

# 习题

设  $p(\mathbf{x}|\Sigma) \sim N(\mu, \Sigma)$  这里  $\mu$  已知， $\Sigma$  未知。如果  $\Sigma$  的最大似然估计为

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^n (x_k - \mu)(x_k - \mu)^t, \text{ 证明以下论述:}$$

- 1) 证明矩阵等式  $a^t A a = \text{tr}[A a a^t]$ ，这里矩阵的迹  $\text{tr}[A]$  表示  $n \times n$  维矩阵  $A$  的对角线元素之和， $a$  为一个向量。

- 2) 证明似然函数可以写为以下形式:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \Sigma) = \frac{1}{(2\pi)^{nd/2}} |\Sigma^{-1}|^{n/2} \exp \left[ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} \sum_{k=1}^n (\mathbf{x}_k - \mu)(\mathbf{x}_k - \mu)^t \right] \right]$$

注: 多维正态分布  $p(\mathbf{x}|\Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} \{ (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \} \right]$

3) 设  $A = \Sigma^{-1} \hat{\Sigma}$  ,  $\lambda_1, \dots, \lambda_d$  为  $A$  的特征值, 证明前面第(2)小题中的概率式

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \Sigma) = \frac{1}{(2\pi)^{nd/2}} |\Sigma^{-1}|^{n/2} \exp \left[ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} \sum_{k=1}^n (\mathbf{x}_k - \mu)(\mathbf{x}_k - \mu)^t \right] \right]$$

可写为

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \Sigma) = \frac{1}{(2\pi)^{nd/2} |\hat{\Sigma}|^{n/2}} (\lambda_1 \dots \lambda_d)^{n/2} \exp \left[ -\frac{n}{2} (\lambda_1 + \dots + \lambda_d) \right]$$

提示:

$$|A| = |\Sigma^{-1}| |\hat{\Sigma}|$$

$$\text{tr}(A) = \sum (A_{ii}) = \sum (\lambda_i)$$

$$|A| = \prod \lambda_i$$

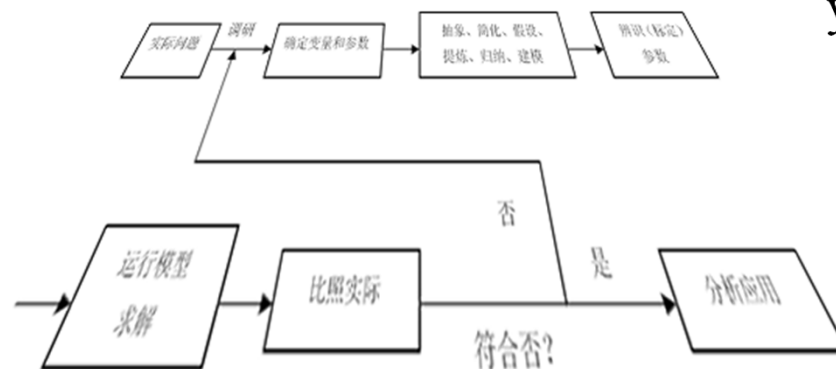
# 数据建模与分析

## 数据概述



姚远

yaoyuan@shu.edu.cn



上海大学机自学院

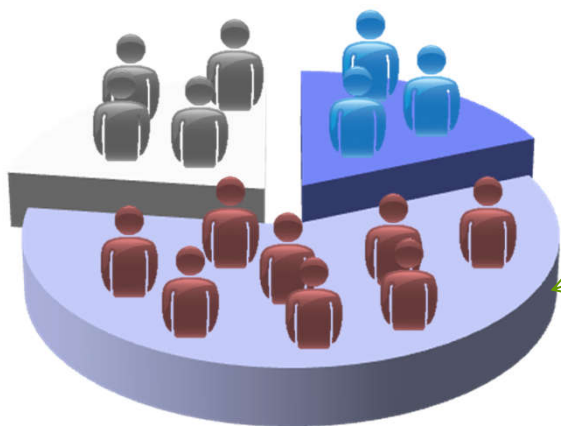
2020/10/11

# 大纲

- 数据的表达
- 数据特征
- 数据预处理
- 数据存储
- 数据分析

# 数据的表达

## □数据-真实世界的表达



# 数据的表达

## □表达形式

- 有组织的数据



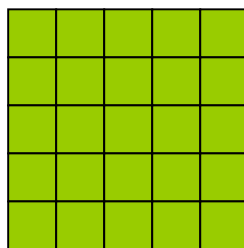
标量

0阶张量



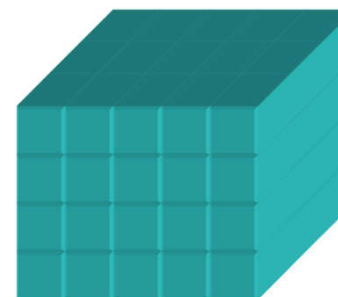
向量

1阶张量



矩阵

2阶张量



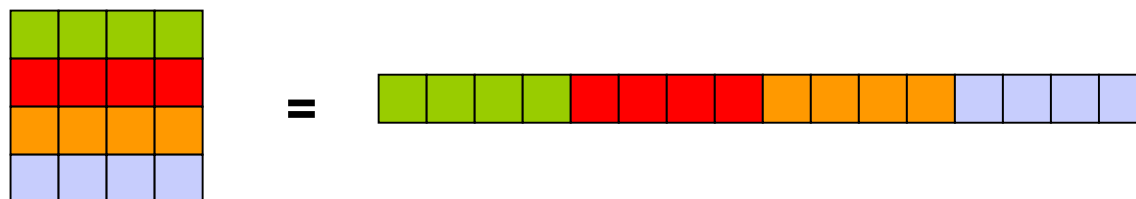
张量 (tensor)

3阶张量

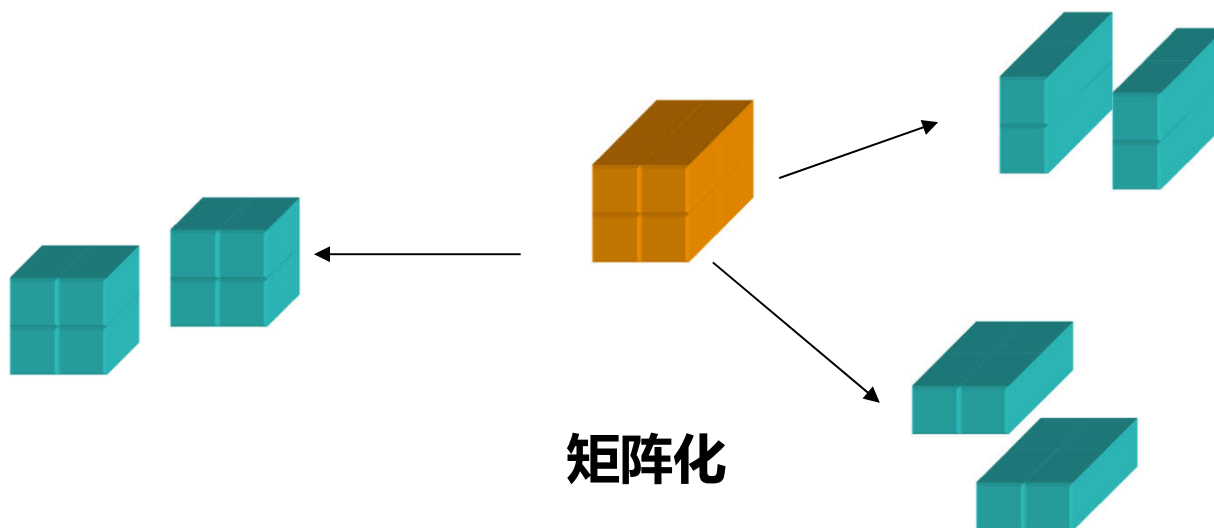
# 数据的表达

## □表达形式

- 有组织的数据



向量化



矩阵化







# 大纲

- 数据的表达
- 数据特征
- 数据预处理
- 数据存储
- 数据分析

# 数据的常识

---



我们先要了解一些有关数据的常识













# 数据属性

- 数据对象的特征（Characteristics）  
或特性（feature）
- 别名：
  - 特征
  - 维度
  - 变量
- 属性集合：属性向量

# 数据属性

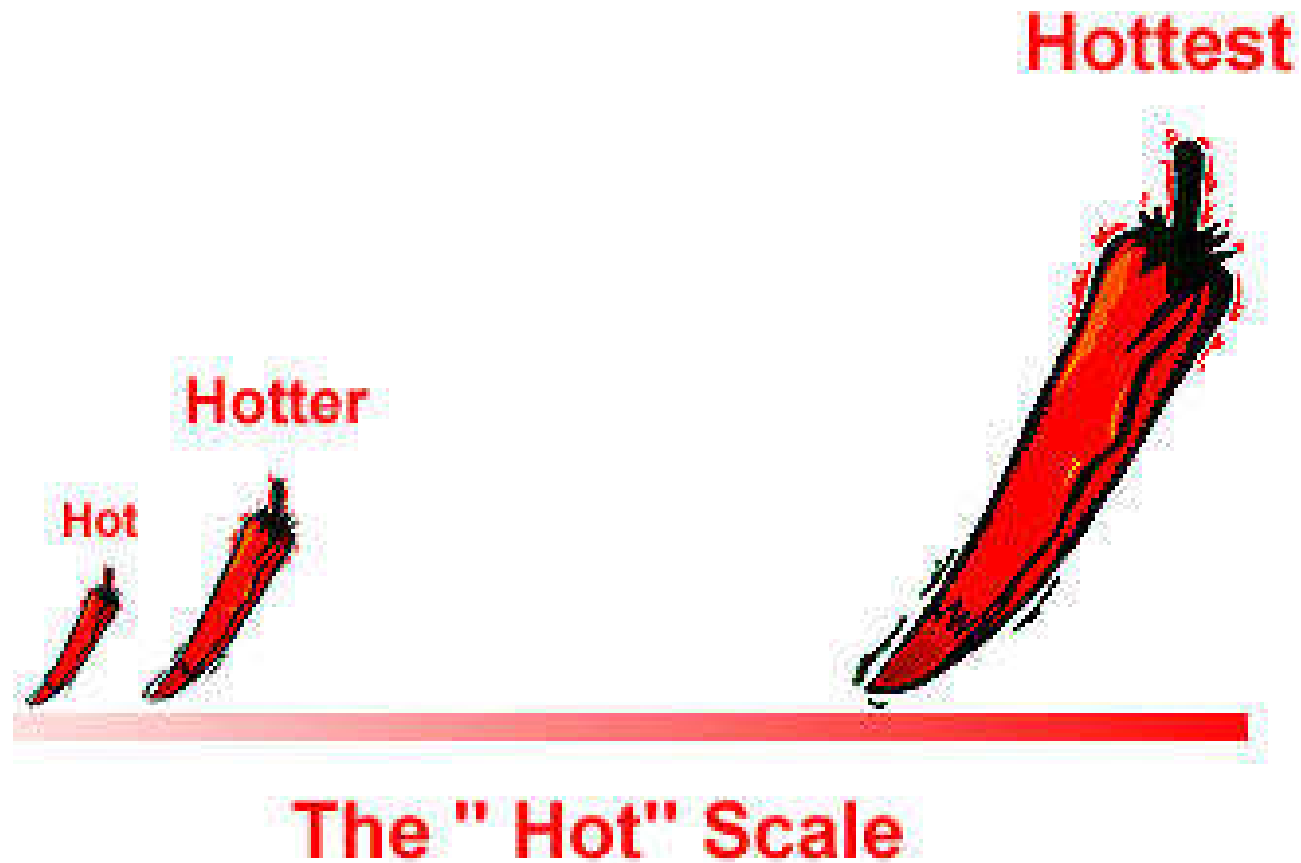
- 类别型属性

## Nominal Data

<b>Point</b>	airport 	town 	mine 	capital 
<b>Line</b>	river 	road 	boundary 	pipeline 
<b>Area</b>	orchard 	desert 	forest 	water 

