



# Data Mining – Chapter 1

---

Kyuseok Shim

Seoul National University

<http://kdd.snu.ac.kr/~shim>

Extended from the slides of the book "Data Mining:  
Concepts and Techniques (3rd ed.)" provided by Jiawei  
Han, Micheline Kamber, and Jian Pei

# 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

# 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

# Moving Toward Information Age

---

- Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day.
- This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools.
- **Businesses worldwide generate gigantic data sets**, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback.
- This explosively growing, widely available, and gigantic body of data makes our time truly the **data age**.
- Powerful and versatile **tools are needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge**.
- This necessity has led to the birth of data mining.

# Data Mining turns large Data into Knowledge!

---

- A search engine receives hundreds of millions of queries every day.
- Each query can be viewed as a transaction where the user describes her or his information need.
- Some patterns found in user search queries can disclose invaluable knowledge.
  - Google's Flu Trends uses specific search terms as indicators of flu activity.
  - It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms.
  - Using aggregated Google search data, Flu Trends can estimate flu activity up to two weeks faster than traditional systems can.

# Data Mining as the Evolution of Information Technology

---

- The database industry evolved in the development of several critical functionalities.
  - Data collection and database creation
  - Data management (storage, retrieval and transaction processing)
  - Advanced data analysis (data warehousing and data mining)

# Data Mining as the Evolution of Information Technology

---

- Since the 1960s, database technology has evolved from primitive file processing systems to sophisticated and powerful database systems.
- Since the 1970s, database systems progressed from early hierarchical and network database systems to relational database systems.
  - Convenient and flexible data access through query languages, user interfaces, query optimization, and transaction management.
- After the establishment of database systems (since the late 1980s), technology moved toward the development of **advanced database systems**, **data warehousing**, and **data mining** for advanced data analysis.
- The progress of computer hardware technology in the past three decades boosts to the database industry, and it enables a large number of databases repositories to be available data analysis.
- Data can now be stored in many different kinds of databases and information repositories.

# Data Warehouse

---

- It is a **repository of multiple heterogeneous data sources organized under a unified schema at a single site.**
- Data warehouse technology includes
  - Data cleaning
  - Data integration
  - Online analytical processing (OLAP) — analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles.
- Although OLAP tools support multidimensional analysis and decision making, additional data mining tools are required for in-depth analysis
  - Data classification
  - Clustering
  - Outlier/anomaly detection
  - Characterization of changes in data over time.



# Data Mining as the Evolution of Information Technology

---

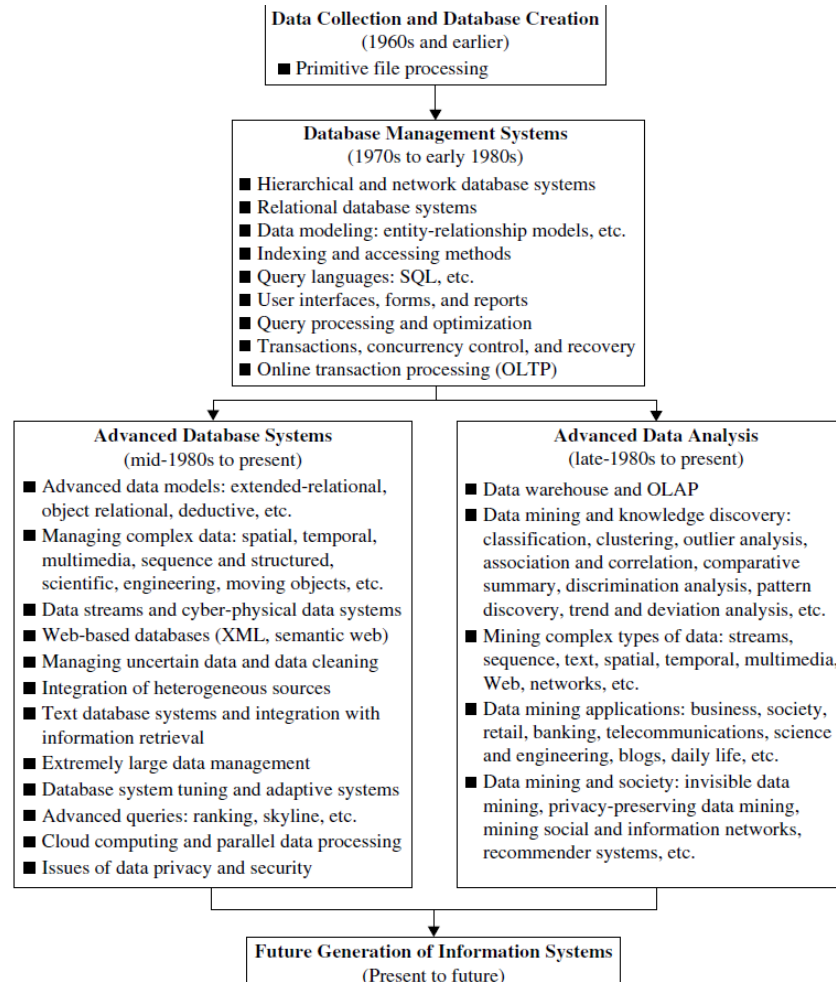
- During the 1990s, the World Wide Web and web-based databases (e.g., XML databases) began to appear.
- Internet-based global information bases, such as the WWW and various kinds of interconnected, heterogeneous databases, have emerged and play a vital role in the information industry.
- Efficient analysis of data from such different forms of data by data integration, data mining, and network analysis is a challenging task.

# Evolution of Database Technology (Recap.)

---

- 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

# The Evolution of Database Technology




# 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

# 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 (non-trivial, implicit, previously unknown and potentially useful) 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



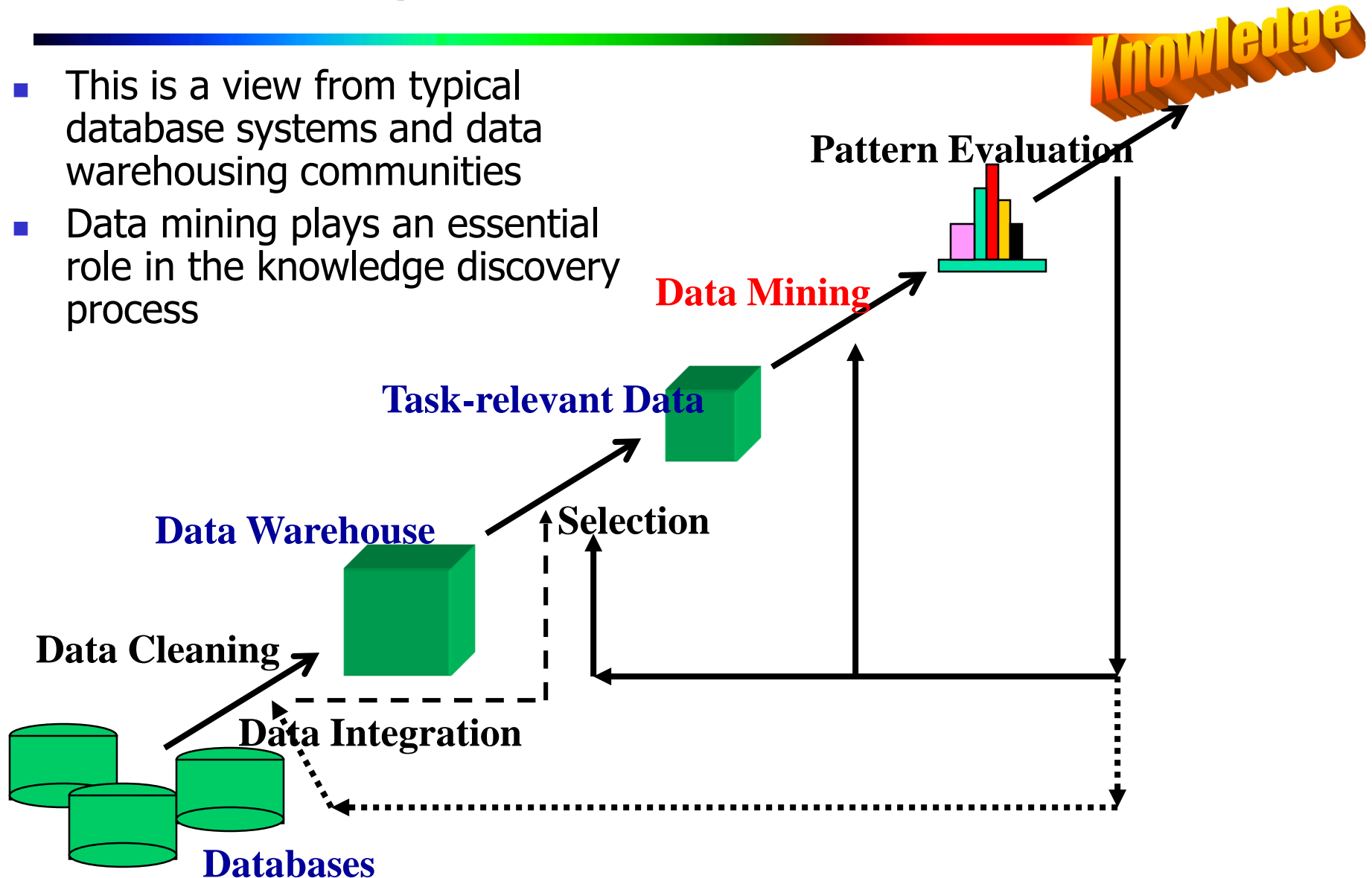
# Data Mining – a Step in Process of Knowledge Discovery

