DATA MINING AND BUSINESS INTELLIGENCE - LECTURE 01

Dr. Mahmoud Mounir

mahmoud.mounir@cis.asu.edu.eg

COURSE OUTLINE

Data Mining and Business Intelligence

- Overview of Data Mining.

- Data Exploration

  - Data Preprocessing.

  - Outlier Analysis.

- Association Analysis

  - Mining Frequent Patterns: Apriori Algorithm, FPGrowth, ECLAT.

  - Associations Rules Evaluation and Correlations.

COURSE OUTLINE

- Correlation Analysis

  o Probability and Statistics Basic Concepts.

  o Correlation and Linear Regression.

- Unsupervised Learning (Cluster Analysis)

  o Centroid-Based Methods: K-Means.

  o Distance-Based Methods: Hierarchical Clustering.

  o Density-Based Methods: DBSCAN.

  o Clustering Model Evaluation.

COURSE OUTLINE

- Supervised Learning (Classification)

  o Decision Trees.

  o Naïve Bayes Classification.

  o K-Nearest Neighbors (KNN).

  o Rule-based classification.

  o Classification Model Evaluation and Selection.

- Graph-Based Mining and Social Network Analysis.

COURSE OUTLINE

A- Methods Used			
	1.	Midterm	
	2.	Quizzes	
	3.	Project	
	4.	Assignments	
	5.	Lecture Tasks	
	6.	Final Exam	
B- Assessment Schedule			
	1.	Midterm	week 8
	2.	Quizzes	week5 ,week10
	3.	Project	week9-week11
	4.	Assignments	week4, week9
	5.	Lecture Tasks	week3 ,week6
	6.	Final Exam	week 15
C- Weighting of Assessments			
	1.	Midterm	25%
	2.	Quizzes	10%
	3.	Project	10%
	4.	Assignments	10%
	5.	Lecture Tasks	5%
	6.	Final Exam	40%
		Total	100%

REFERENCES

❑ Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques", 3rd edition, The Morgan Kaufmann Series in Data Management Systems

❑ Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining", 2nd Edition, Addison Wesley, 2014.

Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 1 —

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University

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?

- 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

Large-scale Data is Everywhere!
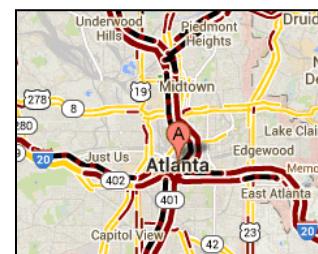
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies

- New mantra
  - Gather whatever data you can whenever and wherever possible.

- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.
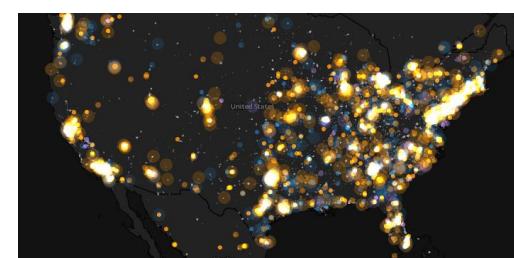
Cyber Security
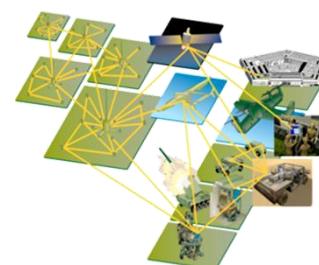
E-Commerce
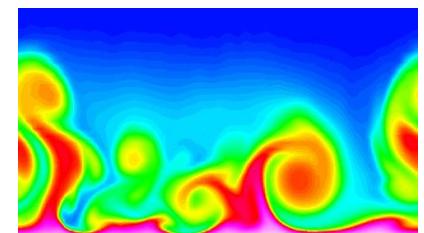
Traffic Patterns