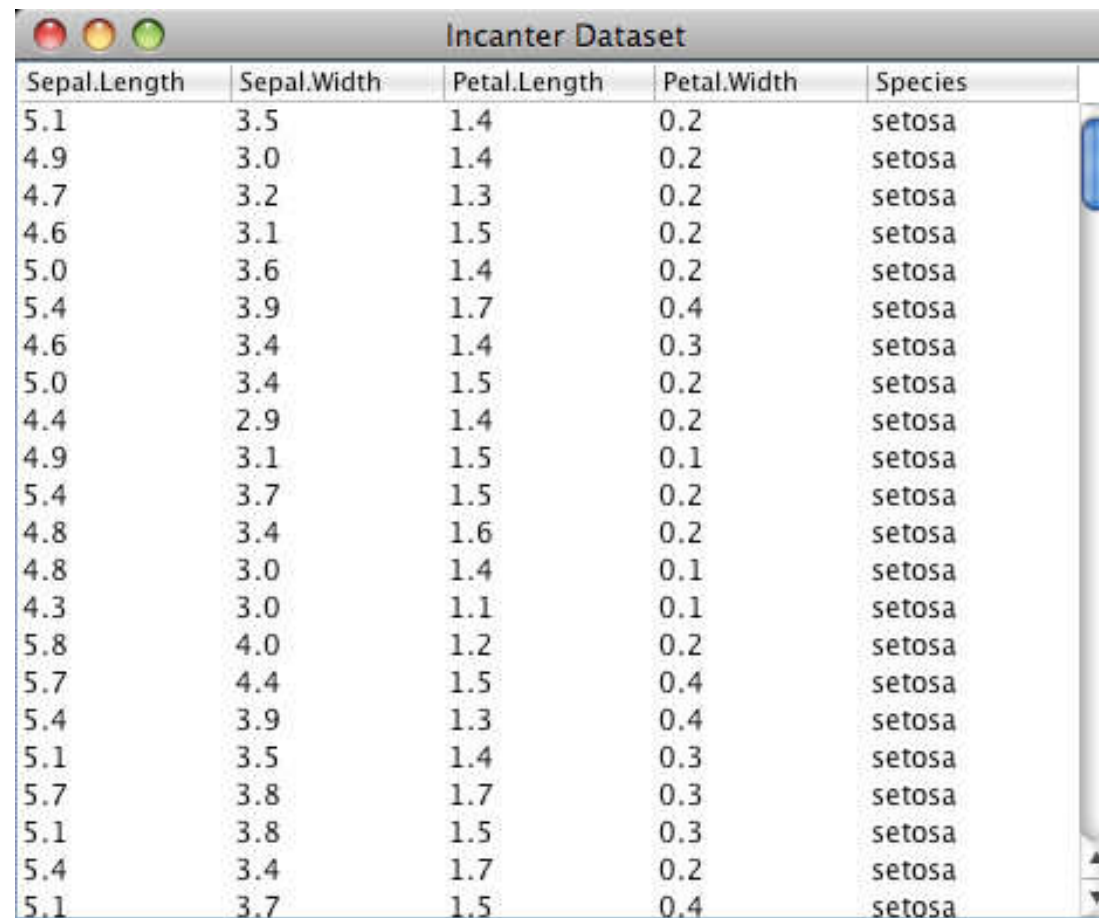
# ■ 数据属性

- 有序型属性



# 属性类型

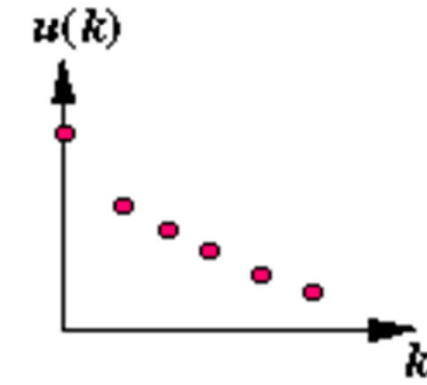
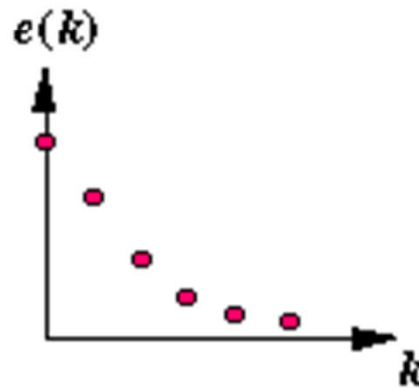
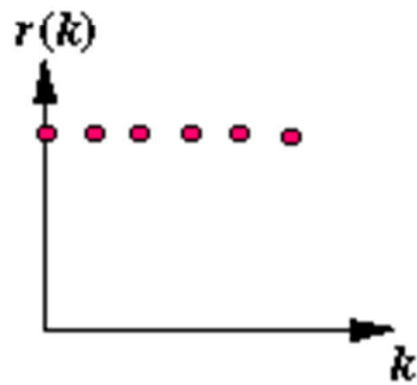
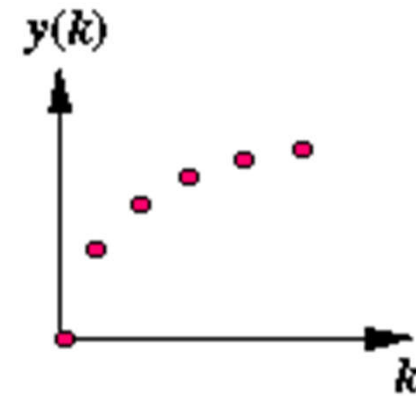
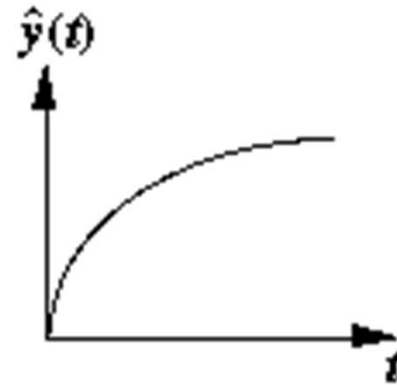
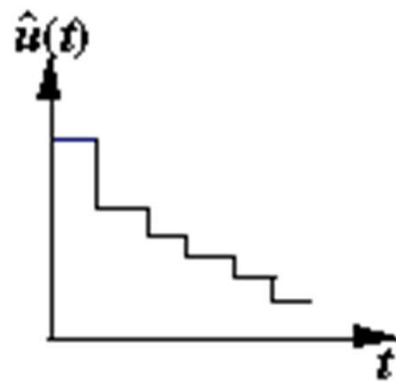
- 数值型属性



Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa

# 数据属性

- 离散型和连续型





# 数据属性分析

---

- 通过相似度能够对数据本身进行评价
- 相似度的定义与适用领域有关
- 相似度取值一般为[0-1]

# 类别型数据的相似度

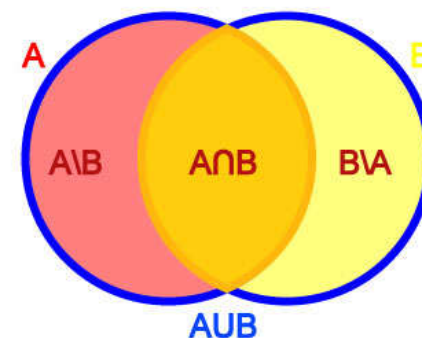
- 不匹配率：

$$d(i, j) = \frac{p - m}{p}$$

$p$ 为两对象间对等属性的个数， $m$ 为两对象对应属性相等的个数。

- （二值类别型数据）Jaccard系数：

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



$q$ 为两字符串对应位置为同一值的个数； $r$ 为两串对应位置上，第一个字符串为1，第二个字符串为0的个数； $s$ 与 $r$ 相反。

# 类别型数据的相似度

- (长度相等的字符串类别型数据) Hamming 距离

"RapidManufacture" and "RapooManufacture" = 2

2173896 and 2233796 = 3

- (长度不等的字符串类别型数据) 距离?

Smith–Waterman algorithm

# 相似度和相异度矩阵

---

- 数据间关系的度量
- 经常在统计和数据挖掘中使用

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

▪ 相异度矩阵

## 数值型属性间的距离：明科夫斯基距离系

- 欧氏距离 (L2)

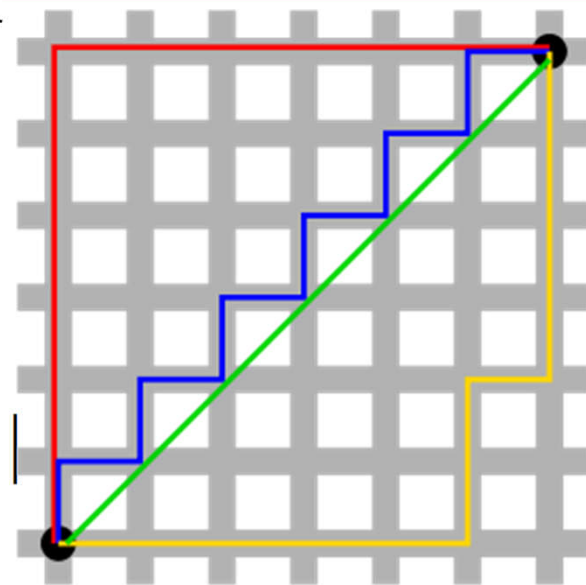
$$d_{Euc} = \sqrt{\sum_{i=1}^d |P_i - Q_i|^2}$$

- 曼哈顿距离 (L1)

$$d_{CB} = \sum_{i=1}^d |P_i - Q_i|$$

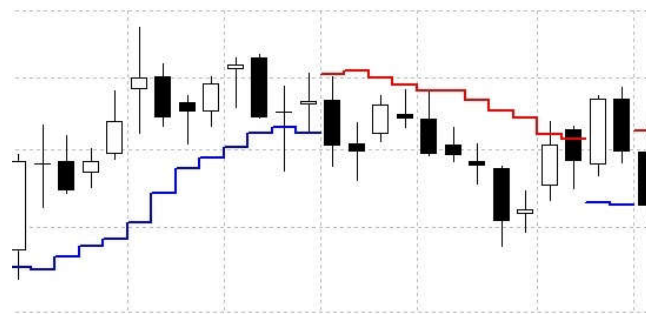
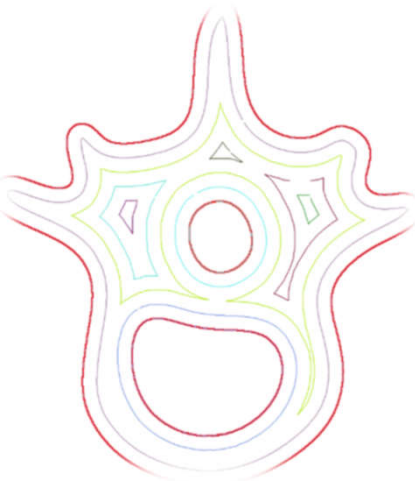
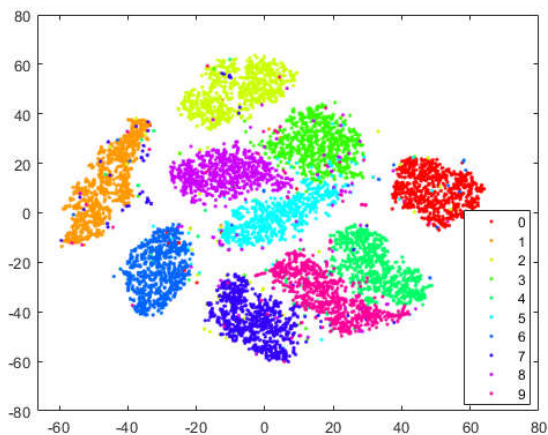
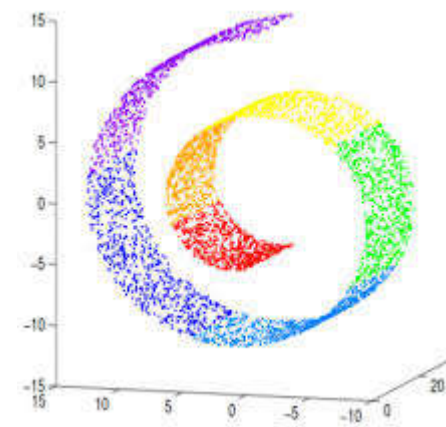
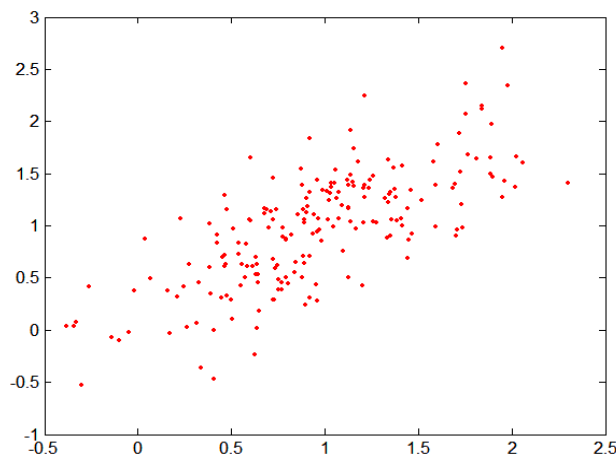
- 明科夫斯基距离 (LP)

$$d_{Mk} = \sqrt[p]{\sum_{i=1}^d |P_i - Q_i|^p}$$



# 一些特殊的“距离”评价

- 一些差异是总体上的，需要一些统计描述



# 基本统计描述

---

- 均值

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

- 中位数

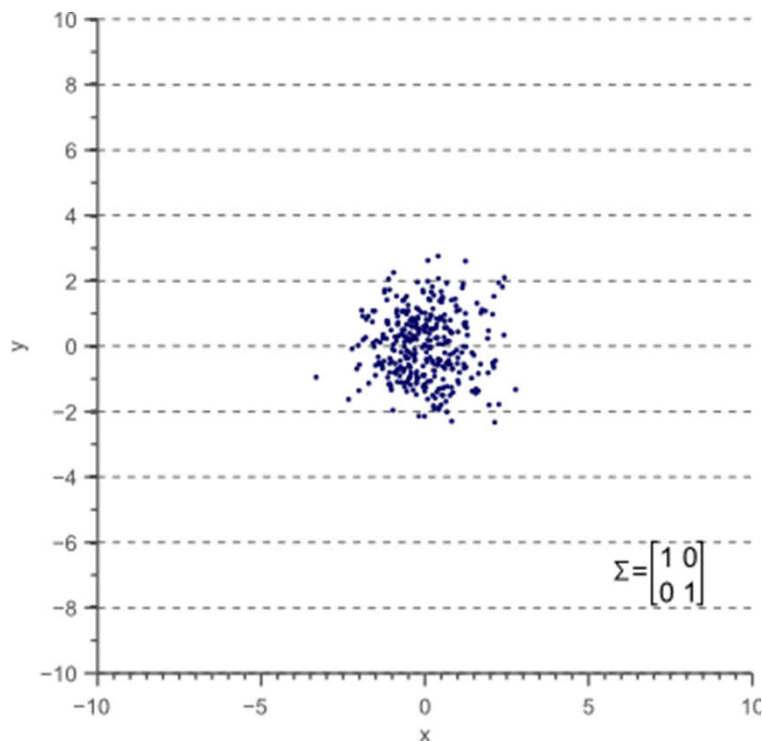
$$Q_{\frac{1}{2}}(x) = \begin{cases} x'_{\frac{n+1}{2}}, & \text{if } n \text{ is odd.} \\ \frac{1}{2}(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}), & \text{if } n \text{ is even.} \end{cases}$$

