模式识别与机器学习

01 绪论
（第1部分）

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

目录

01

PART ONE

引言

Introduction

课程简介

❑ Instructor: Qian Liu (刘倩)

(创新园大厦A0822, Email: qianliu@dlut.edu.cn)

- 助教

  蔺虎虎、穆胜达

- 学习基础
  - ✓ 矩阵与数值分析（线性代数），概率与统计，最优化理论
- 交叉课程
  - ✓ 图像处理，计算机视觉，数据挖掘，自然语言处理，多媒体技术
- 参考书目
  - ✓ 《机器学习》，周志华等，清华大学出版社
  - ✓ 《统计学习方法》，李航等，清华大学出版社
  - ✓ 《模式分类（第2版）》(Pattern Classification)，[美]迪达等著；
  - ✓ Pattern Recognition and Machine Learning》, Christopher Bishop, Springer
  - ✓ 《深度学习》，Ian Goodfellow等，人民邮电出版社

■主要授课内容

- ➢模式识别与机器学习基本概念
- ➢传统模式识别与机器学习算法

  监督学习，无监督学习，半监督学习

  线性回归，逻辑斯蒂回归，Boosting，K-Means, Naive Bayes，

  支持向量机，...
- ➢稀疏表示与低秩矩阵，神经网络与深度学习，强化学习
- ➢机器学习前沿：迁移学习，对抗学习...

考核方法

■笔试（40%）

- 闭卷
- 考试时间：待定

■小作业（20%）

- 课程过程中会留4次课后作业
- 提交时间：每周一上课前
- 提交方式：线上——电子版/扫描版；线下——纸质版

■ 大作业（40%）

- **Poster**：具体要求另行通知，在15周-18周之间组织Poster大赛
- **Project**（编程大作业）

■期刊

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- IEEE Transactions on Image Processing (TIP)
- Pattern Recognition (PR)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)

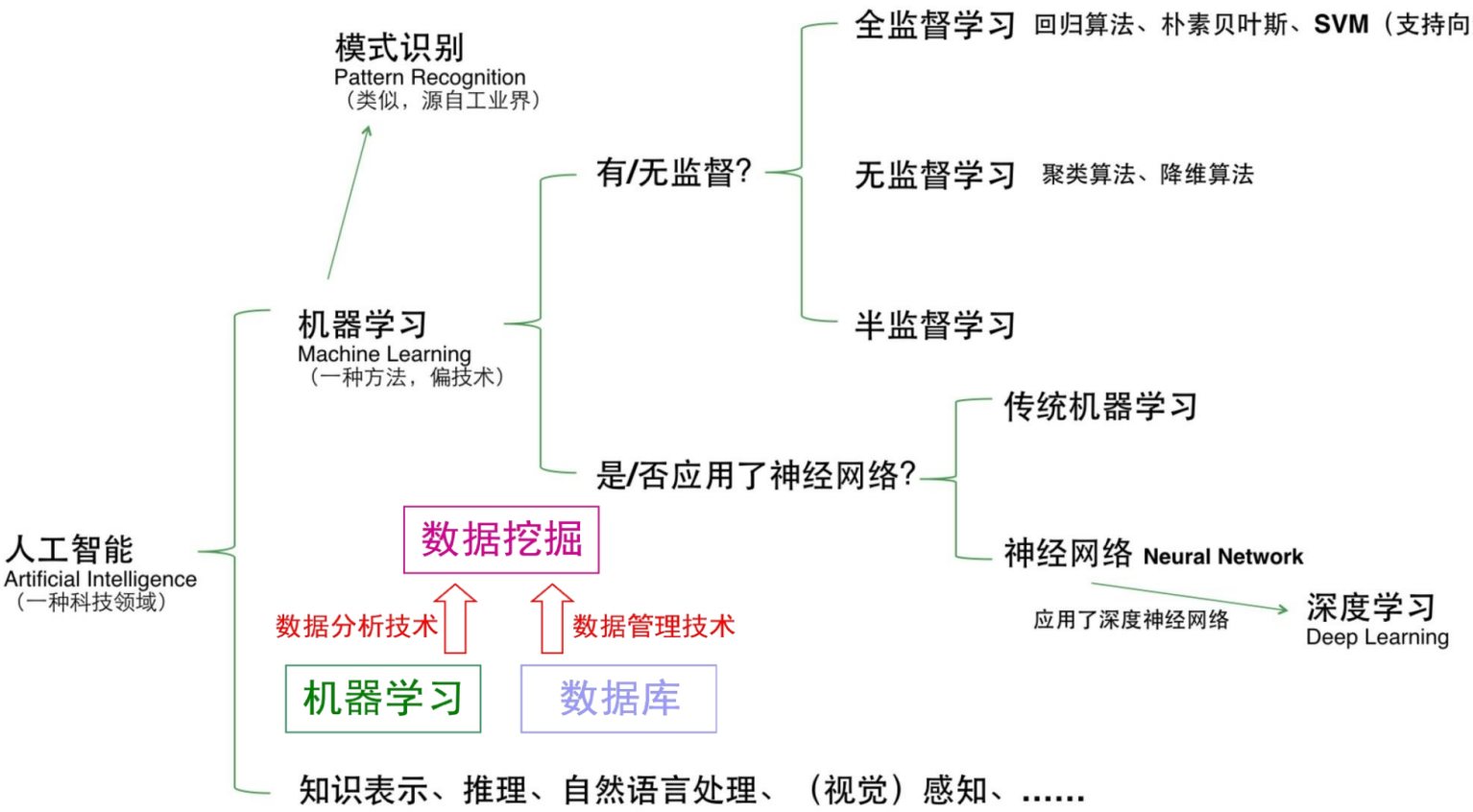
■ 会议

- ICML (International Conference on Machine Learning)
- NeurIPS (Neural Information Processing Systems)
- CVPR (IEEE Conference on Computer Vision and Pattern Recognition)
- ICLR (International Conference on Learning Representations)
- ICCV (International Conference on Computer Vision)

….

- Arxiv: https://zh.wikipedia.org/wiki/ArXivarXiv

模式识别vs机器学习



➢**模式识别vs机器学习**。两者的主要区别在于前者是从工业界发展起来的概念，后者则主要源自计算机学科。在著名的《Pattern Recognition And Machine Learning》这本书中，Christopher M. Bishop在开头是这样说的“模式识别源自工业界，而机器学习来自于计算机学科。不过，它们中的活动可以被视为同一个领域的两个方面，同时在过去的10年间，它们都有了长足的发展”。

模式识别vs机器学习

- “Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field.”

