

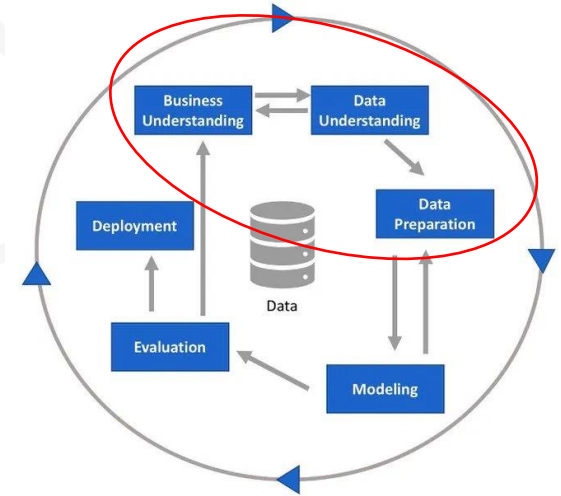
# Préparation, Analyse et visualisation des données

# Préparation

- Introduction
- Compréhension de la problématique métier
- Connaissance de(s) base(s) de données
- Opérations que les valeurs
- Opérations sur les colonnes/variables
- Opérations sur les lignes/observations

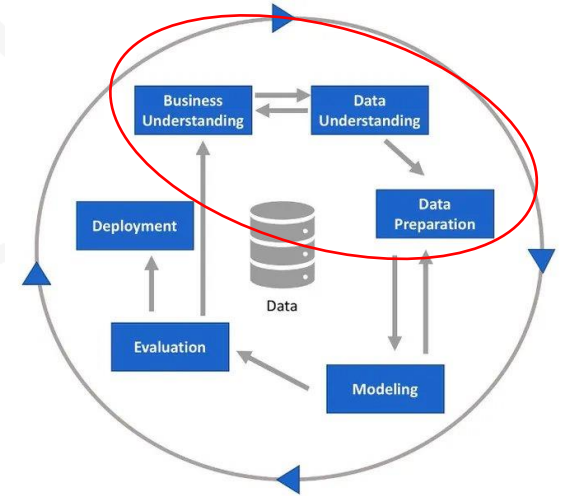
# Problématique métier

- Comprendre et formaliser la problématique métier
- Faire le point des bases de données et des variables d'intérêts
- Définir une stratégie d'analyse



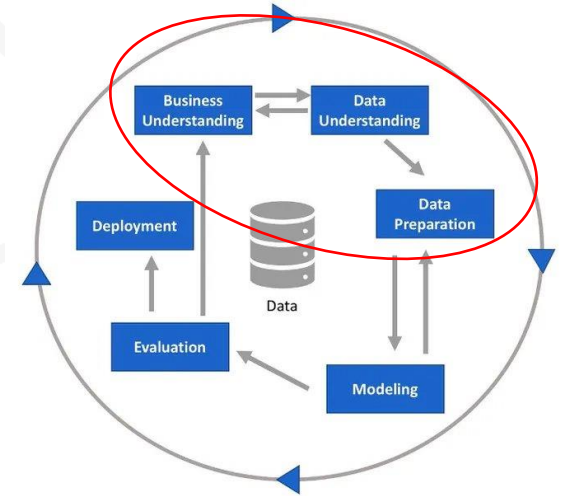
# Préparation des données

- Connaissance des bases de données et des variables d'intérêts (Datamap ou dictionnaire des données)
- Récupérer/collecter/charger les données
- Nettoyage des données



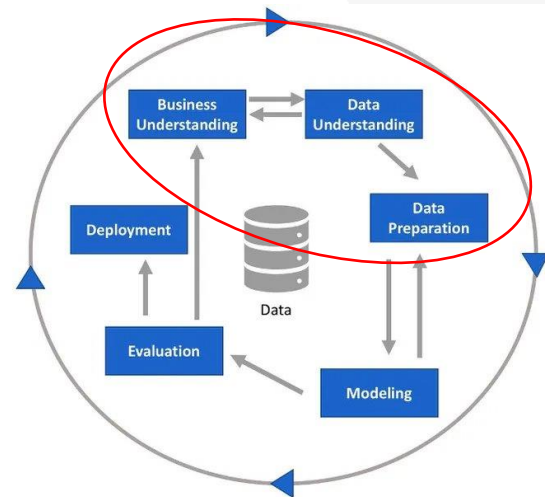
# Préparation des données

- Nettoyage des données
  - Faire le point des variables
  - Connaitre le type des variable
  - Détecter les valeurs manquantes
  - Détecter les valeurs aberrantes
  - Connaitre les relations entre les variables
  - Définir le champ d'étude
  - Normaliser les valeurs



