

# Data and Information Quality Projects

# Evaluation

---

You must deliver a **.zip** folder named with your surnames  
(example: Sancricca\_Sancricca\_Sancricca.zip) containing:

1. A **report** of few pages — more details on writing the report at the end of the document
2. The **code** you made (.py, or .ipynb) — better if well-commented
3. The **dataset** you have **cleaned**

# Deadline

---

The first exam call (17/01/2025)

# The Project

---

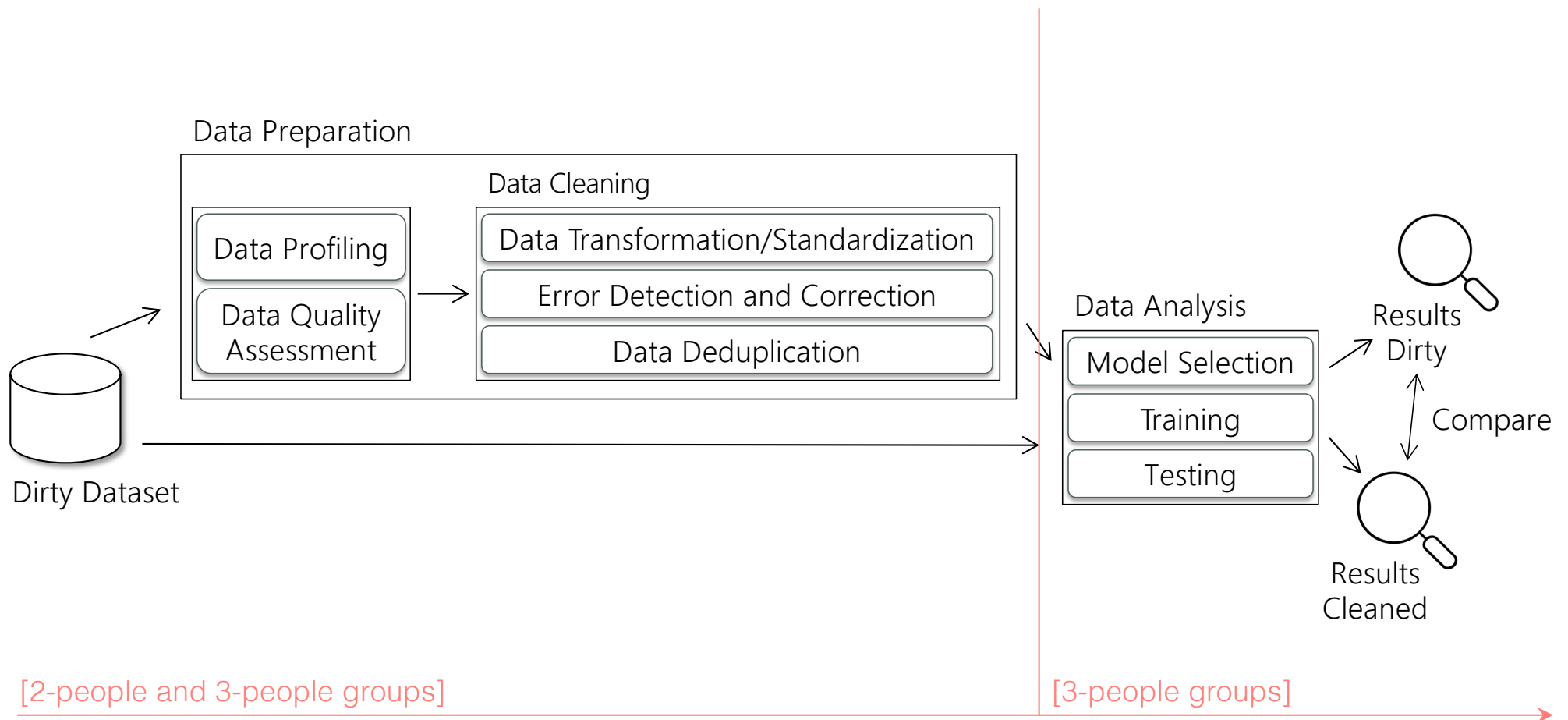
1/2-people groups: Data Preparation Pipeline

3-people groups: Data Preparation + Data Analysis Pipelines

A) We will assign a different **dirty dataset** to each group.

B) You must **execute** a complete **Data Preparation Pipeline** on the assigned dataset.

# Project Pipelines



## Data Analysis [only for 3-people groups]

---

- a. Choose the type of analysis (classification-regression-clustering):
  - i. Choose one column as the target column (categorical = **classification** OR numerical = **regression**)
  - OR
  - ii. Perform unsupervised **clustering** analysis
- b. Perform a data analysis pipeline on (1) the dirty dataset and (2) the cleaned dataset (model selection, training and testing)
- c. Compare the results using the right performance metrics (Precision, Recall, F1, etc. [Classification], MSE, RMSE, etc. [Regression], Silhouette, etc. [Clustering])

**N.B.** Some datasets that we will assign are not specifically made for machine-learning analysis! It is OK if the performance is very low. The important thing is that the dataset is cleaned properly and the pipeline is complete.

# Project Report (few pages)

---

PROJECT ID

ASSIGNED DATASET

STUDENTS (NAME SURNAME ID)

## 1. SETUP CHOICES

Describe the setup choices made: libraries, data preparation techniques used, etc.

## 2. PIPELINE IMPLEMENTATION

Describe all the pipeline steps in detail: what did you find from the data exploration? How did you decide to use it in the data preparation phase? Why did you use specific data preparation technique?

## 3. RESULTS

Discuss the main results obtained: verify the desired quality level has been achieved, compare the data analysis results [only for 3-people groups]

**Very important** Justify your choices! (for example, why you have chosen a specific data preparation technique for a specific column than all those seen in the lectures?)



[camilla.sancricca@polimi.it](mailto:camilla.sancricca@polimi.it)

for any additional information write me 😊