大数据系列丛书

大数据可视化技术

周苏 张丽娜 王文 编著

清华大学出版社

大数据系列丛书

大数据可视化技术

周苏 张丽娜 王文 编著

清华大学出版社
北京

内 容 简 介

这是一个大数据爆发的时代。面对信息的激流，多元化数据的涌现，大数据已经为个人生活、企业经营，甚至国家与社会的发展都带来了机遇和挑战，成为IT信息产业中最具潜力的蓝海。

大数据可视化这种新的视觉表达形式是应信息社会蓬勃发展而出现的——因为我们不仅要呈现世界，更重要的是要通过呈现来处理更庞大的数据，理解各种各样的数据集合，表现多维数据之间的关联。换句话说，就是归纳数据内在的模式、关联和结构。复杂数据可视化既涉及科学也涉及设计，它的艺术性实际上是使用独特手法展示万千世界的某个局部，从而提出问题。大数据可视化，位于科学、设计和艺术三学科的交叉领域(准确地说，应该是位于三个不同维度的人类活动的交叉领域)，蕴藏着无限可能性。

大数据可视化技术是一门理论性和实践性都很强的课程。本教材针对计算机、信息管理和其他相关专业学生的发展需求，系统、全面地介绍了关于大数据技术及其可视化的基本知识和技能，详细介绍了数据可视化之美、Excel数据可视化方法、Excel数据可视化应用、Tableau应用初步、Tableau数据管理与计算、Tableau可视化分析、Tableau地图分析、Tableau预测分析、Tableau仪表板、Tableau故事、Tableau分享与发布以及课程设计与实验总结等内容，共12章。各章均配套设计了导读案例、实验与思考等部分，具有较强的系统性、可读性和实用性。

本书是为高等院校相关专业“大数据可视化”、“数据媒体设计”等课程全新设计编写的，具有丰富实践特色的主教材，也可供有一定实践经验的软件开发人员、管理人员参考和作为继续教育的教材。

与本书配套的教学PPT课件等文档可从清华大学出版社网站的下载区下载，欢迎教师与作者交流并索取为本书教学配套的相关资料：zhousu@qq.com，QQ：81505050，个人博客：http://blog.sina.com.cn/zhousu58。

图书在版编目(CIP)数据

大数据可视化技术/周苏，张丽娜，王文编著. —北京：清华大学出版社，2016
(大数据系列丛书)
ISBN 978-7-302-44978-2

Ⅰ. ①大… Ⅱ. ①周… ②张… ③王… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第216210号

责任编辑：张　玥　薛　阳
封面设计：常雪影
责任校对：梁　毅
责任印制：李红英

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
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京鑫丰华彩印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：16.75　　**彩　插**：6　　**字　数**：415千字
版　　次：2016年11月第1版　　**印　次**：2016年11月第1次印刷
印　　数：1～2000
定　　价：44.50元

产品编号：071417-01

前言

PREFACE

大数据(Big Data)的力量,正在积极地影响着我们社会的方方面面,它冲击着社会的各行各业,同时也正在彻底地改变人们的学习和日常生活。如今,通过简单、易用的移动应用和基于云端的数据服务,人们能够追踪自己的行为以及饮食习惯,还能提升个人的健康状况。因此,有必要真正理解大数据这个极其重要的议题。

然而,仅有数据是不够的。对于身处大数据时代的企业而言,成功的关键还在于找出大数据所隐含的真知灼见。“以前,人们总说信息就是力量,但如今,对数据进行分析、利用和挖掘才是力量之所在。”

大数据可视化这种新的视觉表达形式是应信息社会蓬勃发展而出现的——因为我们不仅要呈现世界,更重要的是要通过呈现来处理更庞大的数据,理解各种各样的数据集合,表现多维数据之间的关联。换句话说,就是归纳数据内在的模式、关联和结构。复杂数据可视化既涉及科学也涉及设计,它的艺术性实际上是使用独特手法展示万千世界的某个局部,从而提出问题。大数据可视化,位于科学、设计和艺术三学科的交叉领域(准确地说,应该是位于三个不同维度的人类活动的交叉领域),蕴藏着无限可能性。

