

# 数据挖掘与机器学习

潘斌

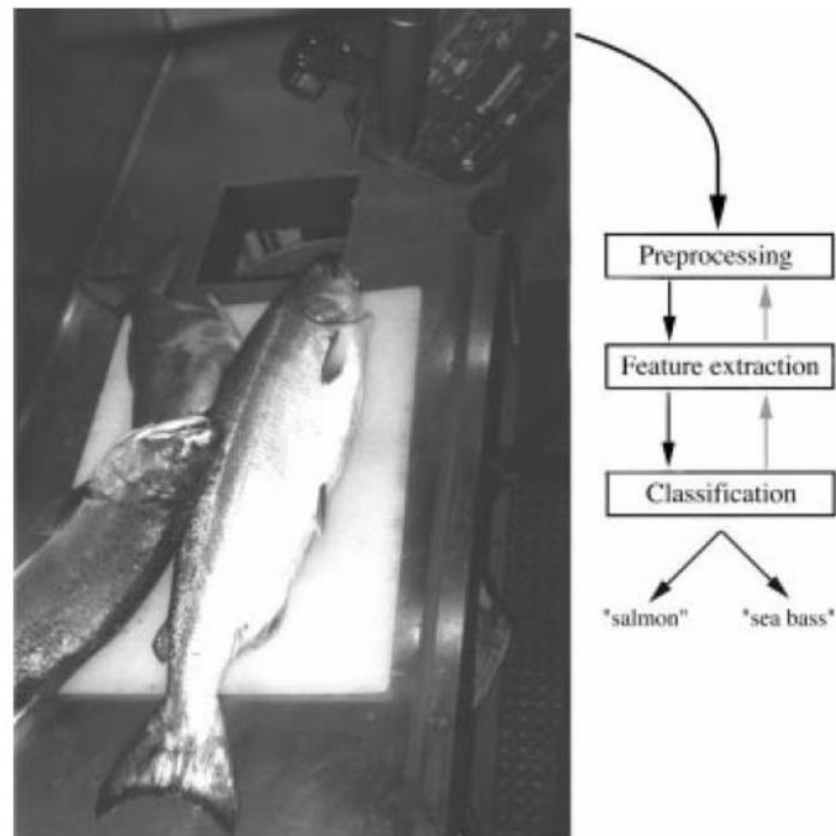
panbin@nankai.edu.cn

范孙楼227

1

# 上节回顾

- 什么是数据挖掘
  - 数据挖掘是在大型数据存储库中，自动的发现有用信息的过程
- 什么是机器学习
  - 可自动发现有用信息的手段即为机器学习算法
- 什么是大数据
  - 大数据具有4V特征



# 本节提要

- 数据的特点
  - 数据集的一般特性
  - 数据质量
  - 数据预处理
- 特征学习
  - 特征提取
  - 特征选择
- 概念学习
  - 总体、目标、样本、假设





# 数据

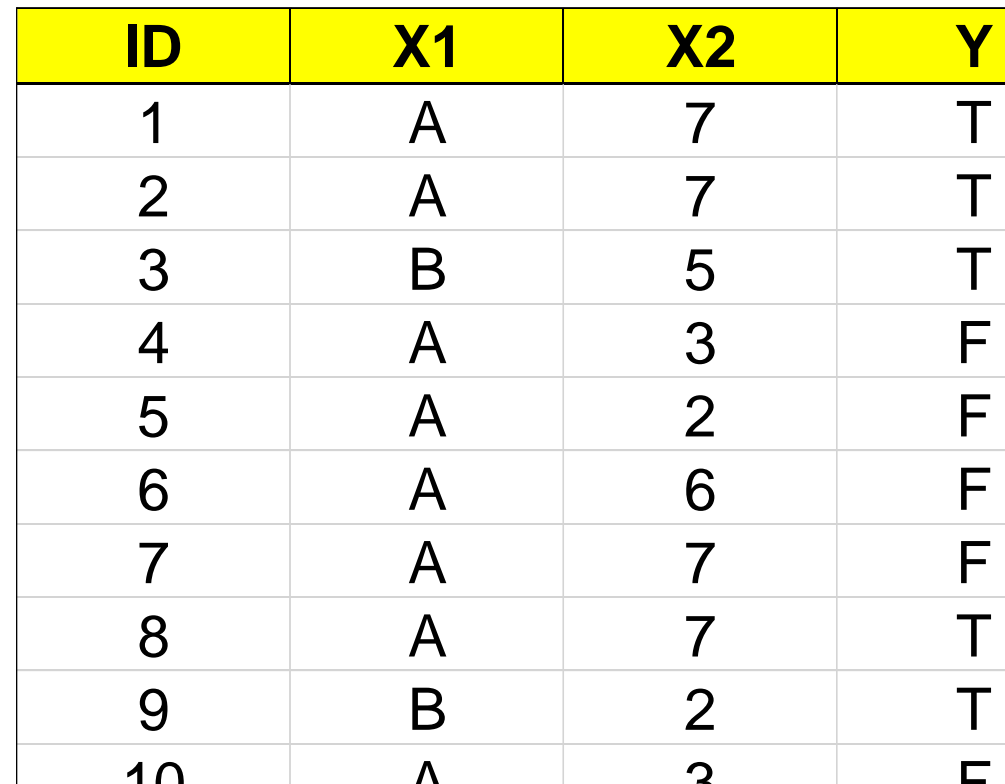
The Data

# 属性

- 属性（Attributes）是对象（Objects）的性质或特征
- 也称为变量、特性、字段、特征（子集）、或维
  - 因对象而异
    - 人眼颜色：符号属性 { 棕色，黑色，蓝色，绿色，淡褐色，..... }
  - 随时间变化
    - 物体温度：数值属性，可取无穷多个值
- 一组属性集刻画对象的基本特征——数据对象
  - 记录、点、向量、模式、事件、案例、样本、观察或实体

# Attributes

## Objects



ID	X1	X2	Y
1	A	7	T
2	A	7	T
3	B	5	T
4	A	3	F
5	A	2	F
6	A	6	F
7	A	7	F
8	A	7	T
9	B	2	T
10	A	3	F

# 属性类型

## ■ 用值的个数描述属性

- 离散的
  - 具有有限个值或无限可数个值
- 连续的
  - 取实数值的属性

## ■ 非对称的属性

- 出现非零属性值才是重要的

## ■ 对属性量化

实际的空间结构

A	B
C	D



抽象的空间相邻  
矩阵 (Rook 型)

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$



# 数据对象类型

- 记录（Record）数据：数据集是记录（样本）的汇集，每个记录包含固定的数据字段（特征）集。
- 图形（Graph）数据：基于图形的数据。
- 有序（Ordered）数据：数据间存在序的联系。

# 记录数据

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

- 数据矩阵 ( Data Matrix )
  - 所有数据对象都有相同的数值属性集，则可将数据对象视为多维空间中的点（向量），其中每一维代表对象的一个不同属性
  - 数据对象集可以用一个  $m \times n$  矩阵表示，一个对象一行（ $m$ 行），一个属性一列（ $n$ 列）

# 记录数据

- 文档数据（Document Data）
  - 每一文档为一词（term）向量
    - 每个词是向量的一个分量（属性）
    - 每个分量的值对应词在文档中出现的次数
    - 字典学习法

	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

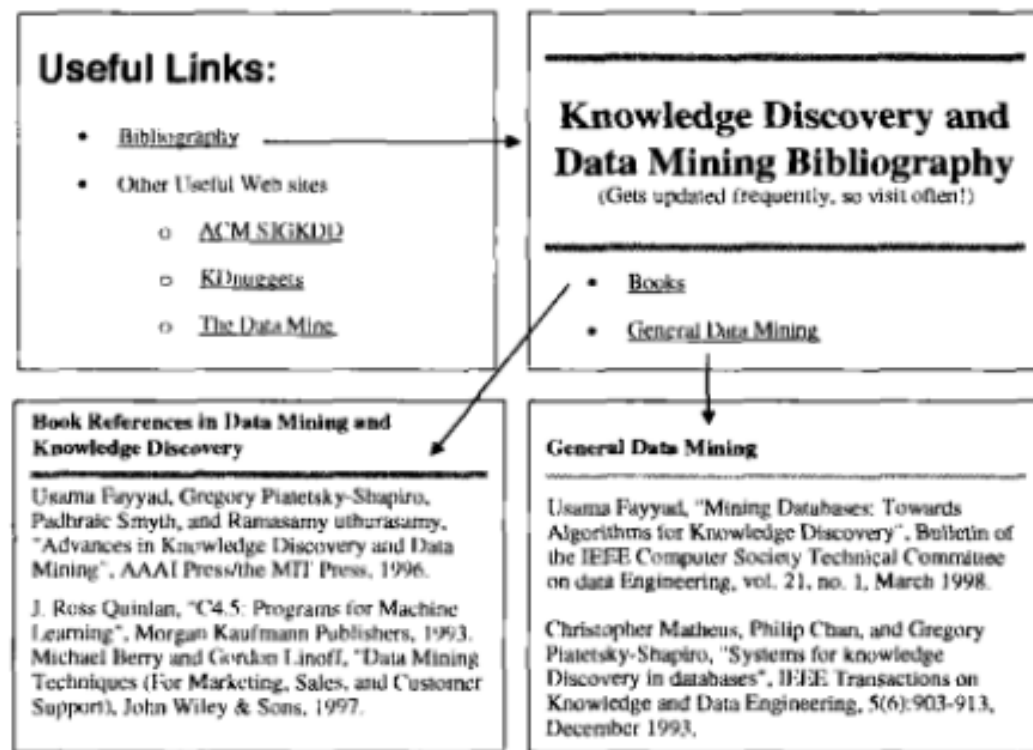
# 记录数据

- 事务数据 ( Transaction Data )
  - 特殊类型的记录数据，每个记录 ( 事务 ) 涉及一系列的项
  - 记录中的项是顾客“购物篮”中的商品——购物篮数据 ( market basket data )

