Digital Bridge RDC

# Science des données: introduction

December 2022

# Définition de la science des données

- La science des données est le domaine qui combine des méthodes, des processus et des systèmes scientifiques pour extraire des connaissances et des informations à partir de données structurées et non structurées.

# Pourquoi est-il important de rester concentré sur la mission de l'entreprise tout au long d'un projet de science des données?

- **Amélioration de l'efficacité:** En restant concentré sur le problème de l'entreprise, vous pouvez éviter de perdre du temps et des ressources sur des activités qui ne contribuent pas directement à résoudre le problème. Cela peut aider à s'assurer que le projet est achevé en temps opportun et de manière rentable.
- **Meilleure prise de décision:** Rester concentré sur le problème métier permet de s'assurer que les résultats du projet de science des données sont pertinents et exploitables, ce qui peut conduire à une meilleure prise de décision.
- **Adhésion accrue des intervenants:** En restant concentré sur le problème métier, vous pouvez démontrer la valeur du projet de science des données aux parties prenantes et augmenter leur adhésion et leur soutien.
- **Amélioration des résultats du projet:** En fin de compte, le succès d'un projet de science des données se mesure à sa capacité à résoudre le problème métier qui l'a initié. En restant concentré sur le problème de l'entreprise, vous pouvez augmenter les chances d'obtenir un résultat positif.
- **En résumé :** il est important de se concentrer sur le problème métier, car cela permet de s'assurer que le projet de science des données reste sur la bonne voie, fournit des résultats significatifs et aboutit à un résultat positif.

# Composantes de la science des données

- Collecte et stockage des données
- Préparation et nettoyage des données
- Exploration et visualisation des données
- Modélisation des données et apprentissage automatique
- Communication et présentation des données

# Collecte et stockage des données

- La science des données repose sur des données, qui peuvent provenir de diverses sources telles que des bases de données, des capteurs et des médias sociaux.
- Les données doivent être collectées, stockées et organisées de manière à les rendre accessibles et utiles pour l'analyse.
- Attention aux données biaisées et aux questions éthiques, ainsi qu'à « GIGO »

# GIGO: Garbage in – Garbage Out


- GIGO, ou « Garbage In – Garbage Out », est un terme utilisé pour décrire le phénomène de l'utilisation de données de mauvaise qualité ou non pertinentes comme entrée, conduisant à une sortie imparfaite ou dénuée de sens.
- Dans le contexte de la science des données, GIGO est un concept important car il souligne l'importance d'utiliser des données pertinentes et de haute qualité dans le processus d'analyse et de modélisation des données.
- Si les données d'entrée sont de mauvaise qualité ou ne sont pas pertinentes pour le problème à résoudre, les résultats de l'analyse et de la modélisation seront probablement erronés ou trompeurs. Cela peut avoir de graves conséquences, car les décisions prises sur la base de résultats aussi erronés pourraient entraîner de mauvais résultats ou même des dommages.
- D'autre part, si les données d'entrée sont de haute qualité et pertinentes pour le problème à résoudre, les résultats de l'analyse et de la modélisation sont plus susceptibles d'être précis et utiles. Cela peut aider les organisations à prendre des décisions éclairées qui mènent à de meilleurs résultats et à l'amélioration de l'efficience et de l'efficacité de leurs opérations.
- En résumé, GIGO souligne l'importance d'utiliser des données pertinentes et de haute qualité dans le processus de science des données, car cela a un impact significatif sur l'exactitude et l'utilité des résultats.

# Collecte et stockage des données

- La science des données repose sur des données, qui peuvent provenir de diverses sources telles que des bases de données, des capteurs et des médias sociaux.
- Les données doivent être collectées, stockées et organisées de manière à les rendre accessibles et utiles pour l'analyse.
- Attention aux données biaisées et aux questions éthiques, ainsi qu'à « GIGO »

# Préparation et nettoyage des données

- Les données brutes sont souvent désordonnées et doivent être nettoyées et préparées pour l'analyse.
- Cela implique des tâches telles que l'imputation des valeurs manquantes, la détection et l'élimination des valeurs aberrantes et l'ingénierie des fonctionnalités.

# Exploration et visualisation des données

- L'exploration des données implique la compréhension des caractéristiques des données et l'identification de modèles et de relations.
- La visualisation est un outil important pour l'exploration des données car elle permet l'interprétation facile d'ensembles de données complexes.

