

新媒体传播 理论与应用

精·品·教·材·译·丛

数据可视化实用教程

VISUAL INSIGHTS

A Practical Guide to Making Sense of Data

[美]凯蒂·伯尔纳 戴维·E. 波利 | 著

嵇美云 余虹姗 支庭荣 | 译

C⁺

R



清华大学出版社

新媒体传播 理论与应用

精·品·教·材·译·丛

数据可视化实用教程

VISUAL INSIGHTS

A Practical Guide to Making Sense of Data

[美]凯蒂·伯尔纳 戴维·E.波利 | 著

嵇美云 余虹姗 支庭荣 | 译



C+

清华大学出版社

北 京

VISUAL INSIGHTS: A Practical Guide to Making Sense of Data by Katy Börner & David E. Polley © 2014 Massachusetts Institute of Technology. All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

北京市版权局著作权合同登记号 图字: 01-2016-3541

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

数据可视化实用教程 / (美)凯蒂·伯尔纳, (美)戴维·E.波利 著; 嵇美云, 余虹姗, 支庭荣 译. —北京: 清华大学出版社, 2017

(新媒体传播理论与应用精品教材译丛)

书名原文: VISUAL INSIGHTS: A Practical Guide to Making Sense of Data

ISBN 978-7-302-47912-3

①数… II. ①凯… ②戴… ③嵇… ④余… ⑤支… III. ①可视化软件—教材 IV. ①TP31

中国版本图书馆 CIP 数据核字(2017)第 193249 号

责任编辑: 陈 莉 高 岫

封面设计: 周晓亮

版式设计: 方加青

责任校对: 牛艳敏

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

印 装 者: 三河市铭诚印务有限公司

经 销: 全国新华书店

开 本: 185mm×260mm 印 张: 14.75 插 页: 16 字 数: 282 千字

版 次: 2017 年 10 月第 1 版 印 次: 2017 年 10 月第 1 次印刷

定 价: 49.80 元

产品编号: 064291-01

编 委 会

主任：林如鹏 暨南大学

主编：支庭荣 暨南大学

编委(按姓氏音序排列)：

李 彪 中国人民大学

李良荣 复旦大学

刘 涛 暨南大学

陆 地 北京大学

沈 阳 清华大学

谢耘耕 上海交通大学

张志安 中山大学

钟 瑛 华中科技大学

祝建华 香港城市大学

这是一个新兴媒体高歌猛进的时代。中国接入国际互联网二十多年，见证了网络社会的异军突起。“互联网+”计划和国家大数据战略的实施，进一步提升了新媒体的增长空间。截至2015年6月，全国的互联网普及率趋近50%，智能手机普及率超过七成。作为对比，北京地区电视机开机率保持在六成以上，从理论上说，如果电视机全部消失，对城市的影响已不太大，尽管还是会影响到相当一部分乡村地区的收视需求；同样，如果报纸全部消失，对大部分读报人口来说影响也不太大，尽管其阅读体验可能会下降不少。互联网和手机对于传统报纸和电视的替代性，越来越强。只要有手机在，没有报纸的日子并非难以忍受；只要有电脑、平板电脑和互联网，没有大屏幕彩电的日子也没那么难熬。人们对移动和社交的迷恋，甚至已逐渐成为一种“文化症候”。新媒体，正在成为人体的新延伸。

曾几何时，世界上最大的免费物品是空气和阳光，如今可能就要数互联网上的信息。网络信息的市场均衡价格，近乎为零。当然，免费也是世界上最昂贵的东西。免费带动付费，以至于数字经济蓬勃、野蛮生长。专业机构和众包生产参差不齐的内容，一起被投进了免费的染缸，难分彼此。在报纸的黄金时代，读者挑错的来电来函络绎不绝。在互联网时代，用户对低劣信息的容忍度却增加了，见猎心喜，愿意忍受免费、新奇而营养价值或许欠奉的内容。总之，文明虽终将驯化野蛮，野蛮却正在征服文明。互联网以及整个新媒体家族，作为巨大的分布式的数据生产、复制工厂和推送、分享空间，具有一种吞噬性的力量。几乎人类有史以来创造的所有内容，都可以用极低的成本迅速数字化。这样一种近乎“黑洞”般的传播能力，使得任何单体的模拟制式的传播者黯然失色。新媒体以不可阻挡之势，席卷了内容、娱乐和各种各样的应用市场。

