

# Gen-AI: Technical and Social

Lecture 01: Introduction to Gen-AI

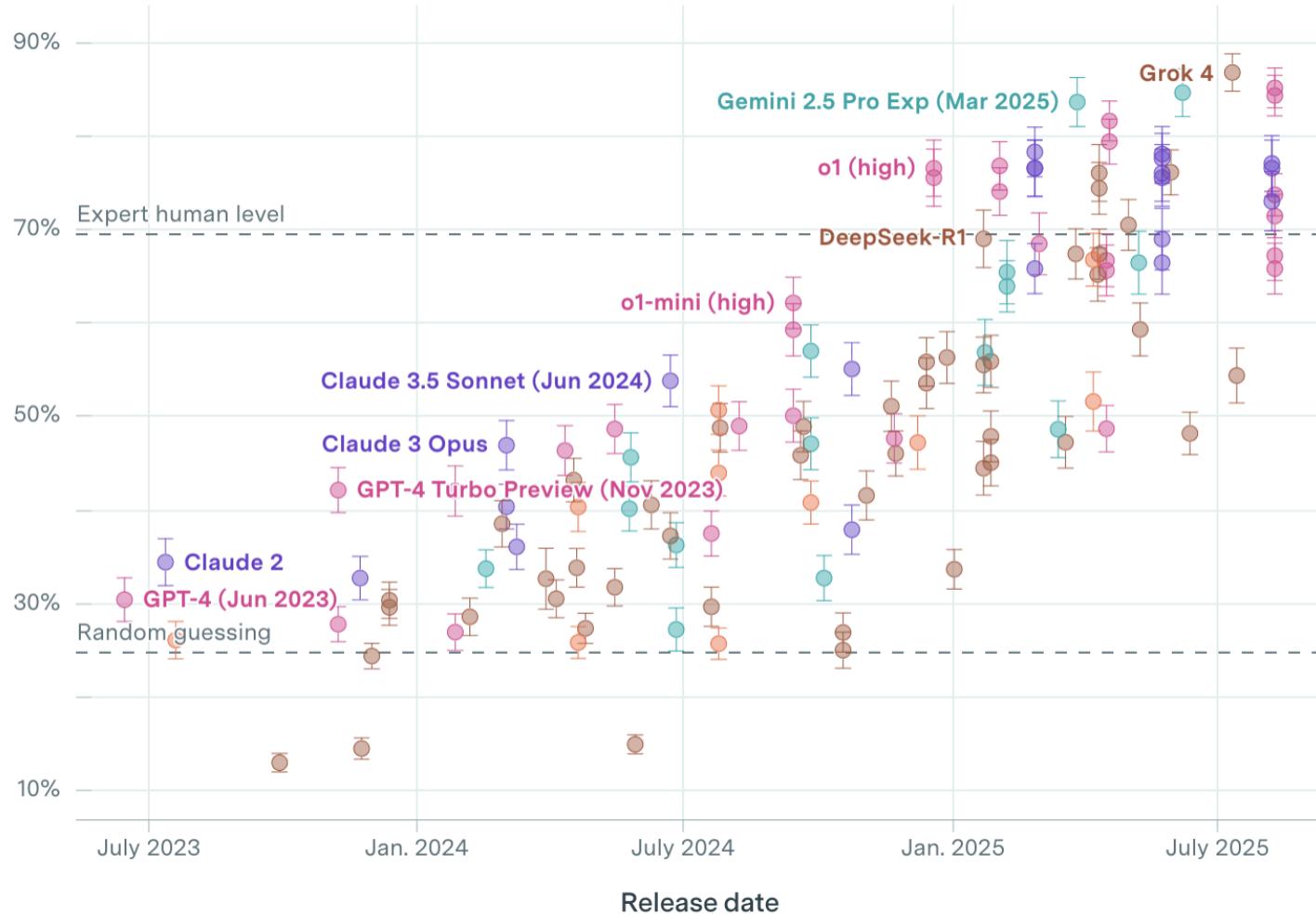
# AI performance on a set of Ph.D.-level science questions

GPQA Diamond accuracy ⓘ

122 Results ⓘ

Organization

- OpenAI
- Anthropic
- Google ⓘ
- Meta AI ⓘ
- Other



2023

2025

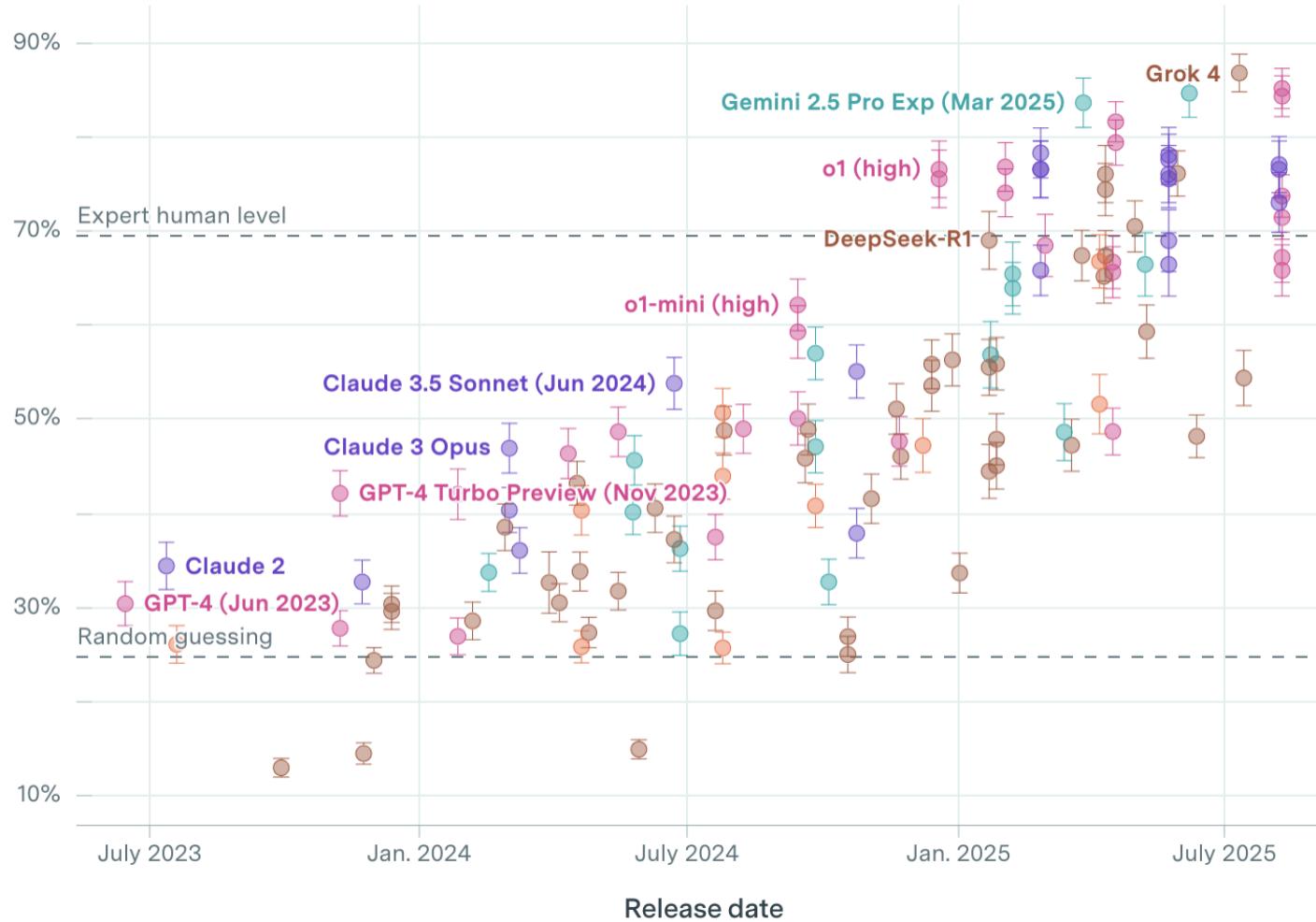
## AI performance on a set of Ph.D.-level science questions

GPQA Diamond accuracy ⓘ

122 Results ⓘ

Organization

- OpenAI
- Anthropic
- Google ⓘ
- Meta AI ⓘ
- Other



Why is AI making progress this fast?

Are we in a normal time of history?

How can I contribute to this transformation?

Should we plan our careers as usual?

Will we live in a better world if AGI is achieved?

# Instructor - Tianyi Peng

- Assistant Professor at DRO
- Education
  - PhD: MIT OR
  - Undergraduate: Tsinghua CS (Yao Class)
- Research interests:
  - AI-driven decision making
- Advising Gen-AI startups in
  - e-commerce
  - data labeling
  - digital twin



# TA - Haonan Wang

- PhD Student advised by Prof. Eugene Wu
- Undergrad at Peking University, MS at CMU; worked @ OtterTune (ML-for-databases startup from CMU research)
- Research interests:
  - Data Agent / Deep-Research Agent for Tabular Datasets
  - Computer-Use Agent
  - System for Agents



# Today's class

1. Course Overview and Logistics
2. Introduction to Gen-AI
3. Pytorch, Kaggle, and Text Classification

# About this course

- **Generative AI: Technical and Social**
- PhD Seminar
- The class is split into two parts
  - **The Technical:** lecture on building LLM and AI agents
    - (Subject to Change)
    - Fine-tuning Transformers
    - Prompt Engineering
    - Retrieval Augmented Generation (RAG) and RLHF
    - Agentic Framework and Evaluation
    - LLM Serving
    - Digital Twin
    - MCP, Vibe-coding ....

# About this course

- **Generative AI: Technical and Social**
- PhD Seminar
- The class is split into two parts
  - **The Technical:** lecture on building LLM and AI agents
  - **The Social/Practice:** industrial speakers talking about AI startups and AI practice
    - Cofounders/CEO/CTO of AI startups
    - Anthropic and LinkedIn

# Syllabus

Swaps between Kravis 490 (business school) and CSB 451 (Mudd)

Date	Week	Location & Time	Theme	Notes	Assignments
9/2/25	1	Kravis 490 6-9PM	Lecture: Introduction and Transformer		
9/9/25	2	CSB 451 7-9PM	Speaker: Aaron Vontel	Anthropic, Member of the Technical Staff; Former Founder & CEO for Regression Games	
9/16/25	3	Kravis 490 6-9PM	Lecture: Text Generation and Agent Benchmark		Due: HW1
9/23/25	4	CSB 451 7-9PM	Speaker: Tom Effland	Noetica, CTO & Co-Founder	
9/30/25	5	Kravis 490 6-9PM	Lecture: Prompt Engineering in the Agent Ara		Due: HW2
10/7/25	6	CSB 451 7-9PM	Speaker: Xiaofeng Wang	LinkedIn, leading AI agent platform development	
10/14/25	7	Kravis 490 6-9PM	Lecture: RAG and Memory Layer		Due: HW 3
10/21/25	8	CSB 451 7-9PM	Speaker: Sushanth Raman	Pallet, CEO & Founder	
10/28/25	9	Kravis 490 6-9PM	Lecture: Test-of-time compute and Efficient Serving		Due: Project Proposal
11/4/25	10	CSB 451 7-9PM	Speaker: Eno Reyes	Factory, CTO & Co-Founder	
11/11/25	11	Kravis 490 6-9PM	Lecture: Reinforcement learning; Environment; and Digital Twin		
11/18/25	12	CSB 451 7-9PM	Speaker: Anish Agarwal	Traversal, CEO & Co-Founder	
11/25/25	13	Kravis 490 6-9PM	Student Presentations		Final Project Presentation
12/2/25	14	CSB 451 7-9PM	Speaker: Anish Das Sarma	Reinforce Labs, Founder and CEO	Due (12/5): Final Project Writeup

# Course Logistics

# Course Philosophy

This is a course in the intersection between foundational Gen-AI technology and its business/social applications

This is **not** a course about:

- Math and theoretical analysis of deep learning
- Nor is it a coding heavy class on cutting edge LLM engineering
- Nor is it a passive learning class where instructors spoon feed you the answers

Most importantly: Enjoy learning, enjoy exploring, and have fun!

