

# 数据预处理

## Data preprocessing

2019 年 1 月

# 第2章 数据获取与预处理

- 数据获取
- 数据评估与统计
- 数据清洗
- 数据整理
- 数据可视化
- 数据获取与预处理实践

## 数据的概念

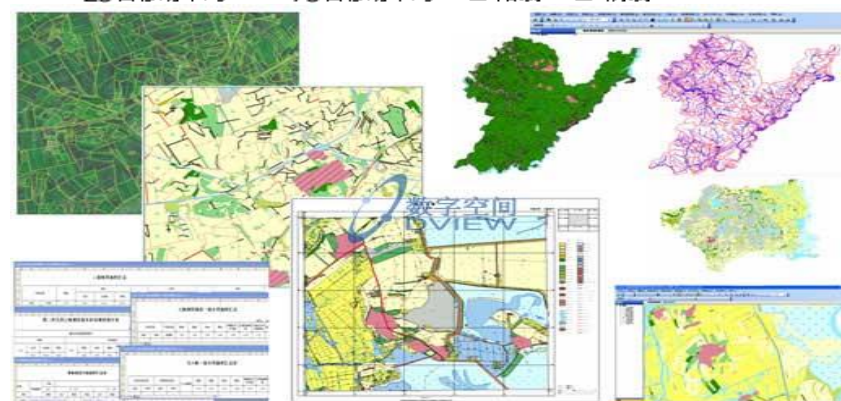
- 事务数据库
- 多媒体数据库
- 异构和遗留数据库
- 时序数据序列
- 互联网数据
- 空间数据库

	A	B	C	D	E	F	G	H	I	J
1	目的IP	记录数	源IP数	源IP熵	源端口数	源端口熵	目的端口数	目的端口熵	上行数据包	上行数据
2	ip	D1_redCou	D1_sIpNum	D3_sIpH	D1_sPortN	D3_sPortH	D1_dPortN	D3_dPortH	D3_pkgUpA	D3_pkgUp
3	220.181.7	54	11	0.762	54	1	1	0	1656.4	3812.
4	218.92.22	1	1	0	1	0	1	0	9571811	
5	118.248.1	1	1	0	1	0	1	0	373	
6	123.125.1	157	1	0	157	1	1	0	1938.9	6030.
7	123.150.1	2	1	0	2	1	1	0	1183.5	941.
8	220.181.1	2	1	0	2	1	1	0	473	

10 [000533] 三个月 日K 2016/2/5



— 25日移动平均 — 75日移动平均 ■ 阳线 ■ 阴线

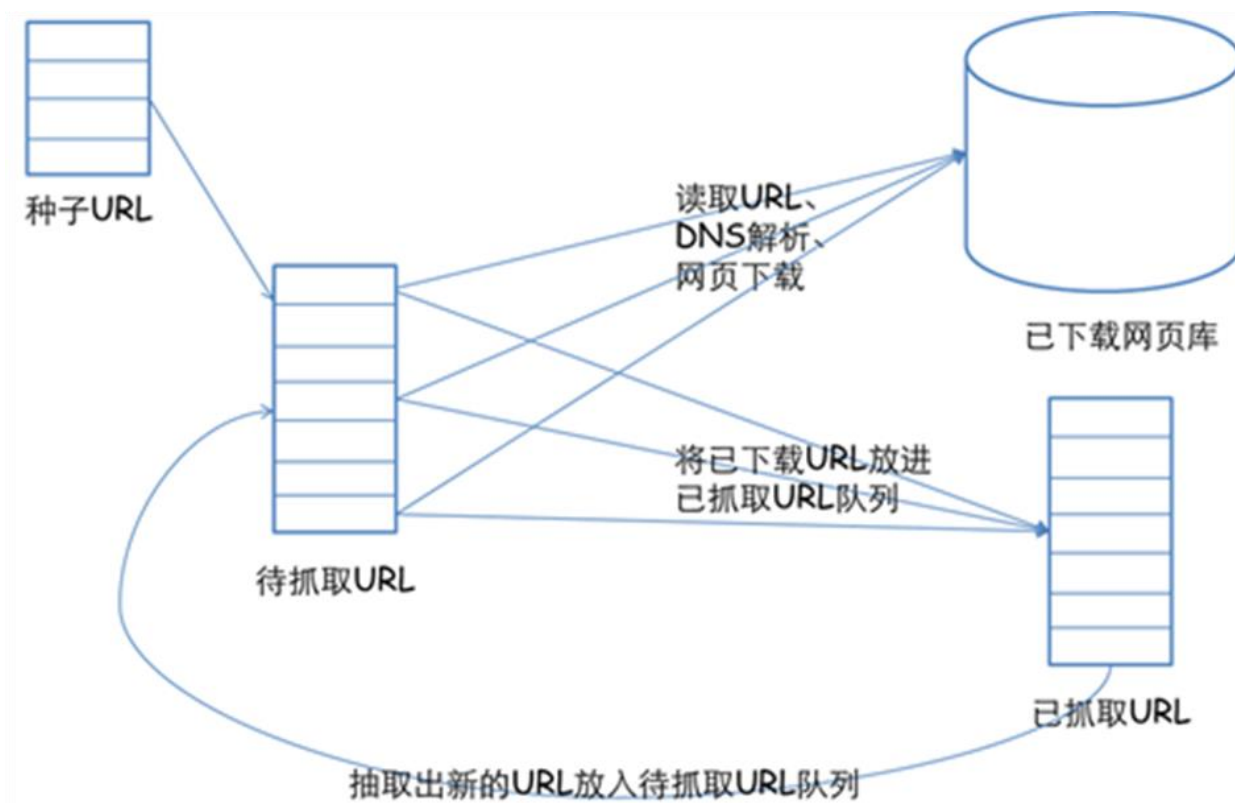


## 爬虫技术

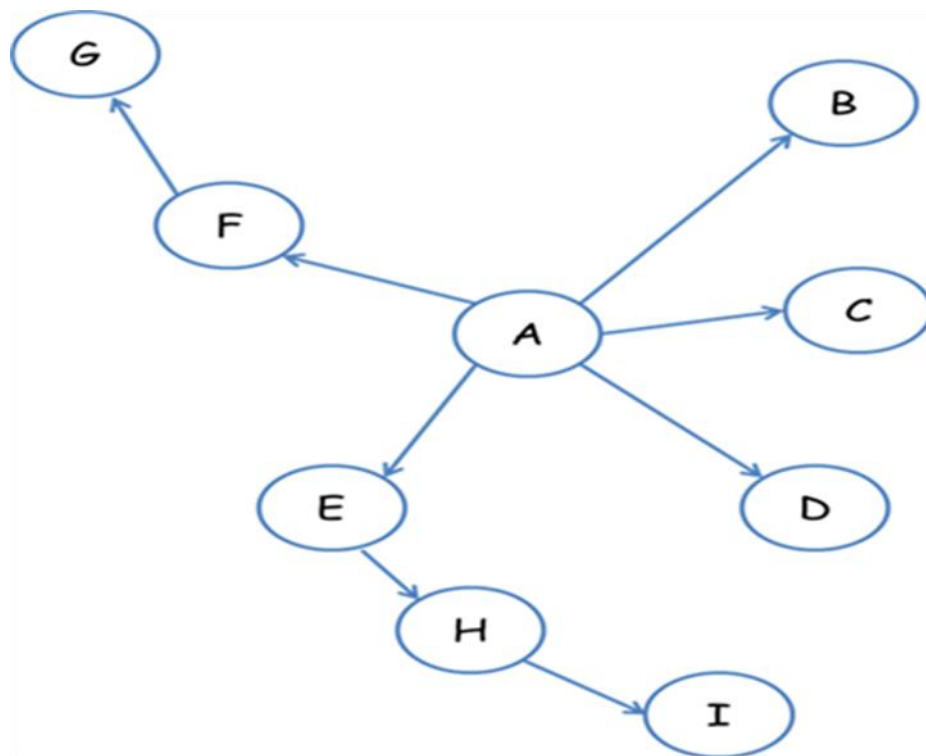
- 网络爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁，自动索引，模拟程序或者蠕虫。
- 网络爬虫是搜索引擎抓取系统的重要组成部分。爬虫的主要目的是将互联网上的网页下载到本地形成一个或联网内容的镜像备份。



## 爬虫框架

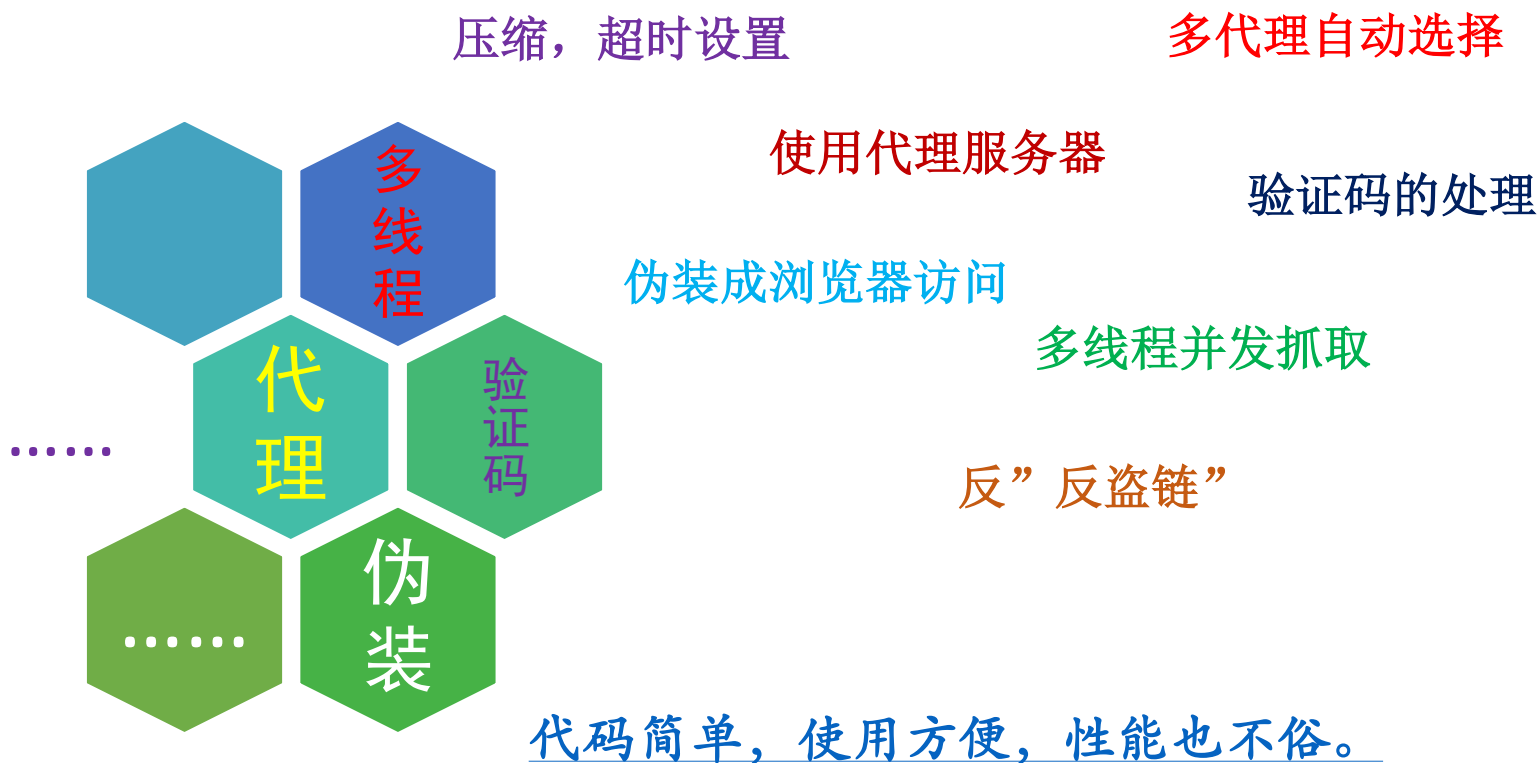


## 爬取策略



- 深度优先遍历策略
  - A-F-G E-H-I B C D
- 宽度优先遍历策略
  - A-B-C-D-E-F G H I
- 反向链接数策略
- Partial PageRank策略
- OPIC策略
- 大站优先策略
- robots.txt (爬虫协议)
- sitemap设置

