Data Science

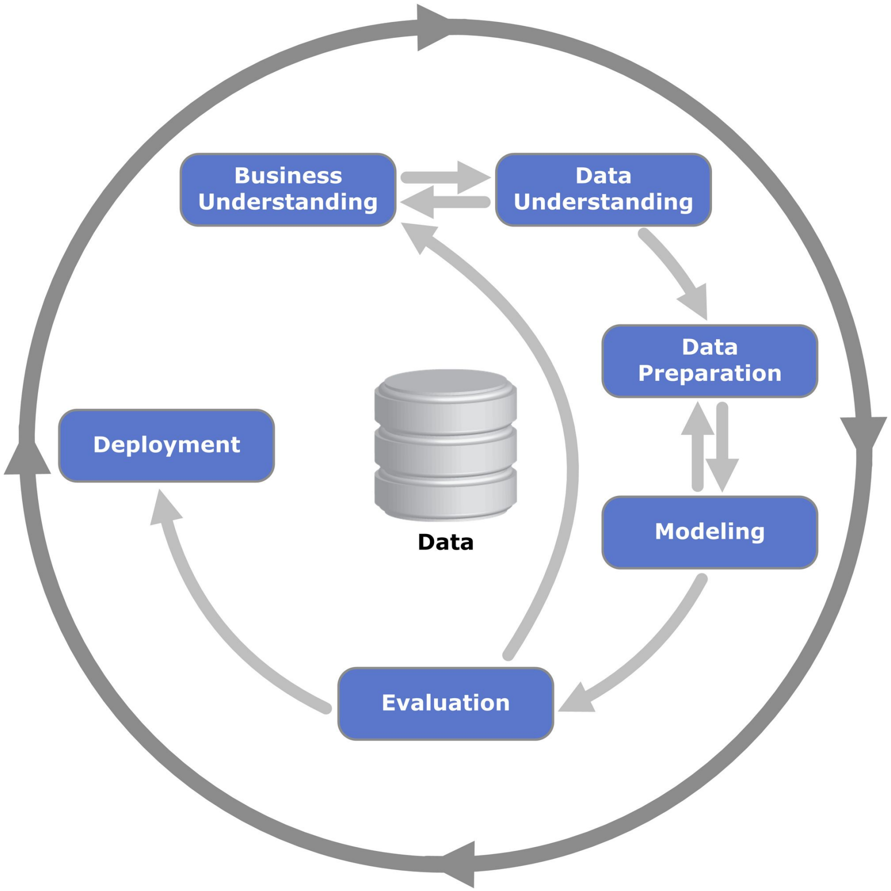
Session 3 - Cleaning data & Missingness

hadrien.salem@centralelille.fr

introduction-to-data-science

Introduction

What did we do last time?



The CRISP-DM method

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

➔ Published in 1999
➔ Common in the industry
➔ Still relevant today

The CRISP-DM method to carry out data-driven projects

(Image source: Wikipedia)

Course outline

Data science course

Session 1: Understanding data

Session 2: Clean code & Git

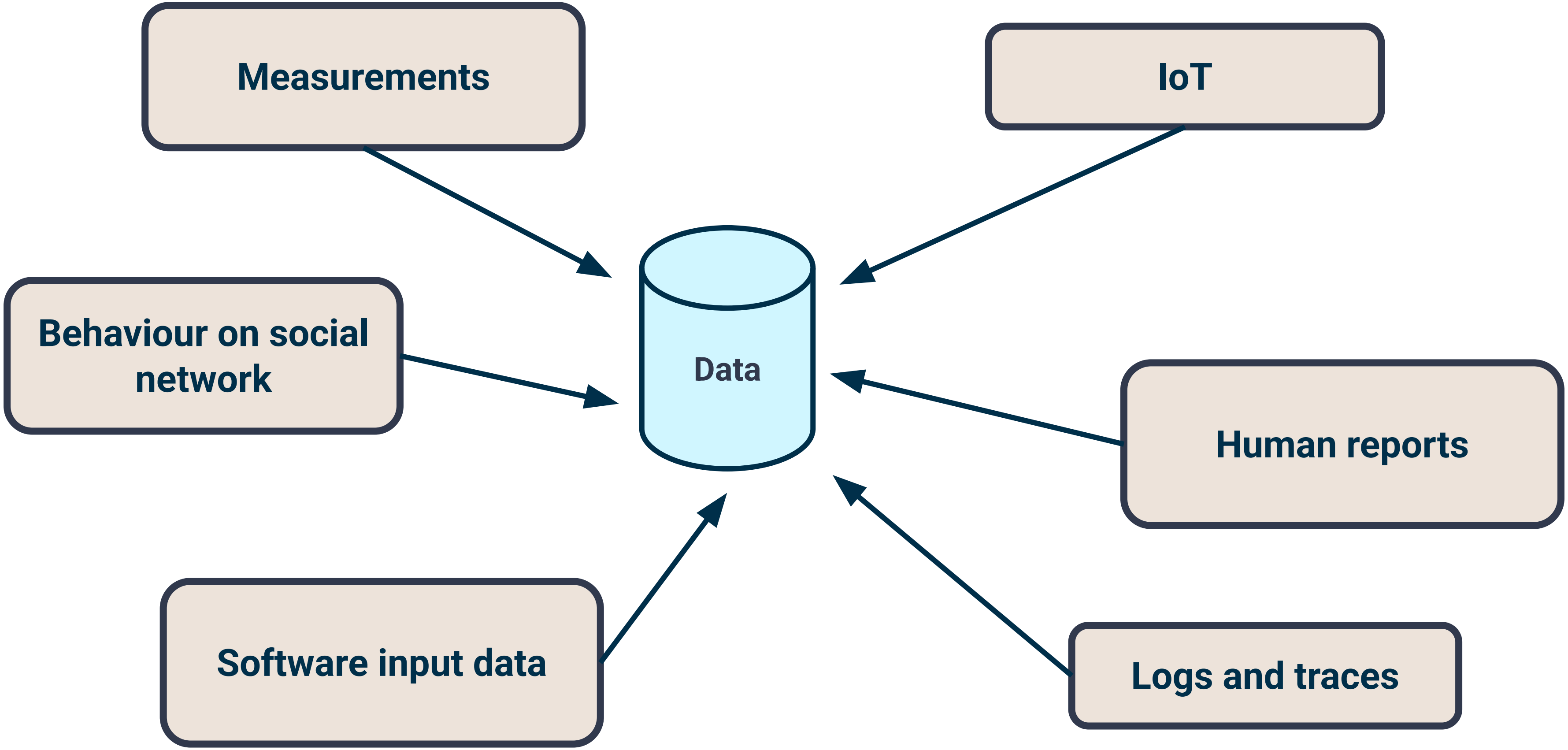
Session 3: Preparing data - Cleaning & Missingness

