

# 数据挖掘与机器学习

潘斌

panbin@nankai.edu.cn

范孙楼227

1

# 课程安排

- 讲授+实验
- 1-17周，每周讲授3课时
- 第5周起安排上机课程，共5次实验+1次课程大作业
- 成绩=闭卷考试（50）+实验（20）+大作业（25）+考勤（5）

# 实验安排

- 利用本课程学习的算法编程完成5项数据挖掘/机器学习任务（待续），并提交实验报告1份（5.16/5.23）
- 5-13周单周安排上机课，助教答疑
- 14、15周实验课随堂考核，针对报告提问

# 课程大作业

- 实现一项 **全链路** 机器学习任务（如目标检测、文字识别等）
- 做 **课堂PPT展示**，汇报研究思路、算法设计、结果，20分钟汇报+5分钟提问，暂定**15-17周**
- 分组完成，4-6人一组，1人为组长，汇报时说明每个人的贡献

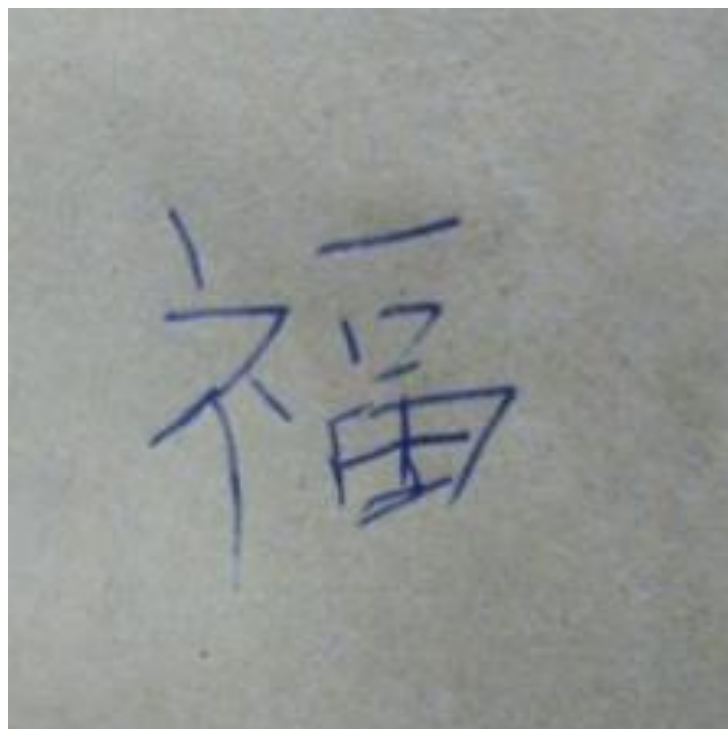
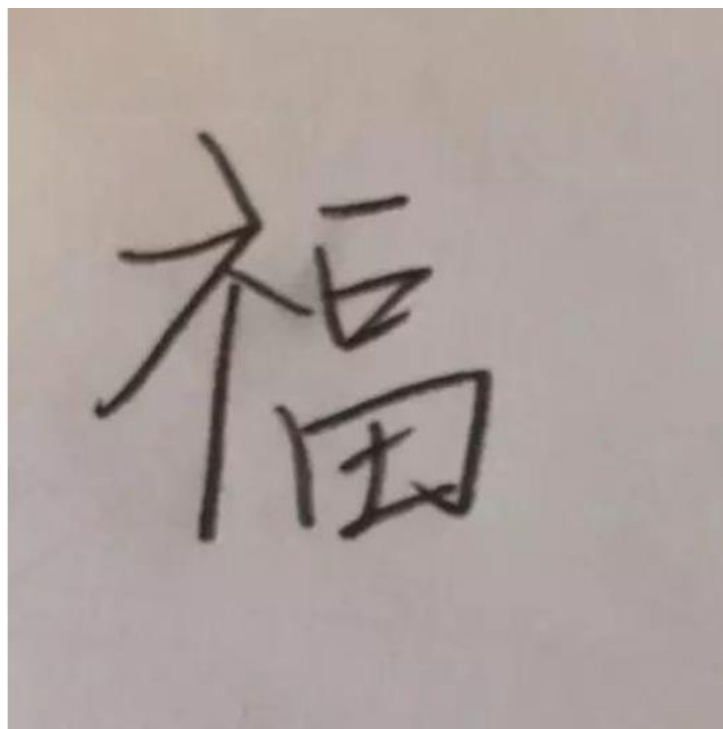
# 课程大作业（四选一）

