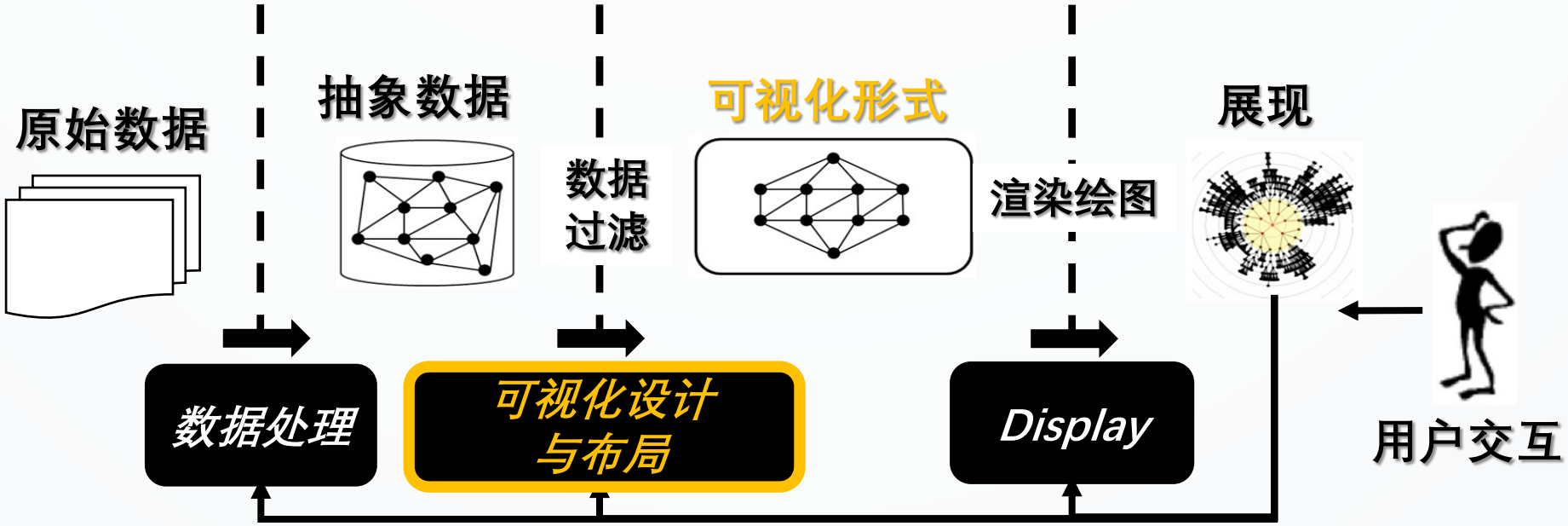
多维度数据可视化

信息可视化

曹楠（教授）
https://idvxlab.com
同济大学

智能大数据可视化实验室

怎样对数据进行可视化?



信息可视化参考模型

课程大纲

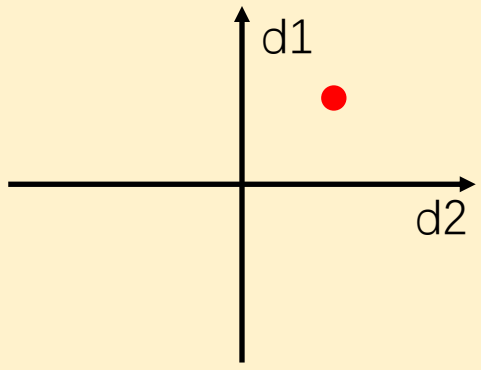
- 多维度数据的可视化
  - 基于不同坐标系的可视化方法(Coordinate Systems)
  - 基于像素的可视化方法 (Pixel Oriented)
  - 基于图标的可视化方法（Icon Based)
  - 基于网格的分视图展示方法（Small Multiple)
- 树的可视化
- 图的可视化

课程大纲

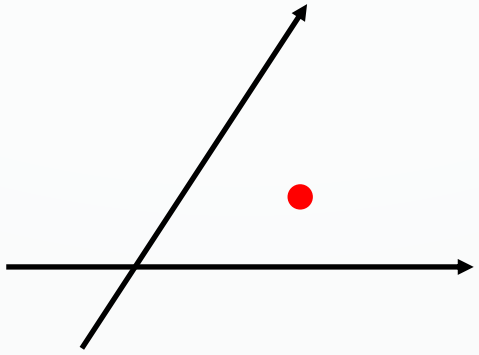
- 多维度数据的可视化
  - 基于不同坐标系的可视化方法(Coordinate Systems)
  - 基于像素的可视化方法 (Pixel Oriented)
  - 基于图标的可视化方法（Icon Based)
  - 基于网格的分视图展示方法（Small Multiple）
- 树的可视化
- 图的可视化

基于不同坐标系的可视化方法

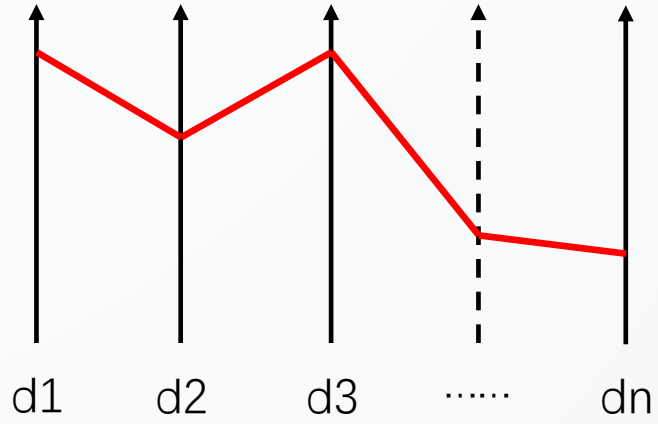
- 常用的三种不同可视空间



正交投影空间



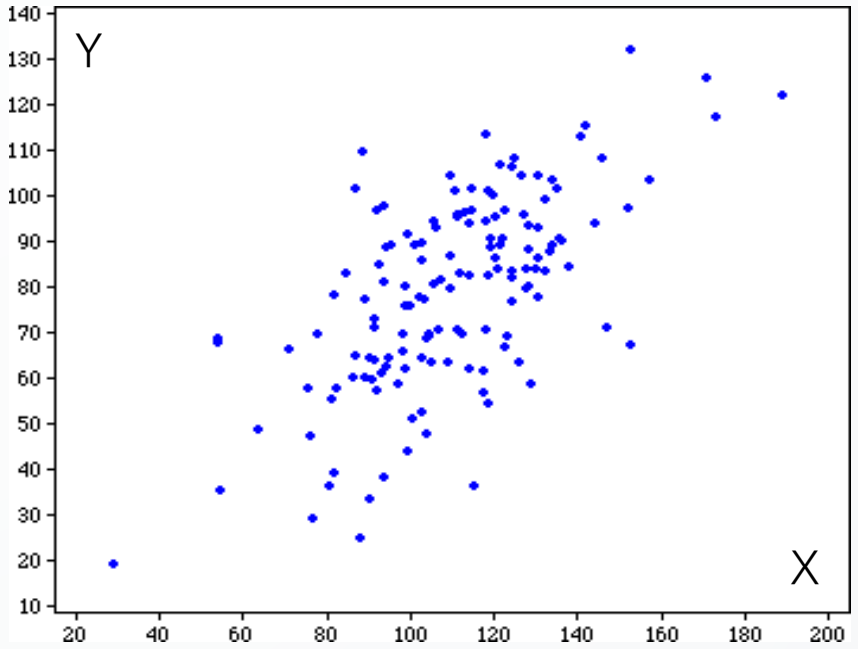
非正交投影空间

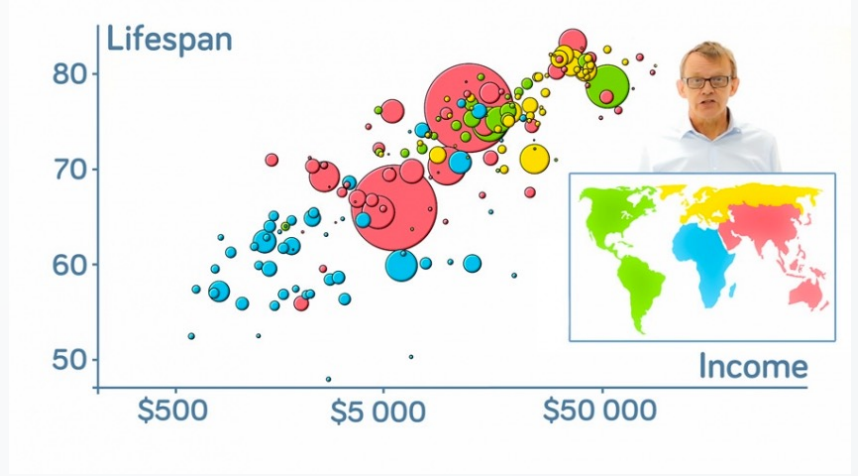


平行坐标轴系统

基于不同坐标系的可视化方法

- **正交投影空间 – 散点图 （Scatter Plot ）**
  - 直接利用笛卡尔坐标系显示多维度数据，可以同时表示数据元素在坐标轴所确定的两个维度上的分布
  - 每一个点代表数据集中的一个数据元素
  - 点的位置，由其在对应维度上的取值所决定
  - 可以通过点的大小、颜色等属性展示其他维度的数据信息，所得到的可视化也被称为“气泡图”，气泡图最多能够显示四个维度的信息
  - 更多维度怎么办?

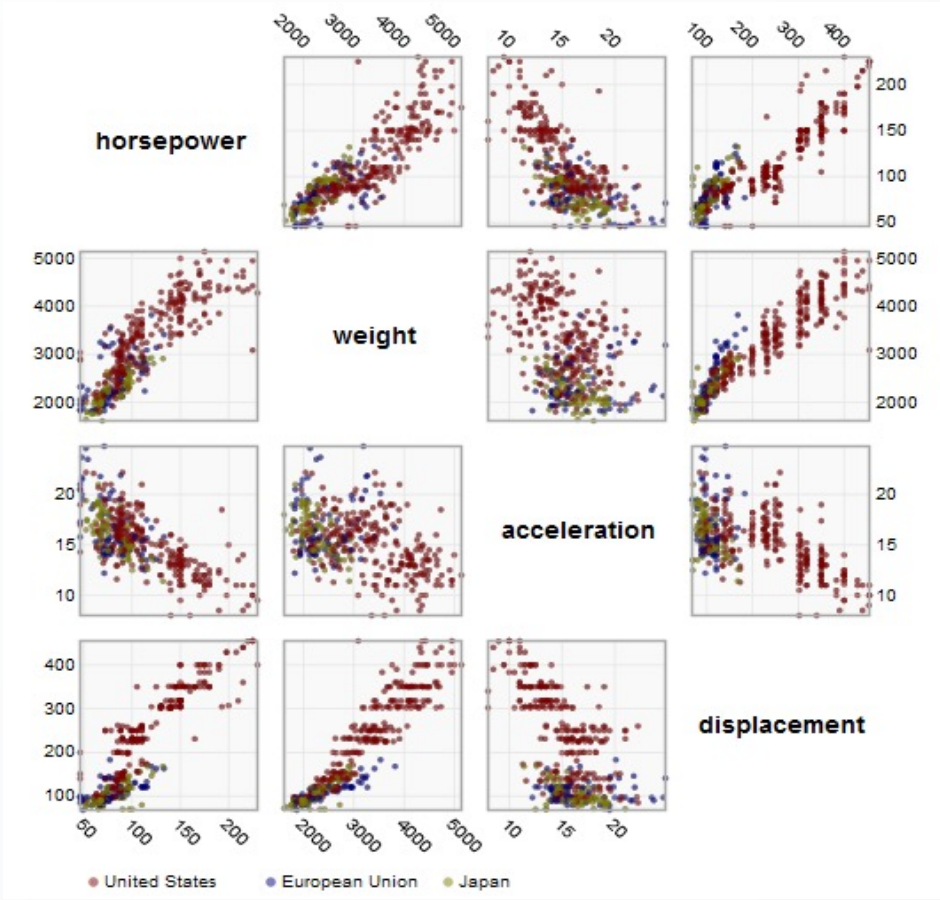




基于不同坐标系的可视化方法

- **散点图矩阵（Scatter Plot Matrix）**

  - 矩阵的每一行、每一列均为数据的一个维度
  - 矩阵的中的元素为 一个 散点图，散点图的横轴与纵轴所对应的维度分别由其所对应的矩阵的行与列所决定
  - 矩阵是对称的
  - N个维度 对应着 N (N - 1)/ 2 个散点图，因此无法显示高纬度数据



基于不同坐标系的可视化方法

- **正交投影空间 – PCA 降维**

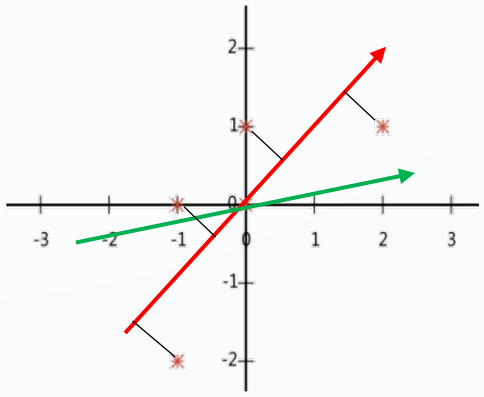
  - 投影的方法是将高维度数据 投影到低维度空间， 能够在低维度空间展示数据在所有维度上的整体分布情况，但是无法明确的展现有 实际意义的 坐标信息

  - PCA （Principle Component Analysis）即主成分分析，是一种常见的正交投影方法

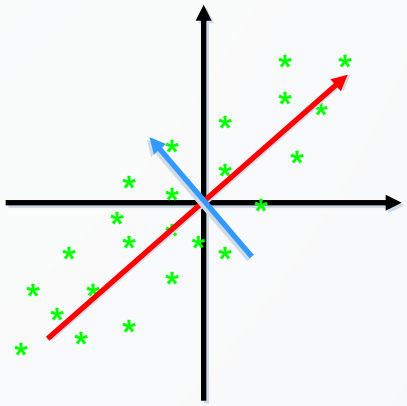
  - 该算法旨在将高维度数据投影到低纬度空间，同时数据在低维空间的分布能够最大程度的保持数据在原有高维空间中的差异

  - 当被投影在一维空间是，数据元素的差异可以用 **方差** 来计算，当被投影在二维空间时，可以用 **协方差** 来计算

  - 用 PCA 进行降维就是在空间中选择两个相互正交的投影方向，使得当数据投影到 由着两个方形所构成的空间中时，他们的协方差是最大的



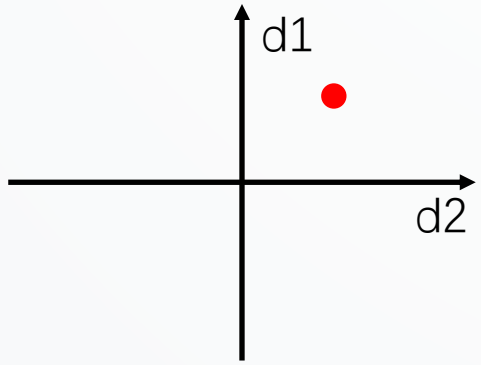
映射到1维空间就是选择一个方向进行投影，使得数据在该方向上的分布的方法最大（例如 红色的方向）



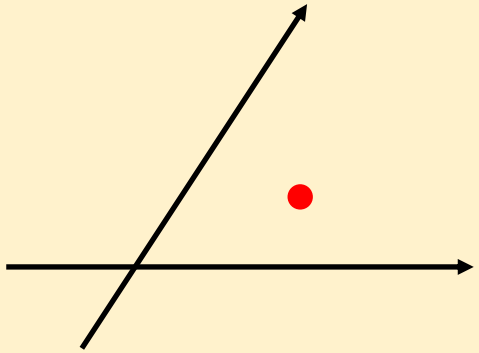
映射到2维空间就是选择连个相互正交的向进行投影，使得数据在该方向上的分布的协方差最大（例如 红色 与 蓝色的方向）

