数据科学与工程导论

Introduction to Data Science and Engineering

- 据《纽约时报》和《英国观察家报》报道，2016 年美国总统大选期间，剑桥分析公司（Cambridge Analytica）与特朗普竞选团队合作，获取了总计超过 5000 万名 Facebook 用户的数据，采用独家的心理统计模型分析用户行为，对用户进行完整画像，帮助特朗普竞选团队定制从政治立场到竞选口号的一切，并精准投放数字广告，预测和影响民意的选择，一定程度上扭转了特朗普的糟糕形象，帮助他赢得了大选。

竞选作弊器?

- 奥巴马2012年竞选团队对每一个地区的支持者都进行分析，发现哪一个州选票下降，就到哪个州去加大竞选力度、做演讲、拉选票、筹款。有段时间，有心的选民发现奥巴马与著名影星乔治•克鲁尼总是在一起，不免奇怪，原来，他的竞选团队通过社交媒体数据分析发现，奥巴马最缺的就是加州的、有钱的、中年妇女的支持选票，进一步研究分析发现，加州中年女富婆的偶像是乔治•克鲁尼，于是立即进行竞选策划，频频展现奥巴马与乔治在一起打球、休闲的照片。自从他们俩在一块以后，竞选筹款的速度立即飚升，广告界不愧是业界的敏感精英，当年度《广告学人》杂志稳稳地评选奥巴马成为为年度最佳广告人。

竞选作弊器?

第 13 章

数据挖掘基础

1 初识数据挖掘

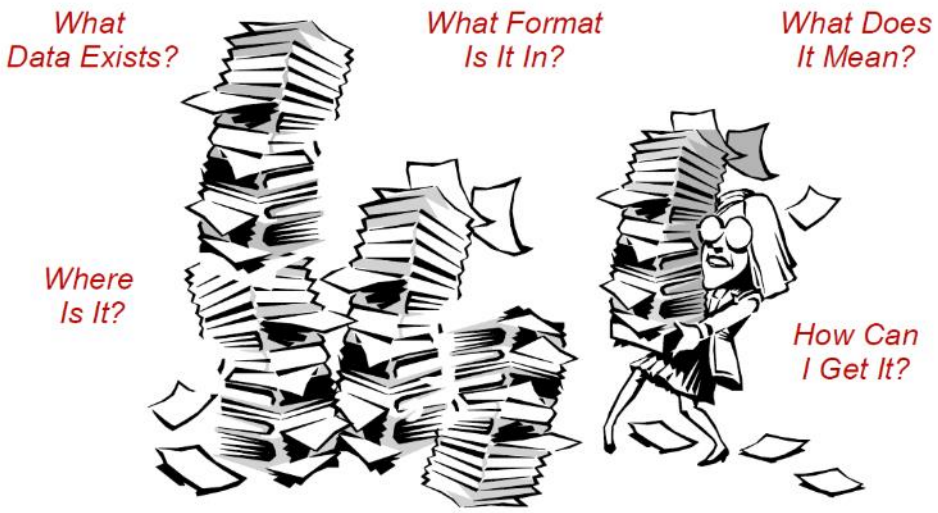
2 数据挖掘标准流程

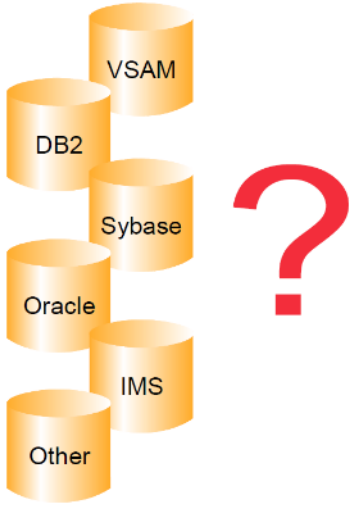
3 数据挖掘的技术

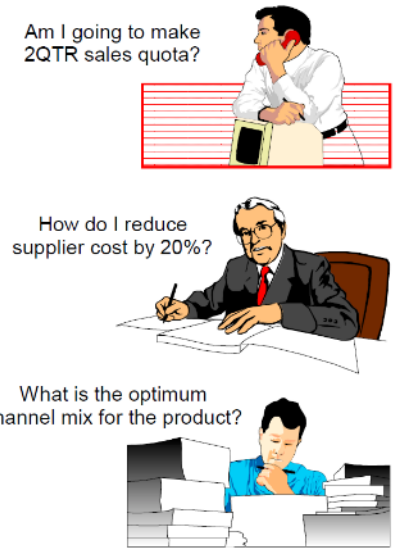
4 大数据挖掘

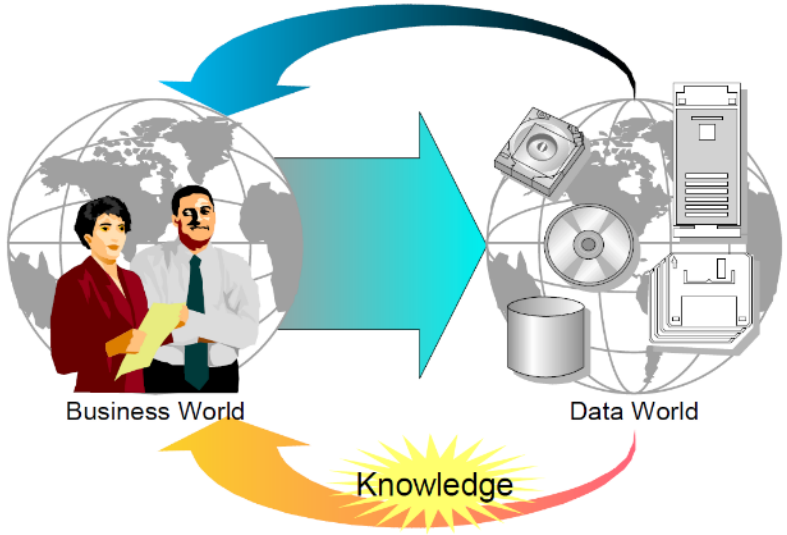
- 数据的爆炸性增长：从TB到PB
  - 数据的收集和数据的可获得性
    - 自动数据收集工具、数据库系统、WEB、计算机化的社会
  - 丰富数据的来源
    - 商业：WEB、电子商务、交易数据、股市...
    - 科学：遥感、生物信息学、科学模拟
    - 社会及每个人：新闻、数码相机、YouTube
- 我们被数据所淹没，但却渴望知识
- “需要是发明之母”，数据挖掘：海量数据的自动分析技术

为什么要数据挖掘









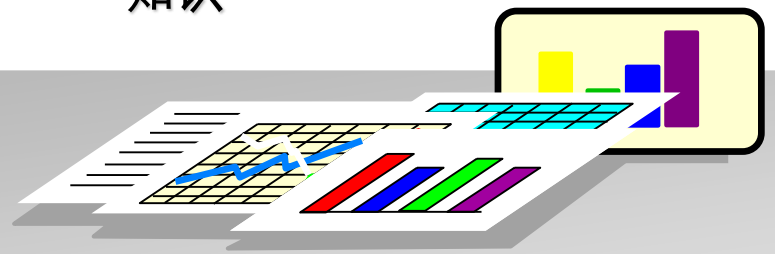
数据挖掘产生于商业高层的决策问题

为什么要数据挖掘

数据

- 金融
- 经济
- 政府
- POS.
- 人口统计
- 生命周期

知识



- 模式
- 趋势
- 事实
- 关系
- 模型
- 关联规则
- 序列

决策

- 目标市场
- 资金分配
- 贸易选择
- 在哪儿做广告
- 销售的地理位置

数据爆炸 知识贫乏

为什么要数据挖掘

- 数据挖掘（从数据中发现知识）
  - 从大量的数据中挖掘哪些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识
- 数据挖掘的替换词
  - 数据库中的知识挖掘（KDD）
  - 知识提炼
  - 数据/模式分析
  - 数据考古
  - 数据捕捞、信息收获等等

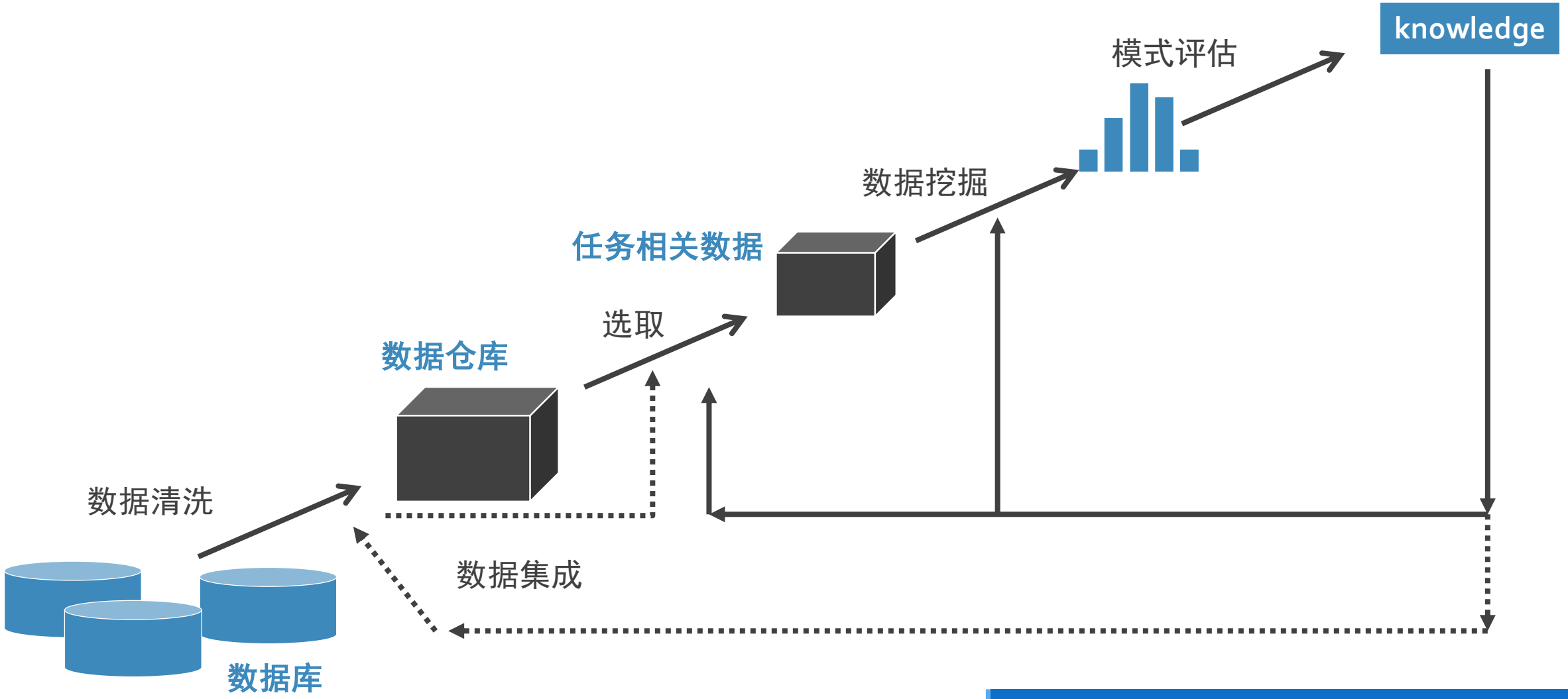
什么是数据挖掘

Data Mining

Knowledge Discovery in Databases (KDD)

A process of discovering patterns
in a large set of data

什么是数据挖掘



数据挖掘过程的核心

- 海量数据
  - 算法必须有高度的可扩展性，以有效处理**TB**级数据
- 高维数据
  - 可高达数万个不同的维
- 数据的高度复杂性
  - 流数据和传感数据
  - 时间数据、序列数据、时序数据
  - 图、社会网络、多关系数据
  - 异构数据库和遗产数据库
  - 空间数据、时空数据、多媒体、文本和WEB数据
- 新的、复杂的应用

