HZ BOOKS
华章IT

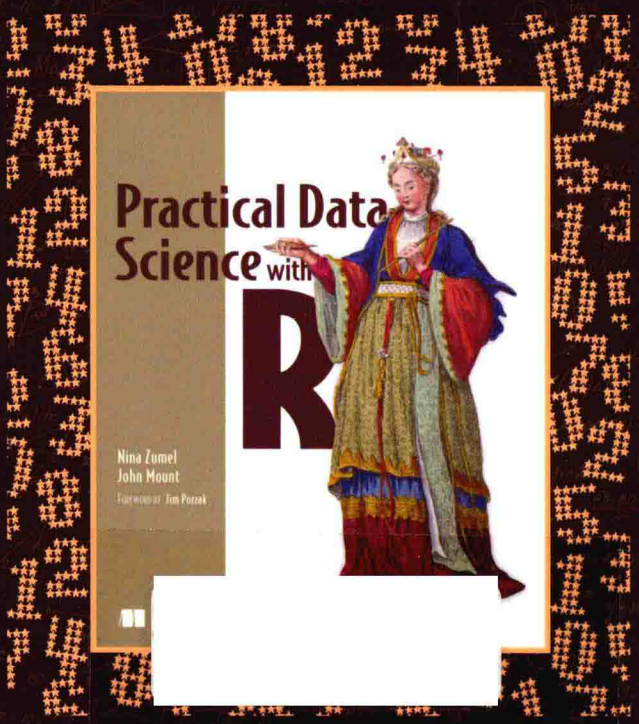
MANNING

数据科学与工程技术丛书

数据科学

理论、方法与R语言实践

[美] 尼娜·朱梅尔（Nina Zumel）
约翰·芒特（John Mount） 著
于戈 鲍玉斌 王大玲 等译



PRACTICAL DATA
SCIENCE WITH R

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PRACTICAL DATA SCIENCE WITH R

数据科学

理论、方法与R语言实践

[美] 尼娜·朱梅尔（Nina Zumel）
约翰·芒特（John Mount） 著

于戈 鲍玉斌 王大玲 等译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据科学：理论、方法与R语言实践 /（美）朱梅尔（Zumel, N.），（美）芒特（Mount, J.）著；于戈等译 . —北京：机械工业出版社，2016.3
（数据科学与工程技术丛书）
书名原文：Practical Data Science with R

ISBN 978-7-111-52926-2

I. 数… II. ① 朱… ② 芒… ③ 于… III. 数据处理 IV. TP274

中国版本图书馆CIP数据核字（2016）第041772号

本书版权登记号：图字：01-2015-7586

本书从实用的角度较为全面地展现了数据科学的主要内容，并结合大量的实际项目案例，利用R语言详细地讲解了数据项目的开发过程和关键技术。本书包括三个部分共11章的内容，主要介绍了数据科学项目的处理过程、选择合适的建模方法，也讨论了bagging算法、随机森林、广义加性模型、核和支持向量机等高级建模方法。此外，还讨论了文档编制和结果部署，以及如何向组织内不同的受众展现项目结果。

本书适合作为高等院校高年级本科生和研究生及从事数据管理与分析的工程技术人员的主要参考书。

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：缪　杰	责任校对：董纪丽
印　刷：三河市宏图印务有限公司	版　次：2016年4月第1版第1次印刷
开　本：185mm×260mm　1/16	印　张：21.25
书　号：ISBN 978-7-111-52926-2	定　价：69.00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

译 者 序

数据科学是关于如何从数据中获取知识的一门新兴学科，主要通过统计学、机器学习和计算机科学等方法，开发面向应用领域的数据科学项目，在数据的基础上建立预测模型，并部署到实际生产环境中，用于解决生产经营、行政管理、科学研究等许多重要领域中的实际问题。数据科学为大数据分析和应用提供理论基础和方法学，在当今大数据时代中，数据科学显得尤为重要。

