南京大学
NANJING UNIVERSITY

# 人工智能导论

## Introduction

郭兰哲

南京大学 智能科学与技术学院

https://www.lamda.nju.edu.cn/guolz/IntroAI/fall2025/index.html

Email: guolz@nju.edu.cn

# 大纲

- 课程信息
- 什么是人工智能
- 人工智能简史
- 本课程内容

# 课程定位

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

# 课程信息

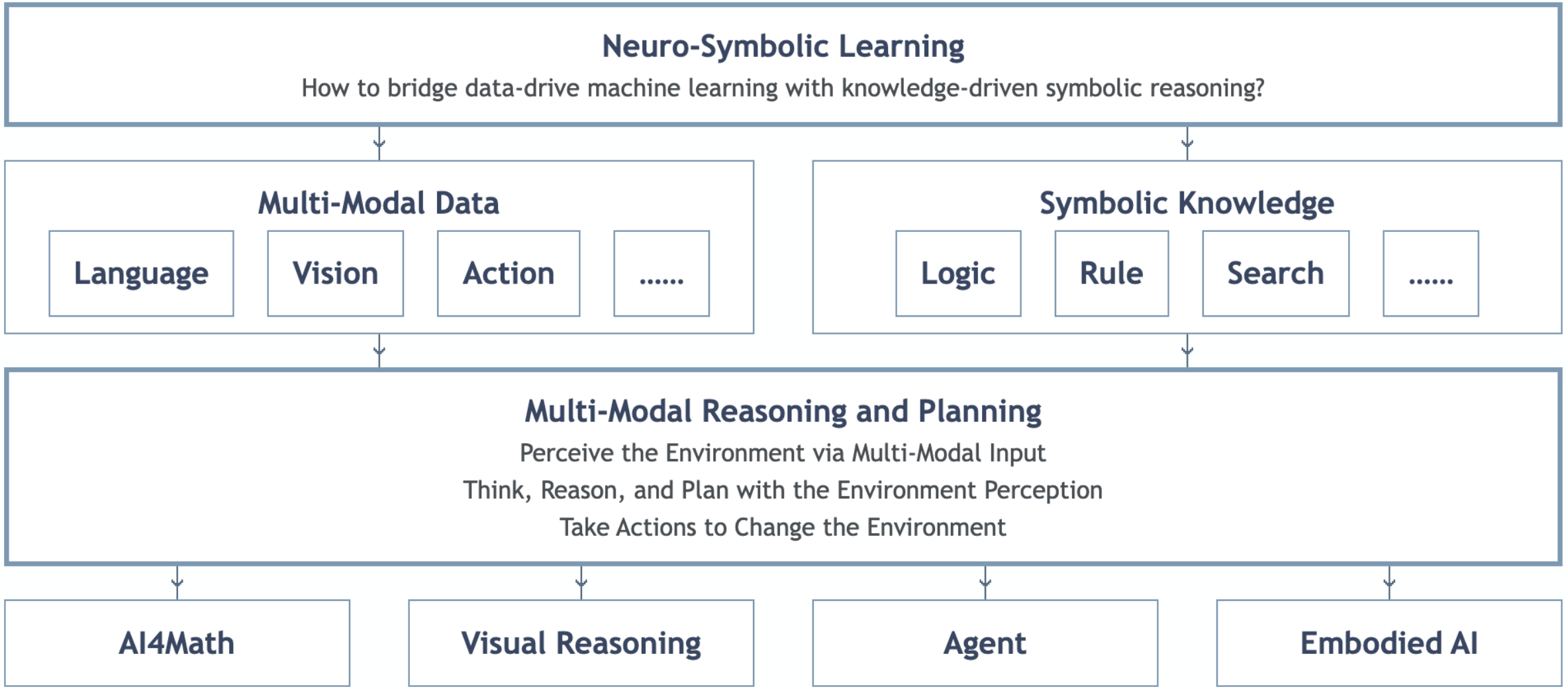
- 上课时间：
  - ✓ 周五2-4节 1-18周 南雍-西 209
- 课程主页：
  - ✓ http://www.lamda.nju.edu.cn/guolz/introai.html
  - ✓ 课程讨论QQ群：285851583

# 课程信息

- 主讲教师：
  - ✓ 郭兰哲，准聘助理教授，博士生导师
  - ✓ 智能科学与技术学院
  - ✓ 机器学习与数据挖掘研究所（LAMDA）
  - ✓ 研究方向：神经符号学习、大模型/多模态推理与规划
  - ✓ 邮箱：guolz@nju.edu.cn
  - ✓ 个人主页：https://www.lamda.nju.edu.cn/guolz/
  - ✓ 办公室：南雍楼-东523

# 课程信息

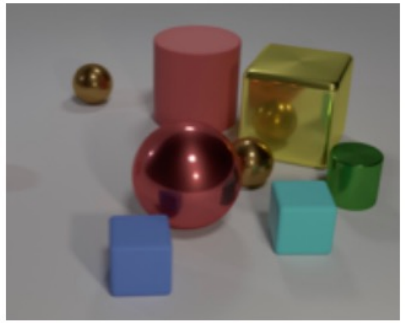
- 研究方向：神经符号学习、大模型/多模态推理与规划



# 课程信息

- 研究方向：神经符号学习、大模型/多模态推理与规划

## 视觉推理



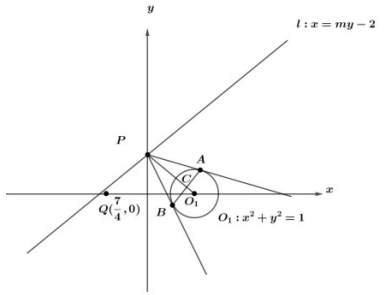
What is the shape of the object closest to the large cylinder?

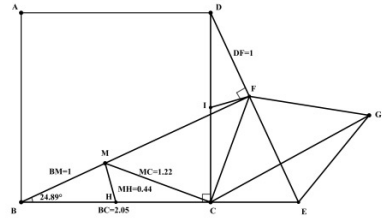
Are there more trees than animals?

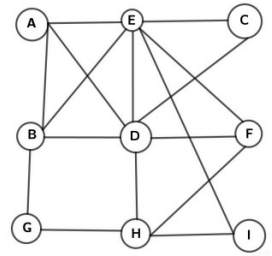
How many blocks are on the right of the three-level tower?

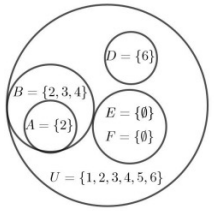
Will the block tower fall if the top block is removed?

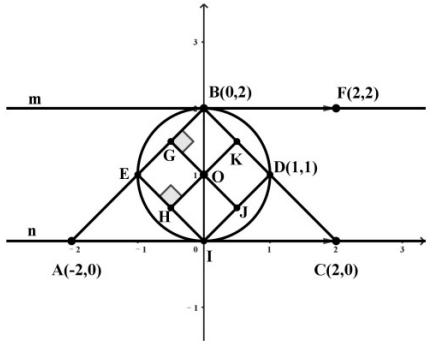
## 数学推理

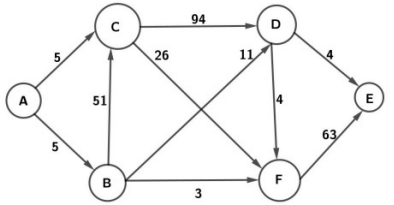


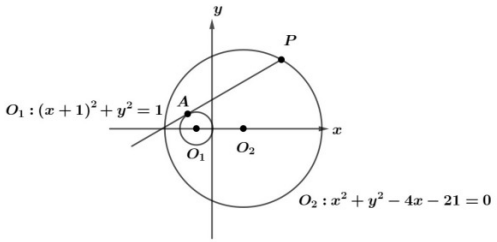



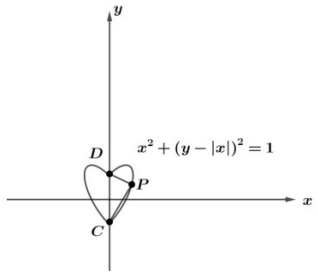



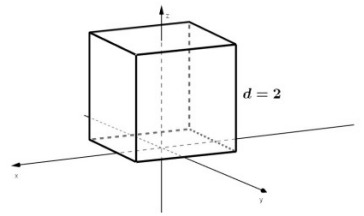



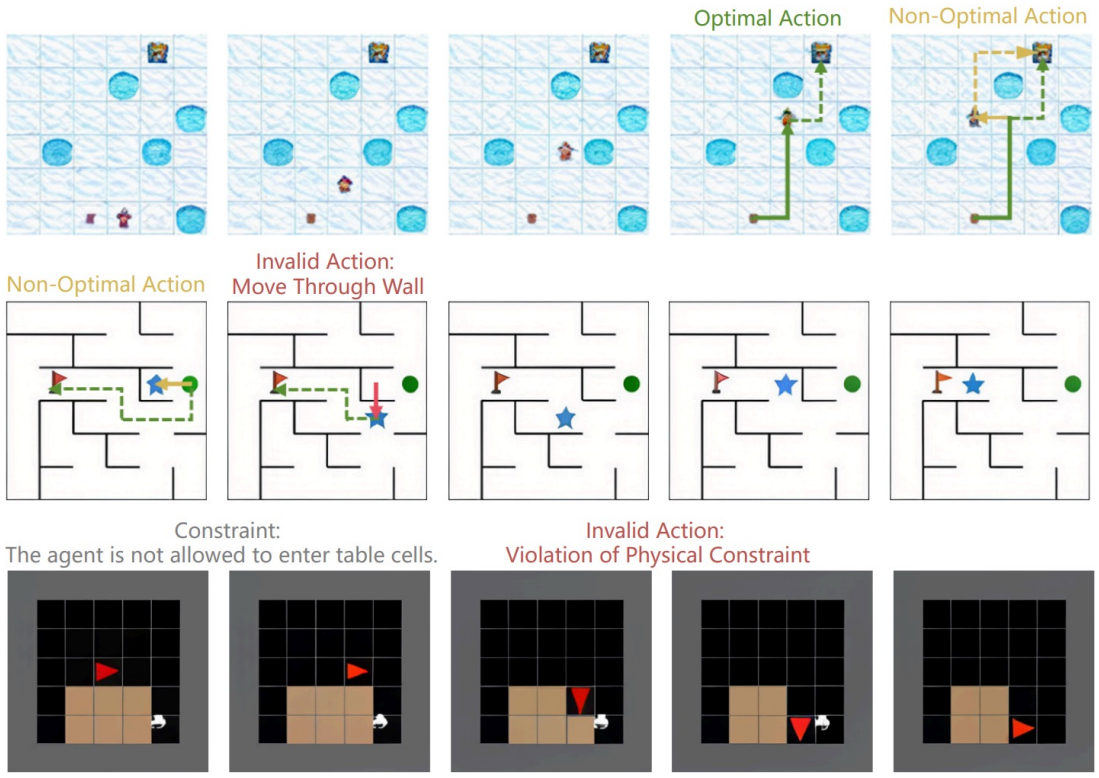



# 课程信息

- 研究方向：神经符号学习、大模型/多模态推理与规划

## 视觉规划

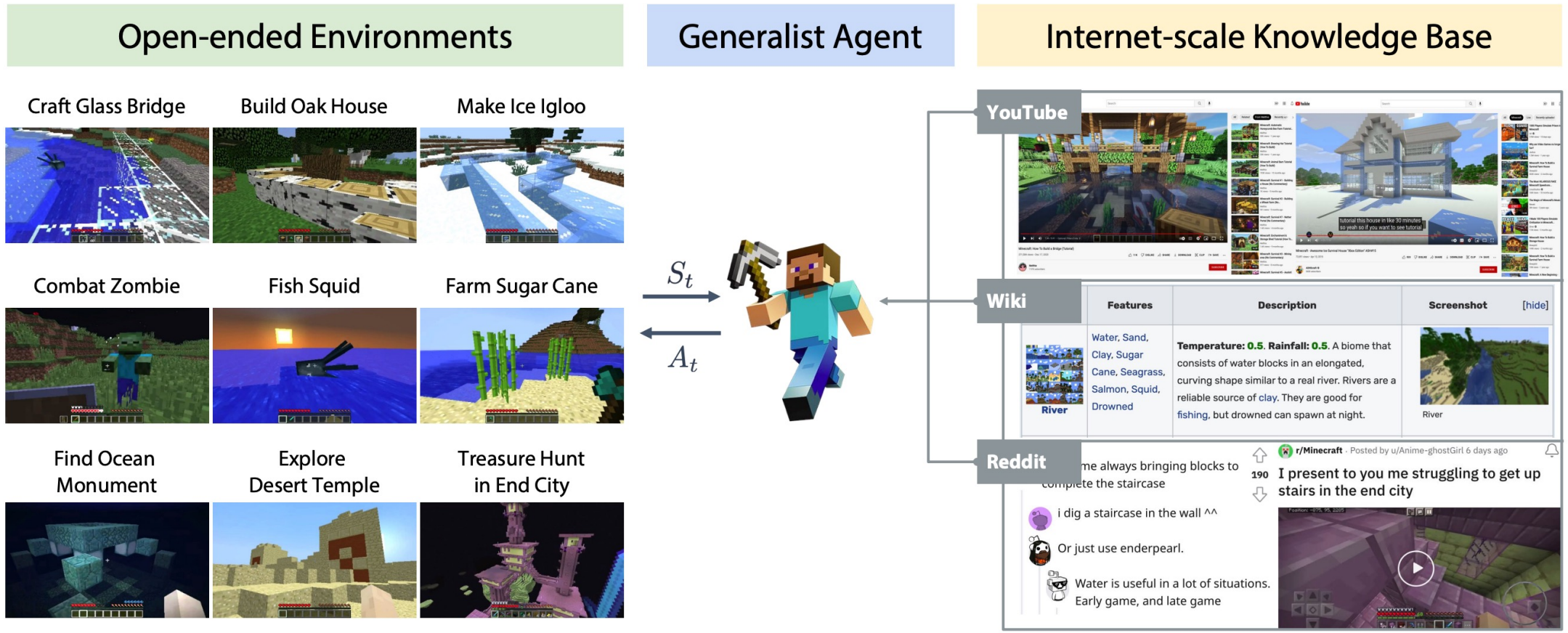



## 具身规划

# 课程信息

- 研究方向：神经符号学习、大模型/多模态推理与规划

## 多模态复杂任务规划的游戏智能体



# 课程信息

- 研究方向：神经符号学习、大模型/多模态推理与规划

**多模态复杂任务规划的游戏智能体**

# 课程信息

## 课程助教

葛凌岳

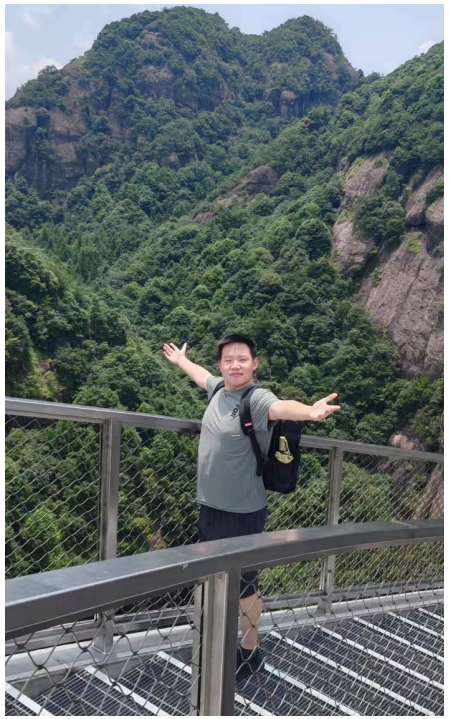
211300025@smail.nju.edu.cn


朱睿

zhurui@smail.nju.edu.cn

吴伟铭

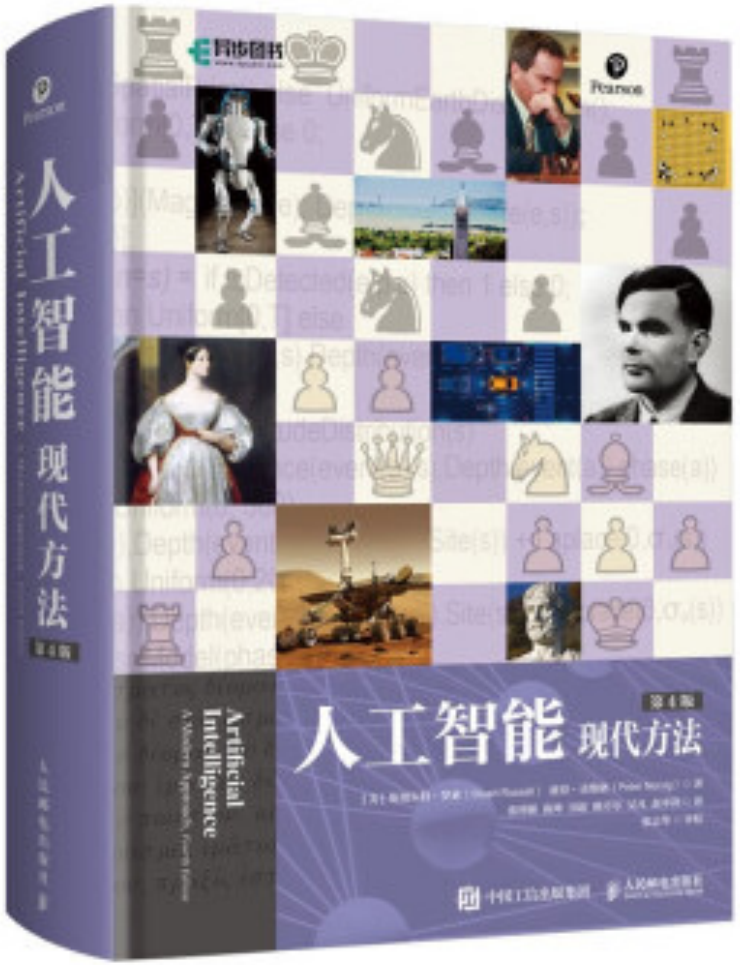
wuwm23@smail.nju.edu.cn



叶晋

231880419@smail.nju.edu.cn

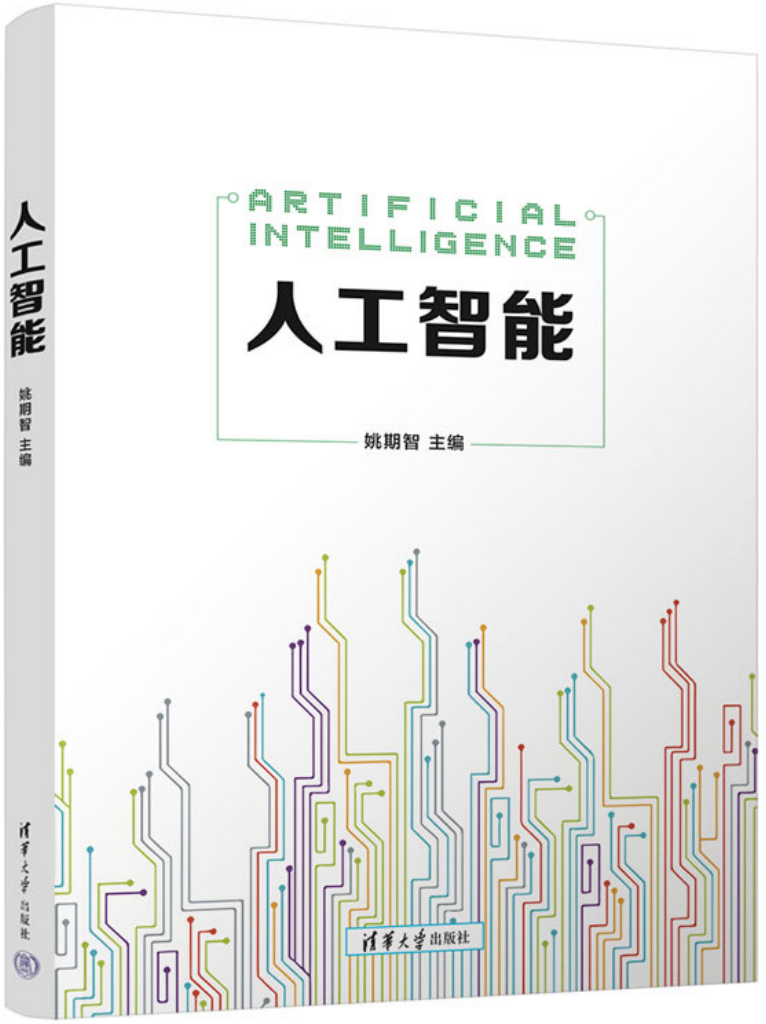
# 参考书籍




《人工智能：一种现代的方法》
Stuart J. Russell, Peter Norving

# 参考书籍


人工智能的底层逻辑
张长水◎著
清华大学出版社


ARTIFICIAL INTELLIGENCE
人工智能
姚期智 主编
清华大学出版社

# 课程考核

- 成绩核算：
  - ✓ 没有期中、期末考试
  - ✓ 平时作业：理论+实践，60%
  - ✓ 课程设计：自主选题，组队完成，40%
  - ✓ bonus：课堂问答奖励3分

# 课程考核

- 课程设计：
  - ✓ 寻找人工智能相关的科研或工程项目，可以通过阅读论文、产业报告、人工智能相关的竞赛确定题目
  - ✓ 创新性30%，完成度30%，答辩情况20%，项目报告20%

- 建议：
  - ✓ 结合个人兴趣，努力研发具备实用性的项目
  - ✓ 希望以大创、竞赛为目标
  - ✓ 每个队伍不超过5人，分工明确，避免划水

# 课程考核

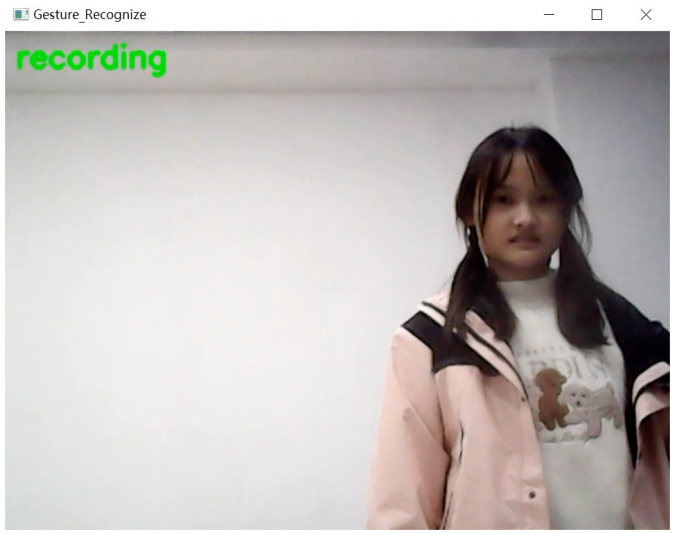
- 时间安排：
  - ✓ 第一周，自由组队，每组不超过5人，选定小组长，**下周三前**由小组长把小组成员信息发送至葛凌岳助教邮箱，**注明组长信息**
  - ✓ 第二至三周，组内成员讨论并自拟课题，与老师协商后确认课题
  - ✓ 通力合作完成课程项目，交付：
    - ✓ 答辩PPT、项目成果演示视频（如有）、项目报告PDF
  - ✓ 项目答辩

# 课程考核

- 课程设计参考：
  - ✓ 游戏AI：棋牌、星际争霸、王者荣耀……
  - ✓ 垂类大模型：医疗、法律、金融……
  - ✓ Agent：旅行规划、教育……
  - ✓ 在领域评测基准上做前沿论文复现与优化
    - ✓ AI4Math、视觉推理、具身智能、游戏智能体……

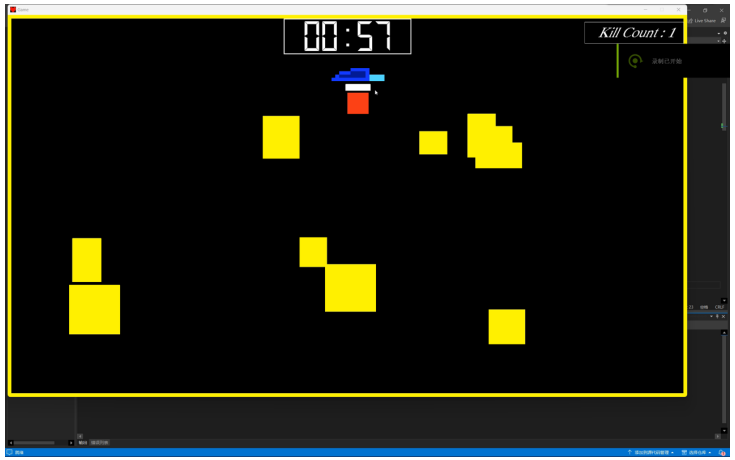
**需能够体现团队的工作和创新，可以参考改进但不允许直接抄袭开源项目**

# 参考案例




手语识别

斗地主AI




自制生存游戏AI


云顶之弈AI

异常姿态提醒桌宠

魔方还原

# 学术诚信

1. 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程

2. 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用**

3. 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**，因此，请主动防止自己的作业被他人抄袭

# 什么是人工智能？

# 先看看人工智能

# 可以做什么？

# 文本对话

**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CEOSamAltman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

# 视频生成

一个全副武装的宇航员
在沙漠中踩着滑板冲浪

Rockefeller center is overrun by golden retrievers! everywhere you look, there are golden retrievers

# 艺术创作


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

# 棋类博弈



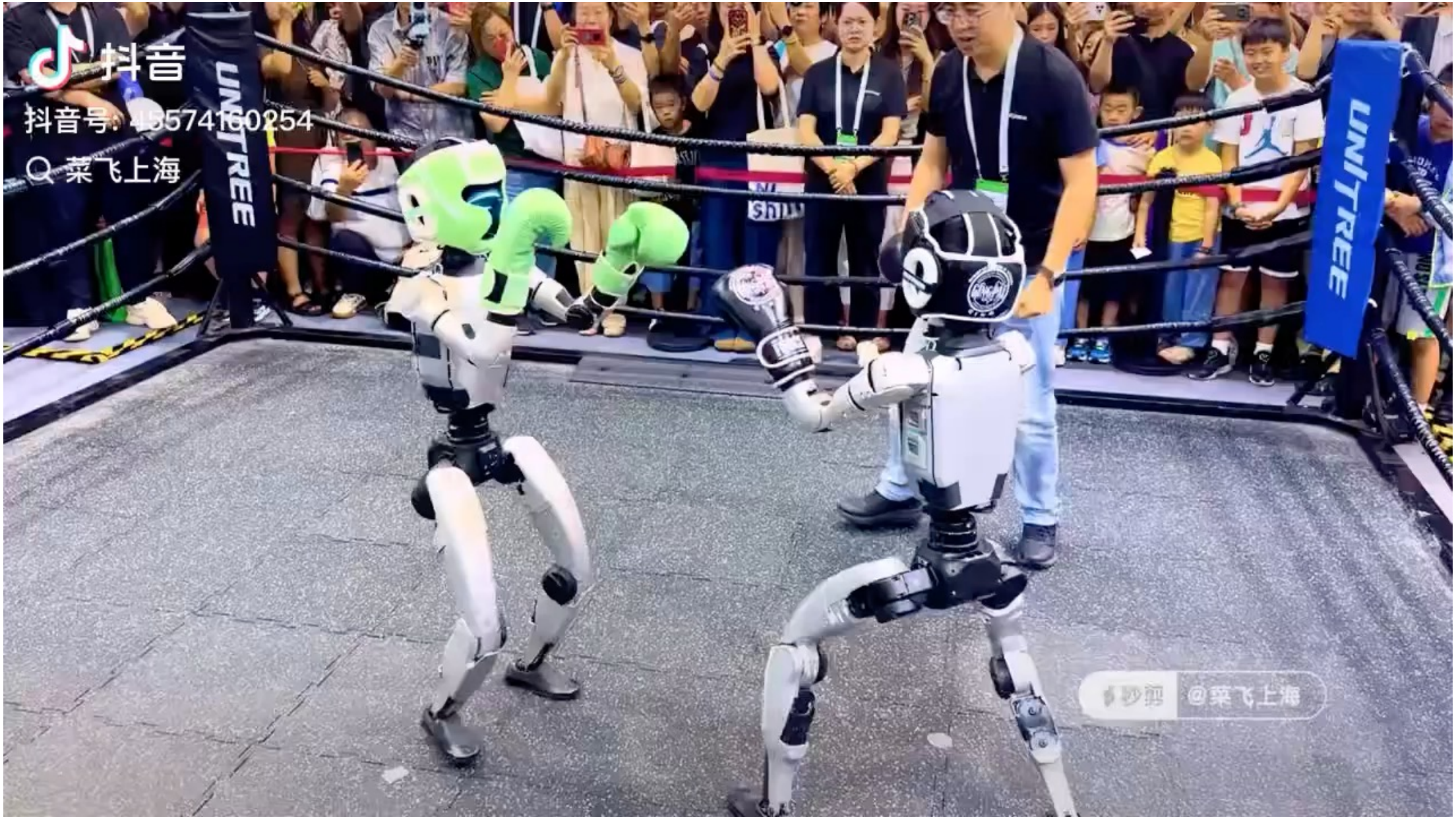
计算/预测出较高胜率的走法?

