大数据专业应用型人才培养规划教材

瑞翼教育

# 大数据可视化技术

杨尚森　许桂秋　主编

浙江科学技术出版社

瑞翼教育

## 大数据专业应用型人才培养规划教材

- 大数据导论
- Python编程基础与应用
- NoSQL数据库原理与应用
- Hadoop大数据技术与应用
- Spark大数据技术与应用
- 大数据预处理技术
- 数据挖掘与机器学习
- 大数据可视化技术
- 商务智能方法与应用

ISBN 978-7-5341-8895-4

9 787534 188954 >

定 价：58.00元

大数据专业应用型人才培养规划教材

瑞翼教育

# 大数据可视化技术

杨尚森　许桂秋　主编

浙江科学技术出版社

**图书在版编目（CIP）数据**

大数据可视化技术 / 杨尚森，许桂秋主编. —杭州：浙江科学技术出版社，2020.1

大数据专业应用型人才培养规划教材

ISBN 978-7-5341-8895-4

Ⅰ. ①大… Ⅱ. ①杨… 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(2020)第002078号

**主　　编**　杨尚森　许桂秋

**副 主 编**　李海涛　张　辉　赖联锋　胡　朋

宋军领　李　勇　陈献聪

**丛 书 名**　大数据专业应用型人才培养规划教材

**书　　名**　**大数据可视化技术**

**主　　编**　杨尚森　许桂秋

---

**出版发行**　**浙江科学技术出版社**

杭州市体育场路347号　邮政编码：310006

办公室电话：0571-85176593

销售部电话：0571-85176040

网　址：www.zkpress.com

E-mail：zkpress@zkpress.com

**排　　版**　杭州真凯文化艺术有限公司

**印　　刷**　杭州广育多莉印刷有限公司

**经　　销**　全国各地新华书店

---

| | | | | | |
|---|---|---|---|---|---|
| **开　　本** | 787×1092　1/16 | | **印　　张** | 11.5 | |
| **字　　数** | 266 000 | | | | |
| **版　　次** | 2020年1月第1版 | | **印　　次** | 2020年1月第1次印刷 | |
| **书　　号** | ISBN 978-7-5341-8895-4 | | **定　　价** | 58.00元 | |

---



**责任编辑**　罗　璀　　**策划编辑**　张祝娟　　**责任校对**　顾旻波

**责任美编**　金　晖　　**责任印务**　崔文红

# 前　言

随着计算机技术与互联网技术的快速发展，使用互联网思维解决问题的方式让人们的生活变得越来越便利，由此也积累了种类繁多、体量巨大的数据。这些数据存在于我们生活的每一个角落，反映着真实的世界，人们都希望挖掘出数据背后蕴藏的巨大价值。可视化技术是大数据分析挖掘的最直观表达，是探索和理解大数据最有效的途径之一。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，大数据正在悄然改变着人们的生活，正在开创一个新时代。

计算机图形学与硬件的进步让大数据的数据呈现更加丰富多彩。一个充满艺术设计美感的视觉图像不仅包含了要表达的信息，也让信息变得更加生动形象。因此，数据可视化的基本理论和方法显得尤为重要。

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，由浅入深地介绍了大数据可视化技术的基本概念和基础知识，并结合实践案例，带着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章介绍了时间数据可视化，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章介绍了比例数据可视化，比例数据的基础图形包括饼图、环形图、堆叠图等；第5章介绍了关系数据可视化，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章介绍了文本数据可视化，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章介绍了复杂数据可视化，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，为第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，也可作为从事数据可视化，数据分析相关工作技术人员的参考书。本课程建议安排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，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。

编著者

2019年12月

# 目　录

# 第1章 数据可视化概述

数据和文字是抽象的，图形却是具体的，正所谓“能用图就不用表，能用表就不用文字”。好的图形或分析报告，应该直观易懂又不失专业性，数据跃然纸上，分析一语中的。数据可视化，官方释义，是关于数据之视觉表现形式的研究；旨在借助于图形化手段，清晰有效地传达与沟通信息。本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。

