Chapitre I : Introduction aux sciences de données

I. Introduction

La science des données est une discipline, qui n'est certes pas nouvelle, mais qui a pris une grande ampleur dans les dernières années. Comme son nom l'indique, la science des données ne concerne pas un domaine en particulier, mais elle s'intéresse plutôt à tous les aspects liés aux données notamment la collecte, le stockage, l'analyse, le transfert, le nettoyage, le filtrage, l'anonymisation, le cryptage, la visualisation, etc. Bien que les données et leur traitement existaient bien avant l'émergence du concept de science des données, le volume très important de données disponibles présentement (données massives - big data) et leur démocratisation, ainsi que l'augmentation des puissances de calcul nous amènent vers une nouvelle ère d'analyse des données qui n'était certainement pas possible à l'époque.

II. La science des données

La science des données permet d'extraire des connaissances à partir de données en utilisant diverses méthodes, algorithmes et processus scientifiques. Cette technologie nous permet de traduire un problème commercial en projet de recherche, puis de le traduire en solution pratique. Ce champ a pour but principal d'identifier des tendances, des motifs, des connexions et des corrélations dans les larges ensembles de données. La science des données englobe une large variété d'outils et de techniques telles que la programmation informatique, l'analyse prédictive, les mathématiques, les statistiques ou l'intelligence artificielle. Désormais, la Data Science inclut aussi les algorithmes de l'apprentissage automatique.

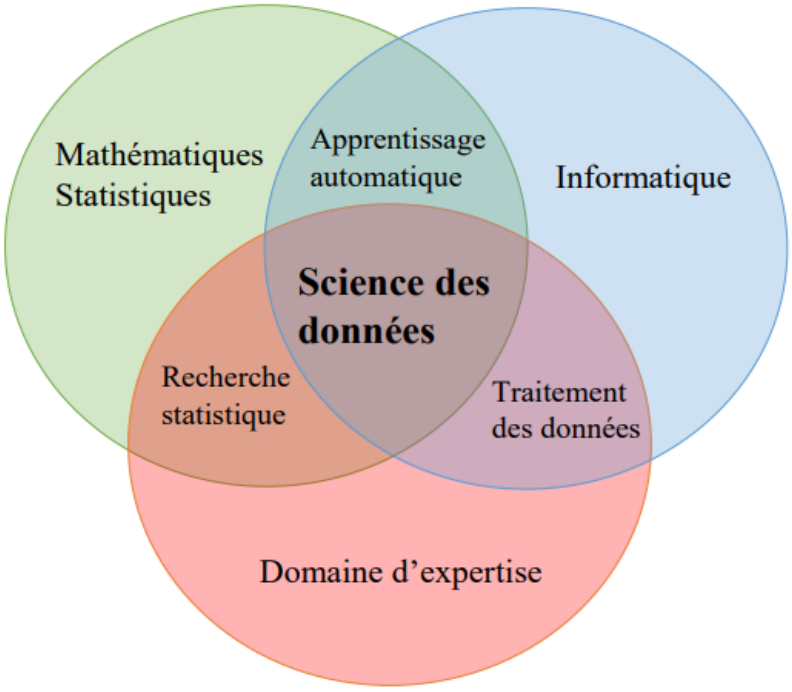

Figure 1 La science des données

III. Origines et enjeux de la science des données

III.1. Origines de la Science des Données

1. **Évolution des Données** :
    - **1970s-1980s** : Début de l'informatique avec la création de bases de données et l'émergence du SQL. Les entreprises ont commencé à collecter et stocker de grandes quantités de données.
    - **1990s** : Croissance d'Internet et explosion des données disponibles. L'analyse de données devient cruciale pour les entreprises.
2. **Statistiques et Mathématiques** :
    - Les concepts statistiques, tels que les tests d'hypothèses et la régression, ont servi de fondement à l'analyse des données.
    - Le développement d'outils mathématiques pour l'optimisation et l'analyse a permis des avancées dans les algorithmes d'apprentissage automatique.
3. **Apprentissage Automatique** :
    - L'apprentissage automatique est né de la convergence entre l'intelligence artificielle et les statistiques. Les premiers algorithmes ont été développés dans les années 1950, mais leur adoption massive a eu lieu dans les années 2000 grâce à la disponibilité de grandes quantités de données et de puissantes capacités de calcul.
4. **Big Data** :
    - Avec l'essor du big data dans les années 2010, la science des données s'est professionnalisée. Les entreprises ont commencé à embaucher des data scientists pour exploiter les grandes quantités de données non structurées.

