Data Mining

Module 1

Introduction to Data Mining

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

I

Data Management: Its Evolution

- 1960s:
  - File management and network DBMS
- 1970s:
  - Relational DBMS
- 1980s:
  - Non-first normal form, extended-relational, OO, deductive databases and application-oriented DBMS (spatial, scientific, CAD/CAM, etc.)
- 1990s - present:
  - Data mining, digital library, and Web databases
  - Cloud databases, data science, and Big Data

Data Mining: Its Definition

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names:
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, etc.

Data Mining: A Multidisciplinary Field

- Pattern Recognition
- Machine Learning
- Databases
- Statistics
- Information Visualization

Data to be mined

- Text databases
- Web databases
- Scientific and biological databases
- Transactional databases

Knowledge to be discovered

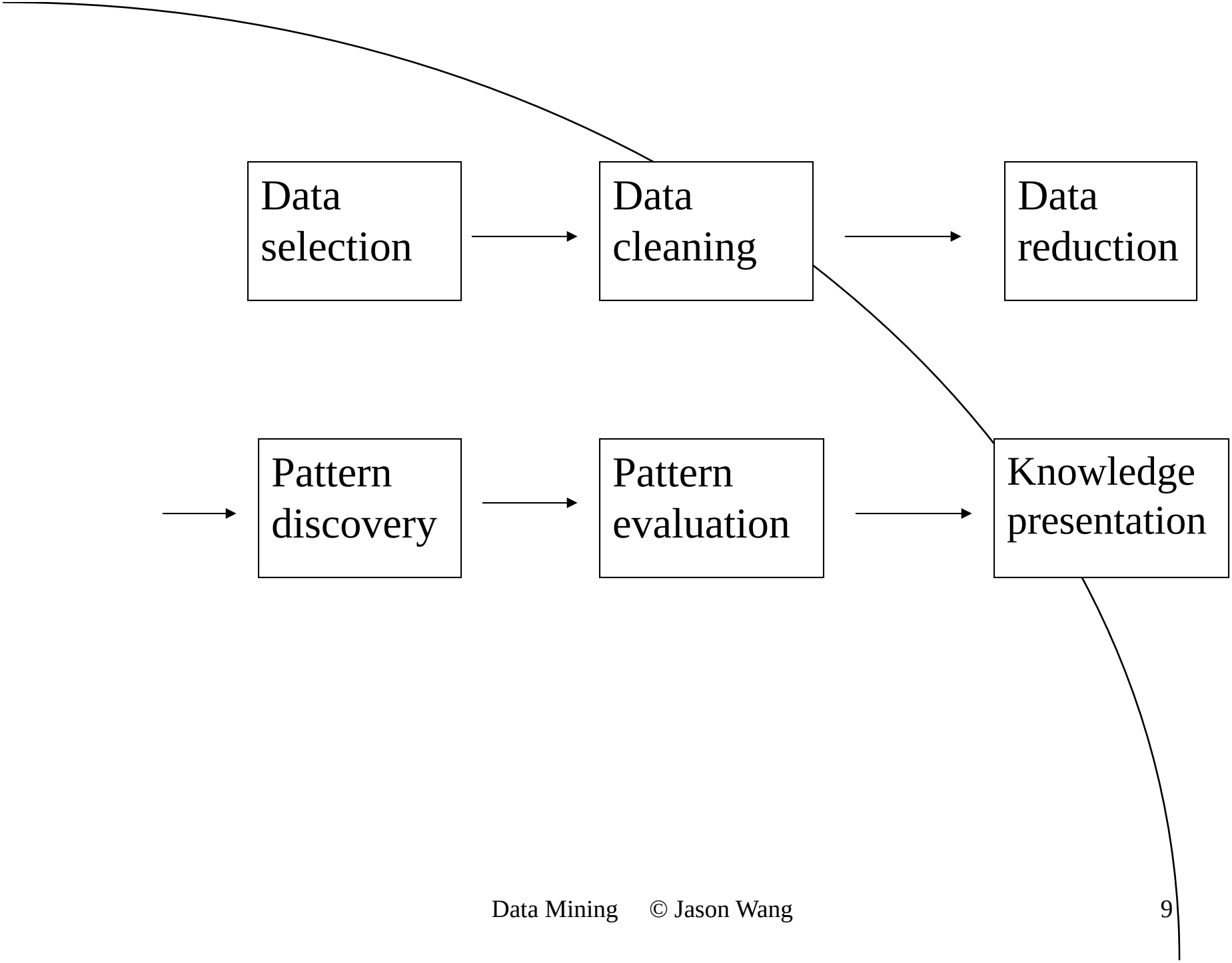
- Association (correlation)
  - Multi-dimensional vs. single-dimensional association
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(X, “computer”) → contains(X, “software”) [1%, 75%]

Knowledge to be discovered (cont.)

- Classification
  - Finding models (functions) that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on climate, or classify cars based on gas mileage
- Clustering
  - Class label is unknown: Group data to form clusters

Interesting patterns

- **Many patterns can be discovered.**
- **Interestingness measures**: A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm.
- **Objective vs. subjective interestingness measures:**
  - Objective: based on support, confidence, etc.
  - Subjective: based on user's judgement


Data selection
Data cleaning
Data reduction
Pattern discovery
Pattern evaluation
Knowledge presentation

Data Mining: Its Applications

- Market analysis and management
- Risk analysis and forecasting
- Fraud detection and management
- Text mining (news group, email, documents) and Web analysis
- Bioinformatics data mining
- Security informatics data mining

Market Analysis and Management

- Credit card transactions - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- Data mining can tell you what types of customers buy what products (clustering or classification).
- Data mining can identify the best products for different customers.

Risk Analysis and Management

- Summarize and compare the resources and spending
- Monitor competitors and market directions
- Set pricing strategy in a highly competitive market

Fraud Detection and Management

- Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Example: detect a group of people who stage accidents to collect on auto insurance

Other Applications

- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining techniques.
- Bioinformatics
  - Many bioinformatics companies use data mining techniques to find genes in DNA and to classify protein sequences.

Security informatics

-- RPI researchers use data mining to find hidden groups on the Internet.

A Summary of Data Mining

Databases to be mined

Knowledge to be discovered

Techniques to be utilized

Domains to be applied

End of Module 1