Chapitre 2

Généralités

2.1 Introduction

La data science est une branche de l'intelligence artificielle qui englobe les domaines interconnectés de la statistique, des méthodes scientifiques et de l'analyse des données. Ces éléments sont utilisés pour extraire du sens et des perspectives à partir des données. Avec l'avènement d'Internet, le monde est devenu extrêmement connecté, où chaque objet manipulé (voitures, réfrigérateurs, vêtements, réseaux sociaux, etc.) génère quotidiennement des millions de données supplémentaires qui s'ajoutent à un vaste océan de données. Toutes ces données peuvent être exploitées pour offrir des services personnalisés et immédiats, répondant ainsi aux attentes des utilisateurs. Mais comment pouvons-nous transformer cet océan de données apparemment infini en un flux régulier d'informations pertinentes ? La réponse réside dans l'intelligence artificielle.

L'intelligence artificielle (IA), ou AI en anglais pour "Artificial Intelligence", englobe un ensemble de techniques visant à permettre aux machines de simuler une forme d'intelligence semblable à celle des êtres humains. L'IA est appliquée dans divers domaines tels que la médecine, le transport, la photographie, etc., avec pour objectif le développement de machines capables de comportements intelligents.

Cependant, un problème se pose actuellement : la croissance exponentielle des données dépasse les capacités des bases de données traditionnelles. Ceci est dû au volume considérable des données, à leur diversité et au temps de traitement qui doit rester dans des délais acceptables

pour les utilisateurs. Les développeurs sont donc confrontés au défi de mettre au point des technologies capables de traiter d'énormes quantités de données variées en un temps réduit, ce que l'on appelle le Big Data.

La **Figure 2.1** illustre l'interrelation entre l'intelligence artificielle (IA), la science des données (Data Science) et le Big Data, qui joue un rôle essentiel dans l'adoption de la transformation numérique. L'IA utilise des algorithmes complexes pour traiter les données, la science des données collecte, traite et analyse les données pour alimenter l'IA, tandis que le Big Data fournit les quantités massives de données nécessaires à l'IA.

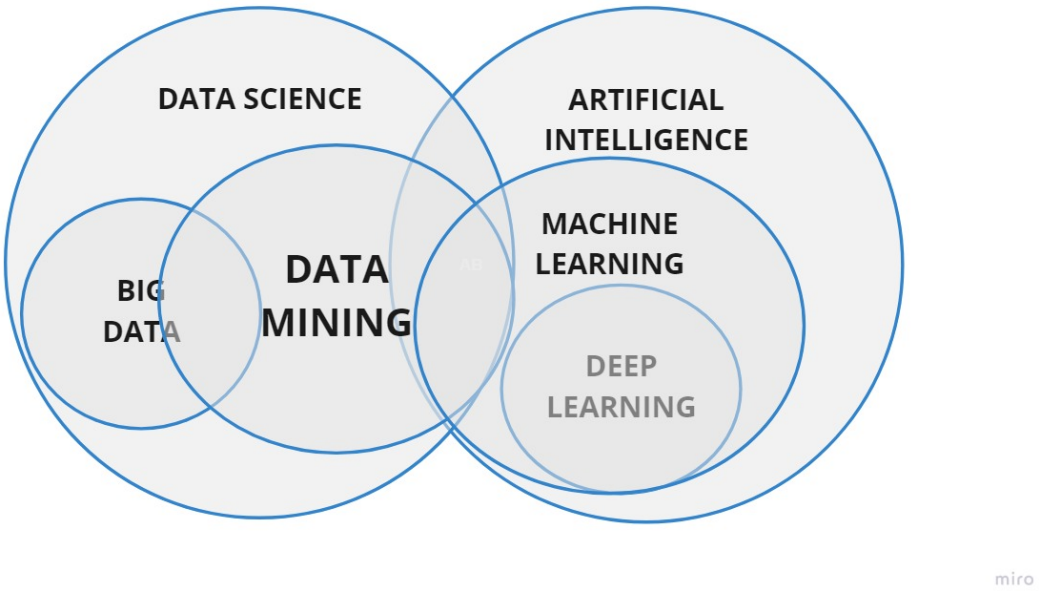

FIGURE 2.1 – Interaction entre l'Intelligence Artificielle, la data science et le big data

Dans la continuation de ce chapitre, nous allons introduire les notions essentielles liées au domaine du "Big Data" et examiner également les concepts de l'apprentissage automatique (Machine Learning).

