
- Cela implique la collecte, l'analyse et l'interprétation de données pour découvrir des tendances, faire des prédictions et pour la prise de décision.
- C'est le mélange de compétences en programmation, en statistiques et en expertise métier pour résoudre des problèmes complexes à partir de données.

Qu'est-ce que la science des données ?

La science des données est un domaine interdisciplinaire qui utilise les mathématiques, les statistiques, le calcul scientifique, les méthodes scientifiques, les algorithmes et les systèmes informatiques automatisés pour extraire et extrapoler des connaissances à partir de grandes quantités de données structurées et non structurées.

https://fr.wikipedia.org/wiki/Science_des_données



https://fr.wikipedia.org/wiki/Science_des_données

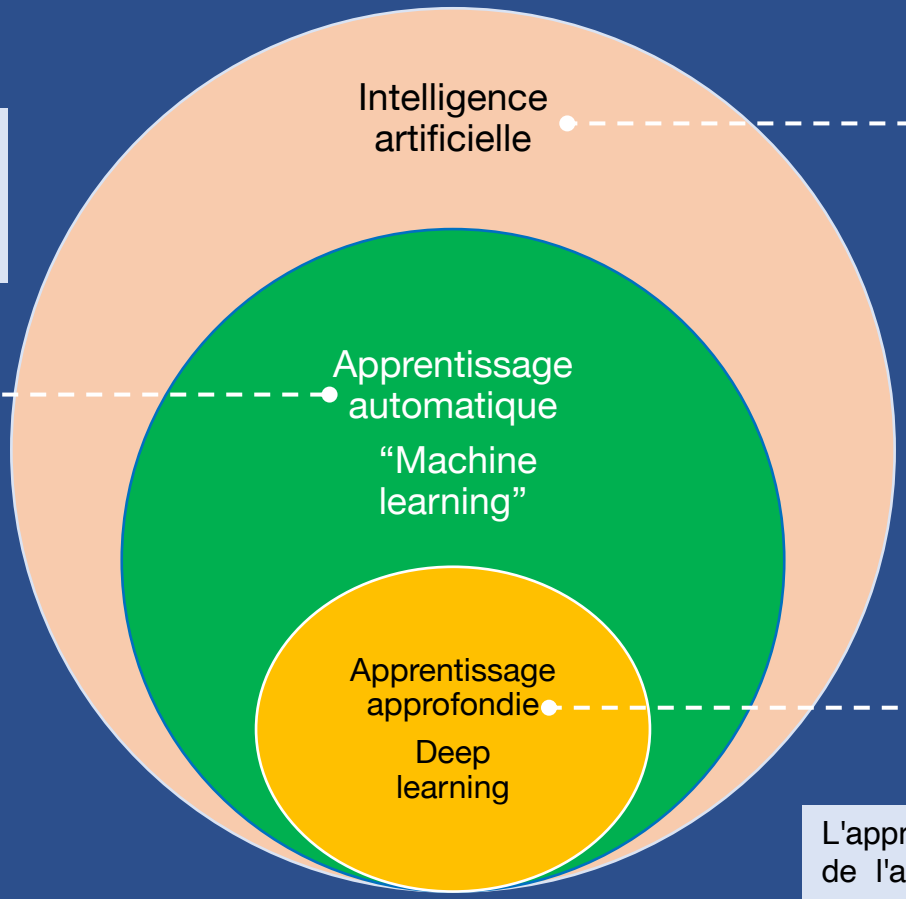
Quelques définitions de domaines liés au science de données
Qu'est ce que le Big data ?

- On parle depuis quelques années du phénomène de big data , que l'on traduit souvent par « données massives ».
- Le big data est un domaine qui a émergé pour traiter les immences quantité de données que nous générons chaque jour.
- Le terme Big Data se réfère à une accumulation de données très larges et très complexes pour être traitées par les outils classiques de gestion des bases de données.

Qu’est ce que l’intelligence artificielle, l’apprentissage automatique et l’apprentissage approfondie ?

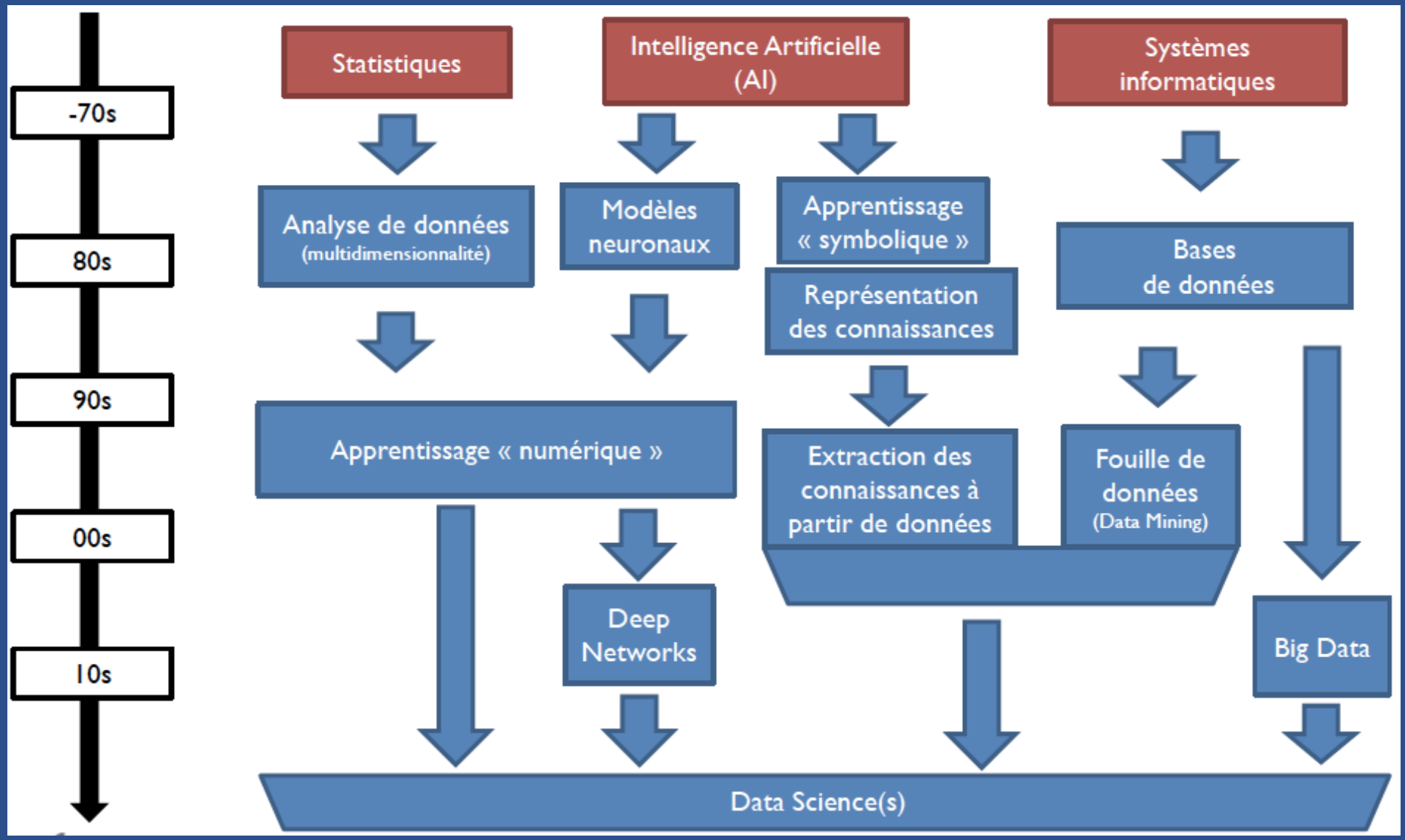
L’IA représente tout outil utilisé par une machine afin de « reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité

