数据库系统概论新技术篇

数据可视化与可视分析

覃雄派

中国人民大学信息学院

2016年12月

数据可视化与可视分析

这一讲的具体内容，包括

1. 可视化的定义及其意义
2. 可视化的一般过程
3. 科学可视化与信息可视化
4. 可视化的若干原则
5. 可视化的若干实例与特色可视化应用
6. 高维数据可视化
7. 可视分析
8. 可视化的挑战和趋势
9. 可视化工具介绍

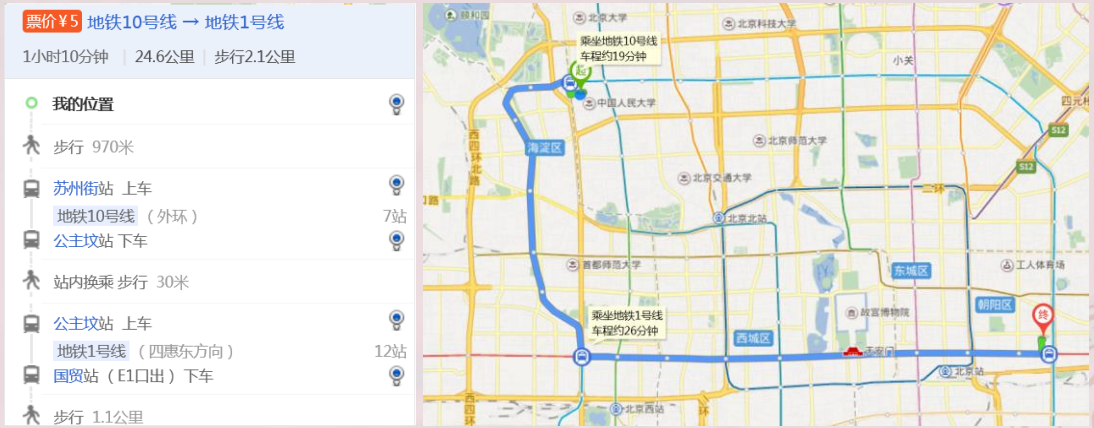
1.可视化的定义及其意义

- 数据可视化由来已久，但是作为一个独立的领域，其发轫于1987年的美国国家科学基金会的“图形、图像处理和工作站”讨论组的一篇里程碑式的报告《科学计算中的可视化》。这个报告，明确提出了将可视化发展成为一个研究领域。自该报告发布以来，可视化领域的研究蓬勃发展。
- 可视化，是数据的可视表现形式(Visual Representation of Data)以及交互技术的总称。它通过图形化的方式把数据给表现出来，方便用户进行观察和理解，并且帮助用户对数据进行探索(exploration)，发现(discovery)数据里面隐藏的模式，获得对数据的洞察力(insight)和理解。

1.可视化的定义及其意义(续)

- 人们常说，“一幅图胜过千言万语”，其意思就是说某些事物，用文字/数字来表达，相当地繁琐，但是用图形来表现，则更加容易把握和理解。
- 比如，通过百度地图，查找从“人民大学”到“国贸三期”的公交路线，总共查出5条线路。

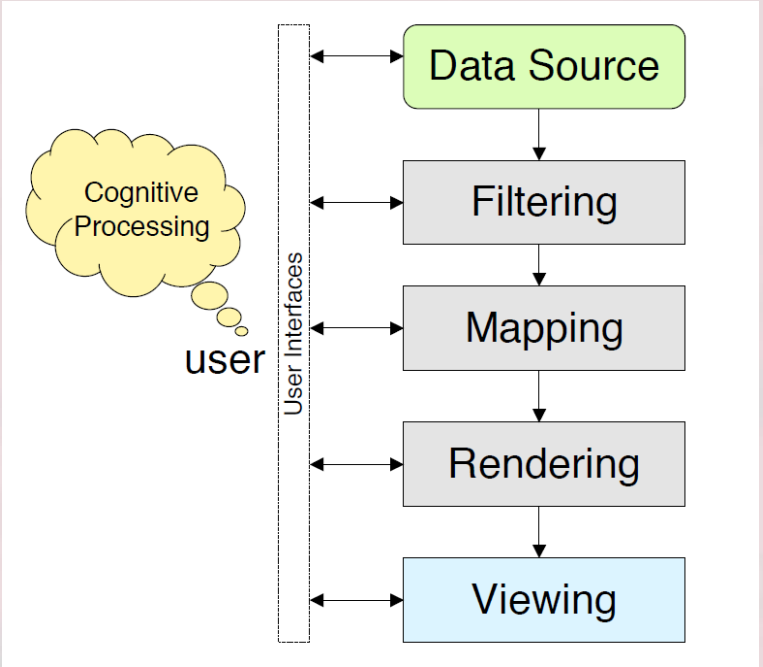
其中一条路线的文字描述和地理信息可视化效果，如下所示：



(a) 文字描述　　(b) 地理信息可视化

2.可视化的一般过程

- 可视化的一般过程，包括：
- (1) 过滤，是选取原始数据集(Raw Dataset)的一部分进行可视化
- (2) 映射(Mapping)，是指将抽象数据，转换为可视化表示的过程
- (3) 渲染(Rendering)，是通过图形渲染库和显示卡的帮助，把经过映射的数据，以二维或者三维图形的形式绘制出来
- (4) 交互(Interaction)，是指计算机对用户的某种特定动作，做出反应。 比如，计算机可以识别用户的手势，适时地改变渲染的效果。

