epsi

Data Science et Machine Learning

PLA-COMES Marine

Objectifs

- Comprendre la différence entre Data science et le machine Learning
- Concevoir un processus de valorisation de données en identifiant les besoins en données et leur faisabilité
- Mettre en forme les données pour les rendre exploitables par des algorithmes d'apprentissage machine
- Comprendre la différence entre ML supervisé et non supervisé

Objectifs

- Comprendre l'impact de la non-représentabilité des données sur les résultats obtenus
- Déterminer les facteurs influençant une variable à prédire.
- Méthodes prédictives : arbres de décision, forêts aléatoires, boosting, bayésien naïf.
- Méthodes non prédictives : K-means, Clustering hiérarchique, règles d'association.

Syllabus

I. Qu'est-ce que la Data science ?

II. Machine Learning non supervisé

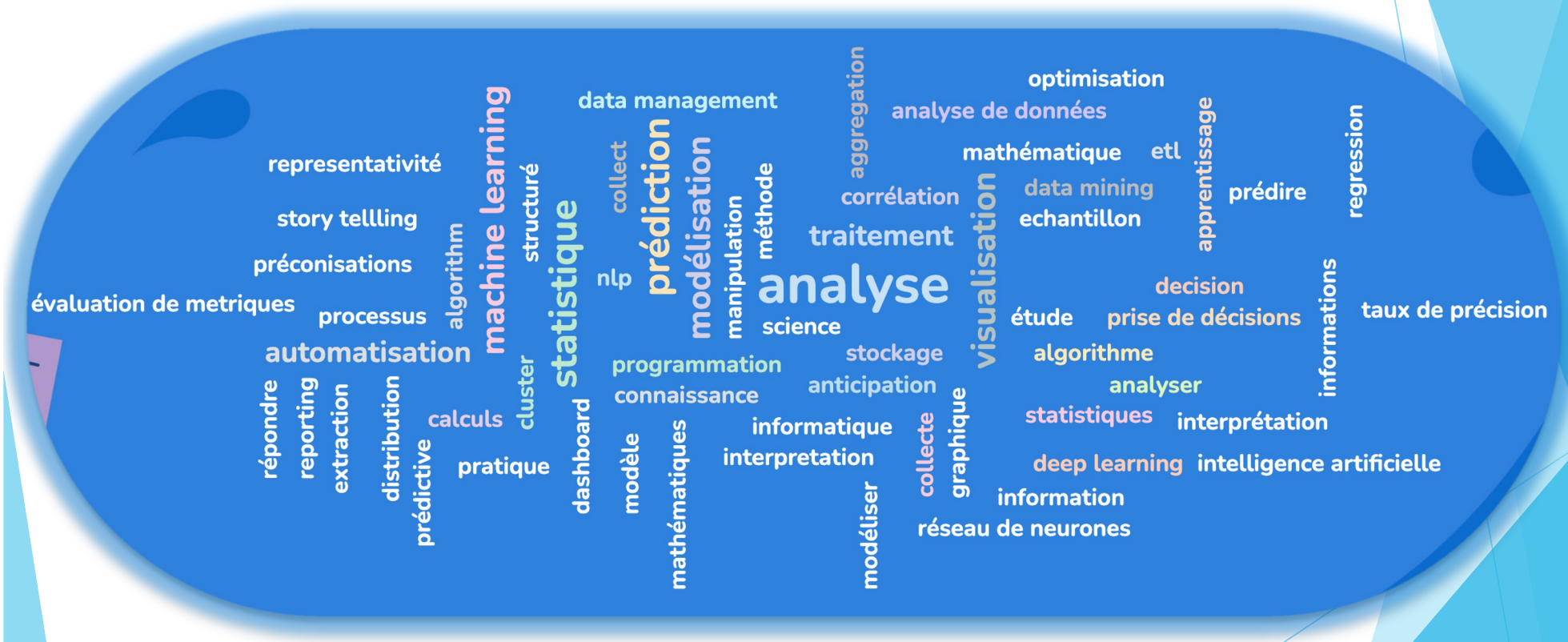
III. Machine Learning Supervisé

IV. Apprentissage par renforcement

epsi

Qu'est-ce que la Data science ?

La Data science – votre vision



La Data science

Pourquoi la Data science ?

- Elle combine des outils, des méthodes et des technologies visant à générer du sens à partir de données.
- Le monde croulent sous les données. Il existe une prolifération d'appareils qui peuvent collecter et stocker automatiquement des informations. Les systèmes en ligne et les portails de paiement capturent davantage de données dans les domaines de l'e-commerce, de la médecine, des finances et de tous les autres aspects de la vie humaine. Nous disposons de données textuelles, audio, vidéo et d'images en grande quantité.

Les données sont le pétrole du monde d’aujourd’hui.

À quoi sert la Data science ?

