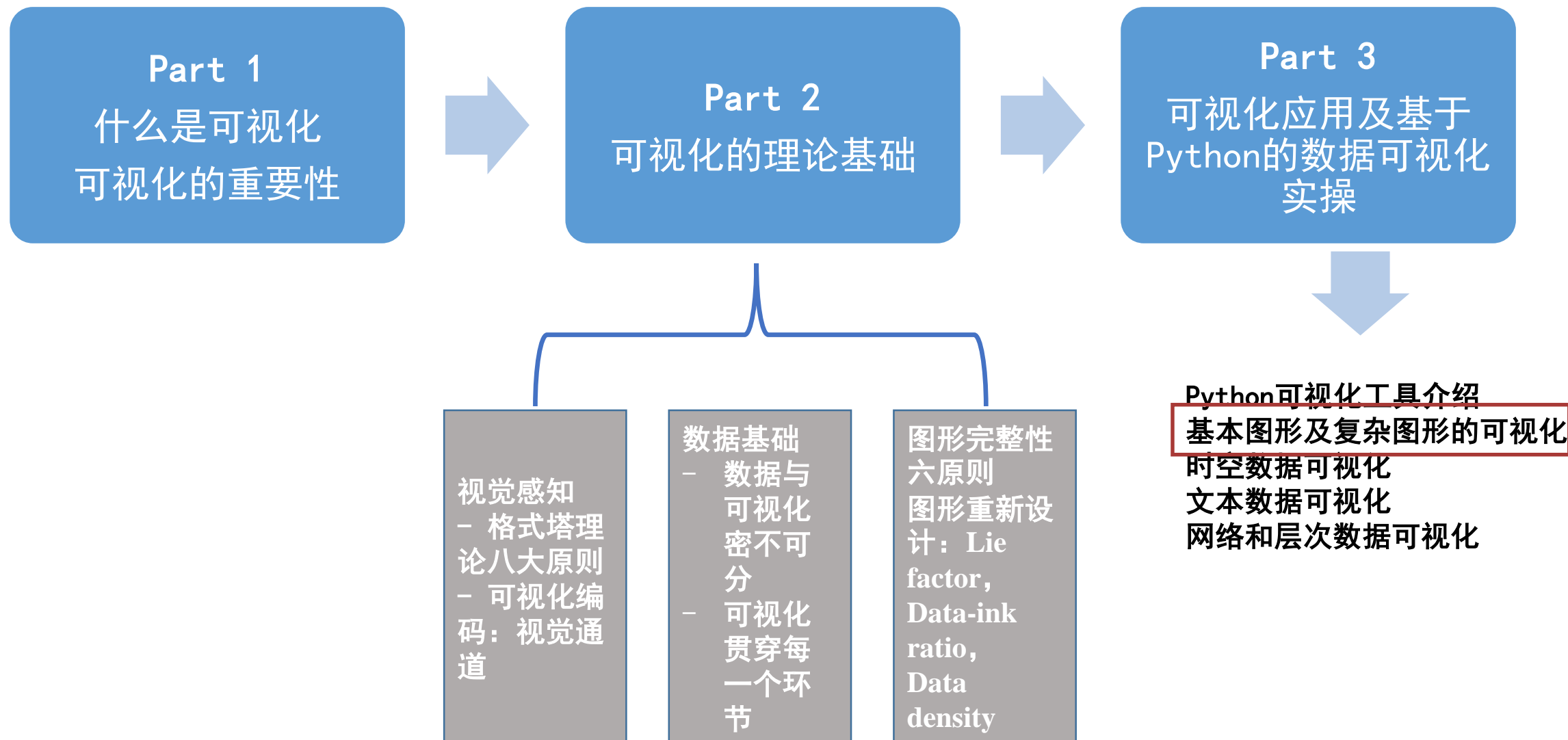




数据可视化

第六讲：可视化方法-高维多元数据

Our structure

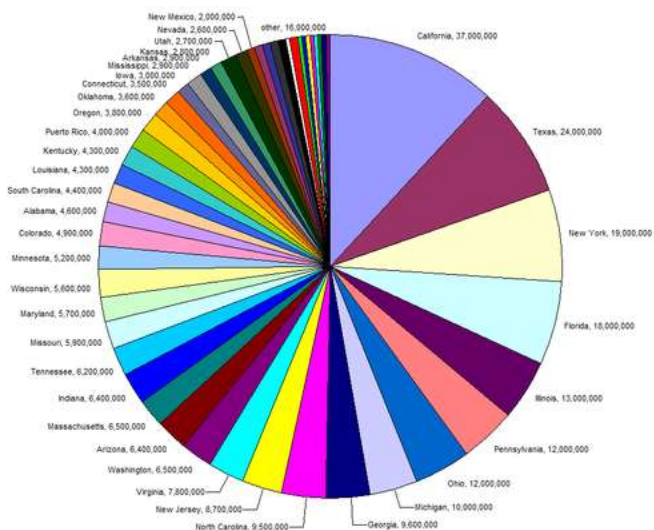


数据类型

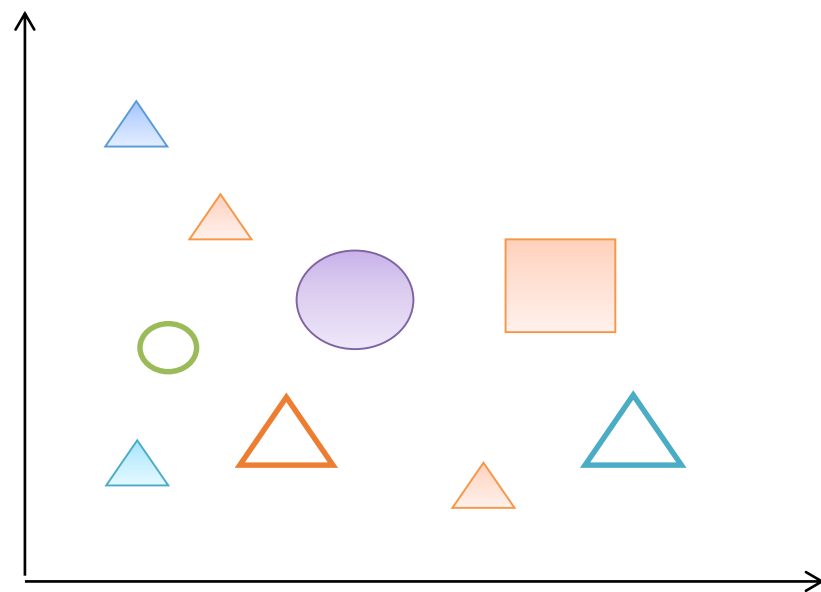
- 数值型数据
 - 连续型 (0.1, 0.11, 2.34, ...)
 - 离散型 (1, 2, 3...)
- 类别型数据
 - 有序的 (星期一, 星期二)
 - 无序的 (红蓝绿)
- 结构化数据
 - 时间
 - 空间
- 无结构化或半结构化数据
 - 文本数据

高维多元数据可视化

- 统计和基本分析为主的可视化系统分析能力不足
- 简单思路
 - 增加视觉通道，以表达更多的属性信息
 - 多视图协调关联



散点的形状、填充形式、颜色、大小等



高维数据可视化

