新视野·数据科学

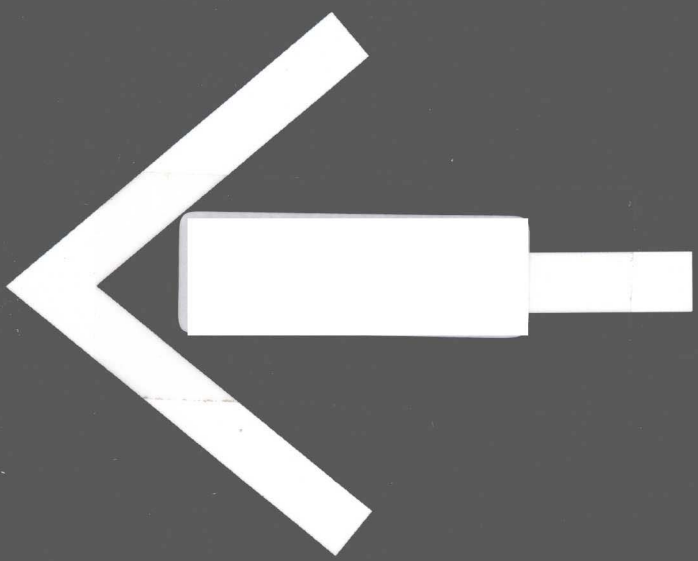
R语言应用系列

数据科学中的R语言

R in Data Science

李 舰 肖 凯 著

吴喜之 审校



西安交通大学出版社
XI'AN JIAOTONG UNIVERSITY PRESS

R语言应用系列

R in Data Science

数据科学中的R语言

李 舰 肖 凯 著
吴喜之 审校

西安交通大学出版社
Xi'an Jiaotong University Press

内容简介

本书是一本R语言实战类书籍，目标群体为缺乏编程或者统计基础，但希望能从零开始深入地理解并应用R语言的读者。囊括了行业应用的真实案例是本书的亮点，涉及从传统的统计分析领域如新药研发、金融分析到当前最热门的大数据、社交网络等应用的例子。作者把从业以来积累的R语言在各行业中的应用案例第一次公开出版奉献给读者。

图书在版编目(CIP)数据

数据科学中的R语言/李舰，肖凯著. —西安：西安交通大学出版社，2015.3
(R语言应用系列)
ISBN 978-7-5605-7082-2

Ⅰ.①数… Ⅱ.①李…②肖… Ⅲ.①程序语言-程序设计 Ⅳ.①TP312

中国版本图书馆CIP数据核字(2015)第028978号

书　　名	数据科学中的R语言
著　　者	李　舰　肖　凯
审　　校	吴喜之
责任编辑	李　颖
责任校对	贺峰涛
出版发行	西安交通大学出版社 (西安市兴庆南路10号　邮政编码710049)
网　　址	http://www.xjtupress.com
电　　话	(029)82668357　82667874(发行中心) (029)82668315(总编办)
传　　真	(029)82668280
印　　刷	陕西宝石兰印务有限责任公司
开　　本	720mm×1000mm　1/16　　**印张** 26
印　　数	0001～3000册　　**字数** 465千字
版次印次	2015年7月第1版　2015年7月第1次印刷
书　　号	ISBN 978-7-5605-7082-2/TP·654
定　　价	79.00元

读者购书、书店添货、如发现印装质量问题，请与本社发行中心联系、调换。
订购热线：(029)82665248　(029)82665249
投稿热线：(029)82665397
读者信箱：banquan1809@126.com

序　言

无论从数据科学的角度，从编程语言的角度，还是从应用的角度，这本书是给读者的一个完全意外的礼物。

这本书如此简明，使用最少的文字清楚明白地传达了大量的信息；这本书的内容如此丰富，鲜有包含这样多资源的涉及 R 语言的文献；这本书的作者站得高，绝不纠缠那些繁琐而非必要的细节，读者可以很容易地看到问题的全貌和整体结构，而这是获得任何知识的关键；这本书的安排对于性急的人非常方便，若干分钟就可以获得通常几天才能获得的信息。

