北京交通大学
BEIJING JIAOTONG UNIVERSITY




《大数据概论》
大数据分析与挖掘

鲍鹏
软件学院

目录

- 数据理解与特征工程
- 常用数据挖掘算法
- 高级数据建模技术
- 数据可视化技术
  - 大数据可视化概述
  - 大数据可视化方法与技术
  - 可视化工具

大数据可视化概述

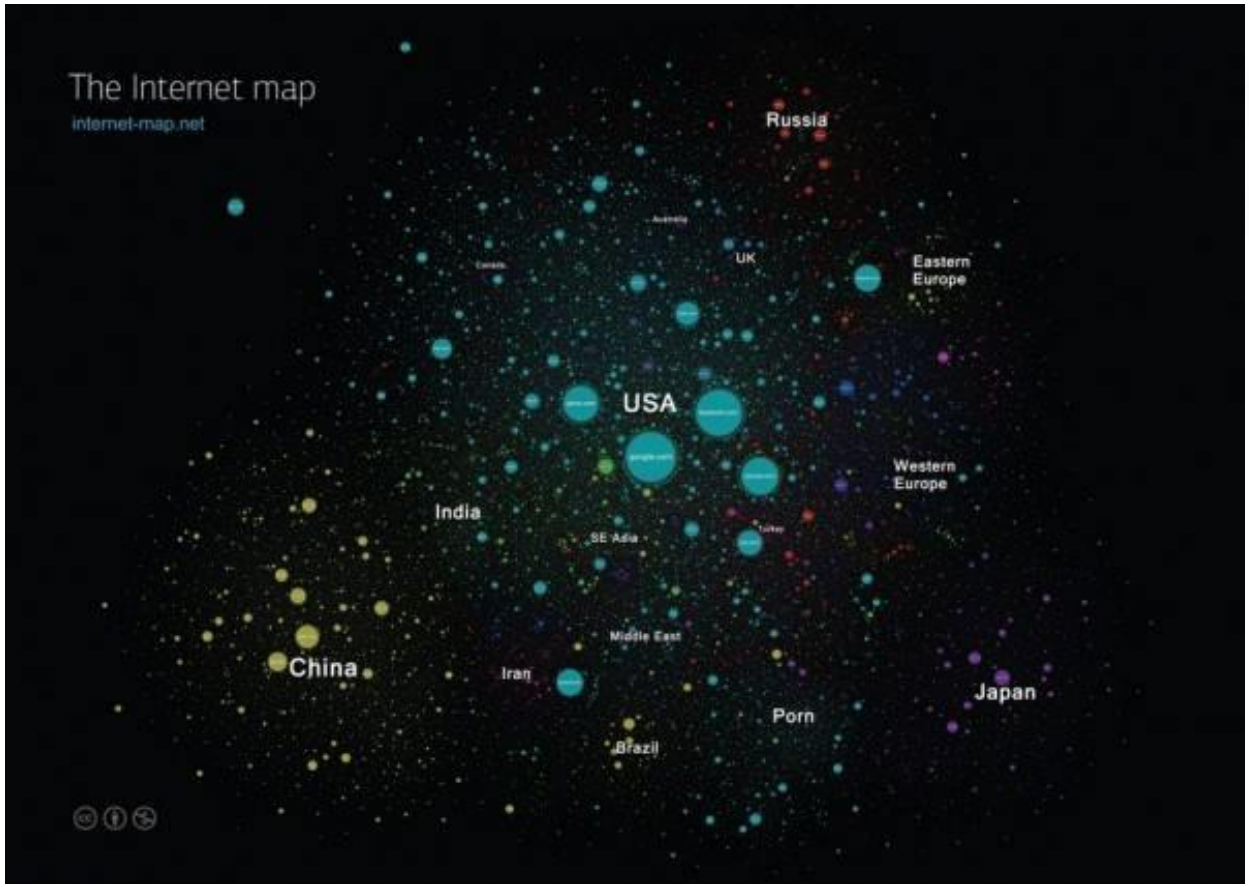
- 数据分析的过程往往离不开机器和人的相互协作和优势互补，大数据分析的理论和方法研究可以从两个维度展开：
  - 从计算机角度出发，强调机器的计算能力和人工智能，以各种高性能算法、智能搜索与挖掘算法等为主要研究内容。
  - 从人作为分析主体和需求主体的角度出发，强调基于人机交互的、符合认知规律的分析方法，如大数据可视化分析技术。

大数据可视化概述

- 数据可视化是指对抽象数据使用计算机支持的、交互的、可视化的表示形式以增强认知能力。
- 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。

