

2024-2025学年 第1学期(秋)



数据挖掘

Data Mining

袁晓彤

南京大学智能科学与技术学院

南雍楼西519 xytuan@nju.edu.cn

2024 年 9 月

课 程 简 介

课程基本信息

□ **年级专业：** 2022级 智能科学与技术

□ **课时：** 32

□ **学分：** 2

□ **考核方式：** 随堂表现 (10%) + 平时作业 (30%) + 期末考试 (60%)

与数据挖掘相关的课程

□ 基础课程：

- 高等数学、 概率与统计、 线性代数、 离散数学
- 数据结构与算法、 程序设计

□ 相关课程

- 机器学习、 人工智能、 模式识别
- 图像处理、 自然语言处理
- 数据库
- 计算机体系结构
- ...

教学方法

- 着重讲述数据挖掘的**基本概念**，**基本方法**和**算法原理**。
- 注重理论与实践紧密结合
 - **实例教学**：通过大量实例讲述如何将所学知识运用到实际应用之中
 - **前沿进展**：介绍数据挖掘技术在互联网、气象、安全等领域的前沿研究和应用进展
- 避免引用过多的、繁琐的数学推导。

教学目标

- 掌握数据挖掘的基本概念和方法
- 有效地运用所学知识和方法解决实际问题
- 为进一步学习数据科学相关理论和方法打下基础

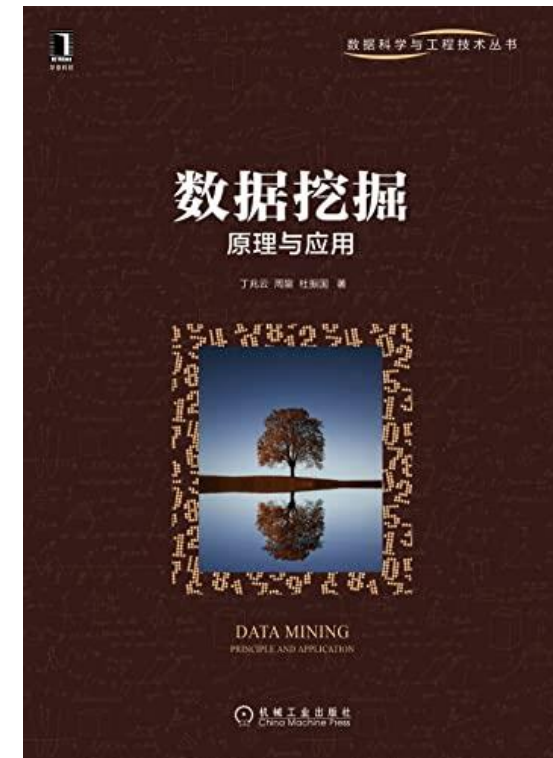
课程教材

□ 主要教材

- 丁兆云，周鋈，杜振国，*数据挖掘：原理与应用*，机械工业出版社，2022

□ 参考教材

- Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Elsevier, 2011
- 喻梅，于健，王建荣，李雪威，*数据分析与挖掘技术*，清华大学出版社，2020
- 欧高炎、朱占星、董彬、鄂维南，*数据科学导引*，高等教育出版社，2017





南京大學
NANJING UNIVERSITY

目录

01

概述

02

数据挖掘的内涵

03

数据挖掘的内容

04

相关资源

什么是数据？

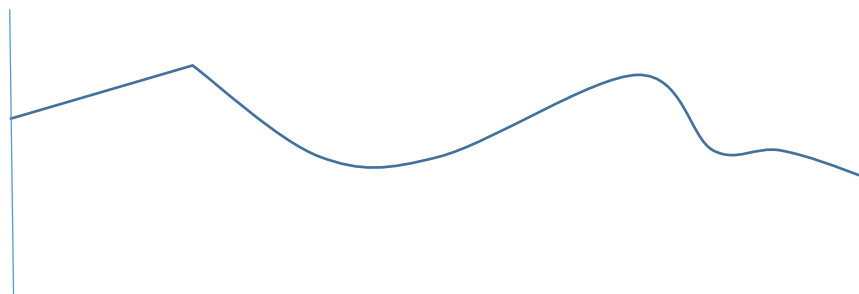
- 数据是所有能输入到计算机并被计算机程序处理的符号的总称
 - 人们通过观察现实世界中的自然现象、人类活动，都可以形成数据



28度



序号	温度
1	27
2	28
3	32
4	33
5	28
6	27
...	...



数据的分类

□ 结构化

Id	Name	Age	Gender
1	小明	12	男
2	小白	13	女
3	小奇	18	男

□ 半结构化

小明是个男孩，今年12岁；小白今年13岁，她是个女孩；小奇18岁了，他是个男生。

□ 非结构化



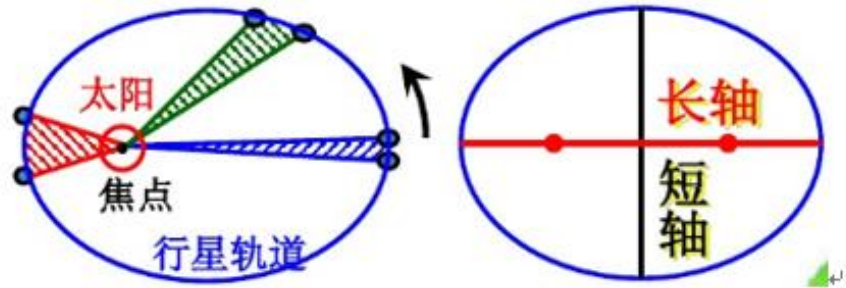
为什么要用数据挖掘

- 数据蕴含价值

表 1-1 太阳系八大行星绕太阳运动的数据^[4]

行星↵	周期/年↵	平均距离↵	周期/距离↵	周期 ² /距离 ³ ↵
水星↵	0.241↵	0.39↵	0.62↵	0.98↵
金星↵	0.615↵	0.72↵	0.85↵	1.01↵
地球↵	1.00↵	1.00↵	1.00↵	1.00↵
火星↵	1.88↵	1.52↵	1.24↵	1.01↵
木星↵	11.8↵	5.20↵	2.27↵	0.99↵
土星↵	29.5↵	9.54↵	3.09↵	1.00↵
天王星↵	84.0↵	19.18↵	4.38↵	1.00↵
海王星↵	165↵	30.06↵	5.49↵	1.00↵

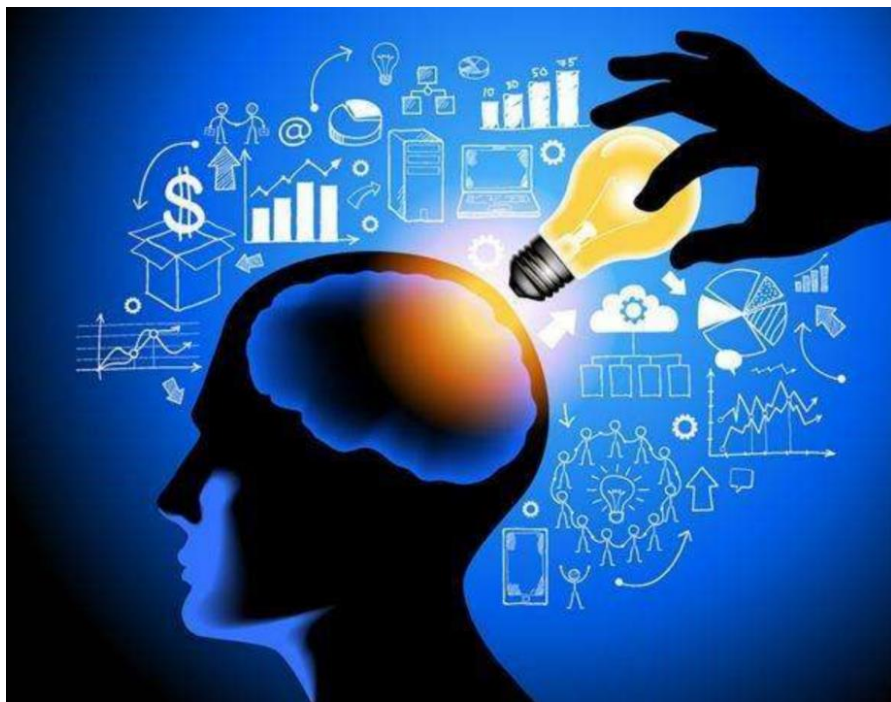
- ① 椭圆定律：所有行星绕太阳的轨道都是椭圆，而且太阳在椭圆的一个焦点上；↵
- ② 面积定律：行星和太阳的连线在相等的时间间隔内扫过相等的面积；↵
- ③ 调和定律：所有行星绕太阳一周的恒星时间(T_i)的平方与它们椭圆轨道半长轴(R_i)的三次方成正比，即 $R_i^3/T_i^2 = k$ (k 是常数)。↵



为什么要用数据挖掘

- 数据爆炸但知识贫乏

人们积累的数据越来越多。但是，传统数据处理方式还仅仅停留在数据的录入、查询、统计等功能，无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，导致了“数据爆炸但知识贫乏”的现象。



