# Contents

# Chapter 1

# Introduction

This book is an introduction to the young and fast growing field of *data mining* (also known as *knowledge discovery from data*, or *KDD* for short). The book focuses on fundamental data mining concepts and techniques for discovering interesting patterns from data in various applications. In particular, we emphasize prominent techniques for developing effective, efficient and scalable data mining tools.

This chapter is organized as follows. In Section 1.1, you will learn why data mining is in high demand and how it is part of the natural evolution of database technology. Section 1.2 defines data mining with respect to the knowledge discovery process. Next, you will learn about data mining from many aspects, such as the kinds of data that can be mined (Section 1.3), the kinds of knowledge to be mined (Section 1.4), the kinds of technologies to be used (Section 1.5), and targeted applications (Section 1.6). In this way, you will gain a multidimensional view of data mining. Finally, Section 1.7 outlines major issues in data mining research and development.

## 1.1 Why Data Mining?

*Necessity, who is the mother of invention.* — Plato

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. Section 1.1.1 looks at how data mining can meet this need by providing tools to discover knowledge from data. In Section 1.1.2, we observe how data mining can be viewed as a result of the natural evolution of information technology.

### 1.1.1 Moving Towards the Information Age

"*We are living in the information age*" is a popular saying, however *we are actually living in the data age*. Tera- or peta-bytes[1] of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, science and engineering, medicine, and almost every other aspect of our daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools. Businesses worldwide generate gigantic data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback. For example, large stores like Wal-Mart handle hundreds of millions of transactions per week at thousands of branches around the world. Scientific and engineering practices generate high-orders of peta-bytes of data in a continuous manner, from remote sensing, process measuring, scientific experiments, system performance, engineering observations, and environment surveillance. Global backbone telecommunication networks carry tens of petabyes of data traffic every day. The medical and health industry generates tremendous amounts of data from medical records, patient monitoring, and medical images. Billions of Web searches supported by Web search engines process tens of petabytes of data every day. Communities and social media have become increasingly important data sources, producing digital pictures and videos, Web blogs, Web communities, and various kinds of social networks. The list of sources generating huge amounts of data is endless.

This explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from such tremendous amounts of data, and to transform such data into organized knowledge. This necessity led to the birth of data mining. The field is young and promising. Data mining has and will contribute to make great strides in our journey from the data age towards the coming information age.

Example 1.1 **Data mining turns a large collection of data into knowledge.** A Web search engine like Google receives hundreds of millions of queries every day. Each query can be viewed as a transaction where the user describes her/his information need. What novel and useful knowledge can a Web search engine learn from such a huge collection of search queries collected from users over time? Interestingly, some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone. For example, Google *Flu Trends* uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms. A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, Google Flu Trends can

[1]A petabyte is a unit of information or computer storage equal to one quadrillion bytes, or 1000 terabytes, or 1,000,000 gigabytes.

estimate flu activity up to two weeks faster than traditional systems[2]. This example shows how data mining can turn a large collection of data into knowledge that can help meet a global challenge of our times.

### 1.1.2 Data Mining as the Evolution of Information Technology

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 (Figure 1.1): *data collection and database creation, data management* (including data storage and retrieval, and database transaction processing), and *advanced data analysis* (involving data warehousing and data mining). The early development of data collection and database creation mechanisms served as a prerequisite for the later development of effective mechanisms for data storage and retrieval, as well as query and transaction processing. Nowadays numerous database systems offer query and transaction processing as common practice. Advanced data analysis has naturally become the next step.

Since the 1960s, database and information technology has evolved systematically from primitive file processing systems to sophisticated and powerful database systems. The research and development in database systems since the 1970s progressed from early hierarchical and network database systems to relational database systems (where data are stored in relational table structures; see Section 1.3.1), data modeling tools, and indexing and accessing methods. In addition, users gained convenient and flexible data access through query languages, user interfaces, query optimization, and transaction management. Efficient methods for online transaction processing (OLTP), where a query is viewed as a read-only transaction, contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data.

After the establishment of database management systems, database technology moved towards the development of *advanced database systems*, *data warehousing and data mining* for advanced data analysis, and *Web-based databases*. Advanced database systems, for example, resulted from an upsurge of research from the mid-1980s onwards. These systems incorporate new and powerful data models such as extended-relational, object-oriented, object-relational, and deductive models. Application-oriented database systems have flourished, including spatial, temporal, multimedia, active, stream and sensor, scientific and engineering databases, knowledge bases, and office information bases. Issues related to the distribution, diversification, and sharing of data have been extensively studied.

Advanced data analysis sprang up from the late 1980s onwards. The steady and dazzling progress of computer hardware technology in the past three decades

[2]This is reported in [GMP$^+$09]

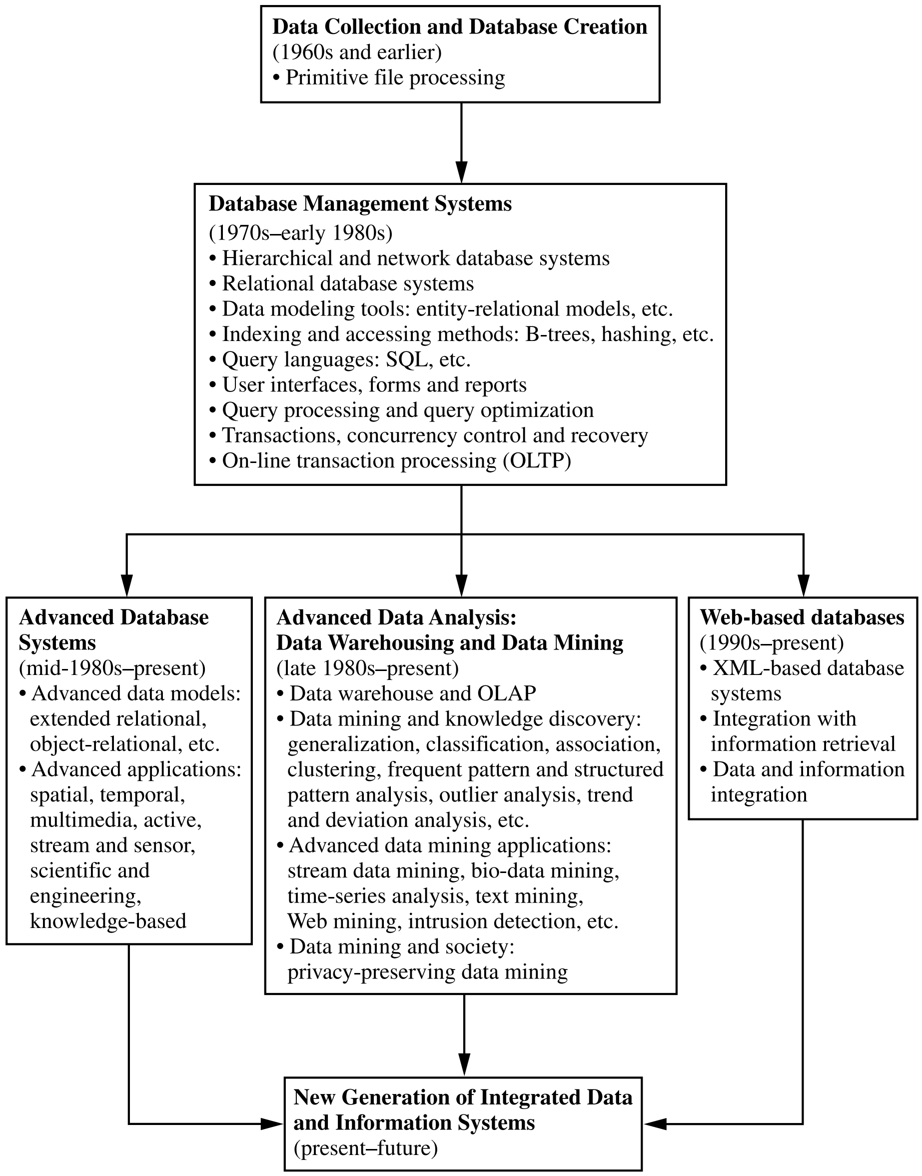


Figure 1.1: The evolution of database system technology.

led to large supplies of powerful and affordable computers, data collection equipment, and storage media. This technology provides a great boost to the database and information industry, and enables a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis. Data can now be stored in many different kinds of databases and information repositories. One emerging data repository architecture is the **data warehouse** (Section 1.3.2). A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making. Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data mining tools that provide data classification, clustering, outlier/exception detection, and the characterization of changes in data over time.