# Prerequisites

- Familiarity with fundamental machine learning concepts (such as training/validation/testing, neural net, and activation function etc)
- Familiarity with Python at a basic level
  - <https://www.kaggle.com/learn/python>
  - <https://www.kaggle.com/code/legendadnan/numpy-tutorial-for-beginners-data-science>

# Course format

Alternate between

- Three-hour lectures at CBS
- Two-hour industrial seminars at the main campus
  - 7:00-8:30pm talk
  - 8:30-9:00pm happy hour

# Industrial Seminar

- Pre-seminar, a short summary
  - The background of the speaker
  - The background of the company
  - The questions you are interested in
- After-seminar, a short summary
  - Does it answer your questions?
  - What you have learnt

Speaker: Aaron Vontel	Anthropic, Member of the Technical Staff; Former Founder & CEO for Regression Games
Speaker: Tom Effland	Noetica, CTO & Co-Founder
Speaker: Xiaofeng Wang	LinkedIn, leading AI agent platform development
Speaker: Sushanth Raman	Pallet, CEO & Founder
Speaker: Eno Reyes	Factory, CTO & Co-Founder
Speaker: Anish Agarwal	Traversal, CEO & Co-Founder
Speaker: Anish Das Sarma	Reinforce Labs, Founder and CEO

# Course project

Explore fun ideas through coding, experiments, designing an MVP, or modeling and theory

We **encourage** you to

- Think of how AI can change your field
- Be ambitious and creative
- Be concrete and grounded
- The project doesn't have to be fully complete before the ddl, but should be summarized in a research-paper format
- Timeline
  - Proposal due: Oct 28
  - In-class presentations: Nov 25
  - Final project due: Dec 05

# Grading

- Class participation 10%
- Homework+Seminar Summary 40%
- Project 50%

# Computing Resource

- Unfortunately we won't be able to reimburse computing resources in this class due to policy restriction
- Prepare to spend a few hundred dollars for this class
- Kaggle and Google Colab have some free GPU resources
- GPU Server: Lambda.AI, Vast.AI
- LLM API: OpenRouter, free credits from Google, etc

# Community

- Slack
  - Technical questions
  - Interesting Ideas
  - Sharing thoughts/news/materials



# Introduction to Gen-AI

# What LLM/AI agents can do now

- Vibe-coding website
  - Bolt.new
  - V0
  - Lovsky
  - ...

The screenshot shows the LullaGPT web application. At the top, it says "Welcome to LullaGPT" and "Your intelligent baby care assistant". Below that, a message from the AI says:

Hi Baby Nova! 😊  
Today is your 125th day in this world!  
That's 17 weeks of joy and love  
Current time in New York: 2025/09/01 22:55:27

A user message "hi" is shown at 10:55 PM. The AI responds:

Good evening! I'm Lulla, your baby care assistant. How can I help you today?  
10:55 PM

The AI asks:

Hello! How can I assist you today with baby Nova's care? 😊  
10:55 PM

On the right, there are three cards showing recent activity:

- Last Feeding 1h ago 120ml formula
- Last Pumping 3h ago 80ml collected
- Last Diaper 1.5h ago Wet only

Below that is a "Recent Activity" section with three items:

- Morning feeding 120ml formula about 1 hour ago 9:55 PM
- Diaper change Wet only about 2 hours ago 9:25 PM
- Morning pumping 80ml collected about 3 hours ago 7:55 PM

# What LLM/AI agents can do now

- Coding agent
  - Claude Code
  - Cursor Agent
  - ...

Let me check what format these part files are in:

```
head -n 2 /Users/tianyipeng/Documents/Operations_Research_Code/  
25_08_Tencent_Competition/Tencent_Competition/creative_emb/emb_81_32/  
part-0000  
  
7050765454769135, 0.2262369692325592, 0.19315361976623535, -0.23622460663318634, -0.248  
17163711786270142, -0.1931087076663971, 0.1656423658132553, -0.16091668605804443, 0.133  
.24578072130680084, 0.22131569683551788, 0.25304272770881653, -0.06331134587526321, -0.  
-0.043621841818094254, -0.14693573117256165, 0.005575160961598158, -0.0935561731457710  
0545, -0.08533520251512527, -0.206947460770607, 0.2643642723560333, -0.1181817501783371  
4648, 0.24119947850704193, -0.2542913556098938, 0.106258325278759, 0.021744949743151665  
  
{"anonymous_cid": 20001627000, "emb": [0.025650011375546455, 0.24831151962280273, -0.16  
0.26538681983947754, 0.12499596923589706, 0.026517320424318314, -0.06490027904510498, 0  
, -0.03005300648510456, -0.006170713808387518, -0.26344579458236694, 0.1723612844944000  
0348, -0.19932417571544647, -0.11119282245635986, -0.23378856480121613, -0.164482533931  
559259415, 0.2487211674451828, -0.1363193243741989, 0.2521066665649414, 0.2497043013572  

```

Perfect! The data is in JSON format with `anonymous_cid` and `emb` fields. The code should be able to handle this format, but it's specifically looking for a `.pkl` file for feature ID '81'. Let me modify the dataset loading code to handle this case properly:

Read [dataset.py L440-474](#)

I see the issue! The code has a special case for feature ID '81' that tries to load from a `.pkl` file, but your data is in the `part-*` format. Let me fix this by modifying the dataset loading logic to handle the `part-*`

The screenshot shows the GitHub repository page for 'LLM Batch Helper'. It includes the README, Apache-2.0 license, and a brief description of the package. The repository has 2 contributors: TiansiPeng and claude. The languages used are Jupyter Notebook (57.1%) and Python (42.9%).

**LLM Batch Helper**

pypi package 0.2.0 downloads 2k downloads/month 2k docs passing License MIT

A Python package that enables batch submission of prompts to LLM APIs, with built-in async capabilities, response caching, prompt verification, and more. This package is designed to streamline applications like LLM simulation, LLM-as-a-judge, and other batch processing scenarios.

[Complete Documentation](#) | [Quick Start Guide](#)

Contributors 2

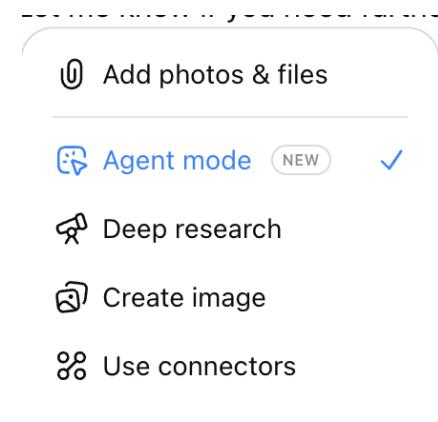
- TiansiPeng Tiansi Peng
- claude Claude

Languages

- Jupyter Notebook 57.1%
- Python 42.9%

# What LLM/AI agents can do now

- Deep Research/Agent Mode



i want to collect the detail information for each CEO/CTO below, can you go to their company websites, linkedin, wikipedia, and other sources to collect those information and generate a long document for me:

Robert E. Switz  
William Douglas Parker, BEC, M.B.A.  
Robert D. Isom, Jr.  
William Joseph Post, CPA  
Donald E. Brandt, CPA

# What LLM/AI agents can do now

- Deep Research/Agent Mode

## 3. Robert D. Isom Jr. – CEO of American Airlines

**Education and formative experiences.** Raised near Detroit, Robert Isom earned dual bachelor's degrees in mechanical engineering and English from the University of Notre Dame and later an MBA from the University of Michigan [en.wikipedia.org](#). This combination of technical rigor and humanities training hints at a balanced cognitive style—capable of quantitative analysis and clear communication. A Michigan Ross interview revealed that he tells aspiring leaders to sit in the front row, show energy and enthusiasm, and be someone others can rely on [michiganross.umich.edu](#); this emphasis on engagement and dependability reflects how he wants to be perceived by colleagues.

**Personal interests and endurance.** Public information about Isom's private life is scarce, but a 2022 *Hockaday Fourcast* feature recounts that he and his daughter Lola swam 1.4 miles from Alcatraz Island across San Francisco Bay and later swam under the Golden Gate Bridge  [hockadayfourcast.org](#). The article notes that this open-water challenge is a family tradition, suggesting that Isom values endurance, discipline and shared family experiences. Such athletic pursuits may translate into a willingness to tackle difficult tasks and persevere through adversity.

# Isom swims San Francisco Bay

Junior joins family tradition after intense training

by [Jane Clark](#), Editor-in-Chief

May 13, 2022

Months of cold water training pays off as Lola Isom jumps into the chilly California waters at the break of dawn, ready to carry on a years-long family tradition.

Lola Isom swam with her dad, Robert Isom, approximately 1.4 miles from Alcatraz Island into the San Francisco Bay on April 9, continuing a family tradition that her mother and siblings have completed in the past. The following day, they swam 1.2 miles underneath the Golden Gate Bridge from tower to tower, proving that her intense training and practice paid off.

“The first one I swam at around 6 a.m. so I got to watch the sunrise, which was nice,” Lola said.

Although it seems extreme, the swim does not have sharks and distance-wise many swimmers go much farther every day in practice.

There is no evidence from this article that Robert Isom (the father of Lola) is the same Robert for CEO of AA





# **Isom swims San Francisco Bay**

Junior joins family tradition after intense training

“Lola’s sister, Sophie, and I did it first,” Robert Isom said. “The next year, Lola’s brother, Henry, and Mrs. Isom did it.”

