

高级大数据人才培养丛书

大数据可视化

BIG DATA

刘鹏 张燕 ○ 丛书主编 何光威 ○ 主编



 中国工信出版集团

 电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

高级大数据人才培养丛书

大数据可视化

丛书主编：刘 鹏 张 燕

主 编：何光威

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书是针对当前大数据应用、可视化分析研究和应用的新形势，专门为大数据专业本科生编写的大数据可视化教材。本书全面诠释了大数据可视化的内涵与外延，详细介绍了大数据可视化概述、可视化的类型与模型、数据可视化基础、数据可视化的常用方法、大数据可视化的关键技术、可视化交互、大数据可视化工具、大数据可视化系统及大数据可视化的行业案例等内容。为了便于学习，每章都附有习题，题型设计也充分考虑到大数据可视化课程教学的需要。

本书适合作为大数据专业本科生和研究生教材，高职高专学校也可以选用部分内容开展教学。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

大数据可视化 / 何光威主编. —北京：电子工业出版社，2018.3
（高级大数据人才培养丛书）

ISBN 978-7-121-33549-5

I. ①大… II. ①何… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 015431 号

策划编辑：董亚峰

责任编辑：杨秋奎

特约编辑：丁福志

印 刷：涿州市京南印刷厂

装 订：涿州市京南印刷厂

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：787×1092 1/16 印张：17.25 字数：420 千字

版 次：2018 年 3 月第 1 版

印 次：2018 年 3 月第 1 次印刷

定 价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zltz@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254755。

编写组

丛书主编：刘鹏 张燕

主 编：何光威

副主编：郑志蕴 梁英杰 朱琼琼

编 者：李 钝 王海涛 马 斌

宋燕燕

基金支持：

金陵科技学院高层次人才科研启动基金（40610186）

国家自然科学基金（61472005）

江苏高校软件工程品牌专业建设工程（PPZY2015B140）

国家自然科学基金（61701517）

国家自然科学基金青年基金（51304078）

江苏省高等学校自然科学研究面上项目（14KJD510011）

中国传媒大学南广学院教改基金

江苏省高校自然科学研究重大项目（16KJA520003）

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井喷式发展。全球多家研究机构统计数据显示，大数据产业将迎来发展黄金期：IDC 预计，大数据和分析市场将从 2016 年的 1300 亿美元增长到 2020 年的 2030 亿美元以上；中国报告大厅发布的大数据行业报告数据也说明，自 2017 年起，我国大数据产业迎来发展黄金期，未来 2~3 年的市场规模增长率将保持在 35% 左右。

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问题的凸显。麦肯锡预测，每年数据科学专业的应届毕业生将增加 7%，然而仅高质量项目对于专业数据科学家的需求每年就会增加 12%，完全供不应求。根据《人民日报》的报道，未来 3~5 年，中国需要 180 万数据人才，但目前只有约 30 万人，人才缺口达到 150 万之多。

以贵州大学为例，其首届大数据专业研究生就业率就达到 100%，可以说“一抢而空”。急切的人才需求直接催热了大数据专业，国家教育部正式设立“数据科学与大数据技术”本科新专业。目前已经有两批共计 35 所大学获批，包括北京大学、中南大学、对外经济贸易大学、中国人民大学、北京邮电大学、复旦大学等。估计 2018 年会有几百所高校获批。

不过，就目前而言，在大数据人才培养和大数据课程建设方面，大部分高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，尚未形成完善的大数据人才培养和课程体系，院校缺“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，开展大数据教学科研工作缺少“原材料”。

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》教材也正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（chinacloud.cn）（在各大大搜索引擎“云计算”关键词中排名第一），2010 年出版了《云计算（第 1 版）》、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首选教材,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、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等。

近几年,面对日益突出的大数据发展难题,我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”关键词排名中位居第一;为了解决大数据师资匮乏的问题,我面向全国院校陆续举办多期大数据师资培训班。2016年年末至今,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。2017年5月,还举办了全国千所高校大数据师资免费讲习班,盛况空前。

