第 7 章　机器学习

教材： 王万良《人工智能导论》（第4版）

https://www.icourse163.org/course/ZJUT-1002694018

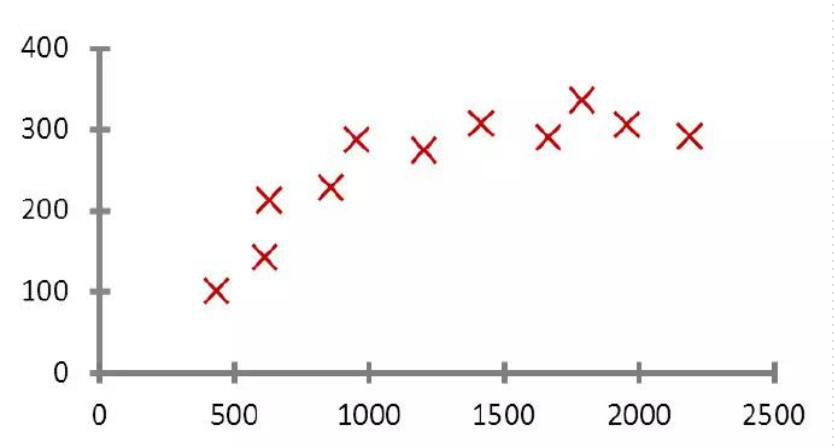
社区资源： https://github.com/Microsoft/ai-edu

参考MOOC： 商务数据分析（赵卫东、上海复旦）

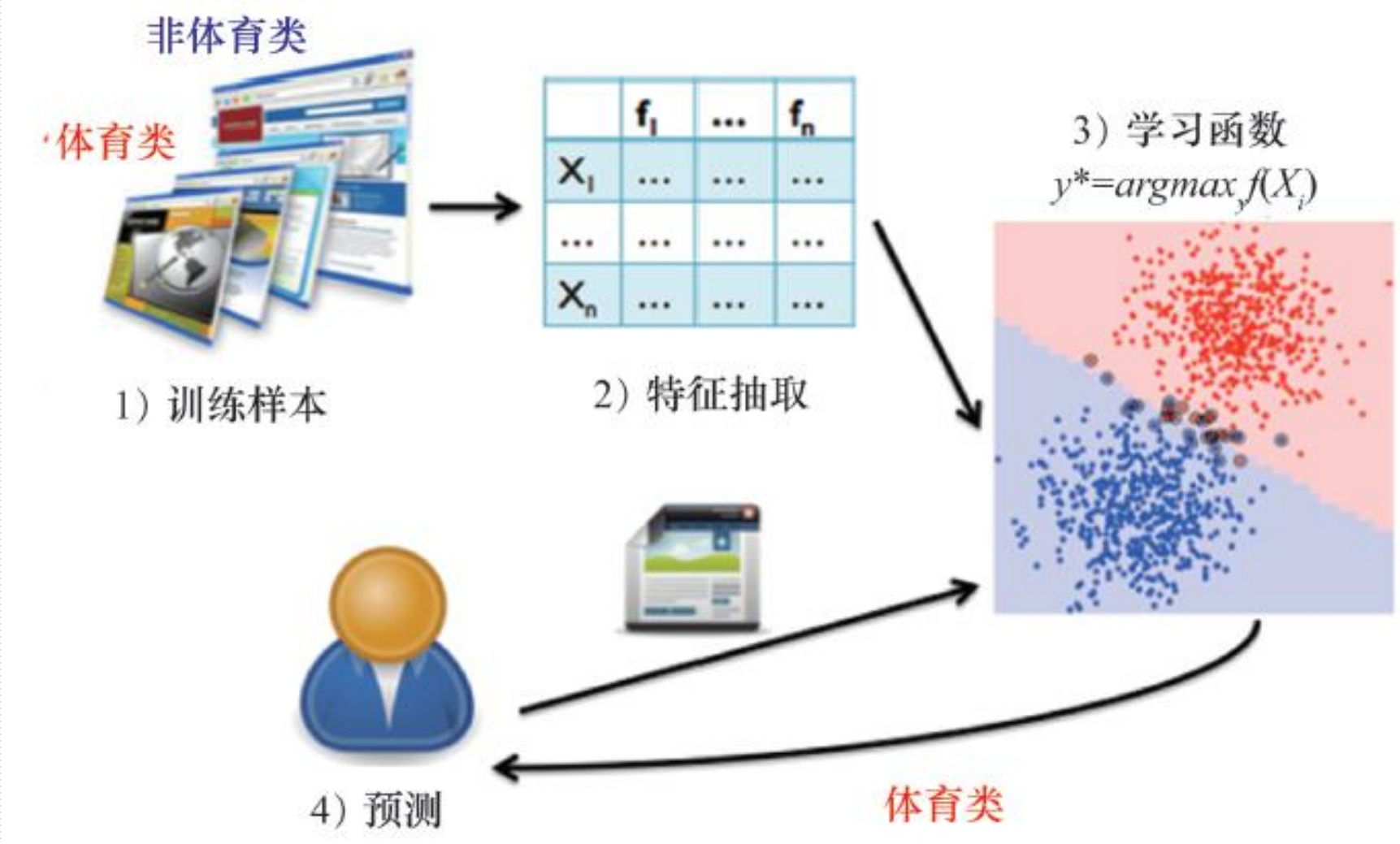
参考课件： 百度深度学习培训

开篇实例：房屋出售

问题：现在我手里有一栋房子需要售卖，我应该给它标上多大的价格？房子的面积是100平方米，价格是100万，120万，还是140万？



网页分类问题



第7章 专家系统与机器学习

- 7.5 机器学习
- 7.6 知识发现与数据挖掘

7.5 机器学习

- 学习是人类具有的一种重要智能行为，但究竟什么是学习，长期以来却众说纷纭。
  - 社会学家、逻辑学家和心理学家都各有其不同的看法。
  - 至今，还没有统一的“机器学习”定义，而且也很难给出一个公认的和准确的定义。
- 机器学习是从人工智能中产生的一个重要学科分支，是实现智能化的关键

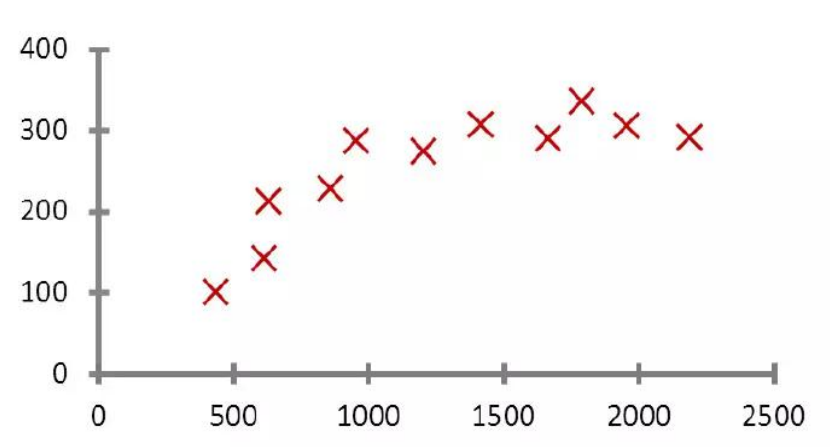
什么是机器学习？

- 机器学习（Machine Learning）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科 。专门研究计算机怎样模拟或实现人类的学习行为，以获取新知识或技能，重新组织已有的知识结构使之不断改善自身的性能。——百度百科
- Machine learning is the study of algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of sample data, known as “training data”, in order to make predictions or decisions without being explicitly programmed to perform the task.——Wikipedia

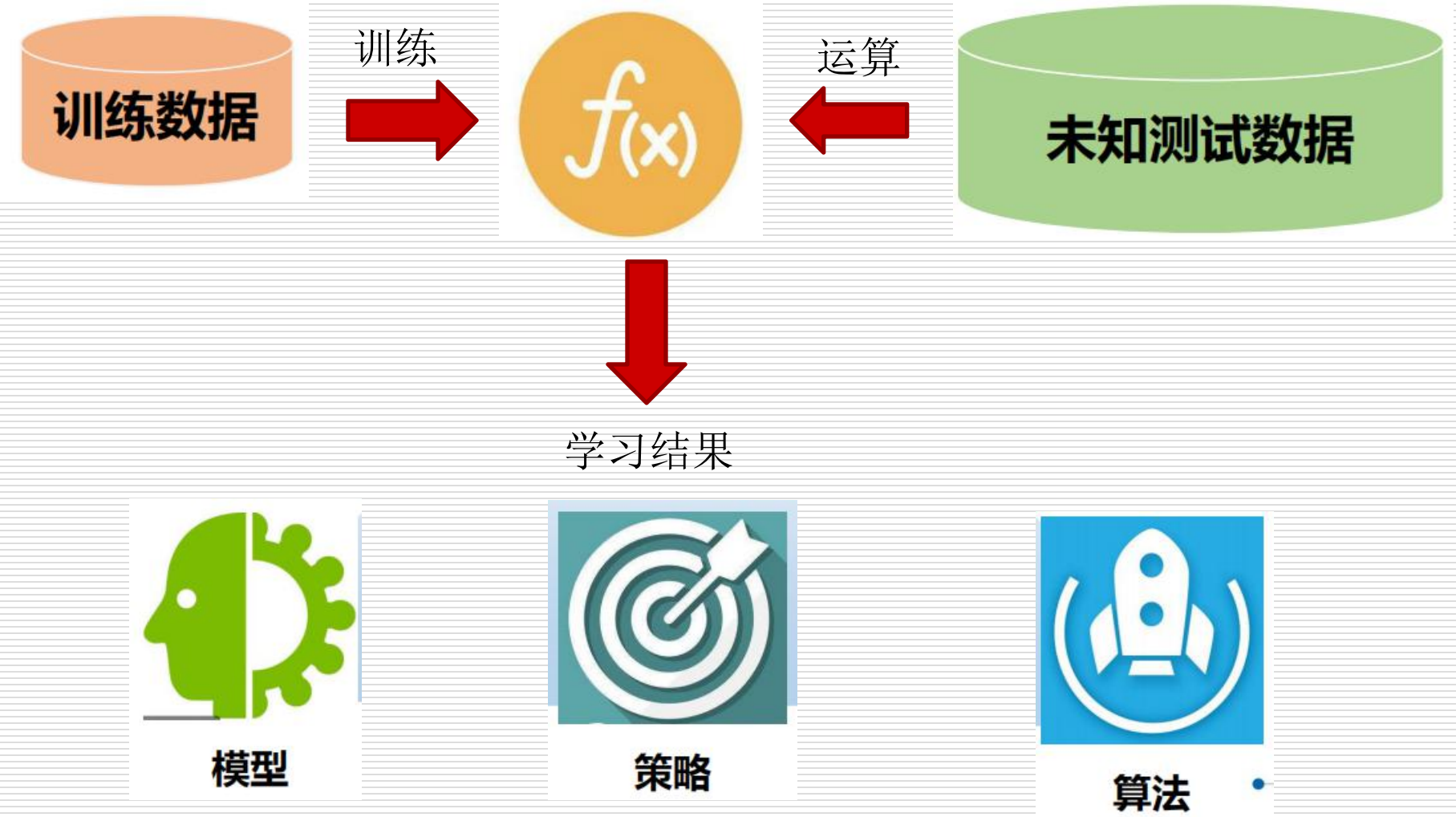
什么是机器学习？

- 从广义上来说，机器学习是一种能够赋予机器学习的能力以此让它完成直接编程无法完成的功能的方法。
- 但从实践的意义上来说，机器学习是一种通过利用数据，训练出模型，然后使用模型预测的一种方法。

