# GenAI as a New Innovation Platform

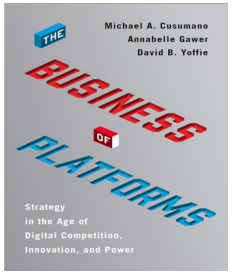
**August 11, 2025**
**MIT Sloan Executive Education**
**Endeavor Program**


## Michael A. Cusumano

**MIT Sloan School of Management**

**cusumano@mit.edu**


MIT MANAGEMENT SLOAN SCHOOL



# Michael A. Cusumano

cusumano@mit.edu

- Specializes in strategy, product development & entrepreneurship in the computer software industry, as well as automobiles and consumer electronics
- Teaches courses on *Software & Internet Entrepreneurship* and *Advanced Strategic Management.*
- 14 books, most recently Strategy Rules (2015, 18 translations) and The Business of Platforms (2019)

## Education

- B.A. degree (Princeton, 1976) and Ph.D. (Harvard, 1984)

Postdoctoral fellowship in Technology & Operations Management (Harvard Business School, 1984-86)

Two Fulbright Fellowships and Japan Foundation Fellowship

## Other Activities etc.

Consulted for some 90 companies and organizations globally

Former editor-in-chief and chairman of the *MIT Sloan Management Review;* column on Technology Strategy and Management for *Communications of the ACM*

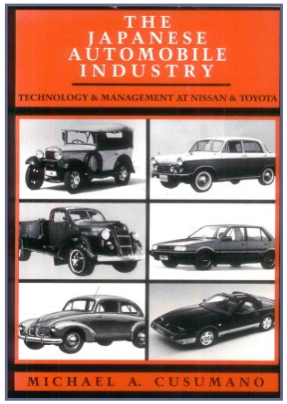
Named one of the most influential people in technology & IT by Silicon.com in 2009

**MIT Sloan School of Management**

SMR Distinguished Professor & former Deputy Dean

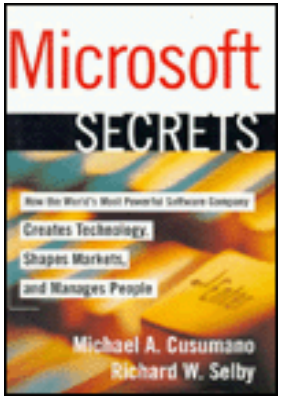
Faculty Director, M.T. Center for MIT Entrepreneurship

Co-Director, MIT System Design & Management Program (SDM)

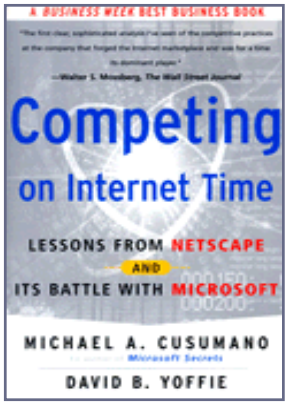



1985

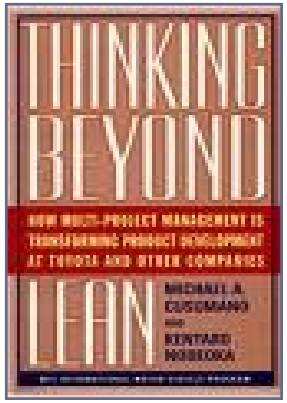

1991



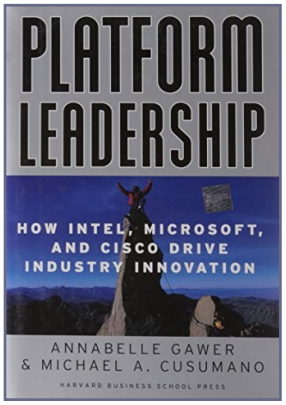

1995



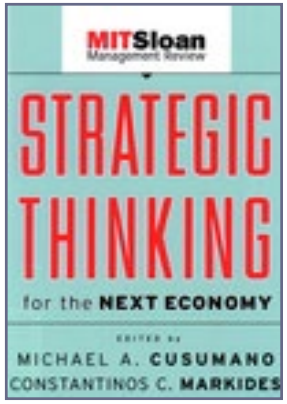

1998



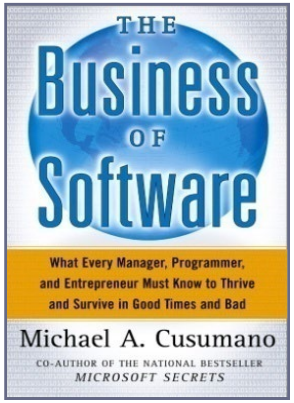

1998



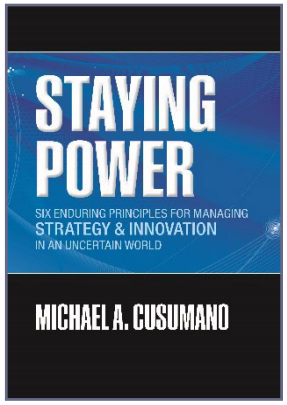

2002



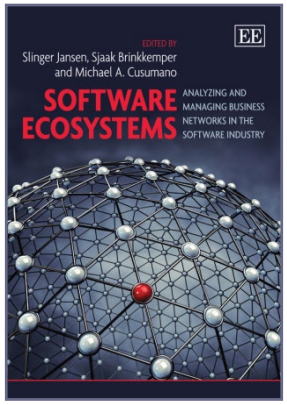

2002



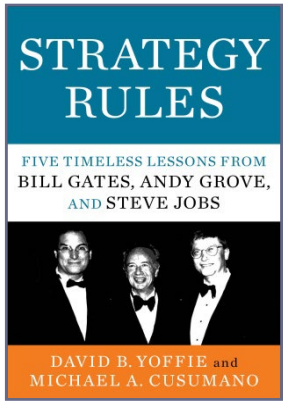

2004



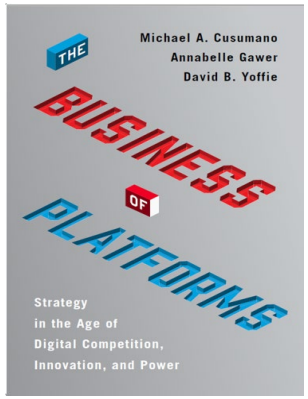

2010




2013




2015




2019

# “Platforms” are Everywhere!

| Platform | Examples |
|---|---|
| **Computers** | Mainframes … Windows, Macintosh … Android, Linux … Quantum |
| **Smartphone OS** | Google, Apple… Microsoft … Huawei, Xiaomi, Samsung |
| **Social Media** | Facebook, Google, Twitter, Instagram, Snapchat, WeChat, TickTock |
| **Video Games** | Microsoft, Sony, Nintendo … Epic Games |
| **Enterprise Software** | SAP, Oracle, Microsoft… Salesforce, Intuit |
| **Microprocessors** | Intel, ARM, Qualcomm, Nvidia |
| **Sharing Economy** | Uber, Airbnb, Lyft, HomeStay, TaskRabbit |
| **Messaging** | WhatsApp, WeChat, Line, Kakao |
| **Payments** | PayPal, Apple Pay, AliPay, WeChat … Bitcoin et al. |
| **Web Services** | Amazon, Microsoft, Google, IBM … |
| **Internet of Things** | GE, IBM, Oracle, Cisco, Amazon, Microsoft, Salesforce … |
| **AI/ML - LLMs** | Open AI/ChatGPT, Google, Microsoft CoPilot, DeepSeek, etc. |

**And many more platforms or platforms on top or alongside other platforms!**

# Platforms Are Among the World's Most Valuable Firms

- **Nvidia** ($4.46 trillion)
- **Microsoft** ($3.88 trillion)
- **Apple** ($3.40 trillion)
- **Amazon** ($2.38 trillion)
- **Alp/Goog** ($2.44 trillion)
- **TSMC** ($1.25 trillion)
- **Meta/FB** ($1.93 trillion)
- **Tencent** ($650 billion)
- **Alibaba** ($287 billion) + **Ant Financial** ($78 billion)

Uber ($187 billion)

Airbnb ($ 74 billion)