ANALYSE DESCRIPTIVE

ANALYSE PRÉDICTIVE

ANALYSE DIAGNOSTIQUE

ANALYSE PRESCRIPTIVE

Exemples de projet

- Détection de fraude
- Maintenance prédictive
- Moteurs de recommandations
- Détection d'objets
- Chatbots
- Analyse de sentiments

Histoire de la data science

- 60’s : synonyme de statistiques
- 70’s : évolution dans le milieu universitaire
- 90’s : formalisation du terme
  DS = conception + collecte + analyse
- +10’s : terme utilisé hors du milieu universitaire

Les 5V qui definissent le BIG DATA

VOLUME

DE DONNÉES
À TRAITER

VITESSE

DE COLLECTE,
D'ANALYSE, ET
D'EXPLOITATION EN
TEMPS RÉEL

VARIÉTÉ

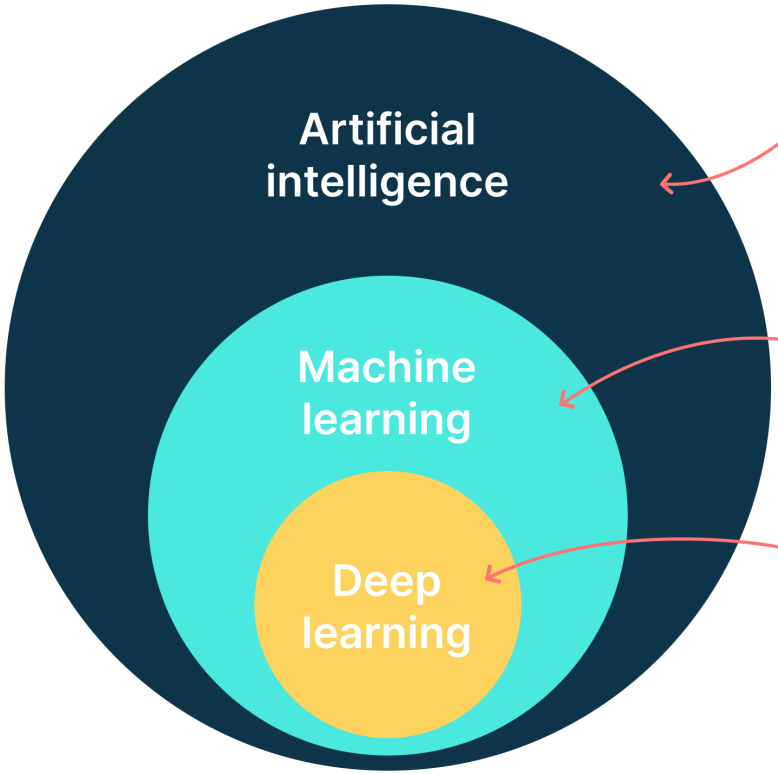
DES SOURCES ET
DES FORMATS

VALEUR

AJOUTÉE DES
DONNÉES POUR LES
CLIENTS ET LES
ENTREPRISES

VÉRACITÉ

DANS LA COLLECTE
DE DONNÉES
FIABLES



Any system that **leverages human capacities** for learning, perception and interaction, all at a level of complexity that ultimately supersedes our own abilities.

A subset of AI that involves programming system to **perform a specific task** without having to code rule-based instructions.

A subset of ML where systems can **learn hidden patterns from data** by themselves, combine them together and build much more efficient decision rules.

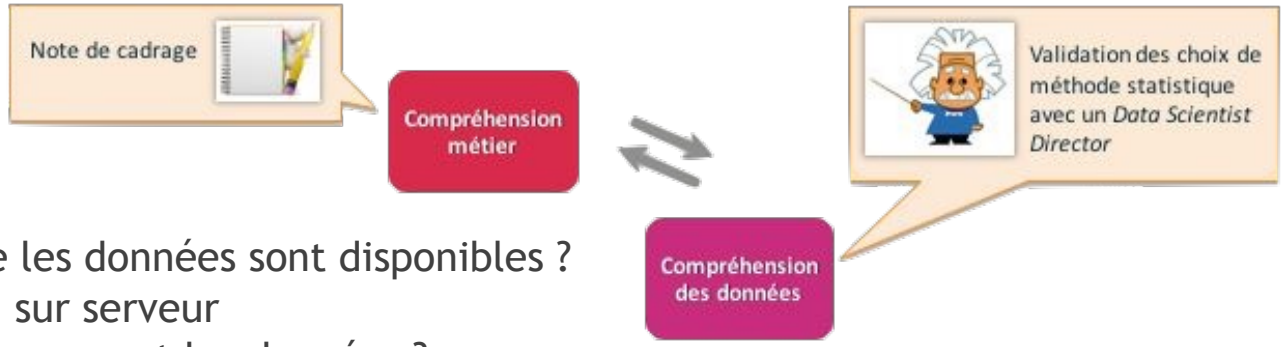
Méthodologie CRISP

Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)

- Apprendre le vocabulaire propre au domaine
- Comprendre les objectifs (définir les KPI, condition de réussite)
- Définir le périmètre (POC ? MVP ?)

