# CHAPTER 1

# INTRODUCTION

## 1.1 OVERVIEW

Almost every discipline from biology and economics to engineering and marketing measures, gathers, and stores data in some digital form. Retail companies store information on sales transactions, insurance companies keep track of insurance claims, and meteorological organizations measure and collect data concerning weather conditions. Timely and well-founded decisions need to be made using the information collected. These decisions will be used to maximize sales, improve research and development projects, and trim costs. Retail companies must determine which products in their stores are under- or over-performing as well as understand the preferences of their customers; insurance companies need to identify activities associated with fraudulent claims; and meteorological organizations attempt to predict future weather conditions.

Data are being produced at faster rates due to the explosion of internet-related information and the increased use of operational systems to collect business, engineering and scientific data, and measurements from sensors or monitors. It is a trend that will continue into the foreseeable future. The challenges of handling and making sense of this information are significant

*Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*,
Second Edition. Glenn J. Myatt and Wayne P. Johnson.
© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

because of the increasing volume of data, the complexity that arises from the diverse types of information that are collected, and the reliability of the data collected.

The process of taking raw data and converting it into meaningful information necessary to make decisions is the focus of this book. The following sections in this chapter outline the major steps in a data analysis or data mining project from defining the problem to the deployment of the results. The process provides a framework for executing projects related to data mining or data analysis. It includes a discussion of the steps and challenges of (1) defining the project, (2) preparing data for analysis, (3) selecting data analysis or data mining approaches that may include performing an optimization of the analysis to refine the results, and (4) deploying and measuring the results to ensure that any expected benefits are realized. The chapter also includes an outline of topics covered in this book and the supporting resources that can be used alongside the book's content.

## 1.2 SOURCES OF DATA

There are many different sources of data as well as methods used to collect the data. Surveys or polls are valuable approaches for gathering data to answer specific questions. An interview using a set of predefined questions is often conducted over the phone, in person, or over the internet. It is used to elicit information on people's opinions, preferences, and behavior. For example, a poll may be used to understand how a population of eligible voters will cast their vote in an upcoming election. The specific questions along with the target population should be clearly defined prior to the interviews. Any bias in the survey should be eliminated by selecting a random sample of the target population. For example, bias can be introduced in situations where only those responding to the questionnaire are included in the survey, since this group may not be representative of a random sample of the entire population. The questionnaire should not contain leading questions—questions that favor a particular response. Other factors which might result in segments of the total population being excluded should also be considered, such as the time of day the survey or poll was conducted. A well-designed survey or poll can provide an accurate and cost-effective approach to understanding opinions or needs across a large group of individuals without the need to survey everyone in the target population.

Experiments measure and collect data to answer specific questions in a highly controlled manner. The data collected should be reliably measured; in other words, repeating the measurement should not result in substantially different values. Experiments attempt to understand cause-and-effect phenomena by controlling other factors that may be important. For example, when studying the effects of a new drug, a double-blind study is typically used. The sample of patients selected to take part in the study is divided into two groups. The new drug is delivered to one group, whereas a placebo (a sugar pill) is given to the other group. To avoid a bias in the study on the part of the patient or the doctor, neither the patient nor the doctor administering the treatment knows which group a patient belongs to. In certain situations it is impossible to conduct a controlled experiment on either logistical or ethical grounds. In these situations a large number of observations are measured and care is taken when interpreting the results. For example, it would not be ethical to set up a controlled experiment to test whether smoking causes health problems.

As part of the daily operations of an organization, data is collected for a variety of reasons. *Operational databases* contain ongoing business transactions and are accessed and updated regularly. Examples include supply chain and logistics management systems, customer relationship management databases (CRM), and enterprise resource planning databases (ERP). An organization may also be automatically monitoring operational processes with sensors, such as the performance of various nodes in a communications network. A *data warehouse* is a copy of data gathered from other sources within an organization that is appropriately prepared for making decisions. It is not updated as frequently as operational databases. Databases are also used to house historical polls, surveys, and experiments. In many cases data from in-house sources may not be sufficient to answer the questions now being asked of it. In these cases, the internal data can be augmented with data from other sources such as information collected from the web or literature.