基于不同坐标系的可视化方法

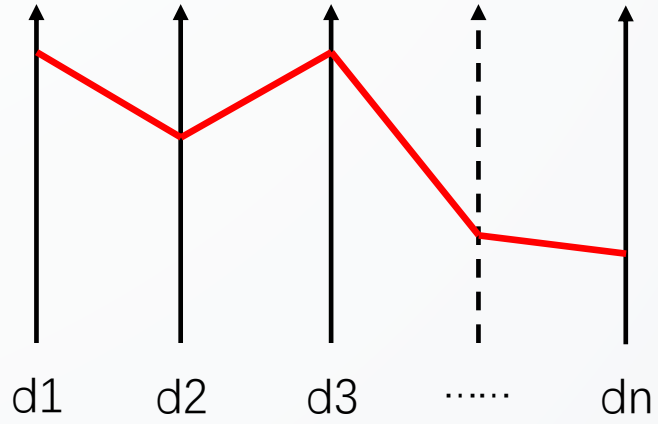
- 常用的三种不同可视空间



正交坐标系



非正交投影



平行坐标轴系统

基于不同坐标系的可视化方法

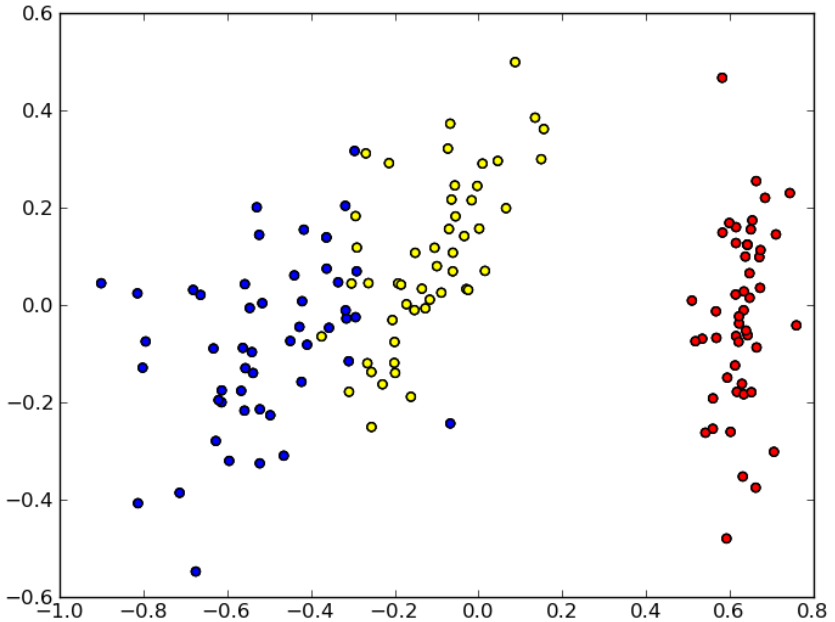
- **非正交投影空间 – MDS**

  - 将高纬度数据根据特定目标及限制条件投影到低纬度空间之上，投影过程并不保证投影方向相互垂直，而是有限确保能够达到目标约束及条件

  - MDS（Multidimensional Scaling）即相似度结构分析，便是可视化领域经常使用的针对多维度数据的投影方法

  - MDS 的目标是 尽可能在低维度空间中保持数据元素在高维度空间中的两两距离，其优化目标为右边的目标函数

  - 投影过程不涉及具体的投影方向，是对数据整体在不同空间中尺度上的调整



$$min \sum_{i<j} \mu_{ij}(d_p(x_i, x_j) - d_m(f(x_i), f(x_j)))^2$$

数据点X在高维度空间中的距离

数据点X 映射低维度空间中的距离

基于不同坐标系的可视化方法

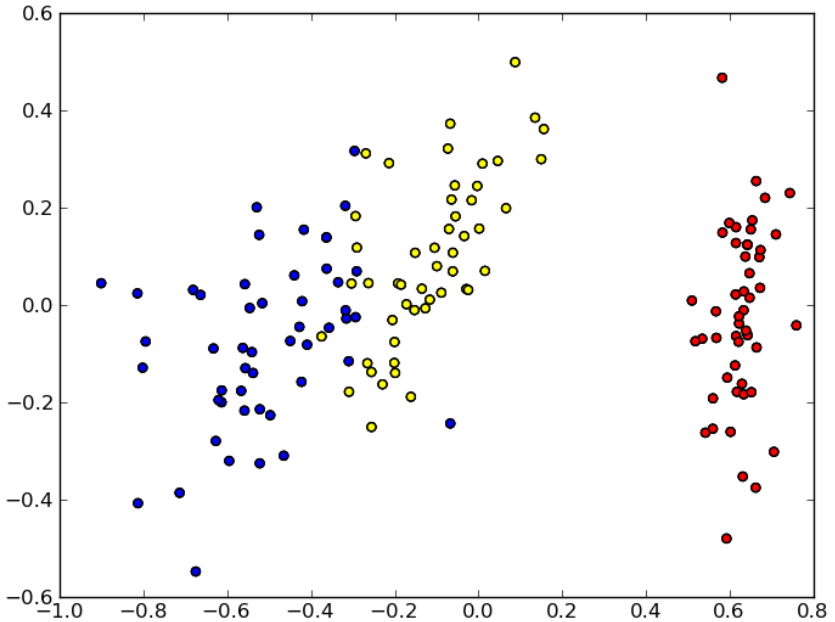
- **非正交投影空间 – MDS**

  - 将高纬度数据根据特定目标及限制条件投影到低纬度空间之上，投影过程并不保证投影方向相互垂直，而是有限确保能够达到目标约束及条件

  - MDS（Multidimensional Scaling）即相似度结构分析，便是可视化领域经常使用的针对多维度数据的投影方法

  - MDS 的目标是 尽可能在低维度空间中保持数据元素在高维度空间中的两两距离，其优化目标为右边的目标函数

  - 投影过程不涉及具体的投影方向，是对数据整体在不同空间中尺度上的调整



$$min \sum_{i<j} \mu_{ij}(d_p(x_i, x_j) - d_m(f(x_i), f(x_j)))^2$$

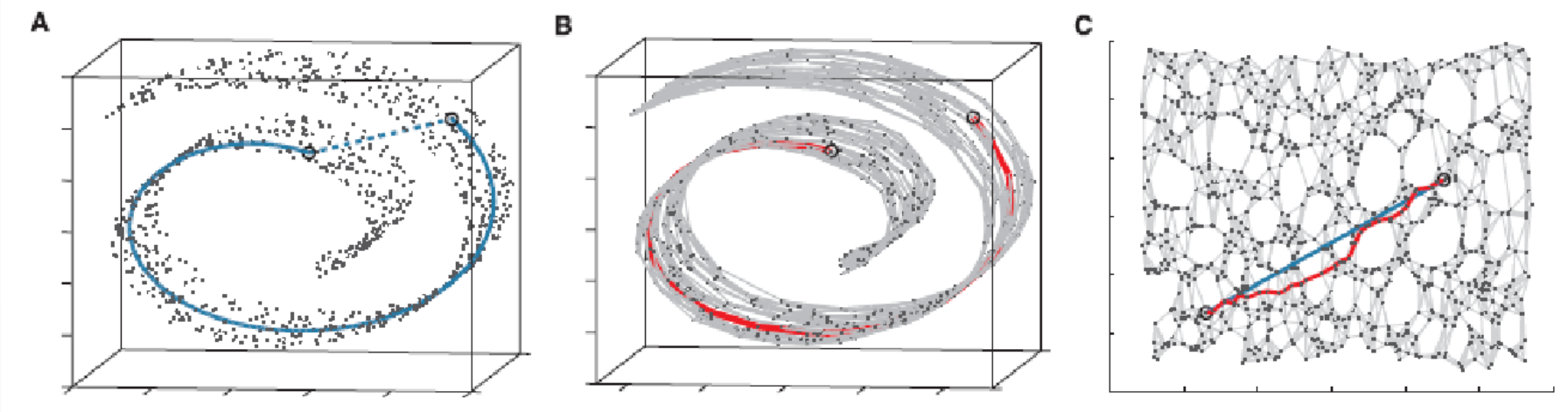
数据点X在高维度空间中的距离

数据点X 映射低维度空间中的距离

基于不同坐标系的可视化方法

- **非正交投影空间 – ISO Map**

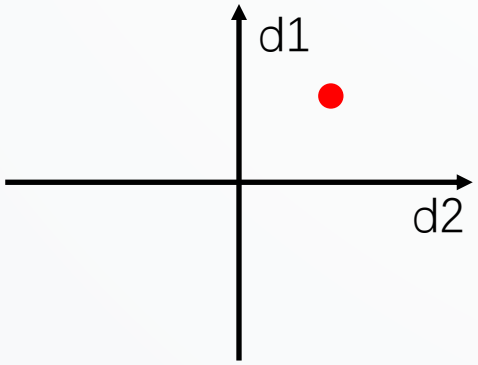
  - ISO Map 是对 MDS 在距离测度上的扩展
  - 用测地线距离（图 A 中的蓝色实曲线）取代了欧式距离（图 A 中的蓝色虚直线），从而确保了当数据在高维度空间不规则分布时，仍然能够正确的描绘数据元素之间的相关性 （图C）



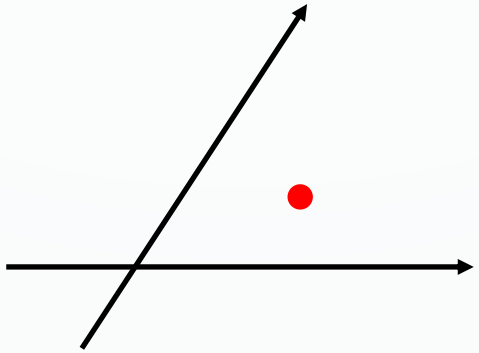
测地线距离　　近似计算方法　　投影展开后的结果

基于不同坐标系的可视化方法

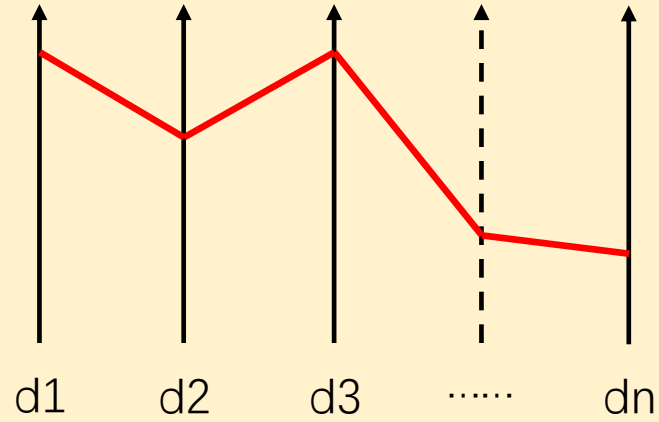
- 常用的三种不同可视空间



正交坐标系



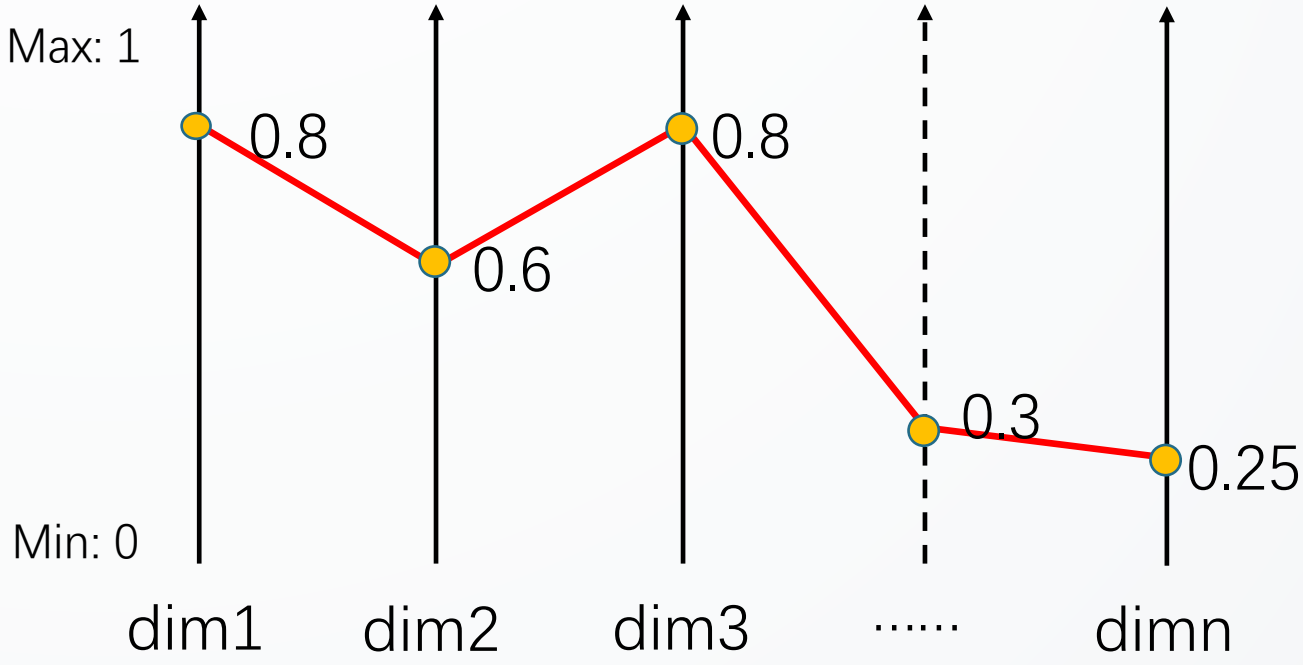
非正交投影



平行坐标轴系统

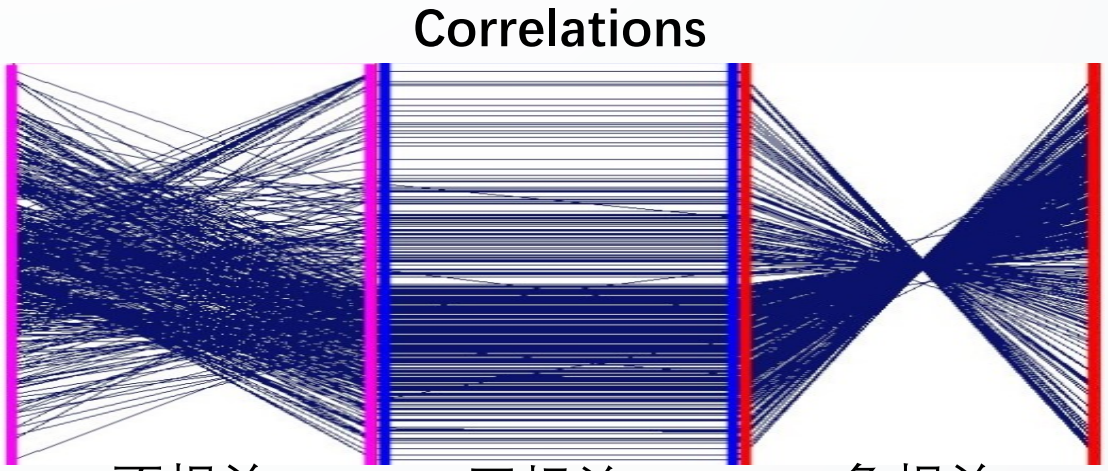
基于不同坐标系的可视化方法

- 平行坐标系
  - 数据的不同维度被显示为平行的坐标轴
  - 数据元素被表示为一根折线
  - 折线与数据轴的交点为数据元素在该维度的取值

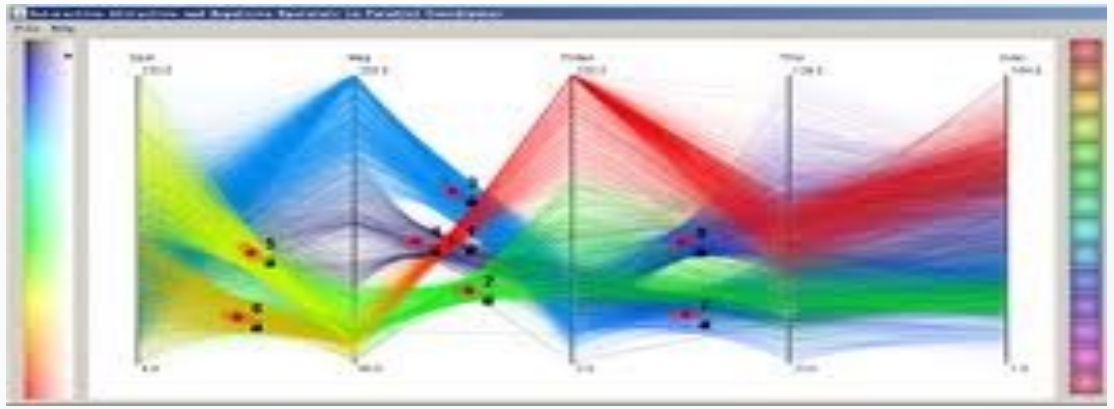


基于不同坐标系的可视化方法

- 平行坐标系中的数据模式
  - 数据的相关性
    - 正相关 – 平行线
    - 负相关 – 交叉线
    - 不相关 – 杂乱无章的交线
  - 数据中的聚类关系

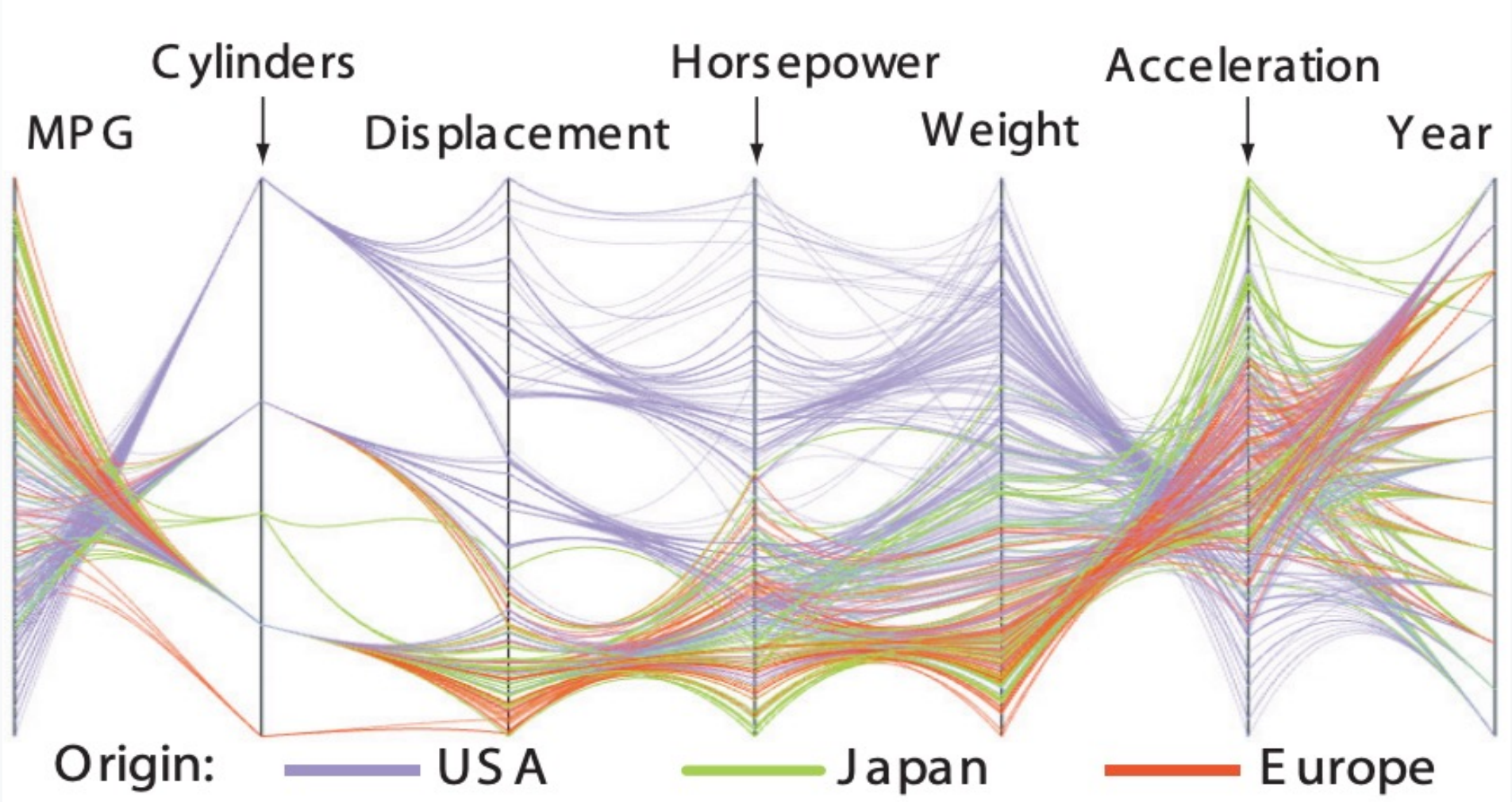


不相关　　正相关　　负相关



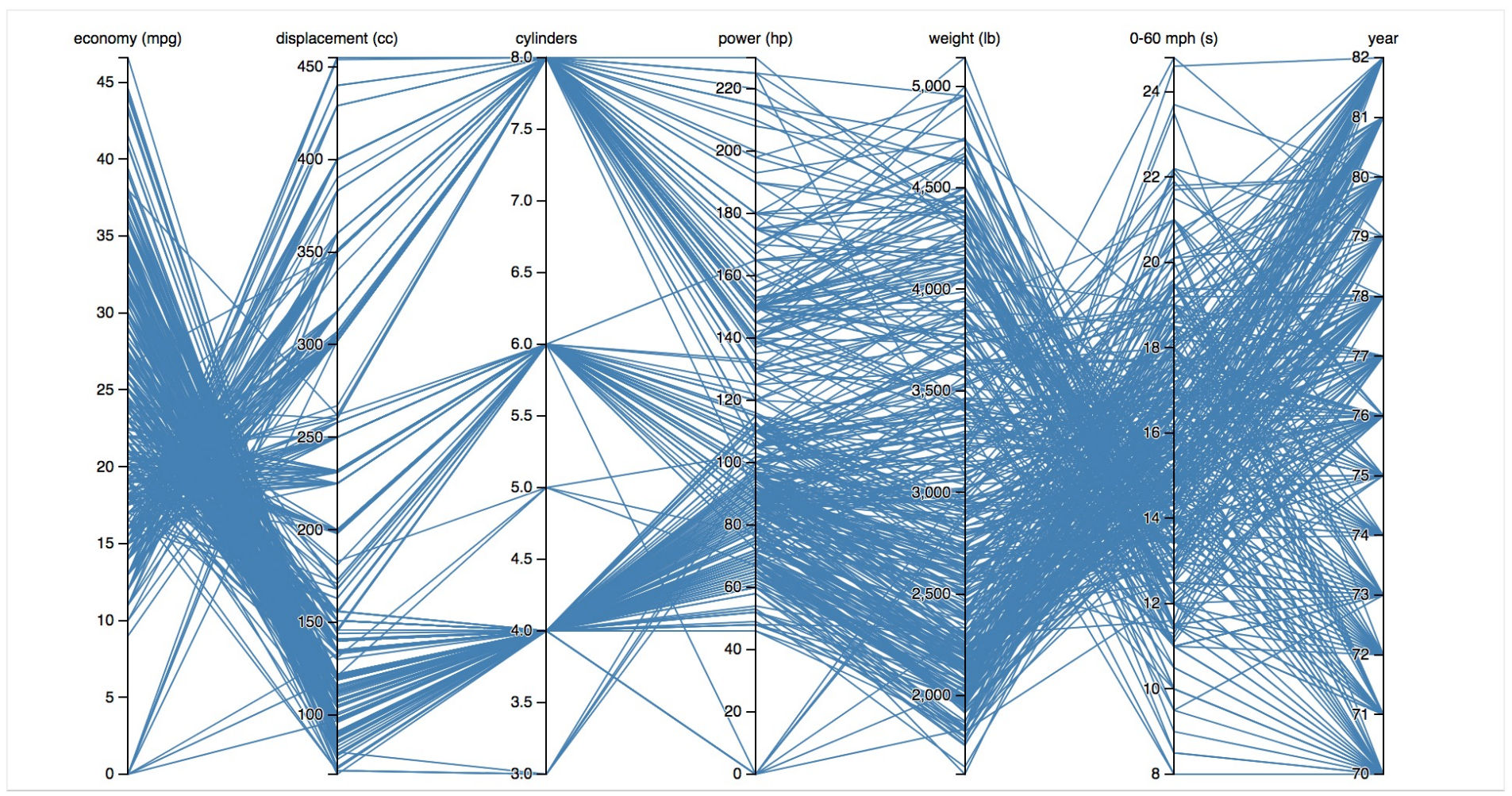
聚类信息

基于不同坐标系的可视化方法



Cars dataset: 392 data items, 7 dimensions

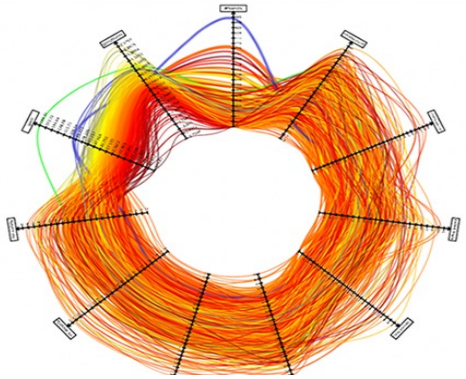
Demo: Interactions in Parallel Coordinates



