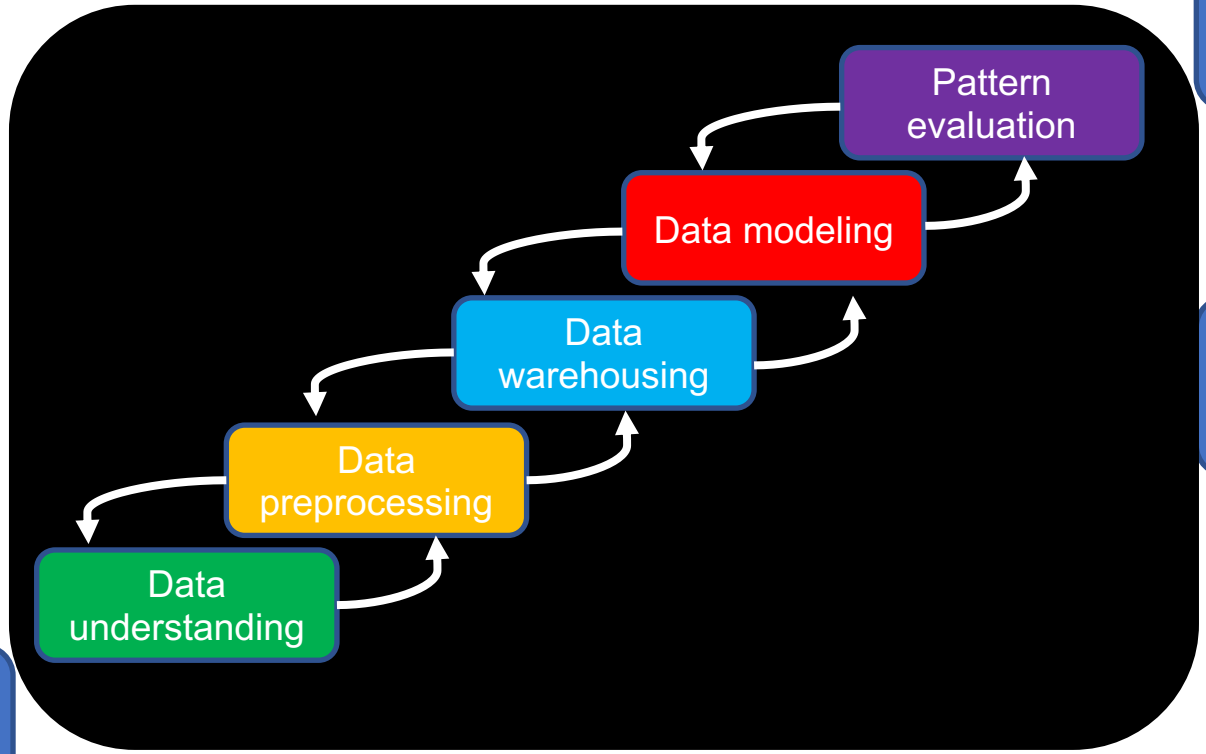
Data Preprocessing

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



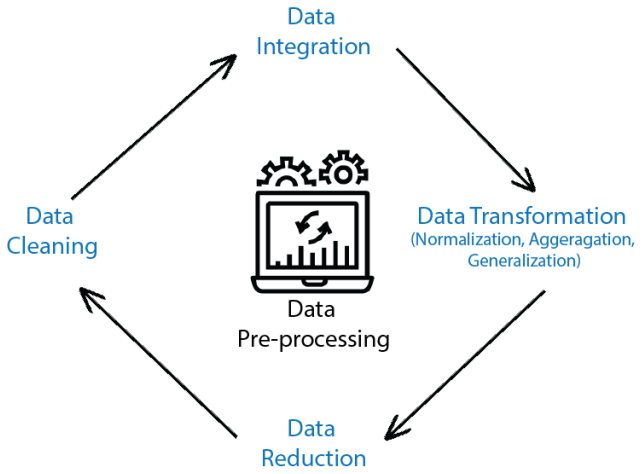
Application

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 Quality

- Relevance
- Accessibility
- Interpretability
- Reliability
- Timeliness
- Accuracy
- Consistency
- Precision
- Granularity
- Completeness

Issues in Real-world Data

➢ Incomplete

- Missing values, missing attributes

➢ Noisy

- Imprecision, errors, outliers: e.g., age = “-10”

➢ Inconsistent

- E.g., age vs. birthday, rating scale

Causes of Data Issues

- ➢ Data collection/transmission/processing
- ➢ Human, hardware, and software
  - Limitations, errors, multiple sources
- ➢ Changes over time
  - Updated survey, new sensing capabilities

Data Cleaning

- **Incomplete**: Remove or fill in missing values
- **Noisy**: Smooth noisy data, identify/remove errors/outliers
- **Inconsistent**: Resolve inconsistencies

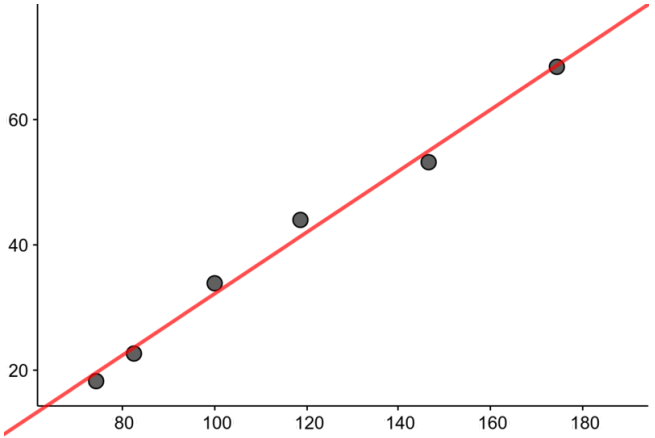
Incomplete Data

- ➢ Remove objects/attributes
- ➢ Manually fill in missing values
- ➢ Automated methods
  - Global constant, attribute mean, class mean
  - Estimated value: regression, kNN, probabilistic, …

