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, we learn what is data mining and why data mining is in high demand. Section 1.2 links data mining with the overall knowledge discovery process. Next, we learn about data mining from multiple 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 relationship between data mining and other disciplines (Section 1.5), and data mining applications (Section 1.6). Finally, we discuss the impact of data mining to society (Section 1.7).

1.1 What is data mining?

Necessity, who is the mother of invention.
– Plato

We live in a world where vast amounts of data are generated constantly and rapidly.

"*We are living in the information age*" is a popular saying; however, *we are actually living in the data age*. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various kinds of devices every day from business, news agencies, society, science, engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful computing, sensing, and data collection, storage, and publication tools.

Businesses worldwide generate gigantic data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback. Scientific and engineering practices generate high orders of petabytes of data in a continuous manner, from remote sensing, to process measuring, scientific experiments, system performance, engineering observations, and environment surveillance. Biomedical research and the health industry generate tremendous amounts of data from gene sequence machines, biomedical experiment and research reports, medical records, patient monitoring, and medical imaging. Billions of Web searches supported by search engines process tens of petabytes of data daily. Social media tools have become increasingly popular, producing a tremendous number of texts, pictures, and videos, generating various kinds of Web communities and social networks. The list of sources that generate huge amounts of data is endless.

Data Mining. https://doi.org/10.1016/B978-0-12-811760-6.00011-4

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 the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining.

Essentially, ***data mining*** *is the process of discovering interesting patterns, models, and other kinds of knowledge in large data sets*. The term, *data mining*, as a vivid view of searching for *gold nuggets* from data, appeared in 1990s. However, to refer to the mining of gold from rocks or sand, we say *gold mining* instead of rock or sand mining. Analogously, 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. Thus, such a misnomer carrying both "data" and "mining" became a popular choice. In addition, many other terms have a similar meaning to data mining—for example, *knowledge mining from data*, *KDD (i.e., Knowledge Discovery from Data)*, *pattern discovery*, *knowledge extraction*, *data archaeology*, *data analytics*, and *information harvesting*.

Data mining is a young, dynamic, and promising field. It has made and will continue to make great strides in our journey from the data age toward the coming information age.

Example 1.1. Data mining turns a large collection of data into knowledge. A search engine (e.g., Google) receives billions of queries every day. What novel and useful knowledge can a search engine learn from such a huge collection of 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's *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, *Flu Trends* can estimate flu activity up to two weeks faster than what traditional systems can.[1] This example shows how data mining can turn a large collection of data into knowledge that can help meet a current global challenge. □

1.2 Data mining: an essential step in knowledge discovery

Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**, whereas others view data mining as merely an essential step in the overall process of knowledge discovery. The overall knowledge discovery process is shown in Fig. 1.1 as an iterative sequence of the following steps:

1. **Data preparation**
   a. **Data cleaning** (to remove noise and inconsistent data)
   b. **Data integration** (where multiple data sources may be combined)[2]

[1] This is reported in [GMP+09]. The *Flu Trend* reporting stopped in 2015.

[2] 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.

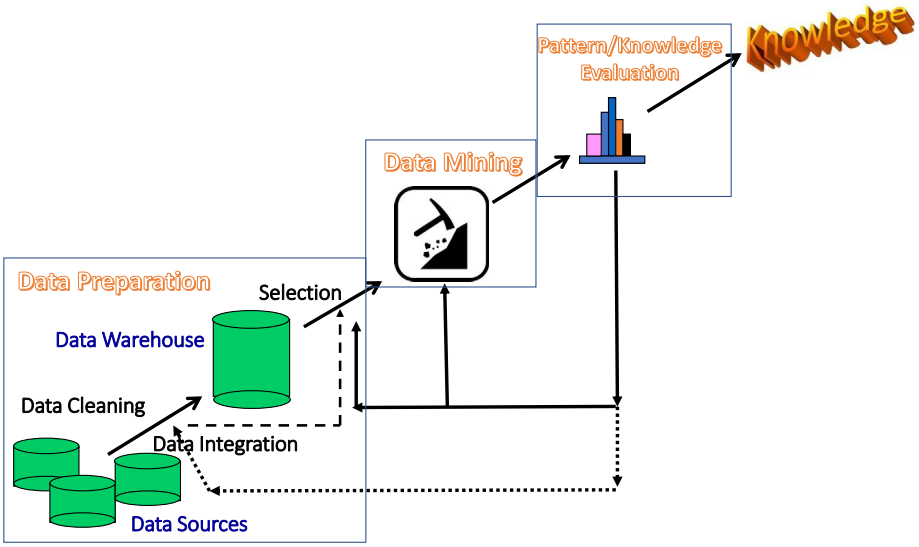

FIGURE 1.1

Data mining: An essential step in the process of knowledge discovery.

c. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)[3]
d. **Data selection** (where data relevant to the analysis task are retrieved from the database)

2. **Data mining** (an essential process where intelligent methods are applied to extract patterns or construct models)
3. **Pattern/model evaluation** (to identify the truly interesting patterns or models representing knowledge based on *interestingness measures*—see Section 1.4.7)
4. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1(a) through 1(d) are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with a 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 preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns or models 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 is shorter than *knowledge discovery from data*). Therefore, we adopt a broad view of data mining functionality: *Data mining is the* process *of discovering inter-*

[3] Data transformation and consolidation are often 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.

esting patterns and knowledge from large *amounts of data*. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

1.3 Diversity of data types for data mining

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. However, different kinds of data may need rather different data mining methodologies, from simple to rather sophisticated, making data mining a rich and diverse field.