## 1.3 PROCESS FOR MAKING SENSE OF DATA

### 1.3.1 Overview

Following a predefined process will ensure that issues are addressed and appropriate steps are taken. For exploratory data analysis and data mining projects, you should carefully think through the following steps, which are summarized here and expanded in the following sections:

1. **Problem definition and planning**: The problem to be solved and the projected deliverables should be clearly defined and planned, and an appropriate team should be assembled to perform the analysis.

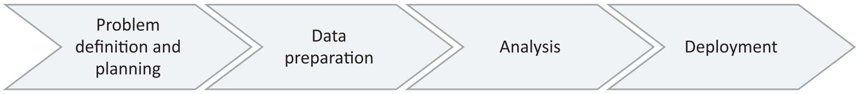

**FIGURE 1.1** Summary of a general framework for a data analysis project.

2. **Data preparation**: Prior to starting a data analysis or data mining project, the data should be collected, characterized, cleaned, transformed, and partitioned into an appropriate form for further processing.
3. **Analysis**: Based on the information from steps 1 and 2, appropriate data analysis and data mining techniques should be selected. These methods often need to be optimized to obtain the best results.
4. **Deployment**: The results from step 3 should be communicated and/or deployed to obtain the projected benefits identified at the start of the project.

Figure 1.1 summarizes this process. Although it is usual to follow the order described, there will be interactions between the different steps that may require work completed in earlier phases to be revised. For example, it may be necessary to return to the data preparation (step 2) while implementing the data analysis (step 3) in order to make modifications based on what is being learned.

### 1.3.2 Problem Definition and Planning

The first step in a data analysis or data mining project is to describe the problem being addressed and generate a plan. The following section addresses a number of issues to consider in this first phase. These issues are summarized in Figure 1.2.

**Problem definition and planning**

- Identify the problem or need to be addressed
- List the project's deliverables
- Generate success factors
- Understand each resource and other limitations
- Put together an appropriate team
- Create a plan
- Perform a costs/benefits analysis

**FIGURE 1.2** Summary of some of the issues to consider when defining and planning a data analysis project.

It is important to document the business or scientific problem to be solved along with relevant background information. In certain situations, however, it may not be possible or even desirable to know precisely the sort of information that will be generated from the project. These more open-ended projects will often generate questions by exploring large databases. But even in these cases, identifying the business or scientific problem driving the analysis will help to constrain and focus the work. To illustrate, an e-commerce company wishes to embark on a project to redesign their website in order to generate additional revenue. Before starting this potentially costly project, the organization decides to perform data analysis or data mining of available web-related information. The results of this analysis will then be used to influence and prioritize this redesign. A general problem statement, such as "make recommendations to improve sales on the website," along with relevant background information should be documented.

This broad statement of the problem is useful as a headline; however, this description should be divided into a series of clearly defined deliverables that ultimately solve the broader issue. These include: (1) categorize website users based on demographic information; (2) categorize users of the website based on browsing patterns; and (3) determine if there are any relationships between these demographic and/or browsing patterns and purchasing habits. This information can then be used to tailor the site to specific groups of users or improve how their customers purchase based on the usage patterns found in the analysis. In addition to understanding what type of information will be generated, it is also useful to know how it will be delivered. Will the solution be a report, a computer program to be used for making predictions, or a set of business rules? Defining these deliverables will set the expectations for those working on the project and for its stakeholders, such as the management sponsoring the project.

The success criteria related to the project's objective should ideally be defined in ways that can be measured. For example, a criterion might be to increase revenue or reduce costs by a specific amount. This type of criteria can often be directly related to the performance level of a computational model generated from the data. For example, when developing a computational model that will be used to make numeric projections, it is useful to understand the required level of accuracy. Understanding this will help prioritize the types of methods adopted or the time or approach used in optimizations. For example, a credit card company that is losing customers to other companies may set a business objective to reduce the turnover rate by 10%. They know that if they are able to identify customers likely to switch to a competitor, they have an opportunity to improve retention through additional marketing. To identify these customers, the company decides to build a predictive model and the accuracy of its predictions will affect the level of retention that can be achieved.

