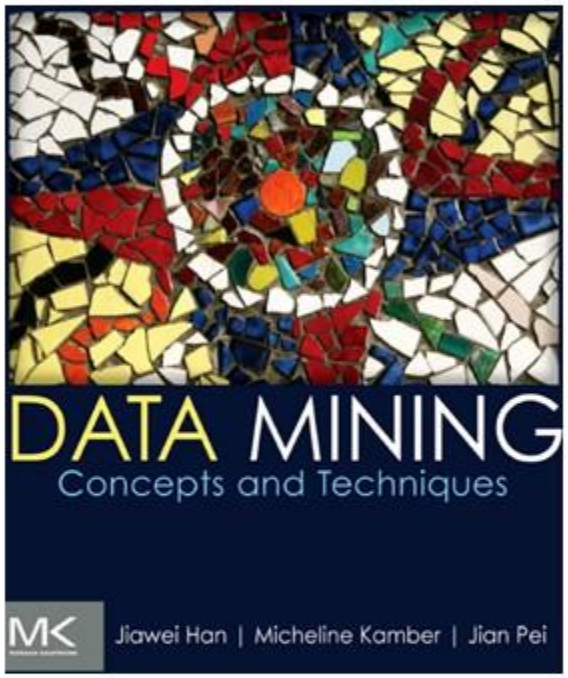
Data Mining



Instructor:

Dr. Mohamed Hassan

Textbook (s)

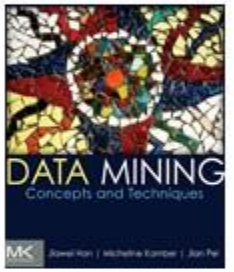
Main textbook,

- **Data Mining Concepts and Techniques** (3rd ed.)

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University



- **Introduction to Data Mining,** 2nd Edition

Tan, Steinbach,

Karpatne, Kumar

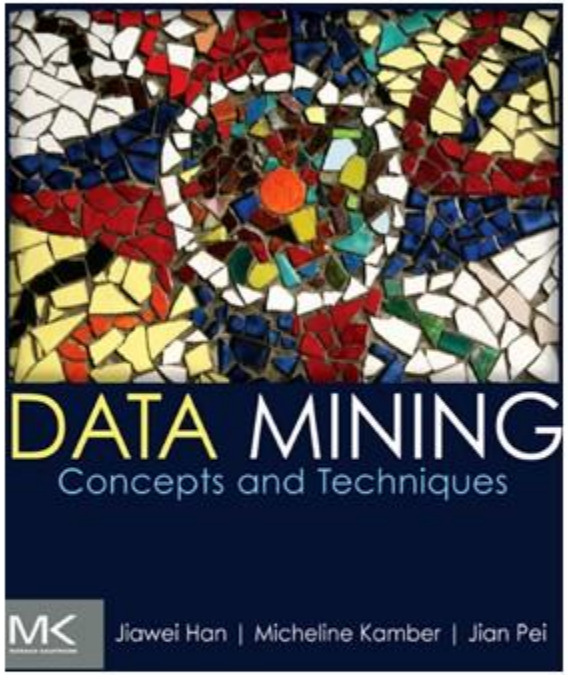
Modified for Introduction to Data Mining by Dr. Mohamed Hassan

Course Outline

- ★ Part 1: Introduction
- ★ Part 2: Getting to Know Your Data
- ★ Part 3: Data Preprocessing
- ★ Part 4: Association Analysis: Basic Concepts
- ★ Part 5: Classification Basic Concepts

Chapter 1

- INTRODUCTION



Chapter 1 LEARNING OBJECTIVES

- Why Data Mining?
- What Is Data Mining?
- A Multi-DimensionalView 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

Why Data Mining?