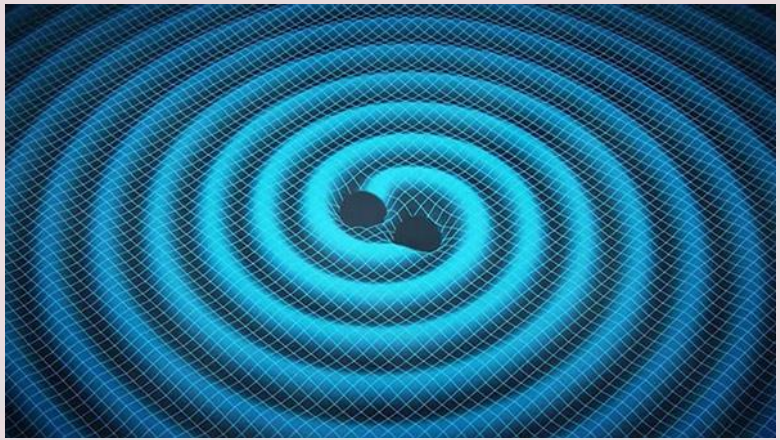


3.科学可视化与信息可视化

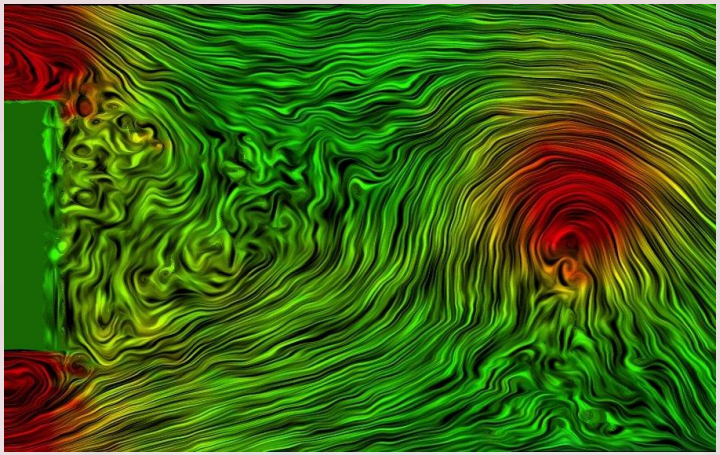
- 目前，可视化领域包括三个主要的分支，分别是科学可视化、信息可视化、以及可视分析等
- 科学可视化是其中最成熟的一个研究分支，它主要面向自然科学实验、探测活动所产生的数据，进行建模、操作和处理。
- 20世纪90年代以来，随着互联网的发展和信息的爆炸，数据可视化的另外一个分支—信息可视化逐渐兴起。
- 信息可视化要处理的数据类型丰富多样，可以是数值型数据，也可以是 类别数据，数据具有不同的结构，如层次结构、网状结构等。具体包括时间序列数据、文本、地图、社交网络等，数据来自不同的行业，比如新闻、电商、股票市场、社交网站等。

3.科学可视化与信息可视化(续)

- 科学可视化实例



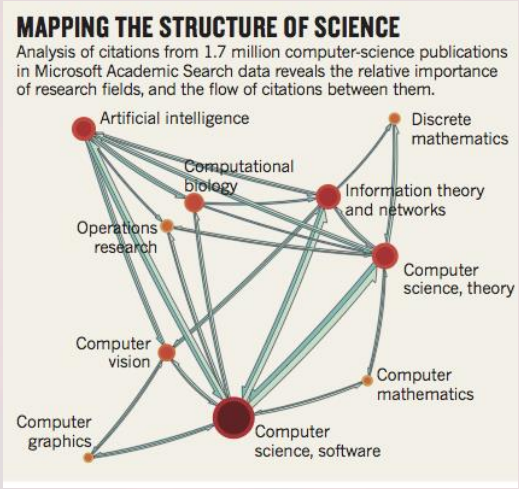
(a)黑洞碰撞与引力波

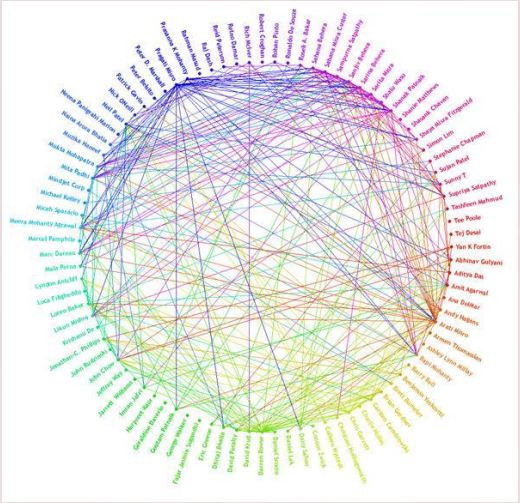


(b) 液体的流动

3.科学可视化与信息可视化(续)

- 信息可视化实例





(a) 计算机科学的学科结构

(b) 企业家在社交媒体上的六度距离

3.科学可视化与信息可视化(续)

科学可视化与信息可视化的比较

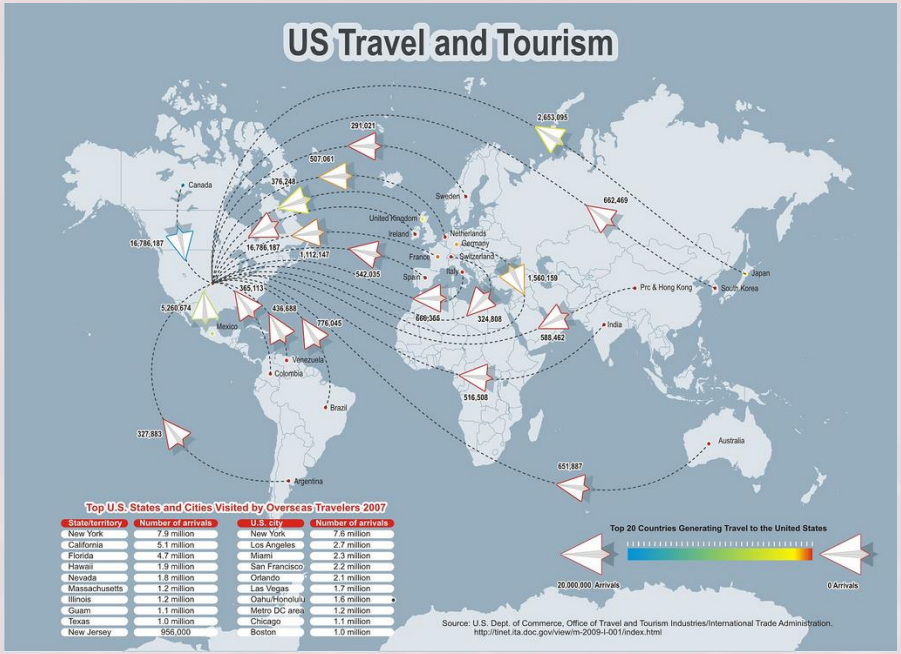
对比项	科学可视化	信息可视化
目标任务	研究科学问题，深入理解自然界中的现象	探索、发现信息之间的关系，发现隐藏的模式
应用领域	气象、高能物理、天文学、生物学、医学、地质学、流体力学…	传感器网络、电子商务、金融、社交网络、新闻、博客、反恐…
数据来源和类型	→来自科学实验、观测、仿真 →结构化数据，具有物理、几何属性	→来自各个领域 →结构化数据和非结构化数据，一般不具有物理、几何属性
主要方法与要求	→预处理、映射、渲染、交互 →准确反映数据中的物理、几何关系	→数据挖掘与机器学习、映射、渲染、交互、以及可视化分析 →把抽象复杂的信息及其关系，映射为有效的可视化表示(Representation)，寻找合适的可视化形式
面向用户	→面向科学家	→面向非技术人员、普通用户、管理人员

4. 可视化的若干原则

- 可视化的目的，是把复杂数据有效地展示出来，首要的原则是**准确和清晰**(precision, Clarity)。准确是指可视化结果反映的是数据的本来面目或者本质(substance)，清晰是指可视化结果，所表达的含义要明确。 此外，
- (1) 我们希望在更小的空间里(Less Space)，用最少的图形(可视化并非越繁琐越好，而是越简洁越好, Less Ink)，在最短的时间里(Less Time)，传达给用户最多的信息(More Ideas)。对数据进行合理简化，突出重点。
- (2) 可视化的结果，需要阐明事物之间的相互关系，以及事物的变化趋势，对于类似的事物要方便用户进行比较。
- (3) 使用用户熟悉的事物，对需要比较的数据进行比较。
- (4) 构建实物场景，生动展现数据。
- (5) 在可视化设计过程中，要考虑把交互方式和动画效果加进去。动画效果可以从时间和空间维度对事物的发展变化过程进行刻画，以便给用户创造沉浸式的体验。

4. 可视化的若干原则(续)

