第1章 绪论

王博：自动化（人工智能）学院
wangbo@hdu. edu. cn

章节目录

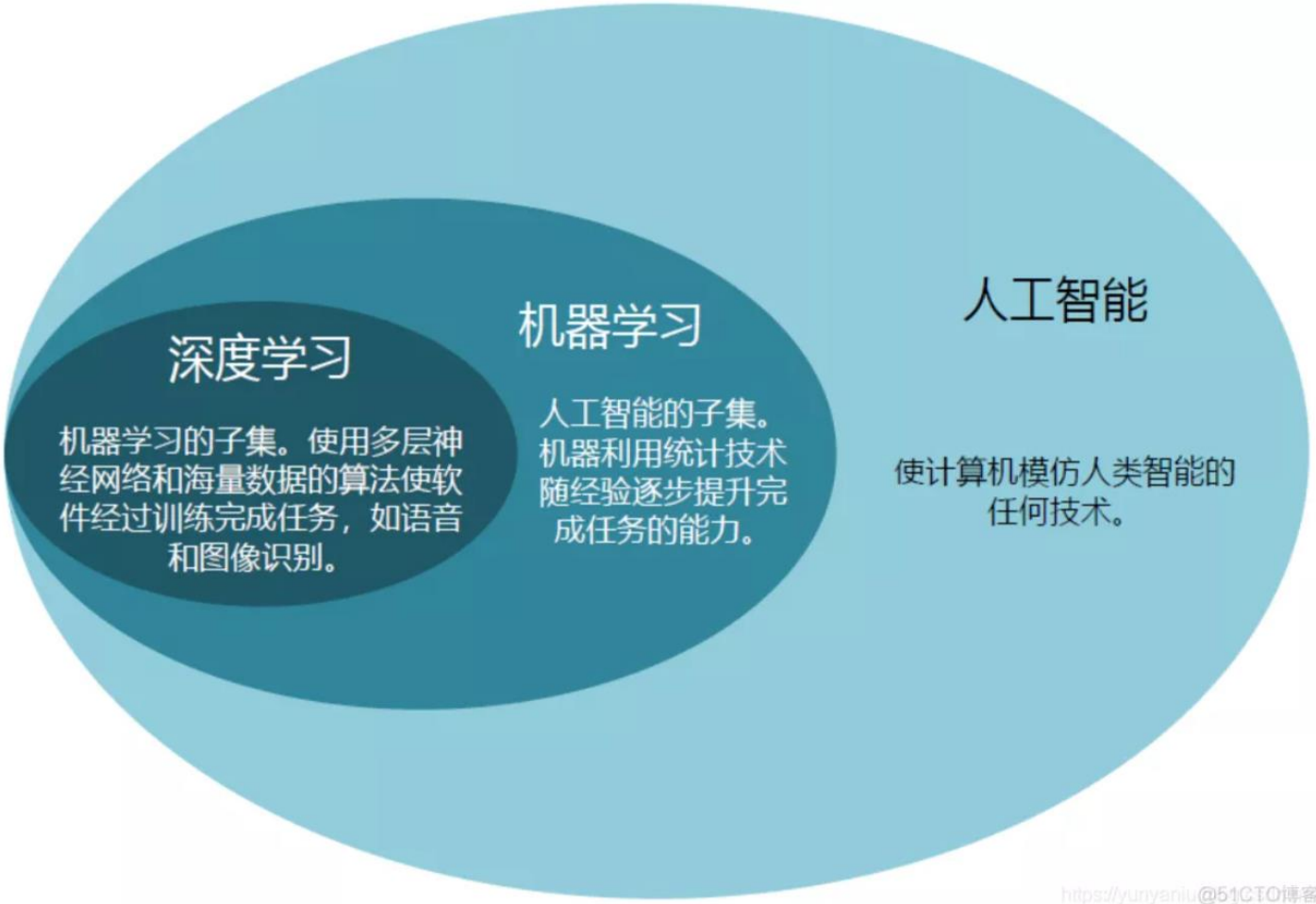
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状