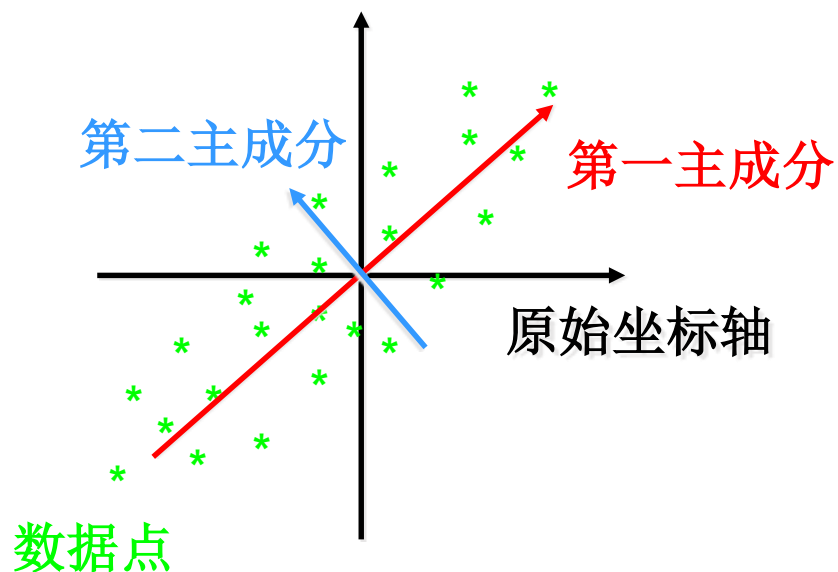
- 数据降维
- 可视化呈现方法
- 与数据交互

数据降维

- 线性方法
 - 主成分分析 (PCA)
 - 多维尺度分析 (MDS): 基于数据集相似程度的降维方法
 - 如果将数据点的相似性定义为数据点之间的欧式距离, 则MDS等价于PCA
 - `from sklearn.manifold import MDS`
 - 非负矩阵分解
- 非线性方法
 - 局部线性嵌入 (LLE)
 - ISOMAP
 - T-SNE (计算速度慢)
 - sklearn包中的tsne (函数)
 - `from sklearn.manifold import TSNE`

主成分分析 (PCA)

- 最小化投影后的损失
- 最大化投影后的方差
- sklearn包中的主成分分析函数PCA (`from sklearn.decomposition import PCA`)