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

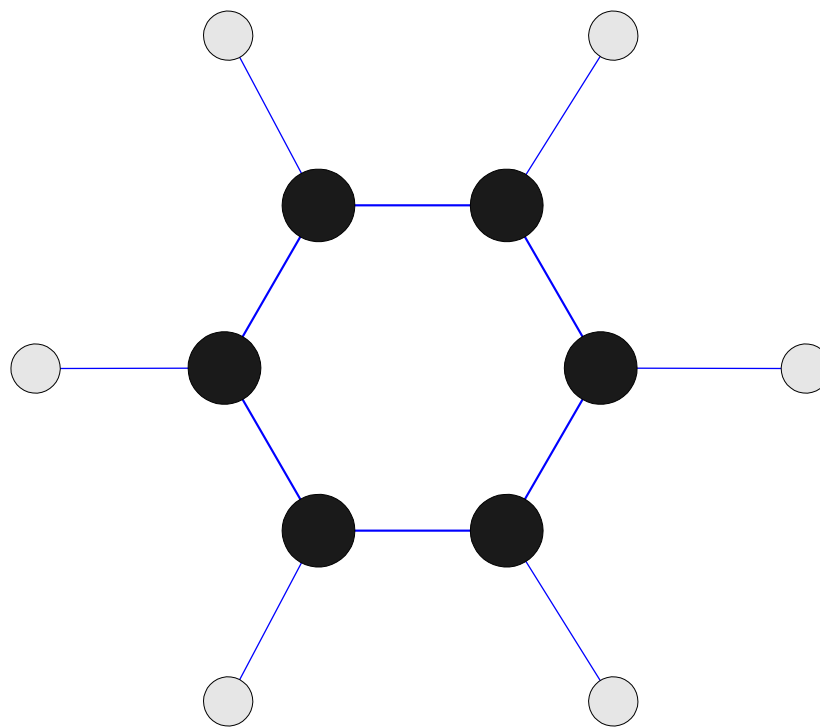
# 图形数据

- 带有对象之间联系的数据
  - 对象之间的联系带有重要信息



# 图形数据

- 具有图形对象的数据
  - 对象具有结构



苯分子结构：C<sub>6</sub>H<sub>6</sub>

# 有序数据

- 时间数据（Temporal Data），也称时序数据（Sequential Data）
  - 每个记录包含一个与之相关的时间

时间	顾客	购买的商品
t1	C1	A,B
t2	C3	A,C
t2	C1	C,D
t3	C2	A,D
t4	C2	E
t5	C1	A,E

# 有序数据

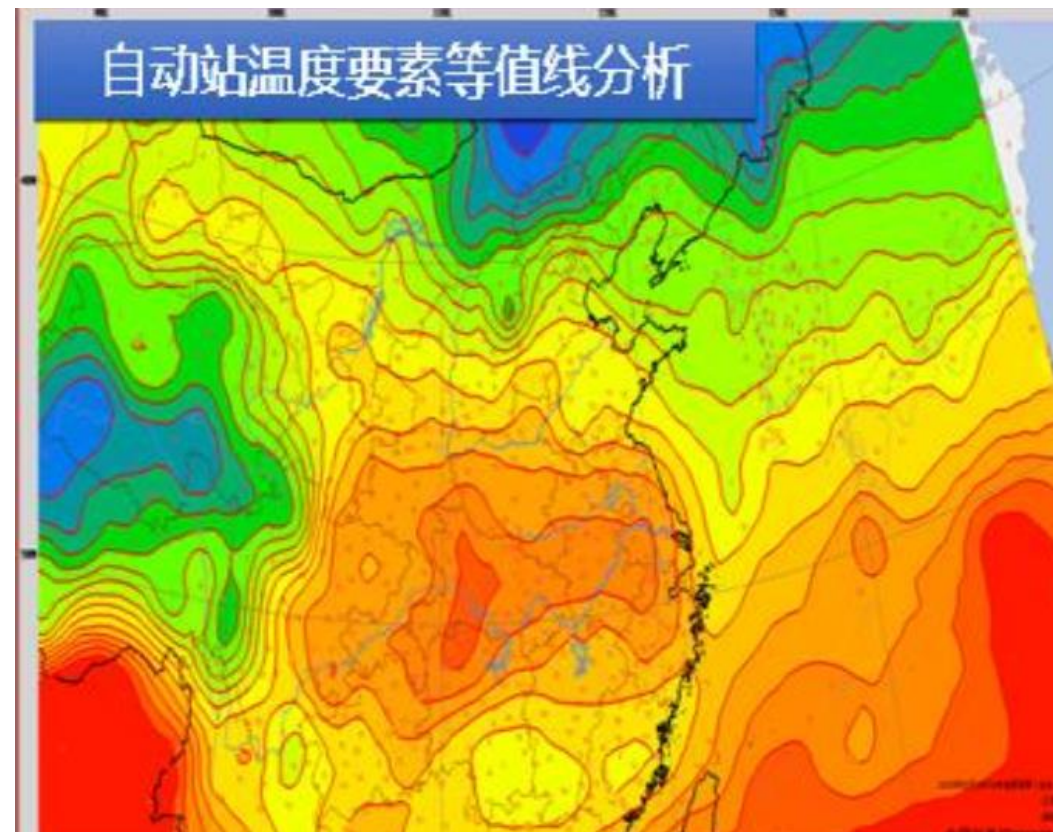
- 序列数据 ( Sequence Data )
  - 有序序列考虑项的位置
  - 与时序数据非常相似
    - 基因序列数据

GGTTC CGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG



# 有序数据

- 空间数据（Spatial Data）
  - 具有空间属性，如位置或区域



# 数据集的一般特性

- 维度 ( Dimensionality )
  - 数据集中的对象具有的属性数目
  - 维数灾难 ( Curse of Dimensionality )
- 稀疏性 ( Sparsity )
  - 一个对象的大部分属性上的值都为0
- 分辨率 ( Resolution )
  - 不同分辨率下数据的性质不同
  - 模式依赖于分辨率水平



# 数据探索 ( DATA EXPLORATION )

- 拿到数据的 **第一件事**
- 对数据初步研究，以更好理解其特殊性质
  - 有助于选择合适的预处理技术和数据分析技术
  - 可以处理一些通常由数据挖掘解决的问题
    - 如，特征的设计
  - 理解和解释数据挖掘的结果



# 鸢尾花 ( IRIS ) 数据集

- 包含150个鸢尾花信息
- 取自三个物种
  - 山鸢尾 ( Setosa ) ; 维吉尼亚鸢尾 ( Virginica ) ; 变色鸢尾 ( Versicolour )
- 特征用五种属性描述
  - 萼片长度 ( cm ) ; 萼片宽度 ( cm ) ; 花瓣长度 ( cm ) ; 花瓣宽度 ( cm ) ; 类 ( 属种 )



萼片长度	萼片宽度	花瓣长度	花瓣宽度	类
5.1	3.5	1.4	0.2	<i>setosa</i>
4.9	3	1.4	0.2	<i>setosa</i>
.....				
5.7	2.9	4.2	1.3	<i>versicolor</i>
5.7	2.8	4.1	1.3	<i>versicolor</i>
.....				
5.8	2.7	5.1	1.9	<i>virginica</i>
7.1	3	5.9	2.1	<i>virginica</i>
.....				