---

- Data cleaning
  - Remove noise and inconsistent data
- Data integration
  - Combine multiple data sources
- Data selection
  - Retrieve relevant data from the database
- Data transformation
  - Transform and consolidate data into forms appropriate for mining
- Data mining
  - Apply intelligent methods to extract data patterns
- Pattern evaluation
  - Identify the truly interesting patterns based on interestingness measures
- Knowledge presentation
  - Present mined knowledge to users

# 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



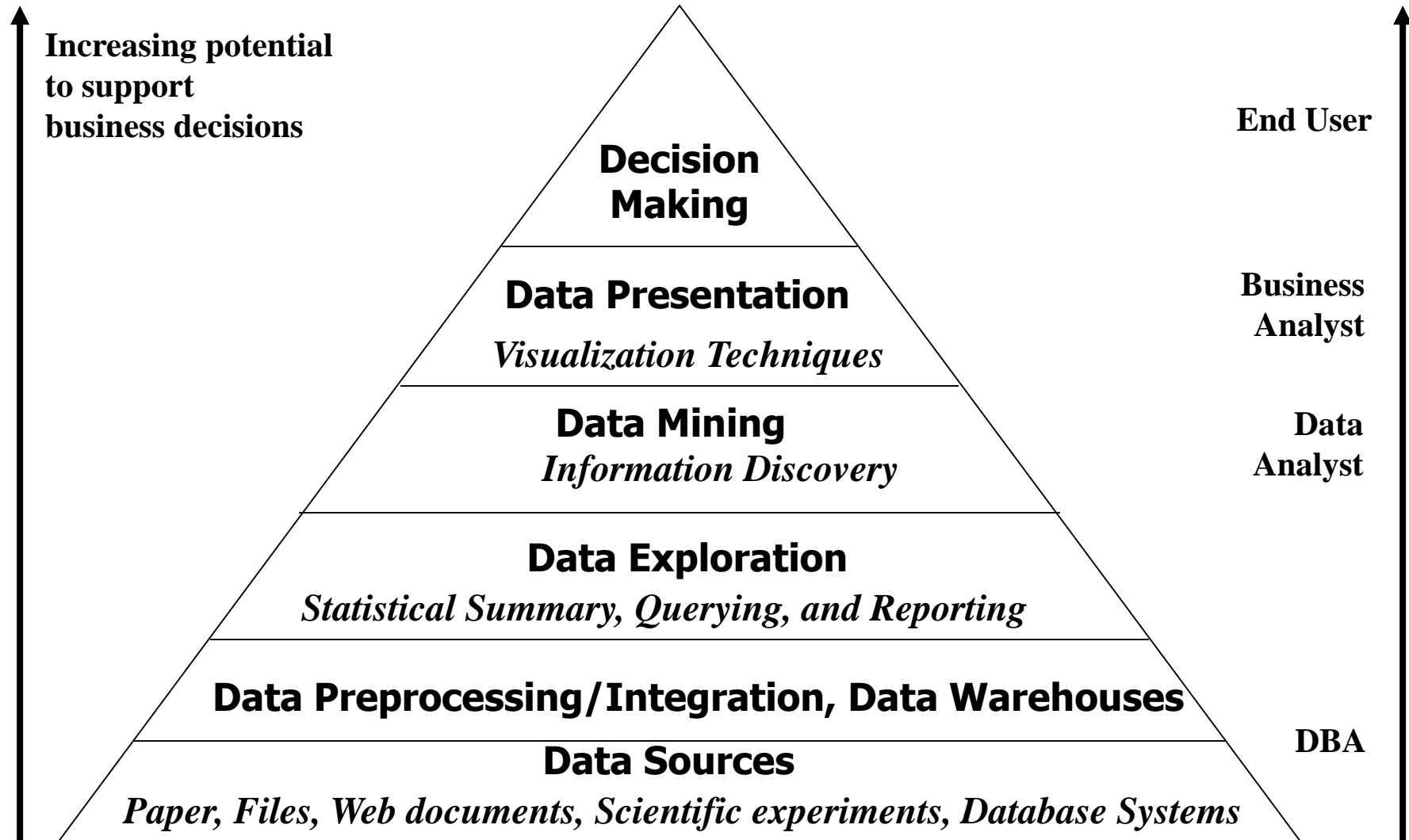


# Example: A Web Mining Framework

---

- 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

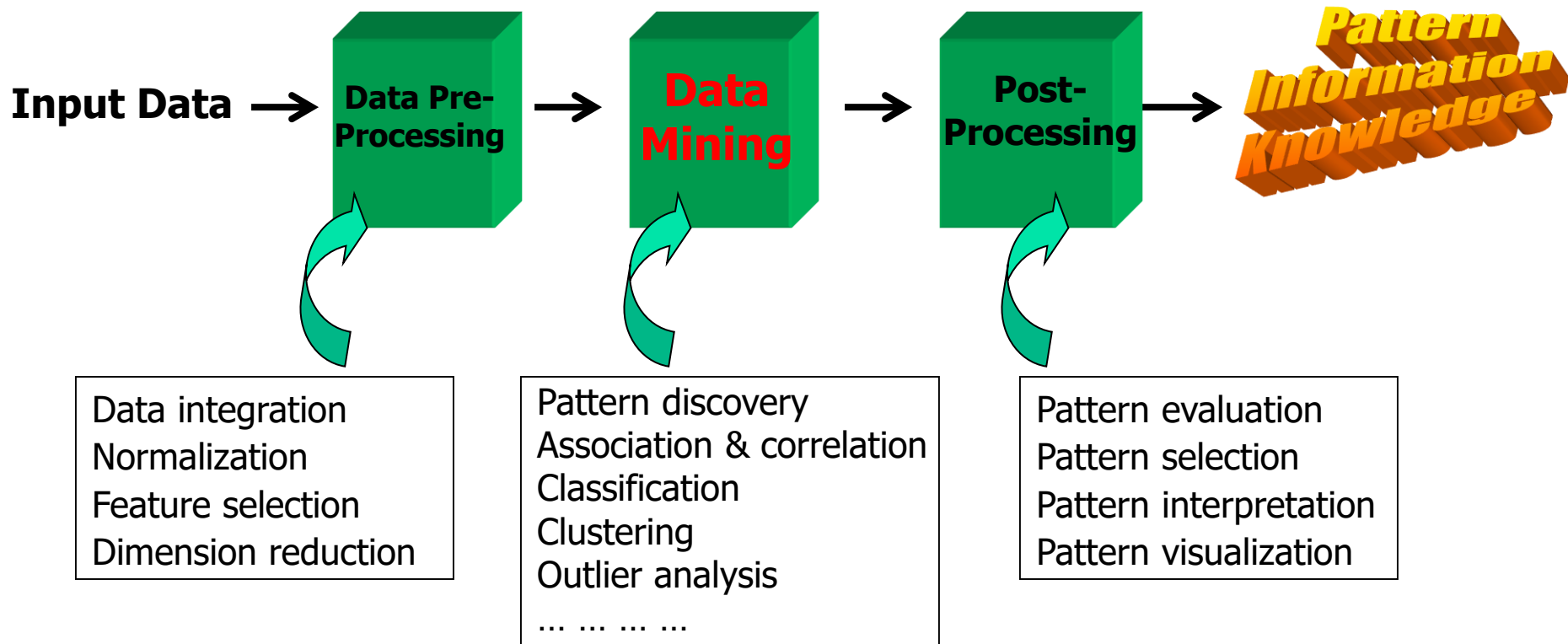


# Example: Mining vs. Data Exploration

---

- Business intelligence view
  - Warehouse, data cube, reporting but not much mining
- Business objects vs. data mining tools
- Supply chain example: tools
- Data presentation
- Exploration

# KDD Process: A Typical View from ML and Statistics



- 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

---

## ■ Data to be mined

- 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

## ■ Knowledge to be mined (or: Data mining functions)

- Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
- Descriptive vs. predictive data mining
- Multiple/integrated functions and mining at multiple levels

## ■ Techniques utilized


- Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

## ■ Applications adapted

- 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?

---

- Data mining can be applied to any kind of data as long as the data are meaningful for a target application.
- 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

# Database Management System (DBMS)

---

- It consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- The software programs provide mechanisms for
  - Defining database structures and data storage
  - Specifying and managing concurrent, shared, or distributed data access
  - Ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.

# Relational Database

---

- It is a collection of tables.
- Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
- A semantic data model, such as an entity-relationship (ER) data model, is often constructed to represent the database as a set of entities and their relationships.

# A Example of Relational Schema for AllElectronics

---

- customer (cust ID, name, address, age, occupation, . . . )
- item (item ID, brand, category, type, price, place made, supplier, . . .)
- employee (empl ID, name, category, group, salary, commission, . . . )
- branch (branch ID, name, address, . . . )
- purchases (trans ID, cust ID, empl ID, date, time, method paid, amount)
- items\_sold (trans ID, item ID, qty)
- works\_at (empl ID, branch ID)

# Relational Databases

---

- Relational databases are one of the most commonly available repositories.
  - A major data form in the study of data mining.
- Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces.
- A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing.
- A query allows retrieval of specified subsets of the data.