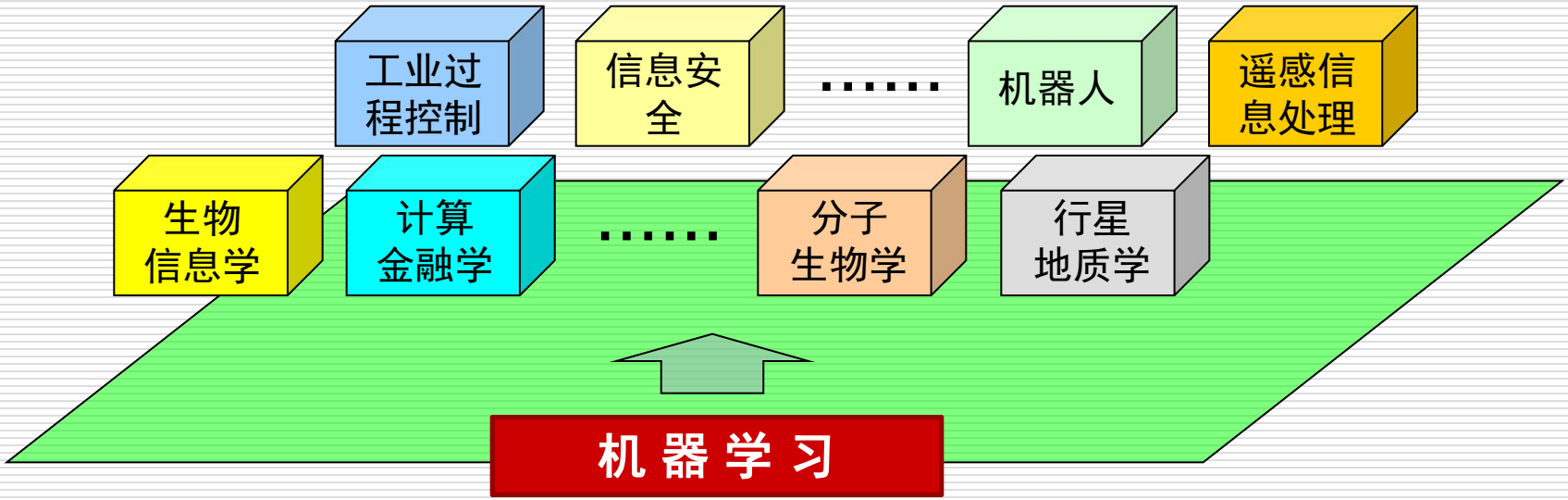
问题：现在我手里有一栋房子需要售卖，我应该给它标上多大的价格？房子的面积是100平方米，价格是100万，120万，还是140万？



机器学习的一般过程



机器学习的重要性



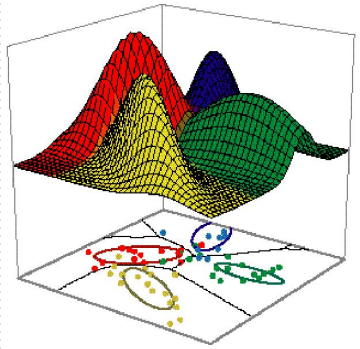
美国JPL实验室的科学家在《Science》（2001年9月）上撰文指出：

- 机器学习对科学研究的整个过程正起到越来越大的支持作用，……，该领域在今后的若干年内将取得稳定而快速的发展

Machine Learning

- We believe machine learning will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, machine learning has the potential to amplify every aspect of a working scientist's progress to understanding. It will also, for better or worse, endow intelligent computer systems with some of the general analytic power of scientific thinking.

——*Science, 14 September, 2001*



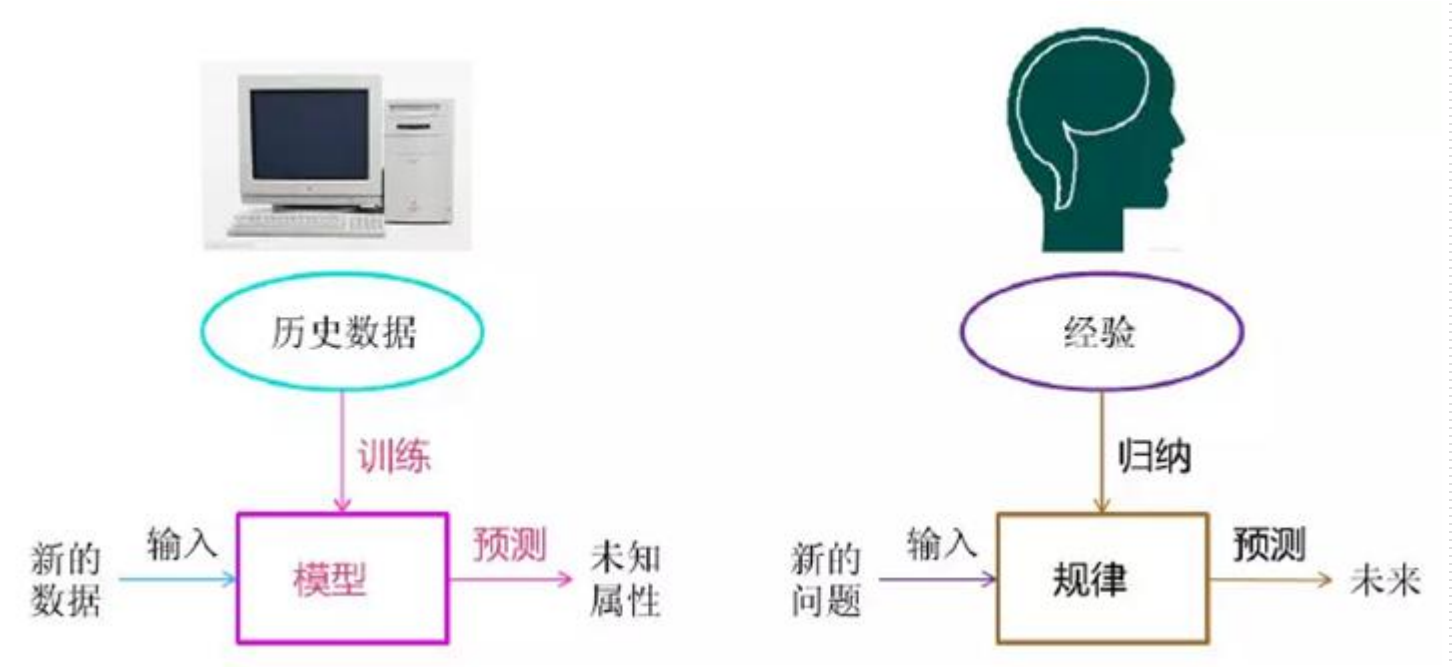
典型机器学习应用领域

- 机器学习能够显著提高企业的智能水平，增强企业的竞争力，人工智能对于各行业的影响越来越大
- 机器学习应用的典型领域有网络安全、搜索引擎、产品推荐、自动驾驶、图像识别、识音识别、量化投资、自然语言处理等。
- 随着海量数据的累积和硬件运算能力的不断提升，机器学习的应用领域还在快速地延展。

网络安全

- 反垃圾邮件
- 反网络钓鱼
- 上网内容过滤
- 反诈骗
- 防范攻击
- 活动监视
- 密码破解
- 无边界攻击模型 & 限制边界攻击模型

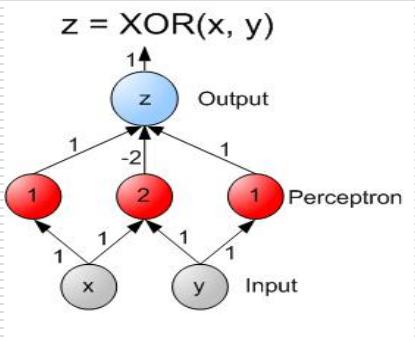
机器学习与人类思考的类比



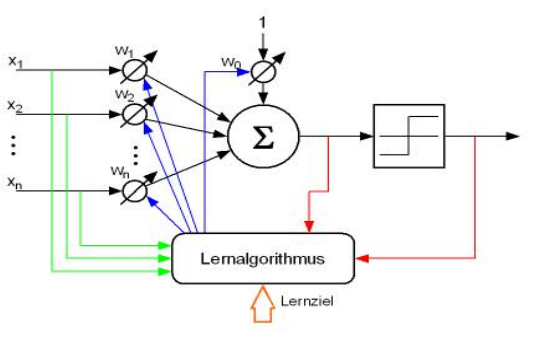
发展历程

- 推理期（20世纪50-70年代初）
  - 认为只要给机器赋予逻辑推理能力，机器就能具有智能
  - A. Newell 和 H. Simon 的“逻辑理论家”、“通用问题求解”程序，获得了1975年图灵奖
- 知识期（20世纪70年代中期）
  - 认为要使机器具有智能，就必须设法使机器拥有知识
  - E.A. Feigenbaum 作为“知识工程”之父在 1994 年获得了图灵奖
- 学科形成（20世纪80年代）
  - 20 世纪 80 年代是机器学习成为一个独立学科领域并开始快速发展、各种机器学习技术百花齐放
  - 1980 年美国卡内基梅隆大学举行第一届机器学习研讨会
  - 1990 年《机器学习:风范与方法》出版
- 繁荣期（20世纪80年代-至今）
  - 20 世纪 90 年代后，统计学习方法占主导，代表为SVM
  - 2006 -，大数据分析的需求，神经网络又被重视，成为深度学习理论的基础
  - 2018年，图灵奖得主是 Yoshua Bengio、Geoffrey Hinton 和 Yann LeCun 三位深度学习巨头。