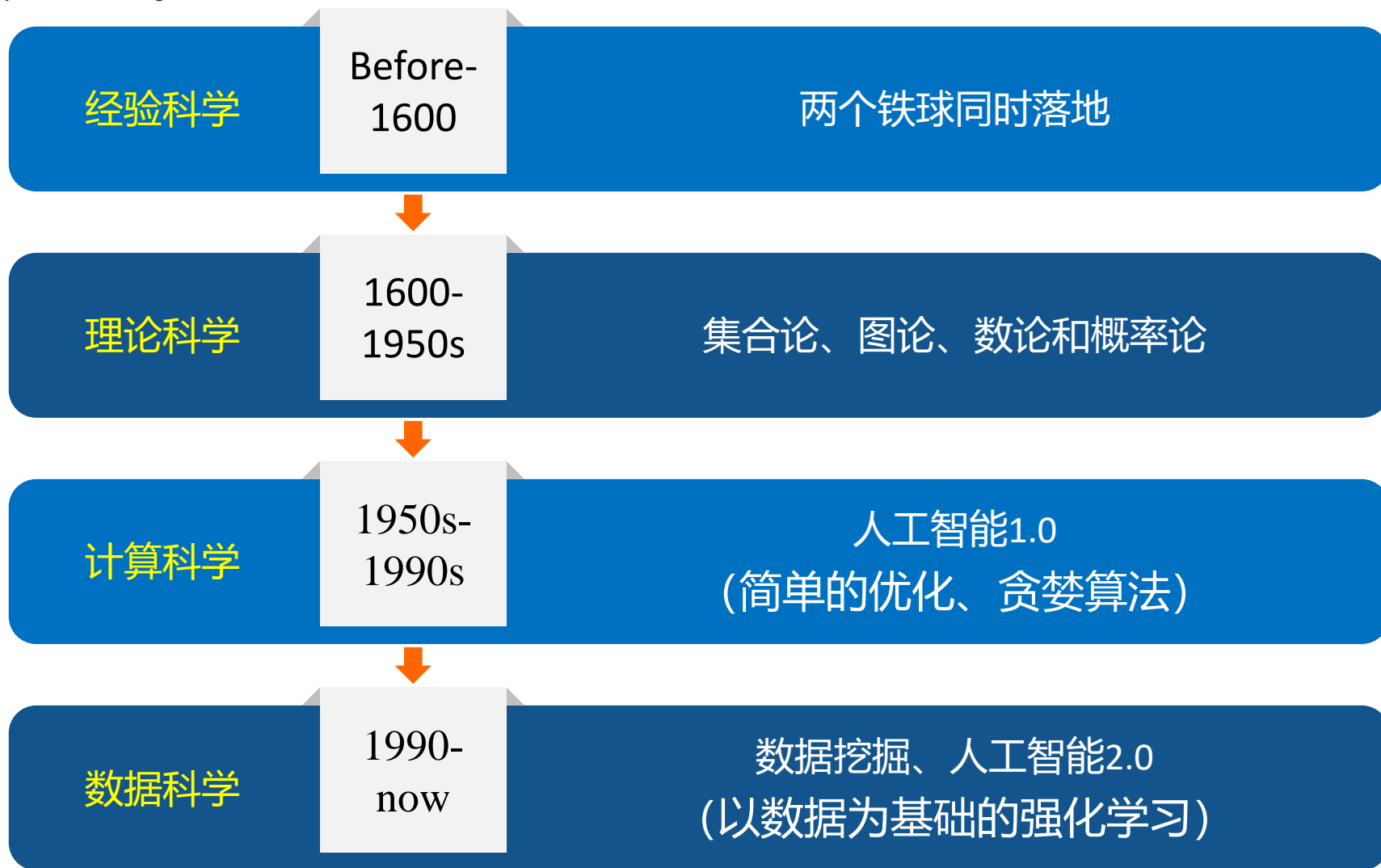
为什么要用数据挖掘

● 从商业数据到商业智能的进化

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集 (60年代)	“过去五年中我的总收入是多少？”	计算机、磁带和磁盘	IBM CDC	提供历史性的、静态的数据信息
数据访问 (80年代)	“在新英格兰的分部去年三月的销售额是多少？”	关系数据库(RDBMS) 结构化查询语言(SQL) ODBC	Oracle Sybase IBM Microsoft	在记录级提供历史性的、动态数据信息
数据仓库 决策支持 (90年代)	“在新英格兰的分部去年三月的销售额是多少？波士顿据此可得出什么结论？”	联机分析处理(OLAP) 多维数据库 数据仓库	Pilot Comshare Cognos Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘 (正在流行)	“下个月波士顿的销售会怎么样？为什么？”	高级算法 多处理器计算机 海量数据库	Pilot Lockheed IBM SGI 其他初创公司	提供预测性的信息

为什么要用数据挖掘

- 科学发展范式



KDD的出现

- 基于数据库的知识发现（KDD）一词首次出现在1989年举行的国际人工智能联合大会IJCAI-89 Workshop。
- 1995年在加拿大蒙特利尔召开了第一届KDD国际学术会议（KDD'95）。
- 由Kluwers Publishers出版，1997年创刊的《Knowledge Discovery and Data Mining》是该领域中的第一本学术刊物。



KDD2024
BARCELONA, SPAIN

数据挖掘的定义

- 数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- Data Mining is the process of automatically extracting interesting and useful hidden patterns from usually massive, incomplete and noisy data. [Wikipedia]



数据挖掘是多学科交叉的产物



数据挖掘是智能技术的核心





南京大學
NANJING UNIVERSITY

目录

01

概述

02

数据挖掘的内涵

03

数据挖掘的内容

04

相关资源

数据、信息、知识

客户信息表

年龄	收入（万）	工作时间（年）	顾客类型
25	10	2	优
29	12	3	优
32	9	6	良
38	7	12	良
36	18	13	中
30	15	4	优
...

数据、信息、知识

客户信息表

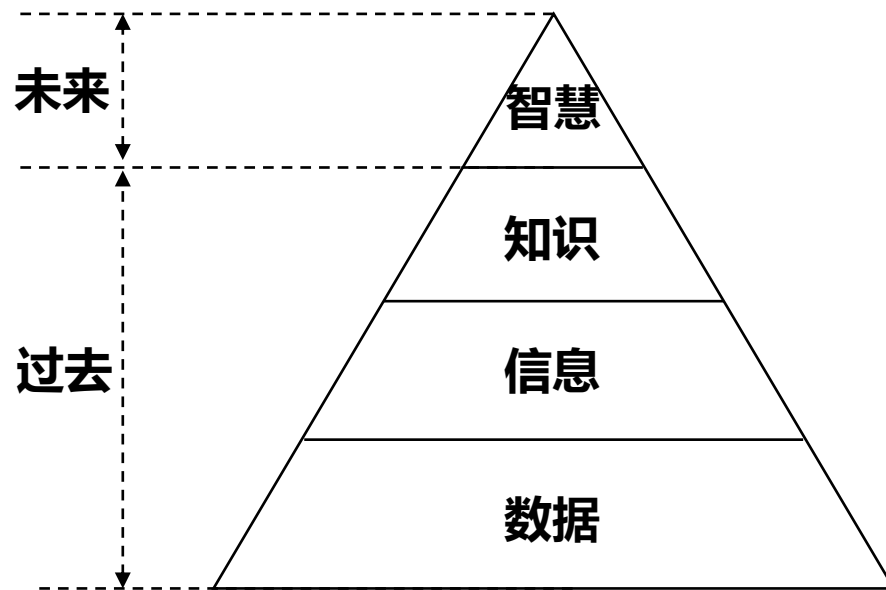
年龄	收入 (万)	工作时间 (年)	顾客类型
25	10	数据 2	优
29	12	3	优
32	9	6	良
38	7	12	良
36	18	信息 13	中
30	15	4	优
...

