Chapter 1

Introduction

For the keen data miner, Chapter 2 provides a quick-start guide to data mining with Rattle*, working through a sample process of loading a dataset and building a model.*

Data mining is the art and science of intelligent data analysis. The aim is to discover meaningful insights and knowledge from **data**. Discoveries are often expressed as **models**, and we often describe data mining as the process of building models. A model captures, in some formulation, the essence of the discovered knowledge. A model can be used to assist in our **understanding** of the world. Models can also be used to make **predictions**.

For the data miner, the discovery of new knowledge and the building of models that nicely predict the future can be quite rewarding. Indeed, data mining should be exciting and fun as we watch new insights and knowledge emerge from our data. With growing enthusiasm, we meander through our data analyses, following our intuitions and making new discoveries all the time—discoveries that will continue to help change our world for the better.

Data mining has been applied in most areas of endeavour. There are data mining teams working in business, government, financial services, biology, medicine, risk and intelligence, science, and engineering. Anywhere we collect data, data mining is being applied and feeding new knowledge into human endeavour.

We are living in a time where data is collected and stored in unprecedented volumes. Large and small government agencies, commercial enterprises, and noncommercial organisations collect data about their businesses, customers, human resources, products, manufacturing pro-

G. Williams, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*, Use R, DOI 10.1007/978-1-4419-9890-3_1, © Springer Science+Business Media, LLC 2011

cesses, suppliers, business partners, local and international markets, and competitors. Data is the fuel that we inject into the data mining engine.

Turning data into information and then turning that information into knowledge remains a key factor for "success." Data contains valuable information that can support managers in their business decisions to effectively and efficiently run a business. Amongst data there can be hidden clues of the fraudulent activity of criminals. Data provides the basis for understanding the scientific processes that we observe in our world. Turning data into information is the basis for identifying new opportunities that lead to the discovery of new knowledge, which is the linchpin of our society!

Data mining is about building models from data. We build models to gain insights into the world and how the world works so we can predict how things behave. A data miner, in building models, deploys many different data analysis and model building techniques. Our choices depend on the business problems to be solved. Although data mining is not the only approach, it is becoming very widely used because it is well suited to the data environments we find in today's enterprises. This is characterised by the volume of data available, commonly in the gigabytes and terabytes and fast approaching the petabytes. It is also characterised by the complexity of that data, both in terms of the relationships that are awaiting discovery in the data and the data types available today, including text, image, audio, and video. The business environments are also rapidly changing, and analyses need to be performed regularly and models updated regularly to keep up with today's dynamic world.

Modelling is what people often think of when they think of data mining. Modelling is the process of turning data into some structured form or model that reflects the supplied data in some useful way. Overall, the aim is to explore our data, often to address a specific problem, by modelling the world. From the models, we gain new insights and develop a better understanding of the world.

Data mining, in reality, is so much more than simply modelling. It is also about understanding the business context within which we deploy it. It is about understanding and collecting data from across an enterprise and from external sources. It is then about building models and evaluating them. And, most importantly, it is about deploying those models to deliver benefits.

There is a bewildering array of tools and techniques at the disposal of the data miner for gaining insights into data and for building models.

This book introduces some of these as a starting point on a longer journey to becoming a practising data miner.

1.1 Data Mining Beginnings

Data mining, as a named endeavour, emerged at the end of the 1980s from the database community, which was wondering where the next big steps forward were going to come from. Relational database theory had been developed and successfully deployed, and thus began the era of collecting large amounts of data. How do we add value to our massive stores of data?

The first few data mining workshops in the early 1990s attracted the database community researchers. Before long, other computer science, and particularly artificial intelligence, researchers began to get interested. It is useful to note that a key element of "intelligence" is the ability to learn, and machine learning research had been developing technology for this for many years. Machine learning is about collecting observational data through interacting with the world and building models of the world from such data. That is pretty much what data mining was also setting about to do. So, naturally, the machine learning and data mining communities started to come together.

However, statistics is one of the fundamental tools for data analysis, and has been so for over a hundred years. Statistics brings to the table essential ideas about uncertainty and how to make allowances for it in the models that we build. Statistics provides a framework for understanding the "strength" or veracity of models that we might build from data. Discoveries need to be statistically sound and statistically significant, and any uncertainty associated with the modelling needs to be understood. Statistics plays a key role in today's data mining.

Today, data mining is a discipline that draws on sophisticated skills in computer science, machine learning, and statistics. However, a data miner will work in a team together with data and domain experts.

