DATA MINING

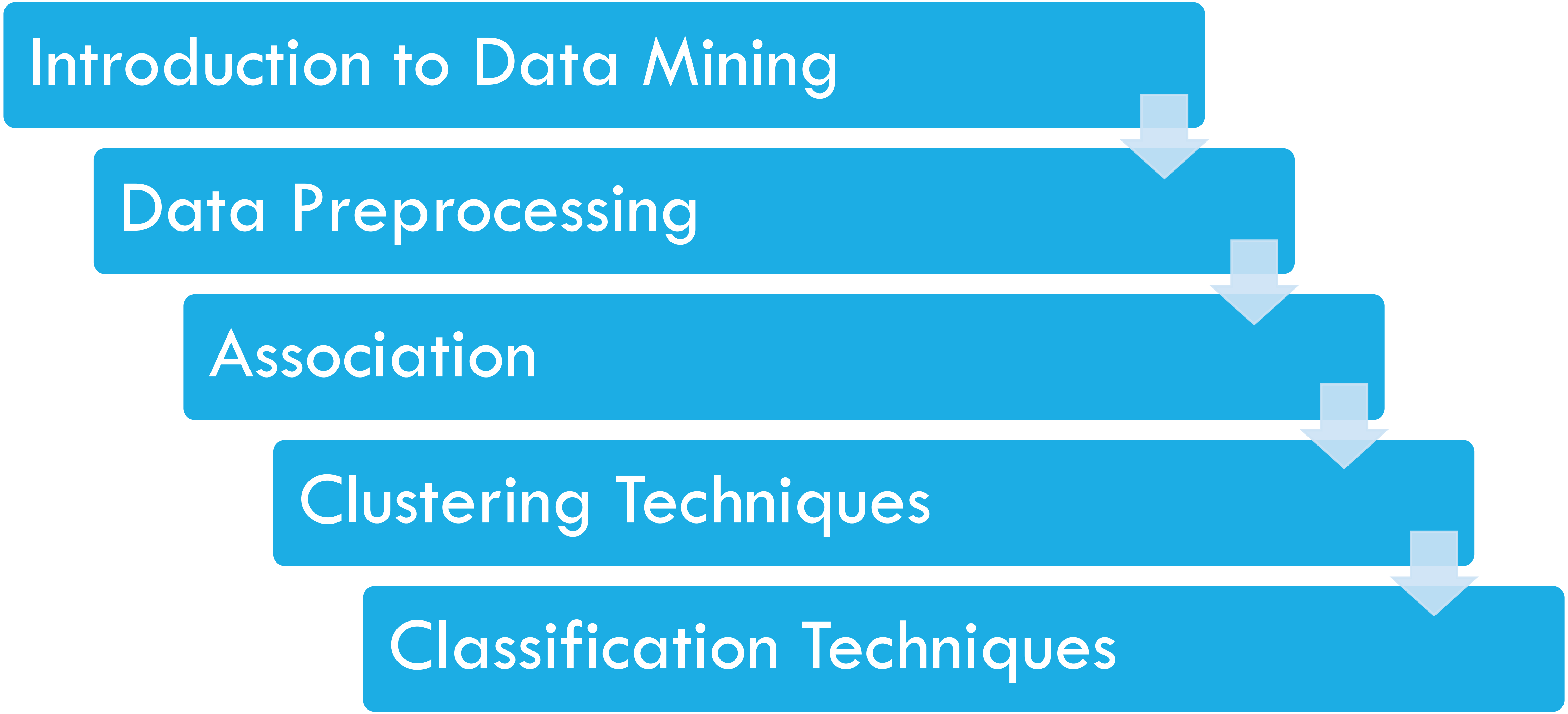
CHAPTER 1 - INTRODUCTION

Dr. Ahmed Said

COURSE GRADING

- MIDTERM (20%)
- YEAR WORK (20%) “Assignments & Attendance”
- FINAL EXAM (60%)

COURSE OUTLINE



Introduction to Data Mining

Data Preprocessing

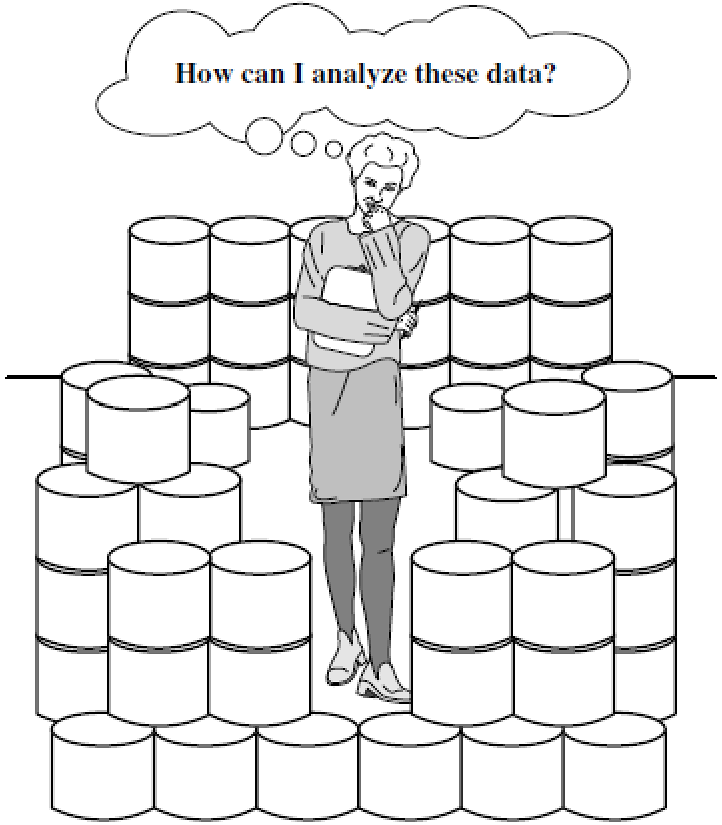
Association

Clustering Techniques

Classification Techniques

WHY DATA MINING?