## 本章重点

◎ 数据可视化的概念和作用。
◎ 数据可视化的发展历史。
◎ 数据可视化的未来。

## 1.1 什么是数据可视化

通俗地说，数据可视化就是“你给我一些数据，我给你一些图片”。更学术地说，就是一个把信息映射成视觉效果的过程。

最开始的数据可视化方法就是一些简单的表格，但是现在可视化的方法可以说是五花八门。大家比较熟悉的可视化的方法有条形图、折线图（图1-1）、饼图、直方图、散点图、箱线图等。

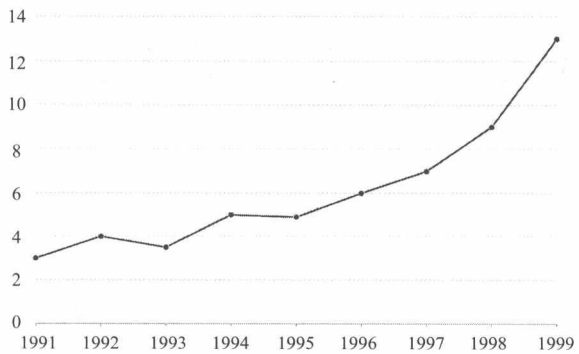



图1-1 折线图

不太熟悉的可能有平行坐标图（图1–2）、并行集图（图1–3）、力导向图、标签云图（图1–4）等。

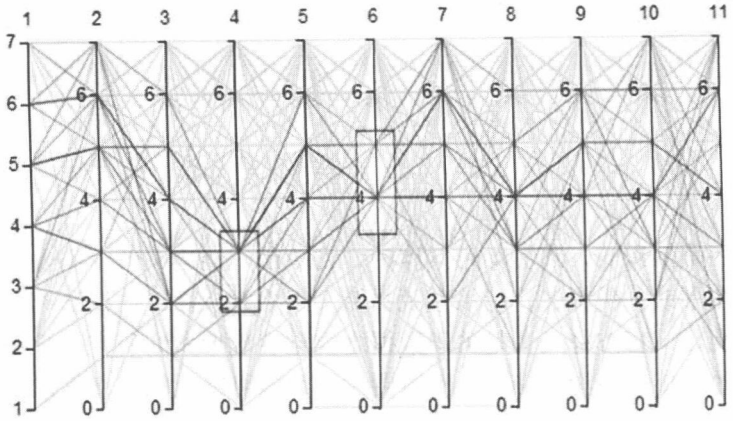



图1–2　平行坐标图

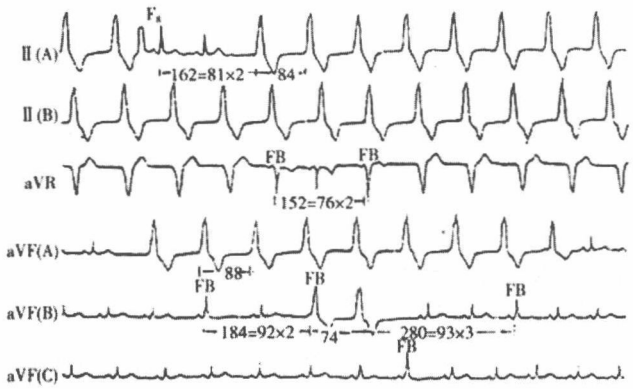



图1–3　并行集图


图1–4　标签云图

但并不是随便什么图片都可以，而是需要能够高效地对这些数据进行描述、探索、总结。它们往往会由点、线、坐标轴、符号、单词、阴影和颜色等视觉元素所构成。

大家可以观察一下图1–5中的数据。

| | 1 | | 2 | | 3 | | 4 | |
|---|---|---|---|---|---|---|---|---|
| | X | Y | X | Y | X | Y | X | Y |
| | 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| | 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| | 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| | 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| | 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| | 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| | 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| | 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| | 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| | 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| | 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |
| Mean | 9.0 | 7.5 | 9.0 | 7.5 | 9.0 | 7.5 | 9.0 | 7.5 |
| Variance | 10.0 | 3.75 | 10.0 | 3.75 | 10.0 | 3.75 | 10.0 | 3.75 |
| Correlation | 0.816 | | 0.816 | | 0.816 | | 0.816 | |

