第11章 复杂高维多元数据的可视化

挑战

- 呈现、理解和应用海量复杂数据是数据可视化与分析面临的挑战
  - 对于高维多元数据，以统计和基本分析为主的可视化系统分析能力不足
  - 数据复杂度大大增加。包括非结构化数据和从多个数据源采集、整合而成的异构数据，传统单一的可视化方法无法支持对此类复杂数据的分析
  - 数据的规模大，超越一般计算处理能力的极限，需要采用全新思路来解决数据大尺度的挑战
  - 数据获取和处理中的数据质量问题，其中特别的，数据的不确定性
  - 数据快速动态变化，常以流式数据形式存在。对流数据的实时分析与可视化是急需解决的问题

1 高维多元数据

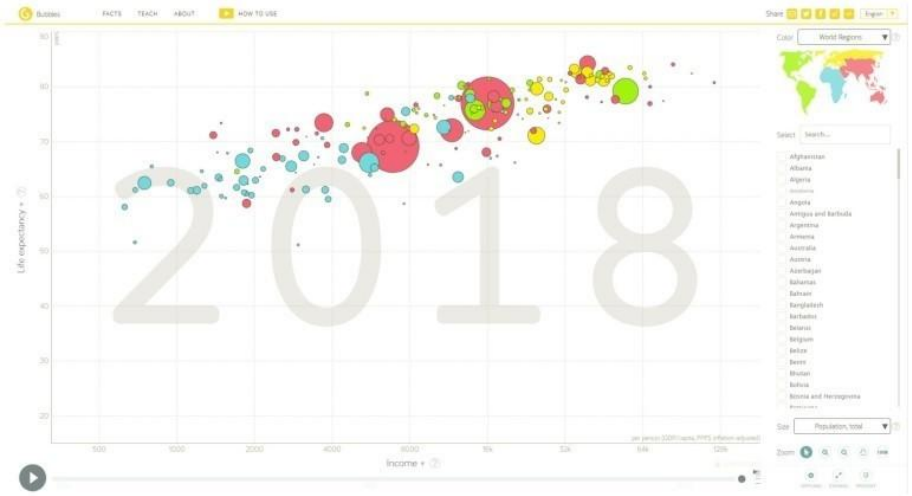
高维多元数据(Multidimensional Multivariate Data)

- 高维多元数据，指每个数据对象有两个或两个以上独立或者相关属性的数据
  - 高维(Multidimensional)，数据具有多个独立属性
  - 多元(Multivariate)，数据具有多个相关属性
  - 当数据同时具有独立和相关属性时，高维多元数据是较为科学、准确的描述

- 用多元数据指代所有的高维多元数据，用维度指代数据属性的数量

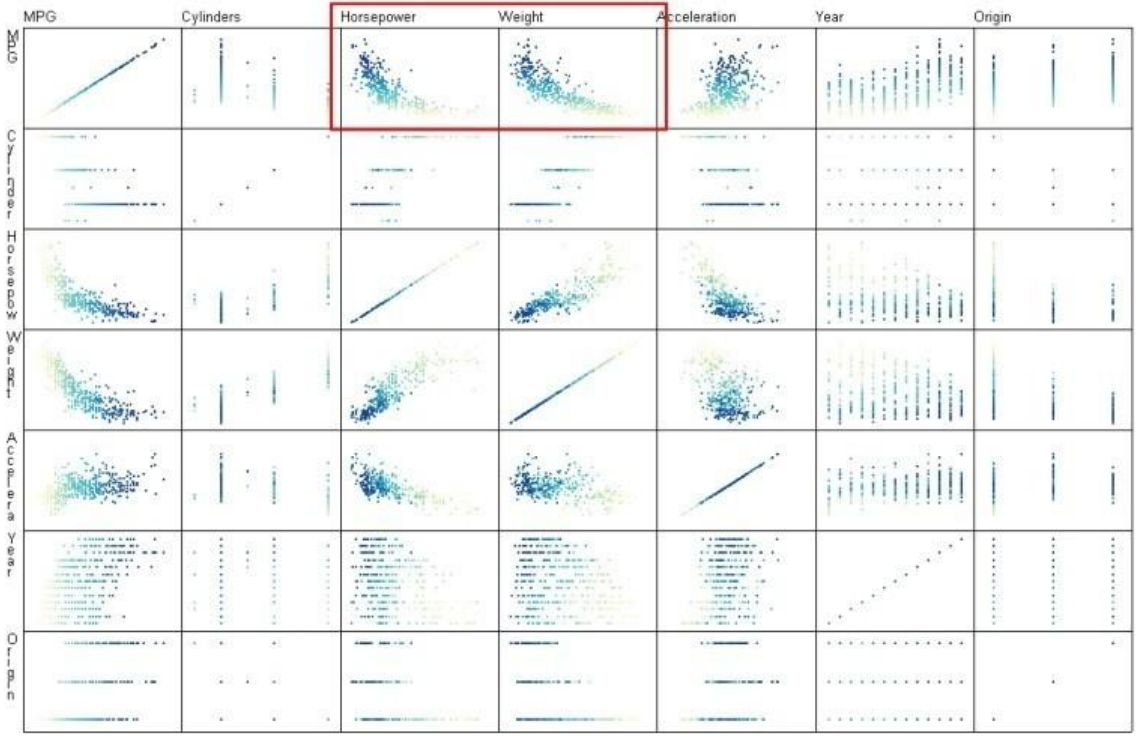
散点图(Scatter Plot)

- 散点图可视化
  - 将各个属性的值映射到不同的坐标轴，并确定各数据点在坐标系中的位置
- 维度超过三维时，可通过视觉编码（颜色、大小、形状等）来表示额外的属性。但
  - 首先，视觉编码的种类有限
  - 其次，过多或者过于复杂的视觉编码会降低可视化的可读性



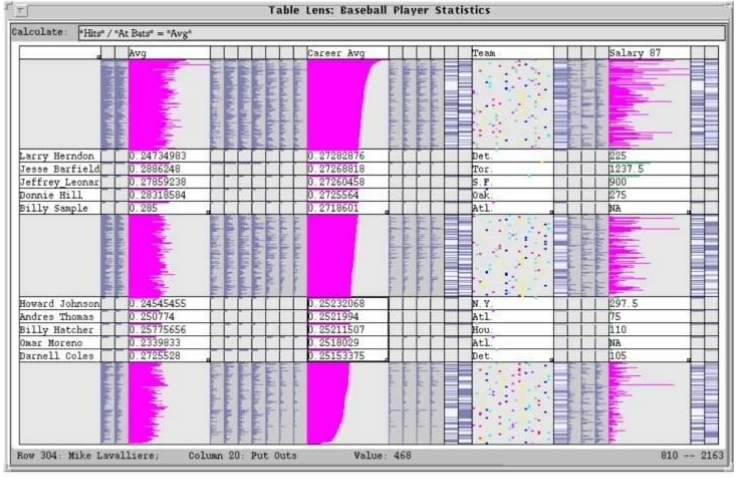
散点图及散点图矩阵

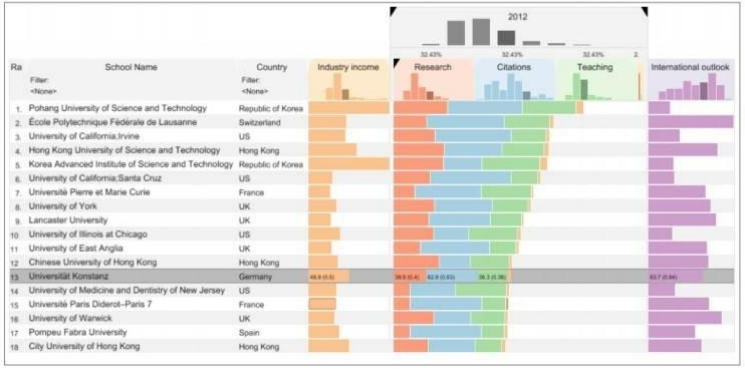
- 散点图矩阵是散点图的扩展
  - 在有限的屏幕空间中显示过多散点图大大降低可视化的可读性
  - 交互式选取感兴趣属性进行可视化和分析



表格透镜

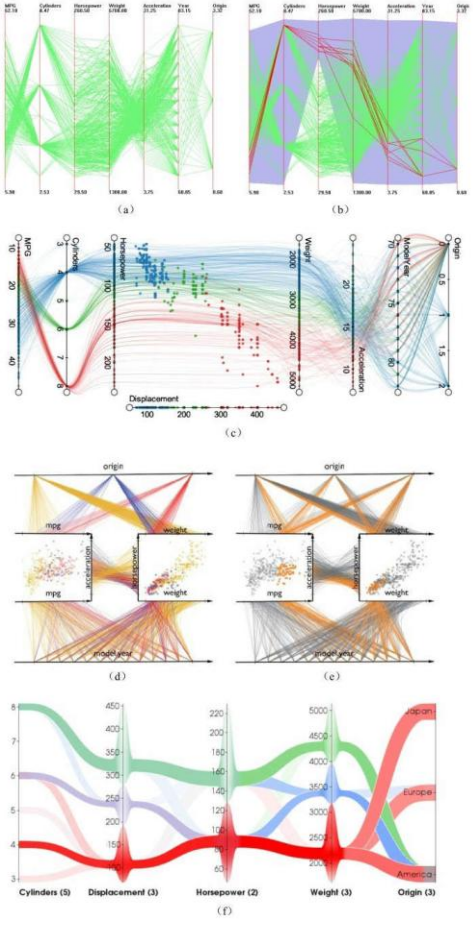
- 表格透镜
- 每个数据对象由一行表示，每列表示一个属性
- 数据在每个维度上的值用水平横条或者点表示
  - 在有限的屏幕空间中表示大量的数据和属性，同时方便用户对数据对象和各个属性进行快速比较





平行坐标(Parallel Coordinates)

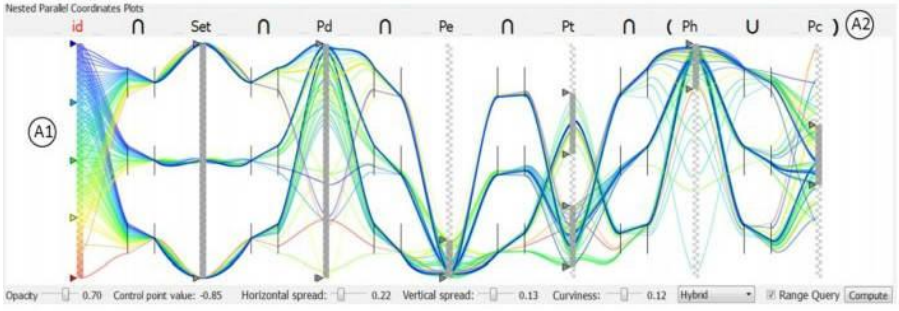
- 平行坐标方法采用相互平行的坐标轴，每个坐标轴代表数据的一个属性，每个数据对象对应一条穿过所有坐标轴的折线