25<年龄<30, 收入>10, 工作时间>2年的消费者是优质顾客

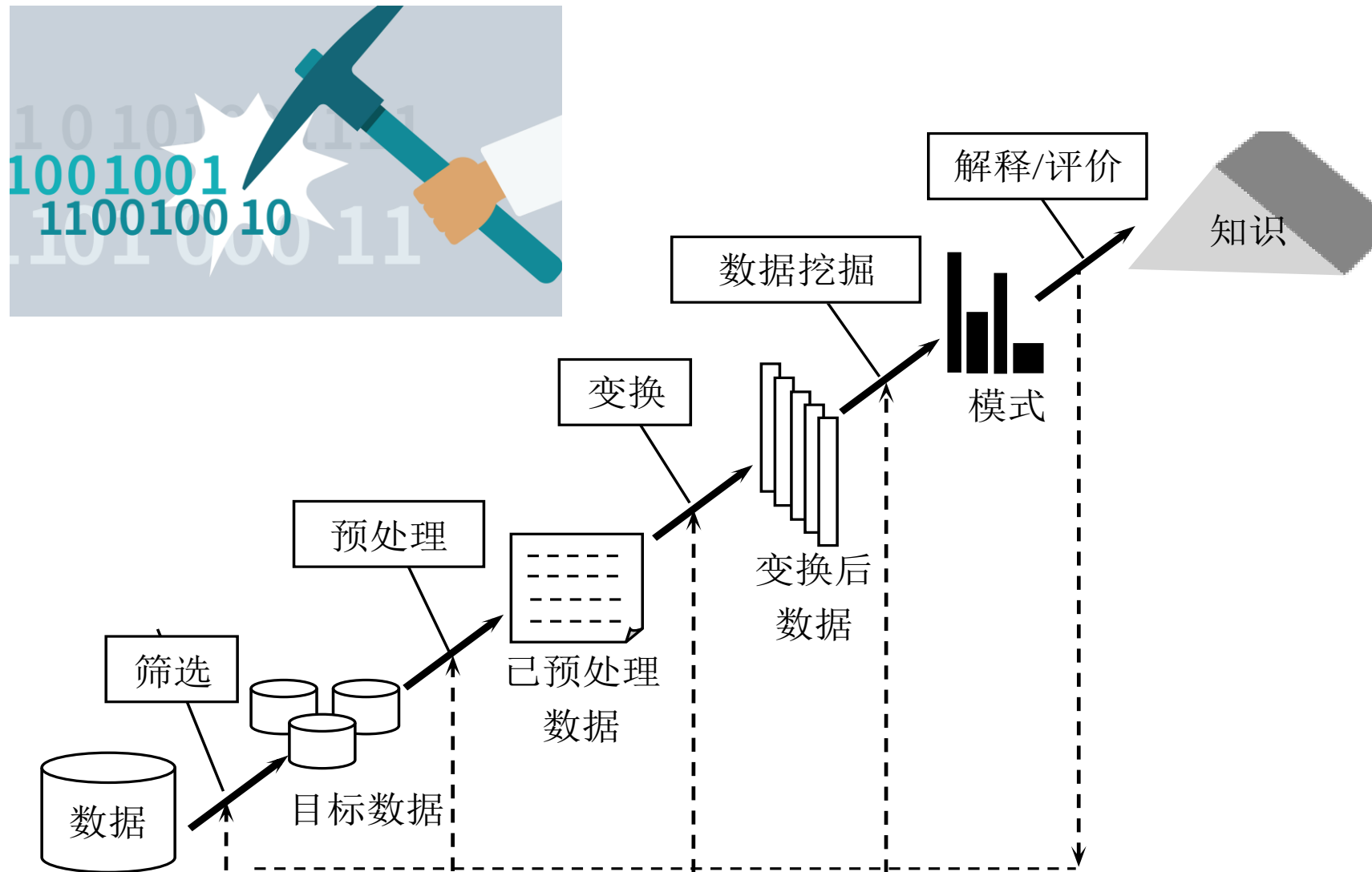
知识

DIKW金字塔

- 数据、信息、知识、智慧

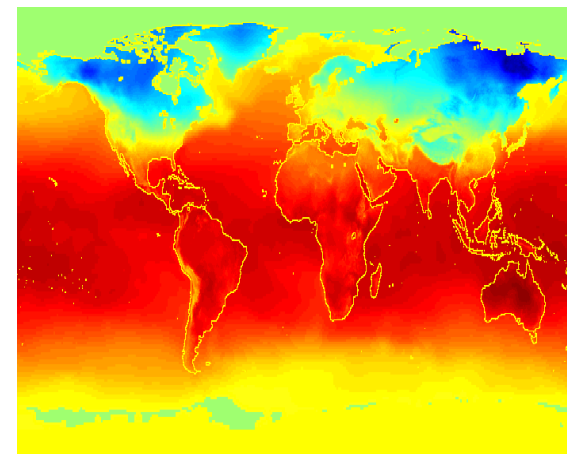
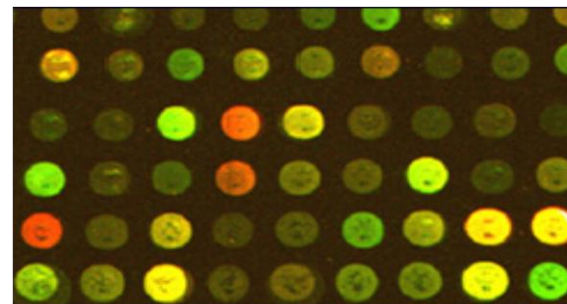
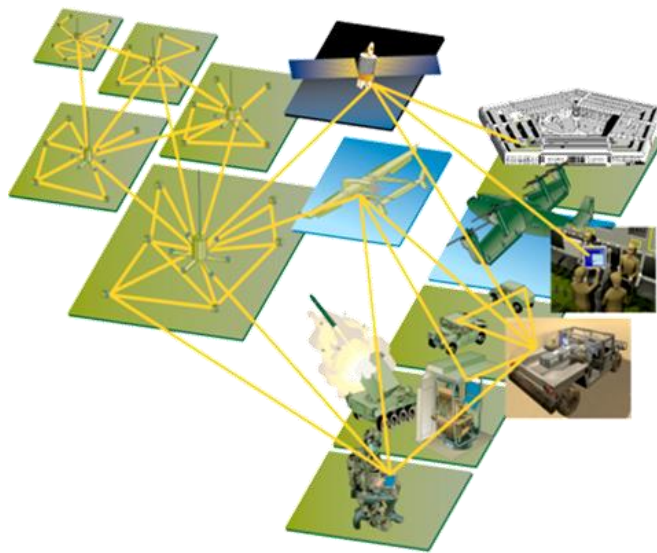


数据挖掘过程



数据来源

- 关系数据库
- 数据仓库
- 事务数据库
- 空间、时间数据库
- 文本和多媒体数据（异构的）
- 各种结构化、半结构化的数据
- 互联网、移动互联网数据源
-





