第十九章　可视化基础

可视化是将数据、信息或概念以图形、图表、图像或其他视觉形式呈现的过程. 通过可视化，人们可以更容易地理解复杂的信息、发现模式、识别趋势和进行数据分析. 可视化是数据科学、数据分析、信息传达和决策支持的重要工具. 可视化通常使用图形、图像、颜色、符号和其他视觉元素来传达信息，从而使信息更易于理解和消化.

19.1　可视化案例

下面举几个例子来展示信息可视化.

19.1.1　伦敦霍乱地图

十九世纪中页, 霍乱在伦敦几度流行, 四万多人死于瘟疫, 当时的医学界普遍认为瘟疫是靠笼罩在伦敦上空的” 瘴气” 传播的, 而 John Snow 却认为霍乱是通过水源传播. 为了证明这一点, 在伦敦爆发霍乱的 1854 年, 他冒着生命危险, 走进病情高发的街区, 挨家挨户的调查了整片街区的居民死亡情况, 并绘制了一张死亡地图.

图 19.1: 医生 John Snow 在伦敦地图上标注死者住过的地点 ©From www.theguardian.com

如图19.1，医生 John Snow 使用点分布图 (Dot Distribution Map) 在地图上标出了所有死者曾经居住过的确切地点, 于是他可以直观的在图上看到疾病爆发的密度和分布. 当 John 走访过发病的一整个街区之后, 他从图上发现了异常——有一幢房屋的死亡人数远高

于其他，而这幢房子，紧挨着一个生活水源. 经过调查, 这个水泵连通河水, 那里也是生活污水的排放场所. 饮用水被污染了.

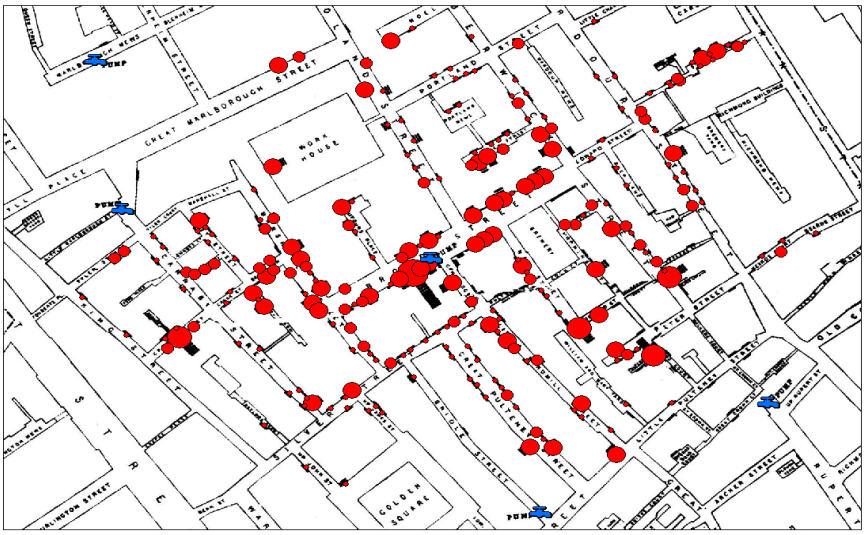

图 19.2: 为了纪念, 有人根据 John Snow 的数据, 重新标注了疫情区域 ©From www.r-bloggers.com

这张地图配合 John 调研的数据, 为证明霍乱是经由受污染的水源传播提供了足够的证据. 于是受污染的水源被拆掉把手, 人们无法再从这里取水, 不久后, 整个街区的疾病流行得到了遏制. 为了纪念, 有人根据 John Snow 的数据, 重新标注了疫情区域，见图19.2.

John Snow 的研究在公共卫生与健康地理学中有重大意义, 并被视为流行病学的发端. 而死亡地图对人类做出的贡献远远超出了医学范畴. 这种方法后来被广泛应用于疾病传播, 犯罪, 地理分析, 建筑学等诸多领域的研究, 甚至衍生了一门专门绘制信息地图的学科：GIS(Geographic Information System).

19.1.2　拿破仑征俄图

1869 年，法国工程师查尔斯 • 约瑟夫 • 米纳德（Charles Joseph Minard）绘制了 1812 年拿破仑征俄图（Map of Napolean's Russian Campaign of 1812），如图19.3，描述了拿破仑在 1812 到 1813 年进攻俄国时所遭受的灾难性损失. 此图将法军东征俄国的过程，精确而巧妙地通过数据可视化的方式展现出来，让人直观感受到拿破仑的 40 万大军，如何在长途跋涉和严寒之中逐步溃散.

线条宽度代表拿破仑的军队人数变化，黄色为进军路线，黑色为撤退路线. 各地理位置连线反映时空关系（从立陶宛到莫斯科军队位移经纬度），文字标明了行军途经的特定地点、河流以及具体人数. 底部温度折线从右到左反映了撤退途中的温度变化.