- The Explosive Growth of Data:
  - Data collection and data availability
    - Automated data collection tools, database systems, web
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: bioinformatics, scientific simulation, medical research …
    - Society and everyone: news, digital cameras, …
- Data rich but information poor!
  - What does those data mean?
  - How to analyze data?

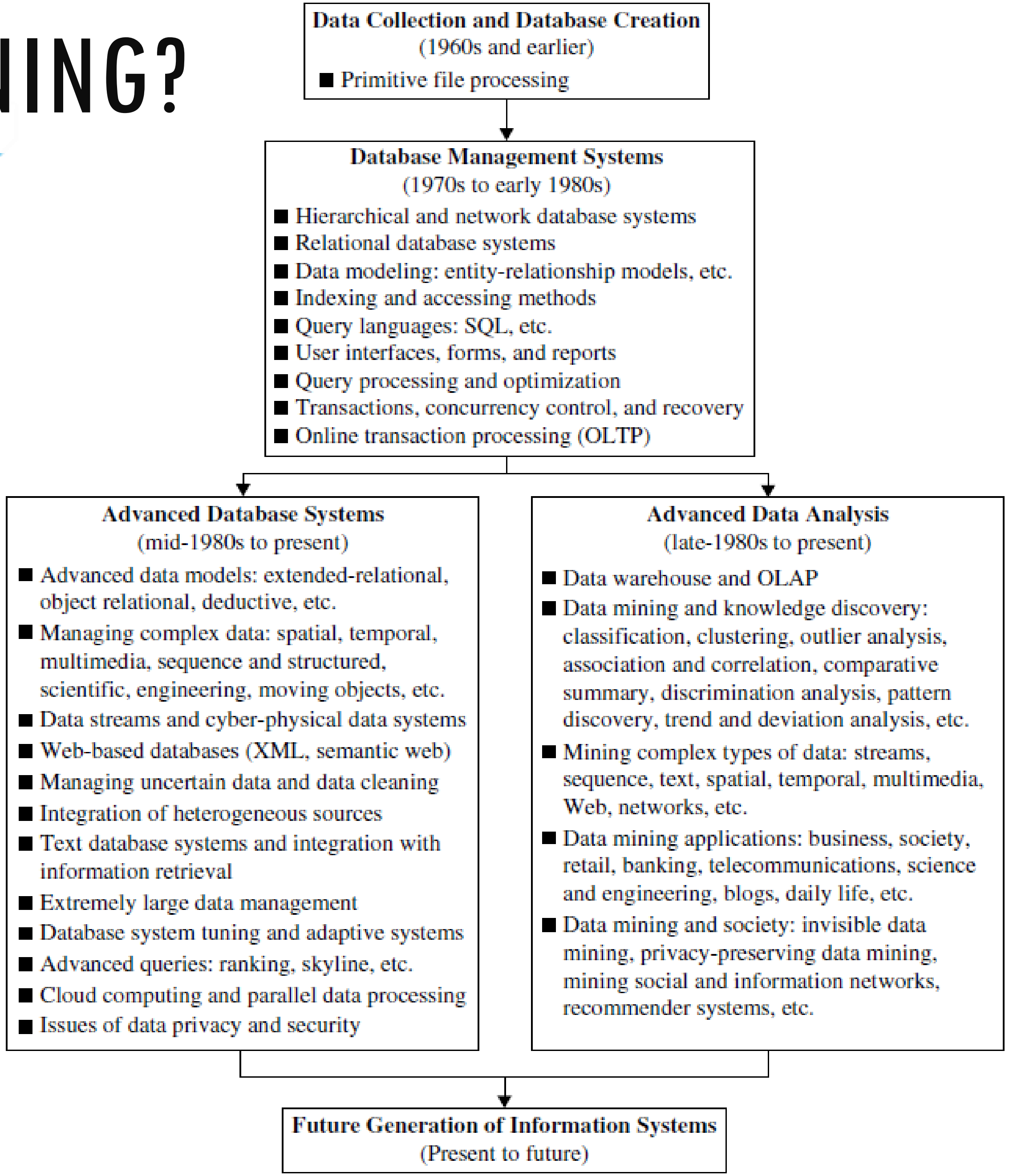


- Data mining — Automated analysis of massive data sets

WHY DATA MINING?

- Data mining can be viewed as a result of the natural evolution of information technology.
- The database and data management industry evolved in the development of several critical functionalities:
  - Data collection and database creation
  - Data management (including data storage and retrieval and database transaction processing)
  - Advanced data analysis (involving data warehousing and data mining)

WHY DATA MINING?



