



第2章 Data



数据从何而来

■ 一、数据由自身系统获得

- 酷狗：获得登录用户的播放音乐列表，类型等..
- 腾讯：社交网络应用
- 淘宝：推荐商品
- 英雄联盟：游戏匹配...
-

■ 优点：相对比较自由，数据权限较大，分析空间较大，领域知识明确...

■ 缺点：需要一定的数据积累，需要时间，用户，使用记录等积累



数据从何而来

- 二、大公司开放的数据接口
 - 第三方登录
 - 支付宝第三方支付接口
 -
- 优点：无需积累，直接借用大公司提供的数据接口进行数据访问
- 缺点：依赖接口程度非常大，只能获取开放接口的一些数据



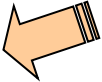
数据从何而来

- 三、爬虫软件，自己爬（很多中小公司应用）
 - 一淘
 -
- 优点：无需积累，从网页上，或者特定方式爬取下来，降低数据维护成本
- 缺点：需要跟随爬取目标的展示方式的变化而变化，需要大量的程序维护人员

数据从何而来

- 四、大公司公开真实、离线、历史数据
 - <https://aws.amazon.com/cn/public-datasets/>
 - <http://snap.stanford.edu/data/index.html>
 - <http://archive.ics.uci.edu/ml/>
 - http://www.sogou.com/labs/resource/list_pingce.php
 - <https://tianchi.shuju.aliyun.com/?spm=5176.100068.1234.1.HkXwVp>
- 优点：无需积累，数据真实可靠，一次性离线下载
- 缺点：数据部分公开，数据格式数据结构需要行业背景，一般是历史的

Chapter 2: 了解数据

- 数据对象和属性类型 **Data Objects and Attribute Types** 
- 数据的(基本)统计描述 **Basic Statistical Descriptions of Data**
- 数据可视化 **Data Visualization**
- 测量数据相似性和相异性 **Measuring Data Similarity and Dissimilarity**

数据集合的类型

■ 记录Record

- 关系记录
- 数据矩阵, **e.g.**, 数值矩阵, 交叉表
- 文档数据: 文本文档:词频向量
term-frequency vector
- 交易数据

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

■ 图 and 网络

- 万维网
- 社会或信息网络
- 分子结构**Molecular Structures**

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



数据集合的类型

- 有序的 Ordered
 - 视频数据: **sequence of images**
 - 时间数据: 时间序列 **time-series**
 - 序列数据: 交易序列 **transaction sequences**
 - 遗传序列数据
- 空间, 图像image and 多媒体multimedia:
 - **Spatial data: maps**
 - **Image data:**
 - **Video data:**



结构数据的重要特征

- 维度 **Dimensionality**
 - 维数灾难 **Curse of dimensionality**
- 稀疏 **Sparsity**
 - 只有计数 **Only presence counts**
- 分辨率 **Resolution**
 - 模式依赖于尺度
- 分布 **Distribution**
 - 中心性和分散 **Centrality and dispersion**



数据对象

- 数据集由数据对象构成
- 一个数据对象代表一个实体
- 例子:
 - 销售数据库**sales database**: 客户 / 顾客, 商店物品, **sales**
 - 医学数据库: **patients, treatments**
 - 大学数据库: **students, professors, courses**
- 10 又称为 样本, 事例, 实例, 数据点, 对象, 元组 *tuples*.



属性

- 属性Attribute (or维度, 特征, 变量): 一个数据字段, 表示一个数据对象的某个特征.
 - *E.g., customer_ID, name, address*
- 类型:
 - 名词性**Nominal**
 - 二元的
 - 数字的**Numeric**: 数量的
 - **Interval-scaled**
 - **Ratio-scaled**

属性类型

- 名词性Nominal:类别,状态, or “名目”
 - ***Hair_color = {auburn, black, blond, brown, grey, red, white}***
 - 婚姻状态, 职业**occupation, ID numbers, zip codes**
- 二元
 - 只有**2**个状态的名词性属性 (**0 and 1**)
 - 对称二元**Symmetric binary**: 同样重要的两相
 - e.g., gender
 - 非对称**Asymmetric binary**: 非同等重要
 - e.g., 医疗检查 (positive vs. negative)
 - 惯例**Convention**: assign 1 to most important outcome (e.g., HIV positive)



数值属性的类型

- 数量Quantity (integer or real-valued)
- 区间Interval
 - 在某个同等大小的一个尺度单位上**Measured on a scale of equal-sized units**
 - 值有序
 - ***E.g., temperature in C° or F° , calendar dates***
 - 没有真正的零点
- Ratio
 - 有真正的零点
 - 可以讲值是被测量单位一个数量级 (**10 K° is twice as high**



离散 vs. 连续属性

■ Discrete Attribute

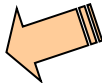
- 一个有限的或可数无限集值
 - **E.g., zip codes, the set of words in a collection of documents**
- 有时,表示为整数变量
- 注: 二元属性是离散属性的一个特殊情况

■ Continuous Attribute

- 属性值为实数
 - **E.g., temperature, height, or weight**
- 实际上, 实值只能使用有限位数进行测量和代表

14. ■ 连续属性通常表示为浮点变量

Chapter 2:数据的统计描述

- **Data Objects and Attribute Types**
- 数据的(基本)统计描述 
- 数据可视化
- 测量数据相似性和相异性 **Measuring Data Similarity and Dissimilarity**
- **Summary**



数据的(基本)统计描述

- 动机

- 为了更好的理解数据:集中趋势, 变异和传播

- 数据特征

- 中位数, 最大, 最小, 分位数, 离群点, 方差, 等.

- 针对排序区间的数值维

- 数据离散度: 多个粒度上的精确分析
 - 排序区间的盒图/分位数图分析

- 某计算侧度下的离散度分析

- 折叠为某数值维度下
 - 转化立方体上的盒图/分位数图



分布度量/代数度量/整体度量

- 从数据挖掘角度，需要考察如何在大型数据库中有效计算度量。
- 分布式度量 **distributive measure**
 - 可通过如下方法计算的度量（函数）：将数据划分成较小子集，计算每个子集的度量，合并计算结果得到整个数据集的度量值。
 - **sum(),count(),min(),max()**
- 代数度量 **algebraic measure**
 - 可用一个函数于一个或多个分布度量计算的度量
 - **average()/mean()**
- 整体度量 **holistic measure**
 - 必须对整个数据集计算的度量
 - **median(),mode()**

度量数据的中心趋势

- 均值 (代数度量) (样本 vs. 总体):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

- 加权算术均值；截断均值：去掉高低极端值

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- 中位数:

- 奇数则为有序集的中间值, 否则为中间两个数的平均

- (基于分组数据)可以插值估计

$$median = L_1 + \left(\frac{n/2 - (\sum freq)_{small}}{freq_{median}} \right) width$$

- 众数Mode

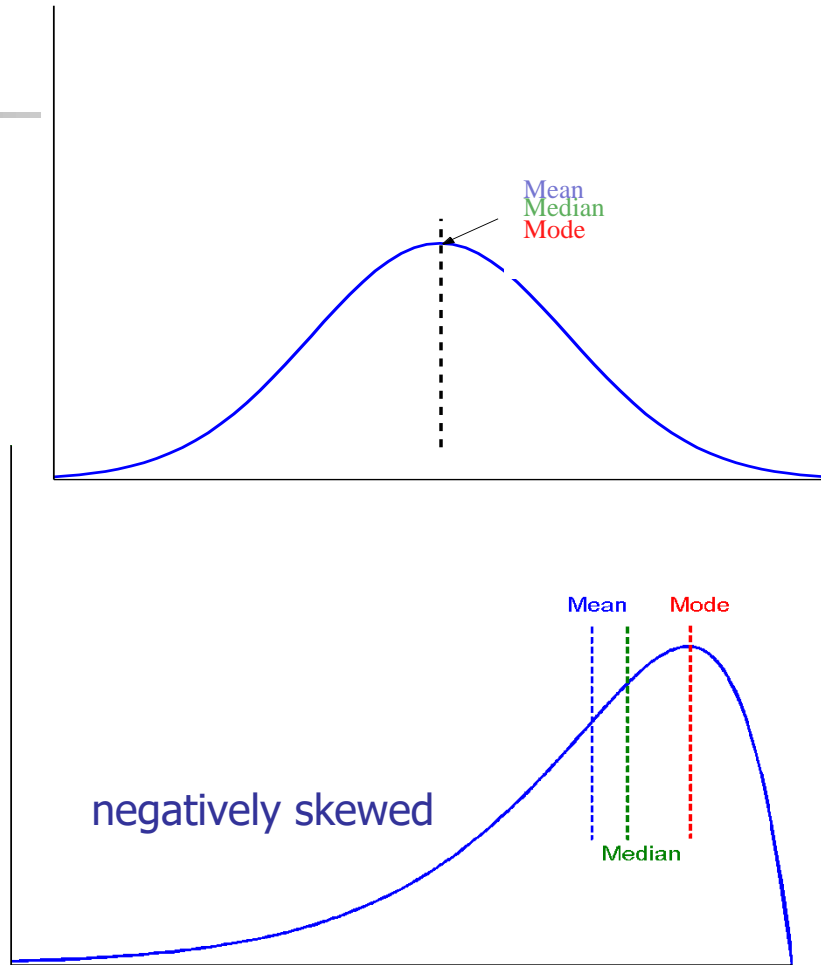
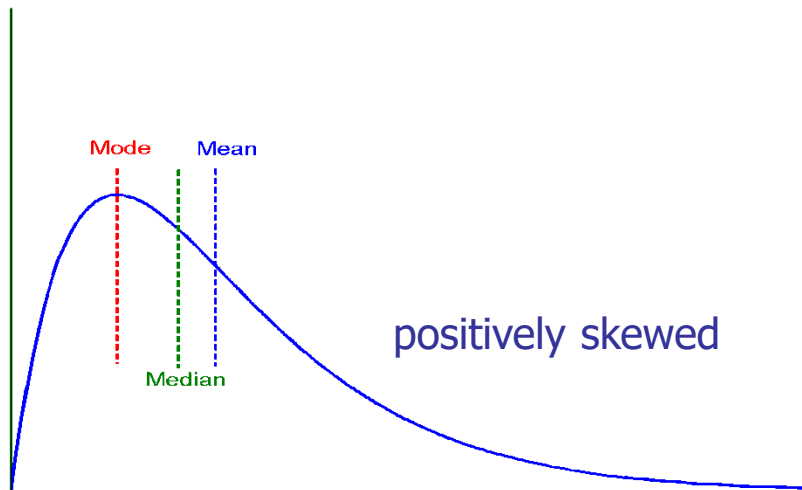
- 出现频率最高的值(不惟一/每个值出现一次则没有众数)

