Data Mining & Business Intelligence (IT647)

Selected Topics In Data Science (IT688)

Information Technology Department

College Of Computer

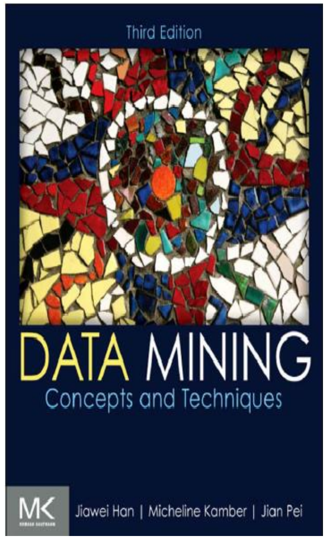
Qassim University

DR.MONA ALSALEEM

Data Mining

— Chapter 1 — Introduction

Text Book: “Data Mining: Concepts and Techniques”, Third Edition



- Why Data Mining?
- History Of Data Mining
- What Is Data Mining?
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Summary

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

We are drowning in data, but starving for knowledge!

“Necessity is the mother of invention”—***Data mining***—Automated analysis of massive data sets

History Of Data Mining

- 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.)
- 1990s-now: Data mining, data warehousing, multimedia databases, and Web databases

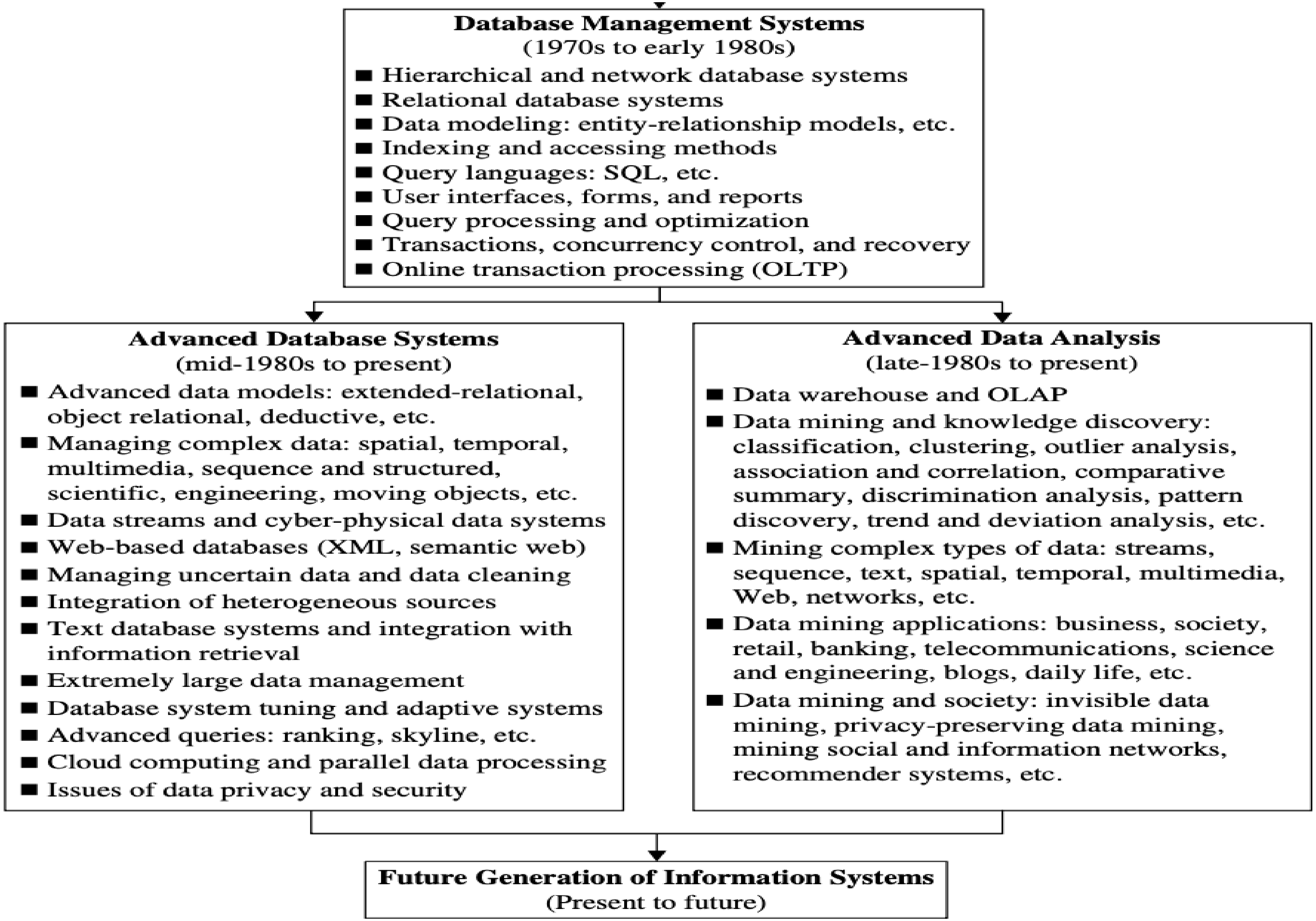
  The ability to store and manage petabytes of data online :