大数据可视化概述

- 互联网星际图



大数据可视化概述

- 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，据此发现霍乱爆发的根源是一个被污染的水泵。

大数据可视化概述

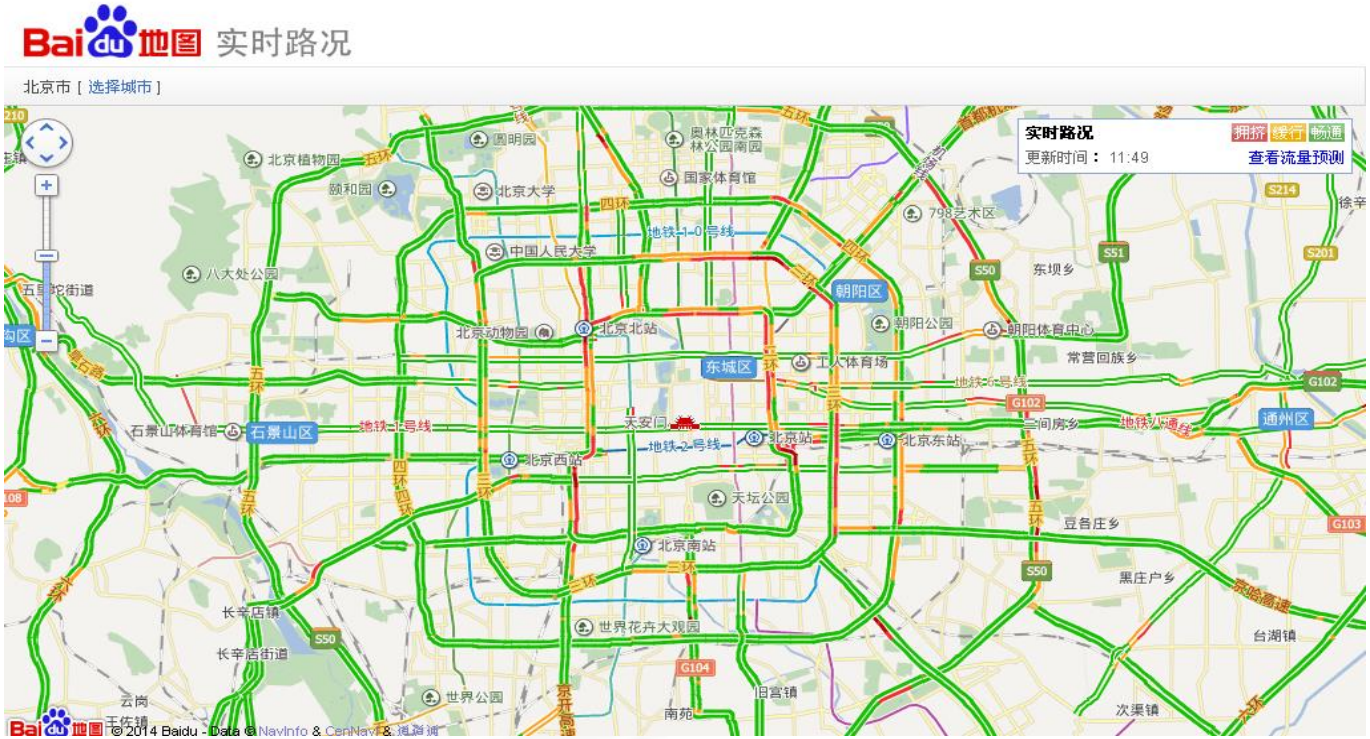
- 大数据可视化分析
  - 定义：指在利用大数据自动分析挖掘方法的同时，利用支持信息可视化的用户界面以及支持分析过程的人机交互方式与技术。

大数据可视化概述

- 大数据可视化分析
  - 有效融合计算机的计算能力和人的认知能力，以获得对大规模复杂数据的洞察力。
  - 将掘取信息和洞悉知识作为目标，根据信息的特征把信息可视化技术分为一维、二维、三维、多维、层次、网络、时序等信息可视化。
  - 随着大数据的兴起与发展，互联网、社交网络、地理信息系统、企业商业智能、社会公共服务等主流应用领域逐渐催生了几类特征鲜明的信息类型，包括文本、网络或图、时空、多维数据等，这些与大数据密切相关的信息类型，将成为大数据可视化的主要研究领域。

