



大数据技术与应用专业规划教材

数据可视化

原理及应用

◎ 樊银亭 夏敏捷 编著

《程序源码》

《电子课件》

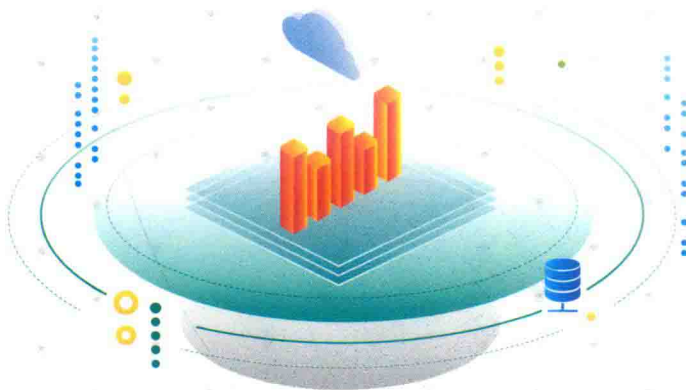
《教学大纲》

《50个知识点案例》

《2个完整项目案例》

《生成词云》

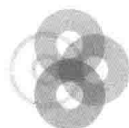
《舆情系统》



清华大学出版社



非
外
借



大数据技术与应用专业规划教材

数据可视化

原理及应用

◎ 樊银亭 夏敏捷 编著



清华大学出版社
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、统计、大数据处理及相关专业开设“数据可视化”课程而编写的。全书分为两篇:原理篇和应用篇。其中,原理篇从数据可视化发展历程、数据可视化数据的度量和可视化组件、可视化流程等方面讲解可视化基础理论和概念,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、文本数据、层次数据)介绍相应的可视化方法;应用篇着重介绍可视化工具 D3 的综合应用,同时介绍 Python 语言在可视化上的应用,最后一章用实例讲解一个 D3 在微信公众号舆情系统中的可视化应用。

本书可作为高等院校计算机、统计、大数据处理及相关专业高年级本科生和研究生的教学用书,也适用于 D3 语言学习者、可视化设计人员和数据分析人员,对于从事数据可视化、数据分析、视觉艺术开发和应用人员也有较大的参考价值。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

数据可视化原理及应用/樊银亭,夏敏捷编著. —北京:清华大学出版社,2019

(大数据技术与应用专业规划教材)

ISBN 978-7-302-53177-7

I. ①数… II. ①樊… ②夏… III. ①数据处理—高等学校—教材 IV. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 112002 号

责任编辑:陈景辉 张爱华

封面设计:刘 键

责任校对:李建庄

责任印制:沈 露

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:清华大学印刷厂

经 销:全国新华书店

开 本:185mm×260mm 印 张:21

字 数:508 千字

版 次:2019 年 10 月第 1 版

印 次:2019 年 10 月第 1 次印刷

印 数:1~1500

定 价:59.90 元

产品编号:079492-01

数据可视化(Data Visualization)起源于 18 世纪。William Playfair 在出版的书籍 *The Commercial and Political Atlas* 中第一次使用了柱状图和折线图,当时是为了表示国家的进出口量,今天柱状图和折线图依然在使用。19 世纪初,他出版了 *Statistical Breviary* 一书,里面第一次使用了饼状图。这三种图形都是至今常用的、著名的可视化图形。19 世纪中叶,数据可视化主要被用于军事,用来表示军队死亡原因、军队的分布图等。进入 20 世纪,数据可视化有了飞跃性的发展。1990 年,在人机界面学会上,它作为信息可视化原型的技术被发表。1995 年,IEEE Information Visualization 正式创立,信息可视化作为独立的学科被正式确立。随着 2012 年世界进入大数据时代,数据可视化作为大量数据的呈现方式,成为当前重要的课题。

数据可视化是指将大型数据集中的数据以图形图像形式表示,并利用数据分析和开发工具发现其中未知信息的处理过程。数据可视化的目的是对数据进行可视化处理,以使能够明确、有效地传递信息。比起枯燥乏味的数值,人类对于大小、位置、浓淡、颜色、形状等能够有更好、更快的认识,经过可视化之后的数据能够加深人类对于数据的理解和记忆。

全书分为两篇:原理篇和应用篇。其中,原理篇从数据可视化发展历程、数据可视化数据的度量和可视化组件、可视化流程等方面讲解可视化基础理论和概念,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、文本数据、层次数据)介绍相应的可视化方法;应用篇着重介绍可视化工具 D3 的综合应用,同时介绍 Python 语言在可视化方面的应用,最后一章用实例讲解 D3 在微信公众号舆情系统中的可视化应用。