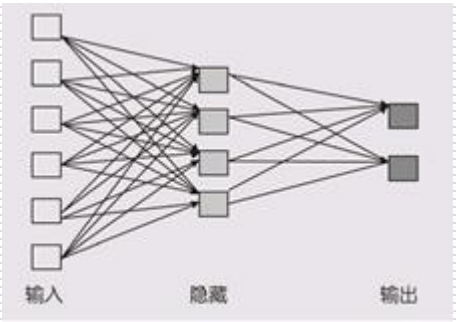
机器学习的发展史



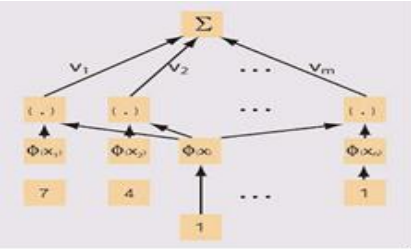
感知机



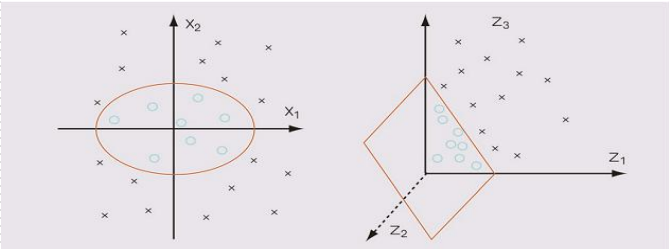
线性适应元

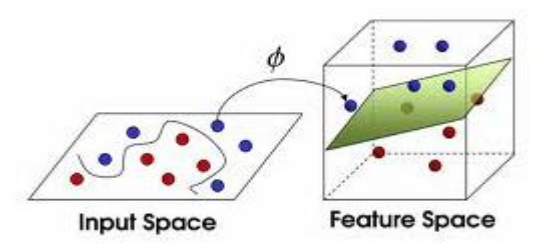


连接主义学习模型



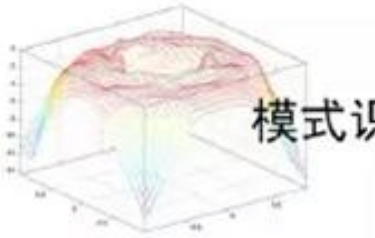
统计学习模型





“核方法”机器学习

机器学习与相关学科



模式识别

计算机视觉

数据挖掘

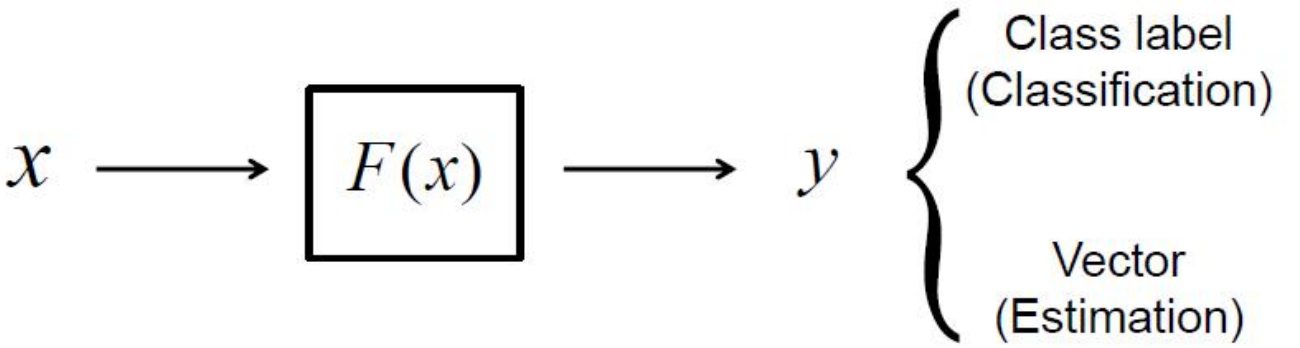
机器学习

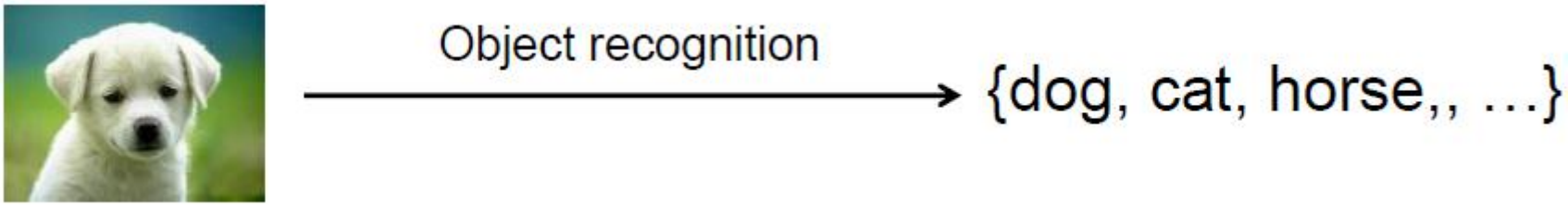
语音识别

统计学习

自然语言处理

机器学习的基本任务

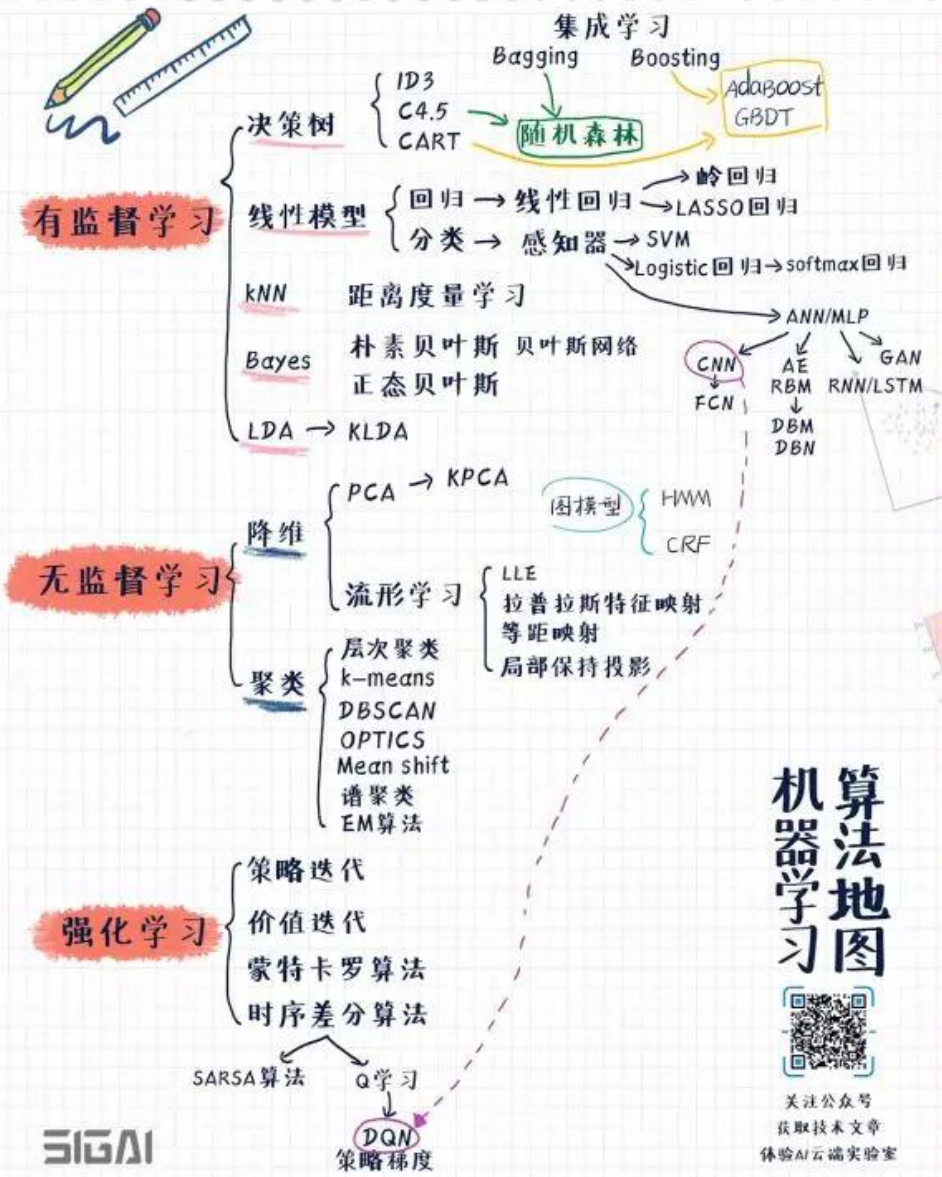

x
$F(x)$
y
Class label
(Classification)
Vector
(Estimation)


Object recognition
{dog, cat, horse,, …}

机器学习的方法

- 按照训练的数据有无标签，可以将机器学习方法分为监督学习算法和无监督学习算法，但推荐算法较为特殊，既不属于监督学习，也不属于非监督学习，是单独的一类。
  - 监督学习算法：线性回归，逻辑回归，神经网络，SVM
    - 监督学习是从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。其训练集要求包括输入和输出，也可以说是特征和目标。训练集的目标通常是由人标注的
  - 无监督学习算法：聚类算法，降维算法
    - 不需要人力来输入标签
  - 强化学习
    - 通过观察来学习做成如何的动作。每个动作都会对环境有所影响，学习对象根据观察到的周围环境的反馈来做出判断。强化学习强调如何基于环境而行动，以取得最大化的预期利益。

机器学习的方法



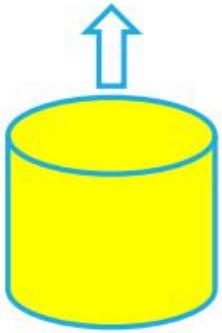
算法	类型	是否支持多分类	预测函数	优化目标	求解算法
贝叶斯分类器	有监督 生成模型 非线性	是	$\arg\max_y P(x\|y)P(y)$ $\arg\min_c \beta\|x_c\|+(x-\mu_c)^T\Sigma_c^{-1}(x-\mu_c)$	对数似然函数 $\max\sum_{i=1}^{N}\ln p(x_i;\theta)$	公式解
决策树	有监督 判别模型 非线性	是	分段常数函数	$\max\frac{\sum N_{t,i}}{N_t}+\frac{\sum N_{R,i}}{N_R}$	贪心法 穷举搜索
kNN	有监督 判别模型 非线性	是	$C_i=\text{vote}(x)$ $\arg\max_i C_i$	无	无
PCA	无监督 线性		$y=Wx$	特征值问题 $Se=\lambda e$	QR算法
LLE (局部线性嵌入)	无监督 非线性			$\min_{w_{ij}}\sum_{i=1}^{N}\|x_i-\sum_{j=1}^{k}w_{ij}x_j\|^2$ $\min_{y_i}\sum_{i=1}^{N}\|y_i-\sum_{j=1}^{k}w_{ij}y_j\|^2$	公式解
LDA	有监督 线性		$y=Wx$	$S_{\beta}w=\lambda S_w w$	QR算法
人工神经网络	有监督 判别模型 非线性	是	多层复合函数 $f_2(w^{(2)}f_1(w^{(1)}\cdots)+b^{(2)})$	$\min_w\frac{1}{L}\sum_{i=1}^{L}L(x_i,y_i,w)$	梯度下降法 +反向传播算法
SVM	有监督 判别模型 非线性	不直接支持	$\text{Sgn}(\sum_i a_i y_i K(x_i^T x)+b)$	$\min\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j y_i y_j K(x_i^T x_j)-\sum_{i=1}^{N}a_i$ $0\le a_i\le C$ $\sum_{i=1}^{N}a_i y_i=0$	SMO算法
logistic回归	有监督 判别模型 线性	否	$\text{Sgn}(w^T x+b)$	$\min\sum_{i=1}^{L}(y_i\log h(x_i)+(1-y_i)\log(1-h(x_i)))$	梯度下降法 牛顿法
softmax回归	有监督 判别模型 线性	是	$h_\theta(x)=\frac{1}{\sum_{i=1}^{k}e^{\theta_i^T x}}\begin{bmatrix}e^{\theta_1^T x}\\ \cdots\\ e^{\theta_k^T x}\end{bmatrix}$	$\min-\sum_{i=1}^{L}\sum_{j=1}^{k}\left[I_{y_i=j}\log\frac{\exp(\theta_j^T x_i)}{\sum_{l=1}^{k}\exp(\theta_l^T x_i)}\right]$	梯度下降法
随机森林	有监督 判别模型 非线性	是	弱学习器投票	同决策树	同决策树
Adaboost	有监督 判别模型 非线性	否	$F(x)=\sum_{i=1}^{T}a_i f_i(x)$ $\text{Sgn}(F(x))$	$\min E(\exp(-yF(x)))$	分阶段优化 公式解
CNN	有监督 判别模型 非线性	是	多层复合函数 Cov pooling	$\min_w\frac{1}{L}\sum_{i=1}^{L}L(x_i,y_i,w)$	梯度下降法 BP
RNN	有监督 判别模型 非线性	是	$h_t=f(W_{x_h}X_t+W_{hh}h_{t-1}+b_h)$ $y_t=g(W_{oh}h_t+b_o)$	$L=\sum_{t=1}^{T}L_t$ $L_t=L(y_t^*,y_t)$	BPTT 梯度下降法
GAN	有监督 生成模型 判别模型		多层复合函数	$\min_G\max_D V(D,G)$ $=E_{x\sim P_{data}(x)}[\log D(x)]+$ $E_{z\sim P_G(z)}[\log(1-D(G(z)))]$	反向传播算法 梯度下降法 分阶段优化
DQN	强化学习		多层复合函数	$\min E[(R+\gamma\max_{a'}Q(s',a',\theta)-Q(s,a,\theta))^2]$	梯度下降法 反向传播算法

监督学习的重要元素

监督学习：损失函数

训练映射函数f

使得$f(x_i)$预测结果尽量等于y_i



训练数据集

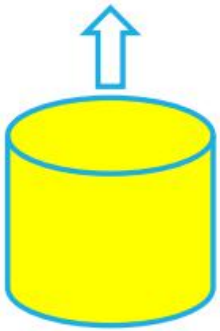
$(x_i, y_i), i = 1, \ldots, n$

- 训练集中一共有n个标注数据，第i个标注数据记为(x_i, y_i),其中第i个样本数据为x_i，y_i是x_i的标注信息。
- 从训练数据中学习得到的映射函数记为f，f对x_i的预测结果记为$f(x_i)$。损失函数就是用来计算x_i真实值y_i与预测值$f(x_i)$之间差值的函数。
- 很显然，在训练过程中希望映射函数在训练数据集上得到“损失”之和最小，即$\min \sum_{i=1}^{n} Loss(f(x_i), y_i)$。

监督学习：损失函数

训练映射函数f

使得$f(x_i)$预测结果尽量等于y_i



训练数据集

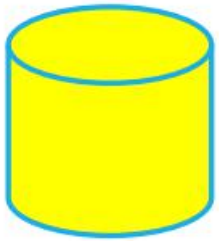
$(x_i, y_i), i = 1, \ldots, n$

损失函数名称	损失函数定义
0-1损失函数	$Loss(y_i, f(x_i)) = \begin{cases} 1, f(x_i) \neq y_i \\ 0, f(x_i) = y_i \end{cases}$
平方损失函数	$Loss(y_i, f(x_i)) = (y_i - f(x_i))^2$
绝对损失函数	$Loss(y_i, f(x_i)) = \lvert y_i - f(x_i) \rvert$
对数损失函数/对数似然损失函数	$Loss(y_i, P(y_i \lvert x_i)) = -logP((y_i \lvert x_i)$

