Machine Learning in Finance

Zichao Yang
yang_zichao@outlook.com
www.yzc.me

What is Machine Learning?

“Learning is any process by which a system improves performance from experience.” - Herbert Simon

Definition by Tom Mitchell (1998):

Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E.

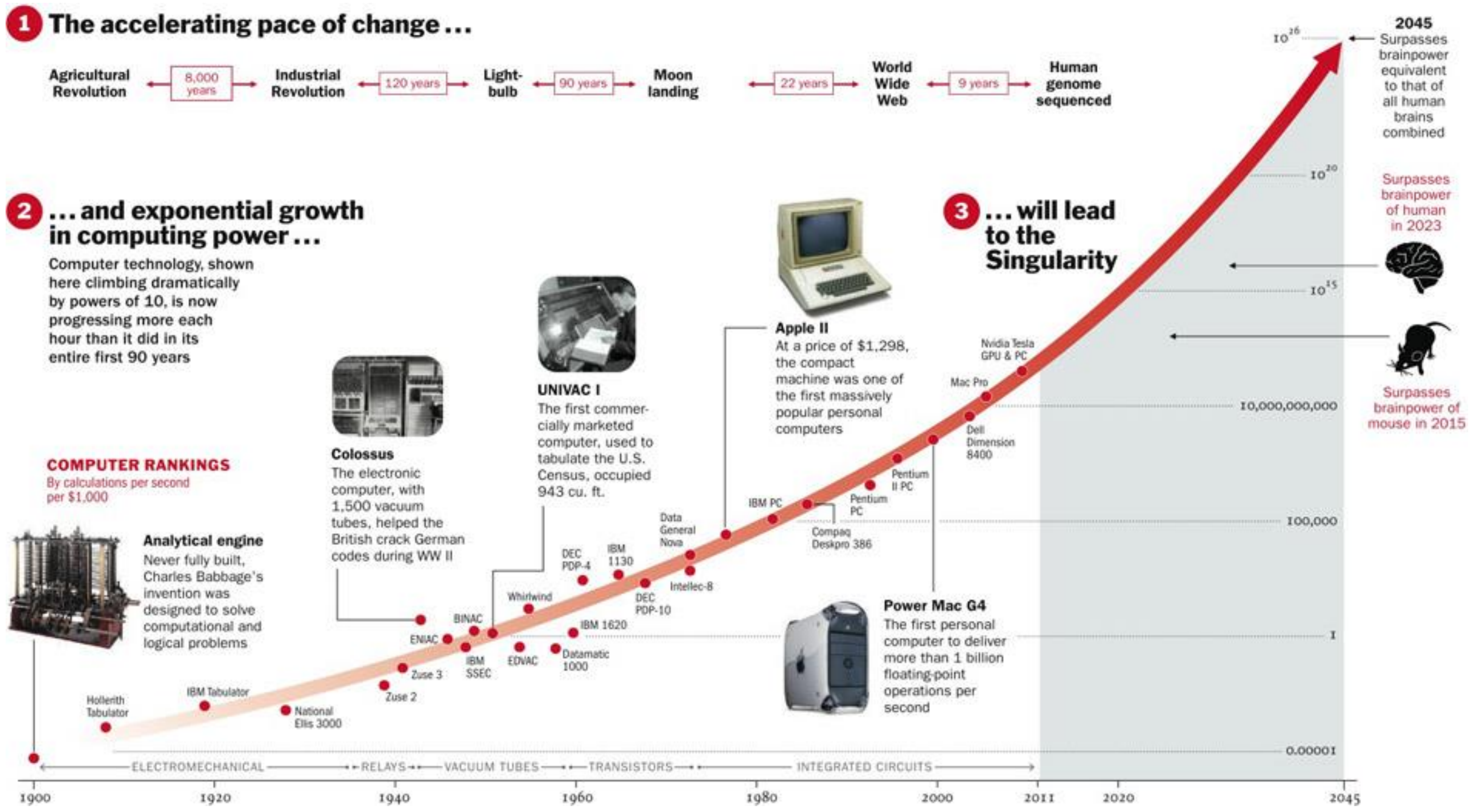
A well-defined learning task is given by <P, T, E>.

What is Machine Learning?

Computers can assist us to perform complicated tasks in two different ways:

- Knowledge-based: a computer program whose logic encodes a large number of properties of the world, usually developed by a team of experts over many years. (e.g., trading algorithm, regex)

- Learning-based: machine learning models extract information directly from historical data and extrapolate to make predictions



