Introduction to Data Mining Pipeline

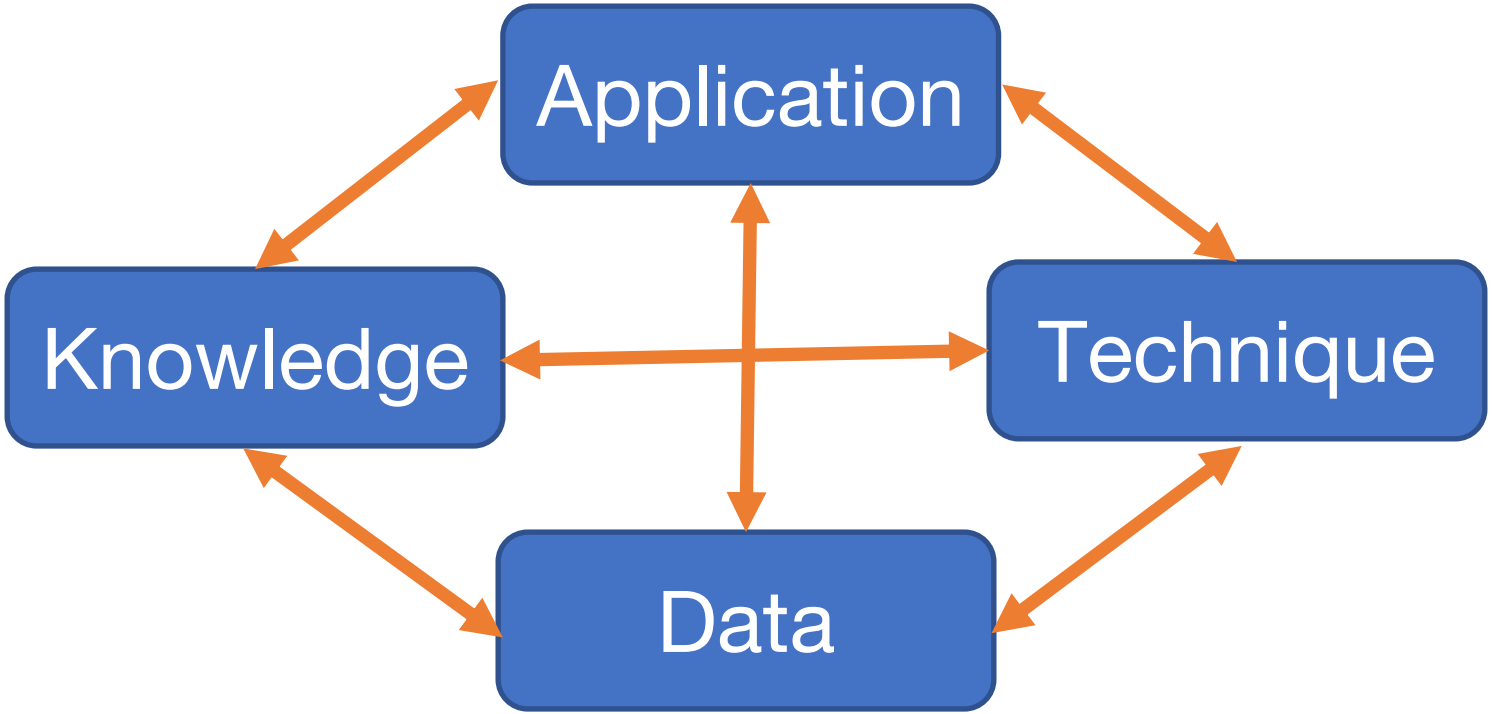
Data Mining: Data Mining Pipeline

with **Dr. Qin Lv**

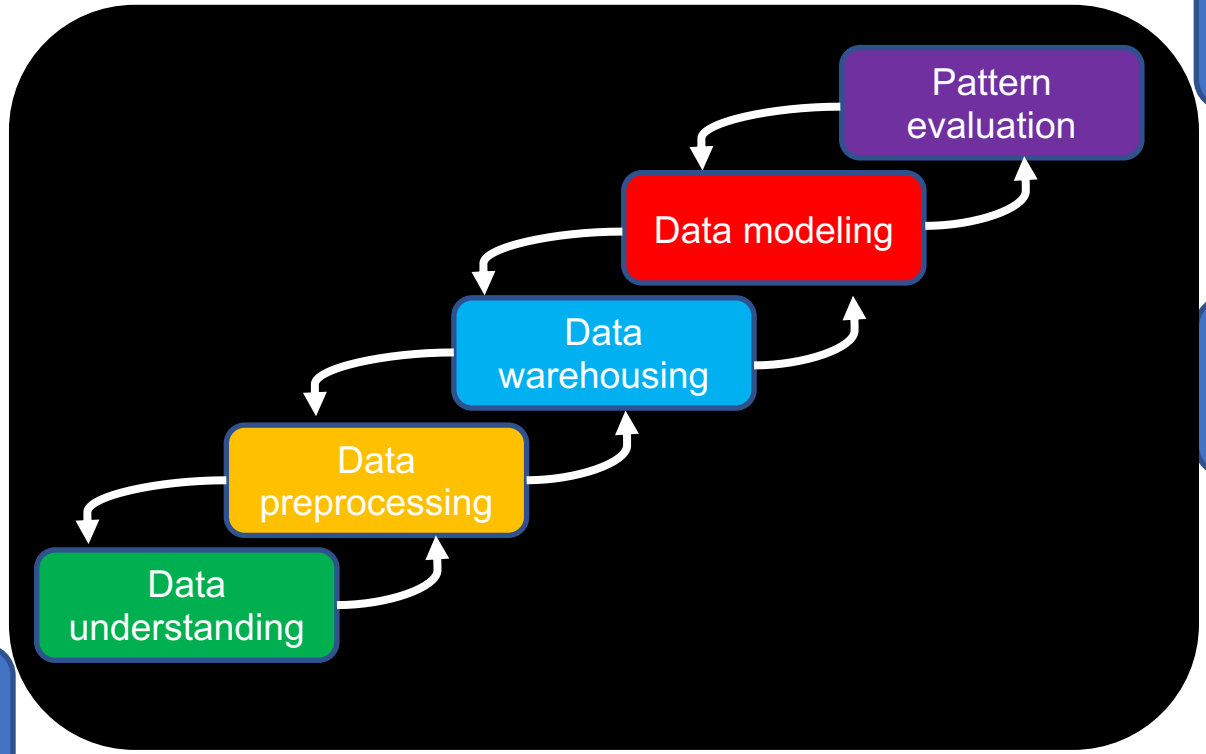
Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Identify the key components of the data mining pipeline. Describe how the components of the data mining pipeline are related.

