# Data Mining

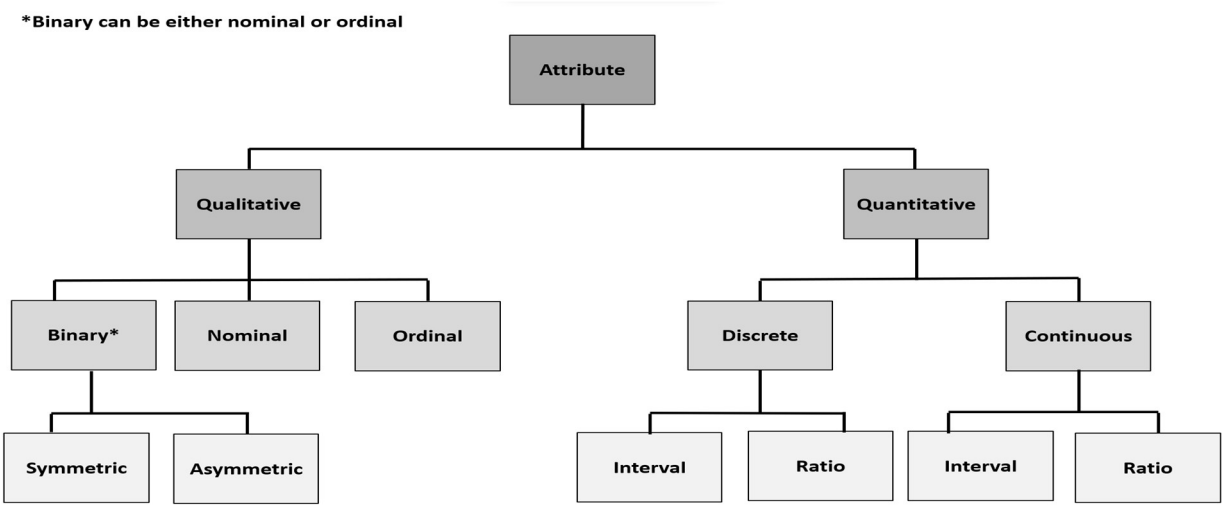
## Week 1: Introduction

- What is Data Mining (identify patterns, trends, and relationships from data)
- Motivation for Data Mining (extract new information and improve decision-making)
- Difference and Relationship between Data Mining, Machine learning, and AI
- Implementation process of Data Mining**:** business and data understanding, cleaning, modeling, …
- Data Mining techniques**:** classification, regression, clustering, association rules, anomaly detection, …
- Some applications of Data Mining in business and scientific research: finance, healthcare, biology, …

## Week 2: Data – part 1

### Data

- What is a Dataset (collection of Attributes and Objects)
- Types of Attributes**:** Nominal, Ordinal, Interval, Ratio, Discrete, Continuous, Asymmetric, …
- Which type of operation we can perform for each type of attribute**:** =, ≠, <, >, +, -, *, /
- Handle important characteristics of Data**:** dimensionality, sparsity, resolution, size, and distribution




### Data Preprocessing

- Data preprocessing tasks**:** data integration, data **cleaning**, data **transformation**, data reduction
- Data cleaning techniques**:** handle problems in data (noise, outliers, missing values, duplicates, …)
- Data transformation techniques**:** normalization and discretization (attributes), sampling (objects)

### Similarity metrics

- Similarity metrics between objects based on the type of attribute (nominal, ordinal, interval, …)
- Similarity for continuous vector of attributes**:** Euclidean and Minkowski distance, Cosine similarity
- Similarity for binary vector of attributes: Simple Matching Coefficient (SMC) and Jaccard Coefficient
- Linear correlation between two continuous attributes
- Properties of Similarity (identity, symmetry) and Distance (non-negativity, symmetry, triangle inequality)
- How to choose a proximity and similarity method