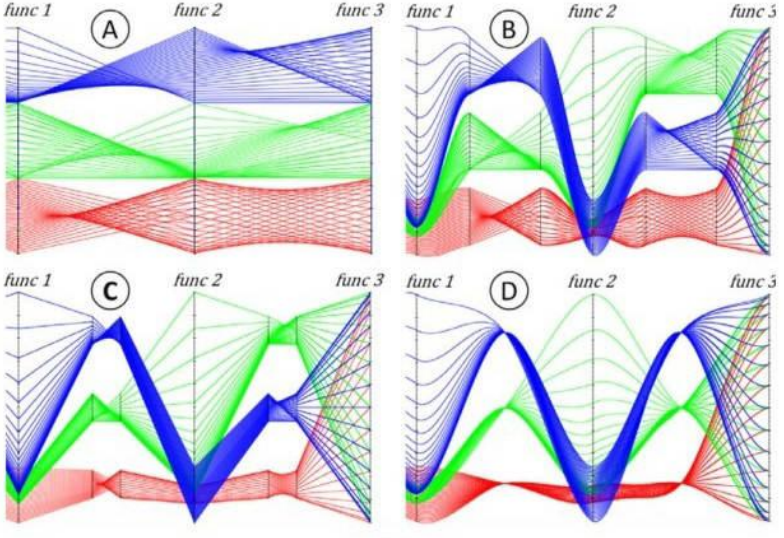
平行坐标

- 由于平行坐标的坐标轴是顺序排列的，对于非相邻属性之间关系的表现相对较弱，不易于同时表现多个维度之间的关系。交互地选取部分感兴趣的数据对象，并用高亮显示是一种常见的解决方法
- 为了应对不同尺度、不同参数计算获得的集合数据(Ensemble Data)，嵌套平行坐标在主要轴之间放置次要轴，展现不同参数下的运行结果

➢ 嵌套平行坐标

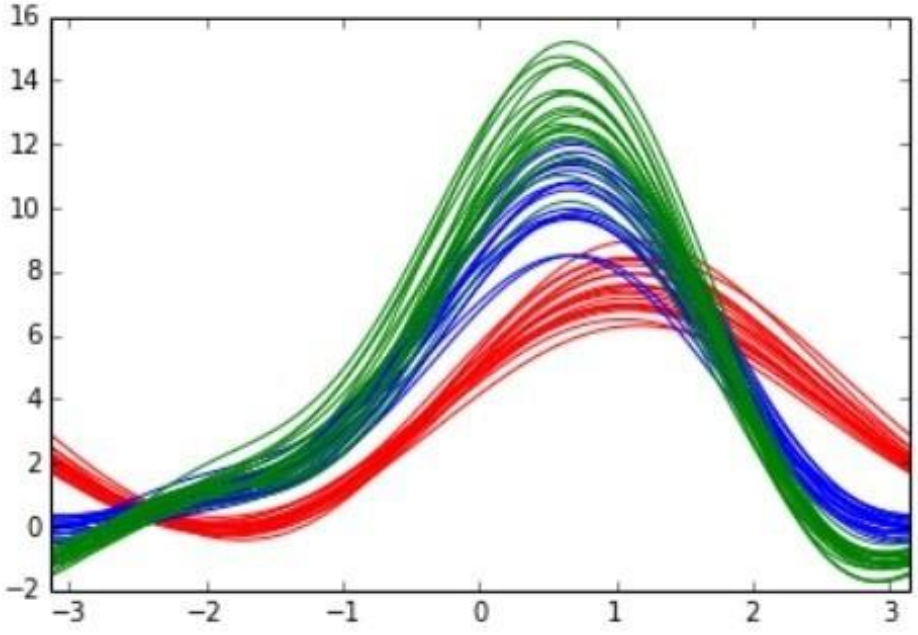

Nested Parallel Coordinates Plots
id
∩
Set
∩
Pd
∩
Pe
∩
Pt
∩
(Ph
∪
Pc)
A2
A1
Opacity
0.70
Control point value: -0.85
Horizontal spread:
0.22
Vertical spread:
0.13
Curviness:
0.12
Hybrid
Range Query
Compute

(a)


func 1
A
func 2
func 3
func 1
B
func 2
func 3
func 1
C
func 2
func 3
func 1
D
func 2
func 3

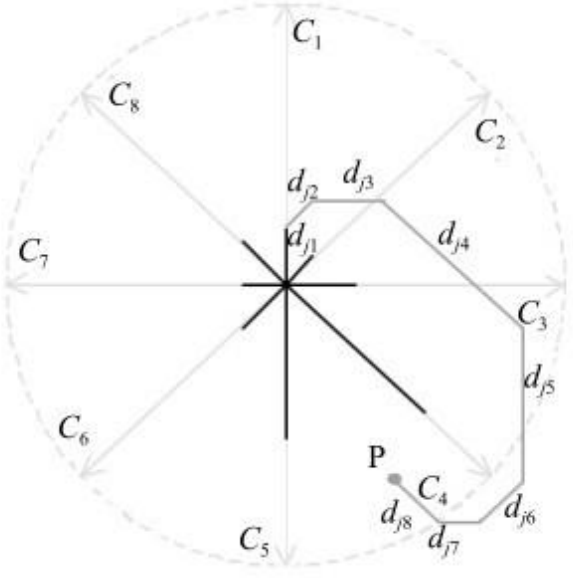
(b)

安德鲁斯曲线



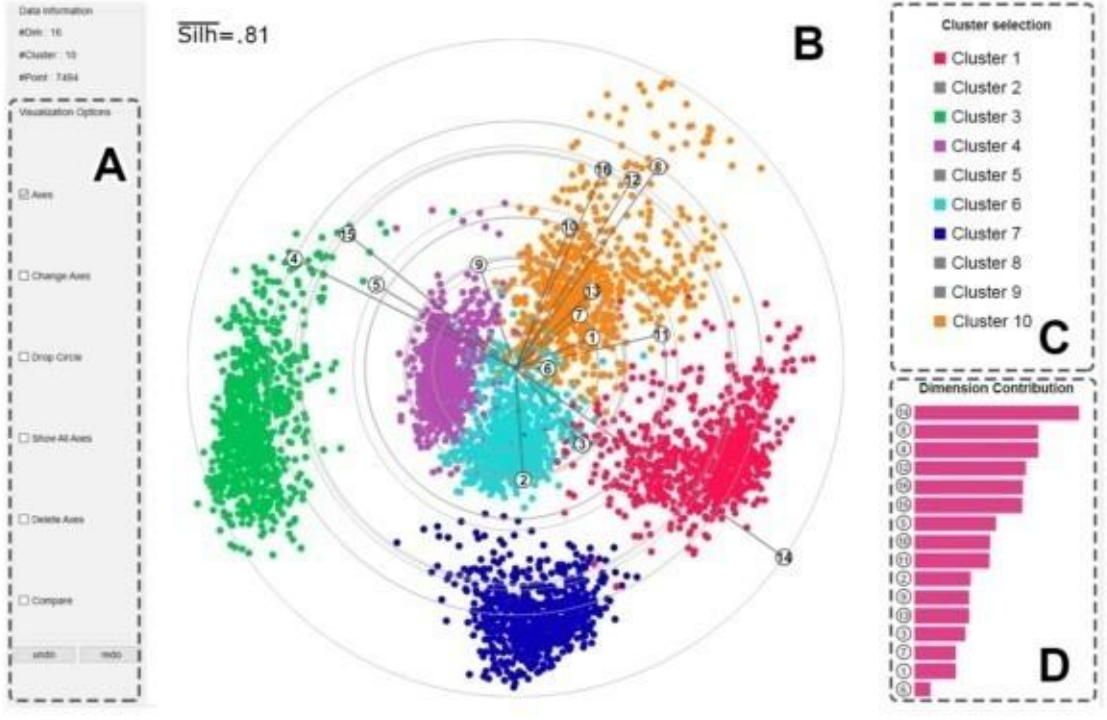
径向布局

- 星形坐标(Star Coordinates)
- 假设二维平面上有一个点作为原点$O_n(x,y)=(O_x,O_y)$,以及一系列的二维向量作为轴。$A_n=\langle\vec{a}_1,\vec{a}_2,\cdots,\vec{a}_i,\cdots,\vec{a}_n\rangle$。数据集D中的一个数据点$D_j(d_{j0},d_{jl},\cdots,d_{ji},\cdots,d_{jn})$映射到二维平面上的结果$P_j(x,y)$是所有轴的单位向量$\vec{u}_i=(u_{xi},u_{yi})$乘以这个数据点在该位置值$d_{11}$以后的向量之和
- 令$\min_i=\min\{d_{ji},0\leqslant j<|\boldsymbol{D}|\},\max_i=\max\{d_{ji},0\leqslant j<|\boldsymbol{D}|\}$,它们分别代表数据集里第$i$维的最小值和最大值
- $P_j(x,y)=\left(ox+\sum_{i=1}^{n}u_{xi}\cdot\left(d_{ji}-\min_i\right),oy+\sum_{i=1}^{n}u_{yi}\cdot\left(d_{ji}-\min_i\right)\right)$



星形坐标

- 星形坐标支持两种交互操作其中一种是轴的伸缩；另一种是旋转。通过这些交互，能让用户看到哪些属性(轴)对数据的聚类有更多或更小的影响，也能合并或拆分不同的聚类



降维

- 数据维度很高时，可视方法无法清晰表示所有细节
  - 通过变换将数据投影(Project)或嵌入(Embed)至低维空间(通常二维或三维)，并尽量在低维空间保持数据在多元空间中的关系或特征。该类策略称为降维(Dimension Reduction)
  - 线性方法
    - 主元分析(Principal Component Analysis,PCA)
    - 多维尺度分析(Multidimensional Scaling,MDS)
    - 线性判别分析(Linear Discriminant Analysis,LDA)
  - 非线性方法
    - 局部线性嵌入(Locally Linear Embedding,LLE)
    - Isomap
    - SNE和t-SNE

主元分析(Principal Components Analysis, PCA)

目标：在减少数据集维度的同时，保持对数据集方差贡献最大的特征

PCA定义一个正交线性变换，将数据变换到新的低维坐标系统，使得数据投影的第一大方差在第一个坐标上，第二大方差在第二个坐标上。

假定X_…X_N是N个多元数据，其维度大小为M。PCA算法的基本计算：

(1)重组数据。将给定的N个数据组合成M × N的矩阵X：每一列表示一个多元数据，每一行代表一个数据维度

(2)计算每个数据属性的均值，得到一个大小为M × 1的均值向量$\boldsymbol{u}$:$\boldsymbol{u}_i = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{X}_{ij}$

(3)将样本数据中心化，以保证在每个属性上的偏移都以0为基点

(4)计算B的协方差矩阵$C = \frac{1}{N}\sum BB^{\mathrm{T}}$

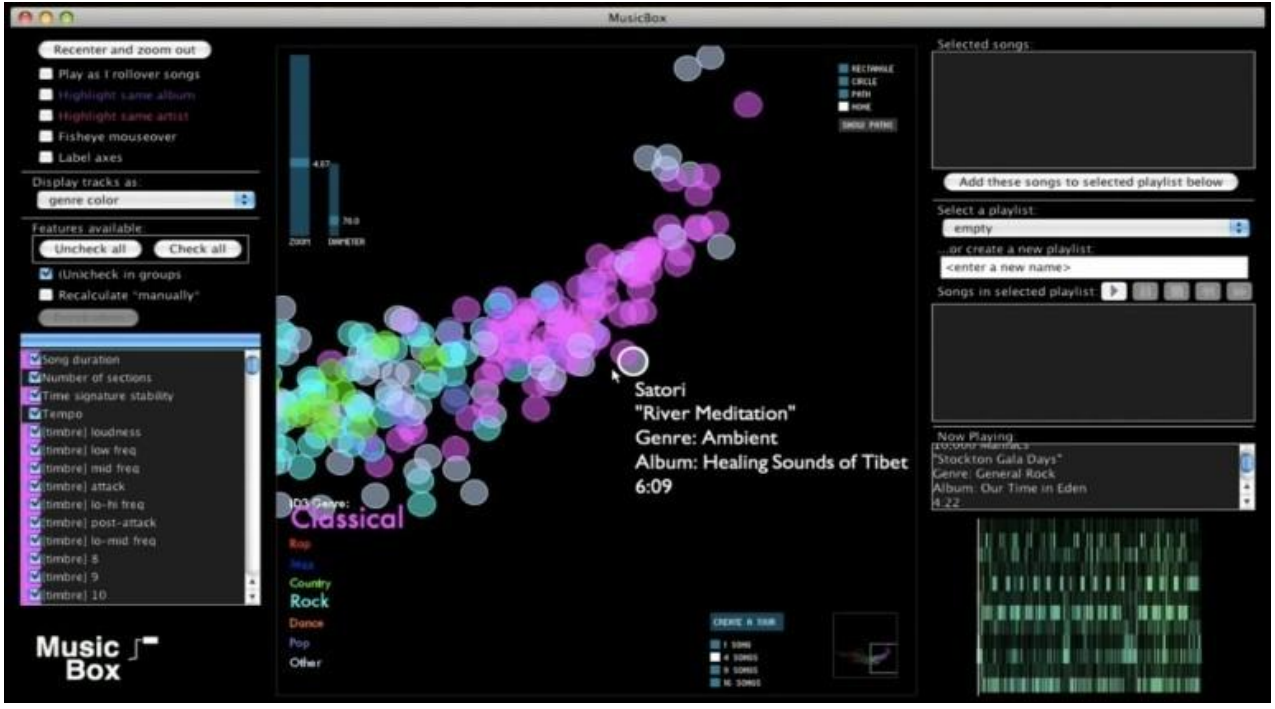
(5)特征分解:C = QΛQ^{-1},其中，Q是由特征向量组成的方阵，Λ是对角矩阵，对角线上的元素为对应的特征值

