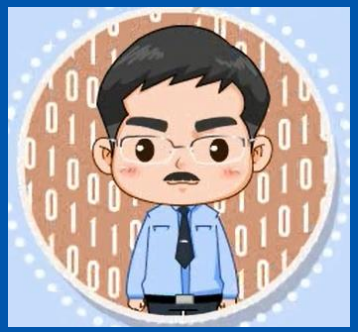


《大数据导论》

教材官网： http://dblab.xmu.edu.cn/post/bigdata-introduction/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第8章 数据可视化

（PPT版本号：2020年秋季学期）

扫一扫访问教材官网

林子雨 博士/副教授

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页： **http://dblab.xmu.edu.cn/post/linziyu**

课程教材

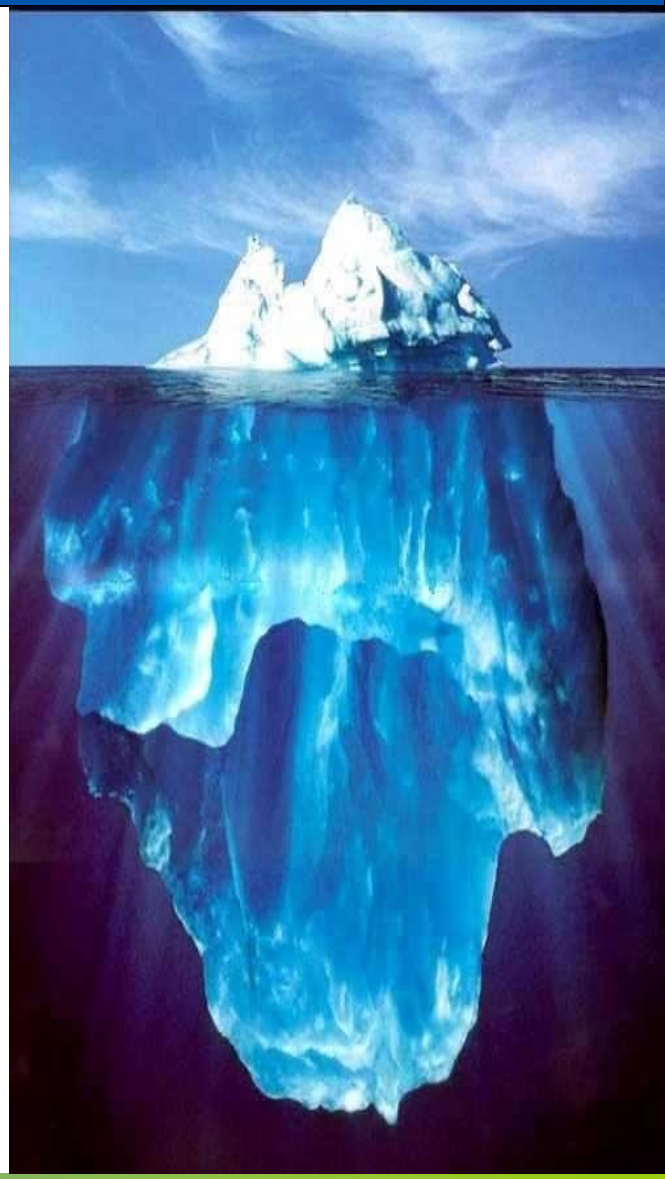
- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

扫一扫访问教材官网

提纲

- 8.1 可视化概述
- 8.2 可视化图标
- 8.3 可视化工具
- 8.4 可视化典型案例



8.1可视化概述

8.1.1 什么是数据可视化
8.1.2 可视化的发展历程
8.1.3 可视化的重要作用

8.1.1什么是数据可视化

- 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
- 数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示，大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析

8.1.2 可视化的发展历程

霍乱地图分析了霍乱患者分布与水井分布之间的关系，发现在有一口井的供水范围内患者明显偏多，据此找到了霍乱爆发的根源是一个被污染的水泵

图8-1 反映霍乱患者分布与水井分布的地图

8.1.2 可视化的发展历程

数据可视化历史上的另一个经典之作是1857年“提灯女神”南丁格尔设计的“鸡冠花图”(又称玫瑰图)，它以图形的方式直观地呈现了英国在克里米亚战争中牺牲的战士数量和死亡原因，有力地说明了改善军队医院的医疗条件对于减少战争伤亡的重要性

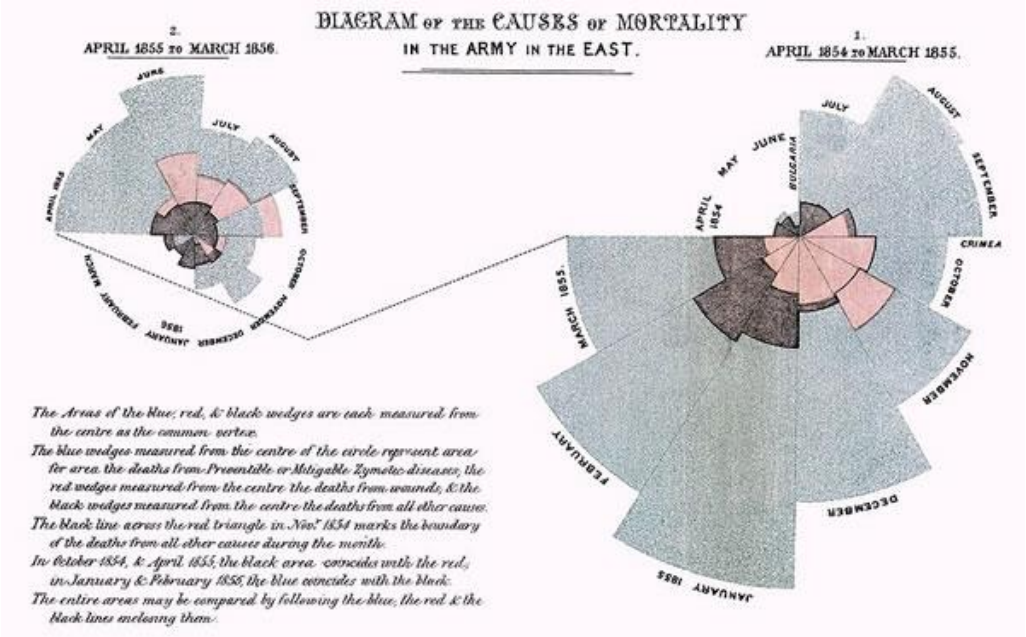


图8-2 “提灯女神”南丁格尔设计的“鸡冠花图”