# Relational Databases

---

- Suppose that you want to analyze the AllElectronics data.
- Through the use of relational queries, you can ask things like, “Show me a list of all items that were sold in the last quarter.”
- Relational languages also use aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum).
- Using aggregates allows you to ask
  - Total sales of the last month, grouped by branch
  - How many sales transactions occurred in the month of December
  - Which salesperson had the highest sales

# Mining Relational Databases

---

- When mining relational databases, we can search for trends or data patterns.
  - e.g., analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information.
- We can also detect deviations—that is, items with sales that are far from those expected in comparison with the previous year.
- Such deviations can then be further investigated.
  - e.g, we may discover that there has been a change in packaging of an item or a significant increase in price.

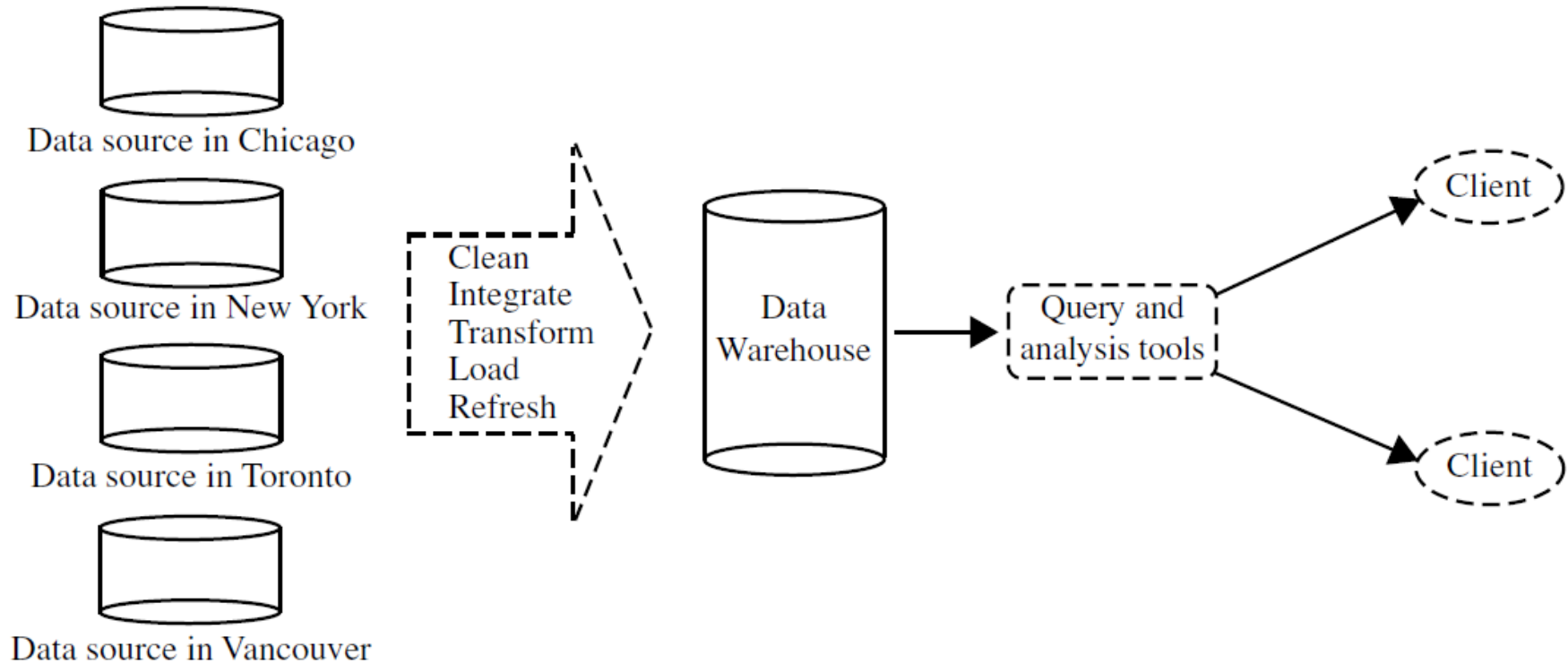
# Data Warehouses

---

- Suppose that AllElectronics is an international company with branches around the world.
- Each branch has its own set of databases.
- We want to analyze the company's sales per itemtype per branch for the third quarter.
- The relevant data are spread out over several databases physically located at numerous sites.
- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.



# Typical framework of a data warehouse for AllElectronics



# Data Warehouses

---

- To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity).
- The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized.
  - e.g., rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region.
- A data warehouse is modeled by a multidimensional data structure, called a **data cube**.
  - Each **dimension** corresponds to an attribute or a set of attributes in the schema, and each **cell** stores the value of some aggregate measure such as count.
- A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

# An Example of Data Cube for AllElectronics

---

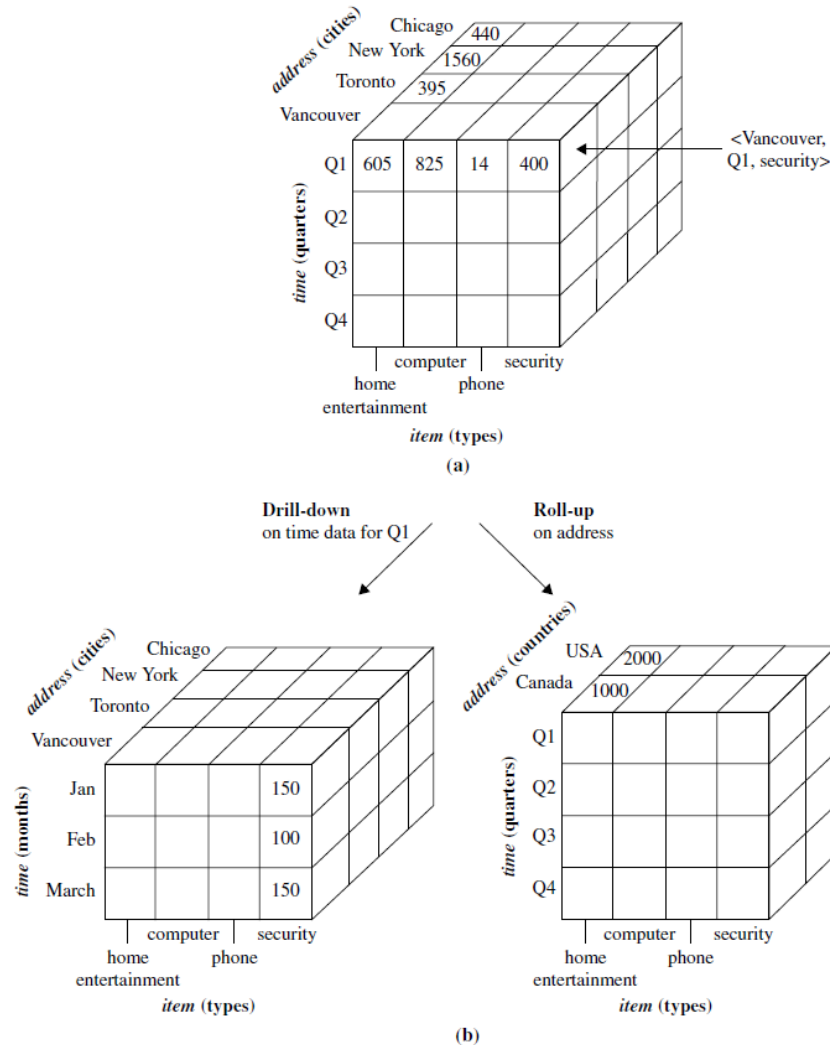
- The cube has three dimensions
  - Address (with city values Chicago, New York, Toronto, Vancouver)
  - Time (with quarter values Q1, Q2, Q3, Q4)
  - Item (with itemtype values home entertainment, computer, phone, security).
- In each cell of the cube has sales amount (in thousands).
  - e.g., in cell <Vancouver, Q1, security>., \$400, 000 is stored and represents the total sales for the first quarter, for the items of security systems in Vancouver
- Cubes stores the aggregate values obtained using different SQL group-bys.

# Data Warehouses

---

- Support for OLAP by multidimensional data views and the precomputation of summarized data, .
- Online analytical processing operations use background knowledge regarding the domain of the data to allow the presentation of data at different levels of abstraction.
- Such operations accommodate different user viewpoints.
  - Drill down on sales data summarized by quarter to see data summarized by month.
  - Roll up on sales data summarized by city to view data summarized by country.

# A data cube for summarized sales data of AllElectronics



# Multidimensional Data Mining

---

- Performs data mining in multidimensional space in an OLAP style.
- Allows the exploration of multiple combinations of dimensions at varying levels of granularity in data mining
- Has greater potential for discovering interesting patterns representing knowledge.

# Transactional Database

---

- Each record in a **transactional database** captures a transaction
- A transaction includes
  - a unique transaction identity number (trans ID)
  - a list of the items making up the transaction, such as the items purchased in the transaction.
- A transactional database have additional tables which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.