Structured vs. unstructured data

Based on whether data have clear structures, we can categorize data as *structured* vs. *unstructured data*.

Data stored in *relational databases*, *data cubes*, *data matrices*, and many *data warehouses* have uniform, record- or table-like structures, defined by their data dictionaries, with a fixed set of attributes (or fields, columns), each with a fixed set of value ranges and semantic meaning. These data sets are typical examples of highly structured data. In many real applications, such strict structural requirement can be relaxed in multiple ways to accommodate *semistructured* nature of the data, such as to allow a data object to contain a set value, a small set of heterogeneous typed values, or nested structures or to allow the structure of objects or subobjects to be defined flexibly and dynamically (e.g., XML structures).

There are many data sets that may not be as structured as relational tables or data matrices. However, they do have certain structures with clearly defined semantic meaning. For example, a *transactional data set* may contain a large set of transactions each containing a set of items. A *sequence data set* may contain a large set of sequences each containing an ordered set of elements that can in turn contain a set of items. Many application data sets, such as shopping transaction data, time-series data, gene or protein data, or Weblog data, belong to this category.

A more sophisticated type of semistructured data set is *graph or network data*, where a set of nodes are connected by a set of edges (also called links); and each node/link may have its own semantic description or substructures.

Each of such categories of structured and semistructured data sets may have special kinds of patterns or knowledge to be mined and many dedicated data mining methods, such as sequential pattern mining, graph pattern mining, and information network mining methods, have been developed to analyze such data sets.

Beyond such structured or semistructured data, there are also large amounts of unstructured data, such as text data and multimedia (e.g., audio, image, video) data. Although some studies treat them as one-dimensional or multidimensional byte streams, they do carry a lot of interesting semantics. Domain-specific methods have been developed to analyze such data in the fields of natural language understanding, text mining, computer vision, and pattern recognition. Moreover, recent advances on deep learning have made tremendous progress on processing text, image, and video data. Nevertheless, mining hidden structures from unstructured data may greatly help understand and make good use of such data.

The real-world data can often be a mixture of structured data, semistructured data, and unstructured data. For example, an online shopping website may host information for a large set of products, which

can be essentially structured data stored in a relational database, with a fixed set of fields on product name, price, specifications, and so on. However, some fields may essentially be text, image, and video data, such as product introduction, expert or user reviews, product images, and advertisement videos. Data mining methods are often developed for mining some particular type of data, and their results can be integrated and coordinated to serve the overall goal.

Data associated with different applications

Different applications may generate or need to handle very different data sets and require rather different data analysis methods. Thus when categorizing data sets for data mining, we should take specific applications into consideration.

Take sequence data as an example. *Biological sequences* such as DNA or protein sequences may have very different semantic meaning from *shopping transaction sequences* or *Web click streams*, calling for rather different sequence mining methods. A special kind of sequence data is time-series data where a *time-series* may contain an ordered set of numerical values with equal time interval, which is also rather different from shopping transaction sequences, which may not have fixed time gaps (a customer may shop at anytime she likes).

Data in some applications can be associated with spatial information, time information, or both, forming *spatial, temporal, and spatiotemporal data*, respectively. Special data mining methods, such as spatial data mining, temporal data mining, spatiotemporal data mining, or trajectory pattern mining, should be developed for mining such data sets as well.

For graph and network data, different applications may also need rather different data mining methods. For example, social networks (e.g., Facebook or LinkedIn data), computer communication networks, biological networks, and information networks (e.g., authors linking with keywords) may carry rather different semantics and require different mining methods.

Even for the same data set, finding different kinds of patterns or knowledge may require different data mining methods. For example, for the same set of software (source) programs, finding plagiarized subprogram modules or finding copy-and-paste bugs may need rather different data mining techniques.

Rich data types and diverse application requirements call for very diverse data mining methods. Thus data mining is a rich and fascinating research domain, with lots of new methods waiting to be studied and developed.

Stored vs. streaming data

Usually, data mining handles finite, stored data sets, such as those stored in various kinds of large data repositories. However, in some applications such as video surveillance or remote sensing, data may stream in dynamically and constantly, as infinite *data streams*. Mining stream data will require rather different methods than stored data, which may form another interesting theme in our study.

1.4 Mining various kinds of knowledge

Different kinds of patterns and knowledge can be uncovered via data mining. In general, data mining tasks can be put into two categories: **descriptive data mining** and **predictive data mining**. Descriptive mining characterizes properties of the interested set of data, whereas predictive mining performs induction on the data set in order to make predictions.