典型的损失函数

监督学习：训练数据与测试数据

从训练数据集学习
得到映射函数f

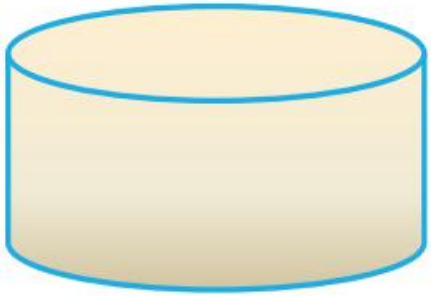


训练数据集
$(x_i, y_i), i = 1, \ldots, n$

在测试数据集
测试映射函数f

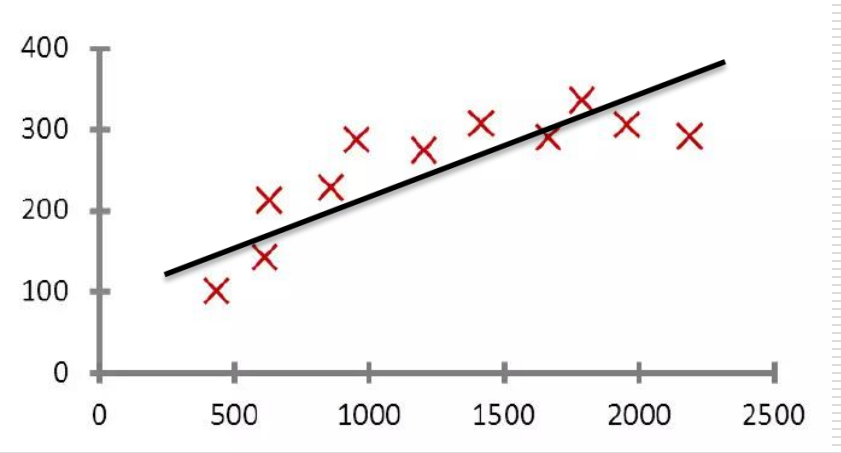
测试数据集
$(x_i', y_i'), i = 1, \ldots, m$

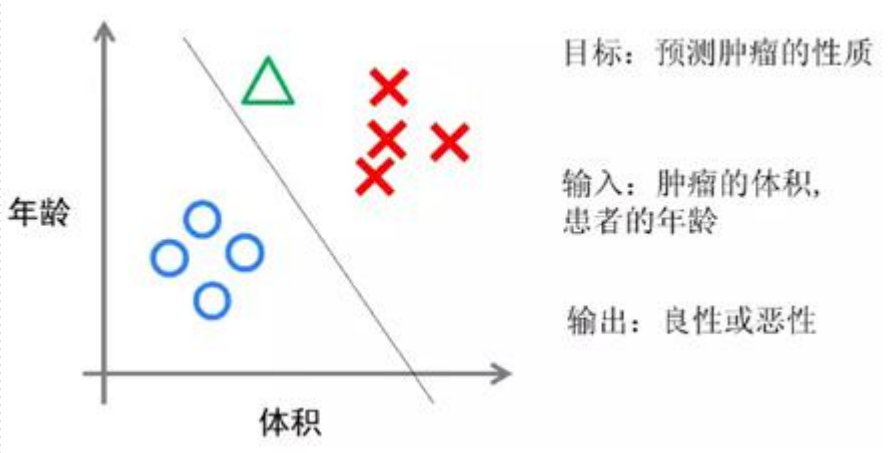
未知数据集
上测试映射函数f



回归方法

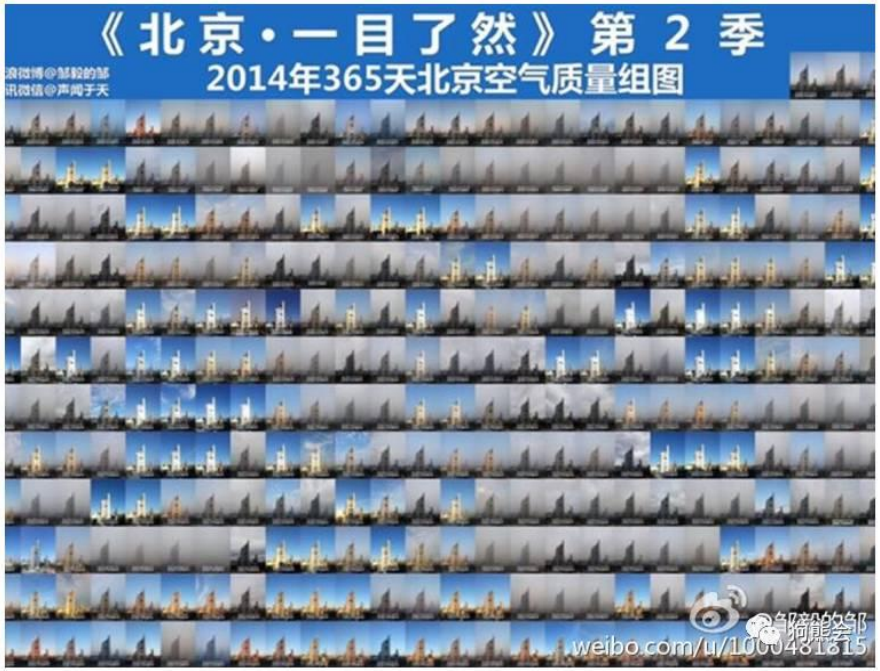
- 线性回归就是前面说过的房价求解问题。如何拟合出一条直线最佳匹配我所有的数据？例如：“最小二乘法”来求解。
  - 分析一个变量与其他一个（或几个）变量之间的相关关系的统计方法就称为回归分析
- 回归算法有两个重要的子类：即线性回归和逻辑回归。





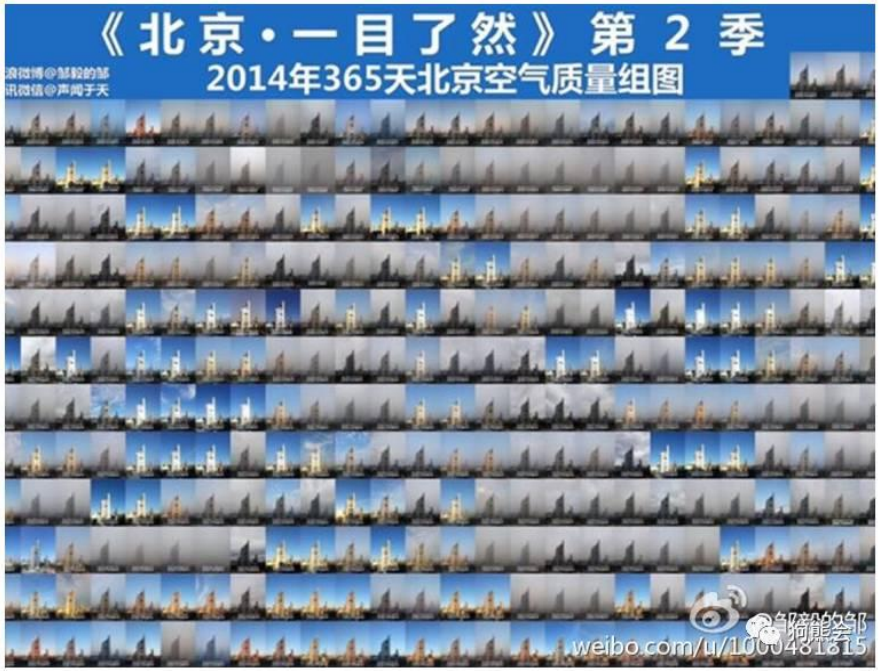
预测城市PM2.5(1)

- 通过图片识别PM2.5



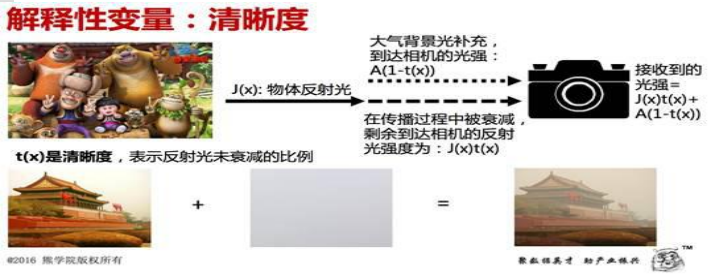
预测城市PM2.5(1)

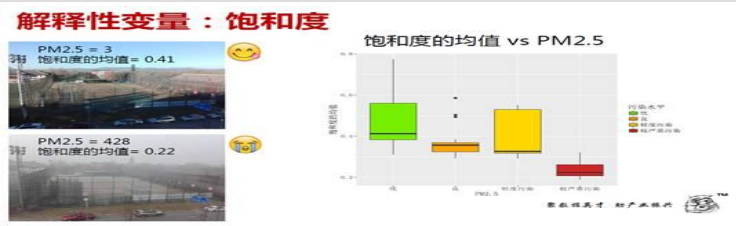
- 通过图片识别PM2.5



预测城市PM2.5(2)

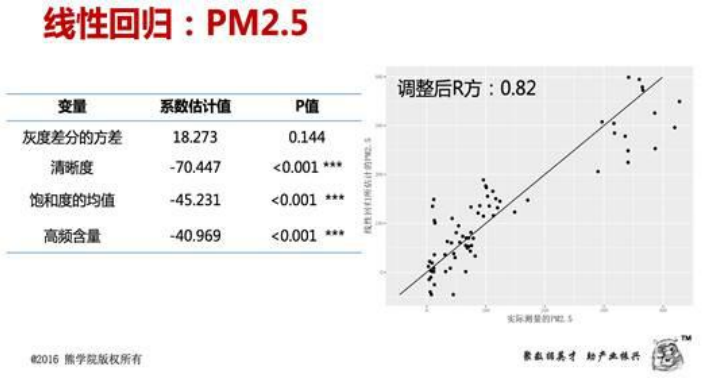
- 从衡量图像清晰程度的角度出发，对图像特征进行观察和分析，得到4个解释性变量：灰度差分的方差、清晰度、饱和度、高频含量等

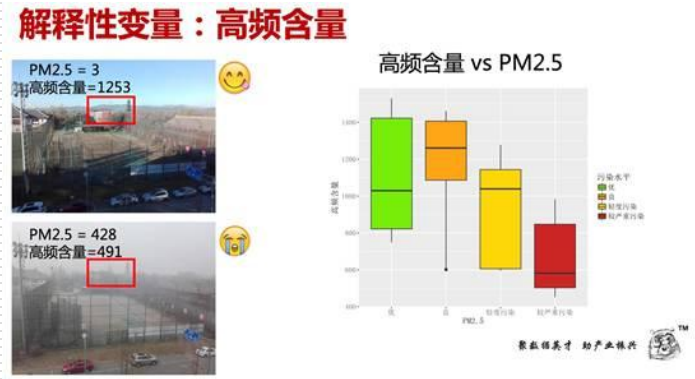




预测城市PM2.5(3)

- 多元线性回归的拟合优度为0.82





逻辑回归

- Logistic回归建立了一个多项式对数回归模型，用于预测二值变量的值(0或1)。相对于独立变量 $x_1, x_2, \ldots, x_n$ ，变量y等于1的概率定义如下：

$$p(y=1 \mid x_1, x_2, \ldots, x_n) = \frac{e^{-(a_1x_1 + a_2x_2 + \ldots + a_nx_n + \mu)}}{1 + e^{-(a_1x_1 + a_2x_2 + \ldots + a_nx_n + \mu)}}$$

- Logistic回归在数据挖掘中很有用，特别是解决两类的数据概率打分问题，如顾客流失风险打分等。

逻辑回归

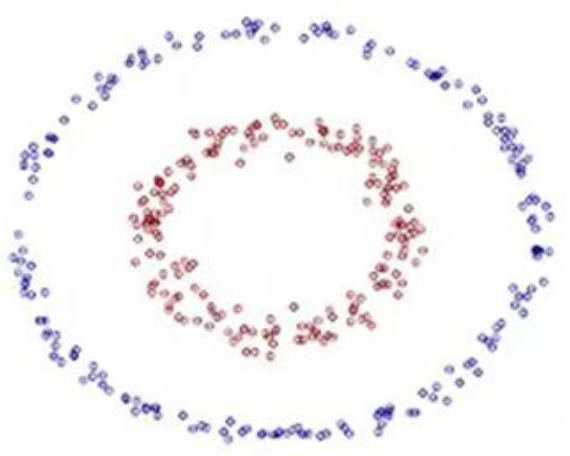
- 应用逻辑回归模型预测银行顾客是否拖欠贷款
  - 根据历史数据识别银行拖欠顾客的特征，预测潜在信贷顾客是否拖欠贷款。这里选取700个信贷顾客的历史记录，其中21.5%是拖欠顾客。这里选择顾客性别（sex）、收入（income）、年龄（age)、edcation(文化程度)，employ(现单位工作年数)，debtinc(负债率)和creddebt（信用卡债务）等作为自变量，顾客拖欠贷款与否作为因变量：1代表拖欠，0代表正常。选择70%历史记录进行训练，剩下30%历史数据用于验证，建立一个预测因变量取1的概率的logistic回归模型，以对新的潜在顾客是否拖欠贷款进行预测。
  - 影响顾客拖欠的自变量比较多，这里采用Forward/Backward方式用于剔除不重要的自变量，例如收入水平、文化程度和年龄等对顾客信用的影响不显著，拖欠概率的回归方程如下：

