

一、基础

主讲教师：周志华

机器学习

智能化是信息科学技术发展的主流趋势，机器学习是实现智能化的关键

经典定义：利用经验改善系统自身的性能 [T. Mitchell 教科书, 1997]



经验 → 数据



随着该领域的发展，目前主要研究智能数据分析的理论和方法，并已成为智能数据分析技术的源泉之一

图灵奖连续授予在该方面取得突出成就的学者



Leslie Valiant
(1949 -)
(Harvard Univ.)

“计算学习理论” 奠基人

2010
年度

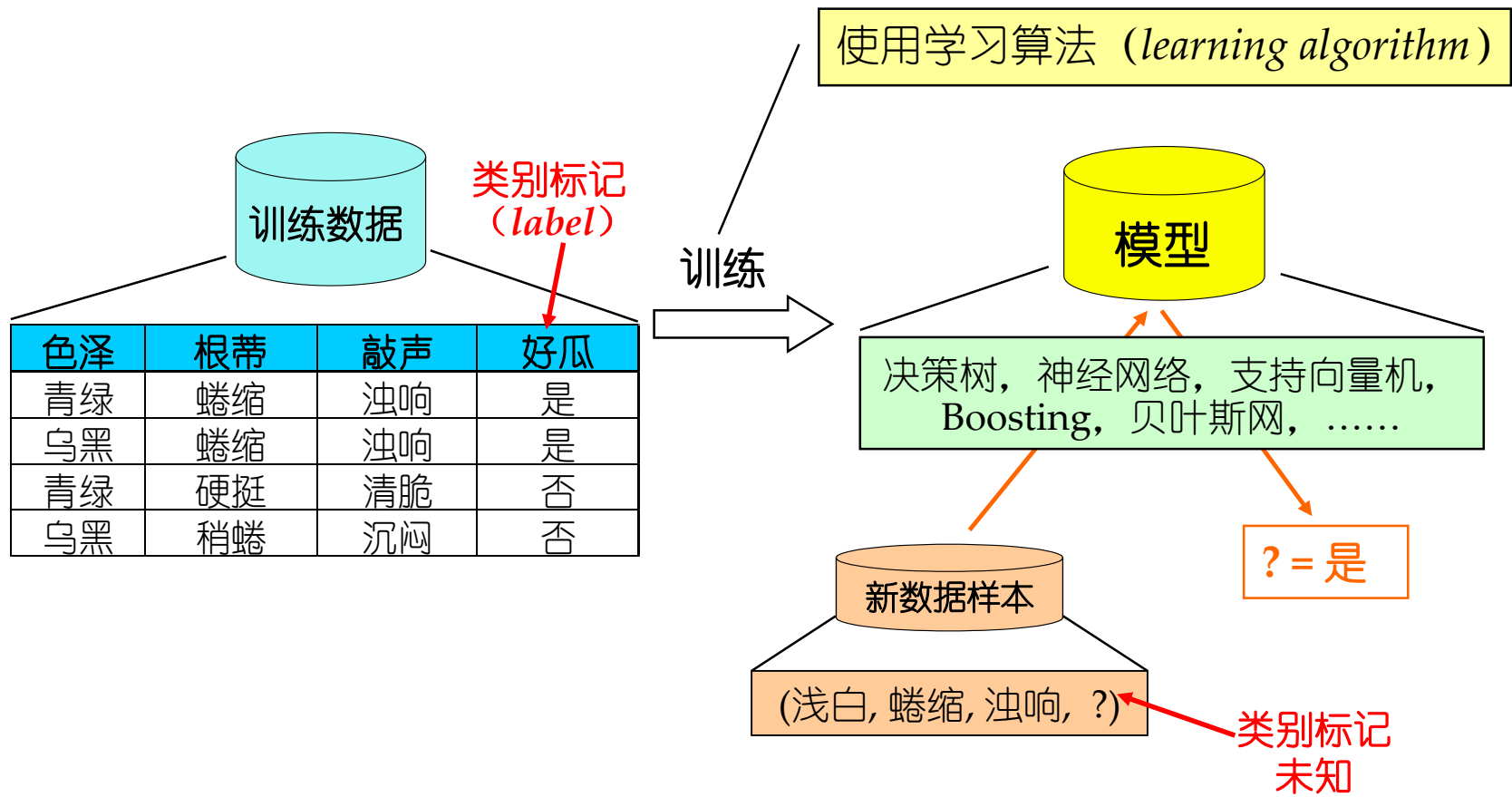


Judea Pearl
(1936 -)
(UCLA)

