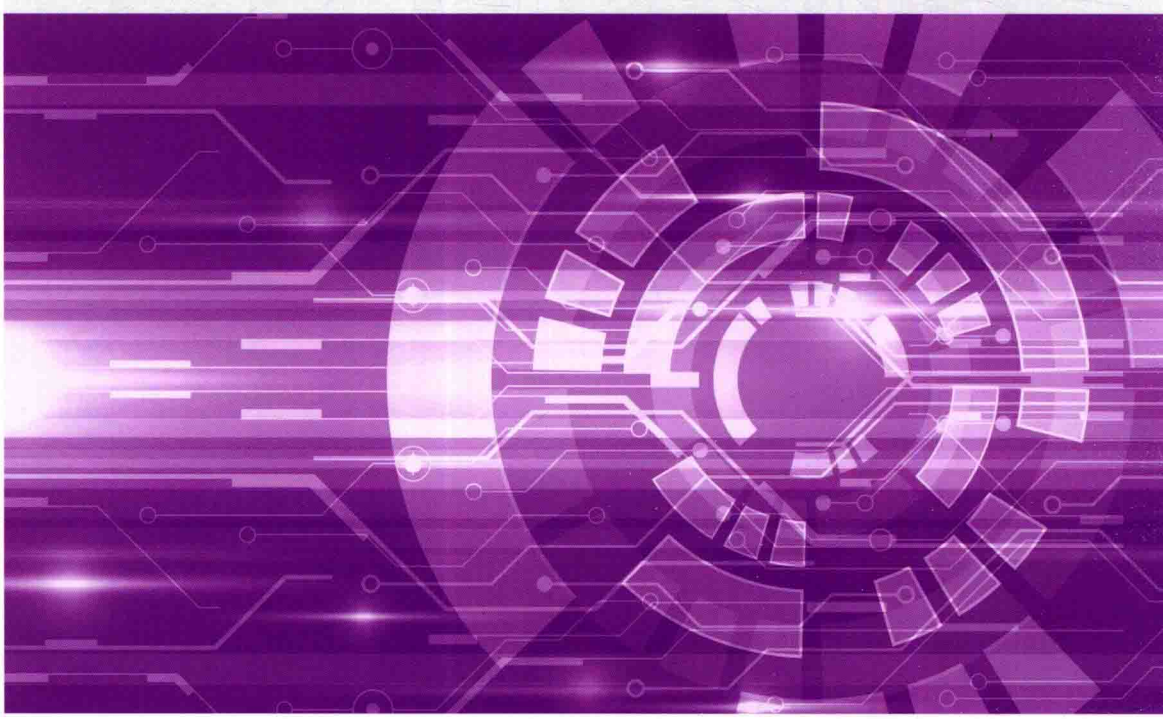
• 大数据应用人才培养系列教材 •

R 语言

■ 总主编◎刘　鹏　张　燕　■ 主编◎程显毅　■ 副主编◎刘　颖　朱　倩



清华大学出版社

大数据应用人才培养系列教材

R语言

总主编　刘　鹏　张　燕
主　编　程显毅
副主编　刘　颖　朱　倩

清华大学出版社
北　京

内 容 简 介

近年来，R 语言可谓是数据分析的热门语言，相关的资料五花八门，让读者难以抉择。本书简洁、精练，以理论与实践相结合的方式让大家快速掌握 R 语言。

全书共 14 章，第 1 章为绪论，从数学、统计学和逻辑学 3 个方面探讨了树立正确数据思维的一些原则。其余各章分为基础篇（第 2～10 章）、应用篇（第 11、12 章）和进阶篇（第 13、14 章）。基础篇按照数据分析过程，主要讨论了 R 的数据结构、数据导入/导出、数据清洗、数据变换、可视化、高级语言编程和常用建模方法。应用篇通过对 2 个经典案例的分析，使读者能够把学到的 R 基础知识应用到解决实际问题中，把数据变成价值。进阶篇解决如何用 R 处理大数据的一些关键技术。

