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. Data? Science? Data Science!

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?

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*

*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.

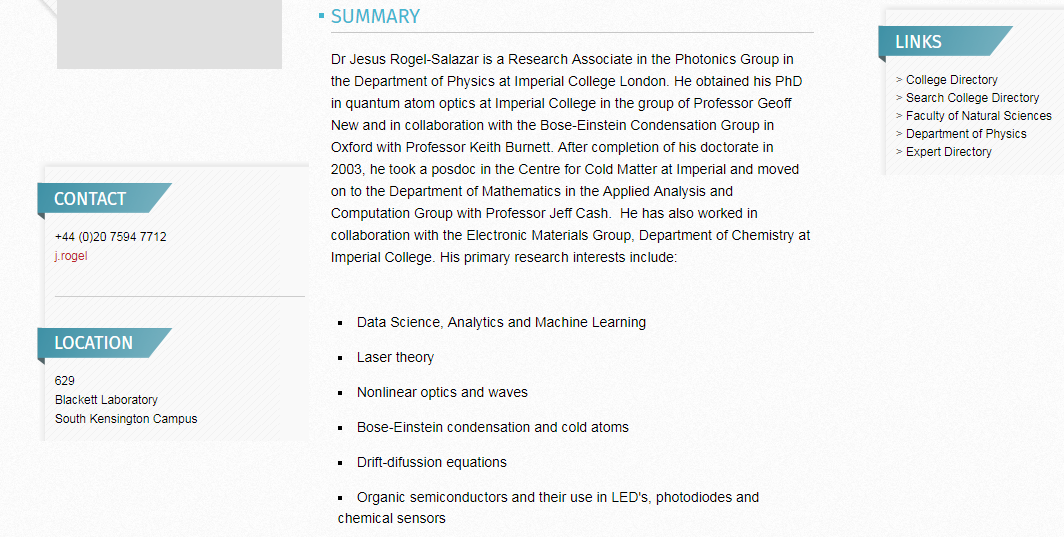
从这个意义上说，数据科学家的角色不仅仅是收集和报告数据；它必须包括从多个有利的角度看待业务应用程序或流程，确定主要问题和后续行动，以及如何使用手头数据的最佳实践。

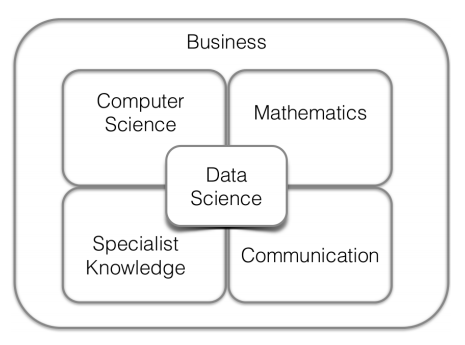
P44

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。)

(1.2.1 Characteristics of a Data Scientist and a Data Science Team)

(1.2.1数据科学家和数据科学团队的特点)

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

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 the latest developments in the field of analytics and data science, and are able to extract actionable insights from our data.

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:

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

• 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

(Some considerations when choosing appropriate tools.)

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.

(Security of the data is very important.)

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).

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

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

• 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)

• 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

(Streaming data collection technologies)

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

(Job scheduling technologies)

• 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

(Big data query languages)

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

(Data stores)

1.3.1 Open Source Tools

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:

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

• 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

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 organisations 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.

(Not all data science is about big data.)

R is seen as the programming language for statistical computing. It is not characterised 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.

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.

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

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