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

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　山
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

丛书总序

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一的内容，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

作者的话

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。

前　言

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

目录

第1章 可视化架构和工作流程设计……1

(一) 理论部分……1
1.1 可视化架构……2
1.2 工作流程设计……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 工作流程设计……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 工作流程设计……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 工作流程设计……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 工作流程设计……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

(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 聚类和主干识别 ……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

第8章 案例研究 …… 174

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

第1章

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)。

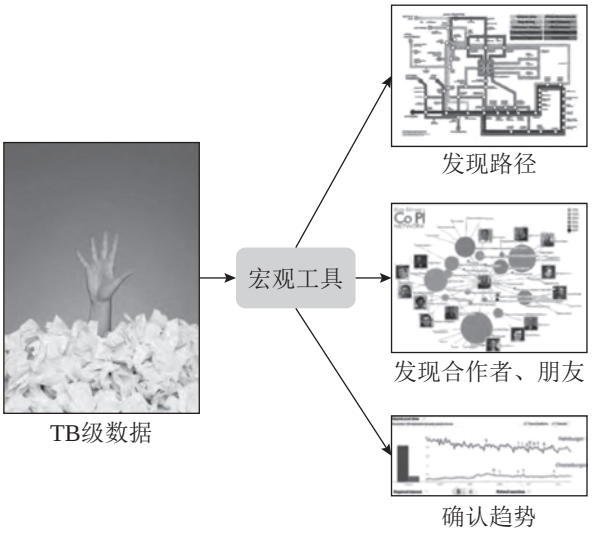


图1.1　将数据转换为随时备用的深刻见解

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 Knowledg*[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。

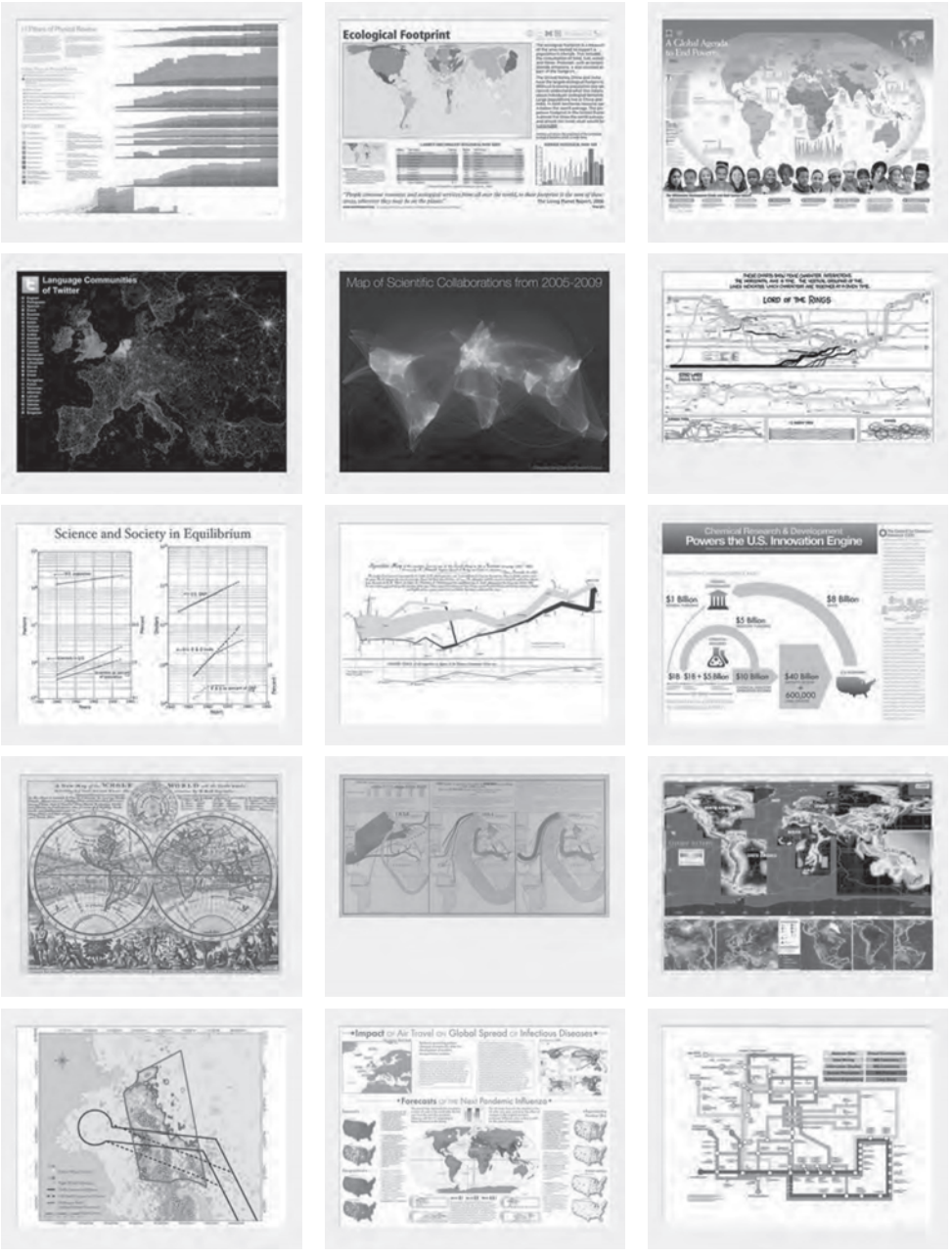


图1.2　《场所与空间：科学图集》中展示的地图集(1)

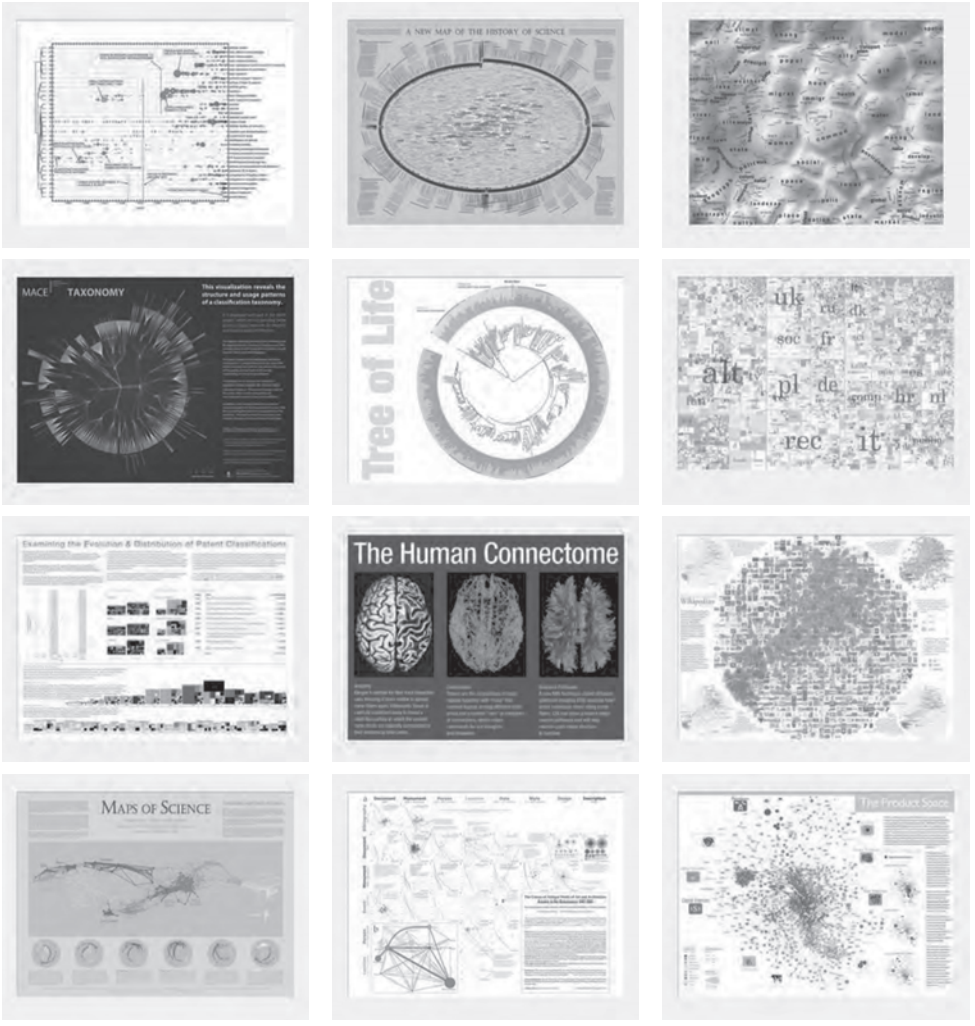
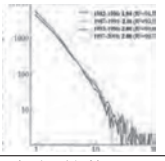
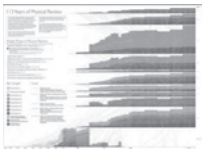
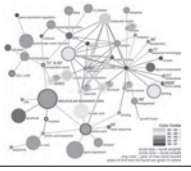
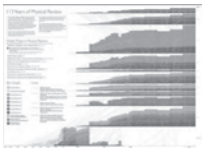


