*Preface*

# 序言

This book is the result of very interesting discussions, debates and dialogues with a large number of people at various levels of seniority, working at startups as well as long-established businesses, and in a variety of industries, from science to media to finance. The book is intended to be a companion to data analysts and budding data scientists that have some working experience with both programming and statistical modelling, but who have not necessarily delved into the wonders of data analytics and machine learning. The book uses Python1 as a tool to implement and exploit some of the most common algorithms used in data science and data analytics today.

本书是与各级资历人员，创业公司以及历史悠久的企业以及各种行业（从科学，媒体到金融）进行非常有趣的讨论，辩论和对话的结果结晶。 本书旨在成为数据分析师和新兴数据科学家的伴侣指南，因为他们虽然在编程和统计建模方面拥有一定的工作经验，但并未深入研究数据分析和了解机器学习所能创造的奇迹。 本书使用Python作为工具来实现和利用当今数据科学和数据分析中使用的一些通用算法。

(1 Python Software Foundation (1995). Python reference manual. <http://www.python.org>)

（1 Python Software Foundation（1995）.Python 参考书.http：//www.python.org）

It is fair to say that there are a number of very useful tools and platforms available to the interested reader such as the excellent open source R project2 or proprietary ones like SPSS® or SAS®. They are all highly recommended and they have their strengths (and weaknesses). However, given the experience I have been lucky to have had in implementing and explaining algorithms, I find Python to be a very malleable tool. This reminds me of a conversation with an experienced analyst at a big consultancy firm who mentioned that doing any machine learning or data science related task in Python was impossible. I politely disagreed. It is true though that there may be more suitable tools for certain tasks, but it would be a truly Herculean labour to present them all in one single volume. With that in mind, the choice of using Python throughout this book suggested itself: Python is a popular and versatile scripting and object-oriented language, it is easy to use and has a large active community of developers and enthusiasts, not to mention the richness of the iPython/Jupyter Notebook, as well as the fact that it has been used by both business and academia for some time now.

平心而论，感兴趣的读者可以找到很多非常有用的工具和平台，例如优秀的开源R项目或SPSS® SAS®等专业项目。这些都非常推荐，他们有各自的优势（和弱点）。但是，鉴于我在实现和解释算法方面的经验，我发现Python是一个非常具有可塑性的编程工具。这让我想起与在一家大型咨询公司工作的经验丰富的分析师的对话，他提到用Python进行任何机器学习或数据科学相关的任务是不可能的，我其实并不赞同他的观点。诚然，虽然某些特定任务可能有更合适的工具，但将它们全部集中在一个单一的工具中去实现确实是一项艰巨的工作。考虑到这一点，本书中使用Python语言来证明我的选择并没有错：Python是一种流行的多功能脚本和面向对象的语言，它易于使用，并且拥有一个由开发人员和爱好者组成的大型活跃社区，还有丰富多彩的iPython / Jupyter notebook交互式计算环境支持，它已经被商界和学术界使用了很长一段时间。

(We shall show in this book that doing machine learning or data science with Python is indeed possible.)

（我们将在本书中展示用Python进行机器学习以及数据科学分析的无限可能。）

(iPython/Jupyter Notebook is a flexible web-based computational environment that combines code, text, mathematics and plots in a single document. Visit http: //ipython.org/notebook.html)

（iPython / Jupyter Notebook是一个灵活的基于Web的计算环境，它将代码，文本，数学和绘图结合在一个文档中。详见http：//ipython.org/notebook.html）

The main purpose of the book is to present the reader with some of the main concepts used in data science and analytics using tools developed in Python such as Scikit-learn3, Pandas4, Numpy5 and others. The book is intended to be a bridge to the data science and analytics world for programmers and developers, as well as graduates in scientific areas such as mathematics, physics, computational biology and engineering, to name a few. In my experience, the background and skills acquired by the readers I have in mind are a great asset to have. However, in many cases the bigger picture is somewhat blurred due to the sharp specialisms required in their day-to-day activities. This book thus serves as a guide to exploit those skills in the data science and analytics arena. The book focusses on showing the concepts and ideas behind popular algorithms and their use, but it does not get into the details of their implementation in Python. It does, however, use open source implementations of those algorithms.

本书的主要目的是向读者介绍数据科学和分析中使用的一些主要概念，使用Python开发的工具，如Scikit-learn，Pandas，Numpy等。本书旨在为程序员和开发人员以及数学，物理，计算生物学和工程等科学领域的毕业生提供数据科学和分析领域的桥梁，仅举几例。根据我的经验，我想读者目前的背景和技能是非常宝贵的财富。然而，在许多情况下，由于日常活动所需的纵向专业局限，横向知识面难免有些模糊。因此，本书可以作为在数据科学和分析领域的相关指南。我将着重于展示流行算法及其使用背后的概念和思想，但不会探讨它们在Python中的实现细节。如果你确实有相关需求，请参考使用这些算法的开源是如何实现的。

(3 Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 )

(4 McKinney, W. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O’Reilly Media 5 Scientific Computing Tools for Python (2013). NumPy. http://www.numpy.org)

(2 R Core Team (2014). R: A language and environment for statistical computing. http: //www.R-project.org)

The examples contained in this volume have been tested in Python 3.5 under MacOS, Linux and Windows 7, and the code can be run with minimal changes in a Python 2 distribution. For reference, the versions of some of the packages used in the book are as follows:

本卷中包含的示例已在Mac OS，Linux和Windows 7下的Python 3.5中进行了测试，并且代码可以在Python 2发行版中进行最少的更改。 作为参考，本书中使用的一些软件包的版本如下：

• Python - 3.5.2

• Pandas - 0.19.1

• NumPy - 1.11.2

• Scikit-learn - 0.18

• StatsModels - 0.6.1

In particular I have chosen to use the Anaconda Python distribution6 provided by Continuum Analytics as it offers installations in all of the three computer systems mentioned above, plus having the advantage of offering a rich ecosystem of libraries readily available directly from the distribution itself, and most importantly it is available to all. There are a few other ways of obtaining Python as well as other versions of the software: For instance directly from the Python Software Foundation, as well as distributions from Enthought Canopy, or from package managers such as Homebrew. Anaconda offers an easy environment to install and maintain the software, with minimum hassle for the user. I assume that the reader is working with the computer via scripts as well as interactively in a shell.

值得注意的是我选择使用Continuum Analytics提供的Anaconda Python发行版6，因为它提供了上面提到的所有三种计算机系统中的安装，并且具有可以直接从发行版本中提供丰富的类库生态系统的优势。 重要的是它适用于所有人。 还有一些其他方法可以获得Python以及其他版本的软件：例如，直接来自Python软件基金会，以及来自Enthought Canopy的分发，或者来自Homebrew等软件包管理器。 Anaconda提供了一个简单的安装和维护软件环境，为用户减少了很多麻烦。 我假设读者通过脚本以及在shell中交互式地工作。

(6 Continuum Analytics (2014). Anaconda 2.1.0. https://store. continuum.io/cshop/anaconda/)

(Python Software Foundation <https://www.python.org>)

(Python 软件基础 https://www.python.org)

(Enthought Canopy下载 https://www.enthought.com/products/epd/)

(Homebrew下载 http://brew.sh)

The book shows the use of computer code by enclosing it in a box as follows:

本书会将代码封装在如下的框中来显示：

Page28图

We have made use of a diple (>) to denote the command line terminal prompt shown in the Python shell. Please note that the same commands can be used in the iPython interactive shell or iPython/Jupyter notebook, although the look and feel may be quite different. As you may have already noticed, the book uses margin notes, such as the one that appears to the right of this paragraph, to highlight certain areas or commands, as well as to provide some useful comments.

我们使用了一个尖括号（>）来表示Python shell中显示的命令行终端提示符。 请注意，iPython交互式shell或iPython / Jupyter笔记本中可以使用相同的命令，但外观可能会有很大不同。 正如您可能已经注意到的那样，本书使用了边幅注释，例如本段右侧的注释那样，用来突出显示某些区域或命令，以及提供一些有用的解释。

（This is an example of the margin notes used throughout this book.）

（这是本书中使用的边幅注释说明的一个例子。）

The book is organised in a way that individual chapters are sufficiently independent from each other so that the reader is comfortable using the contents as a reference rather than a textbook. Inevitably, there will be occasions where certain topics make reference to other parts of the book and I will point out when that may be the case. I would also like to take this opportunity to mention that the implementations presented are by no means the only or best way to do things. Programming is pretty similar to the creative process of writing: The fact that you have a set of words does not imply that we all write reports in a poetic manner. I would be delighted to hear from you all about the implementations and changes you make to the code presented here. Do get in touch!

本书的组织方式使各个章节彼此之间相互独立，以便读者可以轻松地将章节内容用作参考，而不像教科书那样上下文联系紧密。 不可避免地，某些章节会引用本书的其他章节，出现这种情况我会适时指出。 我还想借此机会声明，我所提供的代码实现绝不是唯一或最好的编程方式。 编程非常类似于写作的创作过程：我们都知道26个英文字母如何拼写，并不意味着我们都能写出莎士比亚诗。 我很高兴收到大家关于我代码实现的反馈意见。 欢迎联系！

（Programming is a creative process, and as such there is more than one way to do things.）

（编程是一个创造性的过程，因此条条大路通罗马。）

We start in Chapter 1 with a discussion of what data science and analytics are, from the point of view of the process and results obtained. We pay particular attention to the data exploration process as well as the data munging that needs to be carried out prior to the application of algorithms and analysis.

我们从第1章开始，从使用过程和最终期望获得的结果角度讨论数据科学和分析是什么。我们特别关注数据探索过程以及在应用算法和分析之前需要执行的数据清洗。

（The data science workflow is discussed on Chapter 1.）

（数据科学工作流程将在第1章讨论。）

In Chapter 2 we take the opportunity to remind us of some important features of the Python language. The aim is to revisit some important commands and instructions that provide the base for the rest of the book. This will also give us the opportunity to revise some commands and instructions used in later chapters.

在第2章中，我们借此机会复习一下Python语言的一些重要特性。 目的是重新审视一些重要的命令和说明，为本书的其余部分提供基础。 这也将使我们有机会了解一下后面章节中将会使用的一些命令和指令。

（A Python primer is given in Chapter 2.）

（第2章给出了Python入门）

In Chapter 3 we cover basic elements of machine learning, pattern recognition and artificial intelligence that underpin the algorithms and implementations we will use in the rest of the book.

在第3章中，我们将介绍机器学习，模式识别和人工智能的基本要素，这些要素是我们将在本书其余部分使用的算法和实现的基础。

（Chapter 3 covers the basics of machine learning, pattern recognition and artificial intelligence.）

（第3章介绍了机器学习，模式识别和人工智能的基础知识。）

By the time Chapter 4 is reached we will have the necessary foundations to implement regression analysis using Python via both StatsModels and Scikit-learn. The main points in the usage of generalised linear models for regression are covered in this chapter.

到达第4章时，我们已拥有必要的基础，通过StatsModels和Scikit-learn使用Python实现回归分析。 本章介绍了使用广义回归线性模型的要点。

（Chapter 4 covers various regression algorithms）

（第4章介绍各种回归算法）

In Chapter 5 we talk about clustering techniques, whereas Chapter 6 covers classification algorithms. These two chapters are central to the data science workflow: Clustering enables us to assign labels to our data in an unsupervised manner; in turn we can use these labels as targets in a classification algorithm.

在第5章中我们讨论了聚类技术，而第6章则讨论了分类算法。 这两章是数据科学工作流程的核心：聚类使我们能够以无人监督的方式为我们的数据分配标签; 反过来，我们可以将这些标签用作分类算法中的目标。

（Chapters 5 and 6 cover clustering and classification techniques, respectively.）

（第5章和第6章分别介绍了聚类和分类技术。）

In Chapter 7 we introduce the use of hierarchical clustering, decision trees and talk about ensemble techniques such as bagging and boosting. It is worth pointing out that ensemble techniques have become a common tool among data scientists and you are highly recommended to check this section out.

在第7章中，我们介绍了层次聚类，决策树的使用以及关于Bagging和Boosting等集成学习法技术的讨论。 值得注意的是，集成技术已成为数据科学家的常用工具，强烈建议您查看本节。

（Chapter 7 deals with hierarchical clustering decision trees and ensemble techniques）

（第7章讨论层次聚类决策树和集成技术）

Dimensionality reduction techniques are discussed in Chapter 8. There we will cover algorithms such as principal component analysis and singular value decomposition. As an application we will talk about recommendation systems.

降维技术将在第8章中讨论。我们将介绍主成分分析和奇异值分解等算法。 作为一个应用程序，我们将讨论推荐系统。

（Chapter 8 talks about dimensionality reduction.）

（第8章讨论降维。）

Last but not least, in Chapter 9 we will cover the support vector machine algorithm and the all important Kernel trick in applications such as regression and classification.

（Chapter 9 deals with support vector machines.）

最后也很关键一章，在第9章中，我们将介绍支持向量机算法以及回归和分类等应用程序中所有重要的核心技巧。

（第9章涉及支持向量机。）

The book was made possible, as I mentioned before, thanks to discussions, presentations and exchanges with colleagues both in academia as well as in business. I am very grateful for their input and suggestions. I would also like to thank my editor at CRC Press, Randi Cohen, as well as the technical reviewers for their comments and suggestions. Finally, the encouragement that my family and friends have given me to take up yet another writing project has been invaluable. This goes to you all!

正如我之前提到的那样，这本书的诞生得益于我与学术界和商界同事的讨论，演讲和交流。 我非常感谢他们的意见和建议。 我还要特别感谢CRC Press，Randi Cohen编辑以及技术评审员的意见和建议。 最后，进行这个写作项目也离不开家人和朋友的鼓励。 这本书献给你们所有人！

*Reader’s Guide*

*读者概览*

This book is intended to be a companion to any jackalope data scientist from beginners to seasoned practitioners. The material covered here has been developed in the course of my interactions with colleagues and students and is presented in a systematic way that builds upon previous material presented.

（Read Chapter 1 to understant the Jackalope reference.）

本书旨在成为从初学者到经验丰富的从业者以及任何数据科学家(jackalope)的伴侣。 这里介绍的材料是在我与同事和学生互动的过程中开发的，并以系统的方式呈现，并以之前提供的材料为基础。

（阅读第1章以了解鹿角兔Jackalope的参考文献。）

I highly recommend reading the book in a linear manner. However, I realise that different readers may have different needs, therefore here is a guide that may help in reading and/or consulting this book:

我强烈建议按照章节顺序阅读本书。 但是，我意识到不同的读者可能有不同的需求，因此下列指南可能有助于阅读或咨询本书：

• Managers and readers curious about Data Science:

– Start by reading Chapter 1 where you will learn what Data Science is all about

– Follow that by reading Chapter 3 where an introduction to machine learning awaits you

– Make sure you understand those two chapters inside out; they will help you to understand your jackalope data scientists.

•对数据科学感兴趣的经理和读者：

- 首先阅读第1章，了解数据科学的全部内容

- 通过阅读第3章，了解机器学习的介绍等待着你

- 确保从内到外理解这两章; 他们将帮助你了解你的数据科学家jackalope。

• Beginners:

– If you do not have a background in programming, start with Chapter 2, where a swift introduction to Python is presented

– Follow that by reading Chapter 1 and Chapter 3 to understand more about what Data Science is and the principles of machine learning.

•初学者：

- 如果你没有编程背景，请从第2章开始，其中介绍了Python的快速入门

- 通过阅读第1章和第3章来了解更多有关数据科学和机器学习原理的信息。

• Readers familiar with Python:

– You can safely skip Chapter 2 and go directly to Chapter 4

•熟悉Python的读者：

- 您可以大胆跳过第2章，直接进入第4章

• Seasoned readers may find it easier to navigate the book by themes or subjects

– Regression is covered in Chapter 4, including:

\* Ordinary least squares

\* Multivariate regression

\* LASSO and Ridge regression

\* Support vector machines for regression are covered in Section 9.1.3

– Clustering:

\* K-means is covered in Chapter 5

\* Hierarchical Clustering is covered in Section 7.1

•经验丰富的读者可能会发现按主题或主题导航书更容易

- 第4章介绍了回归，包括：

\* 普通最小二乘法

\*多元回归

\* LASSO和Ridge回归

\*第9.1.3节介绍了用于回归的支持向量机

- 聚类：

\* K-means在第5章中介绍

\*分层聚类在7.1节中介绍

– Classification is generally covered in Chapter 6 including:

\* KNN

\* Logistic regression

\* Naïve Bayes

\* Support vector machines for classification are covered in Section 9.1.4

– Decision Trees and Ensemble Techniques are discussed in Chapter 7

– Recommendation Systems are introduced in Section 8.4

– Text manipulation examples are provided in Section 6.4.2 where tweets are used as the main data source.

– Image manipulation examples are provided in Sections 8.2.1 and 8.3.1

- 第6章通常涉及分类，包括：

\* KNN

\*逻辑回归

\*朴素贝叶斯

\*第9.1.4节介绍了用于分类的支持向量机

- 决策树和集成技术将在第7章中讨论

- 推荐系统在第8.4节中介绍

- 文本操作示例在第6.4.2节中提供，其中推文用作主要数据源。

- 图8.2.1和8.3.1中提供了图像处理示例

*About the Author*

*关于作者*

Dr Jesús Rogel-Salazar is a Lead Data Scientist with experience in the field working for companies such as AKQA, IBM Data Science Studio, Dow Jones and others. He is a visiting researcher at the Department of Physics at Imperial College London, UK and a a member of the School of Physics, Astronomy and Mathematics at the University of Hertfordshire, UK. He obtained his doctorate in Physics at Imperial College London for work on quantum atom optics and ultra-cold matter.

Dr Jesús Rogel-Salazar 博士作为首席数据科学家，在AKQA、IBM Data Science Studio、Dow Jones等大型公司有丰富的工作经验。 他是英国伦敦帝国理工学院物理系的访问研究员，也是英国赫特福德大学物理，天文学和数学学院的成员。 他在伦敦帝国理工学院获得物理学博士学位，从事量子原子光学和超冷物质方面的工作。

He has held a position as senior lecturer in mathematics as well as a consultant and data scientist in the financial industry since 2006. He is the author of the book Essential Matlab and Octave, also published with CRC Press. His interests include mathematical modelling, data science and optimisation in a wide range of applications including optics, quantum mechanics, data journalism and finance.

自2006年以来，他一直担任数学高级讲师以及金融行业的顾问和数据科学家。他是《Essential Matlab和Octave》一书的作者，同时也出版了《CRC Press》。 他的兴趣包括数学建模，数据科学和广泛应用的优化，包括光学，量子力学，数据新闻和金融等方面

Data Science and Analytics with Python

用Python进行数据科学与分析

1. Trials and Tribulations of a Data Scientist

数据科学家的试验与磨难

The ever increasing availability of data requires the use of tools that enable businesses and researchers to draw conclusions and make decisions based on the evidence provided by the data itself. From performing a regression analysis to determining the relationship between data features, or improving on recommendation systems used in e-commerce, data science and analytics are used every day by all of us. This book is intended to provide those interested in data science and analytics a perspective into the subject matter using Python, a popular programming language available for various platforms and widely used both in business and academia.

工具的使用使得数据能够不断发挥越来越大的价值，帮助企业和研究人员基于数据本身提供的证据来做出结论和决定。从进行回归分析到确定数据特征之间的关系，或用于改进电子商务所用的推荐系统，我们每个人每天都会用到数据科学和分析。本书旨在为那些对数据科学和分析感兴趣的人提供一个使用Python的视角来了解主题，Python是一种流行的编程语言，可用于各种平台，并广泛应用于商业和学术界。

In this chapter we will cover what data science is and how it related to various disciplines from mathematics to business intelligence and from programming to design. We will discuss the characteristic that make a good data scientist and the composition of a data science team. We will also provide an overview of the typeical workflow in a data science and analytics project and shall see the trials and tribulations in the work cycle of a data scientist.

在本章中，我们将介绍什么是数据科学，以及它与从数学到智能商业，从编程到设计的各个学科之间的关系。我们将讨论如何成为一个好的数据科学家和构建一个数据科学团队。我们还将概述数据科学和分析项目中的典型工作流程，并将以此感受到数据科学家在工作中的所遇到的各种挑战和磨难。

## 1.1 Data? Science? Data Science!

## 1.1 数据？科学？数据科学！

The use of Data as evidence in support for decision making is nothing new. You only have to take a look at the original meaning of the word statistics as the analysis and interpretation of information relating to states such as economic and demographic data. Nowadays, the word statistics is either understood as a branch of mathematics that deals with the collection, analysis, interpretation and presentation of data; or more colloquially as a fact or figure obtained from a study based on large quantities of data. Simply take a look at the news on any given day and you will surely get to hear about statistics, proportions and percentages, all in support (or not) of a new initiative, plan or recommendation. The power of data is all around us and we use it all the time.

使用数据作为支持决策的证据并非什么新鲜事。“统计学”的原始本意，就是分析和解释与国家有关的信息，如经济和人口数据。现在，“统计学”被理解为数学的一个分支，它涉及数据的收集、分析、解释和呈现；或者更通俗地说是基于大数据得研究从而获得某些事实或直观的数据呈现。只要看一看任意某天的新闻，你一定会听到有关统计数字、比例和百分比的消息，关于支持（或不支持）某项新举措，计划或者提案。由此可见，数据的力量无处不在，我们一直在使用它。

Now, what about the word science? Well, you may remember from your school days that science is system that enables the organization of knowledge, based on testable evidence and predictions. Notice that key word evidence mentioned there again.