  - C. M. Bishop (author of PRML)

- “模式识别称为70年代，80年代和90年代初的“智能”信号处理是合适的。决策树、启发式和二次判别分析等全部诞生于这个时代。而且，在这个时代，模式识别也成为了计算机科学领域的小伙伴搞的东西，而不是电子工程。从这个时代诞生的模式识别领域最著名的书之一是由Duda & Hart执笔的模式分类（Pattern Classification）”；
- 通俗的说，“机器学习”这个名词比“模式识别”这个名词更时髦一些；
- “模式识别和机器学习的区别在于：前者喂给机器的是各种**特征**描述，从而让机器对未知的事物进行判断；后者**可以**喂给机器的是某一事物的海量样本，让机器通过样本来**自己发现特征**，最后去判断某些未知的事物。”

Reference：https://blog.csdn.net/weixin_39910711/java/article/details/79475729

■ 机器学习

https://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0

机器学习是人工智能的一个分支。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。显然，机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。机器学习在近30多年已发展为一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。

机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为统计学习理论。算法设计方面，机器学习理论关注可以实现的，行之有效的学习算法。很多推论问题属于无程序可循难度，所以部分的机器学习研究是开发容易处理的近似算法。

机器学习已广泛应用于数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA序列测序、语音和手写识别、战略游戏和机器人等领域。

■ 机器学习

- "Learning is any process by which a system improves performance from experience."

——Herbert Simon

- Definition by Tom Mitchell (1998):

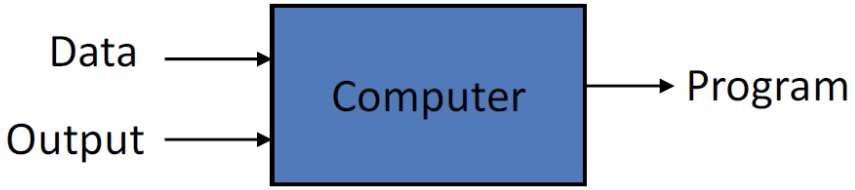
Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E.

A well-defined learning task is given by <P, T, E>.

Traditional Programming

Machine Learning



什么是机器学习

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

经典定义：利用经验改善系统自身的性能

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算法，并已成为智能数据分析技术的源泉之一

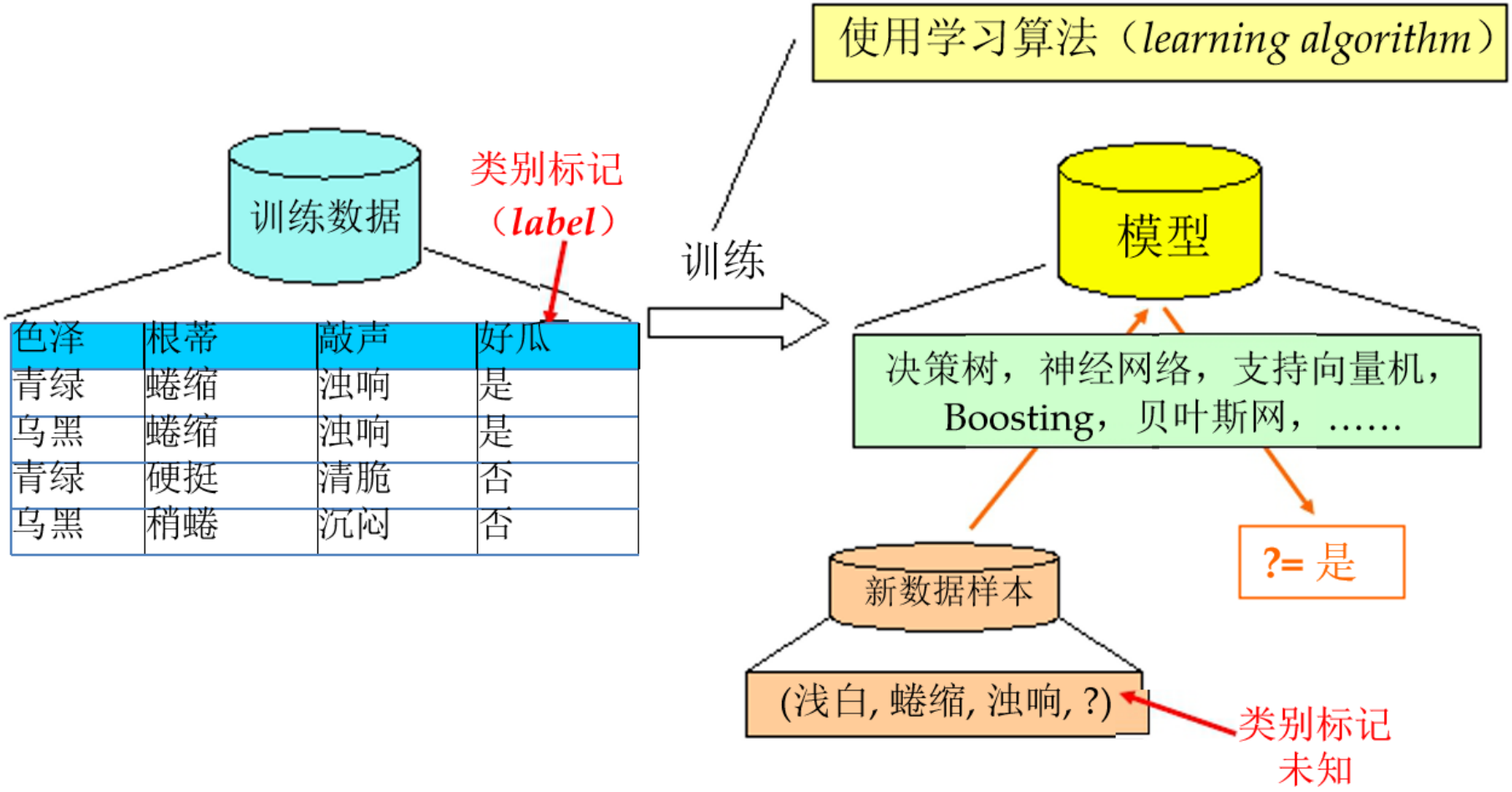
近期机器学习相关ACM图灵奖：

2011, Leslie Valiant, “计算学习理论”

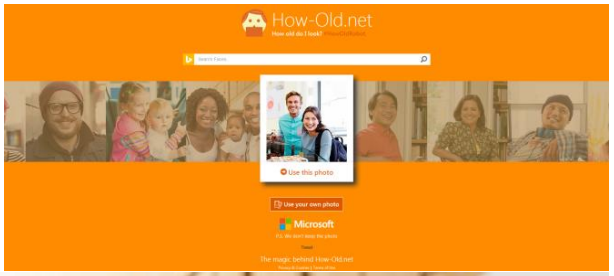
2012, Judea Pearl, “图模型学习方法”

