Analyse de données

Séance 2 - Préparer un dataset (1/2)

Master MIAS - M1
hadriensalem@gmail.com

Des questions par rapport à la séance précédente ?

Nous avons travaillé sur l’analyse des données

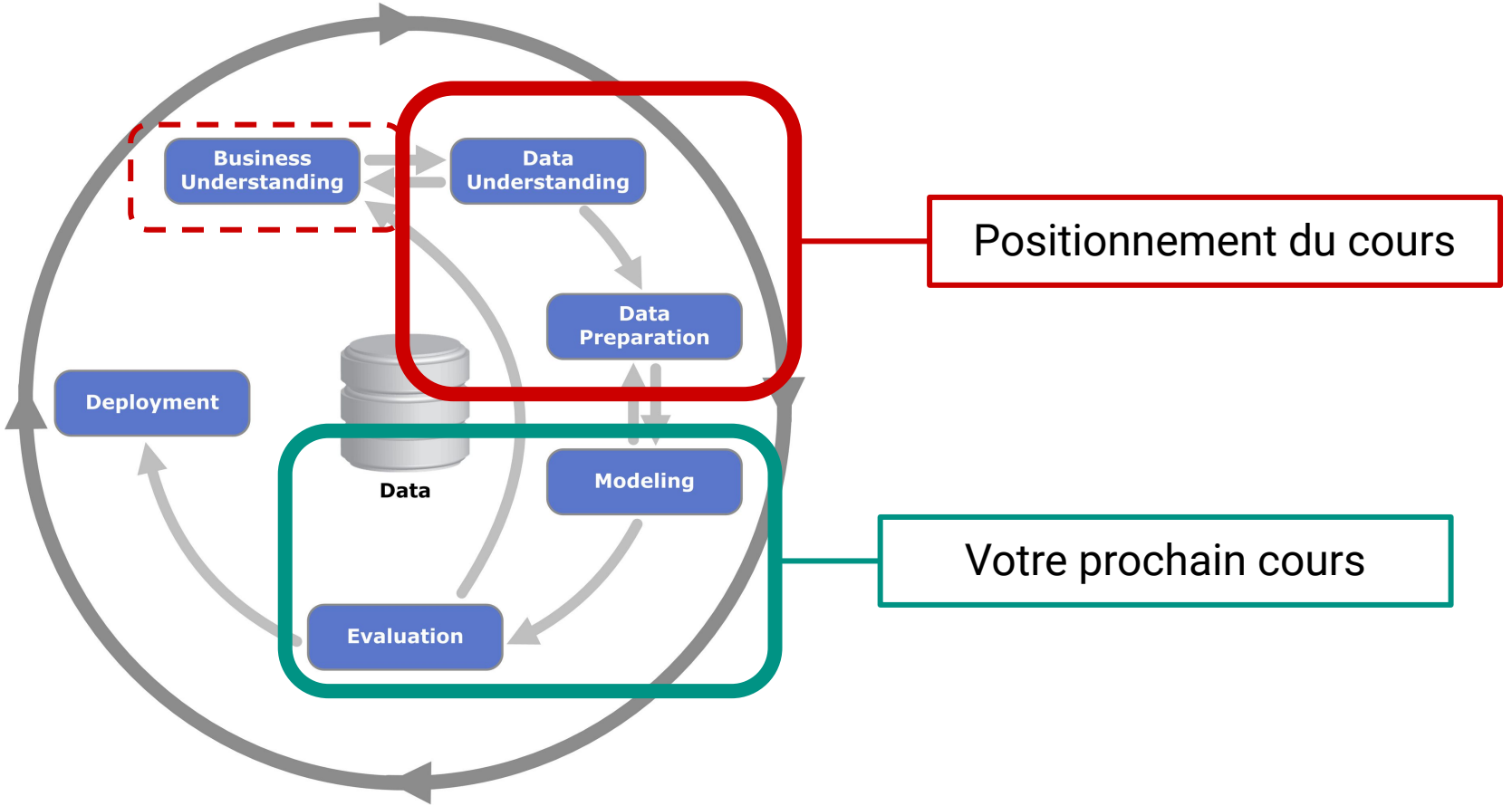
Introduction
Rappels - Positionnement

La méthode CRISP-DM

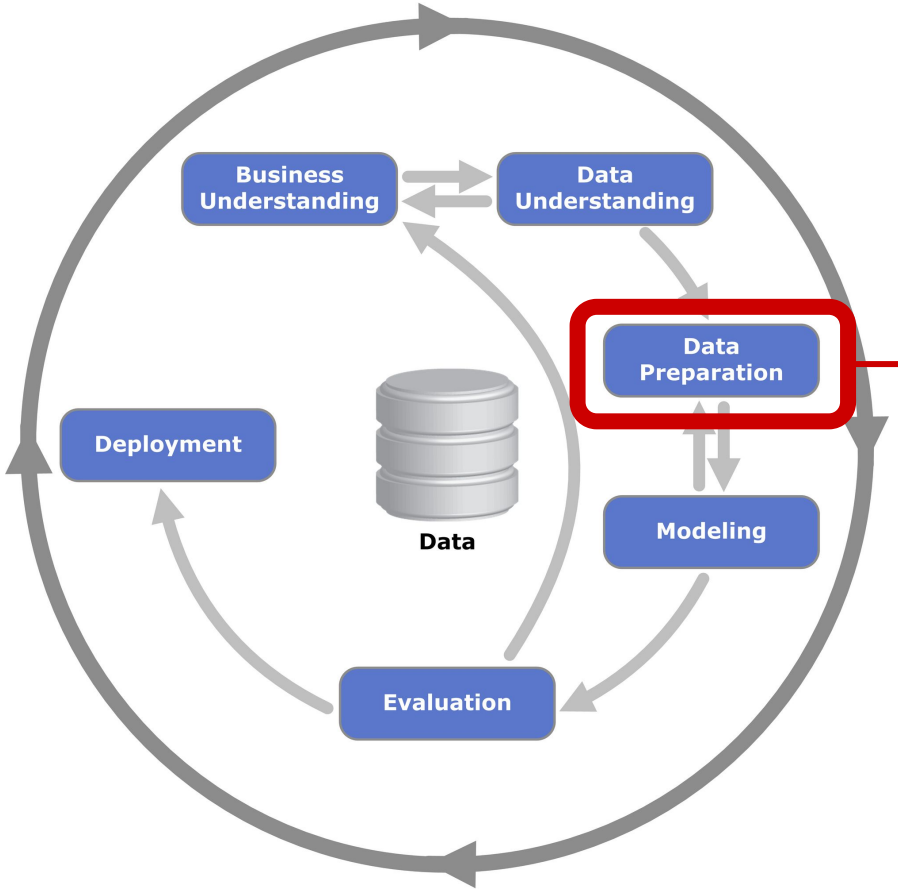
Cross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