- 方差

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \left( \frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2,$$

# 基本统计描述：协方差

## ■ 协方差 (Covariance)



## 衡量两个变量间的总体误差

$$\Sigma = \begin{bmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{bmatrix}$$

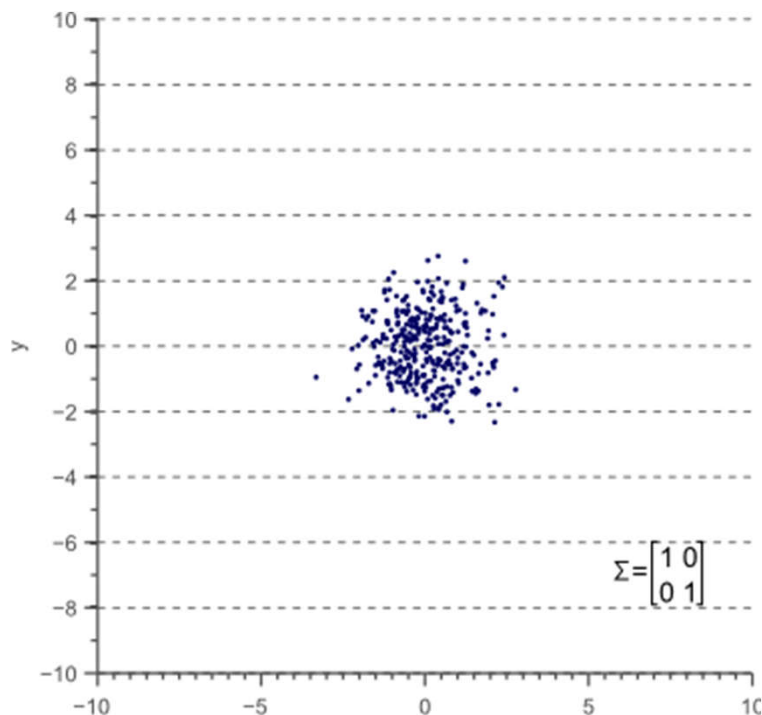
## ■ 协方差的计算

$$\begin{aligned} \Sigma_{ij} &= cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \frac{1}{n-1} \sum_{k=1}^n (X_{ik} - \mu_i)(X_{jk} - \mu_j) \end{aligned}$$



# 基本统计描述：协方差

## ■ 协方差 (Covariance)



$$\begin{aligned}\Sigma_{ij} &= \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \frac{1}{n-1} \sum_{k=1}^n (X_{ik} - \mu_i)(X_{jk} - \mu_j)\end{aligned}$$

## 衡量两个变量间的总体误差

$$\Sigma = \begin{bmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{bmatrix}$$

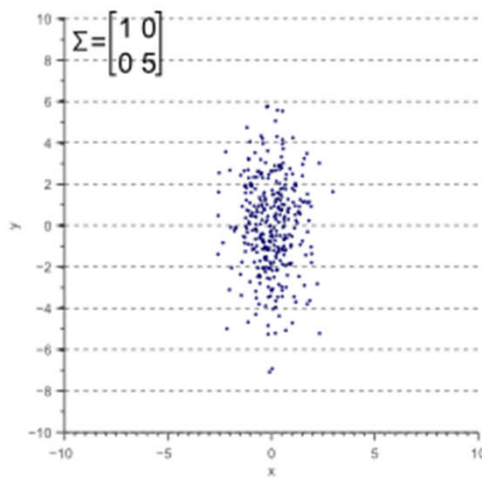
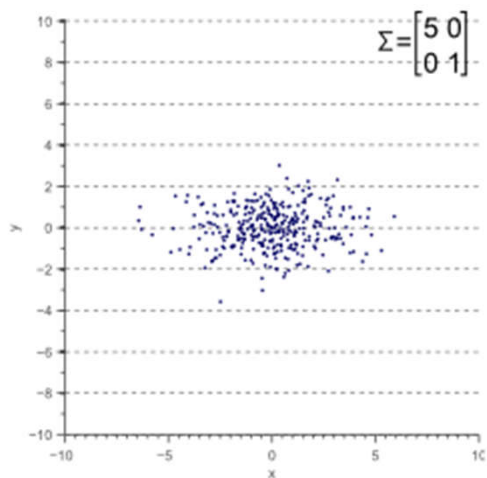
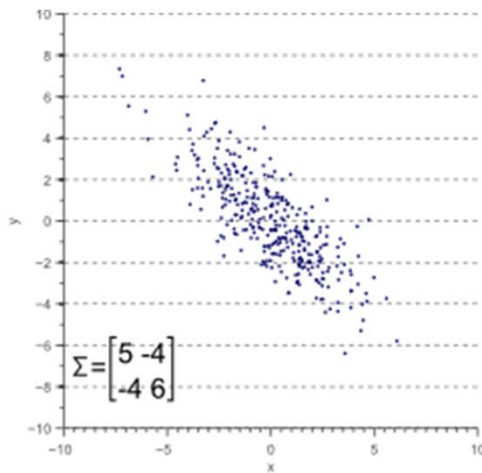
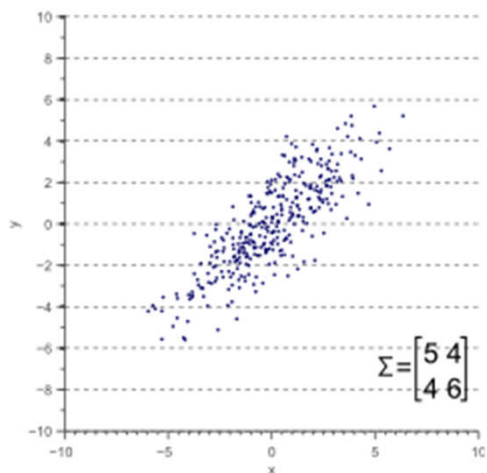
## ■ 对称矩阵

$$\text{cov}(X, X) = \text{var}(X),$$

$$\text{cov}(X, Y) = \text{cov}(Y, X),$$

$$\text{cov}(aX, bY) = ab \text{cov}(X, Y)$$

# 基本统计描述：协方差



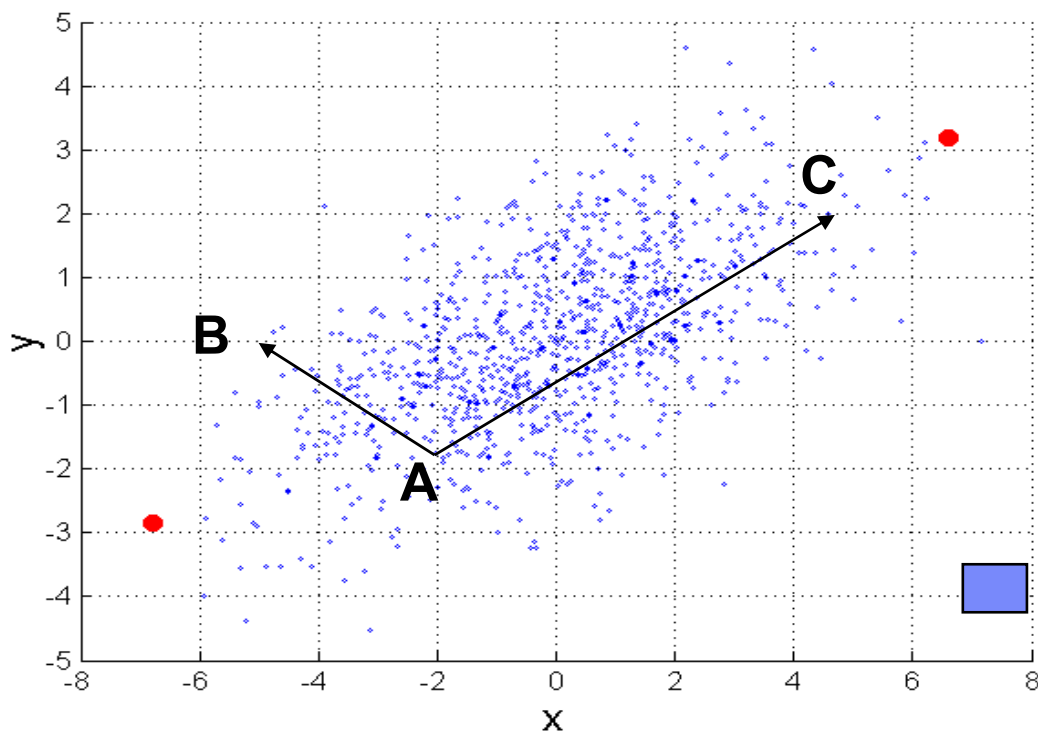
协方差矩阵反应了数据在多维空间中的形态

$$\Sigma = \begin{bmatrix} \sigma(x, x) & \sigma(x, y) \\ \sigma(y, x) & \sigma(y, y) \end{bmatrix}$$

▪哪一维的数据差异大?

# 数值型属性间的距离：马氏距离

- Mahalanobis  $d(X, Y) = \sqrt{(X - Y)^T \Sigma^{-1} (X - Y)}$



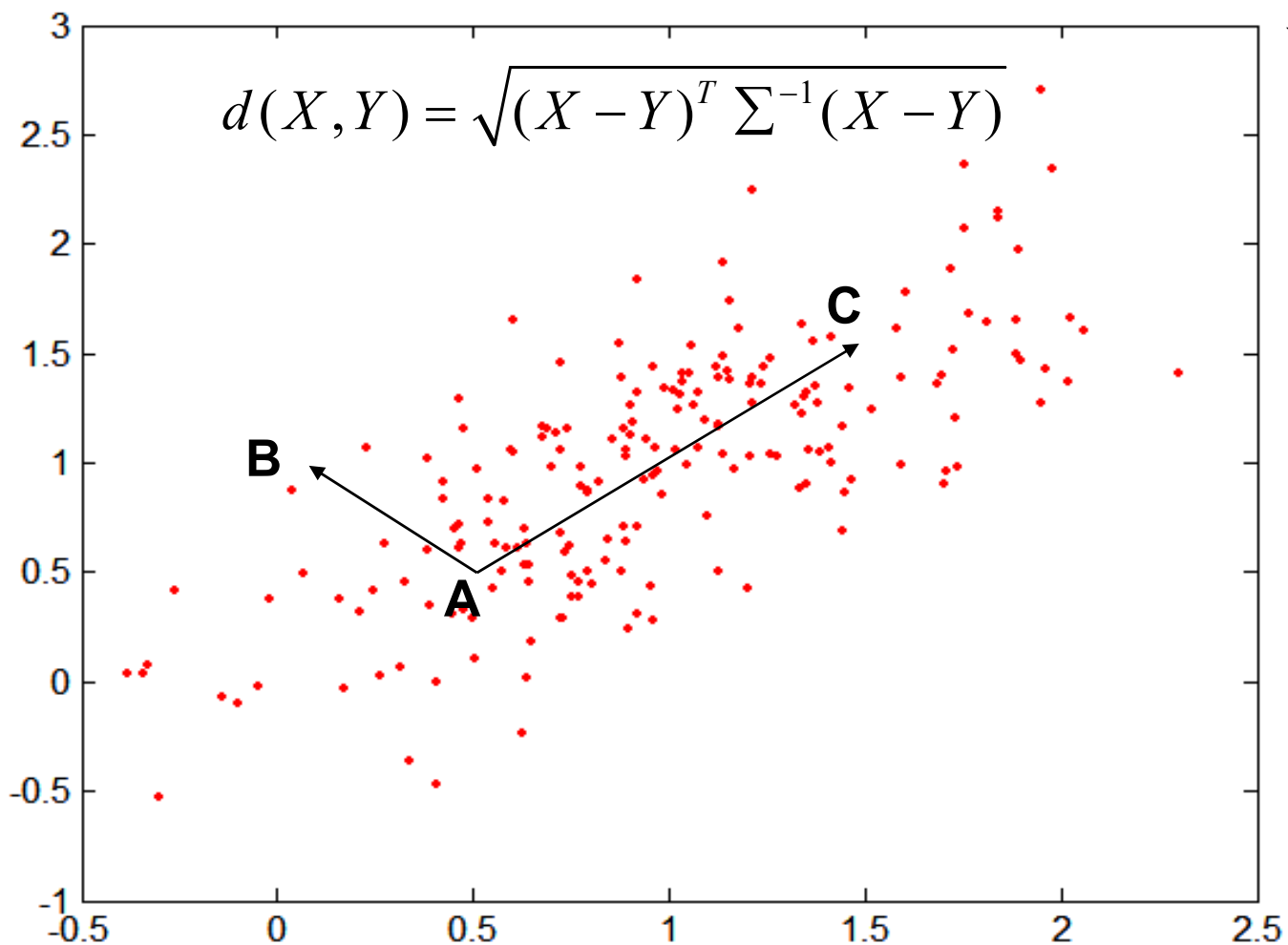
$\Sigma$  输入数据  $X$  的协方差矩阵

$$\begin{aligned}\Sigma_{ij} &= \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \frac{1}{n-1} \sum_{k=1}^n (X_{ik} - \mu_i)(X_{jk} - \mu_j)\end{aligned}$$

当协方差矩阵是单位矩阵时， $d(X, Y)$  等于欧氏距离

- 可用于去除数据中的外点。
- 问题：  $d(A, C) > d(A, B)$ ?