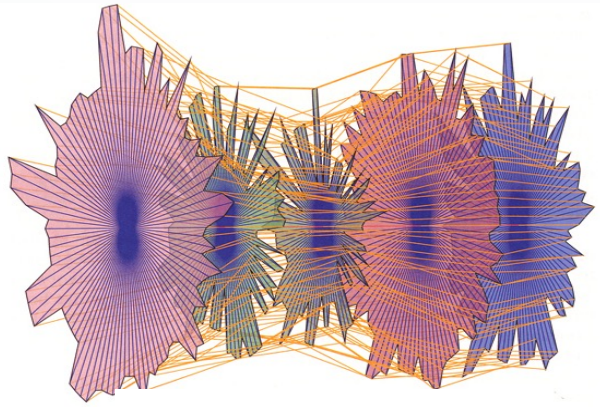
https://bl.ocks.org/jasondavies/1341281

基于不同坐标系的可视化方法

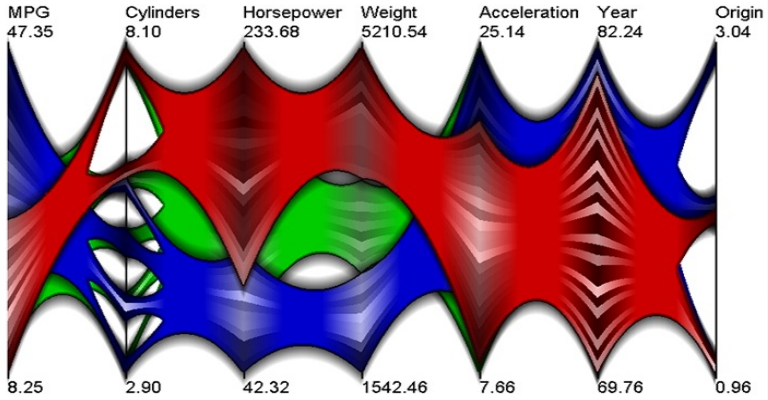
- 平行坐标轴有很多设计上的变化



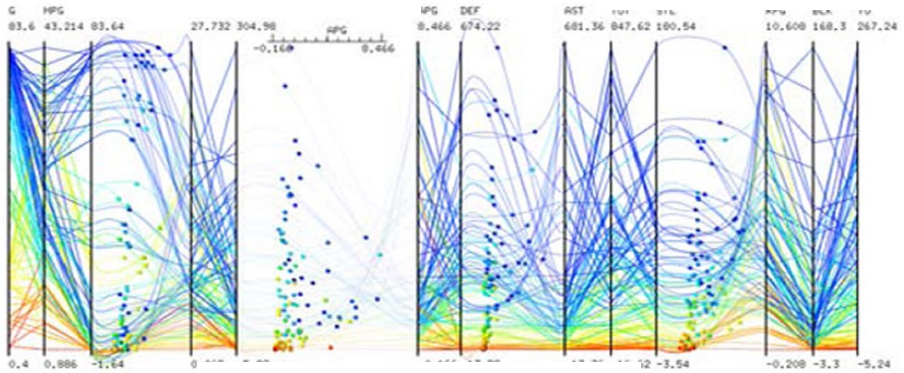
(Homan,1977)



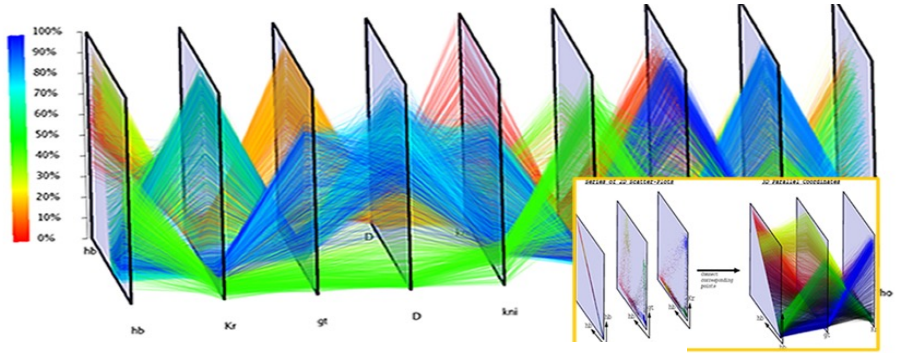
(Fanea et al., 2005)



(McDonnell & Mueller, 2008)



(Yuan et al., 2009)



(Rubel et al., 2006)

基于不同坐标系的可视化方法

- 平行坐标系
  - 看上去只要屏幕可以无限扩张，看似平行坐标轴便能够显示无穷多的数据维度
  - 但是事实并非如此，数据维度过多会导致视觉混乱（如图2）
  - 消除平行坐标系视觉混乱的基本方法
    - 基于 数据过滤 及 聚类的方法
    - 基于 坐标轴排序优化 的方法
    - 基于 视觉增强 的方法

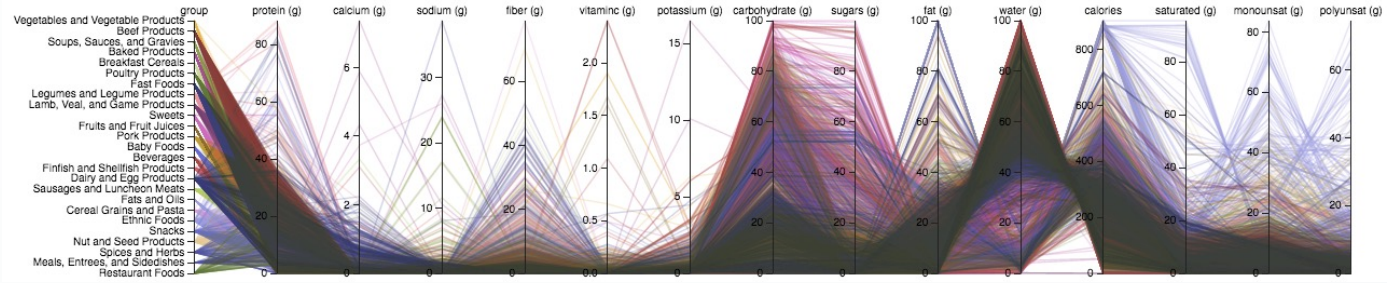


图1：平行坐标系可以显示较多的数据维度

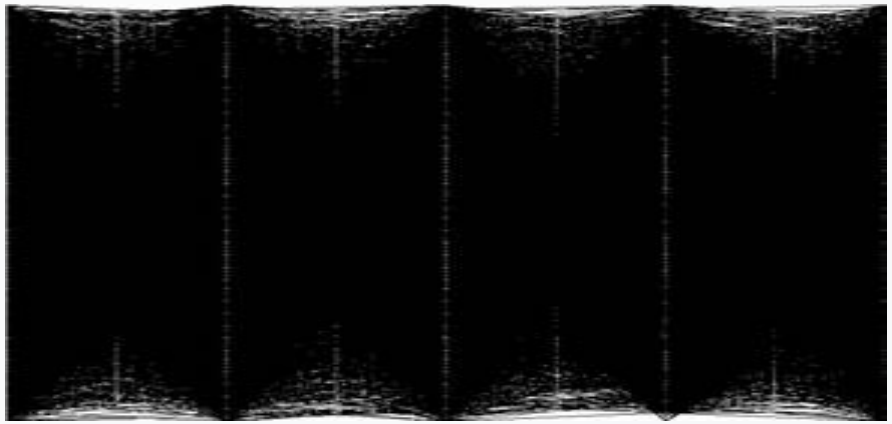
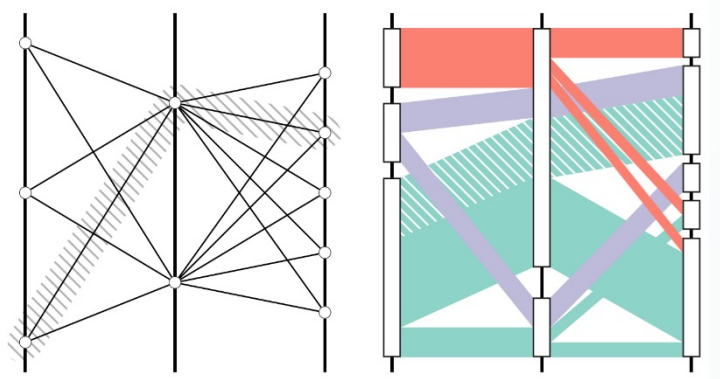
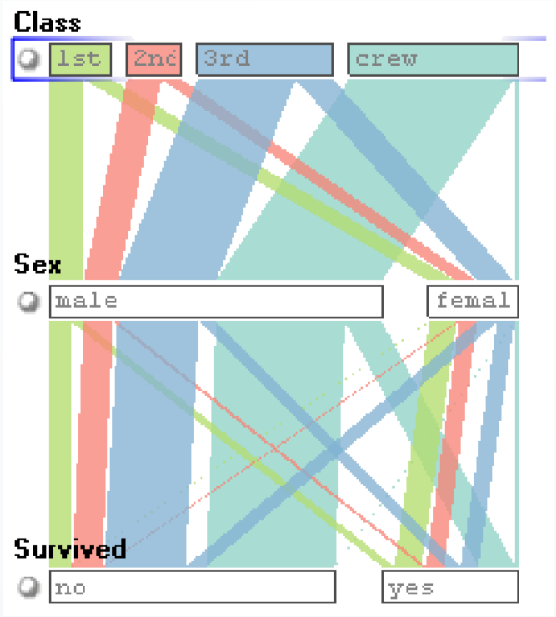


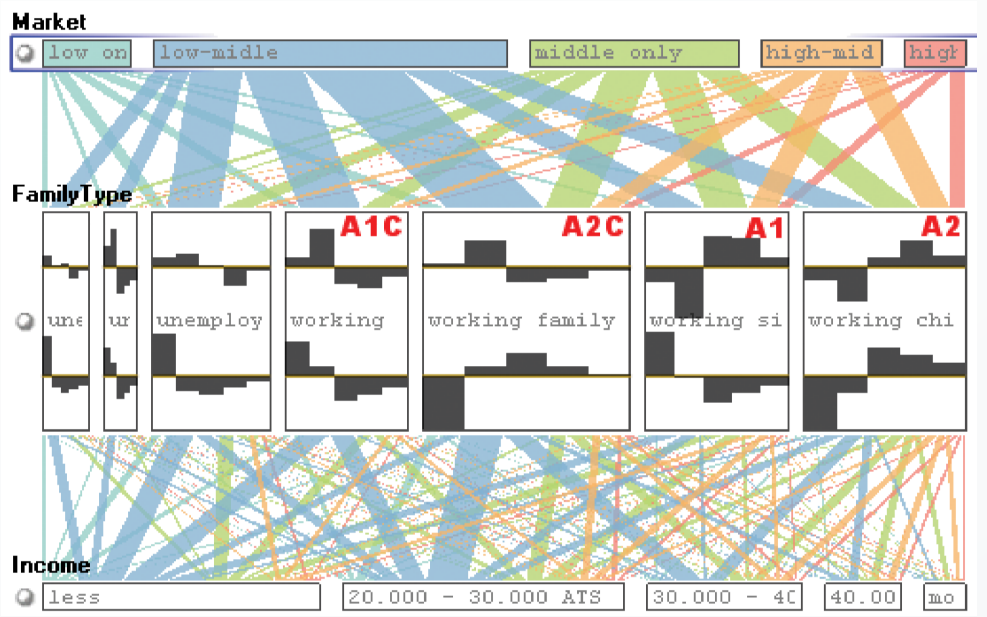
图2：平行坐标系中的视觉混乱

Extension of Category Data: Advizor

- Suitable for category properties
- The coordinate sequence has a greater effect on the results
- Can be used with Parallel Coordinates



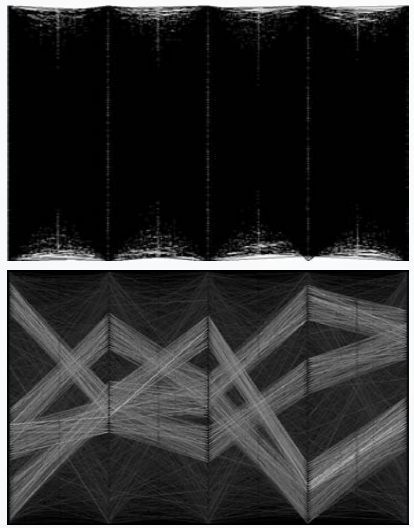




基于不同坐标系的可视化方法

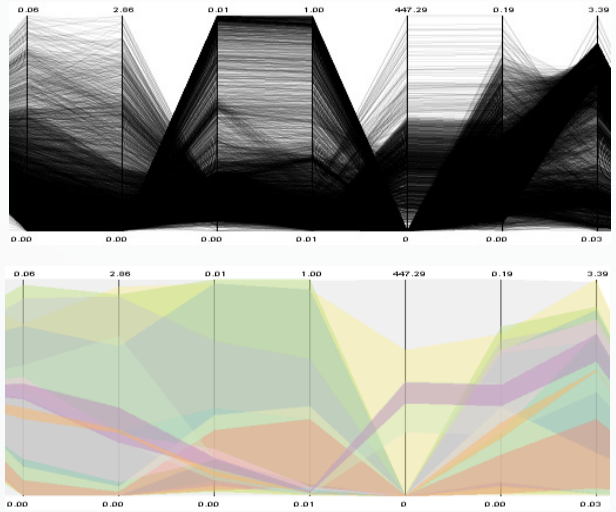
- 消除视觉混乱的基本方法 - “数据过滤”及 “聚类”

基于数据密度的过滤



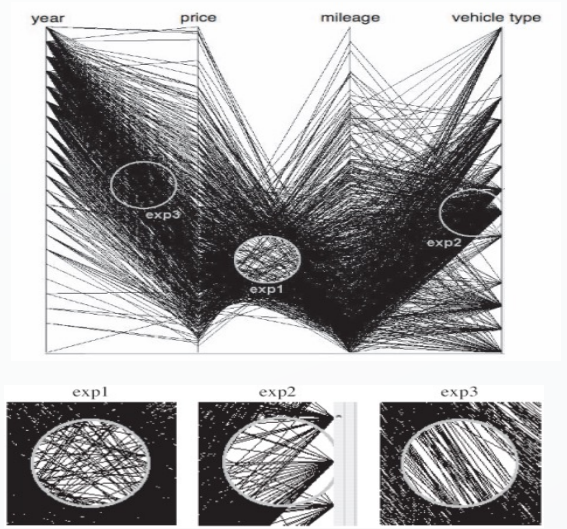
(Artero et al., 2004)

数据聚类



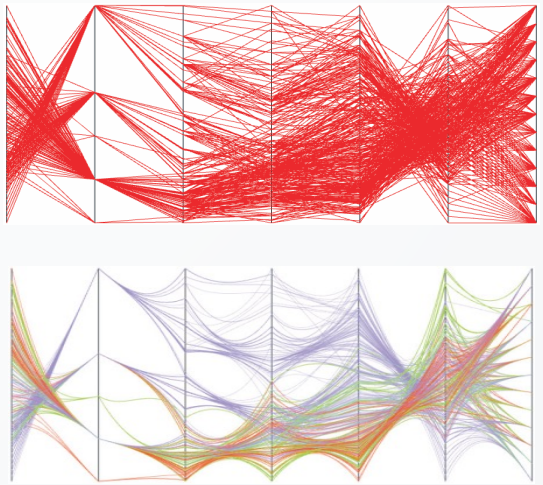
(Novotny et al., 2004)

数据采样



(Ellis & Dix, 2006)

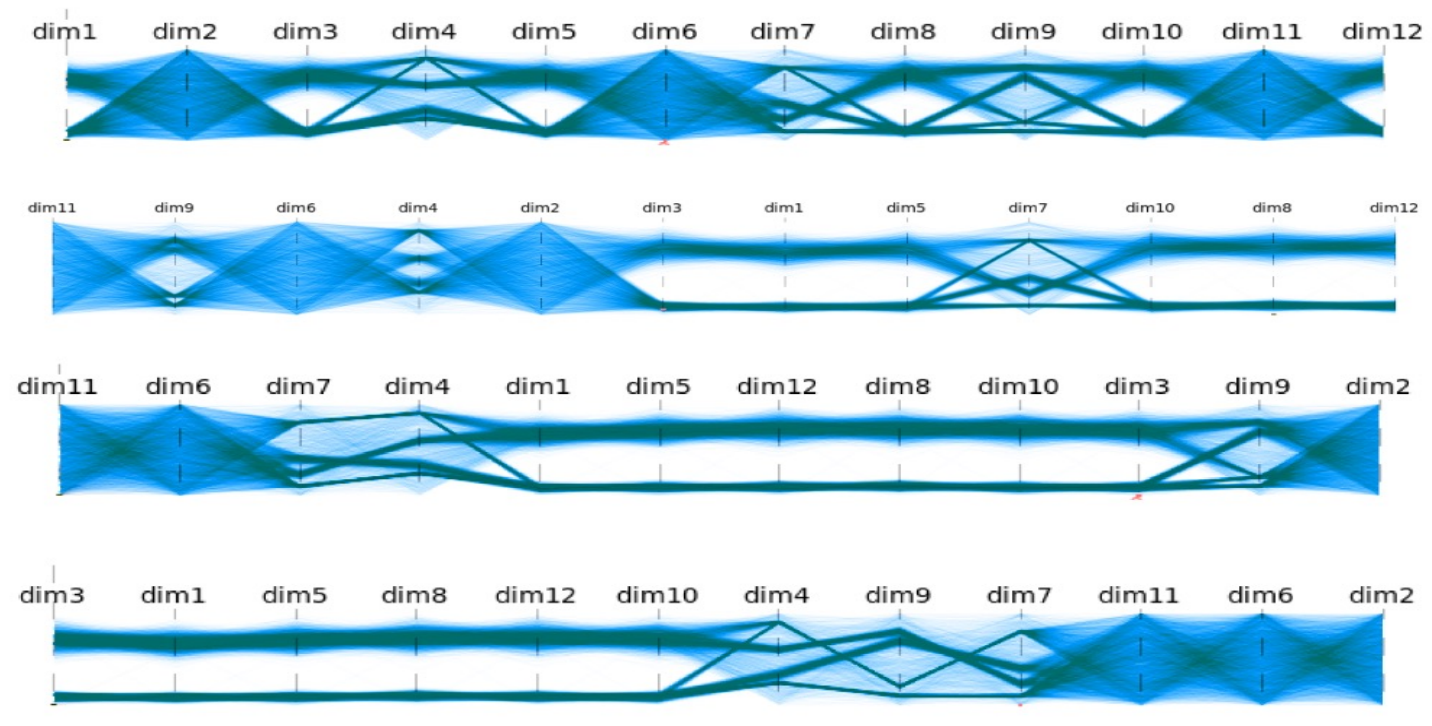
可视化元素聚类



(Zhou et al., 2008)