从图上可以看出，出征时军队人数 42.2 万人，到达莫斯科时还有 10 余万人，而活着返回法国的只有 1 万余人，足以见得拿破仑东征俄国遭受的灾难性损失. 观察黄黑两线交汇处，可以发现活下来的士兵大都中途走岔路返回，前进的大部分都牺牲了. 结合温度变化、河流位置、军队人数，可以看到低温和渡河是士兵牺牲的两大因素.

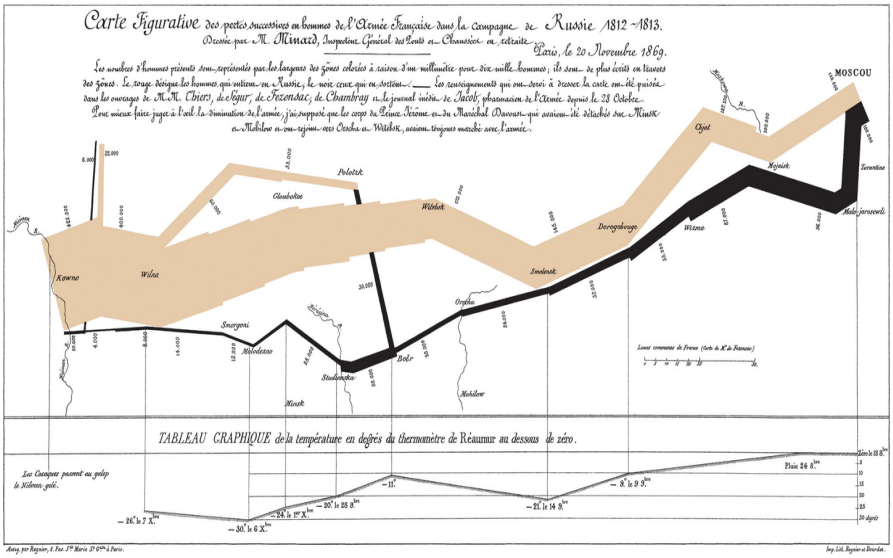

图 19.3: 拿破仑征俄图

	军队规模	进军路线	移动方向	距离远近	河流	温度
元素	线条	线条	线条	线条	线条	点
属性	粗细	方向	颜色	长短	位置	位置
备注	文字	文字	无	图例	文字	文字

19.1.3　南丁格尔玫瑰图

南丁格尔玫瑰图诞生于 19 世纪的克里米亚战争时期，该图表由一位名叫南丁格尔的英国护士长汇总数据并绘制完成. 这张图表不仅清楚地显示了士兵死亡的原因和时间分布，还揭示出医疗条件的不足和卫生环境的恶劣，南丁格尔玫瑰图因此成为了统计图表的先驱之一，被广泛地运用于各个领域.

如图19.4，这里展示了 1854 年 4 月到 1855 年 3 月这一年间士兵的死亡情况. 图中又分为两张小图左图表示 1855 年 4 月到 1856 年 3 月的死亡人数，右图表示 1854 年 4 月到 1855 年 3 月的死亡人数. 对比两张图，可清楚地看到这两年军队死亡人数的变化.

从图中可看出，这一年时间里，死亡人数最多的并不是在战争中受枪伤 (红色部分)，大部分士兵是死于可预防疾病 • (蓝色部分)，特别是冬天的时候 (1854 年 11 月-1855 年 2 月)，死于可预防疾病的士兵人数大幅增加. 由此，可知军队伤亡的真正原因: 影响战争伤亡的并非战争本身，而是由于军队缺乏有效的医疗护理. 正是因为南丁格尔玫瑰图的应用，发现军队伤亡的真正原因，从而，推动军队医疗卫生的改善，挽救更多可预防疾病的士兵，对社会有着现实意义.

19.1.4　2020 美国大选可视化

图19.5和图19.6分别展示了两种不同的选票可视化结果，可以看到图19.5原始地图直接映射的结果在观感上与实际选举结果不一致，而图19.6的可视化根据选票数量调整各州面积，观感上与投票数一致.