图1–5　图表数据集

我们从图1–5中很难发现数据的一些特点，但如果换作图1–6，情况就会变得不一样。

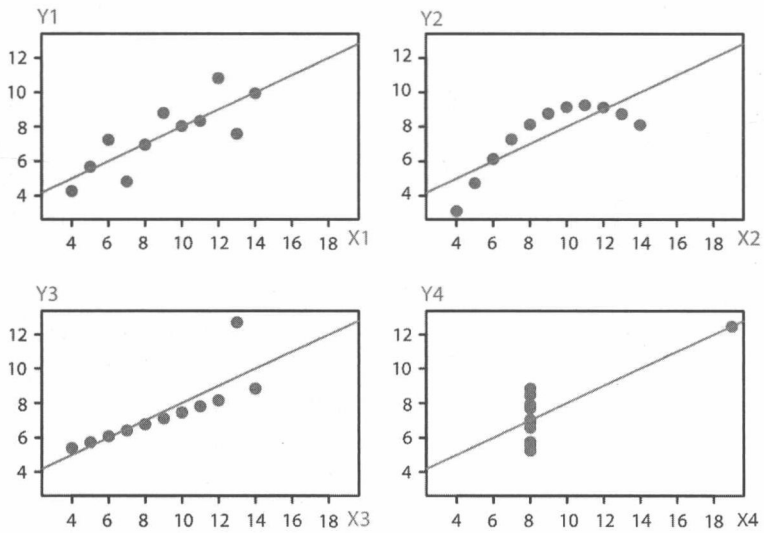



图1–6　数据可视化

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。

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1-7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。该图有以下两个非常明显的特征：

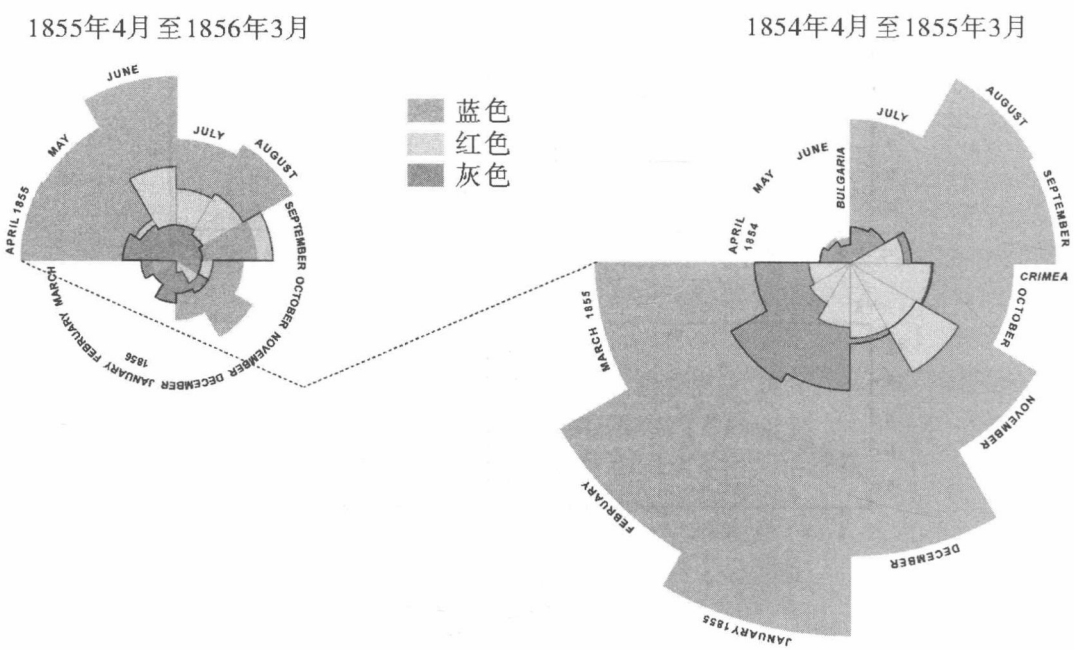
（1）两幅图中蓝色区域的面积明显大于其他颜色的面积。

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。




（a）卫生委员到达后　　（b）卫生委员到达前

图1-7　南丁格尔玫瑰图

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1-8展示了数据可以用多种维度的图表来表示。

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1-9所示，图中大部分点都集中在一个区域，极少量点分散在其他区域，这些都属于异常值，需要特殊处理。