L’apprentissage automatique est un champ d’étude de l’intelligence artificielle qui vise à donner aux machines la capacité d’« apprendre » à partir de données, via des modèles mathématiques.



L'apprentissage profond est un proccédé et sous-ensemble de l'apprentissage automatique, utilisant des réseaux de neurones possédant plusieurs courches de neurones cachées.

Les origines scientifiques de la science de données

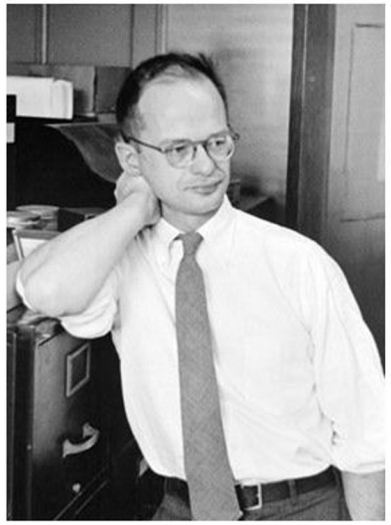


http://www.metz.supelec.fr/metz/personnel/pennerath/Cours/DataScience/

Historiques liés à la science de données
1943
Le neuron formel de McCulloch et Pitts

Warren Sturgis McCulloch

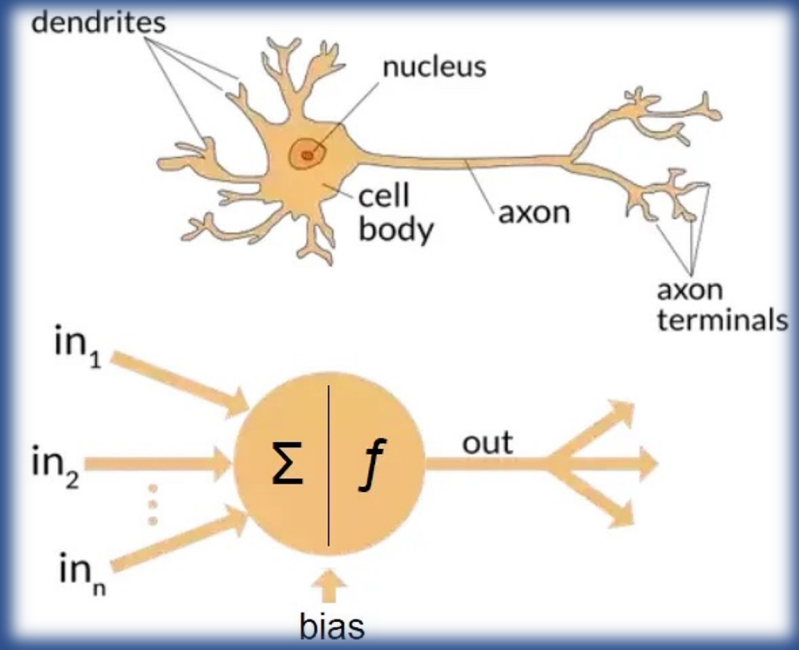
(1898 – 1969)



Walter Harry Pitts, Jr.

(1923 – 1969)