章节目录

- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状

机器学习

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。

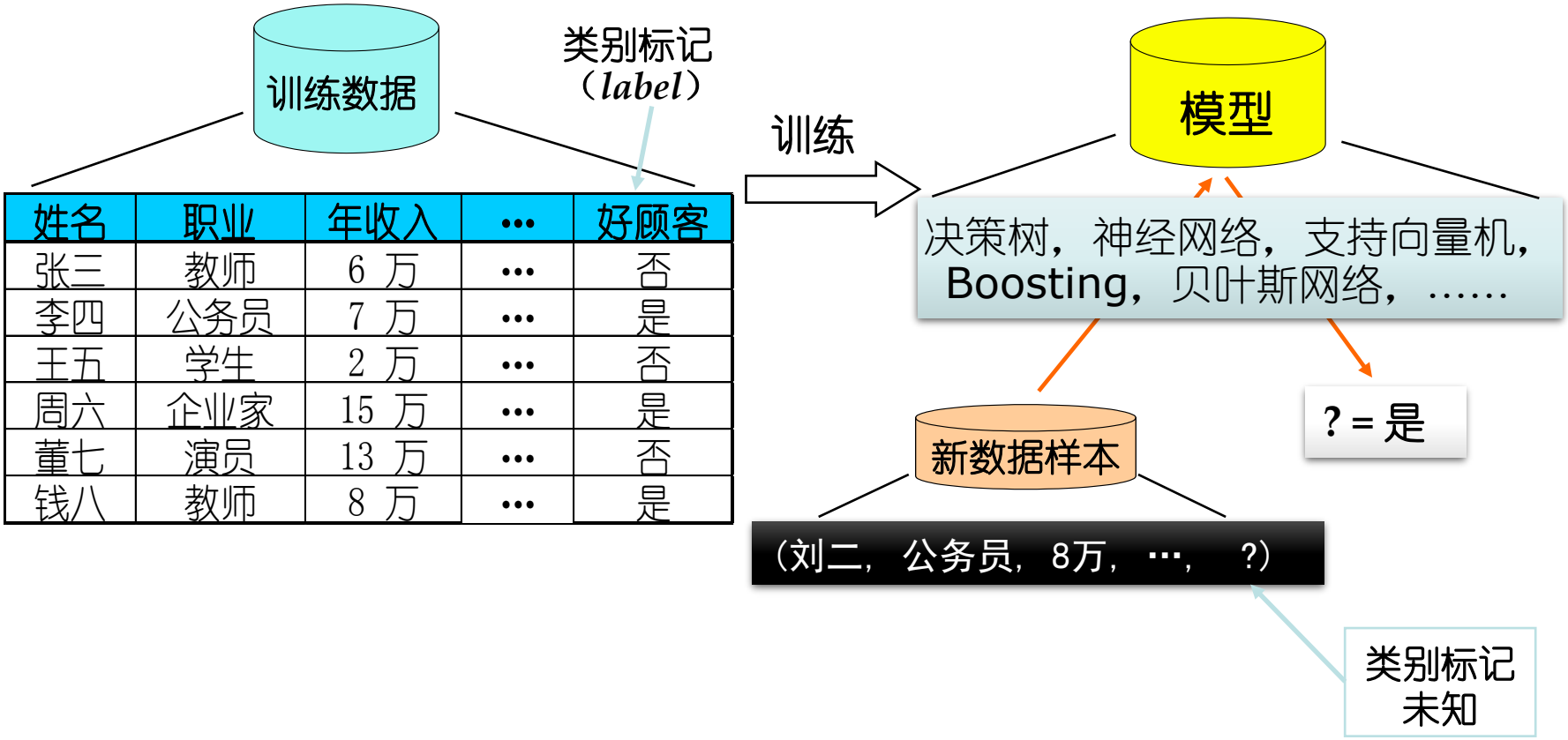


基本术语–数据

	特征				标记
	编号	色泽	根蒂	敲声	好瓜
训练集	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
测试集	1	青绿	蜷缩	沉闷	?