本书可用作培养应用型人才的课程教材，也可作为数据分析爱好者的参考资料。

图书在版编目（CIP）数据

R 语言/程显毅主编. —北京：清华大学出版社，2019
（大数据应用人才培养系列教材）
ISBN 978-7-302-49432-4

Ⅰ. ①R… Ⅱ. ①程… Ⅲ. ①程序语言-程序设计-教材 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字（2018）第 014938 号

责任编辑：贾小红
封面设计：刘　超
版式设计：魏　远
责任校对：赵丽杰
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：16.5　　字　　数：319 千字
版　　次：2019 年 1 月第 1 版　　印　　次：2019 年 1 月第 1 次印刷
定　　价：59.80 元

产品编号：075183-01

总　序

短短几年间，大数据就以一日千里的发展速度快速实现了从概念到落地，直接带动了相关产业的井喷式发展。数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才缺口问题的凸显。根据《人民日报》的报道，未来3～5年，中国需要180万名数据人才，但目前只有约30万人，人才缺口达到150万名之多。

大数据是一门实践性很强的学科，在其呈现金字塔型的人才资源模型中，数据科学家居于塔尖位置，然而该领域对于经验丰富的数据科学家需求相对有限，反而是对大数据底层设计、数据清洗、数据挖掘及大数据安全等相关人才的需求急剧上升，可以说占据了大数据人才需求的80%以上。比如数据清洗、数据挖掘等相关职位，需要源源不断的大量专业人才。

巨大的人才需求直接催热了相应的大数据应用专业。2018年1月18日，教育部公布“大数据技术与应用”专业备案和审批结果，已有270所高职院校申报开设“大数据技术与应用”专业，其中共有208所职业院校获批“大数据技术与应用”专业。随着大数据的深入发展，未来几年申请与获批该专业的职业院校数量仍将持续走高。同时，对于国家教育部正式设立的“数据科学与大数据技术”本科新专业，在已获批的35所大学之外，2017年申请院校也高达263所。

即使如此，就目前而言，在大数据人才培养和大数据课程建设方面，大部分专科院校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，院校尚未形成完善的大数据人才培养和课程体系，缺乏“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学实验工作缺少“原材料”。

对于注重实操的“大数据技术与应用”专业专科建设而言，需要重点面向网络爬虫、大数据分析、大数据开发、大数据可视化、大数据运维工程师的工作岗位，帮助学生掌握大数据技术与应用专业必备知识，使其具备大数据采集、存储、清洗、分析、开发及系统维护的专业能力和技能，成为能够服务区域经济的发展型、创新型或复合型技术技能人才。无论是缺“人”、缺“机制”、缺“机器”，还是缺少“原材料”，最

终都难以培养出合格的大数据人才。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002 年，我与其他专家合作的《网格计算》教材正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（在各大搜索引擎“云计算”关键词中排名名列前茅），2010 年出版了《云计算》，2011 年出版了《云计算》（第 2 版），2015 年出版了《云计算》（第 3 版），每一版都花费了大量成本制作并免费分享了对应的几十个教学 PPT。目前，这些 PPT 的下载总量达到了几百万次之多。同时，《云计算》一书也成为国内高校的优秀教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。

除了资料分享，在 2010 年，我们也在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴和 360 等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长、中国大数据应用联盟人工智能专家委员会主任等。

近几年，面对日益突出的大数据发展难题，我们也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我们于 2013 年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关键词排名中名列前茅；为了解决大数据师资匮乏的问题，我们面向全国院校陆续举办多期大数据师资培训班，致力于解决“缺人”的问题。

2016 年年末至今，我们在南京多次举办全国高校/高职/中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。

其中，为了解决大数据实验难的问题而开发的大数据实验平台，正在为越来越多高校的教学科研带去方便，帮助解决“缺机器”与“缺原材料”的问题。2016 年，我带领云创大数据（www.cstor.cn，股票代码：835305）的科研人员，应用 Docker 容器技术，成功开发了 BDRack 大数据实验一体机，它打破了虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、Storm 集群等，自带实验

所需数据，并准备了详细的实验手册（包含 42 个大数据实验）、PPT 和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。