Huge volumes of data have been accumulated beyond databases and data warehouses. During the 1990s, the World Wide Web and Web-based databases (such as XML databases) began to appear. Heterogeneous database systems and Internet-based global information bases such as the World Wide Web have emerged and play a vital role in the information industry. The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies is a challenging task.


Figure 1.2: We are data rich, but information poor.

In summary, the abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation (Figure 1.2). The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large data repositories become "data tombs"—data archives that are seldom visited. Consequently, important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. Although some efforts have been made on developing expert system and knowledge-base technologies, which typically rely on users or domain experts to *manually* input knowledge into knowledge bases, unfortunately, the manual knowledge input procedure is prone to biases and errors, and is extremely costly and time-consuming. The widening gap between data and information calls for a systematic development of *data mining tools* that can turn data tombs into "golden nuggets" of knowledge.

## 1.2 What Is Data Mining?

It is no surprise that data mining, as a truly interdisciplinary subject, can be defined in many different ways. Even the term "data mining" itself does not really present all major components in the picture. To refer to the mining of gold from rocks or sand, we say *gold mining* instead of rock or sand mining. In analog, data mining should have been more appropriately named "knowledge mining from data", which is unfortunately somewhat long. However, the shorter term,"knowledge mining", may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material (Figure 1.3). Thus, such a misnomer carrying both "data" and "mining" became a popular choice. In addition, many other terms have a similar meaning as data mining, like *knowledge mining from data*, *knowledge extraction*, *data/pattern analysis*, *data archaeology*, and *data dredging*.

Many people treat data mining as a synonym for another popularly used term, **Knowledge Discovery from Data**, or **KDD**, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is depicted in Figure 1.4 as an iterative sequence of the following steps:

1. **Data cleaning** (to remove noise and inconsistent data)

2. **Data integration** (where multiple data sources may be combined)[3]

3. **Data selection** (where data relevant to the analysis task are retrieved from the database)

[3]A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.

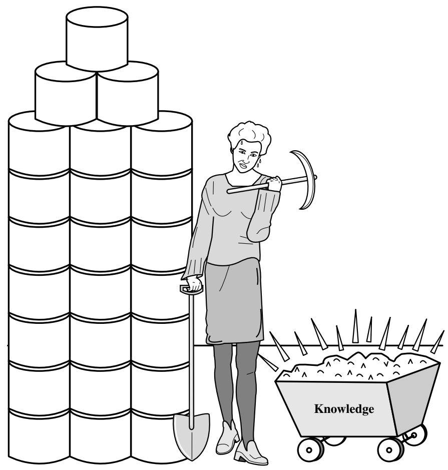


Figure 1.3: Data mining—searching for knowledge (interesting patterns) in your data.

4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)[4]

5. **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns)

6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on *interestingness measures*; Section 1.4.6)

7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Steps 1 to 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

The above view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term 'data mining' is often used to refer to the entire knowledge discovery process (perhaps because the term 'data mining' is shorter than 'knowledge discovery from data'). Therefore, we adopt a broad view of data mining functionality: **Data mining** is the *process* of discovering interesting patterns and knowledge

[4]Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. *Data reduction* may also be performed to obtain a smaller representation of the original data without sacrificing its integrity.

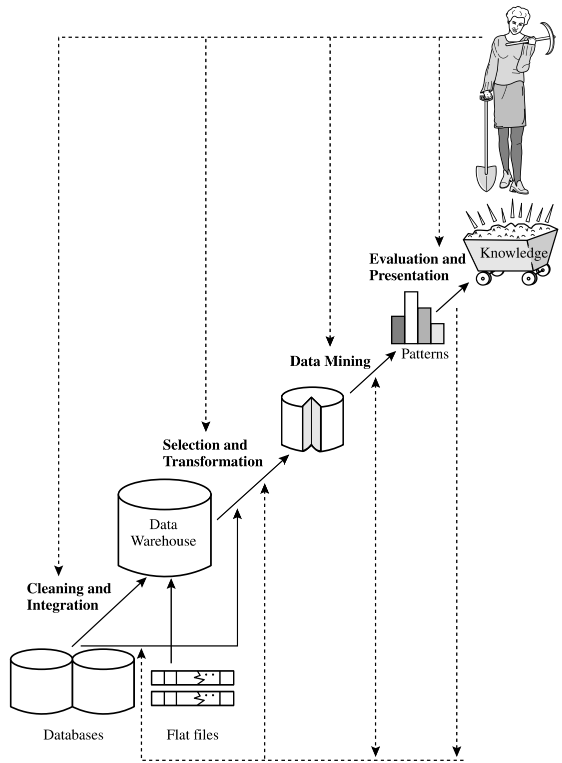


Figure 1.4: Data mining as a step in the process of knowledge discovery.

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.

## 1.3 What Kinds of Data Can Be Mined?

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data (Section 1.3.1), data warehouse data (Section 1.3.2), and transactional data (Section 1.3.3). The concepts and techniques presented in this first volume of the book focus on such data. Data mining can also be applied to other forms of data, such as data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the World Wide Web. We present an overview of such data in Section 1.3.4. Techniques for mining on these kinds of data are covered in volume two of this book. Data mining will certainly continue to embrace new data types as they emerge.

| | |
|---|---|
| *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*.

### 1.3.1 Database Data

A database system, also called a **database management system** (**DBMS** for short), consists of a collection of interrelated data, known as a **database**, and a set of software programs to manage and access the data. The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.

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. A semantic data model, such as an **entity-relationship (ER)** data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

Example 1.2 **A relational database for *AllElectronics*.** The fictitious *AllElectronics* store is used to illustrate concepts throughout the text. The company is described by the following relation tables: *customer, item, employee*, and *branch.* Headers of the tables described here are shown in Figure 1.5. (A header is also called the *schema* of a relation).

- The relation *customer* consists of a set of attributes describing the customer information, including a unique customer identity number (*cust_ID*), customer name, address, age, occupation, annual income, credit information, and category.
- Similarly, each of the relations *item*, *employee*, and *branch* consists of a set of attributes describing the properties of these entities.
- Tables can also be used to represent the relationships between or among multiple entities. In our example, these include *purchases* (customer purchases items, creating a sales transaction that is handled by an employee),

*items_sold* (lists the items sold in a given transaction), and *works_at* (employee works at a branch of *AllElectronics*).

Relational data can be accessed by **database queries** written in a relational query language, such as SQL, or with the assistance of graphical user interfaces. A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing. A query allows retrieval of specified subsets of the data. Suppose that your job is to analyze the *AllElectronics* data. Through the use of relational queries, you can ask things like "Show me a list of all items that were sold in the last quarter." Relational languages also use aggregate functions such as `sum`, `avg` (average), `count`, `max` (maximum), and `min` (minimum). Using aggregates allows you to ask things like "Show me the total sales of the last month, grouped by branch," or "How many sales transactions occurred in the month of December?" or "Which sales person had the highest amount of sales?"

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. Data mining systems may also detect deviations, such as items whose sales are far from those expected in comparison with the previous year. Such deviations can then be further investigated. For example, data mining may discover that there has been a change in packaging of such items, or a significant increase in price.

Relational databases are one of the most commonly available and rich information repositories, and thus they are a major data form in our study of data mining.

### 1.3.2 Data Warehouses

Suppose that *AllElectronics* is a successful international company, with branches around the world. Each branch has its own set of databases. The president of *AllElectronics* has asked you to provide an analysis of the company's sales per item type per branch for the third quarter. This is a difficult task, particularly since the relevant data are spread out over several databases, physically located at numerous sites.

If *AllElectronics* had a data warehouse, this task would be easy. 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. This process will be discussed in Chapters 3 and 4. Figure 1.6 shows the typical framework for construction and use of a data warehouse for *AllElectronics.*

To facilitate decision making, the data in a data warehouse are organized around *major subjects*, such as customer, item, supplier, and activity. The data are stored to provide information from a *historical perspective*, such as in the

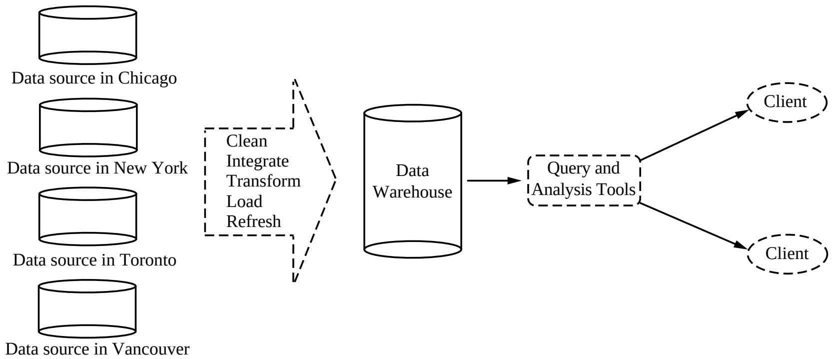


Figure 1.6: Typical framework of a data warehouse for *AllElectronics*.

past 5–10 years, and is typically *summarized*. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region.