Data Mining: Four Views



Data Mining Pipeline



Application

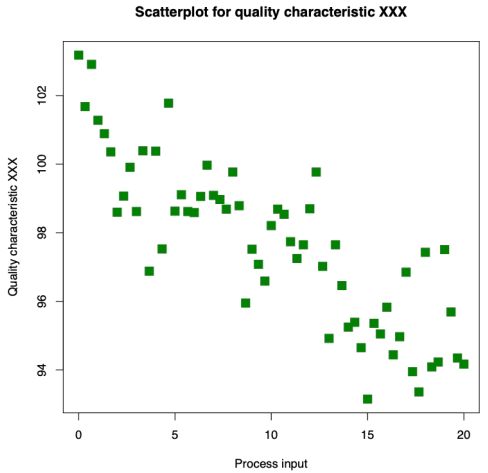
Knowledge

Technique

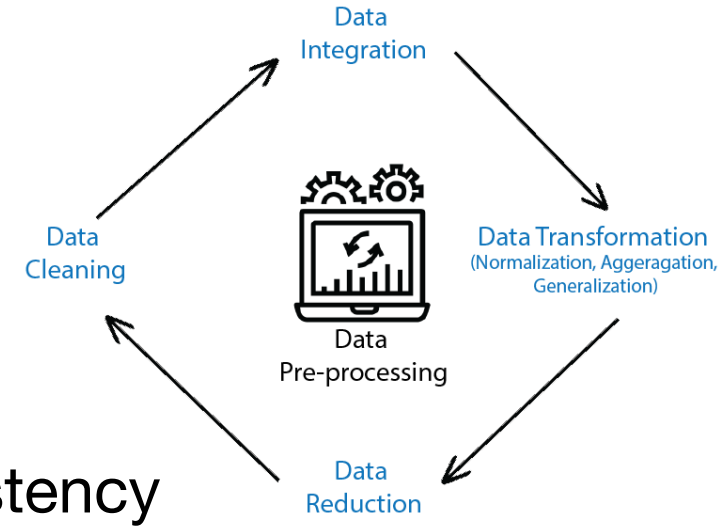
Data

Data Understanding

- What types of data?
- What do they look like?
- Statistics & visualization
- Similarity vs. dissimilarity
- General patterns vs. anomalies



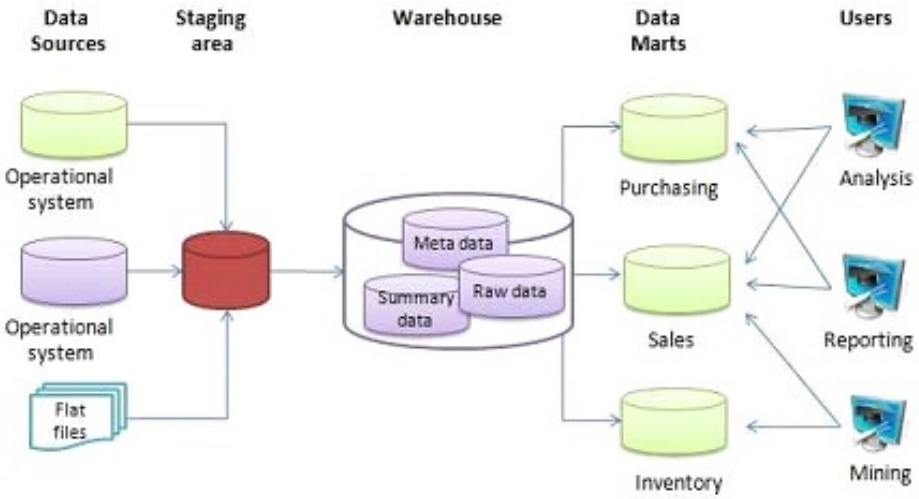
Data Preprocessing



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

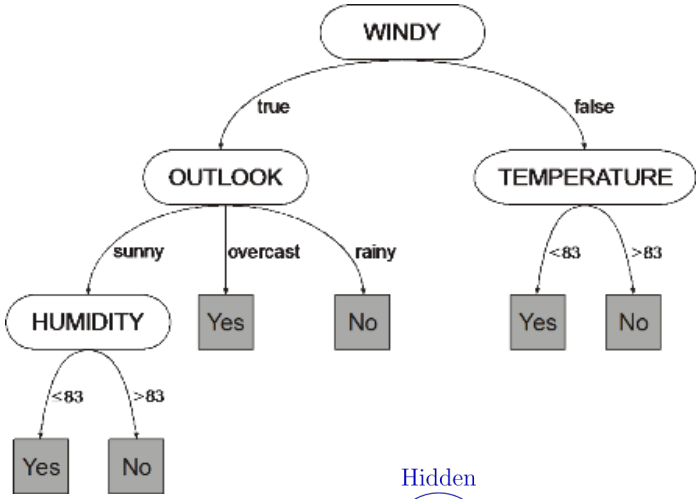
Data Warehousing

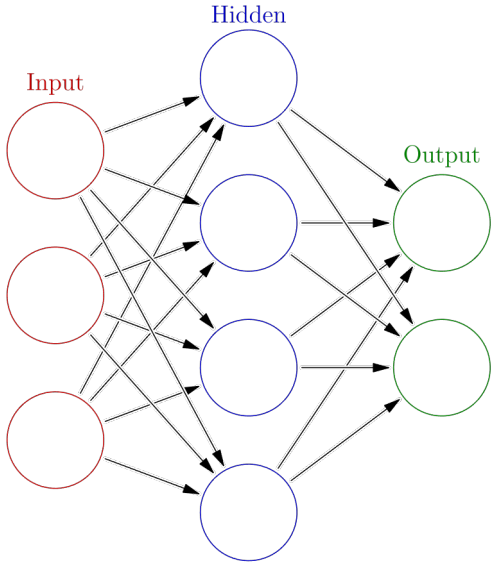
- ➢ Data warehouse
  - vs. operational data
- ➢ Data cube & OLAP
  - Multi-dimensional data management
- ➢ Data warehouse architecture