Méthodologie CRISP

Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)

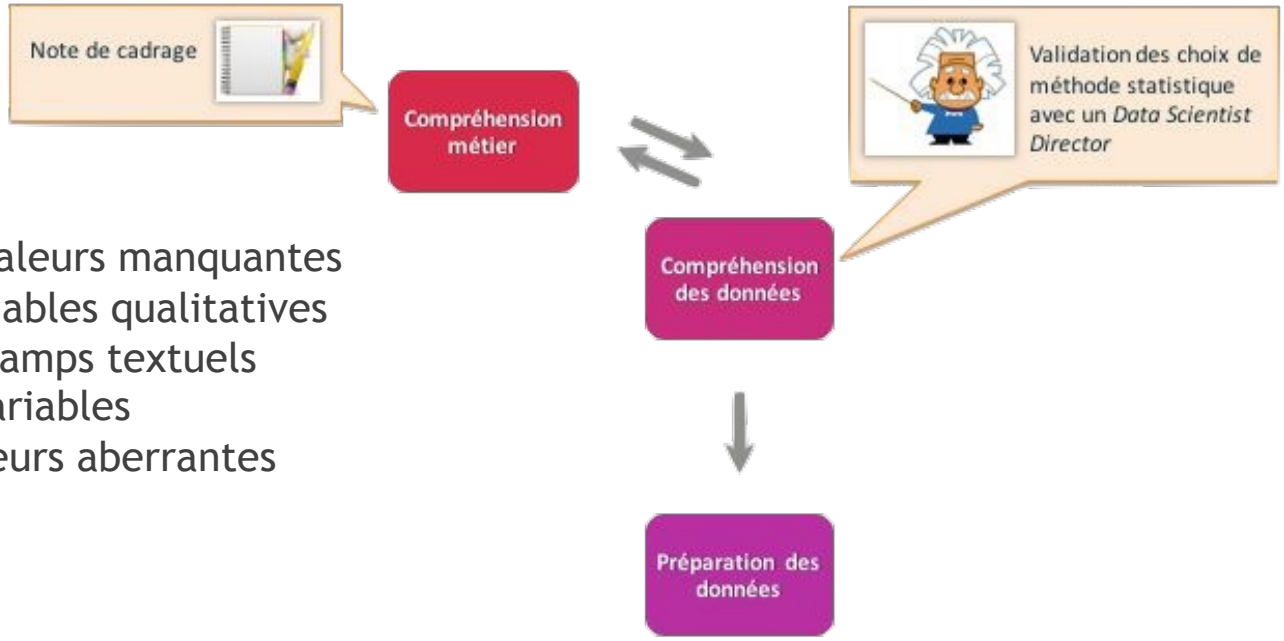


- Où est-ce que les données sont disponibles ?
  en local ou sur serveur
- Sous quelle forme sont les données ?
  base de données ou fichiers plats

- Que signifie la variable f_32 ?
- Valeurs d'un variable toujours supérieure à 100 ?
- Valeurs absente dans une variable ?

