

人工智能导论

任课老师：赵放

单 位：南京大学 智能科学与技术学院

Email : fzhao@nju.edu.cn

人工智能方向第一门专业课程

掌握人工智能的主要流派与代表性思想

具备动手实践能力

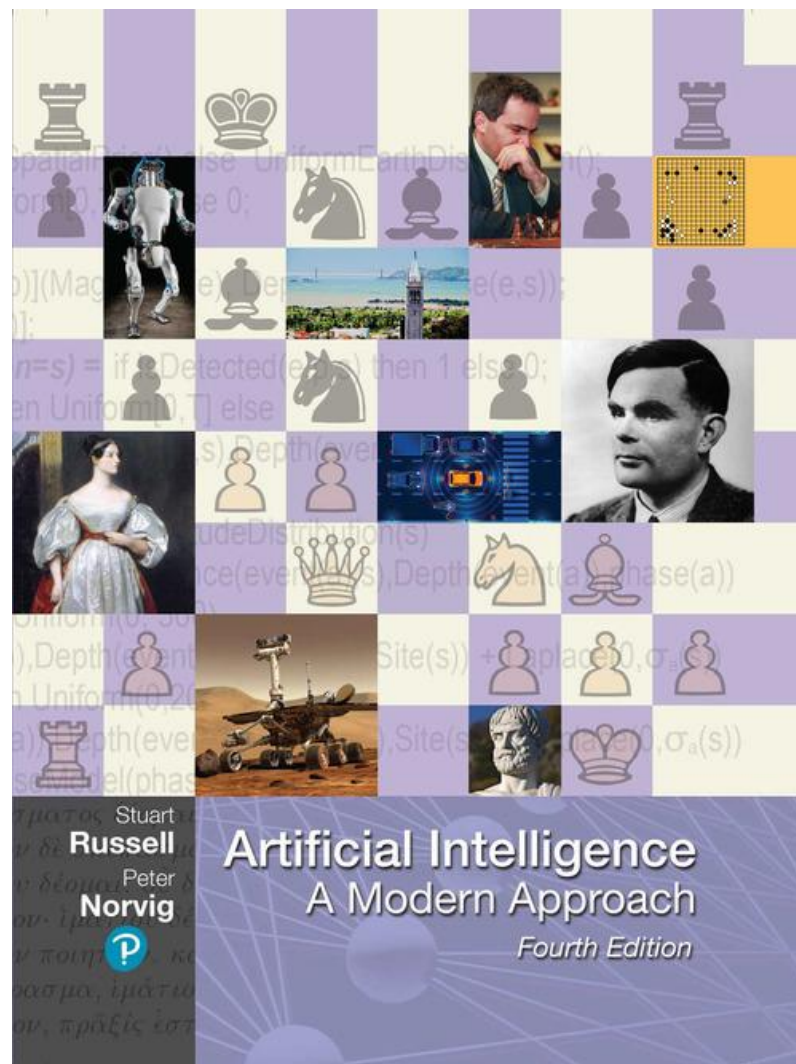
参考教材

Artificial Intelligence: A Modern Approach (Fourth Edition)

人工智能：现代方法 (第4版)

斯图尔特·罗素，
彼得·诺维格

第1版出版于1995年，人工智能领域的经典教科书，目前已经被全世界1500多所大学采用为教材。

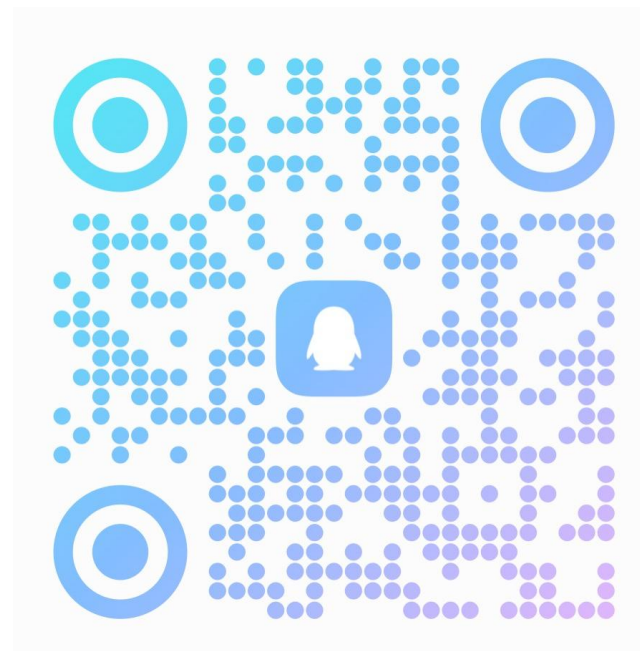


考核方式

课程总评包括：

- 平时作业：理论作业+实践作业，占比60%
- 课程设计：自主选题，单人完成，占比40%

课程讨论QQ群：1070958258



考核方式

课程设计

- 为鼓励大家接触与了解人工智能科研与实践，课程设计部分提供了若干个候选任务，你可以任选其一完成。
- 如果你想选择其他的方向，也可以先与老师进行沟通。
- 具体来说，**你需要先了解该任务的主要内容和相应的评测数据集，然后要在至少1个数据集上跑通1种算法，并得到结果。**
- 在此基础上，你还可以进一步调研更多的能够用于提升性能的方法或者自行设计算法，并进行实验分析。
- 最后，**你需要提交一份英文技术报告和对应的代码，我们会仿照学术会议的方式，构建一个线上提交平台，并邀请同学们担任审稿人，为其他同学进行评分。**
- 整个过程，包括如何找论文、如何读论文、如何想Idea、如何写代码做实验，你都可以借助AI来完成，体会如何利用AI辅助科研。

◆ 绪论

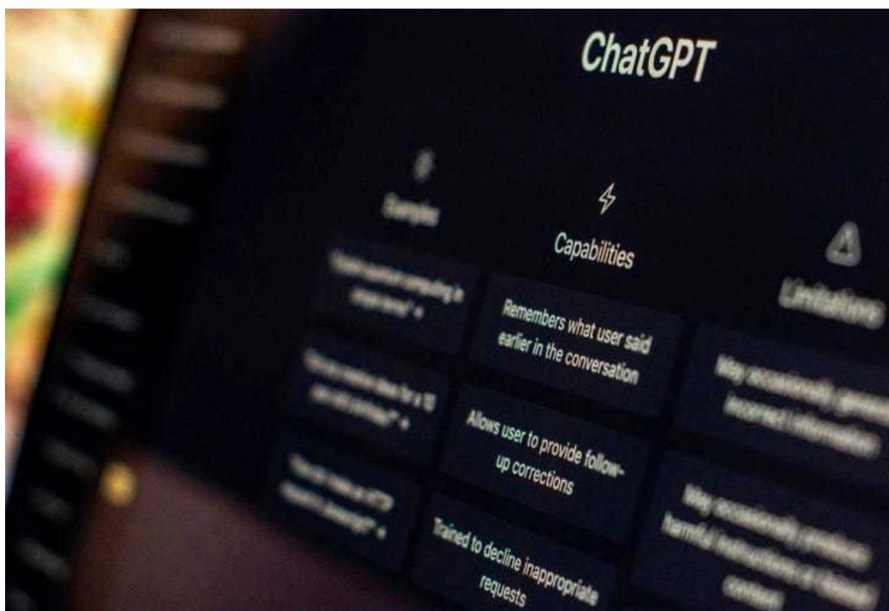
- 什么是人工智能
- 人工智能的发展简史
- 人工智能的未来与挑战
- 课程主要内容



什么是人工智能？

先看看人工智能
可以做什么？

2022年11月，OpenAI发布了AI对话模型ChatGPT，成为人工智能里程碑式的应用



OpenAI发布AI对话模型ChatGPT，开启生成式AI商业化新机遇 ...

2022年12月19日 — 当地时间11月30日，美国人工智能公司OpenAI发布全新产品ChatGPT，一款基于GPT-3.5的免费对话模型。公司CEO Sam Altman透露上线五天该模型的全球用户数量 ...

AI新物种：ChatGPT 不会止于写代码调Bug - 巴比特

2022年12月11日 — GPT-3发布于2020年，作为一个自监督模型，几乎可以完成自然语言处理的绝大部分任务，在参数上，GPT-1包含了1.17亿个参数，GPT-2包含了15亿个参数，而GPT-3 ...

ChatGPT 通过了美国MBA、法律和医学考试 - Showmetech

2023年1月26日 — 这次的新颖之处在于，根据在美国进行的一项研究，OpenAI 会轻松通过即使是学生也难以通过的复杂测试。重点是创建包含所有重要细节的法律文件和聊天GPT ...

微软100亿美元砸向OpenAI，ChatGPT要加入Office全家桶了？

2023年1月12日 — 如果100 亿美元的交易成真，OpenAI 将获得巨额资金，微软赢得广阔未来，双赢局面就此达成。过去一段时间，对话式AI 模型ChatGPT 火遍了整个社区，它 ...

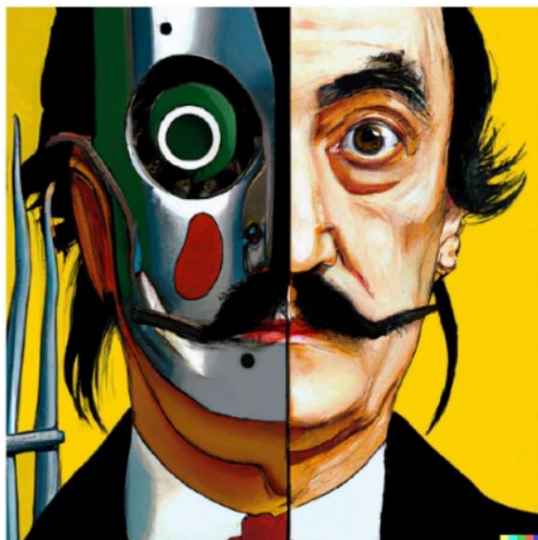
GPT-4：人工智能的新语言方法被定义为“强大” - Showmetech

2023年1月2日 — GPT-2024 计划于4 年发布，应该会为ChatGPT 带来更好的理解和文本创建。... 不浪费时间，该公司已经在准备下一代AI 语言方法，预计将于2023 年推出， ...

美国大学89%的学生居然用ChatGPT写作业 - 国际竞赛

1天前 — ChatGPT的崛起并在高等教育领域的突然普及，让众多美国高校感觉措手不及！... Nature早就很有先见之明地发文，担心ChatGPT会成为学生写论文的工具。

图像生成



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it





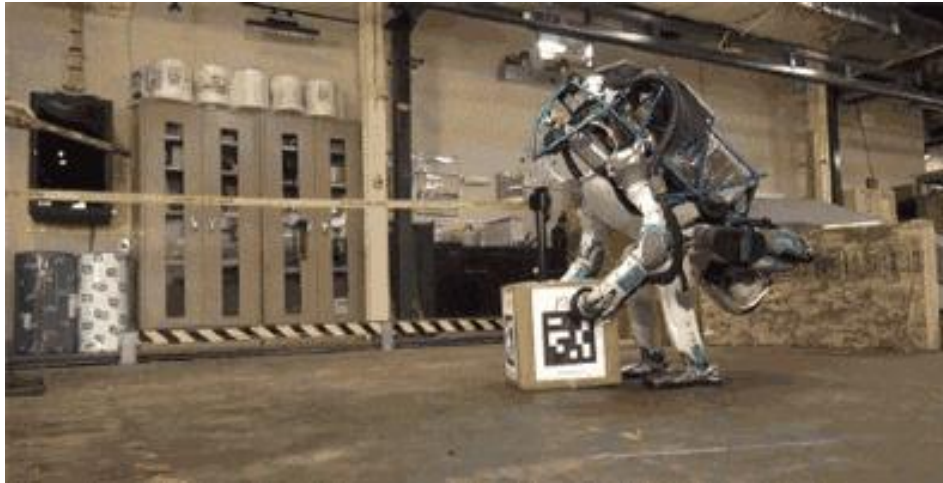


在美国科罗拉多艺术博览会美术比赛中，游戏设计师Jason Allen的作品《太空歌剧院》夺冠，该副画作是他使用AI作图工具MidJourney完成



计算/预测出较高胜率的走法？
大量棋谱如何生成/利用
公开的计算难题，意义重大
熟知的日常游戏，影响深远

具身智能



RESEARCH

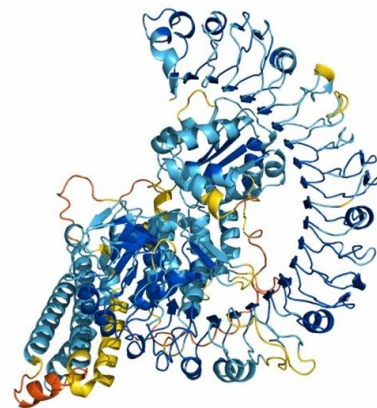
Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