图1.2 《场所与空间：科学图集》中展示的地图集(2)

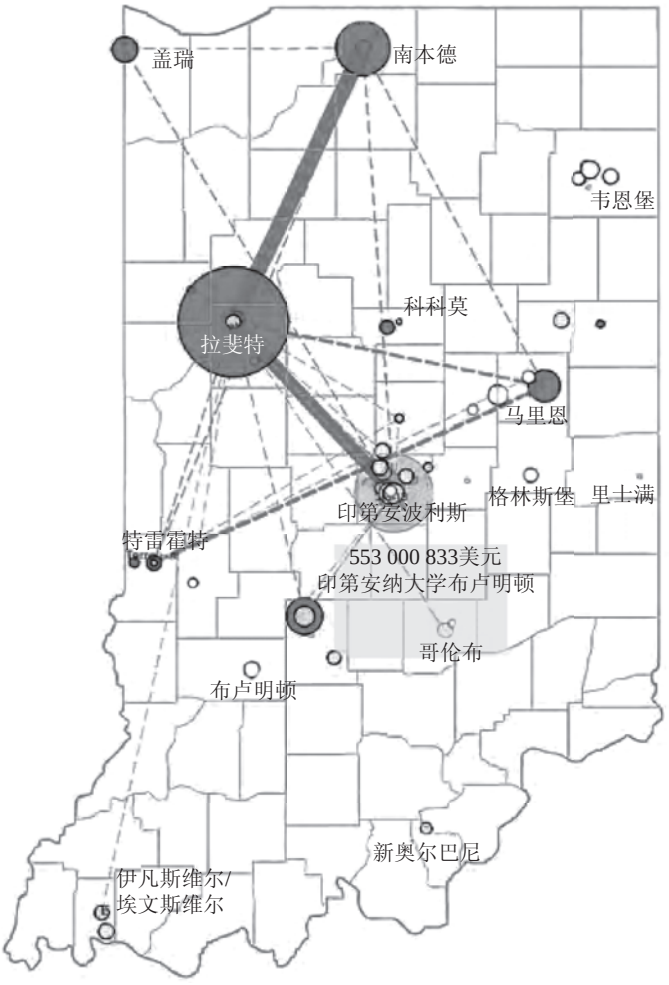
在每一层级的分析中，可能有五种分析类型：统计分析/剖析、时间分析、地理空间分析、主题(topical)分析和网络分析。每种类型的分析都试图回答一种具体类型问题。例如，时间分析有助于回答关于“何时”的问题，地理空间分析则回答“何地”的问题。可以用一个表格展示分析的层级和类型，最左边一列为分析的类型，表格上面一行显示的是分析的层级(见表1.1)。

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

有些项目需要回答多个问题。例如，在对超过113年的物理研究主题演变进行可视化时，时间和主题都得到了回答。表1.1中的几个可视化样本出现了两次。随后，我们讨论了表1.1中精选的全部可视化形式。对于每一种可视化形式，其时间性的、地理空间性的、主题性的以及网络数据的类型和覆盖范围，以及相应的分析类型和层级，均呈现在该图形下方。



数据类型和覆盖范围	
时间范围	2001—2006
区域	印第安纳州
主题领域	生物医药学研究
网络类型	学术-产业合作

分析类型/层级	•	●	⬤
时间的		×	
空间的		×	
主题的			
网络的		×	

图1.3 图绘印第安纳州知识分子空间分布

范例

第一幅可视化形象，如图1.3所示。我们创建的这一图像，旨在展示美国印第安纳州的创新情况以及将创新理念转化为产品的诸多途径。该可视化图像示例表明了中观层级的地理空间分析和网络分析。

对于这一可视化，我们既使用了基金资助的项目数据集，也呈现了未获资助的投标(proposals)。每一个项目都需要业界和学术界的合作，目的是鼓励将研究成果转化为可盈利的产品。将产业、学术机构根据地理空间进行编码的可视化，并将它们的位置和合作网络叠加在印第安纳州的地图上，据此我们创建了这幅可视化图像。也就是说，图中的黑色节点代表学术机构，深灰色节点代表的是产业合作者。节点大小是根据全部收益总金额进行编码的。节点之间的连接表示合作；黑色的连接线表示学界与学界之间的合作，深灰色线表示业界与业界之间的合作，浅灰色线表示学界与业界之间的合作。