https://images.slideplayer.com/22/6379712/slides/slide_8.jpg

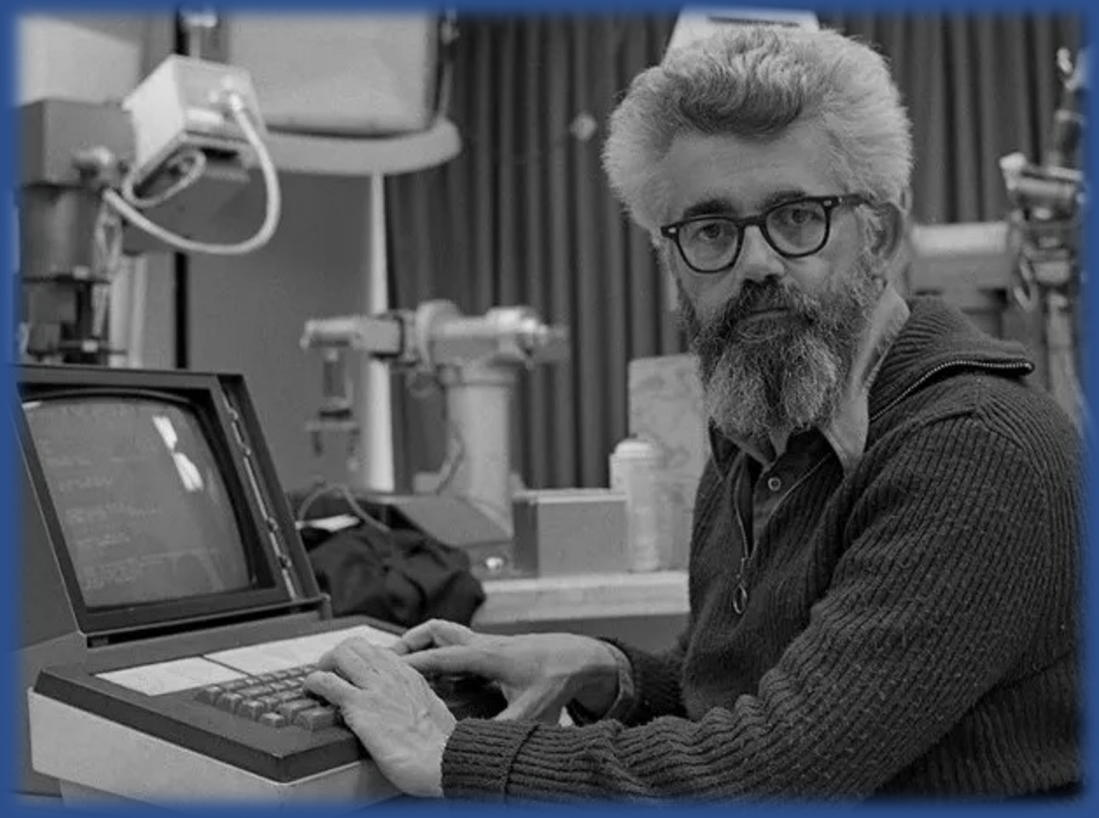


https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7

1956
John McCarthy

Artificial Intelligence

“The science and engineering of making intelligent machines, especially intelligent computer programs”. -John McCarthy-



https://www.independent.co.uk/news/obituaries/john-mccarthy-computer-scientist-known-as-the-father-of-ai-6255307.html

1959
Arthur Samuel

Qu’est ce que le machine learning ?

le *machine learning est* le champ d’étude visant à donner la capacité à une machine d’apprendre sans être explicitement programmée

https://www.ibm.com/history/early-games

1963
Alan Turing

- **Alan Turing** propose une approche innovante pour développer l'intelligence artificielle.
- Il suggère de créer un programme qui imite le cerveau d'un enfant. Par la suite ce programme sera eduqué pour atteindre la maturité intellectuel d’un adult.

https://fr.wikipedia.org/wiki/Alan_Turing

1997

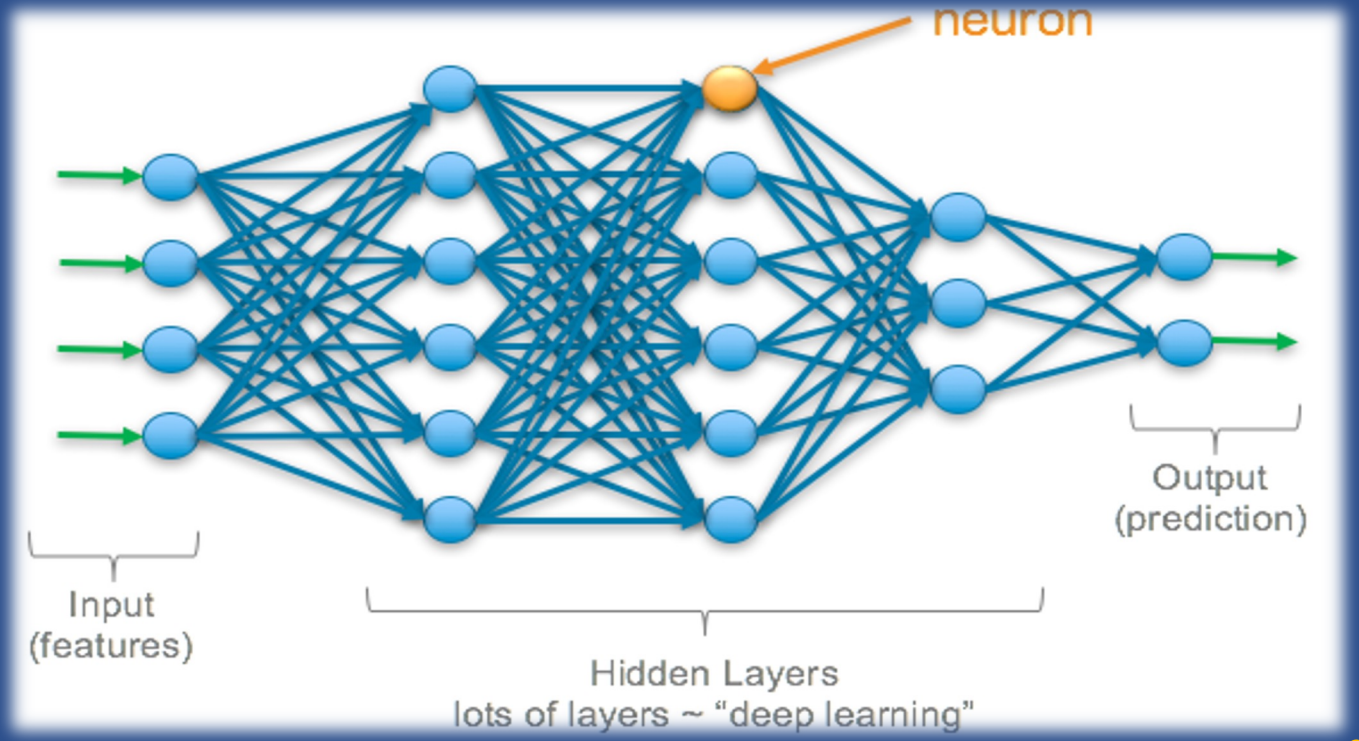
Le jour où Deep Blue a battu Garry Kasparov aux échecs

https://www.cscience.ca/le-jour-ou-deep-blue-a-battu-garry-kasparov-aux-echecs/

2005
Apprentissage Profond (Deep Learning)

Geoffrey Hinton(né le6 décembre 1947) est un chercheur canadien spécialiste de l'intelligence artificielle et plus particulièrement des réseaux de neurones artificiels. Il fait partie de l'équipe Google Brain et est professeur au département d'informatique de l'Université de Toronto. Il a été l'un des premiers à mettre en application l'algorithme de rétropropagation du gradient pour l'entraînement d'un réseau de neurones multi couches. Il fait partie des figures de proue de la communauté de l'apprentissage profond.

https://fr.wikipedia.org/wiki/Geoffrey_Hinton



https://srnghn.medium.com/deep-learning-common-architectures-6071d47cb383

2010
Big Data

Le terme Big Data se réfère à une accumulation de données très larges et très complexes pour être traitées par les outils classiques de gestion des bases de données.

