



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

CS3317: Artificial Intelligence

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

<https://taohuang.info/cs3317>

<https://oc.sjtu.edu.cn/courses/89538>

AI Today: Chat Assistant

What does SJTU stand for?



Thinking Process 1.97s >

What Does SJTU Stand For?

SJTU stands for **Shanghai Jiao Tong University** (上海交通大学).

Key Facts About Shanghai Jiao Tong University

- **Location:** Shanghai, China
- **Type:** Public research university
- **Founded:** 1896
- **Reputation:** One of China's oldest and most prestigious universities, often ranked among the top universities in China and Asia

Notable Features

1. **Engineering Focus:** The university is particularly renowned for its engineering and technical programs
2. **History:** It was originally named "Nanyang Public School" (南洋公学) when founded in 1896
3. **Campuses:** Main campus is located in the Xuhui District of Shanghai
4. **Ranking:** Consistently ranks in the top 3-5 universities in China and among the top 100-200 worldwide

SJTU is also known for its strong programs in computer science, electrical engineering, mechanical engineering, business administration, and medicine.

Write a C++ code of quick sort.



Received. I am now generating the C++ code for quick sort.

Thinking Process 15.30s >

I'll write a clean and efficient C++ quick sort implementation for you.

Quick Sort Implementation in C++

Here's a complete, well-commented implementation of the quick sort algorithm:

```
cpp
#include <iostream>
#include <vector>
#include <cstdlib>
#include <ctime>

using namespace std;

// Function to swap two elements
void swap(int& a, int& b) {
    int temp = a;
    a = b;
```

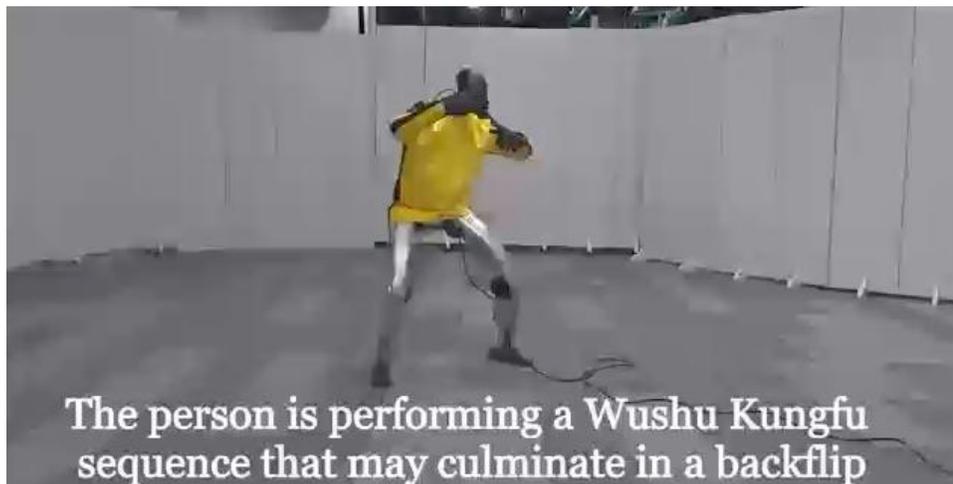


MiniMax Large Language Model

AI Today: Video Generation



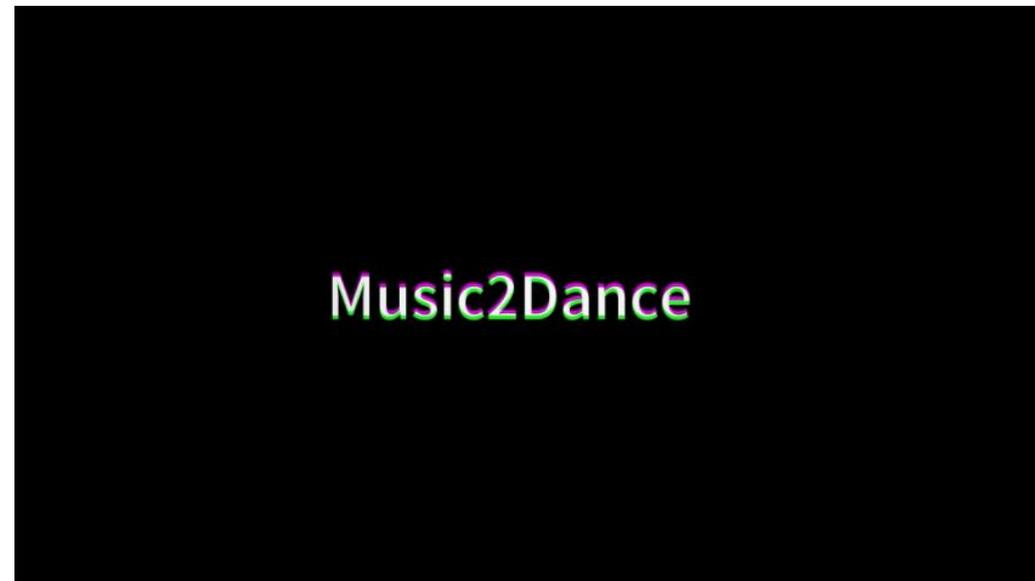
AI Today: Embodied AI



Text-to-locomotion



Video-to-locomotion



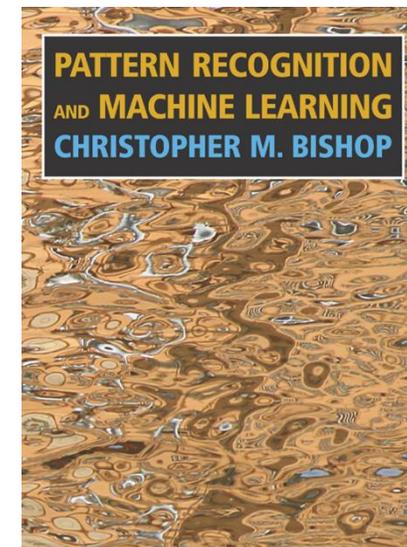
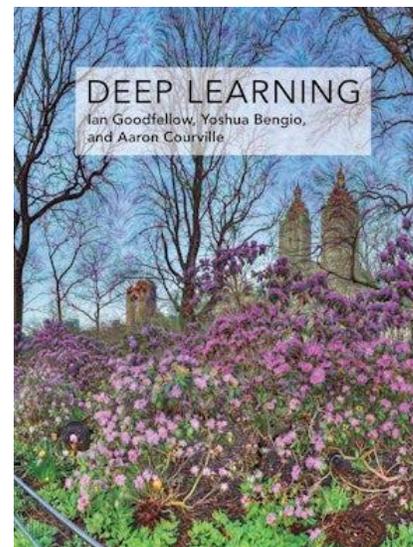
Music-to-locomotion

Self Introduction

- Position
 - Tenure-Track Assistant Professor 2025-now
- Education
 - PhD in Computer Science, The University of Sydney
 - Bachelor in Computer Science, Huazhong University of Science and Technology
- Research Interests
 - Large Foundation Models
 - Image / Video Generation
 - Embodied AI
- **Office Hour:** Mon. 4-7pm at 电院3号楼301b, or t.huang@sjtu.edu.cn

References

- Ian Goodfellow, Yoshua Bengio & Aaron Courville, Deep Learning (MIT Press).
<https://www.deeplearningbook.org>
- 周志华 《机器学习》 清华大学出版社，2016.
- Christopher Bishop, Pattern Recognition and Machine Learning.
<https://www.microsoft.com/en-us/research/wp-content/uploads/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>
- CS221: Artificial Intelligence: Principles and Technique. <https://stanford-cs221.github.io/>



Goal

- Understand what is AI and core AI principles
- Master foundational ML and deep learning methods
- Gain insight into modern AI paradigms: generative models, RL, ...
- Develop the ability to analyze, design, and evaluate AI systems

Grading

- No exam 没有笔试
- Attendance and participation: 10%
- Homework (written & programming): 40%
- Project: 40%
- Presentation: 10%

Course Outline

1. Introduction to AI
2. Search Algorithms
3. Machine Learning Foundations
4. Deep Learning
5. Generative AI
6. Reinforcement Learning
7. Frontier AI



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

Lecture 1: Introduction to Artificial Intelligence

Tao Huang

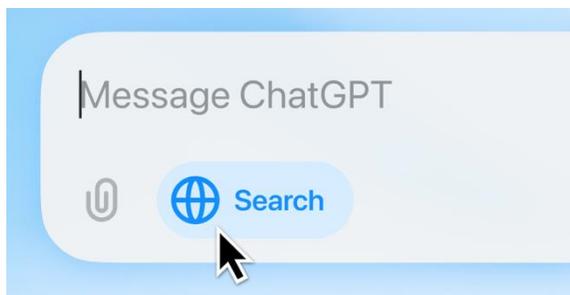
John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

<https://taohuang.info/cs3317>

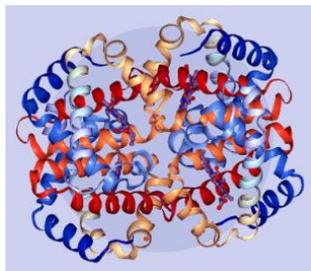
1. What is AI?

Why Are We Here?

- AI is transforming science, industry, and society
- From ChatGPT to autonomous driving

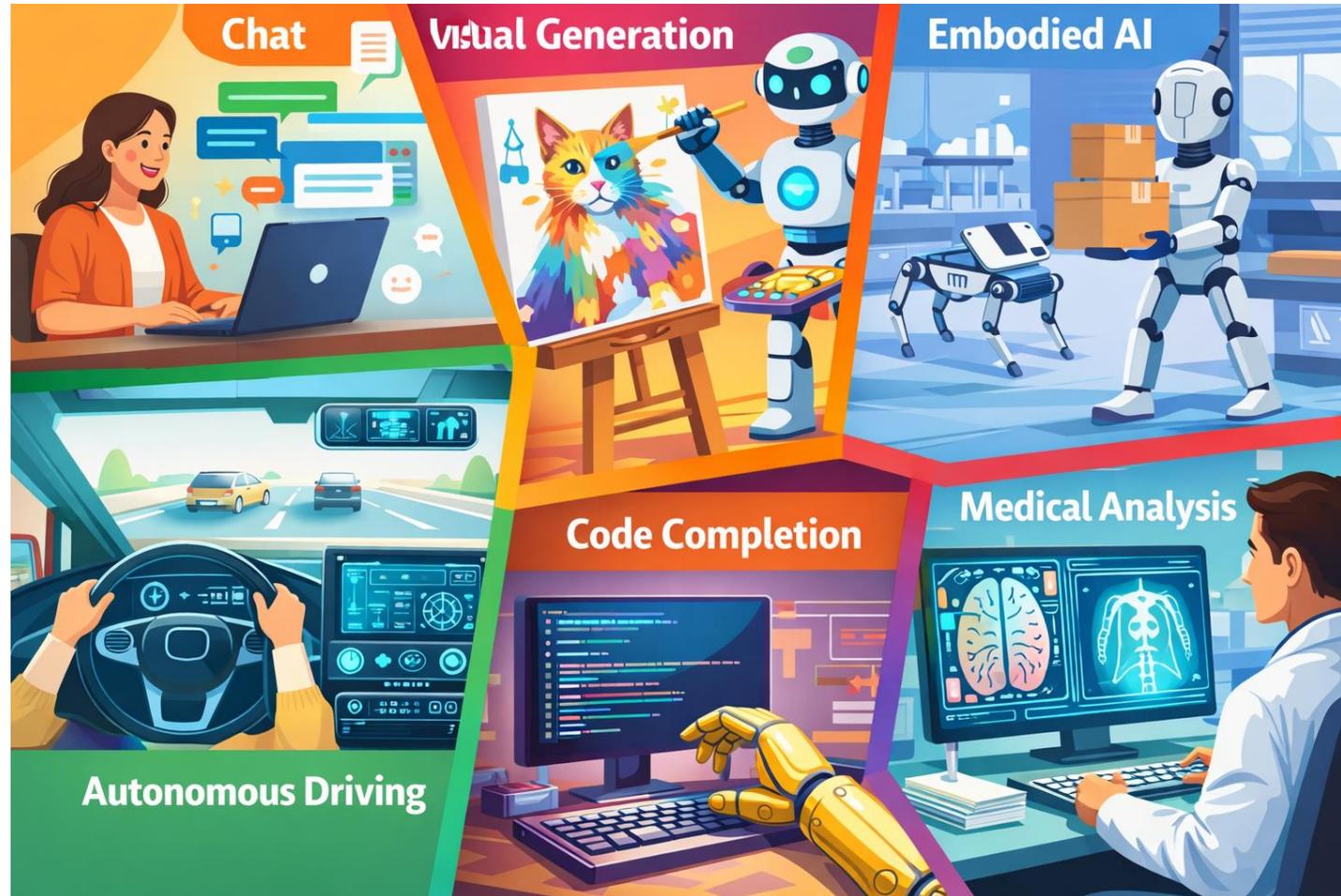


- From protein folding to robotics



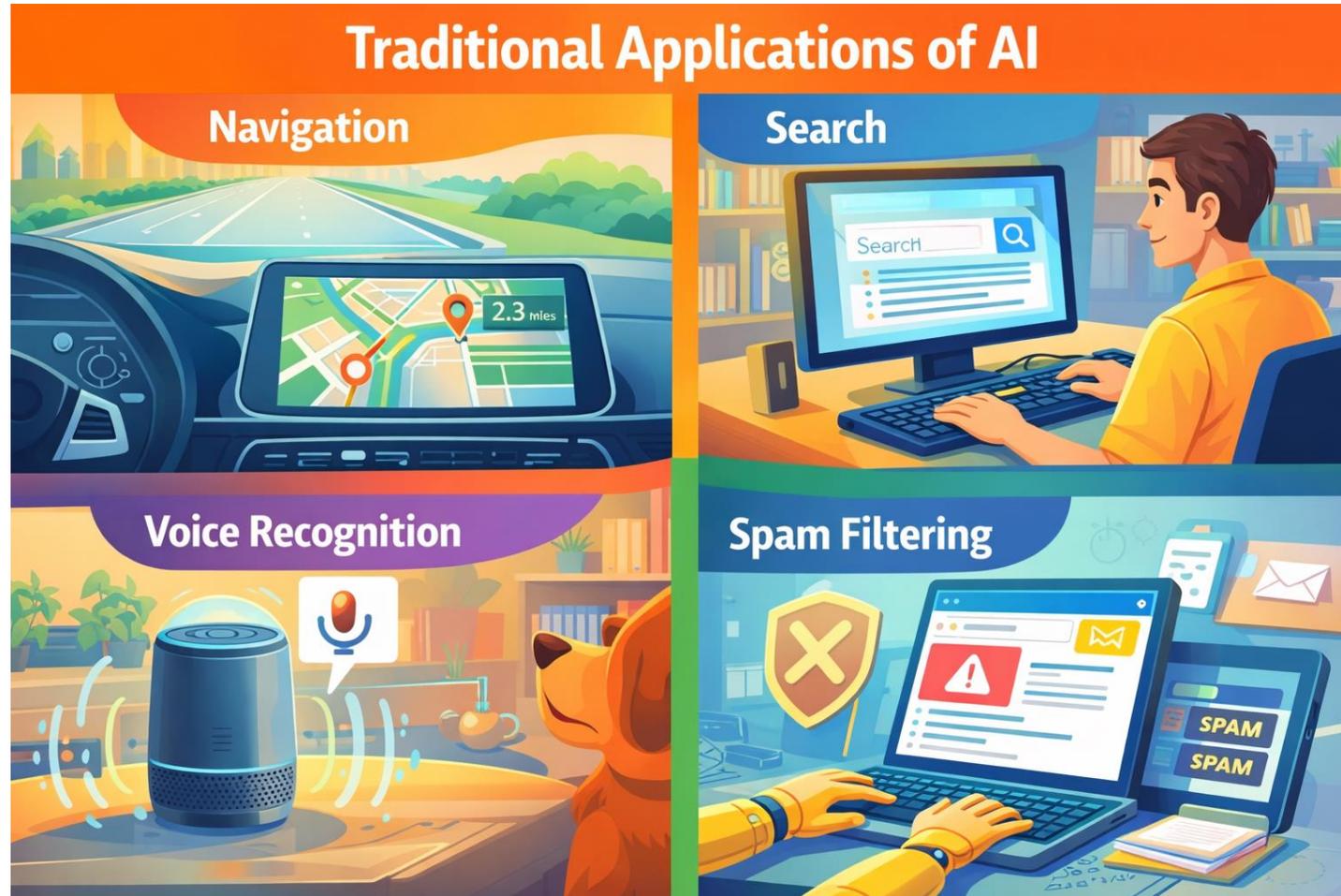
- But... what exactly is AI?