本书正是一本介绍如何在组织机构中使用 R 语言将数据科学的理论和方法应用到实际生产中，并对其数据进行管理和分析预测的书籍。正在从事数据科学及相关技术研究的我们，在看到本书的英文版时，立刻被其中的内容所吸引——本书从数据科学处理过程的描述出发，以读者容易理解的现实生活中的实际案例为分析对象，以预测模型的构建及评价为主线，最终落实到处理结果的交付。本书丰富的内容以及这种独特的组织结构从一个全新的实战视角为我们展现了一幅从数据科学理论、R 语言和系统到它们在工程实践中的应用，有理论、有方法、有案例、有分析、有结果的精彩画卷，将抽象的统计分析理论和机器学习的方法，利用 R 语言和系统生动直观地展现在读者面前。通过翻译这本书，我们从中收获颇多、受益颇丰。

本书的作者 Zumel 博士和 Mount 博士在数据科学项目的咨询、分析、设计和开发方面，具备系统的理论知识和丰富的实践经验，本书从实用的、面向实践的角度较为系统、全面地展现了数据科学的主要内容，并结合大量的实际项目案例，详细地讲解了数据科学项目的开发过程和关键技术。本书无论对于初学者还是有经验者，都是一本非常有价值的参考书。

本书具有以下几个特点：

1. 内容系统全面。详尽介绍了数据科学项目中基础数据的准备技术、预测模型的建模技术，以及数据模型的部署和应用技术。

2. 表达通俗易懂。按照数据科学项目开发的过程，循序渐进、由浅入深地介绍了数据科学项目的基础理论、开发步骤和开发技术。

3. 实践指导性强。结合真实数据集的分析和处理，给出了大量 R 语言源代码，为读者今后开发实际项目提供了宝贵的参考和借鉴。

本书由东北大学计算机科学与工程学院于戈、鲍玉斌、王大玲、张一飞、冷芳玲、张

天成、赵志滨、寇月、聂铁铮翻译。其中，前言和第 7 章由王大玲负责，第 1 章由于戈负责，第 2 章和第 6 章由鲍玉斌负责，第 3 章和第 10 章由张一飞负责，第 4 章和第 9 章由冷芳玲负责，第 5 章和第 8 章由张天成负责，第 11 章由赵志滨负责，附录 A 和附录 B 由寇月负责，附录 C 由聂铁铮负责，词汇和索引表由鲍玉斌、于戈负责。全书由于戈和鲍玉斌统稿和审校。

本书涉及数据库、软件工程、机器学习、统计学等多个领域，理论和实践方面的内容较多，尽管译者在数据库管理、数据仓库、数据挖掘、机器学习等方面具有一定的教学和科研经验，但毕竟水平有限，难免存在不足之处，敬请专家和读者批评指正。

译　者

2016 年 1 月

序　言

如果你是一名刚入门的数据科学家，或者是一个想要从事数据科学工作的新手，那么本书（以下简称 PDSwR）将为你的起步提供帮助。如果你正在从事数据科学工作，PDSwR 一书将填补你的知识缺口，甚至使你重新审视日常使用的工具——就我本人体会而言确实如此。

虽然目前已有很多介绍如何使用 R 语言进行统计和建模的优秀书籍，但却很少有介绍如何在组织机构中应用数据科学进行管理的好书。在将坚实的技术内容与具体的实际相结合并脚踏实地指导如何进行实践方面，本书是独一无二的，我对此的期待丝毫不亚于本书作者 Nina 和 John 这两位专家。

我初识 John 时，他正在早期的 Bay Area R Users Group（湾区 R 用户群）中分享使用 R 语言过程中的悲欢苦乐。此后，我和 Nina、John 在同一家公司合作完成了几个项目。在比 Bay Area R Users Group 更大的群以及 Berkeley R-Beginners（伯克利 R 初学者群）中，John 讲解了 PDSwR 一书中的早期思想。基于作为一名数据科学家所取得的实践经验，John 直率地表达了有关数据科学工作的明确看法。PDSwR 一书反映出 John 和 Nina 关于如何从事数据科学的确切观点——使用哪些工具、采用的处理过程、使用的重要方法、人际交流的重要性等，这些都讲得一清二楚。