数据挖掘 VS 传统数据分析

关系数据库

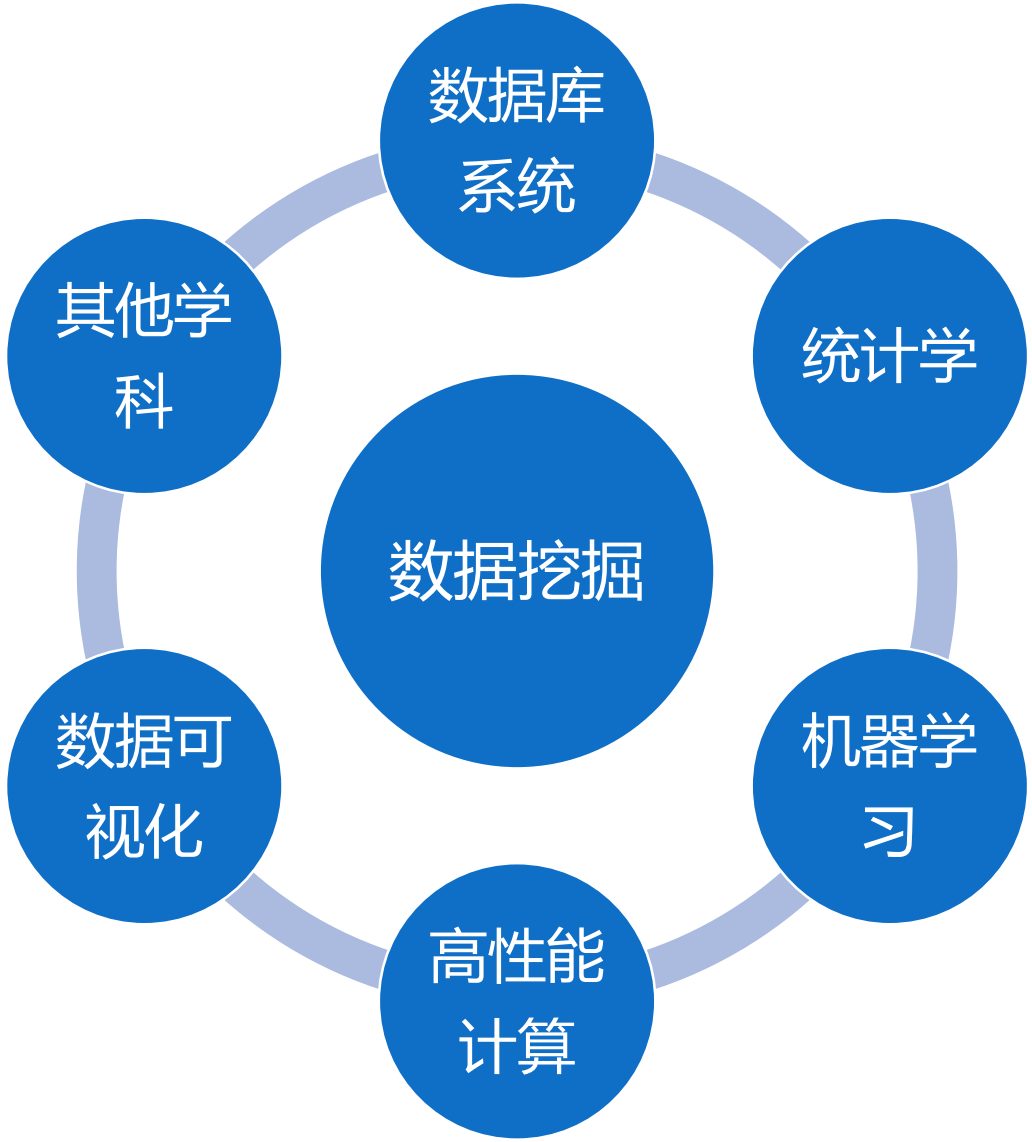
数据仓库

事务数据库

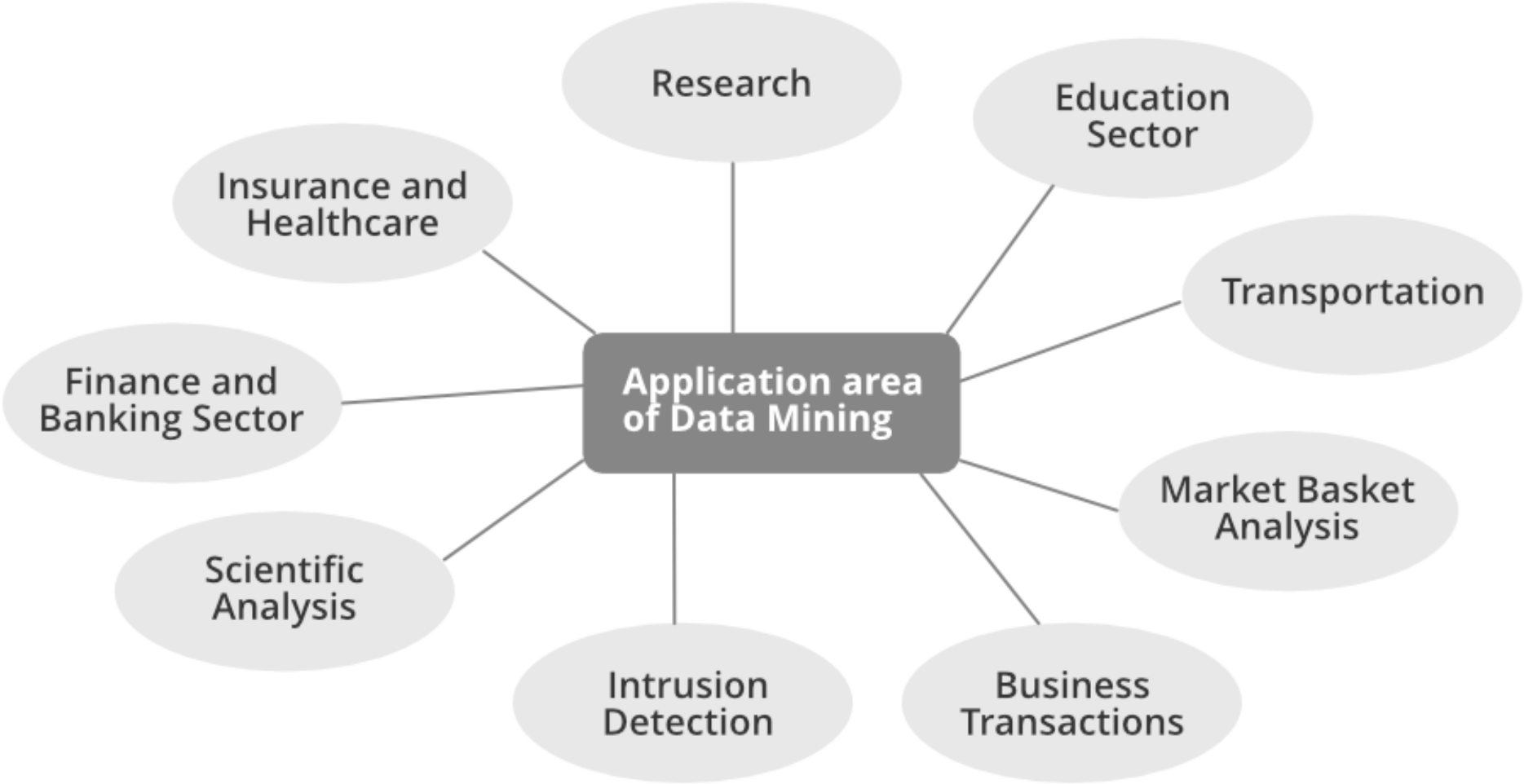
高级数据库和信息库

- 面向对象数据库
- 空间和时态数据库
- 时序数据
- 流数据
- 多媒体数据库
- 异种数据库
- 文本数据库

数据挖掘的数据源



数据挖掘：多学科融合


Research
Education Sector
Insurance and Healthcare
Transportation
Finance and Banking Sector
Application area of Data Mining
Market Basket Analysis
Scientific Analysis
Intrusion Detection
Business Transactions