Data Collection and Database Creation
(1960s and earlier)

- Primitive file processing

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)

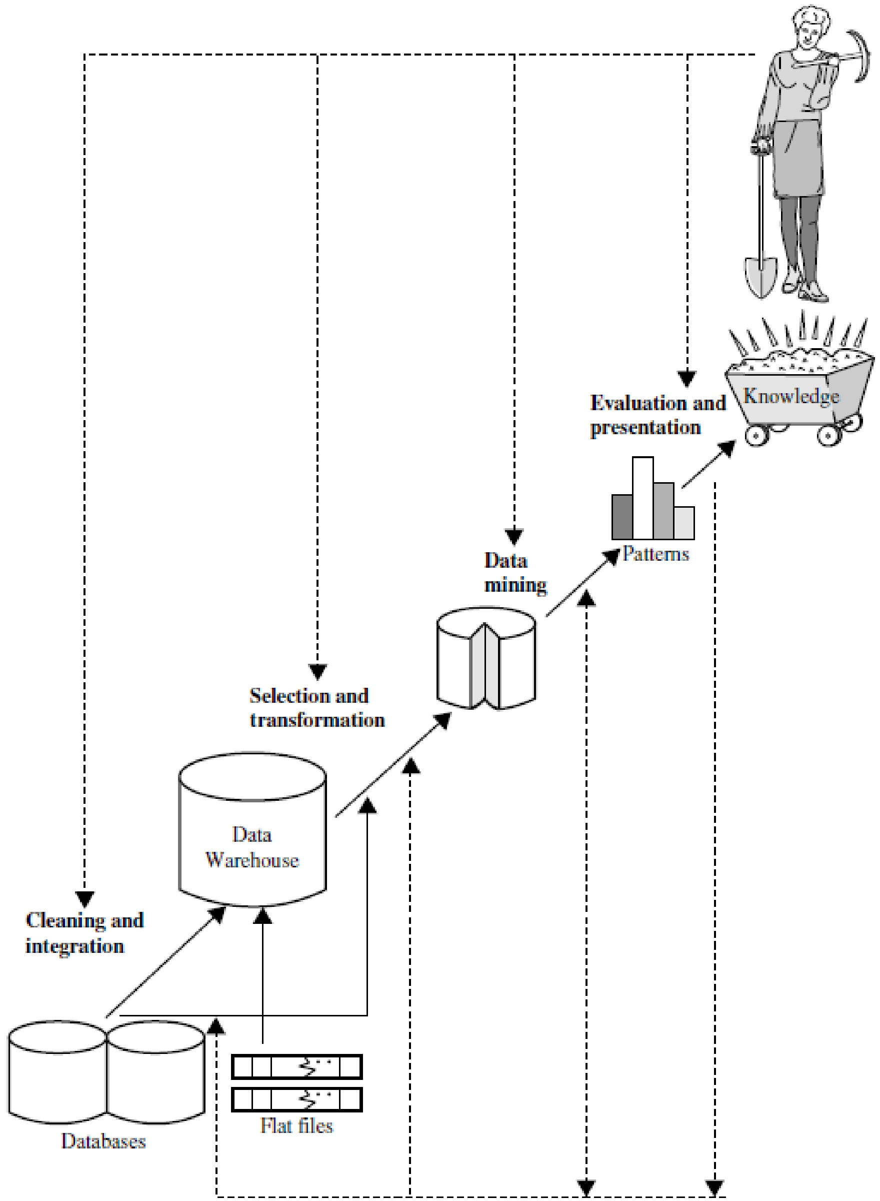
WHAT IS DATA MINING?

- Data Mining refers to extracting or “mining" knowledge from large amounts of data. “knowledge mining from data”
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Other Names:
  - Knowledge discovery (mining) in databases (KDD)
  - Knowledge extraction
  - Data/pattern analysis
  - Information harvesting
  - Business intelligence
  - etc.

WHAT IS DATA MINING?

- Knowledge discovery is an iterative sequence of the following steps:
  1. **Data Cleaning** (remove noise and inconsistent data)
  2. **Data Integration** (where multiple data sources may be combined)
  3. **Data Selection** (where data relevant to the analysis task are retrieved from the data warehouse)
  4. **Data Transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
  5. **Data Mining** (an essential process where intelligent methods are applied in order to extract data patterns i.e., knowledge discovery)
  6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
  7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the users)

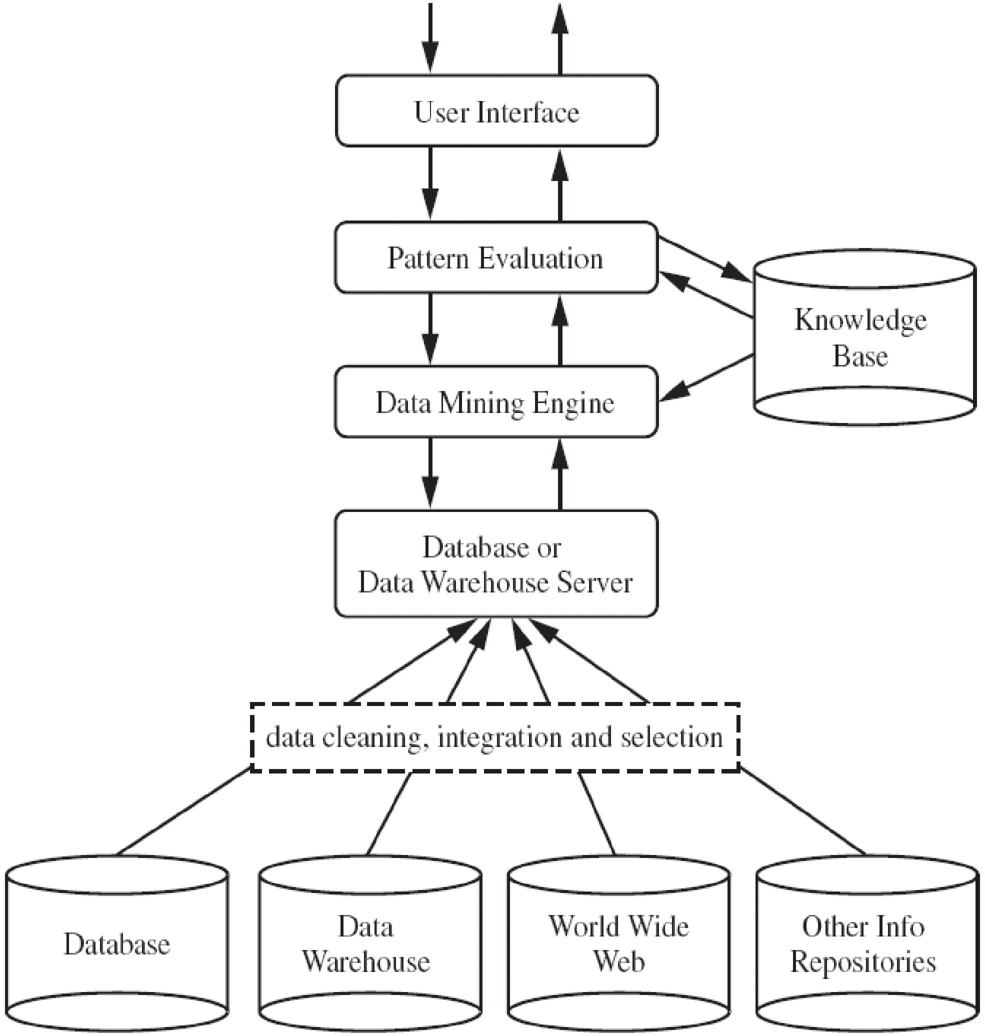
WHAT IS DATA MINING?