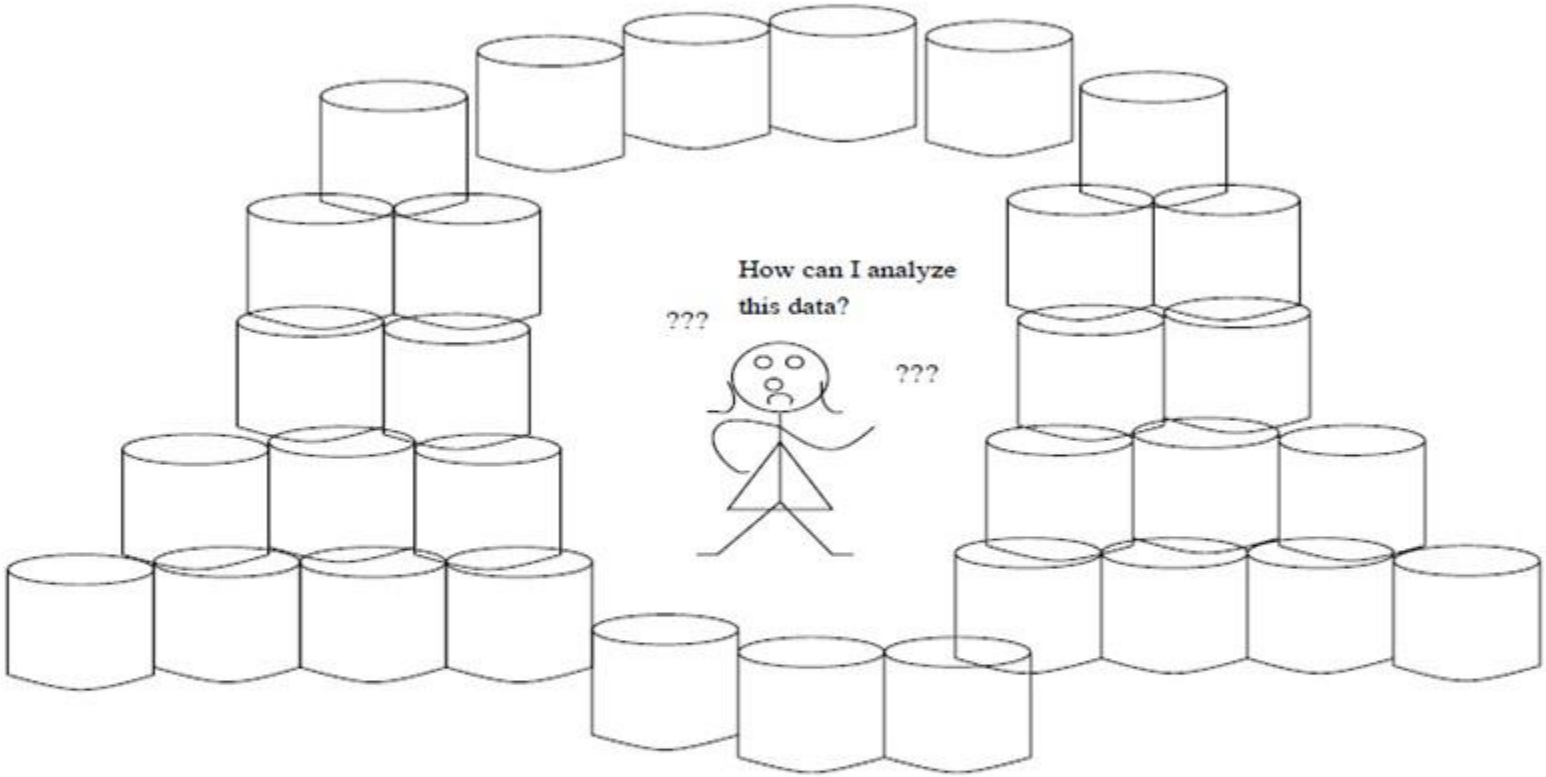
Evolution of Database Technology

- 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

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 data**
    - **Business**: Web, e-commerce, transactions, stocks, …
    - **Science**: Remote sensing, bioinformatics, scientific simulation, …
    - **Society** and everyone: news, digital cameras, YouTube

Lots of Available Data



We are drowning in data, but starving for knowledge!

"Necessity is the mother of invention"

Data mining — Automated analysis of massive data sets

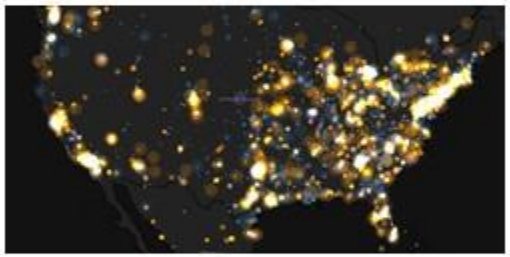
Large-scale Data is Everywhere!

- There has been enormous data growth in both **commercial** and **scientific** databases due to advances in **data generation and collection technologies**

Cyber Security

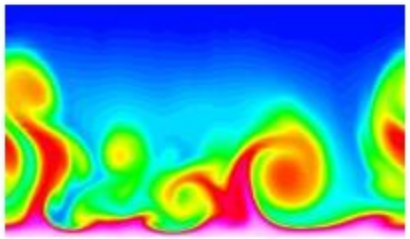
E-Commerce

Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

Why Data Mining? Commercial Viewpoint

- **Lots of data is being collected and warehoused**
  - **Web data**
    - Yahoo 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**

Google

facebook

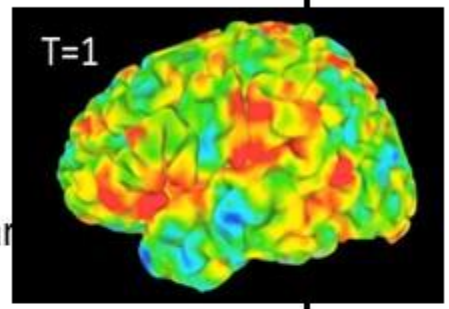
YAHOO!

amazon.com

- **Computers have become cheaper and more powerful**
- **Competitive Pressure is Strong**
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

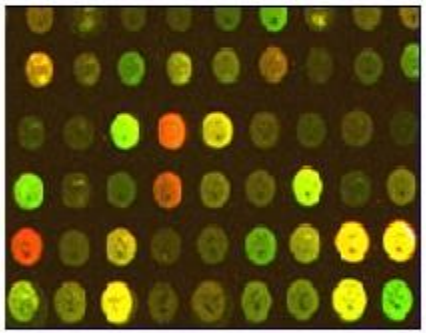
Why Data Mining? Scientific Viewpoint

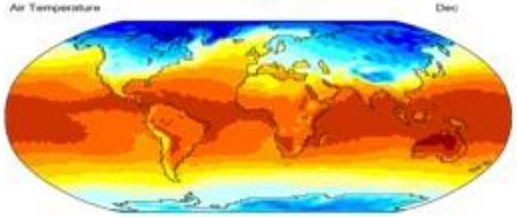
- **Data collected and stored at enormous speeds**
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- **Data mining helps scientists**
  - in automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain

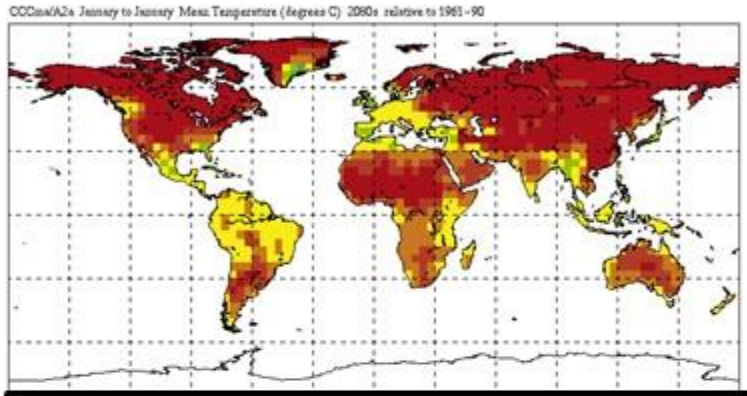

Sky Survey Data


Gene Expression Data


Surface Temperature of Earth

Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

What is Data Mining?

- **Many Definitions**
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data.
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful and hidden patterns
  - Simply stated, extracting or mining knowledge from large amounts of data
- Data mining (**knowledge discovery from data**) **KDD**
- **Alternative names**
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.