从产业结构层面来看，互联网以及新媒体世界的控制力，掌握在技术取向的大型平台和超级运营商的手中，这些大型平台和超级运营商，如谷歌、苹果、百度、腾讯、阿里等，逐渐囊括了信息聚合、信息储存、信息搜索、社交娱乐、地理位置服务、数据挖掘、智能制造、电子商务等环信息经济圈。新闻，只是它们的副业之一。

技术相对于内容的霸权，在目前这一信息技术革命不断升级的阶段是相当明显的。但是，人类社会终究由人们的认知、心态、想法、观念所主导，而非技术

的奴隶。移动终端不过是增加了一些优越感和幸福感而已。好的内容，优质的新闻产品，始终有它的独特价值，并且能够在技术标准逐渐成熟后，再一次恢复自己的崇高声望。因此，技术不可或缺，内容也依然重要，它们彼此纠缠。计算机科学技术不等于新媒体的全部，新媒体传播的理论和应用，仍有许多独特的规律等待人们去探求。

大致说来，用户对新闻信息需求的核心本质，是对周围环境和未来不确定性恐惧的消除，相关联地，也包括交流和娱乐。如果人性不变，那么需求会长期存在。至于满足需求的方式、介质，新传播技术正在并还将创造出很多种可能。看起来，新媒体传播与传统新闻工作有着一定的相似之处，它们都取决于一个个睿智头脑的即时生产，标准化作业即使有，也是有一定限度的。语言的隔阂、用户的地缘兴趣随着距离的增加而衰减，决定了行业的规模边界。但是，机器人对人工操作的取代，在财经、天气等领域已初显身手。智能化技术将会解决很大一部分初级信息的生产和传播问题。技术的含量，与内容、产品、营销等类目相比，如果不是更重要，至少需要得到同等程度的重视。

与此同时，新媒体传播的理论和应用，也对深化和拓展传统新闻传播学的地盘提出了新要求。从历史的角度看，是互联网的出现承接、替代了媒体的功能，而不是媒体创造了网络。媒体是网络时代的追随者，是数字革命的后知后觉者，媒体恐怕做不到掌控网络的命运。互联网为各种各样的企业提供底层平台，也推动了商业、教育、娱乐和新闻信息等应用平台的成长。具有强大商业能力、创新能力的企业，乃是网络时代的弄潮儿。当媒体汇入了互联网的洪流中，意味着新闻行业就像文艺复兴之后的教会一样，必须适应这一商业化和世俗化进程，意味着新闻业的变革成为必然。实践呼唤着理论的回应，新媒体传播学科的进一步发展成为必需。

当然，人们不应忘记，渠道越发过剩，数据越发富集，信息越发泛滥，而优秀的产品始终稀缺。这是新媒体传播的价值和命脉所在。

鉴于时代的新变化和人才培养的新需求，我们与清华大学出版社又一次携手合作，瞄准世界前沿，组织了一套“新媒体传播理论与应用精品教材译丛”，以飨国内的读者。前路漫漫而修远，求索正未有穷期。

支庭荣

这本书是2014年1月修读《信息可视化》慕课(<http://ivmooc.cns.iu.edu>)的学生配套资源。为此,我们优先及时地推出了这本教材,虽然此教材已经由几位专家审读和两位编辑多次校对过,操作流程也已经由诸多新手和高级用户在不同的操作平台多次测试过,我们仍然感激那些将错误和故障报告发送到网址(cns-sci2-help-l@iulist.indiana.edu)的用户,这些问题将在本教材修订时得到纠正。

我们希望在本书中探讨的概念还有许多,但是因篇幅所限未能如愿。视觉感知、认知加工以及如何执行人类主体实验(human subject experiments)只是希望探讨的诸多概念中的几个。

最后,我们承认本书中尺寸比较大的以及具有互动性质的一些视觉画面,在印刷形式下并不能很好地对其进行探究。因此,我们在网站中增加了链接到高分辨率图形的页面(<http://cns.iu.edu/ivmoocbook14>),按照章节和图形编号可以很方便地查询到这些图形。本书绿色放大镜图标表示在网上可查到这些图形,在图形标题中可以发现相关链接。