因此，这对我而言是非常完美的，特别是我与他们持有 98% 的相同观点（我唯一吹毛求疵之处是在 SQL 方面——但那仅是因为我的工作经历不同，而不是存在什么根本性分歧）。他们用清晰文字所表达出的意义，使你更加关注数据科学的技艺，而不必为选择什么工具或什么方法而纠结，正是这种缜密使得 PDSwR 一书非常实用。下面让我们来看些细节。

实用工具集：R 语言本身是规定好了的，RStudio 是一个精选的 IDE（集成开发环境）。自从 RStudio 推出后，我们就一直在使用它，现在它已经发展成一个出色的工具——在最新的版本中集成了调试工具。PDSwR 一书中第三个主要工具是 Hadley Wickham 的 ggplot2，尽管 R 语言在传统上提供了优秀的图形和可视化工具，但 ggplot2 将其可视化提高到一个新的水平（我的实用心得：仔细查看 Hadley 或他的学生的任何 R 工具包）。除了这些主要工具之外，PDSwR 一书还介绍了必要的辅助工具：适合于大型数据集的 SQL

DBMS，支持源代码版本控制的 Git 和 GitHub，用于文档生成的 knitr 工具包。

实用数据集：学习数据科学的唯一途径是通过实践来掌握它。在典型的教学数据集到真实世界数据之间存在很大的距离，PDSwR 一书在供学习所用的（简单）数据集与真实世界中杂乱的数据集之间做了一个很好的折中。PDSwR 一书引领你学会如何通过探索一个新的数据集去发现数据中的问题，以及如何在必要时对数据进行清洗和转换。

实用人际关系：数据科学是一门关于为你的客户解决真实世界问题的科学——这些客户可能是一名咨询师或者你的机构中的一员。无论是哪种情形，你都要与多种个性的人打交道，他们拥有各自的动机、技能和责任。作为承担实际工作的咨询师，Nina 和 John 对此有深入的理解。在数据科学项目进行过程中，要充分理解这些角色，PDSwR 一书在强调这个重要性方面也是独一无二的。

实用建模技术：PDSwR 一书的大部分内容是关于如何建模的。首先，全面概述了建模处理过程，包括如何选择要使用的建模方法，以及在完成模型后如何度量该模型的质量。该书指导你掌握今后工作中所需要的最实用的建模方法，并直观地讲解了每种建模方法的基础理论。本书使用具体的实例贯穿全书——在作者的 GitHub 网站上提供了代码和数据。最重要的是，本书介绍了使用技巧和陷阱，并在每一节的结尾给出了实用的要点。

简而言之，本书是所有数据科学家都应该拥有的一部独一无二、举足轻重的书籍。

Jim Porzak

资深数据科学家、Bay Area R Users Group 联合创始人

前　言

本书正是我们在自学数据科学时所希望得到的那本书，我们可从中了解哪些主题和技能的集合构成了**数据科学**。它也是我们希望发给客户和同行的书。本书的目的是讲解统计学、计算机科学和机器学习等学科中对于数据科学极为重要的内容。

数据科学利用了来自实验科学、统计学、报表技术、分析技术、可视化技术、商务智能、专家系统、机器学习、数据库、数据仓库、数据挖掘和大数据技术等各个领域的工具，正是因为我们需要如此多的工具，因此需要一个能够涵盖全部内容的学科。数据科学本身与这些工具和技术的区别，在于数据科学的核心目标是将有效的决策支持模型部署到实际生产环境中。