大数据可视化概述

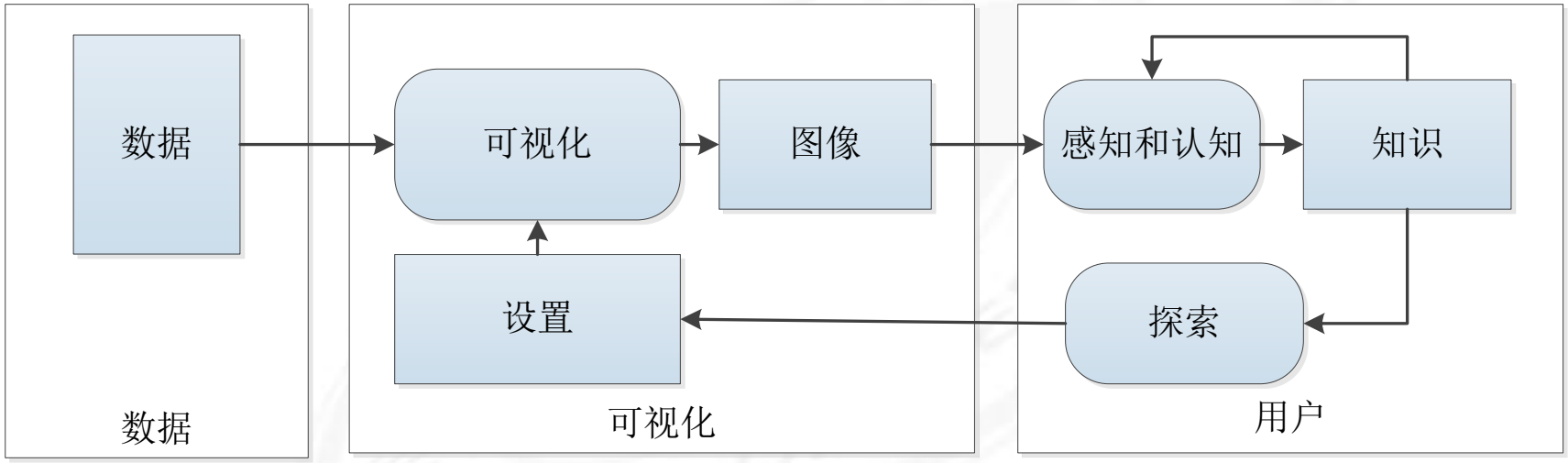
- 大数据可视化技术的应用——观测、跟踪数据



百度地图显示的北京市实时交通路况信息

大数据可视化概述

- 大数据可视化技术的应用——分析数据



用户参与的可视化分析过程

大数据可视化概述

- 大数据可视化技术的应用——辅助理解数据

微软“人立方”展示的人物关系图

大数据可视化概述

- 大数据可视化技术的应用——增强数据吸引力

一个可视化的图表新闻实例

大数据可视化方法与技术

- 文本可视化
- 网络（图）可视化
- 时空数据可视化
- 多维数据可视化

大数据可视化方法与技术

- 文本可视化
  - 文本信息是大数据时代非结构化数据类型的典型代表。如图所示，典型的文本可视化技术是标签云。

大数据可视化方法与技术

- 文本可视化
  - 标签云：将关键词根据词频或其他规则进行布局排列，用大小、颜色、字体等图形属性对关键词进行可视化。
  - 目前，大多数方法用字体大小代表该关键词的重要性。在互联网的应用中，多用于快速识别网络媒体的主题热度，当关键词规模不断增大时，若不设置阀值，将出现布局密集和重叠覆盖等问题，此时需提供交互界面允许用户对关键词进行操作。

大数据可视化方法与技术

- 文本可视化
  - 动态文本时序信息可视化：有些文本的形成和变化过程与时间紧密相关，如何将时序信息进行可视化展示，是文本可视化的重要内容之一。
  - 常见的技术以河流图居多，河流图可以划分为主题、文本及事件河流图等。

大数据可视化方法与技术

- 文本可视化
  - 河流图：河流从左至右的流淌代表时间序列，文本主题按不同颜色带表示，频度以色带宽窄表示。此外，还可以展示主题的合并和分支关系。

Fig.6 ThemeRiver and EventRiver[60,62]

大数据可视化方法与技术

- 文本可视化
  - 结构语义可视化：有些文本中通常蕴含着逻辑层次结构和一定的叙述模式，需要对结构语义进行可视化。如图所示，前者DAViewer将文本以树的形式进行可视化，同时展现了相似度统计，修辞结构以及相应的文本内容。后者DocuBurst以放射状层次圆环的形式展示文本结构。

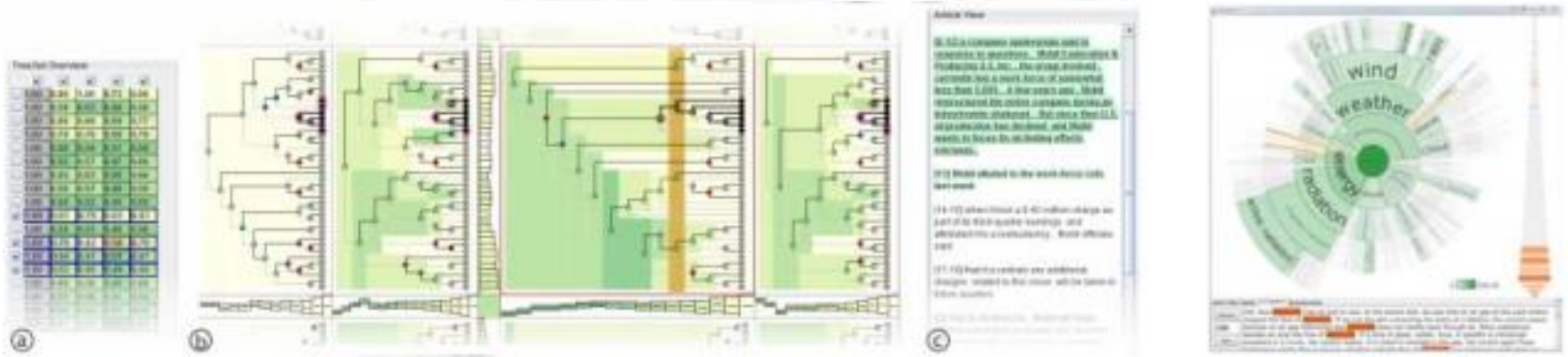


Fig.5 DAViewer and DocuBurst[57,58]