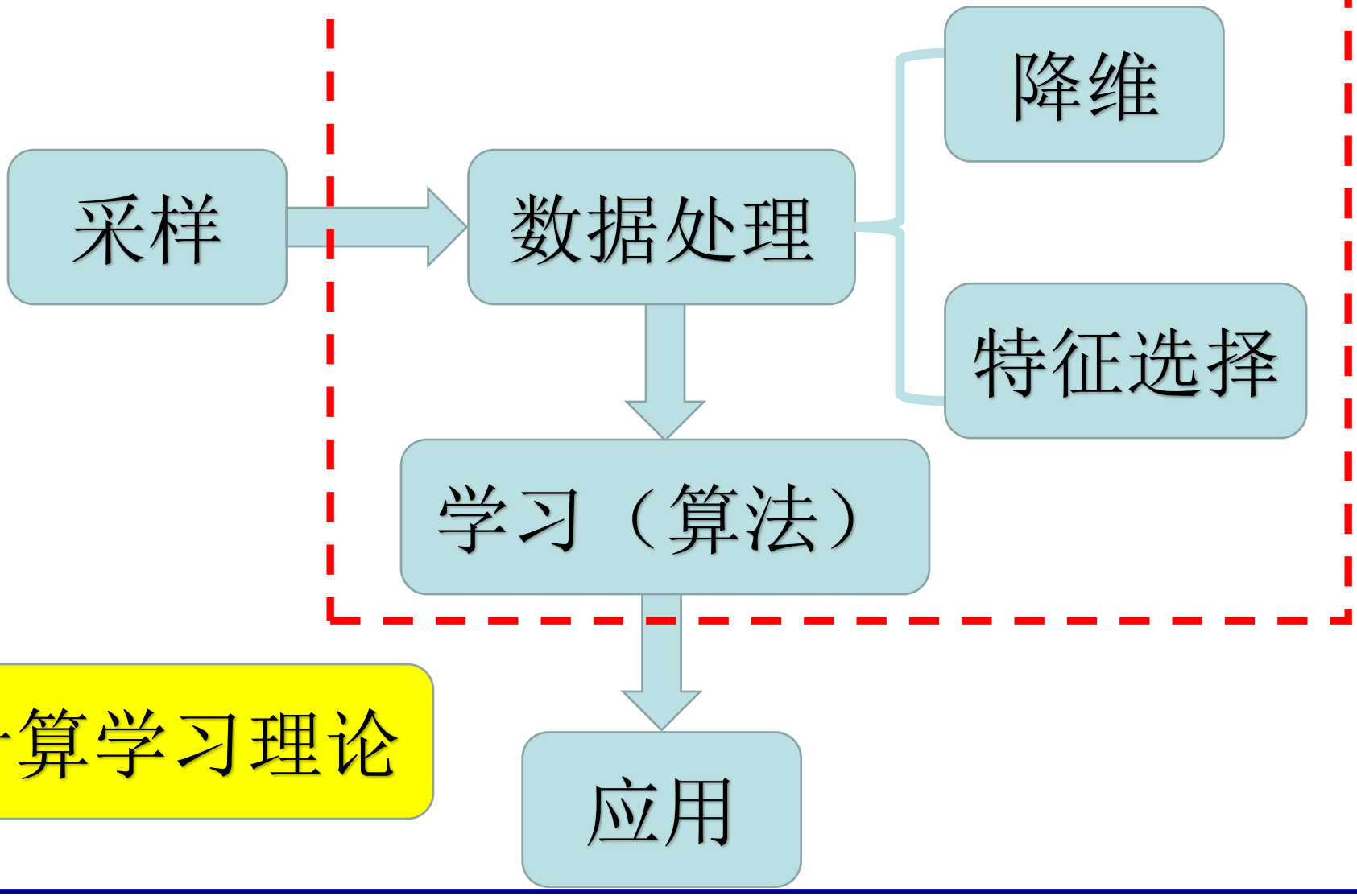
基本术语–数据–模型

使用学习算法（*learning algorithm*）

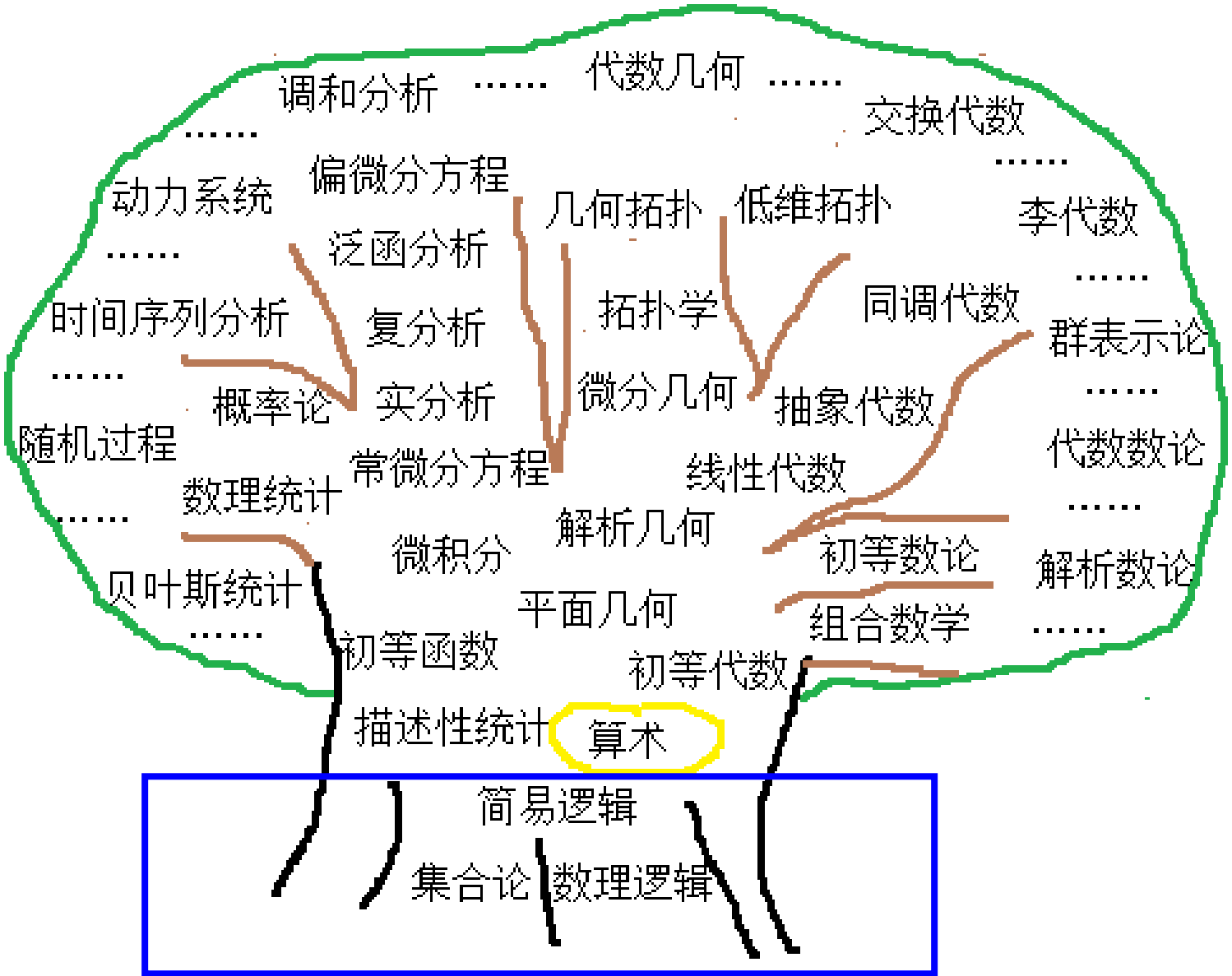


姓名	职业	年收入	…	好顾客
张三	教师	6 万	…	否
李四	公务员	7 万	…	是
王五	学生	2 万	…	否
周六	企业家	15 万	…	是
董七	演员	13 万	…	否
钱八	教师	8 万	…	是