这本书的成功在于作者的经历：统计 — 计算机软件 — IT 相关业界。没有这样的背景，不可能对问题驱动的数据科学如此明白，也不可能对软件要素的理解如此清楚，更不可能对众多应用如此轻松地介绍。这本书的成功还在于作者多年的实际经验。实际经验比众多的文凭、奖状、职位等更重要，这本书的写法和内容体现了作者本身能力和知识增长的历程，在实践中获得的知识和能力远非课堂灌输式教学所能比拟的。作者的轻松幽默的心境为这本书画龙点睛，这来自于他们的智慧，使读者能够在一种令人享受的心情下阅读这本书。

这本书使我得到相当的满足和愉快，相信读者也会有同样的感觉。

吴喜之

2015 年元旦

前　言

僭称科学家我本来是不敢的，不过如今人们对数据的研究和应用的主战场在业界，“数据科学家”通常指的是一个职位的名称。我的部门现在新招的职位都是“Data Scientist”，所以我自称数据科学家应该还好。从我本科进入中国人民大学学习统计学专业开始到现在的 10 多年时间里，我所有的求学经历和职业生涯都在和数据打交道，在数据应用的最前线感受到了业界对于数据价值理解的巨大变化。也亲身经历了从数据被冷遇到如今“大数据”成为显学这一激动人心的变革。这些年的很多经验都化成了这本书中的内容。在这里，我回顾自己在数据科学家道路上的一些经历，用自己的视角来总结这个数据时代的变化，也作为这本书的前言。

我少年时的志向和很多无名的儒生一样，“为天地立心，为生民立命，为往圣继绝学，为万世开太平”，结果也一样，就是越长大越失望、越难有新的目标，对什么事情都不执着，常被推着走。当然也不会否定自己，习惯顺其自然。就这样不小心走上了数据科学家的道路，在这条路上我经历了很多次对知识和技术的被动接受与主动融合。

我们那时高考是先估分再报志愿，很适合我，我对自己的估分很有把握，所以敢于填报一直心仪的人大，最后果然实际得分和估分只差两分，可是离最低录取线也只差两分。于是我被人大录取了，但是专业是我填报的第五志愿，也就是当时还算冷门的统计学，前面四个专业我就不提了，怕被骂黑子。

入学时是 2001 年，刚进入新的世纪。我对统计最初的看法和很多人一样，以为就是算 GDP 和物价指数。后来我在很多报告中都会讲一个段子，“搞计算机的人最烦的就是被叫去修电脑，搞统计最烦的就是一桌人吃完饭后被要求算一下账单是否对”，即使到现在场下都还有不少人笑，说明人们对统计的误解还没完全消失。遥想我自己当年，大一时去系里主办的“资料采矿研讨会”当志愿者，听到谢邦昌老师介绍资料采矿时我还在纳闷“难道台湾也搞采掘业?”后来看到数学公式才明白不是我想的那样。陈希孺老师出来时我和同学都以为可以听懂，因为报告题目是基尼系数，这个词在经济学教材中我们都懂，结果幻灯片出来后没一个人懂，陈老师开口后我比同学多懂了一点点，因为我能听懂一点点

的长沙话。从此以后我知道了统计不是以前想的那样。

大二的时候，吴喜之老师给我们上了统计计算的课，这是我求学阶段最庆幸的事。当时赶上非典，这门课被拆成了两部分。一半的人逃回家了所以暑期回来补课。在家期间我以研究吴老师教的 R 语言为乐，回来后考试得了 100 分，这是我学生生涯中期末考试的唯一满分，当时高兴了好久，完全不会想到现在 R 成为了我工作中的主要工具。我们系大三时就有自己的导师，我跟着吴老师从学年论文写起。算起来我学 R 的起点还是很高的，当年把吴老师那本《非参数统计》中的例子全部用 R 代码实现了一遍，开始尝到了编程的甜头，因为之前觉得像天书般的非参数统计在我写完程序后就觉得什么都懂了，而且对吴老师的引言里的内容有了更深刻的理解，直到今天，我这本书的引言里都还借鉴了好多吴老师的思想。我从当年的程序中选了一个符号检验放在本书附带的 R 包中，在“假设检验”一节中还用到了，这个函数的源代码完全可以当作编程规范的反面例子，但是我丝毫不害怕丢脸，因为即使当年这么弱的水平，也可以写 R 程序，而且有用，直到今天还能用，这就是 R 最大的价值，当然也是数据科学的价值所在。从那时起我就发现了一条学统计的捷径，遇到任何不懂的地方，拿到数据后写程序算个结果再来看书。