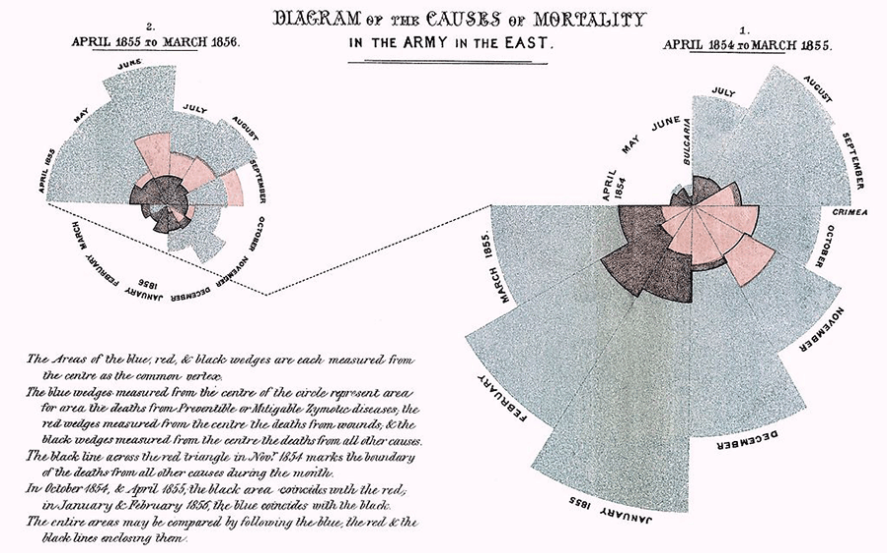

图 19.4: 南丁格尔玫瑰图；蓝色表示死于可预防疾病的士兵人数，红色表示死于枪伤只的人数，黑色表示死于其他意外的人数.

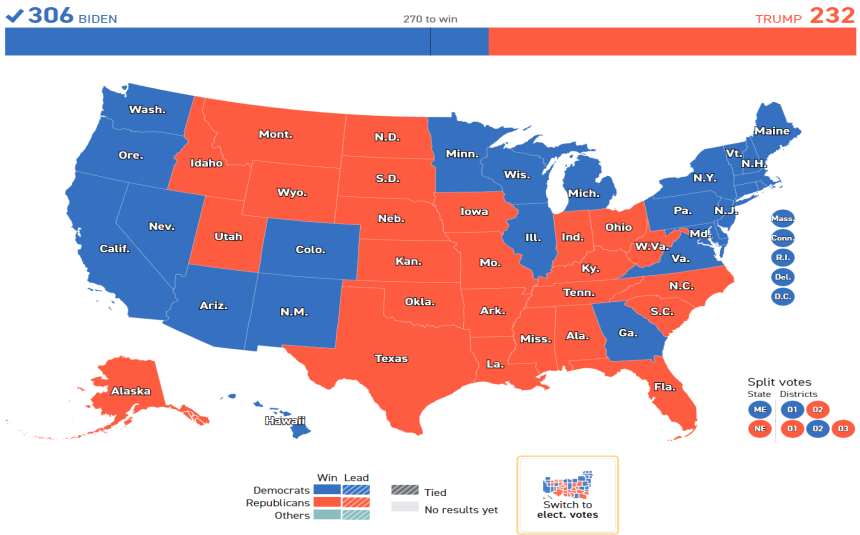

图 19.5: 原始地图直接映射，观感上与实际结果不一致 ©https://www.politico.com/2020-election/results/president/

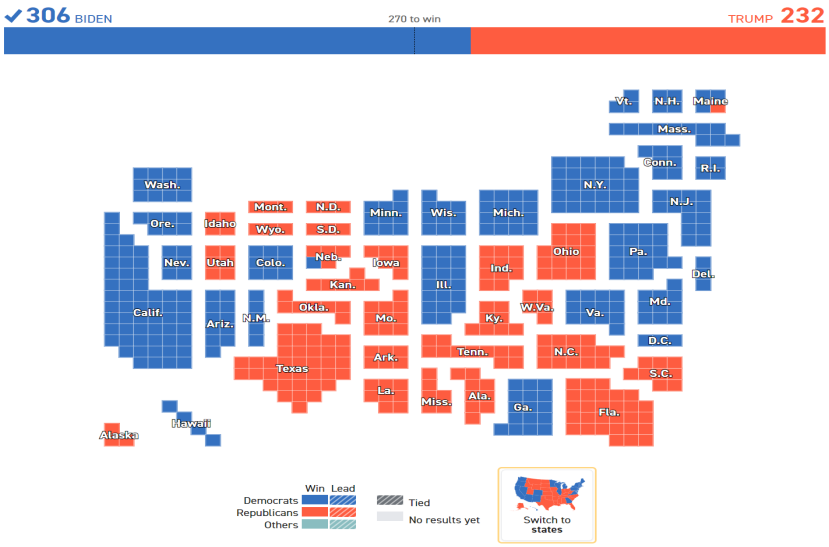

图 19.6: 根据选票数量调整各州面积，观感上与投票数一致

19.1.5　地球演化历史可视化

图19.7出自 2013 年 Juan David Martinez 及其团队的设计，将地球近 46 亿年的进化过程浓缩成这样一张五颜六色的螺旋图.

19.1.6　小结

可视化的历史可以追溯到古代文明，但现代可视化技术的发展主要发生在近几个世纪以来. 随着计算机技术的发展，信息可视化变得更加广泛和复杂. 计算机图形学的兴起使人们能够创建高度复杂的三维可视化和模拟. 借助现代工具和软件，使人们能够轻松地创建各种类型的数据可视化. 可视化的历史反映了人类对于将数据和信息以可视方式呈现的不断追求，以帮助更好地理解世界、做出决策并传达见解. 从最早的手绘图表到现代的交互式可视化，可视化技术一直在不断演进，为各种领域的研究、决策和传播提供了有力工具.