对于在校大学生来说,大数据及其可视化的理念、技术与应用是一门理论性和实践性都很强的“必修”课程。在长期的教学实践中,我们体会到,坚持“因材施教”的重要原则,把实践环节与理论教学相融合,抓实践教学促进理论知识的学习,是有效地改善教学效果和提高教学水平的重要方法之一。本书的主要特色是:理论联系实际,结合一系列了解和熟悉大数据可视化理念、技术与应用的学习和实践活动,把大数据可视化的相关概念、基础知识和技术技巧融入在实践当中,使学生保持浓厚的学习热情,加深对大数据及其可视化技术的兴趣、认识、理解和掌握。

本书是为高等院校相关专业,尤其是计算机、信息管理类专业开设“大数据”相关课程而全新设计编写、具有丰富实践特色的主教材,也可供有一定实践经验的 IT 应用人员、管理人员参考和作为继续教育的教材。

本书系统、全面地介绍了大数据可视化技术的知识和应用技能,详细介绍了数据可视化之美、Excel 数据可视化方法、Excel 数据可视化应用、Tableau 应用初步、Tableau 数据管理与计算、Tableau 可视化分析、Tableau 地图分析、Tableau 预测分析、Tableau 仪表板、Tableau 故事、Tableau 分享与发布以及课程设计与实验总结等内容,共 12 章。各章均配套设计了导读案例、实验与思考等部分,具有较强的系统性、可读性和实用性。

结合课堂教学方法改革的要求,全书设计了课程教学过程,为每章教学内容都针对性地安排了导读案例和课后实验与思考等环节,要求和指导学生在课前、课后阅读课文、网

络搜索浏览的基础上，延伸阅读，深入理解课程知识内涵。

本课程的教学进度设计见课程教学进度表，该表可作为教师授课参考和学生课程学习的概要。实际执行时，应按照教学大纲编排教学进度，按照校历考虑本学期节假日安排来确定本课程的教学进度。

课程教学进度表

（20 —20 学年第 学期）

课程号：_____ 课程名称：大数据可视化技术 学分：2 周学时：2

总学时：34 （其中理论学时（课内）：34 （课外）实践学时：24 ）

主讲教师：_____

序号	校历周次	章节（或实验、习题课等）名称与内容	学时	教学方法	课后作业布置
1	1	引言与第 1 章 数据可视化之美	2	导读案例 课堂教学 实验与思考 课程设计	实验与思考
2	2	第 1 章 数据可视化之美	2		
3	3	第 2 章 Excel 数据可视化方法	2		实验与思考
4	4	第 3 章 Excel 数据可视化应用	2		
5	5	第 3 章 Excel 数据可视化应用	2		实验与思考
6	6	第 4 章 Tableau 应用初步	2		实验与思考
7	7	第 5 章 Tableau 数据管理与计算	2		
8	8	第 5 章 Tableau 数据管理与计算	2		实验与思考
9	9	第 6 章 Tableau 可视化分析	2		实验与思考
10	10	第 6 章 Tableau 可视化分析	2		
11	11	第 7 章 Tableau 地图分析	2		实验与思考
12	12	第 8 章 Tableau 预测分析	2		实验与思考
13	13	第 8 章 Tableau 预测分析	2		
14	14	第 9 章 Tableau 仪表板	2		实验与思考
15	15	第 10 章 Tableau 故事	2		实验与思考
16	16	第 11 章 Tableau 分享与发布	2		实验与思考
17	17	第 12 章 课程设计与实验总结	2		课程设计与实验总结

填表人（签字）： 日期：

系（教研室）主任（签字）： 日期：

本课程的教学评测可以从以下几个方面入手，即：

（1）每章的导读案例（11 次）；

（2）每章的实验与思考（11 次）；

（3）课程设计与实验总结（第 12 章）；

(4) 结合平时考勤；

(5) 任课老师认为必要的其他考核方法。

与本书配套的教学 PPT 课件等文档可从清华大学出版社网站的下载区下载，欢迎教师与作者交流并索取为本书教学配套的相关资料并交流：zhousu@qq.com，QQ：81505050，个人博客：http://blog.sina.com.cn/zhousu58。

本书的编写得到了浙江大学城市学院、浙江商业职业技术学院、浙江安防职业技术学院等多所院校的支持，张健、吴林华、阚晓初、张丽娜等也参与了本书的部分编写工作，在此一并表示感谢！

周　苏

2016 年初夏于西子湖畔

目 录

CONTENTS

第1章

数据可视化之美

【导读案例】

南丁格尔“极区图”

