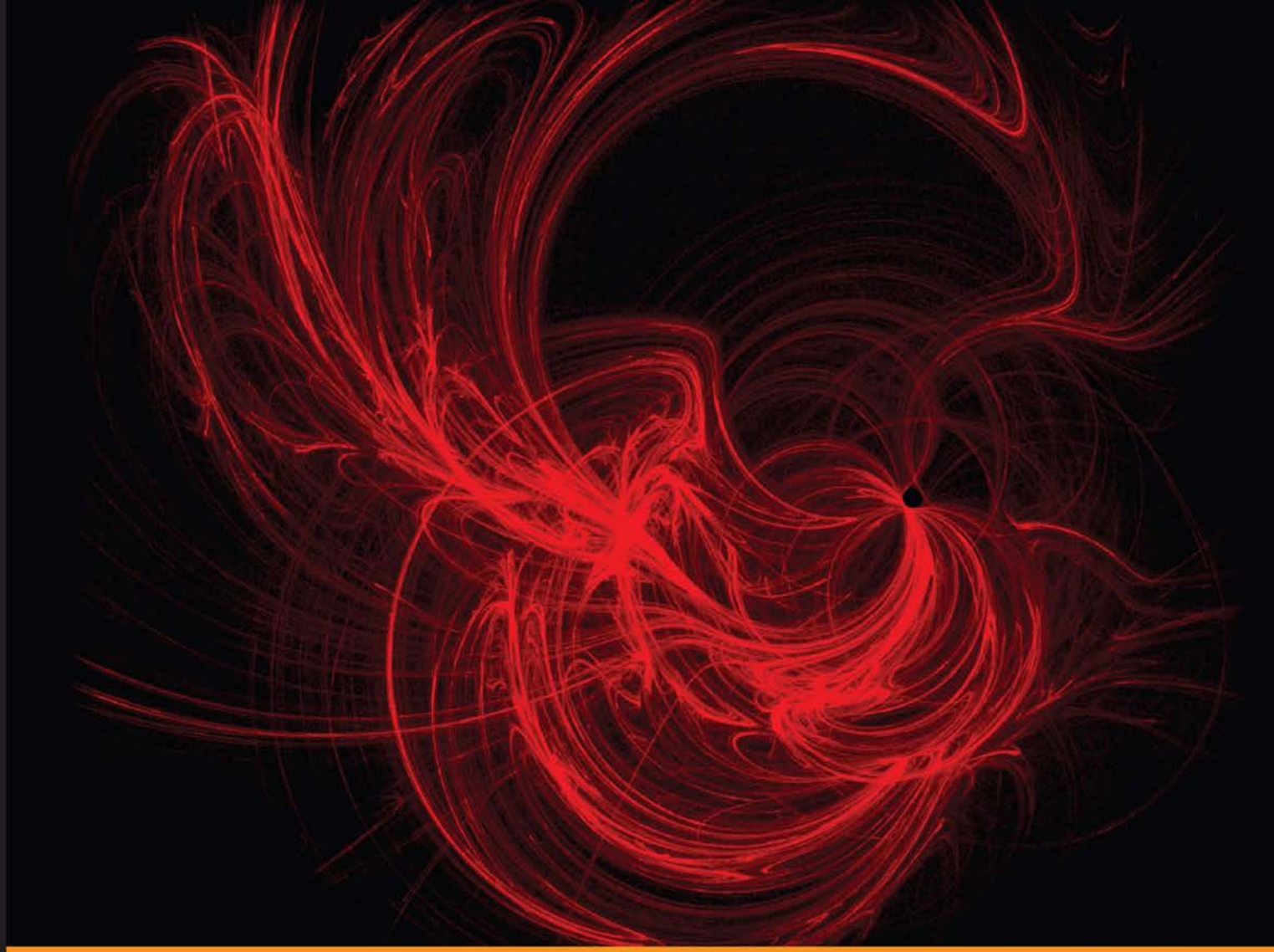


Community Experience Distilled

# Exploring Data with RapidMiner

Explore, understand, and prepare real data using RapidMiner's practical tips and tricks

