1. 数据科学引论

1 什么是数据科学？

我们的生活中都充满了数据。不管你是否意识到，我们每时每刻都在产生数据。从出行轨迹、身体状况、消费活动到音乐喜好，社交网络等等。在2000年以前，诸如交通轨迹，身体状况，消费活动此类的数据，既无法获得，也无法高效处理计算并应用到生产生活中去。然而随着科技发展，我们获取数据的能力在急速增加，这些数据为我们带来了巨大的发展机遇与挑战。如果能够安全高效的运用数据，我们可以在许多领域实现一些传统方法无法达到的优化与进步。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。

同时，我们也面临一系列的挑战，数据量呈现指数增长，同时数据的种类也更加丰富多元。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。

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。当然数据科学的应用范围很广，在整个数据科学行业内部也有不同的分工。那么，如果要成为一个数据科学分析师需要具备哪些技能呢？

2. 数据科学家的能力结构

数据科学最早是统计学家、计算机科学家和机器学习专家在使用数据进行建模来做出预测、决策的实践过程中产生的。因此，最早一批数据科学家，来自于统计学、机器学习、计算机科学。因此，计算机编程、统计建模自然是数据科学的核心技能。然而，在某一个具体的场景下，解读模型的输出结果，必须结合领域知识。这并非统计学家、计算机科学家或机器学习专家所长。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，数据科学是一个跨学科的研究工作领域。一个数据科学家应当具备三方面的能力，如图1.1。

数据科学家应该具有一定的计算机编程知识。有一些处理数据的经验的人，不难发现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。

现有的图形界面的数据处理软件并不能够完全高效的完成数据清理、可视化、建模等数据处理流程的每一个环节。比如Excel可以完成一些数据清理和可视化的任务，但是在数据建模上则乏善可陈。而SPSS在数据建模上相对有优势，但是在数据整理和可视化上相对薄弱。另一个问题是，图形界面的数据处理程序缺乏对每一个分析步骤的记录，在结果的可重复性，主要是处理步骤的透明度上存在一定缺陷。

数据科学家应该具备一定的统计建模和计算机建模的书专业知识。建立计算模型预测决策是数据科学最重要的任务。在数据科学的能力中，统计建模是必不可少的。

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。

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)

3. 数据科学工作流程

数据科学和一般的数据分析重要的区别是，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。

同时，数据分析的过程不是线性的，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1.2a中重点描述了数据处理的部分。

![A picture containing clock, drawing

Description automatically generated]()

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

数据整理是指将各种类型的表格化，并且使得每一列是一个变量(variable)，每一行是一条记录(observation)。这样表格化的数据又被称为洁净数据(Tidy data)。这种数据结构的好处在于，你可以轻松的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。

数据转化则包括基于洁净数据表格，筛选缩小所观察的记录条目，基于现有的变量创造新的变量（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）。

其实通过数据转换，我们已经开始对数据进行探索性分析。探索性分析是基于描述性分析的基础上，进一步寻找变量的趋势和变量之间的相关性和其他关系，并用此来产生新的研究思路、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的方式更加直观的了解数据集的各个方面。当我们对于集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，我们可以选择合适的推断性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。通过数据的转化（描述性统计等）和可视化，可能会揭示一些我们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。

![A close up of text on a white background

Description automatically generated]()![A close up of text on a black background

Description automatically generated]()