Mining and Querying Database

T-id	Items	Total Cost	Date
1	Butter, Bread, Cereal, milk	35	15/10/2012
2	Diapers, eggs, sugar	90	15/10/2012
3	Sugar, eggs, milk, cereal	100	16/10/2012
4	Milk, Cereal, bread, Diapers, bread	130	16/10/2012
5	Eggs, milk, bread, cereal, sugar	68	16/10/2012

- **Select items from tale where total cost > 40**
- **No such query can bring relation between:**

Cereal, milk

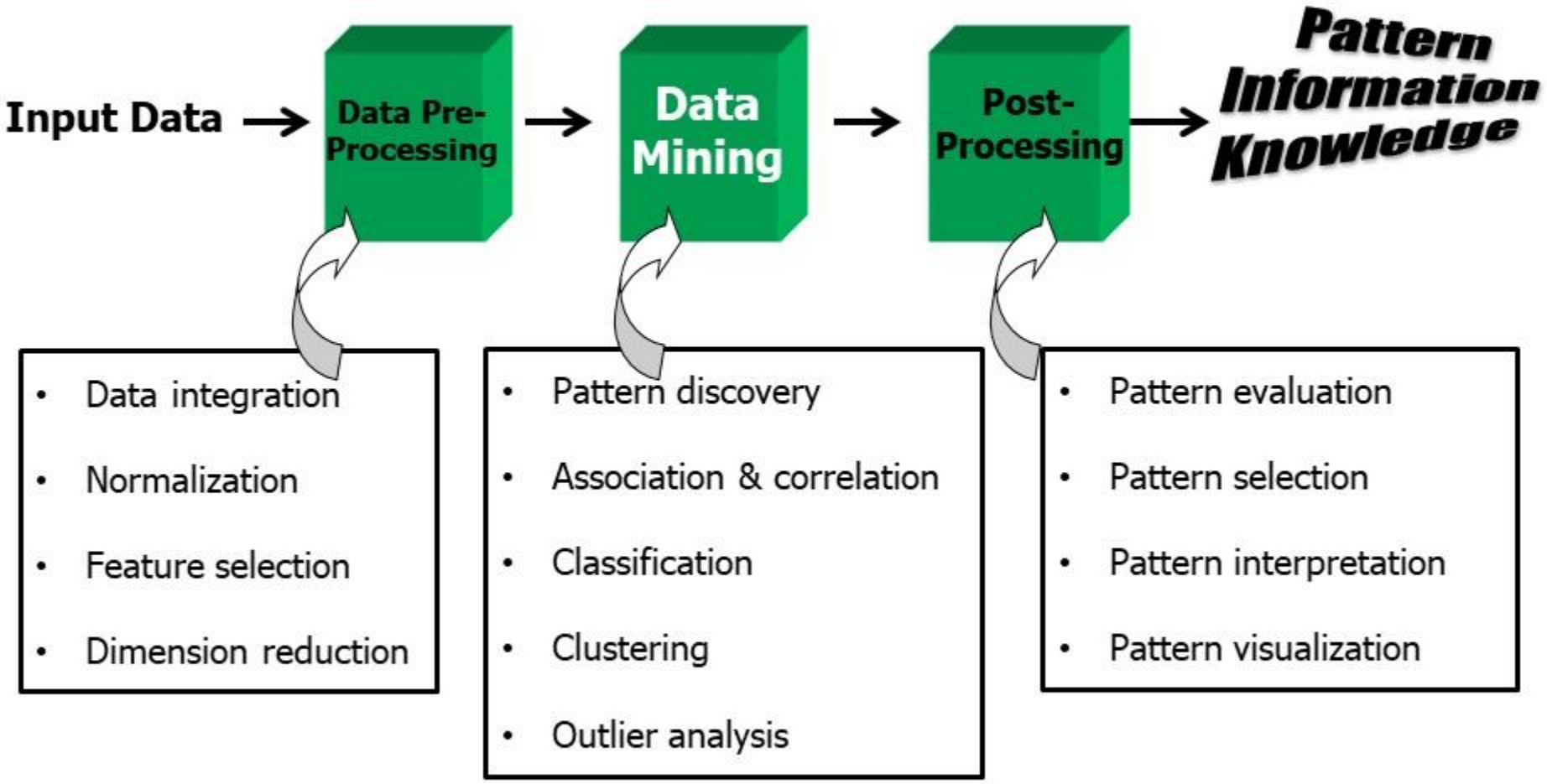
Eggs, sugar

What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

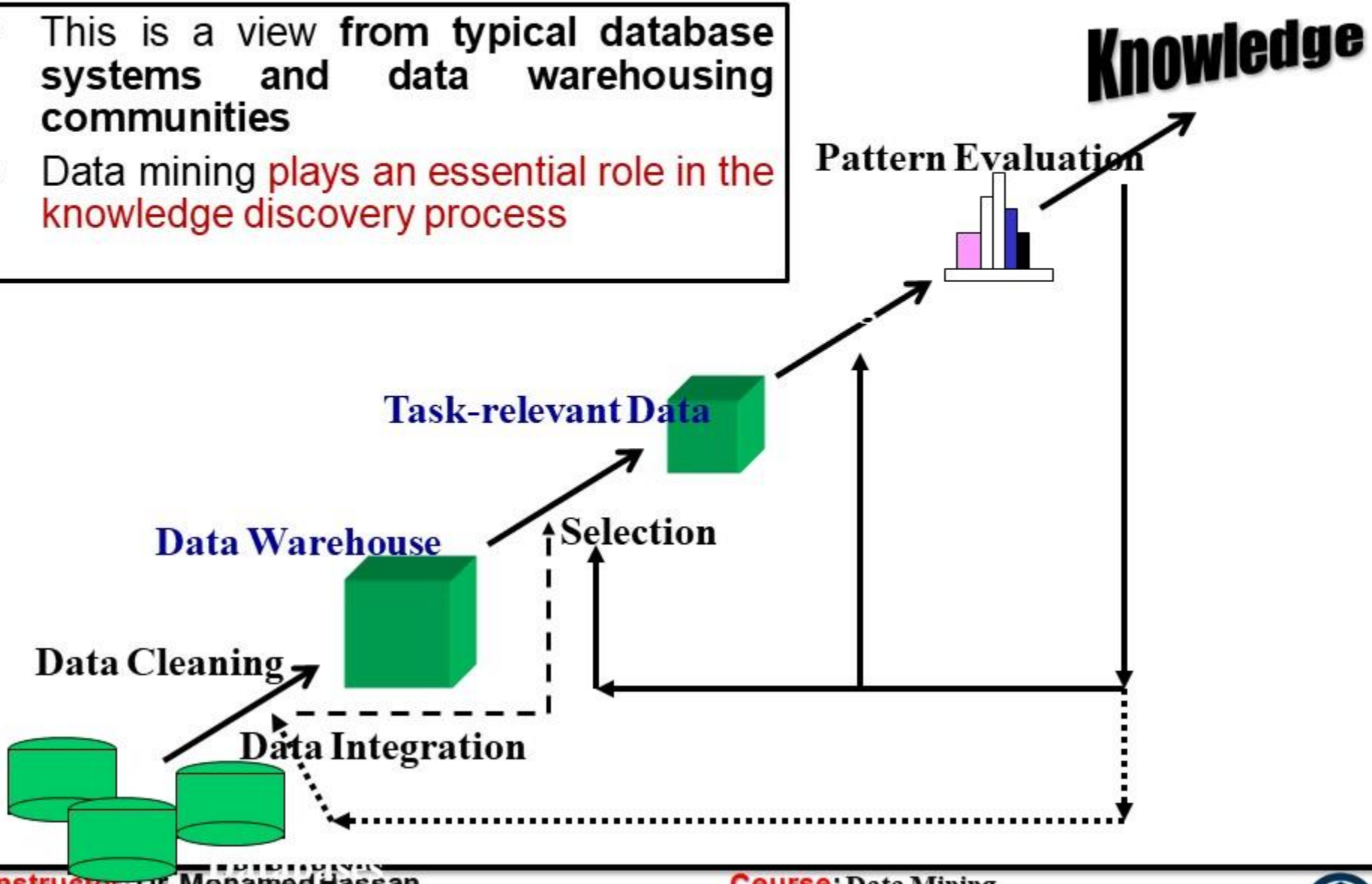
- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rourke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