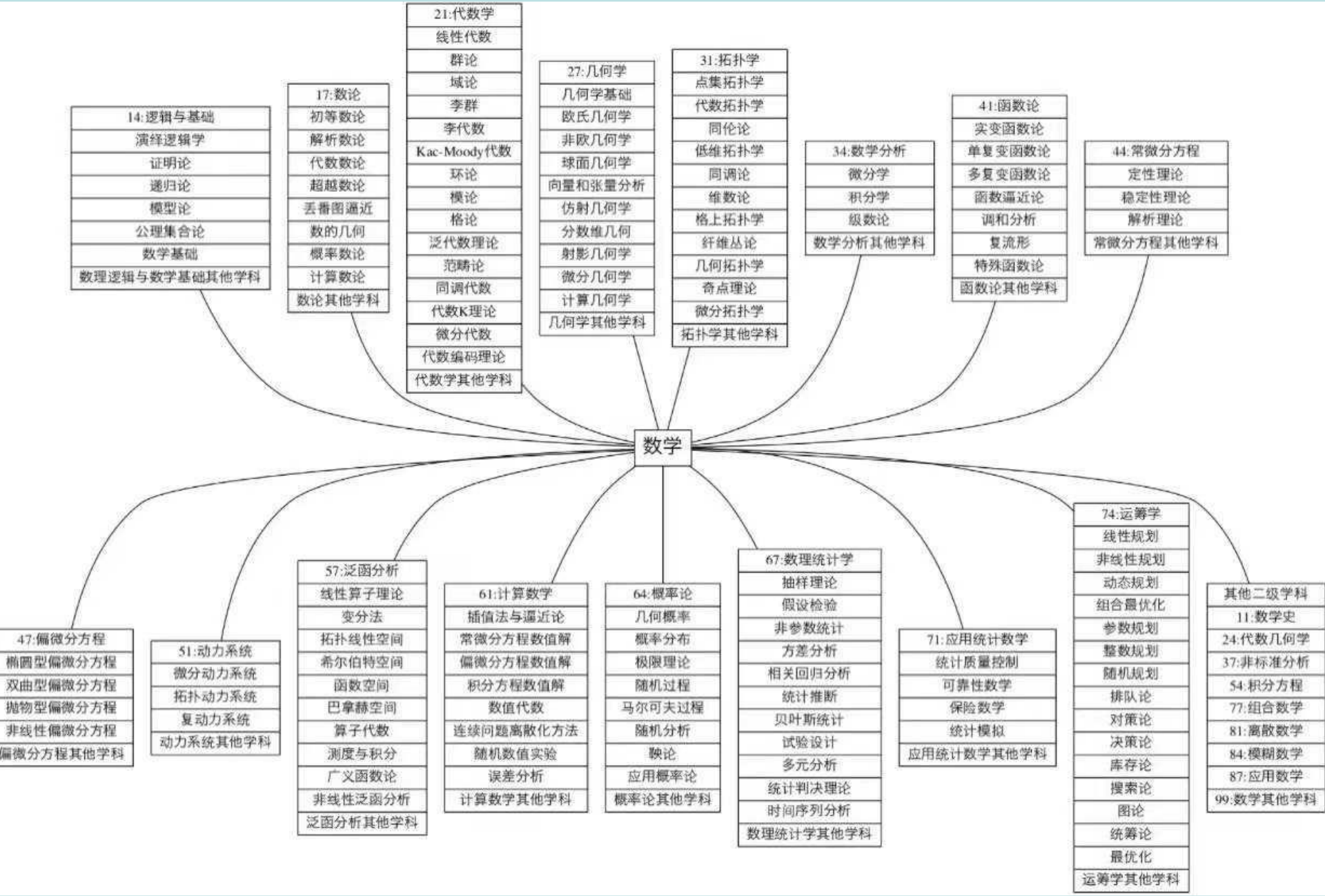
基本内容

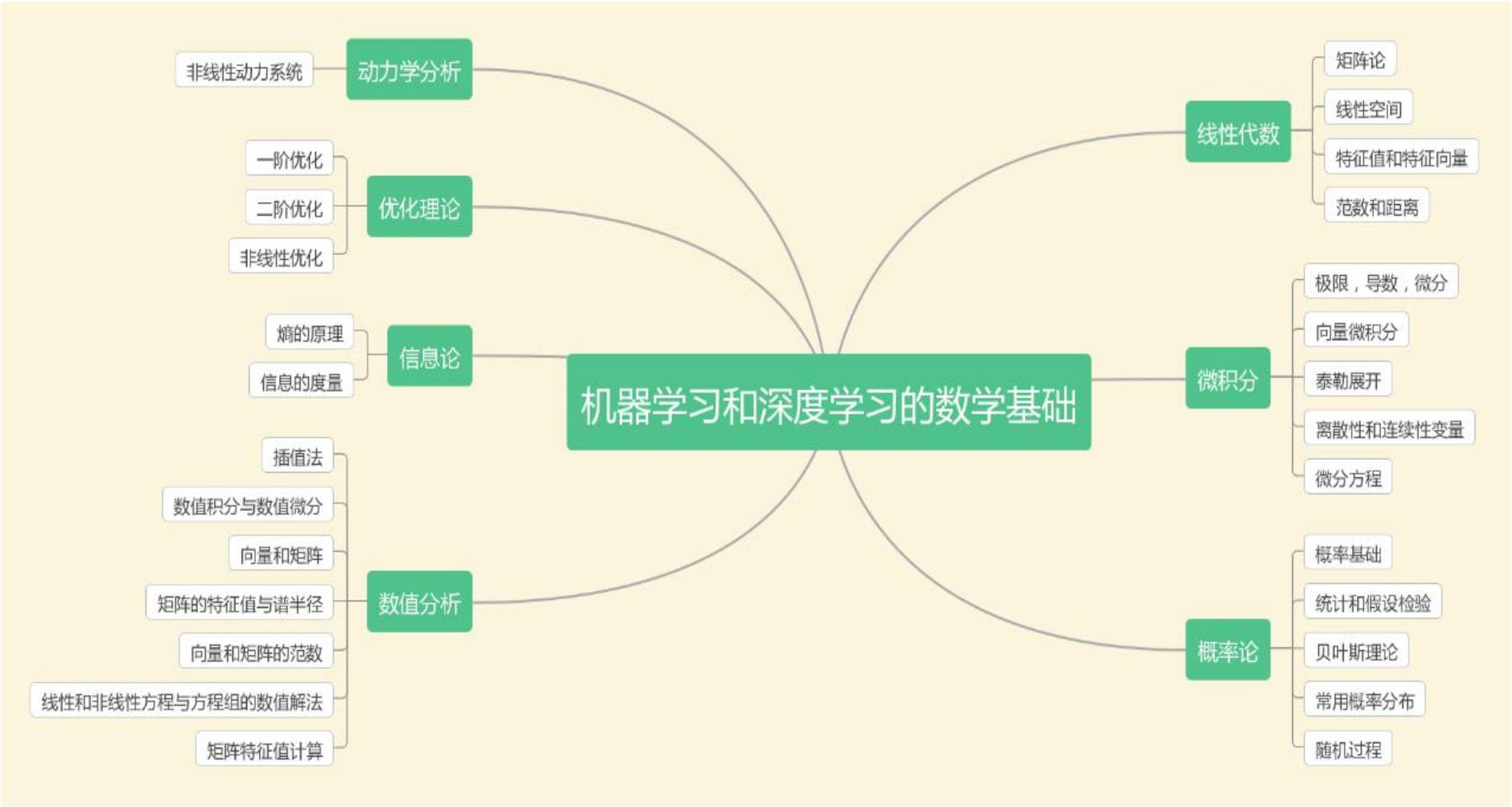


数学分支简介


调和分析
……
代数几何
……
交换代数
……
动力系统
……
偏微分方程
泛函分析
几何拓扑
低维拓扑
李代数
……
时间序列分析
……
复分析
实分析
拓扑学
同调代数
群表示论
……
概率论
随机过程
……
常微分方程
微分几何
抽象代数
代数数论
……
数理统计
……
线性代数
解析几何
微积分
初等数论
解析数论
……
贝叶斯统计
……
平面几何
组合数学
初等函数
初等代数
描述性统计
算术
简易逻辑
集合论
数理逻辑