Andrew Chisholm

[PACKT] PUBLISHING open source* community experience distilled

# Exploring Data with RapidMiner

Explore, understand, and prepare real data using RapidMiner's practical tips and tricks

**Andrew Chisholm**

[PACKT] PUBLISHING
open source
community experience distilled

BIRMINGHAM - MUMBAI

# Exploring Data with RapidMiner



First published: November 2013

Production Reference: 1181113

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78216-933-8

www.packtpub.com

Cover Image by Suresh Mogre (suresh.mogre.99@gmail.com)

# Credits


**Author**
Andrew Chisholm

**Reviewer**
Venkatesh Umaashankar
Ingo Mierswa

**Acquisition Editor**
Pramila Balan

**Commissioning Editor**
Poonam Jain

**Technical Editors**
Pragnesh Bilimoria
Arwa Manasawala
Anand Singh

**Copy Editor**
Alisha Aranha
Roshni Banerjee
Brandt D'Mello
Mradula Hegde
Dipti Kapadia
Kirti Pai

**Project Coordinator**
Suraj Bist

**Proofreader**
Maria Gould

**Indexer**
Rekha Nair

**Graphics**
Ronak Dhruv

**Production Coordinator**
Pooja Chiplunkar

**Cover Work**
Pooja Chiplunkar

# About the Author

**Andrew Chisholm** completed his degree in Physics from Oxford University nearly thirty years ago. This coincided with the growth in software engineering and it led him to a career in the IT industry. For the last decade he has been very involved in mobile telecommunications, where he is currently a product manager for a market-leading test and monitoring solution used by many mobile operators worldwide.

Throughout his career, he has always maintained an active interest in all aspects of data. In particular, he has always enjoyed finding ways to extract value from data and presenting this in compelling ways to help others meet their objectives. Recently, he completed a Master's in Data Mining and Business Intelligence with first class honors. He is a certified RapidMiner expert and has been using this product to solve real problems for several years. He maintains a blog where he shares some miscellaneous helpful advice on how to get the best out of RapidMiner.

He approaches problems from a practical perspective and has a great deal of relevant hands-on experience with real data. This book draws this experience together in the context of exploring data—the first and most important step in a data mining process.

He has published conference papers relating to unsupervised clustering and cluster validity measures and contributed a chapter called *Visualizing cluster validity measures* to an upcoming book entitled *RapidMiner: Use Cases and Business Analytics Applications*, Chapman & Hall/CRC

I would like to thank my family, and in particular my wife Jennie for putting up with me while I wrote this book.

# About the Reviewer

**Venkatesh Umaashankar** is an analytics professional with a rich experience in implementing data mining and machine learning systems. His main areas of interest are machine learning and big data. He is also an avid learner and follower of new developments in the field of machine learning and its practical application. He blogs about machine learning at `http://intelligencemining.blogspot.com`.

# www.PacktPub.com

## Support files, eBooks, discount offers and more

You might want to visit `www.PacktPub.com` for support files and downloads related to your book.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.


`http://PacktLib.PacktPub.com`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across Packt's entire library of books.

## Why Subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print and bookmark content
- On demand and accessible via web browser

## Free Access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view nine entirely free books. Simply use your login credentials for immediate access.

# Table of Contents


**Preface** 1

**Chapter 1: Setting the Scene** 7

- **A process framework** 8
- **Data volume and velocity** 10
- **Data variety, formats, and meanings** 11
- **Missing data** 12
- **Cleaning data** 12
- **Visualizing data** 13
- **Resource constraints** 13
- **Terminology** 14
- **Accompanying material** 15
- **Summary** 16

**Chapter 2: Loading Data** 17

- **Reading files** 17
  - Alternative delimiters 20
  - Reading complete lines 21
  - Reading large numbers of attributes 21
  - Splitting files into smaller pieces 23
- **Databases** 25
  - The Read Database operator 25
  - Large datasets 27
- **Using macros** 27
- **Summary** 28

