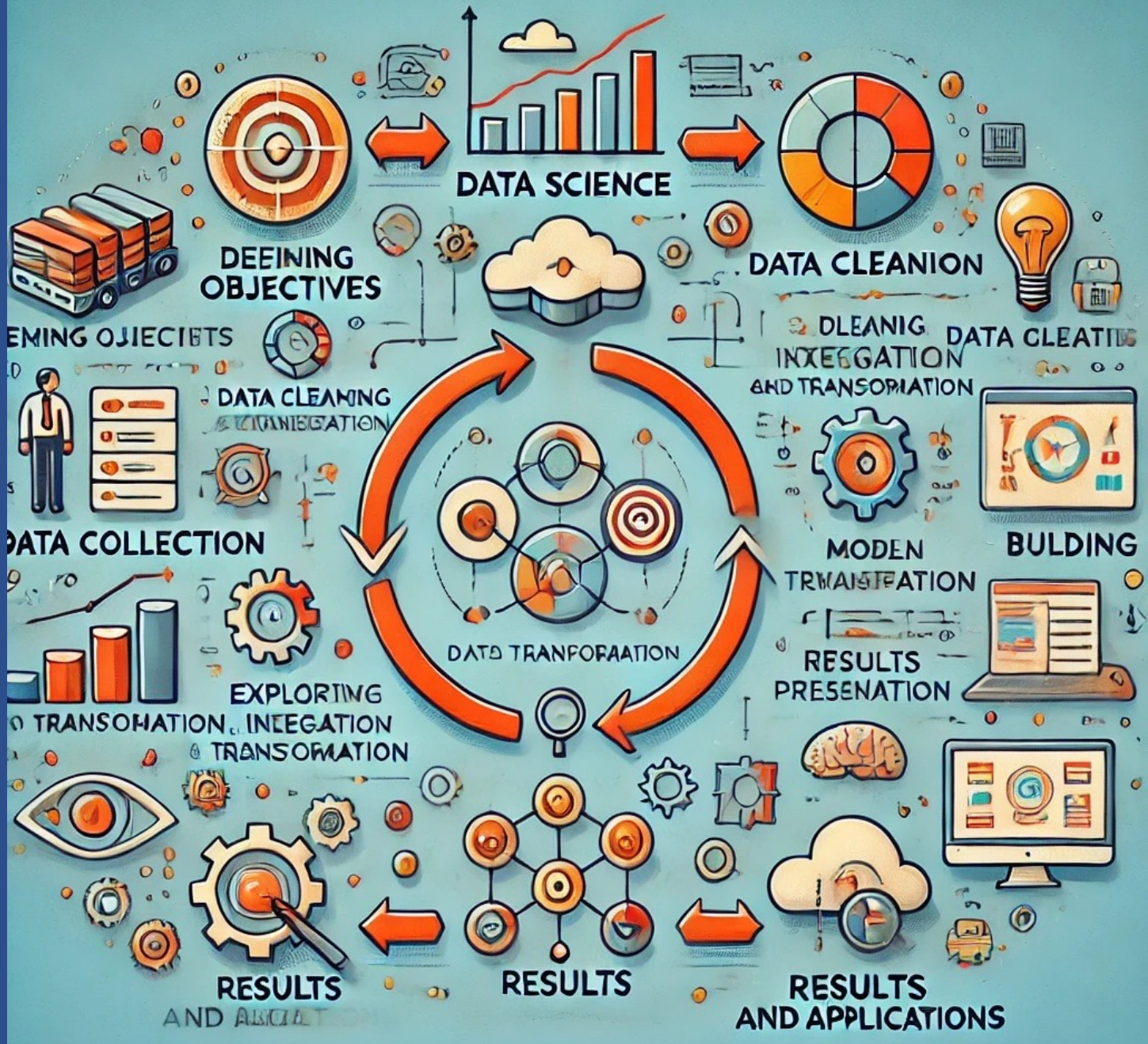
Chapitre 2

Le processus de science des données

Présenté par :
Dr. Bilal Dendani



جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY

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

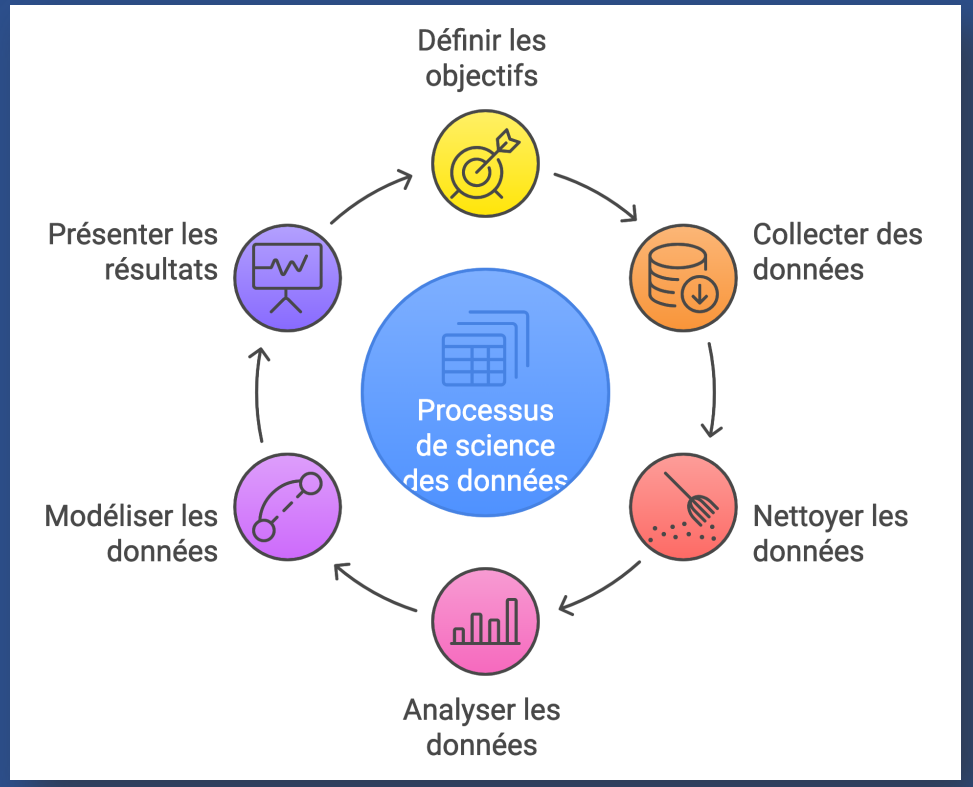
- Étape 1 : Définir les objectifs de recherche et créer une charte de projet
- Étape 2 : Récupération des données
- Étape 3 : Nettoyer, intégrer et transformer les données
- Étape 4 : Analyse exploratoire des données
