西安交通大学 XI'AN JIAOTONG UNIVERSITY | IAIR Est. 1986 Institute of Artificial Intelligence and Robotics | 人工智能学院 College of Artificial Intelligence, XJTU

《高级机器学习》第一章

人工智能与机器学习引论

Introduction to AI and ML

魏平
西安交通大学人工智能学院
人工智能与机器人研究所

课程目标

- 掌握机器学习的概念、原理和方法
- 使用机器学习方法解决人工智能中的基本问题
- 领会并内化建模、学习、推理、优化、随机等思维方式，建立“学习”的思维习惯
- 洞悉人工智能与机器学习未来发展趋势，为今后进一步从事相关专业奠定基础

主要内容

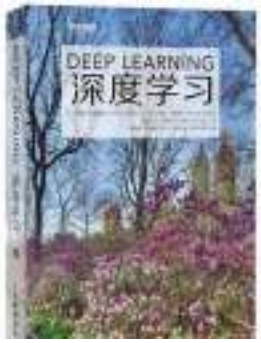
周次	主题	内　容
1	引论	机器学习背景、基础知识
2	线性模型与核方法	线性模型的概念和方法、支撑向量机、结构化支撑向量机、隐变量支撑向量机、核方法
3	随机图模型	随机图模型、贝叶斯分类、马尔科夫随机场、隐马尔科夫模型、条件随机场、图模型推理
4	集成学习	决策树、集成学习、Boosting、随机森林
5	压缩感知	稀疏性与压缩感知
6	蒙特卡罗方法	概率采样、蒙特卡洛方法
7	神经网络	深度学习、卷积神经网络、循环神经网络、图神经网络、
8	深度注意力模型	注意力模型、Transformer、大模型

教材及考核

教材

Christopher Bishop,《Pattern Recognition and Machine Learning》

周志华 著，《机器学习》清华大学出版社，2016



Ian Goodfellow, Yoshua Bengio and Aaron Courville,《Deep Learning》

基础知识

- 矩阵/线性代数
- 高等数学
- 概率/随机过程
- 优化理论

考核

课程考核项目：70%，平时成绩（考勤、作业）：30%

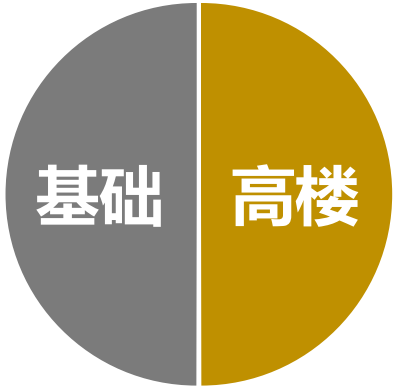
课程要求

- 严格遵守学术规范
- 作业和报告不得抄袭
- 上课不录音、不录像、不直播
- 不要把课程的音视频、课件、作业等发到网上

基础方法与最新方法的关系

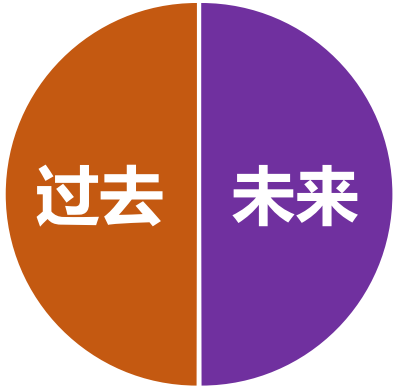
人工智能都这么强大了，为什么还要学习传统机器学习方法?

深度学习是当前AI应用的主要方法，为什么还要学习其他方法?



基础**决定**高度
高楼**反映**基础

内功**支撑**技法
技法**促进**内功



过去**启发**未来
未来**延续**过去

What is past is prologue. 凡是过往，皆为序章 — 莎士比亚

IAIR Est. 1986 Institute of Artificial Intelligence and Robotics

人工智能学院 College of Artificial Intelligence, XJTU

CONTENTS

什么是智能 (Intelligence)

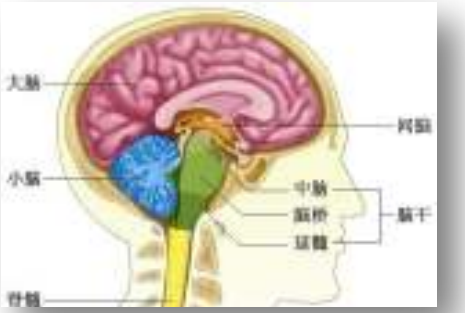
智能的含义

- 智能指人的智慧和行动能力 —— 现代汉语字典
- The ability to learn, understand, and make judgments or have opinions that are based on reason —— 剑桥词典
- 智能是指个体或系统在特定环境中，通过感知、学习、推理、决策等能力，高效解决问题或实现目标的一种综合素养 —— Deepseek-R1
- 所以知之在人者谓之知，知有所合谓之智。所以能之在人者谓之能，能有所合谓之能 —— 《荀子·正名篇》
- 智能之士，不学不成，不问不知 ——《王充·论衡》

人类智能 (Human Intelligence)

- 人类的“智”和“能”不可分离
  - 智：人对事物的认识能力
  - 能：人的行动能力，包括各种技能和习惯

- **人类智能的载体—大脑**
  - 大脑是内外环境信息获得、存储、处理、加工和整合的中枢
  - 大脑的活动所需能量相当于20瓦灯泡的功率
  - 大脑约有1000亿个神经细胞，有可能超过100万亿连接
  - 人脑神经细胞回路比全世界电话网络还复杂1400多倍，最快的神经冲动传导速度为400多公里/小时
  - 每天可处理8600万条信息，其记忆贮存的信息超过任何一台电子计算机



人类的多元智能

人类的多元智能

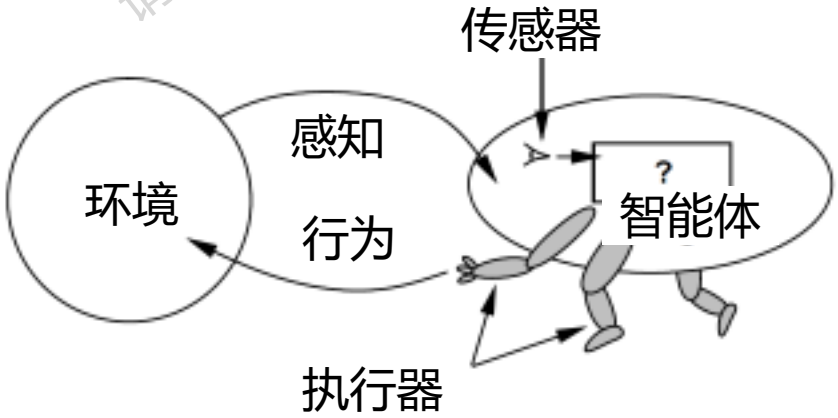
哈佛大学心理学、教育学教授加德纳 “心灵框架：多元智能理论”（Frames of Mind: The Theory of Multiple Intelligences）（1983）：将人类智能区分为8个多特定的“模态”，而不是将智力视为由单一的一般能力主导

什么是人工智能?

- 人工智能是由机器展示的智能，研究领域定义为对“智能体”的研究：任何能够感知环境、并采取行动最大化其在某个目标上成功机会的机器 --- 维基百科

- 人工智能寻找一种方法，将智能映射到机械硬件，并使一个结构进入该系统以形式化思维 --- 《人工智能 - 一种现代方法》

- 人工智能是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学 --- 科学百科



智能体 = 结构 + 程序

让机器拥有像人一样并超越人的智能

人工智能的三个层次

超级人工智能

- 具有自我意识的人工智能

电影《终结者》

通用人工智能

- 在各种任务中都具有智能

大语言模型助力

弱人工智能

- 为特定任务设计的智能

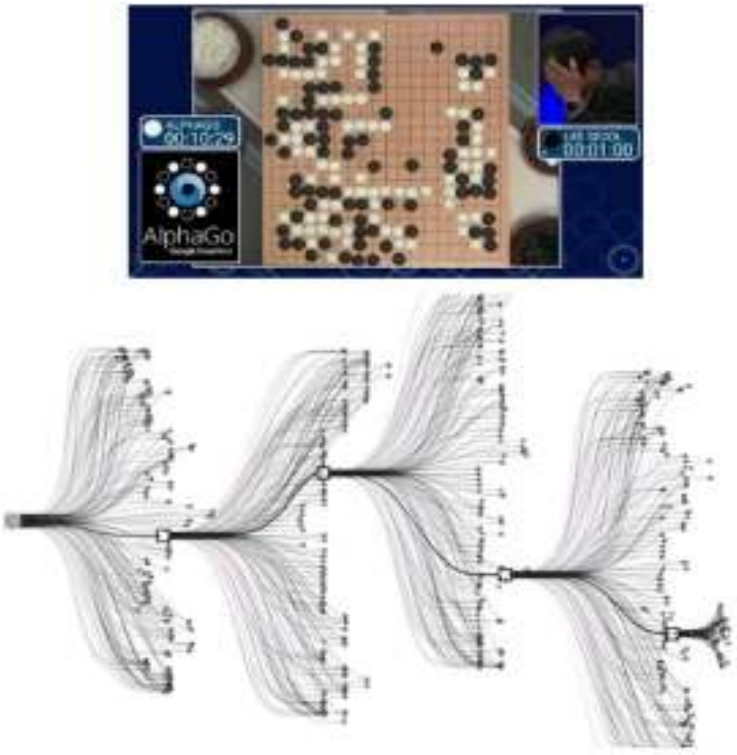
人脸识别

人工智能的学科交叉与渗透性

- **人工智能的交叉领域**
  - 数学 Mathematics
  - 社会学 Sociology
  - 哲学 Philosophy
  - 计算机科学 Computer Science
  - 心理学 Psychology
  - 神经科学 NeuroScience
  - 生物学Biology
  - … …

人工智能的复杂性 1

计算复杂性: 指数增长



AlphaGo, 2016

- 假设人类执黑子，机器执白子，棋盘上还有200个空位，白棋计算最优落子位置，理论上要搜索200!数量级的可能性

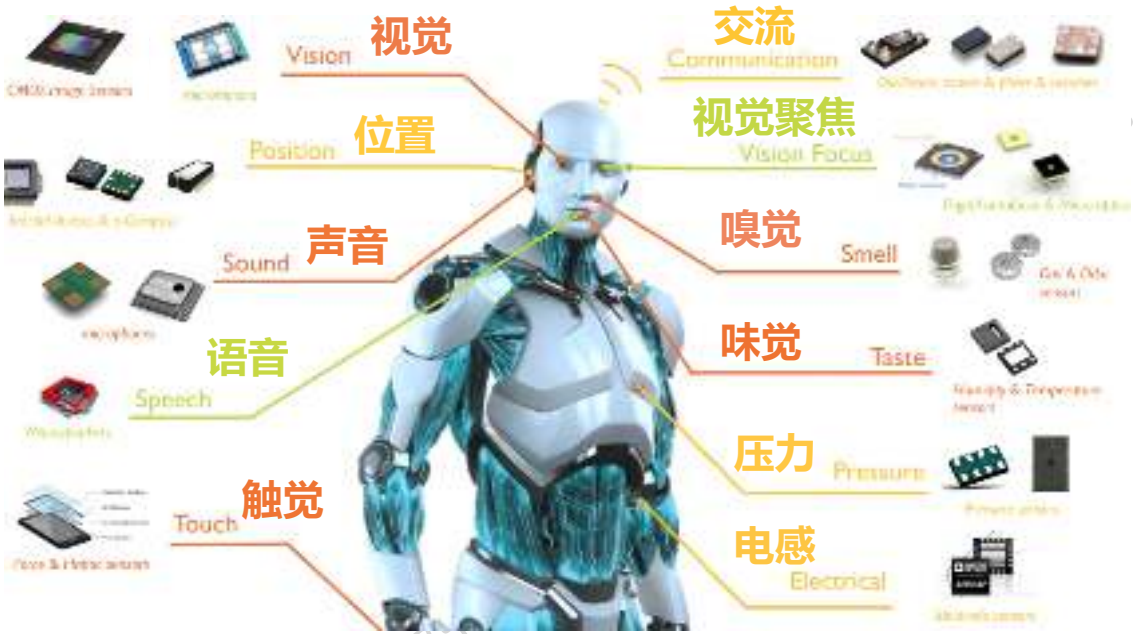
$$200! \approx 10^{50}$$

- 1光年=9.46× 10^{12}千米
- 地球上所有沙子的数量级约10^{23}
- 地球上海洋中的水分子总数约10^{47}

- 人类大脑的功率约20瓦
- 中国普通家庭的用电功率约5KW
- AlphaGo机器功率2000KW

人工智能的复杂性 2

信息复杂性：观测信息的局限性



感知系统

局部观测

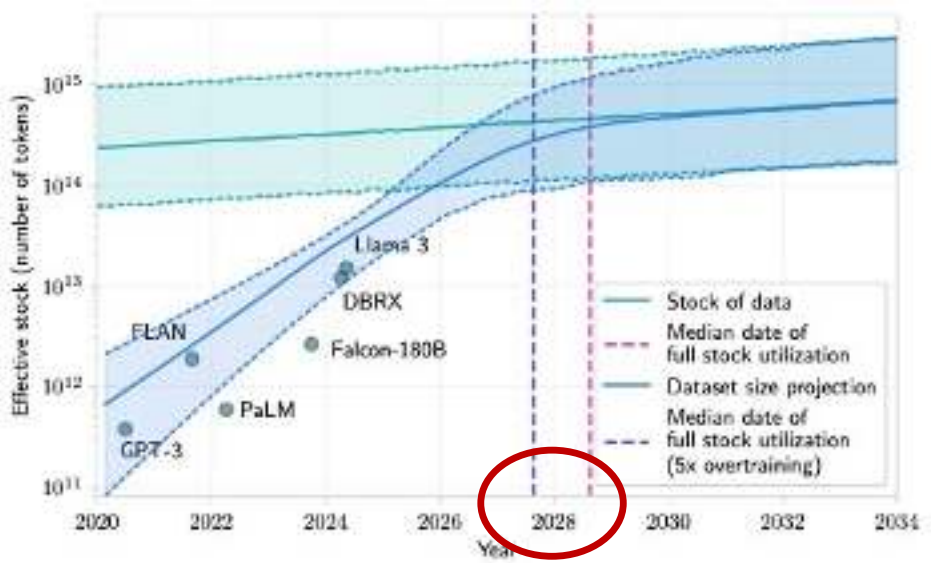
人工智能的复杂性 3

数据复杂性：海量性和稀缺性

全球每天产生的数据量

5亿TB

- 1TB=40万本《红楼梦》文字
- 谷歌图书所有的数据量约10万TB
- 全国图书馆文本数据量约200万TB



MIT研究估计2028年左右，大模型将用尽现有所有人类存储的高质量数据

人工智能的脆弱性

人工智能对人类的意义甚至比火与电更深刻

霍金的警告

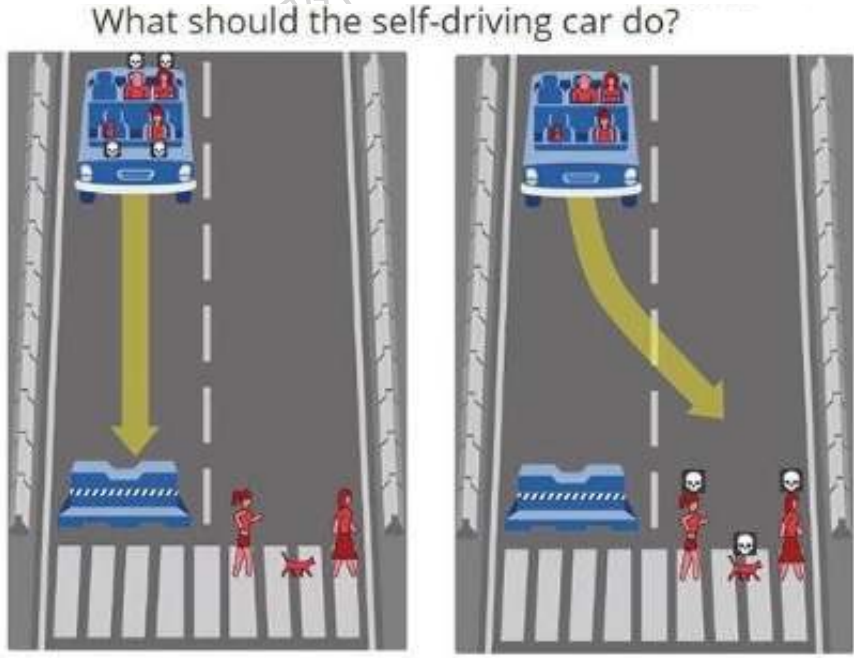
“人工智能的成功有可能是人类文明史上最大的事件。但人工智能也有可能是人类文明史的终结，除非我们学会如何避免危险。我曾经说过，人工智能的全方位发展可能招致人类的灭亡。”

