# 数 据 挖 掘

## Data Mining

主讲: 张仲楠 教授

厦門大學
XIAMEN UNIVERSITY

壹

# 绪论

# 目 录

# 1.数据时代

## 一分钟能做什么?

微信用户每一分钟发布46.52万张图片；每一分钟发起22.91万次视频通话；每一分钟会有54.16万人进入朋友圈。

百度用户每一分钟进行416.6万次搜索。每一分钟有6.94万次语音播报。

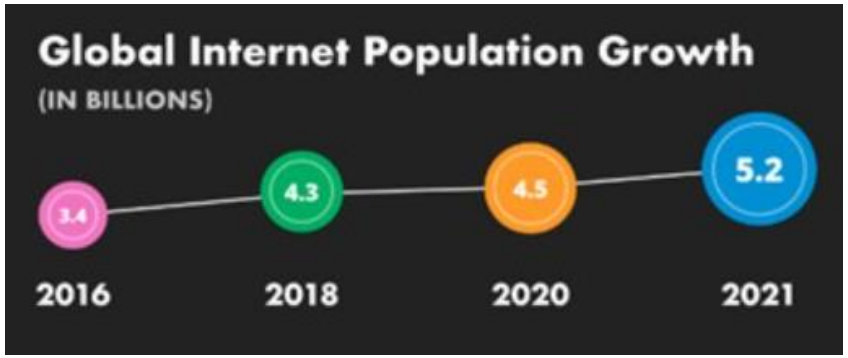
抖音海外版 TikTok 用户一分钟观看了 1.67 亿条视频

淘宝每一分钟会有658.8万人民币销售额。天猫每分钟会有767.59万销售额。

B站每分钟会有83.3万次播放。

# 1.数据时代

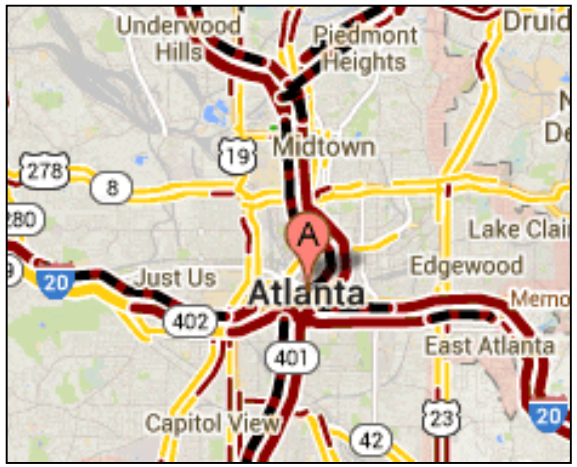
## 一分钟能做什么?




2021年互联网用户总数为50亿，比 2020 年的45亿增长了大约增长了 5 亿，也就是每一分钟增加 950 名新用户。

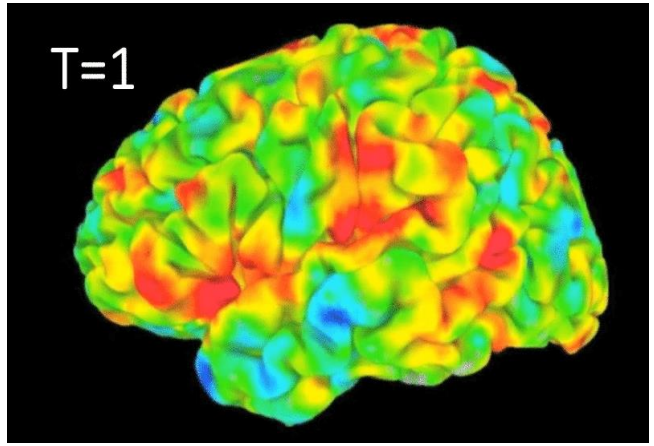
# 1.数据时代

## 交通数据




## 物联网数据

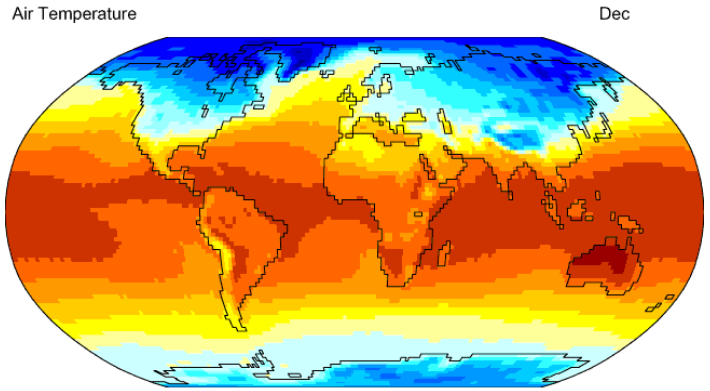
## 医学数据




## 天文观测数据

## 生命科学数据


## 地球观测数据



# 1.数据时代

**数据挖掘发展的动力**

- 数据爆炸：数据采集工具和成熟的数据存储技术使得大量的数据被收集和存储
- 我们拥有丰富的数据，但是缺乏有用的信息
- 迫切需要将传统的数据分析方法与用于处理大量数据的复杂算法相结合

# 1.数据时代

## 数据挖掘的价值 --- 电子商务领域

- 1. 个性化推荐系统：通过分析用户的历史购买记录和行为习惯，为用户提供个性化的产品推荐，提高购物体验和销售额。
- 2. 交易风险评估：基于大量的历史交易数据，通过数据挖掘的方法分析交易模式，提前识别潜在的诈骗和欺诈行为。
- 3. 市场细分：将用户数据进行分析，识别不同的市场细分，制定相应的销售和营销策略，提高销售效率。
- 4. 购物篮分析：分析不同产品之间的关联关系，为电商平台提供交叉销售的机会，提高销售额。