Session 4: Preparing data - Dimensionality reduction

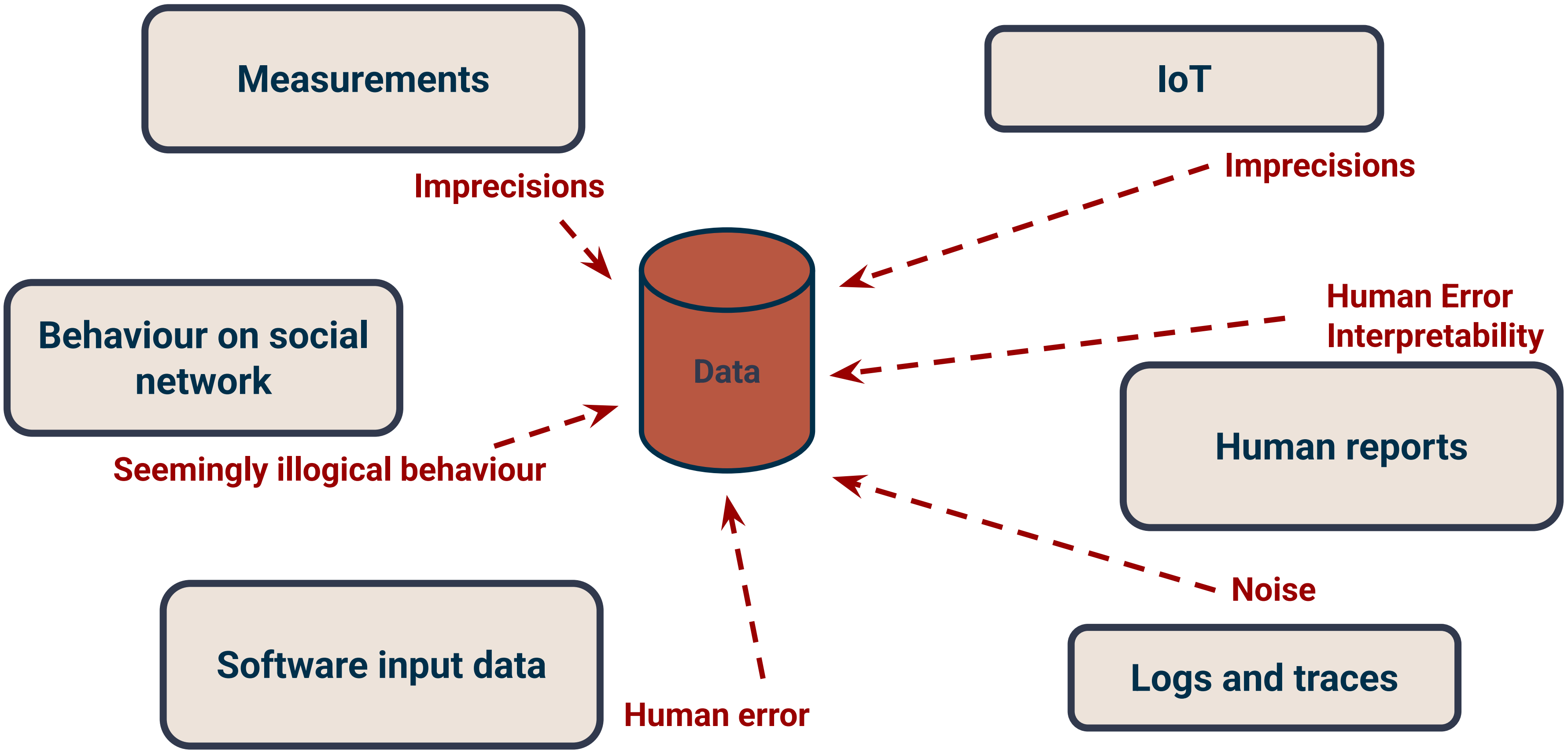
Session 5: Preparing data - Data imbalance

Machine learning course

What does it mean to prepare data?