数据挖掘的应用

第 13 章

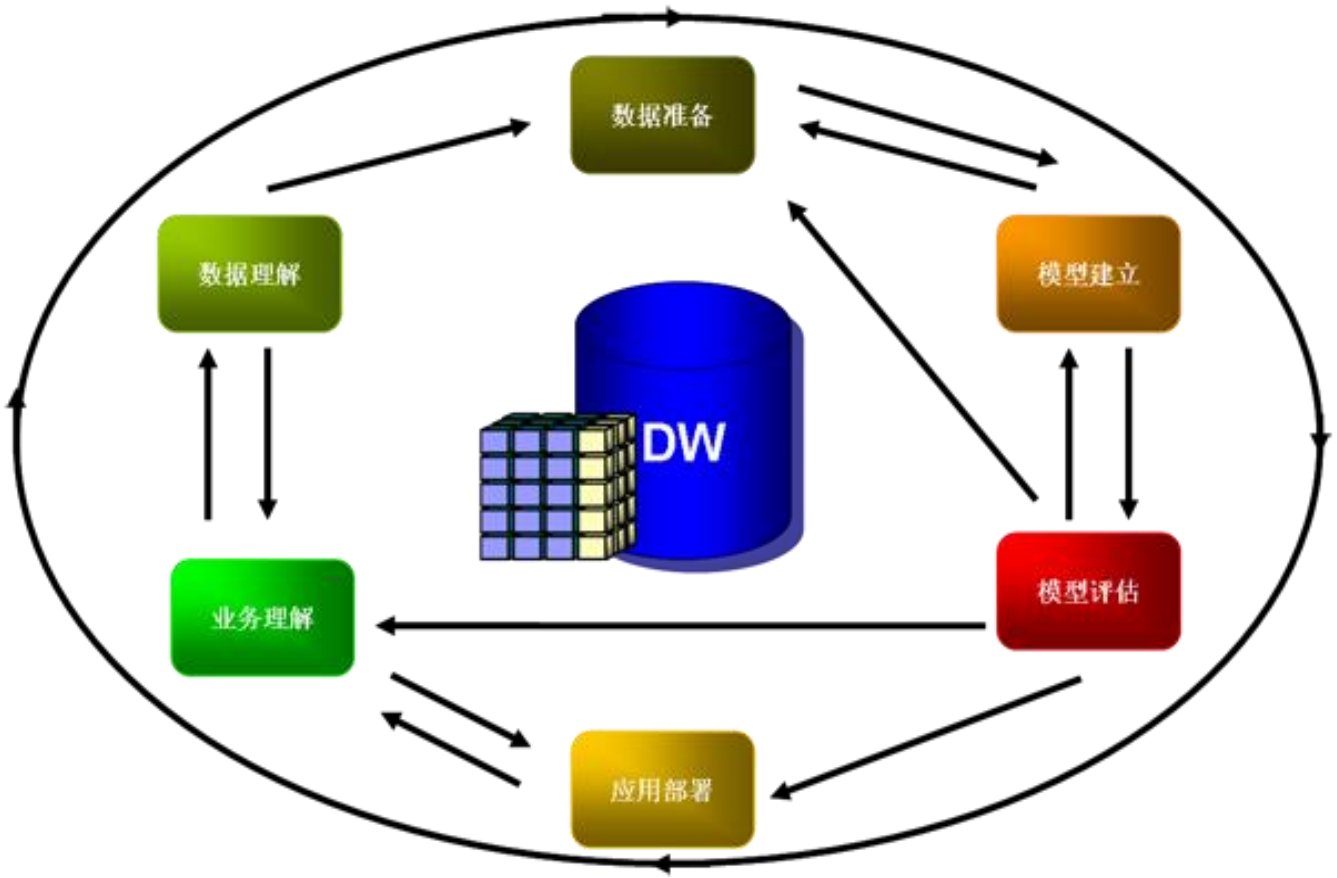
数据挖掘基础

1 初识数据挖掘

2 **数据挖掘标准流程**

3 数据挖掘的技术

4 大数据挖掘



商业理解

数据理解

数据准备

建立模型

模型评估

模型发布

数据挖掘流程

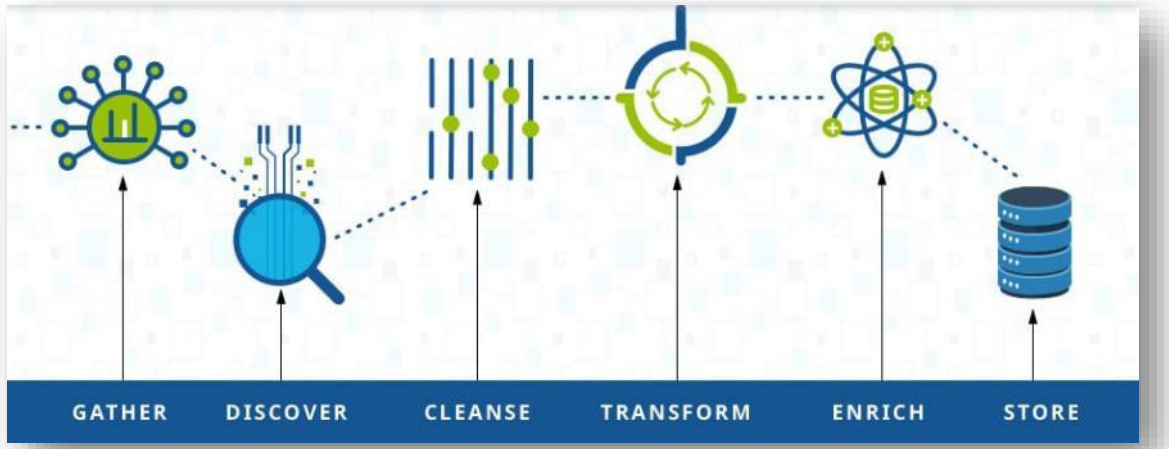
商业理解 | 数据理解 | 数据准备 | 建立模型 | 模型评估 | 模型发布

- 找问题—确定商业目标
- 对现有资源的评估
- 确定问题是否能够通过数据挖掘来解决
- 确定数据挖掘的目标
- 制定数据挖掘计划

数据挖掘流程

- 确定数据挖掘所需要的数据
- 对数据进行描述
- 数据的初步探索
- 检查数据的质量

- 选择数据
- 清理数据
- 对数据进行重建
- 调整数据格式使之适合建模



数据挖掘流程

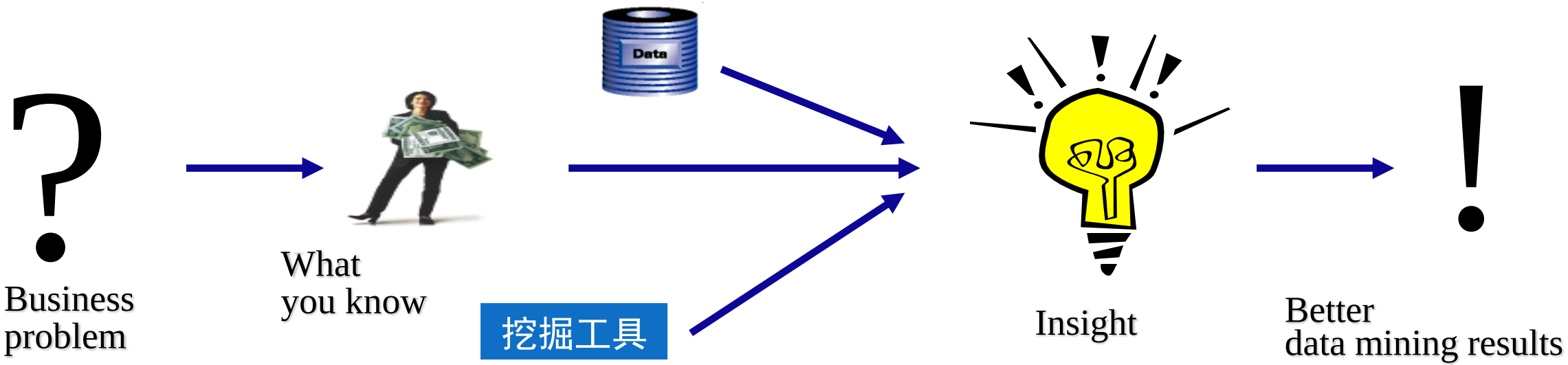
- 对各个模型进行评价
- 选择数据挖掘模型
- 建立模型

- 评估数据挖掘的结果
- 对整个数据挖掘过程的前面步骤进行评估
- 确定下一步怎么办？是发布模型？还是对数据挖掘过程进行进一步的调整，产生新的模型

- 把数据挖掘模型的结果送到相应的管理人员手中
- 对模型进行日常的监测和维护
- 定期更新数据挖掘模型

数据挖掘流程

商业理解 数据理解 数据准备 建立模型 模型评估 模型发布



把业务经验融入数据挖掘过程是数据挖掘成功的关键

数据挖掘流程

第13章
数据挖掘基础

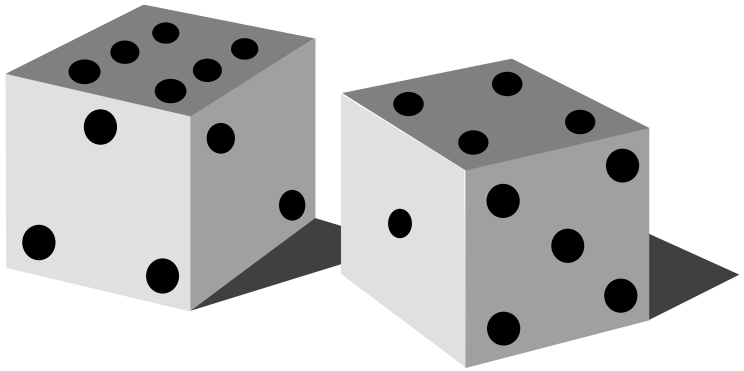
1 初识数据挖掘

2 数据挖掘流程

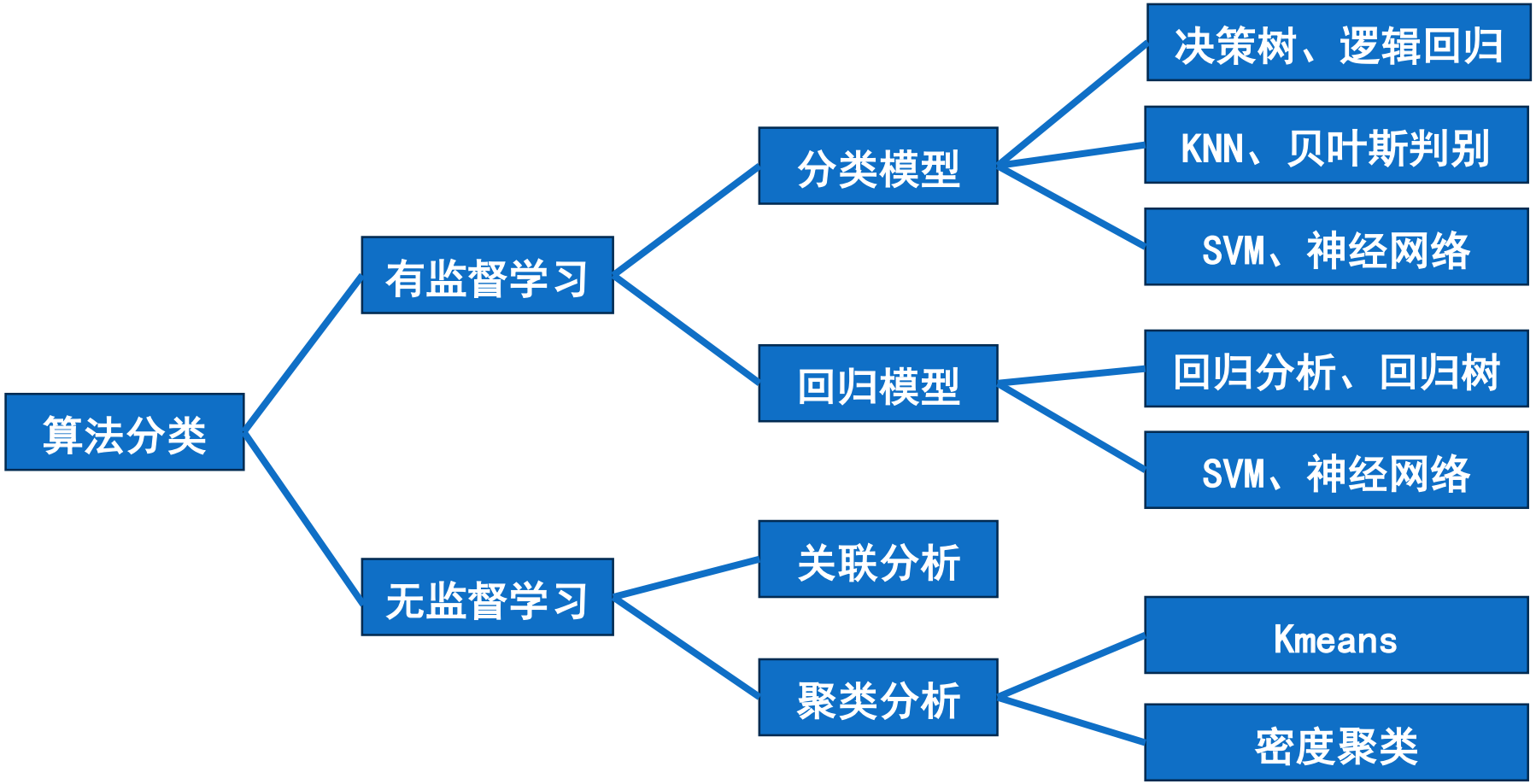
3 **数据挖掘的技术**

4 大数据挖掘

- 从分析目的分类：
  - 描述（Description）：了解数据中潜在的规律
  - 预测（Predication）：用历史预测未来
- 从技术类型分类：
  - 分类
  - 聚类
  - 关联分析
  - 异常检测



技术分类


算法分类
有监督学习
无监督学习
分类模型
回归模型
关联分析
聚类分析
决策树、逻辑回归
KNN、贝叶斯判别
SVM、神经网络
回归分析、回归树
SVM、神经网络
Kmeans
密度聚类

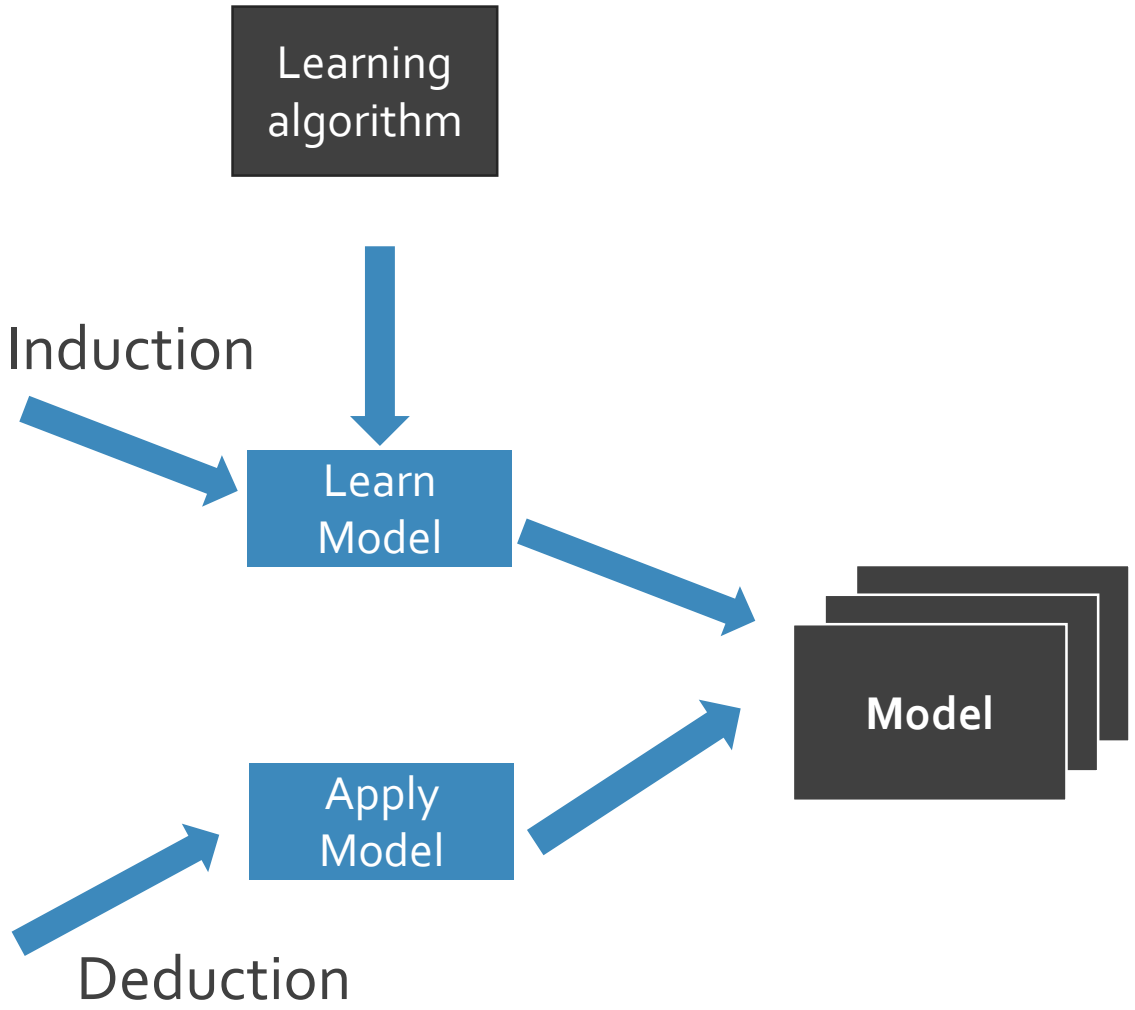
算法类型

- 给定一个记录(样本)集合 (称为训练集)
  - 每条记录有一些属性组成, 其中一个属性为类别
  - $(x_1, x_2, \ldots, x_n, y)$
- 找到一个将类别属性表示为其他属性的函数的模型
  - 如 $y = f(\boldsymbol{x})$
- 目标：未见过的记录尽可能准确地被分类.
  - 一个测试集用来确定模型的精度.
  - 通常, 给定的数据集被分成训练集和测试集, 训练集用于建立模型, 而测试集用于检验该模型.

分类

TID	Attrib1	Attrib2	Attrib3	class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

TID	Attrib1	Attrib2	Attrib3	class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

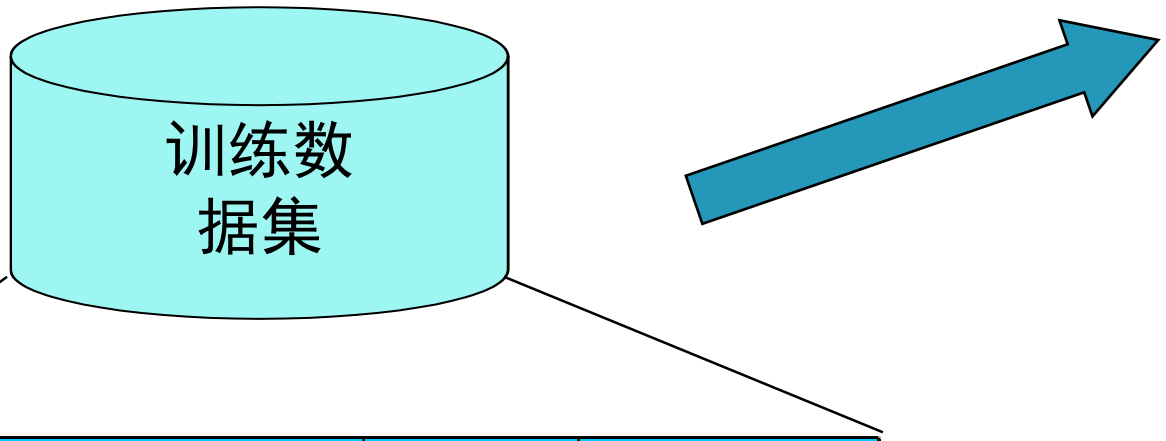


分类

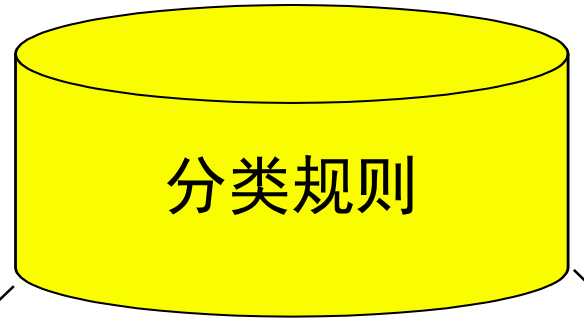
- 第一步，建立一个模型，描述预定数据类集和概念集
  - 假定每个元组属于一个预定义的类，由一个类标号属性确定
  - 基本概念
    - 训练数据集：由为建立模型而被分析的数据元组形成
    - 训练样本：训练数据集中的单个样本（元组）
  - 学习模型可以用分类规则、判定树或数学公式的形式提供
- 第二步，使用模型，对将来的或未知的对象进行分类
  - 首先评估模型的预测准确率
    - 对每个测试样本，将已知的类标号和该样本的学习模型类预测比较
    - 模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比
    - 测试集要独立于训练样本集，否则会出现“过分适应数据”的情况

分类：两步过程

第一步：建立模型



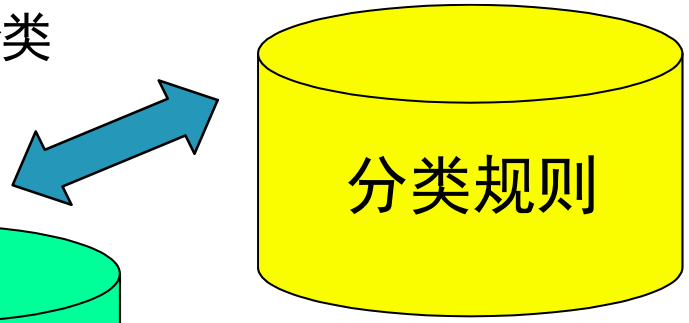
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

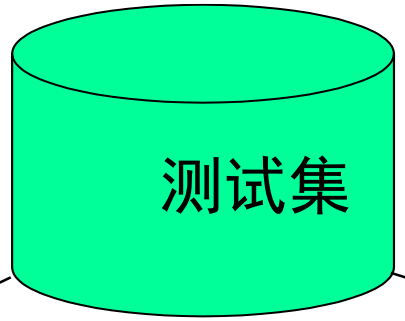


IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

分类：两步过程

第二步：用模型进行分类





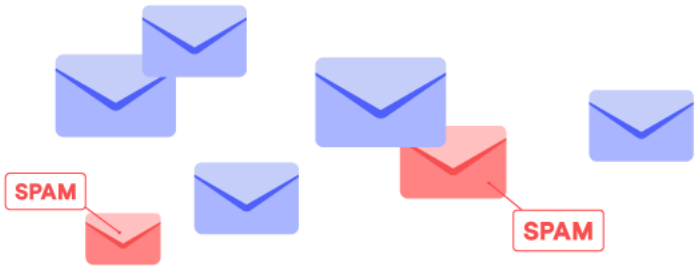
NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4)

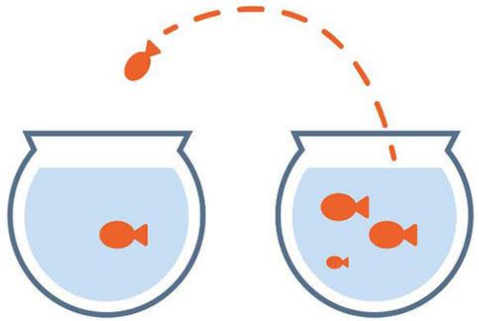
Tenured?