**Chapter 3: Visualizing Data 29**
**Getting started 29**
**Statistical summaries 30**
**Relationships between attributes 32**
Scatter plots 32
Scatter 3D color 34
Parallel and deviation 35
Quartile color 38
**Time series data 39**
Plotting series 39
Using the survey plotter 42
**Relations between examples 43**
Using histograms 44
Using block plots 45
**Summary 47**
**Chapter 4: Parsing and Converting Attributes 49**
**Generating attributes 50**
Date functions 51
Regular expression functions 53
Generating extracts 54
Regular expressions 54
XPath 57
**Renaming attributes 59**
Searching and replacing attribute values 59
Using the Map operator 59
Using the Replace operator 60
Using the Replace (Dictionary) operator 60
**Summary 62**
**Chapter 5: Outliers 63**
**Manual inspection 63**
Increasing the data volume 68
Rules for handling outliers 68
**Automated detection of example outliers 69**
The Detect Outlier (Distances) operator 69
The Detect Outlier (Densities) operator 73
The Detect Outlier (LOF) operator 74
The Detect Outliers (COF) operator 75
**Summary 76**

**Chapter 6: Missing Values** **77**
**Missing or empty?** **77**
**Types of missing data** **78**
Missing completely at random 78
Missing at random 78
Not missing at random 79
**Categorizing missing data** **79**
Finding MCAR data 83
Finding MAR data 85
Finding NMAR data 86
A cautionary note 87
**Effect of missing data** **88**
**Options for handling missing data** **88**
Returning to the root cause 89
Ignoring it 89
Manual editing 89
Deletion of examples 90
Deletion of attributes 90
Imputation with single values 90
Modeling 91
**Summary** **91**
**Chapter 7: Transforming Data** **93**
**Creating new attributes** **94**
**Aggregation** **98**
**Using pivoting** **100**
**Using de-pivoting** **102**
**Summary** **106**
**Chapter 8: Reducing Data Size** **107**
**Removing examples using sampling** **107**
**Removing attributes** **108**
Removing useless attributes 109
Weighting attributes 111
Selecting attributes using models 114
**Summary** **119**

**Chapter 9: Resource Constraints 121**
**Measuring and estimating performance 121**
Measuring performance 122
**Adding memory 129**
**Parallel processing 130**
**Restructuring processes 131**
**Summary 131**
**Chapter 10: Debugging 133**
**Breakpoints in RapidMiner Studio 133**
**Logging data in RapidMiner Studio 134**
**RapidMiner Studio console printing 135**
**Groovy scripts 136**
Outputting macros example 137
Console logging with Groovy 137
**Regex tools 138**
**Using XPath effectively 138**
**Summary 139**
**Chapter 11: Taking Stock 141**
**Exploring new techniques 142**
Time series 142
Web mining 142
Using R 142
Java or Groovy 142
Third-party components 143
RapidMiner Server 143
**Where to go next 143**
**Index 145**

# Preface

This book is a practical guide to exploring data using RapidMiner Studio. Something like 80 percent of a data mining or predictive analytics project is spent importing, cleaning, visualizing, restructuring, and summarizing data in order to understand it. This book focuses on this vital aspect and gives practical advice using RapidMiner Studio to help with the process.

A number of techniques are illustrated and it is the nature of exploratory data analysis that they can be re-used and modified in different places. By drawing these techniques together into a context, the reader will get a better sense of how RapidMiner Studio can be used in general and gain more confidence to use it.

## What this book covers

*Chapter 1*, *Setting the Scene*, describes the main challenges when mining real data. These challenges arise because data is big and, in the real world, it is unstructured, difficult to visualize, and time consuming to bring order to.

*Chapter 2*, *Loading Data*, describes the different ways of loading data into RapidMiner Studio and the advanced techniques sometimes needed to transform raw unstructured data into a common format.

*Chapter 3*, *Visualizing Data*, describes the visualization techniques available in RapidMiner Studio to help make sense of data.

*Chapter 4*, *Parsing and Converting Attributes*, explains that data is rarely in precisely the right format and, therefore, needs to be parsed to extract specific information or converted into a different representation.

*Chapter 5*, *Outliers*, explains that real data contains values that do not seem to fit the rest of the data. There are many reasons for this and it is important to have a strategy for identifying and dealing with them, otherwise model accuracy risks can be severely compromised.