—— 斯蒂芬·霍金

人工智能的风险问题

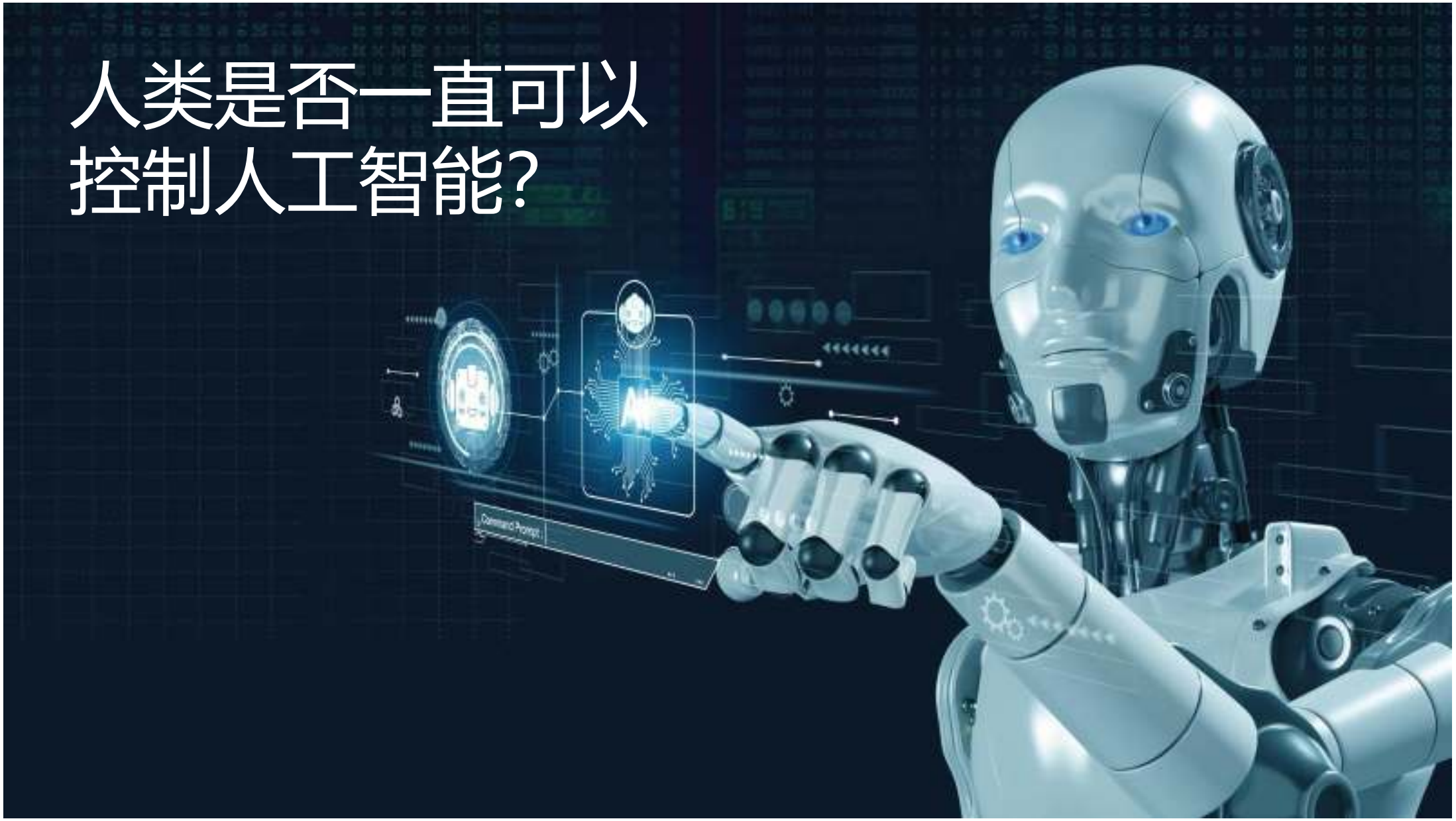
- 人工智能自身没有道德、伦理，人工智能技术及智能机器所引起的危害人类伦理、道德、法律、安全的问题
- 2024年全球人工智能风险突发事件增加50%
- 现阶段弱人工智能技术引发的风险
  - 算法歧视与偏见：人工智能算法在训练数据中的质量偏颇及隐含信息等，可能导致方法结果产生违法人类道德规范的后果，或者带来或扩大社会中的歧视，造成社会问题
  - 权利与利益侵害：人工智能会导致算法黑箱问题，使决策不透明或难以解释，从而影响公民知情权、程序正当及公民监督权，而且人工智能的滥用可能威胁公民隐私权、个人信息权
  - 误导与欺骗问题：假新闻撰写和智能化定向传播、深度伪造等滥用可能导致信息茧房、虚假信息泛滥等问题，虚假新闻的精准推送还可能加大影响人们对事实的认识和观点，进而可能煽动民意、操纵商业市场和影响政治及国家政策
  - 智能不公平问题：掌握技术、数据、算力的群体可以肆意使用人工智能技术，而没有这些资源的群体无法享受AI带来的变化，进而在影响自身发展；AI加速社会分工和职业重塑，影响收入分配机制和收入水平

无人驾驶车的道德困惑



无人驾驶车在危险情况下需要实时决定是否优先考虑乘客的安全，而不是行人或道路上的其他司机？

人类是否一直可以控制人工智能?



UBTECH 优必选 | 9880.HK 人形机器人第一股

全球首个实现自主换电的人形机器人

Walker S2

强人工智能伦理问题

- 未来人工智能是否可能产生自我意识，具有像人一样的自由意志，甚至可能开发出其自己与人类意愿相违背的机器智能系统，从目前来看都是未知

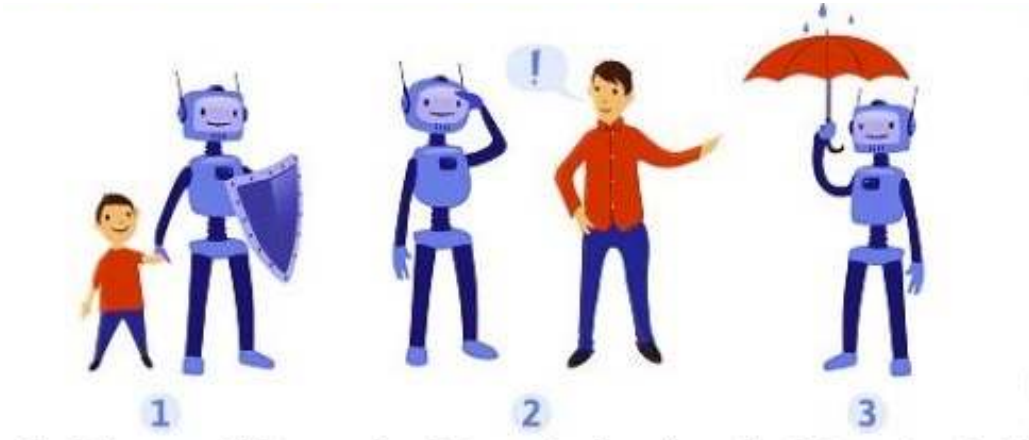
1. 具备自我意识的智能机器的社会地位：智能机器产生自我意识，需认可其某些人类属性(情感)及相应权益，要在哲学层次解决其身份、人格、受伤害时的“人道主义”、在人类社会中的角色等问题

2. 具有自我意识的智能机器的权利义务：智能机器与人类相比权力、社会劳动和责任的划分，法律人格、人身伤害责任界定、折磨或虐待的保护等有待解决

3. 自我意识的智能机器与人类复杂关系：具有自我意识的智能机器与人类之间的关系非常复杂，对社会传统的个体、家庭及机构企业间的关系带来复杂影响，需要从法律、哲学等角度探讨研究

阿西莫夫机器人三定律

- 1942年，美国科幻作家艾萨克·阿西莫夫（Isaac Asimov）在“Runaround”小说中提出机器人三定律（通常简称为三定律或称为阿西莫夫定律）



Zeroth Law:			
A robot may not harm humanity, or, by inaction, allow humanity to come to harm.	A robot may not injure a human being or, through inaction, allow a human being to come to harm.	A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.	A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
机器人不得伤害人类，或不得因不作为而让人受到伤害	**机器人不得意外伤害人，或不得因不作为而让人受到危害**	**机器人必须服从人的命令，除非这些命令与第一定律相冲突**	**机器人必须保护自己的存在，只要这种保护不违反第一或第二定律**

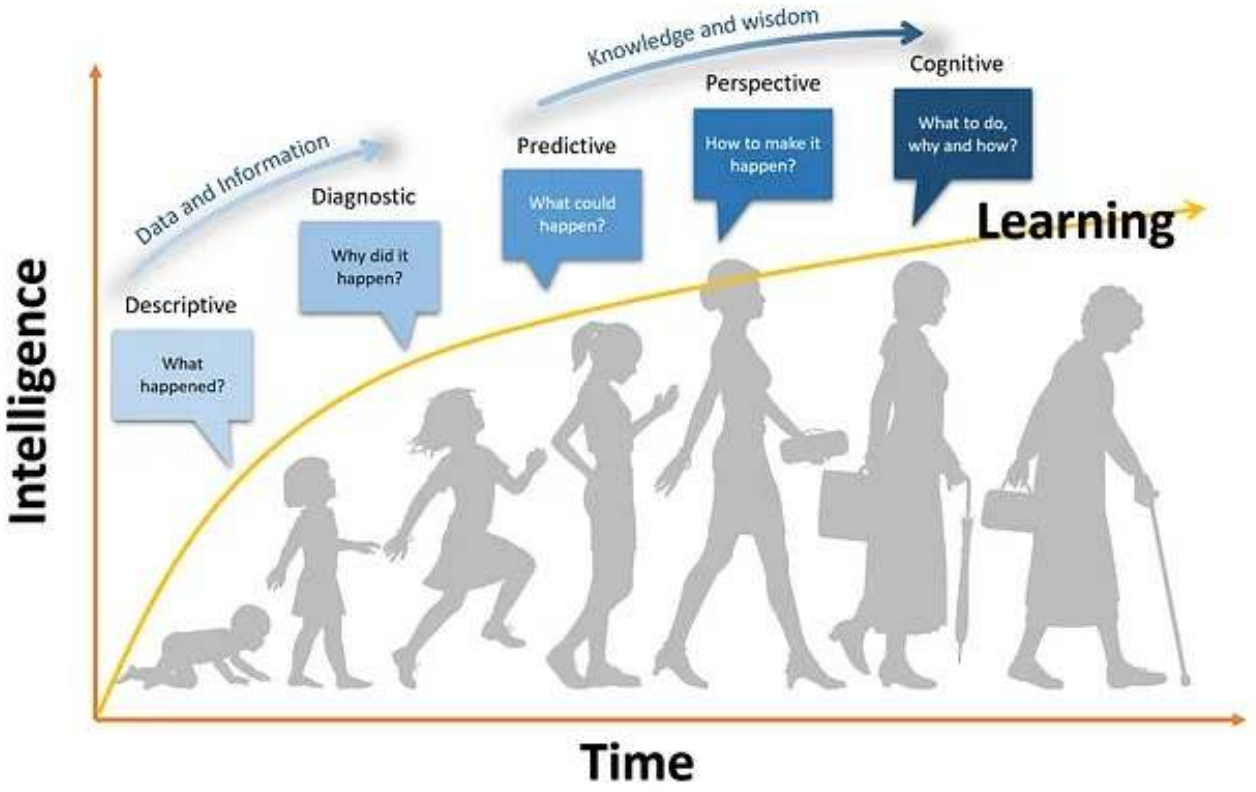
- 该机器人定律可被看作最早看待人与机器人之间伦理关系的基本原则，对机器人的行为以及机器人与人之间的保护与被保护关系做了简明规定
- 该定律具体内容可作为指导原则应用于机器人系统设计、开发、测试、实施、使用和维护

