Data Mining

Module 4

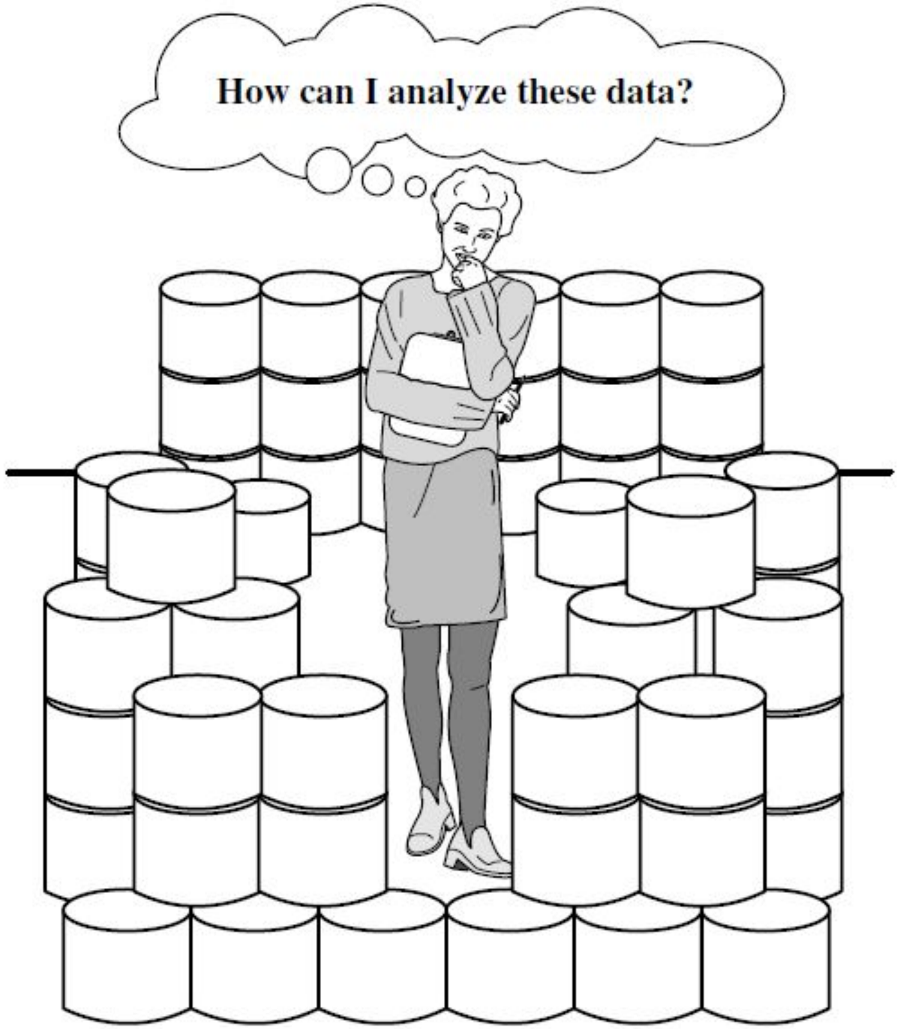


Figure 1.2 The world is data rich but information poor.

Figure 1.3 Data mining—searching for knowledge (interesting patterns) in data.

What is Data Mining?

- Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets.

What is a data warehouse?

A data warehouse, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI) and machine learning.

What Is Data Mining?

- Knowledge mining from data
- Knowledge extraction
- Data/pattern analysis
- Data archaeology, and
- Data dredging.

The knowledge discovery process

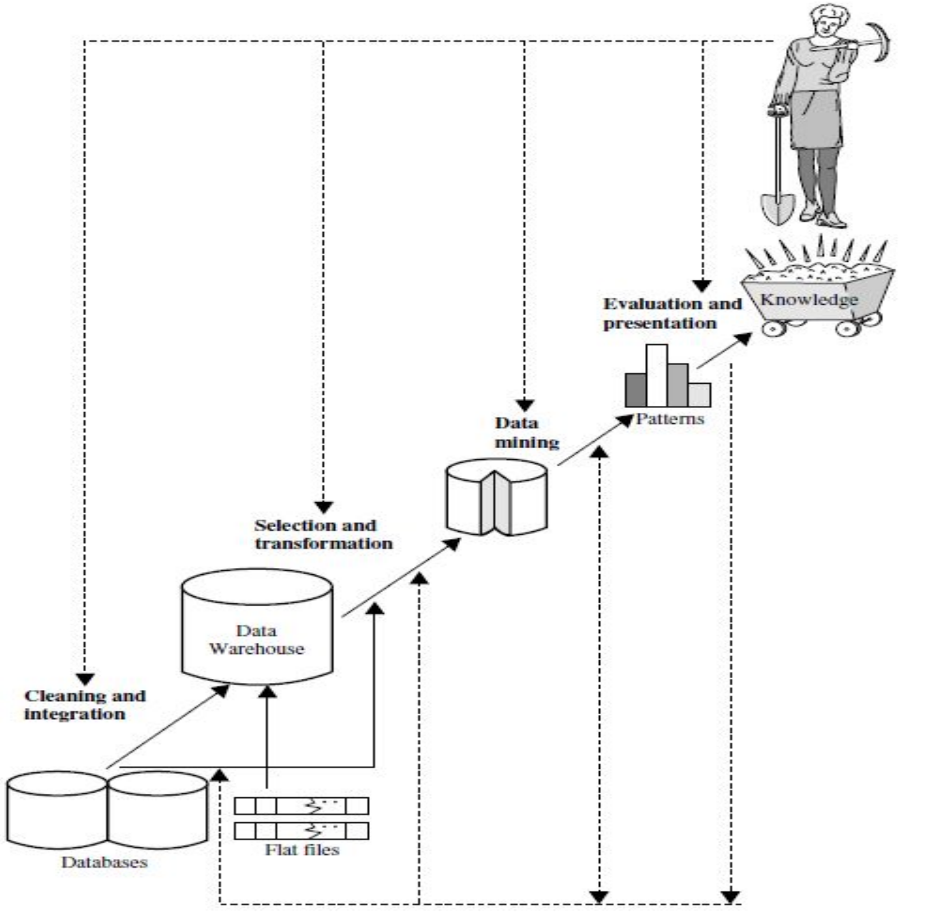
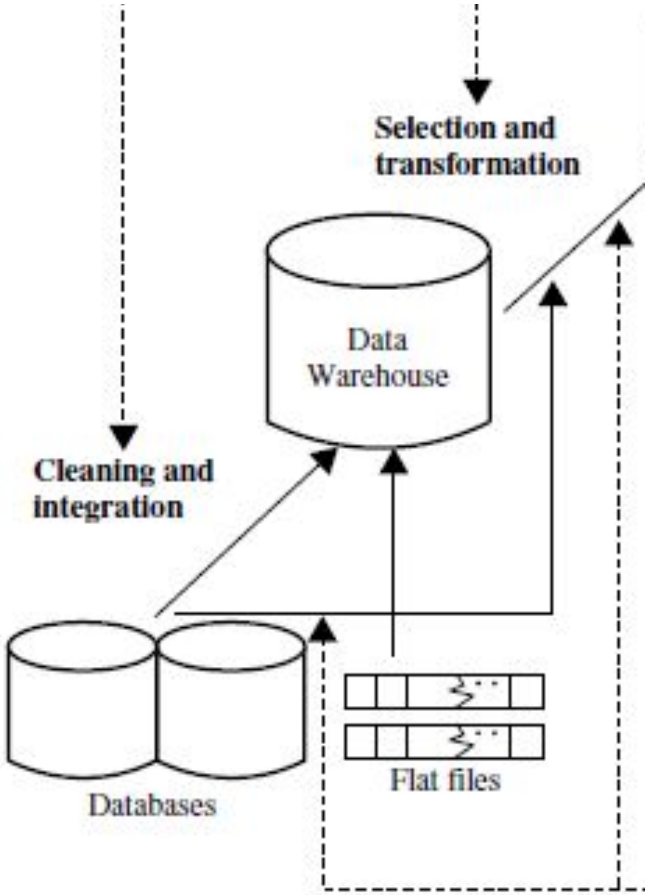


Figure 1.4 Data mining as a step in the process of knowledge discovery.