- **1/2/3**个众数-〉单峰的, 双峰的, 三峰的

- 单峰频率曲线的经验关系: $mean - mode = 3 \times (mean - median)$

对称/偏斜数据

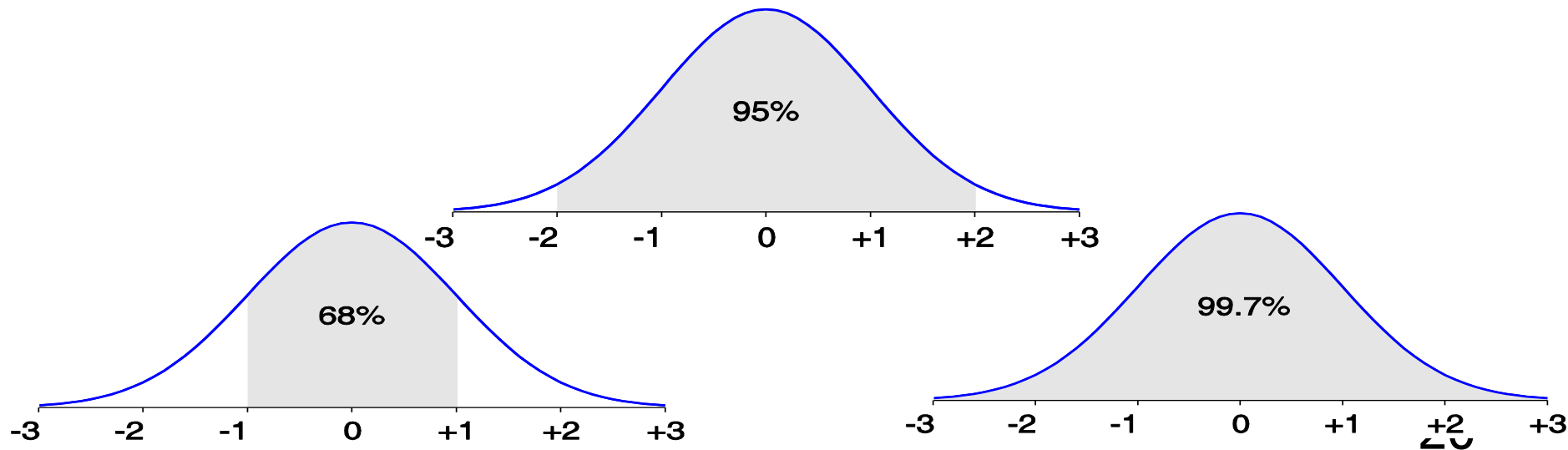
- 中位数(Median), 均值(Mean), 众数(Mode): 对称, 正倾斜和负倾斜数据



正态分布曲线的性质

■ 正态分布曲线

- $[\mu - \sigma, \mu + \sigma]$: 含有约**68%**的测量(μ : 均值, σ : 标准差)
- $[\mu - 2\sigma, \mu + 2\sigma]$: contains about **95%** of it
- $[\mu - 3\sigma, \mu + 3\sigma]$: contains about **99.7%** of it



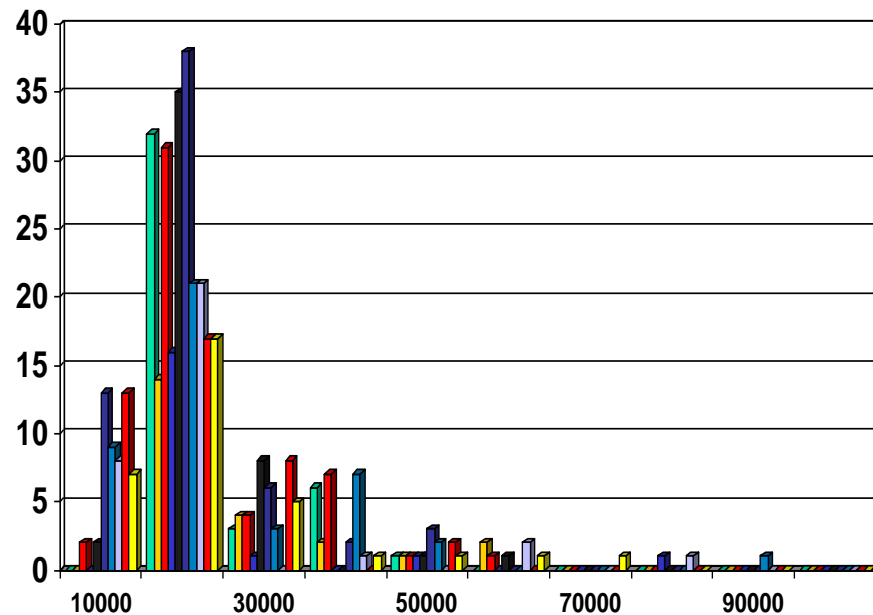


基本统计说明的图形显示

- **Boxplot:** 五数概括的图形
- **Histogram**直方图: 值x-axis, y-axis表示频率
- **Quantile plot**分位数图: 值 x_i 与 f_i (表明近似 $100 f_i \%$ 的数据 $\leq x_i$)成对
- **Quantile-quantile (q-q) plot:** 对着另一个分位数, 绘制一个单变量分布的分位数
- **Scatter plot**散布图: 每个值对 为一个坐标点绘于平面上

直方图分析

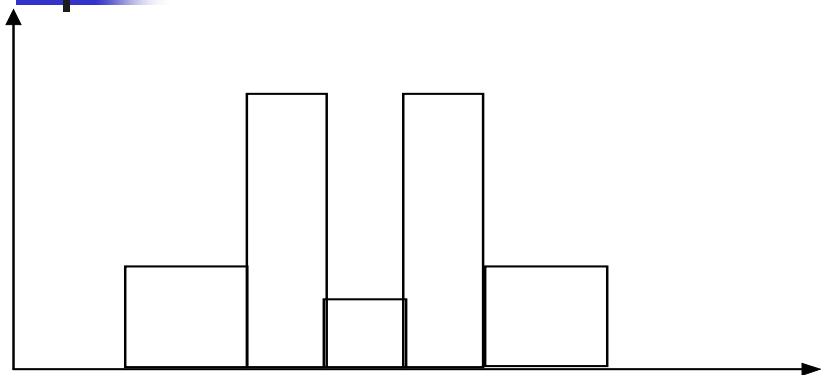
- **Histogram:** 图形显示每个列值的频率，条形图所示
- 显示有多大比例的点下落入每个类别
- 类别并不是均匀的宽度时有别于条形图一个关键：条形图的面积表示值而不是条形图的高度



■ **a bar chart** 柱状图/柱形图

- 类别通常指定为变量的一些非

Histograms Often Tell More than Boxplots

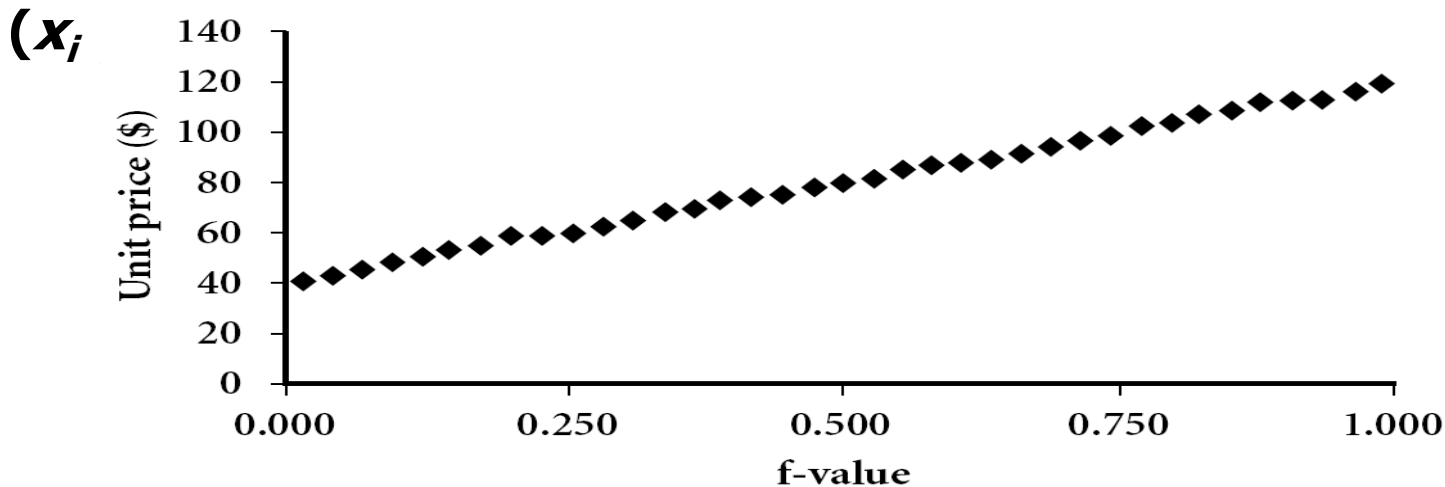


- 两个直方图显示在左边有同样的 **boxplot** 表示
 - 相同的值: min, Q1, median, Q3, max
- 他们拥有的是不同的数据分布

But they have

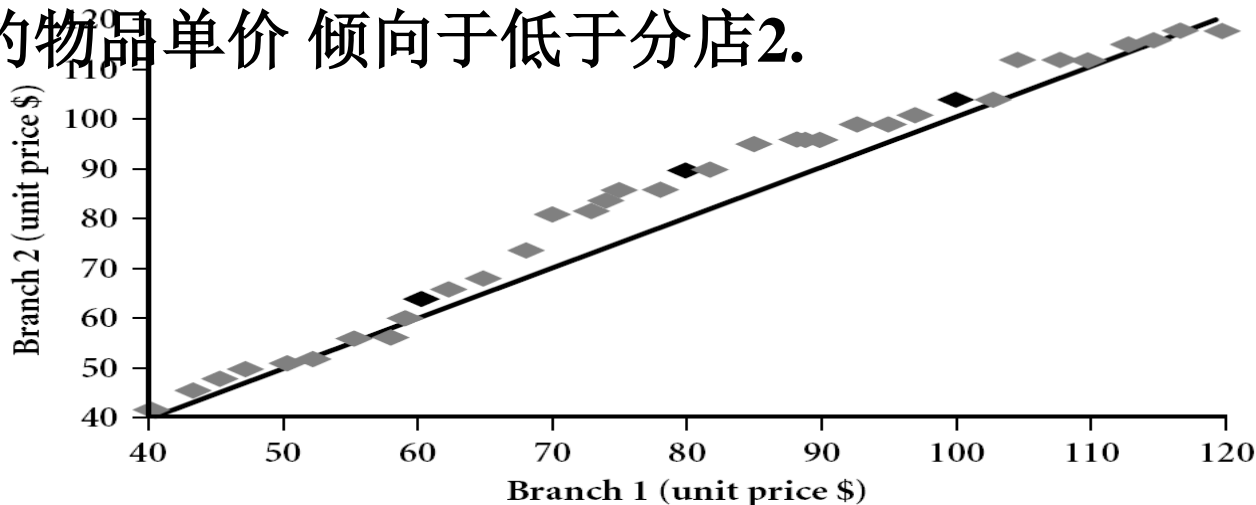
分位数图Quantile Plot

- 显示所有数据 (允许用户评估全部行为和不寻常的事件)
- Plots quantile information
 - 对于升序中的值点 x_i , f_i 表明近似 **100 f_i %** 的数据 $< x_i$, 成对绘制