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



# 数学竞赛

RESEARCH

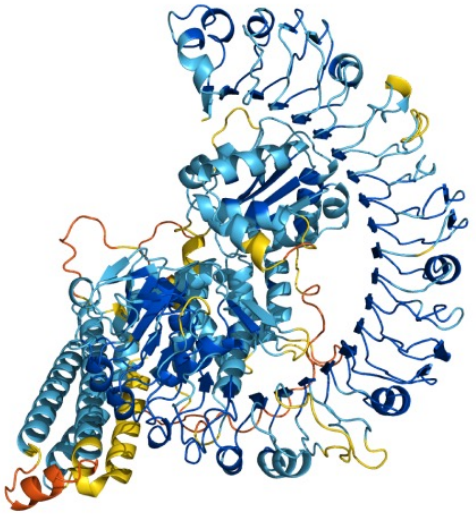
## Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

Thang Luong and Edward Lockhart

# 科学研究

- 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

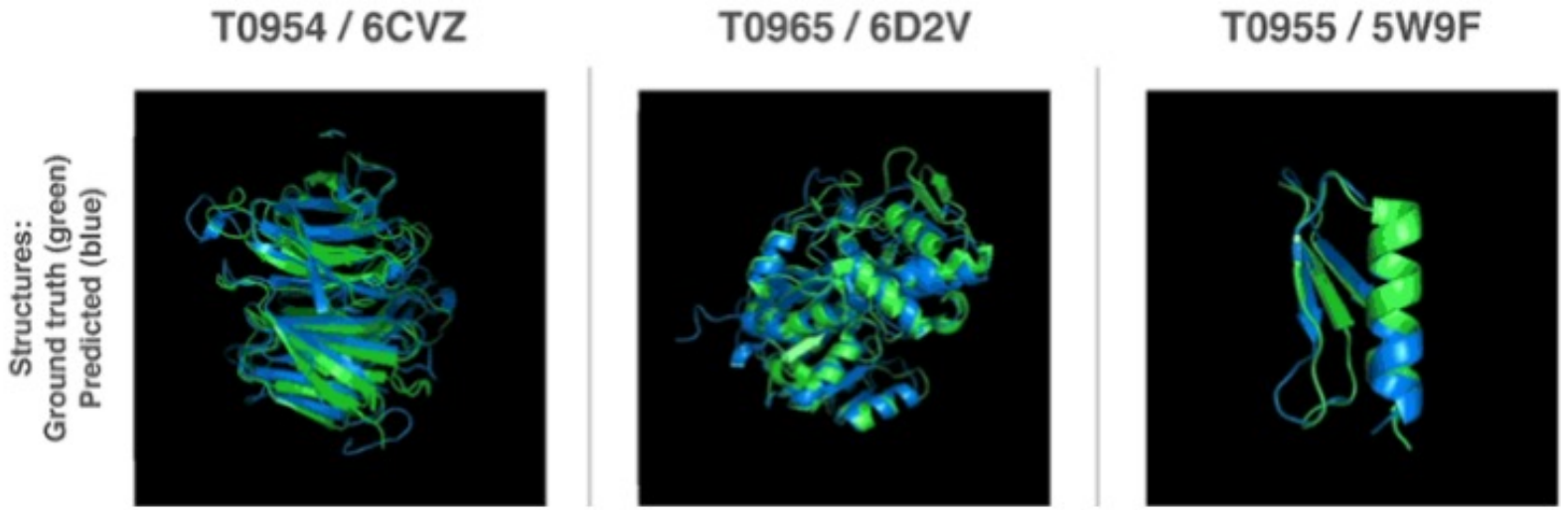


- 2021年8月，DeepMind宣布已将人类的98.5%的蛋白质预测了一遍；此前，已知氨基酸顺序的蛋白质分子的三维结构被看清的不到0.1%

- 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

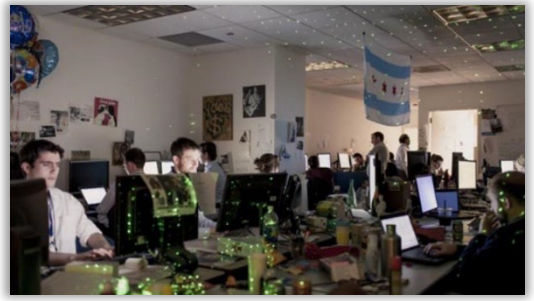
# 科学研究

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

# 美国总统大选


这个团队行动保密，定期向奥巴马报送结果；被奥巴马公开称为总统竞选的“核武器按钮”（“They are our nuclear codes”）

通过人工智能模型

◆ 个性化宣传

喜欢宠物？

奥巴马也有宠物！

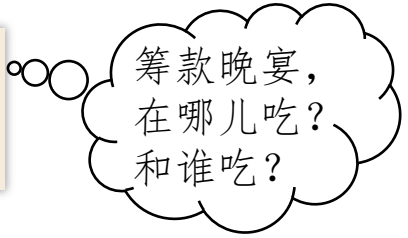

喜欢篮球？

奥巴马也是篮球迷！

◆ 广告购买

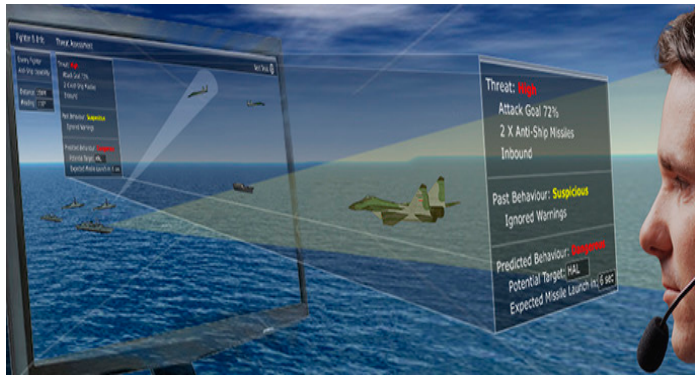
精准定位不同选民群体，建议购买冷门广告时段，广告资金效率比2008年提高14%

◆ 筹款




和乔治克鲁尼/奥巴马共进晚餐对于年龄在40-49岁的美西地区女性颇具吸引力…… 乔治克鲁尼为奥巴马举办的竞选筹资晚宴成功募集到1500万美元

# 战场战术（美）




眼镜蛇系统：

Coastal Battlefield Reconnaissance and Analysis (COBRA)

用于频海战斗舰，执行无人空中战术侦察。在两栖攻击之前，于海浪区和海滩区探测和定位雷区和障碍物

http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=1237&ct=2

http://www.navysbir.com/n15_1/N151-049.htm

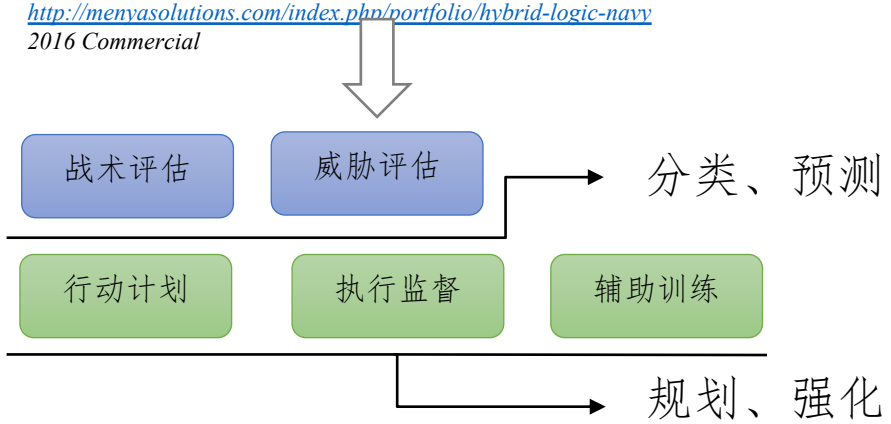
*2015 US Navy Official*

## HybridLogic Navy:

一套自动的基于机器学习的代理，帮助人类和无人机理解战术状况，及时做出最佳决策，以对付海军作战中的威胁

http://menyasolutions.com/index.php/portfolio/hybrid-logic-navy

*2016 Commercial*




United States Navy
Fact File

AN/DVS-1 COASTAL BATTLEFIELD RECONNAISSANCE AND ANALYSIS - (COBRA)

The mission of the AN/DVS-1 Coastal Battlefield Reconnaissance and Analysis (COBRA) system is to conduct unmanned aerial tactical reconnaissance in the littoral battlespace for detection and localization of minefields and obstacles in the surf zone and beach zone prior to an amphibious assault. The COBRA airborne payload will be carried on the MQ-8 Fire Scout unmanned air system. This allows operators and other personnel to remain at a safe distance from the mine and obstacle belts and enemy direct and indirect fire. COBRA will be embarked in the Littoral Combat Ship (LCS) as part of the Mine Countermeasures (MCM) Mission Package (MP).

DESCRIPTION: The Coastal Battlefield and Reconnaissance (COBRA) program (Ref 1) is interested in technologies that facilitate automated target recognition (ATR) capabilities in aerial multi-spectral images for previously unseen environments and target types. Targets of interest include minefields and obstacles in various land and marine environments. Typically, ATR algorithms are developed offline (post-mission) using previously acquired test data sets. These algorithms are based on supervised learning methods (Ref 2) that incorporate data from a limited set of test fields. When data is acquired in new environments, the algorithms often must be re-optimized to have good performance in that environment, as well as maintain performance in previously seen environments. The process for performing this offline re-optimization is often costly since it requires the efforts of expert analysts to assimilate data sets, determine target truth, analyze target features, train the ATR classifiers and evaluate performance.

There is a need for innovative methods that can 1) incorporate information from new data sets into the ATR system as they are acquired, and 2) re-optimize ATR algorithms quickly across all known environments, including those of newly acquired data. Online Machine Learning (OML) algorithms (Ref 3-5) can potentially be used to "learn" in the field based on operator-provided results without affecting prior performance. The information collected online can be used to refine the prediction hypothesis (classifier) used in the ATR algorithms. In addition, the information may provide input for automated methods of optimizing ATR performance across all known data sets.

The proposed effort will develop innovative OML algorithms for ATR that can incorporate human operator decisions to optimize probability of detection and probability of false alarm performance in new environments and for new target types. These algorithms will be integrated into mission and post-mission analysis systems in which operators review acquired images. The algorithms will be implemented as object-oriented C++ code for insertion into the operator systems. Development of the online learning algorithms must be combined with identification of how the operator will interact with them to provide updated decision information. Robust optimization of the ATR algorithms may be performed post-mission, which will require the development of separate software tools for processing historical data sets. The OML algorithms and optimization tools developed in this effort will reduce program costs by minimizing the time required for optimizing ATR algorithms to perform well in unseen operational environments.

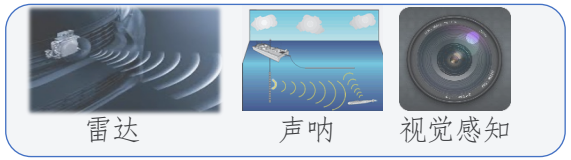
自动目标识别、监督学习以及在线学习技术被作为核心技术并多次提及

# 战场战术（英）

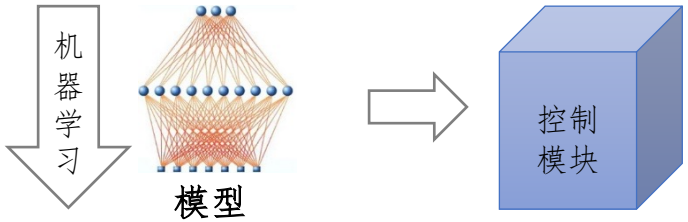
## 无人侦察快艇:

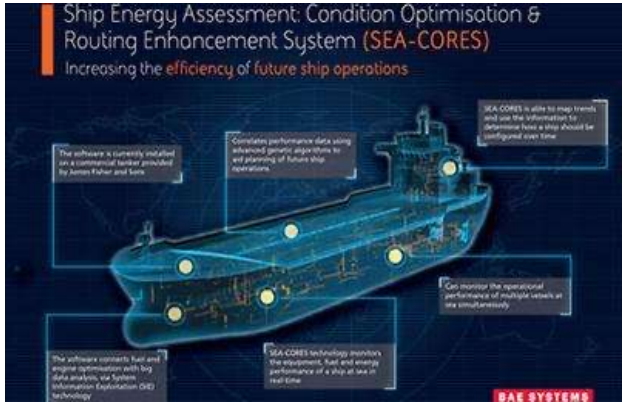
无人控制的情况下以50公里时速追踪快速目标并自动避障，进行跟踪、监视和间谍活动，或者用于海岸巡逻

[http://www.telegraph.co.uk/news/2016/09/05/navy-unveils-robot-spy-speedboat/](http://www.telegraph.co.uk/news/2016/09/05/navy-unveils-robot-spy-speedboat/) *2016 Royal Navy Official /Commercial*




多源感知




应对现代军舰日益复杂系统结构、针对其系统产生的海量数据而开发，能够有机组织军舰各个子系统，最终优化全舰效能

[https://phys.org/news/2016-09-software-ship-maintenance.html](https://phys.org/news/2016-09-software-ship-maintenance.html)

*2016 Royal Navy Official / Commercial*

## 船舶能源评估-条件优化和路由增强系统

**Software to transform ship maintenance**

September 21, 2016

SEA-CORES. Credit: University of Southampton

Researchers from the University of Southampton are to develop software that can monitor the equipment, fuel and energy performance of a ship at sea.

The University is part of the Ship Energy Assessment – Condition Optimisation & Routing Enhancement System (SEA-CORES) consortium, which provides a live model of ship performance on global operations. The development of the software is led by BAE Systems and is sponsored by Innovate UK.

SEA-CORES is able to correlate variables that could affect a ship's performance, such as energy consumption and different weather conditions. Using genetic algorithms to track and capture the live data, SEA-CORES provides those on board with a greater understanding of the vessel's capabilities across a wide range of operations.

Researchers from Electronics and Computer Science at the University of Southampton will work on monitoring loads on the ship and applying novel machine learning techniques to a domain that has largely been data poor.

Dr Sarvapali Ramchurn, who is leading the Southampton research group, said: "Unleashing such technologies on the marine sector is likely to have a huge impact. The work we are doing at Southampton in terms of autonomous systems and machine learning will help improve the efficiency of ships and detect potential issues before they cause major damage."

BAE Systems is developing and testing SEA-CORES on a commercial tanker provided by James Fisher Marine Services. The trial will analyse the vibration and trim performance of the vessel, its hull state and monitor the integrity of the ship's superstructure.

Chris Courtaux, Head of Engineering and Energy Services at BAE Systems, said: "SEA-CORES is able to consider all of the important components which affect the performance of a vessel during deployment.

"For instance, reducing speed may save fuel but increase the wear to the engine if below its optimum performance. This could in turn increase the maintenance requirements for these vessels and reduce their availability. It is crucial that we continue to analyse what more can be done to maintain these vessels in an efficient manner and increase the number of ships available for the Royal Navy fleet."

The software connects technologies in delivering fuel and engine optimisation through the use of the BAE Systems' Ship Energy Assessment System (SEAS), together with big data analysis by using System Information Exploitation (SIE) technology.

SEA-CORES has been developed in response to the increasing complexities of modern warships and the amount of data their systems produce. The technology could transform how the Royal Navy and BAE Systems maintain and support warships in the future by using the genetic algorithms to identify the relationships between a ship's systems, calculate their different permutations and ultimately recommend a strategy to optimise the vessel's performance.

遗传算法以及其他一些机器学习方法用于获取追踪数据和确定舰船子系统关联的任务中

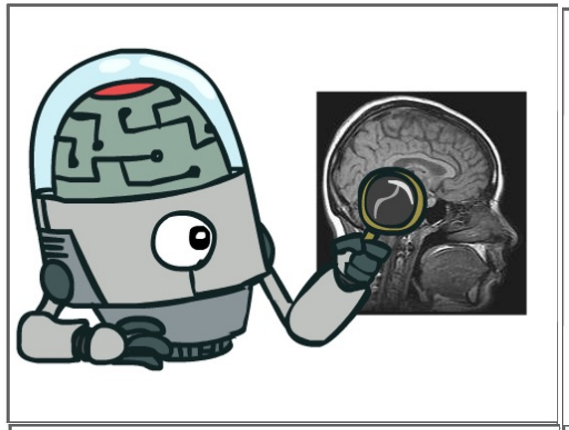
从日常对话、艺术创作，

到科学研究、军事政治

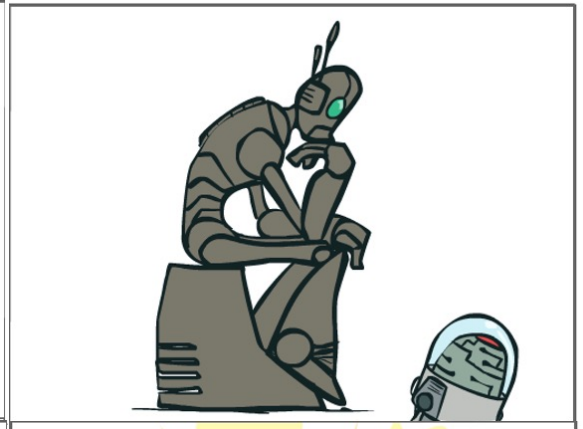
所以，到底什么是人工智能?

# 什么是人工智能

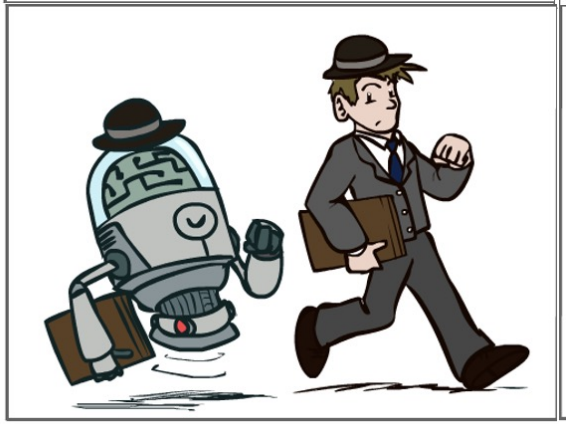
Think like people



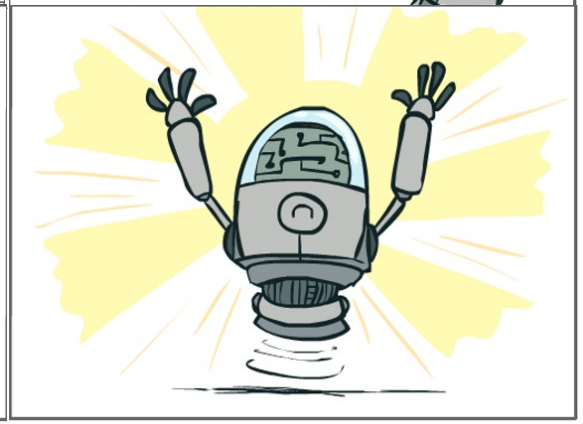
Think rationally



Act like people



Act rationally

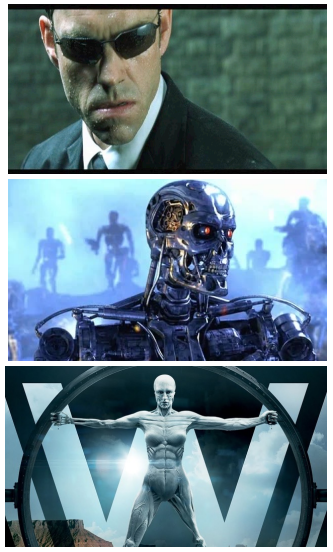


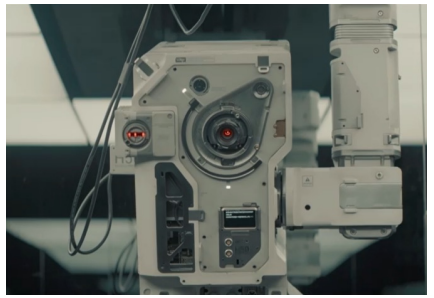
人工智能专注于研究和构建做正确的事情的智能体

# 两种不同的人工智能

- 强人工智能（“科幻人工智能”）

研制出和人一样聪明，甚至比人更聪明的机器





重要特征：

- 具有自主意识
- 全面达到，甚至超过人类智能水平
- …

# 两种不同的人工智能

- 弱人工智能（“科学人工智能”）

  让机器做事时聪明一点

  “人工智能就是让机器来完成那些如果由人来做则需要智能的事情的科学”

Marvin Minsky
(1927-2016)
人工智能奠基者之一
1969年图灵奖

解读：

- 如果某件事情需要智能，通过机器来做，就是人工智能
- 不要求“全面”达到人类智能水平
- “做事”就行，不必具备“自主意识” “情感” ……
- ……

# 一个类比

人的智能行为

人工智能

人工智能重要，是因为能造出“智能工具”（类比：飞机）

- 造飞机的人不会关心飞机有没有“意识”、会不会“疼”
- 更不会关心飞机是否“全面达到”鸟的能力（例如：下蛋）

# 注意

“强人工智能”与“弱人工智能”

区别:

并非在于“能力有多强”，

而在于“是否拥有自主意识”