There are many sources of data...

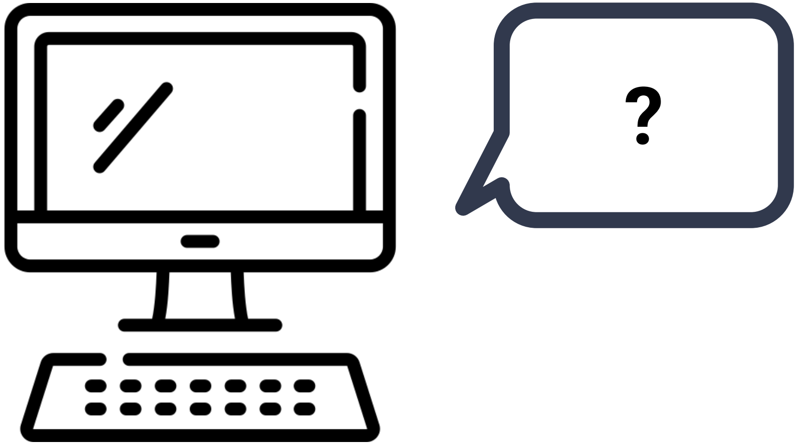


...Which are all subject to uncleanliness

The risks of using unclean data

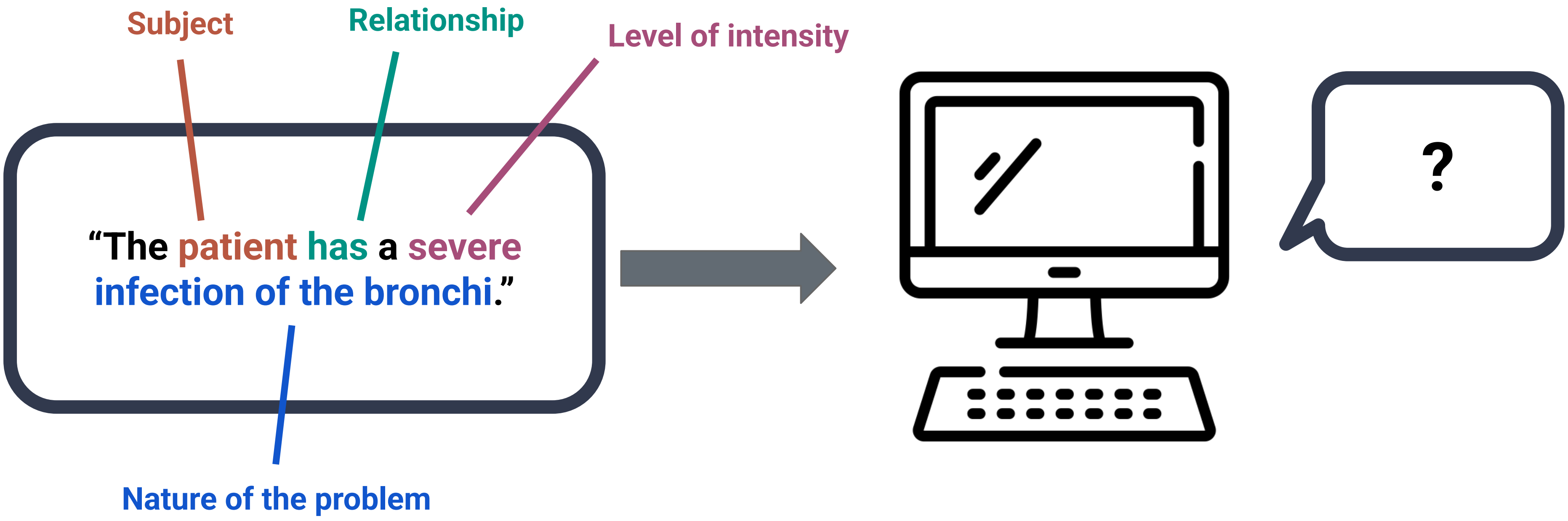
Example #1 : Inability to process data

“The patient has a severe infection of the bronchi.”



