NATIONAL ECONOMICS UNIVERSITY
FACULTY OF MATHEMATICAL ECONOMICS
PYTHON
DATA
Preparation & Visualization

Foundations of Data Preparation for Machine Learning

Lecturer: Dr. Nguyen Tuan Long

Email: ntlong@neu.edu.vn

Mobile: 0982 746 235

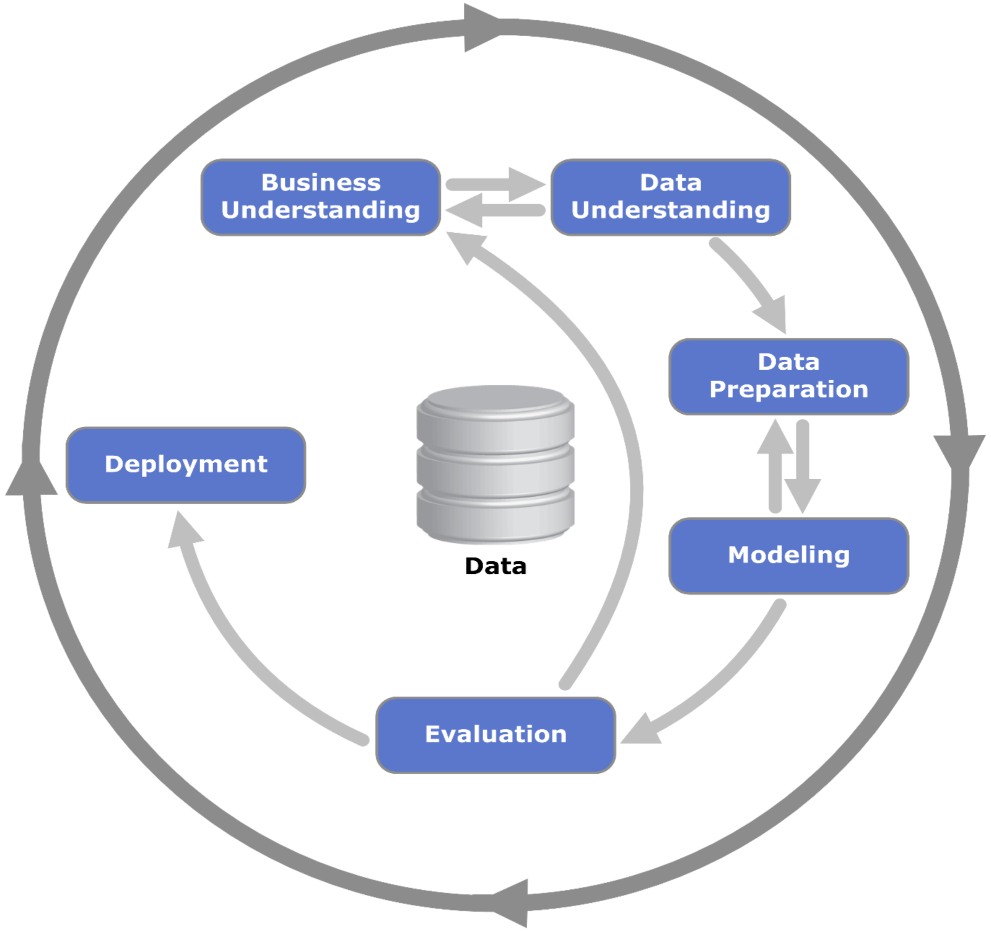
Agenda

Main Content

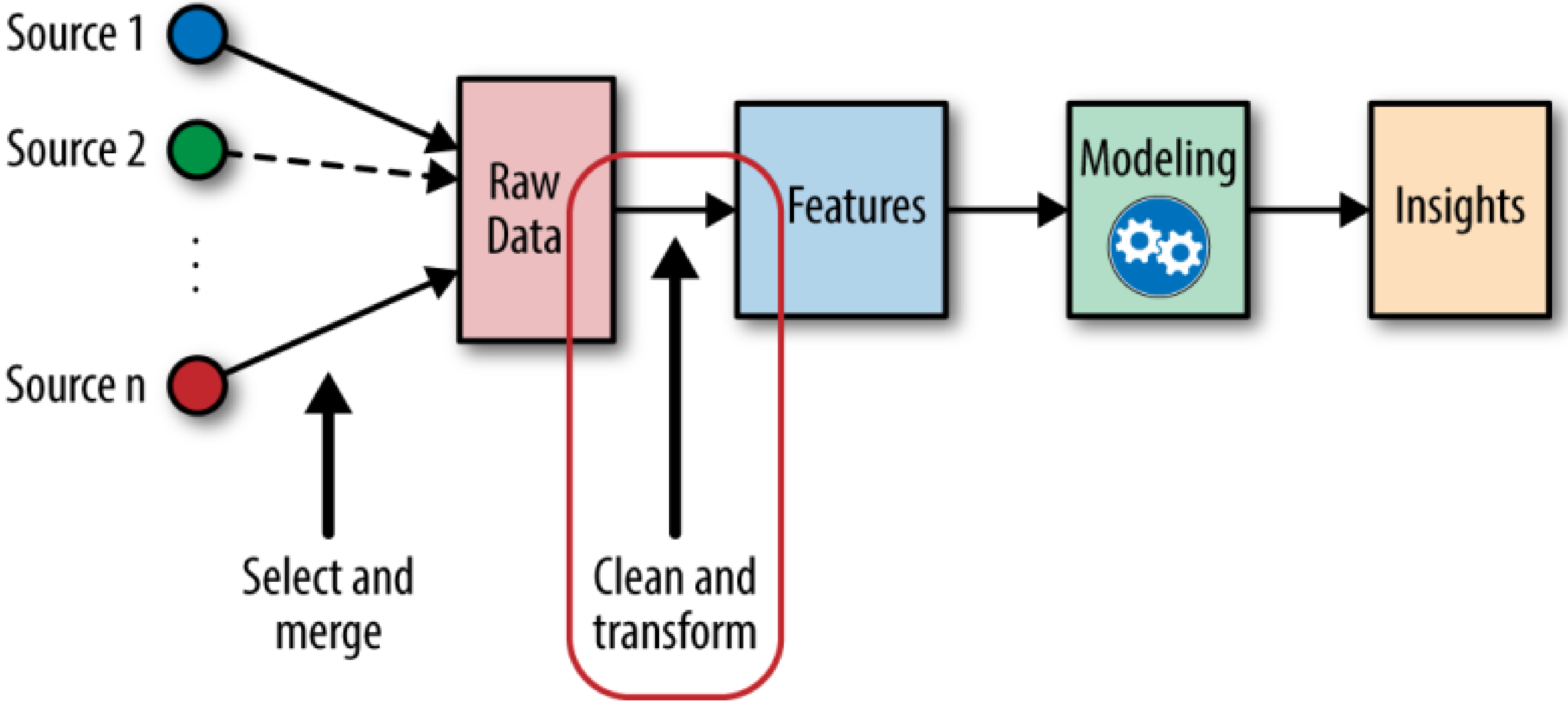
1. **The Role** of Data Preparation in a Machine Learning Project.
2. The Data Preparation **Process** Overview.
3. **Core Concept:** Data Leakage and How to Avoid It.

1. The Role of Data Preparation

A Machine Learning project is not a straight line, but an iterative cycle.



1. The Role of Data Preparation



1. Four Steps in a Project

Step 1: Define Problem

- Understand the problem, collect data, choose a metric.

Step 2: Prepare Data

- Clean, transform, and select features.
- The **foundation** for the next steps.

Step 3: Evaluate Algorithms

- Test multiple models, use cross-validation.

Step 4: Finalize Model

- Train the final model and deploy it.