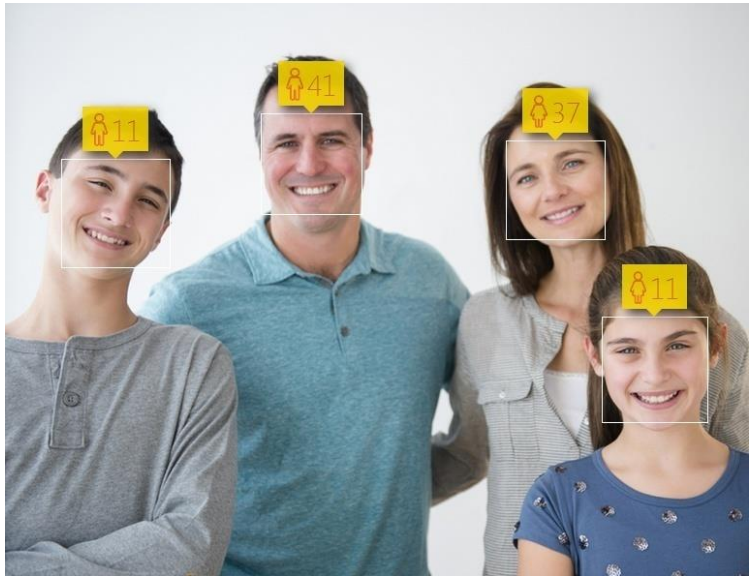
2018, **Geoffrey Hinton, Yoshua Bengio, Yann LeCun**, “神经网络与深度学习”

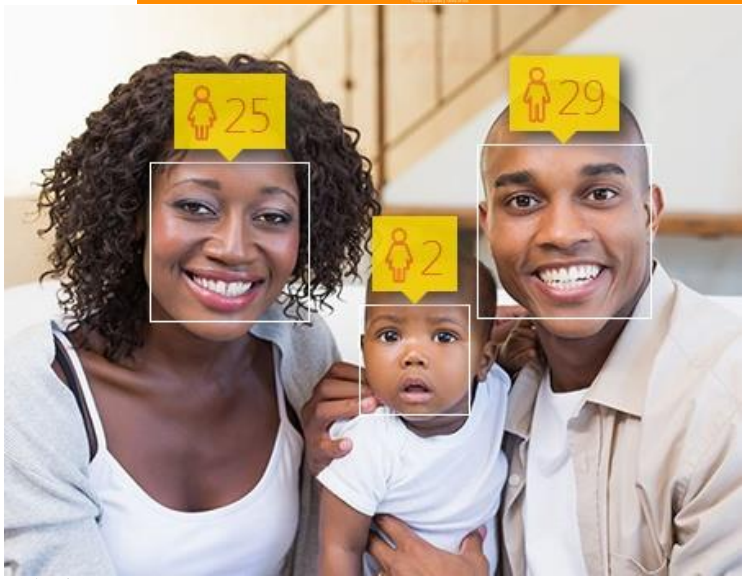
典型的机器学习过程


使用学习算法（learning algorithm）
训练数据
类别标记
（label）
训练
模型
色泽 根蒂 敲声 好瓜
青绿 蜷缩 浊响 是
乌黑 蜷缩 浊响 是
青绿 硬挺 清脆 否
乌黑 稍蜷 沉闷 否
决策树，神经网络，支持向量机，
Boosting，贝叶斯网，……
新数据样本
?= 是
(浅白, 蜷缩, 浊响, ?)
类别标记
未知