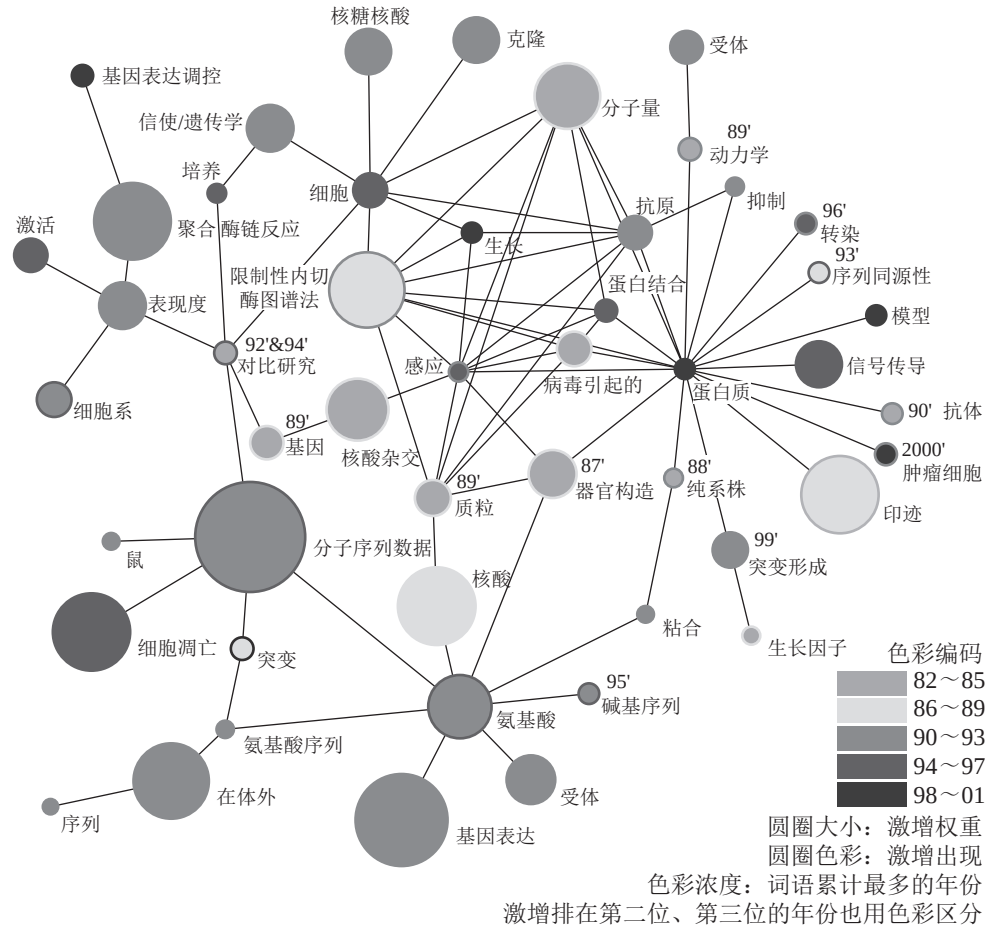
由于此数据集主要涵盖生物医学研究，拉斐特的普渡大学以及印第安纳波利斯的普渡大学分校区(简称IUPUI)在该研究领域具有多种学术和产业活动，所以节点是最大的。拉菲特和印第安纳波利斯之间，以及拉菲特和南本德(圣母大学主校区)之间有着最强的合作关系。

互动性的在线页面支持搜索(例如，可以根据名称和主题查找，按照年份和机构进行筛选，以及用于检索特定的项目和投标的详细信息)。该图像有助于回答以下问题：创意转换为产品的主要途径是什么？何种创意出自何处？这些创意是如何转换为产品并产生收入的？

第二个可视化形象(见图1.4)旨在传达一个数据集的激增情况，该数据集取自美国国家科学院院刊(PNAS)20年的论文发表情况。首先，我们选择了被引用次数最高的前10%的出版物/论文。然后我们通过在第2.4节详细解释的激增检测算法，识别出排名前50最频繁的词语和“激增”[①]词语。我们列出这50个词语及其共同出现的关系作为一个网络，再简化此网络，然后对其进行可视化编码(有关完整的分步工作流程参见第2.6节)。每一个圆圈代表50个单词中的一个，圆圈的大小是根据词语激增权重编码而得的(例如，突然发生频率的增加)。节点的颜色

① Kleinberg，Jon. 2002.“Bursty and Hierarchical Structure in Streams.” In *Proceedings of the 8th ACM/SIGKDD International Conference on Knowledge Discovery and Data Mining*，91–101. New York: ACM.

代表激增情况开始出现的年份。如果需要确定特定节点爆发出现的年份，则需要使用该可视化图像右下角的说明。节点边框的颜色对应的是该词语累计最多频次的年份，同时提供了激增情况排在第二位和第三位的年份。值得注意的是，节点圆圈内的颜色通常都要比外部边框颜色更亮，这表示该词语词第一次出现后，获得了更广泛的使用。



时间类型和覆盖范围	
时间范围	1982—2001
区域	全球
主题领域	生物医学为主
网络类型	同时出现的词语

分析类型/层级	•	●	⬤
时间的		×	
空间的			
主题的		×	
网络的		×	

图1.4　图绘美国国家科学院院刊主题激增情况

如果你将该数据集与2001—2012年间收集的相同数据集做个对比，以“蛋白质”这个词为例，看看在随后的十年里，该词是否得到了广泛的使用。目前，可视化仅限于一种期刊中论文的发表情况；如果增加更多数据以反映一个研究团队所有论文的发表情况，或者一个学科领域中的全部论文的发表情况，那么将具有更大的价值。

第三个可视化形象(见图1.5)使用相同的20年间美国国家科学院院刊的数据集，以了解在互联网时代学者的地理位置是否仍然重要。学者是否仍然需要在主要的研究机构中学习和工作，才能在研究领域中取得高度成功(例如，通过被引用的次数进行测量)？由于数据集捕捉到了互联网的引入和广泛使用情形，我们可能期望看到随着时间的推移，互联网的问世，彼此地理距离无关联的各大研究机构相互引用的数量也会随之增加(见图1.5中右侧的对数坐标图)。换言之，这条曲线将会变得更扁平，因为互联网改善了学者获取研究文章的机会，相应地，他们的引用也会跨越愈加遥远的距离。然而，随着时间的推移，这条曲线的升降却变得更加急剧——随着互联网的问世，研究者对当地文献的引用更多了。

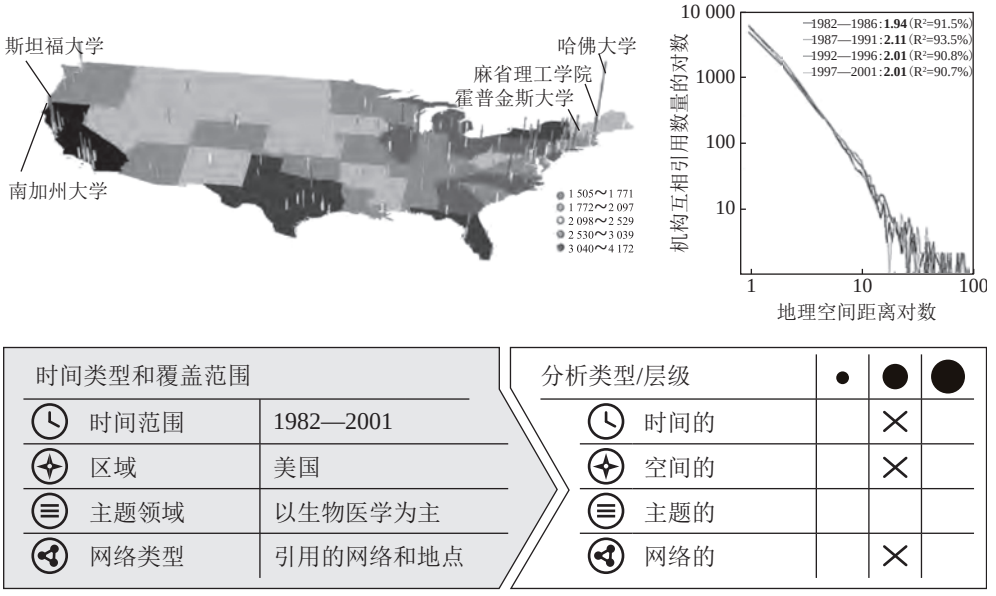


图1.5　美国主要研究机构中时空信息的生产和消费

这个结果与大家的直觉相反，其中，支持该结果的最令人信服的理由是由巴里·威尔曼(Barry Wellman)、霍华德·怀特(Howard White)和南希·纳扎尔

(Nancy Nazer)提出的，他们使用不同的数据获得了类似的结论[①]。他们认为，随着我们越来越多地被信息淹没，社交网络的重要性日益增加。 在本地创建和维护社交网络更容易，因此人们倾向于引用周围同事的论文。