A TYPICAL DATA MINING SYSTEM ARCHITECTURE

- Database, data warehouse, WWW or other information repository (store data)
- Database or data warehouse server (fetch and combine data)
- Knowledge base (turn data into meaningful groups according to domain knowledge)
- Data mining engine (perform mining tasks)
- Pattern evaluation module (find interesting patterns)
- User interface (interact with the user)

A TYPICAL DATA MINING SYSTEM ARCHITECTURE



WHAT KINDS OF DATA CAN BE MINED?

- Database-oriented data sets and applications
  - Relational database, Data warehouse, Transactional database
- Advanced data sets and advanced applications
  - Object-Relational Databases
  - Temporal Databases, Sequence Databases, Time-Series databases
  - Spatial Databases and Spatiotemporal Databases
  - Text databases and Multimedia databases
  - Heterogeneous Databases and Legacy Databases
  - Data Streams
  - The World-Wide Web

RELATIONAL DATABASE

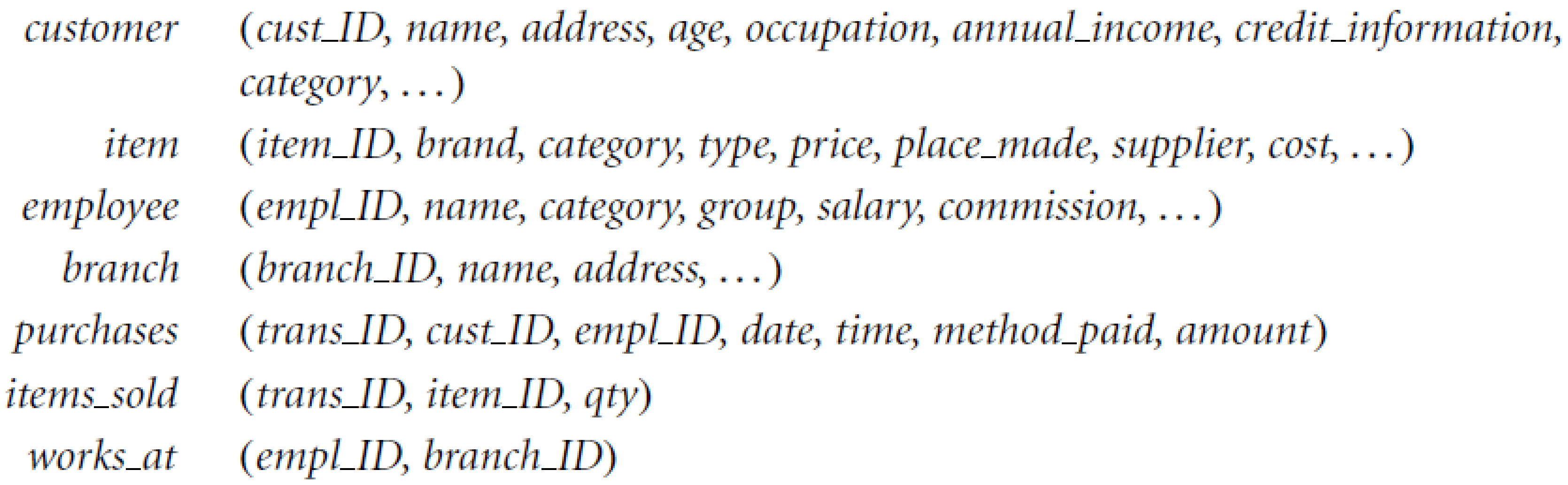
- **DBMS** – database management system, contains a collection of interrelated databases

  e.g., Faculty database, student database, publications database
- Each database contains a collection of tables and functions to manage and access the data

  e.g., student_bio, student_graduation, student_parking
- Each table contains columns and rows, with columns as attributes of data and rows as records
- Tables can be used to represent the relationships between or among multiple tables