Source: Time Magazine

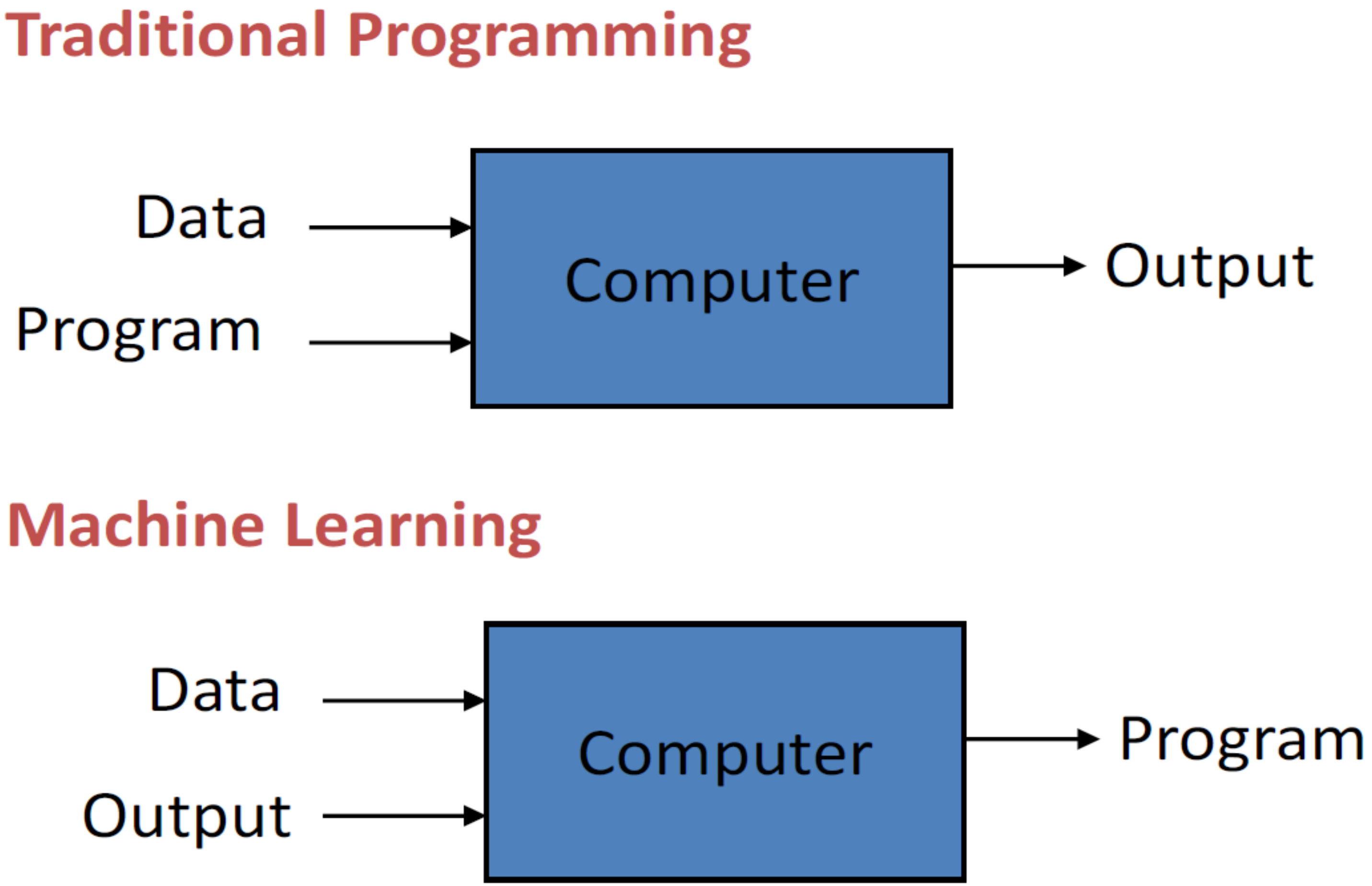
Machine Learning VS. AI

- The term “machine learning” is usually connected with “artificial intelligence (AI)”

- AI does not always imply machine learning, rule based system, tree search, or….even OLS can be called AI.

- Learning based system (machine learning model) extract information directly the data, good at solving pattern recognition problems.

Machine Learning VS. Traditional Programming



Credit: Pedro Domingos

History of Machine Learning

- 1957 - Perceptron algorithm (implemented as a circuit!)
- 1959 - Arthur Samuel wrote a learning-based checkers program that could defeat him.
- 1969 - Minsky and Papert's book Perceptrons (limitations of linear models)
- 1980s – Some foundational ideas are proposed
  - Connectionist psychologists explored neural models of cognition
  - 1984 - Leslie Valiant formalized the problem of learning as PAC learning
  - 1988 - Backpropagation (re-)discovered by Georey Hinton and colleagues
  - 1988 - Judea Pearl's book *Probabilistic Reasoning in Intelligent Systems* introduced Bayesian networks

Credit: Roger Grosse, Chris Maddison, Juhan Bae, Silviu Pitis

History of Machine Learning

- 1990s - the “AI Winter”, a time of pessimism and low funding, but looking back, the 90s were also sort of a golden age for ML research:
  - Markov chain Monte Carlo
  - variational inference
  - kernels and support vector machines
  - boosting
  - convolutional networks
  - reinforcement learning

- 2000s – applied AI fields (vision, NLP, etc.) adopted ML

Credit: Roger Grosse, Chris Maddison, Juhan Bae, Silviu Pitis

History of Machine Learning

- 2010s – deep learning
  - 2010-2012: neural nets smashed previous records in speech-to-text and object recognition, ML increasingly adopted by the tech industry
  - 2016: AlphaGo defeated the human Go champion
  - 2018-now: generating photorealistic images and videos
  - 2020: GPT3 language model
- Now - increasing attention to ethical and societal implications

Credit: Roger Grosse, Chris Maddison, Juhan Bae, Silviu Pitis

Why deep learning did not work back then?

