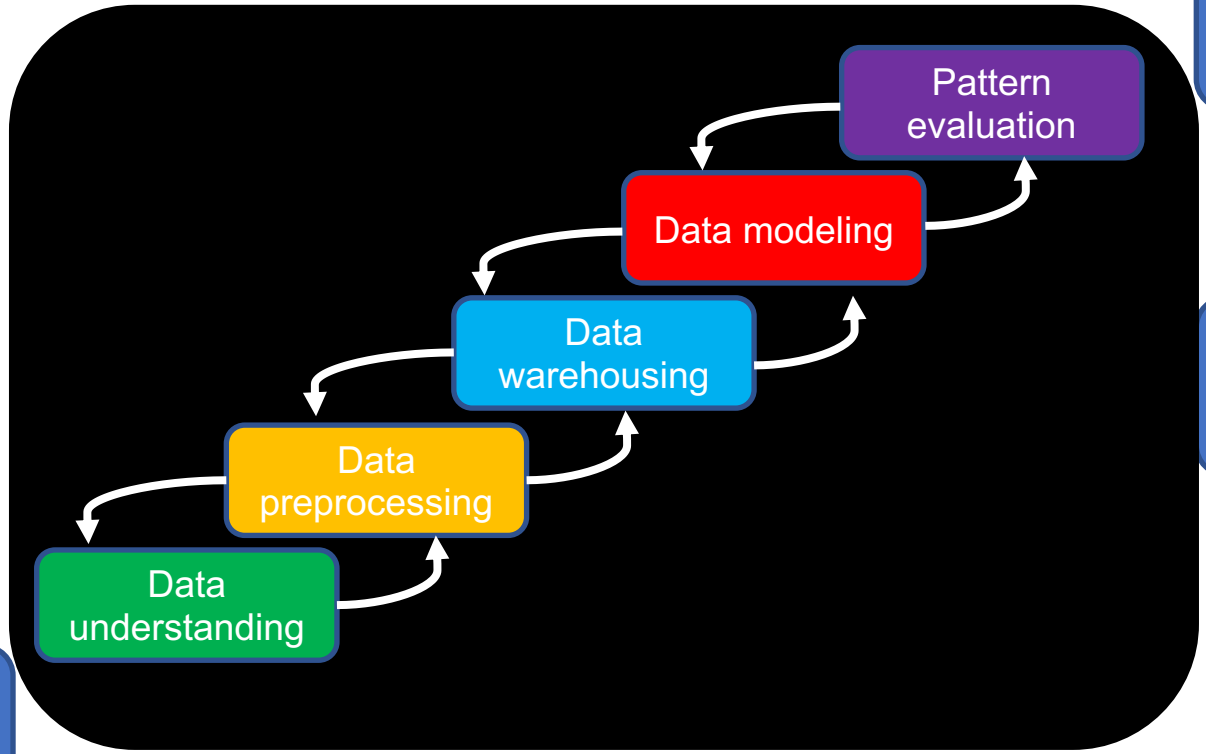
Data Preprocessing

Data Mining:
Data Mining Pipeline
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Identify potential issues in datasets. Apply techniques to preprocess data for data mining tasks.

Data Mining Pipeline



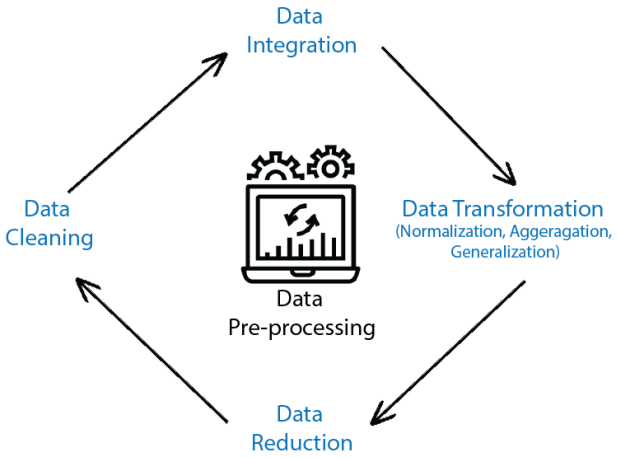
Application

Knowledge

Technique

Data

Data Preprocessing



- Potential issues with data
  - E.g., missing data, errors, inconsistency, availability
- Preparing data for the mining process
  - Data cleaning, integration, transformation, reduction
- No good data, no good data mining!

