

# 机器学习

——宋明丽

——计算机与网络空间安全学院



# 关于这门课

---

- 教材：机器学习，周志华，2016
- 参考书：机器学习，Mitchell,2008,等
- 成绩：平时40+期末60%
- 我：songmingli@cuc.edu.cn,48教A906
- 注：特殊时期，在家学习可以不买教材！我会提供课件与学习资料。



# 大纲

---

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 研究方向
- 阅读材料



# 机器学习

“假设用 $P$ 来评估计算机程序在某任务类 $T$ 上的性能，若一个程序通过利用经验 $E$ 在 $T$ 中任务上获得了性能改善，则我们就说关于 $T$ 和 $P$ ，该程序对 $E$ 进行了学习”

机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能，从而在计算机上从数据中产生“模型”，用于对新的情况给出判断。

问题：什么是机器学习？



# 机器学习与数据挖掘

---

数据分析技术

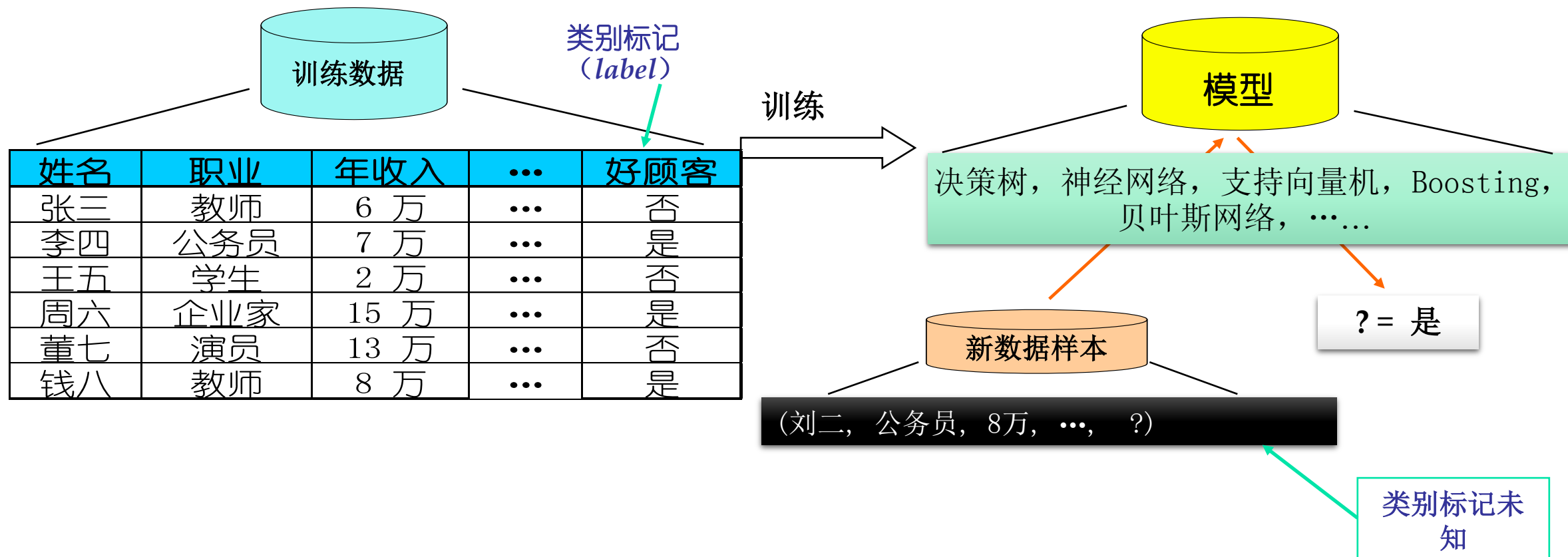


数据挖掘

机器学习

# 典型的机器学习过程

使用学习算法 (*learning algorithm*)



# 基本术语-数据

特征:feature or attribute

标记:label

训练集:training data

测试集:testing data

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否
1	青绿	蜷缩	沉闷	?