2.2 Le Big Data et son analyse

Le monde est constamment alimenté par des données et fait l'objet d'une analyse perpétuelle. Le domaine de l'analyse des données (DAD) joue un rôle essentiel dans tous les secteurs en permettant d'extraire du sens des données collectées, ouvrant ainsi la voie à un avenir

extraordinaire. Par exemple, il permet la conception de voitures autonomes sécurisées, le développement de médicaments efficaces et l'amélioration de nos prises de décision grâce à des machines intelligentes, etc.

Bien que l'acronyme de l'analyse des données (ADD) puisse différer de celui du Big Data, il est la clé pour donner du sens à toutes les informations que nous collectons.

2.2.1 Définition du Big Data

Le concept de "Big Data" a suscité plusieurs définitions qui n'ont pas été universellement adoptées en raison de sa complexité et de sa variation selon les utilisateurs et les fournisseurs de services impliqués. Parmi ces définitions, nous pouvons citer :

Contrairement aux données traditionnelles, le terme "Big Data" fait référence à des ensembles de données volumineux et en constante croissance, comprenant des formats hétérogènes tels que des données structurées, non structurées et semi-structurées. Les mégadonnées ont une nature complexe qui nécessite des techniques puissantes et des algorithmes avancés. Ainsi, les outils traditionnels de business intelligence statique ne sont plus efficaces dans le contexte des applications Big Data [8].

Le Big Data, également connu sous le nom de mégadonnées, englobe un ensemble de technologies, d'architectures et de procédures qui permettent d'analyser et de traiter d'importantes quantités de données hétérogènes, afin d'extraire des informations pertinentes à un coût raisonnable [9].

Le terme "Big Data", traduit littéralement par "grosses données" ou "données massives", fait référence à l'explosion de données. On peut également le comparer à la notion de "datamasse" en faisant une analogie avec la biomasse, qui représente un écosystème complexe à grande échelle.

Une autre définition couramment utilisée est celle proposée par IBM[1] :

Le Big Data fait référence aux quantités exponentielles de données, à la fois structurées et non structurées, qui sont si vastes et complexes qu'elles dépassent les capacités des systèmes traditionnels d'information pour les collecter, les stocker, les gérer et les analyser efficacement

afin d'en tirer des informations précieuses.

2.2.2 Évolution historique du Big Data

L'évolution du Big Data peut être divisée en différentes périodes, marquées par des avancées technologiques majeures et des changements dans l'utilisation des données par les entreprises.

- **Les années 1960 et 1970 :** Pendant cette période, les premières bases de données et les premiers systèmes de gestion de bases de données (SGBD) ont été développés. Bien que coûteux et complexes, ces systèmes ont permis aux entreprises de stocker et de traiter de grandes quantités de données.
- **Les années 1970-2000 :** Avec l'augmentation des volumes de données, les entreprises ont commencé à utiliser des SGBD pour stocker et gérer leurs données. Cependant, ces systèmes étaient limités en termes de capacité et de vitesse de traitement.
- **Au début des années 2000 :** Avec l'avènement d'Internet et le développement des technologies de l'information, les entreprises ont commencé à collecter de plus en plus de données en ligne, telles que les données de navigation Web, les données de vente et les données marketing. Cependant, les systèmes informatiques existants étaient incapables de traiter ces quantités massives de données, ce qui a nécessité le développement de nouvelles technologies [10].
- **En 2004 :** Google a publié un article scientifique sur la méthode MapReduce, utilisée pour le traitement des données à grande échelle sur des clusters informatiques [11].
- **En 2006 :** Doug Cutting et Mike Cafarella ont développé le système de gestion de données distribué Hadoop, qui a permis aux entreprises de collecter, stocker et analyser d'énormes volumes de données de manière efficace.
- **Les années 2010-2015 :** L'expression "Big Data" est apparue pour décrire le défi croissant de la gestion des grandes quantités de données. Les entreprises ont adopté des technologies telles que le stockage distribué et le traitement parallèle pour traiter les données massives.
- **Au cours des années 2015 :** Les entreprises continuent d'adopter des solutions Big Data pour collecter, stocker, gérer et analyser les données, en utilisant des technolo-

1. source Site office du BIM : https ://www.ibm.com/cloud/learn/big-data

gies telles que le machine learning et l'IA [9]. Les défis du Big Data comprennent la qualité des données, la sécurité des données et l'analyse efficace des données massives. L'impact de la pandémie de COVID-19 a également entraîné une augmentation de la collecte de données pour surveiller les tendances et les effets de la pandémie. En 2021, les entreprises continuent d'investir massivement dans le Big Data, avec une prévision de dépenses mondiales d'environ 274 milliards de dollars pour l'année [9].