鸢尾花数据集（部分）



# 汇总统计 ( SUMMARY STATISTICS )

- 用单个数或数的小集合捕获可能很大的值集的各种特征
- 频率：给定一个在 $\{v_1, v_2, \dots, v_k\}$ 取值的分类属性 $x$ 和 $m$ 个对象的集合，值 $v_i$ 的频率定义为 
$$\text{frequency}(v_i) = \frac{\text{具有属性值 } v_i \text{ 的对象数}}{m}$$
- 众数：具有最高频率的值



一所假想大学中各年级学生人数

年级	人数	频率
一年级	200	0.33
二年级	160	0.27
三年级	130	0.22
四年级	110	0.18

则年级属性的众数为“一年级”。



# 汇总统计

- 众数（统计量）的分辨率问题
  - 对于连续属性，按照目前的定义，众数通常没有用。
  - 以毫米为单位，20个人的身高通常不会重复，但如果以分米为单位，则某些人很可能具有相同的身高。
  - 众数可以用来估计缺失值。





# 汇总统计

- 百分位数

- $x_p$ : 对应一个  $x$  值, 使得  $x$  的  $p\%$  观测值小于  $x_p$

萼片长度、萼片宽度、花瓣长度和花瓣宽度的百分位数 (所有的值都以厘米为单位)

百分位数	萼片长度	萼片宽度	花瓣长度	花瓣宽度
0	4.3	2.0	1.0	0.1
10	4.8	2.5	1.4	0.2
20	5.0	2.7	1.5	0.2
30	5.2	2.8	1.7	0.4
40	5.6	3.0	3.9	1.2
50	5.8	3.0	4.4	1.3
60	6.1	3.1	4.6	1.5
70	6.3	3.2	5.0	1.8
80	6.6	3.4	5.4	1.9
90	6.9	3.6	5.8	2.2
100	7.9	4.4	6.9	2.5



# 汇总统计

- 位置度量：均值和中位数

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

- $$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

萼片长度、萼片宽度、花瓣长度和花瓣宽度的均值、中位数和截断均值  
(所有值都以厘米为单位)

度量	萼片长度	萼片宽度	花瓣长度	花瓣宽度
均值	5.84	3.05	3.76	1.20
中位数	5.80	3.00	4.35	1.30
截断均值 (20%)	5.79	3.02	3.72	1.12



# 汇总统计

- 散布度量：极差和方差

- 极差：  $\text{range}(x) = \max(x) - \min(x) = x_{(m)} - x_{(1)}$

- 方差：  $\text{variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$

- 绝对平均偏差：（ absolute average deviation ）

$$\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$$

- 中位数绝对偏差（ median absolute deviation ）

$$\text{MAD}(x) = \text{median} \left( \{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\} \right)$$

- 四分位数极差（ interquartile range ）

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$



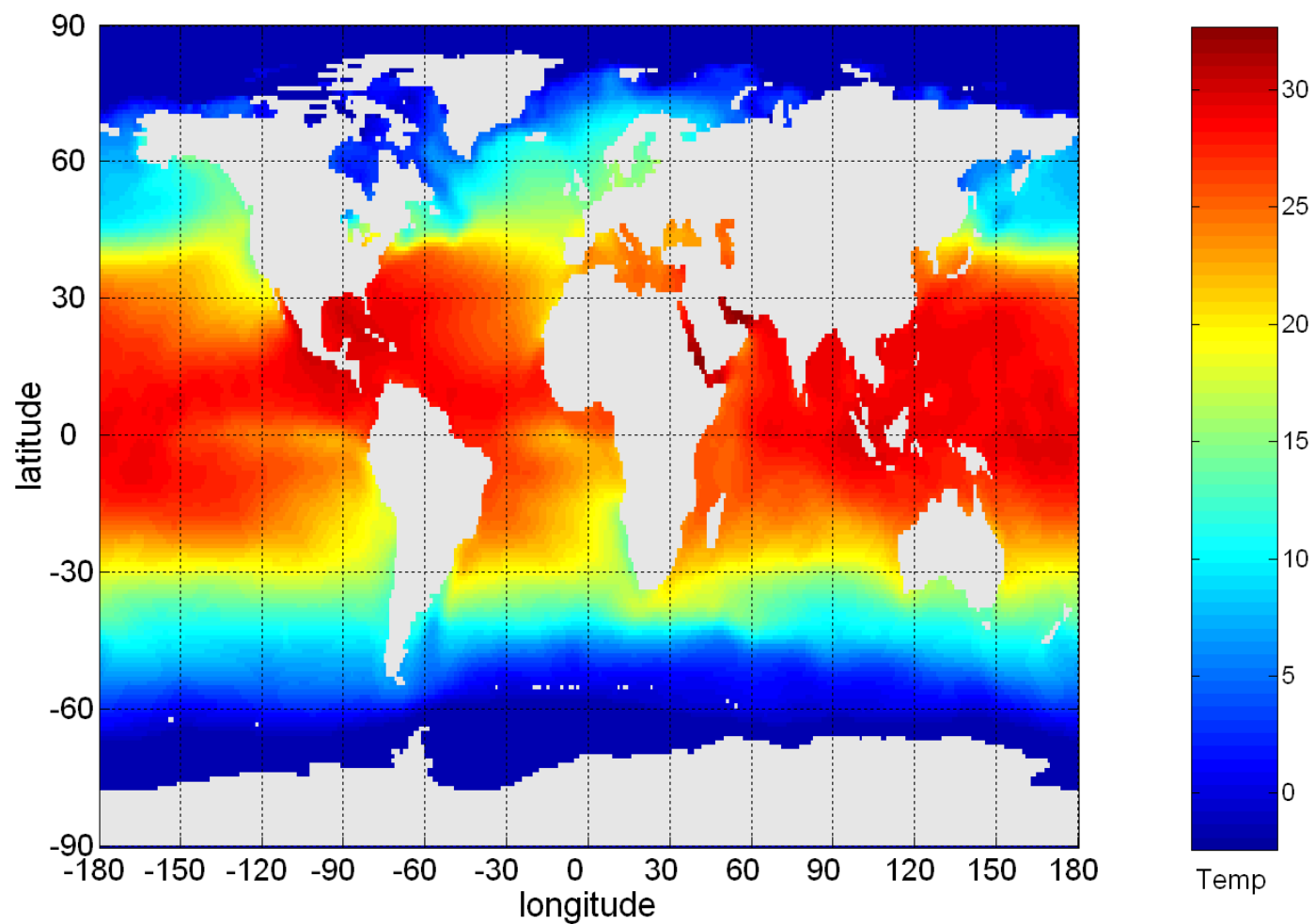
# 汇总统计

萼片长度、萼片宽度、花瓣长度和花瓣宽度的极差、标准差（std）、绝对平均偏差（AAD）、中位绝对偏差（MAD）和中间四分位数极差（IQR）（所有值都以厘米为单位）

度量	萼片长度	萼片宽度	花瓣长度	花瓣宽度
极差	3.6	2.4	5.9	2.4
std	0.8	0.4	1.8	0.8
AAD	0.7	0.3	1.6	0.6
MAD	0.7	0.3	1.2	0.7
IQR	1.3	0.5	3.5	1.5



# 可视化 ( VISUALIZATION )



- 颜色
- 标尺



# 可视化

## ■ 重新安排数据的重要性

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

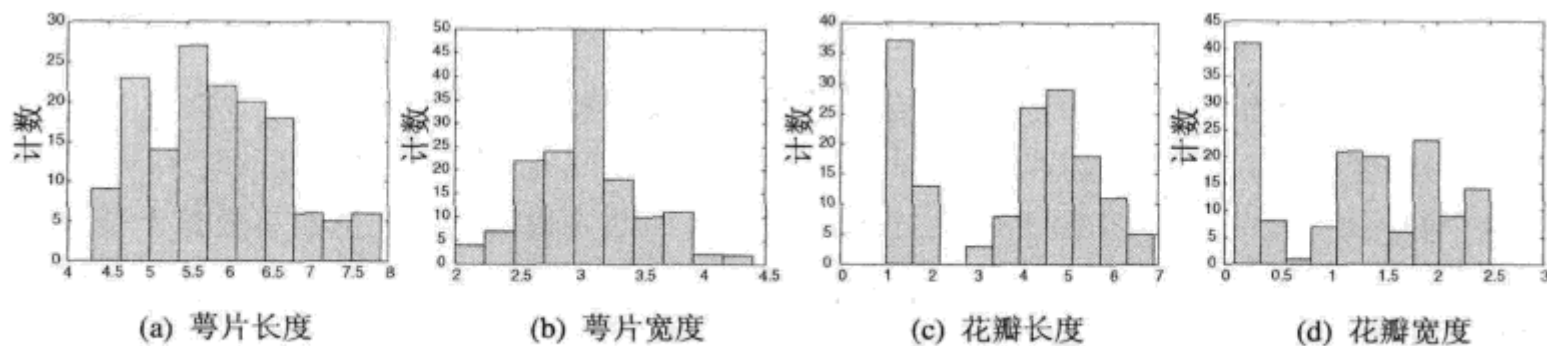


	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

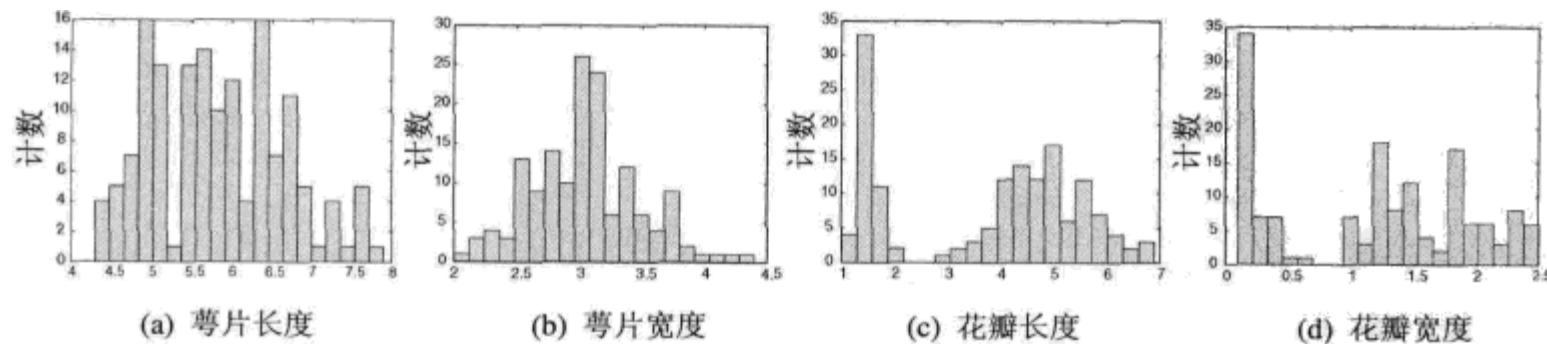


# 可视化

## ■ 直方图 ( Histogram )



四个鸢尾花属性的直方图 (10 个箱)

