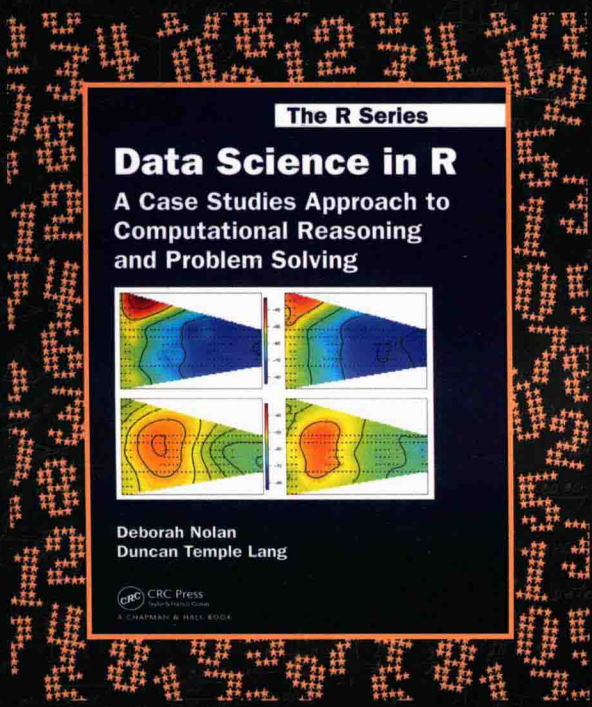
HZ BOOKS 华章IT
CRC Press Taylor & Francis Group
数据科学与工程技术丛书

数据科学R语言实践

面向计算推理与问题求解的案例研究法

[美] 德博拉·诺兰（Deborah Nolan） 邓肯·坦普·朗（Duncan Temple Lang）著
加州大学伯克利分校 加州大学戴维斯分校

于戈 赵志滨 聂铁铮 等译
东北大学



DATA SCIENCE IN R

A CASE STUDIES APPROACH TO COMPUTATIONAL REASONING AND PROBLEM SOLVING

机械工业出版社
China Machine Press

数据科学与工程技术丛书

DATA SCIENCE IN R

A CASE STUDIES APPROACH TO COMPUTATIONAL REASONING AND PROBLEM SOLVING

数据科学R语言实践

面向计算推理与问题求解的案例研究法

[美] 德博拉·诺兰（Deborah Nolan）
加州大学伯克利分校
邓肯·坦普·朗（Duncan Temple Lang）
加州大学戴维斯分校
著

于戈 赵志滨 聂铁铮 等译
东北大学

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据科学R语言实践：面向计算推理与问题求解的案例研究法/（美）德博拉·诺兰（Deborah Nolan），（美）邓肯·坦普·朗（Duncan Temple Lang）著；于戈等译．—北京：机械工业出版社，2017.6
（数据科学与工程技术丛书）
书名原文：Data Science in R：A Case Studies Approach to Computational Reasoning and Problem Solving

ISBN 978-7-111-57111-7

I. 数… II. ①德… ②邓… ③于… III. 程序语言－程序设计 IV. TP312

中国版本图书馆CIP数据核字（2017）第123922号

本书版权登记号：图字：01-2016-2916

本书带领读者身临其境地体验数据科学领域的日常工作，书中的12章即为12个鲜活的实践案例，包括航班延误数据分析、股票配对交易仿真以及二十一点纸牌游戏策略开发等，涵盖统计学、数据库、机器学习和可视化技术等众多知识点。本书的重点是计算推理和问题求解的思维过程，而不涉及具体编程语言的语法细节。

本书适合作为统计计算、数据挖掘等相关课程的补充案例教材，也适合该领域的技术人员阅读参考。

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：朱秀英　　责任校对：殷　虹
印　　刷：北京诚信伟业印刷有限公司　　版　　次：2017年6月第1版第1次印刷
开　　本：185mm×260mm　1/16　　印　　张：28.25（含0.25印张彩插）
书　　号：ISBN 978-7-111-57111-7　　定　　价：119.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