基于不同坐标系的可视化方法

- 消除视觉混乱的基本方法 - 坐标轴排序优化



数据集的原始排序

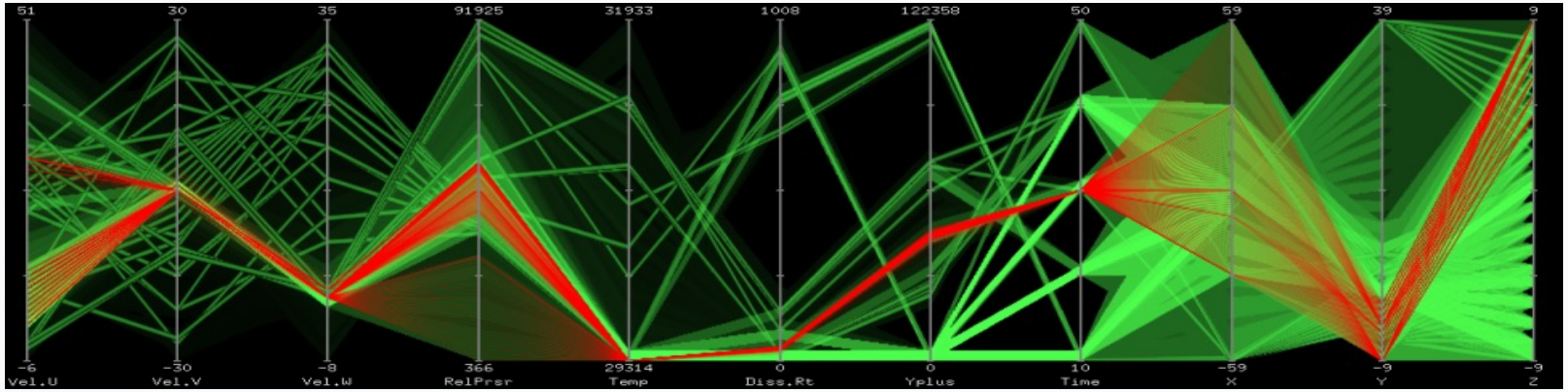
算法1 (Wang et al, 2004)

算法2 (Ankerst et al, 2004)

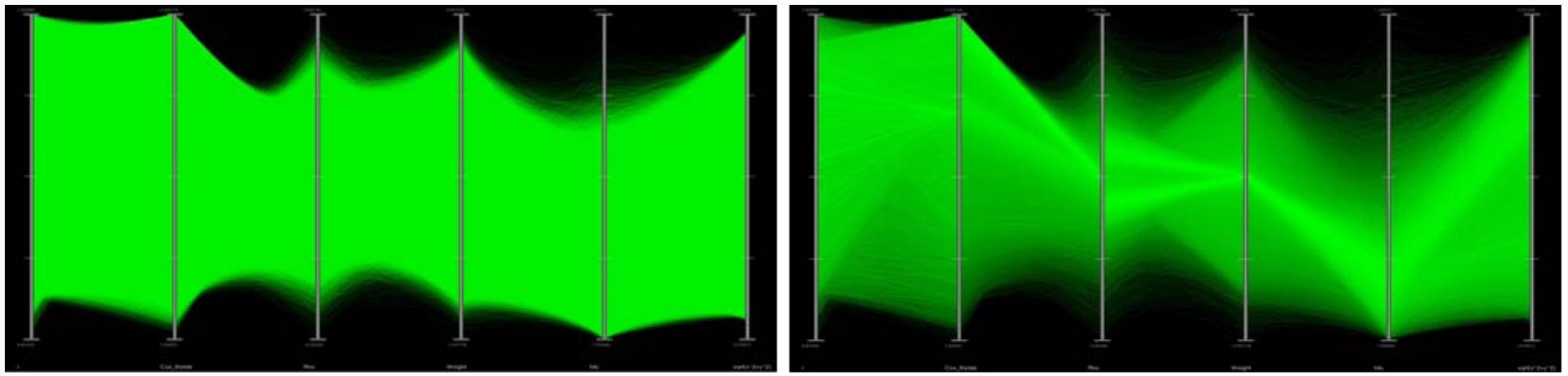
算法3 (Ferdosi & Roerdink, 2010)

基于不同坐标系的可视化方法

- 消除视觉混乱的基本方法 - 视觉增强

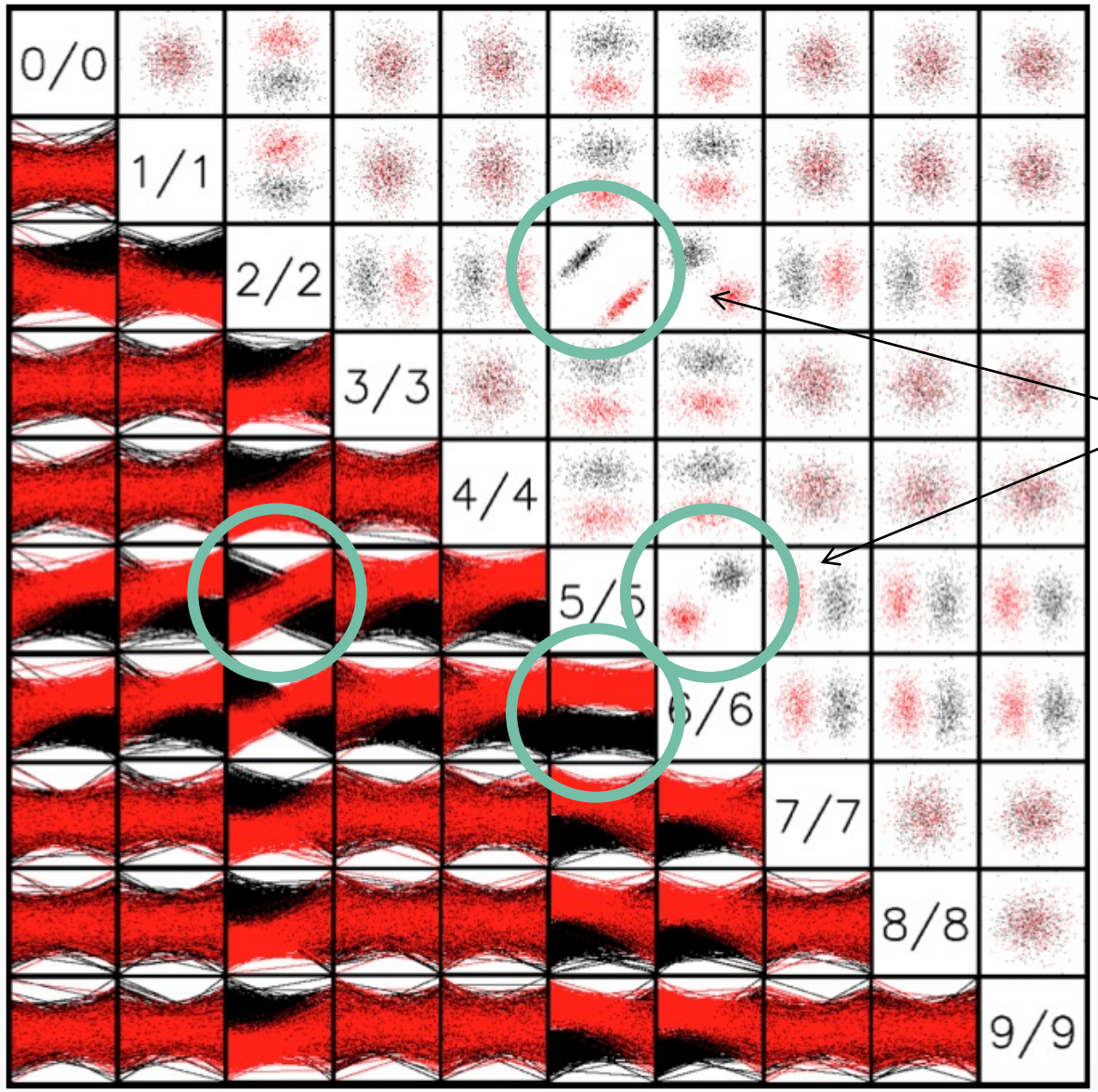


(Novotny & Hauser, 2006)



(Chad Jones et al. 2008)

根据数据密布调整颜色的透明度，从而达到增强视觉效果的目的

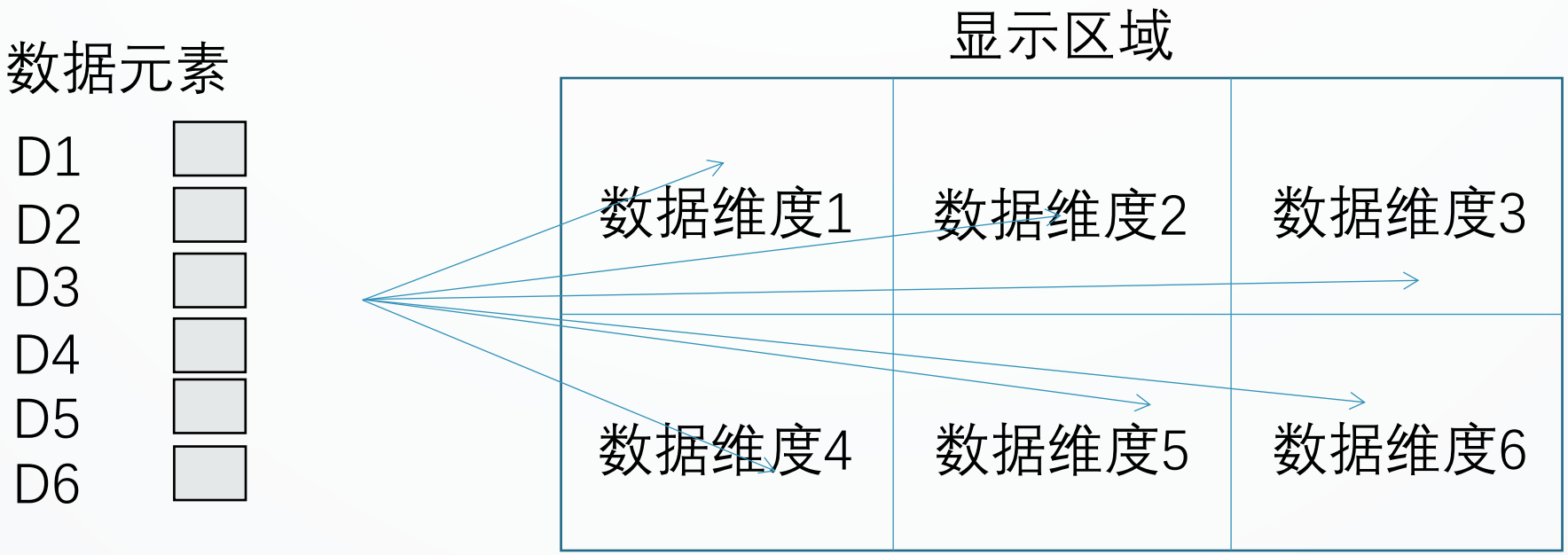


Clusters in scatter-plots

课程大纲

- 多维度数据的可视化
  - 基于不同坐标系的可视化方法(Coordinate Systems)
  - 基于像素的可视化方法 (Pixel Oriented)
  - 基于图标的可视化方法（Icon Based)
  - 基于网格的分视图展示方法（Small Multiple）
  - 多维度数据的可视化诊断
- 树的可视化
- 图的可视化

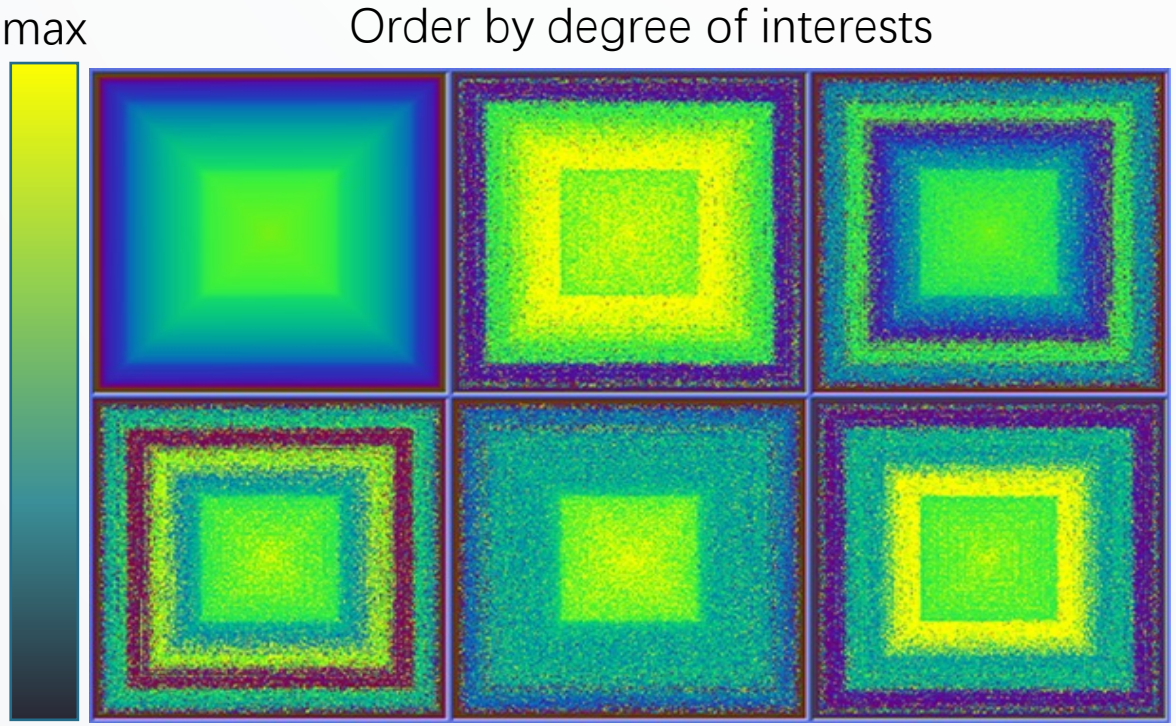
基于像素的可视化方法



- 一个包含有6个维度的多维度数据集合
- 将数据的显示区域划分为6个部分，每个部分分别对应着数据一个维度
- 数据元素中在每一个维度上的属性被表示为 对应区域中的一个像素，像素颜色映射了该属性的取值，不同的像素排列方法，对应了不同的可视化设计

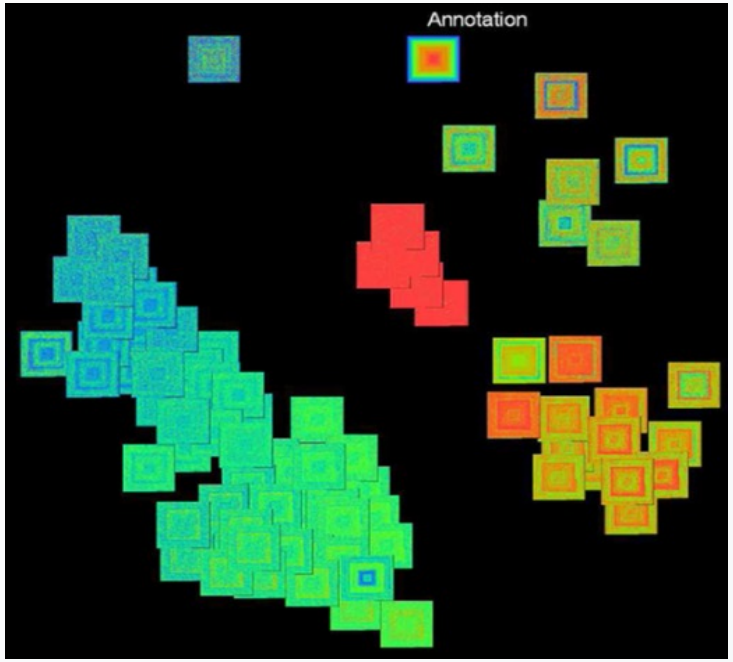
基于像素的可视化方法

- 不同的像素排列方法，对应了不同的可视化设计， 与数据关联



(Keim & Kriegel, 1994; 1996)

对数据库中的一张拥有6个维度表格进行可视化

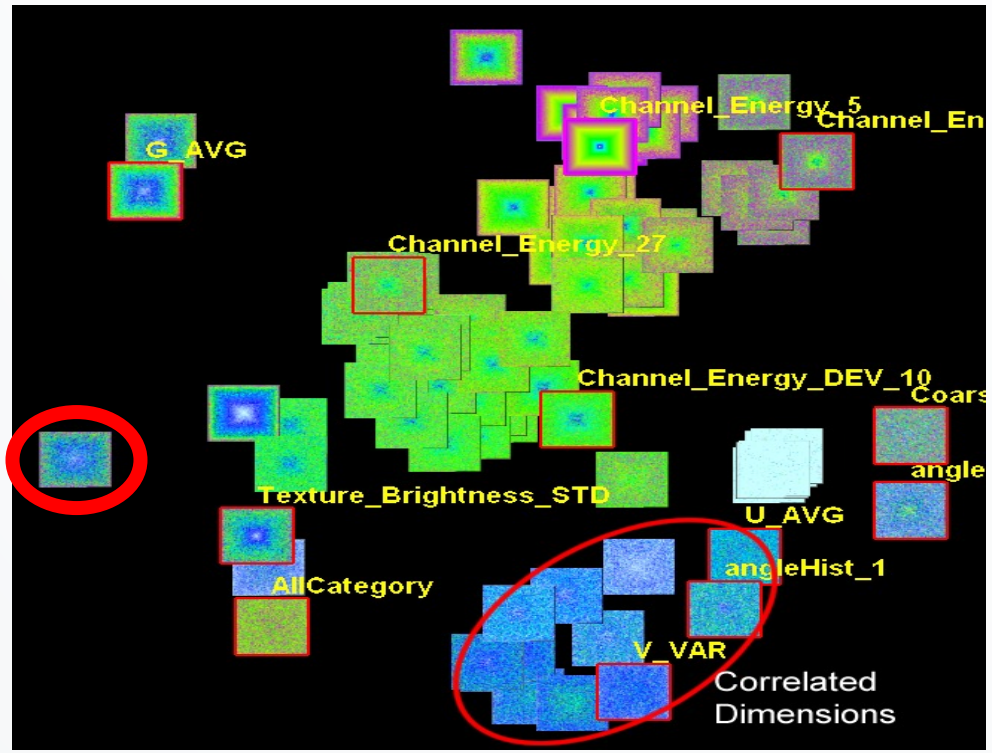


(Yang et al., 2006)

将多维度数据中的每一个维度可视化成一个由像素构成的方块，其中每一个像素对应了一个数据元素，通过MDS 映射，这些方块的位置反应出了数据维度之间的相关性

Pixel Oriented Techniques

Value & Relation View



(Yang et al., 2007)

Visual attributes in each dimension as an pixel icon

Project the icons based on MDS

Similar dimensions are clustered

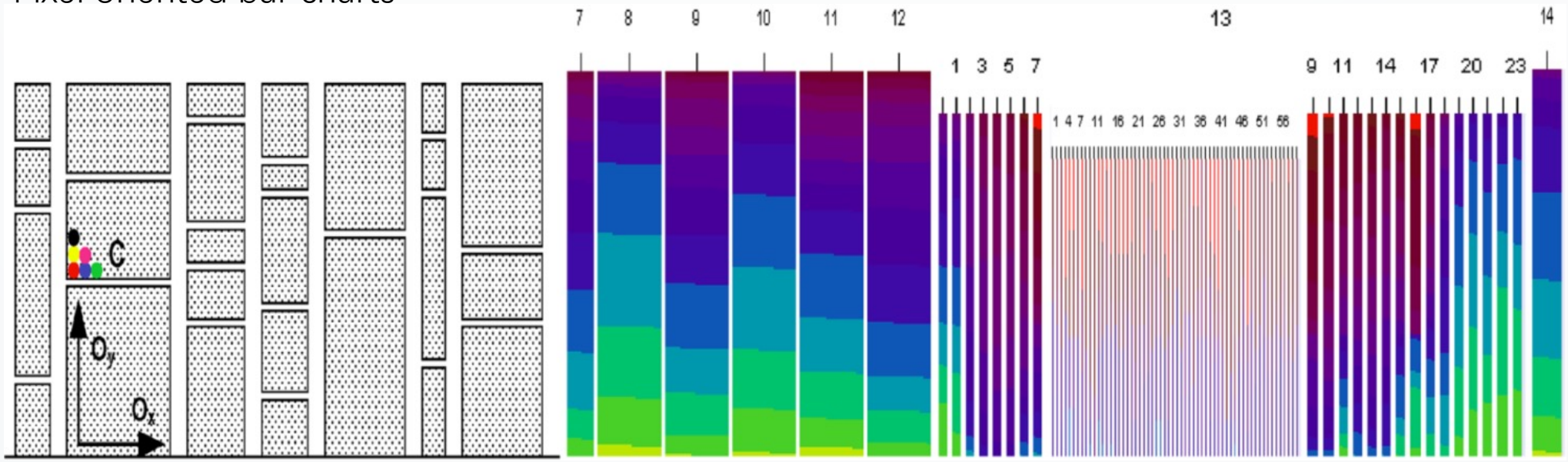
28

Merge the similar dimensions
Delete the outliers

基于像素的可视化方法

- 不同的像素排列方法，对应了不同的可视化设计， 与数据关联

Pixel oriented bar charts



(Keim et al., 2001; 2002)