大数据可视化方法与技术

- 网络（图）可视化
  - 网络关联关系是大数据中最常见的关系，如互联网与社交网络。层次结构也属于网络信息的一种特殊情况。
  - 基于网络节点和连接的拓扑关系，直观地展示网络中潜在的模式关系，例如节点或边的连通性，是网络可视化的主要内容之一。
  - 对具有海量节点和边的大规模网络，如何在有限的屏幕空间中进行可视化，将是大数据时代面临的难点。
  - 除了对静态的网络拓扑关系进行可视化，大数据相关的网络往往具有动态演化性，因此，如何对动态网络的特征进行可视化，也是不可或缺的研究内容。

大数据可视化方法与技术

- 网络（图）可视化
  - 图中主要展示了具有层次特征的图可视化的典型技术，例如树状图、气球图、 放射图、 三维放射图、双曲树

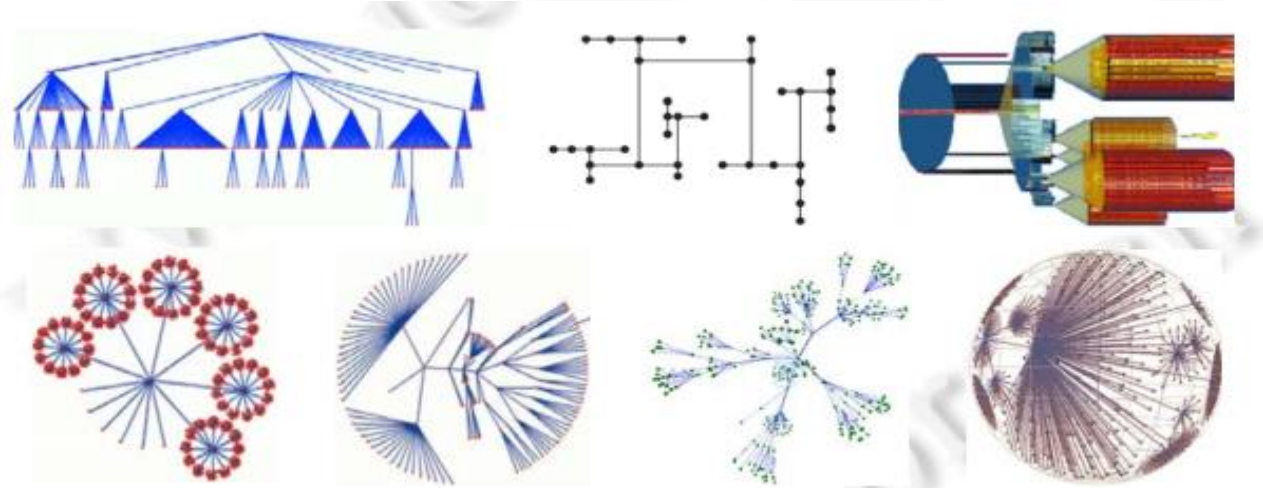


Fig.8 Graph and tree visualization by node-link diagram[64]

大数据可视化方法与技术

- 网络（图）可视化
  - 对于具有层次特征的图, 空间填充法也是常采用的可视化方法, 例如树图技术Treemaps及其改进技术。如图所示是基于矩形填充、Voronoi图填充、嵌套圆填充的树可视化技术。Gou等人综合集成了上述多种图可视化技术, 提出了TreeNetViz，综合了放射图、基于空间填充法的树可视化技术。