1.2 The Data Mining Team

Many data mining projects work with ill-defined and ambiguous goals. Whilst the first reaction to such an observation is that we should become better at defining the problem, the reality is that often the problem to

be solved is identified and refined as the data mining project progresses. That's natural.

An initiation meeting of a data mining project will often involve data miners, *domain experts*, and *data experts*. The data miners bring the statistical and algorithmic understanding, programming skills, and key investigative ability that underlies any analysis. The domain experts know about the actual problem being tackled, and are often the business experts who have been working in the area for many years. The data experts know about the data, how it has been collected, where it has been stored, how to access and combine the data required for the analysis, and any idiosyncrasies and data traps that await the data miner.

Generally, neither the domain expert nor the data expert understand the needs of the data miner. In particular, as a data miner we will often find ourselves encouraging the data experts to provide (or to provide access to) all of the data, and not just the data the data expert thinks might be useful. As data miners we will often think of ourselves as "greedy" consumers of all the data we can get our hands on.

It is critical that all three experts come together to deliver a data mining project. Their different understandings of the problem to be tackled all need to meld to deliver a common pathway for the data mining project. In particular, the data miner needs to understand the problem domain perspective and understand what data is available that relates to the problem and how to get that data, and identify what data processing is required prior to modelling.

1.3 Agile Data Mining

Building models is only one of the tasks that the data miner performs. There are many other important tasks that we will find ourselves involved in. These include ensuring our data mining activities are tackling the right problem; understanding the data that is available, turning noisy data into data from which we can build robust models; evaluating and demonstrating the performance of our models; and ensuring the effective deployment of our models.

Whilst we can easily describe these steps, it is important to be aware that data mining is an agile activity. The concept of agility comes from the agile software engineering principles, which include the evolution or incremental development of the problem requirements, the requirement

for regular client input or feedback, the testing of our models as they are being developed, and frequent rebuilding of the models to improve their performance.

An allied aspect is the concept of pair programming, where two data miners work together on the same data in a friendly, competitive, and collaborative approach to building models. The agile approach also emphasises the importance of face-to-face communication, above and beyond all of the effort that is otherwise often expended, and often wasted, on written documentation. This is not to remove the need to write documents but to identify what is really required to be documented.

We now identify the common steps in a data mining project and note that the following chapters of this book then walk us through these steps one step at a time!

1.4 The Data Mining Process

The Cross Industry Process for Data Mining (CRISP-DM, 1996) provides a common and well-developed framework for delivering data mining projects. CRISP-DM identifies six steps within a typical data mining project:

1. Problem Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

The chapters in this book essentially follow this step-by-step process of a data mining project, and Rattle is very much based around these same steps. Using a tab-based interface, each tab represents one of the steps, and we proceed through the tabs as we work our way through a data mining project. One noticeable exception to this is the first step, problem understanding. That is something that needs study, discussion, thought, and brain power. Practical tools to help in this process are not common.

1.5 A Typical Journey

Many organisations are looking to set up a data mining capability, often called the analytics team. Within the organisation, data mining projects can be initiated by the business or by this analytics team. Often, for best business engagement, a business-initiated project works best, though business is not always equipped to understand where data mining can be applied. It is often a mutual journey.

Data miners, by themselves, rarely have the deeper knowledge of business that a professional from the business itself has. Yet the business owner will often have very little knowledge of what data mining is about, and indeed, given the hype, may well have the wrong idea. It is not until they start getting to see some actual data mining models for their business that they start to understand the project, the possibilities, and a glimpse of the potential outcomes.

We will relate an actual experience over six months with six significant meetings of the business team and the analytics team. The picture we paint here is a little simplified and idealised but is not too far from reality.

Meeting One The data miners sit in the corner to listen and learn. The business team understands little about what the data miners might be able to deliver. They discuss their current business issues and steps being taken to improve processes. The data miners have little to offer just yet but are on the lookout for the availability of data from which they can learn.

Meeting Two The data miners will now often present some observations of the data from their initial analyses. Whilst the analyses might be well presented graphically, and are perhaps interesting, they are yet to deliver any new insights into the business. At least the data miners are starting to get the idea of the business, as far as the business team is concerned.

Meeting Three The data miners start to demonstrate some initial modelling outcomes. The results begin to look interesting to the business team. They are becoming engaged, asking questions, and understanding that the data mining team has uncovered some interesting insights.

