数据中国“百校工程”项目系列教材
数据科学与大数据技术专业系列规划教材

瑞翼教育

大数据
可视化技术

姜枫 许桂秋 ◉ 主编
杨馥如 潘巧智 王大伟 李丛 徐曼 ◉ 副主编

BIG DATA
Technology

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

数据中国“百校工程”项目系列教材
数据科学与大数据技术专业系列规划教材

瑞翼教育

大数据可视化技术

姜枫 许桂秋 ◉ 主编
杨馥如 潘巧智 王大伟 李丛 徐曼 ◉ 副主编

BIG DATA
Technology

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

大数据可视化技术 / 姜枫，许桂秋主编. -- 北京 : 人民邮电出版社，2019.4
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-50349-7

Ⅰ. ①大… Ⅱ. ①姜… ②许… 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030240号

内 容 提 要

本书是一本系统介绍大数据可视化技术的教材。本书首先阐述了大数据可视化技术的基本概念以及相关的基础理论知识；然后，采用理论与实践相结合的方式，针对实际应用中的各种不同类型的数据，介绍相应的可视化理论和操作方法；最后，介绍了数据可视化在各个领域中的应用。

本书实例丰富，图文并茂，叙述简明，重点突出。本书可以作为高等院校计算机、数据科学与大数据技术等相关专业的教材，也可作为从事数据可视化、数据分析的相关技术人员的参考书。

◆ 主　　编　姜　枫　许桂秋
副 主 编　杨馥如　潘巧智　王大伟　李　丛　徐　曼
责任编辑　邹文波
责任印制　陈　犇

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：10　　2019 年 4 月第 1 版
字数：226 千字　　2019 年 4 月河北第 1 次印刷

定价：39.80 元

读者服务热线：(010) 81055256　印装质量热线：(010) 81055316
反盗版热线：(010) 81055315
广告经营许可证：京东工商广登字 20170147 号

前言

数据反映着真实的世界，人们希望用数据寻求真相，分析数据间的关联，了解这个世界正在发生什么，从而解决真实世界的各种问题，比如，降低犯罪率，提高公共卫生意识，改善交通状况，或者增长个人的见识，等等。数据存在于我们生活的每一个角落，等待着我们去有效地利用。人们都希望挖掘出数据背后蕴藏的信息。可视化技术正是探索和理解大数据的最有效的途径之一。将数据转化为视觉图像，能帮助我们更加容易地发现和理解其中隐藏的模式或规律。

计算机呈现数据的方式多种多样，充满艺术设计感的视觉图像遍布我们的生活。这些美观的图像包含了要表达的数据信息，而且十分引人注目。因此，人们总结出数据可视化的一些基本理论和方法，通过一些操作简单的软件或编程语言，就能高效且直观地获得数据中蕴含的信息。

本书是大数据可视化技术的入门教材，采用理论与实践相结合的方式，循序渐进地介绍了大数据可视化技术中的各种理论知识，结合各种实践案例，引导读者运用所学知识解决现实中的问题。

全书共 9 章，可分为三个部分。

第一部分是基础理论，包括第 1 章和第 2 章。第 1 章阐述了数据可视化的定义、作用和发展历史，以及数据可视化所面临的挑战和未来的发展方向；第 2 章详细介绍了数据可视化的基础知识，包括视觉感知与认知的基本原理和可视化编码原则，数据可视化的基本框架、基本原则、基本图表以及相关工具。

第二部分是数据分析，包括第 3～8 章。这 6 章详细介绍了时间数据、比例数据、关系数据、文本数据、复杂数据以及用户交互的各种可视化理论和方法。其中，第 3 章介绍时间数据可视化，时间数据又分为连续型时间数据和离散型时间数据，基础图形包括阶梯图、折线图、散点图和柱形图等；第 4 章介绍比例数据可视化，比例数据的基础图形包括饼图、环形图、堆叠图等；第 5 章介绍关系数据可视化，关系数据的基础图形包括散点图、气泡图、直方图、密度图等；第 6 章介绍文本数据可视化，文本数据可视化需要对文本中的内容进行提取、分析，再使用各种可视化方法进行展示；第 7 章介绍复杂数据可视化，复杂数据包括高维多元数据、非结构化数据以及不确定性数据；第 8 章详细介绍了数据可视化交互的原则、分类和技术，通过交互技术能让用户更好地理解和分析数据。