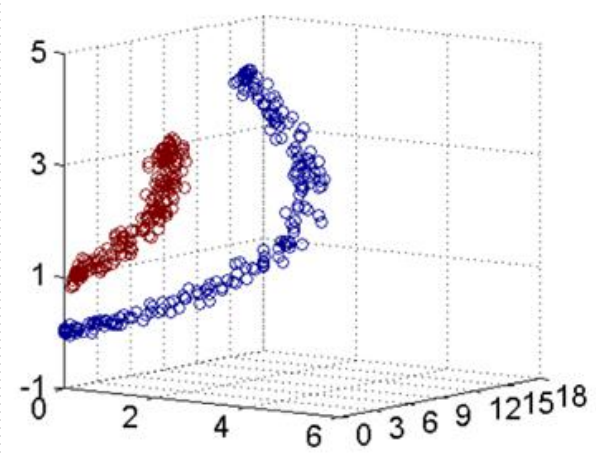
$$\ln\frac{p}{1-p} = -0.76 - 0.249employ - 0.069address + 0.08debtinc + 0.594creddebt$$

  - 对模型进行显著性检验以及回归模型与样本数据的拟合程度以及模型预测精度进行评价，回归模型满足一定要求即可部署使用。从中可以发现拖欠贷款客户的特征：工作不稳定、住址经常变动、债务比率高、信用卡债务多的客户，拖欠贷款的概率较大。

SVM（支持向量机）

- 支持向量机诞生于统计学习界，从某种意义上来说是逻辑回归算法的强化：通过给予逻辑回归算法更严格的优化条件，支持向量机算法可以获得比逻辑回归更好的分类界线。通过跟高斯“核”的结合，支持向量机可以表达出非常复杂的分类界线，从而达成很好的的分类效果。





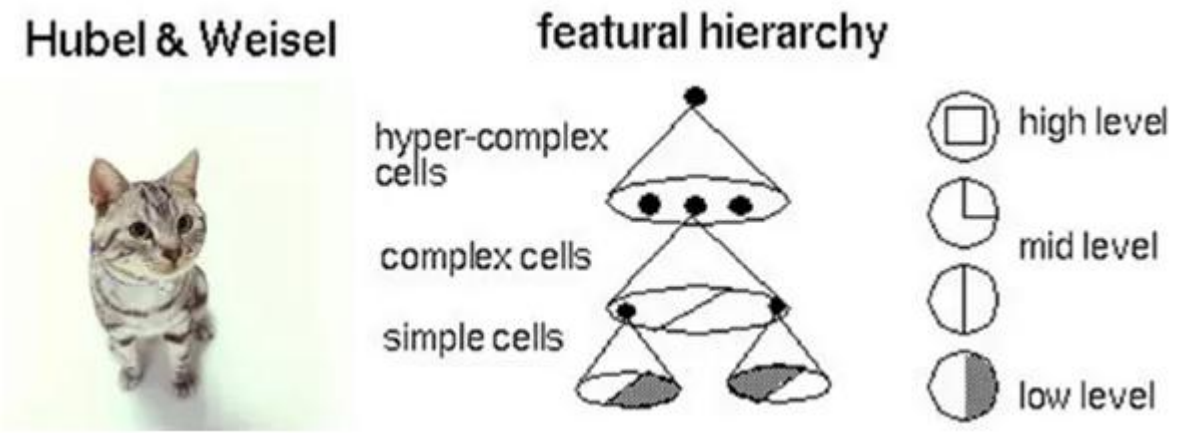
回归与分类的区别

- 两者均是学习输入变量和输出变量之间潜在关系模型，基于学习所得模型将输入变量映射到输出变量

- 监督学习分为回归和分类两个类别。
  - 在回归分析中，学习得到一个函数将输入变量映射到连续输出空间，如价格和温度等，即值域是连续空间。
  - 在分类模型中，学习得到一个函数将输入变量映射到离散输出空间，如人脸和汽车等，即值域是离散空间。

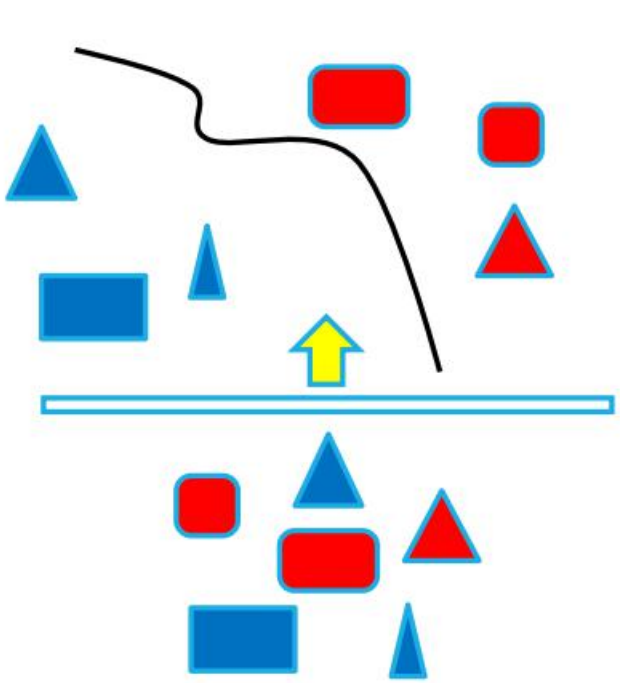
神经网络

- 神经网络(也称之为人工神经网络，ANN)，是80年代机器学习界非常流行的方法，其诞生起源于对大脑工作机理的研究。简单来说，就是分解与整合。

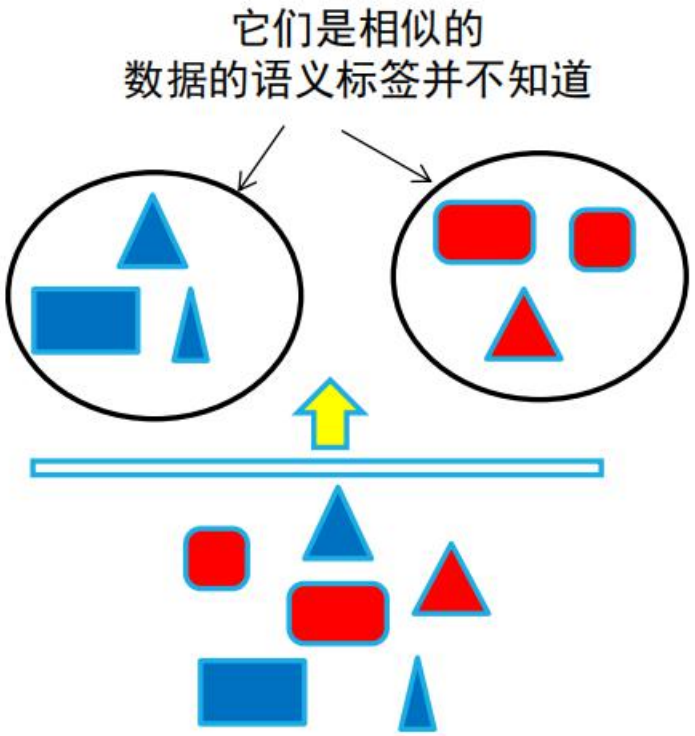


Hubel-Wiesel试验与大脑视觉机理

监督学习 versus 无监督学习



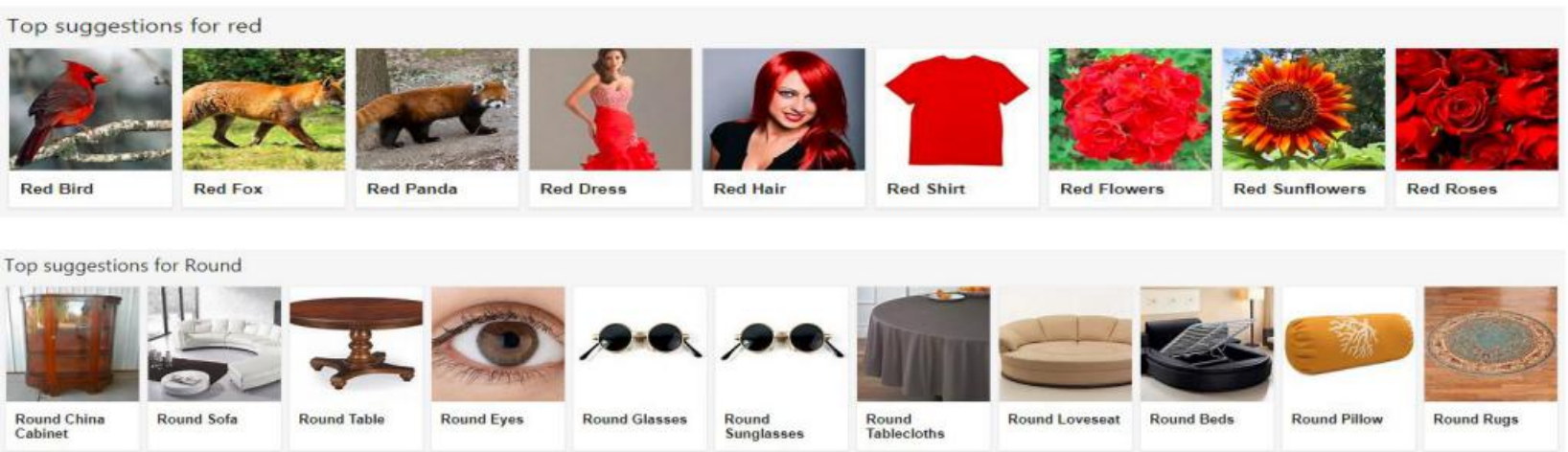
左：监督学习



右：无监督学习

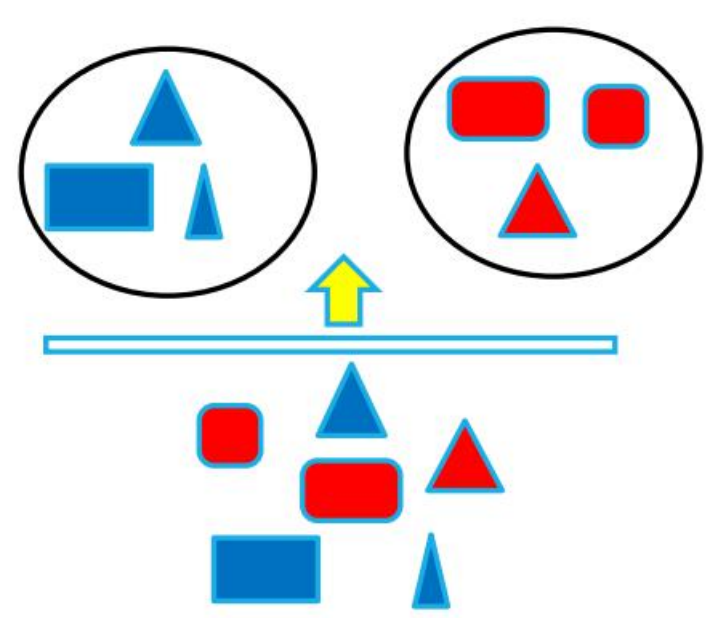
无监督学习的重要因素

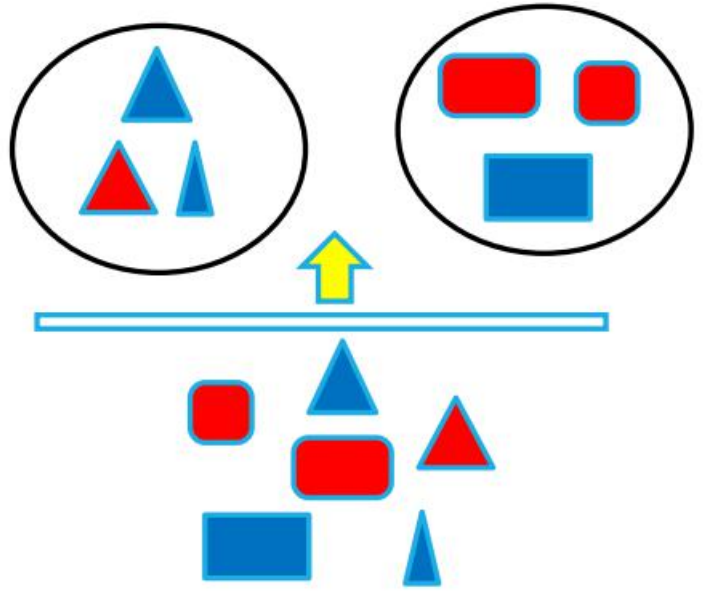
数据特征	图像中颜色、纹理或形状等特征	听觉信息中旋律和音高等特征	文本中单词出现频率等特征
相似度函数	定义一个相似度计算函数，基于所提取的特征来计算数据之间的相似性		