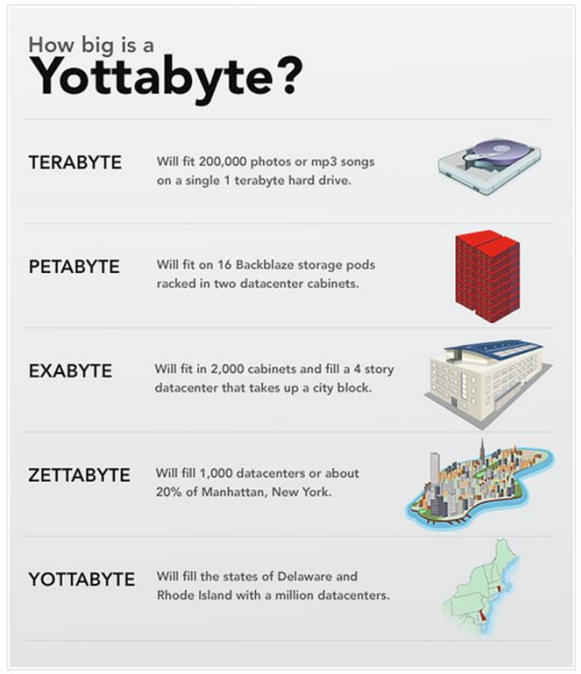
- 可视化的结果，需要阐明事物之间的相互**关系**，以及事物的变化**趋势**，对于类似的事物要方便用户进行比较。



美国(来自世界各国的旅游者)

4. 可视化的若干原则(续)

- 使用**用户熟悉的事物**，对需要比较的数据进行比较。



Yottabyte到底有多大

1TB硬盘

两个“数据中心机柜”

2000个“数据中心机柜”，一个街区

1000个“数据中心”，20%曼哈顿区

1000000个“数据中心”，填满特拉华 Delaware州和罗德岛

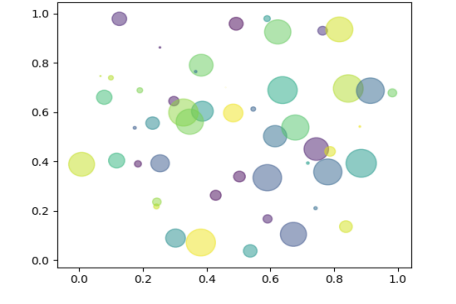
4. 可视化的若干原则(续)

- 构建**实物**场景，生动展现数据。

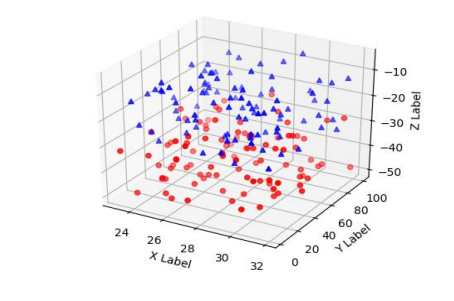
土地使用情况(不同用途的土地的面积)

5.可视化的若干实例与特色可视化应用

- 散点图(Scatter Plot)是对点数据(Point Data, 即向量)的集中趋势、分布形状、离散趋势进行把握的基本的可视化形式。
- 集中趋势，是指数据向中心点靠拢的趋势。分布形态包括数据的分布形状、数据的分布是对称和还是非对称的、平缓的还是比较陡峭的。离散趋势，指的是数据离开中心点的趋势。



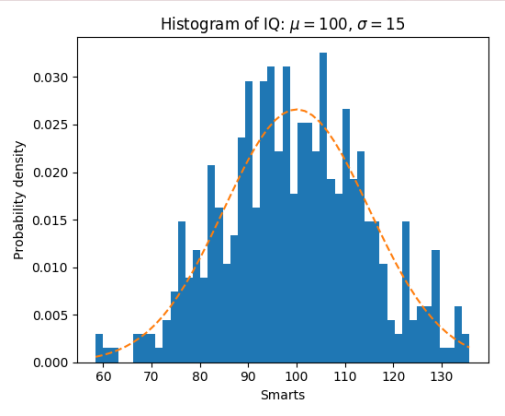
(a) 2d散点图



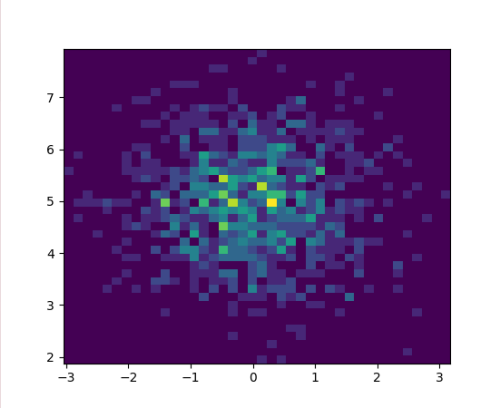
(b) 3d散点图

5.可视化的若干实例与特色可视化应用

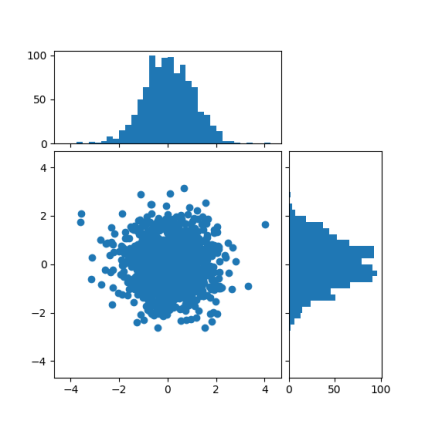
- 直方图，也称为频率直方图(frequency histogram)。它是统计学中用于表示频率分布的图形。



(a) 1维直方图



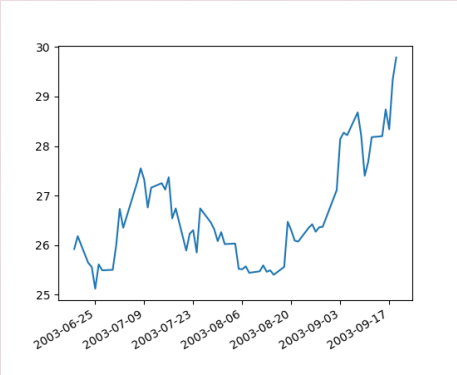
(b) 2维直方图



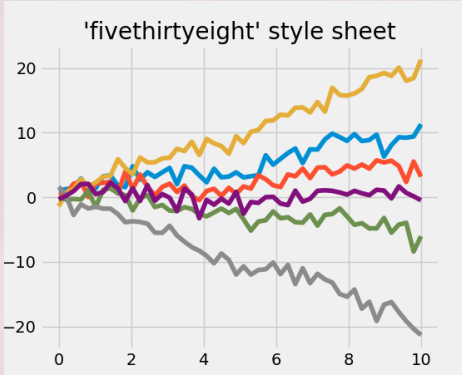
(c) 1维直方图和散点图的组合

5.可视化的若干实例与特色可视化应用

- 线图通过画折线、或者样条曲线，把若干个数据点连接起来。线图分单线图(Line graph)和多线图(Multiple line graph)。



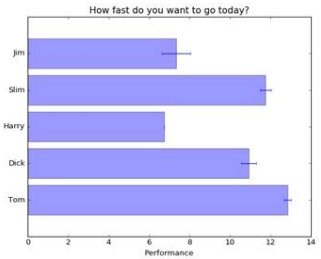
(a) 单线图



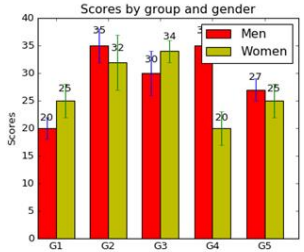
(b) 多线图

5.可视化的若干实例与特色可视化应用

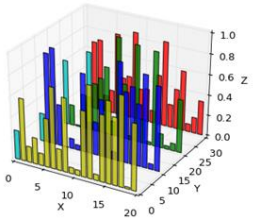
- 柱状图和饼图，一般用来显示一个数据系列里各个数值之间的相对大小关系。柱状图的各个柱子的高度的比例关系、以及饼图的每个扇面的大小的比例关系，反应了数据系列中各个数值之间的大小关系。