Aujourd'hui, le Big Data est un domaine en constante évolution et les entreprises cherchent à exploiter les avantages de l'analyse des données pour améliorer la prise de décision, le marketing et les opérations commerciales.

2.2.3 Caractéristiques du Big Data

Les traits fondamentaux des mégadonnées ou Big Data, illustrés dans la Figure2.2, sont couramment désignés sous le nom des "Cinq V" :

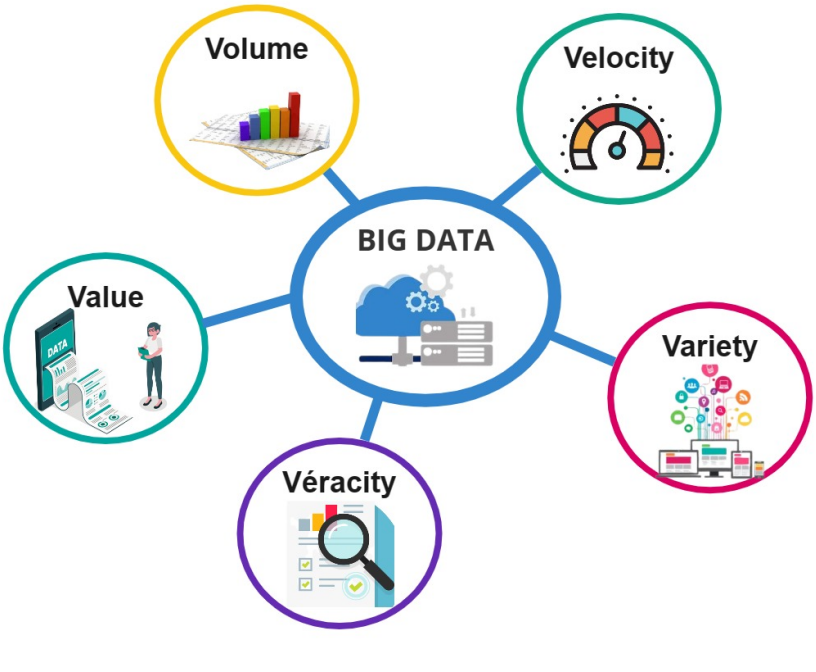

FIGURE 2.2 – Les cinq Vs du Big Data.

Volume :

Le Big Data implique le traitement d'énormes volumes de données non structurées à faible densité. Ces données peuvent provenir de diverses sources, telles que les flux de clics sur des

sites Web, les activités sur les réseaux sociaux ou les mesures de capteurs. Pour certaines organisations, cela peut représenter des dizaines de téraoctets de données, tandis que pour d'autres, cela peut atteindre des centaines de pétaoctets.

Vélocité (Vitesse) :

La vélocité fait référence à la vitesse à laquelle les données sont générées et traitées. Les données à grande vitesse nécessitent une ingestion et un traitement en temps réel. Par exemple, les systèmes de surveillance en temps réel ou les plateformes d'analyse de données en streaming doivent être capables de traiter et de réagir aux données en quasi-temps réel.

Variété :

La variété concerne les différents types de données auxquels le Big Data fait face. Les données traditionnelles étaient principalement structurées et stockées dans des bases de données relationnelles. Cependant, avec l'avènement du Big Data, les données peuvent être non structurées, telles que des données textuelles, audio ou vidéo. Le traitement de ces données nécessite des techniques spécifiques pour les interpréter et les exploiter, ainsi que pour gérer les métadonnées associées.

Véracité :

La véracité se réfère à l'exactitude et à la fiabilité des données. Pour obtenir de la valeur à partir des données, il est essentiel de nettoyer les données pour éliminer les erreurs et les incohérences. La véracité des données est essentielle pour garantir la qualité des analyses et des décisions basées sur ces données.

