DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

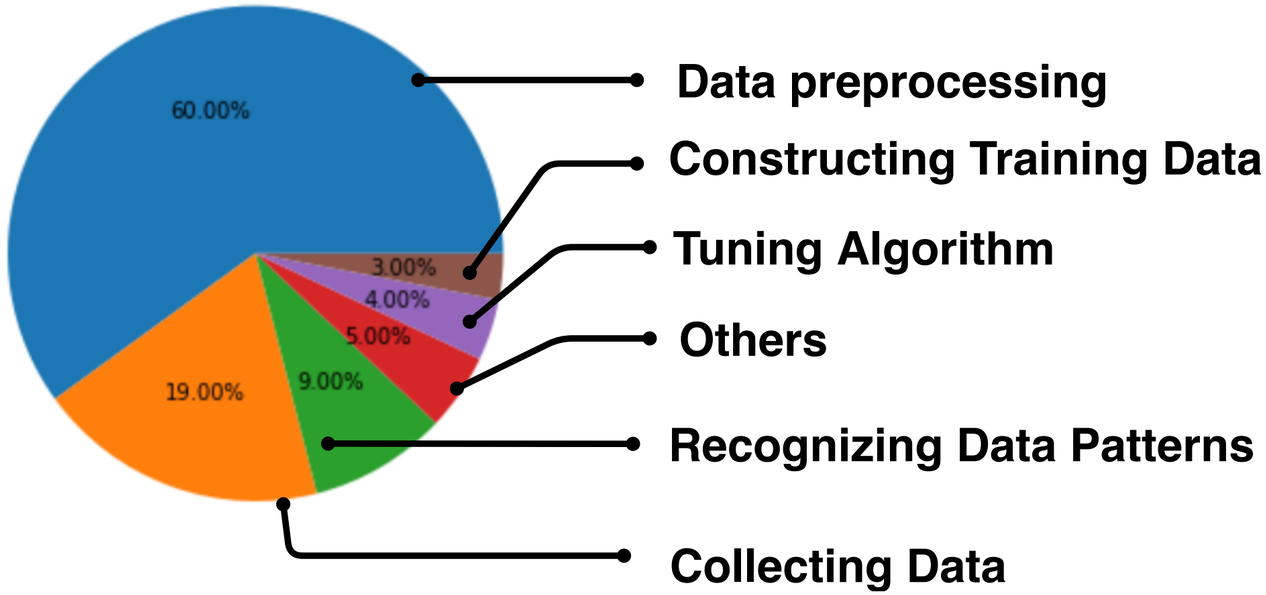
DLI Accelerated Data Science Teaching Kit

Lecture 3.1 - Introduction to Data Pre-processing

Data Pre-processing

Where do data scientists spend their most of time for data analytics?

- Data scientists spend more than 50% of their time on **Data preprocessing.**
- **Collecting data** is the second most time-consuming component.
- **Tuning algorithm** occupies a small part.



Why Do We Need Data Pre-processing?

Data is rarely clean and often has data quality issues.

- **Incomplete**: data lacks attributes or contains missing values.
- **Noisy**: data contains incorrect records such as outliers.
- **Inconsistent**: data contains conflicting records or discrepancies.

Why Do We Need Data Pre-processing?

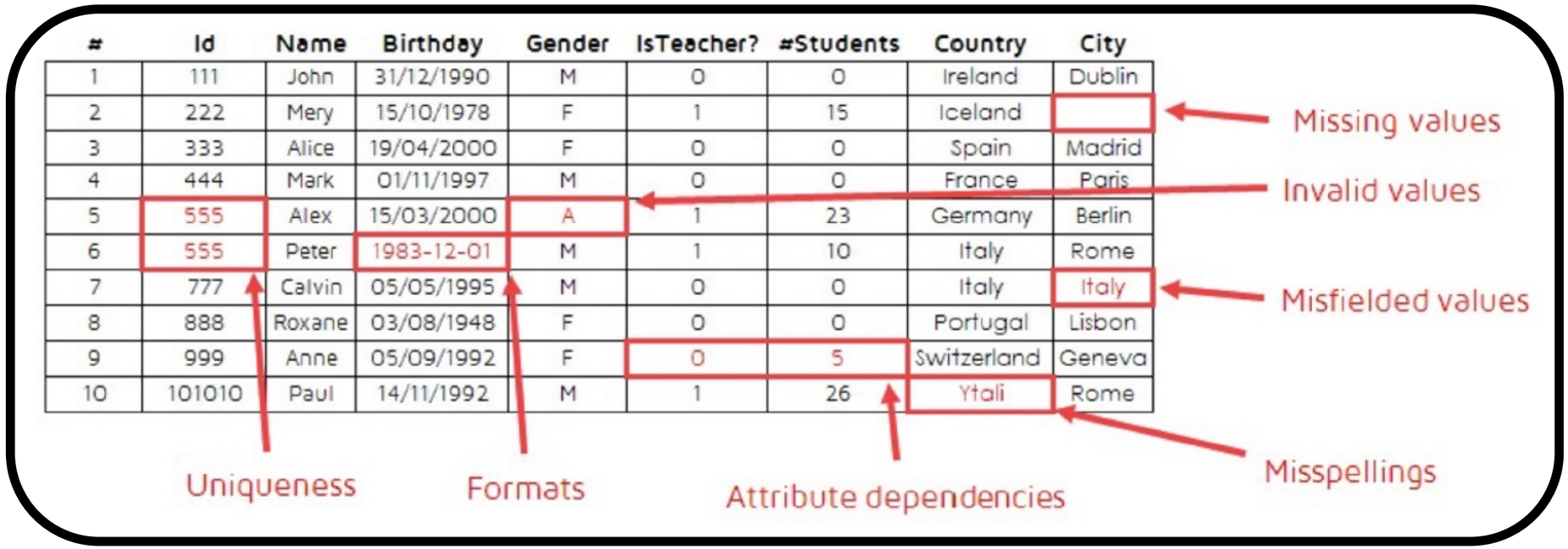
Data from the real world is dirty.

