# 第一章 数据科学引论

我们的生活中都充满了数据。不管是否意识到，我们每时每刻都在产生数据。比如，早上起床时，看一眼运动手环上昨晚的睡眠时间，同时看一下这一周的睡眠趋势图。这个简单的过程就包括了数据收集，数据整理和可视化。如果你的健康设备，结合你的睡眠及身体机能指标，给你打分并提出健康建议，那就还涉及到模型建立和预测。随着科技发展，移动终端使得我们的数据收集能力得到极大提升，我们的出行轨迹、身体状况、消费活动、音乐喜好，社交网络等等都被记录下来，并且存储在日益庞大的存储系统中。这样的海量数据已经超过了传统统计学和计算机科学的处理能力，为我们带来了巨大的发展机遇与挑战。商业领域和研究领域对于数据分析有了更高的需求和期待。数据科学作为一门融合了统计学、计算机编程和专业知识的新兴学科应运而生。

## 数据科学的诞生

早在1962年，统计学家John W. Tukey 在题为《数据分析的未来》的 报告中提出数据分析以及相关的统计部分，拥有科学而非数学的特点。数据分析本质上是一种实证科学。 这说明数据分析不仅仅是简单计算，更包含了从数据中提取信息进而解决问题的过程。

1974， Peter Naur 在《计算机方法简述》 （Concise Survey of Computer Methods） 中使用了数据科学（data science）这一术语，用来表示处理数据的科学，同时他认为数据所代表的意义应该归属于其他学科。

1996 ，国际分类协会International Federation of Classification Societies (IFCS) 在其第五次会议论文集标题(“Data science, classification, and related methods”)中使用了数据科学这一术语。

进入21世纪之后，数据科学逐渐兴起，并且慢慢走向公众视野。Cleveland （2001） 在文章中分析传统的统计学家缺乏利用计算机计算处理大规模数据的能力，而计算机科学家缺乏数据处理的统计知识。这两者造成的数据处理能力的局限。 他提出在融合统计学，计算机科学和数据挖掘的基础上建立数据科学学科。 随后，Data Science Journal （2002年创立）和Journal of Data Science（2003年创立）数据科学的两本杂志先后创立，使得数据科学家有了专属的交流平台，数据科学作为一个学科也渐渐地发展起来。在学术圈之外，在工业界，数据科学成为炙手可热的职业。比如，在商业领域，可以通过消费者数据，优化服务，提高利润；在公共卫生领域，通过人口数据和居民身体健康数据，优化资源分配，提供个性化健康服务。

数据科学作为一个新兴的名词，虽然炙手可热，但是它的定义却很宽泛。

Data science is the art and science of acquiring knowledge through data…… Data science is all about how we take data, use it to acquire knowledge, and then use that knowledge to do the following: Make decisions; Predict the future; Understand the past/present; Create new industries/products (Ozdemir, 2016)

数据科学是从数据中获取知识的艺术和科学……数据科学关于我们如何利用数据获取知识，并且用取得的知识来做出决策，预测未来，理解过去和现在，创造新的行业和产品。

那么，要成为一个数据科学分析师需要具备哪些技能呢？

## 数据科学家的能力结构

大数据时代，我们面临一系列的数据处理的挑战。数据量呈现指数增长，同时数据的种类也更加丰富多元。数据不仅包括数值型数据，还包括文本和语音，以及图片和影响。如何整理清洗数据，使之可以满足建立统计模型或者机器学习模型的要求？如何科学系统地探索数据特征？如何规范数据处理的整个流程使其能更加科学、透明、可复制、可迭代更新？在应对大数据所带来的挑战过程中，在利用数据解决各种实际问题的过程中，数据科学应运而生。

数据科学最早是为了弥补统计学家不善于使用计算机处理数据而计算机科学家不善于统计建模解决问题而产生的。最早一批数据科学家来自于统计学、机器学习、计算机科学。

清洗、整理、转化数据需要计算机编程能力。在实际的数据分析过程中，常常需要面对不同的数据类型，结构等。统计学家所期望的干净标准的能够直接被用于统计建模数据几乎是不存在的。原始数据充满着各种错漏、形式上的不规则，而将这些数据进行清洗整理，使其可以满足统计模型、机器学习模型的要求，是数据科学家的一个重要职责。面对大量数据，自动化数据整理，清洗和转化是必然要求，编程能力是自动化、科学性的保证。