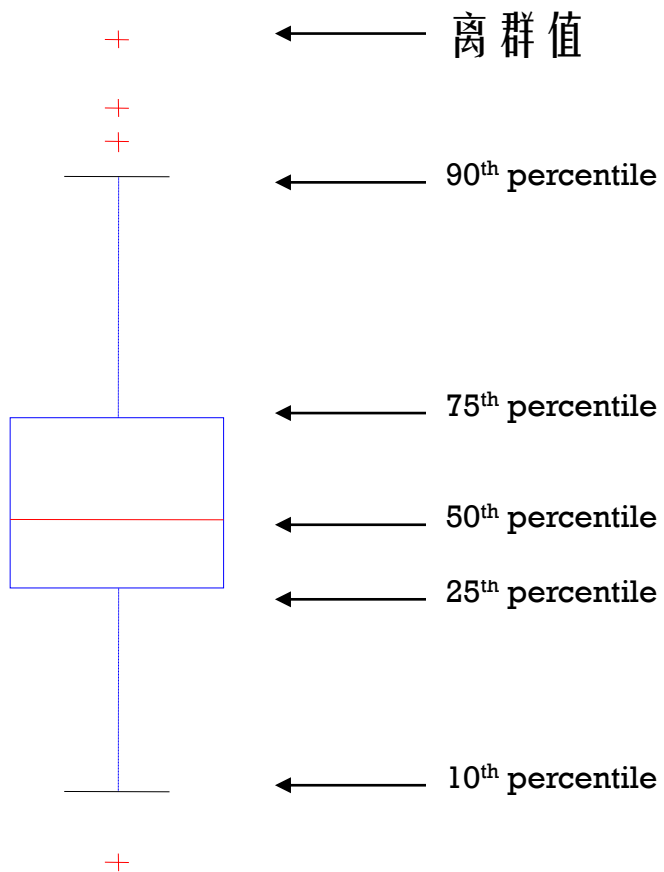


四个鸢尾花属性的直方图 (20 个箱)



# 可视化

## ■ 箱状图 ( box plot )

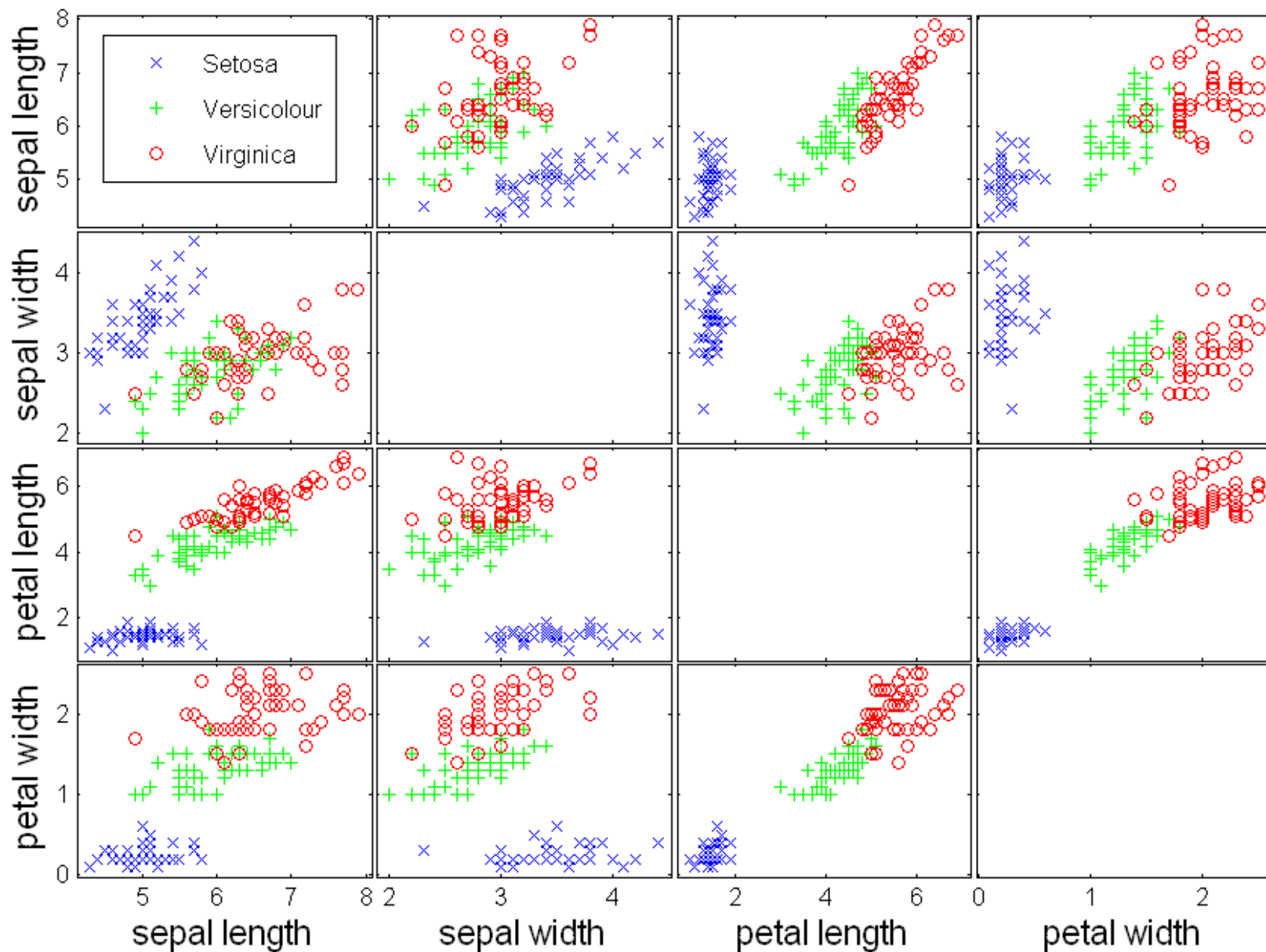




# 可视化

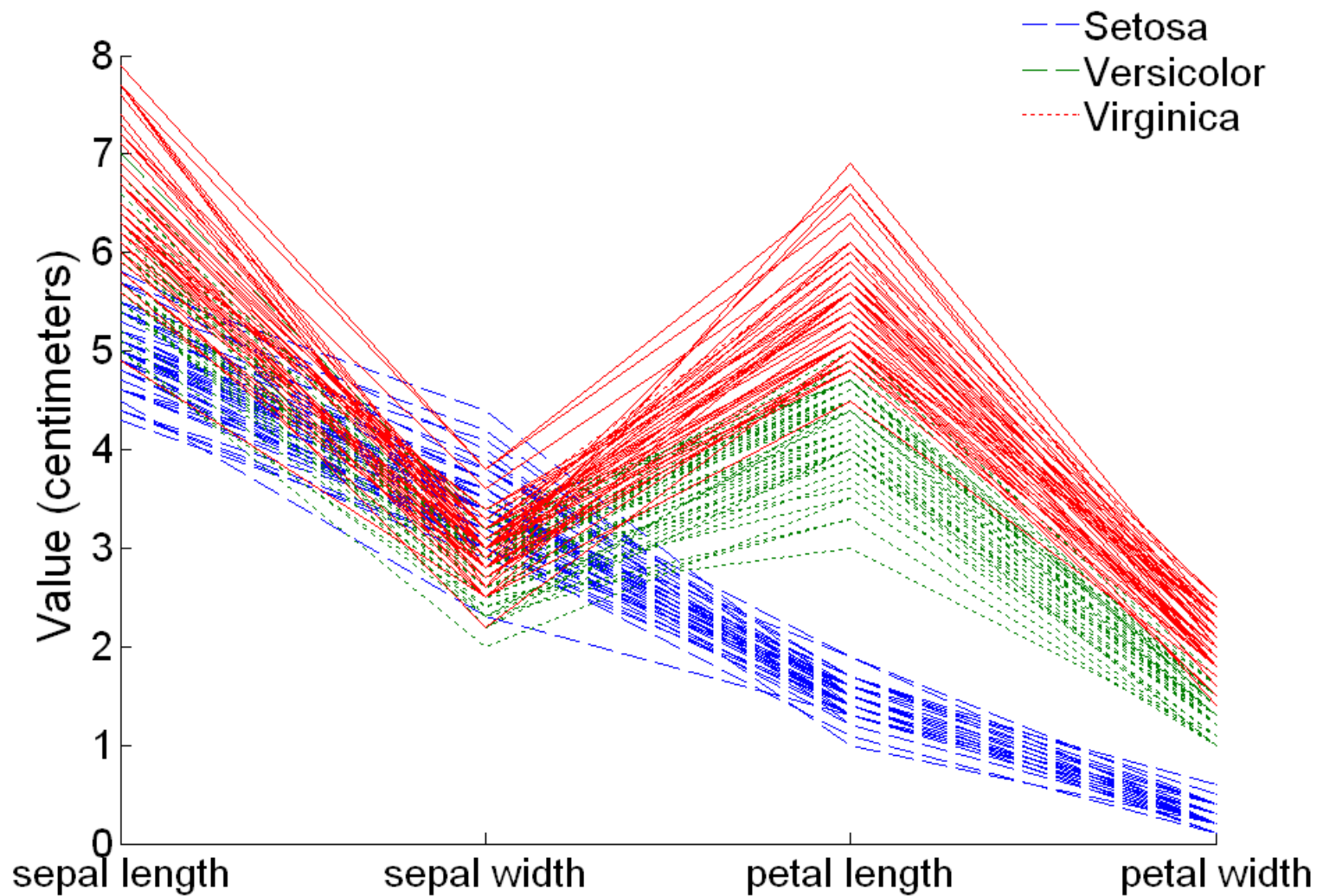
## ■ 散布图 ( Scatter plots )