  Data mining **is a major new challenge!**

- 2000s

Stream data management and Data mining with its applications

Web technology (XML, data integration) and global information systems



Database Management Systems
(1970s to early 1980s)

- Hierarchical and network database systems
- Relational database systems
- Data modeling: entity-relationship models, etc.
- Indexing and accessing methods
- Query languages: SQL, etc.
- User interfaces, forms, and reports
- Query processing and optimization
- Transactions, concurrency control, and recovery
- Online transaction processing (OLTP)

Advanced Database Systems
(mid-1980s to present)

- Advanced data models: extended-relational, object relational, deductive, etc.
- Managing complex data: spatial, temporal, multimedia, sequence and structured, scientific, engineering, moving objects, etc.
- Data streams and cyber-physical data systems
- Web-based databases (XML, semantic web)
- Managing uncertain data and data cleaning
- Integration of heterogeneous sources
- Text database systems and integration with information retrieval
- Extremely large data management
- Database system tuning and adaptive systems
- Advanced queries: ranking, skyline, etc.
- Cloud computing and parallel data processing
- Issues of data privacy and security

Advanced Data Analysis
(late-1980s to present)

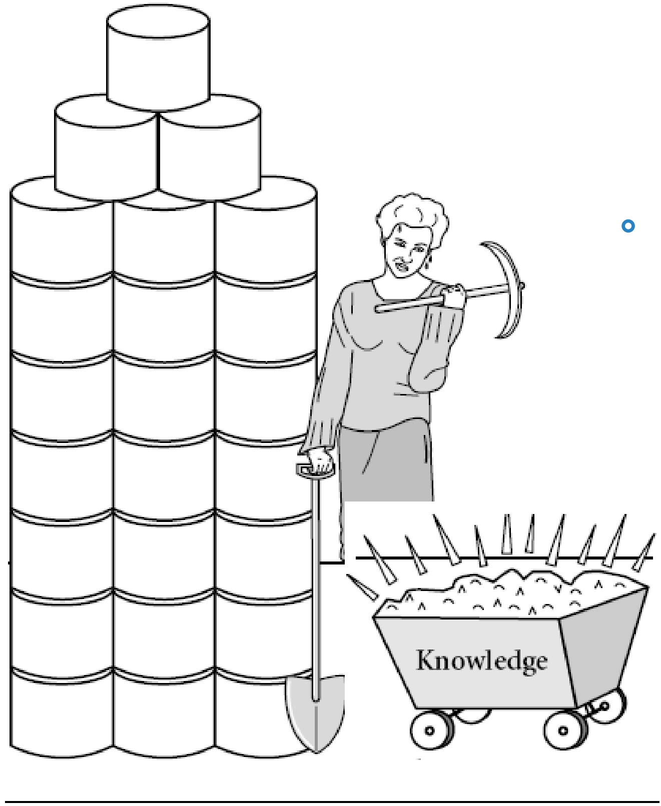
- Data warehouse and OLAP
- Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
- Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
- Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
- Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.