https://iwconnect.com/algorithms-for-solving-big-data-use-cases/

2015
Alpha Go

Alphago est un programme informatique capable de jouer au jeu de go, développé par l'entreprise britannique deepmind et racheté en 2014 par google.
en octobre 2015, il devient le premier programme à battre un joueur professionnel (le français fan hui) sur un goban de taille normale (19×19) sans handicap. il s'agit d'une étape symboliquement forte puisque le programme joueur de go est alors un défi complexe de l'intelligence artificielle

En octobre 2015, AlphaGo a gagné par 5 à 0 un match contre Fan Hui

https://fr.wikipedia.org/wiki/Match_AlphaGo_-_Lee_Sedol#/media/Fichier:FanHui.jpg

AlphaGo

AlphaGo a gagné Lee Sedol toutes les parties sauf la quatrième, entre le 9-15 Mars 2016.

https://fr.wikipedia.org/wiki/Lee_Sedol

2022
Open AI

Openai (« ai » pour artificial intelligence, ou intelligence artificielle) est une entreprise américaine d'intelligence artificielle (ia) fondée en 2015 à san francisco en californie. sa mission est de développer et de promouvoir une intelligence artificielle générale « sûre et bénéfique à toute l'humanité ».

Open AI a lancé de Chatgpt en novembre 2022 a déclenché un intérêt mondial pour les agents conversationnels et l'ia générative, attirant 100 millions d'utilisateurs en à peine 2 mois

https://fr.wikipedia.org/wiki/OpenAI

2023
Midijourney

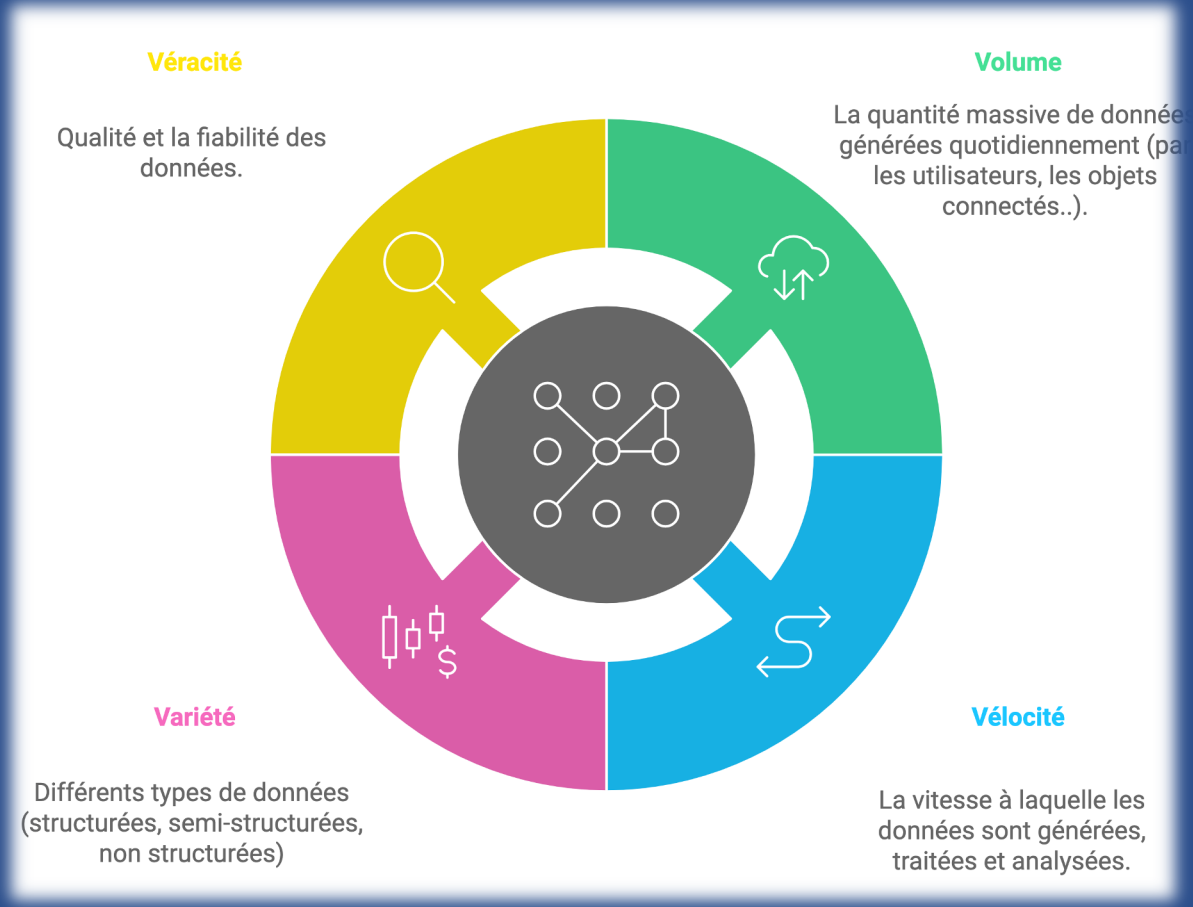
midjourney est un laboratoire de recherche indépendant. il produit un programme d'intelligence artificielle du même nom, lequel permet de créer des images à partir de descriptions textuelles, suivant un fonctionnement similaire à celui de dall-e d'openai4,5.

https://www.midjourney.com/home

Enjeux de la science de données

1. Le Volume et la Complexité des Données:

- **Big Data**: La croissance exponentielle des données rend leur stockage, leur traitement et leur analyse de plus en plus complexes.
- **Variété**: Les données peuvent être structurées, semi-structurées ou non structurées, ce qui nécessite des outils et des méthodes adaptés.
- **Vitesse**: La nécessité de traiter les données en temps réel ou quasi réel pour prendre des décisions rapides.



Enjeux de la science de données

2. La Qualité des Données:

- **Précision**: Des données erronées ou incomplètes peuvent conduire à des résultats biaisés.
- **Cohérence**: Les données doivent être cohérentes entre elles pour garantir la fiabilité des analyses.
- **Pertinence**: Il est essentiel de sélectionner les données pertinentes pour répondre à une question spécifique.

Enjeux de la science de données

3. Biais et éthique dans l'analyse des données

- **Biais dans les algorithmes** : Les modèles de machine learning et d'intelligence artificielle sont souvent biaisés si les données d'entraînement sont elles-mêmes biaisées.