“When we lived in **Arizona**, the kids’ swim club sponsored a training event,” Robert Isom said.

can you do a deep dive into the internet

Thought for 2m 54s

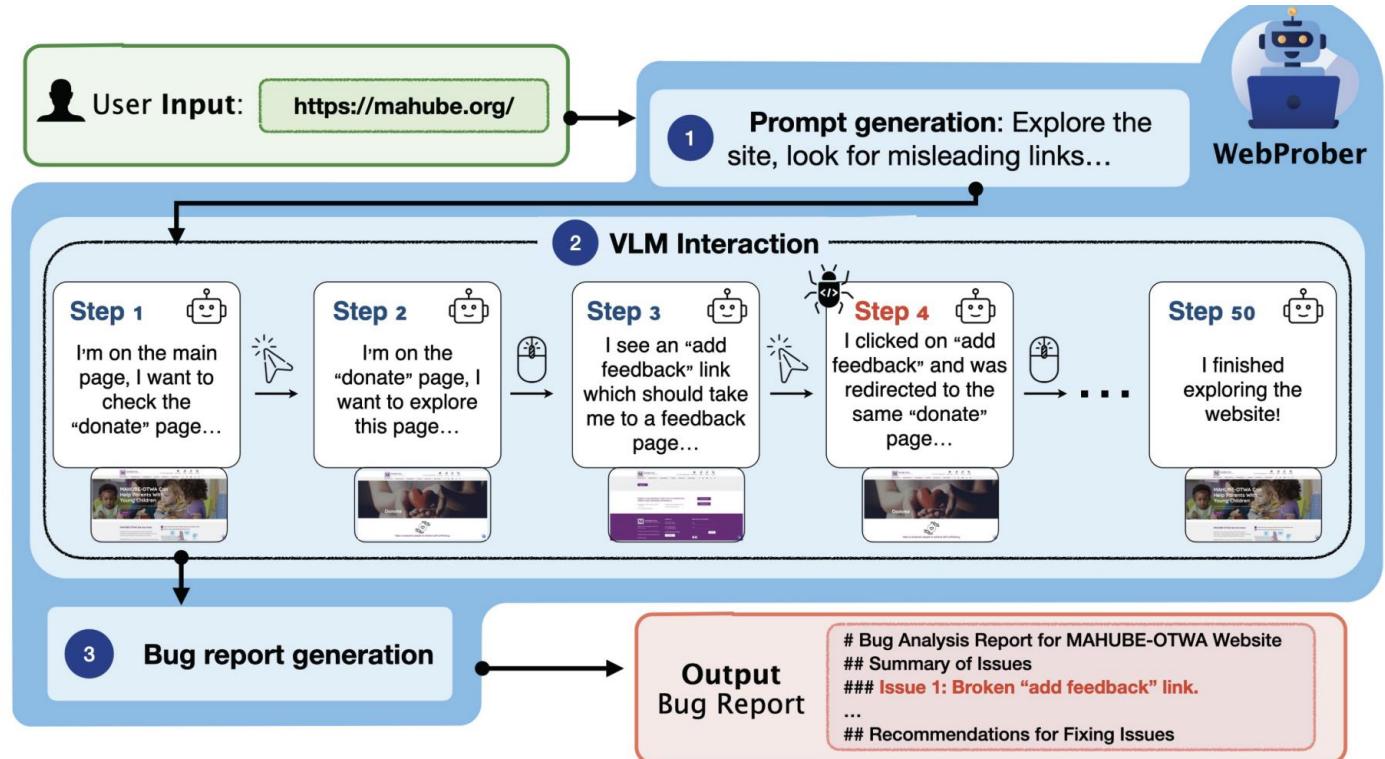
Short answer: it's very likely the **same Robert Isom** (the American Airlines CEO).

Why I say that (public, cross-checked breadcrumbs):

- The article you shared is from **The Hockaday School's student paper** and names **Lola Isom** and her dad **Robert Isom**. The Fourcast
- Hockaday's donor roll lists "**Amy and Robert Isom.**" The AA CEO's spouse is **Amy** (appears in his public posts/bios). Same couple name, same city (Dallas area). hockaday.org Instagram
- A 2007 SEC press release about Robert D. Isom (then at US Airways) says he and **his wife Amy** have **three children**—which matches the Hockaday article's references to Lola's **siblings Sophie and Henry**. SEC The Fourcast
- Independent school/athletics pages show **Lola Isom** (Hockaday '23, Yale track) and **Sophie Isom** (Hockaday '18, Yale/Michigan track), and **Henry Isom** (ESD soccer)—all Dallas-based. This aligns with the CEO's DFW base. Yale University +1 Hudl
- The American Airlines bio confirms **Robert D. Isom** as CEO since March 2022—the same timeframe as the May 2022 Hockaday piece. AA

# What LLM/AI agents can do now

- Web/App Testing
  - OpenAI Agent
  - Claude Computer-Use
  - Browser-Use
  - Manus
  - ...



Ye, Naimeng, Xiao Yu, Ruize Xu, Tianyi Peng, and Zhou Yu. "AI Agents for Web Testing: A Case Study in the Wild." (2025).

# What LLM/AI agents can do now

- Web/App Testing
  - OpenAI Agent
  - Claude Computer-Use
  - Browser-Use
  - Manus
  - ...

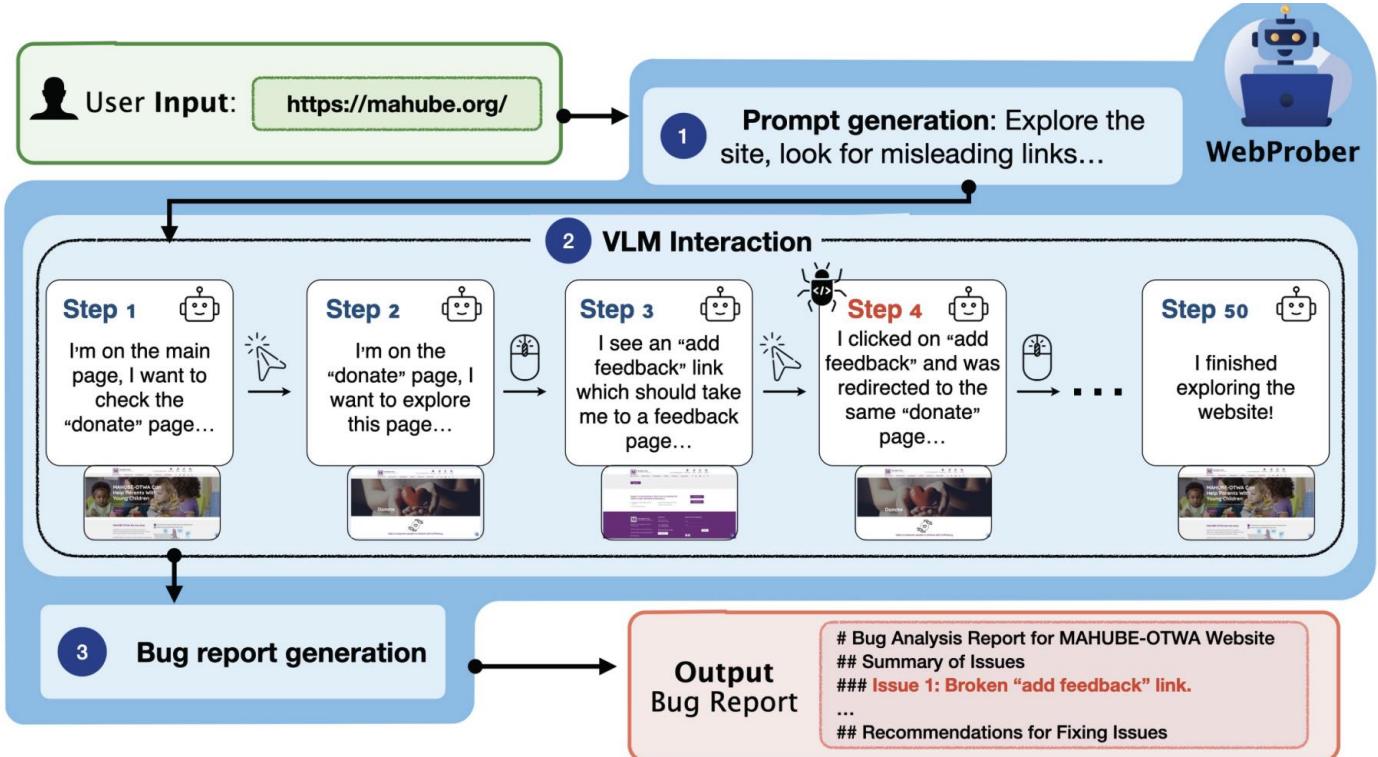
Detect 29 usability issues for 120 academic personal websites

introduce a method to forecast potential risks across orders of magnitude test during evaluation. We make forecasts by studying each query's elicitation probability the query produces and demonstrate that elicitation probability can predict the emergence of chemical synthesis query volume. Our before they manifest such as a up to three actively anti

**Agent: Link directs to the wrong paper!**

[Read more here](#)

(a) Misdirected link



Ye, Naimeng, Xiao Yu, Ruize Xu, Tianyi Peng, and Zhou Yu. "AI Agents for Web Testing: A Case Study in the Wild." (2025).

# Many other business/social use-cases

License: CC BY 4.0  
arXiv:2507.21983v1 [cs.LG] 29 Jul 2025

## Improving Generative Ad Text on Facebook using Reinforcement Learning

Daniel R. Jiang<sup>†,\*</sup>, Alex Nikulkov<sup>†</sup>, Yu-Chia Chen, Yang Bai,  
Zheqing Zhu  
Meta Platforms, Menlo Park, California, USA.

[Submitted on 4 Aug 2025]

## What Is Your AI Agent Buying? Evaluation, Implications and Emerging Questions for Agentic E-Commerce

Amine Allouah, Omar Besbes, Josué D Figueroa, Yash Kanoria,  
Akshit Kumar



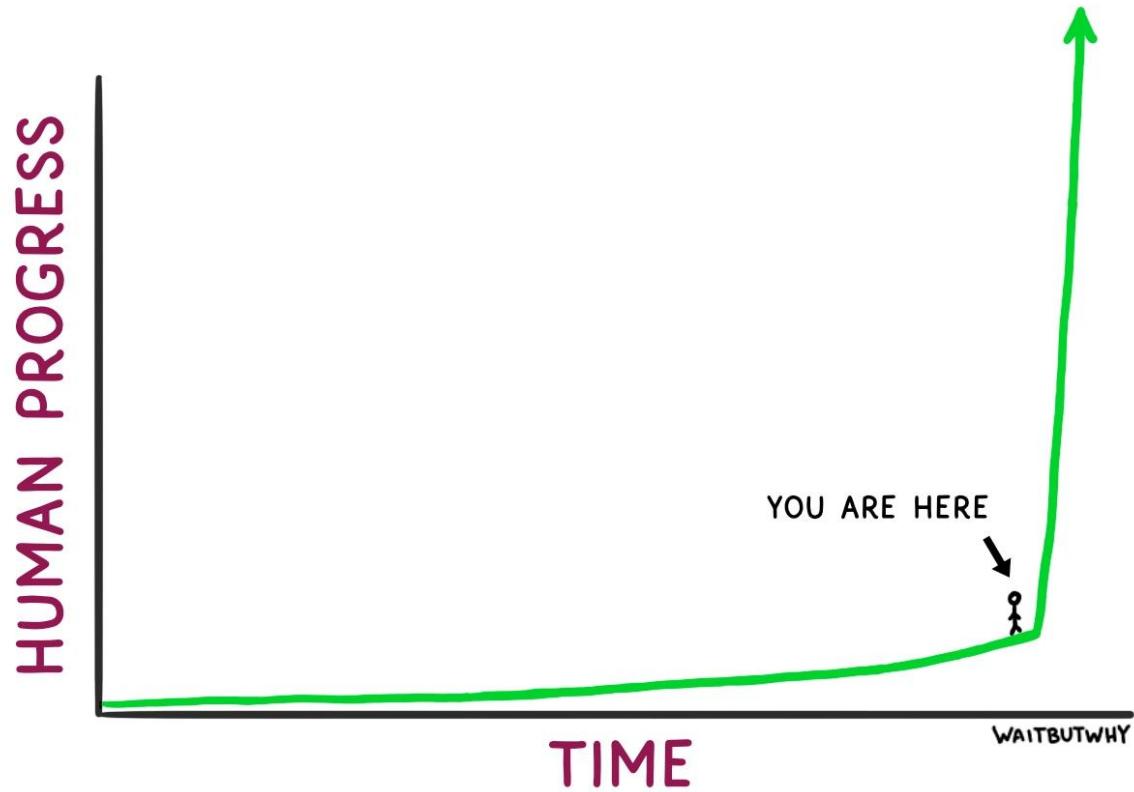
The Wall Street Journal  
<https://www.wsj.com/articles/grindr-aims-to-build-th...> ::

### Grindr Aims to Build the Dating World's First AI 'Wingman'

Oct 5, 2024 — The dating app is testing an AI bot for gay and bi men that would scout for long-term relationship prospects, set up dates—and even date ...

