数据挖掘导论
Introduction to Data Mining

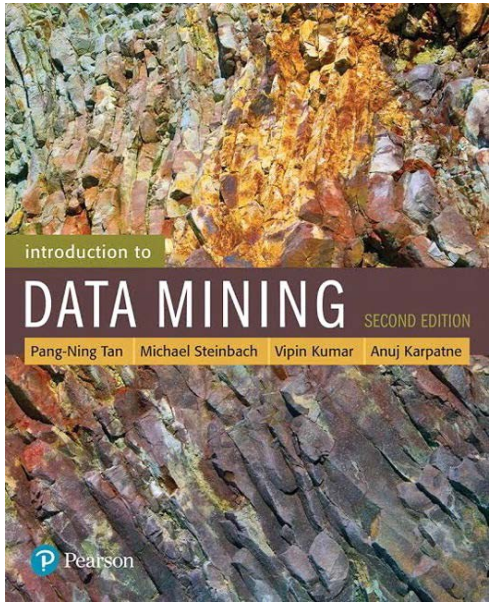
第一章：绪论

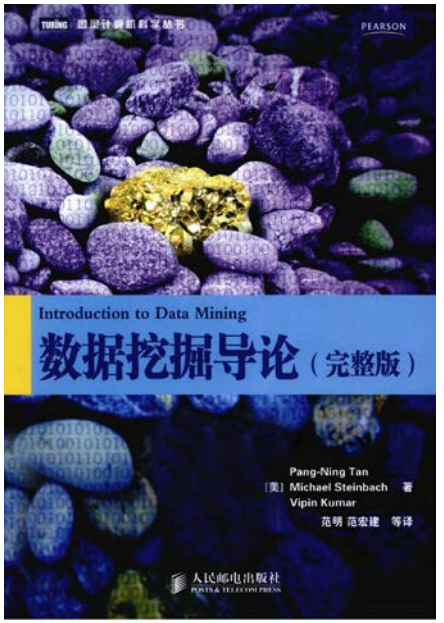
王浩
Email：haowang@szu.edu.cn

深圳大学 SHENZHEN UNIVERSITY
电子信息与工程学院

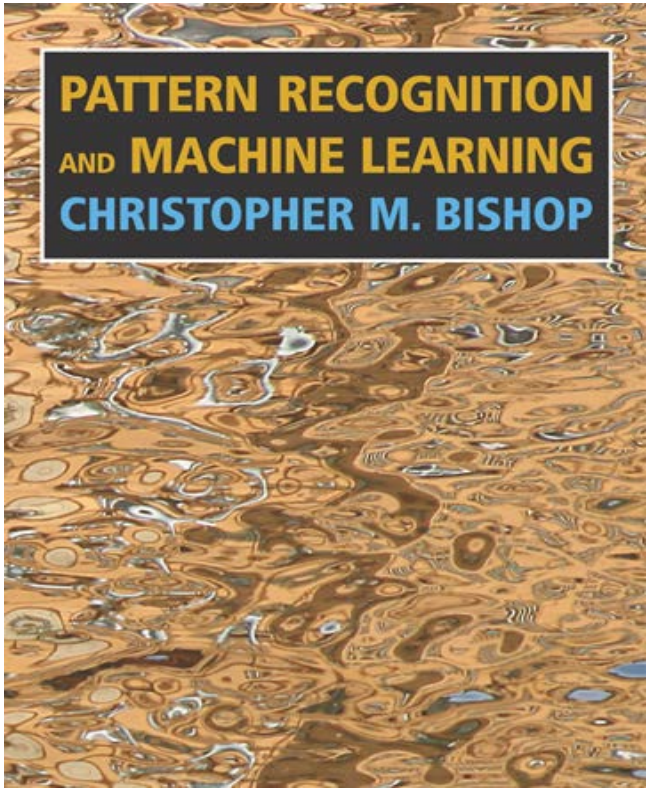
参考教材

- 数据挖掘导论（Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley ）

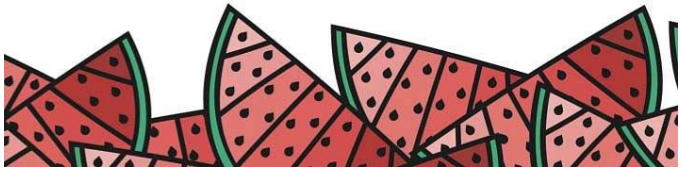




课外读物


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


周志华 著
MACHINE
LEARNING
机器学习
清华大学出版社



目标：

- 大数据背景简介
- 数据挖掘是什么？
- 为什么要进行数据挖掘？
- 数据挖掘有哪些任务？
- 机器学习是什么？
- 数据挖掘与机器学习之间有什么关系？
- 数据挖掘实战举例
- 数据挖掘领域重要会议

大数据——产生背景

- 计算机互联网时代的发展。
  - 数据记录逐步脱离了纸笔的限制，大量的数据可以按0或1的二进制方式存储半导体材料内，大数据热潮诞生的先决条件是计算机存储能力的迅速扩大和成本的一再降低。

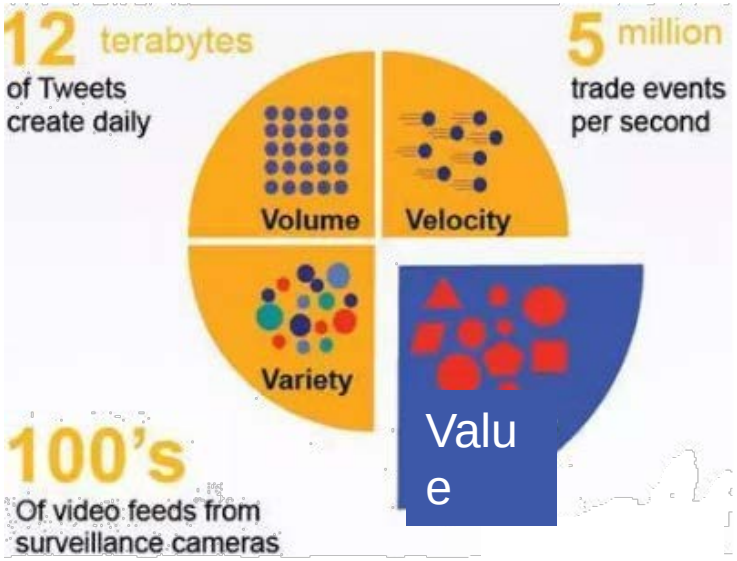
- 数据的概念的变化。
  - 传统的数据是指用数字或文字描述的内容，统称为结构化数据，而大数据时代涌现出了大量新型数据的、非结构化的数据。例如人群之间看不见的社交关系（Social Relationships），移动设备发射的GPS位置，网络传播的图像、视频信号，可穿戴设备采集的健康数据等。对这些各种各样的数据的采集、挖掘、运用，也是现代大数据挖掘的重要研究课题。

- 人类创造数据的能力也同样在高速增长。

- 我们对数据进行挖掘和处理的能力的提升。
  - 这些IT技术在数据产生、存储、挖掘、运用方面的逐步成熟，让数据驱动产生价值的门槛越来越低，终于大数据时代的脚步匆匆到来了。

大数据——4V要素

- **Volume**：具备超出典型数据库软件收集、存储、管理和分析能力的数据集；
- **Variety**：具备多样性的，结构化、半结构化、非结构化等多种类型的数据形式；
- **Velocity**：具备快速、实时的数据处理能力；
- **Value**：具备从稀疏的数据中挖掘高价值内容的意义。



大数据——价值

数据容量、速度、多样性、复杂度方面在今天来看无法想象的事情，几年之后都将完全被颠覆；唯一不变的，是对数据的思考和分析的方法，和利用数据来产生价值的出发点。

1980s年代及以前，企业的各类业务、财务数据都是通过账簿记录，这种方式查阅和统计的效率都很低，可靠性也不高。

1990s年代末开始，金融业、电信业、大型零售等行业企业率先将核心交易数据电子化

2000年以后随着IT技术的进步，越来越多的企业将信息化纳入议程，ERP（Enterprise Resource Planning）、MIS（Management Information System）系统蓬勃发展，设计、制造、进存销等业务管理数据化，这些数据被大家意识到是企业最宝贵的资产，随之而起的统计报表技术也渐渐完善.