第四个例子(见图1.6)展示了微观层级上的分析，了解的是与凯蒂·伯尔纳(Katy Börner)这位学者在项目中具有合作关系的情况。通过研究她在2001—2006年期间获得的国家科学基金资助的全部项目的数据，我们提取出一个共同研究者网络。这个二重网络的节点有两种类型：项目(绿色)和研究者(用小的白色节点或者如果可以提取到的话，用他们的照片表示)。项目节点通过颜色来编码。颜色是基于该项目获得立项的年份而定，从黑色的节点(2001年)到亮绿色的节点(2006年)。节点大小则与项目获得资助的总额相关。通过设计，数据集中包含了既定时间范围内伯尔纳获得的所有项目(例如：她的节点是最大的且位于网络的中心位置)。只有部分项目是与他人合作的。不过，很容易看出谁在什么项目中是合作关系。例如，努希尔·康崔科特(Noshir Contractor)在该时间范围内参与了三个项目。其中之一，最左边节点标签为NSF IIS-0513650[②]的，是由网络工作台(Network Workbench)资助开发的，这是第一个基于OSGi/CIShell的宏观工具。[③]

第五个可视化例子(见图1.7)显示了1974—2004[④]年之间，在IEEE信息可视化会议上合作发表文章的作者网络。每一个作者节点由其姓氏作为标签，节点大小代表发表论文的数量，颜色基于作者被引用的数量。连接线表示合作关系，线条的粗细代表在论文方面合作的数量多少，颜色则表示他们第一次合作的年份。1986—1990年之间的合作用橙色表示。

总体上看来，马里兰大学的本·史奈德曼(Ben Shneiderman)是最大的节点，这表示在所选取的数据集中，他发表的论文最多。之外，他的节点又呈暗色，这

---

① White，Howard D.，Barry Wellman，and Nancy Nazer. 2004. “Does Citation Reflect SocialStructure? Longitudinal Evidence From the ‘Globenet’ Interdisciplinary Research Group.” *Journal of The American Society for Information Science and Technology* 55，2: 111–126.

② National Science Foundation. 2005. “NetWorkBench: A Large-Scale Network Analysis，Modeling， and Visualization Toolkit for Biomedical，Social Science，and Physics Research，” Award no. 0513650. http://www.nsf.gov/awardsearch/showAward?AWD_ID=0513650 (accessed September 4，2013).

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

④ Ke，Weimao，Katy Börner，and Lalitha Viswanath. 2004. “Major Information Visualization Authors，Papers and Topics in the ACM Library.” Analysis and Visualization of the IV 2004 Contest Dataset. Presented at IEEE Information Visualization Conference，Houston，Texas，October 10–12，2004. This entry won first prize.

说明他的文章得到了广泛引用。而其他的节点，例如约翰·斯特思科(John Stasko)的节点很大但颜色很浅。换言之，像他这类的研究者发表了大量的论文，但是被引用的次数较少。就斯特思科的情况而言，他的论文大部分是近期发表的，还没有出现引用情况。如果再纳入更多年份的数据，这张可视化图像看起来可能完全不同。

整个网络由众多在同一机构工作的研究者紧密联系的子网络组成。只有少数机构型(地理空间型)的网络是互联的。例如，卡内基梅隆大学通过斯蒂芬·艾克(Stephen Eick)与帕洛阿尔托研究中心(缩写为PARC)网络相连接。本·史奈德曼与乔克·麦金利(Jock Mackinlay)和斯图尔特·卡德(Stuart Card)有着紧密的合作关系，而他的学生本·贝德森(Ben Bederson)与贝尔实验室合作很紧密，这进一步增强了相互之间的联系以及马里兰大学团队的影响。

注意史奈德曼在大学的合作网络与乔克·麦金利的“三角”合作关系相当不同。斯图尔特·卡德和乔治·罗伯森在帕洛阿尔托研究中心工作。大多数大学里只有为数很少或者根本没有信息可视化研究人员与其学生一起工作和发表论文，而大部分国家级实验室则具有专门的、已经密切合作了几十年的可视化研究团队。

第六个可视化(见图1.8)使用了相同的数据集，试图了解今天的科学是否是由个别多产的专家或是有很高影响力的合作团队推动的[①]。这张合作网络节点边线条的粗细代表所引用的程度(例如：粗的线条代表合作很成功，被更多的人引用)。这种表示引用程度的新方法，使得考察这一领域研究者相互之间的合作导致的被引用的程度更清晰了，使得在特定领域里，彼此有着合作关系的具有学术成就的研究者之间的引用情况成为了可能。原始论文还提出了一种更新颖的以作者为中心的平均信息量(entropy)，用以识别真正多产的研究团队。结果表明，有效合作的研究人员比单独研究的人员，获得引用的次数更多。

第七个可视化例子(见图1.9)呈现了113年来《物理学评论》(*Physical Review*)的论文发表情况[②]。此图展示了宏观层面的主题分析和引用分析，分析范围是

---

① Börner，Katy，Luca Dall’Asta，Weimao Ke，and Alessandro Vespignani. 2005.“Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams.” In “Understanding Complex Systems，” special issue，*Complexity* 10，4: 57–67.

② Herr II，Bruce W.，Russell J. Duhon，Katy Börner，Elisha F. Hardy，and Shashikant Penumarthy.2008.“113 Years of Physical Review: Using Flow Maps to Show Temporal and Topical CitationPatterns.” In *Proceedings of the 12th International Conference on Information Visualization，London，UK，July 9–11*，421–426. Los Alamitos，CA: IEEE Computer Society Press.

全球性的。从1893年到2005年，左至右显示的《物理学评论》发表的389 899篇论文。1893—1976年的论文数量占据了所绘图像左边的1/3。1977—2000年的217 503篇论文，其中部分是物理学引用情况及天文学分类方案(缩写为PACS)[①] 的数据，占据了图像中间部分的1/3。2001—2005年的80 634篇论文，完整的引用情况和可得到的PACS数据占据了该图像最后的1/3。PACS编码数字在图中最右侧迅速上升。对于每一个不同的PACS编码(例如PACS0：总体)，我们可以看出有多少论文发表在什么期刊上(右边的编码颜色图例)。那些在PACS编码之外的论文则在图像的底部出现。基础性图像的上方，涵盖了2005年出版的论文的所有引用情况。有趣的是，物理学家引用文献在时间层面上跨度很大，上可追溯至19世纪。

每年汤森路透(Thomson Reuters)预测的三位诺贝尔物理学获奖者，都是基于其发表论文的被引频次、高影响力论文以及值得特别重视的论题或发现。小小的诺贝尔奖牌表明了所有获得诺贝尔奖的论文价值，而汤森路透的正确预测再次强调了其价值。

第八个可视化例子(见图1.10)旨在展示美国国家卫生研究院资助的不同项目类型的影响，由研究者发起的资助烟草用途研究出版的论文数量和合著者网络演变与烟草使用跨学科研究中心(缩写为TTURC)资助的同一时间跨度及研究主题项目的对比情况。图1.10中的顶端显示的是确认得到R01基金资助的论文中提取的合著者网络。图中相当密集的合著者网络是从TTURC资助的研究项目中提取的。两个网络中的节点颜色编码代表不同的项目。例如，所有确认受LR01-5资助的(在图1.10的最左上角)作者呈现为蓝色。这其中的一部分作者与多位承认受R01-13资助的作者的合作则用黄色节点表示。

研究者发起的资助和TTURC的资助都推进了烟草使用的研究。虽然研究者发起的资助似乎具有较高的投资回报率，但此可视化图像中却显示每支出一美元引用的次数[②]，而TTURC资助产生了更密集的、更多跨学科的网络，这些网络可以支持跨学科领域信息的更快扩散。

最后一个可视化示例(见图1.11)展示的是北京的中国科学院(CAS)的研究人员

---

① PACS was introduced in 1977.