It is also important to understand the consequences of answering questions incorrectly. For example, when predicting tornadoes, there are two possible prediction errors: (1) incorrectly predicting a tornado would strike and (2) incorrectly predicting there would be no tornado. The consequence of scenario (2) is that a tornado hits with no warning. In this case, affected neighborhoods and emergency crews would not be prepared and the consequences might be catastrophic. The consequence of scenario (1) is less severe than scenario (2) since loss of life is more costly than the inconvenience to neighborhoods and emergency services that prepared for a tornado that did not hit. There are often different business consequences related to different types of prediction errors, such as incorrectly predicting a positive outcome or incorrectly predicting a negative one.

There may be restrictions concerning what resources are available for use in the project or other constraints that influence how the project proceeds, such as limitations on available data as well as computational hardware or software that can be used. Issues related to use of the data, such as privacy or legal issues, should be identified and documented. For example, a data set containing personal information on customers' shopping habits could be used in a data mining project. However, if the results could be traced to specific individuals, the resulting findings should be anonymized. There may also be limitations on the amount of time available to a computational algorithm to make a prediction. To illustrate, suppose a web-based data mining application or service that dynamically suggests alternative products to customers while they are browsing items in an online store is to be developed. Because certain data mining or modeling methods take a long time to generate an answer, these approaches should be avoided if suggestions must be generated rapidly (within a few seconds) otherwise the customer will become frustrated and shop elsewhere. Finally, other restrictions relating to business issues include the *window of opportunity* available for the deliverables. For example, a company may wish to develop and use a predictive model to prioritize a new type of shampoo for testing. In this scenario, the project is being driven by competitive intelligence indicating that another company is developing a similar shampoo and the company that is first to market the product will have a significant advantage. Therefore, the time to generate the model may be an important factor since there is only a small window of opportunity based on business considerations.

Cross-disciplinary teams solve complex problems by looking at the data from different perspectives. Because of the range of expertise often required, teams are essential—especially for large-scale projects—and it is helpful to consider the different roles needed for an interdisciplinary team. A *project leader* plans and directs a project, and monitors its results. *Domain experts* provide specific knowledge of the subject matter or business problems, including (1) how the data was collected, (2) what the data values mean, (3) the accuracy of the data, (4) how to interpret the results of the analysis, and (5) the business issues being addressed by the project. *Data analysis/mining experts* are familiar with statistics, data analysis methods, and data mining approaches as well as issues relating to data preparation. An *IT specialist* has expertise in integrating data sets (e.g., accessing databases, joining tables, pivoting tables) as well as knowledge of software and hardware issues important for implementation and deployment. *End users* use information derived from the data routinely or from a one-off analysis to help them make decisions. A single member of the team may take on multiple roles such as the role of project leader and data analysis/mining expert, or several individuals may be responsible for a single role. For example, a team may include multiple subject matter experts, where one individual has knowledge of how the data was measured and another has knowledge of how it can be interpreted. Other individuals, such as the project sponsor, who have an interest in the project should be included as interested parties at appropriate times throughout the project. For example, representatives from the finance group may be involved if the solution proposes a change to a business process with important financial implications.

Different individuals will play active roles at different times. It is desirable to involve all parties in the project definition phase. In the data preparation phase, the IT expert plays an important role in integrating the data in a form that can be processed. During this phase, the data analysis/mining expert and the subject matter expert/business analyst will also be working closely together to clean and categorize the data. The data analysis/mining expert should be primarily responsible for ensuring that the data is transformed into a form appropriate for analysis. The analysis phase is primarily the responsibility of the data analysis/mining expert with input from the subject matter expert or business analyst. The IT expert can provide a valuable hardware and software support role throughout the project and will play a critical role in situations where the output of the analysis is to be integrated within an operational system.

With cross-disciplinary teams, communicating within the group may be challenging from time-to-time due to the disparate backgrounds of the members of the group. A useful way of facilitating communication is to define and share glossaries defining terms familiar to the subject matter experts or to the data analysis/data mining experts. Team meetings to share information are also essential for communication purposes.