# 数值型属性间的距离：马氏距离



协方差矩阵

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

▪A: (0.5, 0.5)

▪B: (0, 1)

▪C: (1.5, 1.5)

▪Mahal(A,B) = 5

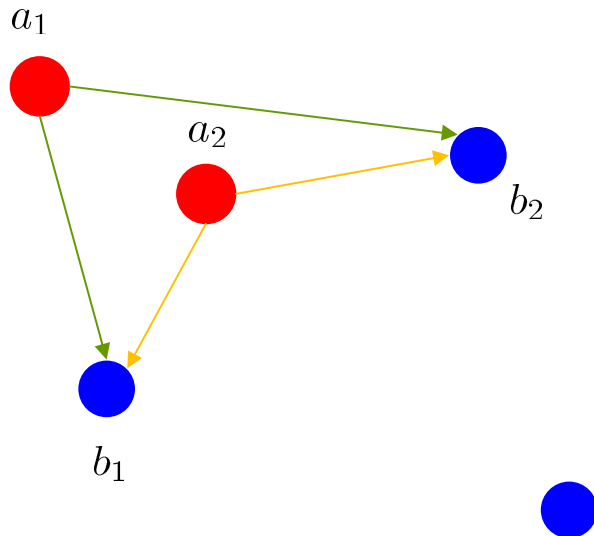
▪Mahal(A,C) = 4

# 数据集间的距离：豪斯多夫距离

- Hausdorff = max(h(A,B), h(B,A))

用于度量两个拓扑结构之间的距离

$$h(A, B) = \max_{a \in A} \{ \min_{b \in B} \{ d(a, b) \} \}$$



▪  $d(a,b)$ : 可以采用任何形式的距离

▪ 计算方法:

-  $h(A,B) = \text{For } j=1:N \text{ Find } \min_{b \in B} (d(a_j, b_i))$

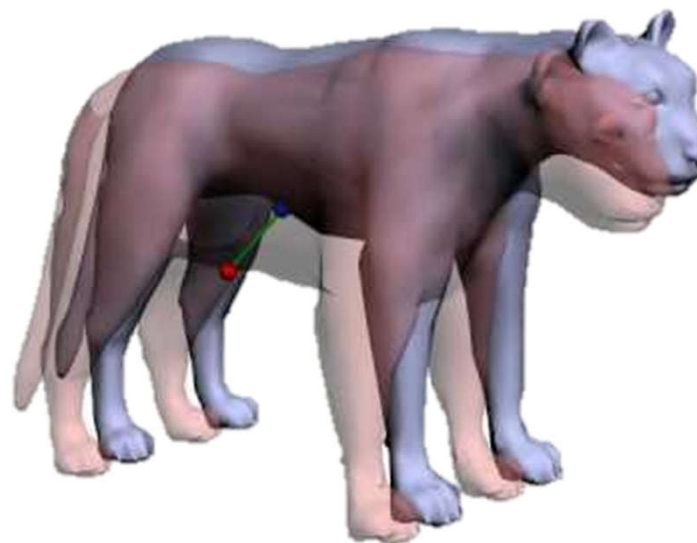
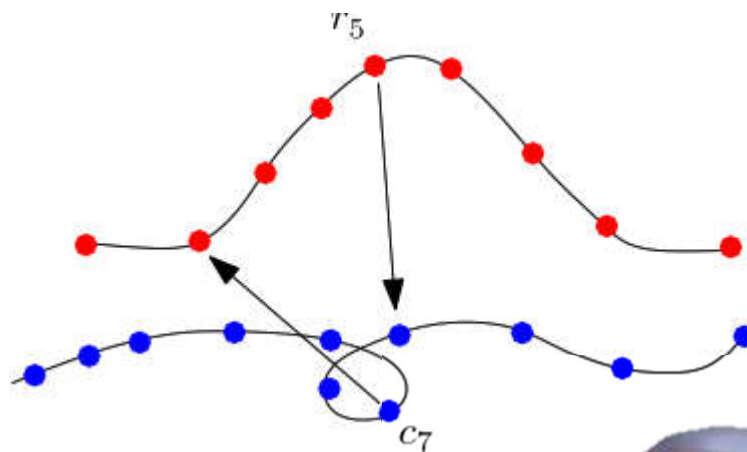
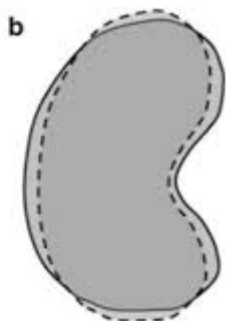
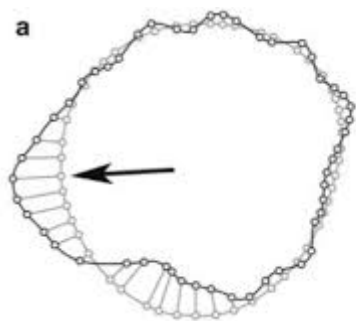
-  $h(B,A) = \text{For } j=1:N \text{ Find } \min_{b \in B} (d(b_1, a_i))$

- 取max(h(A,B), h(B,A))

# 数据集间的距离：豪斯多夫距离

■ Hausdorff =  $\max(h(A,B), h(B,A))$

最大化最小距离



用途：配准、生物相似性

# 统计型数据属性分析

---

- 数据的统计特征一般分为三类

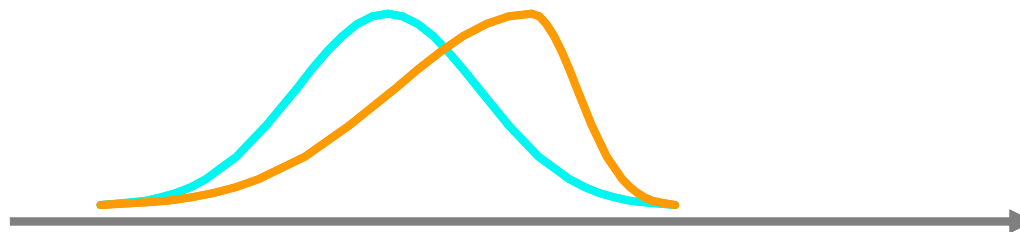
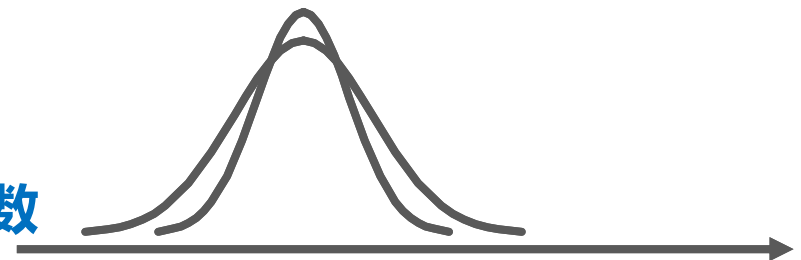
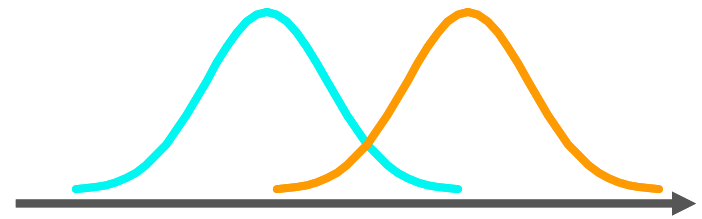
- ~ **集中趋势度量 (位置)**

- 均值、中位数、众数

- ~ **离中趋势度量 (分散程度)**

- 极差、标准差、变异系数、四分位数

- ~ **数据分布形状(偏态和峰态)**



## 小结

---

- 表达不同物理对象要采用不同的数据类型
- 距离是数据属性分析的重要概念
- 统计表达能够提供对数据的整体认识



# 大纲

- 数据的表达
- 数据特征
- 数据预处理
- 数据存储
- 数据分析

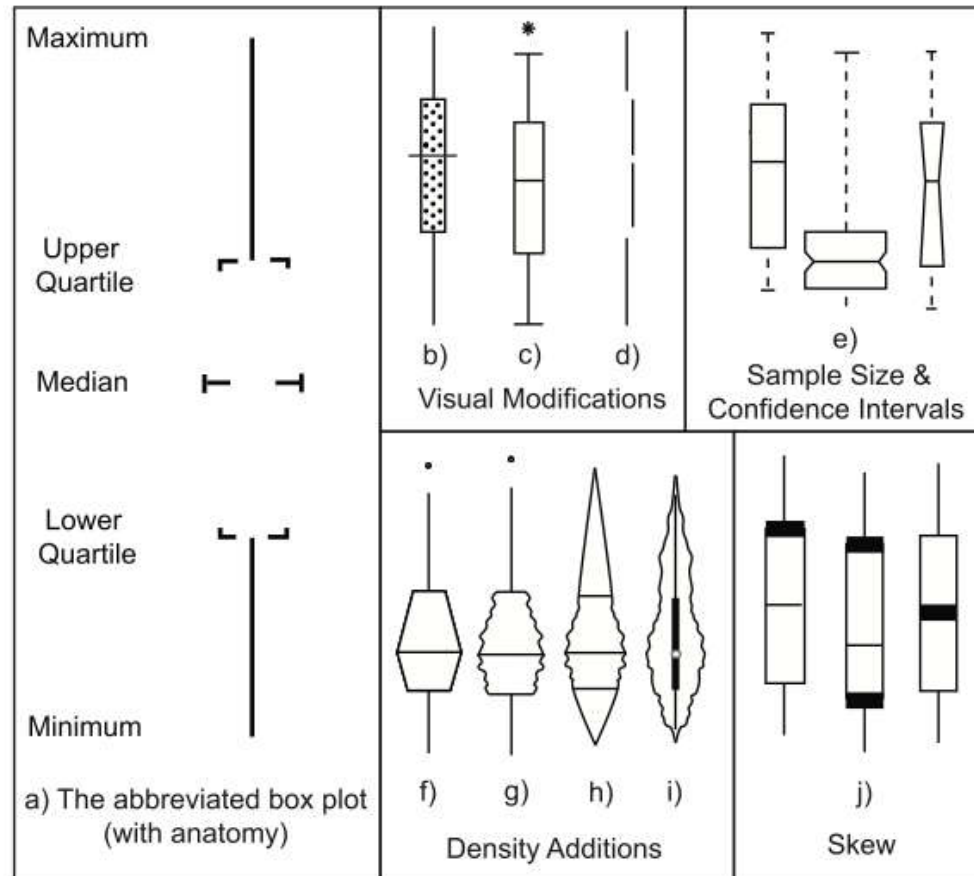
# 数据不确定性

## ■ 分类

- 物理不确定性
- 属性不确定性

## ■ 来源

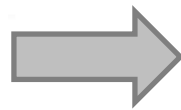
- 本身误差
- 精度转换
- 特定应用需求
- 缺失值
- 数据集成



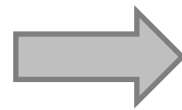
# “Garbage in, garbage out.”