What is AI?



AI can do a lot of things today

These are also AI



Can Machines Be Intelligent?

- What is intelligence?
 - Is it Learning?
 - Reasoning?
 - Creativity?
 - Decision-making?

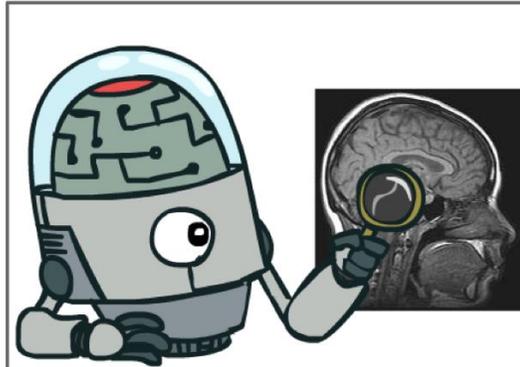
- **Intelligence = ability to perceive, learn, reason, adapt.**

What is Artificial Intelligence?

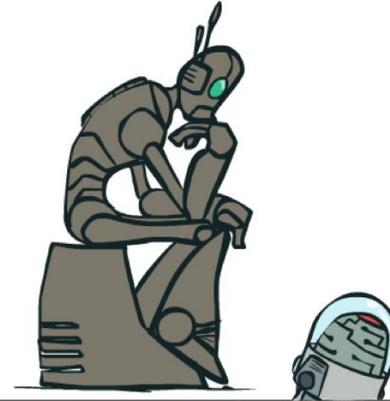
- Some definitions that have been proposed:

AI is the science of making machines that

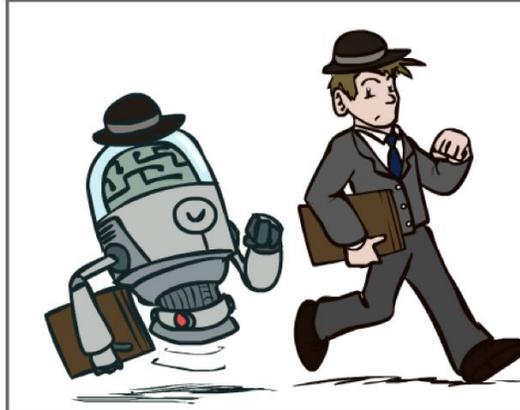
A. Think like humans



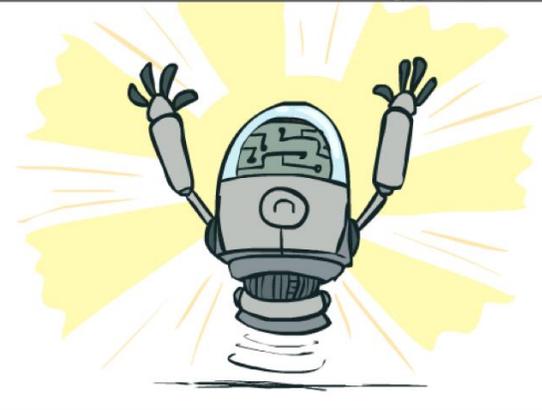
C. Think rationally



B. Act like humans

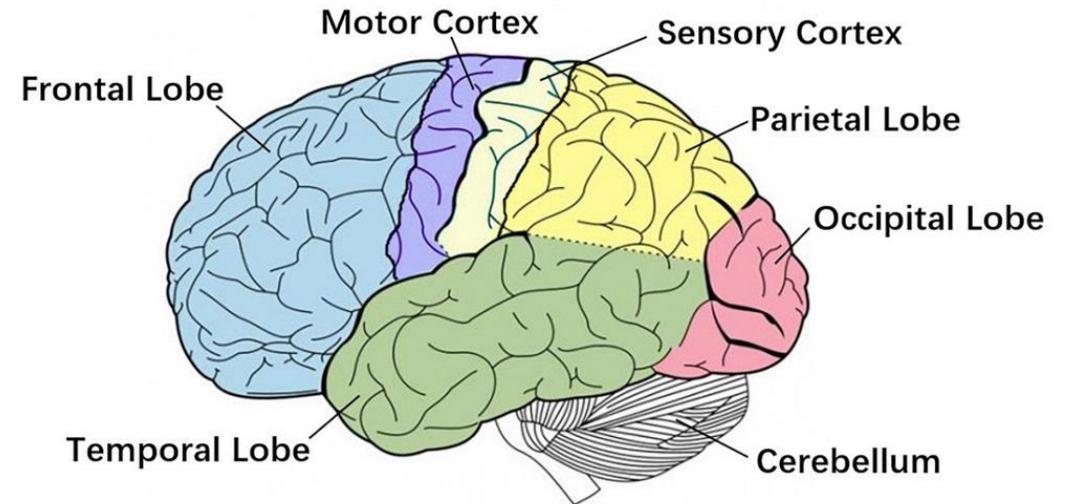


D. Act rationally



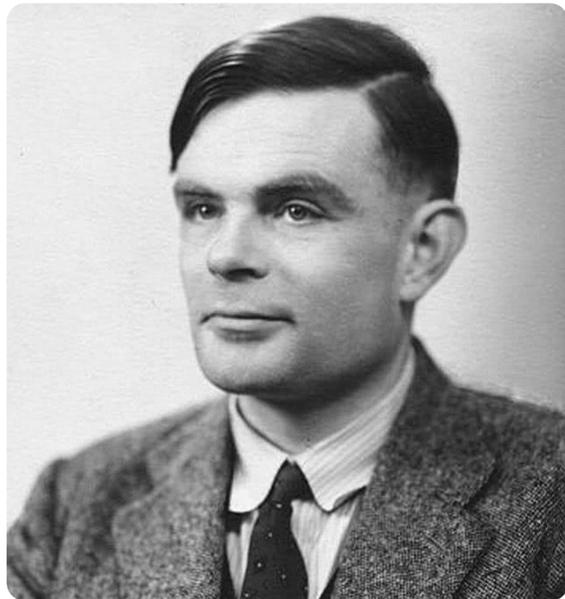
Thinking Humanly: Cognitive Science

- Requires knowledge of how humans think
- Study:
 - Psychological: observation of human behavior
 - Neurological: observation of brain activity
- Today: separate from mainstream AI

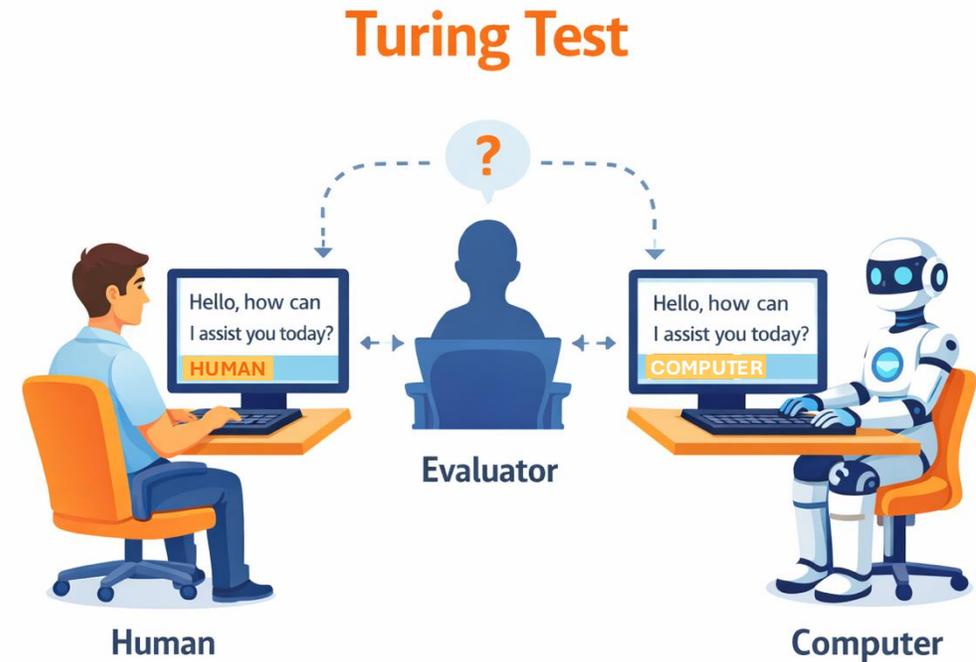


Acting Humanly: The Turing Test Approach

- In 1950s, Turing defined Turing Test to check if a machine's intelligent behavior is indistinguishable from a human's.



Alan Turing (1912-1954)

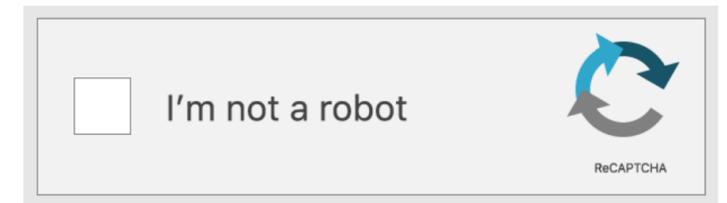
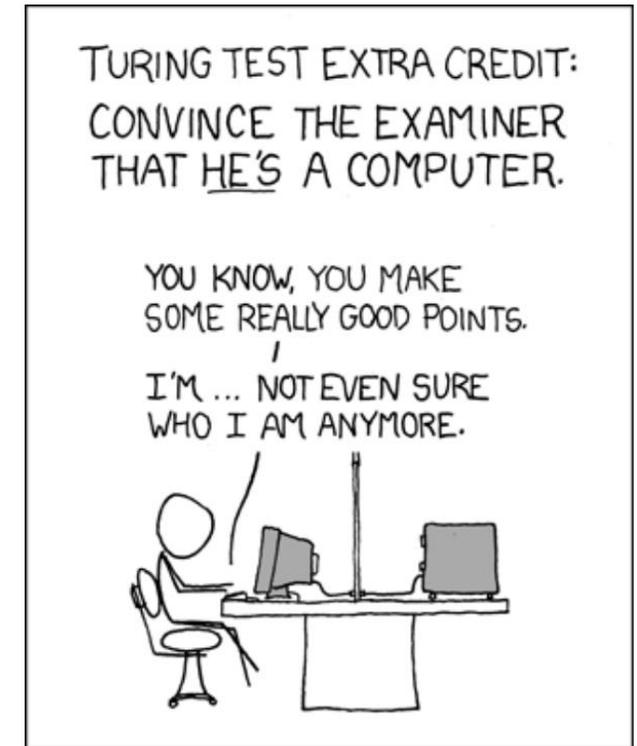


If a judge cannot distinguish machine from human → machine passes.

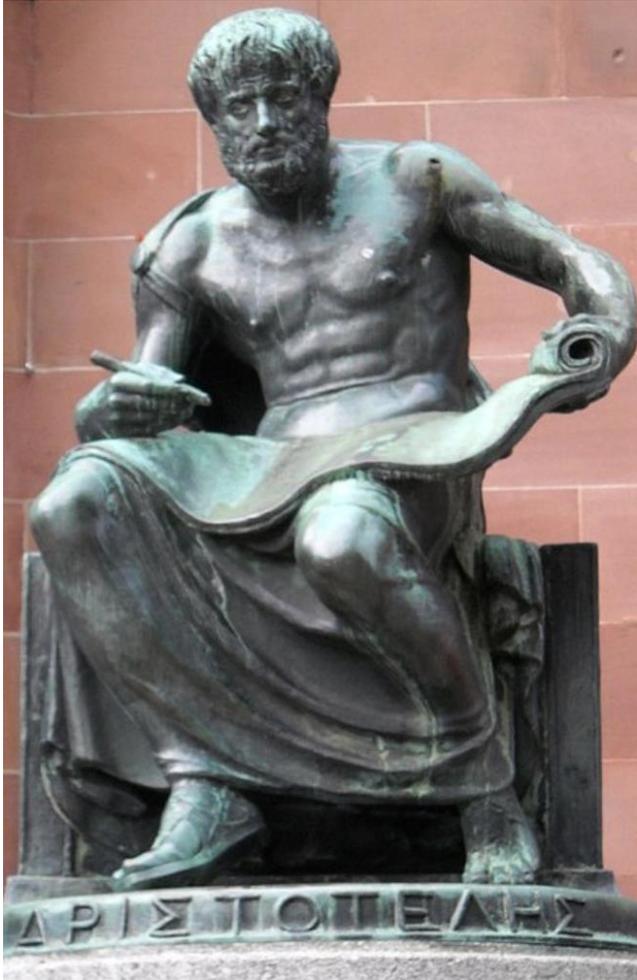
Limits of the Turing Test

- Human actions **not always intelligent**
- Mimicking humans \neq intelligence
- Hard to reproduce

=> Not important in AI “mainstream”

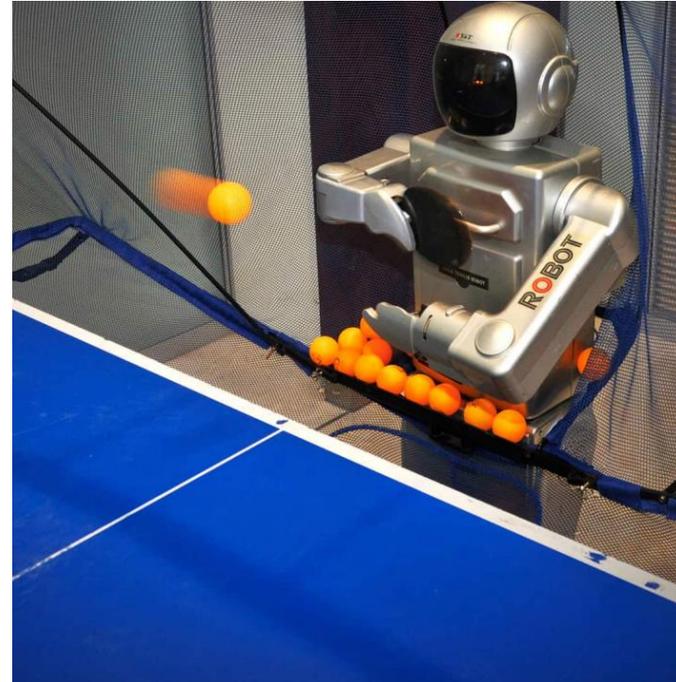
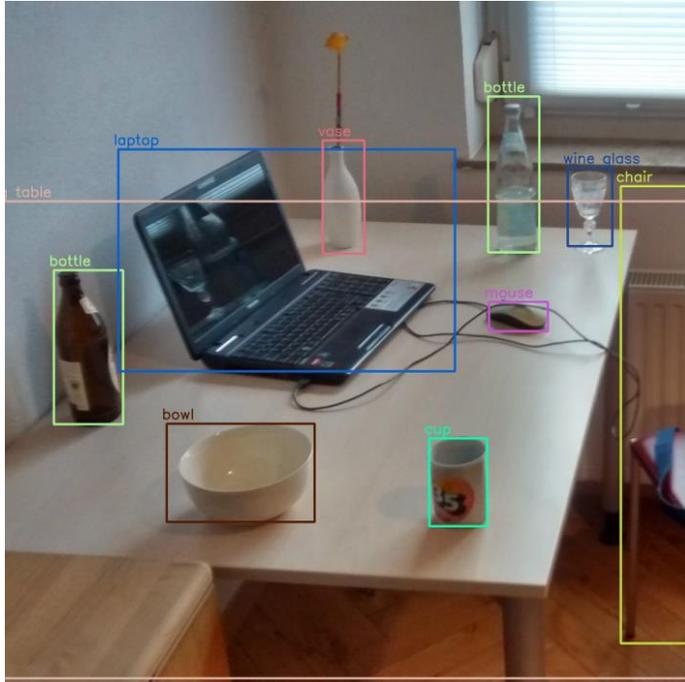


Thinking Rationally: Laws of Thought



- Thinking rationally: find out how correct thinking must proceed: **the laws of thought.**
- Aristotle syllogism (三段论):
 - Socrates is a human.
 - All humans are mortal.
 - Therefore Socrates is mortal.
- The logicians hope to build on logic systems to create intelligent systems.