Data Modeling

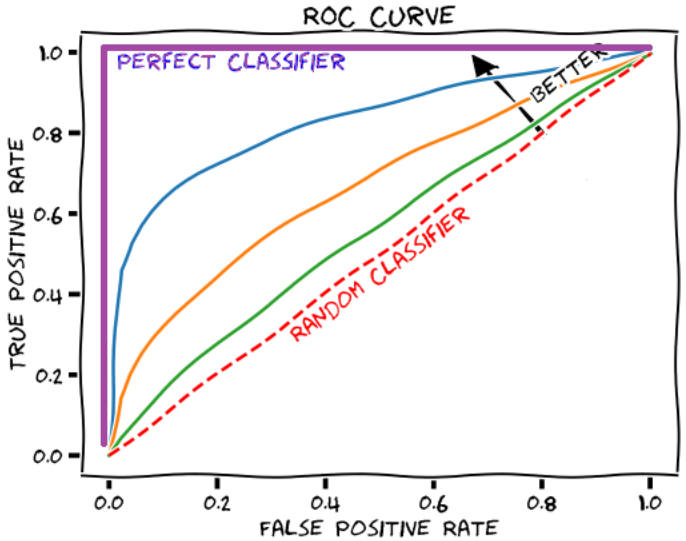
- Frequent pattern analysis
- Classification, prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis





Pattern Evaluation

- ➢ Finding interesting patterns from data
  - New, valid, generalizable, useful, explainable
- ➢ Evaluation metrics
  - Accuracy, error rate
  - False positive/negative rate
  - Efficiency, latency, …
- ➢ Model selection



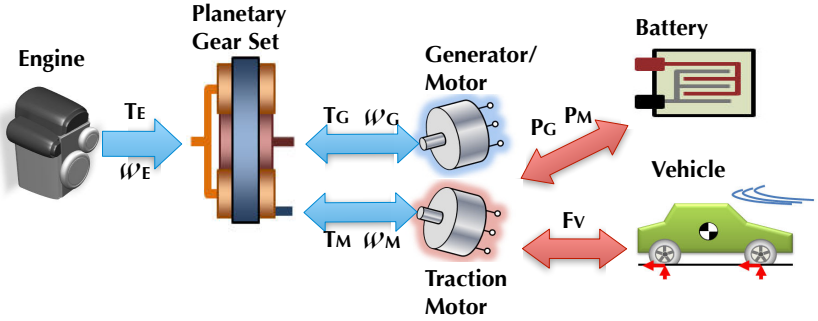
Data Mining In the Real World

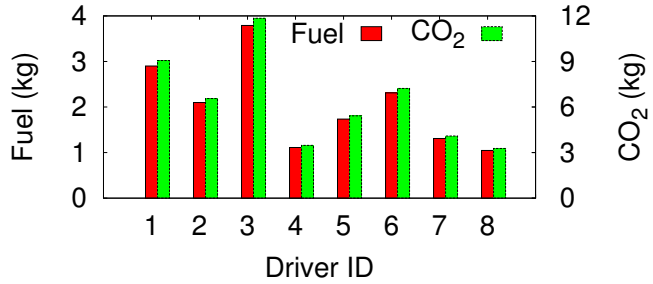
- Integrated views
  - Data, application, knowledge, technique
- Data mining pipeline
  - Data understanding, preprocessing, warehousing, modeling, evaluation
- Analytical reasoning!