III.2. Enjeux de la Science des Données

1. **Prise de Décision Basée sur les Données** :
    - Utilisation des analyses pour guider les décisions stratégiques, améliorer les performances et optimiser les opérations.
2. **Compétitivité Économique** :
    - Les entreprises qui adoptent des approches basées sur les données peuvent mieux comprendre leurs clients, anticiper les tendances du marché et innover plus rapidement.
3. **Éthique et Protection de la Vie Privée** :
    - Les questions éthiques autour de l'utilisation des données, de la confidentialité et des biais algorithmiques sont de plus en plus importantes. La transparence et l'éthique dans la collecte et l'analyse des données sont essentielles.
4. **Qualité des Données** :
    - La qualité des données est cruciale pour des analyses fiables. Les entreprises doivent investir dans des pratiques de gestion des données pour garantir leur intégrité.

5. **Interdisciplinarité** :
    - La science des données nécessite des compétences variées (statistiques, programmation, expertise métier). La collaboration entre différentes disciplines est essentielle pour tirer le meilleur parti des données.
6. **Innovations Technologiques** :
    - L'essor de l'intelligence artificielle, du machine learning et des outils de big data change constamment le paysage des sciences des données, ouvrant de nouvelles opportunités et défis.

IV. Facettes et types de données

La data science est une démarche empirique qui se base sur des données pour apporter une réponse à des problèmes. Une donnée est « le résultat d'une observation faite sur une population ou sur un échantillon». Une donnée est donc un nombre, une caractéristique, qui m'apporte une information sur un individu, un objet ou une observation. Par exemple, 33 est un nombre sans intérêt, mais si quelqu'un vous dit « J'ai 33 ans », 33 devient une donnée qui vous permettra d'en savoir un peu plus sur lui

Généralement, on lie les données à des variables parce que le nombre/la caractéristique varie si on observe plusieurs objets/individus/observations. En effet, si on s'intéresse à l'âge de tous les lecteurs d'un livre, on sait qu'il est défini par un nombre compris entre, disons, 15 et 90, et qu'il variera d'un lecteur à l'autre. Ainsi, si l'on nomme Xâge la variable « âge du lectorat », les données mesurant cet âge sont égales à x1âge, x2âge, …, xmâge, où x1âge est l'âge du lecteur 1, x2âge l'âge du lecteur 2, et ainsi de suite jusqu'à xmâge, où m représente le nombre total de lecteurs.

Tel est le matériau brut que va manipuler le data scientist : des variables exprimées concrètement par des données et qui lui permettent de décrire un ensemble d'objets/individus/observations. Ces données peuvent prendre diverses formes que nous allons désormais détailler. Les principaux types de données. On distingue généralement les données quantitatives des données qualitatives. Les données quantitatives sont des valeurs qui décrivent une quantité mesurable, sous la forme de nombres sur lesquels on peut faire des calculs (moyenne, etc.) et des comparaisons (égalité/ différence, infériorité/supériorité, etc.). Elles répondent typiquement à des questions du type « combien ». On fait parfois la différence entre :

• les données quantitatives continues, qui peuvent prendre n'importe quelle valeur dans un ensemble de valeurs : la température, le taux de chômage.

• **L**es données quantitatives discrètes, qui ne peuvent prendre qu'un nombre limité de valeurs dans un ensemble de valeurs : le nombre d'enfants par famille, le nombre de pièces d'un logement, etc. Les données qualitatives décrivent quant à elles des qualités ou des caractéristiques. Elles répondent à des questions de la forme « quel type » ou « quelle catégorie ». Ces valeurs ne sont plus des nombres, mais un ensemble de modalités. On ne peut pas faire de calcul sur ces valeurs, même dans l'éventualité où elles prendraient l'apparence d'une série numérique. Elles peuvent toutefois être comparées entre elles et éventuellement triées. On distingue :

• Les données qualitatives nominales (ou catégorielles), dont les modalités ne peuvent être ordonnées. Par exemple : la couleur des yeux (bleu, vert, marron, etc.), le sexe (homme, femme),

• Les données qualitatives ordinales, dont les modalités sont ordonnées selon un ordre « logique ». Par exemple : les tailles de vêtements (S, M, L, XL), le degré d'accord à un test d'opinion (fortement d'accord, d'accord, pas d'accord, fortement pas d'accord). Le tableau 1 résume ces différents types de données ainsi que les opérations qu'ils supportent.