# 1.数据时代


## 数据挖掘的价值 --- 金融领域

- 1. 信用评估：通过分析个人或企业的历史信用信息和各类数据，进行信用评估，帮助金融机构决策。
- 2. 欺诈检测：通过对交易数据进行分析，发现潜在的欺诈模式和异常行为，预防金融欺诈事件的发生。
- 3. 股市预测：通过对历史股票数据进行挖掘，建立股市预测模型，辅助投资者做出投资决策。
- 4. 风险管理：通过对大规模的金融数据进行挖掘，分析不同风险因素对投资组合的影响，提供风险管理策略。

# 1.数据时代

## 数据挖掘的价值 --- 医疗健康领域

- 1. 疾病预测：基于患者的个人信息和病历数据，预测患者可能患上某些疾病的概率，提前进行干预和治疗。
- 2. 药物研发：通过对已有研究和药物数据进行分析，挖掘新的药物治疗方案和疗效评估方法。
- 3. 医疗资源优化：通过对医疗数据进行挖掘，医院可以更好地管理和调整资源，提高医疗服务的效率和质量。
- 4. 健康风险评估：基于个人健康数据和生活习惯，评估个体的健康风险，并提供相应的健康管理建议。

# 1.数据时代

## 数据挖掘的价值 --- 交通运输领域

- 1. 智能交通调度：通过分析交通流量数据和道路网络信息，优化交通调度和信号灯配时，减少交通拥堵和延误。
- 2. 交通事故预测：通过对历史交通事故数据和天气数据进行挖掘，建立交通事故预测模型，预防和减少交通事故发生。
- 3. 公共交通优化：通过分析公共交通数据和乘客出行模式，优化公共交通线路和班次，提高乘客出行效率。

# 1.数据时代

## 数据挖掘的价值 --- 社交媒体领域

- 1. 用户情感分析：通过分析用户在社交媒体上的发帖内容和情感表达了解用户对不同话题和事件的态度和情感倾向。
- 2. 舆情监测：通过对社交媒体上的数据进行挖掘，了解公众对某一事件或话题的态度和舆论趋势。
- 3. 社交关系分析：通过分析社交媒体上的用户行为和互动，了解用户之间的社交关系和用户群体特征。

# 1.数据时代

## 数据挖掘的价值 --- 制造业领域

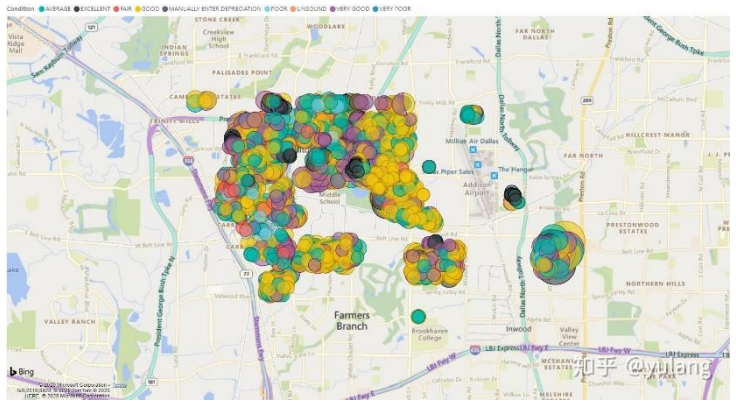
- 1. 产品质量控制：通过分析制造过程中的传感器数据和生产参数，实时监控产品质量，提前发现和预防质量问题。
- 2. 故障预测与预防：通过对设备故障数据和维修记录进行挖掘，预测设备故障的概率和时间，提前进行维修和保养，减少生产中断时间。
- 3. 供应链优化：通过分析供应链上的数据，了解供应链网络、库存和物流的状态，优化供应链的运作效率。

# 1.数据时代

## 数据挖掘的价值 --- 教育领域

- 1. 学生表现预测：通过分析学生的个人信息和学习行为数据，预测学生未来的学习成绩和表现，提供个性化教育方案。
- 2. 学习模式分析：通过对学生的学习行为和学习轨迹数据进行挖掘了解不同学生的学习模式和学习风格，辅助教师做出个性化的教学决策。

# 1.数据时代




## 数据挖掘的价值 --- 政府与公共服务领域

- 1. 犯罪预测：通过分析犯罪历史数据和环境因素，预测不同区域的犯罪概率和类型，提前采取措施维护社会治安。
- 2. 政府支出分析：通过分析政府财政数据和项目投资数据，优化政府支出结构和预算分配，提高投资效益。
- 3. 疫情预测与控制：通过对传染病数据和人口流动数据进行挖掘，预测疫情的传播趋势和高风险区域，制定相应的疫情防控策略。

# 1.数据时代

## 数据挖掘的价值 --- 农业领域

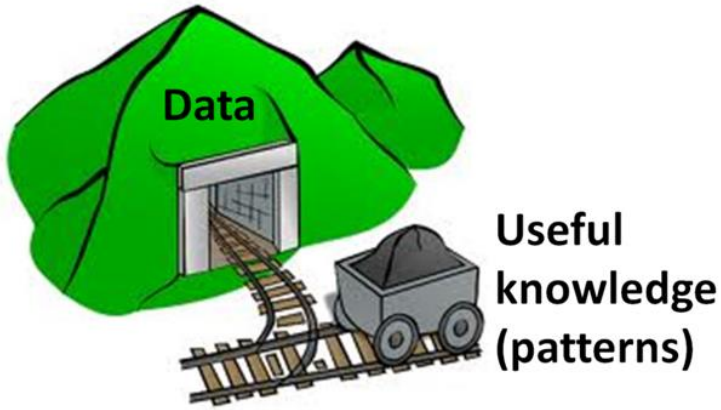
- 1. 作物病虫害预警：通过分析农作物生长环境数据和病虫害历史数据提前预警并采取防控措施，保护作物产量。
- 2. 农产品价格预测：通过对市场供需数据和经济因素的分析，预测农产品价格水平和趋势，为农民和市场参与者提供参考依据。