- Étape 5 : Construire les modèles
- Étape 6 : Présentation des résultats et création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

Processus de la science de données

- Le processus de science des données consiste en une série d'étapes systématiques, depuis la définition des objectifs jusqu'à la présentation des résultats.

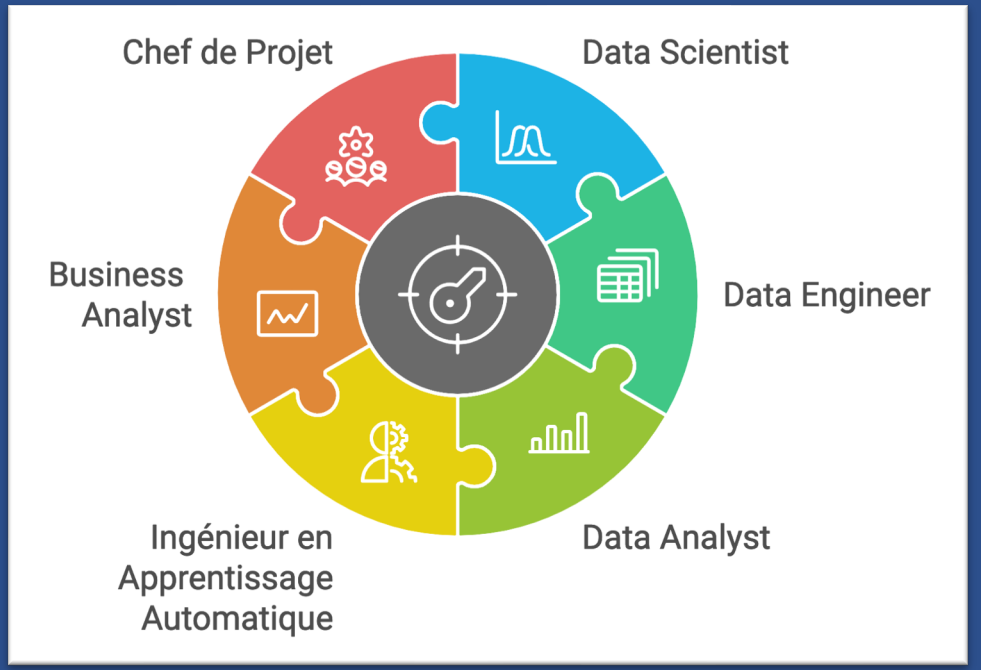
- L'ensemble structuré de ces étapes nous guident pour transformer des données brutes en informations exploitables