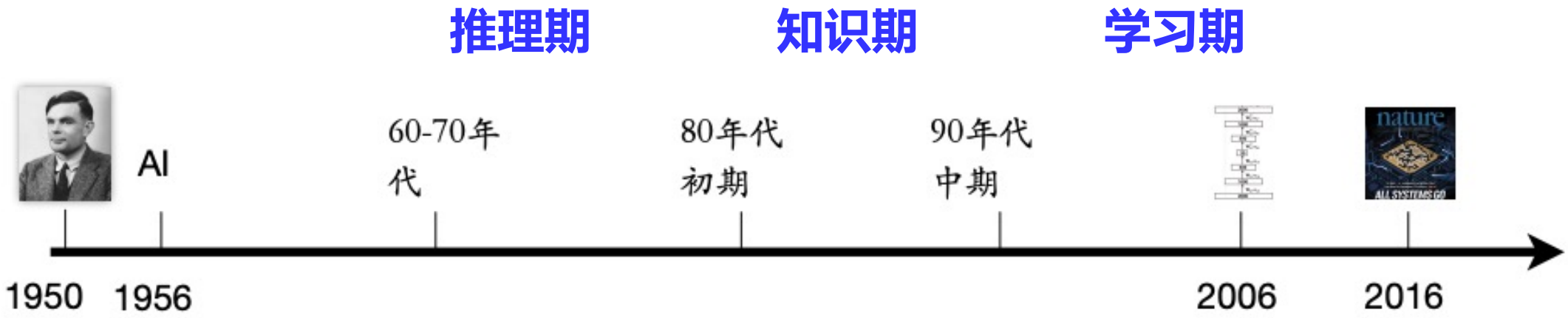
*简单地说:*

*“强人工智能”目的是“造类人”*

*“弱人工智能”目的是“造工具”*

# 人工智能是如何发展到现在的？

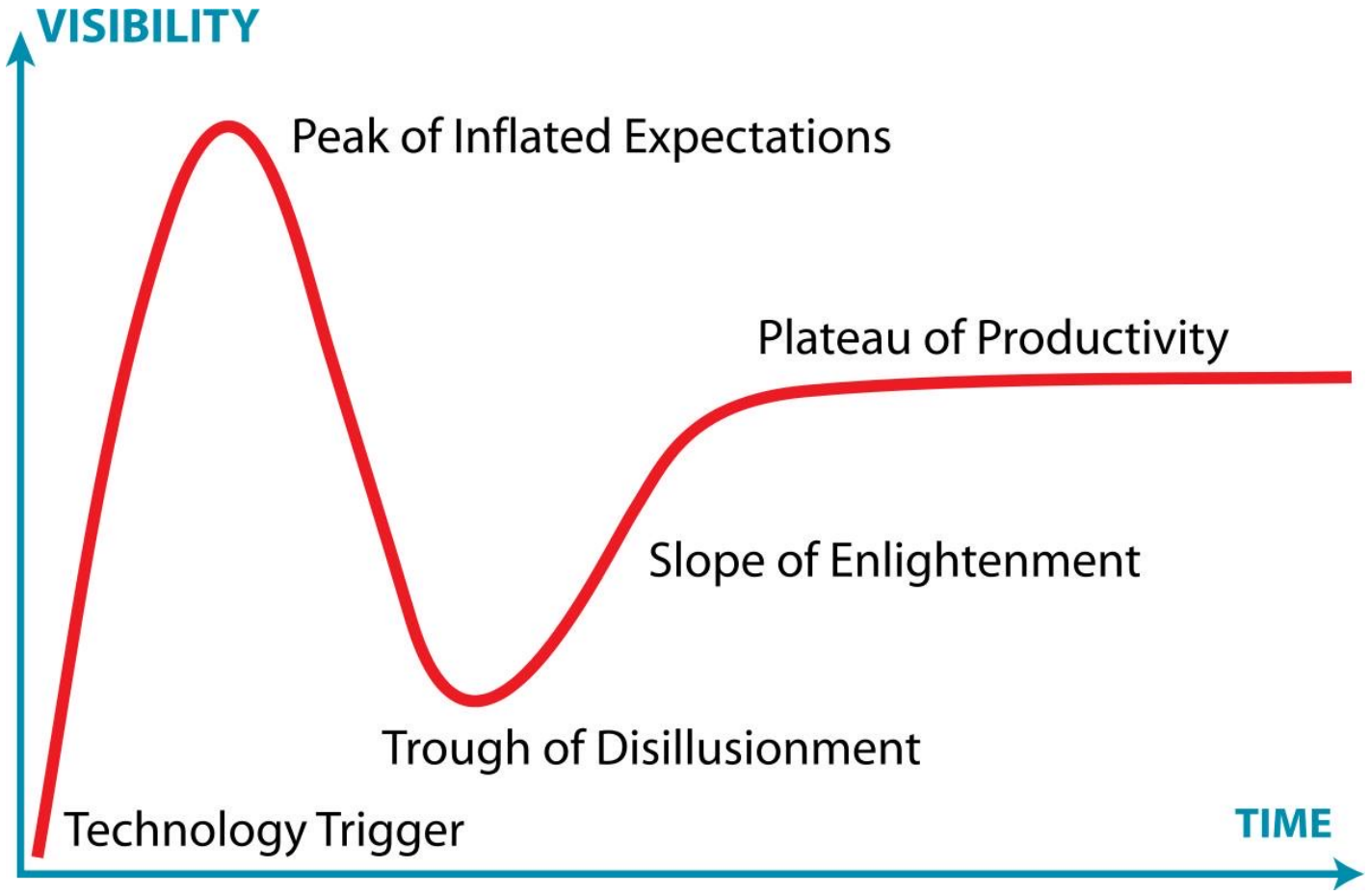
# 人工智能简史


推理期
知识期
学习期
AI
60-70年代
80年代初期
90年代中期
nature
1950
1956
2006
2016

# AI Summer & Winter

- 社会对人工智能领域的热情、投入和研究方法呈周期性起伏

- 每一个AI的夏天，都孕育了有深远影响的方法论



# Overview

- 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**

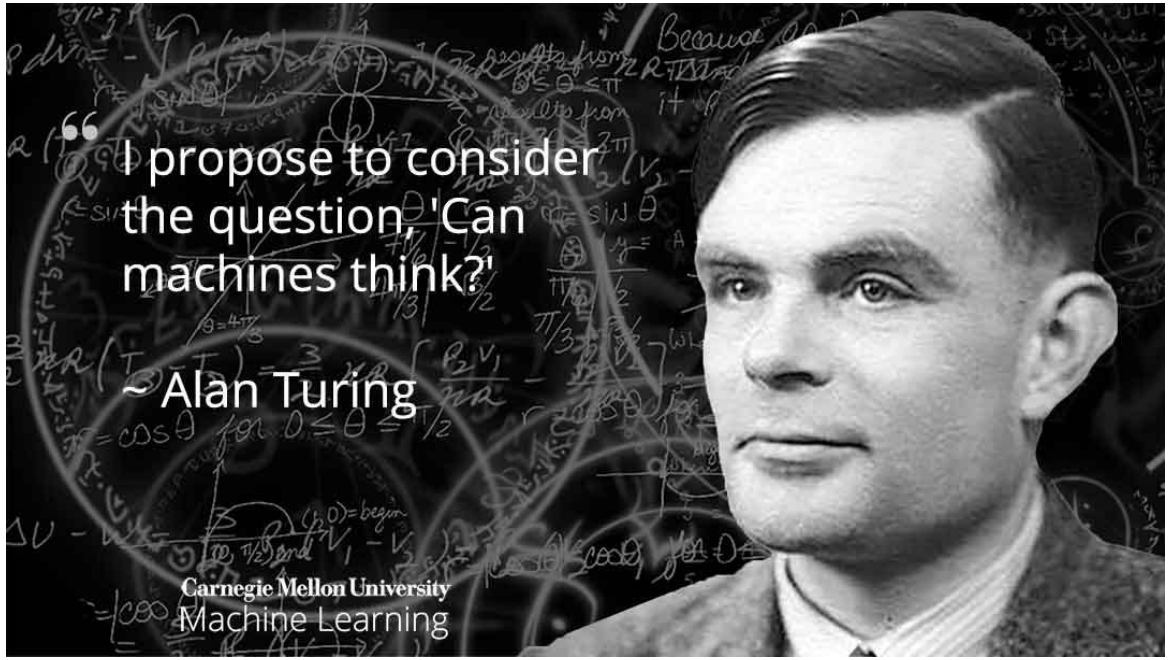
**计算机器与智能**

**1950年**

**艾伦·图灵**

**“Can machine think?”**

**机器能思考吗?**

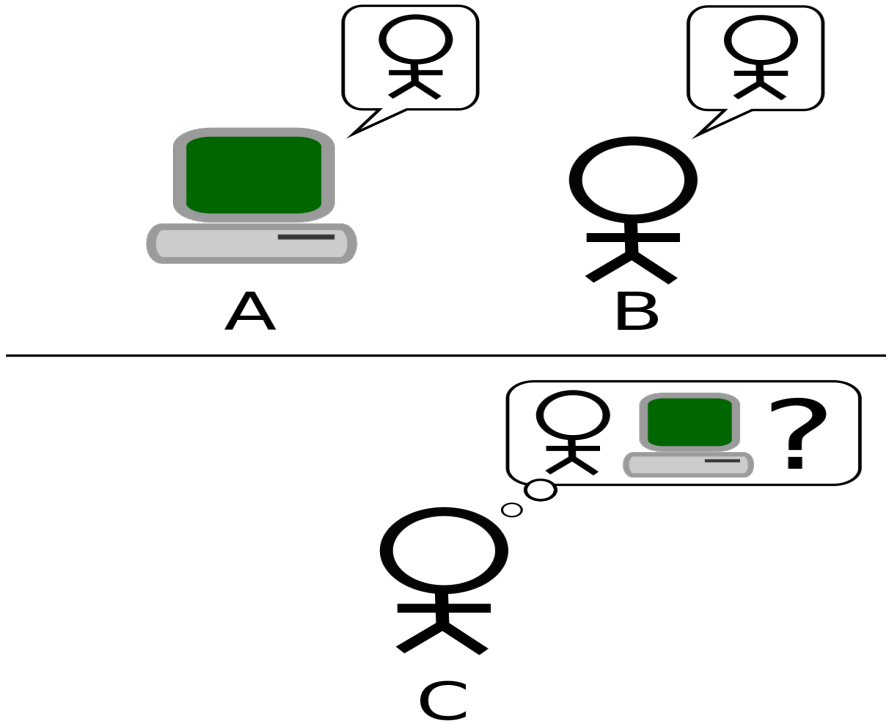



**艾伦·图灵**

**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被测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

# 人工智能的起源

- “We may hope that machines **will eventually compete with men in all purely intellectual fields**. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the **playing of chess**, would be best.”

- “I believe that in **about fifty years** time it will be possible, to program computers, with a **storage capacity of about 10^9 [one gigabyte]**, to make them **play the imitation game** so well that an average **interrogator will not have more than 70% chance of making the right identification** after five minutes of questioning.”

# 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

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

*John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon*

The 1956 Dartmouth summer research project on artificial intelligence was initiated by this August 31, 1955 proposal, authored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The original typescript consisted of 17 pages plus a title page. Copies of the typescript are housed in the archives at Dartmouth College and Stanford University. The first 5 papers state the proposal, and the remaining pages give qualifications and interests of the four who proposed the study. In the interest of brevity, this article reproduces only the proposal itself, along with the short autobiographical statements of the proposers.

guage, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1. Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack

《人工智能达特茅斯夏季研究项目提案》

# 1956人工智能元年

*50年后*

# 第一次浪潮：1956年到20世纪70年代初

## 1956-1970s：推理期

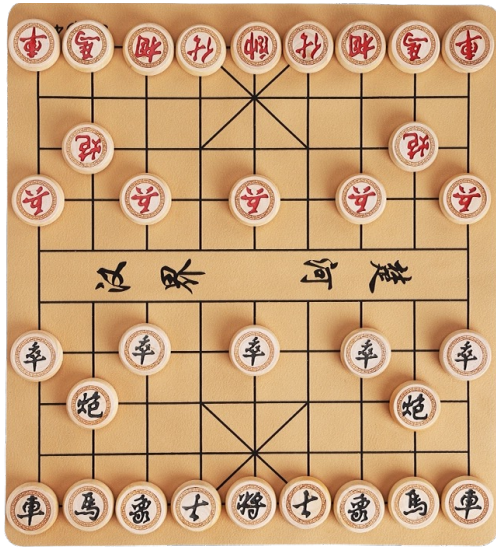
◆出发点：“数学家真聪明！”

◆**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



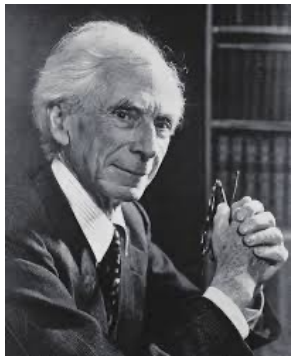
- 棋盘（符号：棋盘格）
- 棋子（符号：车、炮、兵…）
- 每个棋子怎么走（规则：马走日，象走田…）
- 基本策略（启发式规则：开局先出炮…）

**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

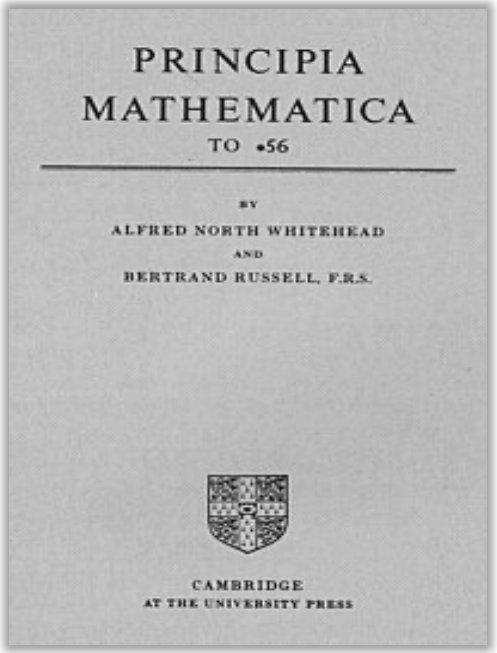
# 第一次浪潮：1956年到20世纪70年代初

- **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）**

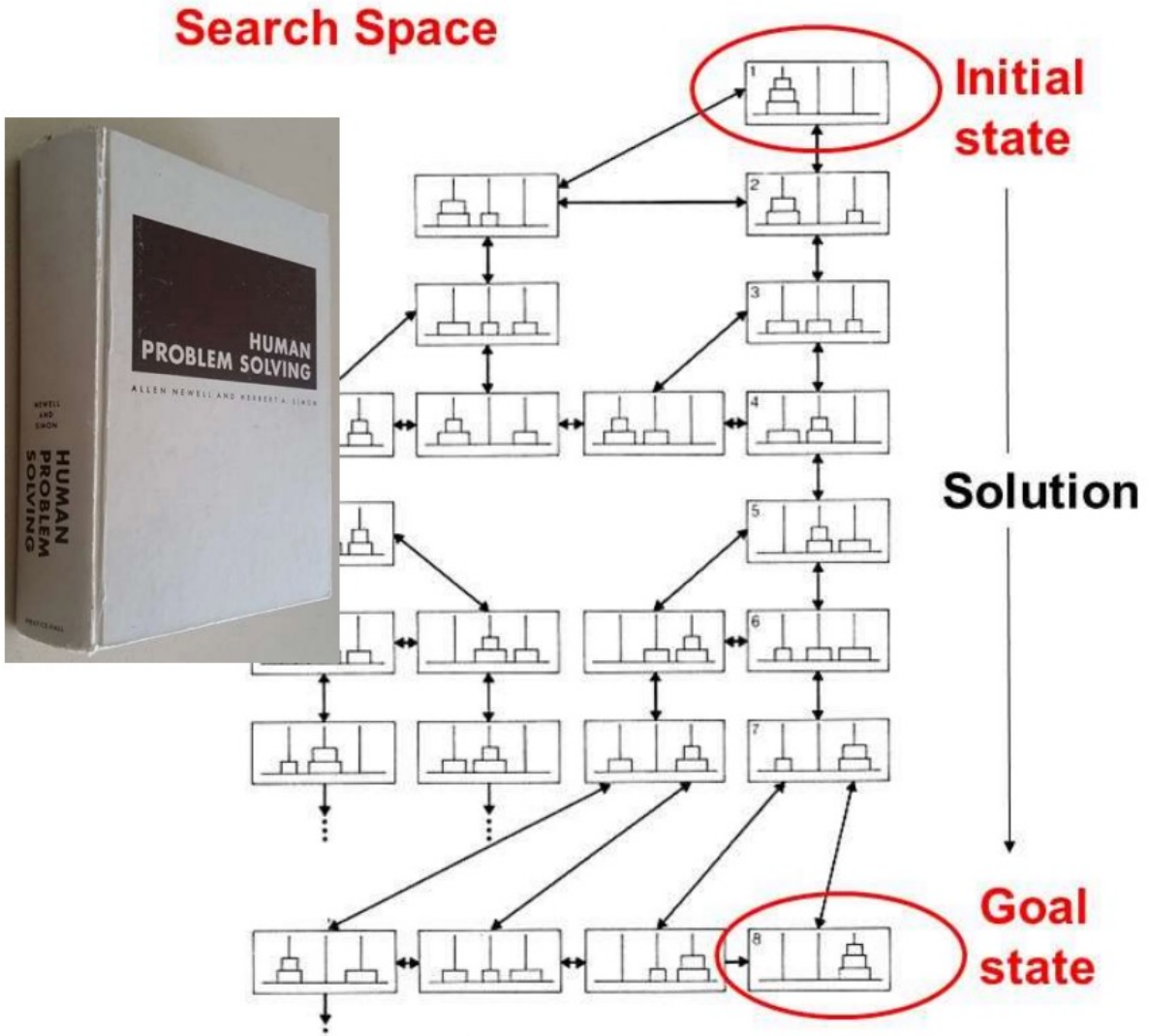
**Problem Solving is a search problem**

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**

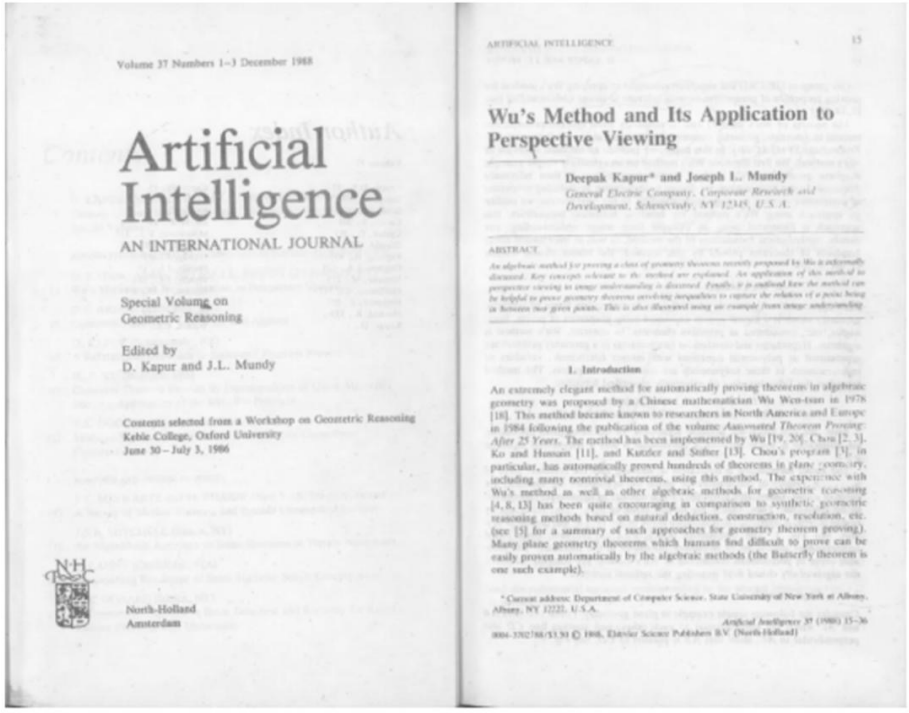
**王浩自动定理证明里程碑大奖**

# 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”



Volume 37 Numbers 1–3 December 1988

Artificial Intelligence

AN INTERNATIONAL JOURNAL

Special Volume on Geometric Reasoning

Edited by D. Kapur and J.L. Mundy

Contents selected from a Workshop on Geometric Reasoning Kebie College, Oxford University June 30 – July 3, 1986

North-Holland Amsterdam

ARTIFICIAL INTELLIGENCE 15

Wu's Method and Its Application to Perspective Viewing

Deepak Kapur* and Joseph L. Mundy

General Electric Company, Corporate Research and Development, Schenectady, NY 12345, U.S.A.

ABSTRACT

1. Introduction

An extremely elegant method for automatically proving theorems in algebraic geometry was proposed by a Chinese mathematician Wu Wen-tsun in 1978 [18]. This method became known to researchers in North America and Europe in 1984 following the publication of the volume *Automated Theorem Proving: After 25 Years*. The method has been implemented by Wu [19, 20], Chou [2, 3], Ko and Hussain [11], and Kutzler and Stifter [13]. Chou's program [3], in particular, has automatically proved hundreds of theorems in plane geometry, including many nontrivial theorems, using this method. The experience with Wu's method as well as other algebraic methods for geometric reasoning [4, 8, 13] has been quite encouraging in comparison to synthetic geometric reasoning methods based on natural deduction, construction, resolution, etc. (see [5] for a summary of such approaches for geometry theorem proving). Many plane geometry theorems which humans find difficult to prove can be easily proven automatically by the algebraic methods (the Butterfly theorem is one such example).

*Current address: Department of Computer Science, State University of New York at Albany, Albany, NY 12222, U.S.A.

Artificial Intelligence 37 (1988) 15–36

0004-3702/88/$3.50 © 1988, Elsevier Science Publishers B.V. (North-Holland)

**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

## 第 2 章　自动定理证明兴衰纪

As a material machine economises the exertion of force, so a symbolic calculus economises the exertion of intelligence ... the more perfect the calculus, the smaller the intelligence compared to the results.

就像机器能省体力一样，符号演算能省脑力。

演算越完美，付出的脑力就越少。

——W. E. Johnson（**约翰逊**）

Proof is cultivated reasoning.

证明就是讲究的推理。

——Bruno Buchberger（**布赫贝格尔**）

https://www.ituring.com.cn/book/tupubarticle/19224

# 第一次浪潮：1956年到20世纪70年代初

https://arxiv.org/pdf/2404.09939

https://machine-learning-for-theorem-proving.github.io/

Published as a conference paper at COLM 2024

## A Survey on Deep Learning for Theorem Proving

**Zhaoyu Li**[1], **Jialiang Sun**[1], **Logan Murphy**[1], **Qidong Su**[1], **Zenan Li**[2], **Xian Zhang**[3]
**Kaiyu Yang**[4]*, **Xujie Si**[1,5]
[1]University of Toronto, [2]Nanjing University, [3]Microsoft Research Asia, [4]Meta FAIR,
[5]CIFAR AI Chair
{zhaoyu, six}@cs.toronto.edu

### Abstract

Theorem proving is a fundamental aspect of mathematics, spanning from informal reasoning in natural language to rigorous derivations in formal systems. In recent years, the advancement of deep learning, especially the emergence of large language models, has sparked a notable surge of research exploring these techniques to enhance the process of theorem proving. This paper presents a comprehensive survey of deep learning for theorem proving by offering (i) a thorough review of existing approaches across various tasks such as autoformalization, premise selection, proofstep generation, and proof search; (ii) an extensive summary of curated datasets and strategies for synthetic data generation; (iii) a detailed analysis of evaluation metrics and the performance of state-of-the-art methods; and (iv) a critical discussion on the persistent challenges and the promising avenues for future exploration. Our survey aims to serve as a foundational reference for deep learning approaches in theorem proving, inspiring and catalyzing further research endeavors in this rapidly growing field. A curated list of papers is available at https://github.com/zhaoyu-li/DL4TP.

## NeurIPS Tutorial on Machine Learning for Theorem Proving

**Video Recording**

### Overview

Machine learning, especially large language models (LLMs), has shown promise in proving formal theorems using proof assistants such as Coq, Isabelle, and Lean. Theorem proving is an important challenge for machine learning: Formal proofs are computer programs whose correctness can be verified. Therefore, theorem proving is a form of code generation with rigorous evaluation and no room for the model to hallucinate, opening up a new avenue for addressing LLMs' flaws in factuality.

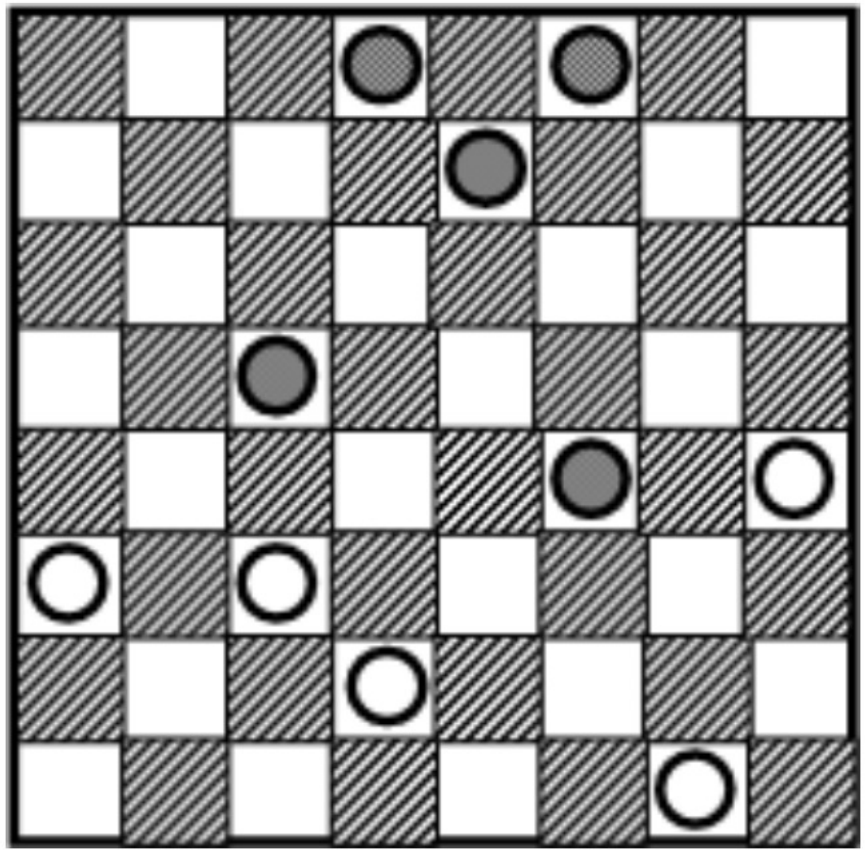
Despite its potential, learning-based theorem proving has significant entry barriers, primarily due to the steep learning curve for proof assistants. This tutorial aims to bridge this gap and make theorem proving accessible to researchers with a general machine learning background. To that end, our presentation will contextualize theorem proving from a machine learning perspective and demonstrate how to develop LLMs for theorem proving, using newly available open-source tools that provides interfaces to proof assistants without requiring in-depth knowledge of their internals. Furthermore, we will cover advanced topics and open problems in learning-based theorem proving, including its synergies with natural language processing and software verification.