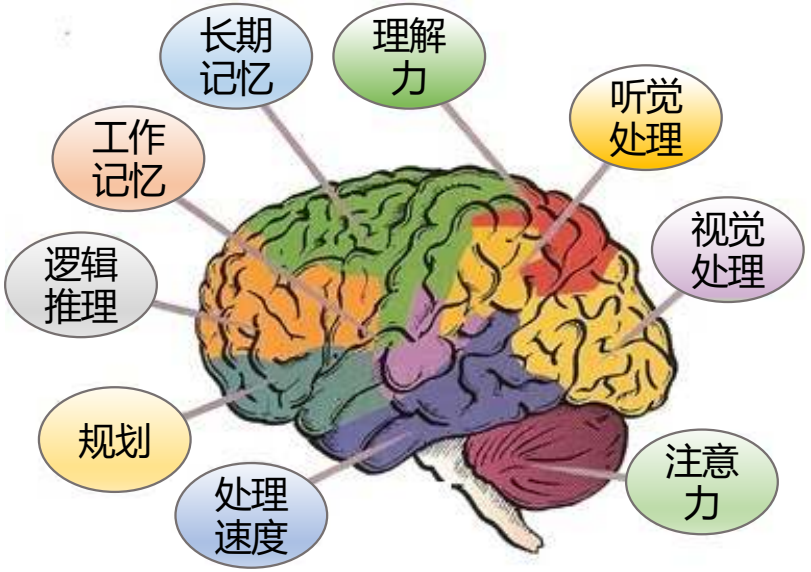
CONTENTS

- 什么是人工智能
- 为什么需要机器学习
- 什么是机器学习
- 机器学习的历史
- 机器学习的典型应用
- 人工智能、机器学习和深度学习

人类的学习能力

- 学习能力是人类的基本能力，是人类智能的根本特征，人类天生具有学习能力





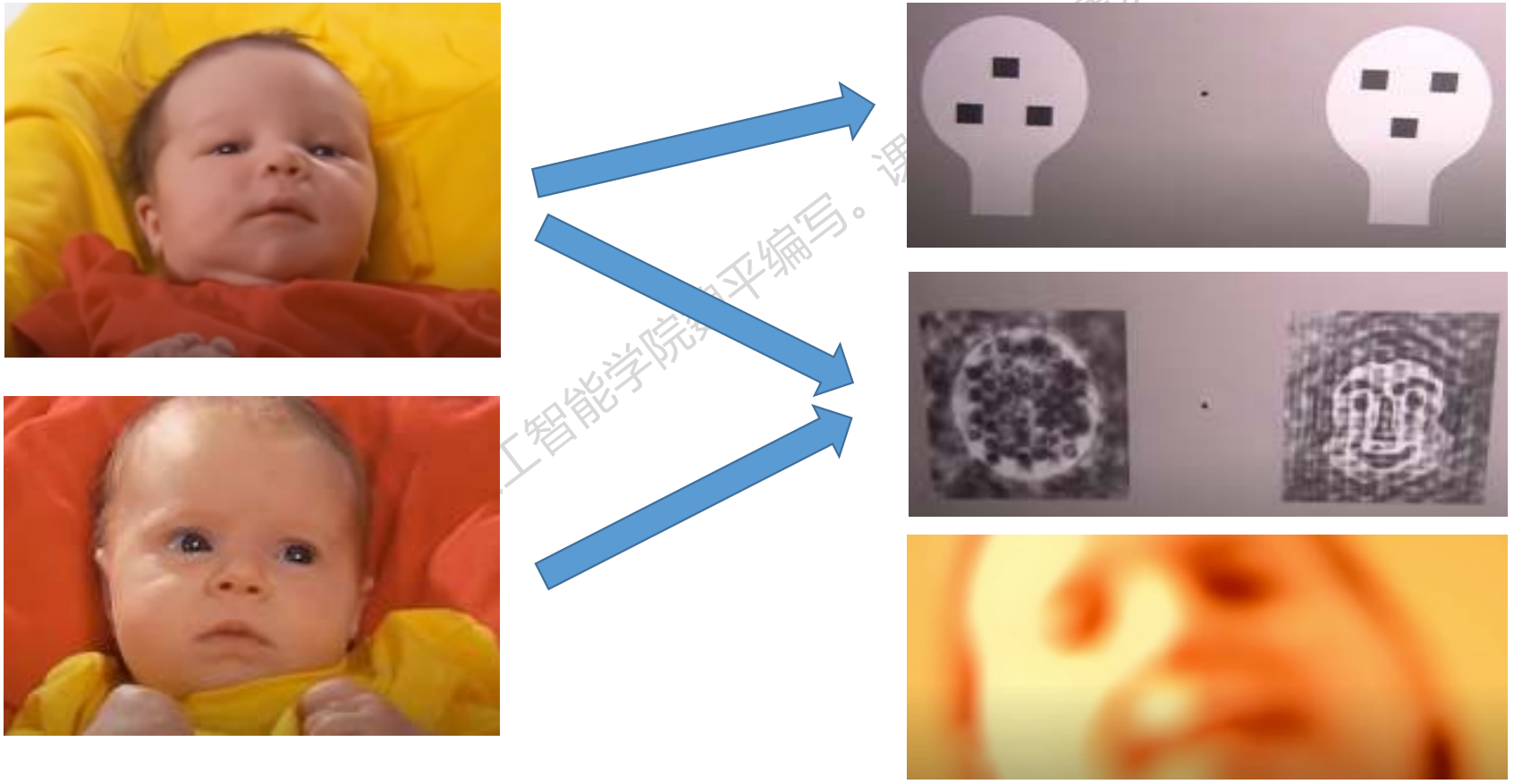
人脑学习的几种方式

人类的学习能力

- 学习能力是人类的基本能力，人类天生具有学习能力

人类的学习能力

- Face Recognition Experiment, Daphne Maurer



动物的“学习”行为

- 动物在遗传因素和环境因素作用下，通过“观察”和“经验”积累，在自然竞争中某些行为和能力得到了选择，并不断发展、进化

动物在自然环境中使用工具

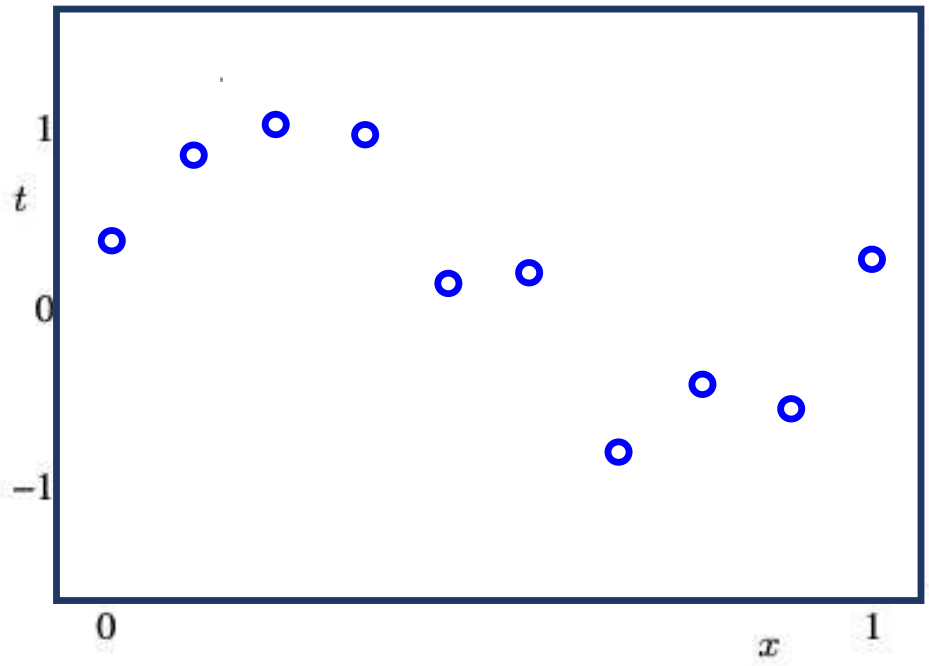
动物在复杂环境中适应和学习

动物的学习能力

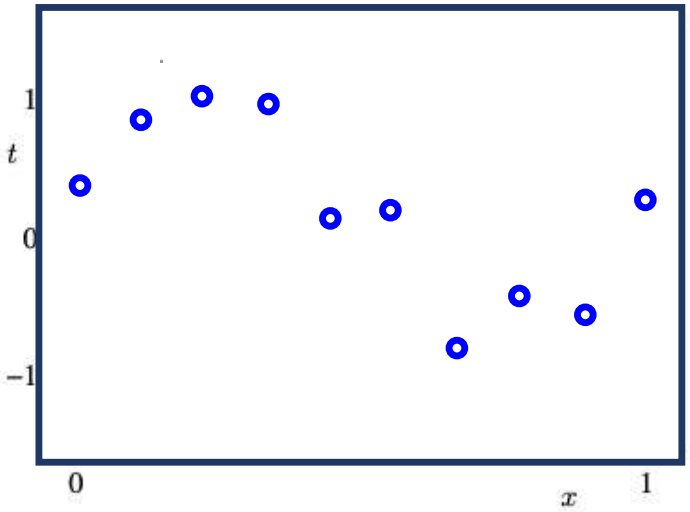
- 一般认为学习能力是人区别其他动物的重要标准，最新研究表明动物的学习能力可以不断拓展，为了保证信息纯净度，它们选择小范围的学习

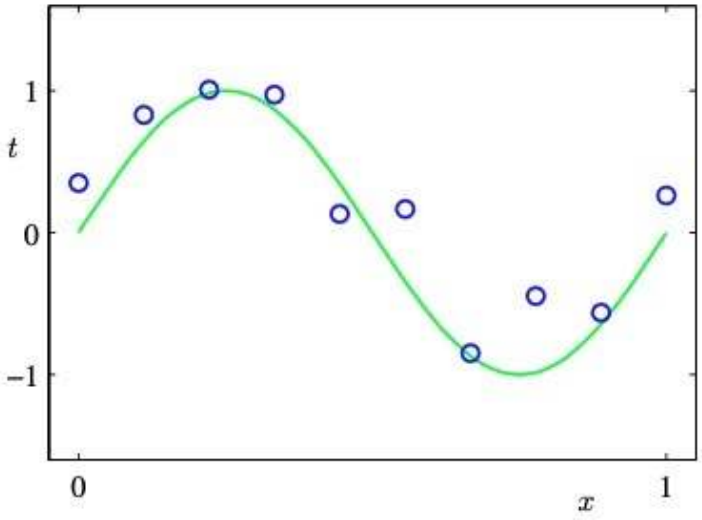
例子 1 — 曲线拟合

给定N个输入数据 $(x_1, \mathrm{x}_2, \ldots \mathrm{x}_N)$以及对应函数值 $(t_1, t_2, \ldots t_N)$，求一个新的x对应的t值是多少?



例子 1 — 曲线拟合





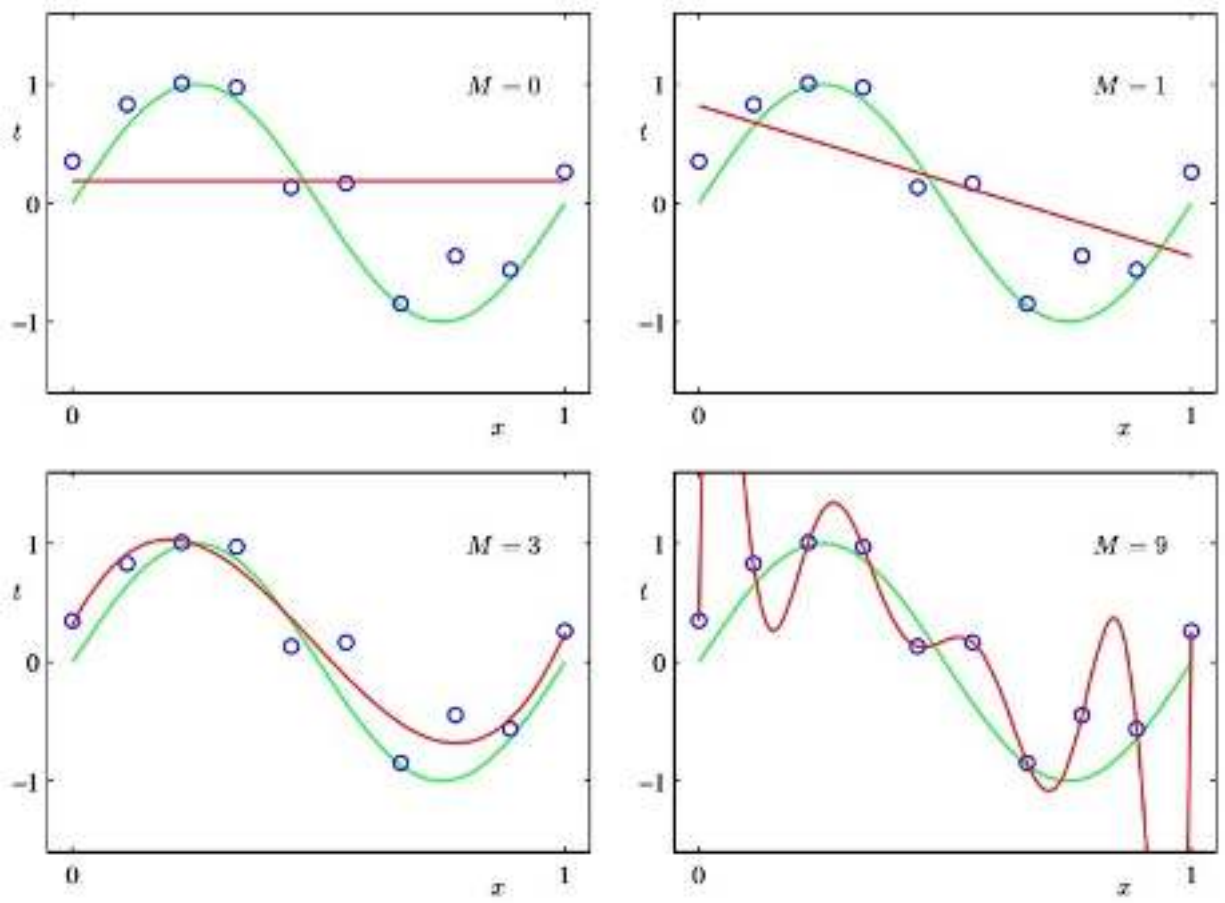
$t = \sin(ax)$ → $t = \sin(2\pi x)$

例子 1 — 曲线拟合

其他函数可以吗?

如何评价哪个函数最优?

如何从已知数据求函数的参数?



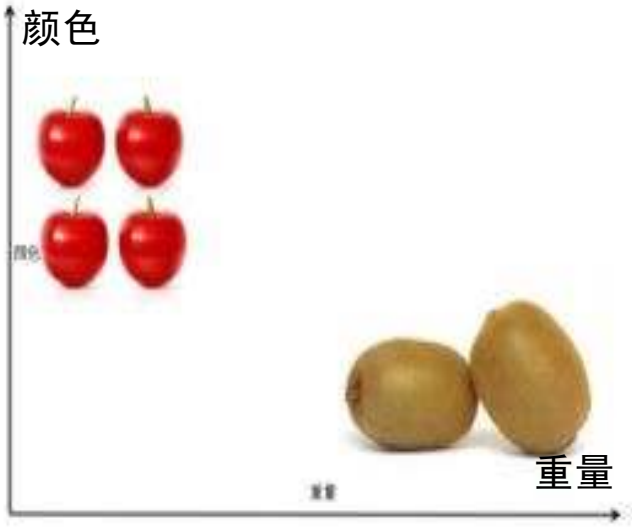
例子 2 — 如何区分樱桃和猕猴桃

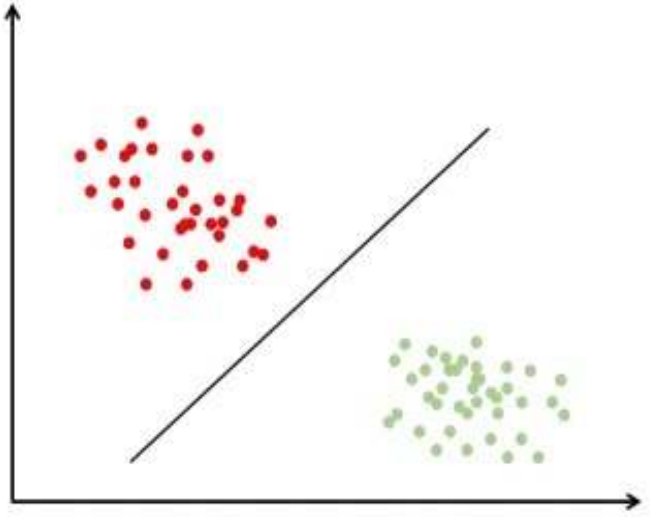
已知多个猕猴桃和樱桃图像样本

问题：

这张图片中的对象是樱桃还是猕猴桃?