A data warehouse is 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.

Example 1.3 **A data cube for *AllElectronics*.** A data cube for summarized sales data of *AllElectronics* is presented in Figure 1.7(a). The cube has three dimensions: *address* (with city values *Chicago, New York, Toronto, Vancouver*), *time* (with quarter values *Q1, Q2, Q3, Q4*), and *item* (with item type values *home entertainment, computer, phone, security*). The aggregate value stored in each cell of the cube is *sales_amount* (in thousands). For example, the total sales for the first quarter, *Q1*, for the items related to security systems in Vancouver is $400,000, as stored in cell ⟨*Vancouver, Q1, security*⟩. Additional cubes may be used to store aggregate sums over each dimension, corresponding to the aggregate values obtained using different SQL group-bys (e.g., the total sales amount per city and quarter, or per city and item, or per quarter and item, or per each individual dimension).

By providing multidimensional data views and the precomputation of summarized data, data warehouse systems can provide inherent support for online analytical processing (OLAP). OLAP operations make use of background knowledge regarding the domain of the data being studied in order to allow the presentation of data at *different levels of abstraction*. Such operations accommodate different user viewpoints. Examples of OLAP operations include

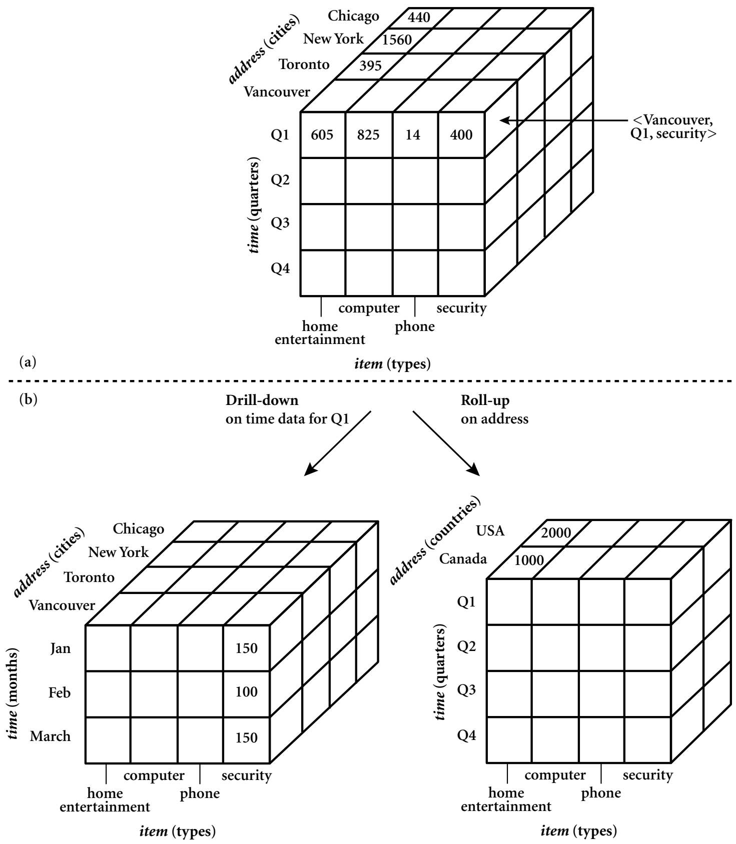


Figure 1.7: A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for *AllElectronics* and (b) showing summarized data resulting from drill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.

**drill-down** and **roll-up**, which allow the user to view the data at differing degrees of summarization, as illustrated in Figure 1.7(b). For instance, we can drill down on sales data summarized by *quarter* to see the data summarized by *month*. Similarly, we can roll up on sales data summarized by *city* to view the data summarized by *country*.

Although data warehouse tools help support data analysis, additional tools for data mining are often needed for in-depth analysis. **Multidimensional data mining** (also called **exploratory multidimensional data mining**) performs data mining in multidimensional space in an OLAP-style. That is, it allows the exploration of multiple combinations of dimensions at varying levels of granularity in data mining, and thus has greater potential for discovering interesting patterns representing knowledge. An overview of data warehouse and OLAP technology is provided in Chapter 4. Advanced issues regarding data

| *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*.

cube computation and multidimensional data mining are discussed in Chapter 5.

### 1.3.3 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 sales person or the branch, and so on.

Example 1.4 **A transactional database for *AllElectronics*.** Transactions can be stored in a table, with one record per transaction. A fragment of a transactional database for *AllElectronics* is shown in Figure 1.8. From the relational database point of view, the *sales* table in Figure 1.8 is a nested relation because the attribute "*list of item_IDs*" contains a set of *items*. Because most relational database systems do not support nested relational structures, the transactional database is usually either stored in a flat file in a format similar to that of the table in Figure 1.8 or unfolded into a standard relation in a format similar to that of the *items_sold* table in Figure 1.5.

As an analyst of *AllElectronics*, you may ask, "Which items sold well together?" This kind of *market basket data analysis* would enable you to bundle groups of items together as a strategy for boosting sales. For example, given the knowledge that printers are commonly purchased together with computers, you could offer certain printers at a steep discount (or even for free) to customers buying selected computers, in the hopes of selling more computers (which are often more expensive than printers). A traditional database system is not able to perform market basket data analysis. Fortunately, data mining on transactional data can do so by mining *frequent itemsets*, that is, sets of items that are frequently sold together. The mining of such frequent patterns from transactional data is discussed in Chapters 6 and 7.

### 1.3.4 Other Kinds of Data

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

Various kinds of knowledge can be mined from such kinds of data. Here, we list just a few. Regarding temporal data, for instance, we can mine banking data for trends of changes, which may aid in the scheduling of bank tellers according to the volume of customer traffic. Stock exchange data can be mined to uncover trends that could help you plan investment strategies (e.g., when is the best time to purchase *AllElectronics* stock?). We could mine computer network data streams to detect intrusions based on the anomaly of message flows, which may be discovered by clustering, dynamic construction of stream models, or comparing the current frequent patterns with those at a previous time. With spatial data, we may look for patterns that describe changes in metropolitan poverty rates based on city distances from major highways. The relationships among a set of spatial objects can be examined in order to discover which subsets of objects are spatially auto-correlated or associated. Mining text data, such as the literature on the subject "data mining" from the past ten years, we can identify the evolution of hot topics in the field. By mining user comments on products (which are often submitted as short text messages), we can assess customer sentiments and understand how well a product is embraced by a market. From multimedia data, we can mine images to identify objects and classify them by assigning semantic labels or tags. By mining video data of a hockey game, we can detect video sequences corresponding to goals. *Web mining* can help us learn about the distribution of information on the Web in general, characterize and classify Web pages, and uncover Web dynamics and the association and other relationships among different Web pages, users, communities, and Web-based activities.

It is important to keep in mind that, in many applications, multiple types of data are present. For example, in Web mining, there often exist text data on Web pages, multimedia data (*e.g.*, pictures and videos) on Web pages, graph data like Web graphs, and map data on some Web sites. In bioinformatics, genomic sequences, biological networks, and 3-D spatial structures of genomes may co-exist for certain biological objects. Mining multiple data sources of complex data often leads to fruitful findings due to the mutual enhancement and consolidation of such multiple sources. On the other hand, it is also challenging due to the difficulties in data cleaning and data integration, as well as the complex interactions among the multiple sources of such data.

While such data require sophisticated facilities for efficient storage, retrieval, and updating, they also provide fertile grounds and raise challenging research and implementation issues for data mining. Data mining on advanced types of data is the topic of volume two of this book.

## 1.4 What Kinds of Patterns Can Be Mined?

We have observed various types of data and information repositories on which data mining can be performed. Let us now examine the kinds of patterns that can be mined.

There are a number of *data mining functionalities*. These include characterization and discrimination (Section 1.4.1); the mining of frequent patterns, associations, and correlations (Section 1.4.2); classification and regression (Section 1.4.3); clustering analysis (Section 1.4.4); and outlier analysis (Section 1.4.5). 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**. 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.

Data mining functionalities, and the kinds of patterns they can discover, are described below. In addition, Section 1.4.6 looks at what makes a pattern interesting. Interesting patterns represent *knowledge*.