Meeting Four The data miners are the main agenda item. Their analyses are starting to ring true. They have made some quite interesting discoveries from the data that the business team (the domain and data experts) supplied. The discoveries are nonobvious, and sometimes intriguing. Sometimes they are also rather obvious.

Meeting Five The models are presented for evaluation. The data mining team has presented its evaluation of how well the models perform and explained the context for the deployment of the models. The business team is now keen to evaluate the model on real cases and monitor its performance over a period of time.

Meeting Six The models have been deployed into business and are being run daily to match customers and products for marketing, to identify insurance claims or credit card transactions that may be fraudulent, or taxpayers whose tax returns may require refinement. Procedures are in place to monitor the performance of the model over time and to sound alarm bells once the model begins to deviate from expectations.

The key to much of the data mining work described here, in addition to the significance of communication, is the reliance and focus on data. This leads us to identify some key principles for data mining.

1.6 Insights for Data Mining

The starting point with all data mining is the data. We need to have good data that relates to a process that we wish to understand and improve. Without data we are simply guessing.

Considerable time and effort spent getting our data into shape is a key factor in the success of a data mining project. In many circumstances, once we have the right data for mining, the rest is straightforward. As many others note, this effort in data collection and data preparation can in fact be the most substantial component of a data mining project.

My list of insights for data mining, in no particular order, includes:

1. Focus on the data and understand the business.
2. Use training/validate/test datasets to build/tune/evaluate models.
3. Build multiple models: most give very similar performance.
4. Question the "perfect" model as too good to be true.
5. Don't overlook how the model is to be deployed.
6. Stress repeatability and efficiency, using scripts for everything.
7. Let the data talk to you but not mislead you.
8. Communicate discoveries effectively and visually.

1.7 Documenting Data Mining

An important task whilst data mining is the recording of the process. We need to be vigilant to record all that is done. This is often best done through the code we write to perform the analysis rather than having to document the process separately. Having a separate process to document the data mining will often mean that it is rarely completed. An implication of this is that we often capture the process as transparent, executable code rather than as a list of instructions for using a GUI.

There are many important advantages to ensuring we document a project through our coding of the data analyses. There will be times when we need to hand a project to another data miner. Or we may cease work on a project for a period of time and return to it at a later stage. Or we have performed a series of analyses and much the same process will need to be repeated again in a year's time. For whatever reason, when we return to a project, we find the documentation, through the coding, essential in being efficient and effective data miners.

Various things should be documented, and most can be documented through a combination of code and comments. We need to document our access to the source data, how the data was transformed and cleaned, what new variables were constructed, and what summaries were generated to understand the data. Then we also need to record how we built models and what models were chosen and considered. Finally, we record the evaluation and how we collect the data to support the benefit that we propose to obtain from the model.

Through documentation, and ideally by developing documented code that tells the story of the data mining project and the actual process as well, we will be communicating to others how we can mine data. Our processes can be easily reviewed, improved, and automated. We can transparently stand behind the results of the data mining by having openly available the process and the data that have led to the results.

1.8 Tools for Data Mining: R

R is used throughout this book to illustrate data mining procedures. It is the programming language used to implement the Rattle graphical user interface for data mining. If you are moving to R from SAS or SPSS,

then you will find Muenchen (2008) a great resource.[1]

R is a sophisticated statistical software package, easily installed, instructional, state-of-the-art, and it is free and open source. It provides all of the common, most of the less common, and all of the new approaches to data mining.

The basic modus operandi in using R is to write scripts using the R language. After a while you will want to do more than issue single simple commands and rather write programs and systems for common tasks that suit your own data mining. Thus, saving our commands to an R script file (often with the `.R` filename extension) is important. We can then rerun our scripts to transform our source data, at will and automatically, into information and knowledge. As we progress through the book, we will become familiar with the common R functions and commands that we might combine into a script.

Whilst for data mining purposes we will focus on the use of the Rattle GUI, more advanced users might prefer the powerful Emacs editor, augmented with the ESS package, to develop R code directly. Both run under GNU/Linux, Mac/OSX, and Microsoft Windows.

We also note that direct interaction with R has a steeper learning curve than using GUI based systems, but once over the hurdle, performing operations over the same or similar datasets becomes very easy using its programming language interface.

A paradigm that is encouraged throughout this book is that of *learning by example* or *programming by example* (Cypher, 1993). The intention is that anyone will be able to easily replicate the examples from the book and then fine-tune them to suit their own needs. This is one of the underlying principles of Rattle, where all of the R commands that are used under the graphical user interface are also exposed to the user. This makes it a useful teaching tool in learning R for the specific task of data mining, and also a good memory aid!