The extent of the project plan depends on the size and scope of the project. A timetable of events should be put together that includes the preparation, implementation, and deployment phases (summarized in Sections 1.3.3, 1.3.4, and 1.3.5). Time should be built into the timetable for reviews after each phase. At the end of the project, a valuable exercise that provides insight for future projects is to spend time evaluating what did and did not work. Progress will be iterative and not strictly sequential, moving between phases of the process as new questions arise. If there are high-risk steps in the process, these should be identified and contingencies for them added to the plan. Tasks with dependencies and contingencies should be documented using timelines or standard project management support tools such as Gantt charts. Based on the plan, budgets and success criteria can be developed to compare costs against benefits. This will help determine the feasibility of the project and whether the project should move forward.

### 1.3.3 Data Preparation

In many projects, understanding the data and getting it ready for analysis is the most time-consuming step in the process, since the data is usually integrated from many sources, with different representations and formats. Figure 1.3 illustrates some of the steps required for preparing a data set. In situations where the data has been collected for a different purpose, the data will need to be transformed into an appropriate form for analysis. For example, the data may be in the form of a series of documents that requires it to be extracted from the text of the document and converted to a tabular form that is amenable for data analysis. The data should be prepared to mirror as closely as possible the target population about which new questions will be asked. Since multiple sources of data may be used, care must be taken not to introduce errors when these sources are brought together. Retaining information about the source is useful both for bookkeeping and for interpreting the results.

**Data Preparation**

- Access and combine data tables
- Summarize data
- Look for errors
- Transform data
- Segment data

**FIGURE 1.3** Summary of steps to consider when preparing the data.

It is important to characterize the types of attributes that have been collected over the different items in the data set. For example, do the attributes represent discrete categories such as color or gender or are they numeric values of attributes such as temperature or weight? This categorization helps identify unexpected values. In looking at the numeric attribute *weight* collected for a set of people, if an item has the value "low" then we need to either replace this erroneous value or remove the entire record for that person. Another possible error occurs in values for observations that lie outside the typical range for an attribute. For example, a person assigned a weight of 3,000 lb is likely the result of a typing error made during data collection. This categorization is also essential when selecting the appropriate data analysis or data mining approach to use.

In addition to addressing the mistakes or inconsistencies in data collection, it may be important to change the data to make it more amenable for data analysis. The transformations should be done without losing important information. For example, if a data mining approach requires that all attributes have a consistent range, the data will need to be appropriately modified. The data may also need to be divided into subsets or filtered based on specific criteria to make it amenable to answering the problems outlined at the beginning of the project. Multiple approaches to understanding and preparing data are discussed in Chapters 2 and 3.

### 1.3.4 Analysis

As discussed earlier, an initial examination of the data is important in understanding the type of information that has been collected and the meaning of the data. In combination with information from the problem definition, this categorization will determine the type of data analysis and data mining approaches to use. Figure 1.4 summarizes some of the main analysis approaches to consider.

**Analysis**

- Summarizing data
- Exploring relationships between attributes
- Grouping the data
- Identifying non-trivial facts, patterns, and trends
- Building regression models
- Building classification models

**FIGURE 1.4** Summary of tasks to consider when analyzing the data.

One common category of analysis tasks provides summarizations and statements about the data. *Summarization* is a process by which data is reduced for interpretation without sacrificing important information. Summaries can be developed for the data as a whole or in part. For example, a retail company that collected data on its transactions could develop summaries of the total sales transactions. In addition, the company could also generate summaries of transactions by products or stores. It may be important to make statements with measures of confidence about the entire data set or groups within the data. For example, if you wish to make a statement concerning the performance of a particular store with slightly lower net revenue than other stores it is being compared to, you need to know if it is really underperforming or just within an expected range of performance. Data visualization, such as charts and summary tables, is an important tool used alongside summarization methods to present broad conclusions and make statements about the data with measures of confidence. These are discussed in Chapters 2 and 4.