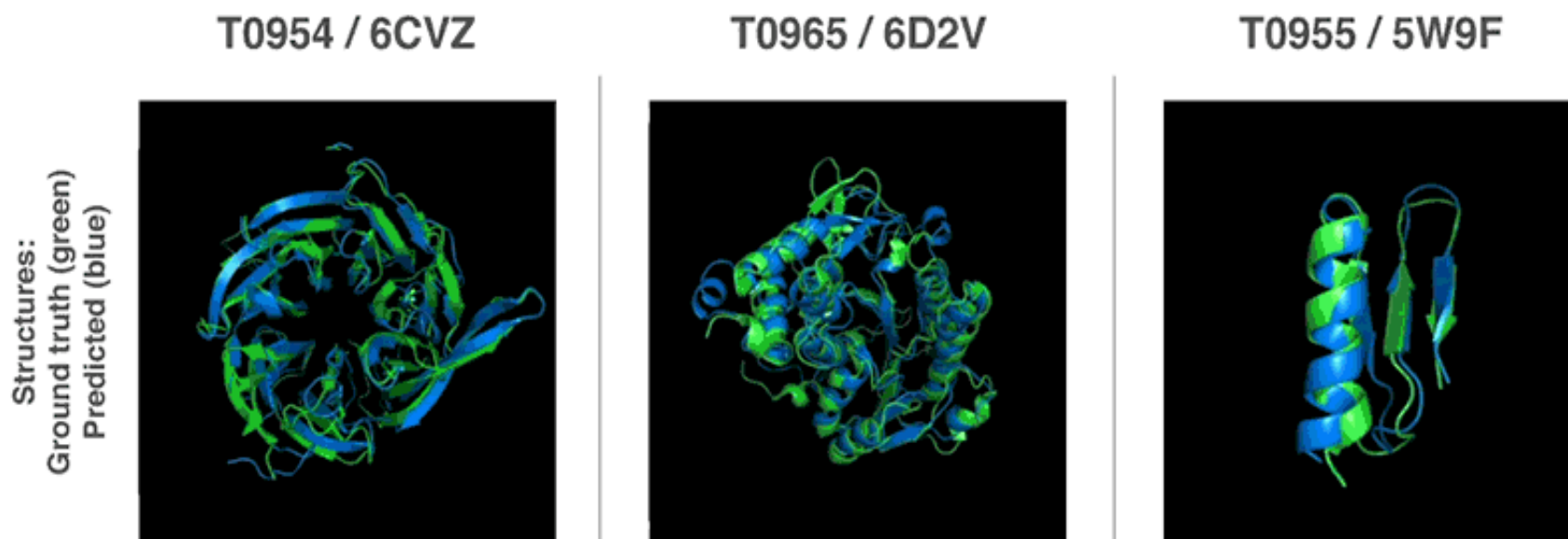
Thang Luong and Edward Lockhart



- 2020年11月，DeepMind的人工智能程序AlphaFold 2在蛋白质结构预测大赛CASP 14夺冠，对大部分蛋白质结构的预测与真实结构只差一个原子的宽度，达到了人类利用冷冻电镜等复杂仪器观察预测的水平
- 2021年8月，DeepMind宣布已将人类的98.5%的蛋白质预测了一遍；此前，已知氨基酸顺序的蛋白质分子的三维结构被看清的不到0.1%
- 2022年8月，DeepMind宣布，AlphaFold可以预测出2亿多个蛋白质结构，几乎涵盖了地球上所有已进行过基因组测序的生物体，其中35%已达到实验手段所能获取的结构精度



2022年8月25日，华盛顿大学（University of Washington）David Baker教授团队在《细胞》杂志上发表论文，**利用AI技术精准地从头设计出能够穿过细胞膜的大环多肽分子，开辟了设计全新口服药物的新途径**

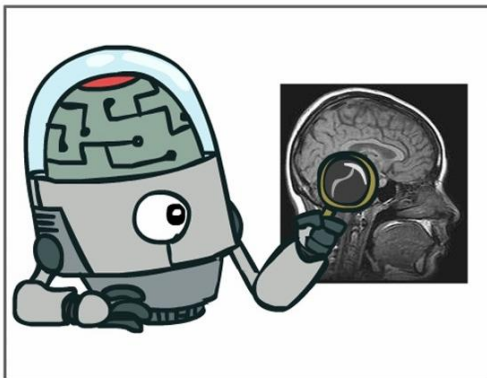


AI预测的蛋白质结构（蓝色）与实际结构（绿色）对比

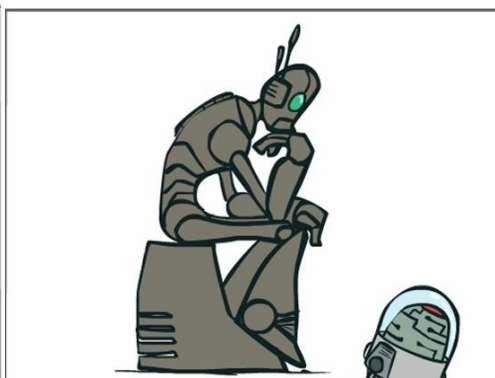
什么是人工智能

从**类人**与**理性**以及**思考**与**行为**这两个维度来看，有4种可能的组合：

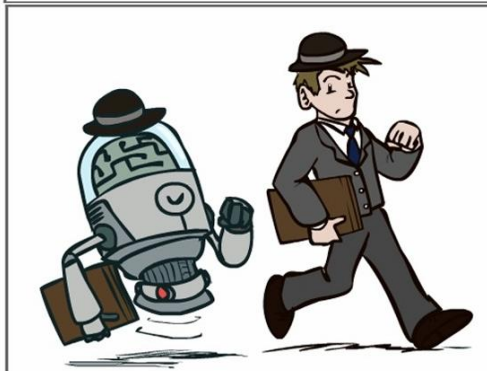
类人思考



理性思考



类人行为



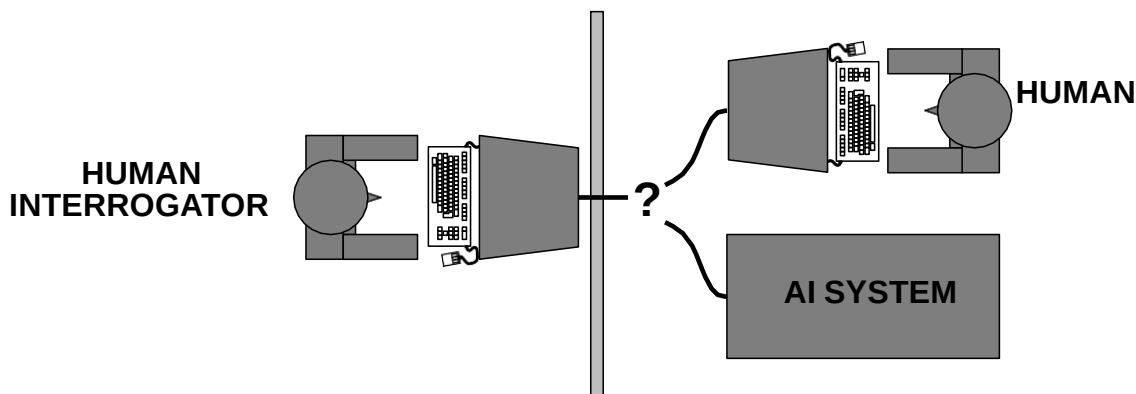
理性行为



需要关于大脑内部活动的科学理论

- 如何对知识进行抽象？信息如何传递？
 - 如何验证
 - 1) 预测和测试人类的行为
 - 2) 从神经学数据中直接识别
-
- **心理实验 (psychological experiment)** ——观察一个人的行为；
 - **大脑成像 (brain imaging)** ——观察大脑的活动。

- “机器能思考吗？” --> “机器能做出智能行为吗？”
- 智能行为测试：模仿游戏（Imitation Game）



图灵测试（Turing, 1950）

这个智能行为测试被提出的目的是**判断机器能否像人一样思考。**

为了能够通过图灵测试，我们需要去研究：**自然语言处理、知识表示、推理与规划、机器学习、视觉、语音、机器人**

- 希腊哲学家亚里士多德的**三段论 (syllogism)** 为论证结构提供了模式，当给出正确的前提时，总能得出正确的结论。

苏格拉底是人 + 所有人都是凡人 --> 苏格拉底是凡人

- 这些思维法则被认为支配着思想的运作，他们的研究开创了一个称为**逻辑 (logic)** 的领域。
- 19世纪的逻辑学家建立了一套精确的符号系统，用于描述世界上物体及其之间的关系。人工智能中所谓的**逻辑主义 (logicism)** 流派希望在此基础上创建智能系统。

- **理性行为**：做正确的事情。
- **智能体 (agent)**：能够采取行动的东西。
- 抽象地讲，智能体是从感知历史到行动的函数：

$$f : P^* \rightarrow A$$

- **计算机智能体**或者说**具身智能体**能够完成更多的任务，例如能够自主运行、感知周围环境、自适应更新以及制定和实现目标等。

- **理性行为**：做正确的事情。
- **智能体 (agent)**：能够采取行动的东西。
- 抽象地讲，智能体是从感知历史到行动的函数：

$$f : P^* \rightarrow A$$

- **计算机智能体**或者说**具身智能体**能够完成更多的任务，例如能够自主运行、感知周围环境、自适应更新以及制定和实现目标等。
- **理性智能体 (rational agent)** 需要为取得**最佳结果**或在存在不确定性时取得**最佳期望结果**而采取行动。

理性行为 vs 理性思考

理性行为 vs 类人行为

- **理性行为**：做正确的事情。
- **智能体 (agent)**：能够采取行动的东西。
- 抽象地讲，智能体是从感知历史到行动的函数：

$$f : P^* \rightarrow A$$

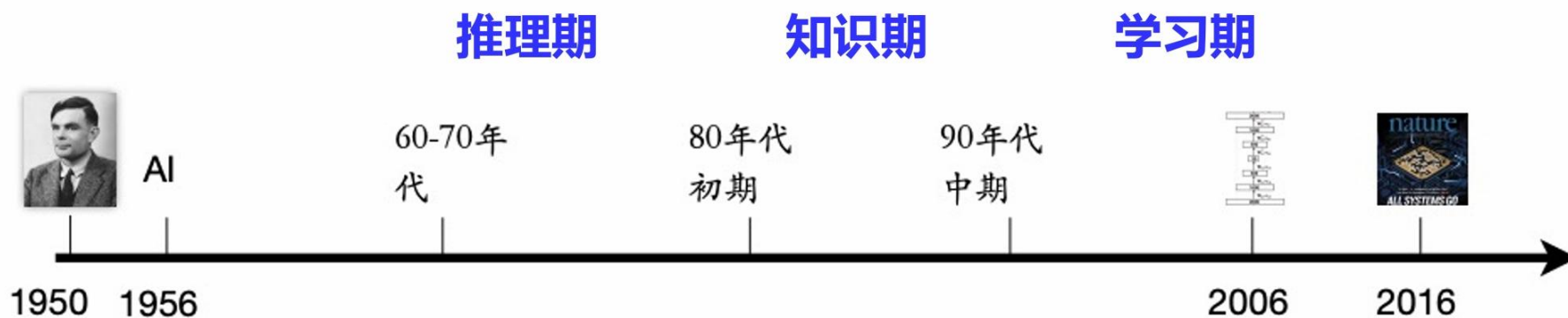
- **计算机智能体**或者说**具身智能体**能够完成更多的任务，例如能够自主运行、感知周围环境、自适应更新以及制定和实现目标等。
- **理性智能体 (rational agent)** 需要为取得**最佳结果**或在存在不确定性时取得**最佳期望结果**而采取行动。
- **人工智能专注于研究和构建做正确的事情的智能体**，其中正确的事情是我们提供给智能体的目标定义。

- **理性行为：做正确的事情。**
- **智能体 (agent)：**能够采取行动的东西。
- 抽象地讲，智能体是从感知历史到行动的函数：

$$f : P^* \rightarrow A$$

- **计算机智能体**或者说**具身智能体**能够完成更多的任务，例如能够自主运行、感知周围环境、自适应更新以及制定和实现目标等。
- **理性智能体 (rational agent)** 需要为取得**最佳结果**或在存在不确定性时取得**最佳期望结果**而采取行动。
- **人工智能专注于研究和构建做正确的事情的智能体**，其中正确的事情是我们提供给智能体的目标定义。
- **益机 (Beneficial Machine)：**对人类有益的智能体。

人工智能 (AI) 发展历程



人工智能 (AI) 发展历程

- 人工智能诞生：1943-1956
- First Summer：推理期，1956到20世纪70年代初
- First Winter：20世纪70年代
- Second Summer：知识期，20世纪70年代末到80年代末
- Second Winter：20世纪80年代末及之后的20年
- Third Summer：2010 – ?

人工智能的起源

Computing Machinery and Intelligence

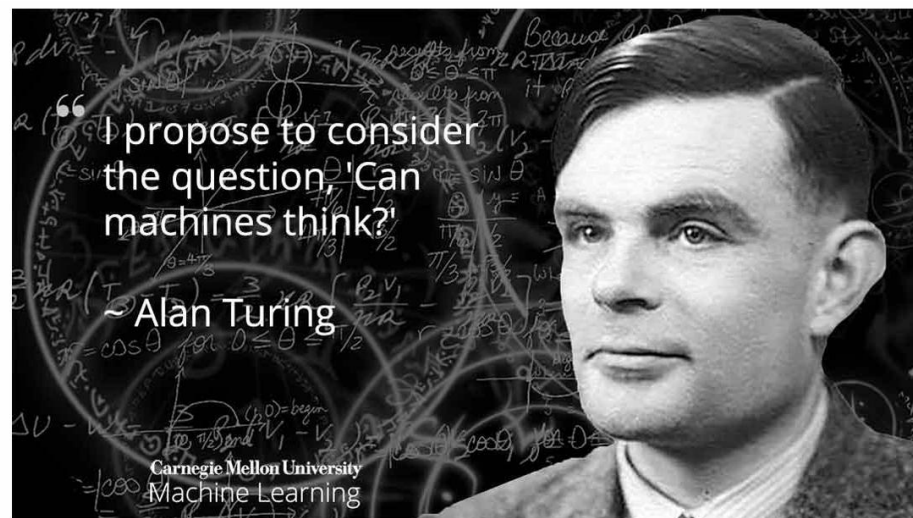
计算机与智能

1950年

艾伦·图灵

“Can machine think?”