What is your use-case?

# How we are getting here and where we are heading to?



“We are past the event horizon  
... the takeoff has started.”

Sam Altman, June 2025

# Brief intro to AI history

Disclaimer: this lecture is developed based on MIT Hands-on Deep Learning course

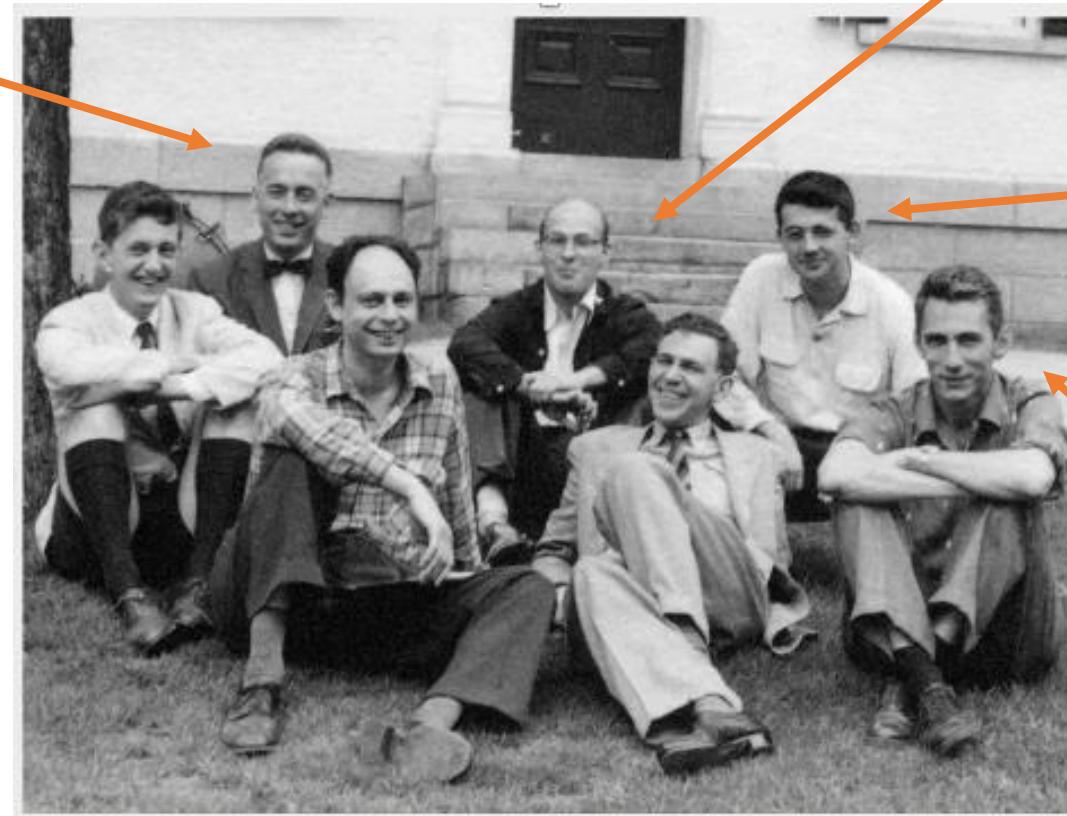
# The “official” beginning of AI



The field of Artificial Intelligence originated in 1956  
<https://spectrum.ieee.org/dartmouth-ai-workshop>

# The “official” beginning of AI

Nathaniel Rochester  
Chief architect of IBM 701  
Wrote the first assembler



Marvin Minsky  
1969 Turing Award  
Co-founded CSAIL at MIT

John McCarthy  
1971 Turing Award for “AI”  
Inventor of Lisp

Claude Shannon  
“father of information theory”

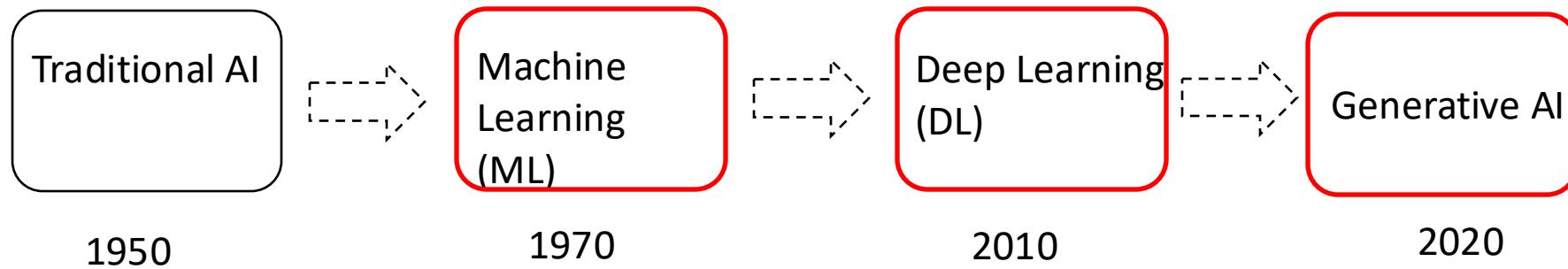
The field of Artificial Intelligence originated in 1956  
<https://spectrum.ieee.org/dartmouth-ai-workshop>

# The “official” beginning of AI

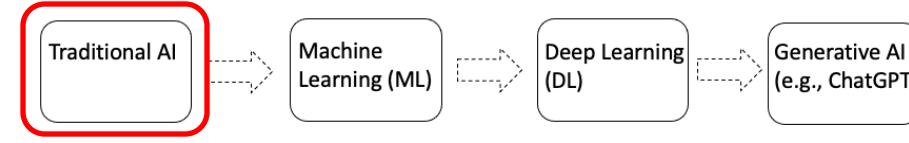
- “*We propose that a 2-month, 10-man study of **artificial intelligence** be carried out during the summer of 1956 at Dartmouth College... An attempt will be made to find how to make machines use language, from abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. 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 proposal of 1956 AI Workshop

# Today is a result of several “breakthroughs”



# The traditional approach to AI



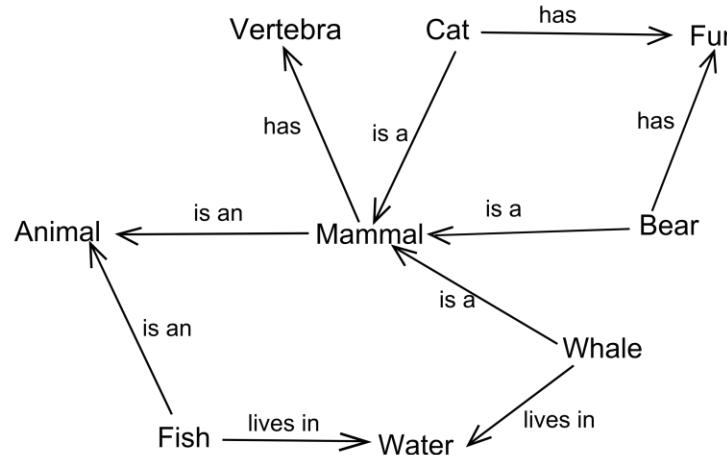
*The Goal: make machines solve problems now reserved for humans, and improve themselves*

Traditional approach:

# The traditional approach to AI



*The Goal: make machines solve problems now reserved for humans, and improve themselves*



Rule-based semantic network

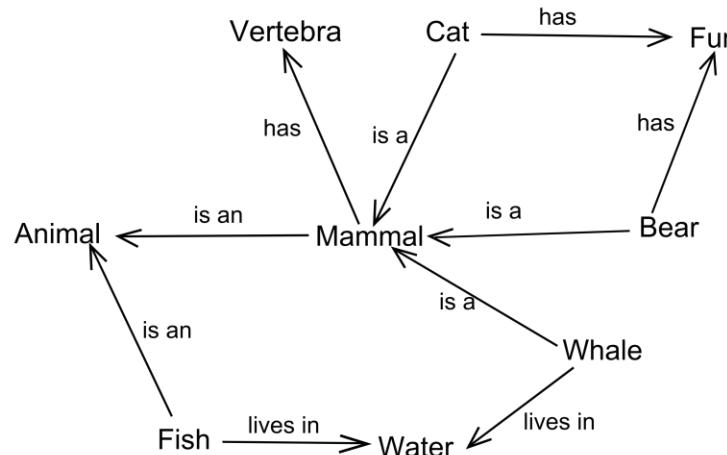
Traditional approach:

Ask human experts how they do it, write it down as IF-THEN rules, explicitly program these rules into the computer

# The traditional approach to AI

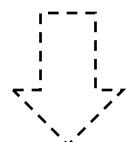


*The Goal:* make machines solve problems now reserved for humans, and improve themselves



Rule-based semantic network

*Traditional approach:*  
Ask human experts how they do it, write it down as IF-THEN rules, explicitly program these rules into the computer



Success in only a few areas

# Why is this so difficult?

- “We know more than we can tell” (Polanyi’s Paradox)
  - We can do lots of things easily but find it very hard to describe how exactly we do them, e.g., driving cars, face recognition, etc..
- We can’t write down if-then rules to cover all situations, edge cases etc. (i.e., we can’t generalize to new situations)

To address this problem, a different approach was developed

Instead of explicitly telling the computer  
what to do ...

# To address this problem, a different approach was developed

## Data Driven!

Instead of explicitly telling the computer what to do ...