② Note that TTURC funding is also used for infrastructure development，education，training，or workshops.

全球合作网络[①]。该学院的合作联系围绕在国家层面(例如：中国与美国之间的合作关系呈现为从北京到美国之间大部分相连的点)。国家的颜色是基于对数规模上合作的数量而定，从红色过渡到黄色。北京的中国科学院最深的红色，表示中国科学院的研究者们与全球各地的研究者有3395次合作关系。流向图的布局用于勾勒边界，增强了易读性。每一根线的粗细程度与研究人员之间合作关系的数目成比例。正如可视化显示的那样，中国科学院与美国合作的次数最多。

综上所述，数据分析和可视化可以表现从微观到宏观的不同层级。合著者网络(见图1.6和图1.11)可以通过网络布局或基本的地理空间地图来呈现。在创造可视化图像之前，需要确定旨在回答什么问题以及达到何种层级的分析，确定其时间性、地理空间性、主题和网络范围，因为这有助于阐释可视化并最终将所传达的信息用于决策过程。

1.2　工作流程设计

本节将介绍如何设计满足用户需求的工作流程。我们探究的可视化类型包括表格、图形、网络和地理空间地图。数据叠加，可以加载在不同的基本的地图上。数据叠加可以使用图形的可变类型对附加信息进行编码，比如颜色、形状和大小的编码。深刻理解可视化类型、数据叠加和图形可变类型，不仅是可视化设计的需要，也是解释可视化的需要。

需求驱动工作流程设计的迭代过程，如图1.12所示。第一步是要找出相关决策者在其生活中的所想和所需，以及数据挖掘和可视化技术如何能够改善他们的决策过程。在了解用户种种需求之后，就可以确定分析的类型和层级。下一步是获得最高品质和覆盖最多的数据，而这往往受预算和时间所限而不得不做出一些妥协。数据经常由相关决策者提供，或者从重要的数据提供商处购买，或从公开的资源中检索获得。需要解析和阅读数据(**读取**)，也可能需要广泛地整理和预处理数据。需要做时间的、主题的以及其他类型的分析，以确认趋势和模式(**分析**)。可视化阶段(**可视化**)包括三个主要步骤。首先，必须确定适当的参考系统。该参考系统需要成为基础性地图以稳定地承载分层数据。第二，参考系统可能

① Duhon，Russell Jackson. 2009.“Understanding Outside Collaborations of the Chinese Academyof Sciences Using Jensen-Shannon Divergence.”In *Proceedings of the SPIE Conference onVisualization and Data Analysis 2009，San Jose，CA，January 19*，edited by Katy Börner and JinahPark. Bellingham，WA: SPIE Press.

被调整(例如，可能会将美国的一张地图做变形处理，以显示选举结果；具体例子可以见第3.2节的图3.9)。第三，使用不同的图形对附加的数据变量进行视觉编码。最终，必须将可视化图像发布出去(印刷出来、在网上发布等)，最后，呈现的可视化要能够为相关决策者所验证和解释。

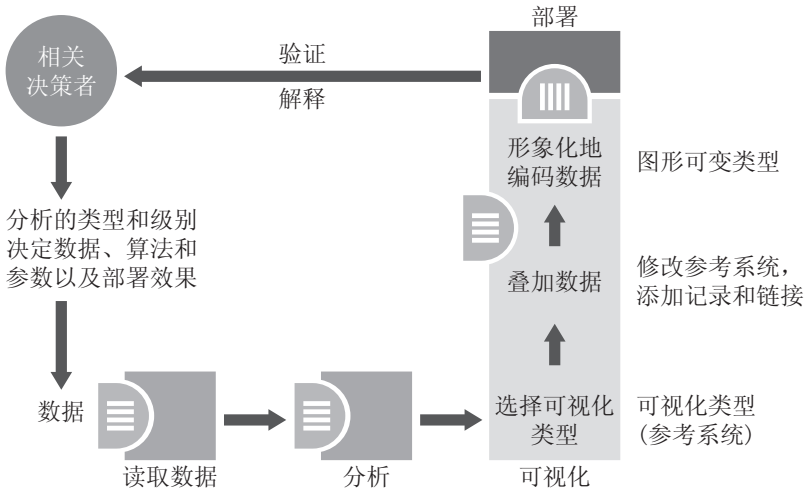


图1.12　需求驱动的工作流程设计

重要的是，应记住信息可视化是一个迭代过程，在验证和解释阶段，相关决策者可能会认识到缺失的数据或提出新的要求。随着用户需求的变化，需要不断重复设计过程：获取新数据，读取数据，进行分析，重新设计可视化和再次发布，最后与相关决策者分享新的可视化效果。

可视化类型

在本书中，我们讨论五种主要的可视化类型：**图表**、**表格**、**图形**、**地理空间地图和网络图**。图1.13中分别展示了这五种可视化类型。

图表是指无固有参考系统的可视化方式(见图4.5)，如饼图和文字云。

表格是一种简单但非常有效的传达数据的方式(见图1.9)。表格单元格可以通过颜色进行编码或排序，也可以包含图形符号或者微型图标(见图4.6)。有些表格支持使用分类轴浏览数据层次结构(见图4.6)。

图形是最常用的可视化类型，可以显示定量或定性的数据，如时间轴、条形图和散点图。本书提供了许多图形的例子。

地理空间地图使用经纬度作为参考系统，如世界地图、城市地图和地形图等。

图1.13　各种可视化结构下的可视化效果

第五种类型是**网络图**，如层次结构或分类的树状图、社会网络或迁移流动的网络图等。

每个可视化类型定义了自身的参考系统或基本地图，并将数据呈现在可视化结果中。例如一个机构必须提供其所在纬度和经度信息，才能将其呈现在一幅地理空间地图上。

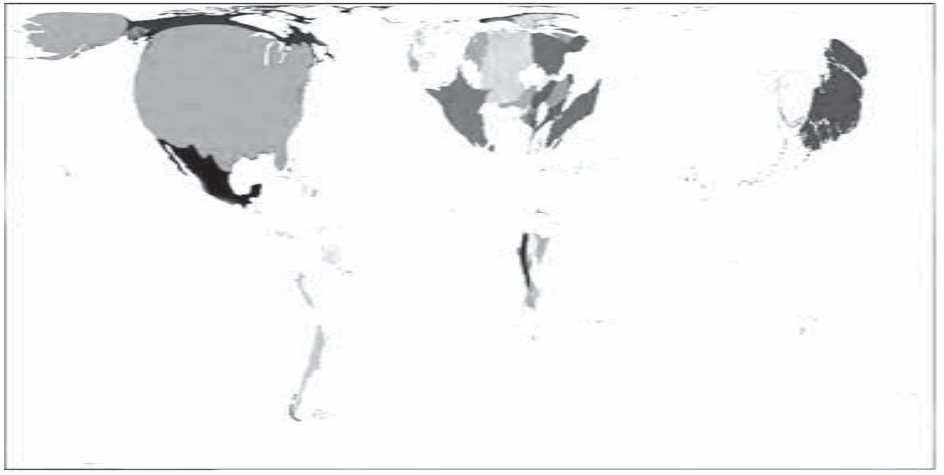
确定适当的可视化类型和参考系统后(见图1.12的**选择可视化类型**步骤)，就可以叠加数据了。对数据属性进行视觉编码的方式主要有三种： (1)对基本地图进行变形；(2)场所记录；(3)通过视觉变量类型进行编码。在图1.12所示的工作流程设计中，类型(1)和类型(2)称为**叠加数据**，而类型(3)称为**视觉编码数据**。