---

原始数据通常含有杂质



▪处理  
▪过程



# 数据质量

---

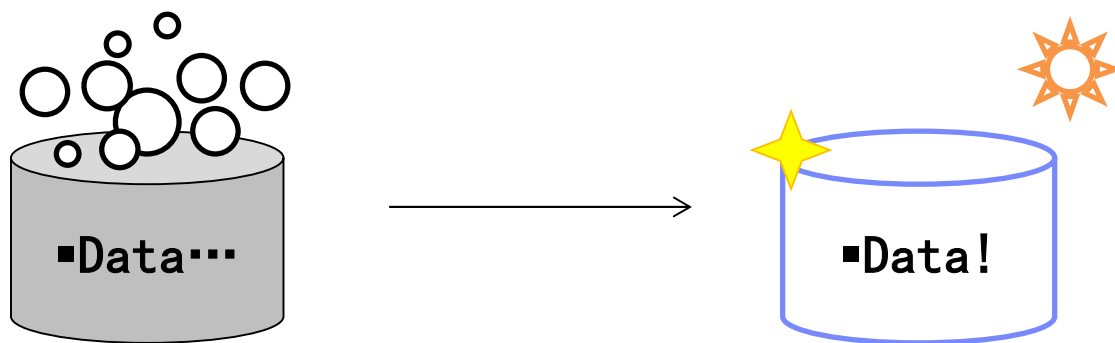
数据质量：数据质量高 -> 对目标用途的符合度高

- 精确性 (Accuracy)
- 完整性 (Completeness)
- 一致性 (Consistency)
- 适时性 (Timeliness)
- 可信性 (Believability)
- 可解释性 (Interpretability)

# 数据清理

---

数据清理：检测和清除数据中的错误和不一致，以提高数据质量



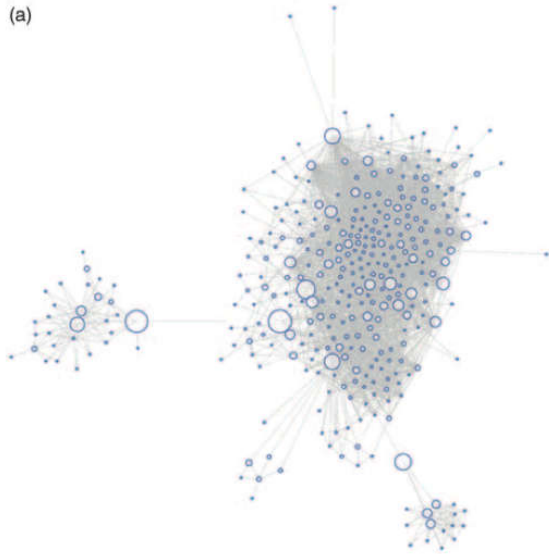
## 措施：

- 使用常量代替缺失值
- 使用属性平均值填充
- 利用回归、分类等方式去噪声
- 人工填充

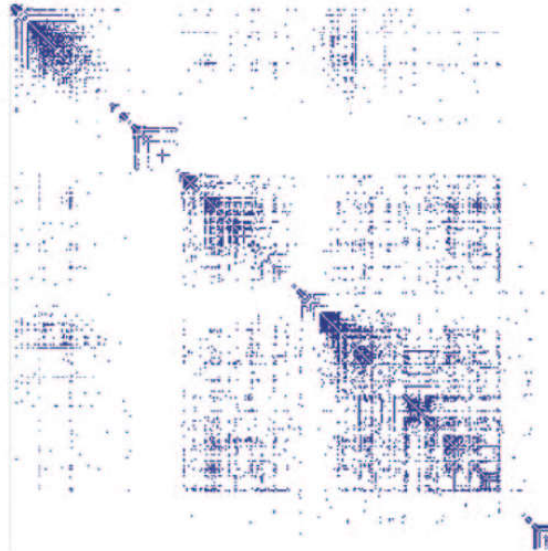
# 可视数据清洗

使用可视化工具进行数据清洗

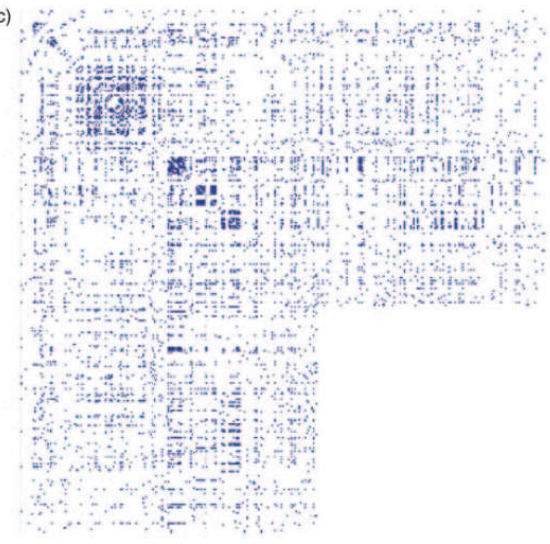
(a)



(b)



(c)



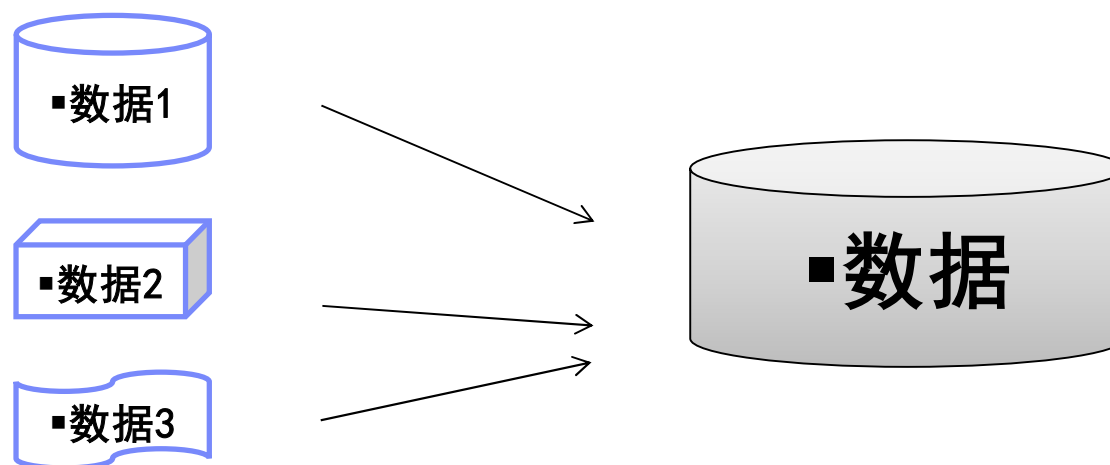
# 数据整合

---

数据整合包括：

- (1) 合并来自多个数据源的数据
- (2) 向用户提供一个关于这些数据的统一视图

管理来自多个数据源的数据



# 多数据源

---

结构冲突 (structural conflicts) :

不同的模式 (schema) 等

数据冲突 (data conflicts) :

重复的记录, 冲突的记录属性等



## 数据整合实例(2)

---

对同一篇论文，来自不同论文数据库的引用格式可能存在不同

整合为某种统一格式

- R. Agrawal, R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB-94, 1994.
- Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994.

# 数据整合实例(1)

## 客户列表1

<i>Cno</i>	<i>LastName</i>	<i>FirstName</i>	<i>Gender</i>	<i>Address</i>	<i>Phone/Fax</i>
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

## 客户列表2

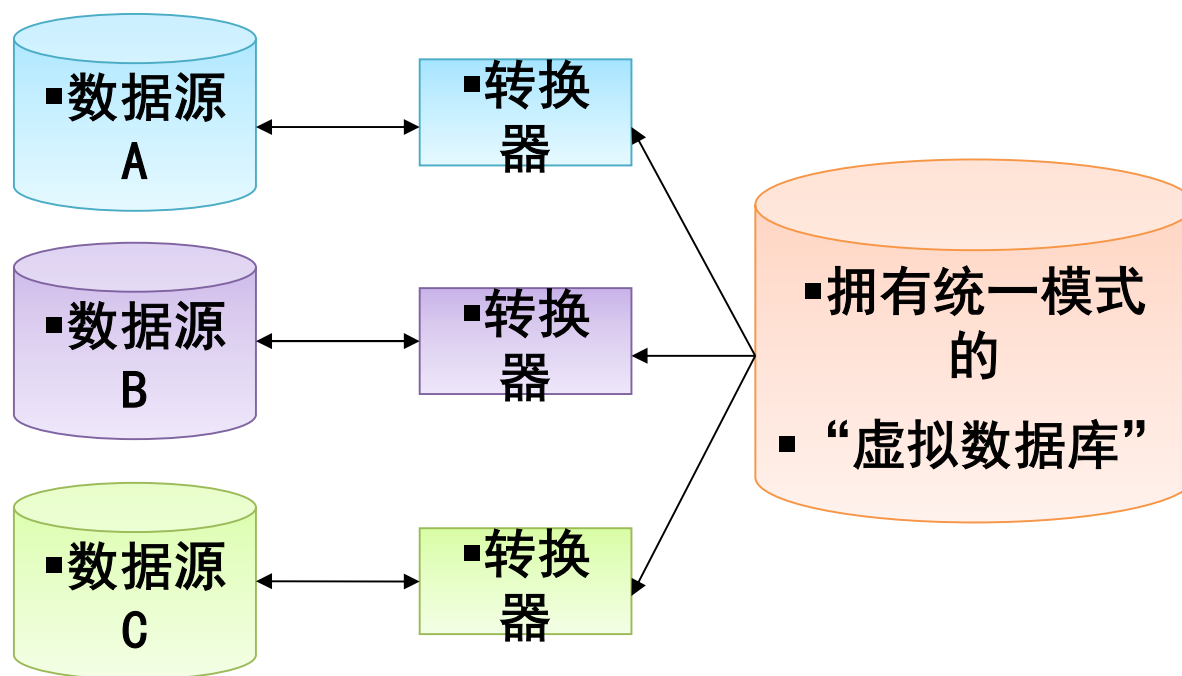
<i>CID</i>	<i>Name</i>	<i>Street</i>	<i>City</i>	<i>Sex</i>
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

## 整合结果

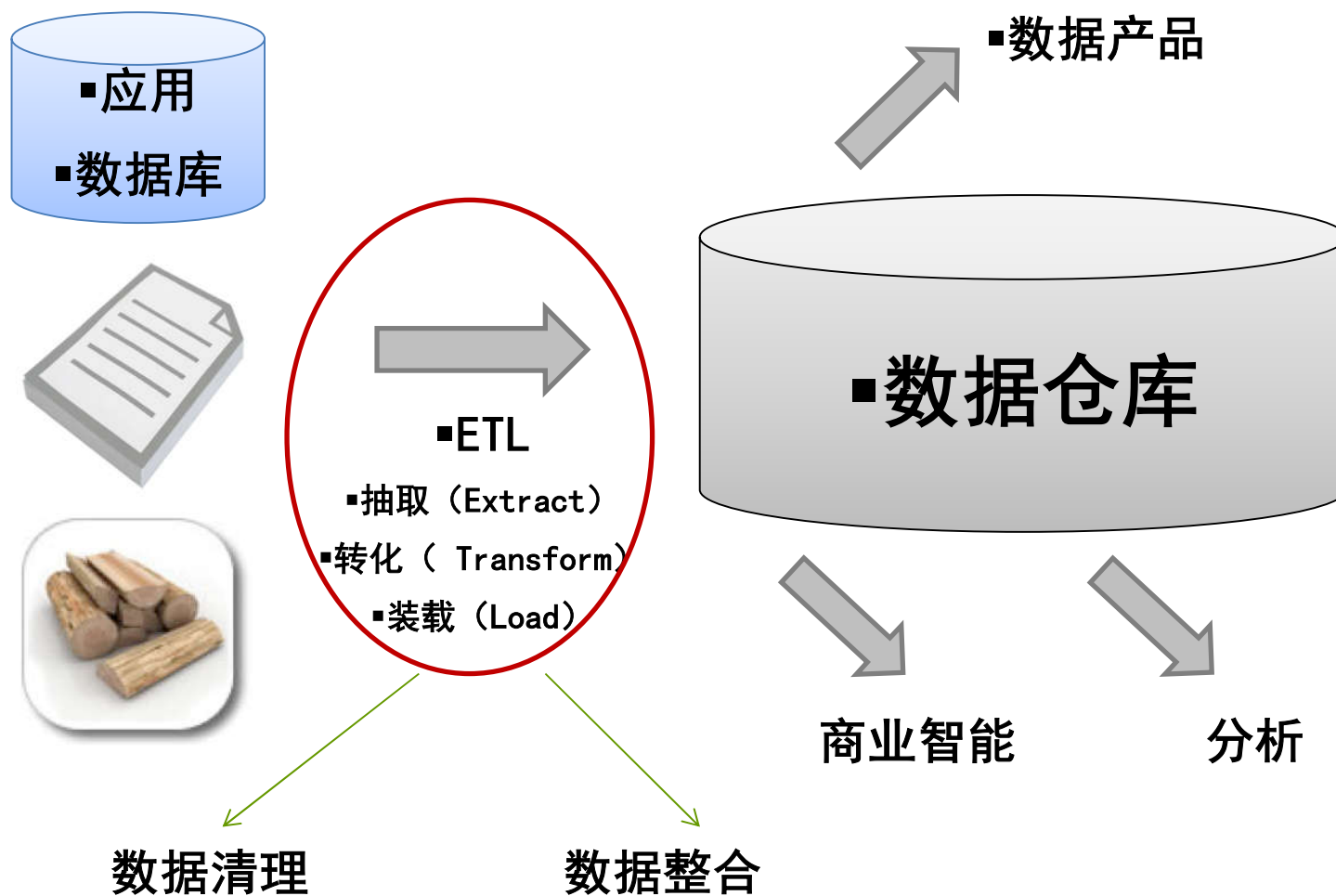
<i>No</i>	<i>LName</i>	<i>FName</i>	<i>Gender</i>	<i>Street</i>	<i>City</i>	<i>State</i>	<i>ZIP</i>	<i>Phone</i>	<i>Fax</i>	<i>CID</i>	<i>Cno</i>
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