# 目 录

# 2.数据挖掘概念

## 什么是数据挖掘

- 数据挖掘是在大型数据库中自动发现有用信息的过程
  - 探查大型数据库
  - 发现先前未知的有用模式（pattern）
- 可以预测未来的观测结果
- 并非所有的信息发现任务都被视为数据挖掘
  - 数据库的查询任务

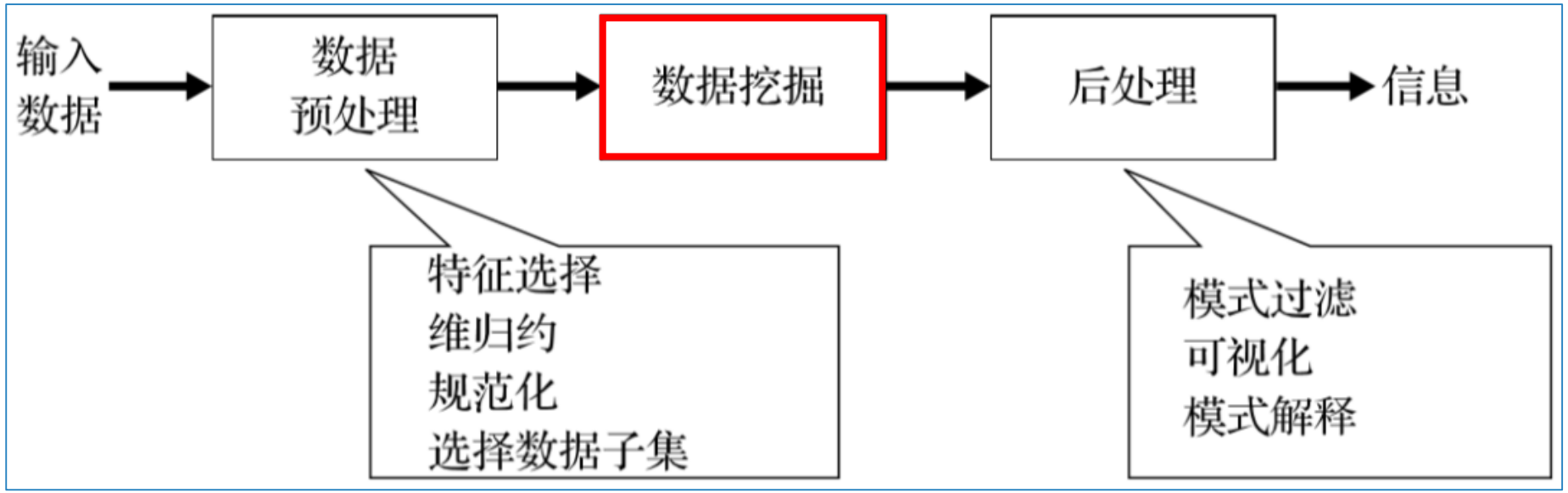


# 2.数据挖掘概念

## 什么是数据挖掘

- 数据挖掘是数据库中知识发现（KDD）不可缺少的一部分
- KDD是识别数据中有效的、新颖的、潜在有用的和易于理解的模式的非平凡过程(Goebel, 1999)