KDD Process: A Typical View from ML and Statistics



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 Discovery Process (cont.)

Data Cleansing: remove noise or irrelevant data and outliers

Data Integration: multiple data sources are combined (sometimes combined with data cleansing named preprocessing step)

Data Selection: data relevant to the analysis task are retrieved from the database

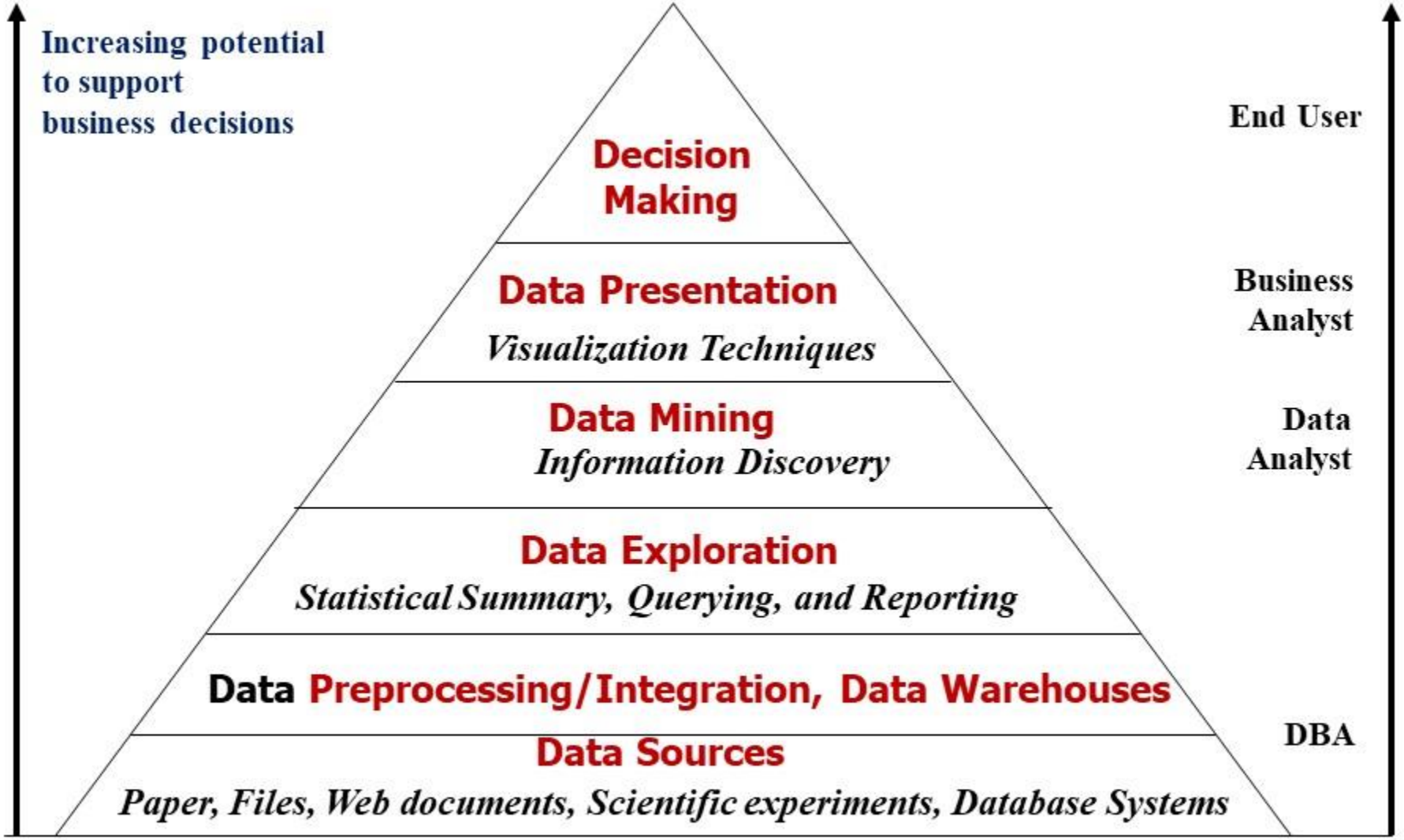
Data Transformation: data are transformed into forms appropriate for mining by performing aggregation or summary.(sometimes this step is performed before data selection)