例子 2 — 如何区分樱桃和猕猴桃





直线方程

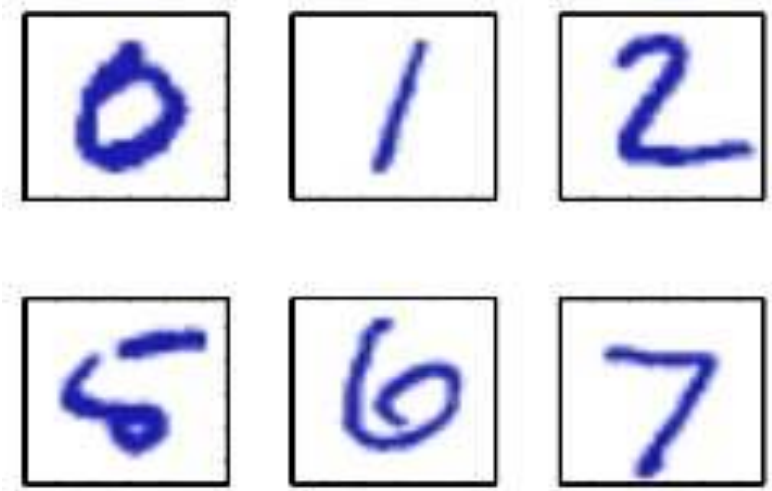
$ax + by + c = 0$

樱桃：$ax + by + c > 0$

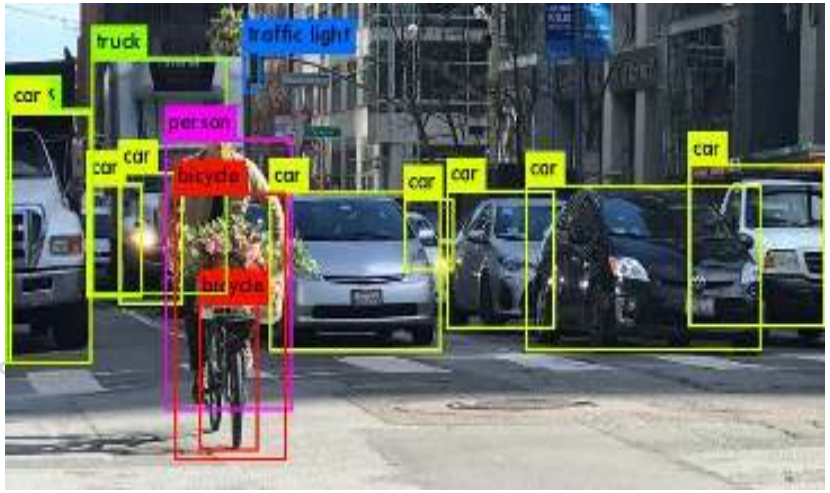
猕猴桃：$ax + by + c < 0$

- 选择两个明显特征-颜色x，重量y，组成特征向量(x, y)
- 计算给定样本的特征向量

其他例子



机器如何识别数字



无人车如何检测与识别目标

机器如何下围棋

机器人如何行走

- **数据包含了大量的有用信息**

  数据能给我们提供什么？如何使用数据？如何表达数据？

- **复杂的任务需要足够的知识和经验去才能完成**

  如何获得知识？如何表达知识？如何使用知识？

- **复杂的任务需要智能，一个系统是否具有学习能力成为是否具有“智能”的重要标志**

  如何获得智能？如何体现智能？如何表达智能？

机器学习是关于数据、知识、智能的学科

CONTENTS

- 什么是人工智能
- 为什么需要机器学习
- 什么是机器学习
- 机器学习的历史
- 机器学习基础概念

什么是机器学习?

- Field of study that gives computers the ability to learn without being explicitly programmed. ----
  Arthur Samuel (1956)

Arthur Lee Samuel
1901 – 1990,
Bell, IBM, Stanford

plays checkers with an IBM computer

- A computer program is said to learn from experience *E* with respect to some task *T* and some performance measure *P*, if its performance on *T*, measured by *P*, improves with experience *E*.
  ----Tom Mitchell (1997)

什么是机器学习?

- 对于某种任务 *T* 和性能度量 *P*，一个计算机程序被认为可以从经验 *E* 中学习是指：利用经验*E*，它在任务 *T*上由性能度量 *P* 衡量的性能提升

实现机器学习的三大要素

- 任务 *T* (Task)
- 性能 *P* (Performance)
- 经验 *E* (Experience)

机器学习是研究算法的学科，这些算法

- 利用经验 E，
- 使得在某种任务 T 中，
- 性能 P 得到提高

机器学习任务T

- 任务T是智能系统执行的、实现某种目标的工作，也可描述为智能系统处理一个样本(example) 的工作
- 样本(example)是从对象或事件中收集的已经量化的特征(feature)集合
- 特征(feature)是从对象或事件中提取的描述某种属性的度量

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \end{bmatrix} \in \mathbb{R}$$

机器学习基本任务

分类 (Classification)

- 判断输入属于已知的k个类别的哪一类
- 机器学习算法的目标是产生一个函数
- 常见的分类任务：目标识别

$$f: \mathbb{R}^n \rightarrow \{1, \ldots, k\},$$
$$y = f(\boldsymbol{x})$$

{dog, cat, tiger, bird} ?

分类器 $y = f(\boldsymbol{x})$

cat

机器学习基本任务

输入缺失分类 (Classification with missing inputs)

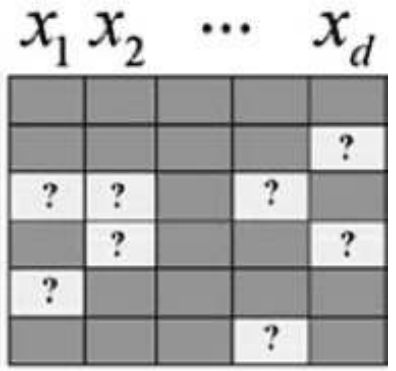
- 判断输入属于已知的k个类别的哪一类
- 输入向量的每个度量不能保证被观测
- 机器学习算法的目标是学习一组函数，每个函数对应具有不同缺失子集的输入

例子：疾病诊断

Table 1 [illegible]

[illegible]		[illegible]		[illegible]		[illegible]	
年龄	2	[illegible]	2	[illegible]	168	肌酐	110
[illegible]	2	[illegible]	2	[illegible]	300	TSH	110
HDLC	8	[illegible]	16	PCI前CK	300	PCI前CKMB	300
[illegible]	16	[illegible]	19	白蛋白	301	白细胞	303
[illegible]	19	[illegible]	20	LPa	310	[illegible]	115
[illegible]	22	[illegible]	27	BNP	328	[illegible]	141
[illegible]	31	心率	48	[illegible]	122		
[illegible]	48	[illegible]	48				
[illegible]	80	[illegible]	1				

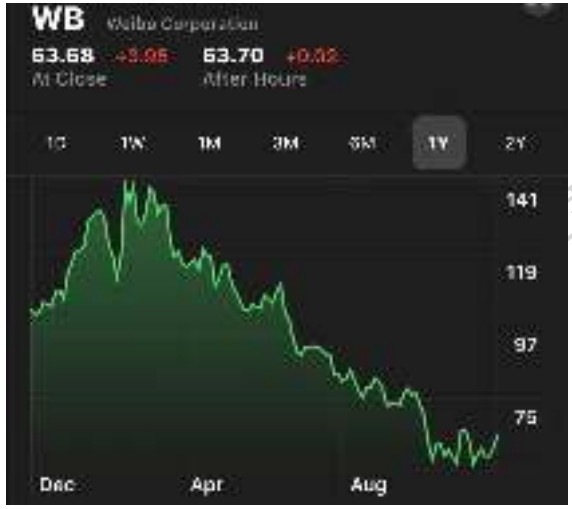
$$\mathbf{x} = [x_1, x_2, \ldots, x_i, \ldots, x_d]^T$$



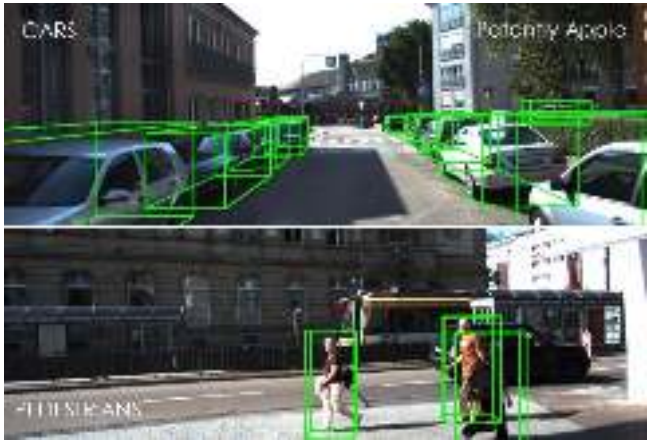
机器学习基本任务

回归 (Regression)

- 根据输入预测输出一个连续数值
- 机器学习算法的目标是产生一个函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}, \quad y = f(\boldsymbol{x})$
- 常见的回归任务：股价预测、年龄估计、距离估计、位置计算



预测：11月份的股票价格是多少？

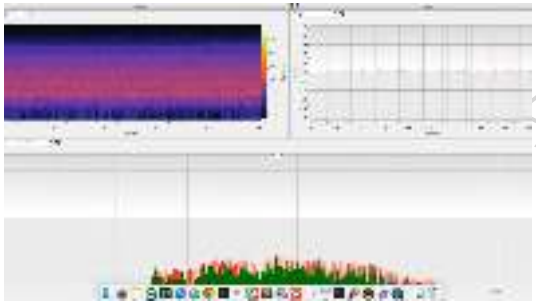


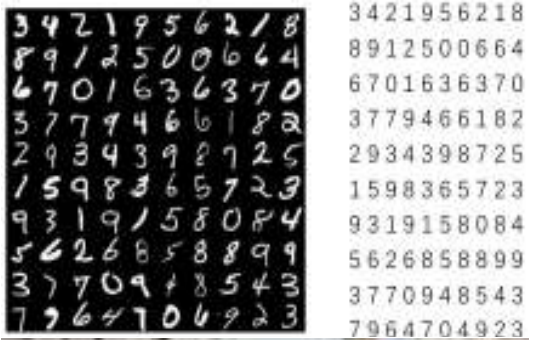
3D Object Detection

机器学习基本任务

转录 (Transcription)

- 输入为非结构化的某种数据或数据序列
- 输出为离散的文本符号
- 常见的转录任务：语音识别、字符识别

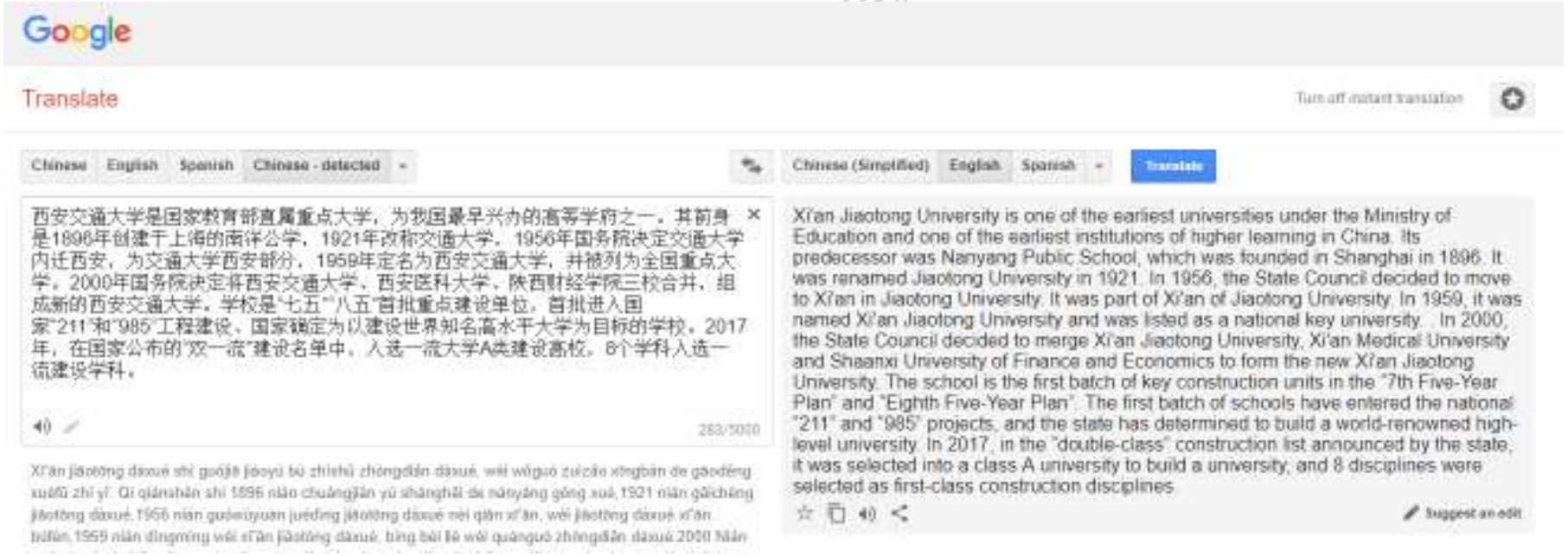




机器学习基本任务

机器翻译 (Machine translation)

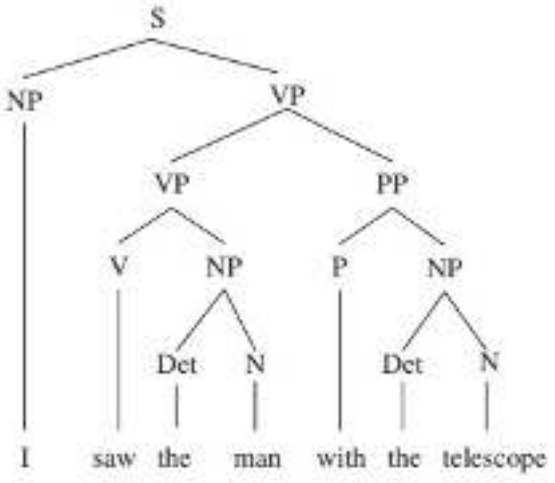
- 输入为某种语言符号序列
- 输出为另一种语言符号序列
- 常见的翻译任务：汉译英、英译汉、google翻译

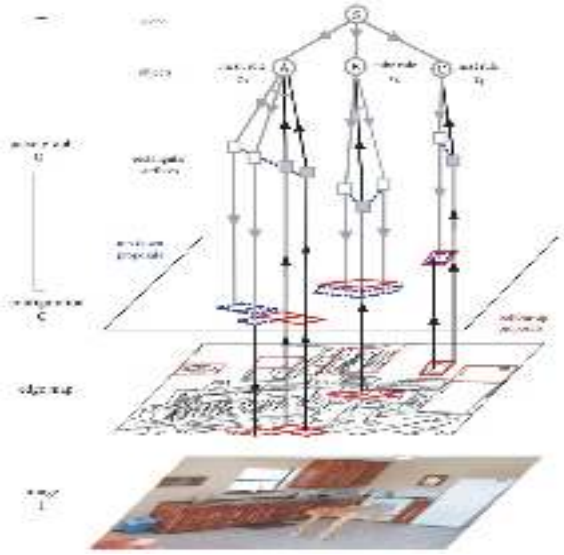


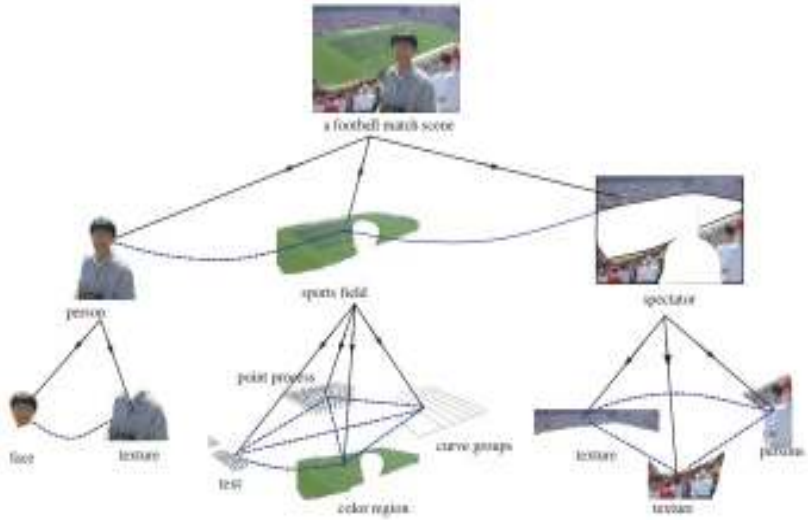
机器学习基本任务

结构输出 (Structured output)