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: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85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成功应用,并广受校方好评。该平台也以云服务的方式在线提供(大数据实验平台,https://bd.cstor.cn),帮助师生通过自学,用一个月左右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。

同时,为了解决缺乏权威大数据教材的问题,我所负责的南京大数据研究院,联合金陵科技学院、河南大学、云创大数据、中国地震局等多家单位,历时两年,编著出版了适合本科教学的《大数据》《大数据库》《大数据实验手册》等教材。另外,《数据挖掘》《大数据可视化》《深度学习》《虚拟化与容器》《Python语言》等本科教材也将于近期出版。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,我们还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，目前也已经陆续进入定稿出版阶段。

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，与时俱进且日臻完美，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。他的严谨治学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。我的邮箱：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

前言

本书编写背景及依据

(1) 在电影《侏罗纪世界》中，公园指挥中心对每只恐龙和每名警员的实时位置、身体状况等进行监控，所有监控画面和统计数据均实时呈现。这类系统和指挥中心，可使一切尽在掌握——这就是数据可视化的魅力和价值所在！科幻大片似乎有一种神奇的魔力，总能给观众留下神奇向往的印象，观众也难免会幻想用科技来改变世界。一图胜千言，数据可视化在表述方面胜过千万条枯燥的数据。近年来，虽然数据思维已经深刻影响并改变了各行各业的传统思维方式，但现有的可视化技术还远远无法满足用户的期望。数据密集型科学研究已上升到与科学实验、理论分析、计算模拟并列的科学研究“第四范式”，也拓展了数据可视化学科的内涵和外延，大大推动了数据可视化学科向着更深层次不断迈进。(2) 可视化是一种媒介。当今这个互联网时代，客观世界和虚拟社会正源源不断产生大量的数据，而人类视觉对数字、文本等形式存在的非形象化数据的处理能力远远低于对形象化视觉符号的理解。如果直接面对这些数据，很可能让人无从入手。可视化技术为大数据分析提供了一种更加直观的挖掘、分析与展示手段，有助于发现大数据中蕴含的规律，有助于大数据应用落地。特别是一些监控中心、指挥中心、调度中心等重要场所，大屏幕显示系统已经成为大数据可视化不可或缺的核心基础系统。近年来，全球智能技术发展突飞猛进，发展人工智能已经提升到国家战略高度。大数据可视化在智慧城市、智慧交通、网络安全、航天领域等得到了更加广泛的应用。(3) 可视化是对数据的一种完美的诠释。我们不仅要呈现出数据美好的一面，更重要的是要透过数据，理解数据的丰富内涵，洞察数据中蕴含的奥秘，为更深刻地理解世界、帮助辅助决策提供技术支持，在各行业及生活中，实现用数据说话，用数据讲道理。(4) 数据既可以是抽象的，也可以是异常美丽的。例如，伦敦地铁图、拿破仑进军莫斯科流图及春运迁徙图等，都诠释了数据的美。可视化技术为大数据分析提供了一种更加直观的挖掘、分析与展示手段，有助于发现大数据中蕴含的规律。(5) “可视化”是目的，其技术手段复杂多样。因此，数据可视化的概念应加以扩展，学习数据可视化不应仅仅局限于“可视化”，可视化过程中的数据采集、分析、治理、管理、挖掘等各个技术环节都要融会贯通。(6) 在众多高校陆续开展大数据相关专业的背景下，为向广大读者更好地诠释大数据可视化的内涵与外延，本书应运而生。

本书特色

与国内外同类书比较，本书最大的特点是突出大数据可视化这一艺术与科技融合的

特点,融合数字信号处理、人的视觉特性、可视分析学、大数据可视化渲染等基本理论,展示了大数据技术概貌,构建了大数据可视化的知识逻辑,同时强调实践,具有鲜明的理论与实践并重的特色。本书在强调大数据可视化的基础性原理的同时,融入真实案例分析,注重实用性,使读者真正学会大数据可视化的工具,运用大数据思维,解决实际问题。本书把握大数据可视化应用的趋势,强调多视图整合,强调所有数据视图交互联动,除了原有的饼状图、柱形图、热图、地理信息图等数据展现方式,还可以通过图像的颜色、亮度、大小、形状、运动趋势等多种方式在一系列图形中对数据进行分析,通过交互挖掘数据之间的关联,利用数据推动决策。