➔ Publiée en 1999
➔ Méthode suivie dans l'industrie
➔ Toujours d'actualité

(Source de l'image: Wikipedia)



(Source de l'image: Wikipedia)



Séance d'aujourd'hui

Exploitation des données : la méthode CRISP-DM

(Source de l'image: Wikipedia)

Plan de cours

Séance 1 : Comprendre un dataset

- Étude exploratoire des données
- Visualisation des données

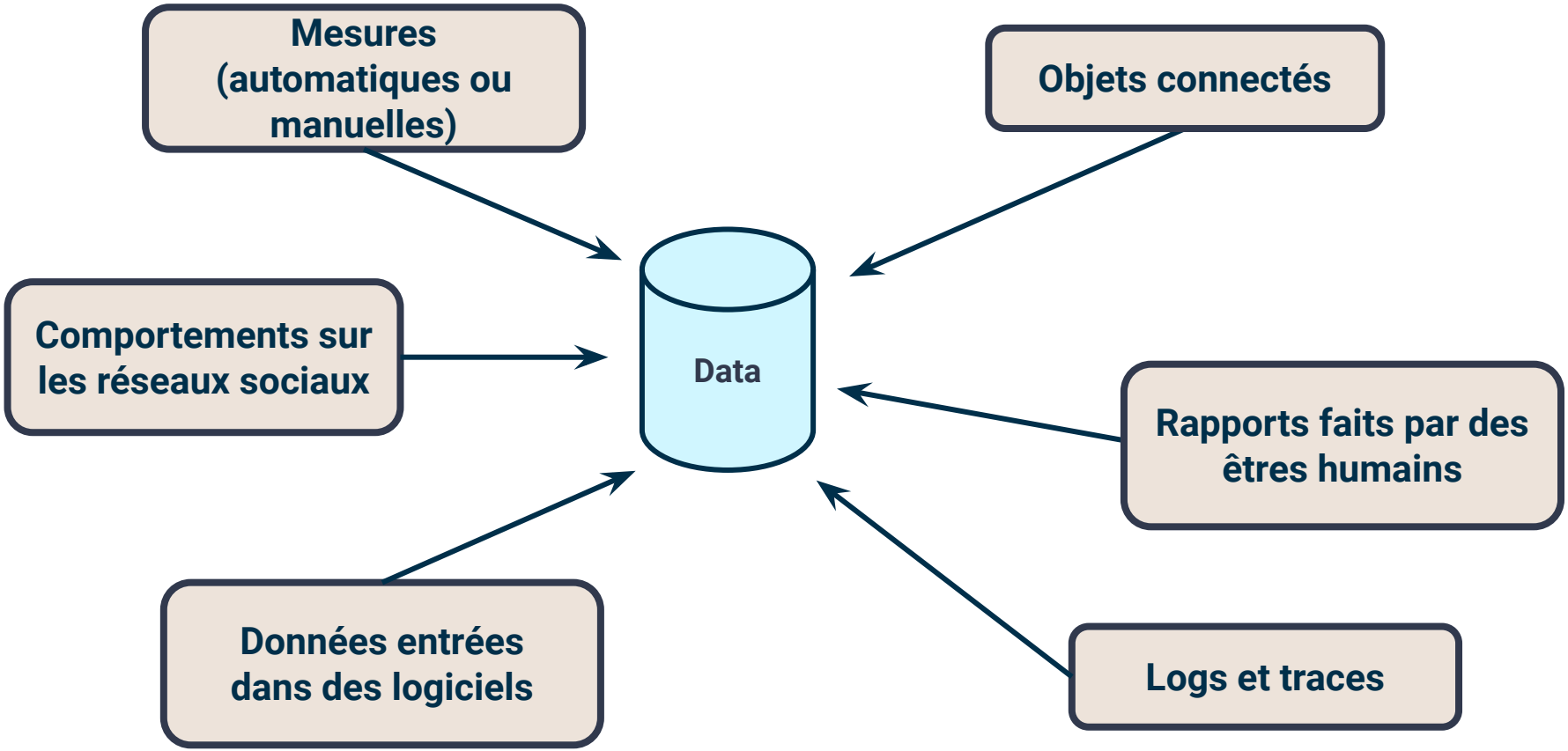
Séance 2 : Préparer un dataset (1/2)

- Overview des types de pré-traitement des data
- Gestion des valeurs manquantes & absurdes

