Ambarella
AI envisioned™

GenAI's Move to the Edge

Yutao Feng
China General Manager
Ambarella

First Things First: Define the Terms First

GenAI = Generative AI, but here, refers to current SOTA

- Generative: what is generated?
- Auto-regressive next token prediction, transformer based

AGI, singularity

AI on Cloud vs. Edge and Device

- Centralized Compute, for large model training or mass parallel inferences
- Edge: de-centralized, on-premise AI compute
- Device: IoT sensors, e.g. cameras

Recent Advances (2012-2024)

Convolutional Neural Network (CNN)

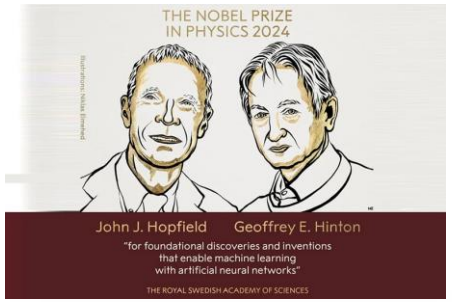
AlexNet
ImageNet

2012

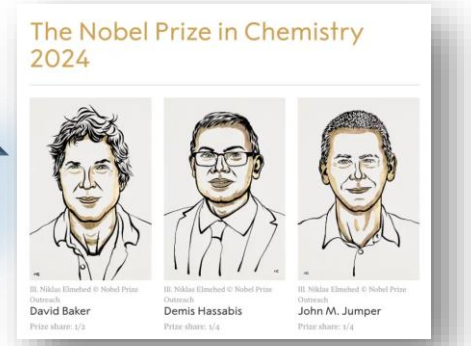
Transformer

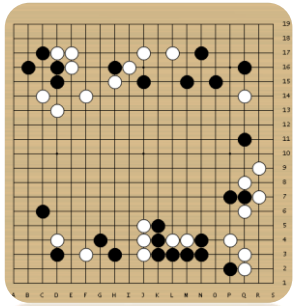
Attention is All You Need!
Google

2017



2024





2016

AlphaGo

DeepMind
Google

2022

ChatGPT 3.5

OpenAI
Microsoft

Roads to AGI

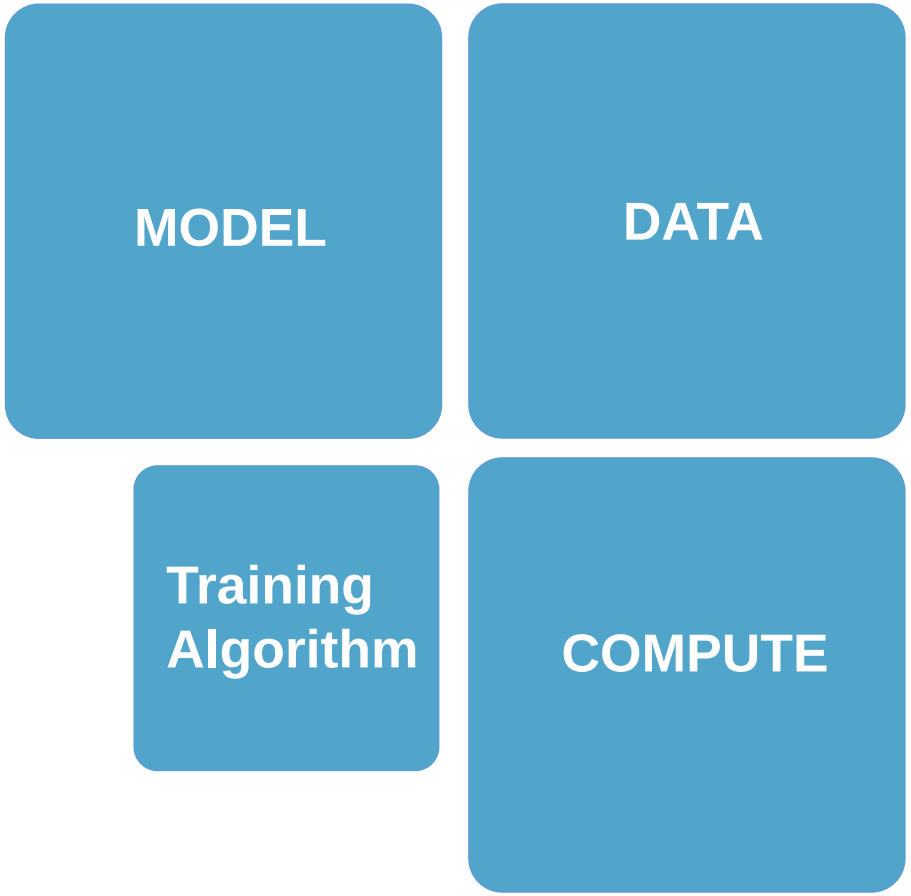
- **How to get there?**
  - **Believers of the “scaling law”: scale everything up, intelligence will “emerge”**
  - **OpenAI, Microsoft, Google, Meta…**
  - **Tencent, ByteDance, Ali, Baidu**