此外，为了给可视化图像提供必要的语境，时常需要添加标题、标签、图例和说明文字。适当地交代图形创作者，用户可以向作者反馈其意见和建议，或者让地图创作者更好地为其服务。

以下是来自《场所与空间：测绘科学》中展示的例子(http://scimaps.org)。第一个例子(见图1.14)[①]显示了一个地理区域范围，也称为**统计地图**的变形情况。统计地图可视化的前提假设是熟悉地理区域的实际形状和范围。在这里，大地图通过对国家大小的变形来代表人们的生态足迹，也就是这些人实际上需要多少面积才能维持其生活方式。把变形过的地图和位于左下角的原始地图进行比较，可以看到美国还有相当大的面积，而非洲许多国家只有非常小的面积。

① Dorling. Daniel. Mark E. J. Newman，and Anna Barford. 2010. *The Atlas of the Real World: Mappingthe Way We live*. Revised and expanded. London: Thames & Hudson.

生态足迹



生态足迹是以支持人口生活方式所需面积而衡量的。这包括食物、燃料、木材和纤维的消费。污染，如二氧化碳排放，也被视为生态足迹的一部分。美国、中国、印度生态足迹最大。不了解人口规模，我们无法理解这对个人生态需求意味着什么。中国和印度有着庞大的人口。这两个地区资源利用低于世界平均水平。美国的人均生态足迹几乎是世界平均水平的5倍，差不多是可持续发展的10倍。

区域的大小显示出世界各地现存的生态足迹的比例。

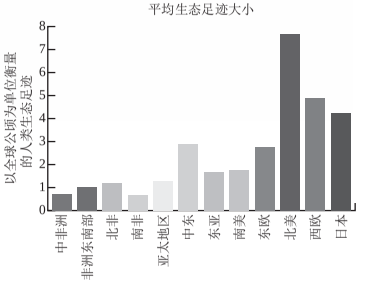
土地面积

技术说明
- 数据来自世界自然基金会国际和动物学研究所
- 生态足迹以全球公顷衡量，一全球公顷就是具有世界平均生物生产力的面积
- 更详细信息请参阅网站

最大和最小生态足迹面积

排名	区域	值	排名	区域	值
1	阿拉伯联合王国	10.6	191	尼泊尔	0.61
2	美国	9.7	192	刚果共和国	0.58
3	格陵兰	7.7	193	赞比亚	0.58
3	巴哈马	7.7	194	刚果	0.58
5	加拿大	7.5	195	马拉维	0.57
6	科威特	7.4	196	海地	0.57
7	澳大利亚	7.0	197	柬埔寨	0.55
8	芬兰	6.8	198	孟加拉国	0.47
9	爱沙尼亚	6.1	199	索马里	0.23
10	新西兰	6.1	200	阿富汗	0.11

2002年人均全球公顷生态足迹



“人们消耗的是世界各地的资源和生态，无论他们位于地球何处，人们的生态足迹是这些地区的总和。”

2006年地球生命力报告

www.worldmapper.org ©Copyright 2006 SASI Group (University of Sheffield) and Mark Newman (University of Michigan)

图1.14 《生态足迹(2006)》，作者是丹尼·多林(Danny Dorling)、马克E. J. 纽曼(Mark E. J. Newman)、格雷厄姆·奥尔索普(Graham Allsopp)、安娜·巴福德(Anna Barford)、本·惠勒(Ben Wheeler)、约翰·普里查德(John Pritchard)和大卫·道灵(David Dorling)(http://scimaps.org/IV.6)

我们还可以使用世界地图和视觉编码的地图创建一个**地区分布图**。例如，由世界银行数据中心(World Bank Data Group)、国家地理和联合国联袂出品的《千年发展目标》(*The Millennium Development Goals Map*)(见图1.15)[①]，使用颜色编码显示2004年全球相对收入水平，以美元计的人均国民总收入(GNI)。颜色从红色过渡到深绿色，代表低收入(825美元或更低的)地区过渡到高收入(10 106美元或更高)的地区。

下一个例子(见图1.16)，将推特(Twitter)的数据做地理定位并叠加在欧洲地图上。图中的各个点是按照语言进行颜色编码，如有许多荷兰语的推文(淡蓝色)大多数分布在荷兰。同样，意大利语的推文(深蓝色)主要分布在意大利。此外，法国的巴黎和其他省会城市在这些地图上占主导地位，显然，城市地区的推文多于农村地区。

下一个例子[②]展示了2005—2009年间，世界性的科学合作联系图叠加在一幅世界地图上(见图1.17)。利用爱思唯尔的斯高帕斯(Elsevier Scopus)数据库确认谁与谁存在合作关系，可以立即识别出一些合作模式。例如，美国的研究人员相互合作的现象相当常见，欧洲和亚洲也是如此。不仅各大洲之间合作十分普遍，而且不同大陆间的合作也很多，特别是那些使用相同语言的人们之间。

到目前为止，讨论的所有示例都使用了一个或几个定义清晰的参考系统，如图表、表格、图形、地理空间图和网络图等。数据叠加的主要方式包括修改基本图(例如，统计地图)、场所记录和联系、运用图形变量类型进行视觉变量编码(点、线、区域、外观)。

图形变量类型

我们已经介绍了多种不同的视觉编码方式并用于数据可视化的编码。对于不同方式详细的评论和比较，请参阅《知识图集》[③]。通常，图形变量类型包括**位置**、**形式**和**颜色**。不太常用的图形变量类型包括**质地**(texture)和**光学效果**，参见图1.18的综述部分。

---

① Department of Public Information，United Nations. 2010. *We Can End Poverty 2015: Millennium Development Goals.* http://www.un.org/mullenniumgoals (accessed August 2，2013).

② Beauchesne，Olivier H. 2011. *Map of Scientific Collaborations from 2005 to 2009*. http://collabo.oIinb.com/ (accessed August 2，2013}.

③ Burner，Katy. 2014. *Atlas of Knowledge: Anymore Can Map*. Cambridge，MA.:The MIT Press.