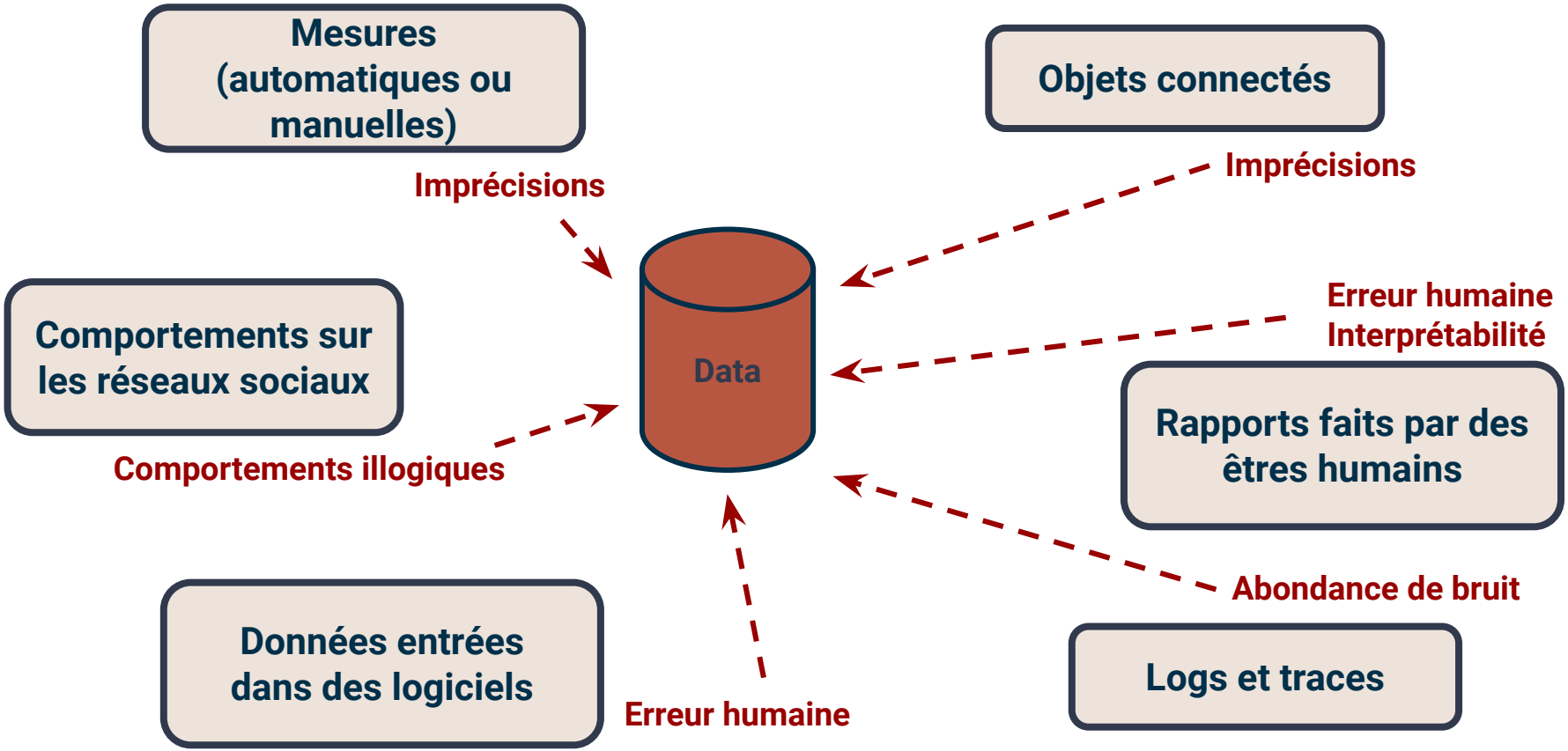
Séance 3 : Préparer un dataset (2/2)

- Introduction à la réduction de dimension
- Typologie des algorithmes de machine learning

Qu’est-ce que la préparation des données ?



Les données “brutes” peuvent avoir de nombreuses sources ...

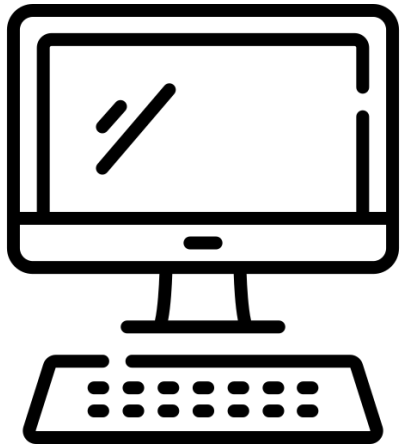


Qui sont toutes sujettes à des erreurs ou des imprécisions !

Les risques d’utilisation de données “sales”