## ■ 任务1：遥感图像飞机检测



# 课程大作业（四选一）

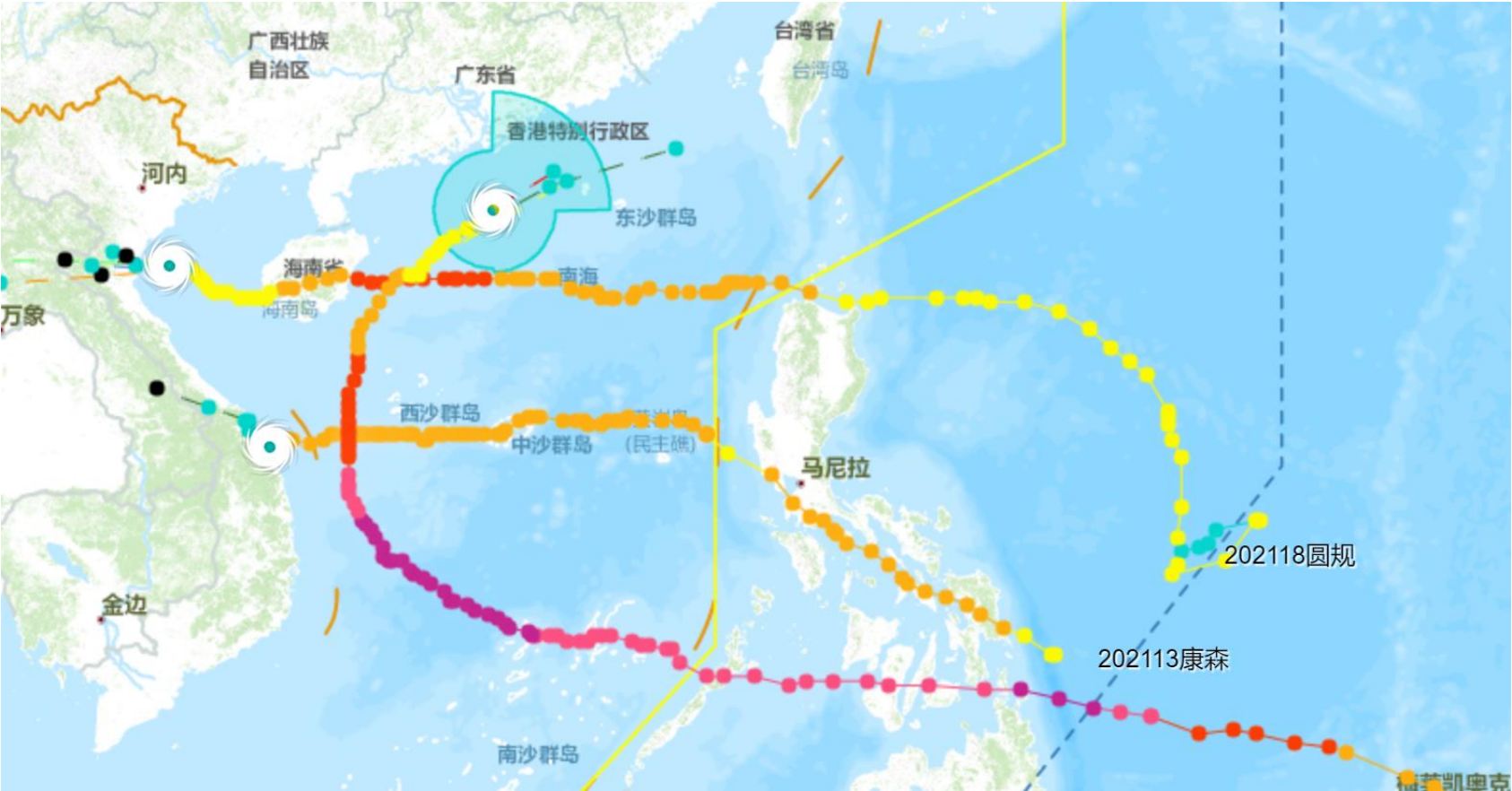
## ■任务2：“福”字识别-解决类别不平衡问题





# 课程大作业（四选一）

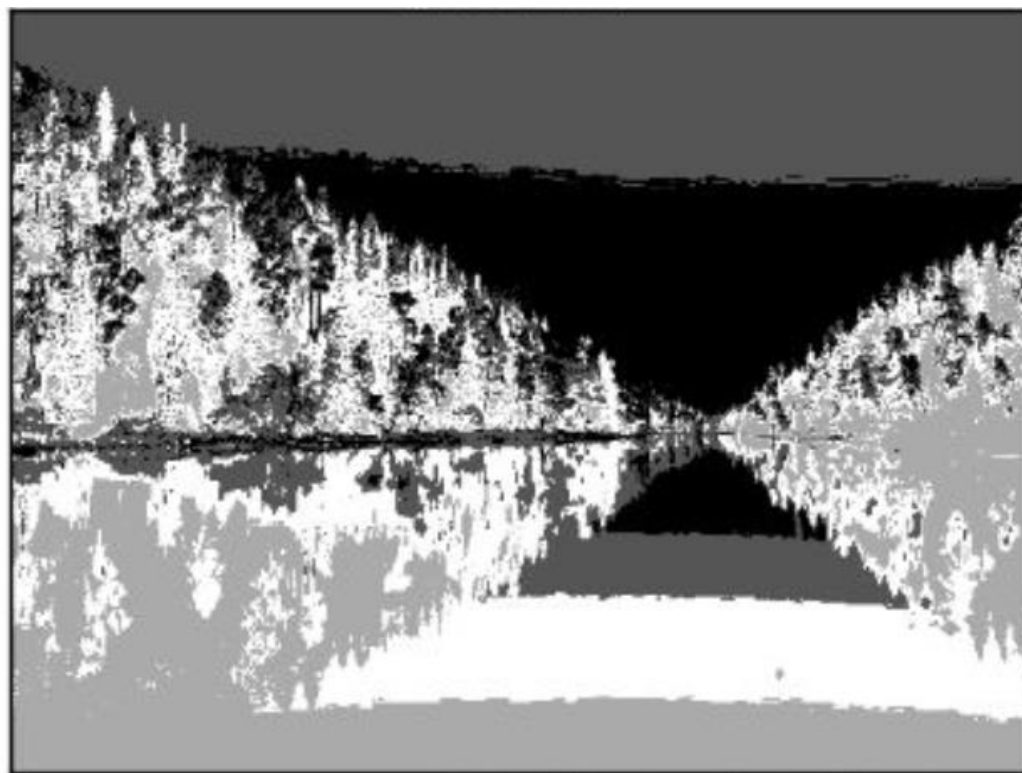
## ■任务3：台风预报



66666	1815	21	0018	1815	0	6	LEETPI
2018081012	1	174	1448	1006		13	
2018081018	1	179	1444	1004		15	
2018081100	1	185	1441	1004		15	
2018081106	1	192	1438	1004		15	
2018081112	2	200	1435	1000		18	
2018081118	2	208	1432	995		20	
2018081200	2	217	1427	990		23	
2018081206	2	227	1420	990		23	
2018081212	3	235	1415	982		28	
2018081218	3	245	1407	982		28	
2018081300	3	253	1398	982		28	

# 课程大作业（四选一）

- 任务4：图像区域分割提取-如何保持空间相关性？





# 课程大作业

- 无需实验报告，只PPT展示即可
- 上述任务随机4选1，优先遵照意愿
- 第11周前完成分组（4.27），可在飞书群里自行组队

■ 助教：潘奕安，微信：pan-yi-an

# 参 考 文 献

- 李航，统计学习方法，清华大学出版社
- 周志华，机器学习，清华大学出版社
- 张学工，模式识别，清华大学出版社
- 其他文献.....
- 网络资源.....



# 绪论

Overview of Data Mining & Machine Learning

- Intel公司创始人之一戈登·摩尔（Gordon Moore）发现芯片的容量每18~24个月增加一倍。他据此推断，按此趋势发展下去，在较短时间内计算能力将呈指数增长。——“摩尔定律”（Moore's Law）。
- John Roth在联合国世界电信论坛上又提出了一个关于网络科技的观点：互联网宽带每9个月会增加一倍的容量，但成本降低一半，比芯片的变革速度还快。——“光纤定律”（Optical Law）。



- 1KB ( Kibibyte, 千字节 ) = 1024B,
- 1MB ( Mebibyte, 兆字节, 简称“兆” ) = 1024KB,
- 1GB ( Gigabyte, 吉字节, 又称“千兆” ) = 1024MB,
- 1TB ( Terabyte, 万亿字节, 太字节 ) = 1024GB,
- 1PB ( Petabyte, 千万亿字节, 拍字节 ) = 1024TB,
- 1EB ( Exabyte, 百亿亿字节, 艾字节 ) = 1024PB,
- 1ZB ( Zettabyte, 十万亿亿字节, 泽字节 ) = 1024EB,
- 1YB ( Yottabyte, 一亿亿亿字节, 尧字节 ) = 1024ZB,
- 1BB ( Brontobyte, 一千亿亿亿字节 ) = 1024YB

# 天文学



- 2000年斯隆数字巡天（Sloan Digital Sky Survey）项目启动的时候，位于新墨西哥州的望远镜在短短几周内收集到的数据，已经比天文学历史上总共收集的数据还要多。到了2010年，信息档案已经高达 $1.4 \times 2^{42}$ 字节（200PB）。

# 沃尔玛

- 沃尔玛每隔**1小时**就要处理超过100万客户的交易,信息录入量超过**2,5PB**;



# 互 联 网

- 根据 Visual Capitalist 统计，2019年平均每分钟 Facebook 有100万账户登录，Youtube 有450万次视频播放，谷歌有380万次搜索，以及100万美元的线上消费！
- 短视频、直播、自媒体等新一代互联网数据近年来迎来了大爆发。抖音2016年初创，2018年月活1亿，2019年5亿，2020年8亿，接近微信。



