# 人工智能基础

吉建民

USTC
jianmin@ustc.edu.cn

2024 年 2 月 21 日

# Used Materials

Disclaimer: 本课件采用了 S. Russell and P. Norvig's Artificial Intelligence –A modern approach slides, 徐林莉老师课件和其他网络课程课件，也采用了 GitHub 中开源代码，以及部分网络博客内容

# Table of Contents

# 《人工智能基础》

- 教材
  - Artificial Intelligence – A Modern Approach (3rd Edition)
    S. Russell and P. Norvig
  - 人工智能——一种现代方法
- 参考书
  - 机器学习 周志华 (2016)
  - 动手学深度学习（第二版）
- 课程考核
  - 学期总评 = 期末考试 (60%) + 书面作业 (15%) + 实验部分 (25%)
- 课程主页
  http://staff.ustc.edu.cn/~jianmin/2024/ai2024/
- 课件下载
  http://staff.ustc.edu.cn/~jianmin/lecture/ai2024/lec1.pdf
  http://staff.ustc.edu.cn/~jianmin/lecture/ai2024/lec2.pdf

⋮

# 课程内容

- 第一部分：人工智能概述 / Introduction and Agents (chapters 1, 2)
- 第二部分：问题求解 / Search (chapters 3 $\sim$ 6)
  - Solving Problems by Searching, Informed Search, Constraint Satisfaction Problems (CSP), Game Planning
- 第三部分：知识、推理与规划 / Logic, Knowledge, Reasoning, and Planning (chapters 7 $\sim$ 12)
  - Logical Agents, FOL and Inference in FOL, Planning and Knowledge Representation
- 第四部分：不确定知识与推理 / Uncertainty and Decision Making (chapters 13 $\sim$ 17)
  - Uncertainty and Bayesian Networks, Decision Making (MDPs, POMDPs, and Stochastic Games)
- 第五部分：学习 / Learning (chapters 18 $\sim$ 21)
  - Machine Learning, Deep Learning, Reinforcement Learning
- 第六部分：应用 / NLP, Perception, and Robotics (chapters 22 $\sim$ 25)

# 课程内容

- 广义上"人工智能"包括很多经典工作，如搜索、知识表示与推理、规划、MDPs、贝叶斯网等。
- 由于近期深度学习的成功，现在提到"人工智能"很多人主要指机器学习，尤其是深度学习、大模型方面的工作。
- 《人工智能基础》这门课的定位是：
    - 面向计算机专业的同学，介绍 IT 工程师应该掌握的人工智能基础技术和知识，要求大家熟练掌握；
    - 同时，作为索引介绍人工智能的各方面工作，尤其是机器学习、神经网络、大模型等方面的新工作，仅要求了解。
- 上完这门课，期望大家了解人工智能的全貌，掌握人工智能的基础技术和知识，了解机器学习的一些工作。
    - 想系统入门深度学习，建议自学"动手学深度学习"，实际动手 coding，听李沐的在线课程：`https://zh.d2l.ai/`
    - 想入门以后再精深，建议选定 CV、NLP、多模态等某个具体领域，从综述开始系统阅读核心论文、持续学习该领域的最新工作，自己动手实现系统，同时培养科研能力和工程能力；
    - 想精深后再成为某方向专家，建议加入相关实验室，培养自己对该方向的洞见，开展前沿研究。

# 助教

- 王若麟 violinwang@mail.ustc.edu.cn
- 许悦娇 xyj1017560205@mail.ustc.edu.cn
- 盛玉 Shengyu724@mail.ustc.edu.cn
- 杨环宇 yanghuanyu@mail.ustc.edu.cn
- 冯子洋 fzy617@mail.ustc.edu.cn
- 李豪雨 lihaoyu31801@mail.ustc.edu.cn
- QQ 群: 344989701

# Table of Contents

Different people think of AI differently

| | |
|---|---|
| "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ..." (Bellman, 1978) | "The study of mental faculties through the use of computational models" (Charniak+McDermott, 1985) |
| "The study of how to make computers do things at which, at the moment, people are better" (Rich+Knight, 1991) | "The branch of computer science that is concerned with the automation of intelligent behavior" (Luger+Stubblefield, 1993) |

Views of AI fall into four categories:

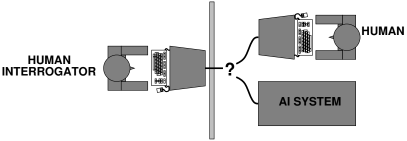
| | |
|---|---|
| Thinking humanly | Thinking rationally |
| Acting humanly | Acting rationally |

The textbook advocates "acting rationally (理性的)"

# Acting humanly: The Turing Test 图灵测试

Turing (1950) "Computing machinery and intelligence":

- "Can machines think?" $\longrightarrow$ "Can machines behave intelligently?"
- Operational test for intelligent behavior: the Imitation Game




- Predicted that by 2000, a machine might have a 30% chance of fooling a lay person for 5 minutes
- Anticipated all major arguments against AI in following 50 years
- Suggested major components of AI: knowledge (知识), reasoning (推理), language understanding (语言理解), learning (学习)

Problem: Turing test is not reproducible, constructive, or amenable to mathematical analysis