## 另一种数据整合方式：虚拟化

---



# 标准商业数据系统架构



# 数据清洗和整合步骤

---

**初步分析：** 在操作之前进行数据分析

**冲突解析：** 解析数据源间的数据冲突

**定义数据转换 workflow 和转换规则：** 使用 workflow 方式完成模式（schema）配准和转换

**workflow 验证：** 验证 workflow 中的步骤是否正确

**数据转换：** 开始流程

# 大纲

- 数据的表达
- 数据特征
- 数据预处理
- 数据存储
- 数据分析

# 装载并存储数据

---

基于文件的存储

数据库 & 数据库管理系统

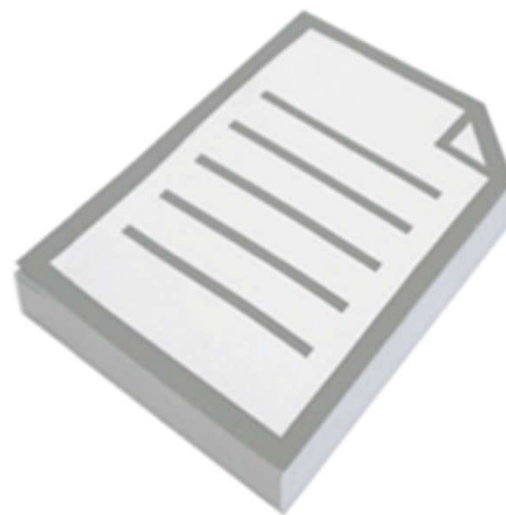
数据仓库

# 最简单的方法

---

直接将数据存储为文件形式

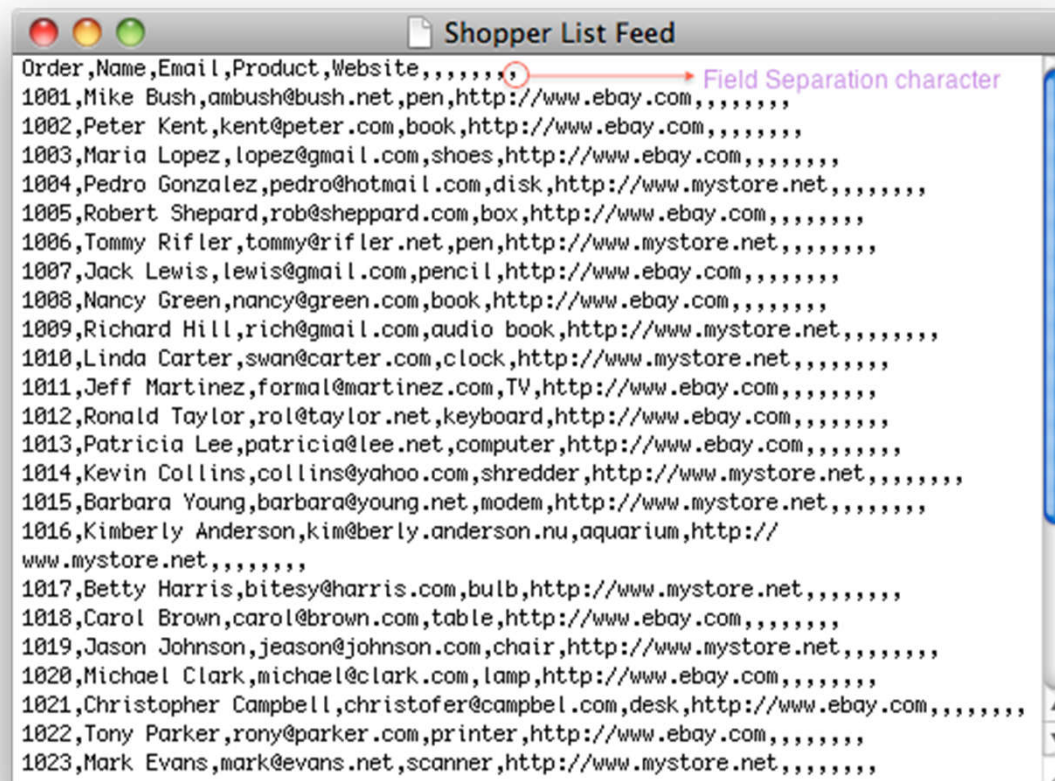
简单、方便





# 电子表格类型：CSV文件

## 逗号分隔值 (comma-separated values)



```
Order,Name,Email,Product,Website,,,,,
1001,Mike Bush,ambush@bush.net,pen,http://www.ebay.com,,,,,
1002,Peter Kent,kent@peter.com,book,http://www.ebay.com,,,,,
1003,Maria Lopez,lopez@gmail.com,shoes,http://www.ebay.com,,,,,
1004,Pedro Gonzalez,pedro@hotmail.com,disk,http://www.mystore.net,,,,,
1005,Robert Shepard,rob@sheppard.com,box,http://www.ebay.com,,,,,
1006,Tommy Rifler,tommy@rifler.net,pen,http://www.mystore.net,,,,,
1007,Jack Lewis,lewis@gmail.com,pencil,http://www.ebay.com,,,,,
1008,Nancy Green,nancy@green.com,book,http://www.ebay.com,,,,,
1009,Richard Hill,rich@gmail.com,audio book,http://www.mystore.net,,,,,
1010,Linda Carter,swan@carter.com,clock,http://www.mystore.net,,,,,
1011,Jeff Martinez,formal@martinez.com,TV,http://www.ebay.com,,,,,
1012,Ronald Taylor,rol@taylor.net,keyboard,http://www.ebay.com,,,,,
1013,Patricia Lee,patricia@lee.net,computer,http://www.ebay.com,,,,,
1014,Kevin Collins,collins@yahoo.com,shredder,http://www.mystore.net,,,,,
1015,Barbara Young,barbara@young.net,modem,http://www.mystore.net,,,,,
1016,Kimberly Anderson,kim@berly.anderson.nu,aquarium,http://
www.mystore.net,,,,,
1017,Betty Harris,bitesy@harris.com,bulb,http://www.mystore.net,,,,,
1018,Carol Brown,carol@brown.com,table,http://www.ebay.com,,,,,
1019,Jason Johnson,jeason@johnson.com,chair,http://www.mystore.net,,,,,
1020,Michael Clark,michael@clark.com,lamp,http://www.ebay.com,,,,,
1021,Christopher Campbell,christofer@campbel.com,desk,http://www.ebay.com,,,,,
1022,Tony Parker,rony@parker.com,printer,http://www.ebay.com,,,,,
1023,Mark Evans,mark@evans.net,scanner,http://www.mystore.net,,,,,
```

# 结构化文件格式

---

通用格式：XML（可扩展标记语言，eXtensible Markup Language）

- `<employer>`
- `<id>23</id>`
- `<name>Alice</name>`
- `<city>CA</city>`
- `<dptid>1</dptid>`
- `</employer>`

ID	Name	City	Dpt. ID
23	Alice	CA	1
24	Bob	NY	2

# XML的扩展

---

**IVOA VOTable: 用于交换天文学领域表格数据的XML扩展**



# XML的扩展

---

## Keyhole Markup Language (KML)：在基于web的二维或三维地图上表达地理标注信息

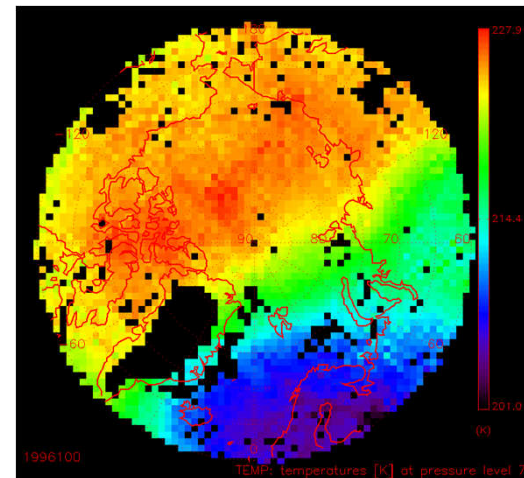


```
▪<?xml version="1.0" encoding="UTF-8"?>
▪<kml xmlns="http://www.opengis.net/kml/2.2">
▪<Document>
▪<Placemark>
▪  <name>New York City</name>
▪  <description>New York City</description>
▪  <Point>
▪    <coordinates>-74.006393,40.714172,0</coordinates>
▪  </Point>
▪</Placemark>
▪</Document>
▪</kml>
```

# 特殊用途文件格式

---

**HDF (Hierarchical Data Format): 组织和存储大量的数值型数据，特别是科学计算数据**



# 数据库

---

“A database is a collection of data, typically describing the activities of one or more related organizations.”

（数据库是数据的集合，通常用来描述多个相关组织结构的活动。）

▪--*Raghu Ramakrishnan and Johannes Gehrke, “Database Management System”*



# 关系数据库管理系统(RDBMS)

---

数据的关系模型是现代数据库系统的标准—

最小化应用程序与机器表示间的耦合度

高级数据语言：数据定义语言（Data Definition Language），结构化查询语言（Structured Query Language）

# 关系模型

---

表（关系）

列（属性）

行（记录）

约束

键：主键，外键等

索引

*Employee*

Name	EmpId	DeptName
Harry	3415	Finance
Sally	2241	Sales
George	3401	Finance
Harriet	2202	Sales

*Dept*

DeptName	Manager
Finance	George
Sales	Harriet
Production	Charles

使用结构化查询语言SQL (Structured Query Language)

**Select \* from Employee where Name = 'Harry'**



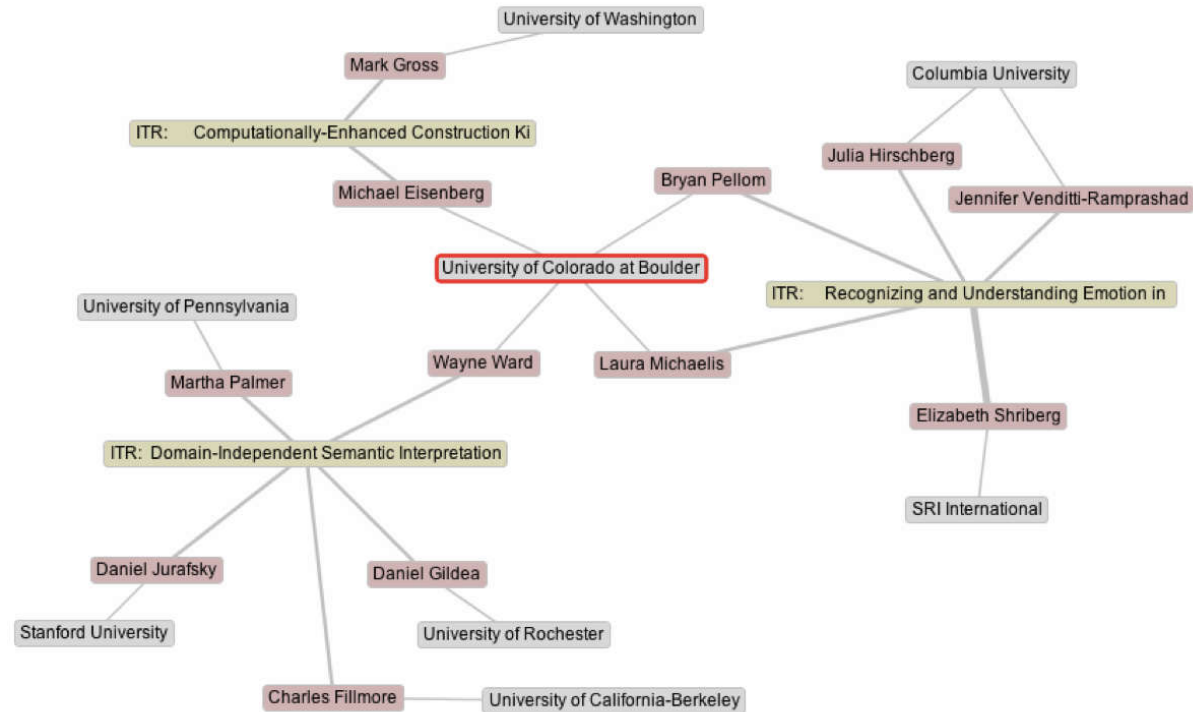
---

“When people use the word database, fundamentally what they say is that the data should be self-describing and it should have a schema. That’s really all the word database means.”

