AI Slides Viewer & Checker

Sunghee Yun

Co-founder / CTO - AI Technology & Product Strategy

Erudio Bio, Inc.

Table of contents

About Speaker

- *Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio, CA, USA*
- Advisory Professor, Electrical Engineering and Computer Science @ DGIST
- Adjunct Professor, Electronic Engineering Department @ Sogang University
- Technology Consultant @ Gerson Lehrman Gruop (GLG), NYC, USA
- *Co-founder / CTO & Chief Applied Scientist @ Gauss Labs, CA, USA – 2023*
- Senior Applied Scientist @ Amazon, Vancouver, Canada – 2020
- Principal Engineer @ Software R&D Center of DS Division - Samsung – 2017
- Principal Engineer @ Strategic Marketing Team of Memory Business Unit – 2016
- Principal Engineer @ DT Team of DRAM Development Lab. - Samsung – 2015
- Senior Engineer @ CAE Team - Samsung – 2012
- M.S. & Ph.D. - Electrical Engineering @ Stanford University – 2004
- B.S. - Electrical Engineering @ Seoul National University – 1998

Highlight of career journey

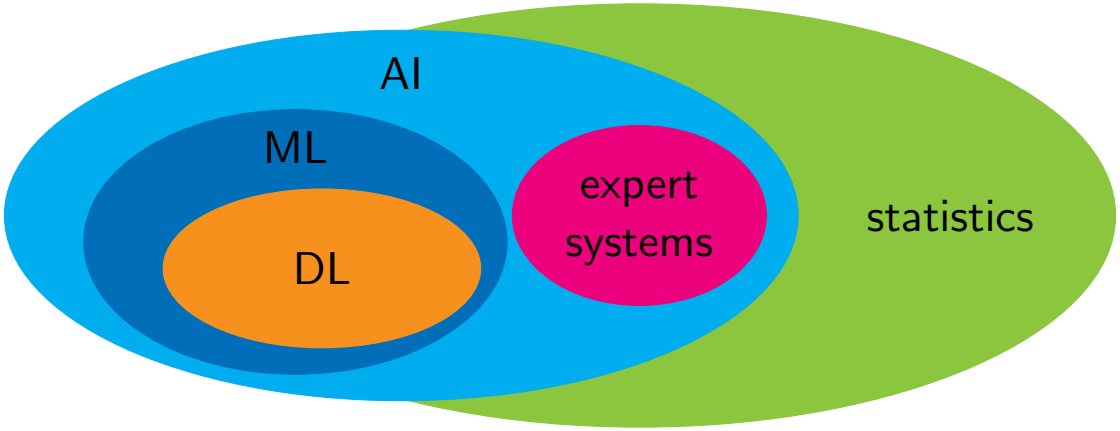
- B.S. in EE @ SNU, M.S. & Ph.D. in EE @ Stanford Univ.
  - *Convex Optimization - theory / algorithms / applications - under supervision of Prof. Stephen P. Boyd*
- Principal Engineer @ Memory Design Technology Team
  - AI & optimization partnering with *DRAM/NAND Design/Process/Test teams*
- Senior Applied Scientist @ Amazon
  - *S-Team Goal (Bezos's) project - better customer shopping experience via Amazon shopping app using AI - increased sales by $200M*
- Co-founder / CTO & Chief Applied Scientist @ Gauss Labs
  - *R&D industrial AI products & technology, market/product/investment strategies*
- Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio
  - *biotech - AI technology & product strategy*

Artificial Intelligence

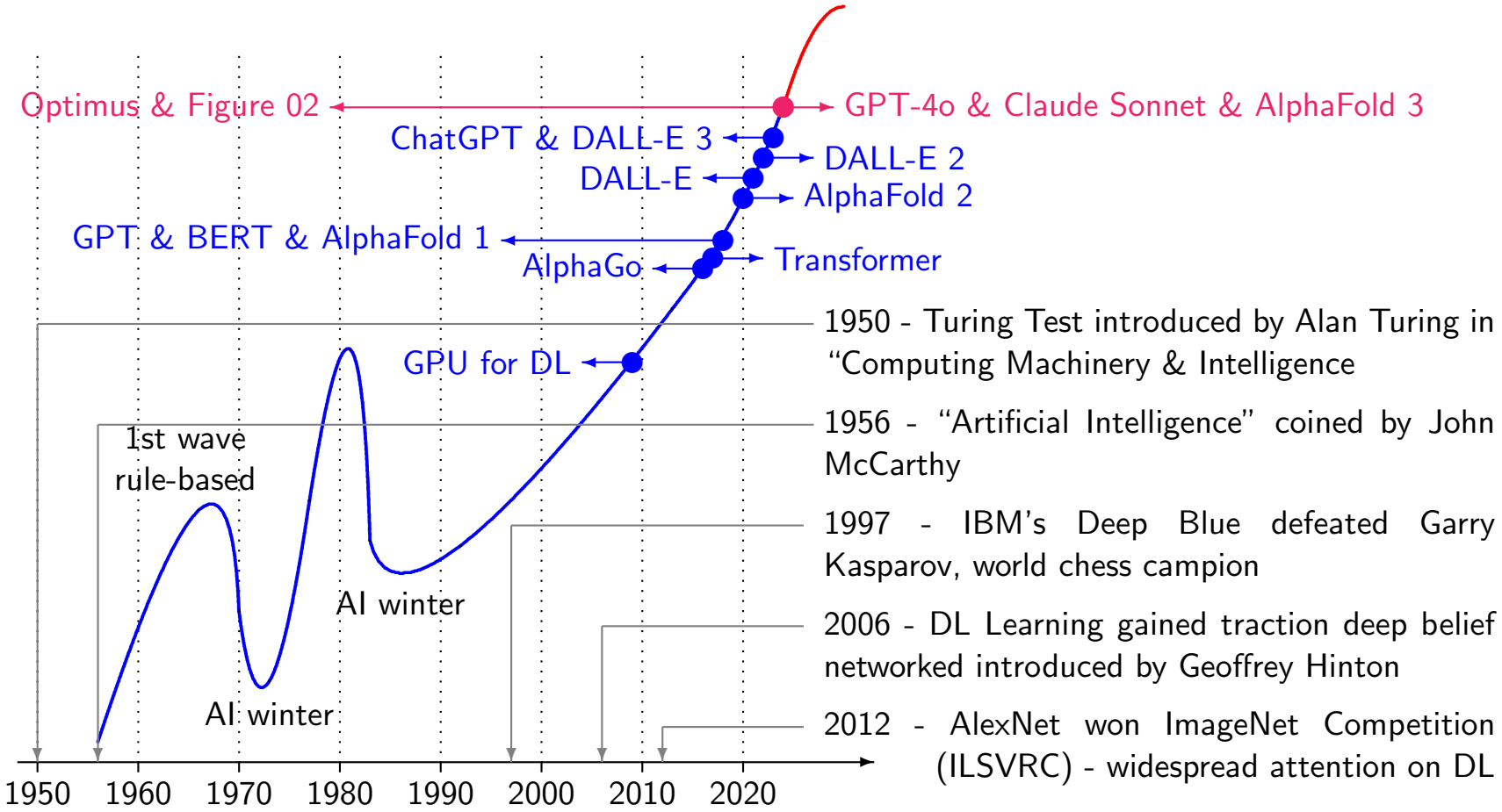
Definition and History

Definition of AI

- AI is
  - technology enabling machines to do tasks requiring human intelligence, such as learning, problem-solving, decision-making & language understanding
  - *not* one thing - encompass range of technologies, methodologies & applications
- relationship of AI, statistics, ML, DL, NN & expert system [HGH+22]



History of AI



Significant AI Achievements - 2014 – 2024

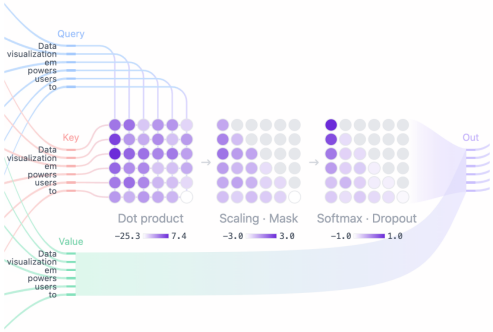
Deep learning revolution

- 2012 – 2015 - Deep Learning Revolution[1]
  - CNNs demonstrated exceptional performance in image recognition, $e.g.$, *AlexNet's victory in ImageNet competition*
  - widespread adoption of DL learning in CV transforming industries
- 2016 - AlphaGo Defeats Human Go Champion
  - DeepMind's AlphaGo defeated world champion in Go, extremely complex game *believed to be beyond AI's reach*
  - significant milestone in RL - AI's potential in solving complex & strategic problems

[1] DL: deep learning, CNN: convolutional neural network, CV: computer vision, RL: reinforcement learning

Transformer changes everything

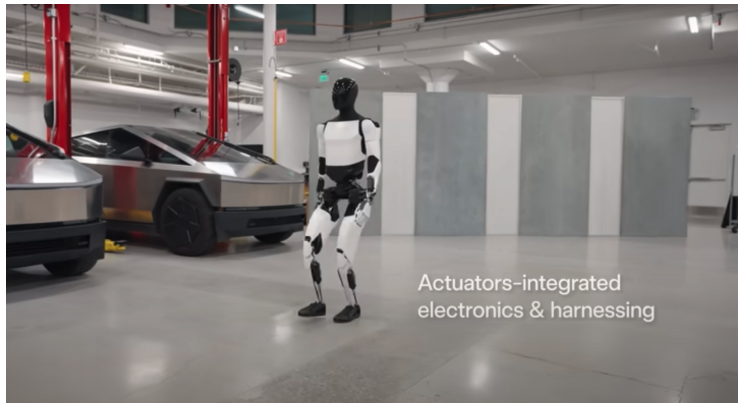
- 2017 – 2018 - Transformers and NLP breakthroughs[2]
  - *Transformer (e.g., BERT & GPT) revolutionized NLP*
  - major advancements in, *e.g.*, machine translation & chatbots
- 2020 - AI in Healthcare – AlphaFold and Beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting 3D protein structures with remarkable accuracy
  - accelerates drug discovery and personalized medicine - offering new insights into diseases and potential treatments



