普通高等教育“十二五”规划教材

高等院校重点推荐教材

数据可视化的基本原理与方法

陈 为 张嵩 鲁爱东 编著

彭群生 主审

科学出版社

普通高等教育“十二五”规划教材

高等院校重点推荐教材

数据可视化的基本原理与方法

陈　为　张　嵩　鲁爱东　编著

彭群生　主审

科学出版社

北　京

内 容 简 介

本书是面对当前科学可视化、信息可视化、可视分析研究和应用的新形势，专门为计算机、数据处理、视觉设计、统计、数学、航空航天、建筑、遥感影像等专业本科生开设数据可视化课程而编写的一本教材。全书划分为三篇：基础篇、数据篇和应用篇。其中，基础篇从人、数据、可视化流程等三个层面阐述数据可视化的基础理论和概念；数据篇则针对实际应用中遇到的不同类型的数据，包括时空数据、地理信息数据、高维非空间数据、层次和网络数据介绍相应的可视化方法；应用篇着重介绍可视化综合应用及实用系统。为了便于学习，每章后都附有习题和参考文献。

本书的特点是内容完整，叙述简明，重点突出；以数据类型为导向，以行业应用为目标。作者专门收集和整理了相关的课程教案、典型数据、精彩案例、可视化作品、课程附属视频和动画材料，将在课程网站 http://www.cad.zju.edu.cn/home/vagblog 上发布并实时更新。

本书可作为高等院校计算机、数据处理及相关专业高年级学生和研究生的教学用书，对于从事数据可视化、数据分析、视觉艺术开发和应用的科技人员也有较大的参考价值。

图书在版编目(CIP)数据

数据可视化的基本原理与方法/陈为，张嵩，鲁爱东编著. —北京：科学出版社，2013

（普通高等教育“十二五”规划教材·高等院校重点推荐教材）

ISBN 978-7-03-037488-2

Ⅰ.①数… Ⅱ.①陈…②张…③鲁… Ⅲ.①可视化软件-高等学校-教材 Ⅳ.①TP31

中国版本图书馆 CIP 数据核字（2013）第 101398 号

责任编辑：鞠丽娜/责任校对：柏连海

责任印制：吕春珉/封面设计：三函设计

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

http://www.sciencep.com

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2013 年 6 月第 一 版 开本：B5（720×1000）