- Un processus bien structuré permet d'assurer la qualité, la cohérence et la fiabilité des résultats obtenus.



Rôles et Responsabilités dans un Projet de Science des Données

- Un projet de science des données est collaboratif et implique plusieurs rôles spécialisés.
- Chacun de ces rôles contribue à différentes étapes du projet pour garantir des analyses de qualité et des résultats exploitables.



Équipe de projet en science des données

Chef de projet

Coordonne les activités du projet, fixe les objectifs et surveille les ressources.

Data Scientist

Analyse les données et crée des modèles prédictifs pour extraire des insights.

Machine Learning Engineer

Implémente et déploie des modèles dans des environnements de production.

Data Engineer

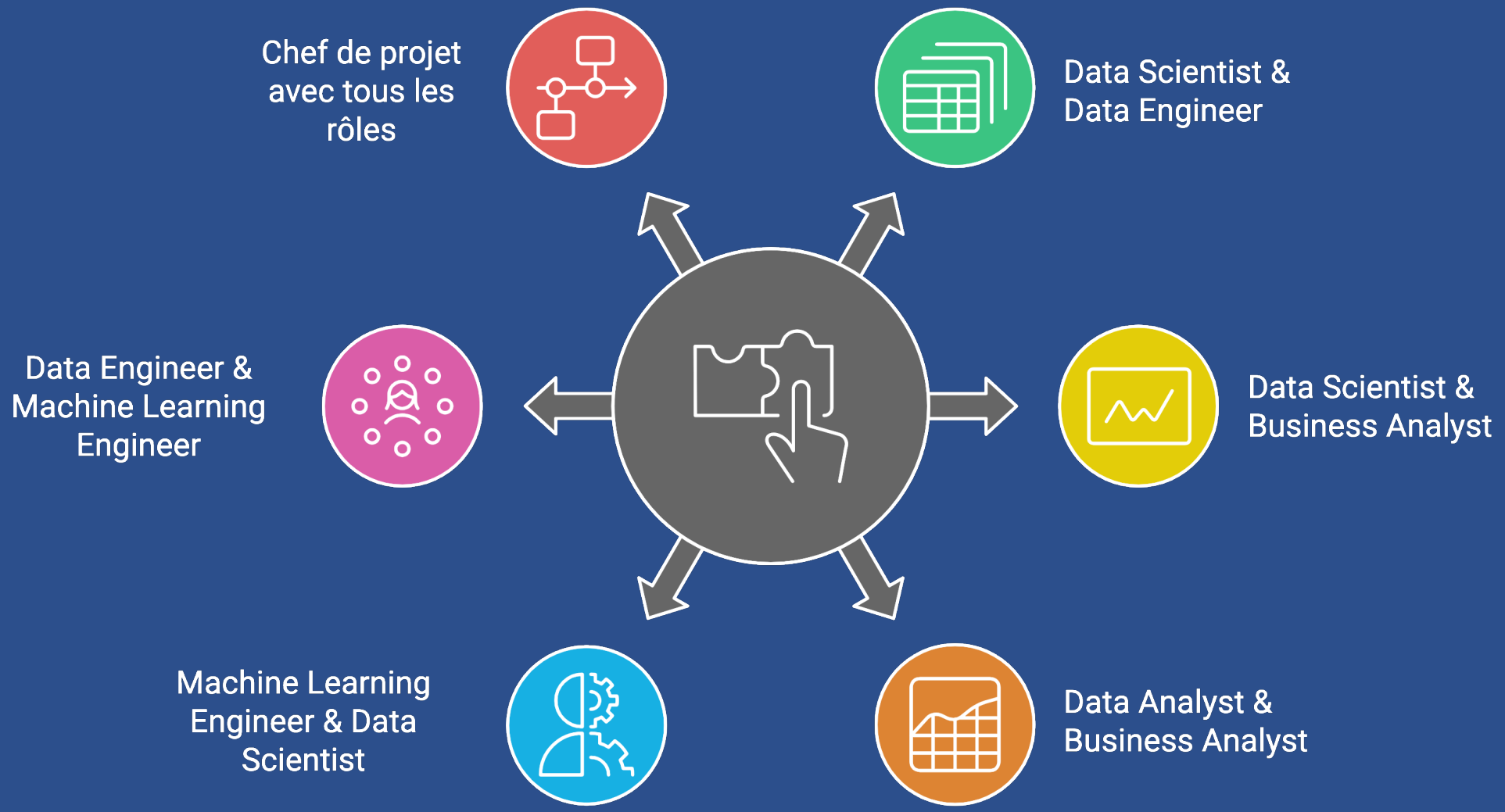
Gère l'infrastructure de données, collecte et transforme les données pour les rendre exploitables.