DeepMind/Google

- **AlphaGo (2016)**
- **AlphaFold 1/2/3 (2020 – 2024)**
- **AlphaChip (2020 – 2024)**
- **AlphaProteo, AlphaProof, AlphaGeometry…**

Challenges

- **Resource drain: only for the super rich/big**
- **Government regulations**
- **Safety: political, cultural, IP**
- **Doubt: will AGI happen down the road?**



Move to the Edge

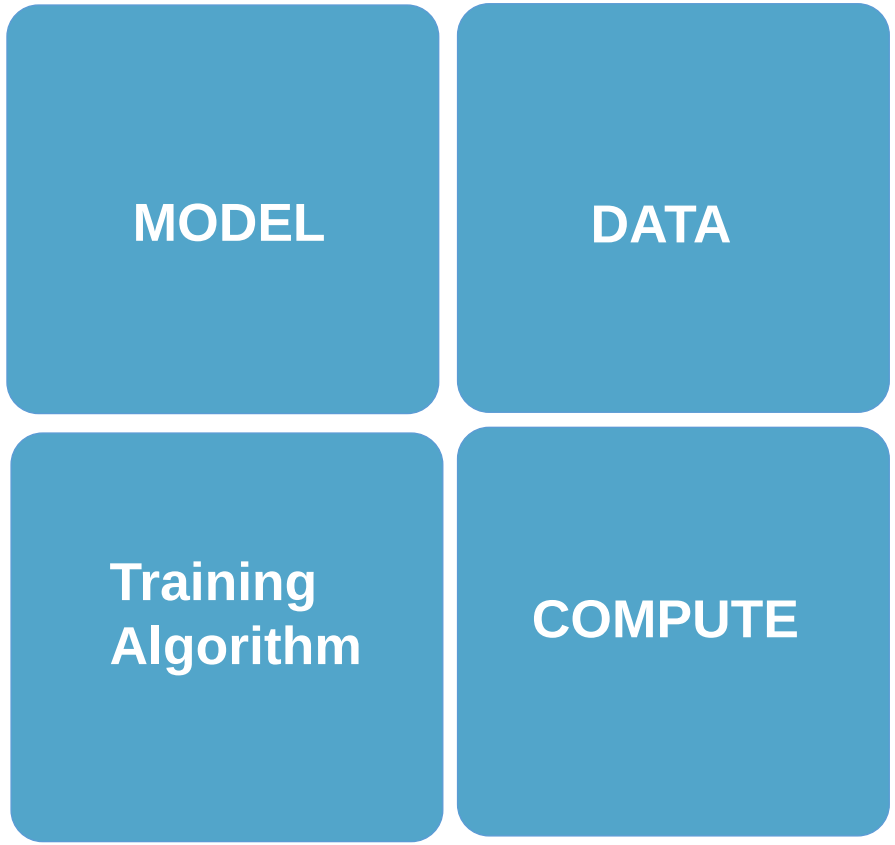
- **Cannot handle the “big game”**
- **Do not believe the current “LLM based” approach**