*Chapter 6*, *Missing Values*, explains that real data inevitably contains missing values. Simple deletion of rows containing missing values can quickly lead to a significant reduction in the performance of a data mining algorithm. Much better techniques exist.

*Chapter 7*, *Transforming Data*, covers techniques to restructure the data into new representations that can assist its exploration and understanding.

*Chapter 8*, *Reducing Data Size*, explains that reducing the number of rows will generally speed up processing but will reduce accuracy. Balancing this is important for large datasets. Reducing the number of columns of data can often improve model accuracy and for large datasets it is doubly valuable as it can speed up processing in general.

*Chapter 9*, *Resource Constraints*, explains that processing large amounts of data requires a lot of physical processing power and memory, to say nothing of the amount of time. This chapter gives some techniques to help measure process performance. Sometimes, it is not possible to process the data using available resources and in this situation, some techniques can be adopted to persuade the process to complete.

*Chapter 10*, *Debugging*, explains that when something goes wrong, it can be frustrating and time consuming to detect and resolve the problem. This chapter gives some generic methods for making this process a bit easier.

*Chapter 11*, *Taking Stock*, explains that having reached this point, the reader will have a greater visibility of the possibilities to process, clean, and explore data as part of the data mining process. This will be a stepping stone to more complex data mining techniques.

# What you need for this book

You will need some basic previous exposure to RapidMiner.The latest version of RapidMiner is now RapidMiner Studio which adds some templating features to help analysts get going more quickly as well as some changes to the look and feel of the GUI in general. This book uses the latest version and it is assumed that you have installed it so that you can download and try the example processes that are worked through in the text.

# Who this book is for

This book is for data analysts with some experience of RapidMiner who wish to use it to explore real data as part of an overall data mining or business intelligence objective. It is very likely that the analyst may have spent some initial time on mining data but could not get the results they wanted. This book gives some real examples and helps to build a context in which data exploration can be done.

# Conventions

In this book, you will find a number of styles of text that distinguish between different kinds of information. Here are some examples of these styles, and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "To identify missing attributes, the `Filter Examples` operator can be used."

**New terms** and **important words** are shown in bold. Words that you see on the screen, in menus or dialog boxes for example, appear in the text like this: "The **featureNames** attribute shows the attributes that were used to create the various performance measurements."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

# Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or may have disliked. Reader feedback is important for us to develop titles that you really get the most out of.

To send us general feedback, simply send an e-mail to `feedback@packtpub.com`, and mention the book title via the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide on `www.packtpub.com/authors`.

# Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

## Downloading the example code

You can download the example code files for all Packt books you have purchased from your account at `http://www.packtpub.com`. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

## Downloading the color images of this book

We also provide you a PDF file that has color images of the screenshots/diagrams used in this book. The color images will help you better understand the changes in the output.You can download this file from: `https://www.packtpub.com/sites/default/files/downloads/9338OS_Images.pdf`.

## Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you would report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **errata submission form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded on our website, or added to any list of existing errata, under the Errata section of that title. Any existing errata can be viewed by selecting your title from `http://www.packtpub.com/support`.

# Piracy

Piracy of copyright material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works, in any form, on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors, and our ability to bring you valuable content.

# Questions

You can contact us at `questions@packtpub.com` if you are having a problem with any aspect of the book, and we will do our best to address it.

# 1
# Setting the Scene

You have data, you know it has hidden value, and you want to mine it. The problem is you're a bit stuck.

The data you have could be anything and you have a lot of it. It is probably from where you work, and you are probably very knowledgeable about how it is gathered, how to interpret it, and what it means. You may also know a domain expert to whom you can turn for additional expertise.

You also have more than a passing knowledge of data mining and you have spent a short time becoming familiar with **RapidMiner** to perform data mining activities, such as clustering, classification, and regression. You know well that mining data is not just a case of using a spreadsheet to draw a few graphs and pie charts; there is much more.

Given all of this, what is the problem, why are you stuck, and what is this book for?