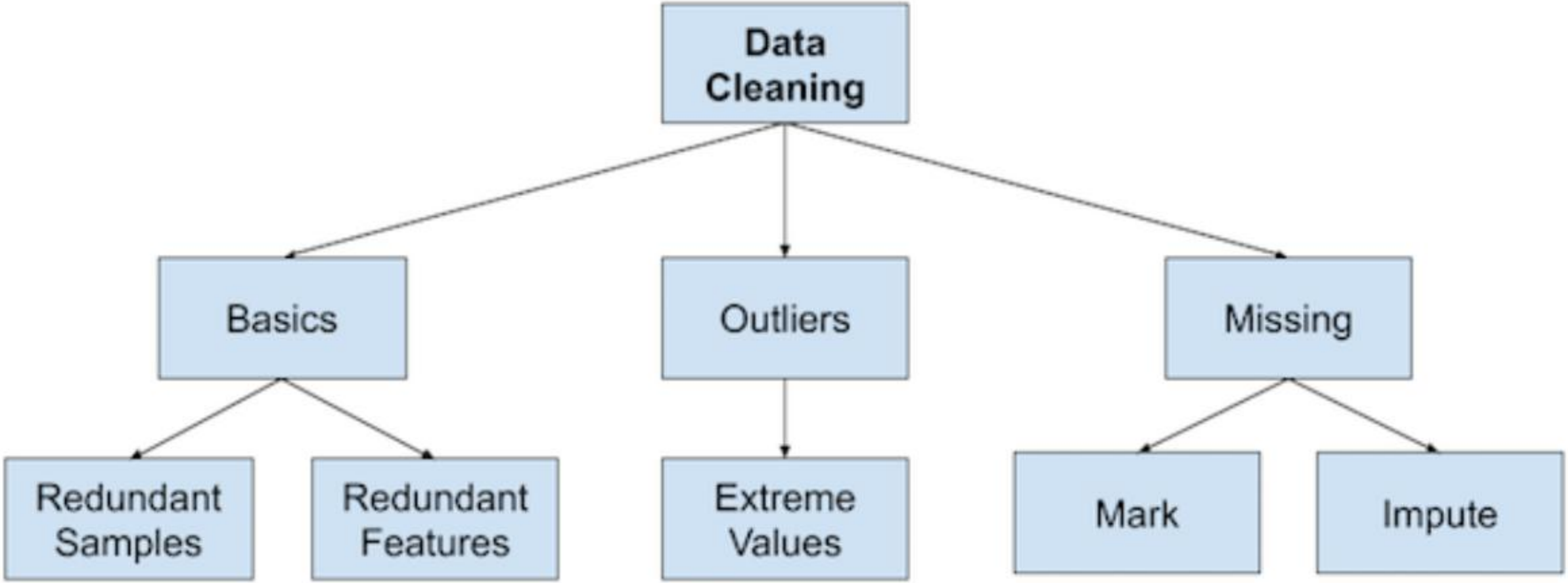
2. The Data Preparation Process

5 Main Task Groups

1. Data Cleaning
2. Feature Selection
3. Data Transforms
4. Feature Engineering
5. Dimensionality Reduction

2.1. Data Cleaning

Overview of Data Cleaning



Data Cleaning

Basics

Outliers

Missing

Redundant Samples

Redundant Features

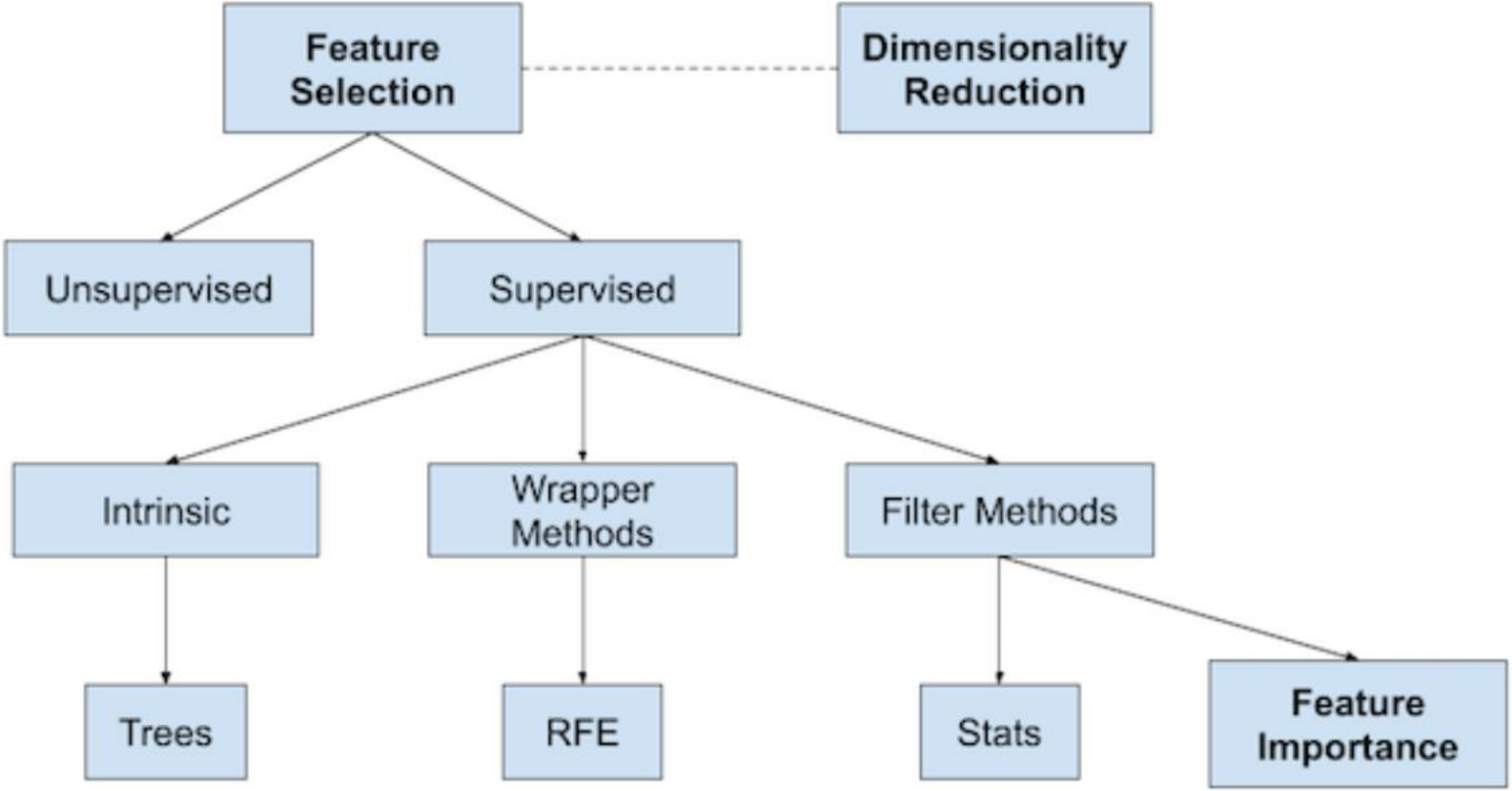
Extreme Values

Mark

Impute

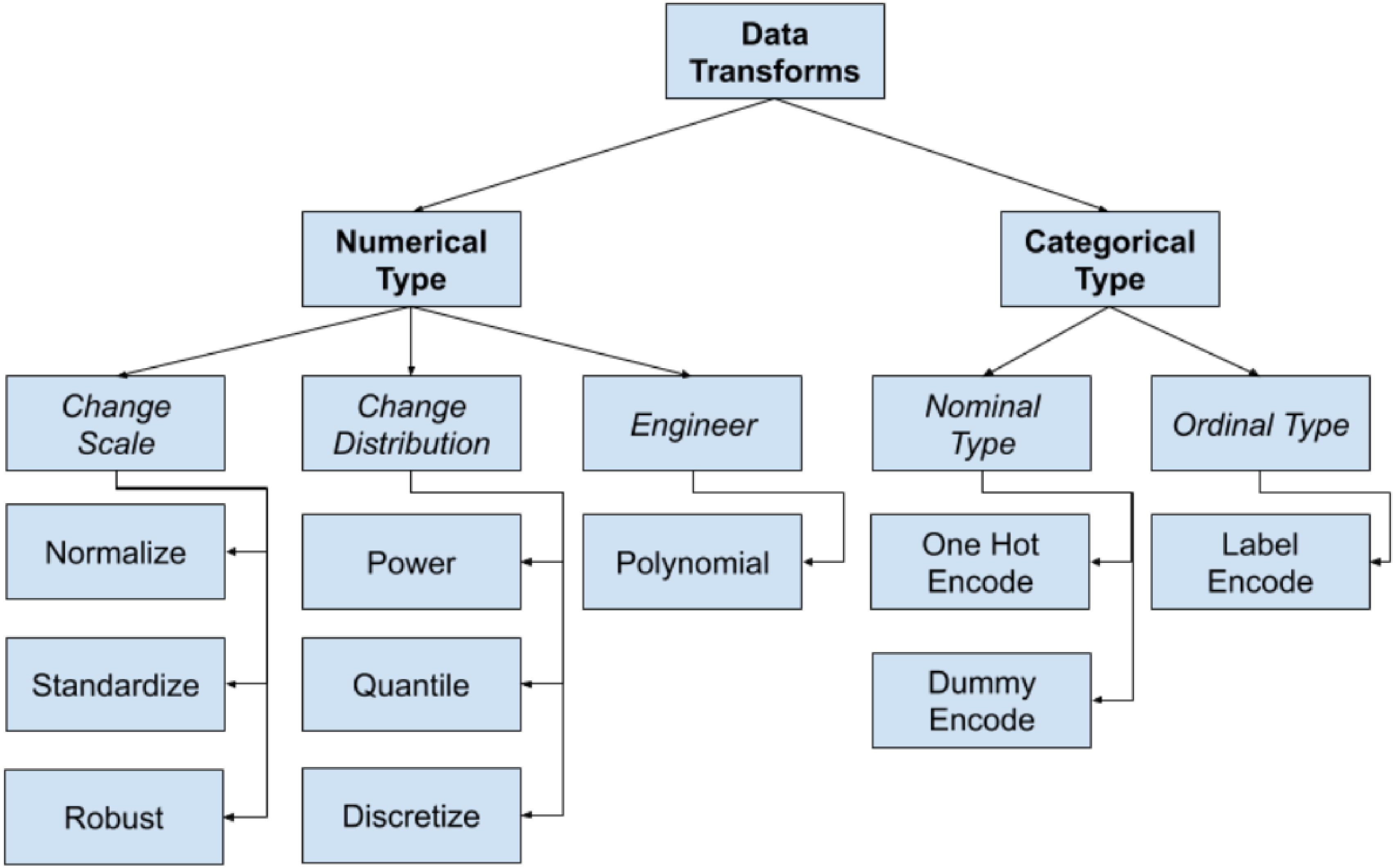
2.2. Feature Selection

Overview of Feature Selection Techniques



2.3. Data Transforms

Overview of Data Transforms



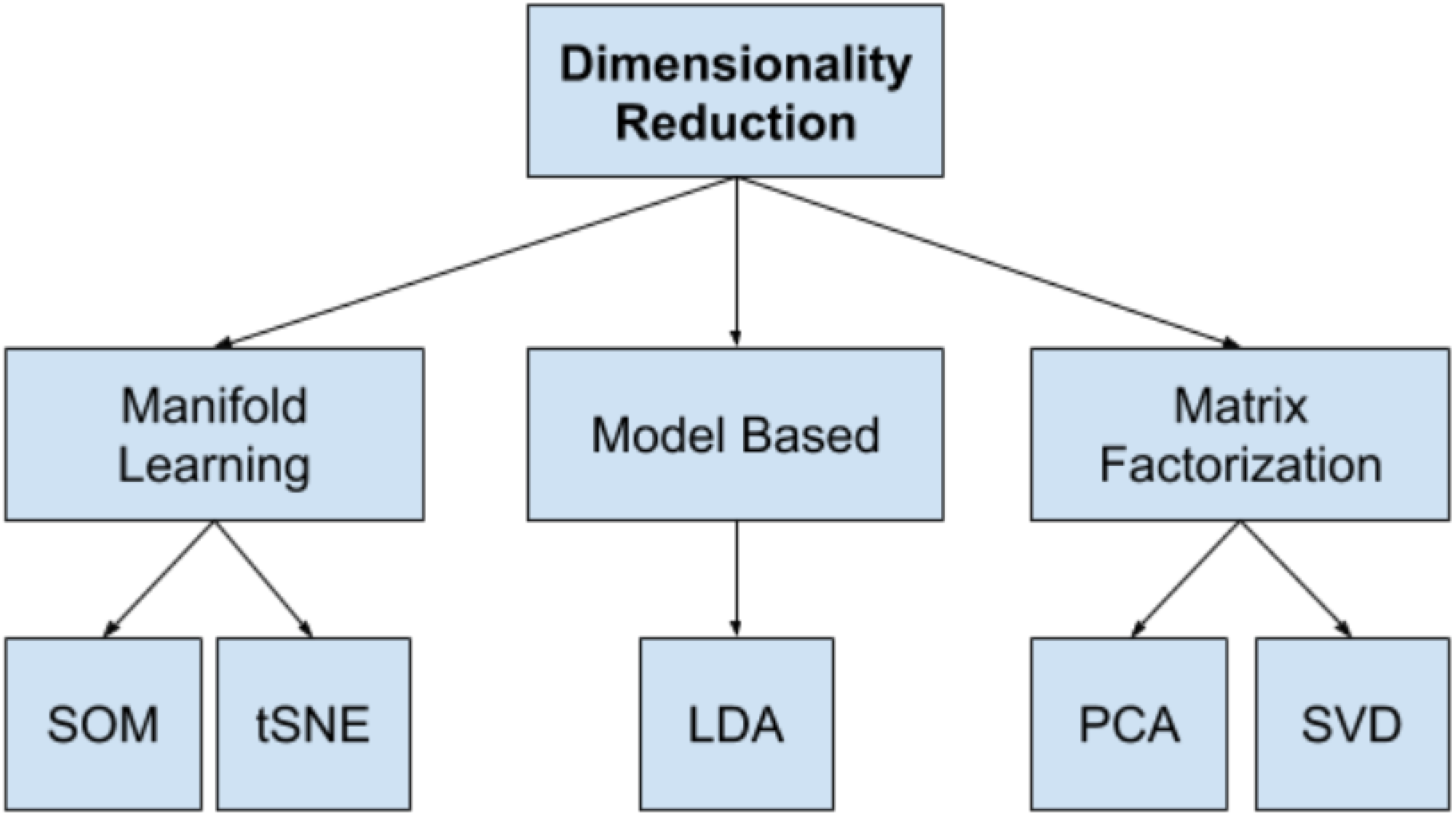
- Data Transforms
