Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 1 —

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Data Explosion Problem

- According to DOMO’s 6th Annual Report “**Data Never Sleeps 6.0**” June 5th 2018:

- "Over 2.5 quintillion bytes of data are created every single day, and it’s only going to grow from there. By 2020, it’s estimated that 1.7MB of data will be created every second for every person on earth.“

- “90% of the world’s data has been created in the last two years”

Data Explosion Problem

- Automated data collection tools and mature database technology led to tremendous amounts of data accumulated
- Amount of digital data recording and storage exploded during the past decades

  BUT

- Number of scientists, engineers, and analysts available to analyze the data has not grown correspondingly
- The world is data-rich but information-poor
- We are drowning in data, but starving for knowledge

Why Data Mining?

- Data collected in large data repositories became “data tombs"—data archives that are seldom visited
- Expert Systems were developed that rely on domain experts to manually input knowledge into systems (Costly, Time Consuming, Biases & Errors)
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
- To turn “data tombs” into “golden nuggets” of knowledge

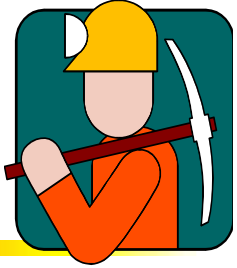
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, hierarchical and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, object-relational models)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and web-based databases e.g., XML databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications expanded

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?



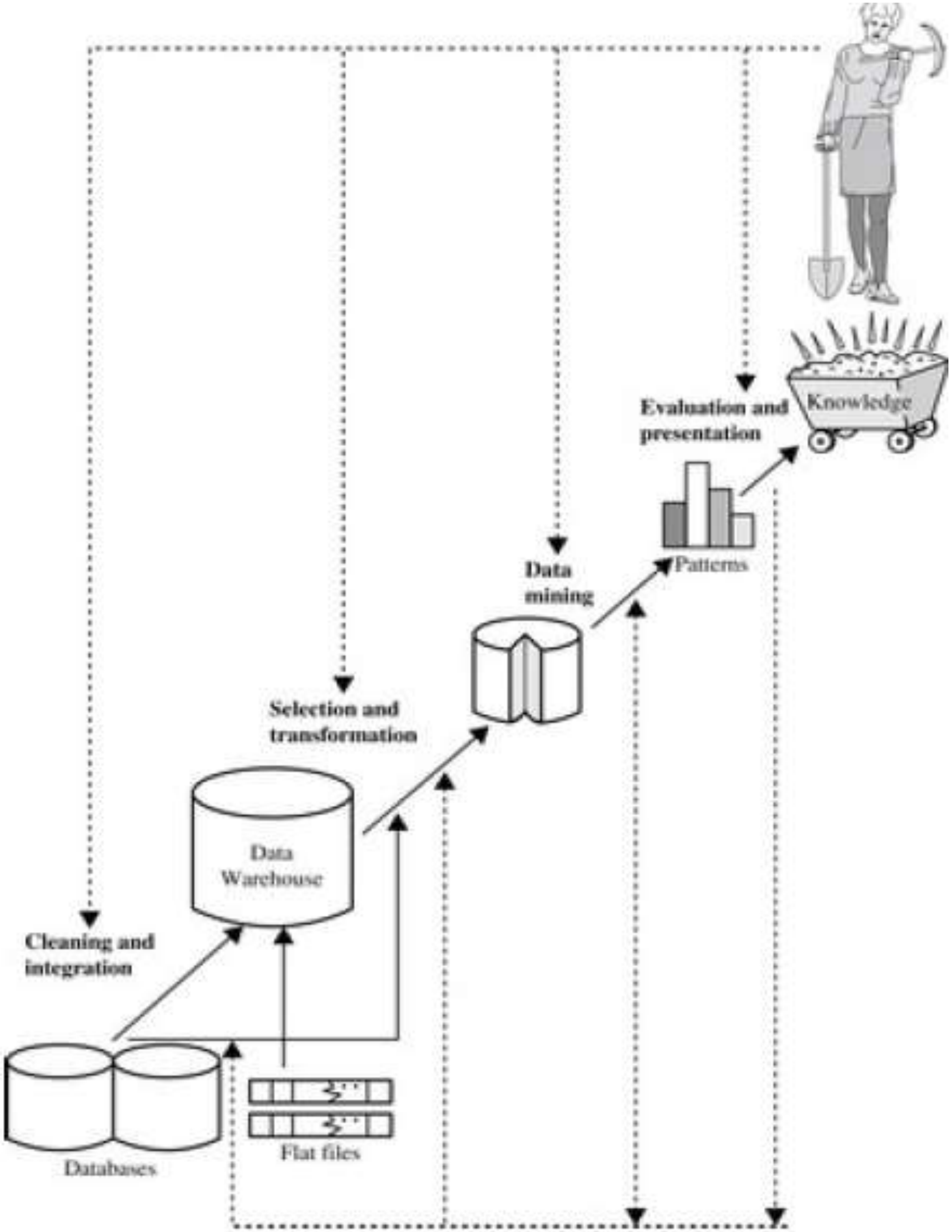
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing

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'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context

Data Mining as a step in the Knowledge Discovery (KDD) Process



Chapter 1. Introduction

- Why Data Mining?
- What Is 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 Kinds of Data Can Be Mined?