# Thinking humanly: Cognitive Science (认知科学)

- 1960s "cognitive revolution": information-processing psychology replaced prevailing orthodoxy of behaviorism
- Requires scientific theories of internal activities of the brain
  - What level of abstraction? "Knowledge" or "circuits"?
  - How to validate? Requires
    - Predicting and testing behavior of human subjects (top-down), or
    - Direct identification from neurological data (bottom-up)
- Both approaches (roughly, Cognitive Science and Cognitive Neuroscience) are now distinct from AI

# Thinking rationally: Laws of Thought

- Normative (or prescriptive) rather than descriptive
- Aristotle (亚里士多德): what are correct arguments/thought processes?
- Several Greek schools developed various forms of logic:

  *notation and rules of derivation for thoughts; may or may not have proceeded to the idea of mechanization*

- Direct line through mathematics and philosophy to modern AI
- Problems:
  - Not all intelligent behavior is mediated by logical deliberation
  - What is the purpose of thinking? What thoughts should I have?
  - Logical systems tend to do the wrong thing in the presence of uncertainty
    - 部分学者观点，部分学者不同意

# Acting rationally

- Rational behavior: doing the right thing
  - The right thing: that which is expected to maximize goal achievement, given the available information
  - Doesn't necessarily involve thinking—e.g., blinking reflex—but thinking should be in the service of rational action
  - Aristotle (Nicomachean Ethics, 亚里士多德所著《尼各马可伦理学》):

    > *Every art and every inquiry, and similarly every action and pursuit, is thought to aim at some good*

- Entirely dependent on goals!
- Irrational $\neq$ insane, irrationality is sub-optimal action
- Rational $\neq$ successful
- Our focus here:
  - Systems which make the best possible decisions given goals, evidence, and constraints
  - In the real world, usually lots of uncertainty
    - ... and lots of complexity
  - Usually, we're just approximating rationality

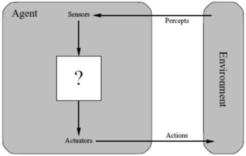
# Rational agents

*Maximize Your Expected Utility*

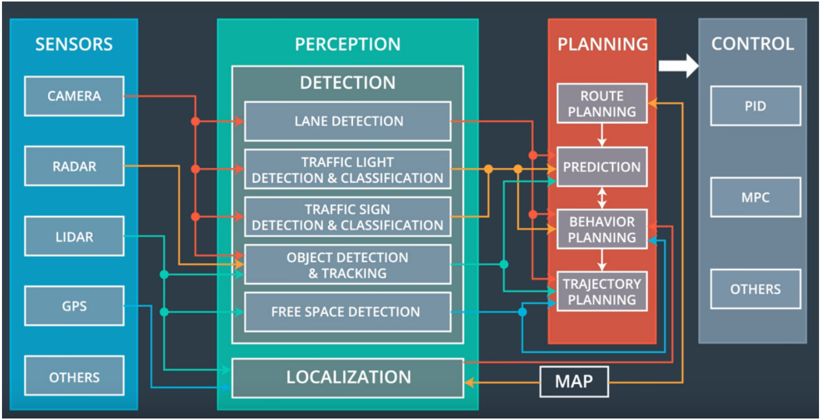
- An agent is an entity that perceives and acts
- This course is about designing rational agents
- Abstractly, an agent is a function from percept histories to actions:

$$f : \mathcal{P}^* \rightarrow \mathcal{A}$$

- For any given class of environments and tasks, we seek the agent (or class of agents) with the best performance
- Caveat: computational limitations make perfect rationality unachievable $\rightarrow$ design best program for given machine resources

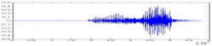


# Self Driving Car Architecture



# 机器学习 ≈ 拟合一个函数

- Speech Recognition

$f($  $)=$ "How are you"

- Image Recognition

$f($  $)=$ "Cat"

- Playing Go

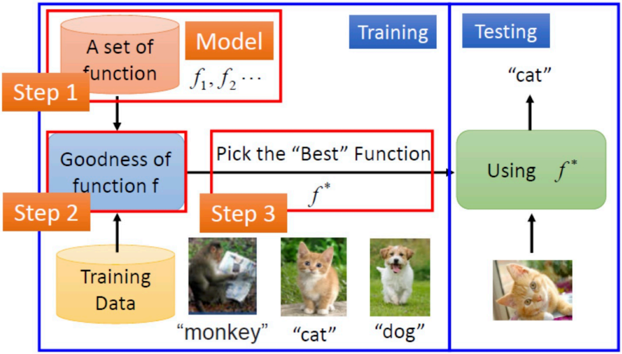
$f($  $)=$ "5-5" (next move)

- Dialogue System

$f($ "Hi" (what the user said) $)=$ "Hello" (system response)

# 学习 = 表示 + 评价 + 优化

- 表示（Representation）：确定假设空间（hypothesis space）
- 评价（Evaluation）：评价函数（目标函数、打分函数）来判断优劣
- 优化（Optimization）：需要一个搜索方法，能够在假设空间中找到评价函数得分最高的函数