可视化在各个领域中都具有重要的意义，它可以提供以下多方面的价值. 可视化将抽象的数据和信息转化为图形、图表、图像等可视元素，使人们更容易理解和解释数据. 通过可视化，人们可以快速看到数据的趋势、模式和关系，而无需深入研究原始数据. 可视化帮助决策者更好地分析信息，从而做出更明智的决策. 例如，在商业中，可视化仪表板可以显示关键业务指标，帮助管理层迅速识别问题并采取行动. 可视化可以揭示数据中的潜在模式和趋势，这些模式可能在纯文本或表格数据中难以察觉. 通过数据可视化，人们可以更容易地发现新的见解和发展新的假设. 可视化可以用来讲述故事或传达信息. 通过将数据和信息组织成视觉叙述，人们可以更生动地传达信息，引起观众的兴趣和共鸣. 可视化是一种强大的沟通工具，可以帮助将复杂的概念和数据向非专业人士传达. 它还允许人们轻松共享见解和信息，以促进合作和协作. 可视化可以用于检测异常和异常情况. 通过观察图表和图形，人们可以更容易地识别与预期不符的情况，从而采取适当的行动. 科学家使用可视化来呈现实验结果、模拟和模型. 这有助于科学家们更好地理解自然现象，并与同行分享研究成果. 总之，可视化不仅是一种强大的工具，用于数据分析和决策制定，还是一种有效的传达和理解信息的方式. 它在各种领域中都具有广泛的应用，有助于提高工作效率、推动创新和改进决策质量.

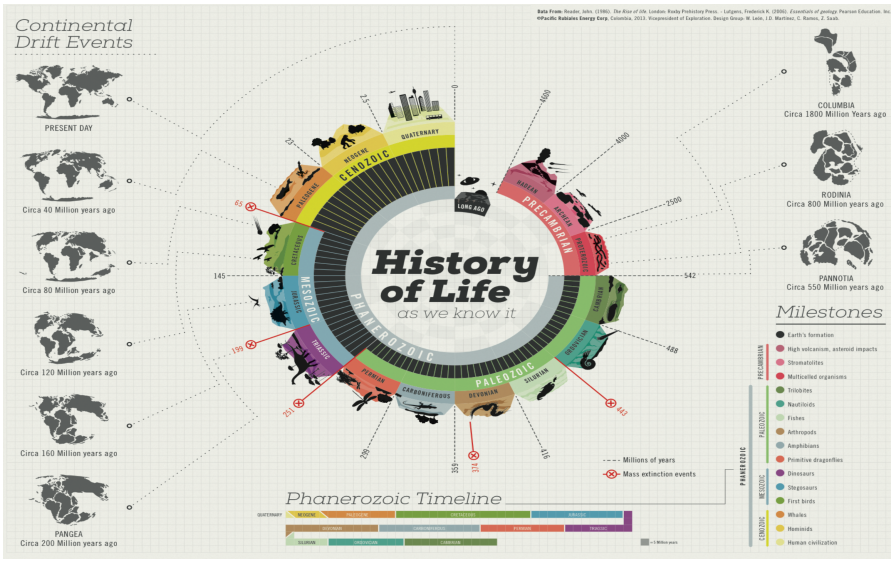

图 19.7: History of Life©https://www.behance.net/gallery/10901127/History-of-Life

19.2　常见可视化工具

有许多可视化工具可供选择，具体取决于数据类型和可视化需求．以下是一些常见的可视化工具：

- 条形图和柱状图：用于比较不同类别的数据，如图 19.8
- 折线图：用于显示数据随时间变化的趋势，如图 19.9
- 散点图：用于显示两个变量之间的关系，如图 19.10
- 饼图：用于显示数据的组成部分，如图 19.11
- 热力图：用于显示数据的密度和分布
- 地图：用于可视化地理数据
- 雷达图：用于比较多个变量之间的关系
- 树状图和网络图：用于可视化层次结构和关系数据，如图19.12和图19.13

19.3　可视化的流程

- 数据分析：数据准备用于可视化（例如，通过应用平滑滤波器、插值缺失值或校正错误测量）．这通常是计算机中心的，几乎没有用户交互．
- 过滤：选择要可视化的数据部分，通常是用户中心的．
- 映射：聚焦数据被映射到几何图元（例如，点、线）及其属性（例如，颜色、位置、大小）；实现表达和效果的关键步骤．
- 渲染：几何数据被转换为图像数据．