The knowledge discovery process



The knowledge discovery process

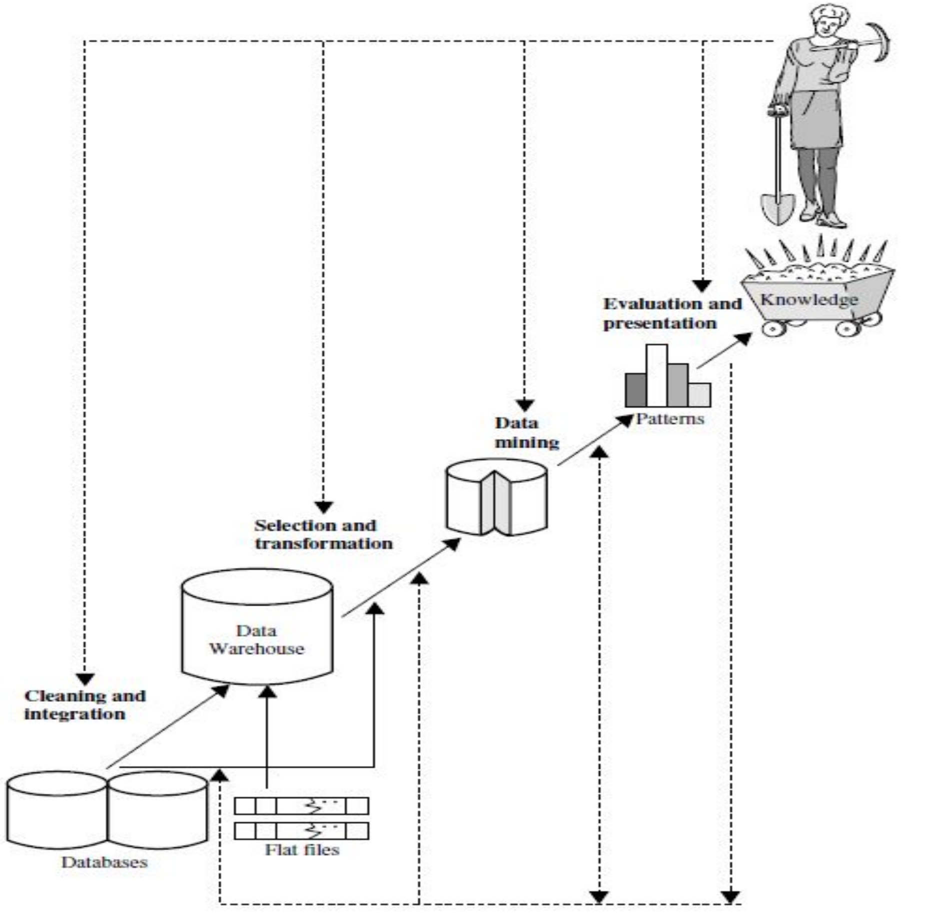
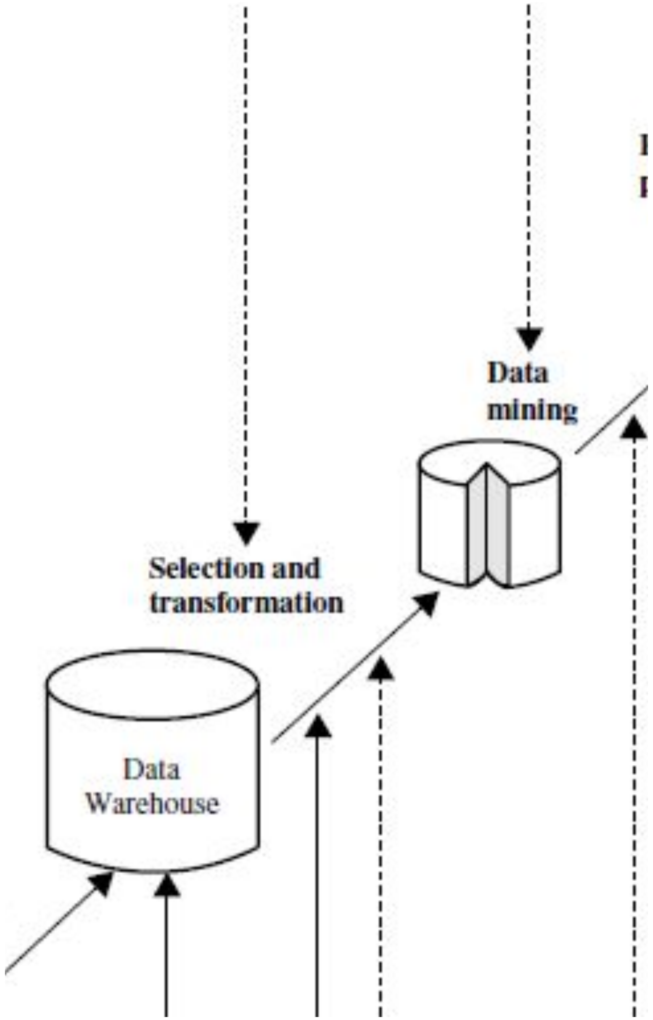


Figure 1.4 Data mining as a step in the process of knowledge discovery.