# 2018 *This Is What Happens In An Internet Minute*

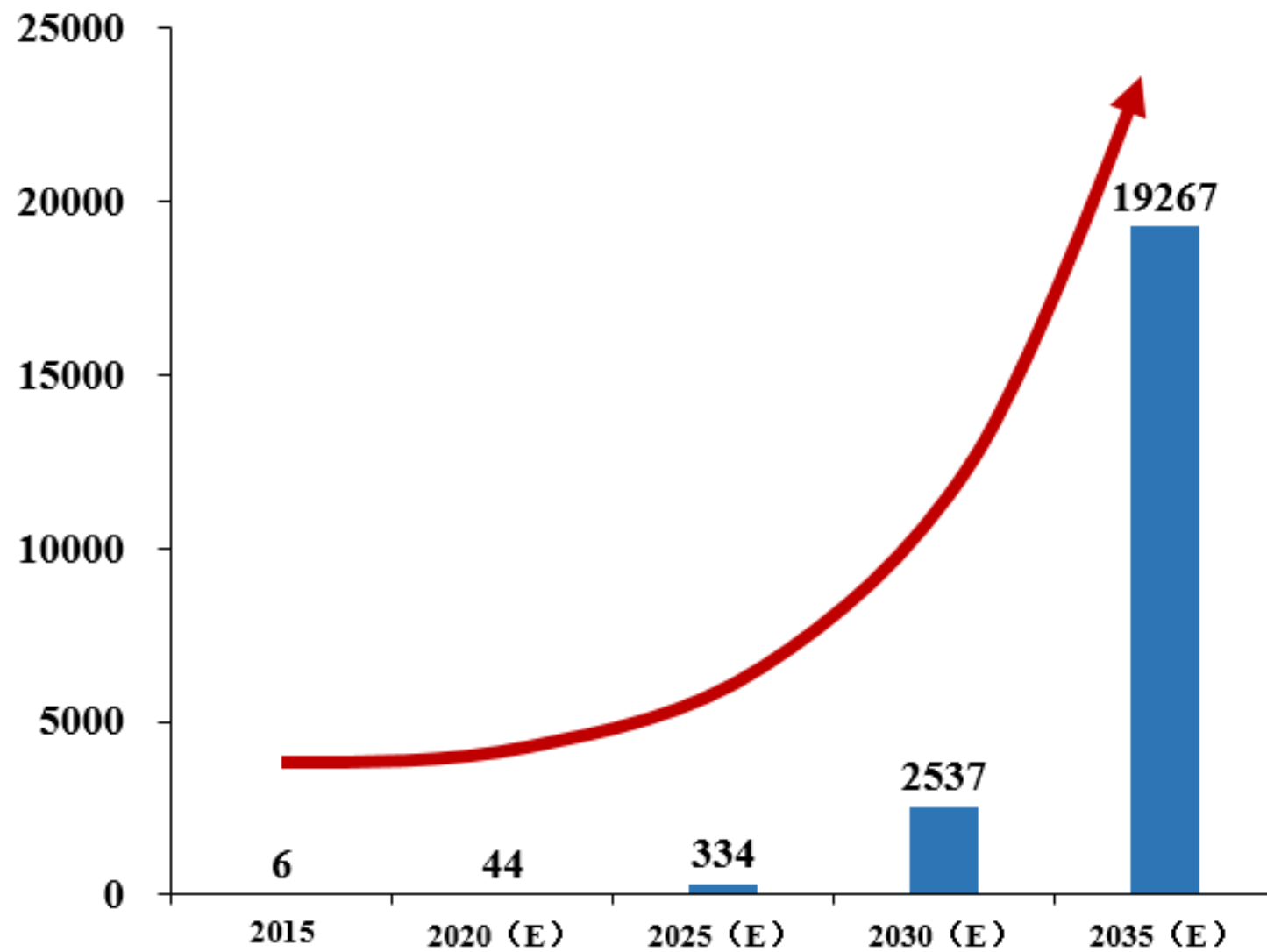


# 2019 *This Is What Happens In An Internet Minute*





■ 全球数据总量 ( ZB )



数据来源：IDC，中国电子学会整理

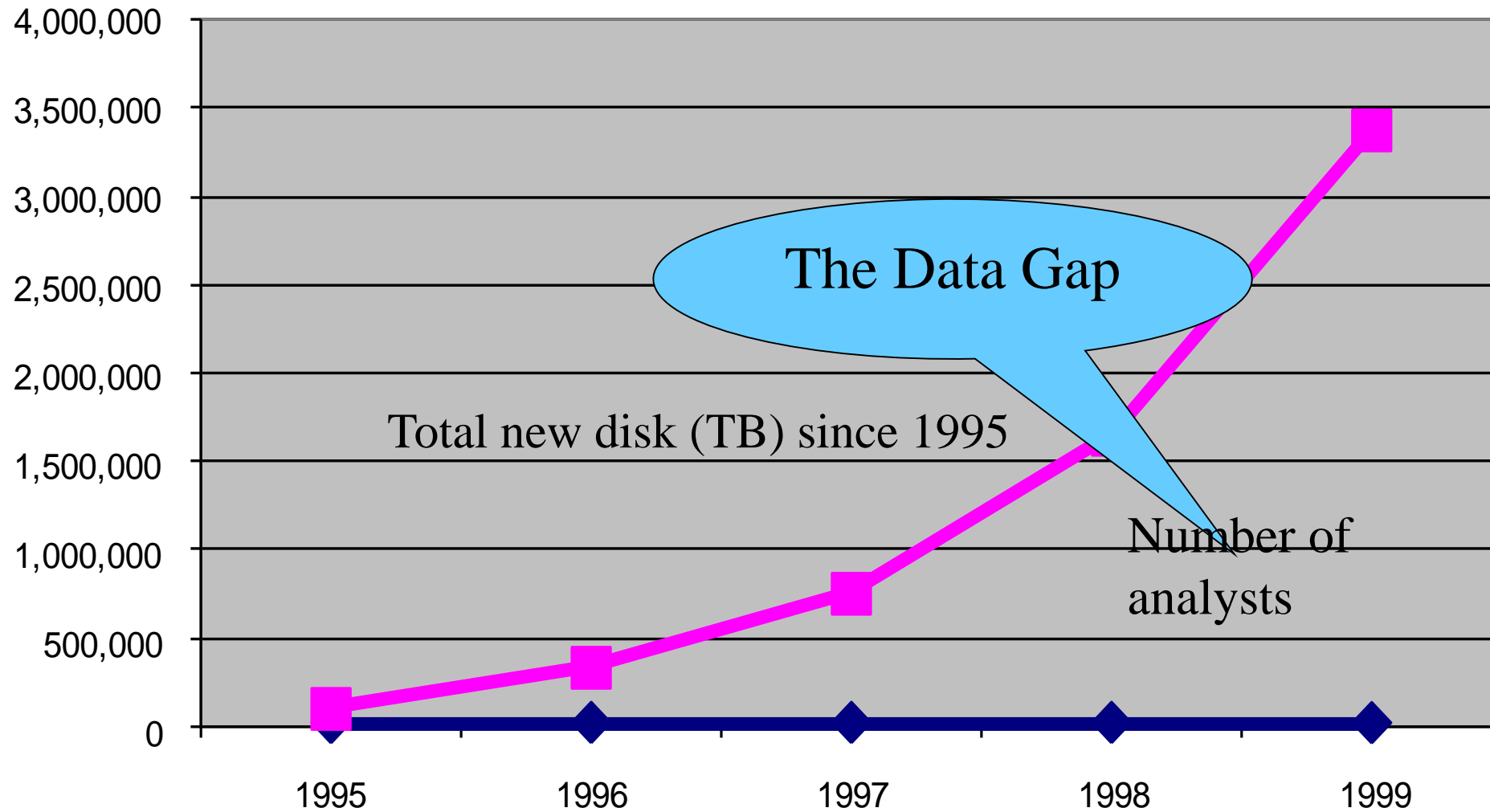
[http://www.sohu.com/a/252632116\\_179557](http://www.sohu.com/a/252632116_179557)

- 我们已经进入了“Big Data”时代！



# 大数据的特点（4V）

- 数据量 **太大** Volume
- 数据本身 **多样性强** Variety
- 大数据的产生非常 **高速** Velocity
- **价值低** Value——历史数据变成“数据坟墓”



***“We’re drowning in information, but starving for knowledge.”***

***- John Naisbett***



# 新数据的特点及对分析方法的要求（1）

- 海量数据集越来越普遍

- 如，处理的数据可能不能放入内存
- 算法必须是可伸缩的（scalable）

- 常遇到高维数据集

- 如，生物学领域已产生涉及数千特征的基因表达数据
- 算法需要具有高维性（稀疏）

# 新数据的特点及对分析方法的要求（2）

- 很多数据带有**异构属性**
  - 含有半结构化文本和超链接的Web页面集；描述目标属性是所使用的.doc文件和.mp4文件；包含地球表面不同位置上的时间序列测量值（温度、气压等）的气象数据
  - 算法要能处理这些非传统数据集

# 新数据的特点及对分析方法的要求（3）

- 需要分析的数据并非存放在一个站点或归属一个机构，而是**地理上**分布在属于**多个机构**的资源中（贵州大数据）
  - 需要开发分布式数据分析技术
    - 降低执行分布式计算所需的信息量
    - 有效地统一从多个资源得到的数据分析结果
    - 处理数据安全性问题

# 新数据的特点及对分析方法的要求（4）

- 待分析的数据集通常**不是精心设计的实验的结果**
  - 通常代表数据的时机性样本（opportunistic sample）  
而非随机性样本（random sample）
- 数据集常涉及非传统的数据类型和数据分布
- 数据分析的任务需要产生和评估数千种假设

- **数据挖掘**（Data Mining）是在大型数据存储库中，自动的发现**有用信息**的过程。
- 发现先前未知的有用模式（**描述**任务）
- 预测未来观测结果（**预测**任务）