***And nearly half of all "Unicorns" and many other startups try to be platform businesses!***

# Product vs. Industry Platforms

*A. Gawer / Research Policy 43 (2014) 1239–1249*

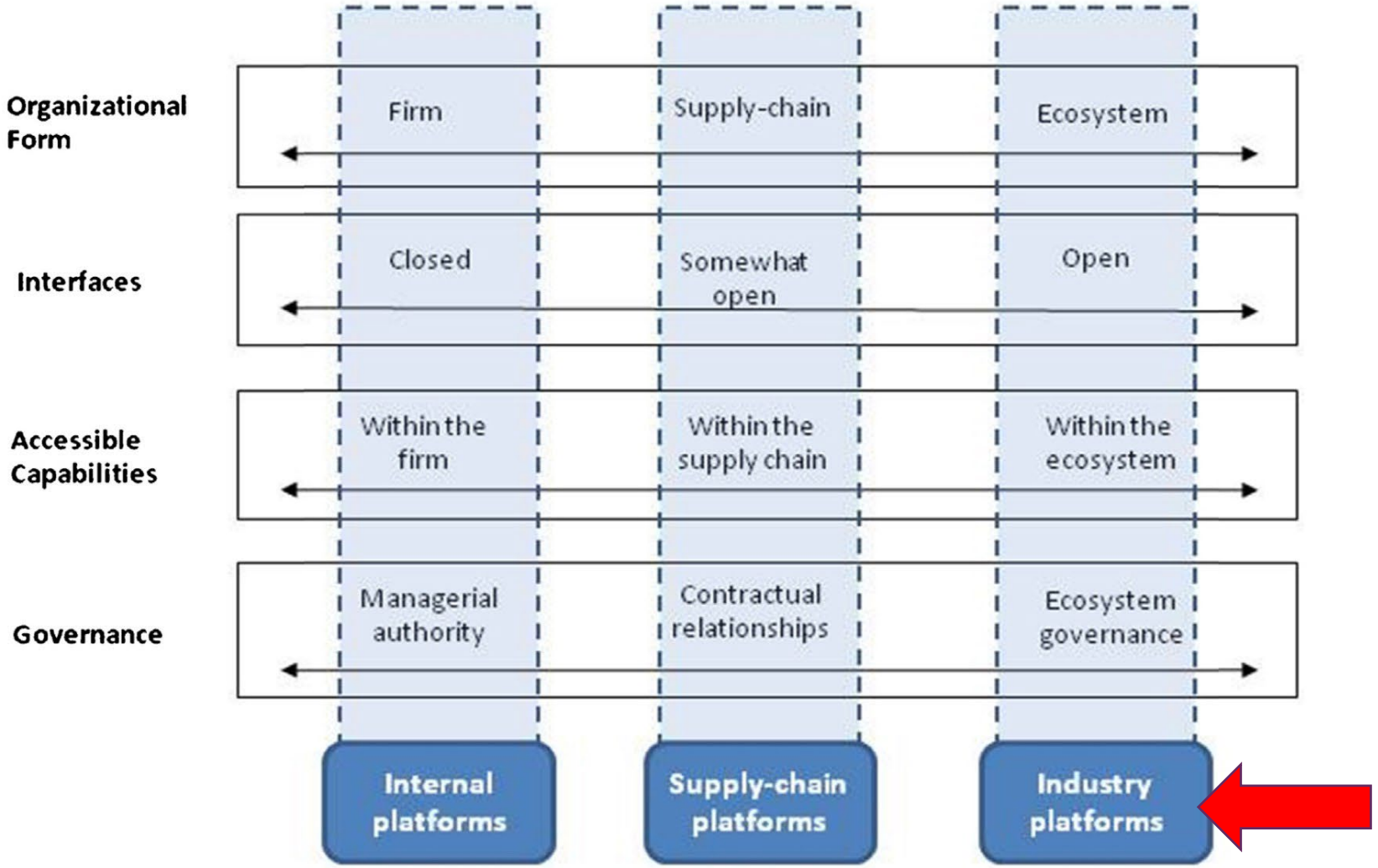
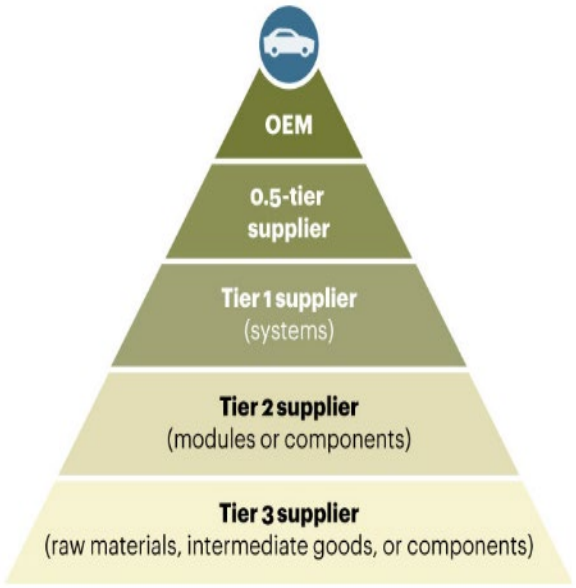
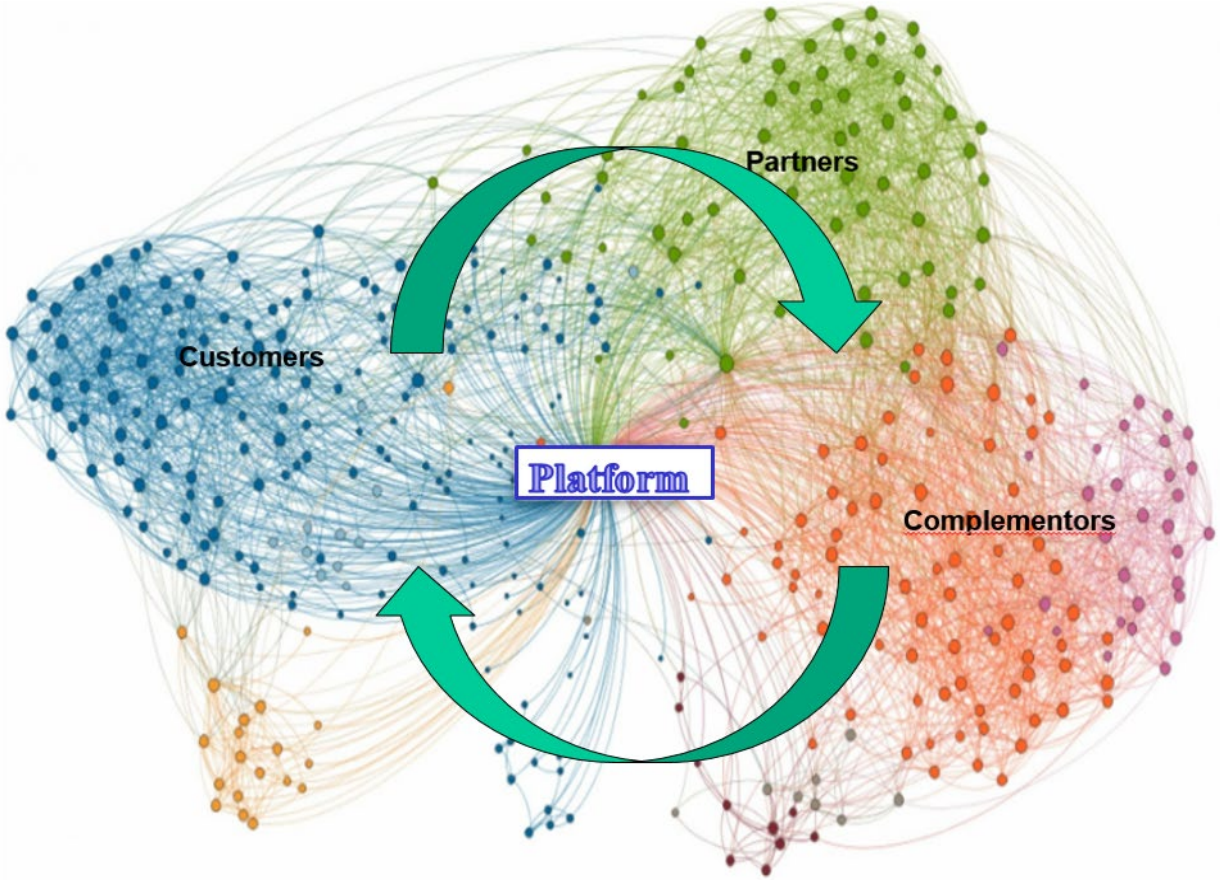



**Fig. 1.** The organizational continuum of technological platforms.

# Product (+ Supply Chain) vs. Industry Platform (+ Ecosystem)

**Product Platform** = reusable modules, black box or white box parts, managed by contracts




**Industry Platform** = Global Network of customers, complementors & partners, driven by network effects

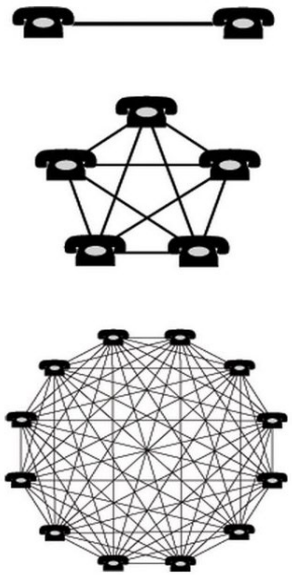
# ***Industry Platform*** **Definition?**

- A product or service that serves as a common foundation to exchange information & goods or to enable 3rd-party "complementary" products/services
  - ***User interactions & complementary innovations would not occur, or not occur so easily, without the platform.***
  - ***The more users or "complements," the more valuable the platform becomes (= network effects)***
    - ***Self-reinforcing positive feedback loops***
    - ***Growth potential is geometric or exponential, not linear***
    - ***Loss or decline potential also non-linear …***

# Network Effects – Definition?

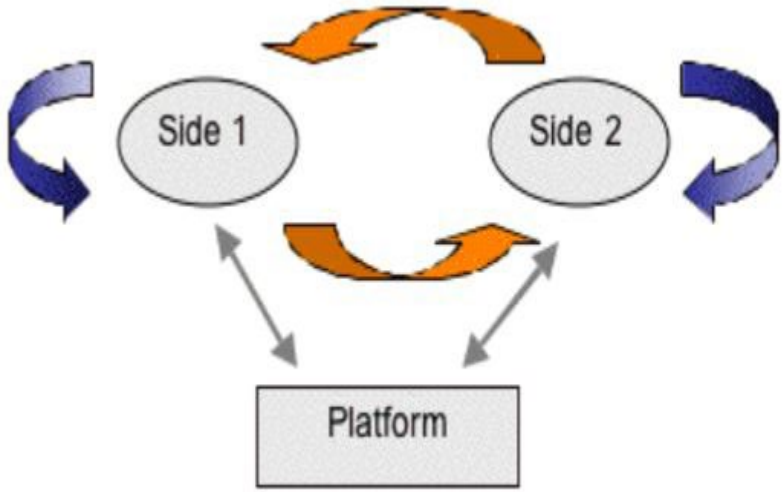
## ***Self-reinforcing positive feedback loops***

**where the potential value increases with each additional user or complement**



**Metcalf's "Law"**
$n(n-1)/2$

| Nodes | Connections |
|---|---|
| 2 | = 1 |
| 5 | = 10 |
| 10 | = 45 |
| 1000 | = ca. 500K |
| 1M | = ca. 500B |




**Historical Examples:**

telephone, railroads, Yellow Pages, credit cards, computers, fax machines, VCRs, Internet browsers, IoT, cryptocurrencies

# Industry Platforms:
# *What Do They Have in Common?*

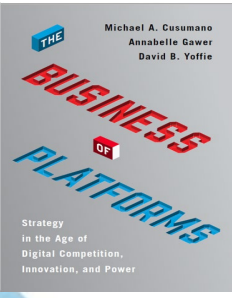
1. **Bring together 2 or more "market sides"**
   **(key market participants)**

2. **Generate unique value from "network effects"**
   **(direct/same-side … indirect/cross-side)**

3. **Must solve a "chicken-or-egg problem"**
   **(how create value & get network effects started?)**

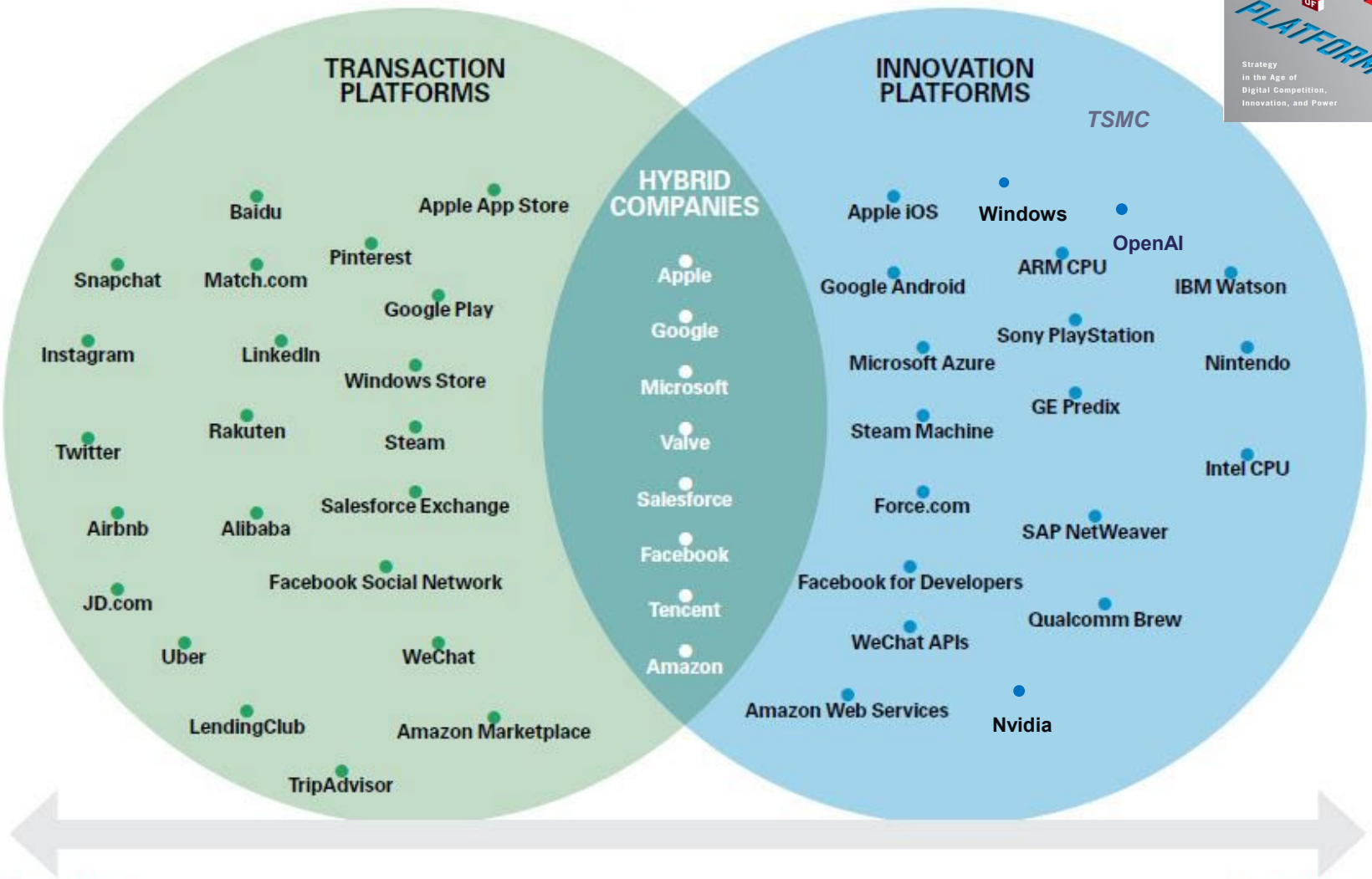
***Complex business models & market dynamics.***