### 1.4.1 Concept/Class Description: Characterization and Discrimination

Data entries can be associated with classes or concepts. For example, in the *AllElectronics* store, classes of items for sale include *computers* and *printers*, and concepts of customers include *bigSpenders* and *budgetSpenders*. 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 via (1) *data characterization*, by summarizing the data of the class under study (often called the **target class**) in general terms, or (2) *data discrimination*, by comparison of the target class with one or a set of comparative classes (often called the **contrasting classes**), or (3) both data characterization and discrimination.

**Data characterization** is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class is typically collected by a query. For example, to study the characteristics of software products whose sales were increased by 10% last year, the data related to such products can be collected by executing an SQL query on the sales database.

There are several methods for effective data summarization and characterization. Simple data summaries based on statistical measures and plots are described in Chapter 2. The data cube–based OLAP roll-up operation (Section 1.3.2) can be used to perform user-controlled data summarization along a specified dimension. This process is further detailed in Chapters 4 and 5, which discuss data warehousing. An *attribute-oriented induction* technique can be used to perform data generalization and characterization without step-by-step user interaction. This technique is also described in Chapter 4.

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**).

Example 1.5 **Data characterization.** A customer relationship manager at *AllElectronics* may raise the following data mining task: "*Summarize the characteristics of customers who spend more than $5,000 a year at AllElectronics*". The result is a general profile of these customers, such as they are 40–50 years old, employed, and have excellent credit ratings. The data mining system should allow the customer relationship manager to drill down on any dimension, such as on *occupation* in order to view these customers according to their type of employment.

**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. 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 like to compare the general features of software products whose sales increased by 10% last year against those whose sales decreased by at least 30% during the same period. The methods used for data discrimination are similar to those used for data characterization.

*"How are discrimination descriptions output?"* The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes. Discrimination descriptions expressed in the form of rules are referred to as **discriminant rules**.

Example 1.6 **Data discrimination.** A customer relationship manager at *AllElectronics* may want to compare two groups of customers—those who shop for computer products regularly (more than twice a month) versus those who rarely shop for such products (i.e., less than three times a year). The resulting description provides a general comparative profile of these customers, such as 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, such as *occupation*, or adding new dimensions, such as *income_level*, may help to find even more discriminative features between the two classes.

Concept description, including characterization and discrimination, is described in Chapter 4.

### 1.4.2 Mining Frequent Patterns, Associations, and Correlations

**Frequent patterns**, as the name suggests, 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 frequently appear together in a transactional data set, such as 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 PC, followed by a digital camera, and then a memory card, is a (*frequent*) *sequential pattern*. A substructure can refer to different structural forms, such as graphs, trees, or lattices, which 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.

Example 1.7 **Association analysis.** Suppose, as a marketing manager at *AllElectronics*, you would like 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

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

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 of 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 above rule can be written simply as "*computer* $\Rightarrow$ *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 of age with an income of 40,000 to 49,000 and have purchased a laptop (computer) at *AllElectronics*. There is a probability of 60% 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**.

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. The mining of frequent patterns, associations, and correlations is discussed in Chapters 6 and 7, where particular emphasis is placed on efficient algorithms for frequent itemset mining. Sequential pattern mining and structured pattern mining are considered advanced topics, which will be discussed in the second volume of this book.

### 1.4.3 Classification and Regression for Predictive Analysis

**Classification** is the process of finding a **model** (or function) that describes and distinguishes 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** (i.e., data objects whose class labels are known).

*"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 flow-chart-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 be converted to classification rules. A **neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and $k$-nearest neighbor classification.

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. **Regression analysis** is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution *trends* based on the available data.

Classification and regression may need to be preceded by **relevance analysis**, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. The other attributes, which are irrelevant, can then be excluded from consideration.

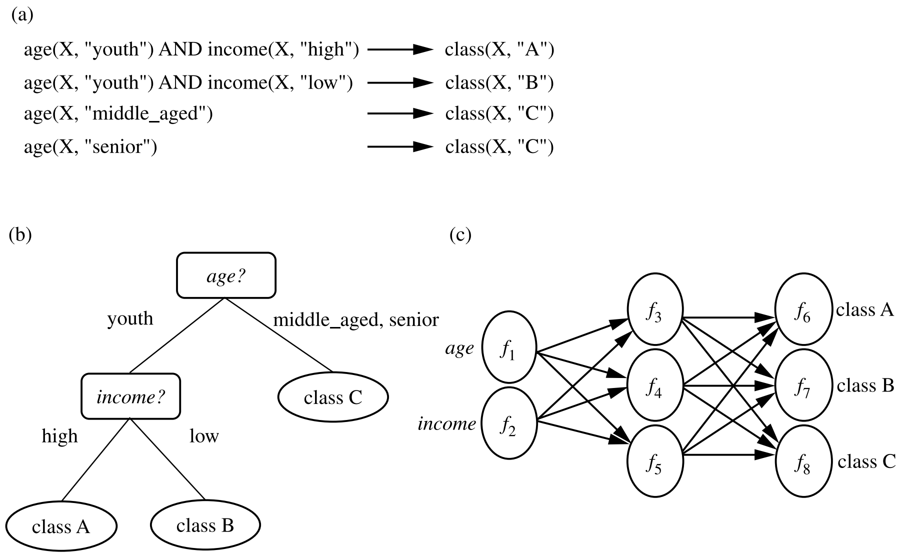


Figure 1.9: A classification model can be represented in various forms, such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network.

Example 1.8 **Classification and regression.** Suppose, as a sales manager of *AllElectronics*, you would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign: *good response*, *mild response*, and *no response*. You would like to derive a model for each of these three classes based on the descriptive features of the items, such as *price*, *brand, place_made, type*, and *category*. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

Suppose that the resulted classification is expressed in the form of a decision tree. The decision tree, for instance, may identify *price* as being the single factor that best distinguishes the three classes. The tree may reveal that, in addition to *price*, other features that help to further distinguish objects of each class from another include *brand* and *place_made*. Such a decision tree may help you understand the impact of the given sales campaign and design a more effective campaign for the future.

Suppose instead, that rather than predicting categorical response labels for each store item, you would like to predict the amount of revenue that each item will generate during an upcoming sale at *AllElectronics*, based on the previous sales data. This is an example of regression analysis because the regression model constructed will predict a continuous function, or ordered value.

Chapters 8 and 9 discuss classification in further detail. Regression analysis will be studied in the second volume of this book.

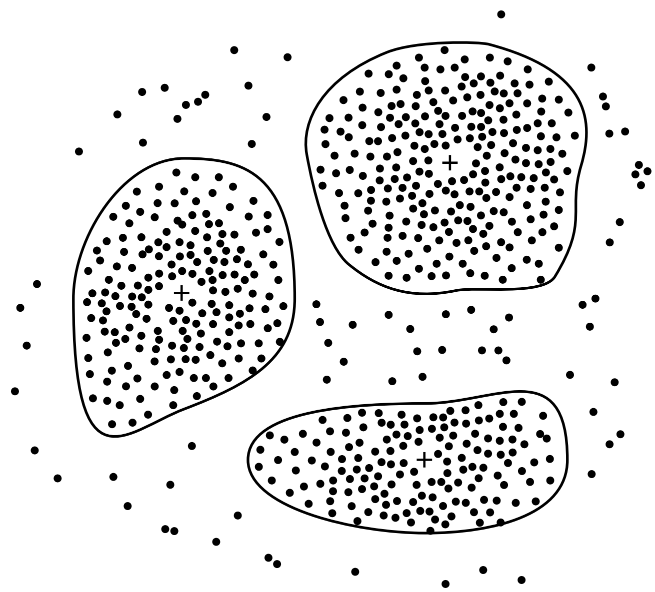

Figure 1.10: A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster "center" is marked with a "+".
[TO EDITOR This figure will be improved.]

### 1.4.4 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 that is 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.

Example 1.9 **Cluster analysis.** Cluster analysis can be performed on *AllElectronics* customer data in order to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing. Figure 1.10 shows a 2-D plot of customers with respect to customer locations in a city. Three clusters of data points are evident.

Cluster analysis forms the topic of Chapters 10 and 11.

### 1.4.5 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 such as 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, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

Example 1.10 **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.

Outlier analysis is discussed in Chapter 12.

### 1.4.6 Are All Patterns Interesting?

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

*"So,"* 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? Can a data mining system generate only those interesting patterns?"*

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**.

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

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

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

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.

Other objective interestingness measures include *accuracy* and *coverage* for classification (IF-THEN) rules. In general terms, accuracy tells us the percentage of data samples that are correctly classified by a rule. Coverage is similar to "support", in that it tells us the percentage of data samples 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 to identify interesting patterns, they are often insufficient unless combined with subjective measures that reflect the needs and interests of a particular user. 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 some 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 it to save lives. Patterns that are **expected** can be interesting if they confirm a hypothesis that the user wishes to validate, or 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. The methods involved are examined in detail in Chapter 7.