8.1.2 可视化的发展历程

- 20世纪50年代，随着计算机的出现和计算机图形学的发展，人们可以利用计算机技术在电脑屏幕上绘制出各种图形图表，可视化技术开启了全新的发展阶段。最初，可视化技术被大量应用于统计学领域，用来绘制统计图表，比如圆环图、柱状图和饼图、直方图、时间序列图、等高线图、散点图等，后来，又逐步应用于地理信息系统、数据挖掘分析、商务智能工具等，有效促进了人类对不同类型数据的分析与理解
- 随着大数据时代的到来，每时每刻都有海量数据在不断生成，需要我们对数据进行及时、全面、快速、准确的分析，呈现数据背后的价值，这就更需要可视化技术协助我们更好地理解和分析数据，可视化成为大数据分析最后的一环和对用户而言最重要的一环

8.1.3 可视化的重要作用

在大数据时代，可视化技术可以支持实现多种不同的目标：

（1）观测、跟踪数据

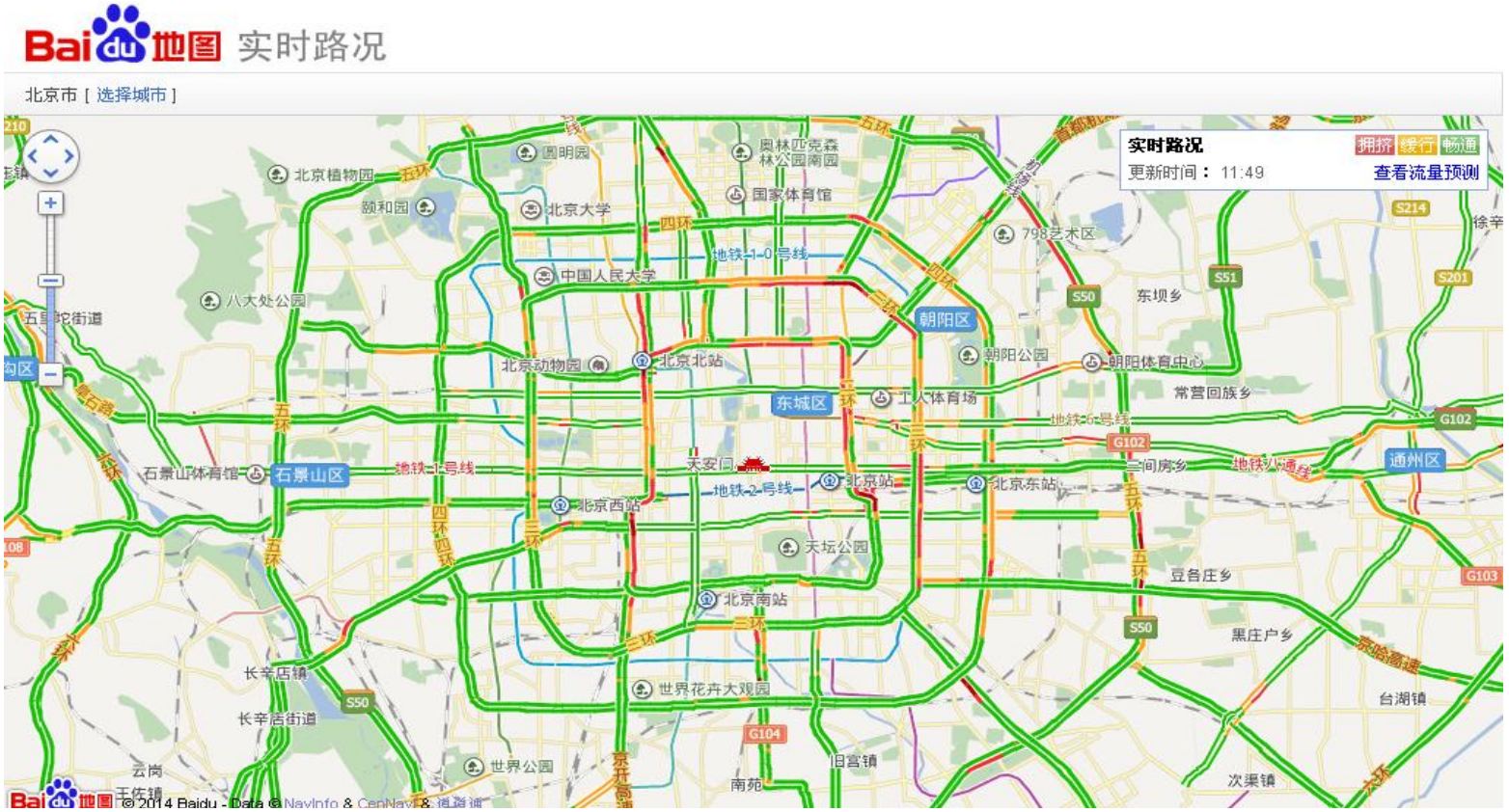


图8-3 百度地图显示的北京市实时交通路况信息

8.1.3 可视化的重要作用

（2）分析数据

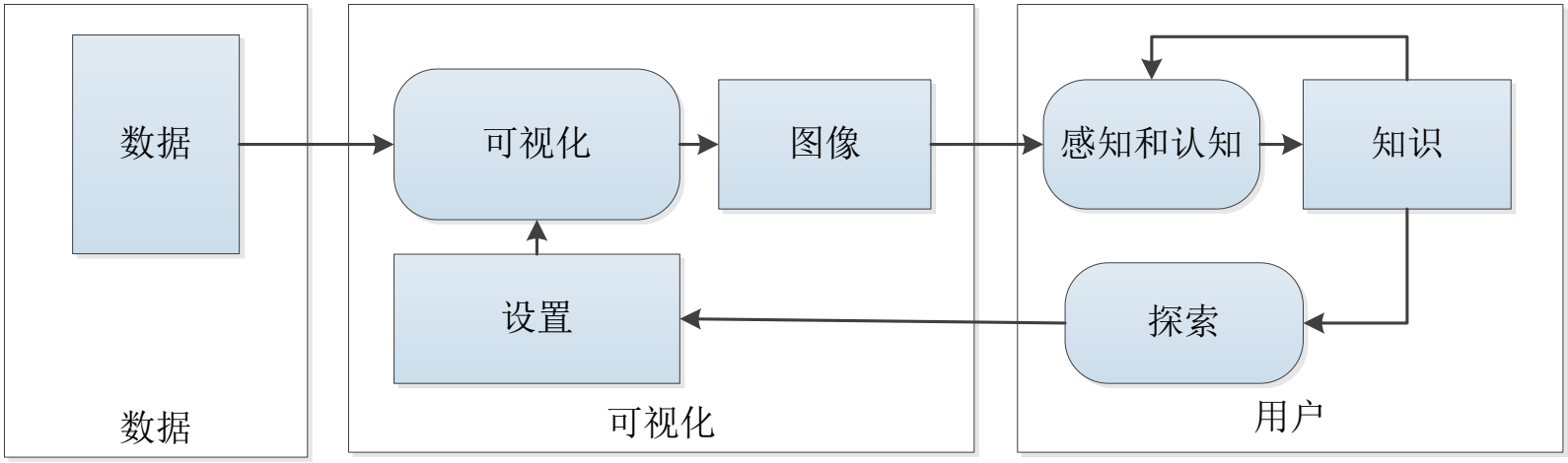


图8-4 用户参与的可视化分析过程

8.1.3 可视化的重要作用

（3）辅助理解数据

图8-5 微软“人立方”展示的人物关系图

8.1.3 可视化的重要作用

（4）增强数据吸引力

图8-6 一个可视化的图表新闻实例

8.2可视化图表

表 最常用的统计图表类型及其应用场景

图表	维度	应用场景
柱状图	二维	指定一个分析轴进行数据大小的比较，只需比较其中一维
折线图	二维	按照时间序列分析数据的变化趋势，适用于较大的数据集
饼图	二维	指定一个分析轴进行所占比例的比较，只适用于反映部分与整体的关系
散点图	二维或三维	有两个维度需要比较
气泡图	三维或四维	其中只有两维能够精确辨识
雷达图	四维以上	数据点不超过6个

8.2可视化图表

除了上述常见的图表以外，数据可视化还可以使用其他图表，具体如下：
（1）漏斗图。漏斗图适用于业务流程比较规范、周期长、环节多的流程分析，通过漏斗各环节业务数据的比较，能够直观地发现和说明问题所在。
（2）树图。树图是一种流行的、利用包含关系表达层次化数据的可视化方法，它能将事物或现象分解成树枝状，因此又称“树型图”或“系统图”。树图就是把要实现的目的与需要采取的措施或手段，系统地展开，并绘制成图，以明确问题的重点，寻找最佳手段或措施。
（3）热力图。以特殊高亮的形式显示访客热衷的页面区域和访客所在的地理区域的图示，它基于GIS坐标，用于显示人或物品的相对密度。