无监督学习：数据特征和相似度函数都很重要

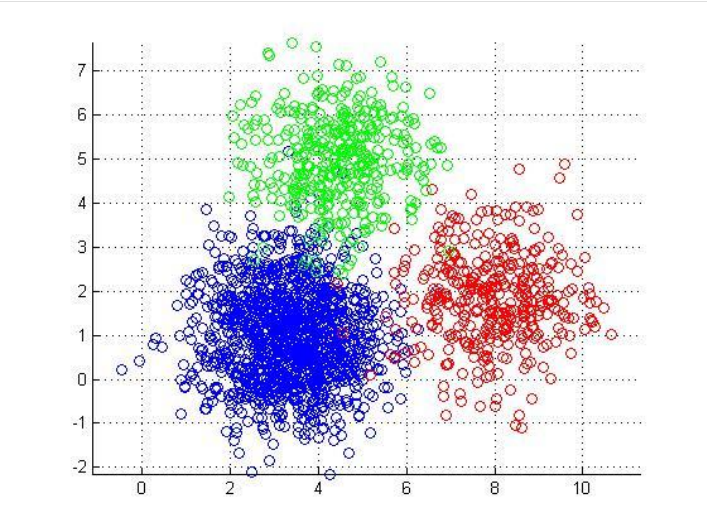
- 相似度函数：颜色相似　　相似度函数：形状相似

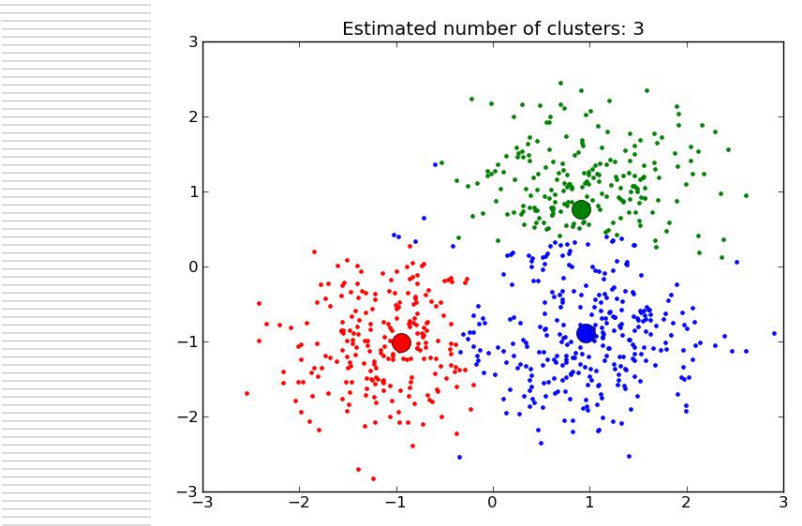




聚类

- 物以类聚，人以群分(《战国策·齐策三》)
- 这类方法有一个统称，即无监督算法，其中最典型的代表就是聚类。聚类就是计算种群中的距离，根据距离的远近将数据划分为多个族群。
- 聚类算法中最典型的代表就是K-Means算法。

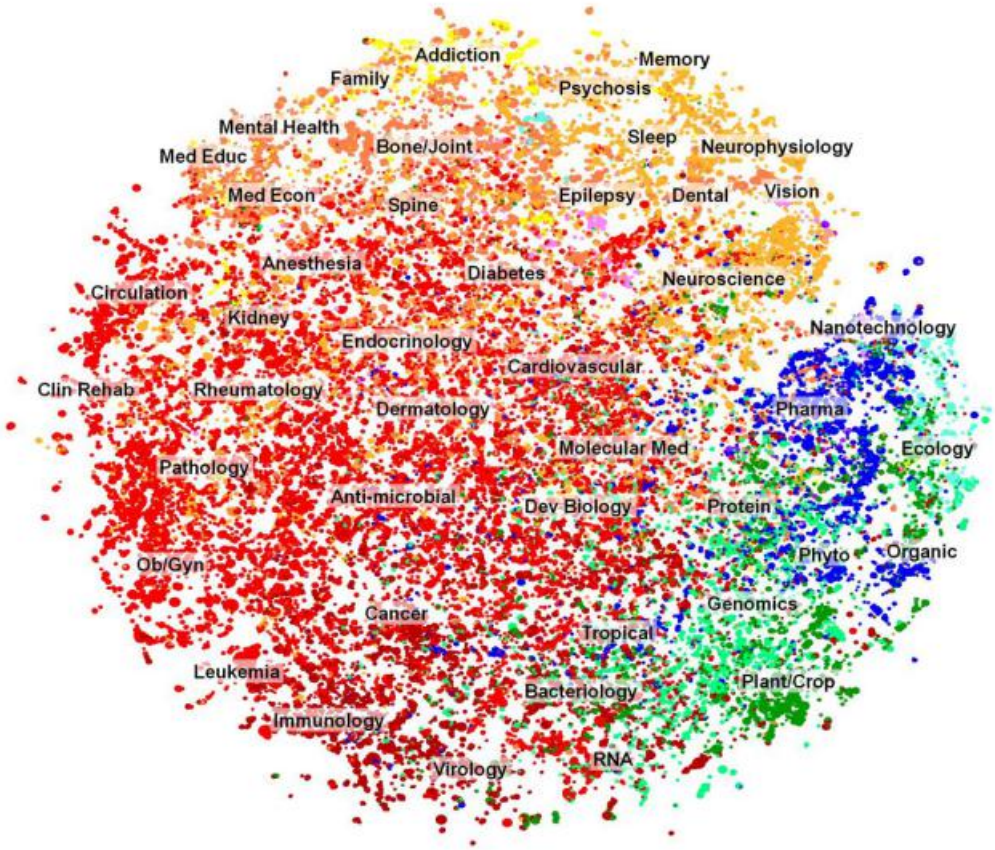




K均值聚类算法的应用

- 图像分类

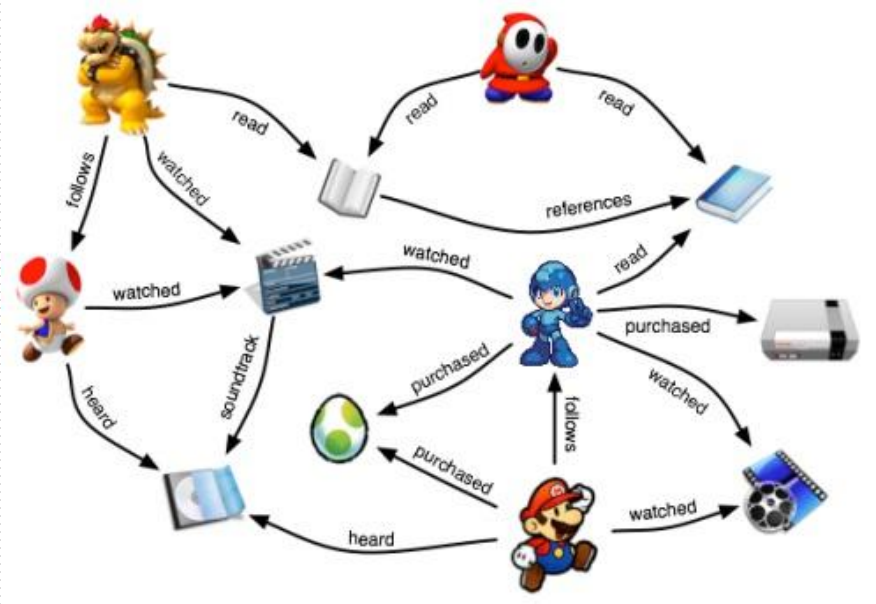
- 文本分类



文本分类：将200多万篇论文聚类到29,000个类别，包括化学、工程、生物、传染疾病、生物信息、脑科学、社会科学、计算机科学等及给出了每个类别中的代表单词

推荐算法

- 推荐算法是目前业界非常火的一种算法，在电商界，如亚马逊，天猫，京东等得到了广泛的运用。推荐算法的主要特征就是可以自动向用户推荐他们最感兴趣的东西，从而增加购买率，提升效益。



机器学习无所不能？

- 问题思考：机器学习是否无所不能？

机器学习面临的难题与挑战

- **数据稀疏性**：训练一个模型，需要大量（标注）数据，但是数据往往比较稀疏。
  - 比如，我们想训练一个模型表征某人“购物兴趣”，但是这个人在网站上浏览行为很少，购物历史很少，很难训练出一个“meaningful model”来预测应该给这个人推荐什么商品等
- **高数量和高质量标注数据需求**：获取标定数据需要耗费大量人力和财力。而且，人会出错，有主观性。如何获取高数量和高质量标定数据，或者用机器学习方法只标注“关键”数据 (active learning) 值得深入研究
- **冷启动问题**：一个好互联网产品，用的人多，得到的数据多；得到的数据越多，模型训练的越好，产品会变得更好用，用的人就会更多 … 进入良性循环。对于一个新产品，在初期，要面临数据不足的冷启动问题
- **泛化能力问题**：训练数据不能全面、均衡的代表真实数据。

机器学习面临的难题与挑战

- **模型抽象困难**：总结归纳实际问题中的数学表示非常困难。
- **模型评估困难**：在很多实际问题中，很难形式化的、定量的评估一个模型的结果是好还是不好？
- **寻找最优解困难**：要解决的实际问题非常复杂，将其形式化后的目标函数也非常复杂，往往在目前还不存在一个有效的算法能找到目标函数的最优值。

机器学习面临的难题与挑战

- Scalability是互联网的核心问题之一。
  - 搜索引擎索引的重要网页超过 100 亿: 如果1台机器每秒处理1000 网页，需要至少100天。所以出现了 MapReduce, MPI, Spark, Pegasus, Pregel, Hama … 等分布式计算构架。选择什么样的计算平台，和算法设计紧密相关 …
- 机器学习面临的难题与挑战
  - 速度是互联网核心的用户体验。线下模型训练可以花费很长时间：比如，Google 某个模型更新一次需要几千台机器，大约训练半年时间。但是，线上使用模型的时候要求一定要 快
- 实时 (real-time)
  - online learning：互联网每时每刻都在产生大量新数据，要求模型随之不停更新，所以online learning是机器学习的一个重要研究方向。

机器学习与数据挖掘

- 数据挖掘是从大量的业务数据中挖掘隐藏、有用的、正确的知识促进决策的执行。
- 数据挖掘的很多算法都来自于机器学习，并在实际应用中进行优化。
- 机器学习最近几年也逐渐跳出实验室，解决从实际的数据中学习模式，解决实际问题。
- 数据挖掘和机器学习的交集越来越大

7.6 知识发现与数据挖掘

- 知识发现和数据挖掘的目的：从数据集中抽取和精化一般规律或模式。
- 知识发现过程分为：数据准备、数据挖掘以及结果的解释评估等三步。
- 1.数据准备：数据选取、数据预处理和数据变换。
  - 数据选取就是根据用户的需要从原始数据库中抽取的一组数据。
  - 数据预处理一般可能包括消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换等。
  - 数据变换是从初始特征中找出真正有用的特征以减少数据开采时要考虑的特征或变量个数。

7.6 知识发现与数据挖掘

- **2.数据挖掘**
  - 数据挖掘阶段首先要确定挖掘的任务或目的是什么，如数据总结、分类、聚类、关联规则或序列模式等。
  - 确定了挖掘任务后，就要决定使用什么样的挖掘算法。同样的任务可以用不同的算法来实现。
- **选择实现算法有两个考虑因素：**
  - 一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；
  - 二是用户或实际运行系统的要求，有的用户可能希望获取描述型的、容易理解的知识，而有的用户系统的目的是获取预测准确度尽可能高的预测型知识。

7.6 知识发现与数据挖掘

- **3.结果解释和评价**
  - 数据挖掘阶段发现的知识模式中可能存在冗余或无关的模式，所以还要经过用户或机器的评价。
  - 若发现所得模式不满足用户要求，则需要退回到发现阶段之前，如重新选取数据，采用新的数据变换方法，设定新的数据挖掘参数值，甚至换一种挖掘算法。
- 由于KDD最终是面向人的，因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂的另一种表示，如把分类决策树转换为“if-then…”规则。

7.6 知识发现与数据挖掘

- 知识发现的任务：
  - 数据总结：对数据进行浓缩，给出它的紧凑描述。
  - 概念描述：从学习任务相关的数据中提取总体特征。
  - 分类：提出一个分类函数或分类模型（也常常称作分类器），该模型能把数据库中的数据项映射到给定类别中的一个。
  - 聚类：根据数据的不同特征，将其划分为不同的类。包括统计方法、机器学习方法、神经网络方法和面向数据库的聚类方法等。
  - 相关性分析：发现特征之间或数据之间的相互依赖关系。
  - 偏差分析：寻找观察结果与参照量之间的有意义的差别。
  - 建模：通过数据挖掘，构造出能描述一种活动、状态或现象的数学模型。