Yes

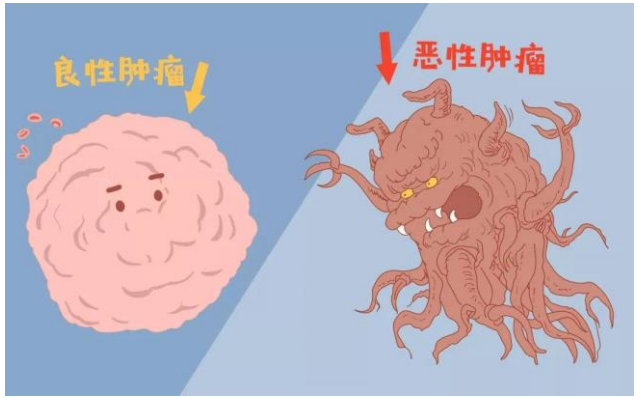
分类：两步过程



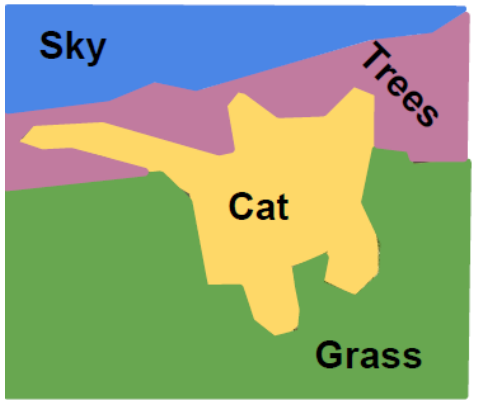
垃圾邮件检测



客户流失预测



肿瘤判断



图像分割

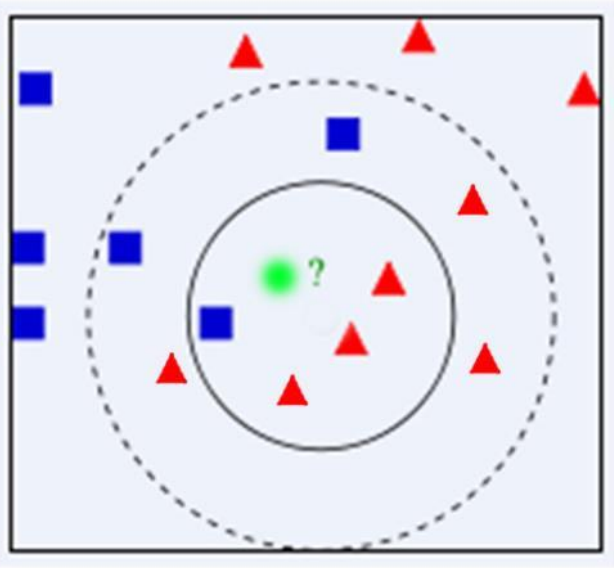
分类 应用

下面图片中只有三种豆，有三个豆是未知的种类，如何判定他们的种类？

1968年，Cover和Hart提出了最初的近邻法

分类：KNN算法示例

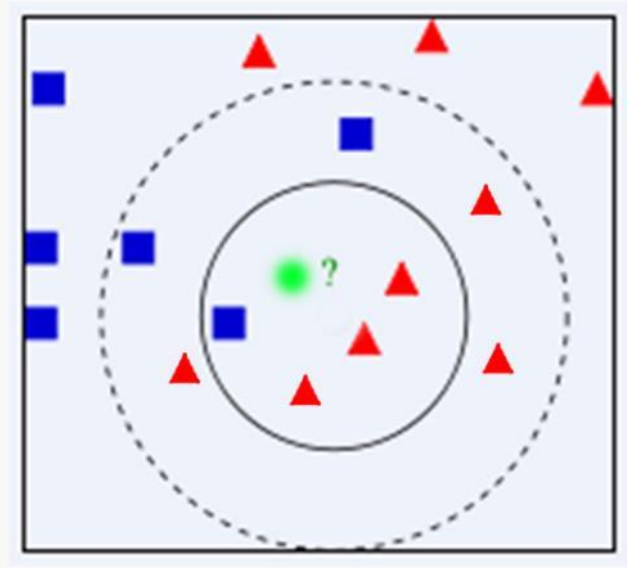
- 提供一种思路，即：未知的豆离哪种豆最近就认为未知豆和该豆是同一种类。
- 由此，我们引出最近邻算法的定义：
  - 为了判定未知样本的类别，以全部训练样本作为代表点，计算未知样本与所有训练样本的距离，并以最近邻者的类别作为决策未知样本类别的唯一依据。
- 但是，最近邻算法明显是存在缺陷的，我们来看一个例子。
  - 我们可以明显发现最近邻算法的缺陷——**对噪声数据过于敏感**。为了解决这个问题，我们可以把位置样本周边的多个最近样本计算在内，扩大参与决策的样本量，以避免个别数据直接决定决策结果，这样的算法又称作K-最近邻算法（KNN）。



绿色点是正方形还是三角形？

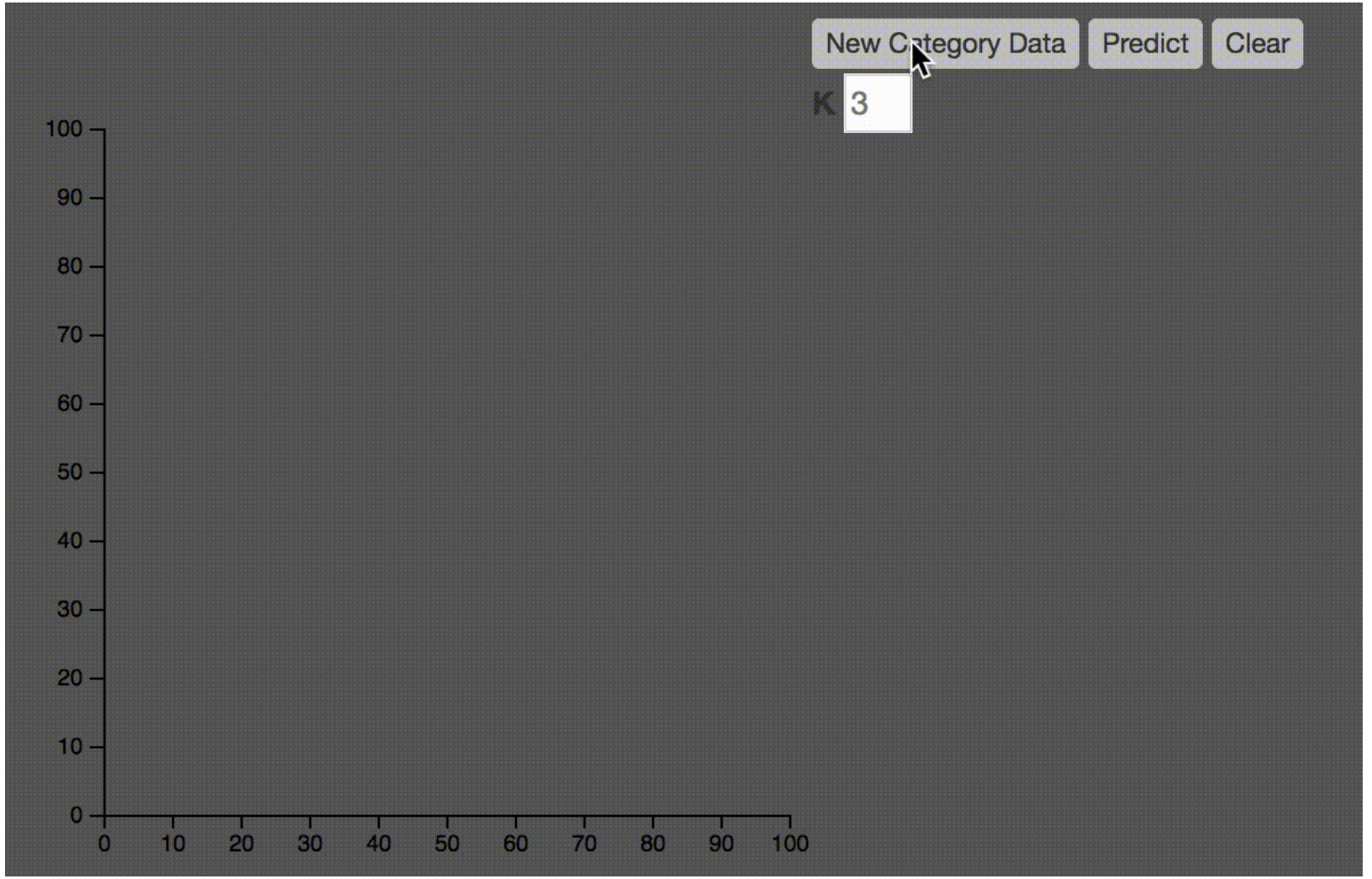
分类：KNN算法示例

- KNN算法是最近邻算法的一个延伸。
- 基本思路是：选择未知样本一定范围内确定个数的K个样本，该K个样本大多数属于某一类型，则未知样本判定为该类型。
- 实现步骤：
  1. 初始化距离为最大值
  2. 计算未知样本和每个训练样本的距离dist
  3. 得到目前K个最临近样本中的最大距离maxdist
  4. 如果dist小于maxdist，则将该训练样本作为K-最近邻样本
  5. 重复步骤2、3、4，直到未知样本和所有训练样本的距都算完
  6. 统计K个最近邻样本中每个类别出现的次数
  7. 选择出现频率最大的类别作为未知样本的类别



绿色点是正方形还是三角形？

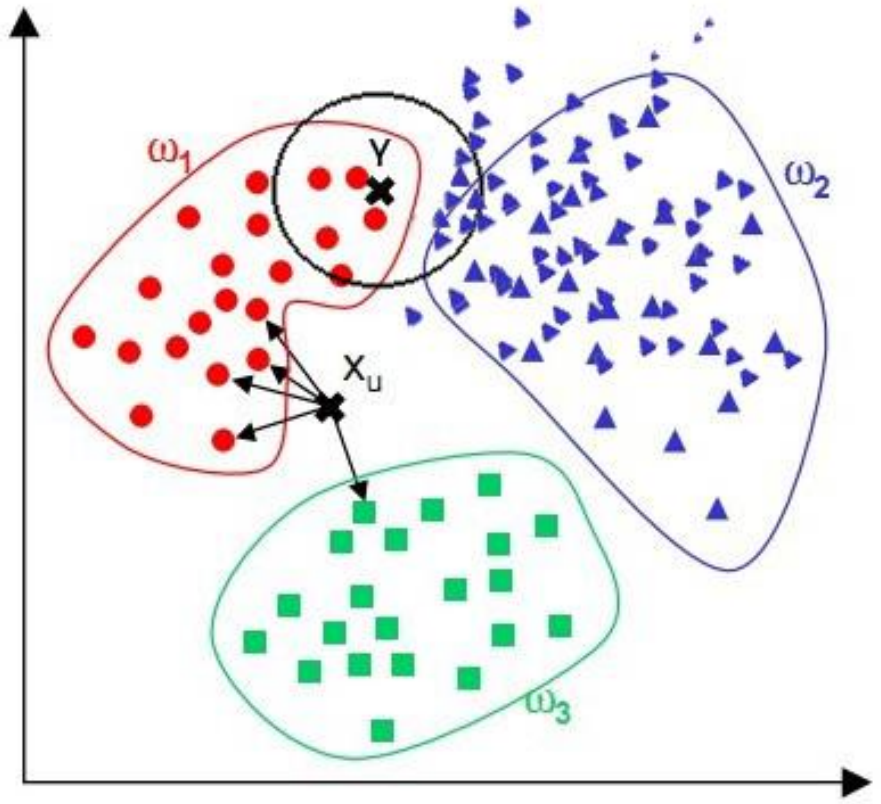
分类：KNN算法示例



https://codepen.io/gangtao/pen/ayPVQz

分类：KNN算法示例

- KNN算法的缺陷
  - 对于位置样本X，通过KNN算法，我们显然可以得到X应属于红点
  - 但对于位置样本Y，通过KNN算法我们似乎得到了Y应属于蓝点的结论
  - 而这个结论直观来看并没有说服力。



分类：KNN算法示例

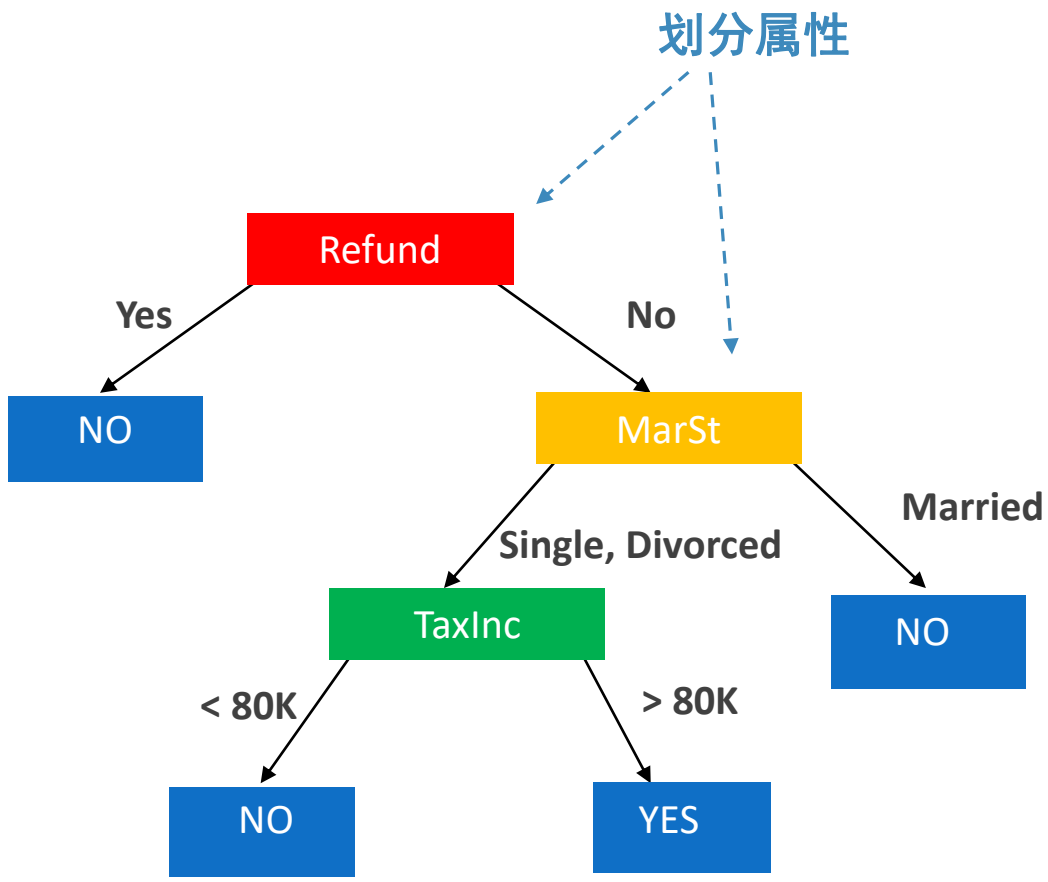
- 由上面的例子可见，该算法在分类时有个重要的不足是，当样本不平衡时，即：一个类的样本容量很大，而其他类样本数量很小时，很有可能导致当输入一个未知样本时，该样本的K个邻居中大数量类的样本占多数。 但是这类样本并不接近目标样本，而数量小的这类样本很靠近目标样本。
- 这个时候，我们有理由认为该位置样本属于数量小的样本所属的一类，但是，KNN却不关心这个问题，它只关心哪类样本的数量最多，而不去把距离远近考虑在内。
- 因此，我们可以采用权值的方法来改进。和该样本距离小的邻居权值大，和该样本距离大的邻居权值则相对较小，由此，将距离远近的因素也考虑在内，避免因一个样本过大导致误判的情况。