# An Example of a Transactional Database for AllElectronics

- Transactions can be stored in a table, with one record per transaction.
- From the relational database point of view, the sales table is a **nested relation** because the attribute list of item IDs contains a set of items.
- Because most relational database systems do not support nested relational structures, the transactional database is usually either stored in a flat file in a format or unfolded into a standard relation.

<i>trans_ID</i>	<i>list_of_item_IDs</i>
T100	I1, I3, I8, I16
T200	I2, I8
...	...



# An Example of a Transactional Database for AllElectronics

---

- You may ask, “Which items sold well together?”
- We can bundle groups of items together for boosting sales.
  - e.g., If printers are commonly purchased together with computers, Offer certain printers at a steep discount to customers buying selected computers, in the hopes of selling more computers.
- A traditional database system is not able to perform market basket data analysis.
- Data mining on transactional data can help
  - e.g., frequent itemsets - sets of items that are frequently sold together.

# Other Kinds of Data

---

- Besides relational database data, data warehouse data, and transaction data, there are many other kinds of data that have versatile forms and structures and rather different semantic meanings.
- Such kinds of data can be seen in many applications:
  - time-related sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data)
  - data streams (e.g., video surveillance and sensor data, which are continuously transmitted),
  - spatial data (e.g., maps)
  - engineering design data (e.g., the design of buildings, system components, or integrated circuits),
  - multimedia data (including text, image, video, and audio data)
  - graph and networked data (e.g., social and information networks)
  - the Web (a huge, widely distributed information repository made available by the Internet)

# Other Kinds of Data

---

- Temporal data
  - We can mine banking data for changing trends, which may aid in the scheduling of bank tellers according to the volume of customer traffic.
- Stock exchange data
  - We can mine to uncover trends that could help you plan investment strategies (e.g., the best time to purchase AllElectronics stock).
- computer network data streams
  - We can mine to detect intrusions based on the anomaly of message flows, which may be discovered by clustering, dynamic construction of stream models or by comparing the current frequent patterns with those at a previous time.
- Spatial data
  - We may look for patterns that describe changes in metropolitan poverty rates based on city distances from major highways.

# Other Kinds of Data

---

- Text data
  - We can identify the evolution of hot topics in the field.
  - By mining user comments on products (which are often submitted as short text messages), we can assess customer sentiments and understand how well a product is embraced by a market.
- Multimedia data
  - We can mine images to identify objects and classify them by assigning semantic labels or tags.
- Video data
  - We can detect video sequences corresponding to goals.
- Web data
  - We can learn about the distribution of information on the WWW in general, characterize and classify web pages, and uncover web dynamics and the association and other relationships among different web pages, users, communities, and web-based activities.


# Other Kinds of Data

---

- In many applications, multiple types of data are present.
  - In web mining, text data and multimedia data (e.g., pictures and videos) on web pages, web graph data, and map data on some web sites.
  - In bioinformatics, genomic sequences, biological networks, and 3-D spatial structures of genomes.
- Mining multiple data sources of complex data leads to fruitful findings due to the mutual enhancement and consolidation of such multiple sources.
- On the other hand, it is also challenging.

# 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 Functionalities

---

- Characterization and discrimination
- Mining of frequent patterns, associations, and correlations
- Classification and regression
- Clustering analysis
- Outlier analysis

# Data Mining Functionalities

---

- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.
- Classified into two categories
  - Descriptive mining – characterize properties of the data in a target data set.
  - Predictive mining - perform induction on the current data in order to make predictions.



# Characterization and Discrimination

---

- Data entries can be associated with classes or concepts.
  - e.g.) classes of items for sale include computers and printers, and concepts of customers include bigSpenders and budgetSpenders.
- Useful to describe individual classes and concepts in summarized, concise, and yet precise terms.
- These descriptions can be derived using
  - Data characterization, by summarizing the data of the class under study (often called the target class) in general terms
  - Data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes)

# Data Characterization

---

- A summarization of the general characteristics or features of a target class of data
- The data corresponding to the user-specified class are typically collected by a query.
  - e.g.) To study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.
- There are several methods for effective data summarization and characterization.
  - Simple data summaries based on statistical measures and plots
  - The data cube-based OLAP roll-up operation to perform user-controlled data summarization along a specified dimension.
  - An attribute-oriented induction technique to perform data generalization and characterization without user interaction.

# An Example of Data Characterization

---

- Summarize the characteristics of customers who spend more than \$5000 a year at AllElectronics.
- The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.
- The data mining system should allow the customer relationship manager to drill down on any dimension, such as on **occupation** to view these customers according to their type of employment.

# Data Discrimination

---

- A comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.
  - e.g.) Compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.
- Discrimination descriptions expressed in the form of rules are referred to as **discriminant rules**.

# An Example of Data Discrimination

---

- Compare two groups of customers buying computer products
  - regularly (more than twice a month)
  - rarely (e.g., less than three times a year).
- The resulting description provides a general comparative profile of these customers
  - 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education,
  - 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.
- Drilling down on a dimension like occupation, or adding a new dimension like income level, may help to find even more discriminative features between the two classes.

# Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns are patterns that occur frequently in data.
- There are many kinds of frequent patterns
  - frequent itemsets
  - frequent subsequences (also known as sequential patterns)
  - frequent substructures.
- A frequent itemset typically refers to a set of items that often appear together in a transactional data set
  - e.g.) milk and bread, which are frequently bought together in grocery stores by many customers.
  - A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.
- A substructure can refer to different structural forms (e.g., graphs, trees, or lattices).
  - If a substructure occurs frequently, it is called a (frequent) structured pattern.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

# Association Rules



- 데이터 상호간의 연관 규칙을 찾아내는 기술
- '{라면, 우유}->{커피}'
  - 라면과 우유를 산 사람은 커피도 같이 산다
  - 지지도 (support)
    - 전체 소비자 중에서 그 규칙을 구성하는 물품을 구매한 소비자의 비율
    - 50% - 4명 중 **라면**, **우유**, **커피**를 구매한 사람은 2명
  - 신뢰도 (confidence)
    - 규칙의 왼쪽에 있는 물품을 산 소비자 중에서 오른쪽에 있는 물품들을 산 소비자의 비율
    - 66.7% - **라면**과 **우유**를 산 사람들은 3명인데 그 중에서 **커피**를 산 사람은 2명

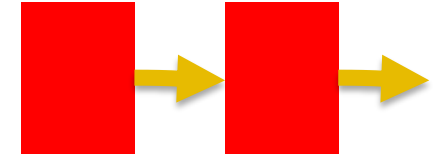
# An Example of Association Rules

- 고장이 발견된 부품들을 통해 아직 발견되지 않는 부품의 고장을 예측
  - 예) {오일, 냉각수}  $\Rightarrow$  {엔진}
    - 오일과 냉각수에 이상이 발견되었다면 엔진에 이상이 발견되지 않았더라도 꼼꼼한 점검이 필요
- 부품 간의 상관 관계를 파악하고 부품 성능 개선에 활용