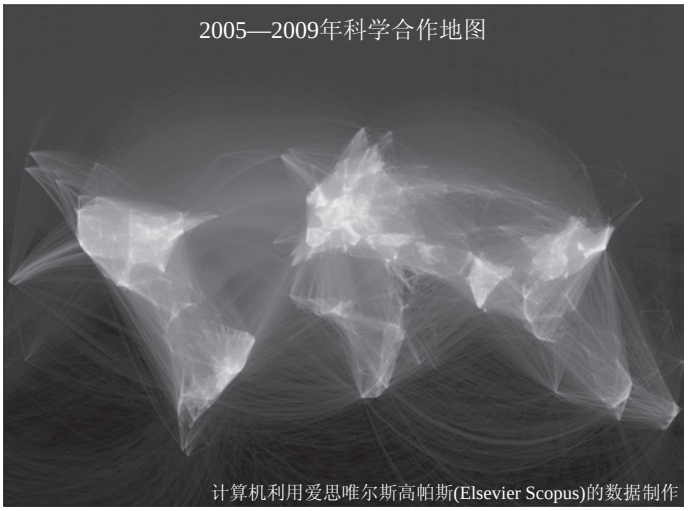


图1.17　奥利维尔·H. 鲍彻斯尼(Olivier H. Beauchesne)制作的世界城市之间的科学合作地图(2012)(http://scimaps.org/VII.6)

位置

如果是三维的可视化，**位置**通常包括*x*，*y*坐标，有时还有*z*坐标。请记住，三维可视化很难阅读，因为很容易出现相互遮挡数据点的情况，透视图呈现的近处对象往往大于远处的对象，还有近处物体的颜色比远处的物体更明亮。如果以三维呈现物体，由于聚光灯和阴影的原因，要想识别真实的颜色难度会更大。

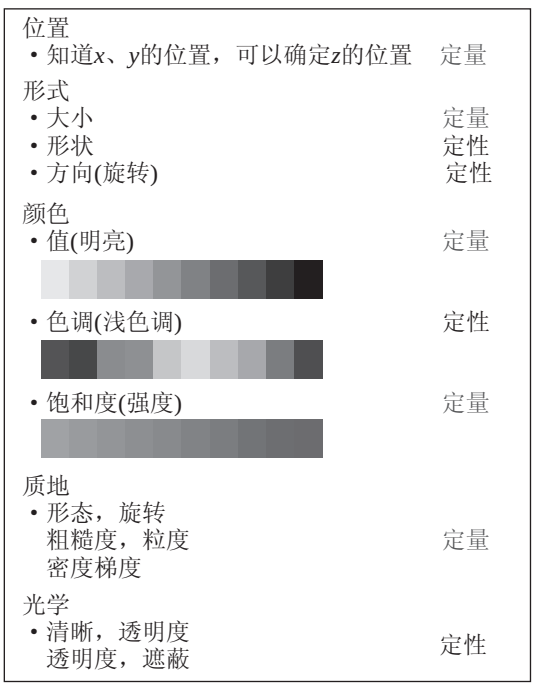


图1.18　使用颜色进行可视化编码

形式

形式包括大小和形状两种特别常用的视觉编码形式。此外，视觉编码可承载一定的意义。例如树形图图标垂直放置的话，则表示继续发挥作用，如果图标是水平放置的，则说明已经被淘汰了。

颜色

颜色有三个属性值：明暗、色调和饱和度(见图1.18)。

质地

质地可以用来编码数据属性的类型，包括形态、旋转、粗糙程度、粒度(size)或密度梯度。

光学效果

光学效果包括清晰、透明和遮蔽等。

数据测量类型

有四种数据测量类型：定类测量、定序测量、定距测量和定比测量(见图1.19)。

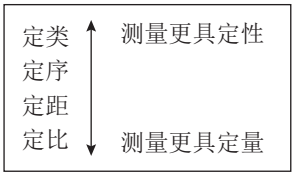


图1.19　从定性到定量的数据变量类型

定类

定类变量是定性的，也称为定类变量或类别变量。一般认为类别是非重叠的，如构成名称及描述人、地、物或事件的词语或数字。

定序

定序变量是定性的，数据以序列或按等级排序，例如星期几。李克特量表也是一种排序类型，比如从非常贫穷到非常富裕。

定距

区间测量，也称为数值数据，是一种定量测量，其中任何两个相邻值或间隔之间的距离相等，但零点是任意的，如以摄氏度为单位的温度或以小时为单位的时间。

定比

定比测量是定量的，被称为比例范围。定比表示数值是按一种顺序排列的，如重量或高度。在定比变量中，有一个具有实际意义的零点。

运用图形变量类型对数据进行视觉编码，必须考虑数据变量类型的测量范围。例如，定类变量，比如职业，可以使用色调进行编码，这是定类和定性测量。性别不应该使用定量的颜色变量，诸如对色值或饱和度进行编码(见图1.18)。更多的例子将在第1.3节讨论。

自我测评

1. 研究学者的职业轨迹需要什么样的层级分析？

a. 微观　　b. 中观　　c. 宏观

2. 什么样的“分析类型”最适合研究和可视化前100名研究型大学的研究人员数量？

a. 时间的　　b. 地理空间的　c. 主题的　　d. 网络的

3. 哪些可视化类型具有量化的或质化的轴线？

a. 图表　　b. 表格　　c. 图形　　d. 地理空间图

e. 网络图

(二) 实践部分

1.3 示例

我们必须知道不同数据变量类型的测量范围，才能有效地应用图形变量类型(见第1.2节)。某些图形变量类型更适合定性数据的编码，因此，我们应该使用其他数据变量对定量数据进行编码。

定性图形变量类型包括色调和定向，我们可以将之用于定类或定序数据的编码。值得注意的是，有一些形状人们无法命名，而且人们不善于对某些形状旋转的程度做出评估。

大小和色值都是定量的，通常将之用于基于定距或定比数据基础上的几何符号的大小和值的编码。

举一个具体的例子，表1.2中列出的不同数据属性——名称、年龄、邮政编码和专业，确定它们是定量的还是定性的。其中有些变量相当明显，如年龄是定量的，专业是定性的。邮政编码看起来是定量的，因为是用数字表示的。然而，美国邮政编码的数字是不连续的，它们是定性的参考编号。只有当邮政编码被转换成经度和纬度的数值时，才会变成定量的。

表1.2　数据量表类型示例

姓名	年龄	邮编	职业
伊娃	30	47401	教师
布鲁斯	46	47405	工程师
定性	定量	定性	定性

表1.3显示不同的数据属性以及可以使用不同的图形变量类型进行视觉编码。在可视化过程中创建出这些图形(见图1.20中的工作流设计)。也就是说，原

始数据或计算处理过的数据变量被绘制到图形变量类型之中。不同数据变量的数据量表类型影响到所采用的视觉编码。

表1.3　图形变量与数据量表类型

	姓名	年龄	邮编	职业
方位		×	×	
范围		×	×	
色调	×			×
外形	×			×
	定性	定量	定量	定性

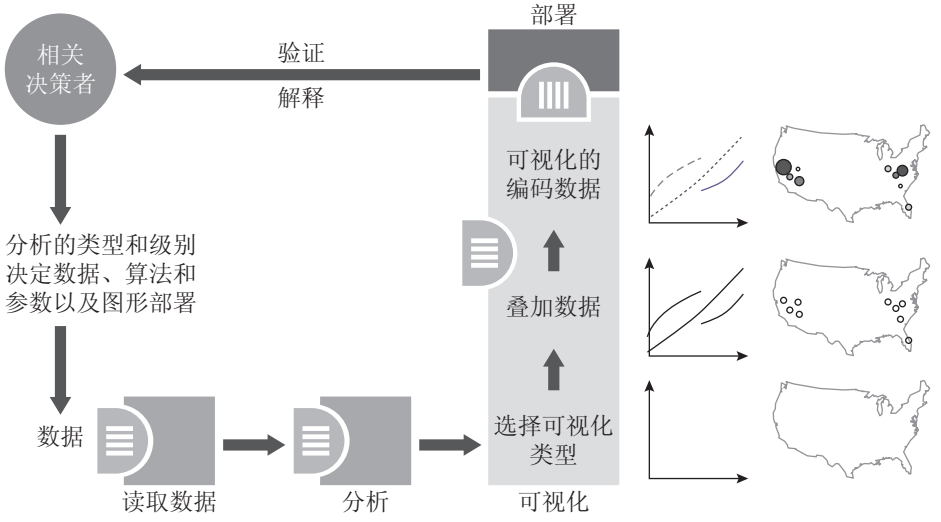


图1.20　在图形和地理空间地图上的视觉编码过程示例

图1.20中例示了以下内容：

(1) 选择可视化参考系统；

(2) 数据叠加；

(3) 使用线型、颜色和大小将以上内容编码到一幅简单的图形和地理空间图中，对数据变量进行视觉编码。遵循下面的步骤，会让可视化过程的设计更容易。

1.4　下载并安装SCI2

我们将使用Sci2工具1.1版本开展时间分析(第2章)、地理空间分析(第3章)、主题分析(第4章)和网络分析(第5、6章)，并呈现视觉效果。Sci2工具是一个所谓