2010年以后，更多种类的数据，包括客户的浏览数据、反馈数据等在一些企业中也都开始记录并逐步进行个性化建模和分析，数据驱动的CRM（Customer Relationship Management）客户关系管理开始在精准运营和个性化服务方面崭露头角，基于数据分析的预测技术也逐步开始出现。

在企业管理中数据价值的提升

大数据——采集

- 大数据时代的数据采集有如下三个特点：

1）数据采集以自动化手段为主，要尽量摆脱人工录入的方式；
2）采集内容以全量采集为主，要摆脱对数据进行采样的方式；
3）采集方式多样化、内容丰富化，摆脱以往只采集基本数据的方式。

- 常见数据采集技术

传统的数据采集方法	面向移动设备的数据采集技术	网络爬虫	物联网	传感器
• 人工录入 • 调查问卷 • 电话随访 • ……	• Android或iOS的采集SDK（Software Develop Kit）	• 进行大规模全网信息采集 • 舆情监控 • 竞品分析	• 无线射频标签（RFID） • 物品信息与互联网实现自动连接	• 携带传感器+大数据平台的智能设备 • 智能医疗，智慧城市等

大数据——存储

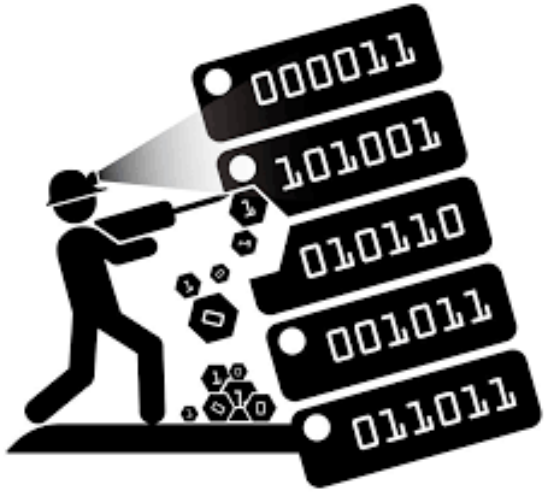
- 随着数据量的惊人增长，已经使用了20余年的传统数据库再也无法支撑起新的存储需求了，所以被Google称为Big Table和Google File System(GFS）的新型存储技术在过去的几年里被发明出来，并在行业中广泛应用，这些技术通过自动调配上万台服务器协同工作，能完成高性能和高可靠的数据存储任务，为大数据的运用铺平了道路。

- 云计算可谓是大数据的最好载体。由于大数据存储和运算非常复杂，传统企业在运作时需要投入很高的人力物力，因此把涉及存储运算的基础设施抽象和独立出来，形成的专门性服务称为云计算（Cloud Computing）。云计算就好比大数据时代的“电”，大数据系统则是“家用电器”——云计算注重服务的通用性，大数据关注实际的用途和效果。

什么是数据挖掘：

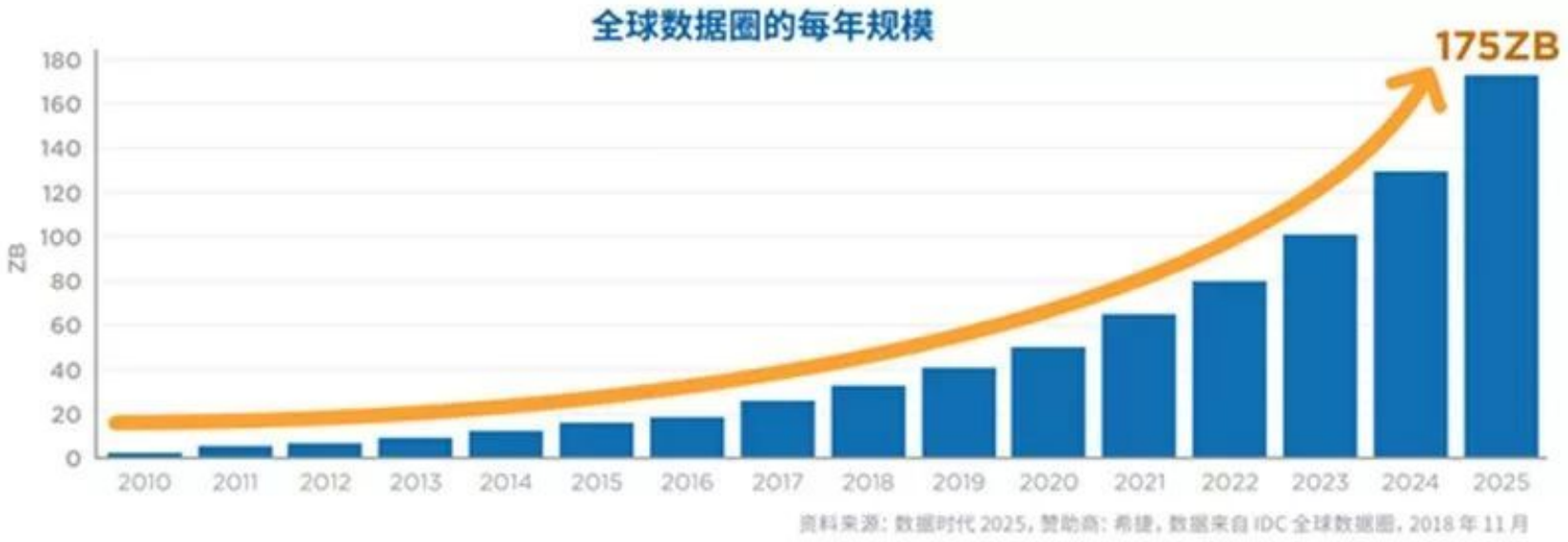
- 初步定义：数据挖掘是从大型数据存储库中，自动地发现有用信息的过程。
- 并非所有的信息发现任务都是数据挖掘，简单的检索、查找并不能称之为信息挖掘。

- 以下哪些属于数据挖掘？
  - 依据性别划分公司的顾客；
  - 根据可盈利性划分公司的顾客；
  - 计算公司的销售总额；
  - 预测掷一对骰子的结果；
  - 使用历史记录预测公司未来股票价格。