2013 年 6 月第一次印刷 印张：19 3/4

字数：361 000

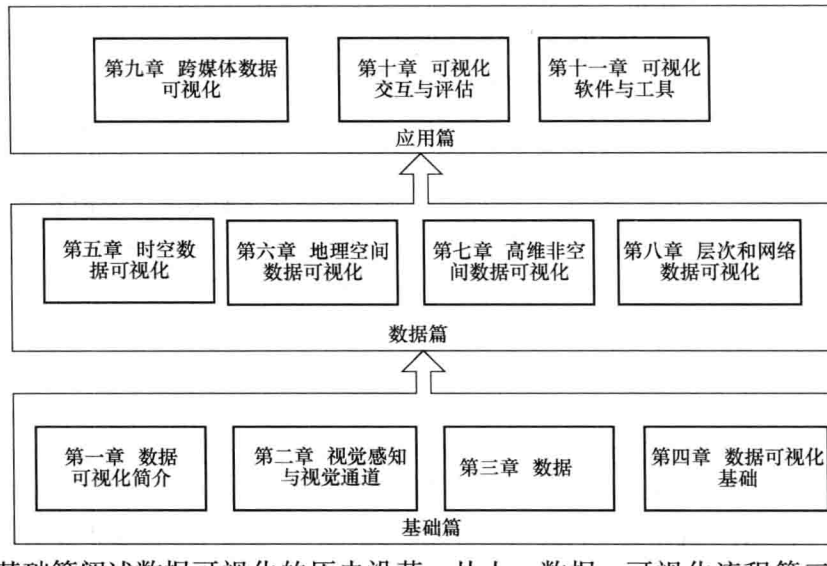
定价：36.00 元

（如有印装质量问题，我社负责调换<双青>）

销售部电话 010-62134988 编辑部电话 010-62135763-2032

前　言

数据是记录科学现象和客观世界的基本信息单元。当今世界每时每刻都在产生大量的数据。数据可视化为人类洞察数据的内涵、理解数据蕴藏的规律提供了重要的手段。我国对数据可视化的研究始于1990年初，部分高校开设了面向研究生的科学可视化课程，一些经典的科学可视化教材相继出版，例如，1996年浙江大学石教英教授等编写的《科学计算可视化》、1999年清华大学唐泽圣教授等编写的《三维数据场可视化》等，为我国培养可视化研究和应用方面的专业人才发挥了重要的作用。进入21世纪，人类获取数据的能力得到迅猛发展。据统计，2010年世界新增信息总量为988艾（10^{18}）字节，该数量以每年至少30%的速率增长。随着数据时代的来临，针对实际科学和社会问题，涌现了许多新的复杂数据的处理与分析方法。面对当前科学可视化、信息可视化、可视分析研究和应用的新形势，高校急需一本新的数据可视化教材。本书正是在这样一个背景下，基于作者多年从事大学本科可视化课程建设和科学研究的积累编撰而成。全书分为三篇，共11章，组织结构分布如下图所示。



基础篇阐述数据可视化的历史沿革，从人、数据、可视化流程等三个层面阐述数据可视化的基础理论和概念。其中，第一章阐述可视化的定义、作用和发展历史，给出了数据可视化的分类，并简略描述了数据可视化设计框架和可

视化流程。第二章详细介绍视觉感知和认知的基本原理和可视化编码原则。数据模型在第三章中介绍，包括数据定义、组织、管理、分析、挖掘等及数据工作流。第四章介绍可视化的基础理论，包括可视化流程、图形符号、视觉变量和评估方法等内容。

根据数据的时空特性，可粗略地将数据分为时空数据和非时空数据。数据篇的第一部分讲述含有空间或时间信息的可视化方法。其中，第五章介绍空间标量场数据可视化，包括一维、二维或三维真实物理空间的标量场数据的可视化方法。当空间数据具有真实的地球地理空间坐标时，需要采用特定的地理信息可视化方法，第六章详细介绍面向地理信息的基于点、线和区域的可视化方法及其应用。

数据篇的第二部分介绍复杂多变量、非结构化、非几何的抽象数据的可视化方法，这类数据来源于真实物理空间、人类社会空间和网络信息空间，例如，物种遗传关系、家谱和社交网络。第七章将介绍复杂多变量数据的可视化方法，主要采用视觉通道编码、降维、变量选择、交互等方法减少复杂多变量的非时空数据可视化产生的视觉混淆对有用信息的干扰。注意到大部分非时空数据具有层次结构关系，例如公司组织结构、社交网络用户关系，第八章详细描述层次结构数据的可视化方法，包括树、图、超图、网络等。

应用篇着重介绍面向不同类型数据的可视化综合应用及实用系统。其中，第九章介绍跨媒体数据（文本、图像、音乐、视频等数据的综合）的可视化方法；第十章介绍可视化中的人机交互方法，包括交互准则、交互的分类和相关评测技术；第十一章介绍可视化系统，包括应用系统、开发工具和研究小组等。

本书可作为大学本科计算机或相关专业的三、四年级学生的数据可视化课程的教科书，课程总学时是 32。具体分配方案是：第四、五、七章每章 4 个学时；其余各章，每章 2 个学时；学期末的 4 个学时安排期末的大课程答辩。授课可采用多媒体投影教学方式，辅助以大量的案例分析、视频材料和互动演示。课程的实践可采用 Processing（http://www.processing.org）和 VTK（http://www.vtk.org）。

本书的附属资料（电子课件、作业、数据、在线资源、视频和图像）和修订信息将在 http://www.cad.zju.edu.cn/home/vagblog 上予以实时更新。

本书由陈为（浙江大学）、张嵩（美国密西西比州立大学）、鲁爱东（美国北卡罗莱纳大学夏洛特分校）合作撰写。陶煜波（浙江大学）、陈伟锋（浙江财经学院）、吴向阳（杭州电子科技大学）及浙江大学 CAD&CG 国家重点实验室部分师生参与了初稿的写作。各章初稿的撰写人是：第一章陈为、陶煜波；第二章陈伟锋；第三章吴向阳、马昱欣；第四、五、十一章张嵩，第 4.3、4.4 节陈伟锋；第六、七、十章鲁爱东，第 10.2 节张嵩；第八章夏菁；第九章黄芯芯、