## 爬虫技巧



# 第2章 数据获取与预处理

- 数据获取
- 数据评估与统计
- 数据清洗
- 数据整理
- 数据可视化
- 数据获取与预处理实践



## 数据质量

### 准确性

- 数据记录的信息是否存在异常或错误

### 完整性

- 数据是否存在缺失，缺失可能是整个记录数据，也可能是记录数据中某个关键字段的缺失

### 一致性

- 数据是否遵循了统一的规范，数据集是否保持统一的格式

### 时效性

- 数据从产生到可以查看的时间间隔

## 数据描述

### 中心趋势度量

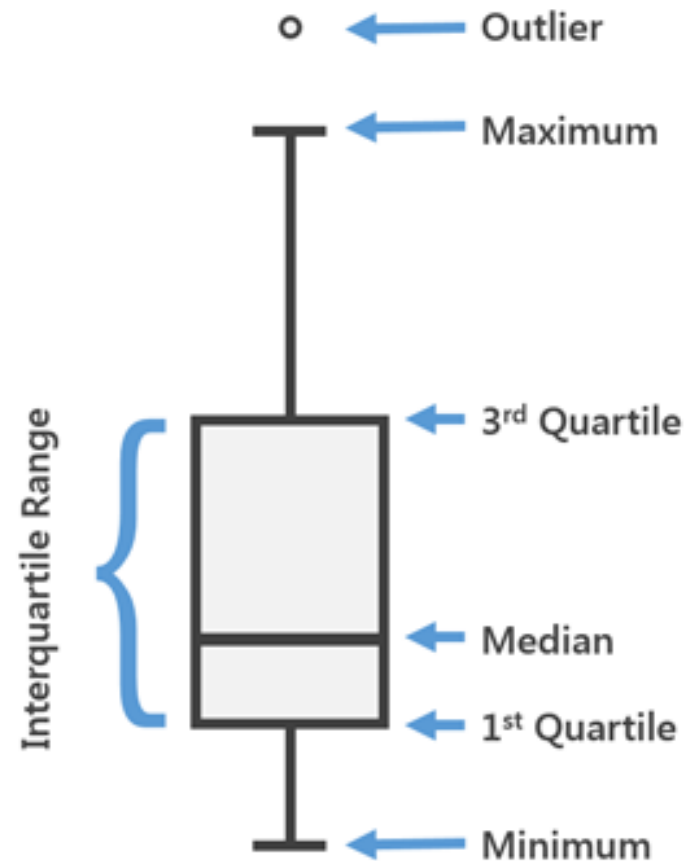
- 算数均值
- 加权平均
- 截断均值
- 中位数
- 众数
- 中列数

### 离散趋势度量

- 方差
- 极差
- 百分位数
- 四分位数
- 四分位距
- 五数概括

## 盒图(Box plot)

- 最小值(min), 下四分位数(Q1), 中位数(median), 上四分位数(Q3), 最大值(max)
- 下四分位数、中位数、上四分位数组成一个“带有隔间的盒子”。上、下四分位数到最大、最小值之间建立一条延伸线
- $IQR = Q3 - Q1$ , 即上四分位数与下四分位数之间的差, 也就是盒子的长度



## 数据相似性

- 欧几里得距离，欧氏距离，Euclidean

$$d(i, j) = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \dots + (X_{in} - X_{jn})^2}$$

- 曼哈顿距离，Manhattan

$$d(i, j) = |X_{i1} - X_{j1}| + |X_{i2} - X_{j2}| + \dots + |X_{in} - X_{jn}|$$

- 切比雪夫距离，上确界距离，Chebyshev

$$d(i, j) = \max\{|X_{i1} - X_{j1}|, |X_{i2} - X_{j2}|, \dots, |X_{in} - X_{jn}|\}$$

- 闵可夫斯基距离，Minkowski

$$d(i, j) = \sqrt[k]{|X_{i1} - X_{j1}|^k + |X_{i2} - X_{j2}|^k + \dots + |X_{in} - X_{jn}|^k}$$

# 第2章 数据获取与预处理

- 数据获取
- 数据评估与统计
- 数据清洗
- 数据整理
- 数据可视化
- 数据获取与预处理实践

## 数据缺失值填充

### 忽略元组

通常当在缺少类标号时，通过这样的方法来填补缺失值

### 用属性的均值填充缺失值

数据属性分为数值属性和非数值属性进行处理，通过利用已存数据的多数信息来推测缺失值

### 人工填写缺失值

数据偏离的问题小，但该方法十分费时,不具备实际的可操作性

### 填充 缺失 值

### 用同类样本的属性均值填充缺失值

利用均值替换缺失值

### 使用一个全局常量填充缺失值

大量采用同一属性值，可能会误导挖掘程序得出有偏差甚至错误的结论

### 使用最可能的值填充缺失值

数据属性分为数值属性和非数值属性进行处理，通过利用已存数据的多数信息来推测缺失值

数据清洗可以视为一个过程，包括检测偏差与纠正偏差两个步骤：



## 1 检查偏差

可以使用已有的关于数据性质的知识发现噪声、离群点和需要考察的不寻常的值。这种知识或“关于数据的数据”称为元数据。



## 2 纠正偏差

即一旦发现偏差，通常需要定义并使用一系列的变换来纠正它们。但这些工具只支持有限的变换，因此，常常可能需要为数据清洗过程的这一步编写定制的程序。

噪声是被测量的变量的随机误差或方差。给定一个数值属性，如何才能使数据“光滑”，去掉噪声？下面给出数据光滑技术的具体内容。

01

分箱

分箱方法通过考察某一数据周围数据的值，即“近邻”来光滑有序数据的值。

02

回归

光滑数据可以通过一个函数拟合数据来实现。线性回归的目标就是查找拟合两个属性的“最佳”线，使得其中一个属性可以用于预测出另一个属性。

03

聚类

离群点可通过聚类进行检测，将类似的值组织成群或簇，离群点即为落在簇集合之外的值。许多数据光滑的方法也是涉及离散化的数据归约方法。



## 分箱法

- 分箱：

就是将数据按照属性值划分的子区间。如果某一属性值处于某个子区间范围内，就称把该属性值放进这个子区间所代表的“箱子”内。

需要确定的两个主要问题：

- 1.如何分箱？
- 2.如何对每个箱子中的数据进行平滑处理？

## 分箱法

分箱的方法有：等深分箱法、等宽分箱法和用户自定义区间法。

等深分箱法：将数据集按记录行数分箱，每箱具有相同的记录数，每箱记录数称为箱子的深度。

**例1**的等深分箱法：设定权重（箱子深度）为4

- 箱1： 800,1000,1200,1500      箱2： 1500,1800,2000,2300
- 箱3： 2500,2800,3000,3500      箱4： 4000,4500,4800,5000
- 箱5： **400000**

## 分箱法

等宽分箱法，使数据集在整个属性值的区间上平均分布，即每个箱的区间范围是一个常量，称为箱子宽度。

**例1**等宽分箱法：设定区间范围（箱子宽度）为1000元人民币

箱1： 800,1000,1200,1500,1500,1800

箱2： 2000,2300,2500,2800,3000

箱3： 3500,4000,4500

箱4： 4800, 5000

箱5： 400000

## 分箱法

用户自定义区间，用户根据需要自定义区间，当用户明确希望观察某些区间范围内的数据分布时，使用这种方法可以方便地帮助用户达到目的。

**例1**用户自定义区间法：将收入划分为1000元以下、1000~2000、2000~3000、3000~4000和4000元以上

箱1： 800      箱2： 1000,1200,1500,1500,1800,2000

箱3： 2300,2500,2800,3000

箱4： 3500,4000      箱5： 4500,4800, 5000 , 400000

## 分箱法平滑

数据平滑方法：

按平均值平滑、按边界值平滑和按中值平滑

例1 用户自定义区间法的均值平滑

箱1：800 → 800

箱2：1000,1200,1500,1500,1800,2000 →  
1500,1500,1500,1500,1500,1500

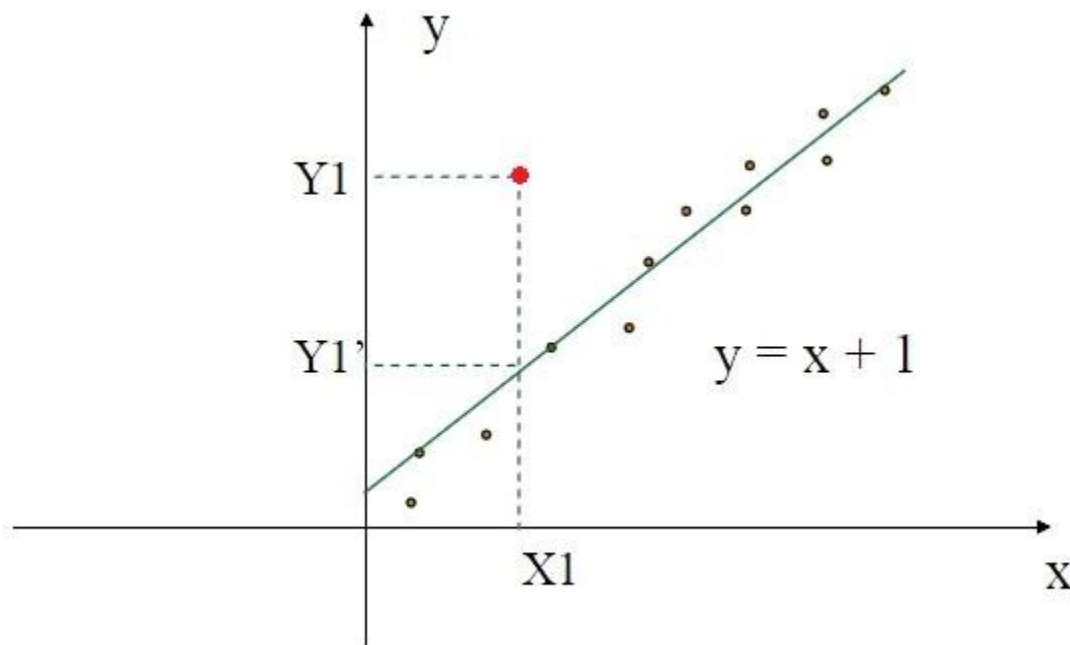
箱3：2300,2500,2800,3000 → 2300,2300,3000,3000