# Sequential Patterns



빈번한 패턴



# An Example of Sequential Patterns

## ■ 의료 데이터마이닝

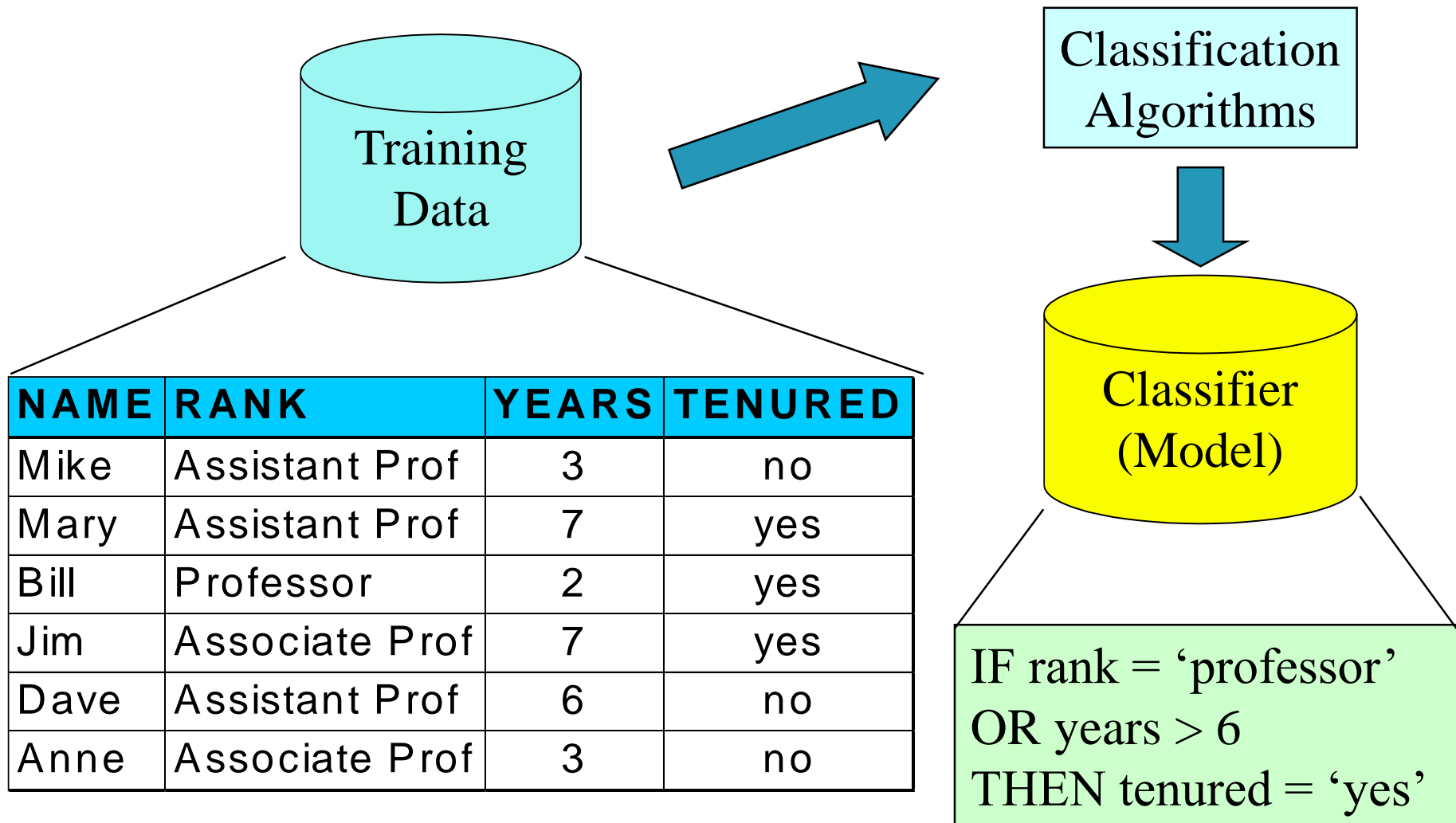
- 환자의 과거 테스트, 진단, 치료, 처방등에 관한 히스토리 데이터로부터 당뇨병에 걸리는 환자의 패턴을 이용하여 앞으로 당뇨병 또는 치매에 걸릴 확률이 높은 환자를 알아냄
- 히스토리 정보를 이용하여 부작용이 발생할 환자를 예측함
- 입원 환자, 신생아, 마취 사망 가능성 예측 : 기존의 입원 환자/신생아 사례(항생제 등)로부터 패턴을 만들고, 이후 입원 환자/신생아 경우에 예상하는 자료에 사용
  - 연관규칙, 순차패턴 모두 적용 가능

# Classification and Regression for Predictive Analysis

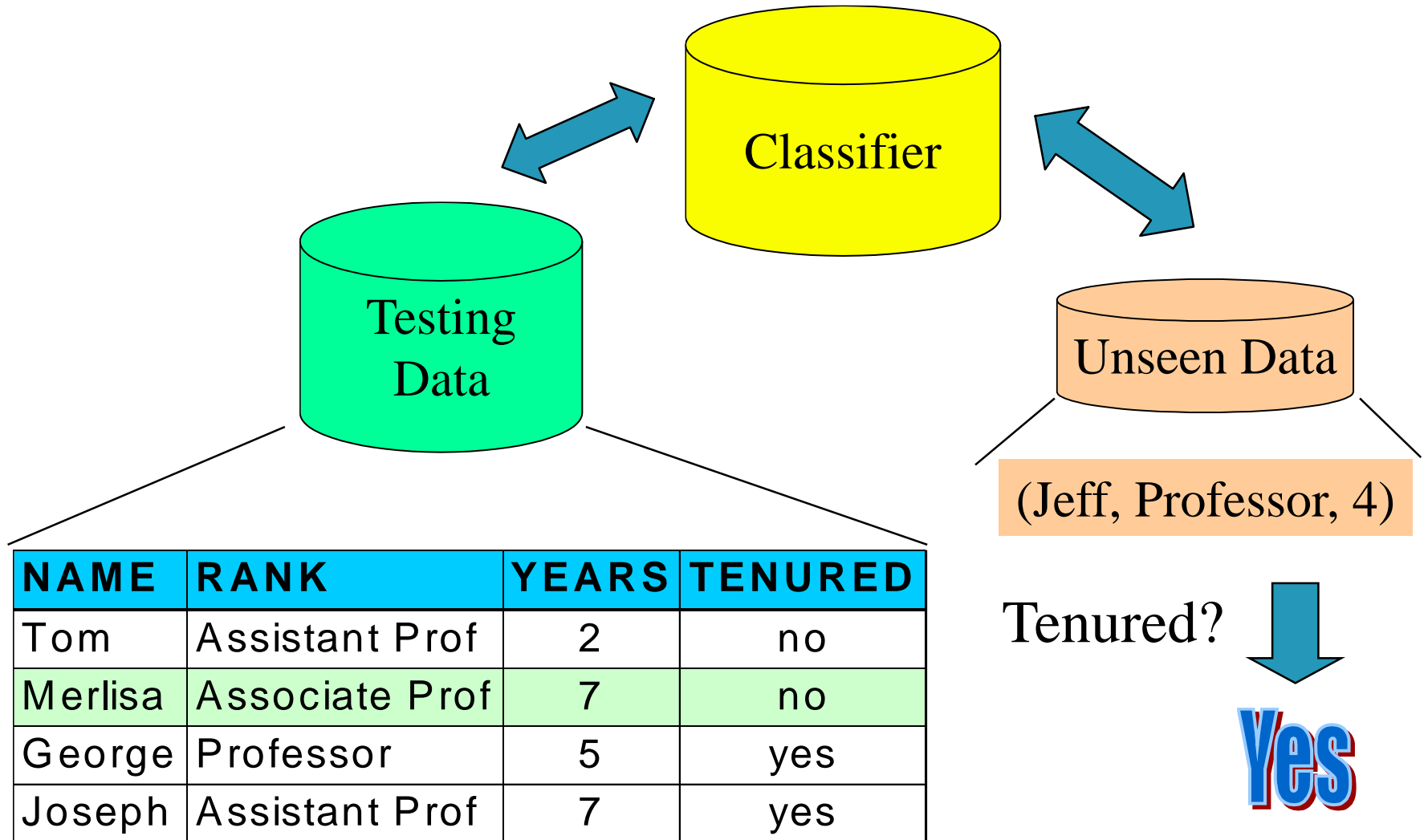
---

- **Classification** is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts.
- The model are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known).
- The model is used to predict the class label of objects for which the class label is unknown.
- The derived model may be represented in various forms
  - classification rules (i.e., IF-THEN rules)
  - decision trees
  - neural networks
  - naive Bayesian,
  - support vector machines
  - k-nearest-neighbor

# Process (1): Model Construction



# Process (2): Using the Model in Prediction



# Classification/Regression

---

- Classification
  - Predicts categorical (discrete, unordered) labels,
- Regression
  - Predicts continuous values
- Classification and regression may need to be preceded by relevance analysis
  - Identify attributes that are significantly relevant to the classification and regression process.
  - Selected such attributes.
  - Excluded other attributes which are irrelevant.

# An Example of Classification

---

- You want to classify a large set of items in the store, based on three kinds of responses to a sales campaign
  - good response, mild response and no response.
- You want to derive a model for each of these three classes based on the descriptive features of the items, such as price, brand, place made, type, and category.
- Suppose that the resulting classification is expressed as a decision tree.
- The decision tree may identify price as being the single factor that best distinguishes the three classes.
- The tree may reveal that, in addition to price, other features that help to further distinguish objects of each class from one another include brand and place made.
- Such a decision tree may help you understand the impact of the given sales campaign and design a more effective campaign in the future.

# An Example of Regression

---

- You want to predict the amount of revenue that each item will generate during an upcoming sale, based on the previous sales data.



# Cluster Analysis

---

- Clustering analyzes data objects without consulting class labels.
- In many cases, class-labeled data may simply not exist at the beginning.
- Clustering can be used to generate class labels for a group of data.
- The objects are clustered or grouped based on the principle of **maximizing the intraclass similarity and minimizing the interclass similarity**.
  - Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.
- Each cluster so formed can be viewed as a class of objects, from which rules can be derived.
- Clustering can also facilitate **taxonomy formation**, that is, the organization of observations into a hierarchy of classes that group similar events together.

# An Example of Clustering

Click log database

User	News <sub>1</sub>	News <sub>2</sub>	News <sub>3</sub>	News <sub>4</sub>	News <sub>5</sub>	News <sub>6</sub>
u <sub>2</sub>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
u <sub>5</sub>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
u <sub>6</sub>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	
u <sub>1</sub>				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
u <sub>3</sub>				<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
u <sub>4</sub>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Cluster<sub>1</sub>  
News<sub>1</sub>, News<sub>2</sub>,  
News<sub>3</sub>

Cluster<sub>2</sub>  
News<sub>4</sub>, News<sub>5</sub>,  
News<sub>6</sub>

# An Example of Clustering

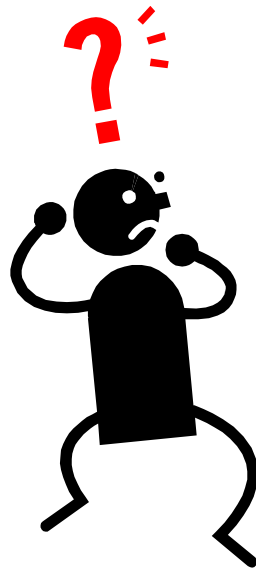
## Clustering Results

Cluster<sub>1</sub>

News<sub>1</sub>, News<sub>2</sub>,  
News<sub>3</sub>

Cluster<sub>2</sub>

News<sub>4</sub>, News<sub>5</sub>,  
News<sub>6</sub>



Recommend  
**News<sub>5</sub>**  
for this new user

A new user

He clicked News<sub>4</sub> and News<sub>6</sub>

# Outlier Analysis

---

- Objects that do not comply with the general behavior or model of the data.
- Many data mining methods discard outliers as noise or exceptions.
- However, in some applications (e.g., fraud detection), the rare events can be more interesting than the more regularly occurring ones.
- Outliers may be detected using
  - statistical tests that assume a distribution or probability model for the data
  - distance measures where objects that are remote from any other cluster are considered outliers
  - density-based methods assuming that outliers are in a local region

# An Example of Outlier Analysis

---

- Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.
- Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.