Valeur :

La valeur représente l'utilité et l'importance des données pour atteindre un objectif spécifique. L'objectif ultime de l'analyse des mégadonnées est d'extraire de la valeur à partir des données collectées. La valeur des données peut également être liée à leur validité et à leur exactitude. Dans certains cas, la valeur dépend également de la rapidité avec laquelle les données peuvent être traitées pour prendre des décisions rapides et éclairées [12].

2.2.4 Structuration du Big Data

La structuration des données est le processus d'organisation et de stockage des données dans un ordinateur de manière à ce qu'elles puissent être référencées et modifiées de manière efficace. Dans le contexte du Big Data, les données collectées, stockées et traitées proviennent de différents domaines et sont générées par de multiples sources de données hétérogènes, ce qui crée une masse de données de natures différentes [13].

2.2.4.1 Données structurées

Les données structurées font référence à des données d'un format et d'une longueur spécifiques, d'une facilité de stockage et d'analyse et d'une organisation élevée. Cela signifie que les données sont organisées dans une structure reconnaissable afin qu'elles puissent répondre aux requêtes pour récupérer des informations à des fins organisationnelles. Une base de données relationnelle comme Structured Query Language (SQL) est un bon exemple de données structurées, elle contient des nombres structurés, des dates, des combinaisons de mots et de nombres appelées chaînes/texte.

En raison de la structure transparente de la base de données, elle peut être recherchée à l'aide d'algorithmes de recherche simples et directs qui peuvent être par type de données dans le contenu réel [14]

2.2.4.2 Données semi-structurées

Les données non structurées sont des informations qui, sous de nombreuses formes différentes, ne correspondent pas aux modèles de données traditionnels et ne conviennent donc généralement pas à une base de données relationnelle traditionnelle. Cela rend le traitement et l'analyse des données non structurées très difficiles et chronophages. Selon Feldman et Sanger, les données non structurées n'ont pas de structure définie.

Les données non structurées incluent généralement des images/objets bitmap, du texte, des e-mails et d'autres types de données qui ne font pas partie de la base de données[15].

Il est important de comprendre la nature et la structure des données afin de déterminer la meilleure approche pour les collecter, les stocker et les analyser. Les solutions Big Data sont capables de gérer différents types de données et de les traiter de manière efficace, permettant

ainsi une analyse plus approfondie et une prise de décision améliorée.

2.2.5 Traitement du Big Data

Le traitement du Big Data comprend généralement les étapes suivantes, illustrées dans la Figure 2.3 :

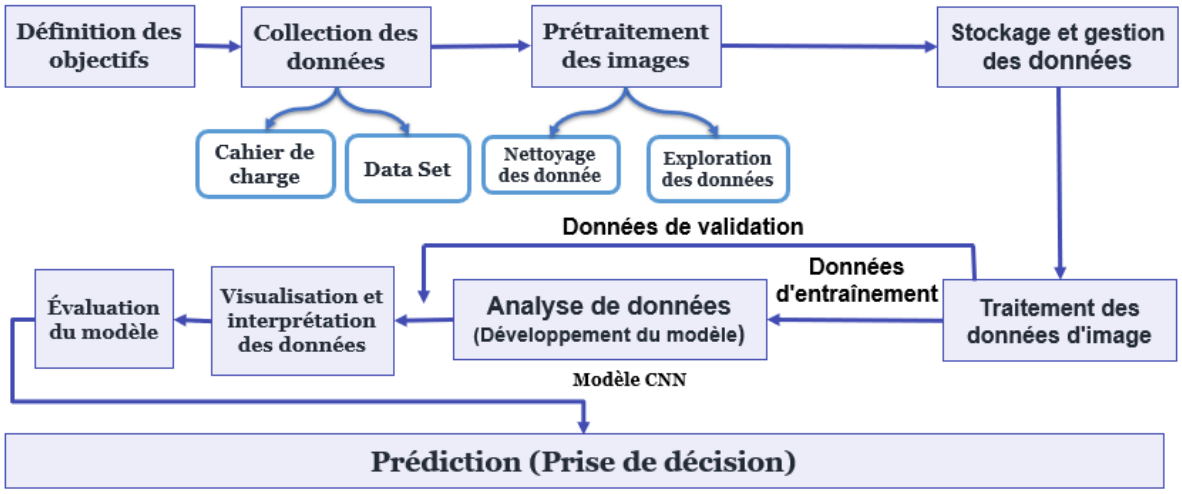

FIGURE 2.3 – Étapes du traitement des Big Data.

