B 大数据丛书
IG DATA SERIES

Apress®

Pro Data Visualization using R and JavaScript

采用R和JavaScript的数据可视化

【美】汤姆·巴克 TOM BARKER 著

刘小虎 邢静 程国建 译

机械工业出版社
CHINA MACHINE PRESS

大数据丛书

采用 R 和 JavaScript 的数据可视化

【美】汤姆·巴克（Tom Barker）著
刘小虎　邢　静　程国建　译

机械工业出版社

Pro Data Visualization using R and JavaScript
By Tom Barker, ISBN: 978-1-4302-5806-3
Original English language edition published by Apress Media.

图书在版编目(CIP)数据

采用 R 和 JavaScript 的数据可视化/（美）汤姆·巴克（Tom Barker）著；刘小虎，邢静，程国建译. —北京：机械工业出版社，2019.1
（大数据丛书）
书名原文：Pro Data Visualization using R and JavaScript
ISBN 978-7-111-62015-0

Ⅰ. ①采… Ⅱ. ①汤… ②刘… ③邢… ④程… Ⅲ. ①程序语言—程序设计②JAVA 语言—程序设计 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字（2019）第 029208 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：王 康 责任编辑：王 康 刘丽敏
责任校对：肖 琳 封面设计：陈 沛
责任印制：张 博
北京铭成印刷有限公司印刷
2019 年 4 月第 1 版第 1 次印刷
169mm×239mm · 13 印张 · 246 千字
标准书号：ISBN 978-7-111-62015-0
定价：49.80 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务
服务咨询热线：010-88361066
读者购书热线：010-68326294

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
金 书 网：www.golden-book.com
教育服务网：www.cmpedu.com

封面无防伪标均为盗版

本书使得日益流行的 R 语言变得平易近人，并促成数据采集和分析理念变为现实。本书介绍如何使用 R 来查询和分析数据，使用 D3 JavaScript 函数库以优雅、富有信息和交互的方式来格式化并显示数据。您将学会如何有效地收集数据、如何理解每种类型图表的方式理念及其实现，并能直观地呈现结果。

本书可作为高校计算机类本科相关课程的教学参考书，也可作为人工智能、机器学习、数据科学等应用系统开发者的参考资料。

致 谢

我要感谢 Ben Renow-Clarke 考虑我承担这个大的项目。我要感谢 Matthew Moodie 和 Christine Rickets 以及 Apress 团队其他成员的指导和帮助。我要感谢 Matt Canning，他帮助我以新的眼光看待程序代码并让我保持诚实。

我要感谢我所在的 Comcast 团队：你们每个人都很棒。这个团队让我变得更好。我要感谢我美丽的妻子 Lynn 和我们漂亮的孩子 Lukas 和 Paloma，他们对我的写作过程给予了耐心和理解。

译 者 序

数据可视化是指采用较为高级的技术方法，利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。数据可视化与信息表征、信息可视化、科学可视化以及统计图形密切相关。在研究、教学和开发领域，数据可视化是一个极为活跃且非常关键的研究领域。特别是在大数据领域，数据可视化工具和技术对于分析大量信息和制定数据驱动决策至关重要。

数据可视化的相关领域包括：数据采集（对现实世界进行采样，以便产生可供计算机处理的数据的过程）、数据分析（为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程）、数据治理（为特定组织机构的数据创建协调一致的企业级视图所需的人员、过程和技术）、数据管理（所有与管理作为有价值资源的数据相关的学科领域）以及数据挖掘（对大量数据加以分类整理并挑选出相关信息的过程）。

随着“大数据时代”进入商业化阶段，可视化越来越成为了解数据的关键工具。数据可视化有助于通过将数据整理成易于理解的形式来讲述故事，突出显示趋势、模式和异常值。良好的可视化可以清晰呈现数据背后的故事，消除数据中的噪声并突出显示有用的信息。数据和视觉效果需要协同工作，才可将细致的分析与精彩的故事情节结合起来，所以数据可视化也可以说是一门艺术。

本书涵盖 R 语言的基本概念与编程方法、基于 R 与 JavaScript 实现的空间与时序数据可视化实现、讲解条形图与散点图的实现技巧、追求速度和质量平衡的可视化技术等。通过本书您将学会如何使用 R 来查询和分析数据，然后使用 D3 JavaScript 以信息聚集和交互的方式来格式化并显示数据。您将学会如何有效地收集数据、如何理解每种类型图表的方式理念及其实现方法，并直观地呈现出数据背后的洞察效果。