# 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  $\rightarrow$  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?

# 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
  - ...

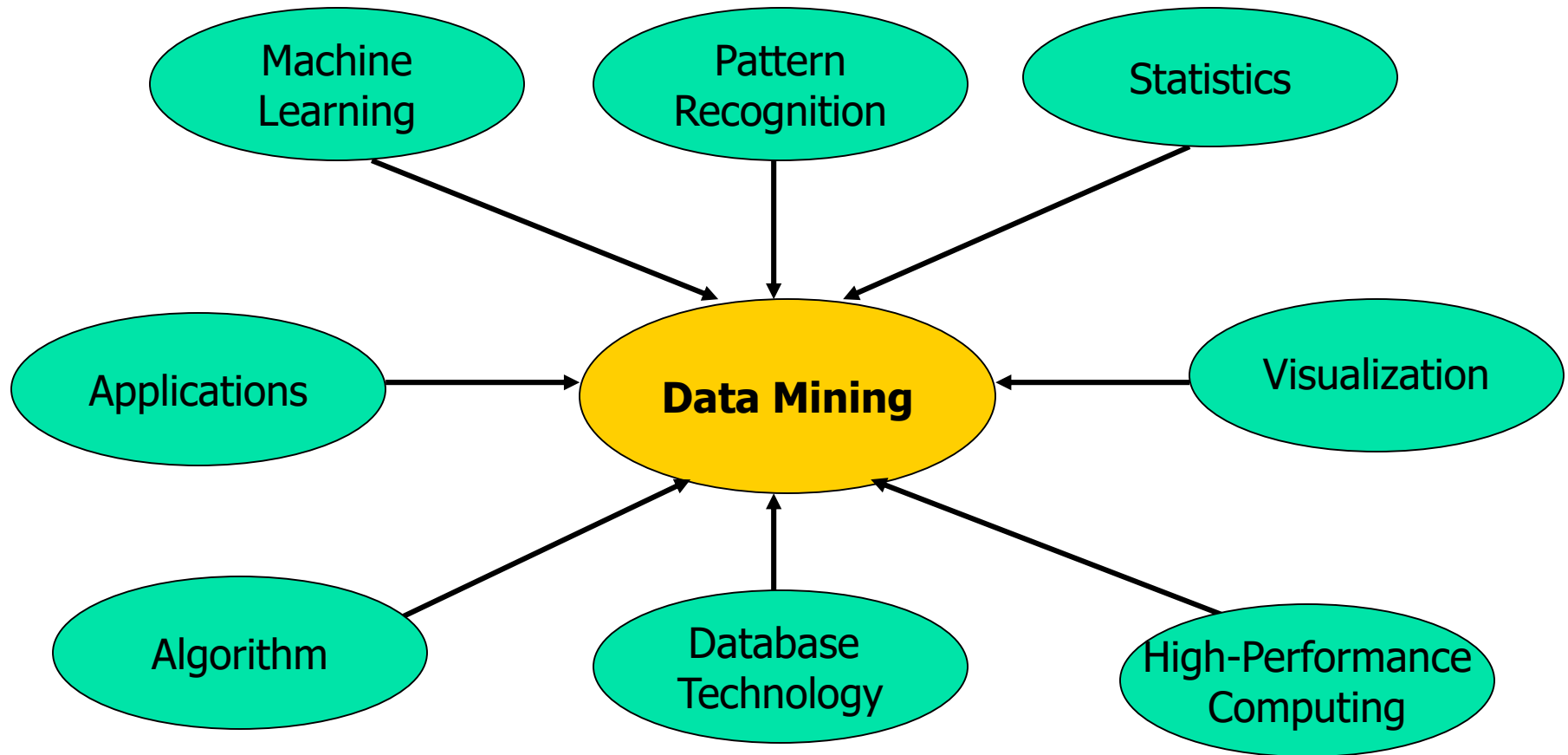
# Chapter 1. Introduction

---

- 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

---



# Statistics



- Statistics studies the collection, analysis, interpretation or explanation, and presentation of data.
- A statistical model is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions.
- Statistical models are widely used to model data and data classes.
  - In data characterization and classification, statistical models of target classes can be built.
- Alternatively, data mining tasks can be built on top of statistical models.
  - Use statistics to model noise and missing data values.
  - Then, when mining patterns in a large data set, the data mining process can use the model to help identify and handle noisy or missing values in the data.



# Statistics

---

- Statistical methods can also be used to verify data mining results.
  - e.g.) After a classification or prediction model is mined, the model should be verified by statistical hypothesis testing.
- A statistical hypothesis test (sometimes called confirmatory data analysis) makes statistical decisions using experimental data.
- A result is called **statistically significant** if it is **unlikely to have occurred by chance**.
- If the classification or prediction model holds true, the descriptive statistics of the model increases the soundness of the model.

# Statistics



- Applying statistical methods in data mining is far from trivial.
- Often, a serious challenge is **how to scale up a statistical method over a large data set**.
- Many statistical methods have high complexity in computation.
- When such methods are applied on large data sets that are also distributed on multiple logical or physical sites, algorithms should be carefully designed and tuned to reduce the computational cost.
- This challenge becomes even tougher for online applications, such as online query suggestions in search engines, where data mining is required to continuously handle fast, real-time data streams.

# Machine Learning

---

- Machine learning investigates how computers can learn (or improve their performance) based on data.
- A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.
- e.g., automatically recognize handwritten postal codes on mail after learning from a set of examples.
- Classic problems in machine learning highly related to data mining.
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning
  - Active learning

# Supervised Learning

---

- A synonym for classification.
- The supervision in the learning comes from the labeled examples in the training data set.
- In the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples.

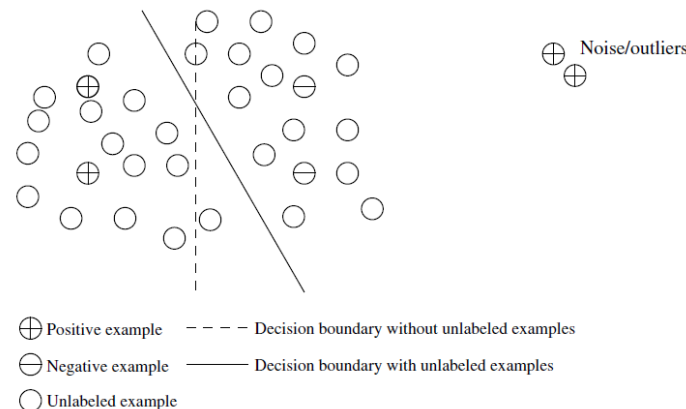
# Unsupervised Learning

---

- A synonym for clustering.
- The learning process is unsupervised since the input examples are not class labeled.
- Typically, we may use clustering to discover classes within the data.
- An example
  - Take, as input, a set of images of handwritten digits.
  - Suppose that it finds 10 clusters of data.
  - These clusters may correspond to the 10 distinct digits of 0 to 9, respectively.

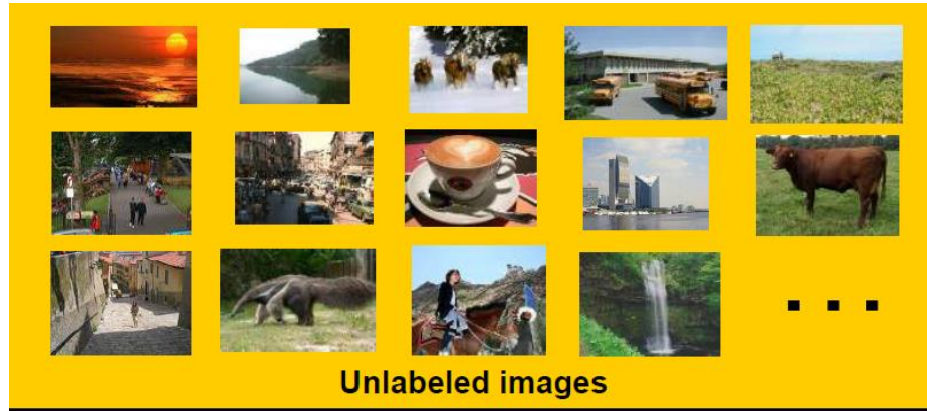
# Semi-supervised Learning

- Use of both labeled and unlabeled examples when learning a model.
- An Example
  - Labeled examples are used to learn class models and unlabeled examples are used to refine the boundaries between classes.