***Like playing "three-dimensional chess"!***




# BASIC PLATFORM TYPES

In the quest for competitive advantage, companies are combining transaction and innovation platforms into a hybrid model.

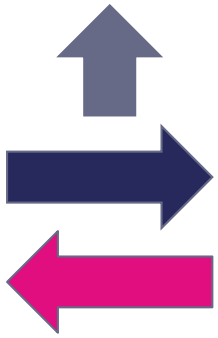



**Transactions**
The platform serves as an intermediary for direct exchange or transactions, subject to network effects.

**Innovations**
The platform serves as a technological foundation upon which other firms develop complementary innovations.

SOURCE: *THE BUSINESS OF PLATFORMS: STRATEGY IN THE AGE OF DIGITAL COMPETITION, INNOVATION, AND POWER* (HARPER BUSINESS, 2019)

# Hybrid Strategy



## Innovation Platform

- Operating systems
- Products with 3rd-Party APIs
- Building blocks (e.g., CPU, GPU)
- Toolkits for complementors

Technology becomes “core”
to solve an industry need

## Transaction Platform

- Marketplaces
- App Stores
- Advertising engines
- Data centers with integration services

Service becomes “core”
to solve an industry need

# The Argument in Brief

**Innovations:** If you can create (more) value by enhancing your product or service with 3rd-party "complements," try/add an ***innovation platform***.

***E.g., Windows OS + applications.; iPhone + App Store, Intuit QuickBooks + third-party apps***

**Transactions:** If you can create (more) value by connecting market sides rather than creating/reselling a product or service, try/add a ***transaction platform***.

***E.g., Amazon Marketplace vs. Walmart, Airbnb vs. traditional hotels, Uber vs. traditional taxis***

# Winner Take All or Most? (WTAoM)

1) **Strong network effects** among users & between platform & complements (new strategy concept)
2) **"Multi-homing" costly** (difficult to use more than one platform as your "home") (new strategy concept)

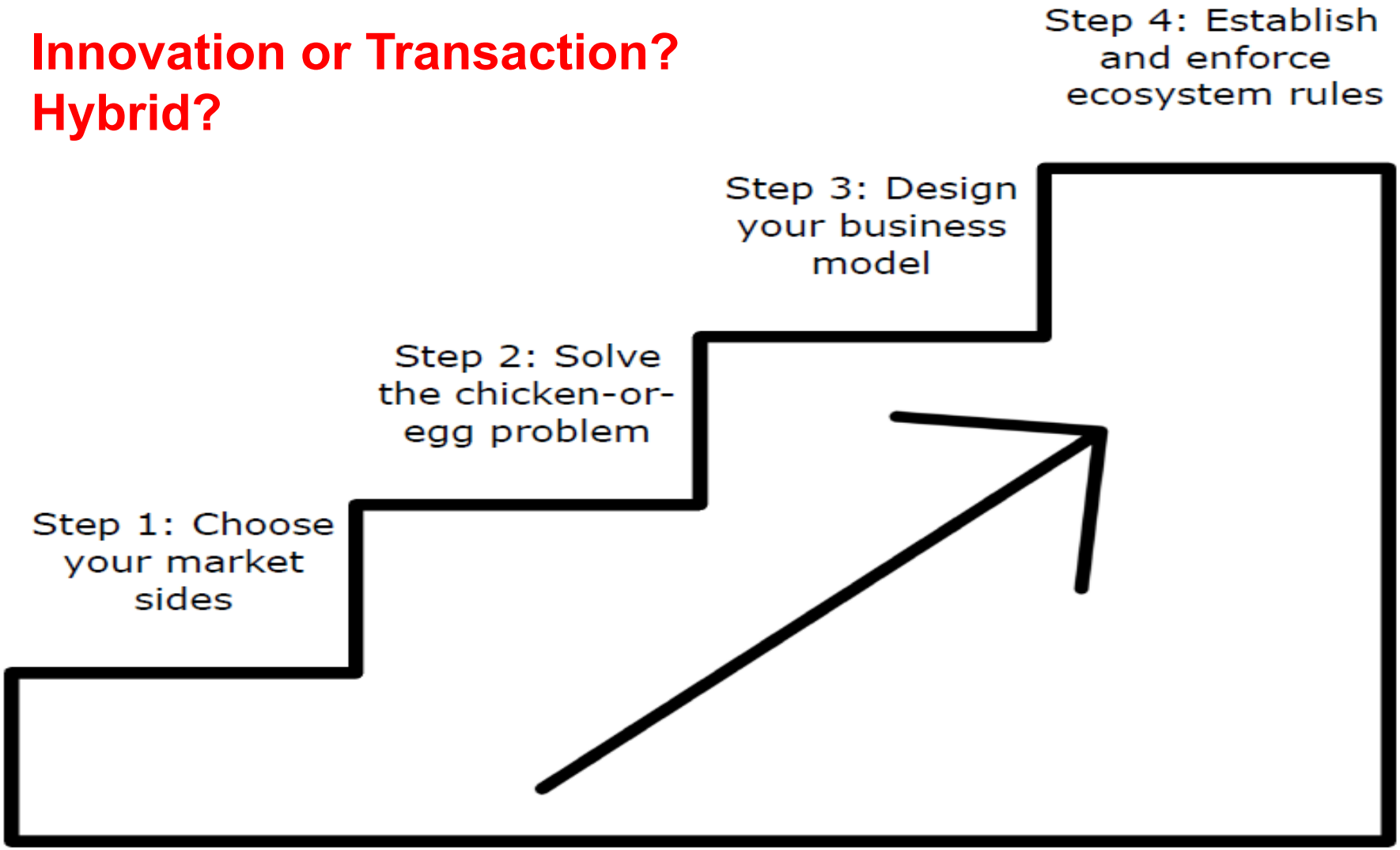
3) **Little differentiation or niches** among competitors
4) **High barriers to entry** for potential new competitors

**MAKE THEM CHOOSE -- YOUR PLATFORM!**

Ref: Eisenmann, Parker, and van Alstyne (2006); Cusumano (2010); Cusumano, Gawer, and Yoffie (2019)

# How to Build a Platform Business

**Innovation or Transaction?**
**Hybrid?**



## Users

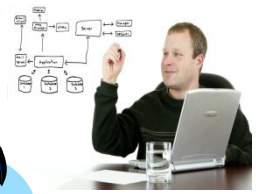
## Technology Partner

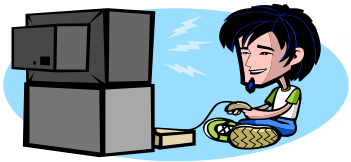
intel

MICROSOFT WINDOWS

## Applications

Microsoft Office





## PC Makers

DELL

Lenovo

HEWLETT PACKARD

## Users & Friends



## Platform Partners

Login with Facebook

CNN

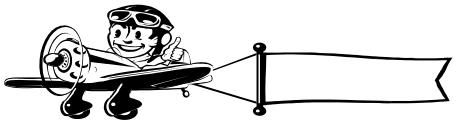
Spotify

Pandora

# Platform

facebook

## Advertisers

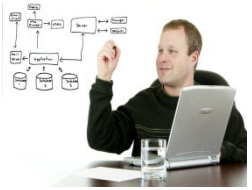


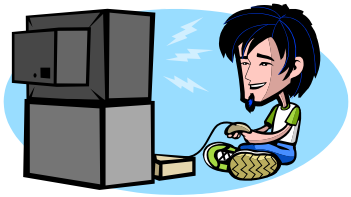
Buy Now!

Coca-Cola

## Applications

zynga





## Users – 150 million active

## B2C Partners


airbnb

*Pricing tools, Photos, etc.*

## Listings – 7 million
## 220 countries & 100k cities

## B2B Partners

# Chicken-or-Egg Solutions?

## 1. Create standalone value for one side first

- Product/Service that does not need 3rd-party complements
- Your solution solves a common industry problem
  - ***E.g., Amazon Store, iPhone, Nvidia GPU, QuickBooks, Open Table***

## 2. Subsidize 1 side (maybe 2 sides temporarily)

- Offer free platform access or financial & technical assistance
- Build or buy key complements & bundle with the platform
  - ***E.g., Windows-Office, Google Android, Facebook, Other?***

## 3. Bring on 2+ sides simultaneously or zig-zag

- Pay (subsidize) 2 sides to connect, maybe temporarily
  - ***E.g., credit cards, eBay, Uber, Bitcoin … Other?***
- Form "partnerships" instead of relying on network effects, early on

# *"Platformania"*

- **"Platformizing" a "bad" business does not make it a "good" business!**
  - E.g. ride sharing or grocery shopping via a **digital platform** does not generate the same high profits as selling **digital goods!**
  - Platforms should solve a **market failure** where there is "good" (profitable) business potential – industry demand-supply imbalance, also with opportunity for new scale/scope economies via digital
- **If the platform needs to subsidize 2 or more sides in order to operate, then it has NOT solved the "chicken-or-egg" problem. *Not growing via network effects!***