Future Generation of Information Systems
(Present to future)

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Simply stated, data mining refers to extracting or “mining” knowledge from large amounts of data”.
  - Data mining have been more appropriately named “*Knowledge mining from data*”.

Data mining—searching for knowledge (interesting patterns) in your data.



How can I analyze my data?

We have a lot of data but lack the information and knowledge

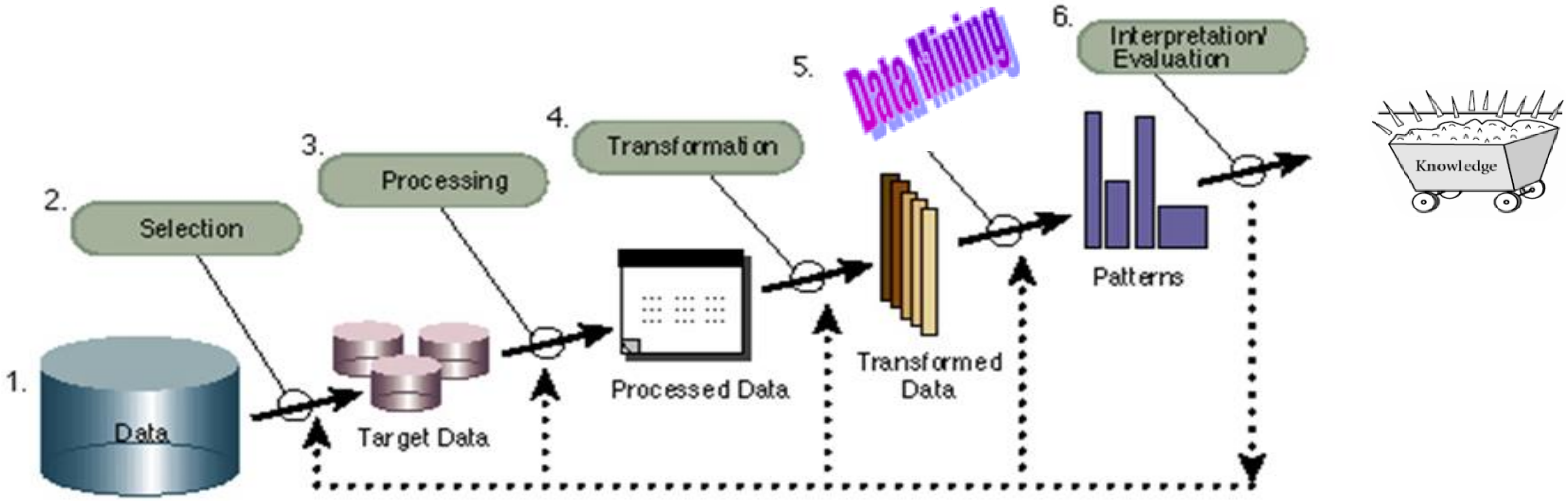
Data Mining is a part of the knowledge Discovery Process, Knowledge Discovery from Data abbreviated as **(KDD)**

Knowledge Discovery (KDD) Process

Typically, the steps of a KDD process are:

- Data cleaning (to remove noise and inconsistent data)
- Data integration (where multiple data sources may be combined)
- Data selection (where data relevant to the analysis task are retrieved from the database)
- Data transformation (converting data into a form more appropriate for mining by performing summary or aggregation operations)
- ***Data mining*** (an essential process where intelligent methods are applied to extract data patterns)
- Pattern evaluation (to identify the interesting patterns representing knowledge)
- Knowledge presentation (where visualization and knowledge presentation techniques are used to present mined knowledge to users)

Knowledge Discovery (KDD) Process



An Overview of the Steps That Compose the KDD Process

Data Mining Scope

- **Finance and business:**

  Fraud detection, Market forecasting

  Basket analysis, Product targeting, Efficient mailing

- **Engineering:**

  Process modeling and optimization

  Machine diagnostics, Predictive maintenance

- **Internet:**

  Text mining, Intelligent query answering

  Web access analysis, Site personalization

- **Medical Informatics and others**

Data Mining: On What Kinds of Data?