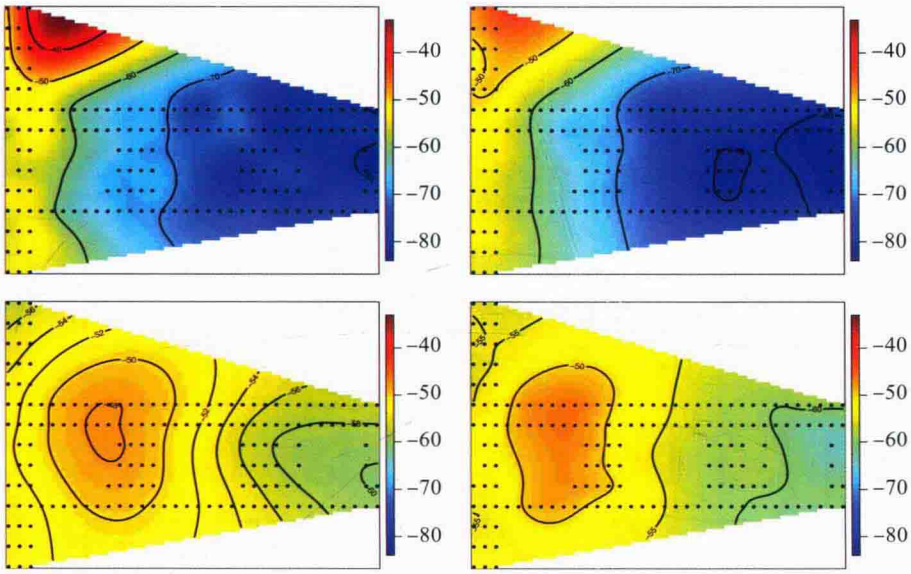


图 1-10　在两个接入点和两个角度上的信号中值。这 4 个热度图提供了信号强度的平滑地形表示。上面的两张地图分别对应于接入点 `00:14:bf:b1:97:90` 的角度 0°（左图）和角度 135°（右图）。底下的两张地图分别对应于接入点 `00:0f:a3:39:e1:c0` 的两个同样角度

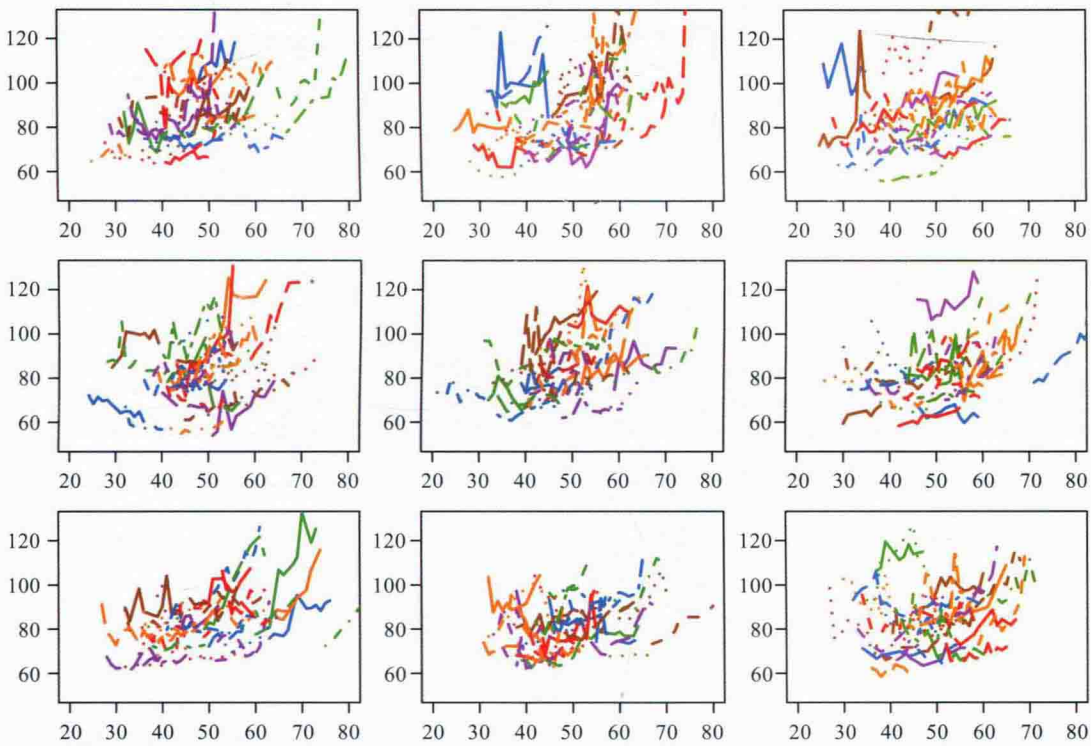


图 2-17　多场赛事的比赛用时。这些曲线图表示至少完成 8 次樱花赛的男选手的跑步时间。每个连接的片段集合对应一位运动员的跑步时间。观察所有的曲线图可以看出，该图和图 2-7 中的散点图呈现相似的形状，例如，图形随着年龄向上弯曲。然而，我们也可以看到单个选手成绩的变化情况。例如，许多中年选手的跑步时间随着年龄的增长而迅速增加，但并不是所有人都这样。他们中的一些人的成绩在进步，而另一些人的成绩变化较慢