Geoffrey Hinton summarized the findings up to today in these four points:

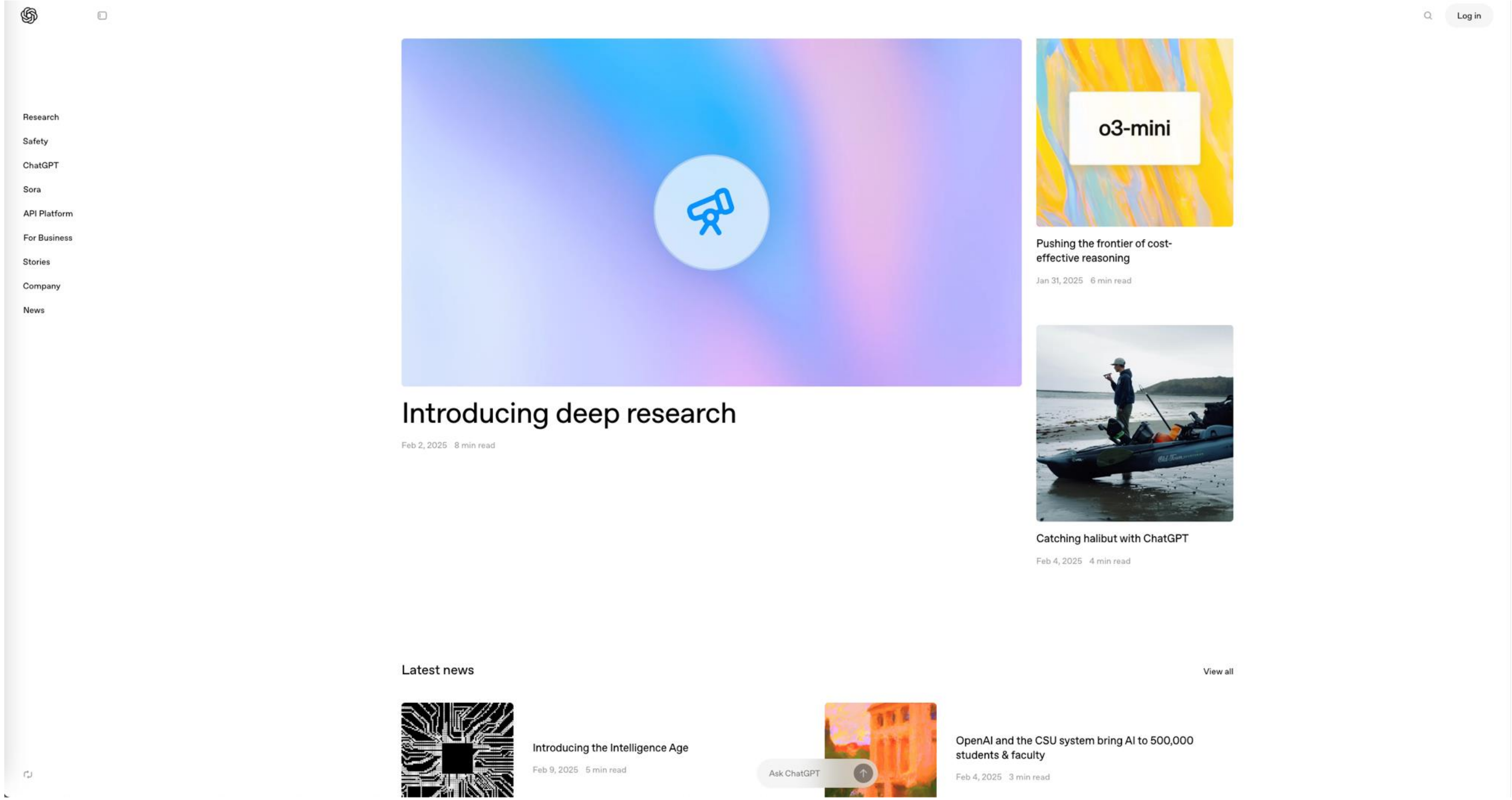
- Our labeled datasets were thousands of times too small.
- Our computers were millions of times too slow.
- We initialized the weights in a stupid way.
- We used the wrong type of non-linearity.

Current Major Players

- OpenAI
- xAI (Elon Musk)
- Alphabet (Google)
- Meta (Facebook)
- DeepSeek
- Moonshot AI
- Universities

- NVIDA (essential)

OpenAI



https://openai.com/

Google

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

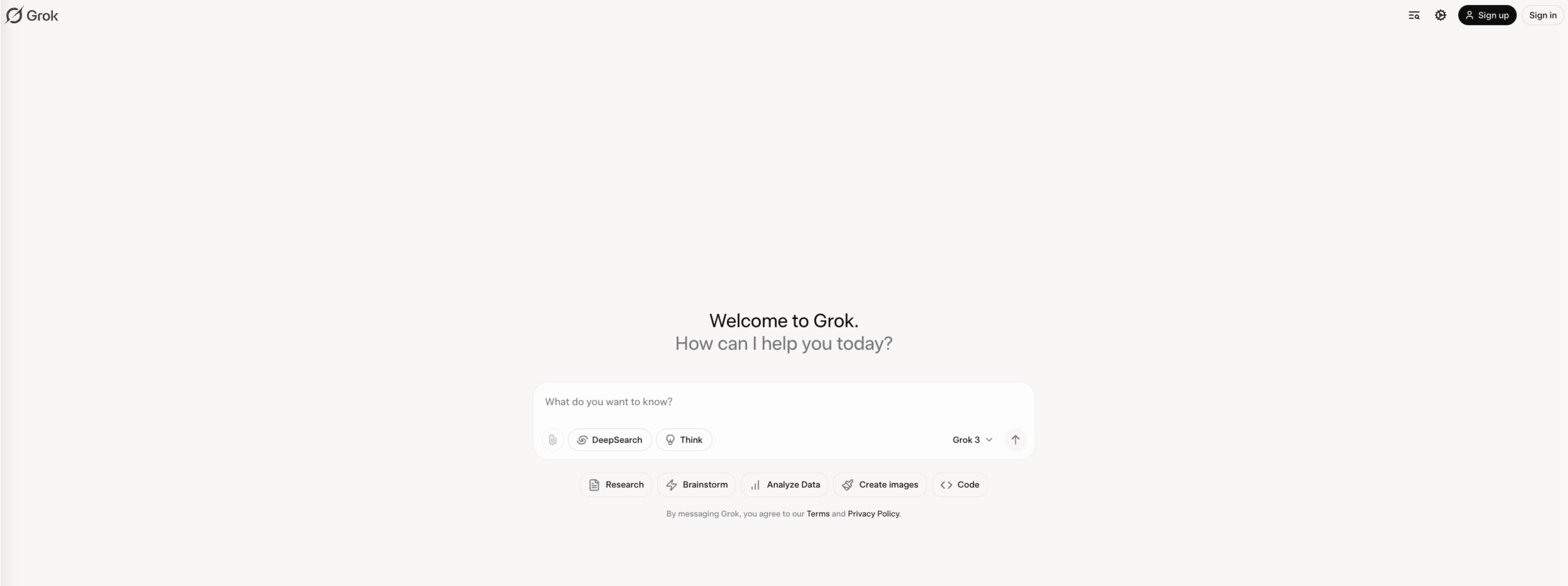
We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