本书的出版得益于机械工业出版社的大力支持与帮助，在此深表谢意。同时也感谢译者们及其研究生的协同努力。

本书的翻译出版得到西安培华学院学术基金的支持，在此表示感谢。

译 者

目 录

第1章 背 景

在互联网发展领域中出现了一个新概念：使用数据可视化作为交流工具。这个概念某种程度上已经在其他领域和部门很好地确立。在公司中，财务部门可能使用数据可视化来表示内部和外部的财务信息；仅仅看看季度收益报告，几乎所有上市公司都是这样做的。这些报告中充满大量图表来显示季度收入、年度收益或者其他历史金融数据，所有这些简单且易于理解的图表设计都是为了展示大量的数据点和潜在的大量数据点。

将 Google 2007 年 Q4 季度收益的条形图（见图 1-1）和表格形式的子集数据（见图 1-2）做比较。

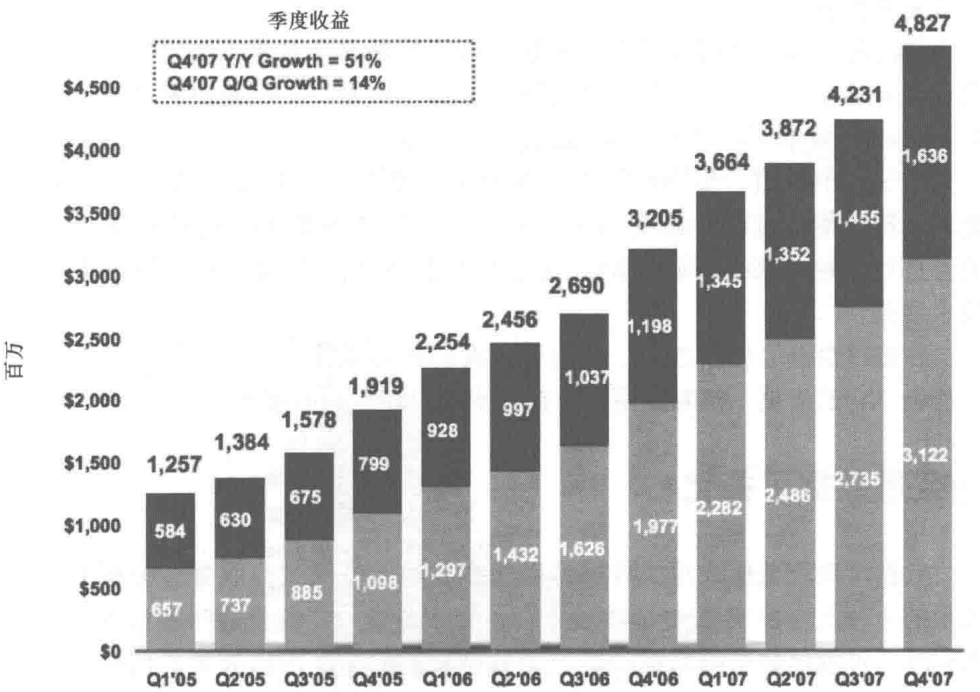

图 1-1　用条形图展示的 Google 2007 年 Q4 季度收益

相比而言，条形图更容易读懂。通过观察条形图的形状可以清楚地看到收益在上升，并且每个季度都在稳步上升。通过颜色标注，可以看到收入的来源；通过注释可以看到，这些颜色标注所代表的精确数字和其每年所占的百分比。