Throughout the presentation, we will highlight several conceptual themes recurring in theorem proving that are also critical for machine learning, such as mathematical reasoning, code generation, and hallucination prevention. The panel will complement the presentation through a broader discussion of related topics such as trustworthy machine learning, LLMs for code, reasoning, and program synthesis.

# 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：**

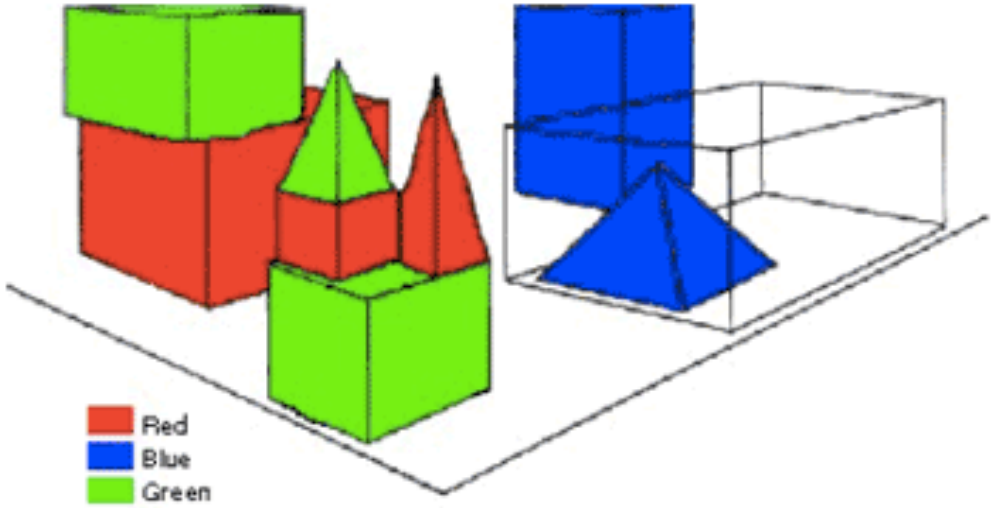
**1.死记硬背：** 把遇到过的局面及其评估分数存起来，下次遇到直接用

**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：

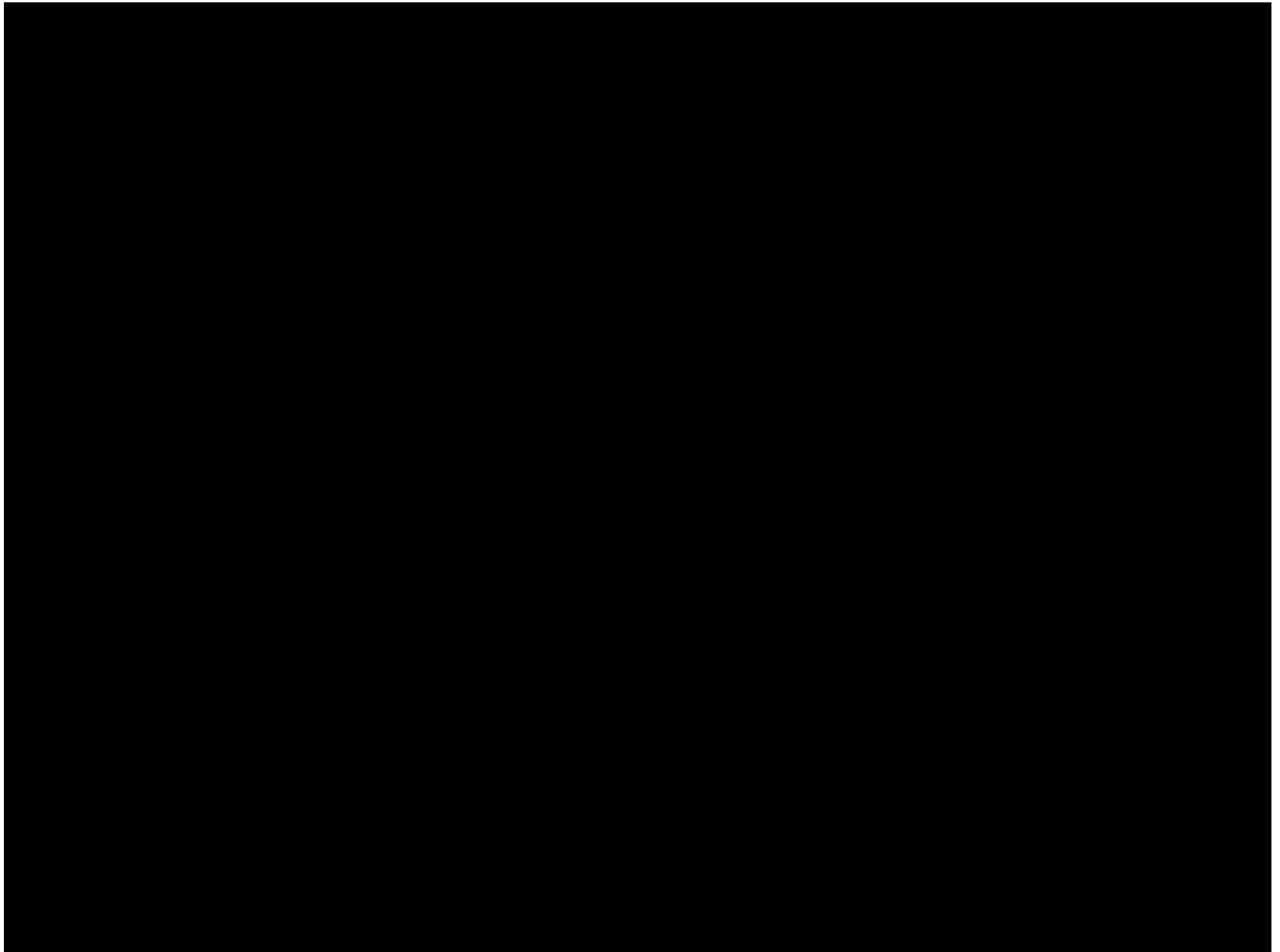
“为什么你感到沮丧？”

# 第一次浪潮：1956年到20世纪70年代初

## 那些年我们吹过的牛

- 赫伯特西蒙在1957年预测：“**不出十年，数字计算机将成为世界象棋冠军**”
- “**不出十年，数字计算机将能发现并证明一个重要的全新数学定理**”
- 马文明斯基在1970年预测：“**在三到八年的时间里，我们将拥有一台具有普通人一般智能的机器**”