- 数据挖掘：目的

- 机器学习：手段

- 大数据：原料

- 大数据技术？

大数据 vs 数据挖掘



- 数据：大量的、不完全的、有噪声的、模糊的、随机的
- 发现：隐含在其中的、人们事先不知道的、但有潜在的有用信息和知识

# 数据挖掘过程

- 数据采集
- 数据预处理
- 数据挖掘
- 评价和呈现

# 描述任务

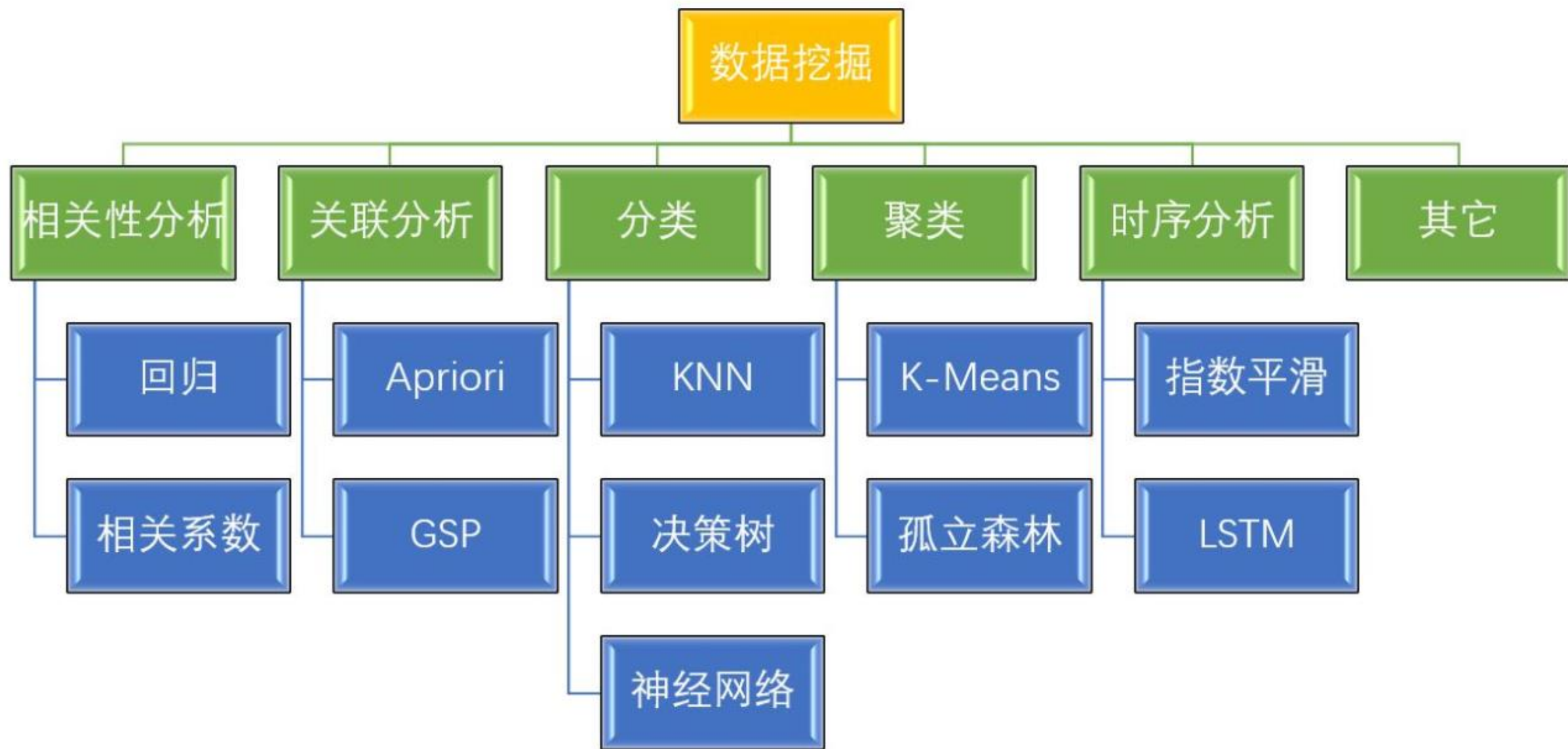
- 导出概括数据中潜在联系的模式（相关、趋势、聚集、轨迹、异常等）
- 本质上是探查性的
- 常需要后处理技术验证和解释结果

# 预测任务

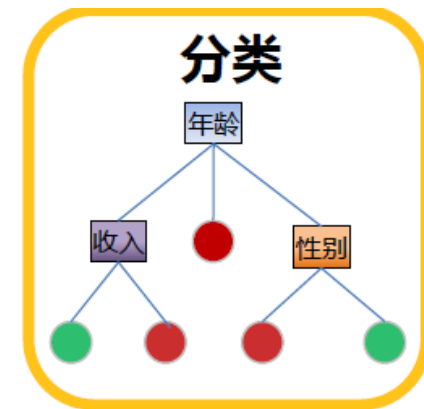
- 根据其他属性的值，预测特定属性的值
- 分类
  - 用于预测离散的目标变量
- 回归
  - 用于预测连续的目标变量

# 主要数据挖掘任务

- 分类（ Classification ） [Predictive]
- 聚类（ Clustering ） [Descriptive]
- 关联规则（ Association Rule Discovery ） [Descriptive]
- 回归（ Regression ） [Predictive]
- 序列模式识别（ Sequential Pattern Discovery ） [Descriptive]
- 异常检测（ Anomaly Detection ） [Predictive]



# 分 类



- 给定一个记录集合
  - 每条记录（样本）由一些属性（特征+标签）组成，其中一个属性为类别。
  - $(x_1, x_2, \dots, x_n, c)$
  - 找到一个将类别属性表示为其他属性的函数的模型（如  $c = f(x)$ ）。
  - 目标：未见过的记录尽可能准确地被分类。



# 分类步骤

- 给定的数据集被分成独立的训练集和验证集，训练集用于建立模型，而验证集用于检验该模型。
- 模型创建：对一个类别已经确定的数据创建模型。
- 模型检验与应用：对创建的模型进行评价，用其预测未来或者类别未知的记录。

- [例] 某鱼类加工厂生产线混合生产鲑鱼、鲈鱼两种鱼类罐头，需要对原材料鲑鱼、鲈鱼进行人工分拣（分类识别）。
- 若用一机器学习系统来进行自动分拣，采用什么样的识别方法和识别系统？

# 问题和要求

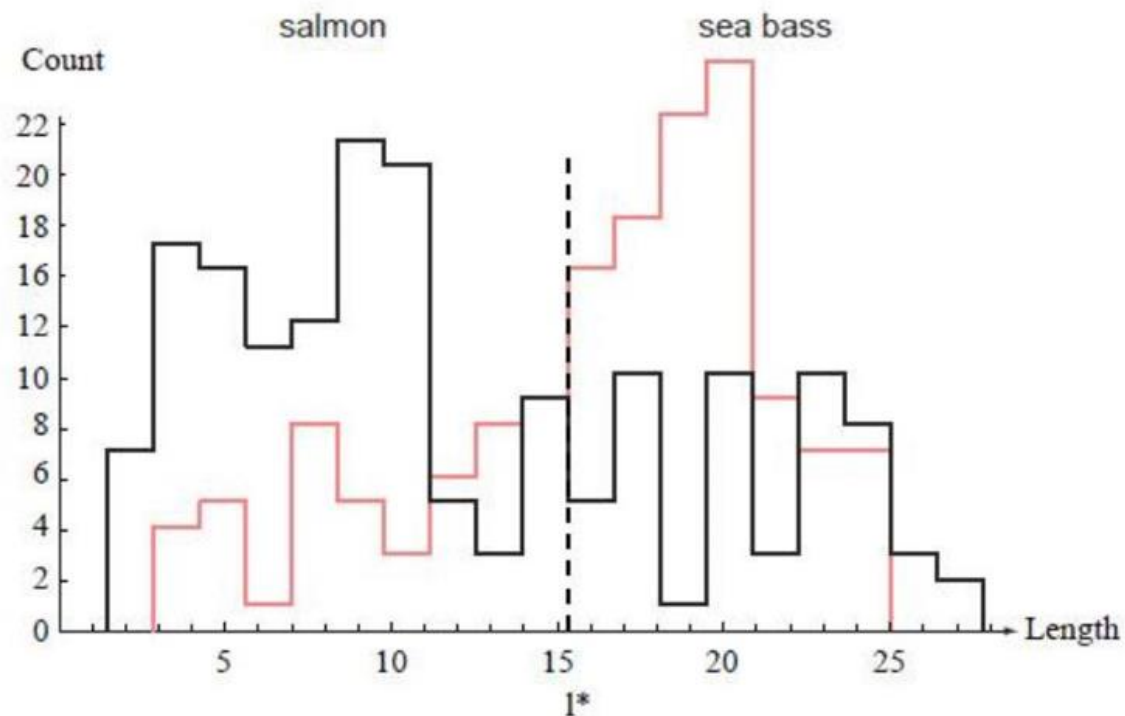


# 机器学习系统概述

- 水产工人目视识别的参考依据
  - 长度差异 (鲑鱼相对鲈鱼短一些)
  - 宽度差异 (鲑鱼相对鲈鱼窄一些)
  - 亮度差异 (鲑鱼相对鲈鱼暗一些)
  - 鳞片形状
  - 嘴的形状
  - 等等