	Class A and Class B Common Stock		Additional Paid-In Capital	Deferred Stock-Based Compensation	Accumulated Other Comprehensive Income	Retained Earnings	Total Stockholders' Equity
	Shares	Amount	Amount	Amount	Amount	Amount	Amount
Balance at January 1, 2005	266,917	$ 267	$ 7,582,352	$ (249,470)	$ 5,436	$ 580,471	$ 2,929,056
Issuance of common stock in connection with follow-on public offering and acquisitions, net	14,869	15	4,316,022	(2,036)	—	—	4,314,001
Stock-based award activity	11,241	11	579,418	132,491	—	—	711,920
Comprehensive income:							
Change in unrealized gain (loss) on available-for-sale investments, net of tax effect of $11,404	—	—	—	—	16,580	—	16,580
Foreign currency translation adjustment	—	—	—	—	(17,997)	—	(17,997)
Net income	—	—	—	—	—	1,465,397	1,465,397
Total comprehensive income	—	—	—	—	—	—	1,463,980
Balance at December 31, 2005	293,027	293	7,477,792	(119,015)	4,019	2,055,868	9,418,957
Issuance of common stock in connection with follow-on public offering and acquisitions, net	7,689	8	3,236,778	—	—	—	3,236,786
Stock-based award activity	8,281	8	1,168,336	119,015	—	—	1,287,359
Comprehensive income:							
Change in unrealized gain (loss) on available-for-sale investments, net of tax effect of $13,280	—	—	—	—	(19,309)	—	(19,309)
Foreign currency translation adjustment	—	—	—	—	38,601	—	38,601
Net income	—	—	—	—	—	3,077,446	3,077,446
Total comprehensive income	—	—	—	—	—	—	3,096,738
Balance at December 31, 2006	308,997	309	11,882,906	—	23,311	5,133,314	17,039,840
Stock-based award activity	3,920	4	1,358,315	—	—	—	1,358,319
Comprehensive income:							
Change in unrealized gain (loss) on available-for-sale investments, net of tax effect of $19,963	—	—	—	—	29,029	—	29,029
Foreign currency translation adjustment	—	—	—	—	61,033	—	61,033
Net income	—	—	—	—	—	4,203,720	4,203,720
Total comprehensive income	—	—	—	—	—	—	4,293,782
Adjustment to retained earnings upon adoption of FIN 48	—	—	—	—	—	(2,262)	(2,262)
Balance at December 31, 2007	312,917	$ 313	$ 13,241,221	$ —	$ 113,373	$ 9,334,772	$ 22,689,679

图 1-2　以表形式展示的相似收益数据

使用表格数据必须看左侧的标签，根据这些标签将右侧数据进行排序，再做分类和对比，然后才能得出结论。使用表格数据获取信息，需要做大量的前期工作，否则很有可能读者并不理解这些数据（因而对数据产生错误的理解）或者完全误解。

不仅财务部门使用可视化来传达大量的数据，有时业务部门也使用图表来发布服务时段，或者客户部门使用图表显示呼叫量。不管是哪种情况，在工程类和 Web 开发中广泛使用可视化已是大势所趋。

对于某个部门、集团和行业，有大量重要的相关数据需要第一时间搞清楚，这样才能改进和提高我们所做的事情；同样也需要将这些数据与利益相关者进行沟通，来表明我们的成功或者验证资源的需求，或者制定下一年的战术线路。

在进行数据可视化之前，必须了解我们正在做什么，需要了解数据可视化是什么、它的历史、何时使用以及在技术和伦理上如何使用。

什么是数据可视化？

数据可视化是什么呢？数据可视化是对经验信息进行收集、分析并图表化表示的一种艺术和实践。有时它被称为信息图形，或者只是叫表或图。不管怎么称呼，可视化数据的目标就是说出数据里的故事。说出故事的前提是在更深层次上理解数据，并通过比较数字里的数据点获得更深入的见解。

下文介绍数据可视化的一些语法和图表形式的模式等概念。本书中每一个重要的图表类型都用一章来介绍。

时间序列表

时间序列表显示数据随时间的变化。如图 1-3 所示的时间序列表，表示的是

Google Trends 中关键词“数据可视化”受欢迎的权重（http：//www. google. com/trends）。

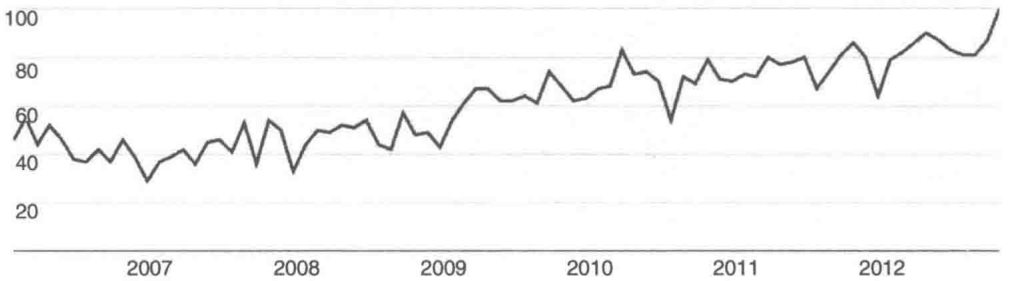

图 1-3 Google Trends 中关键词“数据可视化”的权重趋势时间序列图

注意：垂直的 y 轴表示数字的顺序，从 20 增加到 100。这些数字代表搜索值的权重，其中 100 是测试的峰值搜索值。水平的 x 轴，从 2007 年到 2012 年。这个图表中的曲线代表两个坐标轴所给出每一个数据的搜索值。