7.6 知识发现与数据挖掘

- 知识发现的主要方法：
  - 1.统计方法：从事物的外在数量上的表现去推断事物可能的规律性。常见的有回归分析、判别分析、聚类分析以及探索分析等。
  - 2.粗糙集：粗糙集是具有三值隶属函数的模糊集，即是、不是、也许。常与规则归纳、分类和聚类方法结合起来使用。
  - 3.可视化：把数据、信息和知识转化为图形等，使抽象的数据信息形象化。信息可视化也是知识发现的一个有用的手段。
  - 4.传统机器学习方法：包括符号学习和连接学习。

7.6 知识发现与数据挖掘

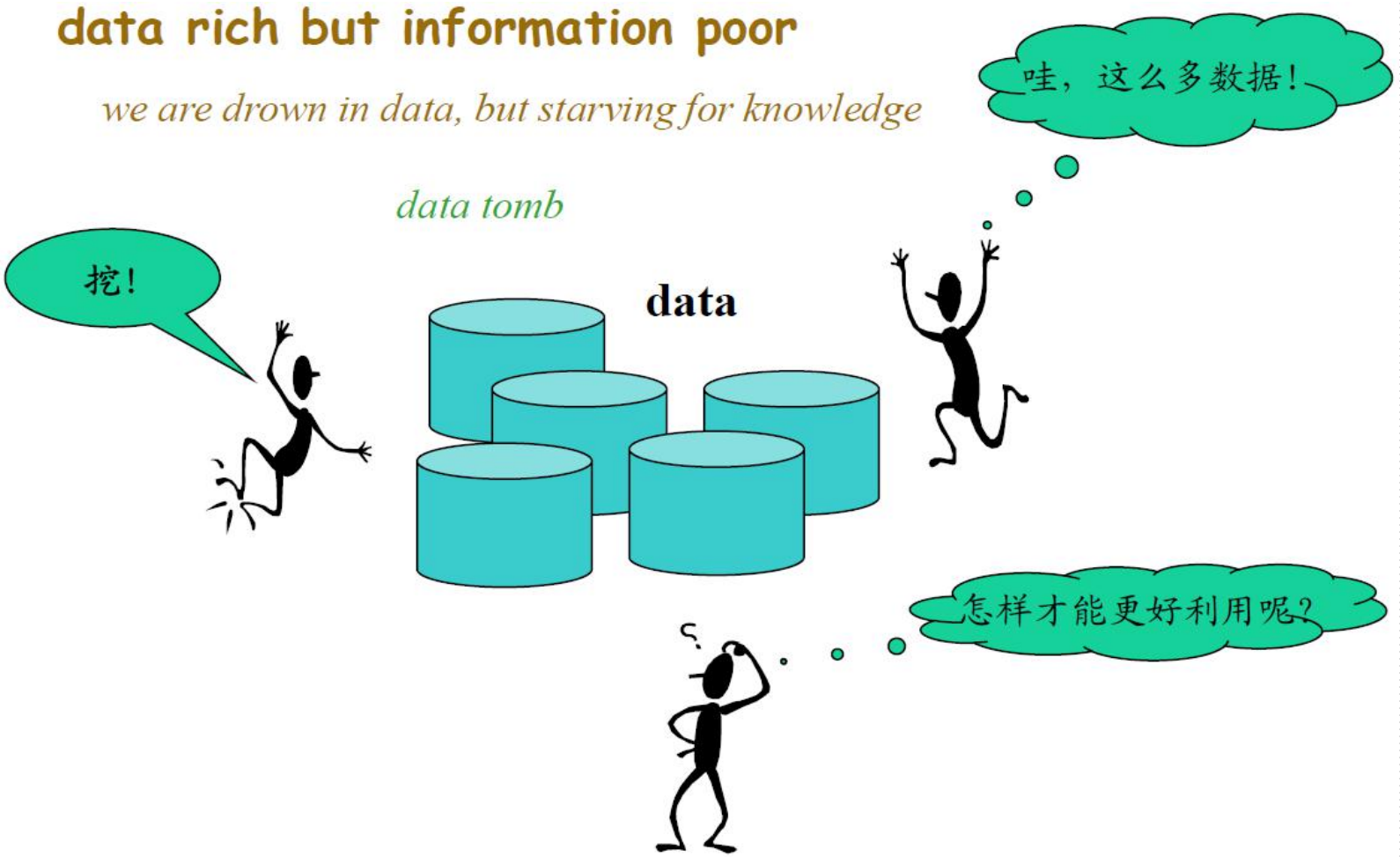
- 知识发现的对象：
  - 1.数据库：当前研究比较多的是关系数据库的知识发现。
  - 2.数据仓库：数据挖掘为数据仓库提供深层次数据分析的手段，数据仓库为数据挖掘提供经过良好预处理的数据源。
  - 3. Web信息：Web知识发现主要分内容发现和结构发现。内容发现是指从Web文档的内容中提取知识;结构发现是指从Web文档的结构信息中推导知识。
  - 4. 图像和视频数据：图像和视频数据中也存在有用的信息。比如，地球资源卫星每天都要拍摄大量的图像或录像。

数据挖掘

为什么要数据挖掘？

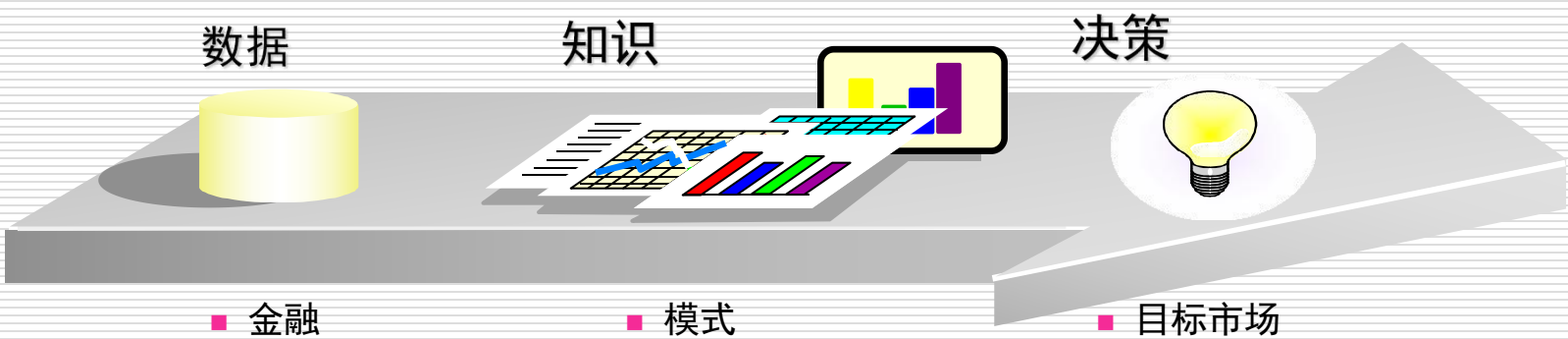
- 数据的爆炸性增长：从TB到PB
  - 数据的收集和数据的可获得性
    - 自动数据收集工具、数据库系统、WEB、计算机化的社会
  - 丰富数据的来源
    - 商业：WEB、电子商务、交易数据、股市...
    - 科学：遥感、生物信息学、科学模拟
    - 社会及每个人：新闻、数码相机、YouTube
- 我们被数据所淹没，但却渴望知识
- “需要是发明之母”，数据挖掘：海量数据的自动分析技术

为什么要数据挖掘？


data rich but information poor
we are drown in data, but starving for knowledge
data tomb
挖!
data
哇，这么多数据!
怎样才能更好利用呢?

数据挖掘的社会需求

苦恼: 淹没在数据中；不能制定合适的决策!