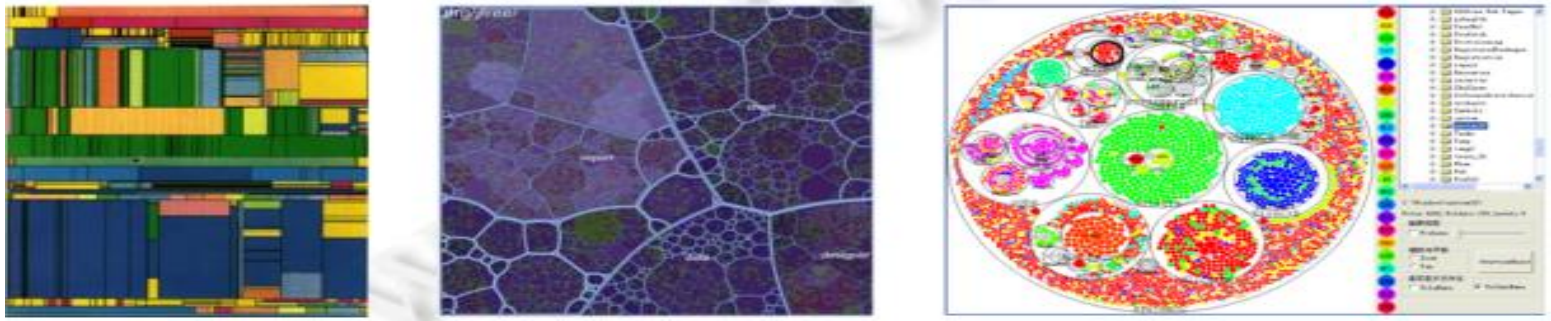


Fig.9 Tree visualization by space-filling diagram[65,67,68]

大数据可视化方法与技术

- 网络（图）可视化
  - 图可视化方法技术的特点是直观表达了图节点之间的关系，但算法难以支撑大规模（如百万以上）图的可视化，并且只有当图的规模在界面像素总数规模范围以内时效果才较好（例如百万以内），当处理大规模图数据时，需要对这些方法进行改进，例如计算并行化、图聚簇简化可视化、多尺度交互等。

大数据可视化方法与技术

- 网络（图）可视化
  - 大规模网络中, 随着海量节点和边的数目不断增多, 例如规模达到百万以上时, 可视化界面中会出现节点和边大量聚集、重叠和覆盖问题, 使得分析者难以辨识可视化效果。
  - 图简化(graph simplification)方法是处理此类大规模图可视化的主要手段。

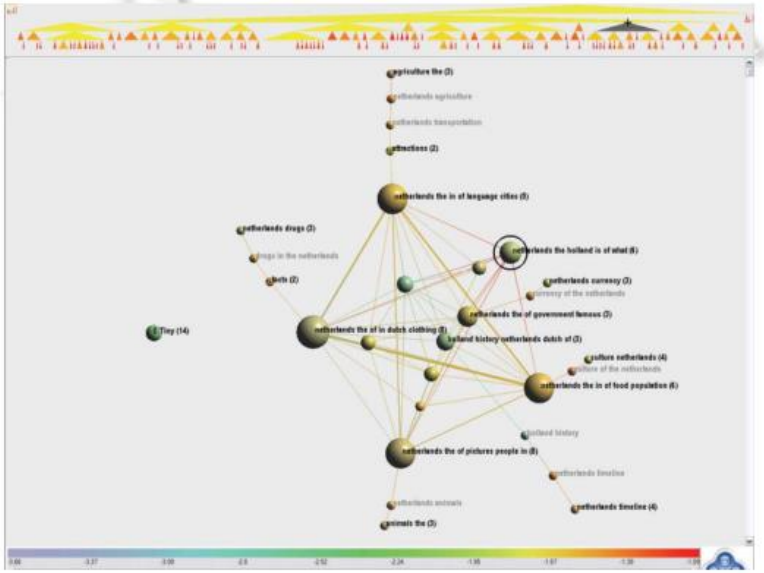
大数据可视化方法与技术

- 网络（图）可视化
  - 一类图简化方法是对边进行聚集处理, 例如基于边捆绑(edge bundling)的方法, 使得复杂网络可视化效果更为清晰。下图展示了3种基于边捆绑的大规模密集图可视化技术。此外，Ersoy等人还提出了基于骨架的图可视化技术, 主要方法是根据边的分布规律计算出骨架, 然后再基于骨架对边进行捆绑。

Fig.10 Graph visualization by edge bundling[70–72]

大数据可视化方法与技术

- 网络（图）可视化
  - 另一类图简化方法是通过层次聚类与多尺度交互, 将大规模图转化为层次化树结构, 并通过多尺度交互来对不同层次的图进行可视化。
  - 例如, 如图所示的**ASK-Graphview**能够对具有**1600**万条边的图进行分层可视化。这些方法技术将为大数据时代大规模图可视化提供有力的支持。交互技术的引入, 将是解决大规模图可视化不可或缺的手段。