那么，什么是“科学”？好吧，你可能还记得在你上学的时候老师教的定义：科学一个[建立](https://baike.so.com/doc/6247046-6460455.html)在可检验的证据和(对[客观事物](https://baike.so.com/doc/1292750-1366818.html)进行)预测的知识系统。注意这里又提到了关键词-“证据”。

No surprises here so far, right? From a very simplified point of view, the scientific method makes use of data and their analysis to acquire, correct and integrate knowledge. Nonetheless, data science is not just simply the direct use data science of statistics, or the systematization of data. How shall we understand that much loved combination of the words data and science?

到目前为止，并没有什么让人吃惊的，对吧？从简单的角度来看，科学的方法是利用数据及其分析来获取、修正和整合知识。然而，数据科学不仅仅是简单的统计数据或直接分类数据那么简单。那么我们应该如何理解“数据+科学”？

* + 1. So, what is Data Science?

### 1.1.1 因此，到底什么才是数据科学？

Data science and analytics are rapidly gaining prominence as some of the more sought after disciplines in academic and professional circles. In a nutshell, data science can be understood as the extraction of knowledge and insight form various sources of data, and the skills required to achieve this range from programming to design, and from mathematics to storytelling.

数据科学和分析在学术和专业领域中正在迅速地获得突出地位。简而言之，数据科学可以理解为从各种数据来源中获取知识和相关洞察力，实现这一目标需要各种技能：从编程到设计，从传统数学到如何生动的讲述该故事(数据)。

There is no doubt that the term data science is a true neologism of our time. The term has started being used and, to a certain extent, even abused. As we have mentioned before data science is rather more than the sum of data on the one hand and science on the other one, although it is inevitably related to both concepts.

毫无疑问，数据科学一词是我们这个时代特有的新词。然而，这个词已经开始被使用，甚至在某种程度上被滥用。正如我们前面提到的，数据科学比数据和科学的涵盖面要广，尽管它不可避免地涉及到这两个概念。

Currently, data science can be considered a budding field with applications in a wide range of areas and industries, as well as in academic research. It is fair to say that it is elusive to define this emerging field, and throughout this book we shall consider data science and analytics as a portmanteau for a number of overlapping tasks related to data-from collection, provision and preparation, analysis and visualization, curation and storage-that exploit tools from empirical sciences, mathematics, business intelligence, machine learning and artificial intelligence. The aim of these tasks is to enable effective, pragmatic and most importantly actionable decisions.

目前，数据科学可以被认为是一个萌芽领域，广泛应用在各个行业，以及学术研究领域。可以公平地说，这个新兴领域的定义是难以捉摸的，在本书中，我们将把数据科学和分析看作是与数据相关的一系列任务的叠加：如数据收集，提供和准备、分析和可视化、管理以及储存，然后利用经验科学、数学、商业智能、机器学习和人工智能来开发相关工具。最终通过这些任务的形成有效、务实以及可采取具体行动的决定。

The motivation for data science and analytics in deriving valuable insights from data is great, and widely welcomed by businesses. However, this is a very challenging task. Companies such as Google, Netflix and Amazon have demonstrated that careful storage and analysis of data delivers a very competitive edge. These days there are easier and cheaper ways to collect large amounts of data than ever before, and mobile is becoming a ubiquitous presence. This have allowed companies, particularly start-ups, to develop in-house data science capabilities.

对于企业来说，如何应用数据科学和分析在数据中获得有价值的洞察力的商机巨大，并且受到广泛欢迎。然而，这是一项非常具有挑战性的任务。谷歌(Google)、Netflix和亚马逊(Amazon)等公司已经证明，对用户数据进行存储和分析将提升品牌的竞争力。如今，收集大量数据的方法比以往任何时候都更容易、更便宜，并且移动设备正在成为一种无处不在的存在。这使得公司，特别是初创企业，具备了能够在家就开发数据科学的能力。

Typical examples of data science products are better explained by the questions they aim to answer; these questions are the drivers to the acquisition and selection of the appropriate data to be interrogated in order to provide insight into an area of interest. I am sure you can come up with a few of examples relevant to you, but there are some that come to mind:

* What product will sell better in conjunction with another popular product?
* Who will be declared Prime Minister (or President, or winner; depending on the flavor of the government system of interest) in the next general election?
* How can customers be encouraged to spend a longer time in an online portal?
* Are there any discernible patterns that allow us to characterize different groups of sales agents, customers or businesses?

|  |  |
| --- | --- |
| * What advertisement should be placed on what site? | Advertising and marketing |
| • Given the interests of a customer, what other products can be recommended to them? | Recommendation systems |
| • What are the latest developments and breakout reports in newspapers and social media that may affect the industry of interest? | Social media analysis |
| • Given someone’s interests and hobbies, who may be suitable potential partners? | Online services |
| • How can we keep potentially sensitive information protected and react proactively to information we store? | Cybersecurity |
| • How can we distinguish valid, relevant documents such as emails (ham), from invalid, irrelevant ones (spam)? | Classification analysis |
| • How to determine if a retail transaction is valid or not? | Fraud prevention |
| • What is the demand for a particular service at a particular time or place? | Demand forecasting |

典型的数据科学产品的例子将更好地解释他们(企业)所想要回答的问题; 这些问题是获取和选择要询问的适当数据的驱动因子，以便对该领域提供有价值的洞察。我相信你也可以拿出一些相关例子，下面是我所想到的：

* 哪种产品与另一种受欢迎的产品结合起来会卖得更好？
* 下一届谁将当选为总理(或称为总统以及胜利者，基于政府利益体系的价值最大化)？
* 如何吸引客户在线门户网站上花更长时间？
* 是否有任何可识别通用模型, 使我们能够描述不同的销售代理、客户或企业组？
* 如何根据广告类型精准投放到对应的网站上？
* 如何根据客户的兴趣爱好，推荐其相关产品？
* 在报纸和社交媒体上有哪些最新报道和大新闻可能会对相应的行业产生相关影响？
* 考虑到某人的兴趣和爱好，谁可能是其合适的潜在伴侣？
* 我们如何保护自己潜在的敏感信息，并对所存储的信息作出积极的反应？
* 我们如何区分有效的文件？(例如在电子邮箱中从无效的、无关紧要的垃圾邮件甄选出有价值的邮件)
* 如何确定零售交易记录是否有效？
* 在特定的时间或地点, 对某一特定服务的需求是什么？

These are not questions that decision-makers, businesses and industries, large and small, have recently started formulating. So, why the resurgence in seeking answers to them? The main answer is the availability of potentially useful data, big or small, together with the impact of technology, computer science and statistics in everyday life.

这些都是决策者、企业和行业(无论大小)长久以来所思考的问题。那么，我们为什么要重新寻找答案呢？主要的答案在与我们想知道相关数据(无论大小) 进行计算机科学和技术统计，如何对日常生活产生影响。

Out of the ingredients mentioned above, accessible datasets may be the most important one since without them the insight provided by technology alone is rather limited. After all, the plural of anecdote is not data. Having said that, it is important to note that this does not mean that every single data science case to be tackled falls into the category of so-called big data, particularly when we take into account that the adjective big can be used in a relative manner. We shall expand on this point later on in Section 1.3.1.

在上述要素之外，可访问的数据集可能是最重要的数据集，因为没有这些数据集，技术本身所提供的洞察力就相当有限。毕竟，逸闻趣事道听途说并非真实数据。尽管如此，需要重点注意的是，这并不意味着每一个(要处理的)数据科学案例都属于所谓的大数据范畴，特别是当我们考虑到形容词Big可以是个相对的概念时。我们稍后将在第1.3.1节中详述这一点。

One important thing to bear in mind about the outputs of data science and analytics is that in the vast majority of cases they do not uncover hidden patterns or relationships as if by magic, and in the case of predictive analytics they do not tell us exactly what will happen in the future.

关于数据科学和分析的结论，需要记住的一件重要的事情是，在绝大多数情况下，它们不会像魔术一样揭示数据里面隐藏的模式或关系，而在预测分析的情况下，它们不会确切地告诉我们未来会将发生什么。

Instead, they enable us to forecast what may come. In other words, once we have carried out some modelling there is still a lot of work to do to make sense out of the results obtained, taking into account the constraints and assumptions in the model, as well as considering what an acceptable level of reliability is in each scenario

相反，它们使我们能够预测将来也许会发生什么。换句话说，一旦我们进行了一些建模，考虑到模型中的约束和假设，以及考虑每个场景中可接受的可靠性水平，仍然有许多工作要做，以便理解所得到的结果。

Similarly, there is the tacit prerequisite of having accurate, timely data that can be readily utilized to make sense out of the modelling results, and reflect the state-of-the-art in an application. It is therefore imperative that decision makers as well as IT and business stakeholders take time to understand the information that will be needed, as well as being prepared to realize that certain data may not be fit for their purpose. It is indeed disheartening to come to terms with the fact that some data may not have the necessary features to be used in building a prediction, for example. Nonetheless, it is better to realize that is the case at an early stage, rather than relying on unsuitable results to make important decisions that impact the business.

同样，拥有准确、及时的数据也是一个必要的前提，这些数据可以很容易地被用来解释建模结果，并实时反映出应用程序的状态。因此，决策者以及IT和业务利益攸关方必须花时间了解所需要的信息，并做好准备:某些数据可能不适合。事实上，令人沮丧的是，有些数据可能缺乏必要的特性来建立预测模型。然而，最好在建模之前就认识到这一点，而不是依靠不适当的结果输出来做出影响企业的重要决定。

P43:

Even if data science may not yet be considered a well-defined subject, the number of academic and training programmers being offered by universities and at various workplaces has seen a healthy increase. This is a natural result of the need that exists for well-informed, capable experts that we get to call data scientists. So what do data scientists do and what do they look like? It will all shall be uncovered.

虽然数据科学可能还没有被认为是一个明确定义的学科，但大学相关的学术研究和很多企业提供的培训程序员的数量却已经出现了健康的增长。这是一个自然而然的结果，因为我们需要信息灵通的、有能力的专家，我们可以把这些专家称为数据科学家。那么，数据科学家究竟是做什么的，他们长什么样？一切谜底都将在本书中被揭开。

## 1.2 The Data Scientist: A Modern Jackalope

## 1.2 数据科学家：现代鹿角兔

The new term used to describe the person that deals with the seemingly disparate array of tasks described above may seem to be yet another, more fashionable way to describe a statistician or a business analyst. However, we can certainly agree that there is a gap between the latter two, and that the skills required by a data scientist involve aspects that include both statistics and a strong business acumen, but also foundations in computer science, mathematics, modelling and programming, not to mention good communication skills. A simplified diagram of these skills and their relationship is shown in Figure 1.1.

用来描述上述一系列看似不相关任务的人的新术语（现代鹿角兔）似乎是描述统计学家或业务分析师的另一种更时尚的方式。然而，我们当然必须承认统计学家或者说业务分析师和数据科学家存在差距，一名数据科学家所需的技能所涉及的方面不但包括统计和强大的商业头脑，而且还需要计算机科学、数学、建模和编程方面的基础，更不用说良好的沟通技巧。图1.1显示了这些技能及其关系的简化图。

In that sense a data scientist role goes beyond the collection and reporting on data; it must involve looking at a business application or process from multiple vantage points and determining what the main questions and follow-ups are, as well as recommending the most appropriate ways to employ the data at hand.

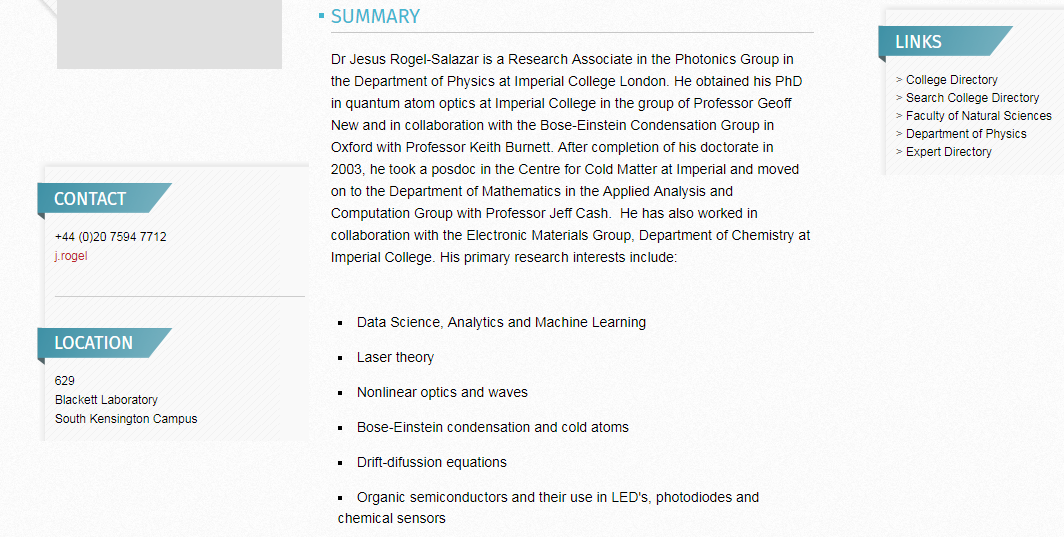
从这个意义上说，数据科学家的角色不仅仅是收集和报告数据；它必须包括从多个有利的角度看待业务应用程序或流程，确定主要问题和后续行动，以及如何使用手头数据的最佳实践。

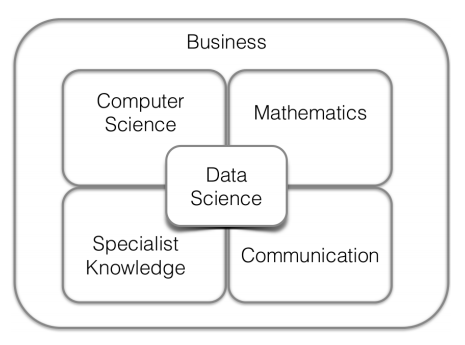
P44

8 j. rogel-salazar

<https://www.researchgate.net/profile/J_Rogel-Salazar2> ？？

<http://www.imperial.ac.uk/people/j.rogel> ？？





In terms of characteristics, a data scientist has an inquisitive mind and is prepared to explore and ask questions, examine assumptions and analyze processes, test hypotheses and try out solutions and, based on evidence, communicate informed conclusions, recommendations and caveats to stakeholders and decision makers.

就特征而言，数据科学家有一种好奇的头脑，随时准备探索和提出问题，审查假设和分析过程，测试假设和尝试解决办法，并根据结论向利益攸关方和决策者传达知情的结论、建议和警告。

In other words, a data scientist is a true new Renaissance woman or man. No wonder that despite being branded the sexiest job of the 21st century, as well as the increasing demand for these individuals, it is hard to find people with the right skills to fill in these roles. This has led to branding data scientists as Unicorns.

换句话说，数据科学家是真正的文艺复兴时期的女性或男性。难怪尽管被评为21世纪最时髦的工作，以及对这些人的总体需求不断增加，但很难找到具备适当技能的人来填补这些职位。这导致数据科学家被称为独角兽(Unicorns)。

To a certain extent, the symbolism of a Unicorn as a creature that is beautiful, mysterious and difficult to tame or even capture may be applicable to describe a data scientist. However, in my opinion, it may not be totally appropriate given the fact that, as majestic as Unicorns can be, they are way too common as far as popular culture goes.

独角兽作为一种美丽、神秘、难以驯服甚至难以捕捉的生物来说，其象征意义在某种程度上可以用来描述数据科学家。然而在我看来，独角兽更具有一种威严庄重感，而数据科学家就流行文化而言，更加普遍通俗一些，因此可能并非完全合适。

(Branding data scientists as Unicorns is a result of the quixotic expectations businesses and industries have and thus is not appropriate.)

(将数据科学家冠以“独角兽”的称号其实是商业和工业对其所抱有不切实际的期望的结果，因此是并不合适。)

The shortage that businesses experience when trying to attract data scientists is more likely due to the fact that they have created internal expectations for the role and that no single individual can fulfil, thus appealing to the magical nature of a common mythical beast. They have created their idea of the Data Scientist Unicorn, and unfortunately the fascination prevails.

企业试图吸引数据科学家的原因往往是因为他们的缺陷，在于该角色他们创造了内在的期望值，目前没有任何个体能够达到该期望，因此寄托于神话怪兽独角兽般存在的数据科学家能够做到这些。不幸的是，这样想法的企业非常普遍。

To tackle the prevailing image, I am convinced that the use of a new symbol is needed. And a silly one at that! There is an allegory I usually propose to colleagues and those that talk about the data science Unicorn. It seems to me to be a more appropriate one than the existing image: It is still another mythical creature, less common perhaps than the Unicorn, but more importantly with some faint fact about its actual existence: a Jackalope. You can see an artistic rendition of a couple of Jackalopes in Figure 1.2.

为了解决这一流行现象，我坚信需要使用一个新形象来代表数据科学家。真是愚蠢的家伙！我通常向同事说那些谈论数据科学是独角兽的人。新形象在我看来，它似乎比现存的更合适：它仍然是神话中的生物，但并不像独角兽那么常见，更重要的是，基于一些模糊的事实，他又似乎曾经真实存在过：这就是鹿角兔。你可以在图1.2中看到一对鹿角兔(Jackalopes)的艺术再现。

( The Jackalope is the one we propose.)

我们建议的形象就是鹿角兔

A Jackalope is said to be a strange beast that looks like a jackrabbit with a pair of stag horns. It is described to be a shy but clever and cunning animal, and if threatened it can be dangerous. If you are ever in the Mountain West in the United States you may stumble into Jackalope heads mounted as trophies; but of course that is not the only place where Jackalopes are endemic; there are tales of the Hasenbock in Austria1 or you may hear the Huichol stories about how the tátSu (rabbit) lost its antlers to the Colorado kaukamali (deer)2

鹿角兔(Jackalope)据说是一种奇怪的野兽，看起来像一只长着一对鹿角的野兔子。它被描述为一种羞怯但聪明又可爱的动物，当然如果受到威胁，它也可能变得很危险。如果你在美国的西部山区，你可能会跌倒在Jackalope的头上，作为奖杯；当然，这并不是Jackalopes所特有的地方；在奥地利的哈森博克也有鹿角兔的故事，甚至在维乔人那里你也可以听到兔子在科罗拉多失去了鹿角的故事

（Figure 1.2: Jackalopes are mythical animals resembling a jackrabbit with antlers.）

图1.2：鹿角兔是神话中的动物

1 Toelken, B. (2013). The Dynamics of Folklore. University Press of Colorado

1托尔肯，B（2013）。动态民俗学。科罗拉多大学出版社

2 Zingg, R., J. Fikes, P. Weigand, and C. de Weigand (2004). Huichol Mythology. University of Arizona Press

2 Zingg，R .，J . Fises，P . Weiand，and C . de Weigrd（2004）。维乔人的神话故事 亚利桑那大学出版社

(A Jackalope is a mythical being similar to a jackrabbit with a pair of stag horns)

鹿角兔是一种神话般的动物，类似于一头长着一对鹿角的野兔子

No need to explain that a Jackalope is indeed an imaginary, mythical being, much like the Unicorn, but it seems to be a better metaphor for the data scientist. We can argue that it is rather difficult at best, and impossible at worst, getting hold of a single individual that is able to be an all rounded ninja programmer, with vast expertise in mathematics, statistics and probability, plus knowledge of computer science and well-versed in business. This offers no solution to businesses interested in getting the benefit of exploiting the available data.

无需解释，Jackalope确实是也一个虚构的、神话般的存在，这点很像独角兽，但对于数据科学家来说，它似乎是一个更好的比喻。因为对于我们来说，要抓住一个能全能型程序员的人，并且他同时在数学、统计和概率方面拥有丰富的专业知识，并且精通商业造诣，是相当困难的一件事，在最坏的情况下几乎是不可能的任务。这对于那些有兴趣利用攫取大数据中商业价值的企业来说等于无法提供解决方案。

（It is indeed difficult to get hold of a Unicorn.）

抓住一个独角兽确实很难。

Well, if you cannot get them in the wild, make them up from various parts - in the best style of Dr Frankenstein and his monster - and that is where the image of the Jackalope comes handy. In 1932 Douglas Herrick did indeed put together his creation when he stuck a pair of deer horns on a dead jackrabbit and mounted it as a trophy3. The rest is history, as the Converse County city of Douglas, Wyoming became the Jackalope capital of the United States.

1932年，道格拉斯·赫里克把一对鹿角贴在一只死掉的豺兔上，并把它作为艺术品悬挂起来，这的确是他的创意。时间一长，这便成了一段历史，美国怀俄明州作为道格拉斯的一个县，因此成了鹿角兔的发源地。

（If you cannot get data scientists in the wild, make them up.）

如果你不能在野外找到数据科学家，那就造一个出来吧

（3 Martin, D. (2003, Jan 19th). Douglas Herrick, 82, Dies; Father of West’s Jackalope. The New York Times）

Furthermore, you do not have to get a fake hunter’s trophy to see a Jackalope. As I mentioned before, there is faint fact to the existence of horned rabbits. That is definitely more than one can say about a one-horned horses. This is thanks to the existence of a virus, the cottontail rabbit papilloma virus (CRPV), which makes infected rabbits grow bone-like structures in their skulls4. The virus was discovered in the 1930s by Richard E. Shope and was the first example a cancer caused by a virus.