In this section, we introduce different data mining tasks. These include multidimensional data summarization (Section 1.4.1); the mining of frequent patterns, associations, and correlations (Section 1.4.2); classification and regression (Section 1.4.3); cluster analysis (Section 1.4.4); and outlier analysis (Section 1.4.6). Different data mining functionalities generate different kinds of results that are often called patterns, models, or knowledge. In Section 1.4.7, we will also introduce the interestingness of a pattern or a model. In many cases, only interesting patterns or models will be considered as *knowledge*.

1.4.1 Multidimensional data summarization

It is often tedious for a user to go over the details of a large set of data. Thus it is desirable to automatically summarize an interested set of data and compare it with the contrasting sets at some high levels. Such summaritive description of an *interested set of data* is called **data summarization**. Data summarization can often be conducted in a multidimensional space. If the multidimensional space is well defined and frequently used, such as product category, producer, location, or time, massive amounts of data can be aggregated in the form of **data cubes** to facilitate user's drill-down or roll-up of the summarization space with mouse clicking. The output of such multidimensional summarization can be presented in various forms, such as **pie charts**, **bar charts**, **curves**, **multidimensional data cubes**, and **multidimensional tables**, including crosstabs.

For structured data, multidimensional aggregation methods have been developed to facilitate such precomputation or online computation of multidimensional aggregations using data cube technology, which will be discussed in Chapter 3. For unstructured data, such as text, this task becomes challenging. We will give a brief discussion of such research frontiers in our last chapter.

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 often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a computer bag, and then other accessories, is a (*frequent*) *sequential pattern*. A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (*frequent*) *structured pattern*. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Example 1.2. Association analysis. Suppose that, a webstore manager wants to know which items are frequently purchased together (i.e., in the same transaction). An example of such a rule, mined from the transactional database, is

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

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 webcam as well. A 1% **support** means

that 1% of all the transactions under analysis show that computer and webcam are purchased together. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as "*computer* $\Rightarrow$ *webcam* [1%, 50%]."

Suppose, mining the same database generates another association rule:

$$age(X, \text{“20..29”}) \wedge income(X, \text{“40K..49K”}) \Rightarrow buys(X, \text{“laptop”})$$
$$[support = 0.5\%, confidence = 60\%].$$

The rule indicates that of all its customers under study, 0.5% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer). There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., *age, income*, and *buys*). Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule**. □

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. Mining frequent itemsets, associations, and correlations will be discussed in Chapter 4. Mining diverse kinds of frequent pattern, as well as mining sequential patterns and structured patterns, will be covered in Chapter 5.

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. The model is derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class labels of objects for which the class labels are unknown.

Depending on the classification methods, a derived model can be in various forms, such as a set of *classification rules* (i.e., *IF-THEN rules*), a *decision tree*, a *mathematical formula*, or a learned *neural network* (Fig. 1.2). A **decision tree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily 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 **feature selection** or **relevance analysis**, which attempts to identify attributes (often called *features*) that are significantly relevant to the clas-

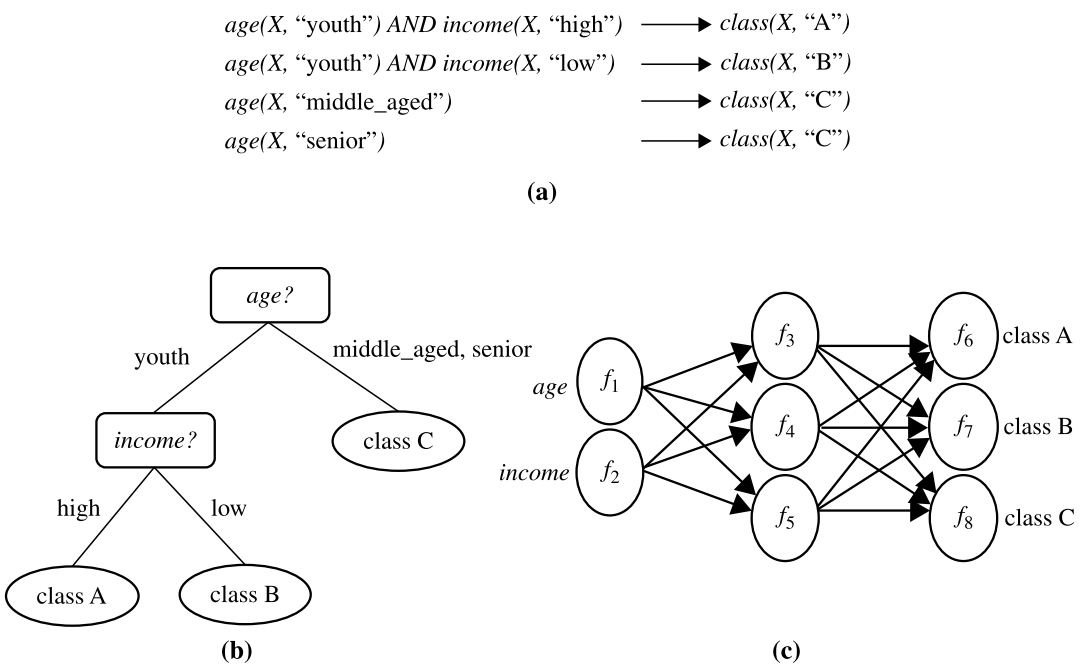

FIGURE 1.2

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

sification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration.