RELATIONAL DATABASE

- ***AllElectronics Store***



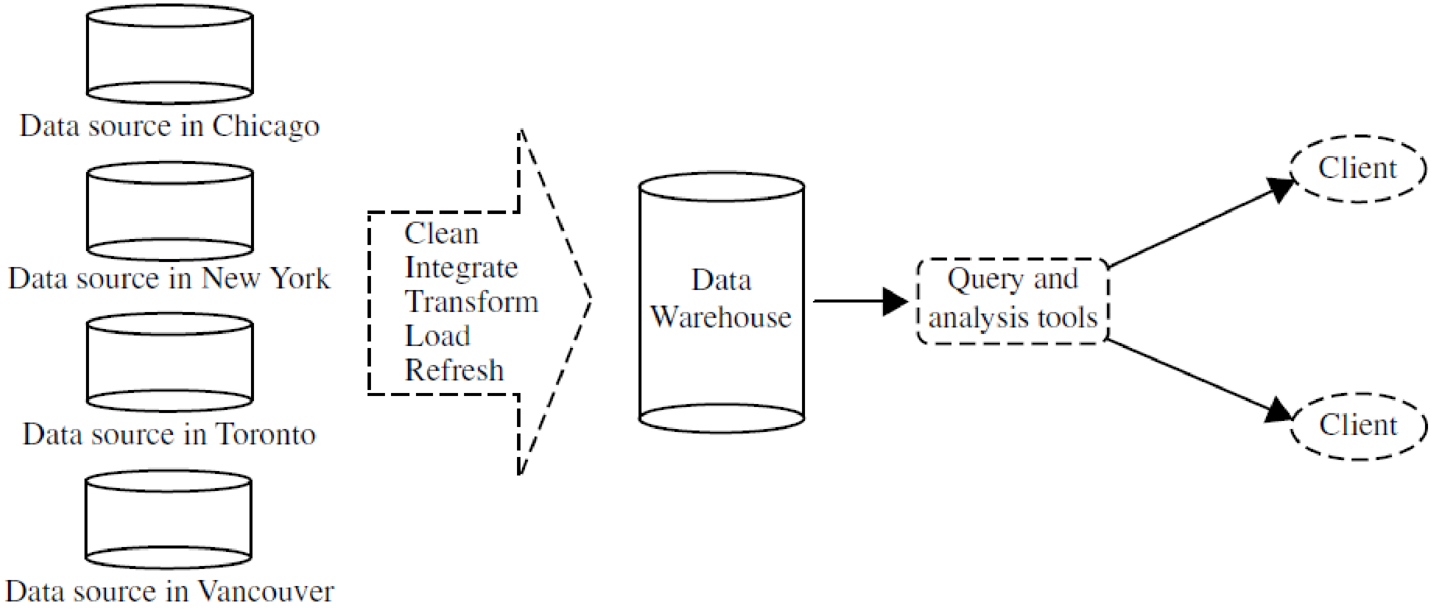
customer	*(cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .)*
item	*(item_ID, brand, category, type, price, place_made, supplier, cost, . . .)*
employee	*(empl_ID, name, category, group, salary, commission, . . .)*
branch	*(branch_ID, name, address, . . .)*
purchases	*(trans_ID, cust_ID, empl_ID, date, time, method_paid, amount)*
items_sold	*(trans_ID, item_ID, qty)*
works_at	*(empl_ID, branch_ID)*

RELATIONAL DATABASE

- With a relational query language, e.g., SQL, we will be able to find answers to questions such as:
  - How many items were sold last year?
  - Who has earned commissions higher than 10%?
  - What is the total sales of last month for Dell laptops?
- When data mining is applied to relational databases, we can search for trends or data patterns.
- Relational databases are one of the most commonly available and richest information repositories, and thus they are a major data form in the study of data mining.

DATA WAREHOUSES

- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.