教学建议

本书共9章,内容包括大数据可视化概述、可视化的类型与模型、数据可视化基础、数据可视化的常用方法、大数据可视化的关键技术、可视化交互、大数据可视化工具、大数据可视化系统——魔镜及大数据可视化的行业案例。本书可作为大学本科计算机及相关专业数据可视化课程的教材,建议总学时为48学时(理论32学时+实践16学时)。具体分配方案:第1~6章每章4学时;第7~9章每章2学时,加上16学时的实践环节。学期末的2学时安排期末的大作业答辩。授课可采用多媒体投影教学方式,辅以大量的案例分析、视频材料和互动演示。本书的附属资料(电子课件、作业、数据、在线资源、视频和图像)将实时更新。

本书作者团队简介

刘鹏,教授,清华大学博士毕业,现任中国信息协会大数据分会副会长、南京大数据研究院院长、中国大数据应用联盟人工智能专家委员会主任、中国大数据专家委员会委员。

张燕,博士,教授,金陵科技学院副校长。现任江苏省计算机学会常务理事、江苏省人工智能学会常务理事、江苏省农学会智慧农业分会理事长。主持市厅级以上科研项目14项,获江苏省教学成果二等奖、江苏省高教研究成果二等奖各1项。发表论文20多篇,合作出版专著1部,主编教材4部。

何光威,中国传媒大学南广学院教授,高级工程师,智能科学与技术 and 电子信息工程专业负责人,传媒科技研究所常务副所长,中国广播电影电视社会组织联合会技术工作委员会委员。指导学生多次获得江苏省大学生实践创新项目立项,以及江苏省计算机设计大赛特等奖、国家级三等奖。主持及参与省部级科研项目6项、国家广电总局部级社科项目3项、江苏省高校自然科学研究面上项目3项、校级科研项目10多项。获软件著作权1项、国家发明专利1项(合作);出版规划教材、专著7部。

郑志蕴,博士,教授,硕士生导师,郑州大学信息工程学院软件工程系主任,中国计算机学会高级会员、ACM郑州分会秘书长、河南省计算机学会理事、河南省高等学校计算机教育研究会理事。研究方向:云计算、大数据处理、语义网络。参与国家发改委项目2项、国家自然科学基金项目2项、教育部项目1项,主持完成河南省科技攻关项目4项、横向项目10项,主持河南省国际科技合作项目1项。作为第一完成人,通过河南省科学技术厅鉴定项目4项;获得河南省科技进步奖三等奖1项、河南省工业和信息化厅

科技成果一等奖 2 项；被评为“河南省教育系统优秀教师”“郑州市市优秀女科技工作者”和“郑州大学三育人先进个人”。获软件著作权 8 项；发表论文 45 篇，其中 21 篇被 EI 收录；出版学术著作 4 部。

梁英杰，博士，解放军海军工程大学教师。主持国家自然科学基金 1 项、国家重点实验室开放课题 1 项、湖北省基金 1 项，参与“863”等科研项目 10 余项。获专利 2 项；发表论文 20 余篇，其中 12 篇被 EI 收录；参编教材 2 部。

朱琼琼，苏州国云数据科技有限公司教育部营销总监。

李钝，博士，郑州大学信息工程学院副教授，硕士生导师。主持国家社会科学基金 1 项、河南省科技公关项目 2 项、河南省教育厅自然基金 1 项。发表论文 20 余篇，其中 10 篇被 EI 收录，参编教材 2 部。

王海涛，博士，解放军陆军工程大学信息管理中心副教授，硕士生导师，计算机学会高级会员。主持和参与完成国家自然科学基金及军内科研项目 10 余项。获军队科技进步二等奖 1 项、三等奖 2 项。获发明专利 4 项；发表论文 50 余篇，其中 15 篇被 EI 收录；编撰、翻译著作和教材 6 部。