Simply put, real data is huge, stored in a multitude of formats, contains errors and missing values, and does not yield its secrets willingly. If, like me, your first steps in data mining involved using simple test datasets with a few hundred rows (all with clean data), you will quickly find that 10 million rows of data of dubious quality stored in a database combined with some spreadsheets and sundry files presents a whole new set of problems. In fact, estimates put the proportion of time spent cleaning, understanding, interpreting, and exploring data at something like 80 percent. The remaining 20 percent is the time spent on mining.

The problem restated is that if you don't spend time cleaning, reformatting, restructuring, and generally getting to know your data as part of an exploration, you will remain stuck and will get poor results. If we agree that this is the activity to be done, we come to a basic question: how will we do this?

The answer to this problem for this book is to use RapidMiner, a very powerful and ultimately easy-to-use product. These features, coupled with its open source availability, means it is very widely used. It does have a learning curve that can seem daunting. Be assured, once you have ascended this, the product truly becomes easy to use and lives up to its name.

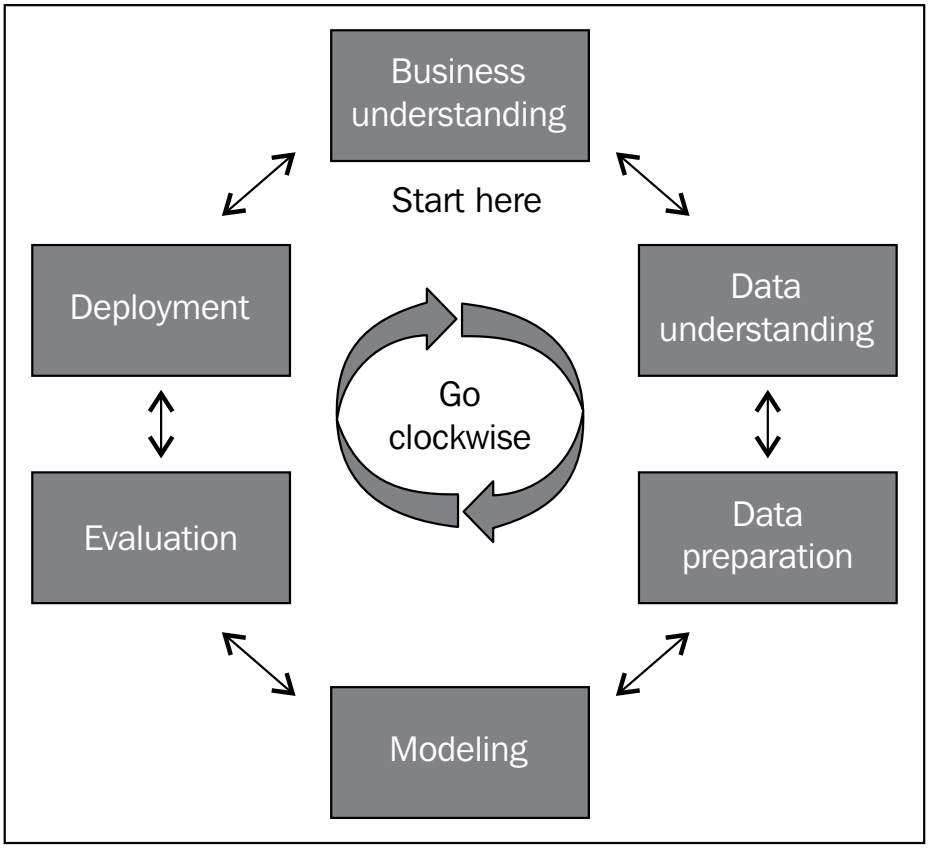
This book is therefore an intermediate-level practical guide to using RapidMiner to explore data and includes techniques to import, visualize, clean, format, and restructure data. This overall objective gives a context in which the various techniques can be considered together. This is helpful because it shows what is possible and makes it easier to modify the techniques for whatever real data is encountered. Hints and tips are provided along the way; in fact, some readers may prefer to use these hints as a starting point.

Having set the scene, let us consider some of the aspects of data exploration raised in this introduction. The following sections explain some of the aspects of data exploration and give references to chapters where these aspects are considered in detail.

# A process framework

It is important to think carefully about the framework within which any data mining investigation is done. A systematic yet simple approach will help results happen and will ensure everyone involved knows what to do and what to expect.