***The bigger the platform gets, the more money it will lose! …until?***

## Riders (150 million)

## B2C Partners

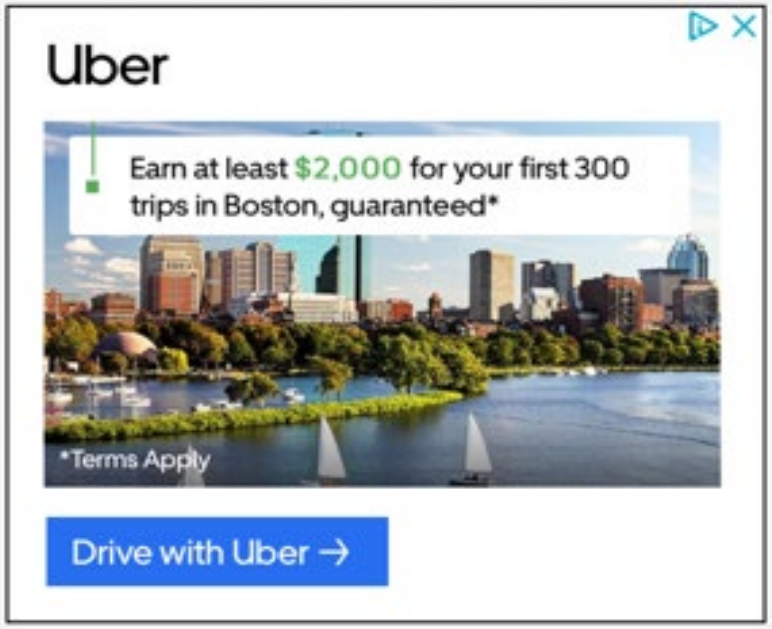


UBER

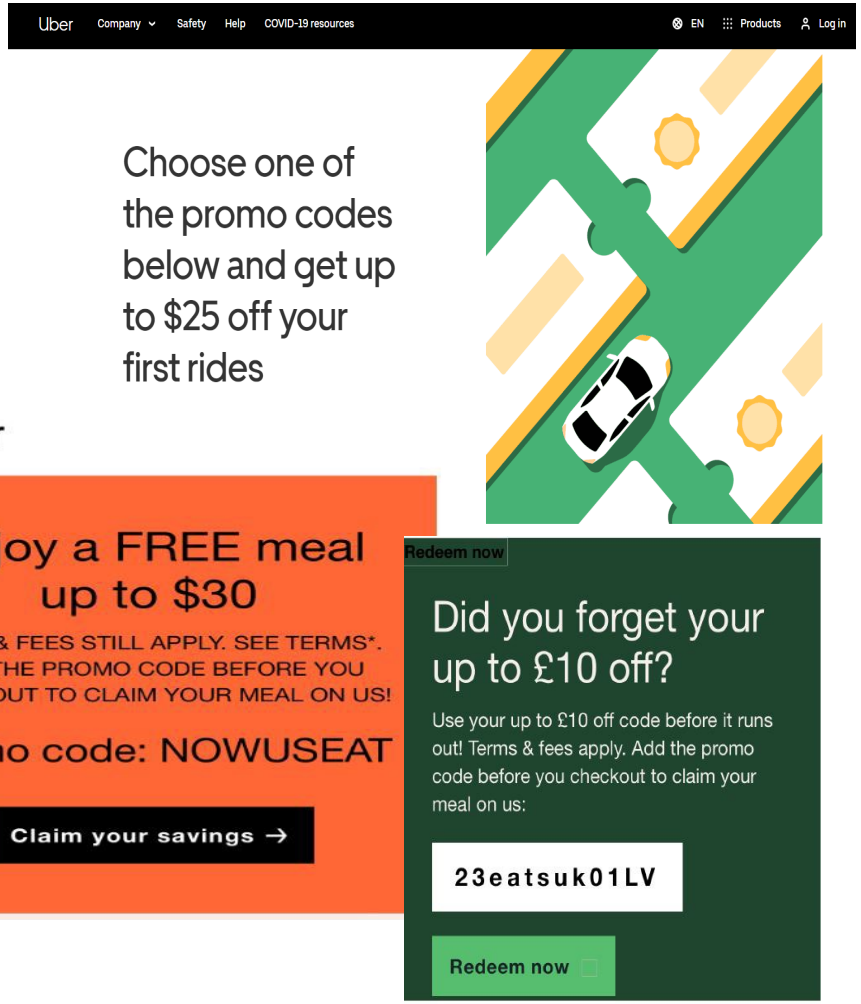
## Drivers (6 million), 70 countries

## B2B Partners

# Driver Side

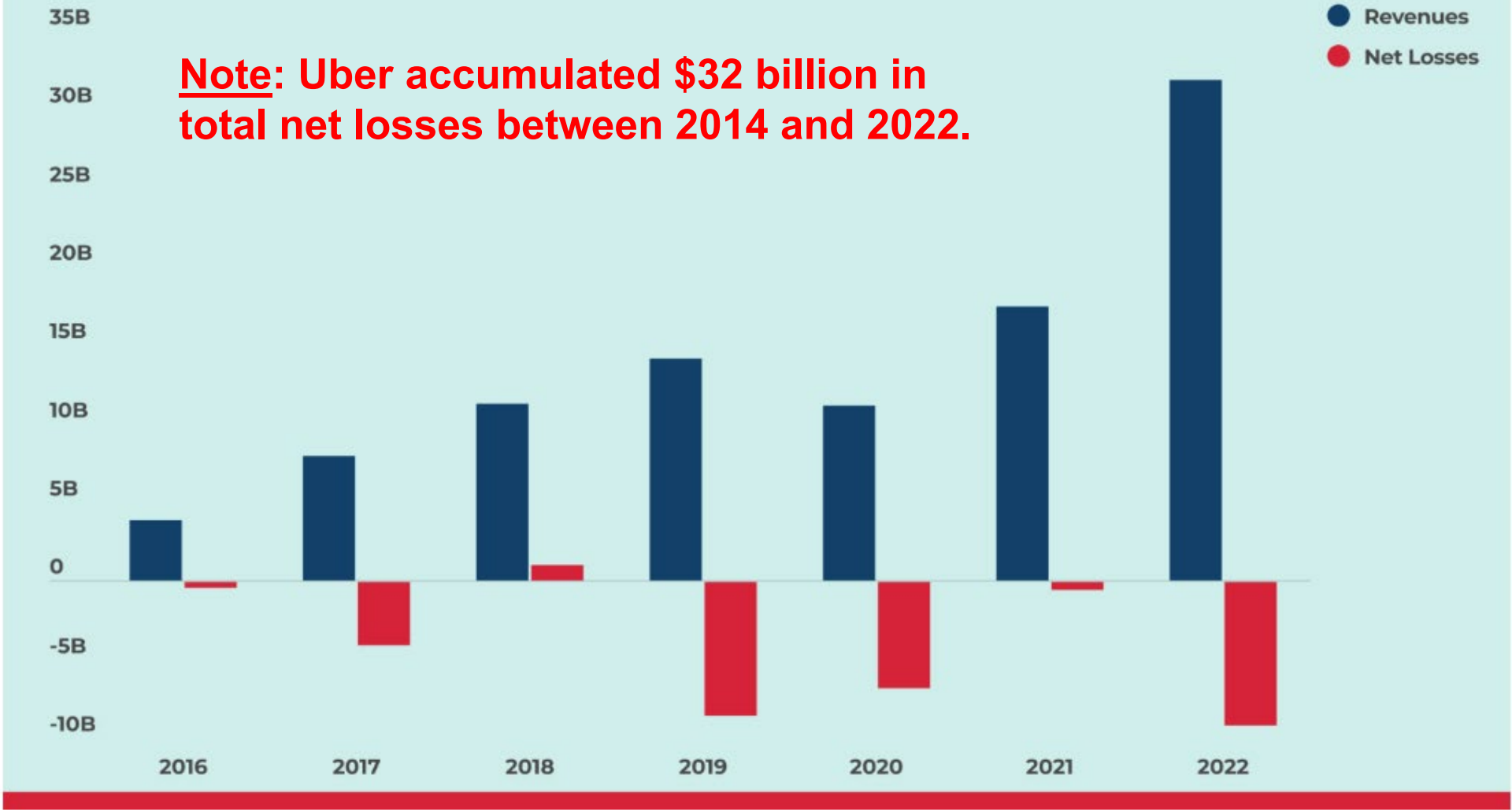



# User Side



# Is Uber Profitable? Uber Net Losses 2016-2022

As of 2022, on net revenues of $31.87 billion, Uber posted a net loss of $9.14 billion. In 2021, Uber posted a lower net loss ($496 million), primary thanks to the business divestitures of various assets. Throughout its history, on an annual basis, Uber has never made a profit. Yet, it has also shown incredible business growth, over the years, with its revenue at $3.8 billion in 2016, to almost $32 billion in 2022.




**Note: Uber accumulated $32 billion in total net losses between 2014 and 2022.**

Source: https://fourweekmba.com/is-uber-profitable/

# Airbnb and Uber Comparison

| Year | Airbnb Revenue ($Billion) | Airbnb Operating Profit ($B) | Uber Revenue ($Billion) | Uber Operating Profit ($B) |
|---|---|---|---|---|
| 2024 | 11 | 2.6 *(23%)* | 44 | 2.8 (6%) |
| 2023 | 10 | 1.5 *(15%)* | 37 | 1.1 (8%) |
| 2022 | 8 | 1.8 *(21%)* | 32 | -1.8 |
| 2021 | 6 | 0.5 *(8%)* | 17 | -3.8 |

Source: Annual Reports

# Platform Business Advice

1. Target an ***industry-wide problem*** related to a demand-supply imbalance: ***"market failure" = opportunity!***

2. Design a ***compelling solution***– standalone with platform potential **or** a platform from the "moment of creation."

3. Design an ***innovation or transaction platform*** – depending on the type of problem and which actors you want to bring together ***who would not otherwise connect***.

4. Attack the chicken-or-egg problem! Offer ***standalone value*** OR ***subsidize the most important side*** (e.g., make access FREE for users or complementors).

5. After launch, ***grow via network effects or partnerships***, ***not permanent 2-sided subsidies or below-cost prices!***

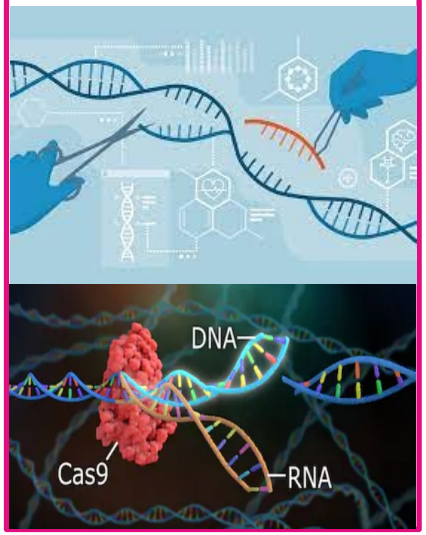
# Platform Trends & New Technologies

1. ***More Hybrids*** — **Innovation + Transaction Platforms**
2. ***More Concentration*** — **Fewer & Bigger Firms** (more regulation!)
3. ***New Enabling tech*** — **AI/ML, Blockchain, IoT, Biotech, Quantum, etc.**