KDD过程



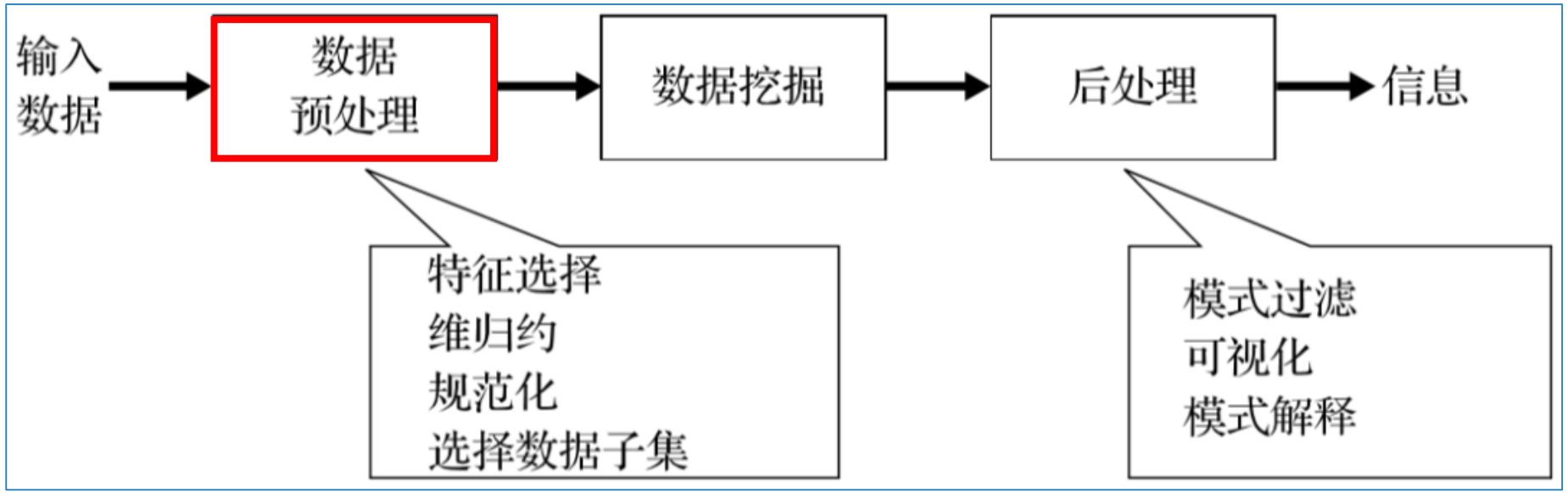
# 2.数据挖掘概念

## 什么是数据挖掘

■预处理是将原始输入数据转换为适当的格式，以便后续分析

■融合多个数据源的数据，数据清洗（消除噪声和重复值），选择与任务相关的记录和特征

KDD过程



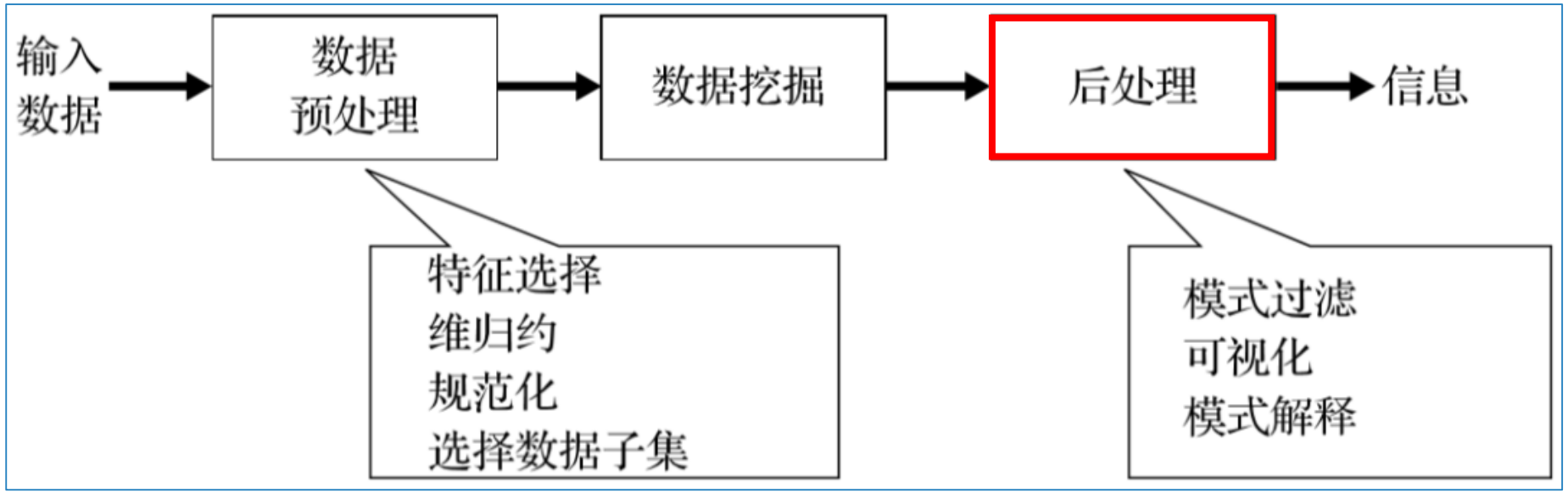
# 2.数据挖掘概念


## 什么是数据挖掘

■后处理确保只将有效和有用的结果集成到决策支持系统(DSS)中

- 可视化: 使得数据分析者从不同的视角探查数据和挖掘结果
- 使用统计度量或假设检验，删除虚假挖掘结果

KDD过程



# 目 录

# 3.数据挖掘要解决的问题

**可伸缩性/可扩展性(Scalability)**

- 代表一种弹性，随着数据量的快速增长，能够保证旺盛的生命力，通过很少的改动甚至只是硬件设备的添置，就能实现处理能力的线性增长，实现高吞吐量和高性能
  - 采样技术：随机采样，均匀采样，水库采样
  - 并行(分布式)计算：Map/Reduce计算框架
  - 在线/增量学习方法：持续学习 (Continual Learning)

# 3.数据挖掘要解决的问题

**高维性(**High Dimensionality**)**

- 维度(Dimension)通常是指描述数据的属性或特征，比如地理位置、时间、类别、性别、职业等，它们是用来划分数据的类别或属性。
- 为低维数据开发的传统分析技术通常不能很好地处理高维数据。
- 某些分析算法，随着维度（特征数）的增加，计算复杂度会迅速增加。

# 3.数据挖掘要解决的问题

**异构和复杂(**Heterogeneous and Complex**)**

- 传统分析方法只处理包含相同类型属性的数据集
- 现在的数据类型更复杂：文本、图像、音频和视频
- 为挖掘复杂对象，需要考虑数据中的联系
  - 时间和空间的自相关性
  - 图的连通性
  - 半结构化数据与XML文档的关系

# 3.数据挖掘要解决的问题

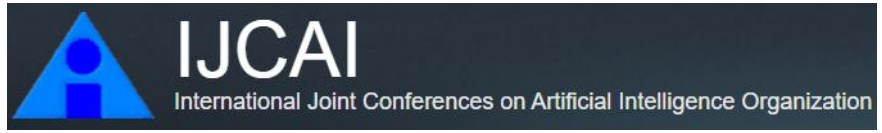
**所有权和分布(**Ownership and Distribution**)**

- 针对在物理上分布式存储的数据需要面临一些列挑战
  - 如何降低分布式计算所带来的额外通信开销
  - 如何有效地从多个数据源获得挖掘结果
  - 如何解决数据安全和隐私问题