本书由樊银亭和夏敏捷(中原工学院)主持编写,樊银亭编写第 1 章和第 2 章,尚展垒(郑州轻工业大学)编写第 5~7 章蔡增玉(郑州轻工业大学)编写第 9~11 章,其余章节由夏敏捷编写。

在本书的编写过程中,为确保内容的正确性,参阅了很多资料,并且得到了中原工学院研究生教材建设项目资助和资深 Web 程序员的支持,宋宝卫、潘惠勇、李娟和李国伟参与了资料整理。

本书的学习资源可以在清华大学出版社网站本书页面中下载。由于编者水平有限,书中难免有疏漏之处,敬请广大读者批评指正。

夏敏捷

2019 年 9 月

原 理 篇

第 1 章 数据可视化简介	3
1.1 数据可视化发展历程	3
1.2 数据可视化的目标和作用	9
1.2.1 数据可视化的目标	9
1.2.2 数据可视化的作用	9
1.3 数据可视化的优势	11
1.4 数据可视化与人机交互技术	11
1.4.1 人机交互的发展历史	12
1.4.2 人机交互的研究内容	13
1.4.3 人机交互的前景	14
第 2 章 数据可视化基础	15
2.1 数据对象与属性类型	15
2.1.1 数据对象	15
2.1.2 属性	15
2.1.3 属性类型	16
2.2 数据的基本统计描述	17
2.2.1 中心趋势度量	17
2.2.2 数据分布度量	19
2.3 数据的相似性和相异性度量	21
2.3.1 数据矩阵与相异性矩阵	21
2.3.2 标称属性的度量	22
2.3.3 二元属性的度量	22
2.3.4 数值属性的度量	22
2.3.5 序数属性的度量	23
2.3.6 文档的余弦相似性	23
2.4 视觉感知	25

2.4.1	视敏度和色彩感知	25
2.4.2	视觉模式识别	26
2.5	视觉通道	29
2.5.1	视觉通道简介	30
2.5.2	视觉通道的类型	34
2.5.3	视觉通道的表现力	34
2.6	可视化的组件	35
2.6.1	坐标系	35
2.6.2	标尺	36
2.6.3	背景信息	38
2.6.4	整合可视化组件	38
第3章	数据可视化过程	40
3.1	数据可视化流程	40
3.2	数据处理和数据变换	41
3.2.1	数据清理	42
3.2.2	数据集成	44
3.2.3	数据变换与数据离散化	45
3.2.4	数据配准	47
3.3	可视化映射	47
3.3.1	图形标记和视觉通道	48
3.3.2	可视化编码的选择	49
3.3.3	源于统计图表的可视化	51
第4章	数据可视化方法	55
4.1	二维标量场数据可视化方法	55
4.1.1	颜色映射法	55
4.1.2	等值线法	56
4.1.3	高度映射法	57
4.1.4	标记法	57
4.2	三维标量场数据可视化方法	58
4.2.1	直接体绘制	58
4.2.2	等值面绘制	60
4.3	向量场可视化方法	60
4.3.1	向量简化为标量	61
4.3.2	箭头表示方法	61
4.3.3	流线、迹线、脉线、时线	61
4.4	时间序列数据可视化方法	61
4.4.1	星状图	62

4.4.2	日历视图	63
4.4.3	邮票图表法	64
4.5	地理空间数据可视化方法	64
4.5.1	地图投影	64
4.5.2	墨卡托投影	64
4.5.3	摩尔威德投影	65
4.5.4	地理空间可视化方法	65
4.5.5	统计地图	66
4.6	文本与文档可视化方法	66
4.6.1	文本可视化的基本流程	67
4.6.2	文本可视化典型案例——词云	67
4.7	层次数据可视化方法	68