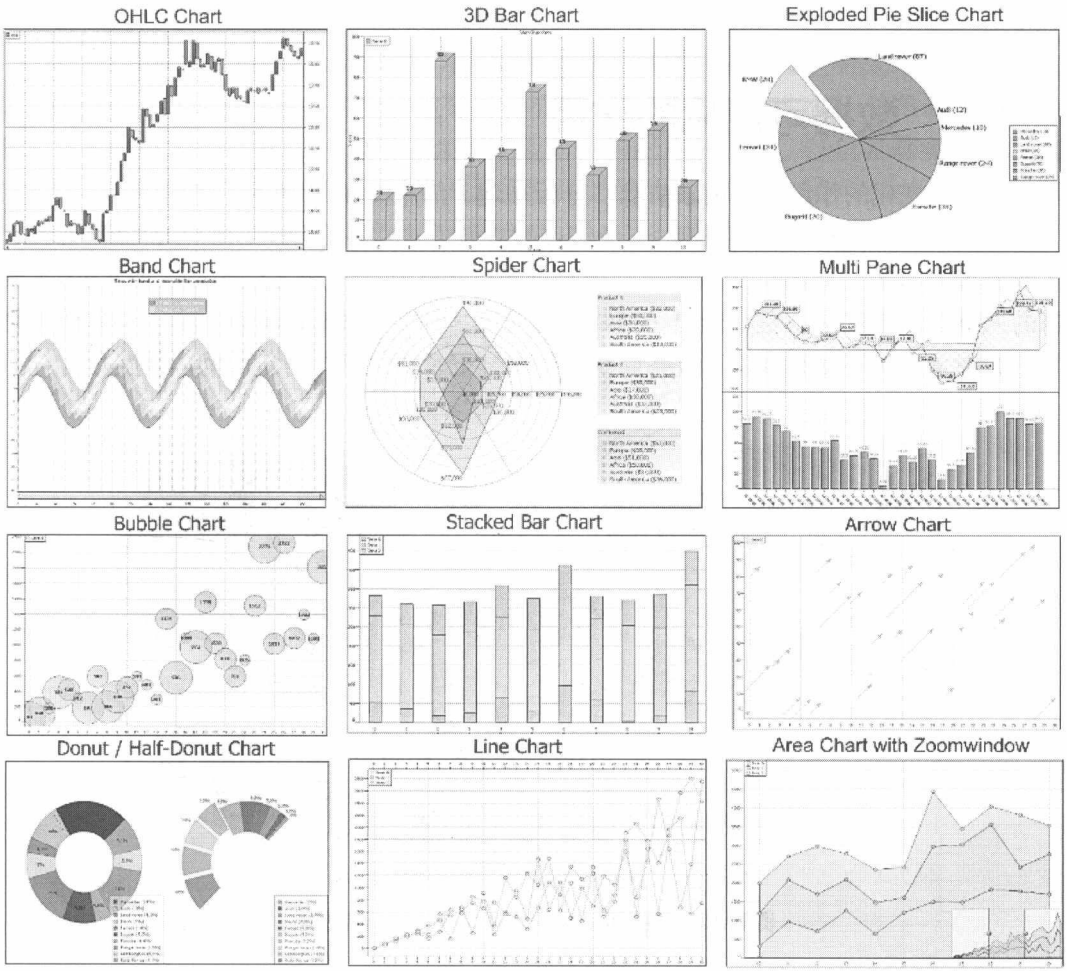



图1-8　基本图表展示数据间的关系

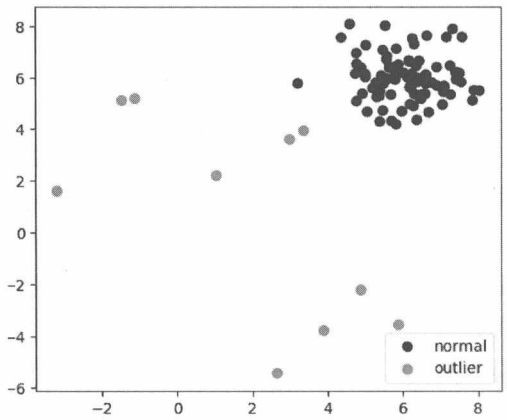



图1-9　异常值

## 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1-10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。

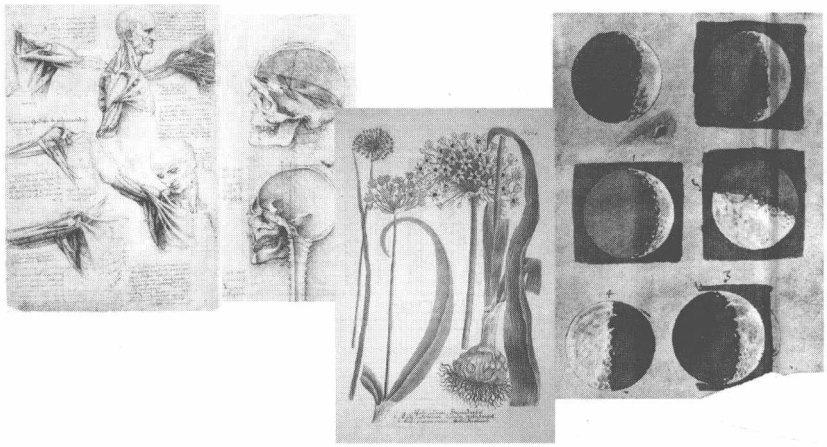


图1-10　数据可视化的作用之一——记录信息

如图1-11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。

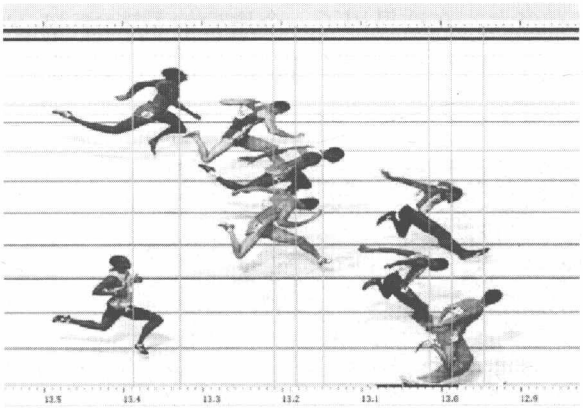