# PAC Learning: 机器学习为什么可能？

- 泛化误差 (generalization error)：给定 $h \in \mathcal{H}$, c$\in \mathcal{C}$，分布 $\mathcal{D}$，$h$ 的泛化误差为

$$R(h) = P_{x \sim \mathcal{D}}[h(x) \neq c(x)]$$

- 经验误差 (empirical error)：给定样本集合 $S = (x_1, \dots, x_m)$

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} [h(x_i) \neq c(x_i)]$$

- 比较经验误差与泛化误差：
  - 经验误差是 $h$ 在训练数据 $S$ 上的平均误差
  - 泛化误差是 $h$ 在分布 $\mathcal{D}$ 上的期望误差
  - 两者存在联系，因为 $S$ 是根据 $\mathcal{D}$ 独立同分布产生的

$$E_{S \sim \mathcal{D}^m}[\hat{R}_S(h)] = R(h)$$

# PAC 学习理论

- PAC 学习理论考虑，能否从假设空间 $\mathcal{H}$ 中学习一个好的假设 $h$
- “好的假设” 需要满足两个条件 (PAC 辨识条件)：
  - 近似正确 (Approximately Correct)： 泛化误差 $R(h)$ 足够小
    - $R(h)$ 越小越好，最好泛化误差能能于 0，但一般是不可能的。那我们就把 $R(h)$ 限定在一个很小的数 $\epsilon$ 之内，即只要假设 $h$ 满足 $R(h) \leq \epsilon$，我们就认为 $h$ 是近似正确的
  - 可能正确 (Probably Correct)： $h$ 在很大概率上近似正确
    - 不指望选择的假设 $h$ 百分之百是近似正确的，即 $R(h) \leq \epsilon$，只要很可能是近似正确的就可以，即我们给定一个值 $\delta$，假设 $h$ 满足 $P(R(h) \leq \epsilon) \geq 1 - \delta$
- 同时学习所需的样本数量不能太大，样本数量是关于 $1/\epsilon$，$1/\delta$，样本大小，$size(c)$ 的多项式函数

# PAC Learnable

## 定义

A concept class $\mathcal{C}$ is said to be *PAC-learnable* if there exists an algorithm $\mathcal{A}$ and a polynomial function function $poly(\cdot, \cdot, \cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distribution $\mathcal{D}$ on $\mathcal{X}$ (containing instances of length $n$) and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$P_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

If $\mathcal{A}$ further runs in $poly(1/\epsilon, 1/\delta, n, size(c))$, then $\mathcal{C}$ is said to be *efficiently PAC-learable*. When such an algorithm $\mathcal{A}$ exists, it is called a *PAC-learning algorithm* for $\mathcal{C}$.

# 霍夫丁不等式（Hoeffding's Inequality）

**定理 (Hoeffding's Inequality)**

*Let* $X_1, \ldots, X_m$ *be i.i.d. random variable in [0, 1], for any* $\epsilon > 0$

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E(X_i)\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}.$$

- 该定理给出了位于区间 [0,1] 的两两随机变量其期望与均值之间满足的关系，在任意分布 $\mathcal{D}$ 下
- 由泛化误差 $R(h)$ 与经验误差 $\hat{R}_S(h)$ 的定义易知 $E(\hat{R}_S(h)) = R(h)$，由此得到

$$P_{S\sim\mathcal{D}^m}\left[\left|\hat{R}_S(h) - R(h)\right| \geq \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

- 令 $\delta = 2e^{-2m\epsilon^2}$，则 Fix a hypothesis $h$: $\mathcal{X} \to \{0, 1\}$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

# 奥卡姆剃刀原则

- Thus, for a finite hypothesis set $\mathcal{H}$,

$$R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\log_2 |\mathcal{H}|}{m}}\right)$$

- For a fixed $|\mathcal{H}|$, to attain the same guarantee as in the consistent case, a quadratically larger labeled sample is needed
- This can also be viewed as an instance of the so-called Occam's Razor principle
  - Plurality should not be posited without necessity, also rephrased as, the simplest explanation is best.

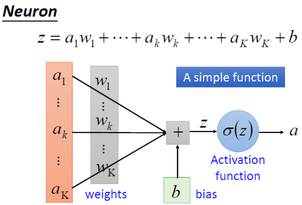
# 机器学习

目前机器学习解决问题的思路

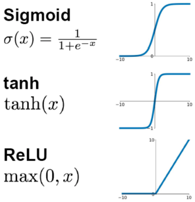

中间的三部分，概括起来就是特征表达。良好的特征表达，对最终算法的准确性起了非常关键的作用，而且系统主要的计算和测试工作都耗在这一大部分。但这块实际中一般都是人工完成的。靠人工提取特征


手工选取特征不太好，那么能不能自动地学习一些特征呢？能！
Deep Learning (Unsupervised Feature Learning)

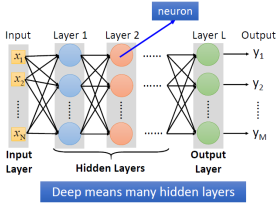
# 深度神经网络



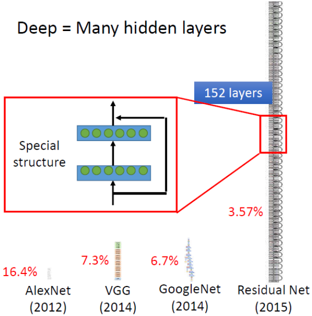
Neuron
$z = a_1 w_1 + \cdots + a_k w_k + \cdots + a_K w_K + b$
A simple function
$a_1$
$a_k$
$a_K$
$w_1$
$w_k$
$w_K$
weights
+
$z$
$\sigma(z)$
$a$
Activation function
$b$
bias

