

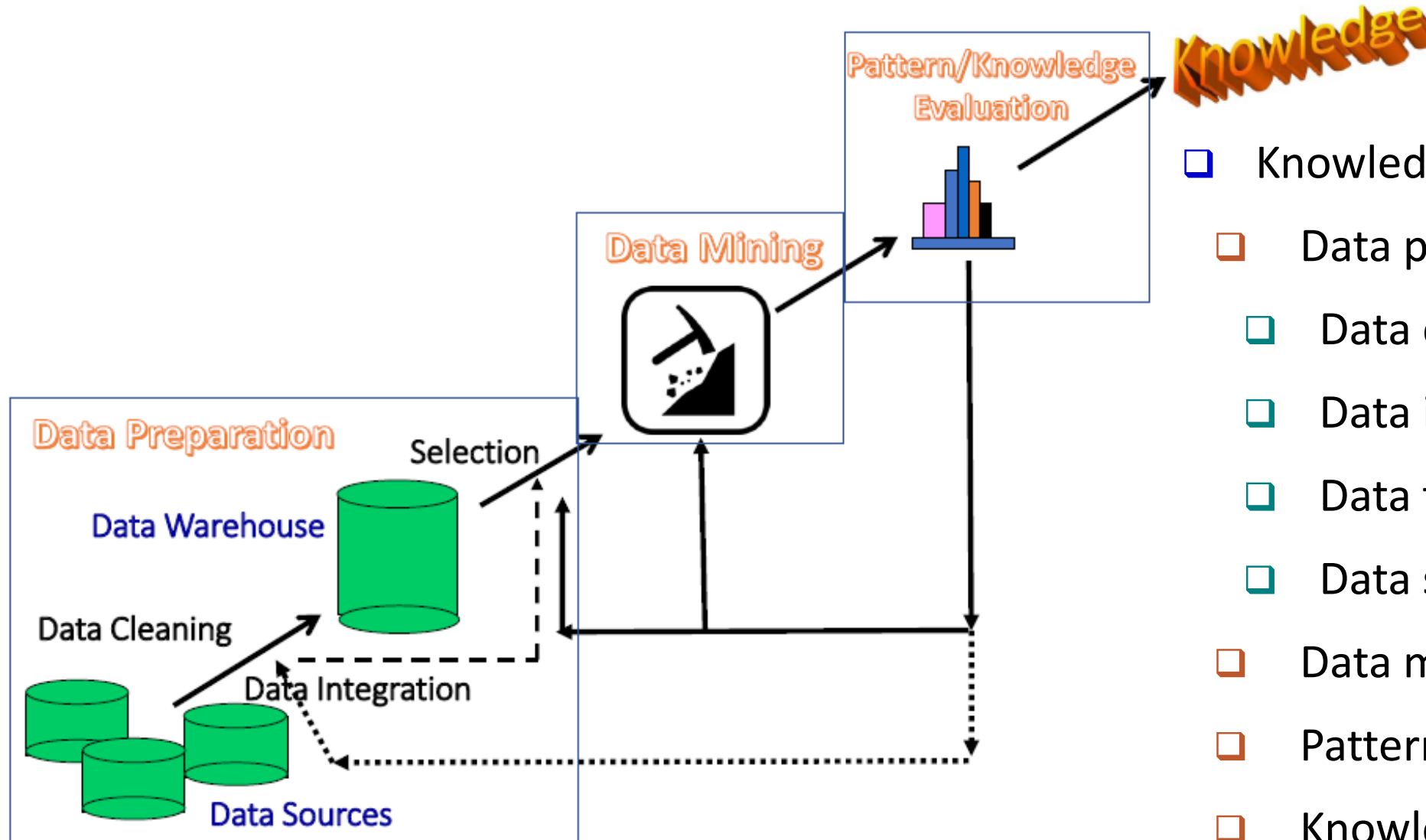
Chapter 1. Introduction

- ☐ What Is Data Mining?
- ☐ Data Mining: An Essential Step in Knowledge Discovery
- ☐ Diversity of Data Types for Data Mining
- ☐ Mining Various Kinds of Knowledge
- ☐ Data Mining: Confluence of Multiple Disciplines
- ☐ Data Mining and Applications
- ☐ Data Mining and Society
- ☐ Summary

What Is Data Mining?