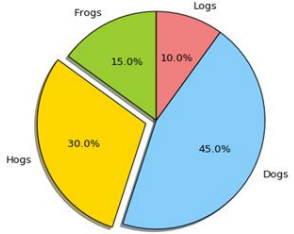
(a) 柱状图



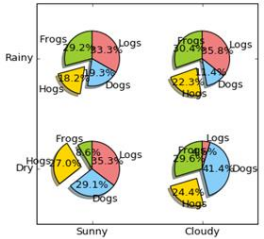
(b) 复式柱状图



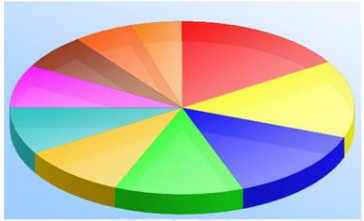
(c) 复式柱状图(三维渲染)



(d) 饼图



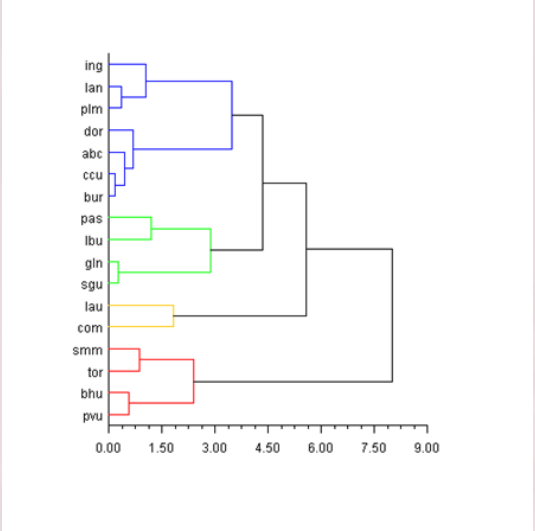
(e) 复式饼图



(f) 饼图(三维渲染)

5.可视化的若干实例与特色可视化应用

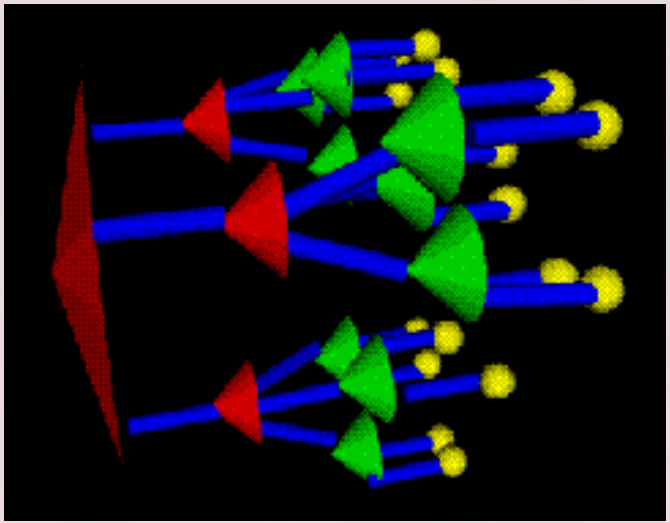
- 树状结构(Tree)，是可视化中应用得最广泛的一种图形结构之一，它一般用于表现某种层级关系，比如某个组织的各个部门、某个家族的族谱等。



凝聚层次聚类的可视化
基于Los Angeles County的17个学区(17 school districts)的若干指标，对学区进行聚类

5.可视化的若干实例与特色可视化应用

- 圆锥树(Cone Tree)，用于对层次结构进行可视化展现。在圆锥树中，层次结构通过3维方式进行展现，以利于最大化使用屏幕空间，以及展现整个层次结构。



圆锥树实例

5.可视化的若干实例与特色可视化应用

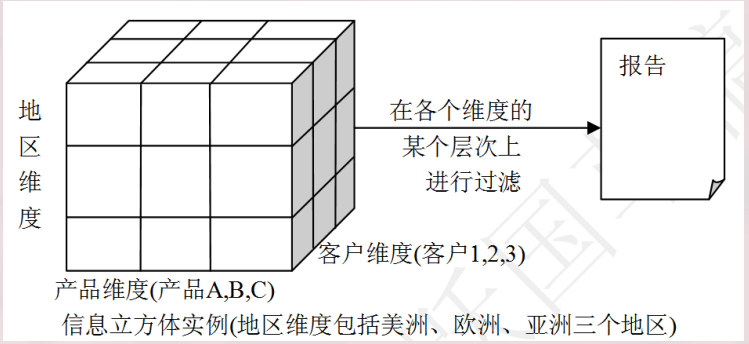
- Tree map是由马里兰大学的Ben Shneiderman教授于20世纪90年代提出的，其最初目的是找到一种有效了解磁盘空间使用情况的方法。
- Smart Money杂志对市场上市值排名靠前的股票进行了展示，他们使用了Treemap。

矩形的大小表示行业、公司的市值。矩形的颜色，表示股票的涨跌程度，用从红到绿的渐变的各种颜色表示。

Smart Money: Map of the Market

5.可视化的若干实例与特色可视化应用

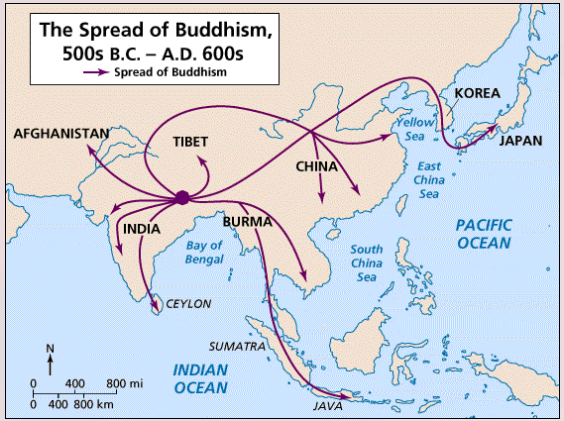
- 信息立方体(Infor Cube)，是一种多维的数据结构，用于从不同维度对数据进行汇总和观察。
- 右图展示了一个关于销售额的信息立方体，它有三个维度，分别是客户维度、产品维度、和地区维度。
- 一般来讲维度具有层次结构，比如产品维度下具有产品大类→产品小类→产品这样的层次。
- 用户可以对各个维度基于某种层次进行过滤，对销售数据进行汇总和观察。
- 比如，我们可以查看某个产品大类各个小类在亚洲地区的销售情况。