“图模型学习方法” 先驱

2011
年度

典型的机器学习过程



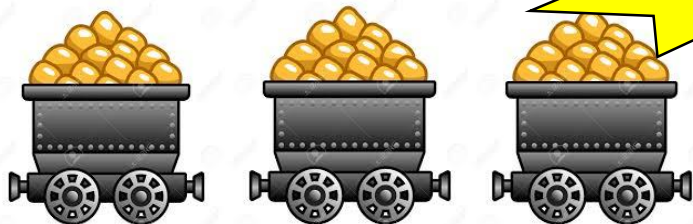
大数据时代



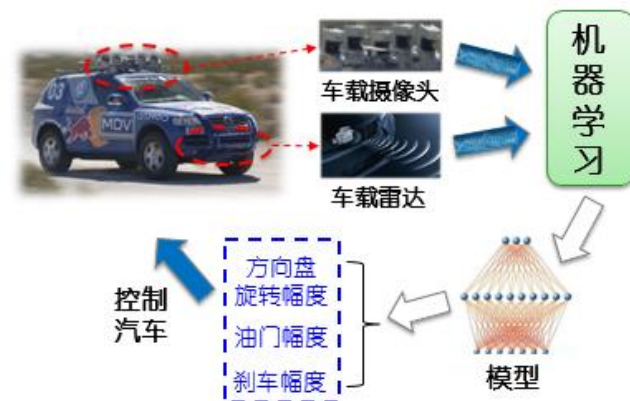
大数据 \neq 大价值

机器学习

有效的数据分析



机器学习已经“无处不在”



机器学习源自“人工智能”

Artificial Intelligence (AI), 1956 -



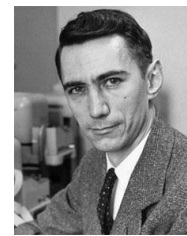
1956年夏 美国达特茅斯学院



J. McCarthy
“人工智能之父”
图灵奖(1971)



M. Minsky
图灵奖(1969)



C. Shannon
“信息论之父”



H. A. Simon
图灵奖(1975)
诺贝尔经济学奖(1978)



A. Newell
图灵奖(1975)

.....
.....

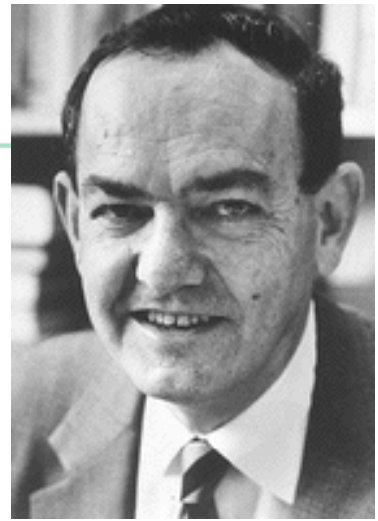
达特茅斯会议标志着人工智能这一学科的诞生

第一阶段：推理期

1956-1960s: Logic Reasoning

- ◆ 出发点：“数学家真聪明！”
- ◆ 主要成就：自动定理证明系统（例如，西蒙与纽厄尔的“Logic Theorist”系统）

渐渐地，研究者们意识到，仅有逻辑推理能力是不够的 …



赫伯特·西蒙
(1916–2001)
1975年图灵奖



阿伦·纽厄尔
(1927–1992)
1975年图灵奖

第二阶段：知识期

1970s -1980s: Knowledge Engineering

- ◆ 出发点：“知识就是力量！”
- ◆ 主要成就：专家系统（例如，费根鲍姆等人的“DENDRAL”系统）



爱德华·费根鲍姆
(1936-)
1994年图灵奖

渐渐地，研究者们发现，要总结出知识再“教”给系统，实在太难了 …

第三阶段：学习期

1990s -now: Machine Learning

- ◆ 出发点：“让系统自己学！”
- ◆ 主要成就：.....