- 输入可以为任何数据形式
- 输出为一个结构，即输出为多个变量、且变量之间具有紧密的交互关系
- 常见任务：语句解析、图像内容解析、场景综合理解





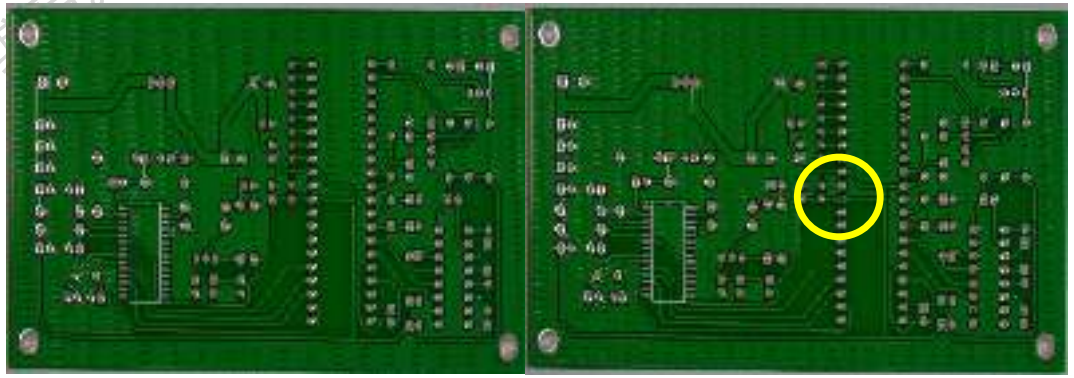


机器学习基本任务

异常检测 (Anomaly Detection)

- 在一组事件或对象中标记不正常或非典型的个体
- 常见任务：信用卡欺诈检测，产品缺陷检测

缺陷检测



异常检测

机器学习基本任务

合成和采样 (Synthesis and Sampling)

- 智能系统输出与训练数据相似的样本
- 常见任务：图像生成、数据合成

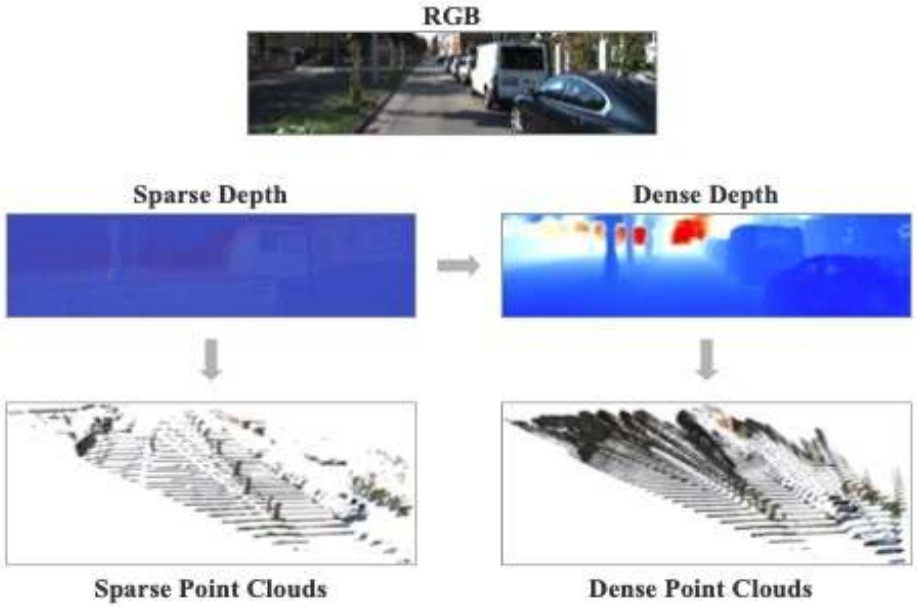
月明清影里，露冷绿樽前
赖有佳人意，依然似故年

我的爱人哪
快把光明的灯擎起来了
那里有美丽的天
问着村里的水流的声音
我的爱人在哪
因为我的红灯是这样的幻变
像是美丽的秘密
她是一个小孩子的歌唱
那时间的距离

机器学习基本任务

缺失值填补 (Imputation of missing values)

- 样本中某些元素缺失，机器学习算法填补这些缺失
- 常见任务：深度补全、图像修复

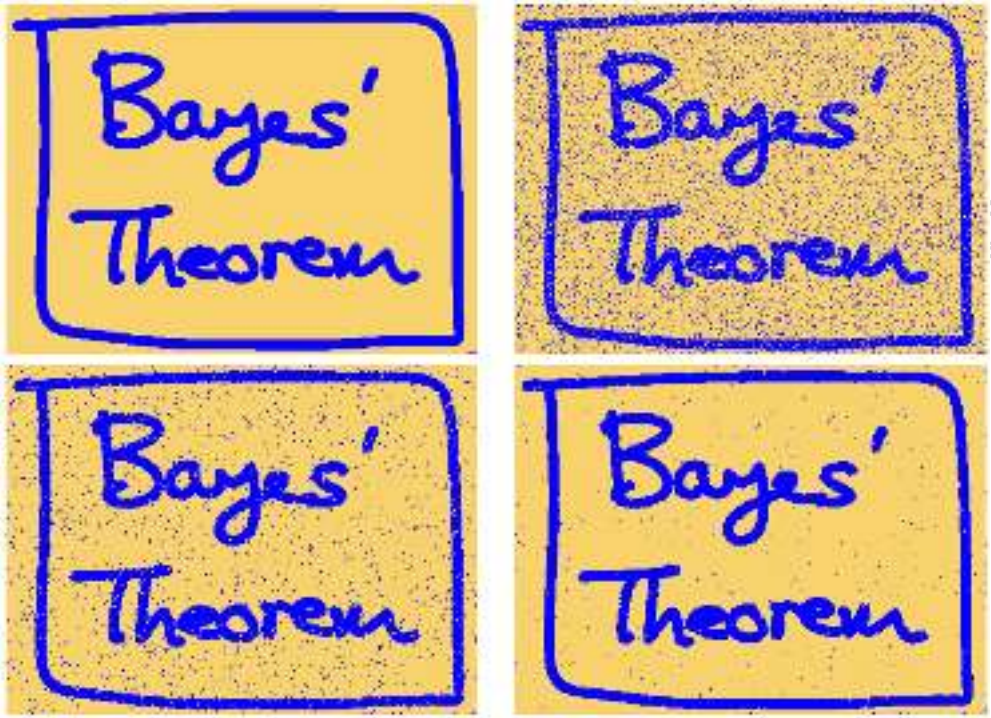


激光数据深度补全

图像修复

机器学习基本任务

- **去噪 (Denoising)**
  - 去除照片中的未知噪声过程，恢复原始图像

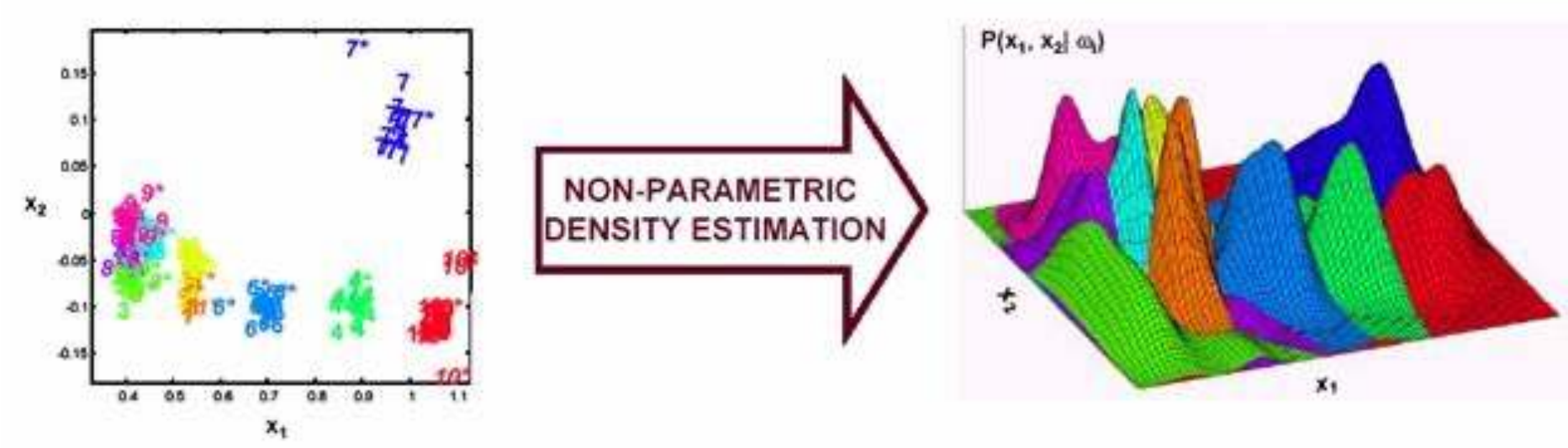


机器学习基本任务

密度估计 (Density estimation)

- 机器学习算法学习样本的概率密度函数

$$p_{\text{model}}: \mathbb{R}^n \rightarrow \mathbb{R}$$



性能度量 P

- **性能度量用来描述机器学习算法的能力，与具体任务相关**

  - 分类任务：正确率 Accuracy 0-1 loss
  - 回归任务：均方误差(mean-square error, MSE)，$E_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{data}}}\|\boldsymbol{y}-f(\boldsymbol{x})\|^2$
  - 概率估计：K-L Divergence，average log-probability
  - 不同任务有不同的性能度量，但不一定能精确定义其性能度量

经验 *E*

- 经验是人们知识积累的抽象
- 在机器学习算法中，经验就是数据集(dataset)
- 数据集是样本(example, sample)的集合；样本也叫数据点(data point)

Dataset Order	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	I. setosa
2	4.9	3.0	1.4	0.2	I. setosa
3	4.7	3.2	1.3	0.2	I. setosa
4	4.6	3.1	1.5	0.2	I. setosa
5	5.0	3.6	1.4	0.2	I. setosa
6	5.4	3.9	1.7	0.4	I. setosa
7	4.6	3.4	1.4	0.3	I. setosa
8	5.0	3.4	1.5	0.2	I. setosa
9	4.4	2.9	1.4	0.2	I. setosa
10	4.9	3.1	1.5	0.1	I. setosa
11	5.4	3.7	1.5	0.2	I. setosa
12	4.8	3.4	1.6	0.2	I. setosa
13	4.8	3.0	1.4	0.1	I. setosa
14	4.3	3.0	1.1	0.1	I. setosa
15	5.8	4.0	1.2	0.2	I. setosa
16	5.7	4.4	1.5	0.4	I. setosa
17	5.4	3.9	1.3	0.4	I. setosa
18	5.1	3.5	1.4	0.3	I. setosa
19	5.7	3.8	1.7	0.3	I. setosa
20	5.1	3.8	1.5	0.3	I. setosa
21	5.4	3.4	1.7	0.2	I. setosa

Iris dataset , *Fisher, 1936*

ImageNet , *Jia Deng, 2009*

机器学习与统计学习

Machine Learning VS. Statistical Learning

- 两者没有本质区别
- 如果说有“很小”的区别，则存在于以下几个方面
  - Statistical Learning主要来自统计学(Statistics)研究者的说法， Machine Learning主要来自AI/CS研究者的说法
  - Machine Learning 属于AI/CS的分支， Statistical Learning属于Statistics分支
  - Machine Learning 的主要目的是构建模型在新的数据上做预测，而Statistical Learning的主要目标是构建模型理解、解释和分析数据