The knowledge discovery process



The knowledge discovery process

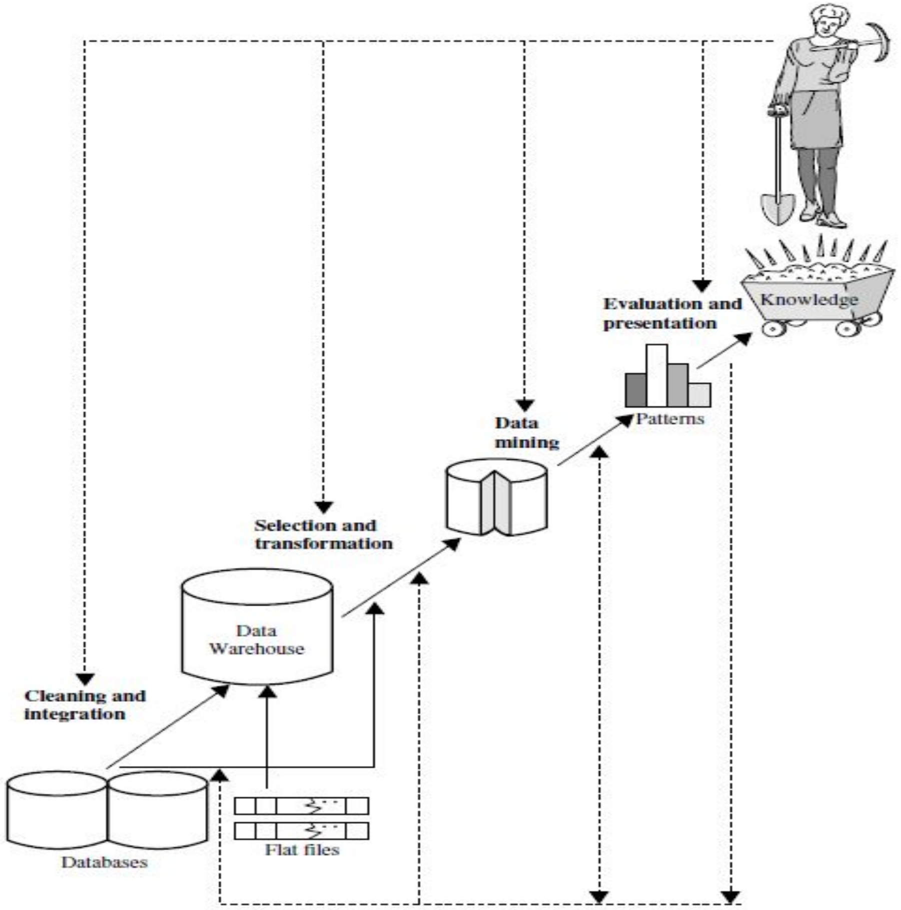
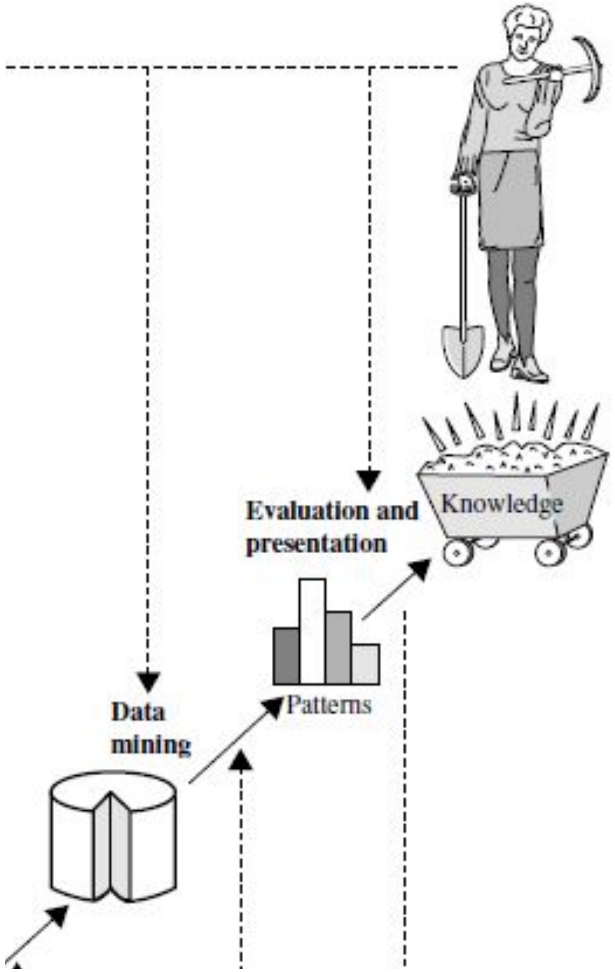


Figure 1.4 Data mining as a step in the process of knowledge discovery.

The knowledge discovery process



The knowledge discovery process

1. Data cleaning (to 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 database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

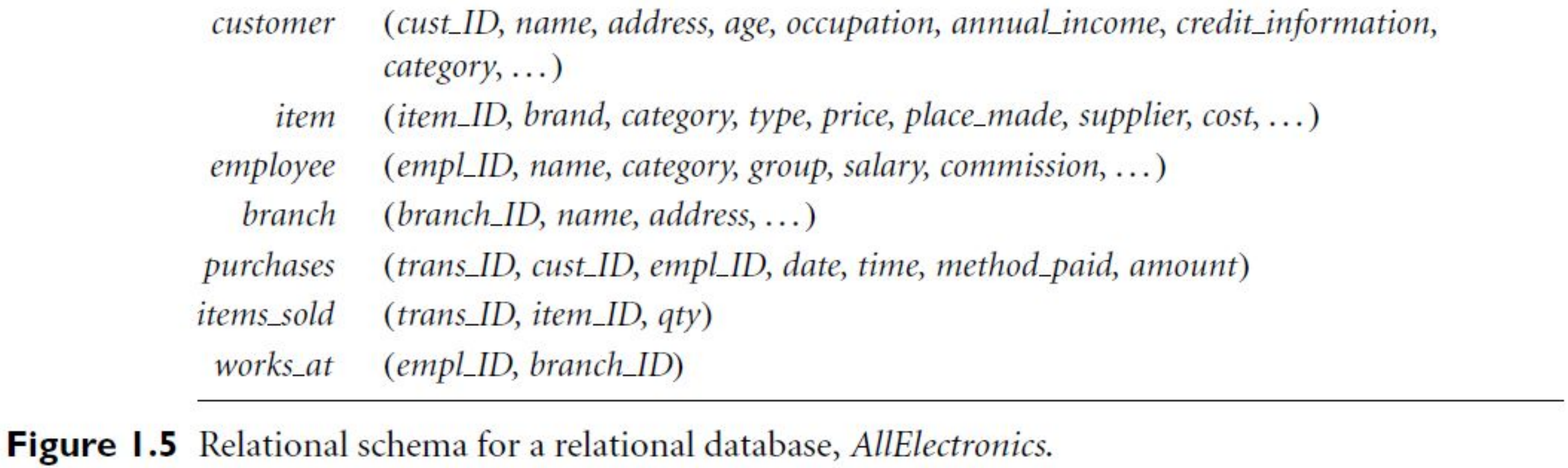
Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

What Kinds of Data Can Be Mined?

1. **Database Data:**
   a. A 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.
   b. A **relational database** is a collection of **tables**, each of which is assigned a unique name.
   c. Each table consists of a set of **attributes** (columns or fields) and usually stores a large set of **tuples** (records or rows).
   d. Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
   e. A semantic data model, such as **an entity-relationship (ER)** data model, is often constructed for relational databases.
   f. An ER data model represents the database as a set of entities and their relationships.

A relational database for AllElectronics



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*)

Figure 1.5 Relational schema for a relational database, *AllElectronics*.

- 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.
- When mining relational databases, we can go further by searching for trends or data patterns.
- For example, data mining systems can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information.

DataWarehouses

- 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.
- 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, and 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.