The risks of using unclean data

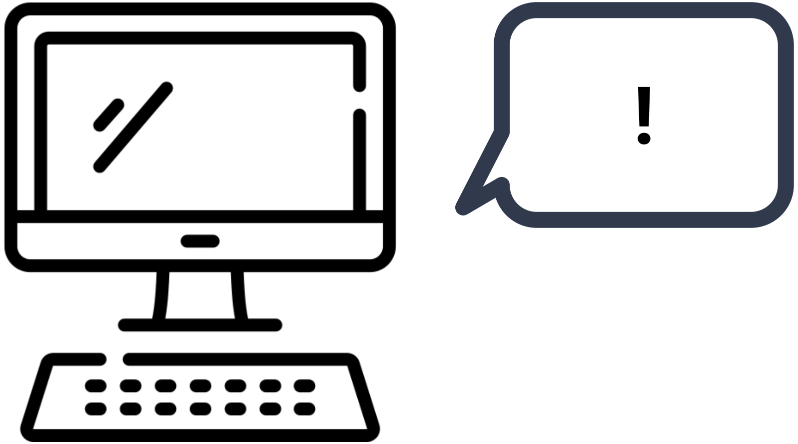
Example #1 : Inability to process data



The risks of using unclean data

Example #1 : Inability to process data

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



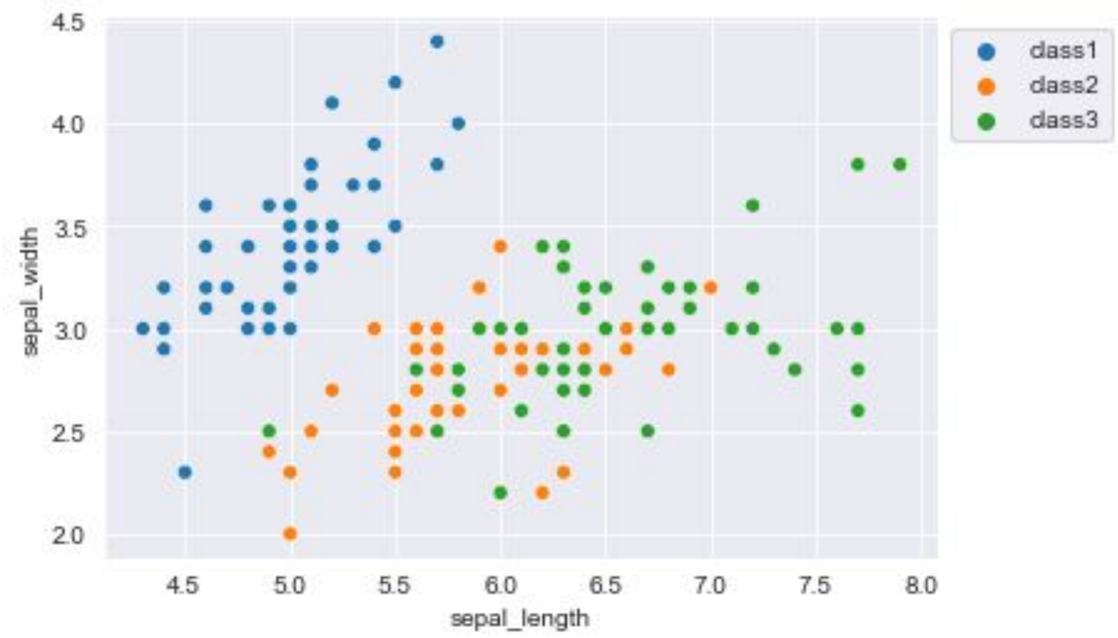
Data must be processed to be usable by a machine

The risks of using unclean data

Example #2 : Difficulty to model the data

An illustration from the first practical

Using only sepal information naively, the classification task is very difficult.