[2] NLP: natural language processing GPT: generative pre-trained transformer

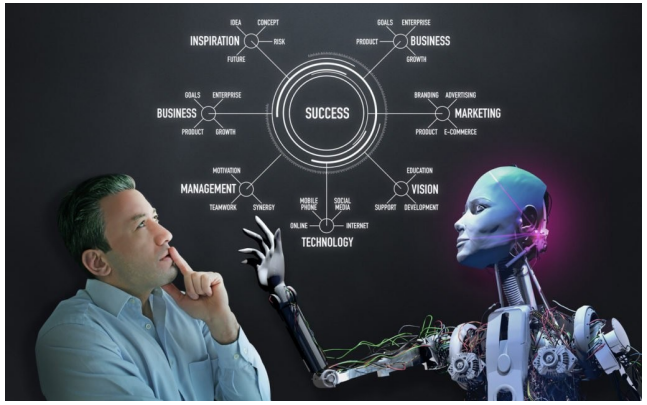
Lots of breakthroughs within 6 months in 2024

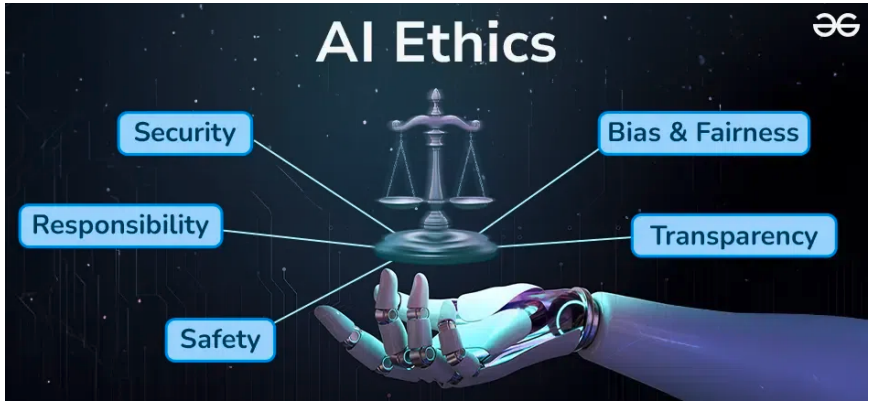
- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Llama 3, Sora
  - *transforming industries* such as content creation, customer service, education, $etc.$
- breakthroughs in specialized AI applications
  - Figure 02, Optimus, AlphaFold 3
  - driving unprecedented advancements in automation, drug discovery, scientific understanding - *profoundly affecting healthcare, manufacturing, scientific research*



Transformative impact of AI - reshaping industries, work & society

- accelerating human-AI collaboration
  - not only reshaping industries but *altering how humans interact with technology*
  - AI's role as collaborator and augmentor redefines productivity, creativity, the way we address global challenges, $e.g.$, *sustainability & healthcare*
- AI-driven automation *transforms workforce dynamics* - creating new opportunities while challenging traditional job roles
- *ethical AI considerations* becoming central not only to business strategy, but to society as a whole - *influencing regulations, corporate responsibility & public trust*

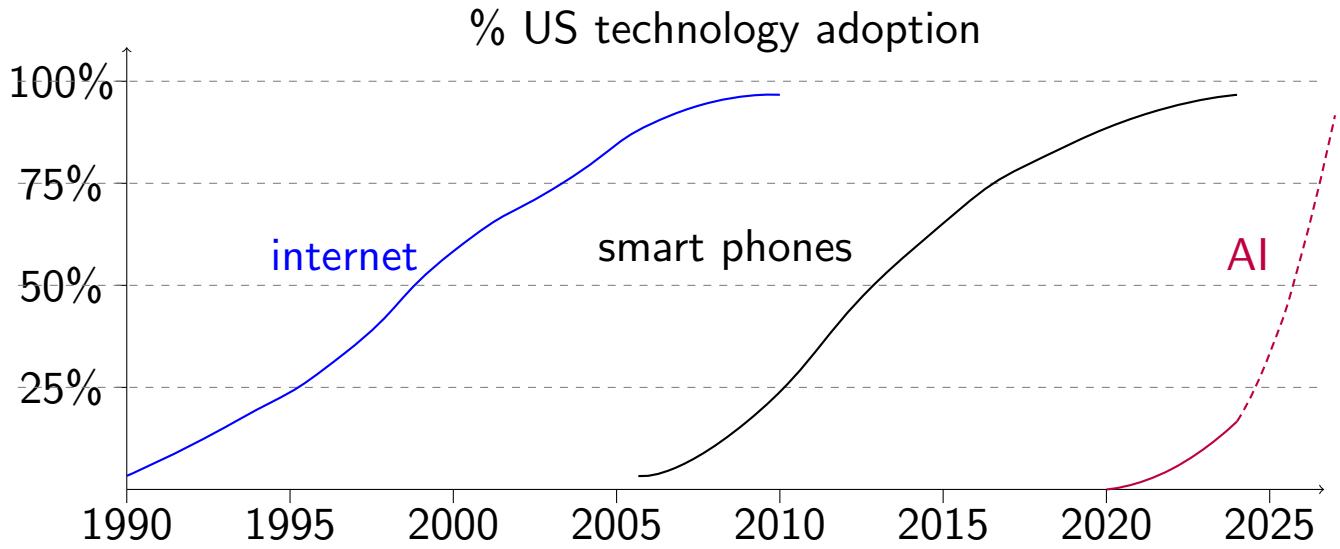




Recent Advances in AI

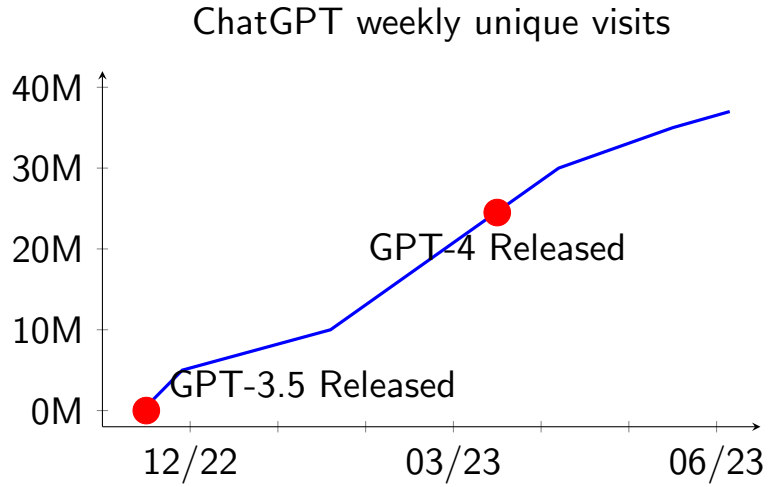
Where are we in AI today?

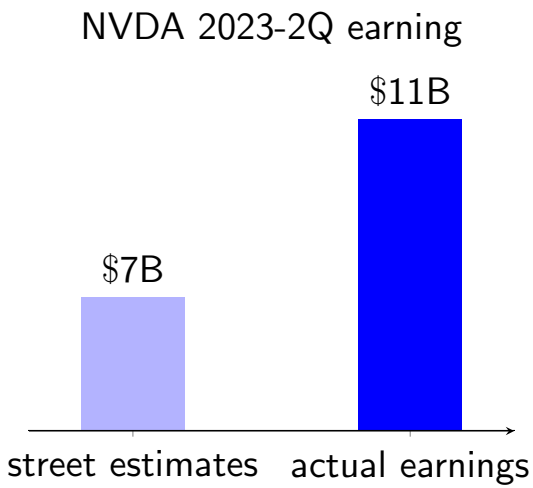
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



Explosion of AI ecosystems - ChatGPT & NVIDIA

- took only *5 months for ChatGPT users to reach 35M*
- NVDIA 2023 Q2 earning exceeds market expectation by big margin - $7B vs $13.5B
  - surprisingly, *101% year-to-year growth*
  - even more surprisingly *gross margin was 71.2%* - up from 43.5% in previous year[3]