1.9 Tools for Data Mining: Rattle

Rattle is built on the statistical language R, but an understanding of R is not required in order to use it. Rattle is simple to use, quick to deploy, and allows us to rapidly work through the data processing, modelling, and evaluation phases of a data mining project. On the other hand,

[1] An early version is available from `http://r4stats.com`.

R provides a very powerful language for performing data mining well beyond the limitations that are embodied in any graphical user interface and the consequently canned approaches to data mining. When we need to fine-tune and further develop our data mining projects, we can migrate from Rattle to R.

Rattle can save the current state of a data mining task as a Rattle project. A Rattle project can then be loaded at a later time or shared with other users. Projects can be loaded, modified, and saved, allowing check pointing and parallel explorations. Projects also retain all of the R code for transparency and repeatability. This is an important aspect of any scientific and deployed endeavour—to be able to repeat our "experiments."

Whilst a user of Rattle need not necessarily learn R, Rattle exposes all of the underlying R code to allow it to be directly deployed within the R Console as well as saved in R scripts for future reference. The R code can be loaded into R (outside of Rattle) to repeat any data mining task.

Rattle by itself may be sufficient for all of a user's needs, particularly in the context of introducing data mining. However, it also provides a stepping stone to more sophisticated processing and modelling in R itself. It is worth emphasising that the user is not limited to how Rattle does things. For sophisticated and unconstrained data mining, the experienced user will progress to interacting directly with R.

The typical workflow for a data mining project was introduced above. In the context of Rattle, it can be summarised as:

1. Load a **Data**set.
2. **Select** variables and entities for exploring and mining.
3. **Explore** the data to understand how it is distributed or spread.
4. **Transform** the data to suit our data mining purposes.
5. Build our **Model**s.
6. **Evaluate** the models on other datasets.
7. **Export** the models for deployment.

It is important to note that at any stage the next step could well be a step to a previous stage. Also, we can save the contents of Rattle's **Log** tab as a repeatable record of the data mining process.

We illustrate a typical workflow that is embodied in the Rattle interface in Figure 1.1.

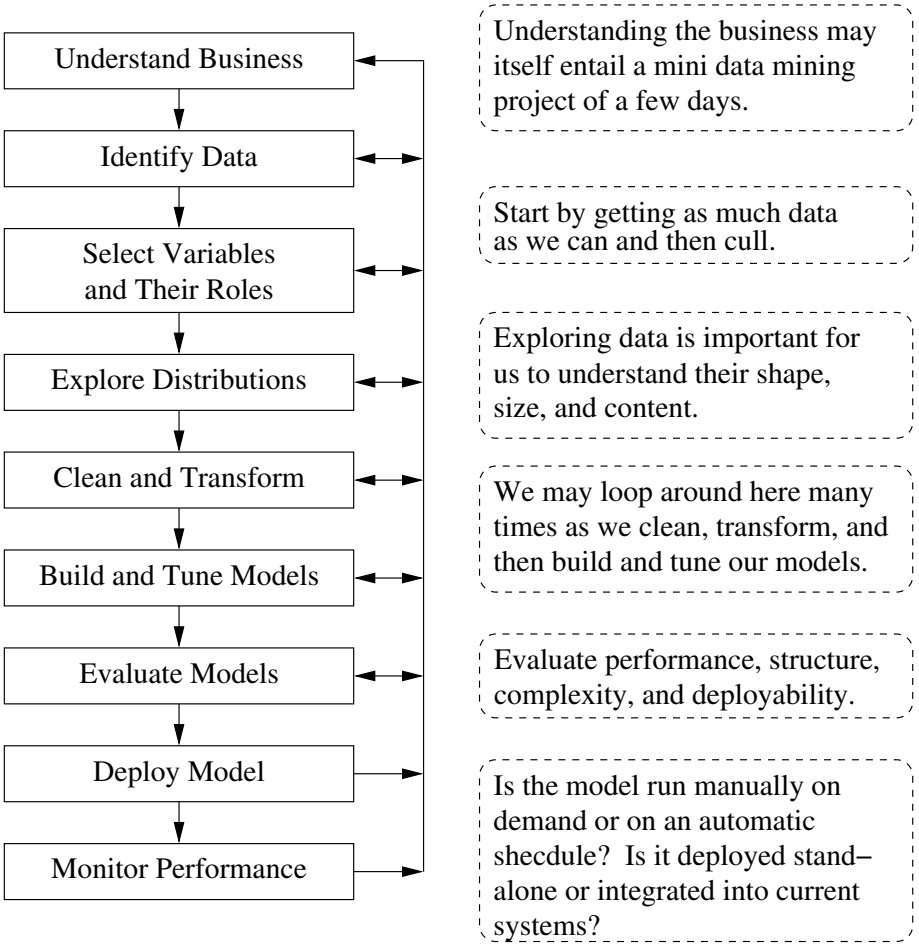

Figure 1.1: The typical workflow of a data mining project as supported by Rattle.