数学分支简介


数学
14:逻辑与基础
演绎逻辑学
证明论
递归论
模型论
公理集合论
数学基础
数理逻辑与数学基础其他学科
17:数论
初等数论
解析数论
代数数论
超越数论
丢番图逼近
数的几何
概率数论
计算数论
数论其他学科
21:代数学
线性代数
群论
域论
李群
李代数
Kac-Moody代数
环论
模论
格论
泛代数理论
范畴论
同调代数
代数K理论
微分代数
代数编码理论
代数学其他学科
27:几何学
几何学基础
欧氏几何学
非欧几何学
球面几何学
向量和张量分析
仿射几何学
分数维几何
射影几何学
微分几何学
计算几何学
几何学其他学科
31:拓扑学
点集拓扑学
代数拓扑学
同伦论
低维拓扑学
同调论
维数论
格上拓扑学
纤维丛论
几何拓扑学
奇点理论
微分拓扑学
拓扑学其他学科
34:数学分析
微分学
积分学
级数论
数学分析其他学科
41:函数论
实变函数论
单复变函数论
多复变函数论
函数逼近论
调和分析
复流形
特殊函数论
函数论其他学科
44:常微分方程
定性理论
稳定性理论
解析理论
常微分方程其他学科
47:偏微分方程
椭圆型偏微分方程
双曲型偏微分方程
抛物型偏微分方程
非线性偏微分方程
偏微分方程其他学科
51:动力系统
微分动力系统
拓扑动力系统
复动力系统
动力系统其他学科
57:泛函分析
线性算子理论
变分法
拓扑线性空间
希尔伯特空间
函数空间
巴拿赫空间
算子代数
测度与积分
广义函数论
非线性泛函分析
泛函分析其他学科
61:计算数学
插值法与逼近论
常微分方程数值解
偏微分方程数值解
积分方程数值解
数值代数
连续问题离散化方法
随机数值实验
误差分析
计算数学其他学科
64:概率论
几何概率
概率分布
极限理论
随机过程
马尔可夫过程
随机分析
鞅论
应用概率论
概率论其他学科
67:数理统计学
抽样理论
假设检验
非参数统计
方差分析
相关回归分析
统计推断
贝叶斯统计
试验设计
多元分析
统计判决理论
时间序列分析
数理统计学其他学科
71:应用统计数学
统计质量控制
可靠性数学
保险数学
统计模拟
应用统计数学其他学科
74:运筹学
线性规划
非线性规划
动态规划
组合最优化
参数规划
整数规划
随机规划
排队论
对策论
决策论
库存论
搜索论
图论
统筹论
最优化
运筹学其他学科
其他二级学科
11:数学史
24:代数几何学
37:非标准分析
54:积分方程
77:组合数学
81:离散数学
84:模糊数学
87:应用数学
99:数学其他学科


机器学习和深度学习的数学基础
动力学分析
非线性动力系统
优化理论
一阶优化
二阶优化
非线性优化
信息论
熵的原理
信息的度量
数值分析
插值法
数值积分与数值微分
向量和矩阵
矩阵的特征值与谱半径
向量和矩阵的范数
线性和非线性方程与方程组的数值解法
矩阵特征值计算
线性代数
矩阵论
线性空间
特征值和特征向量
范数和距离
微积分
极限，导数，微分
向量微积分
泰勒展开
离散性和连续性变量
微分方程
概率论
概率基础
统计和假设检验
贝叶斯理论
常用概率分布
随机过程

基本术语-任务

预测目标：

- 分类：离散值
  - 二分类：好瓜；坏瓜
  - 多分类：冬瓜；南瓜；西瓜
- 回归：连续值
  - 瓜的成熟度
- 聚类：无标记信息

基本术语-任务

有无标记信息

- 监督学习：分类、回归
- 无监督学习：聚类
- 半监督学习：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”，称模型适用于新样本的能力为泛化(generalization)能力。

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

章节目录

- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状

假设空间

假设空间：由输入空间到输出空间的映射的集合。也就是由输入空间 X 到输出空间 Y 的映射 f : X → Y 所构成的集合，该空间是一个函数空间，即由函数所构成的集合（李航-机器学习方法）

集合元素个数数目：$|Y|^{|X|}$

学习过程：在所有假设空间组成的空间中进行搜索的过程。

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

(色泽=?)∧(根蒂=?)∧(敲声=?)↔好瓜

色泽2种、根蒂2种、敲声3种：假设空间：2^(2*2*3)