19.4　可视化资源限制

可视化设计师需要综合考虑三种非常不同的资源限制因素，分别是计算能力、人类因素以及显示器性能．

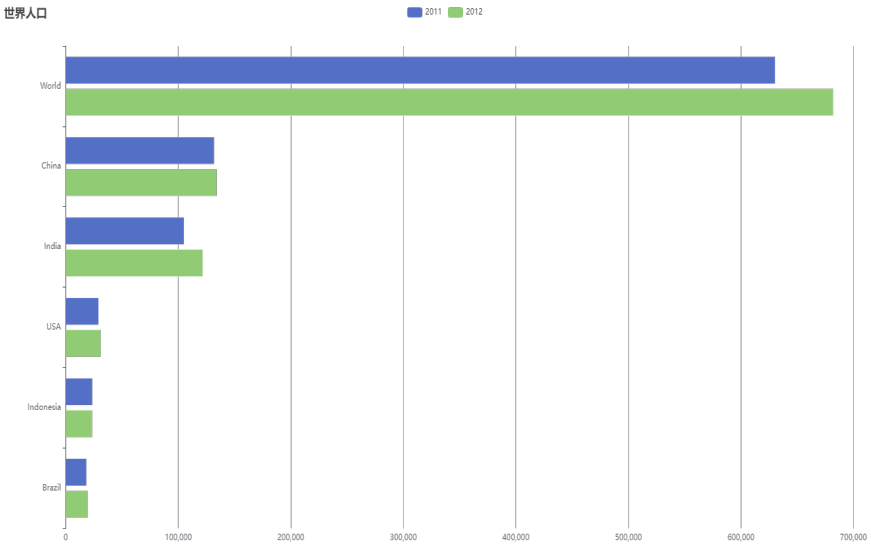

图 19.8: 世界人口数量条形图, ©https://echarts.apache.org/zh/index.html

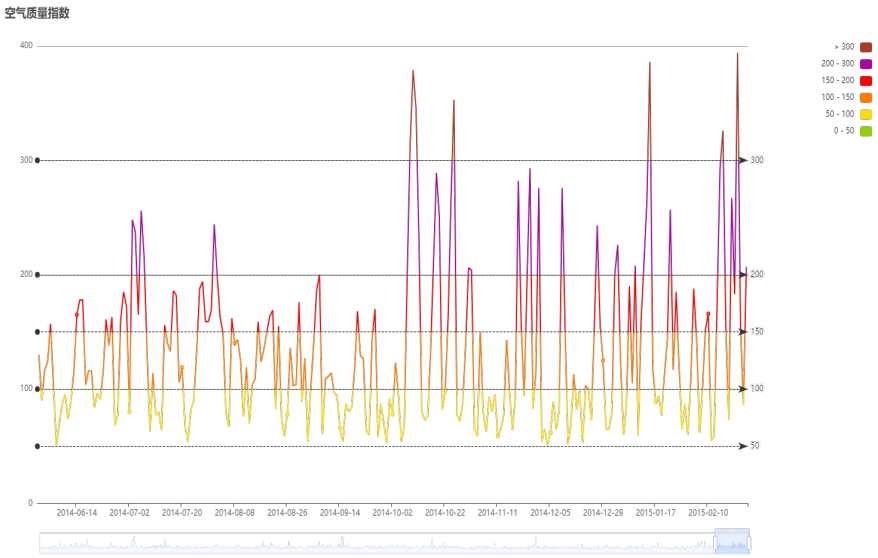

图 19.9: 空气质量折线图, ©https://echarts.apache.org/zh/index.html

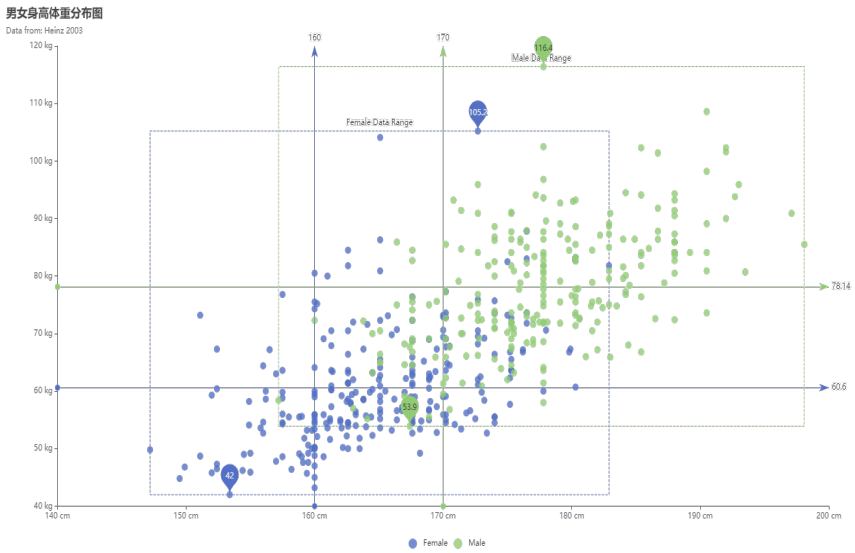

图 19.10: 男女身高体重分布图