另外,为了更好地呈现某些图片,我们将其统一放于本书最后,作为彩插,展现给读者,这些图片是:图1.6、图1.7、图1.8、图1.9、图1.10、图1.11、图1.15、图1.16、图2.4、图2.11、图2.17、图3.2、图3.3、图3.9、图3.12、图3.14、图3.15、图3.17、图3.20、图3.22、图4.2、图4.3、图4.4、图4.12、图4.17、图5.4、图5.8、图5.9、图5.10、图6.2、图6.3、图6.5、图6.12、图6.21、图6.22、图6.41、图7.6、图7.20、图8.1、图8.4、图8.5、图9.2。

2012年9月，我同时接到美国印第安纳大学(IU)信息学和计算机学院院长罗伯特·B. 施纳贝尔(Robert B. Schnabel)以及图书馆与信息科学学院院长黛博拉·拉尔夫·肖(Debora “Ralf” Shaw)的电话，他们都表达出希望我承担2013年春季大型网络开放课程(massive online open course, MOOC, 通常称为慕课)的授课任务。因为这种“开放式教育”理念与网络科学中心(Cyberinfrastructure for Network Science Center, CNS)的网络基础设施建设，与我亲自领导的正在开发和推进的“开放数据”和“开放代码”非常契合，这个独特的机会让我立刻产生了浓厚的探索兴趣。过去十年里，我在多个国家的讲习班中讲授过开放数据和代码，而且十多万用户已经下载过我们的即插即用的宏观工具(macroscope tools)。^①戴维·E. 波利(David E. Polley)最近加入了我们的团队，负责测试和录制软件，同时在印第安纳大学(IU)讲习班和国际研讨会上讲解过这些工具。事实证明，我的博士生斯科特·B. 魏因加特(Scott B. Weingart)不仅是一位出色的研究者和魔术师，还是一位很能鼓舞人心的老师。到一月份开始上课似乎是可行的，特别是得到印第安纳大学广泛的支持，于是我应承在2013年春季学期讲授《信息可视化》慕课(Information Visualization MOOC, IVMOOC)。

我们很快了解到印第安纳大学已决定所有慕课的开发和教学，都使用谷歌课程制作(Google Course Builder, GCB)平台的开放源码。那时候，我们已经使用过GCB一次，向10多万学生讲授《谷歌强力搜索》^②课程。GCB平台不支持发送电子邮件或成绩评定；课程设置和评定采用的是低水平的编码和脚本处理。我希望《信息可视化》慕课的学生可以与我互动，与网络科学中心的其他人互动，与外部客户们互动，我们雇请了迈克·威德纳(Mike Widner)和斯科特·B. 魏因加特为GCB开设内容管理(Drupal)论坛。为了满足评定成绩的需要，罗伯特·P. 莱特(Robert P. Light)为IVMOOC设计了数据库，此数据库不仅可以给学生评分，还可以捕捉到谁与谁合作了，谁看了什么视频，看了多长时间诸如此类的信息。最终，慕课用户需要新的技术和工具才能达到最佳效果——老师需要理解数千名学

① Börner, Katy. 2011. “Plug-and-Play Macroscopes.” *Communications of the ACM* 54, 3: 60–69.

② <http://www.google.com/insidesearch/landing/powersearching.html>

生行为的意义, 学生需要浏览学习材料, 并在跨学科和跨时区之间成功地实现合作学习, 例如进行客户端项目研究(参见第9章慕课可视化分析)。

在为IVMOOC开发和录制材料期间, 我正在研究《知识地图集》(*The Atlas of Knowledge*), 其副标题是“人人皆可绘制”, 此灵感来自厨神奥古斯特·古斯多(Auguste Gusteau)的口号“人人都能当厨师”。该地图集的目标是突出永恒的知识(爱德华·塔夫特称之为“永远的知识”), 原则是无所谓文化、性别、国籍或历史。相比之下, 《信息可视化》慕课突出的是“适时的知识”, 或者说运用当前的数据格式、工具、工作流程将数据转换成深刻见解。