*: IDC发布《数据时代2025》报告

为什么要进行数据挖掘：

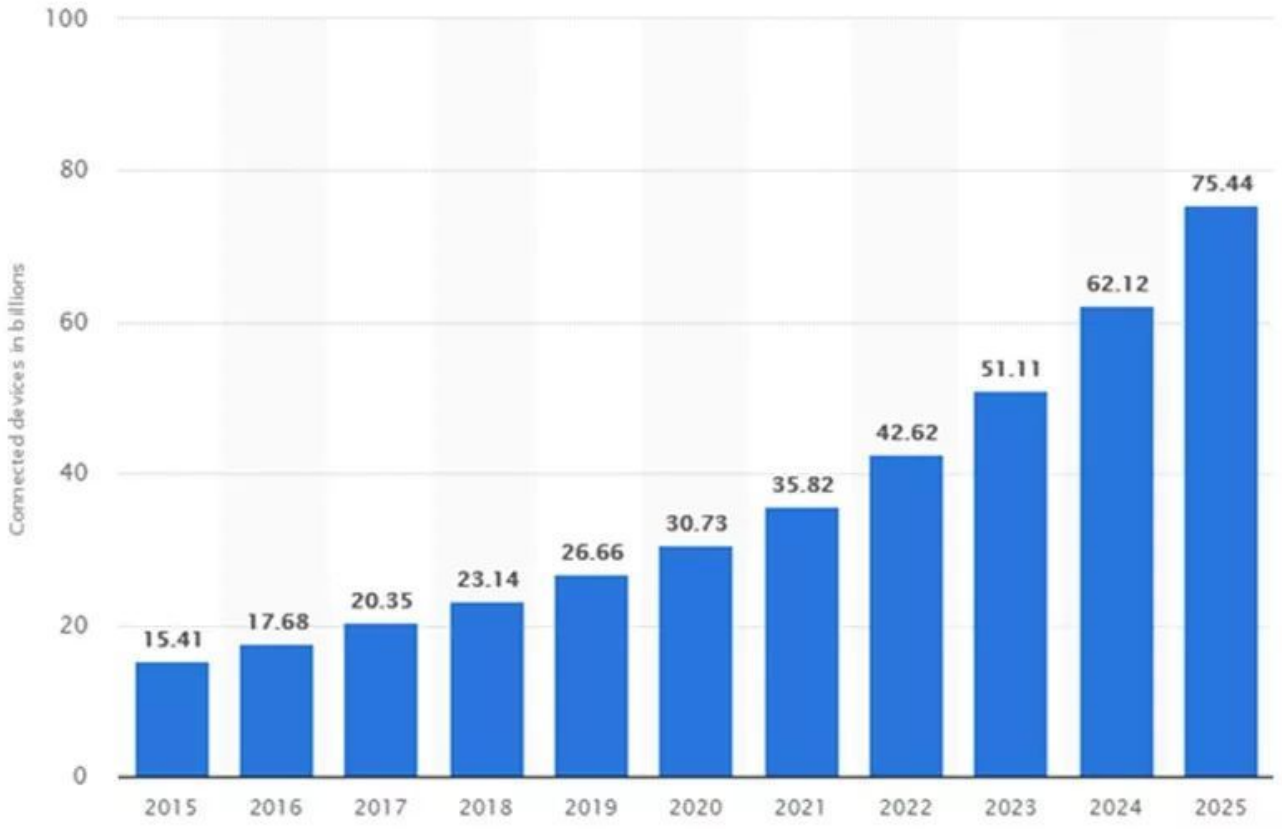


- 2025年，全球每年产生的数据将从2018年的33ZB增长到175ZB，相当于每天产生491EB的数据*；以25MB/秒的速度，要下载完这175ZB的数据，需要2.3亿年；
- 人的大脑细胞数超过全世界人口总数2倍多，每天可处理8600万条信息，其记忆贮存的信息超过任何一台电子计算机。

1GB (Gigabyte 吉字节 又称“千兆”)=1024MB；1TB (Trillionbyte 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。

*: IDC发布《数据时代2025》报告

为什么要进行数据挖掘：



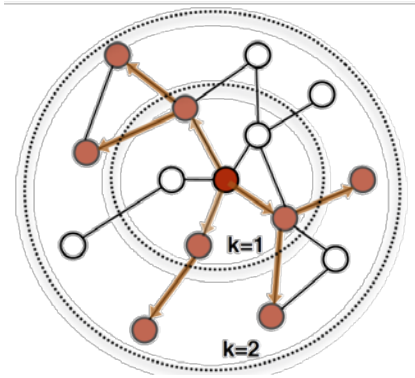
- 据HIS的数据预测，到2025年，全球物联网（IoT）连接设备的总安装量预计将达到754.4亿，约是2015年的5倍；
- 人的平均数量约为37.2万亿个细胞，其中大部分都可以进行信息采集或传递，人的大脑细胞数140亿个左右，完成各类信息的筛选整合和处理。

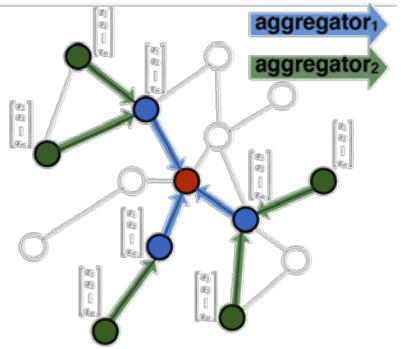
1. IDC发布《数据时代2025》报告

为什么要进行数据挖掘：

3 基于支付网络的用户影响力评价项目 **05/2019-11/2019 Tencent**

- 研究解决基于微信支付网络（10亿节点，千亿边）的用户影响力评价问题。
- 使用Scala、SQL、Python实现了leaderrank、graphsage等多个算法进行影响力评价。
- 项目输出的用户影响力评分作为用户特征已上线使用，帮助多项业务取得30%-50%的效果提升。

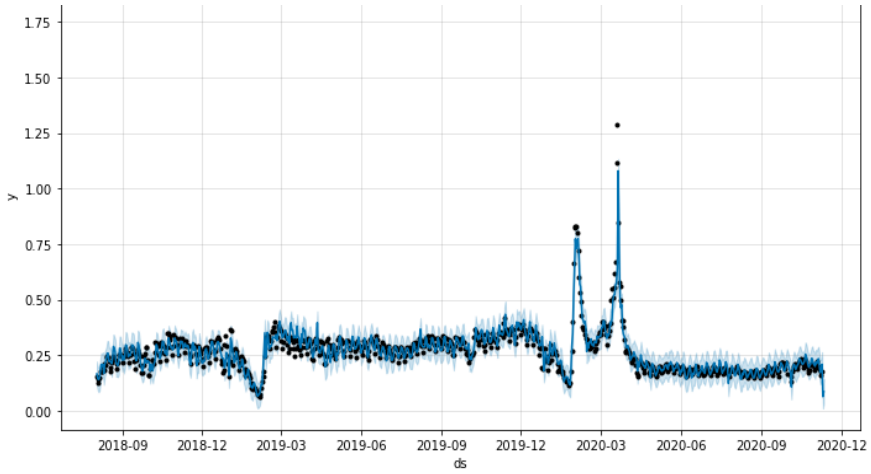


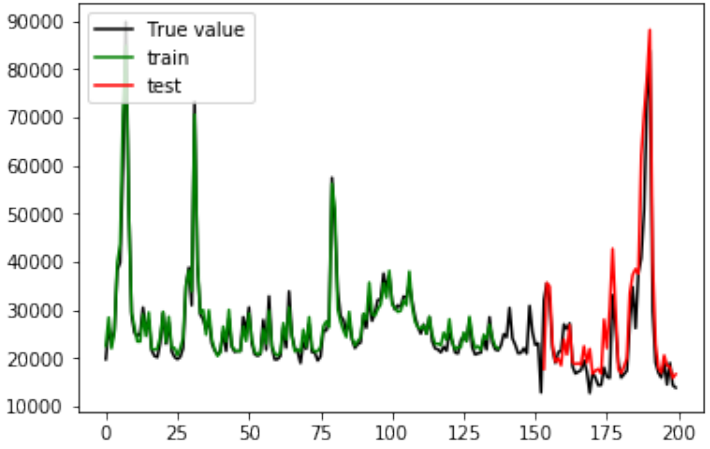


为什么要进行数据挖掘：

4 跨境交易预测（时间序列预测问题） **08/2020-03/2021 Tencent**

- 对微信线上跨境交易场景进行建模，将换汇问题转化为交易量和汇率的预测问题。
- 整合fbprophet、AutoETS等统计模型与LSTNet、TCN、TPA-LSTM等深度学习模型构建时间序列预测框架。
- 基于特征工程和算法改进不断提升预测精度，目前交易量预测误差<5%，已满足模型上线需求。上线后，每日可为企业挽回几百万汇率波动带来的损失。