# Préparation des données

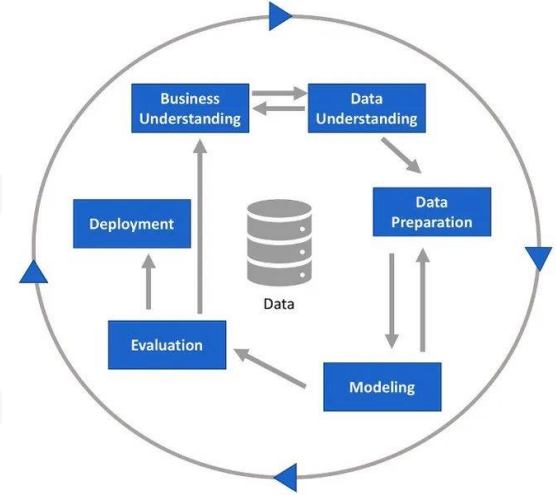
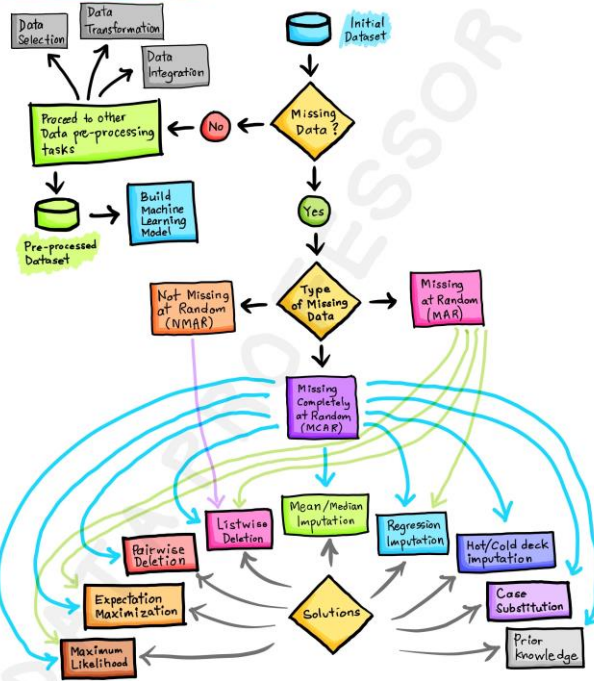
- Traitement des valeurs manquantes (**opérations sur les valeurs**)
  - Les supprimer
  - Faire une imputation
    - Par la moyenne/médiane pour les variables numériques
    - Par le mode les variables catégorielles
    - Par une méthode raisonnée par la connaissance métier
    - Autres méthodes plus sophistiquées



# Préparation des données

## HANDLING MISSING DATA

By: Chanin Nantasenamat



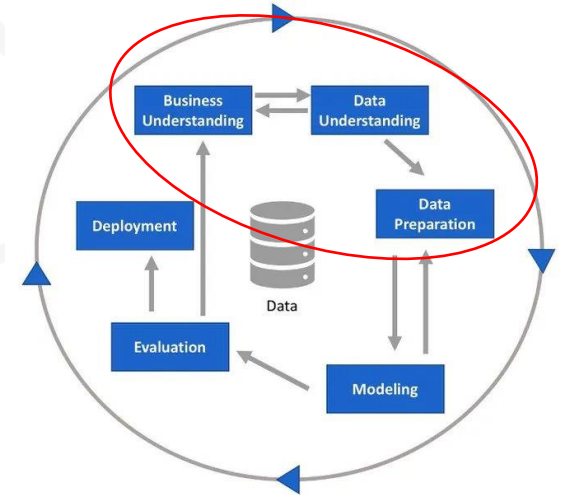
DATA PROFESSOR

<http://youtube.com/dataprofessor>

# Préparation des données

- Traitement des valeurs extrêmes/aberrantes (**opérations sur les valeurs**)
  - Détecter si ce sont des erreurs de saisie ou si ce sont des valeurs bien réelles
  - Si erreurs alors traiter comme des valeurs manquantes
  - Sinon adopter divers stratégies