机器学习算法的分类

- 基于学习策略，机器学习可以分为以下五类

“死记硬背”，信息存储与检索

“老师”教“学生”学习

类比源域与目标域相似性

构造解释，推广解释

通过实例归纳一般描述

机器学习算法的分类

- 基于学习方式，机器学习算法可以分为以下三类

机器学习

有监督学习

学习数据中有监督信息

- 数据有标记
- 直接反馈
- 预测输出

无监督学习

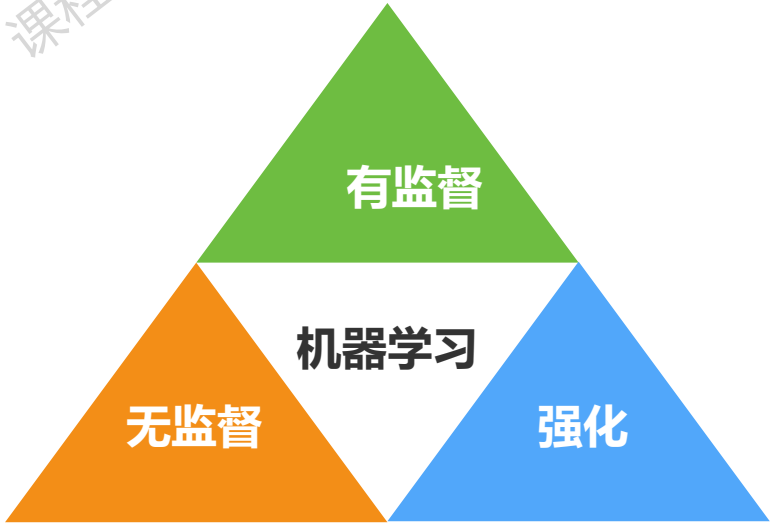
学习数据中无监督信息

- 数据无标记
- 没有反馈
- 发现隐藏规律

强化学习

以反馈(奖/惩)指导行为

- 决策过程
- 奖励系统
- 学习动作序列



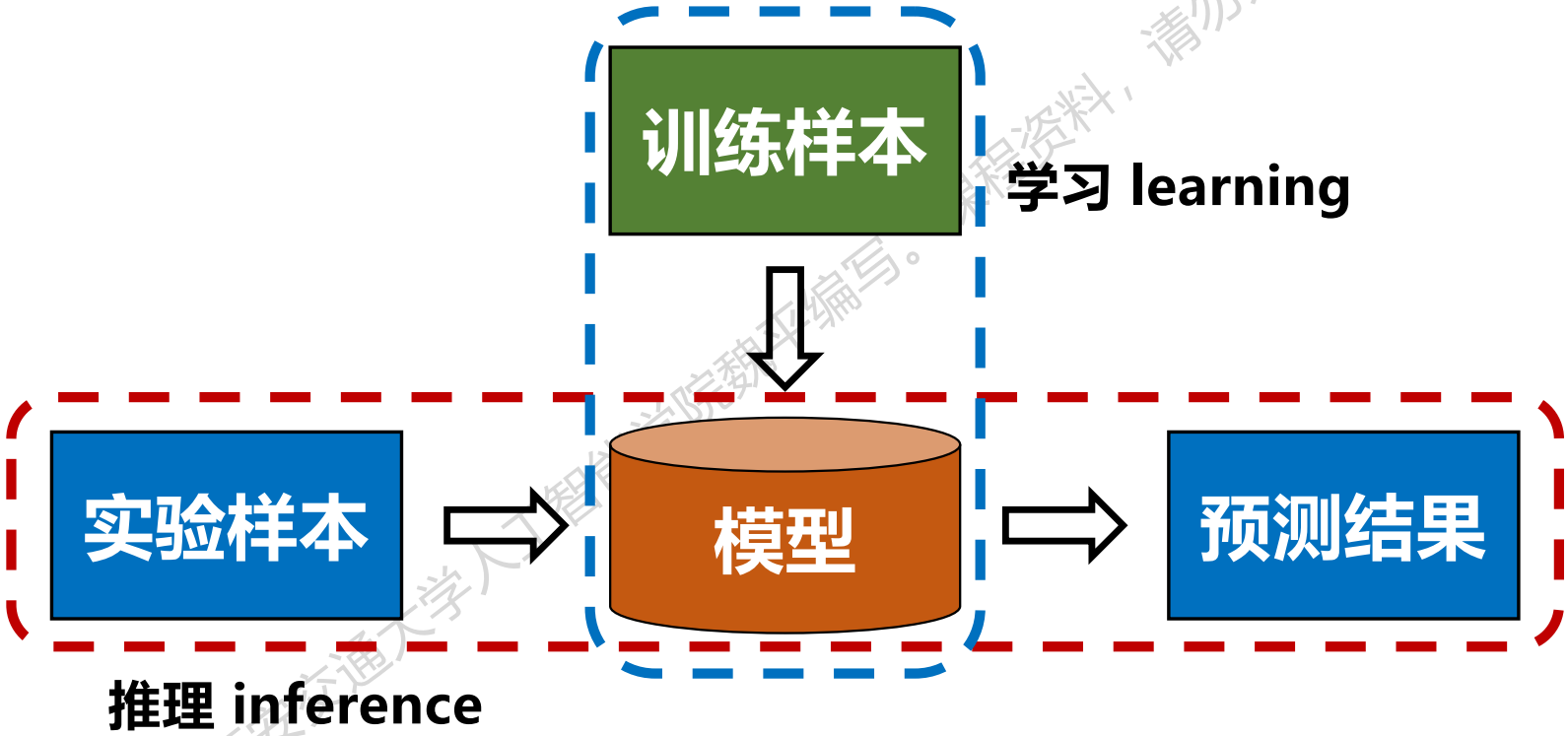
机器学习算法的分类

- 基于监督方式，监督类机器学习算法可以细分为

监督类机器学习

有监督学习 Supervised Learning	有标签
无监督学习 Unsupervised Learning	无标签
半监督学习 Semi-supervised Learning	部分有标签
弱监督学习 Weakly-supervised Learning	有低质量标签
自监督学习 Self-supervised Learning	没有外部标签，利用内部标签

有监督机器学习的一般流程



弱监督学习 (Weakly Supervised)

- 数据标注过程成本太高，很多任务很难获得全部真值标签这样的强监督信息,无监督学习由于学习过程困难，发展较慢
- 弱监督学习：通过较弱的或较低质量的监督信息来构建模型
- 弱监督学习三种典型类型

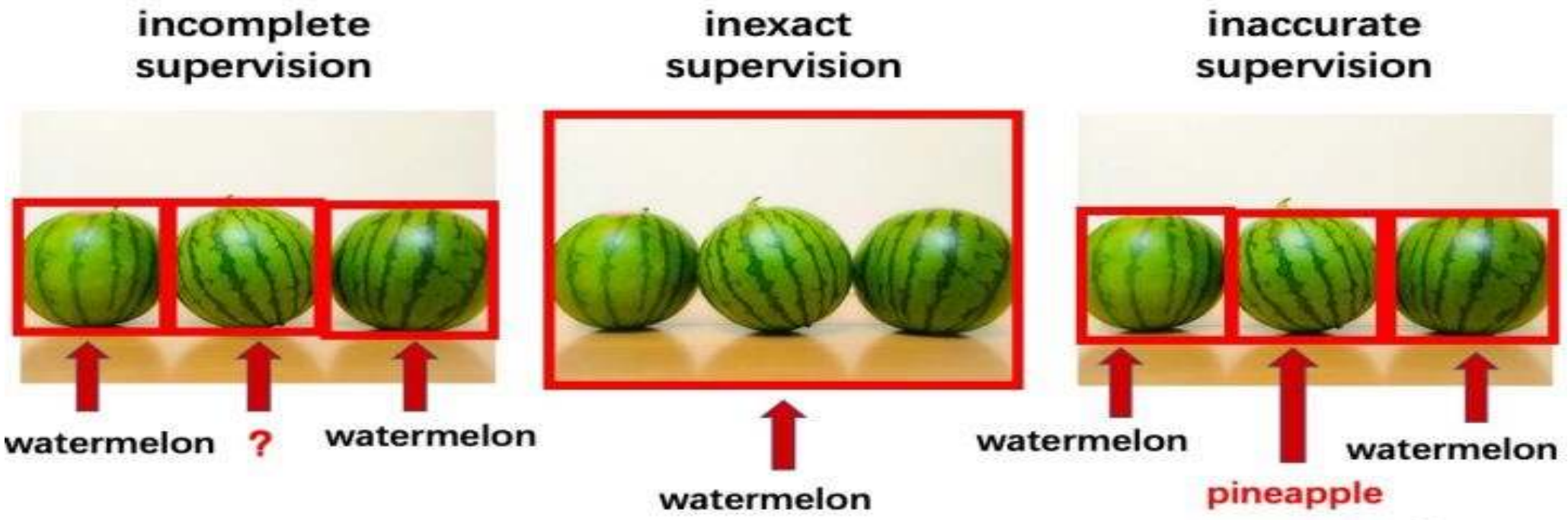
① 不完全监督（Incomplete supervision）

② 不确切监督（Inexact supervision）

③ 不精确监督（Inaccurate supervision）

弱监督学习 (Weakly Supervised)

- 不完全监督：训练数据中只有一部分数据被给了标签
- 不确切监督：训练数据只给出了粗粒度标签
- 不精确监督：给出的标签不总是正确的



CONTENTS

- 什么是人工智能
- 为什么需要机器学习
- 什么是机器学习
- 机器学习的历史
- 机器学习基础概念

机器学习的发展阶段

逻辑推理 Logic Reasoning	知识工程 Knowledge Engineering	统计学习 Statistical Learning	深度学习 Deep Learning
1940 ~ 1970	1970 ~ 1990	1990 ~ 2010	2010 ~ now
1943 Foundation of Neural Networks	1973 British AI Report	1990 Boosting	2012 AlexNet
1950 Turing Test Introduced	1977 EM	1995 AdaBoost, SVM	2013 Deep Reinforcement Learning
1956 Logical Theorist	1980 First ML International Conference	1996 DBSCAN	2013 VAE
1958 Birth of the Perceptron	1982 Hopfield Network	1997 LSTM	2014 GAN
1959 Machine Learning Term	1981 Boltzmann Machine	1997 Deep Blue	2015 ResNet
1959 Pandemonium Model	1985 Bayesian Network	1998 MNIST database	2017 Transformer
1969 Backpropagation Learning Algorithm	1986 Restricted Boltzmann Machine	2000 Neural Probabilistic Language Model	2018 GPT
	1989 Q-learning	2001 CRF, Random Forest	2020 Diffusion Model
	1989 Convolutional Neural Networks (CNNs)	2006 Deep Larning Term	2022 AlphaTensor
		2009 ImageNet	2024 Copilot
			2025 DeepSeek

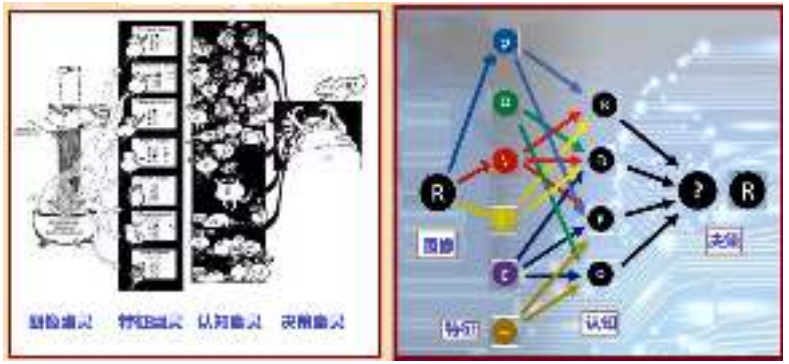
机器学习的诞生

机器学习的产生与人工智能的进程密不可分

- 1950年图灵在论文“Computing machinery and intelligence”中提出了**学习机器(Learning Machine)**名词
- 1959年Arthur Samuel在论文“Some studies in machine learning using the game of checkers”中提出了**机器学习（Machine Learning）**名词
- 1959年Oliver Selfridge在里程碑工作“Pandemonium: A Paradigm for Learning”中提出了“鬼域模型”模型，通过发现事件中的模式来提高模型的性能

Arthur Samuel (1901–1990)
Bell, IBM, Stanford

Oliver Selfridge（1926-2008）
MIT, Lincoln Lab.



鬼域模型示意图

第一阶段（1940-1970）：逻辑推理

- 假设：只要赋予机器逻辑推理能力，机器就具有智能

  - **代表性成果**

    - **Logic Theorist：** 赫伯特·西蒙（Herbert Simon）等人于1956年提出，第一个能够自动进行逻辑推理和定理证明的程序
    - **通用问题求解器（General Problem Solver, GPS）：** 是由艾伦·纽厄尔（Allen Newell）和赫伯特·西蒙（Herbert Simon）于1957年提出，模拟人类问题解决的过程
    - **归结法（Resolution）：** 由约翰·罗宾逊（John Alan Robinson）于1965年提出的一种自动推理方法，主要用于一阶逻辑（谓词逻辑）中的定理证明

  - **缺陷**

仅有逻辑推理能力是远远实现不了人工智能

Herbert Simon

Allen Newell

第二阶段（1970-1990）：知识工程

- **Knowledge Engineering**

  基本假设：要使机器具有智能，就必须设法使机器拥有知识

- **代表性成果：**专家系统
  - **DENDRAL：**由斯坦福大学的爱德华·费根鲍姆（Edward Feigenbaum）、布鲁斯·布坎南（Bruce Buchanan）、约书亚·莱德伯格（Joshua Lederberg）等人开发，是第一个成功的专家系统，用计算机模拟化学家的推理过程
  - **MYCIN：**由斯坦福大学的爱德华·肖特利夫（Edward Shortliffe）和布鲁斯·布坎南（Bruce Buchanan）等人开发，帮助医生诊断细菌感染，推荐合适抗生素

Edward Feigenbaum

Joshua Lederberg

- **缺陷**

由人来把知识总结出来再教给计算机非常困难

第三阶段（1990-2010）：统计学习

- **Statistical Learning (1990-2010)**

基本假设：数据背后存在的规律或模式可以通过统计模型来捕捉

Michael I. Jordan

- **归纳学习**：从具体的训练样本数据中归纳出一般的规则或模型，从而对未见过的数据进行预测或分类
- **强化学习**：核心思想是在问题求解和规划过程中，通过学习获取策
- **类比学习**：将已知的知识或经验与新的情况进行比较，从而学习到新知识或解决问题的方法

Vladimir Vapnik

- **代表性成果**：贝叶斯网络、支持向量机、集成学习、聚类、PCA

可解释性强，属“浅层”模型

第四阶段（2010-至今）：深度学习

- **Deep Learning (2010-今)**

基本假设：让机器自己学

- **深度学习的崛起（2010-2014）：**以 AlexNet 和 GAN 为代表，计算机视觉和生成模型取得重要突破
- **深度学习的爆发（2015-2018）：**以 ResNet和Transformer 为代表，深度学习在多个领域广泛应用
- **深度学习的扩展与深化（2019-至今）：**以 ChatGPT 、扩散模型和Deepseek为代表，模型规模和应用领域进一步扩展

Geoffrey Hinton

Yann LeCun

Yoshua Bengio

- **代表性成果**
  - **计算机视觉：**CNN、R-CNN、YOLO、U-Net、Mask R-CNN
  - **自然语言处理：**RNN、LSTM、Transformer、BERT、GPT、Word2Vec

IAIR Est. 1986 Institute of Artificial Intelligence and Robotics

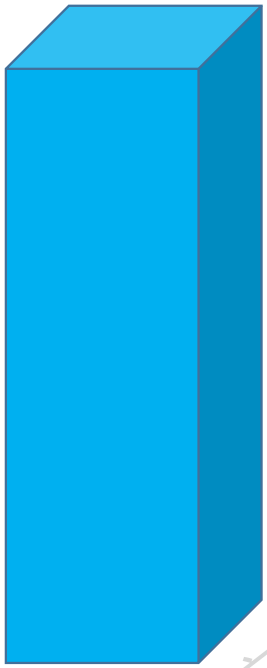
人工智能学院 College of Artificial Intelligence, XJTU

CONTENTS

- 什么是人工智能
- 为什么需要机器学习
- 什么是机器学习
- 机器学习的历史
- 机器学习基础概念

评价方法

- **数据集**



$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

样本数据：$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$

样本标签：y_i

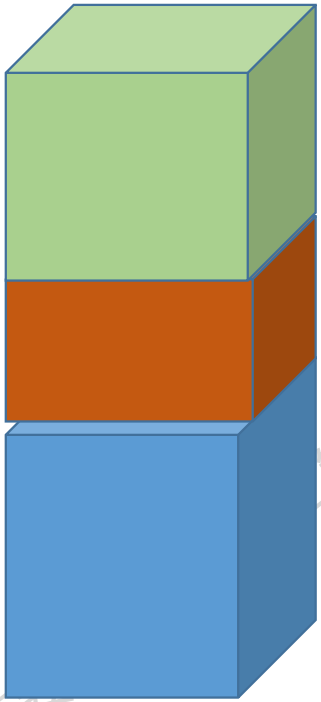
评价方法

- **留出法划分**

测试集- T
test set

验证集- V
validation set

训练集- S
training set



测试误差
test / generalization error

验证误差
validation error

训练误差
training error

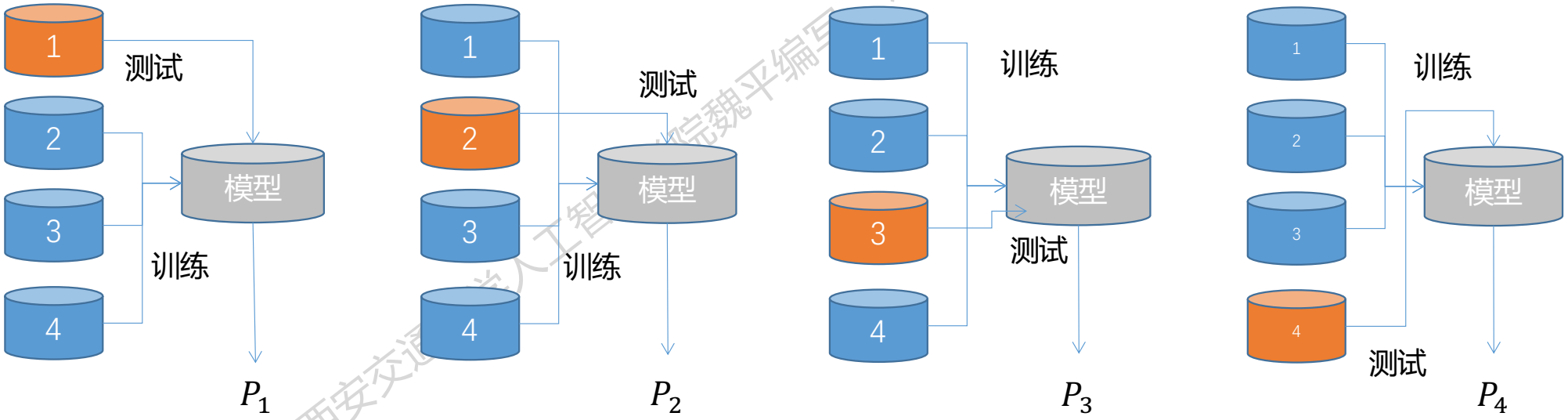
$$D = T \cup V \cup S$$

$$T \cap V = \emptyset$$
$$V \cap S = \emptyset$$
$$T \cap S = \emptyset$$

评价方法

- 交叉验证 (cross-validation)