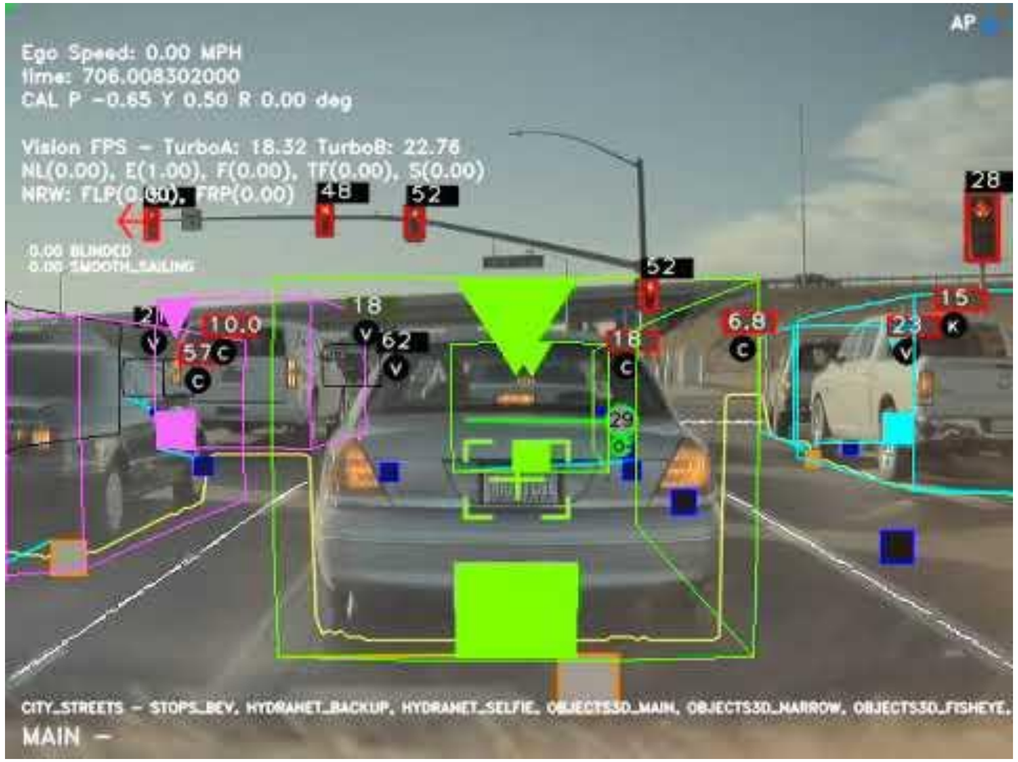
We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers

xAI

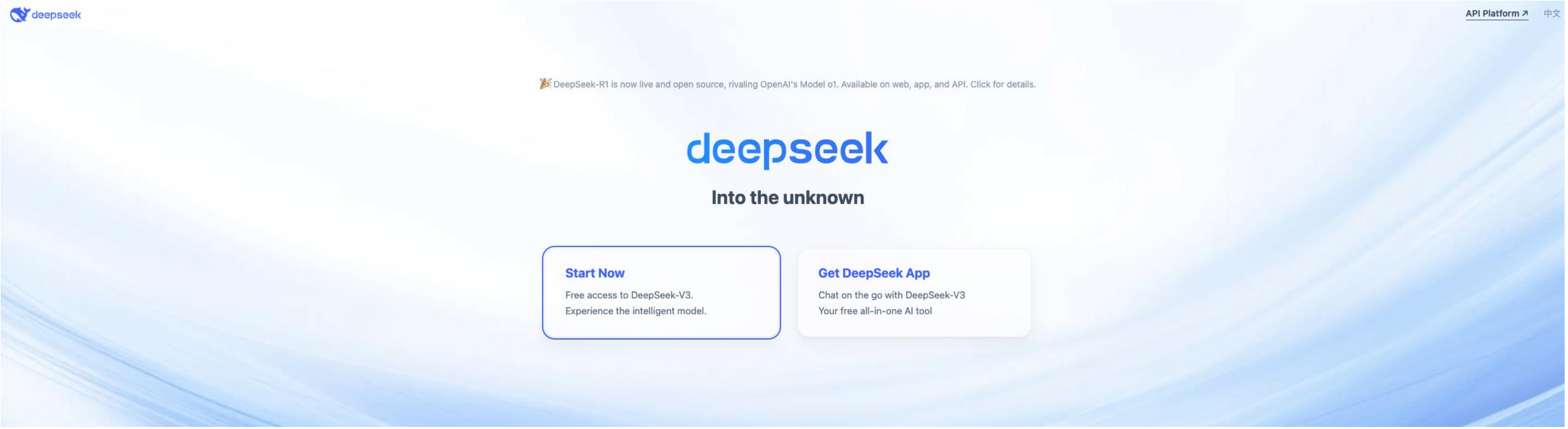


https://grok.com/

Tesla


AP
Ego Speed: 0.00 MPH
time: 706.008302000
CAL P -0.65 Y 0.50 R 0.00 deg
Vision FPS - TurboA: 18.32 TurboB: 22.76
NL(0.00), E(1.00), F(0.00), TF(0.00), S(0.00)
NRW: FLP(0.00), FRP(0.00)
0.00 BLINDED
0.00 SMOOTH_SAILING
CITY_STREETS - STOPS_BEV, HYDRANET_BACKUP, HYDRANET_SELFIE, OBJECTS3D_MAIN, OBJECTS3D_NARROW, OBJECTS3D_FISHEYE,
MAIN -

deepseek



https://www.deepseek.com/

Moonshot AI

https://www.moonshot.cn/

AI competition is fierce…

NVIDIA Ethernet Networking Accelerates World’s Largest AI Supercomputer, Built by xAI

NVIDIA Spectrum-X Makes Colossal NVIDIA Hopper 100,000-GPU System Possible

October 28, 2024

NVIDIA today announced that xAI’s Colossus supercomputer cluster comprising 100,000 NVIDIA Hopper GPUs in Memphis, Tennessee, achieved this massive scale by using the NVIDIA Spectrum-X™ Ethernet networking platform, which is designed to deliver superior performance to multi-tenant, hyperscale AI factories using standards-based Ethernet, for its Remote Direct Memory Access (RDMA) network.

Colossus, the world’s largest AI supercomputer, is being used to train xAI’s Grok family of large language models, with chatbots offered as a feature for X Premium subscribers. xAI is in the process of doubling the size of Colossus to a combined total of 200,000 NVIDIA Hopper GPUs.

The supporting facility and state-of-the-art supercomputer was built by xAI and NVIDIA in just 122 days, instead of the typical timeframe for systems of this size that can take many months to years. It took 19 days from the time the first rack rolled onto the floor until training began.

THE WALL STREET JOURNAL.

English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾

TECHNOLOGY | ARTIFICIAL INTELLIGENCE Follow

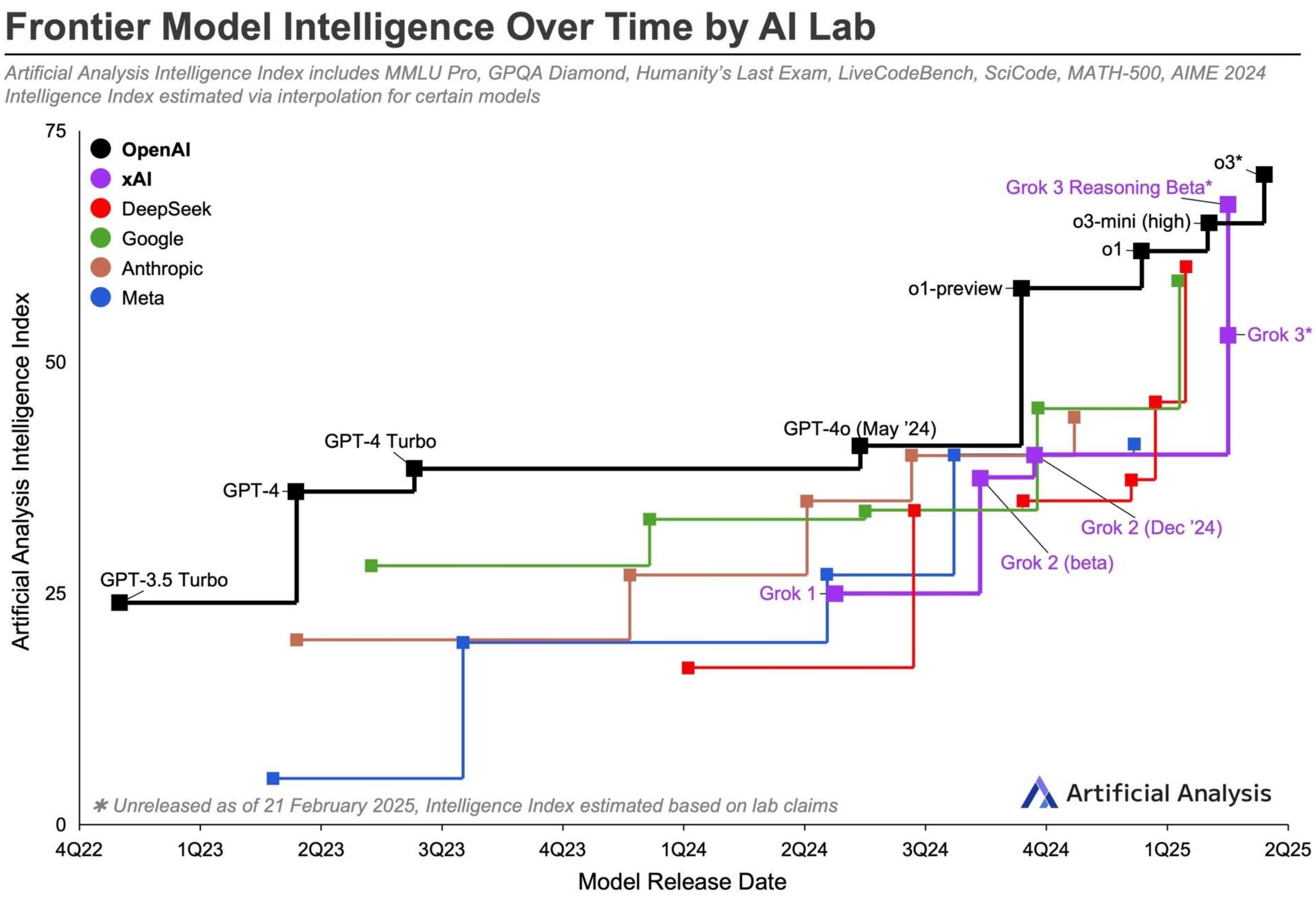
Tech Leaders Pledge Up to $500 Billion in AI Investment in U.S.