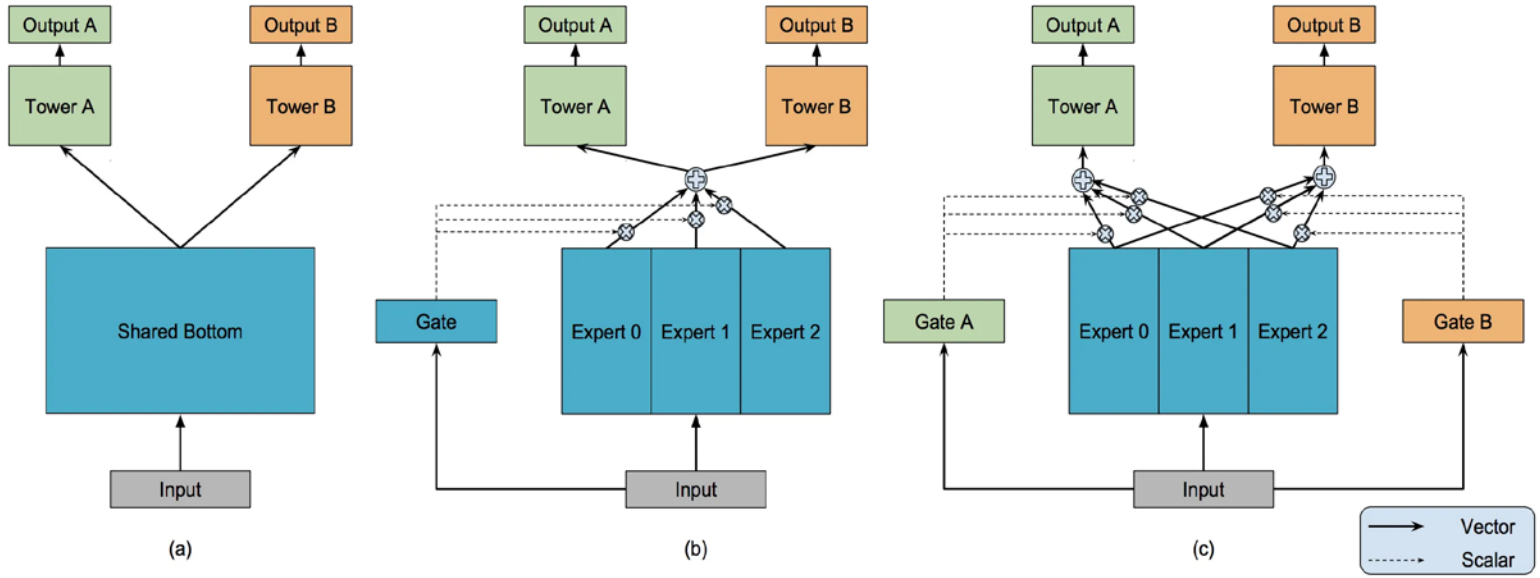




为什么要进行数据挖掘：

5 数据挖掘与分析 **11/2019-03/2021 Tencent**

- 信用卡还款业务流失用户预测、潜力分期用户挖掘
- 基于多任务的浮收基金推荐项目
- 数据分析专项



为什么要进行数据挖掘：

数据层面：
- 原始数据过于庞大
- 原始数据往往无法直观无法理解；

商业层面：
- 数据挖掘可以产生经济效益；

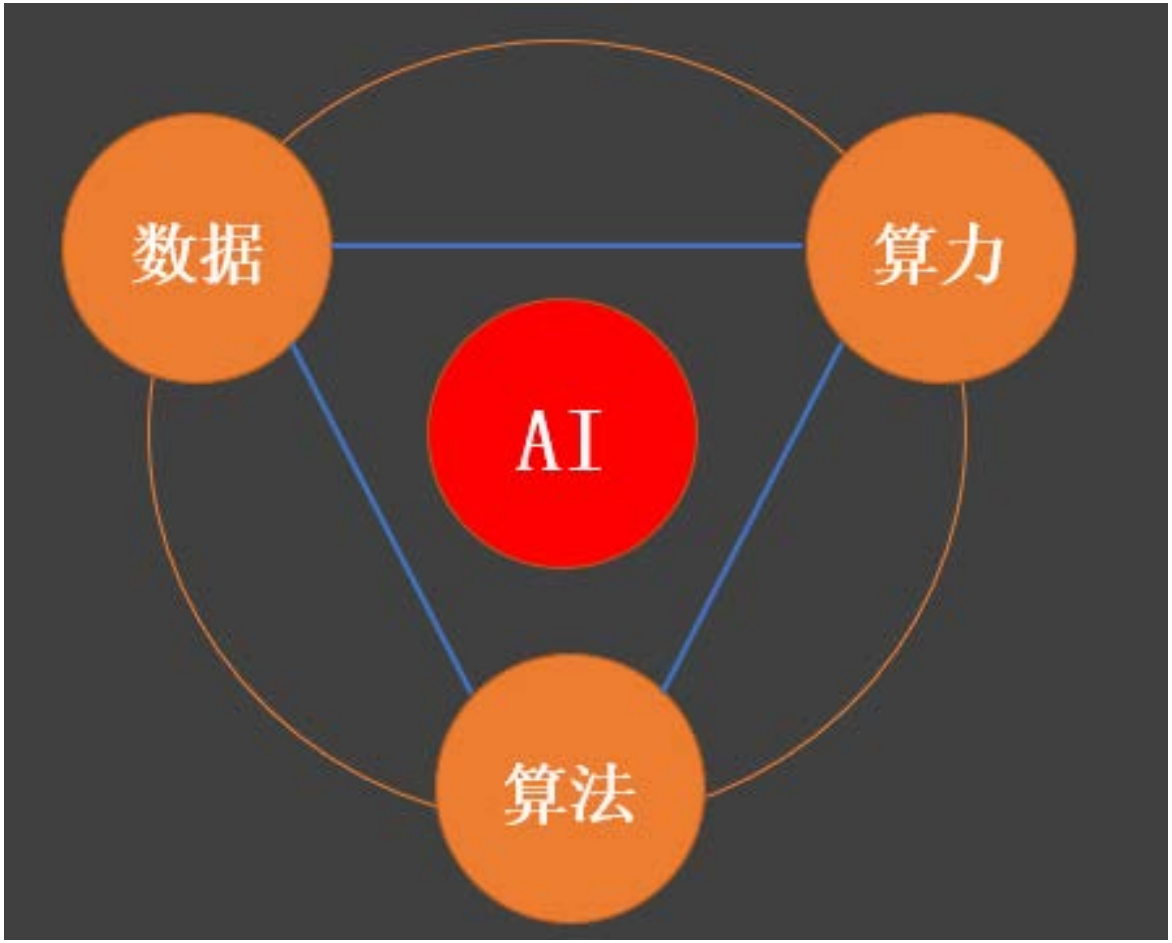
  --例如：Facebook、Google、Amazon、Tencent、Baidu等