# 目 录

# 4.数据挖掘的起源

1989年8月于美国底特律市召开的第十一届国际联合人工智能学术会议上首次提到“知识发现”这一概念；



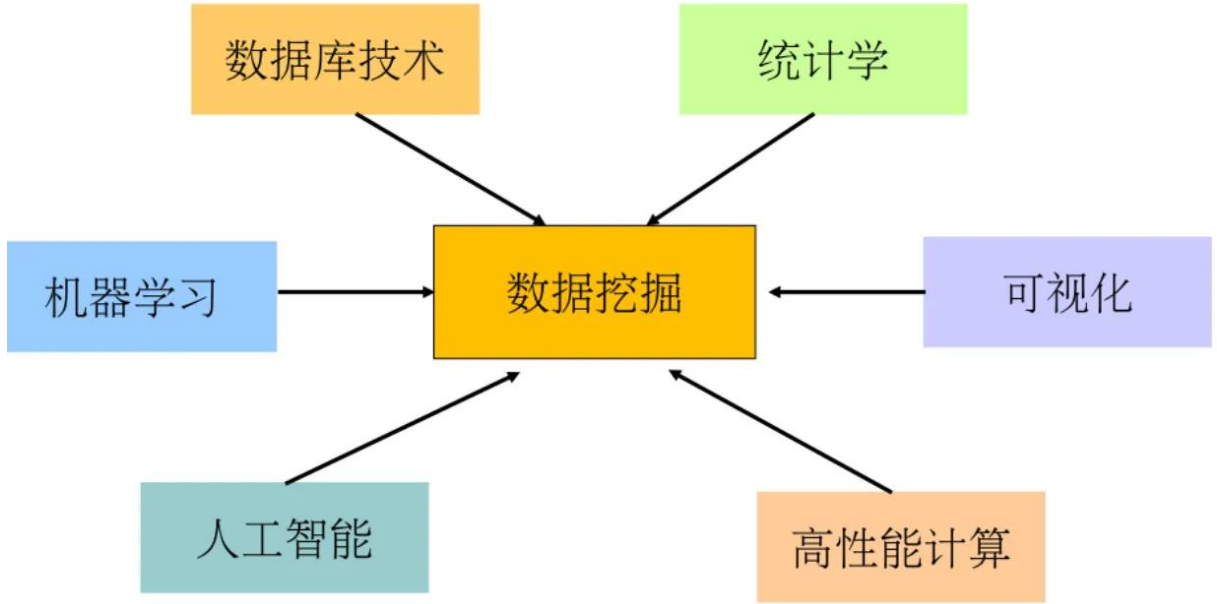

到1993年，美国电气电子工程师学会(IEEE)的知识与数据工程(Knowledge and Data Engineering)汇刊出版了KDD技术专刊，发表的论文和摘要体现了当时KDD的最新研究成果和动态。


1995年在加拿大蒙特利尔召开的首届“知识发现和数据挖掘”国际学术会议上，首次提出了“数据挖掘”这一学科的名称，并把数据挖掘技术分为科研领域的知识发现与工程领域的数据挖掘。

# 4.数据挖掘的起源

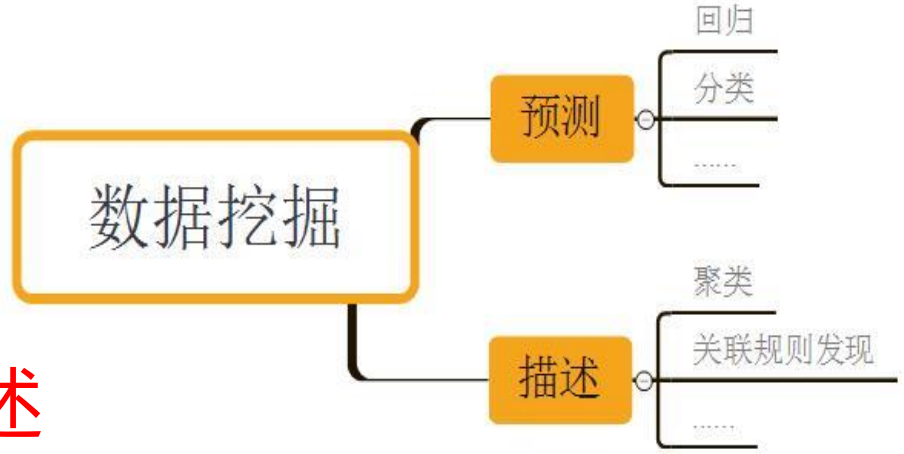
## 多学科的交叉

■涉及数据库技术、人工智能、数理统计、机器学习、模式识别、高性能计算、知识工程、神经网络、信息检索、信息的可视化等众多领域



# 目 录

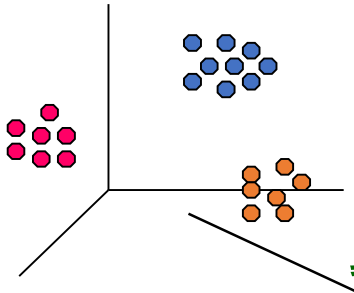
# 5.数据挖掘的任务




## 数据挖掘要做什么

- 数据挖掘的两大基本任务是预测和描述
- 预测任务：根据其他属性的值预测特定属性的值
  - 被预测的属性：目标变量/因变量
  - 用于预测的属性：解释变量/自变量
- 描述任务：导出数据中潜在联系的模式
  - 相关、趋势、聚类、轨迹和异常
  - 探查性的，需要加以验证和解释