（4）关系图。基于3D空间中的点线组合，再加以颜色、粗细等维度的修饰，适用于表征各节点之间的关系。
（5）词云。通过形成“关键词云层”或“关键词渲染”，对网络文本中出现频率较高的“关键词”给予视觉上的突出。
（6）桑基图。也被称为“桑基能量分流图”或“桑基能量平衡图”，它是一种特定类型的流程图，图中延伸的分支的宽度对应数据流量的大小，通常应用于能源、材料成分、金融等数据的可视化分析。
（7）日历图。以日历为基本维度的、对单元格加以修饰的图表。

8.3 可视化工具

8.3.1 入门级工具
8.3.2 信息图表工具
8.3.3 地图工具
8.3.4 时间线工具
8.3.5 高级分析工具

8.3.1 入门级工具

- Excel是微软公司的办公软件Office家族的系列软件之一，可以进行各种数据的处理、统计分析和辅助决策操作，已经广泛地应用于管理、统计、金融等领域

8.3.2 信息图表工具

信息图表是信息、数据、知识等的视觉化表达，它利用人脑对于图形信息相对于文字信息更容易理解的特点，更高效、直观、清晰地传递信息，在计算机科学、数学以及统计学领域有着广泛的应用。

1. Google Chart API

谷歌公司的制图服务接口Google Chart API，可以用来为统计数据并自动生成图片，该工具使用非常简单，不需要安装任何软件，可以通过浏览器在线查看统计图表。

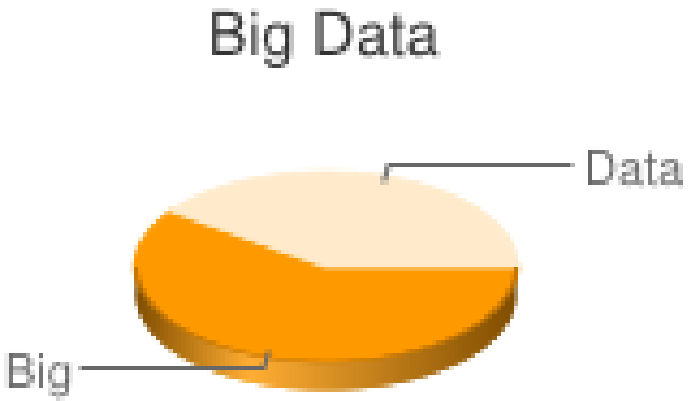


图8-7 通过浏览器在线查看Google Chart统计图表

8.3.2 信息图表工具

2. ECharts

ECharts是由百度公司前端数据可视化团队研发的图表库，可以流畅地运行在PC和移动设备上，兼容当前绝大部分浏览器（IE8/9/10/11、Chrome、Firefox、Safari等），底层依赖轻量级的、Canvas类库ZRender，可以提供直观、生动、可交互、可高度个性化定制的数据可视化图表。

ECharts提供了非常丰富的图表类型，包括常规的折线图、柱状图、散点图、饼图、K线图，用于统计的盒形图，用于地理数据可视化的地图、热力图、线图，用于关系数据可视化的关系图、treemap，用于多维数据可视化的平行坐标，以及用于BI的漏斗图、仪表盘，并且支持图与图之间的混搭，能够满足用户绝大部分分析数据时的图表制作需求。

8.3.2 信息图表工具

3. D3

D3是最流行的可视化库之一，是一个用于网页作图、生成互动图形的JavaScript函数库，提供了一个D3对象，所有方法都通过这个对象调用。D3能够提供大量线性图和条形图之外的复杂图表样式，例如Voronoi图、树形图、圆形集群和单词云等（如图10-8所示）。