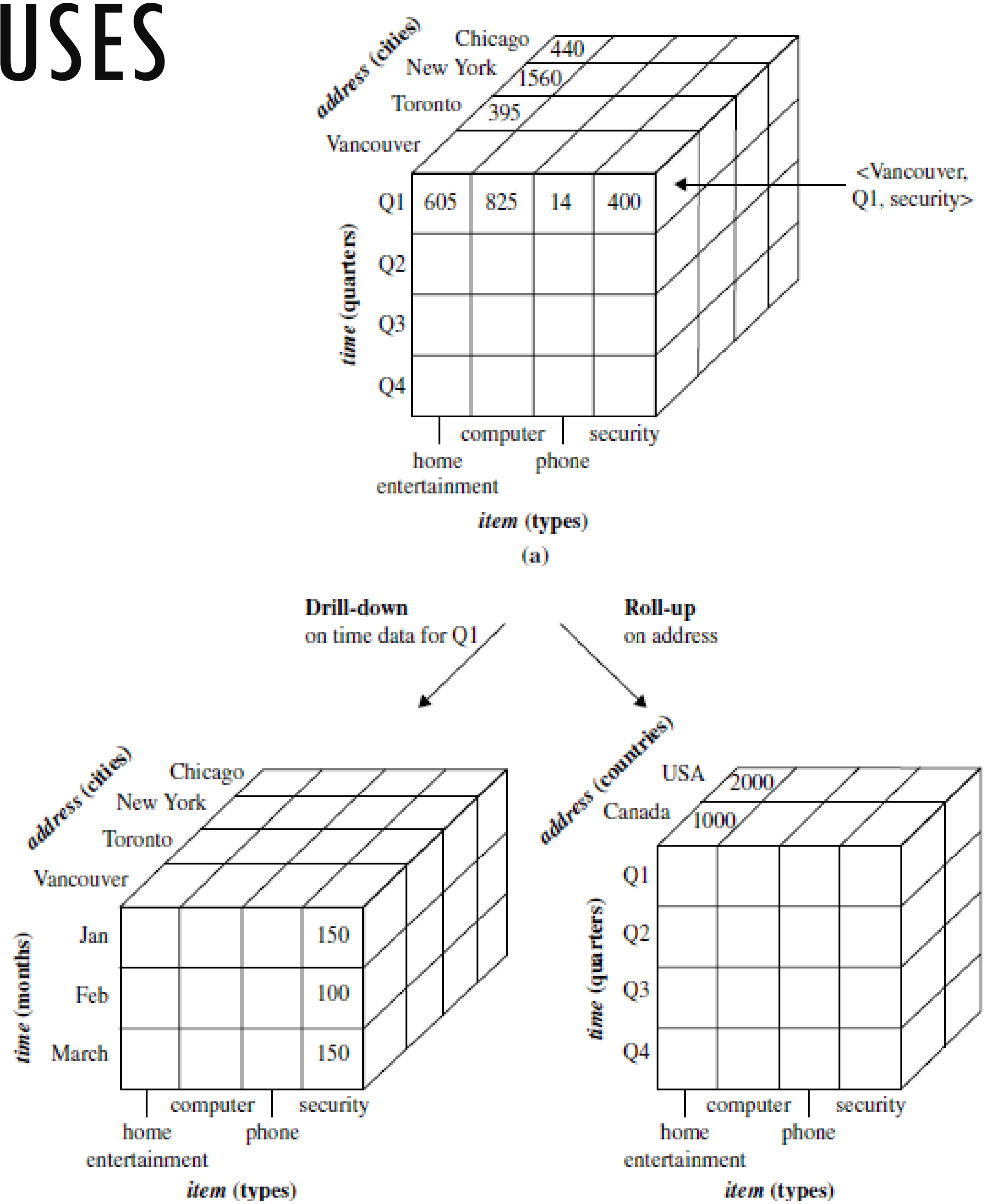
DATA WAREHOUSES

- Data are organized around major subjects, e.g., customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the transactions per item type for each store)

DATA WAREHOUSES

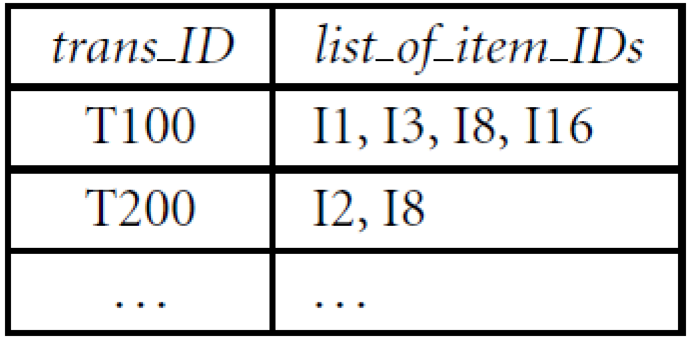
- A data warehouse is usually modeled by a multidimensional data structure, called a **data cube**, in which each dimension corresponds to an attribute or a set of attributes in the schema
- each **cell** stores the value of some aggregate measure such as *count* or *sum(sales_amount)*
- A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data
- User can perform **drill-down** or **roll-up** operations to view the data at different degrees of summarization

DATA WAREHOUSES



TRANSACTIONAL DATABASES

- Consists of a file where each record represents a transaction
- A transaction typically includes a unique transaction ID and a list of the items making up the transaction.



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

- Either stored in a flat file or unfolded into relational tables
- Easy to identify items that are frequently sold together

DATA MINING FUNCTIONALITIES

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- In general, data mining tasks can be classified into two categories:
  - **Descriptive** (characterize the general properties of the data in the database)
  - **Predictive** (perform inference on the current data in order to make predictions)

DATA MINING FUNCTIONALITIES

- Main mining functionalities are described as follows:
  1. Concept/Class description: Characterization and Discrimination
  2. Association analysis
  3. Classification and prediction
  4. Clustering analysis
  5. Evolution and Deviation analysis

CONCEPT/CLASS DESCRIPTION: CHARACTERIZATION AND DISCRIMINATION

- Data can be associated with classes or concepts.
  - Classes of items – computers, printers, …
  - Concepts of customers – bigSpenders, budgetSpenders, …
  - How to describe these items or concepts?
- These descriptions can be derived via :
  - **Data characterization** – summarizing the general characteristics of a target class of data
    - E.g., summarizing the characteristics of customers who spend more than $1,000 a year at *AllElectronics*. Result can be a general profile of the customers, such as 40 – 50 years old, employed, 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** – comparing the target class with one or a set of comparative classes
    - E.g., 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 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.
  - Or both of the above

MINING FREQUENT PATTERNS, ASSOCIATIONS, AND CORRELATIONS

- **Association analysis:** is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data.
- Typical example is market basket analysis.
  - **Frequent itemset:** a set of items that frequently appear together in a transactional data set (e.g. milk and bread)
  - **Frequent subsequence:** a pattern that customers tend to purchase product A, followed by a purchase of product B

MINING FREQUENT PATTERNS, ASSOCIATIONS, AND CORRELATIONS

- A sample analysis result – an association rule:

  buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%]

  - (if a customer buys a computer, there is a 50% chance that she will buy software. 1% of all of the transactions under analysis showed that computer and software are purchased together.)
  - Associations rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.