马晓红和彭帝超。彭群生教授参与了本书的策划，对全书稿做了认真审查，并提出了许多具体的修改建议。

本书编辑鞠丽娜女士对此书的出版给予了大力支持。浙江大学CAD&CG国家重点实验室可视分析小组的朱斯衍、张建霞、邹瑶瑶等同学参与了书稿的准备、讨论和校对。杭州电子科技大学刘真博士帮助审校了部分章节，在此一并致谢。

由于时间仓促，作者水平有限，书中难免存在欠妥之处，敬请读者不吝指正。

作　者

2013年4月

目　录

第一篇　基　础　篇

第二篇　数　据　篇

第三篇　应　用　篇

第一篇　基　础　篇

第一章　数据可视化简介

1.1　可视化的意义

视觉是人类获取外部世界信息的最重要通道。人眼是一个高带宽的并行视觉信号输入处理器，带宽高达每秒 100 兆字节，且具有很强的模式识别能力，对可视符号的感知速度比对数字或文本快多个数量级，且大量的视觉信息处理发生在潜意识阶段。超过 50%的人脑机能都用于视觉感知，包括解码可视信息、高层次可视信息处理和思考可视符号［Ward，2010］。

在计算机学科的分类中，对数据进行交互的可视表达以增强认知的技术，称为可视化。它将不可见或难以直接显示的数据映射为可感知的图形、符号、颜色、纹理等，增强数据识别效率，高效传递有用信息［Hansen，2004］。

可视化的终极目标是洞悉蕴含在数据中的现象和规律，这包含多重含义：发现、决策、解释、分析、探索和学习［Ward，2010］。可视化的一个简明定义是“通过可视表达增强人们完成某些任务的效率”。表 1.1 列出 4 个二维数据点集，它们的单维度均值、最小二乘法回归线方程、误差的平方和、方差的回归和、均方误差的误差和、相关系数等统计属性均相同，无法从统计属性中获得这 4 组数据的差异信息。将实际的数据分布情况用二维可视化呈现（图 1.1），则观察者可迅速从数据中寻找模式和规律。

表 1.1　四组二维数据点集，它们的均值、方差和相关系数均相同

第一组		第二组		第三组		第四组	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.1	14.0	8.84	8.0	7.04

续表

第一组		第二组		第三组		第四组	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.1	4.0	5.39	19.0	12.5
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
均值 9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
方差 10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
相关系数	0.816		0.816		0.816		0.816