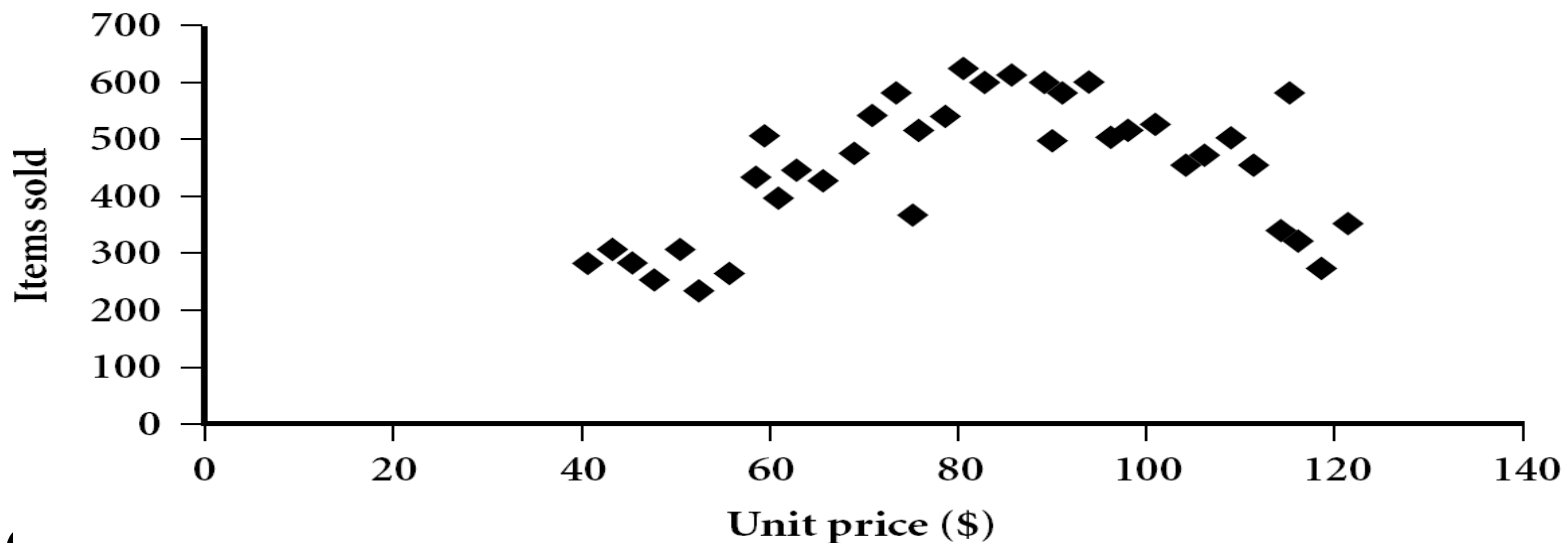
分位数-分位数图 (Q-Q图)

- 对着另一个分位数，绘制一个单变量分布的分位数
- 观察:正从一种分布到另一个种是否有偏移?
- 例子表示分店1出售的物品单价 vs. 分店 2 的每个分位数.分店 1出售的物品单价 倾向于低于分店2.

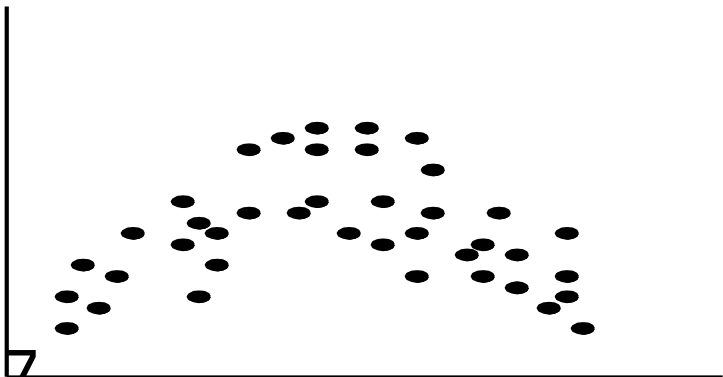
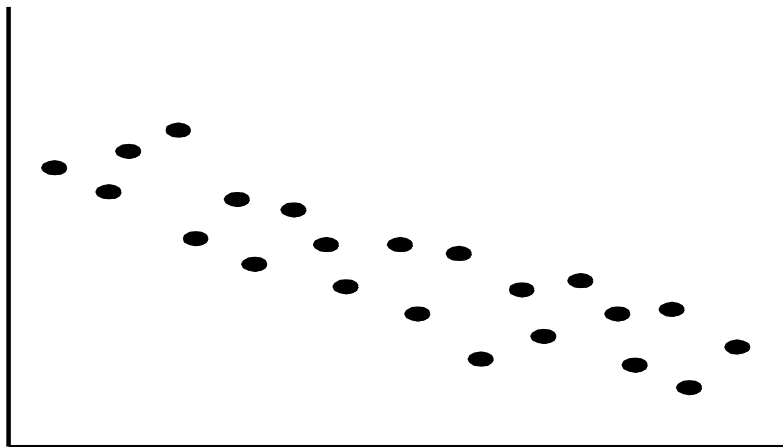
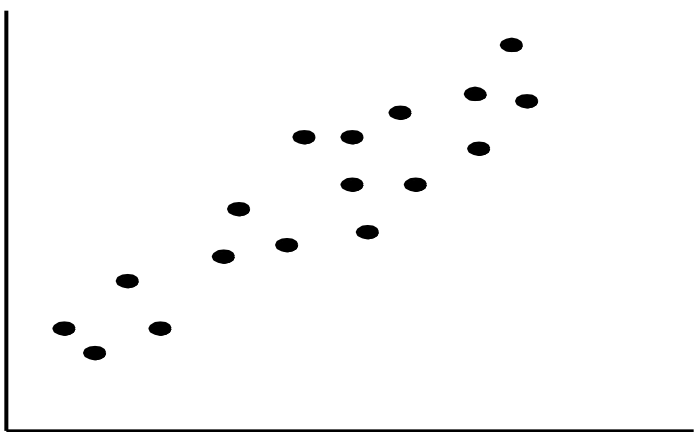


散布图Scatter plot

- 提供双变量的数据的第一印象：点的聚集，离群点，等
- 每个值对作为一个坐标点绘于平面上

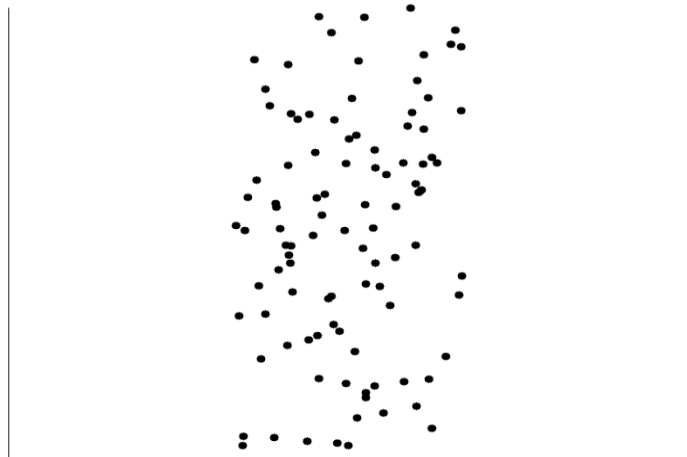
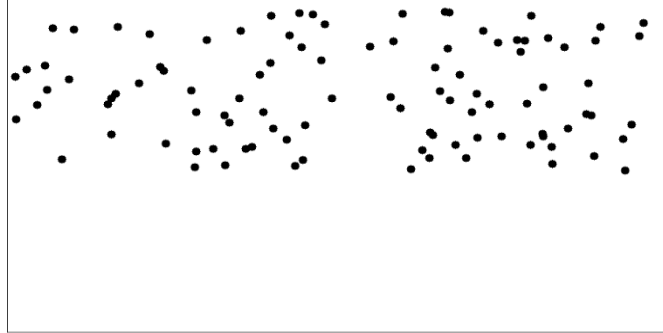
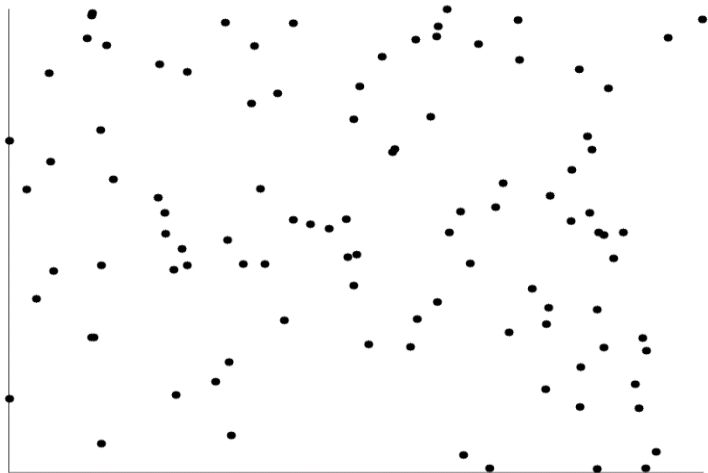


正/负 相关数据



- The left half fragment is positively correlated
- The right half is negative correlated

不相关的数据



散布图的例子

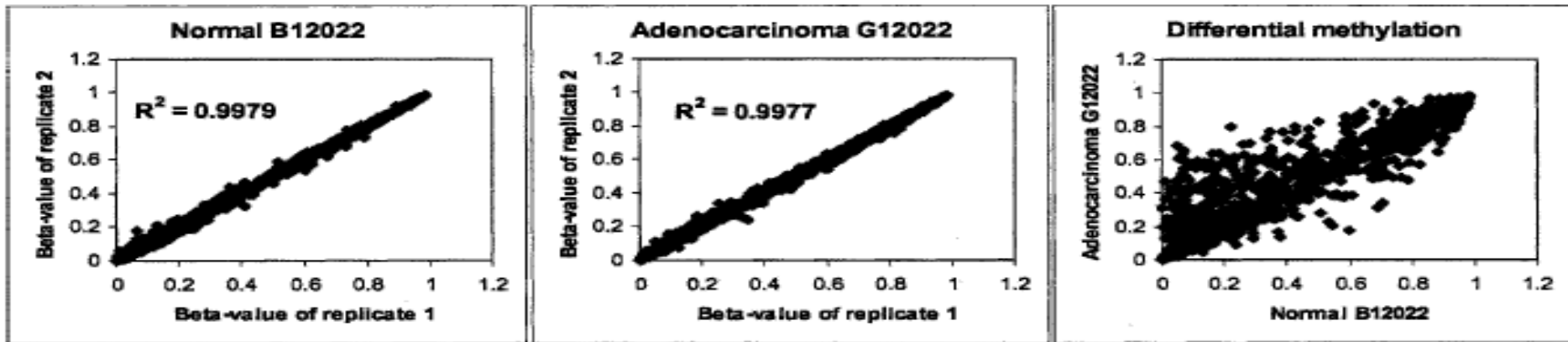
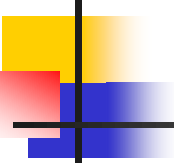



Figure 2. Methylation assay reproducibility and differential methylation detection. Comparison of methylation profiles between lung cancer and matching normal tissue. The β -value (i.e., the methylation ratio measured for all 1536 CpG sites) obtained from one replicate experiment is plotted against



Chapter 2: 数据可视化

- 数据对象和属性类型**Data Objects and Attribute Types**
- 数据的(基本)统计描述**Basic Statistical Descriptions of Data**
- 数据可视化**Data Visualization**
- 测量数据相似性和相异性**Measuring Data Similarity and**



数据可视化

- Why data visualization?