在这个小样本范围内，我们可以看到搜索值从 2007 年开始的 29 增长到 2012 年的上限 100，这个术语的受欢迎度已经超过了 3 倍。

条形图

条形图展示出数据点之间的对比。如图 1-4 所示的条形图，显示了不同国家对关键词“数据可视化”的搜索值，这个数据也来源于 Google Trends。

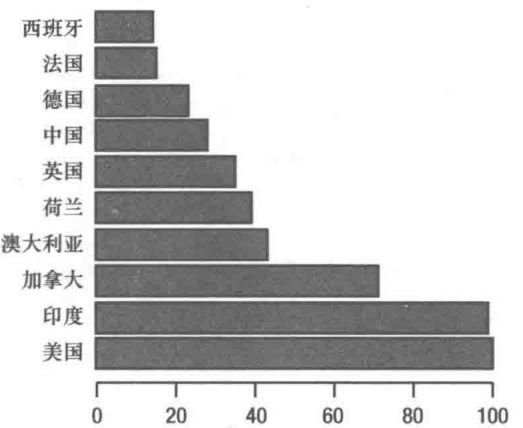

图 1-4 按关键字“数据可视化”搜索并按区域展现的 Google 趋势图

y 轴表示的是国家的名字，x 轴表示的是标准的搜索值，范围为 0 ~ 100。需要注意的是，图中没有给出时间的标准。那么这个表格所表示的到底是 1 天，1

个月，还是 1 年？

同时需要注意，背景中没有给出度量单位是什么。强调这一点不是要去确定它们，而是表明这种特殊图表类型的局限和陷阱。我们必须清楚读者并不具有和我们一样的经验和背景，所以必须努力使可视化中的故事尽可能清楚地显现出来。

直方图

直方图是条形图的一种，经常用来展示数据的分布或在数据中各组信息出现的次数。如图 1-5 所示的直方图是从 1980 年到 2012 年纽约时报每年发表的与数据可视化学科相关的文章的数量。从图表中可以看到从 2009 年起这个学科发表文章的频率一直在上升。

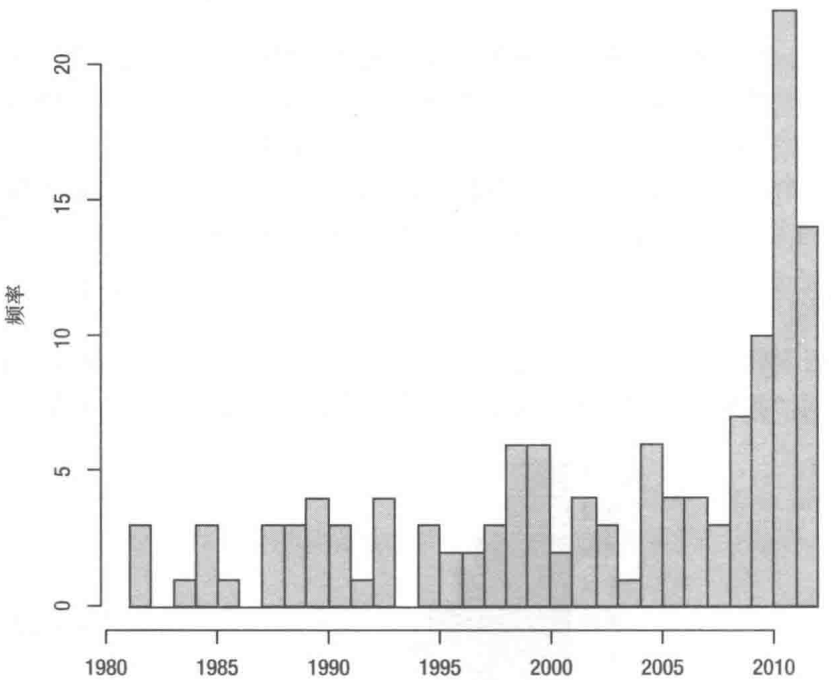

图 1-5　直方图展示的是纽约时报对数据可视化文章分布情况

数据映射

数据映射通常用于展示信息在空间区域中的分布。如图 1-6 所示，数据映射用来表示美国除阿拉斯加州之外其他各州对术语“数据可视化”搜索的兴趣度。

本例中，用深色标注的州表明这个州对搜索的这个术语有较高的关注度，(这个数据也来自 Google Trends，这个兴趣度用在 Google 中搜索“数据可视化”术语的频繁度来体现)。