Dans certains cas, il est nécessaire de normaliser les données

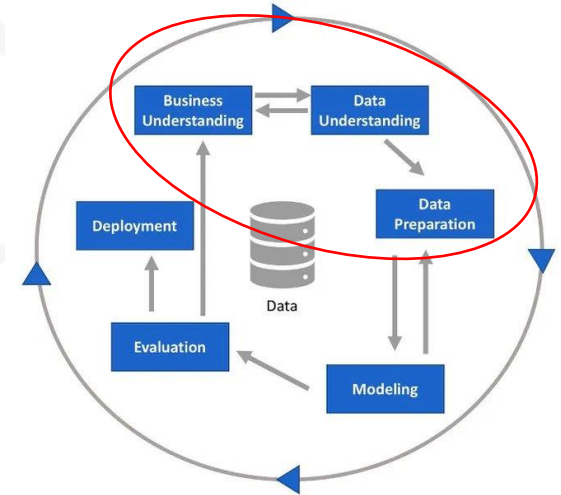




# Préparation des données

- Sélection des variables (**opérations sur les colonnes**)
- Connaitre les relations entre les variables (corrélation, divers liens, etc)
- Variables contenant trop de valeurs manquantes ou aberrantes pour être utilisées
- Eviter les biais de sélection par exemple

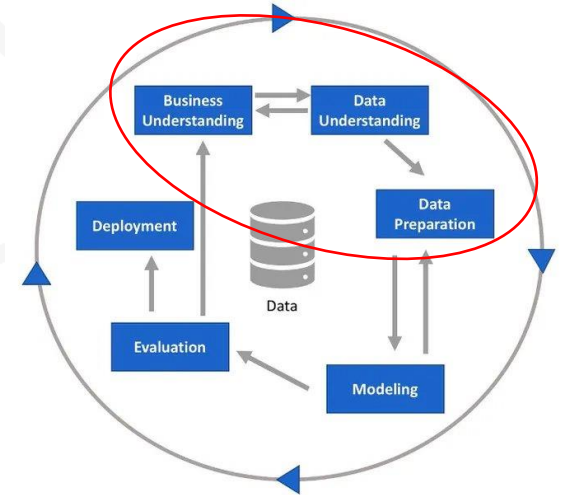
**C'est une étape très importante dans la préparation des données**



# Préparation des données

- Sélection des variables (**opérations sur les colonnes**)
- Connaître les relations entre les variables (corrélation, divers liens, etc)
- Variables contenant trop de valeurs manquantes ou aberrantes pour être utilisées
- Eviter les biais de sélection par exemple

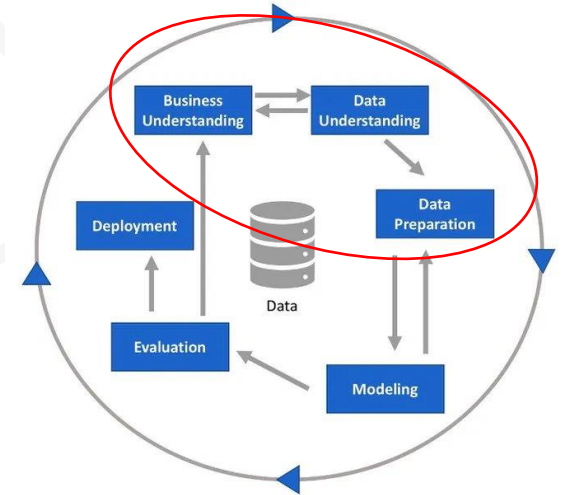
**C'est une étape très importante dans la préparation des données**