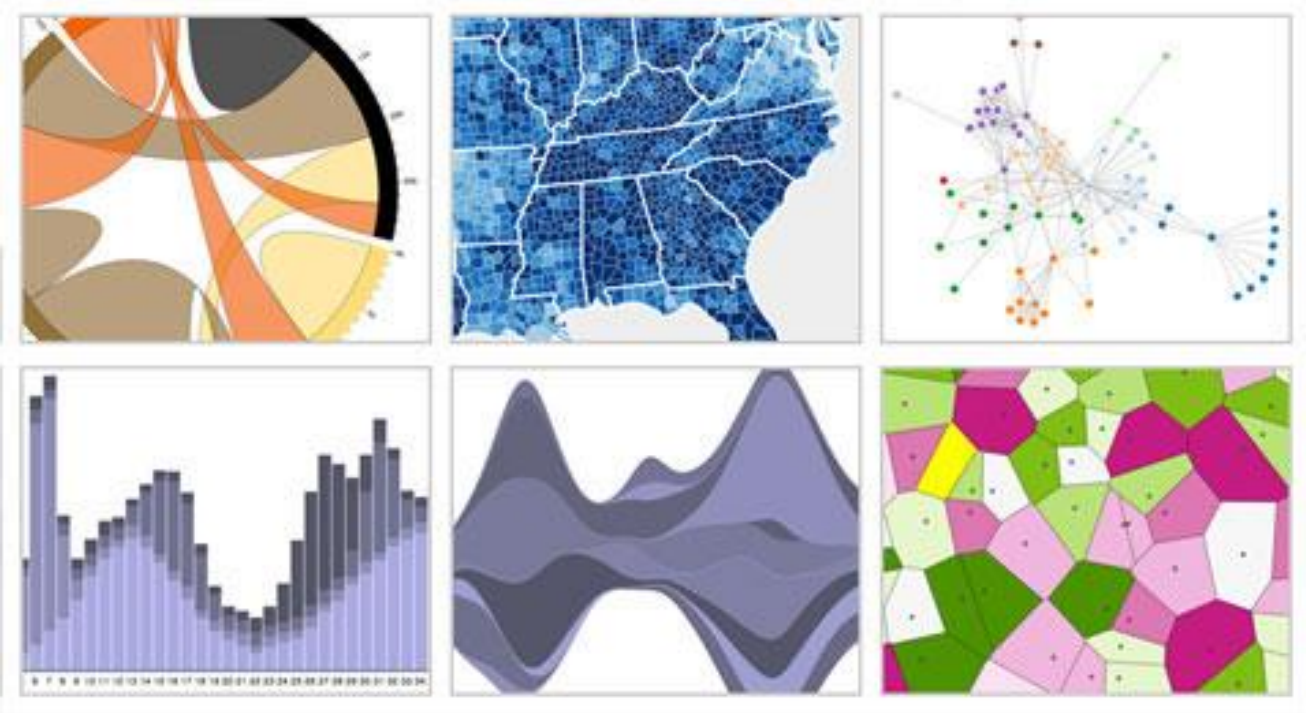


图8-8 D3提供的可视化图表

8.3.2 信息图表工具

4. Tableau

Tableau是桌面系统中最简单的商业智能工具软件，更适合企业和部门进行日常数据报表和数据可视化分析工作。Tableau实现了数据运算与美观的图表的完美结合，用户只要将大量数据拖放到数字“画布”上，转眼间就能创建好各种图表。

5. 大数据魔镜

大数据魔镜是一款优秀的国产数据分析软件，它丰富的数据公式和算法可以让用户真正理解探索分析数据，用户只要通过一个直观的拖放界面就可创造交互式的图表和数据挖掘模型。

8.3.3 地图工具

•地图工具在数据可视化中较为常见，它在展现数据基于空间或地理分布上有很强的表现力，可以直观地展现各分析指标的分布、区域等特征。当指标数据要表达的主题跟地域有关联时，就可以选择以地图作为大背景，从而帮助用户更加直观地了解整体的数据情况，同时也可以根据地理位置快速地定位到某一地区来查看详细数据。图8-9就是以数据地图形式呈现的2017年河南各地区GDP数据。

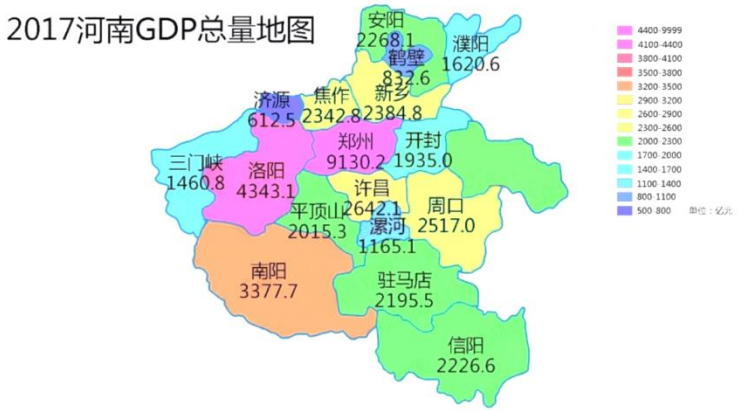


图8-9 2017年河南各地区GDP数据地图

8.3.3 地图工具

- **1. Google Fusion Tables**

Google Fusion Tables让一般使用者也可以轻松制作出专业的统计地图。该工具可以让数据表呈现为图表、图形和地图，从而帮助发现一些隐藏在数据背后的模式和趋势。

- **2. Modest Maps**

Modest Maps是一个小型、可扩展、交互式的免费库，提供了一套查看卫星地图的API，只有10KB大小，是目前最小的可用地图库，它也是一个开源项目，有强大的社区支持，是在网站中整合地图应用的理想选择。