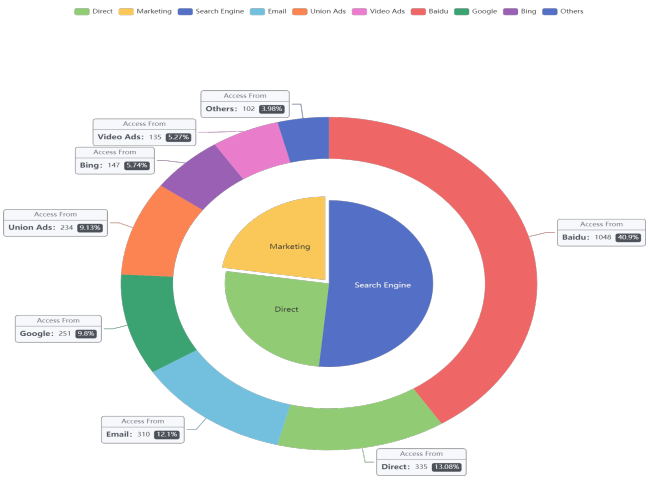

图 19.11: 饼状图

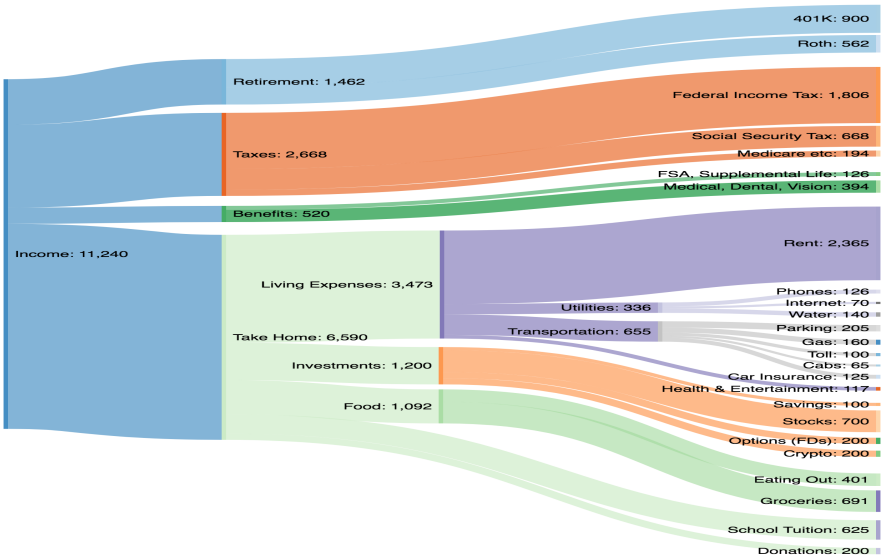

图 19.12: ©https://www.reddit.com/r/dataisbeautiful/comments/bpk5d7/how_my_salary_of_11k_per_month_is_used_in_seattle

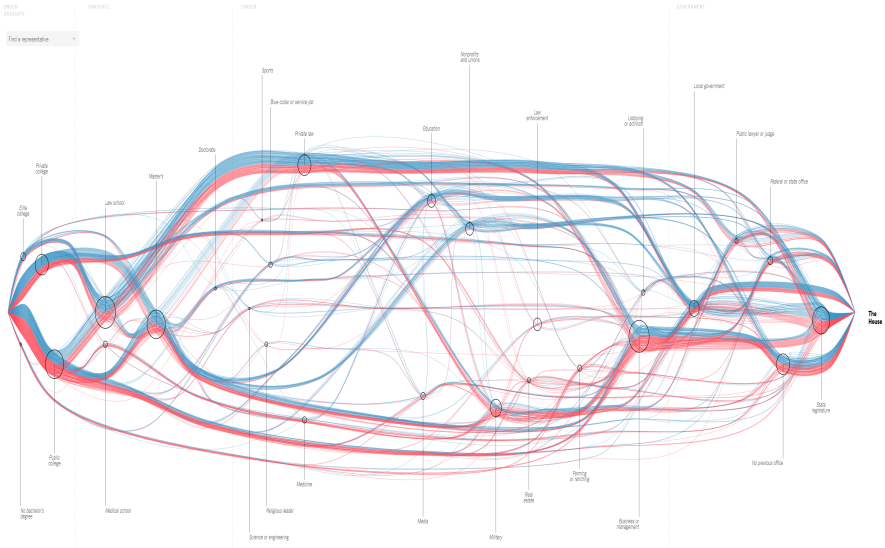

图 19.13: ©https://www.nytimes.com/interactive/2019/01/26/opinion/sunday/paths-to-congress.html