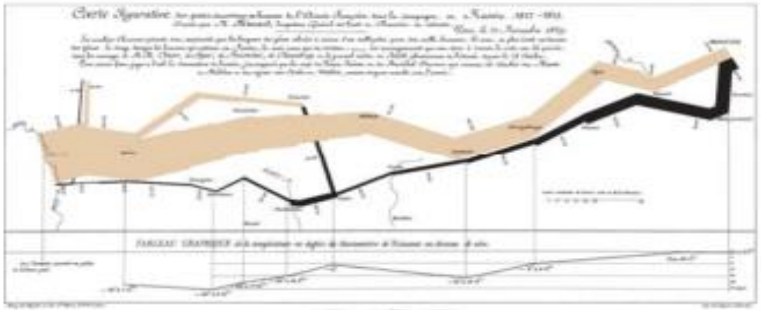
大数据可视化方法与技术

- 时空数据可视化
  - 时空数据是指带有地理位置与时间标签的数据。
  - 传感器与移动终端的迅速普及，使得时空数据成为大数据时代典型的数据类型。
  - 时空数据可视化与地理制图学相结合，重点对时间与空间维度以及与之相关的信息对象属性建立可视化表征，对与时间和空间密切相关的模式及规律进行展示。大数据环境下时空数据的高维性、实时性等特点，也是时空数据可视化的重点。

大数据可视化方法与技术

- 时空数据可视化
  - 时空为反映信息对象随时间进展与空间位置所发生的行为变化，通过信息对象的属性可视化来展现。
  - 流式地图(Flow map)是一种典型的方法，将时间事件流与地图进行融合。下图显示了使用Flow map分别对 1864年法国红酒出口情况以及拿破仑进攻俄罗斯情况可视化的例子。

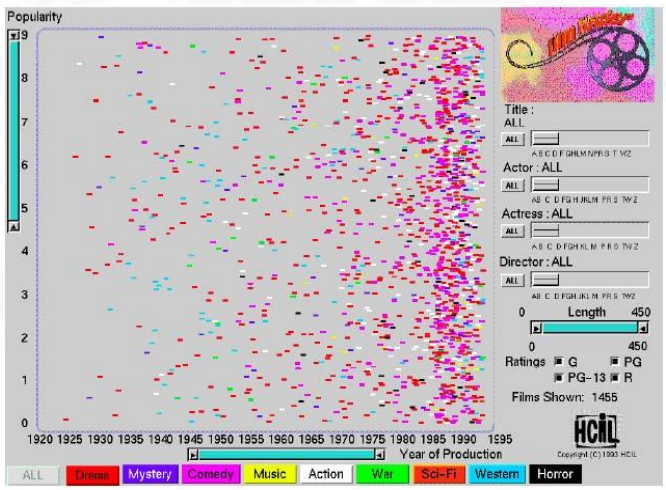
(a) 法国 1864 年红酒出口

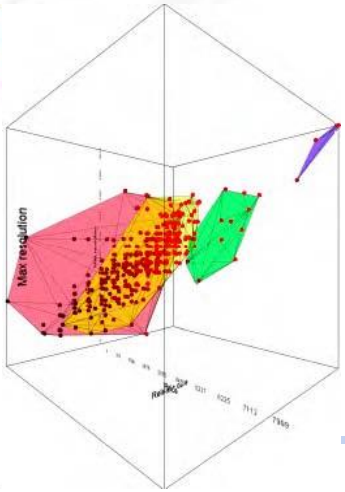


(b) 拿破仑 1812 年进攻俄罗斯

大数据可视化方法与技术

- 多维数据可视化——散点图
  - 时空散点图 (scatter plot) 是最常用的多维可视化方法二维散点图将两个维度属性值集合映射至两条轴，在二维轴确定的平面内通过不同视觉元素来反映其他维度属性值。例如，可通过不同形状、颜色、尺寸等来代表连续或离散的属性值。





大数据可视化方法与技术

- 多维数据可视化——投影
  - 投影(projection) 是能够同时展示多维的可视化方法之一。如下图所示，VaR将各维度属性列集合通过投影函数映射到一个方块图形标记中，并根据维度间的关联度对各个小方块进行布局。基于投影的多维可视化方法一方面反映了维度属性值的分布规律，同时也直观展示了多维度之间的语义关系。

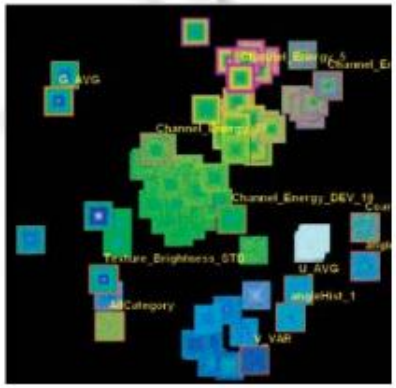


Fig.20 VaR based on projection[91]

图 20 基于投影的多维可视化[91]

大数据可视化方法与技术

- 多维数据可视化——平行坐标
  - 平行坐标(parallel coordinates) 是研究和应用最为广泛的一种多维可视化技术。如图所示，将维度与坐标轴建立映射，在多个平行轴之间以直线或曲线映射表示多维信息。近年来，研究者将平行坐标与散点图等其他可视化技术进行集成，提出了平行坐标散点图 PCP(parallel coordinate plots)。