- **Database Data**
  - Relational database are one of the most commonly available and richest information repositories
  - Major data source in the study of data mining

customer	*(cust_ID, name, address, age, occupation, annual_income, credit_information, category…)*
item	*(item_ID, brand, category, type, price, place_made, supplier, cost, …..)*
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_D, item_ID, qty)*
works_at	*(empl_ID, branch_ID)*

Relational schema for a relational database, *AllElectronics*

What Kinds of Data Can Be Mined?

- **Data Warehouses**
  - Data cleaning, data integration, data transformation, and periodic refreshing
  - Data Cube
    - Allows pre-computation and fast access to summarized data (Multidimensional Data Mining or Exploratory Data Mining)
    - Allows drill-down and roll-up

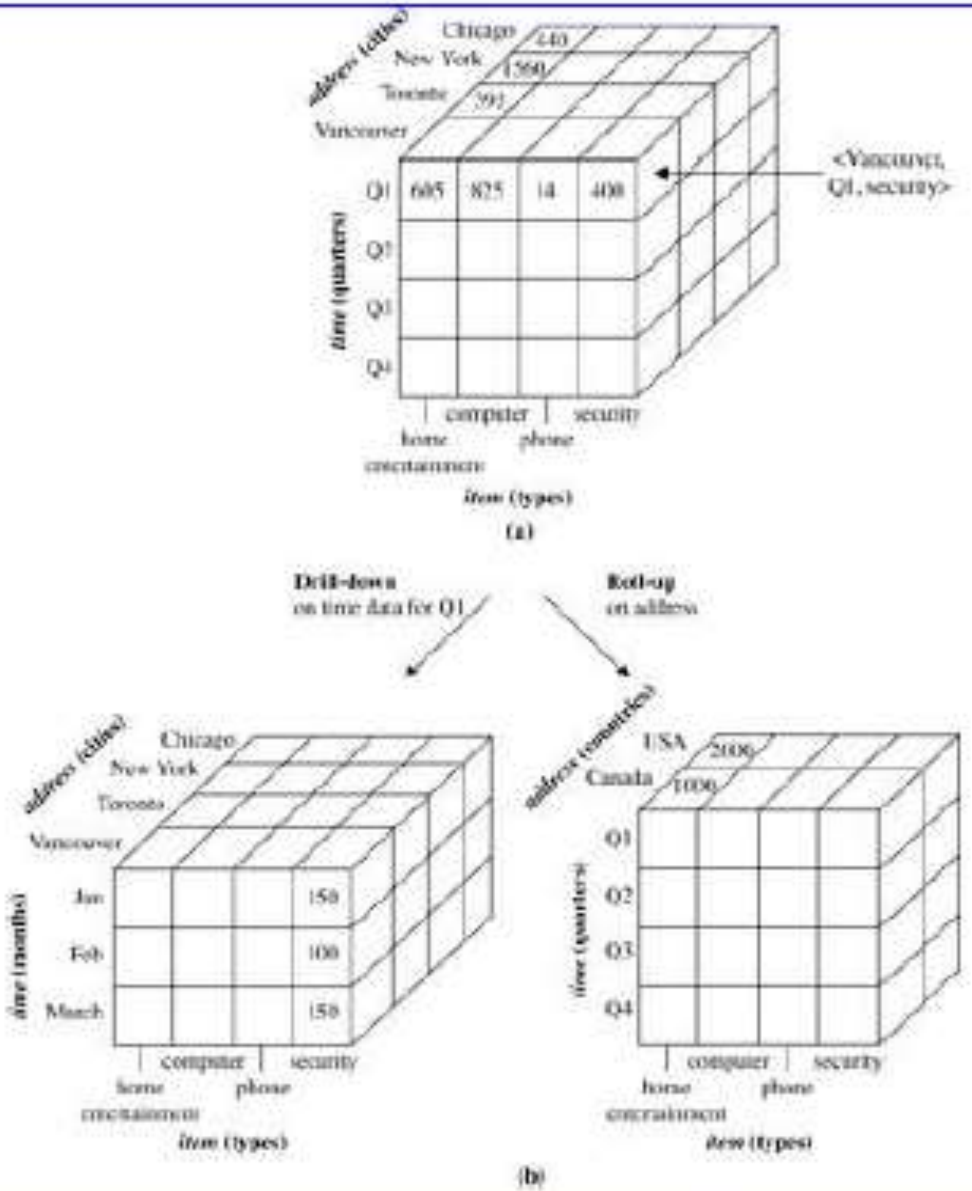


Figure 1.7A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for *AllElectronics* and (b) showing summarized data resulting fromdrill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.

What Kinds of Data Can Be Mined?

- **Transactional Data**
  - Each **record** in a **transactional database** captures a transaction
  - E.g., a customer's purchase, flight booking, user's clicks on a web page
  - Includes a Transaction ID (trans_ID) and list of the items making up the transaction (list_of_item_IDs)
  - Application: Market basket analysis to bundle groups of items together as a strategy for boosting sales (mining *frequent itemsets*)

trans_ID	*list_of_items_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
…	…

Fragment of a transactional database for the sales at *AllElectronics*

What Kinds of Data Can Be Mined?

- **Other Kinds of Data**
  - Temporal/time-related data, mining banking data to schedule bank tellers according to customers' traffic volume
  - Data streams (e.g., video surveillance and sensor data), mining computer network data streams to detect intrusions based on outlier analysis
  - Spatial data (e.g., maps), looking for patterns describing changes in metropolitan poverty rates based on city distances from major highways
  - Text data, mining text data such as literature for the past ten years, one can identify the evolution of hot topics in different eras; sentiment analysis
  - Multimedia data (e.g., image, video, audio data), mining objects to identify objects and classify them by assigning tags; mining video data etc.
  - Web mining, classifying web pages, identifying relationships among different web pages, users, communities, and web-based activities; social media 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

What Kinds of Patterns Can Be Mined?