Data Analyst

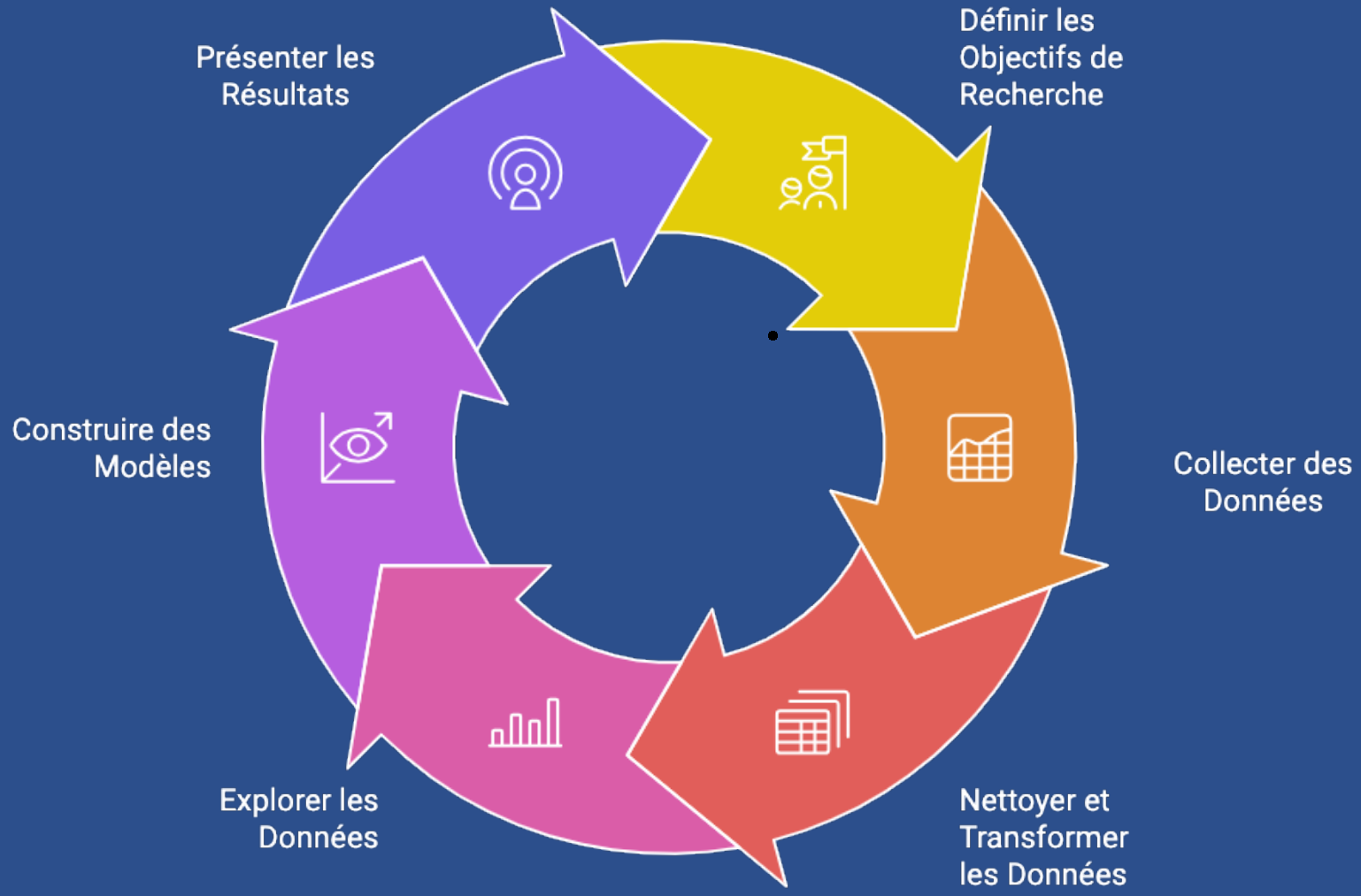
Analyse les données et crée des rapports pour aider à la prise de décision.