2.2.5.1 Définition des objectifs

La première étape du traitement des Big Data consiste à identifier les objectifs à atteindre. Quelles informations ou quelles questions souhaitez-vous obtenir à partir des données ? Il est important de définir clairement les objectifs pour orienter les étapes suivantes.

2.2.5.2 Collecte de données

La collecte de données est la première étape de la gestion des Big Data. Elle implique la récupération de données à partir de différentes sources telles que des bases de données internes, des capteurs connectés, des réseaux sociaux, des fichiers journaux, des données GPS, etc. Ces données peuvent être structurées, semi-structurées ou non structurées. Il est important de collecter suffisamment de données pour obtenir des informations pertinentes, tout en évitant une collecte excessive qui rendrait le processus d'analyse coûteux et complexe [16].

2.2.5.3 Préparation des données

La préparation des données consiste à nettoyer, normaliser et transformer les données pour les rendre utilisables dans le cadre de l'analyse. Cela peut inclure des tâches telles que la suppression des données en double, la correction des erreurs de saisie, la conversion des données non numériques en données numériques, la fusion de données provenant de différentes sources, etc [16].

2.2.5.4 Stockage et gestion des données

Le stockage et la gestion des données font référence aux processus et aux techniques utilisés pour stocker, organiser, sécuriser et gérer les données de manière efficace et fiable. Les Big Data nécessitent souvent des solutions de stockage et de gestion appropriées pour une utilisation efficace. Des solutions telles que les systèmes de fichiers distribués (comme Hadoop HDFS) ou les bases de données NoSQL sont utilisées pour gérer et organiser les données de manière évolutive.

2.2.5.5 Analyse des données

L'analyse des données consiste à utiliser des techniques statistiques et d'apprentissage automatique pour extraire des informations et des insights à partir des données. Des algorithmes d'apprentissage automatique tels que la régression, les arbres de décision, les réseaux neuronaux, etc., peuvent être utilisés pour prédire des tendances, identifier des anomalies, découvrir des relations cachées, etc [16].

2.2.5.6 Visualisation et interprétation des données

La visualisation et l'interprétation des données sont les étapes suivantes de la gestion des Big Data. Cette phase consiste à présenter les résultats de l'analyse sous forme de graphiques, de tableaux et de rapports pour une meilleure compréhension. Des outils de visualisation tels que des graphiques, des tableaux croisés dynamiques et des cartes peuvent être utilisés pour représenter les données de manière claire et concise [16].

2.2.5.7 Intégrité et sécurité des données

La dernière étape de la gestion des Big Data consiste à garantir l'intégrité et la sécurité des données. Il est important de veiller à la qualité, à l'exactitude et à la confidentialité des données en utilisant des méthodes de cryptage, d'authentification et de sauvegarde pour protéger les données contre les risques de sécurité tels que les violations de données, les cyberattaques et les erreurs humaines. Des politiques de sécurité des données doivent être mises en place pour gérer l'accès aux données, les mots de passe et les autorisations d'accès. Il est également important de surveiller régulièrement les données afin de détecter les anomalies et les incohérences [16].

2.3 Apprentissage automatique (Machine Learning)

L'apprentissage automatique, également connu sous le nom de Machine Learning, est une méthode d'analyse de données qui automatise la construction de modèles analytiques. C'est une branche de l'intelligence artificielle (IA) qui repose sur l'idée que les systèmes peuvent apprendre à partir de données, identifier des motifs et prendre des décisions avec un minimum d'intervention humaine. Les algorithmes d'apprentissage automatique peuvent être utilisés pour diverses tâches telles que la reconnaissance d'images et de la parole, le traitement du langage naturel et la prise de décision [8].

Dans cette section, nous examinerons la définition de l'apprentissage automatique et comment elle a évolué au fil du temps, ainsi que quelques types courants du Machine Learning .

2.3.1 Définitions de l'apprentissage automatique

Au fil du temps, la définition de l'apprentissage automatique a évolué pour inclure de nouveaux développements et concepts. Voici quelques-unes de ces définitions :

- En 1997, Tom Mitchell, un chercheur américain, a défini l'apprentissage automatique comme l'étude d'algorithmes informatiques conçus pour effectuer des tâches sans être explicitement programmés pour les accomplir [**?**].
- L'apprentissage automatique est une approche permettant à un ordinateur d'apprendre à effectuer des calculs sans être explicitement programmé. Cette définition a été proposée par Arthur Samuel, un mathématicien américain qui a développé un programme capable

d'apprendre à jouer aux dames en 1959 (Figure 2.4) [1].