# Préparation des données

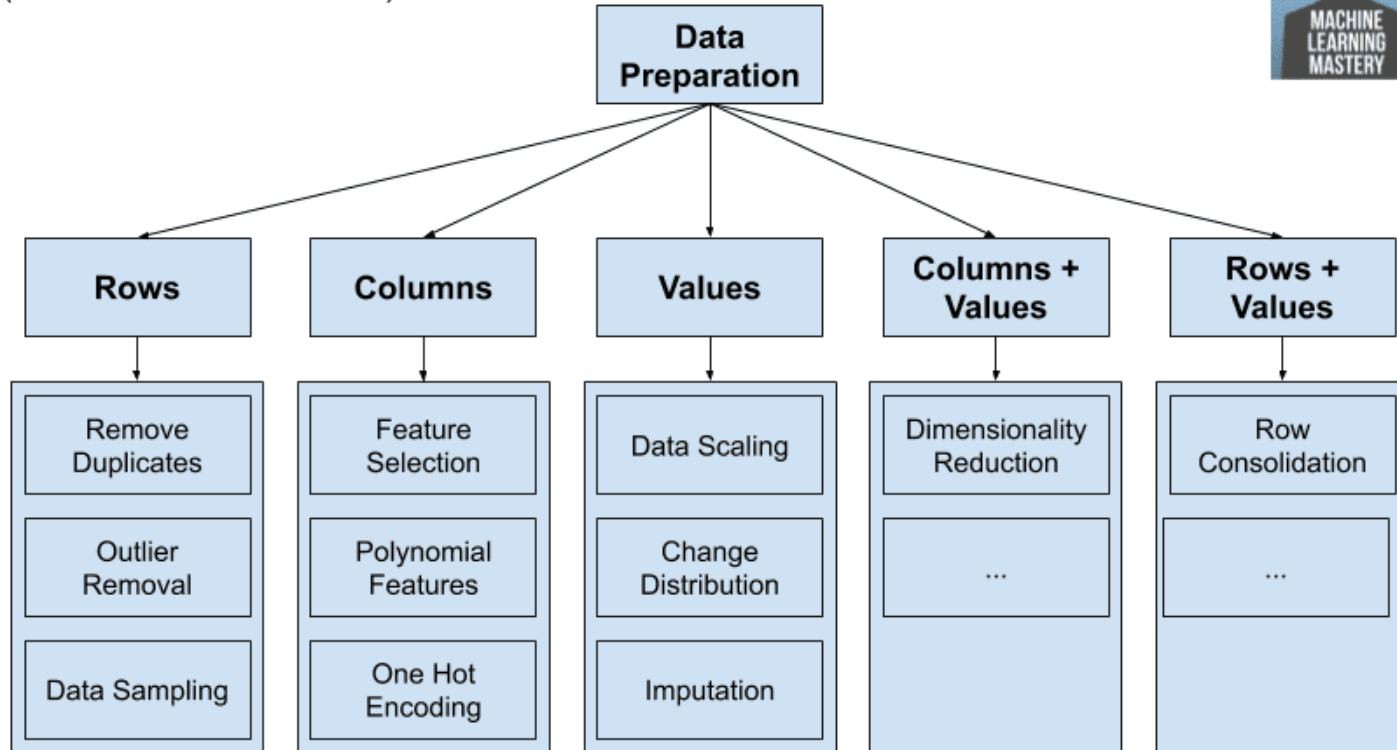
- Définir le champ de la base (**opérations sur les lignes**)
- Supprimer les doublons
- La suppression des valeurs manquantes réduit aussi le champ d'étude. A faire avec parcimonie
- Cette partie de la préparation intrinsèquement liée à la problématique métier

**Cette étape vient clore la préparation des données et ainsi obtenir une base de travail propre**



# Récapitulatif de préparation des données

**Data Preparation Framework**  
(for structured/tabular data)



# Analyse des données

- Calculer et interpréter les statistiques descriptives

- Pour les variables numériques

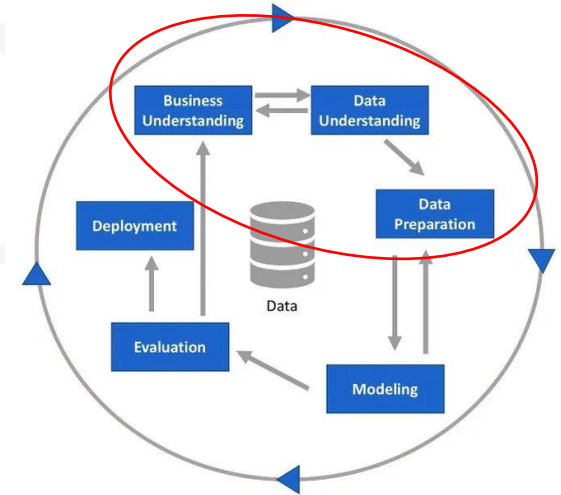
- Moyenne, min, max, écart-type

- médiane, distribution etc.