Noisy Data

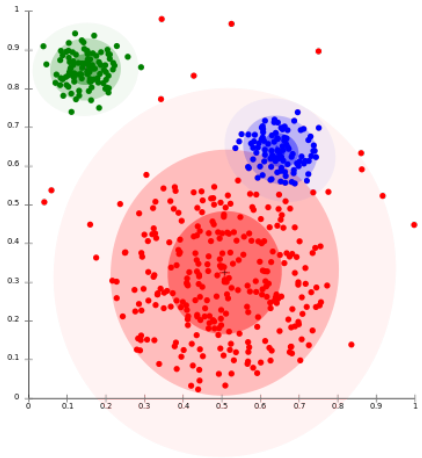
➢ Regression

- Fit data with regression functions
- E.g., linear, polynomial, logistic, …



➢ Clustering

- Group data into clusters
- Detect and remove outliers

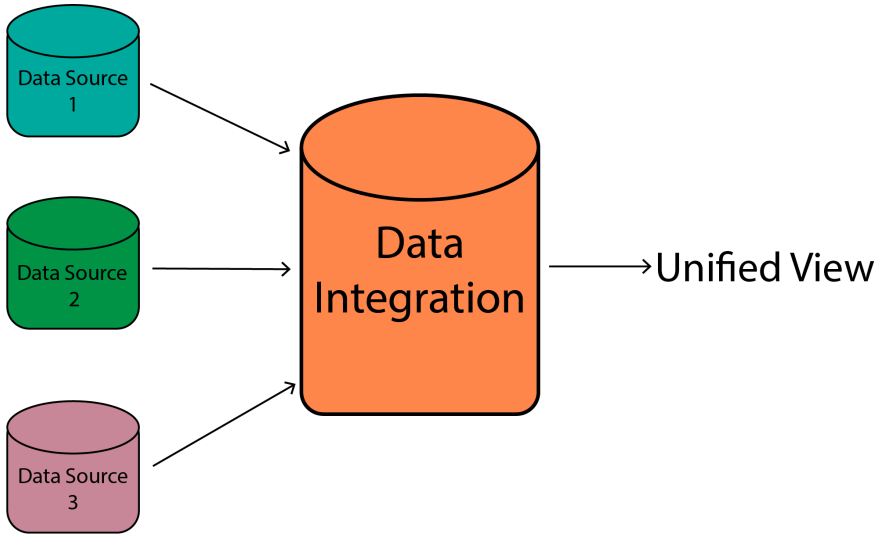


Inconsistent Data

- ➢ Semantic-based checking
  - Metadata, attribute relationships
  - E.g., age vs. DOB
- ➢ Data understanding
  - Statistical analysis, visualization
  - E.g., scatter plot

Data Integration

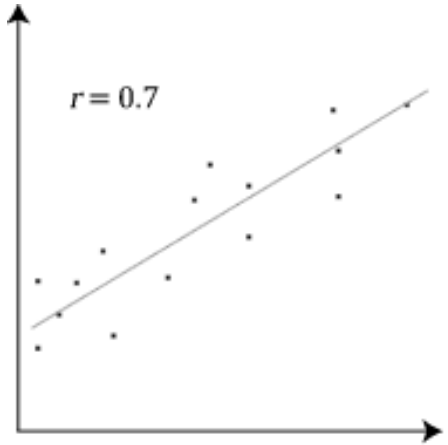
- Combines data from multiple sources
- Entity identification
  - E.g., users, items
- Redundant data
  - E.g., correlation analysis



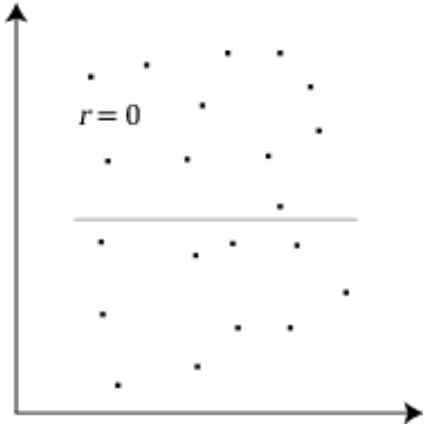
Correlation Analysis (1)

- Numerical attributes: correlation coefficient

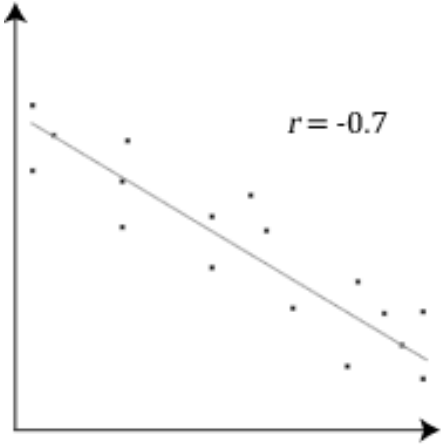
$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$



Positive Correlation



No Correlation



Negative Correlation

Correlation Analysis (2)

- Nominal attributes: chi-square test

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

	B=b1	B=b2
A=a1	O_a1b1	O_a1b2
A=a2	O_a2b1	O_a2b2

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

Correlation Analysis (3)

➢ Does correlation imply causality?

- sleeping with one's shoes on is strongly correlated with waking up with a headache
- the more fireman fighting a damage, the more damage there is going to be
- as ice cream sales increases, the rate of drowning deaths increases sharply
- Correlation does NOT imply causality!