信息立方体实例(地区维度包括美洲、欧洲、亚洲三个地区)

5.可视化的若干实例与特色可视化应用

- 在地图(Map)上进行可视化，可以展示事物的发展涉及的不同地理位置。

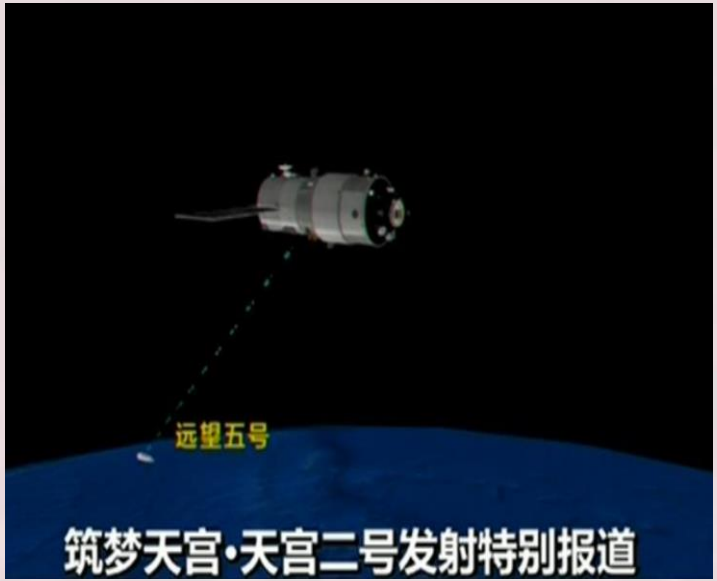


佛教的传播过程(公元前500年-600年)

在谷歌地球上显示航线数据

5.可视化的若干实例与特色可视化应用

- 我们国家在空间站和载人飞船任务中，相关单位研发了“飞行任务轨道三维可视化”软件。
- 飞行任务轨道可视化软件实现了对空间站、飞船在轨飞行任务的图形化展示，直观展示空间站、飞船的运行状态和运行环境，包括对飞船和空间站对接过程的展示和监控。
- 为保障空间站和载人飞船的任务圆满完成，发挥了重要作用。

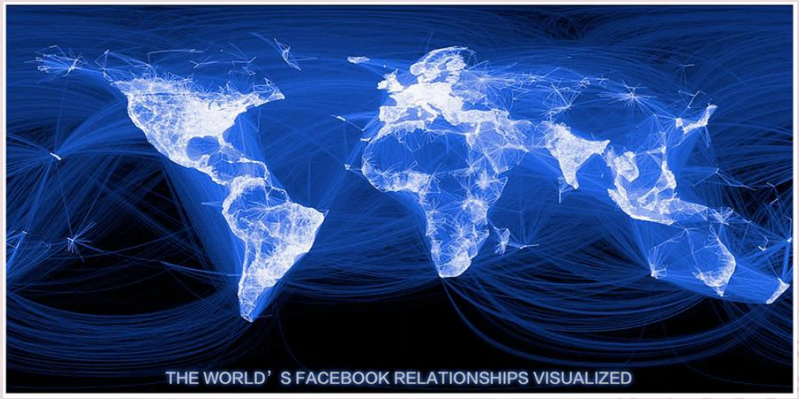


飞行任务轨道可视化(电视屏幕截屏)

5.可视化的若干实例与特色可视化应用

- 可视化提供了观察社交网络(Social Network)的有力工具
- **(1) 世界范围内Facebook的好友关系可视化**
- (2) 博客空间的可视化

展现不同地理区域之间(geography)、国家之间(Country)、不同城市(City)之间好友关系的多少(friendships between them)



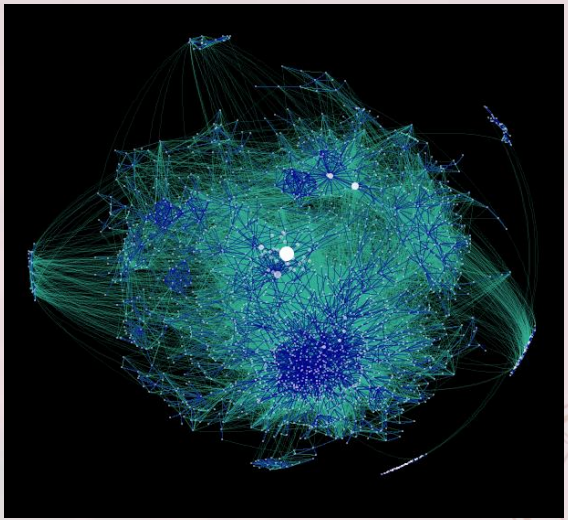
世界范围内Facebook好友关系的可视化

5.可视化的若干实例与特色可视化应用

- 可视化提供了观察社交网络的有力工具
- (1) 世界范围内Facebook的好友关系可视化
- **(2) 博客空间的可视化**

节点的大小表示博客的入链接in links，同样的颜色，则代表隶属同一个域名的博客。

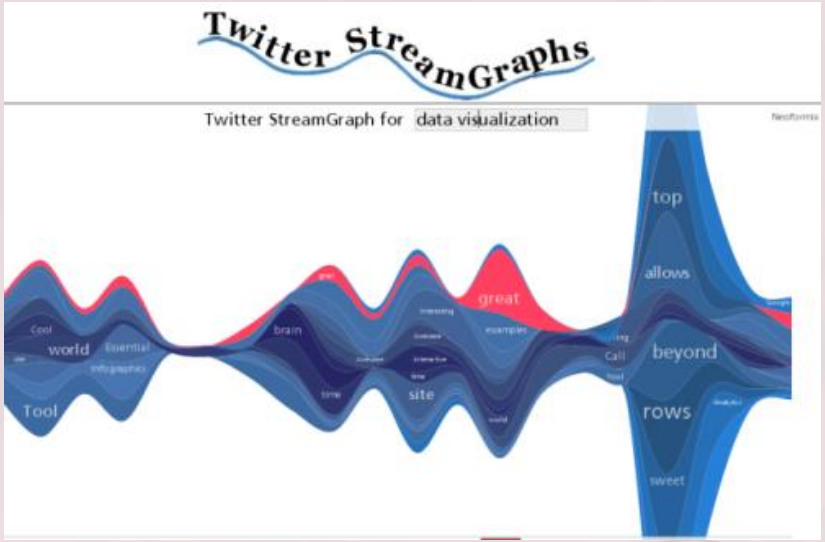
深色的边，表示双向的互相引用关系(reciprocal links, A has cited B and B has cited A)，浅色的边表示单边引用关系(a-reciprocal links)。



博客空间的可视化(Visualization of Blog sphere)

5.可视化的若干实例与特色可视化应用

- Jeff Clark 创建了Twitter Stream Graphs可视化效果。他通过堆叠的河流，显示Twitter数据流里流行的关键字(top trending keywords)随着时间变化的情况。
- 通过这张图，我们可以了解不同的时间段里，Twitter数据流里最流行的一些关键字，以及它们的频率对比。