也可以将显示空间根据数据分类构成 柱状图的形式，从而形成一个基于像素的柱状图可视化

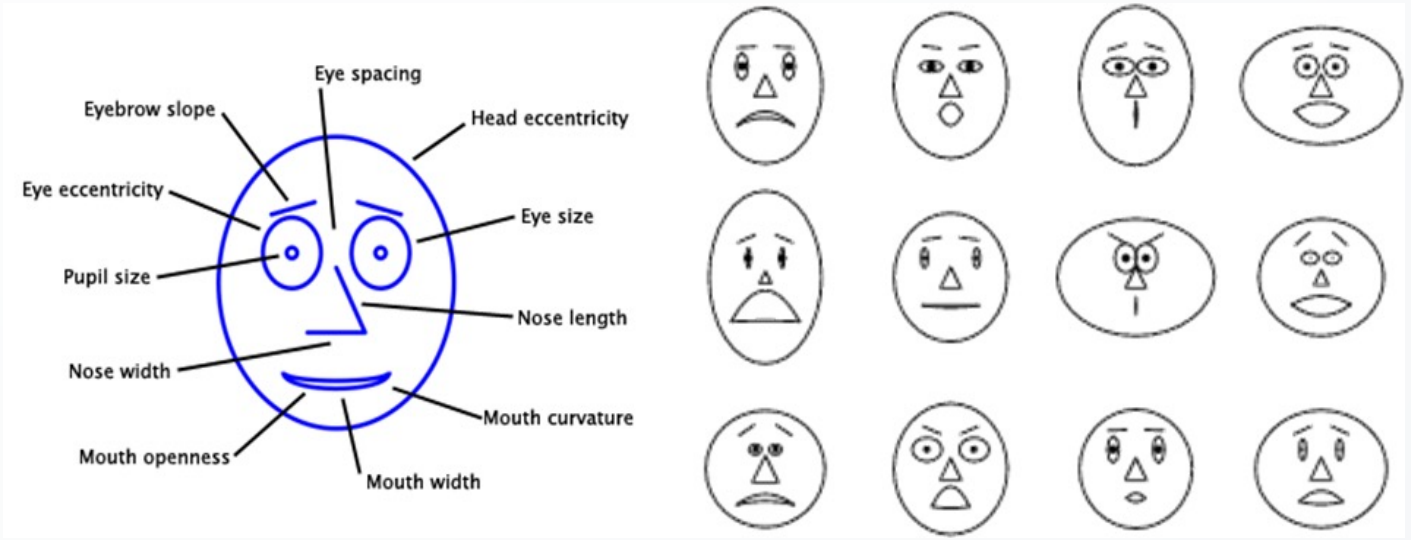
课程大纲

- 多维度数据的可视化
  - 基于不同坐标系的可视化方法(Coordinate Systems)
  - 基于像素的可视化方法 (Pixel Oriented)
  - 基于图标的可视化方法（Icon Based)
  - 基于网格的分视图展示方法（Small Multiple）
  - 多位度数据的可视化诊断
- 树的可视化
- 图的可视化

基于图标的可视化方法

- **切尔诺夫面孔**（Chernoff Faces）

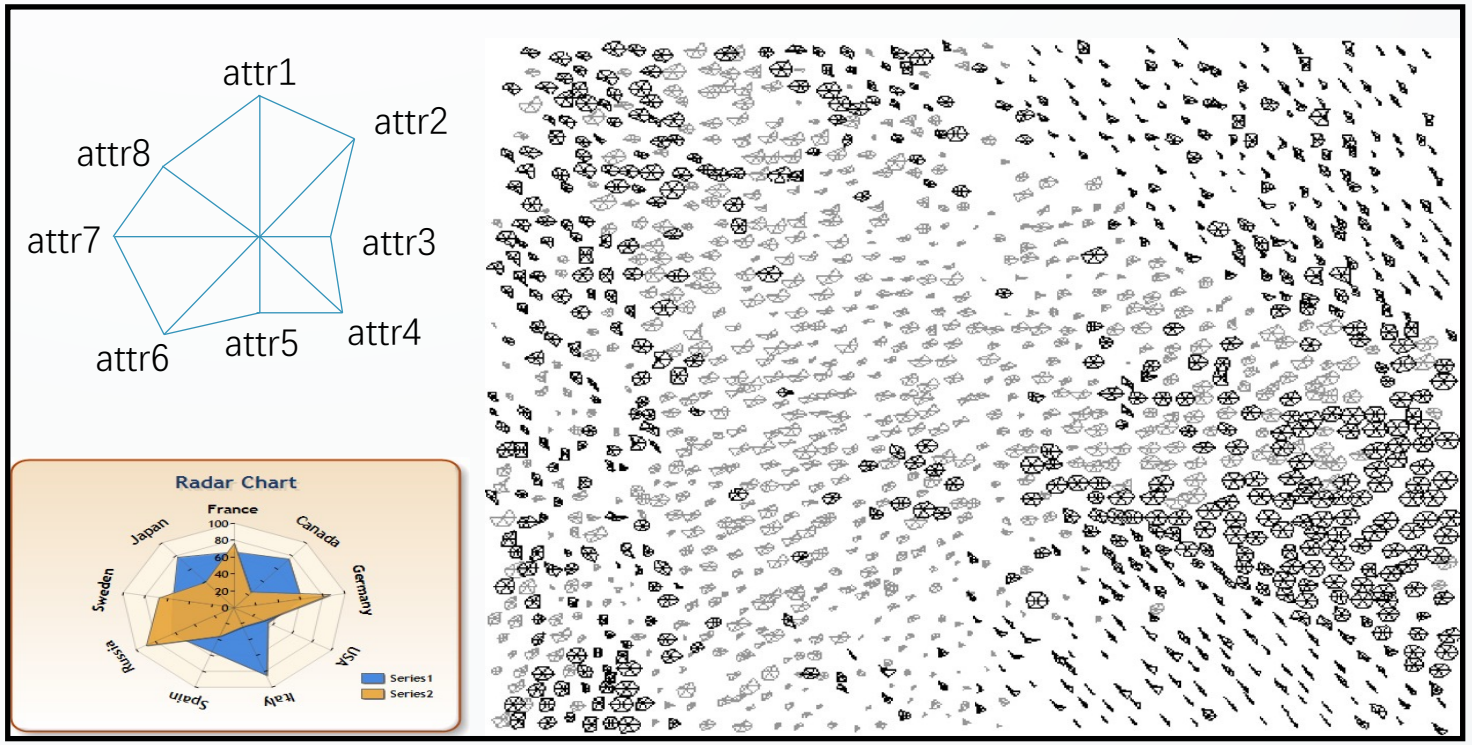
  - 一种利用人的五官特征来展示数据多维度属性的可视化方式
  - 每一个数据元素对应着一张面孔,面孔中眼睛的大小，嘴巴的宽窄、鼻子的高低等五官特征都对应着该数据元素在特定维度上的属性值的大小
  - 利用了人们能够快速辨识不同面孔特征的能力
  - 但是所形成的面部表情 往往与数据无关，导致语义上的误导



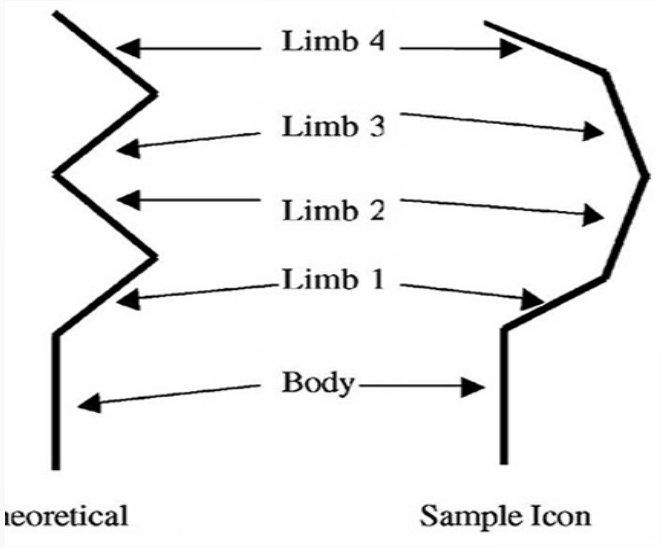
(Chernoff, 1973)

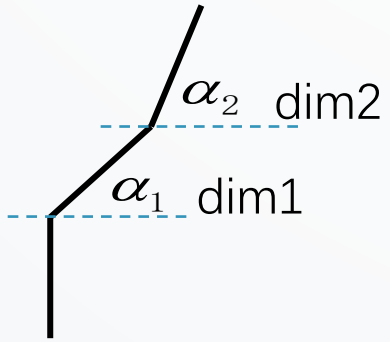
基于图标的可视化方法

- **星形图标**（Star Glyph）
  - 每一个数据元素对应着一个图标
  - 数据的维度被显示为围绕着中心原点的数据轴
  - 数据元素在不同维度上的取值构成了图标的外轮廓形状
  - 当独立重叠使用时，便构成了雷达图；当以图标形式集体使用时，便构成了能够凸显数据分布模式的可视化

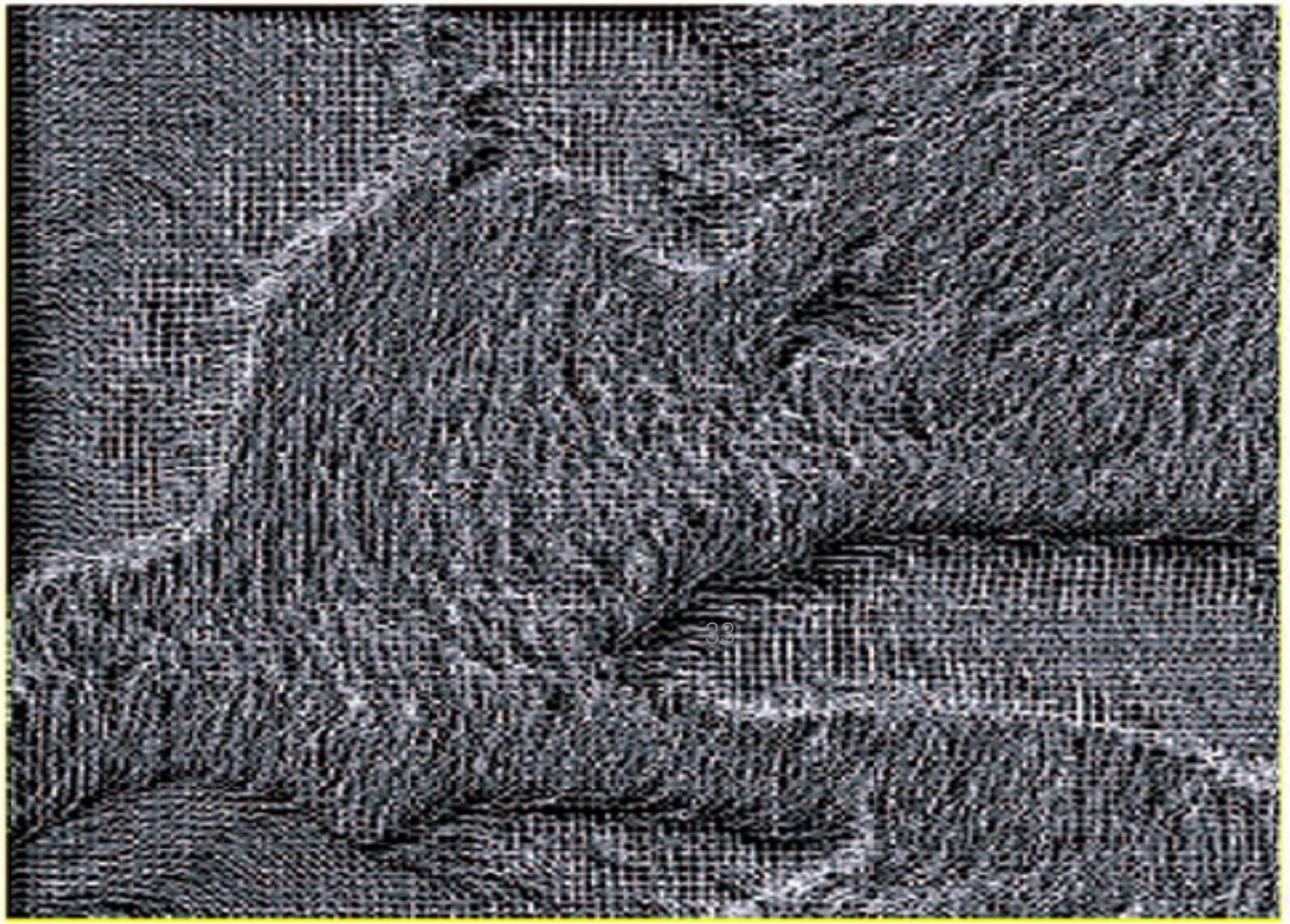


Icon Based Techniques – Stick Figure

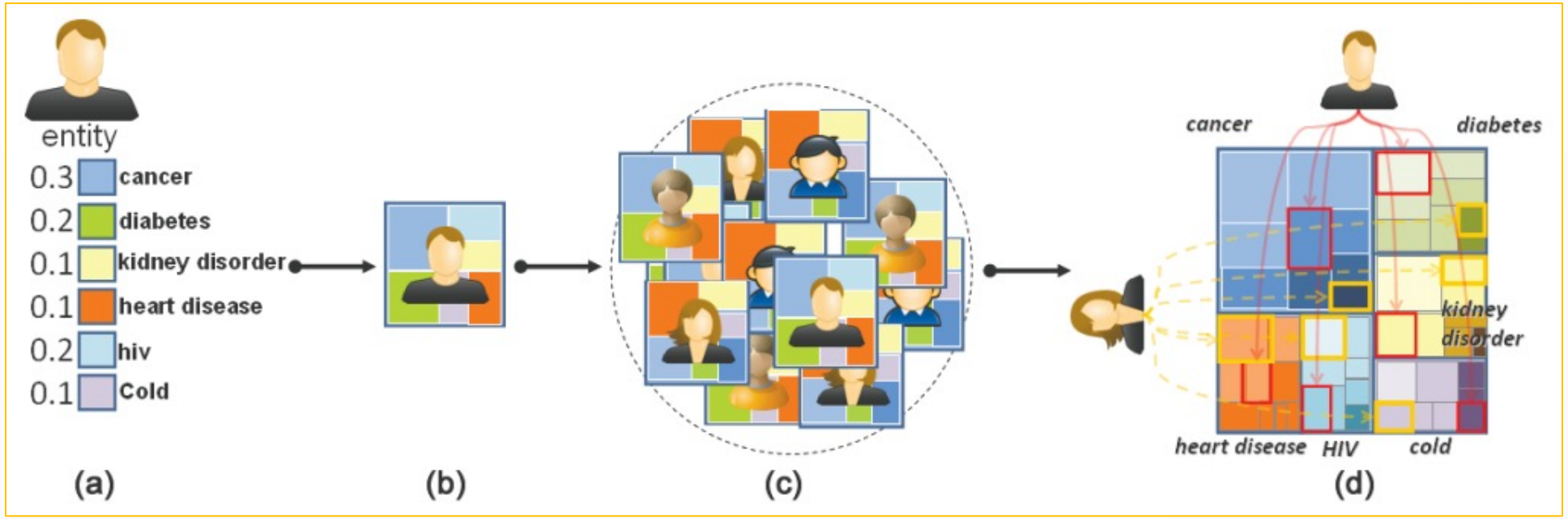




User angles of sticks
to encode dimensions



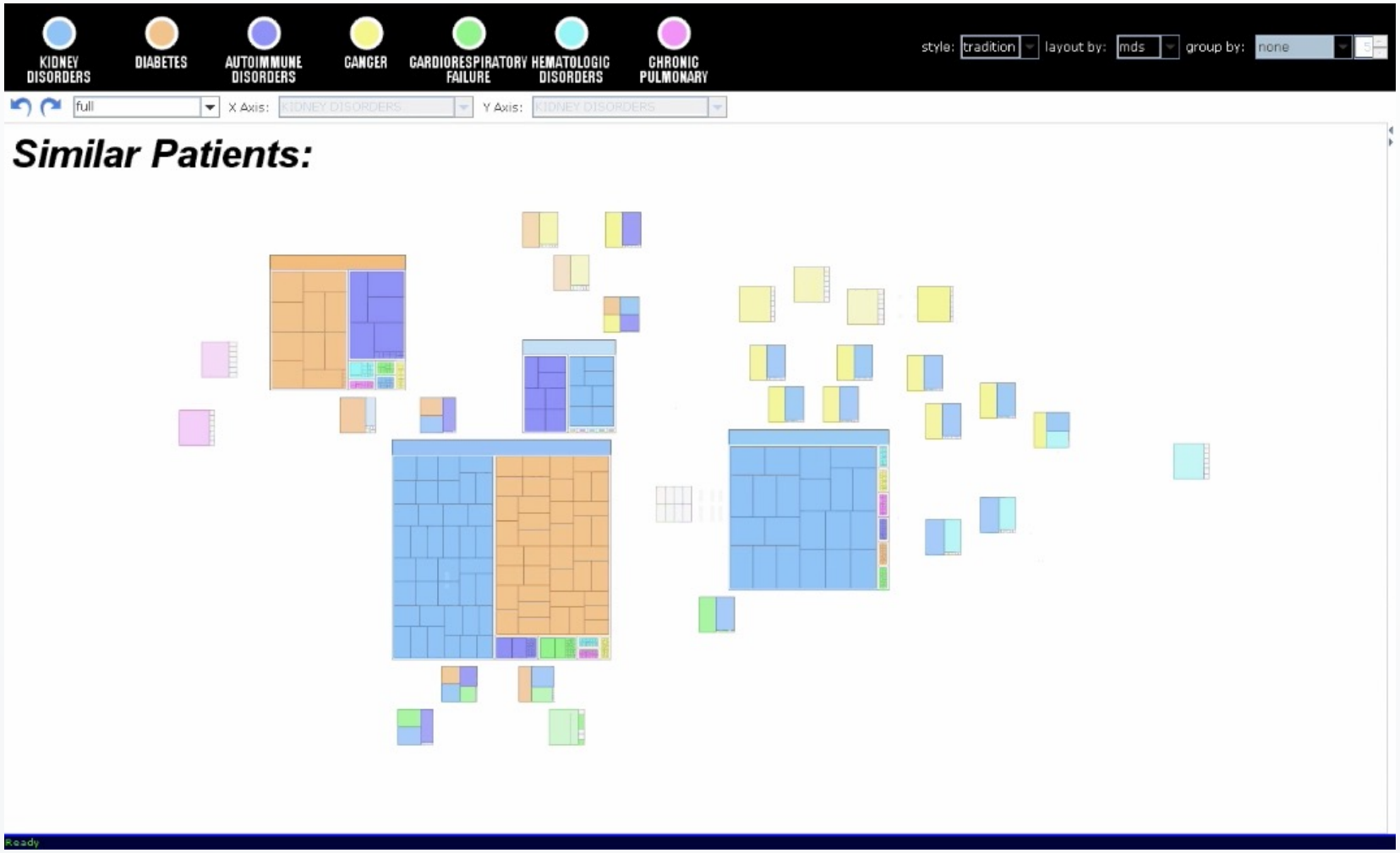
基于图标的可视化方法



- 动态图标技术（DICON）
  - 数据元素用方形图标表示
  - 图标中不同颜色的区域代表了不同的数据维度，区域的大小代表了所对应属性的取值
  - 展示多个数据元素时，可拆分单个数据元素的图标，并重组构成一个更大的代表数据集合图标
  - 重组的方式可以根据数据分组方式的不同进行动态调整