Database-oriented data sets and applications

- Relational database, data warehouse, transactional database
- Object-relational databases, Heterogeneous databases and legacy databases.

Advanced data sets and advanced applications

- Data streams
- Structure data, graphs, social networks and information networks
- Spatial data and spatiotemporal data
- Multimedia database
- The World-Wide Web

Data Mining: On What Kinds of Data?

Database Data

- database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- A relational database is a collection of tables, each of which is assigned a unique name.
- 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.

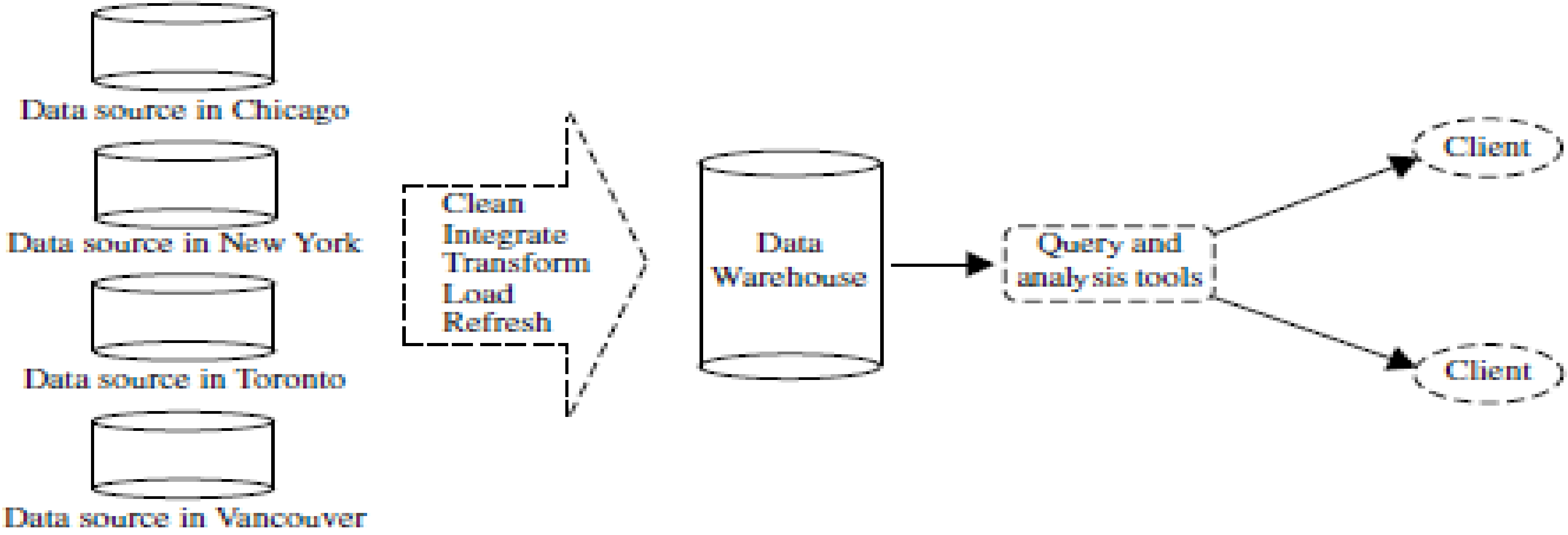
Data Mining: On What Kinds of Data?

Data Warehouses

- 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.
- It is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure.

Data Mining: On What Kinds of Data?