- **Correlation Analysis**: additional analysis to find statistical correlations between associated pairs

CLASSIFICATION AND PREDICTION

- **Classification**
  - The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
  - The derived model is based on the analysis of a set of training data (data objects whose class label is known).
  - The model can be represented in classification (IF-THEN) rules, decision trees, neural networks, etc.
- **Prediction**
  - Predict missing or unavailable numerical data values

CLASSIFICATION AND PREDICTION

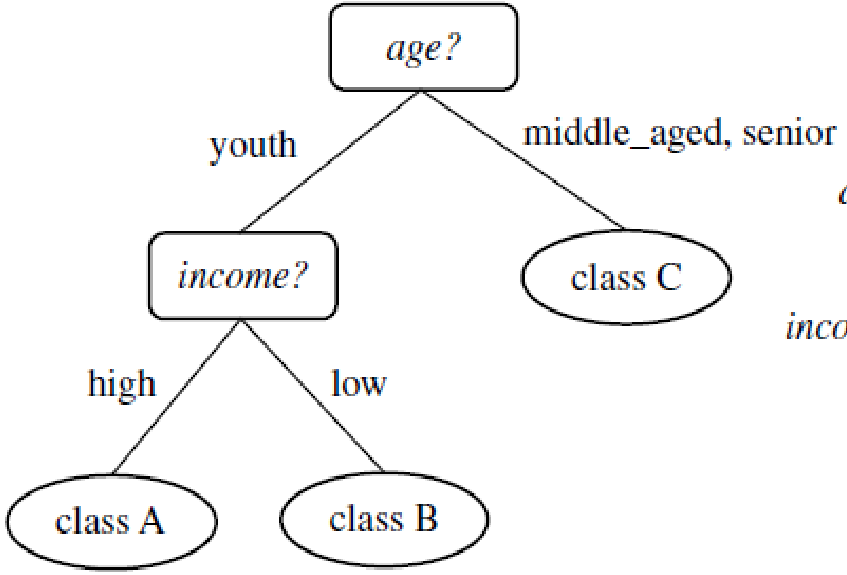
age(X, "youth") AND income(X, "high") → *class(X, "A")*

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

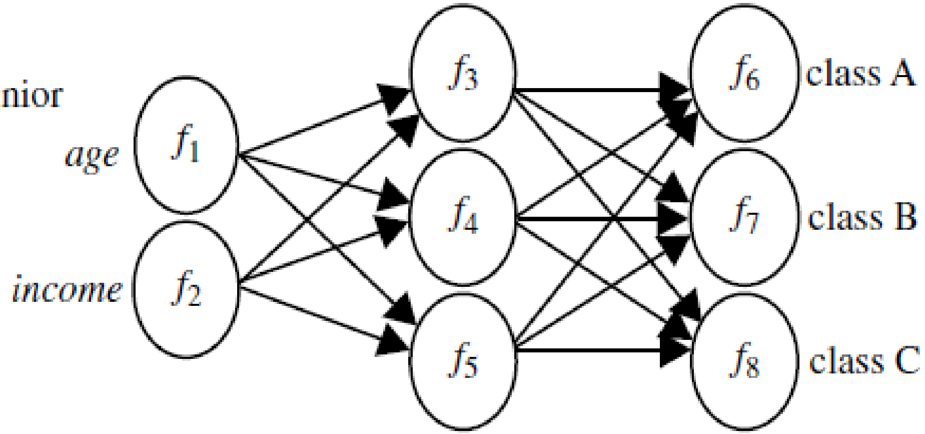
age(X, "middle_aged") → *class(X, "C")*

age(X, "senior") → *class(X, "C")*

(a)



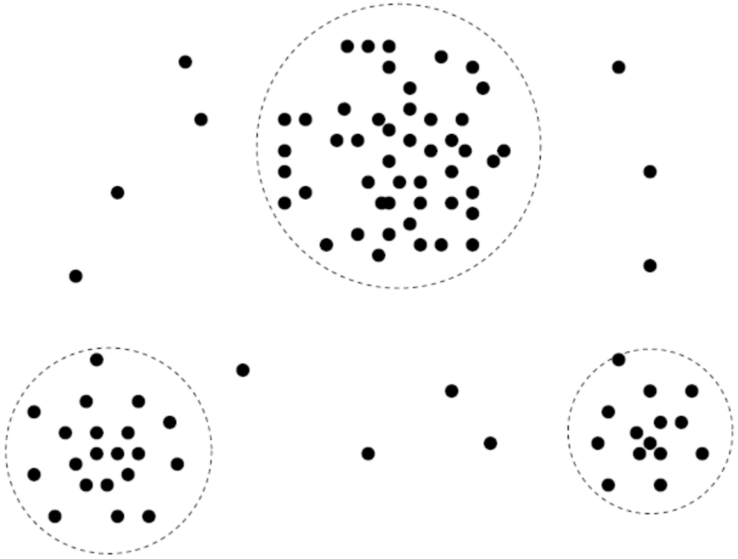
(b)



(c)