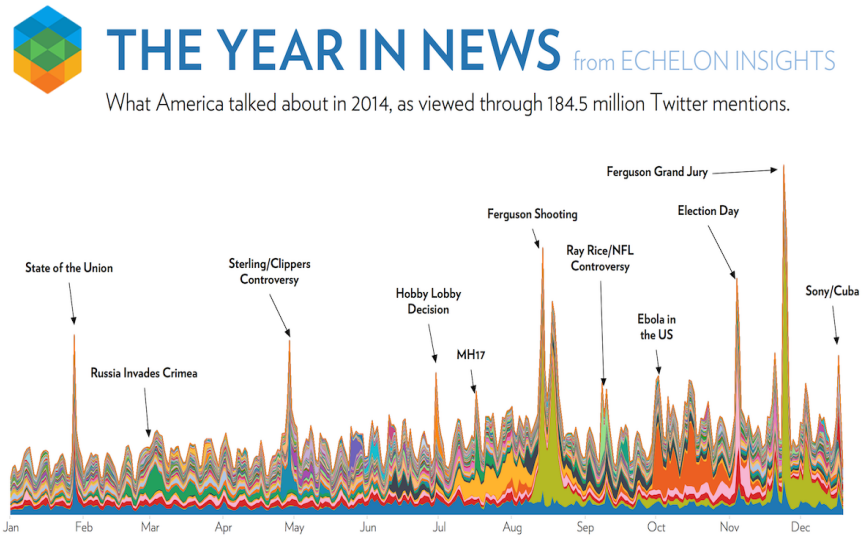

图 19.14: 2014 年美国新闻热度可视化 ©https://echeloninsights.tumblr.com/post/105911206078/theyearinnews-2014

- 计算限制
  - 处理时间：可视化设计必须考虑计算机的处理能力，确保生成和呈现大规模或复杂的数据可视化不会导致性能问题．快速的可视化生成对用户体验至关重要．
  - 系统内存：大型数据集和复杂的图形可能需要大量的内存来存储和处理．设计师需要在可视化设计中谨慎使用内存，以防止系统崩溃或变得缓慢．
- 人的关注和记忆：设计师必须考虑人类的认知能力．信息过载可能导致注意力分散，因此设计必须注重突出显示最重要的数据，以帮助用户集中注意力并记住关键信息．
- 显示限制：像素是宝贵的资源：在设计可视化时，设计师必须优化像素的使用，确保信息清晰可见．不合理的像素使用可能导致图形不清晰或加载速度缓慢．设计师需要在信息的紧凑性和可读性之间取得平衡．太拥挤的图表可能难以解释，而太稀疏的图表可能无法有效传达信息．

19.5　选择合适的可视化类型

根据数据的性质和分析目标，选择合适的可视化类型是创建有力和有效的数据可视化的关键步骤．以下是一些指导原则：

理解数据类型　了解要处理的数据类型是开始选择可视化类型的关键．不同类型的数据需要不同类型的可视化方法．定量数据，如数字和测量值，通常适合使用柱状图、折线图或散点图来表示．而定性数据，如类别或标签，更适合使用饼图、条形图或词云等可视化方式．时序数据，即随时间变化的数据，如股价、气温等，适合用瀑布图、折线图、时间轴、流程图等方式可视化．

目标驱动选择　在决定可视化类型时，首要考虑的是分析目标，例如是否要比较值、显示趋势、发现模式、探索关系、展示分布等．

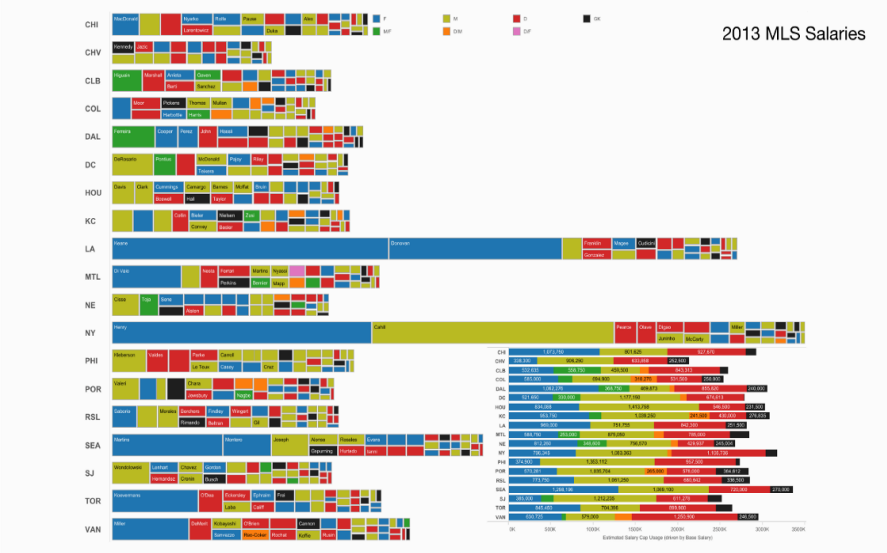

图 19.15: 单个可视化文件中的过多数据会立即使观看者不知所措．当可视化包含太多数据时，信息就会淹没，并且数据会融化成大多数观众无法忍受的图形

考虑受众　考虑可视化的受众是谁，以及他们需要什么样的信息．不同的可视化类型对不同的受众可能更具有说服力．考虑到受众的背景知识，要确保可视化能够满足受众的理解水平．如果受众对数据分析不熟悉，简单直观的可视化可能更有用．