- ***Data Mining Functionalities*** are used to specify the kinds of patterns to be found in data mining tasks
- Two types:
  - **Descriptive**, describes properties of the data in a target data set
  - **Predictive**, performs induction on the current data in order to make predictions

Data Mining Function: Class/Concept Description

- Data entries can be associated with classes or concepts
- For example:
  - **Classes** of items for sale include computers and printers
  - **Concepts** of customers include bigSpenders and budgetSpenders
- Types of class/concept descriptions
  - Data characterization
  - Data discrimination

Data Mining Function: Class/Concept Description: (1) Characterization and Discrimination

- **Data characterization**
  - Summarization of the data of the class under study (target class)
  - Can be represented in the form of pie charts, bar charts, curves, multidimensional data cubes, multidimensional tables (crosstabs)
  - **Example:**
    - DM Task: *Summarize the characteristics of customers who spend more than $5000 a year at AllElectronics*
    - Output: Profile of customers with age [40-50], employed, excellent credit ratings

Data Mining Function: Class/Concept Description: (1) Characterization and Discrimination

- **Data discrimination**
  - Comparison of a class with one or more contrasting classes
  - Example:
    - DM Task: *Compare two groups of customers-those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year)*
    - Output: Profile of customers, such as 80% of ***frequent customers*** age [20-40] with university degree
    - 60% of ***infrequent customers*** are either seniors or youths with no university degree

Data Mining Function: (2) Mining Frequent Patterns, Association Rule Mining

- **Frequent patterns**
  - Patterns that occur frequently in data
  - Leads to the discovery of interesting associations and correlations within data
  - Types:
    - **Frequent itemsets**, set of items that often appear together e.g., milk and bread are frequently bought together
    - **Frequent subsequence** (sequential pattern), frequently occurring subsequence e.g., 1st laptop followed by digital camera followed by memory card
  - Example: A rule mind from AllElectronics transactional database:

buys(X, "computer") → buys(X, "software") [support = 1%, confidence = 50%]

Data Mining Function: (2) Mining Frequent Patterns, Associations and Correlation

buys(X, "computer") → buys(X, "software") [support = 1%, confidence = 50%]

- ***X*** = variable representing a customer
- ***confidence (certainty)*** of 50% means if a customer buys a computer, there is 50% chance that he will buy software also
- ***support*** of 1% means that 1% of all transactions under analysis show that computer & software are purchased together
- ***Single-dimensional association rule***, single attribute or predicate (buys)
- Simply: ***computer → software [1%, 50%]***

Data Mining Function: (2) Mining Frequent Patterns, Associations and Correlation

age(X, "20..29") ^ income(X, "40K..49K") →

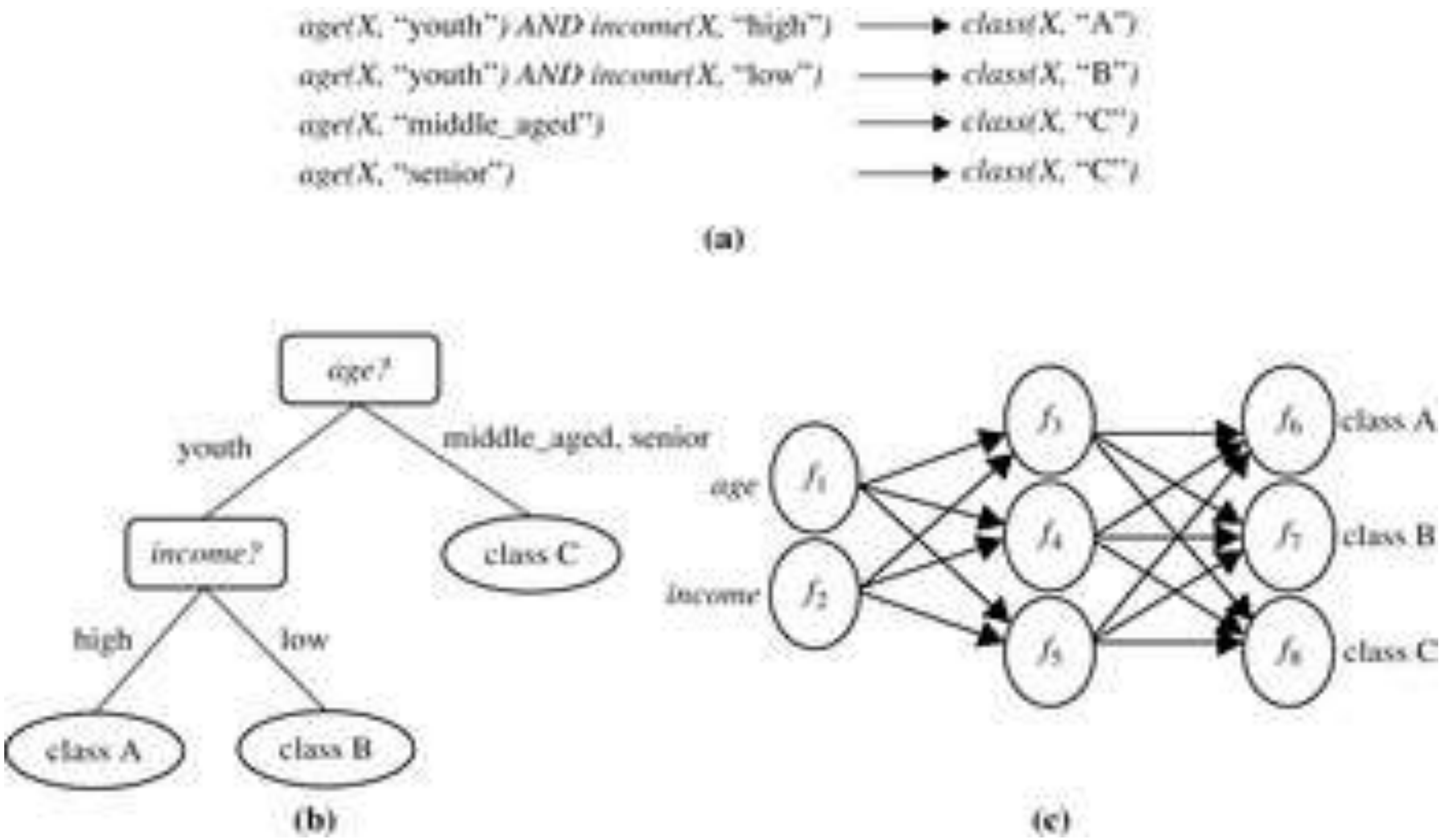
buys(X, "laptop") [support = 2%, confidence = 60%]