（当使用数据库这个词时，人们强调的是数据需要能够自描述，并且拥有模式。这就是“数据库”的含义。）

—Jim Gray, “The Fourth Paradigm”

# 关系数据库可视化表达



## 美国自然科学基金数据库可视化

Z. Liu, S. B. Navathe, and J. T. Stasko, Network-based visual analysis of tabular data, IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 41 - 50, 2011.

# 挑战

---

**胜任交互式任务所需的响应时间（通常为亚秒级）**

**大尺度数据的索引**

**构建数据间的语意关系**

# NoSQL数据库

---

**“Not Only SQL”（不仅仅是SQL）**

**面向海量数据（并且数据不需要关系模型）**

**通常不使用表结构，并且不使用SQL进行查询**

# NoSQL数据库实例

---

文档存储 - CouchDB

图结构存储 - Neo4j

键-值存储 - Redis（内存数据库），MongoDB（磁盘数据库）

表格数据 - Apache HBase（基于Hadoop）



# 数据仓库

---

A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process.

（数据仓库是面向主题的、已整合的、时变且稳定的数据集，用来支持管理的决策过程。）

--W. H. Inmon, "Building the Data Warehouse" . 1996.

- 更大的容量
- 很少修改
- 历史数据
- 多数据源



- 提供决策支持

# 数据仓库

---

Loosely Speaking, a data warehouse refers to a data repository that is maintained separately from an organization's operational databases.

（概括地讲，数据仓库指与企业功能数据库分离维护的数据贮藏系统。）

*--H. Jiawei and M. Kamber, "Data Mining: Concepts and Techniques", 3rd ed., 2011.*

# 数据库和数据仓库的异同

---

	数据库	数据仓库
特点	处理数据操作	处理数据中的信息
面向领域	事务	分析
用户	终端用户：职员，数据库管理员（DBA）	知识工作者：经理，分析师，执行官
功能	日常操作	长期决策支持分析
数据	当前最新的数据	历史数据，时变数据
访问方式	读写平均	（主要）读
聚焦点	数据输入	信息/知识输出
容量尺度	1GB ~ <1TB	>=TB



# 大纲

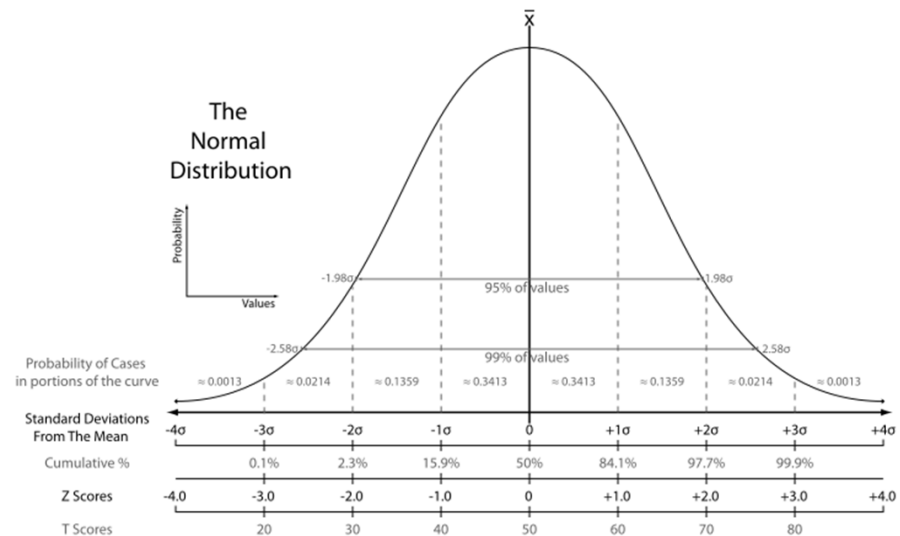
- 数据的表达
- 数据特征
- 数据预处理
- 数据存储
- 数据分析

# 数据分析

## ▪基本统计分析

-参数估计

-假设检验



(基本上讲) 统计分析是现代数据分析的基础

# 数据分析

---

## ▪基本统计分析的局限性

-现实世界及其复杂

-依赖于数据资料本身的性质、统计方法的适用程度，以及统计者自身应用水平

统计决断以概率为基础，获得结论并非绝对正确

# 探索式数据分析

(Exploratory Data Analysis, EDA)

---

从原始数据出发

不拘泥于特定模型的假设

着重方法的稳定性而非概率意义上的精确性

强调采用可视化工具

适用于普通用户

# 数据挖掘

---

**从数据中发现知识（新模式）**

**通过分类与预测、聚类分析、关联分析等方法**

**适用于海量数据**

# 数据挖掘中的方法



## ▪ 统计方法

- (回归分析; 参数估计)



## ▪ 机器学习

- (决策树; 神经网络)



## ▪ 算法方法

- (K-means, K-最近邻)

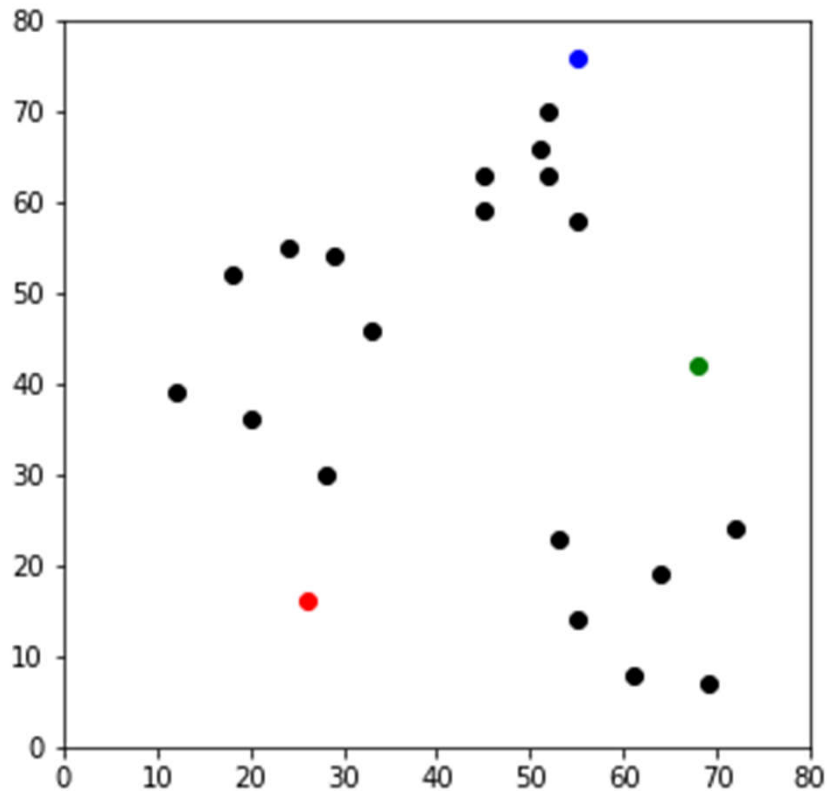
## ▪ 统计学习方法

- (概率模型; 贝叶斯网络)