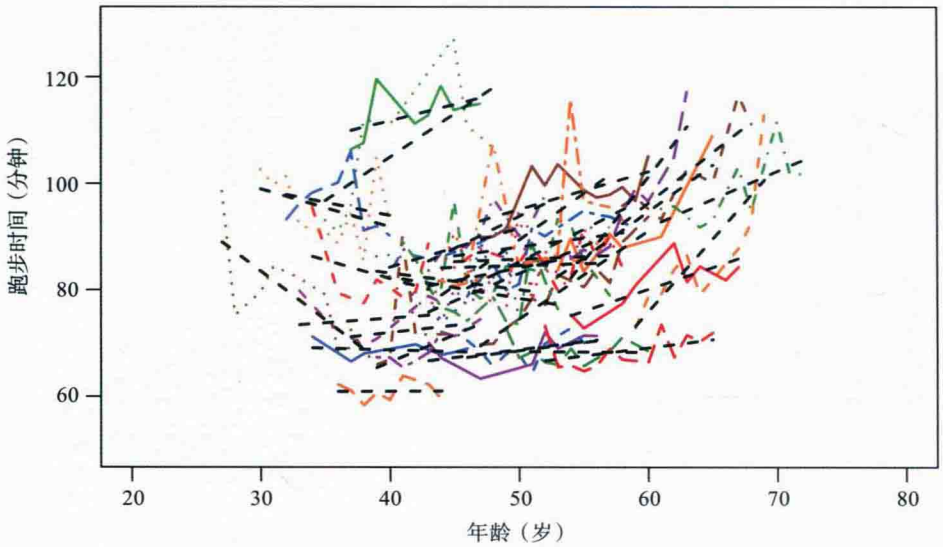


图 2-18　单个选手的跑步时间关于年龄的线性拟合。这里我们用最小二乘法对每个运动员的比赛用时进行拟合，以增强图 2-17 右下角的曲线图。总共有 30 条左右单个选手的黑色虚线段

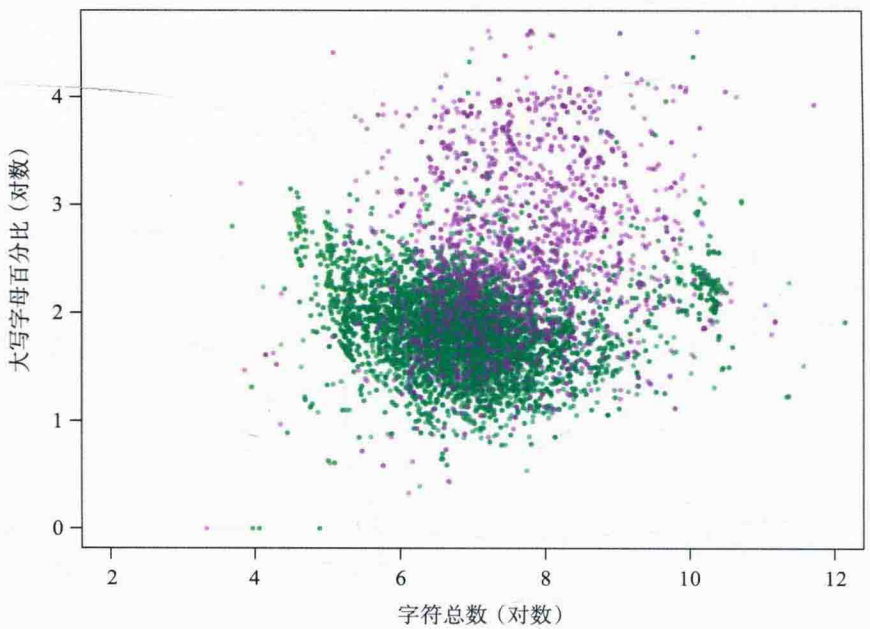


图 3-6　大写字母数量与消息规模的比较。该散点图考察消息中大写字母在所有字母中所占的比例以及该消息中字符的总数之间的关系，其中垃圾邮件用紫色的点表示，非垃圾邮件用绿色的点表示，较暗的颜色表示重叠的点。在图中我们可以看到，与非垃圾邮件相比，垃圾邮件趋于更长且具有更多的大写字母