大四的时候有门专业课叫做“数据挖掘”，这在当时是个极热门的概念，在世纪之交的时候，数据挖掘被各路专家钦点为新世纪最重要的技术。当时谷歌刚刚上市，国内的数据电子化和企业信息化也差不多成熟了，人们的问题由数据不够变成了数据过多，需要借助搜索引擎和数据挖掘才有可能获取有价值的信息。当时比较流行的说法是“知识爆炸”，各路英雄都瞄准了这个激动人心的方向。当时甚至还有数据挖掘和统计学谁主谁次的争论，也有很多认为两者就是一回事的观点。而当时的我对数据挖掘的理解只是关联规则、聚类、分类、神经网络等具体算法的实现。按照我当时的认知水平，我感觉数据挖掘只是对统计方法的补充以及对大数据量的实现，在思维方式上并没有什么不同。当然，那时我对数据挖掘的理解完全是根据其自行描述的理想状态来判断的。

由于统计专业当时仍然不热门，我考研时报了热门的北大光华，当年有个数理金融的方向，我很看好，可惜结果下来后总分出局，不过成绩不算太差，正赶上那时候软件学院刚开张，所有科目和总分都过线的话就可以调剂一个金融信息化方向的双证，于是我又被推着朝数据科学家的方向近了一步。我很喜欢自己的一个优点就是从不抱怨，为了让之前的努力完全没有白费，我就迅速地找到了自己的专业和软件的结合点，在软件学院苦练数据结构和编程技术，也多学了几种编程语言，曾经也怀疑过这样苦练的意义，因为不可能超过同资质下计算机出身的人，后来在我的导师杭诚方老师教的课程中找到了自信，那是一个 OLAP

的练习，我感觉那些立方体实际上就是 R 中的多维数组，于是自己用 R 写了个 OLAP 的工具，从分析结果来看，不比其他同学用商业软件做的差，从此我开始从自己以前的专业中找到了存在感，也可以以一个更轻松的心态来随意学习自己想要的，而不是跟在编程高手后面追赶。

研究生入学的时候是 2005 年，数据挖掘正如日中天，但业界更喜欢热炒的一个词是“商业智能”，简称 BI。BI 主要是厂商提的概念，按道理应该包含数据仓库和数据挖掘。当时大的企业都有 ERP（企业资源规划）系统，小的企业至少也有 MIS（管理信息系统），这些系统都能采集数据，再加上其他各类应用系统，使得数据的内容过于丰富，快速而直观地发现数据中的规律是企业非常现实的需求。数据仓库的思路是将所有系统中的数据存入一个数据库，但这个数据库的设计范式与业务系统的不同，因为其目的是数据分析而不是操作，所以数据的增删改并不重要，而数据的查询非常重要。这样的数据就是数据仓库，可以从数据层面实现企业内所有数据的整合，同时能够快速地访问所需要的数据。所有的数据仓库都包含 OLAP（联机分析处理）系统，基于数据仓库对数据的各个维度进行展现。维度就是统计中分类变量的概念，在行业中也常被写成“纬度”，这和把“阈值”写成“阀值”的是一个门派的。一般来说，BI 项目的核心就是建数据仓库，而建仓库时最大的工作量是 ETL（数据清洗、抽取、转换、加载等），基本上仓库建好后靠 OLAP 就能解决一般企业绝大多数的分析需求，因此很多时候 BI 都不包含数据挖掘。对厂商来说，以 ETL 为主导的 BI 项目可以用比较便宜的人工，同时也容易复制，因此慢慢地在行业内 BI 这个词就变味了，变成了 OLAP 和报表可视化的代称。我那个时候还没有深入业界，虽然有这样的怀疑，但是不敢真往这方面想。

我的实习和第一份工作是在西门子，一个财务相关的部门。同事大多是会计背景，但是他们用 Excel 的能力让我惊叹。我的工作是分析财务数据，但是实际的内容主要是操作 SAP 然后用 VBA 写自动化报告的程序，工作的过程中我也感受到了 Excel 和 VBA 的强大，最重要的领悟就是任何语言都有可能解决任何的问题。网络上喷来喷去的只是弱点，可能影响到效率，但实际工作中，人们最关注的是能与不能，而不是好与不好。用 Excel 的过程中我解决了同事提出的所有问题，有些和交互协作相关的问题就用 JSP 来写，不过当时公司的服务器上没有 Tomcat，于是自学 ASP 也都解决了。毕业后我留在了西门子，并随公司搬到了上海。