- ❑ We live in a world where vast amounts of data are generated constantly and rapidly
- ❑ **Data mining** is the process of discovering interesting patterns, models and other kinds of knowledge in large data sets
 - ❑ “Data mining”: a misnomer? It should be “knowledge mining from data”
 - ❑ Other terms: *Knowledge mining from data, KDD (Knowledge Discovery from Data), pattern discovery, knowledge extraction, data analytics, information harvesting*
- ❑ Data mining is a young, dynamic, and promising field
- ❑ Example: Data mining turns a large collection of data into knowledge
 - ❑ Google’s *Flu Trends* found a close relationship between the number of people who search for flu-related info. and the number of people who have flu symptoms
 - ❑ It can estimate flu activity up to two weeks faster than traditional systems

Data Mining: An Essential Step in Knowledge Discovery



□ Knowledge Discovery Process

□ Data preparation

□ Data cleaning

□ Data integration

□ Data transformation

□ Data selection

□ Data mining

□ Pattern/model evaluation

□ Knowledge presentation

Diversity of Data Types for Data Mining (I)

- ❑ **Structured vs. unstructured data**
 - ❑ *Structured*: uniform, record- or table-like structures, defined by data dictionaries, with a fixed set of attributes, each with a fixed set of value ranges and semantic meaning
 - ❑ Ex. Data stored in *relational databases*, *data cubes*, *data matrices*, and many *data warehouses*
 - ❑ *Semi-structured*: allow a data object to contain a set value, a small set of heterogeneous typed values, or nested structures, or to allow the structure of objects or sub-objects to be defined flexibly and dynamically
 - ❑ Data having *certain structures* with clearly defined semantic meaning, such as *transactional data set*, *sequence data set* (e.g., time-series data, gene or protein data, or Weblog data)
 - ❑ *Graph or network data*: A more sophisticated type of semi-structured data set
 - ❑ *Unstructured data*: text data and multimedia (e.g., audio, image, video) data
- ❑ The real-world data can often be a mixture of structured, semi-structured data and unstructured data

Diversity of Data Types for Data Mining (II)

□ Data associated with different applications

- Different applications: different data sets and require different data analysis methods
 - Sequence data: *Biological sequences vs. shopping transaction sequences*
 - Time-series: ordered set of numerical values with equal time interval
 - Spatial, temporal and spatiotemporal data
 - Graph and network data: Social networks, computer communication networks, biological networks, and information networks may carry rather different semantics
- On the same data set, finding different kinds of patterns: require different mining methods
 - Ex. software programs: finding plagiarized modules vs. finding copy-and-paste bugs

□ Stored vs. streaming data

- Stored data: Finite, stored in various kinds of large data repositories
- Streaming data (e.g., video surveillance or remote sensing): Dynamic, constantly coming, infinite, real-time response—posing challenges on effective data mining

Mining Various Kinds of Knowledge

- ☐ Multidimensional Data Summarization
- ☐ Mining Frequent Patterns, Associations, and Correlations
- ☐ Classification and Regression for Predictive Analysis
- ☐ Cluster Analysis
- ☐ Deep Learning
- ☐ Outlier Analysis
- ☐ Are All Mining Results Interesting?