➢**how-old.net**







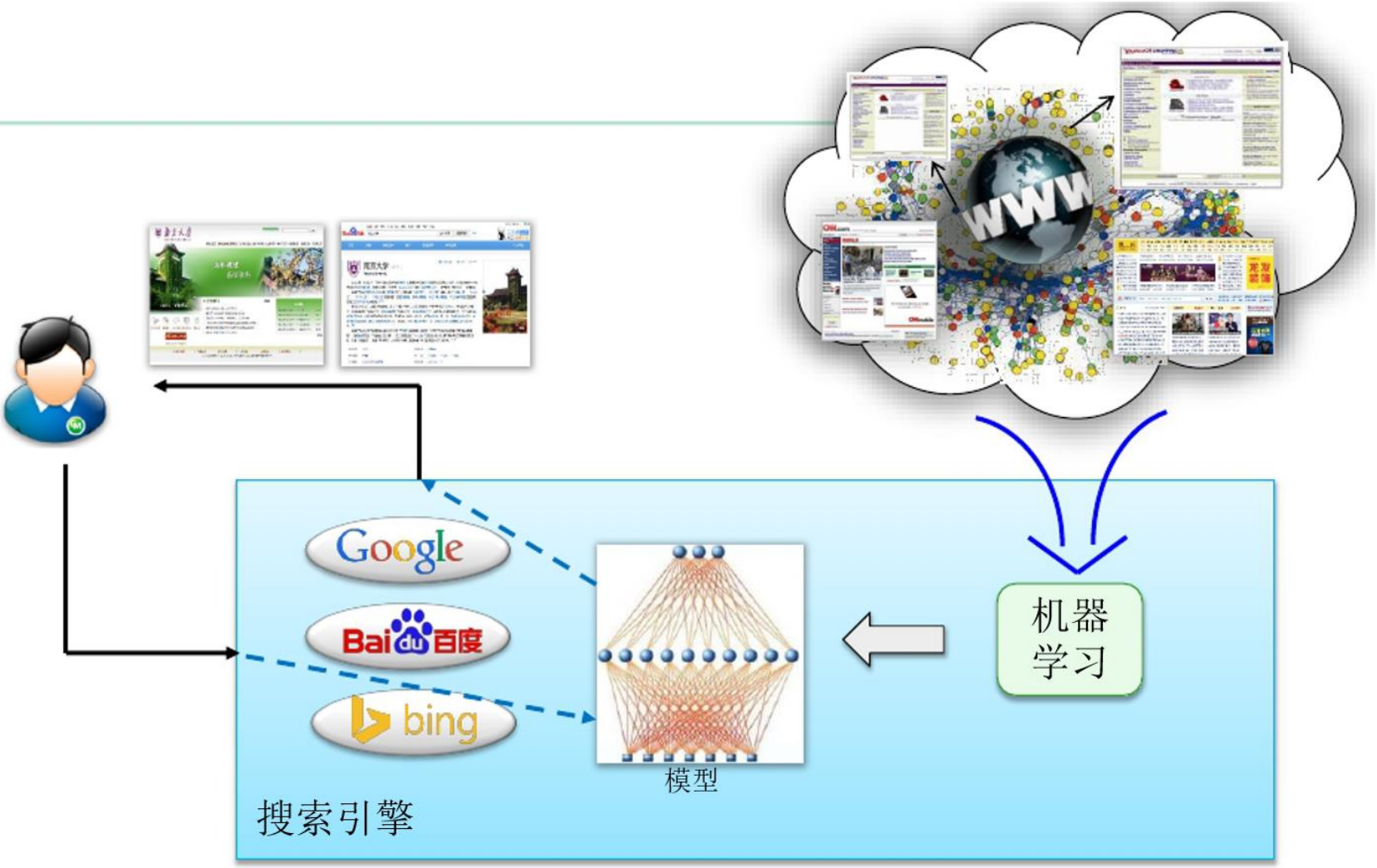
包含子问题：

a) 人脸检测，Face Detection
b) 人脸对齐，Face Alignment
c) 年龄分类，Age Classification

其他子问题：

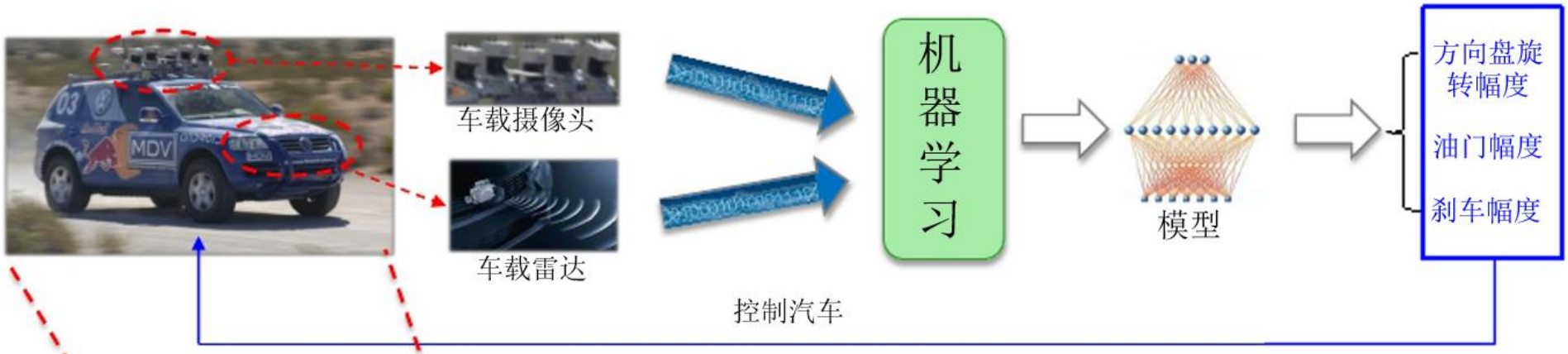
a) 人脸识别，Face Recognition
b) 性别识别，Gender Recognition
c) 表情识别，Expression Recognition
d) 种族识别，Race Recognition

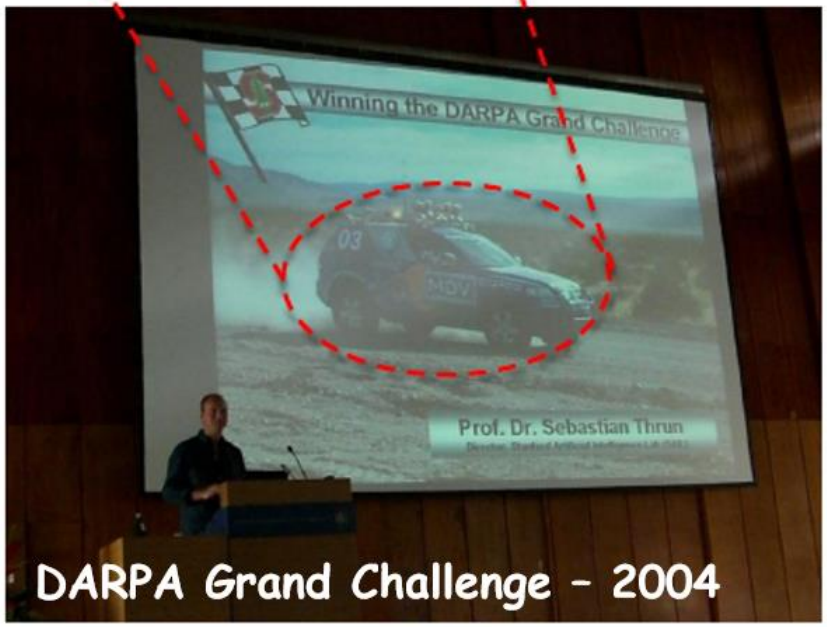
机器学习应用（网络搜索）

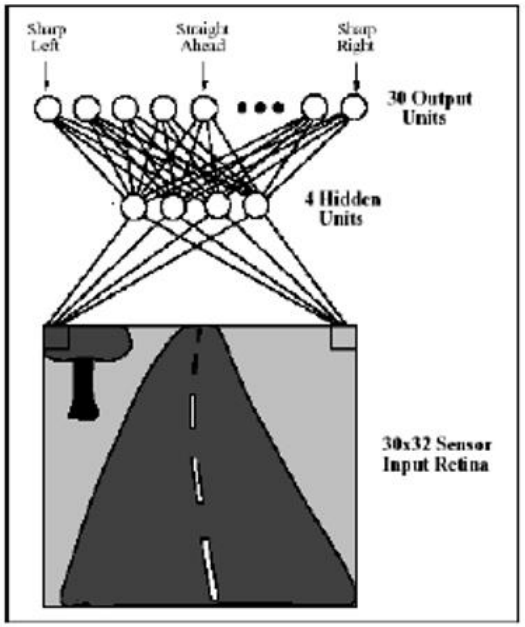


机器学习技术正在支撑着各种搜索引擎

机器学习应用（自动驾驶）







美国在20世纪80年代就开始研究基于机器学习的汽车自动驾驶技术

■ 机器学习源自“人工智能”

1956年夏　　美国达特茅斯学院

J. McCarthy, M. Minsky, N. Lochester, C. E. Shannon,
H.A. Simon, A. Newell, A. L. Samuel 等10余人

约翰 麦卡锡
(1927-2011)
“人工智能之父”
1971年图灵奖

达特茅斯会议标志着人工智能这一学科的诞生

John McCarthy (1927 - 2011):

1971年获图灵奖, 1985年获IJCAI终身成就奖。人工智能之父。他提出了“人工智能”的概念，设计出函数型程序设计语言Lisp，发展了递归的概念，提出常识推理和情境演算。出生于共产党家庭，从小阅读《10万个为什么》，中学时自修CalTech的数学课程，17岁进入CalTech时免修两年数学，22岁在Princeton获博士学位，37岁担任Stanford大学AI实验室主任。

■ 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

爱德华 费根鲍姆
(1936-)
1994年图灵奖

- 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)

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 ...