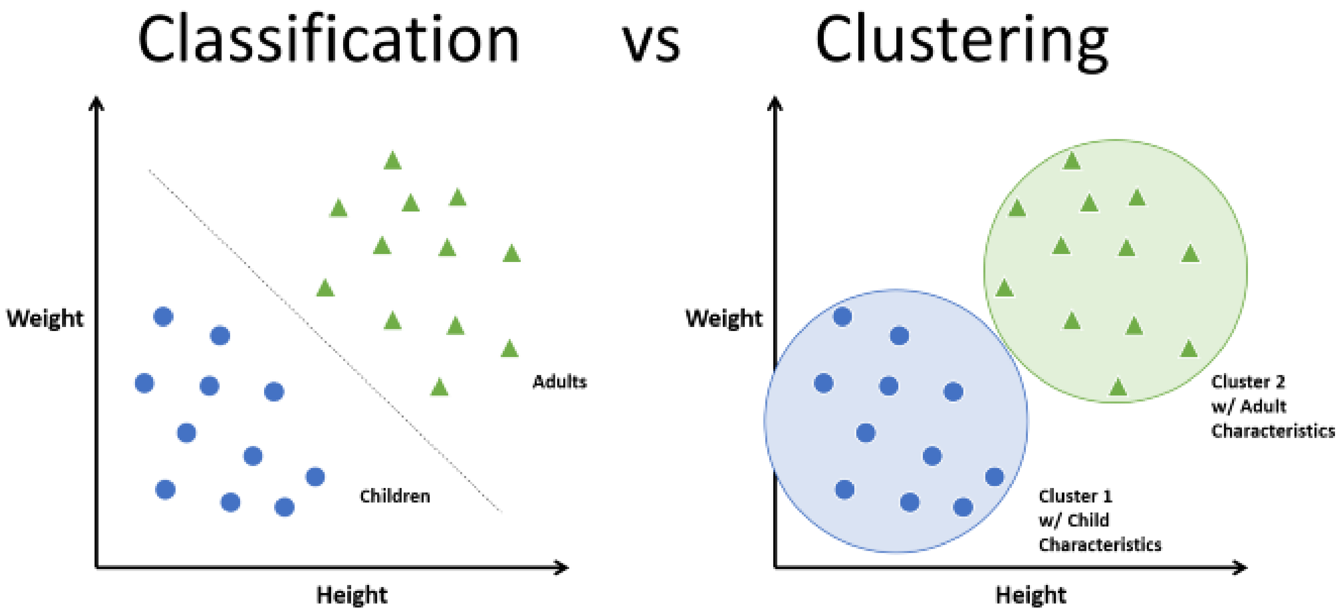
CLUSTER ANALYSIS

- Class label is unknown: group data to form new classes
- Clusters of objects are formed based on the principle of maximizing intra-class similarity & minimizing interclass similarity
  - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.



CLASSIFICATION VS CLUSTERING

Classification	Clustering
Model predicts a probable classification for a given input	Model maps a given input into one of the data clusters
Uses supervised learning	Uses unsupervised learning
Labeled data required for training	Does not require labeled data
Number of categories is known	Number of groups is unknown
Two-step process to train and predict	Single-step process



EVOLUTION AND DEVIATION ANALYSIS

- **Outlier Analysis**
  - Data that do no comply with the general behavior or model.
  - Outliers are usually discarded as noise or exceptions.
  - Useful for fraud detection.
    - E.g., Detect purchases of extremely large amounts
- **Evolution Analysis**
  - Describes and models regularities or trends for objects whose behavior changes over time.
  - E.g. Identify stock evolution regularities for overall stocks and for the stocks of particular companies.

ARE ALL PATTERNS INTERESTING?

- Data mining may generate thousands of patterns: **Not all of them are interesting**
- A pattern is interesting if it is:
  - easily understood by humans
  - valid on new or test data with some degree of certainty
  - potentially useful
  - Novel “unfamiliar”
  - validates some hypothesis that a user seeks to confirm
- An interesting measure represents **knowledge**!

PATTERN INTERESTINGNESS MEASURES

- **Objective measures**
  - Based on **statistics** and **structures** of patterns
    - e.g., support, confidence, etc. (Rules that do not satisfy a threshold are considered uninteresting.)
- **Subjective measures**
  - Reflect the **needs** and **interests** of a particular user.
    - E.g. A marketing manager is only interested in characteristics of customers who shop frequently
  - Based on user's **belief** in the data.
    - e.g., Patterns are interesting if they are unexpected, or can be used for strategic planning, etc
- Objective and subjective measures need to be combined.

ARE ALL PATTERNS INTERESTING?

- Find all the interesting patterns: Completeness
  - Unrealistic and inefficient
  - User-provided constraints and interestingness measures should be used
- Search for only interesting patterns: An optimization problem
  - Highly desirable
  - No need to search through the generated patterns to identify truly interesting ones.
  - Measures can be used to rank the discovered patterns according their interestingness.

MAJOR ISSUES IN DATA MINING

1. Mining methodology and User interaction
   - ○ Mining different kinds of knowledge
     - DM should cover a wide spectrum of data analysis and knowledge discovery tasks
     - Require the development of numerous data mining techniques
   - ○ Interactive mining of knowledge at multiple levels of abstraction
     - Difficult to know exactly what will be discovered
   - ○ Incorporation of background knowledge
     - Allow discovered patterns to be expressed in concise terms and different levels of abstraction
   - ○ Data mining query languages
     - High-level query languages need to be developed
     - Should be integrated with a DB/DW query language
   - ○ Presentation and visualization of results
     - Knowledge should be easily understood and directly usable
   - ○ Handling noisy or incomplete data
     - Require data cleaning methods and data analysis methods that can handle noise
   - ○ Pattern evaluation – the interestingness problem
     - How to develop techniques to access the interestingness of discovered patterns, especially with subjective measures bases on user beliefs or expectations

MAJOR ISSUES IN DATA MINING

2. **Performance Issues**
   - Efficiency and scalability
     - Huge amount of data
     - Running time must be predictable and acceptable
   - Parallel, distributed and incremental mining algorithms
     - Divide the data into partitions and processed in parallel
     - Incorporate database updates without having to mine the entire data again from scratch

3. **Diversity of Database Types**
   - Other database that contain complex data objects, multimedia data, spatial data, etc.
   - Expect to have different DM systems for different kinds of data
   - Heterogeneous databases and global information systems
     - Web mining becomes a very challenging and fast-evolving field in data mining

End of Chapter 1

THANK YOU