Data Mining: Intelligent methods are applied in order to extract data patterns

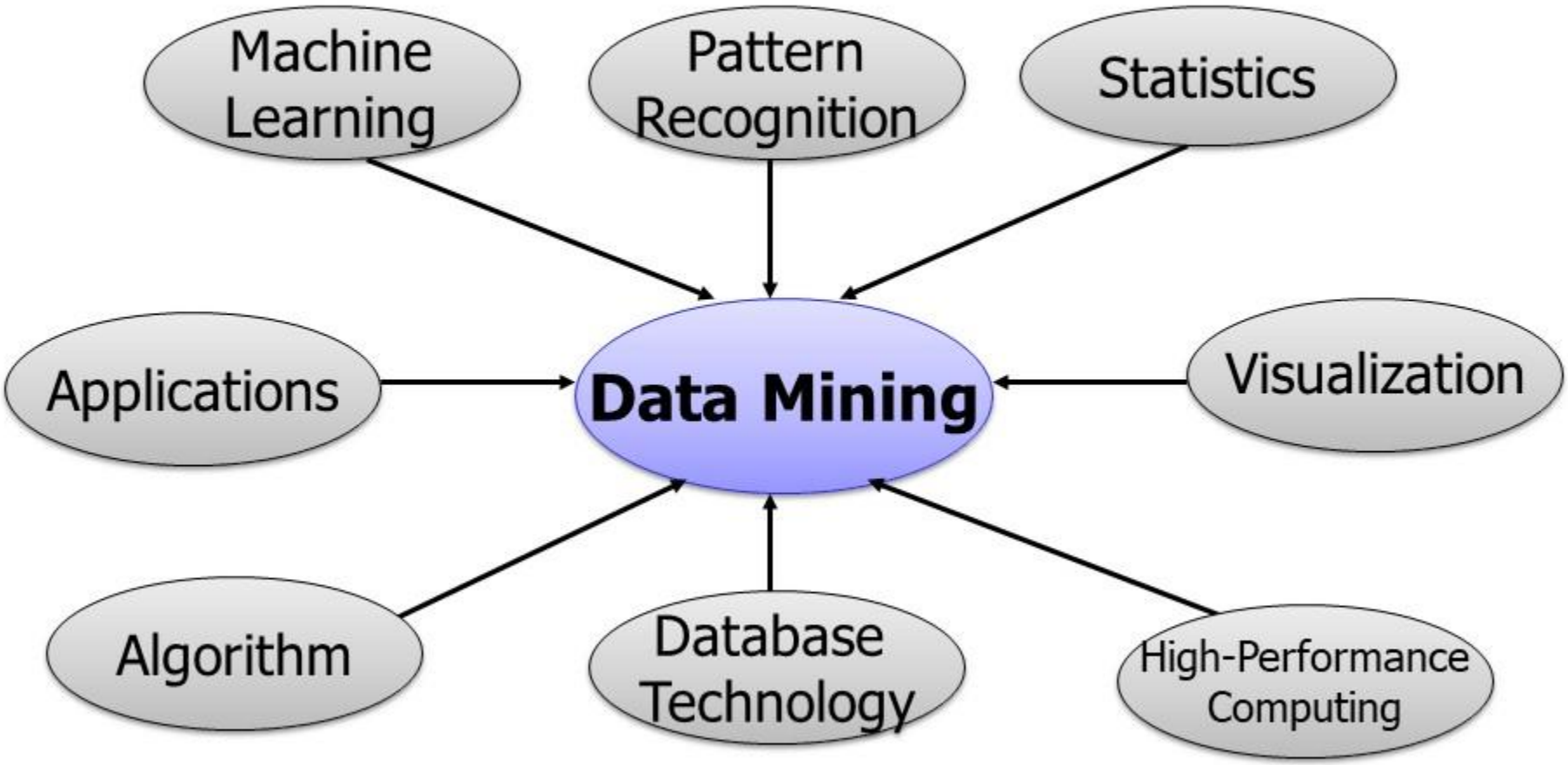
Pattern Evaluation: identify interesting patterns representing knowledge base.

Knowledge Representation: visualization and knowledge representation techniques are used to represent the extracted patterns.

Data Mining in Business Intelligence



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**
  - Micro-array 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**

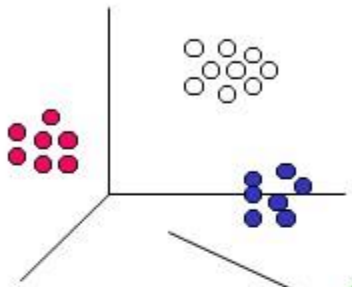
Data Mining Tasks

- **Prediction Tasks**
  - Use some variables to predict unknown or future values of other variables
- **Description Tasks**
  - Find human-interpretable patterns that describe the data.

Common data mining tasks

- Classification [**Predictive**]
- Clustering [**Descriptive**]
- Association Rule Discovery [**Descriptive**]
- Sequential Pattern Discovery [**Descriptive**]
- Regression [**Predictive**]
- Deviation Detection [**Predictive**]

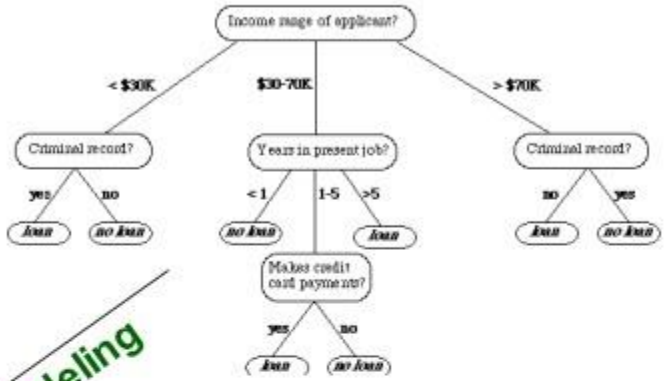
Data Mining Tasks ...



Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



Predictive Modeling

Association Rules

Anomaly Detection

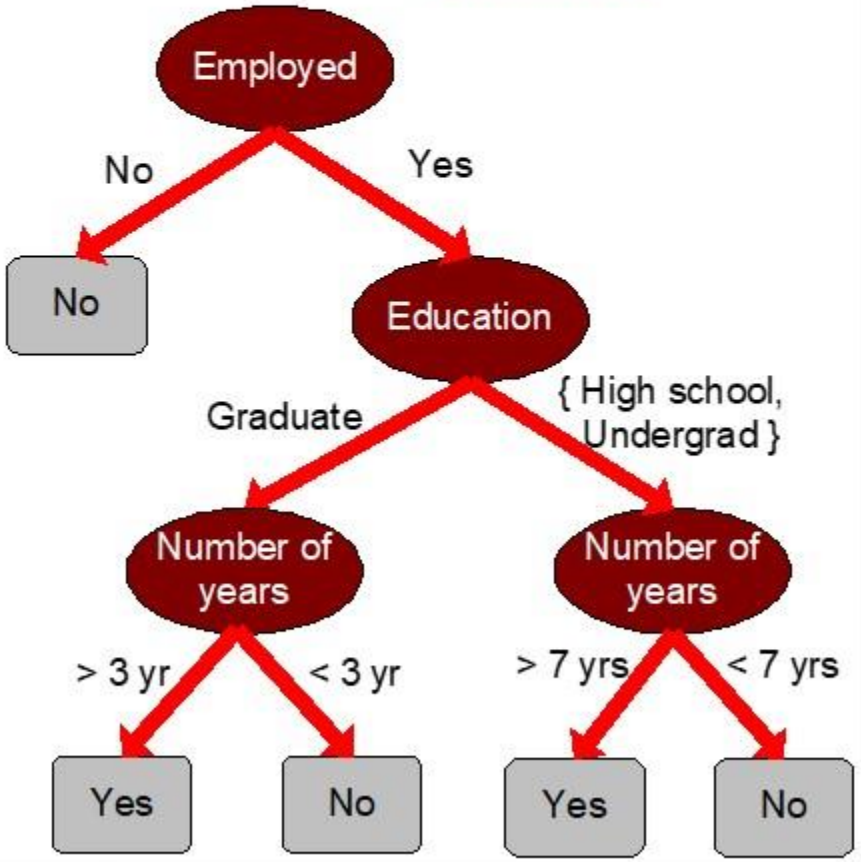
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness

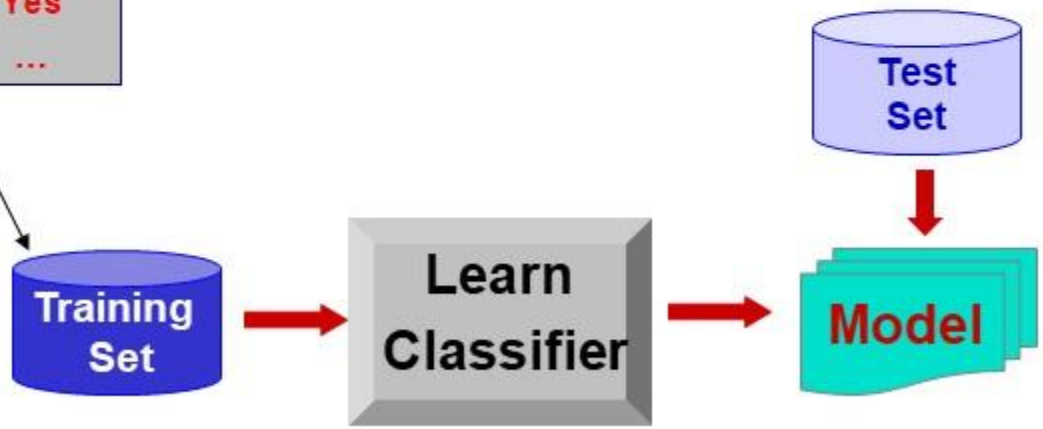


Classification Example

categorical categorical quantitative class

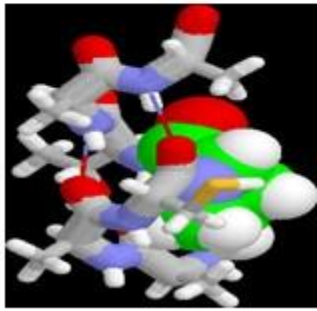
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Examples of Classification Task

- Classifying **credit card transactions** as legitimate or fraudulent
- Classifying **land covers** (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing **news stories** as finance, weather, entertainment, sports, etc
- Identifying **intruders** in the cyberspace
- Predicting **tumor** cells as benign or **malignant**
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



Classification: Application 1

- Fraud Detection
  - **Goal: Predict fraudulent cases in credit card transactions.**
  - **Approach:**
    - Use **credit card transactions** and the **information on its account-holder** as **attributes**.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - **Label past transactions** as **fraud** or fair transactions. This forms the class attribute.
    - **Learn a model** for the class of the transactions.
    - **Use this model to detect fraud** by observing credit card transactions on an account.

Classification: Application 2

- Churn prediction for telephone customers
  - **Goal: To predict whether a customer is likely to be lost to a competitor.**
  - **Approach:**
    - Use **detailed record of transactions with each of the past and present customers**, to find **attributes**.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - **Label the customers as loyal** or disloyal.
    - **Find a model** for loyalty.