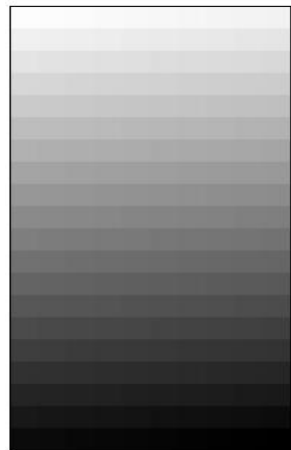
- 把数据映射到图形信息空间中获取视角
- 提供定性的概述(大数据集的)
- 在数据中搜寻 模式, 趋势, 结构, 不规则, 关联
- 为进一步的量化分析发现有意义的区域及合时的参数
- 为衍生的计算机表示提供一个视觉证据

- 可视化方法的分类:

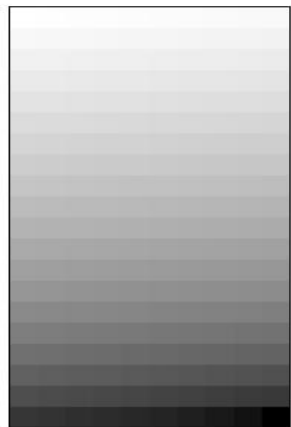
- 基于像素的可视化技术 **Pixel-oriented visualization**
- 几何投影可视化技术 **Geometric projection**

基于像素的可视化技术

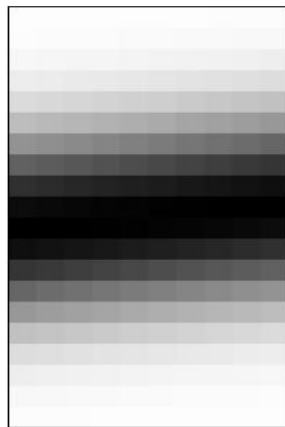
- 对一个维度 m 的数据，在屏幕上产生 m 个窗口,每个维度一个
- 一个记录的 m 维度值被匹配到窗口中对应位置的 m 个像素上
- 像素的颜色值反映了相应的值