  --能够从数据中提取有用的信息是商业化开发的关键。

科学层面：
- 通过数据去发现和感知世界，其乐无穷；

个人层面：
- 这个技能可以养活自己。

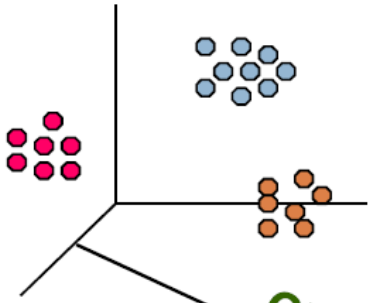
相关学科：



相关学科：

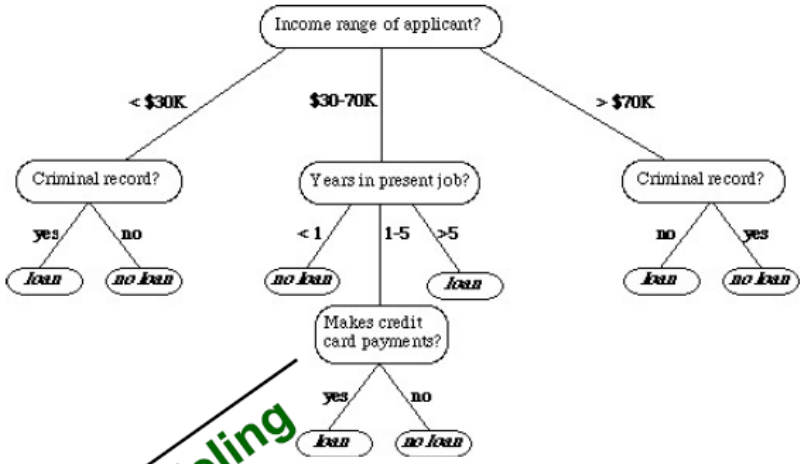
与数据挖掘相关的学科

数据挖掘任务：



Clustering

聚类分析



Predictive Modeling

预测建模

Data

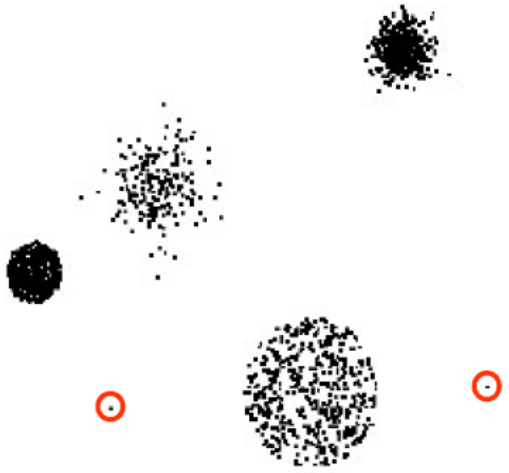
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

关联分析

Association Analysis

异常检测

Anomaly Detection



////2017

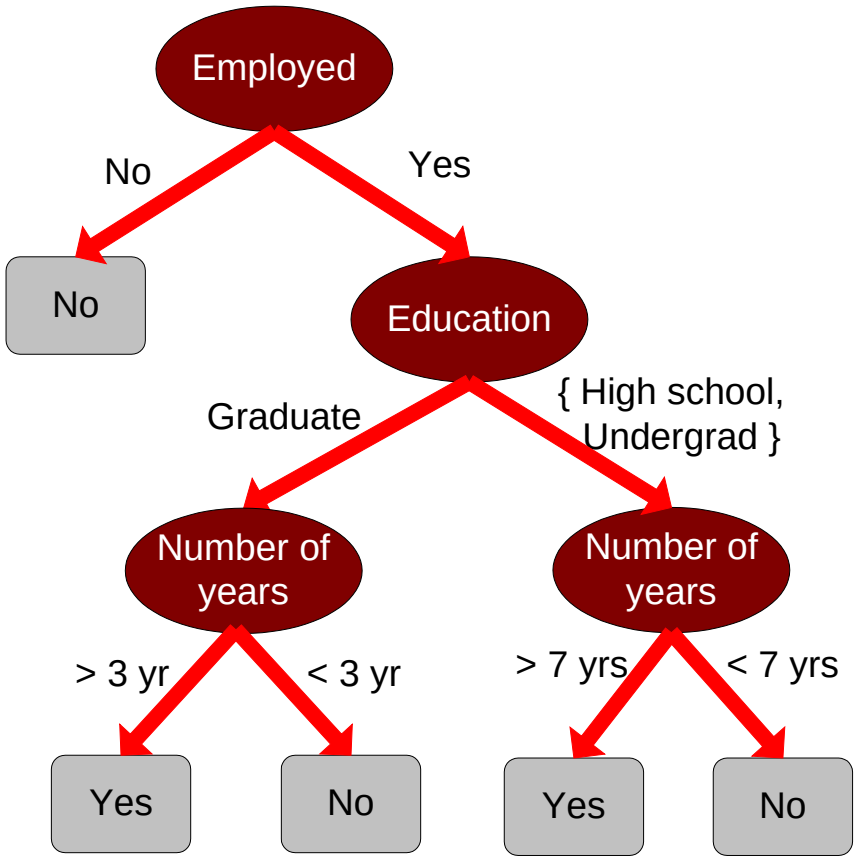
任务1：预测模型-分类

- 构建模型基于样本的属性，预测其类别，属于监督学习的一种。

类别

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

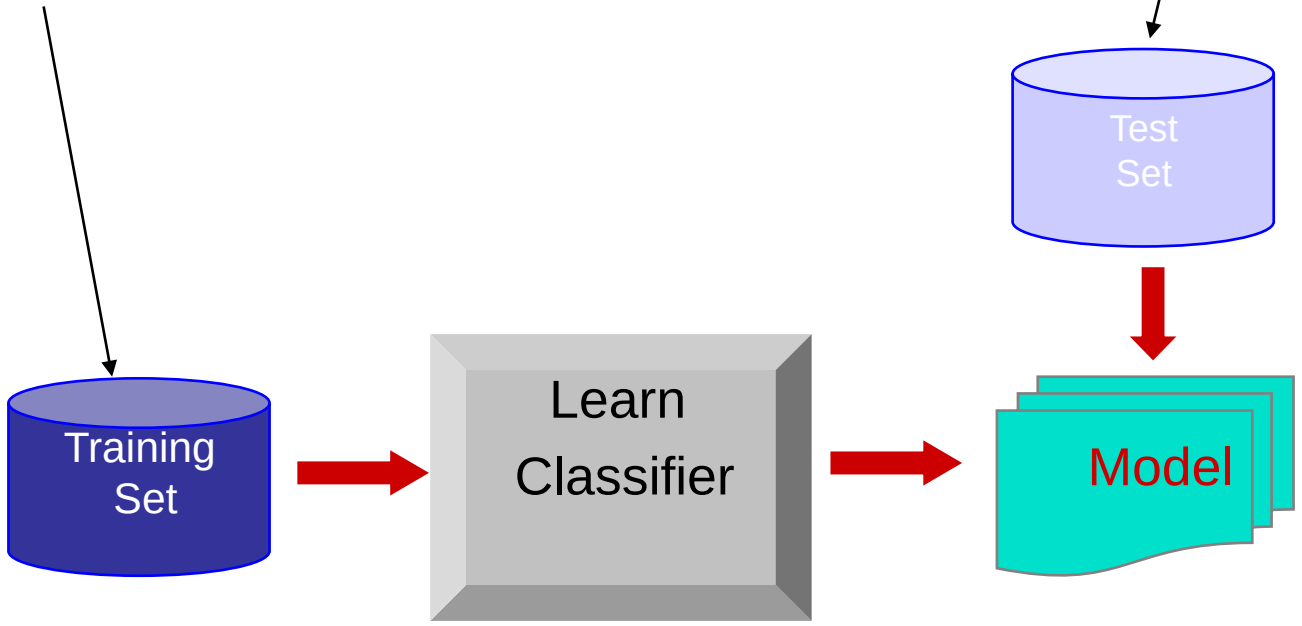
用户信用价值预测



任务1：预测模型-分类

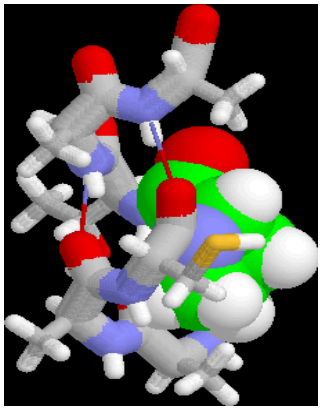
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



任务1：预测模型-分类

- 将信用卡交易归类为合法或欺诈
- 垃圾邮件甄别
- 利用卫星数据对土地覆盖进行分类（水、城市、森林等）
- 把新闻分为财经、天气、娱乐、体育等
- 肿瘤细胞良恶性的预测
- 将蛋白质的二级结构分类