弗洛伦斯·南丁格尔(1820年5月12日出生,1910年8月13日去世,见图1-1)是世界上第一个真正意义上的女护士,被誉为现代护理业之母,5·12国际护士节就是为了纪念她,这一天是南丁格尔的生日。除了在医学和护理界的辉煌成就,实际上,南丁格尔还是一名优秀的统计学家——她是英国皇家统计学会的第一位女性会员,也是美国统计学会的会员。据说南丁格尔早期大部分声望都来自其对数据清楚且准确的表达。

图1-1 南丁格尔

南丁格尔生活的时代各个医院的统计资料非常不精确,也不一致,她认为医学统计资料有助于改进医疗护理的方法和措施。于是,在她编著的各类书籍、报告等材料中使用了大量的统计图表,其中最为著名的就是极区图,也叫南丁格尔玫瑰图(见图1-2)。南丁格尔发现,战斗中阵亡的士兵数量少于因为受伤却缺乏治疗的士兵。为了挽救更多的士兵,她画了这张《东部军队(战士)死亡原因示意图》(1858年)。

这张图描述了1854年4月至1856年3月期间士兵死亡情况,右图是1854年4月至1855年3月,左图是1855年4月至1856年3月,用蓝、红、黑三种颜色表示三种不同的情况,蓝色代表可预防和可缓解的疾病治疗不及时造成的死亡,红色代表战场阵亡,黑色代

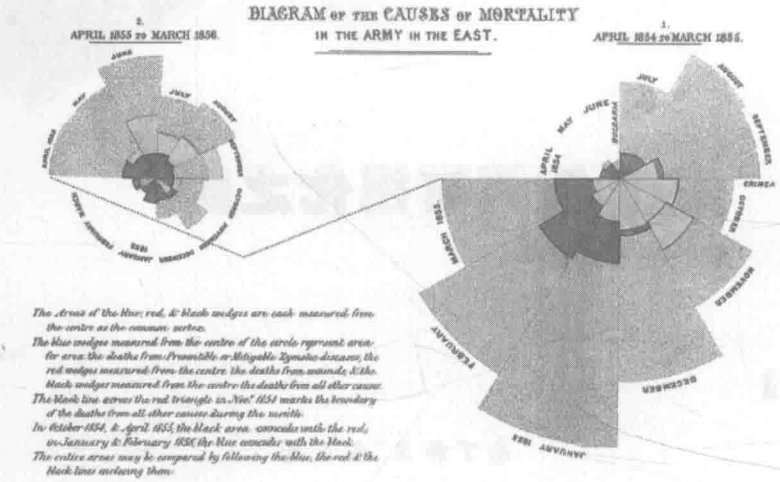

图 1-2　南丁格尔“极区图”

表其他死亡原因。图表各个扇区角度相同，用半径及扇区面积来表示死亡人数，可以清晰地看出每个月因各种原因死亡的人数。显然，1854—1855 年，因医疗条件而造成的死亡人数远远大于战死沙场的人数，这种情况直到 1856 年年初才得到缓解。南丁格尔的这张图表以及其他图表“生动有力地说明了在战地开展医疗救护和促进伤兵医疗工作的必要性，打动了当局者，增加了战地医院，改善了军队医院的条件，为挽救士兵生命做出了巨大贡献”。

南丁格尔“极区图”是统计学家对利用图形来展示数据进行的早期探索，南丁格尔的贡献，充分说明了数据可视化的价值，特别是在公共领域的价值。

图 1-3 是社交网站(Facebook vs. 推特)对比信息图，是一张典型的南丁格尔玫瑰图(极区图)案例。极区图在数据统计类信息图表中是常见到的一类图表形式。

