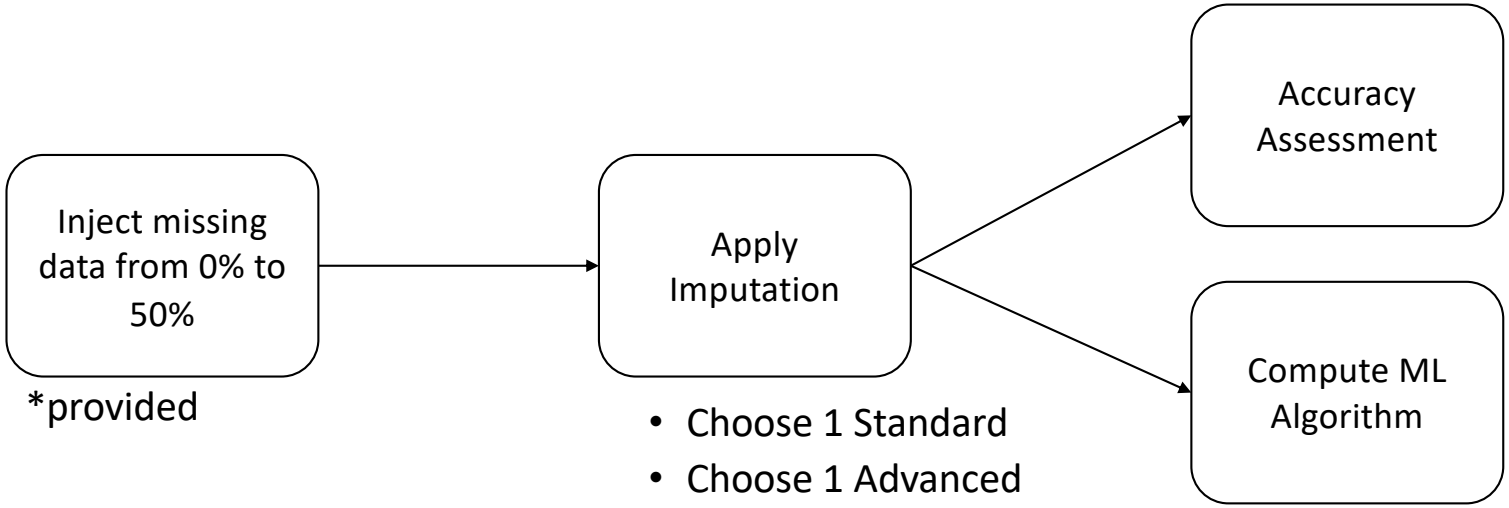
Data and Information Quality Projects

You can choose between:

- Project 1
- Project 2
- Project 3 (only for 2-students projects)

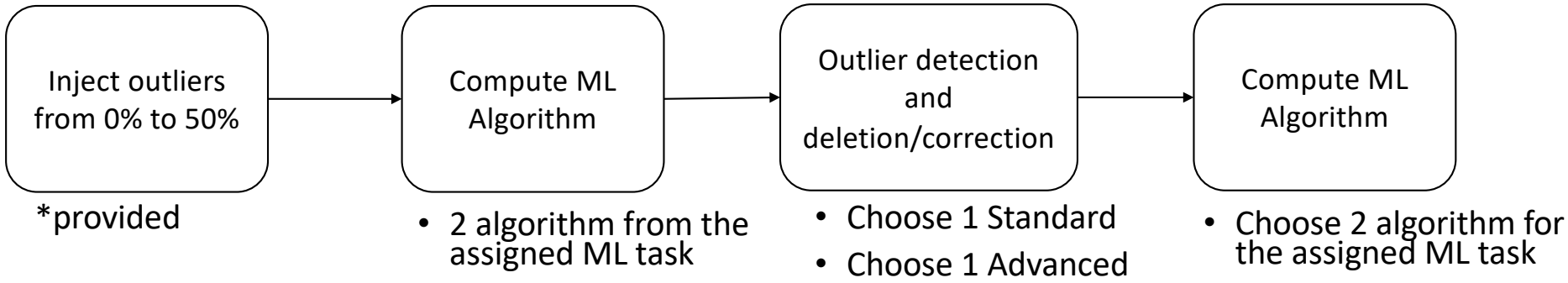
Project 1

- Assignment:
  - 1 Student: 1 Dataset – Classification Or Clustering task
  - 2 Students: 2 Dataset – Classification Or Clustering task



Project 2

- Assignment:
  - 1 Student: 1 Dataset – Classification Or Clustering task
  - 2 Students: 2 Dataset – Classification Or Clustering task



Project 3

- Assignment:
  - 2 Students: 1 Dirty Dataset

Imputation Techniques

STANDARD TECHNIQUES

- Mean
- Median
- Replace with standard values (0, “Missing”, etc)
- Mode
- Previous value
- Next values
- Random

ADVANCED TECHNIQUES

- MICE (Multiple Imputation by Chained Equations)
- Maximum Likelihood (ML)
- EM (Expectation Maximization)
- Linear/Logistic Regression
- Naive Bayes
- Gaussian Processes
- Kernel-based
- PCA (Principal Component Analysis)
- SVD (Singular-Value Decomposition)
- MF (Matrix Factorization)

  Association Rules
- KNN (k-Nearest-Neighbors)
- Decision Trees
- MLP (MultiLayer-Perceptron)
- RF (Random Forest)
- SVM (Support Vector Machines)

Outlier Detection Techniques

STANDARD TECHNIQUES

- Standard Deviation
- Z-score
- Interquartile Range (IQR)
- Histograms, box-plots
- Hypothesis testing
- Local Outlier Factor (LOF)
- Kernel Density Estimation

ADVANCED TECHNIQUES

- KNN (k-Nearest-Neighbours)
- Connective-based Outlier Factor (COF)
- Regression Models
- Gaussian Models
- Regression Methods
- Clustering-based methods (k-Means, DB-Scan, HCA, etc)
- Learning-based methods (Active learning,
- Deep learning, Graph-based, etc)
- Ensemble-based methods (Bagging, Boosting, etc)

Questions?