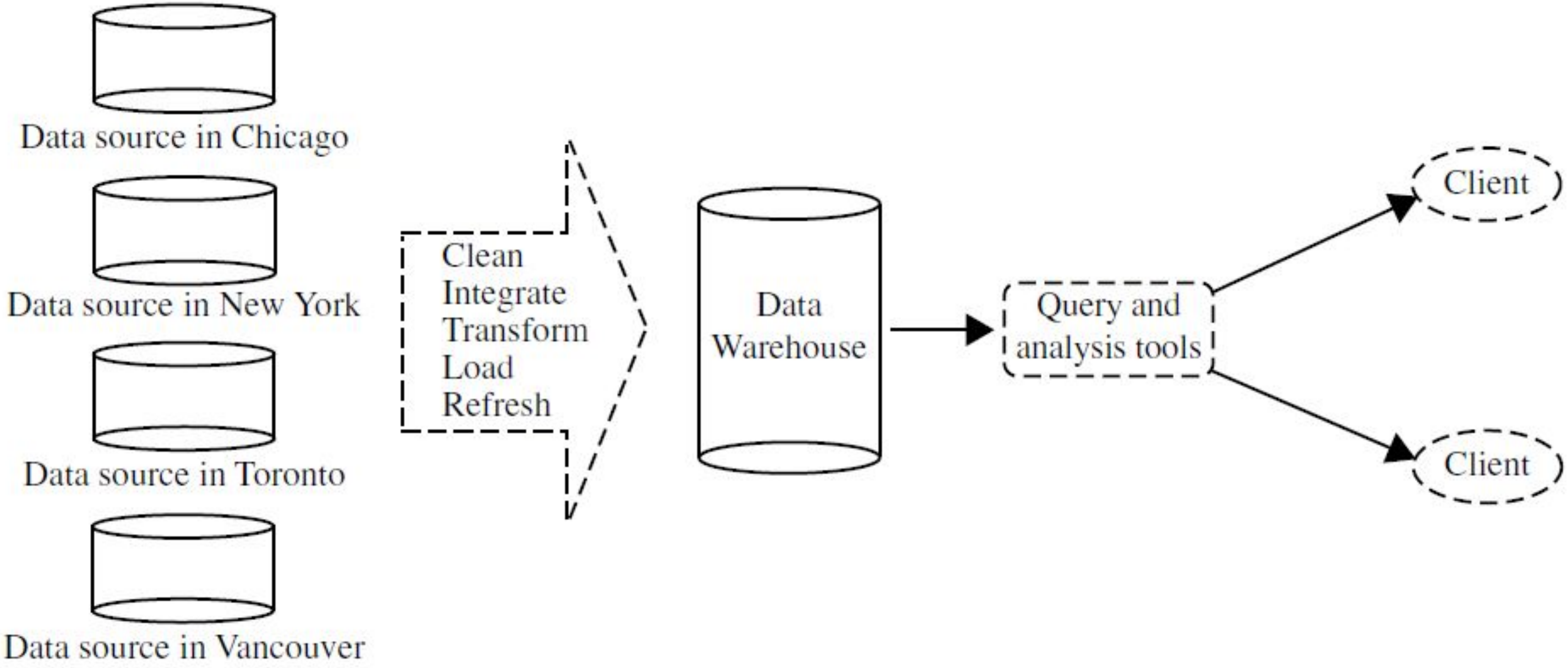


Figure 1.6 Typical framework of a data warehouse for *AllElectronics.*

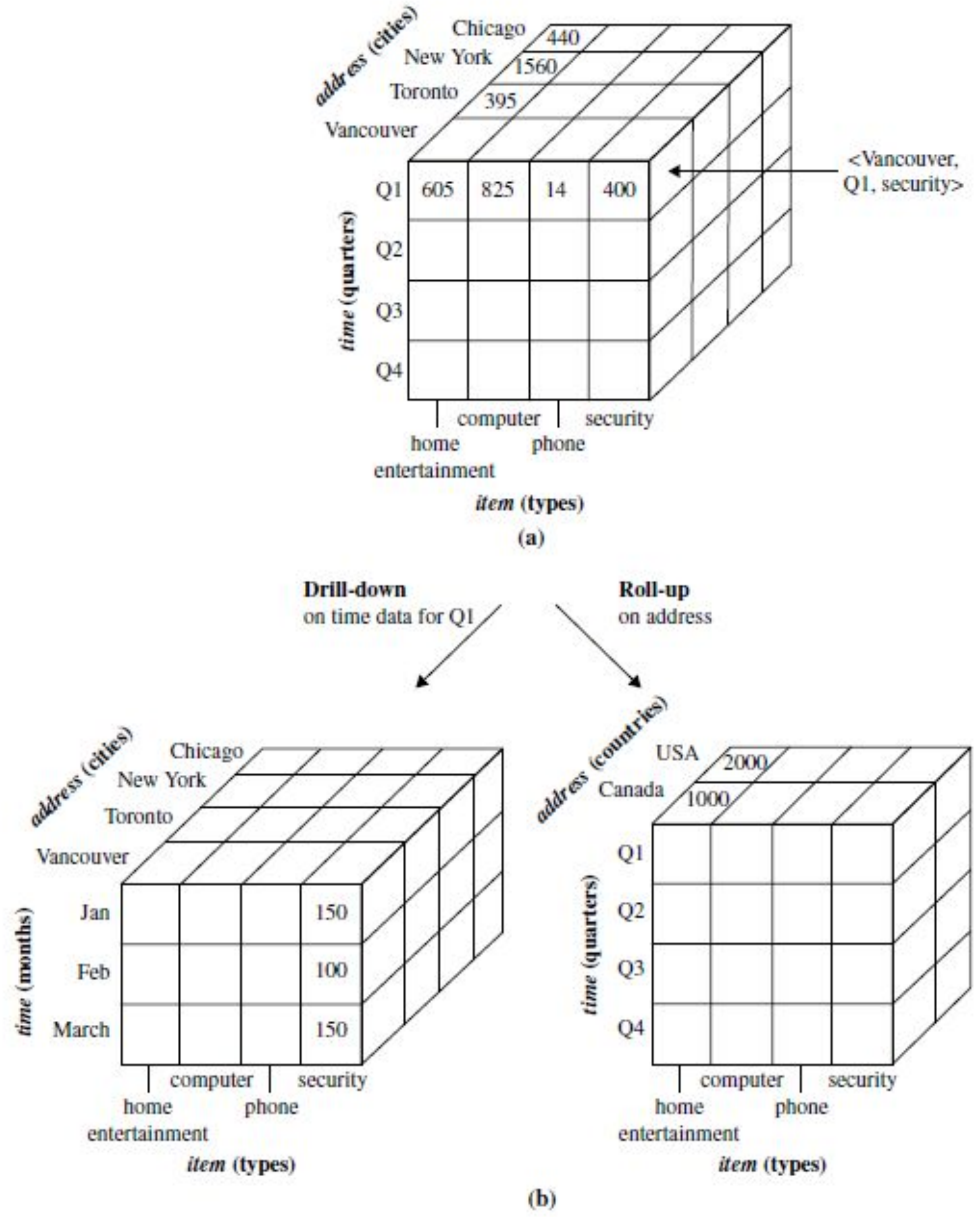
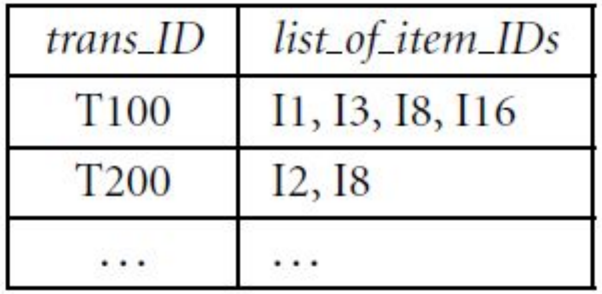


Figure 1.7 A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for *AllElectronics* and (b) showing summarized data resulting from drill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.

Transactional Data

- In general, 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.
- A transactional database may 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.



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

Figure 1.8 Fragment of a transactional database for sales at *AllElectronics*.

Other kind 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 or 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),
  - hypertext and multimedia data (including text, image, video, and audio data),
  - graph and networked data (e.g., social and information networks), and
  - the Web (a huge, widely distributed information repository made available by the Internet).

Other kind of Data

- These applications bring about new challenges, like
    - how to handle data carrying special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity), and
    - how to mine patterns that carry rich structures and semantics.

What Kinds of Patterns Can Be Mined?

- Data mining functionalities
  - 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.
  - Descriptive mining tasks characterize properties of the data in a target data set.
  - Predictive mining tasks perform induction on the current data in order to make predictions.