# 第一次浪潮：1956年到20世纪70年代初



# 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

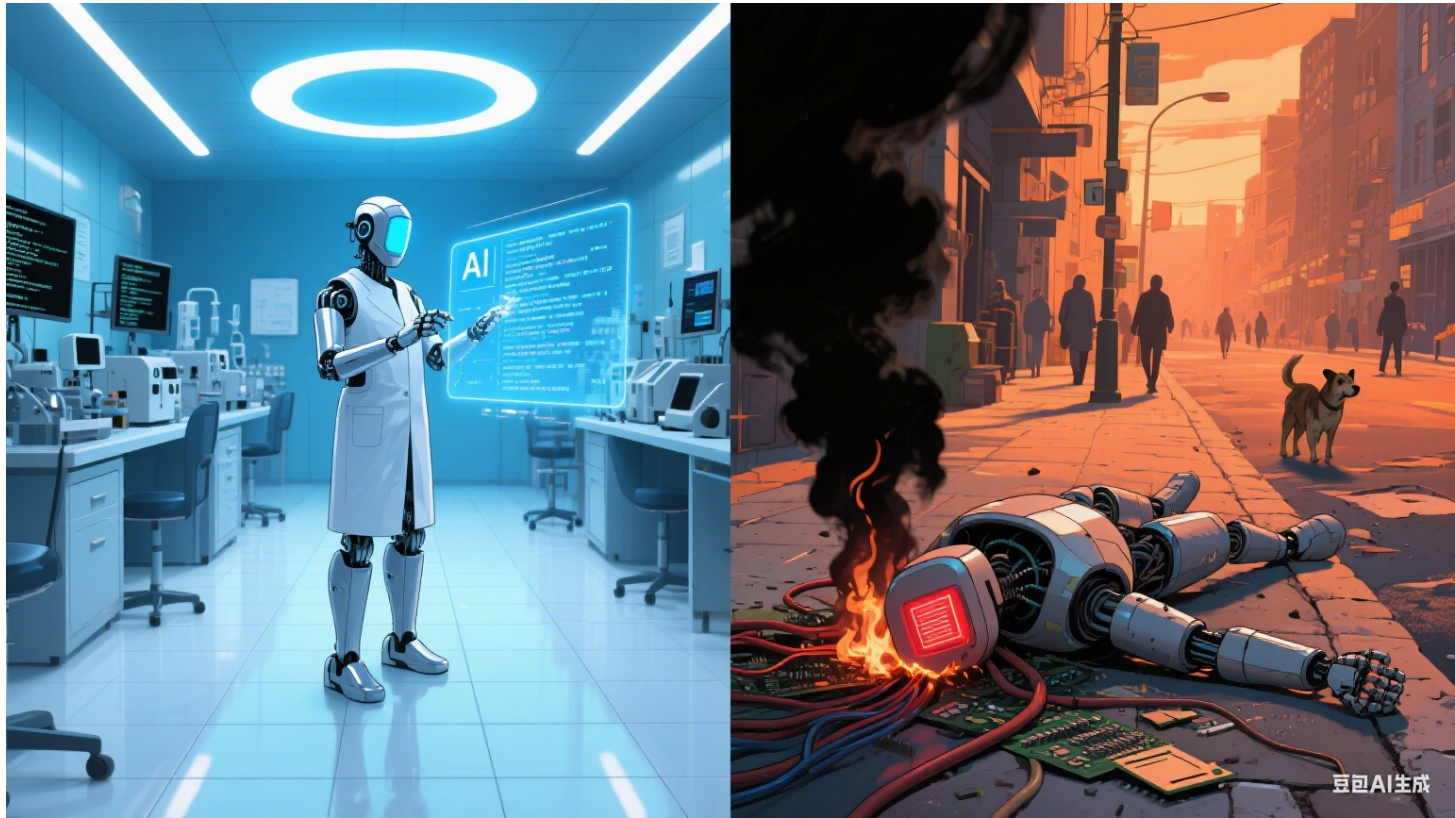
**重要挑战：**

- **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

# 第一次人工智能寒冬

# 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

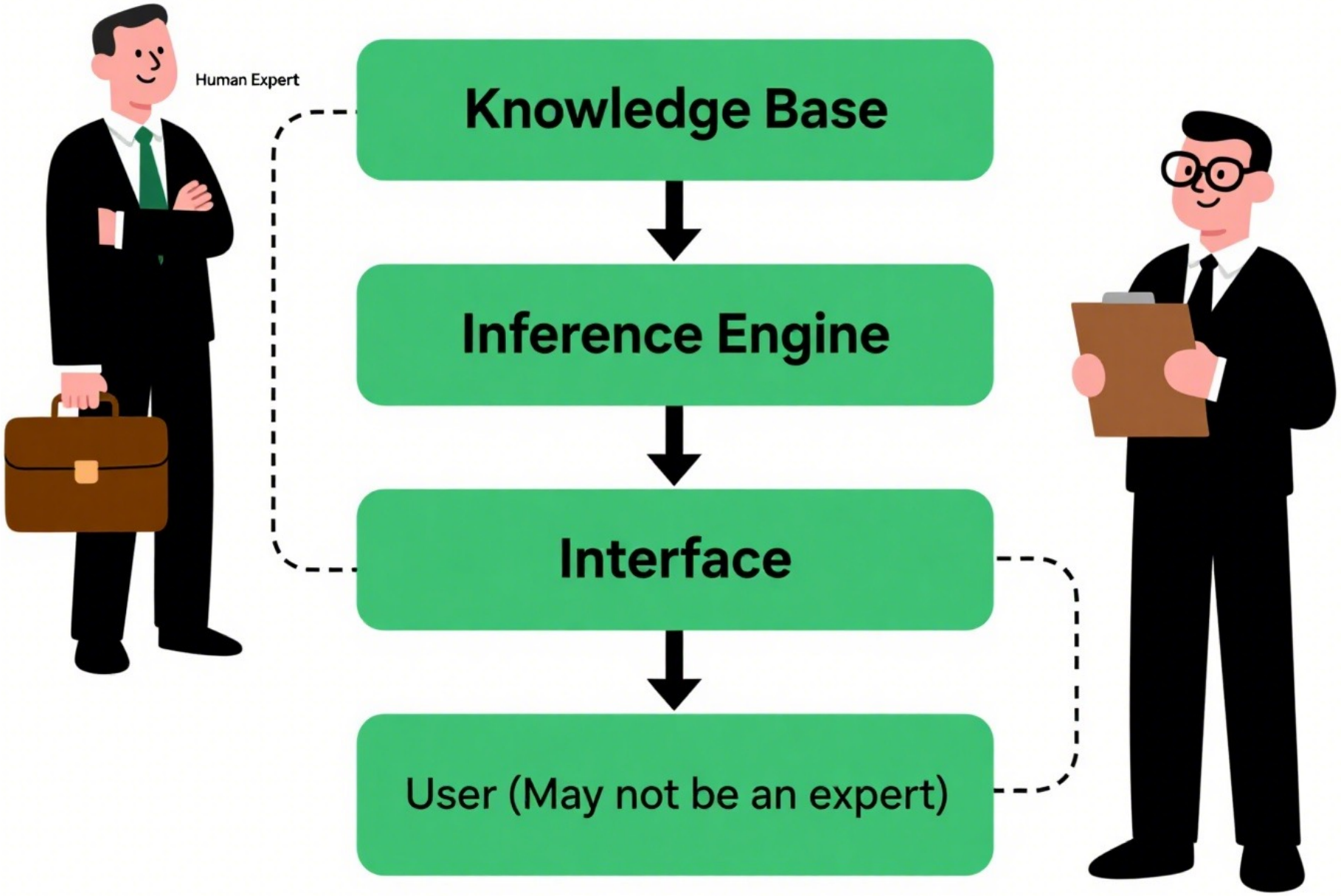
**1970s-1980s: 知识期**

**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

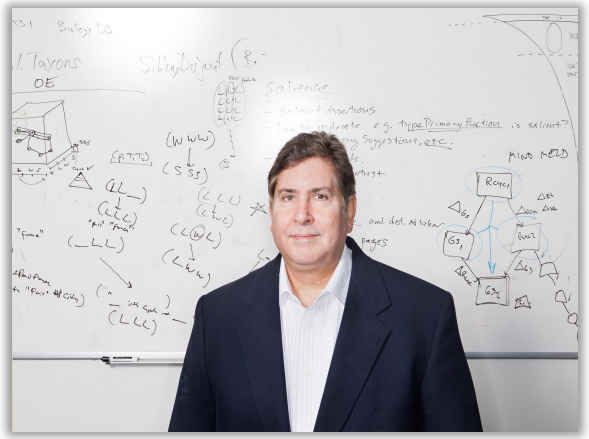
# 第二阶段：20世纪70年代末到80年代末

- **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)**



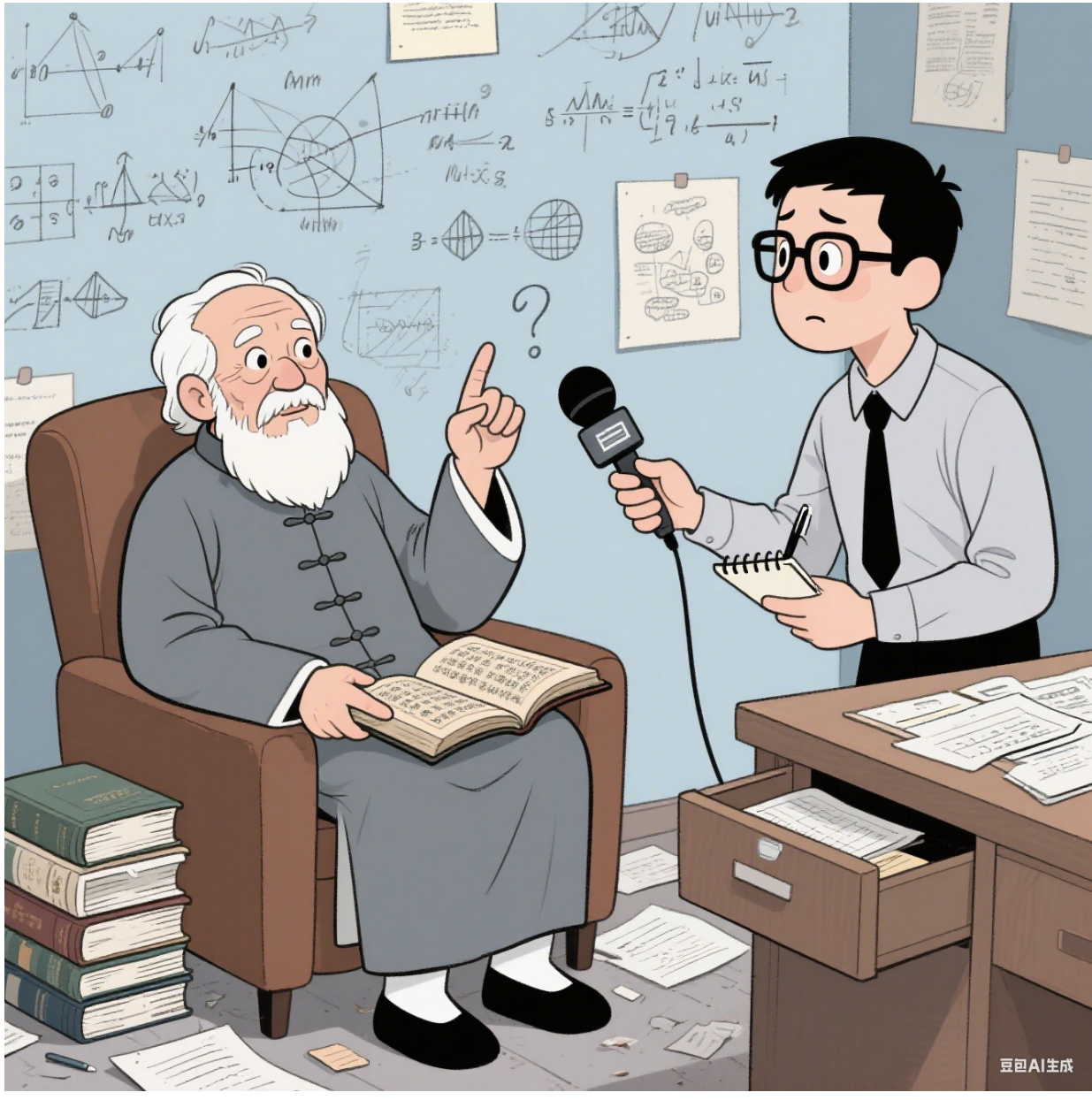
# 第二阶段：20世纪70年代末到80年代末

- **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

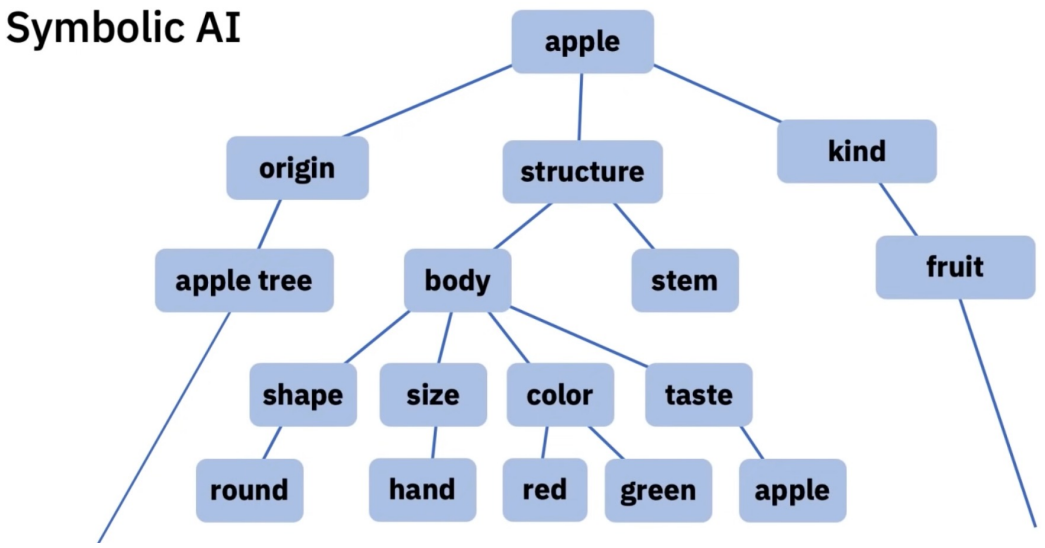


# 第一代人工智能：符号主义人工智能 (Symbolic AI)

## 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**

**如何识别苹果?**




## 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

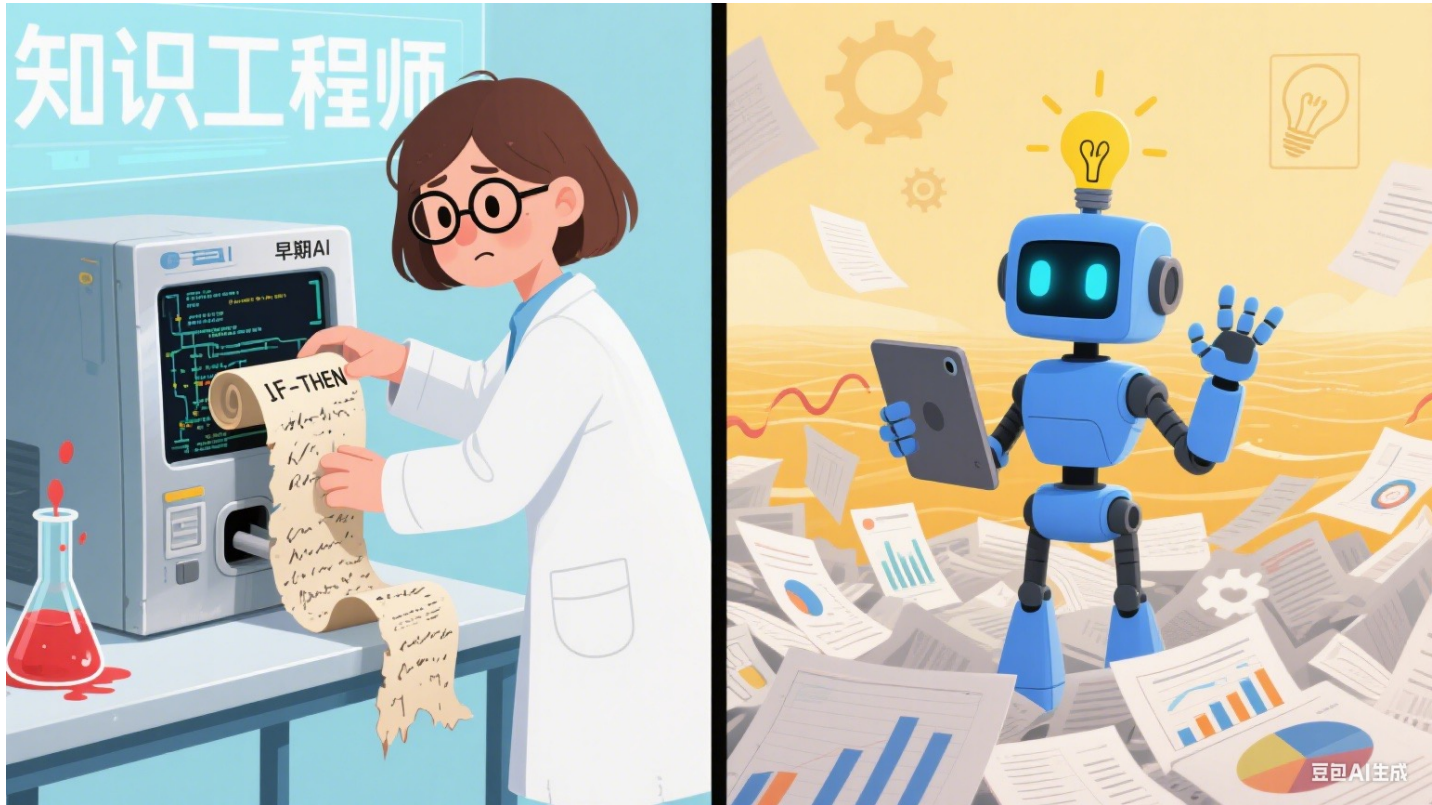
# 第三阶段-20世纪90年代至今

## 1990s-Now: 学习期

**从给我规则到给我数据**

- 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



# 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**

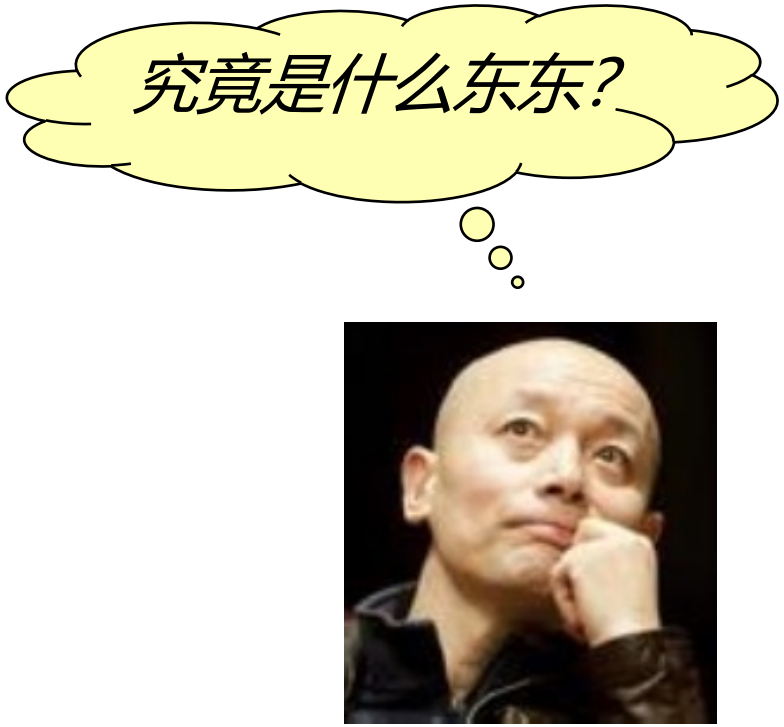
**大数据 ≠ 大价值**

**智能数据分析**

**机器学习**

# 机器学习（Machine Learning)

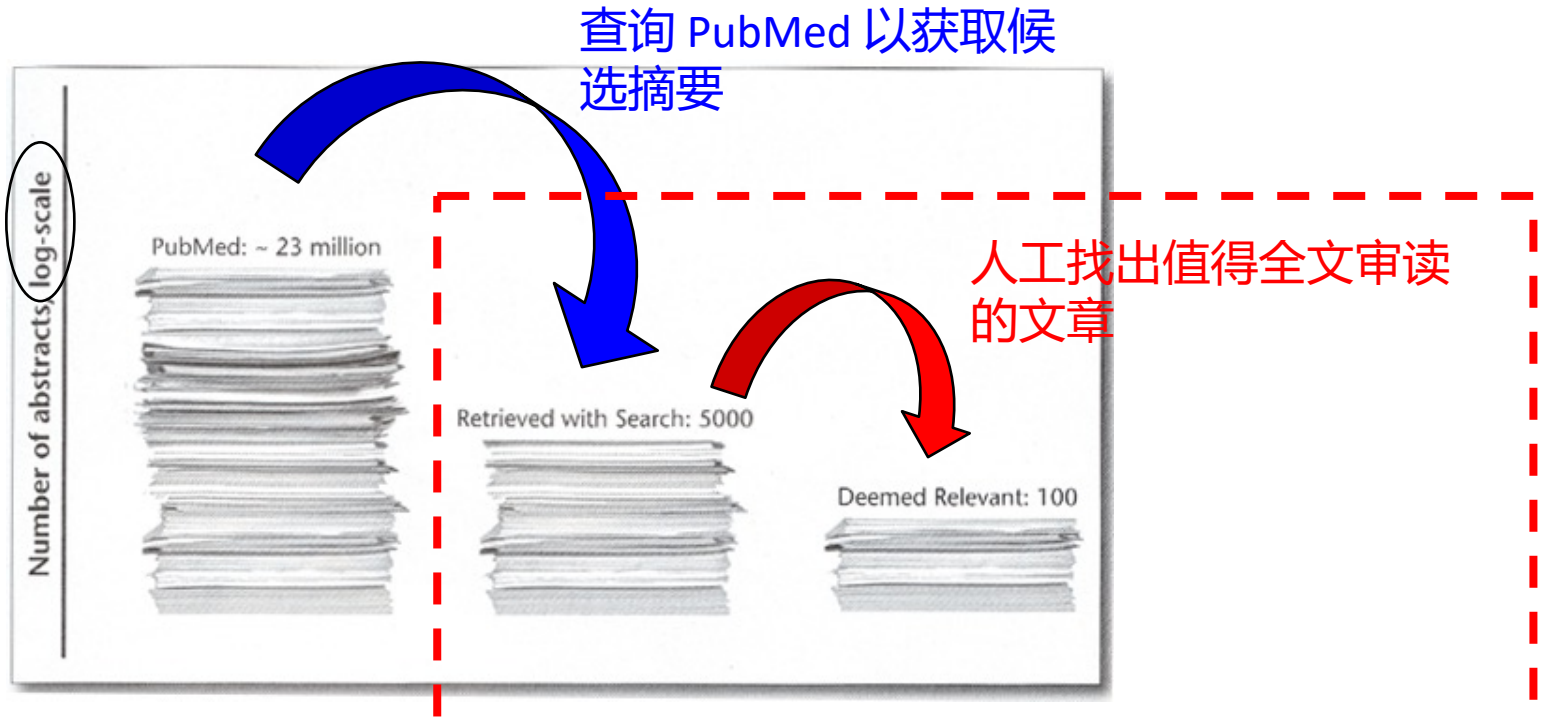
## 机器学习
## （Machine Learning）




看几个例子 ⇨

# 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

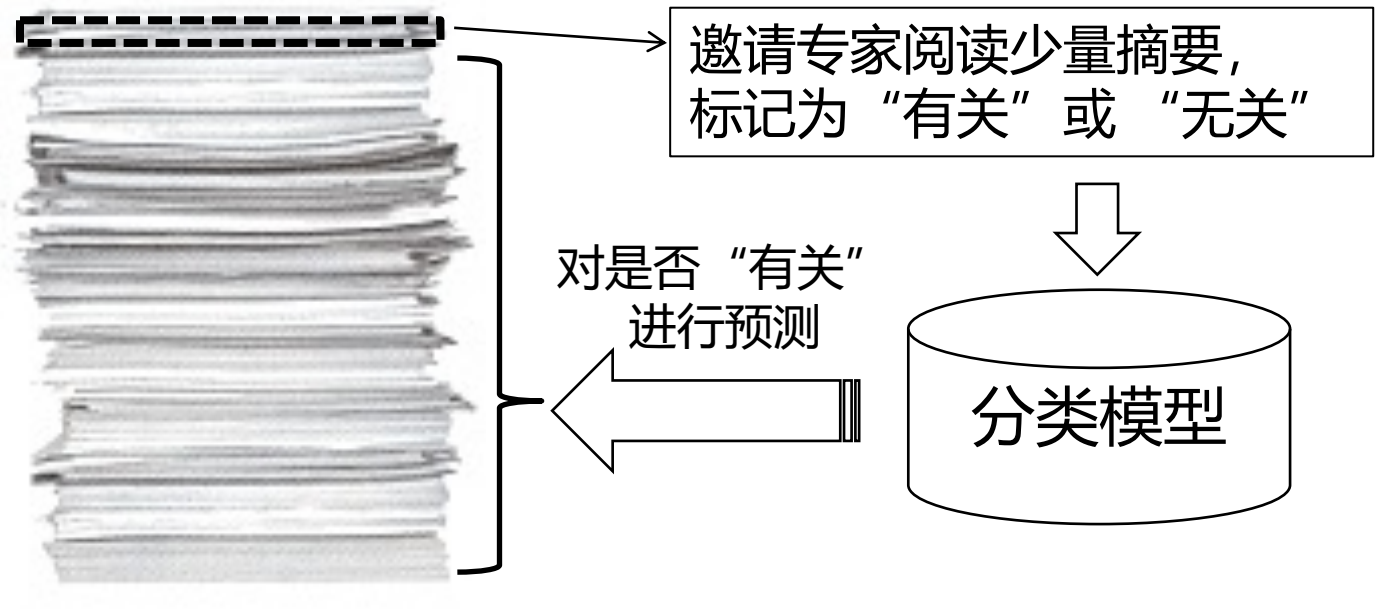
a portion of the 33,000 abstracts

**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)

# 画作鉴别

画作鉴别(painting authentication)：确定作品的真伪

勃鲁盖尔（1525-1569）的作品?

梵高（1853-1890）的作品?

该工作对专业知识要求极高

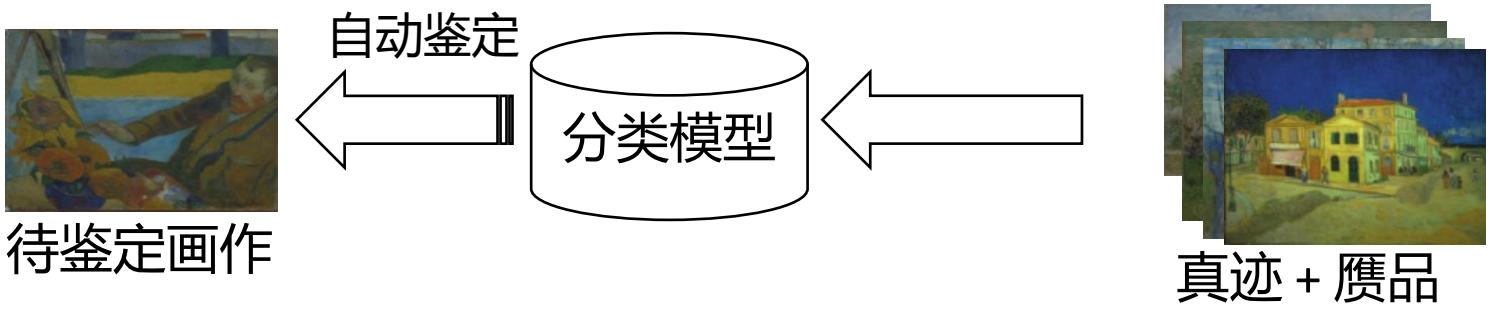
- 具有较高的绘画艺术修养
- 掌握画家的特定绘画习惯

**只有少数专家花费很大精力才能完成分析工作!**

**很难同时掌握不同时期、不同流派多位画家的绘画风格!**

# 画作鉴别

为了降低分析成本，**机器学习**技术被引入




Kröller Müller美术馆与Cornell等大学的学者对82幅梵高真迹和6幅赝品进行分析，自动鉴别精度达 **95%** [C. Johnson et al., 2008]

Dartmouth学院、巴黎高师的学者对8幅勃鲁盖尔真迹和5幅赝品进行分析，自动鉴别精度达 **100%** [J. Hughes et al., 2009][J. Mairal et al., 2012]

(对用户要求低、准确高效、适用范围广)

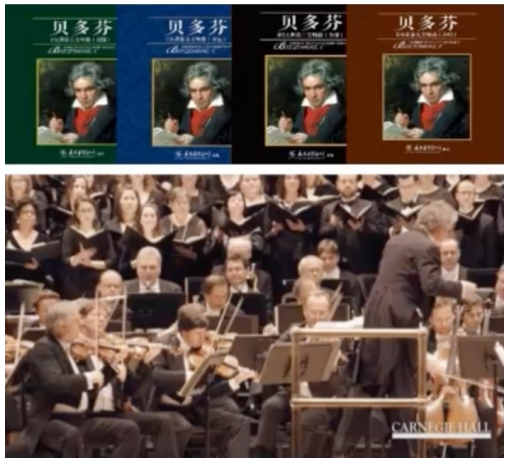
# 艺术创作

贝多芬（1770-1827）

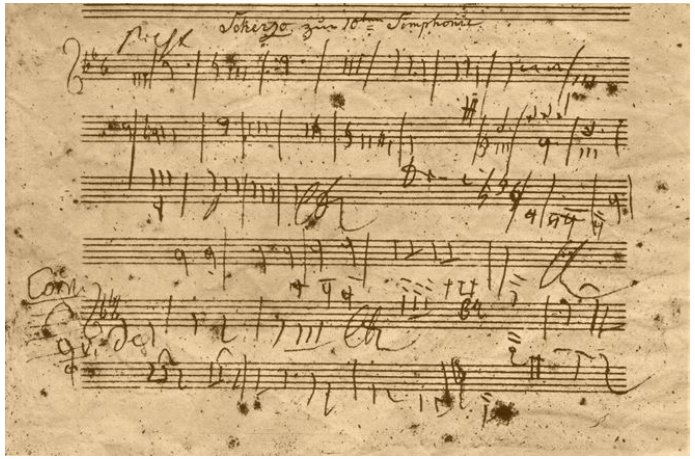
一共创作了9部交响曲：
《英雄》《田园》《命运》…

第十交响曲由于耳疾未能完成，
仅留下少量乐谱存稿

2019年组建团队
AI科学家+音乐家+历史学家




第一至九交响曲



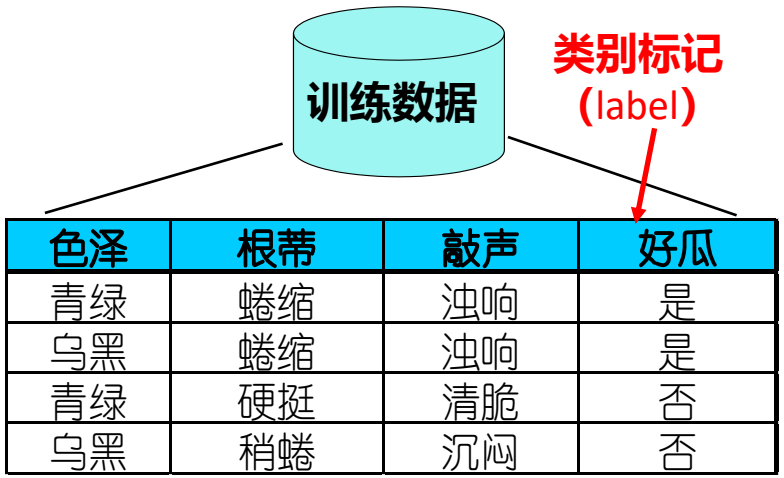
第十交响曲手稿片段

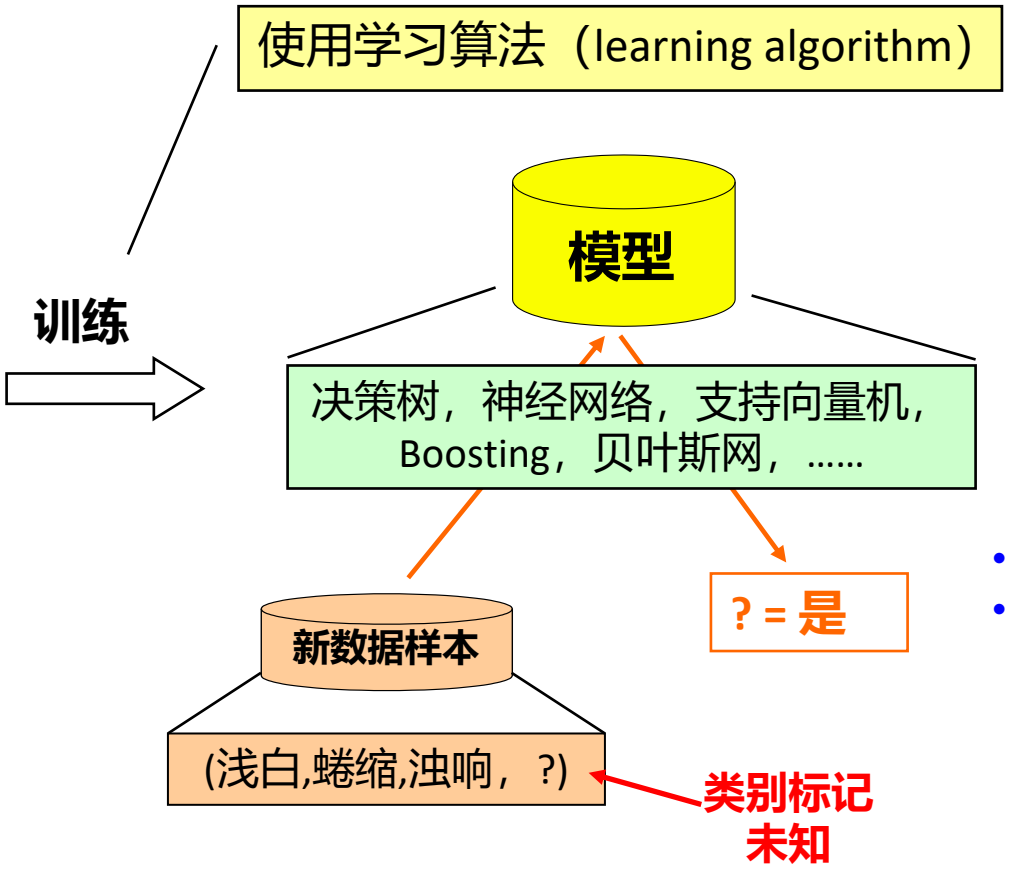
# 大量的机器学习算法

- 线性回归(Linear Regression)
- 决策树(Decision Tree)
- 支持向量机(Support Vector Machine)
- Adaboost
- 神经网络(Neural Network)
- ……

# 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(reinfocement learning)




- 分类, 回归
- 二分类, 多分类

- 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

# 潜在意义

横：输入（数据） -> 输出（标记）

纵：
历史（数据）
->
未来（数据）

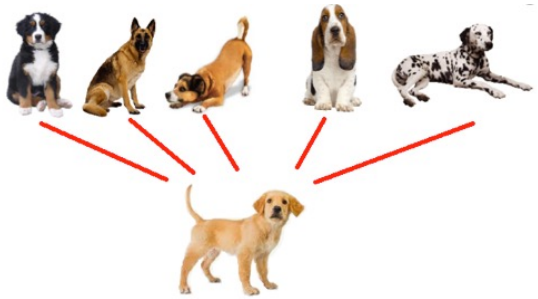
特征　　标记

| | 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 | ? |

机器学习：面向未来的技术

# 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)

# 例：房价预测

任务：训练机器学习模型，能够根据房子的面积预测出房价

训练数据

| 面积 | 房价 |
| --- | --- |
| 100 | 300 |
| 110 | 330 |
| 180 | 540 |

测试数据

| 面积 | 房价 |
| --- | --- |
| 140 | ? |

- 假设$f$是线性函数

$$f = w * x$$

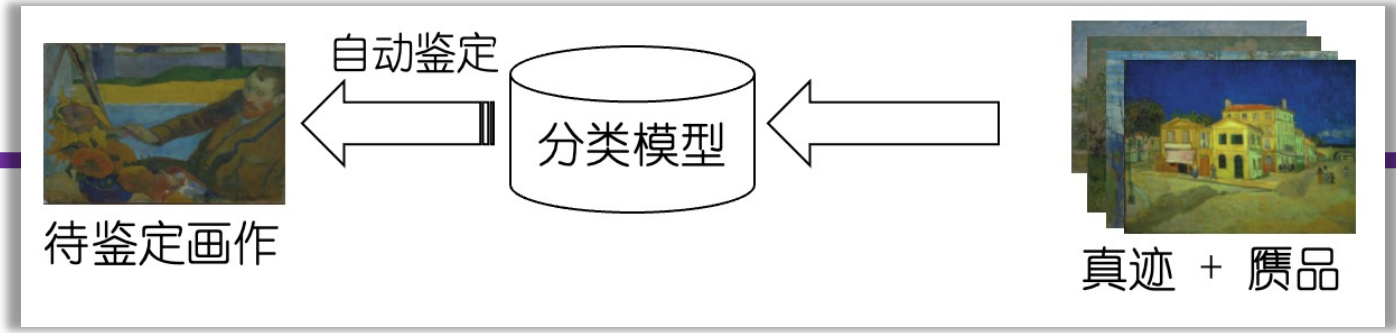
- Objective: 损失函数为均方误差

$$loss(f(x), y) = (y - f(x))^2$$

$$\min_w (300 - w * 100)^2 + (330 - w * 110)^2 + (540 - w * 180)^2$$

求得：$w = 3$

# 例：画作鉴别




假设$f$是线性函数：$f = w^{\top}X + b$

将像素转换为向量

| | | | |
|---|---|---|---|
| 0.2 | -0.5 | 0.1 | 2.0 |
| 1.5 | 1.3 | 2.1 | 0.0 |
| 0 | 0.25 | 0.2 | -0.3 |

$w$

| |
|---|
| 56 |
| 231 |
| 24 |
| 2 |

$x$

| |
|---|
| 1.1 |
| 3.2 |
| -1.2 |

b

| | |
|---|---|
| -96.8 | 梵高 |
| 437.9 | 勃鲁盖尔 |
| 61.95 | 其他人 |

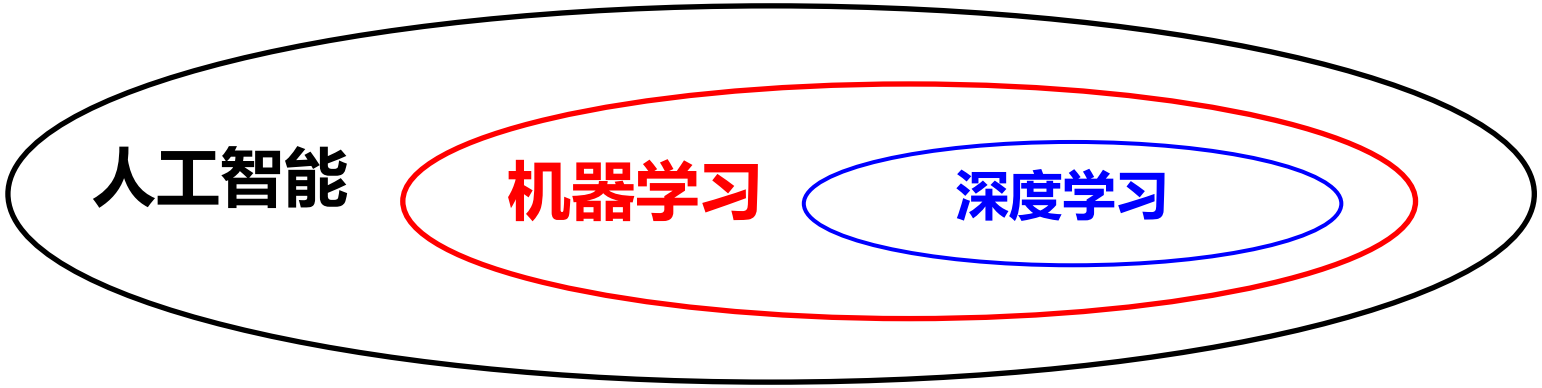
# 2010年以后，深度学习时代

**机器学习是人工智能的核心研究领域（之一）**

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



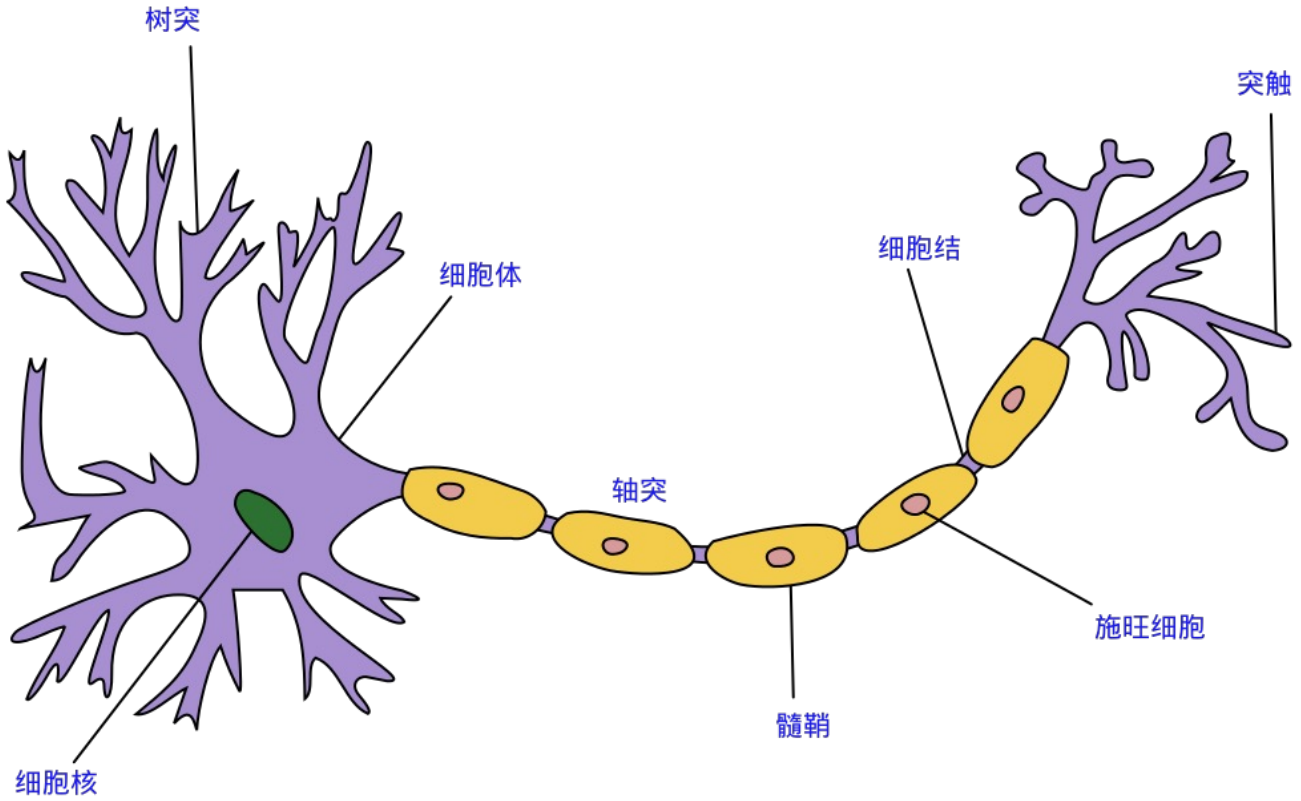
# 联结主义/连接主义

**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

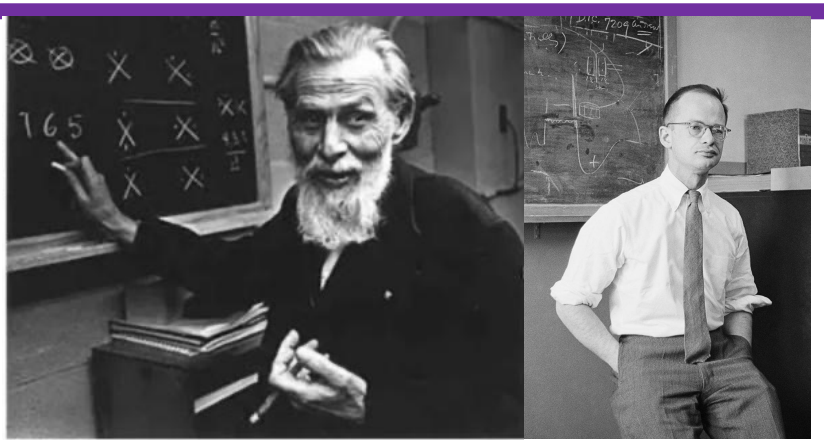
- 每个神经元与其他神经元相连, 当它“兴奋”时, 就会向相连的神经元发送化学物质, 从而改变这些神经元内的电位
- 如果某神经元的电位超过一个“阈值”, 那么它就会被激活, 向其它神经元发送化学物质