箱4：3500,4000 → 3750, 3750

箱5：4500,4800, 5000, 400000 → 4900, 4900, 4900,4900

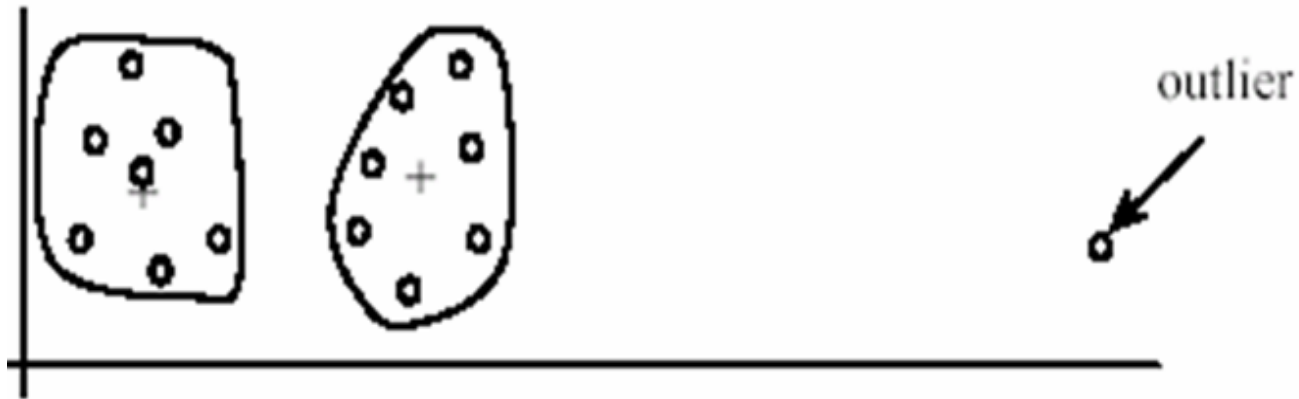
## 回归

发现两个相关的变量之间的变化模式，通过拟合函数来平滑数据，即建立数学模型来预测下一个数值，包括线性回归和非线性回归。



## 聚类

聚类方法将物理的或抽象对象的集合分组为由类似的对象组成的多个簇类.找出并清除那些落在簇之外的值（孤立点），这些孤立点被视为噪声。



# 第2章 数据获取与预处理

---

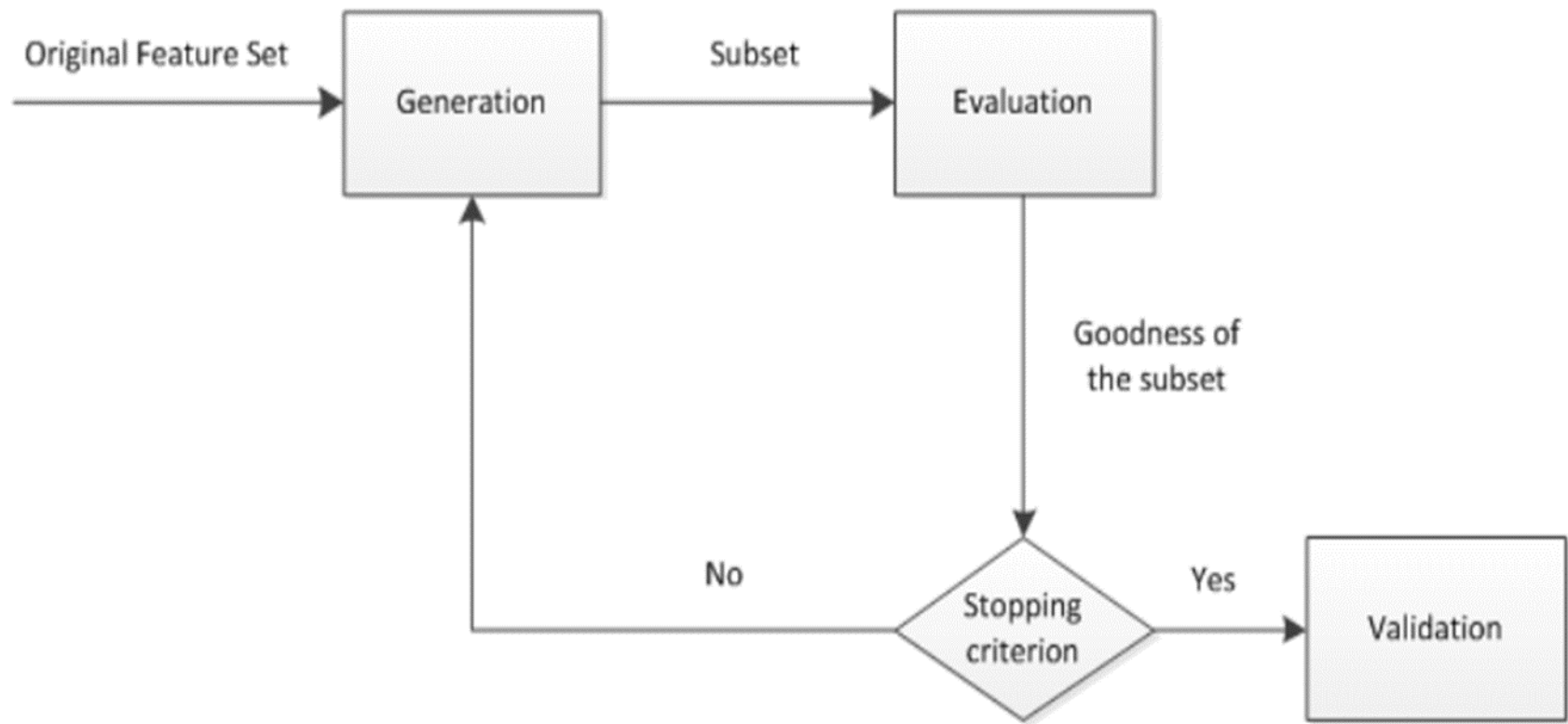
- 数据获取
- 数据评估与统计
- 数据清洗
- 数据整理
- 数据可视化
- 数据获取与预处理实践



## 特征选择

- 什么是特征选择？
- 从全部特征中选取一个特征子集，使得系统的特定指标最优化
- 为什么要进行特征选择？
  1. 剔除不相关(irrelevant)或冗余(redundant)的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的
  2. 可能存在与研究内容不相关的特征，特征之间也可能存在相互依赖
  3. 特征个数多，容易引起“维度灾难”，模型会越来越复杂，其推广能力会下降

## 特征选择



## 产生过程

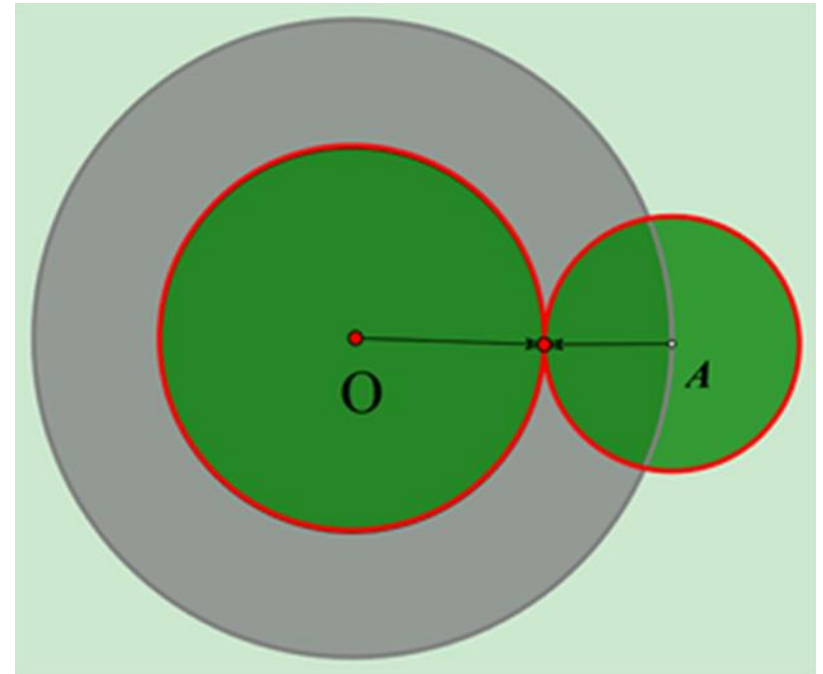
- 搜索特征子空间的过程
- 完全搜索(Complete)
  - 穷举搜索(Exhaustive)
  - 非穷举搜索(Non-Exhaustive)
- 启发式搜索(Heuristic)
- 随机搜索(Random)

## 完全搜索

- 广度优先搜索( Breadth First Search )
- 分支限界搜索( Branch and Bound )
- 定向搜索 (Beam Search )
- 最优优先搜索 ( Best First Search )

## 启发式搜索

- 序列前向选择(SFS)
- 序列后向选择(SBS)
- 双向搜索(BDS)
- 增L去R选择算法 (LRS)
- 序列浮动选择
  - 序列浮动前向选择(SFFS)
  - 序列浮动后向选择(SFBS)
- 决策树(DTM)



## 随机搜索

- 随机产生序列选择算法(RGSS)
- 模拟退火算法(SA)
- 遗传算法(GA)
- 依赖于随机因素，有实验结果难以重现

## 评价函数

- 评价函数的作用是评价产生过程所提供的特征子集的好坏。
- **筛选器(Filter)**通过分析特征子集内部的特点来衡量其好坏；筛选器一般用作预处理，与分类器的选择无关。
- **封装器(Wrapper)**实质上是一个分类器，封装器用选取的特征子集对样本集进行分类，分类的精度作为衡量特征子集好坏的标准。
- 常见的评价函数
  - a.相关性
  - b.距离
  - c.信息增益
  - d.一致性
  - e.分类器错误率

## 评价函数

- 信息增益( **Information Gain** )
- 假设：存在特征子集A和特征子集B，分类变量为C，若 $IG(C|A) > IG(C|B)$ ，则认为选用特征子集A的分类结果比B好，因此倾向于选用特征子集A。x代表某一特征，H代表信息熵。

$$IG(x) = H(c) - H(c/x)$$

$$- \sum_{i=1}^n p(c_i) \log_2 p(c_i) + p(x) \sum_{i=1}^n p(c_i|x) \log_2 p(c_i|x) + p(\bar{x}) \sum_{i=1}^n p(c_i|\bar{x}) \log_2 p(c_i|\bar{x})$$



## 评价函数

- 一致性( **Consistency** )
- 假设：若样本1与样本2属于不同的分类，但在特征A、B上的取值完全一样，那么特征子集{A, B}不应该选作最终的特征集。
- 分类器错误率 ( **Classifier error rate** )
- 使用特定的分类器，用给定的特征子集对样本集进行分类，用分类的精度来衡量特征子集的好坏。

## 特征提取

- 定义

将原始特征集**转换**为一组具有明显物理意义或者统计意义或核的特征。

- 特征选择和特征提取的关系

**共同点：** 1.减少数据存储和输入数据带宽；  
2.减少冗余和部分噪声；  
3.发现更有意义的潜在特征，加深数据理解。

**不同点：** 1.挑选 VS 转换；（方法）  
2.直观 VS 抽象。（结果可读性）

## 特征提取

### 经典算法：

1.线性特征提取: 主成分分析 (PCA), 线性判别分析 (LDA), 独立成分分析 (ICA)。

2.非线性特征提取: 核主成分分析 (KPCA), 核费舍尔判别分析 (KFDA), 流形学习(Manifold Learning), 深度学习(Deep Learning)。

## 主成份分析

主成分分析（**principal components analysis**）也称主分量分析，是由霍特林（Hotelling）于1933年首先提出的。主成分分析是利用降维的思想，在损失很少信息的前提下把多个特征转化为几个综合特征的多元统计方法。通常把转化生成的综合特征称之为主成分。

- 每一个主成分都是各原始变量的线性组合；
- 主成分的数目大大少于原始变量的数目；
- 主成分保留了原始变量绝大多数信息；
- 各主成分之间互不相关。

## PCA形式化解释

- 特征的变异性
- 当一个特征只取一个数据时，这个特征（数据）提供的信息量是非常有限的，当这个特征取一系列不同数据时，我们可以从中读出最大值、最小值、平均数等信息。特征的变异性越大，说明它对各种场景的“遍历性”越强，提供的信息就更加充分，信息量就越大。主成分分析中的信息，就是特征的变异性，用标准差或方差表示它。

## PCA数学描述

设对某一事物的研究涉及  $p$  个特征，分别用  $X_1, X_2, \dots, X_p$  表示，这  $p$  个特征构成的  $p$  维随机向量为  $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ 。设随机向量  $\mathbf{X}$  的均值为  $\mu$ ，协方差矩阵为  $\Sigma$ 。