# 5.数据挖掘的任务

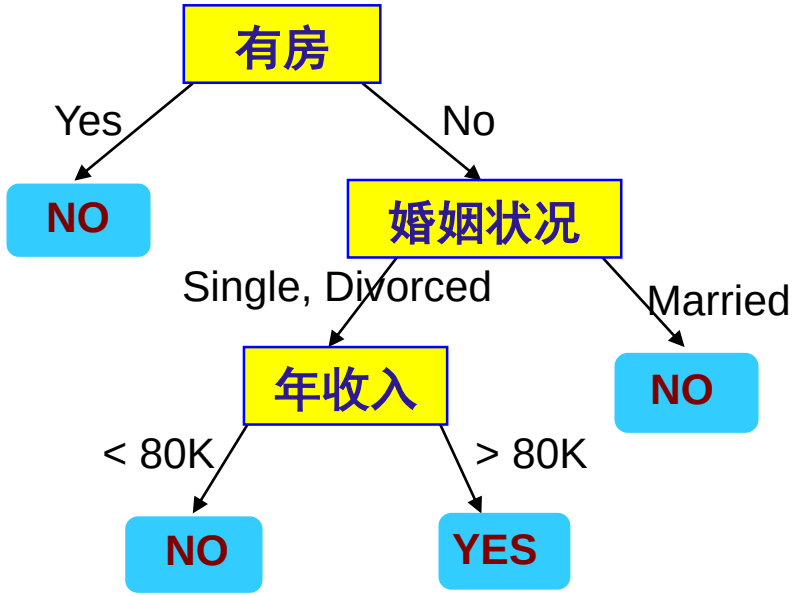


聚类分析

**数据**

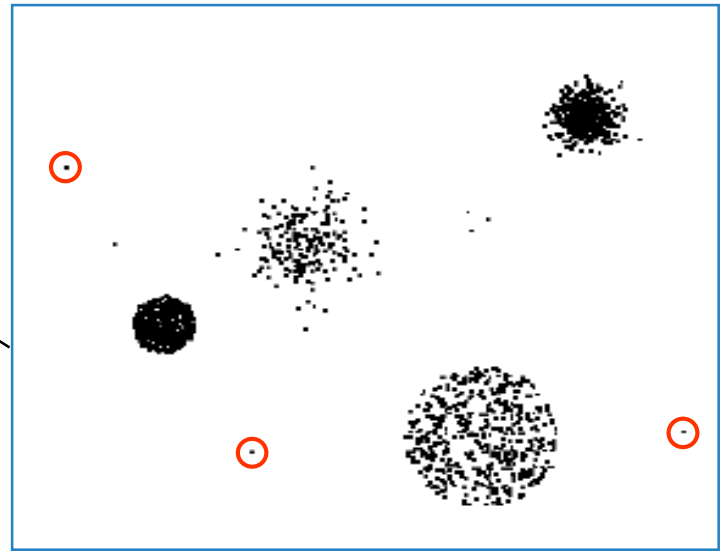
| ID | 有房 | 婚姻状况 | 年收入 | 拖欠贷款 |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

预测建模




关联分析


异常检测



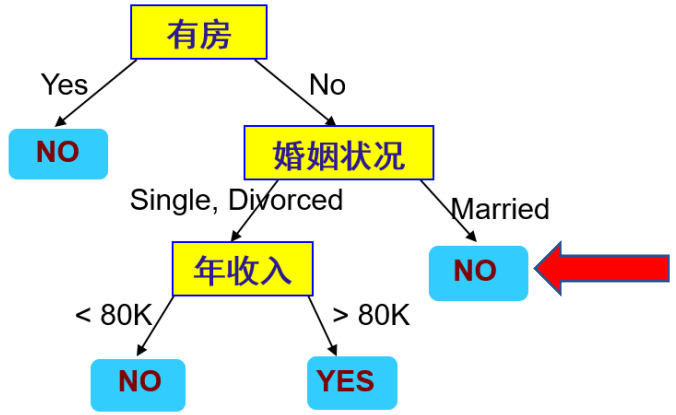
# 5.数据挖掘的任务

## 1. 预测建模

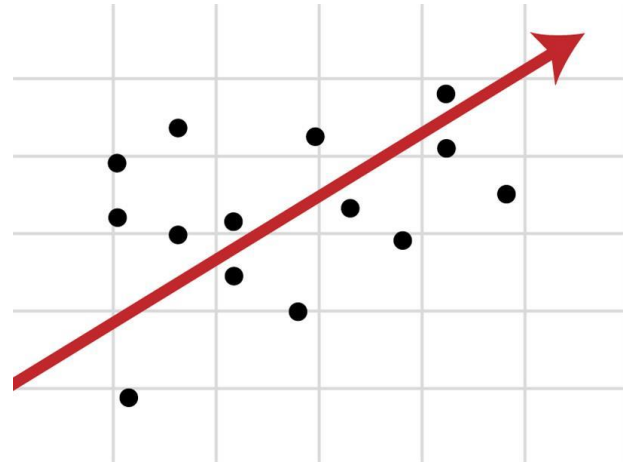
- 为目标变量建立模型，将其作为解释变量的函数
  - 分类：用于预测离散的目标变量
  - 回归：用于预测连续的目标变量
- 任务目标：使目标变量的预测值与实际值之间的误差最小

**分类**

| 有房 | 婚姻状况 | 年收入 | 拖欠贷款 |
|---|---|---|---|
| No | Married | 80K | ??? |



# 5.数据挖掘的任务



| 房子面积 | 占地大小 | 卧室 | 花岗岩 | 卫生间重装 | 销售价格 |
|---|---|---|---|---|---|
| 3529 | 9191 | 6 | 0 | 0 | 205, 000 |
| 3247 | 10061 | 5 | 1 | 1 | 224, 900 |
| 4032 | 10150 | 5 | 0 | 1 | 197, 900 |
| 2397 | 14156 | 4 | 1 | 0 | 189, 900 |
| 2200 | 9600 | 4 | 0 | 1 | 195, 000 |
| 3536 | 19994 | 6 | 1 | 1 | 325, 000 |
| 2983 | 9351 | 5 | 0 | 1 | 230, 000 |
| 3198 | 9669 | 5 | 1 | 1 | ???? |