Data Warehouses



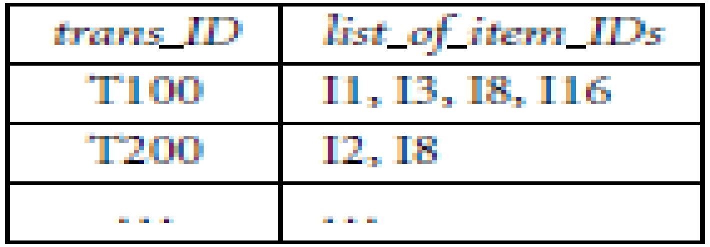
***Typical framework of a data warehouse for AllElectronics*.**

Data Mining: On What Kinds of Data?

Transactional Data

- Each record in a transactional database captures a transaction, such as a customer’s purchase, a flight booking, or a user’s clicks on a web page.

- A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.



trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Fragment of a transactional database for sales at AllElectronics

What Kinds of Patterns Can Be Mined?

There are a number of data mining functionalities. These include:

- characterization and discrimination
- The mining of frequent patterns, associations, and correlations;
- classification and regression;
- clustering analysis
- outlier analysis

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

In general, such tasks can be classified into two categories: descriptive and predictive.

What Kinds of Patterns Can Be Mined?

Data Characterization

- Data characterization is a summarization of the general characteristics or features of a target class of data.

- **For example,** to study the characteristics of software products with sales that increased by 10% in the previous year

Data Discrimination

- Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

- **For example,** a user may want to 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.

What Kinds of Patterns Can Be Mined?

Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

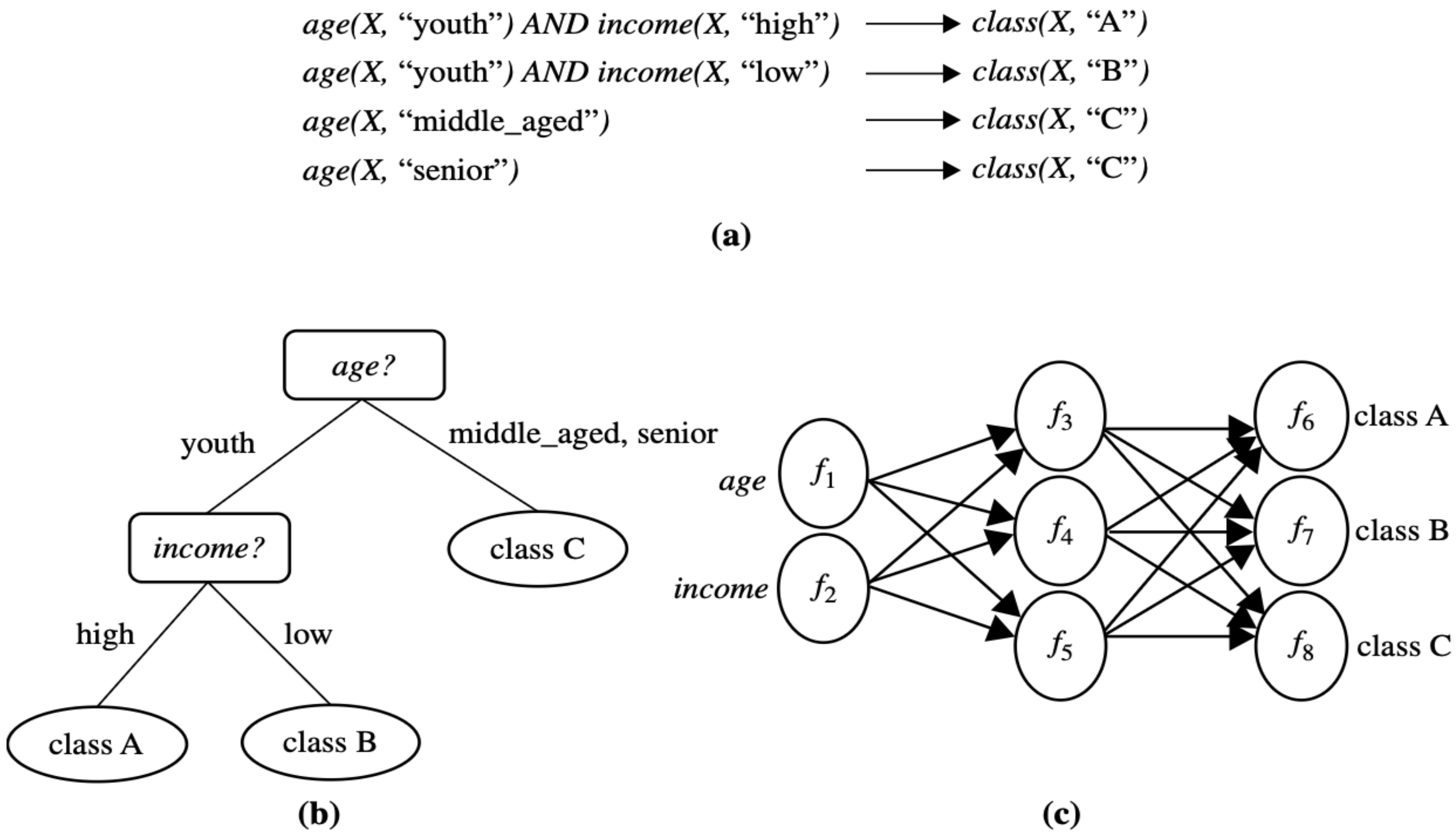
- **Example of association rule :**

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}software\text{”})\ [support = 1\%, confidence = 50\%],$$

- Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.
- Additional analysis can be performed to uncover interesting statistical correlations between associated attribute–value pairs.

Classification

- ○ 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 the class label is unknown.



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)

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

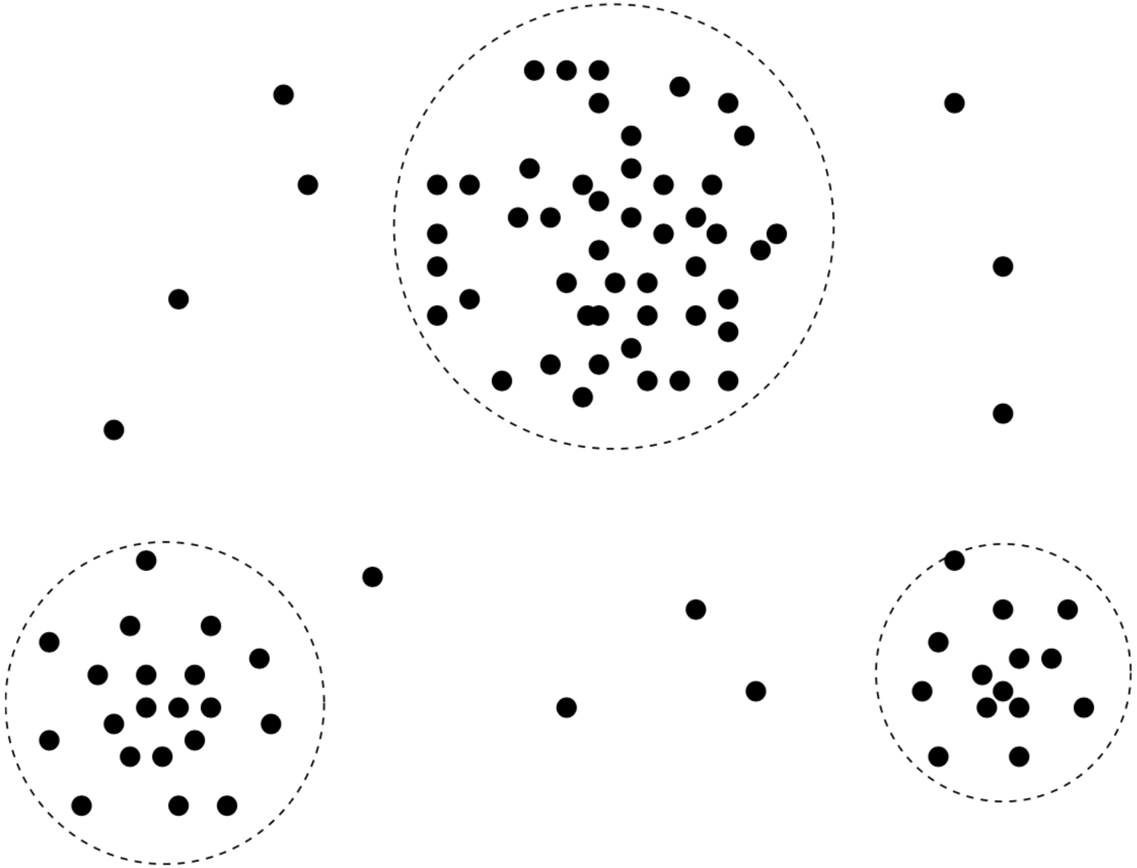
What Kinds of Patterns Can Be Mined?