我们的目标是从实用的、面向实践的角度来展现数据科学，通过在真实数据上的可运行的练习题，我们设法达到这一目标，全书给出了 10 个重要的数据集。我们认为这种方式能举例说明我们到底想要讲授什么，还能演示说明在真实世界项目中所必需的全部预备步骤。

贯穿全书，我们讨论实用的统计学和机器学习概念，给出具体的代码示例，探索如何与非专业人士开展合作以及如何向他们讲解。我们期待，即使你不能在这些主题中发现新意，这本书也能够在你还未想到的其他一两个主题上闪出一道灵光。

关于本书

本书讨论数据科学的概念和方法：数据科学领域主要使用统计学、机器学习和计算机科学的成果来建立预测模型。由于数据科学具有宽泛性，所以有必要对其展开一些讨论并对本书所涉及的方法加以界定。

什么是数据科学

统计学家 William S. Cleveland 将数据科学定义为一个比统计学自身大得多的跨学科领域。而我们定义数据科学为一种管理过程，该过程能够将假设和数据转换成可应用的预测。典型的预测分析目标例子有：预测谁将在选举中获胜、什么样的商品放在一起销售更好、哪些贷款将被拖欠或者什么网上广告将被点击等。数据科学家负责获取数据、管理数据、选择建模技术、编写代码以及验证结果。

由于数据科学领域涉及众多的学科，所以它通常进行“二次调用”。我们遇到的许多

优秀数据科学家原本是程序设计者、统计学家、业务分析师或科技工作者，他们在原有知识储备的基础上再多学一些技术，就成为了优秀的数据科学家。这一观察促成了本书的写法：通过具体地介绍在真实数据上执行的各个通用的项目开发步骤，来介绍数据科学家所需的实用技能。对于这些开发步骤，有的你将比我们懂得更多，有的你会更快地掌握，有的还需要你进一步深入研究。

数据科学的大多数理论基础来源于统计学，但正如我们所知，数据科学强烈地受到技术学和软件工程方法学的影响，并且在计算机科学和信息技术所驱动的各个子领域中得到了极大的发展。下面通过列举一些著名的案例来体会数据科学的若干工程风格：

- Amazon 的商品推荐系统
- Google 的广告评估系统
- LinkedIn 的人脉推荐系统
- Twitter 的趋势话题
- Walmart 的消费者需求预测系统

上述系统有许多共同特点：

- 所有系统均建立在大规模数据集基础之上。它们并非一定属于大数据领域，不过如果仅使用小数据集的话，这些系统将无法成功。为了管理数据，这些系统需要源自计算机科学的概念：数据库理论、并行程序设计理论、流数据技术以及数据仓库。
- 这些系统大多是在线或实时运行的。当数据科学团队部署一个决策程序或打分程序时，目的是要用于直接做出决策或直接向许多终端用户展示结果，而非只是产生单一的报表或分析结果。生产部署阶段是校正结果的最后机会，因为数据科学家不会长期留在现场来解决存在的缺陷。
- 所有系统均允许出错，但出错率的上限是不容讨价还价的。
- 这些系统不需考虑因果关系，如果它们能发现有用的相关性，就算作是成功的。它们不必非要从结果中正确地找出导致该结果的原因。

本书讲授构建这样的系统时所需要的原理和工具，包括：通用的任务、开发步骤和成功地交付这样的项目所使用的工具。我们强调整个工作过程——如何进行项目管理，如何与其他人合作，以及如何对非专业人士展现结果。

导读路线图

本书涵盖如下内容：

- 如何对数据科学处理过程本身进行管理。数据科学家必须有能力来衡量和跟踪他们自己的项目。
- 如何应用在数据科学项目中常用的最强的统计和机器学习技术。可将本书看作一系列有明确工作目标的练习，需使用程序设计语言 R 去实现真实的数据科学工作。
- 如何向各种利益相关者进行结果展现，包括管理人员、用户、部署团队等。必须用具体的术语向混合类型的受众解释你的工作，并且使用他们所熟悉的语言来表