Thinking Rationally: Laws of Thought



Problem: it is not always possible to model thought as a set of rules; sometimes there are imprecisions.

Acting Rationally: Doing the Right Thing

- The right thing: act as to achieve the best outcome
 - Choose actions
 - Maximize expected utility
 - Focus on outcomes



This is how birds fly



Humans tried to mimic birds for centuries

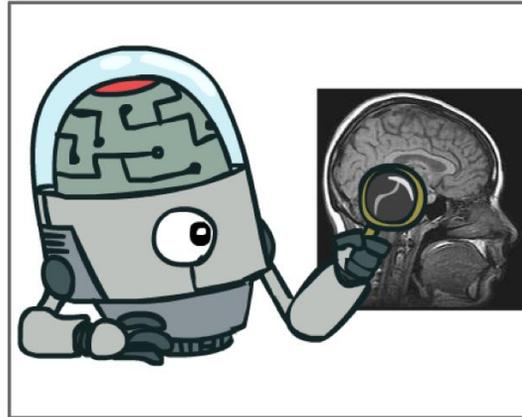


This is how humans finally achieved “artificial flight”

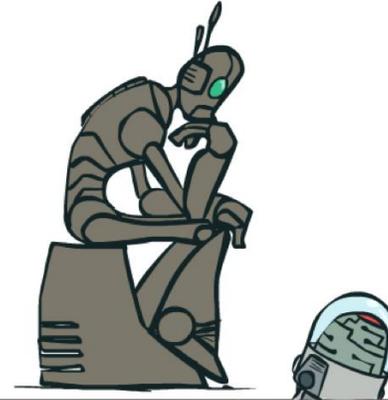
What is Artificial Intelligence?

AI is the science of making machines that

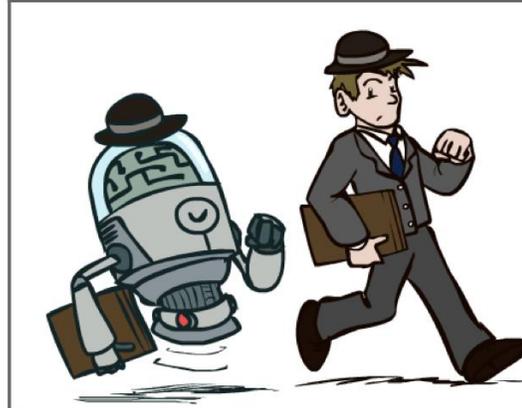
A. Think like humans



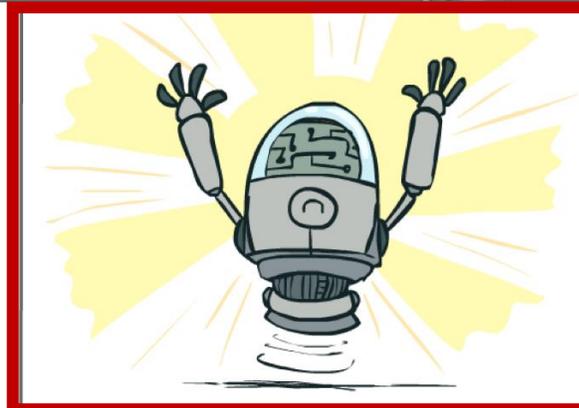
C. Think rationally



B. Act like humans



D. Act rationally
(the most common view today)

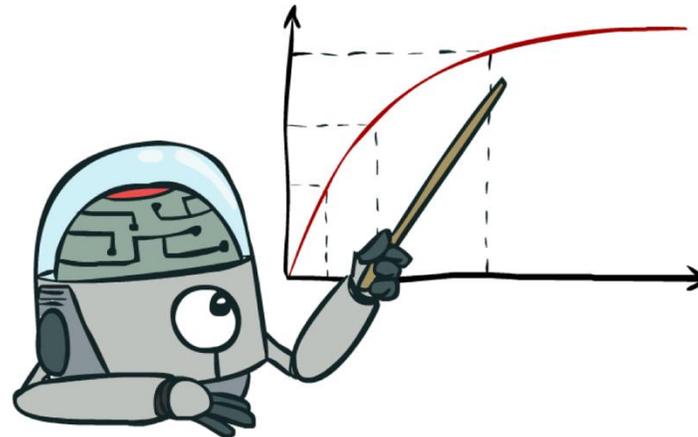


What is Artificial Intelligence?

- **Rational Agent Perspective:**

An AI system is:

An agent that perceives its environment and takes actions to maximize expected utility.

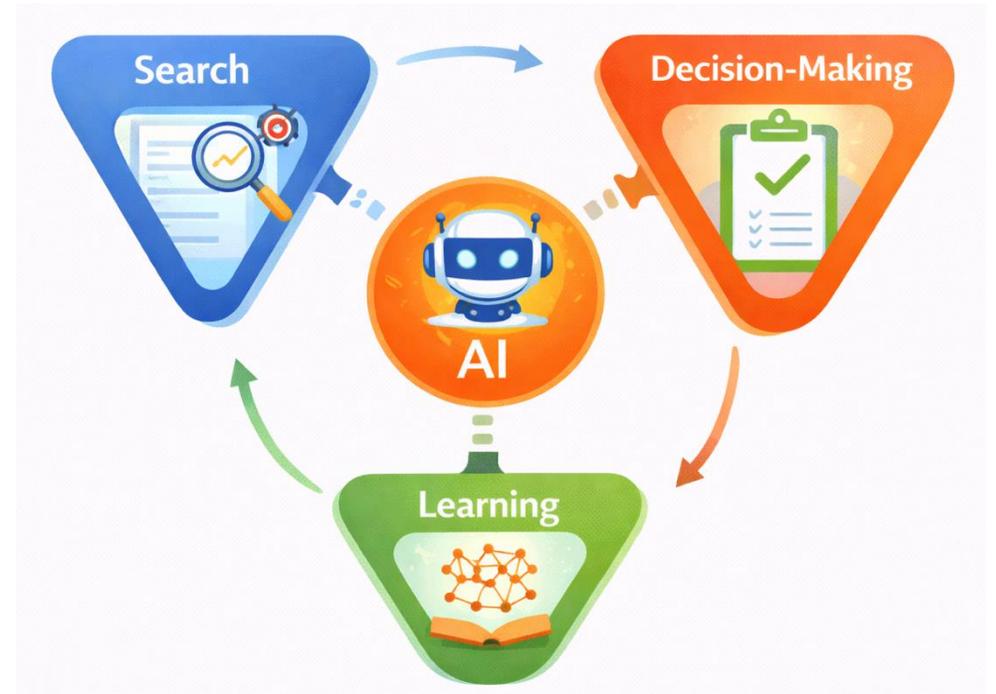


The AI Triad

Most AI systems combine:

- **Search** → exploring possibilities
- **Learning** → improving from data
- **Decision-making** → choosing actions

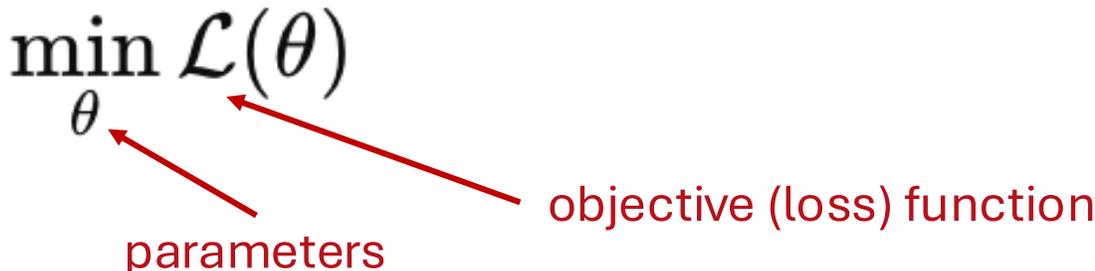
This will structure our entire course.



AI as Optimization

Most modern AI:

- Define objective $\mathcal{L}(\theta)$
- Optimize parameters θ
- Generalize to unseen data

$$\min_{\theta} \mathcal{L}(\theta)$$


The diagram shows the equation $\min_{\theta} \mathcal{L}(\theta)$ with two red arrows. One arrow points from the text 'parameters' to the variable θ in the denominator of the min operator. The other arrow points from the text 'objective (loss) function' to the function $\mathcal{L}(\theta)$.

Course Roadmap Preview

1. Search Algorithms
2. Machine Learning Foundations
3. Deep Learning
4. Generative AI
5. Reinforcement Learning
6. Frontier AI

Course Roadmap Preview

Search Algorithms

Breadth-First Search

Depth-First Search

A* Search

...

Predefined rules

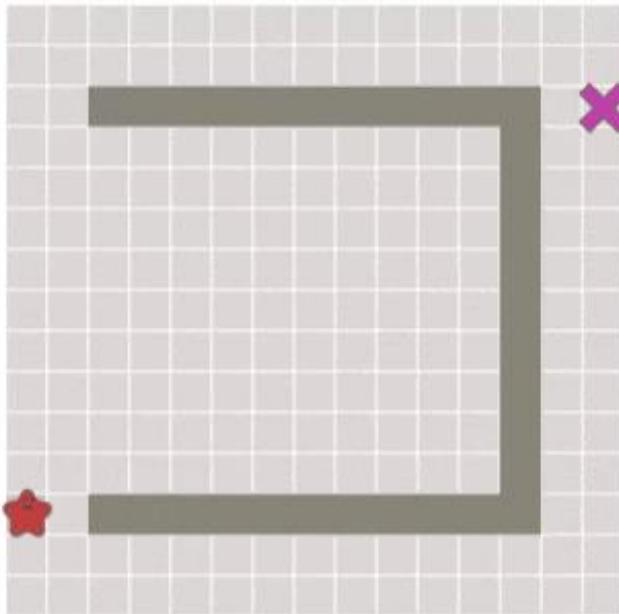


Intelligence from Search

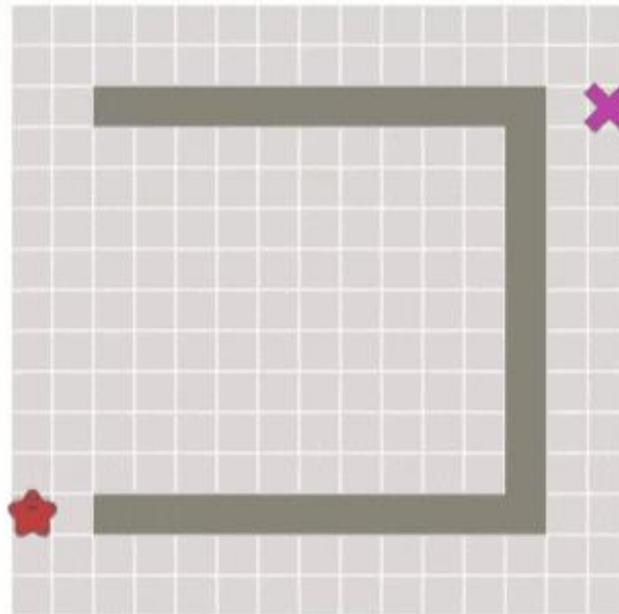
- Search Algorithm

How can we solve problems if the rules are known?

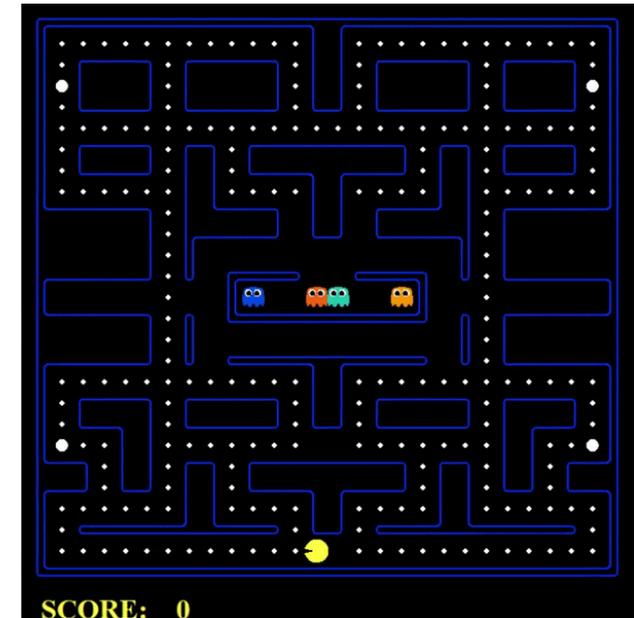
Core idea: solve problems by systematic exploration.



Breadth-First Search



Greedy Search



Pac-Man Game

Course Roadmap Preview

Search Algorithms

Breadth-First Search

Depth-First Search

A* Search

...

Predefined rules



Statistical learning

Machine Learning

Linear Regression

Logistic Regression

Support Vector Machines

...

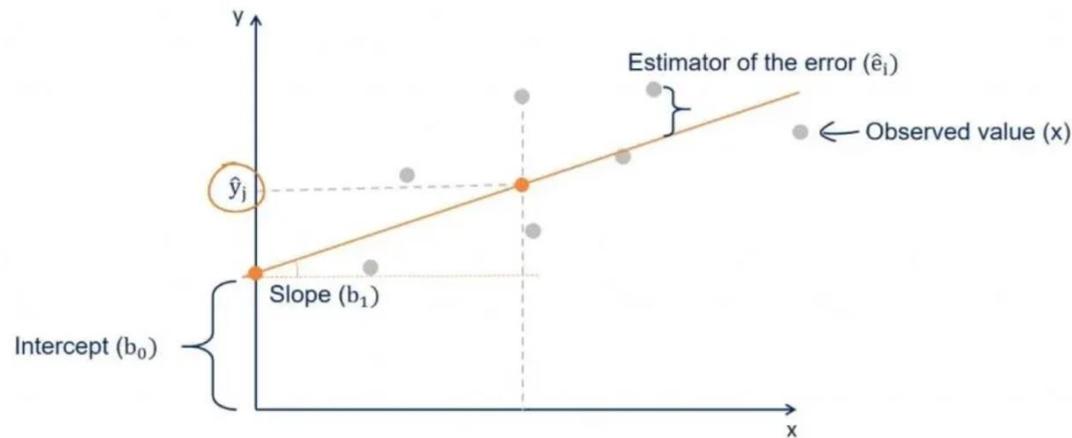
Intelligence from Data

- Machine Learning

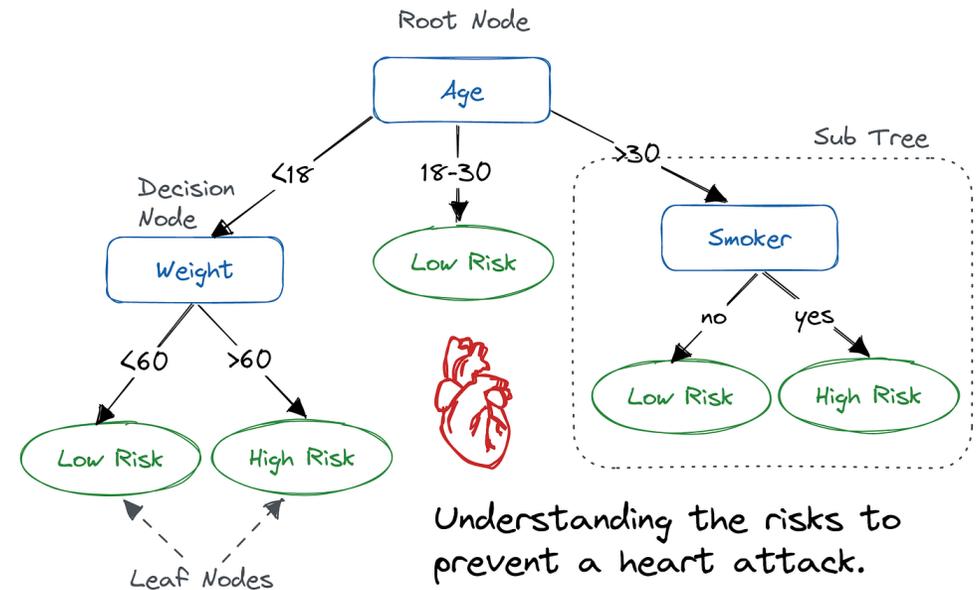
What if we don't know the rules?