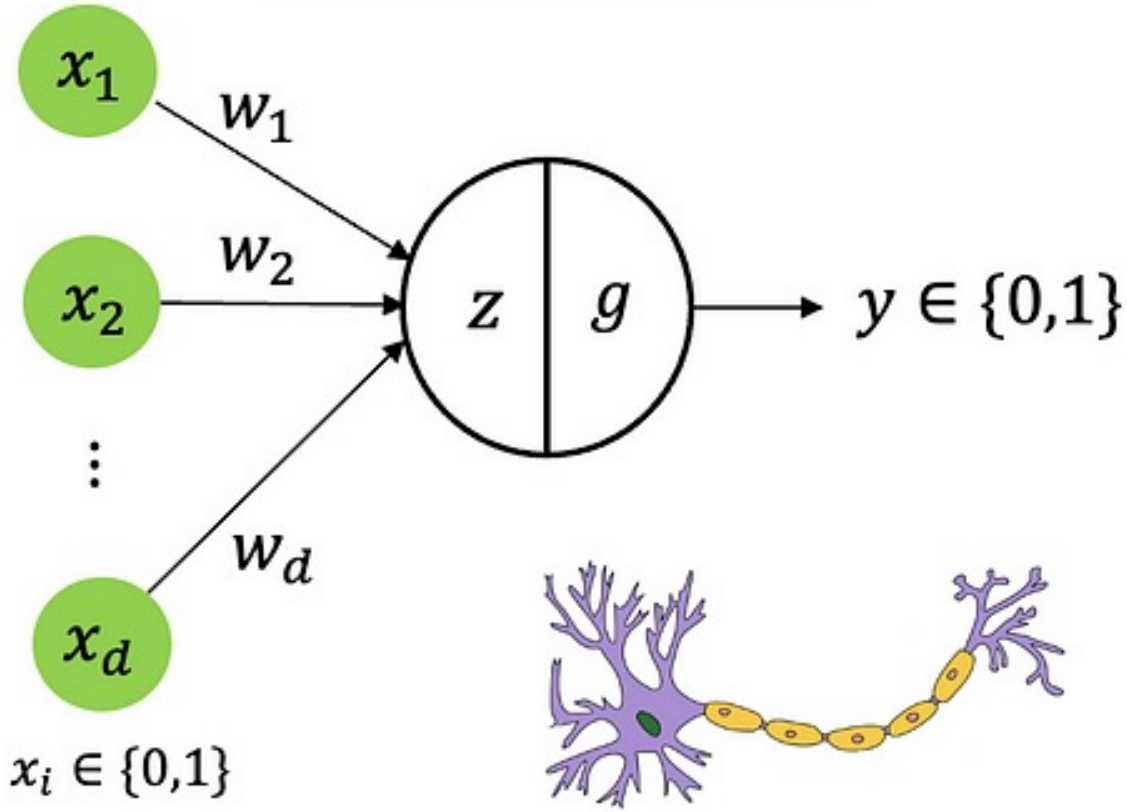
# M-P神经元

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





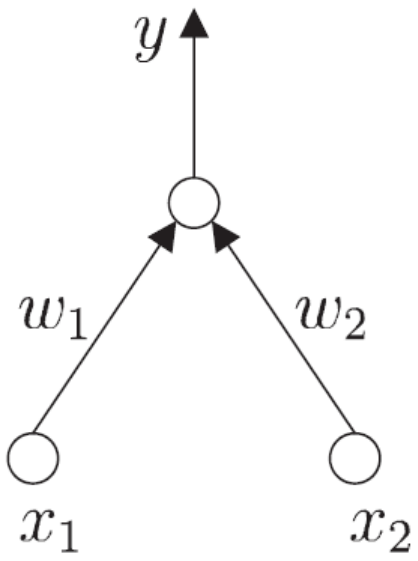

$$x_i \in \{0,1\} \text{ and } y \in \{0,1\}$$

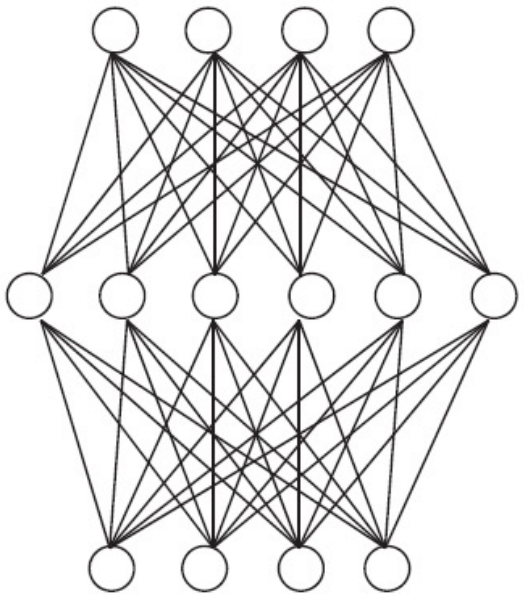
$$z(x_1, x_2, , \ldots, x_d) = z(\mathbf{x}) = \sum_{j=1}^{d} w_j \cdot x_j$$

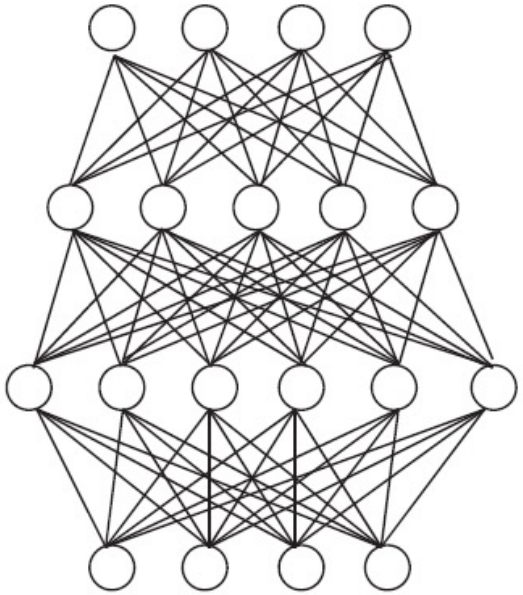
$$y = g(z(\mathbf{x})) = \begin{cases} 1 & \text{if } z(\mathbf{x}) \geq T \\ 0 & \text{if } z(\mathbf{x}) < T \end{cases}$$

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 Williams1986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

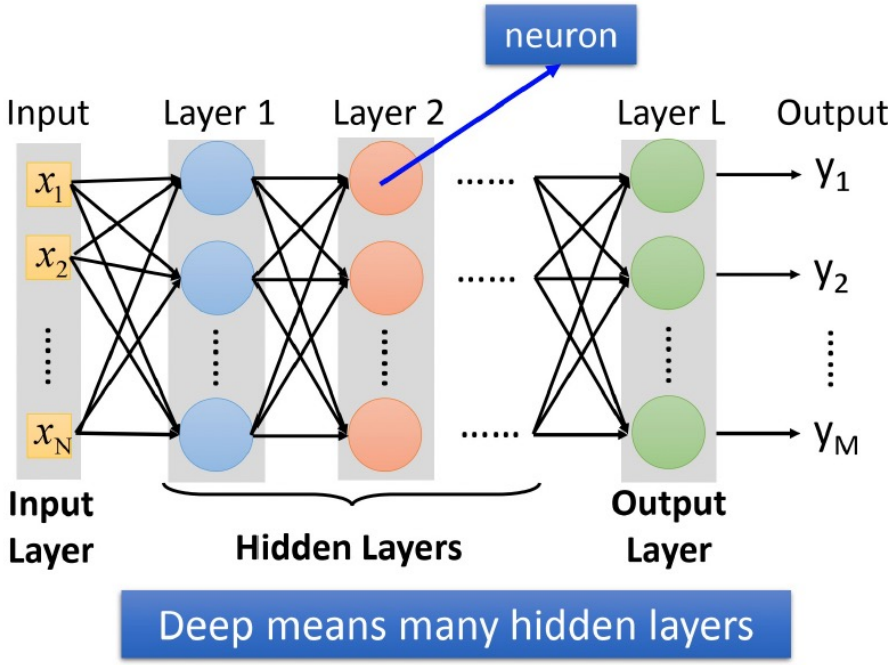
David Rumelhart

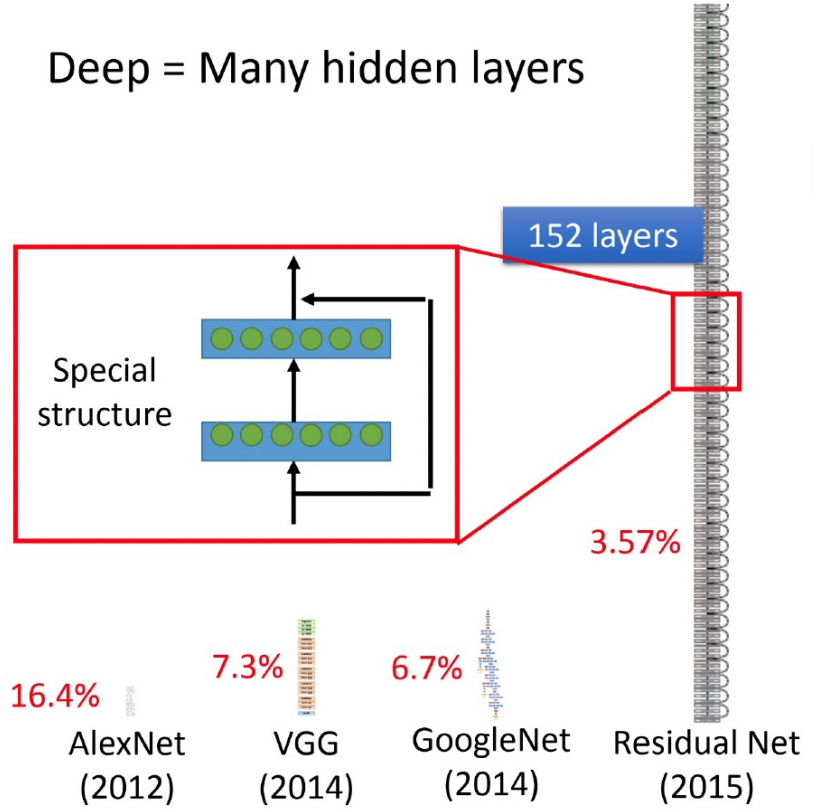
Geoffrey Hinton

Ronald Williams

# 深度学习(Deep Learning)

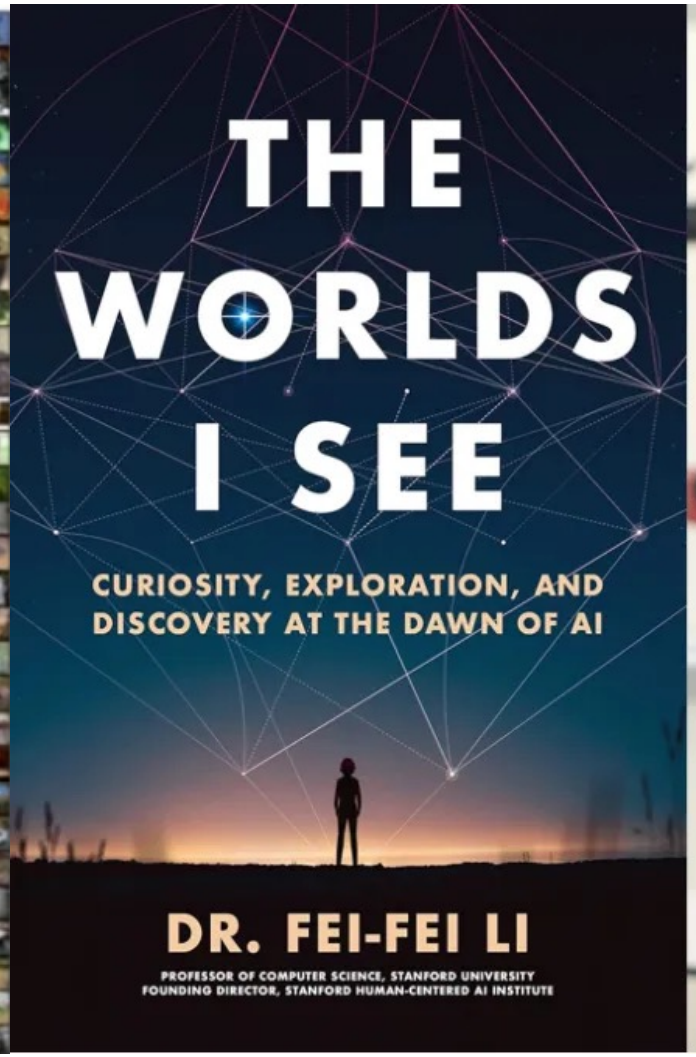
- 2006年, Hinton 在 Science 发表文章《Reducing the dimensionality of data with neural networks》，提出深度信念网络（Deep Belief Networks，DBNs）
- **深度学习模型就是具有很多个隐层的神经网络**



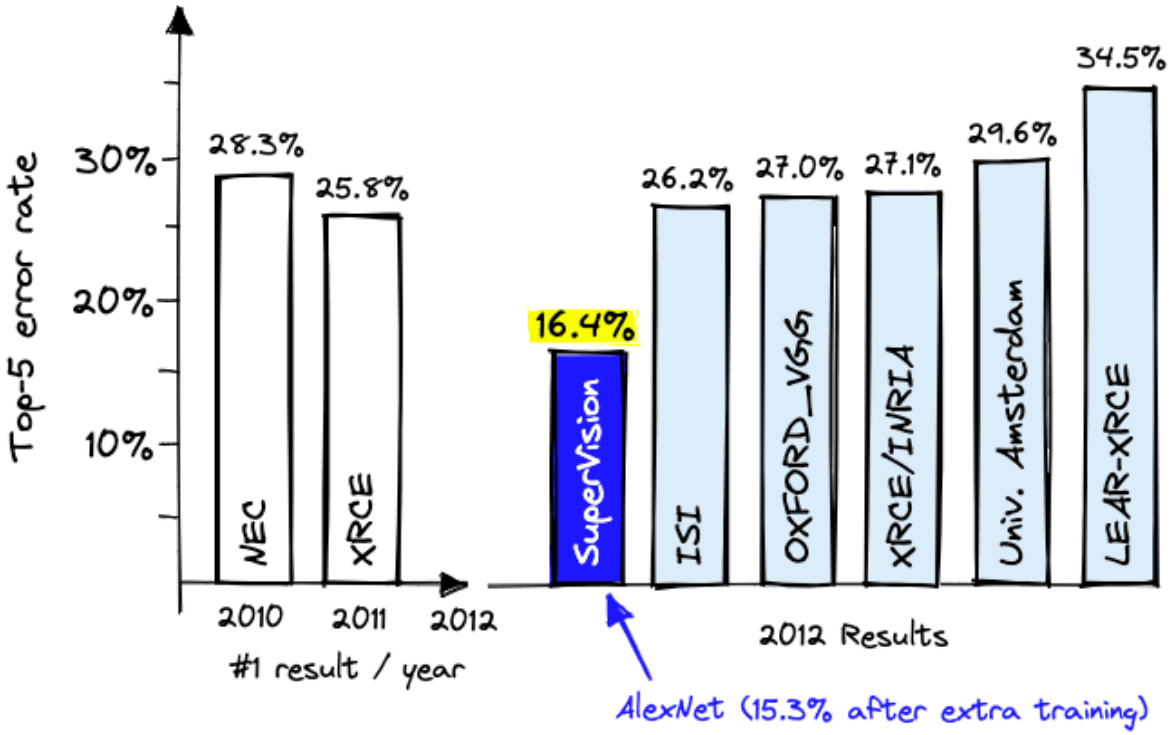
# ImageNet

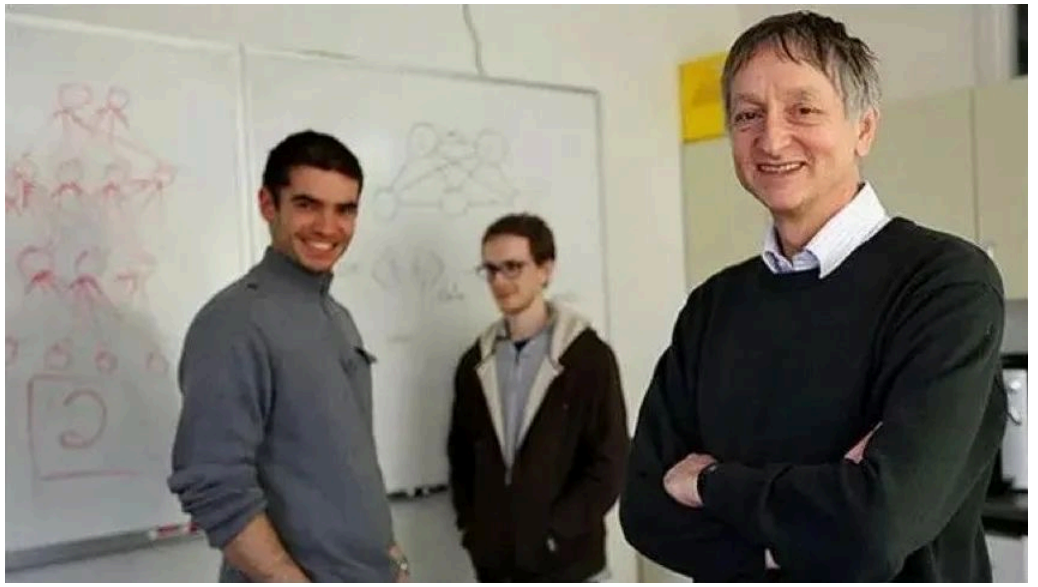

IMAGENET


THE WORLDS I SEE
CURIOSITY, EXPLORATION, AND DISCOVERY AT THE DAWN OF AI
DR. FEI-FEI LI
PROFESSOR OF COMPUTER SCIENCE, STANFORD UNIVERSITY
FOUNDING DIRECTOR, STANFORD HUMAN-CENTERED AI INSTITUTE