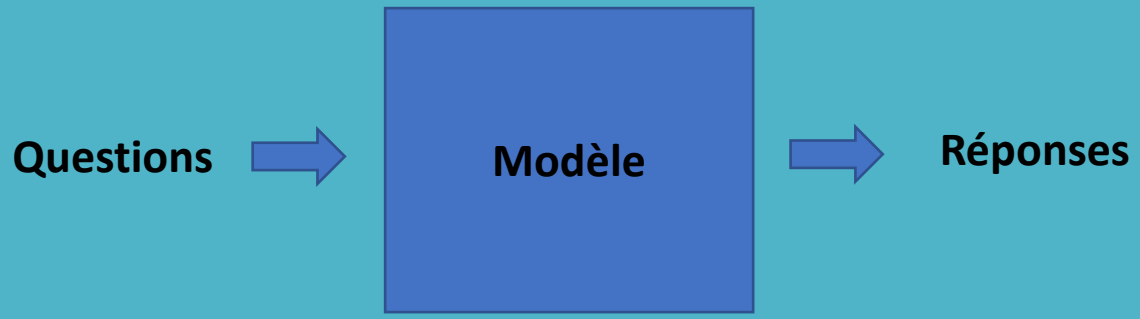
# Modélisation des données et apprentissage automatique

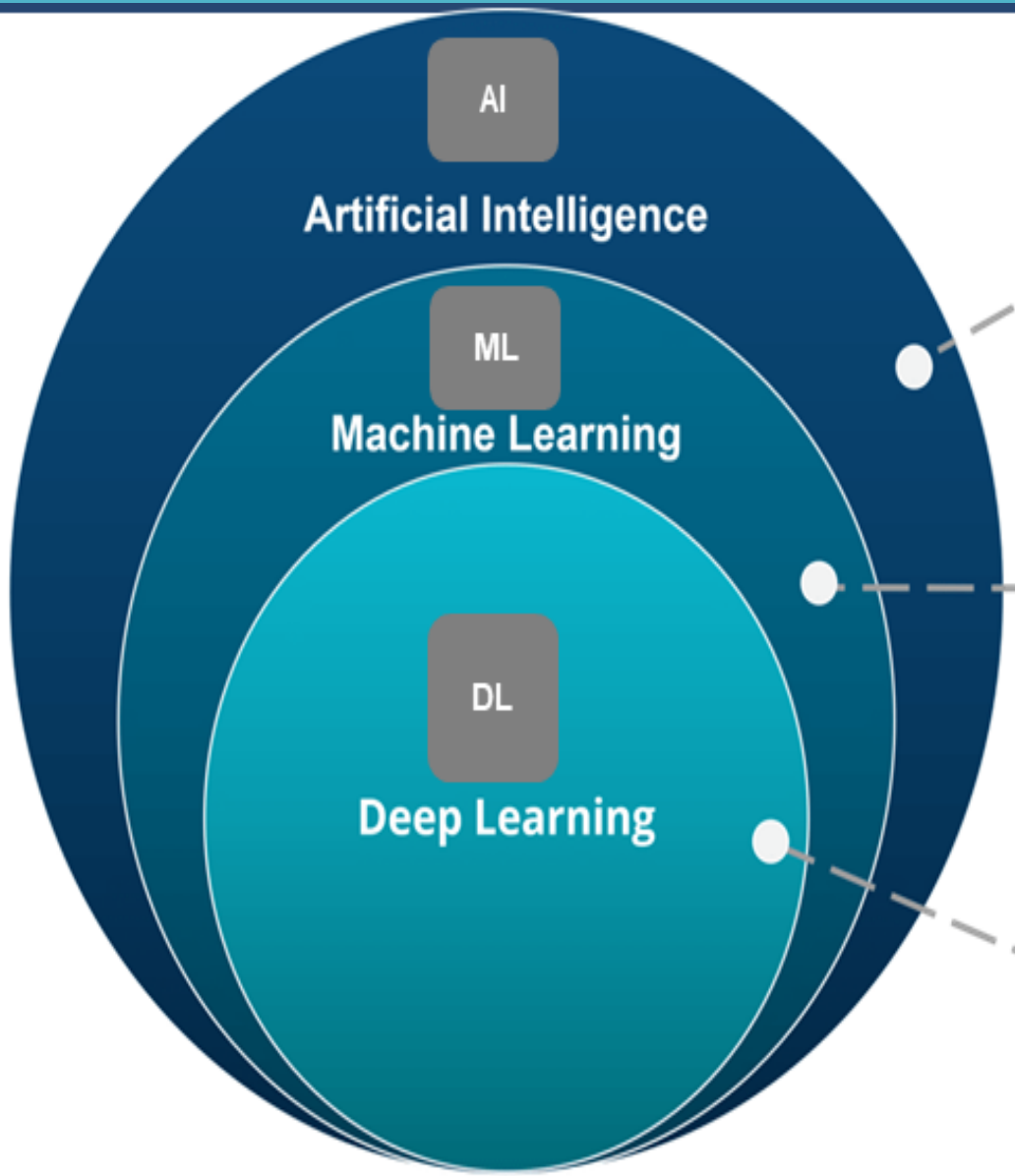
- La modélisation des données implique l'utilisation de techniques statistiques et mathématiques pour construire des modèles capables de faire des prédictions ou de classer des données.
- L'apprentissage automatique est un sous-ensemble de la science des données qui implique la formation d'algorithmes pour apprendre des modèles dans les données et faire des prédictions sans être explicitement programmé.