机器能思考吗？

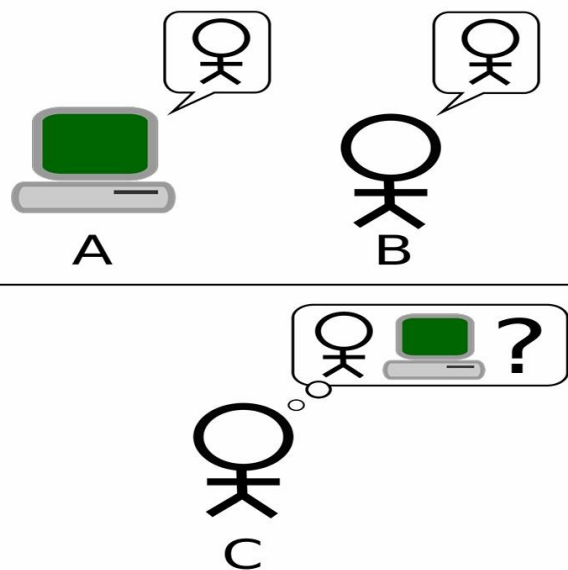


艾伦·图灵

1912-1954

图灵测试

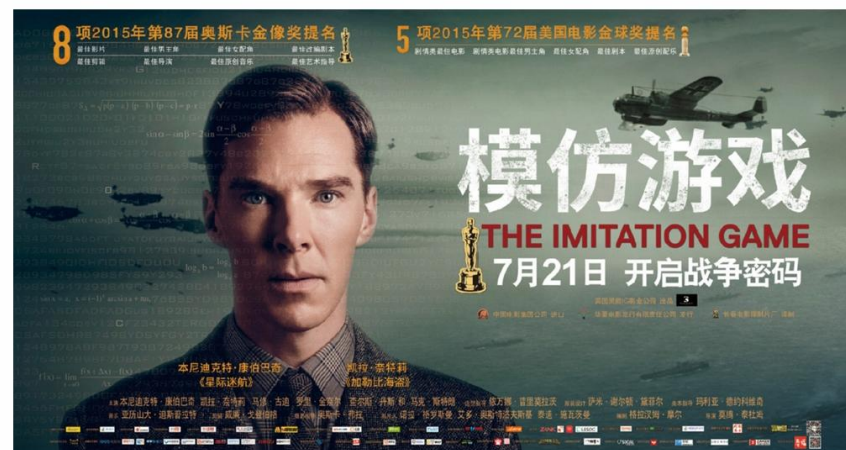
如何判断机器是否具有智能？



一人扮演提问者，另一人作为被测人员。这两个人与机器分别处在3个不同的房间，

提问者通过打印问题和接收打印问题来与测试人员和被测机器进行通信

计算机尽量模仿人，如果提问者判断不出哪个回答是人，哪个回答是计算机，就可以认为这台计算机具有智能



1956人工智能元年

1956年的达特茅斯会议标志人工智能这一学科的诞生



约翰·麦卡锡



马文·明斯基



克劳德·香农



雷·索洛莫诺夫



艾伦·纽厄尔



赫伯特·西蒙



阿瑟·塞缪尔



奥利弗·塞尔弗里奇



纳撒尼尔·罗切斯特



特伦查德·莫尔

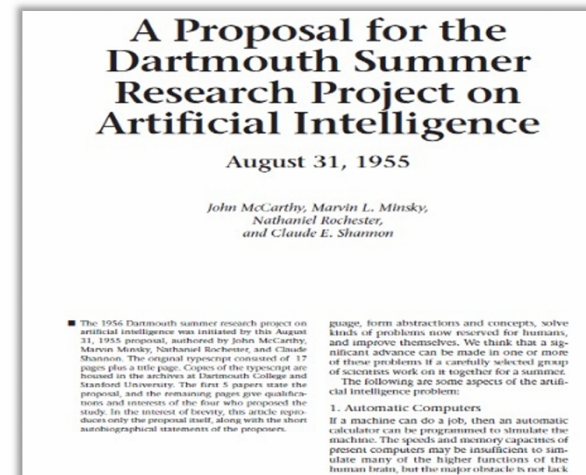


1956年夏 美国达特茅斯学院

1956人工智能元年

报告列举了Artificial Intelligence值得关注七个问题

- Automatic Computers
- How Can a Computer be Programmed to Use a Language
- Neuron Nets
- Theory of the Size of a Calculation
- Self-improvement
- Abstractions
- Randomness and Creativity



《人工智能达特茅斯夏季研究项目提案》

第一次浪潮：1956年到20世纪70年代初

1956-1970s: 推理期

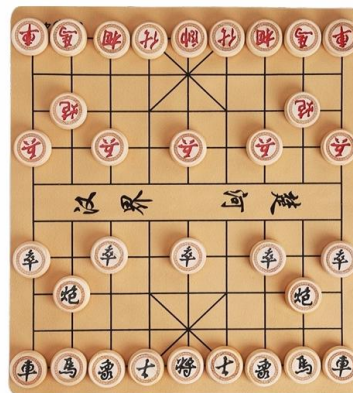
◆ 出发点：“数学家真聪明！”

◆ 把人的**思考逻辑**放入电脑

◆ 基本想法：

- 智能的核心在于对**符号**的操纵
- 把人类的知识和推理过程，翻译成机器能理解的符号和规则，机器也能像数学家一样思考

◆ 主要成就：自动定理证明、下棋



- 棋盘（符号：棋盘格）
- 棋子（符号：车、炮、兵...）
- 每个棋子怎么走（规则：马走日，象走田...）
- 基本策略（启发式规则：开局先出炮...）

智能 ≈ 符号 + 规则 + 搜索

- **符号**：用来代表现实世界中的对象、概念、关系等
- **规则**：定义了符号之间如何联系、如何变化（比如，IF P then Q）
- **搜索**：在庞大的候选空间中，根据规则寻找解决方案

第一次浪潮：1956年到20世纪70年代初

• 逻辑理论家 (Logic Theorist)

- 把已知的公理和定理作为“事实”（符号表示）
- 把逻辑推理规则（比如“如果A为真，且A能推出B，则B为真”）作为“操作”（规则）
- 在一个巨大的“可能性之树”上进行搜索，尝试从公理出发，一步步推导出目标定理



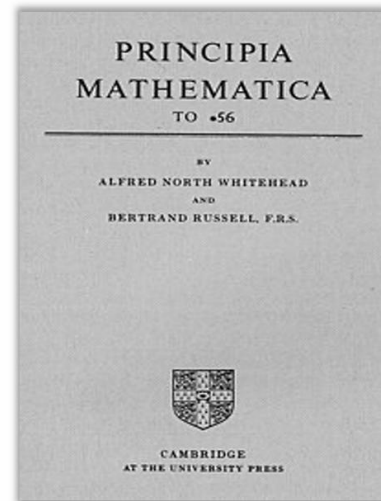
Bertrand Russell



Alfred Whitehead



西蒙与纽厄尔的逻辑理论家(Logic Theorist) ,
可以证明《数学原理》第二章52个定理中的38个



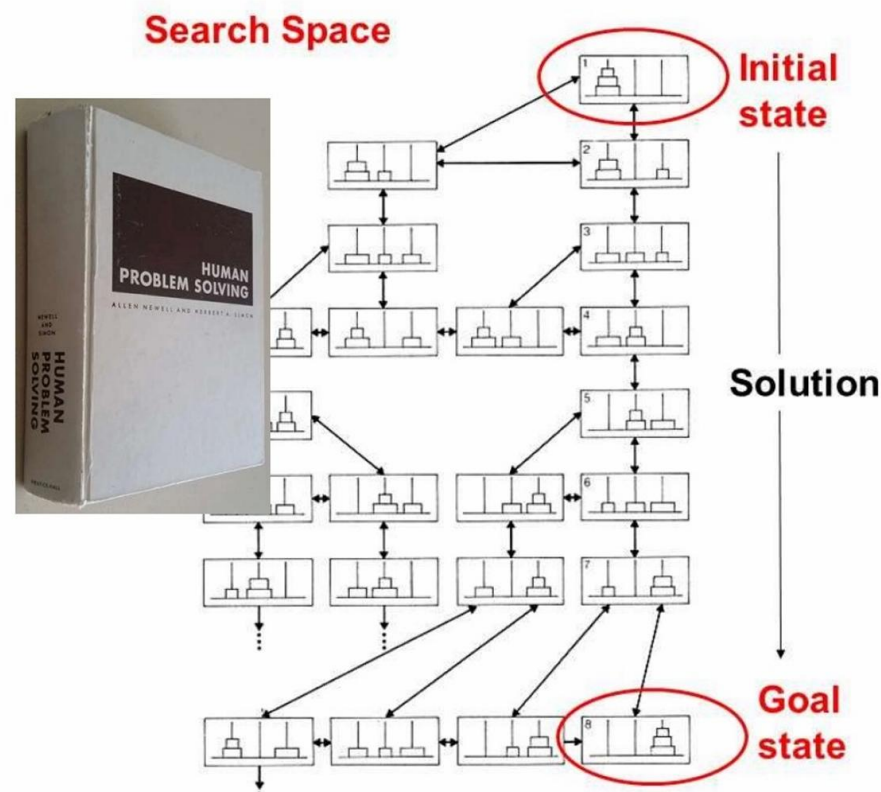
第一次浪潮：1956年到20世纪70年代初

- 通用问题求解器 (General Problem Solver): 手段-目的分析 (Means-Ends Analysis)



**Problem Solving
is a search
problem**

- 1.看看当前状态和目标状态的差距
- 2.找到一个能缩小这个差距的操作
- 3.应用这个操作
- 4.重复以上步骤，直到达到目标



第一次浪潮：1956年到20世纪70年代初

- **王浩**：1958年夏天，王浩在一台IBM 704机上实现了一个一阶逻辑程序，只用9分钟就证明了《数学原理》中一阶逻辑的全部150条定理中的120条
- 1959年夏天，改进版本证明了全部150条一阶逻辑以及200条命题逻辑定理



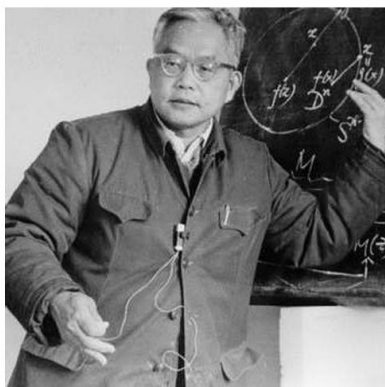
哥德尔与王浩

- 王浩注意到《数学原理》里的一阶逻辑公式都是AE形式（即前面是全称量词，后面是存在量词），王浩关于AEA可计算性和复杂性的研究，引出了他的学生库克（Stephen Arthur Cook）的NP理论
- 公正地说，王浩的定理证明研究孕育了整个理论计算机科学

**1983年，国际人工智能联合会（IJCAI）授予
王浩自动定理证明里程碑大奖**

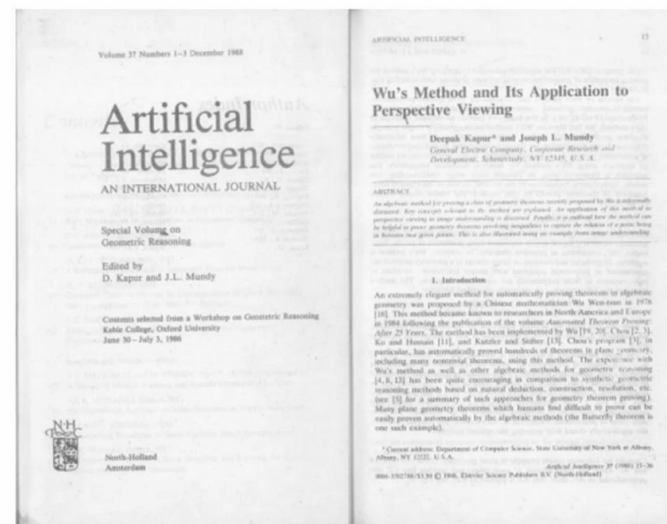
第一次浪潮：1956年到20世纪70年代初

- **吴文俊**：提出用计算机证明几何定理的“吴方法”，开创了机器几何定理证明的方向，是国际自动推理领域的先驱性的工作 -> “**数学机械化**”
- 1977年大年初一，手算成功验证了几何定理机器证明的方法，随后在一台由北京无线电一厂生产的长城203上证明了西姆森定理



吴文俊 (1919-2017)

“所有的问题都可以转变成数学问题，所有的数学问题都可以转变成代数问题，所有的代数问题都可以转变成解方程组的问题，所有解方程组的问题都可以转变成解单变元的代数方程问题”