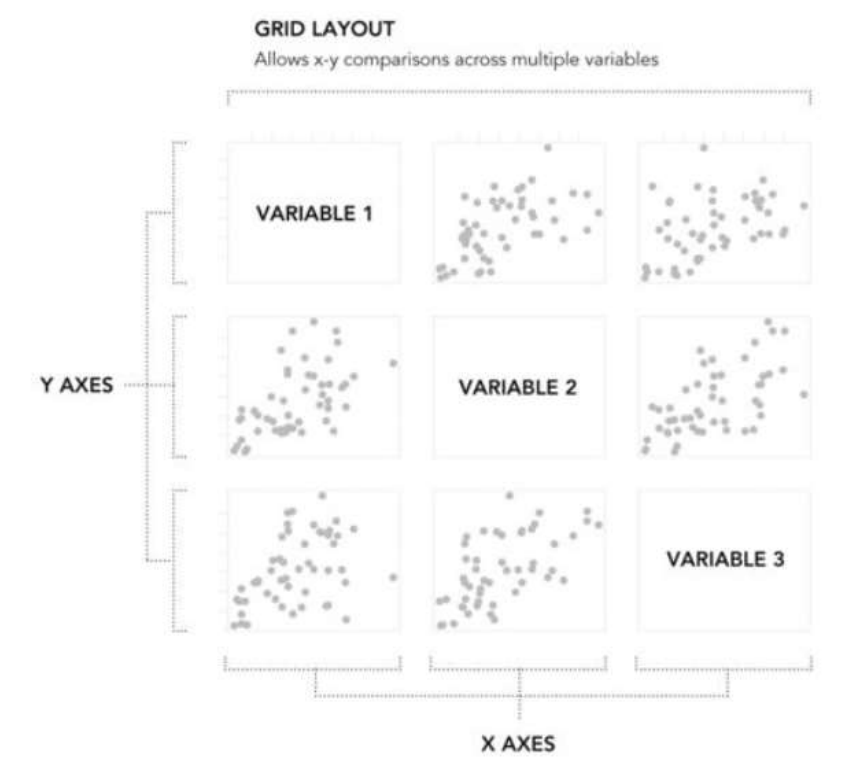


高维多元数据可视化呈现方法

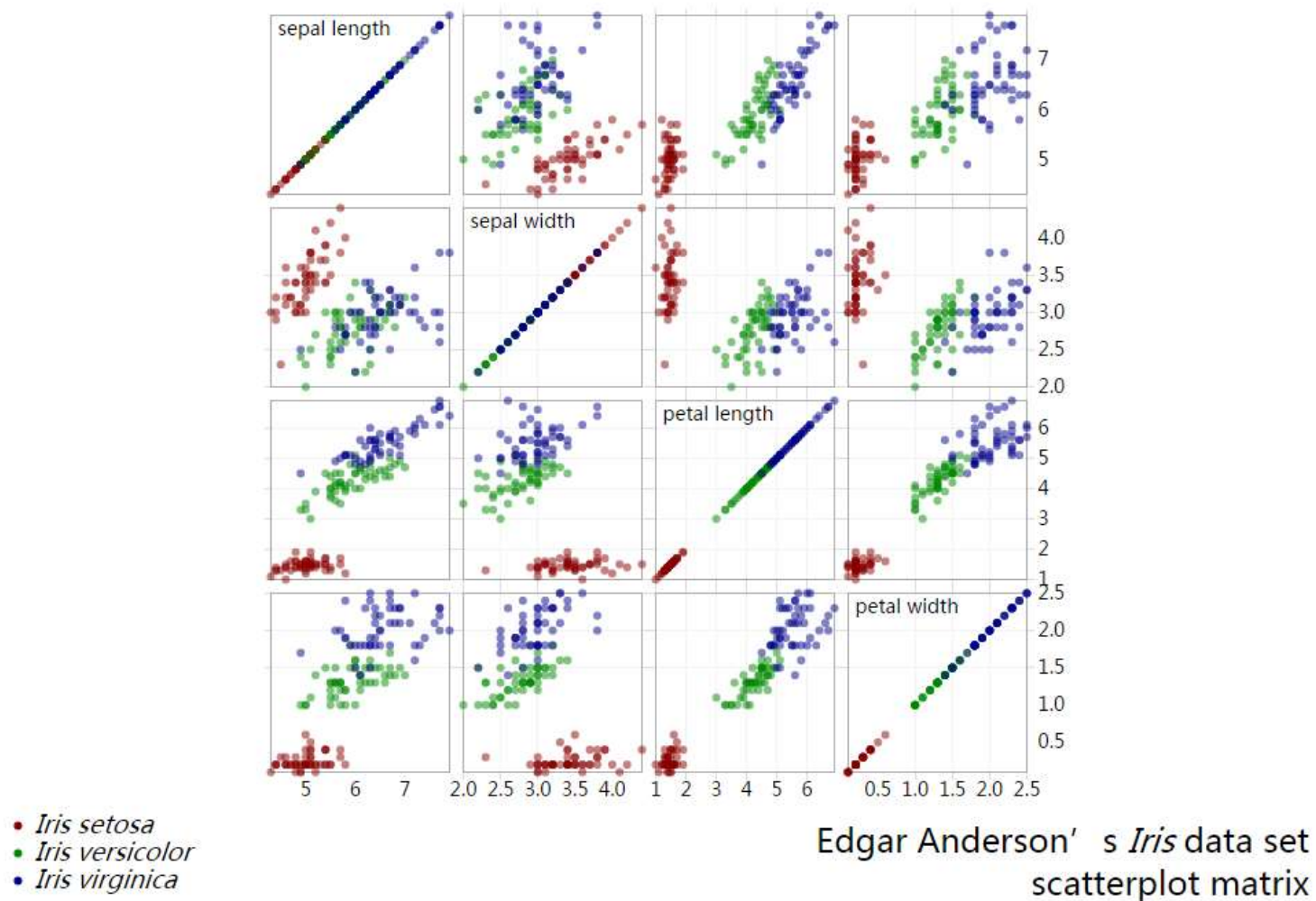
- 基于点的方法
 - 散点矩阵、径向布局
- 基于线的方法
 - 线图、平行坐标、径向轴
- 基于区域的方法
 - 柱状图、表格显示、像素图、维度堆叠、马赛克图
- 基于样本的方法
 - 切尔诺夫脸谱图、邮票图、雷达图

散点图矩阵

- 散点图的高维扩展
 - 每一个散点图代表两个属性之间的两两关系
 - Seaborn包中的pairplot（）函数
 - map_diag()控制对角线上的子图表展示类型
 - map_offdiag()函数控制非对角线外的子图类型