分类：KNN算法示例

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

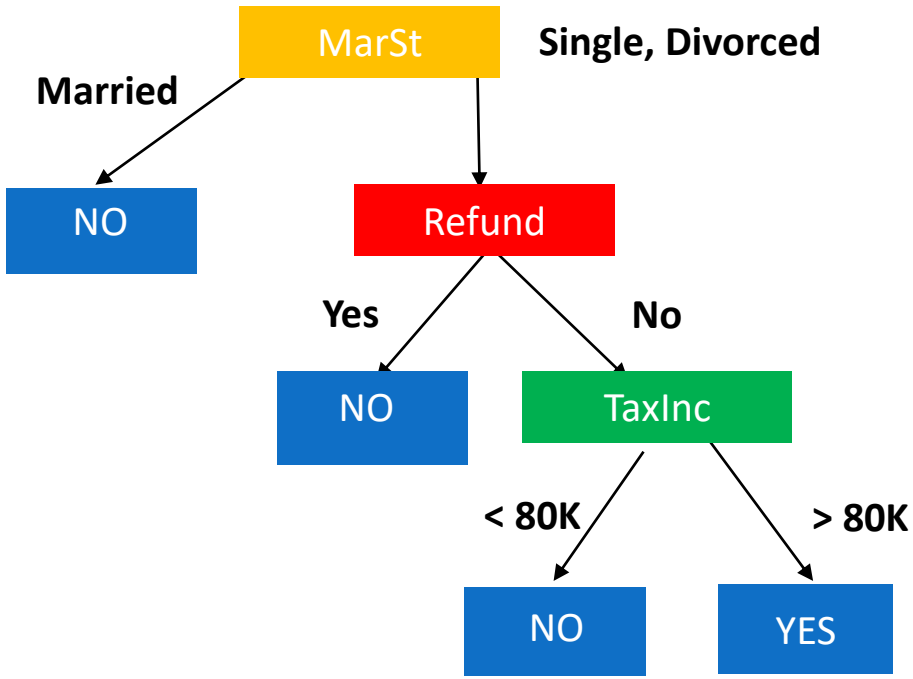
离散值

连续值



分类：决策树示例

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

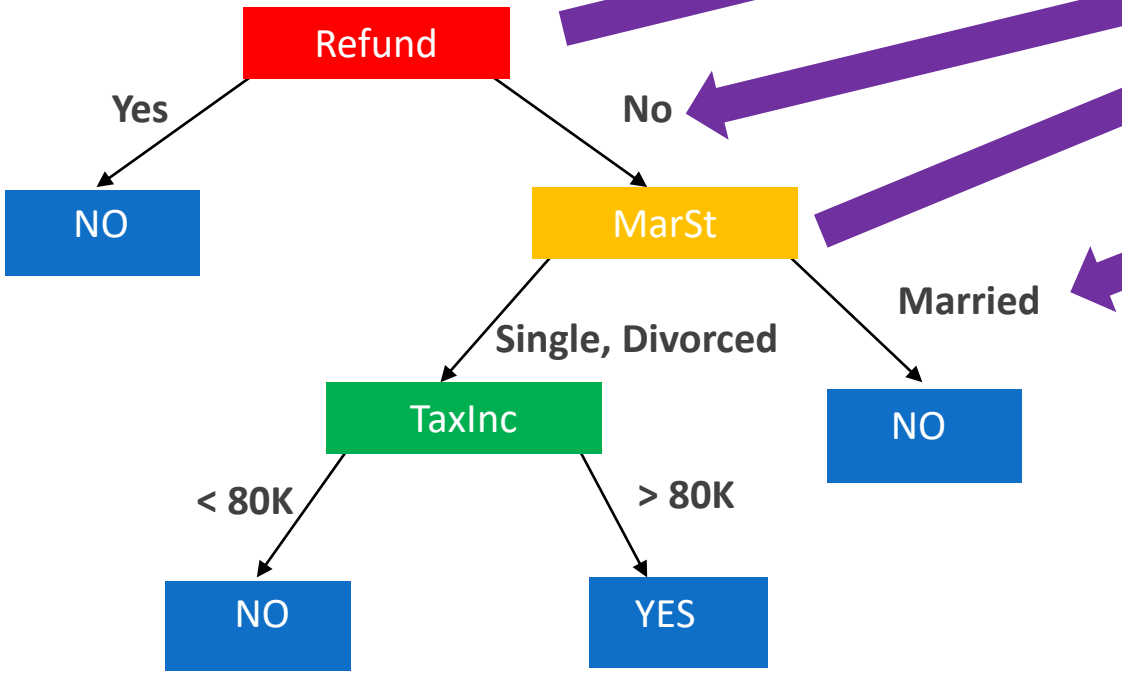


可能有多棵决策树拟合同一个数据集

分类：决策树示例

测试数据

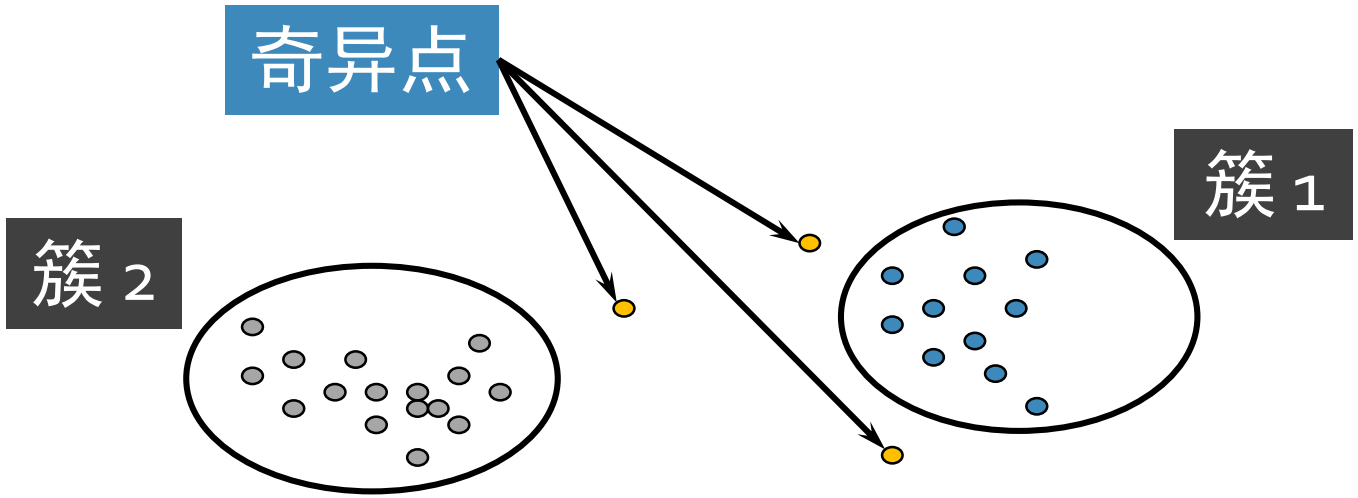
Refund	Marital Status	Taxable Income	Cheat
N	Married	80K	NO



分类：决策树示例

- 聚类分析：
  - 将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程。
  - 最大化类内的相似性和最小化类间的相似性
  - 不像分类和预测分析标号类的数据对象，聚类分析数据对象不考虑已知的标号类
- 聚类是一种无监督分类法: 没有预先指定的类别；
- 例如：
  - 对WEB日志的数据进行聚类，以发现相同的用户访问模式
  - 城市规划: 根据类型、价格、地理位置等来划分不同类型的住宅
  - 市场销售: 帮助市场人员发现客户中的不同群体，然后用这些知识来开展一个目标明确的市场计划

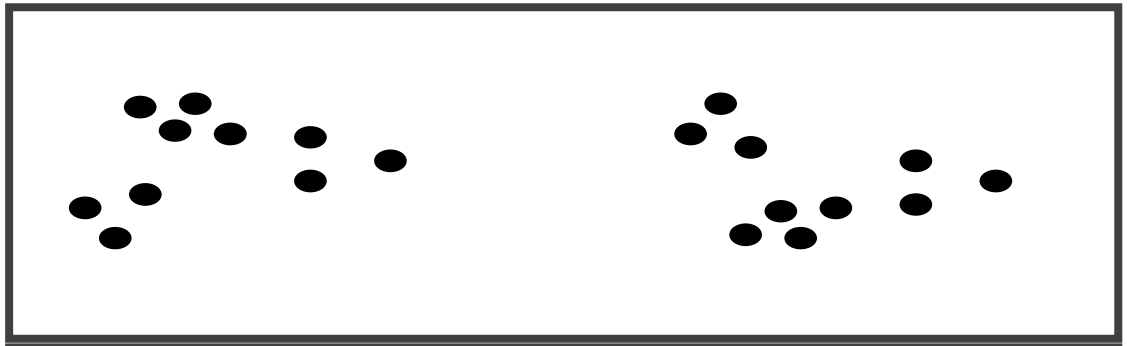
- 把数据聚类成多个簇
  - 同一个簇中的数据相似
  - 不同簇中数据不相似
  - 非监督学习：没有预先定义的类



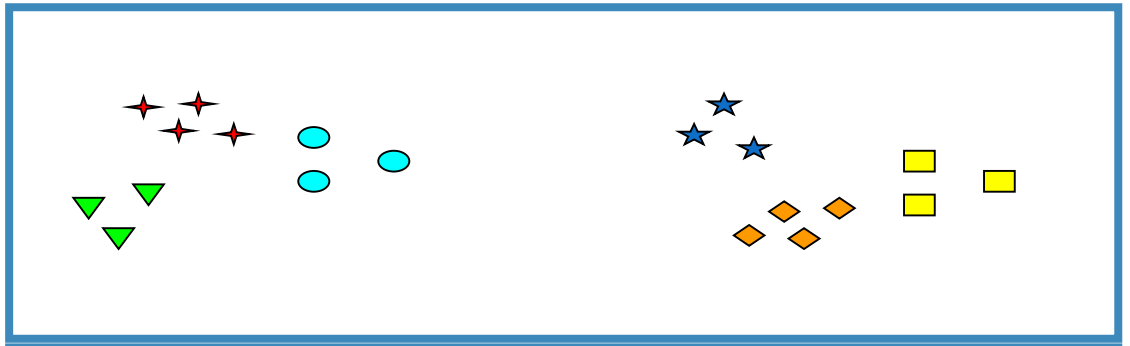
应用

- 发现数据分布
- 模式识别
- 文档聚类
- 空间数据分析
- 市场研究

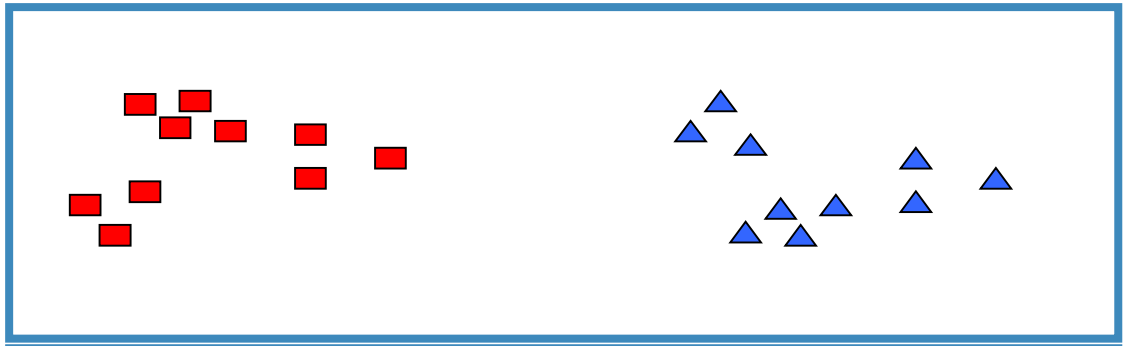
什么是聚类



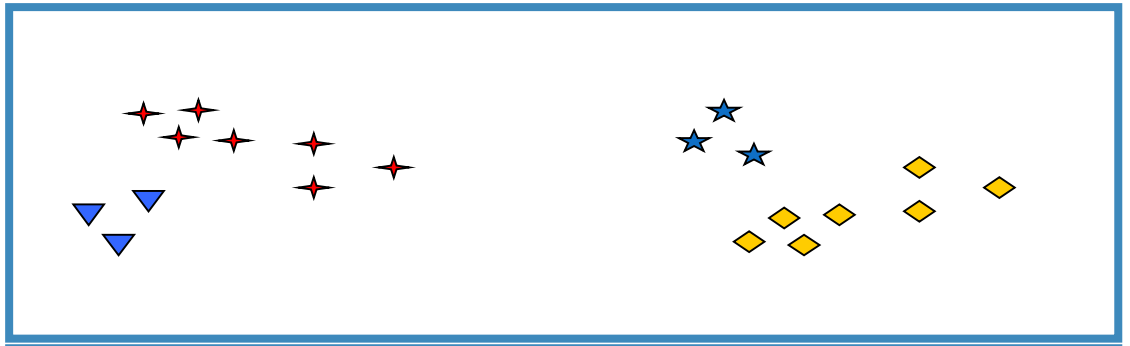
多少个簇?