- Pour les variables catégorielles

- Décomptes des catégories

- mode



# Analyse des données

## ■ Trouver les relations entre les variables

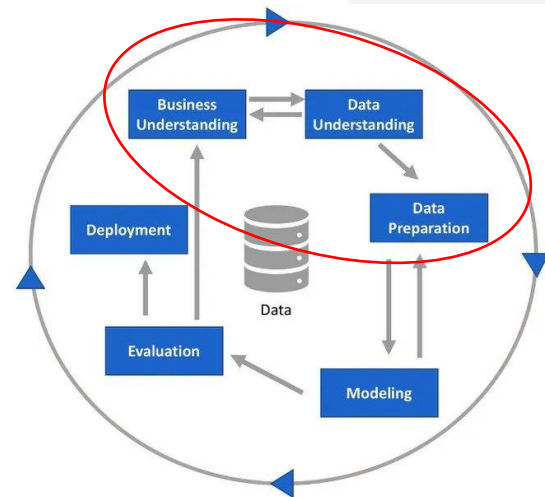
### ■ Pour les variables numériques

#### ■ Table de corrélation

### ■ Pour les variables catégorielles

#### ■ Test de $\chi^2$ et V de Cramer

**L'analyse des données peut amener à utiliser des techniques avancées pour mieux regrouper l'information contenue dans les données**



# Analyse des données

## ■ Trouver les relations entre les variables

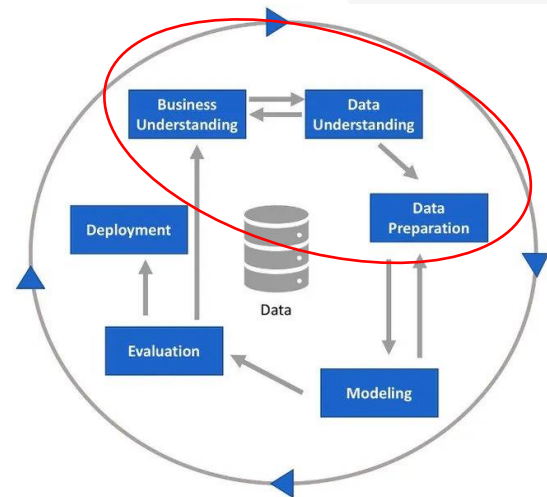
### ■ Pour les variables numériques

#### ■ Table de corrélation

### ■ Pour les variables catégorielles

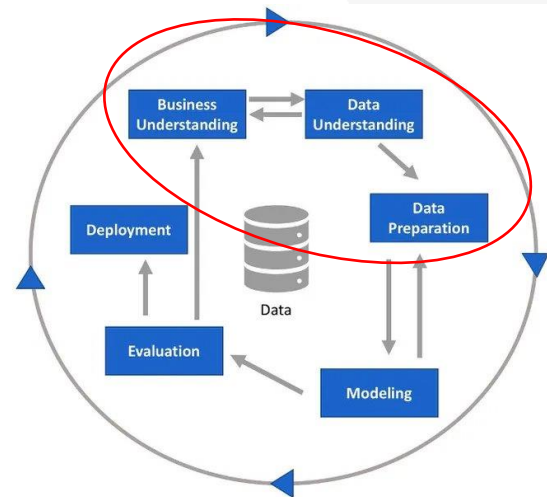
#### ■ Test de $\chi^2$ et V de Cramer

**L'analyse des données peut amener à utiliser des techniques avancées pour mieux regrouper l'information contenue dans les données**



# Visualisation des données

- Une bonne et pertinente visualisation permet de résumer l'information
- Représenter les relations entre les variables
- faire des choix stratégiques dans les divers traitements
- Entrevoir les résultats
- Elle ne doit pas ajouter de la complexité inutile





# Travaux pratiques

**Lewis Hounkpevi**

**0695335936**

**[lewis.dumesnil@gmail.com](mailto:lewis.dumesnil@gmail.com)**