The following diagram shows a simple process framework, derived in part from CRISP-DM (`ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf`):




There are six main phases. The process starts with **Business understanding** and the whole process proceeds in a clockwise direction, but it is quite normal to return, at any stage, to the previous phases in an iterative way. Not all the stages are mandatory. It is possible that the business has an objective that is not related to data mining and modeling at all. It might be enough to summarize large volumes of data in some sort of dashboard, so the **Modeling** step would be ignored in this case.

The **Business understanding** phase is the most important phase to get correct. Without clear organizational objectives set by what we might loosely call the business, as well as its continuing involvement, the whole activity is doomed. The output from this phase is considered the criteria for determining success. For the purpose of this book, it is assumed that this critical phase has been started and this clear view exists.

**Data understanding** and **Data preparation** follow **Business understanding**, and these phases involve activities such as importing, extracting, transforming, cleaning, and loading data into new databases and visualizing and generally getting a thorough understanding of what the data is. This book will be concerned with these two phases.

The **Modeling**, **Evaluation,** and **Deployment** phases concern building models to make predictions, testing these with real data, and deploying them in live use. This is the part that most people regard as data mining but it represents 20 percent of the effort. This book does not concern itself with these phases in any detail.

Having said that, it is important to have a view of the **Modeling** phase that will eventually be undertaken because this will impact the data exploration and understanding activity. For example, a predictive analytics project may try to predict the likelihood of a mobile phone customer switching to a competitor based on usage data. This has implications for how the data should be structured. Another example is using online shopping behavior to predict customer purchases, where a market basket analysis would be undertaken. This might require a different structure for the data. Yet another example would be an unsupervised clustering exercise to try and summarize different types of customers, where the aim is to find groups of similar customers. This can sometimes change the focus of the exploration to find relationships between all the attributes of the data.

**Evaluation** is also important because this is where success is estimated. An unbalanced dataset, where there are few examples of the target to be predicted, will have an effect on the validation to be performed. A regression modeling problem, which estimates a numerical result, will also require a different approach to a classification in which nominal values are being predicted.

Having set the scene for what is to be covered, the following sections will give some more detail about what the **Data understanding** and **Data preparation** phases contain, to give a taste of the chapters to come.

# Data volume and velocity

There is no doubt that data is growing. Even a cursory glance at historical trends and future predictions reveals graphs trending sharply upwards for data volumes, data sources, and datatypes as well as for the speed at which data is being created. There are also graphs showing the cost of data storage going down, linked to the increased power and reduced cost of processing, the presence of new devices such as smartphones, and the ability of standard communication networks such as the Internet to make the movement of data easy.

So, there is more and more data being generated by more and more devices and it is becoming easier to move it around.

However, the ability of people to process and understand data remains constant. The net result is a gap in understanding that is getting wider.

For evidence of this, it is interesting to use Google Trends to look for search terms such as data visualization, data understanding, data value, and data cost. All of these have been trending up to a greater or lesser extent since 2007. This points to the concerns that people have which causes them to search for these terms because they are being overwhelmed with data.

Clearly, there is a need for something to help close the understanding gap to make the process of exploring data more efficient. As the first step, therefore, *Chapter 8*, *Reducing Data Size*, and *Chapter 9*, *Resource Constraints*, give some practical advice on determining how long a RapidMiner process will take to run. Some techniques to sample or reduce the size of data are also included to allow results to be obtained within a meaningful time span while understanding the effect on accuracy.

# Data variety, formats, and meanings

For the purpose of this book, data is something that can be processed by a computer. This means that it is probably stored in a file on a disk or in a database or it could be in the computer's memory. Additionally, it might not physically exist until it is asked for. In other words, it could be the response to a web service query, which mashes up data sources to produce a result. Furthermore, some data is available in real time as a result of some external process being asked to gather or generate results.

Having found the data, understanding its format and the fields within it represents a challenge. With the increase of data volume comes an inevitable increase in the formats of data, owing simply to there being more diverse sources of data. User-generated content, mash-ups, and the possibility of defining one's own XML datatypes means that the meaning and interpretation of a field may not be obvious simply by looking at its name.