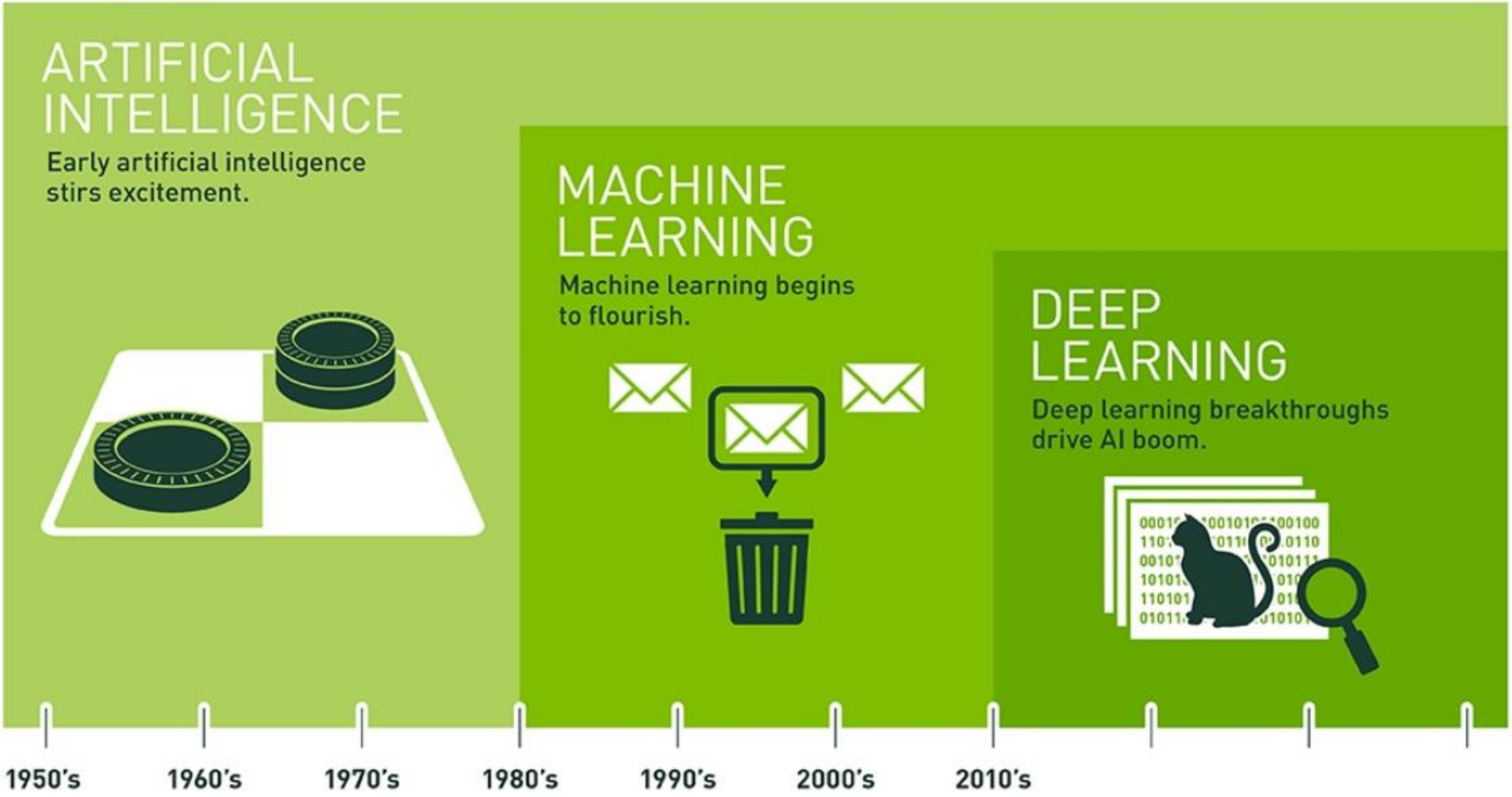
■ 第三阶段：学习期

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


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

奥巴马提出“大数据计划”后，美国NSF进一步加强资助UC Berkeley研究如何整合将”数据”转变为”信息”的三大关键技术——机器学习、云计算、众包(crowd sourcing)

National Science Foundation: In addition to funding the Big Data solicitation, and keeping with its focus on basic research, NSF is implementing a comprehensive, long-term strategy that includes new methods to derive knowledge from data; infrastructure to manage, curate, and serve data to communities; and new approaches to education and workforce development. Specifically, NSF is:

整合三大关键技术

- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next-generation Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for turning data into information - machine learning, cloud computing, and crowd sourcing;
- Funding the first round of grants to support "EarthCube" – a system that will allow geoscientists to access, analyze and share information about our planet;
- Issuing a $2 million award for a research training group to support training for undergraduates to use graphical and visualization techniques for complex data.
- Providing $1.4 million in support for a focused research group of statisticians and biologists to determine protein structures and biological pathways.
- Convening researchers across disciplines to determine how Big Data can transform teaching and learning.