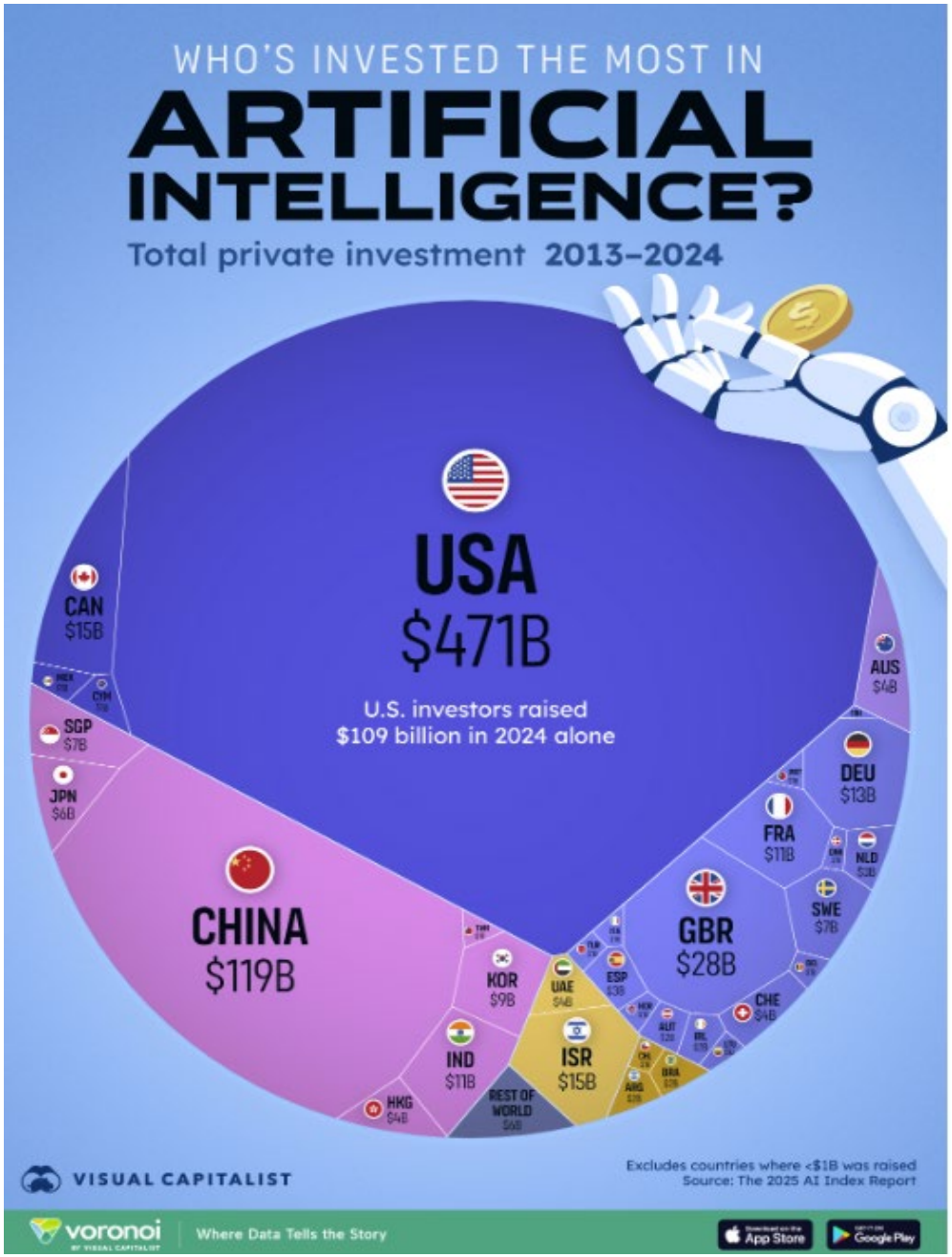
- AI/ML Enterprise & Home


- Self-driving vehicles & IOT


- Gene editing tools & ecosystem




- Quantum computing & communications




| Geographic area | Investment ($B, 2013-2024) |
|---|---|
| US U.S. | $471 |
| CN China | $119 |
| GB UK | $28 |
| CA Canada | $15 |
| IL Israel | $15 |
| DE Germany | $13 |
| IN India | $11 |
| FR France | $11 |
| KR South Korea | $9 |
| SG Singapore | $7 |
| SE Sweden | $7 |
| JP Japan | $6 |
| AU Australia | $4 |
| CH Switzerland | $4 |
| AE UAE | $4 |
| HK Hong Kong | $4 |
| NL Netherlands | $3 |
| ES Spain | $3 |
| AT Austria | $2 |
| BR Brazil | $2 |

Source: https://www.visualcapitalist.com/visualizing-global-ai-investment-by-country/ (April 21, 2025)

| Geographic Area | Number of Newly Funded AI Companies (2013–2024) |
|---|---|
| US United States | 6,956 |
| CN China | 1,605 |
| GB United Kingdom | 885 |
| IL Israel | 492 |
| CA Canada | 481 |
| FR France | 468 |
| IN India | 434 |
| DE Germany | 394 |
| JP Japan | 388 |
| KR South Korea | 270 |
| SG Singapore | 239 |
| AU Australia | 178 |
| CH Switzerland | 154 |
| ES Spain | 117 |
| NL Netherlands | 116 |

| Focus Area | Total Investment ($B, 2024) |
|---|---|
| AI infrastructure/research/governance | $37.3 |
| Data management, processing | $16.6 |
| Medical and health care | $10.8 |
| AV (autonomous vehicles) | $9.4 |
| Fintech | $6.9 |
| Manufacturing | $6.6 |
| Semiconductors | $5.5 |
| NLP, customer support | $4.2 |
| Cybersecurity, data protection | $3.7 |
| Robotics | $3.3 |
| Drones | $2.6 |
| Energy, oil, and gas | $2.0 |
| Marketing, digital ads | $1.6 |
| Business operations | $1.5 |
| Semantic search | $1.4 |
| Supply chain | $1.4 |
| Insurtech | $1.4 |
| AR/VR | $1.3 |
| Retail | $1.2 |
| Ed tech | $1.0 |
| Quantum computing | $1.0 |
| IoT | $0.8 |
| Agritech | $0.8 |
| Content creation/translation | $0.8 |
| Creative, music, video content | $0.7 |

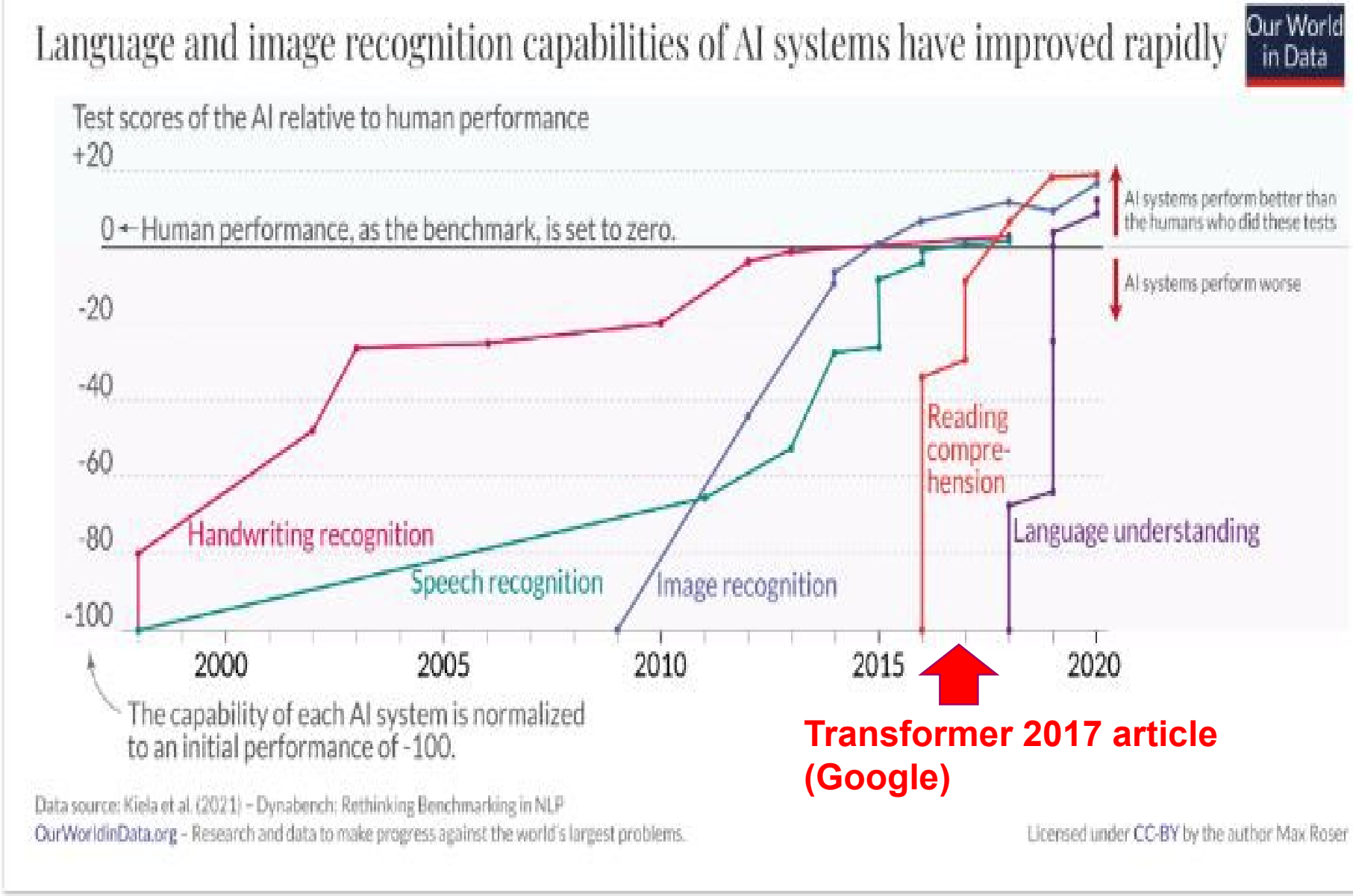
# (1) GenAI Innovation History

- AI/ML using neural networks dates back to the 1950s, but ... ***Inflection point:*** **research done at Google, in 2017**.