Class/Concept Description: Characterization and Discrimination

- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms.
- Such descriptions of a class or a concept are called class/concept descriptions.
- These descriptions can be derived using
  (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms
  (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes)
  (3) both data characterization and discrimination.

Data characterization

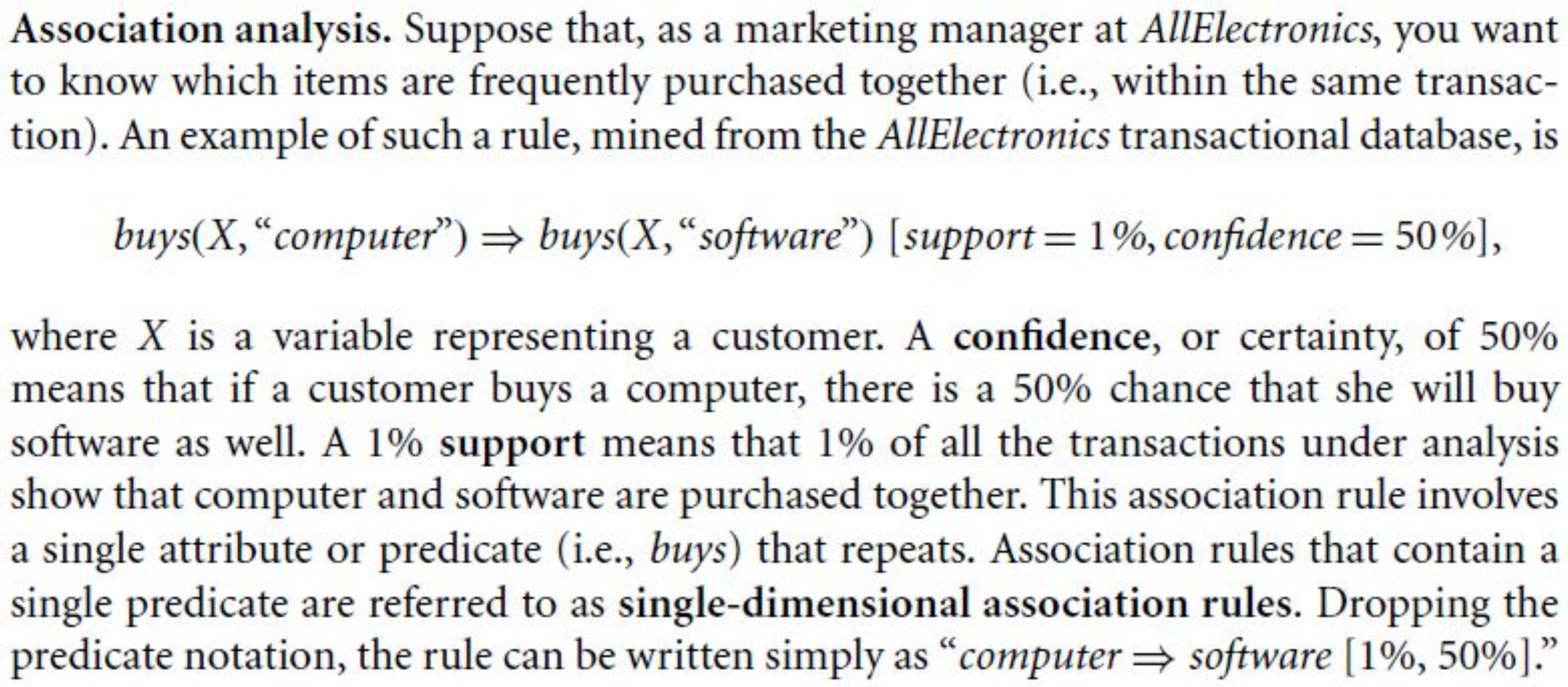
- Summarization of the general characteristics or features of a target class of data.
- The data corresponding to the user-specified class are typically collected by a query.
    - For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.
- The output of data characterization can be presented in various forms.
    - Examples include **pie charts**, **bar charts**, **curves**, **multidimensional data cubes**, and **multidimensional tables**, including crosstabs.
- The resulting descriptions can also be presented as **generalized relations** or in rule form (called **characteristic rules**).

Data discrimination

- 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.
    - 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.
- The methods used for data discrimination are similar to those used for data characterization.
- Discrimination descriptions expressed in the form of rules are referred to as **discriminant rules**.

Mining Frequent Patterns, Associations, and Correlations

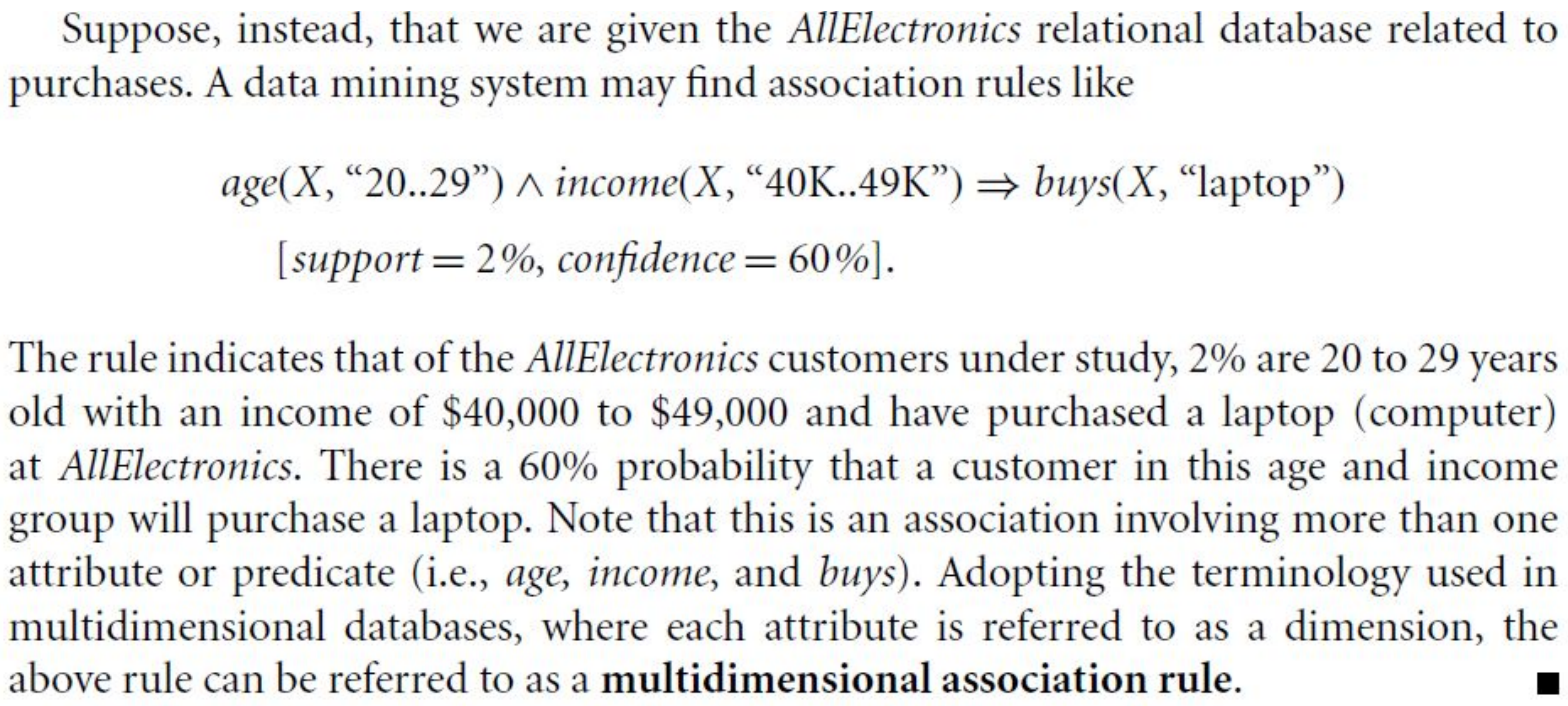
- **Frequent patterns** are patterns that occur frequently in data.
- There are many kinds of frequent patterns, including **frequent itemsets**, **frequent subsequences** (also known as sequential patterns), and **frequent substructures**.
- A *frequent itemset* typically refers to a set of items that often appear together in a transactional data set—for example, 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) that may be combined with itemsets or subsequences. 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 analysis. Suppose that, as a marketing manager at *AllElectronics*, you want to know which items are frequently purchased together (i.e., within the same transaction). An example of such a rule, mined from the *AllElectronics* transactional database, is