Core Idea: Learn decision functions from data.

$$\hat{y}_i = b_0 + b_1 x_i$$



Linear Regression



Understanding the risks to prevent a heart attack.

Decision Trees

Course Roadmap Preview

Search Algorithms

Breadth-First Search

Depth-First Search

A* Search

...

Predefined rules

Deep Learning

Multi-Layer Perceptrons

Backpropagation

Transformers

...

Representation

Narrow

General

Statistical learning

Machine Learning

Linear Regression

Logistic Regression

Support Vector Machines

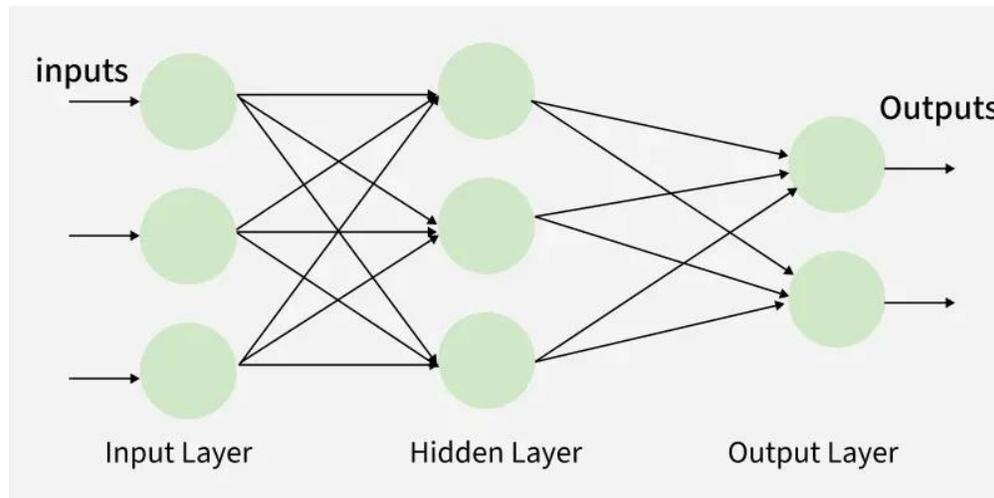
...

Intelligence from Representation

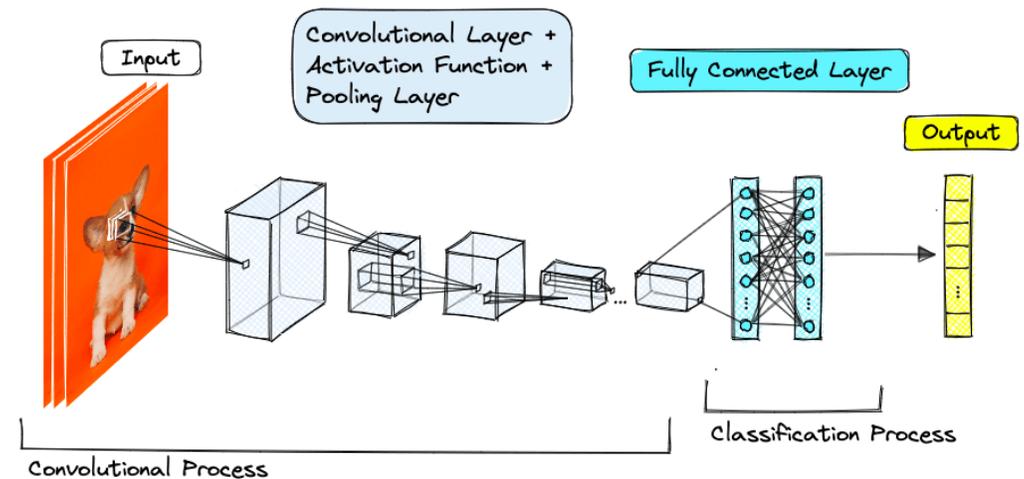
- Deep Learning

How do we automatically learn features?

Core Idea: Learn hierarchical representation in multi-layer neural networks.



Multi-Layer Perception (MLP)



Convolutional Neural Network (CNN)

Course Roadmap Preview

Search Algorithms

Breadth-First Search

Depth-First Search

A* Search

...

Predefined rules

Deep Learning

Multi-Layer Perceptrons

Backpropagation

Transformers

...

Representation

Narrow

General

Statistical learning

Machine Learning

Linear Regression

Logistic Regression

Support Vector Machines

...

Modeling distributions

Generative AI

Autoregressive Models

VAE, GAN

Diffusion Models

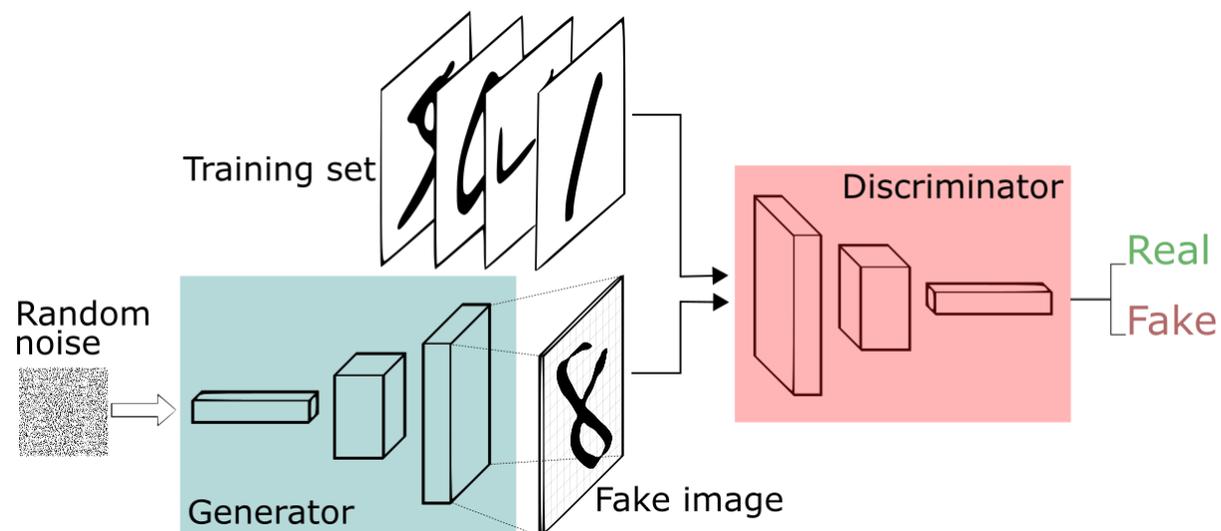
...

Intelligence as World Modeling

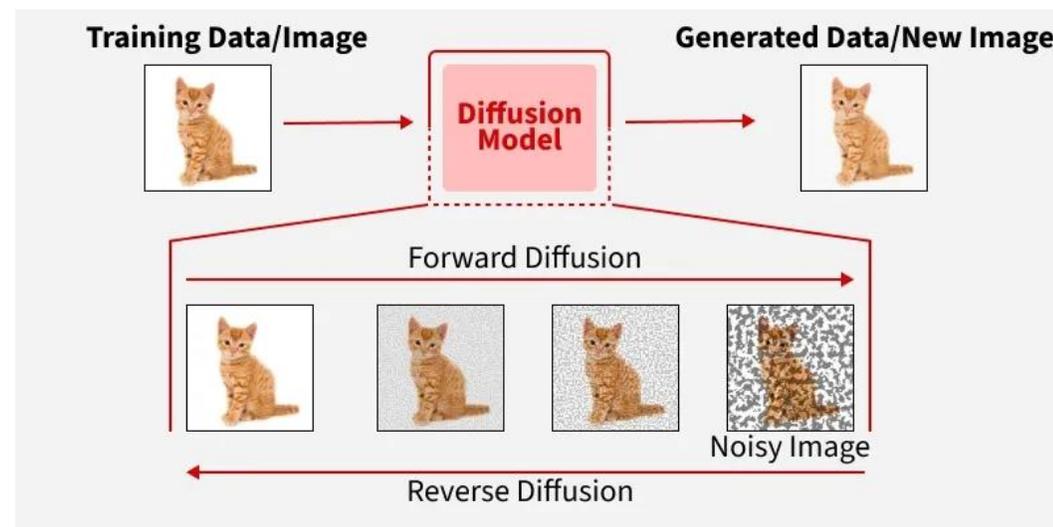
- Generative AI

Can we model the world and create new content?

Core Idea: Model data distributions and generate content.

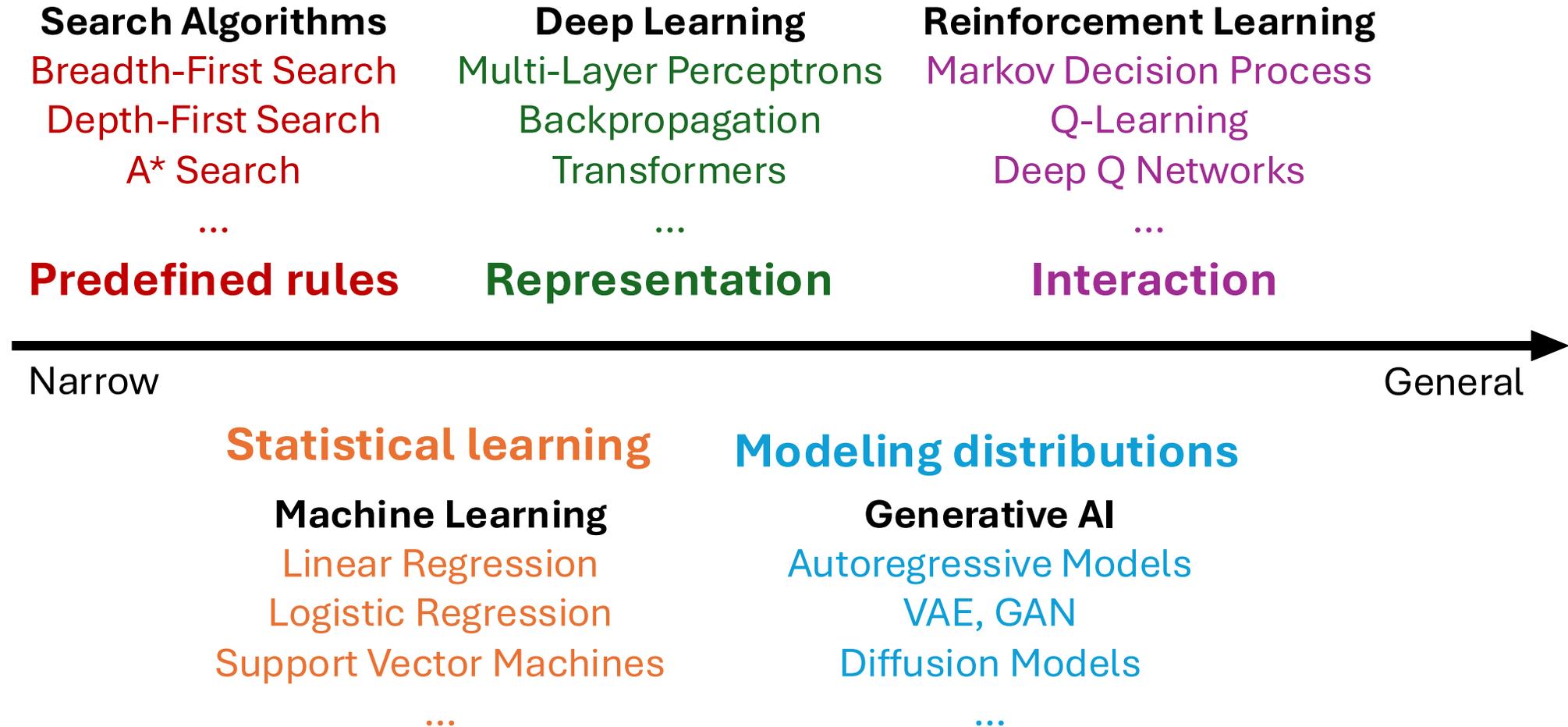


Generative Adversarial Network (GAN)



Diffusion Models

Course Roadmap Preview

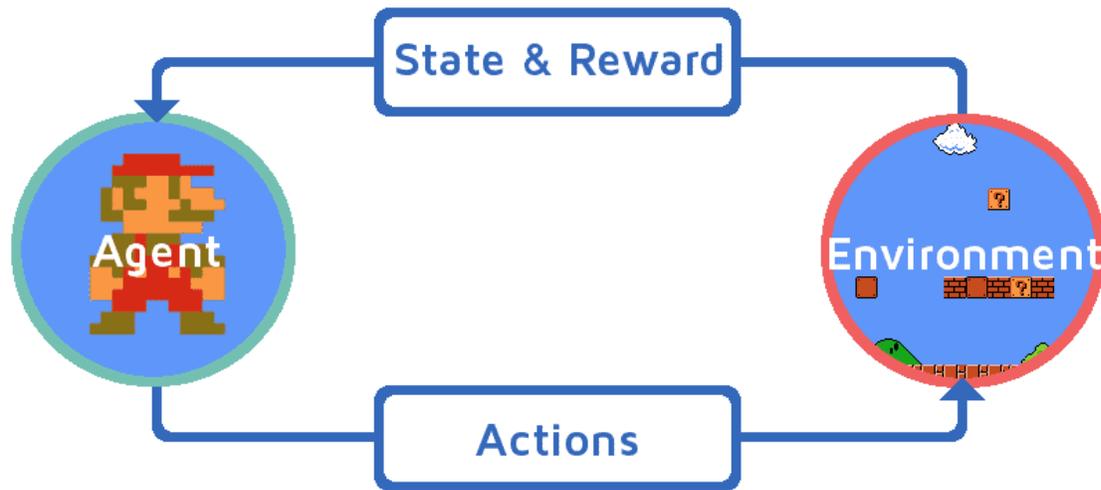


Intelligence through Interaction

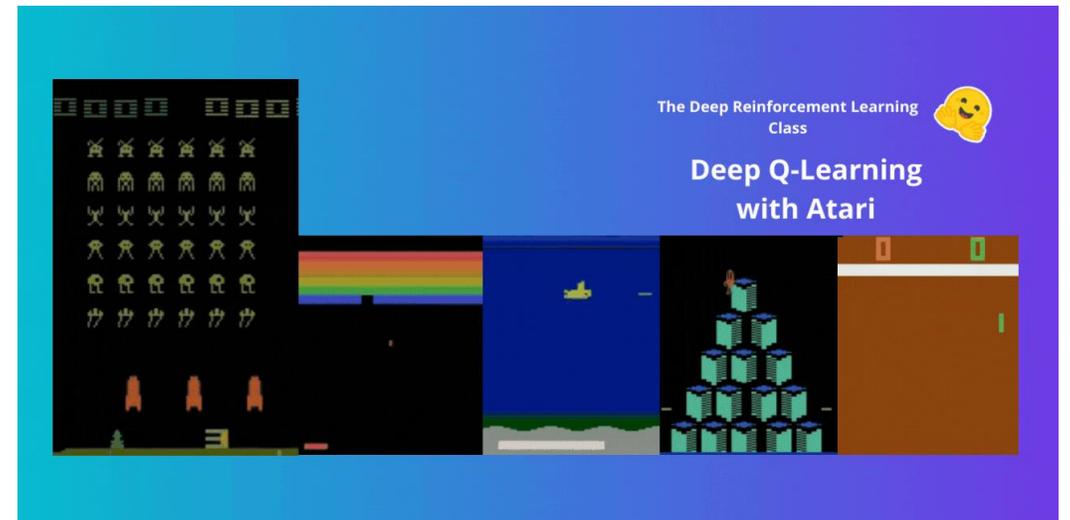
- Reinforcement Learning

How do agents act and learn from feedback?