https://www.youtube.com/watch?v=CIdSwQqRKc4

Collaboration et intéraction entre les roles dans un projet de science de données



Cycle de vie d’un projet en science de données

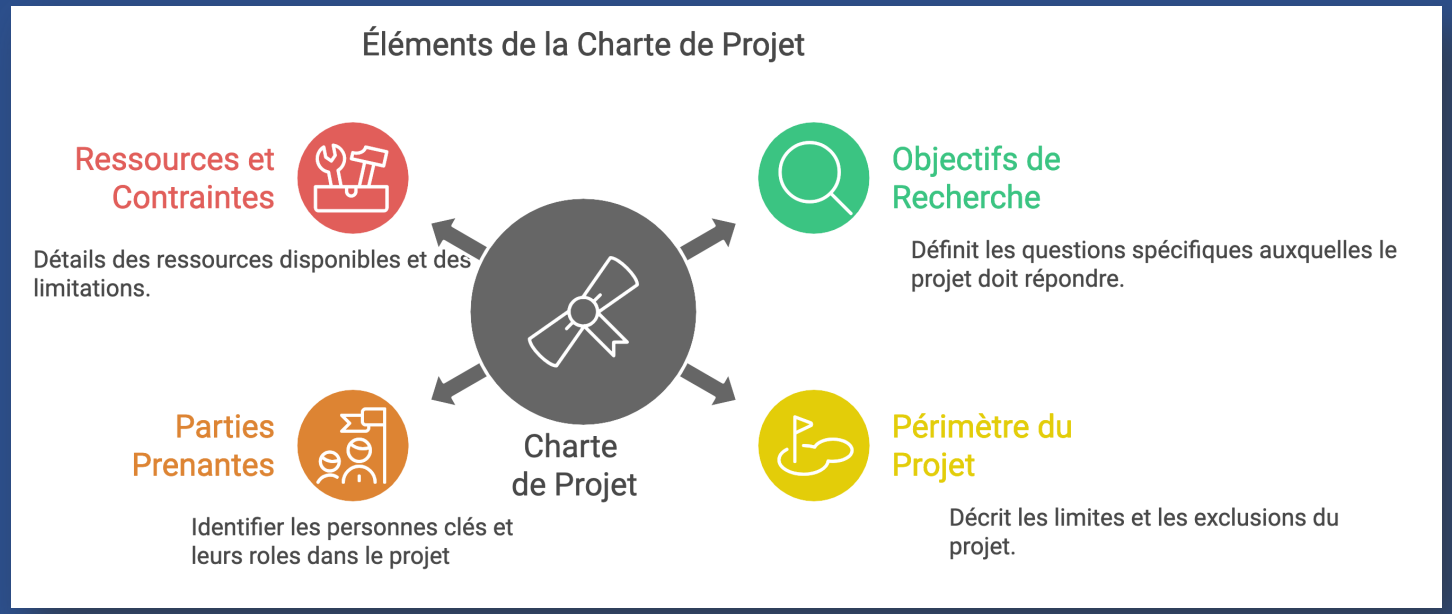


Les Étapes Clés du Cycle de Vie

1. Définir les Objectifs et Créer une Charte de Projet : Déterminer le problème à résoudre et clarifier les objectifs du projet.
2. Collecte des Données : Identifier et rassembler les données pertinentes, en provenance de sources internes et externes.
3. Nettoyage et Transformation des Données : Préparer les données en les nettoyant et en les transformant pour les rendre exploitables.
4. Analyse Exploratoire des Données : Comprendre la distribution et les relations dans les données pour orienter les analyses futures.
5. Modélisation et Construction des Modèles : Appliquer des algorithmes pour créer des modèles prédictifs ou des analyses avancées.
6. Présentation des Résultats : Communiquer les insights et proposer des actions concrètes basées sur les résultats du modèle.