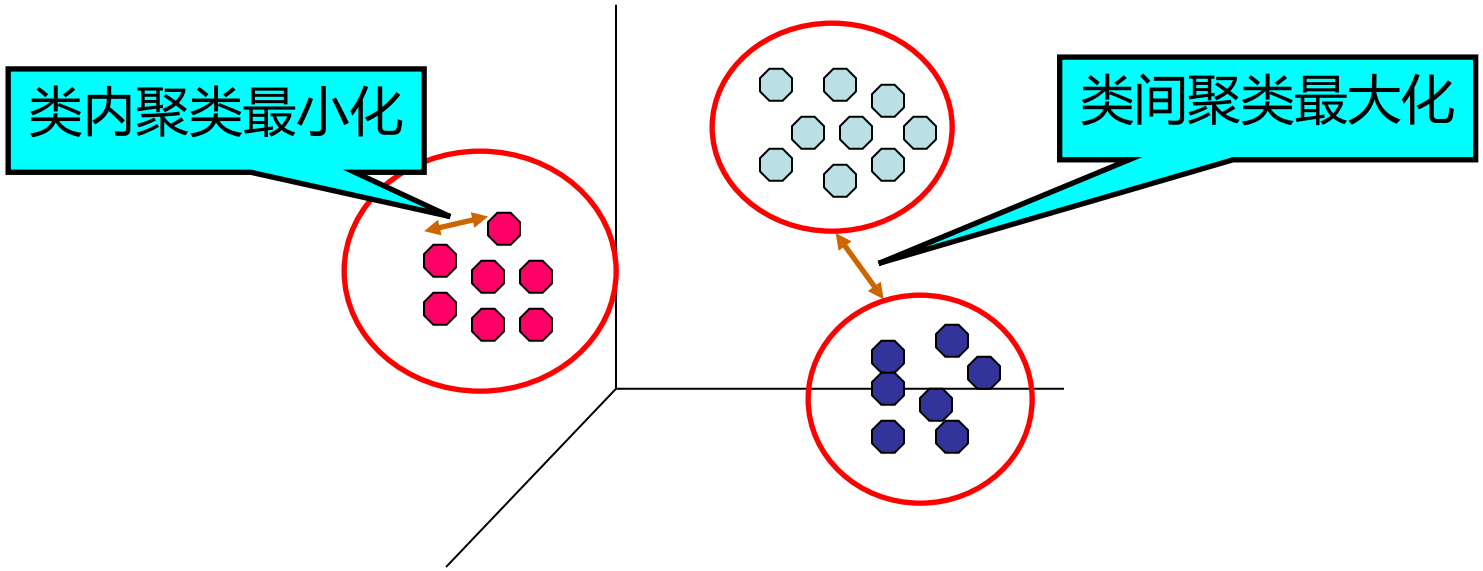
任务1：预测模型-回归

构建模型，根据样本历史数据和其他相关变量，预测给定连续变量的未来走势，也属于监督学习。

- 广泛研究于统计学、神经网络等领域。
- 示例：
  - 根据广告支出预测新产品的销售额；
  - 根据温度、湿度、气压等预测风速；
  - 股票市场指数的时间序列预测；
  - 各币种交易金额预测、汇率走势预测。

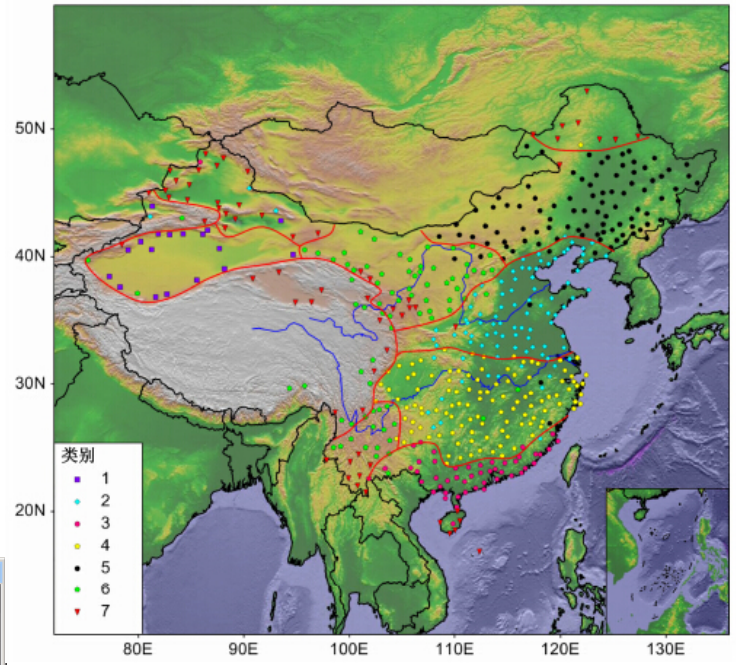
任务2：聚类

聚类(Clustering) 是按照某个特定标准(如距离)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。