- 图形化显示二属  
性之间的关系
- 当类标号给出时，  
考察二属性度



# 可视化

- 平行坐标系 ( Parallel Coordinates )



# 数据质量

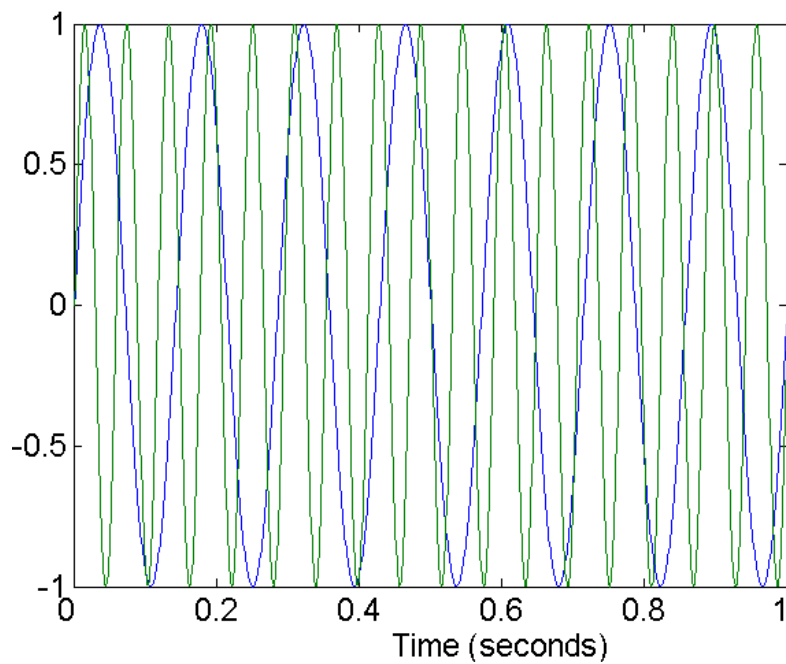
- 数据并非完美
  - 人为错误
  - 设备限制
  - 搜集漏洞
- 无法避免
- 两个方面
  - 数据质量问题的检测和纠正（预处理）
  - 使用可以容忍低质量数据的算法（鲁棒）

# 数据质量

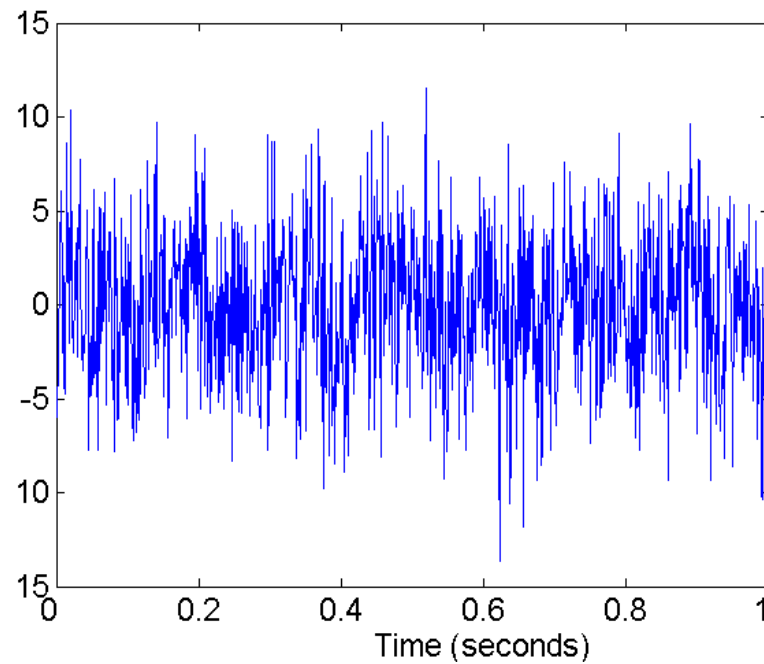
- 噪声：测量误差的随机部分
  - 可通过使用信号或图像处理技术降低噪声，很难完全消除
  - Robust algorithm 的使用能产生可以接受的结果
  - 例：



# 数据质量



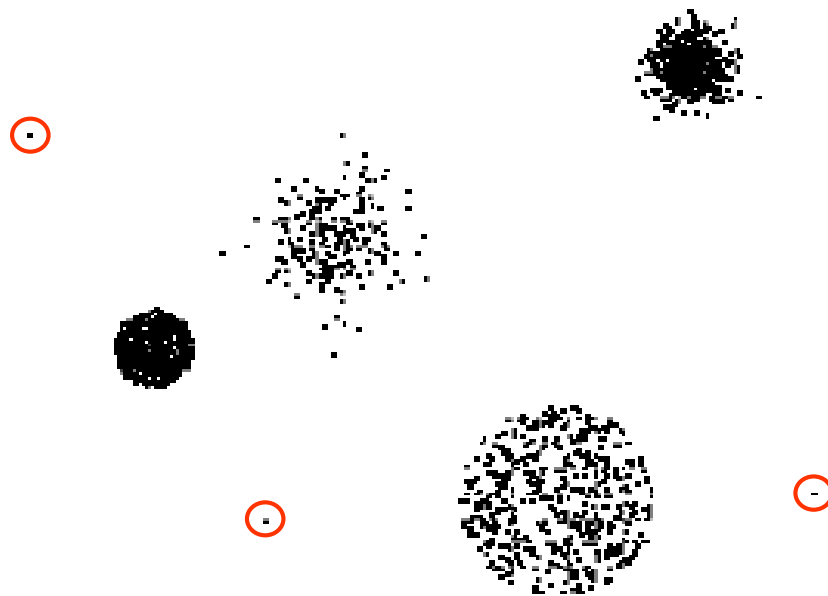
Two Sine Waves



Two Sine Waves + Noise

# 数据质量

- 离群点 ( Outlier )
  - 在某种意义上具有不同于数据集中其他大部分数据对象的特征，或相对于该属性典型值来说不寻常的特征
  - 异常 ( anomalous ) 对象，异常值
  - 可以是合法的数据对象或值
  - 有时是感兴趣的对象



# 数据质量

- 缺失值 ( Missing Values )
  - 信息搜集不全
  - 某些属性并不能用于所有对象
- 处理策略
  - 删除数据对象或属性
  - 估计缺失值
  - 忽略缺失值
  - 用所有可能值代替 ( 可能性为权重 )

# 数据质量

- 重复数据 ( Duplicate Data )
  - 数据集可能包含重复或几乎重复的数据
  - 如，重复邮件
- 不一致的值
  - 如，地址字段列出了邮政编码和城市名，但有的邮政编码区域并不包含在对应的城市中
  - 负数的身高



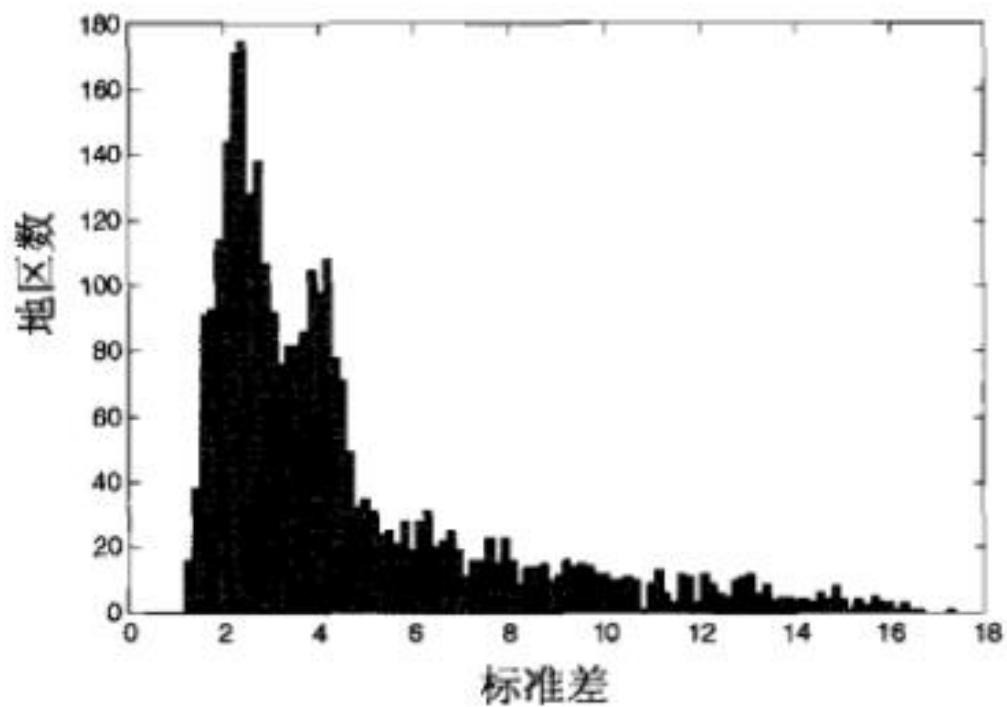
# 数据预处理

- 聚集 ( Aggregation )
- 抽样 ( Sampling )
- 特征创建 ( Feature creation )
- 离散化和二元化 ( Discretization and Binarization )
- 属性变换 ( Attribute Transformation )
- 维归约 ( Dimensionality Reduction )
- 特征子集选择 ( Feature subset selection )