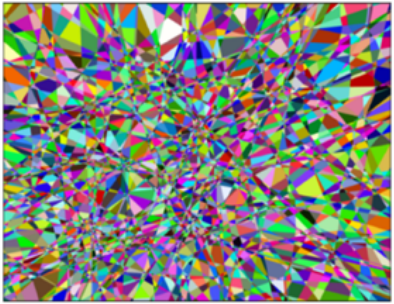
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$
tanh
$\tanh(x)$
ReLU
$\max(0, x)$

neuron
Input
Layer 1
Layer 2
Layer L
Output
$x_1$
$x_2$
$x_N$
$y_1$
$y_2$
$y_M$
Input Layer
Hidden Layers
Output Layer
Deep means many hidden layers

Deep = Many hidden layers
152 layers
Special structure
3.57%
16.4%
7.3%
6.7%
AlexNet (2012)
VGG (2014)
GoogleNet (2014)
Residual Net (2015)

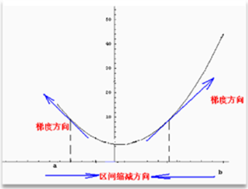



# 深度学习基本方法

- 为了让f(w, x)接近D， 我们定义了损失函数

$$L(w)=\sum_{d_i\in D} f(w,x_i)-y_i$$

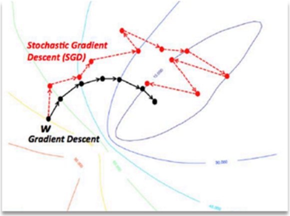
$$w=argmin_w L(w)$$




- 通常我们可以通过“梯度下降”的方法来求解w

$$w_{new}=w_{old}-\eta\frac{\partial L}{\partial w}$$

学习率　　损失函数对参数的梯度




这种算法是机器学习（也是深度学习）的基础

# 计算图

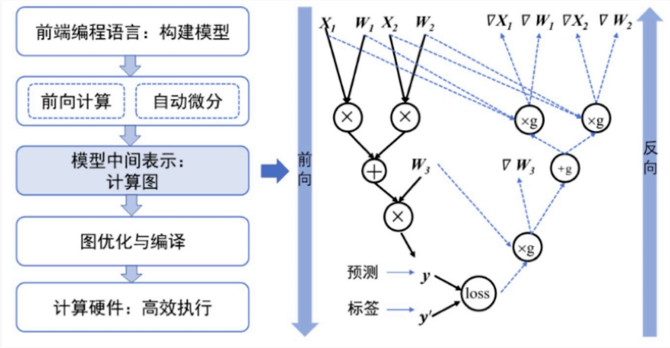
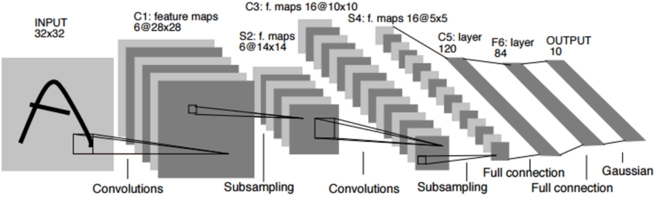



图4.1.1 基于计算图的架构

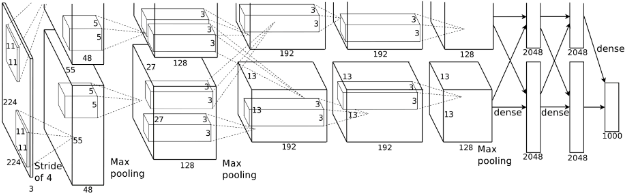
自动微分：将计算机程序中的运算操作分解为一个有限的基本操作集合，且集合中基本操作的求导规则均为已知，在完成每一个基本操作的求导后，使用链式法则将结果组合得到整体程序的求导结果。