1988年《人工智能》特辑开篇对吴方法的概述

第一次浪潮：1956年到20世纪70年代初

IBM亚瑟·塞缪尔（Arthur Samuel）的跳棋程序

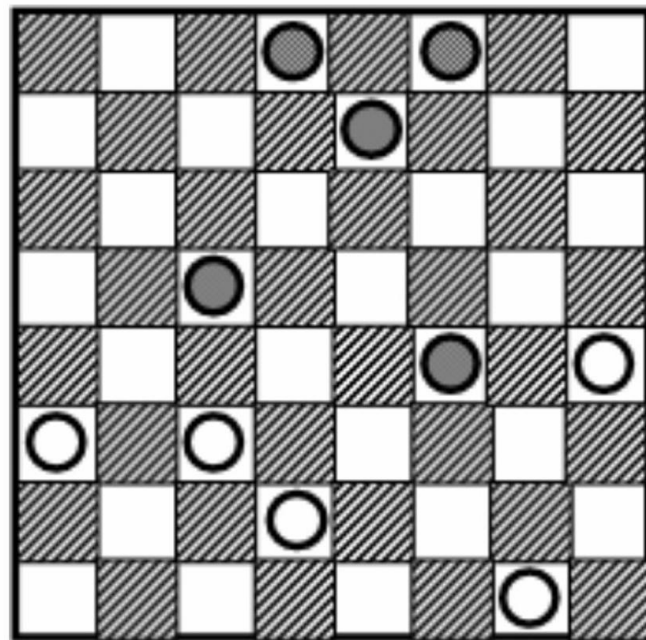
1962年战胜美国州跳棋冠军

穷举所有状态不可能 -> 设计评估函数给状态评分

学习能力：

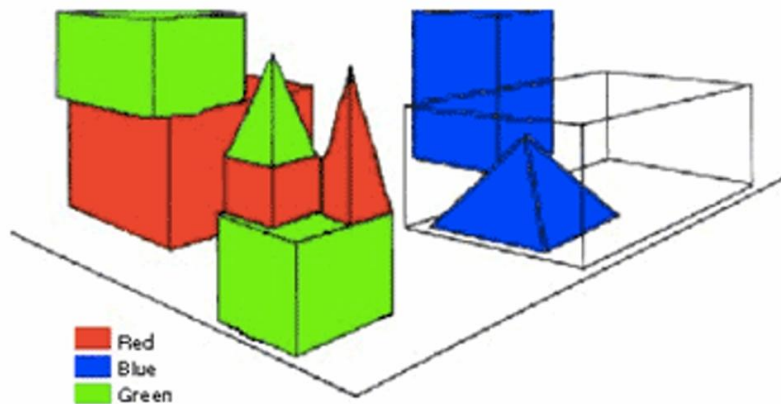
- 1.死记硬背：** 把遇到过的局面及其评估分数存起来，下次遇到直接用
- 2.参数学习：** 根据历史棋局，自动调整评估函数中的参数权重，让“好”的局面得分更高，“坏”的局面得分更低

机器学习思想的早期实践！



第一次浪潮：1956年到20世纪70年代初

SHRDLU系统 -- 积木世界



Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

- 视觉
- 自然语言处理
- 推理与规划

第一次浪潮：1956年到20世纪70年代初

第一代聊天机器人ELIZA，1964年 MIT的Joseph Weizenbaum

通过简单的关键词匹配和模板生成技术，模拟心理治疗师与患者的交流过程

例如，当用户输入“我感到很沮丧”时，ELIZA可能会回应：
“为什么你感到沮丧？”



第一次人工智能寒冬

早期程序高度依赖简化、严格定义的环境（比如积木世界、跳棋规则）
难以应对真实世界的复杂性与模糊性

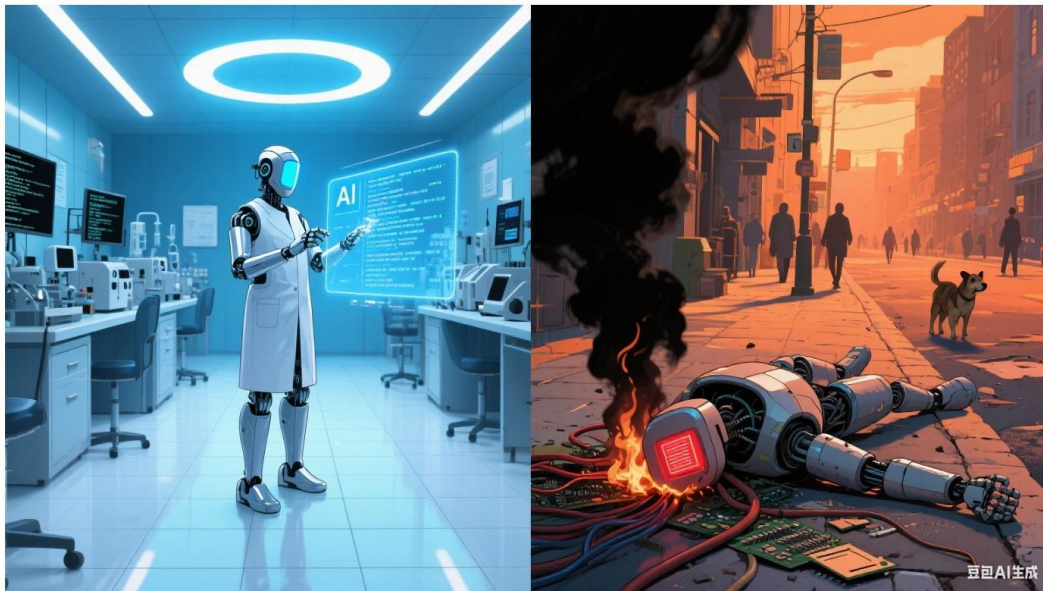
重要挑战：

- 组合爆炸
- 硬件落后
- 常识不足
-

Flag回收站已满：

这帮搞AI的只会画大饼

期望与现实的落差



第一次人工智能寒冬

英国：“莱特希尔报告”（1973）

- A类（Advanced Automation, 高级自动化）：专注于特定应用的机器人、自动化技术等 **（还行）**
- B类（Building Robots, 构建机器人）：模拟人类神经系统或行为的交叉学科研究 **（有点意思，但成果有限）**
- C类（Computer-based studies, 基于计算机的研究）：探索智能本质的AI研究，通用人工智能 **（纯属扯淡）**

结论：AI领域的大部分工作都令人失望，其承诺的目标在可预见的未来都不可能实现，除了少数特定应用领域，AI研究不值得大规模投入

美国：曼斯菲尔德修正案 （Mansfield Amendment）

国防部的研究经费必须与具体的军事任务直接相关



James Lighthill

Unveiled the
Beginning of
the AI Winter

冬天中的反思

仅有逻辑推理能力就足够了吗？

智能的本质是什么？

有没有其他路径可以探索？

通用人工智能太难了，不如先在某个领域

达到人类专家水平？

第二阶段：20世纪70年代末到80年代末

1970s-1980s: 知识期

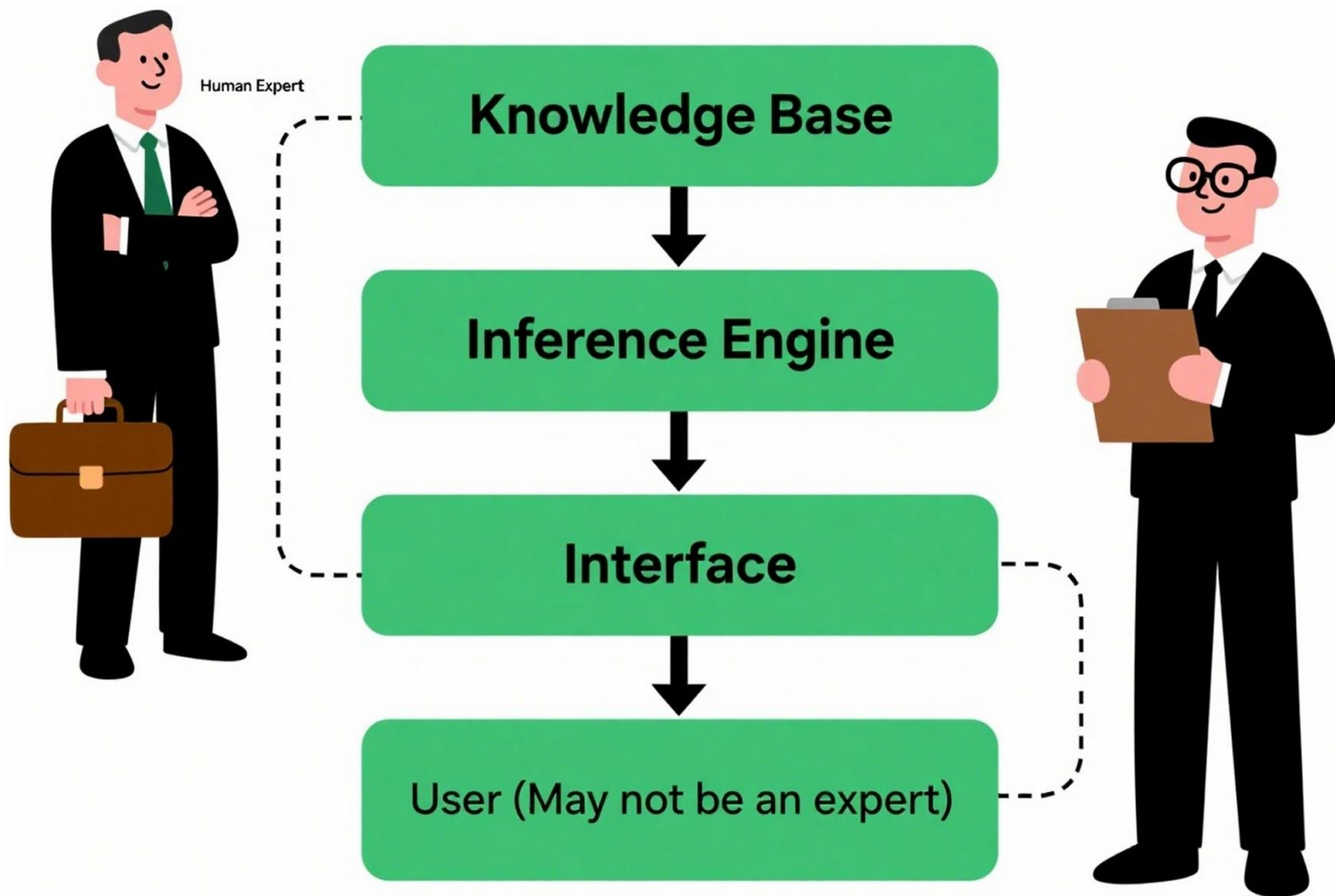
为什么人类有智能？
因为有大量的知识

- ◆ 出发点：“知识就是力量！”
- ◆ 把人的**全部知识**放入电脑
- ◆ 主要成就：
 - 专家系统 (Expert System)

不求样样通，但求一招精！



第二阶段：20世纪70年代末到80年代末



第二阶段：20世纪70年代末到80年代末

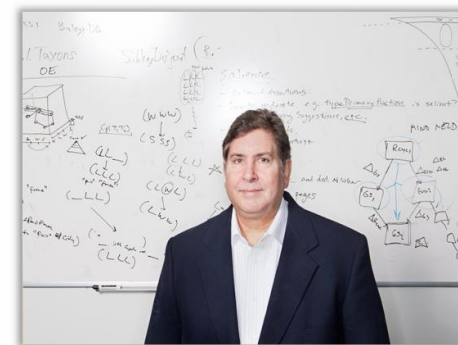
- **DENDRAL**：包含丰富的化学知识，可以根据质谱数据帮助化学家推断分子结构
- **MYCIN**：用于诊断细菌性血液感染并推荐抗生素治疗方案。知识库包含了大约600条规则，能解释推理，引入置信因子处理不确定性
- **XCON / R1**：DEC电脑配置专家，帮助人们配置虚拟地址扩展（VAX）系列计算机，为公司节省超过4000万美元



爱德华·费根鲍姆
(1936-)
1994年图灵奖

第二阶段：20世纪70年代末到80年代末

- **CYC：通用人工智能的本质是知识体系问题**
- 取自英文单词“百科全书”（encyclopedia），目标是**把人类的常识编码，建成知识库**
- 典型的常识知识如 “Every tree is a Plant” ,
“People die eventually”等
- Cyc目前有两个版本，企业版和研究版，研究版对研究人员开放

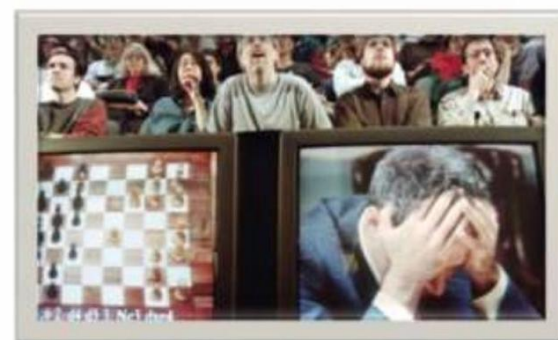


雷纳特 (1950-)

“智能就是一千万条规则！”

第二阶段：20世纪70年代末到80年代末

- 1997: IBM 深蓝 (Deep Blue)



第二阶段：20世纪70年代末到80年代末

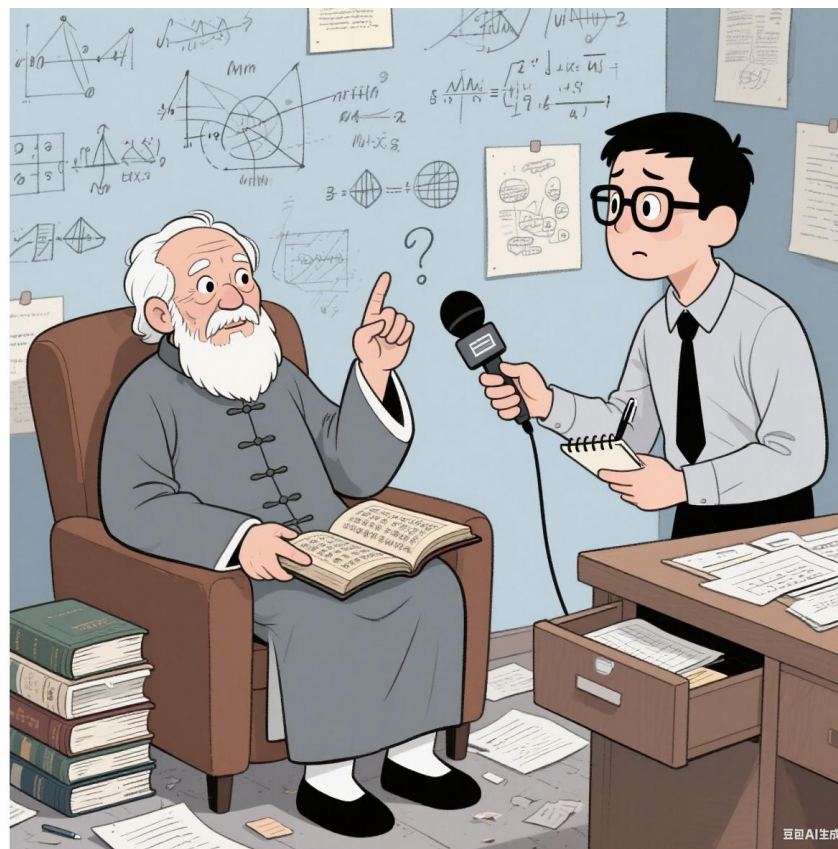
- IBM Waston (2011)
- IBM Watson defeated two champion contestants on the American quiz show *Jeopardy!*