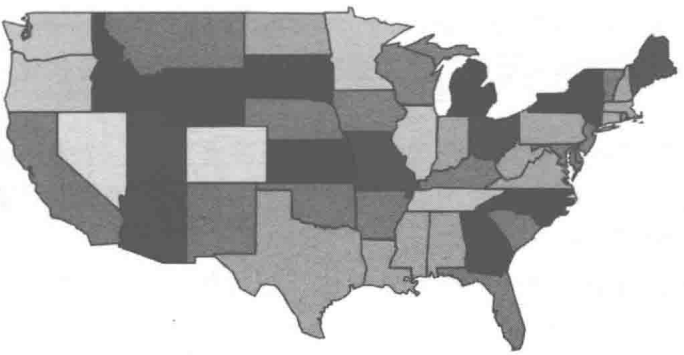

图 1-6 美国各州对“数据可视化”关注度的数据地图（数据来源 Google Trends）

散点图

和条形图一样，散点图经常用于对比数据，但有时是专门用来强调数据的相关性，或者表明在某种程度下这些数据在哪里可能是独立的，或者是相关的。图 1-7

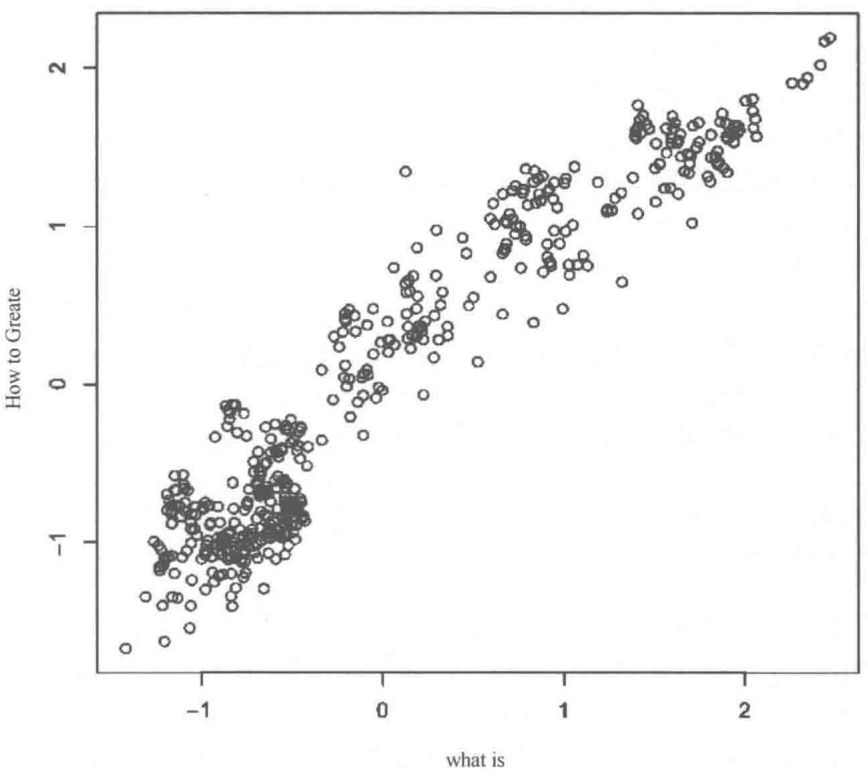

图 1-7 散点图展现了术语“Data Visualization”“How to Create”及其“What is”搜索量之间的关联性

使用来自 Google Correlate 的数据（http：//www. google. com/trends/correlate），观察关键字“什么是数据可视化”和关键字“如何制作数据可视化”的搜索值的关系。

这张图表明这些数据具有相关性，这也意味着其中一个术语兴趣度的增加会使另一个也增加。所以这张图表明越多的人了解数据可视化，则会有越多的人想要学习如何制作数据可视化。