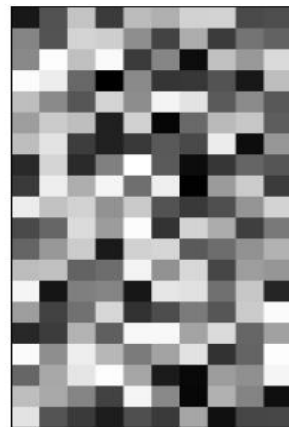
(a) Income



(b) 信用限额



(c) 交易额

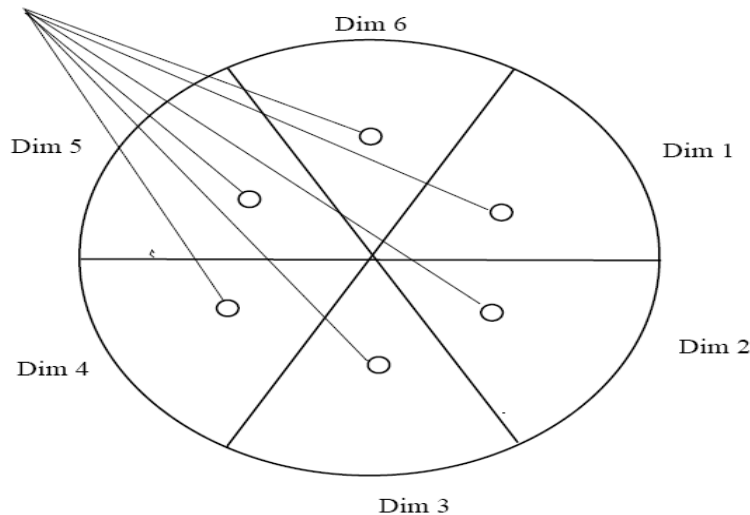


(d) age

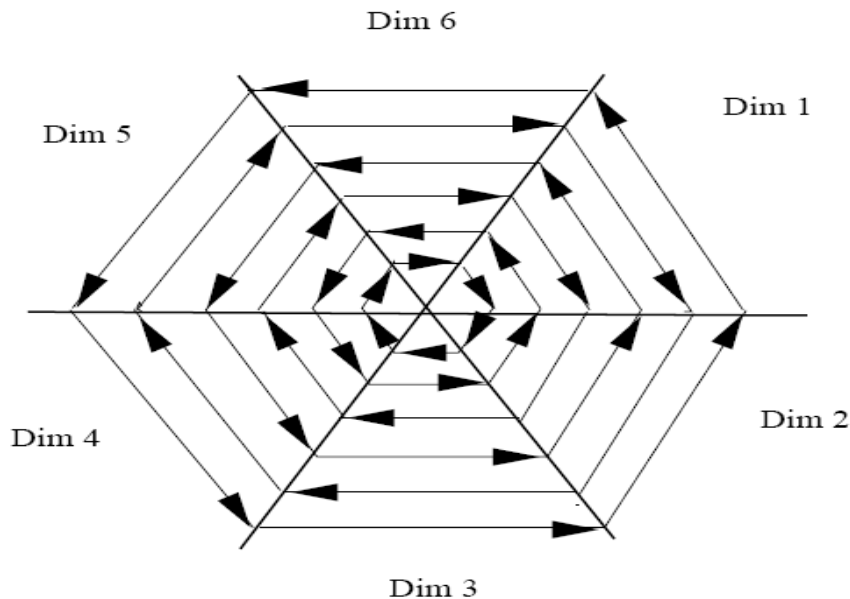
安排像素于圆弧片断

- 为节省空间并显示多个维度间的联系,往往是以一个弧形片段填充空间

one data record

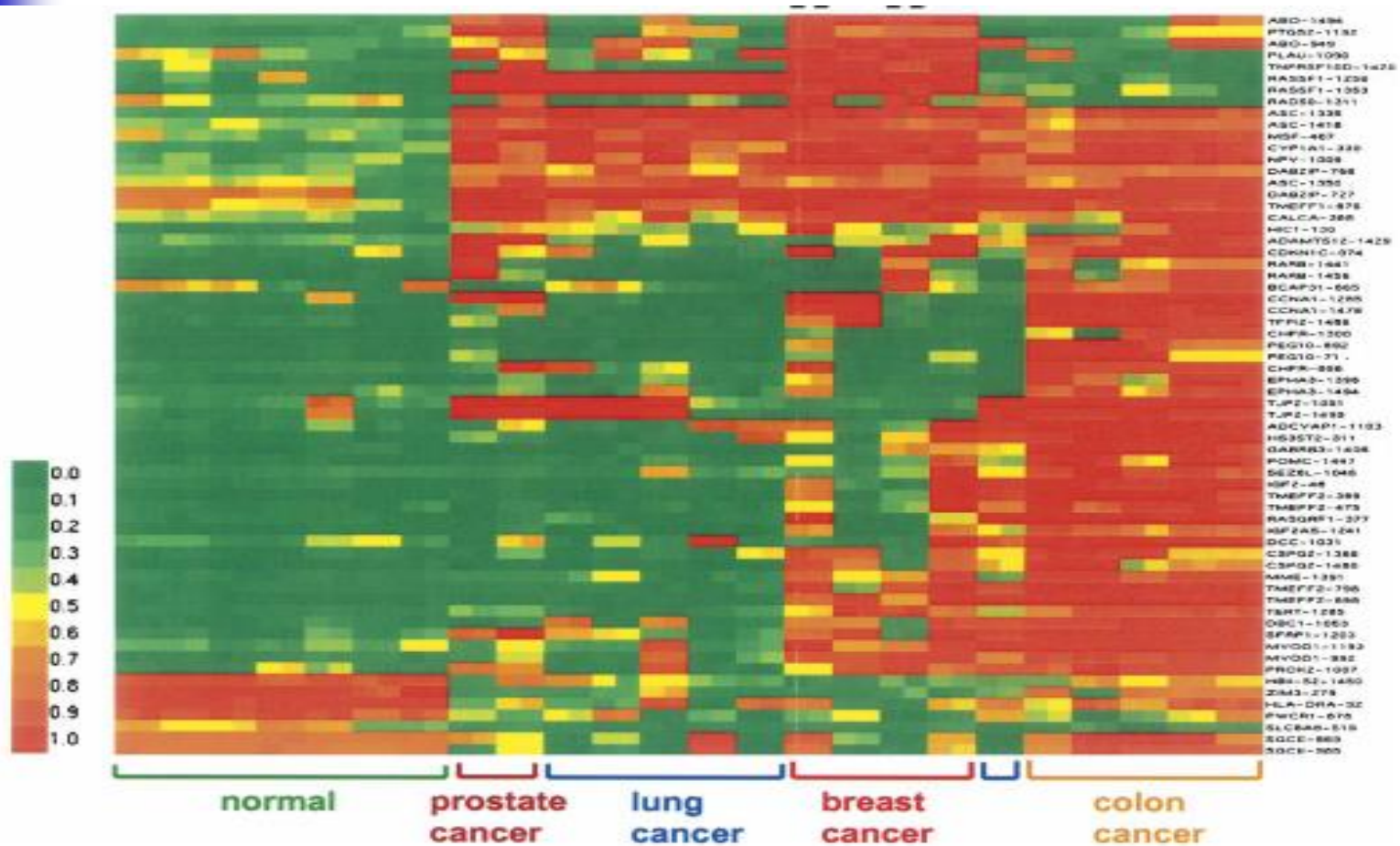


(a) Representing a data record in circle segment



(b) Laying out pixels in circle segment

像素图的例子



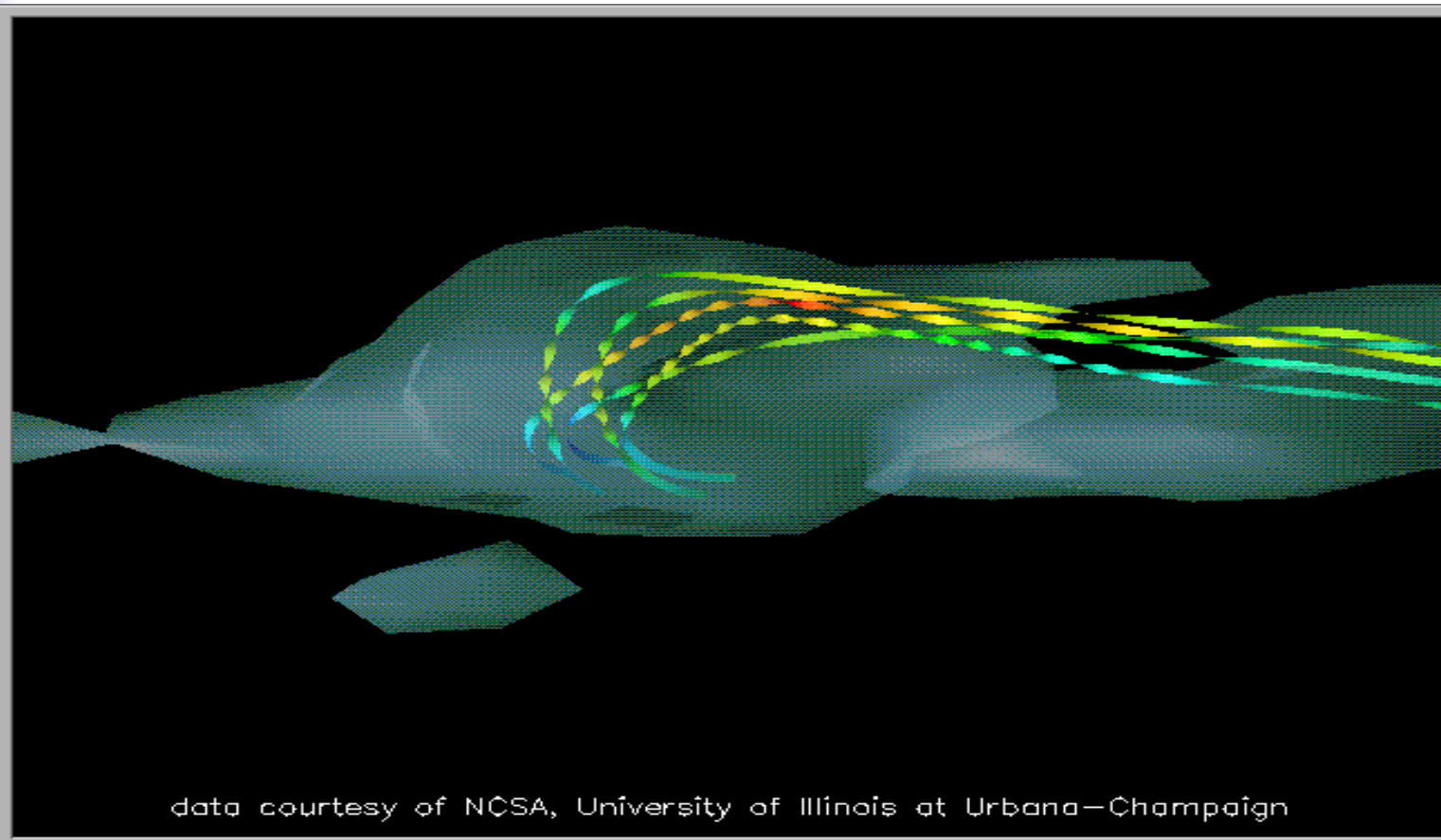


几何投影可视化技术

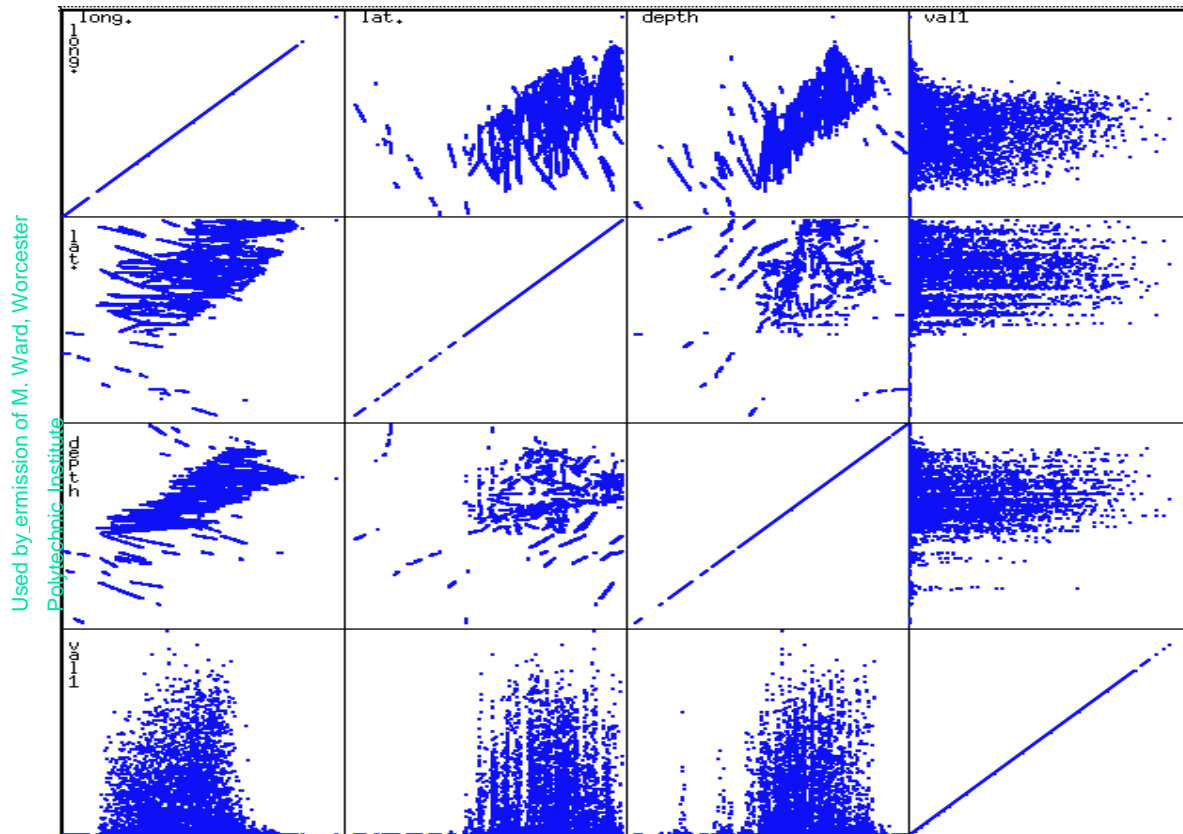
- 可视化数据的几何变换和投影
- 方法
 - 直接可视化
 - 散布图和散布图矩阵 **matrices**
 - 透视地形 **Landscapes**
 - 投影捕获技术: 帮助用户发现有意义的投影 (多维数据上)
 - 解剖视角 **Prosection views-- projections and sections**
 - **sections, i.e., intersections of subspaces with a highdimensional object, can easily display structure of**

直接数据可视化

基于涡度的含扭曲丝带



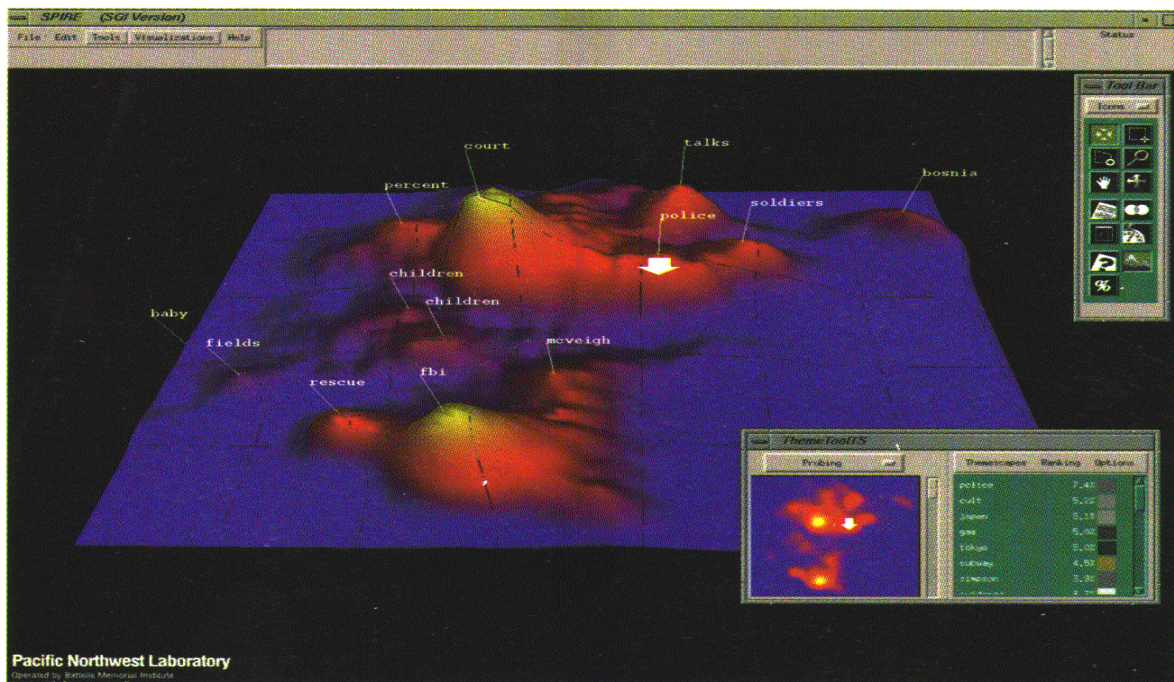
散布图矩阵



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of $(k^2/2 - k)$ scatterplots]

透视地形/景观

Used by permission of B. Wright, Visible
Decisions Inc.