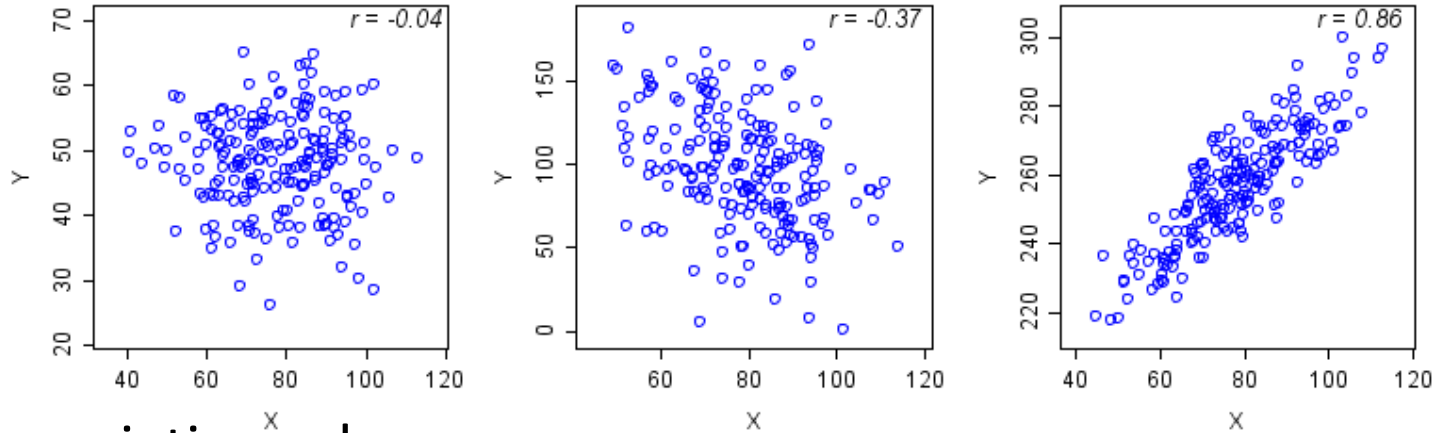
Multidimensional Data Summarization

- ❑ 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



Pattern Discovery: Mining Frequent Patterns, Associations, and Correlations

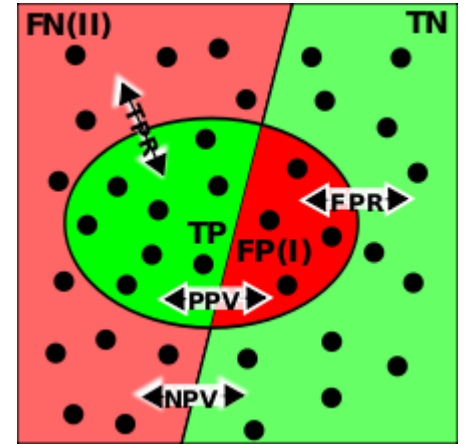
- Frequent patterns (or frequent itemsets)
 - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



- A typical association rule
 - Diaper \rightarrow Beer [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?

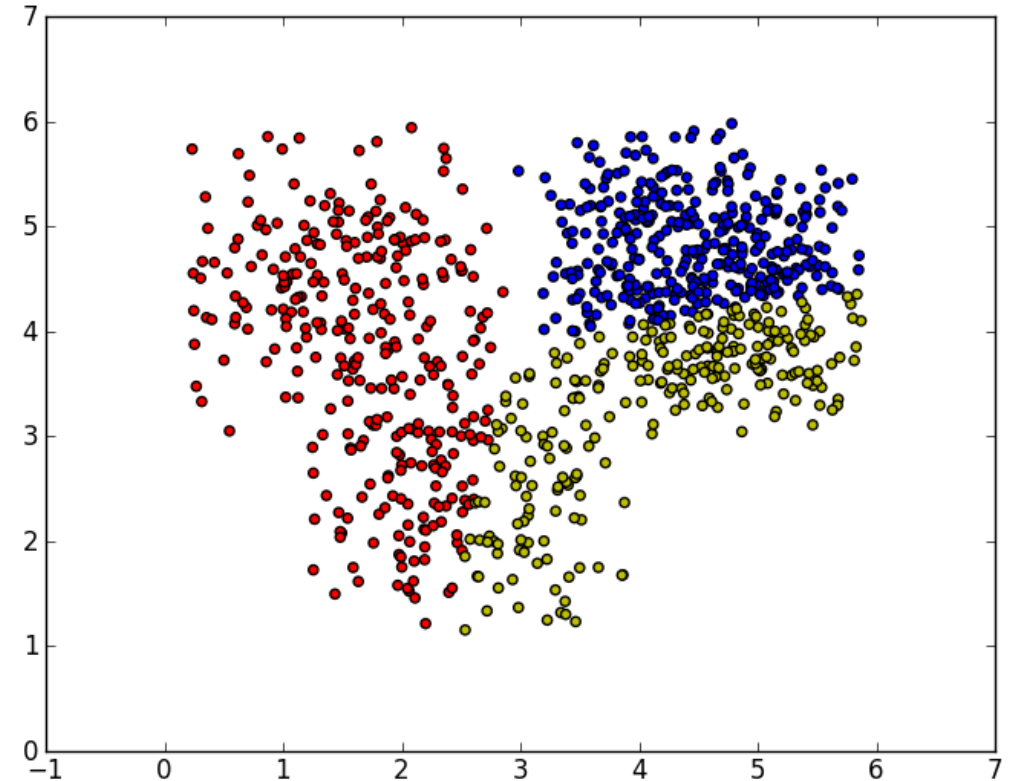
Classification and Regression for Predictive Analysis