对  $\mathbf{X}$  进行线性变换，可以形成新的综合特征，用  $\mathbf{Y}$  表示，也就是说，新的综合变量可以由原来的特征线性表示，即满足下式：

$$\begin{cases} Y_1 = t_{11}X_1 + t_{12}X_2 + \dots + t_{1p}X_p \\ Y_2 = t_{21}X_1 + t_{22}X_2 + \dots + t_{2p}X_p \\ \dots\dots\dots \\ Y_p = t_{p1}X_1 + t_{p2}X_2 + \dots + t_{pp}X_p \end{cases}$$

## PCA数学描述

将线性变换约束在下面的原则之下：

1.  $T_i' T_i = 1$  , 即:  $T_{i1}^2 + T_{i2}^2 + \cdots + T_{ip}^2 = 1$  ( $i = 1, 2, \dots, p$ )

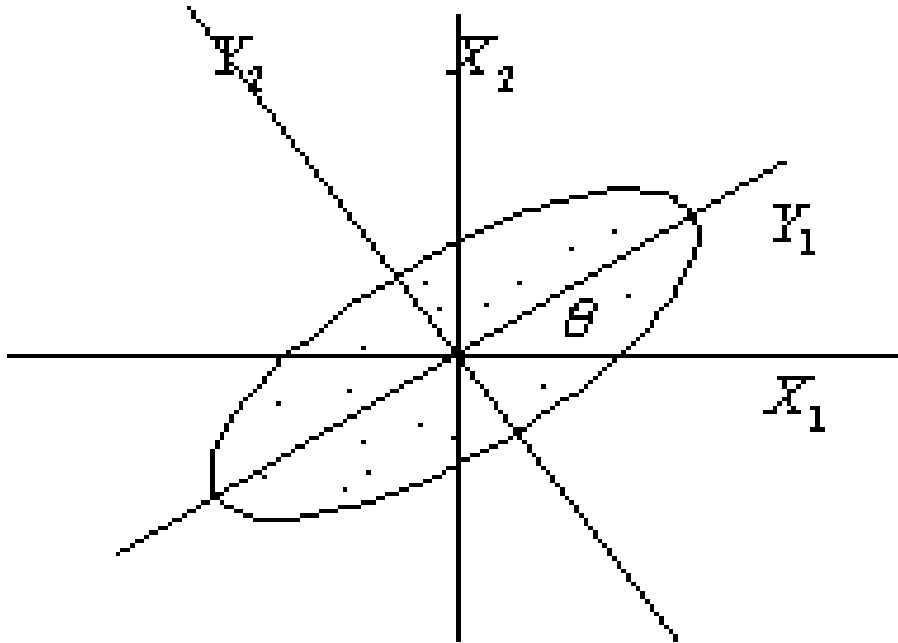
2.  $Y_i$ 与 $Y_j$ 相互无关 ( $i \neq j; i, j = 1, 2, \dots, p$ )

3.  $Y_1$ 是  $X_1, X_2, \dots, X_p$  上一切满足原则1的线性组合中方差最大者； $Y_2$ 是与 $Y_1$ 不相关的所有线性组合中方差最大者；...,  $Y_p$ 是与  $Y_1, Y_2, \dots, Y_{p-1}$  都不相关的所有线性组合中方差最大者。

从数学定义上来讲，PCA回答的问题是：如何找到另一组正交基，它们是标准正交基的线性组合，能够最好的表示数据集

## PCA几何意义

假设共有 $n$ 个样品，每个样品都测量了两个特征（ $X_1$ ,  $X_2$ ），它们大致分布在一个椭圆内如图所示。



问题：  
如果仅考虑 $X_1$ 或 $X_2$ 中的任何一个，那么包含在另一分量中的信息将会损失。

主成分的几何意义



## PCA几何意义

如果我们将该坐标系按逆时针方向旋转某个角度  $\theta$ ，变成新坐标  $y_1 O y_2$ ，这里  $y_1$  是椭圆的长轴方向， $y_2$  是椭圆的短轴方向。旋转公式为

$$\begin{cases} Y_1 = X_1 \cos \theta + X_2 \sin \theta \\ Y_2 = -X_1 \sin \theta + X_2 \cos \theta \end{cases}$$

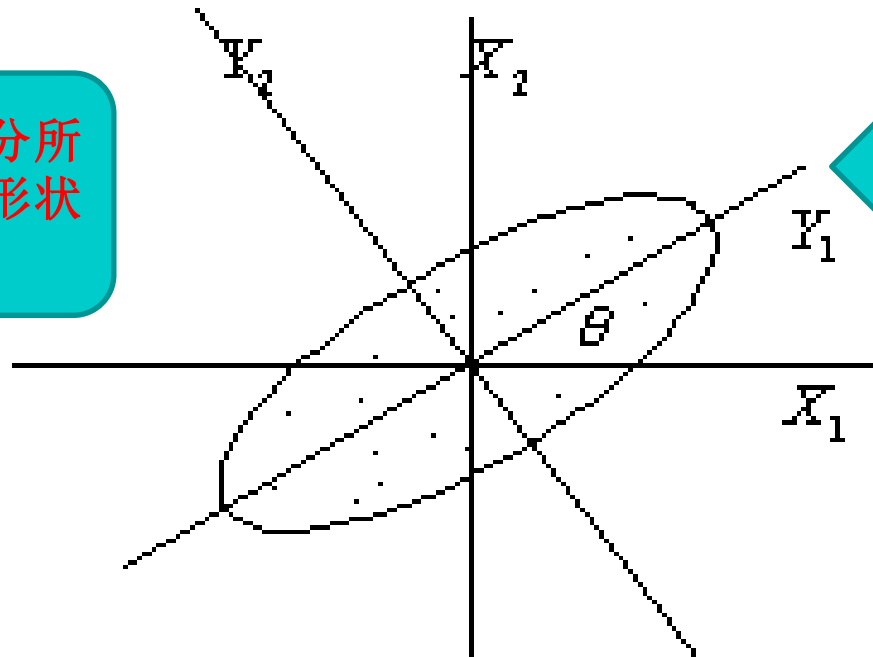
我们看到新变量  $Y_1$  和  $Y_2$  是原变量  $X_1$  和  $X_2$  的线性组合，它的矩阵表示形式为

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = T' X$$

其中， $T'$  为旋转变换矩阵，它是**正交矩阵**，即有  $T' = T^{-1}$  或  $T' T = I$ 。

## PCA几何意义

思考：第一主成分所含信息量与椭圆形状的关系？



$n$ 个点在  $Y_1$  轴上的方差达到最大，即在此方向上包含了有关  $n$  个样品的最大信息量

$n$ 个点在新坐标系下的坐标  $Y_1$  和  $Y_2$  几乎不相关。称它们为原始变量  $X_1$  和  $X_2$  的综合变量。称  $Y_1$  为第一主成分，称  $Y_2$  为第二主成分。

因此，欲将二维空间的点投影到某个一维方向上，则选择  $y_1$  轴方向能使信息的损失最小。

## PCA几何意义

考虑两种极端的情况：

1. 椭圆的长轴与短轴的长度相等，即椭圆变成圆。
2. 椭圆扁平到了极限，变成y1轴上的一条线。

情况1：第一主成分只含有二维空间点的约一半信息，若仅用这一个综合变量，则将损失约50%的信息，这显然是不可取的。造成它的原因是，原始变量 $X_1$ 和 $X_2$ 的相关程度几乎为零，也就是说，它们所包含的信息几乎不重迭，因此无法用一个一维的综合变量来代替。

情况2：第一主成分包含有二维空间点的全部信息，仅用这一个综合变量代替原始数据不会有任何的信息损失，此时的主成分分析效果是非常理想的，其原因是，第二主成分不包含任何信息，舍弃它当然没有信息损失。

## PCA流程

- 主成分分析的步骤对此进行归纳如下：

1. 根据研究问题选取初始分析变量；
2. 根据初始特征的特性判断由协方差阵求主成分还是由相关阵求主成分；
3. 求协差阵或相关阵的特征根与相应标准特征向量；
4. 判断是否存在明显的多重共线性，若存在，则回到第一步；
5. 得到主成分的表达式并确定主成分个数，选取主成分；
6. 结合主成分对研究问题进行分析并深入研究。

## 数据集成

数据挖掘经常需要数据集合并来自多个数据存储的数据。数据还可能需变换成适于挖掘的形式。数据分析任务多半涉及数据集成。

问题

(1) 模式集成和对象匹配问题

(2) 冗余问题

(3) 元组重复

(4) 数据值冲突的检测与处理问题

## 数据转换

数据变换的目的是将数据变换或统一成适合挖掘的形式。数据变换主要涉及以下内容：

1、光滑。去除数据中的噪声

2、聚集。对数据进行汇总或聚集。

3、数据泛化。使用概念分层，用高层概念替换低层或“原始”数据

4、规范化。将属性数据按比例缩放，使之落入一个小的特定区间

5、属性构造。可以构造新的属性并添加到属性集中，以帮助挖掘过程

## 数据转换

- 数据规范化

- 最大-最小规范化 
$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

- z-score 规范化（0均值标准化） 
$$v_i' = \frac{v_i - \text{avg}_A}{\delta_A}$$

- 小数定标规范化 
$$v_i' = \frac{v_i}{10^j}$$

- 数据离散化

- 等距离散化

- 等频离散化

- 聚类离散化

- 有监督的离散化

# 第2章 数据获取与预处理

- 数据获取
- 数据评估与统计
- 数据清洗
- 数据整理
- 数据可视化
- 数据获取与预处理实践



## 大数据可视化

### ◆ 大数据可视化核心问题

随着互联网技术的发展，尤其是移动互联技术的发展，网络空间的数据量呈现出爆炸式增长。如何从这些数据中快速获取自己想要的信息，并以一种直观、形象的方式展现出来？这就是大数据可视化要解决的核心问题。

### ◆ 数据可视化解释

数据可视化，最早可追溯到20世纪50年代，它是一门关于数据视觉表现形式的科学技术研究。数据可视化是一个处于不断演变之中的概念，其边界在不断地扩大，主要指的是技术上较为高级的技术方法，而这些技术方法允许利用图形图像处理、计算机视觉及用户界面，通过表达、建模，以及对立体、表面、属性及动画的显示，对数据加以可视化解释。

## 可视化的基本特征



## 可视化的目标 and 作用

数据可视化的作用主要包括数据表达、数据操作和数据分析3个方面，它是以可视化技术支持计算机辅助数据认识的3个基本阶段。

### 数据表达

数据表达是通过计算机图形图像技术来更加友好地展示数据信息，方便人们阅读、理解和运用数据。常见的形式如文本、图表、图像、二维图形、三维模型、网络图、树结构、符号和电子地图等。

### 数据操作

数据操作是以计算机提供的界面、接口、协议等条件为基础完成人与数据的交互需求，数据操作需要友好的人机交互技术、标准化的接口和协议支持来完成对多数据集合或者分布式的操作。