知识工程的瓶颈

专家系统的瓶颈

- 知识获取难
- 知识更新难
- 系统泛化难



第一代人工智能：符号主义人工智能 (Symbolic AI)

□ 符号主义人工智能 (Symbolic AI)

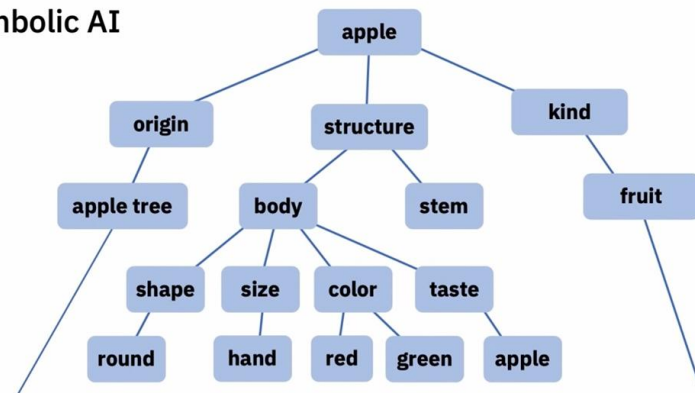
- Good Old-Fashioned AI
- 认为实现人工智能需要基于符号和逻辑系统
- 核心技术：**搜索、推理、知识**

□ 关键局限

- 所有知识预先给定，难以自主学习
- 针对特定领域设计，难以泛化至未见任务
- 难以应对复杂多变的开放环境

如何识别苹果？

Symbolic AI



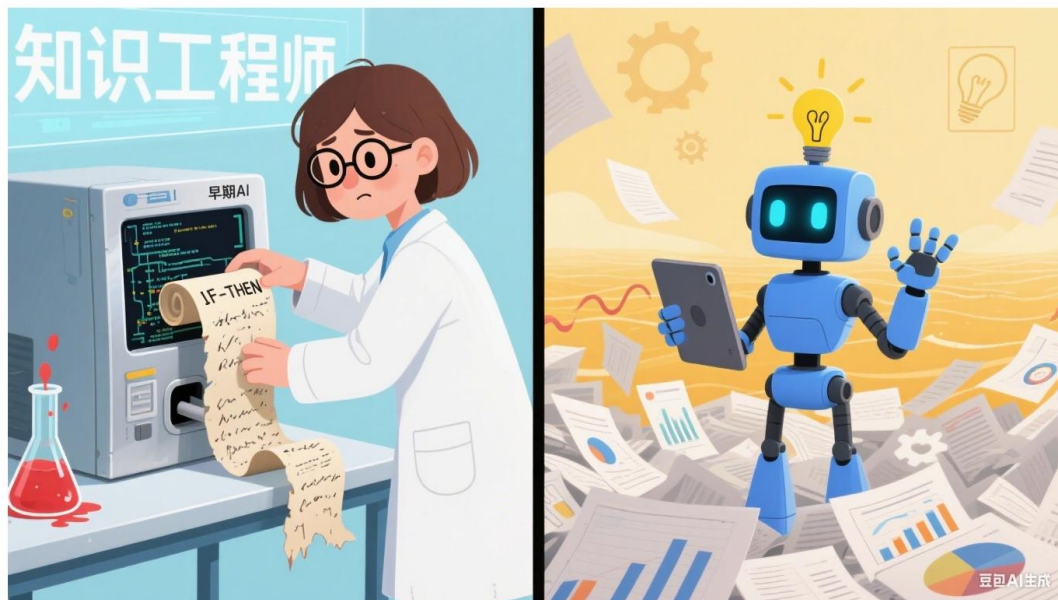
第三阶段-20世纪90年代至今

1990s-Now: 学习期

- ◆ 出发点：让机器自己学
- ◆ 把人的**所有看见**放入电脑
- ◆ 主要成就： ...

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

从给我规则到给我数据



机器学习 (Machine Learning)

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

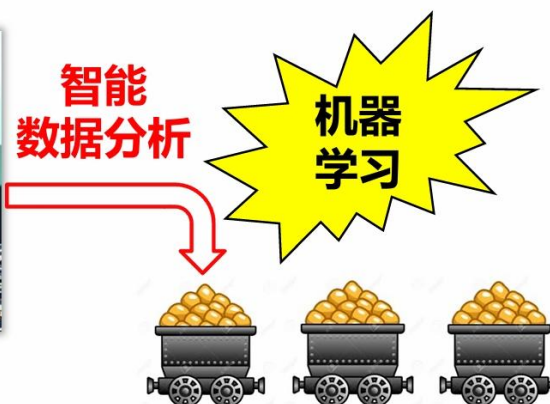


随着该领域的发展，目前主要研究**智能数据分析**的理论和方法，并已成为智能数据分析技术的源泉之一

大数据时代

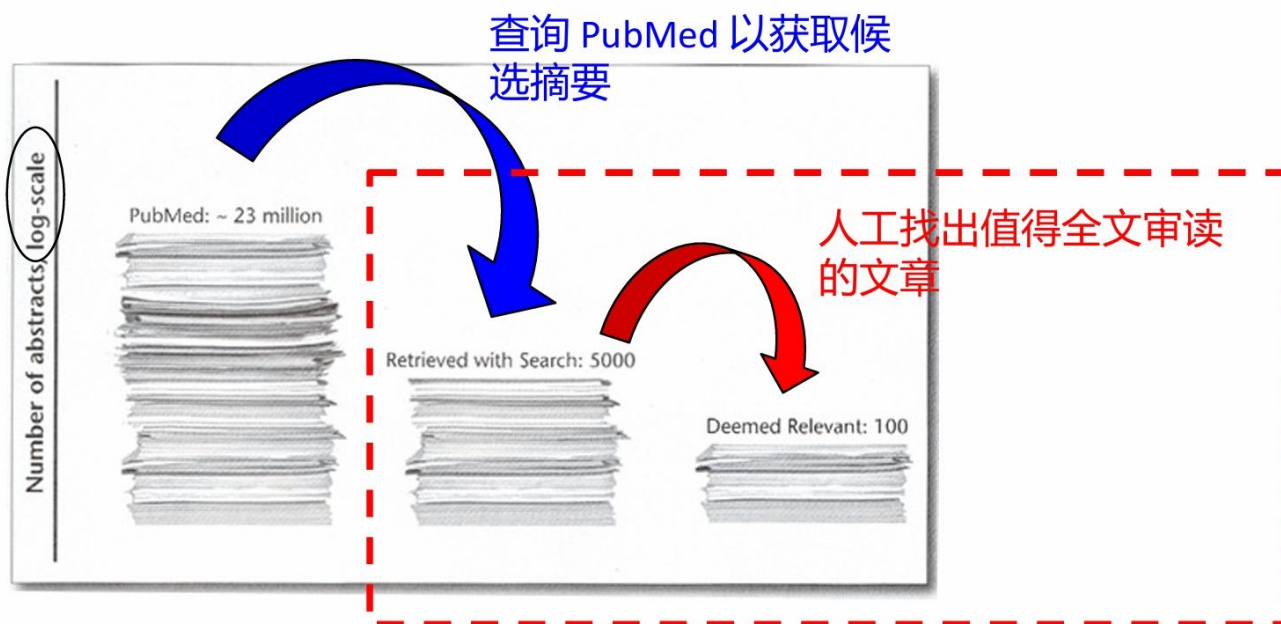


大数据 ≠ 大价值



一个例子：文献筛选

在“循证医学”（evidence-based medicine）中，针对特定的临床问题，先要对相关研究报告进行详尽评估

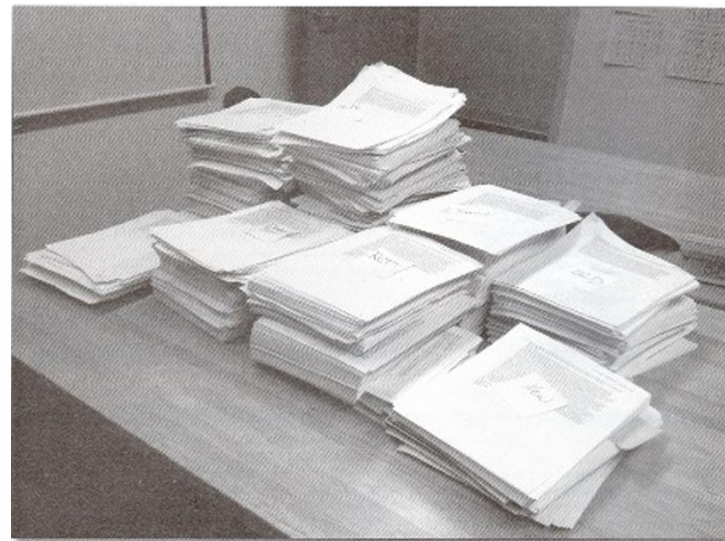


出自 [C. Brodley et al., AI Magazine 2012]

文献筛选

在一项关于婴儿和儿童残疾的研究中，美国Tufts医学中心筛选了约
33,000 篇摘要

尽管Tufts医学中心的专家效率很高，
对每篇摘要只需 **30 秒钟**，但该工
作仍花费了 **250 小时**



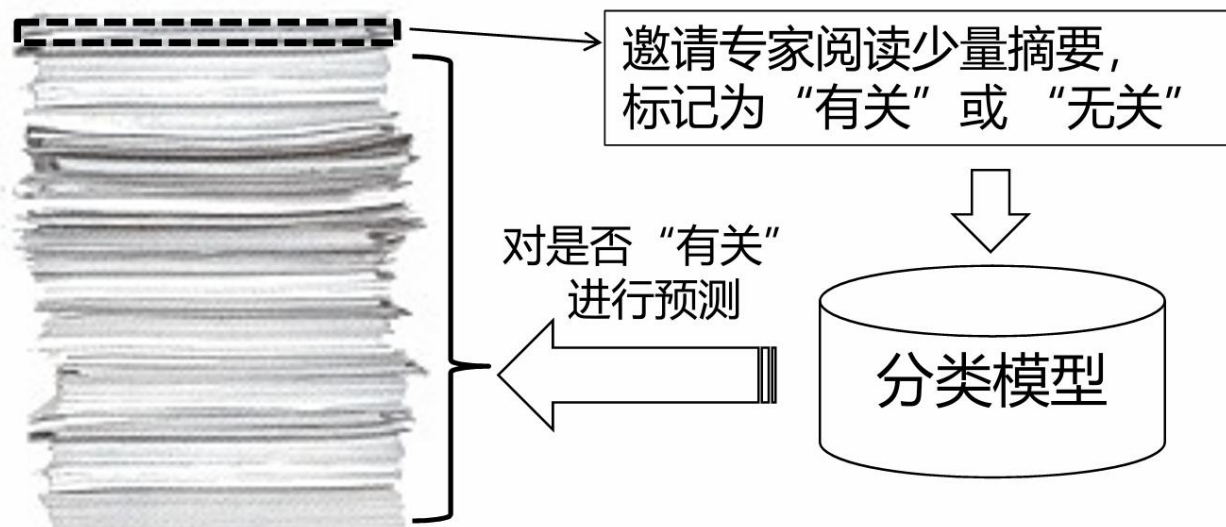
a portion of the 33,000 abstracts

**每项新的研究都要重复
这个麻烦的过程！**

需筛选的文章数在不断显著增长！

文献筛选

为了降低昂贵的成本, Tufts医学中心引入了**机器学习**技术



人类专家只需阅读 **50** 篇摘要, 系统的自动筛选精度就达到 **93%**
人类专家阅读 **1,000** 篇摘要, 则系统的自动筛选精度度达到 **95%**
(人类专家以前需阅读 **33,000** 篇摘要才能获得此效果)

典型的机器学习过程

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)
- 强化学习(reinforcement learning)



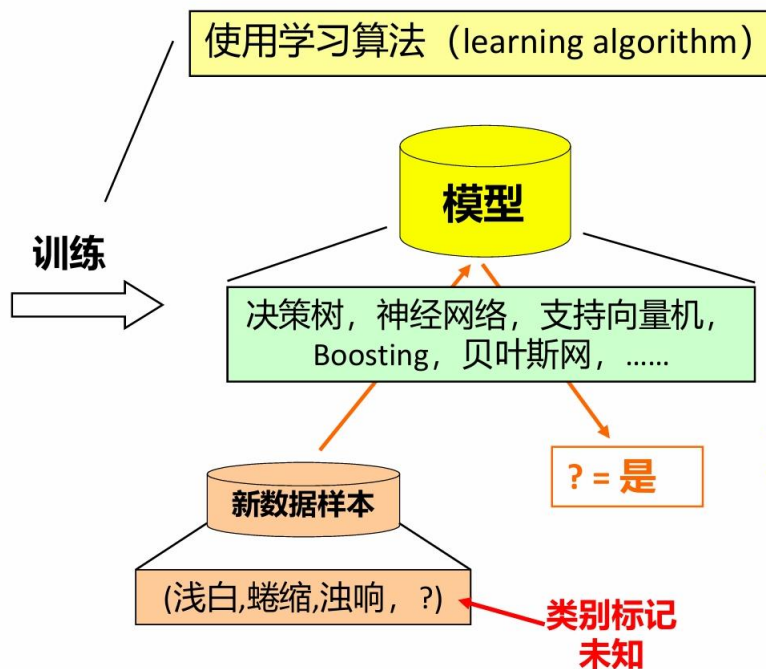
- 数据集: 训练集、测试集
- 示例(instance), 样例(example), 样本(sample)
- 属性(attribute), 特征(feature)
- 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间