Example 1.3. Classification and regression. Suppose a webstore sales manager wants 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 want 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 resulting classification is expressed as a decision tree. The decision tree, for instance, may identify *price* as being the first important factor that best distinguishes the three classes. Other features that help further distinguish objects of each class from one another include *brand* and *place_made*. Such a decision tree may help the manager understand the impact of the given sales campaign and design a more effective campaign in 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, 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 6 and 7 discuss classification in further detail. Regression analysis is covered lightly in these chapters since it is typically introduced in statistics courses. Sources for further information are given in the bibliographic notes.

1.4.4 Cluster analysis

Unlike classification and regression, which analyze class-labeled (training) data sets, **cluster analysis** (also called **clustering**) groups data objects without consulting class labels. In many cases, class-labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate **taxonomy formation**, that is, the organization of observations into a hierarchy of classes that group similar events together.

Example 1.4. Cluster analysis. Cluster analysis can be performed on the webstore customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing. Fig. 1.3 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 8 and 9.

1.4.5 Deep learning

For many data mining tasks, such as classification and clustering, a key step often lies in finding "good features," which is a vector representation of each input data tuple. For example, in order to predict

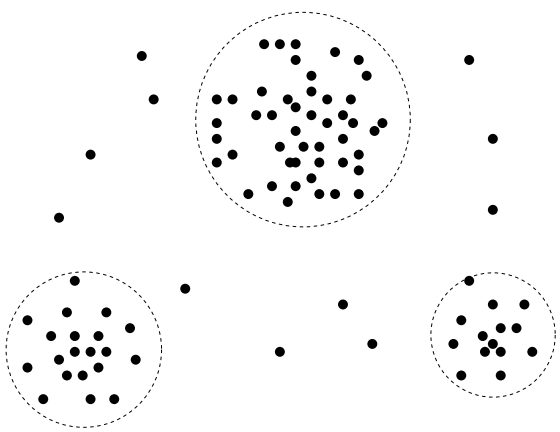

FIGURE 1.3

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

whether a regional disease outbreak will occur, one might have collected a large number of features from the health surveillance data, including the number of daily positive cases, number of daily tests, number of daily hospitalization, etc. Traditionally, this step (called feature engineering) often heavily relies on domain knowledge. Deep learning techniques provide an automatic way for feature engineering, which is capable of generating semantically meaningful features (e.g., weekly positive rate) from the initial input features. The generated features often significantly improve the mining performance (e.g., classification accuracy).

Deep learning is based on *neural networks*. A neural network is a set of connected input-output units where each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights to be able to predict the correct target values (e.g., class labels) of the input tuples. The core algorithm to learn such weights is called *backpropagation*, which searches for a set of weights and bias values that can model the data to minimize the loss function between the network's prediction and the actual target output of data tuples. Various forms (called architectures) of neural networks have been developed, including feed-forward neural networks, convolutional neural networks, recurrent neural networks, graph neural networks, and many more.

Deep learning has broad applications in computer vision, natural language processing, machine translation, social network analysis, and so on. It has been used in a variety of data mining tasks, including classification, clustering, outlier detection, and reinforcement learning.

Deep learning is the topic of Chapter 10.

1.4.6 Outlier analysis

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are **outliers**. Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining**.

Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers. Rather than using statistical or distance measures, density-based methods may identify outliers in a local region, although they look normal from a global statistical distribution view.

Example 1.5. 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 11.

1.4.7 Are all mining results interesting?

Data mining has the potential to generate a lot of results. A question can be, *"Are all of the mining results interesting?"*

This is a great question. Each type of data mining functions has its own measures on the evaluation of the mining quality. Nevertheless, there are some shared philosophy and principles.

Take pattern mining as an example. Pattern mining may generate thousands or even millions of patterns, or rules. You may wonder, *"What makes a pattern interesting? Can a data mining system generate all of the interesting patterns? Or, can the system generate only the interesting ones?"*

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

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.

There are also other objective measures. For example, one may like set of items to be strongly correlated in an association rule. We will discuss such measures in the corresponding chapter.

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

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

The second question—"*Can a data mining system generate* all *of the interesting patterns?*"—refers to the **completeness** of a data mining algorithm. It is often unrealistic and inefficient for a pattern mining system to generate all possible patterns since there could be a very large number of them. However, one may also worry whether one may miss some important ones if the system stops short. To solve this dilemma, user-provided constraints and interestingness measures should be used to focus the search. With well-defined interesting measures and user-provided constraints, it is quite realistic to ensure the completeness of pattern mining. The methods involved are examined in detail in Chapter 4.

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 a data mining system to generate only

interesting patterns. This would be efficient for both the data mining system and the user because the system may spend much less time to generate much fewer but interesting patterns, whereas the user will not need to sift through a large number of patterns to identify the truly interesting ones. Constraint-based pattern mining described in Chapter 5 is a good example in this direction.