具体而言, IVMOOC的材料分为7个单元, 讲授时间需要7周多(参见本书第1~7章)。每周教学单元的理论部分由我负责, 实践部分由戴维·E. 波利负责。第一个理论单元介绍了理论性的可视化框架, 旨在帮助那些非专家级人士整合高级的工作流分析并设计不同的可视化效果。为了更好地解释或者实现效果的最优化, 该框架还可以应用于“剖析可视化”。接下来的5个单元关于工作流程和可视化的介绍, 运用时间、空间、主题以及网络分析方面的多种技能, 回答何时、何地和谁等问题。最后一个单元探讨了动态变化的数据可视化和不同媒介部署方式可视化的最优方案。实践部分深入地介绍了如何浏览和操作适用于信息可视化的软件程序。此外, 学生们在学习种种技能后, 需要将他们自己的数据可视化, 他们还需要创建独特的可视化效果。在相应的地方有在线辅导课^①的指示。理论部分和实践部分都是独立的。参与者们可以观察哪个部分是其首先感兴趣的, 然后再回顾其他部分。观看理论课程视频之后需要做出自我测评, 观看实践课程视频后, 需要完成一份简短的家庭作业。

在修读此课程之前、期间或者之后, 我们都鼓励学生创建并使用推特(Twitter)和闪烁(Flicker)账户, 并贴上“IVMOOC”标签进行图片分享, 同样也可以链接到更富有洞察力的可视化实践、研讨会、活动或者相关的职位, 这些职位面向那些应用数据测量和可视化技术, 回答各种实际问题, 有能力创建独特的、实时数据流的最优的可视化技术人员、专家或公司开放。

此研究生水平的信息可视化课程, 免费向世界各地的参与者开放, 任何一个注册者都能自由访问学术数据库^②, 该数据库拥有26万份论文、专利和授权记

① <http://sci2.wiki.cns.iu.edu>

② <http://sab.cns.iu.edu>

录，以及有100+算法和Sci2 工具^①。学习者还有机会与现实世界中可视化项目的真实客户合作。

IVMOOC最终成绩基于期中考试(30%)、期末考试(40%)和项目/家庭作业(30%)而评定。所有参与者只要掌握超过80%的知识要点，都将获得结业证书。

欢迎登录网站<http://ivmooc.cns.iu.edu>在线注册和学习课程内容。

凯蒂·伯尔纳

印第安纳大学信息学和计算机学院网络科学中心网络基础设施部

^① <http://sci2.cns.iu.edu>

第1章 可视化架构和 workflow 设计..... 1

- (一) 理论部分 1
 - 1.1 可视化架构 2
 - 1.2 workflow 设计 13
- (二) 实践部分 20
 - 1.3 示例 20
 - 1.4 下载并安装SCI2 21

第2章 “何时”：时间数据... 25

- (一) 理论部分 25
 - 2.1 可视化示例 25
 - 2.2 概述和术语 31
 - 2.3 workflow 设计 37
 - 2.4 激增检测 43
- (二) 实践部分 47
 - 2.5 时间条形图：美国国家科学基金会资助概况 47
 - 2.6 出版物标题激增检测 49

第3章 “何地”：地理空间数据..... 57

- (一) 理论部分 57
 - 3.1 可视化示例 57
 - 3.2 概述和术语 61
 - 3.3 workflow 设计 67
 - 3.4 色彩 69

- (二) 实践部分 72
 - 3.5 用比例符号图和地区分布图可视化USPTO的数据 72
 - 3.6 使用选区进行地理编码 75
 - 3.7 运用通用地理编码器编码NSF数据 78

第4章 “什么”：主题数据... 84

- (一) 理论部分 84
 - 4.1 可视化示例 84
 - 4.2 概述和术语 88
 - 4.3 workflow 设计 91
 - 4.4 分类系统的设计和更新：UCSD科学地图 96
- (二) 实践部分 100
 - 4.5 图绘PNAS中的主题激增 100
 - 4.6 UCSD科学地图 102

第5章 “与谁”：树形数据... 104

- (一) 理论部分 104
 - 5.1 可视化示例 104
 - 5.2 概述和术语 109
 - 5.3 workflow 设计 110
- (二) 实践部分 115
 - 5.4 用SCI2可视化目录结构(分层数据) 115