Data Transformation

- Smoothing: noise removal/reduction
- Aggregation: e.g., cities => state (n-to-1)
- Generalization: e.g., city => state (1-to-1)
- Normalization: feature scaling
- Discretization: continuous => intervals
- Attribute construction from existing ones

Normalization (1)

➢ Rescaling (Min-max normalization)

- Map attribute values from [min, max] to [min', max']
- E.g., income range [50K, 200K], map to [0, 1.0]
- v = 100K
- v' = (100K-50K)/(200K – 50K) * (1.0 – 0) + 0 = 0.33

$$v' = \frac{v - min}{max - min}(max' - min') + min'$$

Normalization (2)

➢ Mean normalization

- E.g., income range [50K, 200K], mean = 120K
- v = 100K
- v’ = (100K – 120K) / (200K – 50K) = -0.13

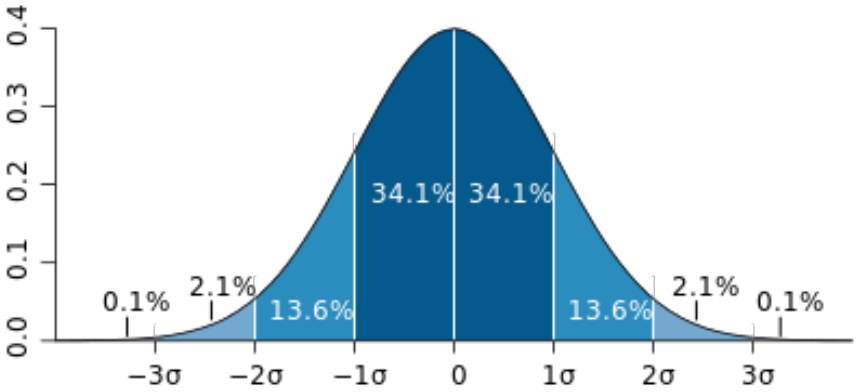
$$v' = \frac{v - mean}{max - min}$$

Normalization (3)

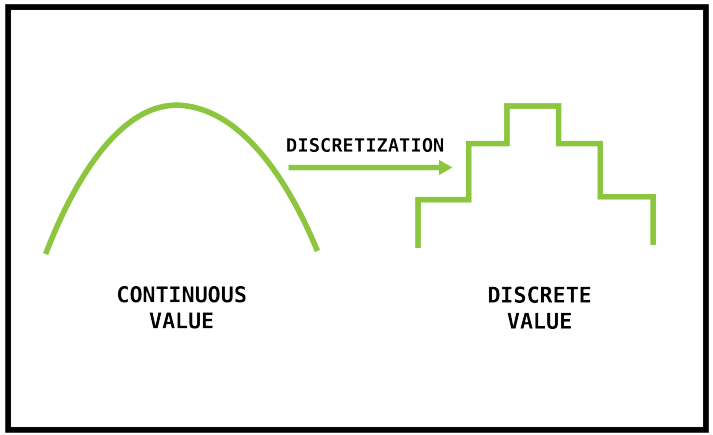
➢ Standardization (z-score normalization)

- E.g., income mean = 120K, stdev = 30K, v = 100K
- v’ = (100K – 120K) / 30K
- = -0.67

$$v' = \frac{v - mean}{stdev}$$



Discretization



- Continuous => intervals
  - E.g., income => 10K increments
- Split or merge
  - Recursive splitting or merging
- Supervised or unsupervised: class labels

Unsupervised Discretization

- Binning and histogram analysis
  - Equal width, equal frequency
  - e.g., grading: 10, 20, … of 100; 10%, 20%, … of class
- Clustering analysis
- Intuitive partitioning

Supervised Discretization

- ➢ Pre-determined class labels
- ➢ Entropy-based interval splitting
  - lower entropy means “purer” class distribution
- ➢ X^2 analysis-based interval merging
  - lower X^2 value means class is independent of interval

Data Reduction

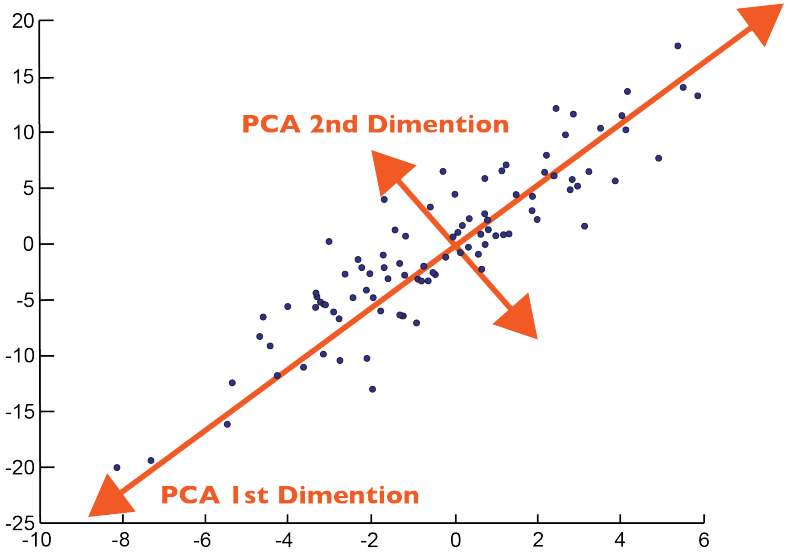
- Large data takes a long time to mine
  - Reduce data to make it more efficient
- Should still find similar patterns
- Dimensionality reduction: #attributes
- Numerosity reduction: #objects