# 基本术语

---

- 属性空间: **attribute space**
- 属性值: **attribute value**
- 样本: **sample or instance**
- 维数: **dimensionality**
- 预测: **prediction**
- 学习器: **learner**
- 假设: **hypothesis**





# 基本术语-任务

---

- 预测目标:
  - 分类(classification): 离散值
    - 二分类: 好瓜; 坏瓜
    - 多分类: 冬瓜; 南瓜; 西瓜
  - 回归(regression): 连续值
    - 瓜的成熟度
  - 聚类(clustering): 无标记信息

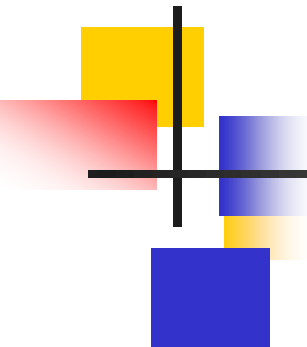


# 基本术语-任务

---

- 有无标记信息

- 监督学习 (supervised learning): 分类、回归
- 无监督学习 (unsupervised learning): 聚类
- 半监督学习 (semi-supervised learning): 两者结合



# 基本术语-泛化能力

机器学习的目标是使得学到的模型能很好的适用于“新样本”，而不仅仅是训练集合，我们称模型适用于新样本的能力为泛化 (generalization) 能力。

通常假设样本空间中的样本服从一个未知分布  $\mathcal{D}$ ，样本从这个分布中独立获得，即“独立同分布” (i.i.d)。一般而言训练样本越多越有可能通过学习获得强泛化能力的模型

# 假设空间

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

$(\text{色泽}=\text{?}) \wedge (\text{根蒂}=\text{?}) \wedge (\text{敲声}=\text{?}) \leftrightarrow \text{好瓜}$

在模型空间中搜索不违背训练集的假设

假设空间大小:  $3*3*4+1=37$



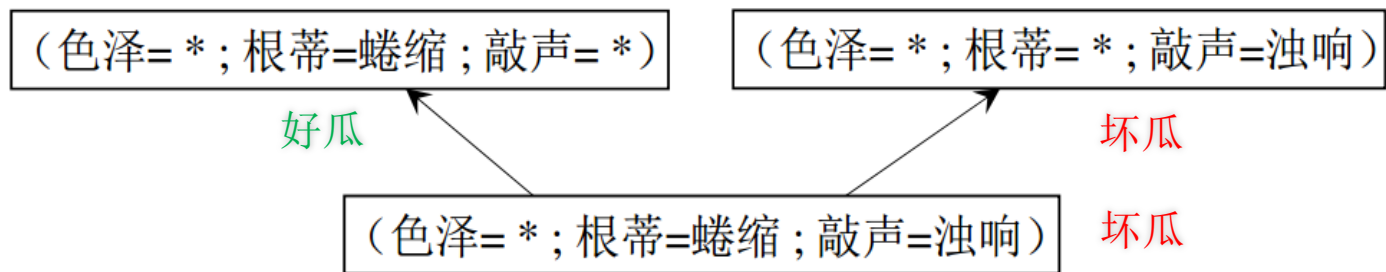
# 假设空间

---

- 归纳: **induction**
  - 从特殊到一般的泛化过程
  - 样例学习
- 演绎: **deduction**
  - 从一般到特殊的特化过程
- 各自可能出现的问题?