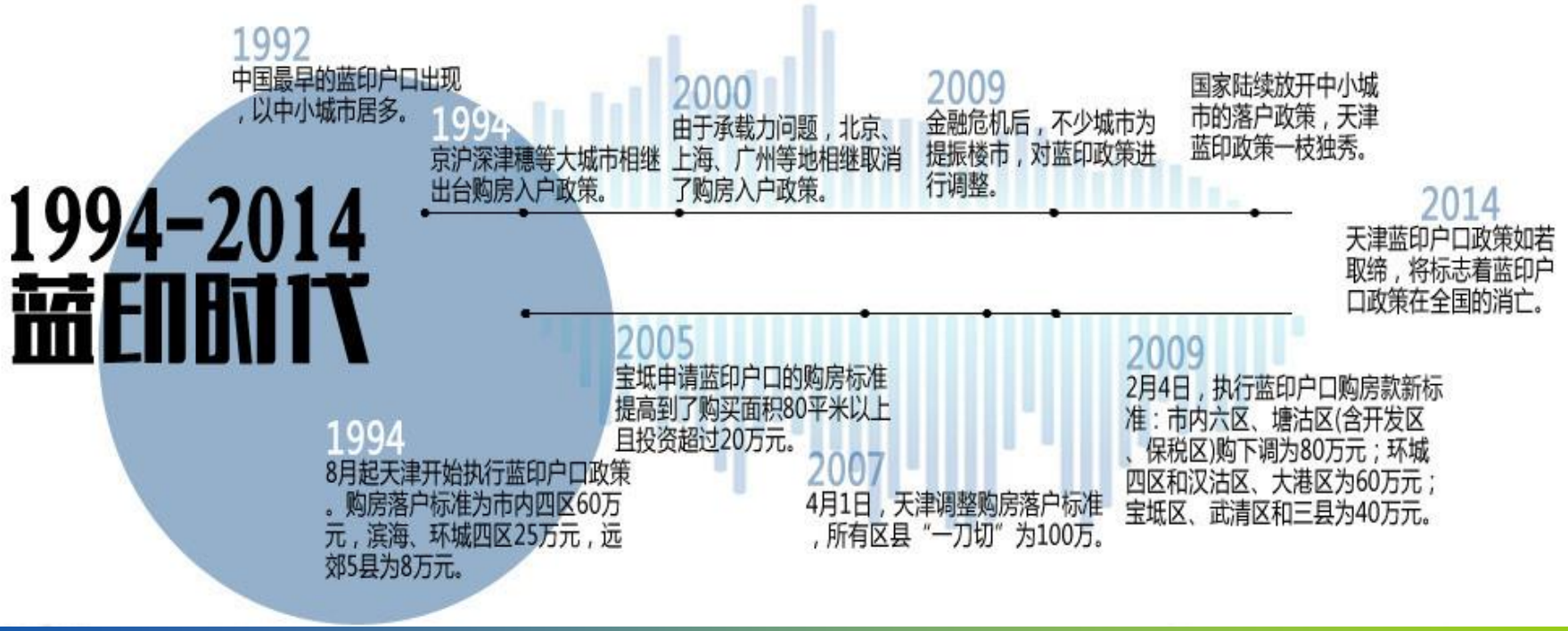
- **3. Leaflet**

Leaflet是一个小型化的地图框架，通过小型化和轻量化来满足移动网页的需要。

8.3.4 时间线工具

时间线是表现数据在时间维度的演变的有效方式，它通过互联网技术，依据时间顺序，把一方面或多方面的事件串联起来，形成相对完整的记录体系，再运用图文的形式呈现给用户。时间线可以运用于不同领域，最大的作用就是把过去的事物系统化、完整化、精确化。自2012年Facebook在F8大会上发布了以时间线格式组织内容的功能后，时间线工具在国内外社交网站中开始大面积流行。

图10-10显示了我国户籍制度在1994年到2014年间随时间的演变情况，



8.3.4 时间线工具

- **1. Timetoast**

Timetoast是在线创作基于时间轴事件记载服务的网站，提供个性化的时间线服务，可以用不同的时间线来记录你某个方面的发展历程、心理路程、进度过程等等。Timetoast基于 flash 平台，可以在类似flash时间轴上任意加入事件，定义每个事件的时间、名称、图像、描述，最终在时间轴上显示事件在时间序列上的发展，事件显示和切换十分流畅，随着鼠标点击可显示相关事件，操作简单。

- **2. Xtimeline**

Xtimeline 是一个免费的绘制时间线的在线工具网站，操作简便，用户通过添加事件日志的形式构建时间表，同时也可给日志配上相应的图表。不同于Timetoast的是，Xtimeline是一个社区类型的时间轴网站，其中加入了组群功能和更多的社会化因素，除了可以分享和评论时间轴外，还可以建立组群讨论所制作的时间轴。

8.3.5 高级分析工具

- **1. R**

R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具，使用难度较高。R的功能包括数据存储和处理系统、数组运算工具（具有强大的向量、矩阵运算功能）、完整连贯的统计分析工具、优秀的统计制图功能、简便而强大的编程语言，可操纵数据的输入和输出，实现分支、循环以及用户可自定义功能等，通常用于大数据集的统计与分析。

- **2. Weka**

Weka是一款免费的、基于Java环境的、开源的机器学习以及数据挖掘软件，不但可以进行数据分析，还可以生成一些简单图表。

- **3. Gephi**

Gephi是一款比较特殊也很复杂的软件，主要用于社交图谱数据可视化分析，可以生成非常酷炫的可视化图形。

8.3.5 高级分析工具

4.Python

Python是一种面向对象的解释型计算机程序设计语言，由荷兰人吉多•范罗苏姆（Guido van Rossum）于1989年发明。Python是纯粹的自由软件，源代码和解释器CPython遵循GPL（GNU General Public License）协议。Python具有丰富和强大的库。它常被称为“胶水语言”，能够把用其他语言制作的各种模块（尤其是C/C++）很轻松地连接在一起。Python也是一种很好的可视化工具，可以开发出各种可视化效果图，Python可视化库可以大致分为：基于matplotlib的可视化库、基于JavaScript的可视化库、基于上述两者或其他组合功能的库。

8.4 可视化典型案例

- 8.4.1 全球黑客活动
- 8.4.2 互联网地图
- 8.4.3 编程语言之间的影响力关系图
- 8.4.4 世界国家健康与财富之间的关系
- 8.4.5 3D可视化互联网地图APP

8.4.1全球黑客活动

安全供应商Norse打造了一张能够反映全球范围内黑客攻击频率的地图（http://map.ipviking.com），它利用Norse 的“蜜罐”攻击陷阱显示出所有实时渗透攻击活动。如图10-11所示，地图中的每一条线代表的都是一次攻击活动，借此可以了解每一天、每一分钟甚至每一秒世界上发生了多少次恶意渗透。