# Apprentissage automatique supervisé

**Phase 1 : Formation – élaboration d'un modèle**


**Phase 2 : Inférence – utilisation du modèle**






## ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

## MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience

## DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible

# Communication et présentation des données

- La science des données n’est pas seulement une question d’analyse, elle consiste également à communiquer les résultats de cette analyse à d’autres.
- Cela implique de créer des visualisations et des rapports clairs et efficaces qui peuvent être compris par un public non technique.

# Conclusion

- La science des données est un domaine multidisciplinaire qui implique un large éventail de compétences et de techniques pour extraire des informations à partir de données.
- Il joue un rôle crucial en aidant les organisations à prendre des décisions basées sur les données et à résoudre des problèmes complexes.

# Utilisation de Python dans les sciences des données

- Python est un langage de programmation populaire et puissant qui est largement utilisé dans le domaine de la science des données. Avec sa syntaxe flexible et son riche écosystème de bibliothèques et de frameworks, Python offre une multitude d'outils et de ressources pour travailler avec des données.
- Il existe de nombreuses bibliothèques et outils disponibles en Python qui peuvent être utilisés pour diverses tâches en science des données, y compris la manipulation de données, la visualisation et l'apprentissage automatique.
- Python est un langage de programmation populaire dans le domaine de la science des données.
- Il existe de nombreuses bibliothèques et outils disponibles en Python qui peuvent être utilisés pour diverses tâches en science des données, y compris la manipulation de données, la visualisation et l'apprentissage automatique.

# Numpy


- NumPy est une bibliothèque fondamentale pour le calcul scientifique en Python. Il fournit des outils puissants pour travailler avec des tableaux et des matrices de données, y compris des fonctions pour les opérations mathématiques, l'algèbre linéaire et la génération de nombres aléatoires.
- **référence:** https://numpy.org/
- **installer avec:** pip install numpy
- **Importer avec:** import numpy as np
- **tuyaux:** http://datacamp-community-prod.s3.amazonaws.com/ba1fe95a-8b70-4d2f-95b0-bc954e9071b0

# Pandas


- Pandas est une bibliothèque pour travailler avec des données tabulaires et rectangulaires en Python. Il fournit des structures de données et des fonctions pour manipuler, nettoyer et analyser des données, y compris des outils pour travailler avec des valeurs manquantes, regrouper et agréger des données, et fusionner et joindre des jeux de données.
- **référence:** https://pandas.pydata.org/
- **installer avec:** pip install pandas
- **importer avec:** import pandas as pd
- **tuyaux:** http://datacamp-community-prod.s3.amazonaws.com/d4efb29b-f9c6-4f1c-8c98-6f568d88b48f