# 卷积神经网络（CNN）

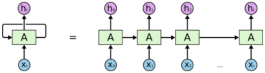
LeCun et al.
1989-1998
LeNet



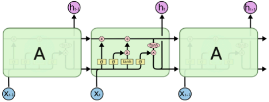

Krizhevksy,
Sutskever,
Hinton 2012
ImageNet



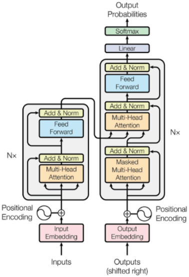
# BERT and GPT-3



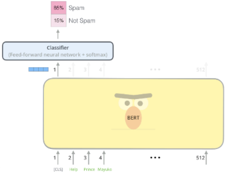
Vanilla RNN



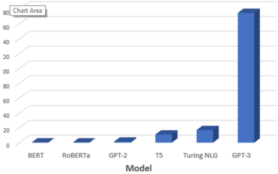
LSTM

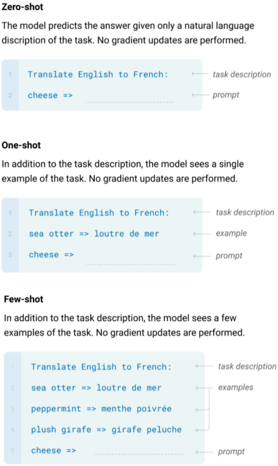



Transformer



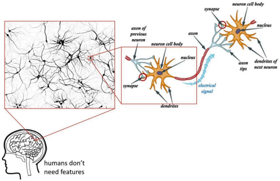

BERT

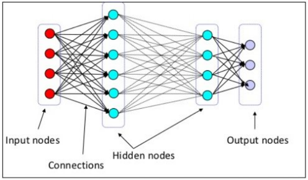


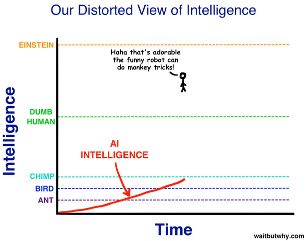



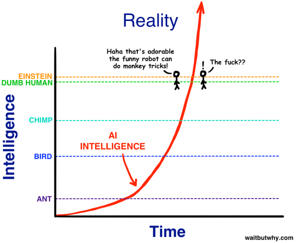

GPT-3

# How to fly? How to be intelligent?



# “New” Novum Organum (“新”新工具)

- 自然科学是怎样可能的？
  - 英国经验主义（如，休谟）：没有真理，只是经验的归纳整理，compact 的表示
  - 大陆理性主义（如，莱布尼茨）：类似几何原本，从几条简单公设出发形式化推理构建完整理论
    - 哥德尔不完备性定理：任何足够复杂的公理系统不能既可靠又完备；在可靠系统中，一定存在真命题没法被证明
    - 不同的公理系统（理论），彼此不能证伪
  - 康德、黑格尔：人通过先验概念理解世界，不是世界一定是这样，而是我们只能理解由公理系统所描述的世界
  - 培根《新工具》(“Novum Organum”)：公理系统 + 科学实验
- “新”新工具：神经网络，端到端学习
  - Universal approximation：任何连续函数都可以被前馈神经网络以无限小的误差近似
  - Turing completeness：任何图灵机都可以被循环神经网络进行模拟

# Table of Contents

# AI prehistory

| | |
|---|---|
| Philosophy | logic, methods of reasoning |
| | mind as physical system |
| | foundations of learning, language, rationality |
| Mathematics | formal representation and proof |
| | algorithms |
| | computation, (un)decidability, (in)tractability |
| | probability |
| Psychology | adaptation |
| | phenomena of perception and motor control |
| | experimental techniques (psychophysics, etc.) |
| Linguistics | knowledge representation |
| | grammar |
| Neuroscience | physical substrate for mental activity |
| Control theory | homeostatic systems, stability |
| | simple optimal agent designs |

# AI history

- The gestation of AI 孕育期 (—1956)
- Reasoning methods 注重推理时期 (1956-1975)
- Knowledge-based system 知识运用时期 (1976- 1988)
- Integration 集成运用 (1989- present)

# The Gestation of AI

- 古希腊 Aristotle (亚里士多德 BC 384-322)，给出形式逻辑的基本规律 Syllogism (三段论)
- 英国 Bacon (培根 1561-1626)，系统地给出 Induction (归纳法)
- 德国 Leibnitz (莱布尼茨 1646-1716) 提出 Symbolic Logic (符号逻辑)
- 英国 Boole (布尔 1815-1864) 提出 Boolean Algebra (布尔代数) 系统，实现了思维符号化和数学化
- 1936 英国 Turing (图灵, 1912-1954): 理想计算机模型 Turing Machine (图灵机)
- 1946 美国 Mauchly (莫克利), Eckert (埃克特): ENIAC
- 1948 美国 Shannon (香农): Information Theory (信息论)
- 1950 Turing Test 图灵测试

# The Birth of AI (1956)

- John McCarthy organized a two-month workshop at Dartmouth (达特矛斯会议) in the summer of 1956, ten young men were there:

  *McCarthy, Minsky, Rochester, Shannon, Moore, Samuel, Selfridge, Solomonff, Simon, Newell*

- They introduced all the major figures to each other and agreed to adopt the name of Artificial Intelligence for the field.

# Abridged history of AI

| | |
|---|---|
| 1943 | McCulloch & Pitts: Boolean circuit model of brain |
| 1950 | Turing's "Computing Machinery and Intelligence" |
| 1952–69 | Look, Ma, no hands! |
| 1950s | Early AI programs, including Samuel's checkers program, Newell & Simon's Logic Theorist, Gelernter's Geometry Engi |
| 1956 | Dartmouth meeting: "Artificial Intelligence" adopted |
| 1965 | Robinson's complete algorithm for logical reasoning |
| 1966–73 | AI discovers computational complexity<br>Neural network research almost disappears |
| 1969–79 | Early development of knowledge-based systems |
| 1980–88 | Expert systems industry booms |
| 1988–93 | Expert systems industry busts: "AI Winter" |
| 1985–95 | Neural networks return to popularity |
| 1987– | AI becomes a science |
| 1995– | The emergence of intelligent agents |
| 2001– | The availability of very large data sets |
| 2010– | Deep learning |

# AI Today

- Mostly about engineering domain-specific solutions rather than creating general theories
- We don't know how to do most of intelligent things, but the rest can be solved pretty well
- A set of "tools" for representing information and using them to solve specific tasks
  - Neural networks, hidden Markov models, Bayesian networks, heuristic search, logic, ...
- There's no magic in AI. It's all about representation, optimization, probability, and algorithms