南京大學
NANJING UNIVERSITY

目录

01

概述

02

数据挖掘的内涵

03

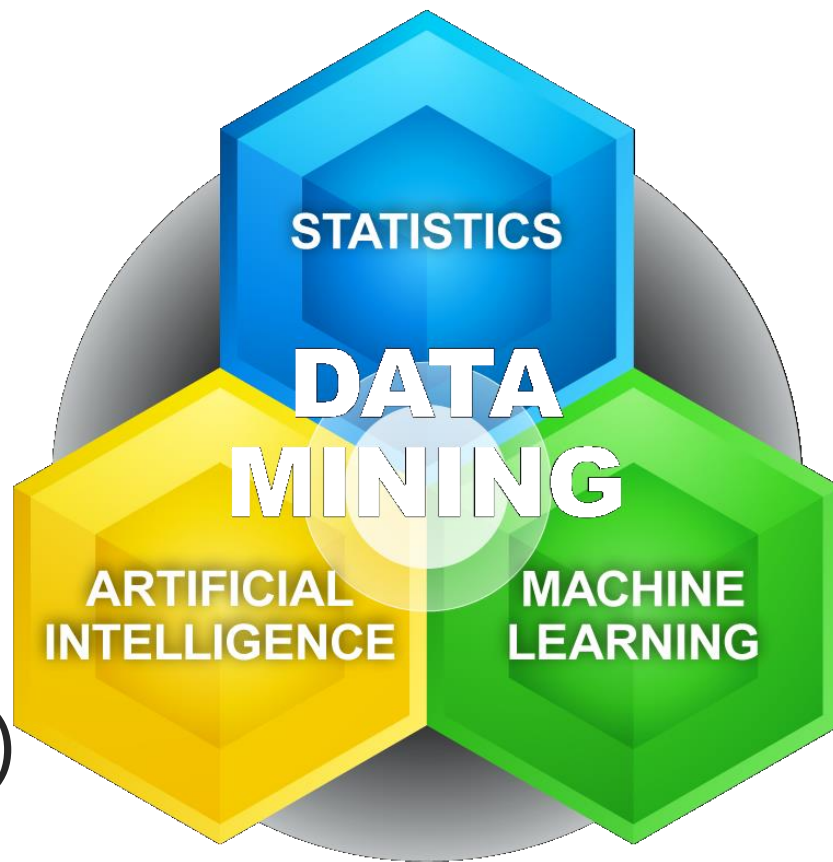
数据挖掘的内容

04

相关资源

数据挖掘的主要内容

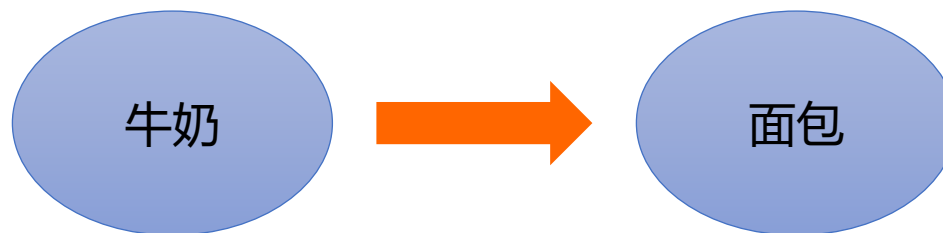
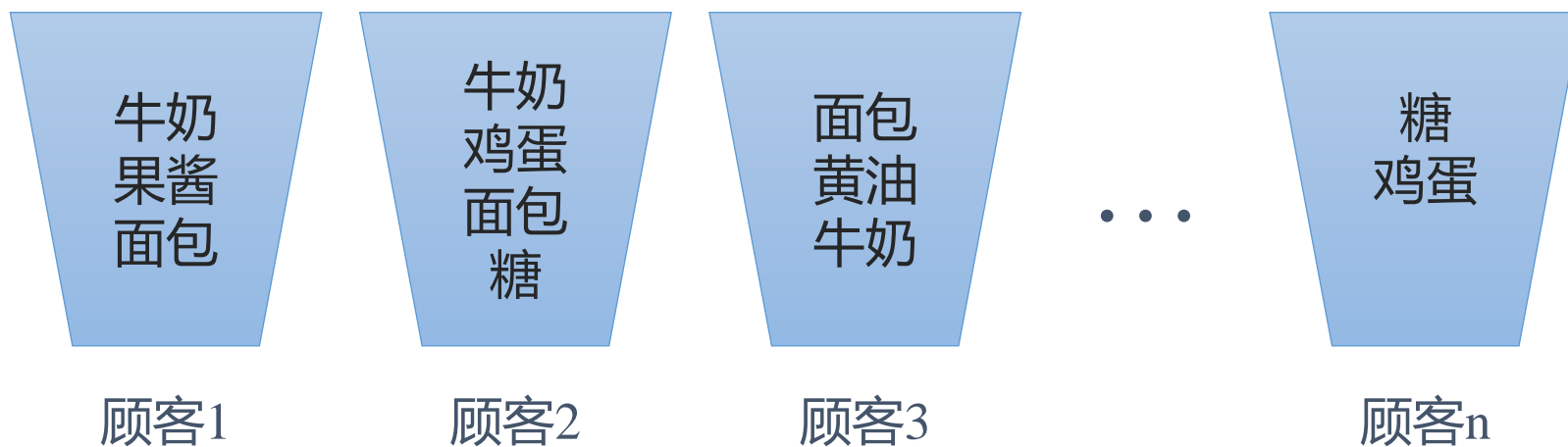
- 关联规则挖掘
- 非监督式机器学习-聚类
- 监督式机器学习
 - 离散标签预测-标签分类
 - 连续标签预测-数值预测 (回归)



关联规则挖掘



关联规则挖掘



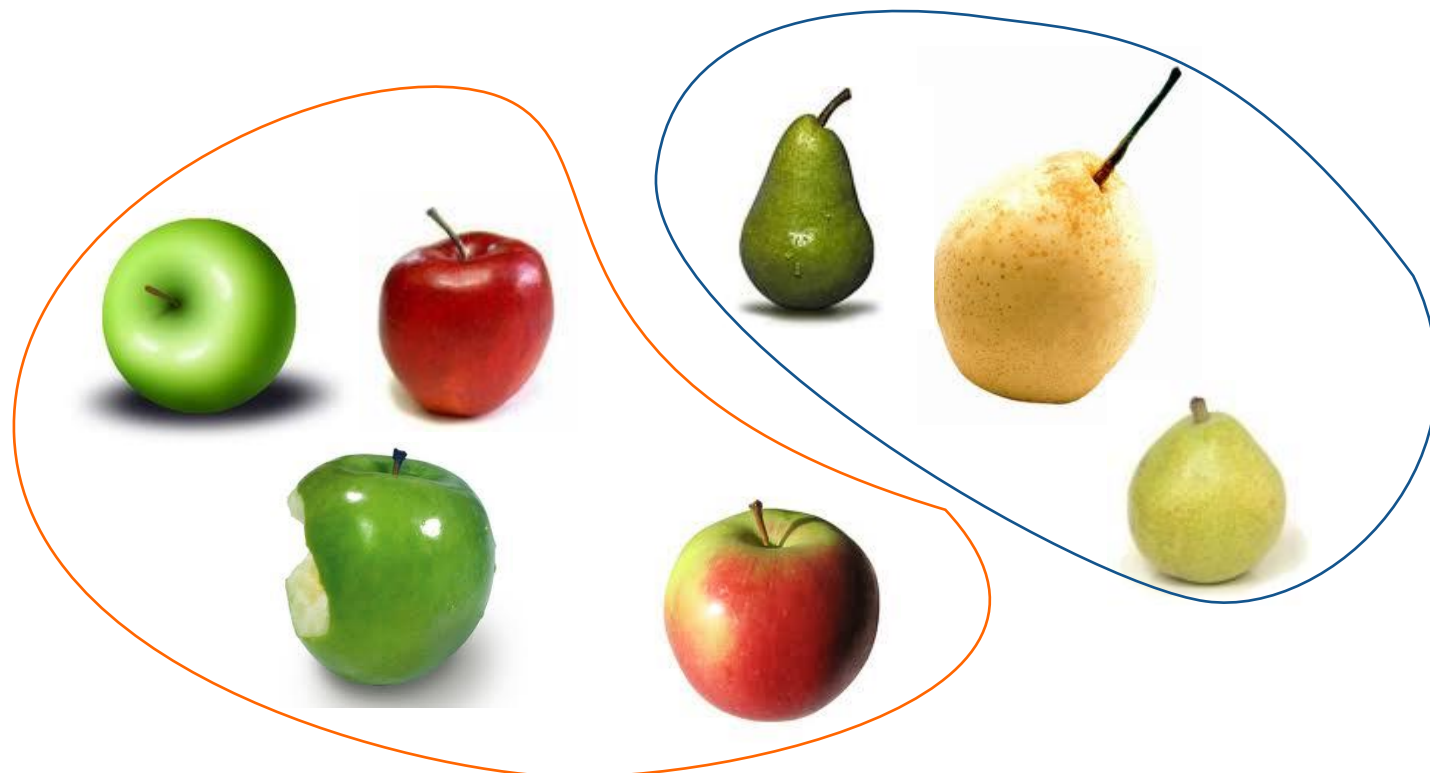
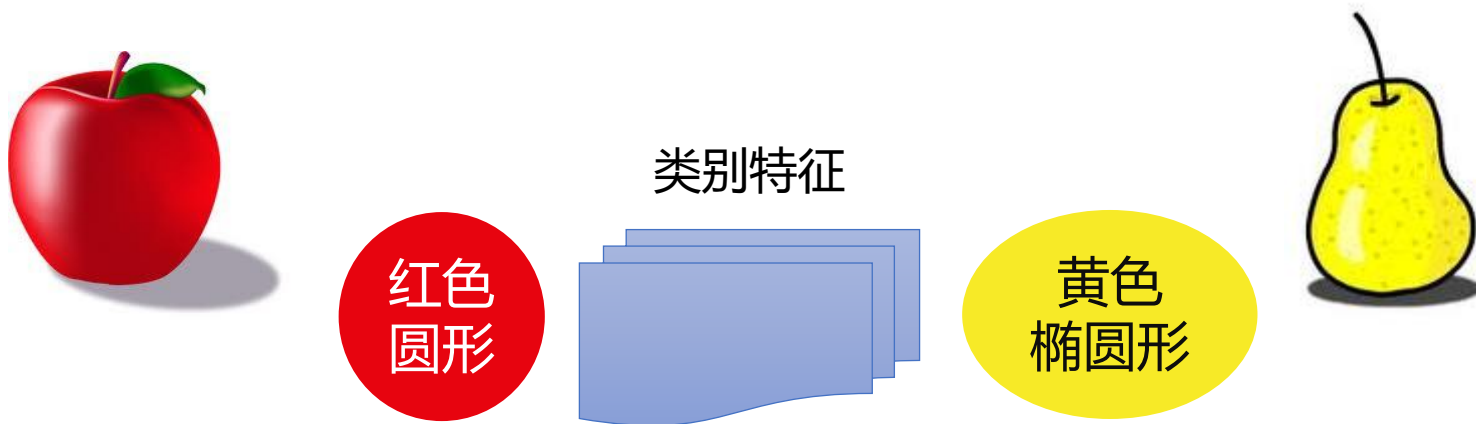
关联规则挖掘

数据库 TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

哪些商品被频繁购买？

分类



分类分析 —— 第一步：学习建模

姓名	年龄	收入	发展评估
汪明	<30	低	一般
王敏	<30	低	良好
李勇	30 ~ 40	高	良好
...