### 数据分析

数据可视化可以有效地表达数据的各类特征，帮助人们推理和分析数据背后的客观规律，进而获得相关知识，提高人们认识数据的能力和利用数据的水平。

## 数据可视化流程

1

### 数据获取

主动式是以明确的数据需求为目的，如卫星影像、测绘工程等；被动式是以数据平台为基础，由数据平台的活动者提供数据来源，如电子商务、网络论坛等。

2

### 数据处理

数据处理是指对原始的数据进行质量分析、预处理和计算等步骤。数据处理的目标是保证数据的准确性、可用性。

3

### 可视化模式

可视化模式是数据的一种特殊展现形式，常见的可视化模式有标签云、序列分析、网络结构、电子地图等。可视化模式的选取决定了可视化方案的雏形。

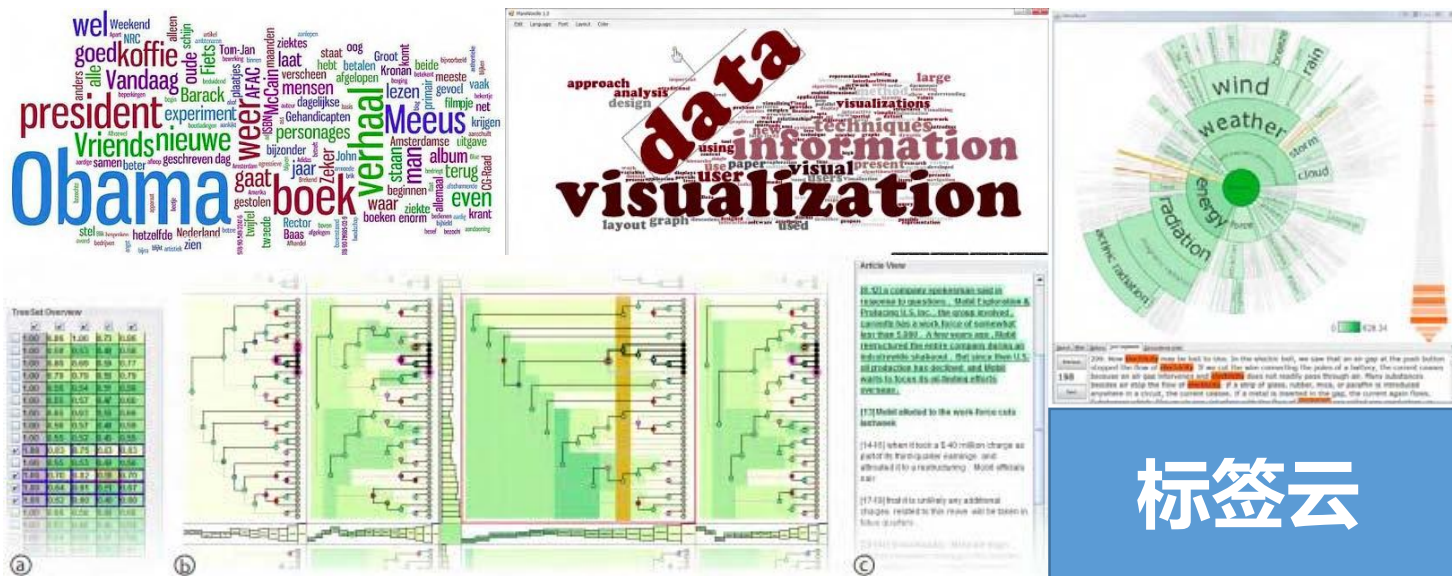
4

### 可视化应用

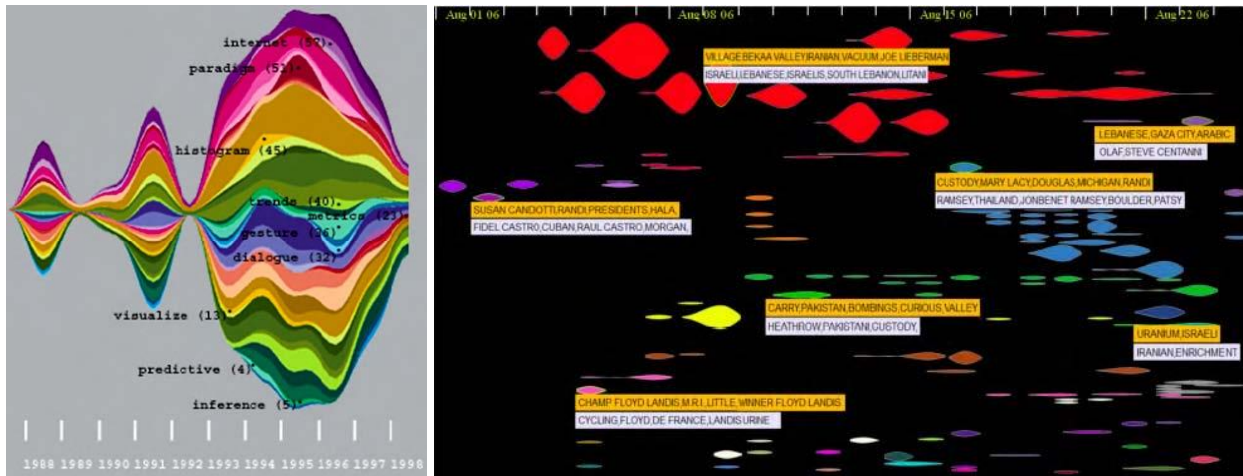
可视化应用主要根据用户的主观需求展开，最主要的应用方式是用来观察和展示，通过观察和人脑分析进行推理和认知，辅助人们发现新知识或者得到新结论。

## 文本可视化

如图所示是一种称为标签云（Word Clouds或Tag Clouds）的典型的文本可视化技术。它将关键词根据词频或其他规则进行排序，按照一定规律进行布局排列，用大小、颜色、字体等图形属性对关键词进行可视化。一般用字号大小代表该关键词的重要性，该技术多用于快速识别网络媒体的主题热度。



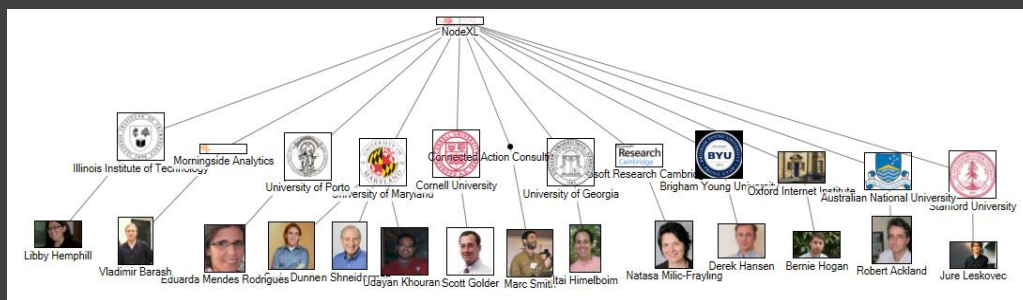
## 动态文本时序信息可视化



有些文本的形成和变化过程与时间是紧密相关的，因此，如何将动态变化的文本中时间相关的模式与规律进行可视化展示，是文本可视化的重要内容。引入时间轴是一类主要方法，常见的技术以河流图居多。河流图按照其展示的内容可以划分为主题河流图、文本河流图及事件河流图等。