# Table of Contents

# Well-known AI applications

- Expert systems (organic chemistry, medicine, geology, configuring computers)
- Speech recognition
- Handwriting recognition
- Game playing (chess, checkers, now Go)
- Robots (automated cars, ping pong player, Honda robot)
- Automated theorem proving
- Web search engines
- Natural language understanding (machine translation, Google)
- Logistics scheduling (military — people, cargo, vehicles)
- Cruise missiles
- Microsoft Answer Wizard

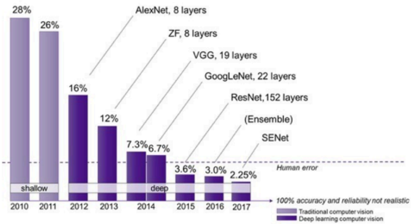
# The state of the art

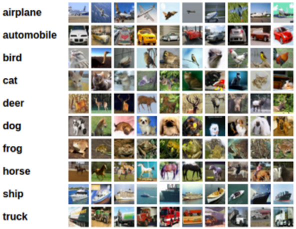
- Google language translation services
- Google automatic news aggregation and summarization
- Nuance voice recognition (behind Apple's Siri)
- Face detection and face recognition systems
- Apple Siri question-answering system
- IBM Watson question-answering system
- IBM Deep Blue chess playing program
- Deepmind AlphaGo
- Microsoft Photosynth
- Google Goggles
- Driverless cars
- OpenAI ChatGPT
- Robotics
- OpenAI Sora

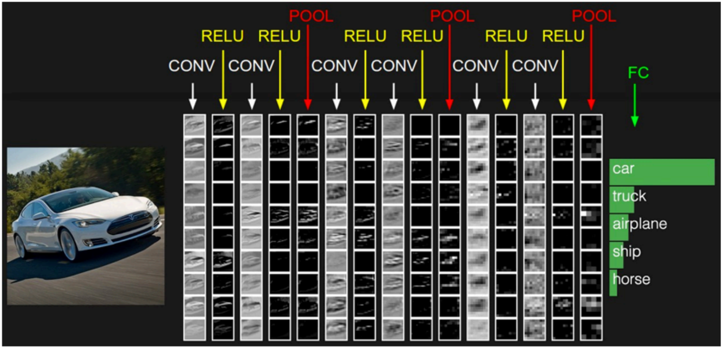
# Question Answering

Feb 2011, Watson（沃森）beat human on the quiz show Jeopardy!. And received the first prize of $1 million.

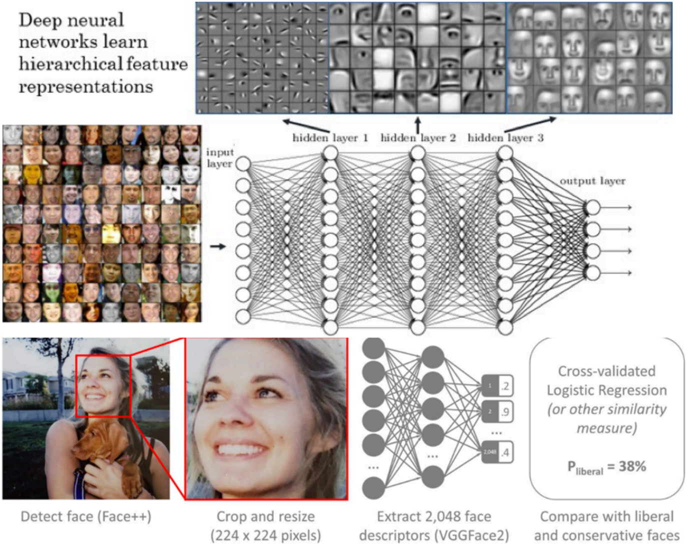
# Visual Recognition



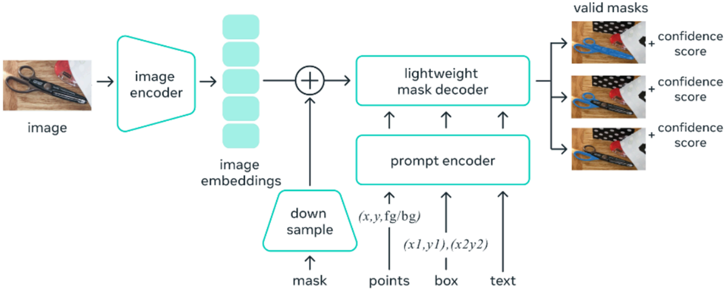



# Face Detection



# Meta AI: Segment Anything Model (SAM)

- SAM: an AI model that can "cut out" any object, in any image, with a single click




# OpenAI: CLIP

- CLIP: Connecting text and images

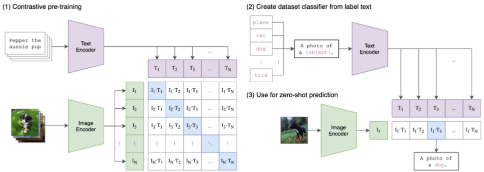
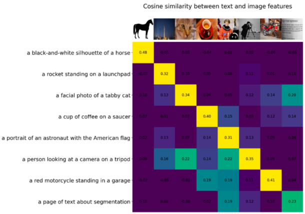
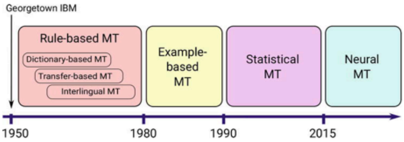
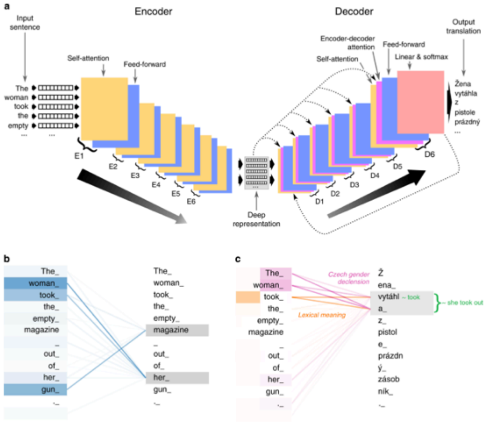



