Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 1 —

By: Jiawei Han, Micheline Kamber, and Jian Pei

Lecturer: Mohammed L. Mahmood

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined (Data Mining Functions)?

- What Technology Are Used?

- What Kind of Applications Are Targeted?

- Major Issues in Data Mining

- A Brief History of Data Mining and Data Mining Society

- Summary

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
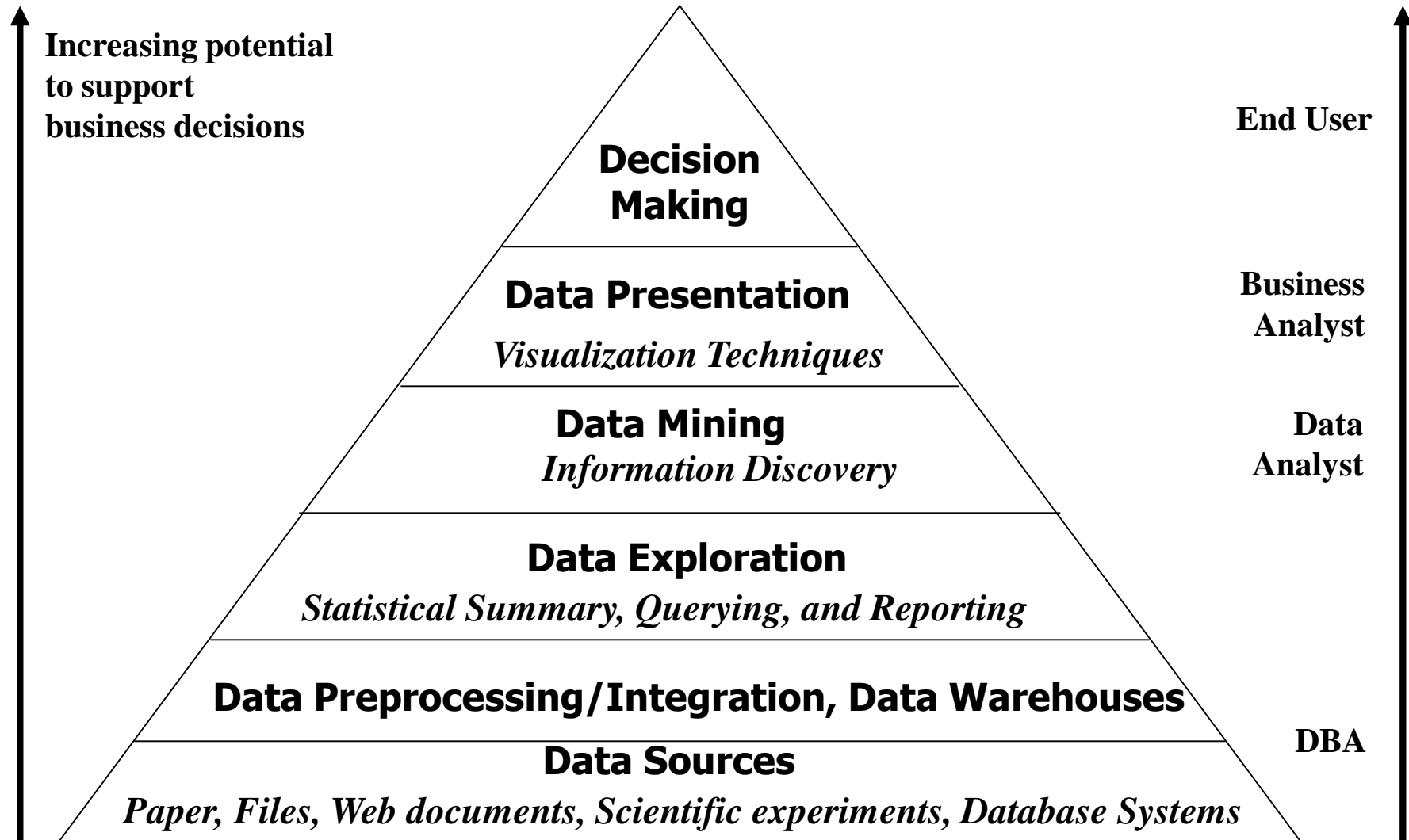 - Web technology (XML, data integration) and global information systems