OpenAI, Oracle and SoftBank unveil AI infrastructure plans at White House

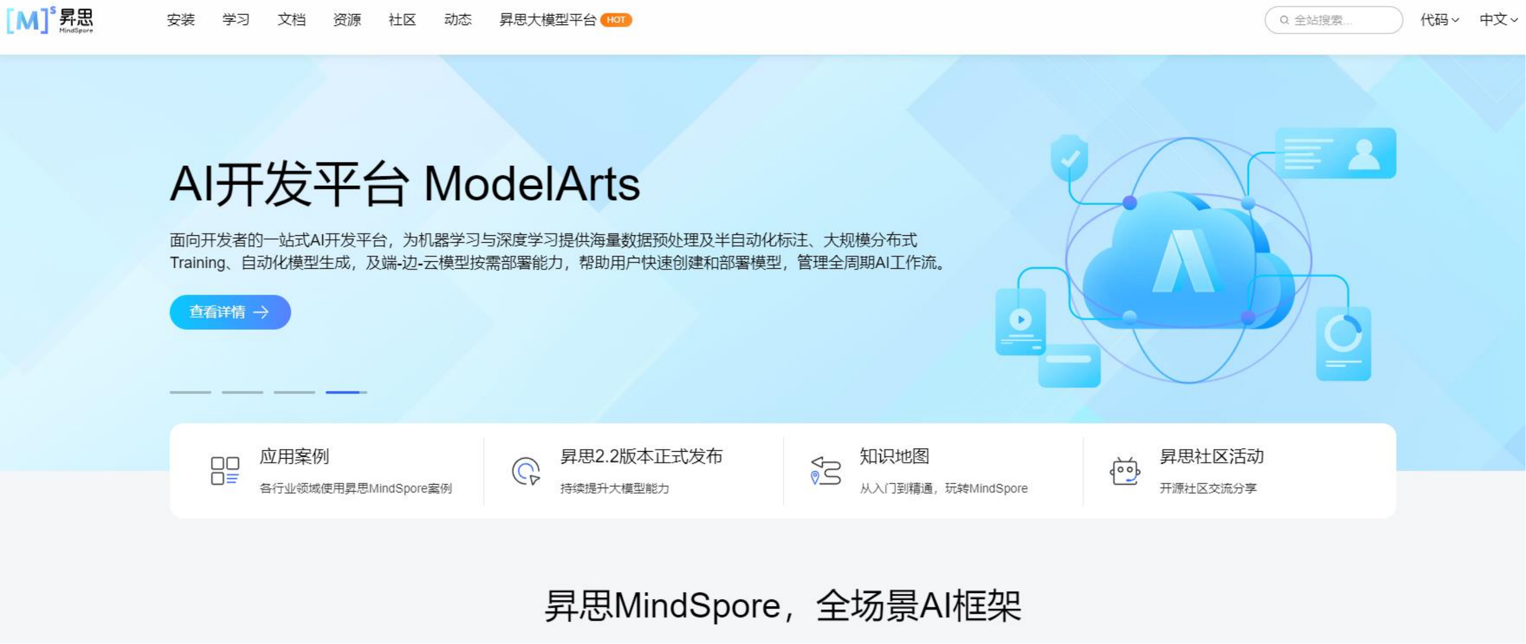
By Deepa Seetharaman Follow *and Tom Dotan* Follow

Updated Jan. 21, 2025 at 8:08 pm ET

AI competition is fierce…



NVIDIA



昇思
MindSpore
安装
学习
文档
资源
社区
动态
昇思大模型平台 HOT
代码
中文
AI开发平台 ModelArts
面向开发者的一站式AI开发平台，为机器学习与深度学习提供海量数据预处理及半自动化标注、大规模分布式Training、自动化模型生成，及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。
查看详情
应用案例
各行业领域使用昇思MindSpore案例
昇思2.2版本正式发布
持续提升大模型能力
知识地图
从入门到精通，玩转MindSpore
昇思社区活动
开源社区交流分享
昇思MindSpore，全场景AI框架