Méthodologie CRISP

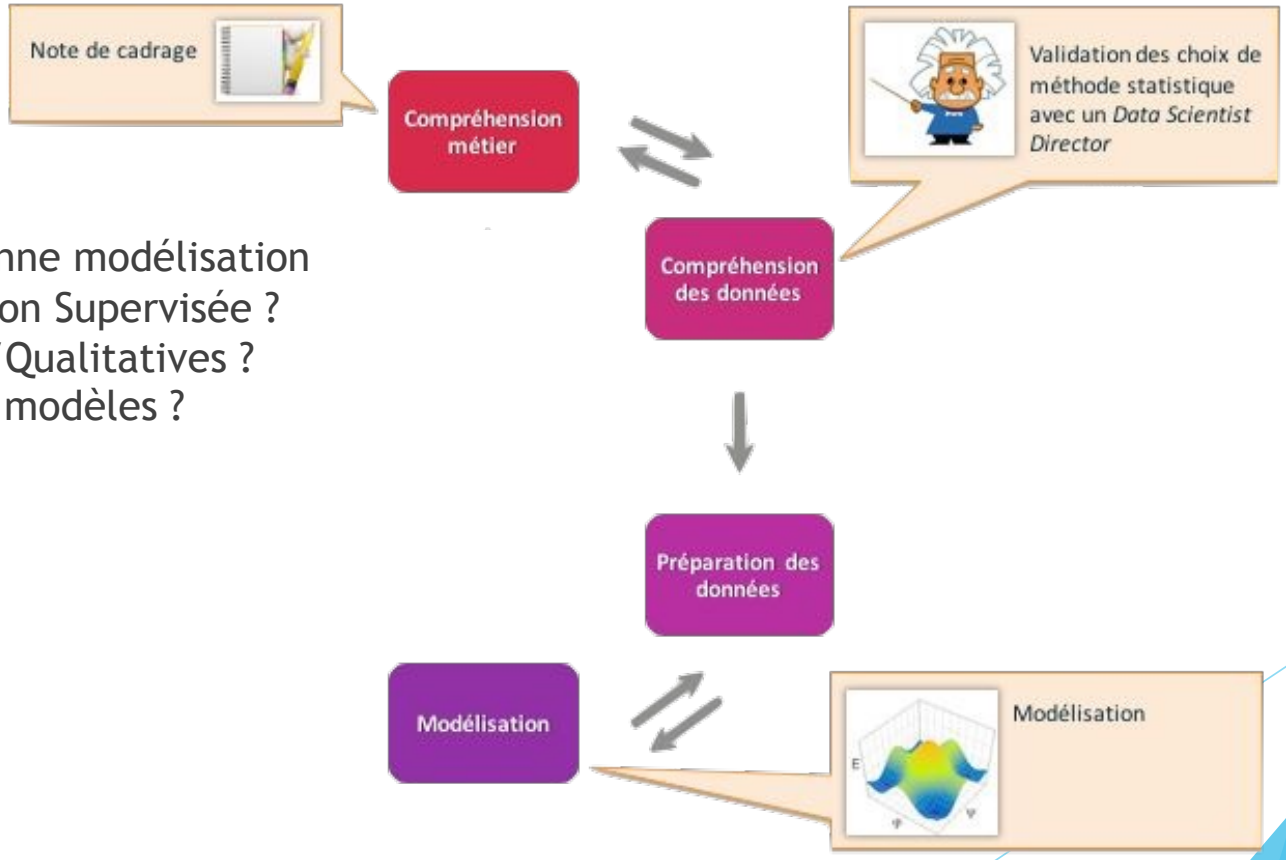
Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)



- Imputer les valeurs manquantes
- Gérer les variables qualitatives
- Traiter les champs textuels
- Normer les variables
- Gérer les valeurs aberrantes

Méthodologie CRISP

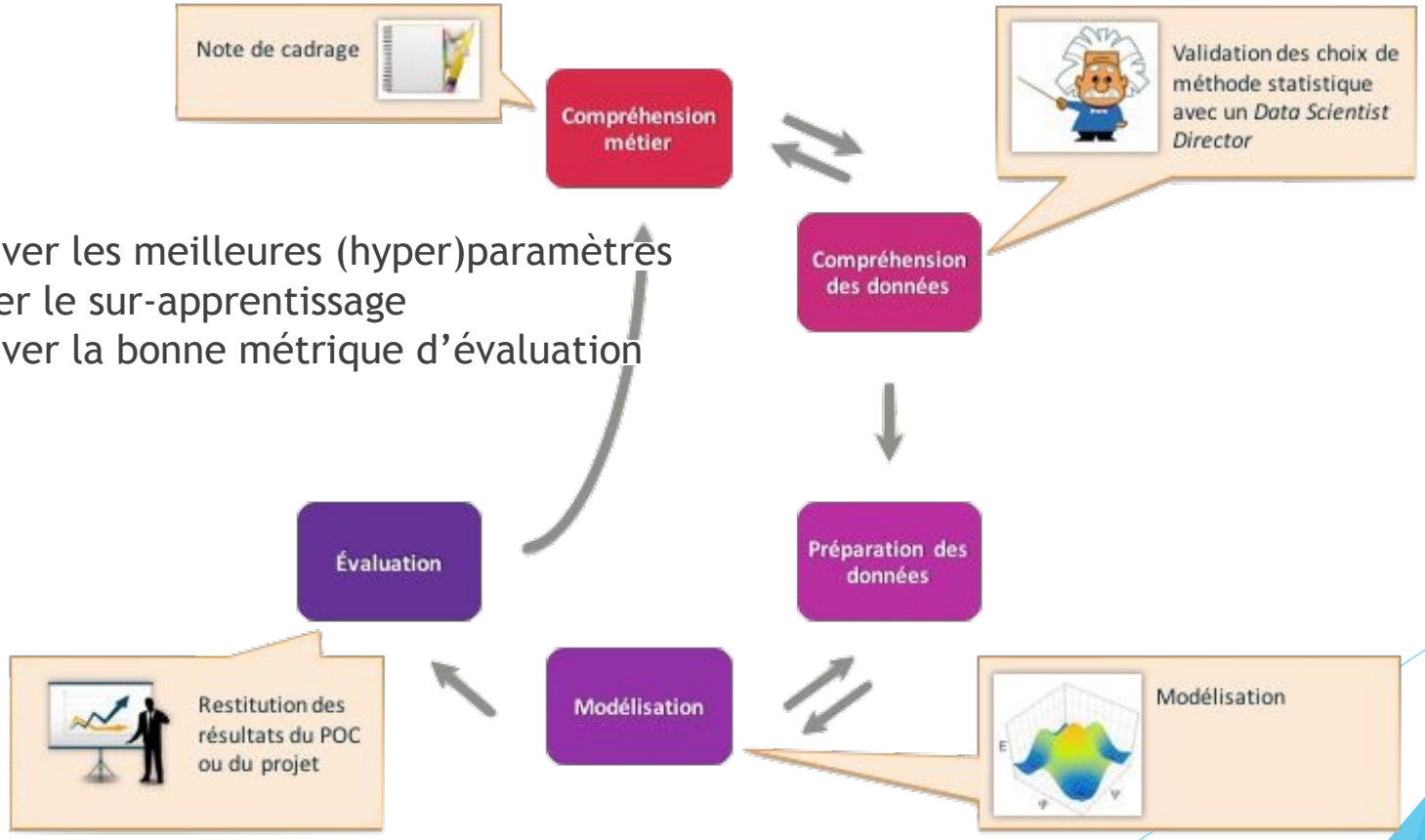
Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)