马斌，博士，华北水利水电大学信息工程学院副教授，硕士生导师。已主持国家自然科学基金 1 项、河南省教育厅项目 2 项，参与国家“863”项目 1 项、国家自然科学基金 3 项、国家“十一五”项目 1 项、河南省科技攻关等科研项目 10 余项。申请专利 1 项，获软件著作权 4 项；发表论文 10 余篇，其中 4 篇被 EI 收录；出版专著 1 部，参编教材 2 部。

宋燕燕，硕士，中国传媒大学南广学院数字媒体技术教研室主任。多次指导学生获得江苏省计算机设计大赛特等奖、国家级三等奖。主持及参与校级科研项目 5 项、江苏省高校自然科学研究面上项目 4 项。

读者对象

无论是高等院校相关专业师生，还是数据工程师、数据科学家等，都是本书的受众群体。本书作者来自高校、科研院所和长期从事数据可视化的工程技术人员，从教学、科研和应用的角度，更好地向读者阐述数据可视化，具有很强的系统性、可读性和实用性。经任课老师对教材内容的取舍，本书也可作为高职高专教材和培训教材。

致谢

本书的编写得到了南京大学郭延文教授和南京信息工程大学毕硕本教授的指导，两位教授对全书稿做了认真审查，并提出了许多具体的修改建议。南京云创大数据科技股份有限公司和苏州国云数据科技有限公司工程师参与了本书部分章节的编写工作。南京金陵科技学院赵海峰博士（高工）和南京云创大数据科技股份有限公司武郑浩经理（高工）做了很多协调工作，也对书稿提出了很好的意见。电子工业出版社编辑为本书的出版付出了辛勤的劳动，提出了宝贵的意见。本书也参考了网络上一些高手分享的相关资料，在此一并致谢。

本书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者表示感谢，虽然各位作者通力合作，反复修改，但书中难免有欠妥或者错误之处，敬请读者不吝指正。

何光威
于南京方山

目 录

第 1 章 大数据可视化概述	1
1.1 大数据可视化的概念	1
1.1.1 科学可视化	4
1.1.2 信息可视化	5
1.1.3 数据可视化	5
1.2 数据可视化的作用与意义	5
1.2.1 数据可视化的作用	5
1.2.2 数据可视化的意义	6
1.3 数据可视化的应用领域	9
1.3.1 在“工业 4.0”中的应用	9
1.3.2 在智能交通中的应用	11
1.3.3 在新一代人工智能领域的应用	16
1.3.4 在其他领域的应用	16
1.4 与相关学科的关系	19
1.4.1 与计算机图形学的关系	20
1.4.2 与计算机视觉的关系	20
1.4.3 与计算仿真的关系	20
1.4.4 与人机交互的关系	21
1.4.5 与数据库的关系	21
1.4.6 与数据分析和数据挖掘的关系	21
习题	22
参考文献	23
第 2 章 可视化的类型与模型	24
2.1 可视化的类型	24
2.1.1 科学可视化	24
2.1.2 信息可视化	30
2.2 可视化的模型	39

2.2.1	顺序模型	39
2.2.2	分析模型	40
2.2.3	循环模型	41
习题	42
参考文献	43
第3章	数据可视化基础	45
3.1	光与视觉特性	45
3.1.1	光的特性	45
3.1.2	三基色原理	46
3.1.3	黑白视觉特性	47
3.1.4	彩色视觉特性	52
3.2	可视化的基本特征	55
3.3	可视化流程	56
3.3.1	可视化的基本步骤	56
3.3.2	可视化的一般流程	57
3.4	可视化设计组件	60
3.4.1	可视化设计模型	60
3.4.2	可视化设计原则	60
3.4.3	可视化的数据	61
3.4.4	可视化的原材料	62
3.4.5	可视化的基本图表	62
3.5	可视化中的美学因素	64
3.6	可视化框架设计整体思路	65
3.6.1	可视化框架的构成	66
3.6.2	数据图形映射的流程	66
习题	67
参考文献	68
第4章	数据可视化的常用方法	70
4.1	视觉编码	70
4.1.1	视觉感知	71
4.1.2	视觉通道	72
4.1.3	数据分类	73
4.1.4	常用的复杂数据处理方法	74
4.2	统计图表可视化方法	74
4.2.1	柱状图	75
4.2.2	条形图	77