(6)选择最大的k个特征值，将其对应的k个特征向量分别作为列向量，组成特征向量大小为M × k 的矩阵L。k是目标低维空间的维数，通常为2或者3

(7)将多元数据点投影到选取的特征向量上，得到在低维空间中的投影

主元分析

- 音乐可视化软件MusicBox
  - 将每首音乐看成一个多元数据
  - PCA将其投影至二维平面。每个点代表一首音乐



多维尺度分析（Multidimensional Scaling,MDS）

1）对于给定的包含M个数据属性和N个数据记录的数据集，计算数据记录之间的相似性，得到一个包含NXN个元素的相似度矩阵D。矩阵D刻画了数据点相互之间的相似性

通常可采用欧式距离、Cosine距离等计算多元数据之间的相似性

2）假定目标降维空间的维度为K，通常K为2或者3，创建大小为NXK的矩阵L’，并初始化。L’的每一行表示多元空间中的数据点在低维空间中的位置

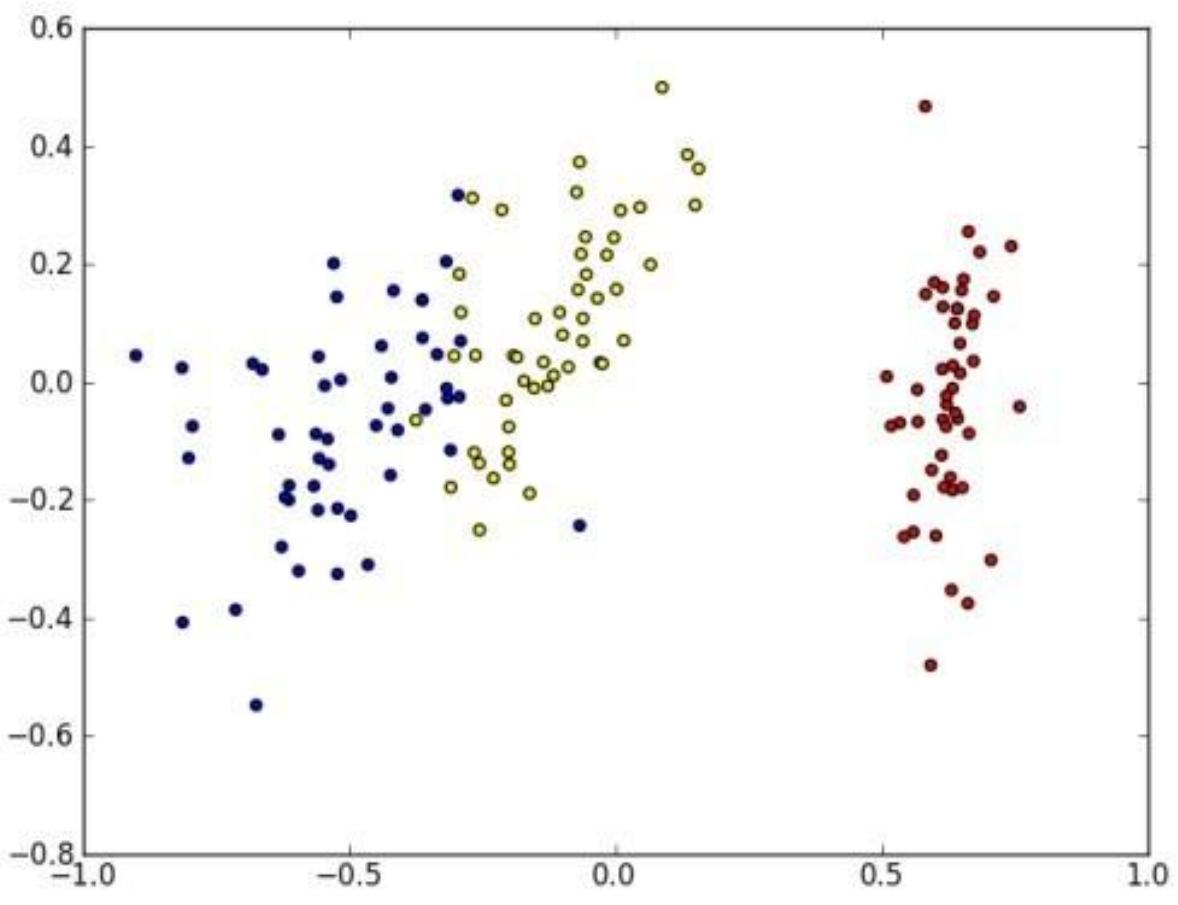
3）计算低维空间中数据点之间的相似度，得到一个大小为NXN的相似度矩阵L

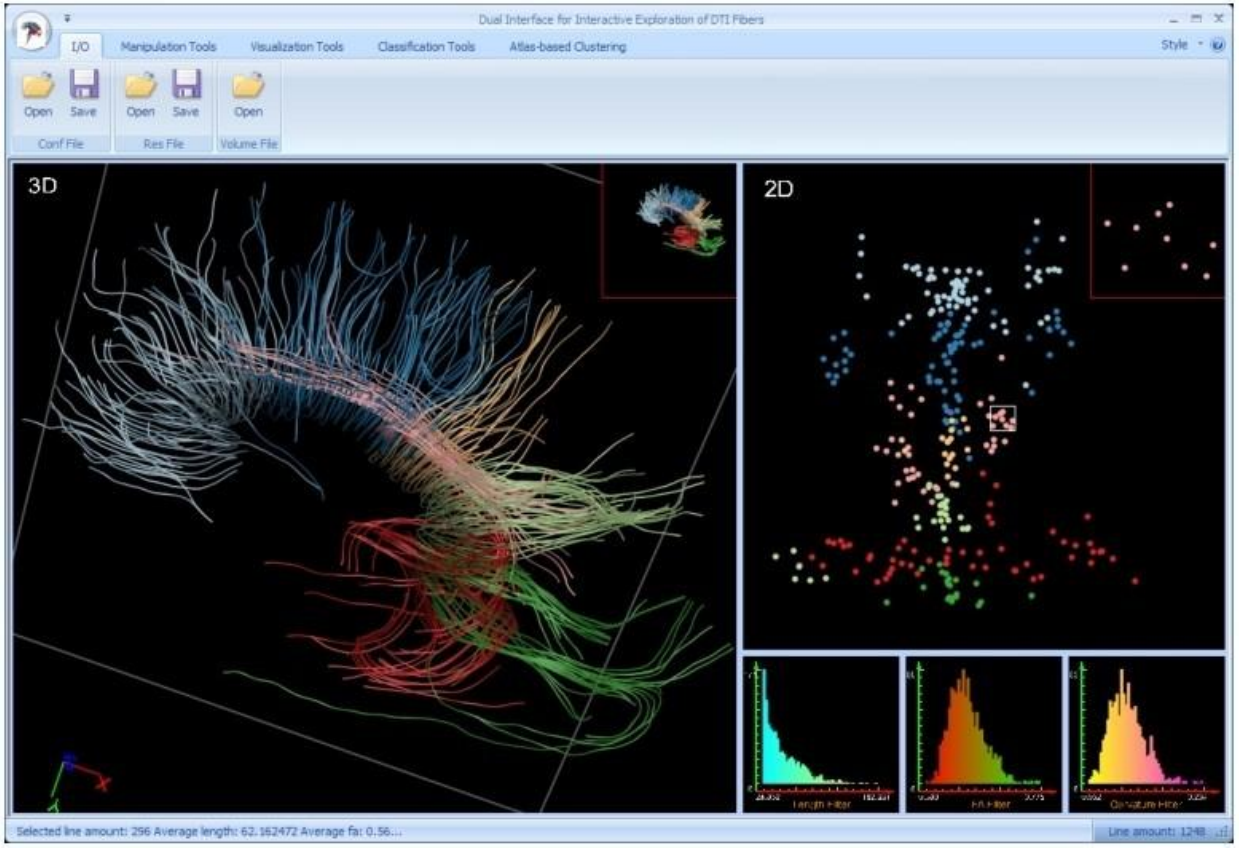
4）计算应力值:$S = \sqrt{\frac{\sum_{ij}\left(D_{ij}-L_{ij}\right)^2}{\sum_{ij} L_{ij}^2}}$ S描述了D与L之间的差异性大小

5）如果S足够小或者收玫，则退出算法

6）否则，沿某个方向，移动低维空间中的点使得S变小

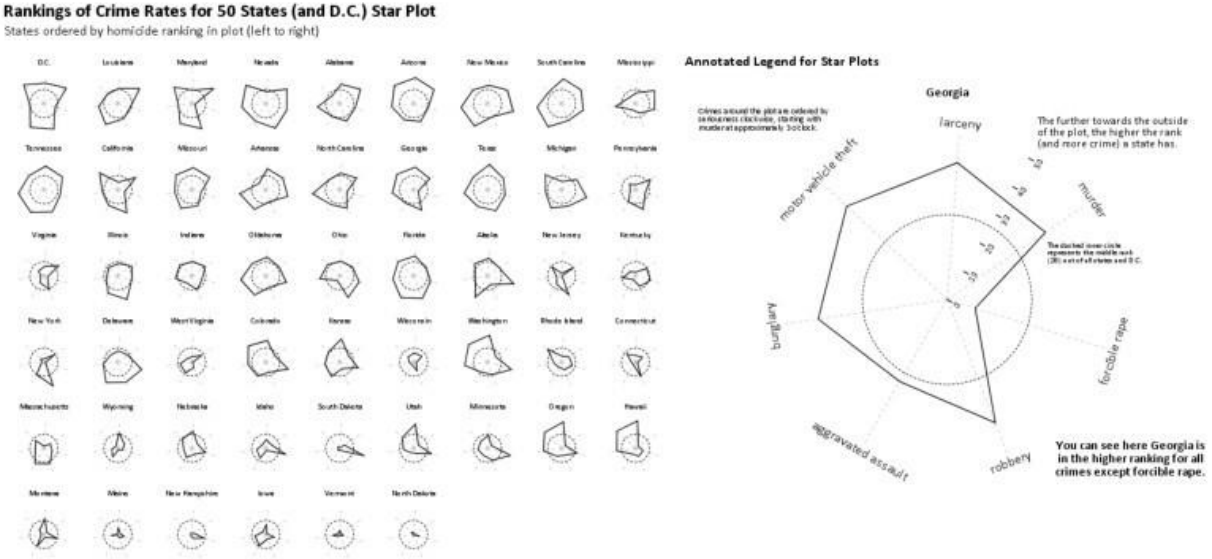
7）返回第3步


0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−0.5
0.0
0.5
1.0


Dual Interface for Interactive Exploration of DTI Fibers
I/O
Manipulation Tools
Visualization Tools
Classification Tools
Atlas-based Clustering
Style
Open
Save
Open
Save
Open
Conf File
Res File
Volume File
3D
2D
Selected line amount: 296 Average length: 62.162472 Average fa: 0.56...
Line amount: 1248