- **"Transformer"** for language translation: Identifies patterns and creates "digital tokens" representing meaning AND context. ***Sequence & context used to predict what "should" come next.***
  - **1 token =** ca. 4 characters or 0.75 words
  - **"parameters"** = variables or weights used to fine-tune tokens, via model prompts

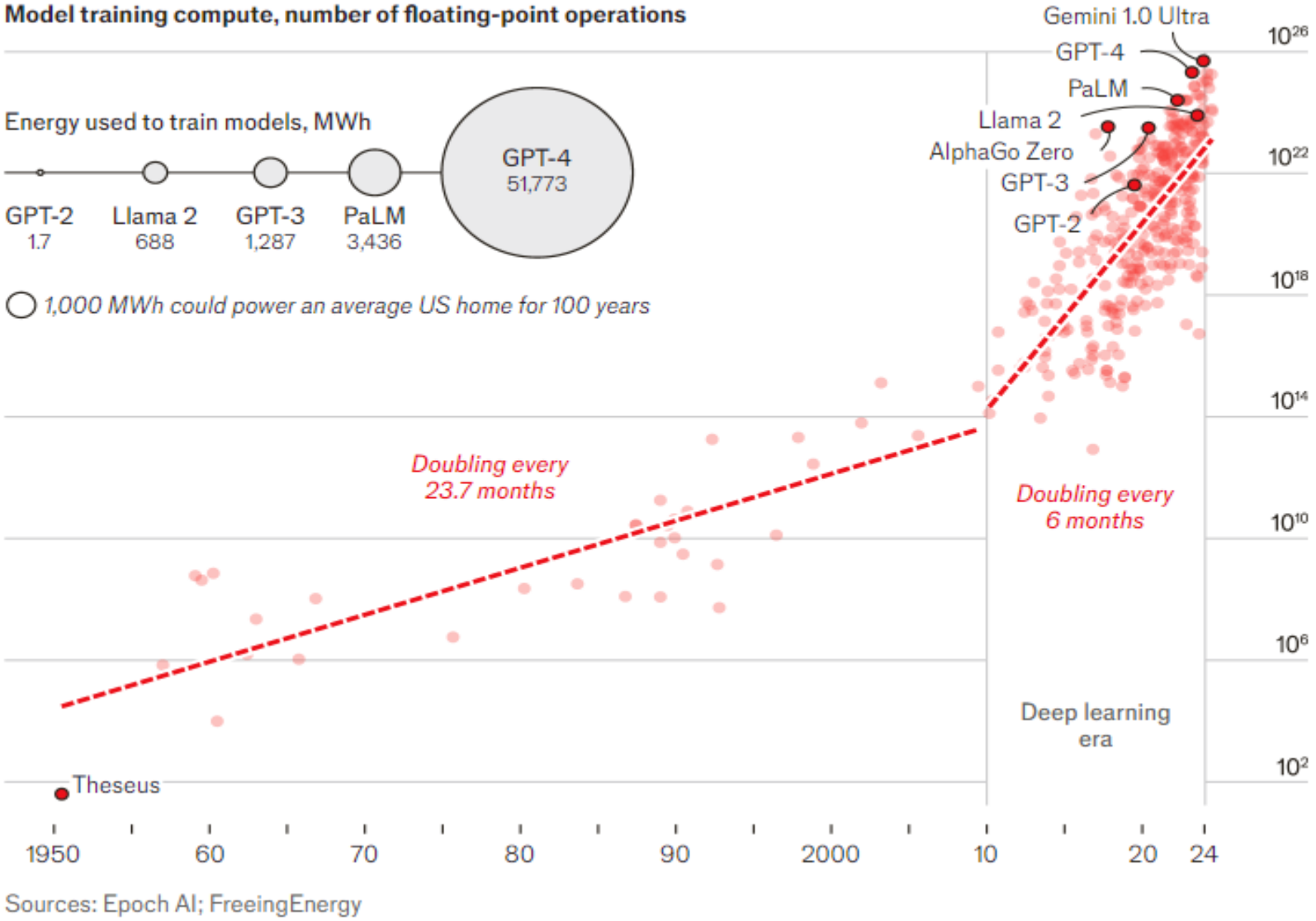
- Trained on HUGE data sets – "whole languages" – and produce language on their own. ***Vast improvement for translation!***

➢ **Google researchers moved to other firms, including OpenAI**. Developed many use cases **beyond language translation**, and "multi-modal" language models (text + images, audio, video).




Source: Max Roser, "The brief history of artificial intelligence," *Ourworldindata.org*, December 6, 2022

## → Re-doubling




Sources: Epoch AI; FreeingEnergy

Source: "The race is on to control the global supply chain for AI chips," *The Economist*, July 30, 2024

# (2) Nvidia Innovation History

- **1993:** Nvidia founded to make ***graphic cards*** for PCs.
  ***Struggled … limited market … and Intel/AMD competition***

- **2006:** Switched from CPU-like design (dozens to hundreds of complex "compute cores") to many thousands of simple cores, all running in parallel at superfast speeds

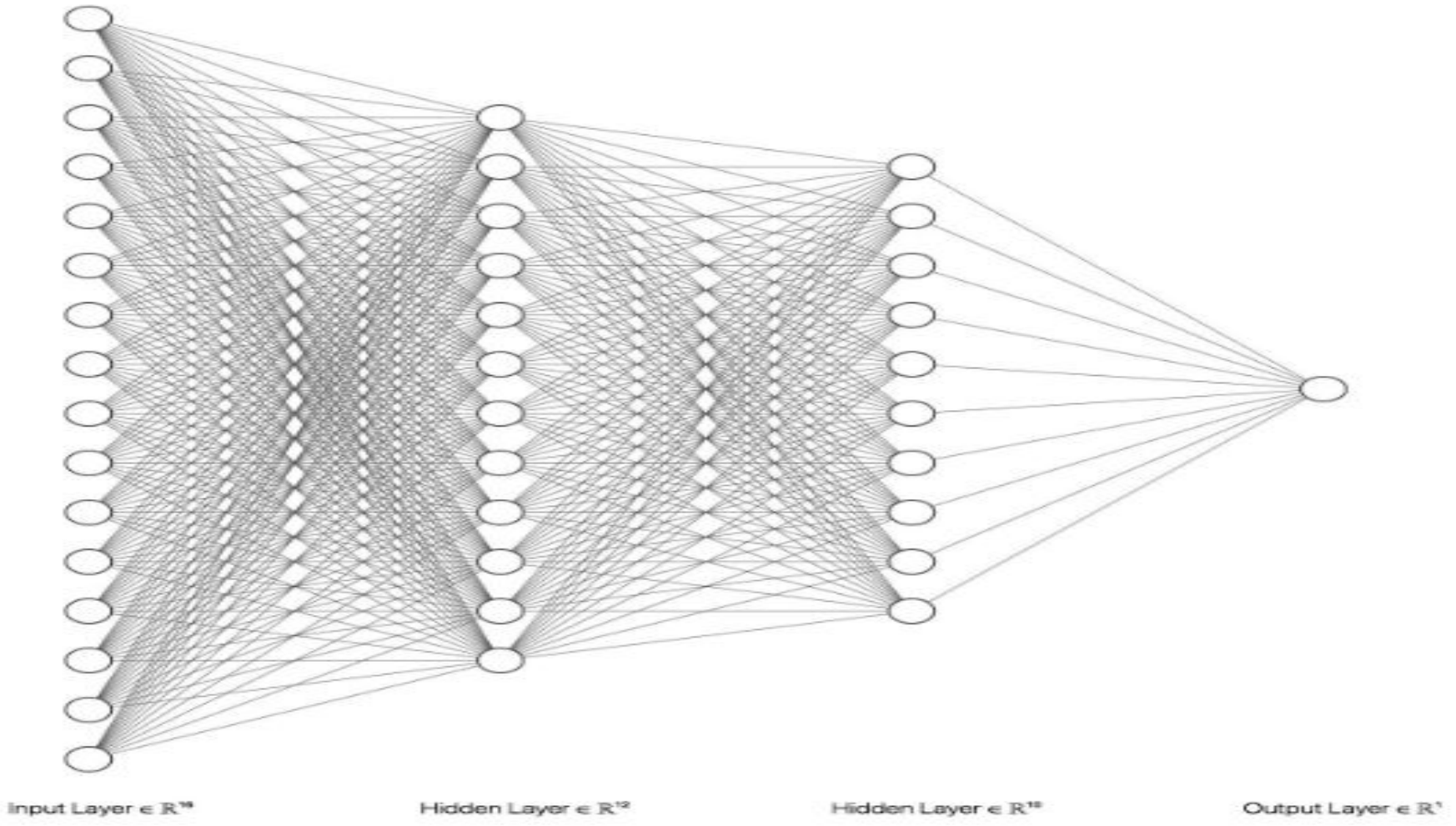
- **2006:** Introduced CUDA *(Compute Unified Device Architecture)* as a **FREE** Software Development Kit (SDK) to program GPUs for super-fast parallel processing

# Nvidia History continued

- **2012:** Independent researchers use Nvidia graphic cards to train neural networks
  - ***Nvidia GPUs & CUDA "perfectly suited" to the thousands of matrix multiplication tasks & logic layers in neural networks***

- **2013:** CEO Huang commits to AI business focus
- **2014:** CuDNN (CUDA Deep Neural Network) SDK
- **2016:** *Nvidia gives away powerful GPU servers to OpenAI … and other researchers & universities*
- **2025:** Nvidia GPUs/CUDA continually GenAI-optimized

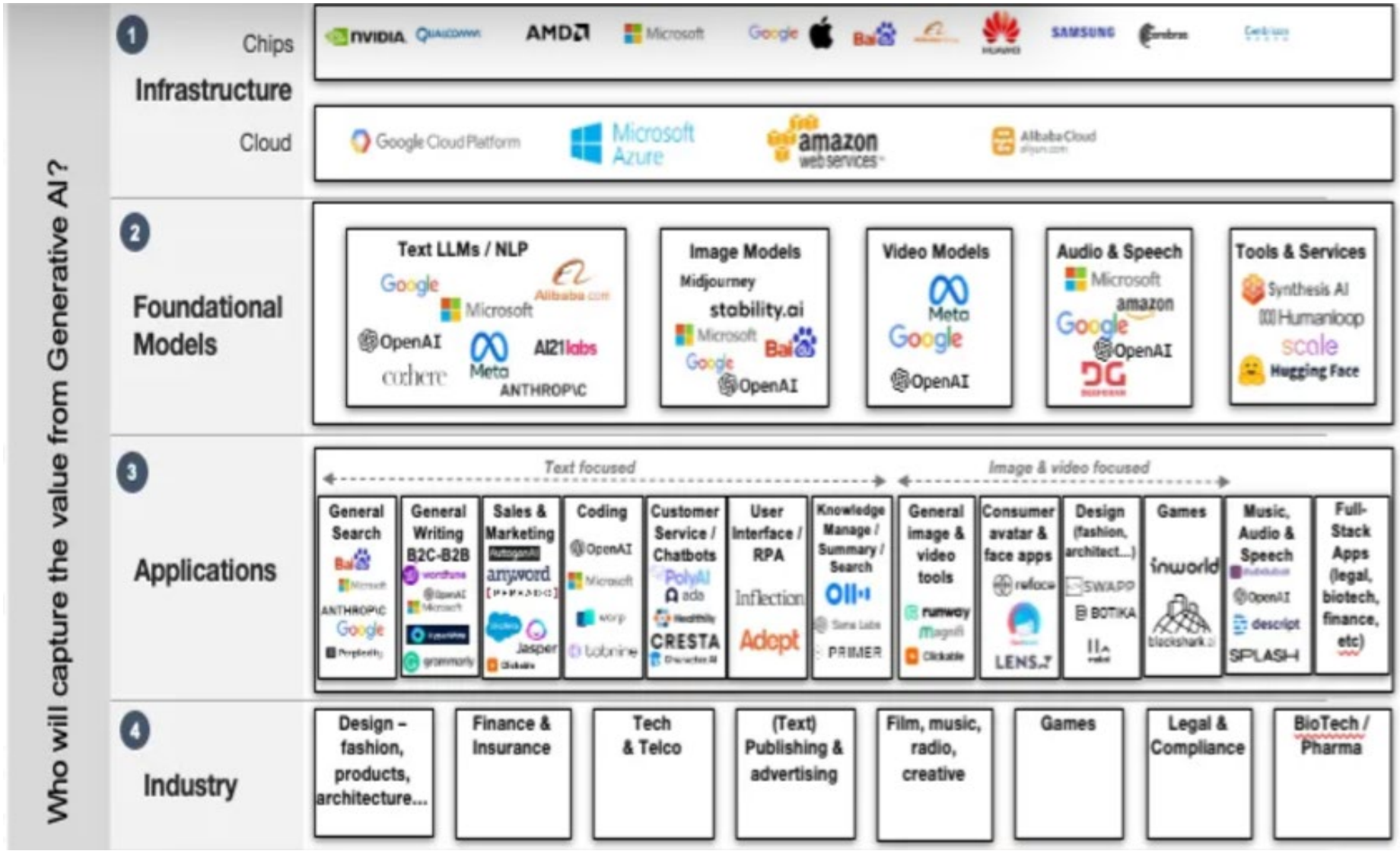
# Neural Network Graphic






Source: https://datascience.stackexchange.com/questions/12851/how-do-you-visualize-neural-network-architectures