- ❑ Classification and label prediction
 - ❑ Construct models (functions) based on some training examples
 - ❑ Describe and distinguish classes or concepts for future prediction
 - ❑ Ex. 1. Classify countries based on (climate)
 - ❑ Ex. 2. 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, ...



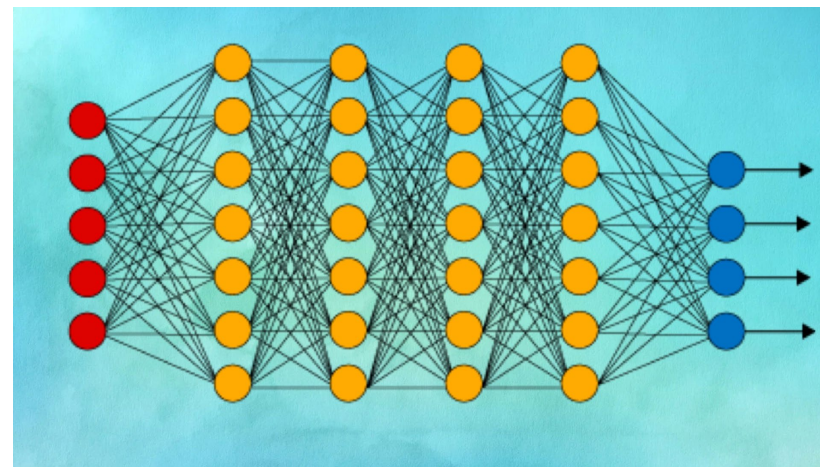
Cluster Analysis

- ❑ 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



Deep Learning

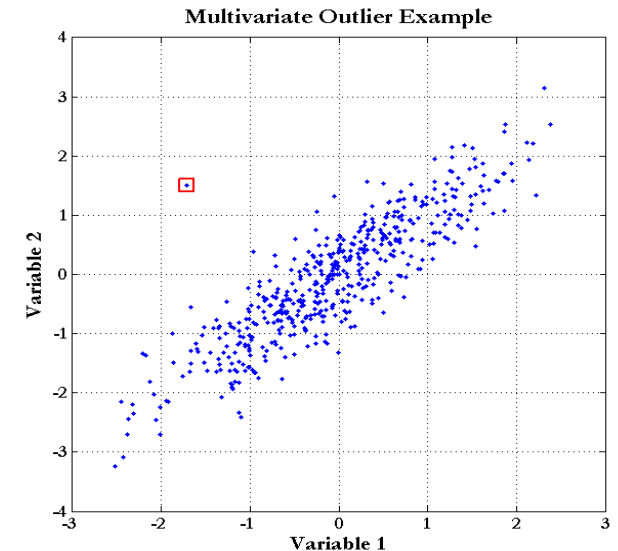
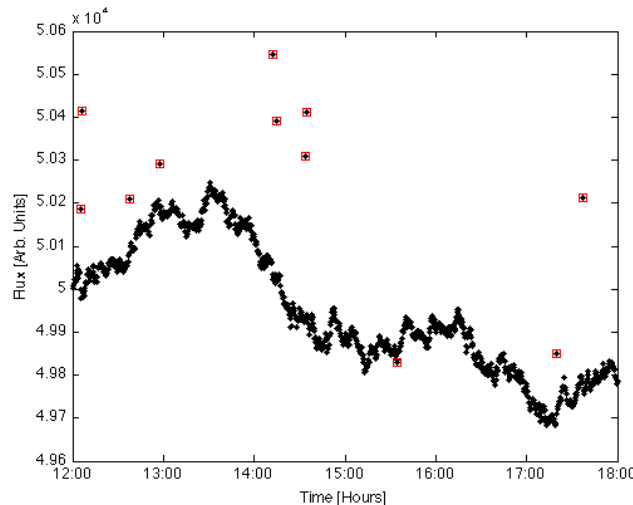
- ❑ Deep learning: A fast expanding dynamic frontier in machine learning
- ❑ Deep learning has developed various *neural network architectures*
 - ❑ Feed-forward neural networks
 - ❑ Convolutional neural networks
 - ❑ Recurrent neural networks
 - ❑ Graph neural networks