Enjeux de la science de données

4. Sécurité des données

- **Cyberattaques et fuites de données** : La science des données repose souvent sur des volumes massifs d'informations stockées dans des bases de données ou des infrastructures cloud. Cela en fait une cible attrayante pour les cyberattaques. Assurer la sécurité des infrastructures, l'intégrité des données, et la résilience des systèmes est donc essentiel.
- **Protection contre les manipulations** : Les données peuvent également être manipulées pour produire des résultats biaisés ou malhonnêtes (comme les fausses nouvelles ou la désinformation).

Facettes et types de données
1. **Données structurées**

Les données peuvent être identifiées selon leur structure.

Définition (Données Structurées)

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

Exemples:

- Données stockées dans des feuilles (e.g., Fichier Excel).
- Enregistrements stockés dans les tables d'une base de données relationnelle.
- Chaque propriété est distinguée facilement des autres.
- Elle correspond à une unité de la structure (e.g., colonne de la table).

2. **Données non structurées**

Définition (Données Non Structurées)

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

Exemples:

- Un texte est non structuré.
- Description des propriétés d'une entité noyée dans un contexte riche.
- Aucun accès direct à ces propriétés.

3. **Données semi-structurées**

Définition (Données semi-structurées)

Une donnée semi-structurée possède une structure où les entitées 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.

Exemples: Document XML, JSON, HTML

Xml

```xml
<book id="bk101">
  <author>Gambardella, Matthew</author>
  <title>XML Developers Guide</title>
  <genre>Computer</genre>
  <price>44.95</price>
  <publish_date>2000-10-01</publish_date>
</book>
```

Json

```json
{
  "nom": "Alice",
  "âge": 30,
  "adresse": {
    "rue": "123 Rue de Paris",
    "ville": "Paris"
  },
  "téléphones": ["123-456-7890", "098-765-4321"]
}
```

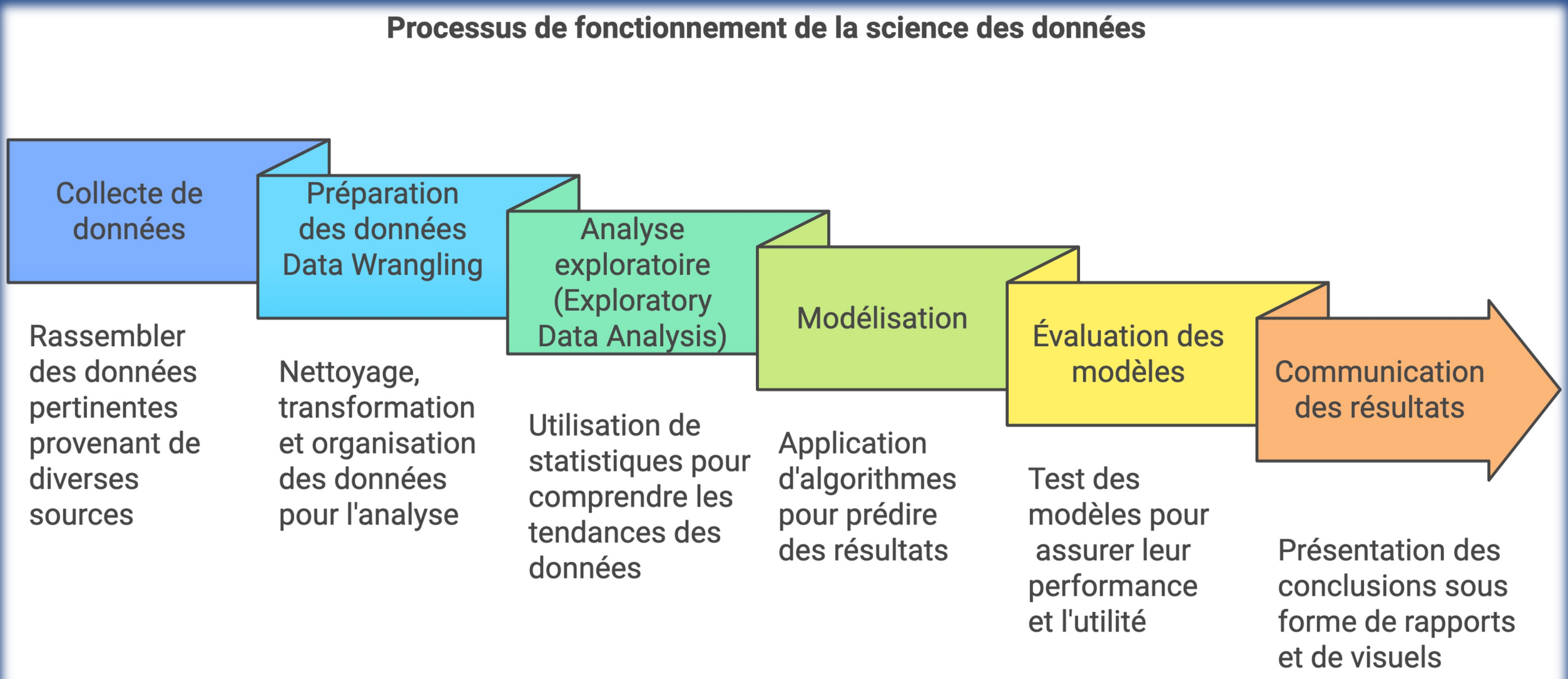
Comment fonctionne la science de données ?

- La science des données est une discipline qui fonctionne en suivant un processus structuré de collecte, traitement, analyse et interprétation des données.
- Elle repose sur un ensemble d'étapes clés pour extraire des informations exploitables à partir des données.