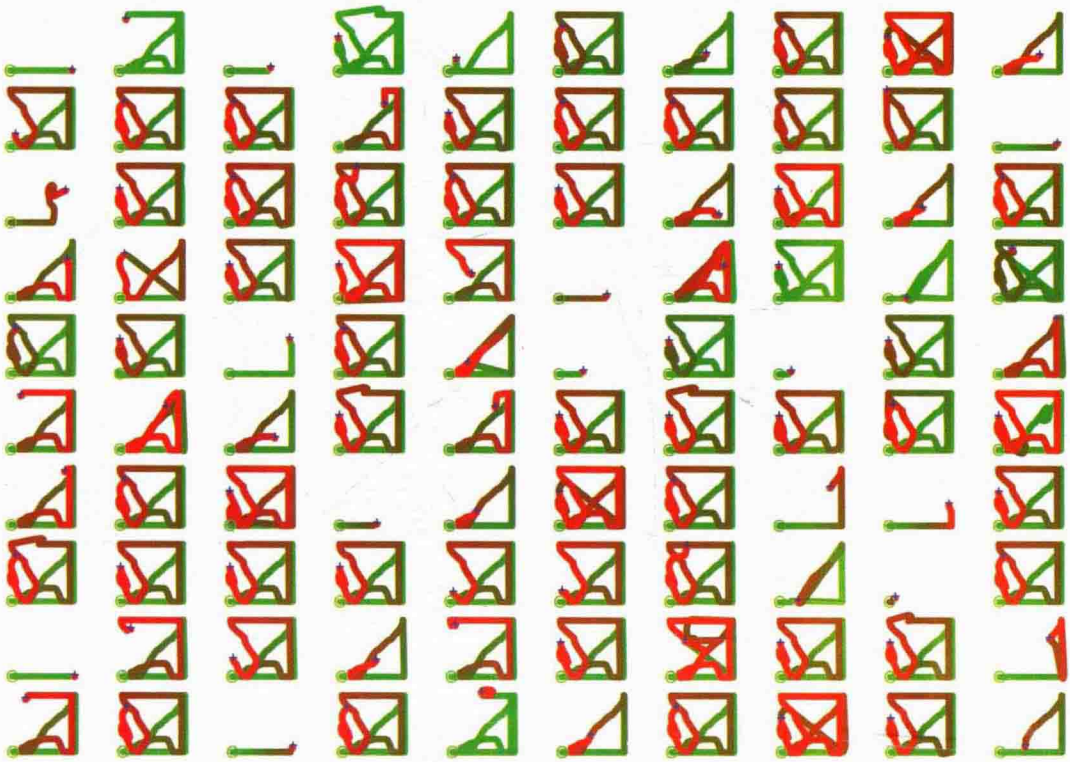


图 4-7　所有实验的展示。本图显示了机器人在 100 次实验中每一次实验的路径，起始点是绿色的圆形，颜色从绿到红的变化对应的是机器人运动的方向。最终的位置被标记为一个蓝色的 +

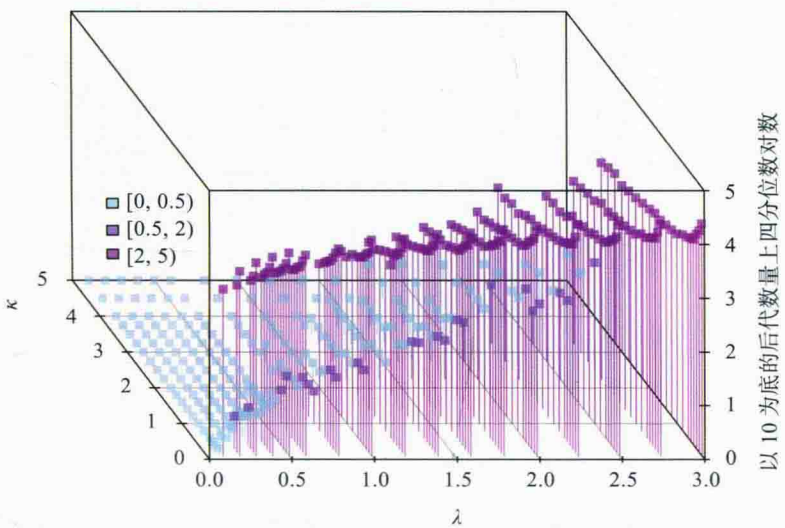


图 7-7　λ 和 κ 值及对应的后代数量三维散点图。散点图中每个点代表一个特定（λ, κ）对的分支过程的 400 个随机结果的上四分位数。后代数量是以 10 为底的对数刻度绘制，因此第一类，即 [0, 0.5)，对应了 1～3 个后代

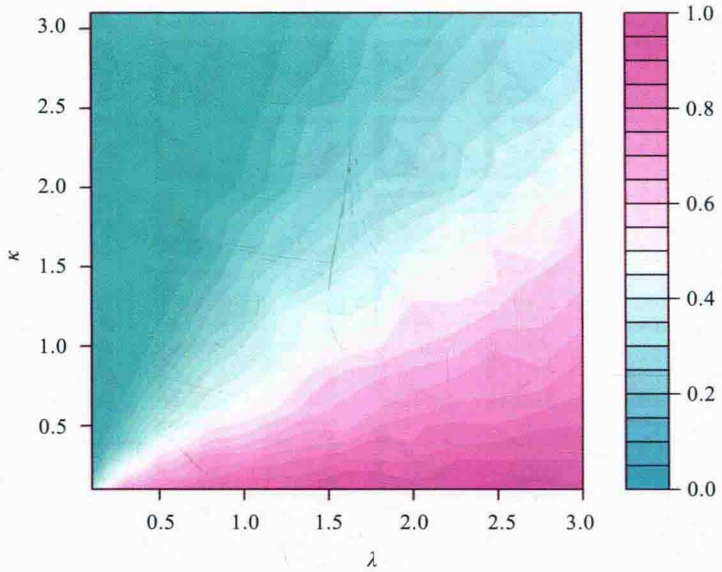


图 7-8　达到仿真限制的副本比例的图像映射图。这个图像映射图显示出 400 次仿真中，达到 20 个世代或 1000 个后代并因此终止的仿真所对应的（λ, κ）点对的比例的平滑轮廓