机器学习是作为“突破知识工程瓶颈”
之利器而出现的



恰好在20世纪90年代中后期，人类发现自己淹没在数据的汪洋中，对自动数据分析技术——机器学习的需求日益迫切

今天的“机器学习”已经是一个广袤的学科领域

例如，这是第33届国际机器学习大会 (ICML 2016) 的“主题领域”

2006年，美国CMU (卡内基梅隆大学) 成立“机器学习系”

- | | |
|--|---|
| <input type="checkbox"/> Active Learning | <input type="checkbox"/> Network and Graph Analysis |
| <input type="checkbox"/> Approximate Inference | <input type="checkbox"/> Neural Networks and Deep Learning |
| <input type="checkbox"/> Bayesian Nonparametric Methods | <input type="checkbox"/> Neuroscience |
| <input type="checkbox"/> Bioinformatics | |
| <input type="checkbox"/> Causal Inference | |
| <input type="checkbox"/> Clustering | |
| <input type="checkbox"/> Computational Learning Theory | |
| <input type="checkbox"/> Computational Social Sciences | |
| <input type="checkbox"/> Computer Vision | |
| <input type="checkbox"/> Cost-Sensitive Learning | |
| <input type="checkbox"/> Digital Humanities | |
| <input type="checkbox"/> Economics and Finance | |
| <input type="checkbox"/> Ensemble Methods | |
| <input type="checkbox"/> Feature Selection and Dimensionality Reduction | |
| <input type="checkbox"/> Gaussian Processes | |
| <input type="checkbox"/> Graphical Models | |
| <input type="checkbox"/> Graphs and Social Networks | |
| <input type="checkbox"/> Health Care | |
| <input type="checkbox"/> Inductive Logic Programming and Relational Learning | |
| <input type="checkbox"/> Information Retrieval | |
| <input type="checkbox"/> Information Theory | |
| <input type="checkbox"/> Kernel Methods | |
| <input type="checkbox"/> Large Scale Learning and Big Data | |
| <input type="checkbox"/> Latent Variable Models | |
| <input type="checkbox"/> Learning and Game Theory | |
| <input type="checkbox"/> Learning and Mechanism Design | |
| <input type="checkbox"/> Learning for Games | |
| | <input type="checkbox"/> Other Models and Methods |
| | <input type="checkbox"/> Parallel and Distributed Learning |
| | <input type="checkbox"/> Planning and Control |
| | <input type="checkbox"/> Privacy, Anonymity, and Security |
| | <input type="checkbox"/> Probabilistic Programming |
| | <input type="checkbox"/> Ranking and Preference Learning |
| | <input type="checkbox"/> Recommender Systems |
| | <input type="checkbox"/> Reinforcement Learning |
| | <input type="checkbox"/> Representation Learning |
| | <input type="checkbox"/> Resource Efficient Learning |
| | <input type="checkbox"/> Robotics |
| | <input type="checkbox"/> Rule and Decision Tree Learning |
| | <input type="checkbox"/> Semi-Supervised Learning |
| | <input type="checkbox"/> Sparsity and Compressed Sensing |
| | <input type="checkbox"/> Spectral Methods |
| | <input type="checkbox"/> Speech Recognition |
| | <input type="checkbox"/> Statistical Learning Theory |
| | <input type="checkbox"/> Statistical Relational Learning |
| | <input type="checkbox"/> Structured Prediction |
| | <input type="checkbox"/> Supervised Learning |

经常被谈到的“深度学习” (Deep Learning) 仅是机器学习中的一个小分支

机器学习很强大，但是.....

并非“一切皆可学”

◆ 特征信息不充分

- 例如，重要特征信息没有获得

◆ 样本信息不充分

- 例如，仅有很少的数据样本

机器学习有坚实的理论基础

计算学习理论

Computational learning theory

最重要的理论模型：

PAC (Probably Approximately Correct,

概率近似正确) learning model [Valiant, 1984]