此外，你并不需要编造一个假猎人的关于鹿角兔的故事。正如前文所述，鹿角兔的存在并非空穴来风。然而对于一匹独角兽来说，并没有人见过其真实的依据。由于存在这样一种病毒，即棉铃虫兔乳头状瘤病毒（CRPV），它使受感染的兔子在其头骨中生长骨样结构。该病毒是在20世纪30年代由Richard E. Shope发现的，也是第一例由病毒所引起的癌症。

（Plus, there is a faint fact to the existence of horned rabbits.）

（因此，鹿角兔的存在并非空穴来风）

（4 Zimmer, C. (2012). Rabbits with Horns and Other Astounding Viruses. Chicago Shorts. University of Chicago Press）

（有角的兔子和其他令人吃惊的病毒。芝加哥新闻猎奇 芝加哥大学出版社）

The use of this allegory is proposed to show how silly it is to simply employ wishful thinking in the pursuit of exploiting data and hoping that a single individual will come to the rescue. What I am trying to say is that one should think optimistically about the prospect of finding capable data scientists if we are prepared to be realistic about distinguishing mythological aspirations from messy reality.

(It is possible to find capable data scientists if we are prepared to be realistic about our expectations.)

使用这个寓言是为了表明简单分析一下数据就能盈利完全是不切实际的想法，就像依靠个体力量就能单枪匹马拯救整个团队一样的愚蠢。我想表达的是：如果我们准备真正区分神话的愿望和混乱的现实，那么我们应该持有积极乐观地态度来思考如何寻找（或是打造）一个有能力的数据科学家。

只要我们满怀期待并加以准备，就有可能找到（打造）一个有能力的数据科学家

What I propose is that the best way to tackle the data science needs of a business - a startup or a large conglomerate - is to put together a rangale of jackalope data scientists, than daydreaming of a bliss of non-existant Unicorns. After all, there are indeed better chances of seeing a Jackalope-like animal than a Unicorn, right?

如何解决企业(无论是一家初创企业抑或是一家大型企业集团)数据科学家需求的最佳方式，我的提议是是聚集一大批鹿角兔般的数据科学家，而不是白日梦般地幻想着去寻找到一个独角兽般存在的数据科学家（因为这样完美的数据科学家是永远不存在的）。毕竟，比起独角兽，看到一只鹿角兔般的动物更加现实，对吧？

(I propose therefore to put together a rangale of jackalope data scientists.)

（因此，我建议把鹿角兔般的数据科学家集合起来。）

The next question is thus related to how the rangale of data scientists should be put together, what roles they should have and what resources to provide them with. These points are perhaps not easy to answer, as they depend to a large extent on the area in a business where the insight is beingsought, and for what purpose (see Section 1.4). Nonetheless,there are ome general guidelines that can be taken into account when tackling the data scientist conundrum.

因此，下一个问题涉及如何将这一类的数据科学家集合在一起，他们应该扮演各自什么角色，以及提供相应的什么资源。这些问题也许不容易回答，因为它们在很大程度上取决于企业所在的领域的洞察力，以及相关需求（见第1.4节）。然而，在处理数据科学家难题时，还是有一些通用准则。

(Not only is it important to know what qualities a data scientist should have, but also what role they are expected to play and what tools they will use to do their jobs.)

(不仅要知道一个数据科学家应该具备什么素质，而且要知道他们应该扮演什么角色，以及他们将使用对应的什么工具来完成相应的工作。)

### 

It seems that everyone loves, or would love to have, a data scientist, and as we have seen, the wishful list of desired characteristics makes it more difficult for businesses to choose among otherwise capable candidates.

似乎每个人都喜欢或者希望拥有一位完美的数据科学家，正如我们所看到的，这种一厢情愿的期望使得企业很难在候选人中做出选择。（因为每个候选者总会在某个领域或多或少有一定的欠缺）

(Everyone would like to have their own data scientist and knowing what is important for the business needs is a major aspect to consider.)

(每个人都希望有自己领域的数据科学家，并且将其对于该领域业务需求的了解作为主要方面的考察方向)

The analogy that comes to mind is that of the everlasting dating puzzle where everyone is waiting for Princess or Prince Charming, but is unable to find “the one”. For a data scientist to be considered “the one” the skills required include those discussed in the previous section and summarised in Figure 1.1.

脑海中浮现的比喻是永恒的相亲之谜：每个人都在等待公主或白马王子，却永远也找不到“那个人”。对于一个被认为是“最佳”的数据科学家来说，所需的技能已经在上一节中讨论并总结在图1.1中。

Let us pause for a moment before we tackle the subject at hand and consider what the purpose of the data science team is or will be. This is a crucial step in building that team as these objectives will help identify the important traits that the data scientists are expected to have.

让我们先停顿片刻，再讨论手头的主题，即数据科学团队的目标究竟是什么。因为这是建立该团队的关键步骤，这些目标的确立将有助于确定大家所期望的数据科学家的重要特征。

(Having a clear idea of how a potential data scientist will fit in the organisation and what they will work on is important.)

（形成一个清楚的概念关于一个有潜力的数据科学家将如何契合企业以及他们的工作将是如何的至关重要）

Furthermore, having a clear idea of how they will fit in the organisation and what problems they are expected to solve will aid in defining the size of the team and the type of expertise needed. It is not uncommon to hear of organisations that are interested in riding the data science wave, but do not have a clear goal regarding the purpose of their data science journey.

此外，清楚地知道他们将如何契合企业，以及期望他们解决什么问题将有助于确定团队的规模大小和所需的业务知识类型。常常听闻很多企业期望驾驭数据科学浪潮，但对其数据科学之旅的目的却没有明确的目标，这并不罕见。

With the objective of the data science team in mind, it becomes much easier to decide what is relevant in a particular case. In general, what makes a good data scientist is a linear combination of some of the following traits:

根据数据科学团队的目标，在特定条件下来决定相关要点要容易得多。一般来说，一个好的数据科学家通常具有以下特征组合：

• Curiosity

• Grasp of machine learning

• Data product building and management

• Effective communication of data insights

• Programming and data visualisation abilities

• Knowledge of statistics and probability (other mathematical areas are welcome)

• Healthy skepticism, in the scientific tradition: Carry out experiments, test hypotheses, etc

* 好奇心
* 掌握机器学习
* 数据产品的建模和管理
* 有效沟通数据洞察力
* 编程和数据可视化能力
* 了解统计和概率（熟悉其他数学领域也可）
* 科学怀疑主义：大胆猜想，进行实验、检验猜想假设等

(Some important traits in a data scientist.)

（数据科学家的一些重要特征）

The important thing to realise here is that the linear combination of the features mentioned above do not necessarily have to be equally weighted, and that is the main reason for the persistance of the Unicorn fallacy we have been discussing. Should your data scientists lack some more developed branches in their antlers, all you need to do is give them a helping hand and provide them with colleagues that will help in developing those skills, but more importantly cover the gap in those desirable features. In other words, much like Mr Herrick, put together your very own Jackalope team with people who have a broad-range of generalist interests, but a deep expertise in a certain area or two.

这里需要认识到的重要一点是，上述特征的组合并非缺一不可，而这正是我们前文一直在讨论的独角兽谬误的主要原因。如果你的数据科学家在某些方面有所欠缺(就像鹿角缺少一些发达的分支那样)，你所要做的就是向他们伸出援助之手，提供同事帮助他们发展这些欠缺的技能，弥补与理想特征之间的差距。换句话说，就像赫里克先生所带队的team那样，把你数据科学家团队打造成大部分人具有广泛的通才，少数几人拥有深厚的专业知识在单个或多个领域即可

(The features mentioned do not have to be combined in equal measures.)

所提到的特征没有固定的组合模式。

(Start a data science team with a solid core, perhaps made out of more than one person.)

（一个稳定的数据科学团队，很可能由多人组成。）

The sensible thing to do is to start with a solid core and not let the list above let you get carried away. In other words, setting the foundations of the data science team is similar to having strong foundations in a building; without them the whole tower may collapse in an instant. Furthermore, use this core to your advantage and bank some of the easy wins to start with. The three pillars in this data science triumvirate I am referring to may include, with variations in the titles, the following main roles:

明智的做法是先从打造一个坚实的团队核心成员开始，而不要被上面的清单所迷惑。换言之，数据科学团队的核心成员如同建筑物中打下坚固的地基；没有它们的存在，整个塔可能在瞬间倒塌。此外，通过核心团队的建议有助于先打赢一些小战役(如签下小单，做演讲demo等)。我所提到的数据科学三大支柱包括下列主要角色：

• Data Science Project Manager

• Lead or Principal Data Scientist

• Data Architect

(The data science triumvirate.)

* 数据科学项目经理
* 首席数据科学家
* 数据架构师

（数据科学三部曲）

Having a person that is able and experienced in managing technical teams is an important role to have in the mix. The main idea is to cover the fact that many a data scientist is far more interested in tackling questions and problems head on, rather than dealing with managing a project from end to end. One way to help them deliver is to have a knowledgeable individual that is able, on the one hand, to keep track of how projects are going, attend meetings and manage relationships. On the other hand, they should have a general understanding of techniques, algorithms and technology to be able to liaise with the team effectively. The project manager does not have to be a ninja programmer, but should be able to understand what the rest of the team are working on and the challenges they may be facing.

在管理团队中有有能力和经验的项目经理是一个至关重要的角色，其存在的主要目的是涵盖这样一个事实：多数数据科学家更感兴趣的是如何直接处理一个接一个的技术问题，而不是从始至终的管理项目。因此帮助他们实现目标的一种方式是需要一个知识全面的人，一方面，能够跟踪项目进展情况，参加会议和管理关系。另一方面，也对算法和各类技术栈有着全面的了解，以便能够有效地与团队进行联系。项目经理不必是忍者型程序员(即能够完全解决问题的程序员)，但是应该能够理解团队的其他成员正在做什么，以及他们可能面临的挑战。

(First, a Data Science Project Manager is needed.)

（首先，需要一个数据科学项目经理。）

The second figure in the triumvirate is that of the principal data scientist. Not only is it necessary to have a good project manager, but also have someone with a strong background in a quantitative field: Physics, mathematics, computer science, etc. Ideally the academic credentials this person would speak for themselves. In terms of programming, this person may not be a developer in the full sense of the word, however, they should have a firm background in coding and solving problems with the use of technology. An important ingredient of the role is to be able to act as an advisor or guide to other data scientists and analysts in the team.

核心三角阵型的第二个关键是首席数据科学家。团队中不但要有一个好的项目经理，而且要有一个在专业领域有很强技术背景的人：物理、数学、计算机科学等。理想情况下，这个人的学术资历就能证明一切。在编程方面，这个人可能不是一个完全意义上的开发人员，但是，他们应该具有使用技术编码和解决问题的坚实能力。这个角色的重要作用在于能够扮演为团队中其他数据科学家和分析师的顾问或向导的角色。

(Followed by a Lead Data Scientist.)

（其次是首席数据科学家。）

The third pillar in the team is the data architect, who will provide expertise in terms of data structures, databases, software engineering and computational capability. It is important for the data architect to be able to disentangle the data resources that the business may (or may not) have, and be able to use their expertise to assess what data is available, when it is available, and manage the constraints that the business, regulation and security impose on the workflow. Ideally, the data architect would be interested in quantitative topics, but most importantly their programming skills must be spot on. Note that the data architect will use the same technology that the data scientists employ in their day-to day activities.

核心团队的第三边是数据架构师，该角色将在数据结构、数据库、软件工程和计算能力方面提供专业知识。对于数据架构师来说，重要的是能够分离业务可能拥有（或可能没有）的数据资源，并且能够使用他们的专业知识来评估哪些数据是可用的，何时是可用的，以及制定合理的工作流程来管理业务、法规和保证项目开发安全。理想情况下，数据架构师会对某些特定主题感兴趣，但是最重要的一点是，他们的编程技能必须是非常精确。需要注意的是，数据架构师和数据科学家在日常工作中使用的技术应该完全相同。

(And finally a Data Architect completes the trio.)

（最后，数据架构师完成核心的第三边。）

Finally, there are four aspects that are important to remember when considering putting together that data science team. First, consider who the main stakeholders of the data science team are, and clarify the lines of reporting. Remember that everyone wants their own data scientist, and confusing or conflicting messages can lead to undesired results.

最后总结一下，如何建立一支数据科学团队有四个重要的方面需要要记住：

首先，考虑数据科学团队的主要利益相关者是谁，并澄清报告的内容。记住，每个利益相关者都想要自己的数据科学家，而混淆或冲突的消息会导致不期望的结果(因此要分清主次)

(Clear reporting lines are also important.)

(明确报告范围也很重要。)

Second, for data scientists to be able to work independently and (more importantly) productively it is important for them to be able to navigate the stack entirely. This enables extracting relevant data with appropriate tools (see Section 1.3). A data science team without strong IT skills or engineering support will have a hard time doing the job they do best.

第二，对于数据科学家来说，能够独立完成工作，如高效地完成数据模型搭建是非常重要的。包括能够用适当的工具提取相关数据（见第1.3节）等。如果没有强大的专业IT工具支持，数据科学团队将很难做到尽善尽美。

（Having appropriate tools to work with is paramount.）

(合适的工具对工作来说至关重要)

Third, once data has been identified for tackling a problem, proper interpretation is not necessarily easy, and misrepresentation of the results can be very damaging. It is not uncommon to see the use of tools such as machine learning algorithms to be seen as a black box; in practice, knowing the capabilities, limitations and trade-offs requires experience.

第三，一旦数据被确定用于处理某个问题，正确的解释（拟合）不一定是容易的事情，错误的结果可能是非常有害的。一些机器学习算法类工具被视为黑箱并不少见，往往只有实践才能出真知.

（It is necessary to have appropriate expertise to interpret and rework results.）

(必须依靠对应专业知识来解释和修正结果。)

Fourth, have the product always in mind: Not only is it important to have the right IT and statistics/machine learning skills, but also the team has to have a clear idea of the final product of their efforts, as well as their target audience. You may be able to come up with the most amazing models and results, but they may not be of much use if the product is of no interest to stakeholders or if the data scientist fails to communicate the results to them.

第四，始终牢记产品准则：拥有正确的IT和统计/机器学习技能固然很重要，但团队必须清楚地了解他们所努力打造的最终产品，以及目标受众是谁。您可能能够提出最令人惊叹的模型和结果，但是如果涉众对产品不感兴趣(未能抓住痛点)，或者数据科学家未能将结果完美展示出来，那么该产品则意味着失败。

（Also have a clear idea of the final product and communicate results clearly.）

（对最终产品有着清晰的认识并能够将其清楚地传达出去）

Consider as well the tools used to present results; in other words, there may be technology out there that lets the data scientist dazzle his/her target audience, but if that audience is not able to even access the technology, then you have lost the battle before starting.

还要考虑用于呈现结果的工具；换句话说，假如数据科学家的产品所采用的技术能够令他/她的目标受众眼花缭乱的，但是如果大家无法从结果展示中看出该技术的价值，那么产品还未展示就意味着已经失败了

（Appropriate technology for presenting and delivering results is also important.）

（适当的呈现和传递产品的技术工具也是很重要的。）

A point in case in my experience is the use of great JavaScript libraries such as D3. I am an advocate for their use as they can be effective and even great fun to use. However, they only work on “modern” browsers and unfortunately a large number of institutions out there only support old browsers unsuitable to render the created assets. This becomes a relevant point when considering the deployment of solutions (dashboards, reports, etc.).

我的经验中是使用JavaScript庞大的类库如D3(Data-Driven Documents)。我倡导使用的原因是因为它们对受众是有效的，甚至使用起来非常的有趣。然而，它们通常只在较新版的浏览器上工作，不幸的是，许多机构使用的浏览器版本比较陈旧。因此如何进行结果解决方案的呈现（如仪表板、报告等）时，就成为了一个需要探讨的话题。

译者注：D3支持的主流浏览器不包括IE8及以前的版本。D3测试了Firefox、Chrome、Safari、Opera和IE9。D3的大部分组件可以在旧的浏览器运行。

Page53

## 1.3 Data Science Tools

## 1.3 数据科学工具

With our newly acquired data science team and the individual high-calibre data scientists and analysts that compose it, we are able to keep abreast of the latest developments in the field of analytics and data science, and are able to extract actionable insights from our data.

通过我们新组建的数据科学团队以及由其组成的个人高素质数据科学家和分析师，我们能够及时了解分析和数据科学领域的最新发展，并能够从我们的数据中提取可操作的洞察力

However, not only do we need to be flexible, agile and expert, we are also required to have the right tools and infrastructure to enable the team to fulfil the objectives agreed with the team sponsors. To that end, there are a number of considerations that we would need to think about in helping the team decide on the tools needed as well as some other points such as:

但是，我们不仅需要灵活，灵活和专业，还需要拥有合适的工具和基础架构，以使团队能够实现与团队赞助商期望的目标。为此，在帮助团队进行工具选型以及其他一些要点时，我们需要考虑许多因素，例如：

(The tools chosen need to enable us to be flexible agile and expert.)

（所选择的工具需要使我们变得灵活灵活和专业）

• Regulatory and security requirements of hosting and manipulating the data

• Locations of data sources - and related subjects such as whether we would need/have immediate access to them, or would get them in batches for upload

• Responsiveness requirements for queries - e.g. Real-time v Fixed Reporting

• Volume of queries/searches to be run

• Format of the data source

• Quality of the data

* 托管和操纵数据的法规和安全要求
* 数据源的位置——以及相关的项目，例如我们是否需要/立即访问它们，还是成批地将它们上传
* 查询的响应能力要求 - 例如实时或定时报告
* 要运行的查询/搜索量
* 数据源格式
* 数据质量

(Some considerations when choosing appropriate tools.)

(选择合适工具时的一些注意事项)

The security consideration above is usually a big question for any business that requires their data to be in a particular jurisdiction and does not plan to create their own cloud service. For instance, Google to date will not guarantee that data will stay in Europe, for example.

对于要求其数据位于特定管辖范围且不打算创建自己的云服务的任何企业而言，上述安全考虑通常是一个大问题。例如，到目前为止，谷歌并不能保证全部数据都留在欧洲。

(Security of the data is very important.)

(数据的安全性非常重要。)

Data science and analytics is all about data, statistical analysis and modelling. It is therefore important to have the technology that enables those functions. A data warehouse, ETL software, statistical, modelling and data-mining tools are necessary. Similarly, an appropriate hardware and network environment are required (perhaps even in the cloud).

数据科学和分析都是关于数据、统计分析和建模。因此，拥有能够实现这些功能的技术非常重要。数据仓库，ETL软件，统计，建模和数据挖掘工具都是必须的。同样，需要适当的硬件和网络环境支撑（甚至可能在云端搭建）。

(A data warehouse, ETL software, statistical, modelling and data mining tools are necessary.)

（数据仓库，ETL软件，统计，建模和数据挖掘工具是必要的。）

The technologies used in the analytics arena have evolved at a fast pace in the last few years, and a number of open source projects, with lots of support have emerged, for instance:

分析领域中使用的技术在过去几年中发展迅速，并且出现了许多开源项目，并提供了大量支持，例如：

• Data Framework: MapReduce, BigQuery, Hadoop, Spark.

Hadoop is probably the most widely deployed (if sometimes under-utilised) framework to process data.

Hadoop is an open source implementation of the MapReduce programming model from Google. Other

technologies are aimed at processing streaming data, such as S4 and Storm. BigQuery (by Google) is a web service that enables interactive analysis of massive datasets and can be used in conjunction with MapReduce.

Enterprise versions of Hadoop are available from vendors such as HortonWorks. More recently the use of Spark has captivated the imagination of the big data connoisseurs

(Data framework technologies)

数据框架：MapReduce, BigQuery, Hadoop, Spark.

Hadoop可能是处理数据最广泛部署（但有时未充分利用）的框架。

Hadoop是谷歌的MapReduce编程模型的一种开源实现。其他的技术旨在处理流数据，例如S4和Storm。BigQuery（由Google提供）是一种Web服务，可以对大量数据集进行交互式分析，并可与MapReduce结合使用。

Hadoop的企业版可以从HortonWorks等供应商处获得。Spark框架最近也受到一些大数据使用者的青睐。

(大数据框架技术)

• Streaming data collection: Kafka, Flume, Scribe. The models may be different but the aim is similar: Collect data from many sources, aggregate it and feed it to a database, or a system like Hadoop, or other clients

流数据收集框架：Kafka, Flume, Scribe。这些模型之间可能有所不同，但是目的相似：从多个源收集数据，聚合数据并将其输出到数据库，或者像Hadoop这样的系统，或者其他客户端以供分析

(Streaming data collection technologies)

(流数据收集技术)

• Job scheduling: Azkaban and Oozie manage and coordinate complex data flows

工作调度：Azkaban和Oozie管理和协调复杂的数据流

(Job scheduling technologies)

(工作调度技术)

• Big Data Query languages: Pig and Hive are languages for querying large non-relational datastores. Big data frameworks such as MapReduce and Hadoop can be made more “user friendly” with them. Hive is very similar to SQL. Pig is a data-oriented scripting language

大数据查询语言：Pig和Hive是查询大型非关系数据库的语言。MapReduce和Hadoop等大数据框架和其配合“用户易用性较好”。Hive与SQL非常相似。Pig是一种面向数据的脚本语言

(Big data query languages)

(大数据查询语言)

• Data stores: Voldemort, Cassandra, Neo4j and HBase. These are data stores designed for good performance on very large datasets

数据存储：Voldemort，Cassandra，Neo4j和HBase，这些是专为在巨型数据集上获得良好性能而设计的数据存储

(Data stores)

（数据存储）

### 1.3.1 Open Source Tools

### 1.3.1 开源工具

The model of developing tools whose source code is made available for contribution has shifted the environment for their deployment both in small and large enterprises.The collaborative nature of the various projects provides a pool of knowledge and quality assurance that is difficult to beat. A rich and wide set of tools in the open source domain has contributed to the expansion of data science.They include tools that process large datasets as well as data visualisation, together with prototyping tools:

开源工具中贡献的源代码模型已经改变了在小型和大型企业中部署的环境。各种项目的协作性质提供了难以击败的知识壁垒和质量保证。开源领域中丰富而广泛的工具为数据科学的扩展做出了贡献。它们包括处理大型数据集以及数据可视化的工具以及原型设计工具：

(There are many open source tools that can be readily used in the data science workflow.)