Twitter Stream Graphs

5.可视化的若干实例与特色可视化应用

- 百度迁徙应用
- 可以让用户观察到全国范围内的迁徙最热路线、迁入和迁出的最热城市和地区。

(a) 百度迁徙

5.可视化的若干实例与特色可视化应用

- 百度通勤图应用(北京版)
- 展现了上班时间、下班时间，白领的通勤路线及其热度。
- 对于公交线路规划，地铁规划，都具有参考意义 。

(c)百度通勤图

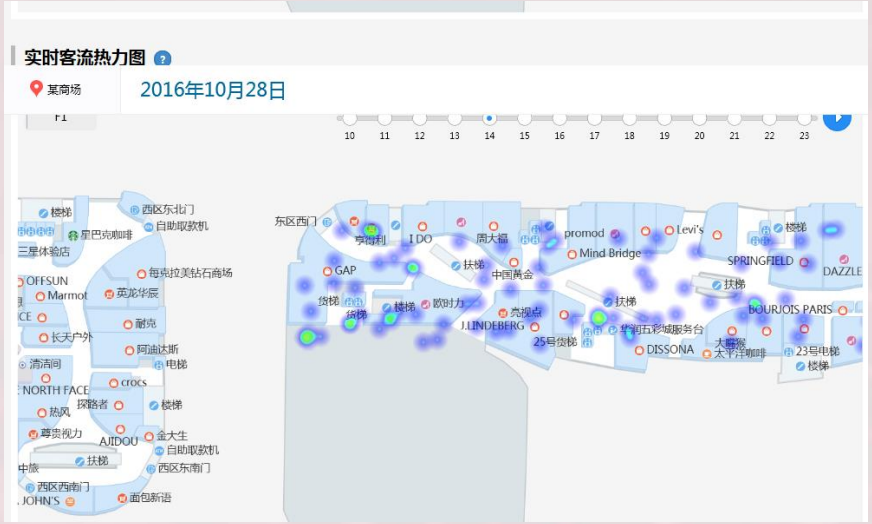
5.可视化的若干实例与特色可视化应用

- 景区热力图应用
- 则让用户实时查看各个热点景区的拥挤程度，提前为出行做出安排。

(d) 百度景区热力图

5.可视化的若干实例与特色可视化应用

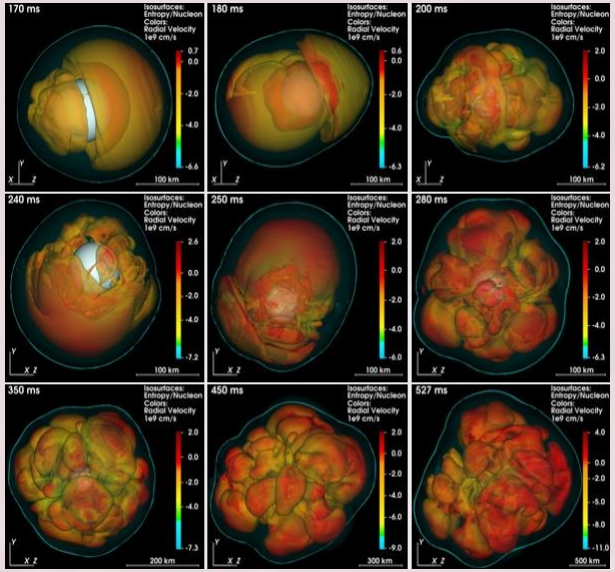
- 百度慧眼则是一项商业服务
- 它是一款位置大数据获取、管理、分析、可视化平台，集成了商业信息、地理信息、人口信息。其中的楼层热力图，对商场的各个楼层的客户流进行实时监控。



(b) 实时客流热力图

6.高维数据可视化

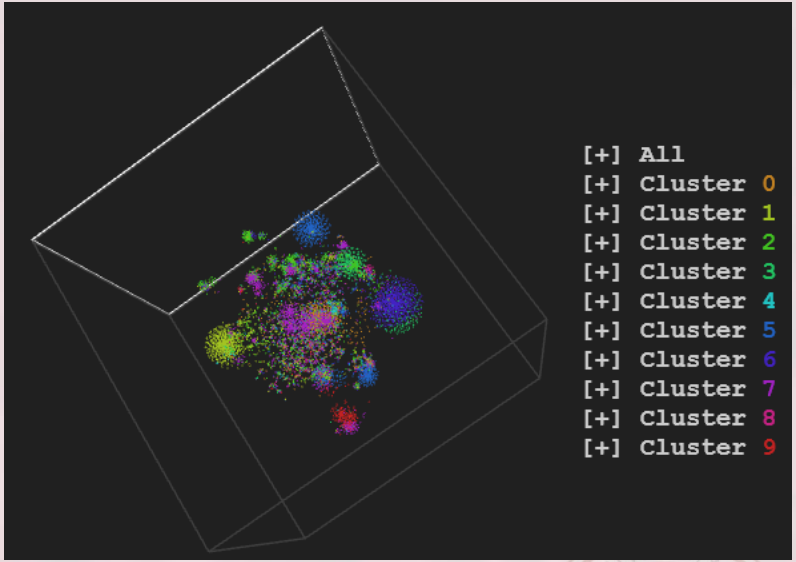
- 如果数据本身的维度为2维或者3维，我们可以方便地设计可视化效果。
- **当数据为4维以上时，而且维度不高的时候，我们可以固定某些维度为某些常数值，然后进行数据的可视化，然后改变这些维度的值，再进行数据的可视化，最后获得一系列的可视化结果。我们通过观察一系列的可视化结果，可以了解事物的发展过程。**
- 高维数据可视化的重要步骤是降维。可以使用的方法包括奇异值分解SVD、多维尺度分析 Multi Dimensional Scaling等方法。
  - 保证高维空间中互相接近的数据点，在低维空间(可视化结果)里也是互相接近的



20个太阳质量的恒星的坍塌过程的可视化

6.高维数据可视化

- 如果数据本身的维度为2维或者3维，我们可以方便地设计可视化效果。
- 当数据为4维以上时而且维度不高的时候，我们可以固定某些维度为某些常数值，然后进行数据的可视化，然后改变这些维度的值，再进行数据的可视化，最后获得一系列的可视化结果。我们通过观察一系列的可视化结果，可以了解事物的发展过程。
- **高维数据可视化的重要步骤是降维。可以使用的方法包括奇异值分解SVD、多维尺度分析 Multi Dimensional Scaling等方法。**
  - **保证高维空间中互相接近的数据点，在低维空间(可视化结果)里也是互相接近的**



根据用户行为(上万维，也就是上万个特征)
对用户进行聚类

7.可视分析简介

- 2005年以来，可视化技术和数据分析技术结合，发展出数据可视化的一个新的分支——可视分析学，它是以可视化交互界面为基础的分析科学。
- 可视分析，综合大量的多模态信息(文本、语音、视频、图像、社交网络等)，并且利用可视化技术以及数据分析技术，帮助人们理解数据。由此可以看出，可视分析是可视化技术、交互技术、分析技术的结合，可视化和交互技术是为分析和推理服务的。