# Scikit-learn


- Scikit-learn est une bibliothèque pour l'apprentissage automatique en Python. Il fournit une large gamme d'algorithmes et d'outils pour la formation, le test et l'évaluation des modèles d'apprentissage automatique, y compris la prise en charge de la classification, de la régression, du clustering et de la réduction de dimensionnalité.
- **référence:** https://scikit-learn.org/stable/
- **installer avec:** pip install scikit-learn
- **importer avec:** import sklearn
- **tuyaux:** http://datacamp-community-prod.s3.amazonaws.com/eb807da5-dce5-4b97-a54d-74e89f14266b

# Matplotlib


- Matplotlib est une bibliothèque puissante pour la visualisation de données en Python. Il fournit un large éventail de fonctions de traçage et d'options de personnalisation pour créer des visualisations statiques et interactives de données.
- **référence:** https://matplotlib.org/
- **installer avec:** pip install matplotlib
- **importer avec:** import matplotlib.pyplot as plt
- **tuyaux:** https://matplotlib.org/cheatsheets/_images/handout-beginner.png

# Seaborn


- Seaborn est une bibliothèque pour créer des graphiques statistiques en Python. Il est construit sur Matplotlib et fournit une interface de haut niveau pour créer des graphiques visuellement attrayants et informatifs, y compris des cartes thermiques, des graphiques en boîte et des graphiques de séries chronologiques.
- **référence:** https://seaborn.pydata.org/
- **installer avec:** pip install seaborn
- **importer avec:** import seaborn as sns
- **tuyaux:** http://datacamp-community-prod.s3.amazonaws.com/263130e2-2c92-4348-a356-9ed9b5034247

# Plotly


- Plotly est une bibliothèque permettant de créer des tracés et des visualisations interactifs basés sur le Web en Python. Il fournit un large éventail d'options de personnalisation et prend en charge plusieurs langages de programmation et plates-formes.
- **référence:** https://plotly.com/
- **installer avec:** pip install plotly
- **importer avec:** import plotly.express as px
- **tuyaux:** https://res.cloudinary.com/dyd911kmh/image/upload/v1668605954/Marketing/Blog/Plotly_Cheat_Sheet.pdf

# TensorFlow


- TensorFlow est une bibliothèque pour l'apprentissage profond en Python. Il fournit des outils et des bibliothèques pour la création, la formation et le déploiement de modèles d'apprentissage automatique, y compris la prise en charge des réseaux neuronaux et d'autres architectures avancées.
- **référence:** https://www.tensorflow.org/
- **installer avec:** pip install tensorflow
- **importer avec:** import tensorflow as tf
- **tuyaux:** https://github.com/kailashahirwar/cheatsheets-ai/blob/master/PDFs/Tensorflow.pdf

# Keras


- Keras est une bibliothèque de haut niveau pour la construction et la formation de réseaux neuronaux en Python. Il fournit une interface simple et intuitive pour définir et entraîner des modèles, et il peut être utilisé avec plusieurs backends, y compris TensorFlow, PyTorch et Theano.
- **référence:** https://keras.io/
- **installer avec:** pip install keras
- **importer avec:** from tensorflow import keras
- **tuyaux:** https://res.cloudinary.com/dyd911kmh/image/upload/v1660903348/Keras_Cheat_Sheet_gssmi8.pdf

# NLTK


- NLTK est une bibliothèque pour le traitement du langage naturel en Python. Il fournit des outils et des ressources pour travailler avec des données textuelles, y compris des fonctions de tokenisation, de stemming et de balisage, ainsi que des ensembles de données pour la formation et l'évaluation des modèles.
- **référence:** https://www.nltk.org/
- **installer avec:** pip install nltk
- **importer avec:** import nltk
- **tuyaux:** https://cheatography.com/murenei/cheat-sheets/natural-language-processing-with-python-and-nltk/pdf/

# Statsmodels


- Statsmodels est une bibliothèque pour la modélisation statistique et l'analyse de données en Python. Il fournit des fonctions pour estimer et tester des modèles statistiques, y compris la régression linéaire, l'analyse de séries chronologiques et les tests d'hypothèses.
- **référence:** https://www.statsmodels.org/stable/
- **installer avec:** pip install statsmodels
- **importer avec:** import statsmodels.api as sm
- **tuyaux:** https://www.statsmodels.org/dev/examples/index.html

Digital Bridge
R D C

# Autres sources

- https://www.kaggle.com/getting-started/78118
- https://www.utc.fr/~jlaforet/Suppl/python-cheatsheets.pdf