第三部分是实际应用，包括第 9 章。第 9 章详细介绍了数据可视化在科研领域、网络领域以及商业领域的各种应用。

本书可以作为高等院校计算机、数据科学与大数据技术等相关专业的数据可视化教材。使用本书作为教材，建议安排 32 课时，教师可根据学生的接受能力以及高校的培养方案选择教学内容。

由于编者水平有限，编写时间仓促，书中难免存在一些疏漏和不足之处，恳请广大读者批评指正。

编者

2019 年 1 月

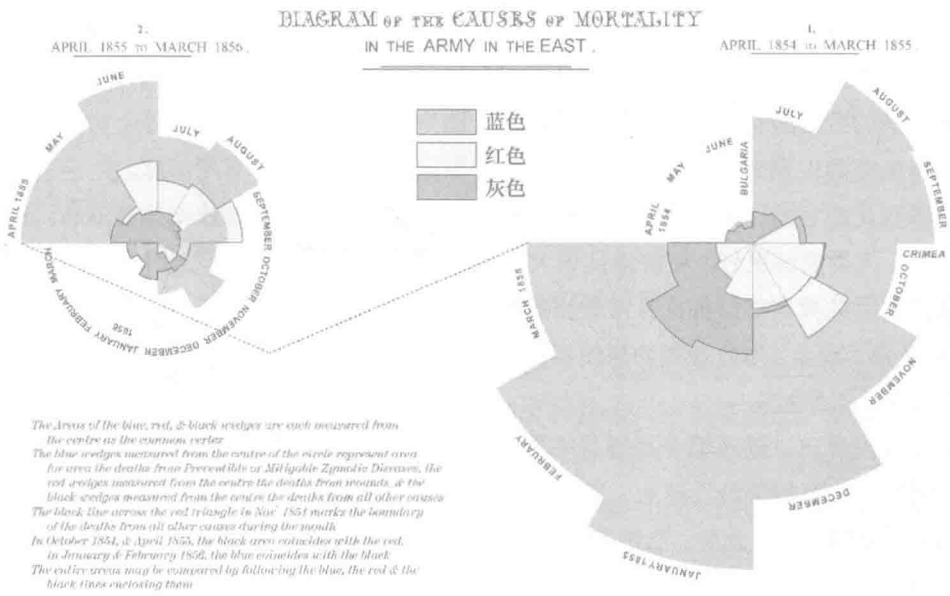
目　录

第一部分　基础理论

第二部分　数据分析

第三部分　实际应用

第一部分
基础理论



第1章 数据可视化概述

本章首先介绍数据可视化的概念和作用，然后介绍数据可视化的发展历史，以及数据可视化的未来。

1.1 什么是数据可视化

人类对图形、图像等可视化符号的处理效率要比对数字、文本的处理效率高很多。有研究表明，绝大部分的视觉信号处理过程通常发生在人脑的潜意识阶段，例如，人们在观看包含自己的集体照时，通常潜意识会第一时间寻找照片中的自己，然后才会寻找其他感兴趣的目标。

可视化对应的英文词汇有 Visualize 和 Visualization。Visualize 是动词，表示生成可视化图像，利用可视化方式传递信息；Visualization 是名词，表示可视化过程，对某个原本不易描述的事物形成一个可感知的画面的过程。在计算机视觉领域，数据可视化是对数据的一种形象直观的解释，实现从不同维度观察数据，从而得到更有价值的信息。数据可视化将抽象的、复杂的、不易理解的数据转化为人眼可识别的图形、图像、符号、颜色、纹理等，这些转化后的数据通常具备较高的识别效率，能够有效地传达出数据本身所包含的有用信息。