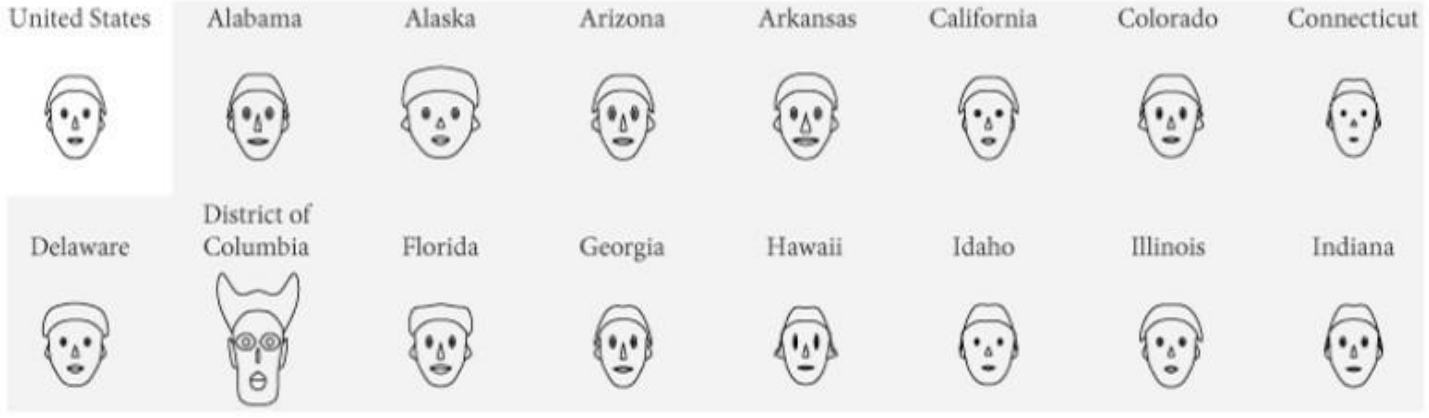
图标法

- 图标（Glyph）表达单个多元数据对象，图标中的不同视觉元素被用来表示数据对象的不同属性。典型的代表有雷达图、Chernoff Faces和DICON
- 雷达图（Radar Chart），又称为星形图（Star Plots）。雷达图可以看成平行坐标的极坐标版本



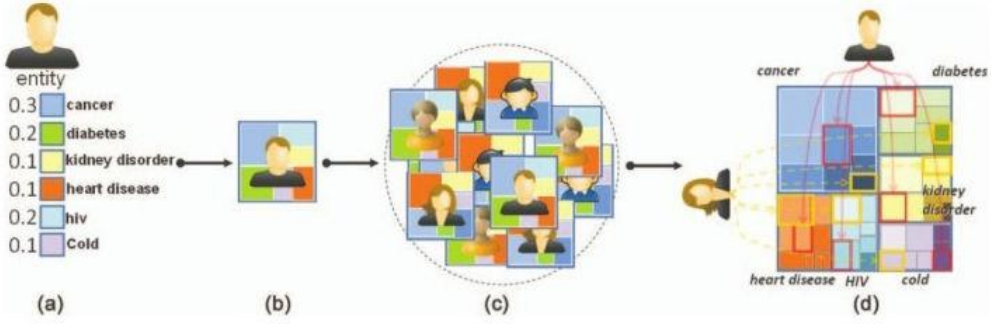
Chernoff Faces

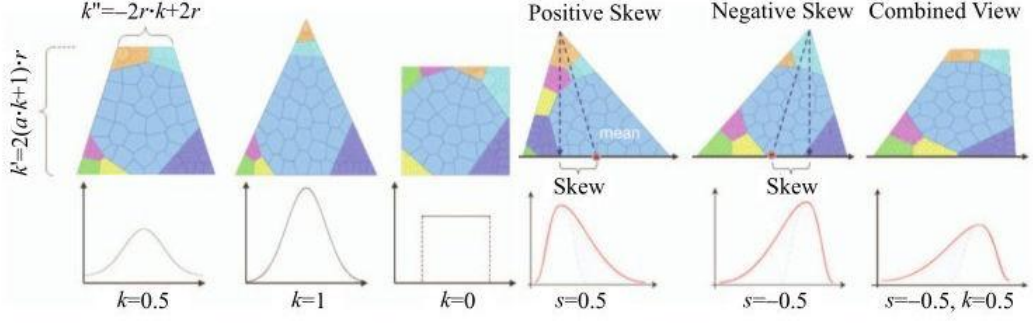
- 用模拟人脸的图标来表示数据对象，不同的属性映射为人脸的不同部位和结构，如脸的大小、眼睛的大小等



DICON

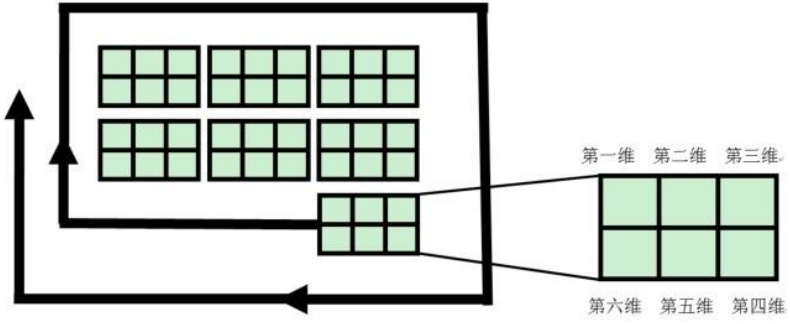
- DICON，一种动态的基于图标的可视化技术，帮助用户理解、评估和调整复杂的高维数据的聚类结果

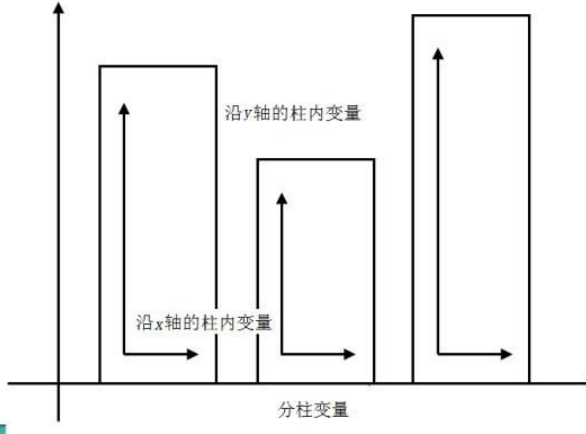


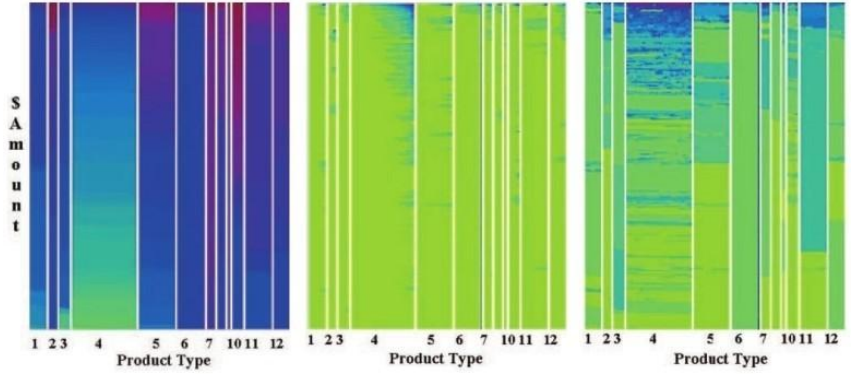


基于像素图的方法

- 利用密集型的不同颜色的像素显示表达存储在大规模数据库中的多元数据

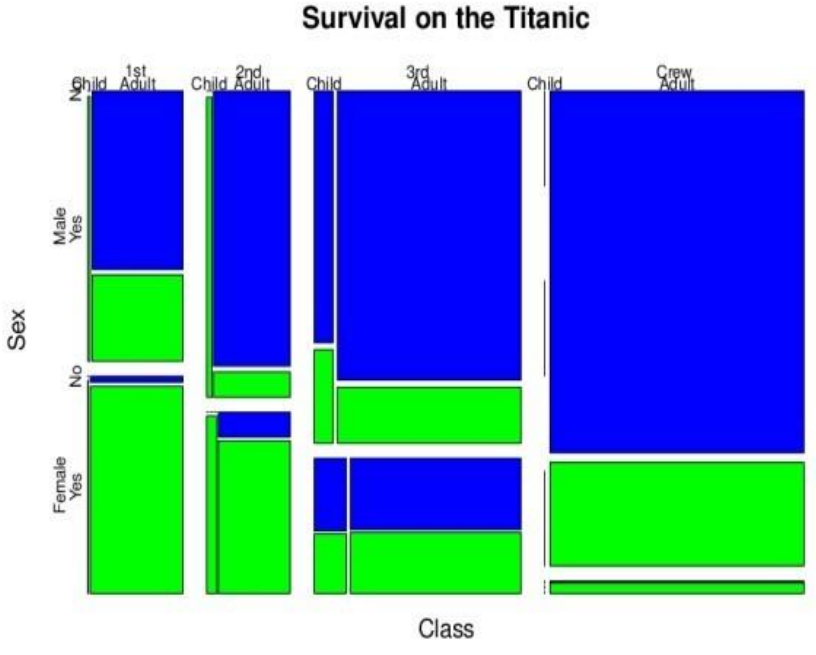






马赛克图（Mosaic Plot）

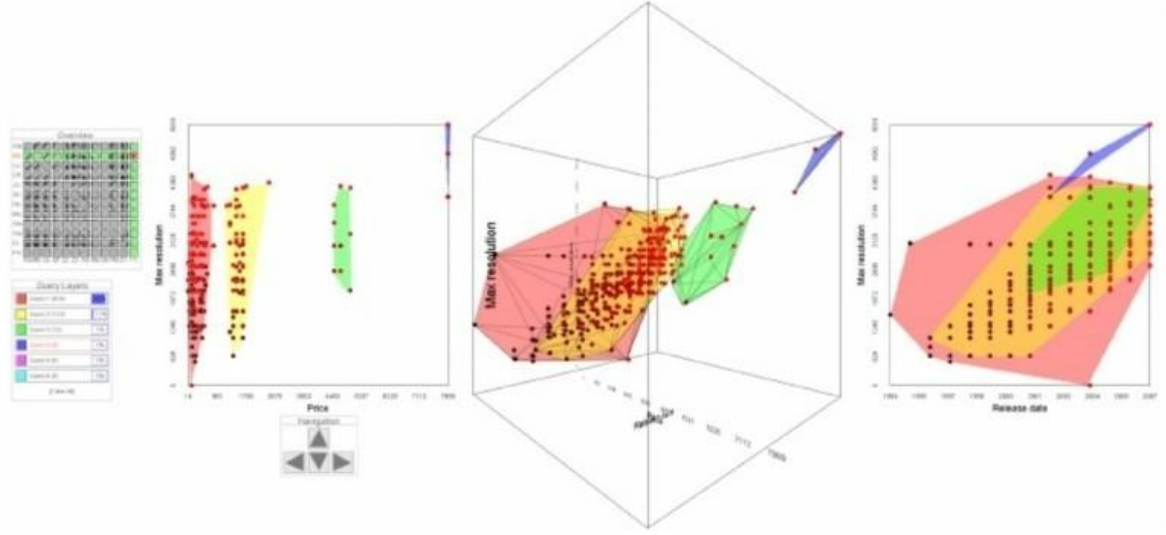
- 马赛克图通过空间剖分的方法展示多元类别型数据的统计信息
  - 从一个单位长度的二维空间（通常为长方形）出发，依据每一个数据维度以x、y轴的次序递归地将二维空间进行层次的剖分。