Social Networking: Twitter

Sensor Networks

Computational Simulations
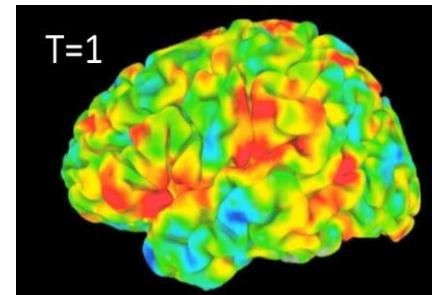
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
    - Web data
        - Google has Peta Bytes of web data
        - Facebook has billions of active users
    - purchases at department/ grocery stores, e-commerce
        - Amazon handles millions of visits/day
    - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
    - Provide better, customized services for an edge (e.g. in Customer Relationship Management)
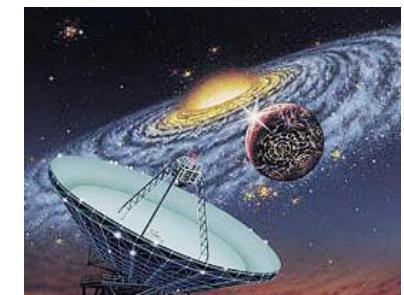
Why Data Mining? Scientific Viewpoint
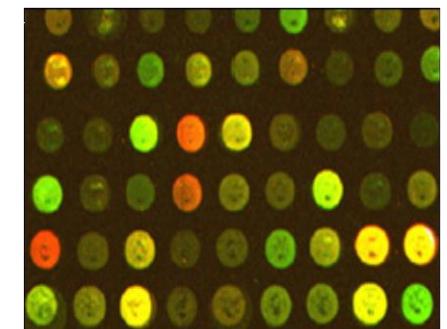
- Data collected and stored at enormous speeds

  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year

  - telescopes scanning the skies
    - Sky survey data

  - High-throughput biological data

  - scientific simulations
    - terabytes of data generated in a few hours

- Data mining helps scientists
  - in automated analysis of massive datasets
 - In hypothesis formation
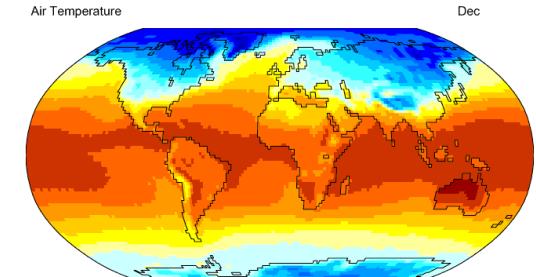
fMRI Data from Brain

Sky Survey Data

Gene Expression Data

Surface Temperature of Earth

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

- **We are drowning in data, but starving for knowledge!**

- "**Necessity is the mother of invention**"—Data mining—Automated analysis of massive data sets

Evolution of Database Technology

- 1960s:
    - Data collection, database creation, Information Management System **(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

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?

- 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

What Is Data Mining?