# 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**

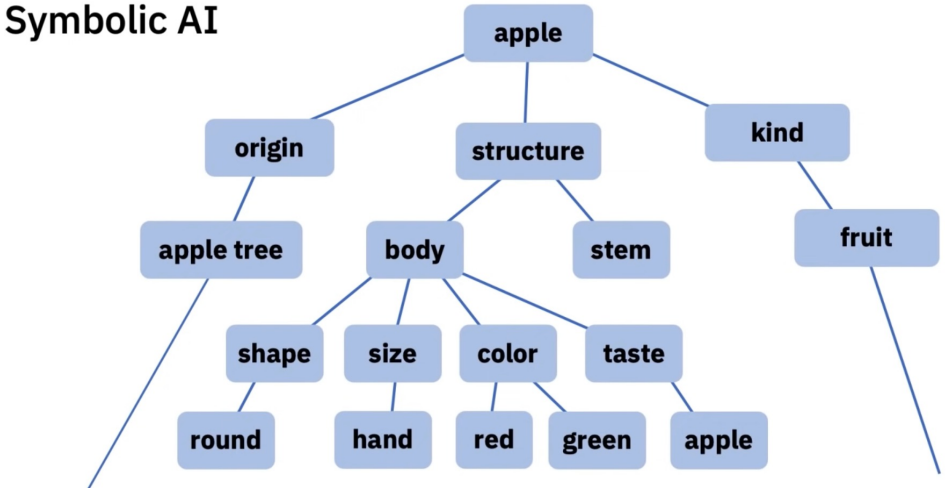
**计算能力提升，缓解了训练效率**

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

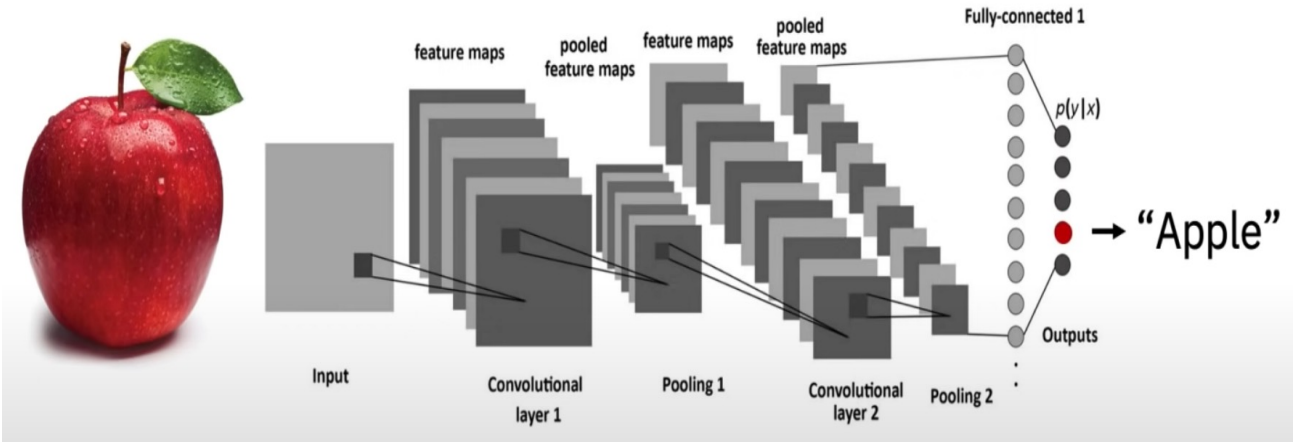
### 知识驱动的人工智能




- 可信可解释
- 与人类推理过程一致

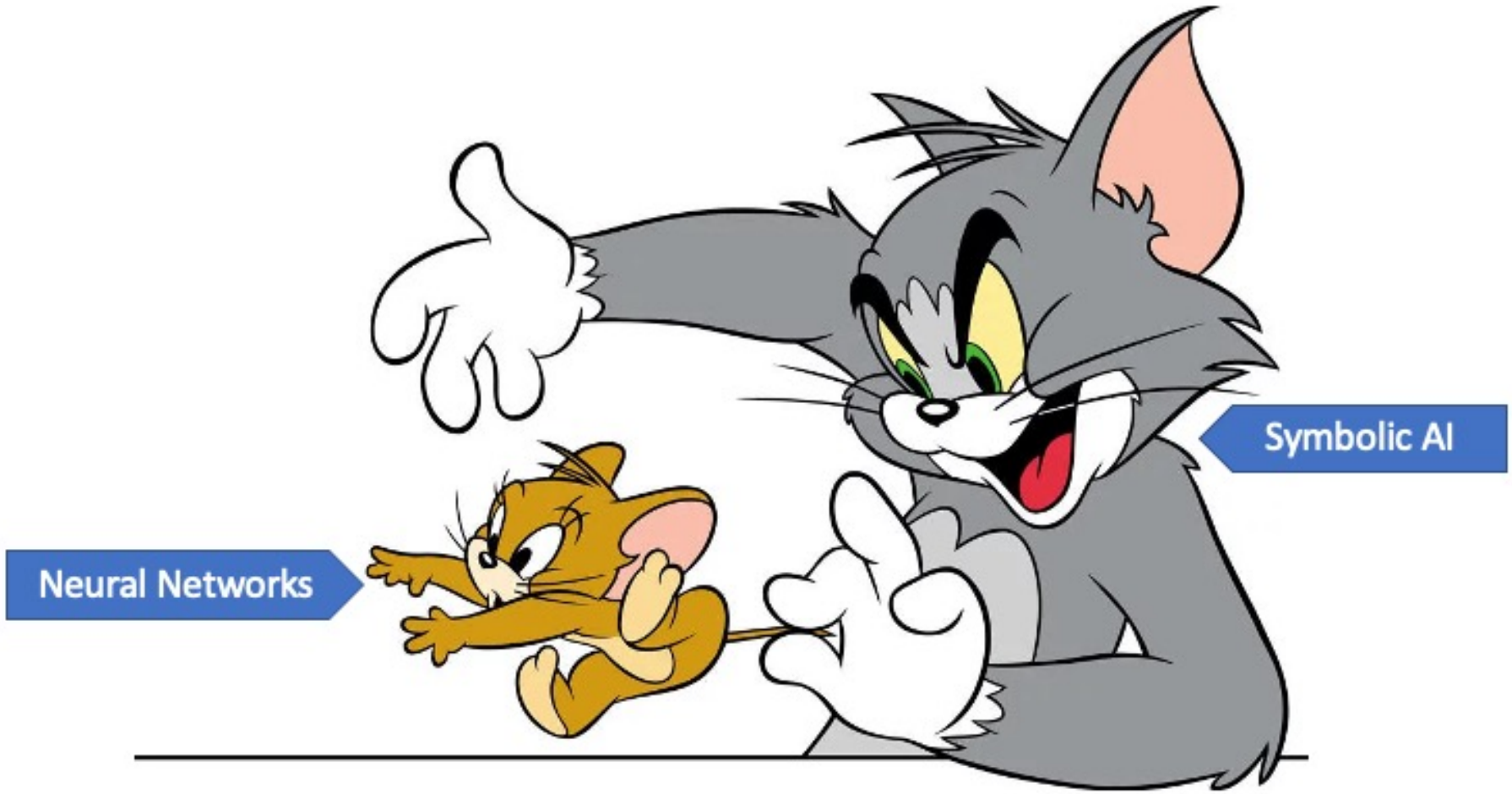
## 连接主义

### 数据驱动的人工智能

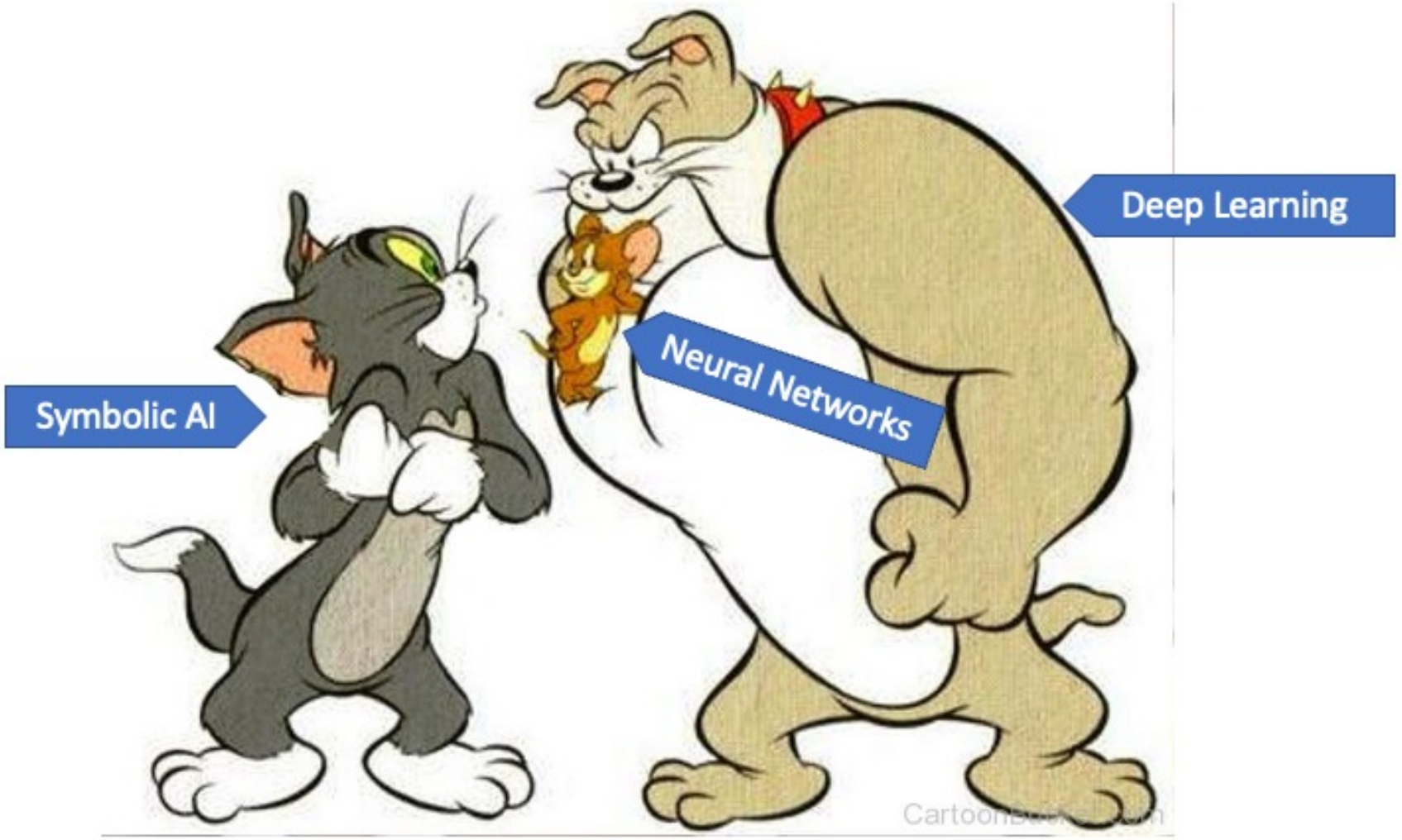



- 善于利用海量数据
- 无需领域知识，门槛低

# 深度学习 (Deep Learning)



# 深度学习 (Deep Learning)



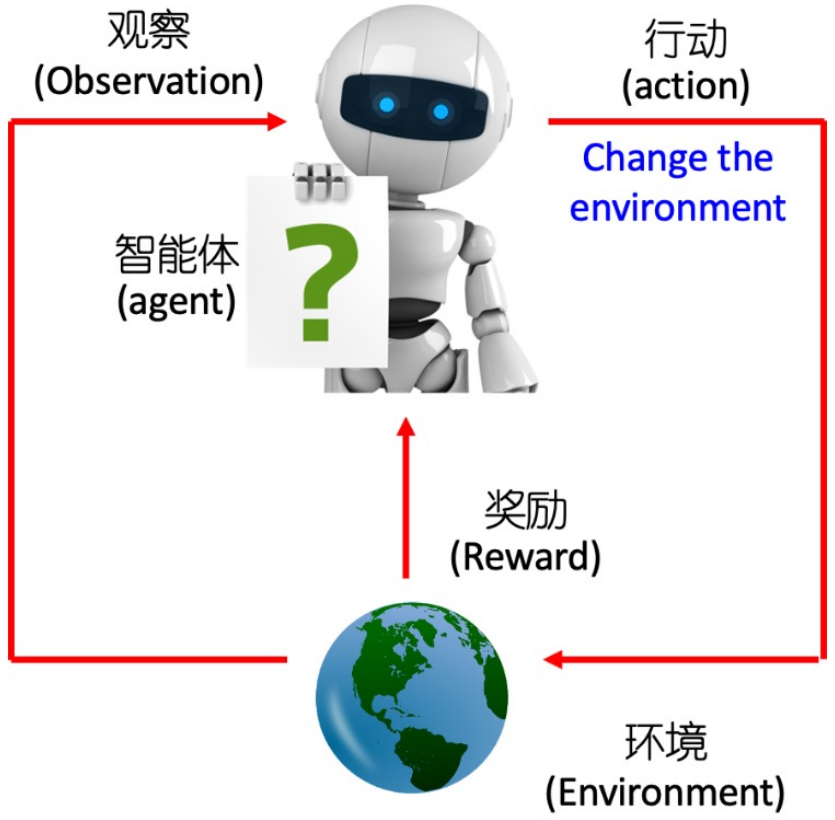
# 行为主义(Behaviorist AI)

**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

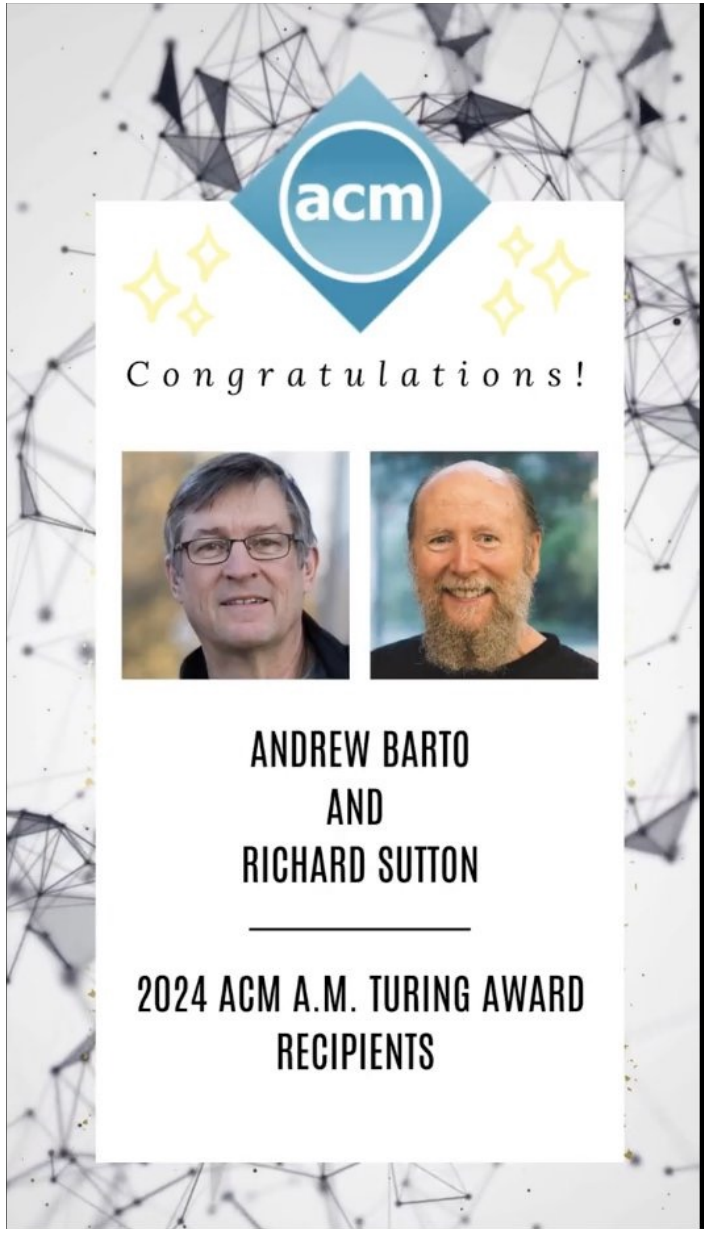
**强化学习**

**(Reinforcement Learning)**



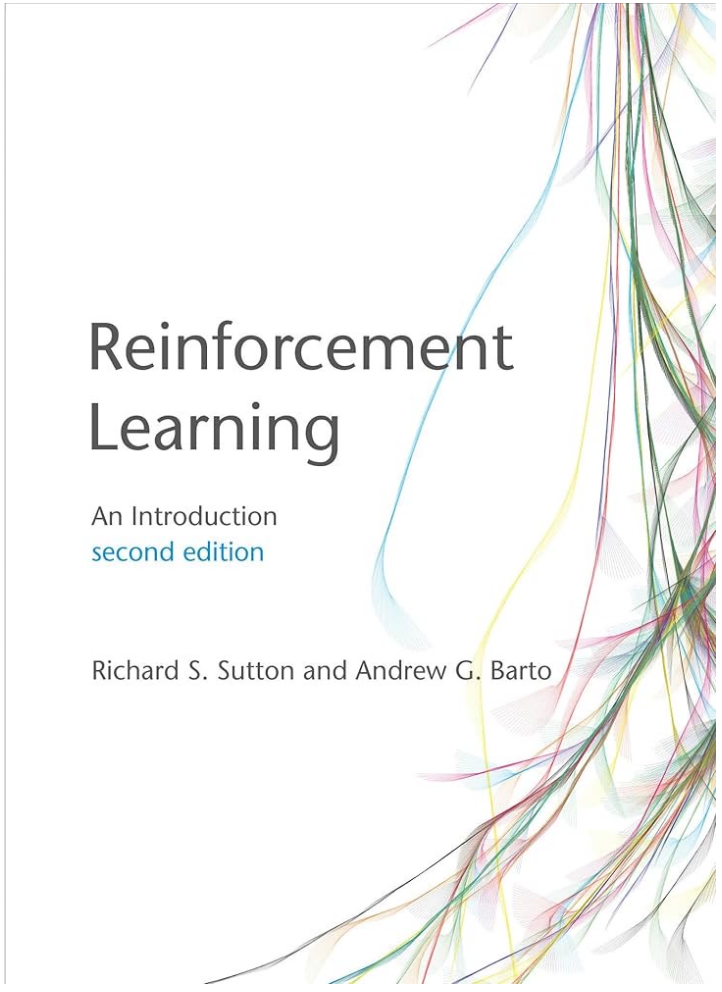
# 强化学习 (Reinforcement Learning)

# 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 **奠定了人工智能领域基础**
- John McCarthy：1971年图灵奖
- Allen Newell，Herbert Simon ： 1975年图灵奖 **问题求解的符号模型**
- Ed Feigenbaum，Raj Reddy：1994年图灵奖 **知识工程与专家系统**
- Judea Pearl：2011年图灵奖 **概率因果推理与贝叶斯网**
- Yoshua Bengio, Geoffrey Hinton, Yann Lecun：2018年图灵奖 **深度学习**
- Andrew Barto，Richard Sutton ： 2024年图灵奖 **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

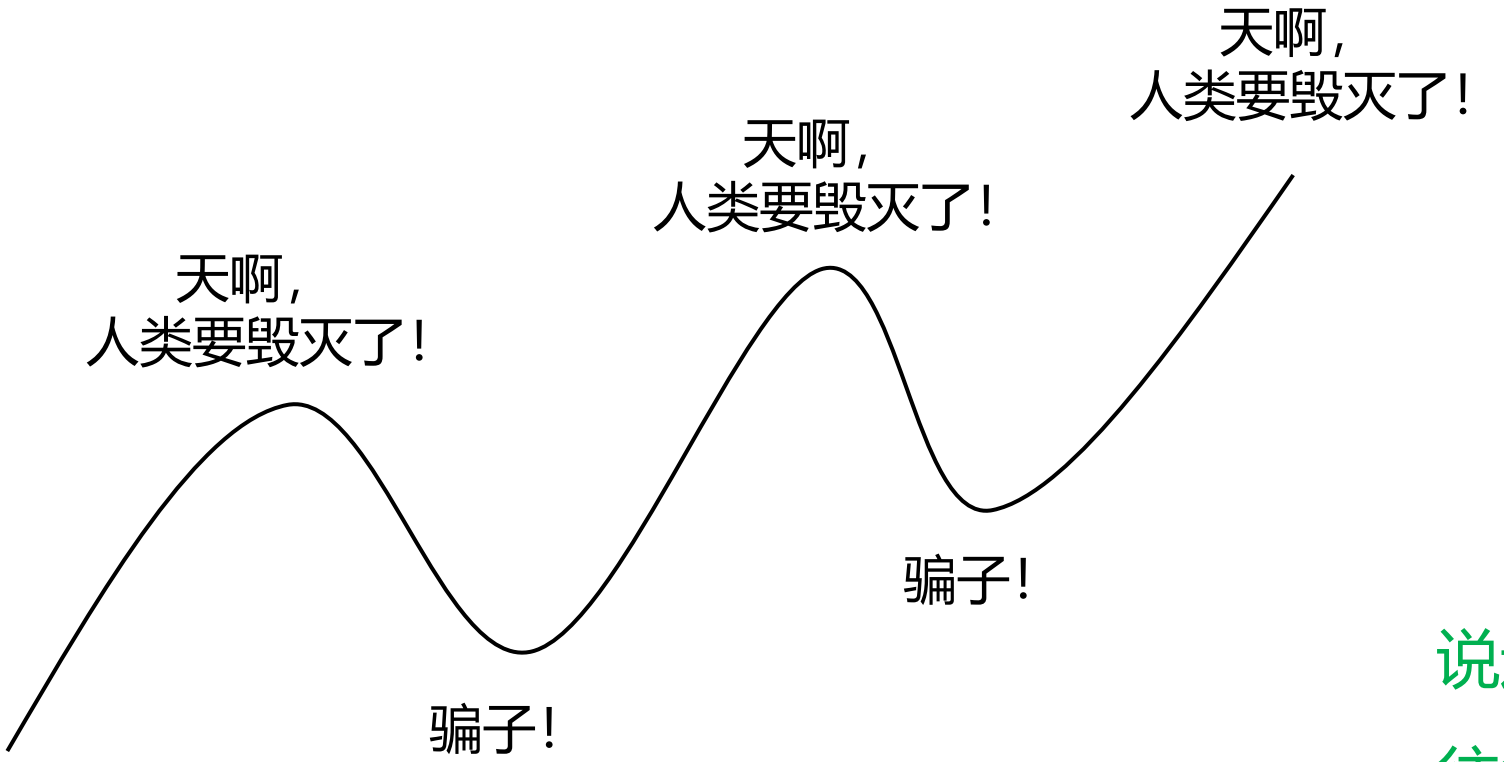
AlphaGo综合使用三种方法：强化学习、蒙特卡洛树搜索、深度学习，
综合行为主义、符号主义、连接主义的技术

**各个流派的融合发展是大势所趋**

# “人工智能”无所不能吗？

# 还会不会有下一次AI寒冬？

# 人工智能发展过程




说这些的往往是同一群人，

往往并不真正了解人工智能

技术的进展和局限

# 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**

# 可信性问题

**You**

请介绍一下林黛玉倒拔垂杨柳的故事

**ChatGPT**

林黛玉倒拔垂杨柳是《红楼梦》中的一个描写，发生在小说的第五十六回。这一场景表现了林黛玉情感的纠葛和心境的复杂。

故事发生在贾府中，林黛玉因为家族事务和个人感情的纷扰而感到烦闷和心情沉重。在贾府的后花园有一片垂杨柳，林黛玉在心情沮丧的情况下，倒拔起这片垂杨柳。这个举动被认为是林黛玉情感的宣泄，同时也反映了她坚韧、叛逆的一面。

垂杨柳在中国文学中通常被赋予柔弱、婉约的象征意义，而林黛玉倒拔垂杨柳的举动则被解读为她个性的张扬和对命运的反抗。整个场景揭示了林黛玉内心的矛盾、痛苦与坚持，是《红楼梦》中一个富有象征意味的描写。这一情节也成为后来研究《红楼梦》时常被引用和评论的经典场景之一。

# 可信性问题

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。

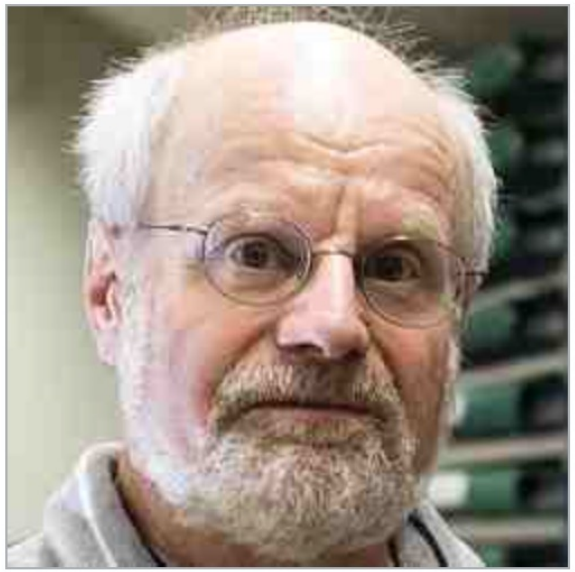
仔细观察图片，红色的球有 8 个，黄色的球有 11 个。

# 可信性问题

## ChatGPT Invents a Lot of Nonsense

By Herbert Bruderer
June 19, 2023
Comments (1)

VIEW AS: SHARE:



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

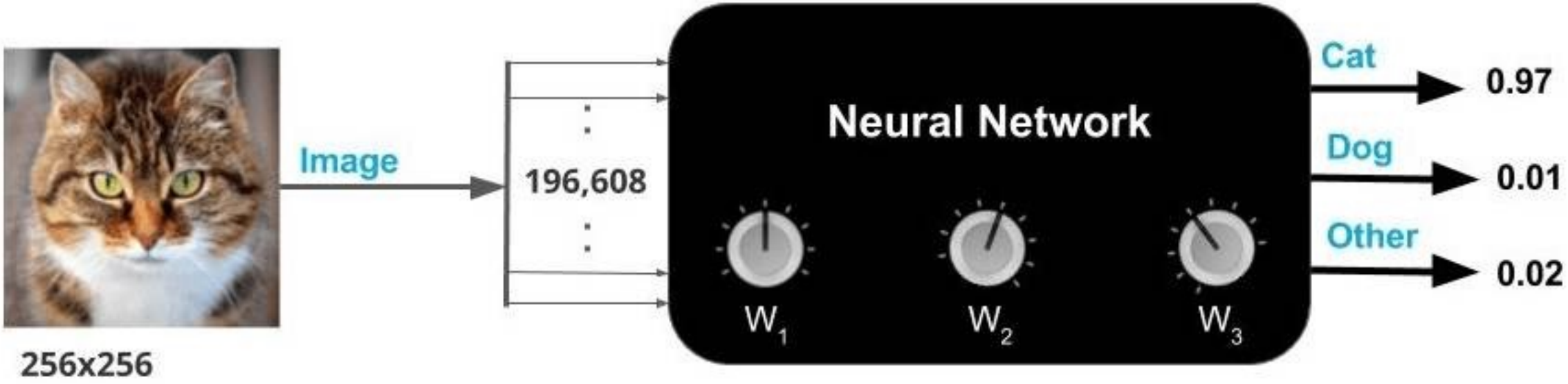
可靠

……?

