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.

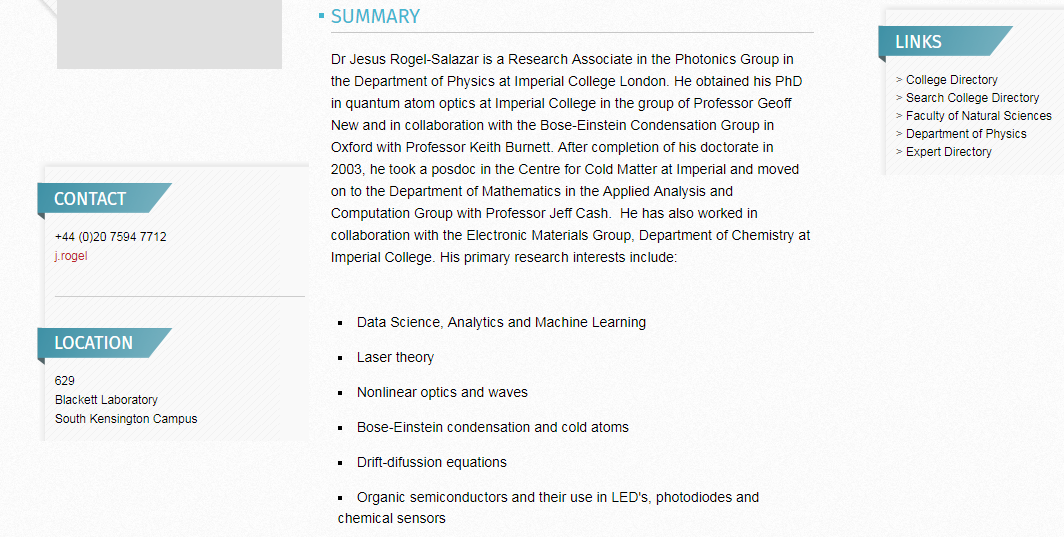
从这个意义上说，数据科学家的角色不仅仅是收集和报告数据；它必须包括从多个有利的角度看待业务应用程序或流程，确定主要问题和后续行动，以及如何使用手头数据的最佳实践。

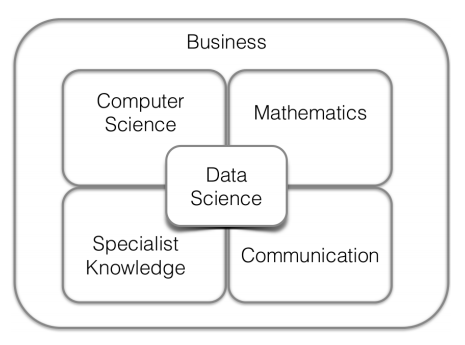
P44

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。