第6章 “与谁”：网络数据… 123 第8章 案例研究 …… 174

(一) 理论部分 …… 123

6.1 可视化示例 …… 123

6.2 概述和术语 …… 128

6.3 工作流程设计 …… 131

6.4 聚类 and 主干识别 …… 139

(二) 实践部分 …… 144

6.5 佛罗伦萨家族网络的
可视化 …… 144

6.6 作者共现(合著者)网络 …… 148

6.7 定向网络：论文-引用网络 …… 154

6.8 对分网络：杰弗里·福克斯
获得NSF资助情况 …… 155

第7章 动态可视化和部署 …… 158

(一) 理论部分 …… 158

7.1 动态可视化 …… 158

7.2 交互式可视化 …… 160

(二) 实践部分 …… 166

7.3 从Gephi中用Seadragon导出
动态可视化 …… 166

案例一 了解非紧急呼叫系统的

扩散 …… 174

案例二 探索《魔兽世界》玩家活动 的成功 …… 178

案例三 用观察相机研究学生-教师 互动 …… 183

案例四 Phylet：交互式“生命 之树”可视化 …… 189

案例五 Isis：绘制科学期刊史地理 空间和主题分布 …… 194

案例六 蜂巢纽约学习网络影响的 可视化 …… 198

第9章 讨论与展望 …… 204

9.1 IVMOOC 评价 …… 204

9.2 IVMOOC 数据分析 …… 206

9.3 IVMOOC 拓展计划 …… 211

附 录 …… 213

可视化架构和工作流程设计

(一) 理论部分

欢迎来到信息时代，在这个时代，我们每个人每天通过微博、电子邮件、新闻以及其他数据流接收到的信息要多于我们在24小时内能处理的信息；任何一次互联网连接都可以访问大部分的人类知识。我们的办公室充满各种信息，我们的电子邮箱也不堪信息泛滥(见图1.1左)。我们迫切地需要更高效的方式了解这些大量的数据，浏览并掌控信息，识别出合作者与朋友，确认模式和趋势(见图1.1右)。

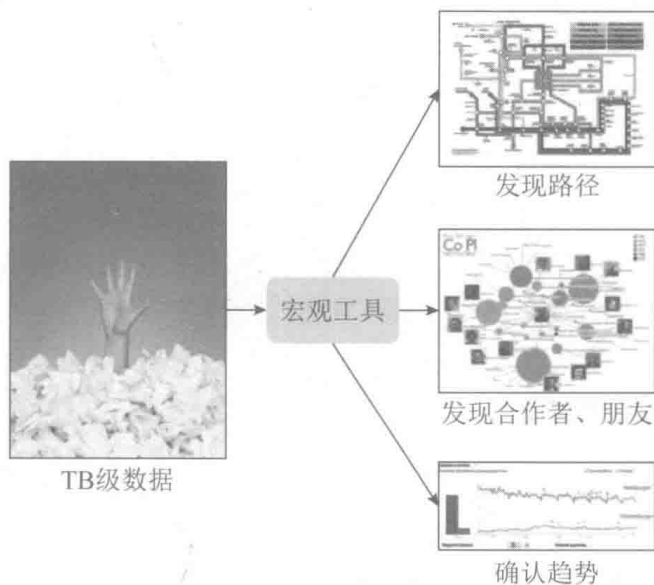


图1.1 将数据转换为随时备用的深刻见解

本书将教你运用先进的数据挖掘技术和可视化技术将数据转化为深刻的见解。每一章都包括理论与实践两个部分，理论部分配有自我测评内容，实践操作部分则包括了课后作业。

第1章介绍了理论性的可视化架构，此架构有助于选择最恰当的算法，以及将这些算法组合到有效的工作流之中。实践操作部分介绍了所谓的宏观工具^①，此工具让任何人都能够阅读、处理、分析数据，并将数据可视化。