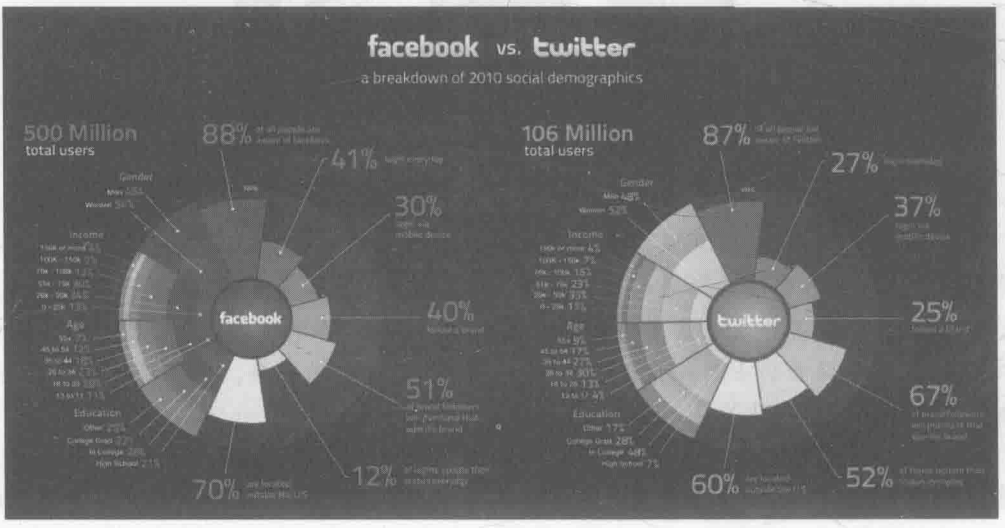

图 1-3　极区图：Facebook vs. 推特

阅读上文，请思考、分析并简单记录：

(1) 你看到过且印象深刻的数据可视化的案例。

答：

(2) 你此前知道南丁格尔吗？你此前是否知道南丁格尔玫瑰图(极区图)？

答：

(3) 发展大数据可视化，那么传统的数据或信息的表示方式是否还有意义？请简述你的看法。

答：

(4) 请简单记述你所知道的上一周发生的国际、国内或者身边的大事。

答：

1.1 数据与可视化

数据是什么？大部分人会含糊地回答说，数据是一种类似电子表格的东西，或者一大堆数字。有点儿技术背景的人会提及数据库或者数据仓库。然而，这些回答只说明了获取数据的格式和存储数据的方式，并未说明数据的本质是什么，以及特定的数据集代表着什么。

1.1.1 数据是什么

要想把数据可视化，就必须知道它表达的是什么。事实上，数据是现实世界的一个快照，会传递给我们大量的信息。一个数据点可以包含时间、地点、人物、事件、起因等因素，因此，一个数字不再只是沧海一粟。可是，从一个数据点中提取信息并不像一张照片那么简单。你可以猜到照片里发生的事情，但如果对数据心存侥幸，认为它非常精确，并和周围的事物紧密相关，就有可能曲解真实的数据。你需要观察数据产生的来龙去脉，并把数

据集作为一个整体来理解。关注全貌，比只注意到局部更容易做出准确的判断。

通常在实施记录时，由于成本太高或者缺少人力，人们不大可能记录下一切，而是只能获取零碎的信息，然后寻找其中的模式和关联，凭经验猜测数据所表达的含义，数据是对现实世界的简化和抽象表达。当你可视化数据的时候，其实是在将对现实世界的抽象表达可视化，或至少是将它的一些细微方面可视化。可视化能帮助人们从一个个独立的数据点中解脱出来，换一个不同的角度去探索它们。

数据和它所代表的事物之间的关联既是把数据可视化的关键，也是全面分析数据的关键，同样还是深层次理解数据的关键。计算机可以把数字批量转换成不同的形状和颜色，但是必须建立起数据和现实世界的联系，以便使用图表的人能够从中得到有价值的信息。数据会因其可变性和不确定性而变得复杂，但放入一个合适的背景信息中，就会变得容易理解了。

1.1.2 数据的可变性

德国物理学家兼业余摄影师克里斯蒂安·克维塞克经常晚上带着相机到小镇的森林里，用长时间曝光摄影，抓拍萤火虫在树丛中飞舞的情景。这种昆虫特别小，在白天几乎看不见，但是在晚上，除了树林里，又很难在别的地方看到。

虽然对观察者来说，萤火虫飞行中的每个时刻都像是空间中随机的点，但克维塞克的照片中还是出现了一个模式。如图 1-4 所示，看上去萤火虫们好像沿着小径，环绕着大树，朝既定的方向飞舞。

图 1-4 萤火虫之路(http://quit007.deviantart.com/)

然而，这些依然是随机的。下一次你可以根据这条飞行路线图猜测萤火虫会往哪儿飞吗？一只萤火虫随时上下左右地飞蹿，这种变化使得萤火虫的每次飞行都是独一无二的。也正因为如此，观察萤火虫才那么有趣，拍出来的照片才那么漂亮。你关心的是萤火虫飞行的路径，而它们的起点、终点和平均位置并没有那么重要。

从这些数据中，可以发现一些模式、趋势和周期，但从 A 点到 B 点往往都不是一条平滑的线路(实际上，几乎从来都不是)。总数、平均值和聚合测量可能很有趣，但它们都只揭示了冰山一角而已。数据中的波动才是最有趣、最重要的部分。