大数据时代，机器学习必不可少！！

收集、传输、存储大数据的目的，

是为了“利用”大数据

没有机器学习技术分析大数据，

“利用”无从谈起

02

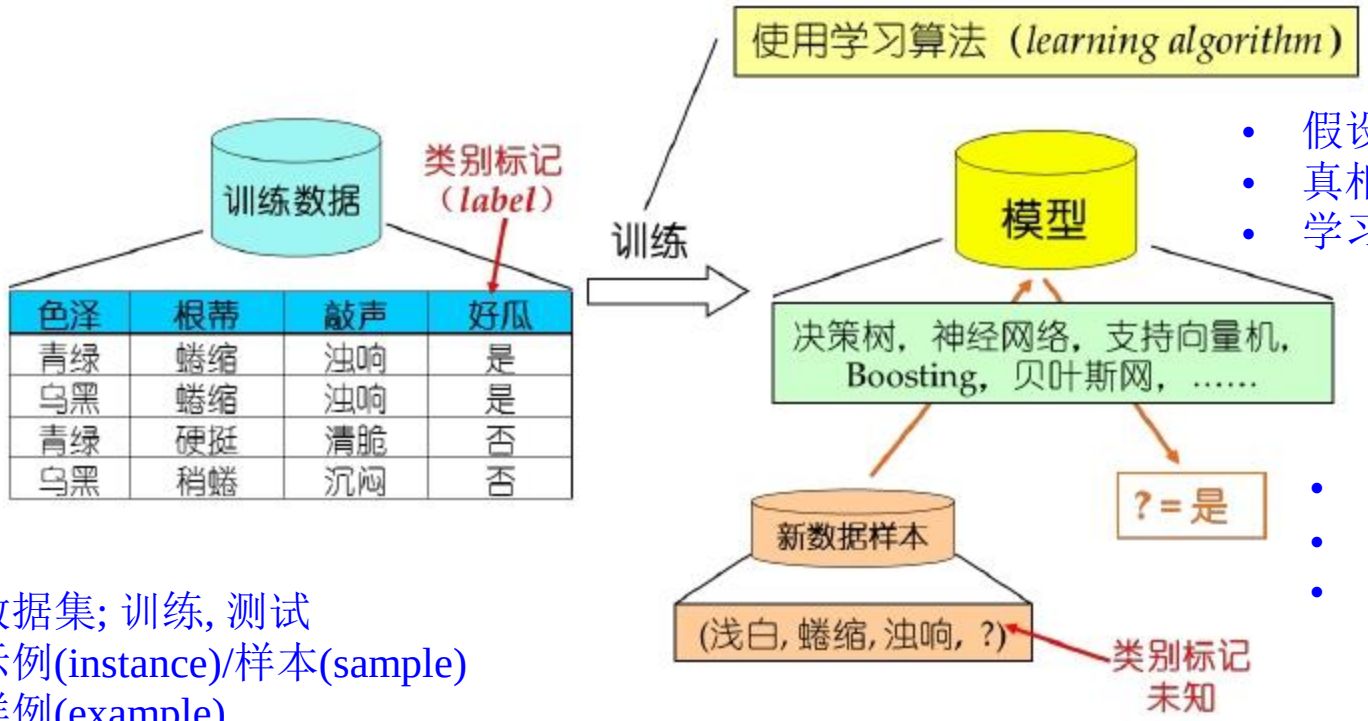
PART TWO

基本术语

Basic Notions

基本术语

- 监督学习(supervised learning)
-



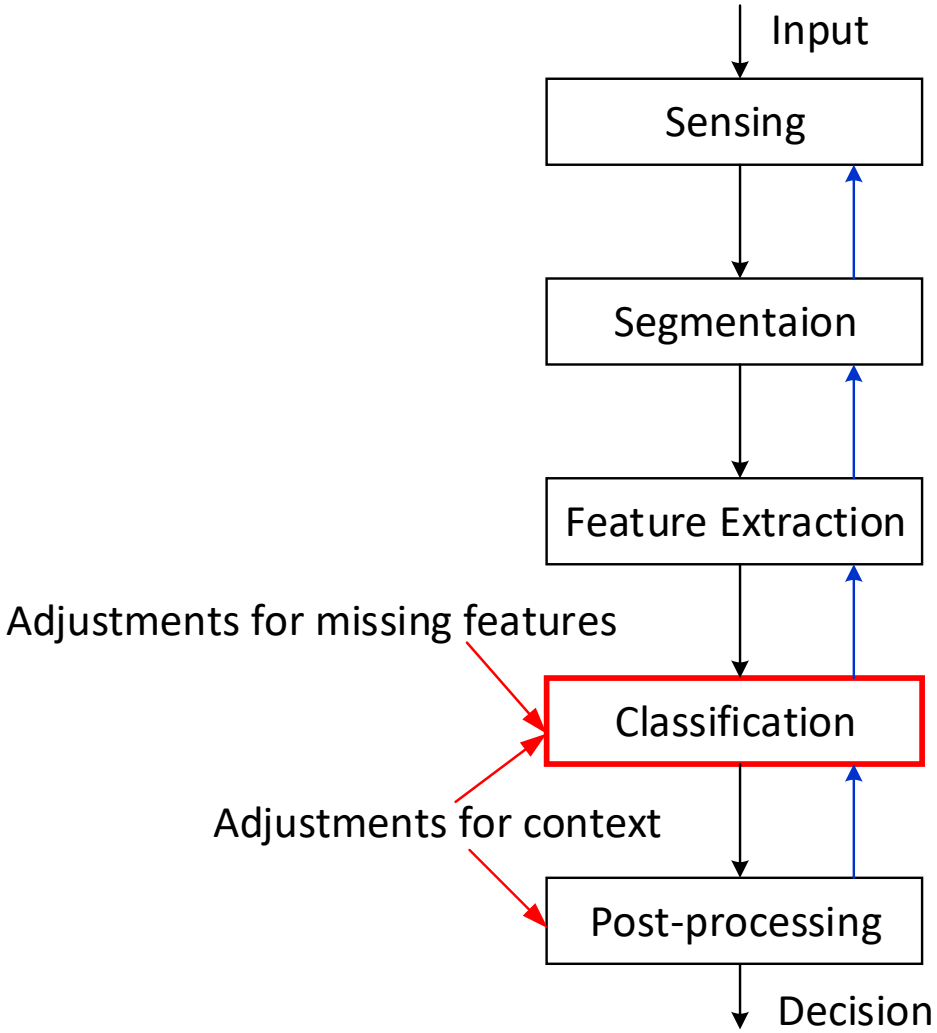
- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

- 分类, 回归
- 二分类, 多分类
- 正类, 负类

- 数据集; 训练, 测试
- 示例(instance)/样本(sample)
- 样例(example)
- 属性(attribute), 特征(feature); 属性值
- 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)**

➢ **参考《机器学习》，周志华著，p.1-3**



Context

Network anchor(网络主播), anchor, female anchor, camgirl,
Internet celebrity(网红)

cele= quick, speed=速度,快 (e.g. accelerate)

celebr与celer同源，表示“更快”，原指在节日里
因为狂欢而心跳加速，后指欢庆（节日）、举行（活动）

celebrity

movie stars

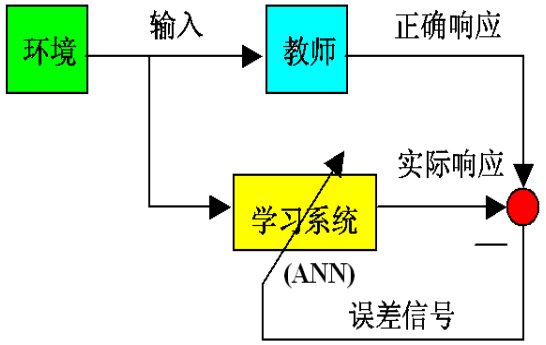
Ground truth

Context: Input dependent information

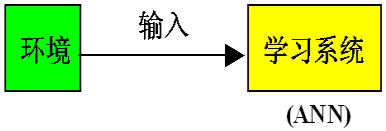
Tne cat → The cat

监督学习/无监督学习

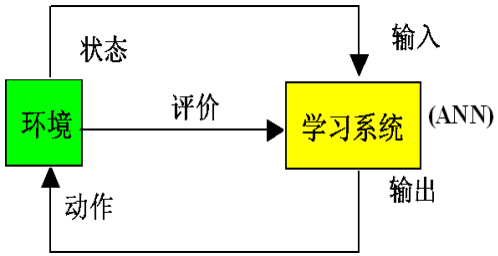
- **监督学习**(Supervised Learning)：监督学习是从标记的训练数据来推断一个功能的机器学习任务。如**分类、回归**。