“LLM based approach” does not work

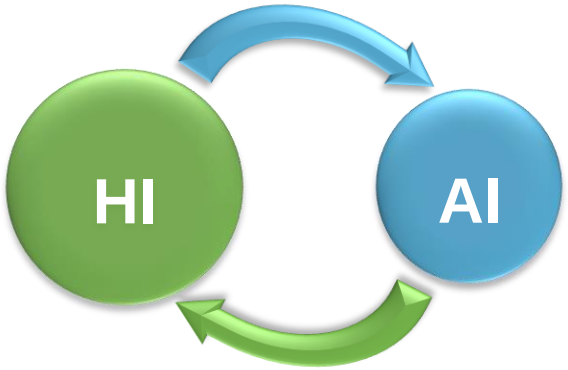
- Limitation due to the “auto-regressive next token prediction” approach
- Multimodal is still based on language model
- Lack of reasoning, lack of real physical world understanding
- Lack of transparency

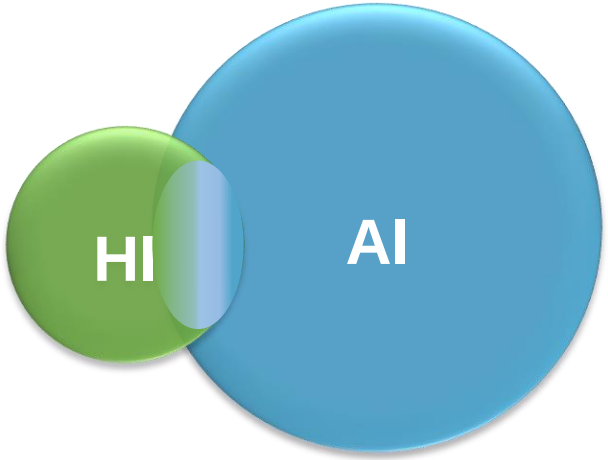
Challenges (questions)

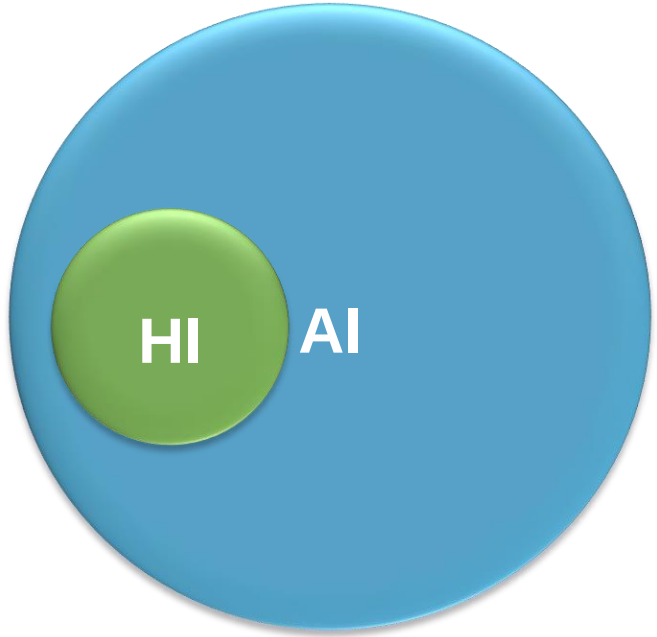
- World model? Human-like
- Training/Inference together
- How to survive the long journey
- What is the right chip for this?