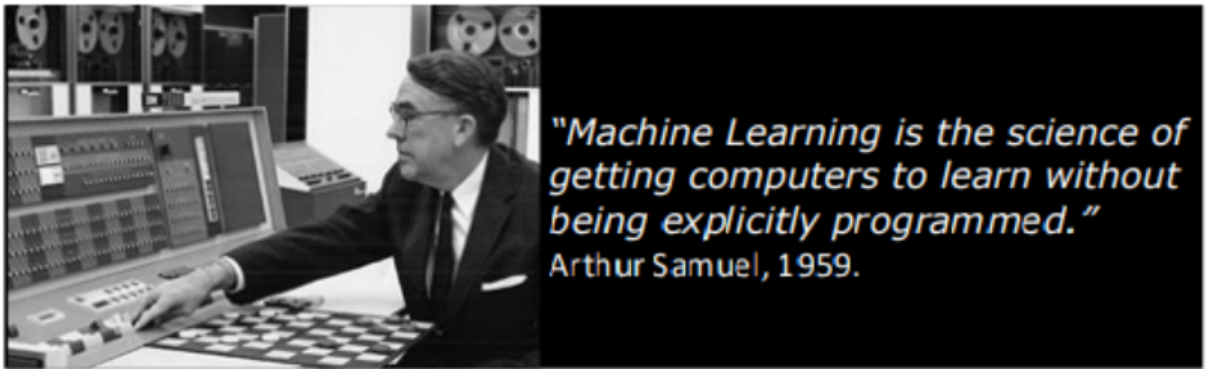

FIGURE 2.4 – Définition de l'apprentissage automatique par Arthur Samuel [1]

- Une définition plus récente, issue d'un sous-ensemble de l'Union européenne en 2020, décrit l'apprentissage automatique comme une approche informatique visant à développer des modèles capables de résoudre des problèmes complexes à partir de données.
  Ces définitions montrent que l'apprentissage automatique est un domaine interdisciplinaire axé sur la création de systèmes informatiques capables d'apprendre de manière autonome à partir de données et de les utiliser pour résoudre des problèmes complexes. Les algorithmes d'apprentissage automatique sont utilisés dans une grande variété d'applications telles que la médecine, le filtrage des e-mails, la reconnaissance vocale et la vision par ordinateur.

2.3.2 Évolution de l'apprentissage automatique

L'histoire de l'apprentissage automatique est riche et remonte à plusieurs décennies. Elle a été influencée par les développements clés dans les domaines de l'intelligence artificielle et de la statistique.
Au début des années 1950, les premiers concepts d'apprentissage automatique ont été proposés par des chercheurs en IA tels que Marvin Minsky et John McCarthy. Dans les années 1960, les algorithmes de régression linéaire et les arbres de décision ont été introduits pour effectuer des prédictions à partir de données. Les algorithmes d'apprentissage supervisé, tels que les réseaux de neurones, sont apparus dans les années 1980 pour traiter des problèmes complexes. Au début des années 1990, les algorithmes d'apprentissage non supervisé, tels que les algorithmes de regroupement (clustering), ont été introduits pour explorer les structures cachées des données. Les algorithmes de Deep

Learning, qui utilisent des réseaux de neurones profonds, ont été développés dans les années 2000 pour traiter des données complexes telles que les images et les textes. Dans les années 2010, les algorithmes d'apprentissage par renforcement, qui permettent aux machines de s'adapter à leur environnement en apprenant à partir de retours d'information, ont été introduits. Plus récemment, les algorithmes d'apprentissage automatique distribué, en temps réel et pour la prise de décision et l'optimisation, sont devenus de plus en plus populaires pour résoudre des problèmes complexes en temps réel.
En résumé, l'histoire de l'apprentissage automatique montre comment ce domaine est devenu un élément clé de l'IA et comment les algorithmes ont évolué pour traiter des données de plus en plus complexes et des problèmes plus difficiles.