图1–11　田径赛运动员冲刺图

（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，伦敦暴发了一场霍乱，英国医生John Snow绘制了一张街区地图，如图1–12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，据此找到了霍乱暴发的根源——一个被污染的水泵。

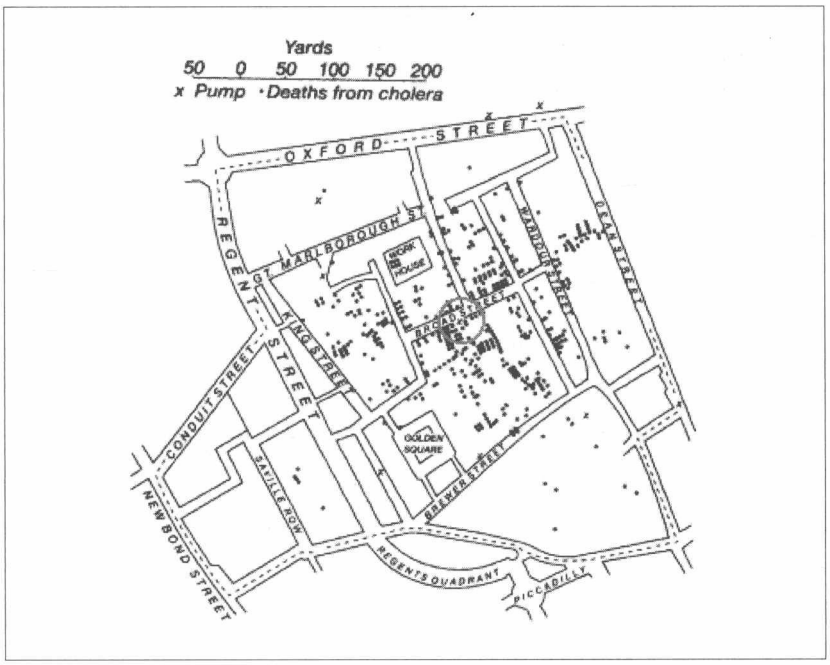



图1–12　伦敦鬼图

（3）信息传播与协同。俗话说“百闻不如一见”“一图胜千言”。

图1–13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A4纸，如果把每一张纸首尾对接，可以绕地球4圈多。由此，就能深刻地感受到雅虎邮箱处理的数据量之大。