6个簇

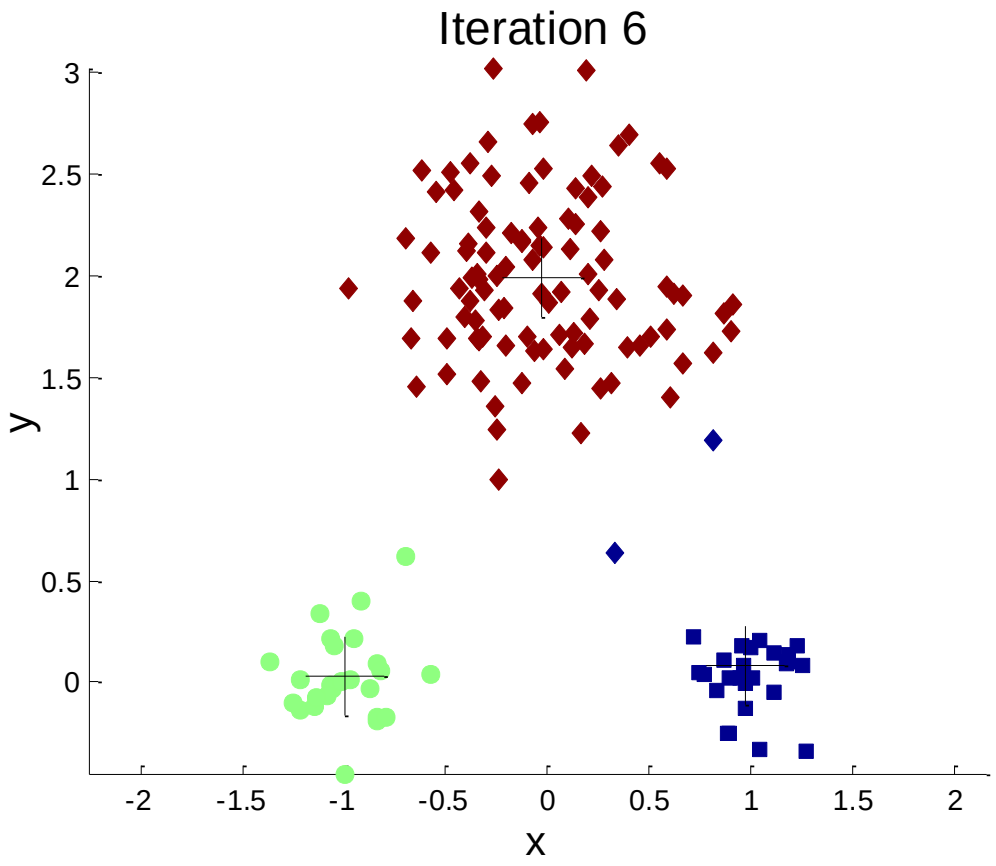


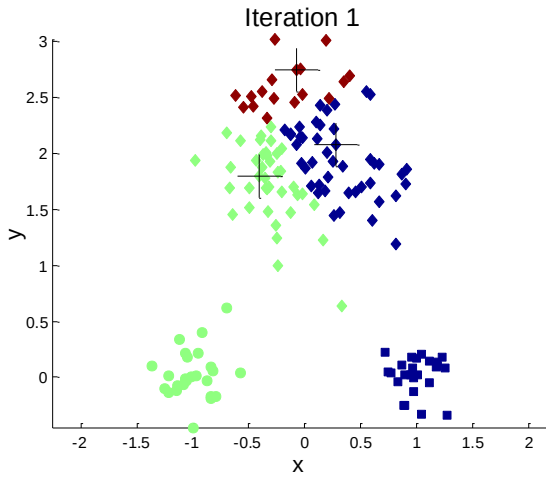
2个簇

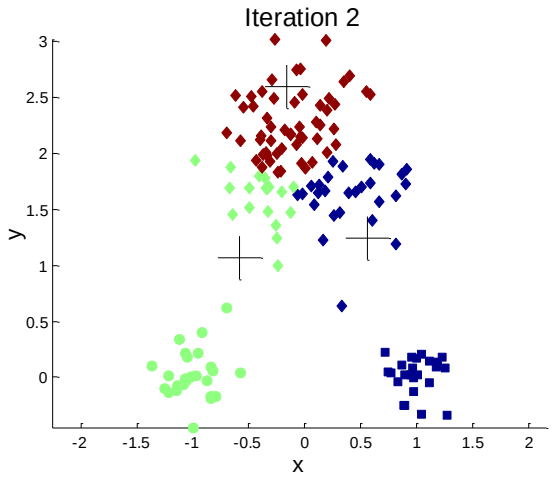


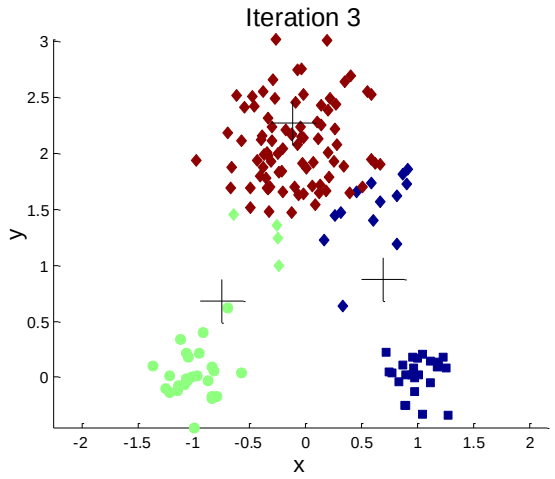
4个簇

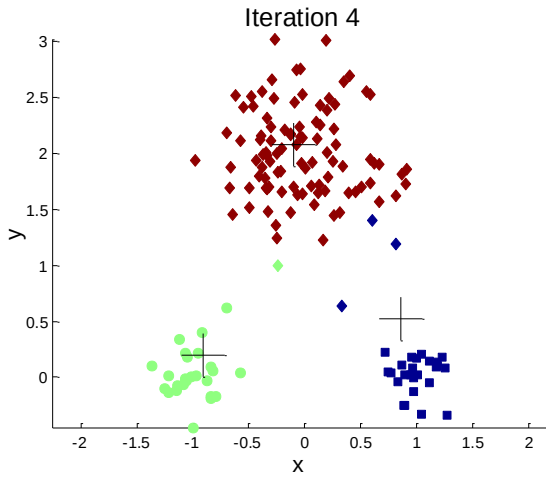
聚类的概念是模糊的

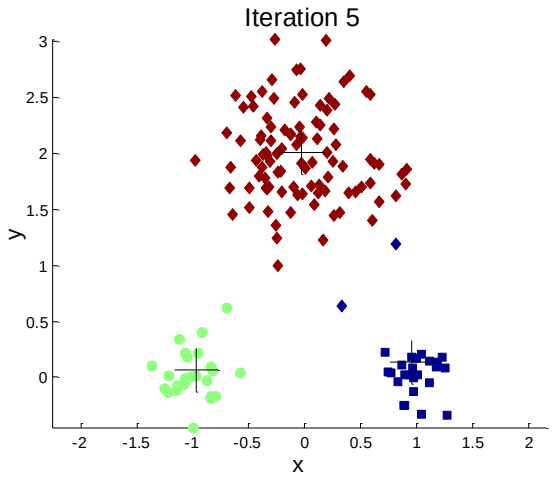

Iteration 6
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

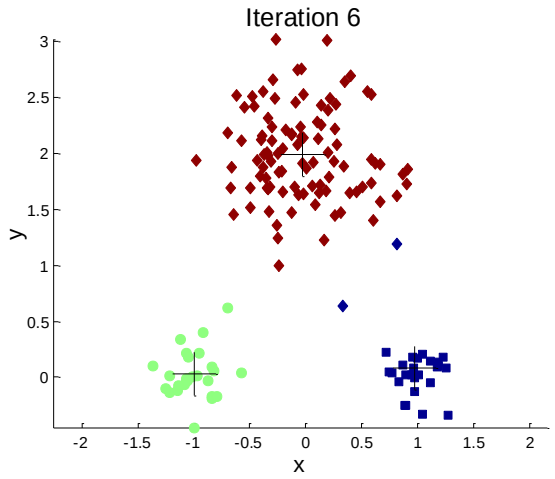

Iteration 1
x
y


Iteration 2
x
y


Iteration 3
x
y

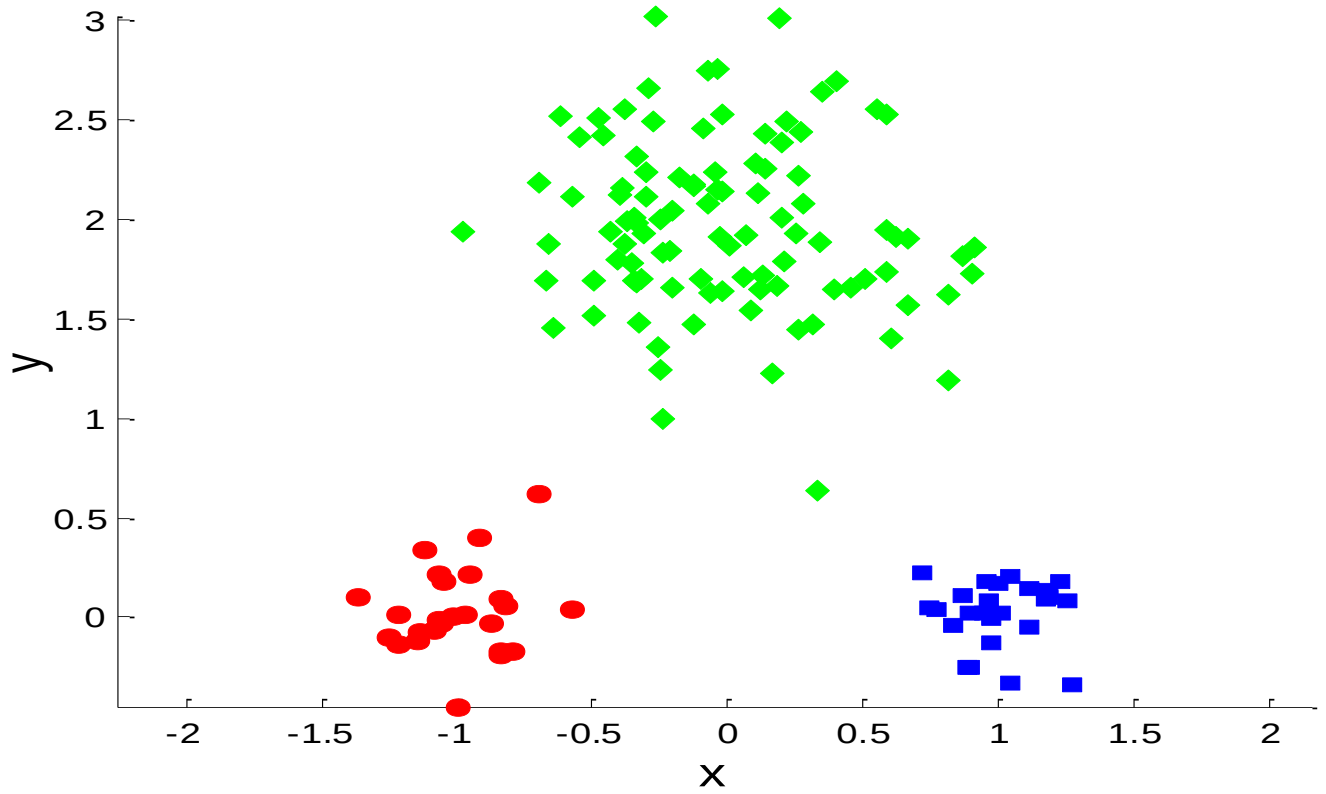

Iteration 4
x
y

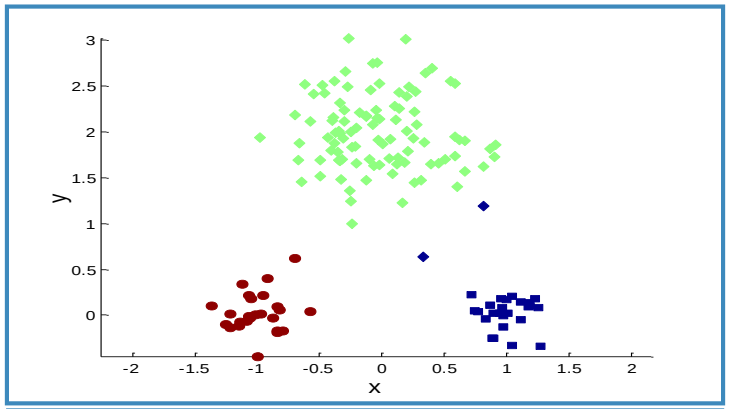

Iteration 5
x
y


Iteration 6
x
y

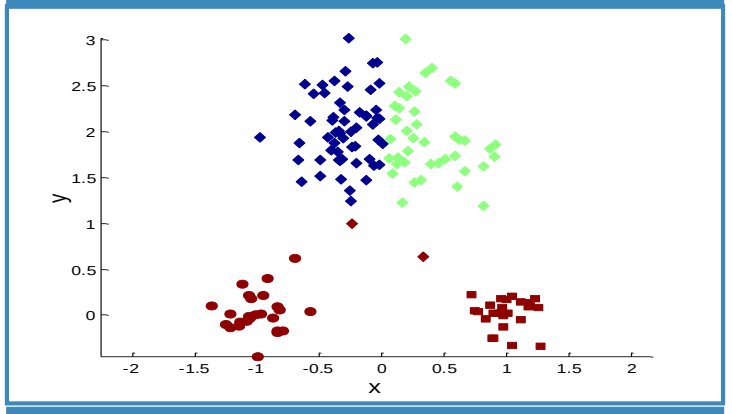
聚类：K-Means

原始数据点



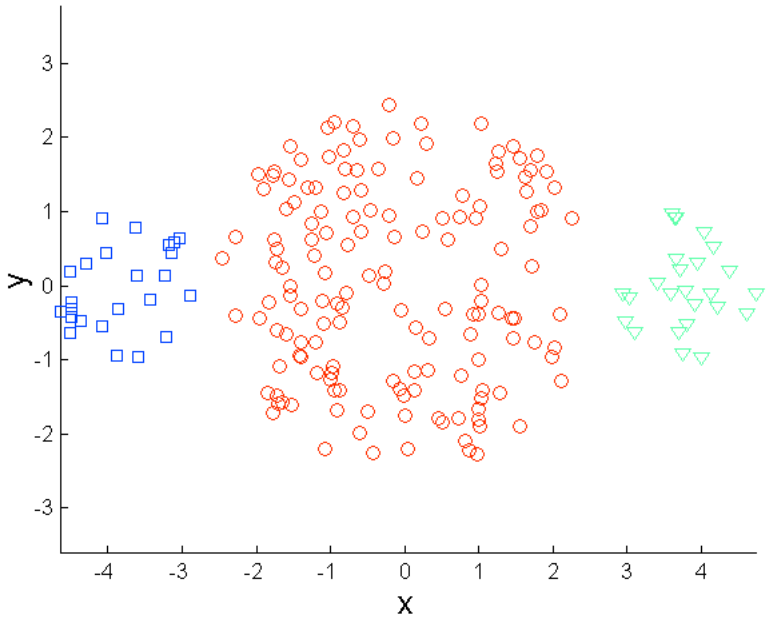


最优聚类

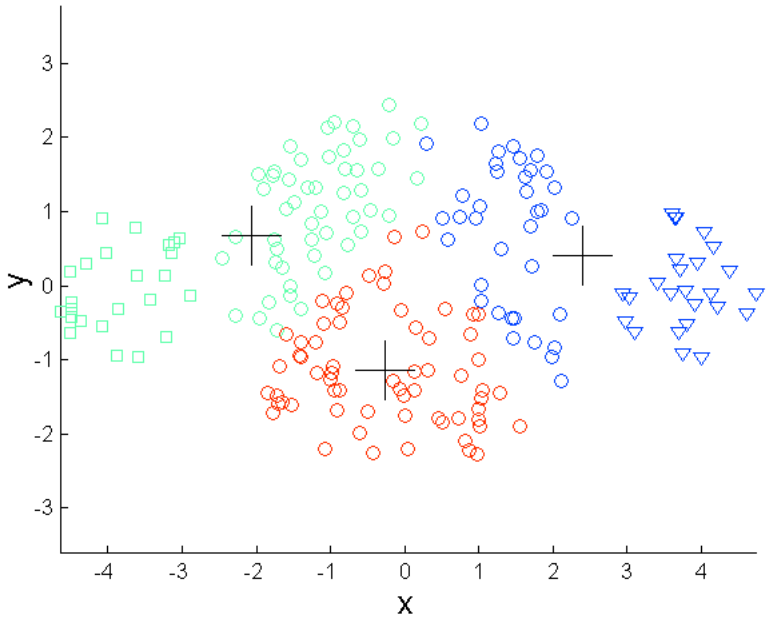