4.2.3	折线图	78
4.2.4	饼图	79
4.2.5	散点图	79
4.2.6	气泡图	79
4.2.7	雷达图	80
4.3	图可视化方法	80
4.3.1	图的类型	80
4.3.2	图论可视化	81
4.3.3	思维导图	81
4.4	可视化分析方法的常用算法	83
4.4.1	可视化分析方法	84
4.4.2	可视分析研究的特点	85
4.4.3	可视分析的应用实例	86
4.4.4	主成分分析	88
4.4.5	聚类分析	90
4.4.6	因子分析	91
4.4.7	层次分析法	91
4.5	可视化方法的选择	92
4.5.1	百度地图开发	93
4.5.2	城市人流走势	93
4.5.3	商圈人流对比	94
4.5.4	D3.js 和 Echarts 选择上的建议	94
4.5.5	优秀作品欣赏	94
	习题	95
	参考文献	95
第 5 章 大数据可视化的关键技术		97
5.1	大数据架构	97
5.1.1	系统协调者	100
5.1.2	数据提供者	100
5.1.3	大数据应用提供者	100
5.1.4	大数据框架提供者	102
5.1.5	数据消费者	103
5.1.6	安全和隐私	103
5.1.7	管理	103
5.2	大数据核心技术	104
5.2.1	数据收集	104
5.2.2	数据预处理	104

5.2.3	数据存储	105
5.2.4	数据处理	107
5.2.5	数据分析	108
5.2.6	数据治理	110
5.3	可视化关键技术	110
5.4	大数据可视化渲染	112
5.4.1	图像相关概念	112
5.4.2	渲染技术概述	114
5.4.3	基于 CPU 的渲染	115
5.4.4	基于 GPU 的渲染	116
5.4.5	集群渲染技术	118
5.4.6	云渲染	122
	习题	123
	参考文献	124
第 6 章	可视化交互	125
6.1	可视化交互方法分类	125
6.1.1	平移和缩放技术	126
6.1.2	动态过滤技术	127
6.1.3	概览和细节技术	128
6.1.4	焦点和上下文技术	129
6.1.5	多视图关联协调技术	130
6.2	可视化交互空间	131
6.2.1	可视化交互空间查询	131
6.2.2	可视化交互空间分析	132
6.2.3	交互空间分类	133
6.3	可视化交互模型	134
6.3.1	交互式信息可视化的用户界面模型	135
6.3.2	支持信息多面体可视分析界面模型 (IMFA)	138
6.3.3	交互式可视化的关联规则挖掘模型	138
6.3.4	基于 Web 的交互式数据可视化模型	140
6.3.5	基于交互技术的知识可视化模型	142
6.4	交互硬件与软件	145
6.4.1	交互硬件	145
6.4.2	交互软件	147
6.4.3	交互系统	148
	习题	151
	参考文献	152

第 7 章 大数据可视化工具	153
7.1 Excel	153
7.1.1 Power Map 简介	153
7.1.2 Power Map 的使用	153
7.1.3 数据可视化示例	155
7.2 Processing	155
7.2.1 Processing 开发环境简介	156
7.2.2 Processing 绘制功能	156
7.2.3 Processing 应用程序的结构	158
7.2.4 数据可视化示例	159
7.3 NodeXL	161
7.3.1 NodeXL 简介	162
7.3.2 系统界面	162
7.3.3 数据获取与编辑	163
7.3.4 数据可视化	164
7.3.5 图形分析与数据过滤	166
7.4 ECharts	166
7.4.1 ECharts 架构及特点	166
7.4.2 基本组成	167
7.4.3 引入 ECharts	169
7.4.4 图表绘制	169
7.5 Tableau	176
7.5.1 Tableau 简介	176
7.5.2 Tableau 的使用	177
7.5.3 Tableau 数据可视化示例	182
习题	187
参考文献	188
第 8 章 大数据可视化系统——魔镜	189
8.1 魔镜简介	189
8.2 系统架构与技术流程	190
8.3 数据处理与分析	191
8.4 数据可视化	194
习题	199
参考文献	199