基于图标的可视化方法

- 动态图标技术（DICON）

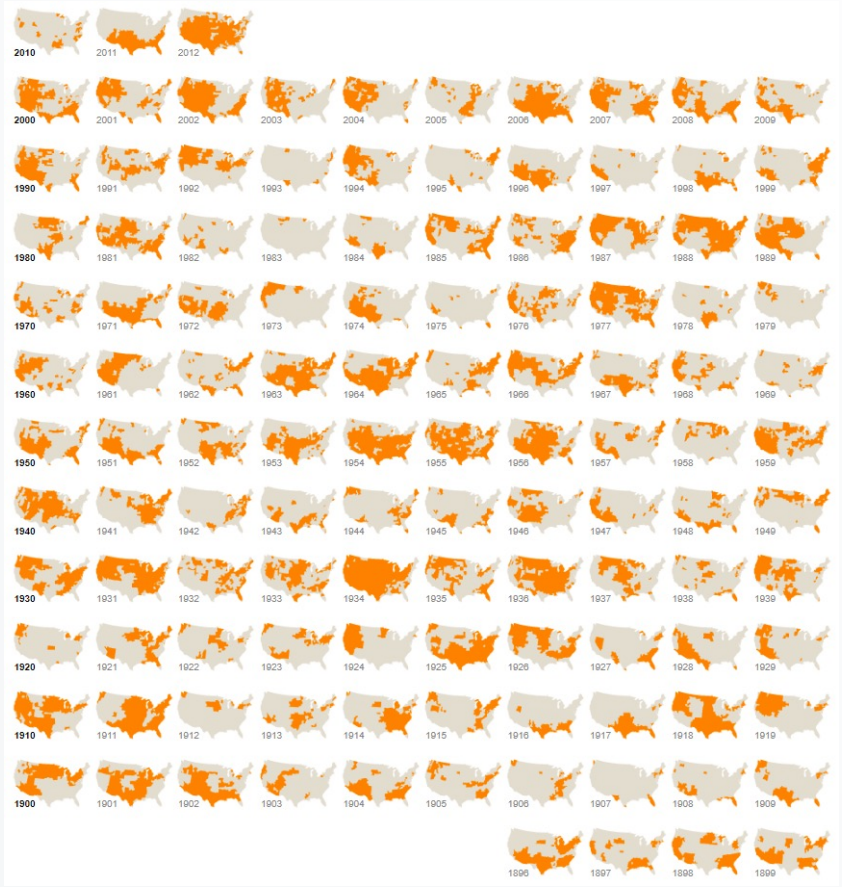


课程大纲

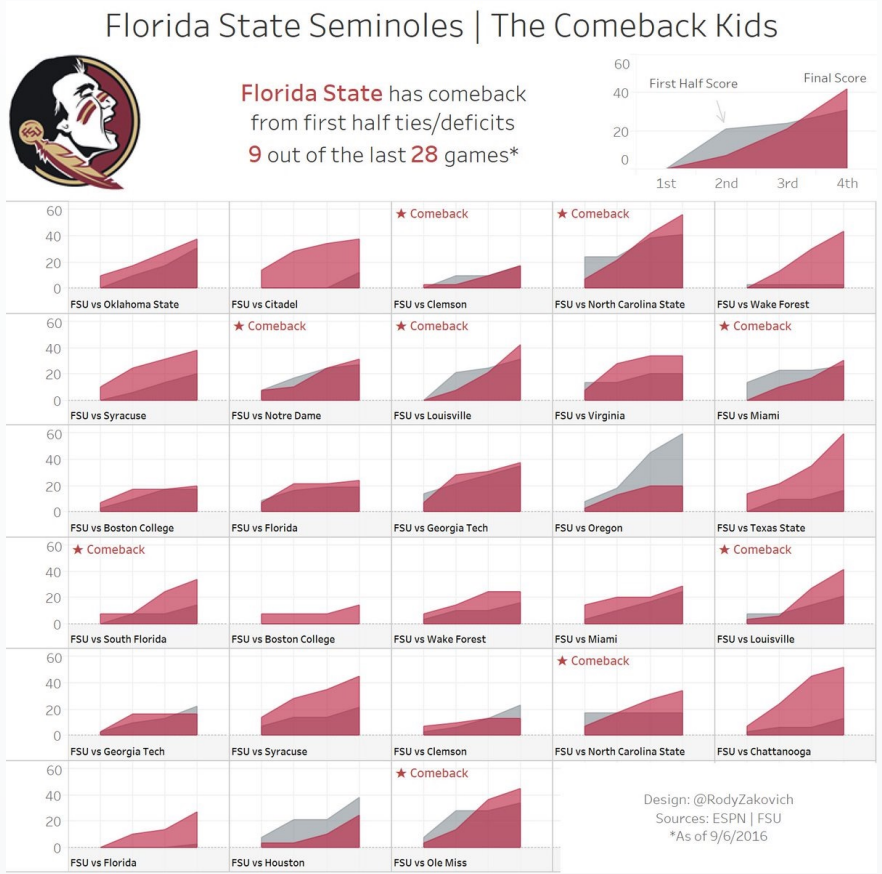
- 多维度数据的可视化
  - 基于不同坐标系的可视化方法(Coordinate Systems)
  - 基于像素的可视化方法 (Pixel Oriented)
  - 基于图标的可视化方法（Icon Based)
  - 基于网格的分视图展示方法（Small Multiple)
  - 多维度数据的可视化诊断
- 树的可视化
- 图的可视化

基于网格的分视图展示方法

- 在低维度可视化的基础上，按照其他数据进行划分，行程矩阵式的展显现方式，增加信息呈现维度



特定数值在空间上的分布，并按时间进行视图划分



按照不同的比赛场次划分展现两只队伍的表现

按照地理空间进行划分

课程大纲

- 多维度数据的可视化
  - 基于不同坐标系的可视化方法(Coordinate Systems)
  - 基于像素的可视化方法 (Pixel Oriented)
  - 基于图标的可视化方法（Icon Based)
  - 基于网格的分视图展示方法（Small Multiple）
  - 多维度数据的可视化诊断
- 树的可视化
- 图的可视化

可视化视图诊断 （Visual Diagnostics）

- Visual diagnostics:
  - Estimates multidimensional visualization views based on a set of measures
  - Recommends the views with interesting visual patterns based on their measurements

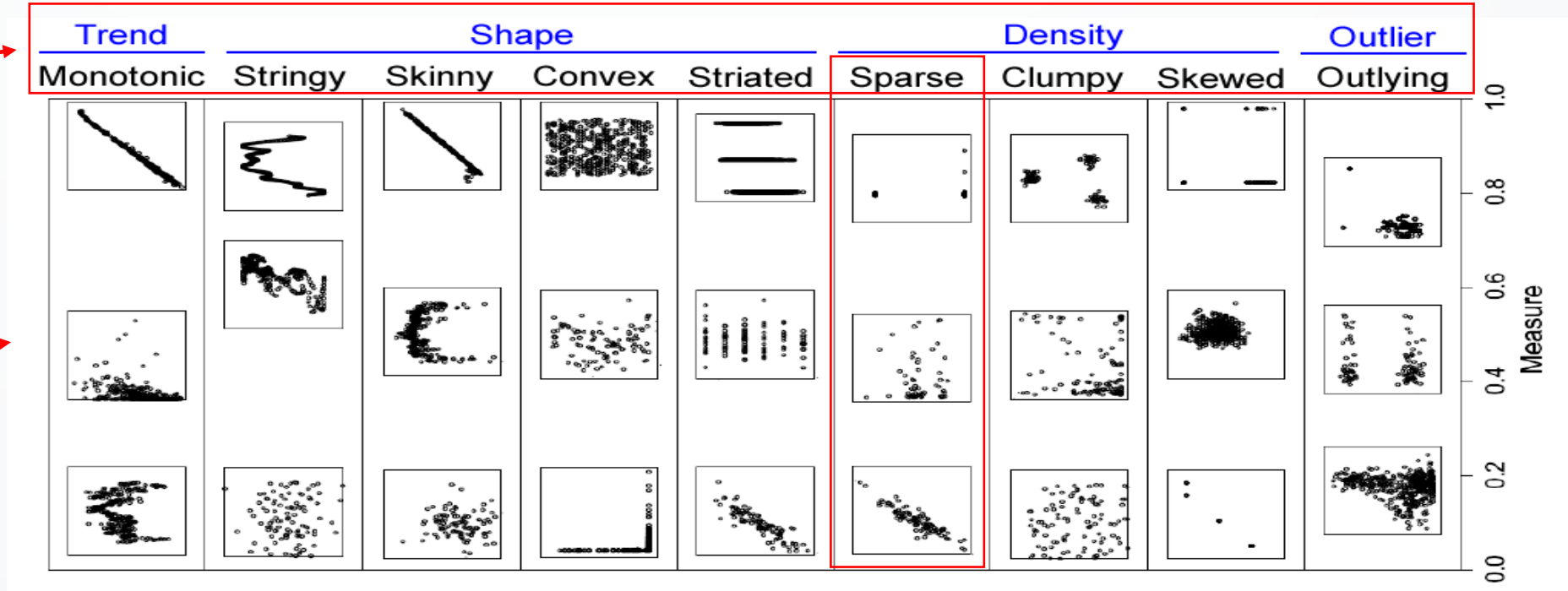
- Applications:
  - Scagnostics (scatter plots + diagnostics)
  - Pixnostics (pixel + diagnostics)
  - Pragnostics (parallel coordinates + diagnostics)

Scagnostics (Scatter plots + Diagnostics)

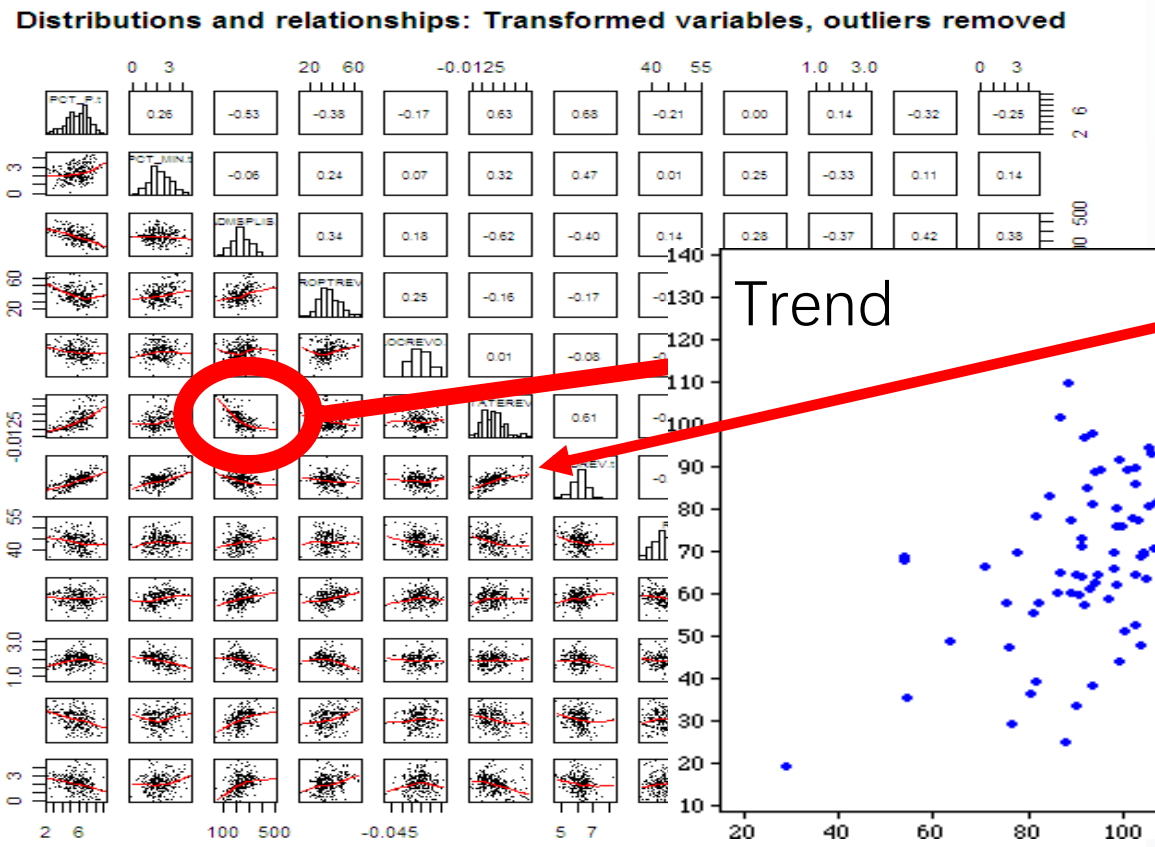
- Scatter plots diagnostics estimates scatter plot views based on a set of pre-defined measurements
- Views are ranked based on their measurements

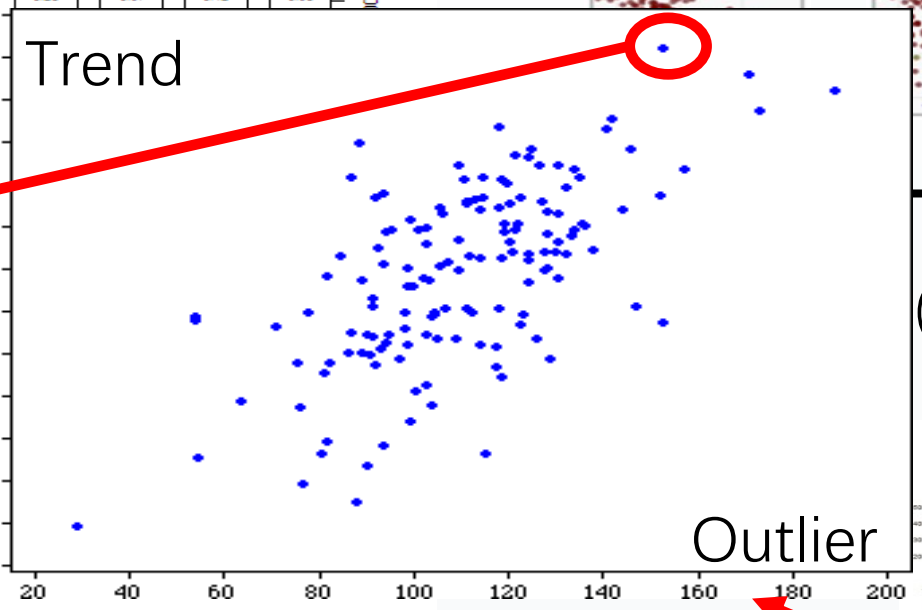
View Measurements

Scatter Plot Views



Scagnostics (Scatter plots + Diagnostics)

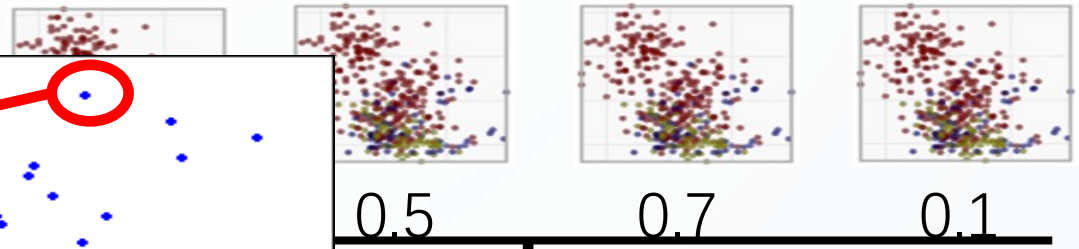




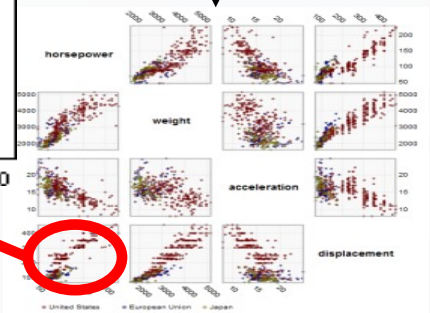
P = n (n – 1) / 2 views for an n-dimensional dataset

Measurements of scatter plot views

Trend | Shape | Density | Outlier

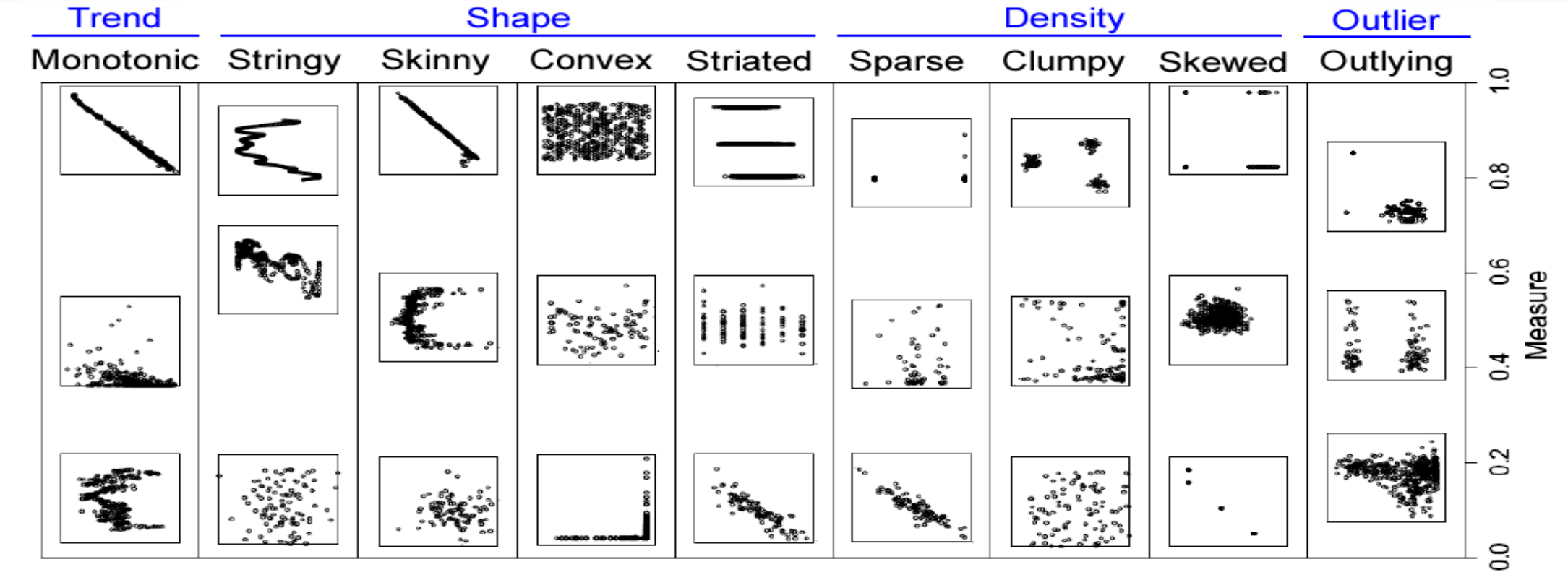


(0.1, 0.5, 0.7, 0.1)