次最优聚类

Limitations of K-means: Differing Sizes

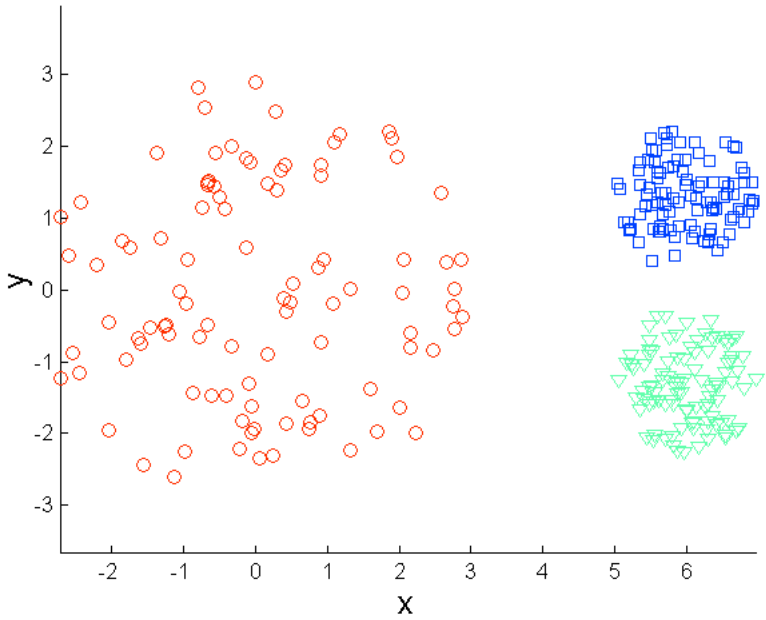


Original Points

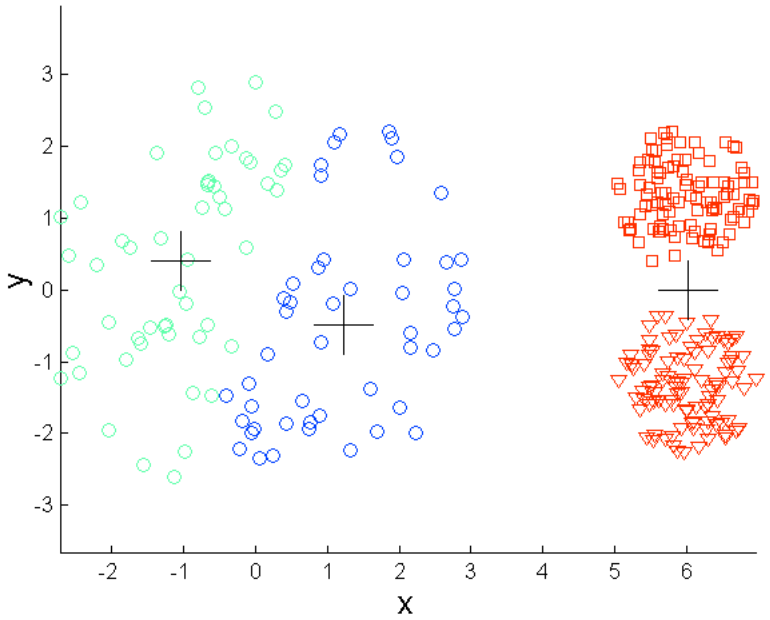


K-means (3 Clusters)

Limitations of K-means: Differing Density

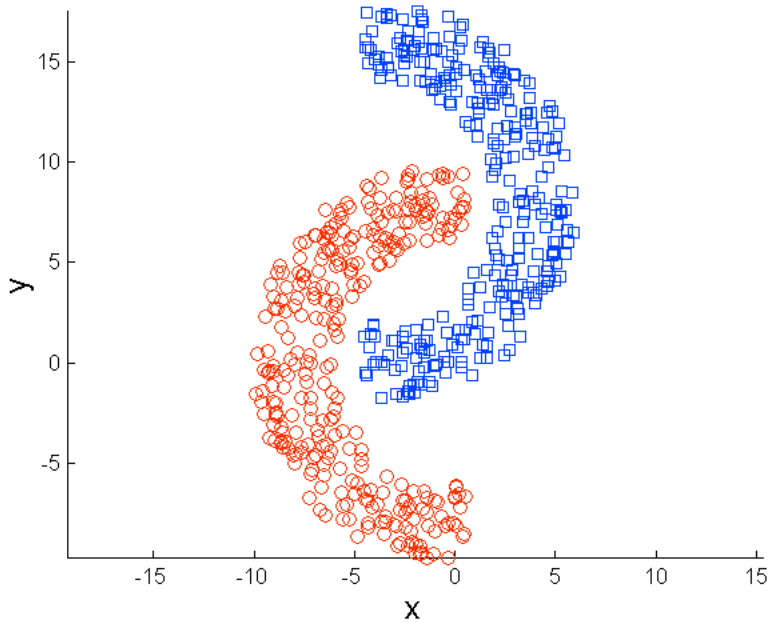


Original Points

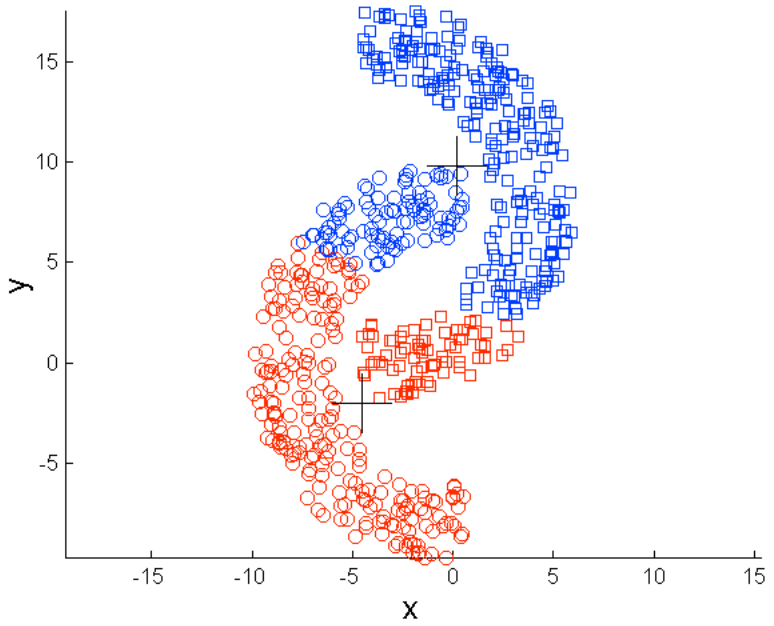


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

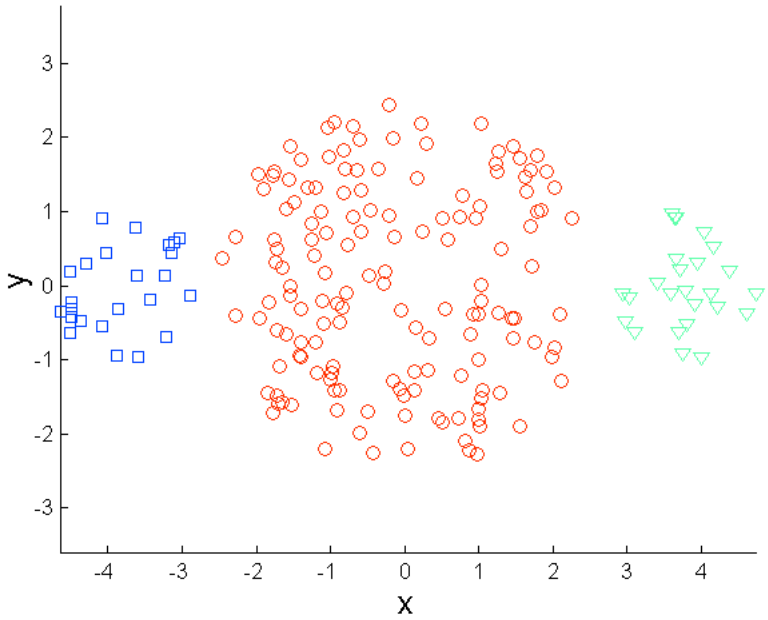


Original Points

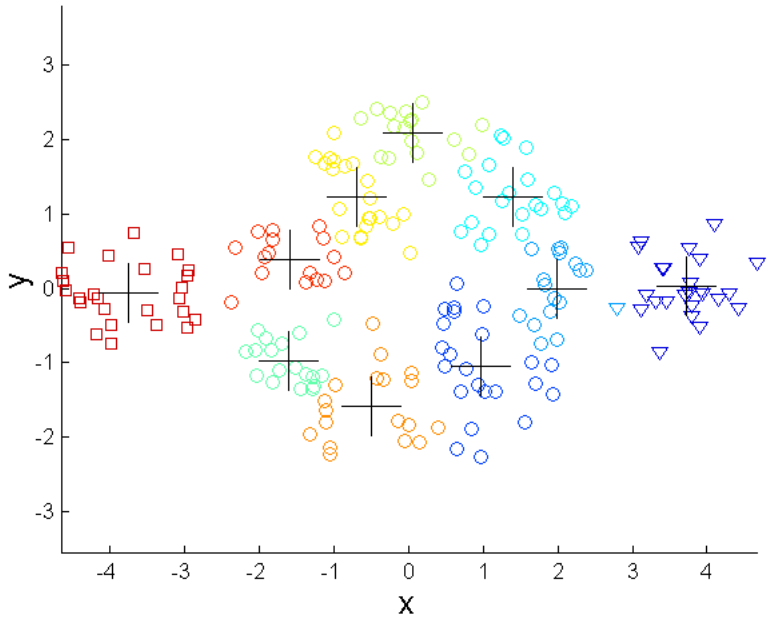


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points

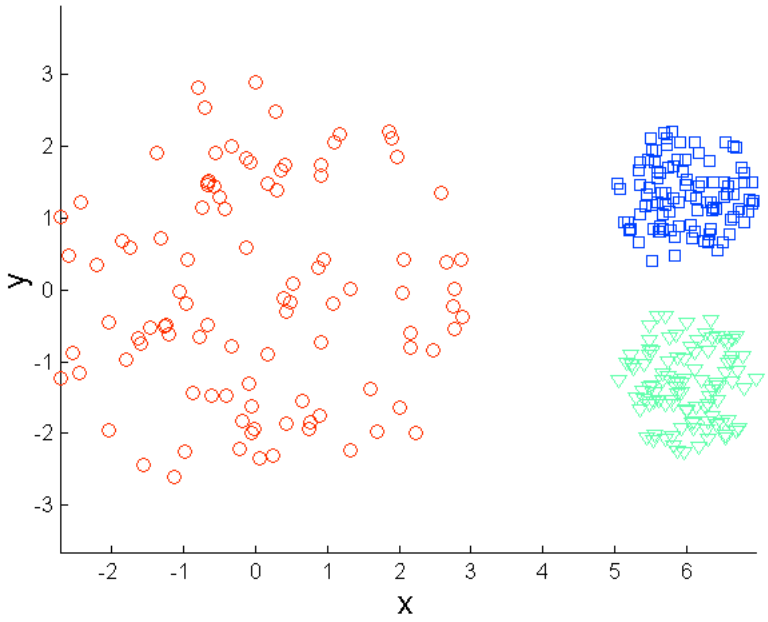


K-means Clusters

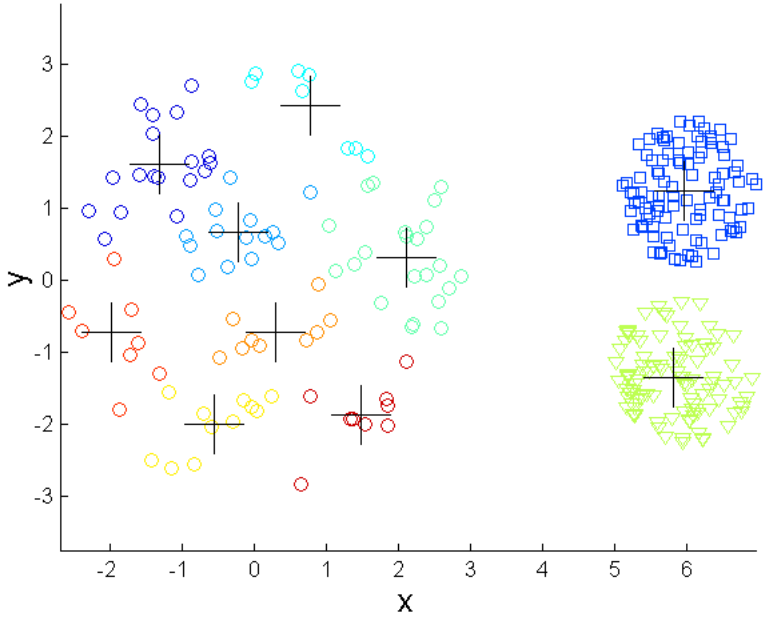
One solution is to use many clusters.

Find parts of clusters, but need to put together.