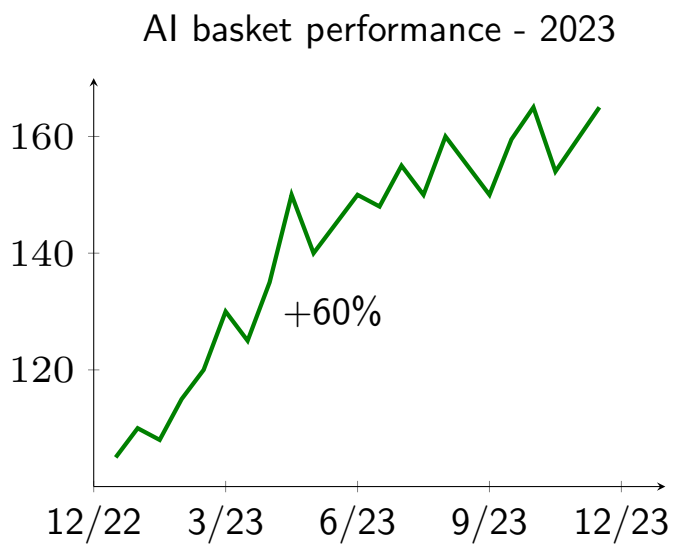


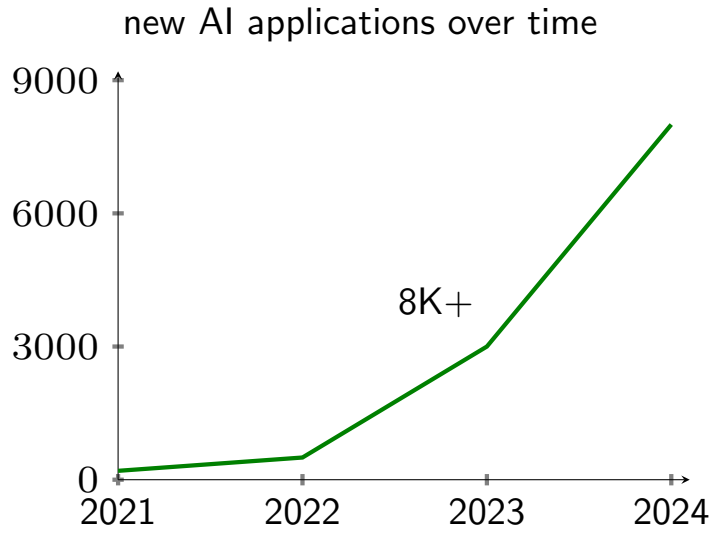


[3] source - Bloomberg

Explosion of AI ecosystems - AI stock market

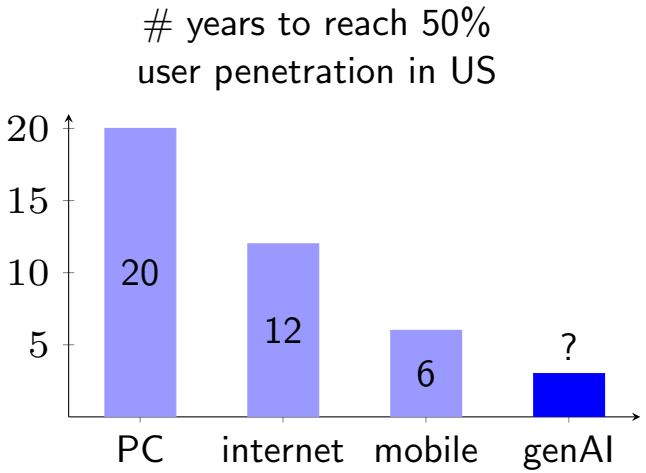
- *AI investment surge in 2023 - portfolio performance soars by 60%*
  - AI-focused stocks significantly outpaced traditional market indices
- *over 8,000 new AI applications* developed in last 3 years
  - applications span from healthcare and finance to manufacturing and entertainment

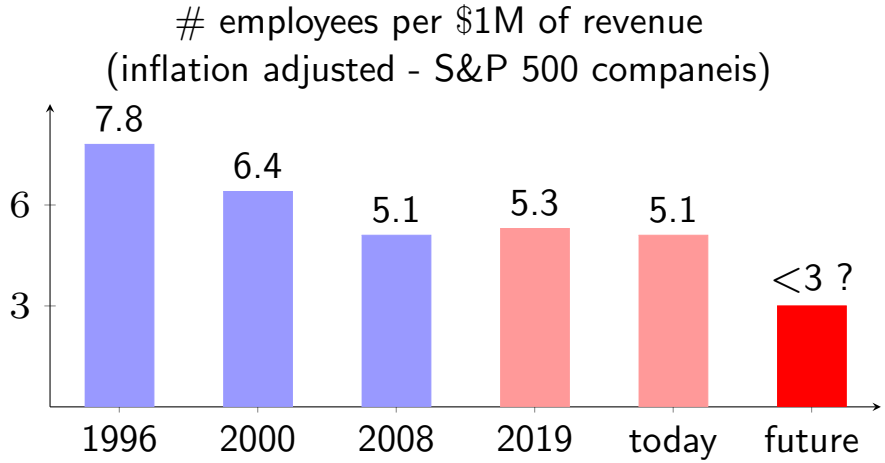




AI's transformative impact - adoption speed & economic potential

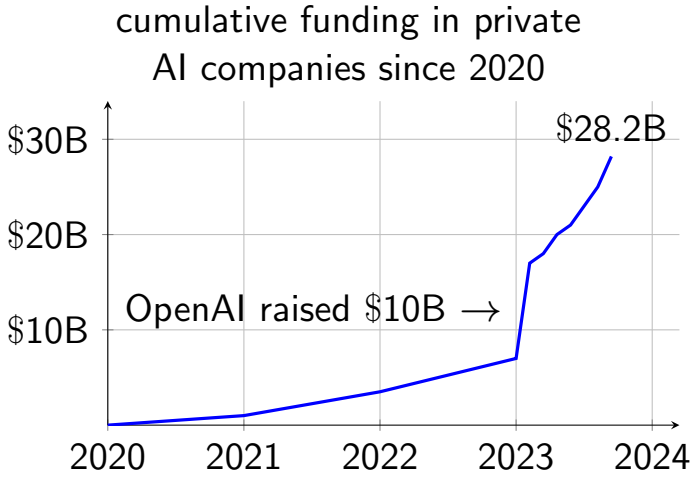
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI's potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater gains expected with AI proliferation

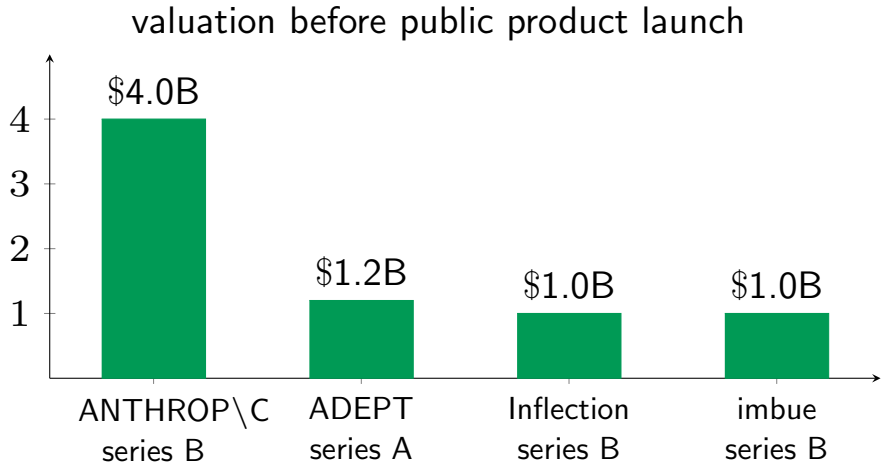




Massive investment in AI

- *explosive growth* - cumulative funding skyrocketed reaching staggering $28.2B
- OpenAI - significant fundraising (=$10B) fueled rapid growth
- *valuation surge* - substantial valuations even before public products for stella companies
- *fierce competition for capital* among AI startups driving innovation & accelerating development
- massive investment indicates *strong belief in & optimistic outlook for potential of AI* to revolutionize industries & drive economic growth





AI Market & Values

Fiber vs cloud infrastructure

- fiber infrastructure - 1990s
  - Telco Co's raised $1.6T of equity & $600B of debt
  - bandwidth costs decreased 90% within 4 years
  - companies - Covage, NothStart, Telligent, Electric Lightwave, 360 networks, Nextlink, Broadwind, UUNET, NFS Communications, Global Crossing, Level 3 Communications
  - became *public good*
- cloud infrastructure - 2010s
  - entirely new computing paradigm
  - mostly public companeis with data centers
  - *big 4 hyperscalers generate* $150B + annual revenue

Cloud stacks

- SaaS dominates cloud stack - account for 40% of total cloud stack market with estimated TAM of $260B
- IaaS and PaaS significant players
- semi-cloud's niche presence

cloud stack	companies	estimated TAM	% total in stack
SaaS apps	Salesforce, Adobe	$260B	40%
PaaS	Confluent, snowflake	$140B	22%
IaaS	AWS, Azure, GCP	$200B	30%
cloud semis	AMD, Intel	$50B	8%

AI stacks

- AI investment landscape - AI sector witnessing significant capital inflow with total funding of approximately $29 billion across various segments
- models lead pack - AI models, particularly those developed by OpenAI and Anthropic, attracted lion's share of investments, accounting for 60% of total funding
- diverse growth - while models dominate funding, other segments like apps, AI cloud, and AI semis also experiencing substantial growth, indicating broadening AI ecosystem

AI stack	companies	total funding	% total in stack
apps	character.io, replit	∼$5B	17%
models	openAI, ANTHROP\C	∼$17B	60%
AIops	Hugging Face, Weights & Biases	∼$1B	4%
AI cloud	databricks, Lambda	∼$4B	13%
AI semis	cerebras, SambaNova	∼$2B	6%

AI model companies

- AI model companies - competing for which AI model companies will dominate 2020s
- venture funding surge - private AI model companies raised approximately $17B since 2020, indicating strong investor confidence
- growing open-source presence - becoming increasingly prevalent, adding competition and innovation to AI landscape
- key players - notable companies in AI model space include Adept, OpenAI, Anthropic, Imbue, Inflection, Cohere, and Aleph Alpha
- outcome uncertain - future success is still to be determined, reflecting dynamic and evolving nature of AI industry

AI advancing much faster

- rapid AI advancement - general AI projected to progress from basic content generation to superhuman reasoning in only 5 years
- significantly outpacing 15-year timeline for fully autonomous vehicles