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

where X is a variable representing a customer. A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as “*computer* ⇒ *software* [1%, 50%].”



Suppose, instead, that we are given the *AllElectronics* relational database related to purchases. A data mining system may find association rules like

$$age(X, \text{“20..29”}) \wedge income(X, \text{“40K..49K”}) \Rightarrow buys(X, \text{“laptop”})$$
$$[support = 2\%, confidence = 60\%].$$

The rule indicates that of the *AllElectronics* customers under study, 2% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer) at *AllElectronics*. There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., *age*, *income*, and *buys*). Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule**. ■

Mining Frequent Patterns, Associations, and Correlations

- 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.
- *Frequent itemset mining* is a fundamental form of frequent pattern mining.

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.

“*How is the derived model presented?*” The derived model may be represented in various forms, such as *classification rules* (i.e., *IF-THEN rules*), *decision trees*, *mathematical formulae*, or *neural networks* (Figure 1.9). A **decision tree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily

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”*)*

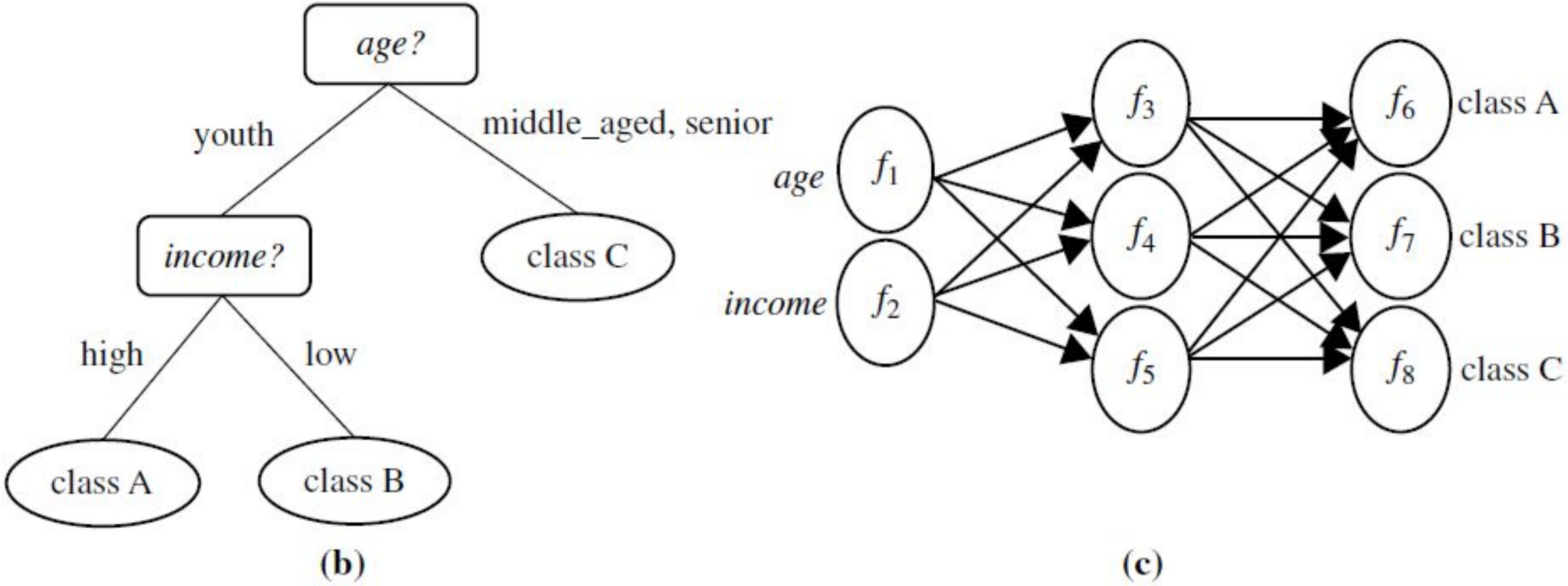


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Cluster Analysis

- Unlike classification and regression, which analyze class-labeled (training) data sets, **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*.
- That is, 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.