应 用 篇

第 5 章	可视化工具 D3 基础	75
5.1	D3 简介和使用	75
5.1.1	D3 简介	75
5.1.2	安装引用 D3	76
5.1.3	D3 需要的预备知识	76
5.1.4	HTML	76
5.2	JavaScript 编程基础	78
5.2.1	在 HTML 中使用 JavaScript 语言	78
5.2.2	JavaScript 的数据类型	79
5.2.3	JavaScript 运算符和表达式	81
5.2.4	JavaScript 控制语句和函数	85
5.2.5	JavaScript 函数	93
5.2.6	JavaScript 类的定义和实例化	99
5.2.7	调试 JavaScript 程序的方法	104
5.3	SVG 基础知识	106
5.3.1	图片存储方式	106
5.3.2	SVG 的概念	106
5.3.3	SVG 的优势	107
5.3.4	向网页添加 SVG XML	107
5.4	DOM	108
5.4.1	DOM 结点树	108
5.4.2	访问修改 HTML 元素	109
5.4.3	添加删除 HTML 元素结点	111
5.4.4	DOM 优点和缺点	111

5.5	Canvas	112
5.5.1	Canvas 元素的定义语法	112
5.5.2	使用 JavaScript 获取网页中的 Canvas 对象	112
5.5.3	绘制图形	113
5.5.4	描边和填充	117
5.5.5	绘制图像	118
5.5.6	图形的操作	121
5.6	CSS 语法基础	123
5.6.1	CSS 基本语句	123
5.6.2	在 HTML 文档中应用 CSS 样式	124
5.6.3	CSS 选择器	124
第 6 章	D3 开发入门	126
6.1	D3 入门实例	126
6.2	选择集与数据绑定	128
6.2.1	选择元素	128
6.2.2	设置和获取属性	131
6.2.3	插入和删除元素	134
6.2.4	绑定数据	136
6.3	enter-update-exit 模型	139
6.4	获取外部数据	143
6.4.1	JSON 数据	143
6.4.2	CSV 数据	146
第 7 章	绘制基本图形	150
7.1	颜色	150
7.1.1	RGB	150
7.1.2	HSL	151
7.1.3	插值	152
7.2	SVG	152
7.2.1	创建基本形状	152
7.2.2	过滤器和渐变	157
7.2.3	SVG 生成文本	159
7.3	图形生成器	161
7.3.1	直线生成器	161
7.3.2	区域生成器	164
7.3.3	弧生成器	166
7.3.4	弦生成器	169
7.3.5	对角线生成器	170

7.3.6 符号生成器	171
7.4 绘制柱状图	173
7.5 绘制折线图	174
第 8 章 比例尺和坐标轴	177
8.1 比例尺	177
8.1.1 什么是比例尺	177
8.1.2 线性比例尺	178
8.1.3 序数比例尺	179
8.1.4 量化比例尺	180
8.1.5 时间比例尺	181
8.1.6 颜色比例尺	182
8.1.7 其他比例尺	182
8.1.8 invert()与 invertExtent()方法	182
8.1.9 给柱状图添加比例尺	182
8.2 坐标轴	183
8.2.1 添加坐标轴	184
8.2.2 坐标轴相关方法	187
8.3 绘制有坐标轴的折线图	189
8.4 绘制有坐标轴的散点图	192
8.5 新版本 4.X 的坐标轴	194
第 9 章 实现动态过渡效果	196
9.1 动态效果	196
9.1.1 什么是动态效果	196
9.1.2 D3 实现动态效果的方法	196
9.2 动态效果实例	197
9.2.1 实现简单的动态效果	197
9.2.2 给柱状图加上动态效果	199
第 10 章 布局的应用	203
10.1 力导向图	203
10.1.1 D3 与其他可视化工具的区别	203
10.1.2 力导向图初步	204
10.1.3 基于力导向图的人物关系图	209
10.1.4 力导向图的事件	215
10.2 分区图	217
10.2.1 分区图的定义与设置参数	217
10.2.2 矩形分区图	219

10.2.3	圆形分区图	222
10.3	直方图	226
10.4	饼状图	229
10.5	弦图	236
10.6	集群图	240
10.7	树状图	244
10.8	堆栈图	246
10.9	矩阵树图	251
10.10	捆图	254
10.11	打包图	259
第 11 章	交互的应用	264
11.1	交互操作	264
11.1.1	鼠标交互的应用	265
11.1.2	键盘交互的应用	268
11.2	拖曳应用	271
11.2.1	单个元素拖曳的应用	271
11.2.2	多个元素拖曳的应用	273
11.3	缩放的应用	275
第 12 章	Python 科学计算和可视化应用	280
12.1	Python 基础知识	280
12.2	NumPy 库的使用	282
12.2.1	NumPy 数组	282
12.2.2	NumPy 数组的算术运算	284
12.3	Matplotlib 绘图可视化	285
12.3.1	Matplotlib, pyplot 模块——快速绘图	285
12.3.2	绘制条形图、饼状图、散点图等	292
12.3.3	绘制图像	298
12.3.4	图像轮廓和直方图	299
12.3.5	交互式标注	300
12.4	文本可视化实战——爬取豆瓣影评生成词云	301
12.4.1	安装 WordCloud 词云	302
12.4.2	使用 WordCloud 词云	302
12.4.3	爬取豆瓣影评生成词云的设计步骤	303
第 13 章	可视化在微信公众号舆情系统中的应用	311
13.1	系统背景意义	311
13.2	系统功能模块	311