达，而不要坚持使用专门领域的技术术语。对于数据科学项目的结果展现，你无法绕开这一障碍。

我们使用循序渐进的方式来安排本书的内容，其详细内容组织如下：

第一部分描述数据科学处理过程的主要目标和技术，强调协作和数据。

第 1 章讨论作为一名数据科学家如何开展工作，第 2 章介绍如何将数据装载到 R 系统，并演示如何启动 R 系统开始工作。

第 3 章讲授首先要在数据中寻找什么，以及用于刻画数据特征和理解数据的重要步骤。在做数据分析之前，必须准备好数据，另外必须修正数据中存在的问题，第 4 章介绍如何处理这些问题。

第二部分从刻画数据特征转到如何构建有效的预测模型上来。第 5 章提供将业务需求映射到技术评价和建模技术的初始词典。

第 6 章讲授如何通过记忆化训练数据构建模型。这种记忆化模型虽然概念上简单却非常有效。第 7 章进展到具有显式加性结构的模型问题，这种功能结构增加了进行有益的内插值和外插值，以及辨识重要变量和效果的能力。

第 8 章描述当项目中没有可用的带标签的训练数据时，还能够做什么。第 9 章介绍用于改进模型预测性能和修正具体建模问题的高级建模方法。

第三部分从建模问题再回到处理过程上来，展示如何交付建模结果。第 10 章演示如何管理、文档编制和部署模型。第 11 章介绍如何针对不同的受众给出有效的展现方法。

附录部分包括关于 R 系统、统计学和其他可用工具的补充技术细节。附录 A 介绍如何安装 R 系统、如何启动工作以及如何运用其他工具（如 SQL）。附录 B 是关于一些重要统计学思想的最新资料。附录 C 讨论附加的工具和研究思路。参考文献提供参考文献资料并介绍今后的研究机遇。

书中的学习材料是根据目标和任务来组织的，相关的工具在需要时才被引入。每一章的主题均以一个带有相关数据集的代表性项目为背景展开讨论。在学习全书的过程中，你将接触 10 个实质性项目。本书提供的所有数据集均保存在本书的 GitHub 资料库中（https://github.com/WinVector/zmPDSwR），你可以下载整个资料库（这是一个 zip 压缩文件，GitHub 服务之一），然后将该库复制到你的机器上，也可以根据需要只复制单个文件。

致读者

为学习和运行本书的例子，你需要熟悉一些 R 语言、统计学以及 SQL 数据库（某些例子涉及）的知识，建议你手头准备一些好的入门教材。在学习这本书之前，你不必是一位 R 语言、统计学和 SQL 方面的专家，但应该能够很轻松地自学本书提及却不能完整讲解的内容。

对于 R 语言，我们推荐参考 Robert Kabacoff 的《R in Action，2nd Edition》（www.manning.com/kabacoff2/）以及与本书相关的网站 Quick-R（www.statmethods.net）。对于统计学，我们推荐参考 David Freedman、Robert Pisani 和 Roger Purves 的《Statistics，4th Edition》。对于 SQL，我们推荐参考 Joe Celko 的《SQL for Smarties，4th Edition》。