（有许多开源工具可以在数据科学工作流程中使用。）

• Python: Data manipulation, prototyping, scripting, and the main focus in this book

• Apache Hadoop: Framework for processing big data

• Apache Mahout: Scalable machine-learning algorithms for Hadoop

• Spark: Cluster-computing framework for data analytics

• The R Project for Statistical Computing: Data manipulation and graphing

• Julia: High-performance technical computing

(We will be using Python in this book.)

(R is a noteworthy software package widely used by the data science community.)

• GitHub, Subversion: Software and model management tools

• Ruby, Perl, OpenRefine: Prototyping and production scripting languages

* Python：数据操作，原型设计，脚本编写，这也是本书的主要重点
* Apache Hadoop：处理大数据的框架
* Apache Mahout：用于Hadoop的可扩展机器学习算法
* Spark：用于数据分析的集群计算框架，用于统计计算
* R项目：数据处理和图形
* Julia：高性能技术计算
* GitHub，Subversion：软件和模型管理工具
* Ruby，Perl，OpenRefine：原型和生产环境脚本语言

（本书中将使用Python）

（R是数据科学界广泛使用的值得注意的软件包。）

As mentioned above, Hadoop is rapidly becoming ubiquitous for processing massive datasets. The framework is scalable for distributed data processing, but as remarked in Section 1.1.1, in my view not all data science problems require big data processing. The Hadoop “hype” has caused many organizations to deploy MapReduce-like systems that are effectively used to dump data - without a big picture of the information management strategic plan or without understanding how all the pieces of a data analytics environment fit together.

如上所述，Hadoop在处理海量数据集方面正迅速变得无处不在；该框架可扩展用于分布式数据处理，但正如1.1.1节所述，在我看来，并非所有的数据科学问题都需要大数据处理；Hadoop“热”已经导致许多企业部署了类MapReduce系统，这些系统可以有效地用于转储数据，从而无需了解信息管理战略计划的大局，也不必了解数据分析环境相关部分如何组合在一起。

(Not all data science is about big data.)

(不是所有的数据科学都关于大数据

R is seen as the programming language for statistical computing. It is not characterized by the beauty of its code, but the results are great. The number of packages that is available in the R repository (CRAN) makes it very flexible.

R被视为统计计算的编程语言。它的特点并非代码的美丽整洁，而在于输出很棒（译者注：此处理解为输出结果绘图方便并且容易与其他编程语言和数据库进行交互）。在R存储库（CRAN）中有各类包使得它非常灵活且易于扩展。（译者注：截至2018年10月16日，CRAN已经收录了各类包13181个，涵盖经济计量、财经分析、人文科学研究以及人工智能。）

The use of scripting languages such as Python provide a professional platform for application development and deployment. It is very well suited for prototyping and testing new ideas. Furthermore it supports various data storage and communication formats, such as XML and JSON, plus there is a large number of open source libraries for scientific computing and machine learning.

Python等脚本语言的使用为应用程序开发和部署提供了专业平台。他非常适合原型设计和测试新的想法，而且它支持各种数据存储和通信格式，如XML和JSON。除此之外还有大量用于科学计算和机器学习的开源库可供选择

(In recent times, Python has seen a resurgence thanks to the data science scene.)

（最近，多亏了数据科学领域的旺盛需求，Python已经复苏并有了王者归来的架势。）

Python has a number of very useful libraries such as SciPy, NumPy and Scikit-learn. SciPy extends Python into the domain of scientific programming. It supports various functions, including parallel programming tools, integration, ordinary differential equation solvers, and even extensions for including C/C++ code within Python code. Scikit-learn is a Python-based machine learning package including many algorithms for supervised learning (support for vector machines, naïve Bayes), unsupervised learning (clustering algorithms), and other algorithms for dataset manipulation. It is for these reasons that we will use Python in the rest of this book.

Python有许多非常有用的库，如SciPy、NunPy和Scikit-learn。SciPy将Python扩展到科学编程领域。它支持各种功能，包括并行编程工具、集成、常微分方程求解器，甚至在代码中支持C/C++代码的扩展。Scikit-Learn是基于Python的机器学习包，包括许多用于监督学习的算法（支持向量机、朴素贝叶斯）、无监督学习（聚类算法）和其他用于数据集操作的算法。出于以上原因，我们将在本书的剩余部分使用Python进行演示。

(Python is a well-supported language with a wide variety of modules and libraries.)

（Python是一个被各行各业广泛支持的语言，具有各种各样的模块和库。）

Page58

## 1.4 From Data to Insight: the Data Science Workflow

## 1.4 从数据到洞察力：数据科学工作流

As we have seen above, the role of a data scientists is an interesting one, and at times a challenging one. Not only do we need the right combination of skills (either in a team or an individual), but also the right tools and business questions. In this section we will address the steps that adata science project may follow. It is important to emphasise that although we may categorise and separate the various steps, the workflow is not necessarily a linear one as we shall see.

正如我们上面所看到的，数据科学家的角色是一个有趣的角色，有时是一个具有挑战性的角色。这不仅需要正确的技能组合（无论是团队还是个人），还需要正确的工具和业务问题。在这一节中，我们将讨论数据科学项目可能遵循的步骤。需要强调的是，尽管我们对各个步骤已经进行了归类和标序号，但是实际工作流并不一定是线性的，正如下图我们将看到的那样。

（The workflow of the data science expert is worth discussing too.）

（数据科学专家的工作流程也值得讨论。）

With our newly acquired data science team and the right combination of skills, we are ready to tackle out first assignment, and it is now when key measurements of success for the project should be identified. Furthermore, I tis important to realize from the start that in the vast majority of cases there will not be a unique, final answer. It is thus better to frame the problem as an iterative process where a better solution is reached on each iteration. The various steps in the data science workflow include:

具备新组建的数据科学团队和各种正确的技能组合后，我们已准备好解决首次任务，现在应该确定项目成功的关键衡量标准。此外，从一开始我们就应该意识到，在大多数情况下，不会有唯一的、最终的答案。因此，最好将问题框架化为迭代过程，在每个迭代中达到更好的解决方案。数据科学工作流中的各个步骤包括：

（The data science process is iterative.）

（数据科学过程是迭代的。）

• Question identification

• Data acquisition

• Data munging

• Model construction

• Representation

• Interaction

（The steps in the data science workflow）

问题识别

数据采集

数据挖掘

模型构建

表征

交互作用（互作）

（数据科学工作流中的步骤）

Page 59 图

Figure 1.3: The various steps involved in the data science workflow.

图1.3：数据科学工作流中涉及的各个步骤。

The fact that they have been listed in that order does not mean that they have to be followed one after the other. In some cases you may start, for example, with an internal dataset and immediately create some plots even before cleaning the data. Also, once you have started the project, you may move between steps in every iteration. Figure 1.3shows a diagram of the steps mentioned above, note that they do not necessarily follow one another in the order listed above.

（The workflow outlined above is not necessarily followed in sequence.）

上述步骤按照该顺序列出的事实并不意味着它们必须一个接一个地被遵循。在某些情况下，您可以使用内部数据集开始，甚至在清理数据之前立即创建一些图表。此外，一旦启动项目，您可以在每次迭代中的步骤之间移动。图1.3显示了上述步骤的图表，注意它们不一定按照上面列出的顺序彼此跟随。

（上述工作流程不一定按顺序执行。）

### 1.4.1 Identify the Question

### 1.4.1识别问题

There is no such a thing as insight without a clear and concise question, as well as having a way to measure the success or failure of the answer obtained after running your favourite machine learning algorithms. It is not a matter of getting a dataset and simply massaging it and creating some plots. On the contrary, let the questions point out the potential datasets that may be useful in answering them and to what extent.

（Without a clear question, there is no insight.）

没有清晰简洁的问题就不会形成洞察力这样的东西，也没有办法衡量在运行你最喜欢的机器学习算法后得到的答案是成功还是失败。这不是获取一个数据集和信息分析后形成一个点位图那么简单。相反，通过问题引导出那些潜在有用的数据集来并通过这些深层数据来形成洞察力（洞见）

（没有明确的问题，就无法形成洞察力）

Another important thing to bear in mind is the fact that although we may encapsulate a problem in a single question, in many cases it is much easier to break it down into smaller parts that can be tackled in a more straightforward manner. Furthermore, at every iteration there may be more, smaller or larger, follow-up questions that will also require answers. Remember that it is an iterative process!

（Breaking down the problem into smaller questions is useful.）

另一件需要牢记的重要事情是，尽管我们可以将问题封装在单个问题中，但在许多情况下，将其分解为可以更直接的方式处理的较小部分要容易得多。此外，在每次迭代中，可能会有更多、更小或更大的后续问题也需要答案。记住这是一个反复的过程！

（把问题分解成更小的问题是非常有用的。）

### 1.4.2 Acquire Data

### 1.4.2获取数据

Once you have a problem to tacke, the first thing that needs doing is figuring out if you or your organisation has the appropriate data that may be used to answer the question. If the answer is no, you will need to find appropriate sources for suitable data externally - web, social media, government, repositories, vendors, etc. Even in the case where the data is available internally, the data may be in locations that are hard to access due to technology, or even for regulatory and security reasons.

（Identify appropriate sources of suitable and useful data.）

一旦遇到问题，需要做的第一件事就是确定您或您的企业是否具有可用于回答问题的适当数据。如果答案是否定的您将需要在外部找到适当数据的适当来源 - 网络，社交媒体，政府，存储库，供应商等。即使数据在内部拥有，但也可能位于由于技术原因难以访问到，甚至是出于监管和安全的考虑而无法访问。

（确定合适的有用数据来源。）

### 1.4.3 Data Munging

### 1.4.3数据挖掘

If there is no insight without a question, then there is no data without data munging. Munging, or wrangling data is actually the most time-consuming task in the data science workflow. According to the New York Times’ Steve Lohr data scientists may spend anything between 50 and 80 percent of their time doing “data janitor work”5 and I can definitely attest to that.

（There is no data without data munging.）

如果说没有问题就没有洞察力，那么没有数据挖掘，数据也就不存在了。在数据科学工作流中，数据挖掘和数据整理实际上是最耗时的任务。根据《纽约时报》的史蒂夫·洛尔的调研：数据科学家们可能会把50%到80%的时间花在“看门人”工作(数据挖掘和整理)上，关于这一点我完全赞同。

（没有数据挖掘，数据也就不存在了）

Data preparation is key to the extraction of valuable insight and although some may prefer to concentrate only on the much more fun modelling part, the fact that you get to know your dataset inside out while munging it implies that any new or follow-up questions can probably be attained with less effort.

（5 Lohr, S. (2014, Aug 17th). For Big-Data Scientists, ’Janitor Work’ Is Key Hurdle to Insights. The New York Times）

数据准备是提取有价值的洞察力的关键，虽然有些人可能更愿意只关注建模部分（因为觉得那更有趣），但事实上当你经过数据挖掘后会对数据有了从内到外的深层次理解，这也意味着任何新的或后续的问题都可能事半功倍。

（5 Lohr, S.（2014，8月17日）对于大数据科学家来说，“看门人工作”(数据挖掘和整理)是洞察力的关键障碍。纽约时报

### 1.4.4 Modelling and Evaluation

### 1.4.4建模与评价

Having a clean dataset to feed to a machine learning or statistical model is a good start. Nonetheless, the question remains regarding what the most appropriate algorithm to use is. A partial answer to that question is that the best algorithm depends on the type of data you have, as well as its completeness. It also depends on the question you decided to tackle. Once the model has been run through the so-called training dataset the next thing to do is to evaluate how effective and accurate the model is against the testing dataset and decide if the model is suitable for deployment.

(Every model needs to be evaluated.)

将一个干净的数据集输入到机器学习或统计模型是一个良好的开端。尽管如此，问题仍然在于最适合使用的算法是什么。这个问题的部分答案是，最佳算法取决于您所拥有的数据类型，以及它的完整性。也取决于你决定要解决的问题。一旦模型已经建立并训练数据集测试通过，接下来要做的是评估模型与测试数据集的有效性和准确性，并决定模型是否适合部署。

（每一个模型都需要进行评估）

（有关训练和测试数据集的更多信息，请参见第3.11节）

（For more in training and testing datasets, see Section 3.11.）

### 1.4.5 Representation and Interaction

### 1.4.5表征与互动

They say that a picture is worth a thousand words, and it does stand to reason that the vast majority of us get further information from a couple of well-presented plots than by looking at rows and rows of data. Data visualization is more of an art than a science, and much has been written (or drawn) about by brilliant designers and data journalists. Simply take a look at the great images produced by David McCandless in his book Information is Beautiful6 or the visualizations produced by Manuel Lima in Visual Complexity7.

俗话说一图胜千言，我们绝大多数人都可以从一些精心设计的图中获得更多信息，而不是查看行和行的数据。数据可视化与其说是一门科学，还不如说是一门艺术，许多优秀的设计人员和数据记者都曾撰写（或绘制）过。例如大卫·麦克坎德莱斯（David McCandless）在《信息是美丽的》（Information is Beautiful6）一书中所创作的伟大图像，或者曼纽尔·利马（Manuel Lima）在《视觉复杂性》（Visual Complexity7）一书中所创作的可视化图像。

（Data visualization is more of an art than a science, but an important one nonetheless.）

（数据可视化更像是一门艺术而不是科学，但却是一门重要的艺术。）

（6 McCandless, D. (2009). Information is Beautiful. Collins）

（6 McCandless, D. (2009). 信息是美丽的. Collins）

（7 Lima, M. (2011). Visual Complexity: Mapping Patterns of Information. Princeton Architectural Press）

（7 Lima, M. (2011).视觉复杂性：映射信息的模式。普林斯顿建筑出版社)

You do not have to produce such artistic beauties, but make sure that the data representation that you decide to use is accurate, simple, and provides clarification to the story you want to communicate. In some cases, there is the possibility of telling the story in a way that the reader/viewer is able to interact with the data representation, for instance in the form of dashboards, reports or interactive plots. These may be fun to use, but the same principles of accuracy, simplicity and clarity do apply.

你不必要求自己的作品也达到艺术美的高度，但至少要确保所使用的数据准确性，简单性，并为你要阐述的观点提供说明。在某些情况下，有可能以读者/观众更容易被数据的直观展现所吸引，例如仪表板、报表或交互式绘图等形式。这些展现方式可能非常有趣，但同样需要遵循准确性，简单性和清晰适用等原则。

### 1.4.6 Data Science: an Iterative Process

### 1.4.6数据科学：一个迭代过程

The simple fact that a machine learning model has been run on clean data does not mean that the work of the data scientist is done and dusted. On the contrary, the effectiveness of the model needs to be carefully monitored as the outcome depends on the data that is fed to them. A simple case of garbage-in-garbage-out. Similarly, any new data with a variety of new features may degrade the accuracy of the initial model, and thus it becomes necessary to adjust parameters or acquire new data.

机器学习模型已经在清洁数据上运行的简单事实并不意味着数据科学家的工作已经完成并且已经尘埃落定。相反，需要仔细监控模型的有效性，因为结果取决于提供给他们的数据。 一个简单的例子就是garbage-in-garbage-out（若输入错误数据则输出亦为错误数据）。类似地，具有各种新特征的任何新数据都可能降低初始模型的准确性，因此有必要调整参数或获取新数据。

（The effectiveness of a model needs to be monitored.）

（一个模型的有效性需要被监控）

Furthermore, even if there were such a thing as a never-changing model, the fact is that new and follow-up questions arise naturally from the data science process. This means that the workflow is better to be treated as an upward spiral where constant iterations provide improvement and new insights.

此外，即使最开始存在一个不需改变的模型，但随着新的和后续的问题在数据科学分析过程中不断涌现，工作流程逐渐向上螺旋盘升，因此需要不断的迭代改进以提供新的洞察力。（迭代过程中模型需要改变）

（Think of this process as an upward spiral.）

（把这个过程想象成一个向上的螺旋。）

I would like to close this chapter with a few questions that the data scientist team and their stakeholders should always have in mind:

• What data was used and why?

• Where was the data acquired from and who owns it?

• Was the entire dataset used? Is a sample representative of the entire population?

• Were there any outliers? Have they been considered in the analysis?

• What assumptions were made when applying the model/algorithm? Are they easily relaxed/strengthened?

• What does the result of the model mean to the process/business/product?

（Some questions to always bear in mind during any data science project.）

我想用一些数据科学家团队及相关负责人应该始终牢记的问题来结束本章：

* 使用什么数据？为什么？
* 数据从何而来，谁拥有它们？
* 是否使用了整个数据集？这个集合能否代表整个样本空间吗？
* 有异常值吗？分析中考虑过它们吗？
* 在应用模型/算法时做出了什么假设条件？它们容易放松/加强吗？
* 模型的结果对流程/业务/产品意味着什么？

（在任何数据科学项目中都要记住的这些问题。）

*1.5 Summary*

*1.5总结*

In this chapter we addressed some crucial aspects that will enable us to implement and acquire that elusive mythical being called a data scientist. We have provided a working definition for the term data science and have described how it is a rapidly evolving multi-disciplinary field encompassing areas in mathematics, computer science, statistics and business expertise.

在本章中，我们讨论了一些关键方面，这些方面将使我们能够实现并获得这种难以捉摸的神话（鹿角兔）即数据科学家。我们为数据科学这个术语提供了一个工作定义，并描述了它是如何迅速发展的多学科领域，包括数学，计算机科学，统计学和商业专业领域。

We mentioned a few examples of data science products and have seen that the main motivation for data science and analytics is deriving valuable and actionable insights from data. Our discussion pointed out that in order to fulfil that motivation, a data scientist with the appropriate skills is needed. Unfortunately, understanding about this subject has created expectations that point at the data scientist role as one akin to the mythical Unicorn. We argued that a better understanding of the role indicates that we should perhaps use a different allegory: A Jackalope. It is still a mythical being, but the fact that you can put one together out of different parts (as in a team), or the hint that there may be a scientific explanation for their potential existence provides a more hopeful panorama for many businesses interested in exploiting their data.

我们提到了一些数据科学产品的例子，并且已经看到数据科学和分析的主要动机是从数据中获得有价值和实际可操作的见解（洞察力）。我们的讨论指出，为了实现这一动力，需要具备适当技能的数据科学家。 不幸的是，对这个主题的理解产生了一些过高的期望，即数据科学家的角色应该类似于神秘的独角兽。 我们认为，更好地理解这个角色应该使用不同的寓言：即鹿角兔(Jackalope）。 它仍然是一个神话般的存在，但是事实是，你可以从不同的部分（如在团队中）拼凑出一个这样的整体(译者注：就像1.2章道格拉斯•赫里克把一对鹿角贴在一只死掉的豺兔上那样)，或者暗示它们可能存在科学的解释，为许多对利用自身大数据感兴趣的企业提供了更有希望的全景展望。

We continued our discussion with the three main pillars that would sustain a productive data science team, namely a data science project manager, a lead or principal data scientist and a data architect. We saw that not only is the team composition important, but also the tooling provided to carry out their tasks in a productive manner. We concluded this chapter by outlining the steps in a simplified data science workflow and explained their importance as part of what should be seen as an iterative process.

我们继续讨论支持高效数据科学团队的三个主要支柱，即数据科学项目经理，团队负责人即首席数据科学家和数据架构师。 我们看到，不仅团队组成很重要，而且提供高效的方式执行任务的工具是也不可或缺的。 在本章的最后，我们概述了简化的数据科学工作流程中的步骤，并解释了这些步骤应该被视为迭代过程一部分的重要性。

In the next chapter we will provide a brief refresher of some important concepts of using Python. This will enable us to have a point of reference for the rest of the book where Python will be used to implement a variety of algorithms that form part of the Jackalope’s bag of tricks.

在下一章中，我们将简要介绍一些使用Python的重要概念。 这将使我们能够为本书的其余部分提供一个参考点，其中Python将用于实现各种算法，这些算法是也是构成数据科学家(Jackalope)技巧包的一部分。

Page65

# 2 Python: For Something Completely Different

# 2 Python：完全不同的东西

There is no shortage of programming languages and paradigms. With that in mind it may seem surprising that what started up as a “hobby programming project” by Guido van Rossum in 1989 has taken on a life of its own, with a very active community, used in a wide variety of applications. Its success is perhaps due to the compactness of its code, or the fact that is open source, or even the variety of toolsets. Whatever the reason, Python is a great tool to have under your jackalope data scientist belt.

（Python started up as a hobby programming project.）

编程语言和范例史上并不缺乏，尽管如此，Guido van Rossum在1989年开始的“业余爱好编程项目”确令人惊讶的慢慢拥有了自己的生命，以及一个非常活跃的社区，并适用于各种各样的场合。 它的成功可能是归功于其代码的紧凑性，或者是开源的事实，甚至是工具集的多样性。不管是什么原因，Python都是作为组成数据科学家(jackalope鹿角兔)不可或缺的工具组的一部分。

(Python起初是一个业余爱好编程项目。)

Python is not named so after the nonvenomous snakes whose family includes some of the largest in the World. Instead it is named after the famous British comedy troupe Monty Python, and it was created to be appealing to Unix/C coders. Today, Python’s emphasis on code readability and expressive syntax has made it into a general-purpose, high level, object-oriented programming language available in multiple platforms and with a plethora of supporting packages and modules.