  k折(k-fold)交叉验证将样本集均匀的分成k份互斥子集$D = D_1 \cup \cdots \cup D_k$，$D_i \cap D_j = \emptyset$，轮流用其中的$k-1$份作为训练集，剩下的1份作为测试集，用$k$次性能指标的均值作为最后的指标



$$P = \frac{P_1 + P_2 + P_3 + P_4}{4}$$

容量、过拟合、欠拟合

泛化能力 （generalization）

- 一个机器学习算法在先前未观测到的输入数据上执行某种任务的能力称作该算法的泛化能力
- 机器学习算法追求的最终目标是在测试集上有更小的误差（更好的性能）！

独立同分布假设（ i.i.d ）

- 理论上独立同分布的训练集与测试集具有相同的误差
- 现实中测试误差大于训练误差

容量、过拟合、欠拟合

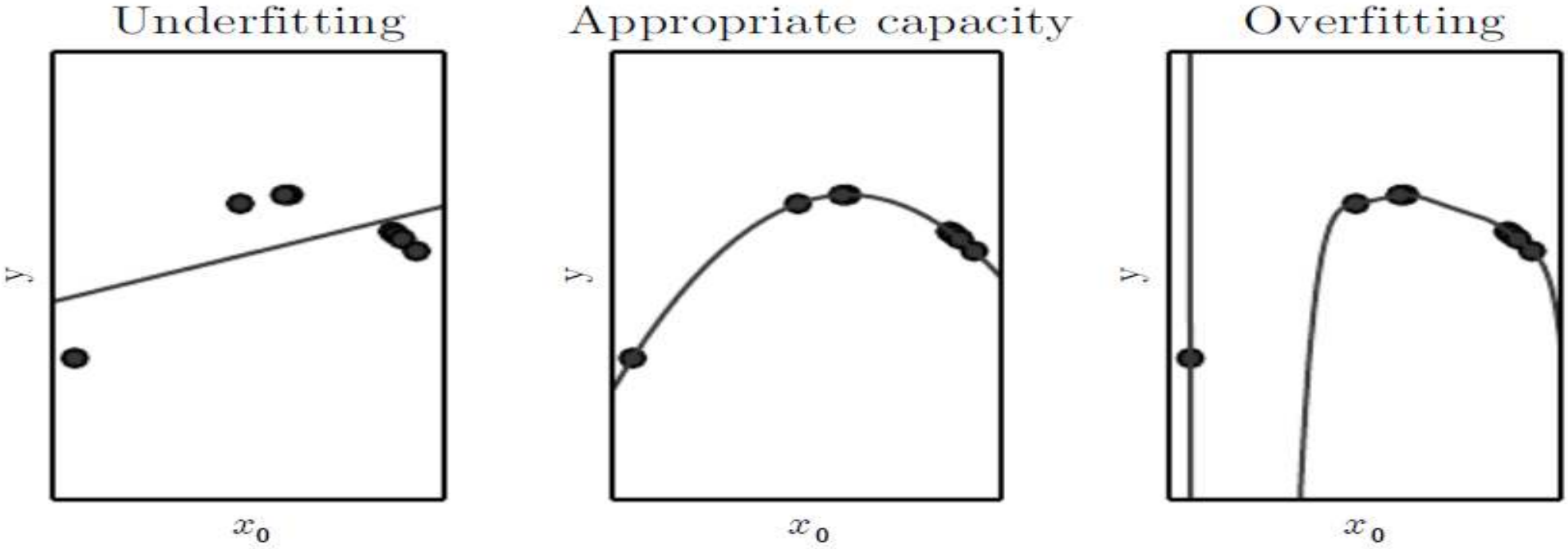
判断机器学习算法效果是否好坏的两个因素

- 训练误差小
- 训练误差和测试误差之间的差距小

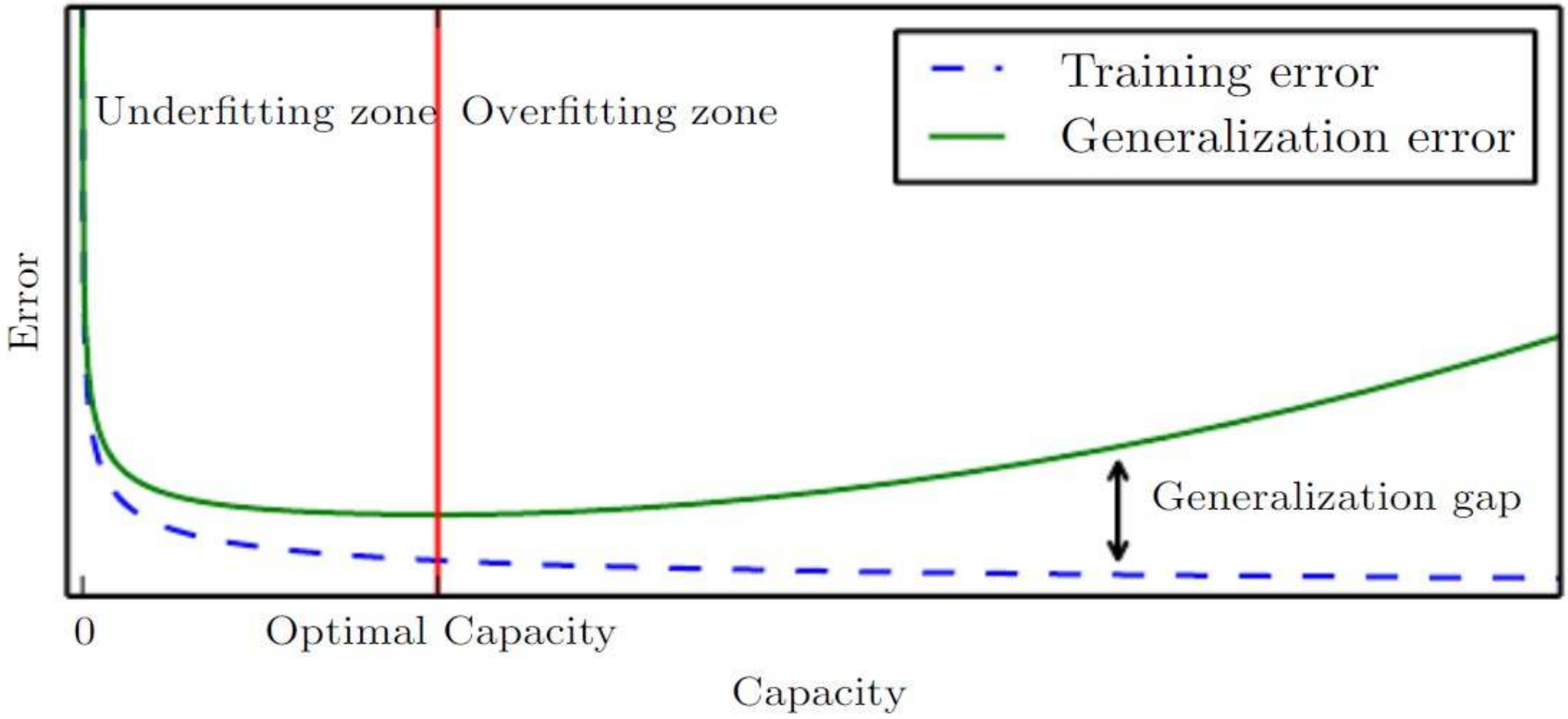
容量、过拟合、欠拟合

模型的容量(Capacity)：拟合函数的能力

- 容量低的模型很难拟合训练集
- 容量高的模型会记住不适用于测试集的训练集的性质



容量、过拟合、欠拟合



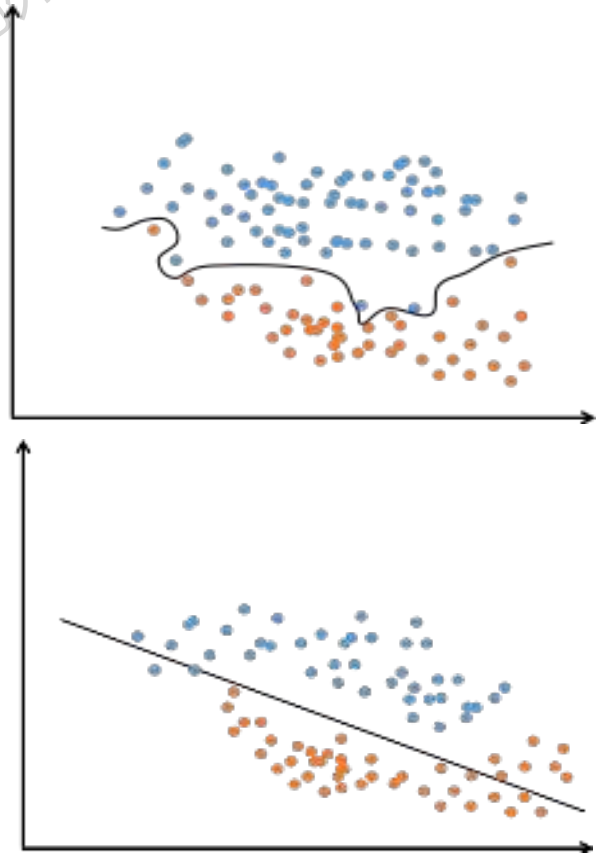
过拟合与欠拟合分析

- 引起过拟合的可能原因有下面几种情况

① 模型本身过于复杂，拟合了训练样本集中的噪声。此时需要选用更简单的模型，或者对模型进行裁剪

② 训练样本太少或者缺乏代表性。此时需要增加样本数，或者增加样本的多样性

③ 训练样本噪声的干扰，导致模型拟合了这些噪声，这时需要剔除噪声数据或者改用对噪声不敏感的模型



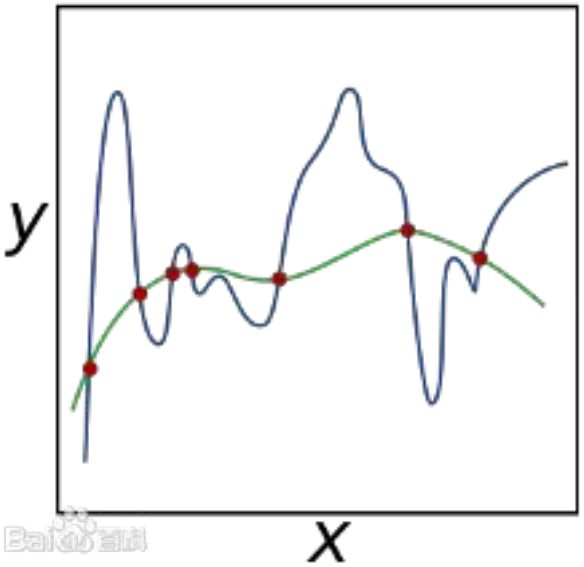
- 导致欠拟合的可能原因有模型简单，特征数太少无法正确的建立映射关系

正则化 (Regularization)

- 为了防止过拟合，可以为损失函数加上一个惩罚项，对复杂的模型进行惩罚，强制让模型的参数值尽可能小以使得模型更简单，以提高泛化性能，这种方法称为正则化技术

- 常用的是L^2正则化，它是L^2范数的平方，加上正则化项后的目标函数为

$$J(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; X, Y) + \frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$$



信息与熵 (Entropy)

- 如何量化描述整个随机系统（分布）的不确定性总量——熵，描述系统的无序(混乱)程度-熵越大，随机系统的混乱程度越大
- 宇宙中的事物都有自发变得更混乱的倾向，也就是说熵会不断增加，这就是熵增原理
- 定义于一个概率分布之上，对概率分布的随机性程度进行度量，反映了一组数据所包含信息量的大小——熵越大，信息量越大；熵越小，信息量越小
- 熵是对随机变量取每个值的信息量的数学期望
- 对于离散型随机变量，熵定义为

$$H(p) = \mathrm{E}_x[-\ln p(x)] = -\sum_{i=1}^{n} p_i \ln p_i \qquad p_i = p(x_i)$$

- 对于连续型随机变量，熵通过积分定义，也称为微分熵

$$H(p) = -\int_{-\infty}^{+\infty} p(x) \ln p(x) dx$$

交叉熵 (Cross-Entropy)

- 定义于两个概率分布之上，反映两个概率分布的差异程度
- 对于离散型随机变量x，$p(x)$和$q(x)$是两个概率分布的概率质量函数，他们的交叉熵定义为

$$H(p,q) = \mathrm{E}_p[-\ln q] = -\sum_x p(x)\ln q\,(x)$$

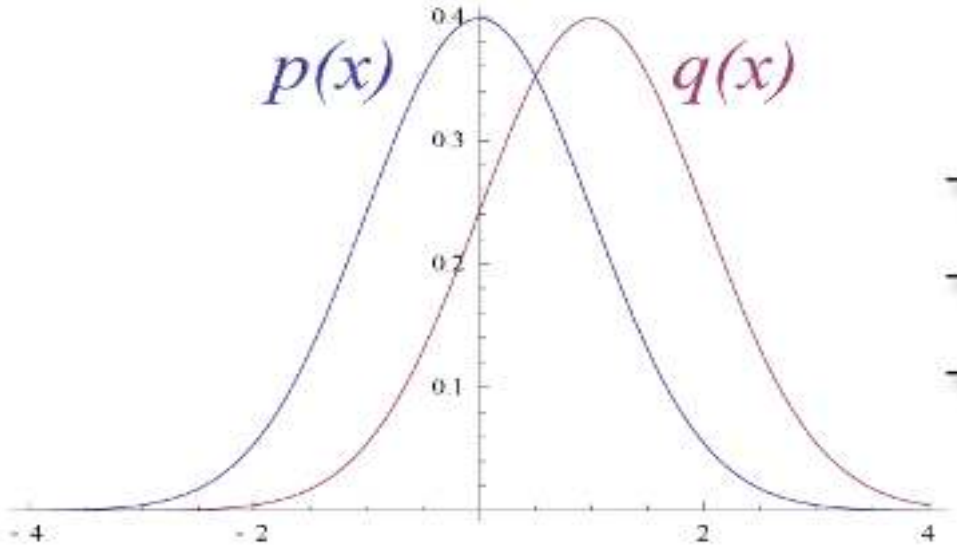
- 对于连续型概率分布，交叉熵定义为

$$H(p,q) = \mathrm{E}_p[-\ln q] = -\int_x p(x)\ln q\,(x)\mathrm{d}x$$

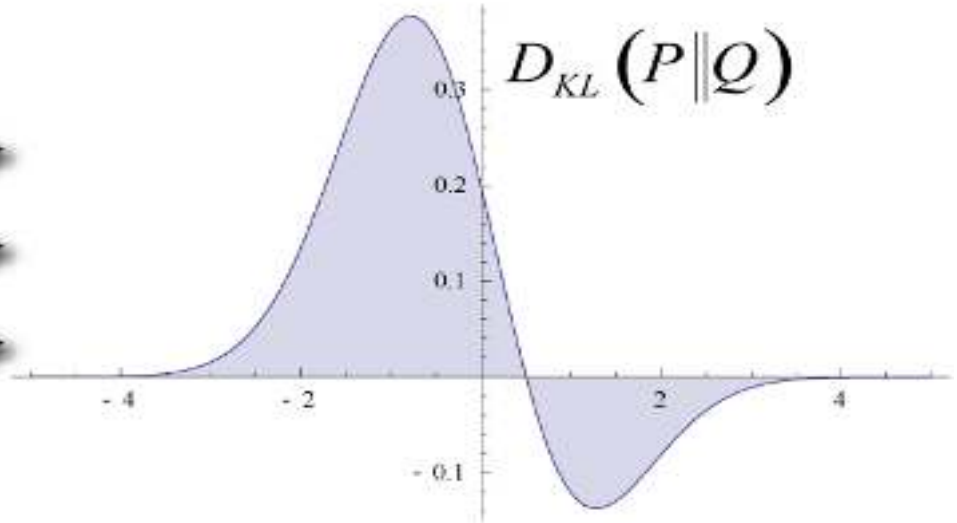
- 如果两个概率分布完全相等，则交叉熵退化成熵
- 交叉熵不具有对称性，$H(p,q) \neq H(q,p)$