Regression

Regression analysis is a statistical methodology that is most often used for numeric prediction.

Clustering

clustering analyzes data objects without consulting class labels.



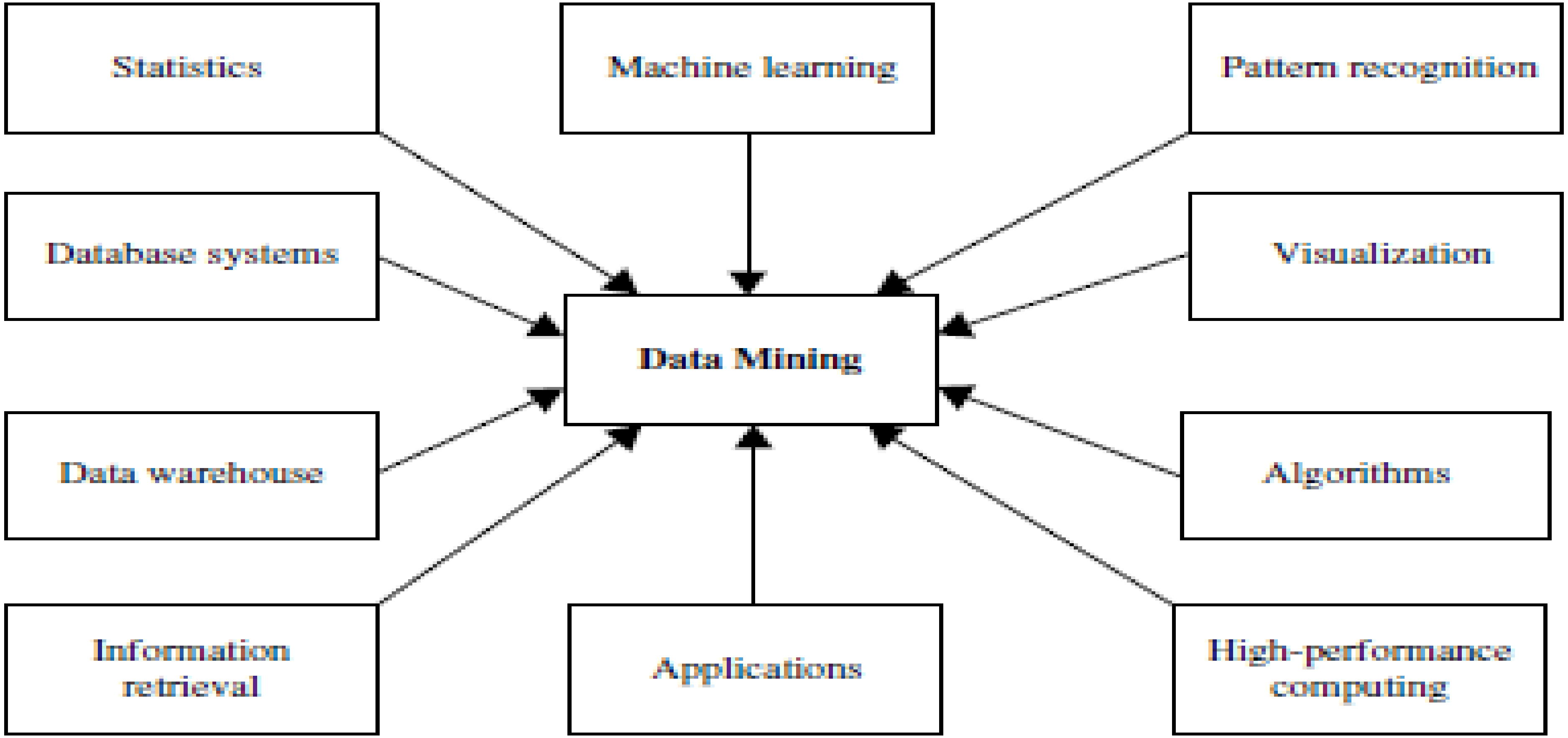
What Kinds of Patterns Can Be Mined?