- 金融
- 经济
- 政府
- POS.
- 人口统计
- 生命周期

- 模式
- 趋势
- 事实
- 关系
- 模型
- 关联规则
- 序列

- 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

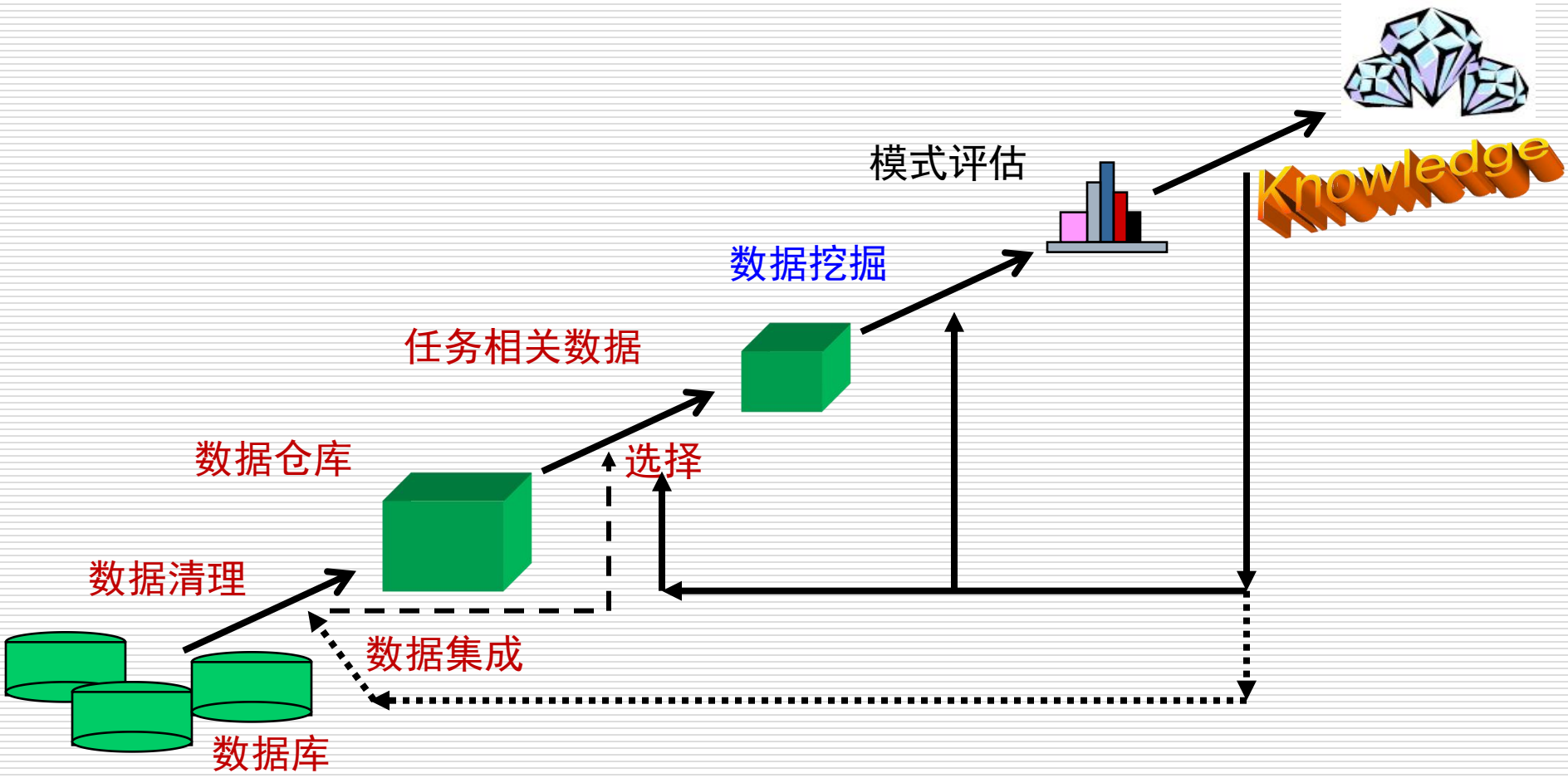
数据爆炸，知识贫乏

什么是数据挖掘？

- 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

数据挖掘：数据库中的知识挖掘（KDD）

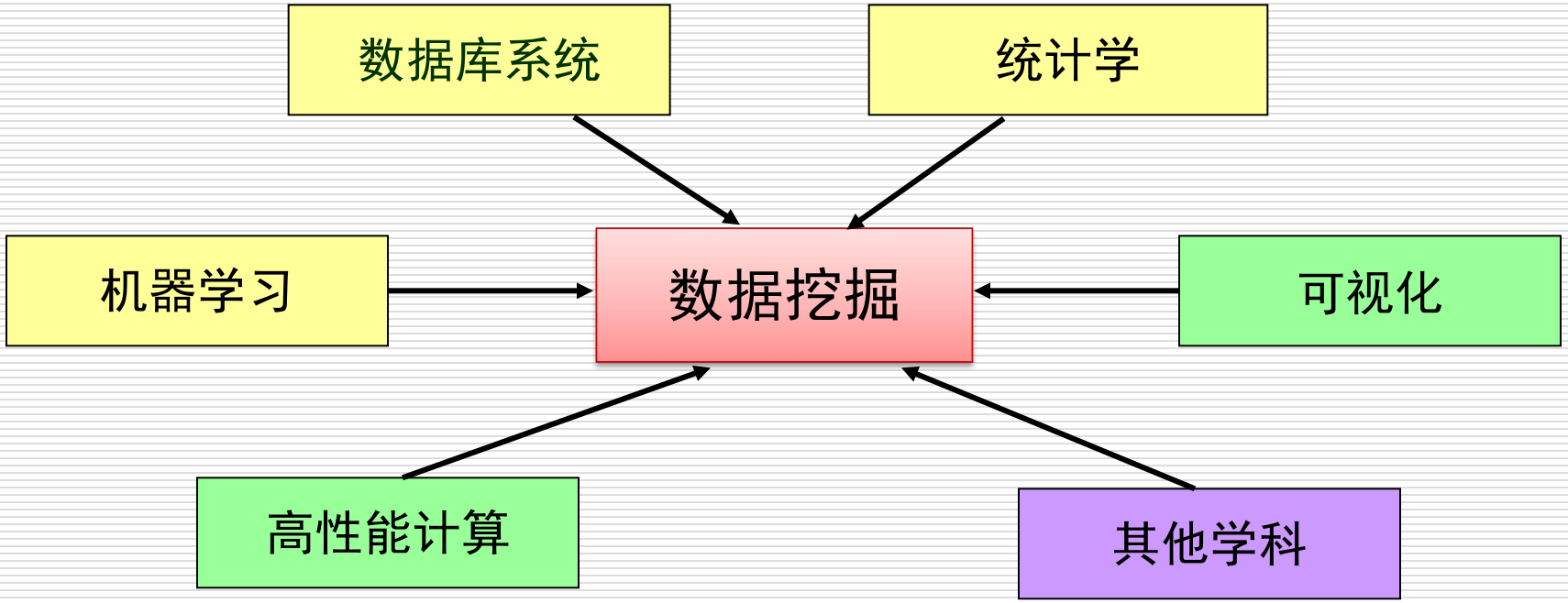
数据挖掘：知识挖掘的核心



为什么不是传统的数据分析？

- 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：多个学科的融合



数据挖掘的主要功能：可以挖掘什么类型的模式？

- 一般功能
  - 描述性的数据挖掘
  - 预测性的数据挖掘
- 通常并不知道在数据中能挖掘出什么，对此会在数据挖掘中应用一些常用的挖掘功能，挖掘出一些常用的模式，包括：
  - 概念/类描述: 特性化和区分
  - 关联分析
  - 分类和预测
  - 聚类分析
  - 孤立点分析
  - 趋势和演变分析

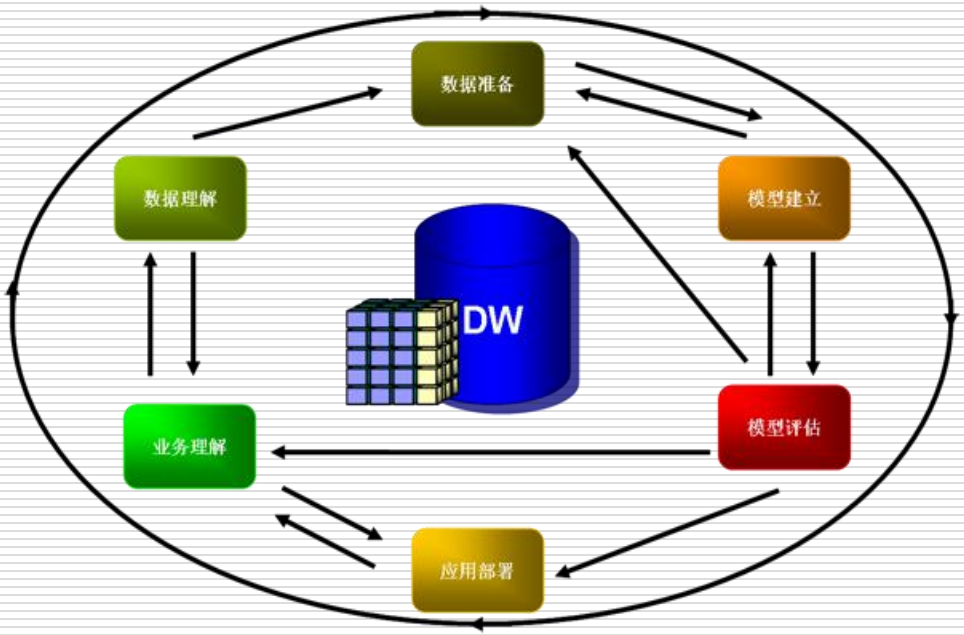
数据挖掘技术

- 技术分类
  - 预言（Predication）：用历史预测未来
  - 描述（Description）：了解数据中潜在的规律
- 数据挖掘技术
  - 关联分析
  - 序列模式
  - 分类（预言）
  - 聚集
  - 异常检测

数据挖掘标准流程

CRISP－DM

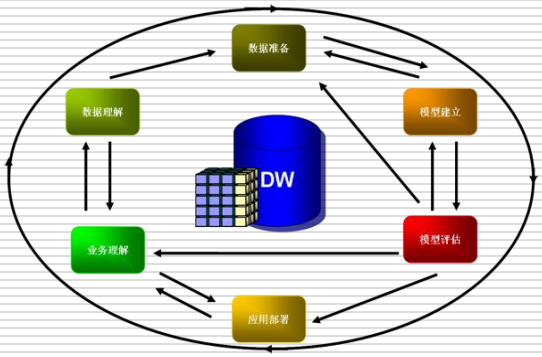
- CRISP－DM是CRoss-Industry Standard Process －Data Mining的缩写



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

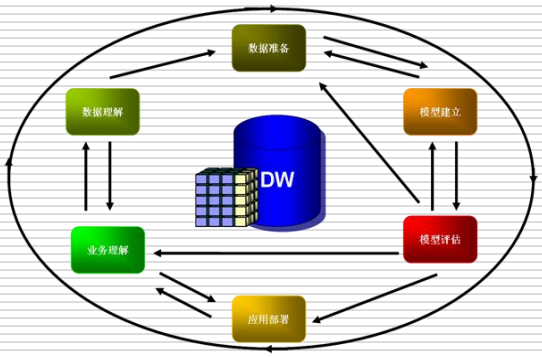
商业理解（Business Understanding)

- 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



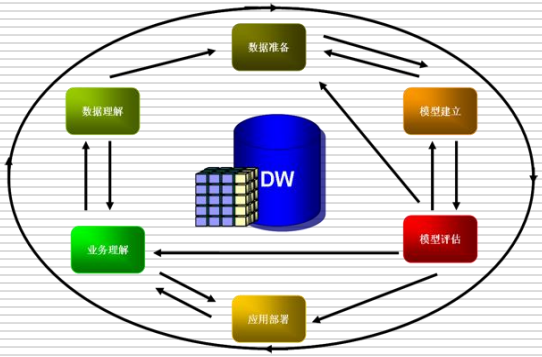
数据理解（Data Understanding）

- 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



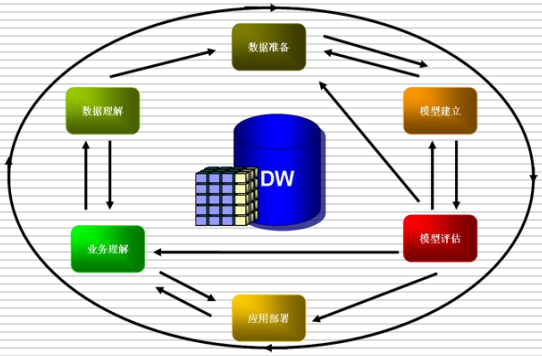
数据准备（Data Preparation）

- 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



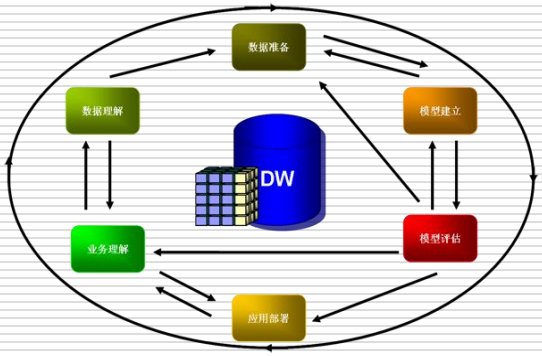
建立模型（Modeling)

- 对各个模型进行评价
- 选择数据挖掘模型
- 建立模型



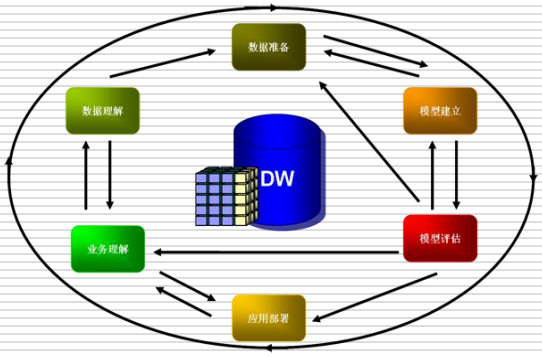
模型评估（Evaluation）

- 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

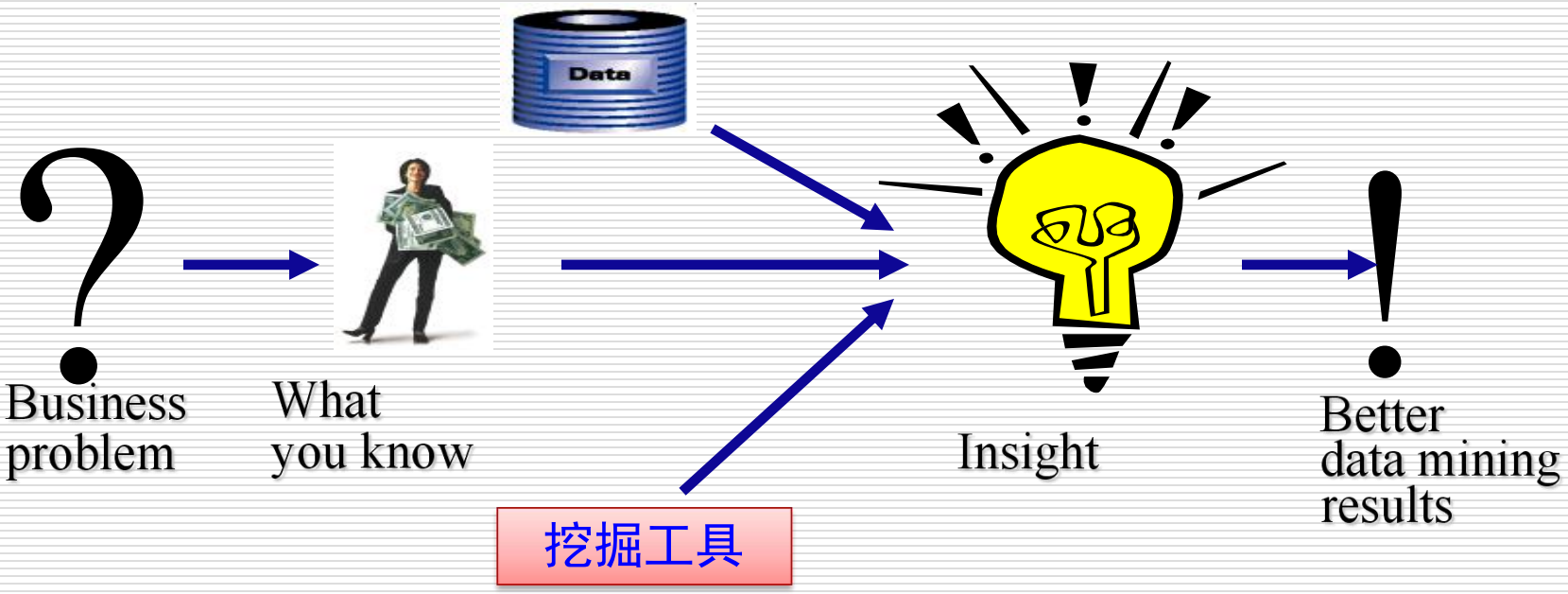


模型发布（Deployment）

- 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



- 把业务经验溶入数据挖掘过程是数据挖掘成功的关键



THE END