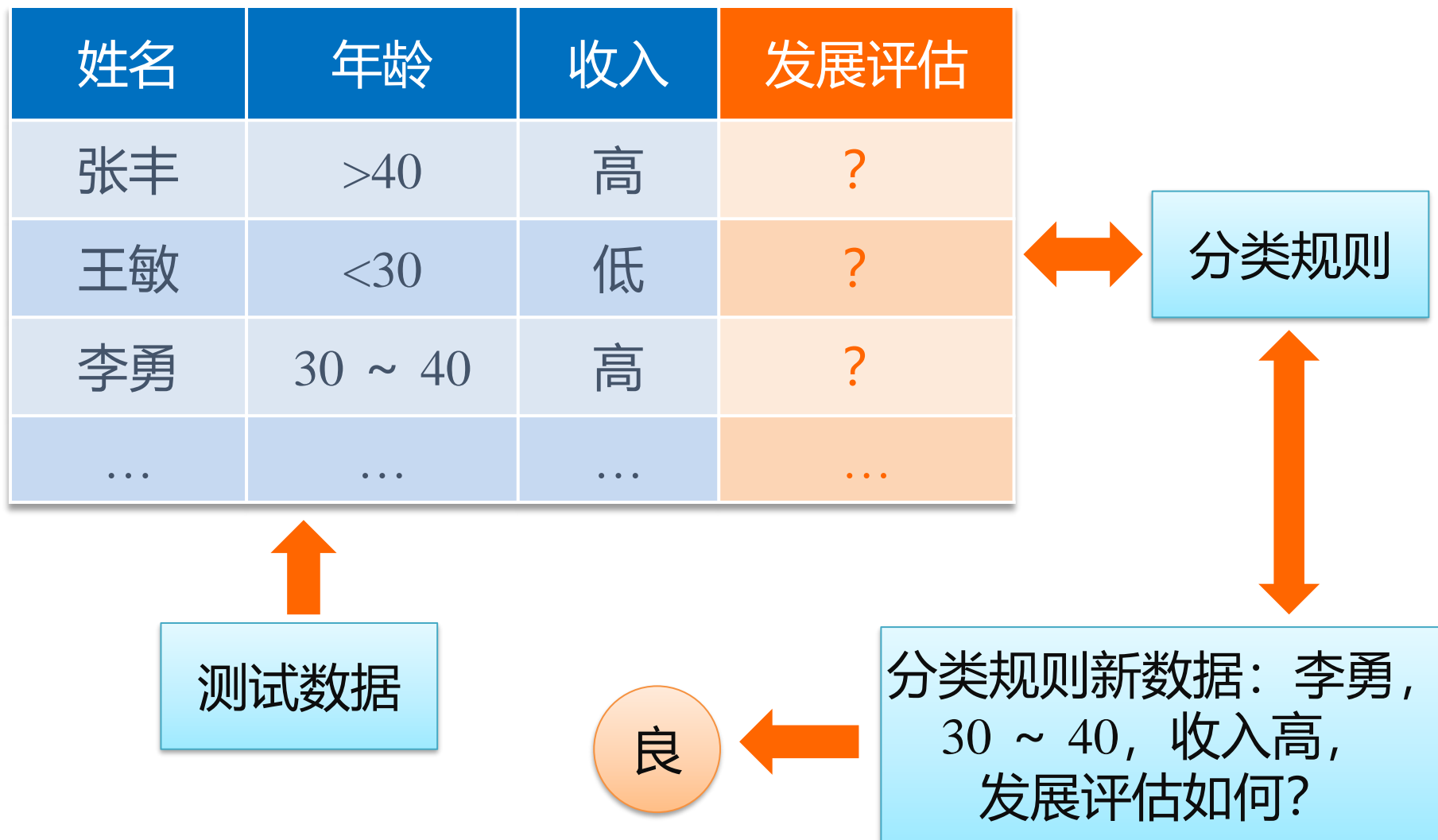
训练样本

分类算法

If age=30 to 40 and
income =高 则发展
评估=良好

分类规则

分类分析 —— 第二步：分类测试



数值预测 —— 第一步：学习建模

姓名	年龄	收入	发展评估值
汪明	<30	低	65
王敏	<30	低	74
李勇	30 ~ 40	高	78
...



训练样本

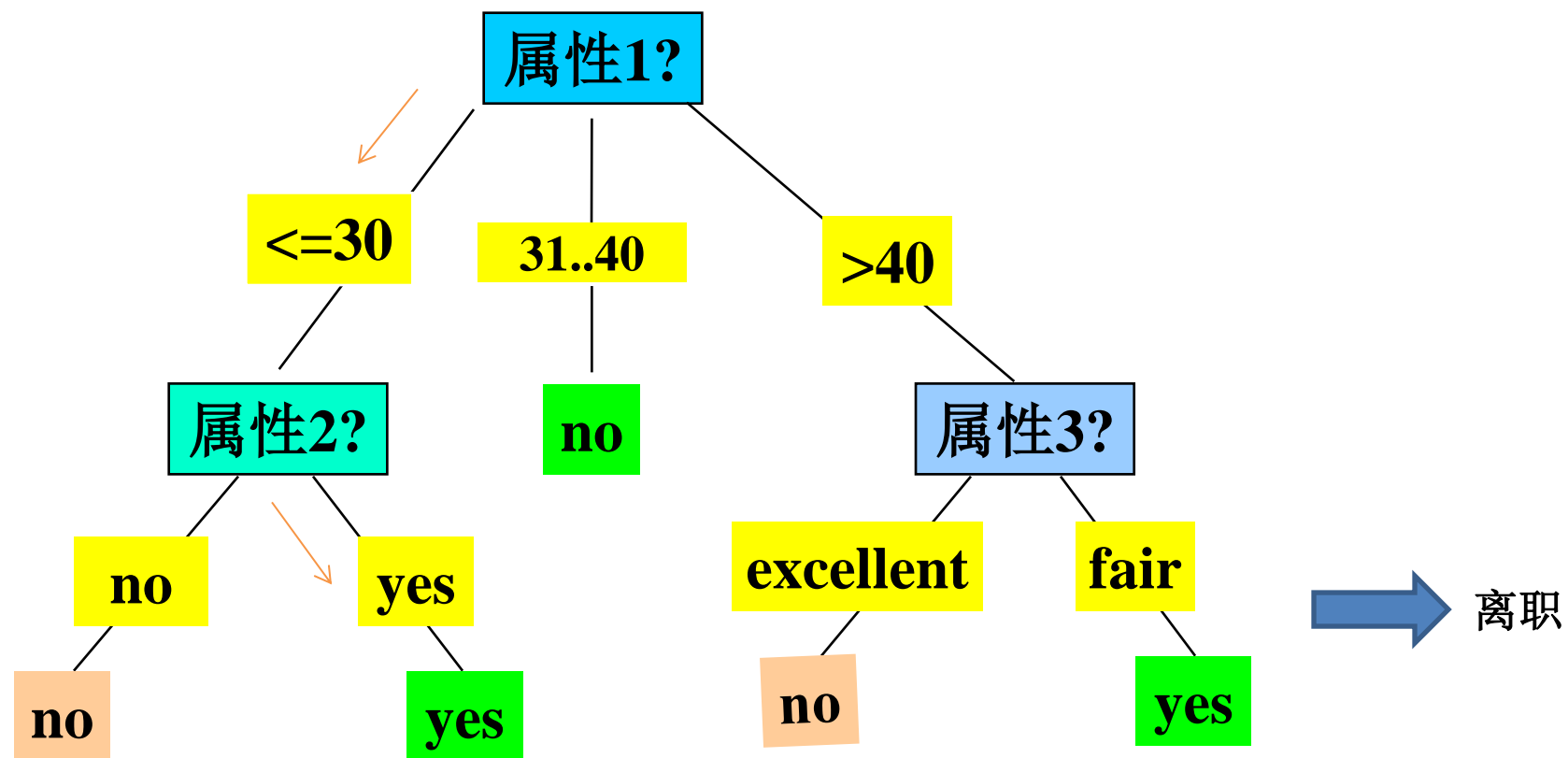
数值预测 —— 第二步：预测测试

姓名	年龄	收入	发展评估值
张丰	>40	高	?
王敏	<30	低	?
李勇	30 ~ 40	高	?
...