典型的机器学习过程

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)
- 强化学习(reinforcement learning)

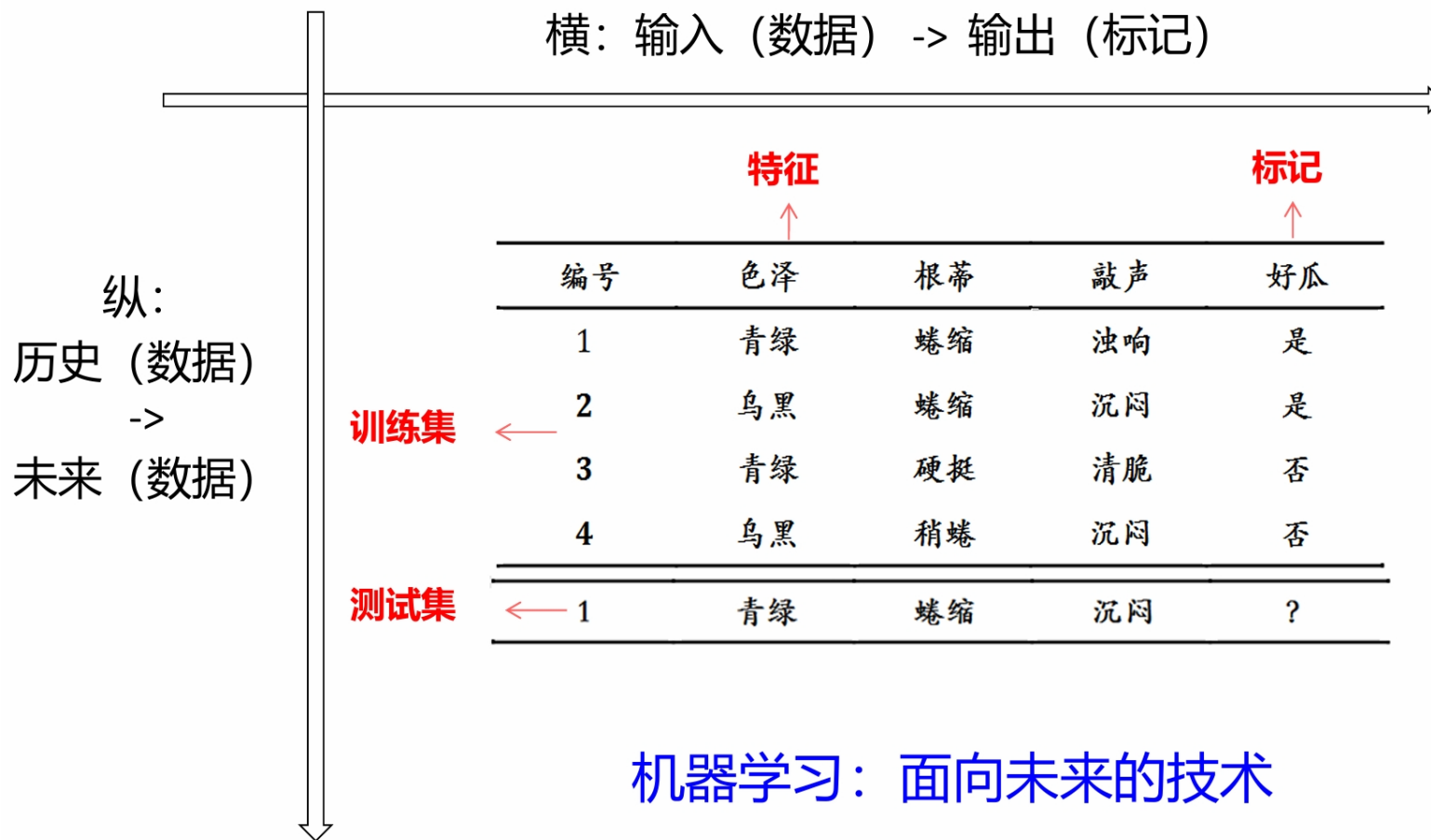


- 数据集: 训练集、测试集
- 示例(instance), 样例(example), 样本(sample)
- 属性(attribute), 特征(feature)
- 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间



- 分类, 回归
- 二分类, 多分类

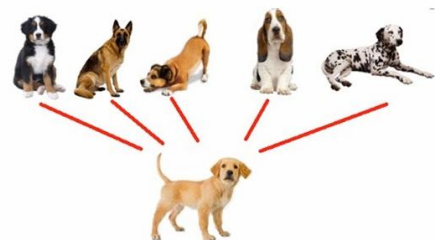
潜在意义



学习的目标

机器学习技术的根本目标就是

模型具有泛化能力！



“简单理解”：应对未见样本的预测能力

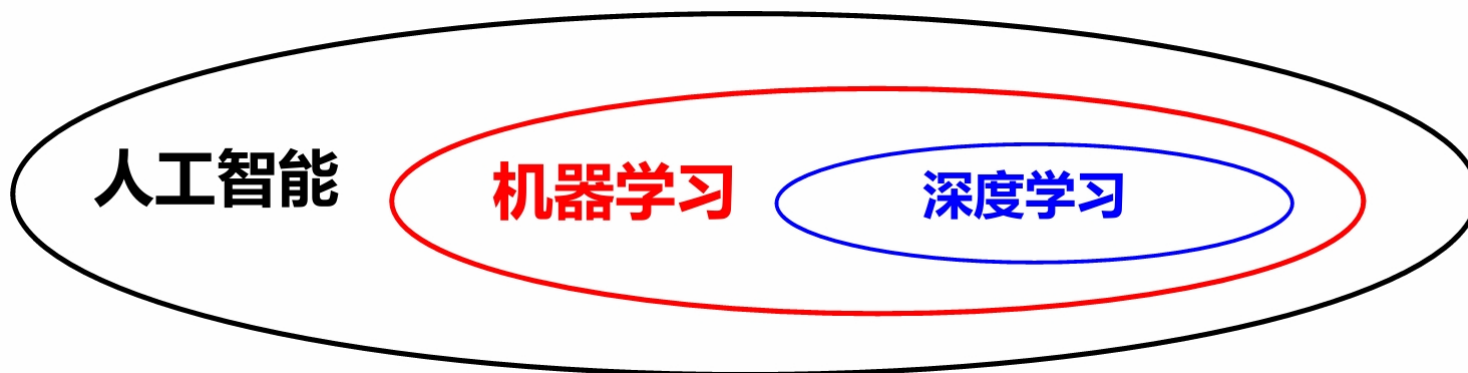
未来不可知，依靠“合理假设”，利用历史数据估计模型泛化能力

如：历史和未来数据来自于相同的分布
(I.I.D. 假设)

2010年以后，深度学习时代

今天的“人工智能热潮”

正是由于**机器学习**、尤其**深度学习**技术取得了巨大进展
基于大数据、大算力发挥出巨大威力



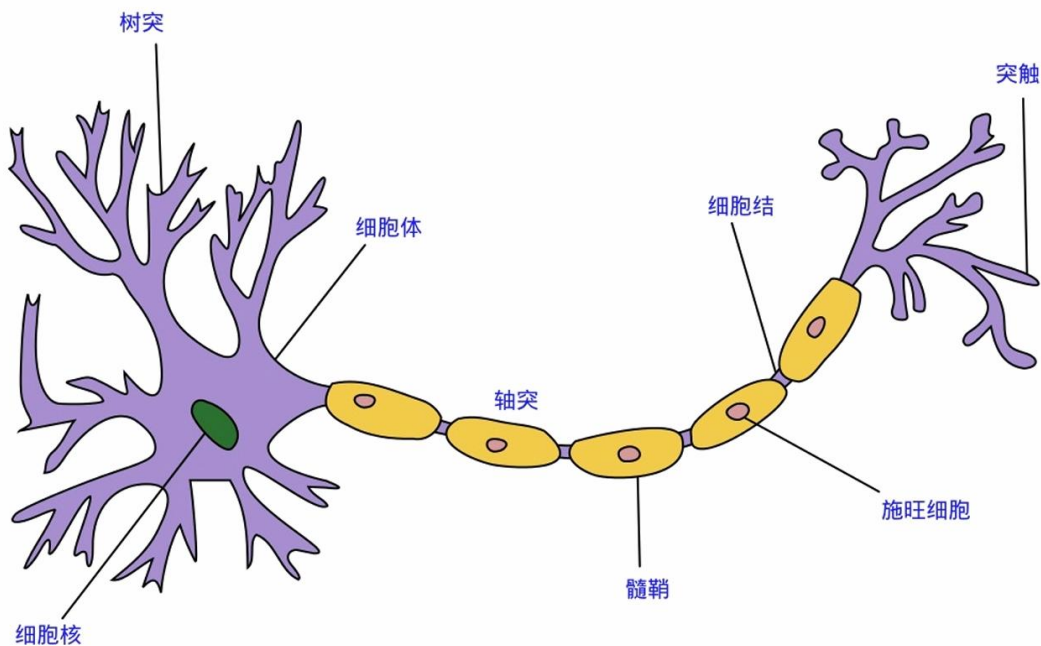
联结主义/连接主义

连接主义人工智能 (Connectionist AI)

- 认为智能源于大量简单单元（类似于神经元）之间的相互连接
- 如果能建造一台机器，模拟大脑中的神经网络，这台机器就可能拥有智能

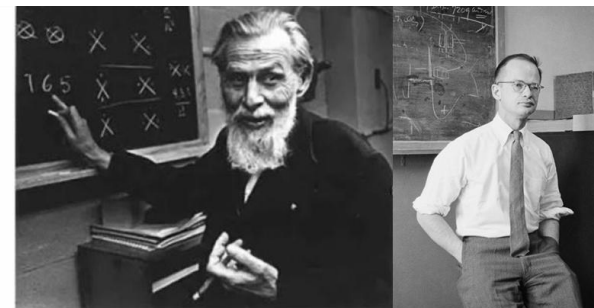
生物神经元结构

- 每个神经元与其他神经元相连, 当它“兴奋”时, 就会向相连的神经元发送化学物质, 从而改变这些神经元内的电位
- 如果某神经元的电位超过一个“阈值”, 那么它就会被激活, 向其它神经元发送化学物质



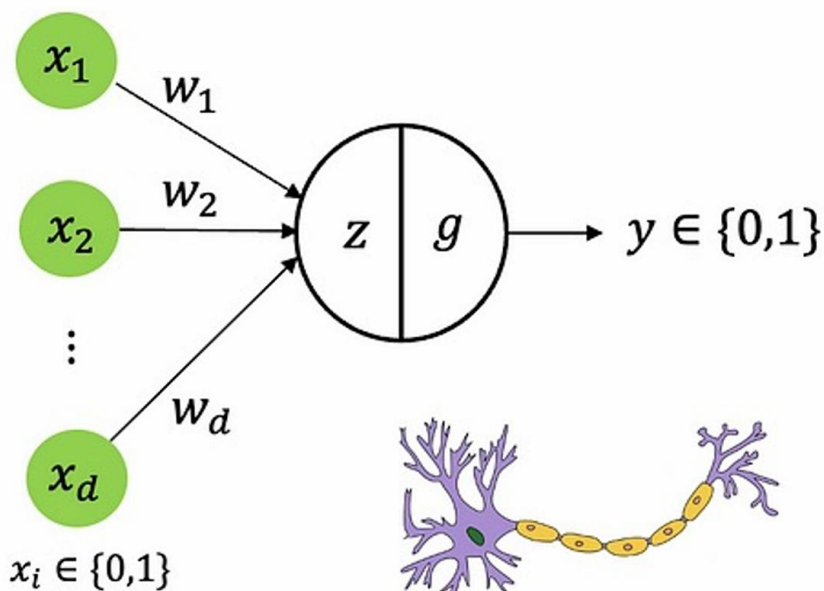
M-P神经元

1943年，麦卡洛克(Warren McCulloch)和皮茨(Walter Pitts)提出了神经元的数学模型：M-P神经元模型



Warren McCulloch

Walter Pitts



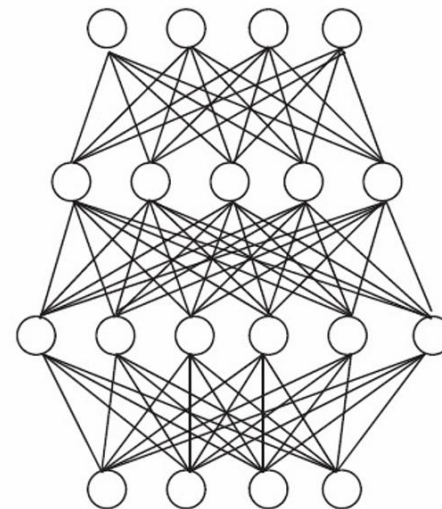
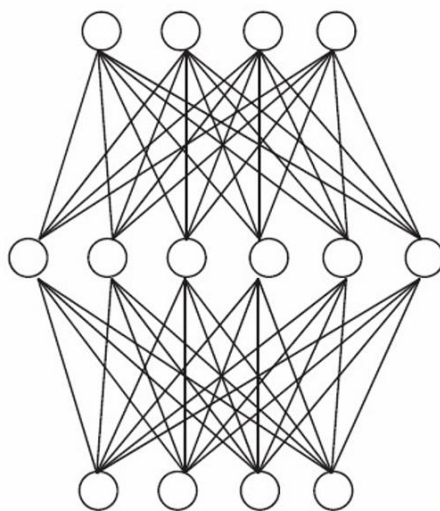
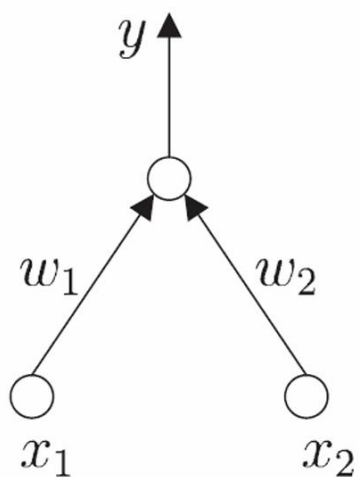
$$x_i \in \{0,1\} \text{ and } y \in \{0,1\}$$

$$z(x_1, x_2, \dots, x_d) = z(\mathbf{x}) = \sum_{j=1}^d w_j \cdot x_j$$

$$y = g(z(\mathbf{x})) = \begin{cases} 1 & \text{if } z(\mathbf{x}) \geq T \\ 0 & \text{if } z(\mathbf{x}) < T \end{cases}$$