7.可视分析简介

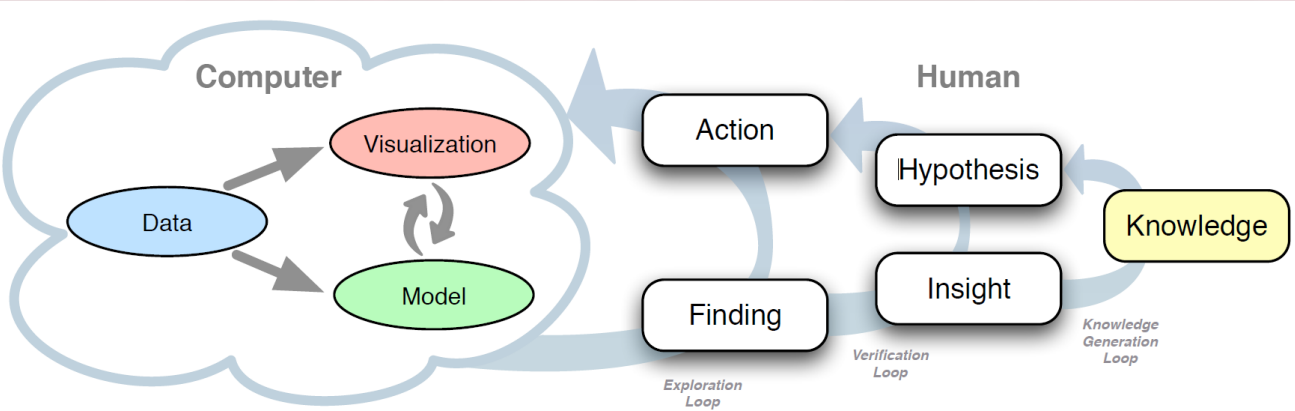
- 可视分析把交互式的可视化过程、以及具体分析技术(包括统计分析方法、数据挖掘方法)结合起来，帮助用户实现高层次的复杂的 (high-level &complex)一系列活动，包括意义构建(sense making)、推理(reasoning)、决策(decision making)。
- 通过推理分析，用户利用其判断能力，基于掌握的证据和假设，获得对数据的理解和洞察力，达成关于数据(数据反映了实际业务)的结论，从而帮助用户完成评估(assessment)和决策(decision)。

7.可视分析简介

- 在可视分析过程中，用户通过如下方式，增强对于数据的认知(cognitive)。
  - (1) 把大量的数据在有限的空间里进行整体展示，使得用户对数据有一个总体的把握和初步的理解。
  - (2) 在时间和空间维度上展示数据的变化，帮助用户对数据的模式(pattern)进行感知和认知。
  - (3) 把复杂的网络关系以可视化的方式展示出来，帮助用户基于感知，进行关系推理(inference of relationships)。
  - (4) 通过交互式分析的方式，使得用户可以调整参数值(parameter values)，通过及时改变的可视化结果，对数据进行探索。

7.可视分析简介

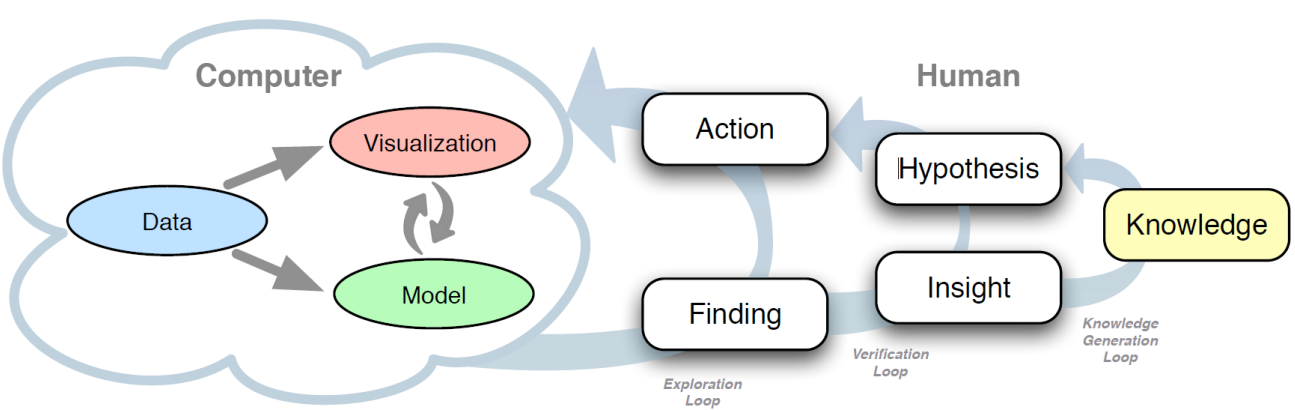
- 可视分析是一个迭代的过程，如图所示。可视分析引导用户进入分析流程，让用户可以通过交互式的可视化界面，将其经验和智能输入到系统中，不断发现规律，建立假设，然后肯定或者否定假设。



可视分析的过程(从数据到模型到知识)

7.可视分析简介

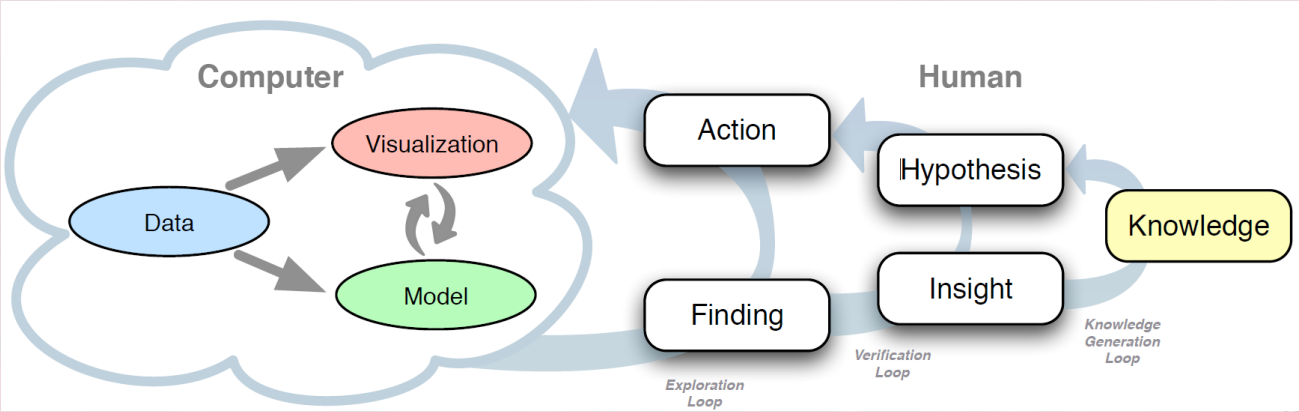
- 可视分析包括三个要素，分别是
  - 数据(Data)，数据是所有可视化分析的基础 。
  - 模型(Model)，包括统计模型、以及机器学习、数据挖掘模型。
  - 利用可视化(Visualization)，探测(Detect)数据中变量之间的关系(relationships) 。



可视分析的过程(从数据到模型到知识)