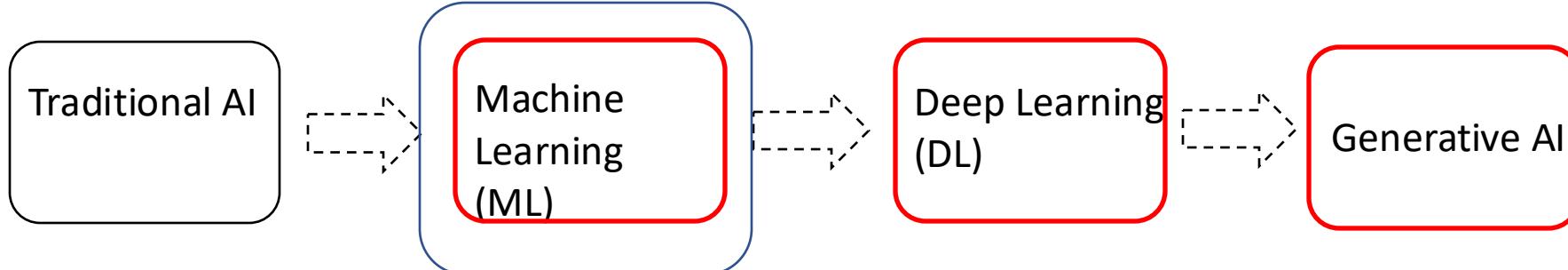
*Provide the computer with lots of examples of inputs-and-outputs and use statistical techniques to learn the relationship between inputs and outputs*

# To address this problem, a different approach was developed

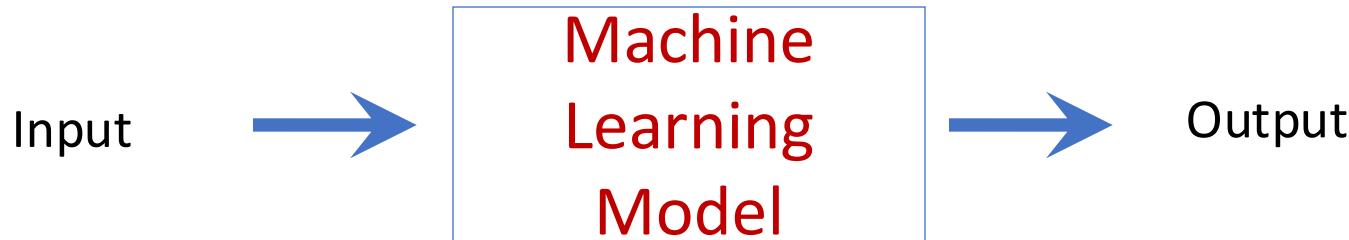
## Data Driven!

Instead of explicitly telling the computer what to do ...

*Provide the computer with lots of examples of inputs-and-outputs and use statistical techniques to learn the relationship between inputs and outputs*



# There are numerous ways\* to create Machine Learning models



- Linear Regression
- Logistic Regression
- Classification and Regression Trees
- Support Vector Machines
- Random Forests
- Gradient Boosted Machines
- Neural Networks
- .....

Data + Right Model

Machine Learning has had tremendous impact and is used worldwide across numerous applications (e.g., credit scoring, loan granting, disease prediction, demand forecasting, ....) where the input data is structured

**Structured** input data = data that can be “numericalized” into a spreadsheet\*

INPUT						OUTPUT
Age	Smoker	Exercise	Cholesterol	Family History	Blood Pressure	Cardiac Arrest
30	No	120	190	Yes	120/80	No
45	Yes	30	220	No	130/90	Yes
50	No	60	210	Yes	125/85	No
35	Yes	45	230	No	135/88	Yes
40	No	150	180	Yes	118/78	No
55	Yes	10	240	Yes	140/92	Yes
28	No	180	170	No	115/75	No
60	Yes	20	250	Yes	145/95	Yes
48	No	90	200	No	128/82	No
53	Yes	35	235	Yes	133/89	Yes

\*informal definition

But the situation is different for unstructured input data (images, videos, text, audio, ...)

Images



Text

*Four score and seven years  
ago our fathers brought forth,  
upon this continent, ...*

Audio

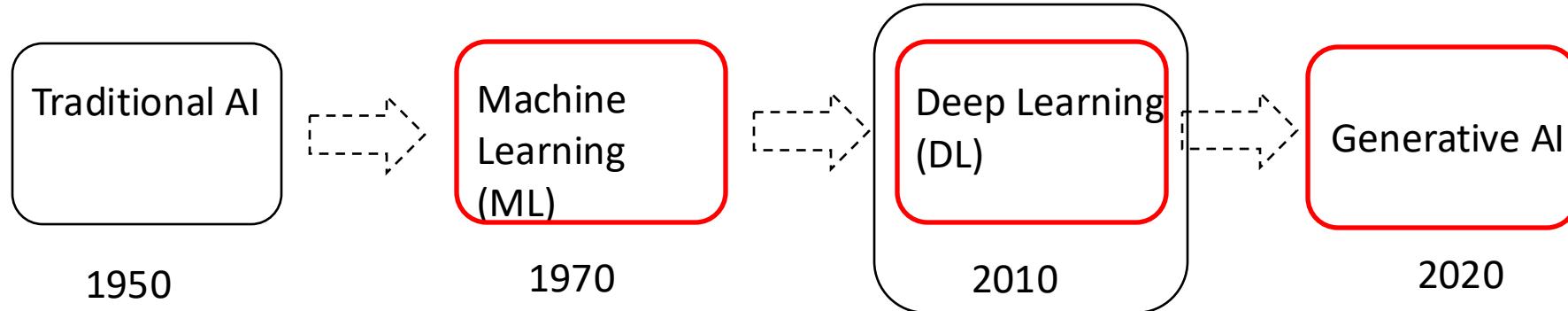
...

# The reason: The “raw form” of unstructured data has no intrinsic meaning



	Red											Green											Blue										
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]		[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]		[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	
[1,]	147	131	138	144	131	134	144	135	133	145		186	171	179	185	171	172	180	171	168	180		251	232	233	237	230	243	255	255	250	246	
[2,]	140	131	141	149	138	138	143	132	136	146		177	169	180	188	176	175	178	167	169	180		248	234	239	245	238	246	255	251	246	243	
[3,]												175	169	174	176	169	172	178	172	171	183		255	241	238	236	229	241	253	249	238	234	
[4,]												186	171	179	185	171	172	180	171	168	180		255	252	243	233	228	237	242	234	218	205	
[5,]												177	169	180	188	176	175	178	167	169	180		255	255	249	231	228	231	224	215	204	166	
[6,]												175	169	174	176	169	172	178	172	171	183		255	255	230	192	189	202	205	205	204	147	
[7,]												175	169	174	176	169	172	178	172	171	183		231	231	188	140	138	152	156	159	177	136	
[8,]												186	171	179	185	171	172	180	171	168	180		155	172	149	114	113	111	93	82	119	115	
[9,]												177	169	180	188	176	175	178	167	169	180		107	130	108	93	113	100	67	66	81	95	
[10,]												175	169	174	176	169	172	178	172	171	183		84	104	90	69	69	61	52	63	59	46	

# Next revolution: Deep Learning



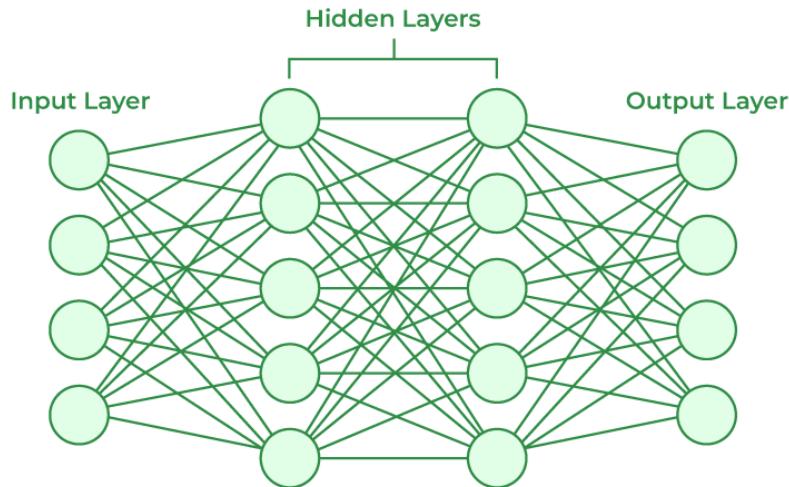
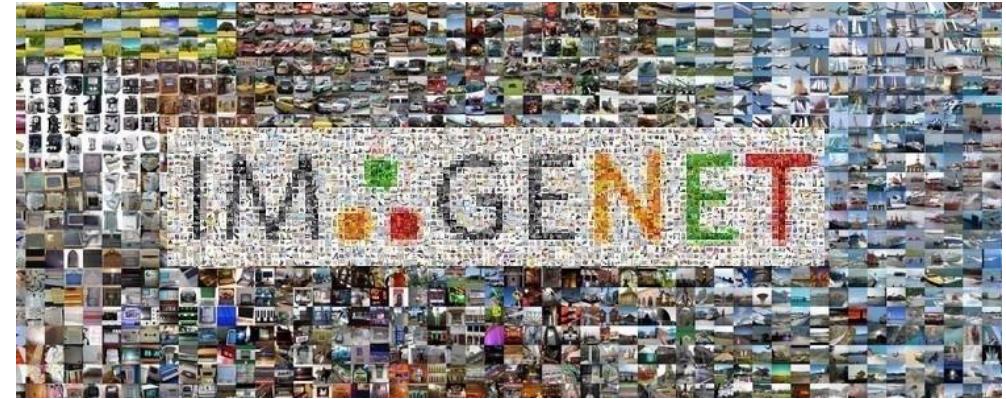
# 2009 ImageNet

- 14M images with 20K categories



# 2009 ImageNet

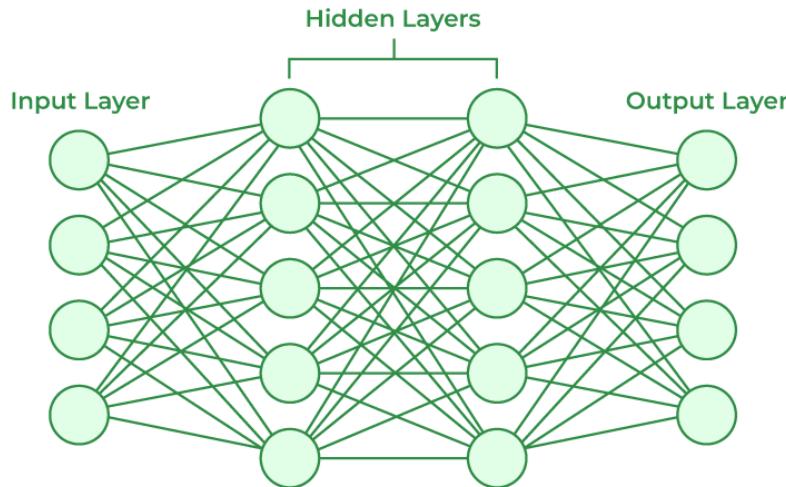
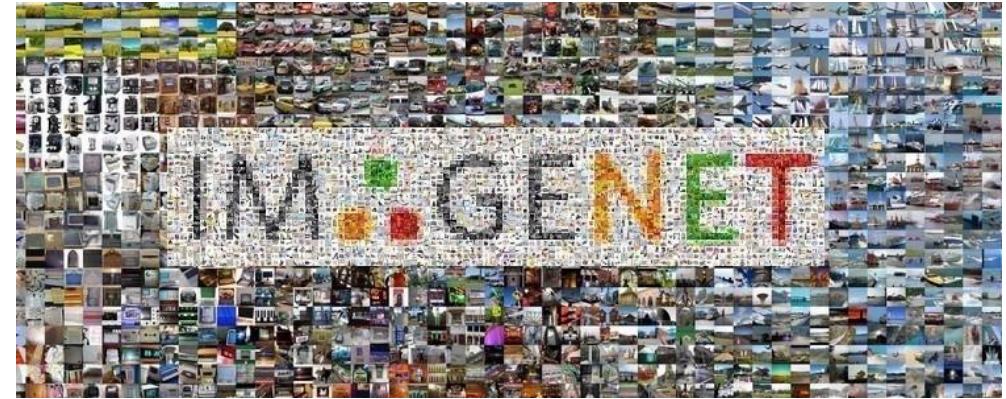
- 14M images with 20K categories



Neural Networks: GPU speedup,  
minimal human-knowledge, ..