2.3.3 Types de Machine Learning

2.3.3.1 Apprentissage supervisé

L'apprentissage supervisé est une méthode d'apprentissage automatique dans laquelle un algorithme informatique est formé à partir d'exemples de données étiquetées pour effectuer une tâche spécifique. Dans un système d'apprentissage supervisé, les algorithmes utilisent les données d'entraînement pour apprendre à faire des prédictions sur de nouvelles données. Les algorithmes apprennent en comparant leurs prédictions à des étiquettes correctes pour les données d'entraînement et en ajustant leurs modèles en conséquence.
L'apprentissage supervisé est utilisé pour de nombreuses tâches, telles que la classification de données, la régression linéaire, la reconnaissance d'images, la reconnaissance de la parole, etc. Il est largement utilisé dans les domaines de la science des données, de l'intelligence artificielle (IA) et du machine learning (ML) [10].

2.3.3.2 Apprentissage non supervisé

L'apprentissage non supervisé (Unsupervised Learning) est une méthode du ML dans laquelle un algorithme informatique est formé à partir de données non étiquetées pour découvrir des structures et des relations dans les données. Dans un système d'apprentissage non supervisé, l'algorithme est libre de découvrir des motifs et des modèles dans les données sans guidance explicite. Les algorithmes peuvent utiliser des techniques

telles que la réduction de dimensions, la clusterisation et la génération de modèles pour découvrir des structures cachées dans les données (voir Figure 2.5).

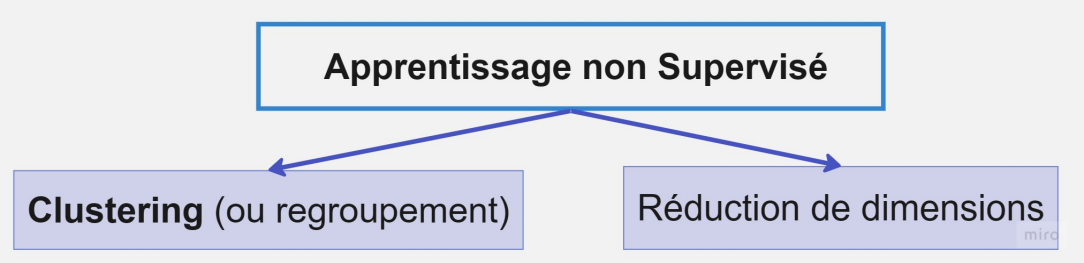

FIGURE 2.5 – Apprentissage non supervisé.

L'algorithme des k-moyennes et les algorithmes Apriori sont des exemples d'apprentissage non supervisé [10].

2.3.3.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une méthode de machine learning qui combine des données étiquetées et non étiquetées pour entraîner un algorithme. Cette approche permet à l'algorithme d'avoir accès à une petite quantité de données étiquetées, qui peuvent fournir des informations précieuses pour guider le processus d'apprentissage, ainsi qu'à une grande quantité de données non étiquetées, qui peuvent améliorer la capacité de généralisation du modèle.
L'apprentissage semi-supervisé se situe entre l'apprentissage supervisé classique, où tous les échantillons sont étiquetés, et l'apprentissage non supervisé, où aucune classe n'est attribuée. Les méthodes d'apprentissage semi-supervisé étendent les techniques d'apprentissage choisi, soit non supervisé ou supervisé, pour inclure des informations supplémentaires typiques de l'autre paradigme d'apprentissage.

2.3.3.4 Apprentissage par renforcement

L'apprentissage par renforcement fait référence à une classe de problèmes d'apprentissage automatique dans lesquels les machines learning essa
ient différentes situations afin de pouvoir déterminer quelles actions sont les plus utiles, et pas seulement recevoir des instructions sur les actions à appliquer, ce qui distingue cette méthode des autres techniques d'apprentissage.
Il est également considéré comme une forme d'apprentissage comportemental. L'algo-

rithme dans ce cas reçoit les informations en analysant les données pour pouvoir orienter l'utilisateur vers les meilleurs résultats. Dans ce type d'apprentissage, le système n'est pas entraîné à partir d'un ensemble de données, mais apprend par l'expérience et utilise les erreurs, par exemple pour les voitures autonomes, ce qui diffère des autres types d'apprentissage supervisé [17].

2.3.3.5 Deep Learning

Le deep learning (DL) est un type de machine learning qui utilise des réseaux de neurones artificiels pour modéliser et résoudre des problèmes complexes. Ces réseaux de neurones sont composés de couches de nœuds interconnectés, ou "neurones" (comme illustré dans la Figure 2.6), conçus pour traiter et analyser de grandes quantités de données d'entrée. Chaque neurone dans un réseau de deep learning reçoit une entrée des neurones de la couche précédente et utilise cette entrée pour effectuer une opération mathématique, appelée fonction d'activation, qui produit une sortie. Cette sortie est ensuite transmise à la couche suivante de neurones, et le processus est répété jusqu'à ce que la sortie finale soit produite.