成人	获救者		未获救者		儿童	获救者		非获救者	
	男性	女性	男性	女性		男性	女性	男性	女性
一等舱	57	140	118	4	一等舱	5	1	0	0
二等舱	14	80	154	13	二等舱	11	13	0	0
三等舱	75	76	387	89	三等舱	13	14	35	17
船员	192	20	670	3	船员	0	0	0	0

基于动画的方法

- 动画过渡效果可用于增强用户对数据点和不同数据可视化图表之间关系的感知
- ScatterDice利用3D的骰子结合散点图矩阵：在散点图矩阵上上下或左右移动一格，其总有一个不变的维度和两个变化的维度，两个变化维度的转移过程用骰子在立体空间中的转动做隐喻



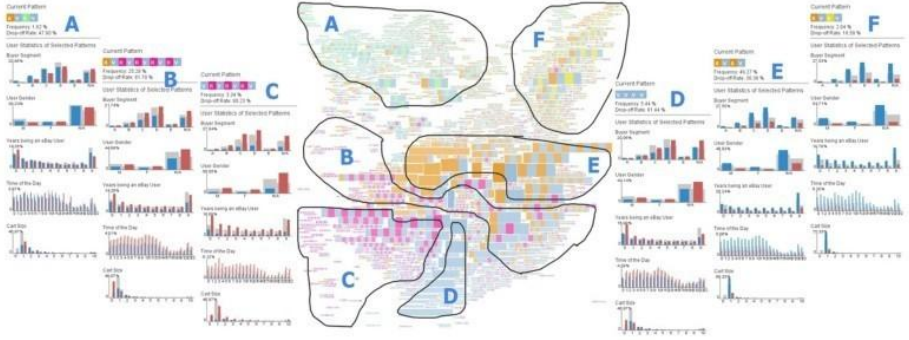
2 非结构化与异构数据的可视化

非结构化数据

- 结构化数据（文本、时间、日志等）无法采用N维表表示

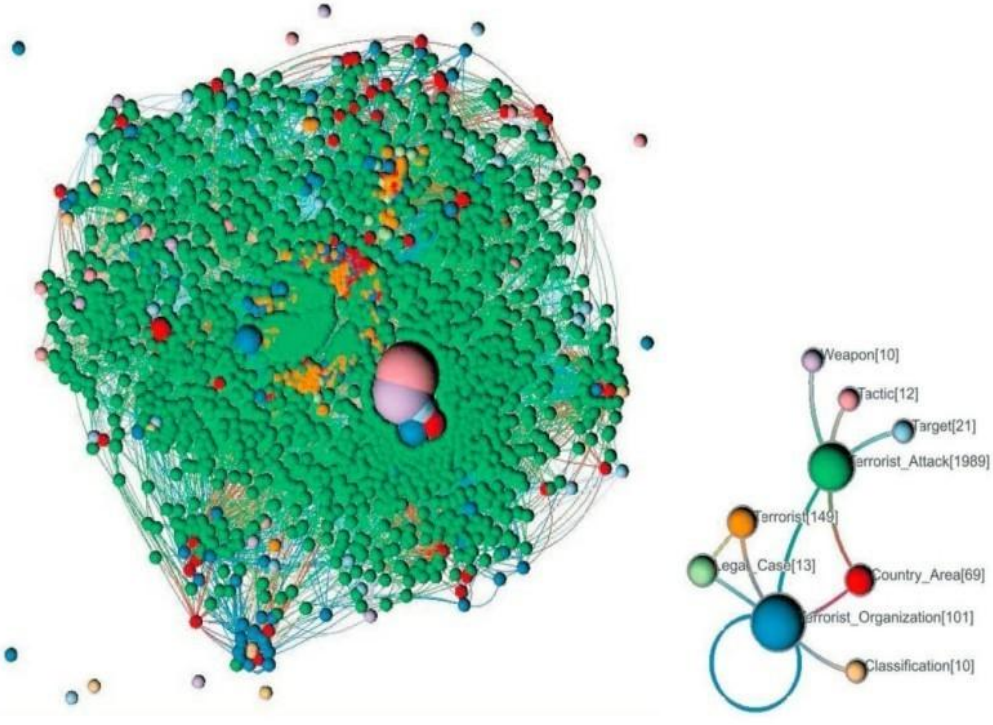
分析日志数据

- 帮助了解用户如何使用目标网站、他们的典型浏览行为等
- 帮助开发者和设计师针对性地改进用户体验，建立用户行为模型，提供更优化和定制化的体验

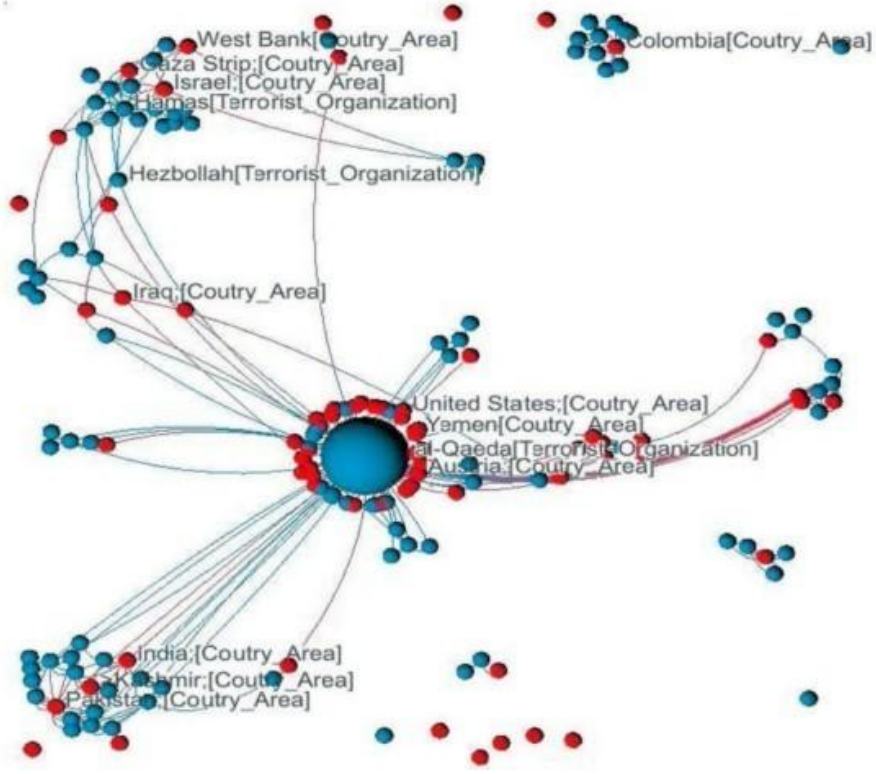


异构数据

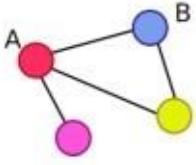
- 异构数据，指同一个数据集中存在结构或者属性不同的数据的情况。存在多种不同类别的节点和连接的网络被称为异构网络
- 异构数据通常可采用网络结构进行表达

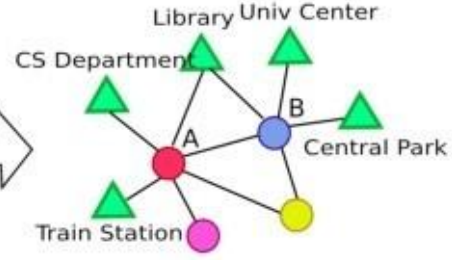


MobiVis


West Bank[Coutry_Area]
Israel;[Coutry_Area]
Hamas[Terrorist_Organization]
Hezbollah[Terrorist_Organization]
Iraq;[Coutry_Area]
United States;[Coutry_Area]
Yemen[Coutry_Area]
Austria;[Coutry_Area]
India;[Coutry_Area]


Central Park
Univ Center
Train Station
Library
CS Department

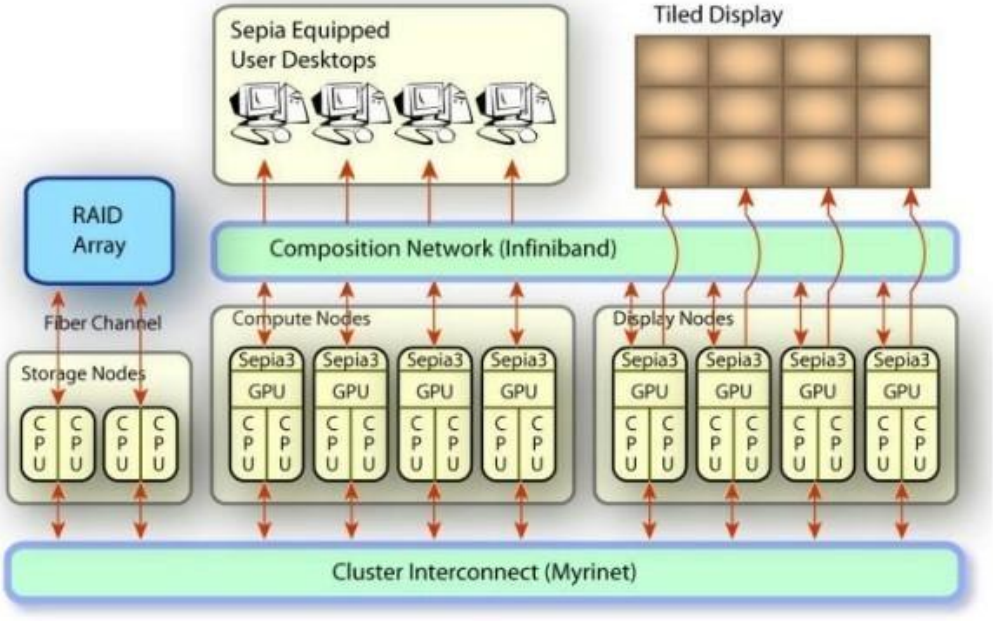

A
B

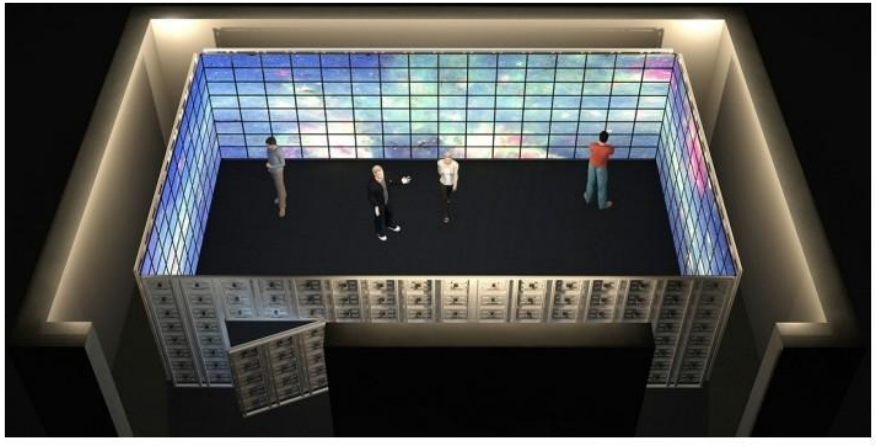

Library
Univ Center
CS Department
B
A
Central Park
Train Station

3 大尺度数据的可视化

基于并行的大尺度数据高分辨率可视化

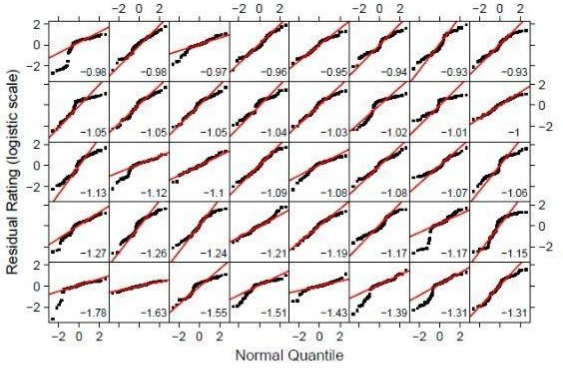
- 全方位显示大尺度数据的所有细节是一个计算密集型的过程
- 大规模计算集群（如分布式多核计算集群、GPU+CPU混合架构集群等）