（We have to thank Monty Python for the name of this flexible programming language.）

Python名字的由来和世界上最大的无毒蛇其实没有任何关系。 相反，它取自的英国喜剧剧团Monty Python，它的创建是为了吸引Unix / C程序员。 今天，Python强调代码可读性和表达性的语法，使其成为一种通用的，高级的，面向对象的编程语言，可在多个平台上使用，并提供大量支持包和模块。

(我们要感谢Monty Python提供这种灵活编程语言的名称。)

译者注：在 python shell 中输入：import this

可以看到如下内容：

The Zen of Python, by Tim Peters

Python 的设计哲学，作者：Tim Peters

Beautiful is better than ugly.

优雅胜于丑陋。

Explicit is better than implicit.

明了胜于晦涩

Simple is better than complex.

简单胜于复杂。

Complex is better than complicated.

复杂胜于凌乱。

Flat is better than nested.

扁平胜于嵌套。

Sparse is better than dense.

间隔胜于紧凑。

Readability counts.

可读性很重要。

Special cases aren't special enough to break the rules.

特例并没有特殊到违背规则

Although practicality beats purity.

尽管实用性胜于纯粹性

Errors should never pass silently.

也不可对错误置之不理。

Unless explicitly silenced.

除非另有明确要求。

In the face of ambiguity, refuse the temptation to guess.

面对模棱两可，拒绝猜测。

There should be one-- and preferably only one --obvious way to do it.

尽量找一种，最好只有一种方法来做一件事。

Although that way may not be obvious at first unless you're Dutch.

虽然这并不容易，除非你是 Python 之父。

Now is better than never.

做总比不做好

Although never is often better than \*right\* now.

但不假思索就动手还不如不做。

If the implementation is hard to explain, it's a bad idea.

如果方案很难解释，那就不是好方案。

If the implementation is easy to explain, it may be a good idea.

如果方案容易解释，那么它很有可能是个好方案。

Namespaces are one honking great idea -- let's do more of those!

命名空间是一个绝妙的理念，我们应当多加利用。

（“Pythonic” coding style guidelines are followed by so-called Pythonistas.）

（所谓的Pythonistas遵循“Pythonic”编码风格指南。）

Perhaps some of the most defining features of the language include the use of indentation for grouping statements, together with was has become to be known as the “pythonic” coding style, i.e. style guidelines and idioms used by veteran Python programmers, aka Pythonistas. These guiding principles can be read in any Python installation by importing the this module. These aphorisms, compiled by the long time pythonista Tim Peters, are shown in the code listing printed above.

（The use of indentation is an important feature of Python.）

上述就是python语言的一些最具定义性的特性包括使用缩进对语句进行分组，以及被称为是“pythonic”编码样式，即由经验丰富的Python程序员使用的样式指南和习惯用法，也称为Pythonistas。通过导入this模块，可以在任何Python安装中读取这些指导原则。这些谚语，由Pyth创始人Tim Peters编写，并显示在上面列出的代码列表中。

（缩进的使用是Python的一个重要特征。）

This book assumes some knowledge of general programming principles and a level of familiarity with Python. Nonetheless, in this chapter we will briefly review some of the concepts and idiosyncrasies of the language that will be used throughout the book. If you are a seasoned Python programmer, you may want to skip this chapter and move on to the next one. If, however, you are interested in a refresher, then go ahead and read the following pages, you might find something completely different. For those of you that are beginning their Python journey, this chapter may whet your appetite to learn more. There are plenty of resources to quench your thirst1,2.

（We will assume a certain familiarity with programming principles and with Python in particular.）

1 Downey, A. (2012). Think Python. O’Reilly Media

2 Langtangen, H. (2014). A Primer on Scientific Programming with Python. Texts in Computational Science and Engineering. Springer Berlin Heidelberg

本书假定读者具有基础的编程原理的一些知识以及对Python有一定的熟悉程度。 尽管如此，在本章中，我们将简要回顾一下将在本书中使用的语言的一些概念和特性。 如果你是一位经验丰富的Python程序员，可以跳过本章并继续下一章。 但是，如果你对复习感兴趣，请继续阅读以下页面，您可能会发现完全不同的东西。 对于那些正准备开始Python之旅的人来说，本章可能会激起你学习更多东西的兴趣。有足够的资源来供你参考见1,2。

（我们假定对编程原理特别是Python有一定的了解。）

1 Downey, A. (2012). Think Python. O’Reilly Media

2 Langtangen, H. (2014). A Primer on Scientific Programming with Python. Texts in Computational Science and Engineering. Springer Berlin Heidelberg

## 2.1 Why Python? Why not?!

## 2.1 为何是Python？为何不是？！

We mentioned above that Python emphasises code readability, and this in turn has an impact on productivity: Not only is a data scientist able to create scripts to be executed as a batch, but also is able to start an interactive console (iPython shell for instance) or notebook (iPython/Jupyter notebook - see Section 2.1.2). Furthermore, the Python ecosystem is supported by various packages that extend and enhance the language.

（Interactive sessions are available through the iPython shell and the Jupyter notebook.）

我们在上文提到了Python强调代码的可读性，这对生产力也是一种促进：不仅数据科学家能够创建脚本作为批处理执行，而且还能够启动交互式控制台（例如iPython shell） ）或笔记本电脑（iPython / Jupyter Notebook - 参见第2.1.2节）。 此外，Python 还有各种扩展和增强语言的包以供选择。

（通过iPython shell和Jupyter notebook可以进行交互式会话操作）

For example, the NumPy module provides functions that enable the manipulation of numeric arrays and matrices. The SciPy module enables functionality usually expected in scientific computing such as optimisation, linear algebra routines, Fourier transformation, etc.

（NumPy provides numerical functions to be used with Python.）

例如，NumPy模块提供了可以操作数值数组和矩阵的函数。 SciPy模块支持科学计算中通常需要的功能，如优化算法（Optimisation），线性代数例程，傅里叶变换等。

（NumPy模块提供了很多供Python使用的科学函数。）

The support that Python has for hierarchical modularity makes it possible for programmers and developers to build further functionality. A good example is the Pandas package, which extends the NumPy arrays into dataframes for ease of data manipulation and analysis (see Section 2.5). We will be using Pandas in the rest of the book. Similarly, in this book we will extensively use packages such as Matplotlib, Statsmodels and Scikit-learn which implement plotting, statistical models and machine learning algorithms, respectively.

（The Pandas package enables data manipulation and analysis. ）

Python对分层模块化的支持使程序员和开发人员可以构建更多功能。 一个很好的例子是Pandas包，它将NumPy数组扩展为数据帧，以便于数据操作和分析（参见第2.5节）。 我们将在本书的其余部分使用Pandas。 同样，在本书中，我们将广泛使用Matplotlib，Statsmodels和Scikit-learn等软件包，分别实现绘图，统计模型和机器学习算法。

（Pandas包可以进行数据处理和分析。）

(Other popular packages are Matplotlib, Statsmodels and Scikit-learn.）

（其他受欢迎的软件包是Matplotlib，Statsmodels和Scikit-learn。）

It is true that Python is an interpreted language and as such code may be often slower than compiled code tailored to a particular machine architecture. In that respect, although the source code in Python is interpreted on the fly, the main advantage is flexibility. This is an important point in the data science workflow as we are interested in the balance between implementation time versus execution: In many cases we are more interested in getting to grips with the raw data rather than in fine-tuning the execution time for a particular machine.

（Python is an interpreted language, in other words, code is read and executed line-by-line.）

确实，Python是一种解释型语言，因此代码可能通常比为针对机器架构先编译再执行的语言执行速度要慢一些。 但是Python中的源代码是动态解释的，主要优点是灵活性。 这是数据科学工作流程中的一个重点，因为我们对实现时间而非执行时间更加关注：在许多情况下，我们更感兴趣的是掌握原始数据而不是针对机器去调优执行时间。

（Python是一种解释型语言，换句话说，代码是逐行读取和执行的。）

That brings us to another advantage: Since Python code is not compiled, we have portability. Scripts developed in one environment can be executed in any other one with the appropriate installation. Python is fast enough for the vast majority of the computational tasks in a data science workflow: It is important to get the logic right in the first place and, if needed, make the execution faster later.

（Python code is flexible and portable.）

这给我们带来了另一个优势：由于Python代码无需编译，因此具有可移植性。 在一个环境中开发的脚本只需适当的安装（Python解释器）即可在任何其他环境中执行。 Python对于数据科学工作流中的绝大多数计算任务来说运行性能足以满足需求：重要的是首先使逻辑正确，如果需要，以后可以更快地执行。

（Python代码灵活且可移植。）

（It is fast enough for the typical data science workflow）

（它对于通常的数据科学工作流程来说性能足够满足需求）

For the purposes of this book, I assume that a suitable Python installation is available to you. Furthermore, I will also assume that the installation is for version Python 3.x rather than Python 2.x. I would like to point out that although version 2 is widely used, more and more users are adopting version 3. Please note that there may be some packages that have not been updated to suit the incompatibilities between the two versions. I hope that in the near future these inconsistencies are ironed out and the community eventually moves to version 3.x.

（We will be working with version 3 of the Python distribution.）

出于本书的目的，我假设你已经安装了合适的Python版本。 此外，我还假设安装版本是Python 3.x版而不是Python 2.x. 我想指出的是，尽管版本2被广泛使用，但越来越多的用户正在使用版本3.请注意，可能有一些软件包尚未更新以适应两个版本之间的不兼容性。 我希望在不久的将来，这些不一致之处得到解决，社区最终会转向版本3.x.

（我们将使用Python发行版的第3版）

In particular, I find that the Anaconda distribution3, built by Continuum Analytics, is robust and complete enough for our purposes. Furthermore, they have suitable distributions for Windows, MacOS and Linux. You can obtain an installation package from http://continuum.io/ and follow the steps provided by the distribution. Please note that any other Python 3.x installation with the appropriate packages is equally suitable in order to follow the discussions in the rest of the book. I will explicitly mention any modules or packages that are required so that the more seasoned user is able to use pip, easy-install, homebrew or any other installation method they are comfortable with.

（3 Continuum Analytics (2014). Anaconda 2.1.0. https://store. continuum.io/cshop/anaconda/）

特别是，我发现由Continuum Analytics构建的Anaconda distribution3非常强大且完整，足以满足我们的需求。 此外，它适用于Windows，MacOS和Linux。 您可以从http://continuum.io/获取安装包，并按照对应版本提供的步骤操作。 请注意，其他软件包的Python 3.x安装都同样适用于本书其余部分的讨论， 我将明确提及所需的模块或软件包，以便经验丰富的用户能够使用pip，easy-install，homebrew或任何其他适合他们的安装方法来完成环境的相应配置。

Page 71

### 2.1.1 To Shell or not To Shell

### 2.1.1 使用Shell或不使用Shell

We have seen above that Python is an interpreted language and as such it is possible to interact with the different commands that have to be executed during the course of a session. Similarly, it is also possible to write all the commands first and execute them in a more conventional development workflow. Both approaches have their advantages and disadvantages, and fortunately you do not have to choose one over the other as Python is flexible enough to enable you to do both.

(Python shell enables interactivity. You can also develop scripts to be executed without user intervention.)

我们在上文已经提及Python是一种解释语言，因此可以与在会话过程中已经执行的不同命令进行交互。 类似地，也可以首先编写所有命令并在更传统的开发工作流程中逐一执行它们。 这两种方法都有其优点和缺点，幸运的是，您不必纠结到底选择其中哪一种方法，因为Python足够灵活，可以让您同时拥有这两种方法。

（Python shell支持交互。您还可以开发脚本，无需用户干预即可自动化执行。）

If you are interested in interacting with the code as you are writing it then starting, for instance, an iPython shell will allow you to type a command and immediately send it to the interpreter for execution. You can then take the output and continue your analysis. This way of working lets us see the results of the commands as we are working. It is an excellent way of prototyping code to be implemented in longer projects.

如果您有兴趣在编写代码时与代码进行交互，那么可以启动iPython shell将允许您键入命令并立即将其发送到解释器以供执行。 然后，您可以获取输出并继续分析。 这种工作方式让我们可以边工作边看到命令的结果。 这是在大型项目中实现代码原型的绝佳方式。

(The interactive shell lets you assess the results of each command you send to the interpreter.)

（交互式shell允许您实时看到发送给解释器的每个命令的结果。）

Unfortunately, working in this way makes the code somewhat ephemeral unless you save the commands that you are issuing to the interpreter. If you are interested in keeping track of your development and execute a series of commands repeatedly, then you can save those commands in a text file to create a Python script which by convention has a “.py” extension. The execution of these scripts does not require a Python shell and instead can be executed directly from the command line using the following syntax:

不幸的是，以这种方式工作只会使代码暂存内存中，除非你发出命令将其保存到解释器。 如果你有兴趣跟踪开发并重复执行一系列命令，则可以将这些命令保存在文本文件中以创建Python脚本，按照惯例，该脚本具有“.py”扩展名。 这些脚本的执行不需要Python shell，而是可以使用以下语法直接从命令行执行：

#python myscript.py

(Python scripts let you save the series of commands that form a program. They are saved with the .py extension.)

（Python脚本允许您保存构成程序的一系列命令。它们以.py扩展名保存。）

where we are assuming that the script called myscript.py is saved in the local path. The use of Python scripts makes it easier to launch longer jobs that do not need input from a user to be executed.

我们假设名为myscript.py的脚本保存在本地路径中。 使用Python脚本可以更轻松地启动而不需要用户重复输入较长的作业指令。

(The command above is launched directly from the terminal; no need for the Python/iPython shell.) （上面的命令直接从终端启动;不需要从Python / iPython shell执行）

In this book we will take the approach of using an interactive shell (code is compatible with Python/iPython shells) so that we are able to use the output given by the interpreter to explain the different steps we are taking. These individual commands can then be saved in a suitable script that can be run later. We will thus present code as follows:

在本书中，我们将采用交互式shell（代码与Python / iPython shell兼容）的方法，以便我们能够及时使用解释器给出的输出来解释我们正在采取的不同步骤。 然后你可以将这些单独的命令保存在可以稍后运行的脚本中。 因此，我们将提供如下代码：

#> 42 + 24

#66

（In this book we will present code as used from the Python shell for ease of presentation.）

（在本书中，我们将提供Python shell中使用的代码，以便于演示。）

Notice that the diple, >, represents the Python shell prompt where the next line of code is expected. Furthermore, if the command sent to the interpreter expects a printed result, the shell will automatically display it. For ease of explanation, in cases where we do not expect a result, or the discussion requires us to break down the code, we will show it in a script style. In other words, no shell prompt will be shown:

请注意，尖括号 >表示Python shell提示符，其中下一行为代码行。 此外，如果发送给解释器的命令需要打印结果，shell将自动显示它。 为了便于解释，在我们不期望结果或讨论要求我们分解代码的情况下，我们将以脚本样式显示它。 换句话说，不会显示shell提示符：

#a = 42 + 24

（In cases where we do not expect a result we will not show the shell prompt.）

（如果我们不期望结果，我们将不会显示shell提示符。）

Comments in Python are indicated with the use of the hash symbol, #. The interpreter will ignore any commands that follow until the end of the line. In the example below we enter a comment after carrying out a division operation

Python中的注释使用哈希符号＃来表示。 解释器将忽略直到行尾的所有命令。 在下面的示例中，我们在执行除法运算后输入注释

（A comment in Python is entered with the hash symbol, #.）

（使用符号#表示输入Python中的注释，＃。）

#> 2/3 # Python 3

#0.6666666666666666

Please note that the behaviour in Python 2 is different, as the operation above will result in an integer divison:

请注意，Python 2、3中的行为是不同的，因为上面的操作将导致整数除法：

#> 2/3 # This integer division returns 0， Python 2

#0

Since we have passed two integers to the interpreter, Python carries out an integer division, returning only the integer part. If we want a real division we can do the following:

由于我们已经将两个整数传递给解释器，因此Python执行整数除法，仅返回整数部分。 如果我们想要真正的除法，我们可以如下做到：

> 2/3.0

> 2/float(3)

（You can avoid integer division by casting integers into floats.）

(你可以通过将整数转换为浮点数来避免整数除法。)

In Python 2 we can import the functionality from Python 3.x with the \_\_future\_\_ module as follows:

在Python 2中，我们可以使用\_\_future\_\_模块从Python 3.x导入功能，如下所示：

> from \_\_future\_\_ import division Python 2

> 2/3

0.6666666666666666

（Alternatively you can use the \_\_future\_\_ module）

(或者你可以使用\_\_future\_\_模块）

### 2.1.2 iPython/Jupyter Notebook

### 2.1.2 使用iPython/Jupyter Notebook

We have seen how the interactive shell in Python allows us to assess the results of the code we are executing. That may be good enough for a number of tasks, but sometimes it may be desirable to present the code in a way that is easier to interact with, not just for the original programmer, but also with other members of a team or general audience.

我们已经看到Python中的交互式shell如何允许我们实时评估我们正在执行的代码的结果。 这对于许多任务来说可能已经足够好了，但有时可能需要以更容易与之交互的方式呈现代码，不仅对于原创程序员，而且对于团队的其他成员或一般受众。

The iPython/Jupyter notebook is a great way to do this. Not only does it let you run your code in the same way as the interactive shell and add comments to your code, but also enables you to document your code, calculations and processing all within a web-based interface. In this book, we have made a conscious decision not to use the iPython/Jupyter notebook for printing purposes, however I would encourage you to make use of it whenever you can.

使用iPython / Jupyter Notebook是一种很好的方法。 它不仅可以让你以交互式shell相同的方式运行代码并为代码添加注释，还可以让你在基于Web的界面中记录代码，计算和处理。 虽然在本书中，我们决定不使用iPython / Jupyter笔记本进行打印，但我鼓励你尽可能地使用它。

（The iPython/Jupyter notebook lets us interact with the code and also add rich media, text and maths.）

（iPython / Jupyter Notebook让我们可以与代码进行交互，还可以添加富媒体，文本和数学。）

The iPython/Jupyter notebook supports the inclusion of text, mathematical expression and inline graphics as well as other rich media such as websites, images, video, maths, etc. At its core, a notebook is a JSON document with the extension .ipynb, which makes the files very light and highly portable. The web interface is very handy, and if required, the notebook can be exported to a number of formats such as HTML, LATEX, PDF, Markdown or even raw Python. Furthermore, the Jupyter project aims to exploit the parts of the code that are not specific to Python and extend their use to other programming languages.

iPython / Jupyter notebook支持包含文本，数学表达式和内嵌图形以及其他富媒体，如网站，图像，视频，数学等。其核心是一个扩展名为.ipynb的JSON文档， 这使得文件非常轻便且便于携带。 Web界面非常方便，如果需要，可以将notebook导出为多种格式，如HTML，LATEX，PDF，Markdown甚至原始Python。 此外，Jupyter项目并非专为Python特定开发，还可以将其用途扩展到其他编程语言。

（A notebook is a JSON document, keeping with the general portability principle of Python.）

（notebook是一个JSON文档，符合Python的可移植性原则。）

Page75

## 2.2 Firsts Slithers with Python

## 2.2 初探Python

We have already had the opportunity to interact with the iPython shell and have seen some simple operations such as addition (+) and division (/). We can continue our exploration of the programming language as an interactive calculator. As we would expect, the rest of the arithmetic operations are supported by Python as shown in Table 2.1. Notice that the exponentiation in Python is represented with two stars, \*\*. So far we have used integers and floating point numbers in the examples presented. It is therefore natural to ask what other types are supported by Python.

我们已经有机会用iPython shell进行交互，并看到了一些简单的操作，如加法（+）和除法（/）。 我们可以继续探索编程语言作为交互式计算器的用法。 正如我们所期望的那样，Python支持其余的算术运算，如表2.1所示。 请注意，Python中的取幂用两颗\*\*表示。 到目前为止，我们在所提供的示例中使用了整数和浮点数。 因此，很自然地会想到Python究竟支持其余哪些类型。

（Exponentiation is denoted with \*\* in Python.）

（取幂在Python中用\*\*表示。）

（Table 2.1: Arithmetic operators in Python.）

（表2.1：Python中的算术运算符。）

### 2.2.1 Basic Types

### 2.2.1基本类型

An important fact to remember about Python is that it is a dynamically typed language. In other words, we do not need to declare variables before we use them and it is not necessary to specify their type. Furthermore, each and every variable that we create is automatically a Python object.