A second category of tasks focuses on the identification of important facts, relationships, anomalies, or trends in the data. Discovering this information often involves looking at the data in many ways using a combination of data visualization, data analysis, and data mining methods. For example, a retail company may want to understand customer profiles and other facts that lead to the purchase of certain product lines. *Clustering* is a data analysis method used to group together items with similar attributes. This approach is outlined in Chapter 5. Other data mining methods, such as *decision trees* or *association rules* (also described in Chapter 5), automatically extract important facts or rules from the data. These data mining approaches—describing, looking for relationships, and grouping—combined with data visualization provide the foundation for basic exploratory analysis.

A third category of tasks involves the development of mathematical models that encode relationships in the data. These models are useful for gaining an understanding of the data and for making predictions. To illustrate, suppose a retail company wants to predict whether specific consumers may be interested in buying a particular product. One approach to this problem is to collect historical data containing different customer attributes, such as the customer's age, gender, the location where they live, and so on, as well as which products the customer has purchased in the past. Using these attributes, a mathematical model can be built that encodes important relationships in the data. It may be that female customers between 20 and 35 that live in specific areas are more likely to buy the product. Since these relationships are described in the model, it can be used to examine a list of prospective customers that also contain information on age, gender, location, and so on, to make predictions of those most likely to buy the product. The individuals predicted by the model as buyers of the product might become the focus of a targeted marketing campaign. Models can be built to predict continuous data values (*regression models*) or categorical data (*classification models*). Simple methods to generate these models include *linear regression*, *logistic regression*, *classification and regression trees*, and *k-nearest neighbors*. These techniques are discussed in Chapter 6 along with summaries of other approaches. The selection of the methods is often driven by the type of data being analyzed as well as the problem being solved. Some approaches generate solutions that are straightforward to interpret and explain which may be important for examining specific problems. Others are more of a "black box" with limited capabilities for explaining the results. Building and optimizing these models in order to develop useful, simple, and meaningful models can be time-consuming.

There is a great deal of interplay between these three categories of tasks. For example, it is important to summarize the data before building models or finding hidden relationships. Understanding hidden relationships between different items in the data can be of help in generating models. Therefore, it is essential that data analysis or data mining experts work closely with the subject matter expertise in analyzing the data.

### 1.3.5 Deployment

In the deployment step, analysis is translated into a benefit to the organization and hence this step should be carefully planned and executed. There are many ways to deploy the results of a data analysis or data mining project, as illustrated in Figure 1.5. One option is to write a report for management or the "customer" of the analysis describing the business or scientific intelligence derived from the analysis. The report should be directed to those responsible for making decisions and focused on significant and actionable items—conclusions that can be translated into a decision that can be used to make a difference. It is increasingly common for the report to be delivered through the corporate intranet.

**Deployment**

- Generate report
- Deploy standalone or integrated decision-support tool
- Measure business impact

**FIGURE 1.5** Summary of deployment options.

When the results of the project include the generation of predictive models to use on an ongoing basis, these models can be deployed as standalone applications or integrated with other software such as spreadsheet applications or web services. The integration of the results into existing operational systems or databases is often one of the most cost-effective approaches to delivering a solution. For example, when a sales team requires the results of a predictive model that ranks potential customers based on the likelihood that they will buy a particular product, the model may be integrated with the customer relationship management (CRM) system that they already use on a daily basis. This minimizes the need for training and makes the deployment of results easier. Prediction models or data mining results can also be integrated into systems accessible by your customers, such as e-commerce websites. In the web pages of these sites, additional products or services that may be of interest to the customer may have been identified using a mathematical model embedded in the web server.

Models may also be integrated into existing operational processes where a model needs to be constantly applied to operational data. For example, a solution may detect events leading to errors in a manufacturing system. Catching these errors early may allow a technician to rectify the problem without stopping the production system.