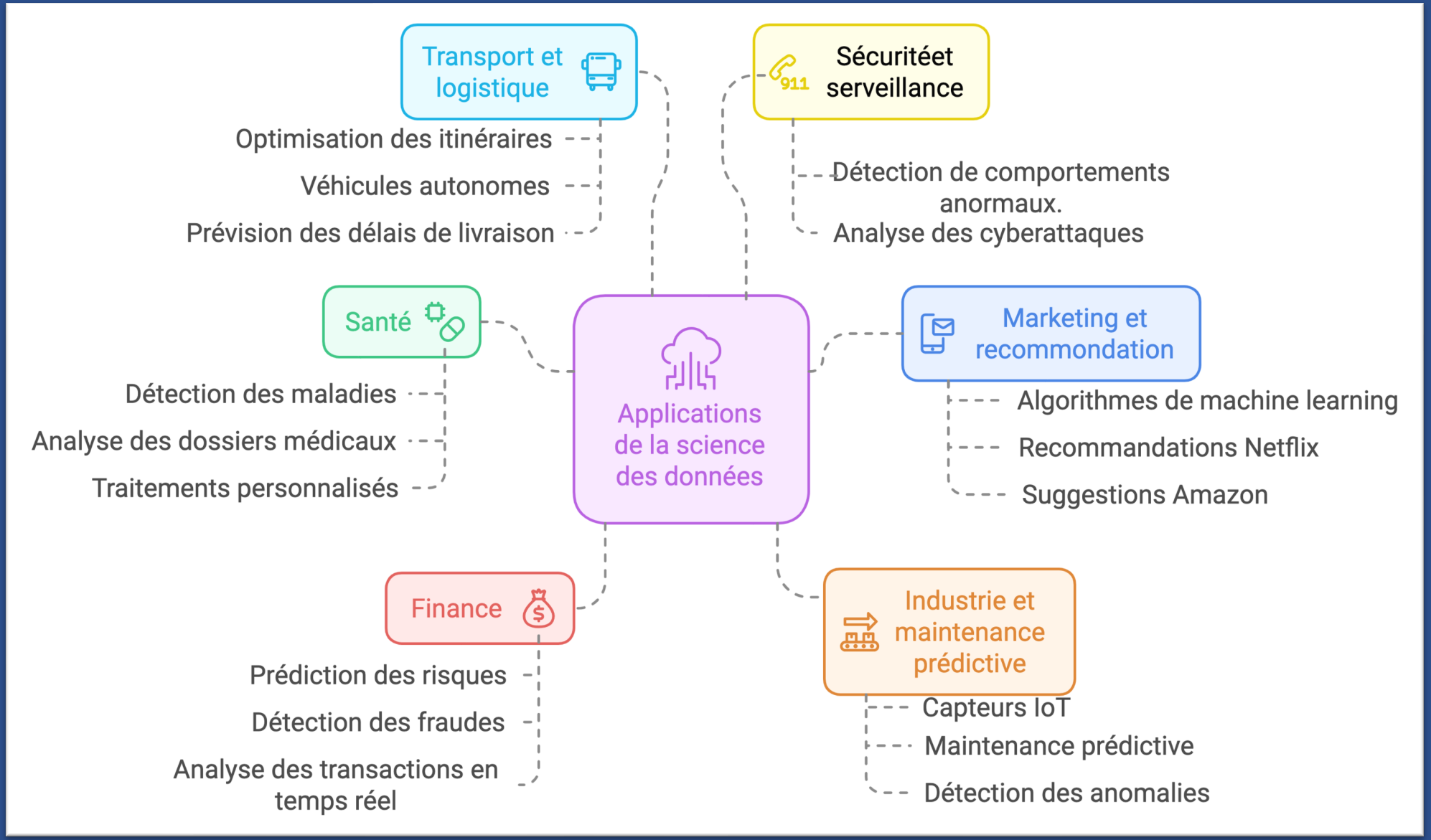
- C'est un peu comme un détective qui cherche à résoudre un mystère en utilisant des indices.



Comment fonctionne la science de données?