图8-11 一张能够反映全球范围内黑客攻击频率的地图

8.4.2 互联网地图

为了探究互联网这个庞大的宇宙，俄罗斯工程师 Ruslan Enikeev 根据 2011 年底的数据，将全球 196 个国家的 35 万个网站数据整合起来，并根据 200 多万个网站链接将这些“星球”通过关系链联系起来，每一个“星球”的大小根据其网站流量来决定，而“星球”之间的距离远近则根据链接出现的频率、强度和用户跳转时创建的链接来确定，由此绘制得到了“互联网地图”（http://internet-map.net），如图10-12所示。

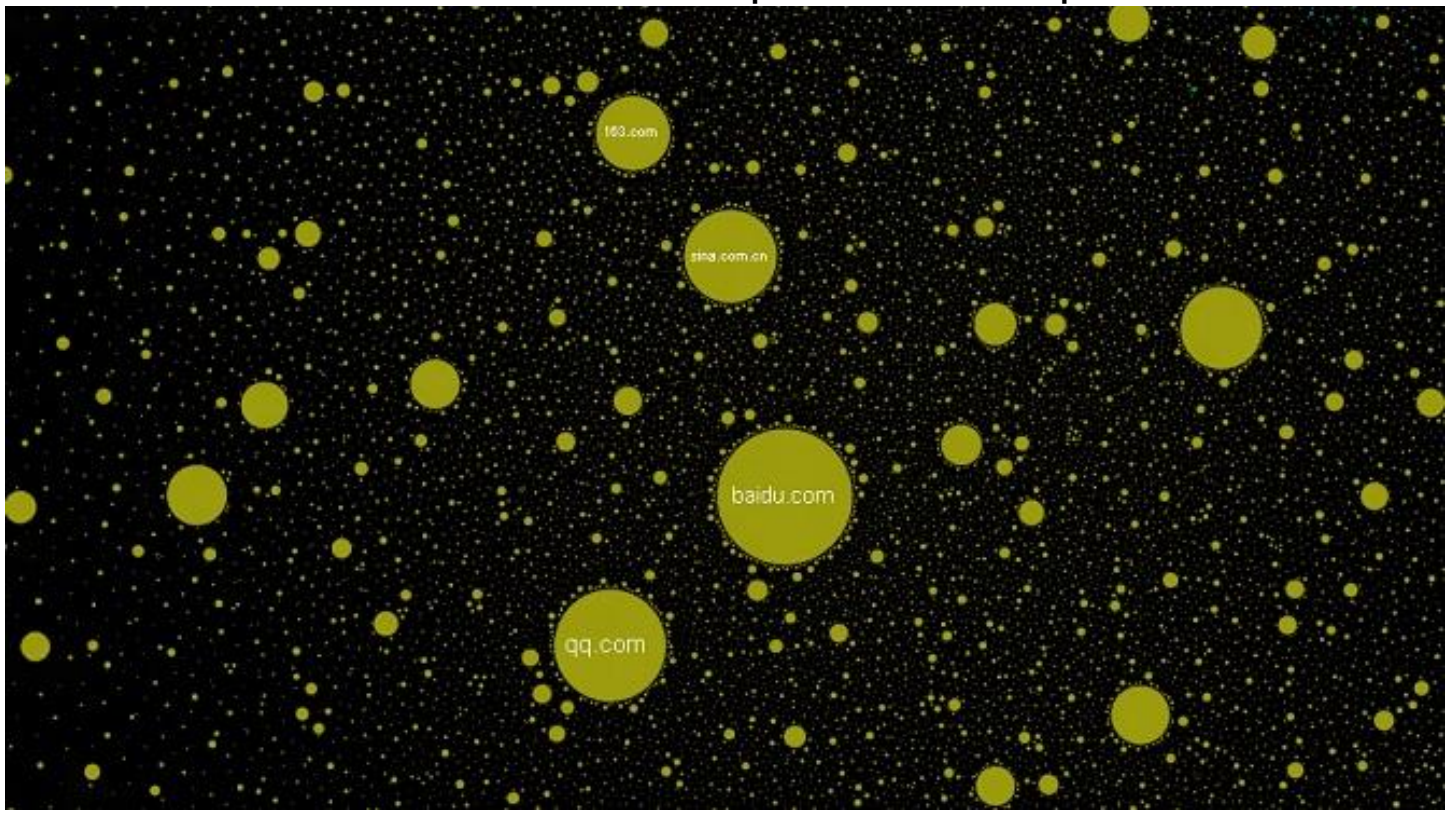


图8-12　俄罗斯工程师绘制的“互联网地图”

8.4.3 编程语言之间的影响力关系图

Ramio Gómez利用来自Freebase上的编程语言维护表里的数据，绘制了编程语言之间的影响力关系图，如图10-13所示，图中的每个节点代表一种编程语言，之间的连线代表该编程语言对其他语言有影响，有影响力的语言会连线多个语言，相应的节点也会越大。