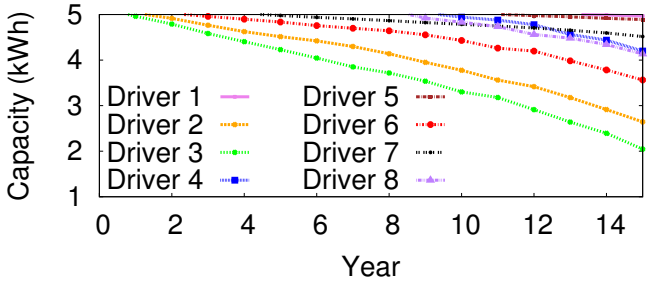
Data Mining Examples

- Business intelligence
  - E.g., customers, products, logistics, promotion, fraud
- Cyberspace
  - E.g., service providers, online social media, security
- Pick examples of your interest

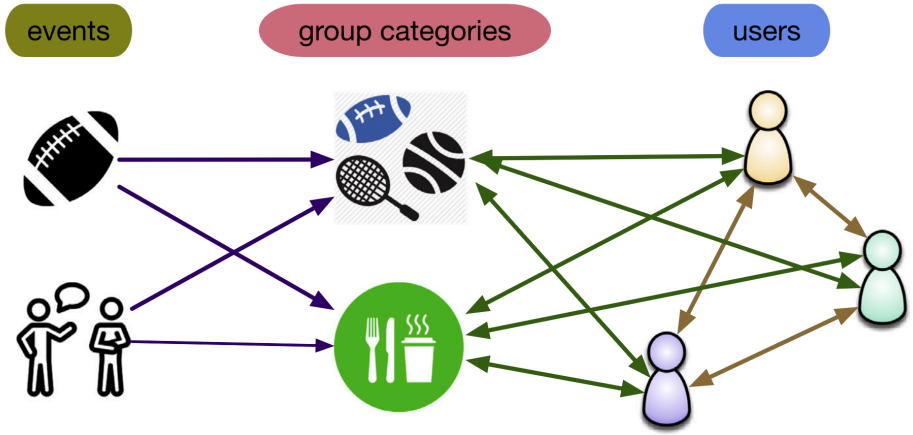
Transportation Electrification