Les algorithmes de deep learning peuvent être formés de manière supervisée ou non supervisée, en fonction du type de problème à résoudre.

L'apprentissage supervisé consiste à fournir à l'algorithme des données d'entraînement étiquetées, où la sortie correcte ou "étiquette" est connue pour chaque entrée. L'algorithme apprend alors à faire des prévisions sur de nouvelles données non vues en se basant sur les modèles qu'il a identifiés dans les données d'entraînement.

L'apprentissage non supervisé, d'autre part, consiste à fournir à l'algorithme des données non étiquetées, et l'algorithme doit identifier des modèles ou des caractéristiques dans les données par lui-même.

Le deep learning est particulièrement adapté aux tâches qui impliquent la reconnaissance d'images et de la parole, le traitement du langage naturel et d'autres types de données non structurées. En raison de sa capacité à apprendre des modèles complexes, il obtient des résultats de pointe dans de nombreux domaines. Cependant, il nécessite également des ressources informatiques importantes et de grandes quantités de données étiquetées.

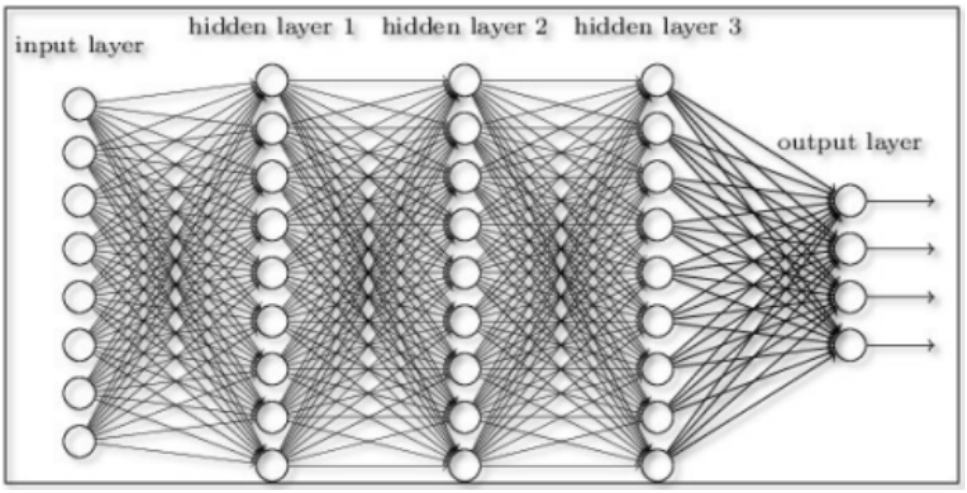

FIGURE 2.6 – Deep learning (réseaux neuronaux). [1]

2.4 Conclusion

En conclusion de ce chapitre, nous avons exploré les concepts essentiels du Big Data et du Machine Learning (ML). Nous avons défini le Big Data comme étant un ensemble de données volumineux et complexes, et nous avons également montré comment il a évolué au fil du temps. De plus, nous avons examiné les caractéristiques clés du Big Data, telles que la variété, la vélocité et la véracité, et nous avons discuté de la structuration et de la gestion de ces données massives, ainsi que de leur importance pour extraire des informations utiles.

En ce qui concerne le Machine Learning, nous avons défini cette technologie comme un sous-domaine de l'intelligence artificielle qui permet aux ordinateurs d'apprendre sans être explicitement programmés. Nous avons examiné les différents types de Machine Learning, y compris l'apprentissage supervisé, non supervisé, semi-supervisé, renforcé et Deep Learning.

En résumé, ce chapitre met en évidence le rôle crucial du Big Data et du Machine Learning dans l'analyse des données massives et la résolution de problèmes complexes. Leur popularité ne cesse de croître et nous pouvons nous attendre à ce que leur utilisation continue de se développer alors que les entreprises cherchent à exploiter pleinement le potentiel de la transformation numérique.

Dans le chapitre suivant, nous nous concentrerons sur les avancées récentes de la recherche dans les techniques de Machine Learning appliquées à l'analyse du Big Data.