挖掘数据的信息需要统计建模以及机器学习建模的能力。数据科学家应该具备一定的统计建模和机器学习建模知识。建立计算模型预测决策是极大的提高了人类处理数据的能力，是数据科学最重要的任务。

在某一个具体的场景下，解读模型的输出结果，指定决策，必须结合领域知识。数据科学家应当具备一定的相关领域的领域知识。模型输出的结果需要结合具体的情境结合一定的理论知识来进行判断。缺乏领域知识而简单的使用统计模型或者是机器学习模型仅仅基于数据进行得出的结论是危险的。

因此，一个数据科学家还需要具备相关项目的领域知识。由此可见，数据科学天然是一个跨学科领域。一个数据科学家应当具备三方面的能力，如图1.1。

值得注意的是在传统的社会科学研究中，我们通常把领域知识（即相关的学科知识）和统计建模相结合来解读数据。由于缺乏一定的计算机编程能力，我们所获取和能够处理的数据在数量和种类上有很大限制。在整个数据处理的流程上，缺乏一定的规范性、标准性、步骤的可重复性。

A picture containing device

Description automatically generated

图1.1数据科学的能力结构(Ozdemir, 2016)

## 数据科学的学习路线图

数据科学家能力的三维结构，可以推断大致有三种不同的学习路线。如果统计学背景人员，需要学习一定的计算机编程（主要用于数据提取、整理），并且在从事某个相关的项目时，需要补充相应的领域知识，或者在团队中需要有相关领域的专家一同参与数据的建模和结论的讨论。同样，计算机科学背景的人士，需要学习统计学知识，同时也需要补充完善基于某一个项目的领域知识或者与该领域的专家进行讨论。

如果是某一专业领域的从业者，希望对自己所在领域的数据进行分析，那么需要掌握基本的计算机编程知识（至少掌握一门常用的数据科学处理语言，例如R语言或者是Python）。同时他还需要掌握一定的统计建模或者机器学习建模的相关知识。当然构建模型时可以求助于相关专业的人事。

本书的目标读者是从事语言研究，主要是实验语言学、量化方式语言研究的研究人员。他们的学术训练背景，已经使他们拥有了丰富的相应主题的语言学知识。他们需要补充的是数据处理中，相关计算机编程技能，以及一定的统计学模型建模知识。因此，本书在后续章节，既提供相应的语言研究的具体场景，也介绍如何实现相应的数据分析过程。本书使用的数据处理语言是R语言。

## 数据科学视域下数据分析工作流程

数据科学和一般数据分析流程重要的区别是，数据科学有一套规范、有序、高效、高度自动化、透明可重复的数据处理流程，可以最大程度上确保结果的准确性和步骤的可重复性。方便团队成员协助，同行合作或核查。数据分析过程不是线性的，而是不断迭代的，特别是数据分析建模的过程，需要不断修正直至达成最优方案。

现有的数据科学工作流程各有不同侧重，图1.2中重点描述了数据处理的部分。

A picture containing shape

Description automatically generated

图1.2 数据科学分析流程(Wickham & Grolemund, 2017)

数据整理（data wrangling）是指将各种类型表格化，使得每一列为一个变量(variable)，每一行为一条记录(observation)。这样表格化数据结构的好处在于，可以轻松地调取各种函数，对变量进行各种计算，进而可以将更多的精力集中在如何回答研究问题。

数据转换（data transformation）基于数据表格，筛选缩小所观察的记录条目，基于现有变量创造新变量（比如基于时间和距离计算速度），基于一些分组条件计算描述性统计量（如按国家分类计算收入平均值和标准差等）。