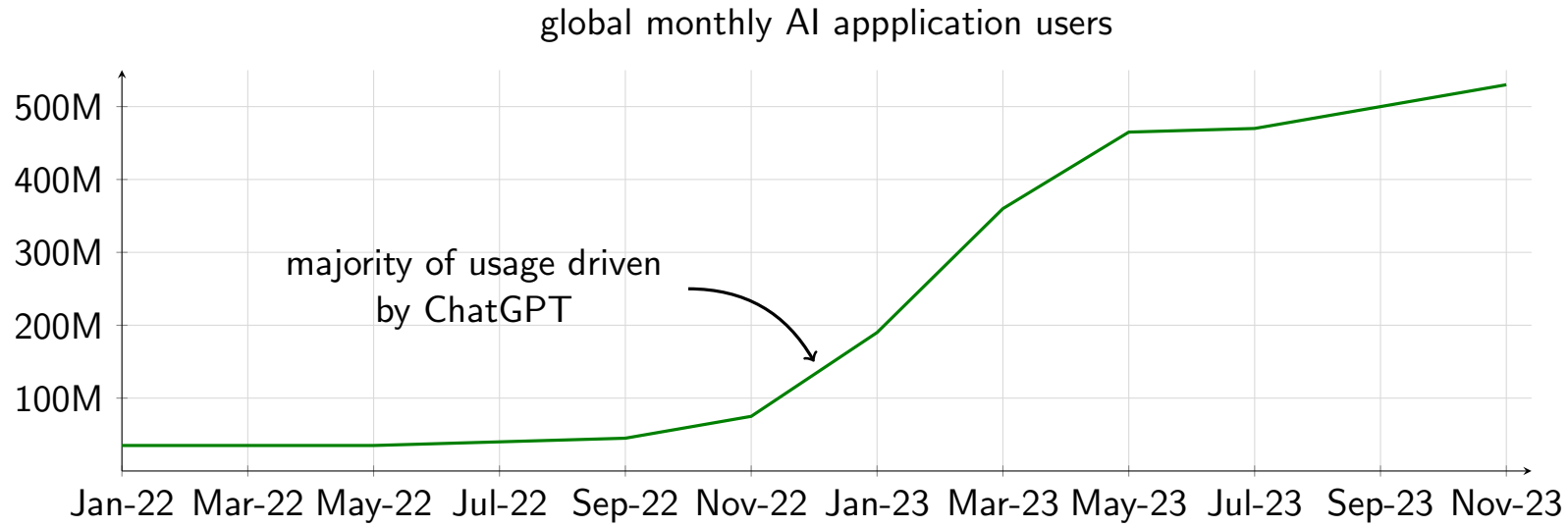
autonomy level	autonomous vehicles	genAI
L5	fullly autonomous	superhuman reasoning & perception
L4	highly autonomous	AI autopilot for complex tasks
L3	self-driving with light intervention	AI co-pilot for skilled labor
L2	Tesla autopilot	supporting humans with basic tasks
L1	cruise control	generating basic content

15 yrs

5 yrs

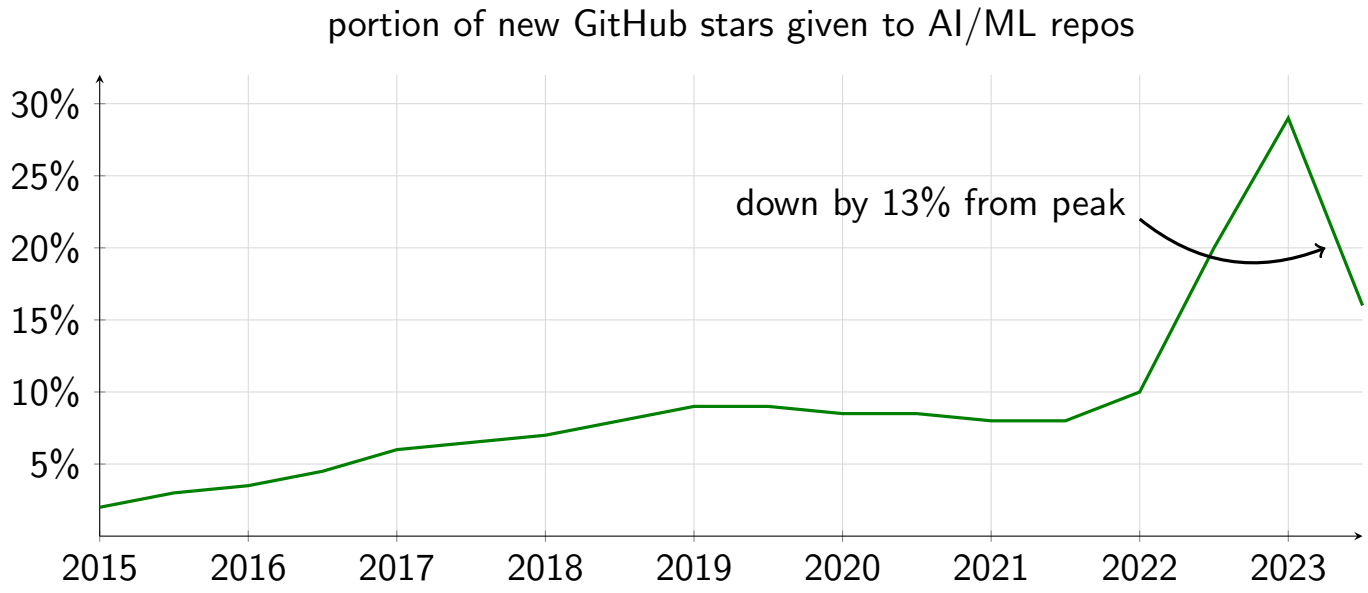
AI interest of users

- AI adoption approaching saturation - initial wave may be nearing saturation
- future growth might come from deeper integration into professional workflows & specialized applications
- potential for market diversification - ChatGPT drove majority of early growth, but now we have other LLMs - Claude, Mistral, Gemini, Grok, Perplexity



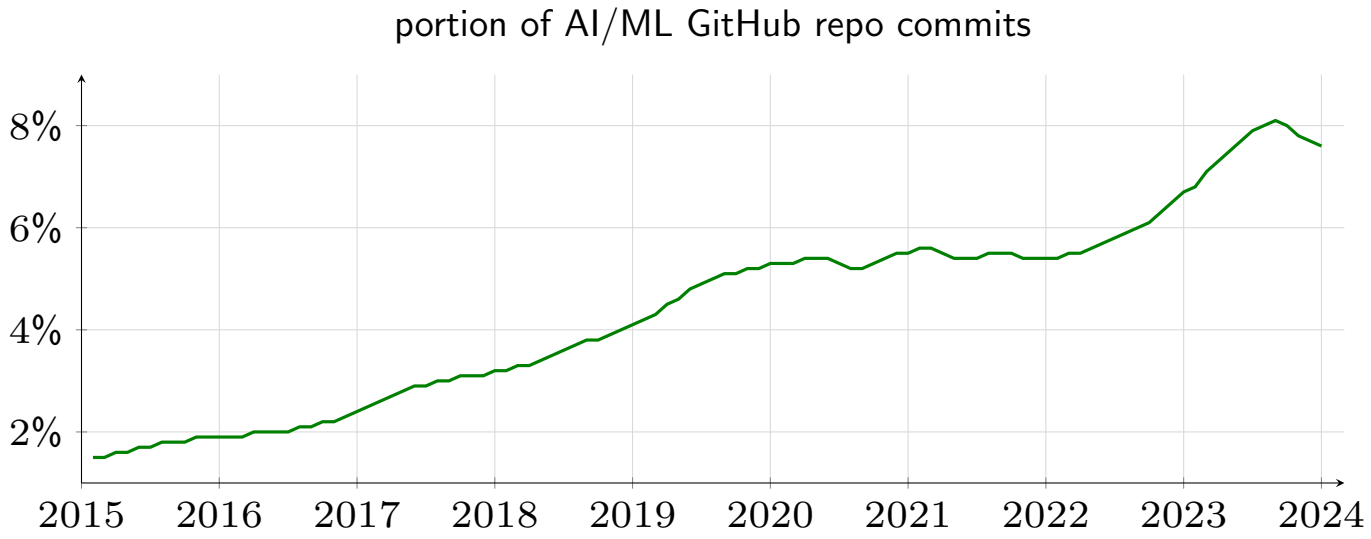
AI interest of developers

- rising popularity - portion of new GitHub stars given to AI/ML repositories steadily increased from 2015 to 2022
- excitement waning & washing out AI "tourists" - decline of 13% from peak in 2022
- could indicate potential factors such as market saturation, economic conditions, or shifts in developer preferences



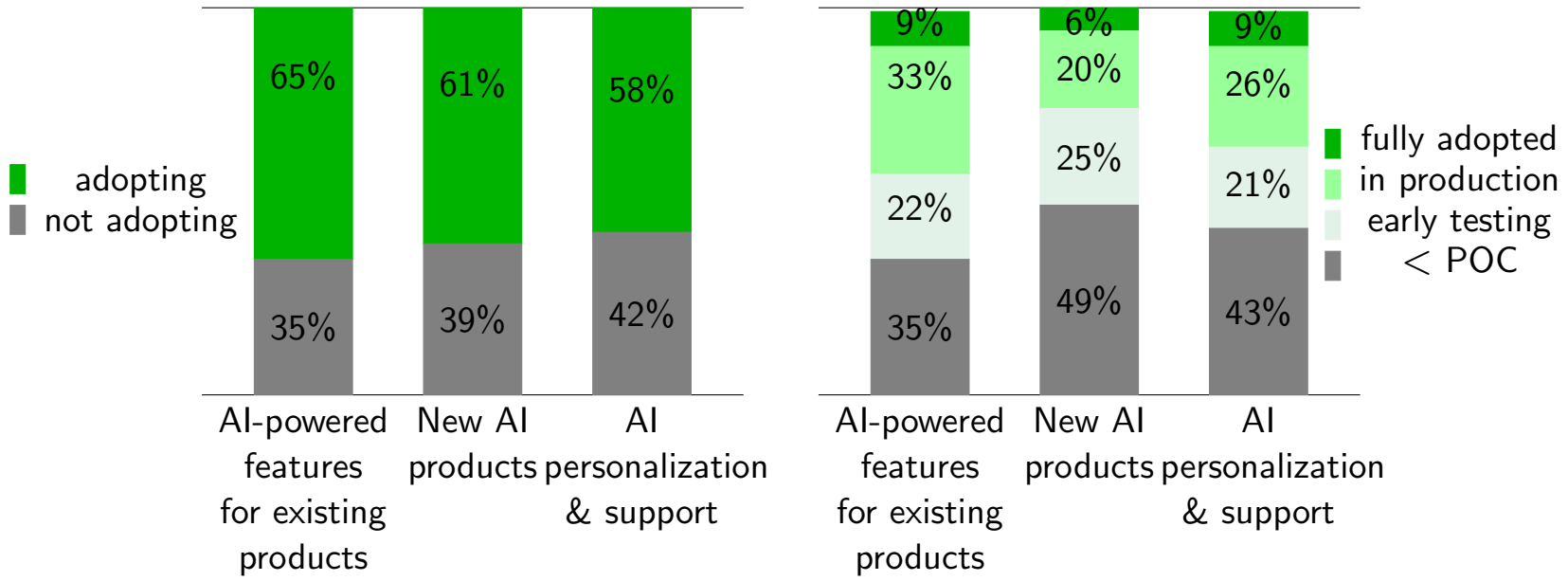
Developers' contribution to software packages