13.3	功能需求	312
13.3.1	系统首页	312
13.3.2	文章热度	313
13.3.3	高频词云	313
13.3.4	文章正负面	314
13.3.5	话题发现	314
13.3.6	信息采集	315
13.4	系统实现	315
13.4.1	登录界面	315
13.4.2	欢迎界面	316
13.4.3	文章热度	317
13.4.4	高频词云	317
13.4.5	文章正负面	318
13.4.6	话题发现	318
13.4.7	信息采集	319
	参考文献	321

原理篇

第 1 章

数据可视化简介

数据可视化旨在借助于图形化手段,清晰、有效地传达与沟通信息。但是,这并不意味着数据可视化就一定因为要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,从而实现对于相当稀疏而又复杂的数据集的深入洞察。然而,设计人员往往并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的目的。

数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。当前,在研究、教学和开发领域,数据可视化乃是一个极为活跃而又关键的方面。“数据可视化”这条术语实现了成熟的科学可视化领域与较年轻的信息可视化领域的统一。

1.1 数据可视化发展历程

数据可视化是数据描述的图形表示,是当今数据分析领域当中发展最快速、最引人注目的领域之一。借助于可视化工具的发展,或朴实,或优雅,或绚烂的可视化作品给我们讲述着各种数据故事。在这个领域当中,科学、技术和艺术完美地结合在一起。

数据可视化一般被认为源于统计学诞生的时代,并随着技术手段、传播手段的进步而发扬光大;事实上,用图形描绘量化信息的思想植根于更早年代人们对于世界的观察、测量和管理的需要。本节将探索数据可视化的发展历程。

1. 数据可视化的起源

欧洲中世纪晚期是一个孕育着新纪元的时代。经济发展和文艺复兴点燃了欧洲人对人文和科学知识的追求,现代科学开始蹒跚起步。同时地理大发现如同大爆炸一般,把一个有待探索的新世界呈现在西欧人的面前,商人和探险家等满怀对财富、贸易或者知识的渴望

登上了驶向远方的航船。面对未知的新世界,很多新的科技,如绘图学、测量学、天文学等在迅速地更新着人们对世界的认识。

在16世纪,天体和地理的测量技术得到了很大的发展,特别是出现了像三角测量这样的可以精确绘制地理位置的技术。到了17世纪,笛卡儿发展了解析几何和坐标系;哲学家帕斯卡发展了早期概率论;英国人 John Graunt 开始了人口统计学的研究。数据的收集整理和绘制开启了系统的发展。这些早期的探索开启了数据可视化的大门。

2. 18 世纪——新的图形符号出现

18世纪是一个科学史上承上启下的时代。在这个世纪开始的时候,牛顿爵士已经在苹果树下发现了天体运动的伟大方程,微积分建立起来了,数学和物理知识开始为科学提供坚实的基础;在这个世纪里,化学也摆脱了炼金术,开始探索物质的组成;博物学家们继续在世界各地探索着未知的事物。社会生活也在发展,在这个世纪稍晚的年代,英国开始了工业革命,从此社会化大生产深刻地改变了整个世界:技术成为科学的另一条主线,社会管理也走向数量化和精确化。

与这些社会和科技进步相伴,统计学出现了早期萌芽。一些和绘图相关的技术也出现了,如三色彩印(1710)和平板印刷(1798)(后者被当今学者称为如同施乐打印机一般伟大的发明)。数据的价值开始为人们所重视,人口、商业等方面的经验数据开始被系统地收集整理,天文、测量、医学等学科的实践也有大量的数据被记录下来。人们开始有意识地探索数据表达的形式,抽象图形和图形的功能被大大扩展,许多崭新的数据可视化形式在这个世纪里诞生了。