在数据转换的过程中，我们已经开始对数据进行探索性分析（Exploratory data analysis）。探索性分析是在描述性统计基础上，进一步寻找变量的趋势和变量之间的相关性或其他关系。探索性分析可以产生新的研究思路、研究问题、研究假设。除了数据转化之外，在探索性分析中，我们还可以通过数据可视化，更加直观地了解数据集的各个方面。当我们对于数据的趋中性趋势等描述性统计值有了一定了解，结合我们感兴趣的研究问题，我们可以选择合适统计模型或者机器学习模型。

值得注意的是，数据探索性分析并不是一个线性过程，而是一个循环迭代的过程。数据转化（分组描述性统计等）和数据可视化可以揭示一些我们在开始阶段并没有意识到的研究问题的一些方面。因此，探索性分析有可能会对研究问题产生反作用，促进我们改进研究问题描述，同时更新整个流程。这种循环迭代的性质，从另一方面要求我们数据处理的流程中，对每一个步骤都有详细的记录，每一个步骤都是可重复、自动化。最终，我们要以一定的方式，例如研究报告、会议报告来展示模型和结论。

数据呈现，在数据科学的输出端，如果是商业或者工业领域那么可能会涉及到模型的产品化。通常在研究领域或者学术研究领域，数据科学的输出也就是研究的成果则通常以研究报告、期刊/会议论文、会议海报、会议演讲等形式呈现。传统的论文的方法论部分通常篇幅有限，不能够完全展现数据处理的每一个细节。然而这些细节又会对研究的结果产生非常重大的影响。因此，如何能够充分的保留数据处理的每一个环节，使得可再生性（reproducibility）最大化，成为现代科学和数据科学共同追求的目标。

科学的发展是基于那些可重复，可再生的发现。某一种变量之间的关系能够被更多的科学家独立观测到，那么我们对于这种关系就更加有信心。可重复性（Replication）是发现科学真理的基石，是指通过开展独立的科学实验能够验证某些发现。和可重复性相关的一个概念是可再生性，是指使用原数据集和数据处理方法科学家们能够独立的计算出相同的数据结果。

|  |  |  |
| --- | --- | --- |
| 数据分析流程 |  | R语言包（Package） |
| 数据导入 | readr.png | readr |
| 数据整理 | tidyr.png | tidyr; |
|  | Image | dplyr |
| 数据转化 | Image | dplyr |
| 数据可视化 | ggplot2.png | ggplot2 |
| 数据建模 |  | lmer |
| 数据汇报 | Image | rmarkdown+knitr |

本书的结构

数据科学引论

语言研究中的数据科学

R语言入门

语言实验数据预处理

语料库构建与数据预处理

语音库构建与数据预处理

探索性数据分析与数据可视化

统计数据建模入门

机器学习建模入门

推荐资源

数据科学引论

E-book: Introduction to Data Science: Data Analysis and Prediction Algorithms with R, Rafael A. Irizarry,<https://rafalab.github.io/dsbook/>

E-book: Introduction to Data Science, Tiffany-Anne Timbers, Trevor Campbell, Melissa Lee. <https://ubc-dsci.github.io/introduction-to-datascience/>

语言研究中的数据科学

R语言入门

实验语言学数据导入与预处理

语料库构建与数据预处理

语音库构建与数据预处理

探索性数据分析与数据可视化

Behrens, J. T. (1997). Principles and Procedures of Exploratory Data Analysis. *Pyschological Methods*, *2*(2), 131–160.

FlowingData 博客介绍了如何使用数据分析和可视化来理解数据并且提供很多鲜活的实例。<https://flowingdata.com/about/>

统计数据建模入门

机器学习建模入门

参考文献

Cady, F. (2017). *The data science handbook*. Wiley.

Leek, J. T., & Peng, R. D. (2015). What is the question? *Science*, *347*(6228), 1314–1315. https://doi.org/10.1126/science.aaa6146

Ozdemir, S. (2016). *Principles of data science*. Packt Publishing.

Peng, R. D., & Matsui, E. (2015). *The art of data science*. Leanpub.

Wickham, H., & Grolemund, G. (2017). *R for data science*. O’REILLY.

思考与练习

请阅读下面数据科学的定义，并讨论数据科学的本质。