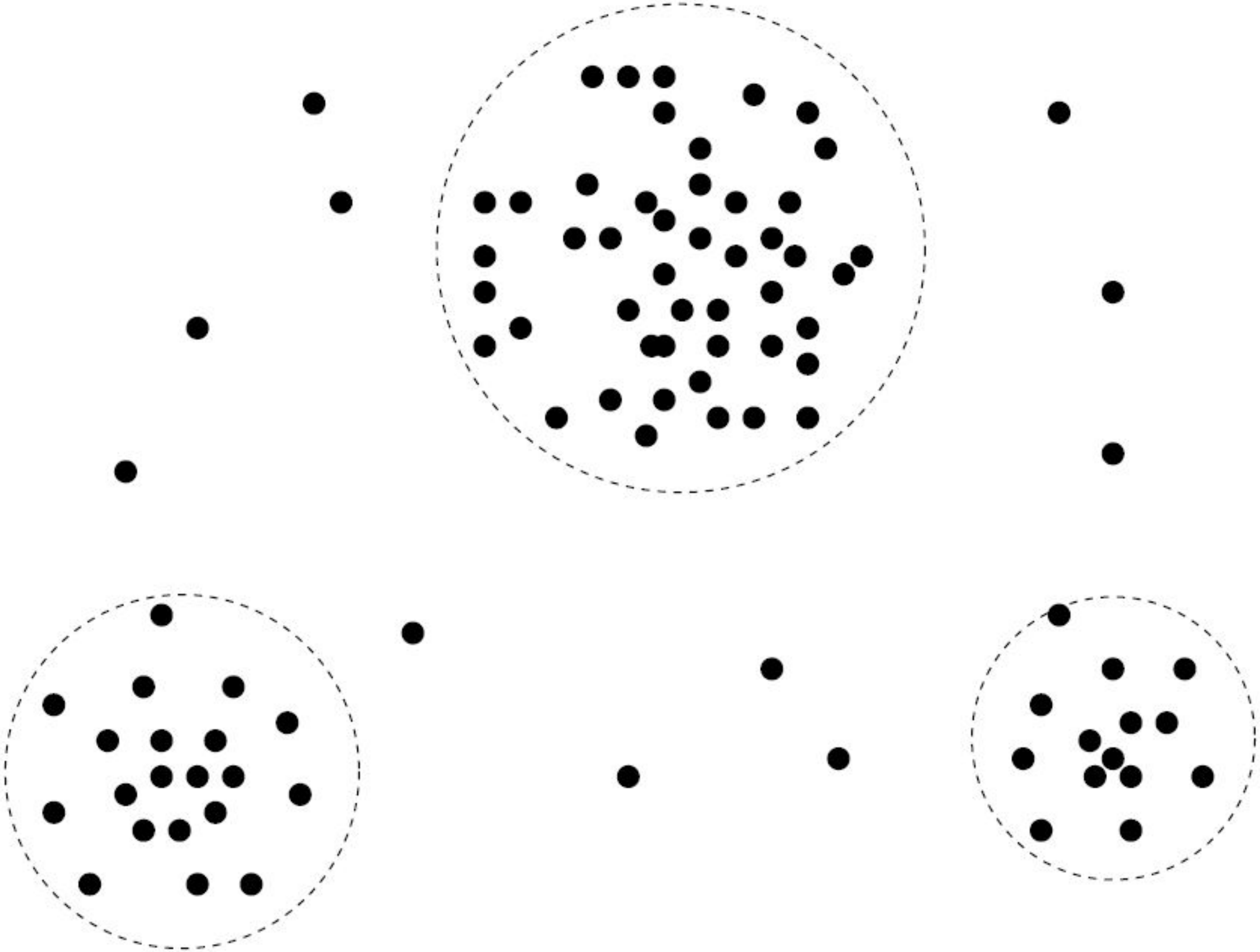


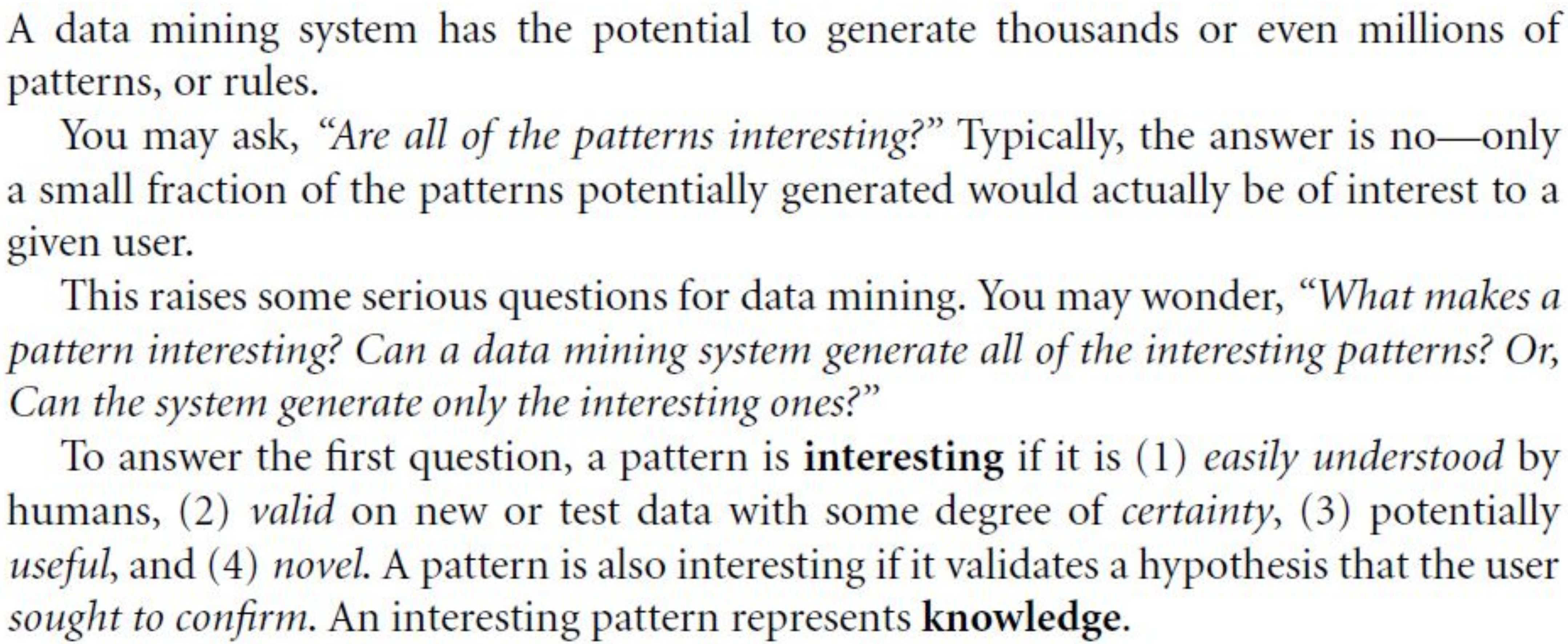
Figure 1.10 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

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) the rare events can be more interesting than the more regularly occurring ones.
- The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining**.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
- Rather than using statistical or distance measures, density-based methods may identify outliers in a local region, although they look normal from a global statistical distribution view.

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. ■

Are All Patterns Interesting?



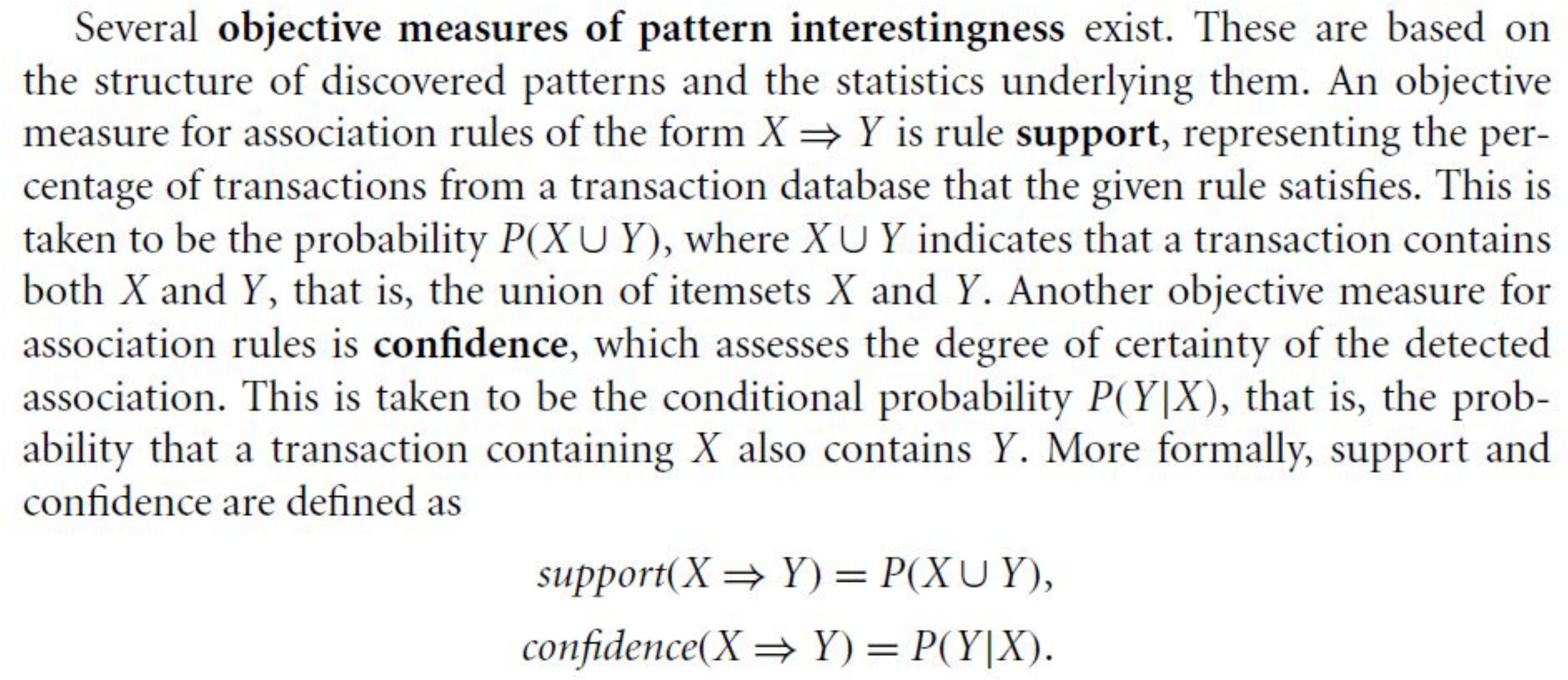
A data mining system has the potential to generate thousands or even millions of patterns, or rules.