- **Data mining (knowledge discovery from data)**
 - Extraction of interesting (<u>non-trivial,</u> <u>implicit</u>, <u>previously unknown</u> and <u>potentially useful)</u> patterns or knowledge from huge amount of data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
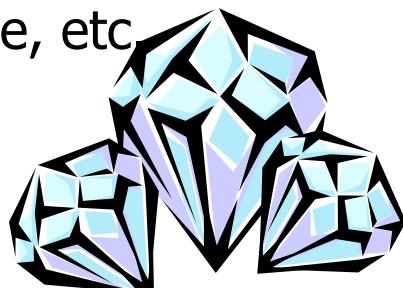
- **Alternative names**
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
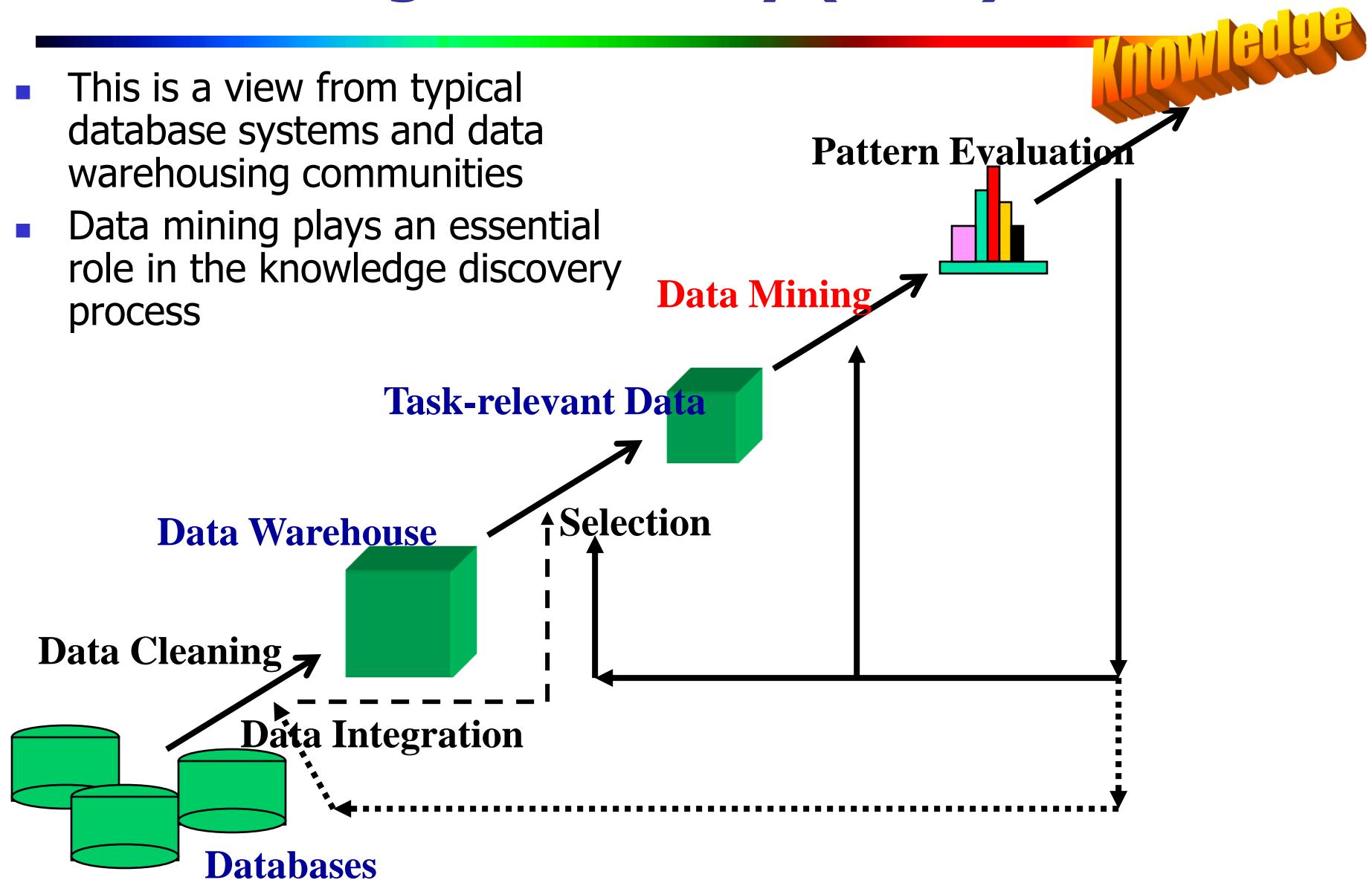
- **Watch out: Is everything "data mining"?**
  - Simple search and query processing
  - (Deductive) expert systems