散点图矩阵



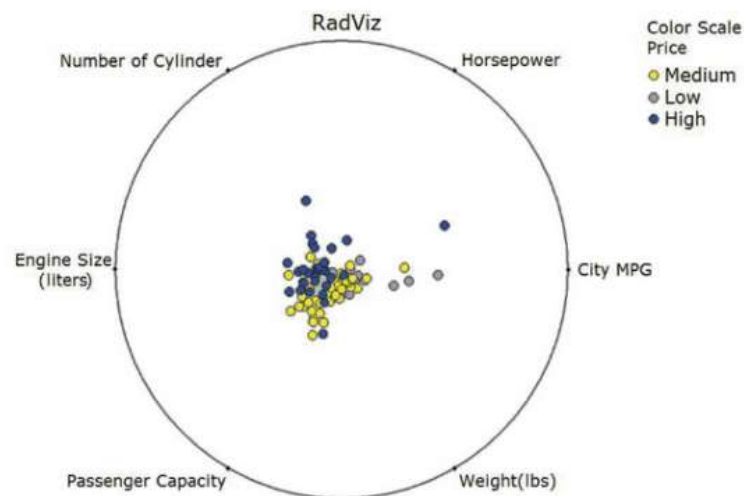
<http://mbostock.github.io/d3/talk/20111116/iris-splom.html>

散点图矩阵

- 优点
 - 便于数据探索 and 比较，有效揭示属性之间的关系
 - 适合于数值型和有序型数据类型
- 缺点
 - 随着属性的增加，其可读性也会受到影响

径向布局法

- 基于弹簧模型的圆形布局方法
- 将代表N维的N个锚点至于圆周上
- 根据N个锚点作用的N种力量将数据点散布于圆内
 - Pandas包中的radviz () 函数
 - `pandas.plotting import radviz`



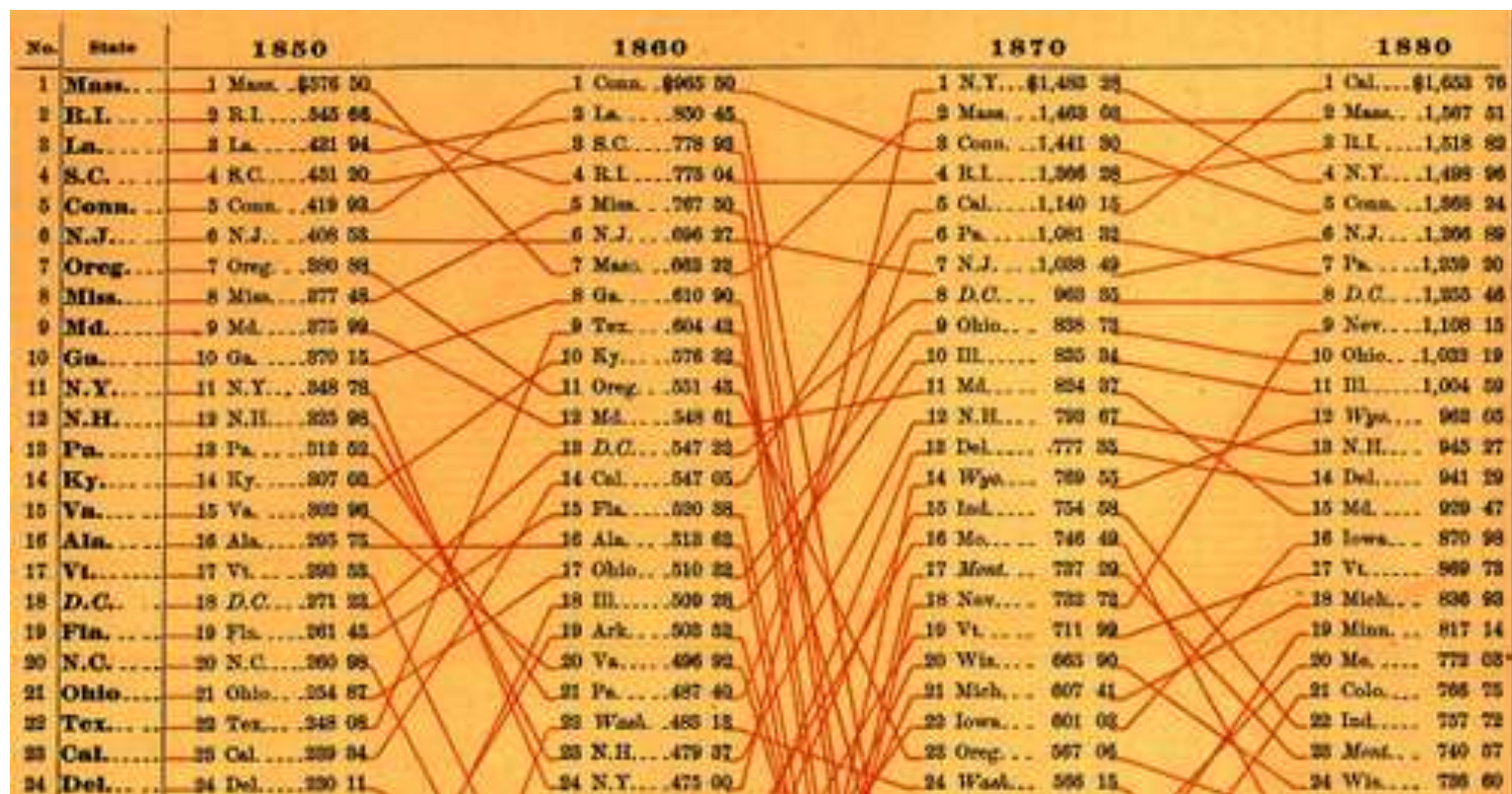
高维多元数据可视化呈现方法

- 基于点的方法
 - 散点矩阵、径向布局
- 基于线的方法
 - 线图、平行坐标、径向轴
- 基于区域的方法
 - 柱状图、表格显示、像素图、维度堆叠、马赛克图
- 基于样本的方法
 - 切尔诺夫脸谱图、邮票图、雷达图