 - ❑ Transformer
- ❑ Deep learning has broad applications in computer vision, natural language processing, machine translation, social network analysis, and so on
- ❑ Deep learning has been reshaping a variety of data mining tasks
 - ❑ Ex. classification, clustering, outlier detection, and reinforcement learning



Outlier Analysis

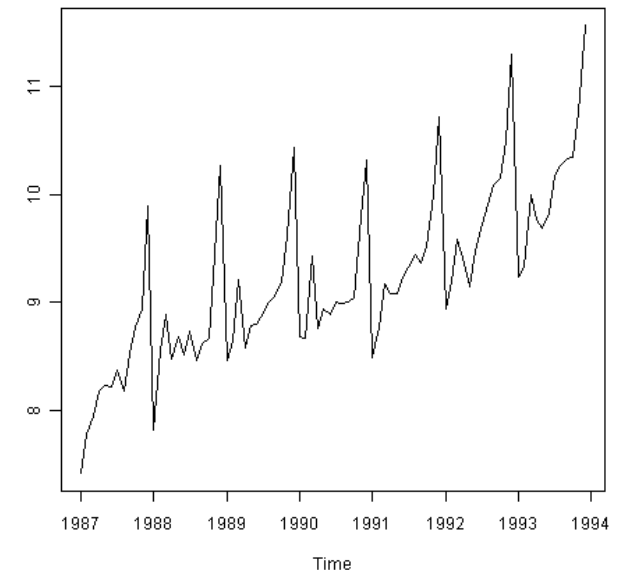
❑ Outlier analysis

- ❑ 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



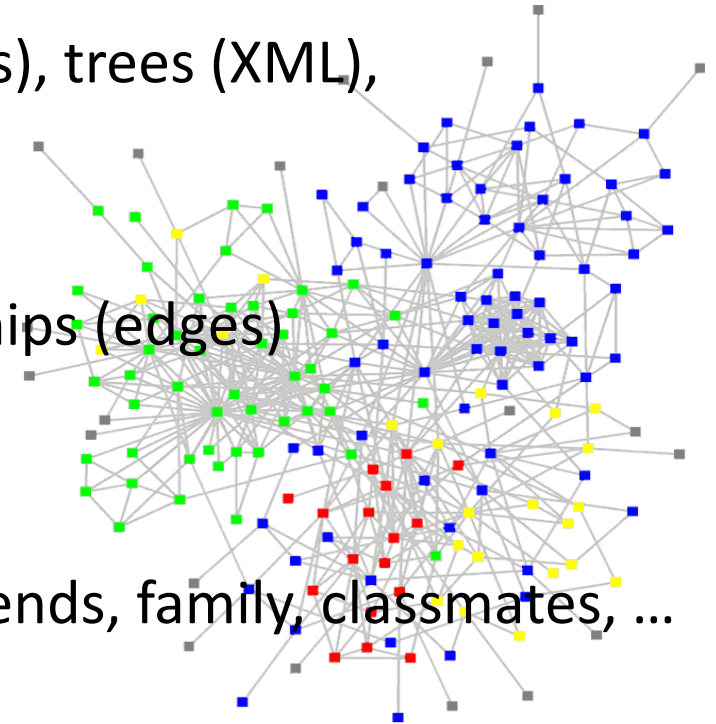
Other Data Mining Functions: Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
 - Trend, time-series, and deviation analysis
 - e.g., regression and value prediction
 - Sequential pattern mining
 - e.g., buy digital camera, then buy large 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