- **X** = variable representing a customer
- ***Multidimensional association rule*** , more than one attributes/predicates (age, income, buys)
- Of the AllElectronics customers under study
- 2% are 20 to 29 years old
- with an income of $40,000 to $49000
- and have purchased a laptop at AllElectonics
- There is 60% probability that a customer in this age and income group will purchase a laptop
- Association rules are **discarded** as uninteresting if they do not satisfy both ***minimum*** support and confidence thresholds

Data Mining Function: (3) Classification and Regression for Predictive Analysis

- Goal: previously unseen records should be assigned a class as accurately as possible
- Given data set is divided into ***training set*** and ***test set***, with training set used to ***build*** the model and test set used to ***validate*** it
- Model construction:
  - The set of tuples used for model construction is ***training set***
  - Each tuple/sample in the ***training set*** is assumed to belong to a predefined class, as determined by the ***class label attribute***
  - The derived model may be represented as classification rules (IF-THEN rules), decision trees, mathematical formulae, or neural networks (figure on next slide)

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network



age(X, "youth") AND income(X, "high") → class(X, "A")

age(X, "youth") AND income(X, "low") → class(X, "B")

age(X, "middle_aged") → class(X, "C")

age(X, "senior") → class(X, "C")

(a)

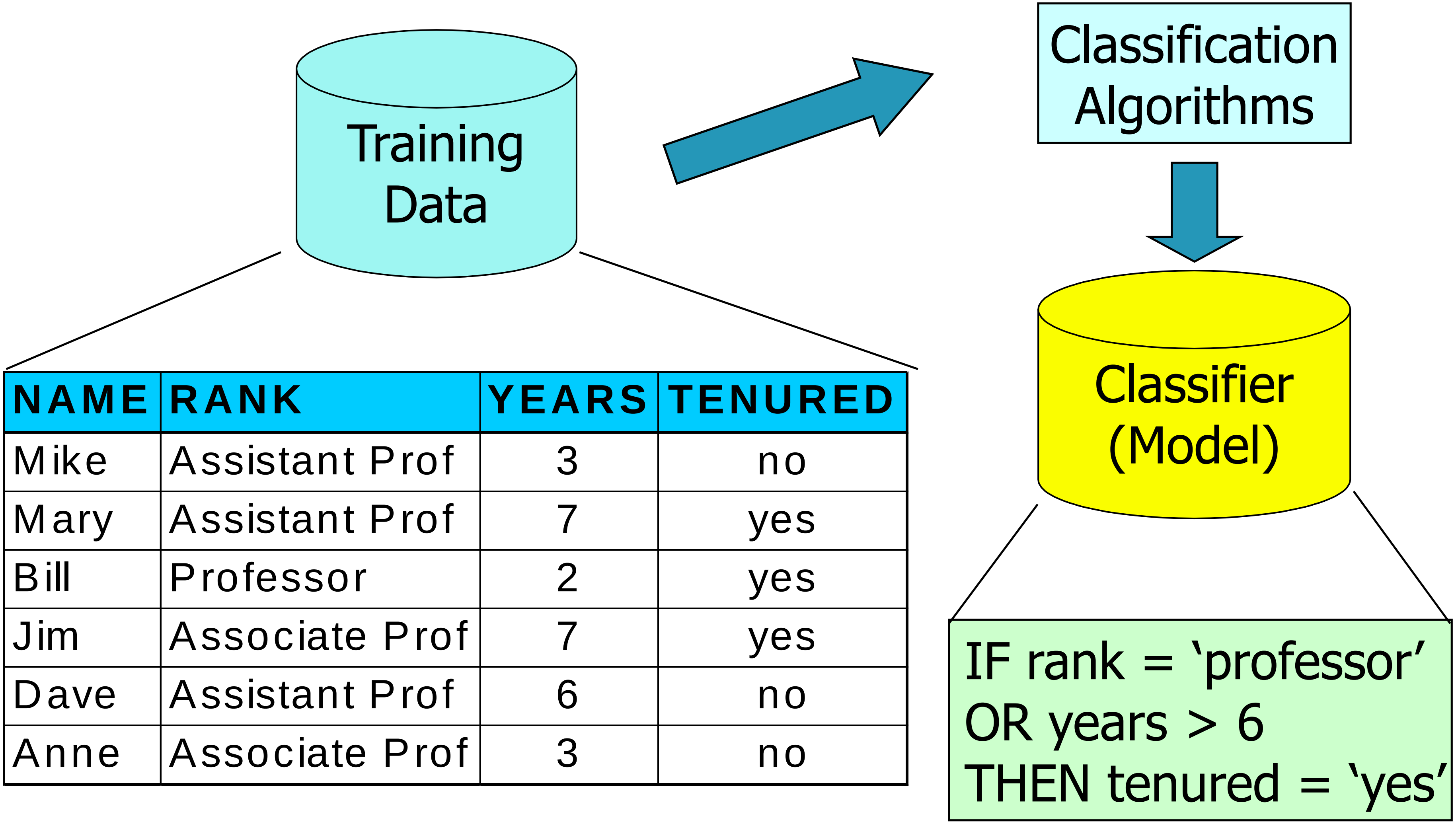
(b)