Exemple 1 : Impossibilité d’exploitation

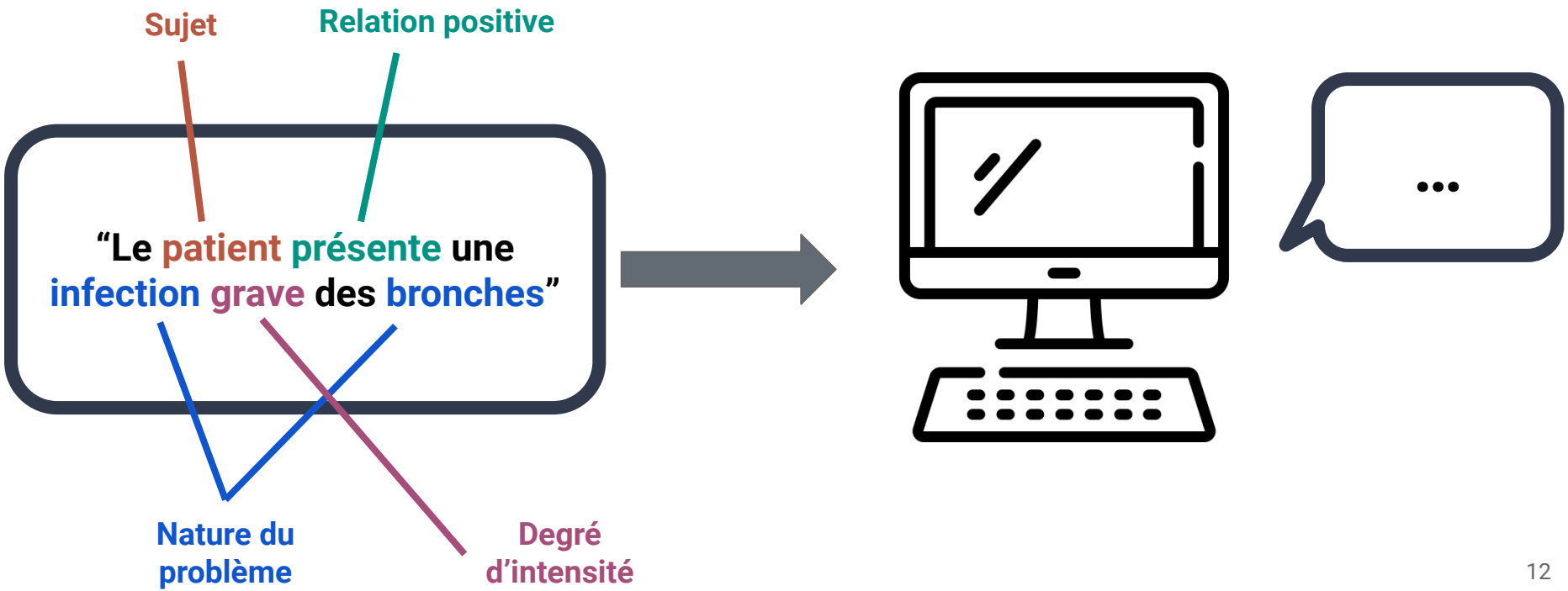
“Le patient présente une infection grave des bronches”



?

Les risques d’utilisation de données “sales”

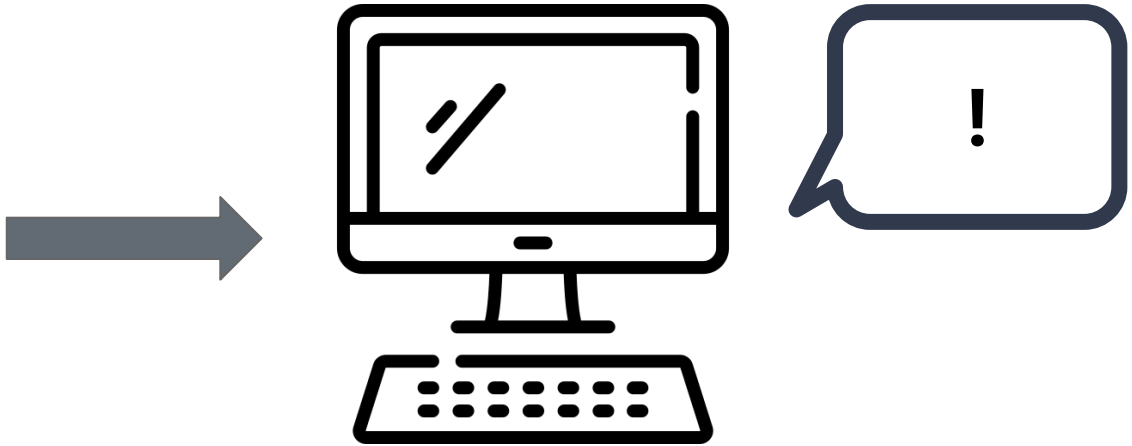
Exemple 1 : Impossibilité d’exploitation



Les risques d’utilisation de données “sales”

Exemple 1 : Impossibilité d’exploitation

TARGET_TYPE: Patient
PROBLEM: Infection
LOCATION: 24
LOCATION_LABEL: Bronchi
INTENSITY: 3

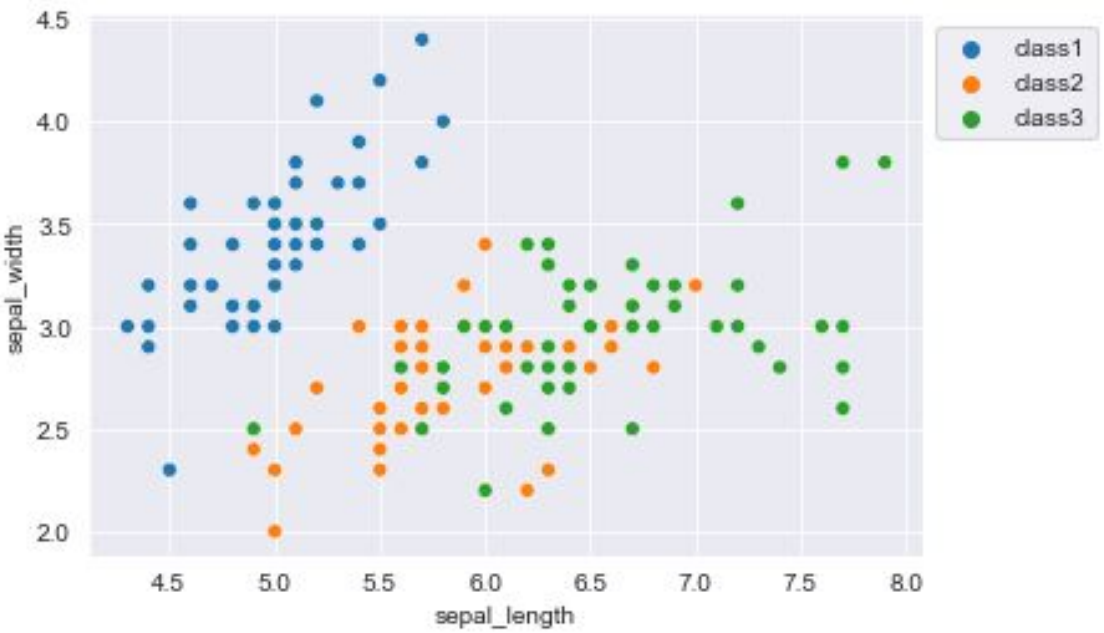


Les données doivent êtres transformées de façon à être interprétables par la machine.