1.10 Why R and Rattle?

R and Rattle are free software in terms of allowing anyone the freedom to do as they wish with them. This is also referred to as open source software to distinguish it from closed source software, which does not provide the source code. Closed source software usually has quite restrictive licenses associated with it, aimed at limiting our freedom using it. This is separate from the issue of whether the software can be obtained for free (which is

often, but not necessarily, the case for open source software) or must be purchased. R and Rattle can be obtained for free.

On 7 January 2009, the New York Times carried a front page technology article on R where a vendor representative was quoted:

> I think it addresses a niche market for high-end data analysts that want free, readily available code. ...We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet.

This is a common misunderstanding about the concept of free and open source software. R, being free and open source software, is in fact a peer-reviewed software product that a number of the worlds top statisticians have developed and others have reviewed. In fact, anyone is permitted to review the R source code. Over the years, many bugs and issues have been identified and rectified by a large community of developers and users.

On the other hand, a closed source software product cannot be so readily and independently verified or viewed by others at will. Bugs and enhancement requests need to be reported back to the vendor. Customers then need to rely on a very select group of vendor-chosen people to assure the software, rectify any bugs in it, and enhance it with new algorithms. Bug fixes and enhancements can take months or years, and generally customers need to purchase the new versions of the software.

Both scenarios (open source and closed source) see a lot of effort put into the quality of their software. With open source, though, we all share it, whereas we can share and learn very little about the algorithms we use from closed source software.

It is worthwhile to highlight another reason for using R in the context of free and commercial software. In obtaining any software, due diligence is required in assessing what is available. However, what is finally delivered may be quite different from what was promised or even possible with the software, whether it is open source or closed source, free or commercial. With free open source software, we are free to use it without restriction. If we find that it does not serve our purposes, we can move on with minimal cost. With closed source commercial purchases, once the commitment is made to buy the software and it turns out not to meet our requirements, we are generally stuck with it, having made the financial commitment, and have to make do.

Moving back to R specifically, many have identified the pros and cons of using this statistical software package. We list some of the advantages with using R:

R is the most comprehensive statistical analysis package available. It incorporates all of the standard statistical tests, models, and analyses, as well as providing a comprehensive language for managing and manipulating data. New technology and ideas often appear first in R.

R is a programming language and environment developed for statistical analysis by practising statisticians and researchers. It reflects well on a very competent community of computational statisticians.

R is now maintained by a core team of some 19 developers, including some very senior statisticians.

The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages.

The validity of the R software is ensured through openly validated and comprehensive governance as documented for the US Food and Drug Administration (R Foundation for Statistical Computing, 2008). Because R is open source, unlike closed source software, it has been reviewed by many internationally renowned statisticians and computational scientists.

R is free and open source software, allowing anyone to use and, importantly, to modify it. R is licensed under the GNU General Public License, with copyright held by The R Foundation for Statistical Computing.

R has no license restrictions (other than ensuring our freedom to use it at our own discretion), and so we can run it anywhere and at any time, and even sell it under the conditions of the license.

Anyone is welcome to provide bug fixes, code enhancements, and new packages, and the wealth of quality packages available for R is a testament to this approach to software development and sharing.

R has over 4800 packages available from multiple repositories specialising in topics like econometrics, data mining, spatial analysis, and bio-informatics.

R is cross-platform. R runs on many operating systems and different hardware. It is popularly used on GNU/Linux, Macintosh, and Microsoft Windows, running on both 32 and 64 bit processors.

R plays well with many other tools, importing data, for example, from CSV files, SAS, and SPSS, or directly from Microsoft Excel, Microsoft Access, Oracle, MySQL, and SQLite. It can also produce graphics output in PDF, JPG, PNG, and SVG formats, and table output for LaTeX and HTML.