- Trouver la bonne modélisation
- Supervisée/Non Supervisée ?
- Quantitative/Qualitatives ?
- Combiner les modèles ?

Méthodologie CRISP

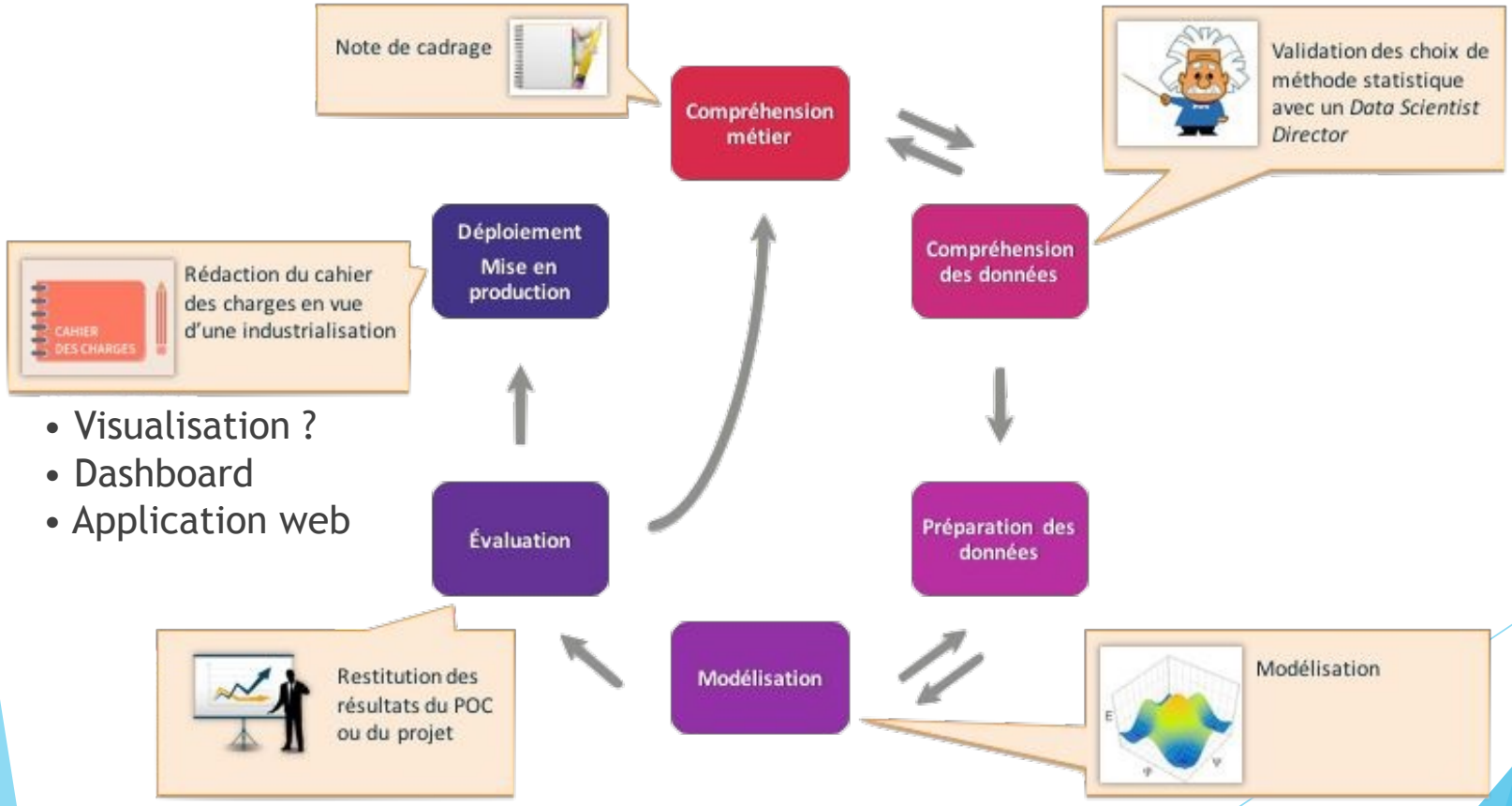
Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)