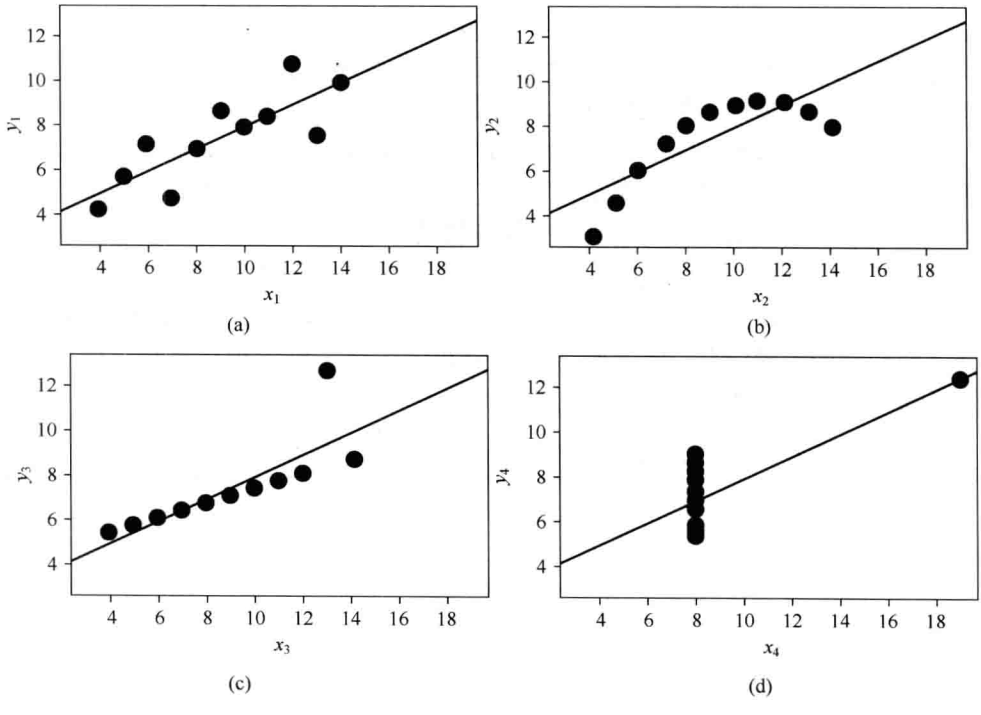


图 1.1 Anscombe 实验的四个二维数据点集的可视化

(a) x_1 数据；(b) x_2 数据；(c) x_3 数据；(d) x_4 数据

图片来源：[Anscombe，1973]

人类执行高效视觉搜索的过程通常只能保持几分钟，无法持久。从信息加

工的角度看，丰富的信息消耗了大量的注意力，可视化作为某种外部内存，在人脑之外保存待处理信息，可补充人脑有限的记忆内存，有助于解决人脑的记忆内存和注意力的有限性的问题。同时，图形化符号可将用户的注意力引导到重要的目标，可高效地传递信息。

1.2　可视化的目标和作用

根据信息传递方式，传统的可视化方法可以大致分为两大类，即探索性可视化和解释性可视化。前者指在数据分析阶段，不清楚数据中包含的信息，希望通过可视化快速地发现特征、趋势与异常，这是一个将数据中的信息传递给可视化设计与分析人员的过程。后者指在视觉呈现阶段，依据已知的信息或知识，以可视的方式将它们传递给公众。

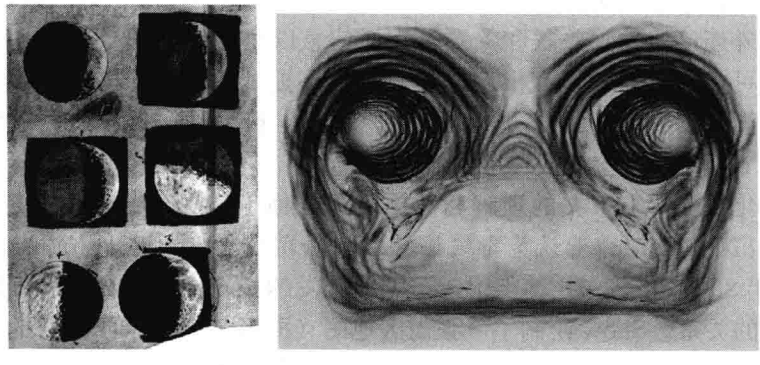
从应用的角度来看，可视化有多个目标：有效呈现重要特征、揭示客观规律、辅助理解事物概念和过程、对模拟和测量进行质量监控、提高科研开发效率、促进沟通交流和合作等。

从宏观的角度看，可视化的三个功能包括：

(1) 信息记录

将浩瀚烟云的信息记录成文，世代传播的有效方式之一是将信息成像或采用草图记载。图 1.2 (a) 展示了可视化的鼻祖之一伽利略的手绘月亮周期可视化图。图 1.2 (b) 展示了流体动力学模拟计算的三维空间数据场的可视化，揭示出原本不可见的飞行器尾部的气流旋涡。

不仅如此，可视化图示能极大地激发智力和洞察力，帮助验证科学假设。



(a)　　(b)

图 1.2　草图记载

(a) 伽利略于 1616 年关于月亮周期的绘图；(b) 基于流体动力学模拟计算数据的飞行器尾部的气流旋涡

例如，20 世纪自然科学最重要的三个发现之一，DNA 分子结构的发现起源于对 DNA 结构的 X 射线衍射照片［图 1.3（a)］的分析：从图像形状确定 DNA 是双螺旋结构，且两条骨架是反平行的，骨架是在螺旋的外侧等［图 1.3（b)］这些重要的科学事实。