219, 358

回　归

# 5.数据挖掘的任务

## 2. 关联分析

- 用来发现描述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变量
- 任务目标：以最有效的方式提取最有趣的模式

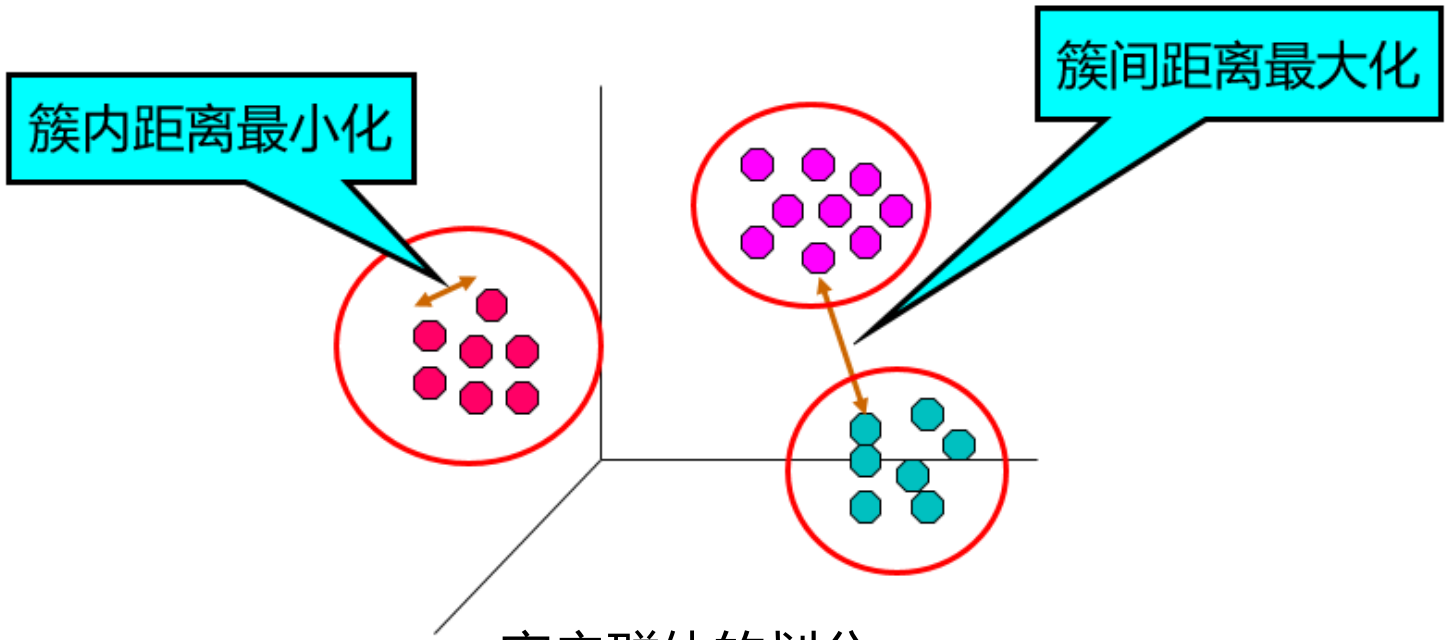
| *TID* | *Items* |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

$\{Milk,Diaper\} \rightarrow \{Beer\}$
$\{Milk,Beer\} \rightarrow \{Diaper\}$
$\{Diaper,Beer\} \rightarrow \{Milk\}$
$\{Beer\} \rightarrow \{Milk,Diaper\}$