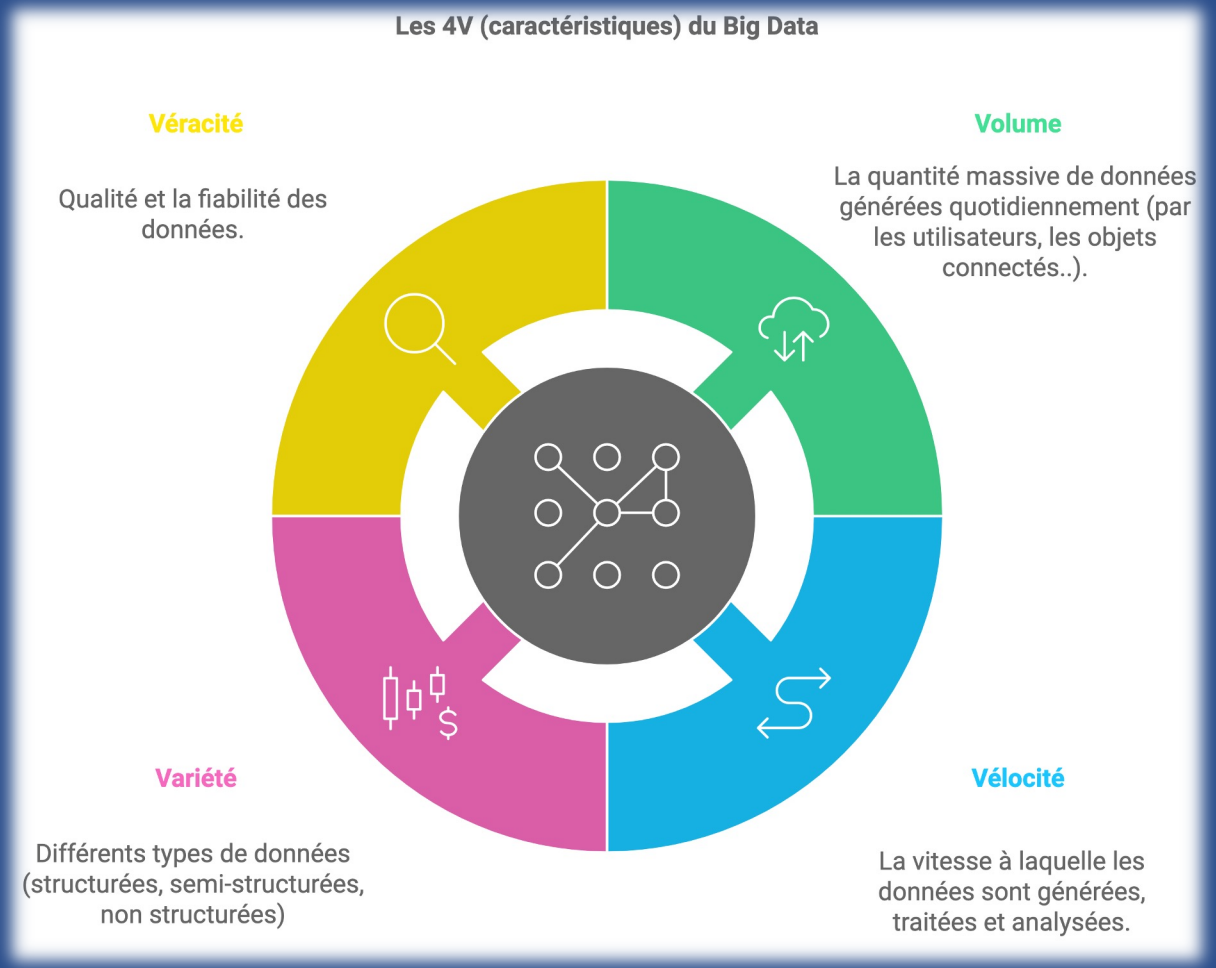
Cas d'usage et domaines d'application



L’écosystème de Big Data et de la science de données

- **Big Data** fait référence aux ensembles de données massifs et complexes, souvent en temps réel.
- Nécessite des technologies et des infrastructures spécifiques pour être traités efficacement.

- Voici les composantes principales de l'écosystème Big Data



L'écosystème du Big Data et de la science des données

- **Collecte de données** : Données issues de sources internes, externes et des objets connectés (IoT).
- **Stockage** : Utilisation de bases de données (SQL, NoSQL), data lakes (stocakge centralisé pour structurés et non structués), et entrepôts de données (stockage de données structués, Amazone redshift,Google Big Query).
- **Traitement** : Technologies distribuées (Hadoop, Spark) et en temps réel (Kafka, Flink).
- **Analyse** : Analyse descriptive, prédictive, et prescriptive, avec outils de machine learning (TensorFlow, Pytorch, Scikit-learn).
- **Visualisation** : Outils comme Tableau, Power BI, les bibliothèques de Python tel que Matplotlib, pandas.. pour rendre les insights accessibles.
- **Sécurité et Gouvernance** : Assurer la qualité et la protection des données.

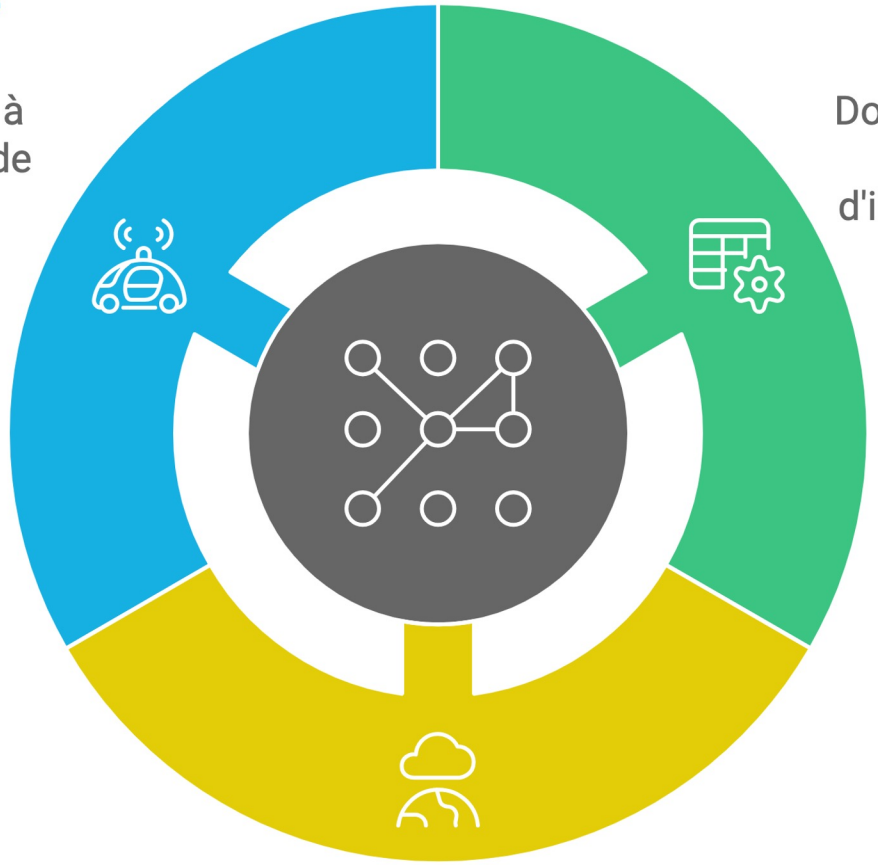
Source de données

Données machine

Données collectées à partir d'appareils et de capteurs IoT

Données internes

Données générées par les systèmes d'information internes de l'organisation



Données externes

Données provenant de sources externes comme les réseaux sociaux et les API

Stockage de données

Stockage des données dans le Big Data

Bases de données NoSQL

Conçues pour des volumes massifs de données non structurées ou semi-structurées.

Data Lakes

Stockage centralisé pour les données brutes dans leur format d'origine.

Bases de données relationnelles

Utilisées pour les données structurées, efficaces pour les petits volumes.

Entrepôts de données

Stocke et organise des données structurées pour des requêtes analytiques.