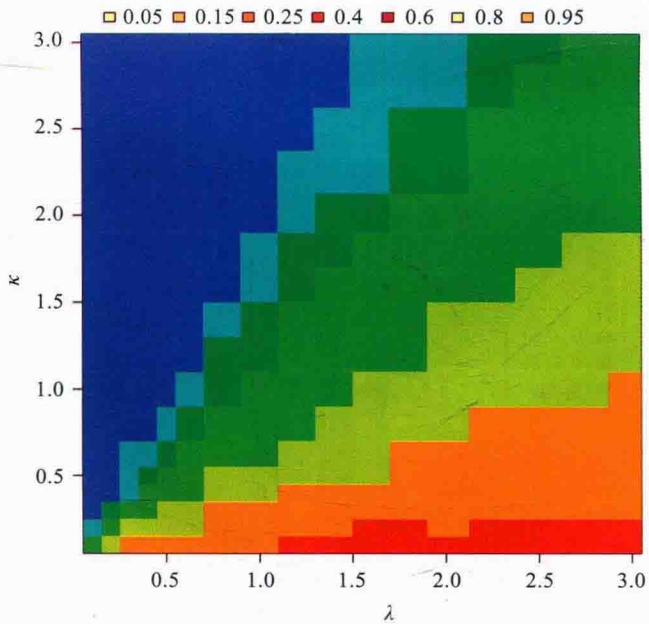


图 7-9　至少有 20 个后代的仿真的比例。这个图像映射图使用了彩虹画板，表现的是 400 个随机结果中，至少有 20 个后代的仿真对应的（λ, κ）值的比例

译 者 序

在过去的短短几年里，随着物联网、Web 2.0 等技术的迅猛发展和普及应用，可用的数据量呈爆炸式增长。可以说，在当今社会，找到数据已经不再是难事，如何有效地使用数据、从数据中挖掘出具有实践指导意义的领域知识才是问题的关键。在这样的背景下，数据科学应运而生。

数据科学主要是综合运用计算机技术、数学和统计学理论，并结合实质性的专业知识，开发面向应用领域的数据科学项目。数据科学为大数据分析和应用提供理论基础和方法。在人类社会迈入大数据时代的今天，数据科学显得尤为重要。

正如作者所言，撰写本书的目的是让读者身临其境地体验解决数据分析实际问题的思维过程，也为开设数据科学课程的教师提供丰富而生动的材料。本书的译者目前正在从事数据科学的研究工作，也深深地感到，在应用领域中实践数据科学时，我们并不缺少成熟的统计学理论和方法，也不缺少精巧的计算机算法和工具，我们面临的最大障碍是难以遵循正确的数据科学思维过程，难以把握数据科学项目中所涉及的各种问题及各种可能的答案。为此，本书精选了 12 个真实的数据分析项目，在一个个具体的案例中说明正确的数据科学思维过程：如何着手处理问题，以及如何考虑采取各种方式实现解决方案。本书注重实战，在各个案例中，首先描绘具体情境，提出初始目标，然后进行代码实现和评测。在此基础上，作者又提出了新的更高层次的目标，并有的放矢地对代码进行修改、精化、扩展和概化，从细节上为读者展示了那些富有经验的数据科学家的日常开发活动。这种渐进式的以问题为引导的内容安排方法，始终使读者目标清晰、兴趣盎然。正因如此，本书虽然内容颇多，但阅读起来轻松愉快！作为译者，我们也感觉翻译此书收获颇丰、受益匪浅。

本书由东北大学计算机科学与工程学院于戈、赵志滨、聂铁铮、申德荣、王大玲、鲍玉斌、张天成、寇月、冯时、冷芳玲、张一飞翻译。其中，前言和第 1 章由于戈负责，第 2 章由张天成负责，第 3 章由王大玲负责，第 4 章由赵志滨负责，第 5 章由鲍玉斌负责，第 6 章由寇月负责，第 7 章由冷芳玲负责，第 8 章由张一飞负责，第 9 章和第 11 章由聂铁铮负责，第 10 章由申德荣负责，第 12 章由冯时负责。全书由于戈、赵志滨、聂铁铮统稿和审校。

本书给出的全部代码由聂铁铮进行了验证性运行，正确无误。

本书涉及数据库、机器学习、可视化技术、统计学等多个领域，12 个经典案例有着截然不同的应用背景，尽管译者长期从事数据管理、数据仓库、数据挖掘、机器学习等方面的教学和研究工作，但终究水平有限，尤其是深入到具体应用领域的背景知识难免不足，敬请专家和读者批评指正。

译者

2017 年 3 月

前　言

我们编写本书有两重目的：一是想让学生能够阅读到计算推理方面的内容以及真实世界中数据分析的细节；二是希望提供有趣而且有用的资料，帮助统计学教师为新型的统计学和数据科学专业的学生讲授一门新拓展课程的重要方面。这门强化型课程是为了揭示数据分析和计算推理方法，而不是注重统计方法学。我们的目标不是提供简短的答案和方案，而是探索在数据科学项目中涉及的各种问题、各种可能的方案以及思维过程。

本书目标

有很多种常用于数据分析和数据科学的编程语言。我们在本书中重点使用 R 语言，但也会使用其他类型的领域专用语言（DSL），甚至还会用到 UNIX shell 语言和 C 语言。本书不打算讲授包括 R 语言在内的任何语言的文法或语义，也不会罗列大量数据科学家常用的 R 语言程序包和函数。本书的编写是为了使读者能够体验数据分析中真实计算问题的**思维过程**。有很多书籍讲解程序设计，所采用的方法是用一个章节介绍重要概念，再用其他章节介绍一些示例。这种方式是非常有用的，可以作为学习的基本出发点。但是，本书中作为示例的程序代码是由专家编写的最终精良版本，我们不会专门为读者说明编写代码的实际**过程**，而是直接给出最终结果代码。我们的目的是要举例说明这样的过程：程序员如何着手处理问题，以及如何考虑采取各种方式实现解决方案。这个过程具有高度的动态性和可重复性。我们首先编写一部分代码，然后测试代码、修改代码、精化代码、扩展代码和概化代码。经常出现的情况是，当从第一次尝试或原型中学到经验后，我们会“从头再来”，重新开发一个更简洁、清晰的版本。在这个过程中，我们需要在简洁性、效率、通用性、可重用性、正确的近似结果等各种要求之间做出折中。我们试图找到的方法是，最小化代码修改，但使得代码执行得更快，也更灵活。本书中，我们想要示范说明这个整体过程，以及成熟的程序员经常会根据丰富的经验做出的那些决定。希望本书能对普通教材做出补充，能为学生、研究者（甚至是教师）简要地展示专业数据科学家如何思考日常计算任务。