- **Missing values:** some attributes in the collected data would have blank or NULL values.
- **Invalid Values:** some well-know attributes such as gender may have incorrect values.
- **Uniqueness**: repeated values of the same identifiers.
- **Misspellings:** incorrectly written values
- … …

Why Do We Need Data Pre-processing?

Data from the real world is dirty.

- **Missing values**
- **Invalid Values**
- **Uniqueness**
- **Misspellings**
- … …

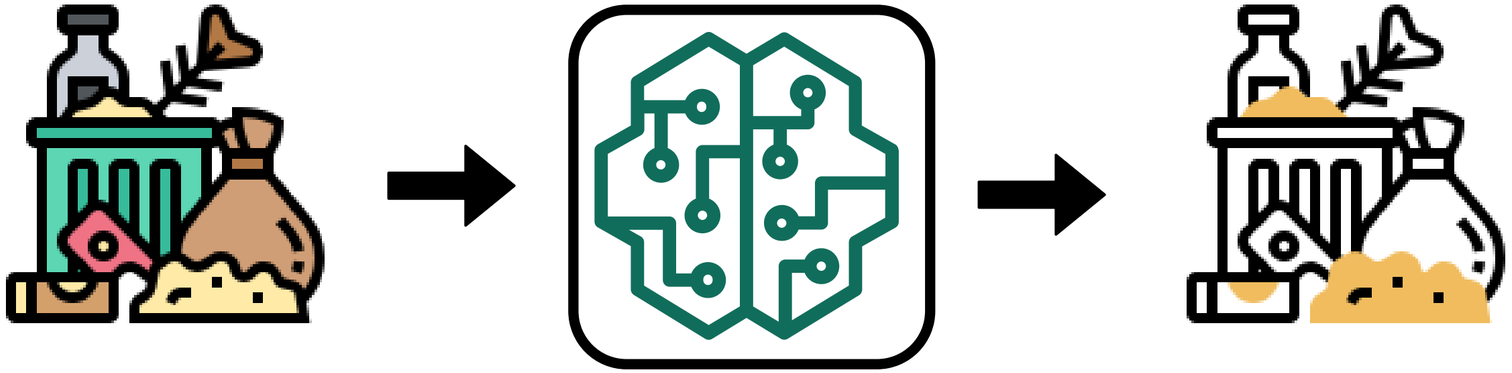


An example of dirty data

Goal of Data Pre-processing

To avoid “Garbage in, Garbage out” for data analytics

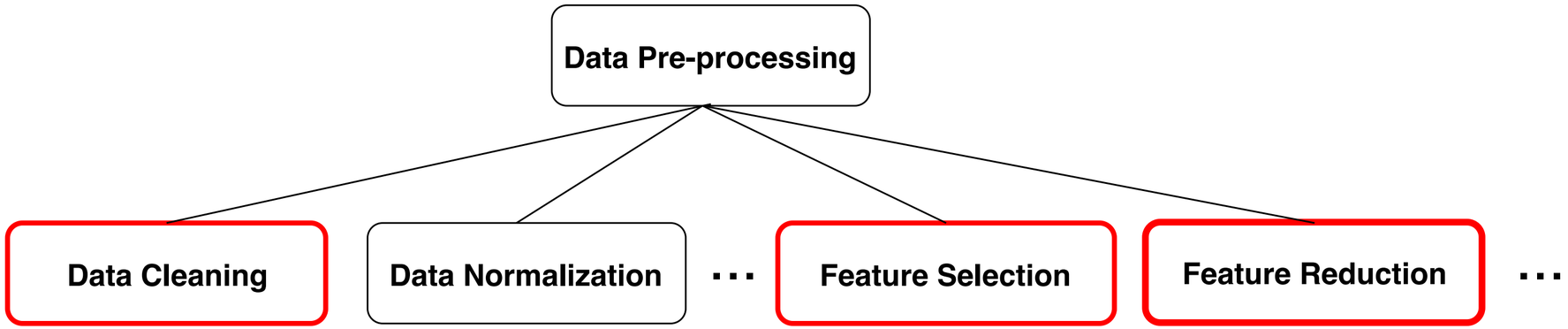
- Dirty data as input
- Useless results as output
- Meaningless work in model construction
- Failed projects/tasks



Goal of Data Pre-processing

To avoid “Garbage in, Garbage out” for data analytics

- Data Cleaning and/or Statistical Preprocessing
- Feature Selection
- Feature Reduction
- … …



DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You