(c)

Data Mining Function: (3) Classification

- Estimate accuracy of the model:
  - A *test set* is used to determine the accuracy of the model
  - The known label of test sample is compared with the classified result from the model
  - Accuracy rate is the percentage of test set samples that are correctly classified by the model
- Model usage:
  - Predict some unknown class labels if the accuracy is acceptable
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks

Process (1): Model Construction



Training Data

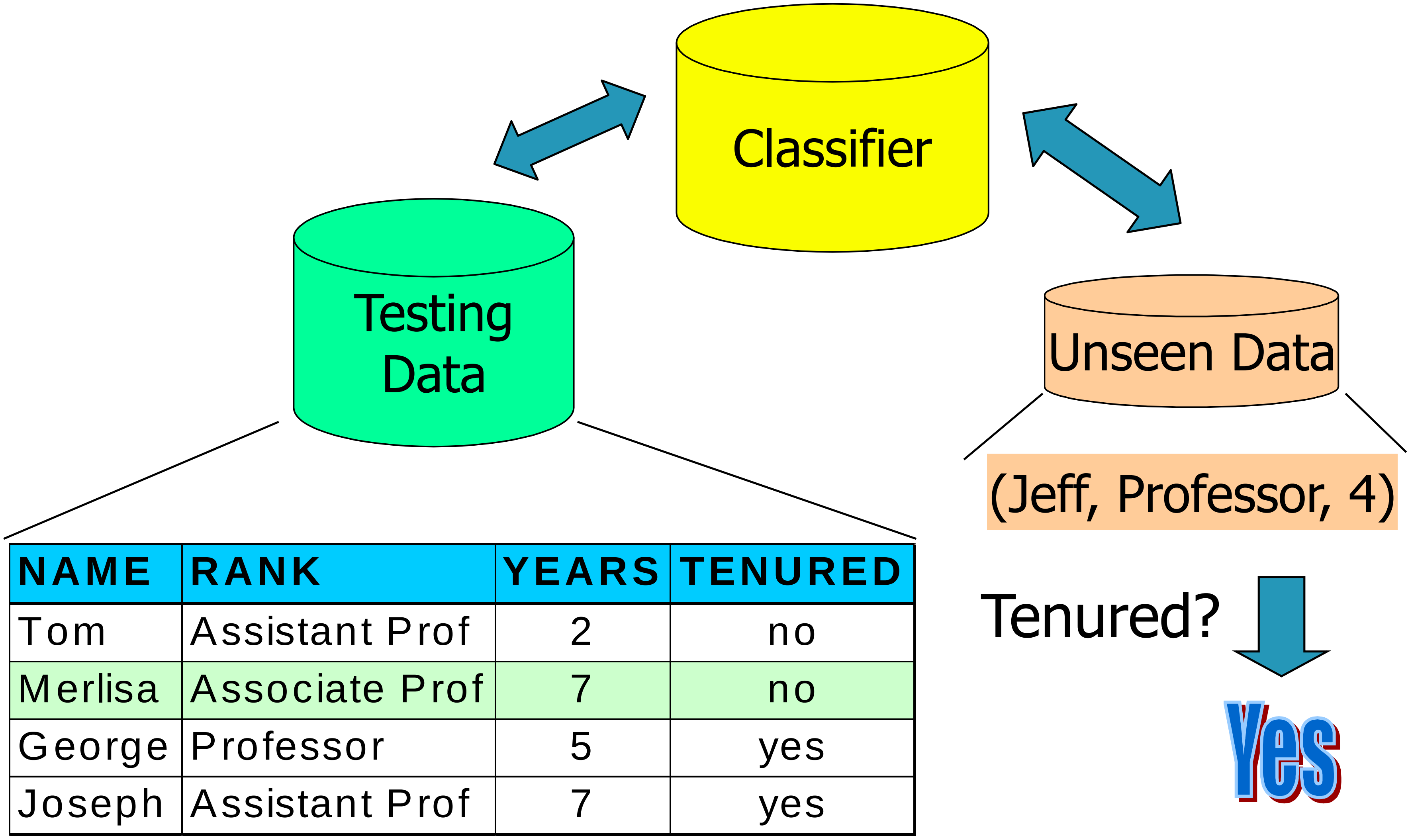
Classification Algorithms

Classifier (Model)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Data Mining Function: (3) Regression