- If we do not consider the unlabeled examples, the dashed line is the decision boundary.
- Using the unlabeled examples, we can refine the decision boundary to the solid line.
- Moreover, we can detect that the two positive examples at the top right corner, though labeled, are likely noise or outliers.

# An Example of Semi-supervised Learning



Andrew Ng

Testing:  
What is this?



# Active Learning



- A machine learning approach that lets users play an active role in the learning process.
- Ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples.
- The goal is to optimize the model quality by actively acquiring knowledge from human users.



# Data Mining vs. Machine Learning

---

- For classification and clustering tasks, machine learning research often focuses on the accuracy of the model.
- In addition to accuracy, data mining research places **strong emphasis on the efficiency and scalability of mining methods on large data sets**, as well as on ways to handle complex types of data and explore new, alternative methods.

# Database Systems vs. Data Mining

---

- Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users.
- Database systems are well known for their high scalability in processing very large and relatively structured data sets.
- Many data mining tasks need to handle large data sets or even real-time, fast streaming data.
- Data mining can make good use of scalable database technologies to achieve high efficiency and scalability on large data sets.
- Moreover, data mining tasks can be used to extend the capability of existing database systems to satisfy advanced users' sophisticated data analysis requirements.

# DataWarehouses

---

- Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities.
- A data warehouse integrates data originating from multiple sources and various timeframes.
- It consolidates data in multidimensional space to form partially materialized data cubes.
- The data cube model not only facilitates OLAP in multidimensional databases but also promotes multidimensional data mining.

# Information Retrieval

---

- Information retrieval (IR) is the science of searching for documents or information in documents.
- Documents can be text or multimedia, and may reside on the Web.
- The differences between traditional information retrieval and database systems are twofold:
- Information retrieval assumes that (1) the data under search are unstructured; and (2) the queries are formed mainly by keywords.
- The typical approaches in information retrieval adopt probabilistic models.
  - A text document can be regarded as a bag of words, that is, a multiset of words appearing in the document.
  - The document's language model is the probability density function that generates the bag of words in the document.
  - The similarity between two documents can be measured by the similarity between their corresponding language models.

# Information Retrieval

---

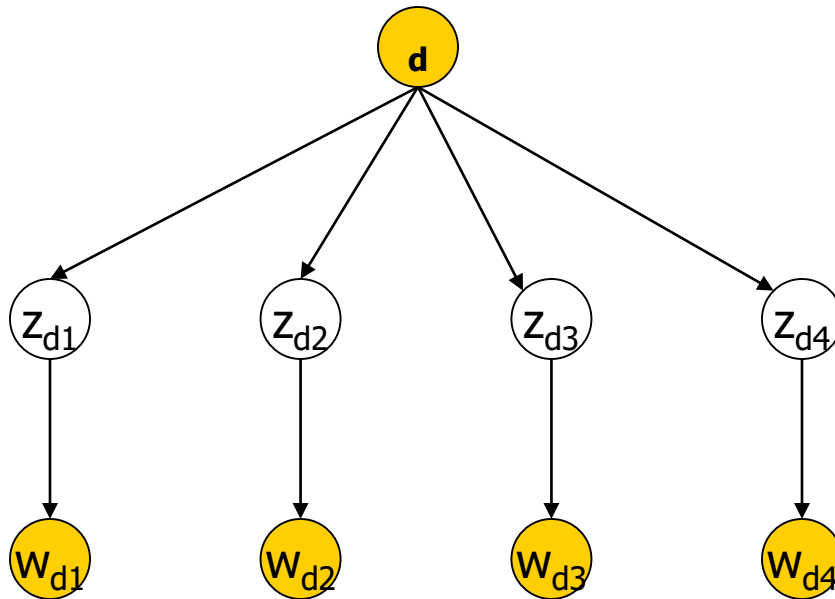
- A topic in a set of text documents can be modeled as a probability distribution over the vocabulary, which is called a topic model.
- A text document, which may involve one or multiple topics, can be regarded as a mixture of multiple topic models.
- By integrating information retrieval models and data mining techniques, we can find the major topics in a collection of documents and, for each document in the collection, the major topics involved.
- Increasingly large amounts of text and multimedia data have been accumulated and made available online due to the fast growth of the Web and applications such as digital libraries, digital governments, and health care information systems.
- Their effective search and analysis have raised many challenging issues in data mining.
- Therefore, text mining and multimedia data mining, integrated with information retrieval methods, have become increasingly important.

# Probabilistic Latent Semantic Indexing (PLSI)

---

- A generative model
- Models each word in a document as a sample from a mixture model
- Each word is generated from a single topic, different words in the document may be generated from different topics
- Each document is represented as a list of mixing proportions for the mixture components

# The PLSI Model: A Generative Model



Probabilistic Latent Semantic  
Indexing (PLSI) Model

For each word of document  $d$  in the training set,

- Choose a topic  $z$  according to a multinomial conditioned on the index  $d$
- Generate the word by drawing from a multinomial conditioned on  $z$

In PLSI, documents can have multiple topics

# Generative Model Illustration

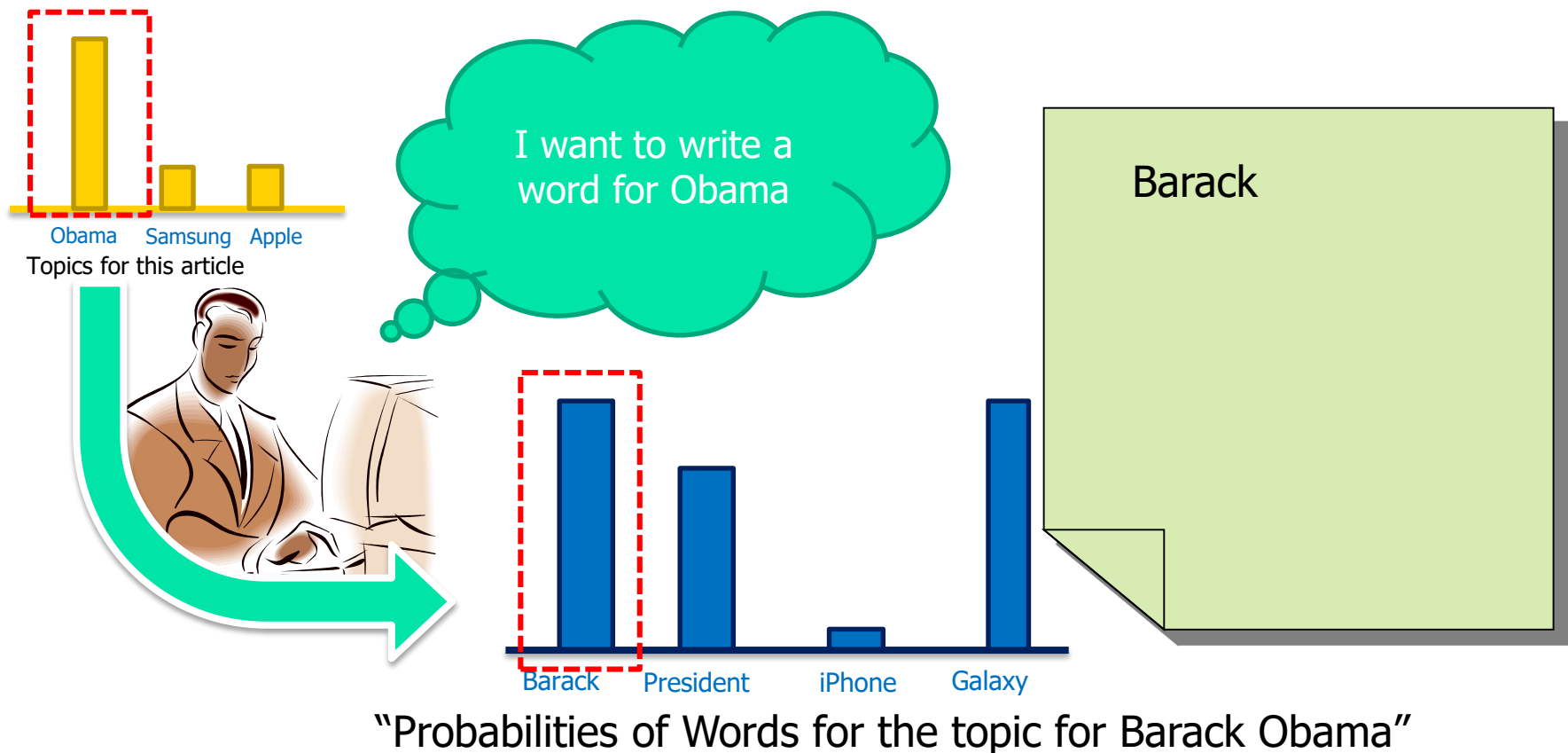
The generative model of PSLI assumes that  
**an article is written according to the following process**