- Trouver les meilleures (hyper)paramètres
- Eviter le sur-apprentissage
- Trouver la bonne métrique d'évaluation

Méthodologie CRISP

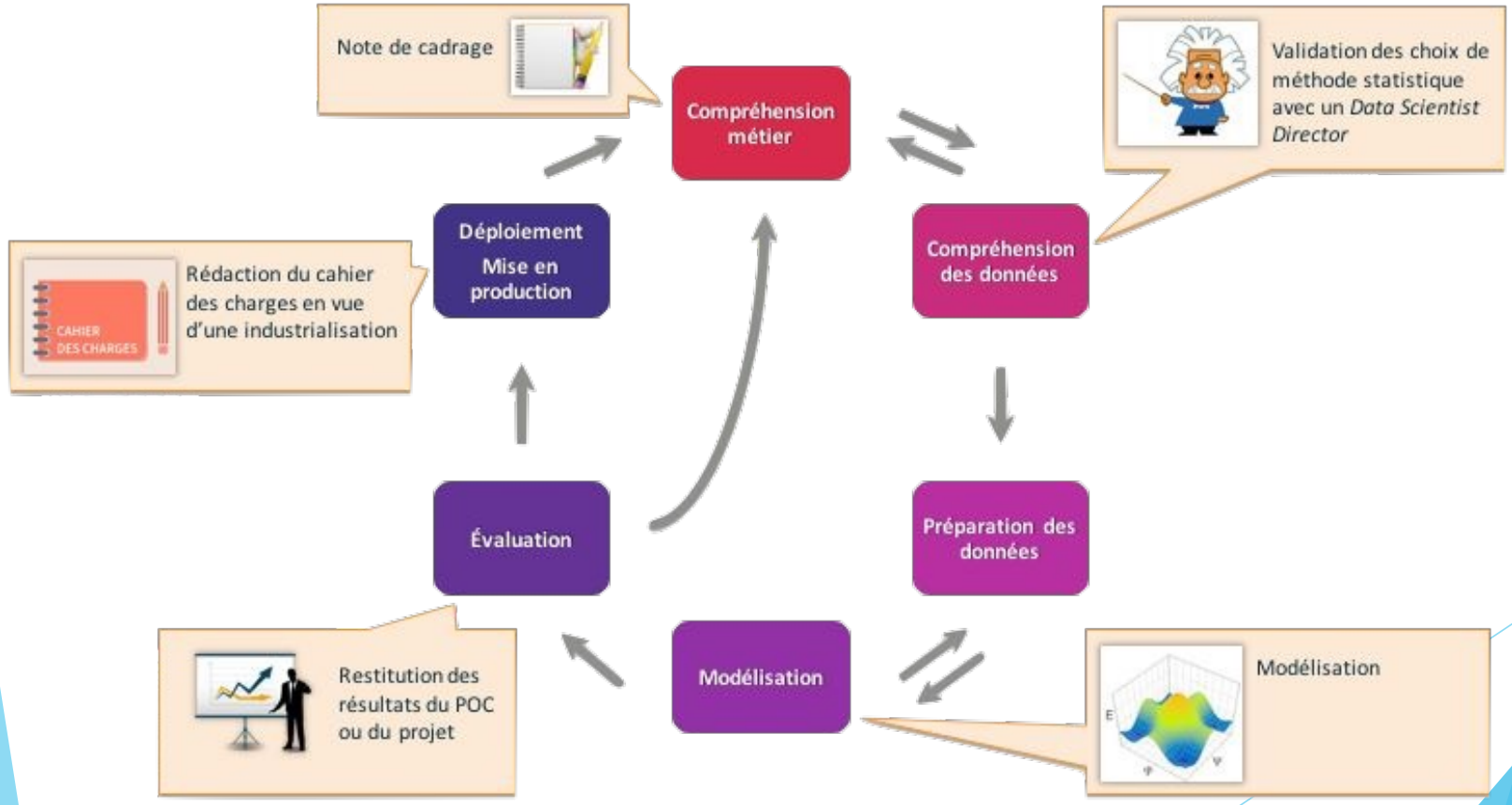
Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)