测试数据

分类案例-员工离职预测



数值预测案例

- 房价预测

房屋销售价格以及房屋的基本信息建立模型，来预测在此期间其他房屋的销售价格

销售日期	销售价格	卧室数	浴室数	房屋面积	停车面积	楼层数	房屋评分	建筑面积	地下室面积	建筑年份	修复年份	纬度	经度
20150302	545000	3	2.25	1670	6240	1	8	1240	430	1974	0	47.6413	-122.113
20150211	785000	4	2.5	3300	10514	2	10	3300	0	1984	0	47.6323	-122.036
20150107	765000	3	3.25	3190	5283	2	9	3190	0	2007	0	47.5534	-122.002
20141103	720000	5	2.5	2900	9525	2	9	2900	0	1989	0	47.5442	-122.138
20140603	449500	5	2.75	2040	7488	1	7	1200	840	1969	0	47.7289	-122.172
20150506	248500	2	1	780	10064	1	7	780	0	1958	0	47.4913	-122.318
20150305	675000	4	2.5	1770	9858	1	8	1770	0	1971	0	47.7382	-122.287
20140701	730000	2	2.25	2130	4920	1.5	7	1530	600	1941	0	47.573	-122.409
20140807	311000	2	1	860	3300	1	6	860	0	1903	0	47.5496	-122.279
20141204	660000	2	1	960	6263	1	6	960	0	1942	0	47.6646	-122.202
20150227	435000	2	1	990	5643	1	7	870	120	1947	0	47.6802	-122.298
20140904	350000	3	1	1240	10800	1	7	1240	0	1959	0	47.5233	-122.185
20140902	385000	3	2.25	1630	1598	3	8	1630	0	2008	0	47.6904	-122.347
20150413	235000	2	1	930	10505	1	6	930	0	1930	0	47.4337	-122.329
20140930	350000	3	1	1300	10236	1	6	1300	0	1971	0	47.5028	-121.77
20150507	1350000	4	1.75	2000	3728	1.5	9	1820	180	1926	0	47.643	-122.299
20140530	459900	3	1.75	2580	11000	1	7	1290	1290	1951	0	47.5646	-122.181
20140723	430000	6	3	2630	8800	1	7	1610	1020	1959	0	47.7166	-122.293
20141003	718000	5	2.75	2930	7663	2	9	2930	0	2013	0	47.5308	-122.184

数值预测案例

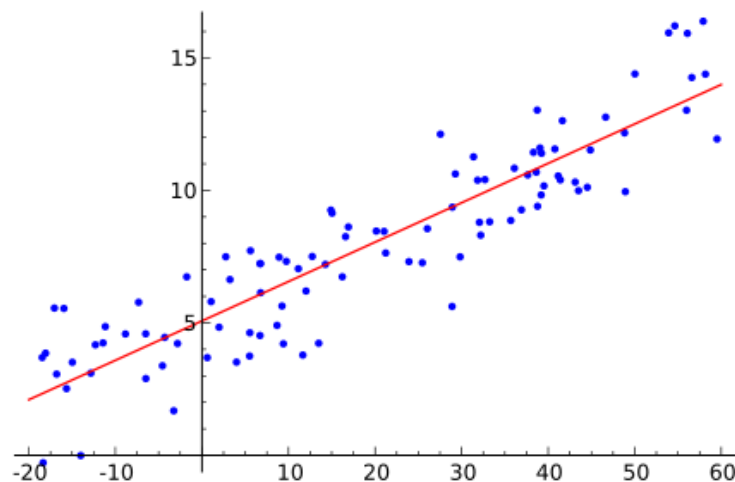
- 房价预测

房屋销售价格以及房屋的基本信息建立模型，来预测在此期间其他房屋的销售价格

$$y = \beta_0 + \beta_1 x$$

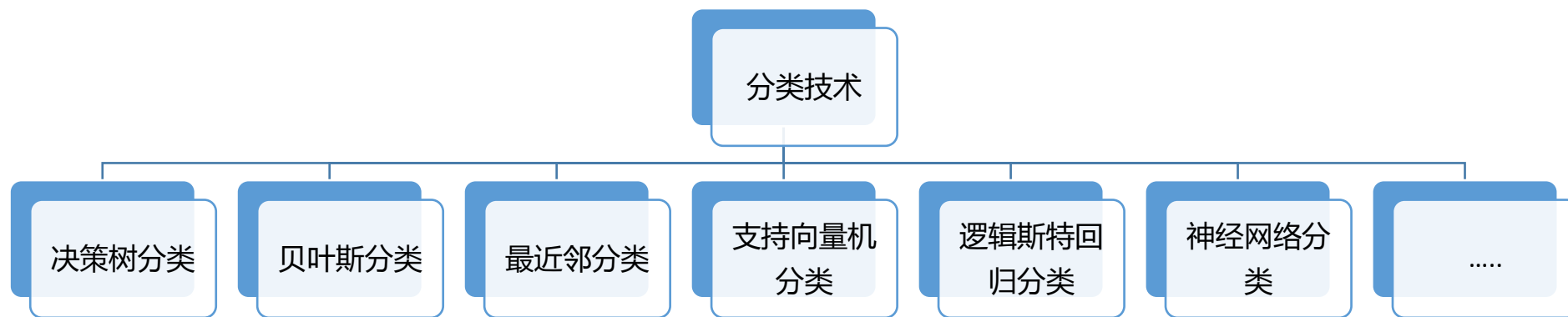
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

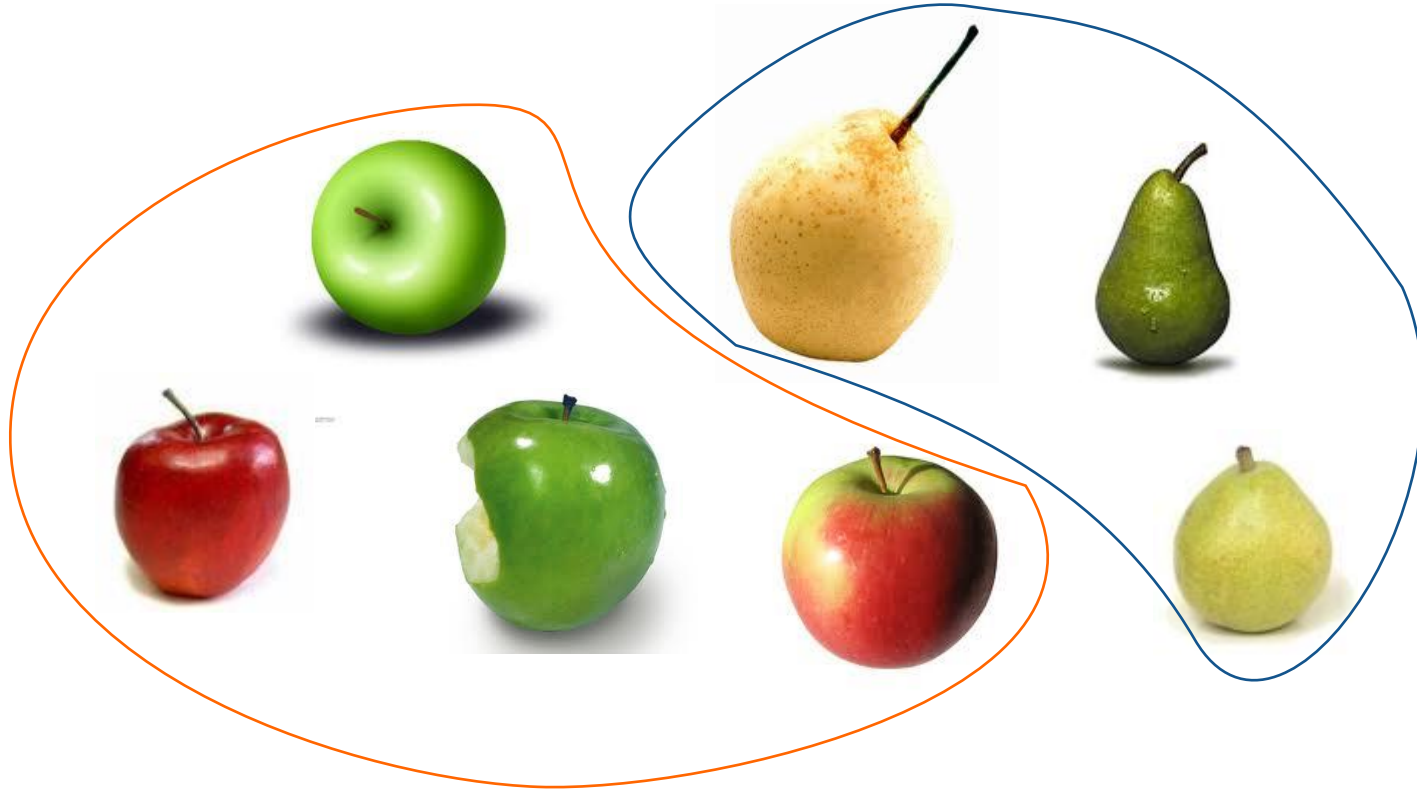


$$y = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

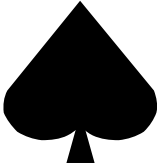

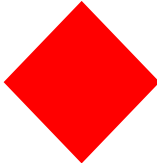

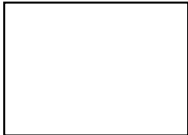
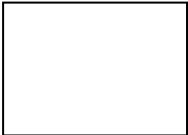
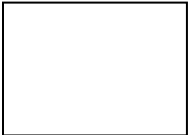