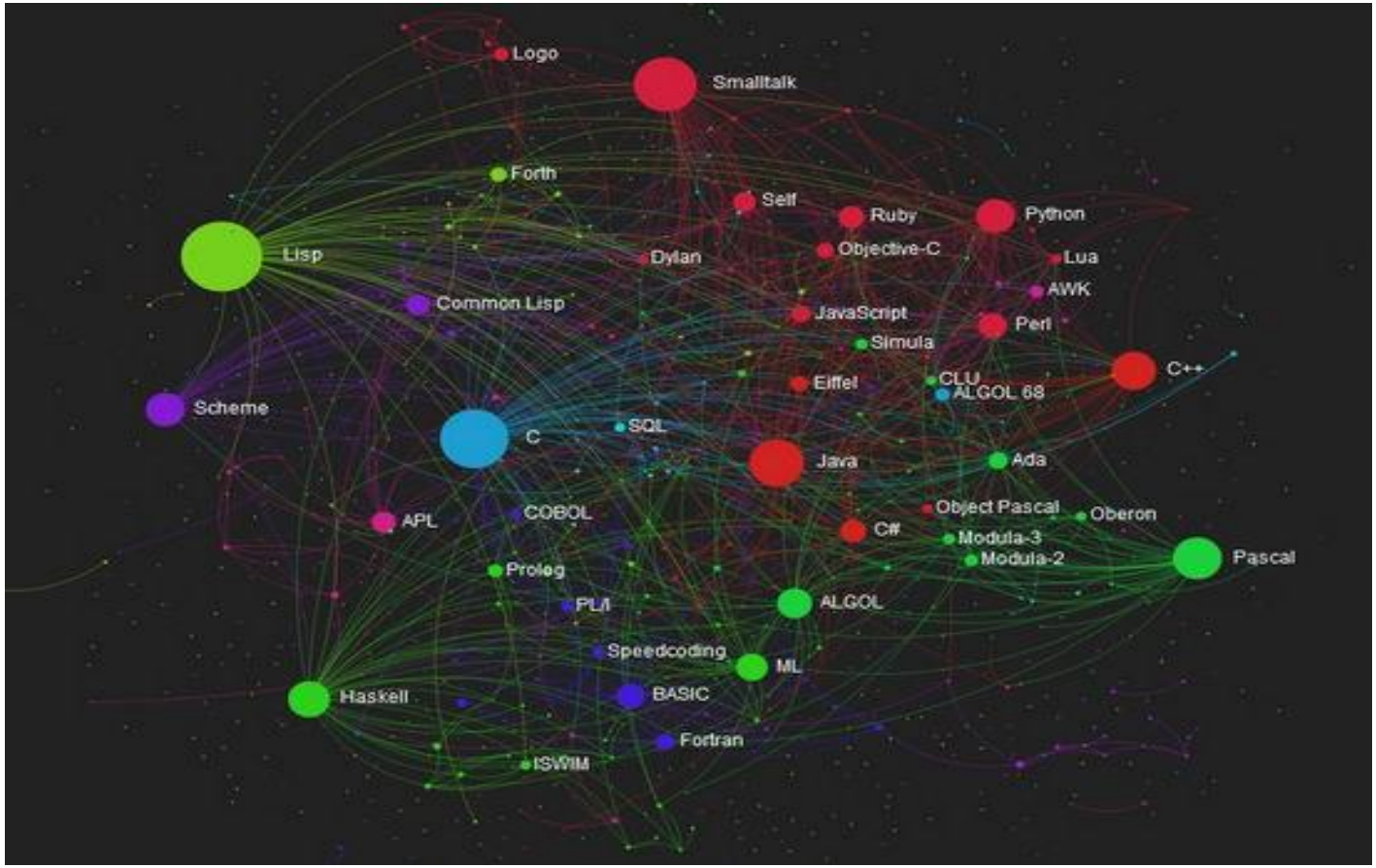


图8-13　编程语言之间的影响力关系图

8.4.4 世界国家健康与财富之间的关系

如图10-15所示，“世界国家健康与财富之间的关系”利用可视化技术，把世界上200个国家，从1810年到2010年历时200年其各国国民的健康、财富变化数据（收集了1千多万个数据）制作成三维动画进行了直观展示（http://www.moojnn.com/Index/whn）。

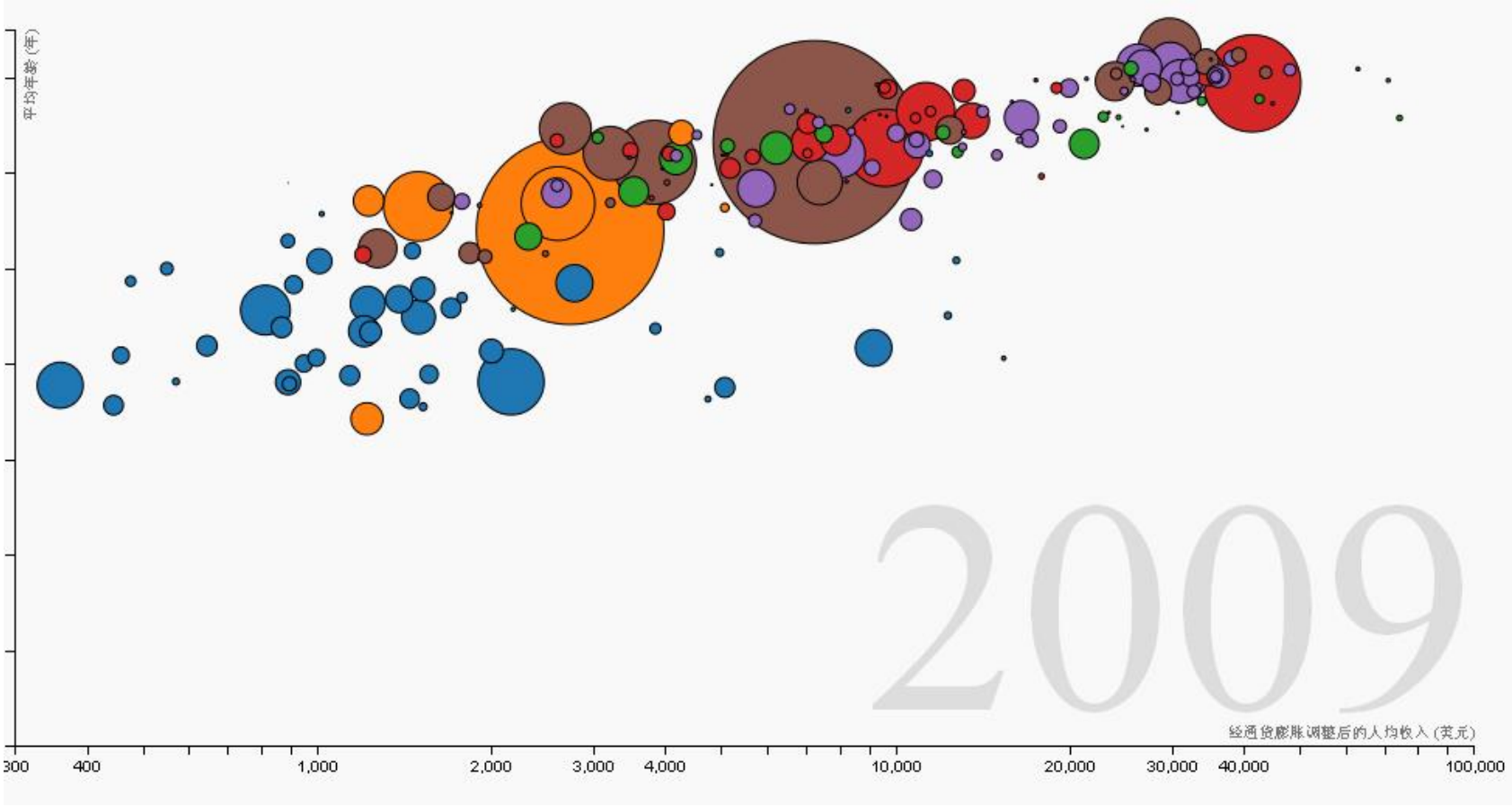


图8-14　世界国家健康与财富之间的关系图

8.4.5 3D可视化互联网地图APP

3D可视化是描绘和理解数据的一种手段，是数据的一种表征形式，并非模拟技术。3D可视化以一种独特的立体视角为用户呈现数据，可以帮助用户发现一些在2D模式下无法察觉的内容。Peer 1开发了一个称为“互联网地图”的APP（如图10-16所示），这是一个建立在小盒子形式上的3D地图。