- Visualisation ?
- Dashboard
- Application web

Méthodologie CRISP

Méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining)

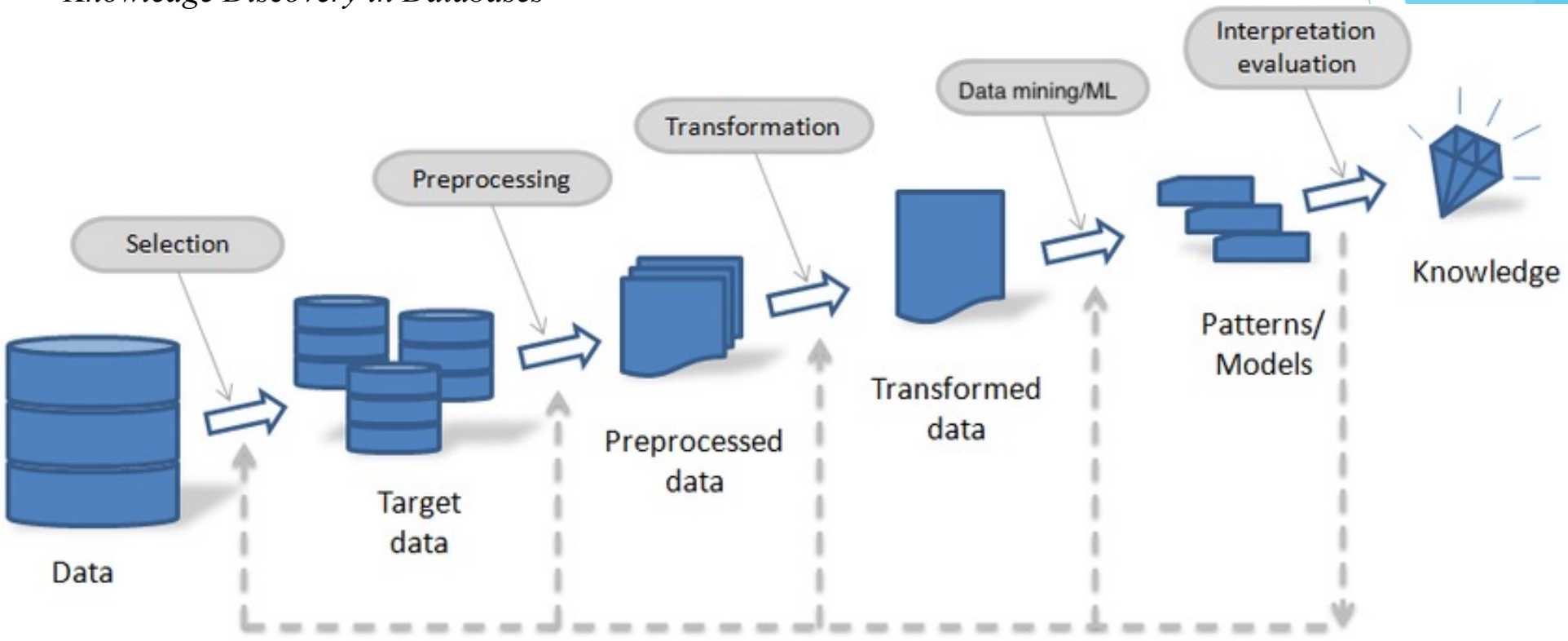


Méthodologie SEMMA

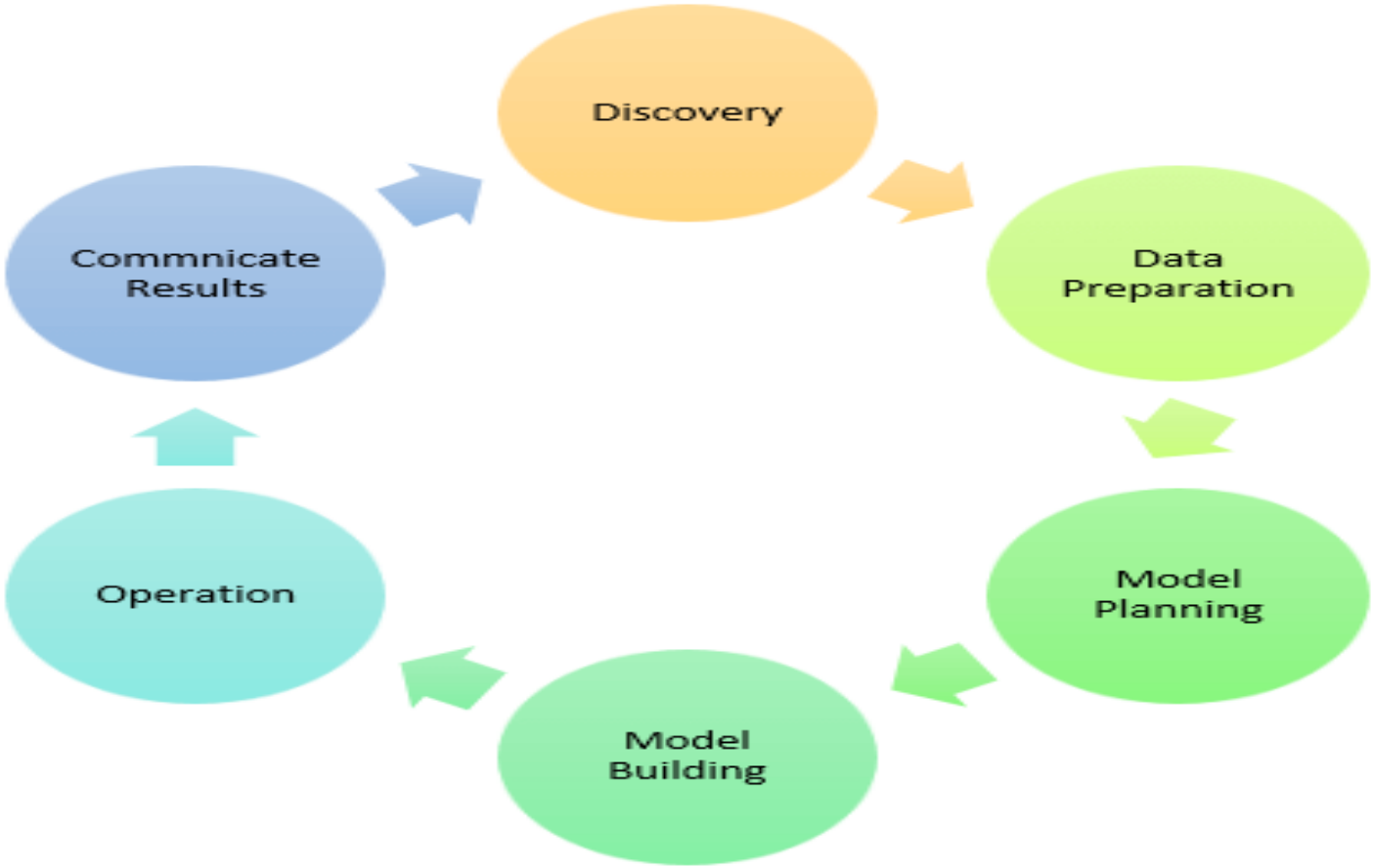
Sampling	Prendre un chantillon significatif pour extraire les modèles
Exploration	Se familiariser avec les données (patterns)
Manipulation	Ajouter des informations, coder, grouper des attributs
Modelling	Construire des modèles
Assessment	Comprendre, valider, expliquer, répondre aux questions

Méthodologie KDD

Knowledge Discovery in Databases



Processus


Discovery
Data Preparation
Model Planning
Model Building
Operation
Commnicate Results

Estimation du temps par étapes

Taches	Charge	Importance dans le projet
Inventaire, préparation et exploration des données	38%	3
Elaboration - Validation des modèles	25%	2