假设空间

可以自行选择（找子集）所考虑问题的假设空间的大小范围，并不一定非要选择包含所有可能性的映射集作为假设空间。（周志华）

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

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

通配符(*)考虑在内，在模型空间中搜索不违背训练集的假设空间大小：

3*3*4+1=37（考虑好瓜不存在）

版本空间

版本空间就是与训练集一致的所有假设所构成的集合，也就是假设空间的一个最大子集，该子集内的每一个元素都不与训练集相冲突。

由于每个假设都是一个完全确定的映射，可以将其与训练集中的每一个样例作比较，所得结果要么有冲突，要么没冲突。有冲突就不放进版本空间，都没冲突就放进版本空间

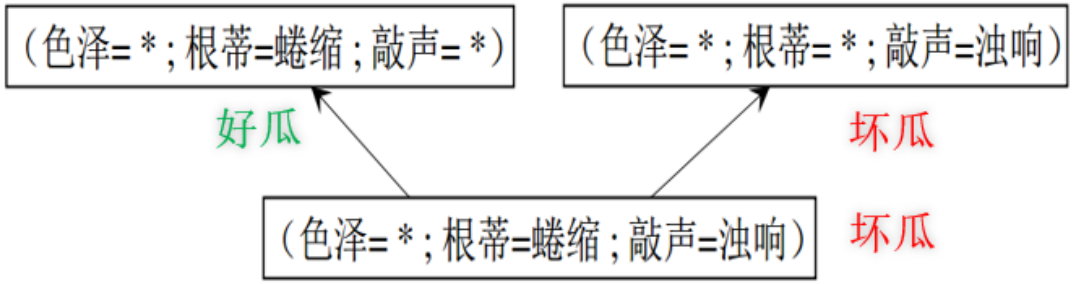
章节目录

- 基本术语
- 假设空间
- **归纳偏好**
- 发展历程
- 应用现状

归纳偏好

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

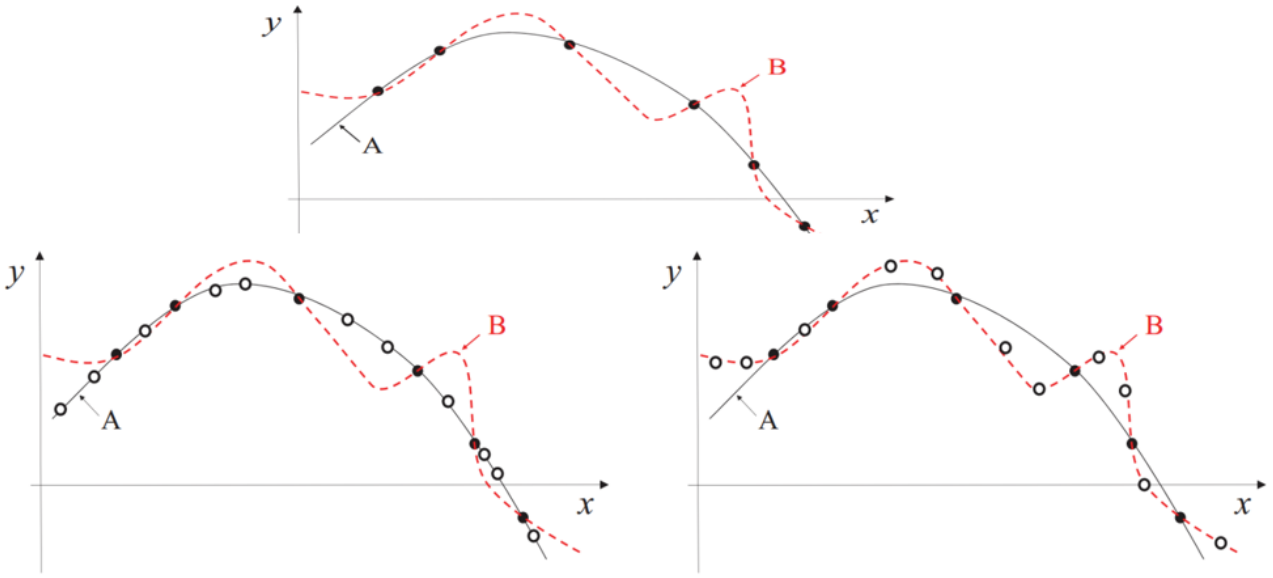
假设空间中有三个与训练集一致的假设，但他们对(色泽=青绿；"根蒂=" 蜷缩；敲声=沉闷)的瓜会预测出不同的结果：



选取哪个假设作为学习模型？

归纳偏好

回归问题



(a) A 优于 B　　(b) B 优于 A

没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

对于一个学习算法La，某些问题是比Lb好，必然在一些问题上不如Lb

归纳偏好

选取哪个假设作为学习模型？什么样的模型更好？

- “奥卡姆剃刀”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”。
- KNN 中假设特征空间中相邻的样本倾向于属于同一类
- SVM 中假设好的分类器应该最大化类别边界距离

- 深度神经网络偏好性地认为，层次化处理信息有更好效果；
- 卷积神经网络 认为信息具有空间局部性，可用滑动卷积共享权重的方式降低参数空间；
- 循环神经网络 则将时序信息纳入考虑，强调顺序重要性；
- 图网络则认为中心节点与邻居节点的相似性会更好地引导信息流动。

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.

归纳偏好

最近邻居	假设在特征空间（feature space）中一小区域内大部分的样本是同属一类。给一个未知类别的样本，猜测它与它最紧接的大部分邻居是同属一类。这是用于最近邻居法的偏置。这个假设是相近的样本应倾向同属于一类别。**KNN就是基于这种思想**
最少特征数	除非有充分的证据显示一个特征是有效用的，否则它应当被删除。这是特征选择算法背后所使用的假设。
最大边界	当要在两个类别间画一道分界线时，试图去最大化边界的宽度。这是用于**支持向量机的偏置**
最小描述长度	当构成一个假设时，试图去最小化其假设的描述长度。假设越简单，越可能为真的。**奥卡姆剃刀**的理论基础
最大条件独立性	如果假说能转成贝叶斯模型架构，则试着使用最大化条件独立性。这是用于**朴素贝叶斯**分类器的偏置。
最小交叉验证误差	当试图在假说中做选择时，挑选那个具有最低交叉验证误差的假说。

脱离问题谈“什么学习算法好”毫无意义！！！

章节目录

- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状

发展历程

推理期（20世纪50-70年代）：

- A. Newell和H. Simon的“逻辑理论家”(Logic Theorist)程序以及此后的“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；