- steep acceleration from 2022 to 2024 correlates with explosion of LLMs & genAI
- suggesting transformative shift in AI landscape beyond gradual growth
- AI/ML still represents relatively small portion (less than 10%)
- indicating significant room for growth and mainstream adoption across various software domains



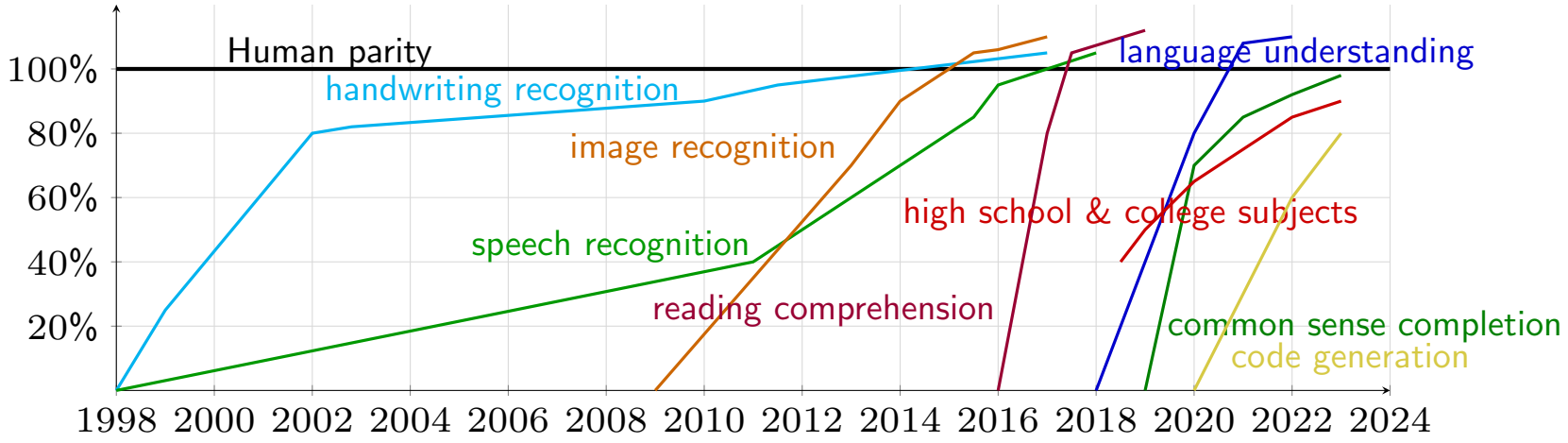
Enterprises adoptiong AI

- more than 60% of enterprises planning to adopt AI
- full adoption rate is less than 10% - will take long time



AI getting better and faster

- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached human parity, while others having still way to go
  - achieving truly human-like capabilities in broad range remains a challenge

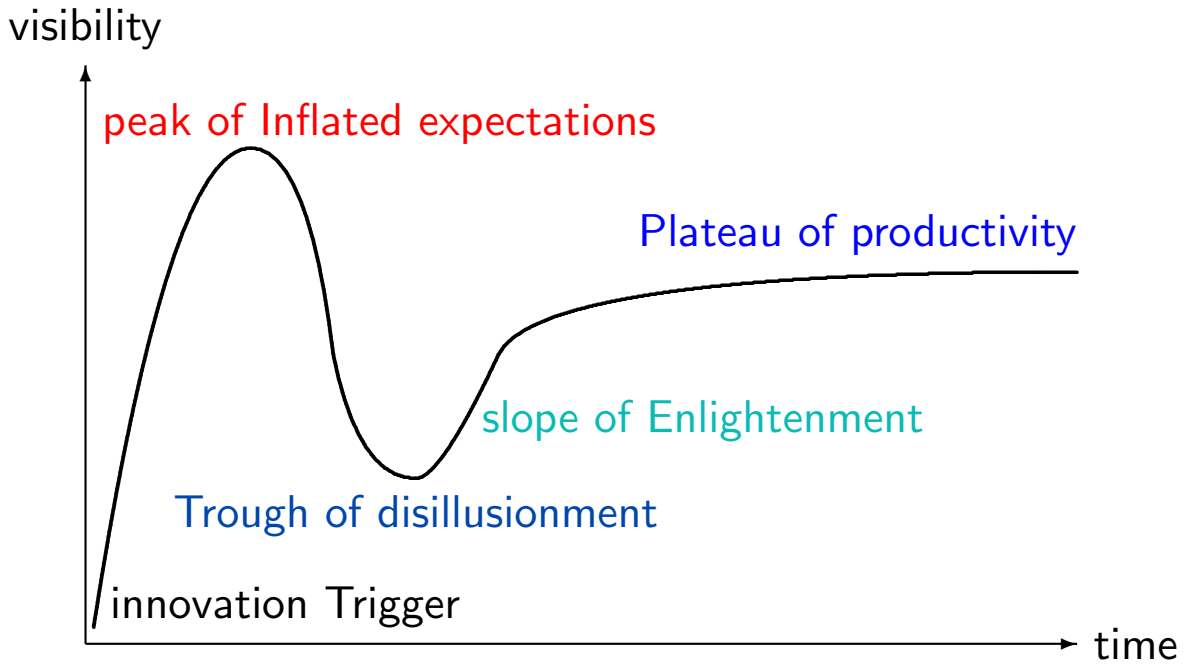


AI delivers game-changing values

- time developers save using GitHub Copilot - *55%*
  - *10M+* cumulative downloads as of 2024 & *1.3M* paid subscribers - *30%* Q2Q increase
  - improves developer productivity by *30%+*
- reduction in human-answered customer support requests - *45%*
  - cost per support interaction - *95%* save / $2.58 (human) vs $0.13 (AI)
  - median response time - *44 min* faster / 45 min (human) vs 1 min (AI)
  - median customer satisfaction - *14%* higher / 55% (human) vs 69% (AI)
- time saved from editing video in runway - *90%*
- AI chat rated higher quality compared to physician responses - *79%*

Is AI hype?

Technology hype cycle



- innovation trigger - technology breakthrough kicks things off
- peak of inflated expectations - early publicity induces many successes followed by even more
- trough of disillusionment - expectations wane as technology producers shake out or fail
- slope of enlightenment - benefit enterprise, technology better understood, more enterprises fund pilots

Yes & No

characteristics of hype cycles	speaker's views
value accrual misaligned with investment	• OpenAI still operating at a loss; business model *still* not clear • gradual value creation across broad range of industries and technologies ($e.g.$, CV, LLMs, RL) unlike fiber optic bubble in 1990s
overestimating timeline & capabilities of technology	• self-driving cars delayed for over 15 years, with limited hope for achieving level 5 autonomy • AI, however, has proven useful within a shorter 5-year span, with enterprises eagerly adopting
lack of widespread utility due to technology maturity	• AI already providing significant utility across various domains • vs quantum computing remains promising in theory but lacks widespread practical utility

AI Research

AI research race gets crazy

- practically impossible to follow all developments announced everyday
  - new announcement and publication of important work everyday!
- *industry leads research - academia lags behind*
  - trend observed even before 2015
- everyone excited to show off their work to the world - conference and `github.com`
  - biggest driving force behind unprecedented scale and speed of advancement of AI together with massive investment of capitalists

AI progress within a month - March, 2024

- UBTECH Humanoid Robot Walker S: Workstation Assistant in EV Production Line
- H1 Development of dance function
- Robot Foundation Models (Large Behavior Models) by Toyota Research Institute (TRI)
- Apple Vision Pro for Robotics
- Figure AI & OpenAI
- Human modeling
- LimX Dynamics' Biped Robot P1 Conquers the Wild Based on Reinforcement Learning
- HumanoidBench: Simulated Humanoid Benchmark for Whole-Body Locomotion and Manipulation - UC Berkeley & Yonsei Univ.
- Vision-Language-Action Generative World Model
- RFM-1 - Giving robots human-like reasoning capabilities

Papers of single company accepted by single conference

- CVPR 2024
  - PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar - MIT, Codec Avatars Lab, & Meta [KXS+24]
    - 3D reconstruction from single-view
  - Nymeria Dataset
    - large-scale multimodal egocentric dataset for full-body motion understanding
  - Relightable Gaussian Codec Avatars - Codec Avatars Lab & Meta [SSS+24]
    - build high-fidelity relightable head avatars being animated to generate novel expressions
  - Robust Human Motion Reconstruction via Diffusion (RoHM) - ETH Zürich & Reality Labs Research, Meta [ZBX+24]
    - robust 3D human motion reconstruction from monocular RGB videos