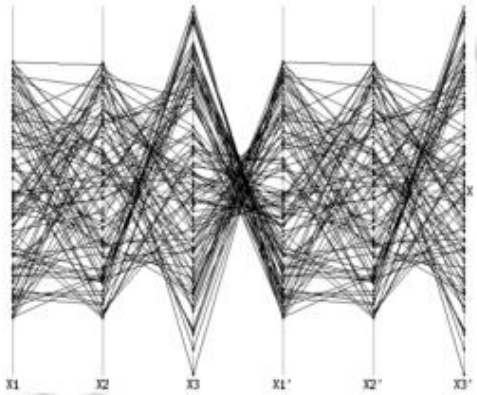


Fig.21 Parallel coordinates[95]

图 21 平行坐标多维可视化技术[95]

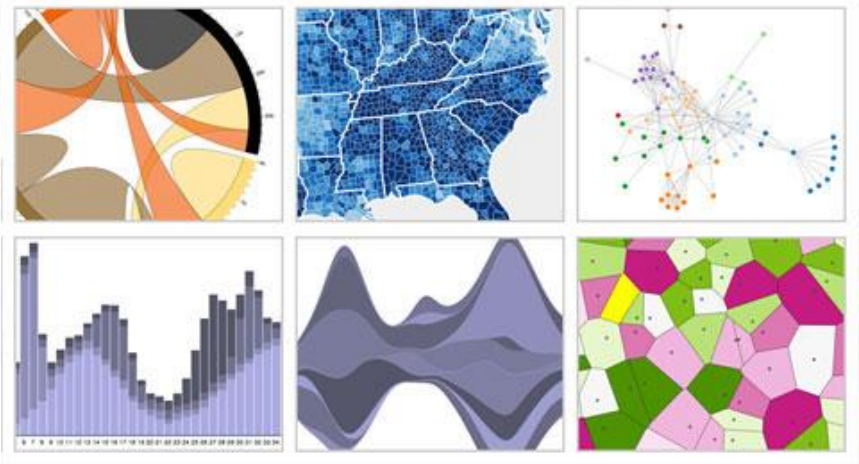
可视化工具

- 常见的大数据可视化工具主要分为三类：
  - 底层程序框架，如OpenGL、Java2D 等；
  - 第三方库，如D3、Echarts、HighChartsVega、OpenLayers、GoogleChart API 等；
  - 软件工具，如 Tableau、Gephi 等。

- 目前常用工具以可方便二次开发的第三方开源库为主。

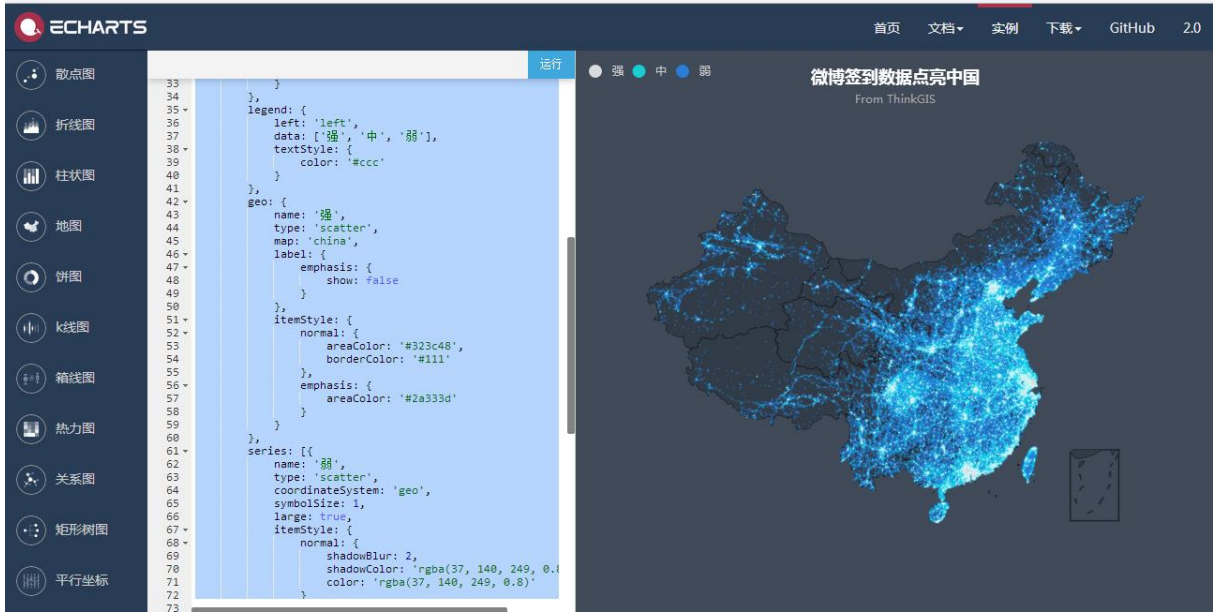
可视化工具

- 可视化工具——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、圆形集群和词云等。D3支持HTML、SVG与CSS。