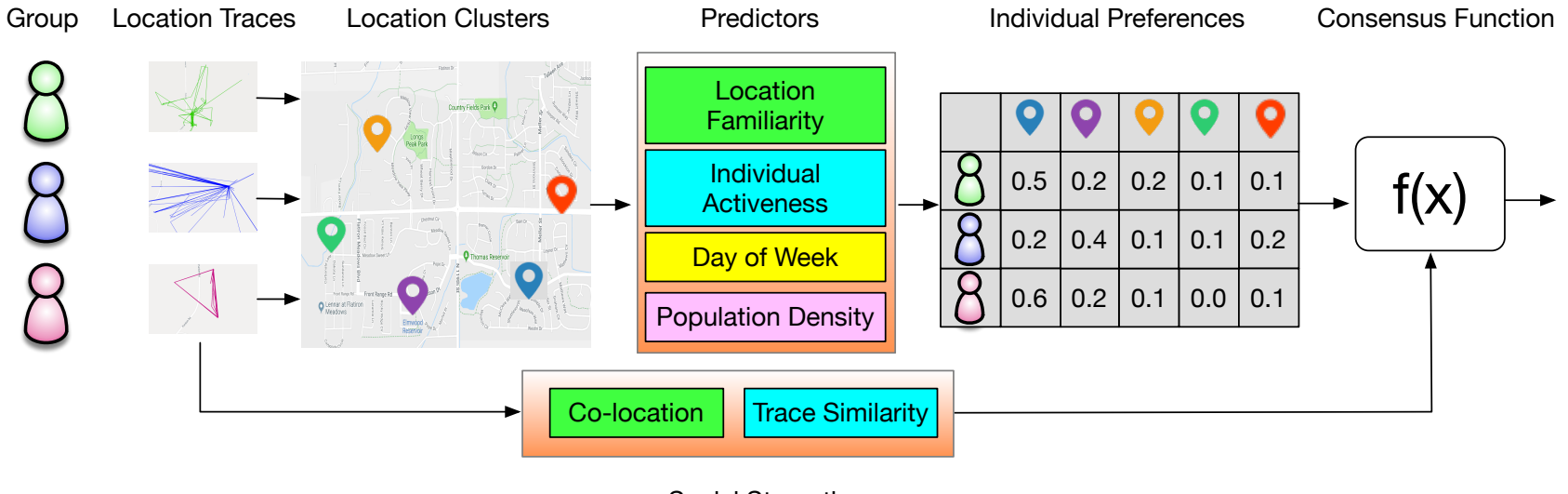




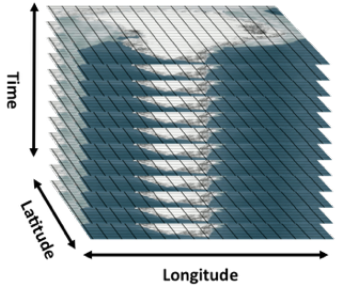


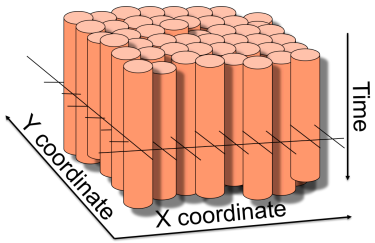
Group Event Scheduling

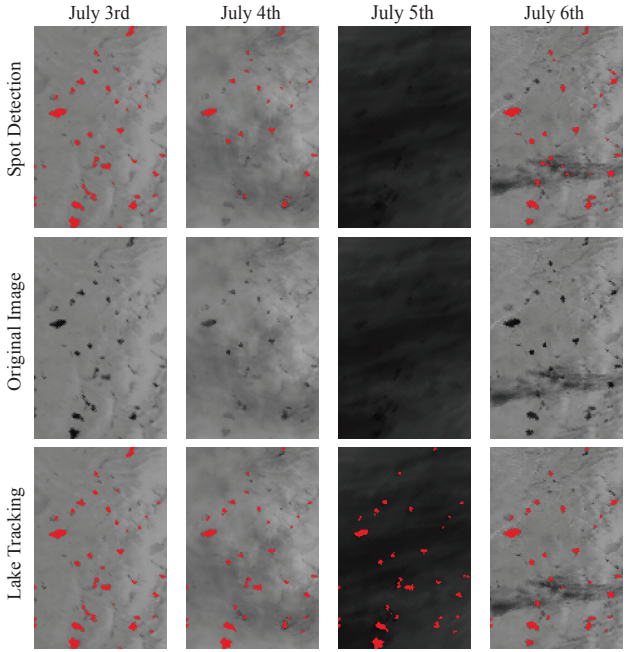


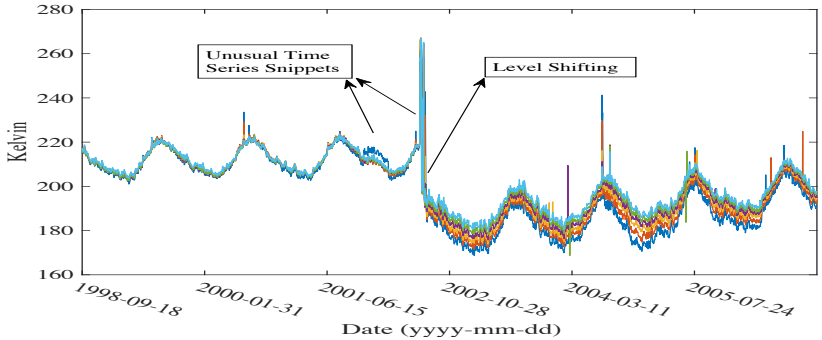


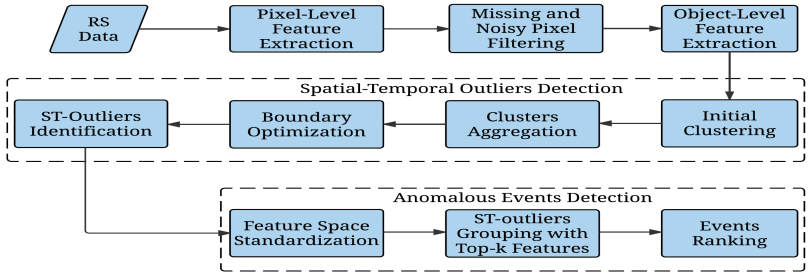
Remote Sensing Data Analysis











Major Issues in Data Mining (1)