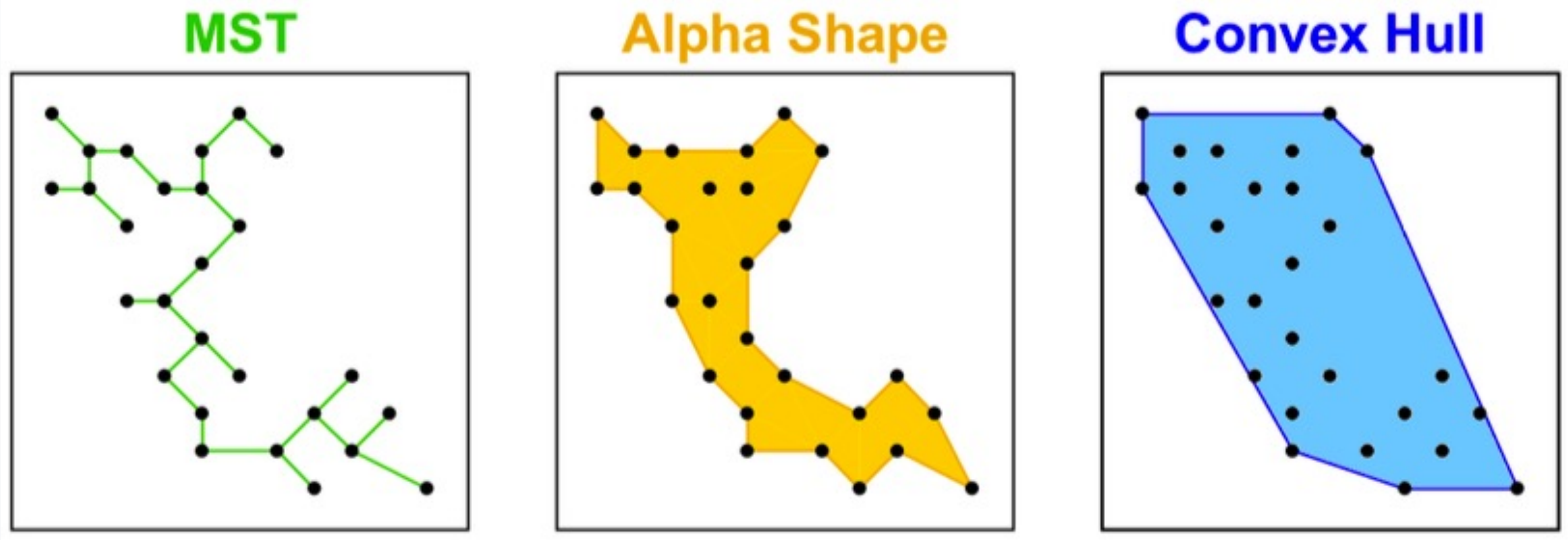


scatter plot matrices in a much smaller scale

Measurements for Scatter Plots

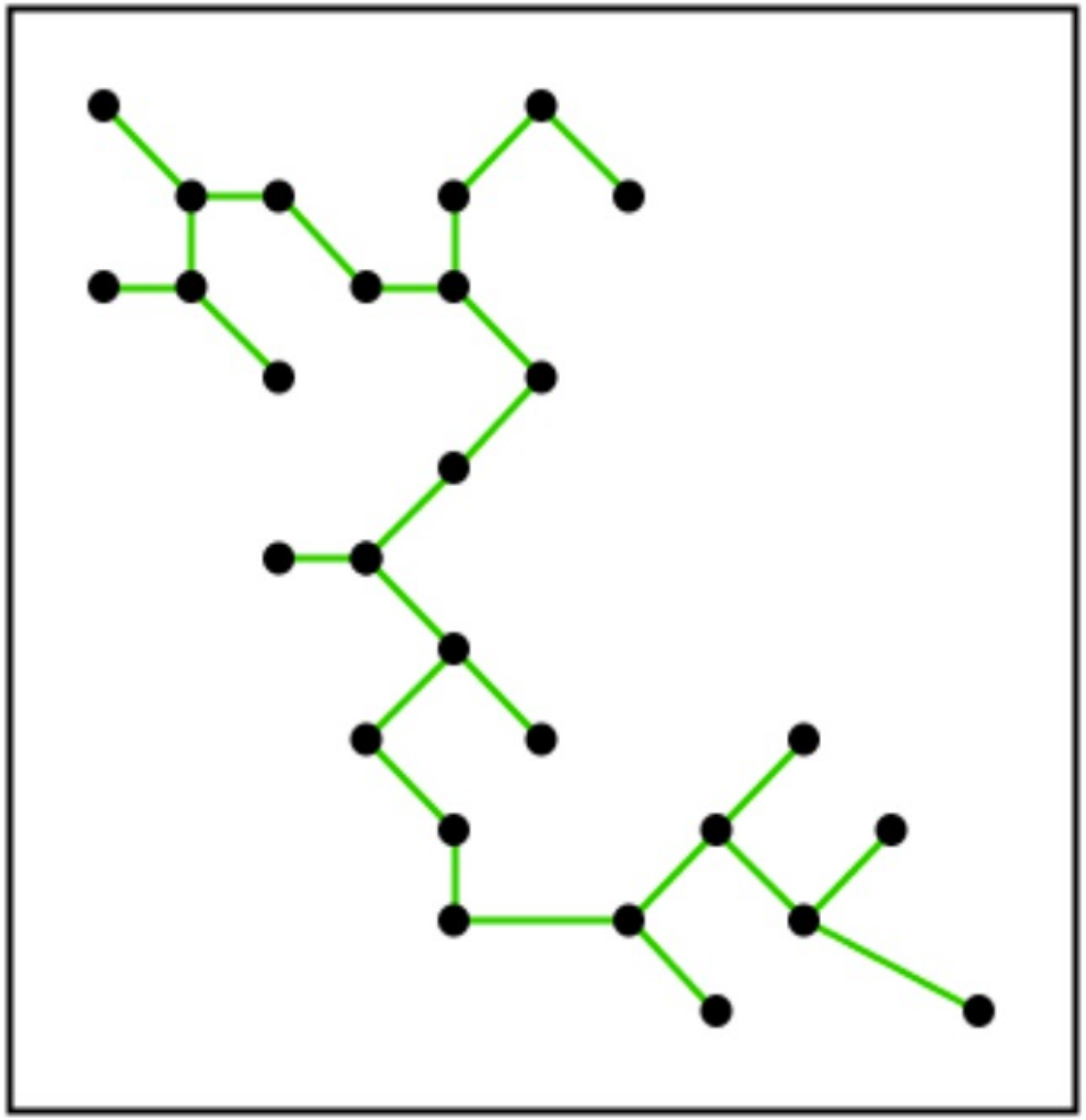


Proximity Graphs



Features are computed based on three types of graphs derived from the scatter plot

MST

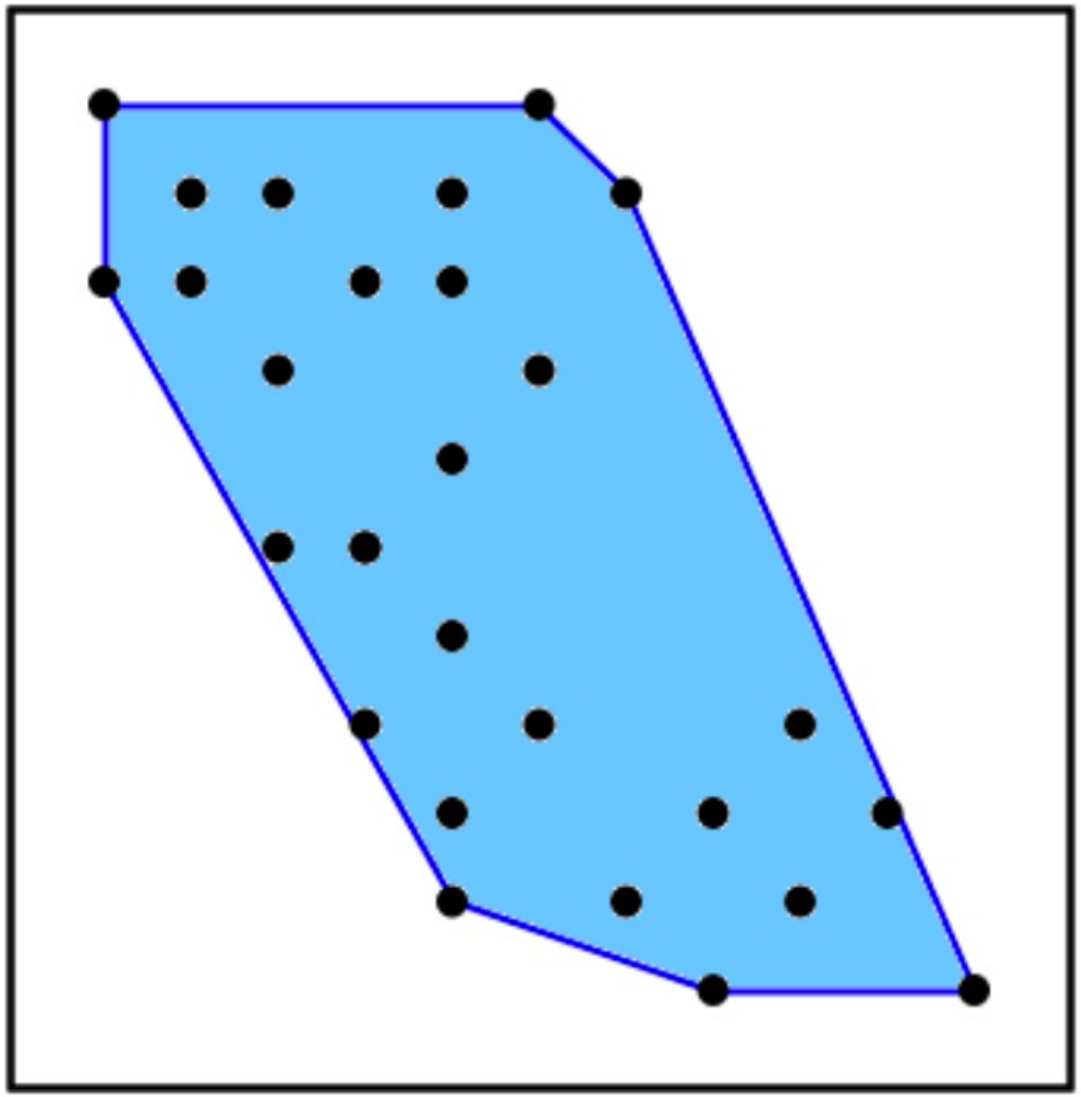


- A path is a list of successively adjacent, distinct edges
- A tree is a graph in which any two nodes are connected by exactly one path
- A spanning tree is an undirected graph whose edges are structured as a tree
- A minimum spanning tree (MST) is a spanning tree whose total length (sum of edge weights) is least of all spanning trees on a given set of points
- The edge weights (edge lengths) of a geometric MST are computed from the distances between its vertices.

Alpha Shape

- A proximity graph (or neighborhood graph) is a geometric graph whose edges are determined by an indicator function based on distances between a given set of points in a metric space.
- To define this indicator function, we use an open disk D.
  - D touches a point if that point is on the boundary of D
  - D contains a point if that point is in D
  - Denote an open disk of fixed radius D(r)
- In an alpha shape graph, an edge exists between any pair of points that can be touched by an open disk D(r) containing no points

Convex Hull



- A hull of a set of points X in R^2 is a collection of the boundaries of one or more polygons that have a subset of the points in X for their vertices and that collectively contain all the points in X
- A hull is convex if it contains all the straight line segments connecting any pair of points in its interior

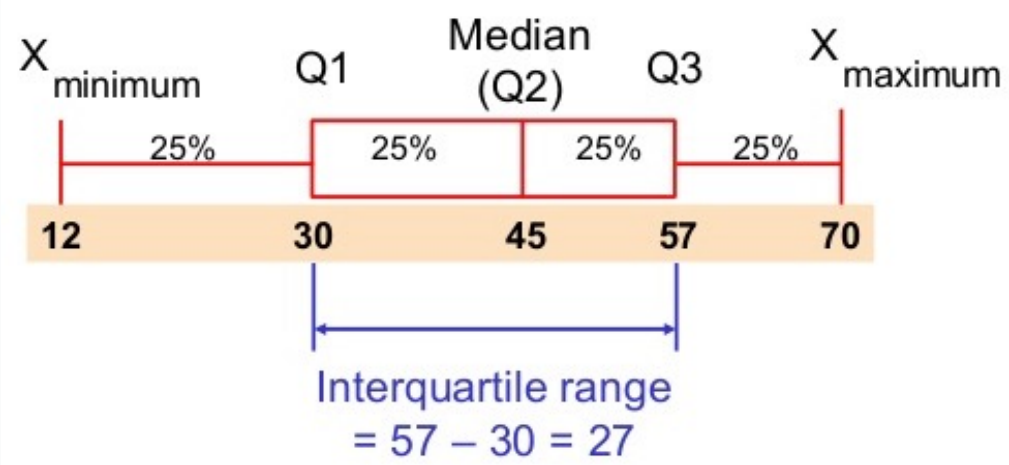
Measurements

- The length of an edge, i.e., the Euclidean distance between its vertices
- The length of a graph, i.e., is the sum of the lengths of its edges
- A path is a list of vertices such that all pairs of adjacent vertices in the list are edges
- A path is closed if its first and last vertex are the same
- A closed path is the boundary of a polygon
- The perimeter of a polygon is the length of its boundary
- The area of a polygon is the area of its interior

Outlying

- We consider an outlier to be a vertex whose adjacent edges in the MST all have a weight (length) greater than w

$$\omega = q_{75} + 1.5(q_{75} - q_{25})$$



- A measure of the proportion of the total edge length due to extremely long edges connected to points of single degree:

$$c_{outlying} = length(T_{outliers})/length(T)$$

Density

- **Skewed**: relatively robust measure of skewness in the distribution of edge lengths

$$c_{skew} = (q_{90} - q_{50})/(q_{90} - q_{10})$$

- **Clumpy**: clusters with small intracluster distances relative to the length of their connecting edge and ignores runt clusters with relatively small runt size:

$$c_{clumpy} = \max_j \left[1 - \max_k [length(e_k)]/length(e_j)\right]$$

Density

- **Sparse**: sparseness statistic measures whether points in a 2D scatterplot are confined to a lattice or a small number of locations on the plane

$$c_{sparse} = q_{90}$$

- **Striated:** We define coherence in a set of points as the presence of relatively smooth paths in the minimum spanning tree.

$$c_{striate} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(\cos \theta_{e(v,a)e(v,b)} < -.75)$$

Shape

- **Convex.** Our convexity measure is based on the ratio of the area of the alpha hull and the area of the convex hull. This ratio will be 1 if the nonconvex hull and the convex hull have identical areas:

$$c_{convex} = area(A)/area(H)$$

- **Skinny.** The ratio of perimeter to area of a polygon measures, roughly, how skinny it is.

$$c_{skinny} = 1 - \sqrt{4\pi area(A)}/perimeter(A)$$

- **Stringy**. A stringy shape is a skinny shape with no branches.

$$c_{stringy} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}$$

# of vertices of degree 2 in MST

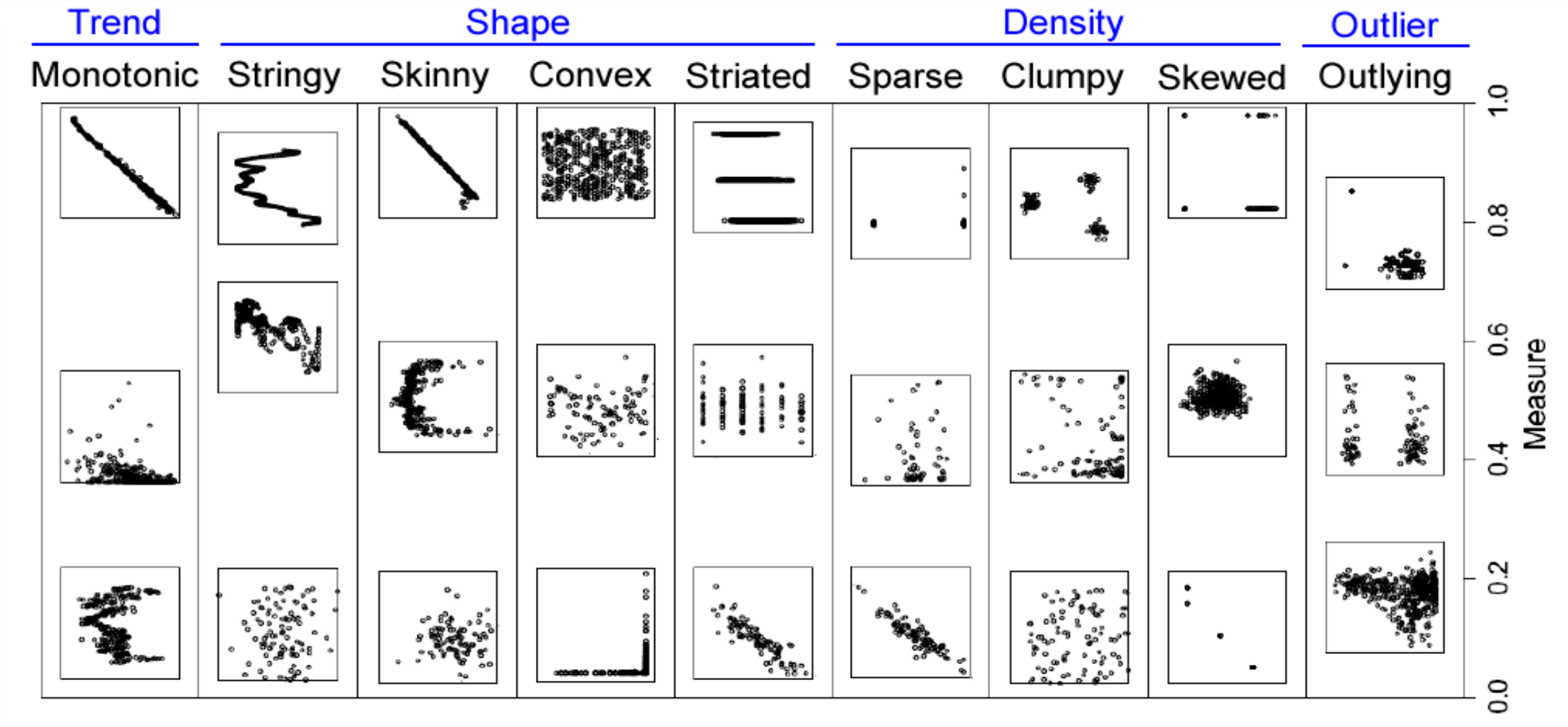
# of vertices - # of single-degree vertices

Association

- Monotonic: the squared Spearman correlation coefficient, which is a Pearson correlation on the ranks of x and y (corrected for ties), to assess monotonicity in a scatterplot