我很庆幸我在西门子的工作经历，可能当时入职时最吸引我的只是五百强的虚荣。但在这样的大企业形成的工作习惯是可以受用一辈子的。虽然效率不是很高，但是任何的工作细节决定了所有的努力不会白费也不会起相反作用，这里

不需要个人英雄主义，只需要所有人的合力。在自己的位置上完成本职工作就是成功。工作的节奏对我这样的急性子来说太慢了，但是慢下来之后和大家的节奏契合之后常常能出一些我之前想象不到的成果，这都是我自行摸索学习不到的东西。

在感到已经没有可学的东西之后，两年过去了，当时已是 2009 年。继续呆在这里只需要深入学习会计和熬资历，一步步升职加薪就能变成真正的外企人，一直成为有用的螺丝钉，我之前的专业和兴趣就要白费了。于是我选择了另一个极端。源略数据是一个当时的创业公司，其理念是融合 IT 和统计，打动了我，一看 Logo 就喜欢，八卦变来的。即使现在我也佩服老板们的远见，当时要搞的就是今天的数据科学。各种类型的项目都做，不限行业不限内容，从满意度调查到 BI，从运筹学到文本挖掘，都是我们的解决方案。在源略的两年是最开心的时间，一群人可以在办公室里搞烧烤，装个卫星锅看世界杯，还一起自驾千里去搞户外，有过这样的经历后现在对创业就没那么向往了。公司在这段时间靠项目过得很不错，但是最终没能迎来大家期望的对数据需求的爆发。可能在很多人看来只是个没成功的理想主义公司，但是这段经历对我来说非常重要。我作为一个资历尚浅的人可以担当很多重要的角色，很多之前的想法在实际项目中一一得到了印证，纯粹地做任何喜欢的事情，以前不确定的地方靠本事也能找到自信。当我想离开的时候，有一种出山的感觉。

在源略数据的两年时光里，数据开始慢慢变得热门。R 语言也开始走入人们的视线，中国 R 语言会议也办起来了。行业里数据的应用仍然是以 BI 为主，但是很多新的应用已经开始兴起。除了具体的技术和工程实践，我开始意识到对数据的理解其实是最重要的能力。纯粹的技术能解决的问题很少，很多时候问题错综复杂，涉及到多个系统之间的复杂关系甚至人与人之间的复杂关系，数据散布其中形成一个又一个难解的结，再前沿的技术也难以成为一把斩断乱麻的刀，只能靠人来抽丝剥茧，然后在不同的阶段和环节选择最有效的或者自己最擅长的工具一个个地解决问题。如果之前没有一个清晰的总体的理解，那么很容易就陷入到局部的死胡同去硬撼各种难题，反之，如果找到了一条正确的道路，就可以用最经济的方式来解决问题。直到今天我都认为这些是数据科学家最重要的技能，实际上也是最容易被忽视的。

下一站是 Mango Solutions，也是此时此刻我的公司。2011 年离开源略后我对自己不再怀疑，开始坚定地向数据应用的巅峰挑战。选择无非只有两条路，读个博士搞学术或者在业界找个更专业的地方搞技术。无论哪种选择，最现实的出路就是找个狭小的领域寻章摘句或者找个狭小的圈子千锤万凿。我选择了后者，因为我信仰数据的价值，但并不执着于方法。在数据应用的领域，学术界和业界

的差别不大，总之数据为王，能更多见到数据的地方就是好地方。Mango 是个专业用 R 的公司，与我的专长非常匹配，更重要的是它可以深入业界去解决一些和数据相关的具体问题，无论大小、无论难易，客户高兴就是最好的度量，这样简单的评价方式是我喜欢的。