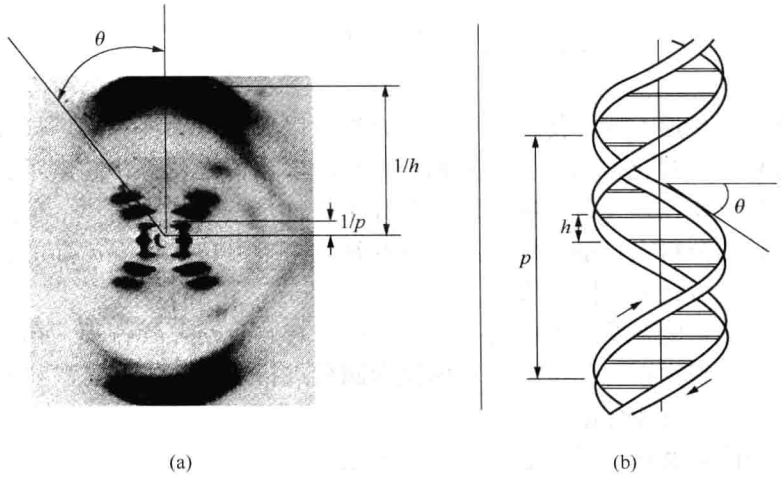


图 1.3 DNA 的分子结构

（a）DNA 的 B 形 51 号 X 射线衍射照片；（b）DNA 的 X 射线衍射照片与双螺旋结构的晶体学解释，为螺旋倾斜角，$h=3.4$ 为基距离，$p=34$ 为螺旋结构单元长度

图片来源：http://home.sandiego.edu/—cloer/bio482/482qanda.html

（2）信息推理和分析

数据分析的任务通常包括定位、识别、区分、分类、聚类、分布、排列、比较、内外连接比较、关联和关系等。将信息以可视的方式呈现给用户，可导引用户从可视化结果分析和推理出有效信息，提升信息认知的效率。这种直观的信息感知机制突破了常规分析方法的局限性，极大地降低了数据理解的复杂度。图 1.4 展示了两个图形化计算的例子。

可视化在支持上下文的理解和数据推理方面也有独到的作用。1831 年起，欧洲爆发霍乱，当时普遍认为毒气瘴气引起了霍乱。英国医生 John Snow 为研究 1854 年 8 月底伦敦布拉德街附近居民区爆发的一场霍乱，调查病例发生的地址和取水的关系。Snow 绘制了一张街区地图（图 1.5），标记了水井的位置，每个地址的病例用条码显示。条码清晰显示出 73 个病例集中分布在布拉德街的水井附近，这就是著名的鬼图（ghost map）。在拆除布拉德街水井的摇把后不久，霍乱停息。

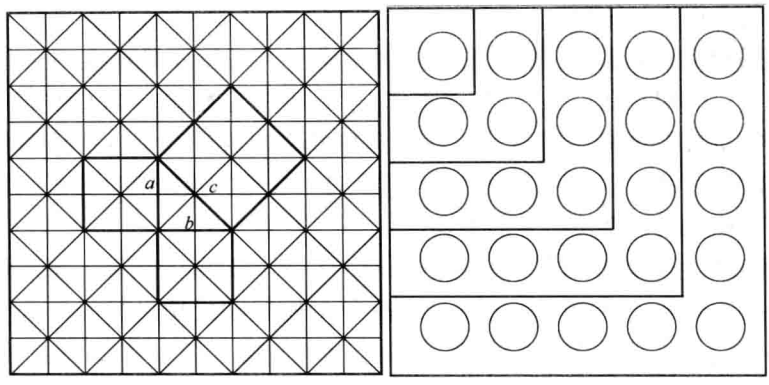


图 1.4　图形化计算例子

(a) 中国古代用于证明勾股定理的图形化不同区域的圆的个数正好是 1，3，5，7，9，它们证明方法：$c^2=a^2+b^2$。从包含的三角形的总和正好是 5 行 5 列的总数目；(b) 奇数和的可视化：1+3+5+7+9 = 25，个数可以直观看出，两个小正方形的面积等于大正方形的面积

图 1.5　伦敦鬼图

图片来源：http://www.datavis.ca/gallery/historical.php