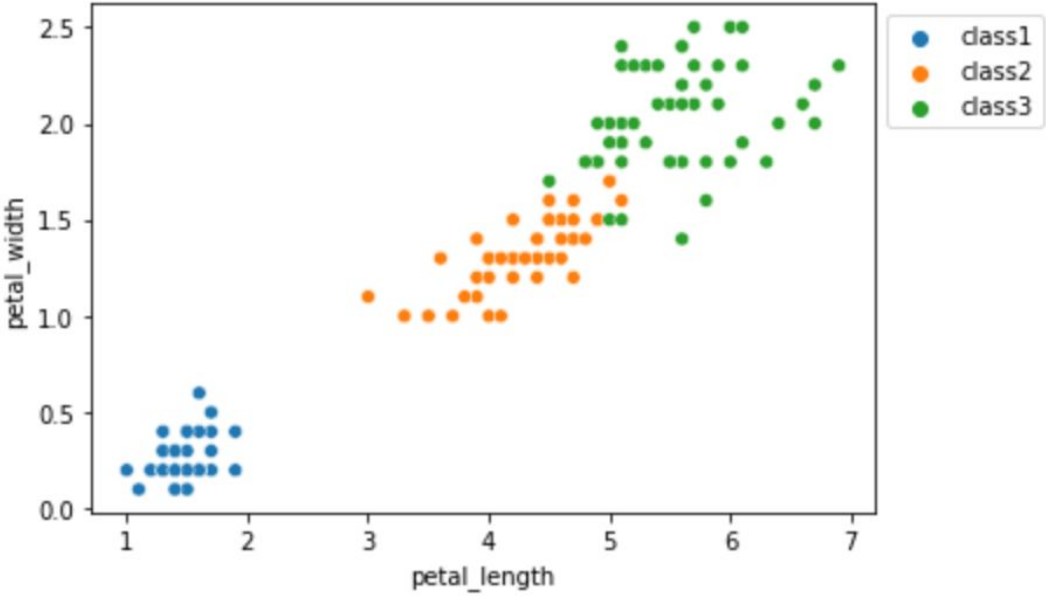
The risks of using unclean data

Example #2 : Difficulty to model the data

An illustration from the first practical

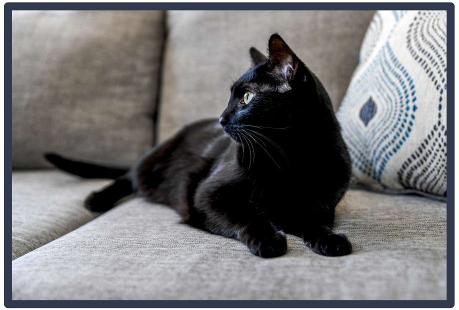
However, using petal information, it is much easier to choose a relevant model for classification.

<u>Using relevant features is essential in machine learning.</u>

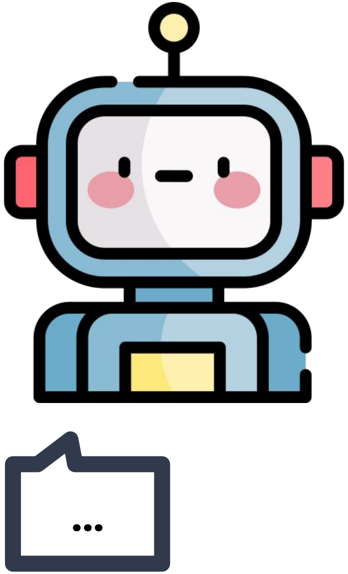


The risks of using unclean data

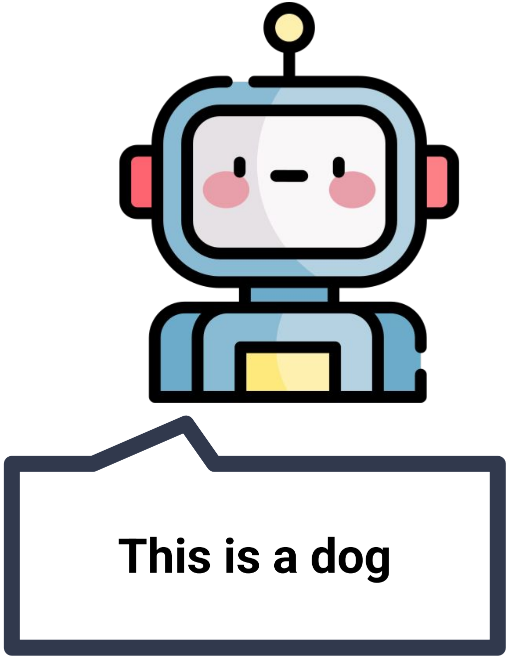
Example #3 : The introduction of bias



Learning



The risks of using unclean data
Example #3 : The introduction of bias



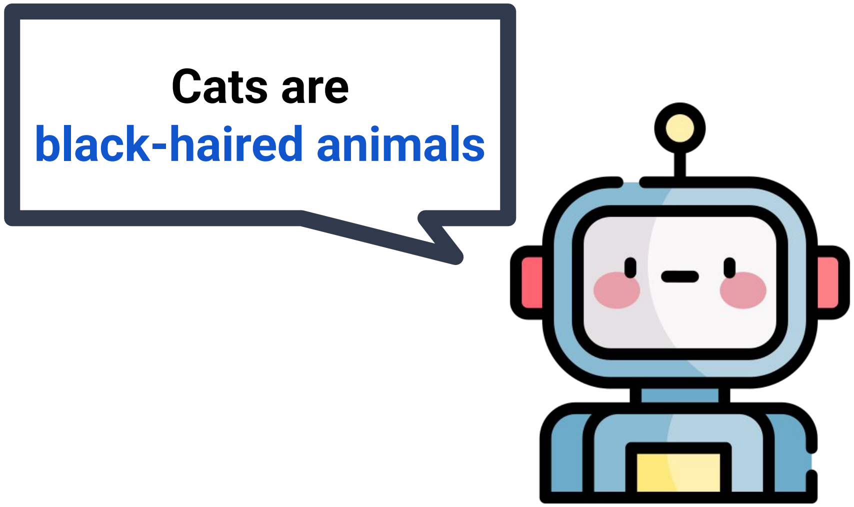
The risks of using unclean data
Example #3 : The introduction of bias

In the previous example, the training set is strongly biased.

Bias can have more severe consequences:

- ❖ Unusability in different regions
- ❖ Discrimination
- ❖ Sexism
- ❖ Maintaining human bias
- ❖ etc.

For the algorithms to **generalize** properly, bias is better avoided in a dataset.