R has active user groups where questions can be asked and are often quickly responded to, often by the very people who developed the environment—this support is second to none. Have you ever tried getting support from the core developers of a commercial vendor?

New books for R (the Springer Use R! series) are emerging, and there is now a very good library of books for using R.

Whilst the advantages might flow from the pen with a great deal of enthusiasm, it is useful to note some of the disadvantages or weaknesses of R, even if they are perhaps transitory!

R has a steep learning curve—it does take a while to get used to the power of R—but no steeper than for other statistical languages.

R is not so easy to use for the novice. There are several simple-to-use graphical user interfaces (GUIs) for R that encompass point-and-click interactions, but they generally do not have the polish of the commercial offerings.

Documentation is sometimes patchy and terse, and impenetrable to the non-statistician. However, some very high-standard books are increasingly plugging the documentation gaps.

The quality of some packages is less than perfect, although if a package is useful to many people, it will quickly evolve into a very robust product through collaborative efforts.

There is, in general, no one to complain to if something doesn't work. R is a software application that many people freely devote their own time to developing. Problems are usually dealt with quickly on the open mailing lists, and bugs disappear with lightning speed. Users who do require it can purchase support from a number of vendors internationally.

Many R commands give little thought to memory management, and so R can very quickly consume all available memory. This can be a restriction when doing data mining. There are various solutions, including using 64 bit operating systems that can access much more memory than 32 bit ones.

1.11 Privacy

Before closing out our introduction to data mining and tools for doing it, we need to touch upon the topic of privacy. Laws in many countries can directly affect data mining, and it is very worthwhile to be aware of them and their penalties, which can often be severe.

There are basic principles relating to the protection of privacy that we should adhere to. Some are captured by the privacy principles developed by the international Organisation for Economic Co-operation and Development—the OECD (Organisation for Economic Co-operation and Development (OECD), 1980). They include:

Collection limitation
Data should be obtained lawfully and fairly, while some very sensitive data should not be held at all.

Data quality
Data should be relevant to the stated purposes, accurate, complete, and up-to-date; proper precautions should be taken to ensure this accuracy.

Purpose specification
The purposes for which data will be used should be identified, and the data should be destroyed if it no longer serves the given purpose.

Use limitation
Use of data for purposes other than specified is forbidden.

As data miners, we have a social responsibility to protect our society and individuals for the good and benefit of all of society. Please take that responsibility seriously. Think often and carefully about what you are doing.

1.12 Resources

This book does not attempt to be a comprehensive introduction to using R. Some basic familiarity with R will be gained through our travels in data mining using the Rattle interface and some excursions into R. In this respect, most of what we need to know about R is contained within the book. But there is much more to learn about R and its associated packages. We do list and comment on here a number of books that provide an entrée to R.

A good starting point for handling data in R is *Data Manipulation with R* (Spector, 2008). The book covers the basic data structures, reading and writing data, subscripting, manipulating, aggregating, and reshaping data.

Introductory Statistics with R (Dalgaard, 2008), as mentioned earlier, is a good introduction to statistics using R. *Modern Applied Statistics with S* (Venables and Ripley, 2002) is quite an extensive introduction to statistics using R. Moving more towards areas related to data mining, *Data Analysis and Graphics Using R* (Maindonald and Braun, 2007) provides excellent practical coverage of many aspects of exploring and modelling data using R. *The Elements of Statistical Learning* (Hastie et al., 2009) is a more mathematical treatise, covering all of the machine learning techniques discussed in this book in quite some mathematical depth. If you are coming to R from a SAS or SPSS background, then *R for SAS and SPSS Users* (Muenchen, 2008) is an excellent choice. Even if you are not a SAS or SPSS user, the book provides a straightforward introduction to using R.

Quite a few specialist books using R are now available, including *Lattice: Multivariate Data Visualization with R* (Sarkar, 2008), which covers the extensive capabilities of one of the graphics/plotting packages available for R. A newer graphics framework is detailed in *ggplot2: Elegant Graphics for Data Analysis* (Wickham, 2009). Bivand et al. (2008) cover applied spatial data analysis, Kleiber and Zeileis (2008) cover applied econometrics, and Cowpertwait and Metcalfe (2009) cover time series, to

name just a few books in the R library.

Moving on from R itself and into data mining, there are very many general introductions available. One that is commonly used for teaching in computer science is Han and Kamber (2006). It provides a comprehensive generic introduction to most of the algorithms used by a data miner. It is presented at a level suitable for information technology and database graduates.