这些新的图形创新涵盖很多图形领域。

在地图中,出现了等值线(Edmund Halley,1701)以及等高线表示的3D地图(Marcellin du Carla-Boniface,1782)。比较国家间差别的几何图形开始出现在地图上(Charles de Fourcroy,1782)。时间线被历史研究者引入,用来表示历史的变迁(Priestley,1765)。

法国人 Marcellin Du Carla-Boniface 绘制的等高线图(见图1-1),用一条曲线表示相同的高程,对于测绘、工程和军事有重大的意义,成为地图的标准形式之一。

特别重要的是,在后来被人们作为基本图形使用的饼状图、圆环图、条形图和线图也出现了。

3. 19 世纪前半叶

19世纪前半叶是最好的时代也是最坏的时代。科技在迅速发展,工业革命从英国扩散到欧洲大陆和北美。但是财富的增加并未同步地改善社会生活,各种革命在这个时代里层出不穷。但对数据可视化来说,这是一个快速发展的好时代。随着社会对数据的积累和应用的需求,以及技术和设计的进步,现代的数据可视化——统计图形和主题图的主要表达方式,在这几十年间基本都出现了。

在这个时期内数据可视化的重要发展包括:在统计图形方面,散点图、直方图、极坐标图和时间序列图等当代统计图形的常用形式都已出现。在主题图方面,主题地图和地图集成成为这个时期展示数据信息的一种常用方式,应用领域涵盖社会、经济、疾病、自然等各个主题。

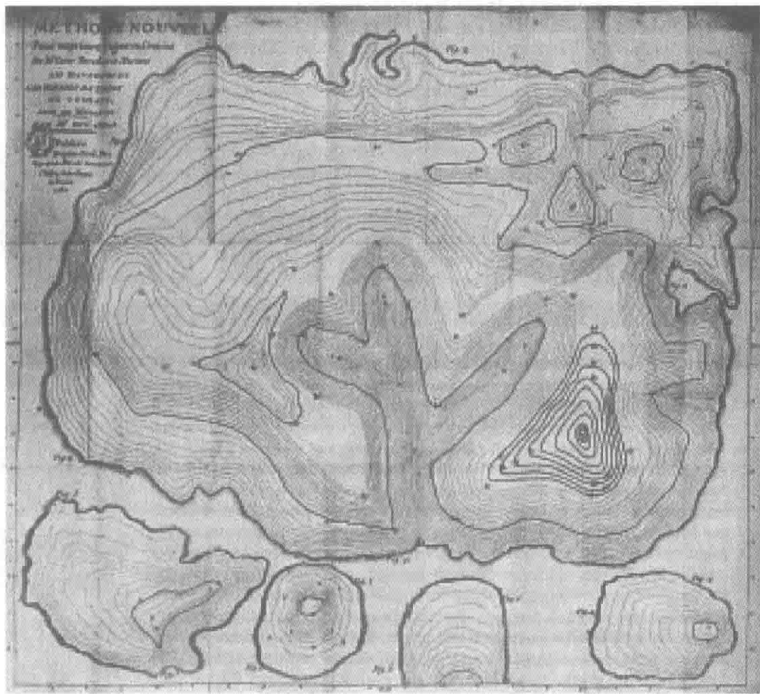


图 1-1 等高线图

1) 主题地图和社会学的发展

在 1801 年,英国地质学家 William Smith(1769—1839)绘制了第一幅地质图,这幅描绘了英格兰地层信息图在 1815 年出版后引起轰动,引领了一场在地图上表现量化信息的潮流。

1826 年,法国男爵 Charles Dupin 发明了使用连续的黑白底纹来显示法国识字分布情况的方法,这可能是第一幅现代形式的主题统计地图。

2) 霍乱地图与传染病的研究

19 世纪上半叶的欧洲,伴随工业迅速发展的是城市的扩张和人口的增长,但是公共管理并未能与时俱进。城市居民极易受到传染病的侵害。1831 年 10 月,英国第一次爆发霍乱,夺走了 5 万余条生命。在 1848—1849 年和 1853—1854 年的霍乱中,死亡人数更多。霍乱传播因何而来又如何传播? 可视化最终给出了答案。

1854 年,英国 Broad 大街大规模爆发霍乱,John Snow 对空气传播霍乱理论表示了怀疑,于 1855 年发表了关于霍乱传播理论的论文。John Snow 采用了点图的方式,图中心东西方向的街道即为 Broad 大街,黑点表示死亡的地点。这幅图揭示了一个重要现象,就是死亡发生地都在街道中部一处水源(公共水泵)周围,市内其他水源周围极少发现死者。通过进一步调查,他发现这些死者都饮用过这里的水。后来证实离这口水泵仅三英尺(1 英尺 $\approx$ 0.3048 米)远的地方有一处污水坑,坑内滋生的细菌正是霍乱发生的罪魁祸首。他成功说服了当地政府废弃那个水泵。这是可视化历史上的一个划时代的事件。