目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校部署应用，并广受校方好评。该平台也可以云服务的方式在线提供（大数据实验平台：https://bd.cstor.cn），实验更是增至 85 个，师生通过自学，可用一个月时间成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的 AIRack 人工智能实验平台、DeepRack 深度学习一体机以及 dServer 人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职高专院校则更偏向于技术性和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R 语言》《数据清洗》《大数据系统运维》《大数据实践》系列教材，帮助解决“机制”欠缺的问题。

此外，我们也将继续在中国大数据（thebigdata.cn）和中国云计算（chinacloud.cn）等网站免费提供配套 PPT 和其他资料。同时，持续开放大数据实验平台（https://bd.cstor.cn）、免费的物联网大数据托管平台万物云（wanwuyun.com）和环境大数据免费分享平台环境云（envicloud.cn），使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第 7 版，与时俱进日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家，他治学严谨，带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘 鹏

于南京大数据研究院

2018 年 5 月

前　言

随着数据分析的需求不断提升，Excel 渐渐无法满足价值挖掘的日常需求，需要更专业化的软件做数据分析。相应的问题就来了，统计学软件那么多，SPSS、R、Python、SAS、JMP、Matlab 等，该选哪一个？目前市场上较为火热的软件是 R 和 Python。

开源软件的 R 能够迅速发展，很大程度上取决于其活跃的社区和各种 R 包的使用。截至目前（2017 年 2 月 25 日），CRAN（Comprehensive R Archive Network）上已经有 10162 个可以获取的 R 扩展包，内容涉及各行各业，可以适用于各种复杂的统计。

数据科学者的工作就是操纵数据，把原始数据加工成建模需求的形状，而 R 语言是帮你实现整理数据的最佳的工具。

该书深入浅出地介绍 R 语言在大数据分析应用中的相关知识，全书分为绪论（第 1 章）、基础篇、应用篇和进阶篇。基础篇（第 2～10 章）学习如何用 R 完成数据处理，包括数据准备、数据探索、数据变换、数据可视化和数据建模等；应用篇（第 11、12 章）学习如何用 R 完成实际的数据分析报告撰写，包括背景与目标、指标设计、描述性分析、模型分析和结论与建议；进阶篇（第 13、14 章）学习如何使用 R 提高大数据处理性能，包括 RHadoop、SparkR。

绪论从数据、统计学和逻辑学 3 个方面探讨了树立正确的数据思维的一些原则。数据分析师的数据思维对于整体分析思路，甚至分析结果都有着关键性的作用。普通数据分析师与高级数据分析师的主要差异就是有正确的数据思维观。正确的数据思维观与数据敏感度有关，类似于情商类的看不见，摸不着的东西。简单来说，正确的数据思维观是一种通过数据手段解决问题的思维。

基础篇，讨论数据处理的 R 环境，包括 R 数据结构（数据框、列表等）、数据导入/导出、数据清洗（处理数据的缺失值、不一致、异常值）、数据变换（汇总、集成、透视表、规约等）、可视化、高级语言编程、数据分析常用建模方法和原理，涵盖了目前数据挖掘的主要算法，包括分类与预测、聚类分析、关联规则、智能推荐和时序模式，利用可视化数据挖掘包 Rattle 进行试验指导。

应用篇，讨论 2 个经典的数据分析报告案例，通过案例分析使读者

能够把学到的R基础知识应用到解决实际问题中，把数据变成价值。

进阶篇，解决R语言在处理大数据时性能低下的问题，讨论了两个R包：RHadoop、SparkR。

本书特点如下：

（1）知识学习的重点是模型的运用而不是模型的原理。第9章既是R语言的重点也是难点，本书利用可视化数据挖掘包Rattle进行试验指导，简化了建模需要具备的数学基础，只要了解相应模型的函数，设置几个参数就可以轻松完成分类与预测、聚类分析、关联规则、智能推荐和时序模式等数据挖掘任务。

（2）注重数据变成价值。数据分析师工作的最重要一环就是写出有情报价值的数据分析报告。直接将分析结果罗列到PPT或Word中，不仅看上去不美观，而且也会影响报告的可读性，使一份数据分析报告成为简单的数据展示。本书通过案例探讨了写出一份具有情报价值的分析报告的技巧。