Les risques d’utilisation de données “sales”
Exemple 2 : Difficulté de modélisation

Une illustration venant du dernier TP

Si on fait une classification “naïve” en n’utilisant que les informations sur les sépales, la séparation entre les espèces n’est pas claire.

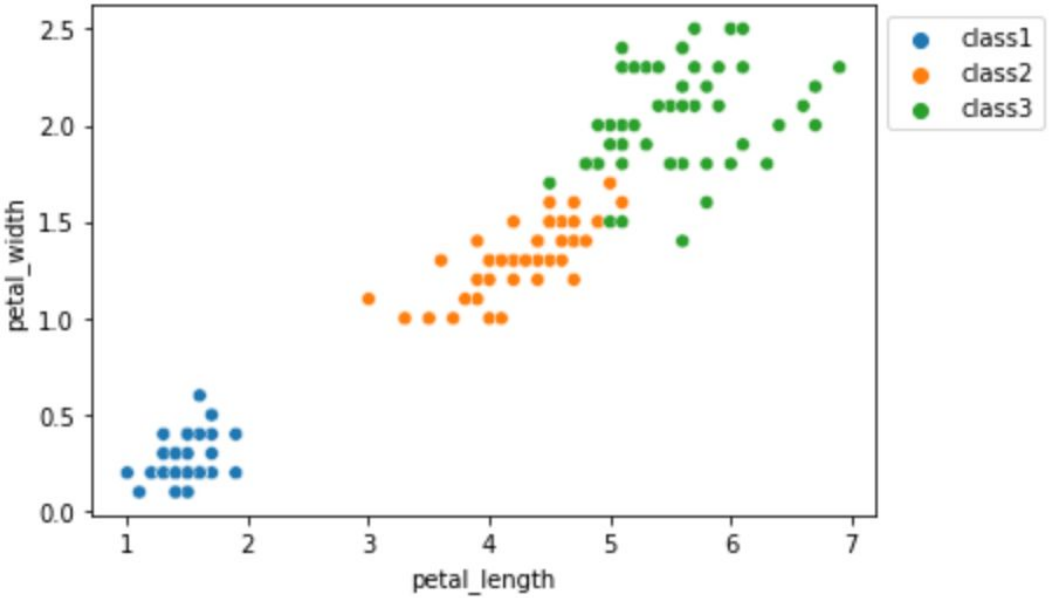


Les risques d’utilisation de données “sales”
Exemple 2 : Difficulté de modélisation

Une illustration venant du dernier TP

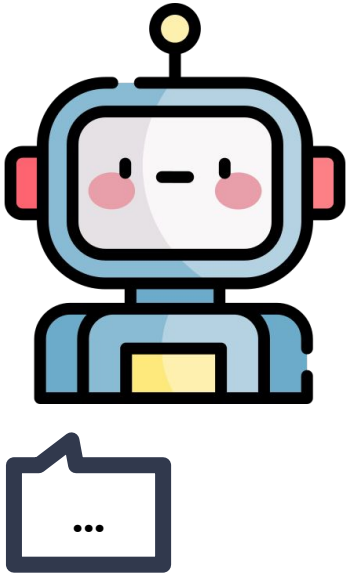
Si en revanche on utilise les pétales, le modèle à utiliser devient beaucoup plus évident.

<u>Utiliser des colonnes pertinentes est primordial pour le machine learning.</u>

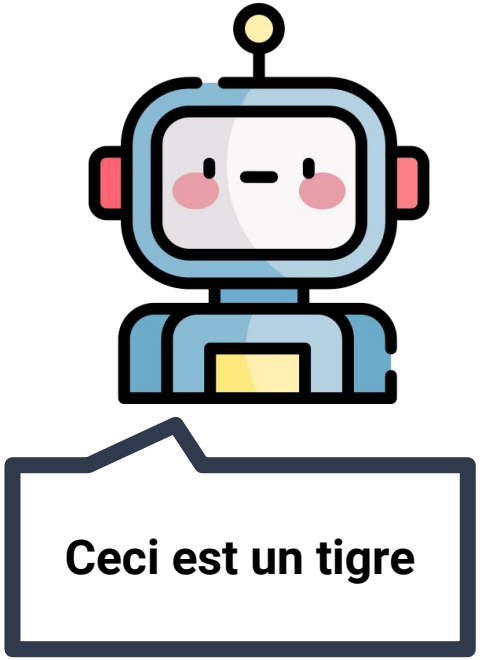


Les risques d’utilisation de données “sales”
Exemple 3 : Création de biais dans les données

Apprentissage



Les risques d’utilisation de données “sales”
Exemple 3 : Création de biais dans les données



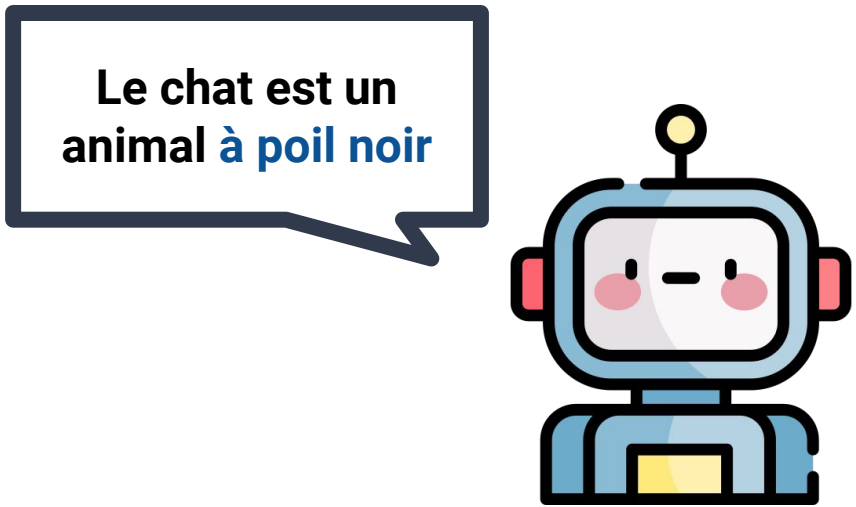
Les risques d’utilisation de données “sales”
Exemple 3 : Création de biais dans les données

Notre base de données d’entraînement a apporté un biais dans l’apprentissage de l’algorithme.