# 数据挖掘方法举例

## K-means

---



■ K均值聚类方法是典型的非监督学习方法

假设有一组数据，需要分为3类

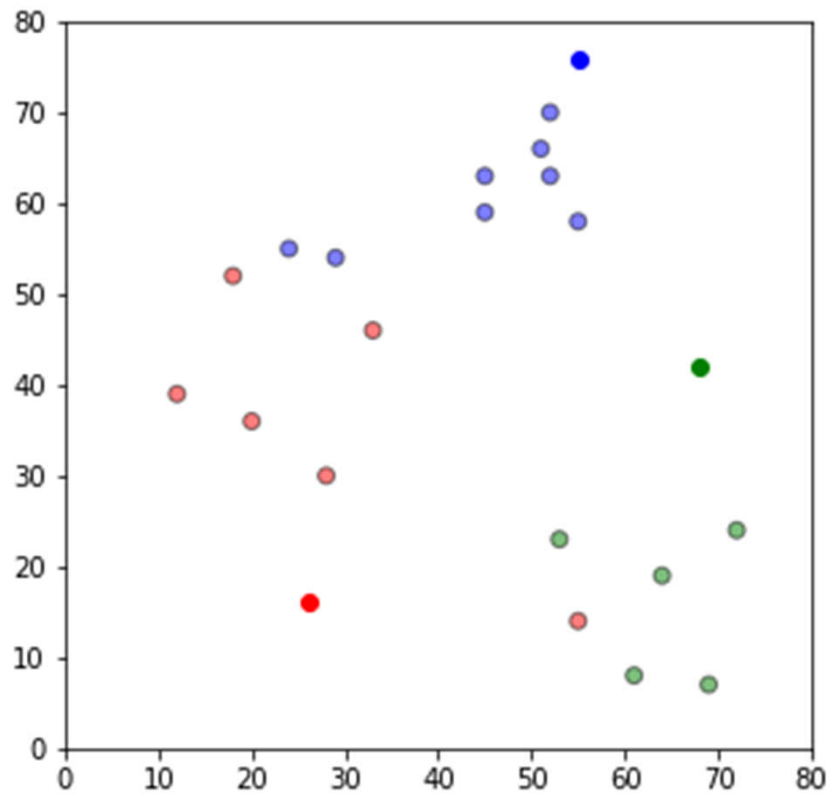
首先给出3个聚类中心，即

$K = 3$

# 数据挖掘方法举例

## K-means

---



▪分类：针对每个点，分别计算与3个类别中心的距离

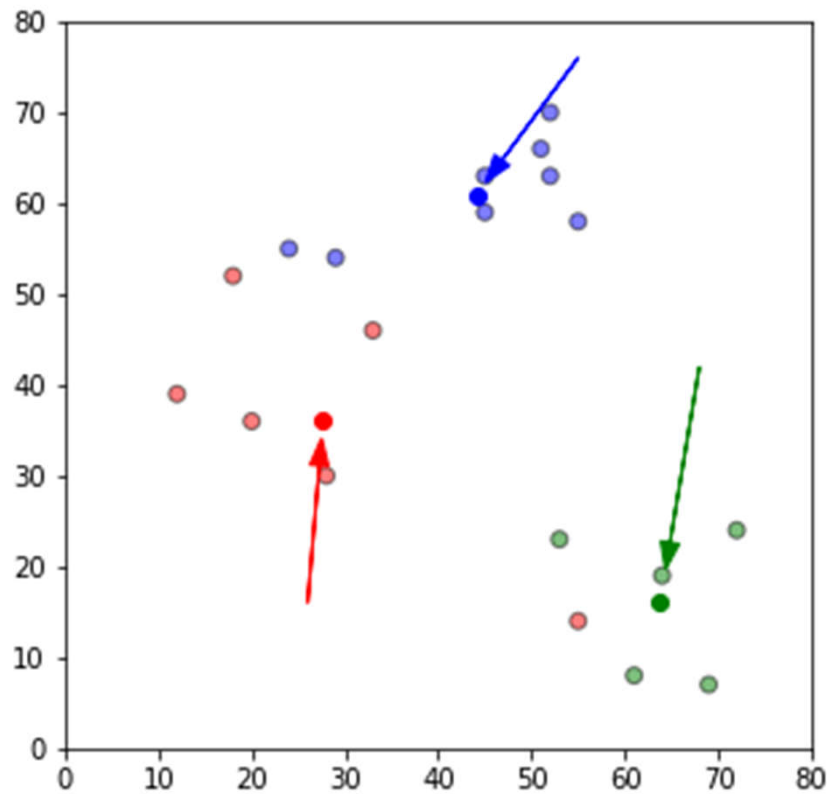
选择最近的类别中心为当前点的类别



# 数据挖掘方法举例

## K-means

---



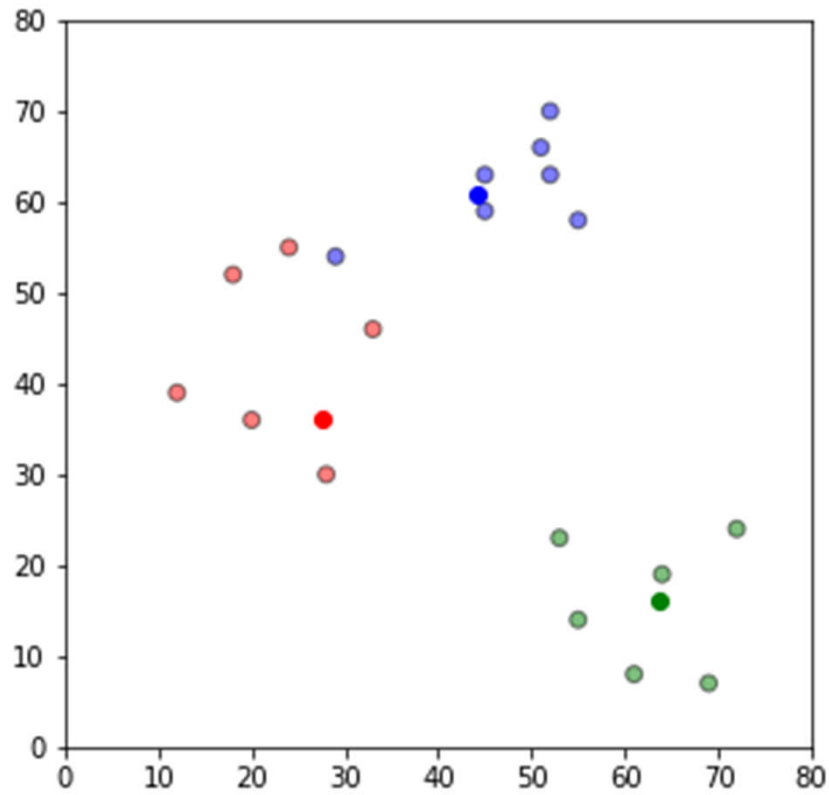
- 更新类别中心坐标

中心坐标=当前类别的几何中心

# 数据挖掘方法举例

## K-means

---

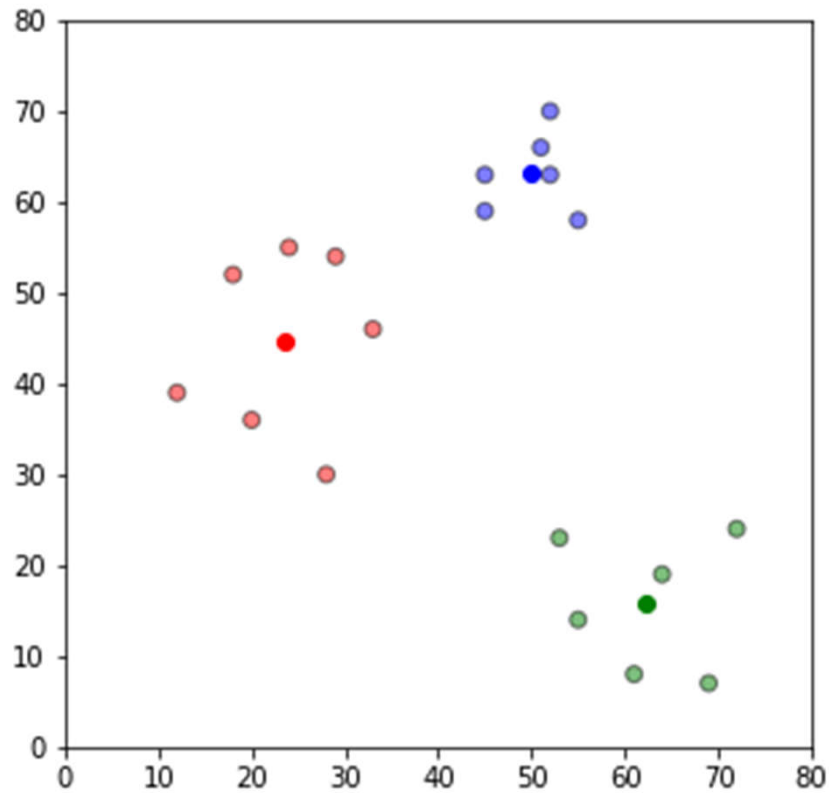


■重新分类

# 数据挖掘方法举例

## K-means

---



- 重新更新聚类中心坐标
- 反复计算，直至收敛

# K-means

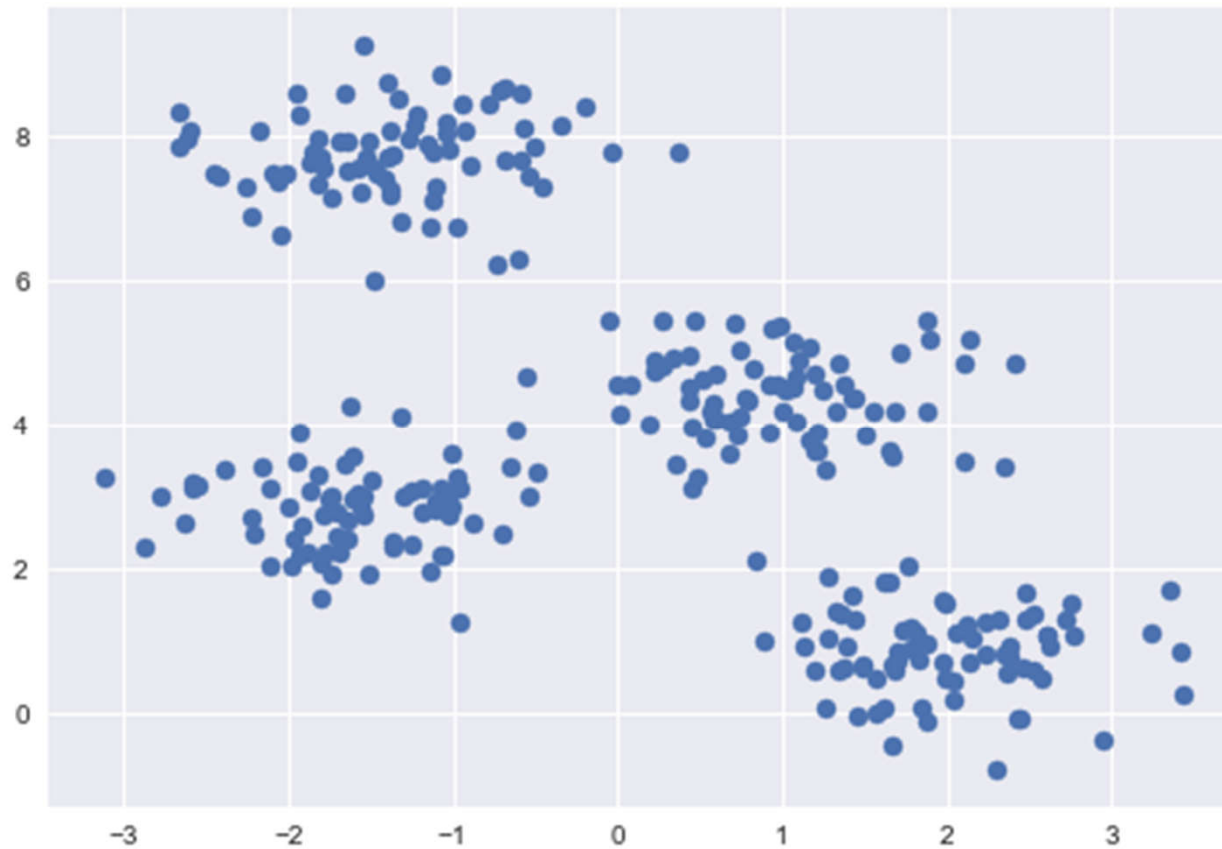
- 距离尺度

**▪ 在示例中使用了什么距离尺度?**

# K-means

- 距离尺度

▪ 在示例中使用了什么距离尺度?



# K-means

- 距离尺度

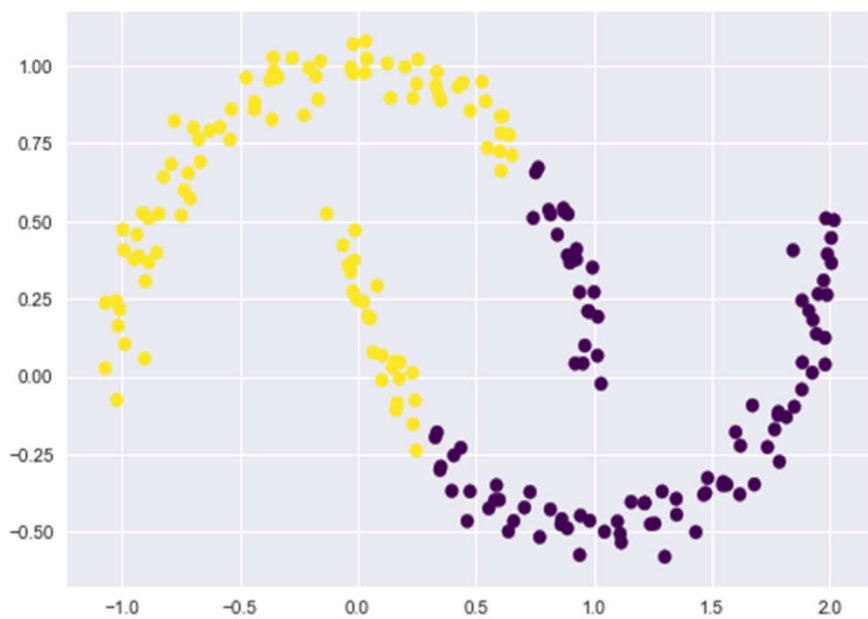
▪ 在示例中使用了什么距离尺度?



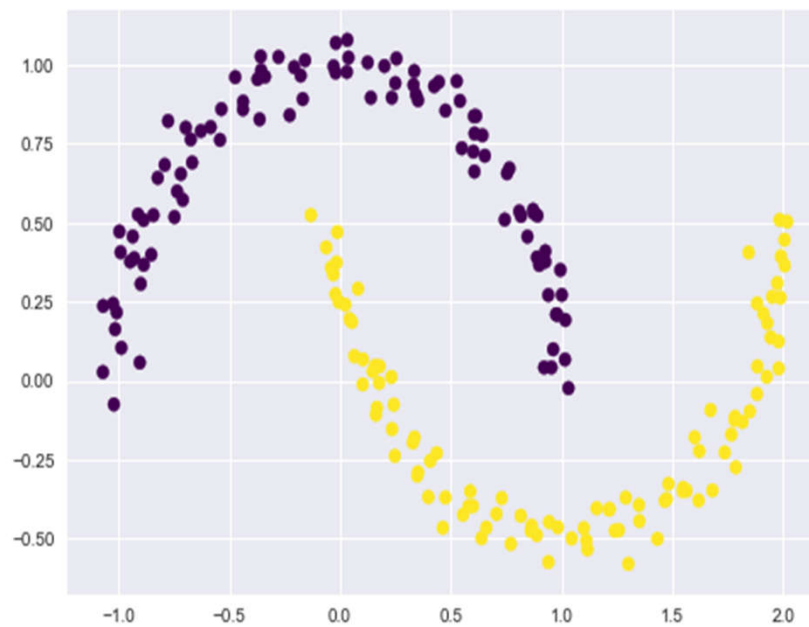
# K-means

- 距离尺度

- 在示例中使用了什么距离尺度?



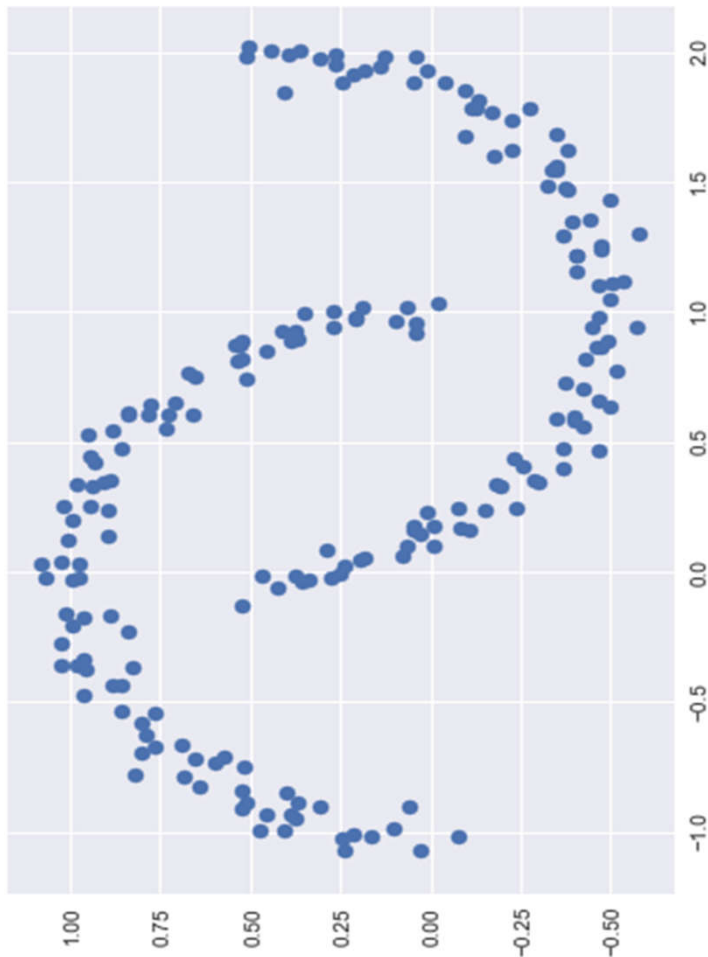
- 欧式距离?



# 可选作业1

## K-means

---



- 根据左面图例生成二维点云数据
- 采用K-means方法聚类,  $K=2$
- 完成聚类算法实现, 可采用 matlab/octave/c++/python
- 分别采用曼哈顿距离/欧式距离/马氏距离作为距离评价方法
- 对比分析算法效果, 给出技术报告。



# 总结：为什么讲数据？

---

数据是一切模型的基础

数据是进行实验设计的参考

数据能检验科学研究中的每一个阶段