在 Mango 一呆就是四年多，已经超过了人大，是我在一个组织内呆的时间最长的了。这四年里，我接触到了欧美很多顶尖的公司和顶尖的人，从他们的项目中学到了很多东西，也帮助他们解决了不少问题，看着自己曾贡献的努力出现到了人们的日常生活中是一种很好的体验，感觉自己的价值得到了实现。这四年的时间也使我从一个青年人变成了中年人，在专业的道路上越走越远，也牺牲了很多原来的兴趣。在这个阶段，我感受到了自己之前所有的技能融会贯通了，统计、编程和沟通能力自不用说，这是基本的技能，即使是会计、市场、销售的能力也感觉很有用武之地。更重要地，我体会到了行业的差异、东西方的差异、文化的差异并没有想象的那么大。能帮到别人，就会是受欢迎的人，能解决难题，就会是令人佩服的人。

这段时间随着互联网行业的成功，“云计算”迅速成了热点。我非常欣赏这种模式，因为“云”是可以对抗传统厂商的绑架的。通过廉价开源的个体聚集成庞大的系统，这就是互联网的精神。但是发起这个概念的人更多的是计算机专家而不是数据的专家，并不是所有分析算法都可以轻松部署到云上的，因此业界的云计算大部分沦为云存储平台。正如之前的数据挖掘变成了关联规则和分类算法、商业智能变成了 OLAP 一样，都是很好的概念被厂商狭义化了。

很快，“大数据”的概念崛起了，迅速占据了最热门的位置，其热度是之前任何时代的热炒概念所不能比拟的。对于大数据，虽然仍然存在很多跟风炒作的，但是不得不承认它确实开创了一个全新的时代。大数据的概念完全是应运而生，因为数据的来源有了翻天覆地的变化，数据的规模完全足够，计算的能力也得到了长足的发展，新的机器学习方法也不断涌现，终于赢来了数据应用的黄金时代。社会上也开始广泛地关注数据的价值和大数据的应用，随后也产生了“数据科学”这一理性的概念。这是所有数据从业者的好时代。

这四年里，数据的价值在国内得到了认可，R 语言也越来越火。工作之余，我和统计之都 [1]一群志同道合的伙伴们也时常探讨数据的价值，也闲聊各类八卦，还组织了规模越来越大的中国 R 语言会议，我们逐渐发现，数据已经融入到了自己的生活和价值观中。理解问题、相信数据、慎用方法、尊重需求，这就是数据科学家的思维方式。数据科学家不是拯救蒙昧的传道者，不是秀智商优越感的“理科生”，不是曲高和寡的“专业人士”，而是真正能用数据来解决问题的

[1]Capital of Statistics，简称 COS，主页是 `cos.name`，是一个旨在推广与应用统计学知识的网站和社区。

实干派。这在本质上与 R 语言是一致的，也是如今大数据时代下这两者越来越火的原因。记得 2012 年北京的 R 语言会议结束之后，郁彬老师给我们作了一次印象深刻的报告，郁老师强调的统计应该跟上现代的节奏、要主动去和计算机结合、要深入到应用领域的观点让我感觉自己做的事情很有意义。

最初有写这本书的想法是在 2012 年上海 R 语言大会时，李颖找到我和肖凯开始谋划一本基于 R 语言与数据实战的原创书。当时肖凯提议起名数据科学时我还从来没听说过这个词，没想到短短两年多的时间后，这个词会变得如此火热。当然，从另一面来看，我们这本书居然写了两年多还没写完。当时我还担心数据科学的书名让人摸不着头脑，不过在读了肖凯写的博客和推荐的链接之后，觉得这个词可以非常精确地描述我们的工作。我们从数据出发，介绍各种方法的原理、在 R 中的实现以及在具体领域中的应用。书中的内容全部来自于我们平时工作中的经验和对 R 语言的感悟，与传统的统计学、R 语言编程或行业实战类书籍都有所不同，命名数据科学是再合适不过了。

感谢中国人民大学的吴喜之老师，从我当年开始学习 R 语言到现在从事专业的数据分析工作，都离不开吴老师悉心的指点，对于本书吴老师也提了很多宝贵的意见，帮助我们改正了不少错误。感谢统计之都的伙伴们，很庆幸有这样一群志趣相投的朋友，大家利用业余时间一同为统计学的普及和应用而努力，平时各类专业问题的讨论和各种各样的八卦是这本书的重要动力和源泉。感谢浙江大学软件学院金融数据分析技术专业 2013 级和 2014 级的全体同学，我在讲授“金融数据分析基础”和“R（语言）及其应用”课程的时候用到了本书中大部分的例子，同学们的参与和反馈为本书的不断完善提供了很大的帮助。