# 5.数据挖掘的任务

## 3. 聚类分析

■旨在发现紧密相关的观测值组群，使得与属于不同簇的观测值相比，属于同一簇的观测值之间尽可能类似模式



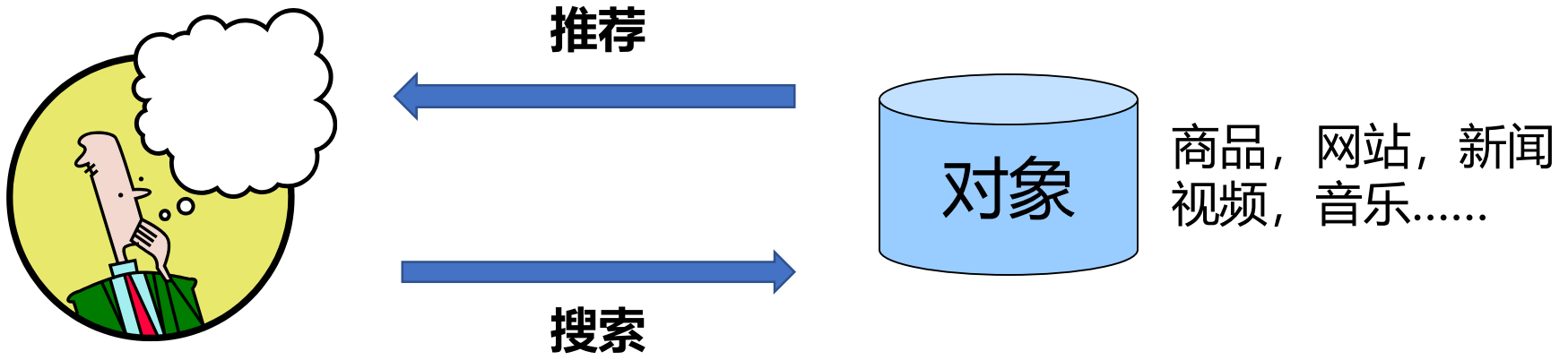

客户群体的划分

# 5.数据挖掘的任务

## 4. 推荐算法

■通过一些数学算法，推测出用户可能喜欢的东西

■任务目标：帮助用户找到真正想要的+降低信息过载

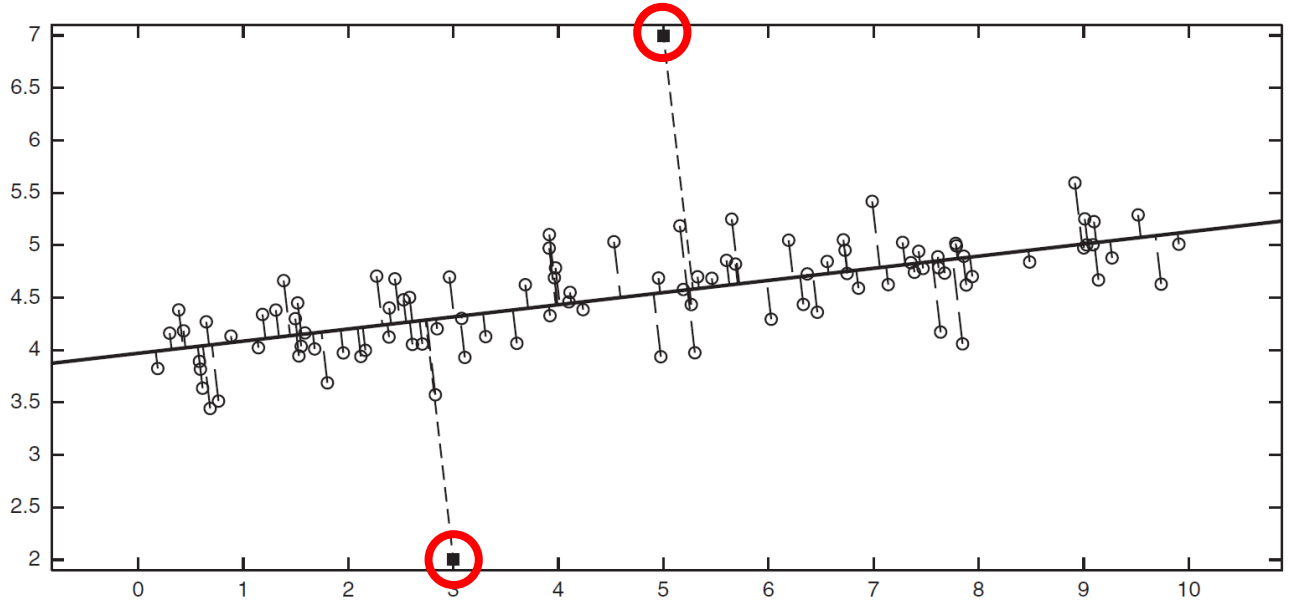


# 5.数据挖掘的任务

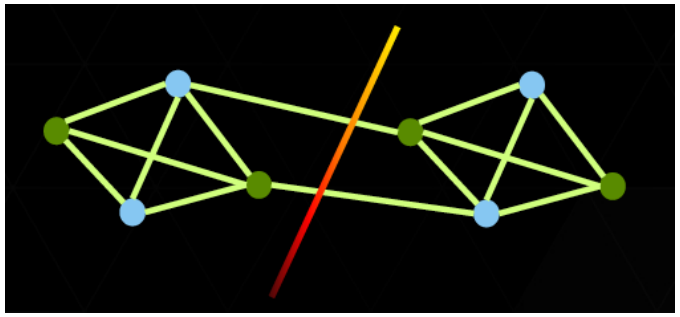
## 5. 异常检测

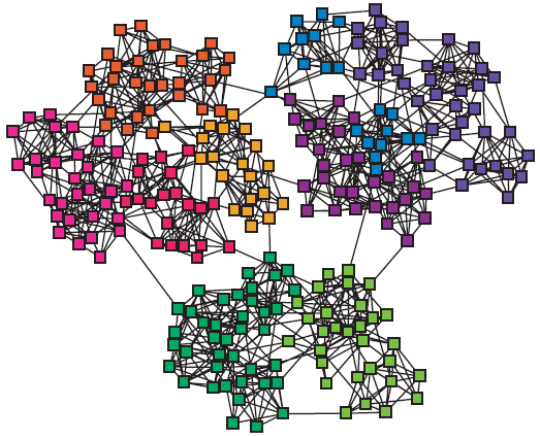
■识别特征显著不同于其他数据的观测值，这样的观测值称为异常点或离群点

■任务目标：具有高检测率和低误报率



# 5.数据挖掘的任务





## 6. 图数据分析

- 图（Graph）是一种数据结构，用于模拟物理、生物、社会和信息系统中的许多类型的关系和过程。
- 图由节点或顶点（表示系统中的实体）组成，这些节点或顶点由边（表示这些实体之间的关系）连接。
- 任务目标：使用特定于图的算法来实现节点级、边级和图级的分析任务。

# 习题

讨论下列每项活动是否是数据挖据任务。

a)根据性别划分公司的顾客。

b)根据可赢利性划分公司的顾客。

c)计算公司的总销售额。

d)按学生的标识号对学生数据库排序。

e)预测掷一对骰子的结果。

f)使用历史记录预测某公司未来的股票价格。

g)监视病人心率的异常变化。

h)监视地震活动的地震波。

i)提取声波的频率