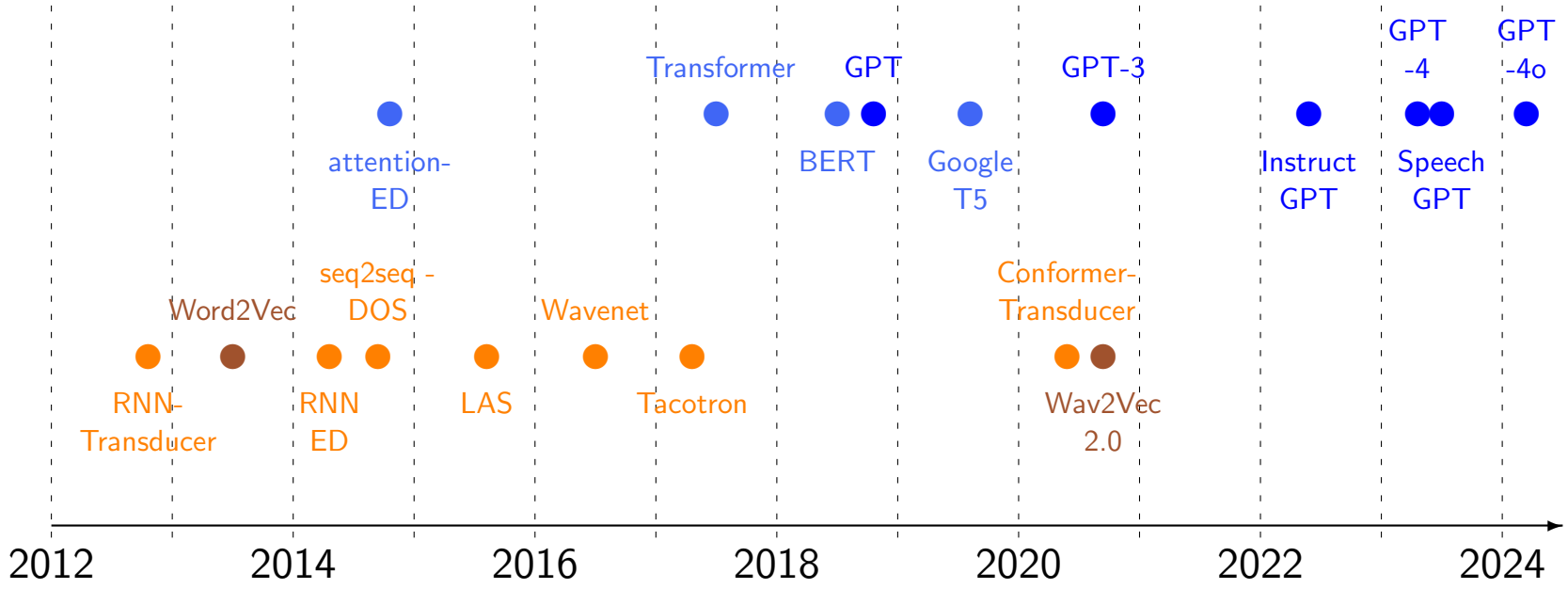
LLM

Language Models

History of language models

- bag of words - first introduced – 1954
- word embedding – 1980
- RNN based models - conceptualized by David Rumelhart – 1986
- LSTM (based on RNN) – 1997
- 380M-sized seq2seq model using LSTMs proposed – 2014
- 130M-sized seq2seq model using gated recurrent units (GRUs) – 2014
- Transformer - Attention is All You Need - A. Vaswani et al. @ Google – 2017
  - 100M-sized encoder-decoder multi-head attention model for machine translation
  - non-recurrent architecture, handle arbitrarily long dependencies
  - parallelizable, *simple* (linear-mapping-based) attention model

Recent advances in speech & language processing

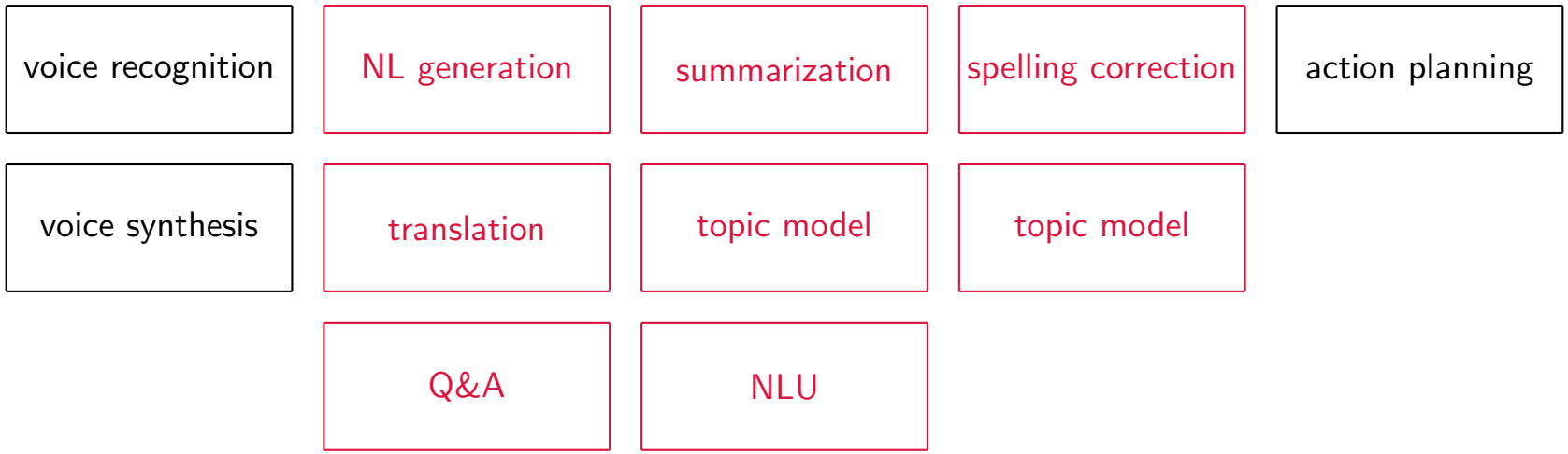


- LAS: listen, attend, and spell, ED: encoder-decoder, DOS: decoder-only structure

Types of language models

- many of language models have common requirements - language representation learning
- can be learned via pre-tranining *high performing model* and fine-tuning/transfer learning/domain adaptation
- this *high performing model* learning essential language representation *is* (lanauge) foundation model

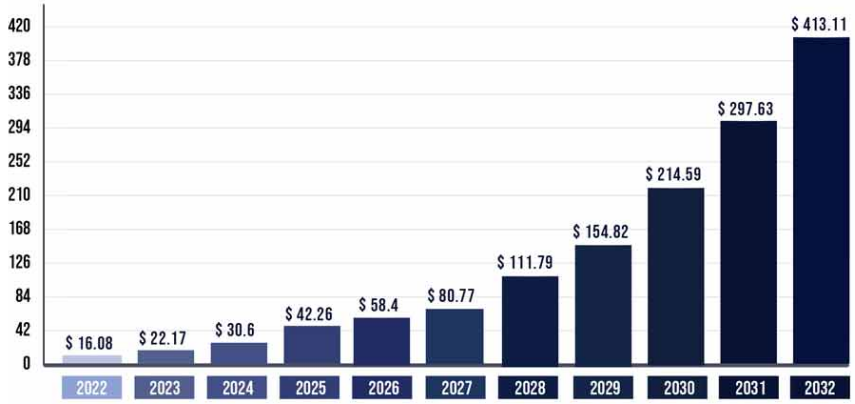
– actually, same for other types of learning, $e.g.$, CV



NLP market size

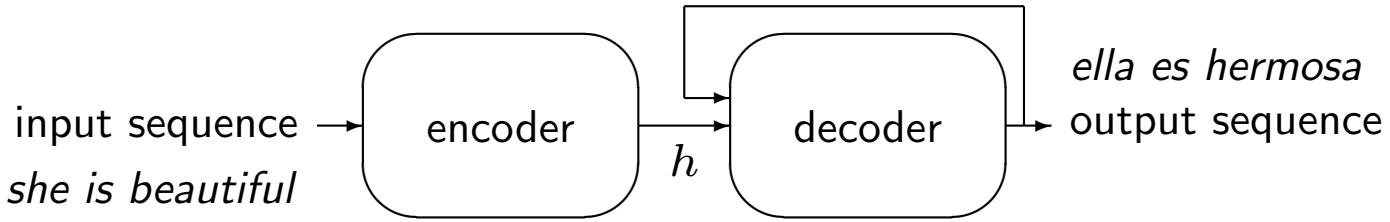
- global NLP market size estimated at USD 16.08B in 2022, is expected to hit USD 413.11B by 2032 - *CAGR of 38.4%*
- in 2022
  - north america NLP market size valued at USD 8.2B
  - high tech and telecom segment accounted revenue share of over 23.1%
  - healthcare segment held a 10% market share
  - (by component) solution segment hit 76% revenue share
  - (deployment mode) on-premise segment generated 56% revenue share
  - (organizational size) large-scale segment contributed highest market share