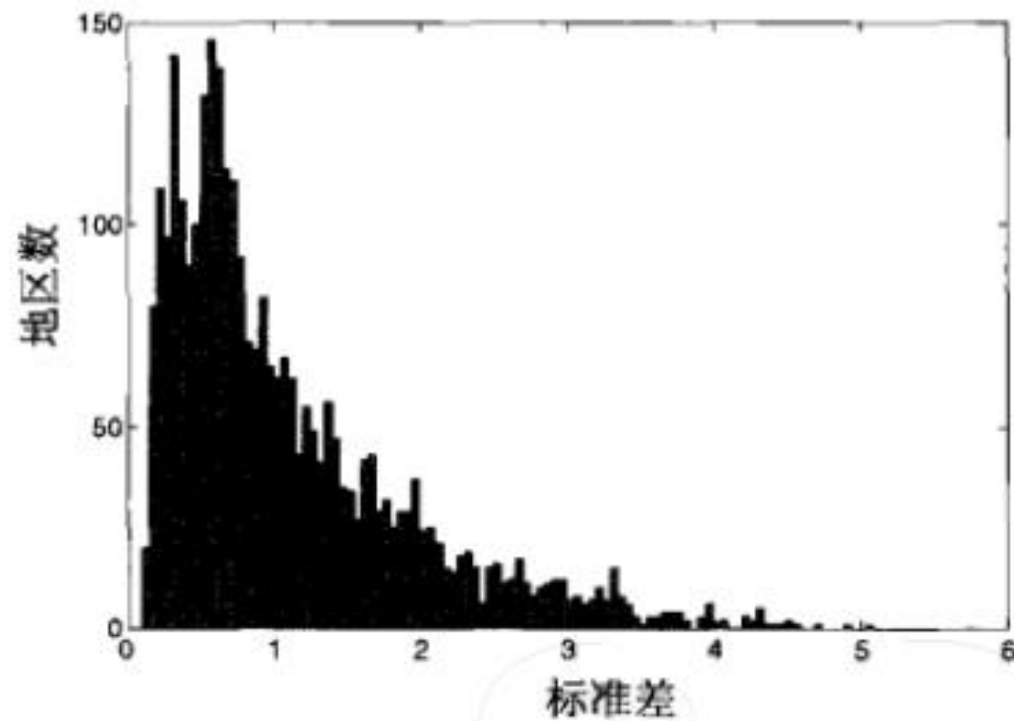
# 数据预处理

- 聚集
  - 将两个或多个对象合并为单个对象
  - 范围或标度的转换
    - 城市聚集为地区、州、国家等
    - 数据由按天记录聚集为按月记录
  - 对象或属性群的行为通常比单个对象或属性的行为更加稳定 ( stable )
    - 平均值、总数等的变异性 ( variability ) 较小
    - 可能丢失有趣的细节

■ 澳大利亚降水量（1982-1993）



(a) 平均月降水量标准差的直方图



(b) 平均年降水量标准差的直方图

# 数据预处理

- 抽样

- 选择数据对象子集进行分析的常用方法

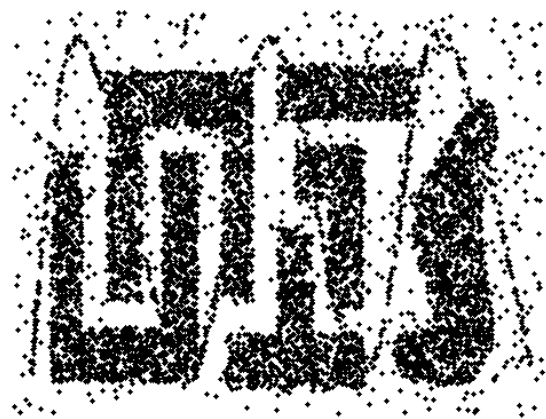
- 统计学中常用

- 二者抽样动机不同

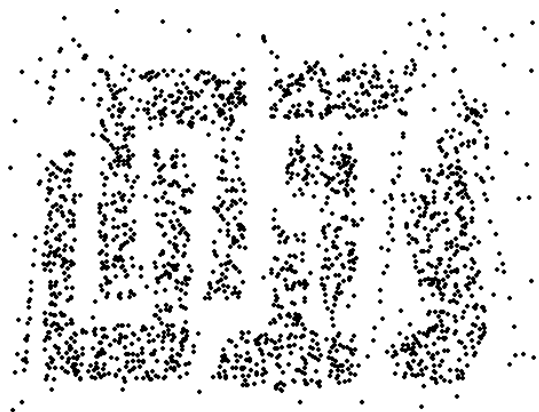
- 统计学：得到感兴趣的整个数据集费用太高、太费时间

- 数据挖掘：处理所有数据的费用太高、太费时间

# 抽样与信息损失



8000 points



2000 Points



500 Points



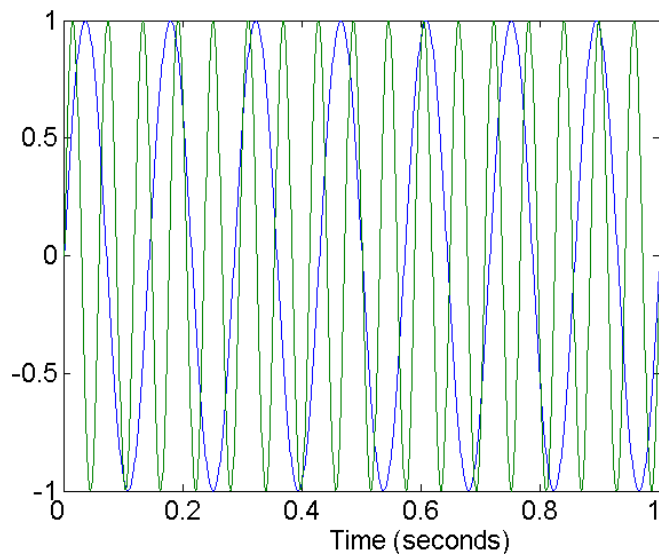
# 数据预处理

## ■ 特征创建

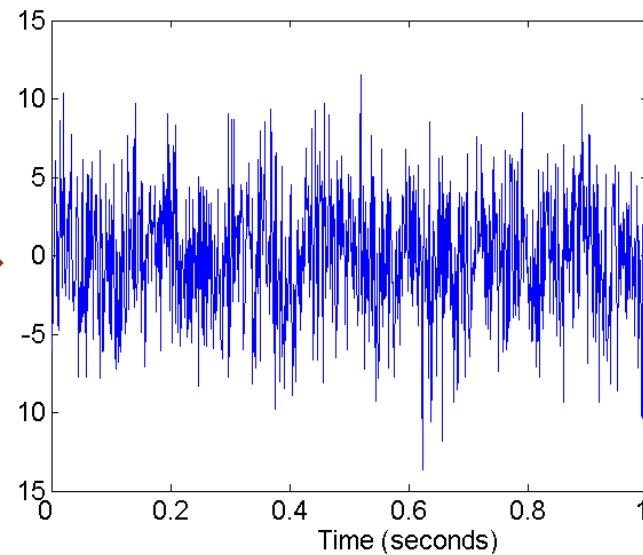
- 由原来的属性创建新的属性集，更有效的捕获数据集中的重要信息
- 常用方法
  - 特征提取 ( Feature Extraction )
    - 针对具体领域，如图像处理
  - 特征构造 ( Feature Construction )
    - 常用专家意见构造特征
  - 映射到新的空间
    - 如傅立叶变换检测时间序列数据的周期模式