关于相关，有很重要的一点需要记住：相关并不表示直接的原因—相关性不是因果关系。

历史

谈到数据可视化的历史，现代的数据可视化的概念最早是由 William Playfair 提出的。William Playfair 是一名工程师、会计师和银行家，经历了文艺复兴时期，他一手创造了时间序列图、条形图和气泡图。Playfair 的图表发表于 18 世纪末和 19 世纪初。他非常清楚他的发明是这类图形的首创，至少也是交流统计信息领域的第一个，所以他在书中花了大量的篇幅来描述如何在思想上取得进步，要明白就像金钱这样的物理事物中所蕴藏的条形图和折线图。

Playfair 因为他的两本书而被公众所熟知：一本是《商业和政治图册》，一本是《统计摘要》。《商业和政治图册》出版于 1786 年，这本书关注的是从国家债务到贸易金额，甚至包括军费支出等其他方面的经济数据。这本书的特色是第一次给出时间序列图和条形图。

《统计摘要》这本书关注的是关于当时欧洲主要国家的资源统计信息，并引入气泡图。

Playfair 使用这些图表还有一些政治目的，或许这些目的是存在争议的，如评论工人阶级消费能力的下降，甚至论证英国对进出口数据有利的平衡，但是，他最深远的目的是用易于接受的、广泛理解的图形传达复杂的统计信息。

声明：这两本书最近得以印刷，这要感谢 Howard Wainer 和 Ian Spence 以及剑桥大学出版社。

与 Playfair 同时代的 John Snow 博士制作了一张本书作者非常喜欢的图表：霍乱地图。这张霍乱地图包含了一个信息图形应具有的一切特点：简单易读，富含信息，更重要的是能解决实际问题。

这张霍乱地图是一个数据图，标出了 1854 年伦敦霍乱暴发时所有确诊案例的地点（见图 1-8）。图中的阴影区域记录了霍乱的致死数，阴影中的圆圈是水井。通过仔细地检查发现，死亡记录似乎是沿着 Broad 街道的水井辐射出去的。

Snow 博士将 Broad 街道的水井封住，霍乱才停止暴发。

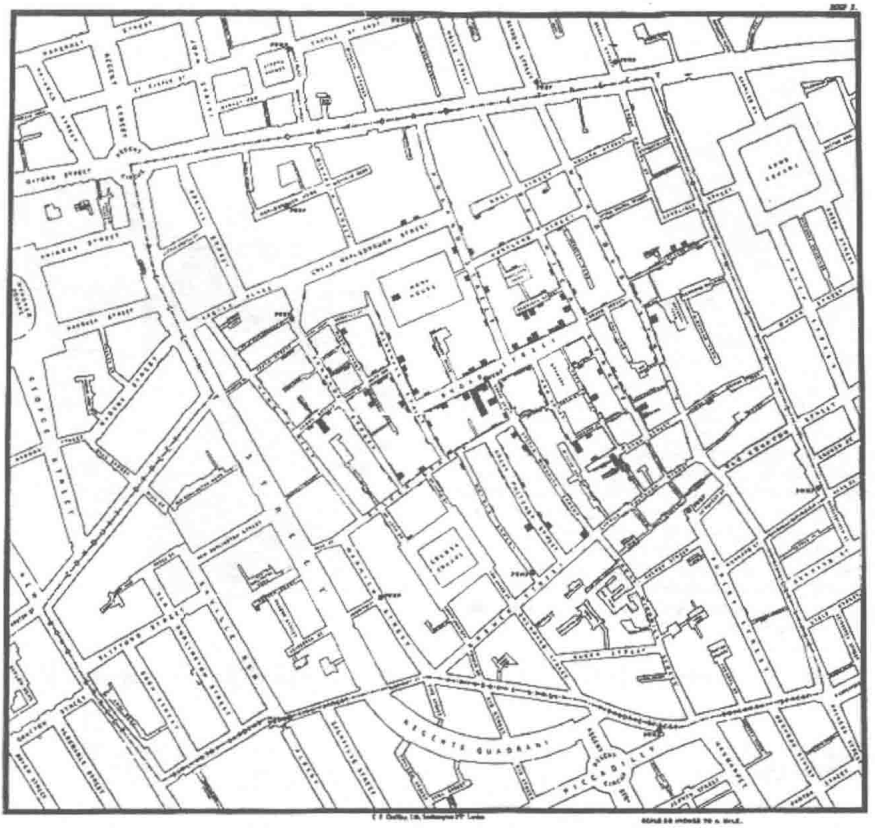

图 1-8 John Snow 制作的霍乱地图

多么完美、精确、有逻辑!