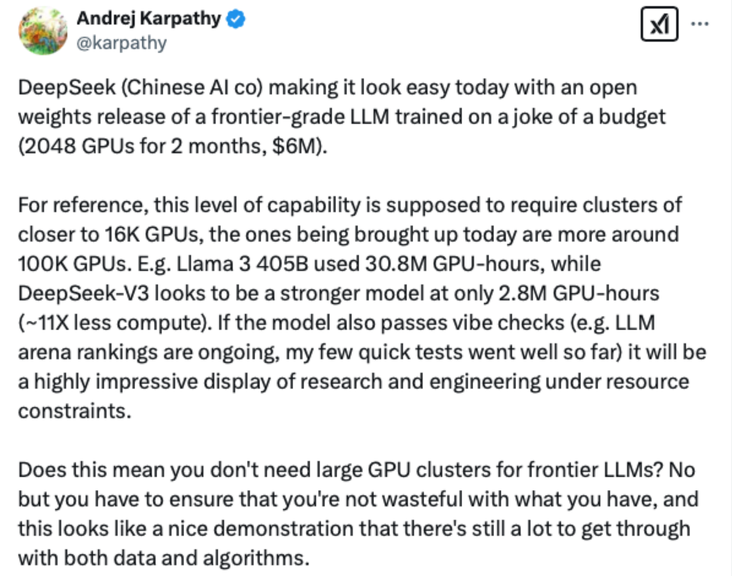
NVIDIA

Andrew Ng @AndrewYNg

Today's "DeepSeek selloff" in the stock market -- attributed to DeepSeek V3/R1 disrupting the tech ecosystem -- is another sign that the application layer is a great place to be. The foundation model layer being hyper-competitive is great for people building applications.

4:16 AM · Jan 28, 2025 · **770.1K** Views

256 1.2K 7.3K 1.2K



Andrej Karpathy @karpathy

DeepSeek (Chinese AI co) making it look easy today with an open weights release of a frontier-grade LLM trained on a joke of a budget (2048 GPUs for 2 months, $6M).

For reference, this level of capability is supposed to require clusters of closer to 16K GPUs, the ones being brought up today are more around 100K GPUs. E.g. Llama 3 405B used 30.8M GPU-hours, while DeepSeek-V3 looks to be a stronger model at only 2.8M GPU-hours (~11X less compute). If the model also passes vibe checks (e.g. LLM arena rankings are ongoing, my few quick tests went well so far) it will be a highly impressive display of research and engineering under resource constraints.

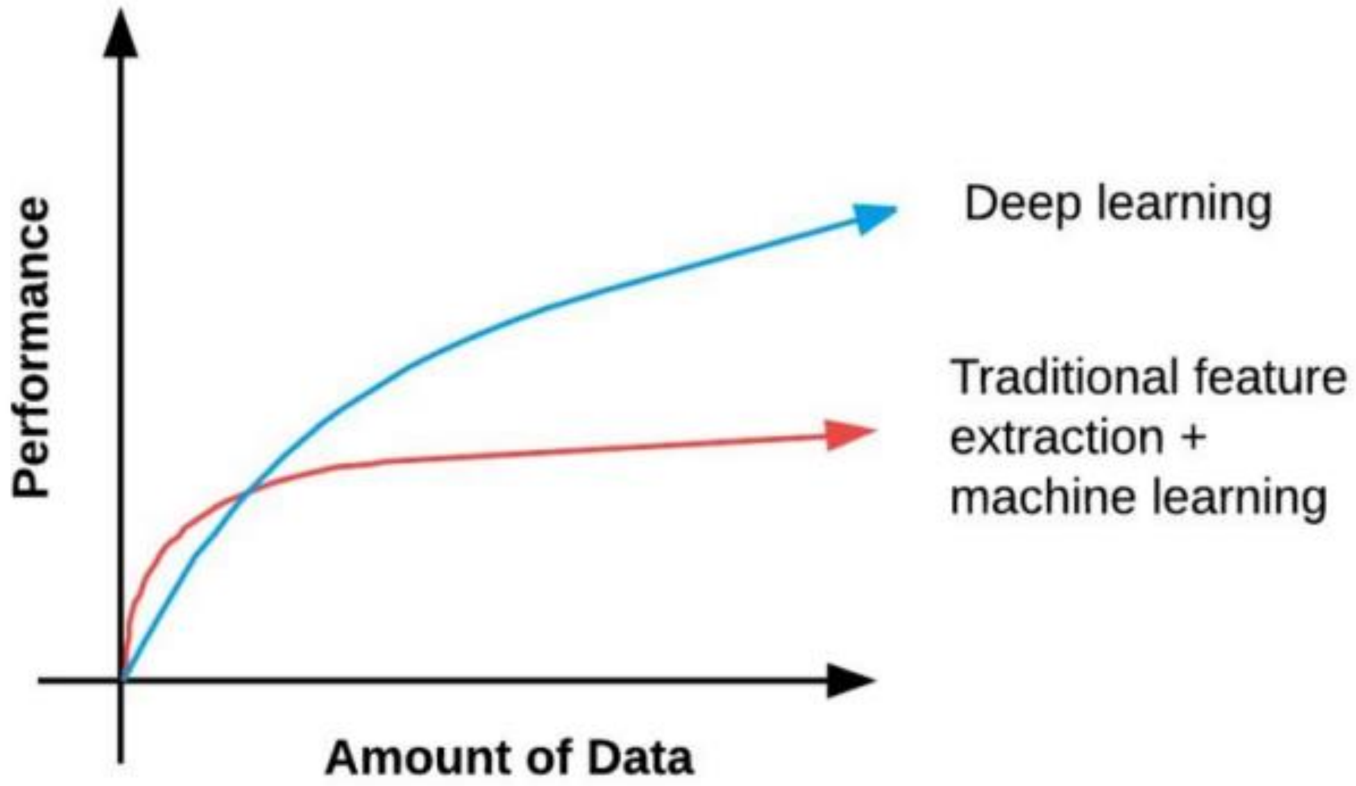
Does this mean you don't need large GPU clusters for frontier LLMs? No but you have to ensure that you're not wasteful with what you have, and this looks like a nice demonstration that there's still a lot to get through with both data and algorithms.