Methods to assess the quality or interestingness of data mining results, and how to use them to improve data mining efficiency, are discussed throughout the book.

1.5 Data mining: confluence of multiple disciplines

As a discipline that studies efficient and effective methods for uncovering patterns and knowledge from various kinds of massive data sets for many applications, data mining naturally serves a confluence of multiple disciplines including machine learning, statistics, pattern recognition, natural language processing, database technology, visualization and human computer interaction (HCI), algorithms, high-performance computing, social sciences, and many application domains (Fig. 1.4). The interdisciplinary nature of data mining research and development contributes significantly to the success of data mining and its extensive applications. On the other hand, data mining is not only nurtured from the knowledge and development of these disciplines, the dedicated research, development, and applications of data mining on various kinds of big data may have substantially impacted the development of these disciplines in recent years as well. In this section, we discuss several disciplines that strongly impact and actively interact with the research, development, and applications of data mining.

1.5.1 Statistics and data mining

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

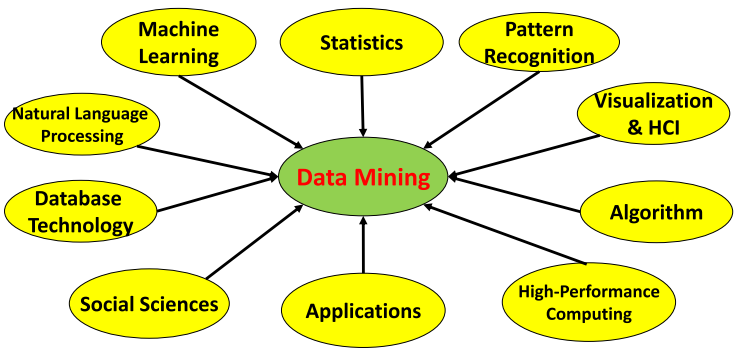

FIGURE 1.4

Data mining: Confluence of multiple disciplines.

models are widely used to model data and data classes. For example, in data mining tasks such as 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. Basic **statistical descriptions** of data are introduced in Chapter 2. Statistics is useful for mining various patterns from data and for understanding the underlying mechanisms generating and affecting the patterns. **Inferential statistics** (or **predictive statistics**) models data 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, then the descriptive statistics of the model increases 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 search engines, where data mining is required to continuously handle fast, real-time data streams.

Data mining research has developed many scalable and effective solutions for the analysis of massive data sets and data streams. Moreover, different kinds of data sets and different applications may require rather different analysis methods. Effective solutions have been proposed and tested, which leads to many new, scalable data mining-based statistical analysis methods.

1.5.2 Machine learning and data mining

Machine learning investigates how computers can learn (or improve their performance) based on data. Machine learning is a fast-growing discipline, with many new methodologies and applications developed in recent years, from support vector machines to probabilistic graphical models and deep learning, which we will cover in this book.

In general, machine learning addresses two classical problems: *supervised learning* and *unsupervised learning*.

- **Supervised learning**: A classic example of supervised learning is classification. The supervision in the learning comes from the labeled examples in the training data set. For example, to automatically recognize handwritten postal codes on mails, the learning system takes a set of handwritten postal code images and their corresponding machine-readable translations as the training examples, and learns (i.e., computes) a classification model.
- **Unsupervised learning**: A classic example of unsupervised learning is clustering. The learning process is unsupervised since the input examples are not class-labeled. Typically, we may use clustering

to discover groups 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 hopefully 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.

As to these two basic problems, data mining and machine learning do share many similarities. However, data mining differs from machine learning in several major aspects. First, even on similar tasks like classification and clustering, data mining often works on very large data sets, or even on infinite data streams, scalability can be an important concern, and many efficient and highly scalable data mining algorithms or stream mining algorithms have to be developed to accomplish such tasks.

Second, in many data mining problems, the data sets are usually large, but the training data can still be rather small since it is expensive for experts to provide quality labels for many examples. Therefore, data mining has to put a lot of effort on developing *weakly supervised methods*. These include methodologies like *semisupervised learning* with a small set of labeled data but a large set of unlabeled data (with the idea sketched in Fig. 1.5), *integration or ensemble of multiple weak models* obtained from nonexperts (e.g., those obtained by crowd-sourcing), *distant supervision*, such as using popularly available and general (but distantly relevant to the problem to be solved) knowledge-bases (e.g., wikipedia, DBPedia), *actively learning* by carefully selecting examples to ask human experts, or *transfer learning* by integrating models learned from similar problem domains. Data mining has been extending such weakly supervised methods for constructing quality classification models on large data sets with a very limited set of high quality training data.