# 归纳偏好(inductive bias)

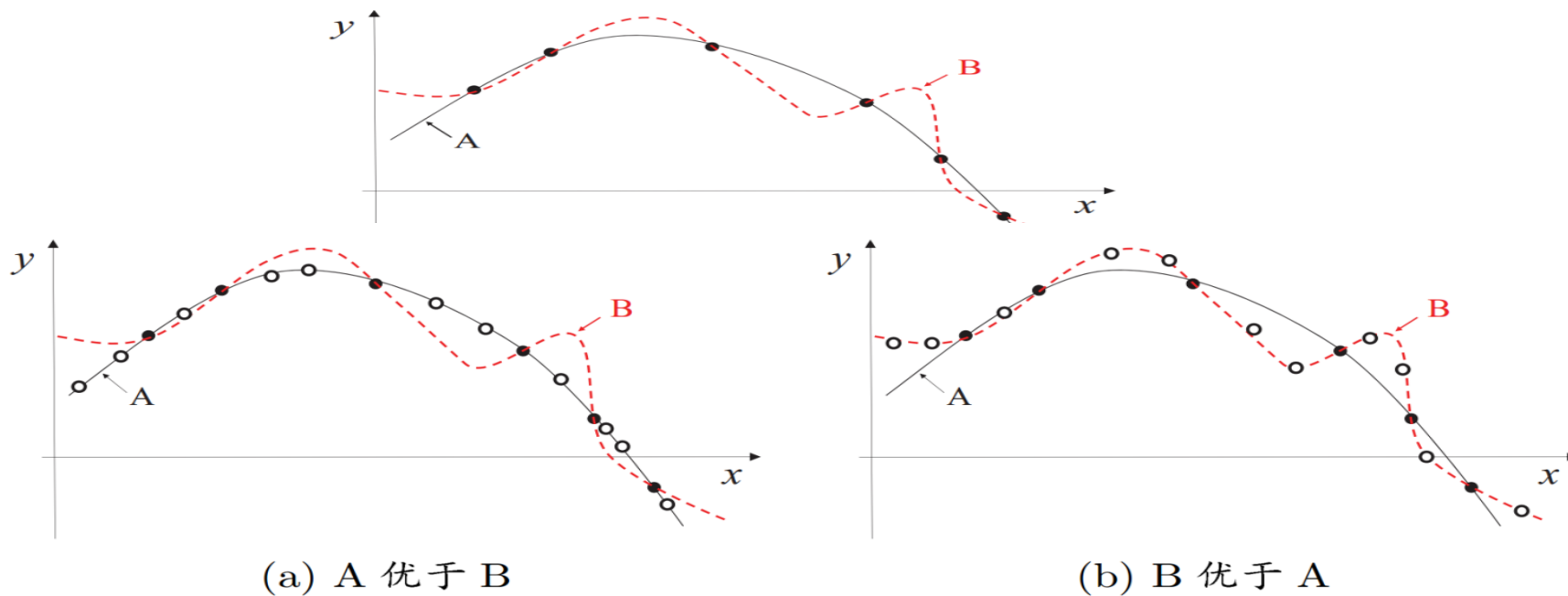
假设空间中有三个与训练集一致的假设，但他们对(色泽=青绿; 根蒂=蜷缩; 敲声=沉闷)的瓜会预测出不同的结果:



选取哪个假设作为学习模型?

# 归纳偏好

学习过程中对某种类型假设的偏好称作归纳偏好



没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)



# 归纳偏好

---

归纳偏好可看作学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或“价值观”。

“奥卡姆剃刀”(Occam)是一种常用的、自然科学研究中最基本的原则，即“若有多个假设与观察一致，选最简单的那个”。

具体的现实问题中，学习算法本身所做的假设是否成立，也即算法的归纳偏好是否与问题本身匹配，大多数时候直接决定了算法能否取得好的性能。





# NoFreeLunch

一个算法  $\xi_a$  如果在某些问题上比另一个算法  $\xi_b$  好, 必然存在另一些问题,  $\xi_b$  比  $\xi_a$  好, 也即没有免费的午餐定理。

简单起见, 假设样本空间  $\mathcal{X}$  和假设空间  $\mathcal{H}$  离散, 令  $P(h|X, \mathfrak{L}_a)$  代表算法  $\mathfrak{L}_a$  基于训练数据  $X$  产生假设  $h$  的概率, 令  $f$  代表要学的目标函数,  $\mathfrak{L}_a$  在训练集之外所有样本上的总误差为

$$E_{ote}(\mathfrak{L}_a|X, f) = \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathfrak{L}_a)$$

$\mathbb{I}(\cdot)$  为指示函数, 若  $\cdot$  为真取值1, 否则取值0



# NoFreeLunch

考虑二分类问题，目标函数可以为任何函数  $\mathcal{X} \mapsto \{0, 1\}$  函数空间为  $\{0, 1\}^{|\mathcal{X}|}$  对所有可能  $f$  按均匀分布对误差求和, 有:

$$\begin{aligned}\sum_f E_{ote}(\mathcal{L}_a | X, f) &= \sum_f \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathcal{L}_a) \\&= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \sum_f \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) \\&= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \\&= 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \cdot 1 .\end{aligned}$$

总误差与学习算法无关!