Les biais peuvent avoir des conséquences plus graves :

- Limitation à une région du monde
- Discrimination
- Sexisme
- Entretien de biais humains
- etc.

Les biais sont à éviter dans une optique de **généralisation des prédictions d’algorithmes.**

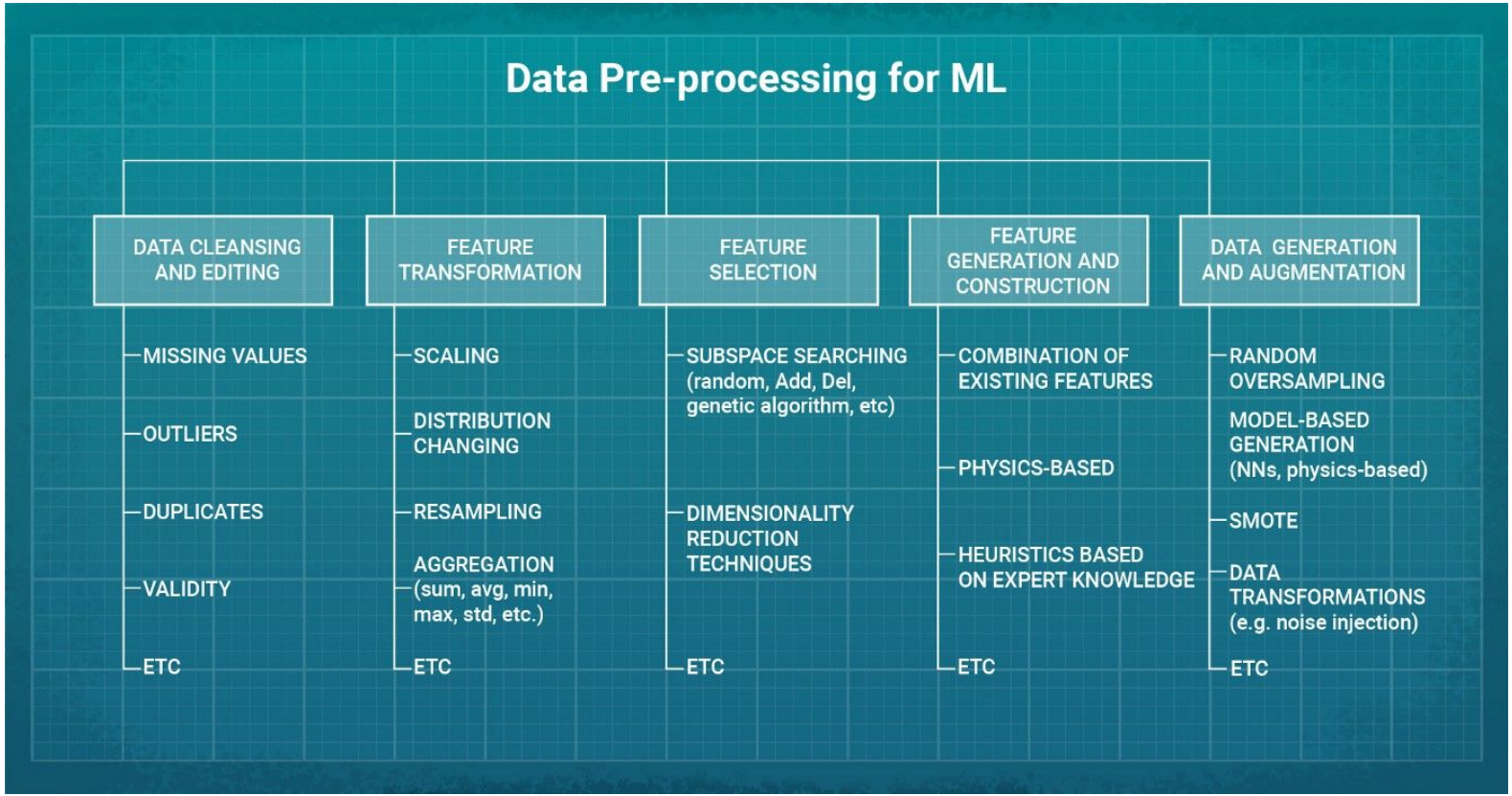


Préparer des données, c’est les rendre exploitables

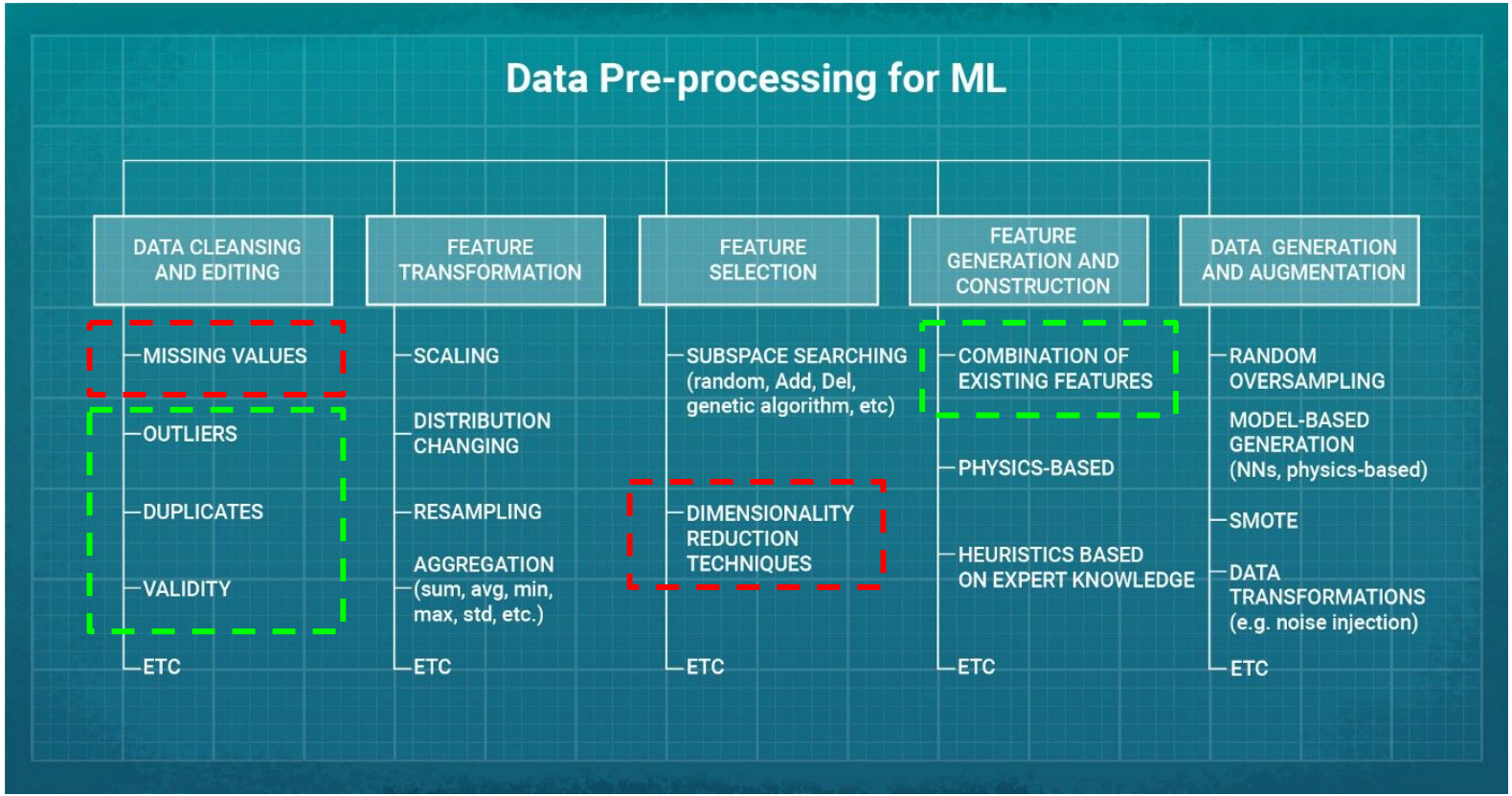
La donnée “à sa source” est (presque) toujours **bruitée**, **peu pratique d’utilisation**, et **difficile à manipuler**.

Un **travail préparatoire** est systématiquement nécessaire en vue d’un apprentissage automatique.

Différents types de traitements



Source de l'image



Nous avons déjà abordé certaines de ces techniques, et nous allons en traiter d'autres

Gestion des valeurs manquantes

Pourquoi y a-t-il des valeurs manquantes ?

Pourquoi y a-t-il des valeurs manquantes ?

Des données peuvent manquer pour des raisons…

- **Techniques**
  - Machines défectueuses
  - Problème d'encodage / de conversion
- **Humaines**
  - Erreurs de saisie
  - Choix délibéré (e.g. sondages)
- **Méthodologiques**
  - Relevés non réalisés pour une partie de la population (e.g. PSA chez les femmes)
  - Manque de moyens de mesure

Pourquoi y a-t-il des valeurs manquantes ?

On distingue trois grandes catégories…

- ***Missing Completely at Random***
  - Pas de motifs explicables
  - Exemple : Oubli humain
- ***Missing at Random***
  - Motifs éventuellement explicables par d'autres colonnes
  - Exemple : Les hommes ont moins tendance à répondre à une certaine question que les femmes