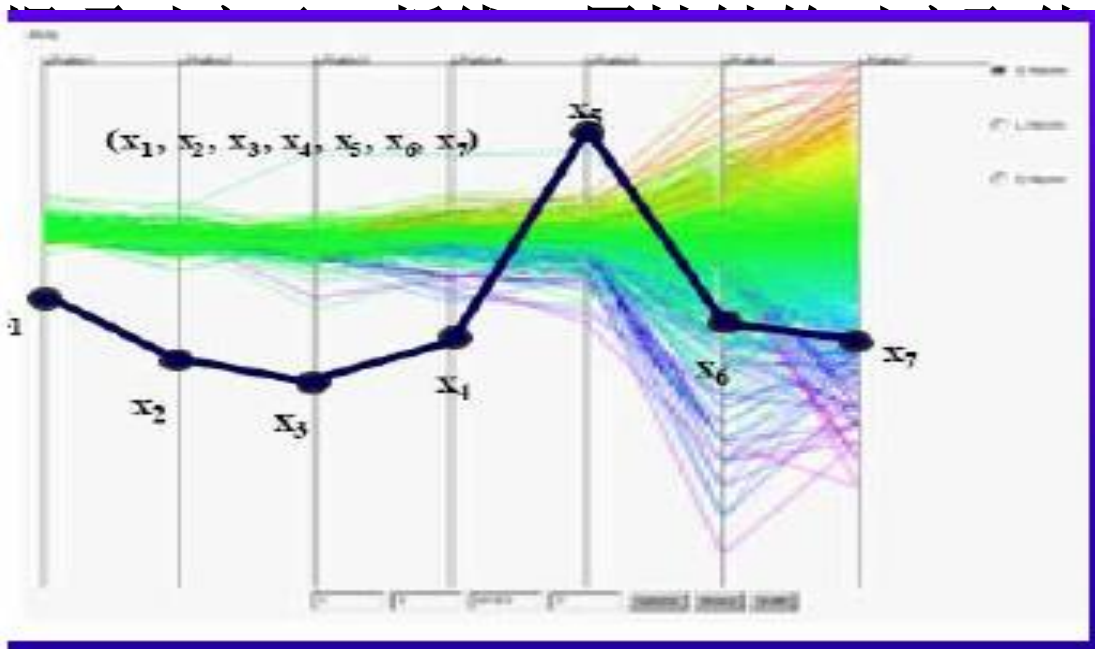


news articles
visualized as
a landscape

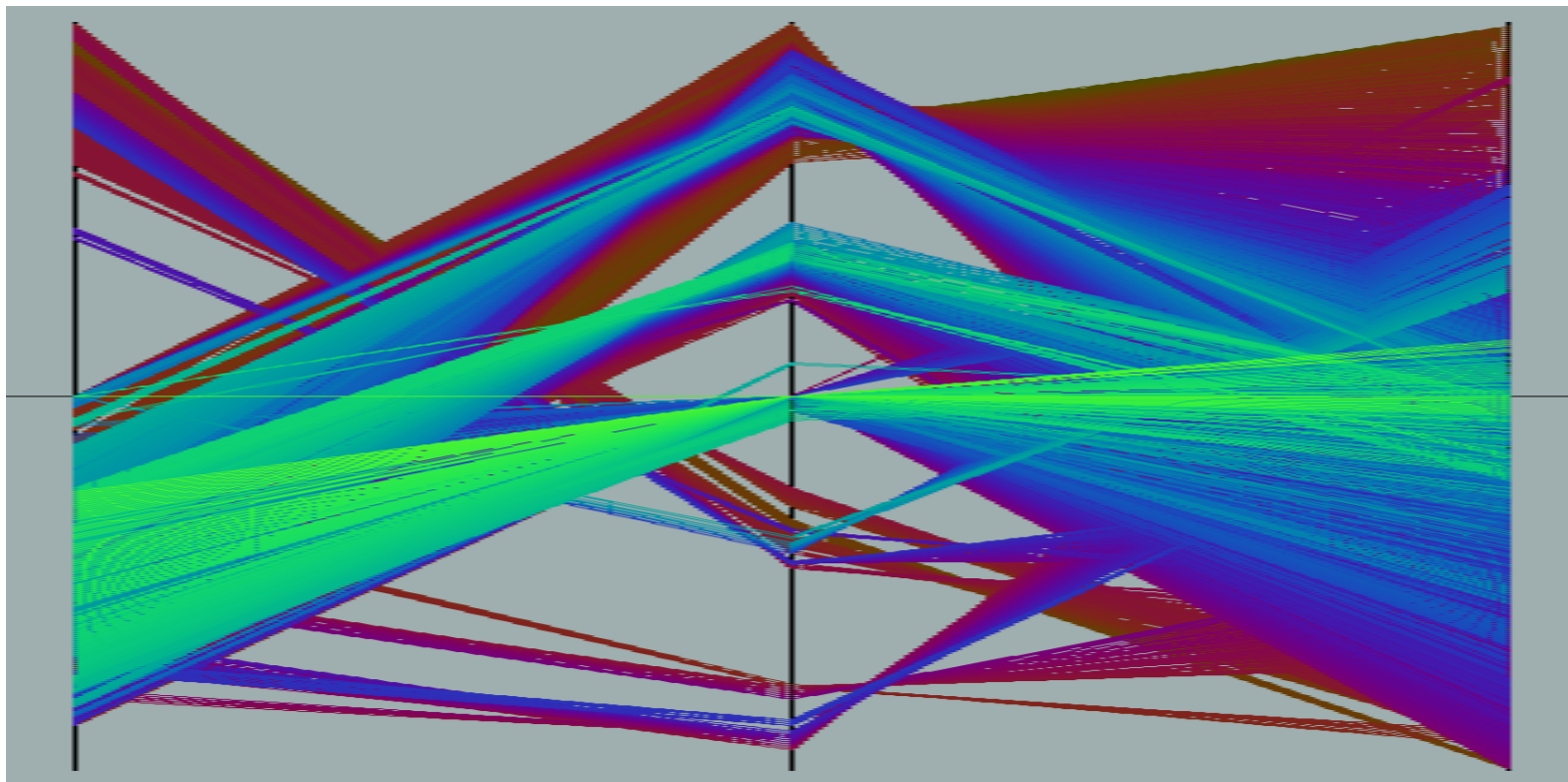
- 透视方式可视化数据
- 数据要被转化为能保持数据特点的二维表示（可能人工）

平行坐标

- 对应于属性的n个等距轴平行于一个屏幕轴
 - 这些轴缩放到[最小值, 最大值]:相应的属性范围
 - 每个数
- 点处相交



一个数据集的平行坐标





基于图标的可视化技术

- 以图标特征可视化数据值
- 典型的可视化方法
 - **Chernoff Faces** 脸谱图
 - **Stick Figures** 棍棒图
- 常用技术
 - 形状编码 **Shape coding**: 使用形状来表示特定信息的编码
 - 颜色图标 **Color icons**: 使用颜色图标编码更多的信息
 - 瓦片条形图 **Tile bars**: 在文档检索中使用小图标代表相

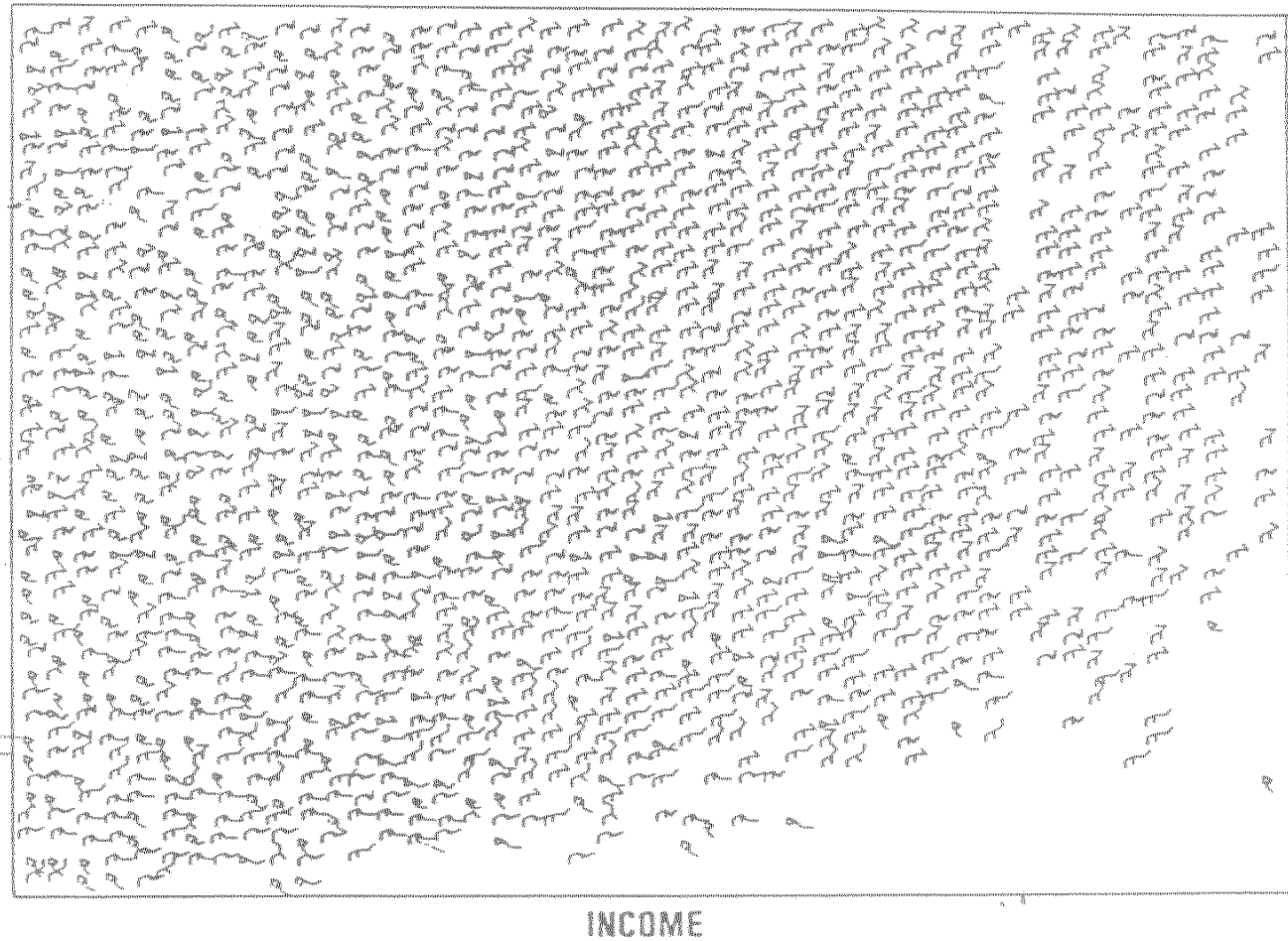
切尔诺夫脸谱图 Chernoff Faces

- 一种方法在二维空间显示变量,如设 X 眉倾斜, Y 是眼睛大小, Z 是鼻子长度等
- 图中的面孔使用10个特点产生-头离心率, 眼睛大小, 眼间距, 眼离心率, 瞳孔大小, 斜眉, 鼻大小, 嘴形, 嘴的大小, 张口程度: Each assigned one of 10 possible values, generated using *Mathematica* (S. Dixon)
- REFERENCE: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics*. New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld*--A Wolfram Web Resource.



mathworld.wolfram.com/ChernoffFace.html

棍棒图Stick Figure



人口普查数据
图显示年龄、
收入、性别、
教育、等

一个5-piece棍
棒图(身体和四
肢)，两个属性
映射到轴，其余
的属性映射到角
度或肢体长度



分层可视化技术

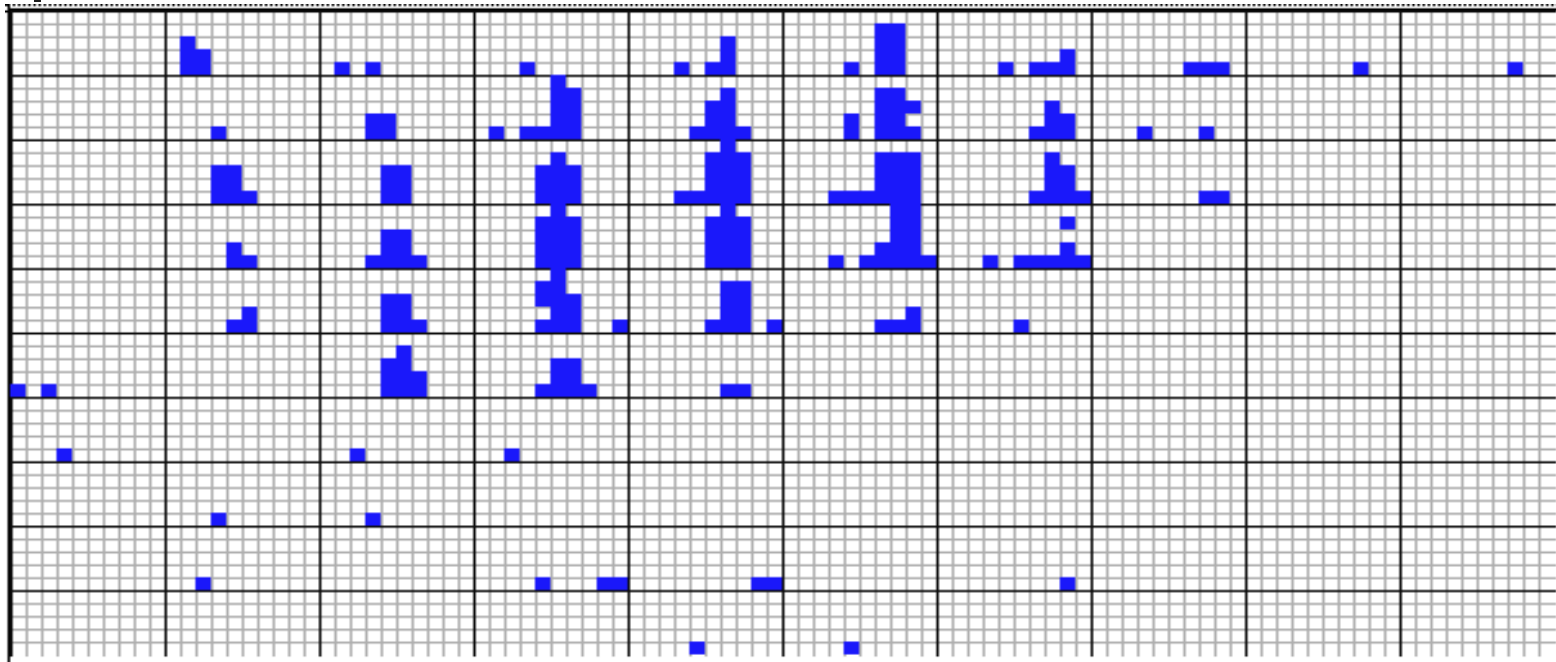
- 使用子空间层次划分可视化数据
- 方法
 - 维数堆叠 **Dimensional Stacking**
 - **Worlds-within-Worlds**
 - **Tree-Map** 树状图
 - **Cone Trees** 锥形树
 - **InfoCube**