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Data Warehouse

Selection

Data Cleaning

Data Integration

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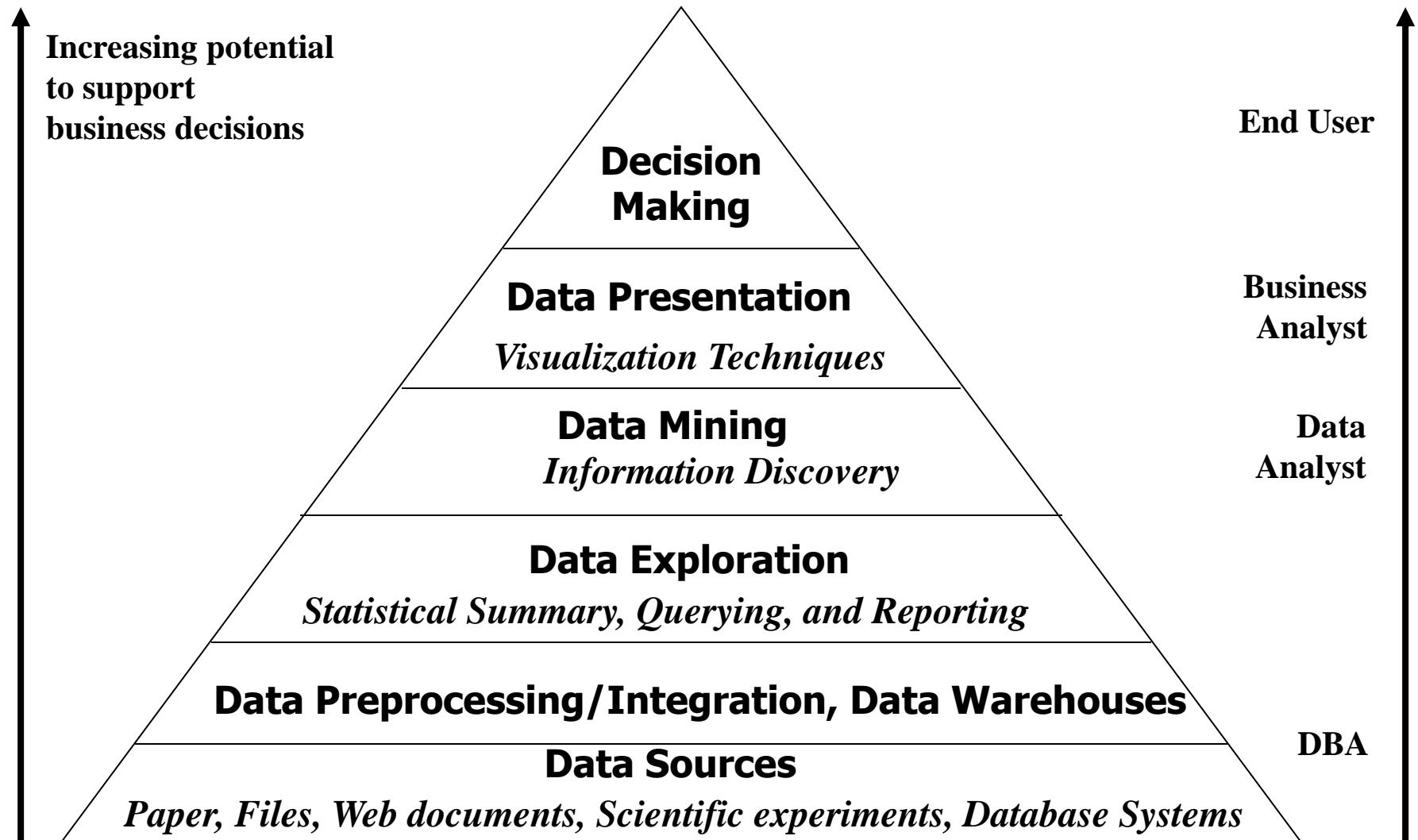