3) 提灯女士的玫瑰图

玫瑰图即极坐标面积图(polar area diagram),将极坐标平面分为若干个角相等但面积不等的区域,适合表示周期循环的数据。这种图形可以被视为饼状图的一个变种,又因为每

个扇区区域面积不同,又称玫瑰图(也称为风玫瑰图)。

在克里米亚战争期间,南丁格尔通过搜集数据发现,很多人死亡的原因并非是“战死沙场”,而是因为在战场外感染了疾病,或是在战场上受伤,却没有得到适当的护理。

为了解释这个原因,并降低英国士兵的死亡率,她绘制了这幅著名的图,并于1858年送到了维多利亚女王手中。这幅图中一个切角是一个月,其中面积最大的灰色块代表着可预防的疾病。这幅图真的很厉害,为什么呢?第一,它用面积直观地表现出了一个时间段内几种死因的占比,让任何人都能看懂;第二,它还很漂亮,像一朵玫瑰花一样。它为什么要那么漂亮?因为这幅图的汇报对象以及最终的决策人是维多利亚女王!南丁格尔的故事告诉我们:数据可视化是为了更好地促进行动,所以要让行动的决策人看懂。

4. 19 世纪下半叶的黄金时期

19 世纪下半叶,系统地构建可视化方法的条件日渐成熟,进入了统计图形学发展的黄金时期。值得一提的是法国人 Charles Joseph Minard,他是将可视化应用于工程和统计的先驱者,其最著名的工作是 1869 年发布的描绘 1812—1813 年拿破仑进军莫斯科大败而归的历史事件的流图。

这幅拿破仑 1812 年的远征图被后世学者称为“有史以来最好的统计图表”。这场战争以法国军队的惨败而告终,侵入俄国的 42 万人最终生还者仅仅数万。造成法军损失惨重的原因,除了俄罗斯人的顽强抵抗,还有恶劣的自然条件,特别是 1812 年冬季的严寒。

这幅远征图反映了这场战争全景,其经典之处在于在一幅简单的二维图上,表现了丰富的信息:法军部队的规模;地理坐标;法军前进和撤退的方向;法军抵达某处的时间以及撤退路上的温度。这张图对 1812 年的战争提供了全面、强烈的视觉表现,如撤退路上在别列津河的重大损失、严寒对法军损失的影响等,这种视觉的表现力是历史学家的文字难以比拟的。

5. 20 世纪上半叶

20 世纪上半叶,数据可视化最重要的影响是在天文、物理、生物和其他科学领域中。图形方法被广泛应用在新发现、新思想和新理论的过程中。其中主要包括:① E. W. Maunder (1904) 的蝴蝶图,研究了太阳黑子随时间的变化。他发现 1645—1715 年,太阳黑子的频率有明显减少。图 1-2 是由 NASA 按照 Maunder 方法绘制的蝴蝶图。② Hertzsprung-Russell 图(1911),作为温度函数的恒星亮度的对数图,解释了恒星的演化,成为现代天体物理的奠基之一。③ Henry Moseley 关于原子序数的发现(1913),这也是基于大量的图形分析。

在这个时期稍晚的阶段,统计和心理学上的一些多维数据可视化的思想和方法提供了超越二维图形表现的动力。

在主题图方面,这个时期的一个有意思的创新是关于伦敦地铁图(见图 1-3)的设计,并由此产生了 Tube Map 这样一种交通简图的表现手法。早期的地铁图与普通地图无异,对乘客来说,地理信息充分但远非简明直观。1931 年,身为电气工程师的 Beck 重新设计了伦敦地铁图,使之具有三个比较明显的特点:以颜色区分路线;路线大多以水平、垂直、45°角三种形式来表现;路线上的车站距离与实际距离不成比例关系。其简明易用的特点使其在 1933 年出版后迅速为乘客接受,并成为今日交通线路图形的一种主流表现方法。

6. 20 世纪下半叶至今——数据可视化的创新思维时代

引领这次大潮的首先是一个划时代的事件——现代电子计算机的诞生。计算机的出现