Preparing data is making it exploitable

Raw data is almost always **noisy** and **impractical**

Preparatory work is systematically necessary for machine learning

How does one prepare data?

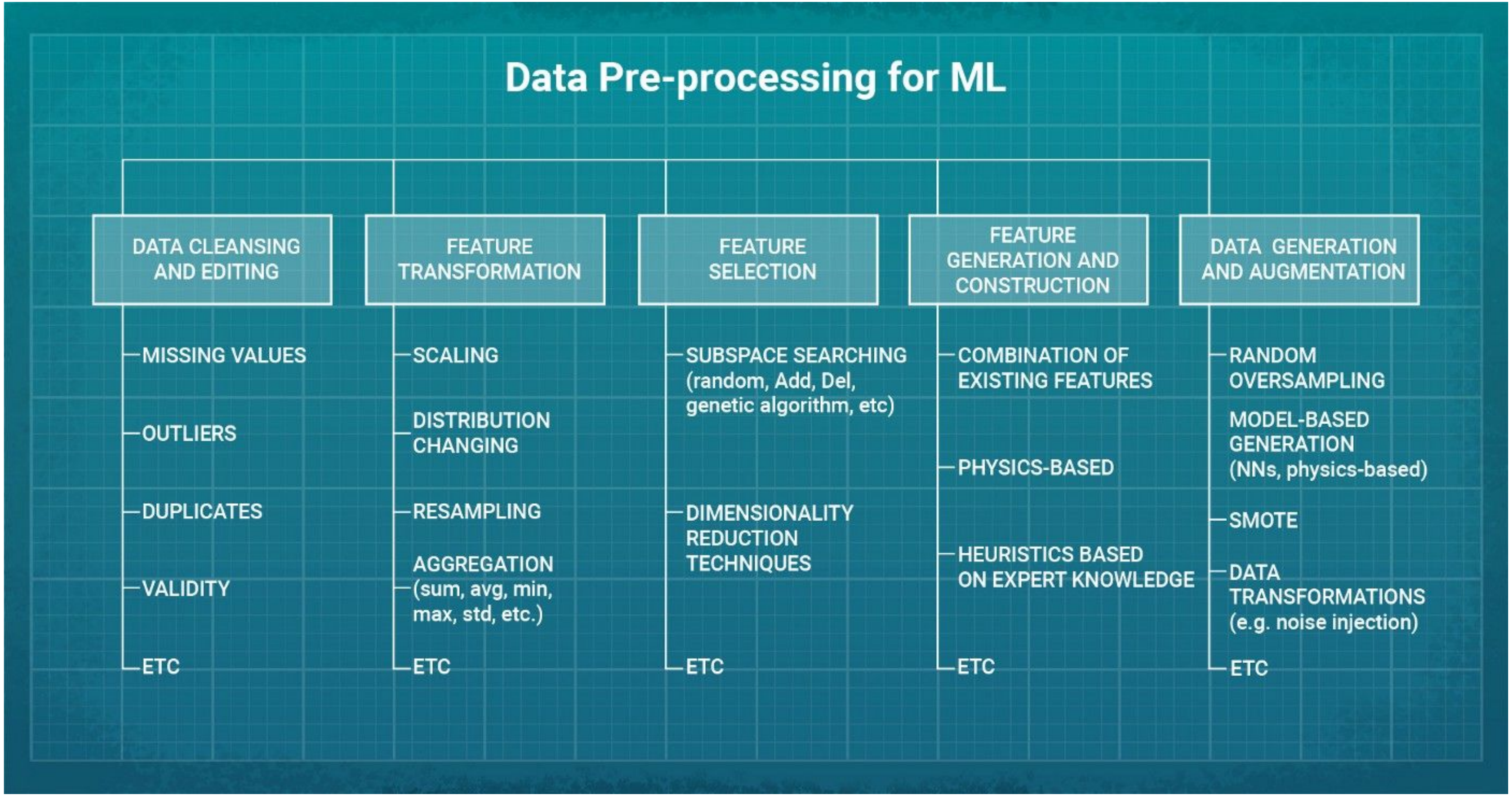
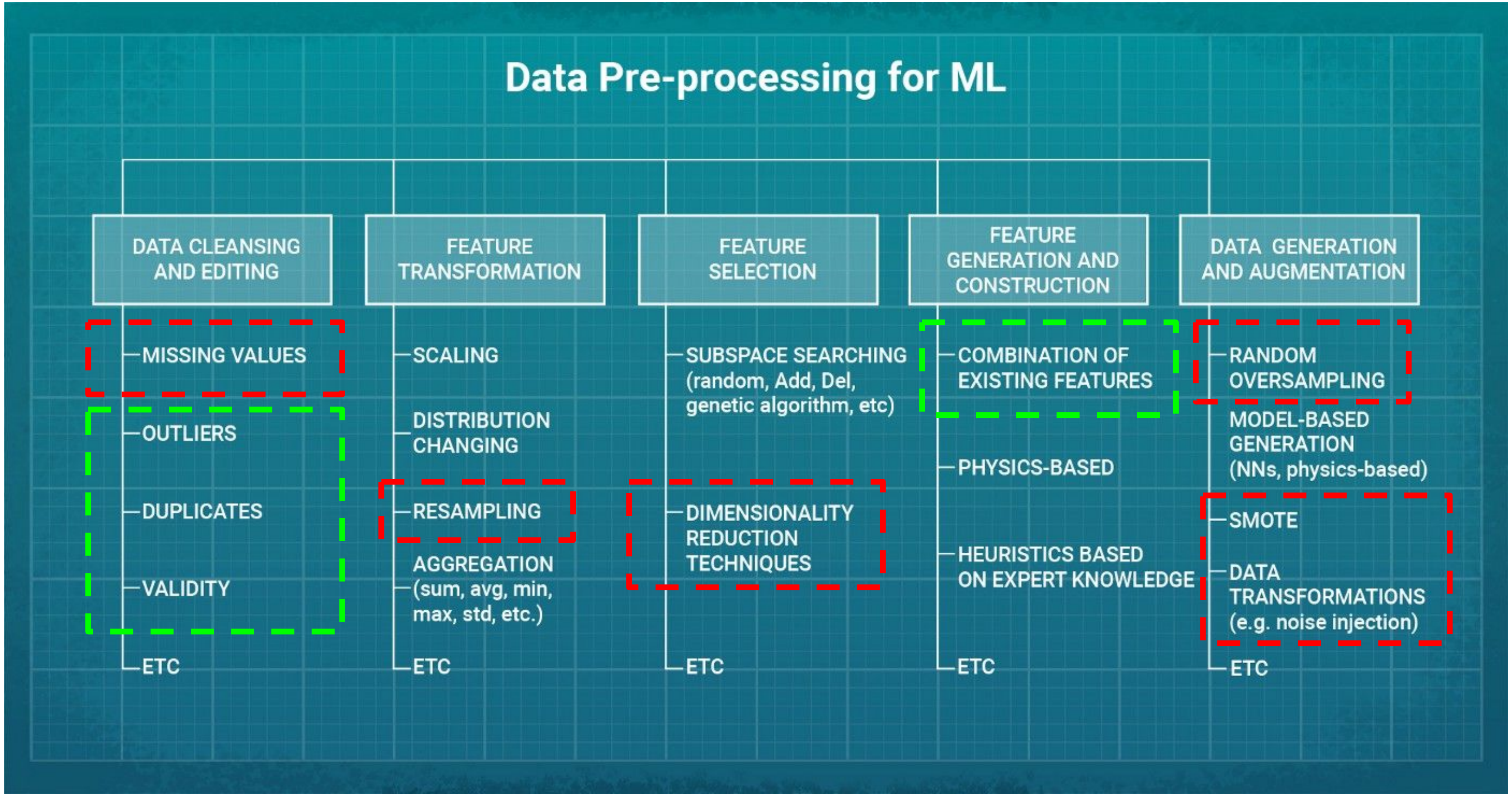