Core Idea: Learn to act by maximizing reward.

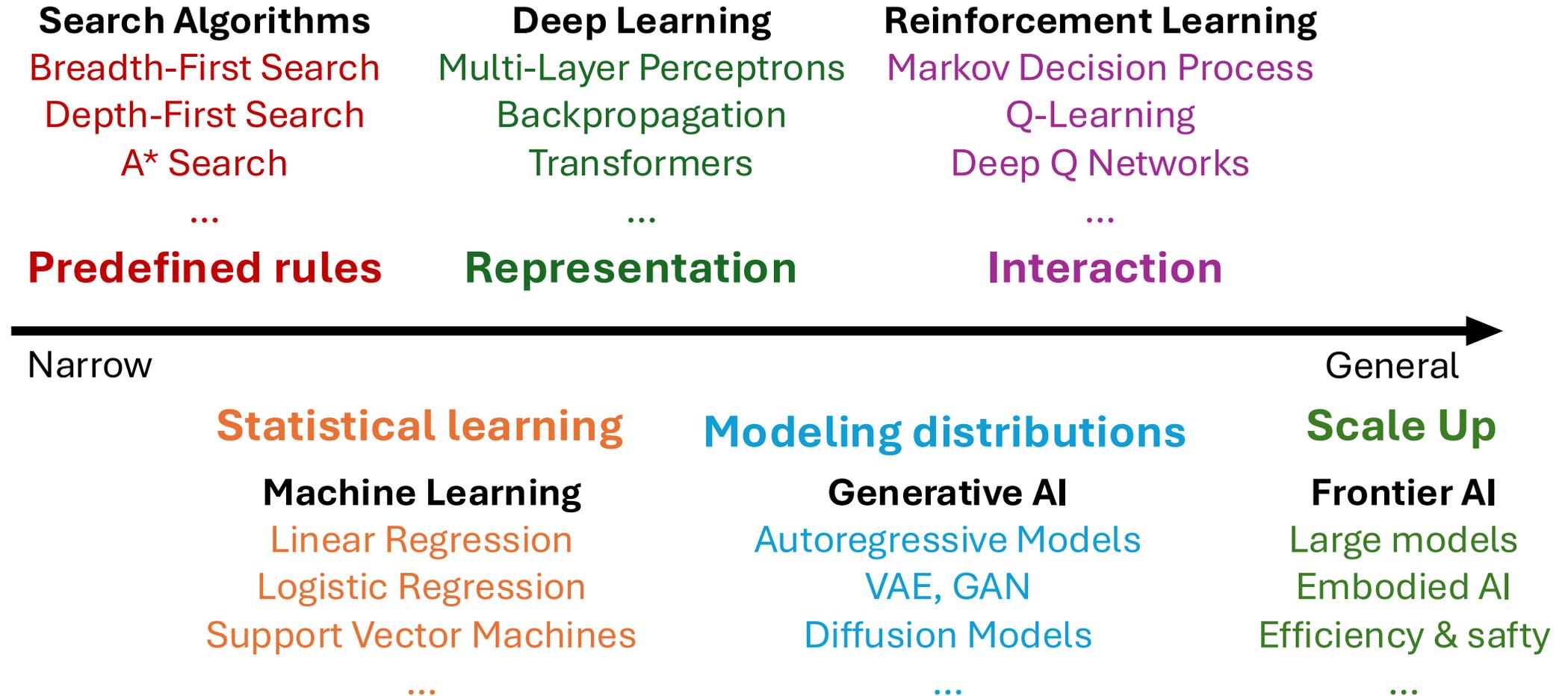


Reinforcement Learning Process



Deep Q-Learning with Atari

Course Roadmap Preview



Intelligence at Scale

- Frontier AI

How do we build general AI systems?

Core Idea: Scale up data and models, build autonomous systems.



World Model - Genie 3 by Google



Embodied AI

Wrap-up

- What did we discuss?

- What is AI?

- An agent that perceives its environment and takes actions to maximize expected utility.

- What does an AI system consist of?

- Search, learning, decision-making.

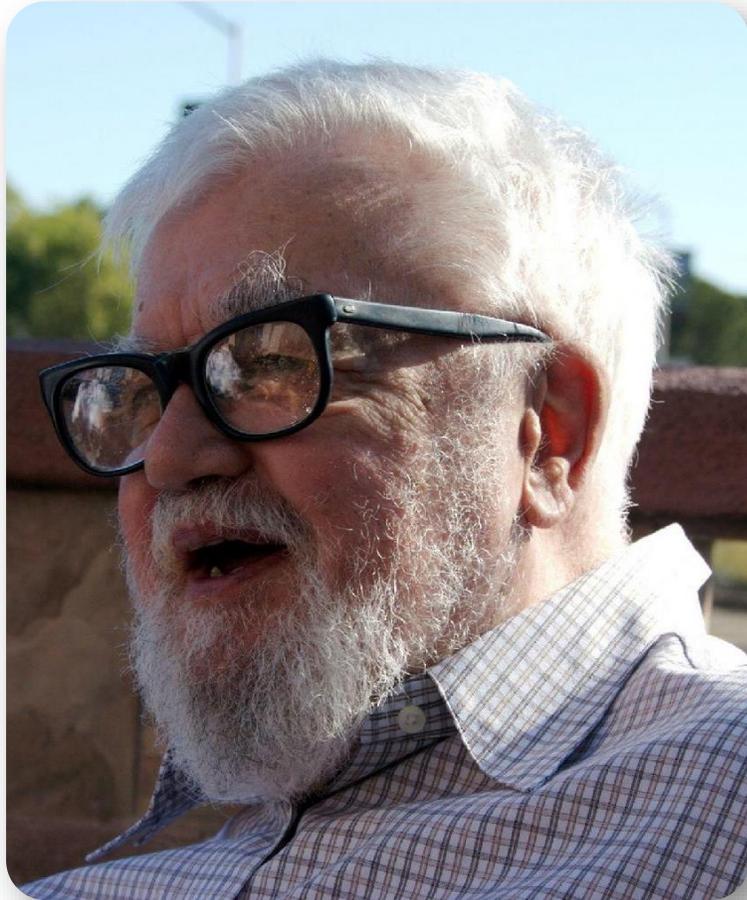
- Discussed the course road map.

- From narrow to general:

- search => ML => DL => Generative AI => RL => Frontier AI

2. A Brief History of AI

The Dartmouth Conference: AI is Born (1956)



- John McCarthy proposed the term “Artificial Intelligence” to unify the field
- AI becomes academic field

Early Optimism (1956-1974)

A period of rapid progress and high expectations in the dawn of AI.

Logic Theorist (1955)

Newell & Simon's theorem-prover, demonstrating machine reasoning.

General Problem Solver (1957)

Universal AI attempt solving diverse problems via means-ends analysis.

ELIZA (1966)

Weizenbaum's chatbot simulating Rogerian psychotherapy.

Microworlds

AI competence in simplified environments like "blocks world".

Optimistic Predictions

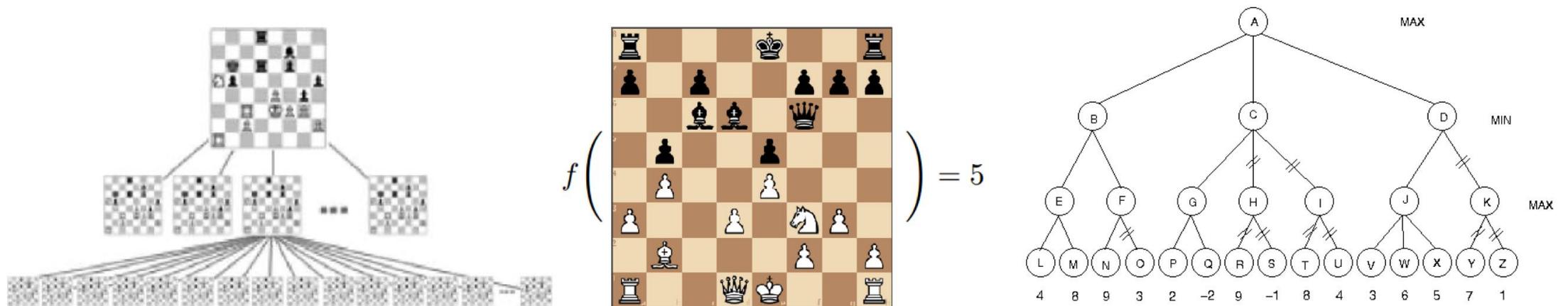
Simon predicted human-level AI soon, driving significant funding.

Milestones

Early successes established AI as a promising field.

Chess as the First Killer App for AI

- Claude Shannon proposed the first chess playing program in 1950
 - It included adversarial search and minimax (later lecture)
 - It also included many heuristics for faster searching



First AI Winter (1974-1980)



Unrealistic Expectations

Early predictions about AI capabilities failed, leading to disillusionment.



Combinational Explosion

Many problems proved hard (combinatorial explosion), and systems failed outside narrow "microworlds".



Hardware too Weak

Hardware was too slow and memory too limited for real-world problems.



The Lighthill Report (1973)

A critical UK report highlighted failures, leading to funding cuts.

"A period of reduced funding and pessimism where the gap between AI's promise and reality became apparent."

Story: The Fall of Machine Translation



The Promise: Cold War Ambitions

Funded heavily by the U.S. government, early MT aimed to automatically translate Russian scientific papers. It was seen as a major AI goal.



The Reality: Poor Translation Results

Simple word-for-word rules led to absurd outcomes. A famous example: "The spirit is willing..." became "The vodka is good..." in Russian.



The ALPAC Report (1966)

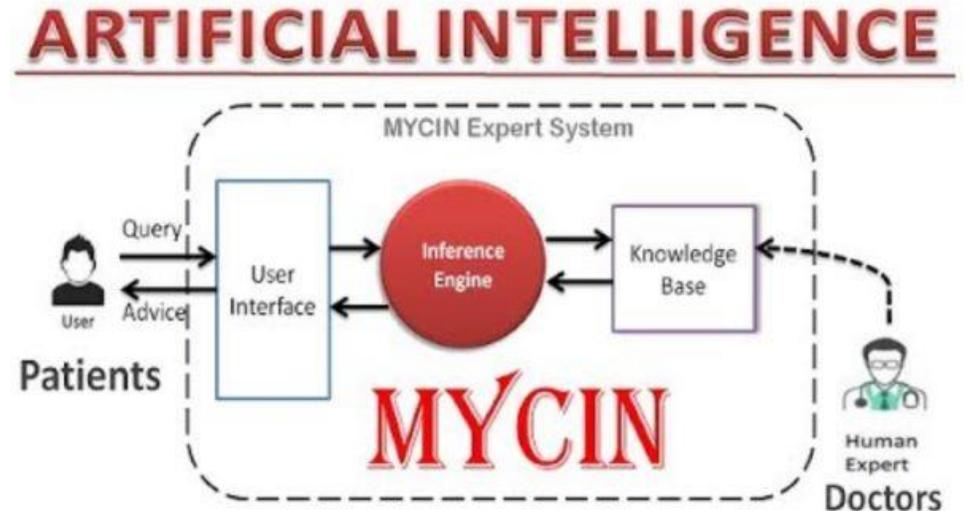
This influential report deemed MT unfeasible, leading to **massive funding cuts** and symbolizing the dashed hopes of the first AI winter.



IBM 701 computer

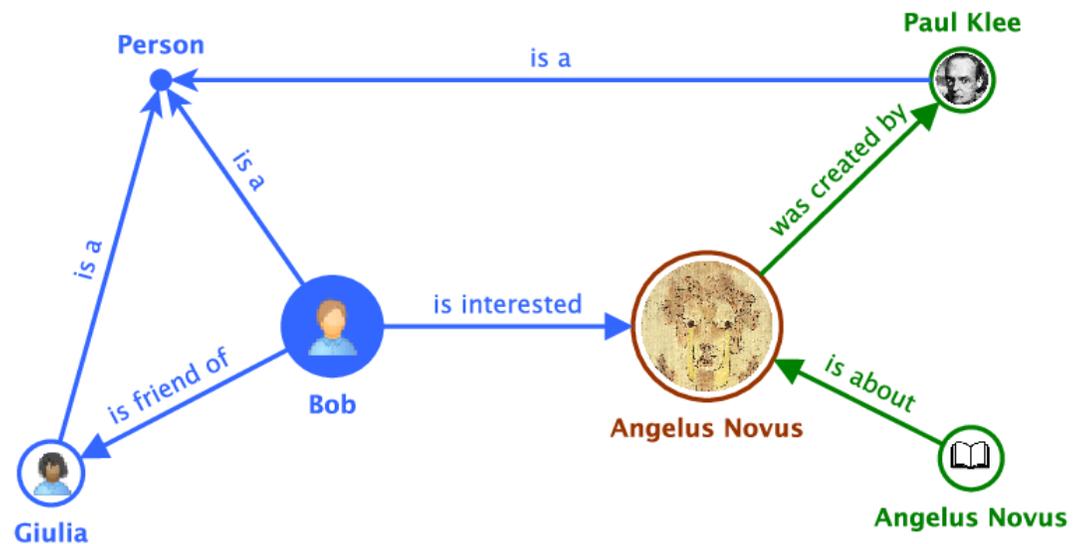
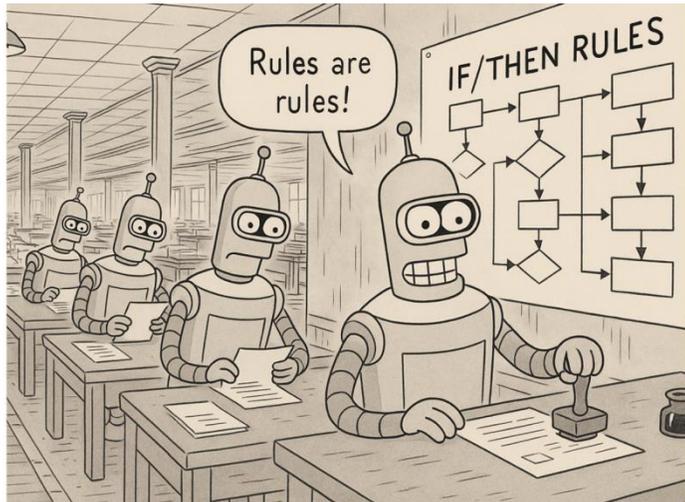
Rise of Expert Systems (1980-1987)

- Core concept & engineering:
 - Narrow domains with expert rules.
 - Knowledge Engineering: Extracting expert knowledge for AI
- Pioneering Systems:
 - MYCIN (1972-1974): Diagnosed bacterial infections.
 - DENDRAL (1965-1976): Determined organic molecule structures.
- Commercial Success:
 - AI moved from academia to industry.
 - Companies like Teknowledge and IntelliCorp drove real-world apps.



Symbolic AI

- **Definition:** AI systems that represent knowledge using explicit symbols (words, numbers, logic) and manipulate them via formal rules.
- Also known as Good Old-Fashioned AI (**GOF AI**)
- **Characteristics:** High interpretability via explicit rules, but brittle with ambiguous real-world data.
- Dominant in 1950s to 1980s



Example: IF (Temperature > 38°C) AND (Cough == True) THEN (Diagnosis = Flu)

Symbolic AI: Successes

- **Deep Blue:** Defeated World Champion Chess Player Garry Kasparov in Chess (1997).