实际问题中，并非所有问题出现的可能性都相同  
脱离具体问题，空谈“什么学习算法更好”毫无意义



# 发展历程

---

## ● 推理期：

- A. Newell和H. Simon的“逻辑理论家” (Logic Theorist)程序以及伺候的“通用问题求解” (General Problem Solving)程序等在当时取得了令人振奋的结果。
- 2006年卡耐基梅隆大学宣告成立第一个“机器学习系”，机器学习奠基人之一T.Mitchell教授任系主任。

## ● 知识期：

- 大量专家系统问世，在很多应用领域取得大量成果；
- 但是由人来总结知识再交给计算机相当困难。



# 发展历程

---

- 学习期:

- 符号主义学习

- 决策树: 以信息论为基础, 最小化信息熵, 模拟了人类对概念进行判定的树形流程
- 基于逻辑的学习: 使用一节逻辑进行知识表示, 通过修改扩充逻辑表达式对数据进行归纳

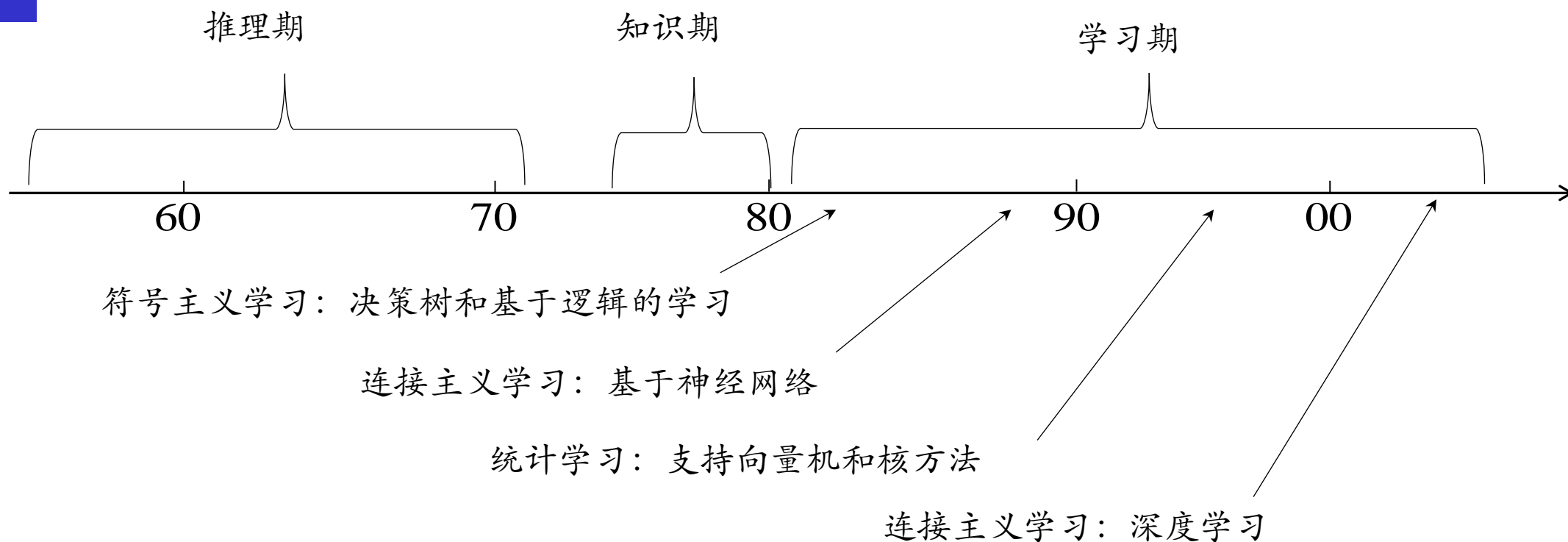
- 连接主义学习

- 神经网络

- 统计学习

- 支持向量机及核方法

# 发展历程





# 应用现状

- 计算机领域最活跃的研究分支之一：
  - NASA\_JPL科学家在Science撰文指出机器学习对科学研究起到越来越大的支撑作用
  - DARPA启动PAL计划，将机器学习的重要性提高到国家安全的高度来考虑
  - 2006年卡耐基梅隆大学宣告成立第一个“机器学习系”，机器学习奠基人之一T.Mitchell教授任系主任。
- 与普通人的生活密切相关：
  - 天气预报、能源勘探、环境监测、搜索引擎、自动驾驶汽车等



# 应用现状

- 影响到人类社会的政治生活：

- 2012美国大选期间奥巴马麾下的机器学习团队，对社交网络等各类数据进行分析，为其提示下一步的竞选行动。

- 具有自然科学探索色彩：

- P.Kanerva在二十世纪八十年代中期提出SDM(Sparse Distributed Memory)模型时并没有刻意模仿脑生理结构，但后来神经科学的研究发现，SDM的稀疏编码机制在视觉、听觉、嗅觉功能的脑皮层中广泛存在，促进理解“人类如何学习”