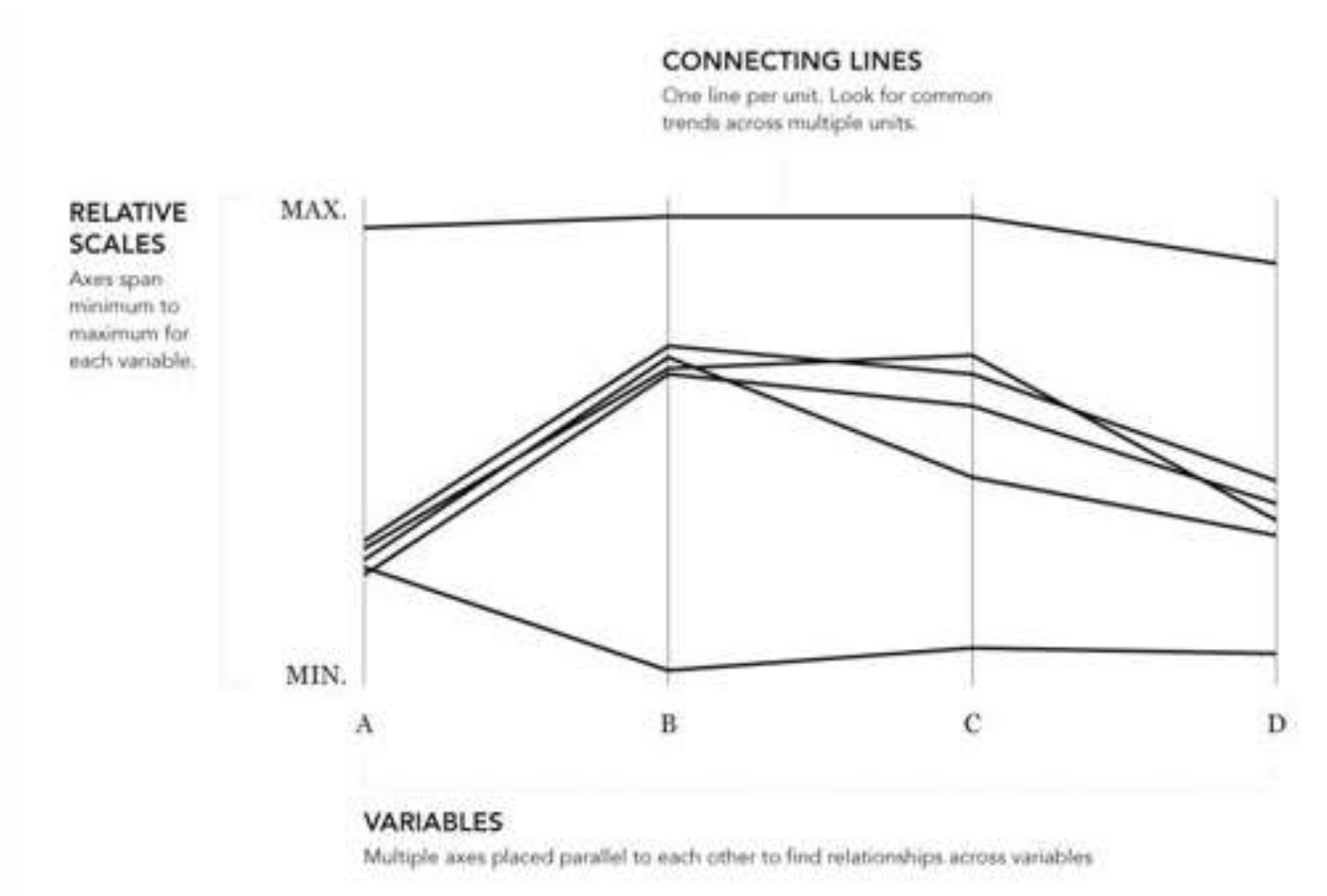
平行坐标

- 1985年由Inselberg提出，用于研究高维集合
- 采用相互平行的坐标轴，每一个坐标轴代表一个属性
- 每一个数据对象对应一条穿过所有坐标轴的折线
- Pandas中的parallel_coordinates()函数
 - `From pandas.tools.plotting import parallel_coordinates`