1.1 可视化架构

本节介绍的这个架构有助于形成结构并将可视化结果分组，还可以支持识别何种类型、何种层级的分析，最适合于处理特定用户的需求。运用分组来形成架构，在信息可视化领域之外具有悠久的历史。事实上，科学往往始于对事物进行分组或者归类。例如，动物学家对动物的分类，老虎、狮子、美洲虎最终都同归为猫科(大型猫科动物)。德米特里·门捷列夫(Dmitri Mendeleev)根据化学性质和原子量将化学元素做了分类，由此创造了化学元素周期表，周期表中留下的那些“洞”，表示尚未发现的元素。同样，系统分析和信息可视化分类有助于设计新的可视化界面，也能诠释在期刊、报纸、书籍以及其他出版物中出现的视觉效果。

有多种方式对可视化进行分组，如图1.2所示。可视化可以根据用户的洞察需求、用户的任务类型，亦或是根据数据的可视化形式进行分组。可视化也可以基于数据挖掘技术，基于能够维持什么样的交互性，或者基于发布的类型(如，可视化是通过报纸、动画发布的，还是在其他交互式显示设备上发布的)。现有分类层级和架构的细节和参考可以在《知识图集》(*Atlas of Knowledge*^②)中找到。在此，我们介绍一个实用的方法，可以教学习者设计出一种有意义的可视化。从用户问题的类型开始，建构支持数据挖掘和可视化工作流程的选项，也支持用于回答这些用户问题的发布选择项。

分析层级和分析类型

可视化架构可以做出三种级别的分析：微观、中观和宏观。可视化架构在微观级别或者个体级别上，由小的数据集组成，通常介于1到100条记录之间，例如一个人以及他/她所有的朋友。其次是中观级别，或者说是群组层级，数据集介于101到10 000条记录之间，如在一所大学或某个研究领域的研究人员的信息。

^① Börner, Katy. 2011. "Plug-and-Play Macroscopes." *Communications of the ACM* 54, 3: 60–69.

^② Börner, Katy. 2014. *Atlas of Knowledge: Anyone Can Map*. Cambridge, MA: The MIT Press.

最后，最广泛的分析层级是宏观级别，有时也称之为全局的或总体层级。这个级别的项目分析数据集往往超过10 000条记录，例如关于整个国家某一方面的数据或所有学科的数据。