# (3) GenAI Platform & Ecosystem




Source: https://medium.com/@sgreenman/who-will-make-money-from-the-generative-ai-gold-rush-9e8a291d2af6 (March 12, 2023)

# Competition in LLMs


**Note: MMLU = "Multi-task Language Understanding" as a benchmark**

Source: Charles River Associates presentation, Academy of Management July 2025

# LLMs: How Differ from OS's?

- **Like *operating systems*. But also like *applications*.**
  - LLMs can be trained to answer specific questions, like in a search engine, or to perform specific tasks, like routing customer service requests or generating software code.
  - "Reasoning" LLMs show and "check" their thought process

- **LLMs vary by capabilities, API cost & policies**
  - Larger LLMs ***usually*** out-perform smaller LLMs, ***until DeepSeek***
  - Smaller, specialized LLMs are cheaper to design, train, operate

- **Open-source LLMs free, but require payments for access to APIs & parameter training weights**
  - Best open-source LLMs (Meta, Mistral, Cohere, DeepSeek) now comparable to OpenAI models

# Powerful Network Effects!

- **Nvidia GPUs = best speed + performance!**
  - Today: Installed base of 500 million servers, 80% market share
- **Nvidia CUDA = best tools/libraries, and FREE!**
  - 300 code libraries & 600 AI models, supporting 3,700 GPU-accelerated applications, used by 5 million developers at 40,000 firms (WSJ, 8/2024). Global startup network of 15,000 firms

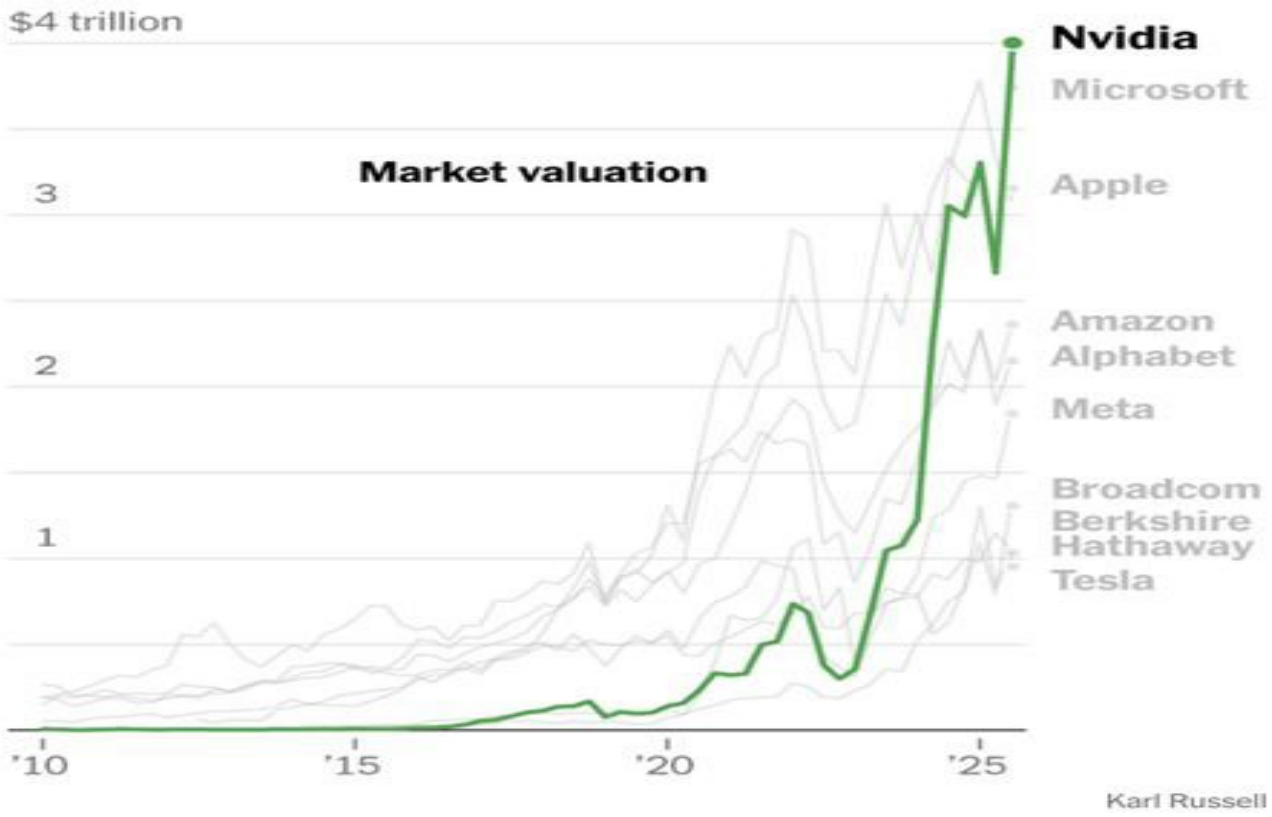
➢ ***More Nvidia GPUs = Can write/run more CUDA software = More demand for Nvidia GPUs, etc.***

❖**But: CUDA proprietary & *relatively difficult* to use.**
Based on C/C++ … porting to Python etc.. Most developers of kernel software prefer CUDA, or program in CUDA first.

# Nvidia Becomes First Public Company Worth $4 Trillion

The A.I. chip maker reached the milestone before Apple and Microsoft, after it jump-started the A.I. frenzy more than any other company.

4 MIN READ




Source: July 10, 2025 ***The New York Times***

# AI Entrepreneurship in China

- **2015 – High-Flyer, a AI/ML hedge fund with $8 billion in assets. Co-founded in Hangzhou by Liang Wenfeng (bn. 1985), graduate of Zhejiang University, studied machine vision.**

  - **2019** – High-Flyer builds supercomputer with 10,000 Nvidia A100 GPUs (cost $139 million, but older, slower & less advanced than H100 or H800)

  - **2021** – After government guidance to reduce high-frequency stock trading & speculation, Liang creates a second team for fundamental AI research

# Liang Wenfeng (bn. 1985), Co-founder of High-Flyer AI/ML Hedge Fund and DeepSeek

# DeepSeek Chronology

- **2023 – DeepSeek spun off as a separate subsidiary (150-200 employees in early 2025, compared to 3500 for OpenAI)**

| | |
|---|---|
| • 2023 November | DeepSeek Coder |
| • 2023 December | DeepSeek-V1 (“general” LLM) |
| • 2024 May | DeepSeek-V2 |
| • **2024 December 26** | **DeepSeek-V3 Technical Report** |
| • **2025 January 10** | **DeepSeek-V3** |
| • **2025 January 20** | **DeepSeek-R1 (“reasoning” LLM)** |
| • **2025 January 22** | **DeepSeek-R1 white paper** |

# DeepSeek-V3

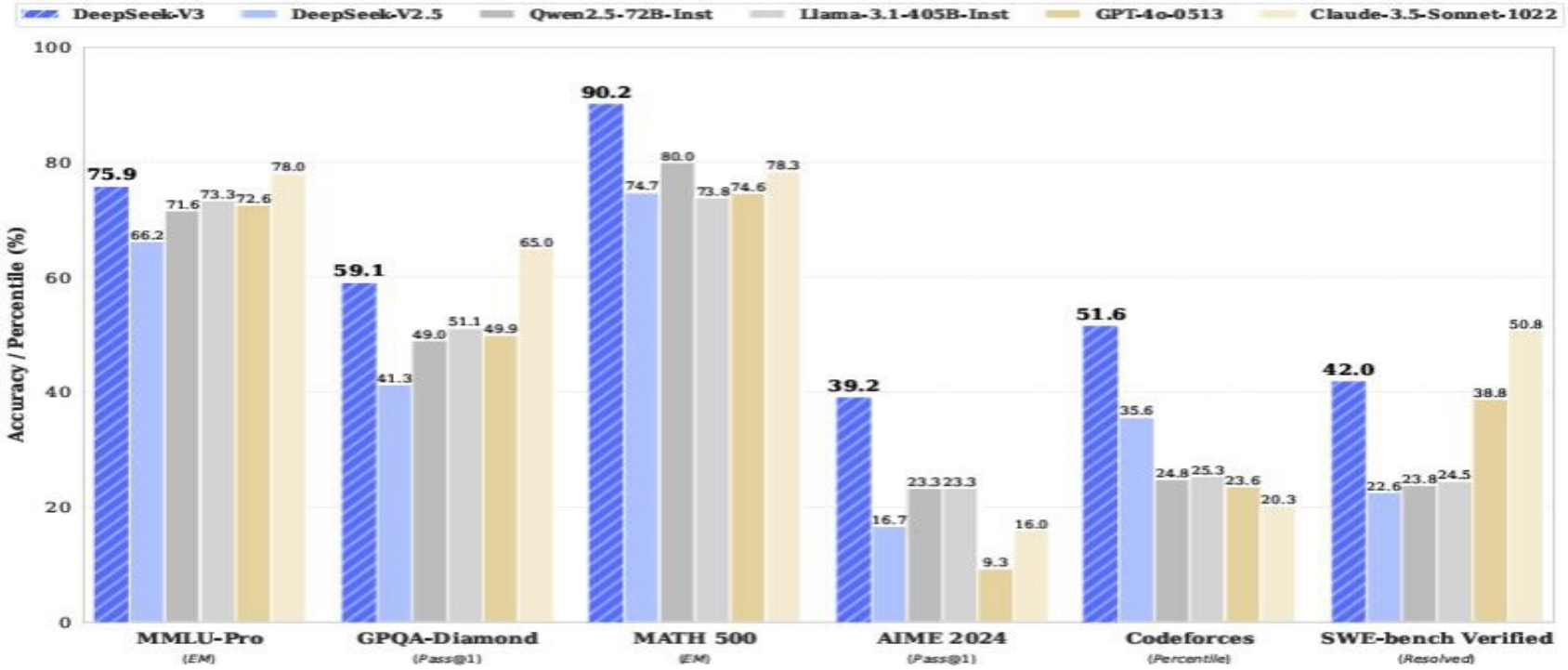



Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

| Training Costs | Pre-Training | Context Extension | Post-Training | Total |
|---|---|---|---|---|
| in H800 GPU Hours | 2664K | 119K | 5K | 2788K |
| in USD | $5.328M | $0.238M | $0.01M | $5.576M |

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

Source: DeepSeek-AI, "DeepSeek-V3 Technical Report," Dec. 27, 2024, pp. 1, 5.