知识期（70年代中期）：

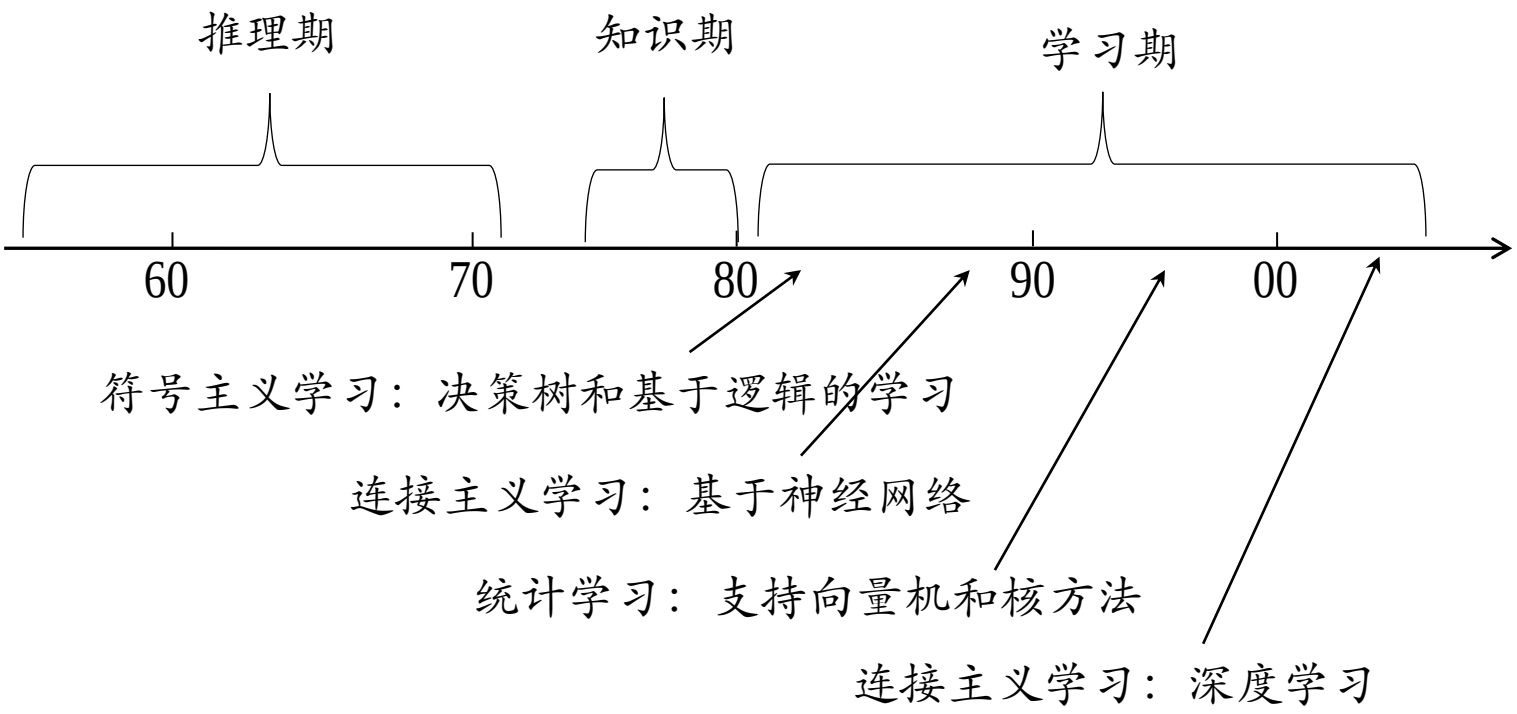
- 大量专家系统问世，在很多应用领域取得大量成果，由人来总结知识再交给计算机相当困难。

发展历程

学习期：

- 符号主义学习（20世纪80年代）
  - 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
  - 感知机（20世纪50年代）
  - Hopfield神经网络（80年代）
  - BP神经网络（80年代）
- 统计学习（90年代）
  - 支持向量（60-70年代）开始研究
  - 支持向量机及核方法

发展历程


推理期
知识期
学习期
60
70
80
90
00
符号主义学习：决策树和基于逻辑的学习
连接主义学习：基于神经网络
统计学习：支持向量机和核方法
连接主义学习：深度学习