（3）关注大数据分析。R语言的最大缺点就是处理大数据的性能较低，无法直接处理TB以上的数据，本书进阶篇讨论的两个R包（RHadoop、SparkR）基本上可以处理任何级别的数据。

（4）向读者提供了书中所用的配套代码、数据及PPT，读者可通过上机实验，快速掌握书中所介绍的R语言的使用方法。

本书由程显毅、刘颖和朱倩负责编写。在本书编写过程中，孙丽丽、季国华、赵丽敏、杨琴和章小华等提供了许多参考资料，在此表示由衷的感谢。由于水平有限，书中可能会有不当之处，希望读者多加指教。本书的编写得到刘鹏教授和清华大学出版社王莉编辑的大力支持和悉心指导，在此深表感谢！

编　者

2018年5月

目　录

第1章

绪　论

1.1　为什么学习R语言

1.1.1　R是什么

R和Python之所以能取得如此的关注，部分原因是大家对其他同类软件的不接受。SPSS的操作可谓“傻瓜级”的，点点鼠标就好了，对编程的要求很弱，与多数人眼中的高级软件有些出入，于是就这样被忽略了。SAS软件是出了名的难安装，在软件安装上就能将一大半的初学者拦在门外，SAS高达8个G的内存占有量，配合着高昂的价格，几乎不适用于个人数据分析。Matlab毕竟不是为专门统计分析而设计的，其他的统计软件相对小众，这样一来，R与Python就因为它们容易安装，编程自由度高的特性脱颖而出。

2008年起，统计之都在中国人民大学举办了第一届中国R语言会议。自此R语言会议（见图1.1），规模越来越大，至今已成功举办了10届。图1.2给出了TIOBE公布的2008年1月编程语言排行榜。

相比于2017年，R是热度增长速度最快的语言，较2017年上升38位（http://www.hangge.com/blog/cache/detail_1925.html）。图1.3是R语言的发展趋势。

图 1.1 R 语言大会会场

Jan 2018	Jan 2017	Change	Programming Language	Ratings	Change
1	1		Java	14.215%	-3.06%
2	2		C	11.037%	+1.69%
3	3		C++	5.603%	-0.70%
4	5	^	Python	4.678%	+1.21%
5	4	v	C#	3.754%	-0.29%
6	7	^	JavaScript	3.465%	+0.62%
7	6	v	Visual Basic .NET	3.261%	+0.30%
8	16	^^	R	2.549%	+0.76%
9	10	^	PHP	2.532%	-0.03%
10	8	v	Perl	2.419%	-0.33%

图 1.2 编程语言排行榜 TOP10 榜单

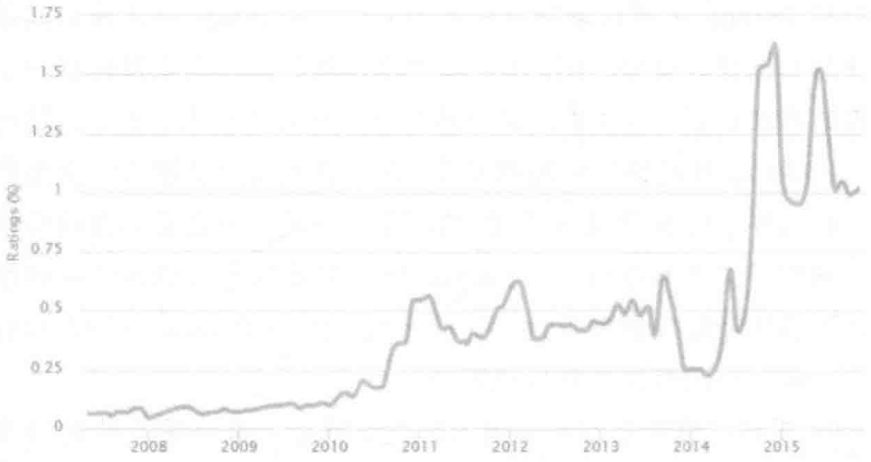


图 1.3 TIOBE Index for R

1.1.2 R 语言主要优势

R 有哪些出色的特征让大家爱不释手呢？

（1）作图美观，完全免费

① R 语言具有卓越的作图功能。既可以画如图 1.4 所示的统计分