的即插即用的宏观工具[①]，可以轻松扩展和定制以满足不同的需求，参见附录如何添加其他插件部分。Sci2是一个独立的桌面应用程序，可在所有常见的操作系统中安装和运行。该软件需要预装Java SE 5 32位1.5.0版本或更高级的版本[②]。

要下载Sci2工具，请转到Sci2主页[③]并单击“下载Sci2工具”按钮。注册是免费的，只需要几分钟就可以安装好下载的工具。单击“立即注册”按钮后，检查电子邮件确认说明，创建密码，通过您选择的操作系统下载工具。Sci2工具及其相关文件将被压缩在一个文件中，保存这个压缩文件，下载完成后解压，双击Sci2图标(sci2.exe)运行程序(见图1.21)。请确保在文件目录中保存和运行该程序，其中该工具具有创建新文件(如，日志文件)的权限。此后桌面将会运行程序，但是程序文件目录有时会出错。

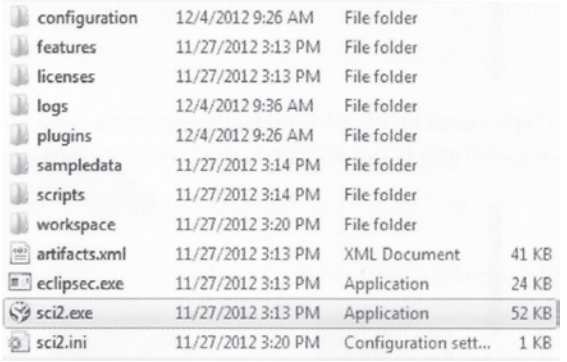


图1.21 Sci2下载目录显示Sci2(sci2.exe)图标被选中

Sci2在所有操作系统中的安装和使用都是一样的，这本书的实践部分是使用PC端进行演示， 但对于Mac和Linux操作系统，每个操作流程都是相同的。

运行Sci2后，闪现画面出现后不久Sci2界面将会弹出。界面分为4个部分，顶部的菜单，底部的“控制台”，左下角的是“调度”，右边则是“数据管理器”(见图1.22)。

该菜单提供了“文件”加载及“数据准备”“预处理”“分析”“建模”和“可视化”算法的便捷访问。主菜单从左到右按照一个正常工作流程中的常见步骤而组织，在每个菜单中，算法按分析类型组织(例如，时间分析，地理空间分析，主题分析，网络分析)。

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

② 可在Java网站运行测试版: http://www.java.com/en/dawnload/installed.jsp

③ http://sci2.cns.iu.edu

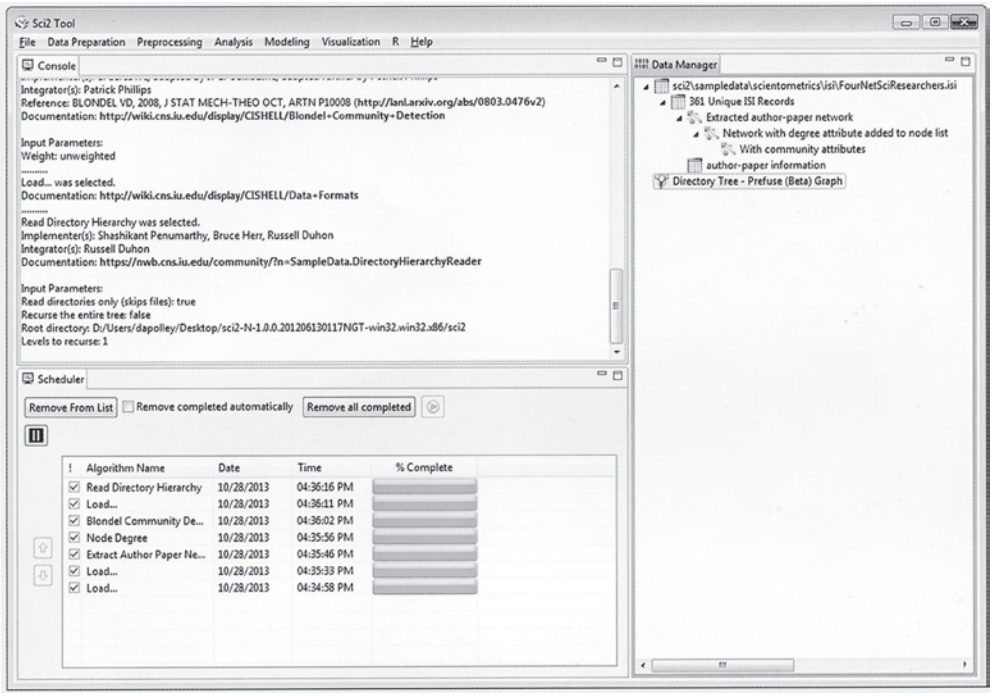


图1.22　Sci2用户界面

数据管理器跟踪全部加载到Sci2的所有文件，以及在分析期间生成的所有后续文件。在一个文件上执行的所有操作都以树状结构嵌套在原始文件之下， 加载文件的类型由其图标说明。

- 文本：文本文件
- 表格：表格数据(电脑化文件)
- 矩阵：数据(Pajek.mat)
- 绘制：可以使用常用指令(Gnuplot)绘制图形的纯文本文件
- 网络：网络数据(内存图形/网络对象或保存为Graph/ML，XGMML，NWB，Pajek.net或Edge列表格式的网络文件)
- 树：树状数据(树状图)

右键单击文件并选择“保存”或“查看”后，可以从“数据管理器”中轻松查看和保存文件。要从“数据管理器”中删除文件，请单击“右键”并选择“删除”。如果删除了一个文件，该文件下嵌套有其他文件，那么所有嵌套文件都将被删除。

“控制台”将会记录会话期间执行的所有操作；它还显示错误消息和警告，致谢该算法原始作者、开发人员、集成商以及参考文献等信息，并在Sci2维基中[①]链接在线文档。同样的信息也储存在你的Sci2目录/日志目录里。

最后，“调度器”显示在工具中执行的操作的进度。

要卸载Sci2工具，只需删除你的Sci2目录。这将删除所有子目录，所以请确保备份所有你想保存的文件。

家庭作业

在Sci2(http://sci2.cns.iu.edu)上注册和下载工具。安装后运行它。使用文件>加载，加载一个电脑化(CSV)文件(如，加载KatyBorner.nsf[②])。在“数据管理器”中右键单击该文件，选择“查看”，用微软电子表格打开它，或选择“查看……”，在另一个电子表格程序打开它。在“获得资助金额日期”这一列的类别中，回答的是：资助金额最高的是什么项目？然后，在“数据管理器”中右键单击该文件，将该文件“重命名”，然后“删除”该文件。

① http://sci2.wiki.cns.iu.edu
② your sci2directory/sample data/scientometrics/nsf