记住Python的一个重要事实是它是一门动态语言。 换句话说，我们使用之前不需要声明变量，也没有必要指定变量的类型。 此外，我们创建的每个变量都自动成为一个Python对象。

（Python is a dynamically typed language: We do not need to specify the variable type in advance.）

（Python是一种动态类型语言：我们不需要提前声明变量类型。

### 2.2.2 Numbers

### 2.2.2数字

As we saw in the previous section, Python supports two types of numbers: Integers and floating point numbers. So we can assign the value of an integer to a variable as follows:

正如我们在上一节中看到的，Python支持两种类型的数字：整数和浮点数。 所以我们可以将一个整数的值赋给变量，如下所示：

Page77 图

（Python supports integers and floating point numbers.）

（Python支持整数和浮点数。）

Notice that assignation does not require Python to print anything as a response. We can check the type of an object with the command type:

请注意，赋值操作不需要Python打印该内容作为响应。 我们可以使用type命令检查对象的类型：

（The command type lets us see the type of an object.）

（命令type让我们看到对象的类型。）

Python will let us know what type of object we are dealing with; in this case the object Universe is of type integer. Let us see an example for a floating point number:

Python会让我们知道我们正在处理什么类型的对象; 在这种情况下，对象Universe的类型为整数。 让我们看一个浮点数的例子：

### 2.2.3 Strings

### 2.2.3 字符串

A string is effectively a sequence of characters. In Python, strings can be defined with the use of either single (’ ’) or double quotes ( ‘‘ ’’) as follows:

字符串实际上是一系列字符。 在Python中，可以使用单引号（''）或双引号（''''）定义字符串，如下所示：

Page78图

（Strings in Python can be defined with single or double quotes.）

（Python中的字符串可以用单引号或双引号定义。）

In the example above we have requested the type of the variable string1, and as expected, Python tells us that indeed it is a string. We can ask Python to print a string as follows:

在上面的例子中，我们已经请求打印变量string1的类型，并且正如预期的那样，Python告诉我们它确实是一个字符串。 我们可以让Python打印一个字符串，如下所示：

（In Python 2 the print statement does not require the brackets.）

（在Python 2中，print语句不需要括号。）

The + operator is overloaded for strings and it results in the concatenation of these objects:

+运算符可以进行字符串对象的连接操作：（python内部重载了该操作符以便于字符串连接操作）

（Concatenation of strings can be achieved with the + symbol.）

（可以使用+符号实现字符串的连接。）

In the example above we have also demonstrated the way in which Python is able to deal with simultaneous assignation in a single line. In other words, the string “Norwegian” is assigned to the variable dead and the string “blue” to the variable parrot. Note that mixing operators between strings and numbers is not allowed, and an error will be thrown if they are. Instead you will have to convert a number to a string using the str function.

在上面的示例中，我们还演示了Python能够在单行中同时分配变量的方式。 换句话说，字符串“Norwegian”被赋值给变量dead，字符串“blue”被赋值给变量parrot。 请注意，不允许在字符串和数字之间混入运算符，如果是，则会抛出错误。 你必须使用str函数将数字转换为字符串才能进行相关操作。

（Python is able to carry out multiple assignation in a single line. This is part of a Pythonic programming style!）

（Python能够在一行中执行多个赋值。 这是Pythonic编程风格的一部分！）

Strings are immutable objects in Python and this means that we cannot change individual elements of a string. We shall discuss more about immutable objects in the context of tuples in Section 2.2.6.

字符串是Python中的不可变对象，这意味着我们无法更改字符串的各个元素。 我们将在2.2.6节中讨论不可变对象元组的更多内容。

（Strings in Python are immutable.）

（Python中的字符串是不可变的。）

### 2.2.4 Complex Numbers

### 2.2.4 复数

Python also supports complex numbers, and it denotes the imaginary number *i* = *p-*1 as j, and so for a number n, nj is interpreted as a complex number.

Python也支持复数，它将虚数i = p-1表示为j，因此对于数n，nj被解释为复数。

Page79 开方

（In Python, the imaginary number i is denoted with the letter j.）

（在Python中，虚数i用字母j表示。）

Let us see an example: If we want to define the complex number z = 2 + 3i we simply tell Python the following:

让我们看一个例子：如果我们想要定义复数z = 2 + 3i，我们只需告诉Python如下：

Page79 图

Please note that although the numbers used in the example above are integers, Python recasts them as floating point numbers to suit the complex number object. In the piece of code shown above we have also demonstrated the fact that we can use the backslash (n) to break a line for code readability.

请注意，尽管上面示例中使用的数字是整数，但Python会将它们重新设置为浮点数以适合复数对象。 在上面显示的代码段中，我们还演示了这样一个事实：我们可以使用反斜杠（\n）来换行提高代码可读性。

（The backslash allows us to break a line.）

（反斜杠允许我们将一行代码分成多行。）

Remember that each and every entity in Python is an object. Each object has a number of possible actions they are able to perform, i.e. methods. In the example above we have called the real and imag methods associated with a complex number object to obtain the real and imaginary parts respectively. Another use of a method is shown in the example above for a string, in this case the format method to tell Python how to format the printing of a string.

请记住，Python中的每个实体都是一个对象。 每个对象都有许多可以执行的操作，即方法。 在上面的例子中，我们调用了与复数对象相关的real和imag方法，分别获得实部和虚部。 上面的示例中显示了一个方法的另一种用法，用于字符串，在本例中是格式化方法，告诉Python如何格式化字符串的并进行打印。

（The method of an object can be invoked by following the name of object with a dot (.) and the name of the method.）

（可以通过跟随带有点（.）的对象名称和方法名称来调用对象的方法。）

### 2.2.5 Lists

### 2.2.5 列表

A list is pretty much self-explanatory: It is a sequence of objects, and these objects can be either of the same type or not. We denote a list with square brackets, [ ]. Lists are mutable objects and therefore it is possible to change individual elements in a list:

列表几乎是不言自明的：它是一系列对象的集合，这些对象可以是相同类型也可以不是。 我们用方括号[]表示一个列表。 列表是可变对象，因此可以更改列表中的单个元素：

Page 80图

（A list is denoted by square brackets [ ].）

（列表用方括号[]表示。）

It is possible to refer to elements of a list using an index that corresponds to their position in the list: 可以使用与列表中的位置对应的索引来引用列表的元素：

（We can refer to elements in a list with an index）

（我们可以引用带索引的列表中的元素）

（Indexing in Python starts at zero.）

（Python中的索引从零开始。）

Indexing in Python starts with the number zero and thus the first element of the numbers list is referred to as numbers[0]. Also, we can refer to a sub-sequence of the list using the colon notation as start:end, where start refers to the first element we want to include in the sub-sequence and end is the last element we want to consider in the slice.

Python中的索引从数字零开始，因此数字列表的第一个元素称为数字[0]。 此外，我们可以使用冒号表示引用列表的子序列，形如start：end，其中start是指我们想要包含在子序列中的第一个元素，end是我们要在其中使用的最后一个元素，从而形成一个切片。

Remember that Python interprets the slicing operation up to, but not including, the last item in the sequence. In the example above, Python reads from index 1 and up to index 3, but not including 3. That is why only the second and third elements of the numbers list are returned.

请记住，Python将切片操作解释为不包括序列中的最后一项。 在上面的示例中，Python从索引1读取到索引3，但不包括3.这就是为什么只返回数字列表的第二个和第三个元素。

（Slicing refers to the subsetting of an array-like object such as lists and tuples.）

（切片指的是类似数组的对象的子集，例如列表和元组。）

Since lists are mutable objects it is possible for us to change elements in a list:

由于列表是可变对象，因此我们可以更改列表中的元素：

Page81 图

（We are able to change the elements of a list because they are mutable objects.）

（我们能够更改列表的元素，因为它们是可变对象。）

We can also add elements to a list with the append method:

我们还可以使用append方法将元素添加到列表中：

（append lets us add elements to a list.）

（append可将元素添加到列表中。）

The new element, 1.4142, is added to the numbers list at the end, increasing the length of the list by one element.

新元素1.4142将添加到数字列表的末尾，同时列表的长度增加一。

Concatenation of lists is easily achieved with the + operator:

使用+运算符可以轻松实现列表的连接：

Page82 图

（The + symbol lets us carry out list concatenation）

（+符号让我们可以进行列表连接）

Notice that if the two lists are numerical, the result using the + operator is again the concatenation of the list elements, not the sum.

请注意，如果两个列表都是数字，则使用+运算符的结果是列表元素的串联，而不是数字总和。

Another useful method of a list is sort, which does exactly what we would expect: It allows us to sort the list’s values. This method will also enable us to see the difference between mutable and immutable objects in our discussion about tuples (Section 2.2.6).

列表的另一个有用方法是sort，它的用法完全符合我们的预期：它允许我们对列表的值进行排序。 这个方法还使我们能够在讨论元组时看到可变对象和不可变对象之间的区别（第2.2.6节）。

（The sort method allows us to sort a list in place.）

（sort方法允许我们对列表进行排序。）

Let us define a list to work with:

让我们定义一个列表来使用：

Page82 图

We can now invoke the sort method as follows:

我们现在可以调用sort方法，如下所示：

（As we can see using sort with a list results in the elements being ordered in ascending order.）

（正如我们所看到的，使用带有列表的sort会导致元素按升序排序。）

There are a couple of things to note here. First, we have called the sort method using the dot (.) notation. When executing the first line in the code above, the interpreter does not return any values, and that is a good sign: It means that the method executed correctly.

这里有几点需要注意。 首先，我们使用点（.）表示法调用sort方法。 在上面的代码中执行第一行时，解释器不返回任何值，这是一个好兆头：这意味着该方法以正确执行。

In order to see what happened we issue the second command, which lets us print the contents of List1. As shown above, the elements of the list are now ordered.

为了看看发生了什么，我们发出了第二个命令，它允许我们打印List1的内容。 如上所示，现在列表的元素已经正确排序。

This takes us to the second point to note. Since lists are mutable, we can change them and in this case the sort method has changed the elements in List1 to be in ascending order. We have sorted the list in place. There was no need to create a copy of the list and sort it.

我们第二点需要注意的地方： 由于列表是可变的，我们可以更改它们，在这种情况下，sort方法已将List1中的元素更改为升序。 我们已将列表排序到位。 无需创建列表副本并对其进行排序。

（Since lists are mutable, we are able to change their elements. In this case sorting the elements for instance.）

（由于列表是可变的，我们可以改变它们的元素。 在这种情况下，例如对元素进行排序。）

Objects in Python also have functions associated with them. Lists are no exception and in this particular case there is a sorted function too. The difference is that a function will create a new object. Let us take a look:

Python中的对象也具有与之关联的方法。 列表也不例外，在这种特殊情况下也有一个排序函数。 不同之处在于该方法函数将创建一个新对象。 我们一起来看看：

Page83 图

（Lists have a sorted function.）

（列表具有排序方法。）

So far so good, nothing has changed, we end up with a sorted list. However, let us take a look at the List1 object one more time

到目前为止一切顺利，和期望一样，我们最终同样得到一个排序列表。 但是，让我们再看一次List1对象

As you can see, the object was not changed! Instead, what the sorted function has done is create a new object with the contents of List1 in ascending order.

如您所见，对象值居然没有改变！ 排序函数所做的是创建一个新对象并返回以升序排列的List1的内容。

（The sorted function creates a new object with the elements of the original list, but in ascending order.）

（排序函数使用原始列表的元素创建一个新对象，但按升序排列其内容。）

We could have assigned the result of the function to a new variable and thus create an object that can be referred to at a later stage.

我们可以将函数的结果分配给一个新变量，从而创建一个可以在稍后阶段引用的对象。

Incidentally, if you require the elements in descending order all you have to do is pass the reverse parameter to either the method:

顺便提一下，如果您需要按降序排列元素，您只需将reverse作为参数传递给方法：

Page 84 图

or the function:

或者使用函数

A very useful pythonic way of constructing lists without the need of a full-blown loop is the so-called list comprehension. A typical usage is in the creation of lists whose elements are the result of some operations applied to each member of another sequence or iterable. For example, let us create a string with a sentence:

一个非常有用的pythonic方法是所谓的列表推导式：构建列表而不需要写出完整循环流程。 典型的用法是创建列表，其元素是应用于另一个序列或可迭代的每个成员的一些操作的结果。 例如，让我们首先创建一个包含句子的字符串：

Page85图

We can use the string above to create a list of lists with each word in the sentence in capital and lower-case letters, as well as determining the length of the word. And we can do all this in a single line of code:

我们可以使用上面的字符串创建一个列表，其中包含句子中的每个单词的大写和小写字母，以及确定单词的长度。 我们可以在一行代码中完成所有这些工作：

Page85图2

（We are using the string methods split(), upper() and lower() to separate the words in the sentence, and convert them to upper- and lower-case.）

（我们使用字符串方法split（），upper（）和lower（）来分隔句子中的单词，并将它们转换为大写和小写。）

（List comprehension is useful when we need to create a list out of operating on elements of another sequence.）

（当我们需要创建一个对另一个序列的元素进行操作的列表时，列表推导式是非常有用的。）

### 2.2.6 Tuples

### 2.2.6 元组

A tuple may be seen as a list by another name: They are also sequences of objects, and they may be of mixed type too. They are indeed closely related to lists and apart from the fact that they are defined with round brackets, ( ), the main difference is that tuples are immutable.

元组可以被视为另一种形式的列表：里面的元素不仅可以是对象的序列集合，也可以是混合类型。 它确实与列表密切相关，除了用圆括号（）定义之外，其主要区别在于元组是不可变的。

（Tuples are defined with round brackets ( ).）

（元组用圆括号（）定义。）

As we have mentioned above, immutable objects cannot be changed. In other words, we cannot add or remove elements and thus, unlike lists, they cannot be modified in place. Let us take a look at some tuples:

如上所述，无法更改不可变对象。 换句话说，我们不能添加或删除元组里面的元素，因此，与列表不同，它们不能在适当的位置进行修改。 我们来看看一些元组：

Page86图

（Tuples are immutable objects）

（元组是不可变对象）

As you can see the only change in the definitions above, compared to the lists in Section 2.2.5, is the use of the round brackets. As with lists, the elements of a tuple can be referred to by their index:

正如您所看到的，与2.2.5节中的列表相比，上述定义中的唯一更改是使用圆括号。 与列表一样，元组的元素也可以通过索引来引用：

Page86图2

（Tuples can also be sliced with the help of an index.）

（元组也可以在索引的帮助下进行切片操作。）

Let us see what happens when we try to change one of the elements of a tuple:

让我们看看当我们尝试更改元组的一个元素时会发生什么：

Page86图2

（We are not able to change elements of a tuple as they are immutable objects.）

This shows that there are manipulations that are not possible to be done with a tuple. What about sorting? Well, the sorted function still works. Let us define a tuple as follows:

（我们无法更改元组的元素，因为它们是不可变对象。）

这表明直接在元组里面进行的操作是不可能的。 排序怎么样？ 好吧，排序的功能仍然有效。 让我们定义一个元组如下：

Page87图

We can now apply the sorted function to the tuple:

我们现在可以将排序函数应用于元组：

Page87图2

（The result of the sorted function on a tuple is a list.）

Not too bad, right?, but have you noticed something odd? Well, it seems that the result is not a tuple anymore, but a list! We can see that thanks to the square brackets, and we can make sure of this by using the type command:

（对元组进行排序函数操作返回的结果是一个列表。）

还不错，对吗？但是你注意到了一些奇怪的东西吗？ 好吧，似乎结果不再是一个元组，而是列表！ 我们可以看到，是方括号而非圆括号，我们可以通过使用type命令来确保这一点：

Page87图3

This is the result of tuples being immutable: The only way to allow for the elements of the tuple to be ordered is by using the mutable nature of a list. Similarly, since the elements of a tuple cannot be changed, there is no point in having a sort method. Let us have a look:

（Since tuples are immutable we cannot change their elements in place.）

这是元组不可变的结果：允许元组元素排序的唯一方法是使用列表的可变特性将其转换为列表再排序。 类似地，由于元组的元素不能改变，因此无法直接对其使用排序方法。 我们来看看：

（因为元组是不可变的，所以我们不能改变它们的元素。）

Page88图

（As stated by the error returned by Python, tuples do not have a sort attribute.）

（正如Python返回的错误所述，元组没有sort属性。）

### 2.2.7 Dictionaries

### 2.2.7 字典

We are all familiar with the concept of a dictionary: If we are interested in finding the meaning of a new or unknown word, we simply open up a book (or access a webpage) that lists words in a specified order (alphabetically, for instance). This enables us to search for the word we are interested in. A dictionary in Python serves the same purpose and it is composed of keys and values.

（In the analogy with actual dictionaries, keys are equivalent to words and values are the definitions.）

我们都熟悉字典的概念：如果我们有兴趣查询新词或未知词的含义，只需打开一本书（或访问网页），里面会按照指定的顺序列出单词（例如，按字母顺序排列））。 这使我们能够索引到我们感兴趣的单词.Python中的字典具有相同的用途，它由键和值组成。

（与实际字典类比，键相当于单词，值是定义。）

A Python dictionary is defined with the use of curly brackets, { }. Furthermore, the key-value pairs are separated by a colon (:) as follows:

使用大括号{}来定义Python的字典类型。 此外，键值对由冒号（:)分隔，例子如下：

Page88图2

（We define a dictionary with curly brackets { }.）

（我们用大括号{}定义一个字典。）

The keywords in a dictionary can be any immutable Python object including numbers, strings and tuples. The value associated with a particular key can be changed by reassigning the new value to the element of the dictionary with the relevant entry. For instance, in our example above we can see that the value to the key spam is 4:

字典中的key（关键字）可以是任何不可变的Python对象，包括数字，字符串和元组。 可以通过索引该条目将新值重新分配给字典的关联元素从而更改相关键所对应的值。 例如，在我们上面的示例中，我们可以看到Key(关键字) “Spam”的对应的值是4：

Page89图

（The keys can be any immutable object: numbers, strings or tuples for example.）

We can change the value of this key by simply reassigning any new value. We can for example reassign the value associated to the spam key:

（键可以是任何不可变对象：例如数字，字符串或元组。）

仅需重新分配任何新值就能更改此键对应的值。 例如，我们可以重新分配与key(关键字)"Spam"相关联的值：

Page89图2

（The values in a dictionary can be modified.）

This can be done repeatedly:

（可以修改字典中的值。）

可以对同一个key重复赋值：

Page89图3

（The modification can be done as many times as required.）

It is possible to access the keys and values in the form of straight lists with the aid of the keys() and values() methods:

（可以根据需要进行多次修改。）

借助keys（）和values（）方法，可以以列表的形式遍历字典的键和值：

Page89图4

（A list of dictionary keys can be obtained with the keys() method. Similarly, values() returns a list of values in a dictionary.）

We can also obtain the key-value pairs in the form of a list of tuples with the items() method:

（可以使用keys（）方法获得字典键列表。 类似地，values（）返回字典中的值列表。）

我们还可以使用items（）方法以元组列表的形式获取键值对：

Page90图

Finally, it is possible to get rid of key-value pairs with the use of the del function:

最后，可以使用del函数删掉字典里面的键值对：

Page90图2

（We can remove entries from a dictionary with del.）

（我们可以用del删除字典中的条目。）

## 2.3 Control Flow

## 2.3 控制流

Not only is it important to understand the types and objects that are available in any programming language, but also how to control the flow of a programme to be able to follow the logic behind the way in which the programme itself is organised, in other words, the order in which the individual statements are executed.

我们不仅需要理解编程语言中可用的类型和对象，而且还需要学会如何遵循程序语言组织方式逻辑来组织控制自己的程序流程。换句话说 ，组织各个语句的先后执行顺序。

In particular it is important to mention that in Python the whitespace is a meaningful character as it enables the definition of blocks of code by having the same level of indentation. Let us see some typical structures to control the flow of a programme in Python.

特别值得一提的是，在Python中，空格是一个有意义的字符，因为它可以通过具有相同级别的缩进来定义代码块。 让我们看一些控制Python程序流程的典型结构。

（Whitespace is a meaningful character in Python.）

（Whitespace在Python中是一个有意义的角色。）

*2.3.1* if... elif... else

Conditional branching enables us to perform different actions depending on the result of boolean operations. If a condition is met, then we apply an operation, otherwise a different action is performed. In Python we can do this as follows:

### 2.3.1 if ... elif ... else

条件分支使我们能够根据布尔运算的结果执行不同的操作。 如果满足条件，则我们应用操作，否则执行不同的操作。 在Python中，我们可以这样做：

Page91图

（The if... elif... else... lets us test various conditions and create branches for our code.）

As you can see, each block of code is indented at the same level. Also, notice that it is possible to nest various conditions with the help of the elif reserved word. The conditions are logical expressions that can test for scalar comparison and thus we can use any of the comparison operators listed in Table 2.2. Let us see an example:

（The conditions to test are logical expression that evaluate to True or False.）

（if ... elif ... else ...让我们测试各种条件并为我们的代码创建分支。）

如你所见，每个代码块都缩进到同一级别。 另外，请注意，可以在elif保留字的帮助下嵌套各种条件。 条件是可以比较的逻辑表达式，因此我们可以使用表2.2中列出的任何比较运算符。 让我们看一个例子：

（测试的条件实质是True或False的逻辑表达式。）

Page92 图

（Table 2.2: Comparison operators in Python.）

Finally, remember that the conditions are tested one by one in the order they are provided in the code. If a condition is met, the rest of the tests are not executed.

（表2.2：Python中的比较运算符。）

最后，请记住，条件是按照代码中提供的顺序逐个执行的。 如果满足条件，则不执行其余部分。

### 2.3.2 while

### 2.3.2 while表达式

A while loop is used when we need to repeat a block of code until a condition is no longer met. The structure of a while loop in Python is:

当我们需要重复执行代码块直到不再满足条件时，可以使用while循环。 Python中while循环的结构是：

Page92 图2

An important thing to remember is that at the very beginning of the while loop, the logical test must evaluate to True, otherwise the block of code is never executed. Also, in order to avoid infinite loops we need to update the control variable inside the block of code.

（The while loop requires a logical test at the beginning of the block.）

We can see how this works by counting down from 10:

要记住的一件重要事情是，在while循环的最开始，逻辑表达式必须求值结果为True，否则永远不会执行该代码块。 另外，为了避免无限循环，我们需要更新代码块内的控制变量。

（while循环需要在块的开头进行逻辑测试。）

我们可以从倒计时10看到这是如何工作的：

Page93 图

（Note that countdown -= 1 is a shorthand for countdown = countdown - 1.）

（注意countdown -= 1 是countdown = countdown - 1的简写）

### 2.3.3 for

### 2.3.3 for表达式

Typically, a while loop is used in cases when we do not know in advance how many times the block of code will need to be executed. If we know how many iterations are needed, we can use a for loop. In Python, a for loop iterates over a sequence: a list, tuple or a string for example.