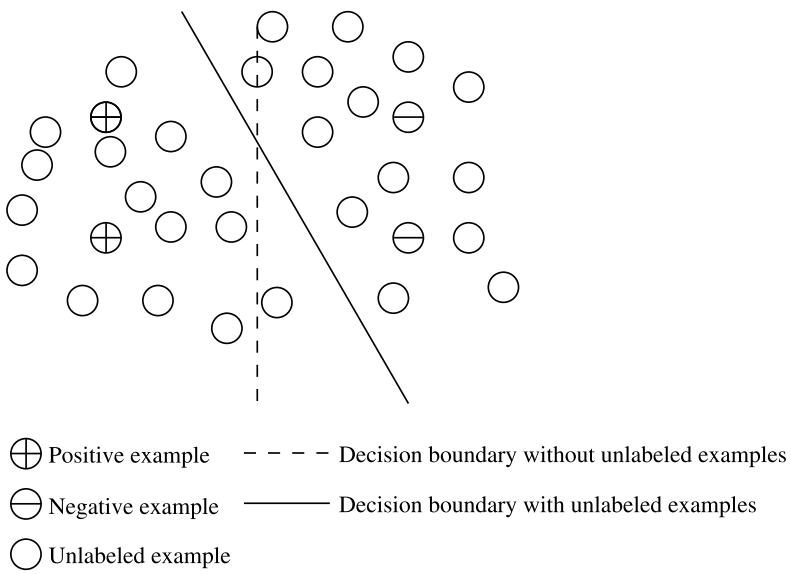

FIGURE 1.5

Semisupervised learning.

Third, machine learning methods may not be able to handle many kinds of knowledge discovery problems on big data. On the other hand, data mining, developing effective solutions for concrete application problems, goes deep in the problem domain, and expands far beyond the scope covered by machine learning. For example, many application problems, such as business transaction data analysis, software program execution sequence analysis, and chemical and biological structural analysis, need effective methods for mining frequent patterns, sequential patterns, and structured patterns. Data mining research has generated many scalable, effective, and diverse mining methods for such tasks. As another example, the analysis of large-scale social and information networks poses many challenging problems that may not fit the typical scope of many machine learning methods due to the information interaction across links and nodes in such networks. Data mining has developed a lot of interesting solutions to such problems.

From this point of view, data mining and machine learning are two different but closely related disciplines. Data mining dives deep into concrete, data-intensive application domains, does not confine itself to a single problem-solving methodology, and develops concrete (sometimes rather novel), effective and scalable solutions for many challenging application problems. It is a young, broad, and promising research discipline for many researchers and practitioners to study and work on.

1.5.3 Database technology and data mining

Database system research focuses on the creation, maintenance, and use of databases for organizations and end-users. Particularly, database system researchers have established well-recognized principles in data models, query languages, query processing and optimization, data storage, and indexing methods. Database technology is well known for its 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. 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 users' sophisticated data analysis requirements.

Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities. A **data warehouse** integrates data originated 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*, which will be further discussed in future chapters.

1.5.4 Data mining and data science

With the tremendous amount of data in almost every discipline and various kinds of applications, big data and data science have become buzzwords in recent years. **Big data** generally refers to huge amounts of structured and unstructured data of various forms, and **data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from massive data of various forms. Clearly, data mining plays an essential role in data science.

For most people, data science is a concept that unifies statistics, machine learning, data mining, and their related methods in order to understand and analyze massive data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and

computer science. For many industry people, the term "data science" often refers to business analytics, business intelligence, predictive modeling, or any meaningful use of data, and is being taken as a glamorized term to re-brand statistics, data mining, machine learning, or any kind of data analytics. So far, there exists no consensus on a definition or suitable curriculum contents in data science degree programs of many universities. Nonetheless, most universities take basic knowledge generated in statistics, machine learning, data mining, database, and human computer interaction as the core curriculum in data science education.

In 1990s, the late Turing award winner Jim Gray envisioned data science as the "fourth paradigm" of science (i.e., from empirical to theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the emergence of massive data. So there is no wonder that data science, big data, and data mining are closely interrelated and represent an inevitable trend in science and technology developments.

1.5.5 Data mining and other disciplines

Besides statistics, machine learning, and database technology, data mining has close relationships with many other disciplines as well.

The majority of the real-world data are unstructured, in the form of natural language text, images, or audio-video data. Therefore, natural language processing, computer vision, pattern recognition, audio-video signal processing, and information retrieval will offer critical help at handling such data. Actually, handling any special kinds of data will need a lot of domain knowledge to be integrated into the data mining algorithm design. For example, mining biomedical data will need the integration of knowledge from biological sciences, medical sciences, and bioinformatics. Mining geospatial data will need much knowledge and techniques from geography and geospatial data sciences. Mining software bugs in large software programs will need to integrate software engineering with data mining. Mining social media and social networks will need knowledge and skills from social sciences and network sciences. Such examples can go on and on since data mining will penetrate almost every application domain.

One major challenge in data mining is efficiency and scalability since we often have to handle huge amounts of data with critical time and resource constraints. Data mining is critically connected with efficient algorithm design such as low-complexity, incremental, and streaming data mining algorithms. It often needs to explore high performance computation, parallel computation, and distributed computation, with advanced hardware and cloud computing or cluster computing environment.