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 in order to identify the truly interesting ones. Progress has been made in this direction, however, such optimization remains a challenging issue in data mining.

Measures of pattern interestingness are essential for the efficient discovery of patterns to target users. Such measures can be used after the data mining step in order to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy prespecified interestingness constraints. Examples of such a constraint-based mining process are described in Chapter 7 (with respect to association mining) and Chapter 11 (with respect to clustering).

Methods to assess pattern interestingness, and their use to improve data mining efficiency, are discussed throughout the book, with respect to each kind of pattern that can be mined.

## 1.5 What Technologies Are Used?


Figure 1.11: Data mining adopts techniques from many domains.

Data mining, as a highly application-driven domain, has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains (Figure 1.11). The interdisciplinary nature of data mining research and development contributes significantly to the success of data mining and its extensive applications. In this section, we give examples of several disciplines that strongly influence the development of data mining methods.

### 1.5.1 Statistics

**Statistics** studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics.

A **statistical model** is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes. For example, in data mining tasks like data characterization and classification, statistical models of target classes can be built. In other words, such statistical models can be the outcome of a data mining task. Alternatively, data mining tasks can be built on top of statistical models. For example, we can use statistics to model noise and missing data values. Then, when mining patterns in a large data set, the data mining process can use the model to help identify and handle noisy or missing values in the data.

Statistics research develops tools for prediction and forecasting using data and statistical models. Statistical methods can be used to summarize or describe a collection of data, which is often known as **descriptive statistics**. (Chapter 2 contains an introduction to descriptive statistics.) Statistics is useful at mining various patterns from data, as well as understanding the underlying mechanisms generating and affecting the patterns. **Inferential statistics** or **predictive statistics** models patterns in a way that accounts for randomness and uncertainty in the observations and is used to draw inferences about the process or population under investigation.

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

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

### 1.5.2 Machine Learning

**Machine learning** investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can recognize handwritten postal codes on mail automatically after learning from a set of examples.

Machine learning is a fast growing discipline. Here, we illustrate classical problems in machine learning that are highly related to data mining.

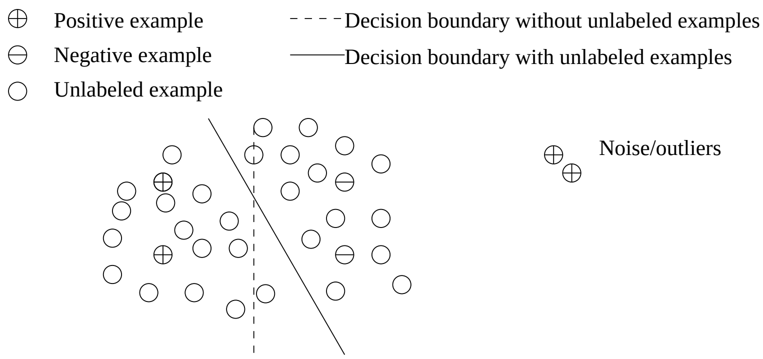


Figure 1.12: Semi-supervised learning.

- **Supervised learning** is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.

- **Unsupervised learning** is essentially a synonym for clustering. The learning process is unsupervised in that the input examples are not class-labeled. Typically, we may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9, respectively. However, since the training data are not labeled, the learned model cannot tell us the semantic meaning of the clusters found.

- **Semi-supervised learning** is a class of machine learning techniques that makes use of both labeled and unlabeled examples when learning a model. In one approach, labeled examples are used to learn class models, and unlabeled examples are used to refine the boundaries between classes. For a two-class problem, we can think of the set of examples belonging to one class as the *positive examples*, and those belonging to the other class as the *negative examples*. In Figure 1.12, if we do not consider the unlabeled examples, the dashed line is the decision boundary that best partitions the positive examples from the negative examples. Using the unlabeled examples, we can refine the decision boundary to the solid line. Moreover, we can detect that the two positive examples at the right top corner, though labeled, are likely noise or outliers.

- **Active learning** is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user such as a domain expert to label an example, which may be from a set of unlabeled examples or synthesized by the learning program. The goal is to optimize the model quality by actively acquiring knowledge from human users, given a constraint on how many examples they can be asked to label.

You can see that data mining and machine learning share a lot of similarity. For classification and clustering tasks, machine learning research often focuses on the accuracy of the model. In addition to accuracy, data mining research places strong emphasis on the efficiency and scalability of mining methods on large data sets, as well as ways to handle complex types of data, and explore new, alternative methods.

### 1.5.3 Database Systems and Data Warehouses

**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. Moreover, data mining tasks can be used to extend the capability of existing database systems to satisfy sophisticated data analysis requirements from advanced users.

Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities. A **data warehouse** integrates data originating from multiple sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes. The data cube model not only facilitates online analytical processing (OLAP) in multidimensional databases, but also promotes *multidimensional data mining* (Section 1.3.2).

### 1.5.4 Information Retrieval

**Information retrieval** (**IR**) is the science of searching for documents or information in documents. Documents can be text or multimedia, and may reside on the World Wide Web. The differences between traditional information retrieval and database systems are twofold: information retrieval assumes that (1) the data under search are unstructured, and (2) the queries are formed mainly by keywords, which do not have complex structures (unlike SQL queries in database systems).

The typical approaches in information retrieval adopt probabilistic models for retrieval. For example, a text document can be regarded as a bag of words, that is, a multiset of words appearing in the document. The **language model** of the document is the probability density function that generates the bag of words in the document. The similarity between two documents can be measured by the similarity between their corresponding language models.

Furthermore, a topic in a set of text documents can be modeled as a probability distribution over the vocabulary, which is called a **topic model**. A text document, which may involve one or multiple topics, can be regarded as a mixture of multiple topic models. By integrating information retrieval models and data mining techniques, we can find the major topics in a collection of documents and, for each document in the collection, the major topics involved.

Increasingly large amounts of text and multimedia data have been accumulated and made available online due to the fast growth of the Web and applications such as digital libraries, digital governments, and health care information systems. Their effective search and analysis has raised many challenging issues in data mining. Therefore, text mining and multimedia data mining, integrated with information retrieval methods, have become increasingly important.

## 1.6 What 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, such as bioinformatics and software engineering, require more in-depth treatment, and are thus left to the second volume of this book. To demonstrate the importance of applications as a major dimension in data mining research and development, we briefly discuss two highly successful and popular application examples of data mining: *business intelligence* and *Web search engines*.

### 1.6.1 Business Intelligence

It is critical for businesses to acquire a better understanding of the commercial context of their organization, such as their customers, the market, supply and resources, and competitors. **Business intelligence** (**BI**) technologies provide historical, current, and predictive views of business operations. Examples include reporting, online analytical processing (OLAP), business performance management, competitive intelligence, benchmarking, and predictive analytics.

*"How important is business intelligence?"* It is predicted that by 2012, more than 35% of the top five thousand global companies may regularly fail to make insightful decisions about significant changes in their business and markets due to the lack of business intelligence.

Clearly, data mining is the core of business intelligence. Online analytical processing (OLAP) 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, for which there are many applications in analyzing markets, supplies and sales. Moreover, clustering plays a central role in customer relationship management, which groups customers based on their similarities. Using characterization mining techniques, we can better understand features of each customer group, and develop customized customer reward programs.

### 1.6.2 Web Search Engines

A **Web search engine** is a specialized computer server that searches for information on the World Wide Web. The search results of a user query are often returned as a list (sometimes called *hits*). The hits may consist of Web pages, images, and other types of files. Some Web search engines also search and return data available in public databases or open directories. Web search engines differ from **Web directories** in that Web directories are maintained by human editors, whereas Web search engines operate algorithmically or are a mixture of algorithmic and human input.

Web search engines are essentially very large data mining applications. Various data mining techniques are used in all aspects of Web search engines, ranging from *crawling*[5] (e.g., deciding which pages should be crawled and the crawling frequencies), indexing (e.g., selecting pages to be indexed and deciding to which extent the index should be constructed), and searching (e.g., deciding how pages should be ranked, which advertisements should be added, and how the search results can be personalized or made "context aware").

Web search engines pose grand challenges to data mining. First, Web search engines have to handle a huge and ever growing amount of data. Typically, such data cannot be processed using one or a few machines. Instead, Web search engines often need to use *computer clouds*, which consist of thousands or even hundreds of thousands of computers to collaboratively mine the huge amount of data. Scaling up data mining methods over computer clouds and large distributed data sets is an area for further research.