It is important to determine if the findings or generated models are being used to achieve the business objectives outlined at the start of the project. Sometimes the generated models may be functioning as expected but the solution is not being used by the target user community for one reason or another. To increase confidence in the output of the system, a controlled experiment (ideally double-blind) in the field may be undertaken to assess the quality of the results and the organizational impact. For example, the intended users of a predictive model could be divided into two groups. One group, made up of half of the users (randomly selected), uses the model results; the other group does not. The business impact resulting from the two groups can then be measured. Where models are continually updated, the consistency of the results generated should also be monitored over time.

There are a number of deployment issues that may need to be considered during the implementation phase. A solution may involve changing business processes. For example, a solution that requires the development of predictive models to be used by end users in the field may change the work practices of these individuals. The users may even resist this change. A successful method for promoting acceptance is to involve the end users in the definition of the solution, since they will be more inclined to use a system they have helped design. In addition, in order to understand

and trust the results, the users may require that all results be appropriately explained and linked to the data from which the results were generated.

At the end of a project it is always a useful exercise to look back at what worked and what did not work. This will provide insight for improving future projects.

## 1.4 OVERVIEW OF BOOK

This book outlines a series of introductory methods and approaches important to many data analysis or data mining projects. It is organized into five technical chapters that focus on describing data, preparing data tables, understanding relationships, understanding groups, and building models, with a hands-on tutorial covered in the appendix.

### 1.4.1 Describing Data

The type of data collected is one of the factors used in the selection of the type of analysis to be used. The information examined on the individual attributes collected in a data set includes a categorization of the attributes' scale in order to understand whether the field represents discrete elements such as *gender* (i.e., male or female) or numeric properties such as *age* or *temperature*. For numeric properties, examining how the data is distributed is important and includes an understanding of where the values of each attribute are centered and how the data for that attribute is distributed around the central values. Histograms, box plots, and descriptive statistics are useful for understanding characteristics of the individual data attributes. Different approaches to characterizing and summarizing elements of a data table are reviewed in Chapter 2, as well as methods that make statements about or summarize the individual attributes.

### 1.4.2 Preparing Data Tables

For a given data collection, it is rarely the case that the data can be used directly for analysis. The data may contain errors or may have been collected or combined from multiple sources in an inconsistent manner. Many of these errors will be obvious from an inspection of the summary graphs and statistics as well as an inspection of the data. In addition to cleaning the data, it may be necessary to transform the data into a form more amenable for data analysis. Mapping the data onto new ranges, transforming categorical data (such as different colors) into a numeric form to be used in a mathematical model, as well as other approaches to preparing tabular or nonstructured data prior to analysis are reviewed in Chapter 3.

### 1.4.3 Understanding Relationships

Understanding the relationships between pairs of attributes across the items in the data is the focus of Chapter 4. For example, based on a collection of observations about the population of different types of birds throughout the year as well as the weather conditions collected for a specific region, does the population of a specific bird increase or decrease as the temperature increases? Or, based on a double-blind clinical study, do patients taking a new medication have an improved outcome? Data visualization, such as scatterplots, histograms, and summary tables play an important role in seeing trends in the data. There are also properties that can be calculated to quantify the different types of relationships. Chapter 4 outlines a number of common approaches to understand the relationship between two attributes in the data.

### 1.4.4 Understanding Groups

Looking at an entire data set can be overwhelming; however, exploring meaningful subsets of items may provide a more effective means of analyzing the data.

Methods for identifying, labeling, and summarizing collections of items are reviewed in Chapter 5. These groups are often based upon the multiple attributes that describe the members of the group and represent subpopulations of interest. For example, a retail store may wish to group a data set containing information about customers in order to understand the types of customers that purchase items from their store. As another example, an insurance company may want to group claims that are associated with fraudulent or nonfraudulent insurance claims. Three methods of automatically identifying such groups—*clustering*, *association rules*, and *decision trees*—are described in Chapter 5.

### 1.4.5 Building Models

It is possible to encode trends and relationships across multiple attributes as mathematical models. These models are helpful in understanding relationships in the data and are essential for tasks involving the prediction of items with unknown values. For example, a mathematical model could be built from historical data on the performance of windmills as well as geographical and meteorological data concerning their location, and used to make predictions on potential new sites. Chapter 6 introduces important concepts in terms of selecting an approach to modeling, selecting attributes to include in the models, optimization of the models, as well as methods for assessing the quality and usefulness of the models using data not used to create the model. Various modeling approaches are outlined, including *linear regression*, *logistic regression*, *classification and regression trees*, and *k-nearest neighbors*. These are described in Chapter 6.

