DIQ Course
Project Assignment

Projects Info

- ➢ The project gives you the opportunity to obtain a maximum of <u>3 additional points</u>

- ➢ Evaluation:
  - We ask you to write a report ..
    - ✓ Setup choices
    - ✓ Pipeline implementation (highlighting the TODO phases)
    - ✓ Results discussion (supported by plots and tables)
  - .. and to deliver the code you made (.py, or .ipynb)

Projects Objective

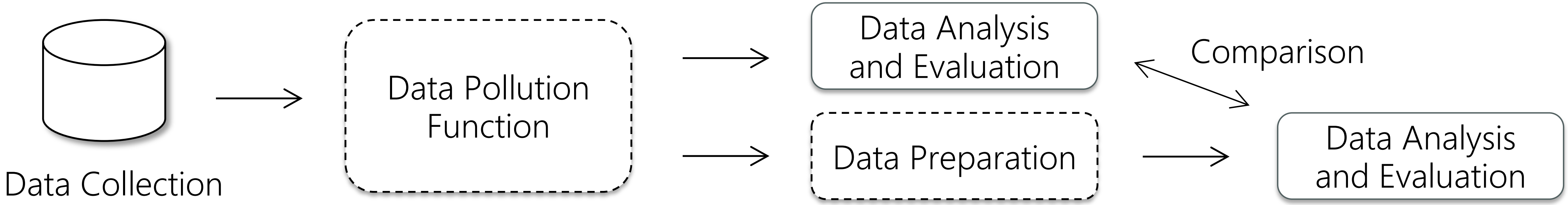
- Data Quality (DQ) is becoming increasingly important for successful Machine Learning (ML) analysis pipelines.

- However, requirements for having a good DQ are changing: we must no longer just ensure a good level of DQ for the traditional aspects, such as Completeness, Accuracy, or Consistency.

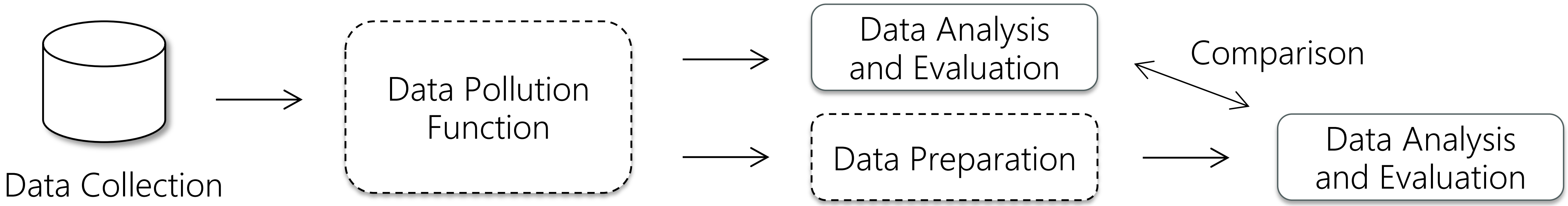
- The success of a ML analysis can depend a multitude of new data issues, such as Dimensionality, Feature Dependency, or Distinctness.

➢ The **goal of the DIQ Project** is to investigate the impact of both the ''traditional'' and ''new'' DQ issues on a ML analysis.

Projects Pipeline

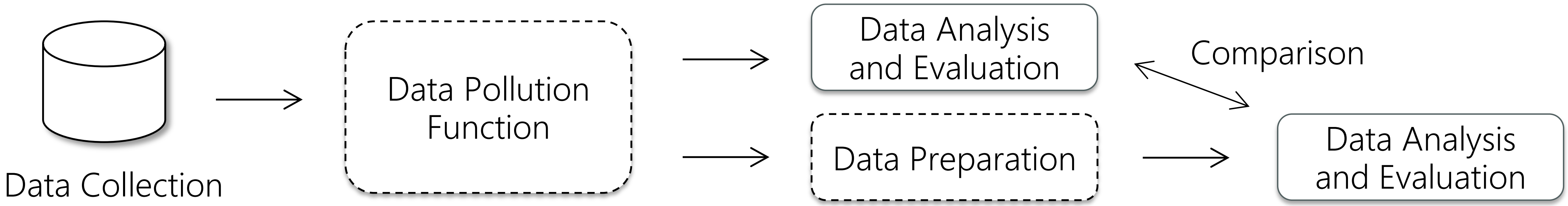


Projects Pipeline



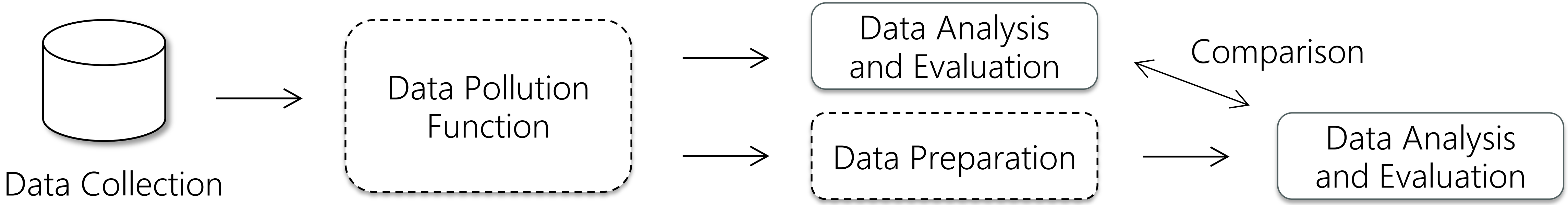
1. **Data Collection** (dataset.make_classification/regression/clustering) GIVEN
   - Fixed default parameters
   - Can be changed according to the needs of the DQ issue/s to be injected)