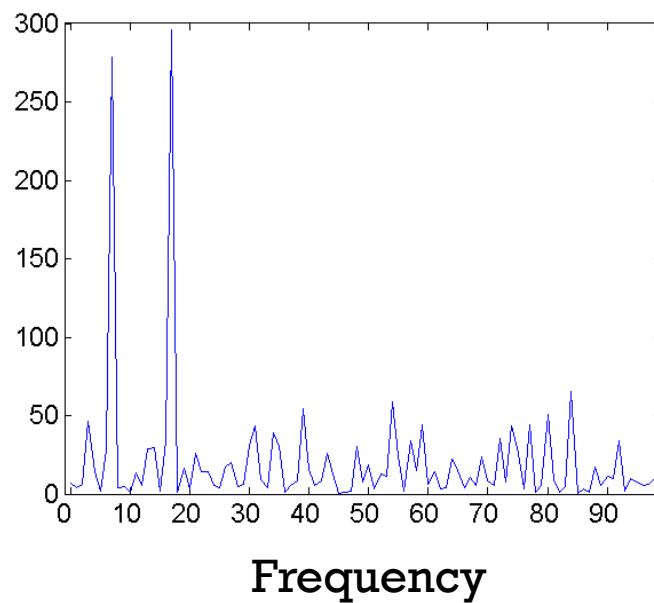
## ■ 映射到新的空间



Two Sine Waves



Two Sine Waves + Noise



# 数据预处理

- 离散化和二元化
  - 将连续属性变换成分类属性（离散化）
  - 离散和连续属性可能都需要变换成一个或多个二元属性
  - 二元化方法
    - 如有 $m$ 个分类值，将每个原始值唯一地赋予区间 $[0, m-1]$ 中的一个整数；如属性是有序的，则赋值必须保持序关系；将这 $m$ 个整数的每一个都变成一个二进制数
    - 需要 $n = \lceil \log_2 m \rceil$ 个二进制位表示这些整数，因此需要 $n$ 个二元属性表示这些二进制数





# 数据预处理

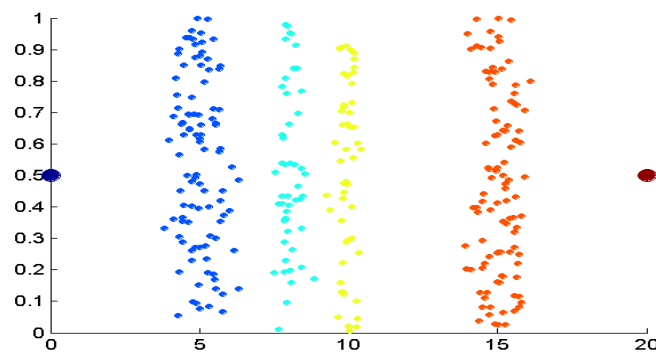
- 例，具有五个值 {awful, poor, ok, good, great} 的分类变量需要三个二元变量
- One-hot

分类值	整数值	$x_1$	$x_2$	$x_3$
<i>awful</i>	0	0	0	0
<i>poor</i>	1	0	0	1
<i>OK</i>	2	0	1	0
<i>good</i>	3	0	1	1
<i>great</i>	4	1	0	0

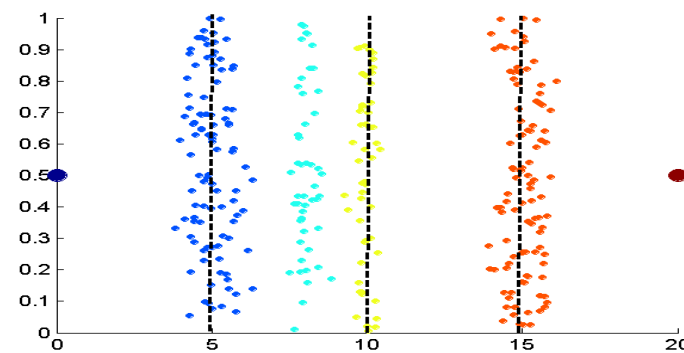


# 数据预处理

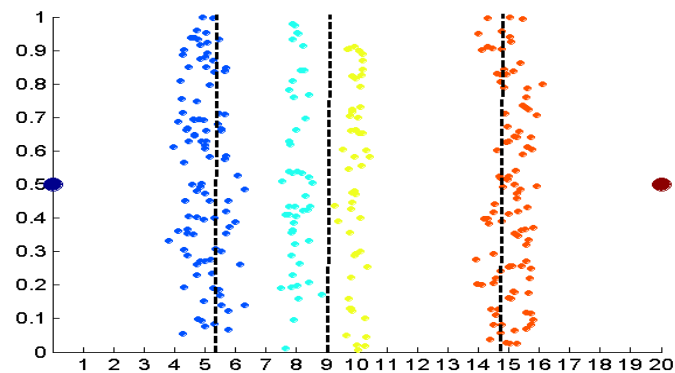
## ■ 离散化方法（非监督）



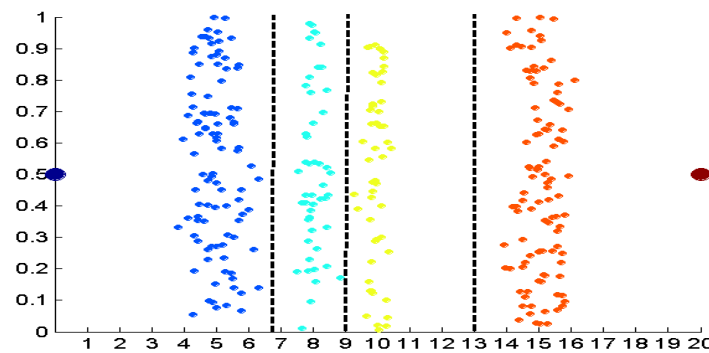
原始数据



等宽离散化



等频率离散化



K-means离散化



# 数据预处理

- 离散化方法（监督）
  - 通常能够产生更好的结果
  - 基于熵的方法
    - 将初始值切分成两部分（候选分割点可以是每个值），让两个结果区间产生最小的熵
    - 取具有最大熵的区间继续分割
    - 重复此分割过程，直到满足终止条件



# 数据预处理

设  $k$  是不同的类标号数,  $m_i$  是某划分的第  $i$  个区间中值的个数, 而  $m_{ij}$  是区间  $i$  中类  $j$  的值的个数。第  $i$  个区间的熵  $e_i$  由如下等式给出

$$e_i = - \sum_{j=1}^k p_{ij} \log_2 p_{ij}$$

其中,  $p_{ij} = m_{ij}/m_i$  是第  $i$  个区间中类  $j$  的概率 (值的比例)。

该划分的总熵  $e$  是每个区间的熵的加权平均, 即

$$e = \sum_{i=1}^n w_i e_i$$

其中,  $m$  是值的个数,  $w_i = m_i/m$  是第  $i$  个区间的值的比例, 而  $n$  是区间个数。



# 数据预处理

- 变量变换
  - 用于变量的所有值的变换
  - 简单函数
    - $x^k, \log(x), e^x, |x|, \sqrt{x}, 1/x, \sin x$
  - 标准化 ( Standardization ) 或 规范化 ( Normalization ) 或 归一化
    - 利用均值和标准差
    - 有时用中位数取代均值，用绝对标准差取代标准差
    - 目的：保持数据分布稳定（如神经网络）；排除数据测度的影响
    - 野点



54

特征

The Feature

数 字 图 像

文 本

音 频

时 序 数 据 等



# § 1.1 从图像到图像处理

---

## ■ 图像和数字图像

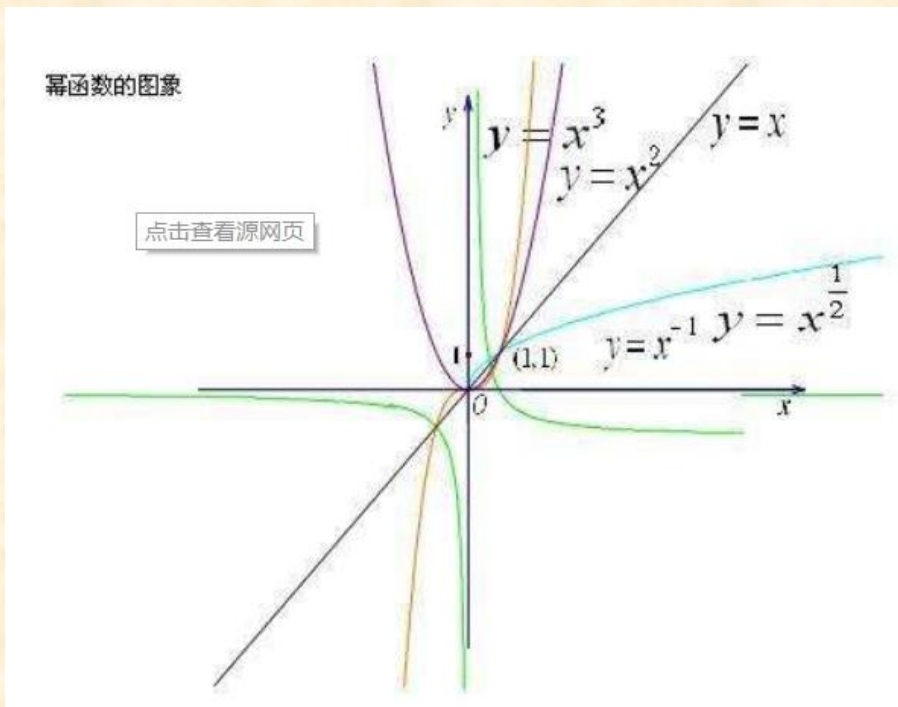
### ■ 图象：

- 用各种观测系统以不同形式和手段观测客观世界而获得的，可以直接或间接作用于人眼并进而产生视觉的实体（**照片**）
- 人类从外界（客观世界）获得的信息约有75%来自视觉系统
- 图象（广义/抽象） $\supset$  图像（狭义/具体）