Data mining is also closely tied with human-computer interaction. Users need to interact with a data mining system or process in an effective way, telling the system what to mine, how to incorporate background knowledge, how to mine, and how to present the mining results in an easy-to-understand (e.g., by interpretation and visualization) and easy-to-interact way (e.g., with friendly graphic user interface and interactive mining).

Actually, nowadays, there are not only many interactive data mining systems but also many more data mining functions hidden in various kinds of application programs. It is unrealistic to expect everyone in our society to understand and master data mining techniques. It is also forbidden for industries to expose their large data sets. Many systems have data mining functions built within so that people can perform data mining or use data mining results simply by mouse clicking. For example, intelligent search engines and online retails perform such **invisible data mining** by collecting their data and user's search or purchase history, incorporating data mining into their components to improve their performance, functionality, and user satisfaction. When your grandma shops online, she may be surprised

when receiving some smart recommendations. This could likely be resulted from such invisible data mining.

1.6 Data mining and applications

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 beyond the scope of this book. To demonstrate the importance of applications of data mining, we briefly discuss a few highly successful and popular application examples of data mining: *business intelligence*; *search engines*; *social media and social networks*; and *biology, medical science, and health care*.

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, business performance management, competitive intelligence, benchmarking, and predictive analytics.

"How important is data mining in business intelligence?" Without data mining, many businesses may not be able to perform effective market analysis, compare customer feedback on similar products, discover the strengths and weaknesses of their competitors, retain highly valuable customers, and make smart business decisions.

Clearly, data mining is the core of business intelligence. Online analytical processing tools in business intelligence rely on data warehousing and multidimensional data mining. Classification and prediction techniques are the core of predictive analytics in business intelligence, 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 multidimensional summarization techniques, we can better understand features of each customer group and develop customized customer reward programs.

Web search engines

A **Web search engine** is a specialized computer server that searches for information on the 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 search engines also search and return data available in public databases or open directories. Search engines differ from **web directories** in that web directories are maintained by human editors, whereas search engines operate algorithmically or by a mixture of algorithmic and human input.

Search engines pose grand challenges to data mining. First, they have to handle a huge and ever-growing amount of data. Typically, such data cannot be processed using one or a few machines. Instead, search engines often need to use *computer clouds*, which consist of thousands or even hundreds of thou-

sands of computers that collaboratively mine the huge amount of data. Scaling up data mining methods over computer clouds and large distributed data sets is an area of active research and development.

Second, Web search engines often have to deal with online data. A search engine may be able to afford constructing a model offline on huge datasets. To do this, it may construct a query classifier that assigns a search query to predefined categories based on the query topic (i.e., whether the search query "apple" is meant to retrieve information about a fruit or a brand of computers). Even if a model is constructed offline, the adaptation 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 search engine wants to provide *context-aware* query recommendations. That is, when a user poses a query, the search engine tries to infer the context of the query using the user's profile and his query history in order to return more customized answers within a small fraction of a second. However, although the total number of queries asked can be huge, many queries may be asked only once or a few times. Such severely skewed data are challenging for many data mining and machine learning methods.

Social media and social networks

The prevalence of social media and social networks has fundamentally changed our life and the way we exchange information and socialize nowadays. With tremendous amounts of social media and social network data available, it is critical to analyze such data to extract actionable patterns and trends from social media and social network data.

Social media mining is to sift through massive amounts of social media data (e.g., on social media usage, online social behaviors, connections between individuals, online shopping behavior, content exchange, etc.) in order to discern patterns and trends. These patterns and trends have been used for social event detection, public health monitoring and surveillance, sentiment analysis in social media, recommendation in social media, information provenance, social media trustability analysis, and social spammer detection.

Social network mining is to investigate social network structures and the information associated with such networks through the use of networks and graph theory and data mining methods. The social network structures are characterized in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, memes spread, friendship and acquaintance networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines.

Social network mining has been used to detect hidden communities, uncover the evolution and dynamics of social networks, compute network measures (e.g., centrality, transitivity, reciprocity, balance, status, and similarity), analyze how information propagates in social media sites, measure and model node/substructure influence and homophily, and conduct location-based social network analysis.

Social media mining and social network mining are important applications of data mining.

Biology, medical science, and health care

Biology, medical science and health care have also been generating massive data at exponential scale. Biomedical data take many forms, from "omics" to imaging, mobile health, and electronic health records. With the availability of more efficient digital collection methods, biomedical scientists and clinicians now find themselves confronting ever larger sets of data and trying to devise creative ways to sift through this mountain of data and make sense of it. Indeed, data that used to be considered large now seems small as the amount of data now being collected in a single day by an investigator can surpass what might have been generated over his/her career even a decade ago. This deluge of biomedical information requires new thinking about how data can be managed and analyzed to further scientific understanding and for improving healthcare.

Biomedical data mining involves many challenging data mining tasks, including mining massive genomic and proteomic sequence data, mining frequent subgraph patterns for classifying biological data, mining regulatory networks, characterization and prediction of protein-protein interactions, classification and predictive analysis of medical images, biological text mining, biological information network construction from biotext data, mining electronic health records, and mining biomedical networks.