# 2009 ImageNet

- 14M images with 20K categories

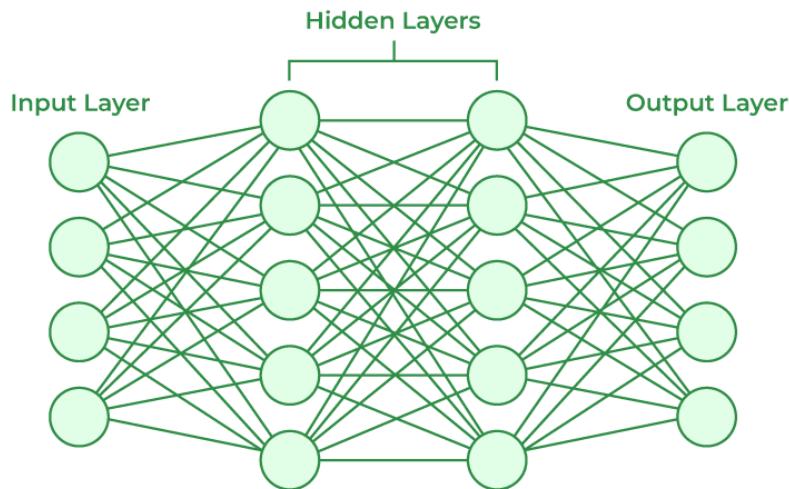
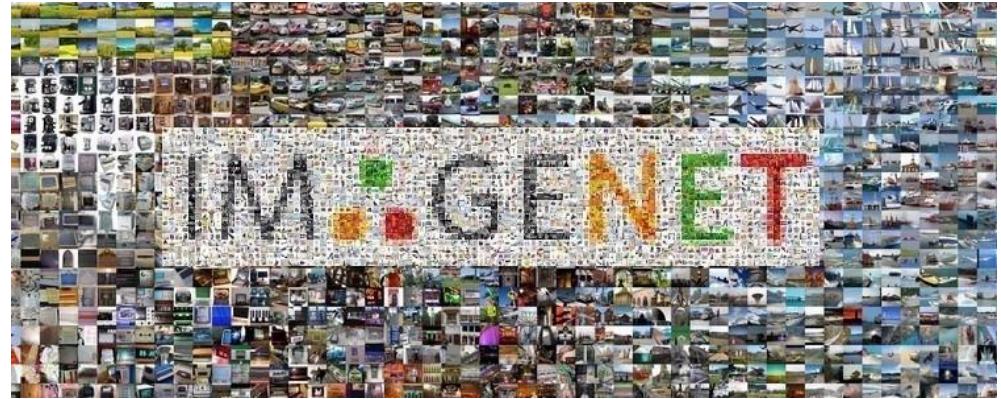


Neural Networks: GPU speedup,  
minimal human-knowledge, ..

Right Data + Right Model!

# 2009 ImageNet

- 14M images with 20K categories



Neural Networks: GPU speedup,  
minimal human-knowledge, ..

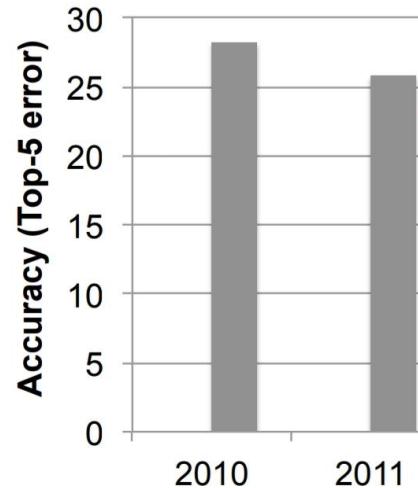
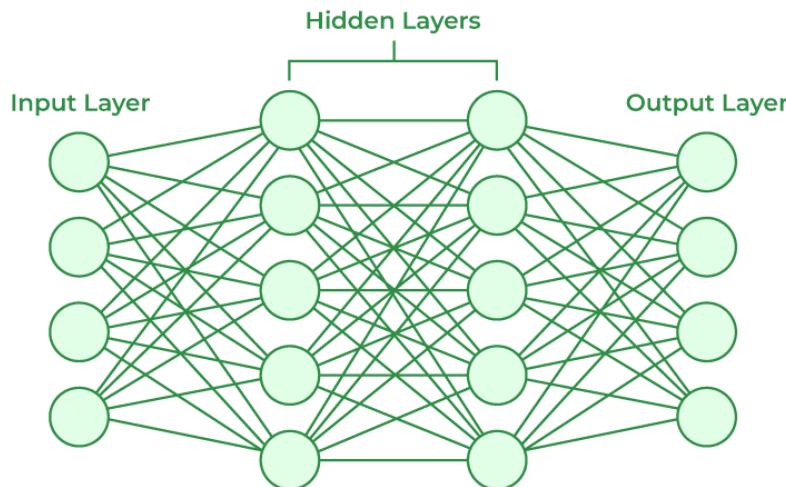
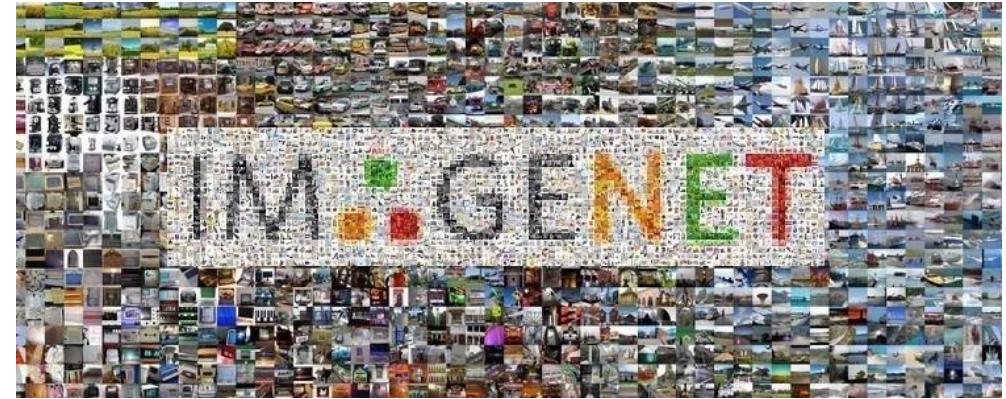


Fig. 7. Results from the ImageNet Challenge [14].

Right Data + Right Model!

# 2009 ImageNet

- 14M images with 20K categories



Neural Networks: GPU speedup,  
minimal human-knowledge, ..

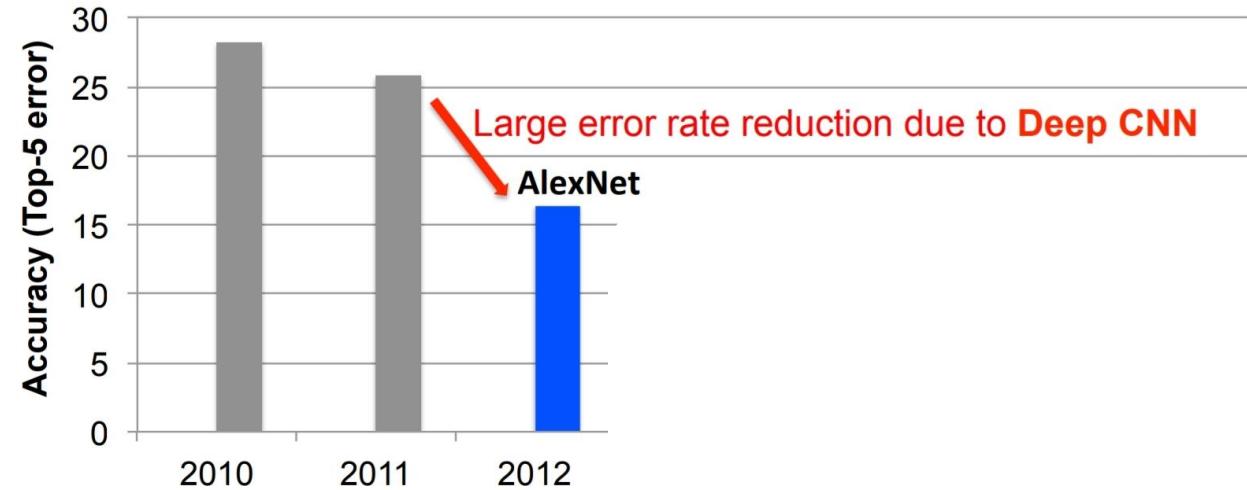
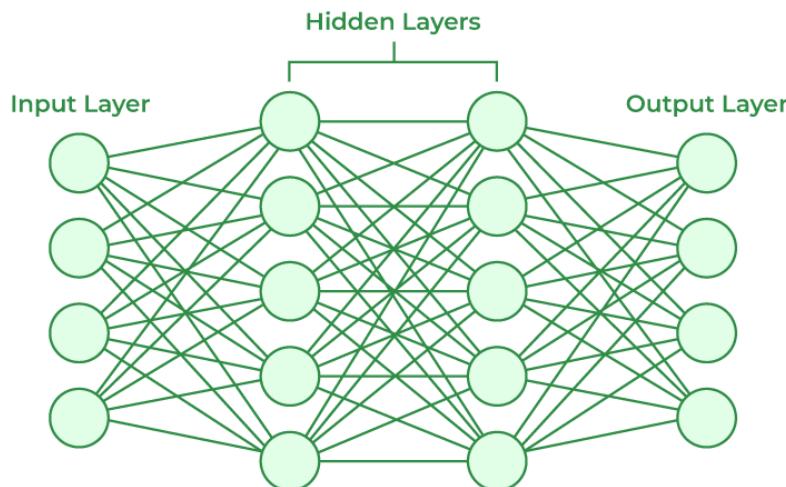
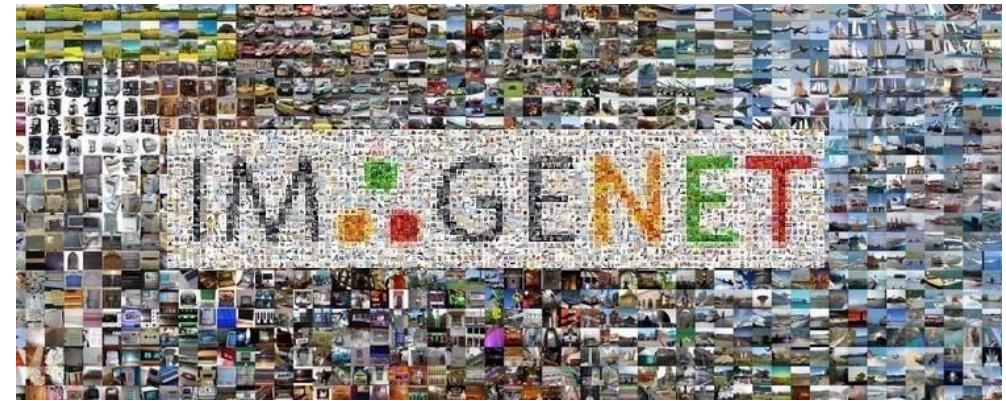


Fig. 7. Results from the ImageNet Challenge [14].