Kullback-Leibler散度的概念

- 也称相对熵、信息增益，用于衡量两个概率分布之间的差距
- 其值越大则说明两个概率分布的差距越大；当两个分布完全相等时KL散度值为0



Original Gaussian PDF's



KL Area to be Integrated

Kullback-Leibler散度的定义

- 对于两个离散型随机概率分布，它们之间的KL散度定义为

$$D_{\mathrm{KL}}(p\|q) = \sum_{x} p(x) \ln \frac{p(x)}{q(x)}$$

- 对于两个连续型概率分布，它们之间的KL散度定义为

$$D_{\mathrm{KL}}(p\|q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

- KL散度不具有对称性

Kullback-Leibler散度的性质

- 两个概率分布相等时KL散度为0，若$p(x)=q(x)$，则$D_{\mathrm{KL}}(p\|q)=0$
  - 证明：$D_{\mathrm{KL}}(p\|q)=\sum_x p(x)\ln\frac{p(x)}{q(x)}=\sum_x p(x)\ln 1=\sum_x p(x)\times 0=0$
- KL散度是非负的，即$D_{\mathrm{KL}}(p\|q)\geq 0$
  - 证明：$$D_{\mathrm{KL}}(p\|q)=-\sum_x p(x)\ln\frac{q(x)}{p(x)}$$
    $$\geq -\sum_x p(x)\left(\frac{q(x)}{p(x)}-1\right)$$
    $$=-\sum_x q(x)+\sum_x p(x)=0$$

不等式：

$\forall x>0,\quad \ln x\leq x-1$

Kullback-Leibler散度与交叉熵的联系

- KL散度 $$D_{\mathrm{KL}}(p\|q)=\int_{-\infty}^{+\infty}p(x)\ln\frac{p(x)}{q(x)}dx$$
- 交叉熵 $$H(p,q)=-\int_{x}p(x)\ln q(x)\mathrm{d}x$$

$$D_{\mathrm{KL}}(p\|q)=\int_{-\infty}^{+\infty}p(x)\ln\frac{p(x)}{q(x)}dx=\int_{-\infty}^{+\infty}p(x)\ln p(x)\,dx-\int_{-\infty}^{+\infty}p(x)\ln q(x)\,dx$$

$$=-\int_{-\infty}^{+\infty}p(x)\ln q(x)\,dx-\left(-\int_{-\infty}^{+\infty}p(x)\ln p(x)\,dx\right)$$

$$=\underbrace{H(p,q)-H(p,p)}_{\text{信息增益}}$$

$H(p,p)$：随机系统真实的信息量

$H(p,q)$：用q代替p描述系统的信息量

最大似然估计(MLE)

最大似然估计 Maximum Likelihood Estimation

模型学习中参数估计一般准则的最常用的一种

$\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)})$ 是m个独立同分布的样本，模型分布为$p(\boldsymbol{x}; \boldsymbol{\theta})$，$\hat{p}_{\text{data}}$为真实分布。参数$\boldsymbol{\theta}$的最大似然估计定义为：

$$\boldsymbol{\theta}_{\mathbf{ML}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}; \boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

最大似然估计(MLE)

$$\boldsymbol{\theta}_{\mathbf{ML}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Maximize the likelihood

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \ln p(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Maximize the log-likelihood

$$\boldsymbol{\theta}_{\mathbf{ML}} = \arg\max_{\boldsymbol{\theta}} E_{\boldsymbol{x}\sim\hat{p}_{\text{data}}} \ln p(\boldsymbol{x}; \boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{\mathbf{ML}} = \arg\min_{\boldsymbol{\theta}} -E_{\boldsymbol{x}\sim\hat{p}_{\text{data}}} \ln p(\boldsymbol{x}; \boldsymbol{\theta})$$

Minimizing the negative log-likelihood

Minimizing the cross entropy

$$\boldsymbol{D}_{\mathbf{KL}}(\hat{p}_{\text{data}} || p) = E_{\boldsymbol{x}\sim\hat{p}_{\text{data}}} [\ln \hat{p}_{\text{data}}(\boldsymbol{x}) - \ln p(\boldsymbol{x})]$$

Minimizing the KL divergence

条件最大似然估计

- **估计条件分布** $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$

$\boldsymbol{X}=(\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(m)})$ 是输入数据，$\boldsymbol{Y}=(\boldsymbol{y}^{(1)},\ldots,\boldsymbol{y}^{(m)})$是对应的目标输出，则条件似然估计为：

$$\boldsymbol{\theta}_{\mathbf{ML}}=\arg\max_{\boldsymbol{\theta}}\ p(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{\mathbf{ML}}=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^{m}\ln p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

最大后验估计

- 最大后验估计(Maximum A Posteriori)是根据经验数据获得对难以观察的量的估计

$\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)})$ 是m个独立同分布的样本，模型分布为$p(\boldsymbol{x}; \boldsymbol{\theta})$，参数$\boldsymbol{\theta}$的最大后验估计为：

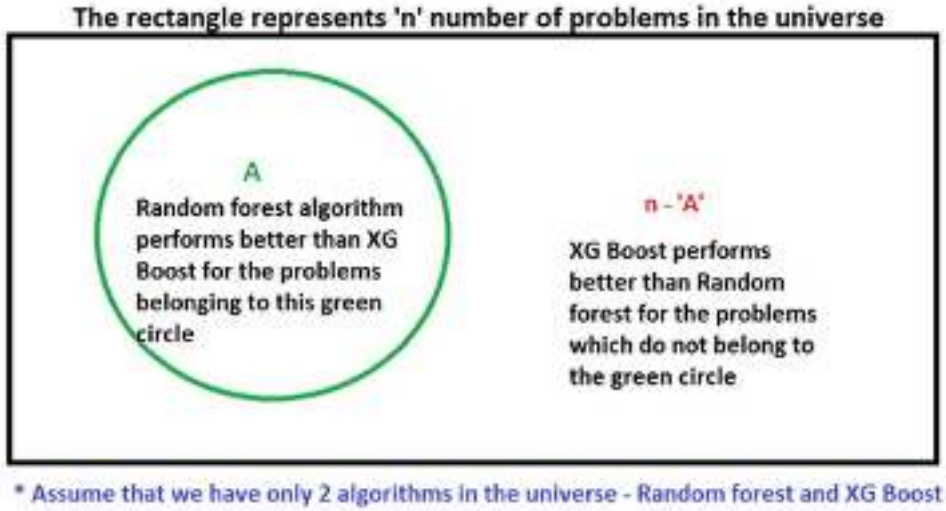
$$\boldsymbol{\theta}_{\mathbf{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{X})$$

$$= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \left(\log p(\boldsymbol{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right)$$

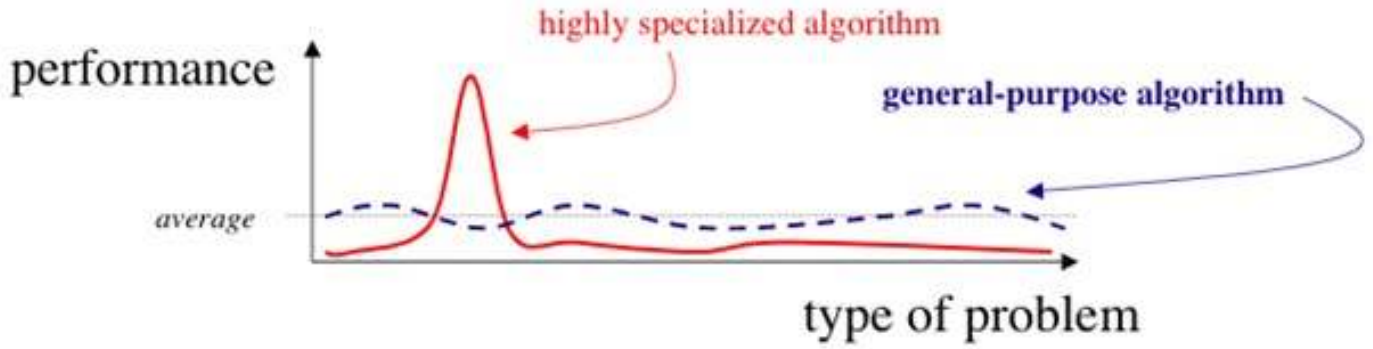
没有免费的午餐 — No Free Lunch Theorem (NFL)

- **NFL定理：对所有可能的目标函数求平均，所有学习算法的测试误差的期望值相同**
  — Wolpert，1996
- 任何模型在所有问题上的总体性能都是相同的，其总误差和模型本身是没有关系的；没有任何一个学习算法或者模型可以在任何任务上总是最好的



没有免费的午餐 — No Free Lunch Theorem (NFL)

- 机器学习研究的目标不是找一个通用学习算法或者绝对最好的学习算法，而是关注什么样的学习算法在我们关注的具体问题、具体数据上效果最好



先验知识很重要，具体问题具体分析：脱离问题的实际情况谈论模型优劣是没有意义的，只有让模型的特点和问题的特征相匹配，模型才能发挥最大的作用

奥卡姆剃刀定律 — Occam's Razor

CORE PRINCIPLES IN RESEARCH

OCCAM'S PROFESSOR

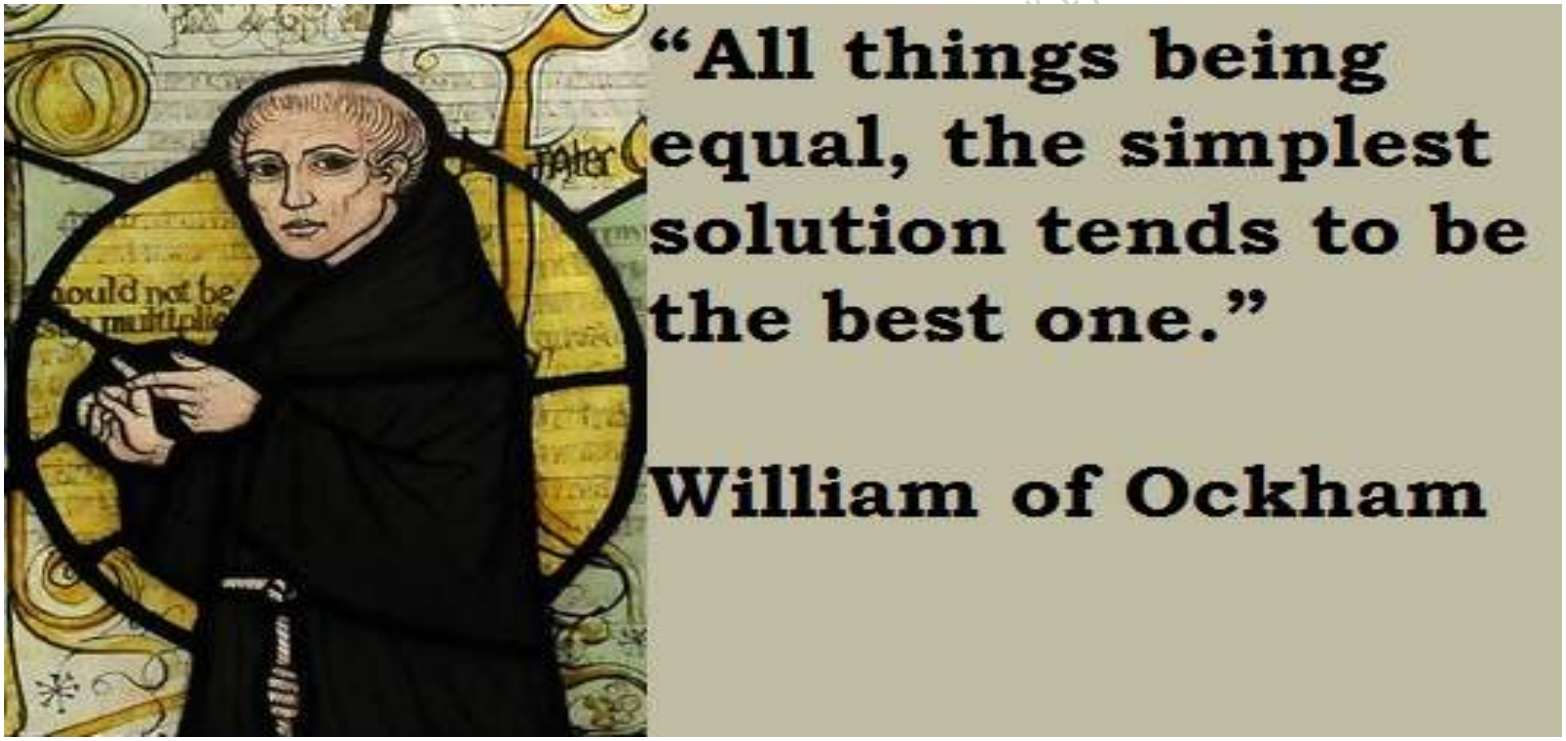
"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."

OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."

奥卡姆剃刀定律 — Occam's Razor

- “如无必要，勿增实体” — Entities should not be multiplied unnecessarily.



奥卡姆剃刀定律 — Occam's Razor

- 简单有效原理：如果有多个原理都能解释观测到的事实，那么应该使用简单的那个；需要假设越少的解释往往是越接近真相的解释
- 如果有多种模型都能够同等程度地符合同一个问题的观测，那就应该选择其中使用假设最少的，也就是最简单的模型；
- 当模型本身过于复杂时，特征和类别之间的关系中所有的细枝末节都被捕捉，主要的趋势反而没有得到应有的重视，这就会导致过拟合的发生

乱花渐欲迷人眼，浅草才能没马蹄

西安交通大學
XI'AN JIAOTONG UNIVERSITY

IAIR Est. 1986
Institute of Artificial Intelligence and Robotics

人工智能学院
College of Artificial Intelligence, XJTU

The End