Second, Web search engines often have to deal with online data. A Web search engine may be able to afford constructing a model offline on huge data sets. To do this, it may construct a query classifier that assigns a Web search query to predefined categories based on the topic of the query (such as whether the Web search query "apple" is meant to retrieve information about a fruit or a brand of computers). Whether or not a model is constructed offline, the application of the model online must be fast enough to answer user queries in real time. Another challenge is maintaining and incrementally updating a model on fast growing data streams. For example, a query classifier may need to be incrementally maintained continuously since new queries keep emerging, and predefined categories and the data distribution may change. Most of the existing model training methods are offline and static, and thus cannot be used in such a scenario.

Third, Web search engines often have to deal with queries that are asked only a very small number of times. Suppose a Web search engine wants to provide *context-aware* query recommendations. That is, when a user poses a query, the Web search engine tries to infer the context of the query using the user's profile and his/her query history in order to return more customized answers. However, although the

[5] A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner.

total number of queries asked can be huge, most of the queries may be asked only a few times. Such severely skewed data are challenging for many data mining and machine learning methods.

## 1.7 Major Issues in Data Mining

*Life is short but art is long.*

Data mining is a dynamic and fast expanding field with great strengths. Here we briefly outline the major issues in data mining research, partitioning them into five groups: *mining methodology, user interaction, efficiency and scalability, diversity of data types*, and *data mining and society*. Many of these issues have been addressed in recent data mining research and development, *to a certain extent*, and are now considered *data mining requirements*; others are still at the research stage. The issues continue to stimulate further investigation and improvement in data mining.

### 1.7.1 Mining Methodology

Researchers have been vigorously developing new data mining methodologies. This involves the investigation of new kinds of knowledge, mining in multidimensional space, integrating methods from other disciplines, and the consideration of semantic ties among data objects. In addition, mining methodologies should consider issues such as data uncertainty, noise, and incompleteness. Some mining methods explore how user-specified measures can be used to assess the interestingness of discovered patterns as well as guide the discovery process. Let's have a look at these various aspects of mining methodology.

- *Mining various and new kinds of knowledge:* Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques. Due to the diversity of applications, new mining tasks continue to emerge, making data mining a dynamic and fast growing field. For example, for effective Web search, the ranking of authoritative Webpages has become a popular research frontier in recent years.

- *Mining knowledge in multidimensional space:* When searching for knowledge in large data sets, we can explore the data in multidimensional space. That is, we can search for interesting patterns among combinations of dimensions (attributes) at varying levels of abstraction. Such mining is known as *(exploratory) multidimensional data mining*. In many cases, data can be aggregated or viewed as a multidimensional data cube. Mining knowledge in cube space can substantially enhance the power and flexibility of data mining.

- *Data mining: An inter-disciplinary 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.

- *Boosting the power of discovery in a networked environment:* Most data objects reside in a linked or inter-connected environment, whether it be the Web, database relations, files, or documents. Semantic links across multiple data objects can be used to advantage in data mining. Knowledge derived in one set of objects can be used to boost the discovery of the knowledge in a "related" or semantically-linked set of objects.

- *Handling uncertainty, noise, or incompleteness of data:* Data often contain noise, errors, exceptions, 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, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

- *Pattern evaluation and pattern- or constraint-guided mining:* Not all the patterns generated by data mining processes are interesting. What makes a pattern interesting may vary from user to user. Therefore, techniques are needed to assess the interestingness of discovered patterns based on subjective measures. These estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. Moreover, by using interestingness measures or user-specified constraints to *guide* the discovery process, we may generate more interesting patterns and reduce the search space.

### 1.7.2 User Interaction

The user plays an important role in the data mining process. Interesting areas of research include *how to interact with a data mining system*, *how to incorporate a user's background knowledge in mining*, and *how to visualize and comprehend data mining results*. We introduce each of these here.

- *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. A user may like to first sample a set of data, explore general characteristics of the data, and estimate potential mining results. Interactive mining should allow users to dynamically change the focus of search, to refine mining requests based on returned results, and to drill, dice, and pivot through the data and knowledge space interactively, dynamically exploring "cube space" while mining.

- *Incorporation of background knowledge:* Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated into the knowledge discovery process. Such knowledge can be used for pattern evaluation as well as to guide the search towards interesting patterns.

- *Ad hoc data mining and data mining query languages:* Query languages (such as SQL) have played an important role in flexible search in that they allow users to pose *ad hoc* queries. Similarly, high-level, data mining query languages or other high-level flexible user interfaces will give users the freedom to define ad hoc data mining tasks. This should facilitate specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns. Optimization of the processing of such flexible mining requests is another promising area of study.

- *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 expressive knowledge representations, user-friendly interfaces, and visualization techniques.

### 1.7.3 Efficiency and Scalability

Efficiency and scalability are always considered when comparing data mining algorithms. As data amounts continue to multiply, these two factors are especially critical.

- *Efficiency and scalability of data mining algorithms:* Data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams. In other words, the running time of a data mining algorithm must be predictable, short, and acceptable by applications. *Efficiency, scalability, performance, optimization*, and the ability to execute in *real-time* are key criteria that drive the development of many new data mining algorithms.

- *Parallel, distributed, and incremental 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-intensive 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. *Cloud computing* and *cluster computing*, which use computers in a distributed and collaborative way to tackle very large scale computational tasks, are also active research themes in parallel data mining. In addition, the high cost of some data mining processes and the incremental nature of input promote **incremental** data mining, which incorporates new data updates without having to mine the entire data again "from scratch." Such methods perform knowledge modification incrementally to amend and strengthen what was previously discovered.

### 1.7.4 Diversity of Database Types

The wide diversity of database types brings about challenges to data mining. These include:

- *Handling complex types of data:* Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi-structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, biological sequences, sensor data, spatial data, hypertext data, multimedia data, software program code, Web data, and social network data. It is 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.

- *Mining dynamic, networked, and global data repositories:* Multiple sources of data are connected by the Internet and various kinds of networks, forming gigantic, distributed, and heterogeneous global information systems and networks. The discovery of knowledge from different sources of structured, semi-structured, or unstructured yet interconnected data with diverse data semantics poses great challenges to data mining. Mining such gigantic, interconnected information networks may help disclose many more patterns and knowledge in heterogeneous data sets than can be discovered from a small set of isolated data repositories. Web mining, multi-source data mining, and information network mining have become challenging and fast-evolving fields in data mining.

### 1.7.5 Data Mining and Society

How does data mining impact society? What steps can data mining take to preserve the privacy of individuals? Do we use data mining in our daily life, without even knowing that we do? These questions raise the following issues:

- *Social impacts of data mining:* With data mining penetrating our everyday life, 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 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.

- *Privacy-preserving data mining:* Data mining will help scientific discovery, business management, economy recovery, and security protection (such as the real-time discovery of intruders and cyber-attacks). However, it poses the risk of disclosing an individual's personal information. Studies on privacy-preserving data publishing and data mining are ongoing. The philosophy is to observe data sensitivity and preserve people's privacy while performing successful data mining.

- *Invisible data mining:* We cannot expect everyone in society to learn and master data mining techniques. More and more systems should have data mining functions built within so that people can perform data mining or use data mining results by simple mouse clicking, without any knowledge of data mining algorithms. Intelligent Web 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 their buying patterns, which it may use to recommend other items for purchase in the future.

These issues and many additional ones relating to the research, development, and application of data mining are discussed throughout the book.

## 1.8 Summary

- *Necessity is the mother of invention.* With the mounting growth of data in every application, data mining meets the imminent need for effective, scalable and flexible data analysis in our society. Data mining can be considered as a natural evolution of information technology and a confluence of several related disciplines and application domains.

- **Data mining** is the process of discovering interesting patterns from massive amounts of data. As a *knowledge discovery process*, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation.

- A pattern is *interesting* if it is valid on test data with some degree of certainty; novel; potentially useful (e.g., can be acted on or validates a hunch about which the user was curious); and easily understood by humans. Interesting patterns represent **knowledge**. Measures of **pattern interestingness**, either *objective* or *subjective*, can be used to guide the discovery process.

- We present a **multidimensional view** of data mining. The major dimensions are **data**, **knowledge**, **technologies**, and **applications**.

- Data mining can be conducted on any kind of **data** as long as the data are meaningful for a target application, such as database data, data warehouse data, transactional data, and advanced data types. Advance data types include time-related or sequence data, data streams, spatial and spatiotemporal data, text and multimedia data, graph and networked data, and Web data.

- A **data warehouse** is a repository for long-term storage of data from multiple sources, organized so as to facilitate management decision making. The data are stored under a unified schema and are typically summarized. Data warehouse systems provide multidimensional data analysis capabilities, collectively referred to as **OLAP** (**online analytical processing**).