Étape 1 - Définir les Objectifs de Recherche et Créer une Charte de Projet

1. Objectif de l'Étape
   - Définir la Direction du Projet : Identifier clairement ce que l'équipe cherche à accomplir.
   - Alignement des Parties Prenantes : S'assurer que tout le monde a une compréhension partagée des objectifs.
2. Élaboration de la Charte de Projet



Étape 2 - Récupération des Données

- Collecter les données nécessaires pour atteindre les objectifs définis dans la charte de projet.
- Les données des différents types et sources
  - Sources de Données :
    - Données internes : Bases de données internes de l'entreprise (ex. : historiques clients, ventes).
    - Données externes : Sources externes comme les réseaux sociaux, données publiques, API tierces.
    - Objets connectés (IoT) : Données générées par des capteurs et appareils connectés (ex. : suivi logistique, données environnementales).
- Types de Données Collectées :
    - Données structurées : Tableaux, bases de données relationnelles.
    - Données semi-structurées : Fichiers JSON, XML.
    - Données non structurées : Images, vidéos, documents texte.

Étape 3 - Nettoyer, Intégrer et Transformer les Données

- L’objectif de cette étape est d’ssurer que les données sont précises, complètes et prêtes pour l'analyse.
- Nettoyage des données :
  - Identification et traitement des valeurs manquantes, et suppression des doublons.
  - Correction des erreurs et des incohérences (ex : erreurs typographiques, formatage).
- Intégration des données :
  - Combinaison de différentes sources de données pour créer un ensemble de données unifié.
  - Harmonisation des formats et des structures (ex : joindre des tables, fusionner des fichiers).
- Transformation des données :
  - Normalisation ou standardisation des valeurs (ex : mise à l'échelle des données numériques).
  - Conversion des types de données si nécessaire (ex : transformation des dates).

Étape 4 - Analyse exploratoire des données

L'Analyse Exploratoire des Données (EDA) :

- l'EDA est une étape fondamentale pour comprendre la structure et les caractéristiques des données, et pour identifier des patterns ou anomalies.

- Objectifs Principaux de l'EDA
  - Identifier des tendances et des motifs dans les données.
  - Détecter des valeurs aberrantes (outliers) et des erreurs potentielles.
  - Comprendre les relations entre différentes variables.
- Techniques et Méthodes Utilisées
  - Statistiques descriptives : Moyenne, médiane, écart-type, etc.
  - Visualisations : Histogrammes, boîtes à moustaches (box plots), graphiques de dispersion (scatter plots), etc.
  - Analyse de corrélation : Identifier les relations linéaires entre variables (matrice de corrélation).

Étape 5 - Construire les Modèles

- Cette étape consiste à utiliser les données pour créer des modèles qui nous aident à répondre aux questions posées en début de projet.
- Types de Modèles :
  - Modèles supervisés : Régression, Classification (ex. : prédire le prix d'une maison ou classifier un client).
  - Modèles non supervisés : Clustering, Réduction de dimensions (ex. : segmenter des clients par comportements d'achat).
- Processus de Modélisation :
  - Sélection du modèle : Choisir le modèle adapté selon les données et les objectifs (ex. : régression pour prédire des valeurs continues).
  - Entraînement : Utiliser un sous-ensemble des données pour former le modèle.
  - Évaluation : Mesurer les performances avec des métriques comme la précision, le rappel, le RMSE, etc.
  - Ajustement et optimisation : Affiner le modèle via des techniques comme la validation croisée ou le tuning d'hyperparamètres.

Étape 6 - Présentation des Résultats et Création d’Applications

- Cette étape permet de Présenter les insights obtenus de manière claire et compréhensible pour les parties prenantes, et développer des applications concrètes.
- Visualisation des données : Utilisation de graphiques, tableaux et dashboards pour illustrer les insights.
- Outils : Outils de visualisation comme Tableau, Power BI, ou Matplotlib pour des visualisations Python.
- Clarté et Simplicité : S'assurer que les résultats sont présentés dans un langage non technique pour faciliter la prise de décision.

Références

- OpenAI. (2024). ChatGPT [AI language model]. Retrieved from https://chat.openai.com