You may ask, *"Are all of the patterns interesting?"* Typically, the answer is no—only a small fraction of the patterns potentially generated would actually be of interest to a given user.

This raises some serious questions for data mining. You may wonder, *"What makes a pattern interesting? Can a data mining system generate all of the interesting patterns? Or, Can the system generate only the interesting ones?"*

To answer the first question, a pattern is **interesting** if it is (1) *easily understood* by humans, (2) *valid* on new or test data with some degree of *certainty*, (3) potentially *useful*, and (4) *novel*. A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*. An interesting pattern represents **knowledge**.

Are All Patterns Interesting?

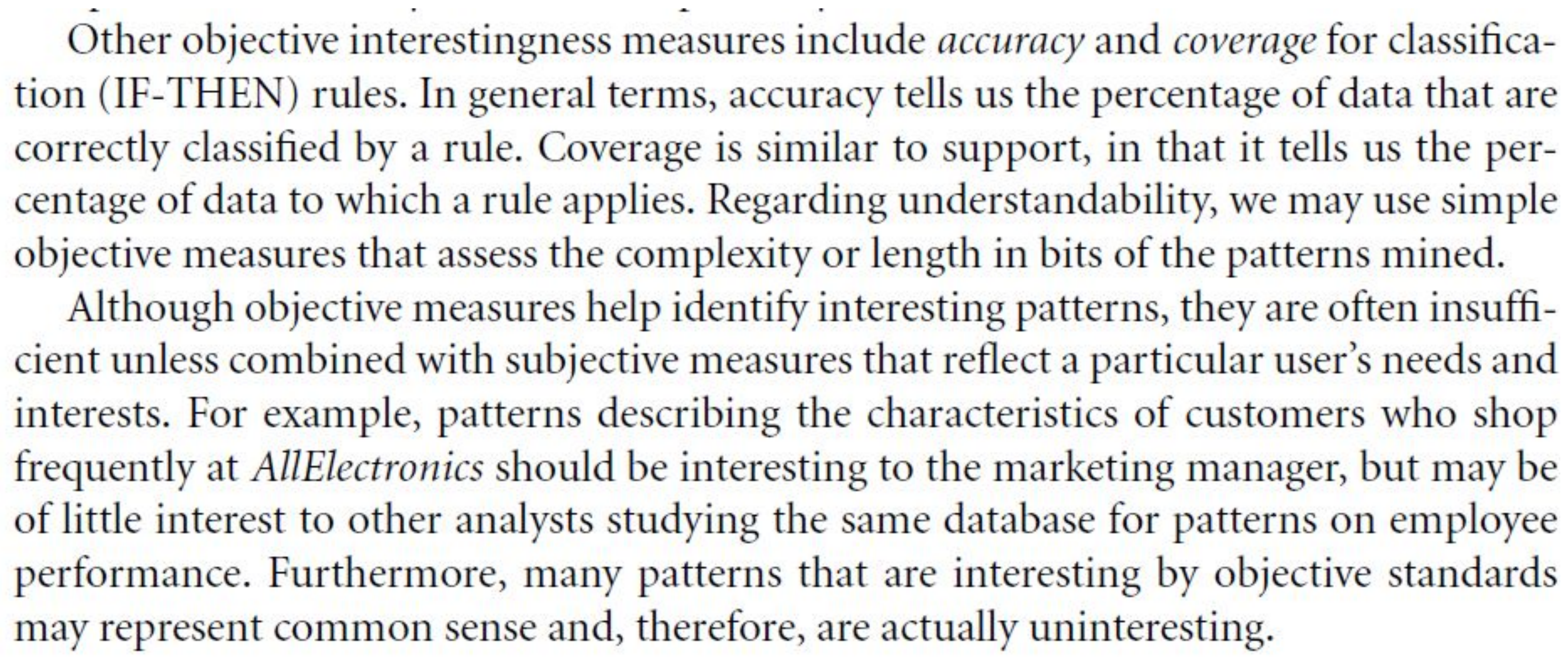


Several **objective measures of pattern interestingness** exist. These are based on the structure of discovered patterns and the statistics underlying them. An objective measure for association rules of the form $X \Rightarrow Y$ is rule **support**, representing the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability $P(X \cup Y)$, where $X \cup Y$ indicates that a transaction contains both X and Y, that is, the union of itemsets X and Y. Another objective measure for association rules is **confidence**, which assesses the degree of certainty of the detected association. This is taken to be the conditional probability $P(Y|X)$, that is, the probability that a transaction containing X also contains Y. More formally, support and confidence are defined as

$$support(X \Rightarrow Y) = P(X \cup Y),$$

$$confidence(X \Rightarrow Y) = P(Y|X).$$

Are All Patterns Interesting?



Other objective interestingness measures include *accuracy* and *coverage* for classification (IF-THEN) rules. In general terms, accuracy tells us the percentage of data that are correctly classified by a rule. Coverage is similar to support, in that it tells us the percentage of data to which a rule applies. Regarding understandability, we may use simple objective measures that assess the complexity or length in bits of the patterns mined.

Although objective measures help identify interesting patterns, they are often insufficient unless combined with subjective measures that reflect a particular user's needs and interests. For example, patterns describing the characteristics of customers who shop frequently at *AllElectronics* should be interesting to the marketing manager, but may be of little interest to other analysts studying the same database for patterns on employee performance. Furthermore, many patterns that are interesting by objective standards may represent common sense and, therefore, are actually uninteresting.

Subjective interestingness measures are based on user beliefs in the data. These measures find patterns interesting if the patterns are **unexpected** (contradicting a user's belief) or offer strategic information on which the user can act. In the latter case, such patterns are referred to as **actionable.** For example, patterns like "a large earthquake often follows a cluster of small quakes" may be highly actionable if users can act on the information to save lives. Patterns that are **expected** can be interesting if they confirm a hypothesis that the user wishes to validate or they resemble a user's hunch.

- The second question—“*Can a data mining system generate* all *of the interesting patterns?*”— refers to the **completeness** of a data mining algorithm.
- It is often unrealistic and inefficient for data mining systems to generate all possible patterns.
- Instead, user provided constraints and interestingness measures should be used to focus the search.
- For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm.
- Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining.

- Finally, the third question—*“Can a data mining system generate only interesting patterns?”*— is an optimization problem in data mining.
- It is highly desirable for data mining systems to generate only interesting patterns.
- This would be efficient for users and data mining systems because neither would have to search through the patterns generated to identify the truly interesting ones.
- Progress has been made in this direction; however, such optimization remains a challenging issue in data mining.