Tableau 1 . Les opérations supportées par chaque type de données

Type de données	Opérations supportées
Quantitatives continues	Calculs, égalité/différence, infériorité/supériorité
Quantitatives discrètes	Calculs, égalité/différence, infériorité/supériorité
Qualitatives nominales	Égalité/différence
Qualitatives ordinales	Égalité/différence, infériorité/supériorité

V. Comment fonctionne la science des données ?

La Data Science couvre une large variété de disciplines et de champs d'expertise. Son but reste toutefois de donner du sens aux données brutes. Pour y parvenir, les Data Scientists doivent posséder des compétences en ingénierie des données, en mathématiques, en statistique, en informatique et en Data Visualisation. Ces compétences leurs permettront de parcourir les vastes ensembles de données brutes pour en dégager les informations les plus pertinentes et les communiquer aux décideurs de leurs organisations. Les Data Scientists exploitent également l'intelligence artificielle, et plus particulièrement le Machine Learning (Apprentissage Mathématiques Statistiques Informatique Domaine d'expertise Apprentissage automatique Traitement des données Recherche statistique Science des données automatique) et le Deep Learning (Apprentissage profond). Ces technologies sont utilisées pour créer des modèles et réaliser des prédictions en utilisant des algorithmes et diverses techniques. La démarche globale des sciences des données consiste à collecter les données, préparer les données, concevoir un modèle prédictif, visualiser les résultats, optimiser le modèle (calibration), déploiement, industrialisation.

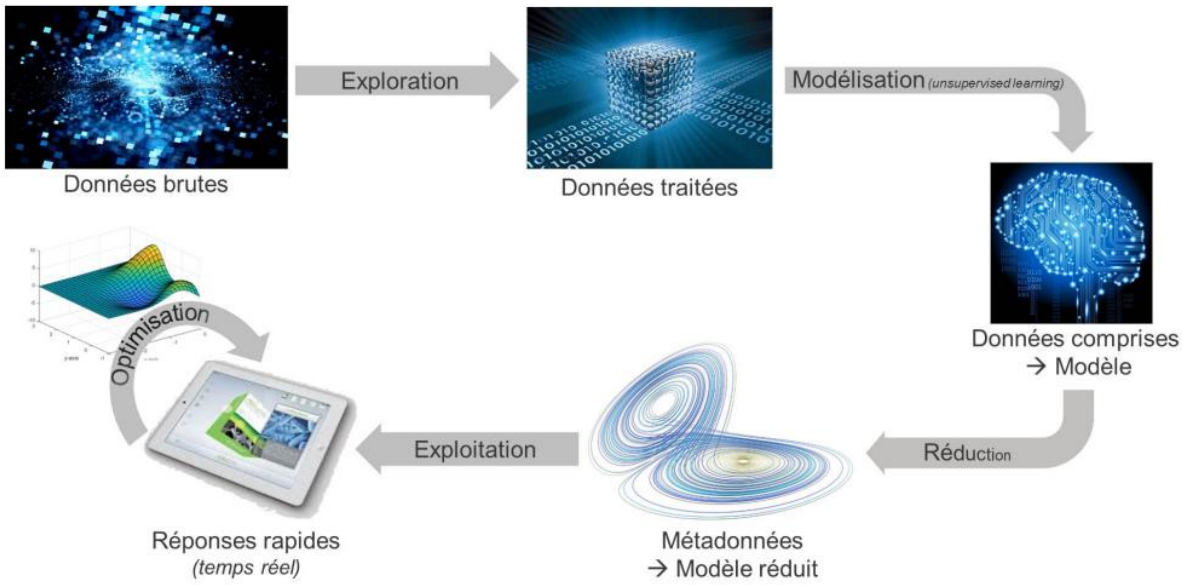

Figure 2 Cycle global de sciences des données

La figure 2 présente le contexte général du cycle de sciences de données. Les différentes étapes présentées :

— On dispose initialement de donnée brute sur lesquelles on dispose de plus ou moins de connaissances ;

— Une première étape consiste à explorer ces données, i.e. à essayer de mieux les appréhender, les connaître pour les transformer en données traitées, i.e. en données faisant sens pour nous ;

— On peut alors entreprendre de modéliser ces données pour les transformer en données comprises ce qui, pour nous, signifie que ces données peuvent être représentées par un modèle. Pour les gens du calcul, c'est souvent directement ici que le travail commence, par la création du modèle.

— Un modèle pouvant être volumineux, il peut être nécessaire de le réduire, i.e. de le décrire à l'aide d'un nombre réduit de variables que l'on appelle métadonnées et qui, elles, constituent les variables du modèle réduit ;

— enfin, disposant d'un modèle réduit, i.e. capable de fournir une bonne approximation du modèle initial dans un temps court, celui-ci peut être exploité à des fins d'optimisation par exemple, ou simplement pour répondre en temps réel dans d'autres applications.