- **Multidimensional data mining** (also called **exploratory multidimensional data mining**) integrates core data mining techniques with OLAP-based multidimensional analysis. It searches for interesting patterns among multiple combinations of dimensions (attributes) at varying levels of abstraction, thereby exploring multidimensional data space.

- *Data mining functionalities* are used to specify the kinds of patterns or **knowledge** to be found in data mining tasks. The functionalities include characterization and discrimination; the mining of frequent patterns, associations, and correlations; classification and regression; cluster analysis; and outlier detection. As new types of data, new applications, and new analysis demands continue to emerge, there is no doubt we will see more and more novel data mining tasks in the future.

- Data mining, as a highly application-driven domain, has incorporated **technologies** from many other domains. These include statistics, machine learning, database and data warehouse systems, and information retrieval. The **interdisciplinary nature of data mining research and development** contributes significantly to the success of data mining and its extensive applications.

- Data mining has many successful **applications**, such as business intelligence, Web search, bioinformatics, health informatics, finance, digital libraries, and digital governments.

- There are many challenging **issues in data mining research**. Areas include mining methodology, user interaction, efficiency and scalability, and dealing with diverse data types. Data mining research has strongly impacted society and will continue to do so in the future.

## 1.9 Exercises

1. What is *data mining*? In your answer, address the following:

   (a) Is it another hype?

   (b) Is it a simple transformation of technology developed from *databases*, *statistics*, *machine learning*, and *pattern recognition*?

(c) We have presented a view that data mining is the result of the evolution of *database technology*. Do you think that data mining is also the result of the evolution of *machine learning research*? Can you present such views based on the historical progress of this discipline? Do the same for the fields of *statistics* and *pattern recognition*.

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

2. How is a *data warehouse* different from a database? How are they similar?

3. Define each of the following *data mining functionalities*: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.

4. Present an example where data mining is crucial to the success of a business. What *data mining functionalities* does this business need (e.g. think of the kinds of patterns that could be mined)? Can such patterns be performed alternatively by data query processing or simple statistical analysis?

5. What is the difference between discrimination and classification? Between characterization and clustering? Between classification and regression? For each of these pairs of tasks, how are they similar?

6. Based on your observation, describe another possible kind of knowledge that needs to be discovered by data mining methods but has not been listed in this chapter. Does it require a mining methodology that is quite different from those outlined in this chapter?

7. *Outliers* are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

8. Describe three challenges to data mining regarding *data mining methodology* and *user interaction issues.*

9. What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)?

10. Outline the major research challenges of data mining in one specific application domain, such as stream/sensor data analysis, spatiotemporal data analysis, or bioinformatics.

## 1.10 Bibliographic Notes

The book *Knowledge Discovery in Databases*, edited by Piatetsky-Shapiro and Frawley [PSF91], is an early collection of research papers on knowledge discovery from data. The book *Advances in Knowledge Discovery and Data Mining*, edited by Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy [FPSSe96], is a collection of later research results on knowledge discovery and data mining. There have been many data mining books published in recent years, including *Predictive Data Mining* by Weiss and Indurkhya [WI98], *Mastering Data Mining: The Art and Science of Customer Relationship Management* by Berry and Linoff [BL99], *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* by Witten and Frank [WF05], *Principles of Data Mining (Adaptive Computation and Machine Learning)* by Hand, Mannila, and Smyth [HMS01], *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman [HTF09], *Data Mining: Introductory and Advanced Topics* by Dunham, *Data Mining: Multimedia, Soft Computing, and Bioinformatics* by Mitra and Acharya [MA03], *Introduction to Data Mining* by Tan, Steinbach and Kumar [TSK05], *Mining the Web: Discovering Knowledge from Hypertext Data* by Chakrabarti [Cha03], and *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* by Liu [Liu06]. There are also books containing collections of papers or chapters on particular aspects of knowledge discovery, such as *Machine Learning and Data Mining: Methods and Applications* edited by Michalski, Brakto, and Kubat [MBK98], *Relational Data Mining* edited by Dzeroski and Lavrac [De01], *Mining Graph Data* edited by Cook and Holder [CH07], *Data Streams: Models and Algorithms* edited by Aggarwal [Agg06], *Next Generation of Data Mining* edited by Kargupta, Han, Yu, and Motwani [KHYM08], *Multimedia Data Mining: A Systematic Introduction to Concepts and Theory* edited by Z. Zhang and R. Zhang [ZZ09], *Geographic Data Mining and Knowledge Discovery* edited by Miller and Han [MH09], and *Link Mining—Models, Algorithms and Applications* edited by Yu, Faloutsos and Han [YFe10]. There are also many tutorial notes on data mining in major database, data mining, machine learning, statistics, and web technology conferences.

*KDD Nuggets* is a regular, free electronic newsletter containing information relevant to knowledge discovery and data mining, moderated by Piatetsky-Shapiro since 1991. The Internet site *KDNuggets* (www.kdnuggets.com) contains a good collection of KDD-related information.

The data mining community started its first international conference on knowledge discovery and data mining in 1995 [Fe95]. The conference evolved from the four international workshops on knowledge discovery in databases, held from 1989 to 1994 [PS89, PS91, FUe93, Fe94]. ACM-SIGKDD, a Special Interest Group on Knowledge Discovery in Databases was set up under ACM in 1998. In 1999, ACM-SIGKDD organized the fifth international conference on knowledge discovery and data mining (KDD'99). IEEE Computer Science Society has organized its annual data mining conference, International Conference on Data Mining (ICDM), since 2001. SIAM (Society on Industrial and Applied Mathematics) has organized its annual data mining conference, SIAM Data Mining conference (SDM), since 2002. A dedicated journal, *Data Mining and Knowledge Discovery*, published by Kluwers

Publishers, has been available since 1997. An ACM journal, *ACM Transactions on Knowledge discovery from Data*, started its first volume in 2007. ACM-SIGKDD also publishes a biannual newsletter, *SIGKDD Explorations*. There are a few other international or regional conferences on data mining, such as the *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD), the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML PKDD), and the *International Conference on Data Warehousing and Knowledge Discovery* (DaWaK).

Research in data mining has also been published in books, conferences, and journals on databases, statistics, machine learning, and data visualization. References to such sources are listed below.

Popular textbooks on database systems include *Database Systems: The Complete Book* by Garcia-Molina, Ullman, and Widom [GMUW08], *Database Management Systems* by Ramakrishnan and Gehrke [RG03], *Database System Concepts* by Silberschatz, Korth, and Sudarshan [SKS05], and *Fundamentals of Database Systems* by Elmasri and Navathe [EN06]. For an edited collection of seminal articles on database systems, see *Readings in Database Systems* by Stonebraker and Hellerstein [HS05]. Many books on data warehouse technology, systems, and applications have been published in the last several years, such as *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* by Kimball and Ross [KR02], *The Data Warehouse Lifecycle Toolkit:* by Kimball, Ross, Thornthwaite et al. [KRTM08], *Mastering Data Warehouse Design: Relational and Dimensional Techniques* by Imhoff, Galemmo and Geiger [IGG03], and *Building the Data Warehouse* by Inmon [Inm96], A set of research papers on materialized views and data warehouse implementations were collected in *Materialized Views: Techniques, Implementations, and Applications* by Gupta and Mumick [GM99]. Chaudhuri and Dayal [CD97] present a comprehensive overview of data warehouse technology.

Research results relating to data mining and data warehousing have been published in the proceedings of many international database conferences, including the *ACM-SIGMOD International Conference on Management of Data (SIGMOD)*, the *International Conference on Very Large Data Bases (VLDB),* the *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, the *International Conference on Data Engineering (ICDE)*, the *International Conference on Extending Database Technology (EDBT)*, the *International Conference on Database Theory (ICDT)*, the *International Conference on Information and Knowledge Management (CIKM)*, the *International Conference on Database and Expert Systems Applications (DEXA)*, and the *International Symposium on Database Systems for Advanced Applications (DASFAA).* Research in data mining is also published in major database journals, such as *IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Database Systems (TODS), Journal of ACM (JACM), Information Systems, The VLDB Journal, Data and Knowledge Engineering, International Journal of Intelligent Information Systems (JIIS)*, and *Knowledge and Information Systems (KAIS).*

Many effective data mining methods have been developed by statisticians and pattern recognition researchers, and introduced in a rich set of textbooks. An overview of classification from a statistical pattern recognition perspective can be found in in *Pattern Classification* by Duda, Hart, Stork [DHS01]. There are also many textbooks covering different topics in statistical analysis, such as *Mathematical Statistics: Basic Ideas and Selected Topics* by Bickel and Doksum [BD01], *The Statistical Sleuth: A Course in Methods of Data Analysis* by Ramsey and Schafer [RS01], *Applied Linear Statistical Models* by Neter, Kutner, Nachtsheim, and Wasserman [NKNW96], *An Introduction to Generalized Linear Models* by Dobson [Dob90], *Applied Statistical Time Series Analysis* by Shumway [Shu88], and *Applied Multivariate Statistical Analysis* by Johnson and Wichern [JW92].