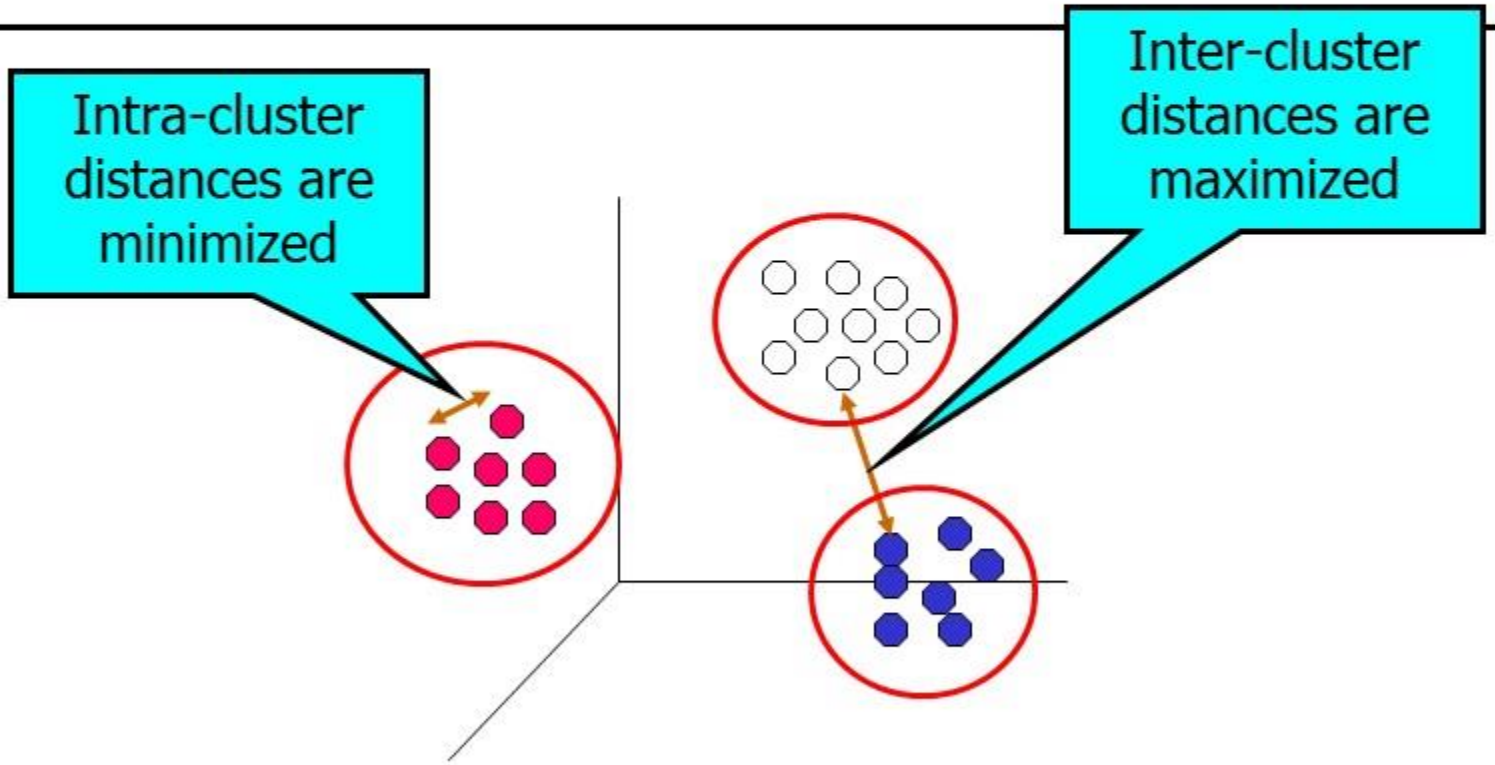
From [Berry & Linoff] Data Mining Techniques, 1997

Regression

- **Predict a value of a given continuous valued variable based on the values of other variables**, assuming a linear or nonlinear model of dependency.
- Extensively studied in **statistics**, **neural network** fields.
- **Examples**:
  - **Predicting sales amounts** of new product based on advertising expenditure.
  - **Predicting wind velocities** as a function of temperature, humidity, air pressure, etc.
  - **Time series prediction** of stock market indices.

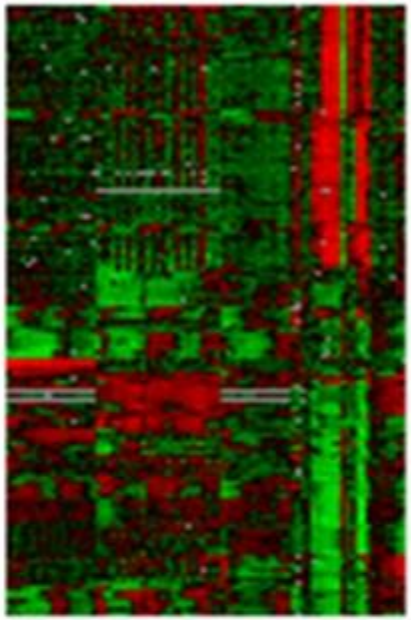
Clustering

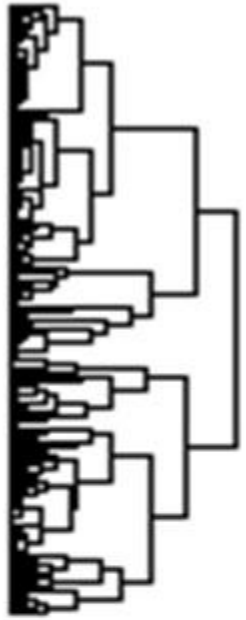
- **Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups**

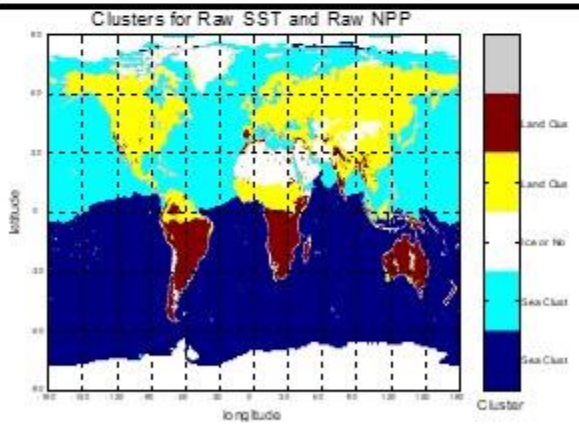


Applications of Cluster Analysis

- **Understanding**
  - **Custom profiling** for targeted marketing
  - **Group related documents** for browsing
  - **Group genes and proteins** that have similar functionality
  - **Group stocks with similar price fluctuations**
- **Summarization**
  - Reduce the size of large data sets







Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.

Courtesy: Michael Eisen

Clustering: Application 1

- Market Segmentation:
  - **Goal: subdivide a market into distinct subsets of customers** where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - **Collect different attributes of customers based on their geographical and lifestyle related information.**
    - **Find clusters** of similar customers.
    - **Measure the clustering quality** by observing buying patterns of customers in same cluster vs. those from different clusters.