图1.3 数据科学流程图（左(Cady, 2017), 右(Peng & Matsui, 2015)）

设定研究目的或问题

数据科学对于数据处理的整个过程是由问题或者研究目的驱动的，数据的收集、整理清洗和分析不是漫无目的的尝试而是围绕问题展开的。

Leek 和Peng （2015)指出在科学研究中，常见的问题有六种：描述型问题、探索型问题、推断型问题、预测型问题、因果型问题、机制型问题。

描述型问题是一些概括描述数据特征的问题。这些问题的答案本身并不需要进一步解读，因为答案本身是一种事实，是数据的某些特征。探索新问题，需要通过分析数据来判断变量之间是否存在某种模式、趋势、相互关系。因此这种数据分析也被认为是“假设产出型”的数据分析。推断性问题是对探索型问题中提出的假设进行一种新的阐述。推断型问题需要一种完全不同的数据来回答。这涉及到对于具有代表性的样本进行采样。预测型问题当中，关注的焦点是什么样的变量能够预测你所感兴趣的某一种现象或者行为。不管这两者之间是否有必然的因果关系。在因果型问题中，需要验证的则是某一个变量的改变会导致总体当中的另一个变量也发生改变。回答这样的问题通常需要在数据采集时采用一定的实验设计，比如说我们需要对总体进行随机采样。机制性问题是最为复杂的问题，问题关于某一个变量是如何导致另外的一些变量的变化。除了一些机械工程研究领域回答机制性的问题需要保持谨慎(Leek & Peng, 2015)。

因此，我们要了解自己需要回答的问题是属于哪种类型，进而决定相应的数据处理方法。同时在数据分析的过程中我们又要不断的修订自己的研究问题。

还有两点值得注意。首先，一种数据分析有可能回答多种类型的问题。比如，我们如果要回答一个推断性的问题，那么在数据的分析过程中，我们势必已经回答了一些与之相关的描述型和探索型的问题。其次，能够回答怎样的问题部分取决于可用的数据。例如，如果要回答一个因果性的问题，必须要有相应的实验设计和数据。不然的话，只能得到探索型的答案，也就是两个变量之间的相关性。

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。

可读性编程（Literate programming）

可读性统计编程（literate statistical programming）是基于可读性编程创造的新的一个概念。可读性统计编程啊统计分析过程对描述和进行那些统计分析过程的代码融合在一个文件当中。这样读者既可以了解到分析过程同时也可以通过作者的数据运行代码得到相同的结果。这样最大程度的保证啊研究数据分析的可再生气。

那么保持较高的研究可再生信有什么好处呢?

首先，研究可再生性可以提高研究的科学性。一个研究能否成为人类知识的一个部分，取决于它是否能够被复制验证。可复制性最低的门槛就是可再生性。

其次，可再生性，可以减少重复研究的浪费。研究者可以利用别人现有的研究过程和代码，在此基础上进一步进行研究。而不需要再从头开发或探索数据分析的过程。

对于个人来说研究的可重复性也是有很大好处的。

首先，一个可再生的研究，会让你在脑海当中更加有效的去计划和组织你的研究。因为可再生性涉及到你如何管理数据？如何管理代码？这种对于可再生性的关注，反过来会让你在研究一开始就思考这些问题。当你想要重新利用过去的某个研究或者修改某个研究的时候，你会发现可再生性的价值。

第二，可再生研究可以提高团队协作的效率。因为数据分析的每一个过程都非常透明易懂，那么你的合作者更加容易融入团队，分工协作。你潜在的合作者也更加容易了解你的工作，并且快速的适应。

第三，研究的过程不是线性的。改动是不可避免的。对于模型的调试，数据的增加减少，都会需要你不断的重复某些数据分析的过程。直至最后你的合作者、期刊的编辑、或者文章的外审专家，都会对于你的研究数据处理过程，提出一些意见。那么能够高效快速的重复之前的数据处理过程，会使得你有更多的精力放在研究问题上，而不是重复的机械劳动上。同时你的修改也会更加准确。

第四，可再生的研究，通常能够让别人更加相信。因此，他们会使用你的数据分析中某些步骤或者代码。这样你的研究影响力就增加了。

数据科学的学习路线图

从数据科学家的能力结构上，我们可以推断大致有三种不同的学习路线。如果你的背景是统计学，那么你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。

如果是计算机背景的人士，那么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，例如二语言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。当然构建模型时他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，主要是量化方式的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。本书使用的语言是R语言。

本书的结构

推荐书目

推荐资源

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

数据可视化

<https://flowingdata.com/about/>

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。

Courses, free textbooks

Introduction to Data Science: Data Analysis and Prediction Algorithms with R, *Rafael A. Irizarry*

<https://rafalab.github.io/dsbook/>

*Introduction to Data Science*, Tiffany-Anne Timbers, Trevor Campbell, Melissa Lee

<https://ubc-dsci.github.io/introduction-to-datascience/>

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.