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (<u>non-trivial</u>, <u>implicit</u>, <u>previously</u> <u>unknown</u> and <u>potentially useful)</u> patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
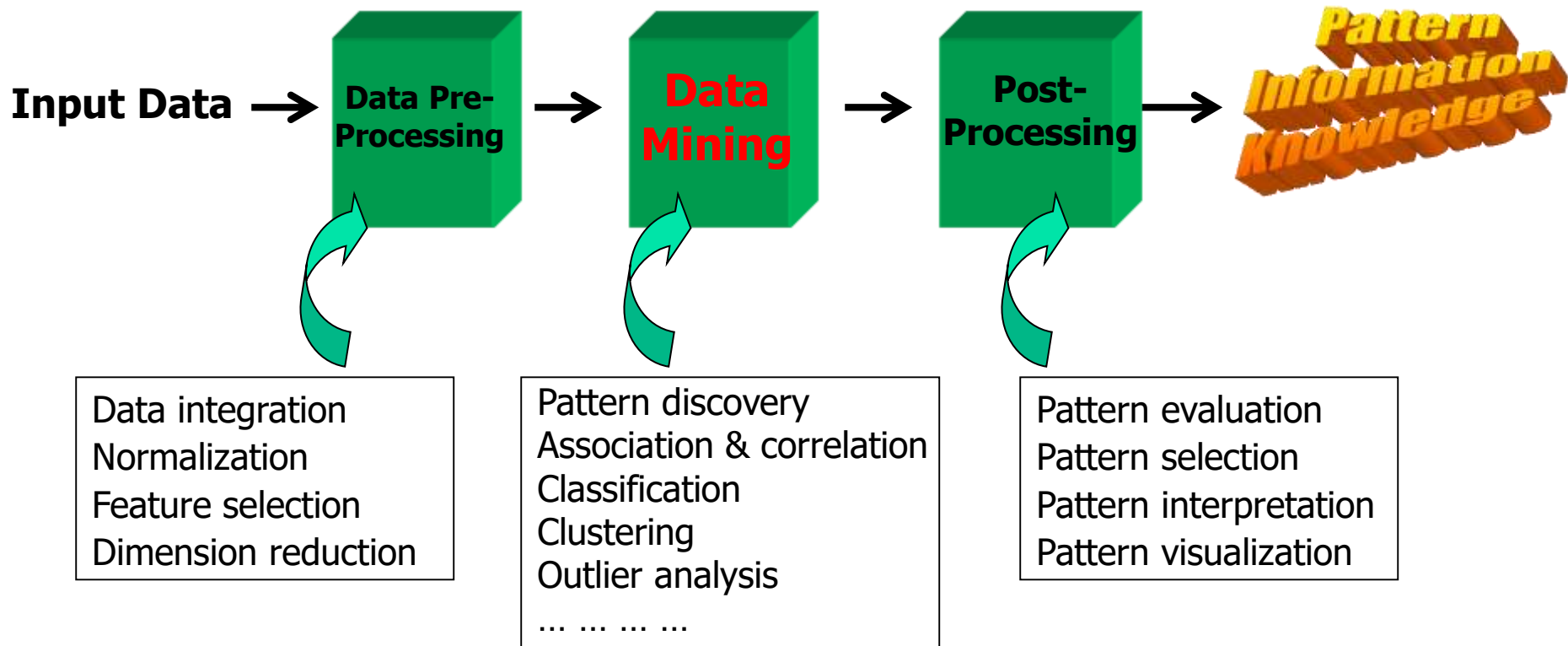 - (Deductive) expert systems
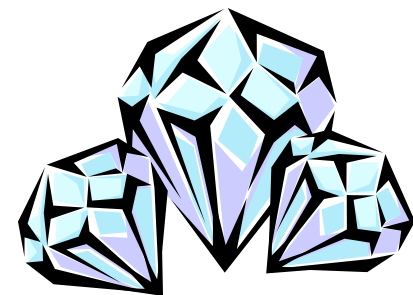
Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence

Increasing potential to support business decisions

End User

Decision Making

Business Analyst

Data Presentation
Visualization Techniques

Data Analyst

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

DBA

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

KDD Process: A Typical View from ML and Statistics

Input Data → **Data Pre-Processing** → **Data Mining** → **Post-Processing** → **Pattern Information Knowledge**

Data integration	Pattern discovery	Pattern evaluation
Normalization	Association & correlation	Pattern selection
Feature selection	Classification	Pattern interpretation
Dimension reduction	Clustering	Pattern visualization
	Outlier analysis	
	… … … …	

- This is a view from typical machine learning and statistics communities

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning

- Preprocessing of the data (including feature extraction and dimension reduction)

- Classification or/and clustering processes

- Post-processing for presentation

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Data mining functions**
  - Characterization, discrimination, association, classification, clustering, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications

  - Relational database, data warehouse, transactional database

- Advanced data sets and advanced applications

  - Data streams and sensor data

  - Time-series data, temporal data, sequence data (incl. bio-sequences)

  - Structure data, graphs, social networks and multi-linked data

  - Object-relational databases

  - Spatial data

  - Multimedia database

  - Text databases

  - The World-Wide Web

Data Mining Function: (1) Generalization

- Information integration and data warehouse construction
    - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
    - Scalable methods for computing (i.e., materializing) multidimensional aggregates
    - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
    - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer
    - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

# Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

What Technologies are used?
Data Mining: Confluence of Multiple Disciplines
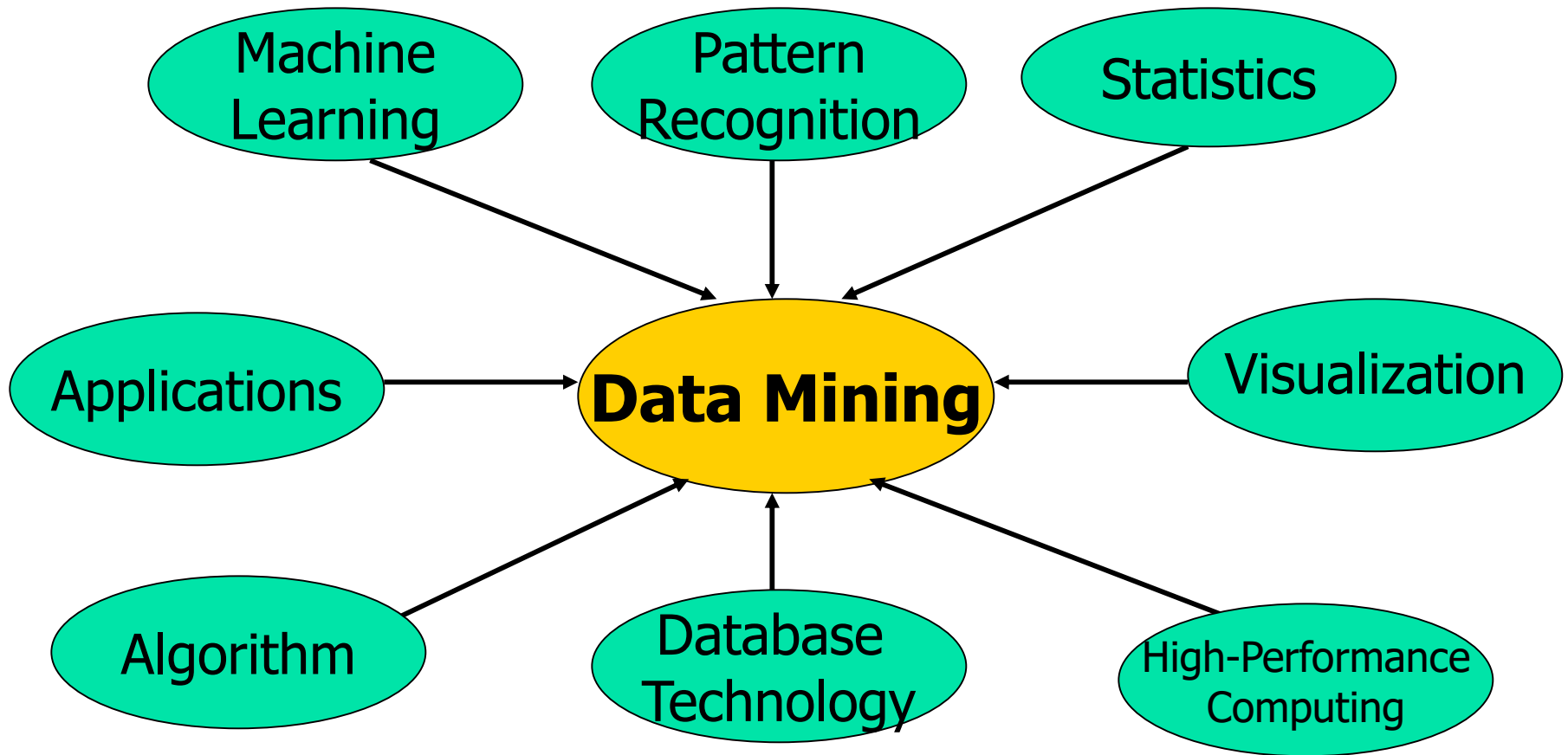
What Kind of Applications are Targeted

- Web page analysis

- Collaborative analysis & recommender systems

- Basket data analysis to targeted marketing

- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis

- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Major Issues in Data Mining (1)

- Mining Methodology

  - Mining various and new kinds of knowledge

  - Mining knowledge in multi-dimensional space

  - Data mining: An interdisciplinary effort

  - Boosting the power of discovery in a networked environment

  - Handling noise, uncertainty, and incompleteness of data

  - Pattern evaluation and pattern- or constraint-guided mining

- User Interaction

  - Interactive mining

  - Incorporation of background knowledge

  - Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data

- A natural evolution of database technology, in great demand, with wide applications

- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation

- Mining can be performed in a variety of data

- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

- Data mining technologies and applications

- Major issues in data mining