Data science

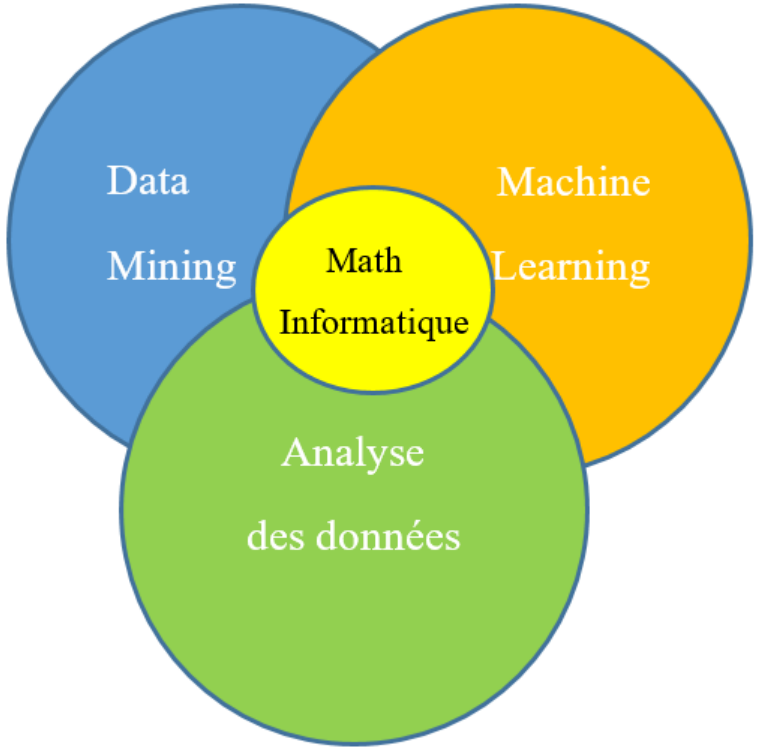
PRÉSENTÉ PAR DR-ING MAROUANE BEN HAJ AYECH

ENSEIGNANT À POLYTECHNIQUE INTERNATIONALE (PI)

Plan

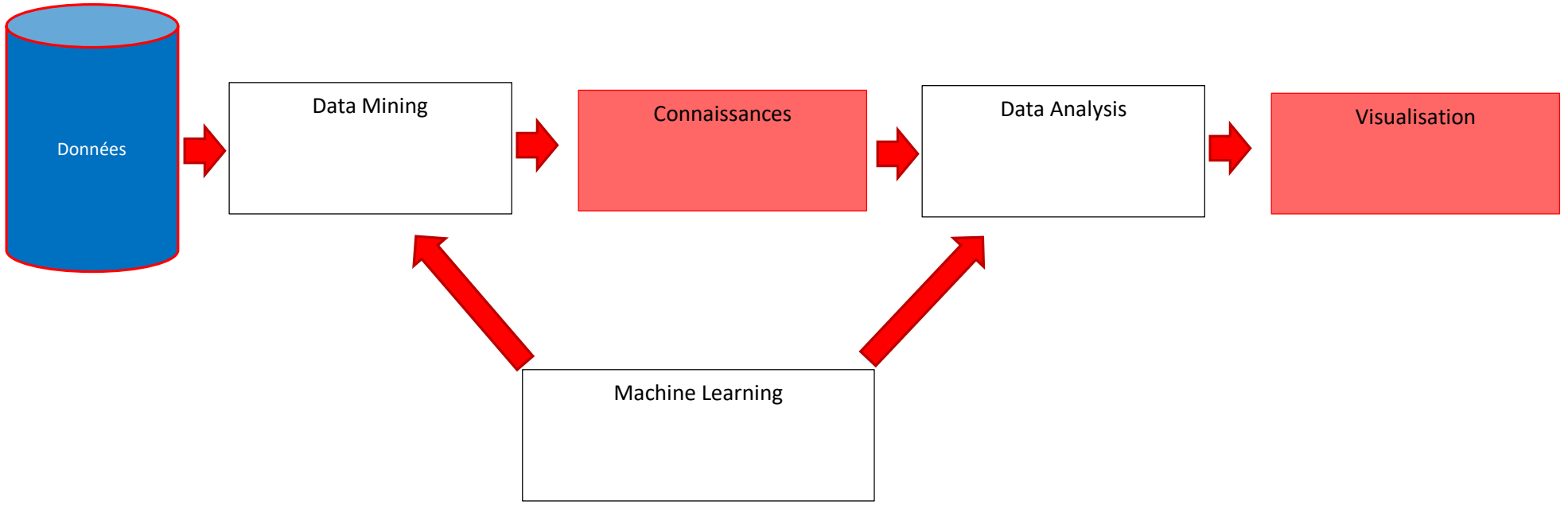
1. Présentation générale de **Data Science**
2. Programmation avec le langage **Python**
3. Calcul matriciel avec le module **Numpy**
4. Visualisation des données avec le module **Matplotlib**
5. Manipulation des données avec le module **Pandas**
6. Apprentissage Automatique avec **Scikit-learn**
7. Projet