- Whereas classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions
- That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels
- The term ***prediction*** refers to both numeric prediction and class label prediction

CLASSIFICATION (Examples)

- Assigning customers to predefined customer segments (good vs. bad)
- Classifying credit applicants as low, medium, or high risk
- Fraud Detection to predict fraudulent cases in credit card transactions
- Sky Survey Cataloging to predict class (star or galaxy) of sky objects
- Instructor rating as excellent, very good, good, fair, or poor
- Weather into rainy/non-rainy

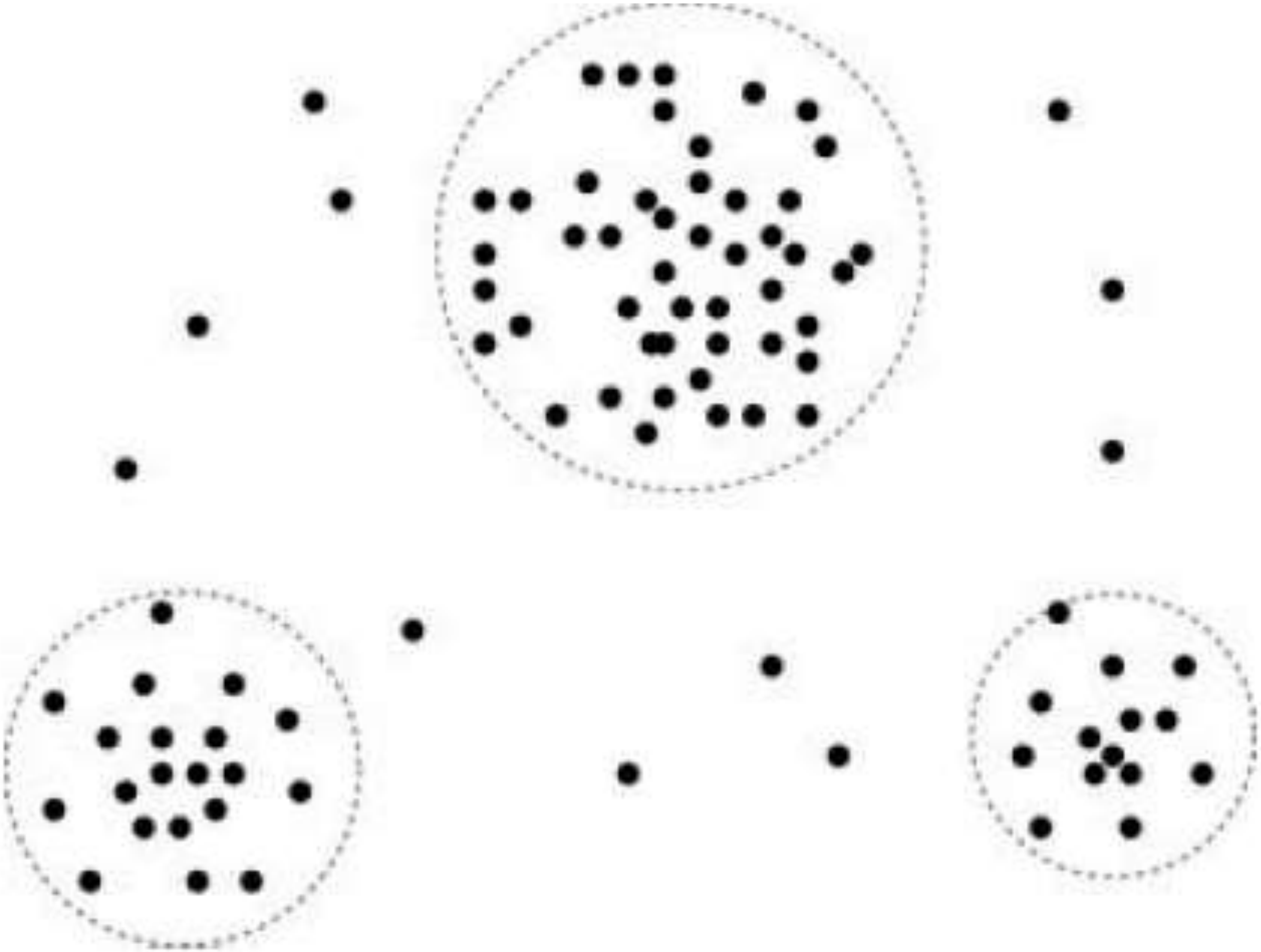
Data Mining Function: (4) Cluster Analysis

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another
- There are no predefined classes and records are grouped together on the basis of self similarity

Data Mining Function: (4) Cluster Analysis

- What distinguishes clustering from classification is that clustering does not rely on predefined classes (i.e., Class label is unknown)
- It is up to you to determine what meaning, if any, to attach to the resulting clusters. Since business decisions are unknown (it is also called unsupervised Learning)
- 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

A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters



Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Clustering: Application

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	***Total Articles***	***Correctly Placed***
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Data Mining Function: (5) Outlier Analysis

- Outlier analysis or Anomaly mining
    - Outlier: A data object that does not comply with the general behavior of the data
    - Methods: by-product of clustering
    - Useful in fraud detection, rare events analysis (e.g., unusually large amount, location, frequency, type of purchase for a given account)

Are All Patterns Interesting?