本书作者李舰撰写了第 1 章、第 5 章、第 7 章、第 9 章、第 12 章、第 13 章、第 15 章、第 16 章和第 11.2 节，作者肖凯撰写了第 2 章、第 3 章、第 4 章、第 6 章、第 8 章、第 10 章、第 14 章和第 11.1 节。全书整体风格的统一、语言的润色和文字的校对由两位作者和编辑李颖共同完成。关于本书的意见和建议请联系作者的邮箱 `rinds.book@gmail.com`。书籍的相关资源和勘误请参见 `http://jianl.org/cn/book/rinds.html`。欢迎任何的建议和指正！

李舰

2015 年 3 月 1 日

目　录

第 1 章　引言：数据科学与 R

1.1　数据科学简介

1.1.1 什么是数据科学？

关于数据科学这个词的渊源，可以追溯到很久以前。Wikipedia [1]上目前最早考据到上个世纪 60 年代 Peter Naur 提出了这个概念。郁彬教授认为上个世纪 40 年代 Turner 和 Carver 等人就提出了数据科学的思想[40]。C.F. Jeff Wu 于 1997 年旗帜鲜明地提出了“Statistics = Data Science?”[37]，那个时期差不多正是数据科学逐渐变得广为人知的开端。通常认为，从 2008 年 DJ Patil 和 Jeff Hammerbacher 把他们在 LinkedIn 和 Facebook 的工作职责定义为“数据科学家”的那段时期开始，数据科学开始在业界流行起来。截至目前，数据科学这个词已经炙手可热，在欧美，“Data Scientist”这个职位已经成为招聘市场的宠儿。

然而“数据科学”在国内还没有这么热门，近两年随着“大数据”的风潮而崛起。由于数据科学这个词在欧美业界的流行程度更甚于学术界，但究其内涵，推崇数据领域全栈的解决方案，很容易借助开源软件来摆脱传统厂商的绑架，所以完全不具备被炒作的特性。因此数据科学在国内的讨论也不是非常多。国内很多媒体和网文中的观点并不是很准确，基本上只要和处理数据有关系的技术都可以被称为数据科学。这样的理解不能说有很大的错误，但是非常容易引起混淆。由于本书专门介绍数据科学及其在行业中的实现方法，因此有必要对“数据科学”这个概念进行明确的界定和阐述。

如果要对数据科学下个定义的话，我们认为数据科学是使用科学方法从数据中获取知识的学科。关于“科学的方法”的介绍，占据了本书的大部分内容。需要注意的是，作者有时候会基于易经来算卦，网络上也有不少人不加理解地拿数据套用某些模型、方法或者工具直接得出结论，这些方法都是从数据中得出的结论，但都不是科学的方法，所以自然也不能称之为数据科学。此外，我们强调了“从数据中获取知识”，所以一些基于演绎和推导得出结论的科学方法也不纳入数据科学的范畴，比如如果某些药物或者食品通过长时间科学的试验，我们从

[1] http://en.wikipedia.org/wiki/Data_science

临床试验数据中得出安全性的结论就是数据科学，但如果直接从其中的生物学、化学原理出发来证明其安全性，就不是数据科学。

关于对数据科学的理解，我们通过图1.1中的韦氏图[2] 来描述。

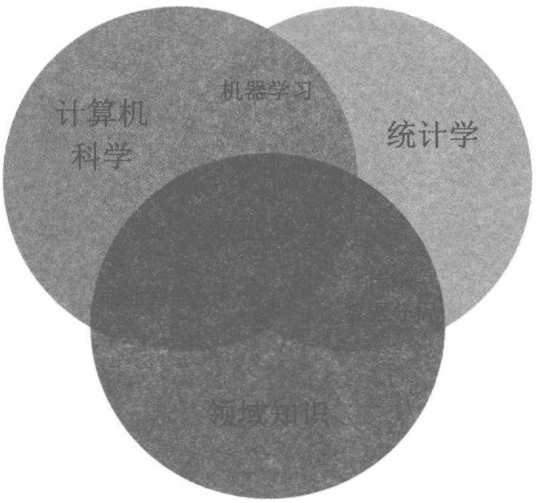


图 1.1 数据科学的韦氏图

这幅图从数据科学涉及到的三个主要领域即统计学、计算机科学、领域知识出发，通过两两结合和三者结合，列出了不同的学科或者行业领域，一共包含七个名词。其中涉及到的一些学科或者领域并不完全与数据有关，所以介绍时只关注其中和数据相关的部分。在这里，作者主要是根据自己的经验进行介绍，可能是一家之言，但是代表了全书的观点。