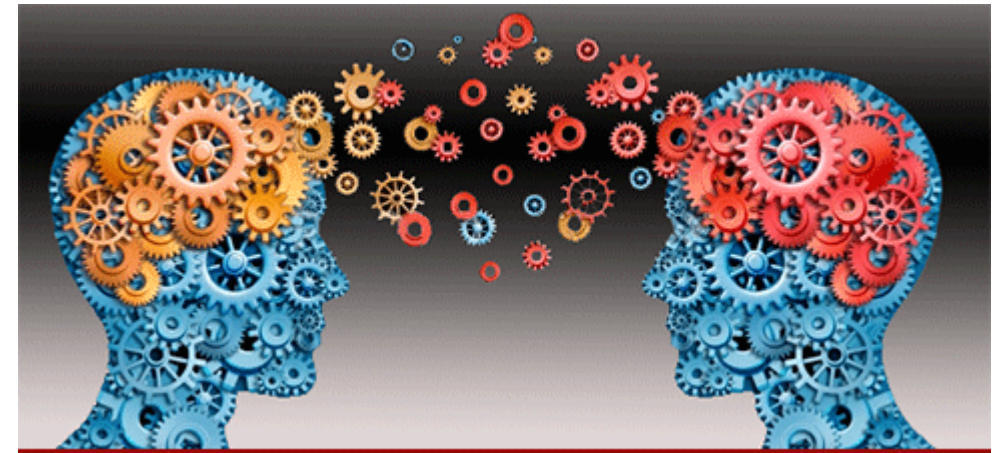
Other Data Mining Functions: Structure and Network Analysis

- 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, ...