HI vs. AI







HI (Biological neural network)

- Multimodal: vision, hearing
- Able to generalize, reason
- Power efficient (relatively)
- Clearly bounded by:
  - Low data I/O bandwidth
  - Sustained high computing load
  - Life span
  - Energy conversion efficiency

AI (Si + Software)

- Today: language based (LLM)
- Does not understand the physical world (yet)
- Ability to form massive parallel computing server
- Unlimited memory and high data bandwidth
- Ability to reproduce at high rate

Why We Need AI on the Edge/Device

We need AI to work for us, not to dominate us (without telling us)

Edge and device AI, such as robots, will “live” in the environment and infrastructure that are designed for human

AI as a tool: to extend human reach

On-device training: individualized AI, device can “grow smarter”

$$W_i + \Delta W_i \rightarrow W_i'$$

$$(W + V_i) + \Delta V_i \rightarrow (W + V_i')$$

Human intelligence is on the “Edge”

HI is still dominating the earth, until ...

HI builds the AI, until...

Will AI keep HI safe, until…

Edge AI Needs the Right Hardware

Sensing and perception

Edge computing hardware:

- Low Power

Interfaces and bandwidth

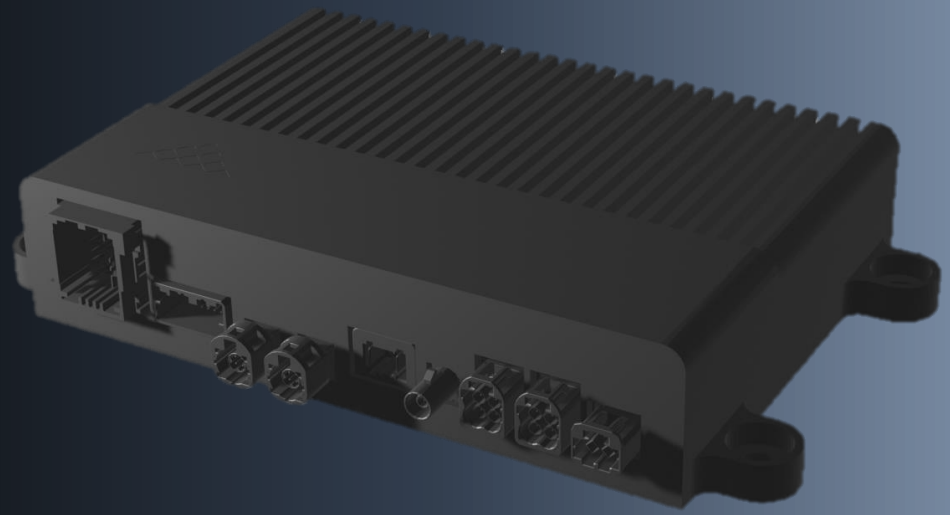
- Low latency, Data privacy, Local Memory

Edge AI Box

Powered by Ambarella

Cooper™ Max

Cooper™ Mini



Ambarella Enabling the Move to the Edge

Software SDK

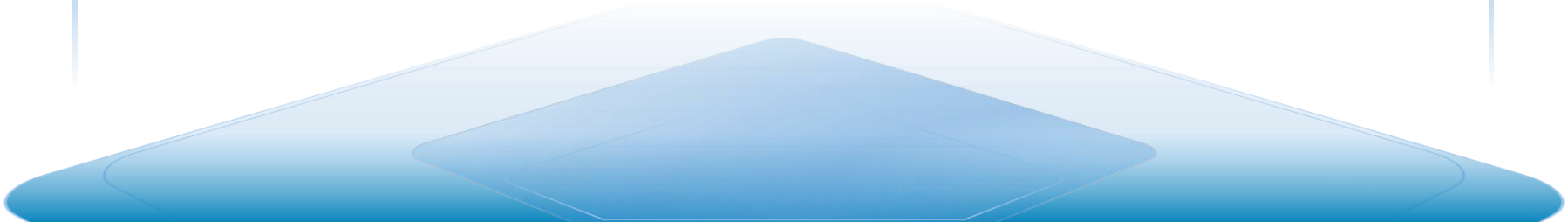
Low Power

“Algorithm First” Approach

Mature Toolchain

Ambarella SoC

Superior Image Quality



Ambarella
AI envisioned™

Thank You

www.ambarella.com