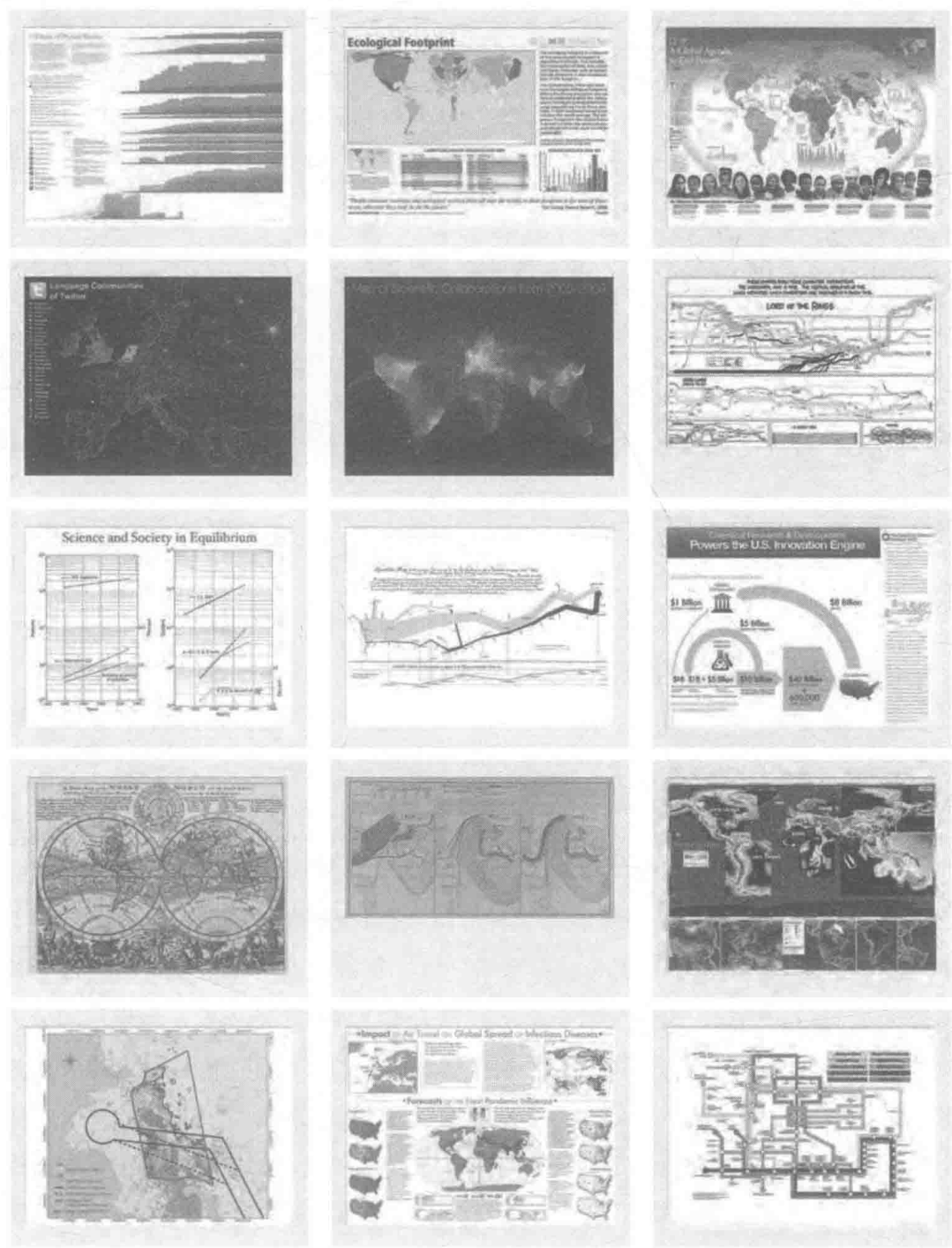


图1.2 《场所与空间：科学图集》中展示的地图集(1)