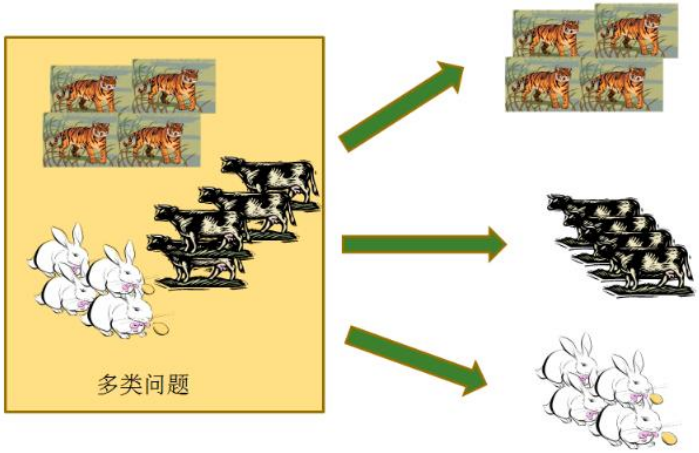
- **无监督学习**(Unsupervised Learning)：无监督学习的问题是，在未标记的数据中，试图找到隐藏的结构。如**聚类**、密度估计。



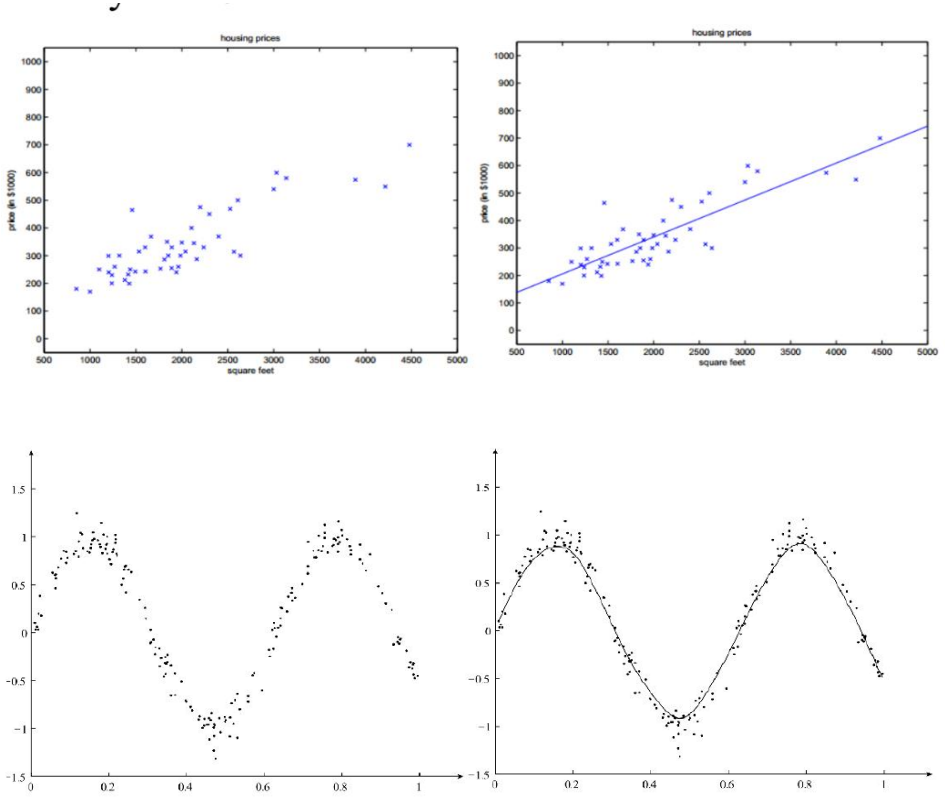
- **强化学习**(Reinforcement Learning)：强化学习是机器学习中的一个领域，强调如何基于环境而行动，以取得最大化的预期利益。



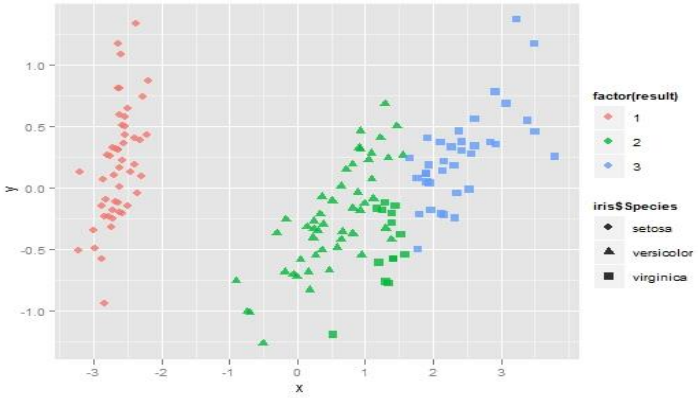
➢ 分类


多类问题

➢ 回归


housing prices
price (in $1000)
square feet
housing prices
price (in $1000)
square feet

➢ 聚类


factor(result)
1
2
3
iris$Species
setosa
versicolor
virginica
x
y

机器学习的基本过程

将数据对象进行特征(feature)化表示

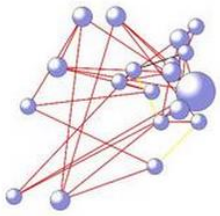
给定一个数据样本集，从中学习出规律（模型）

目标：该规律不仅适用于训练数据，也适用于未知数据(称为泛化能力)