章节目录

应用现状

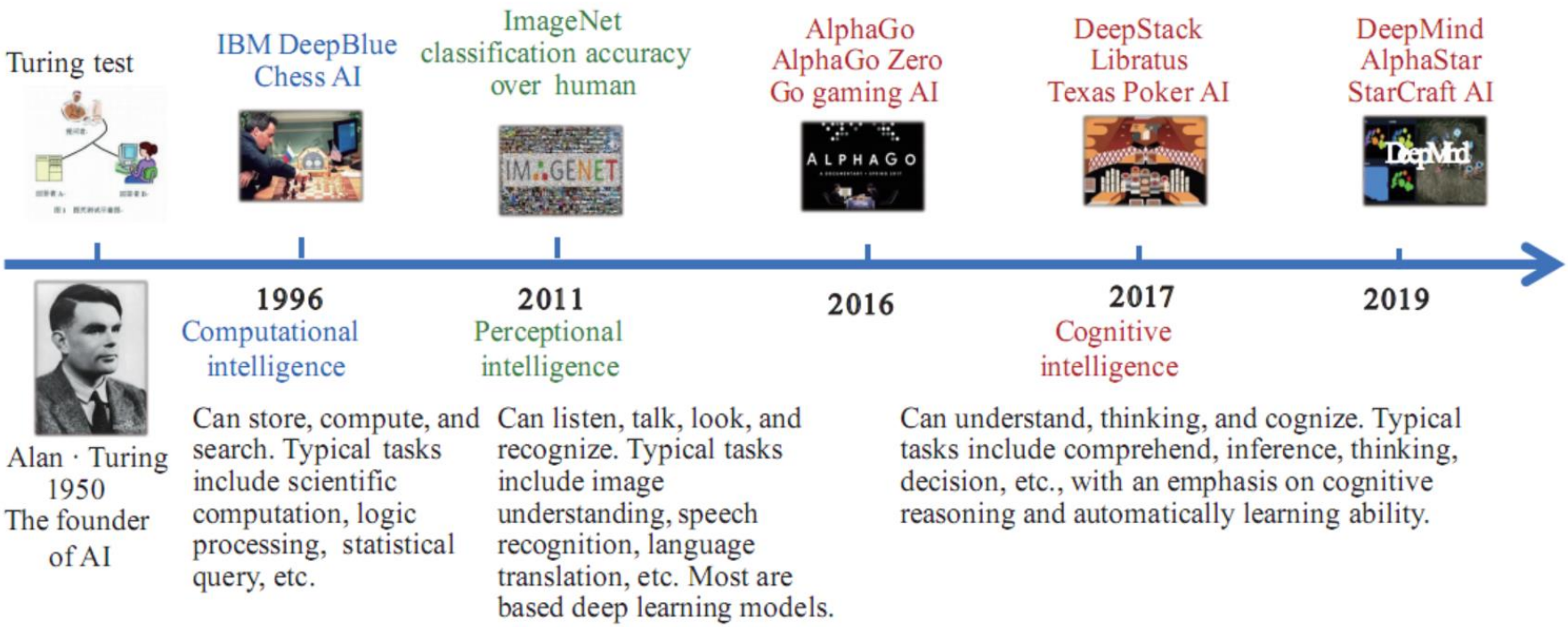
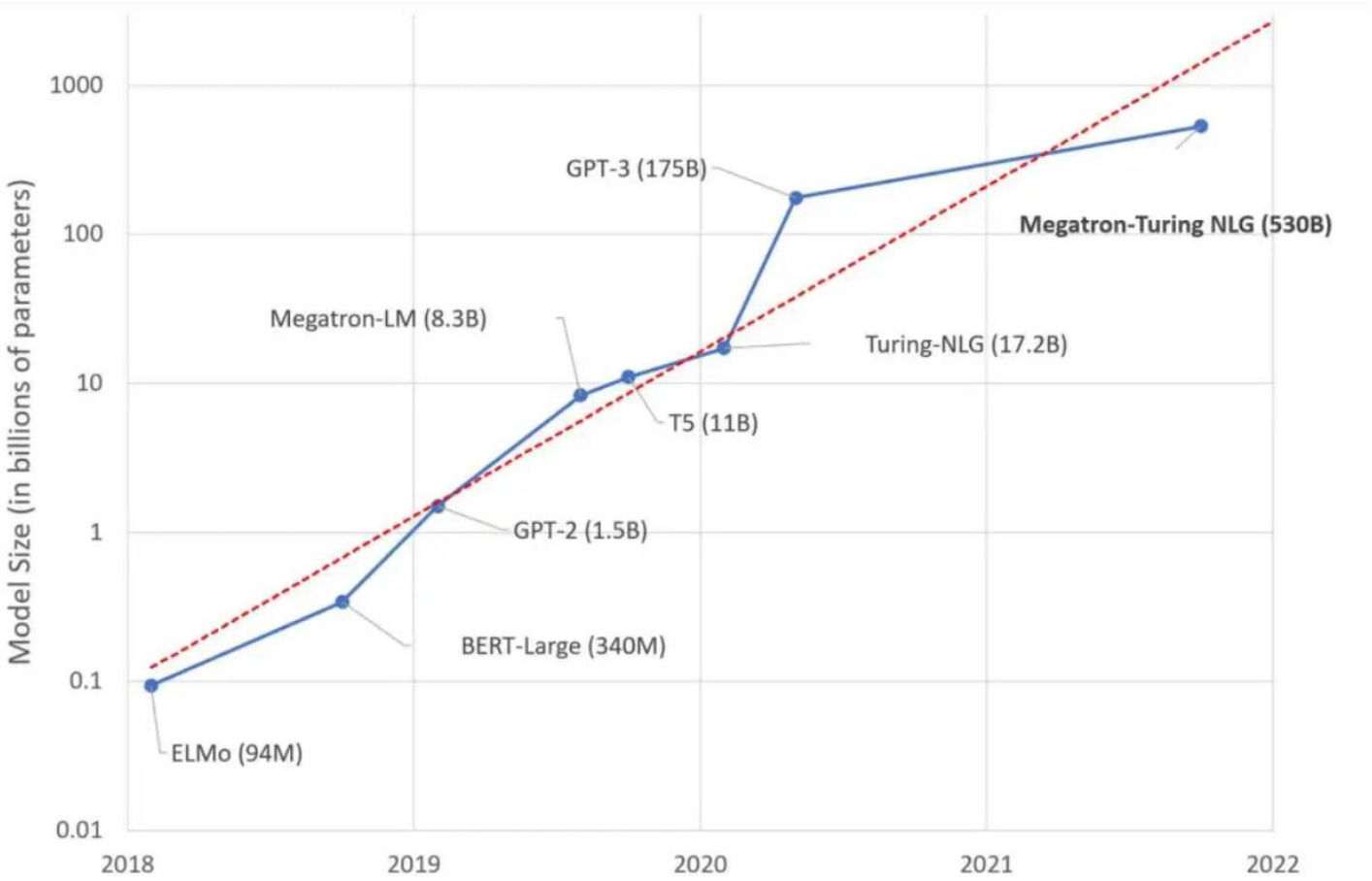


图 1　人机对抗的发展历史

应用现状



大脑的深度学习研究人员估计，人类大脑平均包含860亿个神经元和100万亿个突触。但不是所有的都用于语言。有趣的是，GPT-4预计将有大约100万亿个参数。

应用现状

千亿级参数通用大模型（文本编辑、编程、翻译、算术）