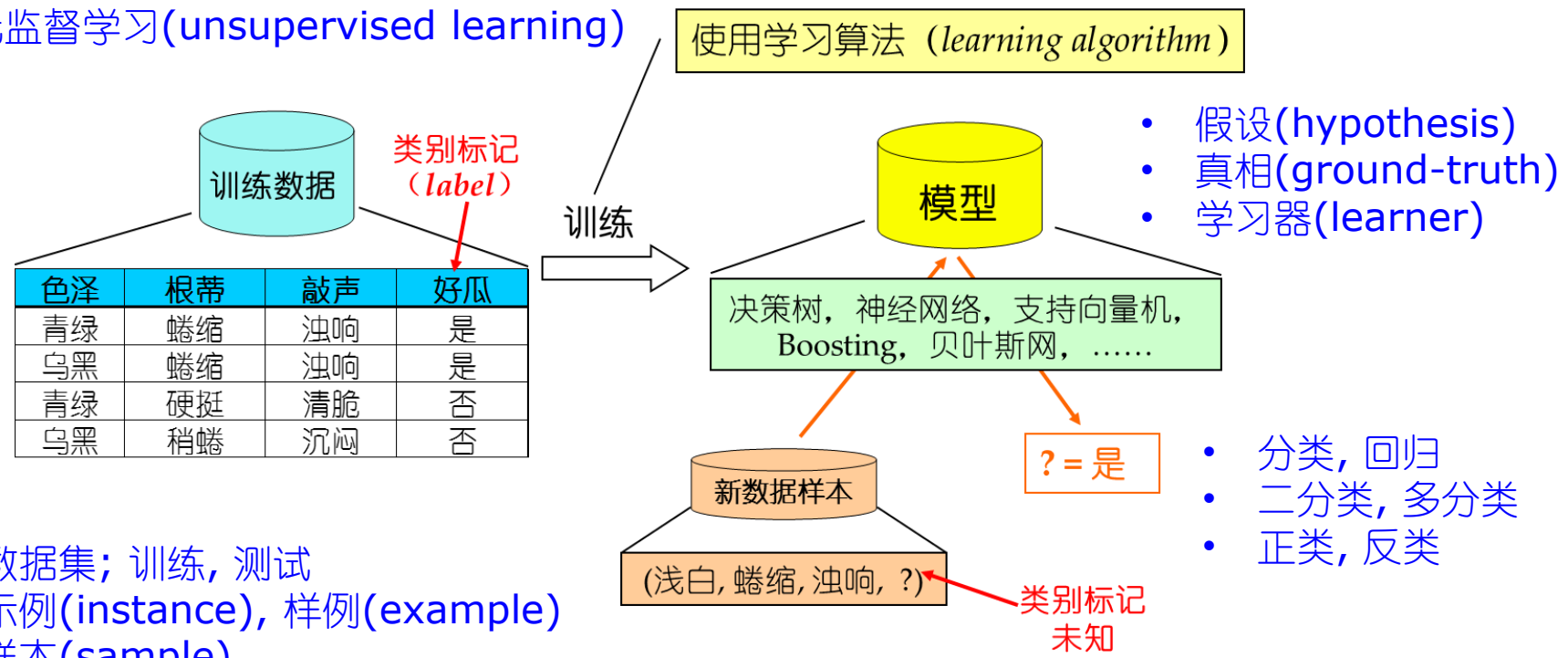
$$P(|f(\mathbf{x}) - y| \leq \epsilon) \geq 1 - \delta$$



Leslie Valiant
(莱斯利·维利昂特)
(1949–)
2010年图灵奖

基本术语

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)

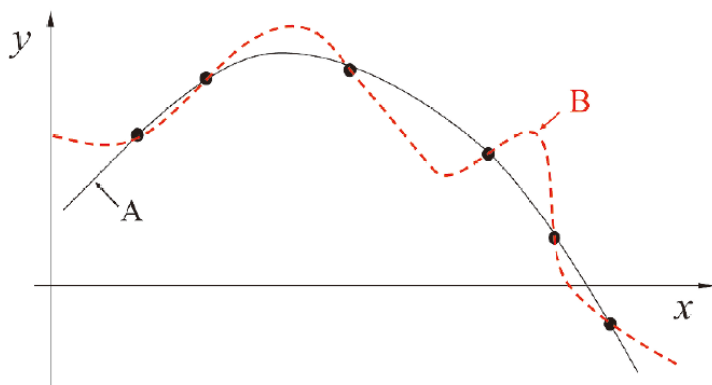


- 数据集; 训练, 测试
- 示例(instance), 样例(example)
- 样本(sample)
- 属性(attribute), 特征(feature); 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间

- 未见样本(unseen instance)
- 未知“分布”
- 独立同分布(i.i.d.)
- 泛化(generalization)

归纳偏好 (inductive bias)