Right Data + Right Model!

# 2009 ImageNet

- 14M images with 20K categories



Neural Networks: GPU speedup,  
minimal human-knowledge, ..

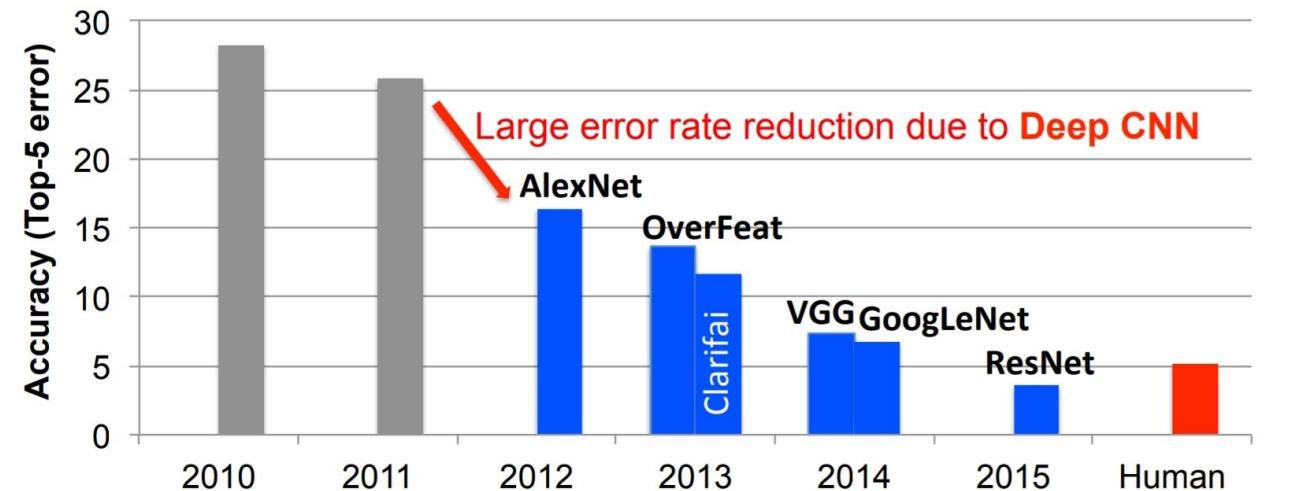


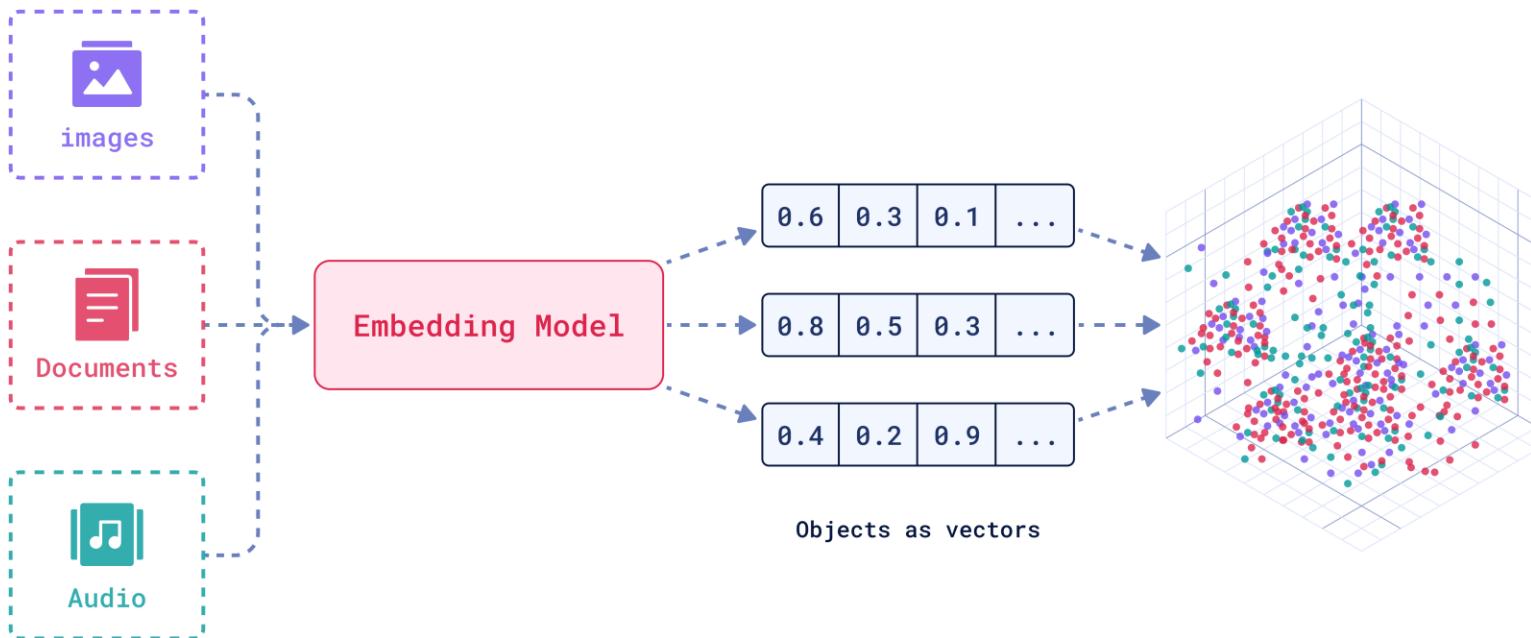
Fig. 7. Results from the ImageNet Challenge [14].

ResNet: 152 layers

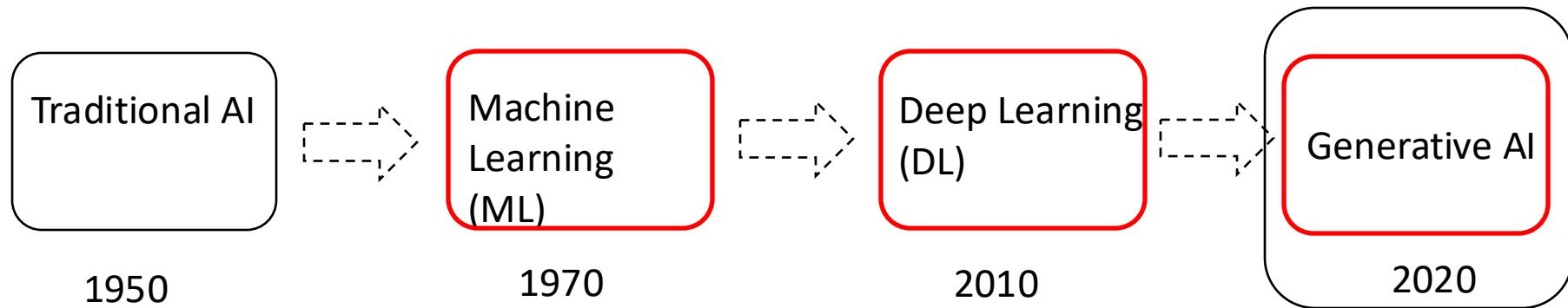
Right Data + Right Model!

# What can Deep Learning do that traditional Machine Learning can't?

- It can automatically extract smart representations from raw, unstructured data.

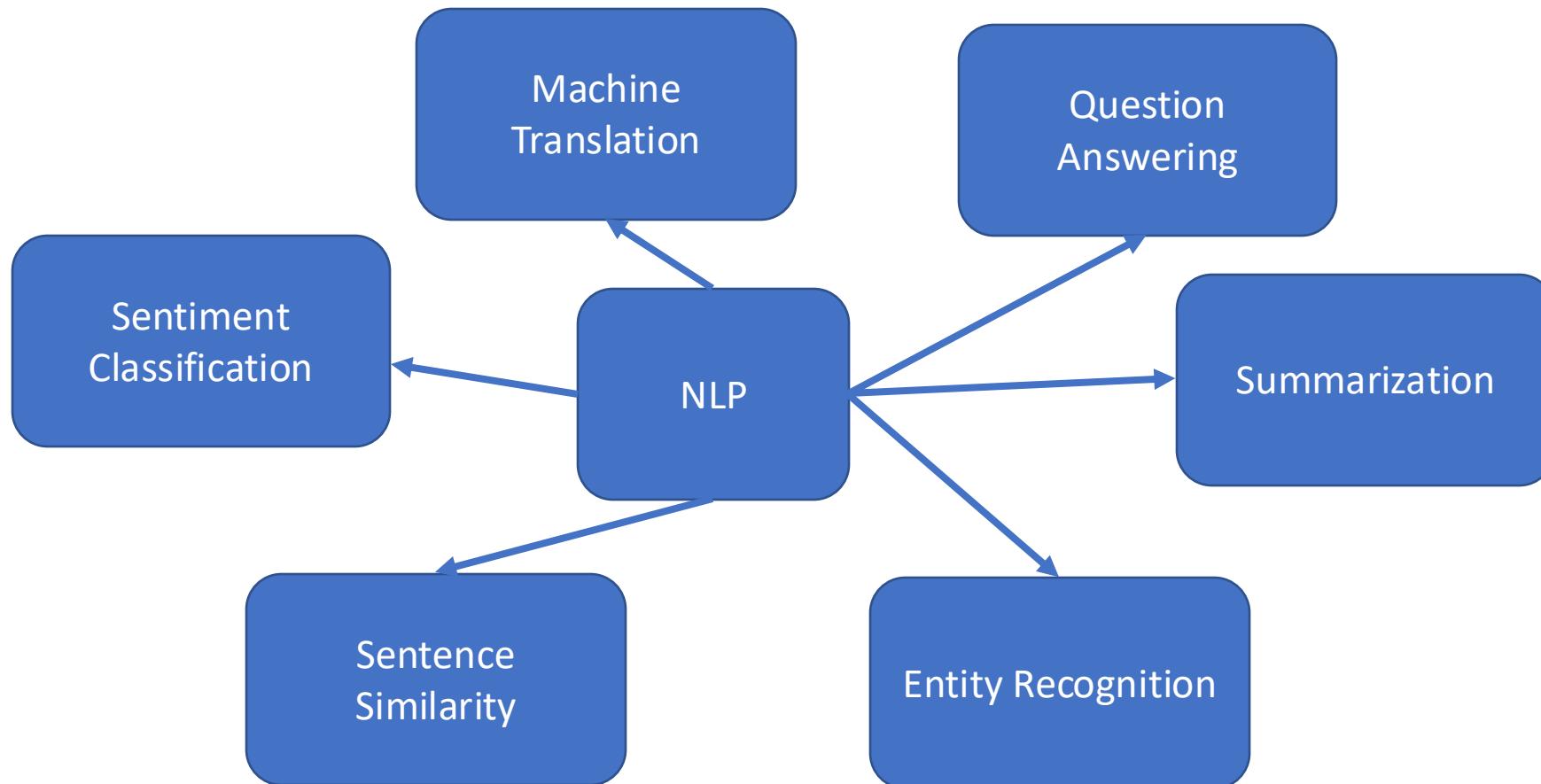


# Next revolution: Gen-AI



# Natural Language Processing (NLP)

- NLP was a bit struggling due to a diverse set of tasks



# Natural Language Processing (NLP)

- A few landmark papers 2017-2018 (Transformer, GPT1, Bert)

# Natural Language Processing (NLP)

- A few landmark papers 2017-2018 (Transformer, GPT1, Bert)
- Observation 1: Pre-training on a large corpus of unlabeled data with a common objective: predict the next/missing token!