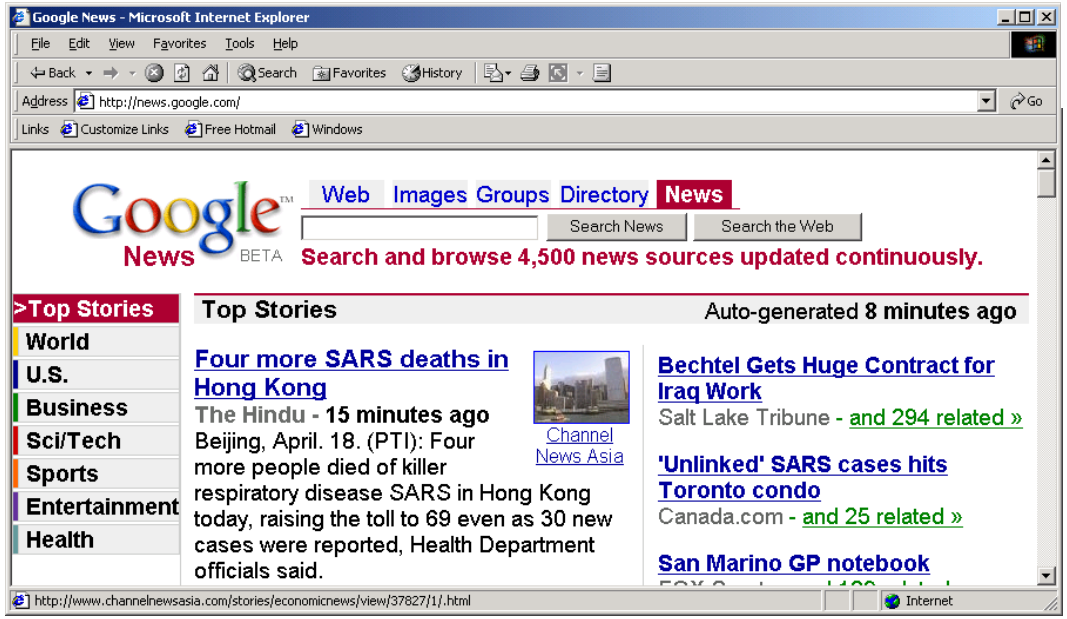


任务2：聚类

- **例子**
  - 消费市场定级
  - 相关新闻分组
  - 将具有相似功能的基因和蛋白质分组
  - 依据价格波动趋势对股票分组

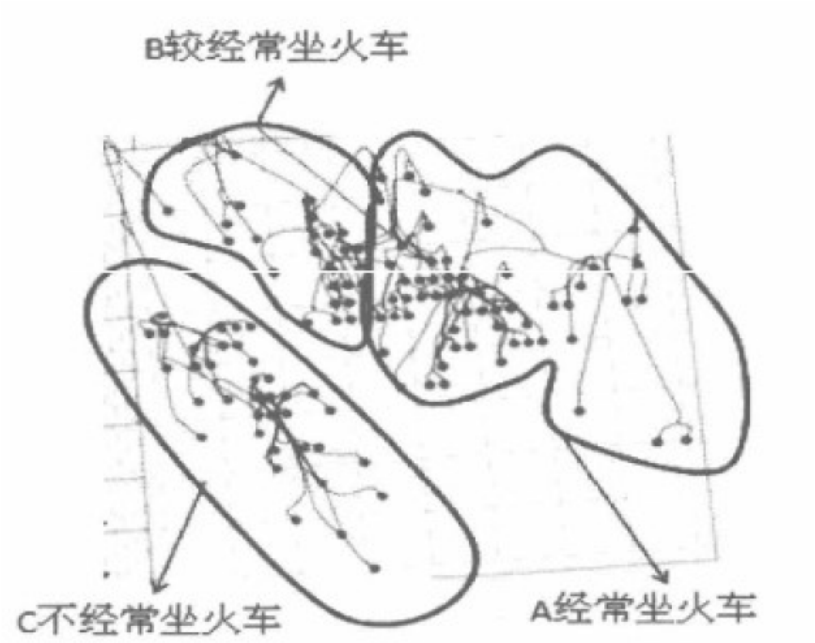


基于聚类算法划分中国温度区



任务2：聚类

- 铁路票价制定
  - 如何制定合适的票价提高上座率？将旅客进行聚类分析，根据旅客乘坐高铁频率的不同提供不同的优惠政策。合适的定价是提高高铁上座率的保障。



任务2：聚类

聚类的结果可能是模糊的。

How many clusters?

Six Clusters

Two Clusters

Four Clusters

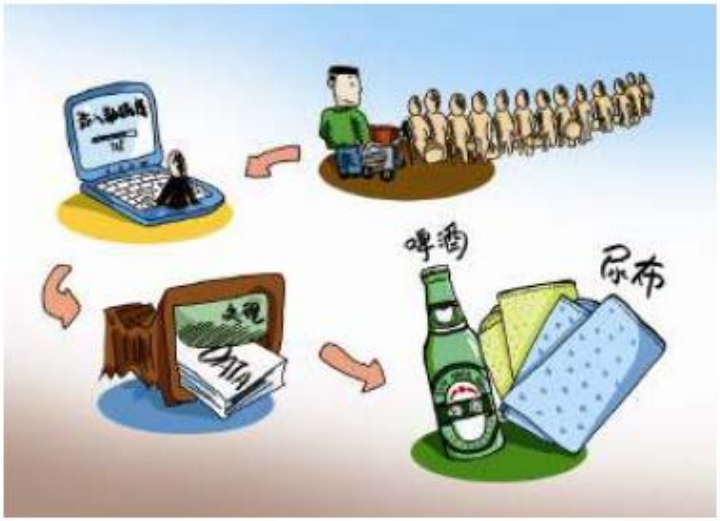
任务3：关联分析

关联分析，也叫关联规则挖掘，属于无监督算法的一种，它用于从数据中挖掘出潜在的关联关系。

零售业--“啤酒与尿布”营销策略

- 在一次圣诞节的顾客消费行为分析中，沃尔玛意外发现跟尿布一起购买最多的商品竟然是啤酒。经过深入分析后，卖场立即对两类商品的空间距离与价格都进行了调整，结果尿布与啤酒销量双双大增。

萨姆·沃尔顿， 沃尔玛公司创始人



轰动一时的啤酒与尿布关联规则

任务3：关联分析

古人的关联分析：马陵之战前，魏攻韩，韩向齐求救，齐国派田盼、孙膑等救援。孙膑运用“围魏救赵”之法，建议齐军直趋魏都大梁，迫使庞涓带领魏军回援。回援途中，庞涓发现一个惊奇的现象，即齐军留下的起火土灶越来越少，第一天有十万个，第二天剩下八万个，而第三天只有两万个，因此判断齐军溃逃。而后庞涓帅兵追击，中孙膑埋伏后死于马陵道。

孙膑

庞涓-马陵之战

任务3：关联分析

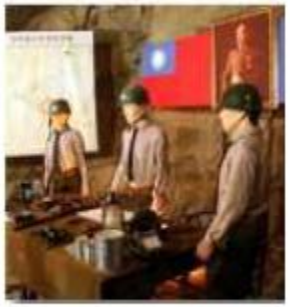
林彪的关联分析：

- 背景：1948年10月东北野战军先克锦州再战辽西， 10月19日，中央军委复电批准了东野的作战计划——就地聚歼廖耀湘兵团于野战之中。
- 一个普通战报的数据引起林彪注意–胡家窝棚
  - “为什么那里缴获的短枪与长枪的比例比其它战斗略高?”
  - “为什么那里缴获和击毁的小车与大车的比例比其它战斗略高?”
  - “为什么在那里俘虏和击毙的军官与士兵的比例比其它战斗略高?”

短枪

小车

军官



指挥所

任务3：关联分析

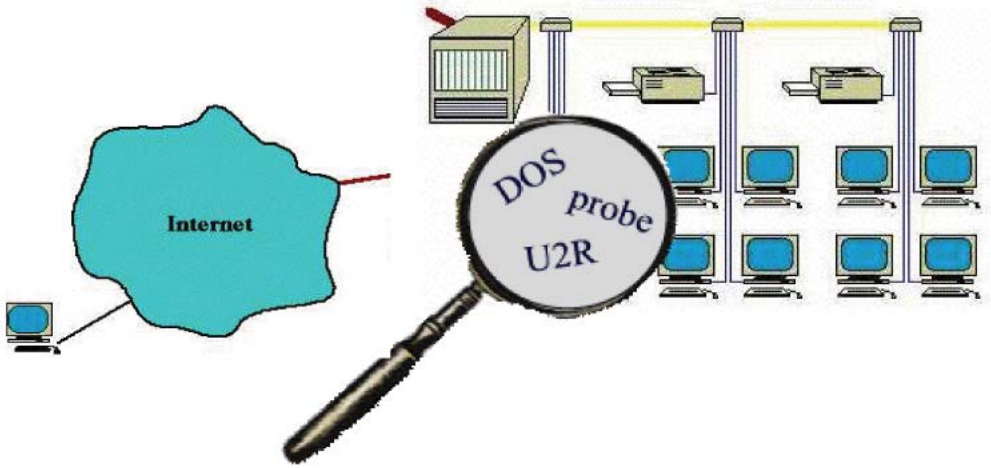
- 市场关联分析
  - 挖掘用于促销、货架管理和库存管理的规则
- 医学信息学
  - 挖掘某些疾病与相关症状、检测结果的关联组合
  - 挖掘癌症等疾病的诱因
- 股指期货关联性分析
  - 挖掘交易额、营收、热度等数据与相应股指期货的关系，借此辅助投资

任务4：异常检测（Deviation/Anomaly Detection）

对不符合预期模式或数据集中的其他项目、事件或观测值进行识别。

典型应用:

- 信用卡反欺诈检测
- 网络入侵检测
- 识别传感器网络中的异常行为，用于监控和监视



相关期刊与会议：

Conferences:

- The birth of data mining/KDD: 1989 IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 Workshops on Knowledge Discovery in Databases
- 1995 – date: International Conferences on Knowledge Discovery and Data Mining (KDD)
- 2001 – date: IEEE ICDM and SIAM-DM (SDM)
- Several regional conferences, incl. PAKDD (since 1997) & PKDD (since 1997)

Journals:

- Data Mining and Knowledge Discovery (DMKD, since 1997)
- Knowledge and Information Systems (KAIS, since 1999)
- IEEE Trans. on Knowledge and Data Engineering (TKDE)
- Many others, incl. TPAMI, TKDD, ML, MLR, VLDBJ …

我们需要什么工具?

–线性代数：向量，矩阵，逆，特征向量，奇异值分解...

–微积分：积分，导数...

–概率统计：随机变量，期望，贝叶斯定理...