*Figure 1.* Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

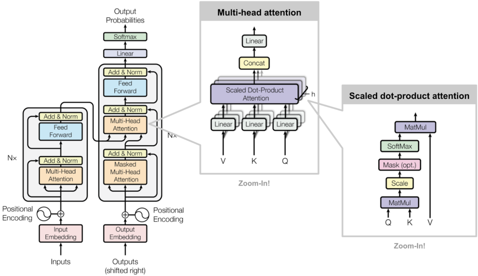


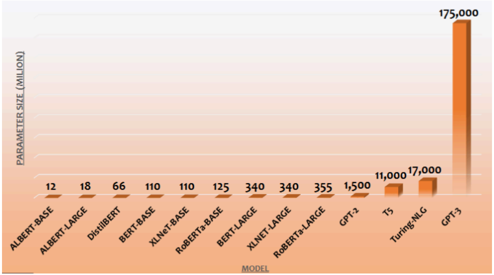
# Machine Translation

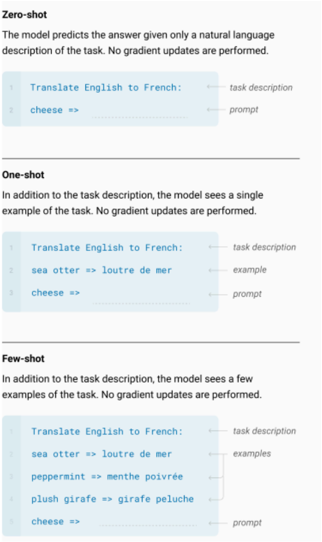


# Language Models

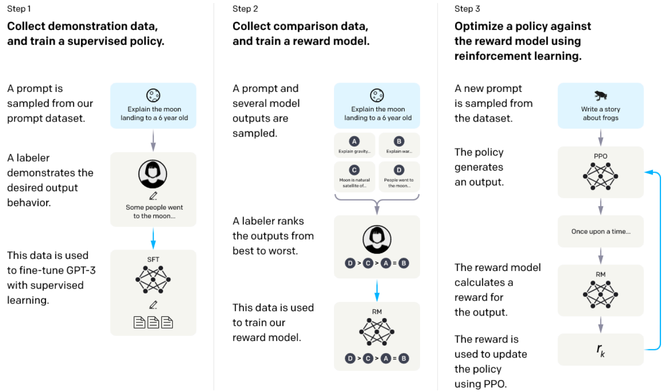


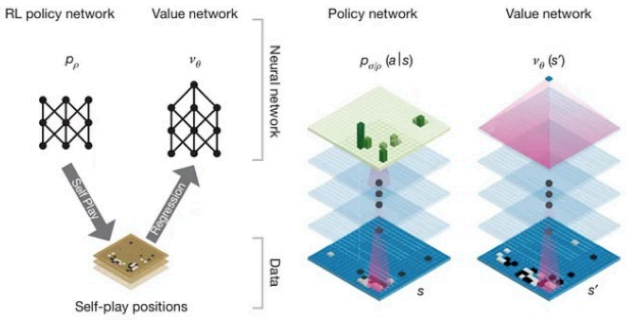
# OpenAI: ChatGPT

- ChatGPT: an artificial intelligence chatbot developed by OpenAI and launched in November 2022
  - GPT-3: Large language model with 175 billion parameters
  - ChatGPT was fine-tuned on top of GPT-3.5 using supervised learning and reinforcement learning

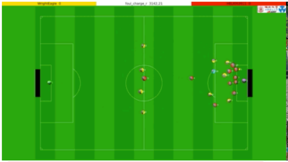


# DeepMind: AlphaGo

- AlphaGo 以 4： 1 的战绩击败李世石，机器第一次在围棋领域战胜人类顶尖高手
- AlphaGo Zero 从 0 学起，在不到 3 天的时间内以 100:0 完虐 AlphaGo



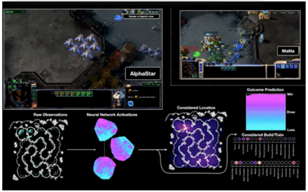
# Games



**RoboCup 2D 仿真平台（算法测试平台）**
多智能体协作、对抗研究的标准平台

- 网易 “易水寒”

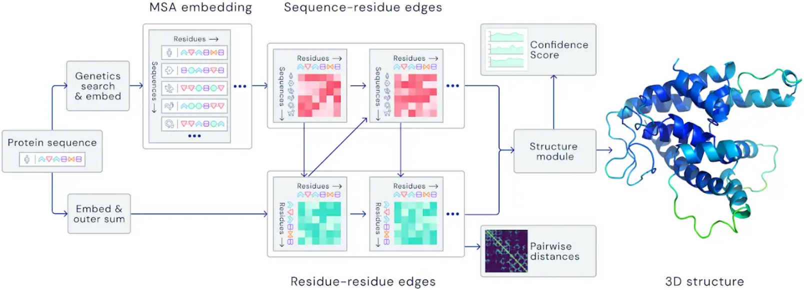
- 网易 “潮人篮球”