# Natural Language Processing (NLP)

- A few landmark papers 2017-2018 (Transformer, GPT1, Bert)
- Observation 1: Pre-training on a large corpus of unlabeled data with a common objective: predict the next/missing token!
- Observation 2: Transformer is the right model
  - Attention is all you need! (130K citations)

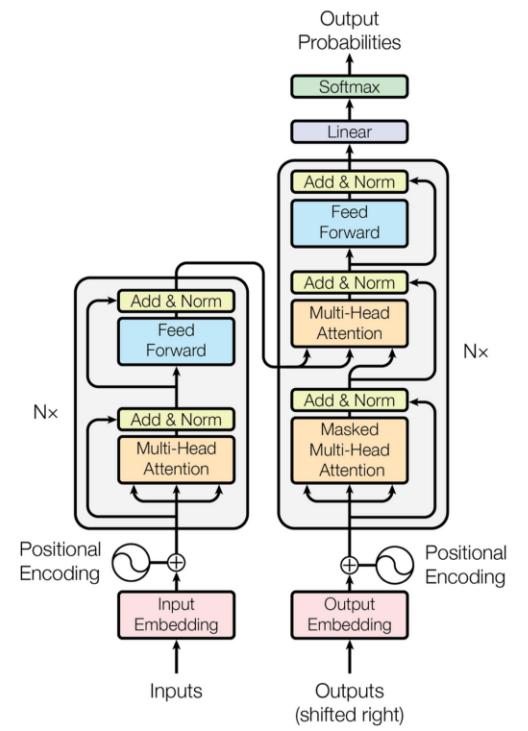


Figure 1: The Transformer - model architecture.

# Natural Language Processing (NLP)

- A few landmark papers 2017-2018 (Transformer, GPT1, Bert)
- Observation 1: Pre-training on a large corpus of unlabeled data with a common objective: predict the next/missing token!
- Observation 2: Transformer is the right model
  - Attention is all you need! (190K citations)

Right Data + Right Model!

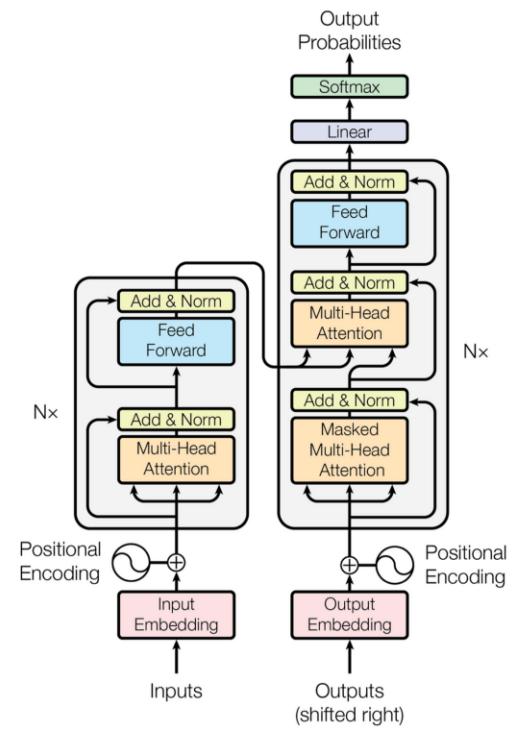


Figure 1: The Transformer - model architecture.

# Generative Pre-trained Transformers (GPT)

- The fine-tuning version significantly outperforms the SOTA

Bert: pre-trained on BookCorpus and English Wikipedia

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>92.7</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>82.1</b>

MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTS
Sentence Deduction	Question-pair similarity	Question-answering	Sentiment classification	Grammar checking	Sentence similarity	Sentence similarity	Sentence Inference

# No-finetuning! Scaling is all you need



## OpenAI's 6 Co-Founders



Elon Musk



Greg Brockman



Sam Altman



Ilya Sutskever



John Schulman



Wojciech Zaremba

	Model Size	Data Size
GPT1	117M	4.8GB (unfiltered)
GPT2	1.5B	40GB
GPT3	175B	570GB filtered from 45TB raw data
GPT4	1.8T	???
GPT5	??	???

Bigger Data + Bigger Model!

# No-finetuning! Scaling is all you need



## OpenAI's 6 Co-Founders



Elon Musk



Greg Brockman



Sam Altman



Ilya Sutskever



John Schulman



Wojciech Zaremba

	Model Size	Data Size
GPT1	117M	4.8GB (unfiltered)
GPT2	1.5B	40GB
GPT3	175B	570GB filtered from 45TB raw data
GPT4	1.8T	???
GPT5	??	???

In 2025:

Bigger data -> creating environments  
Bigger model -> more efficient way of using the parameter space

# Thinking broadly for intelligence

- Computer vision
- NLP
- Long-term planning and decision-making: AlphaGo
- Recommendation/Search Systems: DeepFMM
- Unstructured 3D prediction: AlphaFold
- ...
- One paradigm is emerging from the progress of this century..

# The law of intelligence (the ability of solving complex problems)

- If doing at scale, the intelligence can emerge from
  - Enough rich environments
  - Enough rich parameter spaces
  - A suitable objective function

# The law of intelligence

- If doing **at scale**, the intelligence can emerge from
  - Enough rich environments (nature)
  - Enough rich parameter spaces (gene)
  - A suitable objective function (survive and reproduce)
  - → evolution

# The law of intelligence

- If doing **at scale**, the intelligence can emerge from
  - Enough rich environments (data)
  - Enough rich parameter spaces (model)
  - A suitable objective function (loss function)
  - → training

Paradigm shift:

- we just need to decide the objective and feed the data
- the model will learn itself

# The law of intelligence

- If doing **at scale**, the intelligence can emerge from
  - Enough rich environments (data)
  - Enough rich parameter spaces (model)
  - A suitable objective function (loss function)
  - → training

Paradigm shift:

- we just need to decide the objective ~~and feed the data~~
- the model will learn itself

# The law of intelligence

- If doing **at scale**, the intelligence can emerge from
  - Enough rich environments (data)
  - Enough rich parameter spaces (model)
  - A suitable objective function (loss function)
  - → training
- Empirical Intelligence can be learned efficiently

Search space for AlphaGo

$10^{360}$

# The law of intelligence

- If doing **at scale**, the intelligence can emerge from
  - Enough rich environments (data)
  - Enough rich parameter spaces (model)
  - A suitable objective function (loss function)
  - → training
- Empirical Intelligence can be learned efficiently

Search space for AlphaGo       $10^{360}$

Model parameters      100M

# The law of intelligence

- If doing **at scale**, the intelligence can emerge from
  - Enough rich environments (data)
  - Enough rich parameter spaces (model)
  - A suitable objective function (loss function)
  - → training
- Empirical Intelligence can be learned efficiently
  - Nature evolves in a way that is efficient

# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

 The Economic Times

Company replaces 700  
employees with AI, two years  
later, it's rehiring humans as AI  
falls short

May 19, 2025

Klarna

# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

AI as a human/assistant is there, but AI to manage a whole system?



# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

AI as a human/assistant is there, but AI to manage a whole system?

- requires a new way of organizing information

AI agent/LLM is still slow and costly

- simulate a round of chess is taking 30min-1hour

# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

AI as a human/assistant is there, but AI to manage a whole system?

- requires a new way of organizing information

AI agent/LLM is still slow and costly

- simulate a round of chess is taking 30min-1hour

How to combine LLM with Big-data machine learning

# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

AI as a human/assistant is there, but AI to manage a whole system?

- requires a new way of organizing information

AI agent/LLM is still slow and costly

- simulate a round of chess is taking 30min-1hour

How to combine LLM with Big-data machine learning

How AI will help science: fusion, quantum physics, biology, etc..

# We are still in the early stage of AI...

Many fields are going to change, but in what forms?

- consulting, finance, healthcare, law, education, commerce, ...

AI as a human/assistant is there, but AI to manage a whole system?

- requires a new way of organizing information

AI agent/LLM is still slow and costly

We all have a chance for  
the participation!

- simulate a round of chess is taking 30min-1hour

How to combine LLM with Big-data machine learning

How AI will help science: fusion, quantum physics, biology, etc..

# Where we are heading to with AI?

- The development of AI is another anthem of human civilization.

AI的发展是人类文明的又一赞歌

# Where we are heading to with AI?

- The development of AI is another anthem of human civilization.
- Human advances by making tools and collaboration
  - AI is yet another powerful tool that will free us from labors and let us collaborate better

# Where we are heading to with AI?

- The development of AI is another anthem of human civilization.
- Human advances by making tools and collaboration.
- The law of intelligence: we are assigning values (objectives) to AI

# Where we are heading to with AI?

- The development of AI is another anthem of human civilization.
- Human advances by making tools and collaboration.
- The law of intelligence: we are assigning values (objectives) to AI
- It is not without risk.
  - The current labor economy will be disrupted.
  - Conflicts between nations may be amplified.

# Where we are heading to with AI?

- The development of AI is another anthem of human civilization.
- Human advances by making tools and collaboration.
- The law of intelligence: we are assigning values (objectives) to AI
- It is not without risk.
  - The current labor economy will be disrupted.
  - Conflicts between nations may be amplified.
- Productivity will be sufficient—but how do we ensure everyone benefits and shares the same opportunities?

# Where we are heading to with AI?

- The development of AI is another anthem of human civilization.
- Human advances by making tools and collaboration.
- The law of intelligence: we are assigning values (objectives) to AI
- It is not without risk.
  - The current labor economy will be disrupted.
  - Conflicts between nations may be amplified.
- Productivity will be sufficient—but how do we ensure everyone benefits and shares the same opportunities?
- Let's participate in creating a better world!