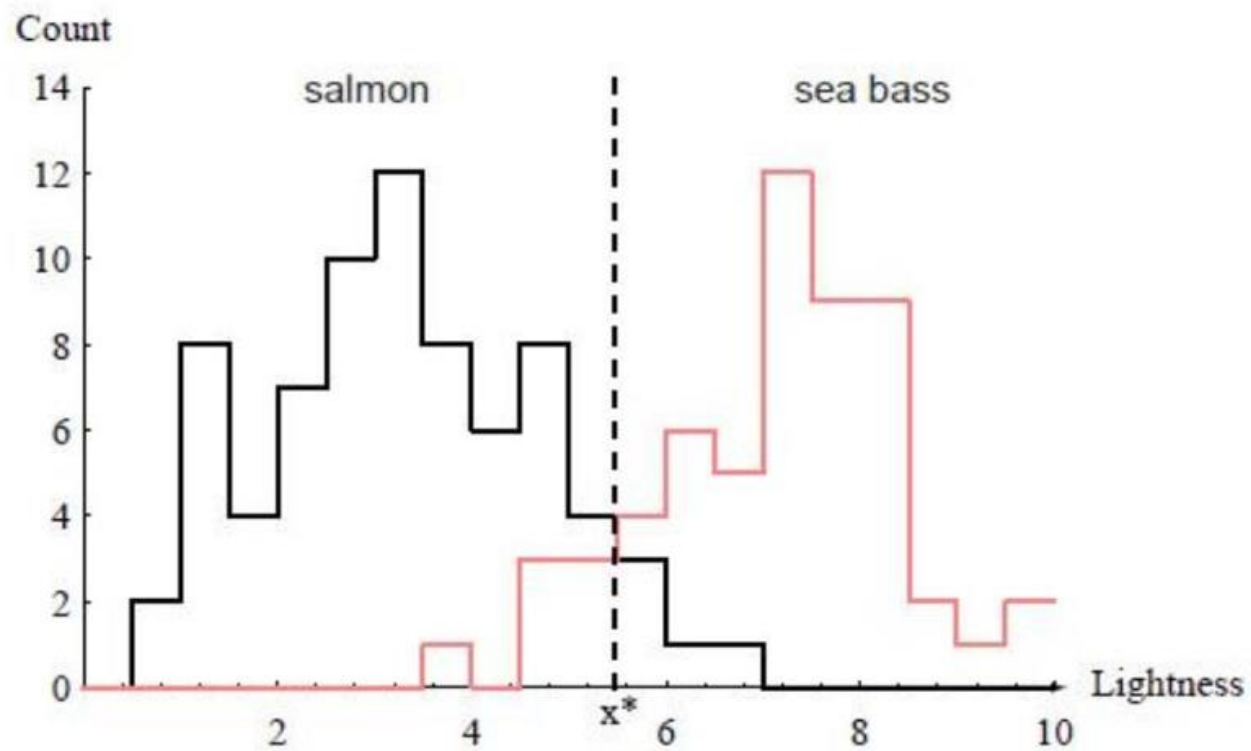
# 机器学习系统概述

- 长度差异（样本统计图）



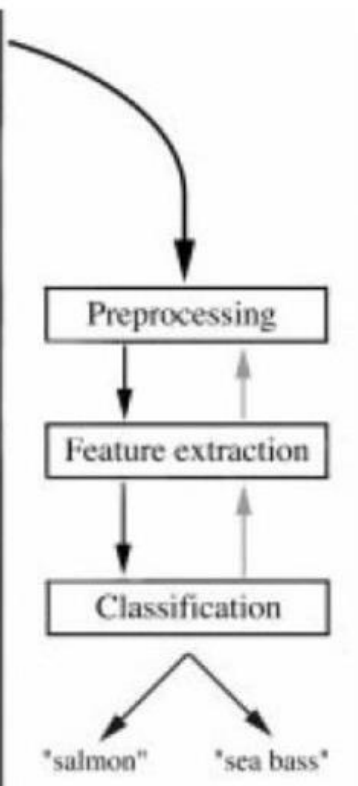
# 机器学习系统概述

- 亮度差异（样本统计图）



# 机器学习系统概述

- 自动识别系统的主要组成部分
  - 数据源：通过相机拍摄获取数字图像
  - 预处理：采用图像处理算法提取鱼的区域
  - 特征计算：采用图像处理算法计算鱼的长度、宽度等
  - 分类：利用已知的两种鱼的长度、亮度等差异进行判别





# 应用问题抽象化

- 类别抽象化

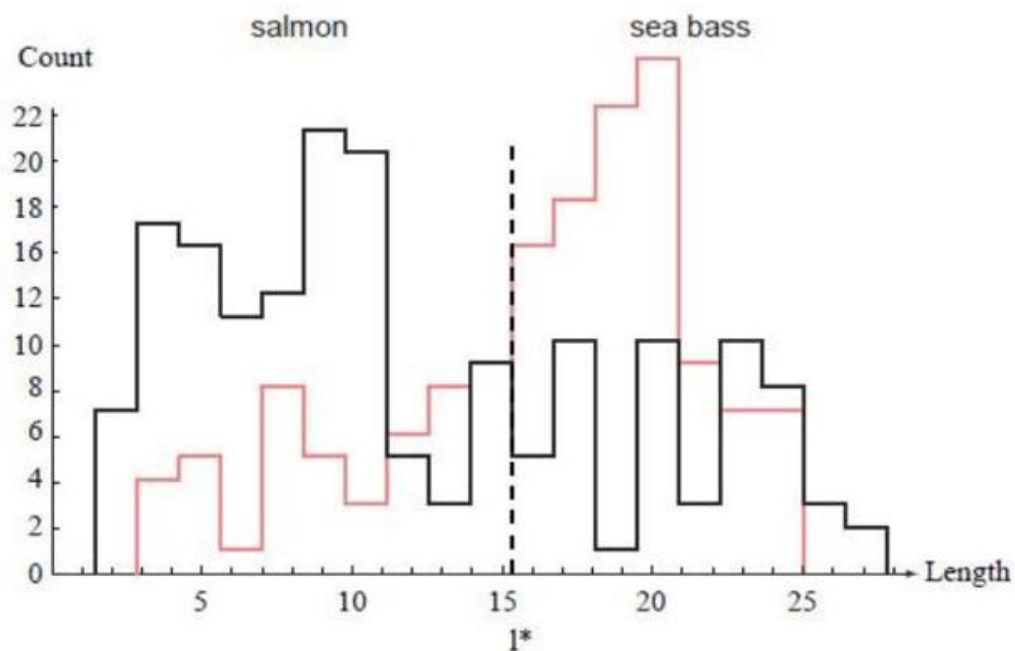
- 鲑鱼—— $\omega_1$ 类
- 鲈鱼—— $\omega_2$ 类

- 分类依据抽象化

- 若采用长度来区分，则定义长度为特征，记为x

# 应用问题抽象化

- 分类依据抽象化
  - 已知目视识别经验
    - 鲑鱼相对较短—— $\omega_1$ 类x值偏小
    - 鲈鱼相对较长—— $\omega_2$ 类x值偏大