Présentation générale de data science

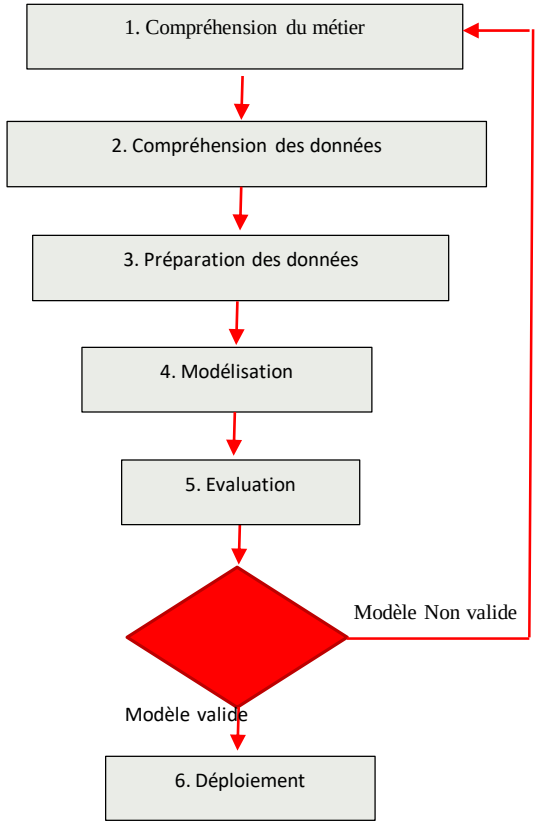


- Data science (**science des données)** est un domaine qui consiste à extraire des connaissances à partir des données (généralement massives) et les analyser afin de les visualiser.
- Ce domaine s'appuie sur des techniques qui proviennent des domaines tels que la statistique, traitement de signal etc mais surtout des techniques de la machine learning.

Présentation générale de data science



Présentation générale de data science Méthodologie



- Dans la dernière décennie (à partir de 2010), CRISP-DM, abréviation du Cross Industry Standard Process for Data Mining, est la méthodologie la plus adoptée par les experts de Data Science (essentiellement data miners) pour mettre en place des projets de Data Science.
- Cette approche est réalisée par un consortum des compagnies pionnières en DM tels que SPSS (Statistical Package for the Social Sciences de IBM), Mercedes benz, …
- Le processus CRISP-DM (schématisé dans la figure ci-dessous) consiste à suivre les 6 étapes indiquées pour réaliser le processus DM.

Programmation avec le langage Python

- **Python** est un langage:
  - Interprété
  - Orienté objet
  - Dynamique

Calcul matriciel avec le module Numpy

- **NumPy** est une extension du langage de programmation Python
- Il est destinée à manipuler des matrices ou tableaux multidimensionnels
- Il permet d'appliquer des fonctions mathématiques opérant sur ces tableaux.

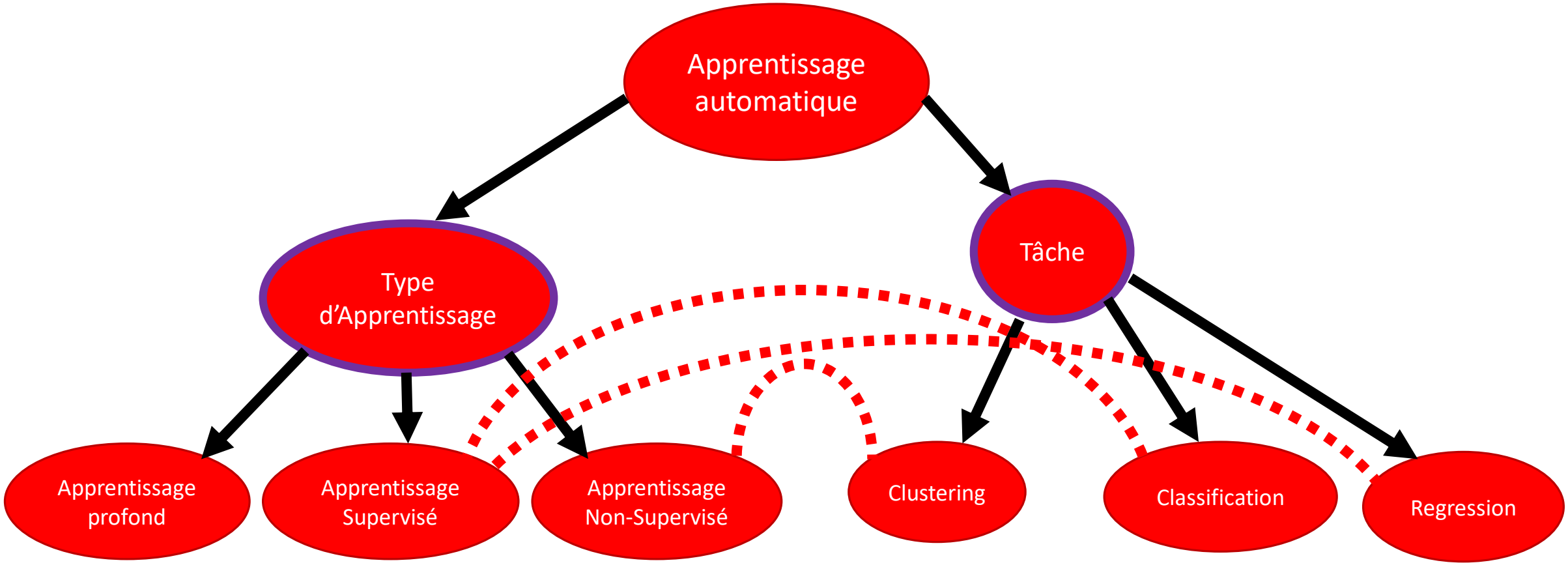
Visualisation des données avec le module Matplotlib