Attribute Selection

- ➢ Forward selection
  - Keep adding (most informative) attributes
- ➢ Backward elimination
  - Keep removing (least informative) attributes
- ➢ Feature engineering
  - Domain knowledge, decision tree induction, …

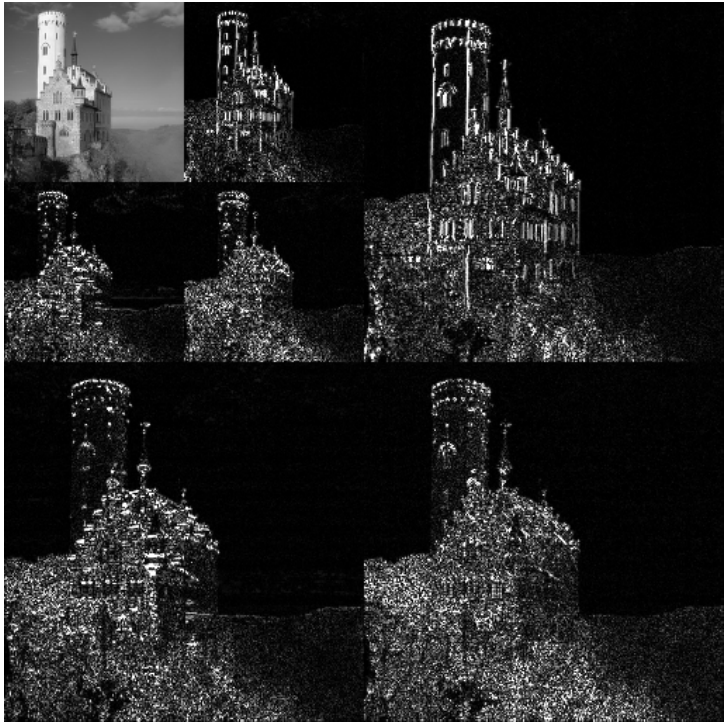
Principle Component Analysis (PCA)

- n-dimensional data
- => use first few orthogonal vectors (principle components)

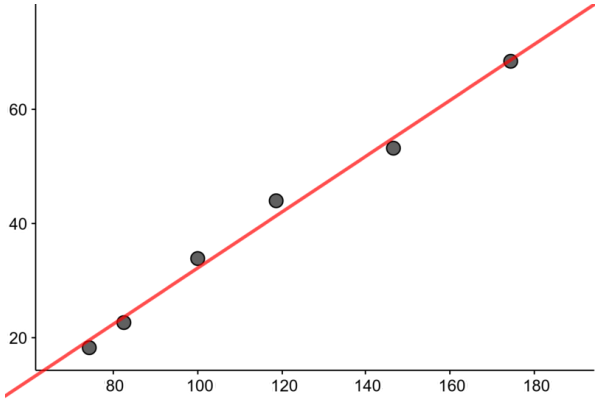


Wavelet Transformation

- Linear signal processing, multi-resolution
- Store a small fraction of the strongest wavelet coefficients