# 应用问题抽象化

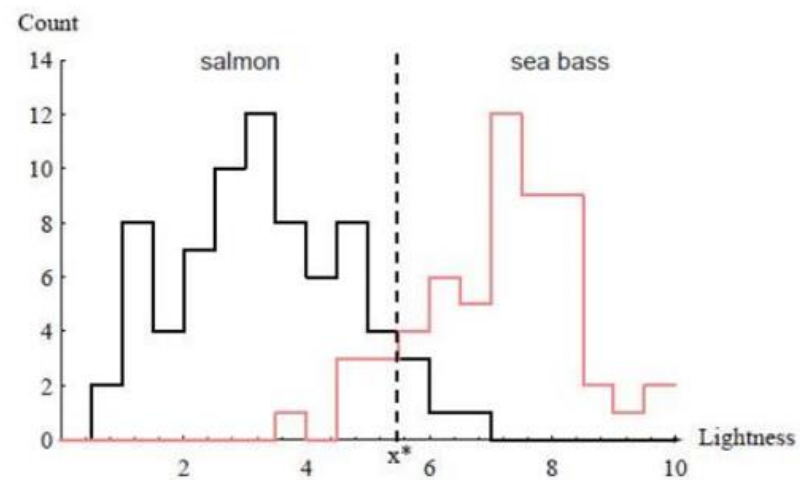
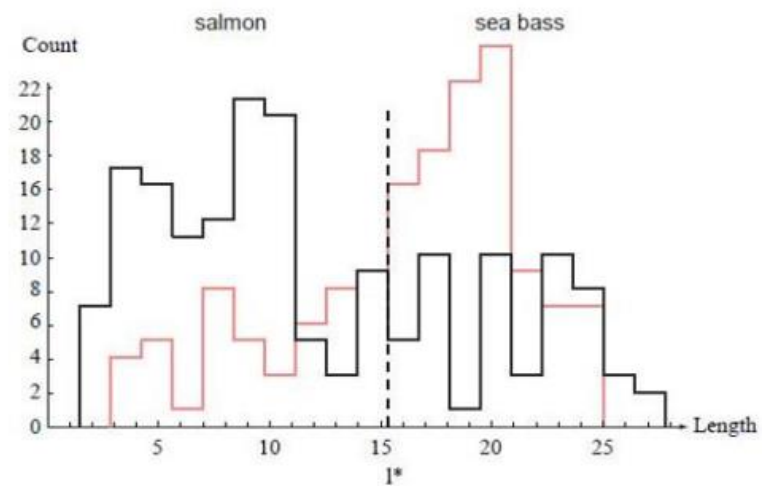
- 分类方法抽象化

- 定义长度特征 $x$ ，统计对象的亮度分布，求出临界点 $x_0$ （阈值）
- 从而确定分类的过程
  - 如果 $x < x_0$ ，则 $x \in \omega_1$ 类（鲑鱼）
  - 如果 $x > x_0$ ，则 $x \in \omega_2$ 类（鲈鱼）
- 上述过程称为决策规则

# 应用问题抽象化

- 特征需要优选
  - 采用长度特征
  - 采用亮度特征
  - 尽量多的选用特征（通常可以提高识别率）
  - 尽量少的选用特征（减少设计和计算难度）

# 应用问题抽象化



# 应用问题抽象化

- 优选结果案例
  - 若采用亮度特征 $x_1$ 和宽度特征 $x_2$
  - 从而确定分类的过程
    - 如果....., 则 $x \in \omega_1$ 类
    - 如果....., 则 $x \in \omega_2$ 类

# 示例

## ■ 待分析数据集

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



## ■ 训练集

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

## ■ 验证集

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	F
8	A	7	T

## ■ 测试集

## ■ 模型创建

分类算法

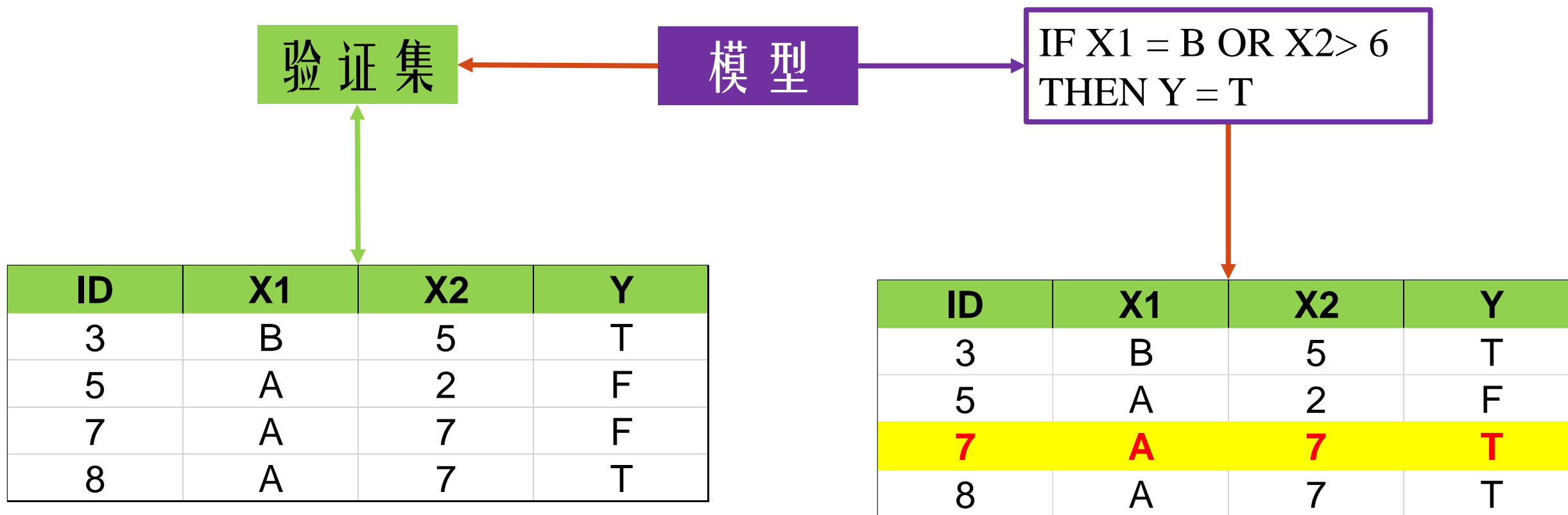
训练集

模型

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

IF X1 = B OR X2 > 6  
THEN Y = T

## ■ 模型检验 (参数、超参数)



## ■ 模型应用



■ 不能在测试集调参！

# 常用分类方法

- 决策树
- KNN
- SVM
- ANN
- Bayes 分类
- .....

# 关联分析

某超市Pos机上记录如下的销售数据：

顾客	购买商品
1	面包，黄油，尿布，啤酒
2	咖啡，糖，小甜饼，鲑鱼，啤酒
3	面包，黄油，咖啡，尿布，啤酒，鸡蛋
4	面包，黄油，鲑鱼，鸡
5	鸡蛋，面包，黄油
6	鲑鱼，尿布，啤酒
7	面包，茶，糖鸡蛋
8	咖啡，糖，鸡，鸡蛋
9	面包，尿布，啤酒，盐
10	茶，鸡蛋，小甜饼，尿布，啤酒

从这个销售数据中可以得出什么结论？

# 关联分析



- 给定一系列记录，找到其中隐含的令人感兴趣的联系，用以根据记录中某些项的出现来预测其他项的产生。
- 在关联规则挖掘算法中，通常的目的是发现数据中强关联特征的模式。
- 所发现的模式通常用蕴含规则或特征子集的形式表示。
- 目标是以有效的方式提取最有趣的模式。



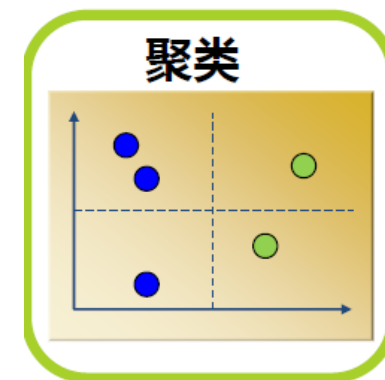
# 示例

## ■ 购物篮分析

商场购物篮事务

<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

关联规则：  
 $\{\text{Diaper}\} \rightarrow \{\text{Beer}\}$   
 $\{\text{Bread}\} \rightarrow \{\text{Milk}\}$



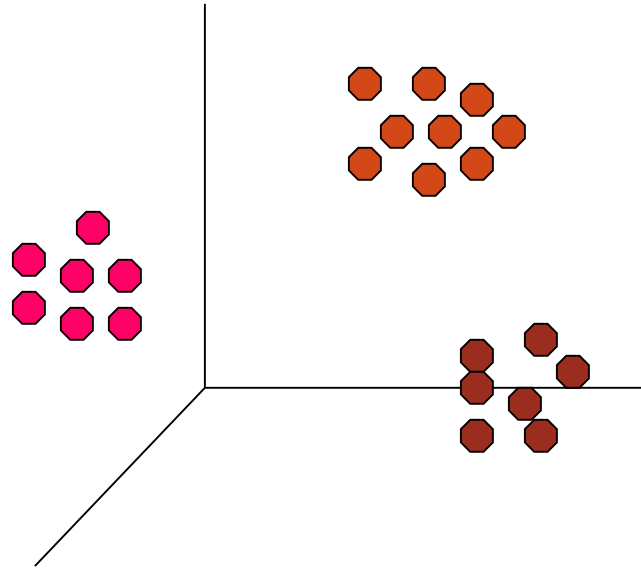
# 聚类

- 聚类是指一组彼此间非常“相似”的数据对象的集合。每一个类中的数据相近，不同类之间的数据相差较大。
- 通过聚类分析可以根据部分数据发现规律，找出对全体数据的描述，可用于异常点检测。

- Euclidean Distance Based Clustering in 3-D space.