1.7 Data mining and society

With data mining penetrating our everyday lives, it is important to study the impact of data mining on society. How can we use data mining technology to benefit society? How can we guard against its misuse? The 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.

Data mining will help scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyberattacks). However, it also poses the risk of unintentionally disclosing some confidential business or government information and disclosing an individual's personal information. Studies on data security in data mining and privacy-preserving data publishing and data mining are important, ongoing research theme. The philosophy is to observe data sensitivity and preserve data security and people's privacy while performing successful data mining.

These issues and many additional ones relating to the research, development, and application of data mining will be 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 and knowledge from massive amounts of data. As a *knowledge discovery process*, it typically involves data cleaning, data integration, data selection, data transformation, pattern and model discovery, pattern or model evaluation, and knowledge presentation.

- A pattern or model 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.
- Data mining can be conducted on any kind of **data** as long as the data are meaningful for a target application, such as structured data (e.g., relational database, transaction data) and unstructured data (e.g., text and multimedia data), as well as data associated with different applications. Data can also be categorized as stored vs. stream data, whereas the latter may need to explore special stream mining algorithms.
- **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; deep learning; 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, is a confluence of multiple disciplines but it has its unique research focus, dedicated to many advanced applications. We study the close relationships of data mining with statistics, machine learning, database technology, and many other disciplines.
- Data mining has many successful **applications**, such as business intelligence, Web search, bioinformatics, health informatics, finance, digital libraries, and digital governments.
- Data mining may already have its strong impact on the society and the study of such impact, such as how to ensure the effectiveness of data mining and in the meantime ensure the data privacy and security, has become an important issue in research.

1.9 Exercises

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

- **a.** Is it a simple transformation or application of technology developed from *databases*, *statistics*, *machine learning*, and *pattern recognition*?
- **b.** Someone believes that data mining is an inevitable result of the evolution of information technology. If you are a database researcher, show data mining is resulted from a nature evolution of database technology. What about if you are a machine learner researcher, or a statistician?
- **c.** Describe the steps involved in data mining when viewed as a process of knowledge discovery.

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

1.3. 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 generated alternatively by data query processing or simple statistical analysis?

1.4. Explain the difference and similarity between correlation analysis and classification, between classification and clustering, and between classification and regression.

1.5. Based on your observations, 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?

1.6. *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 fraud detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

1.7. What are the major challenges of mining a huge amount of data (e.g., billions of tuples) in comparison with mining a small amount of data (e.g., data set of a few hundred tuples)?

1.8. 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 another early collection of research results on knowledge discovery and data mining. There have been many data mining textbook or research books published since then. Some popular ones include *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.) by Witten, Frank, Hall and Pal [WFHP16]; *Data Mining: Concepts and Techniques* (3rd ed.) by Han and Kamber and Pei [HKP11], *Introduction to Data Mining* (2nd ed.) by Tan, Steinbach, Karpatne, and Kumar [TSKK18]; *Data Mining: The Textbook* [Agg15b]; *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (2nd ed.) by Zaki and Meira [ZJ20]; *Mining of Massive Datasets* (3rd ed.) by Leskovec, Rajaraman and Ullman [ZJ20]; *The Elements of Statistical Learning* (2nd ed.) by Hastie, Tibshirani, and Friedman [HTF09]; *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.) by Linoff and Berry [LB11]; *Principles of Data Mining (Adaptive Computation and Machine Learning)* by Hand, Mannila, and Smyth [HMS01]; *Mining the Web: Discovering Knowledge from Hypertext Data* by Chakrabarti [Cha03]; *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* by Liu [Liu06]; and *Data Mining: Multimedia, Soft Computing, and Bioinformatics* by Mitra and Acharya [MA03].

There are also numerous books that contain collections of papers or chapters on particular aspects of knowledge discovery, such as cluster analysis, outlier detection, classification, association mining, and mining particular kinds of data, such as mining text data, multimedia data, relational data, geospatial data, social and information network data, and social media data. However, this list has gone very long over the years and we will not list them individually. There are numerous tutorial notes on data mining in major data mining, database, machine learning, statistics, and Web technology conferences.

KDNuggets is a regular electronic newsletter containing information relevant to knowledge discovery and data mining, moderated by Piatetsky-Shapiro since 1991. The Internet site *KDNuggets* (*https://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. The conference evolved from the four international workshops on knowledge discovery in databases, held from 1989 to 1994. ACM-SIGKDD, a Special Interest Group on Knowl-

edge Discovery in Databases was set up under ACM in 1998 and has been organizing the international conferences on knowledge discovery and data mining since 1999. 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 Springer, has been available since 1997. An ACM journal, *ACM Transactions on Knowledge Discovery from Data*, published its first volume in 2007.

ACM-SIGKDD also publishes a bi-annual newsletter, *SIGKDD Explorations*. There are a few other international or regional conferences on data mining, such as the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD), the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and the International Conference on Web Search and Data Mining (WSDM).

Research in data mining has also been popularly published in many textbooks, research books, conferences, and journals on data mining, database, statistics, machine learning, and data visualization.