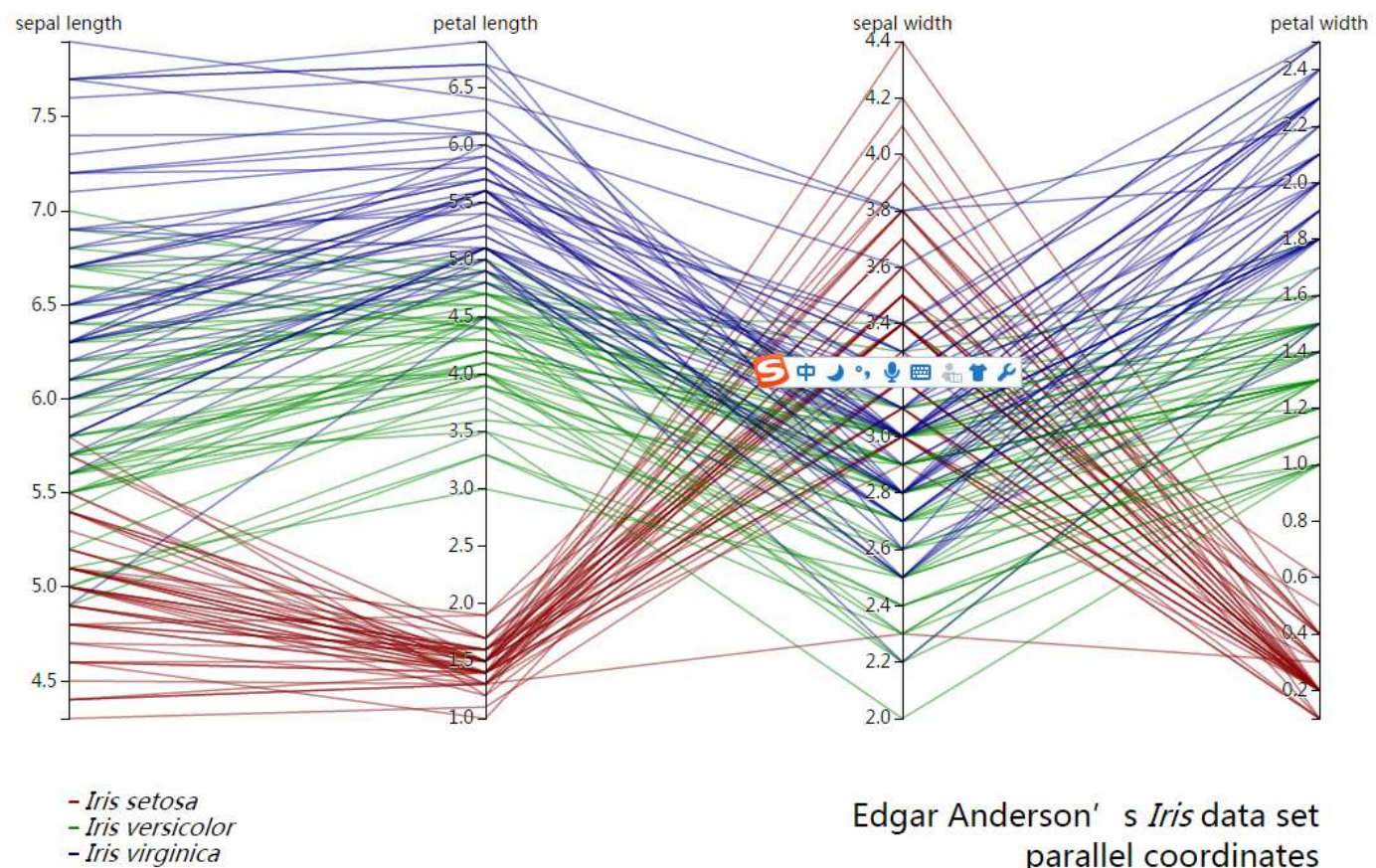
1880年的平行坐标作品



平行坐标框架



Iris实例



<http://mbostock.github.io/d3/talk/20111116/iris-parallel.html>

平行坐标

- 优点
 - 可在二维空间中显示高维度的数据
 - 揭示数据在每个属性上的分布，以及相邻两个属性之间的关系
 - 高数据密度和数据笔墨比例，低畸变因子
- 缺点
 - 不易于同时表达多个维度之间的关系
 - 坐标轴之间的顺序非常重要
 - 增强型平行坐标

径向轴

- 平行坐标的径向排列版本
- 以圆周作为坐标轴，沿圆周绘制线图
- 变种有雷达图，星状图等

高维多元数据可视化呈现方法

- 基于点的方法
 - 散点矩阵、径向布局
- 基于线的方法
 - 线图、平行坐标、径向轴
- 基于区域的方法
 - 柱状图、表格显示、像素图、维度堆叠、马赛克图
- 基于样本的方法
 - 切尔诺夫脸谱图、邮票图、雷达图

基于区域的方法

- 柱状图
 - 最基本的可视化元素（堆叠柱状图）
- 表格显示
 - 热力图：将规则化矩阵数据转换成颜色色调
 - Seaborn中的heatmap（）函数
- 像素图
 - 介于点方法与区域方法的混合方法
- 马赛克图
 - 划分二维空间可视化多维数据
 - 先根据第一维度水平划分空间，再根据第二维度垂直划分空间，重复此过程直到遍历所有维度