另一个历史上很有意义的信息图是中东军队死亡原因的图表，这是由 Florence Nightingale 和 William Farr 共同完成的。这张表如图 1-9 所示。

Nightingale 和 Farr 在 1856 年制作了这张图表，用来表明可控死亡的相对数量，进一步改善军队设施的卫生条件。注意，Nightingale 和 Farr 可视化的是一个固定形状的饼图。饼图通常用一个圆来代表所有给出的数据集合，而圆的每一块代表其所占整体的百分比。饼图的有用性有时会有争论，因为饼图相较于用长度决定的条形图或用直线定位的笛卡尔坐标而言很难认清角度值之间的不同。Nightlife 避免了这种缺陷，用楔子的角度来代表值，同时还变更了相关块的长度，这就避免了内含圆的限制，并且还表示了相关值。

以上所有的例子都是他们尽力去解决的特定目标或问题。

注意：更加丰富综合的历史超出了本书的范围，但是如果您喜欢思考，有敏锐的分析和探索，强烈推荐读 Edward Tufte 的《The Visual Display of Quantitative Information》。

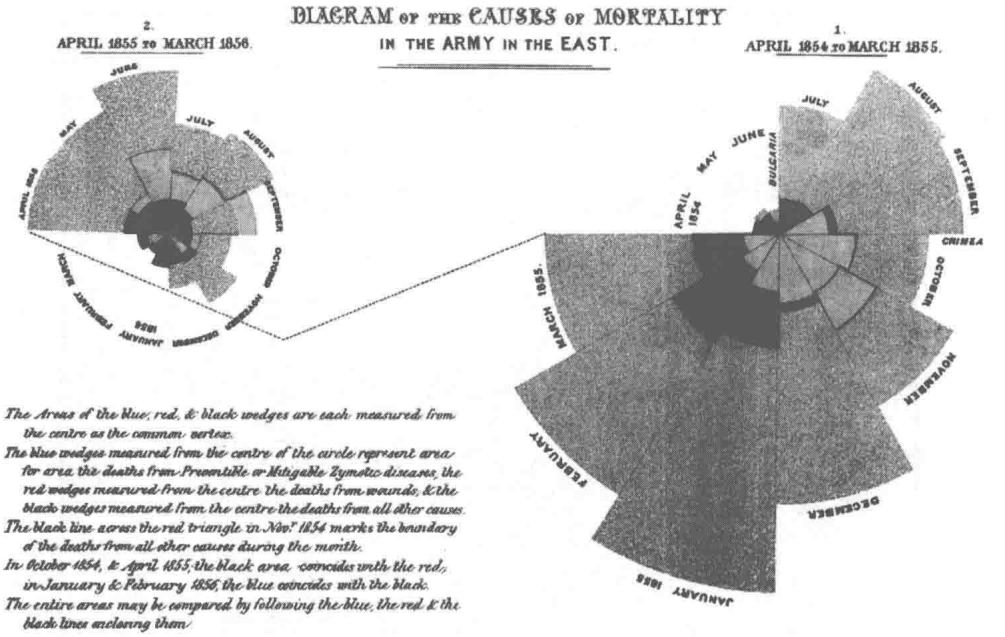

图 1-9　Florence Nightingale 和 William Farr 制作的中东军队死亡原因图表

模型风景画

数据可视化正处于现代复兴之中，是因为用于存储日志的廉价存储空间大量增加，以及有了免费和开源的工具用于分析和记录这些日志信息。

从消费和欣赏的角度来看，有一些网站专注于学习和探讨这些信息图形。这就出现了一些如 FlowingData 的网站，收集和讨论来自周围网站和天文事件大事记的数据，并用国会的议员席来模拟可视化。

Flow About 主页的使命声明（http：//flowingdata. com/about）是："FlowingData 探索如何用数据设计、统计和计算科学从而使我们能更好地理解——主要通过数据可视化"。

也有更加专业的网站如 quantifiedself. com，它主要是收集和可视化关于自身的信息。还有些关于数据可视化的漫画，最好的一个是由 Randall Munroe 制作的 xkcd. com。到目前为止，最著名和局部可视化的一个例子是 Randall 创造的辐射剂量图。如图 1-10 所示辐射剂量图（它在高分辨率下是有效的：http：//xkcd. com/radiation/）。