The obvious example is date formats. The date 1/5/2012 means January 5, 2012 to someone from the US whereas it means May 1, 2012 to someone from the UK. Another example in the context of a measurement of time is where results are recorded in microseconds, labeled as elapsed time, and then interpreted by a person as being in seconds. Yet another example could be a field labeled Item with the value Bat. Is this referring to a small flying mammal or is it something to play cricket with?

To address some aspects of data, *Chapter 2*, *Loading Data*, *Chapter 4*, *Parsing and Converting Attributes*, and *Chapter 7*, *Transforming Data*, take the initial steps to help close the understanding gap mentioned earlier.

# Missing data

Most data has missing values. These arise for many reasons by virtue of errors during the gathering process, deliberate withholding for legitimate or malicious reasons, and simple bugs in the way data is processed. Having a strategy to handle this is very important because some algorithms perform very poorly even with a small percentage of missing data.

On the face of it, missing data is easy to detect, but there is a pitfall for the unwary since a missing value could in fact be a completely legitimate empty value. For example, a commuter train could start at one station and stop at all intermediate stations before reaching a final destination. An express train would not stop at the intermediate stations at all, and there would be no recorded arrival and departure times for these stops. This is not missing data but if it is handled like it is, the data would become unrepresentative and would lead to unpredictable results when used for mining.

That's not all; there are different types of missing data. Some are completely random, while some depend on the other data in complex ways. It is also possible for missing data to be correlated with the data to be predicted. Any strategy for handling missing values has therefore to consider these issues because the simple strategy of deleting records does not only remove precious data but could also bias the results of any data mining activity. The typical starting approach is to fill missing values manually. This is not advisable because it is time consuming, error prone, risks bias, is not repeatable, and does not scale.

What is needed is a systematic method of handling missing values and determining a way to process them automatically with little or no manual intervention. *Chapter 6*, *Missing Values*, takes the first step on this road.

# Cleaning data

It is almost certain that any data encountered in the real world has data quality issues. In simple terms, this means that values are invalid or very different from other values. Of course, it can get more complex than this when it is not at all obvious that a particular value is anomalous. For example, the heights of people could be recorded and the range could be between 1 and 2 meters. If there is data for young children in the sample, lower heights are expected, but isn't a 2-meter five-year-old child an anomaly? It probably is, but anomalies such as these usually occur.

As with missing data, a systematic and automatic approach is required to identify it and deal with it and *Chapter 5*, *Outliers*, gives some details.

# Visualizing data

A picture paints a thousand words, and this is particularly true when trying to understand data and close the understanding gap. Faced with a million rows of data, there is often no better way to view it to understand what quality issues there are, how the attributes within it relate to one another, and whether there are other systematic features that need to be understood and explained.

There are many types of visualizations that can be used and it is also important to combine these with the use of descriptive statistics, such as the mean and standard deviation.

Examples include 2D and 3D scatter plots, density plots, bubble charts, series, surfaces, box plots, and histograms, and it is often important to aggregate data into summaries for presentation because the larger the data gets, the more time it takes to process. Indeed, it becomes mandatory to summarize data as the resource limits of the available computers are reached.

Some initial techniques are given in *Chapter 3*, *Visualizing Data*.

# Resource constraints

There is never enough time and there is never enough money. In other words, there is never enough time to get all the investigation and processing done, both in terms of the capacity of a person to look at the data and understand it as well as in terms of processing power and capacity. To be valuable in the real world, it must be possible to process all the data in a time that meets the requirements set at the outset. Referring back to the overall process, the business objectives must consider and set acceptance criteria for this.

This pervades all aspects of the data mining process from loading data, cleaning it, handling missing values, transforming it for subsequent processing, and performing the classification or clustering process itself.

When faced with huge data that is taking too long to process, there are many techniques that can be used to speed things up and *Chapter 9*, *Resource Constraints*, gives some details. This can start by breaking the process into steps and ensuring that intermediate results are saved. Very often, an initial load of data from a database can dwarf all other activities in terms of elapsed time. It may also be the case that it is simply not possible to load the data at all, making a batch approach necessary.