The Failure of Symbolic AI

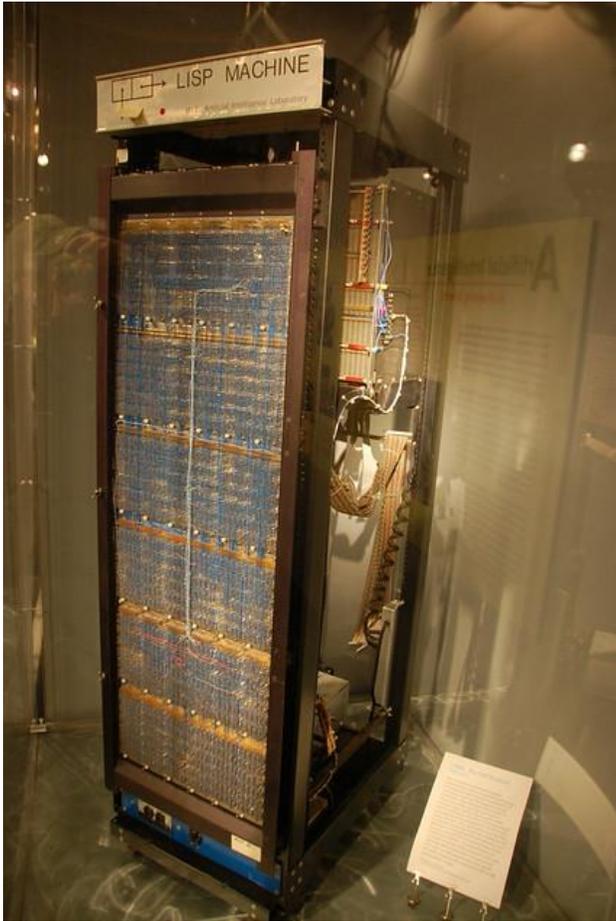
- **Polanyi's Paradox:** "We know more than we can tell."
 - Many human skills (like recognizing a face) are tacit and impossible to fully describe in logic rules.
- **Scalability:** Hand-coding rules for the messy, infinite complexity of the real world is an impossible task.
- **Brittleness:** Symbolic systems often fail completely when inputs deviate even slightly from their rigid rules.
- **Result:** The "AI Winter" (1980s-90s) – withdrawal of funding and loss of academic interest.

Second AI Winter (1987-1993)

The expert systems bubble burst, leading to decline due to high costs, technical limitations, and lost investor confidence.

- **High costs & maintenance:** Expensive to build and required constant updates, making them unsustainable.
- **Technical limitations:** Failed unexpectedly when faced with unprogrammed problems, lacking adaptability.
- **Loss of confidence:** Overpromises led to loss of confidence from investors and governments.

Story: The Collapse of Lisp Machines



The LISP machine in MIT Museum



The Rise: Specialized AI Hardware

In the 1980s, companies like Symbolics built Lisp machines, specialized workstations offering superior performance for AI research and Lisp programming.



The Fall: Cost vs. General-Purpose Computing

Extremely high costs (tens of thousands) made them inaccessible. Meanwhile, affordable workstations from Sun Microsystems rapidly improved, running Lisp efficiently for most applications.

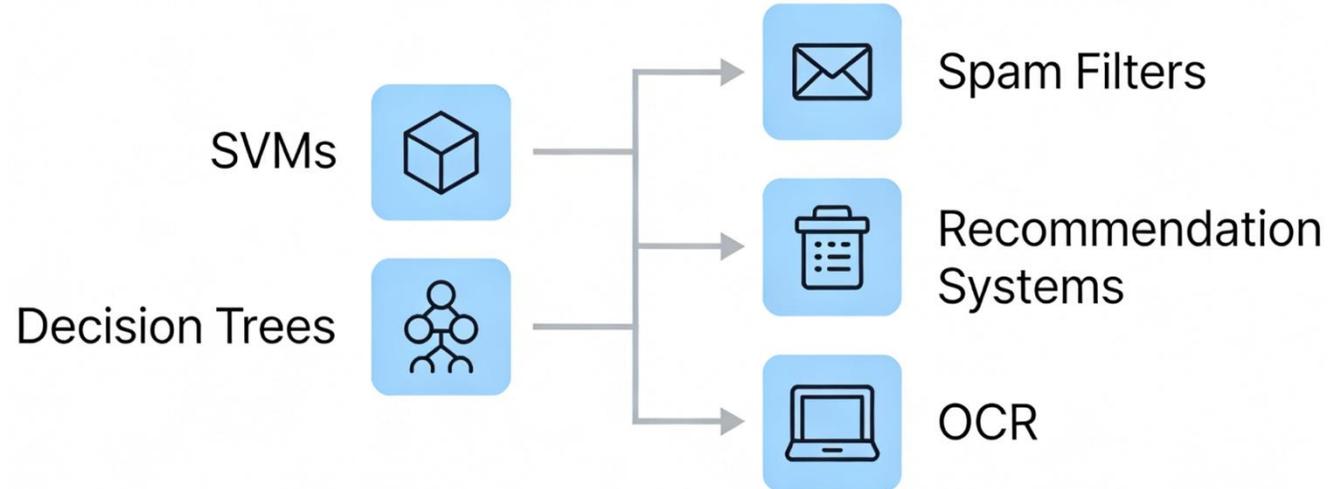


The Impact: Bankruptcy & The AI Winter

The market collapse led to Symbolics' bankruptcy. This commercial failure, combined with expert system limitations, significantly contributed to the second AI winter.

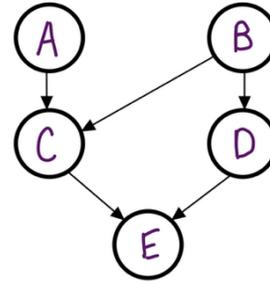
Statistical Learning & Data-Driven Era (1993-2010)

- Paradigm shift:
 - From rules to data
 - Benchmark-driven
- Key machine learning techniques: SVMs, decision trees, etc.

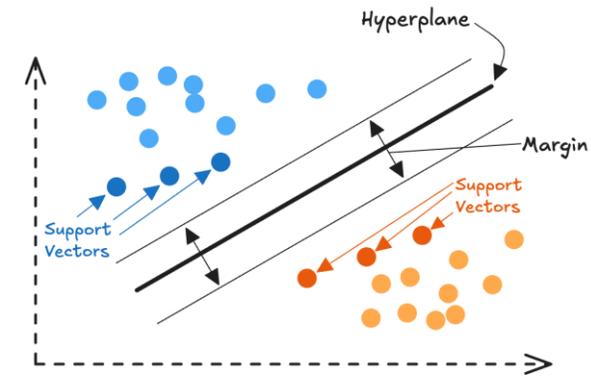


Statistical Machine Learning

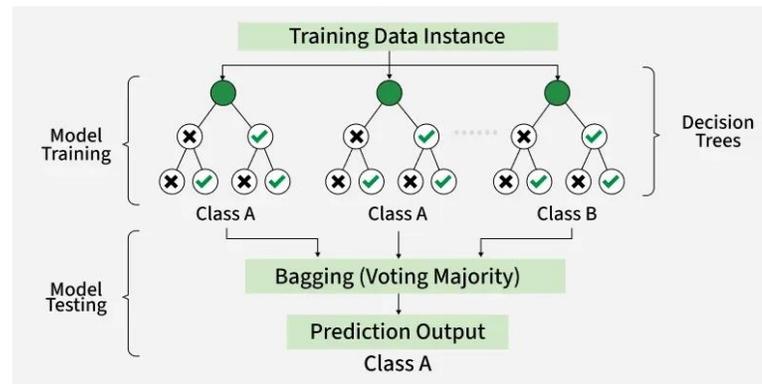
- 1985: Bayesian networks



- 1995: Support vector machines (SVMs)



- 2001: Random forest



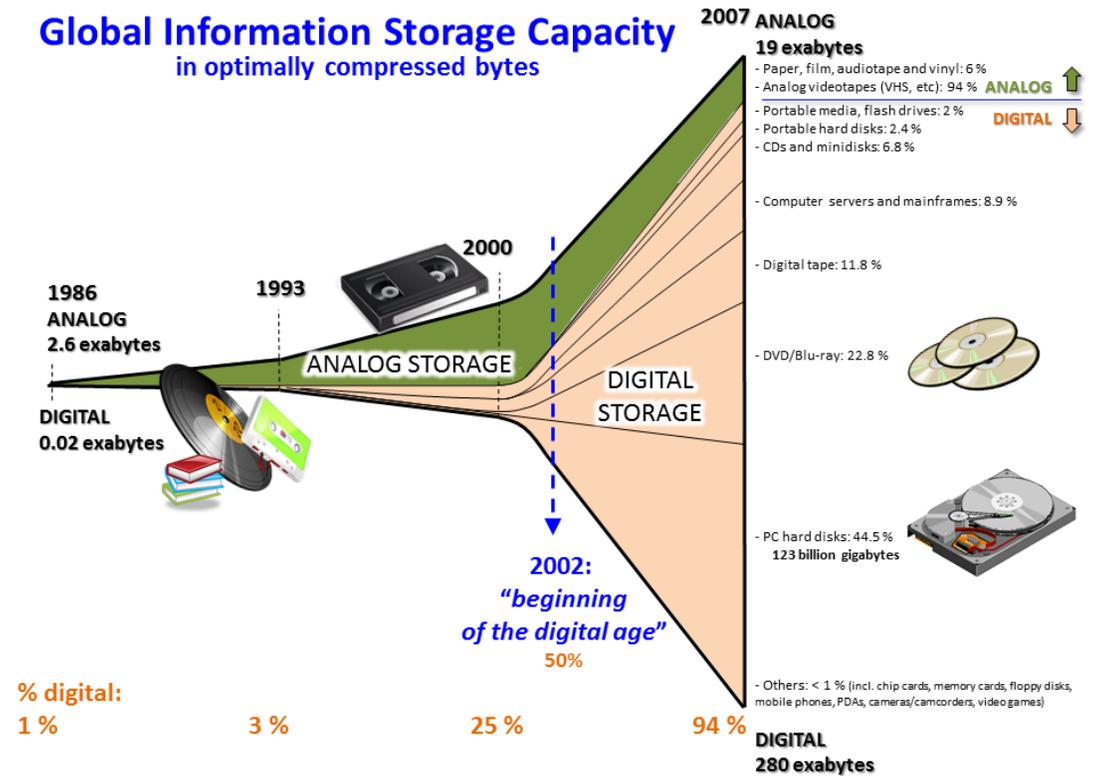
Core Comparison of Symbolic and Statistical AI

Symbolic AI	Statistical AI
Rules	Data
Logic	Optimization
Deduction	Generalization
Human-coded	Learned

- The Old Way (Symbolic):
 - Symbolic AI tried to encode intelligence.
 - Human inputs rules → Computer processes data → Output.
- The New Way (Machine Learning):
 - Statistical AI tries to learn it.
 - Human inputs data + answers → Computer learns rules → Output.

The Internet & Big Data

- Internet -> massive data
- Text, images, audios, videos
- Data fuels ML
- AI integrated into search, social media, e-commerce.



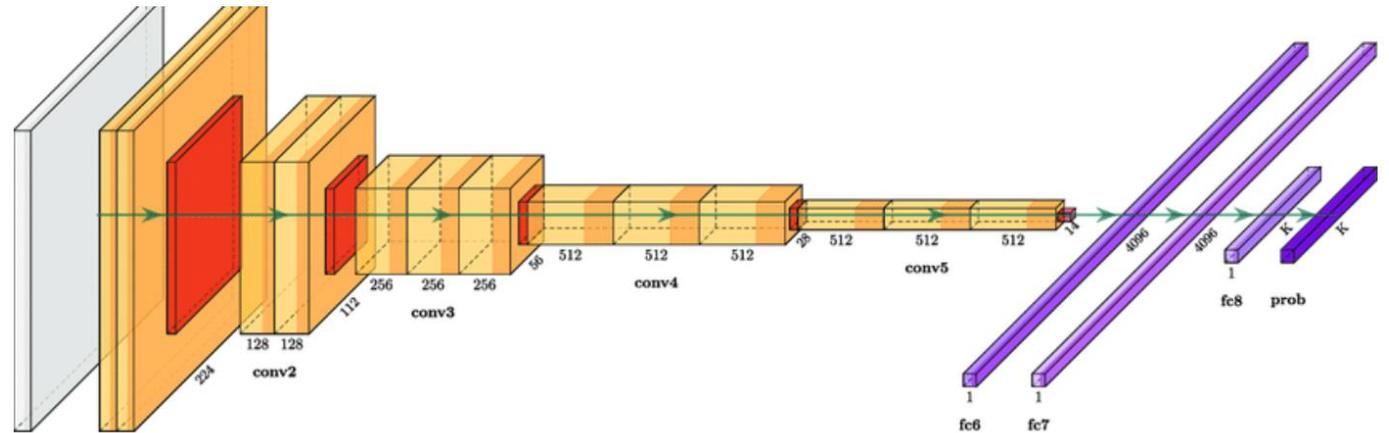
The Increasing Scale of Big Data for Modern AI

- ImageNet (Image Classification):
 - 15 million images
- GPT-3 (Large Language Model):
 - 500 billion tokens
- Stable Diffusion (Image Generation)
 - 5 billion image-text pairs

The Game-Changer: More data + larger models → better performance.

The Deep Learning Revolution (2010-present)

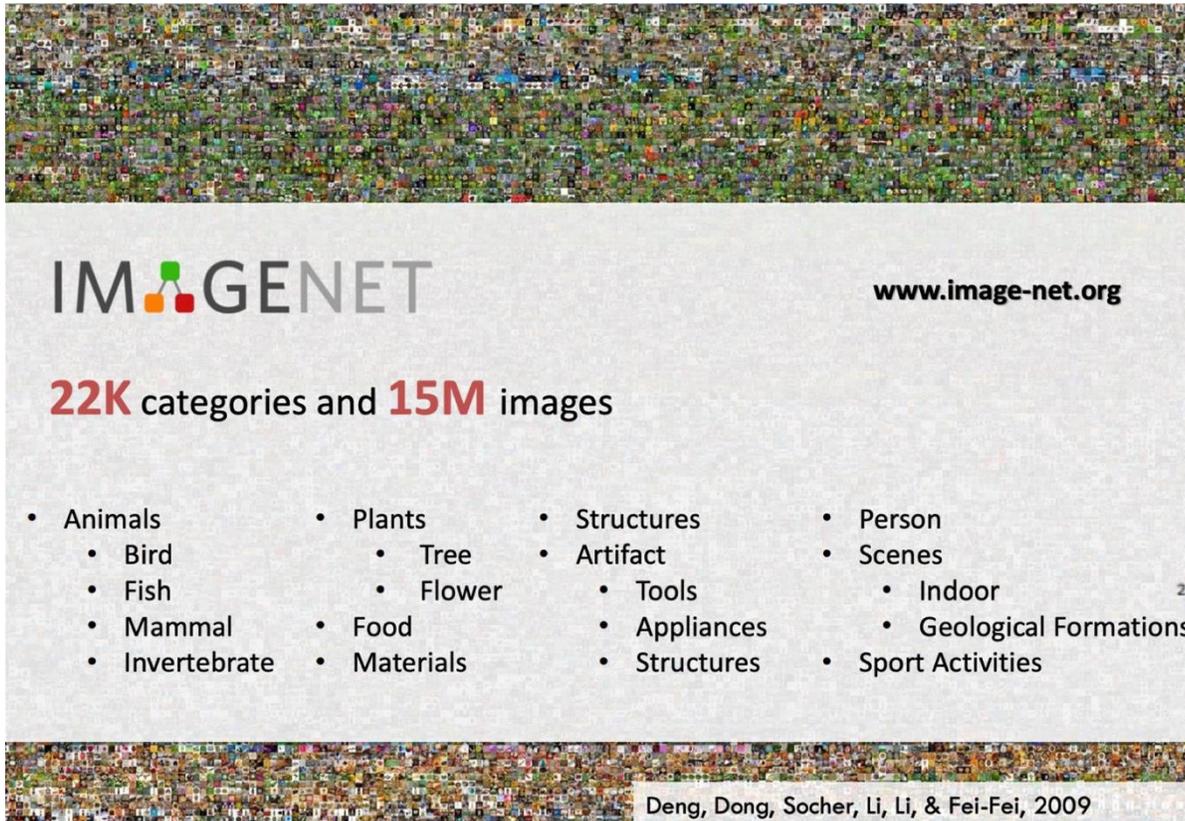
- **Resurgence of Neural Networks:** Deep Learning uses multi-layered Neural Networks to learn complex patterns from data.
- **Why it worked:**
 - **Big data.**
 - **Increasing GPU computation power**
 - **Better training methods:** ReLU, Dropout, BatchNorm, etc.
- **Revolutionized core AI fields:**
 - Computer Vision
 - Natural Language Processing
 - Speech Recognition



VGG-16

Computer Vision: ImageNet, AlexNet

- Key breakthrough: **AlexNet (2012)** won the ImageNet Challenge with a drastically lower error rate (36.7%), proving deep learning's dominance.