该图是针对 2011 年福岛第一核电站灾难而做的，它通过展示来自于他人或香蕉诸如此类辐射源的辐射量的规模差异，直至增加到致命剂量的辐射情形，旨在清除这场灾难造成的比较错误信息及错误解读——根据切尔贝诺利事故附

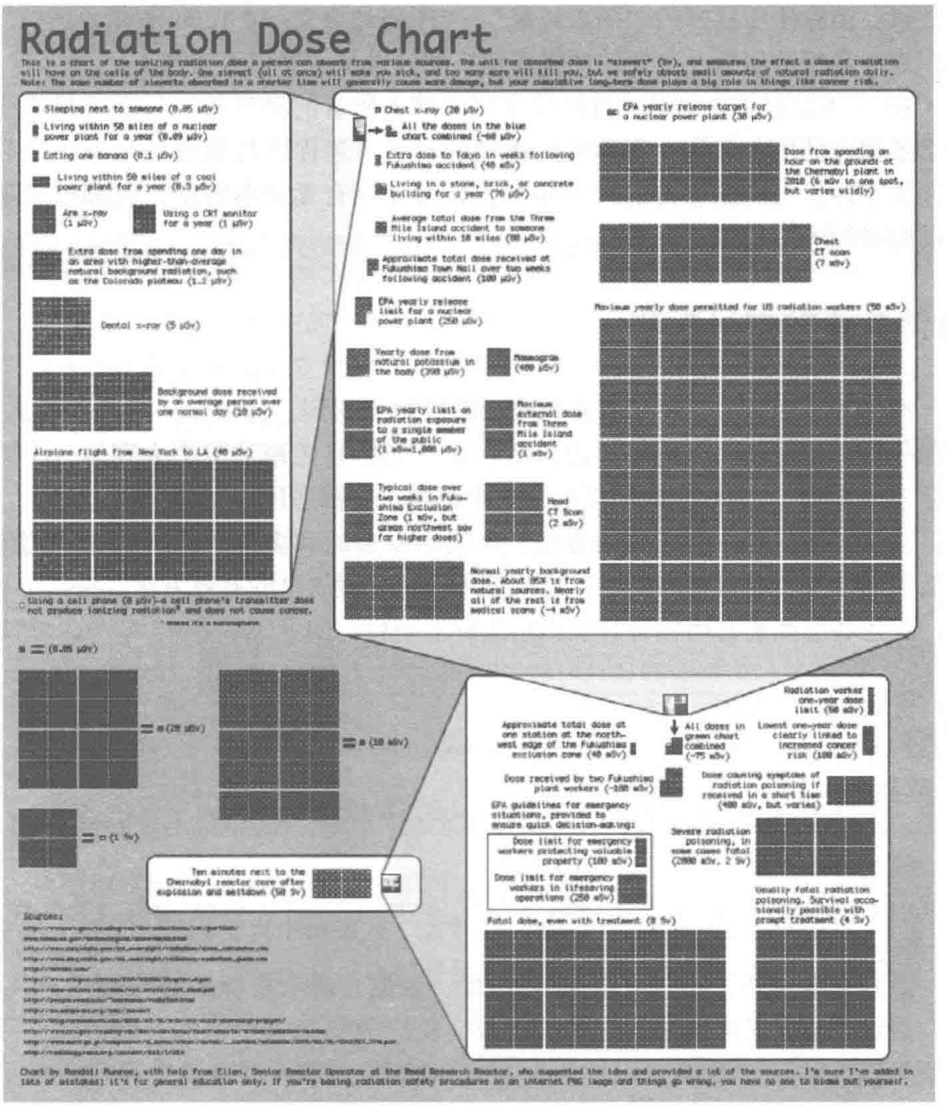

图 1-10 Randall Munroe 制作的辐射剂量图。图中将一个图标分成一些独立的块，将这些块用可视化来表示所代表的范围，这样可以展示一个全新的背景和信息的缩影。按区块比例展现了暴发范围，显示出一个全新的上下文信息关联缩影。此模式一次次重复，展示出令人难以置信的信息深度

近发送的长达 10min 的辐射量作为参考。

在这个世纪的最后一个四分之一结束时，耶鲁大学退休教授 Edward Tufte 一直在研究提高信息图像学中的条形图。他出版了一本开创性的书，列举数据可视化的历史，开始于分类、追踪其起源，甚至超过了 Playfair。在他的原则中有一个观点是将每一个图表中的信息数量最大化，包括增加图表中数据点或变量