# 分类 VS. 预测

- 分类

- 预测分类的类标号（离散的 or 名义上的）
- 分类数据（构建模型）基于训练集与输出值（类标号）在可分类的属性上，用它分类新数据

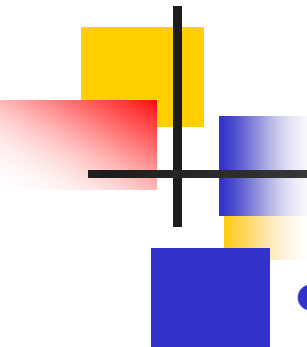
- 预测

- 为连续值函数构建模型，i. e.，预测未知或丢失值

- 典型应用

- 信用证明
- 目标市场
- 医学诊断
- 故障检测

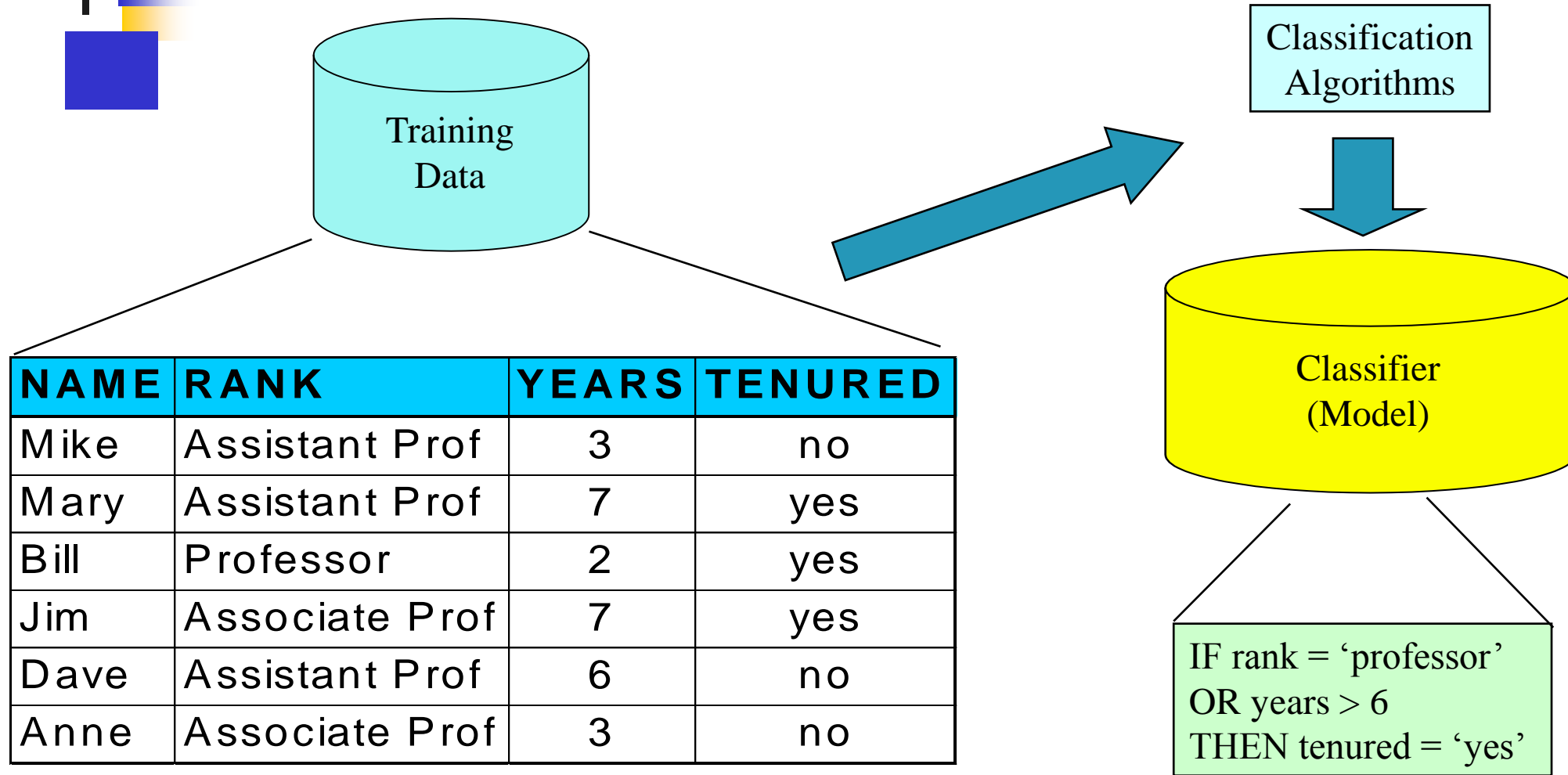




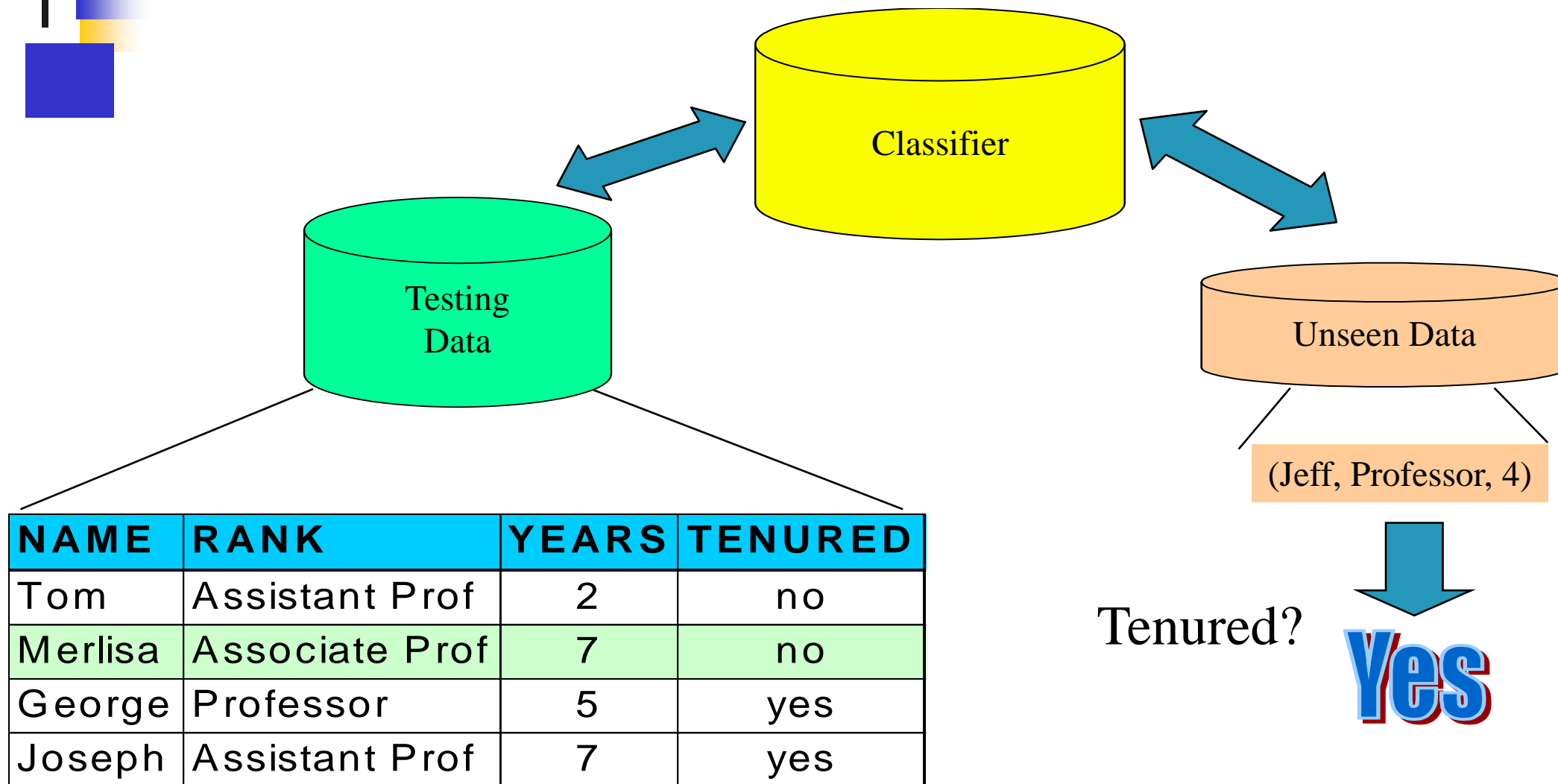
# 分类：两步

- **模型构建**：描述一些已经分好的类
  - 每一个样本都已标好类标号，作为类标号属性
  - 用作建立模型的是训练数据集
  - 模型可以用分类规则，决策树，或其他数学方程表示
- **模型使用**：用来分类后续对象或未知对象
  - 估计模型的准确率
    - ◆ 测试样本的类标号与模型分类的类标号比较
    - ◆ 准确率是测试样本中正确分类的比率