数据可视化的目的，是对数据进行可视化处理，以更明确地、有效地传递信息。比起枯燥乏味的数值，人类能够更好、更快地认识大小、位置、形状、颜色深浅等物体的外在直观表现。经过可视化之后的数据能够加深人对数据的理解和记忆。例如，对于这样一个问题：如果额外给你 10 000 美元现金，你会选择如何使用它？美国投资机构针对三个年龄段的本土公民做出的调研结果如图 1-1 所示。

从图中可以看出，偿还债务是得票率最高的选项。这显然与美国发达的信贷市场和消费结构有关。其中，公民的年龄段越大，还款意愿就越强。除了还款，55 岁以上的美国人还比较倾向于低风险的理财项目，比如，选择高息储蓄或购买债券，或者把钱直接存入退休金账户。35～54 岁的美国人中，绝大多数会选择把这笔钱投资在子女教育上，可见，教育支出也是近二十年来美国人增长最快的财务支出。18～34 岁的美国人既有兴趣加大对自身的教育投入，也愿意尝试高风险的投资产品。我们还可以看到，不动产也是较受美国人欢迎的投资项目之一，其中年

轻人的买房欲望相对而言是最高的。

由此可见，将数据经过图形化展示以后，人们可以从可视化的图形中直观地获取更有效的信息。

图 1-1　不同年龄的人如何进行投资

近年来，随着大数据时代的到来，面对越来越庞大、复杂的数据，数据可视化已经成为各个领域传递信息的重要手段。数据可视化也可以将其理解为一个生成图形、图像符号的过程。更为深层次的理解是，可视化是人类思维认知强化的过程，即人脑通过人眼观察某个具体图形、图像来感知某个抽象事物，这个过程是一个强化认知的理解过程。因此，帮助人们理解事物规律是数据可视化的最终目标，而绘制的可视化结果只是可视化的过程表现。

数据可视化是为了从数据中寻找三个方面的信息：模式、关系和异常。

（1）模式。指数据中的规律。比如，机场每月的旅客人数都不一样，通过几年的数据对比，发现旅客人数存在周期性的变化，某些月份的旅客数量一直偏低，某些月份的旅客数量

则一直偏高。

图 1-2 是著名的南丁格尔玫瑰图，蓝色区域表示死于感染的士兵数量，红色区域表示死于战场重伤的士兵数量，灰色区域表示死于其他原因的士兵数量。该图有如下两个非常明显的特征。

① 两幅图中蓝色区域的面积明显大于其他颜色的面积。

这意味着大多数的伤亡并非直接来自战争，而是来自糟糕医疗环境下的感染。

② 左边这幅中的扇形面积远小于右边这幅图。

说明卫生委员到达后（1855 年 3 月），死亡人数明显下降，成功地展示了医疗卫生条件的改善带来的效果。

这幅图出现在南丁格尔游说英国政府加强公众医疗卫生建设和相关投入的文件里。这幅图让政府官员了解到：改善医院的医疗状况可以显著地降低英军的死亡率。南丁格尔的玫瑰图打动了当时的政府高层（包括军方人士和维多利亚女王），她的医疗改良的提案才得以通过，从而挽救了千万人的生命。