Example: A Web Mining Framework
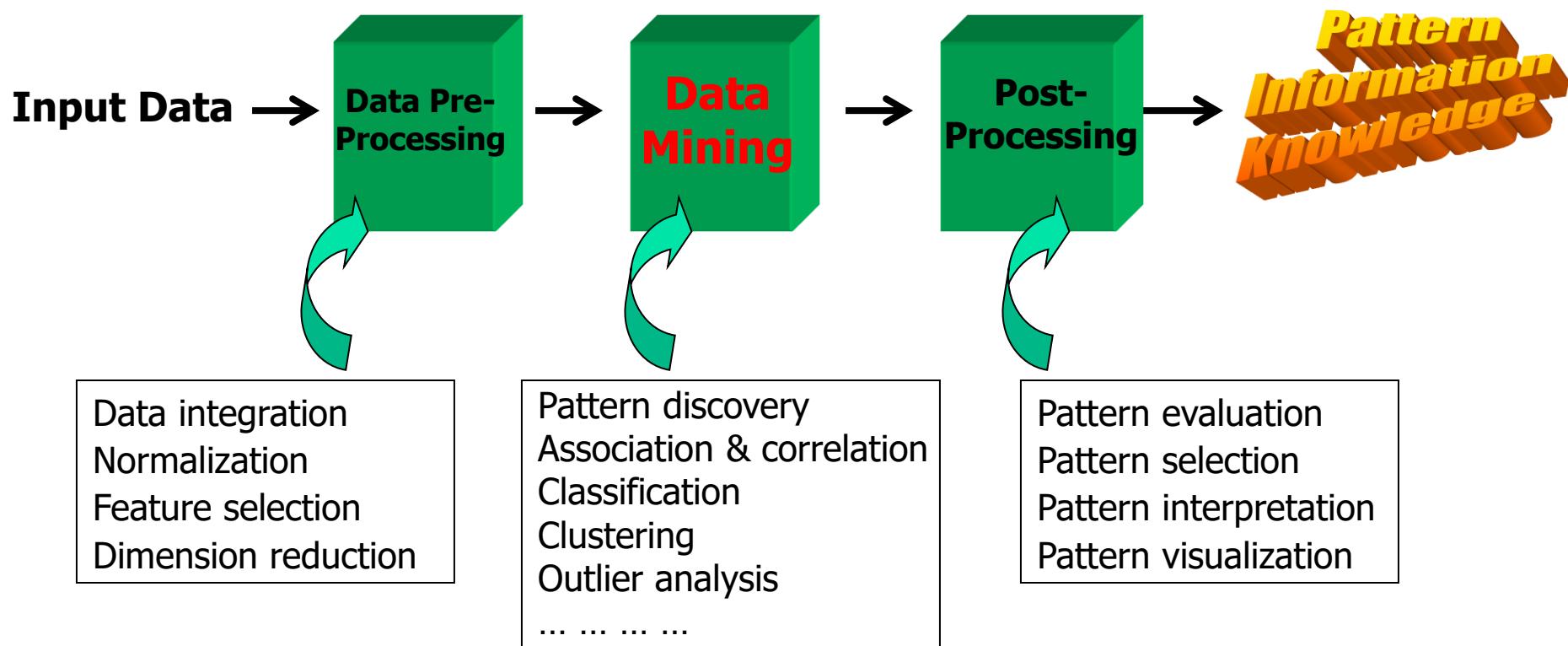
- **Web mining usually involves**
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