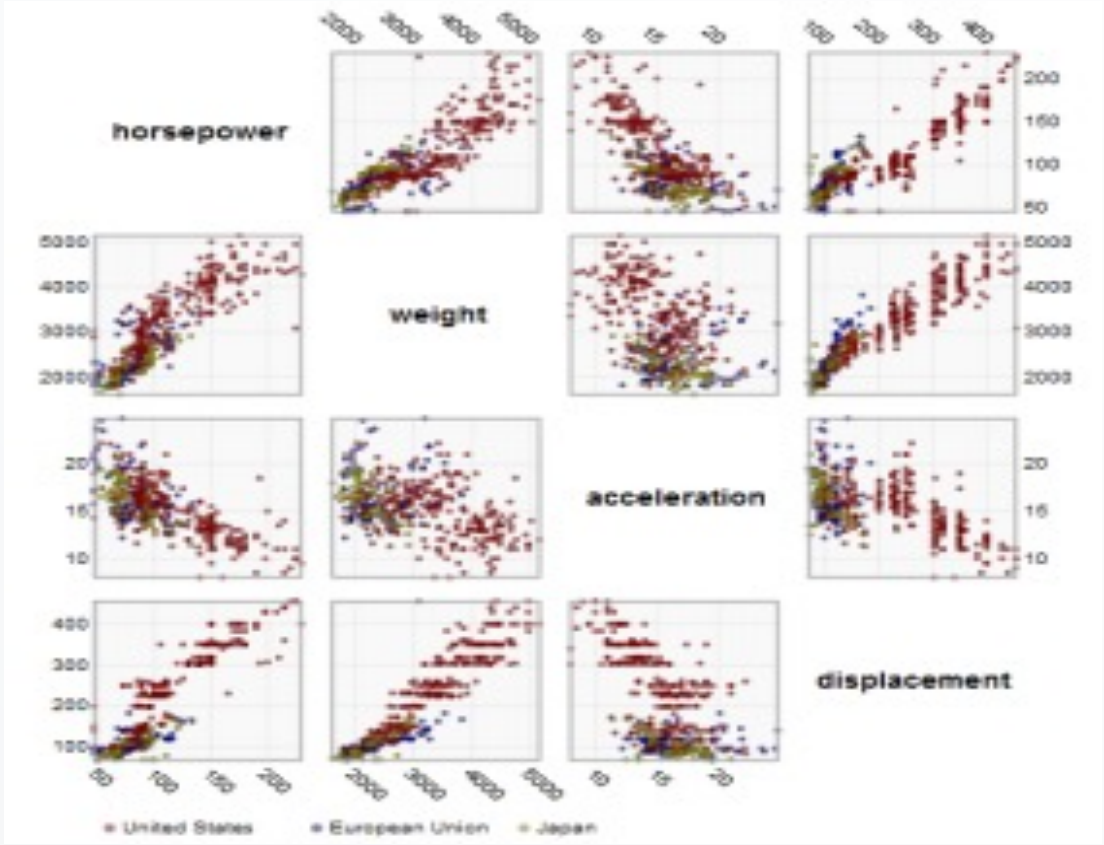
$$c_{monotonic} = r^2_{spearman}$$

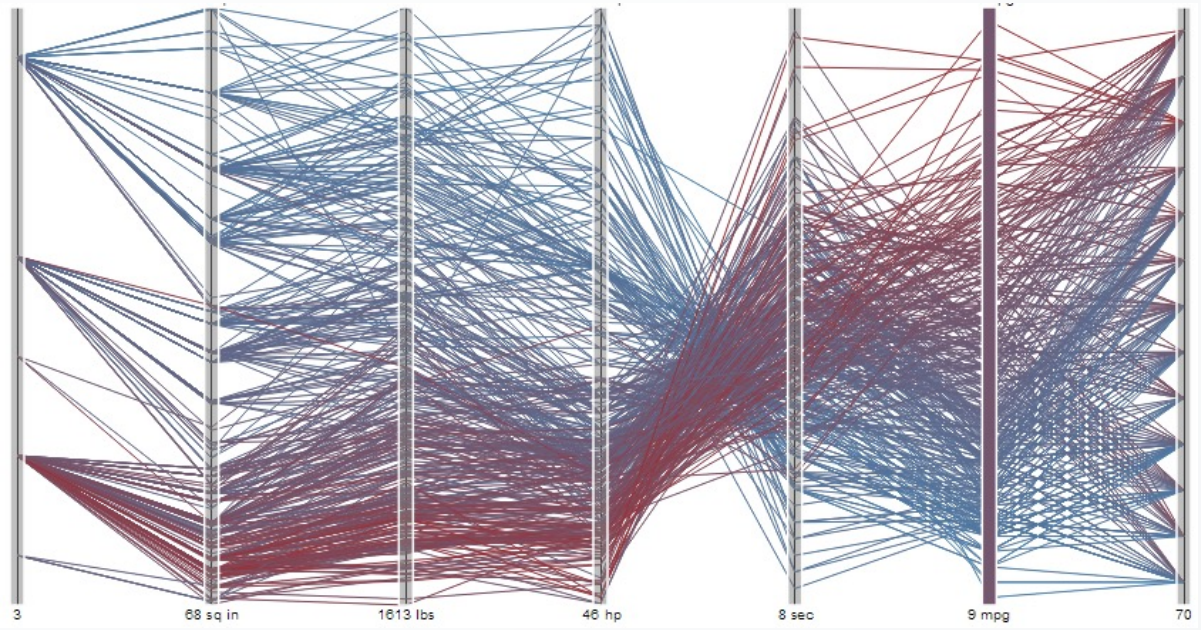
Measurements for Scatter Plots



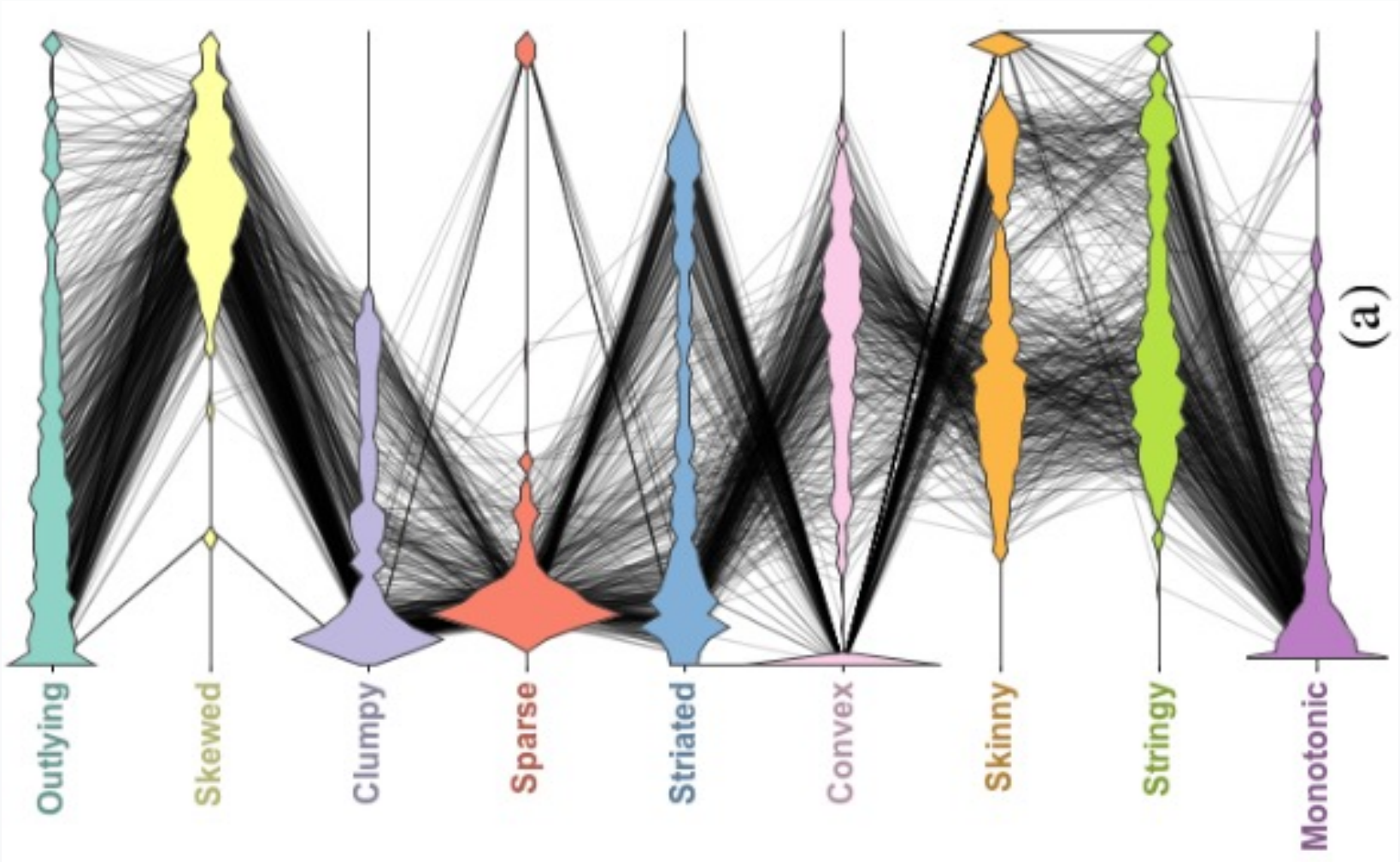
Use 10 minutes to design a visualization to represent the scagnostics results

Visualized in another SPM or PCP

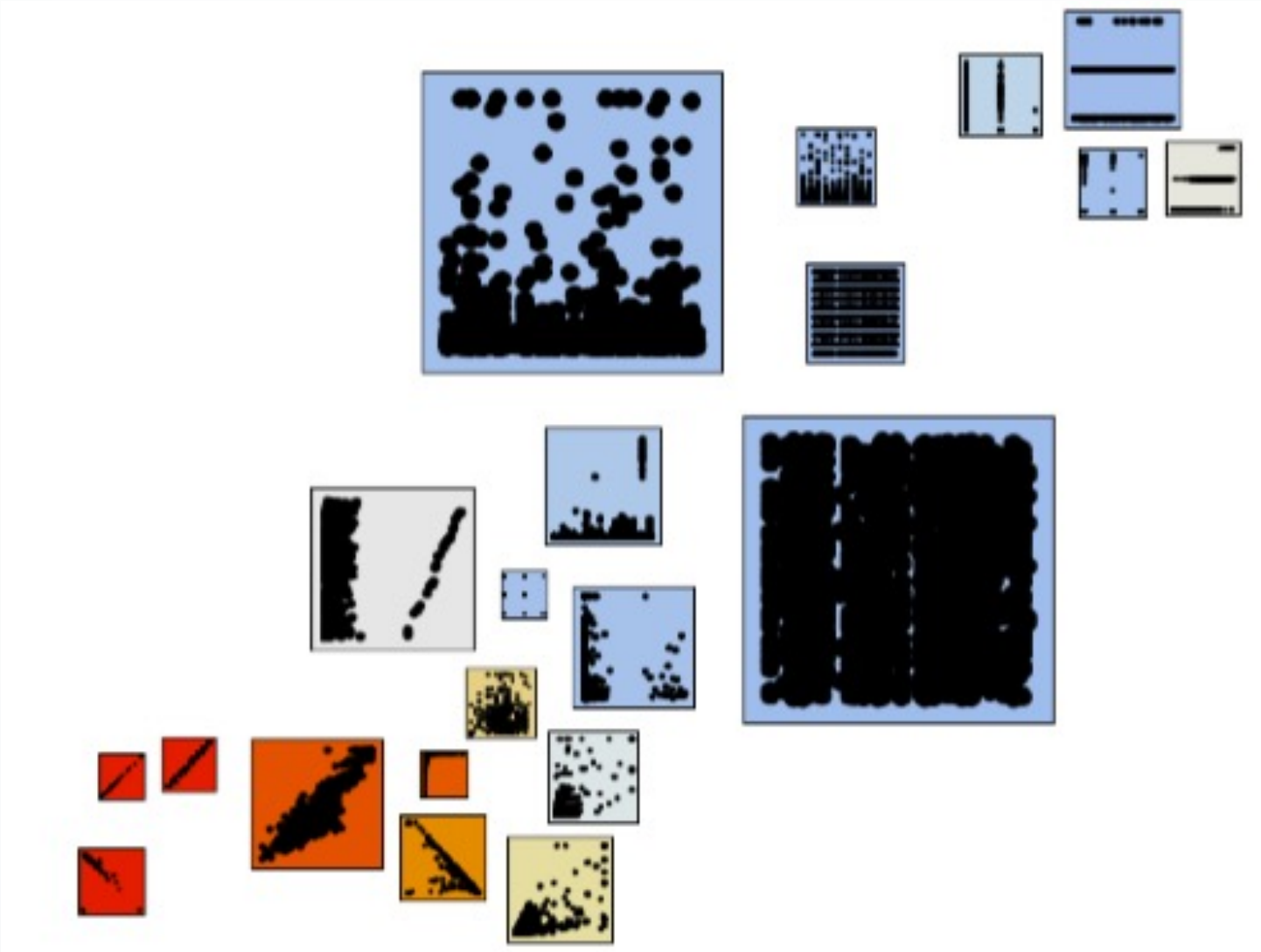




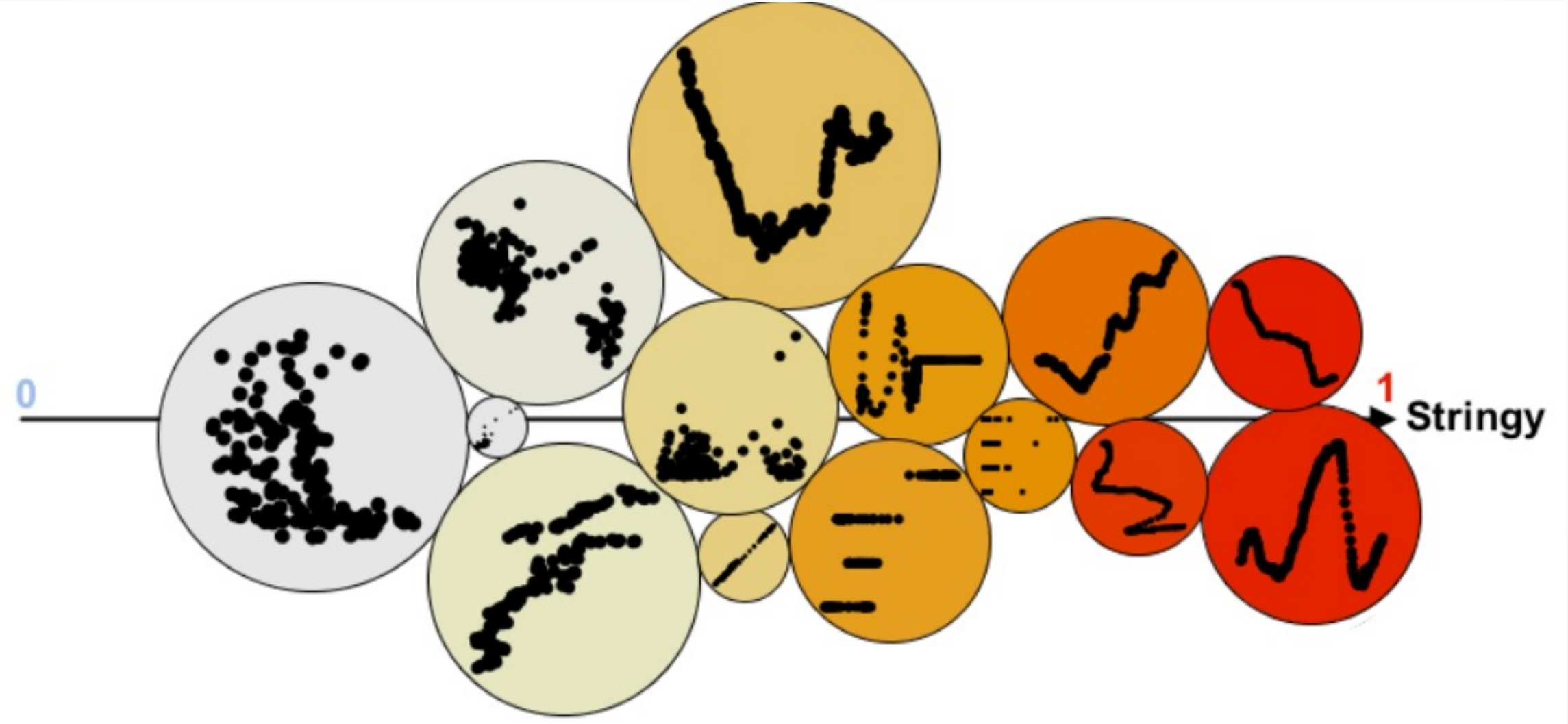
Is there a better approach ?



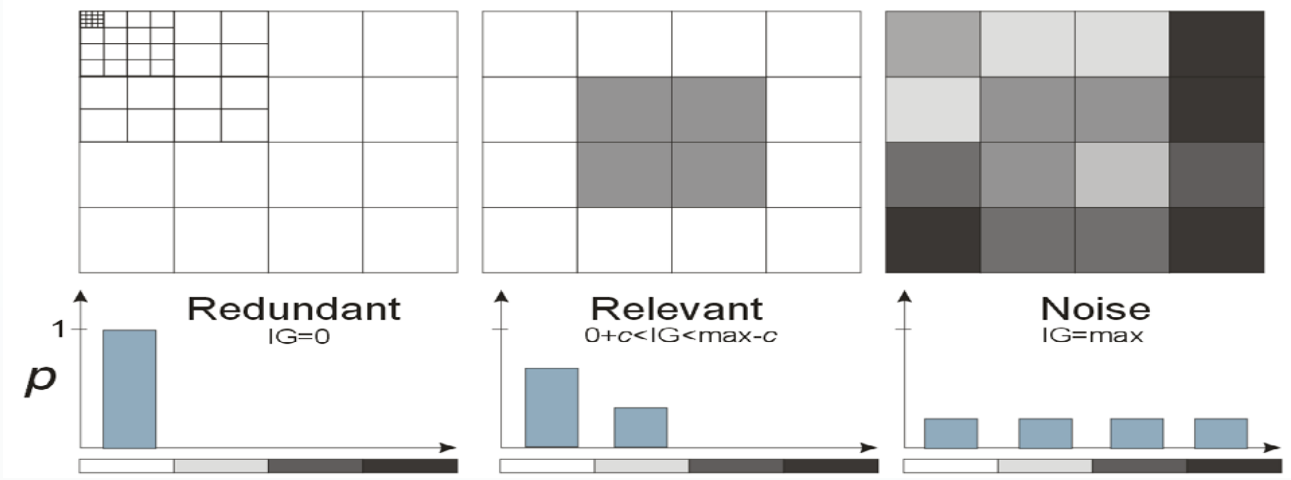
Is there a better approach ?



Is there a better approach ?

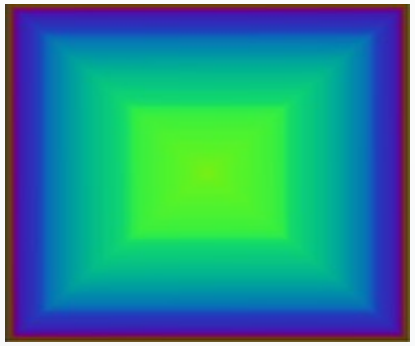


Pixnostics (pixel + diagnostics)

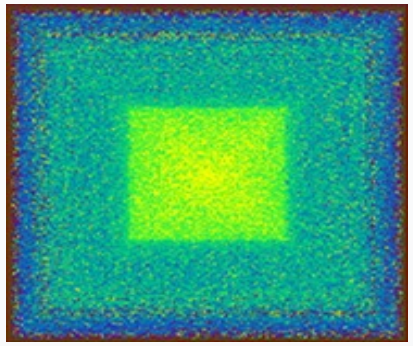


examples of pixel measures

Relevant score:



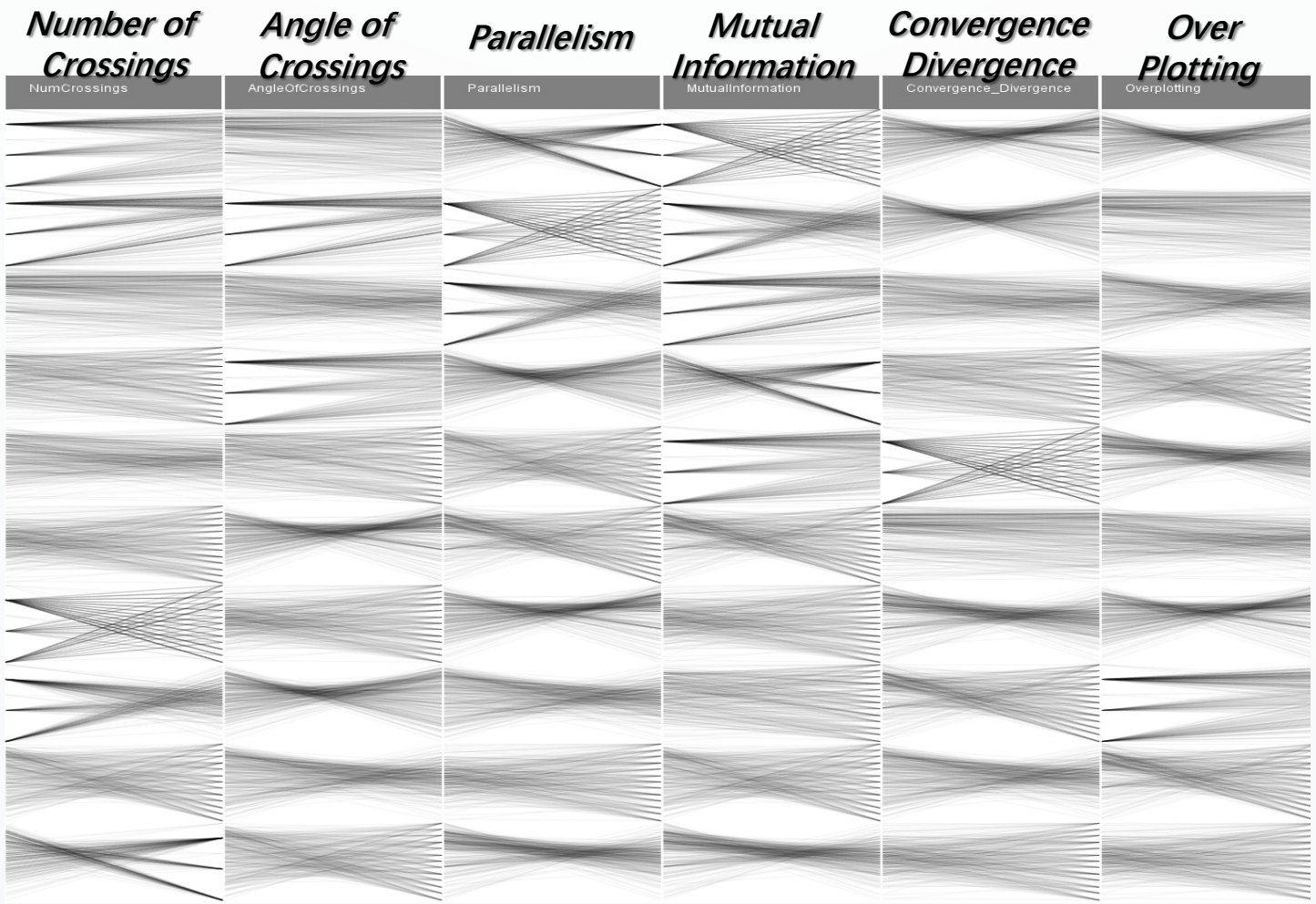
>



0.8 0.5

(schneidewind et al., 2006)

Pragnostics (parallel coordinates + diagnostics)



Ranking the order of dimension pairs
(Dasgupta & Kosara, 2010)