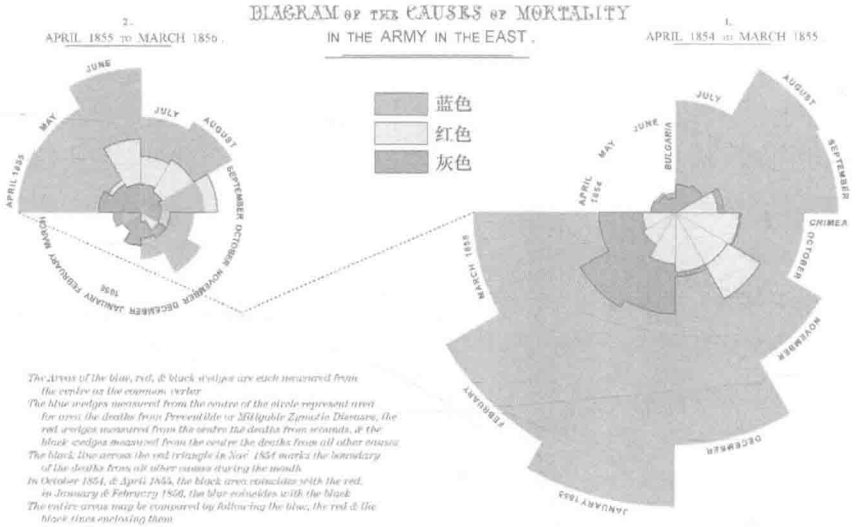


图 1-2 南丁格尔玫瑰图

（2）关系。指数据之间的相关性，在统计学中，通常代表关联性和因果关系。无论数据的总量和复杂程度如何大，数据间的关系大多可分为三类：数据间的比较、数据的构成，以及数据的分布或联系。比如，收入水平与幸福感之间的关系是否成正比，经统计，对于月收入在 1 万元以下的人来说，一旦收入增加，幸福感会随之提升，但对于月收入水平在 1 万元以上的人来说，幸福感并不会随着收入水平的提高而提升，这种非线性关系也是一种关系。图 1-3 展示了基本图表与数据间的关系。

（3）异常。指有问题的数据。异常的数据不一定都是错误的数据，有些异常数据可能是设备出错或者人为错误输入，有些可能就是正确的数据。通过异常分析，用户可以及时发现各种异常情况。如图 1-4 所示，图中大部分点都集中在一个区域，极少数点分散在其他区域，这些