可视化工具

- 可视化工具——ECharts
  - ECharts是一款可视化开发库，底层用的是javascript封装，能在网页HTML中嵌入ECharts代码显示数据图表。



可视化工具

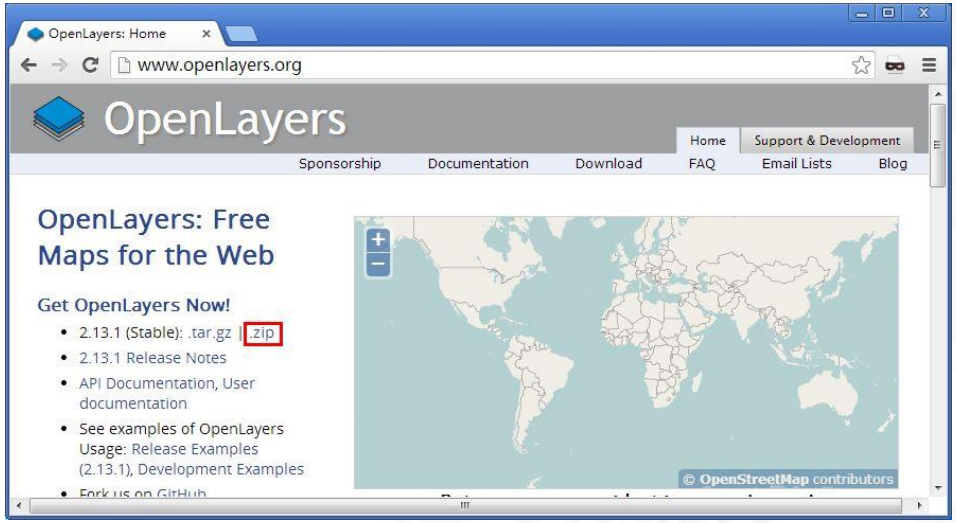
- 可视化工具——ECharts
  - ECharts提供了非常丰富的图表类型，常规的折线图，柱状图，散点图，饼图，K线图，用于统计的盒形图，用于地理数据可视化的地图，热力图，线图，用于关系数据可视化的关系图，多维数据可视化的平行坐标，还有漏斗图，仪表盘等，并且支持图与图之间的混搭，满足绝大部分用户分析数据时的图表制作需求。

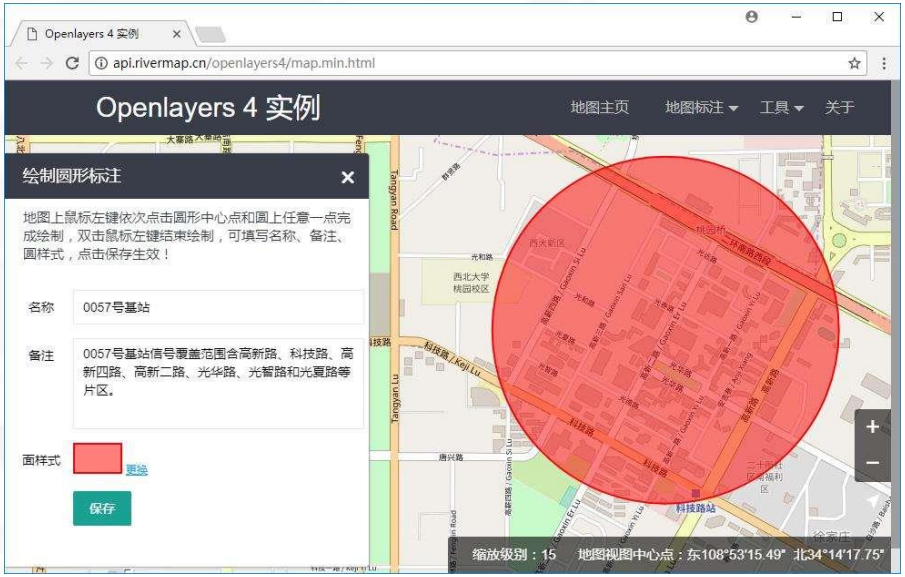
可视化工具

- 可视化工具——OpenLayers
  - OpenLayers 是一个专为Web GIS 客户端开发提供的JavaScript 类库包。
  - 支持的地图来源包括Google Maps、Yahoo、微软Virtual Earth 等，用户还可以用简单的图片地图作为背景图，与其他的图层在OpenLayers 中进行叠加。
  - OpenLayers支持Open GIS 协会制定的WMS（Web Mapping Service）和WFS（Web Feature Service）等网络服务规范，OpenLayers采用面向对象方式开发，使用来自Prototype.js和Rico中的一些组件。

可视化工具

- 可视化工具——OpenLayers





可视化工具

- 可视化工具——Gephi
  - Gephi是一款基于JVM的复杂网络分析软件，主要用于社交图谱数据可视化分析，各种网络和复杂系统，动态和分层图的交互可视化与探测，可生成非常酷炫的可视化图形。