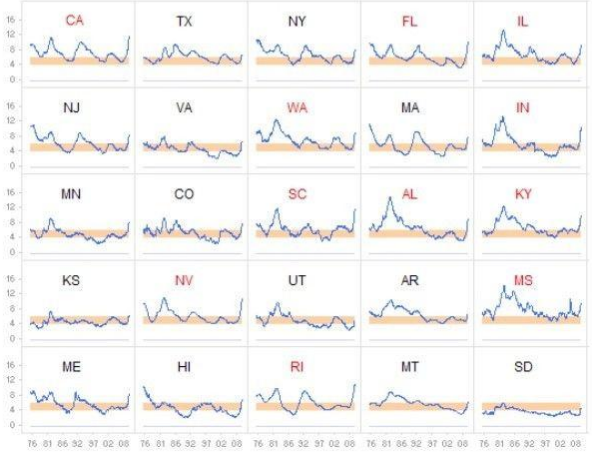




大尺度数据高分辨率可视化

- 采用下采样（Downsampling）的方法，将高精度数据采样为低分辨率，进而在给定分辨率的视图中实现预览式（Preview）可视化。
- 采用层次结构重新组织大尺度数据，并结合多种用户交互方法（如层次细节、聚焦+上下文）实现单一视角下的自适应分辨率选择或多个视角的光滑切换。然而，数据的高分辨率表示增加了对存储空间的额外需求。



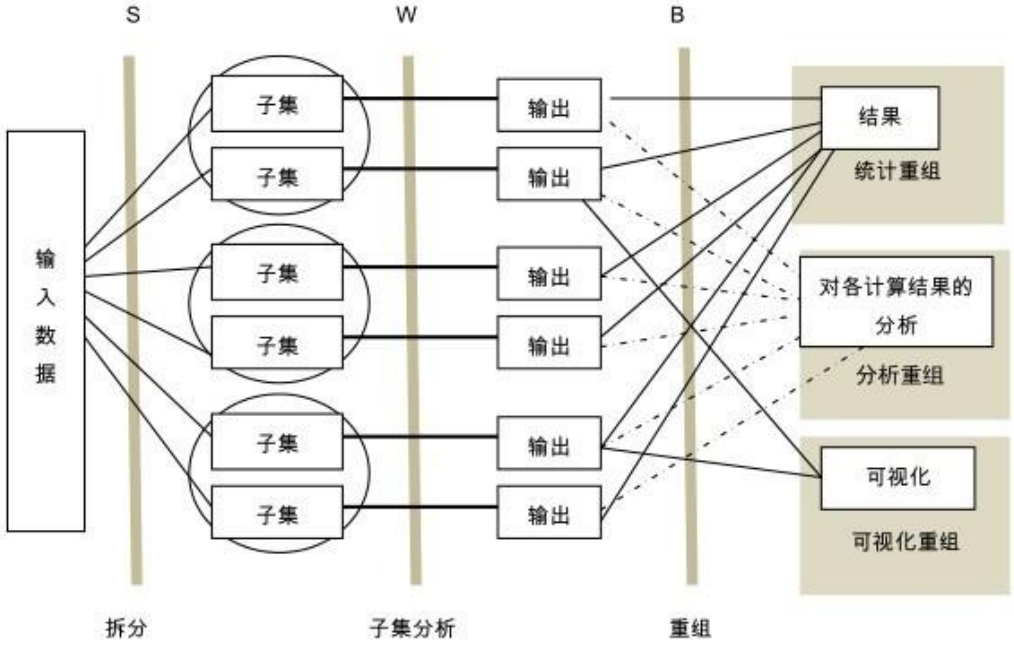


实际应用

- CAVE2混合虚拟现实系统

大尺度数据的分而治之可视化与分析

- 分而治之（Divide and Conquer）：面向大数据的分而重组思想的统计和计算框架



4 数据不确定性的可视化

不确定性（Uncertainty）

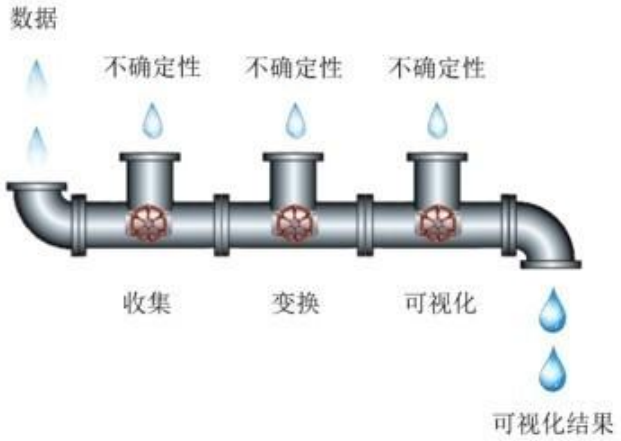
- 数据从采集到使用的过程中不可避免地会带来误差和不确定性
  - 与数据错误和数据矛盾不同，不确定性存在于数据处理的各个环节，甚至会导致新的不确定性
- 通过可视化的方法来呈现不确定性有助于用户准确地理解数据并做出正确的决策。不确定性可视化被认为是数据可视化的关键问题之一
  - 不确定性的清晰表示
  - 降低或避免因不确定性可视化所带来的视觉混乱
  - 降低可视化不确定性所引起的对确定性数据可视化结果的负面影响
  - 不确定性表达的可视隐喻

不确定性的基本定义

- The grammar of graphics列举了众多跟不确定性息息相关的常用概念
- 可变性（Variability），即不一致性
  - 如一组数据的两个或多个元素的值不相等，则这组数据可变的。数据的可变性导致未知的不确定性
- 噪声（Noise）
  - 由平稳随机过程产生的不一致性。如线性系统中的高斯白噪声。噪声通常导致不确定性，噪声有很大的随机性
- 不完整性
  - 设备故障、保密限制等因素是造成数据丢失、不完整的主要原因。不完整的数据往往伴随着不确定性，不可信
- 不定性（Indeterminacy）
  - 针对给定的模型及其相关数据，存在多种参数设置
- 偏倚（Bias）
  - 一种系统偏差。如测量偏倚是真实值与测量值之间偏差
- 误差（Error）
  - 真实值与测量值之间的随机偏差。与偏倚不同，误差可在真实值左右等概率变动；偏倚通常偏向于某一个方向
- 准确性
  - 没有偏倚和误差即为准确。实际的测量值通常是真实值与偏倚、误差的和
- 精确度
  - 没有误差即为精确。一种非常精确的测量通常可能存在一定的偏倚。在实际中，常通过使用更多的有效数字来提高数据的精确度
- 可信度
  - 表示随着时间的变化，测量结果的可重复性。测量结果之间的差异越小，数据越可信
- 有效性
  - 表征了真实测量值与测量过程之间的关系。为了提高有效性，通常不仅需要测量变量本身，还需要测量与实际被测变量相关的一些事物或变量
- 质量
  - 质量是完整性、可信度和有效性的综合