VI. Cas d'usage et domaines d'application

Nombreux champs d'applications actuels et futurs Tous les domaines de la science : climat, physique, épidémiologie, medical, etc.

Secteur privé : Relation clients, marketing ciblé, fréquentation, etc.

Secteur public : amélioration des services, adaptation aux besoins, etc.

• Recherche Internet La recherche Google utilise la technologie de la science des données pour rechercher un résultat spécifique en une fraction de seconde

• Systèmes de recommandation Créer un système de recommandation. Par exemple, « amis suggérés » sur Facebook ou « vidéos suggérées » sur YouTube, tout est fait avec l'aide de la Data Science.

• Reconnaissance d'images et de parole La parole reconnaît des systèmes comme Google Assistant fonctionne sur la technique de la science des données. De plus, Facebook reconnaît votre ami lorsque vous téléchargez une photo avec lui.

• Comparaison de prix en ligne Les comparateurs de prix en ligne travaillent sur le mécanisme de la science des données en récupérant des données sur les sites Web concernés.

VII. L'écosystème du big data et la science des données

La science des données a été, dans les derniers temps, souvent liée aux données massives. Bien que la science des données existait avant l'apparition du terme données massives, le besoin de traiter des données massives et le développement rapide des infrastructures et outils de traitement des données massives font en sorte que la science des données émerge de plus en plus dans pratiquement tous les domaines. La science des données est liée aux données peu importe leur taille ou provenance.

Les données massives (Big Data) désigne à la fois la production de données massives et le développement de technologies capables de les traiter an d'en extraire des corrélations ou du sens . C'est dans les années 1990 que le terme Big Data (données massives) prend sa signification actuelle d'un dé technologique à relever pour analyser de grands ensembles de données, d'abord scientifiques,

mais de plus en plus souvent collectés au quotidien par divers moyens techniques. Le term Big Data (données massives) est lié à trois caractéristiques fondamentales souvent appelées VVV ou 3V (Volume, Variété et Vélocité) comme le montre la figure suivante.

Volume : décrit la quantité de données générées.

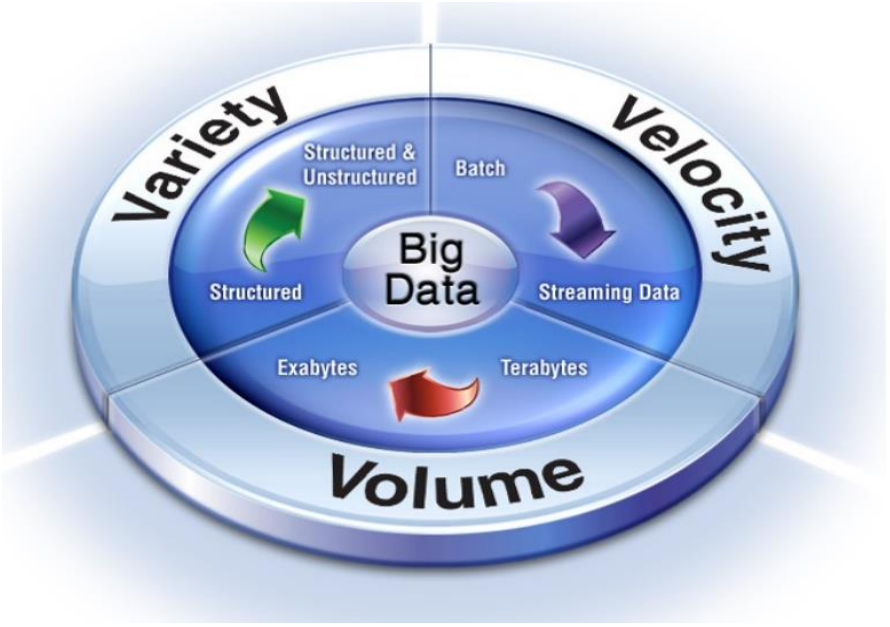

Figure 3 Les trois caractéristiques des données massives

Variété : décrit la diversité des types de données provenant de sources multiples.

Vélocité : décrit la fréquence à laquelle les données sont générées, capturées, partagées et traitées. Les trois caractéristiques des données massives engendrent, donc, beaucoup de dés techniques et computationnels pour les scientifiques de données an d'analyser ces données et d'en extraire des informations pertinentes en temps réel.