案例研究在统计计算课程中的应用

为统计计算（或任何）专业开设一门新的课程，对教师来讲是一项非常耗时的任务。我们常常必须去学习一些新的主题，或起码的基本细节，对它们进行优选和排序，确定哪些主题必须放在课程里，以及按照什么次序排放。我们必须准备大量的作业，以便年复一年地轮换使用。我们还可以布置一些综合性程序设计作业以帮助学生学习，比如矢量化、循环、正则表达式等内容。这些可怕的入门练习对于刚刚接触基本概念的初学者来说是必需的，但这些入门练习不一定要被扩展为大作业或小型项目。我们比较赞成的方法是，在统计计算课程中给学生安排真实的实际数据分析项目，这些项目将新概念紧密结合到常规的数据科学工作流中。我们想为学生揭示数据科学家的日常活动，我们认为学生会对这些内容感兴趣，而且这也有助于他们了解广泛的数据分析应用。进而，我们想要与计算主题一起介绍一些统计方法和概念，这些主题在其他课程中是没有的。基于这些理由，我们的统计计算课程起到了“百宝箱”的作用，囊括数据科学家为了日常工作必须掌握的许多“真实世界”中的主题。

在记住了这些目标后，找到教学上有趣的项目和作业是一项极其有挑战性的任务。要求这些项目和作业能够让学生实际完成并能激发他们的兴趣，还要能够示范专门的主题。在加州大学伯克利分校和戴维斯分校讲授计算课程时，我们花费了数日乃至数周的时间来开发作业，对可能的数据集和数据源产生了许多想法。我们往往需要对 4～5 个相关问题进行“面试”，然后从中筛选出其一并转化为作业。有些问题虽然有趣，但是过于简单或者过于复杂，因而不得不放弃。在进行完数据处理后，有些问题确实成为有趣的统计问题或数据分析问题，而有些问题则不适宜用来讲解那些我们希望学生关注的与计算和统计相关的主题。我们希望本书及其案例研究在将有趣的问题整合到面向数据科学技能的统计课程和计算课程的过程中，为教师扫清障碍。

在当今数据科学时代，我们拥有众多丰富而有趣的数据集可用于研究和教学。Debby Swayne、Paul Murrell 和 Hadley Wickham 等人组织的 Data Expo 竞赛就是一个很好的数据来源，可提供各种有趣的、具有挑战性的、可管理的问题。数据仓储（如加州大学欧文分校（UCI）数据仓储）在数量和多样性方面也在不断增长。一些网站（如 Kaggle.com）也能提供有趣的问题和数据。本书的关注点与它们稍微有所区别。我们尝试从原始数据开始，鉴别和探索有趣的潜在问题，而不是使用规定好的问题或预处理过的数据。让学生既体验如何获取和处理结构化或半结构化数据，也体验如何限定和构造关于这些数据的有趣问题，我们觉得这些是非常重要的。这个动机源自于我们在工业研究实验室（IBM 和贝尔实验室）、暑期学校（如统计学研究中的探索（ESR）暑期学校）以及加州大学伯克利分校和戴维斯分校所进行的教学而积累的经验。

广泛的主题

本书汇集了非传统的作业、样例方案以及练习题。我们专门选择了涉及多种主题、技术和特征的问题，希望学生能够接触和学习如下这些问题，包括：

- 非标准数据格式（机器人日志，邮件消息）；
- 文本处理和正则表达式；
- 新兴的或非传统的技术（Web 抓取、Web 服务、JSON、XML、HTML、KML 和 Google Earth）；
- 统计学方法（分类树、*k*- 近邻、朴素贝叶斯）；
- 可视化和探索式数据分析；
- 关系数据库和 SQL；
- 仿真；
- 算法实现；
- 大规模数据和处理效率；
- 软件设计、开发和测试；
- 使用和连接其他类型的语言，如 UNIX shell、C 和 Python。

我们本来希望涵盖更多其他方面的计算主题，例如，现代统计方法和机器学习方法、版本控制、动态文档、并行计算、Hadoop 和 MapReduce、数据技术、高级文本处理概念，等等，但由于空间和时间的限制，本书中并没有包揽这些内容。

本书包含的案例研究是我们在自己课堂上使用的一部分。它们并不是完美的，欢迎对它们的各种缺陷进行批评。尽管如此，我们希望这些案例研究对于学生和教师来说都是同等有价值的。我们也希望借助于它们能够促进更多的人发布更多的案例研究、问题和数据集等，以帮助学生学习所需要的计算和统计推理技能。