- Diverse data => diverse knowledge
- Data quality issues
- Supervised vs. unsupervised learning
- Performance evaluation
- Effectiveness vs. efficiency

Major Issues in Data Mining (2)

- Incremental, interactive mining
- Integration of domain knowledge
- Visual analytics
- Privacy-preserving mining
- …

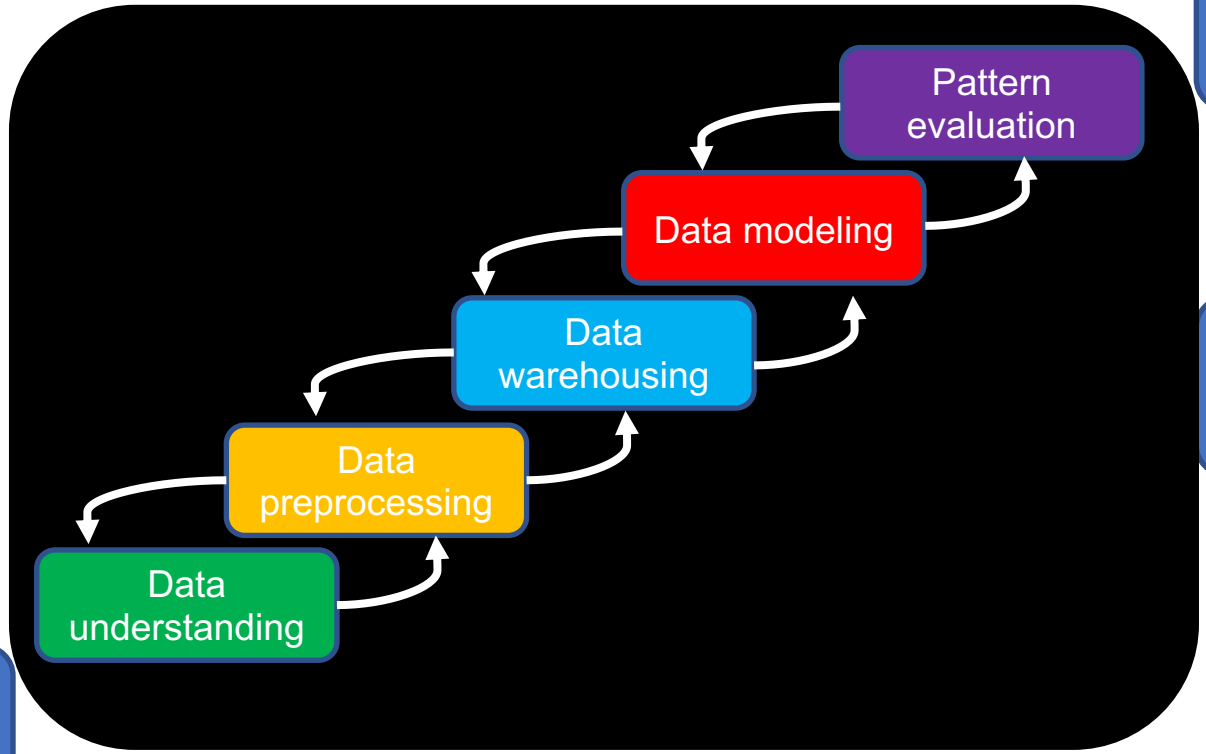
Data Ethics

- Data ownership
- Privacy, anonymity
- Data and model validity
- Data and model bias (algorithmic fairness)
- Interpretation, application, societal consequence

Data Mining Resources

- ACM SIGKDD: https://www.kdd.org/
  - Annual KDD conference: keynotes, research, applied data science, workshops, tutorials, KDD CUP, …
- Other conferences & journals
- Online resources

Data Mining Pipeline



Application

Knowledge

Technique

Data