# DeepSeek Training Costs

- **DeepSeek: $5.6 million to train V3, based on 2788k hours on 2,048 Nvidia H800 GPUs, at $2/hr**
  - **Cf. OpenAI GPT-4 used 16,000 Nvidia H100 GPUs & cost at least $100m for training, excluding prior R&D**
  - Cf. Meta spent ca. $1 billon developing LLaMA 3 series
  - DeepSeek numbers don't include pre-V3 costs or High-Flyer supercomputer ($139m) or other hardware
  - SemiAnalysis estimates DeepSeek spent $500+ million on GPUs & $1.3 billion total – *but* even so, **GPT4 training costs were 20x the cost of DeepSeek V3.**

# DeepSeek V3 to R1

- DeepSeek-V3 the **base model**. Trained on 14.8 trillion tokens, roughly all the high-quality English internet.
- Similar volume of training to Meta's Llama 3 but **twice OpenAI GPT 4's 6.5 trillion tokens.**
- DeepSeek then built **R1 as a "reasoning" version of V3**, with extra machine-learning & conversational features.
- **Performance comparable to OpenAI GPT-4o but with substantially less training costs.**

# DeepSeek-R1

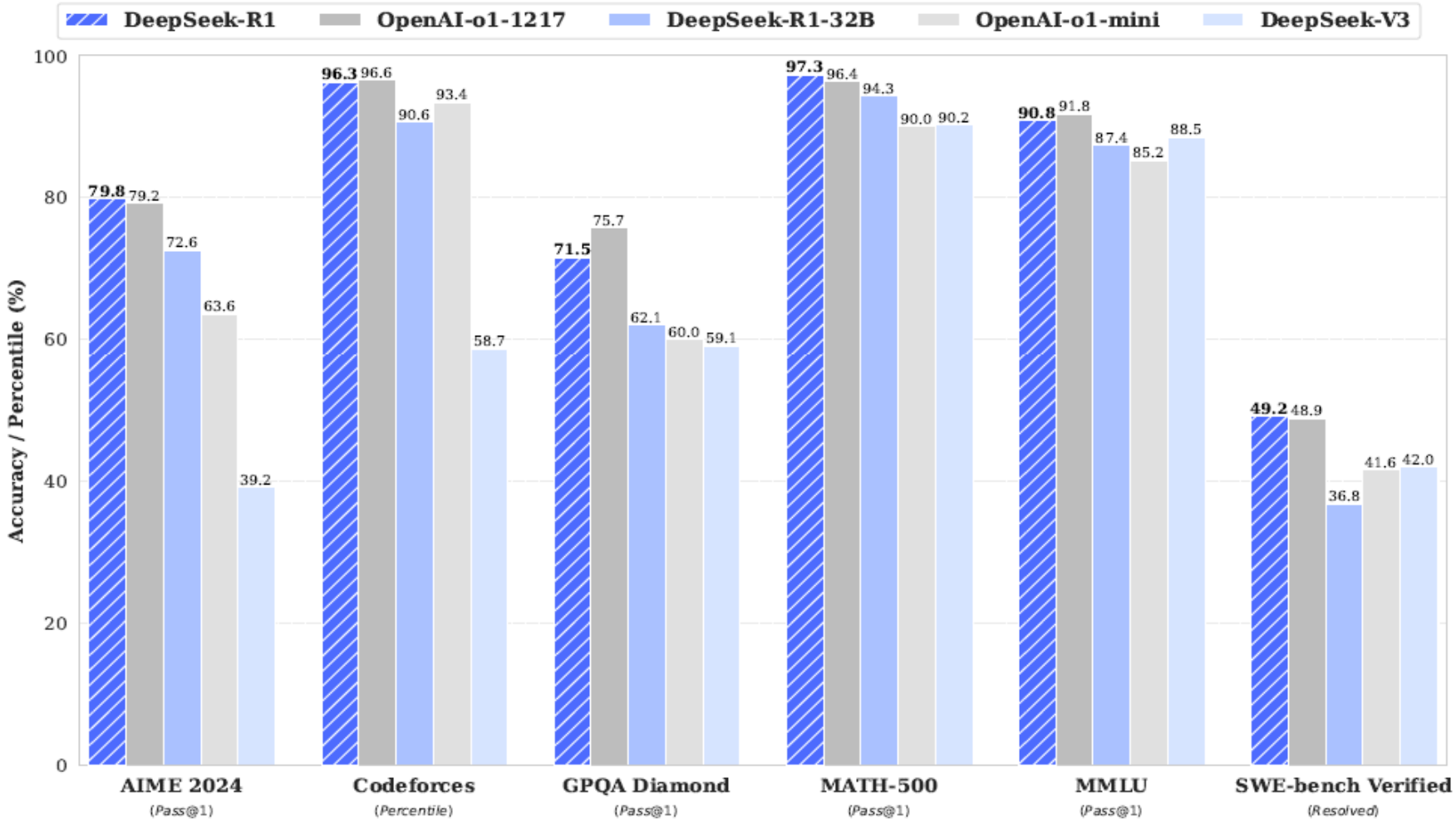



Figure 1 | Benchmark performance of DeepSeek-R1.

Source: DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," January 22, 2025, p. 1.

# V3/R1 Key Technologies

1. **Mixture of Experts (MOE)** – Use of several small specialized language models to spread out training & data analysis.

   a) ***Sparse Neural Network Training*** – Use a subset of parameters; avoid unnecessary time-consuming computations.

   b) ***Parallelization*** – Techniques to minimize resources by sharing data across the MoE models.

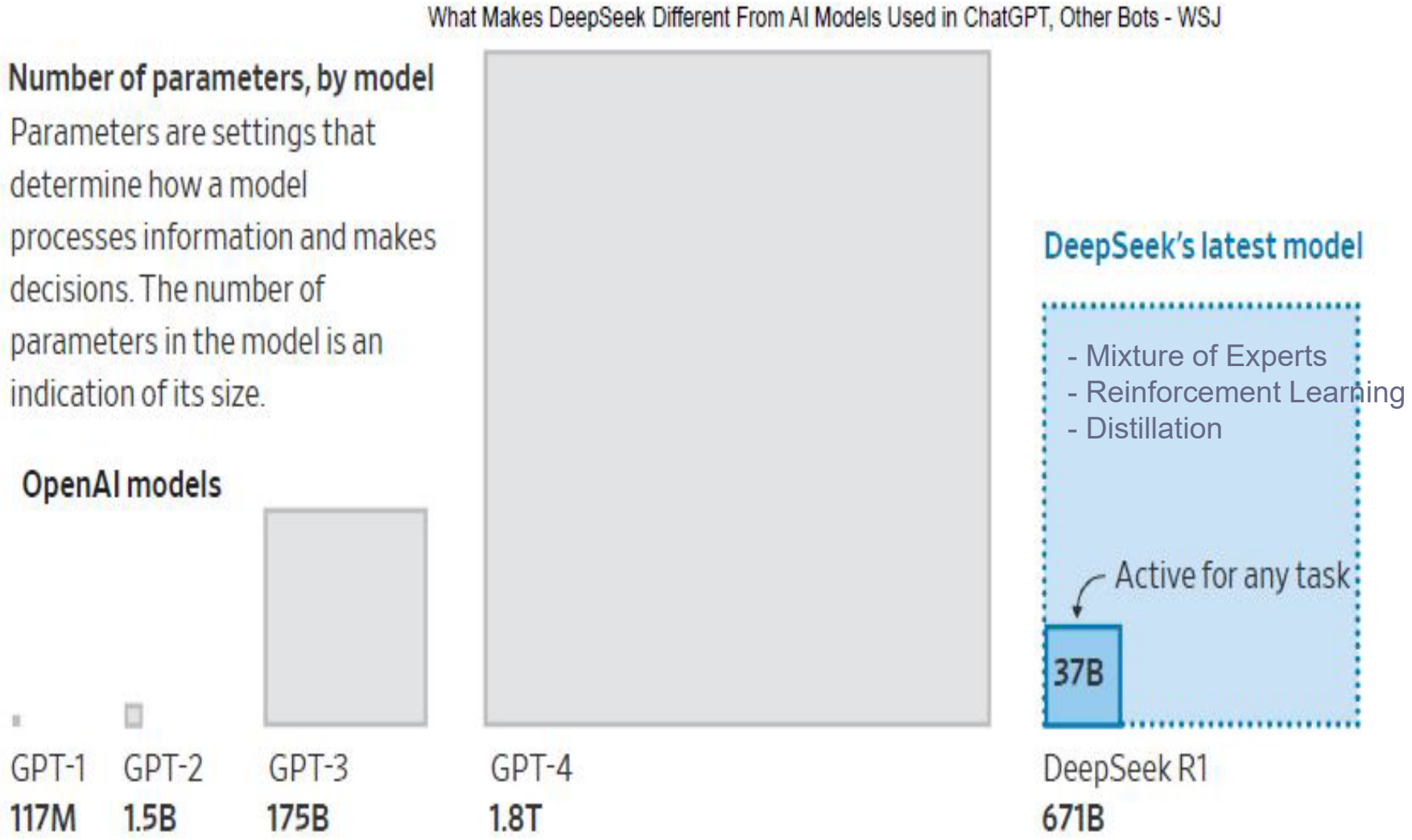
   c) **Quantification** – Limit precision of calculations (decimal points) only to what is necessary.

# LLM "Parameters"

- "Parameters" (like brain "synapses") are variables or weights used to train LLMs and configure their analysis of digital tokens, in different ways.
  - **Like "knobs and dials" to fine-tune the training data and improve outputs, i.e., analysis and predictions.**

➢ **OpenAI's approach to LLM training requires enormous amounts of data and many hours of human supervision and pre-labeling of data to adjust more than a trillion parameters.**

# Cf. Human brain = ca. 200 Trillion synapses ("parameters")




Sources: OpenAI, SemiAnalysis, DeepSeek, quoted in Pipe, Alana, and Rattner, Nate, How DeepSeek's Lower-Power, Less-Data Model Stacks Up, *Wall Street Journal*, February 16, 2025.

# V3/R1 Key Technologies

2. **Reinforcement Learning** – Some post-training "human-supervised fine-tuning" but mostly **machine learning via trial & error with rewards** (e.g. learn chess by rules and playing games)

   a) ***Chain of Thought Reasoning*** – Divide large, complex problems into intermediate steps that are easier to analyze. Showing steps helps RL and fine-tuning.

   b) ***Group Relative Policy Optimization (GRPO)*** – Learn by comparing data from several actions rather than from a two-step "policy-then-critic" model.

# V3/R1 Key Technologies

3. **Distillation** – Use of a larger "teacher" model to transfer outputs to train a smaller "student" model, e.g. on how to answer specific types of questions or problems or play chess without doing the billions of computations itself.

   a) **V3 base model used to build R1,** then R1 to improve V3, and V3 to improve R1. **Avoided expensive training & compute time**, with faster similar results.

   b) Smaller R1 versions built using **distillations of open-source Qwen 2.5 (Alibaba) & Llama 3 (Meta)**.

# Current Trends

- **2020-2023:** GenAI R&D (e.g., OpenAI) relied on ***expensive, time-consuming human pretraining*** with powerful/expensive hardware **(Nvidia!!)**

- **2023-2025:** R&D shifting to ***more economical machine-based reinforcement learning***, with smaller LLMs, less powerful/expensive hardware.

➢ **DeepSeek a great example of this trend!**

(Dario Amodei, Anthropic co-founder & former OpenAI VP)