（A for loop is useful when we know how many times the code needs to be repeated.）

通常，在我们事先不知道需要执行代码块多少次的情况下推荐使用while循环。 如果我们知道需要多少次迭代，我们可以使用for循环。 在Python中，for循环可以遍历序列：例如list，tuple或string。

（当我们知道代码需要重复多少次时，for循环很有用。）

Page93 图2

（This is the same basic structure used in list comprehension.）

The example we used for the while loop in Section 2.3.2 can be written with a for loop as follows:

（这与列表推导式中使用的基本结构相同。）

我们在2.3.2节中使用的while循环的示例也可以用for循环编写，如下所示：

Page94 图

We could simplify the example above by avoiding the explicit definition of the list and instead define a range:

我们可以用列表推导式来通过range生成列表从而简化避免显示的定义一个列表：

Page94 图2

（range enables us to define a sequence of numbers as an object. This means that the values are generated as they are needed.）

（range使我们能够将数字序列定义为对象。 这意味着可以只在需要时再生成值。）

In the example above we used the range(start, end, step) function to generate a sequence of numbers from start to end-1 in steps given by step. In Python 2, similar behaviour is obtained with the xrange function.

在上面的例子中，我们使用range（开始，结束，步长）函数从 start 到 end-1生成指定步长的数字序列。 在Python 2中，xrange函数有类似的行为。

### *2.3.4* try... except

### *2.3.4* try... except异常捕获

It is not unusual to have syntactically correct blocks of code with statements that in certain cases may cause an error during execution. These errors are not necessarily fatal in the execution of a programme and instead they are anomalous or exceptional cases that require special processing.

（Syntactically correct code may cause errors during execution.）

即使一个代码块语法完全正确，但这些语句在某些情况下可能会在执行期间导致错误，这种现象并不奇怪。 这些错误在执行程序时不一定是致命的，而是需要特殊处理的异常或异常情况。

（语法正确的代码可能会在执行期间导致错误。）

Instances such as those described above are called exceptions and when they happen, we are interested in catching them and taking appropriate action, for example by generating an error message. This is what is known as exception handling. In Python this can be done with the try...except structure:

上述实例称为异常，当它们发生时，我们需要将其捕获并采取适当的操作，例如生成错误消息等。 这就是所谓的异常处理。 在Python中，可以通过try ... except结构来完成：

Page95 图

（Exception handling in Python can be done with the try... except structure.）

In the structure above exception1, exception2,... are standard exceptions that Python knows about and that are detailed in the appropriate documentation; we list some common ones in Table 2.3.

（For further information about standard exceptions see https: //docs.python.org/2/library/ exceptions.html）

（Python中的异常处理可以使用try ... except结构完成。）

在上面的结构中，exception1，exception2，...是Python知道的标准异常，在相应的文档中有详细说明; 我们在表2.3中列出了一些常见的。

（有关标准异常的更多信息，请参阅https：//docs.python.org/2/library/ exceptions.html）

For example, we can try to calculate the reciprocal of the elements of a list and print each of the values. However, if the sequence contains the number zero, we can try to catch the exception with ZeroDivisionError:

例如，我们可以尝试计算列表元素的倒数并打印每个值。 但是，如果序列中包含数字0，我们可以尝试使用ZeroDivisionError来捕获异常：

Page96图

Table 2.3: Standard exceptions in Python.

表2.3：Python中的标准异常。

We are using the ZeroDivisionError exception to handle this particular exception.

As you can see, instead of getting an error message and the interpreter halting programme execution, the exception is handled nicely by the code after the appropriate except entry.

我们使用异常ZeroDivisionError来处理这个特殊的异常。

正如你所看到的，通过适当的except方法之后，由代码很好地处理异常，从而避免抛出错误消息导致程序停止执行的情况出现

### 2.3.5 Functions

### 2.3.5 函数

Now that we have covered some of the elementary control flow structures in Python, we can start combining them into logical blocks to carry out specific tasks. In particular we can construct pieces of code that can be repeated when necessary and whose outcome depends on the input parameters provided. In other words, we are talking about functions.

我们已经介绍了Python中的一些基本控制流结构，并且可以开始将它们组合成逻辑块来执行特定任务。 特别地，我们可以构造出在必要时重复执行的代码片段，其结果取决于所提供的输入参数。 接下来，我们谈论的话题是函数。

（A function is a good way to write code that can be repeated, and whose outcome typically depends on inputs provided.）

A function in Python has the following syntax:

（函数是编写可重复执行代码的好方法，其输出通常取决于输入。）

Python中的函数具有以下语法：

Page97图

（The function definition starts with the word def. Remember that code needs to be indented.）

（函数定义以单词def开头。 请记住，代码需要缩进。）

Notice that the function definition starts with the reserved word def and the code inside the function is indicated with appropriate indentation.

请注意，函数定义以保留字def开头，函数内部的代码用适当的缩进表示。

The input parameters for the function are the dummy variables arg1, arg2,... , argn and as you can see it is possible to define default values for some of these parameters. Parameters with default values must be defined last in the argument list.

该函数的输入参数是变量arg1，arg2，...，argn，你可以为其中一些参数定义默认值。 定义具有默认值的参数必须放在参数列表的最后。

The second line in the function definition is called the documentation string and its purpose is to describe the actions that are performed by the function. Finally, notice that it is not necessary for a function to return a result.

函数定义中的第二行称为文档字符串，其目的是说明该函数的执行步骤和目的。 最后，请注意函数不要求必须有返回值。

（A documentation string enables us to provide information about what a function does. Make sure you use it!）

（Let us define a function to calculate the area of a rectangle sides a and b:）

（文档字符串使我们能够提供有关函数功能的信息。 一定要用它！）

（让我们定义一个函数来计算由边a和b所组成的矩形面积:)

Page98图

（We are defining a default value for the parameter b.）

（我们正在为参数b定义一个默认值。）

Notice that the parameter b has been given the default value of 1. If we were to call this function with only one parameter, the function will know how to handle the calculations and use the default values when needed.

请注意，参数b的默认值为1.如果我们调用此函数只传入一个参数，该函数将知道如何处理计算并在需要时使用默认值。

Page98图2

（We can use the function by calling it in the same any other in-built Python function is.）

（函数定义后我们可以就可以通过其名字直接调用了）

In the first line of code above, we are calling the rect\_area function with two parameters, such that we assign the value 20 to a and override the default value of b with 2. As expected the area calculated is 40. Let us try providing only one single value to the function:

在上面的第一行代码中，我们使用两个参数调用rect\_area函数，这样我们就将值20赋值给a并用2覆盖b的默认值。正如预期的那样，计算的面积是40.让我们尝试只提供一个单一值来调用函数：

Page99图

Here we have only passed the value 42.4 to the function. In this case the value is assigned to a and the default value of b = 1 is used in the calculation.

这里我们只将值42.4传递给函数。 在这种情况下，值将分配给a，并在计算中使用默认值b = 1。

We can include control flow structures in our programmes to make them more useful and flexible. Let us for instance implement a simple function to calculate the factorial of a number:

我们可以在程序中包含控制流结构，以使它们更有用和灵活。 举个例子，我们实现一个简单的函数来计算数字的阶乘：

Page99图2

（A function can use any of the other control flow structures of the language.）

（\*= and -= indicate repeated operations with the left-hand-side value.）

（函数可以使用python中的任何控制流结构。）

（\* =和 - =表示使用左侧值重复操作。）

When we pass a number smaller or equal than the one the function expects, it returns the value 1, and when the number is greater than 1 the factorial is calculated with a while loop. Let us use the function:

当我们传入一个小于或等于函数期望值的数字时，它返回值1，当数字大于1时，使用while循环计算阶乘。 让我们使用这个函数：

Page100图

There may be times when it is more convenient to define a simple function on-the-fly, without having to resort of a full def structure. In these cases we can exploit the use of the so-called lambda functions:

有时可能更方便的是即时定义一个简单的函数，而不必使用完整的def结构。 在这些情况下，我们可以利用所谓的lambda函数：

（A lambda function in Python is an anonymous function created at runtime.）

（Python中的lambda函数是在运行时创建的匿名函数。）

Page100图2

where, as before, arg1, arg2,... are the input parameters and statement is the code to be executed with the input provided.

与之前一样，arg1，arg2，...是输入参数，而statement是要根据输入参数执行的代码。

For example, if we needed to calculate the cube of a list of numbers we could try the following code:

例如，如果我们需要计算数字列表的每个元素的立方，我们可以尝试以下代码：

Page100图3

（In this case the object g is a lambda function that can be called as any other function in Python.）

（在这种情况下，对象g是一个lambda函数，可以像Python中的任何其他函数一样调用。）

So far nothing too strange: We have initialised a list with the numbers 1, 3 and 6, and then defined a lambda function that calculates the cube of the argument n. We can now apply this function, for example:

到目前为止，没有什么值得奇怪的地方：我们已经用数字1,3和6初始化了一个列表，然后定义了一个lambda函数来计算参数n的立方体。 我们现在可以应用此功能，例如：

Page101图

Lambda functions may seem very simple, but it is that simplicity that provides their strength, as it shown above. This can be seen employed for instance in the implementations of PySpark, the Python API for Spark, an open-source cluster computing framework.

Lambda函数可能看起来非常简单，但正如上面所示，简单确非常实用。 可以在PySpark的源代码中看到它的身影。PySpark是用于Spark的Python API，一种开源集群计算框架。

（Lambda functions are very useful in frameworks such as Spark.）

（Lambda函数在Spark等框架中非常有用。）

### 2.3.6 Scripts and Modules

### 2.3.6 脚本和模块

With the flexibility provided by the possibility of controlling the flow of a set of instructions, and the repeatability offered by constructing our own functions, it becomes imperative to be able to store programmes in a way that enable us to use and reuse code.

由于我们可以灵活的控制一组指令流，并且通过调用我们自己的函数来重复操作，因此能够将代码保存下来并重复使用函数变得势在必行。

In Python we are able to do this by saving the instructions that make up a programme in a plain text file saved with the extension .py. Furthermore, if we use the interactivity provided by the iPython/Jupyter notebook, it is also possible to save our notebooks in a JSON formatted notebook with the extension .ipynb.

在Python中，可以通过在扩展名为.py的纯文本文件中保存我们的代码。 此外，如果我们使用iPython / Jupyter notebook所提供的交互性，也可以将代码保存在扩展名为.ipynb的JSON格式文件中。

（Python scripts have the extension .py whereas notebooks have the extension .ipynb.）

（Python脚本的扩展名为.py，而notebooks 中保存时扩展名为.ipynb。）

It is then possible to execute a Python script from the command line by calling Python followed by the name of the script to be executed. For instance, we can create a script defining a main function and a call to it. We can save the function in a script called firstscript.py with the following contents:

然后可以通过调用Python+执行脚本的名称在从命令行执行Python脚本。 例如，我们可以创建一个定义为Main主函数和调用它的脚本， 然后将函数保存在名为firstscript.py的脚本中，其中包含以下内容：

Page102图

（We are defining a main function in this programme and calling simply with the command main().）

In this case we are asking the user for a number n with the command input. We then use this number to calculate a sequence given by the square of the numbers from 0 to n and assign it to the variable y. Finally we simply print the list stored in y.

（Notice that we have used n+1 for xrange.）

（我们在这个程序中定义一个main函数，并使用命令main（）调用。）

在这种情况下，我们通过命令行询问用户来输入数字n。 然后，我们使用此数字来计算由0到n的数字的平方给出的序列，并将其分配给变量y。 最后，我们只需打印存储列表y。

（注意我们已经使用n + 1作为xrange。）

Remember that we have saved the script above, but we have not executed it. We can do this by typing the following command in a terminal in the appropriate path:

请记住，我们已经保存了上面的脚本，但我们还没有执行它。 我们可以通过在包含该脚本所在路径的终端中键入以下命令来完成此操作：

Page102图2

（In this case we have given the value n = 4 as an input.）

This is perhaps not the most advanced algorithm to implement, but we can surely see the possibilities. In particular, we can see how we can create scripts to add further functionality to our code and as such the concept of a module becomes natural.

（在这个用例中，我们给出了值n = 4作为输入。）

这可能不是最优算法，但我们可以肯定脚本起码可用。 特别是，我们可以看到如何创建脚本并向其中添加更多功能，因此自然而然的引出了下面模块的概念。

A module is a file or collection of files containing related Python functions and objects to achieve a defined task. These modules enable us to extend the capabilities of the language, and create programmes that enable us to carry out specific tasks. Any user is able to create their own modules and packages and make them available to others. Some of these modules are readily available for us to be used and once appropriate installation is done all we need to do is import them whenever we need to use them.

模块是包含相关Python函数和对象的单一文件或文件集合，用以实现既定的任务。 模块使我们能够扩展python语言的功能，并创建使我们能够执行特定任务的程序。 任何用户都可以创建自己的模块和包，并将其提供给其他人使用。 一旦python安装完成，其中一些模块已经可以随时可供我们使用了，我们所需要做的就是在需要使用它们时导入即可。

(A module is a file containing related Python functions to achieve a specific task.)

For example, we can use the math module to access some common mathematical functions. Let us create for instance a script that implements a function to calculate the area of a circle. In this case we will need the mathematical constant p to carry out the calculations:

（模块是包含相关Python函数的文件，用于实现特定任务。）

例如，我们可以使用数学模块来调用一些常见的数学函数。 让我们创建一个脚本来实现一个计算圆的面积的函数。 在这种情况下，我们需要数学常数π来执行计算：

Page103图

（The math module contains some common mathematical functions.）

（We can use the value of π with math.pi.）

Running the program will result in the following output:

（数学模块中包含一些常见的数学函数。）

（我们可以使用在模块math.pi 中调用常量π）

运行该程序将输出如下：

Page104图

Notice that we need to tell the Python interpreter that the constant p is part of the math module by using the syntax math.pi. In the example above we are importing all the functions of the math module. This can be somewhat inefficient in cases where only specific functionality is needed. Instead we could have imported only the value of p as follows:

请注意，我们需要通过使用语法math.pi告诉Python解释器常量p是数学模块的一部分。 在上面的示例中，我们将导入math模块的所有函数。 在仅需要特定功能的情况下，这可能有些低效。 相反，我们可以只导入π的值，如下所示：

Page104图2

(In some cases it may be more efficient to load only the needed functionality from a module.)

（在某些情况下，仅从模块加载所需的功能可能更有效率。）

A large number of modules are available from the Python Standard Library and more information can be found in https://docs.python.org/2/library/. In the rest of the book we will deal with a few of these modules and packages.

Python标准库提供了大量模块，更多信息可以在https://docs.python.org/2/library/中找到。 在本书的其余部分，我们将讨论其中的一些模块和软件包。

## 2.4 Computation and Data Manipulation

## 2.4计算和数据处理

With the programming structures discussed so far we are ready to take up a large number of tasks, not only in data science, but in more general settings. In our particular case, as we shall see in the rest of the book, computation with data and its manipulation can be managed more effectively and easily with the aid of linear algebra. In this section we will address some basic concepts in both data manipulation and linear algebra with Python.

使用目前所讨论的编程结构，我们已可以承担大量任务，不仅可以用于数据科学中，而且在一般的项目中也没有问题。 但在某些特定情况下，正如我们将在本书的其余部分中看到的那样，借助线性代数可以更有效，更轻松地管理数据计算及进行操作。 在本节中，我们将使用Python涉及数据操作和线性代数中的一些基本概念。

(Data manipulation and computation is a very important step in the data science and analytics workflow.)

（数据处理和计算是数据科学和分析工作流程中非常重要的一步。）

### 2.4.1 Matrix Manipulations and Linear Algebra

### 2.4.1矩阵操作和线性代数

As we have mentioned above, linear algebra enables us to carry out computational tasks with data in a very effective way. It also provides a compact notation to express the type of manipulations we need to do to our data, from pre-processing to presenting results. The use of vectors and matrices is therefore a very important area to cover. Vectors and matrices are arrays of numerical objects with a defined set of operations such as addition, subtraction, multiplication, etc.

正如我们上面提到的，线性代数使我们能够以非常有效的方式执行数据计算任务。 它还提供了一个简洁表示法来表达我们需要对数据进行的操作类型，从预处理到显示结果。 因此，我们需要覆盖向量和矩阵的使用这个非常重要的领域。 向量和矩阵是数值对象的数组，具有一组定义的运算，例如加法，减法，乘法等。

(Linear algebra provides us with an efficient and compact way to carry out complex calculations.)

（线性代数为我们提供了一种高效，简洁的方式来进行复杂的计算。）

An m × n matrix is a rectangular array of numbers having m rows and n columns. In particular when m = 1 we have a column vector and when n = 1 we have a row vector. In general, a matrix A can be represented as follows:

m×n矩阵是具有m行和n列的矩形数组。 特别的：当m = 1时，我们有一个列向量，当n = 1时，我们有一个行向量。 通常，矩阵A可以表示如下：

Page105 图

（A matrix can be thought of as a collection of row (or column) vectors.）

（矩阵可以被认为是行（或列）向量的集合。）

A Python object that may come to mind when thinking about using arrays is the list. For example, we can create two lists as follows:

在考虑使用数组时自然会想到使用Python的列表对象。 例如，我们可以创建两个列表，如下所示：

（See Section 2.2.5 for a discussion about lists.）

（有关列表的讨论，请参阅第2.2.5节。）

Page106 图

However, remember that Python considers these objects as lists and that each type of object has a defined set of operations. For instance, if we tried to add these two arrays in the mathematical sense we will find that Python returns an unexpected answer:

但是，请记住，Python将这些对象视为列表，并且每种类型的对象都有一组已定义的操作。 例如，如果我们尝试在数学概念上叠加这两个数组，我们会发现Python返回一个意外的答案：

（Page106 图2）

（Using the + symbol with lists results in concatenation.）

（使用带有列表的+符号会导致连接。）

Instead of adding the elements of each of the two vectors, Python concatenated the lists. This works because Python has overloaded the + symbol, but an error would be returned if we tried to use subtraction or multiplication.

Python没有叠加两个向量的元素，而是连接了列表。 这是因为Python重载了+符号，但如果我们尝试使用减法或乘法，则会返回错误。

（We covered list concatenation in Section 2.2.5.）

（我们在第2.2.5节中介绍了列表连接。）

（Page106 图3）

（Using other arithmetic symbols with lists results in an error.）

（将其他算术符号与列表一起使用会导致错误。）

It is clear that a list is a good start for the operations we need to execute, and the use of the programming capabilities of Python would enable us to build functions to define mathematical operations on lists to construct arrays. However, although it may be a very good programming practice, rather than building our own functions for this purpose, we can instead exploit the modules that are available to us within Python such as SciPy which provides an ecosystem for mathematics, science, and engineering, and in particular NumPy, a package that supports N-dimensional arrays.

很明显，列表是我们需要执行的操作的良好开端，并且使用Python的编程功能将使我们能够构建函数来定义数学运算并基于列表去构造数组。 然而，尽管它可能是一个非常好的编程实践，但构建我们自己的函数直接操作数组还是有些力不从心。所以我们可以改为利用Python中已有的模块，例如SciPy，它提供了数学，科学和工程的一套完整生态系统， 尤其是NumPy，一个支持N维数组的包。

（Some useful Python modules for array calculations are NumPy and SciPy.）

（用于数组计算的一些有用的Python模块是NumPy和SciPy。）

### 2.4.2 NumPy Arrays and Matrices

### 2.4.2 NumPy 数组和矩阵

NumPy extends the types supported by Python with the definition of arrays to describe a collection of objects of the same type. The dimension of a NumPy array is defined by a tuple of N positive integers called the shape of the array. We can think of arrays as an enhancement on lists and as such we can create arrays with the help of lists:

NumPy支持使用数组定义，扩展了Python支持的类型，该定义用于描述相同类型的对象集合。 NumPy数组的维度由称为数组形状的N个正整数的元组定义。 可以将数组视为列表的扩展，因此我们可以在列表的帮助下创建数组：

（Page107 图）

（NumPy extends the types in Python by including arrays.）

（NumPy通过包含数组扩展了Python中的类型。）

（We define a NumPy array with np.array, where np is a convenient alias used for the NumPy package.）

（我们使用np.array定义NumPy数组，其中np是调用NumPy包的简化别名。）

In the small piece of code above we are importing the NumPy package and using the alias np to refer to the module. With the aid of the array command in NumPy we transform a list into an array object. If we were to print the content of the array C we would obtain the following:

在上面的小段代码片段中，我们将导入NumPy包并使用别名np来引用该模块。 借助NumPy中的array命令，我们将列表转换为数组对象。 如果我们要打印数组C的内容，我们将获得以下内容：

（Page107 图2）

（The use of the + symbol with the arrays defined above results in their addition as expected.）

（使用+符号会导致带有上面定义的两数组按数学定义预期的方式叠加。）

Notice that in this case Python has indeed added the arrays element by element as expected. In the example above we could have simply used the list definitions from the previous section and written the following:

请注意，在这种情况下，Python确实按预期叠加了元素。 在上面的示例中，我们可以简单地使用上一节中的列表定义并编写如下：

（Page108 图）

As we have mentioned above, NumPy extends the functionality of lists in Python to be able to carry out vector arithmetics such as:

• Vector addition: +

• Vector subtraction: -

• Element-wise multiplication: \*

• Scalar product: dot()

• Cross product: cross()

正如我们上面提到的，NumPy扩展了Python中列表的功能，以便能够执行矢量算术，例如：

•矢量加法：+

•矢量减法： -

•元素乘法：\*

•点积：dot（）

•叉乘：cross（）

（These are some of the vector operations that are supported by NumPy arrays.）

You may have noticed that we have been referring to vector operations, but what about matrices? NumPy supports matrices too.