都属于异常值，需要特殊处理。

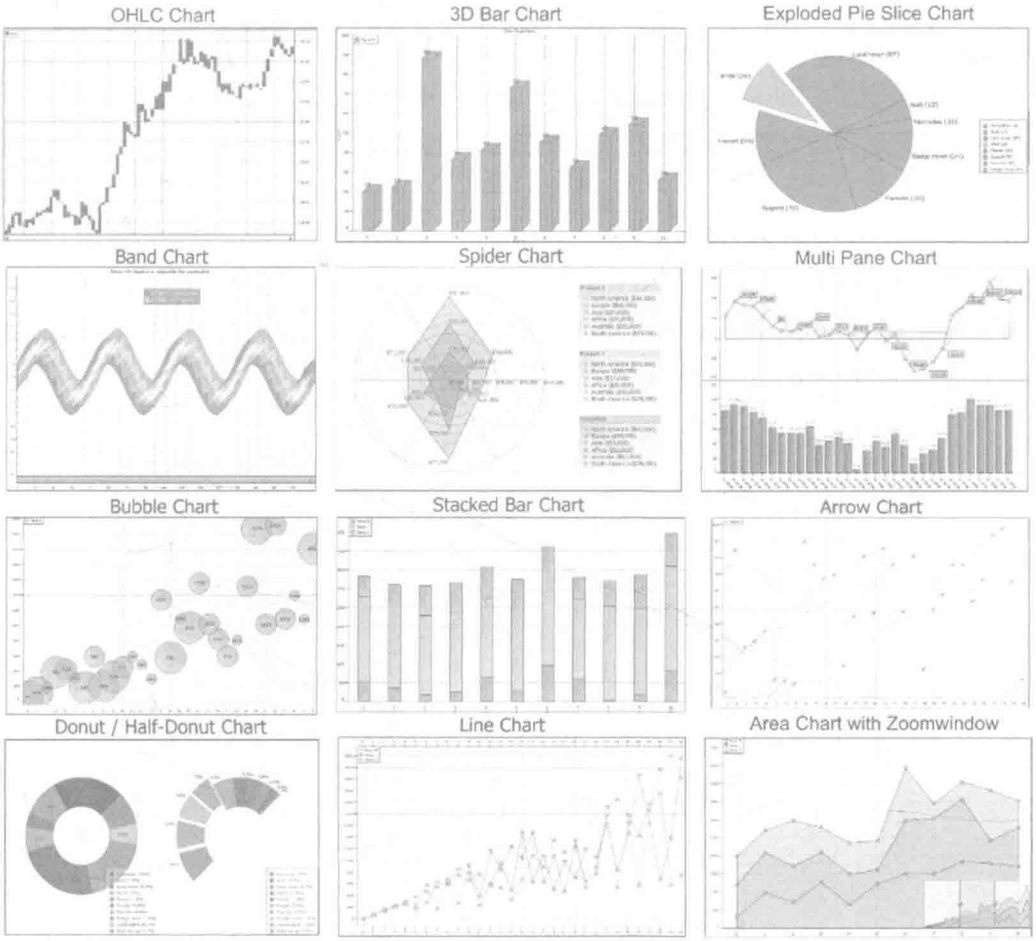


图 1-3　基本图表展示数据间的关系

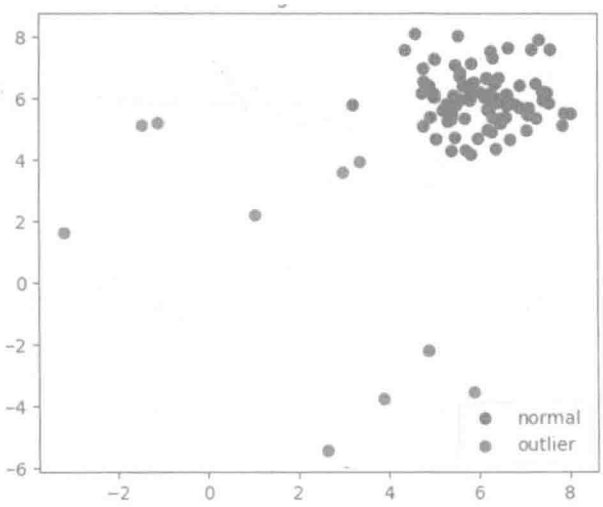


图 1-4　异常值

1.2 数据可视化的作用

数据可视化的作用包括记录信息、分析推理、信息传播与协同等。

（1）记录信息

自古以来，记录信息的有效方式之一是用图形的方式描述各种具体或抽象的事物。如图 1-5 所示，左图是列奥纳多·达芬奇（Leonardo da Vinci）绘制的人体解剖图，中图是自然史博物学家威廉·柯蒂斯（William Curtis）绘制的植物图，右图是 1616 年伽利略关于月亮周期的绘图，记录了月亮在一定时间内的变化。