Outils de traitement de données

Traitement Distribué

Pour de grands volumes de données.
Exemples : Hadoop, Spark

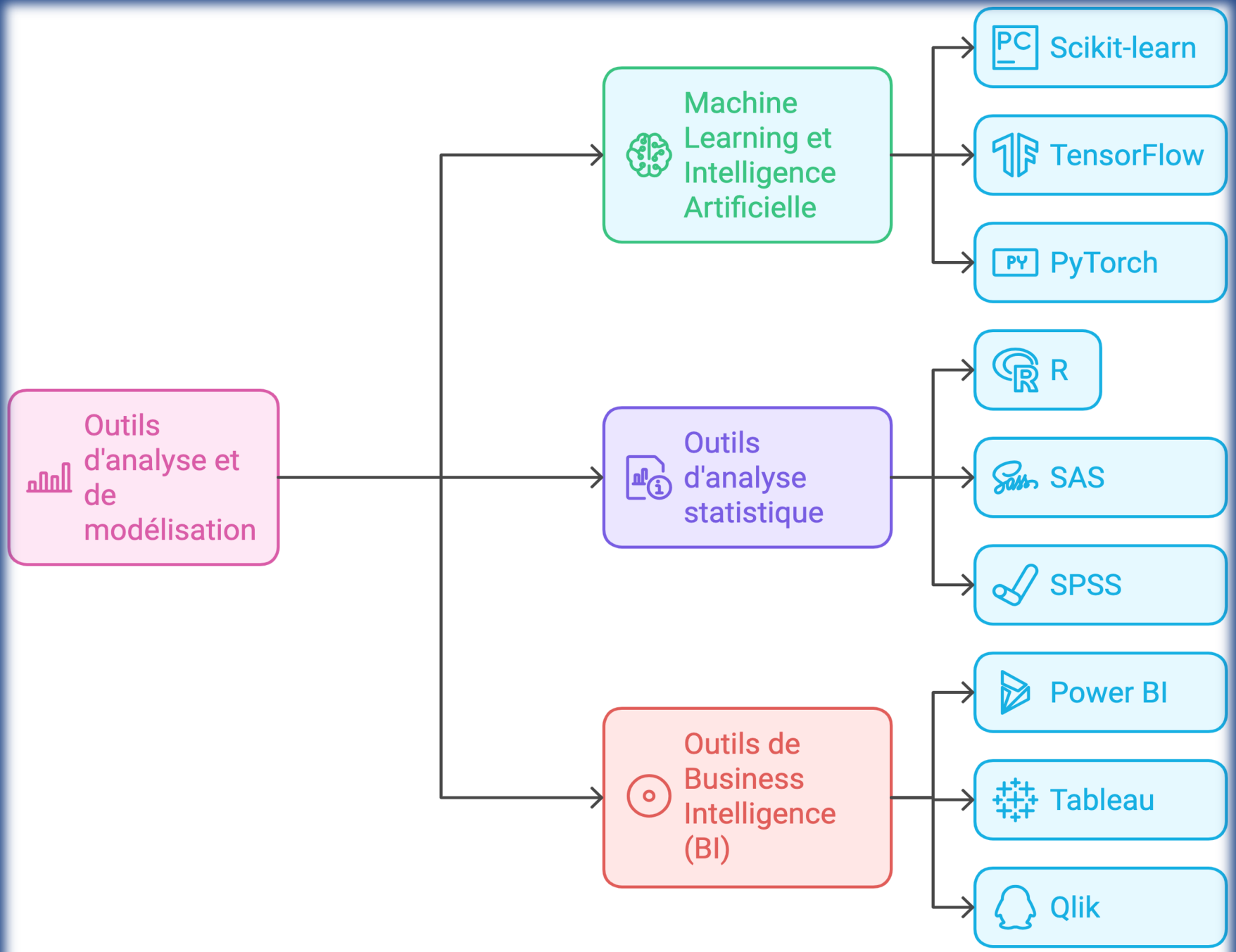
Traitement en Temps Réel

Pour l'analyse des données en temps réel.
Exemple : Apache Kafka, Apache Flink

ETL Extract, Transform, Load

Pour l'extraction, la transformation et le chargement des données.
Exemple : Talend, Apache Nifi, Informatica

Aperçu des Outils et technologies de Traitement des Données

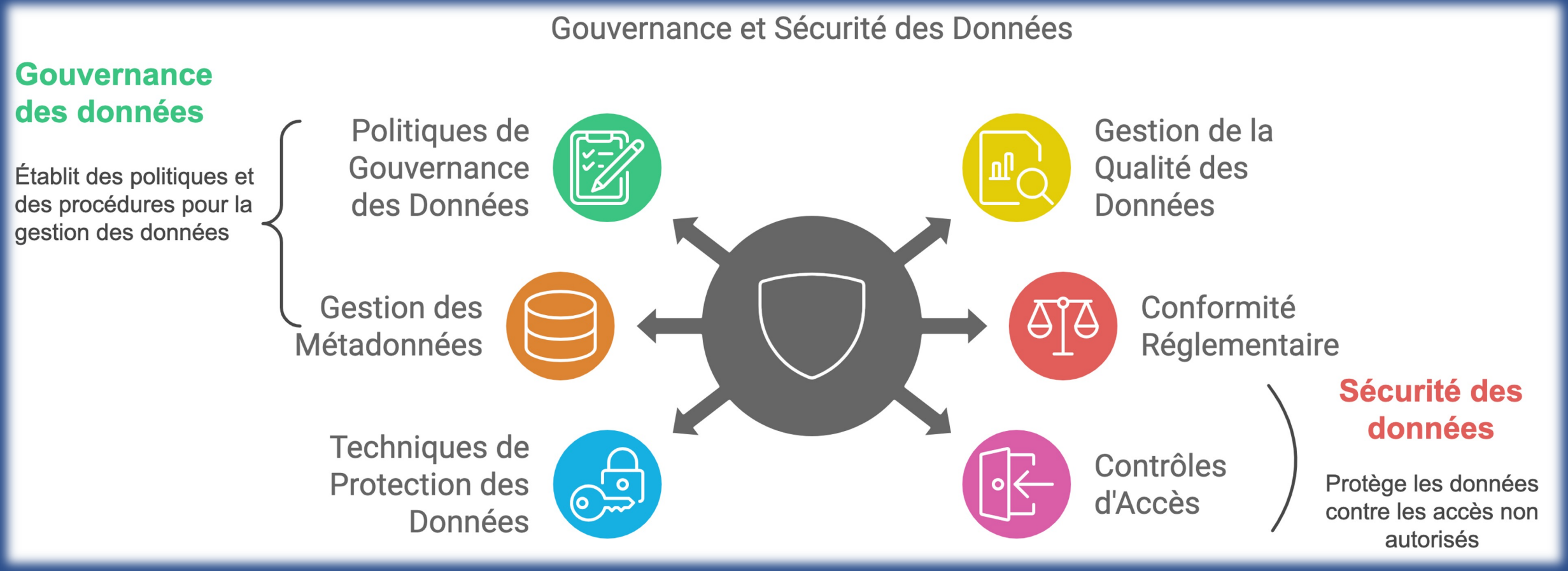


Outils d'analyse et de modélisation

Visualisation des données

- La **visualisation des données** est un élément essentiel pour rendre les résultats des analyses compréhensibles et exploitables. Les outils de visualisation permettent de transformer des résultats analytiques complexes en graphiques, tableaux de bord, ou cartes interactives.
- **Outils de visualisation** :
  - **Tableau de bord interactif** : Power BI, Tableau.
  - **Bibliothèques de visualisation** : Matplotlib, D3.js, Plotly.

Gouvernance et sécurité de données



Cloud Computing dans l’écosystème big data