不确定性的来源

- 测量仪器的优劣
- 测量者知识水平的高低
- 不同的仿真或数值计算模型也将引入一定的不确定性——即使同一数值计算模型，不同的参数设置也会引起数据的不确定性



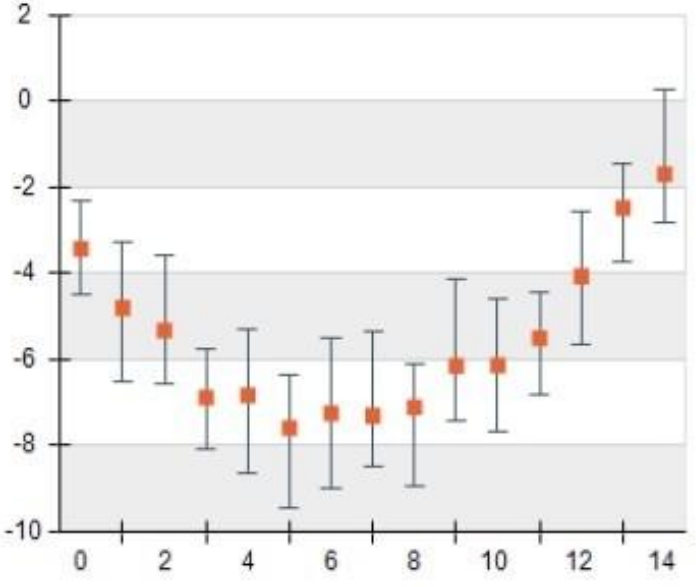
不确定性的可视化方法

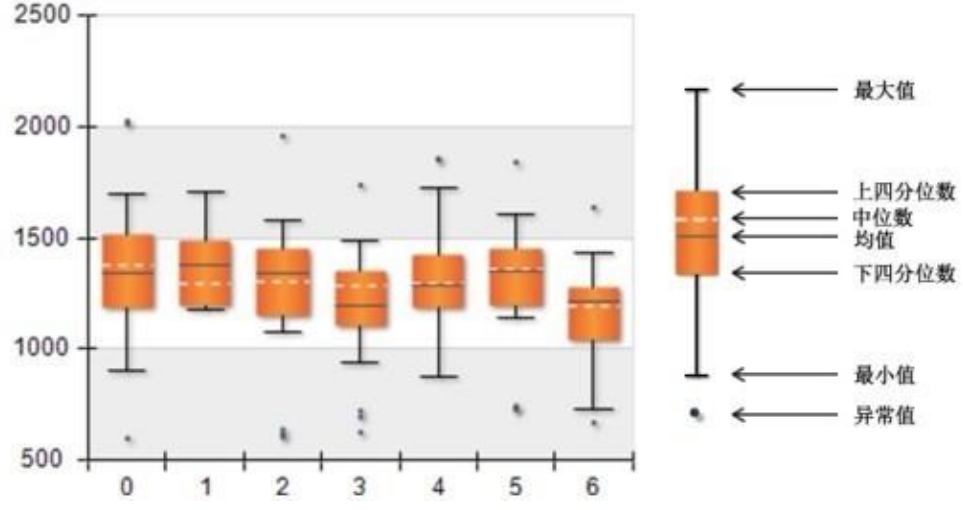
- 不确定性可视化列为可视化十大核心研究问题之一

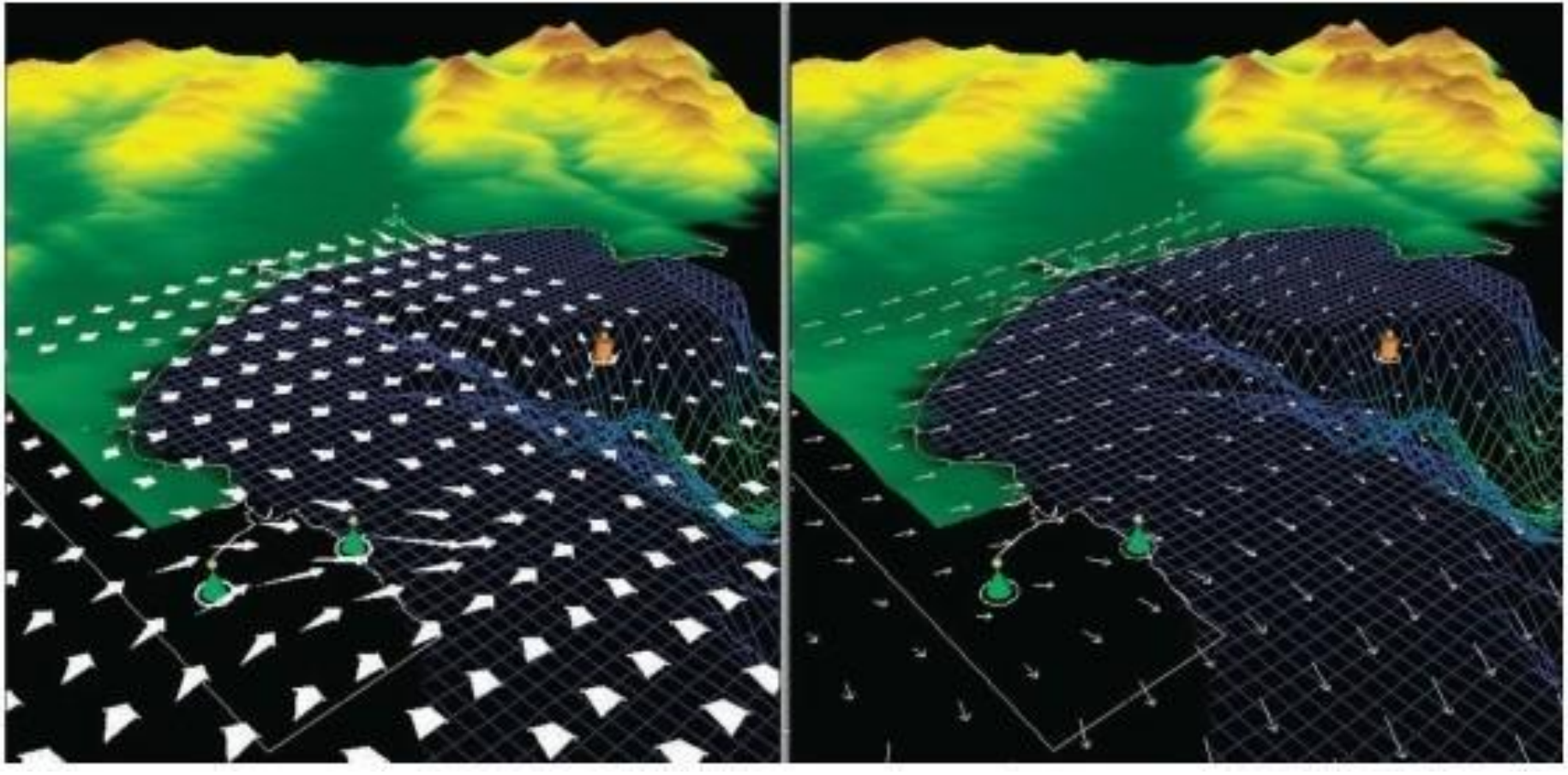
可视化方法	优　势	不　足
图标法	简单、易于理解	易产生视觉混乱等问题
视觉元素编码法	可帮助用户迅速定位可视化结果中不确定性所在的区域和大小	需要精心选择视觉元素才能有效地表达不确定性
几何体表达法	形象、直观，可编码高维的不确定性	易污染原有的确定性数据的可视化结果
动画表达法	可帮助用户更加生动、形象地理解不确定性，提供了更高的自由度调节可视化结果	理解曲线较长，易引起视觉疲劳

图标法

- 误差条图、盒须图

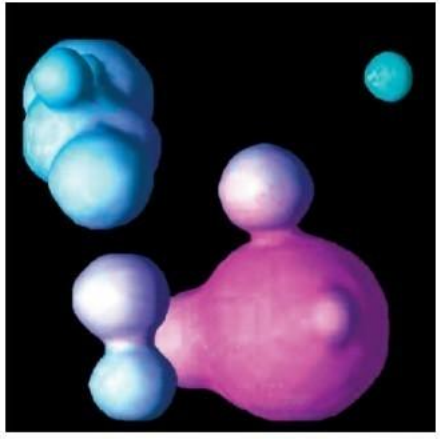
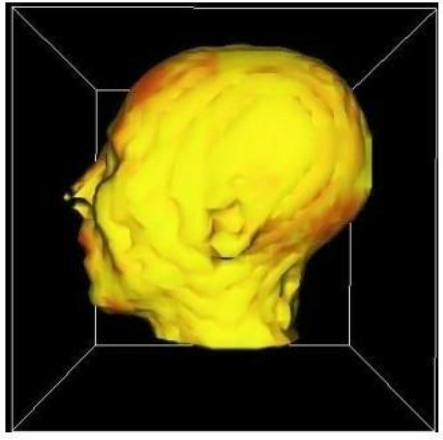
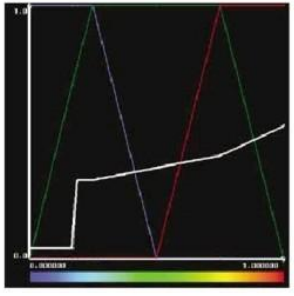
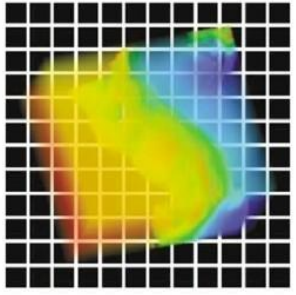
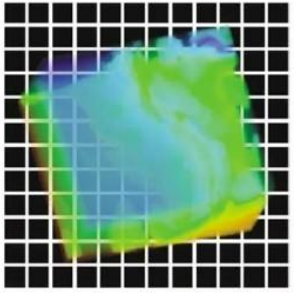

2
0
-2
-4
-6
-8
-10
0
2
4
6
8
10
12
14

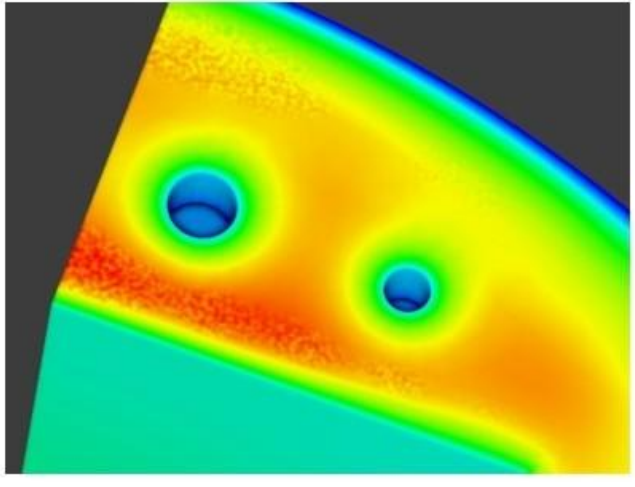

2500
2000
1500
1000
500
0
1
2
3
4
5
6
最大值
上四分位数
中位数
均值
下四分位数
最小值
异常值

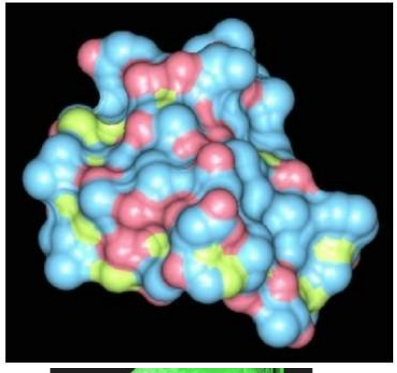


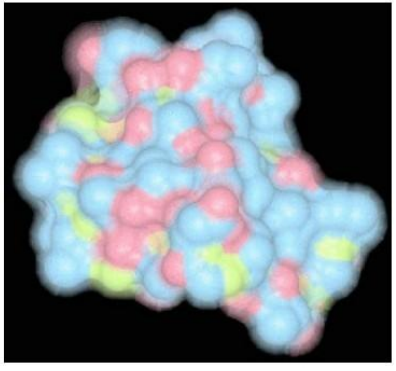
视觉元素编码法

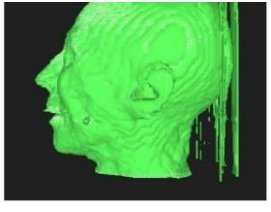
- 以视觉元素作为不确定性编码的基本载体。包括位置、形状、亮度、颜色、方向和纹理等



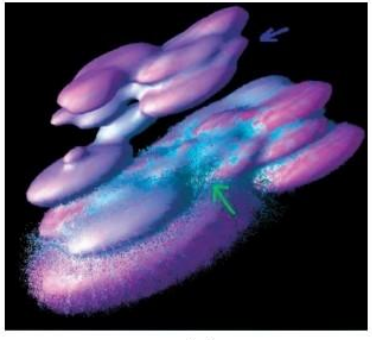




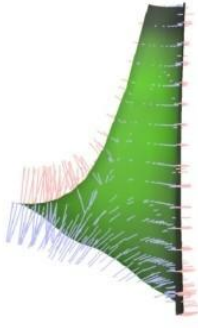


几何体表达法

- 利用代理性的几何物体可提供额外的视觉表达，借以表达数据的不确定性。包括：点、线、面、网格、体等

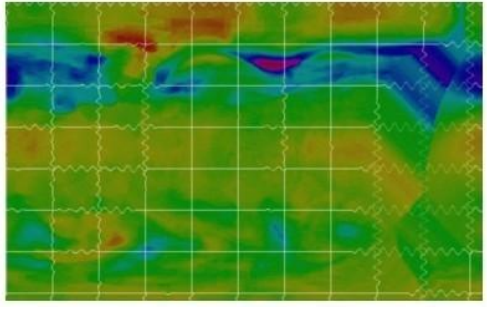


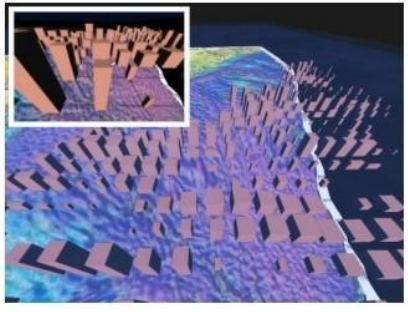
(a)



(b)

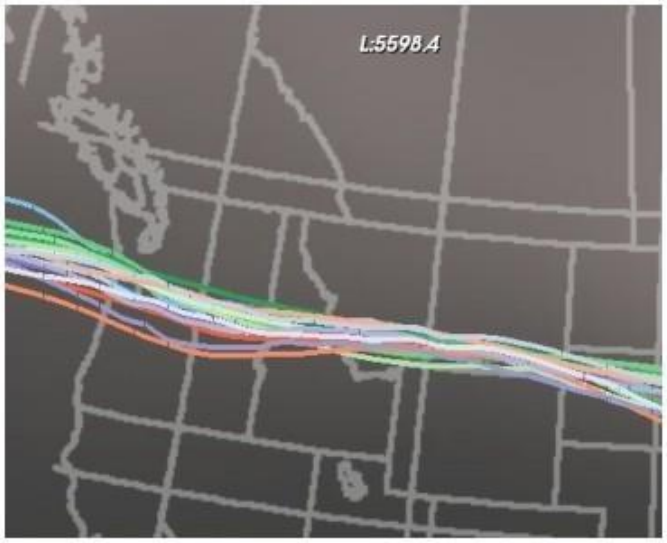
(c)