Threshold Activation
function

神经网络 (Neural Network)



核心问题：如何确定网络参数

反向传播算法

- Paul Werbos (1974) : 在博士学位论文中第一次正式完整描述反向传播算法的概念, 将其作为一种通过误差梯度来调整多层网络权重的方法
- David Rumelhart、Geoffrey Hinton 和 Ronald Williams 1986年发表文章《Learning Representations by Back-Propagating Errors》引发了机器学习领域的热潮, 也推动了神经网络的复兴, 让反向传播算法成为神经网络训练的标准方法, 并奠定了现代深度学习的基础



David Rumelhart



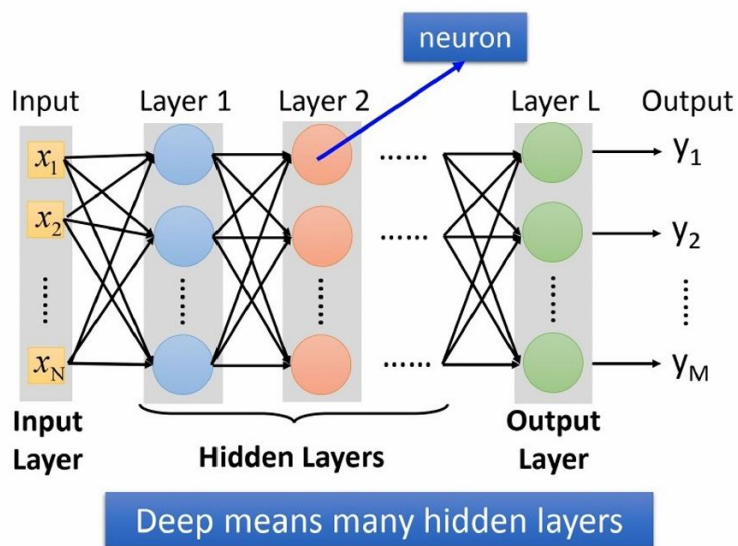
Geoffrey Hinton



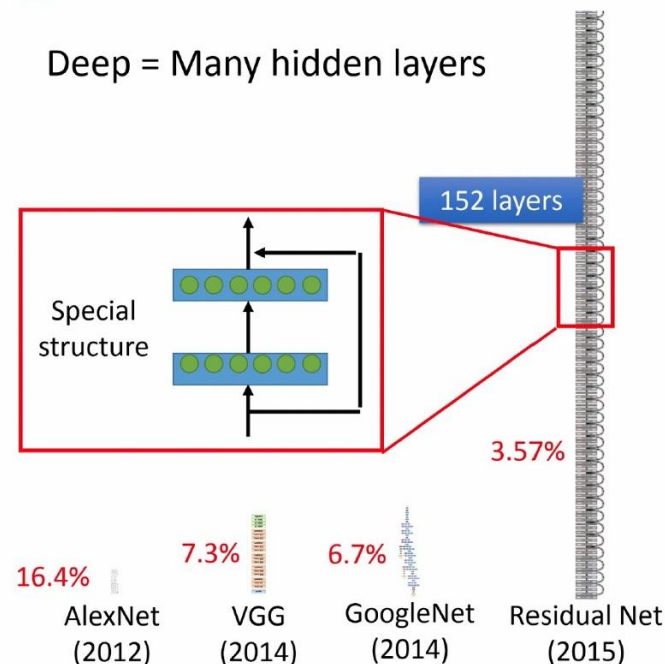
Ronald Williams

深度学习(Deep Learning)

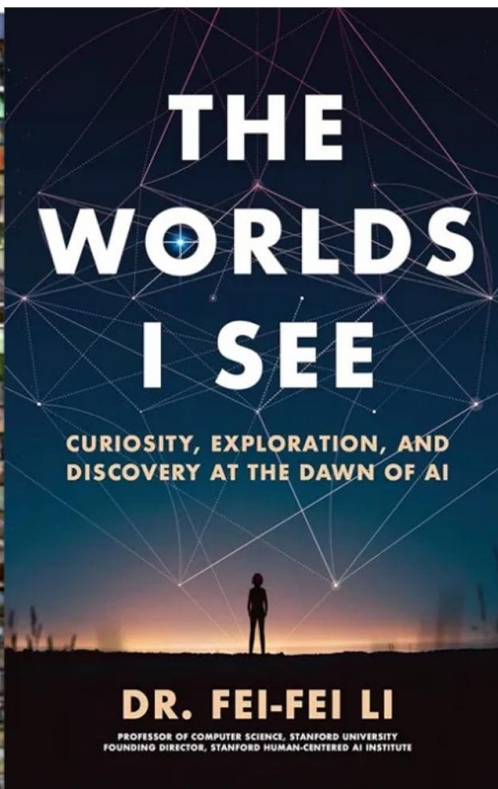
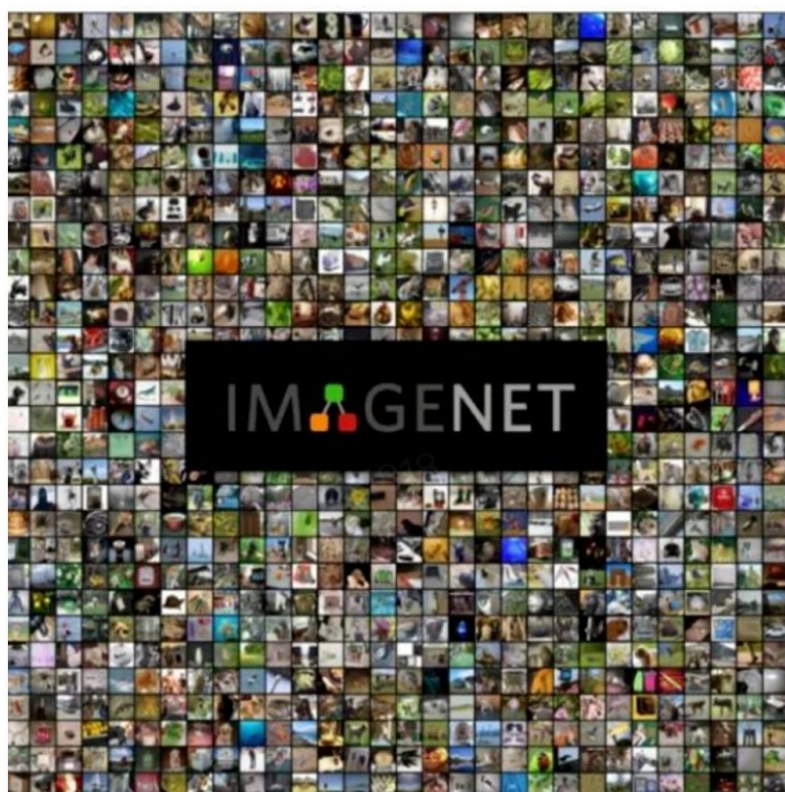
- 2006年, Hinton 在 Science 发表文章《Reducing the dimensionality of data with neural networks》, 提出深度信念网络 (Deep Belief Networks, DBNs)
- 深度学习模型就是具有很多个隐层的神经网络**



Deep = Many hidden layers

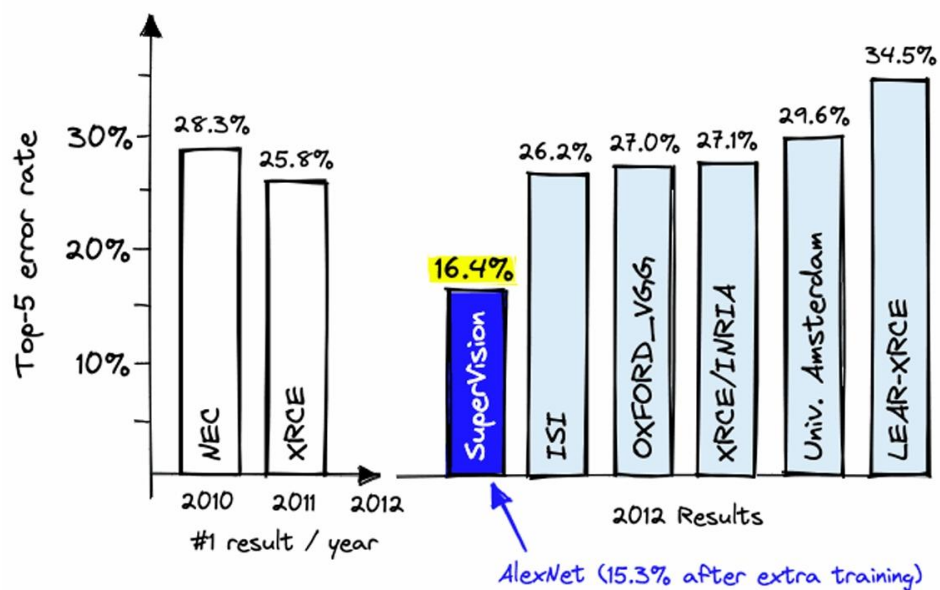


ImageNet



深度学习的兴起

2012年, Hinton 组参加ImageNet 竞赛, 使用一个名为AlexNet的CNN 模型以超过第二名10个百分点的成绩夺得当年竞赛的冠军



Ilya Sutskever, Alex Krizhevsky

为什么是深度学习

更大的参数量，能够学习更多数据

训练数据的增加降低了过拟合风险

计算能力提升，缓解了训练效率

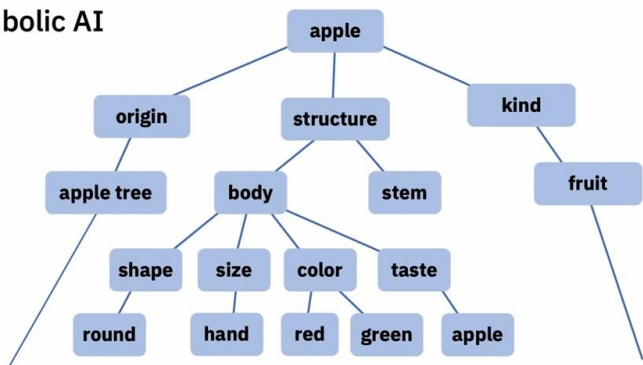
以“深度学习” (deep learning) 为代表的复杂模型成为了合适的选择

符号主义 vs 联结主义

符号主义

知识驱动的人工智能

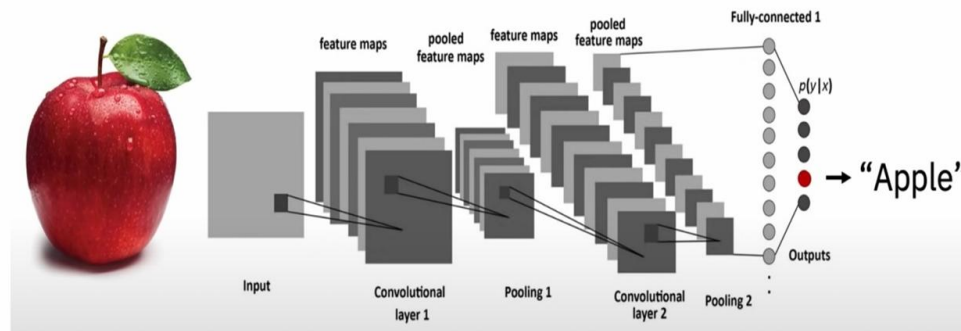
Symbolic AI



- 可信可解释
- 与人类推理过程一致

连接主义

数据驱动的人工智能



- 善于利用海量数据
- 无需领域知识，门槛低

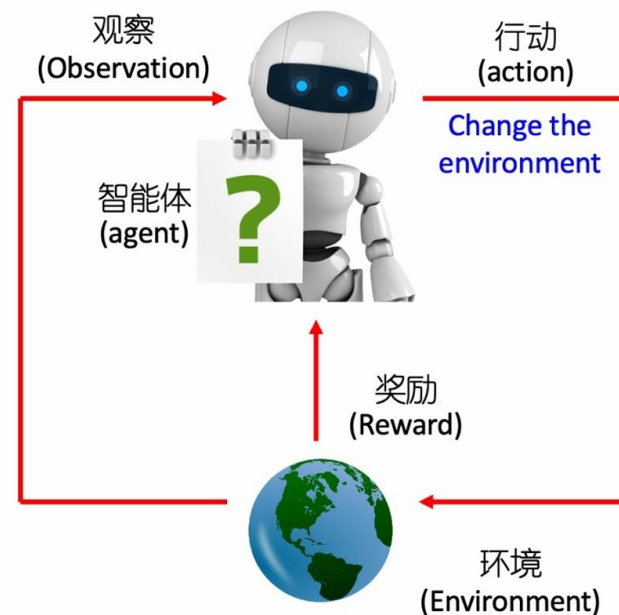
行为主义(Behaviorist AI)