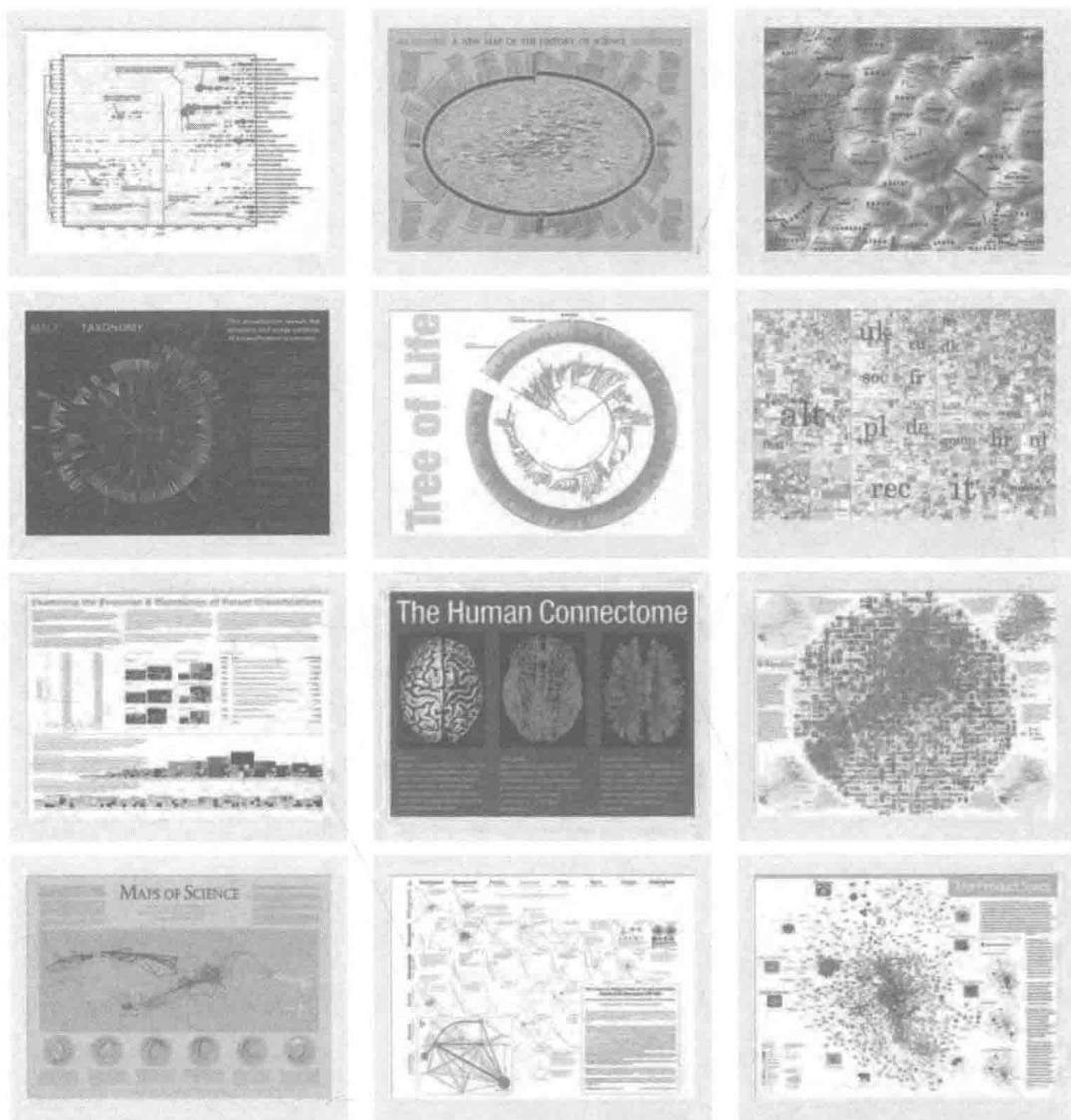
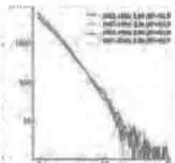
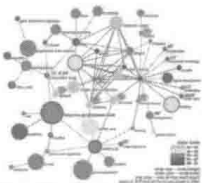
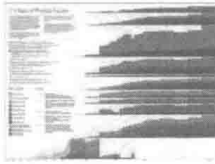
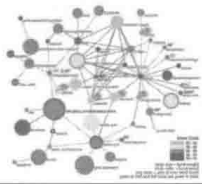
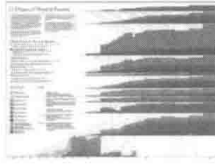
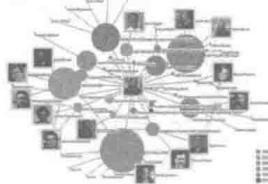


图1.2 《场所与空间：科学图集》中展示的地图集(2)

在每一层级的分析中，可能有五种分析类型：统计分析/剖析、时间分析、地理空间分析、主题(topical)分析和网络分析。每种类型的分析都试图回答一种具体类型问题。例如，时间分析有助于回答关于“何时”的问题，地理空间分析则回答“何地”的问题。可以用一个表格展示分析的层级和类型，最左边一列为分析的类型，表格上面一行显示的是分析的层级(见表1.1)。

微观或个体层级项目的分析偶尔可以手工操作完成。例如，可以用笔和纸绘制出5个朋友的网络。但是，中观或地方级别的项目分析在实际操作中不可能手绘完成，通常需要使用计算机。宏观/全球层级的项目分析，往往需要超级计算机来执行必要的计算。

表1.1 分析类型与分析层级

	● 微观/个体层级 (1~100 记录)	● 中观/地方性层级 (101~10 000 记录)	● 宏观/全球层级 (10 000+ 记录)
统计分析/剖析	个人及出版物、专利、资助金额的数量	大型实验室、研究中心、大学、研究领域或州的专业概况	所有国家科学基金、全美国以及所有学科的统计 
时间分析 (何时) 	个人基金组合的不断发展	20年里美国国家科学院院刊主题激增情况描摹 	113年间的物理研究 
地理空间分析 (何地) 	个人的职业轨迹	图绘某个州的知识景观 	国际合作和引用网络 
主题分析 (做什么) 	从一份出版物汲取的基础知识	20年里美国国家科学院院刊主题激增情况 	113年间的物理研究 
关系网络分析 (和谁) 	国家科学基金会项目负责人网络 	合著者网络 	中国科学院全球合作网络 