    - ◆ 测试样本是独立的，否则会发生过拟合
  - 如果准确率可以接受，则模型科用来预测未知样本

# 分类过程 (1): 建立模型



## 分类过程 (2)：预测中使用





# 有监督学习 VS. 无监督学习

---

- 有监督学习 (classification)

- 监督: 训练数据是有类标号的
- 新数据被分类基于训练数据集

- 无监督学习 (clustering)

- 训练数据集的类标号未知
- 给定的是测量值, 观测值等, 目的是建立一些类或簇



# 阅读材料

---

- [Mitchell, 1997]是第一本机器学习专门教材. [Duda et al., 2001; Alpaydin, 2004; Flach, 2012]为出色的入门读物. [Hastie et al., 2009]为进阶读物, [Bishop, 2006]适合于贝叶斯学习偏好者. [Shalev-Shwartz and Ben-David, 2014]适合于理论偏好者.
- 《机器学习:一种人工智能途径》 [Michalski et al., 1983]汇集了20位学者撰写16篇文章, 是机器学习早期最重要的文献. [Dietterich, 1997] 对机器学习领域的发展进行了评述和展望。



# 阅读材料

- 机器学习领域最重要的国际学术会议是国际机器学习会议(ICML)、国际神经信息处理系统会议(NIPS)和国际学习理论会议(COLT), 重要的区域性会议主要有欧洲机器学习会议(ECML)和亚洲机器学习会议(ACML); 最重要的国际学术期刊是Journal of Machine Learning Research和Machine Learning.
- 国内不少书籍包含机器学习方面的内容, 例如[陆汝钤, 1996]. [李航, 2012]是一统计学习为主题的读物. 国内机器学习领域最重要的活动是两年一次的中国机器学习大会(CCML)以及每年举行的“机器学习及其应用”研讨会(MLA).



# 作业

---

- 写一篇综述：概述机器学习的过去，现在与未来。中文书写，字数不限，手写。列举参考文献至少10篇。请学习IEEE或其他数据库中论文的书写格式。下周课前交。