第9章 大数据可视化的行业案例	200
9.1 电商行业销售数据分析	200
9.1.1 背景分析	200
9.1.2 需求分析	200
9.1.3 大数据分析过程	200
9.1.4 分析结论	212
9.2 广告投放效果分析	212
9.2.1 背景分析	212
9.2.2 需求分析	212
9.2.3 大数据分析过程	212
9.2.4 分析结论	220
9.3 金融行业贷款数据分析	220
9.3.1 背景分析	220
9.3.2 需求分析	220
9.3.3 大数据分析过程	221
9.4 能源行业油井数据分析	232
9.4.1 背景分析	232
9.4.2 需求分析	233
9.4.3 大数据分析过程	233
习题	245
参考文献	245
附录 A 大数据和人工智能实验环境	246

第1章 大数据可视化概述

数据是抽象的，有时也可以是异常美丽的。可视化技术为大数据分析提供了一种更加直观的挖掘、分析与展示手段，有助于发现大数据中蕴含的规律，在各行各业均得到了广泛的应用。可视化和可视分析利用人类视觉认知的高通量特点，通过图形和交互的形式表现信息的内在规律及其传递、表达的过程，充分结合人的智能和机器的计算分析能力，是人们理解复杂现象、诠释复杂数据的重要手段和途径。数据可视化是大数据的主要理论基础，也是大数据的关键技术，已经成为当前大数据分析的重要研究领域。因此，大数据可视化能力是大数据领域的科学家、工程技术人员的核心竞争力之一。本章在讨论数据可视化的基本概念和起源的基础上，重点讨论可视化在大数据系统中地位，可视化目标和作用，介绍大数据可视化应用领域以及可视化与其他学科的关系。

1.1 大数据可视化的概念

数据科学主要以统计学、机器学习、数据可视化及（某一）领域知识为理论基础，其主要研究内容包括数据科学基础理论、数据预处理、数据计算和数据管理。因此数据可视化（Data Visualization）是数据科学的理论基础之一。数据可视化能将复杂的数据转换为更容易理解的方式传递给受众。数据可视化是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式定义为一种以某种概要形式抽取出来的信息，包括相应信息单位的各种属性和变量。它是一个处于不断演变之中的概念，其边界在不断扩大。数据可视化主要指的是利用图形、图像处理、计算机视觉及用户界面技术，通过表达、建模及对立体、表面、属性、动画的显示，对数据进行可视化解释。与立体建模之类的特殊技术方法相比，数据可视化涵盖的技术方法要广泛得多。

数据可视化借助图形化手段，清晰、有效地传达与沟通信息，但这并不意味着数据可视化就一定因为实现其功能用途而令人感到枯燥乏味，或者是为了看上去绚丽多彩而显得极端复杂。为了有效地传达思想观念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，实现对相当稀疏而又复杂的数据集的深入洞察。然而，设计人员往往不能很好地把握设计与功能之间的平衡，从而创造出华而不实的数据可视化形式，无法达到传达与沟通信息这一主要目的。

数据可视化与信息图形、信息可视化、科学可视化及统计图形密切相关。数据可视化与计算机图形学、计算机视觉等学科相比，既有相同之处，也有显著区别。数据可视化主要是通过计算机图形图像等技术手段展现数据的基本特征和隐含规律，辅助人们更好地认识和理解数据，进而支持从庞杂混乱的数据中获得需要的领域信息和知识。

当前，在大数据的研究、教学和开发领域中，数据可视化是一个极为活跃而又关键