机器学习算法在学习过程中对某种类型假设的偏好



A更好？

B更好？

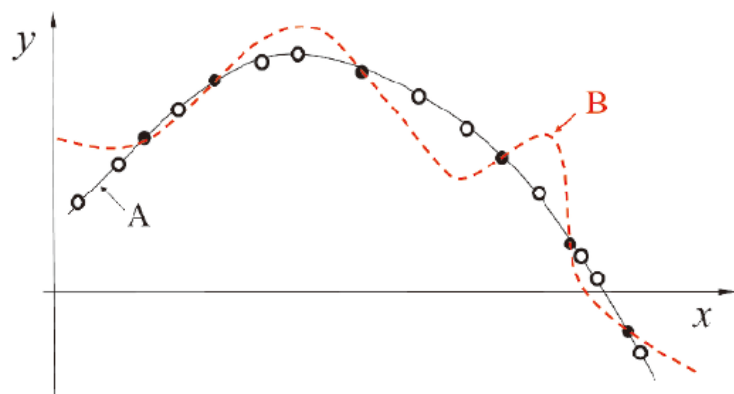
一般原则：
奥卡姆剃刀
(Ocam's razor)

任何一个有效的机器学习算法必有其偏好

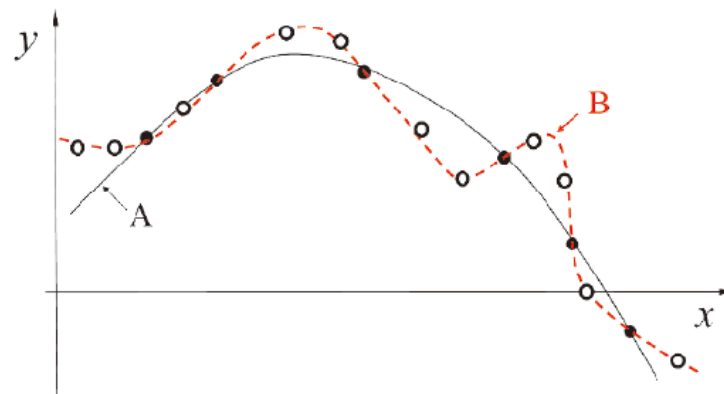
学习算法的归纳偏好是否与问题本身匹配，
大多数时候直接决定了算法能否取得好的性能！

哪个算法更好？

没有免费的午餐！



(a) A 优于 B



(b) B 优于 A

图 1.4 没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

NFL定理：一个算法 \mathcal{L}_a 若在某些问题上比另一个算法 \mathcal{L}_b 好，必存在另一些问题， \mathcal{L}_b 比 \mathcal{L}_a 好。

NFL定理的寓意

NFL定理的重要前提：

所有“问题”出现的机会相同、或所有问题同等重要

实际情形并非如此；我们通常只关注自己正在试图解决的问题

脱离具体问题，空泛地谈论“什么学习算法更好”
毫无意义！

具体问题，具体分析！

把机器学习的“十八般兵器”都弄熟，
逐个试一遍，是不是就OK了？

NO !