Research in statistics is published in the proceedings of several major statistical conferences, including *Joint Statistical Meetings*, *International Conference of the Royal Statistical Society*, and *Symposium on the Interface: Computing Science and Statistics*. Other sources of publication include the *Journal of the Royal Statistical Society*, *The Annals of Statistics*, *Journal of American Statistical Association*, *Technometrics*, and *Biometrika*.

Textbooks and reference books on machine learning include *Machine Learning, An Artificial Intelligence Approach*, Vols. 1–4, edited by Michalski et al. [MCM83, MCM86, KM90, MT94], *C4.5: Programs for Machine Learning* by Quinlan [Qui93], *Machine Learning* by Mitchell [Mit97], *Introduction to Machine Learning* by Alpaydin [Alp04], and *Pattern Recognition and Machine Learning* by Bishop [Bis06]. For an edited collection of seminal articles on machine learning, see *Readings in Machine Learning* by Shavlik and Dietterich [SD90].

Machine learning research is published in the proceedings of several large machine learning and artificial intelligence conferences, including the *International Conference on Machine Learning (ML)*, the *ACM Conference on Computational Learning Theory (COLT)*, the *International Joint Conference on Artificial Intelligence (IJCAI)*, and the *American Association of Artificial Intelligence Conference (AAAI)*. Other sources of publication include major machine learning, artificial intelligence, pattern recognition, and knowledge system journals, some of which have been mentioned above. Others include *Machine Learning (ML), Artificial Intelligence Journal (AI)*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, and *Cognitive Science*.

Textbooks and reference books on information retrieval include Croft, Metzler, and Strohman [CMS09], Manning, Raghavan, and Schutz [MRS08], Grossman and Frieder [GF04], Baeza-Yates and Ribeiro-Neto [BYRN99], and Jones and Willett [Je97].

Information retrieval research is published in the proceedings of several information retrieval and Web search and mining conferences, including the *International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, the *International World Wide Web Conference (WWW)*, the *ACM International Conference on Web Search and Data Mining (WSDM)*, the *Text Retrieval Conference (TREC)* the *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, the *ACM Conference on Information and Knowledge Management (CIKM)*, and the *European Conference on Information Retrieval (ECIR)*. Other sources of publication include major information retrieval, information systems and Word-Wide-Web journals, such as *Journal of Information Retrieval, ACM Transactions on Information Systems (TOIS)*, *Information Processing and Management*, *Knowl-*

*edge and Information Systems (KAIS)*, and *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

# Bibliography


[Agg06] C. C. Aggarwal. *Data Streams: Models and Algorithms.* Kluwer Academic, 2006.

[Alp04] E. Alpaydin. *Introduction to Machine Learning.* MIT Press, 2004.

[BD01] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1.* Prentice Hall, 2001.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[BL99] M. J. A. Berry and G. Linoff. *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons, 1999.

[BYRN99] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press/Addison-Wesley, 1999.

[CD97] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *SIGMOD Record*, 26:65–74, 1997.

[CH07] D. J. Cook and L. B. Holder. *Mining Graph Data.* John Wiley & Sons, 2007.

[Cha03] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data.* Morgan Kaufmann, 2003.

[CMS09] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice.* Addison Wesley, 2009.

[De01] S. Dzeroski and N. Lavrac (eds.). *Relational Data Mining.* Springer, 2001.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification* (2nd ed.). John Wiley & Sons, 2001.

[Dob90] A. J. Dobson. *An Introduction to Generalized Linear Models.* Chapman and Hall, 1990.

[EN06] R. Elmasri and S. B. Navathe. *Fundamental of Database Systems* (5th ed.). Addison Wesley, 2006.

[Fe94] U. M. Fayyad and R. Uthurusamy (eds.). *Notes of AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94)*. Seattle, WA, July 1994.

[Fe95] U. M. Fayyad and R. Uthurusamy (eds.). *Proc. 1995 Int. Conf. Knowledge Discovery and Data Mining (KDD'95)*. AAAI Press, Aug. 1995.

[FPSSe96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.

[FUe93] U. M. Fayyad, R. Uthurusamy, and G. Piatetsky-Shapiro (eds.). *Notes of AAAI'93 Workshop Knowledge Discovery in Databases (KDD'93)*. Washington, DC, July 1993.

[GF04] D. A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*. Springer, 2004.

[GM99] A. Gupta and I. S. Mumick. *Materialized Views: Techniques, Implementations, and Applications*. MIT Press, 1999.

[GMP+09] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, Feb. 2009.

[GMUW08] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book* (2nd ed.). Prentice Hall, 2008.

[HMS01] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[HS05] J. M. Hellerstein and M. Stonebraker. *Readings in Database Systems* (4th ed.). MIT Press, 2005.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer-Verlag, 2009.

[IGG03] C. Imhoff, N. Galemmo, and J. G. Geiger. *Mastering Data Warehouse Design : Relational and Dimensional Techniques*. John Wiley & Sons, 2003.

[Inm96] W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 1996.

[Je97] K. S. Jones and P. Willett (eds.). *Readings in Information Retrieval* (3rd ed.). Morgan Kaufmann, 1997.

[JW92] R. A. Johnson and D. A. Wichern. *Applied Multivariate Statistical Analysis* (3rd ed.). Prentice Hall, 1992.

[KHYM08] H. Kargupta, J. Han, P. S. Yu, and R. Motwani. *Next Generation of Data Mining.* Chapman & Hall/CRC, 2008.

[KM90] Y. Kodratoff and R. S. Michalski. *Machine Learning, An Artificial Intelligence Approach, Vol. 3.* Morgan Kaufmann, 1990.

[KR02] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd ed.). John Wiley & Sons, 2002.

[KRTM08] R. Kimball, M. Ross, W. Thornthwaite, and J. Mundy. *The Data Warehouse Lifecycle Toolkit.* John Wiley & Sons, 2008.

[Liu06] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer, 2006.

[MA03] S. Mitra and T. Acharya. *Data Mining: Multimedia, Soft Computing, and Bioinformatics.* John Wiley & Sons, 2003.

[MBK98] R. S. Michalski, I. Brakto, and M. Kubat. *Machine Learning and Data Mining: Methods and Applications.* John Wiley & Sons, 1998.

[MCM83] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 1.* Morgan Kaufmann, 1983.

[MCM86] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine Learning, An Artificial Intelligence Approach, Vol. 2.* Morgan Kaufmann, 1986.

[MH09] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery (2nd ed.).* Chapman & Hall/CRC, 2009.

[Mit97] T. M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[MRS08] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[MT94] R. S. Michalski and G. Tecuci. *Machine Learning, A Multistrategy Approach, Vol. 4.* Morgan Kaufmann, 1994.

[NKNW96] J. Neter, M. H. Kutner, C. J. Nachtsheim, and L. Wasserman. *Applied Linear Statistical Models* (4th ed.). Irwin, 1996.

[PS89] G. Piatetsky-Shapiro. *Notes of IJCAI'89 Workshop Knowledge Discovery in Databases (KDD'89).* Detroit, MI, July 1989.

[PS91] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–238. AAAI/MIT Press, 1991.

[PSF91] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[RG03] R. Ramakrishnan and J. Gehrke. *Database Management Systems* (3rd ed.). McGraw-Hill, 2003.

[RS01] F. Ramsey and D. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury Press, 2001.

[SD90] J. W. Shavlik and T. G. Dietterich. *Readings in Machine Learning*. Morgan Kaufmann, 1990.

[Shu88] R. H. Shumway. *Applied Statistical Time Series Analysis*. Prentice Hall, 1988.

[SKS05] A. Silberschatz, H. F. Korth, and S. Sudarshan. *Database System Concepts* (5th ed.). McGraw-Hill, 2005.

[TSK05] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2005.

[WF05] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann, 2005.

[WI98] S. M. Weiss and N. Indurkhya. *Predictive Data Mining*. Morgan Kaufmann, 1998.

[YFe10] P. S. Yu, C. Faloutsos, and J. Han (eds.). *Link Mining: Models, Algorithms and Applications*. Springer, 2010.

[ZZ09] Z. Zhang and R. Zhang. *Multimedia Data Mining: A Systematic Introduction to Concepts and Theory*. Chapman & Hall, 2009.