# 第一章 数据科学简介

我们的生活中充满了数据。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，那就还涉及到数据建模和预测。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，都可以被记录下来，并且存储在日益庞大的存储系统中。

在这样的大数据时代，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、语音、图片和影像。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，数据科学作为一门融合了统计学、计算机科学和专业知识的新兴交叉学科应运而生。

## 数据科学发展简史

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统计，不单纯是数学问题，具有科学性。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。1996，国际分类协会International Federation of Classification Societies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，传统统计学家缺乏利用计算机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）的两本数据科学杂志先后创立，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。Ozdemir(2016)认为数据科学是从数据中获取知识的技艺和科学……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、创新行业和产品。

## 数据科学的工作流程

数据科学有一套规范、有序、高效、高度自动化、透明可重复的工作流程，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。

语言研究涉及的数据类型十分丰富，几乎包含了所有的数据处理操作。语言可以包括口语和书面语。语言数据既包括文本数据（含口语转录的文本），也包括语音数据。此外，语言使用者的社会人口学数据（如年龄、性别、出生地等）也是语言研究的重要组成部分。在实验语言学或心理语言学研究中，数据还包括语言使用者加工语言的行为数据和神经活动数据。其中文本和语音数据属于非结构化数据，需要通过一定的加工来提取变量，比如词频和元音的时长。人口学数据和行为实验数据属于结构化数据，可以用表格形式呈现。其中，每一列为一个变量(variable)，每一行为一条记录(observation)。语言研究数据几乎包含了数据的各种类型。这对于语言研究者的数据处理能力提出了很高的要求。

Wickham 和 Grolemund, 2017提出了数据科学的工作流程，分为数据输入、数据整理、数据转化、数据可视化、数据建模以及数据展示（图1.2）。

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）是对结构化数据进行的一系列操作，包括对于数据进行清洗（比如删除异常值）、变量筛选、数据合并（合并被试身份背景数据和行为实验数据）、数据塑形（长宽数据之间的转换）等。数据整理本身不改变数据的取值，也不对数据进行提炼，而是为了获得合适的数据形式。

数据整理完后，我们就可以进行探索性数据分析（Exploratory data analysis）。探索性分析包括数据转换（data transformation）、数据可视化和数据建模，从而寻找变量的趋势和变量之间的关系。探索性分析可以产生新的研究思路、研究问题、研究假设。



数据转换（data transformation）包括两类操作：1）基于现有变量创造新变量（比如基于时间和距离计算速度，或者对数据进行标准化）；2）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）。当我们通过数据转换和数据可视化，对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。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。在数据科学中，我们可以使用R-markdown文件来将数据分析的过程和研究内容有机结合。越来越多的科学家也开始在开放科学平台公开自己研究的数据分析代码。

## 语言数据科学家

语言研究者要想完成语言数据科学家的转变，大体上应在三个能力模块进行学习：数学/统计学，专业知识，计算机编程。

A picture containing device

Description automatically generated

图1.1数据科学的能力结构(Ozdemir, 2016)

首先，数据科学家需要计算机编程能力。使用计算机编程进行数据处理有两个主要优势：效率高和可重复性强。对规模语言数据进行数据加工，必然要求数据处理的效率高。在实际的数据分析过程中，常常需要面对不同的数据类型和数据结构。统计学家所期望的干净的、标准的、能够直接被用于统计建模的数据，几乎是不存在的。原始数据充满着各种错漏和形式上的不规则，而将这些数据进行自动化的清洗整理，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。

可复制性（Replication）是人们熟知的评价科学理论与假设真理性的圭臬。如果某种理论假设，能够被不同的科学家团队通过同样的方法，独立的科学实验来验证（即复制），那么这种假设的真理性就很强。可复制的科学研究结果推动了自然科学发展。然而在语言实证研究领域，完全复制前人的研究结果几乎是不可能，但是语言研究者可以做到的是数据处理过程的可重复性（reproducibility）。从某种程度上来说，可重复性可以看成一个连续体（见图3），最高级别的可重复是数据处理过程和结果的完全复制（Peng 2011）。Roettger等（2019）认为可重复性的最低标准是其他研究者通过使用原数据集和原数据处理（代码）方法，能够独立的计算出相同的数据结果。传统发表过程中仅在研究方法与结论部分描述数据处理过程的做法，可重复性较低。我们认为，随着开放科学运动的推广与发展，在未来社会科学领域（包括语言研究）的数据分析实践中，国际期刊和研究机构会大量推广公开的数据处理代码和相应数据的做法，从而提高研究的可重复性、透明度和影响力。

Shape

Description automatically generated with medium confidence

图3科学研究可重复性连续统（Peng, 2011）

其次，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数据科学最重要的任务。

最后，需要具有语言学领域知识。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。语言数据科学家应当具备一定的语言学相关领域的领域知识，或者具备与领域专家沟通交流的能力。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机械地使用统计模型或者是机器学习模型，从而得出结论是危险的。

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。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），并且在从事语言研究项目时，需要补充相应语言学领域知识，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语言学领域知识或者与该领域的专家进行讨论。

如果是传统的语言研究者，希望对自己所在领域的语言数据进行分析，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人员。

本书的目标读者是从事语言实证研究的研究人员。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。他们需要补充的是计算机编程技能，以及一定的统计学和机器学习建模知识。语言研究者在数据处理方面的优势，最主要的体现在对于研究领域的知识背景熟悉。这对于数据预处理中的一些决策，如判断缺失值，偏异值以及采取相应的措施是十分重要的。此外，对于研究领域的背景知识的了解，对于探索性数据分析和数据可视化的思路上有很大的帮助。如果将语言研究中的数据处理任务外包给统计学或者计算机科学背景的专业人士，他们缺乏对于研究领域的知识，会阻碍数据的处理，探索性分析和建模。因此，新时代的语言研究者具备一定的数据科学知识与技能是十分重要的。

数学和统计学的能力是语言研究者，特别是文科背景的语言研究者担忧的方面。然而，数据科学是一个应用性科学，不同类型的数据处理对于统计能力的要求不同，并不一定都要求十分深厚的数学背景。比如，文本处理和一些语料库的基本数据统计，要求理解描述性统计的基本知识，结合不同的公式，理解其中区别就可以了。再例如一些推断性统计检验，需要理解一些基本的数据分布和概率假设。这些内容理解上并不难，具体的计算则可以通过计算机完成，并不需要非常强的人工计算能力。

编程就是用编程语言将自己需要做的事转化为指令传递给计算机。大部分文科背景的研究者都有学习外语的经历，电脑编程与之类似，需要学习编程语言的函数（相当于词汇）和语法格式。和学习自然语言一样，这是一个过程，并且其中会伴随一些困难，但这些都是暂时的，笔者认为在未来的十年里，在量化模式的语言研究中，通过编程的方式处理语言和语言应用中所产生的数据是大势所趋。

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/>

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/>

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

数据科学诞生的原因是什么？

如果你想要成为数据科学家，需要哪些方面的训练？