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

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?

- 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

Multi-Dimensional View of Data Mining

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

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?

- 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

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

  - Heterogeneous databases and legacy databases

  - Spatial data and spatiotemporal data

  - Multimedia database

  - Text databases

  - The World-Wide Web

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?

- 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

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 → Juice [0.5%, 75%] (support, confidence)
  - 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

Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
   - Web community discovery, opinion mining, usage mining, …
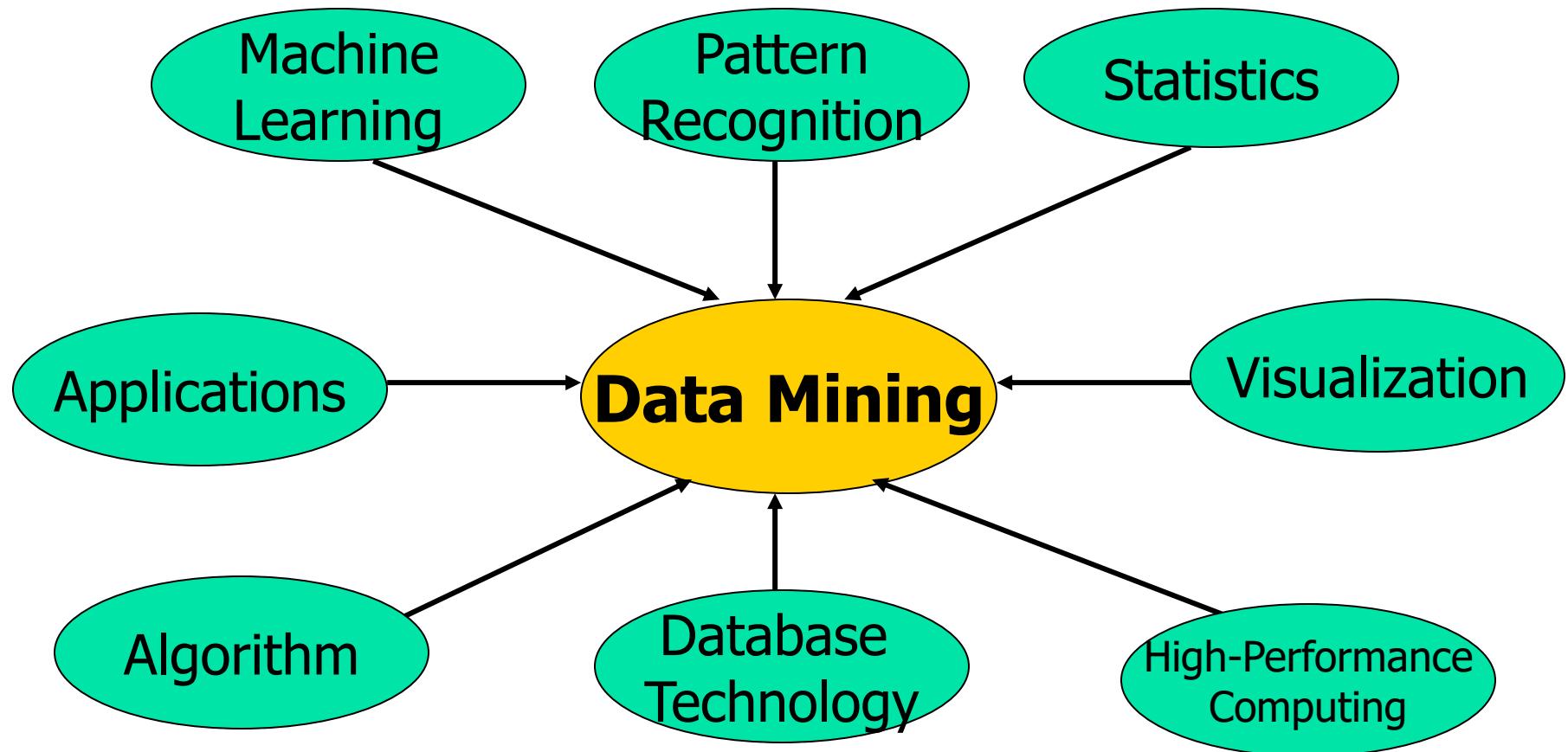
Evaluation of Knowledge

- 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
    - …

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?

- 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

Data Mining: Confluence of Multiple Disciplines

Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Microarray may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

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?

- 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

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms

- 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

- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)

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

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?

- 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

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

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?

- 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

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)

- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD

- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.

- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,

- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.

- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

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?

- 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

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

Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002

- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000

- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996

- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011

- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001

- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009

- B. Liu, Web Data Mining, Springer 2006.

- T. M. Mitchell, Machine Learning, McGraw Hill, 1997

- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991

- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005

- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998

- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005