- source - Precedence Research



RNN-type sequence to sequence (seq2seq) model

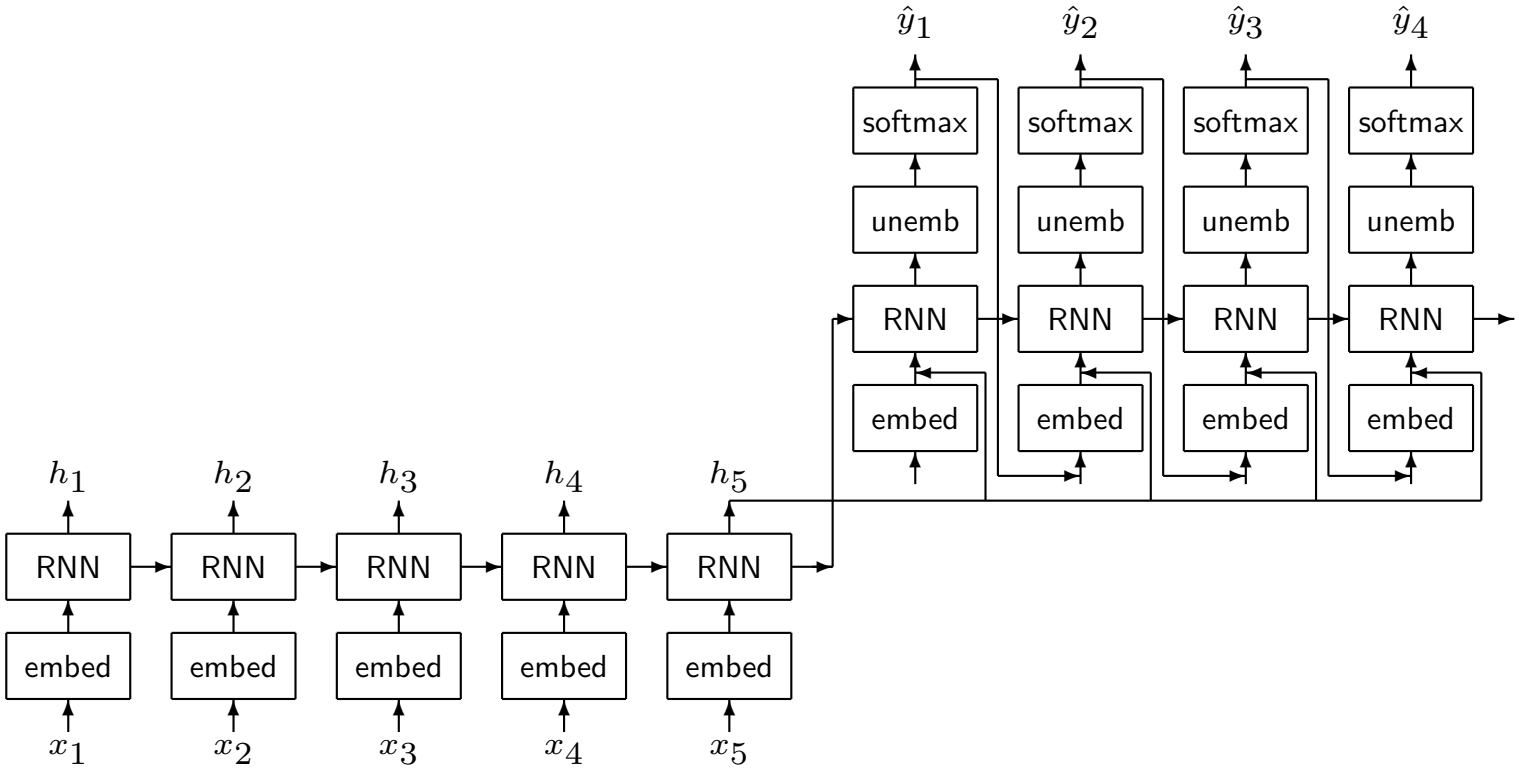
- seq2seq - take sequences as inputs and spit out sequences
- encoder-decoder architecture



  - encoder & decoder is RNN-type model
  - $h \in \mathbf{R}^n$ - hidden state - *fixed length* vector
- (try to) condense and store information of input sequence (losslessly) in (fixed-length) hidden states
  - finite hidden state - not flexible enough, $i.e.$, cannot handle arbitrarily large information
  - memory loss for long sequences
  - LSTM was promising fix, but with (inevitable) limits

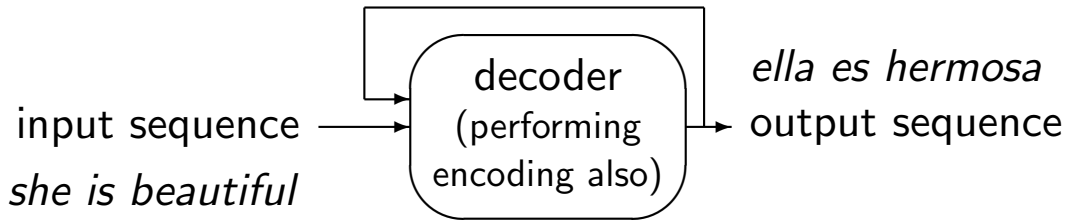
RNN-type encoder-decoder example

- RNN can be basic RNN, LSTM, GRU, $etc.$



Shared encoder/decoder model

- may use single structure to perform both encoding & decoding
- LLMs are built in this way



Large Language Models

LLM

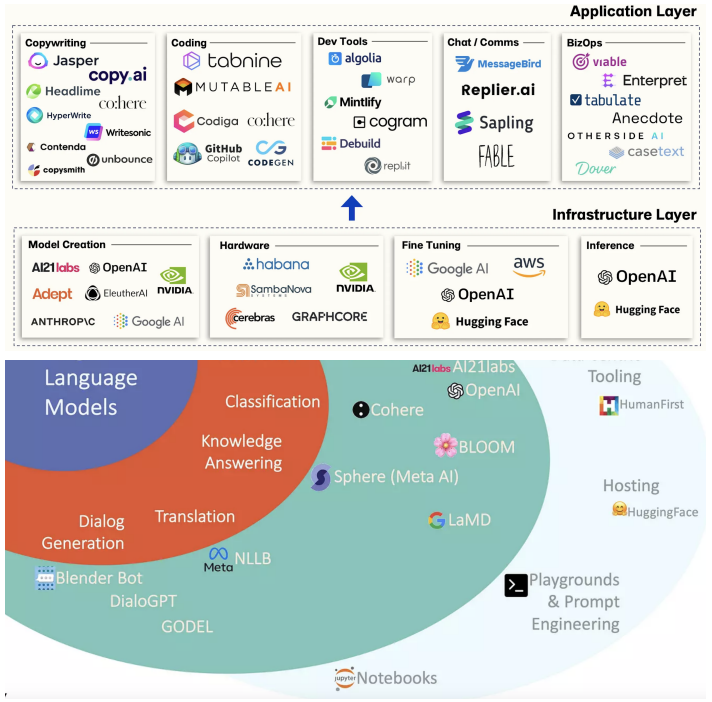
- LLM
  - type of AI aimed for NLP trained on massive corpus of texts & programming code
  - allow learn statistical relationships between words & phrases, $i.e.$, conditional probabilities
  - *amazing performance shocked everyone - unreasonable effectiveness of data (Halevry et al., 2009)*

- applications
  - conversational AI agent / virtual assistant
  - machine translation / text summarization / content creation / sentiment analysis
  - code generation
  - market research / legal service / insurance policy / triange hiring candidates

  \+ virtually infinite # of applications

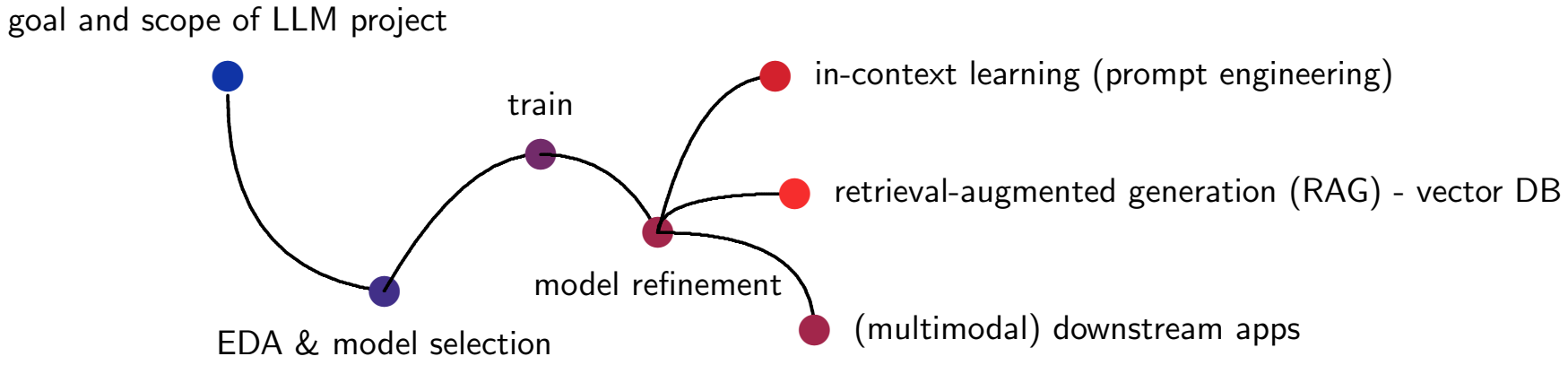
LLMs

- Foundation Models
  - GPT-x/Chat-GPT - OpenAI, Llama-x - Meta, PaLM-x (Bard) - Google
- # parameters
  - generative pre-trained transfomer (GPT) - GPT-1: 117M, GPT-2: 1.5B, GPT-3: 175B, GPT-4: 100T, GPT-4o: 200B
  - large language model Meta AI (Llama) - Llama1: 65B, Llama2: 70B, Llama3: 70B
  - scaling language modeling with pathways (PaLM) - 540B
- burns lots of cash on GPUs!
- applicable to many NLP & genAI applications



LLM building blocks

- data - trained on massive datasets of text & code
  - quality & size critical on performance
- architecture - GPT/Llama/Mistral
  - can make huge difference
- training - self-supervised/supervised learning
- inference - generates outputs
  - in-context learning, prompt engineering



Transformer

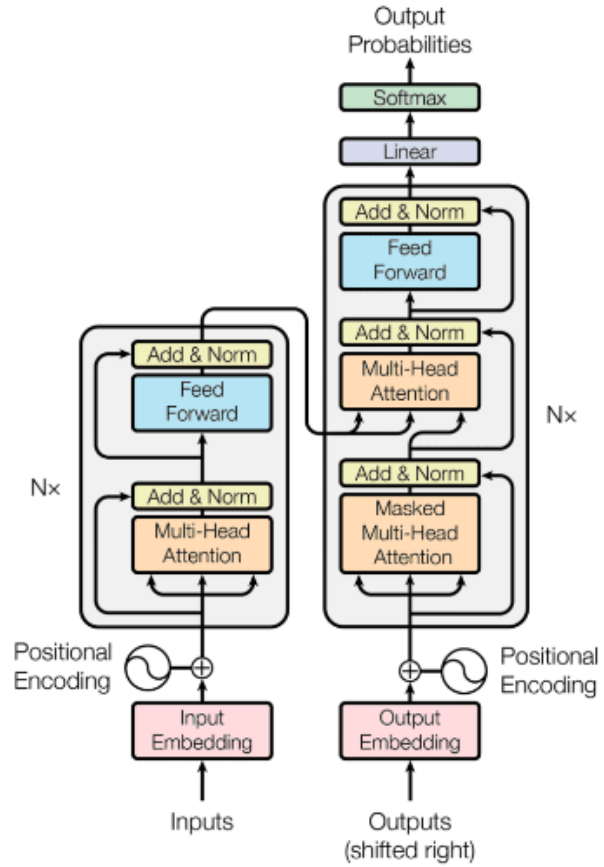
LLM architectural secret (or known) sauce