Image source



We have already seen some of those techniques, and we will see more.

Managing missing data

Why can there be missing data?

Why can there be missing data?

There can be several reasons…

- **Technical**
  - Faulty machines
  - Error during encoding
- **Human**
  - Typing errors
  - Deliberate choices (e.g. surveys)
- **Methodological**
  - Data collection not carried out for a certain part of the population (e.g. PSA for women)
  - Lack of measurement techniques

Why can there be missing data?

There are three main categories of missing data

- ***Missing Completely at Random***
  - There is no explainable pattern
  - Example : Human omitting to input data
- ***Missing at Random***
  - Patterns explainable from other columns
  - Example: A survey where men tend to reply less than women
- ***Missing Not at Random***
  - Explainable patterns, but not by observing the other columns
  - Example : People with lower incomes tend not to respond to questions about their salaries

Why can there be missing data?

There are three main categories of missing data

- ***Missing Completely at Random***
  - There is no explainable pattern
  - Example : Human omitting to input data
  - **Data missing for the training process**

- ***Missing at Random***
  - Patterns explainable from other columns
  - Example: A survey where men tend to reply less than women
  - **Generation of bias: The algorithm will generalize better for women**

- ***Missing Not at Random***
  - Explainable patterns, but not by observing the other columns
  - Example : People with lower incomes tend not to respond to questions about their salaries
  - **Generation of bias: The mean salary in the dataset will be inflated**

How can we deal with missing data?

How can we deal with missing data?

Delete lines

Impute values

How can we deal with missing data?

Delete lines

Simple
Can drastically reduce the amount of data
Can introduce bias

Impute values

More robust with more missing data
We keep the “full” dataset
You have to experiment to find the best method
Can introduce bias or inconsistencies

Practical work

Get the latest version of the notebook from GitHub

Mean imputation

By definition, the mean is a value that makes some kind of sense.

It is however computed from observable data and can be influenced by existing bias.

Possible benefits

- Very simple to implement
- Gives a baseline with little effort

Possible pitfalls

- The mean is sensitive to outliers, especially if they are concentrated on one side of the distribution
- It reinforces the weight of the “mean individual”

Median imputation

In balanced datasets, the median tends to be close to the mean.