Projects Pipeline



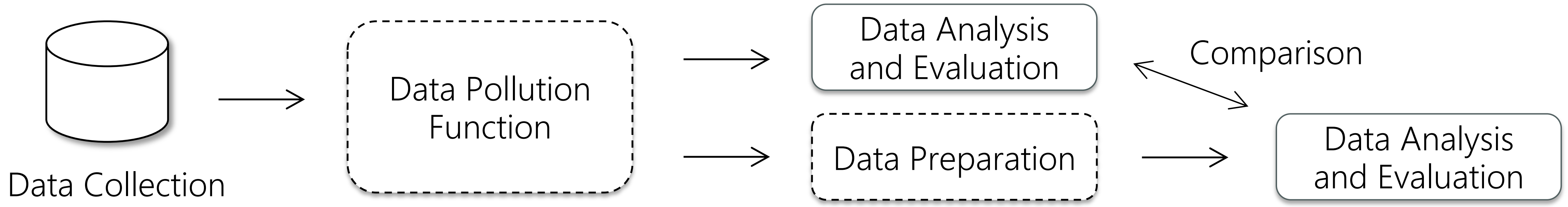
1. **Data Collection** (dataset.make_classification/regression/clustering) GIVEN
2. **Data Pollution Function** TODO
   - ➢ Inject errors/values related to the assigned DQ issue at different (%)
   - ➢ Combined with dataset.make to inject the assigned DQ issue/s

Projects Pipeline



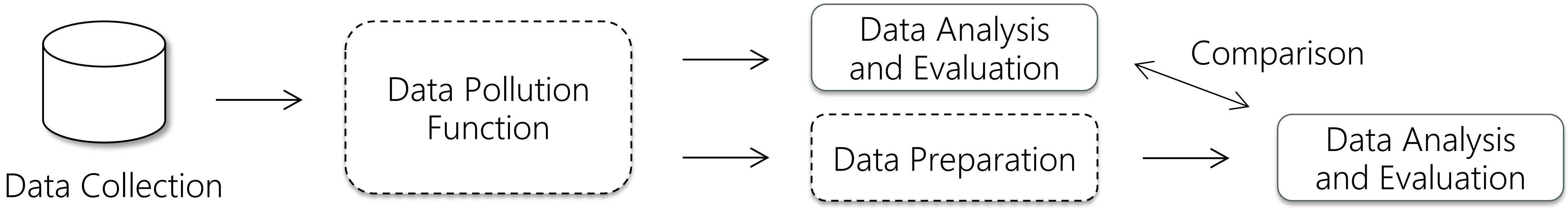
1. **Data Collection** (dataset.make_classification/regression/clustering) GIVEN
2. **Data Pollution Function** TODO
3. **Data Analysis and Evaluation** GIVEN
   - Metrics: Performance, Overfitting, Speed
   - Creation of plots and tables with the numeric results

Projects Pipeline



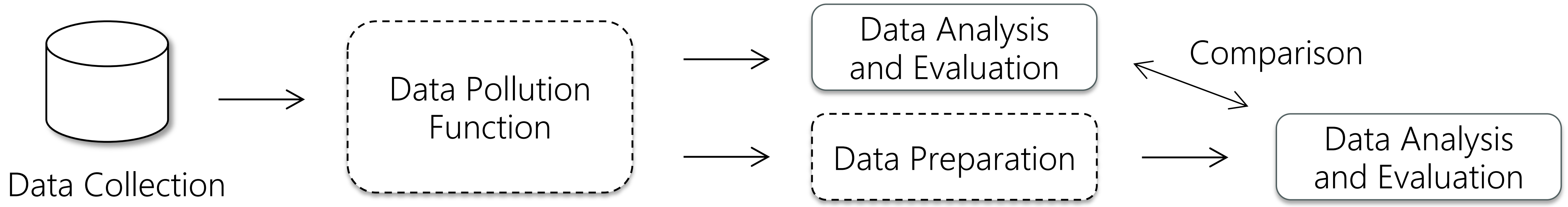
1. **Data Collection** (dataset.make_classification/regression/clustering) GIVEN
2. **Data Pollution Function** TODO
3. **Data Analysis and Evaluation** GIVEN
4. **Data Preparation** TODO
   - Apply different DQ improvements to correct the injected DQ issue
   - Could be requested or not, depending on the assigned DQ issue/s

Projects Pipeline



1. **Data Collection** (dataset.make_classification/regression/clustering) GIVEN
2. **Data Pollution Function** TODO
3. **Data Analysis and Evaluation** GIVEN
4. **Data Preparation** TODO
5. **Data Analysis and Evaluation (again)** GIVEN
   - ➢ Metrics: Performance, Overfitting, Speed
   - ➢ Creation of plots and tables with the numeric results

Projects Pipeline



1. Data Collection (dataset.make_classification/regression/clustering) GIVEN
2. Data Pollution Function TODO
3. Data Analysis and Evaluation GIVEN
4. Data Preparation TODO
5. Data Analysis and Evaluation (again) GIVEN
6. Compare the obtained results TODO

Possible DQ issues

1. Completeness (MNAR and MCAR)
2. Accuracy (Noise)
3. Feature Dependency (Redundancy)
4. Variables types
5. Distinctness (or Irrelevancy)
6. Duplication (not-exact)
7. Dimensionality (#columns, #rows)

ML Tasks

1. Classification (6 algorithms)
2. Regression (6 algorithms)
3. Clustering (5 algorithms)

2 PERSON: 2 DQ ISSUES AND 1 ML TASK

1 PERSON: 1 DQ ISSUE AND 1 ML TASK

➢ We will give you guidelines on what the expected output should look like camilla.sancricca@polimi.it

for any additional information write me ☺