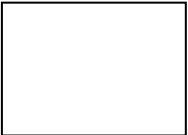



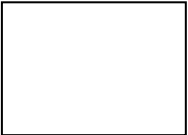




分类技术



聚类

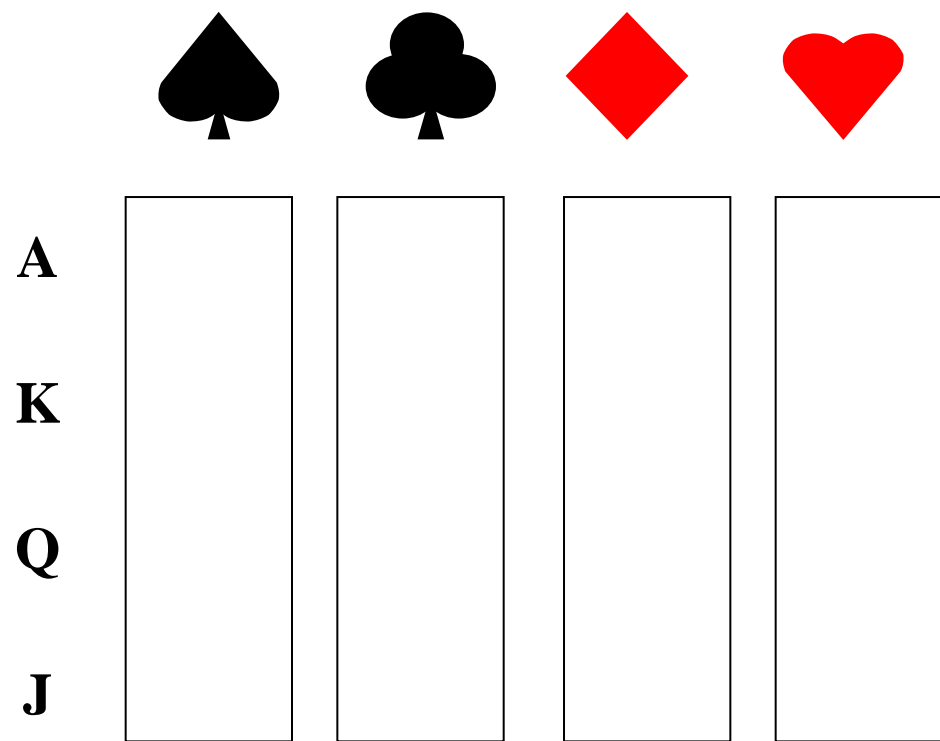


聚类分析原理介绍

				
A				
K				
Q				
J				

以一定的准则将相似的物体聚在一起



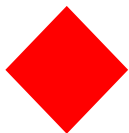

聚类分析原理介绍



花色相同的牌

以**一定的准则**将相似的物体聚在一起

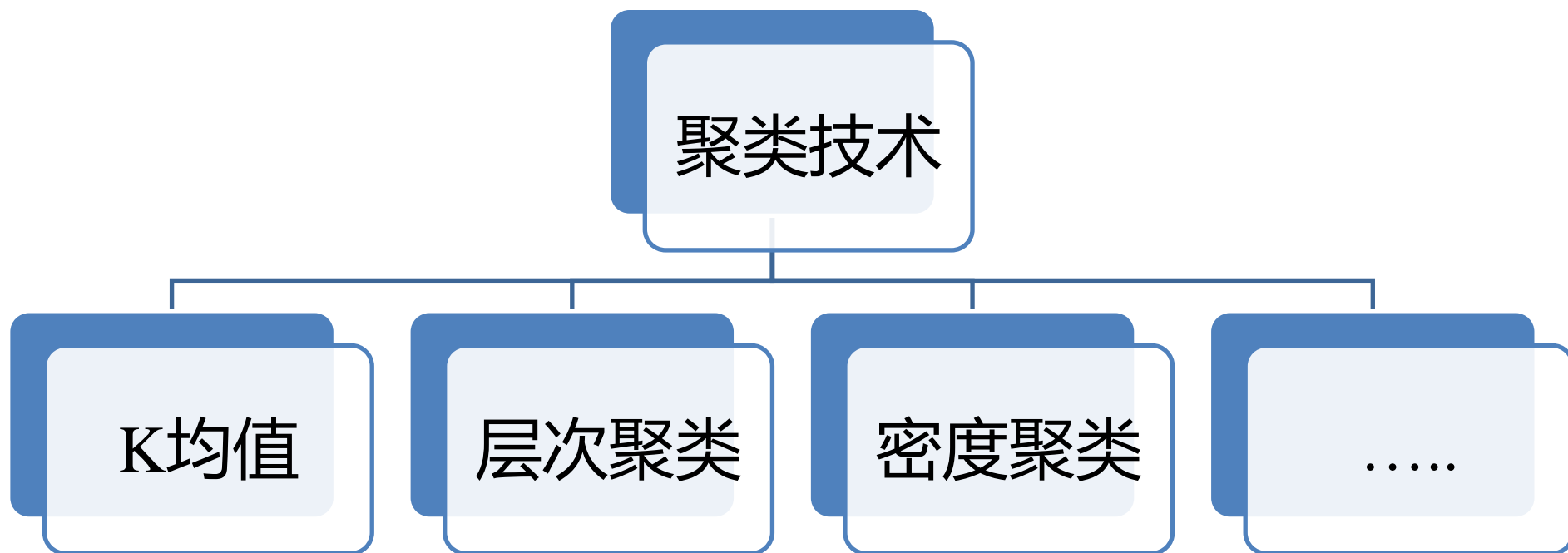
聚类分析原理介绍

				
A				
K				
Q				
J				

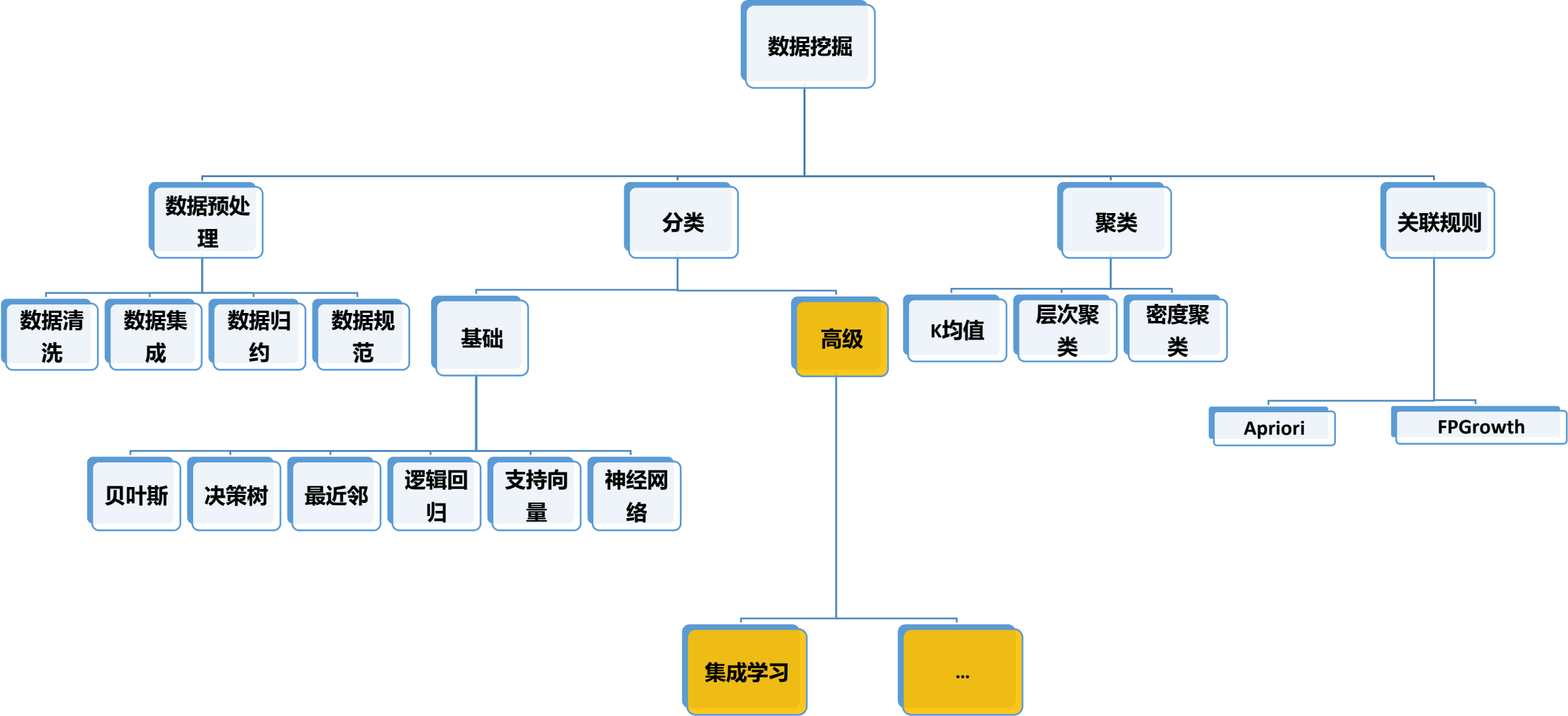
符号相同的牌

以**一定的准则**将相似的物体聚在一起

聚类分析技术



数据挖掘的主要内容





南京大學
NANJING UNIVERSITY

目录

01

概述

02

数据挖掘的内涵

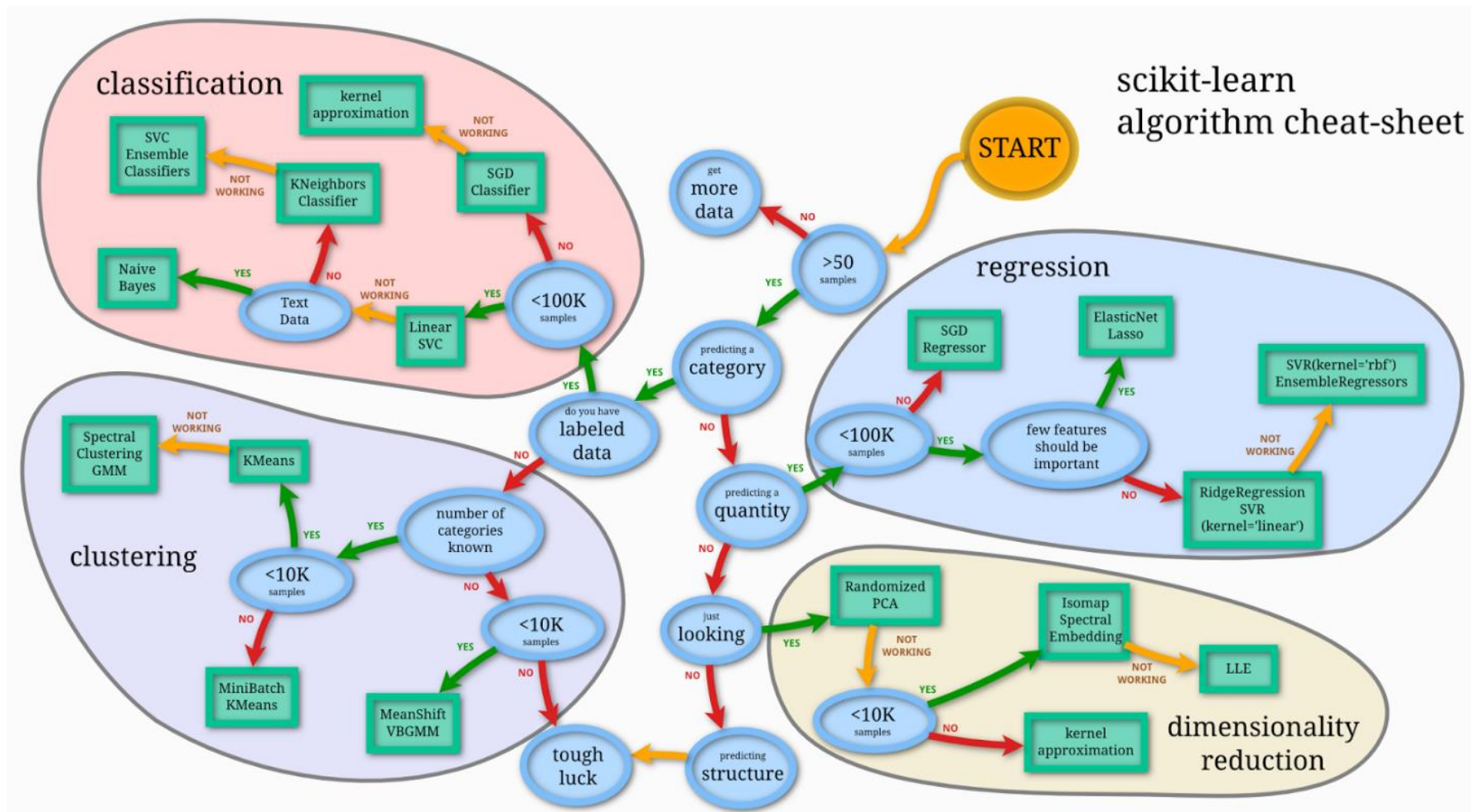
03

数据挖掘的内容

04

相关资源

Python代码资源



scikit-learn 是针对 Python 编程语言的**免费软件机器学习库**。它具有各种**分类、回归和聚类**算法，包括支持向量机，随机森林，梯度提升，k均值等。除此之外，还可以**比较、验证和选择参数及模型**，**数据预处理**等。

<https://scikit-learn.org/stable/>

数据分析三剑客

NumPy 是一个支持大型多维数组和矩阵的 Python 库，并提供了大量的数学函数来操作这些数组。它为 Python 带来了高效的数值数据处理能力，是科学计算的基础包。