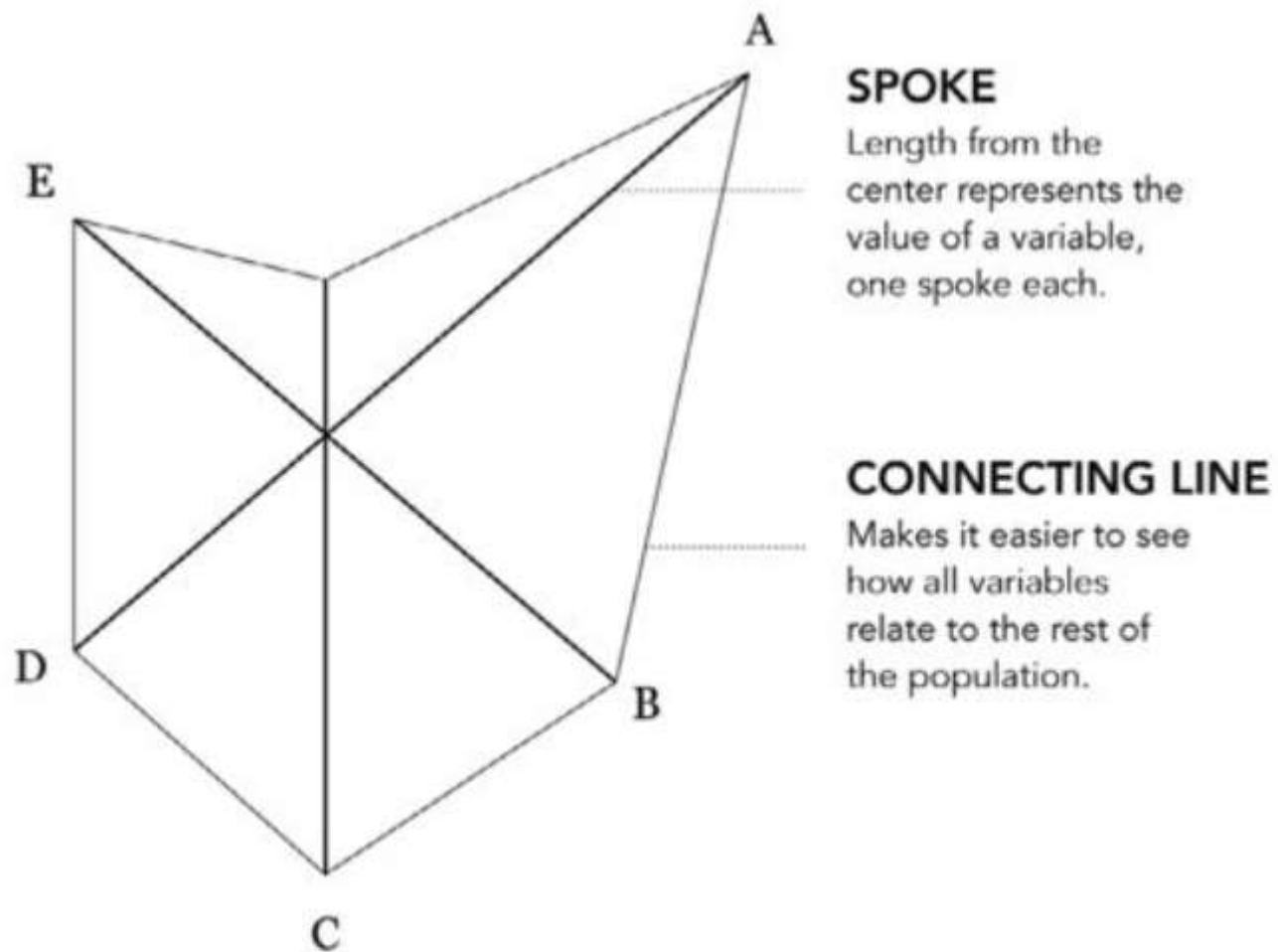
高维多元数据可视化呈现方法

- 基于点的方法
 - 散点矩阵、径向布局
- 基于线的方法
 - 线图、平行坐标、径向轴
- 基于区域的方法
 - 柱状图、表格显示、像素图、维度堆叠、马赛克图
- 基于样本的方法
 - 切尔诺夫脸谱图、邮票图、雷达图

星形图（或雷达图）

- 可视为平行坐标的极坐标版本
- 每一属性由一坐标轴表示，所有坐标轴连到圆心
- 每一坐标轴上的值由数据对象的值与该属性最大值的比例决定，折线连接所有坐标轴上的点