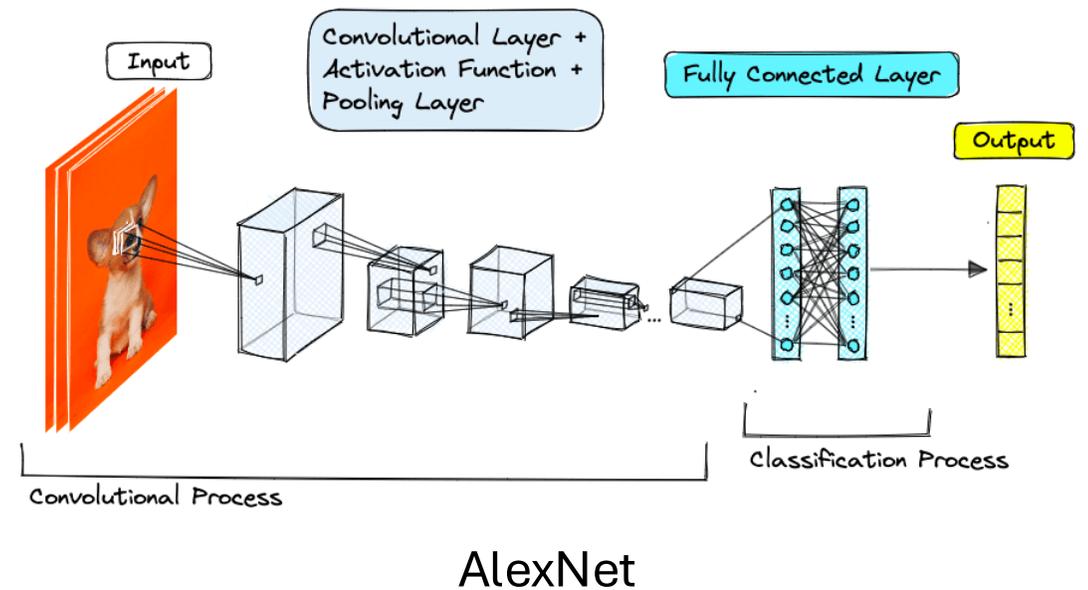


www.image-net.org

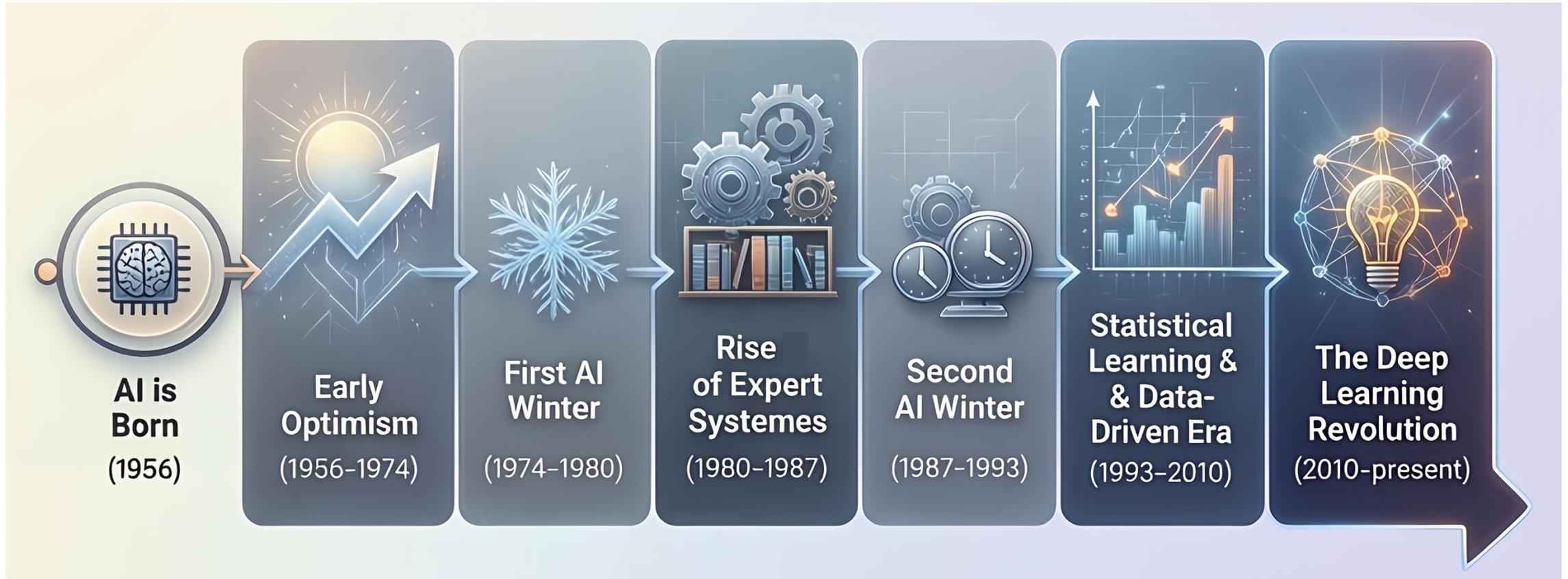
22K categories and **15M** images

- Animals
 - Bird
 - Fish
 - Mammal
 - Invertebrate
- Plants
 - Tree
 - Flower
 - Food
 - Materials
- Structures
 - Artifact
 - Tools
 - Appliances
 - Structures
- Person
 - Scenes
 - Indoor
 - Geological Formations
 - Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009



AI Development Timeline



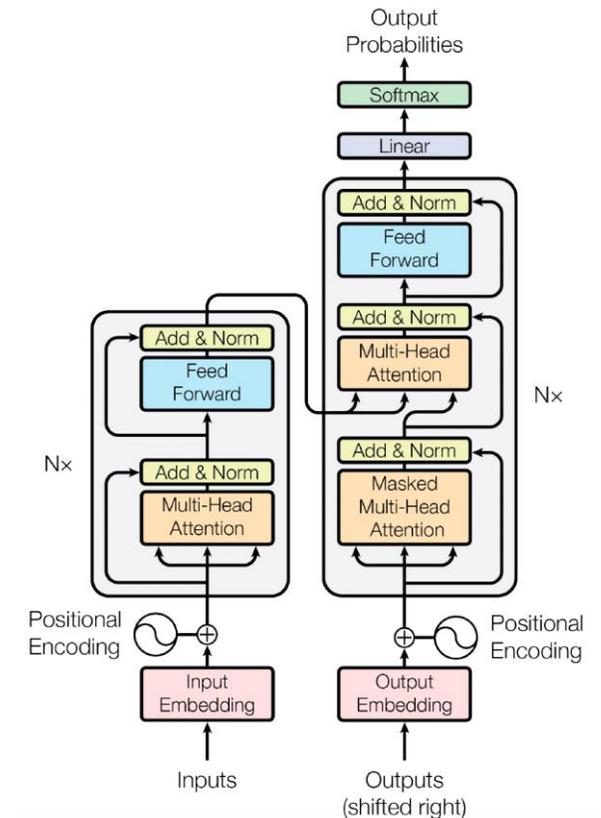
Three Waves of AI

1. Symbolic (1950s-1980s): rule-based systems, logic, reasoning.
2. Knowledge (1980s-1990s): expert systems, encoding human expertise.
3. Deep Learning (2000s-present): data-driven learning, neural networks.

3. Modern AI Landscape

The Transformer Revolution (2017)

- **Mechanism:** "Attention Is All You Need" (Vaswani et al.) introduced the self-attention mechanism.
- **Breakthrough:** Parallel processing of sequences, overcoming the sequential limitations of RNNs.
- **Result:** Enabled efficient training on massive, internet-scale datasets.
- **Impact:** The architectural foundation for all modern LLMs including GPT, BERT, and Llama.



Transformer Architecture

Large Models

Large Language Models

- Billions of parameters
- Trained on web-scale text
- **Emergent abilities:** Capabilities that appear at scale, such as in-context learning, reasoning, and complex coding.

Multimodal Large Models

- Support **text, image, audio, and video**
- Moving towards **understanding the world like humans do**

Large Models

ChatBot

文心一言
 360 智脑
 Baichuan-Chat
 ChatJD
 豆包
 MOSS
 商汤 | SenseChat
 天工
 Kimi Chat
 智谱清言
 知海图Chat

行业大模型应用

客服营销 CeMeta 赤兔机器猫 加我智能 淘宝问问 思涌 AI	代码 Baidu Comate CODEFUSE CodeGeex 通义灵码	协同办公 WPS AI 钉钉智能助手 My AI	其他 文心一格 叮当 保药 光子金融智能助手 魔音工坊 妙鸭相机 Hi Echo 如影 通义法睿 通义仁心 深睿医疗智能多模态科研平台 讯飞智作
--	---	--	---

行业大模型

游戏电商 百川角色大模型 EcomGPT 盘古跨境电商大模型 玉言·游戏行业大模型	社交文娱 阅文妙笔大模型 Social GPT 文曲大模型	传媒影视 文心·传媒/影视行业大模型 盘古媒体大模型 中文逍遥创作大模型	教育 KidsGPT 九章大模型 星辰教育大模型 「子曰」教育大模型	汽车 领克汽车服务大模型 仓颉汽车大模型 DriveGPT 吉利星睿大模型 盘古汽车大模型	金融 文心·金融行业大模型 AntFinGLM 金融行业大模型 LightGPT 马上消费金融大模型 妙想 金融大模型 文因金融大模型 FD-LLM INFINITY 星环无限金融大模型 招联智鹿金融大模型	能源制造 文心·能源/制造行业大模型 星辰矿山大模型 盘古矿山大模型	医疗 百度灵医大模型 BenTsao (本草) HIT-SCIR Health Intelligence CareGPT 叮当HealthGPT DoctorGLM 星火医疗大模型	其他 文心·航天/城市/社科/行业大模型 白玉兰科学大模型1.0 ChatLaw 法律大模型 携程问道 科大讯飞营销大模型 盘古气象大模型 炎南大模型
---	--	---	---	---	--	---	---	---

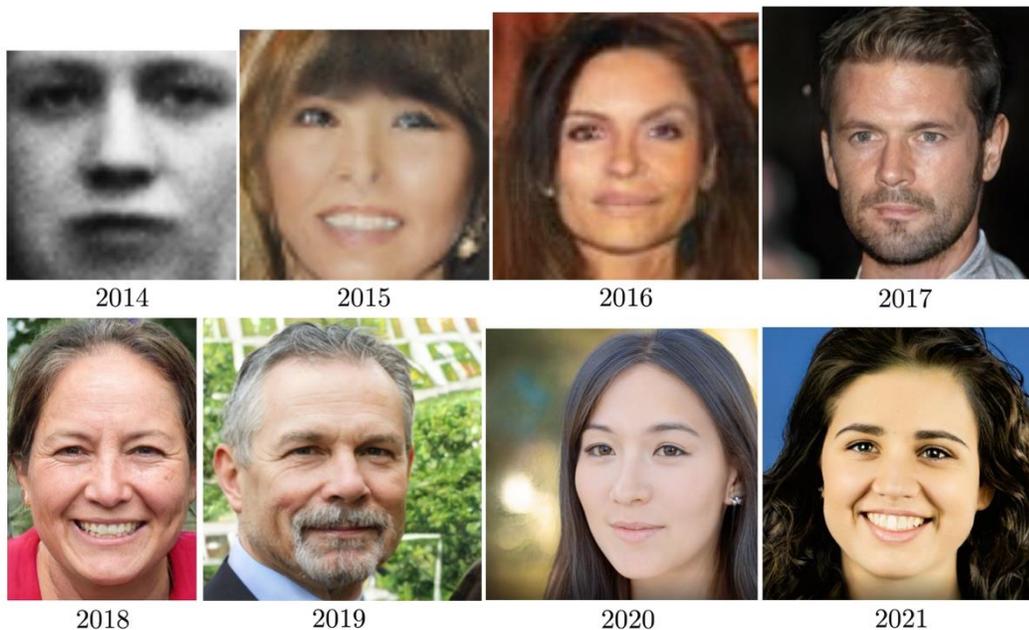
通用大模型

云厂商 HUAWEI 盘古 阿里云 通义 百度 文心 JD.COM 言犀 Tencent 腾讯 混元	服务商 中国移动 九天 中国电信 星河 China unicom 中国联通 鸿湖	科技公司 Baichuan inspur 浪潮 源1.0 MINIMAX 开放平台 基础模型 360 智脑 智谱·AI ChatGLM KUNLUN 天工 山海 孟子 日日新 从容 星海	研究机构 清华大学 PolyLM moss 紫东太初 国家超级计算天津中心 天河天元 idea 二郎神 BAAI 智源研究院 悟道 序列猴子
--	--	--	--

Generative AI: Beyond Text

- **Image Generation:** Diffusion models create high-fidelity visuals.
- **Video Generation:** Progressing towards world modeling.
- **Audio/Speech:** Highly realistic synthesis of human speech and music.

Image Generation



7.5 years of GAN progress on face generation



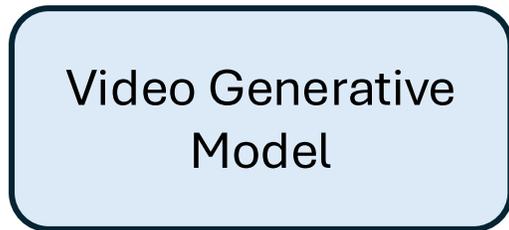
Text-to-image generation with Qwen-Image

“A high-fidelity, ultra-detailed portrait of a young woman from the Han Dynasty (汉朝), standing gracefully in a classical Chinese garden during early morning.”

Text/Image-to-Video Generation

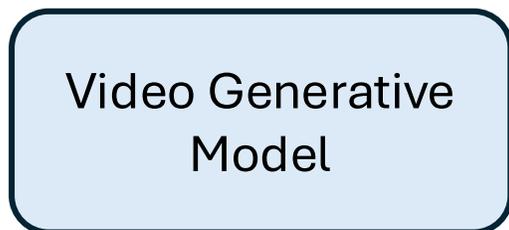


“A running dog.”



万相AI生成

“A running dog.”