Types of Machine Learning

- Statistic Machine Learning Models
  - LASSO
  - SVM
  - Random Forest
  - …

- Deep Learning Models
  - MLP
  - CNN
  - RNN, LSTM (outdated)
  - Transformer
  - …

ML VS. DL



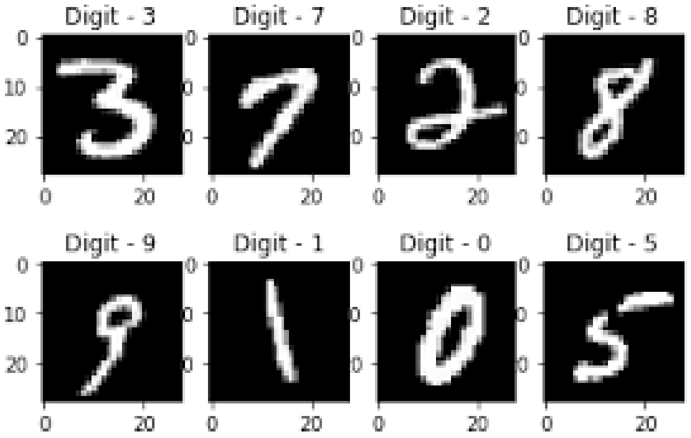
Credit: Adrian Rosebrock. 2017. Deep Learning for Computer Vision With Python. (2017)

Types of Machine Learning

- Supervised Learning
  - training data + desired outputs (labels)
- Semi-supervised Learning
  - A subset of supervised learning, training data + a few labels
- Unsupervised Learning
  - training data (without desired outputs)
- Self-supervised Learning
  - A subset of unsupervised learning, recover patterns from the data
- Reinforcement Learning
  - Rewards from sequence of actions

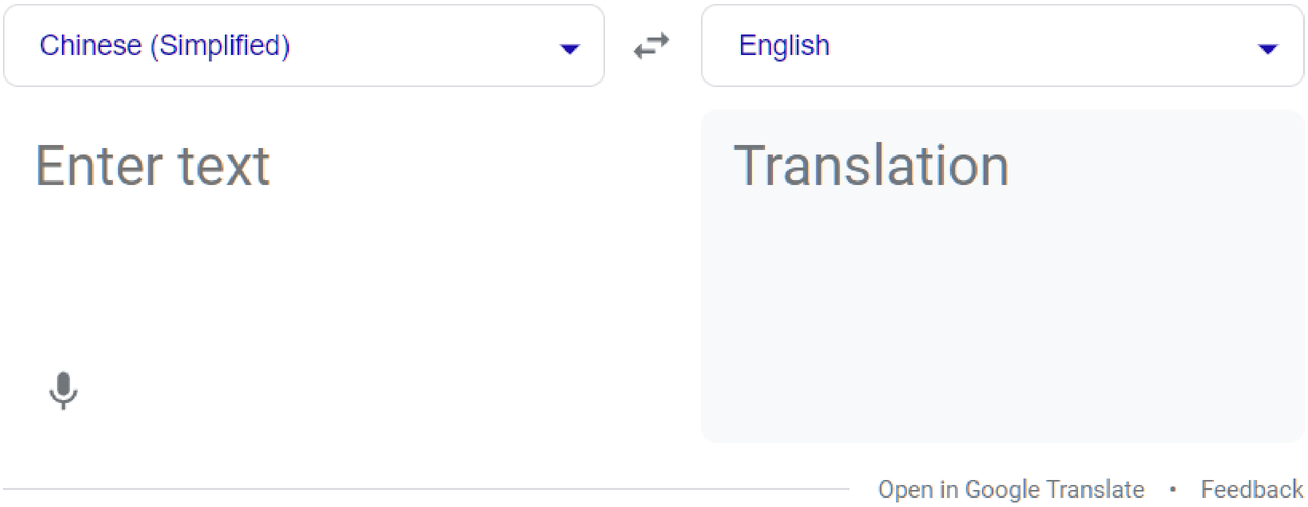
Tasks

- Computer vision: object detection, OCR, semantic segmentation…



Tasks

• NLP: Machine translation, sentiment analysis, topic modeling, spam filtering.



Back to Our Course

We follow this paper to explore the applications of various machine learning models within the finance research field:

Kelly, B., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends® in Finance*, 13(3-4), 205-363.

Download link: https://bfi.uchicago.edu/wp-content/uploads/2023/07/BFI_WP_2023-100.pdf

Reference Books

Nielsen, Michael. *Neural Networks and Deep Learning*, 2019.

周志华 .《机器学习》, 清华大学出版社, 2016.
Zhou, Zhi-Hua. *Machine learning*. Springer Nature, 2021.

Géron, Aurélien. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

Stevens, Eli, Luca Antiga, and Thomas Viehmann. *Deep learning with PyTorch*. Manning Publications, 2020.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Course Requirement

- A functional laptop that can run python
- Attend the class
- A group project, 2-3 people per group, and presentation
- Final report