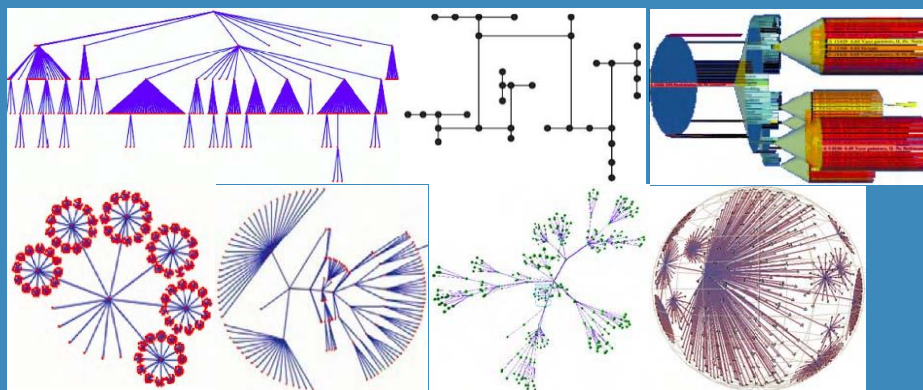


## 网络图可视化

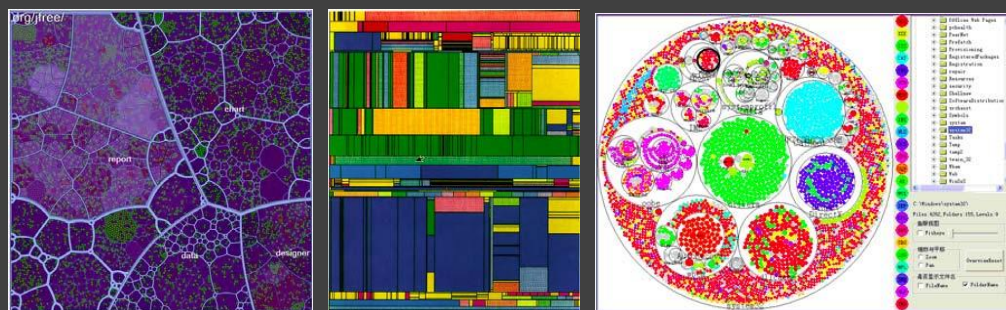


(1) Nodal研究人员及其组织机构社会网络图

(2) 基于节点连接的图和树可视化方法



## 网络图可视化



(3) 基于空间填充的树可视化

(4) 基于边捆绑的大规模密集图可视化

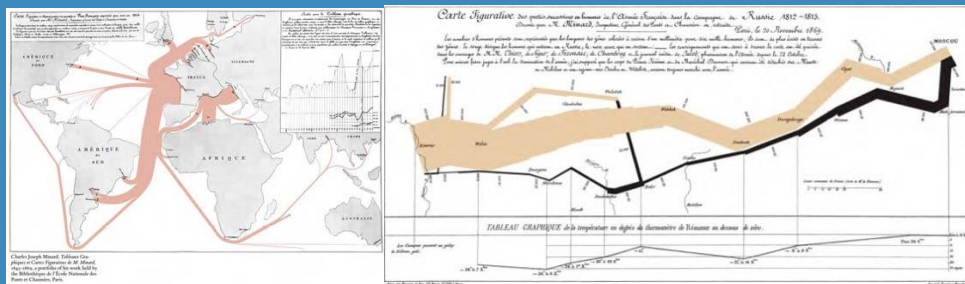




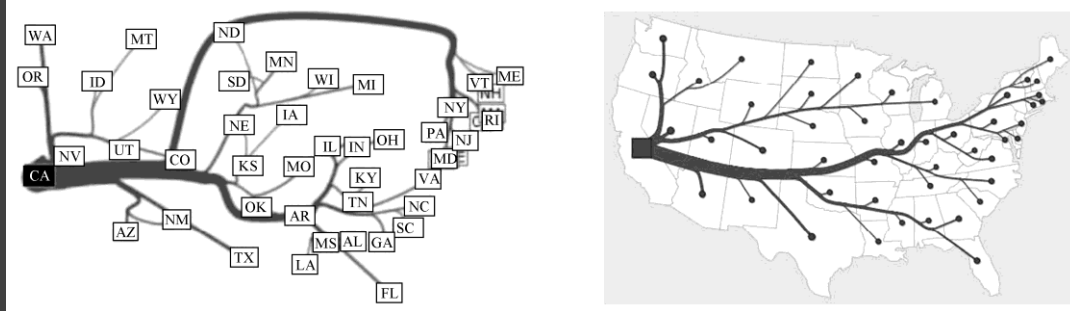
# 数据可视化

## 时空数据可视化

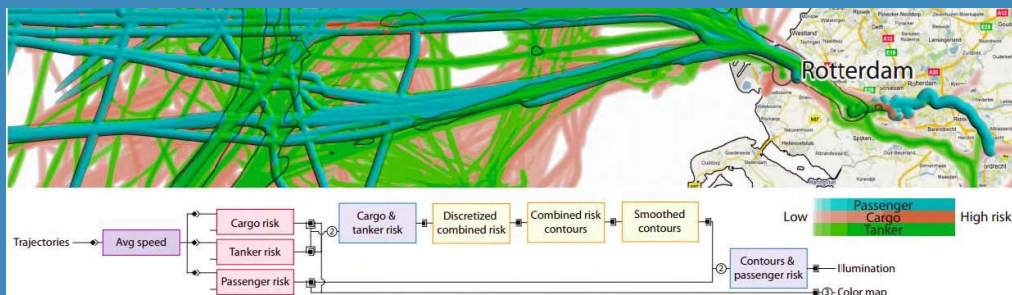
流式地图



结合了捆绑技术的流式地图



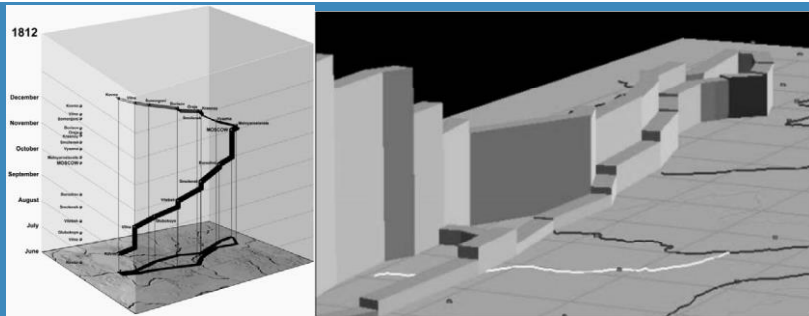
结合了密度图技术的流式地图



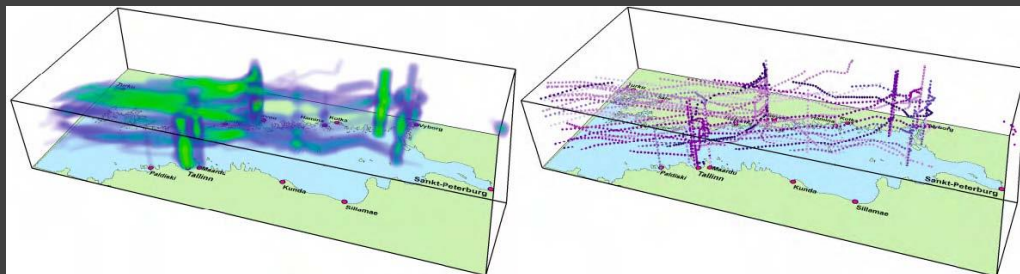
# 数据可视化

## 时空数据可视化

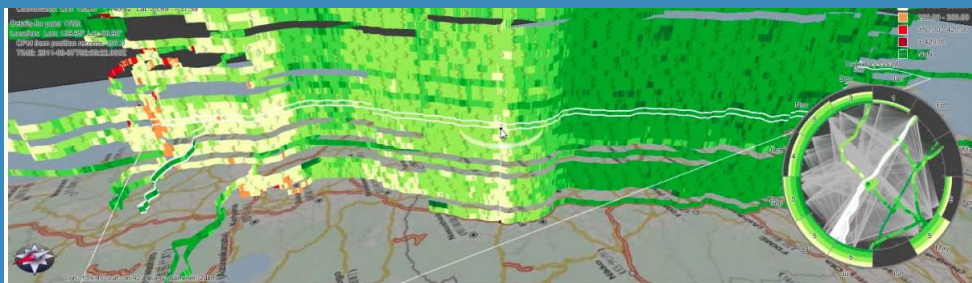
时空立方体



融合散点图与密度图技术的时空立方体



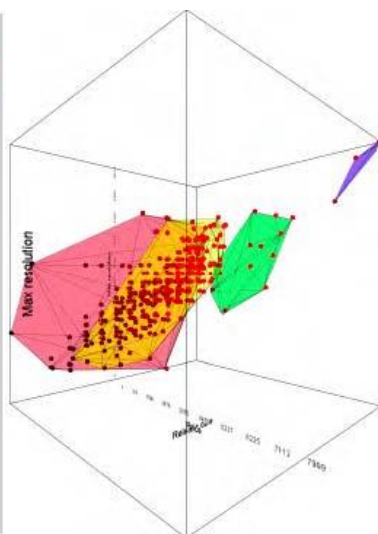
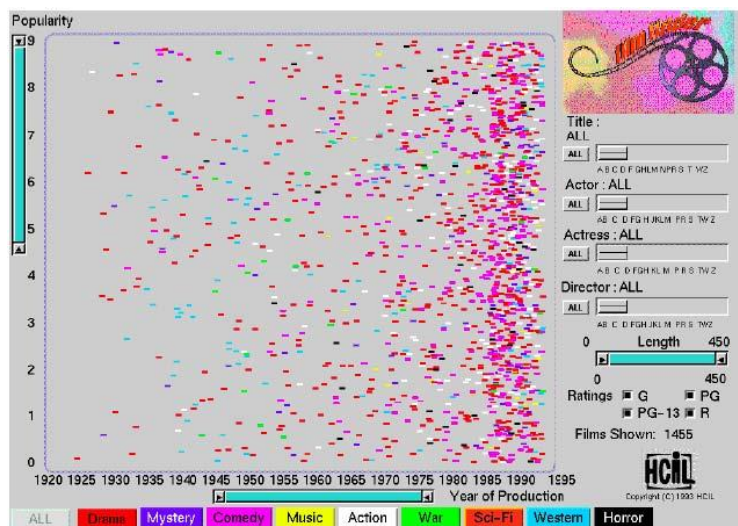
融合堆积图技术的时空立方体



## 多维数据可视化

### 1、散点图（Scatter Plot）

散点图（Scatter Plot）是最为常用的多维可视化方法。二维散点图将多个维度中的两个维度属性值集合映射至两条轴，在二维轴确定的平面内通过图形标记的不同视觉元素来反映其他维度属性值。



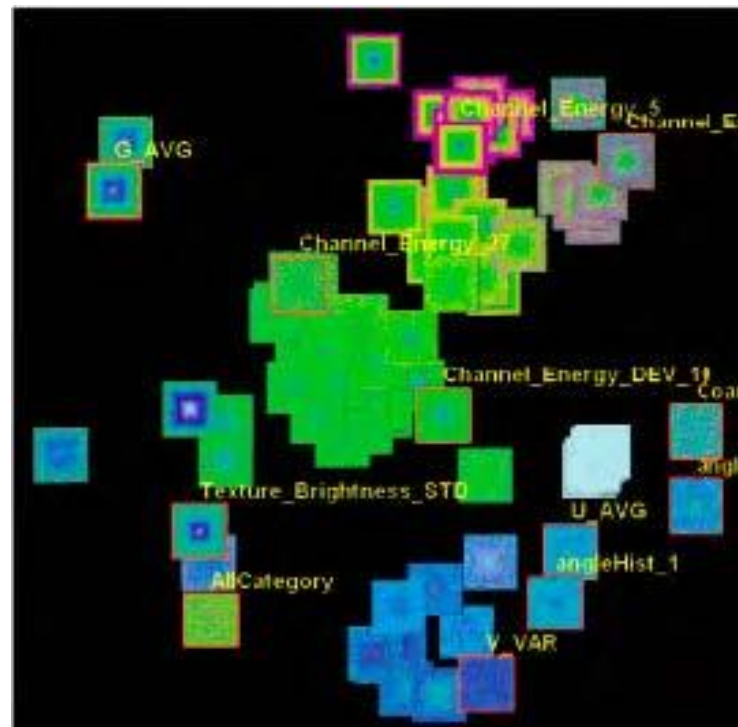
二维散点图能够展示的维度十分有限，研究者将其扩展到三维空间，通过可旋转的Scatter Plot方块（dice）扩展了可映射维度的数目，如图所示。散点图适合对有限数目的较为重要的维度进行可视化，通常不适用于需要对所有维度同时进行展示的情况。

## 多维数据可视化

### 2、投影（Projection）

投影是能够同时展示多维的可视化方法之一。VaR将各维度属性列集合通过投影函数映射到一个方块形图形标记中，并根据维度之间的关联度对各个小方块进行布局。

基于投影的多维可视化方法一方面反映了维度属性值的分布规律，同时也直观地展示了多维度之间的语义关系。

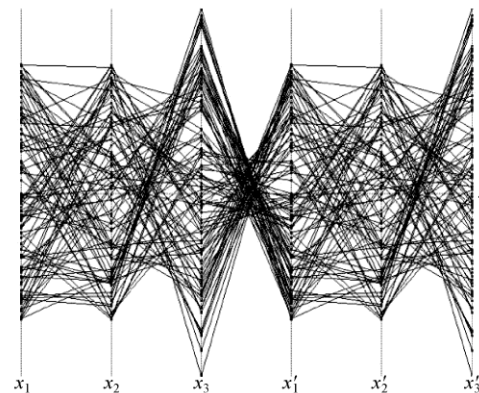
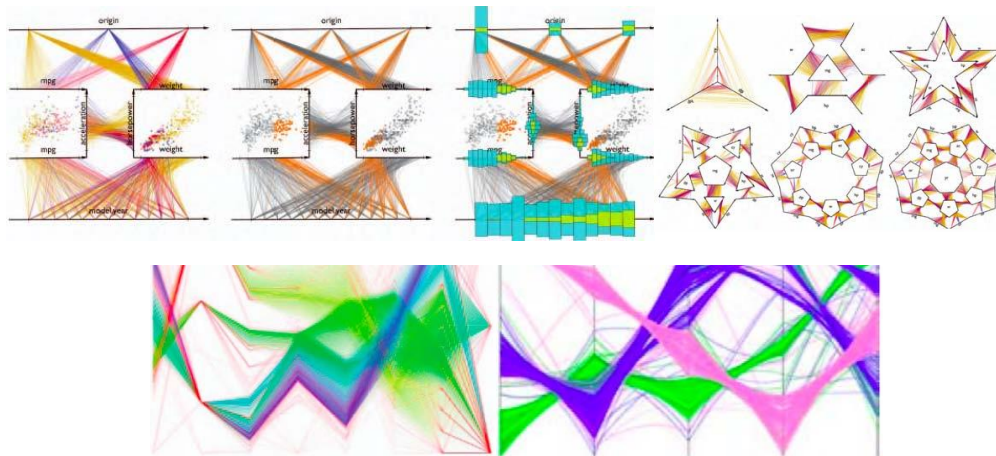




## 多维数据可视化

### 3、平行坐标 (Parallel Coordinates)

平行坐标是研究和应用最为广泛的一种多维可视化技术，将维度与坐标轴建立映射，在多个平行轴之间以直线或曲线映射表示多维信息。



平行坐标多维可视化技术

集成了散点图和柱状图的平行坐标工具

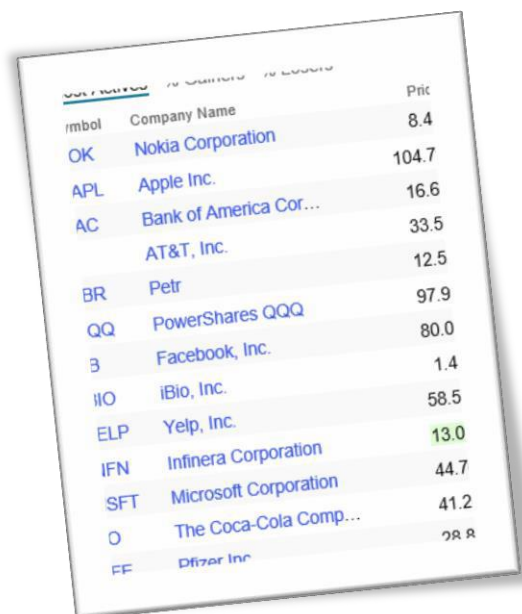
平行坐标图聚簇可视化

# 第2章 数据获取与预处理

- 数据获取
- 数据评估与统计
- 数据清洗
- 数据整理
- 数据可视化
- 数据获取与预处理实践

# 数据获取与预处理实践

## 用Python获取数据



Symbol	Company Name	Price
OK	Nokia Corporation	8.4
APL	Apple Inc.	104.7
AC	Bank of America Cor...	16.6
	AT&T, Inc.	33.5
BR	Petr	12.5
QQ	PowerShares QQQ	97.9
B	Facebook, Inc.	80.0
IBIO	iBio, Inc.	1.4
ELP	Yelp, Inc.	58.5
IFN	Infinera Corporation	13.0
SFT	Microsoft Corporation	44.7
O	The Coca-Cola Comp...	41.2
PF	Pfizer Inc	28.8