Pandas 是基于 NumPy 构建的一个数据分析库，它提供了大量适合动态表格数据（如股票价格或实验数据）的数据结构和数据分析工具。

Matplotlib 是一个全面的库，用于在 Python 中创建静态、动画和交互式可视化。（柱状图，饼状图，折线图，圆饼图等）



开发环境：Anaconda



Anaconda，是一个开源的Python发行版本，其包含了Python解译器以及许多常用的数据科学工具包。

<https://www.anaconda.com/>

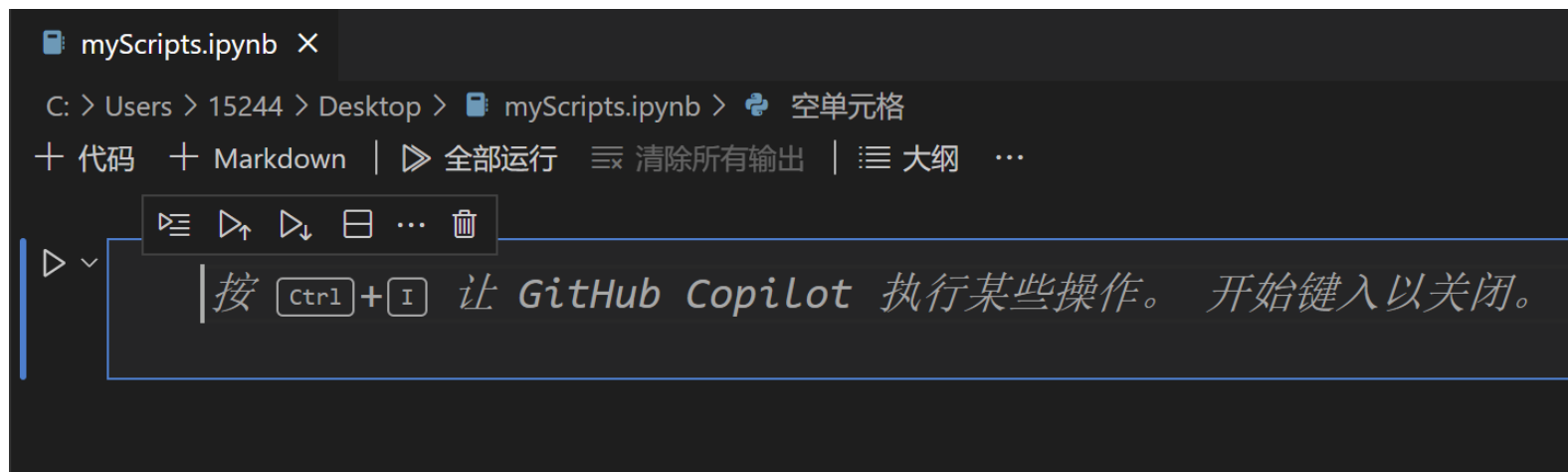
为什么选择 **Anaconda**?

- 方便安装
- 包管理器
- 环境管理
- 集成工具和库
- **Jupyter** 笔记本
- **Spyder** 集成开发环境
- 跨平台性
- 社区支持

根据自己电脑操作系统下载对应版本安装

编程工具：Jupyter Notebook

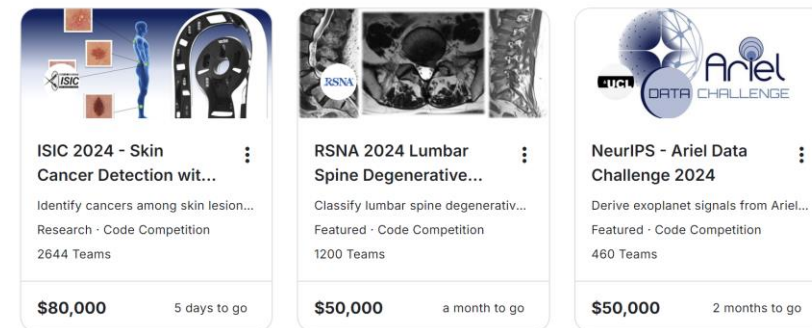
Jupyter是一个开源的交互式计算环境,它允许用户通过 Web 浏览器创建和分享包含代码、方程式、可视化和叙述性文本的文档。Jupyter 支持多种编程语言,如 Python、R、Julia 等,使其成为数据分析、科学计算和机器学习领域的强大工具。



数据/竞赛资源

- Kaggle数据建模与分析竞赛平台

<https://www.kaggle.com/competitions>



- 阿里的天池大数据竞赛

<https://tianchi.aliyun.com/competition/gameList/activeList>



- 百度人工智能竞赛

<https://aistudio.baidu.com/aistudio/competition>



千言数据集：段落检索评测 [进行中](#)

提供大规模的段落检索评测集，挑战机器检索相关段落的能力

标签：信息检索 比赛时间：2022/07/22 - 2025/07/24

举办方：

●