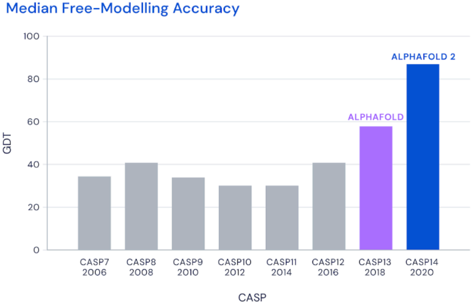



- DeepMind 星际争霸2

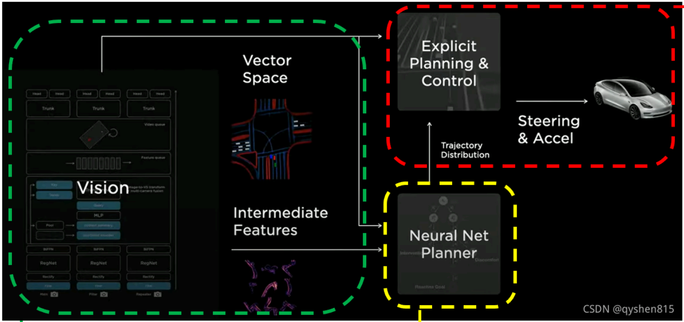

- OpenAI Dota 2

# DeepMind: AlphaFold 2



# Tesla: Self-driving cars

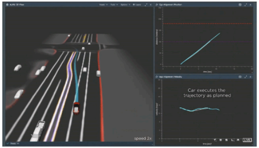



**经典规划和控制**

**Mid-to-Mid**
架构适用于各类无人系统，将感知和规划模块分别简化一张神经网络，可以**数据驱动的方式持续训练更新**

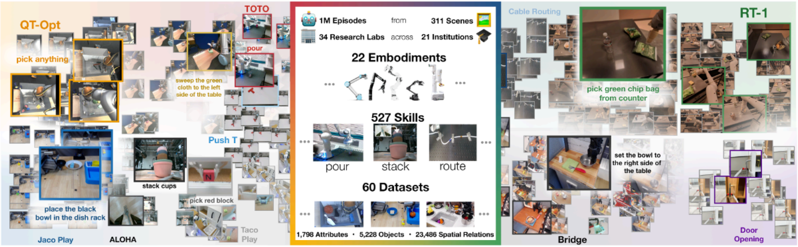
**感知一张网**：输入传感器信息，训练网络输出BEV下的数字化世界模型

**规划一张网**：输入数字化世界模型，训练网络输出可行驶轨迹分布



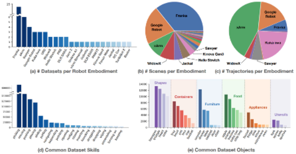
# Open X-Embodiment：机器人界的 Image-Net 时刻

**Open X-Embodiment: Robotic Learning Datasets and RT-X Models**




提供了轨迹数据集和预训练模型：

- **Open X-Embodiment Dataset**: robot learning dataset with *1M+ robot trajectories* from 22 *robot embodiments*.
- **Pre-Trained Checkpoints**: a selection of RT-X model checkpoints ready for inference and finetuning.




数据集：

- 将不同已有数据集统一化
- 输入：规范相机视角，图片形式
- 输出：归一化的7维动作，能适应不同的动作形式：[绝对值，相对值，速度]

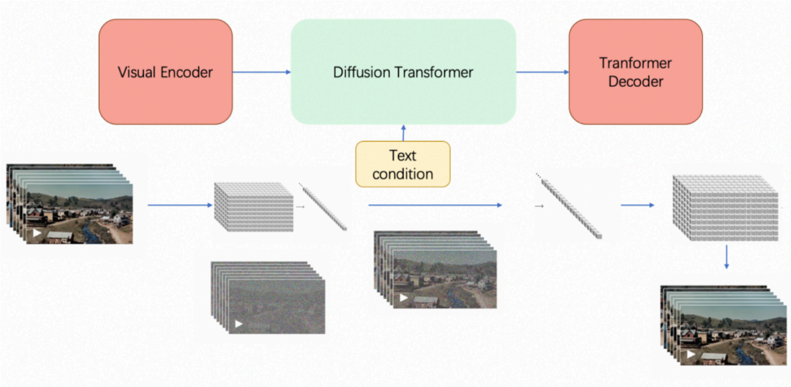
模型：

- RT-1-X 开源
- 输入：RGB图像 + 任务文本描述
- 输出：不同形式的7维动作

通过庞大的开源数据集，训练具身智能大模型，创造一个泛化能力极强的机器人

# OpenAI: Sora

- Sora: an AI model that can create realistic and imaginative scenes from text instructions.



# Summary

- Applications of AI:
  - high-impact (affect billions of people)
  - diverse (language, vision, robotics)
- Challenges: really hard...
  - computation complexity
  - information complexity
- Paradigm: modeling + algorithms