# 缺乏知识

**纯数据驱动，缺乏人类世界的常识**

# 可解释性



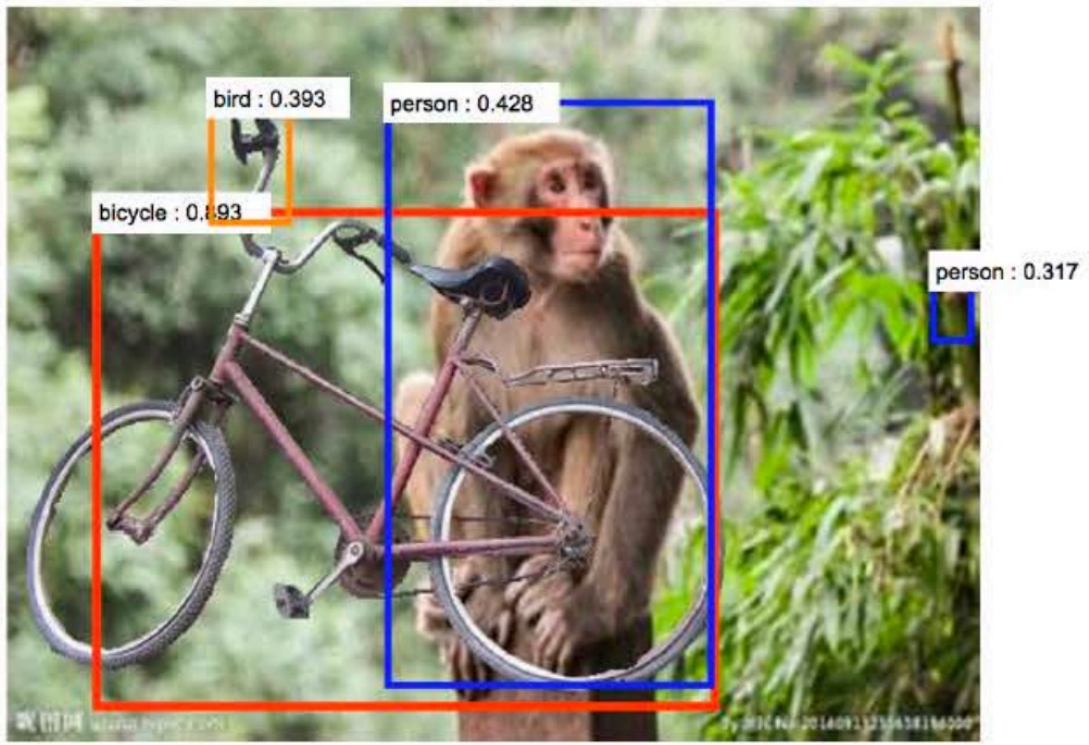

**可解释性挑战：模型黑盒，难以得知模型决策逻辑**

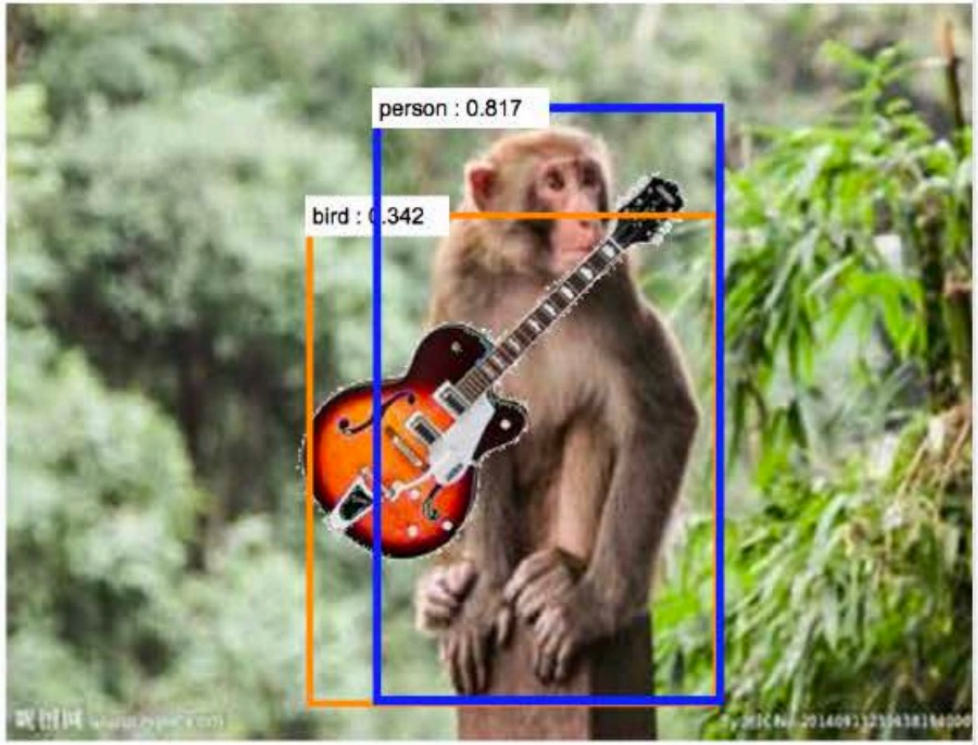
# 泛化性

## Teddy Bear

Meret Oppenheim, *Le Déjeuner en fourrure*

# 泛化性



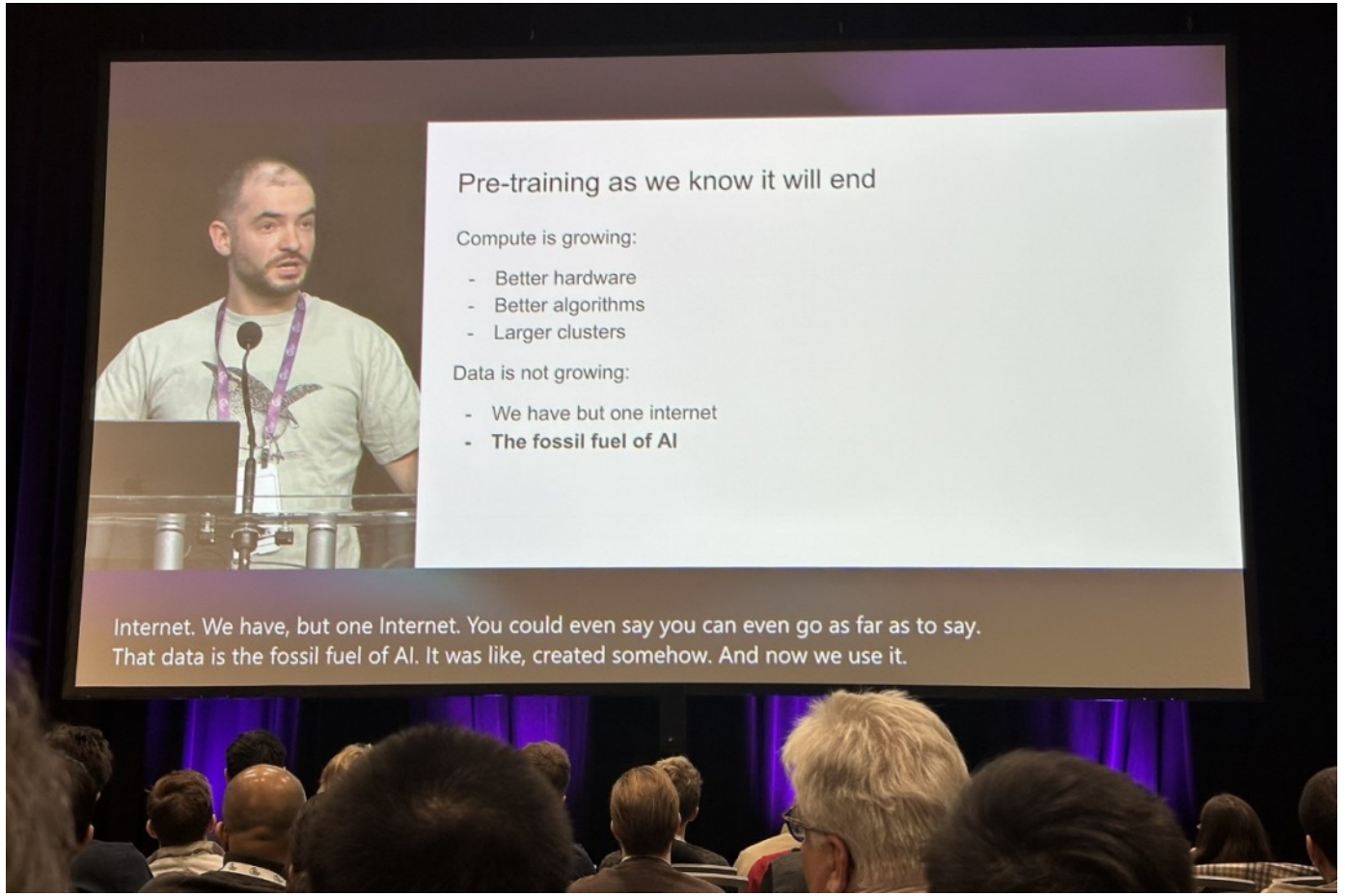

Wang et al. 2018

**泛化性挑战：难泛化到未见场景**

# 数据有限

“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**



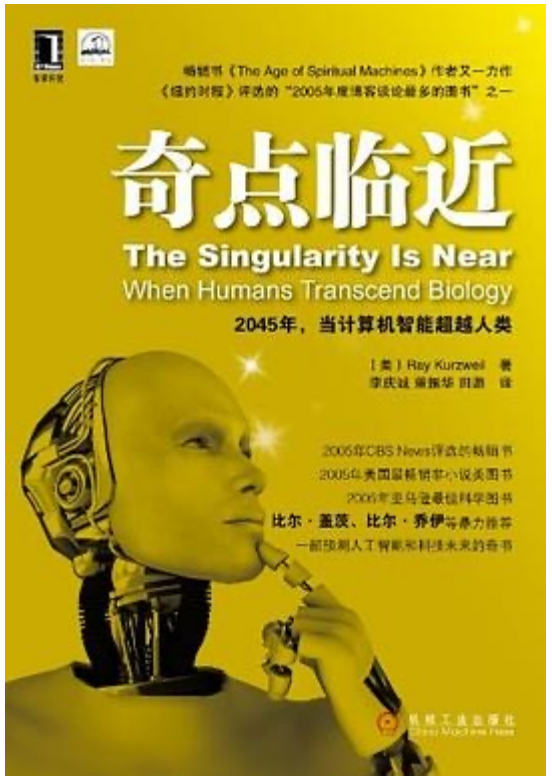

预训练时代将终结

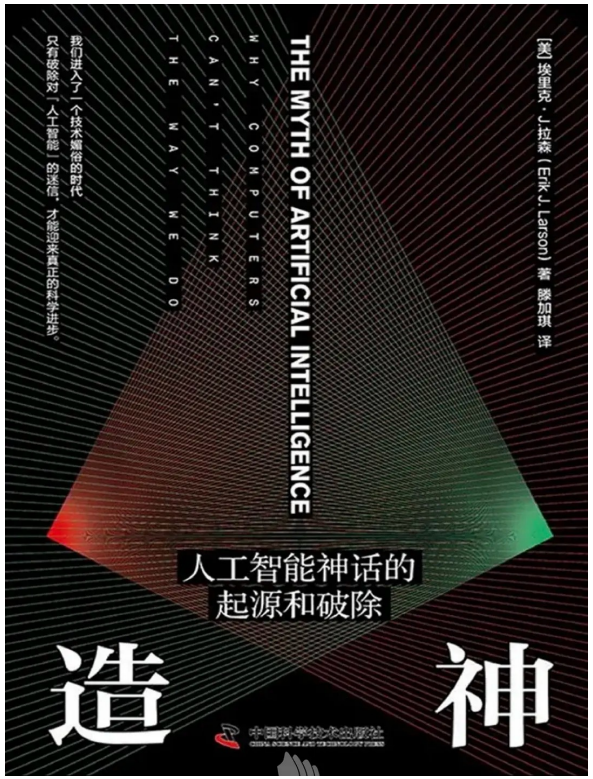
人工智能技术已取得巨大进展

但还有大量困难有待解决

任重而道远

# 一些书籍


奇点临近
The Singularity Is Near
When Humans Transcend Biology
2045年，当计算机智能超越人类
[美] Ray Kurzweil 著


THE MYTH OF ARTIFICIAL INTELLIGENCE
WHY COMPUTERS CAN'T THINK THE WAY WE DO
[美] 埃里克·J·拉森 (Erik J. Larson) 著 滕加琪 译
人工智能神话的起源和破除
造神


CHEERS
湛庐
如何创造可信的AI
Rebooting AI
更懂大脑的人工智能"大牛"盖瑞·马库斯绘制AI的未来路线图
By Gary Marcus Ernest Davis
龙志勇 译
[美] 盖瑞·马库斯 欧内斯特·戴维斯 著
Building Artificial Intelligence We Can Trust

# 本课程内容

Search & Planning 搜索与规划

Knowledge & Reasoning 知识与推理

Machine Learning 机器学习

Applications 应用

# 人工智能的前沿知识如何获取?

# 发展迅速、日新月异

# 仅靠书本内容远远不够

# 会议论文

## 虚假的三大会

ICML


NeurIPS

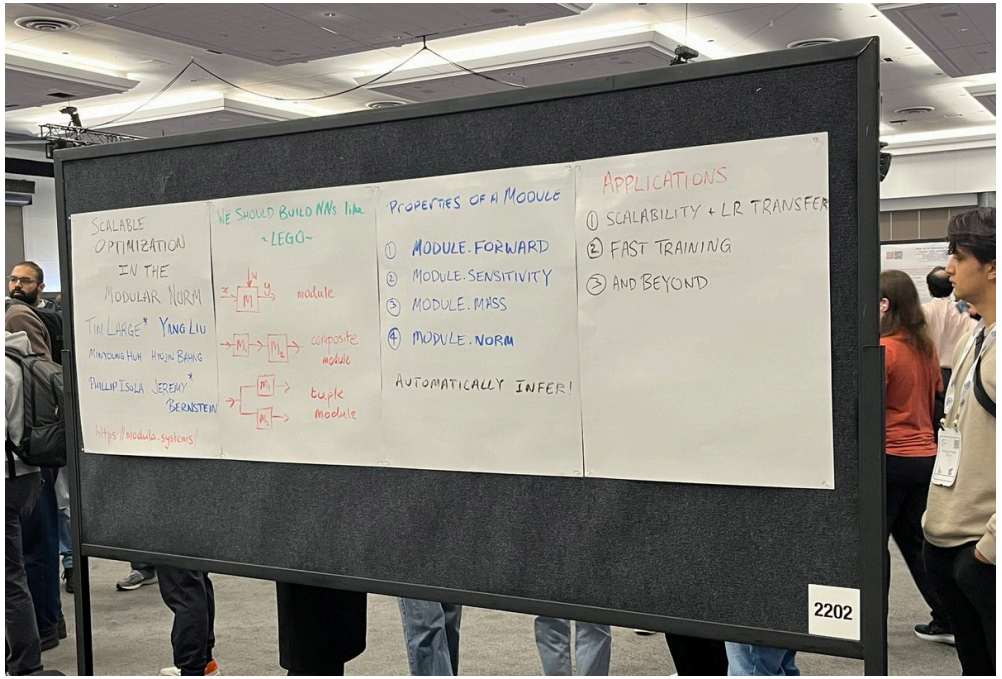

ICLR


## 真正的三大会



# 学术会议

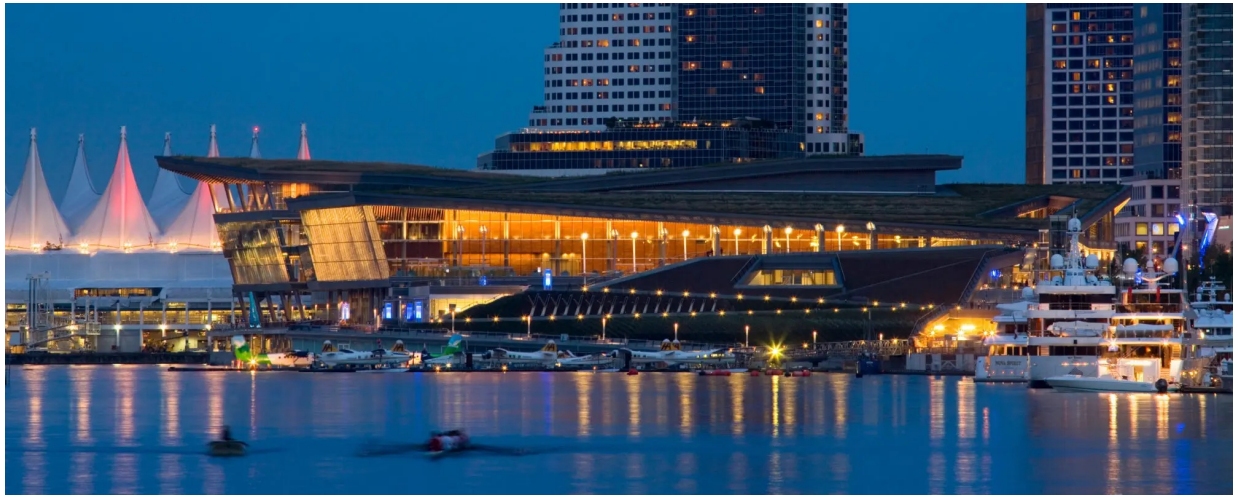
- **Tutorial**：为某个领域提供系统性的知识讲解，帮助参会者理解某个领域的方法与技术，面向初学者或对某个特定主题感兴趣的研究者
- **Main Conference**：Keynote Talks，Oral Presentation, Poster Presentation
- **Workshop**：侧重某个主题的深入研讨



# ICML (International Conference on Machine Learning)

**国际机器学习会议**

- 前身：机器学习研讨会(Machine Learning Workshop)
- 1988年，正式更名为ICML
- 每年在全球不同城市举办一次，一般为每年7-8月份



ICML 2025，加拿大温哥华，2025.07.13-2025.07.19

# ICML (International Conference on Machine Learning)

## 国际机器学习会议

- 前身：机器学习研讨会(Machine Learning Workshop)
- 1988年，正式更名为ICML
- 每年在全球不同城市举办一次，一般为每年7-8月份

ICML
International Conference
On Machine Learning

Topics of interest include (but are not limited to):

- General Machine Learning (active learning, clustering, online learning, ranking, supervised, semi- and self-supervised learning, time series analysis, etc.)
- Deep Learning (architectures, generative models, theory, etc.)
- Evaluation (methodology, meta studies, replicability and validity, human-in-the-loop)
- Theory of Machine Learning (statistical learning theory, bandits, game theory, decision theory, etc.)
- Machine Learning Systems (e.g., improved implementation and scalability, hardware, libraries, distributed methods)
- Optimization (convex and non-convex optimization, matrix/tensor methods, stochastic, online, non-smooth, composite, etc.)
- Probabilistic Methods (Bayesian methods, graphical models, Monte Carlo methods, etc.)
- Reinforcement Learning (e.g., decision and control, planning, hierarchical RL, robotics)
- Trustworthy Machine Learning (causality, fairness, interpretability, privacy, robustness, safety, etc.)
- Application-Driven Machine Learning (innovative techniques, problems, and datasets that are of interest to the machine learning community and driven by the needs of end-users in applications such as healthcare, physical sciences, biosciences, social sciences, sustainability and climate, etc.)

ICML论文集：
https://openreview.net/group?id=ICML.cc&referrer=%5BHomepage%5D(%2F)

# NeurIPS (Conference on Neural Information Processing Systems)

**神经信息处理系统大会**

- 1987–2000: Denver, Colorado, United States
- 2001–2010: Vancouver, British Columbia, Canada
- ……
- 2024: Vancouver, British Columbia, Canada
- **1987 – 2017, NIPS;  2018 – now, NeurIPS**
- MIT Press (1994–2004) and Curran Associates (2005–pre


早期 NeurIPS 涵盖了广泛的主题，从解决纯工程问题到使用计算机模型作为理解生物神经系统的工具，最近主要由机器学习、人工智能和统计学方面的论文主导

# NeurIPS (Conference on Neural Information Processing Systems)

**神经信息处理系统大会**


论文集：https://papers.nips.cc/

Main Track

Dataset and Benchmark Track (2021 - )

The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024) is an interdisciplinary conference that brings together researchers in machine learning, neuroscience, statistics, optimization, computer vision, natural language processing, life sciences, natural sciences, social sciences, and other adjacent fields. We invite submissions presenting new and original research on topics including but not limited to the following:

- Applications (e.g., vision, language, speech and audio, Creative AI)
- Deep learning (e.g., architectures, generative models, optimization for deep networks, foundation models, LLMs)
- Evaluation (e.g., methodology, meta studies, replicability and validity, human-in-the-loop)
- General machine learning (supervised, unsupervised, online, active, etc.)
- Infrastructure (e.g., libraries, improved implementation and scalability, distributed solutions)
- Machine learning for sciences (e.g. climate, health, life sciences, physics, social sciences)
- Neuroscience and cognitive science (e.g., neural coding, brain-computer interfaces)
- Optimization (e.g., convex and non-convex, stochastic, robust)
- Probabilistic methods (e.g., variational inference, causal inference, Gaussian processes)
- Reinforcement learning (e.g., decision and control, planning, hierarchical RL, robotics)
- Social and economic aspects of machine learning (e.g., fairness, interpretability, human-AI interaction, privacy, safety, strategic behavior)
- Theory (e.g., control theory, learning theory, algorithmic game theory)

# ICLR (International Conference on Learning Representations)

## 国际表示学习会议

- ICLR 2013, Scottsdale, Arizona
- ……
- ICLR 2023, Kigali, Rwanda
- ICLR 2024, Vienna, Austria
- ICLR 2025, Singapore


Organization >

People

General Chairs

- Yoshua Bengio, *Université de Montreal*
- Yann LeCun, *New York University*

ICLR
International Conference On
Learning Representations

# ICLR (International Conference on Learning Representations)

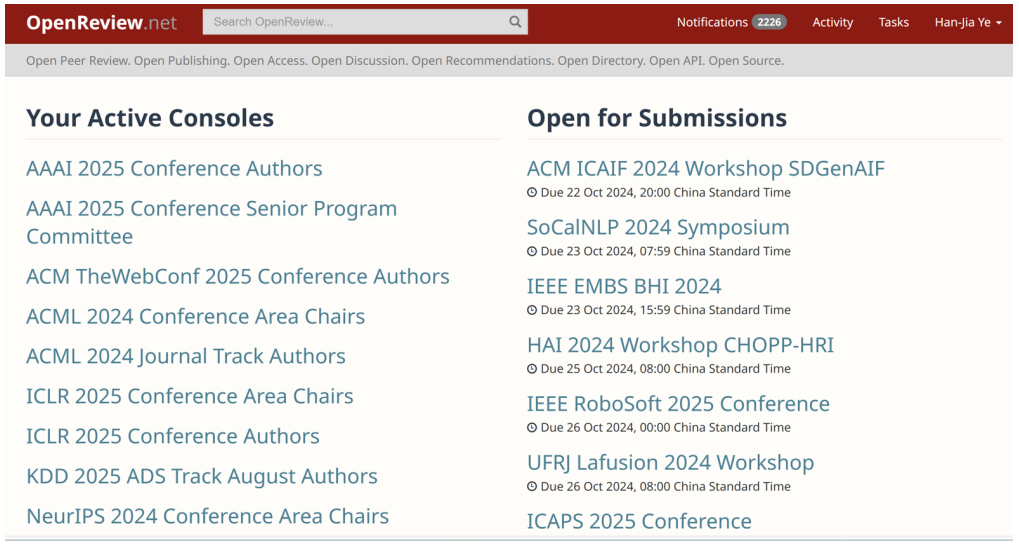
Yann LeCun

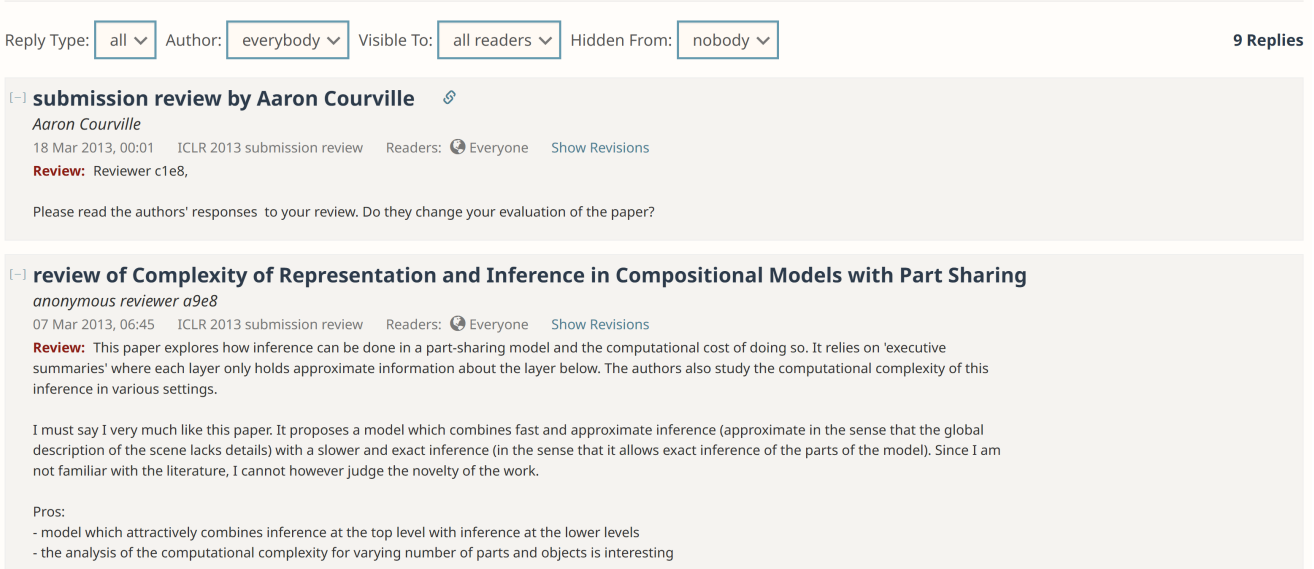
Many computer Science researchers are complaining that our emphasis on highly selective conference publications, and our double-blind reviewing system stifles innovation and slow the rate of progress of Science and technology.

This pamphlet proposes a new publishing model based on an open repository and open (but anonymous) reviews which creates a "market" between papers and reviewing entities.

——A New Publishing Model in Computer Science

OpenReview系统



# ICLR (International Conference on Learning Representations)

## 国际表示学习会议

论文集：
https://openreview.net/group?id=ICLR.cc&referrer=%5BHomepage%5D(%2F)

A non-exhaustive list of relevant topics explored at the conference include:

- unsupervised, semi-supervised, and supervised representation learning
- representation learning for planning and reinforcement learning
- representation learning for computer vision and natural language processing
- metric learning and kernel learning
- sparse coding and dimensionality expansion
- hierarchical models
- optimization for representation learning
- learning representations of outputs or states
- optimal transport
- theoretical issues in deep learning
- societal considerations of representation learning including fairness, safety, privacy, and interpretability, and explainability
- visualization or interpretation of learned representations
- implementation issues, parallelization, software platforms, hardware
- climate, sustainability
- applications in audio, speech, robotics, neuroscience, biology, or any other field

# Others

- 经典AI会议
  - AAAI (Association for the Advancement of Artificial Intelligence)
  - IJCAI (International Joint Conferences on Artificial Intelligence)
- 机器学习会议
  - AISTATS (International Conference on Artificial Intelligence and Statistics)
  - COLT (Conference on Learning Theory)
- 数据挖掘会议
  - KDD (SIGKDD Conference on Knowledge Discovery and Data Mining)
- 计算机视觉会议
  - CVPR (Conference on Computer Vision and Pattern Recognition)
  - ICCV (International Conference on Computer Vision)
- 自然语言处理会议
  - ACL (Annual Meeting of the Association for Computational Linguistics)
  - EMNLP (Conference on Empirical Methods in Natural Language Processing)

# MLA

China Symposium on Machine Learning and Applications

## 中国机器学习及其应用研讨会

"机器学习及其应用"研讨会自2002年开始，先后在上海、南京、北京、西安、天津等地举行。该研讨会每年邀请海内外从事机器学习及相关领域研究的专家与会进行学术交流。研讨会不征文，不收取注册费，欢迎机器学习及相关领域的学者、同行、研究生前来旁听特邀报告并参加讨论。为了促进机器学习及相关领域的研究生之间以及研究生与资深学者之间的交流，2006-2010年在机器学习及其应用研讨会（MLA）期间，同时举行了机器学习及其应用学生研讨会（SSMLA），此后该研讨会融入MLA的Poster session。

以下是各年会议的信息：

| | | |
|---|---|---|
| MLA'23 | | 2023年11月,南京大学 |
| MLA'22 | | 2022年11月,南京大学 |
| MLA'21 | | 2021年12月,南京航空航天大学 |
| MLA'20 | | 2020年11月,南京大学 |
| MLA'19 | | 2019年11月,天津大学 |
| MLA'18 | | 2018年11月,南京大学 |
| MLA'17 | | 2017年11月,北京交通大学 |
| MLA'16 | | 2016年11月,南京大学 |
| MLA'15 | | 2015年11月,南京大学 |
| MLA'14 | | 2014年11月,西安电子科技大学 |
| MLA'13 | | 2013年11月,复旦大学 |
| MLA'12 | | 2012年11月,清华大学 |
| MLA'11 | | 2011年11月,清华大学 |
| MLA'10 | SSMLA'10 | 2010年11月,南京大学 |
| MLA'09 | SSMLA'09 | 2009年11月,南京大学 |
| MLA'08 | SSMLA'08 | 2008年11月,南京大学 |
| MLA'07 | SSMLA'07 | 2007年11月,南京大学、南京师范大学 |
| MLA'06 | SSMLA'06 | 2006年11月,南京大学、南京航空航天大学 |
| MLA'05 | | 2005年11月,南京大学 |

# 学术期刊

- Nature Machine Intelligence
- AIJ 《Artificial Intelligence》
- JMLR 《Journal of Machine Learning Research》
- TPAMI《IEEE Trans. on Pattern Analysis and Machine Intelligence》
- TKDE《IEEE Trans. on Knowledge and Data Engineering》
- MLJ 《Machine Learning》
- TNNLS 《IEEE Trans. on Neural Network and Learning Systems》
- 国内：《中国科学 信息科学》
- …

# 小结

- 了解人工智能的研究范畴
- 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