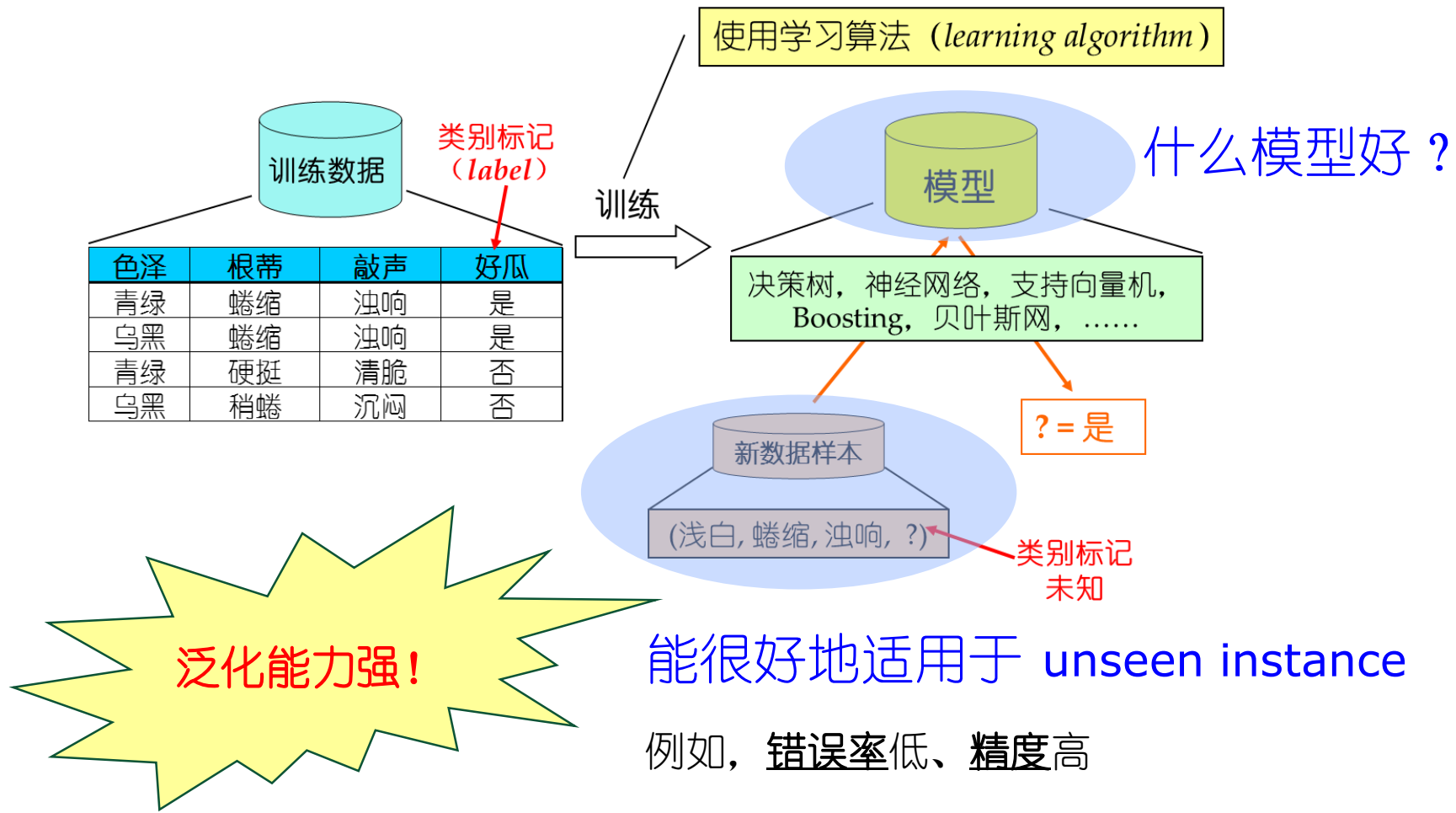
机器学习不是“十八般兵器”的堆积

在现实任务中，很少能“照搬”兵器取得好结果

按需设计、度身定制

模型评估与选择

典型的机器学习过程



然而, 我们手上没有 unseen instance,

泛化误差 vs. 经验误差

泛化误差：在“未来”样本上的误差

经验误差：在训练集上的误差，亦称“训练误差”

- 泛化误差越小越好
- 经验误差是否越小越好？

NO! 因为会出现“过拟合” (overfitting)

过拟合 (overfitting) VS. 欠拟合 (underfitting)

