Chapter 1, Introduction

Young-Rae Cho
Associate Professor
Department of Computer Science
Baylor University

What is Data Mining?

- **Definition**
  - Knowledge discovery from data (KDD)
  - Informative pattern extraction from data
  - Data analysis using specific algorithms or machine learning techniques

- **General View of Data Mining Process**
  - Input Data
  - Pre-Process: Cleaning, Normalization, Integration
  - Data Mining: Pattern Mining, Classification, Clustering
  - Post-Process: Evaluation, Interpretation, Visualization
  - Output Knowledge
Data Mining in Computer Science

- Alternative View of Data Mining Process

Data Mining in Business Intelligence

Increasing potential to support business decisions

CEO
Business Analyst
Data Analyst
DBA

Data Sources

Decision Making

Data Presentation
- Interpretation, Visualization

Data Mining
- Knowledge (Pattern) Discovery

Data Exploration
- Selection, Summarization, Transformation

Data Preprocessing & Integration
- Cleaning, Normalization, Warehousing
Why Need Data Mining?

- **Business**
  - Business data analysis for decision support
    - Market analysis and management
    - Risk analysis and management
    - Fraud detection and security

- **Science and Engineering**
  - Biomedical data analysis
    - Patient treatment, disease diagnosis, and drug discovery
  - WWW data analysis
    - Information retrieval and web management
  - Geographic data analysis
    - City planning and renewal

Why Not Traditional Data Analysis?

- **Explosive Growth of Data**
  - Terabytes or petabytes of data

- **High Dimensionality of Data**
  - Hundreds or thousands of dimensions

- **High Complexity of Data**
  - Stream data, sensor data
  - Time-series data, temporal data
  - Spatial data, spatio-temporal data, multimedia data
  - Structural data, graphic data
  - Combined, heterogeneous data format

Data mining algorithms should handle these data!!
Data Mining Functions: (1) Generalization

- **Data Cleaning and Reduction**
  - Statistical normalization methods
  - Sampling and discretizing techniques

- **Data Integration and Warehousing**
  - Multidimensional data modeling
  - Dimension reduction techniques
  - Data cube aggregation algorithms

- **Data Transformation**
  - OLAP (online analytical process) operations
  - Querying for selection and summarization

Data Mining Functions: (2) Pattern Mining

- **Frequent Pattern Mining**
  - Mining frequently occurred item-sets
  - Mining frequently occurred sequential patterns
  - Mining frequently occurred structural patterns (sub-graphs)

- **Association Rule Mining**
  - Mining one-direction relations between two sets of data

- **Correlation Mining**
  - Mining two-direction relations between two sets of data

- **Coherent Pattern Mining**
  - Mining coherent sequential patterns
  - Mining coherent structural patterns
Data Mining Functions: (3) Classification

- **Supervised Learning**
  - Training data with class labels
  - Prediction of classes of data with no class labels

- **Typical Methods**
  - Decision tree-based induction
  - Naive bayesian classification
  - K-nearest neighbors (kNN)
  - Support vector machine (SVM)
  - Neural network
  - Logistic regression
  - Rule-based classification
  - Pattern-based classification

Data Mining Functions: (4) Clustering

- **Unsupervised Learning**
  - Grouping data with no class labels
  - Prediction of potential members with same class labels

- **Typical Methods**
  - K-means
  - Agglomerative hierarchical clustering
  - Divisive hierarchical clustering
  - Density-based clustering
  - Grid-based clustering
  - Pattern-based clustering
  - Outlier analysis
Origins of Data Mining

- Inter-disciplinary Research Field

Data Mining

- Algorithms
- Machine Learning
- Statistics
- Database Management
- Pattern Recognition
- Application
- Visualization
- High-Performance Computing

Conferences and Journals on Data Mining

**Conferences**
- ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)
- ACM Conf. on Information and Knowledge Management (CIKM)
- SIAM Int. Conf. on Data Mining (SDM)
- European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD)
- Pacific-Asian Conf. on Knowledge Discovery and Data Mining (PAKDD)

**Journals**
- Data Mining and Knowledge Discovery (DMKD) by Springer
- IEEE Transactions on Knowledge and Data Engineering (TKDE)
- ACM Transactions on Knowledge Discovery from Data (TKDD)
Questions?

- Lecture Slides on the Course Website,
  "www.ecs.baylor.edu/faculty/cho/4352"