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection. Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

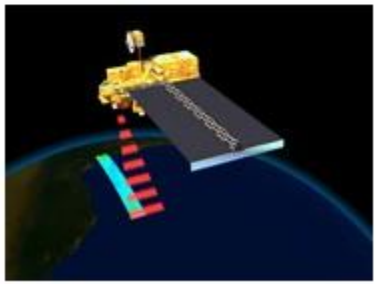
{Diaper, Milk} --> {Beer}

Association Analysis: Applications

- Market-basket analysis
  - Rules are used for **sales promotion, shelf management, and inventory management**
- Telecommunication alarm diagnosis
  - Rules are used to **find combination of alarms that occur together frequently in the same time period**
- Medical Informatics
  - Rules are used to **find combination of patient symptoms and test results associated with certain diseases**

Deviation/Anomaly/Change Detection

- **Detect significant deviations from normal behavior**
- **Applications**:
  - **Credit Card Fraud** Detection
  - **Network Intrusion** Detection
  - Identify **anomalous behavior from sensor networks** for monitoring and surveillance.
  - Detecting changes in the global forest cover.



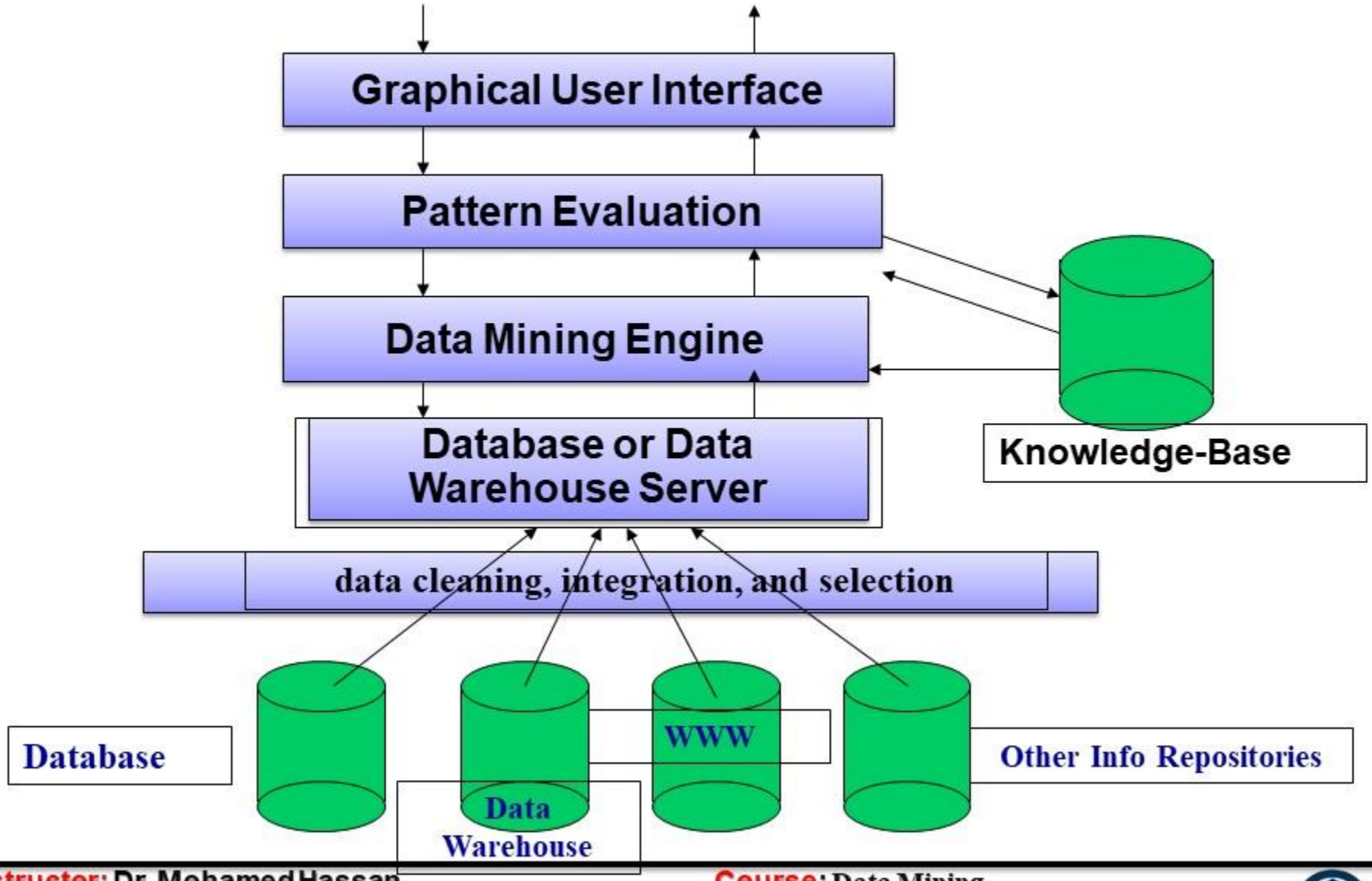
Data Mining: On What Kinds of Data?

- **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

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, …

Architecture: Typical Data Mining System



Major Issues and challenges in Data Mining

Challenges of Data Mining

- Scalability
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Challenges of Data Mining

- Scalability:

  Scalability means the ability of the algorithm to work well on small and large data. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases.

Challenges of Data Mining

- Complex and Heterogeneous Data:

  Recent years have seen the emergence of new and complex source of information like for example **semi-structured data and hyperlinks, XML documents.**

  Relationship between data should be considered like **graph connectivity, temporal and sequence data.**

Challenges of Data Mining

- **Data Quality:**

  It is about the quality of data that will affect the resulting pattern.

  Factors affecting quality are:

  Noise and outliers, duplicate data, missing values and inconsistent data

- **Data Ownership and Distribution:**

  Sometimes data needed for an analysis is not stored in one location or owned by one organization.

  Challenges:

(1) reduce amount of communication needed to perform computation

(2) consolidate data mining results form multiple sources

(3) data security issue

Challenges of Data Mining

- Privacy Preservation:

  Data mining is always accused to threaten the individual privacy. **Data mining is a source of information leakage**

- Streaming Data:

  Data stream is **continuous no stored data like network traffic. This is a big challenge for data miners and there are a lot of algorithms introduced for such kind of data**

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.

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