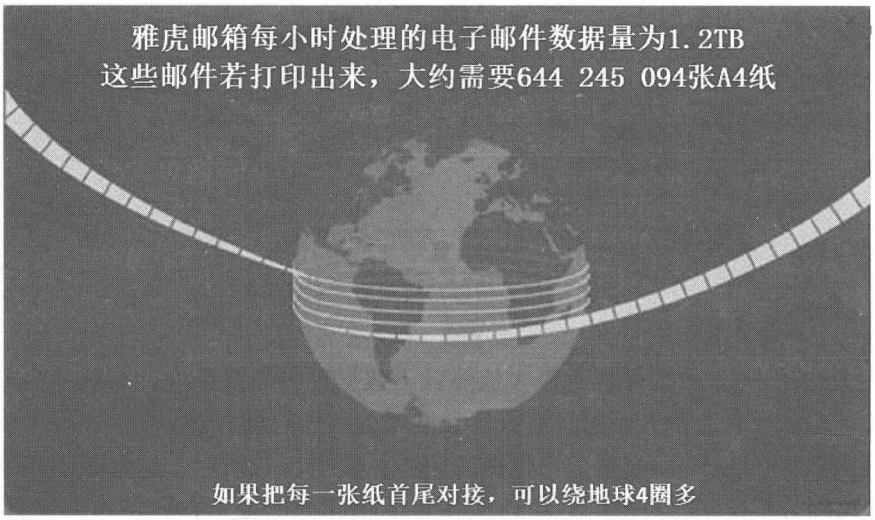



图1-13　雅虎邮箱处理数据量

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，从而能够更快地理解和处理这些信息。

（2）以建设性方式提供决策建议。大数据可视化工具能够用一些简单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柱形图中看到，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，并着手制定改进计划。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　Python常用工具

## 1.3.1　Python中常用的可视化工具

Python在数据科学中的地位，不仅仅是因为numpy、scipy、pandas、scikit-learn这些高效易用、接口统一的科学计算包，其强大的数据可视化工具也是重要组成部分。在Python中，使用的最多的数据可视化工具是matplotlib，除此之外还有很多其他可选的可视化工具包，主要包括以下几大类：

（1）matplotlib以及基于matplotlib开发的工具包：pandas中的封装matplotlib API的画图功能、seaborn、networkx等。

（2）基于JavaScript和d3.js开发的可视化工具：plotly等，这类工具可以显示动态图且具有一定的交互性。

（3）其他提供了Python调用接口的可视化工具：OpenGL、GraphViz等，这类工具各有特点且在特定领域应用广泛。

对于数据科学，用得比较多的是matplotlib和seaborn，对数据进行动态或交互式展示时会用到plotly。

## 1.3.2　Matplotlib与MATLAB

Matplotlib是建立在NumPy数组基础上的多平台数据可视化程序库，John Hunter在2002年提出了matplotlib的初步构想——在Python中画出类似MATLAB风格的交互式图形。

这种接口最重要的特性是有状态的（stateful）：它会持续跟踪当前的图形和坐标轴，所有plt命令（matplotlib pyplot）都可以应用。可以用plt.gcf()（获取当前图形）和plt.gca()（获取当前坐标轴）来查看具体信息。

```
1 import matplotlib as mpl
2 import matplotlib.pyplot as plt
3 mpl.rcParams['axes.linewidth'] = 1.5 #set the value globally, 设置坐标轴线宽
4 import seaborn as sns
5 sns.set()  # 使用seaborn设置绘图风格
```

下面使用MATLAB风格画图，对一组分类变量（categorical variables）进行可视化。

```
1 names = ['group_a', 'group_b', 'group_c']  # 不同分类的名称
2 values = [1, 10, 100]  # 不同分类对应的值
3
4 plt.figure(1, figsize=(9, 3))  # 设置图片大小
5
6 plt.subplot(131)  # 1x3, 第一个子图
7 plt.bar(names, values)  # 柱状图
```