Outlier analysis

- A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers.

- Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) events can be more interesting than the more regularly occurring

Which Technologies Are Used?



Data mining adopts techniques from many domains.

Which Kinds of Applications Are Targeted?

Where there are data, there are data mining applications

- As a highly application-driven discipline, data mining has seen great successes in many applications;
- It is impossible to enumerate all applications where data mining plays a critical role;
- Presentations of data mining in knowledge-intensive application domains require more in-depth treatment.

Which Kinds of Applications Are Targeted?

Business Intelligence

- Business intelligence (BI) technologies provide historical, current, and predictive views of business operations.

- Clearly, data mining is the core of business intelligence. Online analytical processing tools in business intelligence rely on data warehousing and multidimensional data mining.

- Classification and prediction techniques are the core of predictive analytics in business intelligence

Which Kinds of Applications Are Targeted?

Web Search Engines

- A Web search engine is a specialized computer server that searches for information on the Web.

- Various data mining techniques are used in all aspects of search engines, ranging from :

❖ crawling (deciding which pages should be crawled and the crawling frequencies),

❖ indexing (selecting pages to be indexed and deciding to which extent the index should be constructed)

❖ searching (deciding how pages should be ranked,)

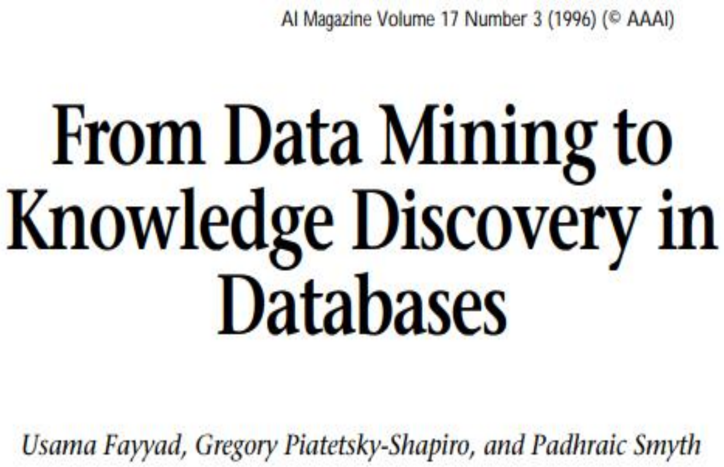
Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank
- Collaborative analysis & recommender systems
- Basket market 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
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information 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, trend and outlier analysis, etc.
- Data mining technologies and applications

Assignment



AI Magazine Volume 17 Number 3 (1996) (© AAAI)

From Data Mining to Knowledge Discovery in Databases

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

This paper is:

- A foundational reference in data science
- Often cited in several research papers
- explain the methodology behind data-driven discovery

Your assignment :

- ➢ Summarize the paper in 2 pages
- ➢ Similar work will have zero

Thank you