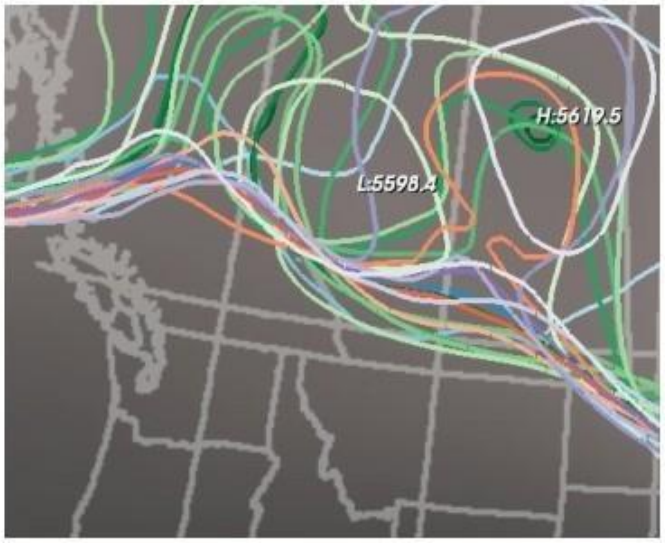




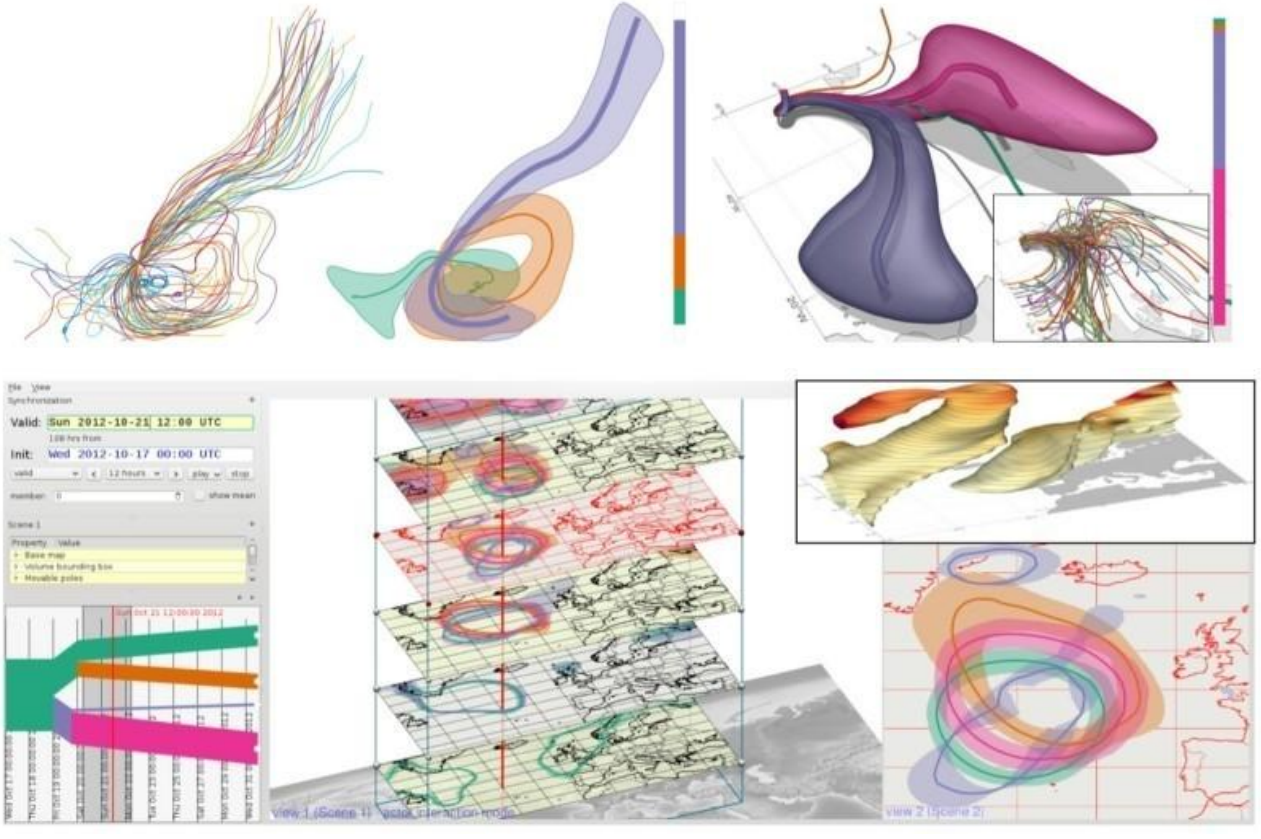
意大利面图（Spaghetti Plot）

- 每条线代表从一个集合成员中提取的等值线，线的颜色编码了数值计算模型，线的杂乱程度间接地描述了集合数据的不确定性



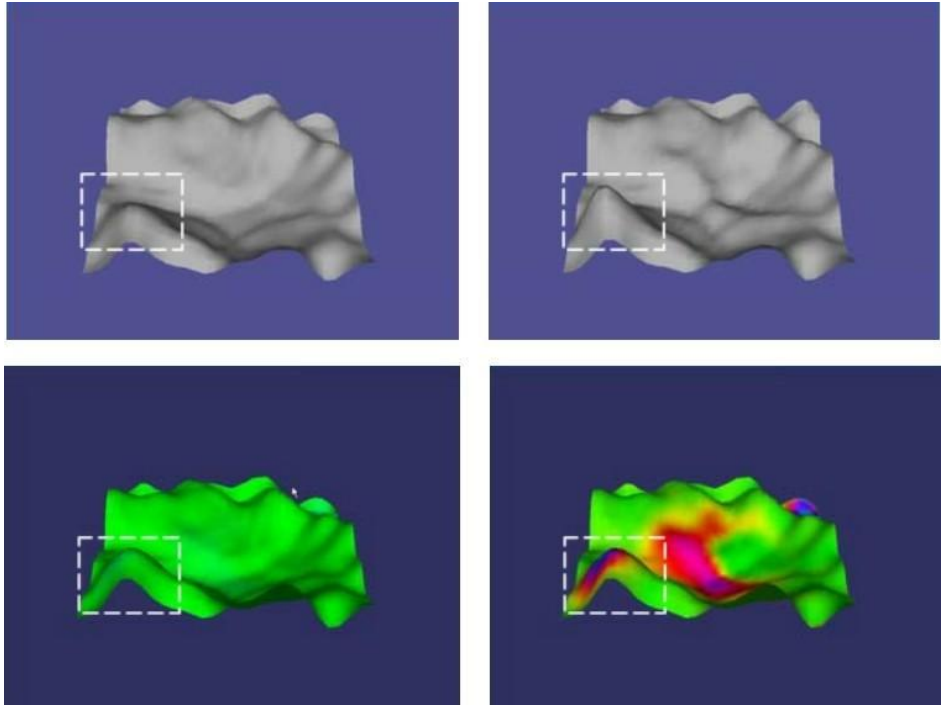


Streamline Variability Plots



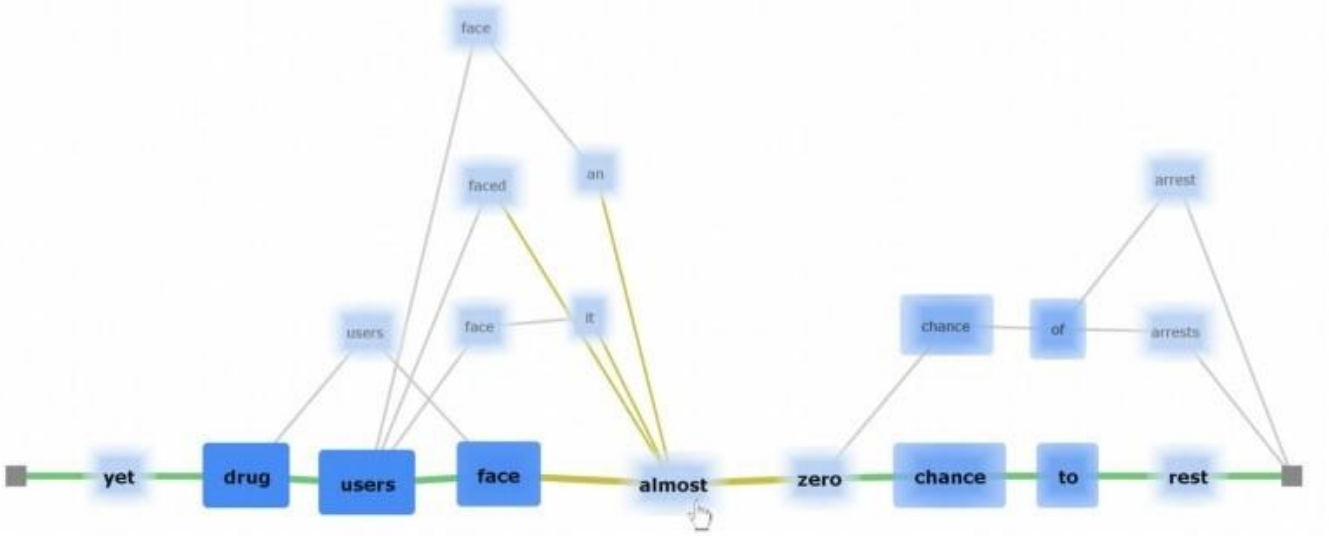
动画表达法

- 动画相关参数可用于编码不确定性
  - 如，速度、持续时间、运动范围、运动顺序、运动模糊、闪烁等



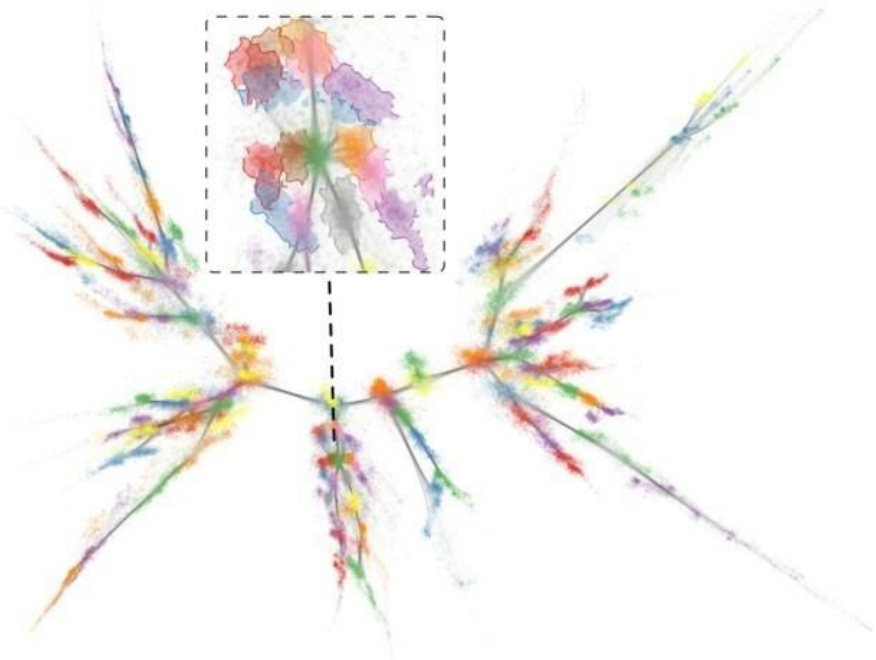
图（网络）的不确定性可视化

- 不确定图
  - 每条路径都是一个句子，每个节点都是一个词
  - 节点的不确定性用其位置、颜色、边缘透明度来表示
  - 根据模型最好的结果路径用绿色展示；而鼠标指针所在节点相连接的边用金黄色展示



概率图

- 同时展示图的拓扑结构和不确定信息的概率分布



AmbiguityVis

- 边长度的歧义性、社群结构的歧义性、点/边聚合的歧义性
- 计算度量，按照用户需求叠加绘制歧义性热力图：蓝色越深，代表用户选择的歧义性越大