总体上，我们所期望的理想读者应该是这样的：

- 对工作示例感兴趣。通过学习这些示例，你将至少学会一种方法，能够完成一个项目的所有步骤。你必须乐于尝试简单的脚本编写和程序设计以充分利用这本书。对于我们给出的每个示例，你应该尝试改变它，并且预料到会有某些失败（你的改变不奏效）和某些成功（你的改变优于示例）。
- 对R语言的统计系统有所了解并且乐于用R语言编写短的脚本和程序。除Kabacoff的书（《R in Action》）外，我们在参考文献中还推荐了几本好书。我们用R语言解决具体的问题。为了理解正在进行什么处理，你需要运行那些示例，并且阅读额外的文档以理解那些在本书中没有展示的变种命令。
- 对概率、均值、标准差和显著性等基本的统计学概念有一些经验。我们在需要时会引入这些概念，对于工作示例，你可能还需要阅读一些额外的参考文献。我们给出某些术语的定义，并提供某些主题的参考文献和合适的博客，但我们认为在某些主题上你需要自己在互联网上进行搜索。
- 一台安装有R系统和其他工具的计算机（OS X、Linux或Windows），以及用于下载有关工具和数据集的互联网。我们强烈地建议你进行示例学习，用R系统help()命令学习各种方法，并且跟踪学习某些补充的参考文献。

书中没有什么

本书不是一本R语言的使用手册。我们使用R语言具体地展示数据科学项目的重要步骤，通过示例讲授足够的R语言知识，但不熟悉R语言的读者需要查阅附录A以及许多优秀的R语言书籍和使用指南。

本书不是一系列案例研究集合。我们更强调方法和技术，在本书中给出案例数据和代码仅仅是为了确保我们给出的建议是具体的、可用的。

本书不是一本大数据方面的书。我们认为大多数有意义的数据科学问题出现在数据库级别或文件级别等可管理的大小规模上（通常比内存更大，但还未大到难以管理的程度）。有价值的数据是能够将测量到的条件映射到依赖于它们的结果上，但产生这些数据往往是代价高昂的，因而在实际应用中通常会限制这些数据的规模。而对于某些报表生成、数据挖掘和自然语言处理任务，才需要进入大数据领域。

本书不是一本理论方面的书。对于任何一种技术，我们不会强调其绝对严格的理论。数据科学的目标应该是支持灵活性，提供很多可用的好技术。并且，当某个技术能够用于解决手头问题时，深入地研究该技术。此外，由于要直接使用R语言代码，所以在本书正文中使用R代码符号，而没用美观的编辑公式。

本书也不是给机器学习多面手使用的。我们只强调那些已经用R语言实现了的方法。对于每种方法，我们介绍其操作的理论并表明该方法有何优点。我们一般不讨论如何实现这些方法（即便这种实现是容易的），因为这些信息是随处可得的。

编码约定及下载

本书是以示例驱动方式叙述的，我们在GitHub资料库（https://github.com/WinVector/zmPDSwR）中提供了准备好的示例数据，它们用R语言进行编码并且链接到初始源，你可以在线查询该库或者将其复制到你自己的机器上。由于从zip压缩文件中复制代码比从本书的电子版中复制和粘贴更容易，我们也提供了产生所有结果的程序代码以及在书中出现的几乎所有的图表（作为一个zip文件）（https://github.com/WinVector/zmPDSwR/raw/master/CodeExamples.zip）。你也可以从Manning出版社的网站（www.manning.com/PracticalDataSciencewithR）下载这些代码。

我们鼓励你在阅读本书时尽力实现这些R代码示例，即便在讨论数据科学中相当抽象的概念时，我们也会用具体的数据和代码来展示示例，在每章均给出了指向该章内容所参考的具体数据集的链接。

在本书中，代码均采用特殊字体书写，以将它们与正常文字区别开来，具体的变量和值采用类似的格式，抽象的数学符号则采用斜体。R是一种数学语言，许多短语都用到了上述两种字体。在我们的示例中，任何提示符（如“>”和“$”）都可以忽略掉。内嵌结果用R的注释符“#”作为前缀来标识。

软硬件要求

为学习示例，需要安装有Linux、OS X或Windows操作系统的计算机，并且安装了相关的软件（安装方法在附录A给出），我们推荐的所有软件都是完全跨平台的、免费使用的、开源的。