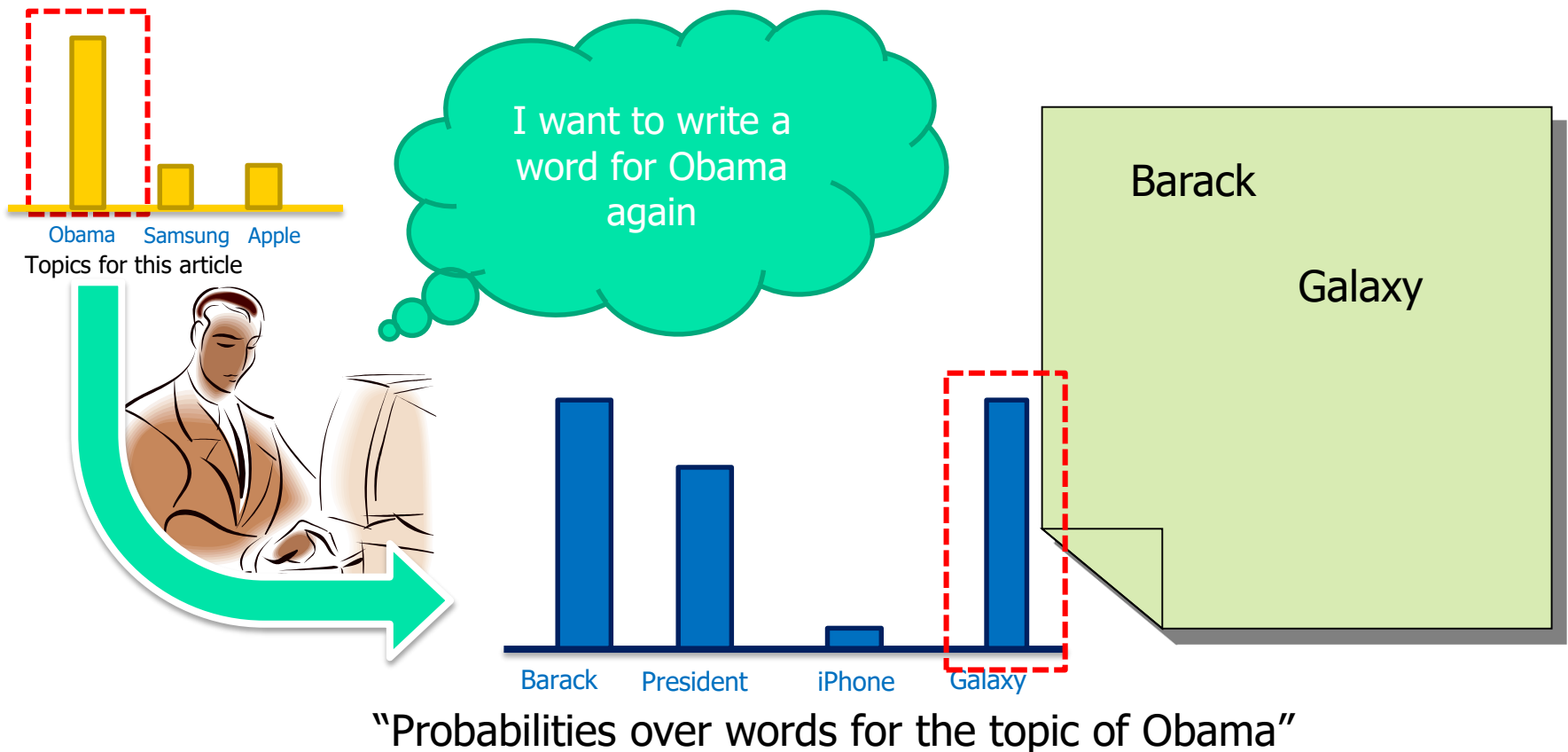
# Generative Model Illustration

The generative model of PSLI assumes that  
**an article is written according to the following process**



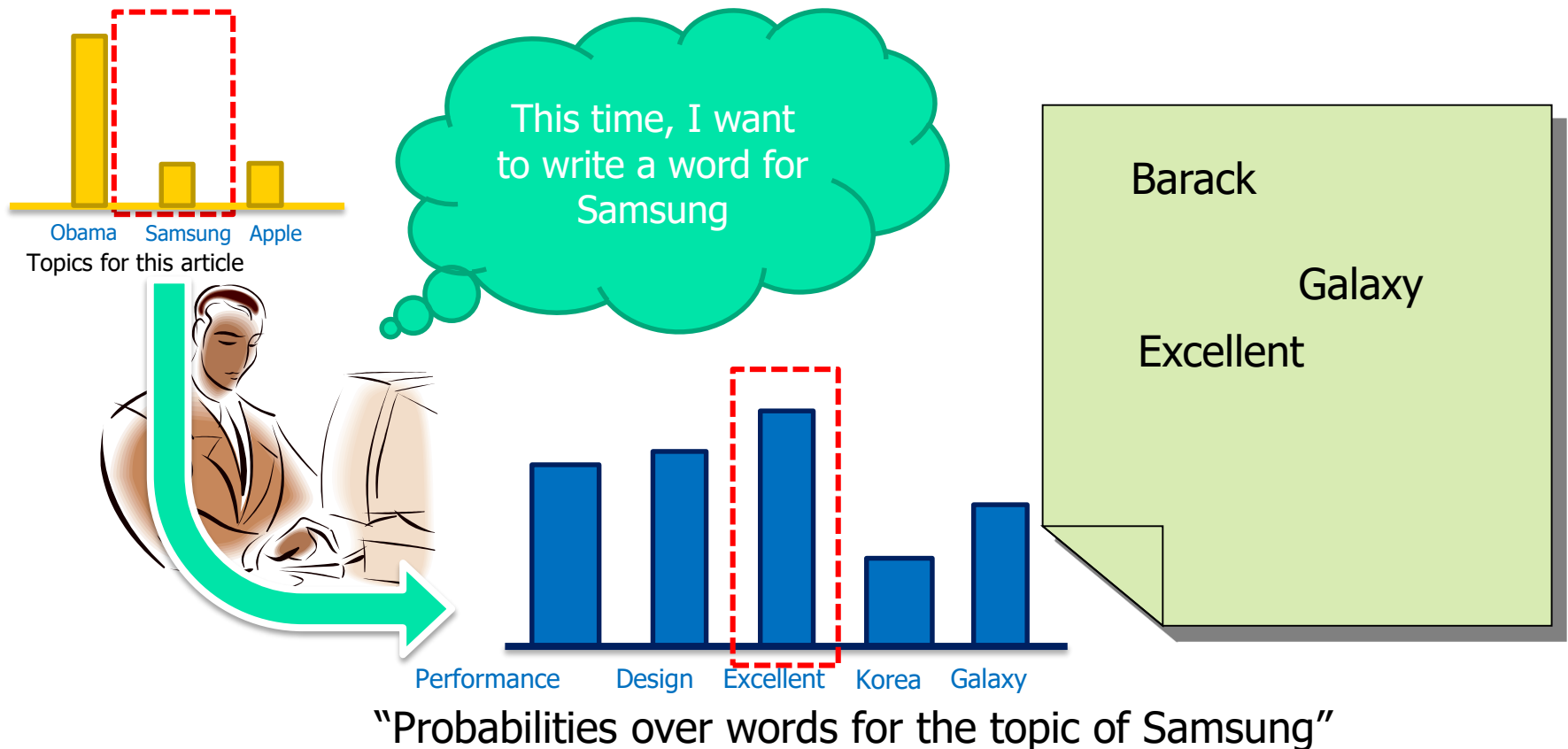
# Generative Model Illustration

The generative model of PSLI assumes that  
**an article is written according to the following process**



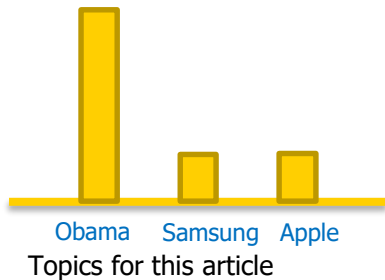
# Generative Model Illustration

The generative model of PSLI assumes that  
**an article is written according to the following process**

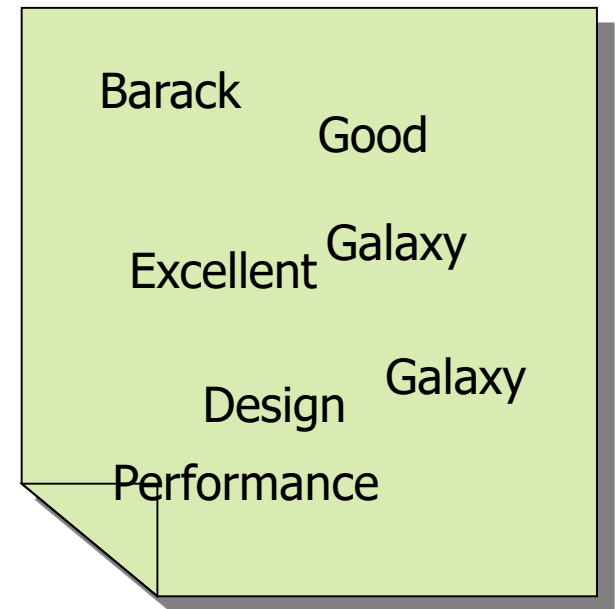


# Generative Model Illustration

Choose words i.i.d. following to the probability distribution



Generate a document




# 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

# 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 Hypertext 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, 3<sup>rd</sup> 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, 2<sup>nd</sup> 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. Indurkha, 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, 2<sup>nd</sup> ed. 2005