# § 1.1 从图像到图像处理

## ■ 图像和图象



仅图象



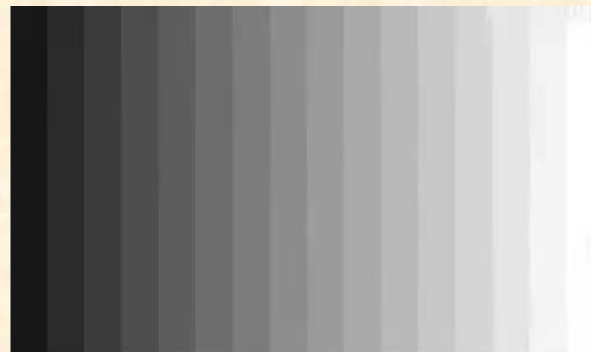
图像或图象

## § 1.1 从图像到图像处理

■ 数字图像概念：由像素组成的矩阵

- 像素：矩阵中的一个元素，像素的值代表该位置的亮度，被称为“灰度值”，灰度值越高这个位置越亮（接近白色），反之越暗（接近黑色）
- 1080P：图像的短边有1080个像素。常见的手机屏幕包含1920\*1080个像素

0	1	2	4	8	16	32	128
1	2	8	16	32	64	128	192
2	8	8	32	32	100	150	208
4	16	32	32	100	150	175	240
8	32	32	100	128	175	200	248
16	64	100	150	175	200	224	252
32	128	150	175	200	224	250	254
128	192	208	240	248	252	254	255



# § 1.1 从图像到图像处理

## ■ 数字图像的数学表示: $f(x,y,\lambda,t)$

■  $x, y$ : 2-D空间中坐标点的**位置**

■  $\lambda$ : 光线的**波长**

■  $t$ : **时间**

□ 其中最基本的形式是静态图像，视觉上就是黑白深浅的区别，表示为 $f(x,y)$ ，称为“**灰度图**”

□ 加入波长，视觉上体现为颜色信息，表示为 $f(x,y,\lambda)$ ，称为“**多光谱图**”，若仅考虑3个可见光波段，就是我们常见的彩图

□ 进一步加入**时间信息**，视觉上体现为“**视频**”



## § 1.1 从图像到图像处理

- 对于静态的彩色图像

- 用红绿蓝三基色表示，图像是一个3个2-D数组

$$f_R(x, y), f_G(x, y), f_B(x, y)$$

- 如果是多光谱图像，图像是一个n个2-D数组

- 对于动态的连续图像，可以是一个时间序列

- 黑白电视信号

$$f_1(x, y), f_2(x, y), \dots, f_k(x, y)$$

- 彩色电视信号

$$f_{Rk}(x, y), f_{Gk}(x, y), f_{Bk}(x, y), k=1, 2, \dots$$



# § 1.1 从图像到图像处理

## ■ 图像处理

### ■ 图像处理

□ 图像处理在广义上是各种与图像有关的技术的总称

□ 主要功能/作用包括：

- 对图像的各种加工（见下）
- 基于加工结果的判断决策和行为规划
- 为此进行的硬件设计及制作

### ■ 图像加工技术

□ 例如：图像的采集、获取、编码、存储和传输，合成和产生，显示和输出，变换、增强、恢复和重建，分割，目标的检测、表达和描述，特征的提取和测量，序列图象的校正，**3-D**景物的重建复原，图像数据库的建立、索引和抽取，图像的分类、表示和识别，图像模型的建立和匹配，图像和场景的解释和理解， .....

# § 1.1 从图像到图像处理

---

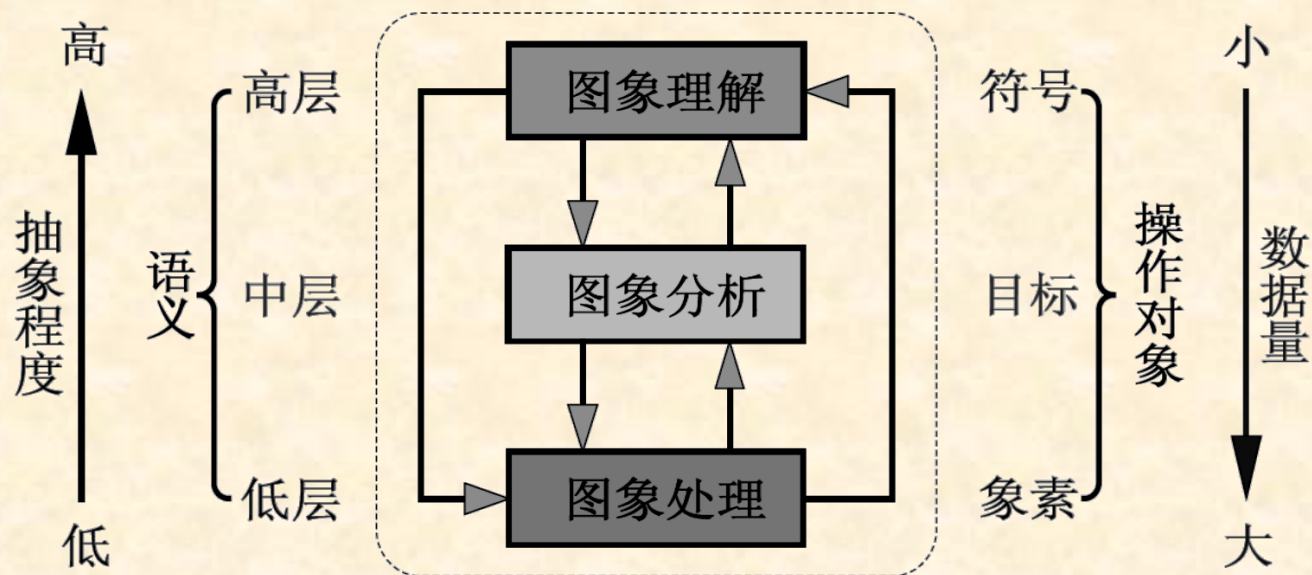
## ■ 图像处理应用领域示例

- 视频通信：可视电话，电视会议，按需电视，远程教育；
- 文字档案：文字识别，过期档案复原，邮件分检，支票，签名辨伪，办公自动化；
- 生物医学：红白学球计数，染色体分析、**X光**、**CT**、**MRI**、**PET**图象分析，医学手术模拟规划，远程医疗；
- 遥感测绘：巡航导弹制导，无人驾驶飞机飞行，精确制导，航天侦查，资源探测，气象预报，自然灾害监测；
- 工业生产：工业检测，工业探伤，自动生产流水线监控，移动机器人，无损探测，金相分析，印刷板质量检验，精细印刷品缺陷检测；
- 军事公安：雷达图像分析、巡航导弹路径规划 / 制导，罪犯脸形合成、识别，指纹、印章的鉴定识别；
- 交通管理：太空探测、航天飞行、公路交通管理。

# § 1.1 从图像到图像处理

## ■ 图像处理三层次：

- 图像预处理（图像——>图像）
- 图像分析（图像——>数据）
- 图像理解（图像——>解释）



# § 1.1 从图像到图像处理

---

■ 图像处理三层次：

□ 图像预处理（狭义图像处理）





# § 1.1 从图像到图像处理

---

## ■ 图像处理三层次：

### □ 图像分析（机器学习）



帽子

人脸

长发

# § 1.1 从图像到图像处理

## ■ 图像处理三层次：

### □ 图像理解（机器学习）



帽子：时尚

人脸：容光焕发

长发：女性标志

图像理解：炯炯有神的美女模特

## § 1.2 图像处理的基本操作

### ■ 图像增强

- 改善图像视觉效果，增强图像的有效信息，消弱噪声的干扰



去雾



去模糊



## § 1.2 图像处理的基本操作

### ■ 图像恢复与重建

■ 恢复图像原本面貌，辐射校正、几何校正



几何校正



## § 1.2 图像处理的基本操作

---

### ■ 图像编码

■ 压缩数据，有效传输，节省空间



450KB



89KB

- 特征形成与计算

- 根据应用领域相关知识决定采用哪些特征，称为原始特征
- 例如细胞图像大小256 x 256，如果全部采用的话，原始特征即为65536维
- 如果改为计算细胞的面积、周长、形状、纹理、核浆比，则特征维数变为5维



- 特征提取的原因

- 机器学习系统的成败，首先取决于所采用的特征是否较好的反映模式的特性以及模式的分类问题
- 原始特征依赖于具体应用问题和相关专业知识的（文字识别和图像识别）
- 希望在保证分类效果前提下，采用尽可能少的特征完成分类



- 原始特征的问题

- 有很多特征可能与要解决分类问题关系不大，但却在后续分类器设计中影响分类器性能
- 即使很多特征与分类问题关系密切，但特征过多导致计算量大、推广能力差。当样本数有限时容易出现病态矩阵等问题

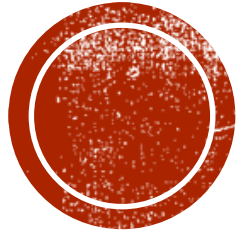




# 数据预处理中的特征提取

- 维归约（特征提取）
  - 维度较低时，许多数据挖掘算法的效果会更好
    - 删除不相关的特征，降低噪声，避免维数灾难
    - 降低了算法的时间和内存需求
    - 模型更易理解，容易让数据可视化
  - 维归约常用方法
    - 主成分分析（PCA）
    - 线性判别分析（LDA）





**THE END !**