7.可视分析简介

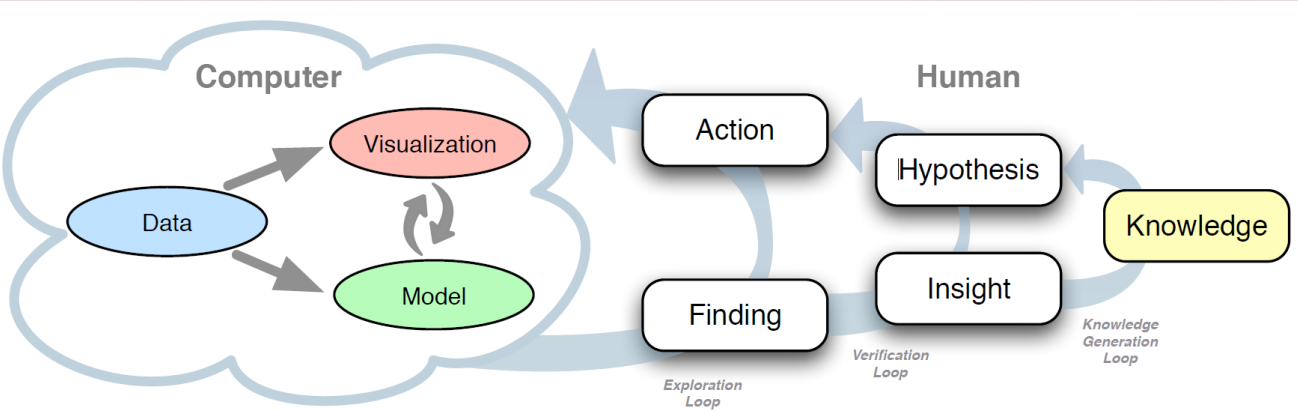
- 可视分析包括三个回路，第1个是探索回路(Exploration Loop)
  - 发现，指的是分析者使用可视化分析系统，获得的一个有趣的观察结果(interesting observation)。
  - 动作，操控可视化效果，改变观察角度，加深对数据的理解。



可视分析的过程(从数据到模型到知识)

7.可视分析简介

- 可视分析包括三个回路，第2个是验证回路(Verification Loop)，第3个是与产生新知识
  - 洞察，对发现进行理解和解释。
  - 假设，是针对问题领域构造了一个假设，以便后续进行验证性的分析。
  - 在可视化分析过程中，分析者为某个假设，寻找证据，或者从数据中学习到了新的知识。从证据到知识，需要一个推理(reasoning)的过程。



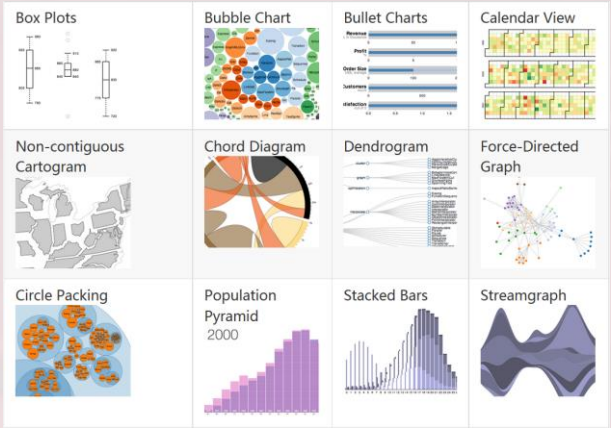
可视分析的过程(从数据到模型到知识)

8.可视化的挑战和趋势

- 海量的异构数据的可视化，对算法设计和硬件基础设施，都提出了更高的要求。
  - 高维的(High Dimensional)、多元的(Multivariate)、多模态的(Multimodalities)、时变的(Time Varying)数据，以及数据不完整(Incomplete Data)，数据里的噪音(Noise)等特点，都给数据的可视化提出了严峻的挑战 。
- 在大数据时代，有几个可视化技术的发展趋势值得注意。
  - (1) 各种新硬件被应用到可视化领域(New Hardware)，可视化系统将支持更高的显示分辨率(Higher Resolution)。(2) 可视化技术被应用到更多的业务领域(More Application Domains)。(3) 可视化技术支持更多样的数据的可视化(More Data Types)。(4) 新的研究热点，是基于可视化、以及可视化分析结果，进行叙事，讲一个故事(Telling a Story)，并且把故事讲完整、讲精彩。(5) 可视化软件提供更加强大的可视分析能力(Advanced Visual Analytics)。

8.可视化工具

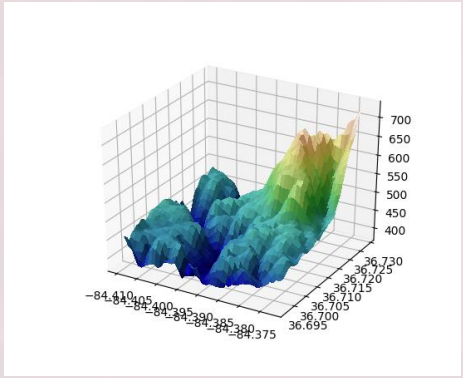
- 接下来，介绍可视化工具
- D3.js
  - D3是一个开源项目，其作者是纽约时报的工程师。目前，D3项目的代码托管于GitHub (https://github.com/d3/d3/wiki)。
  - D3支持大量的可视化效果。

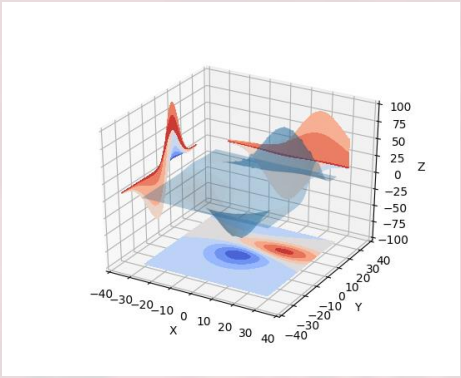


D3.js的可视化实例

8.可视化工具

- **Matplotlib**
  - matplotlib是基于python语言的图形绘制库(plotting library)。使用matplotlib，用户编写少量代码，就可以创建各种常用的图形，比如直方图、散点图、饼图、柱状图。这些图形达到印刷的品质，用户可以导出成PDF文档或者PNG图像文件。

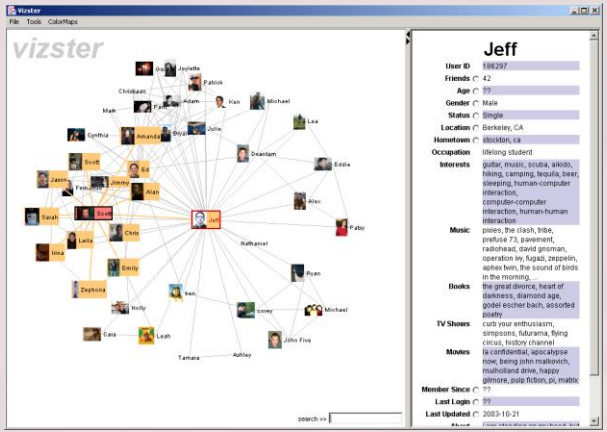




matplotlib的可视化实例

8.可视化工具

- Prefuse
  - Prefuse(http://prefuse.org/)是一款使用Java语言开发的开源的(BSD license)软件工具，用于开发交互式的数据可视化程序。



基于Prefuse的社交网络可视化

回顾

√1. 可视化的定义及其意义
√2. 可视化的一般过程
√3. 科学可视化与信息可视化
√4. 可视化的若干原则
√5. 可视化的若干实例与特色可视化应用
√6. 高维数据可视化
√7. 可视分析
√8. 可视化的挑战和趋势
√9. 可视化工具介绍

结束

The End!