目标读者

由于本书不打算讲解在案例研究中所使用的任何编程语言的基本知识，因此它不是一本可独立使用的教科书。但我们认为它可用作与数据处理有关的所有年级学生的实用大纲，还可用作本科生和一年级研究生的统计计算课程的补充读物。对于那些正在从事数据科学研究的或者新入行的数据科学家，如果还没有正式学过统计计算课程，我们期望这本书对他们有价值。在这些人里面，对于那些寻找导论教材程度之上内容的自学者、本科生、研究生（甚至是教师），我们期望这本书更为有用。此外，对于那些想要对各种数据科学问题中的思维过程及其常规计算细节进行探讨的人来说，我们期望本书中的材料也是很有用的。

三个部分的主题

我们将本书分为三个部分，每个部分对应一个通用主题。虽然它们都注重计算问题，但也注重可视化、数据技术以及平常很少讲授的统计学技术 / 机器学习技术。

第一部分中的案例研究涵盖如何读取和转换原始数据，如何操作和可视化原始数据，以及如何使用统计技术以设法解决问题或者理解变量之间的关系。数据一般是非标准格式或非标准来源（如 Web 页面）。这部分使用的统计技术不是很复杂，但也不同于学生在本科课程里所学到的技术。

第二部分的重点是使用仿真方法理解随机过程本身，同时探索如何使用仿真方法对感兴趣的情景进行建模。这些案例研究也探讨一些高级的计算主题，如参考类、高效率的惯用语法和计算方法等。

第三部分中给出的最后一组案例探索了各种数据技术，包括数据库、使用 KML 进行可视化、使用 HTTP 请求和文本处理技术、从 Web 页面上抓取数据等。

将案例研究划分到哪个章节并不是精确和绝对的。例如，某些仿真主题涵盖数据操作，而某些数据操作和建模章节又涵盖仿真。数据技术方面的案例研究包含了许多数据操作。所有的案例研究都包含可视化，这样做既是为了理解和探索数据，也是为了便于程序代码的调试。

本书的关注点是统计计算，以及如何对数据进行存取、转换、操作、探索、可视化和推理。然而，除了这些技术和计算之外，所有的案例研究都是基于统计学、数学和工程等方面的有趣问题，这些问题本身就是值得研究的。在各章节中，我们将计算细节与统计学和数据分析的概念相融合。对数据和结果的分析，我们有意介绍得很简洁，而不是很详尽。我们的目的是使读者对具体的应用感兴趣，并停留或暂停在某个问题点上认真探索。感兴趣的学生可在这个点上针对用于解决问题的数据和统计方法进行大量的探索工作。本书提供了解决问题的计算基础，并留给学生和教师去做进一步的探讨，同时也提供了许多合适的练习题和各种探索方向。

排版格式约定

在几个案例研究中，我们使用了除 R 语言以外的语言，如 SQL 和 C 语言。尽管上下文已经清楚地指出代码块是使用了 R 之外的语言，但我们还是会在页面的空白处做出说明。例如，UNIX 的 grep 命令显示为：

shell
```
grep position2d JRSPdata_2010_03_10.log | grep -v ' 004 '
```

在编写代码的过程中，我们也介绍错误或差错，作用是让学生学习如何更好地着手处理和解决计算问题。因此，本书中的某些代码会故意写成不正确的或欠缺的（即可以工作但不是好方法）。我们在页边用禁止符号标识这种代码，例如：

```
createGrid(c(3, 5), .5)

Error in grid[pos] = sample(rep(c("red", "blue"), numCars)) :
  (converted from warning) number of items to replace is not
   a multiple of replacement length
```

注意，在本书中给出的用于创建图表的程序代码略不同于实际用于创建可显示图表的代码。在通常的做法中，使用 R 创建绘图时会添加标题，要么是为了交互式观看，要么是为了能够包含到幻灯片和报表中。但是，本书没有这样做，因为图形是显示在图表中，而图表已包含了说明文字和标题。为了避免冗余，代码中删除了对绘图标题的定义。

最后一个约定与练习题有关。有的案例研究将练习题分散在一章的不同位置，有的案例研究将所有练习题集中到一章的结尾。为了帮助识别和查找习题，我们在页边靠近练习题的地方加一个问号。例如：

Q.1 写出在一张图上包含两个数据序列的函数，记住要在你的图表上加标题。

致　谢

我们衷心感谢为本书提供案例研究的所有贡献者。他们在准备各自的章节时投入了大量的时间和精力。这是一个漫长的过程，但他们非常耐心和善解人意。

出版这本关于案例研究的书的初衷源自于一个美国自然科学基金会（NSF）资助的研讨会。我们在2007年主办的这个会议主要是探索计算在统计学课程体系中的作用。其中一个重要想法是共享教学资源，本书是这个想法的成果之一。