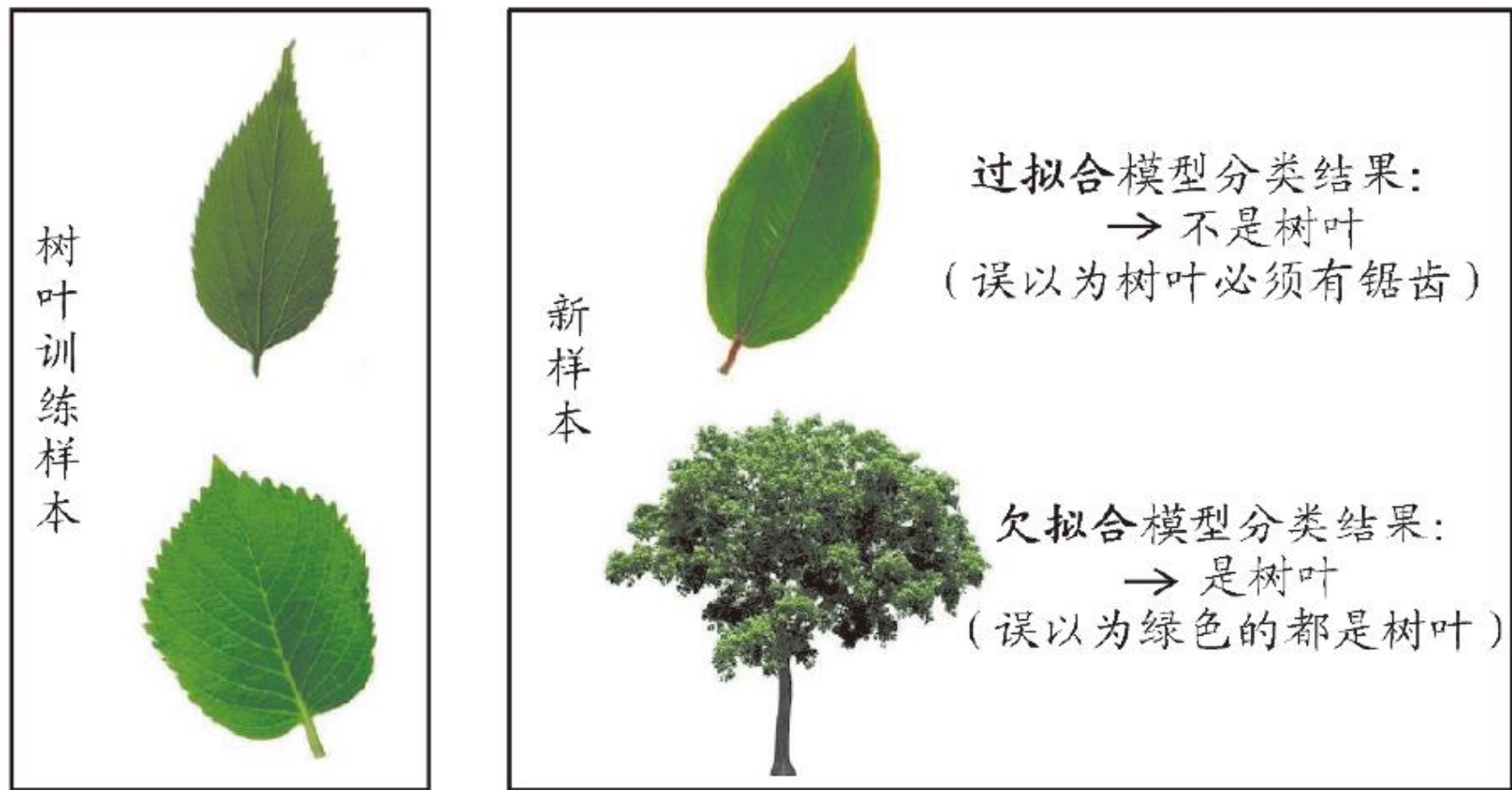
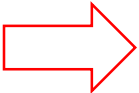
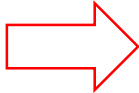
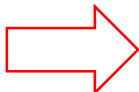


图 2.1 过拟合、欠拟合的直观类比

模型选择 (model selection)

三个关键问题：

- 如何获得测试结果？  评估方法
- 如何评估性能优劣？  性能度量
- 如何判断实质差别？  比较检验

“误差”包含了哪些因素？

换言之，从机器学习的角度看，

“误差”从何而来？

偏差-方差分解 (bias-variance decomposition)

对回归任务，泛化误差可通过“偏差-方差分解”拆解为：

$$E(f; D) = \underbrace{bias^2(\mathbf{x})}_{\text{red}} + \underbrace{var(\mathbf{x})}_{\text{blue}} + \underbrace{\varepsilon^2}_{\text{green}}$$

期望输出与真实输出的差别

$$bias^2(\mathbf{x}) = (\bar{f}(\mathbf{x}) - y)^2$$

同样大小的训练集的变动，所导致性能变化

$$var(\mathbf{x}) = \mathbb{E}_D \left[(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \right]$$

训练样本的标记与真实标记有区别

表达了当前任务上任何学习算法所能达到的期望泛化误差下界

$$\varepsilon^2 = \mathbb{E}_D \left[(y_D - y)^2 \right]$$

泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度共同决定

前往下一站.....