### 1.4.6 Exercises

At the conclusion of selected chapters, there are a series of exercises to help in understanding the chapters' material. It should be possible to answer these practical exercises by hand and the process of going through them will support learning the material covered. The answers to the exercises are provided in the book's appendix.

### 1.4.7 Tutorials

Accompanying the book is a piece of software called *Traceis*, which is freely available from the book's website. In the appendix of the book, a series of data analysis and data mining tutorials are provided that provide practical exercises to support learning the concepts in the book using a series of data sets that are available for download.

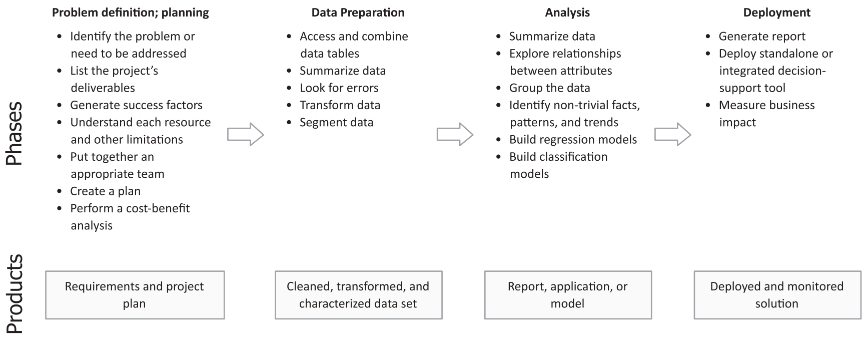
Phases
Problem definition; planning
• Identify the problem or need to be addressed
• List the project's deliverables
• Generate success factors
• Understand each resource and other limitations
• Put together an appropriate team
• Create a plan
• Perform a cost-benefit analysis
Data Preparation
• Access and combine data tables
• Summarize data
• Look for errors
• Transform data
• Segment data
Analysis
• Summarize data
• Explore relationships between attributes
• Group the data
• Identify non-trivial facts, patterns, and trends
• Build regression models
• Build classification models
Deployment
• Generate report
• Deploy standalone or integrated decision-support tool
• Measure business impact
Products
Requirements and project plan
Cleaned, transformed, and characterized data set
Report, application, or model
Deployed and monitored solution

**FIGURE 1.6** Summary of steps to consider in developing a data analysis or data mining project.

## 1.5 SUMMARY

This chapter has described a simple four-step process to use in any data analysis or data mining projects. Figure 1.6 outlines the different stages as well as deliverables to consider when planning and implementing a project to make sense of data.

## FURTHER READING

This chapter has reviewed some of the sources of data used in exploratory data analysis and data mining. The following books provide more information on surveys and polls: Fowler (2009), Rea (2005), and Alreck & Settle (2003). There are many additional resources describing experimental design, including Montgomery (2012), Cochran & Cox (1999), Barrentine (1999), and Antony (2003). Operational databases and data warehouses are summarized in the following books: Oppel (2011) and Kimball & Ross (2013). Oppel (2011) also summarizes access and manipulation of information in databases. The CRISP-DM project (CRoss Industry Standard Process for Data Mining) consortium has published in Chapman et al. (2000) a data mining process covering data mining stages and the relationships between the stages. SEMMA (Sample, Explore, Modify, Model, Assess) describes a series of core tasks for model development in the SAS Enterprise Miner$^{TM}$ software authored by Rohanizadeh & Moghadam (2009). This chapter has focused on issues relating to large and potentially complex data analysis and data mining projects. There are a number of publications that provide a more detailed treatment of general project management issues, including Berkun (2005), Kerzner (2013), and the Project Management Institute (2013). The following references provide additional case studies: Guidici & Figini (2009), Rud (2000), and Lindoff & Berry (2011).