维数堆叠 Dimensional Stacking

- 把n维属性空间剖分为2-D子空间，互相堆叠与一起
- 属性值的范围划分为等级，重要的属性分布在外层.
- 适合次序属性较少的数据
- 超过9个维度时显示困难
- 重要的是匹配维度适当

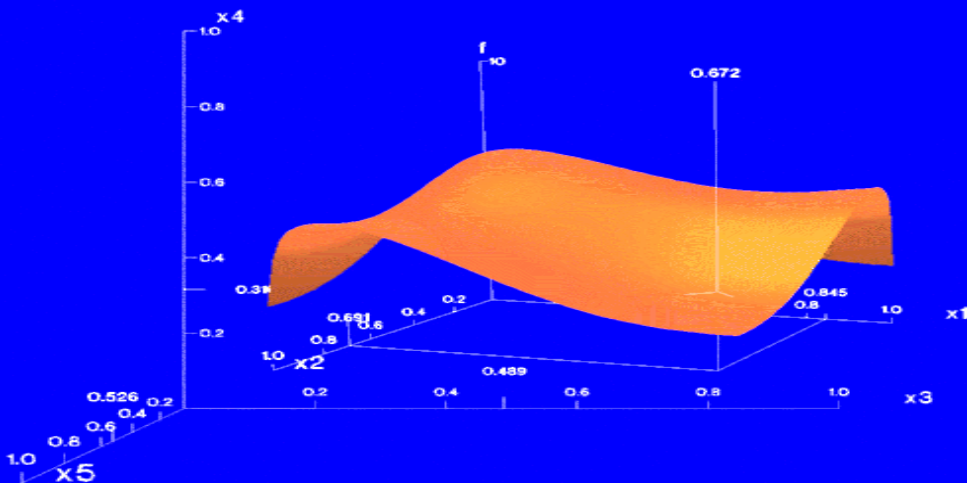
维数堆叠 Dimensional Stacking



可视化石油勘探数据，经度和纬度映射到外x-, y轴，油质和深度映射到内部x-, y-轴

Worlds-within-Worlds

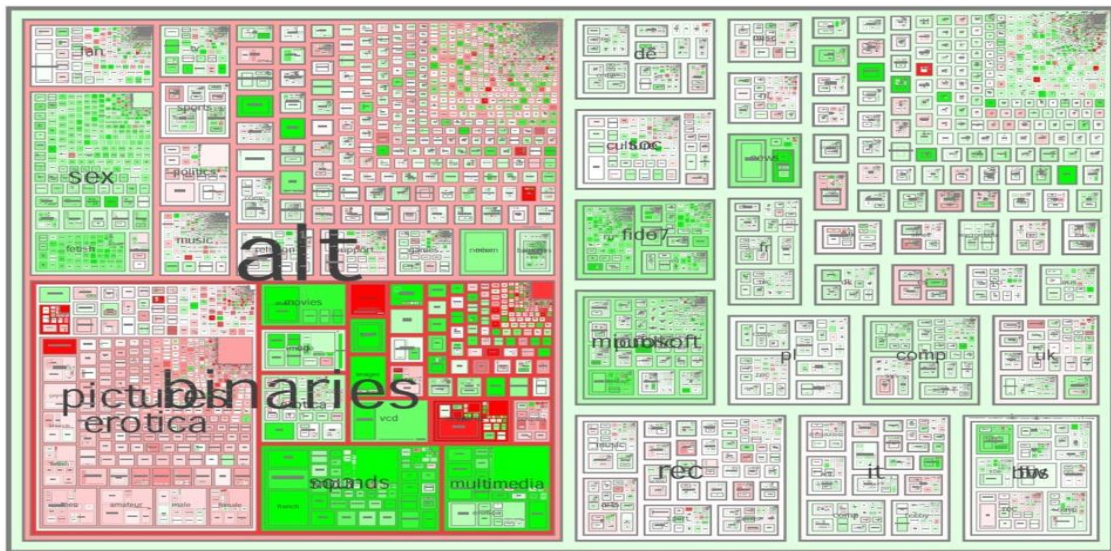
- 分配功能 f 和两个重要参数给内部世界
- 固定其他参数 - draw other (1 or 2 or 3 世界选择他们为坐标轴)
- 使用这种模式的软件
 - **N-vision:**通过数据手套和立体显示以动态互动，包括旋转，缩放（内部）和转换（内/外）
 - 自动视觉：经查询手段静态互动



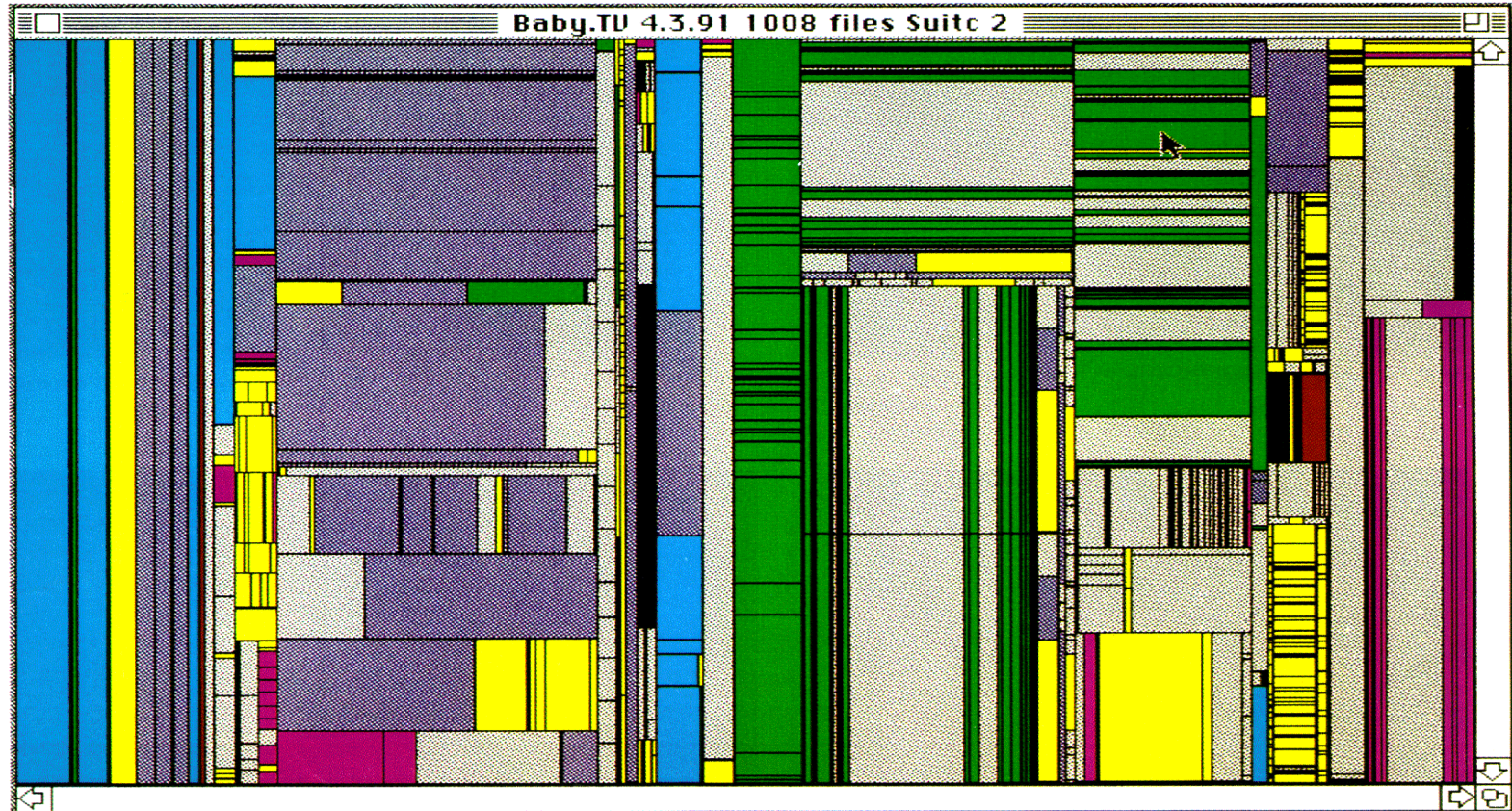
树状图Tree-Map

- 屏幕填充方法：依赖于属性值把 屏幕层次划分为区域
- 根据属性值（类）屏幕的x-y-维交替剖分

MSR Netscan Image



Tree-Map of a File System (Schneiderman)?

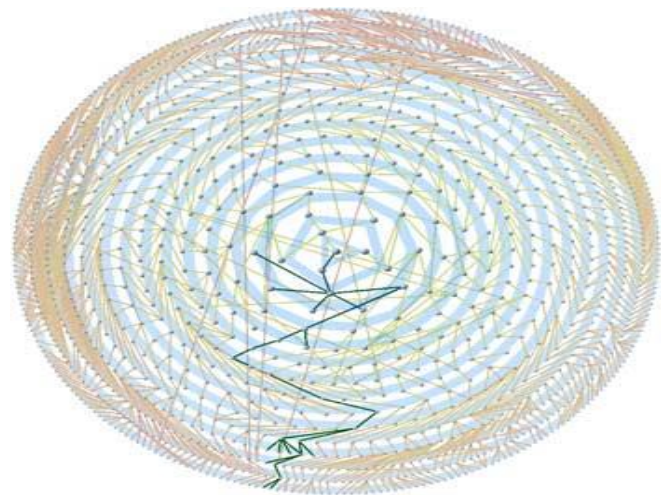
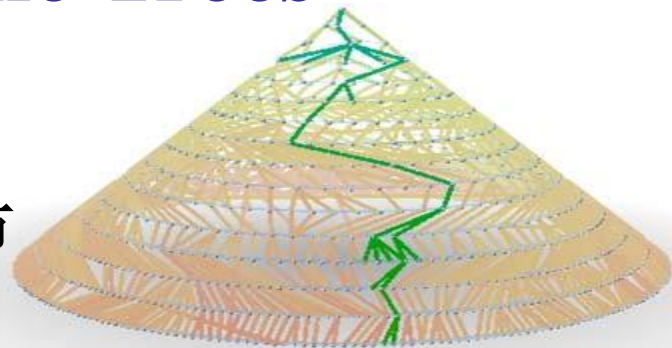