Intracuster distances  
are minimized

Intercluster distances  
are maximized



● 特征空间构建

● 距离度量方式

# 异常检测

- 识别特征显著不同于其他数据的观测值（异常点或离群点）。
- 应用于检测信用卡欺诈、网络攻击、疾病的不同模式、生态系统扰动等。

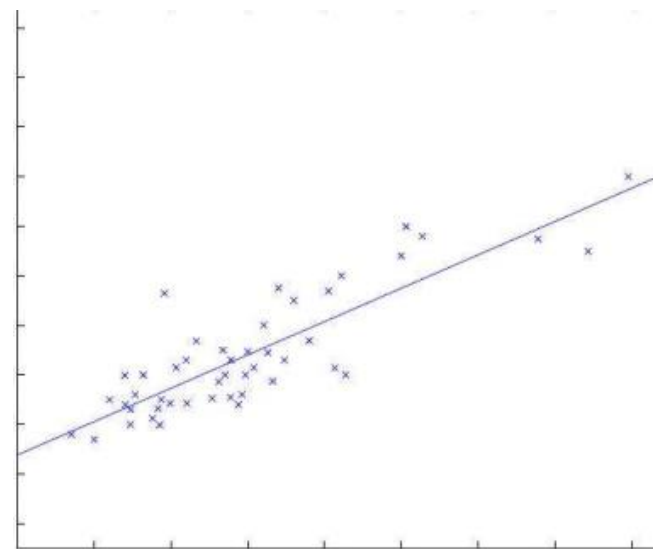
# 回 归

- 监督学习，给定输入  $\mathbf{x}$  和输出  $\mathbf{y}$ ，任务是学习从输入到输出的映射  $\mathbf{y} = \mathbf{f}(\mathbf{x} | \theta)$ 。

线性回归试图学得

$$f(x_i) = wx_i + b, \text{ 使得 } f(x_i) \simeq y_i .$$

$$\begin{aligned} (w^*, b^*) &= \arg \min_{(w, b)} \sum_{i=1}^m (f(x_i) - y_i)^2 \\ &= \arg \min_{(w, b)} \sum_{i=1}^m (y_i - wx_i - b)^2 . \end{aligned}$$



# 序列模式发现

- 主要用于分析数据仓库中的某类与时间相关的数据，搜索类似的序列或子序列，并挖掘时序模式、周期性、趋势和偏离等。
- 例如，它可以导出类似“若A股票连续上涨两天且B股票不下跌，则第三天C股票上涨的可能性为75%”的数据关系。
- 序列模式可以看成是一种特定的关联模型，它在关联模型中增加了时间属性。

# 监督学习和无监督学习

- 监督学习 ( Supervised Learning )
  - 训练集带有类标签，新的数据基于训练集进行分类。
  - 如回归、分类等
- 无监督学习 ( Unsupervised Learning )
  - 数据集没有类标签，提供一组属性，然后寻找出数据集中存在的类别或者聚集
  - 如聚类、降维等

# 监督学习和无监督学习

## Supervised vs Unsupervised

### Supervised Learning

**Data:**  $(x, y)$

$x$  is data,  $y$  is label

**Goal:** Learn a *function* to  
map  $x \rightarrow y$

**Examples:** Classification,  
regression, object detection,  
semantic segmentation, image  
captioning, etc

### Unsupervised Learning

**Data:**  $x$

Just data, no labels!

**Goal:** Learn some *structure*  
of the data

**Examples:** Clustering,  
dimensionality reduction, feature  
learning, generative models, etc.



# 数据挖掘应用

- 电子科大利用大数据寻找校园中最孤独的人
  - 电子科大曾做过一个课题——寻找校园中最孤独的人。他们从3万名在校生中，采集到了2亿多条行为数据，数据来自学生选课记录、进出图书馆、寝室，以及食堂用餐、超市购物等数据。通过对不同的校园一卡通“一前一后刷卡”的记录进行分析，可以发现一个学生在学校有多少亲密朋友，比如恋人、闺蜜。
  - 最后，通过这个课题找到了800多个校园中最孤独的人，他们平均在校两年半时间，一个知心朋友都没有。这些人中的17%可能产生心理疾病，剩下的则可能用意志力暂时战胜了症状，但需要学校和家长重点予以关爱。

# 数据挖掘应用

- 数据新闻让英国撤军
  - 2010年10月23日《卫报》利用维基解密的数据做了一篇“数据新闻”。将伊拉克战争中所有的人员伤亡情况均标注于地图之上。地图上一个红点便代表一次死伤事件，鼠标点击红点后弹出的窗口则有详细的说明：伤亡人数、时间，造成伤亡的具体原因。密布的红点多达39万，显得格外触目惊心。一经刊出立即引起朝野震动，推动英国最终做出撤出驻伊拉克军队的决定。