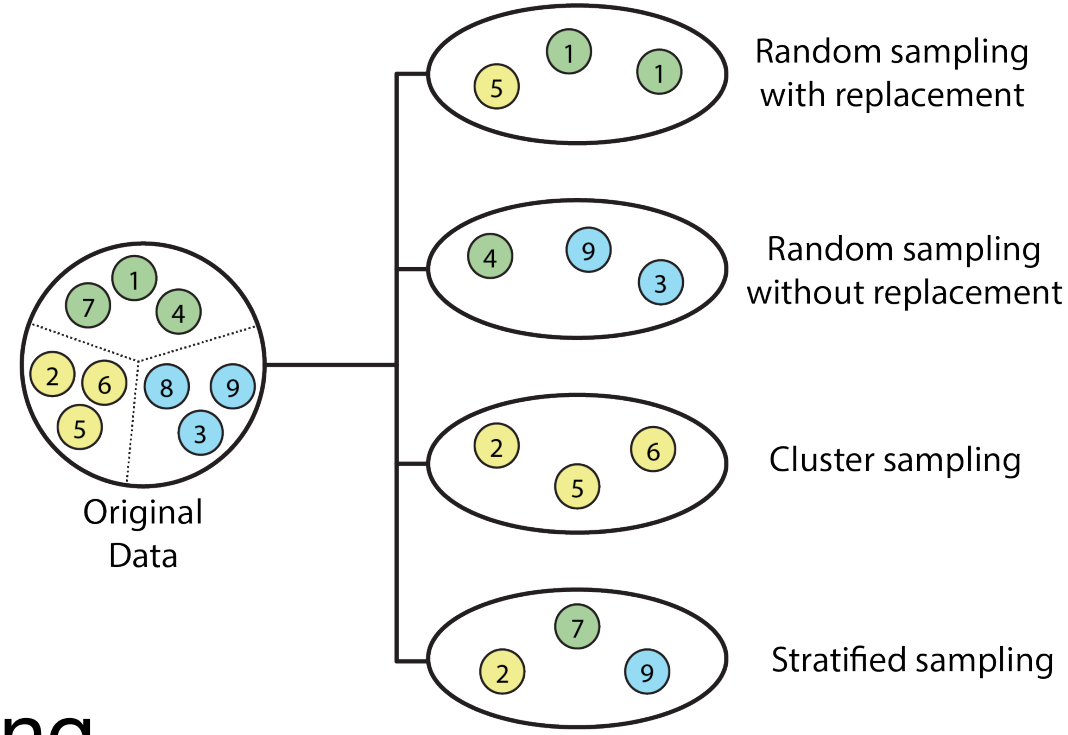
Numerosity Reduction



- ➢ Parametric methods
  - Assume the data fits a certain model
  - Estimate model parameters
  - E.g., linear/multi-linear/log-linear regression
- ➢ Non-parametric methods
  - Do not assume a certain model
  - Use fewer/smaller data representations

Sampling

- Select a representative subset of data points
- Different sampling methods for different scenarios

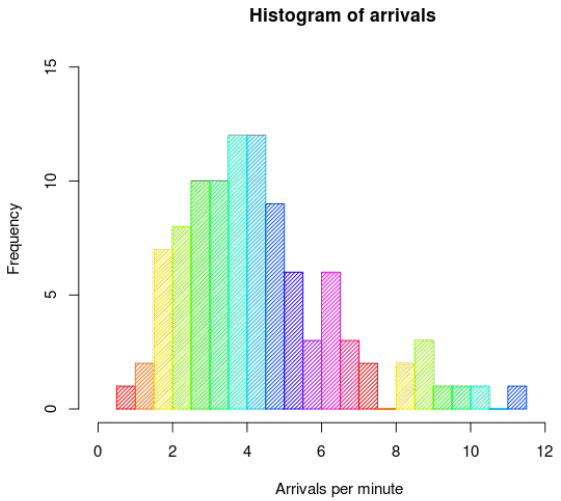


➢ Aggregation

- E.g., daily sales => monthly sales
- E.g., vehicles by city => by state

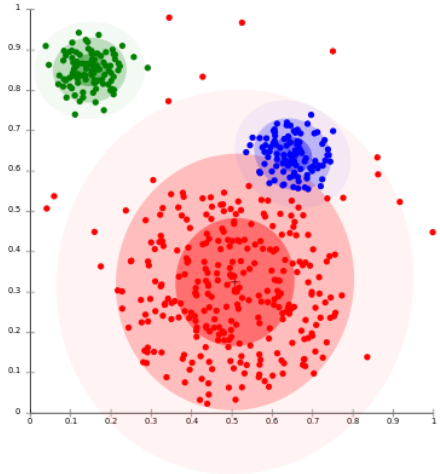
➢ Histogram

- Store bucket intervals & frequencies

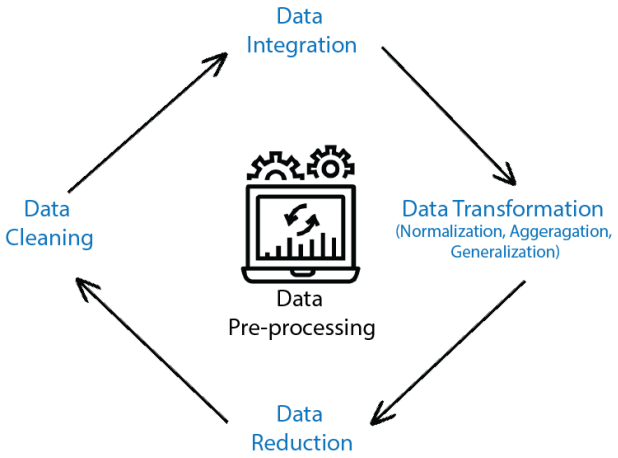


➢ Clustering

- Store cluster representations
- E.g., centroid and radius



Data Preprocessing



- Potential issues with data
  - E.g., missing data, errors, inconsistency, availability
- Preparing data for the mining process
  - Data cleaning, integration, transformation, reduction
- No good data, no good data mining!