数据量和复杂性　数据集的大小和复杂性也是选择可视化类型的重要因素．如果要处理大量数据点或复杂的数据结构，一些高级可视化技术，如热力图、网络图或树状图，可能更适合帮助展示和理解数据．

一致性　保持可视化元素的一致性是重要的，包括颜色、字体、标签和图表样式．一致性有助于降低观众的认知负担，使他们更容易理解和比较不同的数据．

强调关键信息　可视化应该有助于突出展示最重要的数据和趋势．通过颜色、标签、注释和高亮显示等方式，强调关键信息有助于观众快速识别要点．

避免误导　设计师应该努力避免制造误导性的图表或图形．这包括避免截断轴、使用不恰当的比例、选择不合适的图表类型等．可视化应该真实反映数据，而不是歪曲或夸大．

色彩和标签　颜色应该谨慎使用，避免过多的颜色或使用不明确的颜色方案．合适的颜色选择有助于提高可视化的可读性．每个可视化元素都应该具有清晰的标签和标题，以解释图表的含义和数据的来源．这有助于避免混淆和误解．

交互性　交互性在可视化工具中扮演着重要的角色，它使用户能够更深入地探索和理解数据，以及与可视化图形进行互动．如缩放、过滤、阗懦、交互式图例、滑动条和时间轴、链接等．

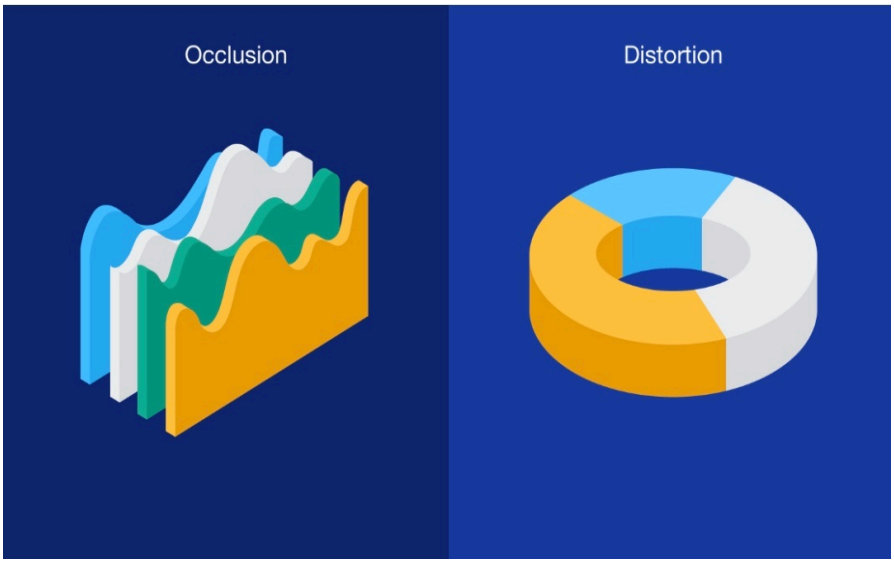

图 19.16: 这里的三维图形造成了遮挡

经验和灵活性　查看类似领域或问题的可视化示例，以获取灵感和最佳实践．最重要的是要灵活，如果一种可视化类型不起作用，尝试其他类型．有时候，根据经验和反馈，可能需要组合多种可视化类型，以更全面地呈现数据．

19.6　常见的可视化评估方式

- 主观评估：主观评估是通过观察和直觉来评估可视化的质量．这包括了设计师和用户的主观看法，例如美观性、可读性、易用性等．主观评估通常涉及用户反馈、焦点小组讨论和专家审查．
- 任务完成度测试：任务完成度测试旨在评估可视化的效用．参与者被要求执行与可视化相关的任务，然后评估他们的任务完成时间和准确性．这可以帮助确定可视化是否支持用户完成任务．
- 眼动追踪：眼动追踪技术记录用户在观看可视化时的眼球运动．这可以帮助确定用户的注意力焦点、浏览模式和视觉搜索路径．眼动追踪可以揭示可视化中的信息流和用户的认知过程．
- 用户反馈和调查：通过收集用户反馈和填写问卷调查，可以了解用户对于可视化的看法和体验．用户反馈可以帮助识别问题和改进设计．

19.7　可视化工具和库

学习可视化最好的方法是实际操作．使用真实数据集创建可视化，尝试不同的图形类型和设计选择，以提高可视化技能．有许多可视化工具和库可供使用，例如：Python 中的 Matplotlib [1]、Seaborn 和 Plotly．JavaScript 中的 D3.js [2]、Chart.js 和 Three.js．商业工具如 Tableau 和 Power BI．

[1] https://matplotlib.org/
[2] https://d3js.org/