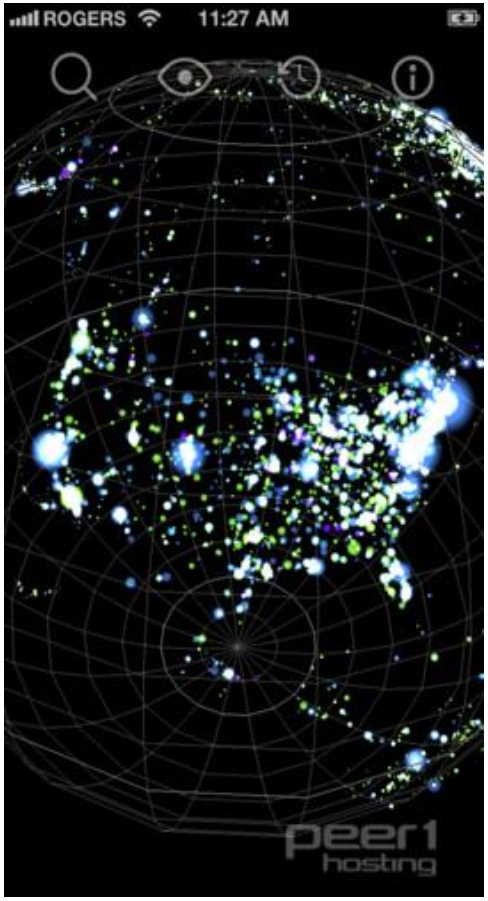


图8-15　Peer 1开发的“互联网地图”

8.5 本章小结

本章介绍了数据可视化的相关知识。数据可视化在大数据分析中具有非常重要的作用，尤其从用户角度而言，它是提升用户数据分析效率的有效手段。

统计图表是可视化图形的常见形式，常见的统计图表包括柱状图、折线图、饼图、散点图、气泡图、雷达图等。

可视化工具包括入门级工具、信息图表工具、地图工具、时间线工具和高级分析工具，每种工具都可以帮助我们实现不同类型的数据可视化分析，可以根据具体应用场合来选择适合的工具。

本章最后介绍了一些典型的数据可视化案例，从中可以深刻感受到数据可视化的魅力和重要作用。

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过200万次，累计访问量超过1000万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元

教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据技术原理与应用》教材

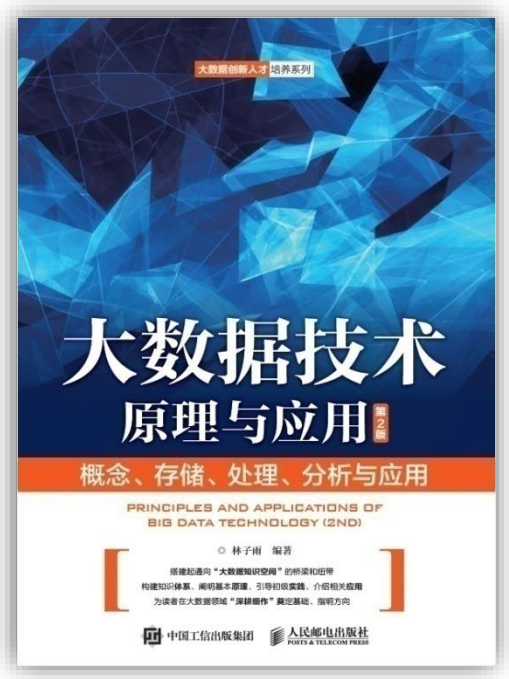
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社　ISBN:978-7-115-44330-4　定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录F：《大数据基础编程、实验和案例教程》

本书是与《大数据技术原理与应用（第2版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

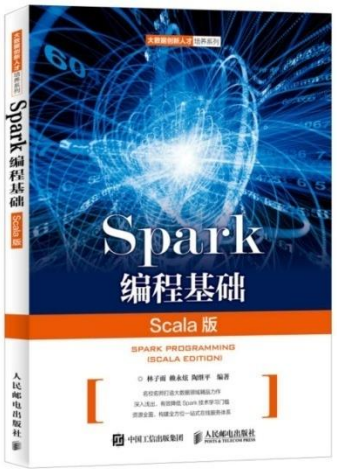
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，五套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程》
清华大学出版社 ISBN:978-7-302-47209-4 定价：59元

附录G：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

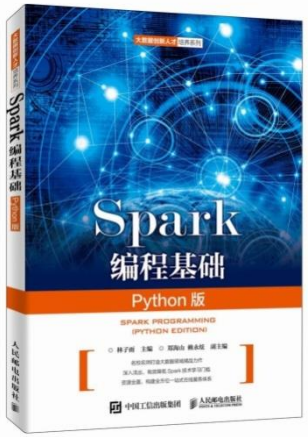
厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录H：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-52439-3
教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录I：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

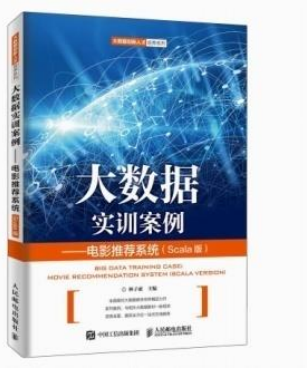
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

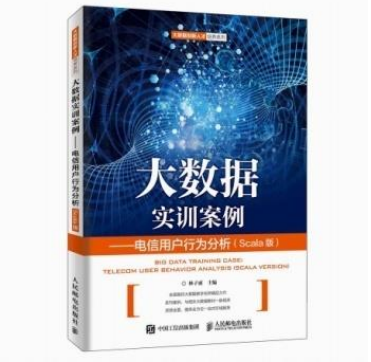
附录J：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页

Department of Computer Science, Xiamen University, 2020