Overcoming K-means Limitations

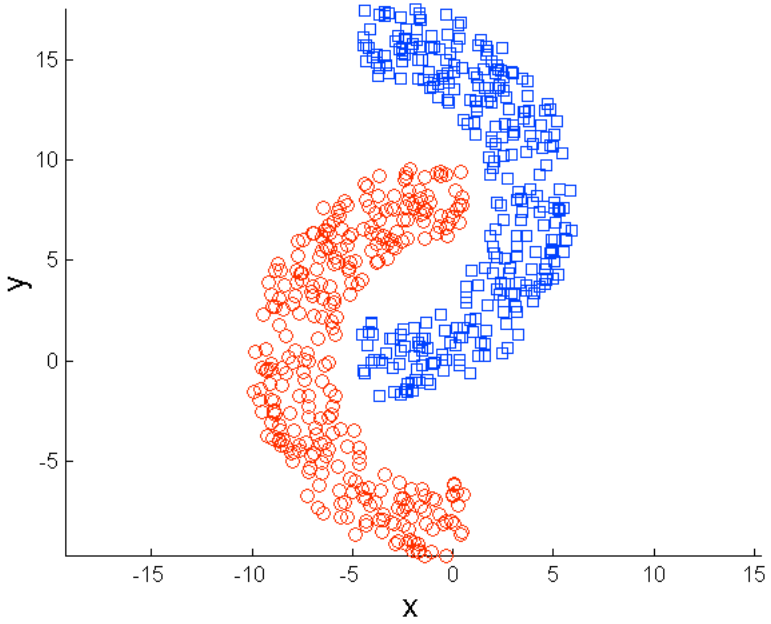


Original Points

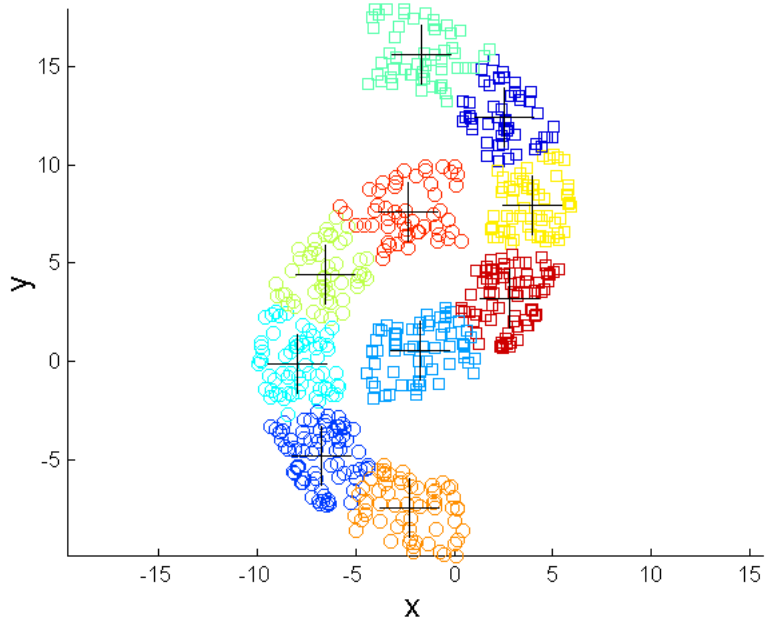


K-means Clusters

Overcoming K-means Limitations

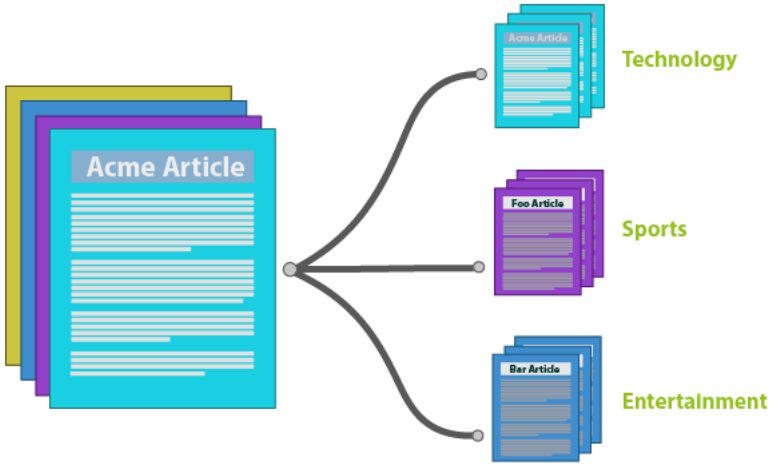


Original Points

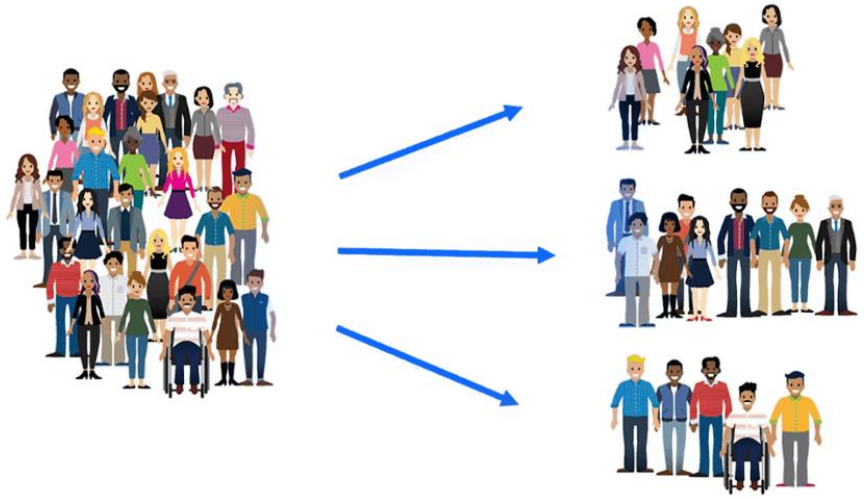


K-means Clusters

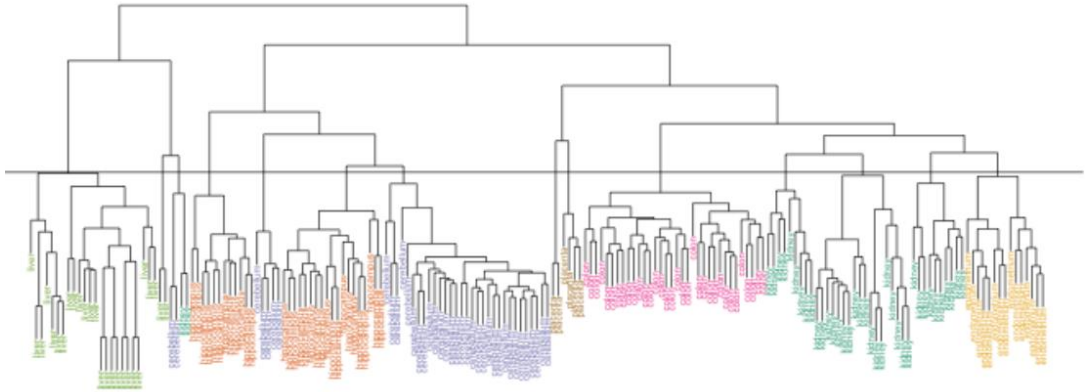
- 主要聚类分析方法分类
  - 划分方法（Partitioning Methods）
  - 分层方法
  - 基于密度的方法
  - 基于表格的方法
  - 基于模型（Model-Based）的聚类方法



Document clustering



Market segmentation

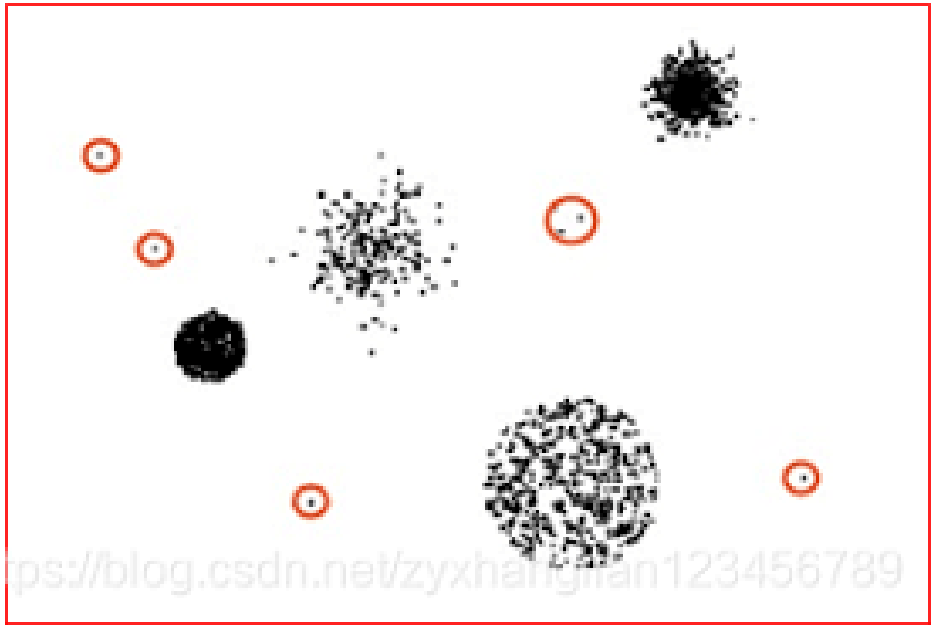


Genomic Clustering

聚类分析的应用

- 关联规则挖掘：
  - 在交易数据、关系数据或其他信息中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性、或因果结构。

- 离群点分析
  - 离群点:一些与数据的一般行为或模型不一致的孤立数据
  - 通常孤立点被作为“噪音”或异常被丢弃，但在欺骗检测中却可以通过对罕见事件进行孤立点分析而得到结论。
- 应用
  - 信用卡欺诈检测
  - 移动电话欺诈检测
  - 客户划分
  - 医疗分析（异常）



离群点分析

- 异常检测是数据挖掘中一个重要方面，用来发现”小的模式”(相对于聚类)，即数据集中间显著不同于其它数据的对象。
- 异常探测应用
  - 电信和信用卡欺骗
  - 贷款审批
  - 药物研究
  - 气象预报
  - 金融领域
  - 客户分类
  - 网络入侵检测
  - 故障检测与诊断等

异常检测

第13章
数据挖掘基础

1 初识数据挖掘

2 数据挖掘标准流程

3 数据挖掘的技术

4 **大数据挖掘**

- “大量”、“多源、异质、复杂”、“动态”、“价值高但价值密度低”的大数据特征决定了大数据挖掘技术不同于之前的数据挖掘技术。大数据挖掘技术包括：
  - 高性能计算支持的分布式；
  - 并行数据挖掘技术；
  - 面向多源、不完整数据的不确定数据挖掘技术；
  - 面向复杂数据组织形式的图数据挖掘技术；
  - 面向非结构化稀疏性的超高维数据挖掘技术；
  - 面向价值高但价值密度低特征的特异群组挖掘技术；
  - 面向动态数据的实时、增量数据挖掘技术等。

大数据挖掘技术

- 大数据的“大”通常是指PB级以上的，这与之前的数据挖掘技术针对的数据对象的规模不同。这一特征需要更高性能的计算平台支持，考虑大规模数据的分布式、并行处理，对数据挖掘技术带来的挑战是I/O交换、数据移动的代价高，还需要在不同站点间分析数据挖掘模型间的关系。
- 大数据环境下，需要新的云计算基础架构支撑（例如，Hadoop、Spark等）。

并行分布式数据挖掘

- 大数据挖掘的数据对象常常具有不确定、不完整的特点，这要求大数据挖掘技术能够处理不确定、不完整的数据集，并且考虑多源数据挖掘模型和决策融合。
- 数据挖掘一直以来重视数据质量。数据的质量决定数据挖掘结果的价值。然而，大数据环境下，数据获取能力逐渐高于数据分析能力。
- 大数据挖掘技术要有更强地处理不确定、不完整数据集的能力。

不确定数据挖掘

- 大数据下，来自文本、图像、视频的数据挖掘应用更加广泛，非结构化数据给数据挖掘技术带来了新的要求，大数据挖掘算法设计要考虑超高维特征和稀疏性。
- 超高维特征分析的需求使得深度学习技术成为热点。
- 大数据环境下，深度学习与大数据的结合，也将成为寻找大数据其中规律的重要支撑技术之一。

超高维数据挖掘

- 时序数据挖掘是数据挖掘领域的一个研究主题。然而，大数据环境下，数据的获取更加高速，关键是处理数据的需求在实时性方面的要求更高。
- 早期的数据挖掘总是能容忍分钟级别，甚至更长时延的响应。现在，许多领域已经使用数据挖掘技术分析本领域数据，各个领域对数据挖掘结果响应需求存在差异，不少领域需要有更到的响应度，例如实时在线精准广告投放、证券市场高频交易等。

实时数据挖掘

- 大数据环境下，产生了新的数据挖掘任务。其中，特异群组是一类低密度高价值的数据，特异群组是指在众多行为对象中，少数对象群体具有一定数量的相同或相似的行为模式，表现出相异于大多数对象而形成异常的组群。
- 特异群组挖掘问题既不是异常点挖掘（只发现孤立点）问题也不是聚类问题（将大部分数据分组），是一类全新的问题。

特异群组挖掘

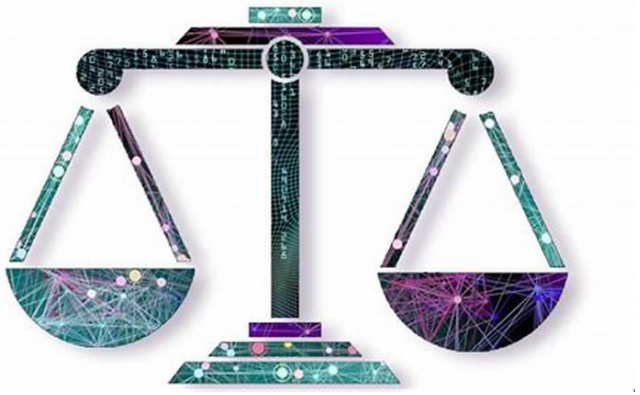
- Google的搜索引擎
  - PageRank采用大量增加输入数据量的方法，不仅仅把文字和标题考虑进去，还考虑了大量的“链接”数据；
  - 谷歌的强不是强在PageRank算法上，而是赢在数据上，利用了比别人多的数据。
- 算法的“好坏”在没有大量有效数据的支撑下是没有意义的。很多算法的结果的质量完全取决于其和真实数据的拟合程度。
- 正方：Google直言，更多的数据胜过更好的算法。
- 反方：数据只是基础，如何建构起有效的算法、模型比数据本身更重要。

数据挖掘对传统算法的挑战

- 滴滴打车平台使得乘客、出租车司机、车辆信息、交通路况、地理信息等相关数据得到了充分的连接；
- 由于市场的变化和巨头的合并，打车软件背后的算法变得单一，决策权开始集中到少数人手上。
- 加价算法模式，开始成为人们唯一的选择。
- 不公平性的产生！

数据挖掘对传统算法的挑战

- 由于每个人看到的内容都是不同的，那么AI是否会把更贵的产品卖给我，或是更偏激的观点推送给我？
- 在现实中是否有一种方法，能够将反歧视——或者说算法公平——植入到数据挖掘与机器学习模型中呢？



大数据“杀熟”？

- 算法有意或无意的把人类的成见、误解和偏见等编码到管理我们生活方方面面的软件系统中，进而重现定义我们现在的生活。
  - 能上哪个学校、在哪里工作、能否获得购车贷款、健康保险的缴费标准是多少等各种决策，越来越多地由数学模型和算法决定；
- 现在使用的很多模型和算法都是不透明的，未受到规制的，明明有错却容不得质疑的。

数据需要透明，数据的使用方式（算法）也需要透明

算法的歧视