- Are all mined knowledge interesting?
  - A mining system can generate thousands or even millions of patterns, or rules
  - However, a small faction of patterns are ***interesting***
- A patterns is interesting if:
  - Easily understood by humans
  - Valid on new or test data with some degree of certainty
  - Potentially useful
  - Novel
- A pattern is also interesting if it ***validates a hypothesis*** that the user ***sought to confirm***
- Interesting pattern represents ***knowledge***

What makes a pattern *interesting*?

- **Objective measures** of pattern interestingness
  - Objective measures for association rules of the form X → Y are:
    - support
    - confidence
    - accuracy
  - ***support*** represents the percentage of transactions that the given rule satisfies, from a transaction database
    - *support* $(X \rightarrow Y) = P (X \cup Y)$
  - ***confidence*** assesses the degree of certainty of the rule
    - *confidence* $(X \rightarrow Y) = P (Y \mid X)$

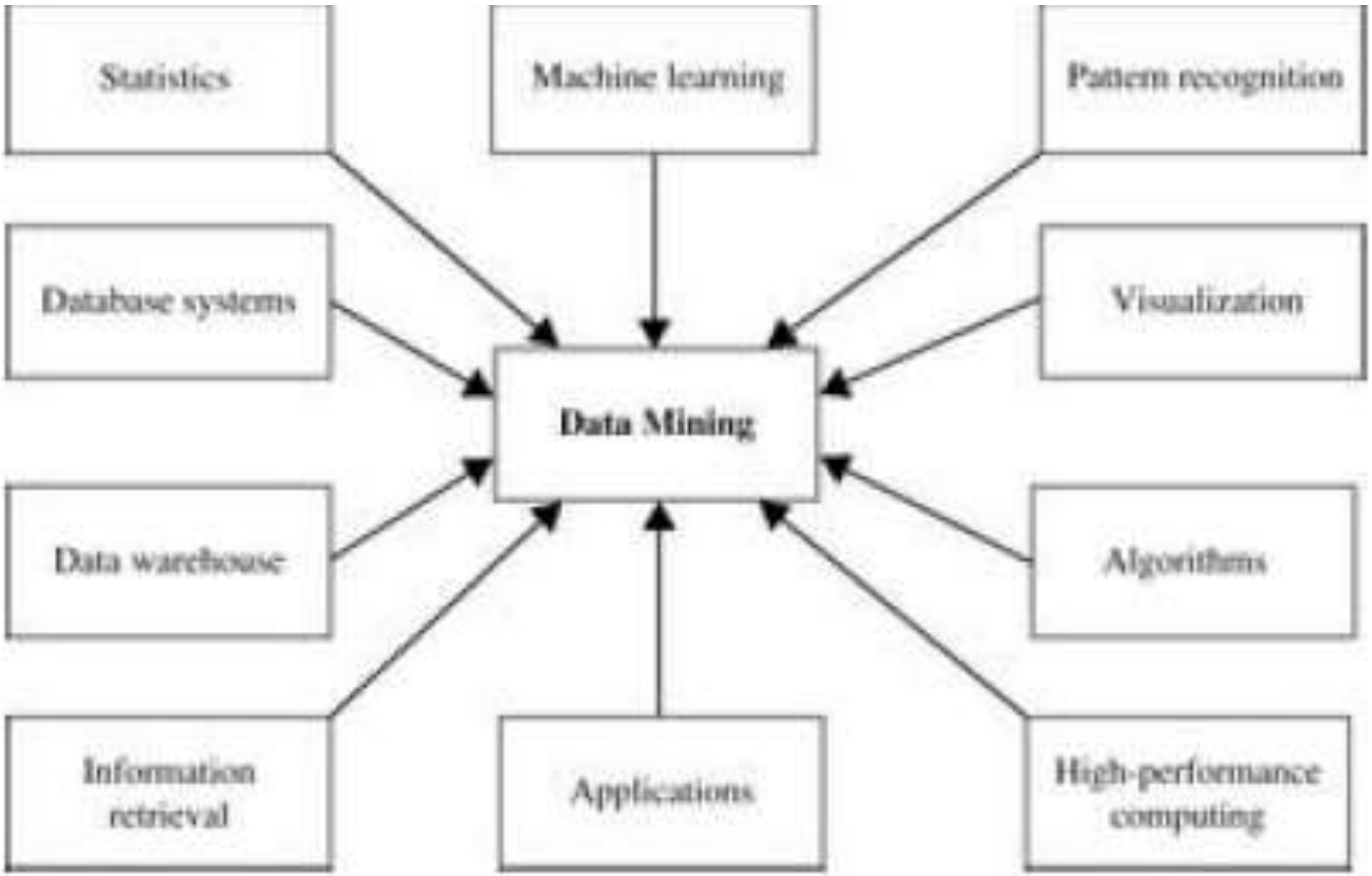
What makes a pattern *interesting*?