# 数据

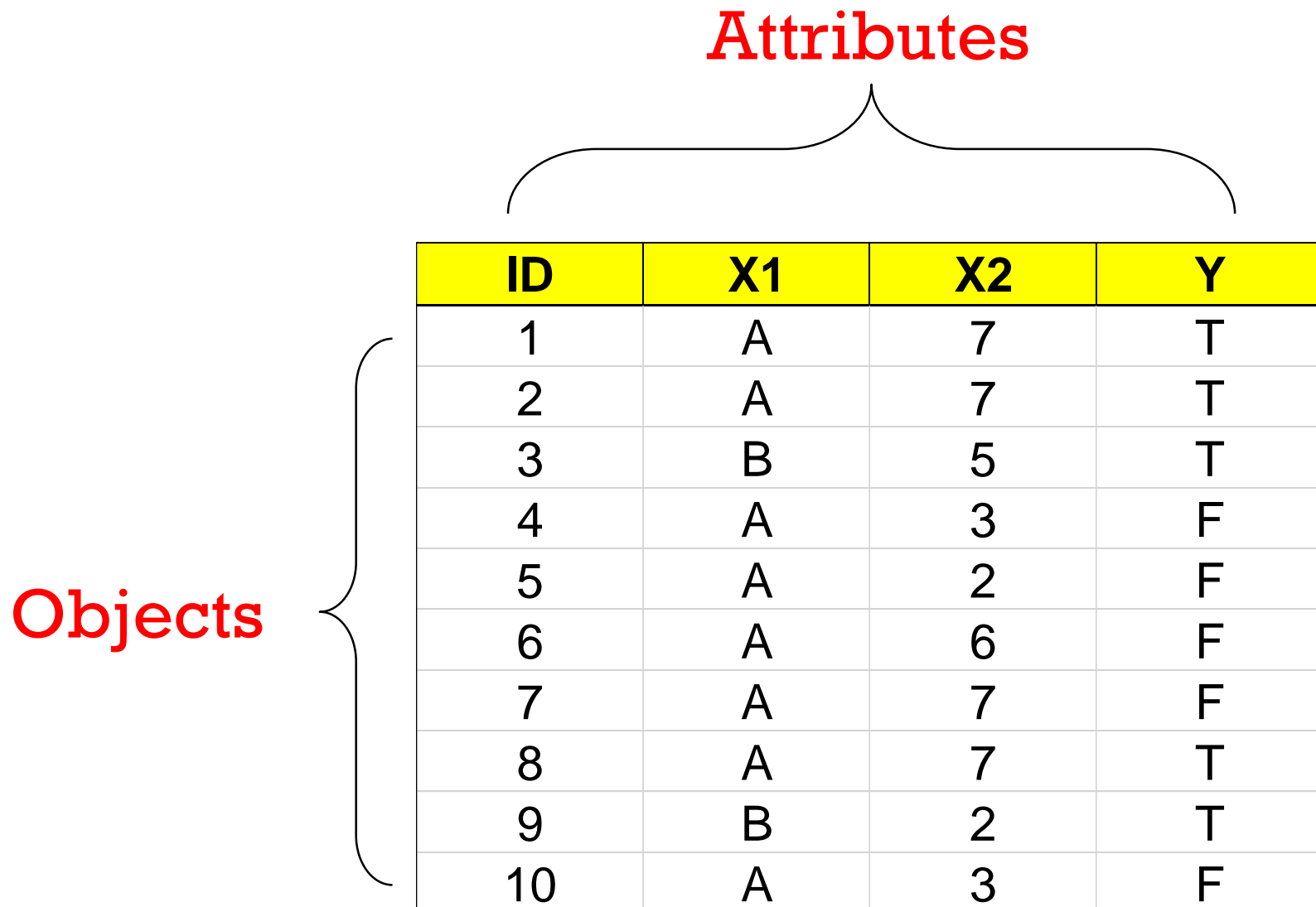
The Data

# 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

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）

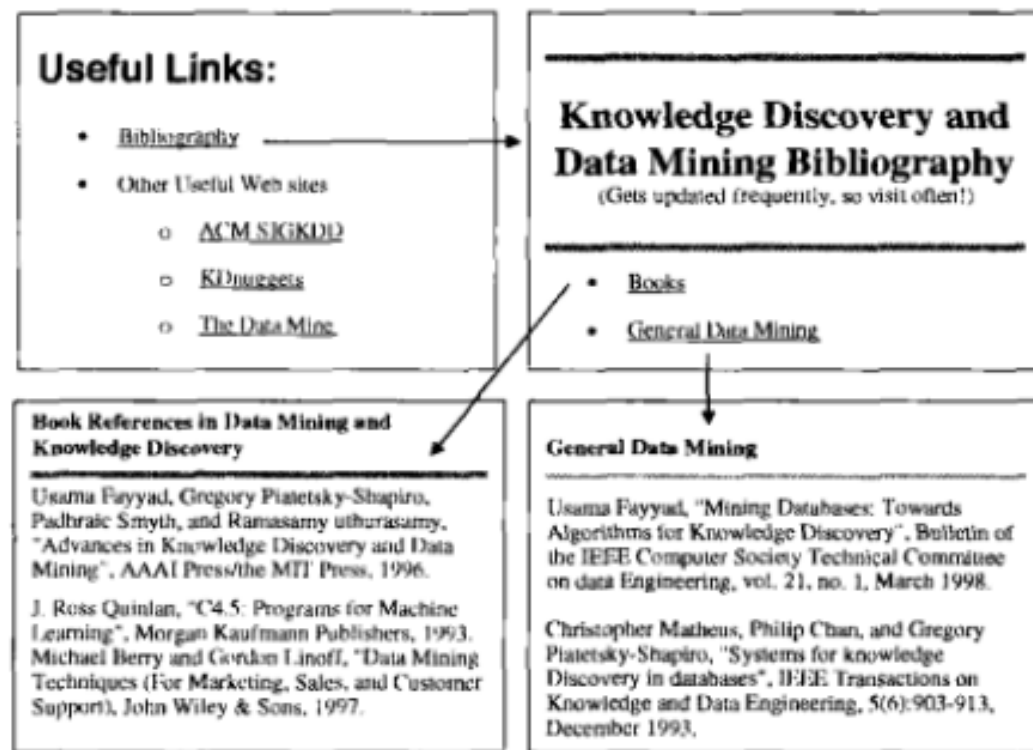
- 特殊类型的记录数据，每个记录（事务）涉及一系列的项

- 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

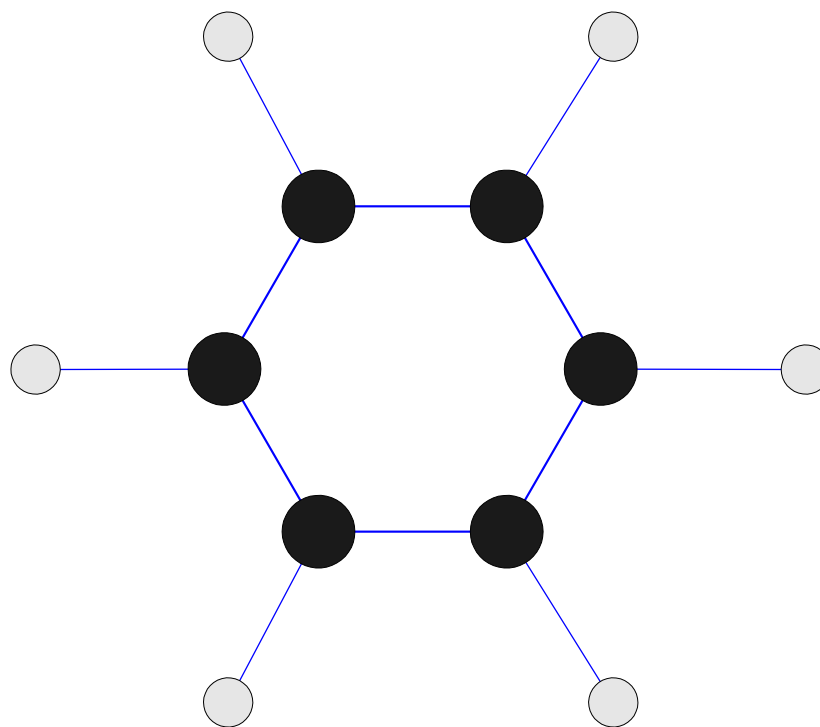
# 图形数据

- 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

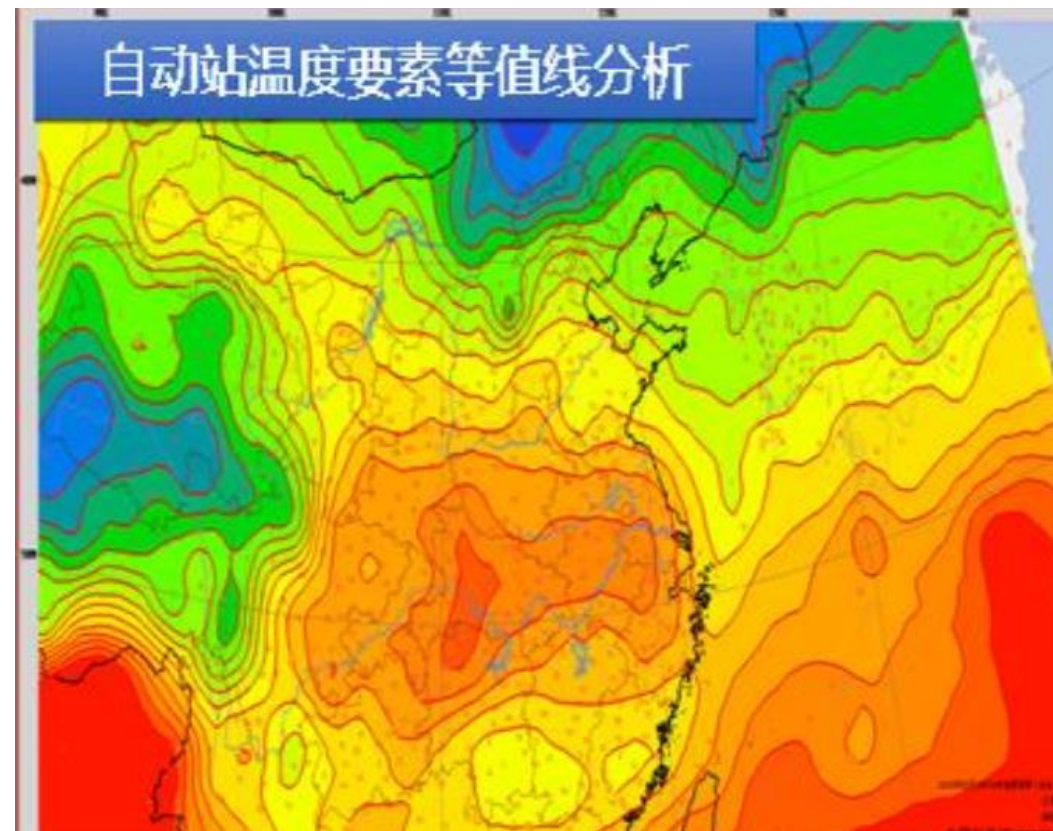
- 序列数据（Sequence Data）

- 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

```
GGTTCCGCCTTCAGCCCCGCGCC  
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







**THE END !**