（1）统计学

统计学是三个领域中唯一一个只和数据相关的学科。可以说统计学是数据科学的核心，因为数据科学的科学性是由统计学体现的。这并不是说其他的领域不是科学，比如计算机科学，当然是科学，但使用计算机的方法进行数据分析的时候并没有专门强调科学性。除了通过简单的分类汇总之类的计算可以得到分析结论以外，复杂的分析常常涉及到推断，关于推断就不能不深入地去研究模型和方法的假设，不能不去透彻地理解数据。

[2] 网络上关于数据科学有一张流传很广的韦氏图：http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram，但是该图遵循 CC Attribution-NonCommercial 协议，无法引用到书籍中。此外，该图与本书作者的观点也有所不同，图1.1是本书作者根据自己的观点使用 R 做的韦氏图，请读者注意其中的区别。关于描述数据科学的韦氏图，类似的还有 IBM 提供的图 https://www.ibm.com/developerworks/jp/opensource/library/os-datascience/。

Efron 说过：“统计是仅有的系统地研究推断的科学”。数据分析最重要的目地是研究过去和推断未来，所有对未知的推断都基于历史数据和各种假设，统计之所以是科学就是在假设数据满足某些前提的条件下通过严格的证明和否定，从而得到科学的结论。如果没有这个过程，直接从数据中调用某个工具或者算法就得到结论的话，很容易就变得不科学或者伪科学。

对于科学来说，一直是在否定中发展的，通常是先提出理论或者假说，然后通过试验来证明或者证伪，每当旧的理论被新的理论推翻，就很有可能实现人类的进步。这就是科学的魅力所在。这也是统计学的思维方式，统计学建模的前提是对真实的世界提出假设，然后通过模型来描述，通过数据来支持或者否定，一旦有新的模型能取代旧的模型，就能更好地分析这个真实的世界[41]。

如果脱离了数据或者真实的世界单纯去研究统计的模型，实质上就成了数学的分支，并不是统计学的目的。数学的方法和数学公式是统计学科学性的保证，但数学的思维方式本质上和统计是不同的。数学的世界是一个理想中完美的世界，大部分的结论都是来自于演绎的思维方式和严密的推导。而统计是在模拟一个真实的世界，其思维方式是基于归纳的，所有的结论都来自于数据，甚至可以说来自于历史，这和人类靠经验做决策是相似的。

在图1.1中，我们并没有提到数学，就是因为在数据分析的过程中，我们把数学当作一种工具或者理论基础，我们直接接触到的方法并不是数学，虽然很多都是数学公式，但数学公式只是载体，我们使用的是它们代表的统计模型或者计算机算法。就好比我们在分析数据时会使用中文对结论进行解释，但我们并不把中文当做分析方法。当然，有些应用数据的分支本身就是分析方法，例如最优化方法，很多统计模型都使用最优化方法进行求解，当然最优化方法也可以直接用来解决很多问题（我们在“第206页: 9 最优化方法”会进行详细的介绍），但此时就成了一种计算机的实现方式。

如果只有统计学而没有图1.1中的另外两个组成部分计算机科学和领域知识，其代表的是一部分传统的统计学者所从事的研究工作。由于统计学是一门应用的学科，虽然在当今的时代下如果离开了计算机能做的事情很少，但也不是没有，可是如果连具体的行业和领域也脱离了，可以说数据都没有了，这部分群体的数目是越来越少的。

（2）计算机科学

计算机科学最主要研究的对象并不是数据。在行业中最重要的目的也不是数据分析。在人类刚进入信息时代时，计算机科学最重要的应用是开发各种软件和应用系统。对于数据分析来说，通过计算机实现了将存储在纸张中的数据电子