### 网络数据如何获取（爬取）？

抓取网页，解析网页内容

- 抓取
  - **urllib**内建模块
    - **urllib.request**
  - **Requests**第三方库
  - **Scrapy**框架
- 解析
  - **BeautifulSoup**库
  - **re**模块

第三方  
API抓取  
+解析

## Requests库

- **Requests**库是更简单、方便和人性化的**Python HTTP**第三方库
- **Requests**官网: <http://www.python-requests.org/>
- 基本方法 `requests.get()` 请求获取指定URL位置的资源，对应HTTP协议的GET方法

遵循网站爬虫协议 robots.txt

```
>>> import requests
>>> r = requests.get('https://book.douban.com/subject/3259440/comments/')
>>> r.status_code
200
>>> print(r.text)
```



# 数据获取与预处理实践

## 网页数据解析

- **BeautifulSoup**是一个可以从**HTML** 或**XML**文件中提取数据的**Python**库
- 官方网站:  
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>



```
soup.find_all('span', 'short')
```

`<span class="short">`昨晚两点开始看《白夜行》一直看到天亮看完。展现出东野圭吾对复杂叙事的掌控能力，精彩绝伦。但是我最欣赏的还是他对恶的动机，那种孜孜不倦的探求，一直向灵魂黑洞最深处走去。他书写的恶往往不是凡俗的恶，而是一种提纯的，高智商的，有分寸的，肃穆的恶。那种恶最终会让人动容，和纯粹的善一样。`</span>`

- **re**正则表达式模块进行各类正则表达式处理

- 参考网站:

<https://docs.python.org/3.5/library/re.html>

```
'<span class="user-stars allstar(.*) rating'
```

```
<span class="user-stars  
allstar50 rating" title="力荐  
></span>
```

`/code/2/book_Crawler.py`

# 数据获取与预处理实践

## 动态网页数据获取

<https://news.sina.com.cn/roll/#pageid=153&lid=2509&k=&num=50&page=1>



新闻中心

新浪新闻 | 新浪首页 | 新浪导航

2019年1月1日 全部滚动新闻

按标题 ▾

全部时间 ▾

搜索


28 秒后刷新

刷新

栏目	标题	时间
[全部]	日美英首次联合军演：日本出云号参演 科目针对性强	01-01 10:50
[全部]	2018基金公司规模:博时广发名次上升 嘉实中银名次降	01-01 10:48
[全部]	新华国际时评：中美建交四十年 一路风雨总向前	01-01 10:44
[全部]	1962年印军被俘后待遇如何？有饭吃有烟抽想呆在中国	01-01 10:39
[全部]	用户携程上买机票后遭诈骗12万 公司被判赔偿5万元	01-01 10:38
[全部]	澎湃新闻新年献词：春天是从冬天开始的	01-01 10:38
[全部]	天弘余额宝四季度缩水1905亿 总规模小幅减少73亿	01-01 10:36
[全部]	为何用枪不用炮轰？中国引进俄军造价4万元狙击步枪	01-01 10:35
[全部]	韩2018年出口额首破6千亿美元创新高	01-01 10:33
[全部]	近10万民众2019年元旦清晨天安门广场看升旗	01-01 10:31

# 数据获取与预处理实践

## 动态网页数据获取

The screenshot displays a web browser window with a news list on the left and a network request inspector on the right. The news list is titled "2019年1月1日 全部滚动新闻" and contains a table of news items. The network inspector shows a request to the Sina News API.

**News List Table:**

栏目	标题	时间
[全部]	日本这次给出神操作：想用自己二手F15换购美全新F35	01-01 10:56
[全部]	2018基金公司规模：平安永赢鹏扬广发规模暴增排名升	01-01 10:55
[全部]	中国固定翼预警机可从滑跃甲板起飞 但付出代价太大	01-01 10:55
[全部]	2019可能是特朗普和梅姨最难熬的一年	01-01 10:55
[全部]	日美英首次联合军演：日本出云号参演 科目针对性强	01-01 10:50
[全部]	2018基金公司规模：博时广发名次上升 嘉实中银名次降	01-01 10:48
[全部]	宿管阿姨退休全院师生送别 500字告别信感动学子	01-01 10:45
[全部]	新华国际时评：中美建交四十年 一路风雨总向前	01-01 10:44
[全部]	默克尔新年献词呼吁共克时艰 强调不再连任	01-01 10:43
[全部]	海南自由贸易账户(FT账户)体系正式上线运行	01-01 10:42
[全部]	中俄朝三国边城民众共迎新年第一缕阳光(图)	01-01 10:42
[全部]	1962年印军被俘后待遇如何？有饭吃有烟抽想呆在中国	01-01 10:39
[全部]	用户携程上买机票后遭诈骗12万 公司被判赔偿5万元	01-01 10:38
[全部]	澎湃新闻新年献词：春天是从冬天开始的	01-01 10:38
[全部]	天弘余额宝四季度缩水1905亿 总规模小幅减少73亿	01-01 10:36
[全部]	为何用枪不用炮轰？中国引进俄军造价4万元狙击步枪	01-01 10:35
[全部]	韩2018年出口额首破6千亿美元创新高	01-01 10:33
[全部]	近10万民众2019年元旦清晨天安门广场看升旗	01-01 10:31
[全部]	北大江俊毅部长离任时间首月 呈打铁还需自身硬	01-01 10:30

**Network Request Details:**

- Name:** get?pageid=153&lid=2509&k...
- General:**
  - Request URL:** https://feed.mix.sina.com.cn/api/roll/get?pageid=153&lid=2509&k=&num=50&page=1&r=0.9809380533643719&callback=jQuery311008812805996428308\_1546310724461&\_=1546310724469
  - Request Method:** GET
  - Status Code:** 200 OK
  - Remote Address:** 113.108.216.230:443
  - Referrer Policy:** unsafe-url
- Response Headers:**
  - Connection:** keep-alive
  - Content-Encoding:** gzip
  - Content-Type:** application/javascript; charset=UTF-8
  - Date:** Tue, 01 Jan 2019 02:53:31 GMT
  - DG-SN-REQID:** 5c2b977f4ebba0ec2925a0112173ae8f
  - DPOOL:** newscenter
  - DPOOL\_HEADER:** web5\_jugg149
  - DPOOL\_LB7\_HEADER:** jugg138

**Console:**

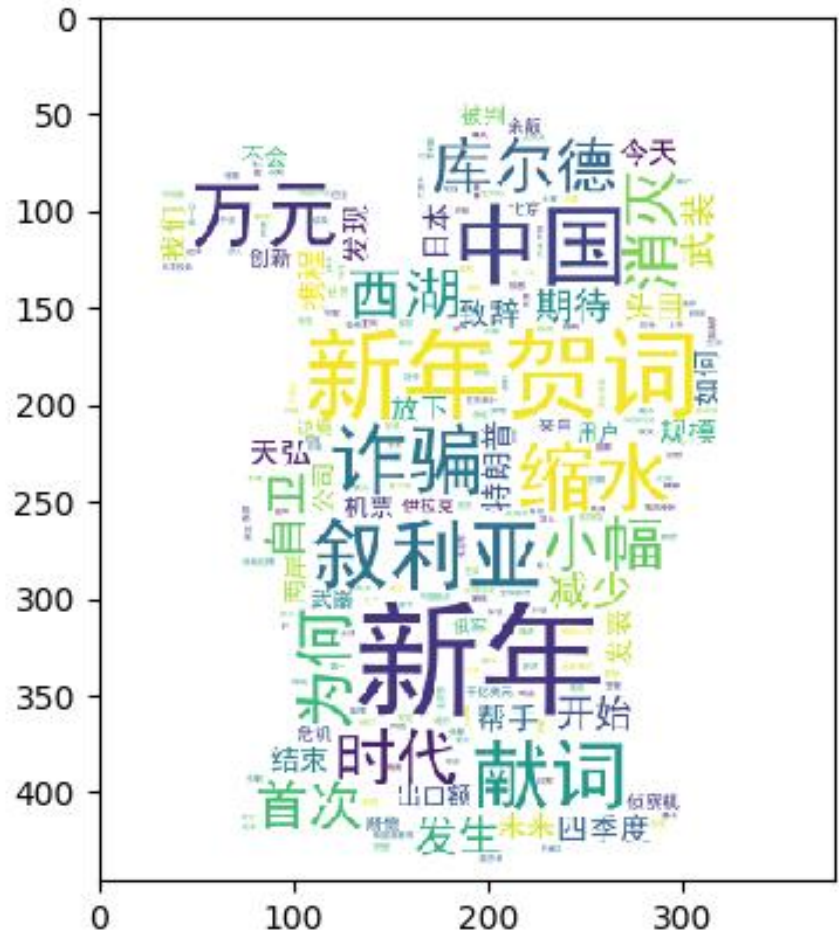
```
pageid=153&lid=2509&k=&num=50&page=1 newlist_channel.js:654
Object newlist_channel.js:790
pageid=153&lid=2509&k=&num=50&page=1 newlist_channel.js:654
Object newlist_channel.js:790
pageid=153&lid=2509&k=&num=50&page=1 newlist_channel.js:654
Object newlist_channel.js:790
pageid=153&lid=2509&k=&num=50&page=1 newlist_channel.js:654
Object newlist_channel.js:790
```

/code/2/news\_title\_mining.py

# 数据获取与预处理实践

## 词云分析

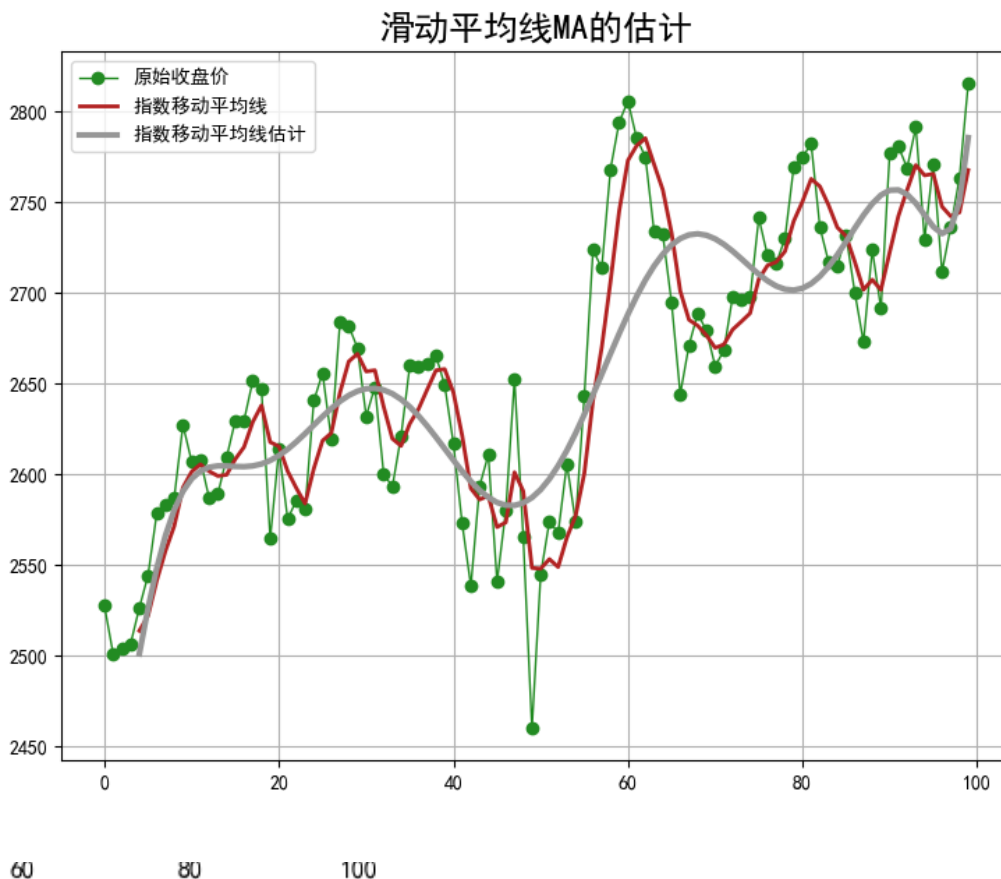
- **标题分词 ( Text Segmentation )**  
要抓热点词首先要将新闻标题进行jieba (分词)，逐行用 jieba 分词  
`cutdata=jieba.cut(data)`
- **根据词频画出词云**  
`mywc=wc.WordCloud(collocations  
font_path=font,mask=catarray,  
background_color="white").genera`