建议至少安装如下软件：

- R系统：http://cran.r-project.org。
- 各种来自CRAN的程序包（由R自身使用`install.packages()`命令安装并使用`library()`命令激活）。
- 版本控制工具Git：http://git-scm.com。
- RStudio：一个集成了编辑器、执行和绘图的开发环境——http://www.rstudio.com。
- 支持系统命令的bash shell，它嵌入在Linux和OS X系统中，能够通过安装Cygwin（http://www.cygwin.com）添加到Windows系统。我们不写任何脚本，所以对于一个经验丰富的Windows shell用户，如果能将我们的bash命令转换成对应的Windows命令，也可以不安装Cygwin。

关于封面插图

本书英文版的封面图片题为“1703年的中国女子服饰”。该插图是从Thomas Jefferys于1757年至1772年在伦敦出版的《各国古今服饰大全（共4卷）》中得到的，其扉页上说明这些都是手工着色的铜版画，用阿拉伯树胶加固。Thomas Jefferys（1719～1771）

被称作“国王乔治三世时代的地理学家”。他是一名英国绘图师，是当时顶级的地图供应商。他为政府和其他公务团体制作和印刷地图，生产了世界各地、特别是北美地区的商业地图和地图集。作为一名绘图师，他对其曾勘查和绘图地区的服饰习俗也感兴趣，这些服饰均出色地展示在这部 4 卷本的服饰大全中。

在 18 世纪，着迷于遥远的世界并为了愉悦而去旅行还是件新事物，类似这样的服饰大全很受欢迎，因为它们能够将其他国家的风土人情介绍给远行的实际旅行者和足不出户的空想旅行家。Jefferys 卷中各种各样的绘图生动地展示了几百年前世界各国的独特性。现在，着装标准发生了变化，在那个时代不同国家和地区之间存在的丰富多彩的差异性已经变得模糊不清，常常难以将一个地区与另一个地区的居民通过服饰区分开来。或许，从乐观角度来看这个问题，我们已经将文化和视觉的多样性转换为形形色色的个体生活——或者是一种更多形式的、有趣的知识技术型生活。

在这个很难将两本不同计算机书籍区分开来的时代，Manning 出版社根据 Jefferys 在 3 个世纪前的图画所重现的国家习俗的丰富多样性，设计了计算机系列丛书的封面，以赞美计算机行业的创造性和主动性。

致谢

感谢所有阅读过本书草稿并提出意见的评论者及同行等，尤其是 Aaron Colcord、Aaron Schumacher、Ambikesh Jayal、Bryce Darling、Dwight Barry、Fred Rahmanian、Hans Donner、Jeelani Basha、Justin Fister、Kostas Passadis 博士、Leo Polovets、Marius Butuc、Nathanael Adams、Nezih Yigitbasi、Pablo Vaselli、Peter Rabinovitch、Ravishankar Rajagopalan、Rodrigo Abreu、Romit Singhai、Sampath Chaparala 和 Zekai Otles。他们的意见、质询和修改大大地改善了这本书的质量。特别感谢 George Gaines，他在这本书出版之前对原稿进行了全面的技术审核。

特别感谢开发编辑 Cynthia Kane，感谢她在照料我们写作过程中给予的有益建议和表现出的无比耐心。同样的感谢送给 Benjamin Berg、Katie Tennant、Kevin Sullivan 以及 Manning 出版公司的其他编辑们，他们竭尽全力，消去了书中的粗糙痕迹，剔除了书中的技术瑕疵。

此外，还要感谢我们的同行 David Steier、UC Berkeley 信息科学学院的 Anno Saxenian 教授、Doug Tygar 以及所有其他有意使用本书作为教材的教师。

还要感谢 Jim Porzak，他邀请作者之一 John Mount 到 Bay Area R Users Group 做演讲。他作为本书的热情支持者，还为本书撰写了序言。在我们疲劳、沮丧甚至怀疑我们为什么要承担这一艰苦任务的日子里，他的关注不断地提醒我们：人们需要我们正在做的这件事，也需要我们做这件事的方法。没有他的鼓励，这本书将难以完成。

目　录