- 3-D可视化技术：层次信息被显示成嵌套的半透明立方体
- 最外层的立方体对应顶层数据, 子节点or低层数据作为稍小的立方体显示于外层立方体中, 以此类推



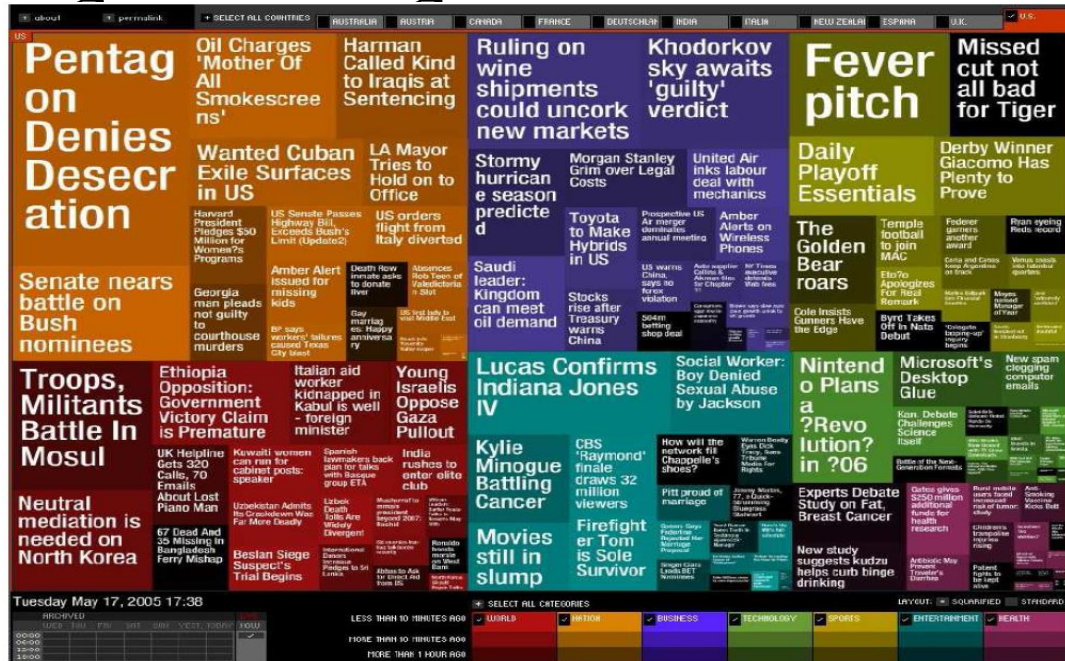
3d锥树 Three-D Cone Trees

- *3D cone tree* 可用于数千个节点
- 先构造 *2D 环形树*，安排节点于根节点为中心的同心圆环
- 投影到2维时将不可避免重叠
- G. Robertson, J. Mackinlay, S. Card.
“Cone Trees: Animated 3D Visualizations of Hierarchical Information”, *ACM SIGCHI'91*
- Graph from Nadeau Software
Consulting website: 可视化社会网络



可视化复杂数据和关系

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
 - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks

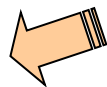


Newsmap: Google News Stories in 2005



Chapter 2:数据相似性和相异性

- Data Objects and Attribute Types
- 数据的(基本)统计描述
- 数据可视化
- 测量数据相似性和相异性Measuring Data Similarity and





相似性和相异性

- **Similarity**

- 数值测量两个数据对象类似程度
- 目标越相似时值越大
- 通常介于 **[0,1]**

- **Dissimilarity (e.g., 距离distance)**

- 数值测量两个数据对象差异程度
- **Lower when objects are more alike**
- **Minimum dissimilarity is often 0**
- **Upper limit varies**

- **邻近度Proximity refers to a similarity or dissimilarity**

数据矩阵和相异度矩阵

■ Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

■ Dissimilarity matrix

- n data points, but registers only the distance

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

名词性属性的邻近度量

- 2个或多个状态, e.g., red, yellow, blue, green (二元属性的推广)
- Method 1: 简单匹配
 - m : p 个变量中匹配的个数, p : 全部变量的个数
- Method 2: 使用一系列的二进制属性
 - 为 M 个名义状态的每一个产生一个新的二进制/二元属性

二进制属性的邻近度量

- 二进制数据的列联表
contingency table

		Object j		
		1	0	sum
Object i	1	q	r	$q + r$
	0	s	t	$s + t$
sum		$q + s$	$r + t$	p

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- 对称二元变量的距离侧度:

$$d(i, j) = \frac{r + s}{q + r + s}$$

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

- 不对称二元变量的距离侧度:

- Note: Jaccard coefficient is the same as "coherence".

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

的相似性侧度):

二进制属性的相异度量

■ Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- 性别是对称属性

- The remaining attributes are asymmetric binary

- 令Y and P 值为1, 且N值为0

$$d(jack, mary) = \frac{2 + 0 + 1}{1 + 1} = 0.5$$
$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

规范数值数据

- **Z-score:** $z = \frac{x - \mu}{\sigma}$
 - **X:** 需标准化的原始数值, **μ:** 总体均值, **σ:** 标准差
 - 在标准偏差单位下, 原始分数和总体均值之间的距离
 - “-”, “+”

- 另一种方法: Calculate the mean absolute deviation
$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

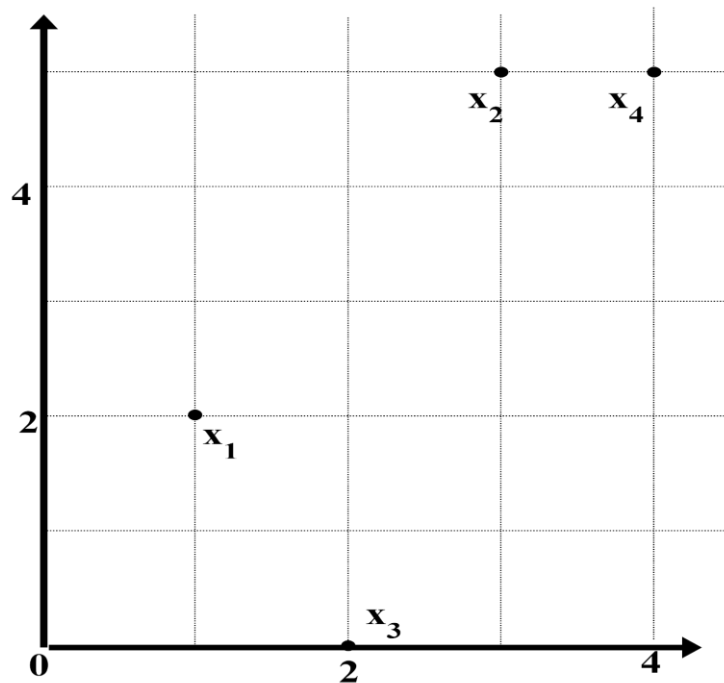
$$m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf}).$$

其中

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- **standardized measure (z-score):**

例：数据矩阵和相异度矩阵



Data Matrix

point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3.61	0		
$x3$	5.1	5.1	0	
$x4$	4.24	1	5.39	0

数值数据的距离: Minkowski Distance

- ***Minkowski distance***: 一种流行的距离测度

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

其中 $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 为两个 p -维数据点, and h is the order (the distance so defined is also called L- h norm)

- 特性
 - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (正定 Positive definiteness)

闵可夫斯基距离特殊形式

- $h = 1$: **Manhattan** (city block, L_1 norm) **distance** 曼哈顿距离 (L1范数)

- E.g., the Hamming distance: the number of bits that are different between two binary vectors

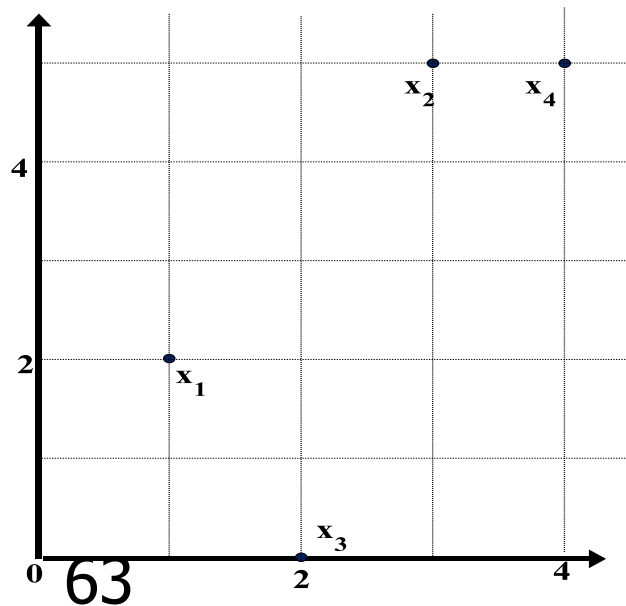
$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- $h = 2$: (L_2 norm) **Euclidean distance**
$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$ 上确界 “**supremum**” (L_∞ norm, L_∞ norm) distance.
$$d(i, j) = \max_{f=1, \dots, p} |x_{if} - x_{jf}|$$
 This is the maximum difference between any

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

有序变量 Ordinal Variables

- 一个序变量可以离散的或连续的
- Order is important, e.g., rank
- Can be treated like interval-scaled $\dots, M_f \}$
 - 用他们的序代替 x_{if}
 - 映射每一个变量的范围于 $[0, 1]$, 用如下支代替第 f -th 变量的 i -th 对象

64 ■ compute the dissimilarity using methods for

混合型属性

- A database may contain all attribute types

- **Nominal, symmetric binary, asymmetric binary, numeric, ordinal**

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- 可以用加权法计算合并的影响

- **f is binary or nominal:**

$$d_{ij}^{(f)} = 0 \text{ if } x_{if} = x_{jf}, \text{ or } d_{ij}^{(f)} = 1 \text{ otherwise}$$

- **f is numeric: use the normalized distance**

- **f is ordinal**

- Compute ranks r_{if} and

余弦相似性 Cosine Similarity

- A document can be represented by the *frequency* of a particular word

Document	team	coach	hockey	baseball
Document1	5	0	3	0
Document2	3	0	2	0
Document3	0	7	0	2
Document4	0	1	0	0

$$\cos(x, y) = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p x_i^2 \cdot \sum_{i=1}^p y_i^2}}$$

- Other vector objects: gene features in micro-arrays, ...
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{\|d_1\| \|d_2\|}$$



Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$,
where \bullet indicates vector dot product, $\|d\|$: the length of vector d
- Ex: Find the similarity between documents 1 and 2.
 $d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
 $d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$
 $d_1 \bullet d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$
 $\|d_1\| = (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$



Summary

- **Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled**
- **Many types of data sets, e.g., numerical, text, graph, Web, image.**
- **Gain insight into the data by:**
 - **Basic statistical data description: central tendency, dispersion, graphical displays**
 - **Data visualization: map data onto graphical primitives**
 - **Measure data similarity**
- **Above steps are the beginning of data preprocessing.**
- **Many methods have been developed but still an active area of research.**



References

- **W. Cleveland, Visualizing Data, Hobart Press, 1993**
- **T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003**
- **U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001**
- **L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.**
- **H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997**
- **D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002**
- **D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999**
- **S. Santini and R. Jain, "Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999**



数据集下载整理

- 下载一套数据集，描述数据集的字段含义，构思根据这些数据集可能得出什么知识？