- ***accuracy*** tells us the percentage of data that are correctly classified by a rule
- In general, each interestingness measure is associated with a threshold, which may be controlled by the user
- For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting
- Rules below the threshold likely reflect noise, exceptions, or minority cases and are probably of less value

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
- Data mining has an inherent connection with statistics
- Statistical models can be the outcome of a data mining task
  - e.g., in data mining tasks like data characterization statistical models of target classes can be built (sum, mean, median etc.)
- Alternatively, data mining tasks can be built on top of statistical models
  - e.g., we can use statistics to model missing data values (regression analysis)

Statistics

- Challenges:
  - 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 applying data mining on real-time fast data streams

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
- For example, a typical machine learning problem is to program a computer so that it can automatically ***recognize handwritten postal codes*** on mail after learning from a set of examples

Database Systems

- Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users
- Particularly, database systems researchers have established highly recognized principles in data models, query languages, query processing and optimization methods, data storage, and indexing and accessing methods
- Database systems are often well known for their high scalability in processing very large, relatively structured data sets
- Many data mining tasks need to handle large data sets or even real-time, fast streaming data
- Therefore, data mining can make good use of ***scalable database technologies*** to achieve high ***efficiency*** and ***scalability*** on large data sets

Data Warehouses

- A Data Warehouse integrates data from multiple sources into a unified schema
- Cleaned data
- Promotes OLAP cubes and 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
- Differences between IR and database systems are two fold:
  1. IR work on structured data DB systems work on unstructured data
  2. Queries are formed mainly by keywords in IR and do not have complex structures (unlike SQL queries in database systems)

Information Retrieval

- 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

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: Efficiency and Scalability of Data Mining Algorithms

- As data amounts continue to multiply, ***efficiency and scalability*** are always considered when ***comparing data mining algorithms***
- The ***running time*** of a data mining algorithm must be ***predictable and short***
- ***Efficiency, scalability,*** and the ability to ***execute in real time*** are key criteria that drive the development of many new data mining algorithms

Major Issues in Data Mining: Parallel and distributed mining algorithms

- The ***humongous size*** of many data sets, the ***wide distribution of data***, and the ***computational complexity*** of some data mining methods are factors that motivate the development of ***parallel and distributed*** data mining algorithms
- Such algorithms first ***partition the data into "pieces."*** Each piece is ***processed, in parallel***, by searching for patterns
- The parallel processes may interact with one another
- The patterns from each partition are ***eventually merged***

Major Issues in Data Mining: Incremental data mining algorithms

- ***Incremental data mining*** incorporates new data updates without having to mine the entire data "from scratch."
- Such methods perform ***knowledge modification incrementally*** to amend and strengthen what was previously discovered

Major Issues in Data Mining: Handling diverse & complex types of data

- Diverse applications generate a wide spectrum of new data types:
  - structured, semi-structured, and unstructured data
  - stable data repositories to dynamic data streams
  - simple data objects to temporal data, biological sequences, sensor data, spatial data, multimedia data, software program code, web data, and social network data
- Unrealistic to expect one data mining system to mine all kinds of data, given the diversity of data types and the different goals of data mining
- Domain or application-dedicated data mining systems are being constructed for in-depth mining of specific kinds of data
- The construction of effective and efficient data mining tools for diverse applications remains a challenging and active area of research

Major Issues in Data Mining: Mining dynamic, networked, and global data repositories

- The discovery of knowledge from different sources of structured, semi-structured, or unstructured yet ***interconnected*** data poses great challenges to data mining
- Mining such gigantic, interconnected information networks may help disclose many more patterns as compared to small sets of ***isolated*** data repositories
- Web mining, information network mining and links analysis have become challenging and fast-evolving data mining fields

Major Issues in Data Mining: Social impacts of data mining

- With data mining penetrating our everyday lives, it is important to study the impact of data mining on ***society***
- How can we use data mining technology to benefit society? How can we guard against its misuse?
- The philosophy is to observe data sensitivity and preserve ***people's privacy*** while performing successful data mining

Major Issues in Data Mining: Privacy-preserving data mining

- DM poses the risk of disclosing an individual's personal information
- Studies on privacy-preserving data publishing and data mining are ongoing
- The improper disclosure or use of data and the potential violation of ***individual privacy*** and data protection rights are areas of concern that need to be addressed

Major Issues in Data Mining: Invisible data mining

- We cannot expect everyone in society to learn and master data mining techniques
- Systems should have builtin data mining functions so that people can perform data mining simply by mouse clicking, ***without any knowledge of data mining algorithms***
- Intelligent search engines and Internet-based stores perform such invisible data mining by incorporating data mining into their components to improve their functionality and performance
- This is done often ***unbeknownst to the user***. For example, when purchasing items online, users may be unaware that the store is likely collecting data on the ***buying patterns*** of its customers, which may be used to recommend other items for purchase in the future

Major Issues in Data Mining: Data mining—an interdisciplinary effort

The power of data mining can be substantially enhanced by ***integrating new methods from multiple disciplines***

For example, to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and ***natural language processing***

As another example, consider the mining of software bugs in large programs

This form of mining, known as ***bug mining***, benefits from the incorporation of ***software engineering*** knowledge into the data mining process

Major Issues in Data Mining: Handling uncertainty, noise, or incompleteness of data

Data often contain noise, errors, exceptions, or uncertainty, or are incomplete

Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns

Data cleaning, data preprocessing, outlier detection and removal are examples of techniques that need to be integrated with the data mining process

Major Issues in Data Mining: Interactive mining

The data mining process should be highly interactive

Thus, it is important to build ***flexible user interfaces*** and an ***exploratory mining environment***, facilitating the user's interaction with the system

Major Issues in Data Mining: Presentation and visualization of data mining results

How can a data mining system present data mining results, vividly and flexibly, so that the discovered knowledge can be easily understood and directly usable by humans?

This is especially crucial if the data mining process is interactive

It requires the system to adopt ***user-friendly interfaces***, and ***visualization techniques***

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

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005