对于一个新的数据样本，利用学到的模型进行预测

■ 向量表示法 $[x_1, \quad x_2, \ldots \quad x_n]$

■ 图表示法



➢ 天气预测：

样本：每一天

问题：如何把每天表示成一个向量？选取哪些特征？

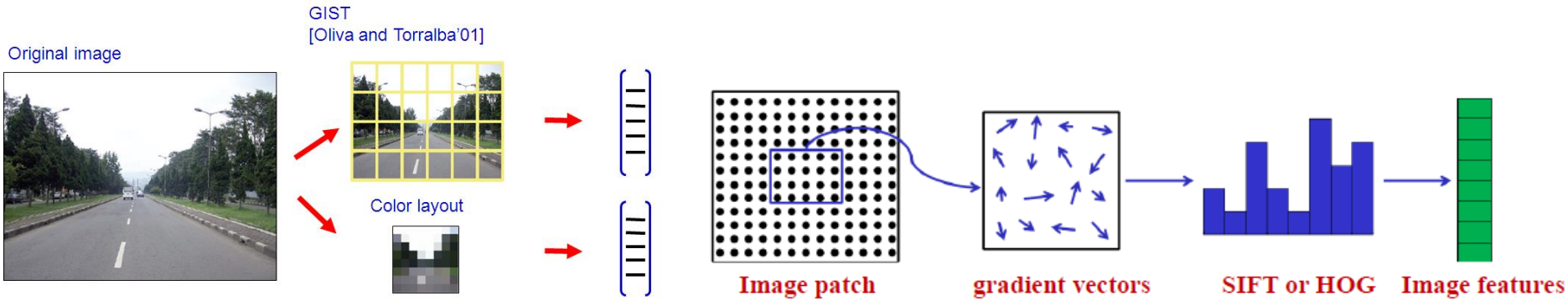
特征：温度，相对湿度，风向，风速，气压

➢ 判断好瓜坏瓜：

样本：每个瓜

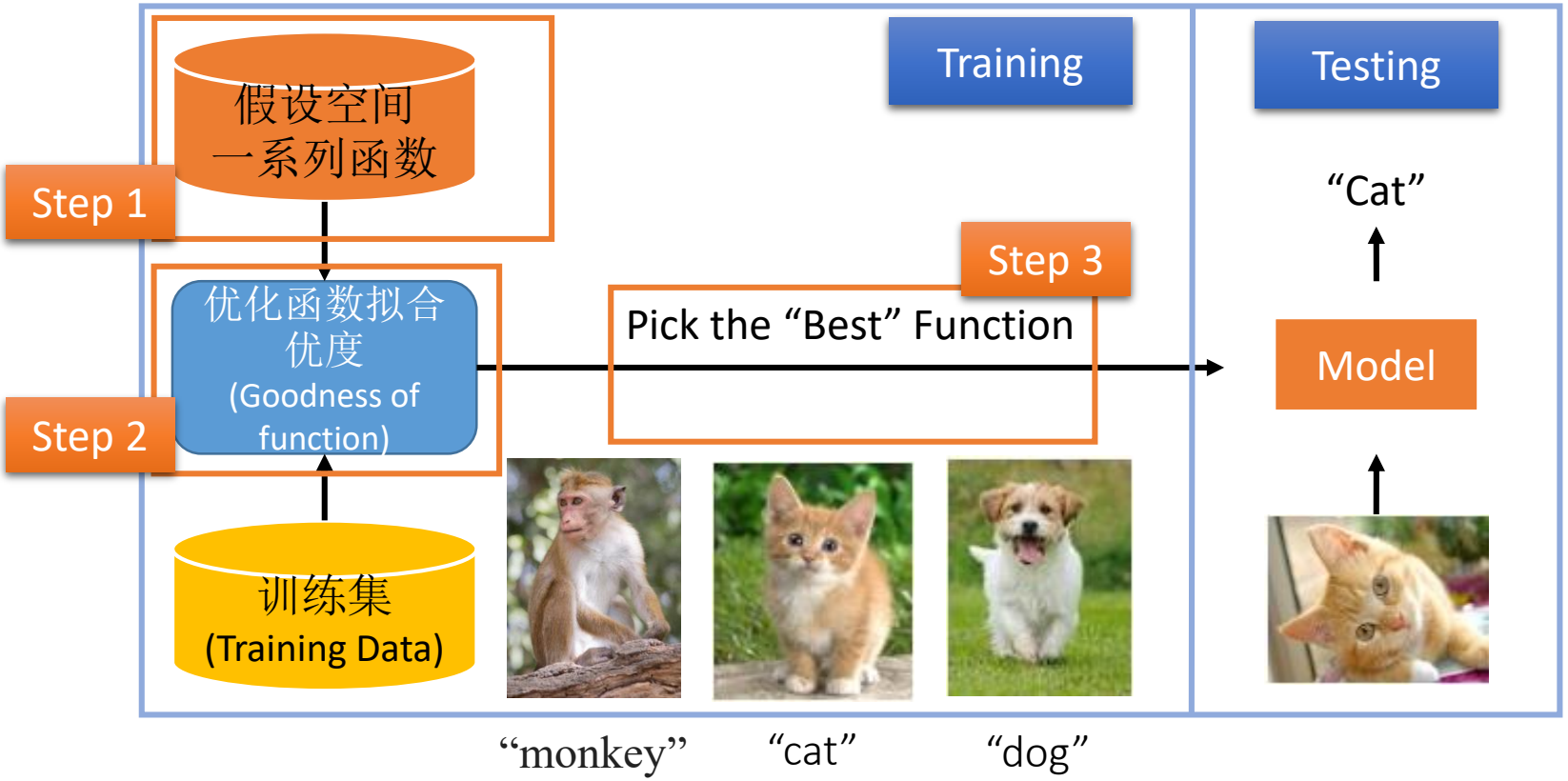
特征：色泽，根蒂，敲声

➢ 图像识别：

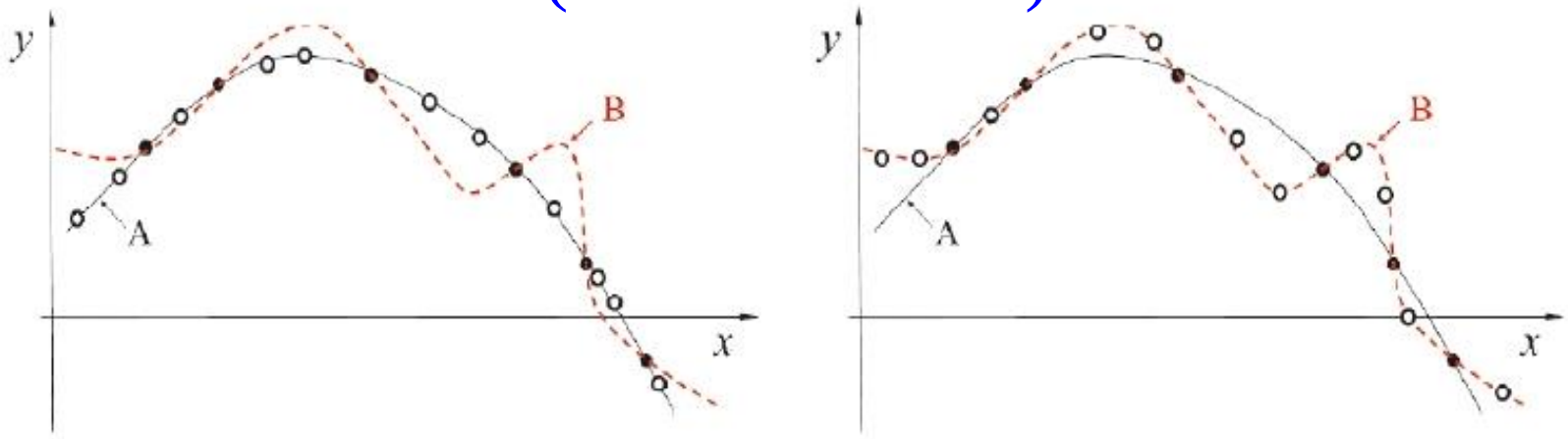


- 机器学习≈寻找函数“ f ”

➢ Image recognition



没有免费的午餐！
(No-free Lunch!)



(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 $\mathfrak{L}_a$ 若在某些问题上比另一个算法 $\mathfrak{L}_b$ 好，必存在另一些问题，$\mathfrak{L}_b$ 比 $\mathfrak{L}_a$ 好。

No free lunch定理寓意

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

不能脱离具体问题，空泛地谈论“什么学习算法更好”