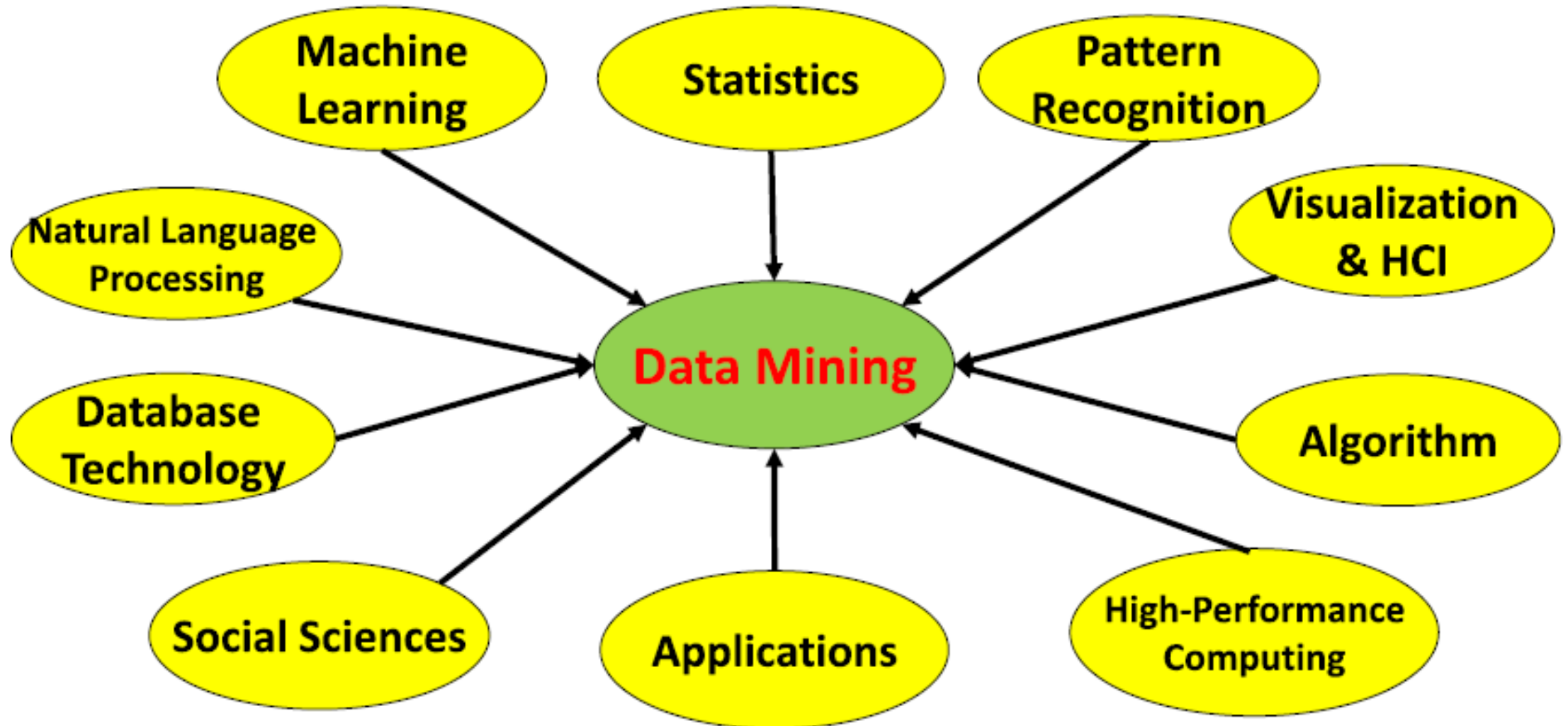


Evaluation of Knowledge

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



Data Mining: Confluence of Multiple Disciplines





Why Confluence of Multiple Disciplines?

- ❑ Tremendous amount of data
 - ❑ Algorithms must be scalable to handle big 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 and information networks
 - ❑ Spatial, spatiotemporal, multimedia, text and Web data
 - ❑ Software programs, scientific simulations
- ❑ New and sophisticated applications

Data Mining and Applications

- ❑ Web page analysis: classification, clustering, ranking
 - ❑ Collaborative analysis & recommender systems
 - ❑ Basket data analysis to targeted marketing
 - ❑ Biological and medical data analysis
 - ❑ Data mining and software engineering
 - ❑ Data mining and text analysis
 - ❑ Data mining and social and information network analysis
 - ❑ Built-in (invisible data mining) functions in Google, Microsoft, LinkedIn, Meta, ...
 - ❑ Major dedicated data mining systems/tools
 - ❑ SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)
- 



Data Mining and Society

- ❑ Data mining technology may benefit society
 - ❑ Ex.: Help scientific discovery, business management, economy recovery, and security protection (*e.g.*, the real-time discovery of intruders and cyberattacks)
- ❑ Need to guard against the misuse of data mining
 - ❑ Data mining also poses the risk of unintentionally disclosing some confidential business or government information and disclosing an individual's personal information
- ❑ Studies on data security in data mining and privacy-preserving data publishing and data mining are important, ongoing research theme
 - ❑ The philosophy is to observe data sensitivity and preserve data security and people's privacy while performing successful data mining
- ❑ These and other related issues will be discussed throughout the book

Summary

- ❑ Data mining: Discovering interesting patterns and knowledge from massive amounts of data
- ❑ A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- ❑ Different data mining method on a wide variety of data
- ❑ Data mining functionalities: summarization, pattern discovery, classification, clustering, deep learning, outlier analysis, trend and outlier analysis, ...
- ❑ Data mining is a confluence of multiple disciplines
- ❑ Data mining has broad applications
- ❑ Promote secure data mining to benefit society