  - Numerical Type
    - *Change Scale*
      - Normalize
      - Standardize
      - Robust
    - *Change Distribution*
      - Power
      - Quantile
      - Discretize
    - *Engineer*
      - Polynomial
  - Categorical Type
    - *Nominal Type*
      - One Hot Encode
      - Dummy Encode
    - *Ordinal Type*
      - Label Encode

2.4. Feature Engineering

- **Purpose:** To create new features from existing data.
- **Examples:**
  - Split a date-month-year column into day, month, year.
  - Combine population and area to create population_density.

2.5. Dimensionality Reduction

Overview of Dimensionality Reduction Techniques



Dimensionality Reduction

Manifold Learning

Model Based

Matrix Factorization

SOM

tSNE

LDA

PCA

SVD

3. Data Leakage

The **MOST IMPORTANT** Concept

- Misunderstanding this can make your model **completely useless** in practice.
- **Data Leakage:** Occurs when information from the **test set** is accidentally "leaked" into the model training process.

3. Process Comparison: Correct vs. Incorrect Process

The Golden Rule: SPLIT FIRST, PREPARE LATER

1. **Split:** Split raw data into train and test sets.
2. **Fit:** Learn the transform parameters **ONLY on the TRAIN set**.
3. **Transform:** Apply the transform to both the train and test sets.
4. Train the model.

1. Take the **ENTIRE** dataset.
2. **Prepare the data** (e.g., Scaling).
3. **Split** into train/test sets.
4. Train the model.

=> **PROBLEM:** Information (e.g., min, max) from the test set has "leaked" into the preparation step.

Summary

Key Takeaways

1. Data preparation is an **iterative process** and the foundation of a project.
2. Master the 5 main task groups: Cleaning, Selection, Transforms, Engineering, & Reduction.
3. **Always split data before preparation** to avoid data leakage.
4. Using a **Pipeline** is a best practice.

Q&A