- **Matplotlib** est une extension du langage de programmation Python
- Il est destinée à tracer et visualiser des données sous formes de graphiques
- Deux fonctions de Matplotlib sont utilisées:
  - **Plot** : utilisée souvent pour visualiser des courbes
  - **Scatter** : utilisée souvent pour visualiser des nuages de points

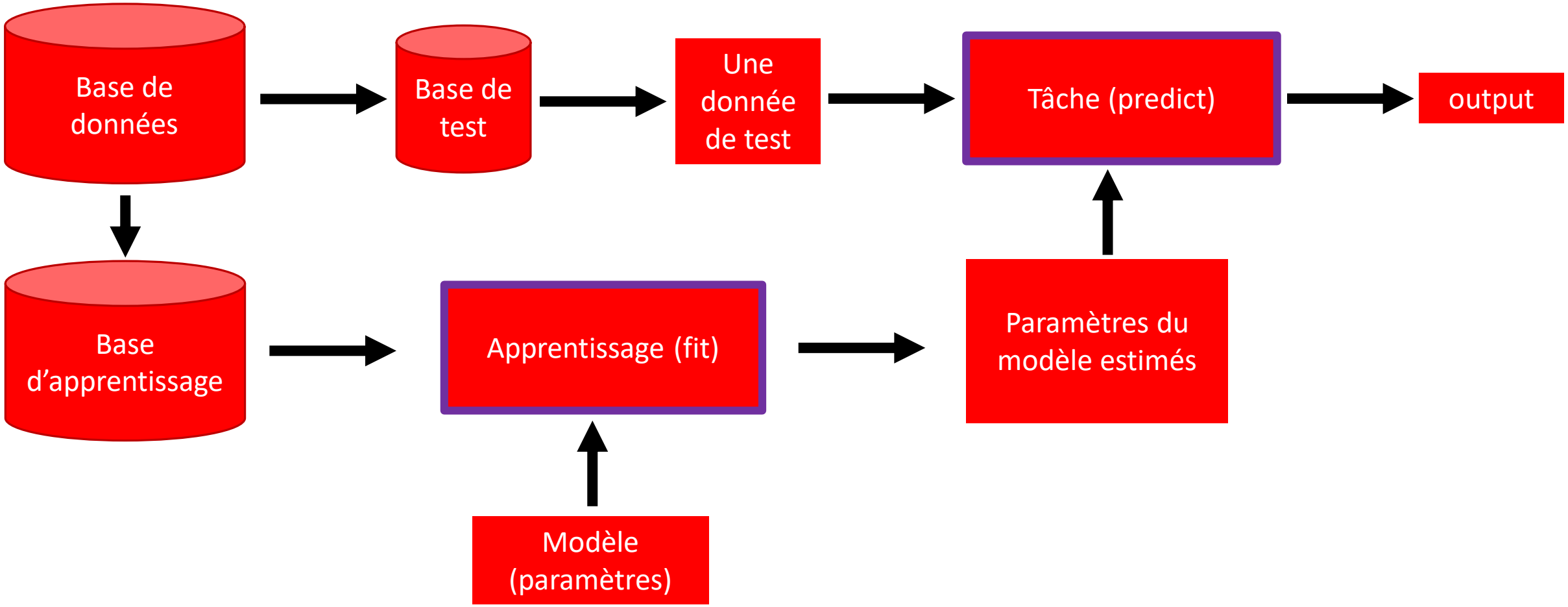
Manipulation des données avec le module Pandas

- **Pandas** est une extension du langage de programmation Python
- Il permet l'analyse de données
- Deux principales structures de données sont:
  - **Séries** : pour stocker des données selon une dimension (grandeur en fonction d'un index)
  - **DataFrames** : pour stocker des données selon 2 dimensions (lignes et colonnes)

Apprentissage Automatique avec Scikit-learn



Apprentissage Automatique avec Scikit-learn



Projet : CV mining