我们感谢各届NSF研讨会的众多与会者。这些研讨会重点关注统计学课程体系中的计算主题，它们可划分为：开发示范课程；建立案例研究；帮助教师讲授现代统计计算。获取所有与会者的研究兴趣、研究热点、反馈和资料是极为重要的。我们特别感谢为案例研究提供新想法和资料的2009年研讨会的与会者，他们是：Samuel Frame（加州州立理工大学，圣路易斯－奥比斯波校区），Robert Gould（加州大学洛杉矶分校），Albyn Jones（里德学院），Michael Kane（耶鲁大学），Daniel Kaplan（麦卡莱斯特学院），Cari Kaufman（加州大学伯克利分校），Guy Lebanon（普渡大学，现就职亚马逊公司），Matt Levinson（加州大学洛杉矶分校），Thomas Lumley（华盛顿大学，现就职奥克兰大学），John Monahan（北卡州立大学），Roger Peng（约翰霍普金斯大学），Andrew Schaffner（加州州立理工大学，圣路易斯－奥比斯波），Luke Tierney（衣阿华大学），Frances Tong（加州大学伯克利分校，现就职Becton Dickinson Technologies公司），John Verzani（纽约市立大学史泰登岛学院），Mark Daniel Ward（普渡大学），Charlotte Wickham（加州大学伯克利分校，现就职俄罗岗州立大学），以及Hadley Wickham（莱斯大学，现就职RStudio公司）。

我们已举办了多年由NSF资助的“统计学研究中的探索”（ESR）暑期研讨会。在这些会议上讲述了一至两天案例研究的各位研究者们，为我们思考如何向本科生讲述高级的现代统计学和数据科学提供了极大帮助。我们特别感谢的讲者有：Andreas Buja（宾州大学沃顿商学院），Amanda Cox（纽约时报），Francesca Dominici（约翰霍普金斯大学，现就职哈佛大学），Chris Genovese（卡耐基梅隆大学），Carie Grimes（谷歌公司），Mark Hansen（加州大学洛杉矶分校，现就职哥伦比亚大学），Dave Higdon（洛斯阿拉莫斯国家实验室），Diane Lambert（朗讯科技公司贝尔实验室，现就职谷歌公司），Dave Madigan（哥伦比亚大学），Doug Nychka（美国国家大气研究中心（NCAR）），Roger Peng（约翰霍

普金斯大学），Katie Pollard（加州大学旧金山分校格莱斯顿研究所），John Rice（加州大学伯克利分校），Patrick Ryan（Janssen Research and Development），Steve Sein（美国国家大会研究中心），Jas Sekhon（加州大学伯克利分校），Terry Speed（加州大学伯克利分校，沃尔特和伊丽莎·霍尔医学研究所），Claudia Tebaldi（美国气候中心），以及 Chris Volinsky（AT&T 香农实验室）。

除了讲者，ESR 研讨会中的许多研究者和教授也帮助准备了教程、职场座谈会等。我们感谢：Joe Blitzstein（哈佛大学），Dianne Cook（衣阿华州立大学），Nick Horton（史密斯学院，现就职阿默斯特学院），David James（朗讯科技公司贝尔实验室，现就职诺华制药公司），Cari Kaufman（加州大学伯克利分校），以及 Debby Swayne（AT&T 香农实验室）。

我们还要感谢为伯克利分校和戴维斯分校的课程以及 ESR 研讨会提供服务的众多助教们。他们通过讲授其中一些案例研究、甄别学生们遇到的状况和问题、提供有价值的反馈，帮助我们解决了一些重要问题。他们是：Gabe Becker，Neal Fultz，Tammy Greasby，Brianna Hirst，Wayne Lee，Erin Melcon，Rakhee Patel，Nick Ulle，Charlotte Wickham。还要感谢为第 2 章的早期版本提供反馈的 Ann Cannon。

我们要感谢本书的编辑 John Kimmel，他始终在支持和鼓励这个项目。

本书部分材料是基于由自然科学基金课题（批准号 DUE-0618865、DMS-0840001 以及 DUE-1043634）所资助的工作，一并致谢。

作者简介

Deborah Nolan（德博拉·诺兰）在改进数学和统计学的教学方法以及为本科生提供拓展服务方面倾注了大量心血。她担任加州大学伯克利分校本科教育的 Zaffaroni Family 主席，获得过伯克利分校的大学杰出教学奖，以及普林斯顿大学杰出教学 William R. Kenan, Jr. 客座教授席位。她是美国统计学会的会士，计算分会和教育分会的前任主席。她也是美国数理统计研究院的会士。她参与指导了数学和理学教师培训计划、加州大学教学培训项目、在职名师培训项目和美国数学教育培训项目。她出版了包括本书在内的多部著作。

Duncan Temple Lang（邓肯·坦普·朗）从事 R 语言和 S 语言程序开发工作 20 余年，开发了 100 多个 R 程序包。他着重探索和开发新的统计计算方法，主要贡献是调研来自其他学科的有发展前景的新范型和新技术，并将其集成到 R 环境中。他当前的研究工作包括：基于 LLVM 方法的 R 语言编译器、R 计算的溯源、类型推导，以及快速、灵活的贝叶斯和似然度计算框架（http://r-nimble.org），还有图形处理器（GPU）。现在担任加州大学戴维斯分校数据科学计划项目的主管。

Nolan 和 Temple Lang 是《XML and Web Technologies for Data Science in R》一书的共同作者。他们组织和领导了多个 NSF 资助的暑期计划，其目的是吸引大学生学习统计学领域的研究生课题，以及参加数据科学方面的小型研讨会。他们合作开发了“数据计算的概念”这门课程并在各自的学校里讲授。他们协作开发了支持交互式和可复制的动态文档、基于 Web 可视化等功能的系统。