–信息论：熵、KL散度、互信息

–优化理论：拉格朗日乘子，对偶性，KKT条件

（不知道，也没关系，用到的时候我们会讲解）

课程预期学习成果（CILOs）
Course Intended Learning Outcomes

1.识别和解释常见的数据；√

2.识别和解释常见的数据挖掘算法；√

3.实现数据挖掘算法并应用于实际问题中； √

4.分析和评价各种算法的有效性与优劣；

5.设计和创建新的机器学习算法，以解决现有算法的缺点和其他待解决特定的问题。

机器学习专题——监督学习

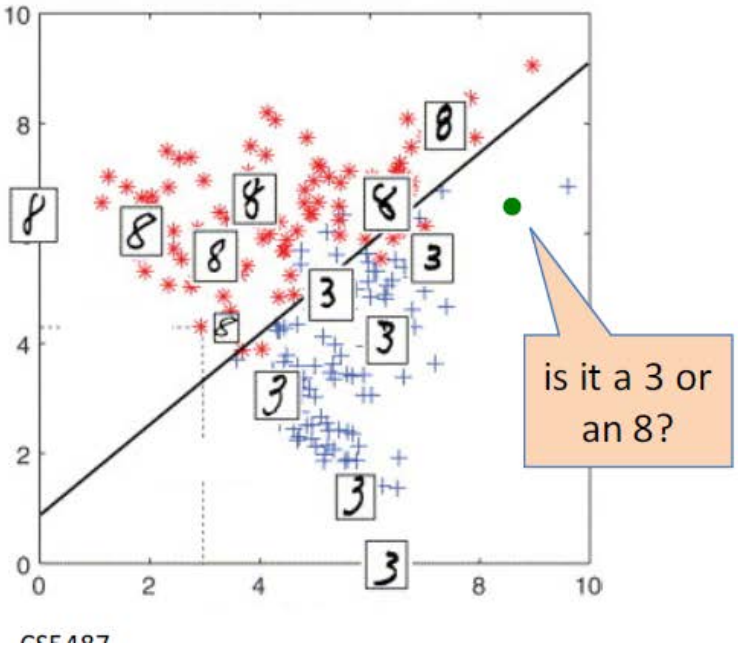
监督学习（*Supervised Learning*）

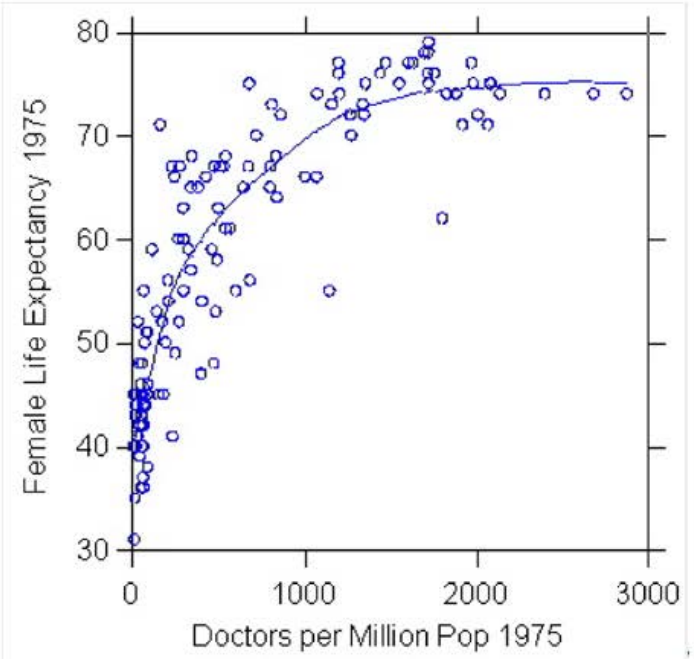
– 训练数据（training data）有输入和输出

- 例如，数字识别（输入=图像，输出=数字）

– 学习将输入映射到输出的函数

- 分类与回归





机器学习专题——无监督学习

无监督学习（*Unsupervised Learning*）

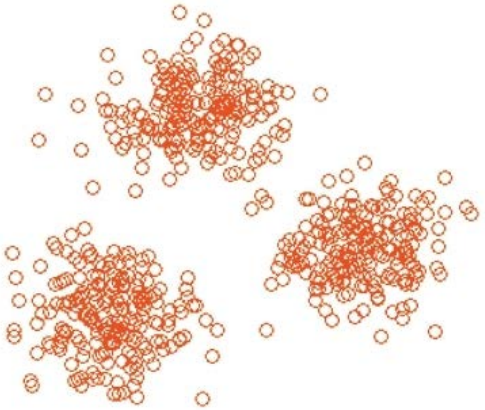
–训练数据只有输入（没有输出），算法在输出未知的情况下学习

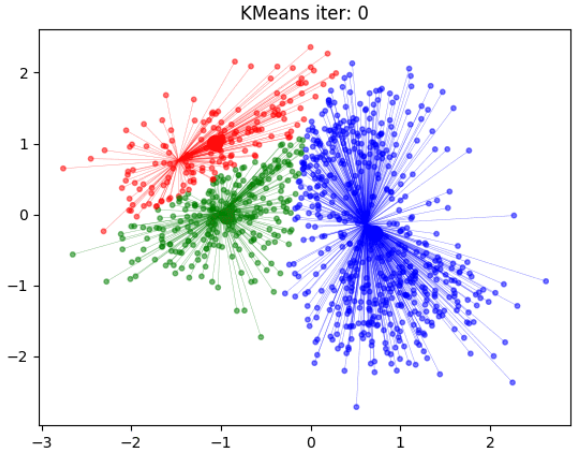
例如，收集网络文件（collection of web documents）

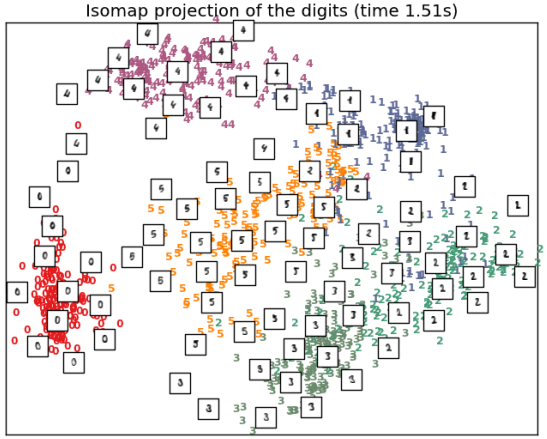
密度估计（density estimation）：了解输入在空间上的密度分布

–聚类：发现挖掘数据中的群组，并将新目标分入相应群组中。

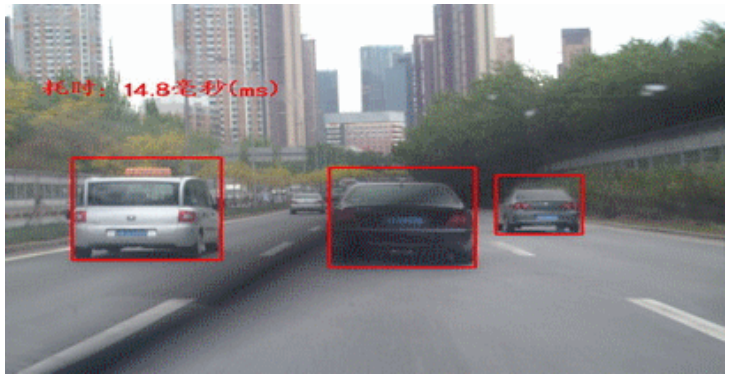
–可视化：将高维数据投影到2维或3维空间。







数据挖掘的应用



数据挖掘的应用


小度智能音箱大金刚
Baidu百度 | 小度
金属外观
红外控制
语音通话
钢炮音质
海量资源
小度


气体传感器
火灾传感器
RIP
彩色触摸屏
温度控制器
空调
电动窗帘开关
水晶面板开关
无线传感器
门磁开关
10寸触摸屏
监控主机

数据挖掘的应用