- ***Missing Not at Random***
  - Motifs explicables, mais l'explication n'est pas toujours observables par d'autres colonnes
  - Exemple : Les gens moins aisés ont tendance à ne pas répondre à des questions sur leurs salaires

Pourquoi y a-t-il des valeurs manquantes ?

Qui peuvent créer différents types de problèmes !

- ***Missing Completely at Random***
  - Pas de motifs explicables
  - Exemple : Oubli humain
  - **Manque de données pour l'apprentissage**
- ***Missing at Random***
  - Motifs éventuellement explicables par d'autres colonnes
  - Exemple : Les hommes ont moins tendance à répondre à une certaine question que les femmes
  - **Biais : l'algorithme généralisera mieux pour les hommes que pour les femmes**
- ***Missing Not at Random***
  - Motifs explicables, mais l'explication n'est pas toujours observables par d'autres colonnes
  - Exemple : Les gens moins aisés ont tendance à ne pas répondre à des questions sur leurs salaires
  - **Biais : le salaire moyen dans la base de données est plus important que le salaire moyen réel**

Comment gérer les valeurs manquantes?

Comment gérer les valeurs manquantes?

Supprimer les lignes

Imputer les valeurs

Comment gérer les valeurs manquantes?

Supprimer les lignes

- Simple
- Peut réduire le nombre de données de façon trop importante
- Peut créer des biais dans les données

Imputer les valeurs

- Plus robuste, surtout quand les données manquantes sont nombreuses
- On garde un dataset “complet”
- Le choix de méthode est empirique
- Peut introduire des biais ou des absurdités dans les données

À vos notebooks !