万相AI生成

Audio Generation

Text-to-audio

Splash in pond



Doorbell rings



Birds singing in
the garden



Text-to-music

Traditional Irish fiddle
plays lively music



Vibrant Brazilian
samba drum rhythm

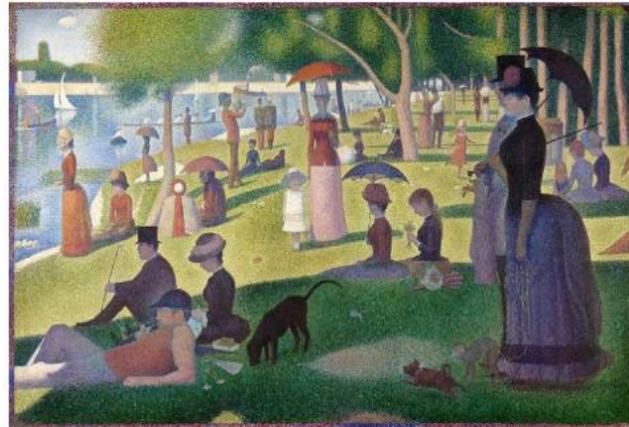


Text-to-speech

Calcraft served the
city of London till
eighteen seventy-four,
when he was
pensioned at the rate
of twenty-five
shillings per week.



Image-to-Audio Generation

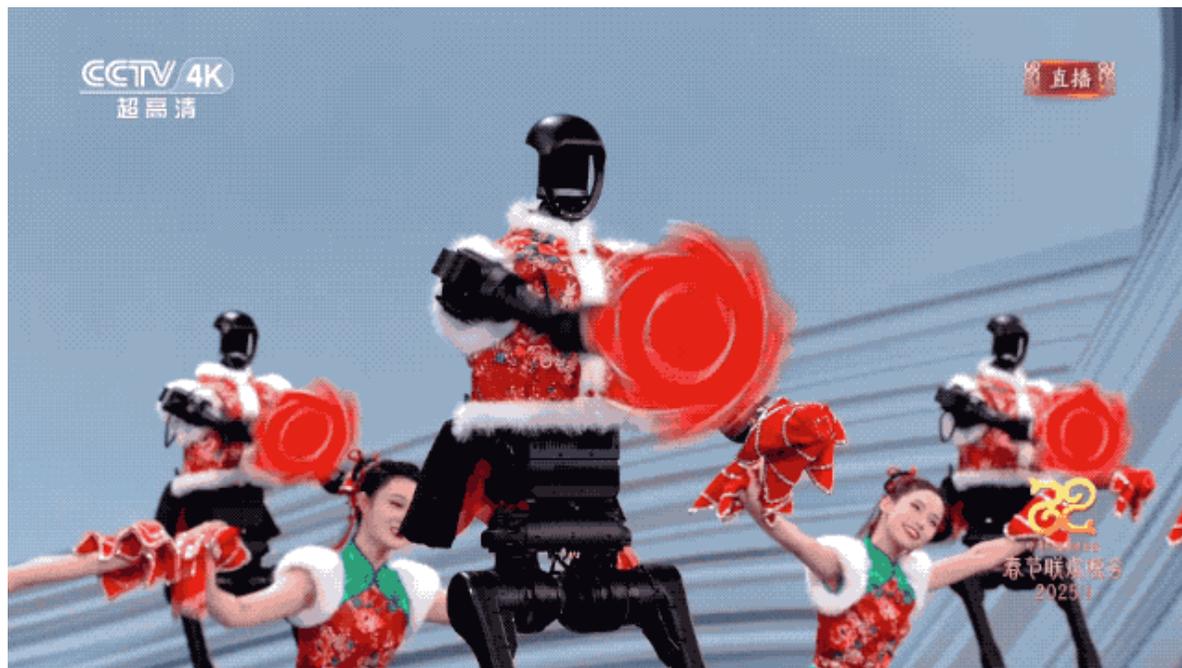


Embodied AI & Robotics

- **Goal:** Bringing intelligence into the physical world through robotic agents.
- **Shift:** From control rules → learned policies.
- **Challenge:** *Moravec's Paradox* – High-level reasoning is relatively easy for AI, but low-level sensorimotor skills are remarkably difficult.

Embodied AI & Robotics

UniTree G1's Evolution

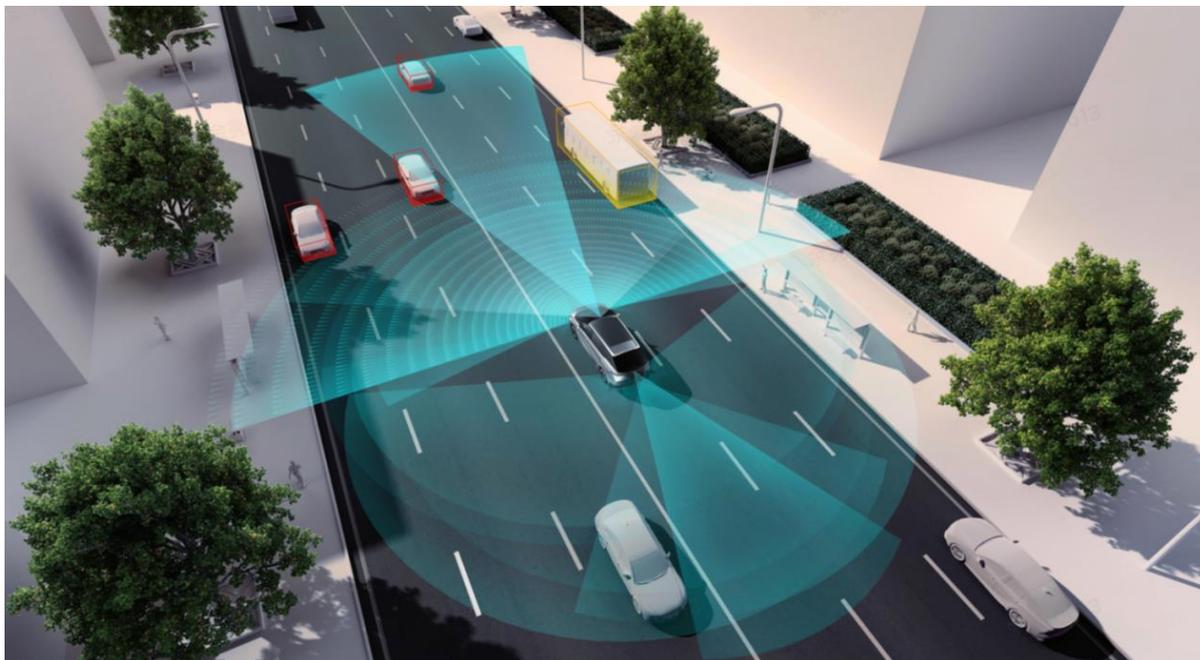


2025



2026

Autonomous Driving



现阶段新能源汽车品牌智驾能力天梯图

新出行-中国领先的新能源车垂直媒体

数据整理至: 2025年2月

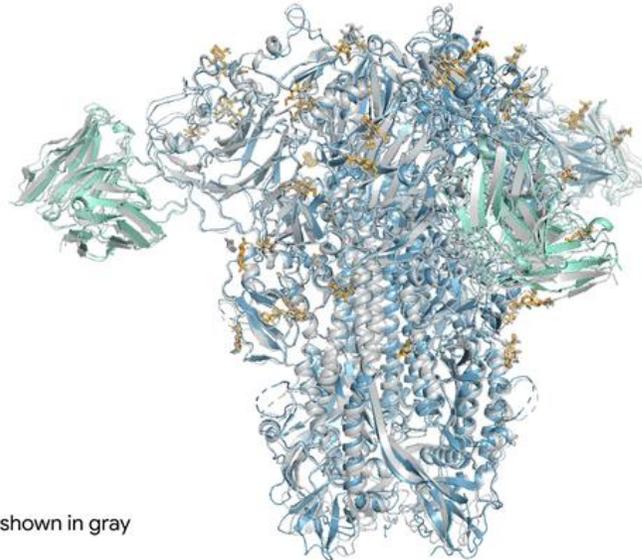


AI for Science

AI is becoming an essential tool to accelerate the rate of scientific discovery.

- **Biology:** AlphaFold solving protein structure prediction, a 50-year-old challenge.

7PNM



Ground truth shown in gray

AlphaFold 3's structural prediction for a spike protein (blue) of a cold virus as it interacts with antibodies (turquoise) and simple sugars (yellow), accurately matches the true structure (gray).

AI for Science

- **Materials:** GNoME discovering millions of new crystal structures for next-gen batteries.



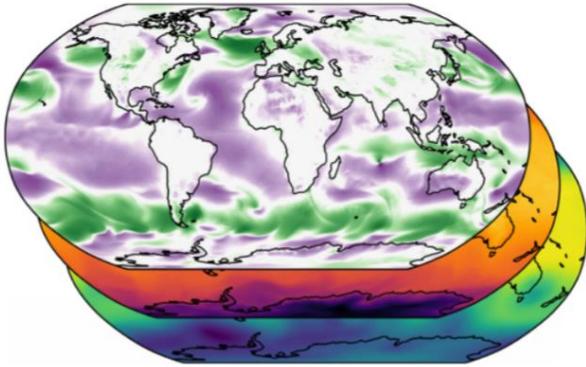
AI tool GNoME finds 2.2 million new crystals, including 380,000 stable materials that could power future technologies.

<https://deepmind.google/blog/millions-of-new-materials-discovered-with-deep-learning/>

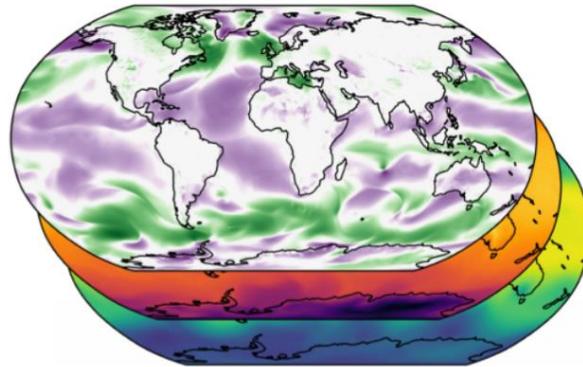
AI for Science

- **Weather:** GraphCast outperforming traditional supercomputer simulations for global forecasts.

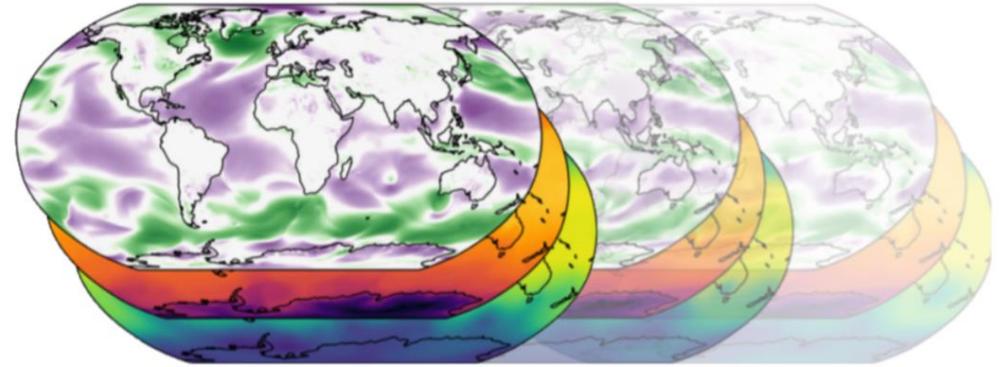
a) Input weather state



b) Predict the next state



c) Roll out a forecast

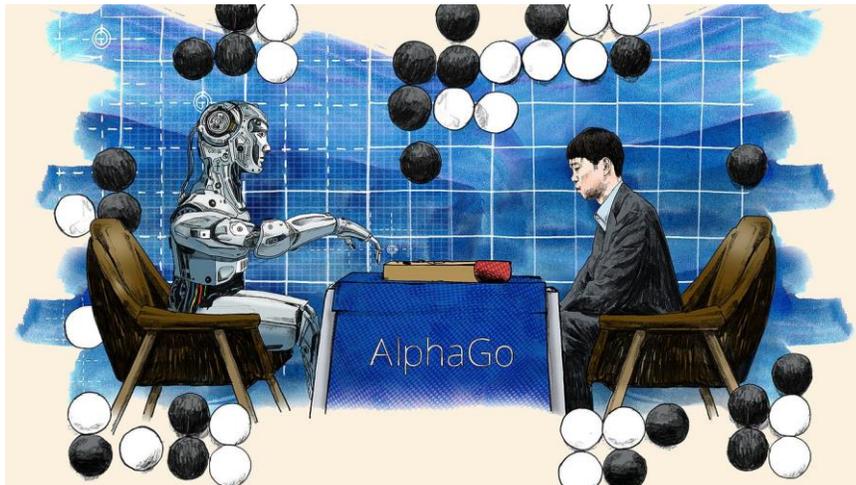


GraphCast delivers 10-day weather predictions at unprecedented accuracy in under one minute. <https://deepmind.google/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/>

4. Weak AI vs. AGI

Weak AI (Artificial Narrow Intelligence)

- **Definition:** AI systems specialized and trained for a **single task** or a limited range of tasks.
- **Characteristics:**
 - Superhuman performance in its specific domain (e.g., Chess, AlphaFold).
 - Fails immediately when applied outside that narrow domain.
 - No consciousness, self-awareness, or genuine understanding of the world.
- **Examples:** Siri, Face Recognition, AlphaGo, Tesla Autopilot.



AlphaGo



Face Recognition

Weak AI: Limitations



Lack of Common Sense

Struggles with basic reasoning, e.g., understanding physical impossibilities.



Transfer Learning Difficulty

Hard to apply knowledge from one domain to another without retraining.



Sensitive to Adversaries

Easily fooled by small input changes, leading to misclassification.



Over-Reliance on Data

Requires massive labeled data, which is expensive and time-consuming.



No Autonomy

Cannot set goals or improve architecture without human help.

These limitations define the boundary between Narrow AI and AGI.

Artificial General Intelligence (AGI)

- **Definition:** A hypothetical AI system with the ability to learn and perform any intellectual task that a human being can do.
- **Key Traits:**
 - **Generalization:** The ability to transfer knowledge and skills from one domain to an entirely different one.
 - **Reasoning:** High-level planning, understanding causality, and applying common sense.
 - **Autonomy:** The ability to set its own goals and improve its own performance without human supervision.

AGI Benchmarks & Tests

- **Turing Test:** Can it fool a human interrogator into thinking it is human during a natural language conversation? (Now considered a low bar for "intelligence").
- **The Coffee Test (Steve Wozniak):** Can a robot enter a random, unfamiliar home, find the kitchen, identify the coffee machine/ingredients, and brew a cup of coffee?
- **The Employment Test:** Can the AI economically replace a median human worker across a wide variety of job types?



The Coffee Test

Are We There Yet?

- **Current State:** We are firmly in the era of **Weak AI**, but the boundaries are beginning to blur as systems become more multi-capable.
- **LLMs as Proto-AGI?:** Models like GPT-4 show impressive generality (coding, poetry, logic), yet still lack a stable world model and reliable long-term planning.
- **Timeline:** Expert estimates for AGI arrival vary wildly – some say as early as 2027, others predict 2050 or later.

AGI: Concerns

- **Safety & Alignment:** Ensuring AGI goals align with human values
- **Economic Impact:** Potential displacement of knowledge work and labor markets.
- **Misuse:** Powerful tools in the hands of bad actors (cyberattacks, disinformation).
- **Existential Risk:** The debate over whether superintelligence poses a threat to humanity.



The "Intelligence" Gap

- **Human Intelligence:**
 - Low energy consumption (~20W).
 - Small data requirements (learns from years of life experience).
 - High efficiency and multi-task generalization.
- **AI Intelligence:**
 - High energy consumption (megawatts for training/inference).
 - Massive data requirements (requires the entire internet).
 - Brute force optimization through enormous compute power.
- **Goal:** Bridging this gap towards more efficient, robust, and truly General Intelligence.

Summary

- **Evolution:** A journey from Symbolic (Logic/Rules) to Statistical (Data/Patterns) and now Generative (Modeling/Creation).
- **Landscape:** LLMs, Robotics, and Science are the current frontiers where AI is having the most profound impact.
- **Goal:** The ongoing transition from specialized Narrow AI towards the ultimate goal of Artificial General Intelligence (AGI).