/code/2/news\_title\_mining.py

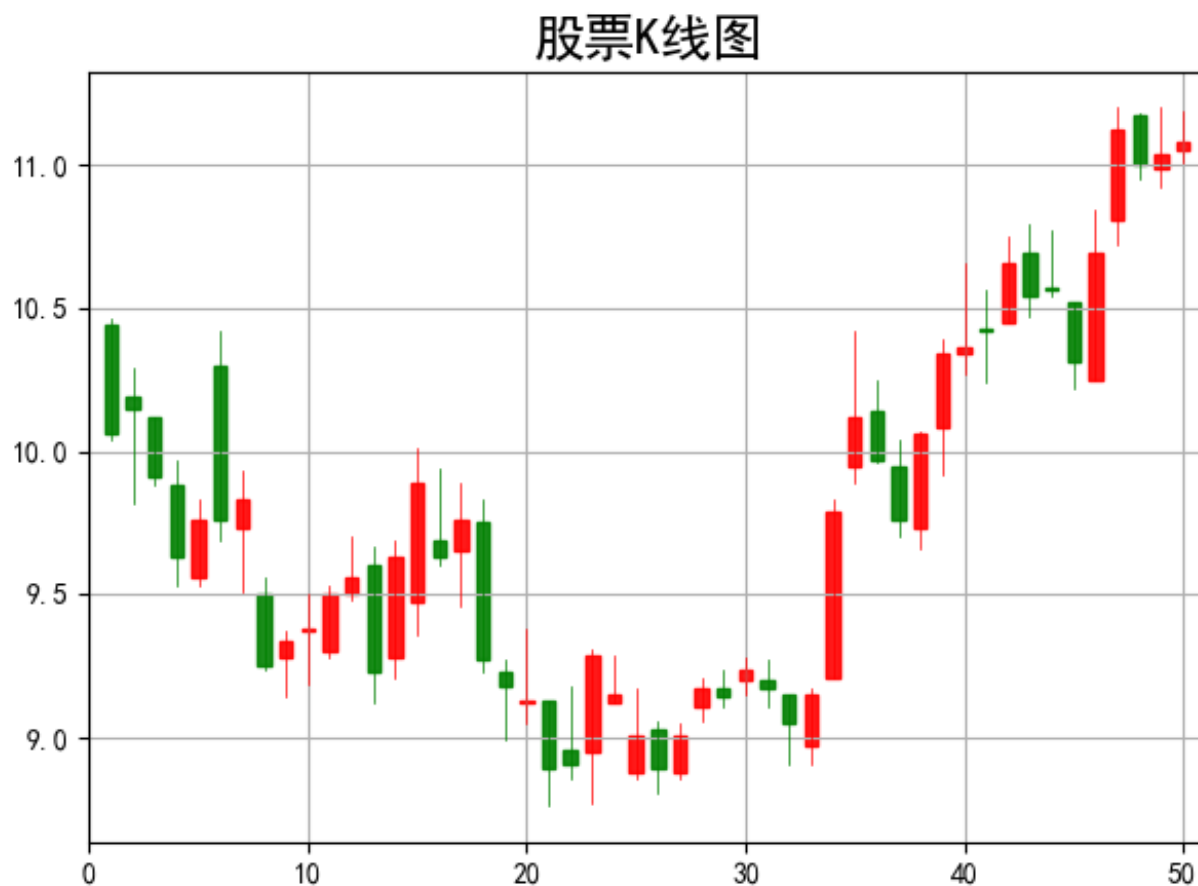
# 数据获取与预处理实践

## 数据平滑



# 数据获取与预处理实践

## 数据可视化

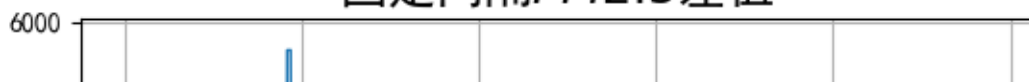


/code/2/stock\_K.py

# 数据获取与预处理实践

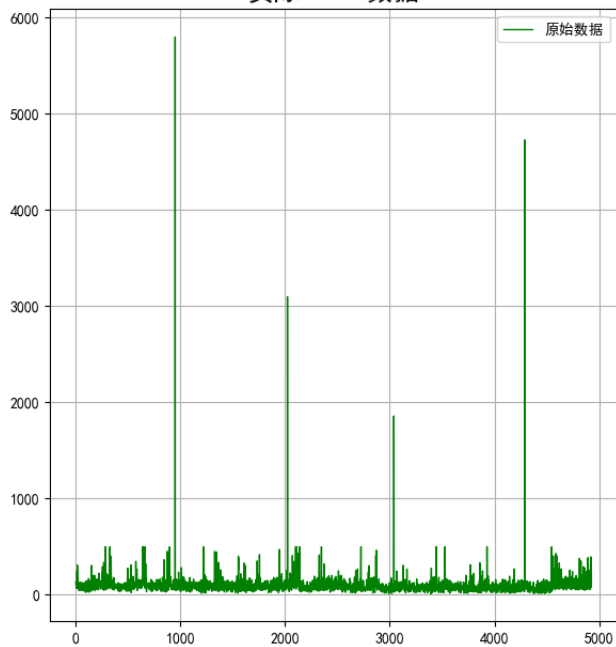
## 数据异常值检测

固定间隔PM2.5差值

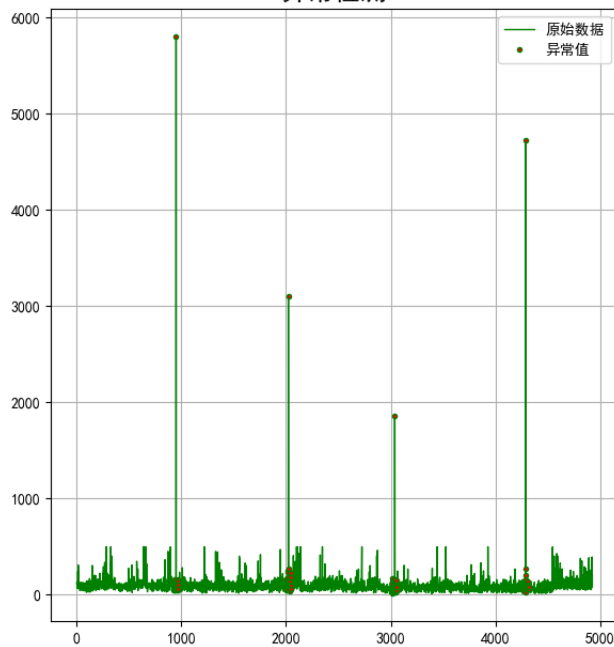


PM2.5的异常值检测与校正

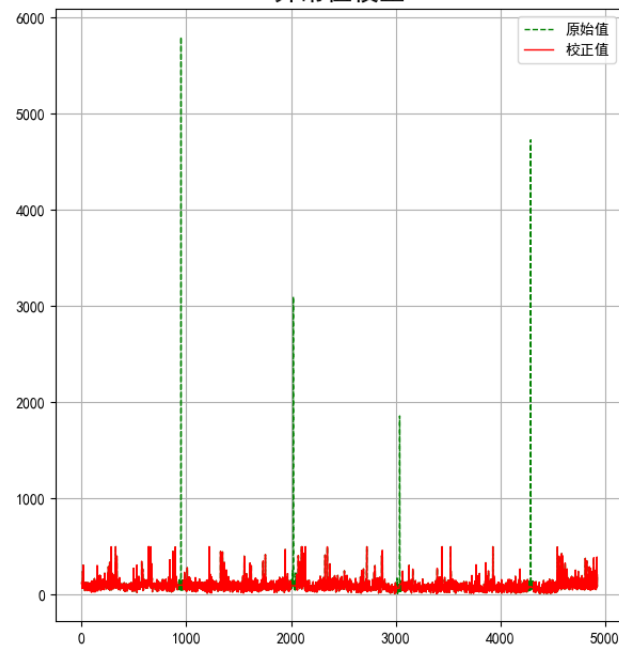
实际PM2.5数据



异常检测



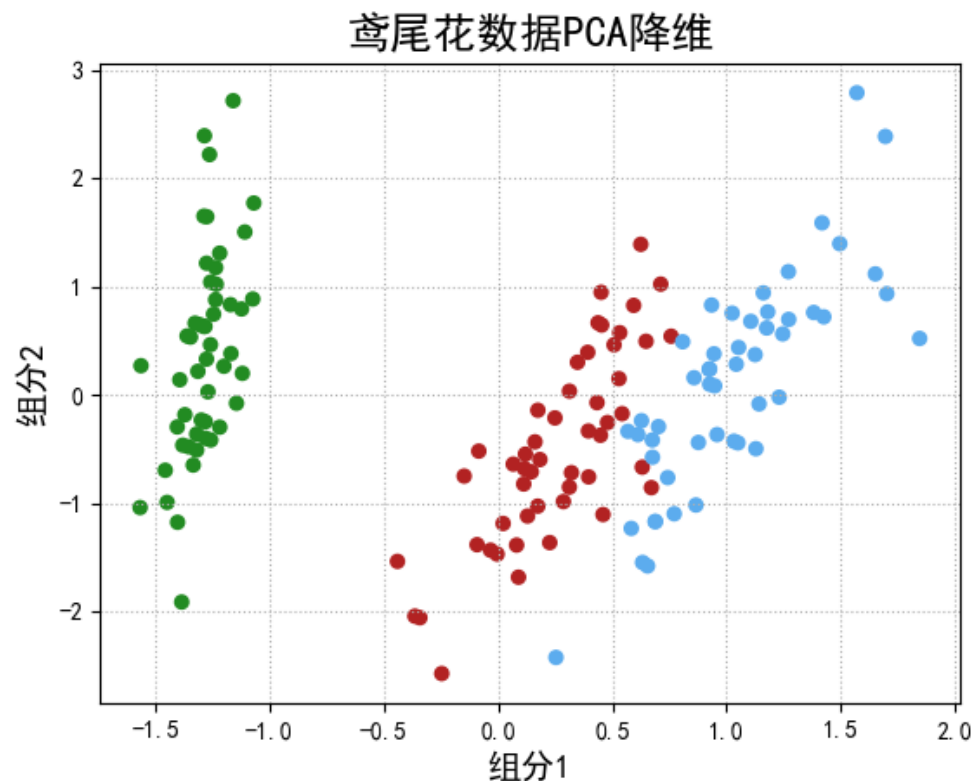
异常值校正



# 数据获取与预处理实践

## PCA降维

5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	4	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setosa
5.1	3.8	1.5	0.3	Iris-setosa
5.4	3.4	1.7	0.2	Iris-setosa
5.1	3.7	1.5	0.4	Iris-setosa
4.6	3.6	1	0.2	Iris-setosa





## 课后作业2

1. 特征选择与特征提取有哪些相同与不同之处？
2. 利用分箱法，按平均值平滑、按边界值平滑和按中值平滑下列数据：
  - 箱1：1000,1400,1800
  - 箱2：500,700,800,800,1100 ,1300
  - 箱3：300,500,800,1000
  - 箱4：1700,3300
  - 箱5：4500,4800, 5000 , 5300
3. Python抓取数据的库有哪些？

# Acknowledgement

《大数据挖掘》 吉林大学 韩霄松

《机器学习》 吉林大学 吴春国、李瑛

《机器学习》 小象学院 邹博

---

Thanks