Restitution des résultats	12%	4
Analyse des premiers tests	10%	3
Définition des objectifs	8%	1
Documentation - présentations	7%	5

Principaux métiers

Data Officer

- Responsable des solutions de mises à disposition de l'entreprise
- Anime et coordonne les différents métiers
- Définit la stratégie de l'Equipe

Data Engineer

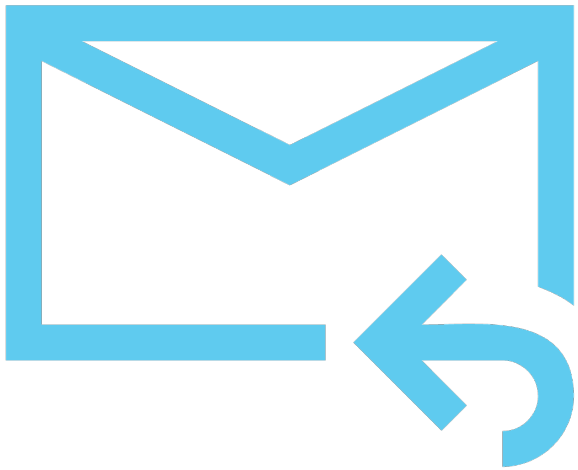
- Responsable des solutions techniques
- Accompagne la mise en œuvre les solution
- Garantit la fiabilité des données

Data Scientist

- Recherche les axes générateurs de valeur en croisant données internes et externes
- Développe des algorithmes d'analyse et/ou de prédictions

Data Analyst

- Identifie les axes de performance basés sur les données internes
- Effectue des dashboarding et des reporting



CONTACT :

PLA-COMES Marine
Marine.placomes@mail-formateur.net