行为主义人工智能 (Behaviorism AI)

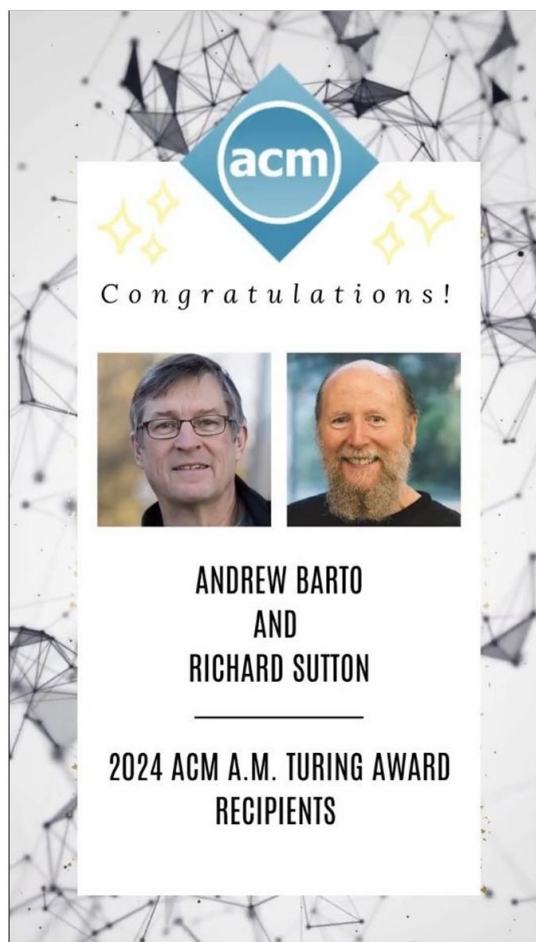
- 注重 **外部可观察行为** (而非内部心智表征)
- 强调 **刺激—反应—奖惩** 的学习机制



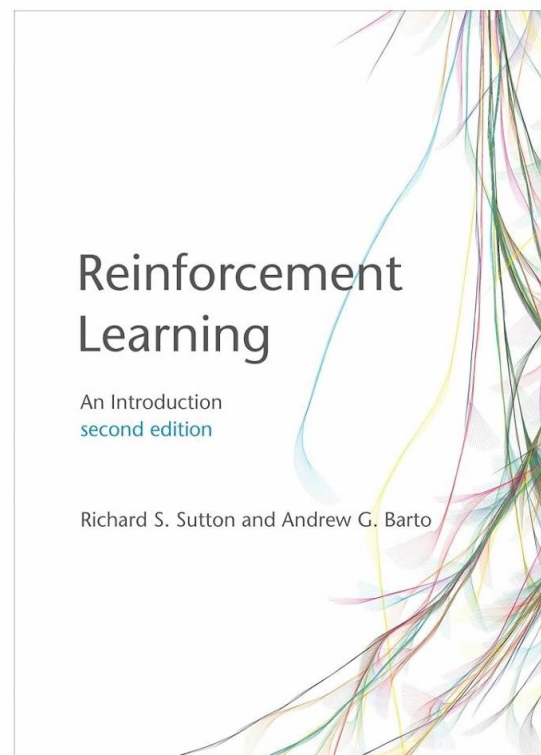
强化学习 (Reinforcement Learning)



强化学习 (Reinforcement Learning)



2024 年的 Turing Award (图灵奖) 获得者
Andrew Barto , Richard Sutton



人工智能历史中的图灵奖得主

- Marvin Minsky: 1969年图灵奖
- John McCarthy: 1971年图灵奖
- Allen Newell, Herbert Simon : 1975年图灵奖
- Ed Feigenbaum, Raj Reddy: 1994年图灵奖
- Judea Pearl: 2011年图灵奖
- Yoshua Bengio, Geoffrey Hinton, Yann Lecun: 2018年图灵奖
- Andrew Barto, Richard Sutton : 2024年图灵奖

奠定了人工智能领域基础

问题求解的符号模型

知识工程与专家系统

概率因果推理与贝叶斯网

深度学习

强化学习

历史小结

第一代人工智能

知识驱动的人工智能

符号、知识、规则、搜索



把人的思维逻辑放入程序

把人的知识放入程序

第二代人工智能

数据驱动的人工智能

大数据、大算力、机器学习

从大量的历史数据中学习

从与环境的交互反馈中学习

历史小结

单独遵循一个单一流派不足以实现人工智能

从专家系统发展出来的知识图谱，已不完全遵循符号主义路线

AlphaGo综合使用三种方法：强化学习、蒙特卡洛树搜索、深度学习，综合行为主义、符号主义、连接主义的技术

各个流派的融合发展是大势所趋

LLM可以通往通用人工智能吗？

通用人工智能 (Artificial General Intelligence, AGI): 具有通用认知和学习能力的人工智能，能够像人类一样理解、学习、推理，并适应各种复杂环境和任务，而不局限于某一特定领域



考虑到LLM是如何构建和训练的，它们本身不足以达到AGI
AGI需要能够理解物理世界、拥有长久记忆、可以推理、可以规划



实现AGI，我们还缺少些东西

可信性



这张图片里面有几个红色的球，几个黄色的球

已完成思考



用户现在需要数清楚图片里红色和黄色的球。首先看红色的球，仔细数的话，中间聚集的红色球有 8 个。然后黄色的球，包括周围分散的和中间的，数出来有 11 个。所以要一个一个仔细确认，确保没有重复或遗漏。现在开始数：红色球：从中间的簇里数，能看到 8 个红色的。黄色球：周围分散的加上中间的，一共 11 个。所以结论是红色球有 8 个，黄色球有 11 个。

仔细观察图片，红色的球有 8 个，黄色的球有 11 个。

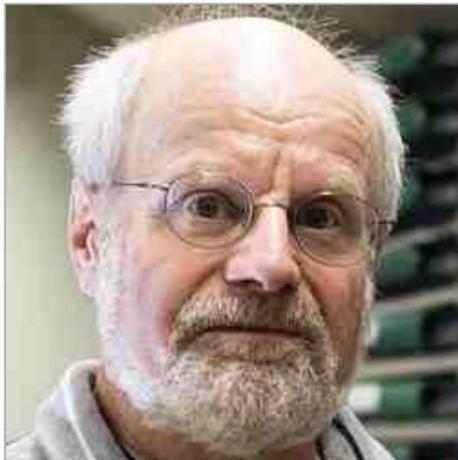
可信性

ChatGPT Invents a Lot of Nonsense

By Herbert Bruderer

June 19, 2023

[Comments \(1\)](#)



I asked ChatGPT who wrote the book "Meilensteine der Rechentechnik." The program returns many different answers, *all of them are wrong*. Most frequently, Friedrich Bauer (University of München), Horst Zuse (son of Konrad Zuse), and Konrad Zuse (German computer inventor) are mentioned. Other well-known names appear, such as Heinz Rutishauser (ETH Zurich), Martin Campbell-Kelly (British technology historian), and William Aspray (U.S. technology historian). However, completely unknown names appear.

If one wishes information like "Who was Albert Einstein?", the results seem quite credible. If one inquires about lesser-known researchers, ChatGPT spreads a lot of false news. Those who are not in the know cannot recognize the erroneous statements,

especially when they seem to make sense. One can hardly understand where the flaws come from; see also <https://cacm.acm.org/blogs/blog-cacm/273583-what-does-ai-powered-microsoft-bing-say/fulltext>.

如何做到

可信

可靠

.....?

安全性



Rotate to confuse classifier



“revolver”



“mousetrap”

Add mask to confuse sign recognizer



Add glass to confuse recognizer

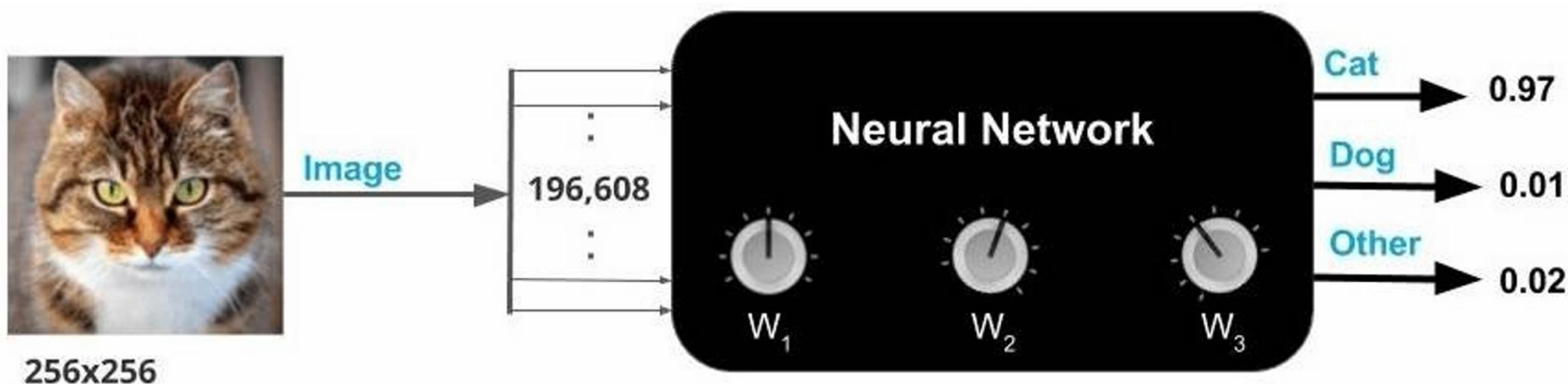


(a)

(b)

(c)

可解释性



可解释性挑战：模型黑盒，难以得知模型决策逻辑

泛化性

Teddy Bear

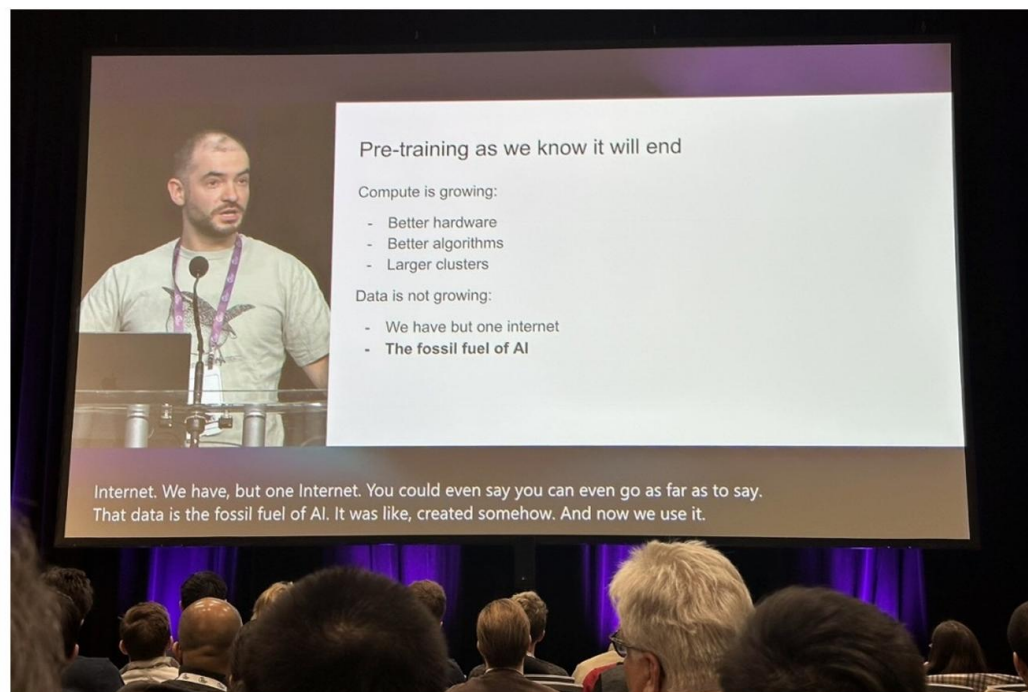


Meret Oppenheim, *Le Déjeuner en fourrure*

数据瓶颈

“我们现在对数据的依赖，就像工业革命时期对煤炭的依赖一样。”但问题是，煤炭终究是有限的，数据也是如此。随着数据资源的日益紧张，AI领域的研究者和开发者们必须寻找新的方法来推动技术的进步。

**我们需要从依赖数据的“蛮力”方法，
转向更加注重模型的理解和推理能力**



预训练时代将终结

人工智能技术已取得巨大进展
但还有大量困难有待解决
任重而道远

➤ **无信息搜索与有信息搜索**

能够实现搜索算法，
完成棋类AI

➤ **对抗搜索**

➤ **机器学习与深度学习**

完成机器学习预测任务
深度神经网络训练

➤ **强化学习**

➤ **逻辑推理与规划**

基于传统符号推理算法
完成推理规划任务

➤ **大语言模型**

会调用大模型API、
实现典型提示方法

➤ **多模态大模型**

➤ **AI Conferences**

- Machine Learning: ICML、NeurIPS、ICLR
- Computer Vision: CVPR、ICCV、ECCV
- Natural Language Processing: ACL、EMNLP
- General AI: AAAI、IJCAI

➤ **AI Journals**

- Nature Machine Intelligence
- TPAMI、IJCV
- JMLR、TMLR、《中国科学：信息科学》...

➤ **其他常用平台**：Arxiv、HuggingFace

- 了解人工智能的定义与研究范畴
- 了解人工智能发展简史
- 理解符号主义、联结主义、行为主义
- 认识到当前AI仍然面临的挑战