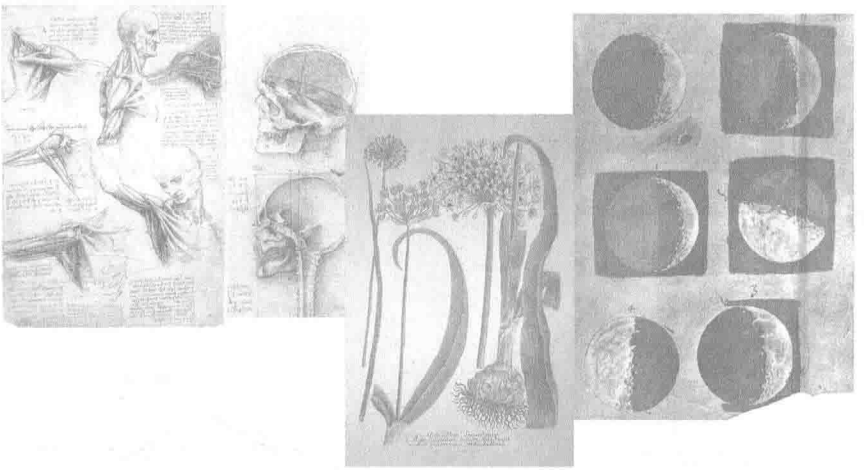


图 1-5 数据可视化的作用之一——记录信息

如图 1-6 所示，田径赛场上的裁判员通过这幅图可以清晰、准确、迅速地判定运动员的名次和成绩。

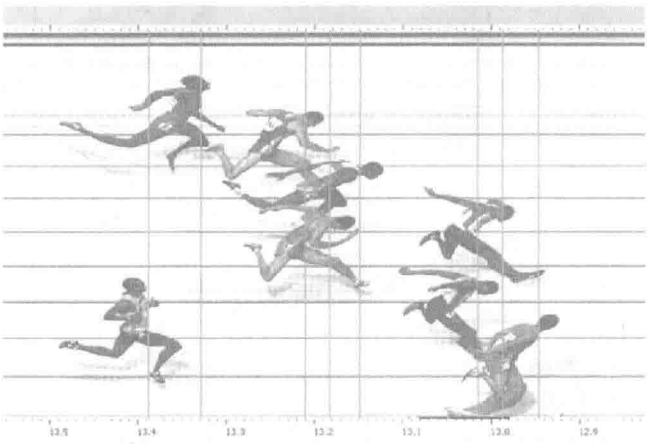


图 1-6 田径赛运动员冲刺图

（2）分析推理

数据可视化极大地降低了数据理解的复杂度，有效地提升了信息认知的效率，从而有助于人们更快地分析和推理出有效信息。1854 年，伦敦爆发了一场霍乱，英国医生 John Snow 绘制了一张街区地图，如图 1-7 所示，这就是著名的“伦敦鬼图”。该图分析了霍乱患者的分布与水井分布之间的关系，发现在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——一个被污染的水泵。

图 1-7　伦敦鬼图

（3）信息传播与协同

俗话说“百闻不如一见”“一图胜千言”。

图 1-8 是介绍中国烟民数量的图形，如果只看左图，可知中国烟民的数量是 320 000 000，这个数据是很大，但具体有多大读者却不能直接感知。结合右图可知，中国烟民数量超过了美国人口总和，通过这种对比，对数据的感知就加深了。

图 1-8　中国烟民的数量

图 1-9 是介绍雅虎邮箱处理数据量的图形，大意是雅虎邮箱每小时处理的电子邮件总量的大小是 1.2TB，这些邮件若打印出来，大约需要 644 245 094 张 A4 打印纸。这也是一个很大的

数据，但到底有多大？在这里用了一个比喻的手法：644 245 094 张纸，如果把每一张纸首尾对接，可以绕地球 4 圈多。由此，读者就能深刻地感受到雅虎邮箱处理的数据量之大。

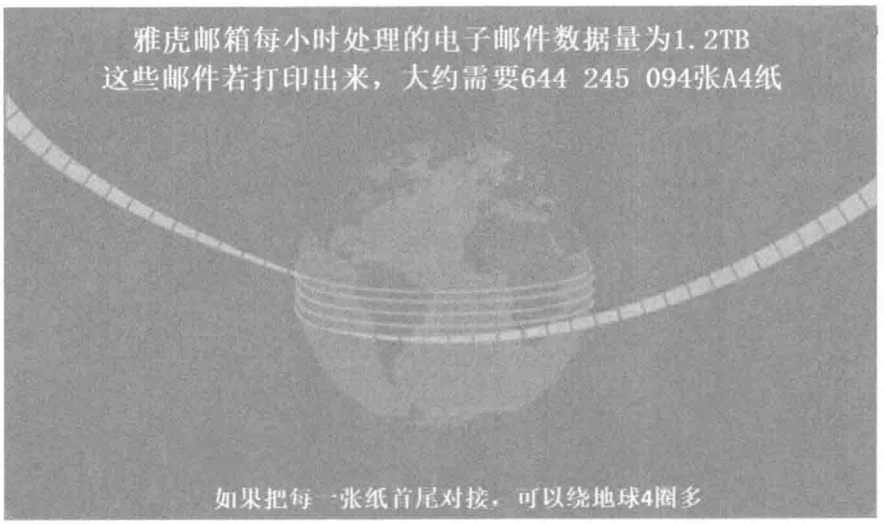


图 1-9　雅虎邮箱处理数据量

随着计算机技术的普及，数据无论从数量上还是从维度层次上都变得日益繁杂。面对海量而又复杂的数据，各个科研机构和商业组织普遍遇到以下问题。

（1）大量数据不能有效利用，弃之可惜，想用却不知如何下手。

（2）数据展示模式繁杂晦涩，无法快速甄别有效信息。

数据可视化就是将海量数据经过抽取、加工、提炼，通过可视化方式展示出来，改变传统的文字描述识别模式，达到更高效地掌握重要信息和了解重要细节的目的。

数据可视化在大数据分析中的作用主要体现在以下几个方面。

（1）动作更快。使用图表来总结复杂的数据，可以确保对关系的理解要比那些混乱的报告或电子表格更快。可视化提供了一种非常清晰的交互方式，从而能够使用户更快地理解和处理这些信息。