（这些是NumPy数组支持的一些矢量操作。）

您可能已经注意到我们一直在说矢量运算，但是矩阵呢？ NumPy当然也支持矩阵操作。

（Page108 图2）

（We can define matrices with the help of np.matrix.）

In this case we are using the command matrix to define the objects and the result of the multiplication is as expected:

（我们通过np.matrix来定义矩阵。）

在这种情况下，我们使用矩阵命令来定义对象，并且乘法的结果是符合预期的：

（Page109 图）

（We can multiply NumPy matrices with the usual multiplication symbol.）

（我们可以将NumPy矩阵与乘法符号相乘来使用。）

An alternative for defining NumPy matrices is to use the mat command to recast NumPy arrays.

A widely used operation in linear algebra is the transposition of a matrix. This can be readily accomplished with the use of the transpose command:

定义NumPy矩阵的另一种方法是使用mat函数来定义NumPy数组。

线性代数中广泛使用的操作是矩阵的转置。 使用transpose命令可以很容易地实现这一点：

（Page109 图2）

Finally, with the SciPy package we can use the linalg methods that will enable us to do some typical linear algebra computations such as matrix inversion:

最后，使用SciPy包，我们可以使用linalg方法，这将使我们能够进行一些典型的线性代数计算，如矩阵求逆：

（Page109 图3）

（Linear algebra methods are included in linalg inside SciPy.）

（We can invert a matrix with the .inv method. ）

（Matrix multiplication with arrays can be done with the dot() function.）

（线性代数方法包含在SciPy模块的linalg包中。）

（我们可以使用.inv方法反转矩阵。）

（使用dot() 函数可以对数组进行矩阵乘法。）

In the code above we have defined a couple of arrays, x and y. We have then calculated n = (xTx)−1 with the help of the .inv command from the linear algebra module. Note that the command .T in the code returns the transpose of a matrix. We then calculated k = xTy, and finally coe f = nk = (xTx)−1xTy.

在上面的代码中，我们定义了几个数组，x和y。 然后我们借助线性代数模块的.inv命令计算n =（xTx）-1。 请注意，代码中的命令.T返回矩阵的转置。 然后我们计算k = xTy，最后是coe f = nk =（xTx）-1xTy。 （数学符号需要敲进去）

（Page110 图）

（We shall come back to this calculation in the context of regression in Chapter 4.）

（我们将在第4章“回归”中详细讨论该计算的细节。）

We have deliberately called the result coe f , as we can think of the result of this simple calculation as the coefficients of a linear regression using the arrays x and y. We will come back to this result in Chapter 4.

我们故意将调用结果命名为coe f，因为我们可以将这个计算结果简单看作是使用数组x和y的线性回归的系数。 我们将在第4章回到这个结果进行进一步讨论。

### 2.4.3 Indexing and Slicing

### 2.4.3索引和切片

As it is the case with lists, it is possible to access the contents of an N-dimensional array by indexing and/or slicing the array. We can do this using the usual notation start:end:step which will extract the appropriate elements starting at start in steps given by step and until end−1.

与列表一样，可以通过索引或切片来访问N维阵列的内容。 我们可以使用通常的表示法start：end：step来执行此操作，该步骤将从starg开始，步长为step提取适当的元素，直到end-1结束。

（Page110 图2）

（Arrays and matrices can be indexed and sliced with the usual colon notation for lists and tuples.）

（对于列表和元组，通常可以使用":"表示法来对数组和矩阵进行索引和切片。）

In the example above we are selecting first the elements from 2 and up to but not including 6. We then ask for the elements from 1 through to 8 in steps of 3.

在上面的例子中，我们索引元素从2开始到6结束（但不包括第6个）。然后我们以3的步长索引从1到8的元素。

The same notation can be used with arrays of more dimensions. Let us see an example:

相同的表示法可以与多维度的数组一起使用。 让我们看一个例子：

（Page111 图）

（The same applies to arrays of more than 1 dimension.）

（这同样适用于超过1维的数组。）

（With the shape command we can see the dimensions of matrices and arrays.）

（使用shape命令，我们可以看到矩阵和数组的维度。）

The array b above is a 2 × 4 array as can be seen from the shape command. We can select all the elements in row zero as follows:

用shape命令可以看出，上面的阵列b是2×4阵列。 我们可以选择第0行中的所有元素，如下所示：

（Page111 图2）

（We are using the colon notation to slice the array. :1 refers to the zero-th row, whereas : indicates all columns.）

（我们使用冒号表示法对数组进行切片。 ：1表示第0行，而：表示所有列。）

So far so good, but until now the arrays, matrices and vectors that we have been dealing with have been numerical. However, in many situations the data that we have to deal with is not necessarily all numbers of a single type. There is therefore a need to find a way to accommodate the manipulation of disparate data types, including categorical and text data. In cases like that the capabilities of NumPy are restricted, nonetheless Python can still help as we shall see in the next section.

到目前为止一切顺利，但我们所处理的数组，矩阵和向量都是数值。 其实，在许多情况下，我们必须处理的数据不一定是单一的数字类型。 因此，需要找到一种方法来适应不同数据类型的操作，如分类和文本数据。 这样NumPy模块的功能会受到限制，尽管如此，Python仍然可以提供帮助，我们将在下一节中看到如何处理。

（The arrays we have discussed so far have all been numerical. In many cases we need to deal with different data types and Python can still help.）

（到目前为止我们讨论过的数组都是数值的。 在许多情况下，我们需要处理不同的数据类型，Python仍然可以提供帮助。）

## 2.5 Pandas to the Rescue

## 2.5 “熊猫”前来救驾

You may be thinking that we have lost the plot and that in the style of the Monty Python troupe we are simply listing animal names as part of a sketch. You would be wrong, as Pandas is actually a powerful library that enables Python to work with structured datasets using panel data or dataframes. Pandas4 started life as a project by Wes McKinney in 2008 with the aim of enabling Python to be a more practical statistical computing environment.

你可能会认为我们已经在Monty Python剧团中神经错乱了，从而将动物名称作为章节的一部分。 可惜你错了，因为此熊猫非彼熊猫。Pandas实际上是一个非常强大的类库，它使Python能够使用面板数据或数据帧来处理结构化数据集。 Pandas库于2008年由Wes McKinney开始作为立项，旨在使Python成为一个更实用的统计数据分析工具（译者注：起初用于金融行业）

（4 McKinney, W. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O’Reilly Media）

（4 McKinney，W。（2012）。 用于数据分析的Python：如何使用Pandas，NumPy和IPython分析数据。 O'Reilly Media）

Pandas is a great addition to the Python stack: It allows us to manipulate indexed structured data with many variables, including work with time series, missing values and multiple datasets. In Pandas, a 1D array is called a series, whereas dataframes are collections of series. The rich assortment of data types that can be held by a dataframe, together with the manipulations that it enables, makes Pandas an indispensable tool for the jackalope data scientist.

Pandas是Python技术栈的一个很好的补充：它允许我们使用各种变量操作索引结构化数据，包括使用时间序列，缺失值和多个数据集。 在Pandas中，一维数组称为Series，而数据框（DataFrame）则是多个Series的容器集合。 数据框（DataFrame）可以保存的各种各样的数据类型并对其进行操作，这使得Python的Pandas库成为数据科学家(jackalope)不可或缺的工具之一。

（Pandas is a powerful library that enables us to carry out complex data manipulation in a very straightforward manner.）

（Pandas是一个功能强大的类库，使我们能够以非常直接的方式执行复杂的数据操作。）

In some sense, we can think of a Pandas series as an extension of a NumPy array, and indeed we can use them to initialise a series:

在某种意义上，我们可以将Pandas序列视为NumPy数组的扩展，实际上我们可以使用它来初始化一个序列：

Page112 图

（A typical alias for the Pandas library is pd.）

（Pandas库的常用别名是pd。）

We could also have used a list or a tuple for initialisation. A very useful feature of Pandas is the ability of using indices and column names to refer to data. Let us consider the data shown in Table 2.4 for some animals detailing their number of limbs and dietary habits:

我们也可以使用列表或元组进行初始化。 Pandas的一个非常有用的功能是使用索引和列名称来引用数据的能力。 让我们看看表2.4中显示的数据，了解一些动物关于其肢体数量和饮食习惯的详细说明：

Page113 图

Table 2.4: Sample tabular data to be loaded into a Pandas dataframe.

表2.4：要加载到Pandas数据框（dataframe）中的表格数据示例。

We can load this data into Python by creating lists with the appropriate information about the two features describing the animals in the table.

我们可以通过创建列表来将这些数据加载到Python中，这些列表包含有关描述表中动物的两个特征的相应信息。

Page113 图2

（We can load data into a Pandas dataframe with lists, dictionaries, arrays, tuples, etc.）

（我们可以将数据加载到带有列表，字典，数组，元组等类型的Pandas数据框中）

Note that we have defined the features limbs and herbivore from Table 2.4 as a dictionary, where the keys will be the names of the columns in our Pandas dataframe, and the values correspond to the entries in the table. Similarly, we are defining a list called animals that will be used as an index to identify each of the rows in the table.

请注意，我们已将表2.4中的肢体数和食草动物的特征定义为字典，其中键值是我们的Pandas数据帧中列的名称，并且值对应于表中的条目。 类似地，我们定义了一个名为animals的列表，它将用作标识表中每一行的索引。

We can have a look at the first three entries in the dataframe df with the command head:

我们可以使用句柄df查看数据帧中的前三个条目：

Page114 图

(The head method lets us see the first few rows of a dataframe. Similarly, tail will show the last few rows.)

（head方法让我们看到数据帧的前几行。同样，tail会显示最后几行。）

As we mentioned above, we can refer to the column data by the name given to the column. For instance, we can retrieve the data about the number of limbs of rows 2 through to 4 using the following command:

如上所述，我们可以通过列的名称来引用列数据。 例如，我们可以使用以下命令检索有关第2行到第4行的肢体数的数据：

Page114 图2

(We can view the contents of a dataframe column by name, and the data can be sliced with the usual colon notation.)

（我们可以按名称查看数据框列的内容，并且可以使用python的冒号表示法对数据进行切片。）

Notice that we have referred to the name of the column as a string. Furthermore, we have use slicing to select the data required. Similarly, the information about a single row can be obtained by locating the correct index:

请注意，我们已将列的名称定义为字符串。 此外，我们使用切片来选择所需的数据。 同样，可以通过查找正确的索引来获取有关单行的信息：

Page114 图3

(The content of a row can be retrieved with the .loc method.)

（可以使用.loc方法检索行的内容。）

There is a number of very useful commands in Pandas that facilitate various tasks to understand the contents in a dataframe. For example, we can get a description of the various columns. If the data is numeric, the describe method will give us some basic descriptive statistics such as the count, mean, standard deviation, etc:

Pandas中有许多非常有用的命令可以帮助各种我们理解数据框中的内容。 例如，我们可以获得各种列的描述。 如果数据是数字，describe方法将给我们一些基本的描述性统计，如计数，平均值，标准差等：

Page115图

(The describe method provides us with descriptive statistics of numerical data.)

（describe方法为我们提供了数值数据的描述性统计。）

Whereas if the data is categorical it provides a count, the number of unique entries, the top category, etc.

并且，如果数据是分类的，则它提供计数，唯一条目的数量，最高类别等。

Page115图2

(We can also obtain useful information of categorical data with describe.)

（我们还可以通过描述获得分类数据的有用信息。）

It is very easy to add new columns to a dataframe. For example, we can add a class to our data above as follows:

将新列添加到数据框非常容易。 例如，我们可以在上面的数据中添加一个类，如下所示：

Page116图

（Adding columns to a Pandas dataframe is very easy.）

（将列添加到Pandas数据帧非常简单。）

Pandas also allows us, among other things, to create groups and aggregations:

除其他外，Pandas还允许我们创建组和聚合：

Page116图2

（Pandas allows us to group data and create aggregations. The method .groups contains the grouped information, .size returns a simple count.）

（Pandas允许我们对数据进行分组并创建聚合。 方法.groups包含分组信息，.size返回简单计数。）

We can also apply aggregation functions. Let us try to calculate the average number of limbs for herbivores and carnivores in our dataset:

我们也可以应用聚合函数（aggregation functions）。 让我们尝试在我们的数据集中计算食草动物和食肉动物的平均肢数：

Page116图3

In this case we are applying the mean function from NumPy to calculate the average per group in our dataset.

在这种情况下，我们应用NumPy中的均值函数(mean)来计算数据集中每组的平均值。

In the example above we used Python itself to input data into a Pandas dataframe. Although this is possible for a small dataset, in reality you may be interested to ingest data from other sources. Fortunately Pandas has a very robust input/output ecosystem and is able to take data from a myriad of sources. Table 2.5 lists some of them:

在上面的例子中，我们使用Python将数据逐行输入到Pandas数据帧中。 虽然这对于小型数据集是可行的，但实际上你可能有兴趣从其他来源摄取数据。 幸运的是，Pandas拥有非常强大的输入/输出生态系统，能够从众多来源获取数据。 表2.5列出了其中一些：

Page117图

（We can import data into a Pandas dataframe from a variety of sources.）

（Table 2.5: Some of the input sources available to Pandas.）

（我们可以从各种来源将数据导入到Pandas数据框中。）

（表2.5：Pandas可用的一些输入源。）

Pandas is a very versatile and rich tool and we have only touched the surface in this brief discussion. We will be using Pandas extensively in the rest of the book and whenever possible we will provide explanations to aid the discussion. Nonetheless, we urge you to take a deeper look into this great library.

Pandas，是一种函数多样且功能丰富的工具，我们在上述简短的讨论中只触及了皮毛而已。 我们将在本书的其余部分广泛使用Pandas，并且只要有可能，我们将提供解释以帮助大家讨论。尽管如此，我们强烈建议你深入了解这个伟大的类库。

（Pandas is a very versatile library and we will continue using it in the rest of the book.）

（Pandas是一个非常多才多艺的类库，我们将在本书的其余部分继续使用它。）

*2.6 Plotting and Visualising: Matplotlib*

*2.6绘图和可视化：Matplotlib*

They say that a picture is worth a thousand words and data visualisation makes the case quite emphatically.

古人云一图胜千言，数据可视化对案例来说非常重要。

There are a number of tools that enable data visualisation in the context of business intelligence such as Tableau and QlikView or Cognos. In Python, there are some really good modules that support very nice visuals such as Seaborn, or interactivity such as Bokeh. For our purposes we will concentrate on the robustness provided by matplotlib and its Matlab-style API called pylab.

有许多工具可以在商业智能环境中实现数据可视化，例如Tableau和QlikView或Cognos。 在Python中，有一些非常好的模块支持非常炫丽的视觉效果，如Seaborn，或者像Bokeh那样具有交互性。 本书中，我们将专注于matplotlib及其称为pylab的Matlab风格的API，因为他们很稳定。

（Python is able to create plots and graphs. Here we will cover some of matplotlib’s functionality.）

（Python能够创建二维线图（plot）和绘制其他各种图形。 这里我们将介绍一些matplotlib的功能。）

In a good pythonic style, matplotlib is an object oriented plotting library that can generate a variety of visualisations: From simple plots, histograms, bar charts, scatterplots and more with a few lines of code. If you are familiar with Matlab or Octave, you will find pylab very easy to use. Let us start by importing the modules:

matplotlib是一种具有良好的pythonic风格，并面向对象的绘图库，可以生成各种可视化图：如简单的图，直方图，条形图，散点图等等，只需几行代码。 如果您熟悉Matlab或Octave，您会发现pylab非常易于使用。 让我们从导入模块开始：

（Matplotlib is an object oriented plotting library. PyLab is a Matlab and Octave inspired API for matplotlib.）

Matplotlib是一个面向对象的绘图库，PyLab是受早年的Matlab和Octave启发，调用Matplotlib的API方式

Page118图

（In an iPython/Jupyter notebook you can use the magic command %pylab inline to load NumPy and matplotlib.）

（在iPython / Jupyter notebook中，您可以使用魔术命令％pylab内联来加载NumPy和matplotlib。）

Let us create a simple figure to plot the following functions:

让我们下例函数来绘图

Page118图2

With the aid NumPy we can create a vector with entries for x and calculate y1 and y2:

借助NumPy，我们可以创建一个带x条目的向量，并计算y1和y2：

Page118图3

（The command linspace lets us create an equally spaced vector with a specified number of points.）

（命令linspace允许我们创建一个具有指定点数的等间距矢量。）

We can create a plot using the plot command as follows:

我们可以使用plot命令创建一个绘图，如下所示：

Page119图

（The commands to create a plot are very similar to those in programming languages such as Octave or Matlab.）

（创建绘图的命令与Octave或Matlab等编程语言中的命令非常相似。）

Remember that matplotlib is an object oriented library and thus we are using objects to create our plots. The commands above are very similar to those used in Matlab and Octave and should you need to take a closer look at the syntax you can consult other resources5. The result of the commands above can be seen in Figure 2.1. Finally, it is possible to save the plot to a file with a single command. In this case we can create a PNG file with the following line of code:

请记住，matplotlib是一个面向对象的库，因此我们使用对象来创建我们的绘图。 上面的命令与Matlab和Octave中使用的命令非常相似，如需要仔细查看可以参考其他资源的语法5。 上面命令的结果如图2.1所示。 最后，可以使用单个命令将绘图保存到文件中。 在这种情况下，我们可以使用以下代码行创建PNG文件：

（5 Rogel-Salazar, J. (2014). Essential MATLAB and Octave. Taylor & Francis）

（5 Rogel-Salazar，J。（2014）。 必不可少的MATLAB和Octave。 泰勒和弗朗西斯）

Page119图2

## 2.7 Summary

## 2.7 小结

In this chapter we have covered some of the most important aspects of programming with Python. We started by looking at some of the advantages of using Python in the data science and analytics workflow as well as covering some of the general pythonista programming style used by Python programmers.

在本章中，我们介绍了使用Python编程的一些最重要的方面。首先是在数据科学和分析工作流程中使用Python的一些优点，以及Python程序员使用的一些通用pythonista编程风格。Page120图

（Figure 2.1: A plot generated by matplotlib.）

（图2.1：matplotlib生成的图。）

We have seen how Python can be used as a scripting language as well interactively in a shell such as iPython. We can also use an enriched ecosystem with the help of the iPython/Jupyter notebook. We also covered the different types that are supported by the language: Numbers, strings, complex numbers, lists, tuples, dictionaries. Similarly, we saw how Python deals with mutable and immutable objects.

我们已经看到Python如何在诸如iPython之类的shell中以交互方式用作脚本语言。 我们还可以在iPython / Jupyter notebook的帮助下使用丰富的生态系统。 我们还介绍了语言支持的不同类型：数字，字符串，复数，列表，元组，词典。 同样，我们看到Python如何处理可变和不可变对象。

We can direct the way in which a programme will execute instructions with the help of control flow structures. Furthermore, Python expands its functionality with a number of modules and packages that can be readily imported and used. In this chapter we covered some modules such as NumPy and SciPy. With the help of the powerful Pandas library we can carry out data analysis and manipulation in a very straightforward manner.

我们可以在控制流结构的帮助下指导程序执行指令的方式。 此外，Python还通过许多可以轻松导入和使用的模块和包（类库）扩展其功能。 在本章中，我们介绍了一些模块，如NumPy和SciPy。 在功能强大的Pandas库的帮助下，我们可以非常直接的方式进行数据分析和操作。

Finally, matplotlib is a module that enables us to create plots and visualisations as part of the analysis we perform. These are by no means the only useful modules available to you as a Python programmer, but they are some of the ones we will use throughout of the rest of the book. We will use other modules we have not covered here and we will explicitly mention this in the appropriate sections.

最后，matplotlib是一个模块，它使我们能够创建绘图并可视化，作为我们执行的分析的一部分。 这些并不是作为Python程序员可用的唯一有用的模块，但它们是我们将在本书其余部分使用的一些模块。 如需使用我们未在此处介绍的其他模块，相应的章节中会有明确提及。

In the next chapter we will cover important concepts from machine learning and pattern recognition that will provide us with the context in which data science and analytics operate. Within that we will present another useful Python library: Scikit-learn.

在下一章中，我们将介绍机器学习和模式识别的重要概念，这些概念是我们进行数据科学和分析运作的基础。 并且，我们将展示另一个有用的Python库：Scikit-learn。

Page122

# **第2章 翻译过程中的注意事项**

## 2.1 合格的译者需要具备哪些基本条件

如果满足一下条件即为一名合格的译者：

1. 认可翻译工作对自身的价值；
2. 具备认真负责的态度，而且至始至终都必须保持；
3. 具备良好的技术功底（信），对书中的技术非常熟悉，有相关工作经验；
4. 具备较强的外语能力（达），能准确理解原文的内容；
5. 具备较强的汉语表达能力（雅），能将原书内容以通俗易懂的方式表达出来。

## 2.4 译稿的文件格式及版式要求

### 2.4.1 译稿的文件格式

因为出版社目前主要使用Microsoft Office Word进行排版，所以译稿的文件格式应该为doc或docx。

### 2.4.2 译稿的版式要求

为了便于编辑加工处理，译者提交的定稿应该在版式的如下几个方面满足相关的要求。

1. 标题

◼ 标题一般三级，最多4级。如“第1章（用阿拉伯数字）”为一级标题、“1.1”为二级标题、“1.1.1”为三级标题，四级标题采用阿拉伯数字（1、2、3，···）表示。四级标题不会在目录中出现，各级标题的字体、字号及样式如下所示。