Le TP se trouve au même endroit que pour la première séance :

https://github.com/SnowHawkeye/mias-data

Remplacement par la moyenne

La moyenne est par définition une “valeur pertinente”.

Attention toutefois, elle est calculée à partir des valeurs observables, et donc d’éventuels biais existants.

Avantages (possibles)

- Extrêmement simple d’implémentation
- Donne une *baseline* avec très peu d’efforts

Inconvénients (possibles)

- La moyenne peut être sensible aux *outliers*, si ceux-ci sont concentrés d’un côté de la distribution
- On renforce le poids de “l’individu moyen”

Remplacement par la médiane

La médiane est souvent proche de la moyenne dans les datasets qui ne sont pas trop déséquilibrés.

Elle a l'avantage d'être moins sensible aux *outliers*.

Remarque : Une autre solution serait de se débarrasser / de recalibrer les outliers

Avantages (possibles)

- Extrêmement simple d'implémentation
- Donne une *baseline* avec très peu d'efforts
- Moins sensible aux *outliers* que la moyenne

Inconvénients (possibles)

- Ignorer les valeurs extrêmes (e.g. dans un dataset à forte variance)
- On renforce le poids de "l'individu médian"

Remplacement par une valeur aléatoire

L'utilisation de valeurs aléatoires en Machine Learning donne parfois des résultats étonnament bons.

Pour tenter d'être plus pertinent, on peut essayer de respecter une distribution proche de celle qui est observable (e.g. distribution normale, exponentielle, etc).

Remarque : D'où encore l'importance de la visualisation (cf. `kdeplot`*) !*

Avantages (possibles)

- Assez facile d'implémentation
- On évite d'augmenter trop le poids des valeurs existantes

Inconvénients (possibles)

- Trouver une distribution pertinente peut être compliqué
- Respecter la distribution observable peut créer des biais dans les données
- Choisir "complètement au hasard" peut introduire des absurdités dans les données

Remplacement par une valeur fréquente

Cette méthode est surtout valable dans le cas de données non numériques.

Pour aller plus loin, et comme pour les données chiffrées, on peut envisager une imputation “intelligente” à partir d’une distribution aléatoire.

Avantages (possibles)

- Extrêmement simple d’implémentation

Inconvénients (possibles)

- Donner trop de poids à la valeur la plus fréquente dans l’apprentissage
- Maintien des biais

Remplacement par une valeur interpolée

L'interpolation est pertinente lorsque la valeur d'une donnée suit une fonction connue.

Les interpolations les plus courantes sont les interpolations linéaires et polynomiales (en particulier, quadratique).

Remarque : Encore une fois, c'est quelque chose qu'on peut observer avec des courbes !

Avantages (possibles)

- Si la distribution suit un profil de fonction proche de celui qu'on a choisi, les résultats peuvent être très bons

Inconvénients (possibles)

- Pas applicable dans tous les cas (les résultats seront mauvais si on ne trouve pas une fonction pertinente)

Imputation par Machine Learning

Il existe des algorithmes basés sur de l'apprentissage, permettant d'imputer des valeurs manquantes.

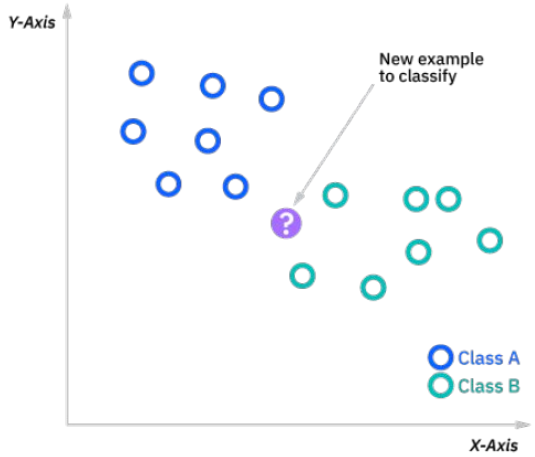
Scikit-learn en propose plusieurs, tels que le `SimpleImputer`, le `KNNImputer`, ou le `IterativeImputer`.

Avantages (possibles)

- Ces méthodes plus avancées évitent certains des inconvénients précités
- Dans beaucoup de cas, elles sont susceptibles de trouver des valeurs plus "proches de la réalité"

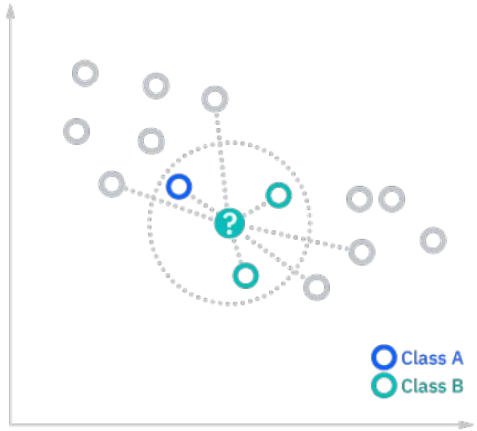
Inconvénients (possibles)

- Le choix de l'Imputer est expérimental
- On réalise déjà une étape de machine learning : si les données n'ont pas été nettoyées, on fait face aux risques habituels



Introduction d'un nouvel exemple

Calcul des distances



Vote majoritaire

L'algorithme des K plus proches voisins permet également d'imputer des valeurs

Source : IBM

K plus proches voisins

L'algorithme KNN (K-nearest neighbours) est l'un des algorithmes de classification les plus basiques, mais il donne de bons résultats dans certains cas.

Il est facilement applicable à l'imputation de données.

Avantages (possibles)

- Implémentation facile
- Peu d'hyperparamètres

Inconvénients (possibles)

- Le choix du calcul de distance n'est pas toujours évident
- Peu devenir cher en ressources machine et en temps de calcul
- Sensible au "mal de la dimension"
- Sensible à l'*overfitting*