（2）以建设性方式提供结果。大数据可视化工具能够用一些简短的图形描述复杂的信息。通过可交互的图表界面，轻松地理解各种不同类型的数据。例如，许多企业通过收集消费者行为数据，再使用大数据可视化来监控关键指标，从而更容易发现各种市场变化和趋势。例如，一家服装企业发现，在西南地区，深色西装和领带的销量正在上升，这促使该企业在全国范围内推销这两类产品。通过这种策略，这家企业的产品销量远远领先于那些尚未注意到这一潮流的竞争对手。

（3）理解数据之间的联系。在市场竞争环境中，找到业务和市场之间的相关性是至关重要的。例如，一家软件公司的销售总监在条形图中看到，他们的旗舰产品在西南地区的销售额下降了 8%，销售总监可以深入了解问题出现在哪里，并着手制订改进计划。通过这种方式，数据可视化可以让管理人员立即发现问题并采取行动。

1.3　数据可视化的分类

数据可视化的处理对象是数据。根据所处理的数据对象的不同，数据可视化可分为科学可视化与信息可视化。科学可视化面向科学和工程领域数据，如三维空间测量数据、计算模拟数据和医学影像数据等，重点探索如何以几何、拓扑和形状特征来呈现数据中蕴含的规律；信息可视化的处理对象则是非结构化的数据，如金融交易、社交网络和文本数据，其核心挑战是如何从大规模高维复杂数据中提取出有用信息。

由于数据分析的重要性，将可视化与数据分析结合，可形成一个新的学科：可视分析学。

1. 科学可视化

科学可视化是可视化领域发展最早、最成熟的一个学科，其应用领域包括物理、化学、气象气候、航空航天、医学、生物学等各个学科，涉及对这些学科中数据和模型的解释、操作与处理，旨在寻找其中的模式、特点、关系以及异常情况。

科学可视化的基础理论与方法已经相对成熟，其中有一些方法已广泛应用于各个领域。最简单的科学可视化方法是颜色映射法，它将不同的值映射成不同的颜色。科学可视化方法还包括轮廓法（Contouring），轮廓法是将数值等于某一指定阈值的点连接起来的可视化方法，地图上的等高线，天气预报中的等温线都是典型的轮廓可视化的例子。

2. 信息可视化

与科学可视化相比，信息可视化的数据更贴近我们的生活与工作，包括地理信息可视化、时变数据可视化、层次数据可视化、网络数据可视化、非结构化数据可视化等。

我们常见的地图是地理信息数据，属于信息可视化的范畴。

时变数据可视化采用多视角、数据比较等方法体现数据随时间变化的趋势和规律。

在层次数据可视化中，层次数据表达各个个体之间的层次关系。树图是层次数据可视化的典型案例，树图是对现实世界事物关系的抽象，其数据本身具有层次结构的信息。

在网络结构数据可视化中，网络数据不具备层次结构，关系更加复杂和自由，如人与人之间的关系、城市道路连接、科研论文的引用等。

非结构化数据可视化通常是将非结构化数据转化为结构化数据再进行可视化显示。

3. 可视分析学

可视分析学被定义为一门以可视交互界面为基础的分析推理科学，综合了图形学、数据挖掘和人机交互等技术。可视分析学是一门综合性学科，与多个领域相关：在可视化领域，与信息可视化、科学可视化、计算机图形学相关；在数据分析相关的领域，与信息获取、数据处理、数据挖掘相关；在交互领域，则与人机交互、认知科学和感知等学科融合。

可视分析学所包含的研究内容非常广泛，其中，感知与认知科学研究人在可视化分析学中的重要作用；数据管理和知识表达是可视分析构建数据到知识转换的基础理论；地理分析、信息分析、科学分析、统计分析、知识发现等是可视分析学的核心分析方法；在整个可视分析过