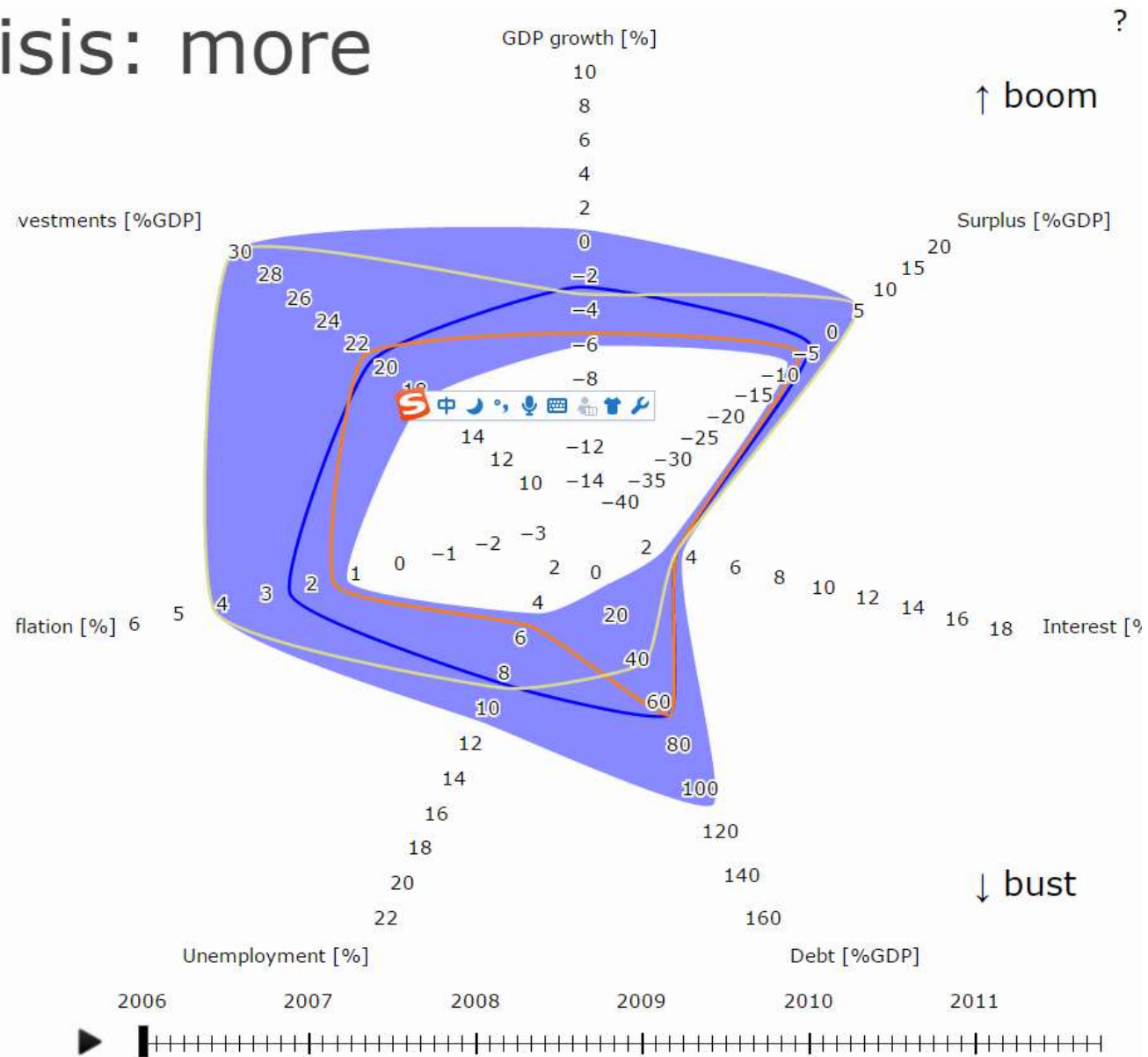
星形图框架



Eurozone crisis: more than debt

- ☒ Austria
- ☐ Belgium
- ☐ Finland
- ☐ France
- ☐ Germany
- ☐ Greece
- ☐ Ireland
- ☐ Italy
- ☐ Luxembourg
- ☐ Netherlands
- ☐ Portugal
- ☒ Spain

January 2006

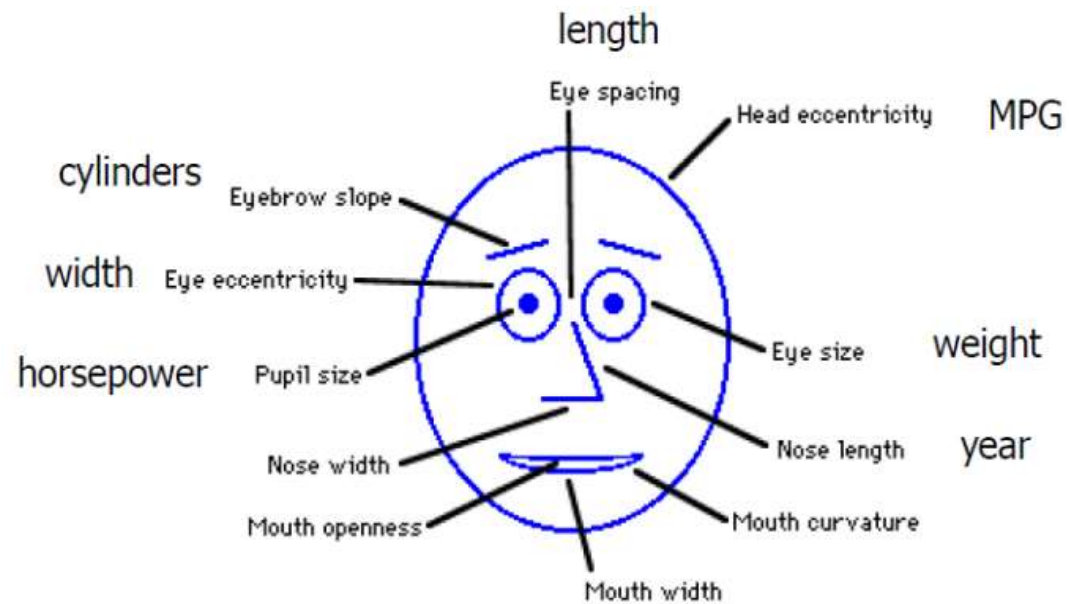


星形图

- 用形状及大小来描述数据
- 数据量少比较有效
- 主要缺点
 - 在一个雷达图中使用多个多边形，会令图表难以阅读
 - 过多变量会导致太多的轴线，增大阅读难度
 - 未能有效地比较每个变量的数值

Chernoff Faces (切尔诺夫脸谱图)

- 与星形图相似，但是模拟人脸的图标
- 不同的属性映射到人脸的不同部位和结构
- R中aplpack包中的faces()函数



Chernoff Faces (切尔诺夫脸谱图)

- 采取了易于被感知的视觉元素
- 人类对于人脸上各个不同部分或者特征的感知程度是不一样的



邮票图表法

- 在一个视图上同时展示多个小图序列
 - 如将时间序列数据按时间点生成一系列图表，并在一个视图空间内有序地平铺展示
- 可用分面图实现
 - Seaborn, plotnine等都有相应实现方法

高维多元数据可视化呈现方法

- 基于点的方法
 - 散点矩阵、径向布局
- 基于线的方法
 - 线图、平行坐标、径向轴
- 基于区域的方法
 - 柱状图、表格显示、像素图、维度堆叠、马赛克图
- 基于样本的方法
 - 切尔诺夫脸谱图、邮票图、雷达图

其他图形

- 南丁格尔玫瑰图
 - 极坐标柱形图，圆弧的半径长度代表
 - 缺点：面积较大的外围部分会更加引人注目，畸变因子较大
- 词云图
 - 每个字的大小与其出现频率成正比
 - Wordcloud包
-



高维数据可视化

- 数据降维
- 可视化呈现方法
- 与数据交互
 - 交互的可视化方法

资源

- Nathan Yau, *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*, Wiley Publishing, 2011.
- Mike Bostock, *Data Driven Documents* (D3.js)
 - <https://github.com/mbostock/d3/wiki/Gallery>
 - <http://bl.ocks.org/mbostock>