It is less sensitive to outliers.

NB: Outliers could also be managed specifically (removed or adjusted).

Possible benefits

- Very simple to implement
- Gives a baseline with little effort
- Less sensitive to outliers than the mean

Possible pitfalls

- Ignoring extreme values can be problematic in a dataset with high variance
- It reinforces the weight of the “median individual”

Random value imputation

Using random values can give surprising good results in machine learning.

Studying the distribution of features can help choose a probability distribution to draw from.

NB: This shows the importance of data visualization (cf. `kdeplot`*) !*

Possible benefits

- Not too difficult to implement
- The weight of existing values is not excessively increased

Possible pitfalls

- Finding a relevant probability distribution can be difficult
- The observable distribution could be biased
- Inconsistencies can be introduced in the data

Frequent value imputation

This method is mostly used for non-numerical data.

Similar to numerical data, more intelligent imputation methods can be implemented by studying the distribution of this data.

Possible benefits

- Extremely simple to implement

Possible pitfalls

- Giving too much weight to the most frequent value
- Introducing or maintaining bias

Interpolated value imputation

Interpolation is very useful when the value of a featured is determined by a known function.

Linear and polynomial interpolations are the most common.

NB: Here again, visualization can help.

Possible benefits

- If the actual distribution is close to the function we choose for interpolation, results can be very good

Inconvénients (possibles)

- Not always applicable in practice

Advanced imputation

Scikit-learn offers several imputers, such as `SimpleImputer`, `IterativeImputer`, or `KNNImputer`.

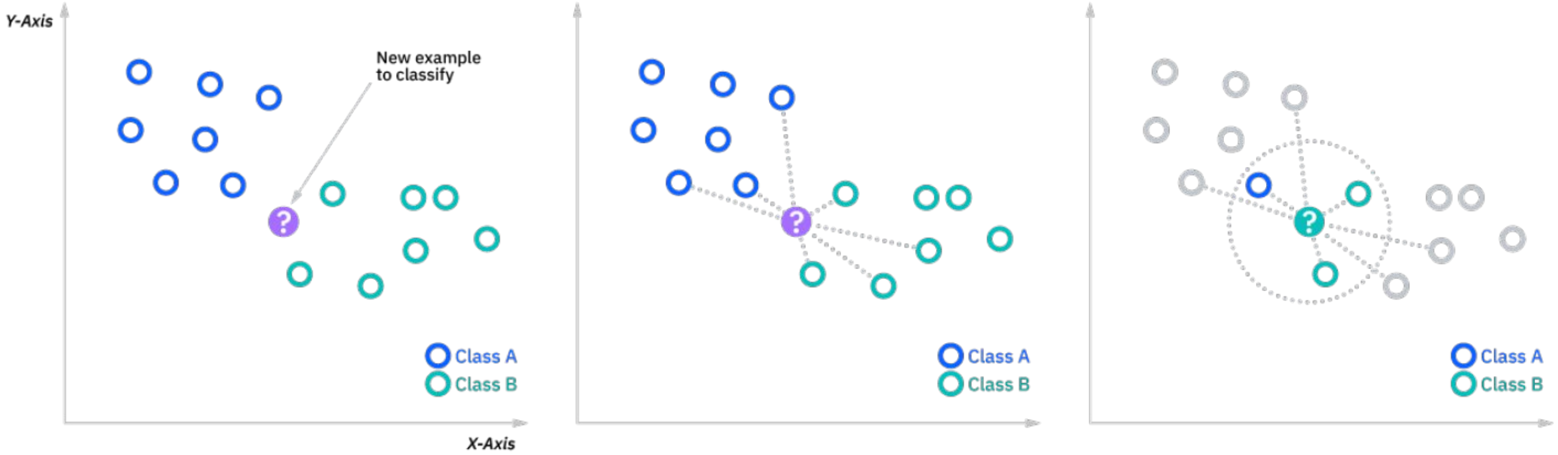
The `SimpleImputer` lets you do what we presented before, whereas the other two are based on machine learning.

Possible benefits

- These methods can help prevent the pitfalls listed before
- They are susceptible to find values that are close to the real ones

Inconvénients (possibles)

- Choosing the imputer is difficult
- The use of machine learning requires data to have been processed to some extent



Introduce a new example

Compute distances

Majority vote

The K-nearest-neighbours (KNN) algorithm for data imputation

Source : IBM

K-nearest-neighbours

KNN is a very simple classification algorithm that can provide good results in some cases.

It can be used for data imputation.

Possible benefits

- Easy to implement
- Few hyperparameters

Possible pitfalls

- Choosing a distance is not always easy
- Can become computationally expensive
- Sensitive to the curse of dimensionality
- Sensible to overfitting

Don’t forget to upload your work!

Debrief

Debrief

What did we learn today?

What could we have done better?

What are we doing next time?

Data Science

Session 3 - Cleaning data & Missingness

hadrien.salem@centralelille.fr

introduction-to-data-science