Transformer - simple parallelizable attention mechanism

A. Vaswani, et al. Attention is All You Need, 2017

Transformer architecture

- encoding-decoding architecture
  - input embedding space → multi-head & mult-layer representation space → output embedding space
- additive positional encoding - information regarding order of words @ input embedding
- multi-layer and multi-head attention followed by addition / normalization & feed forward (FF) layers
- *(relatively simple) attentions*
  - single-head (scaled dot-product) / multi-head attention
  - self attention / encoder-decoder attention
  - masked attention
- benefits
  - *evaluate dependencies between arbitrarily distant words*
  - has recurrent nature w/o recurrent architecture → parallelizable → fast w/ additional cost in computation



Single-head scaled dot-product attention

- values/keys/queries denote value/key/query *vectors*, d_k & d_v are lengths of keys/queries & vectors
- we use *standard* notions for matrices and vectors - not transposed version that (almost) all ML scientists (wrongly) use

- output: weighted-average of values where weights are attentions among tokens
- assume n queries and m key-value pairs

$$Q \in \mathbf{R}^{d_k \times n}, K \in \mathbf{R}^{d_k \times m}, V \in \mathbf{R}^{d_v \times m}$$

- attention! outputs n values (since we have n queries)

$$\text{Attention}(Q, K, V) = V \text{softmax}\left(K^T Q / \sqrt{d_k}\right) \in \mathbf{R}^{d_v \times n}$$

- *much simpler attention mechanism than previous work*
  - attention weights were output of complicated non-linear NN

Single-head - close look at equations

- focus on ith query, $q_i \in \mathbf{R}^{d_k}$, $Q = \begin{bmatrix} - & q_i & - \end{bmatrix} \in \mathbf{R}^{d_k \times n}$
- assume m keys and m values, $k_1, \ldots, k_m \in \mathbf{R}^{d_k}$ & $v_1, \ldots, v_m \in \mathbf{R}^{d_v}$

$$K = \begin{bmatrix} k_1 & \cdots & k_m \end{bmatrix} \in \mathbf{R}^{d_k \times m}, V = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \in \mathbf{R}^{d_v \times m}$$

- then

$$K^T Q / \sqrt{d_k} = \begin{bmatrix} & \vdots & \\ - & k_j^T q_i / \sqrt{d_k} & - \\ & \vdots & \end{bmatrix}$$

$e.g.,$ dependency between ith output token and jth input token is

$$a_{ij} = \exp\left(k_j^T q_i / \sqrt{d_k}\right) / \sum_{j=1}^{m} \exp\left(k_j^T q_i / \sqrt{d_k}\right)$$

- value obtained by ith query, q_i in $\mathrm{Attention}(Q, K, V)$

$$a_{i,1} v_1 + \cdots + a_{i,m} v_m$$

Multi-head attention

- evaluate h single-head attentions (in parallel)
- d_e: dimension for embeddings
- embeddings

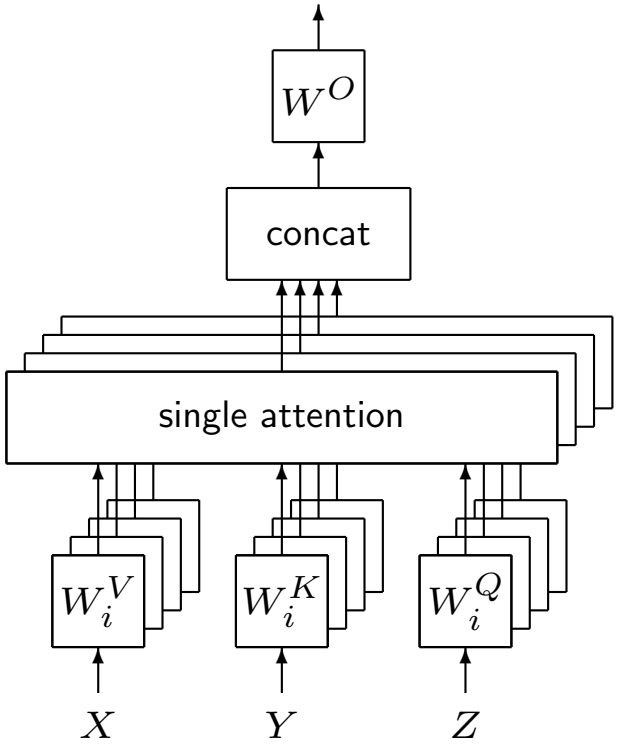
$$X \in \mathbf{R}^{d_e \times m},\ Y \in \mathbf{R}^{d_e \times m},\ Z \in \mathbf{R}^{d_e \times n}$$

  $e.g.$, n: input sequence length & m: output sequence length in machine translation

- h key/query/value weight matrices: $W_i^K, W_i^Q \in \mathbf{R}^{d_k \times d_e}$, $W_i^V \in \mathbf{R}^{d_v \times d_e}$ $(i = 1, \ldots, h)$
- linear output layers: $W^O \in \mathbf{R}^{d_e \times h d_v}$
- *multi-head attention!*

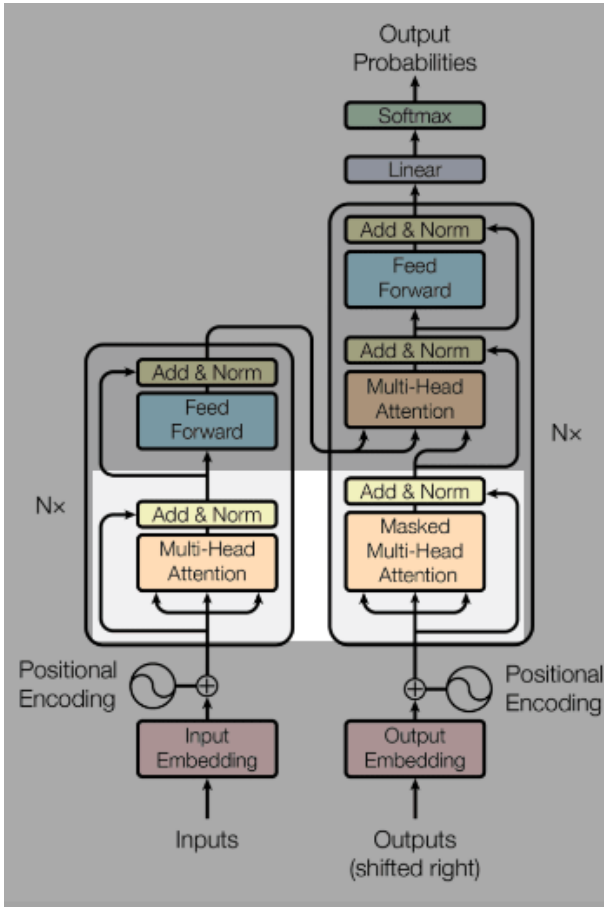
$$W^O \begin{bmatrix} A_1 \\ \vdots \\ A_h \end{bmatrix} \in \mathbf{R}^{d_e \times n},$$

$$A_i = \mathrm{Attention}(W_i^Q Z, W_i^K Y, W_i^V X) \in \mathbf{R}^{d_v \times n}$$



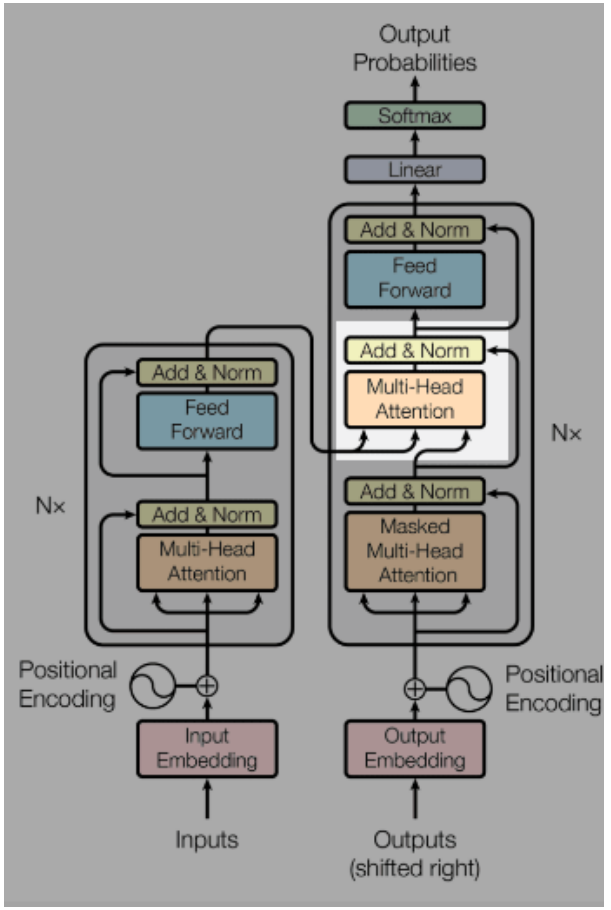
Self attention

- $m = n$
- encoder
  - keys & values & queries (K, V, Q) come from same place (from previous layer)
  - every token attends to every other token in input sequence
- decoder
  - keys & values & queries (K, V, Q) come from same place (from previous layer)
  - every token attends to other tokens up to that position
  - prevent leftward information flow to right to preserve causality
  - assign $-\infty$ for illegal connections in softmax (masking)



Encoder-decoder attention

- m: length of input sequence
- n: length of output sequence
- n queries (Q) come from previous decoder layer
- m keys / m values (K, V) come from output of encoder
- every token in output sequence attends to every token in input sequence

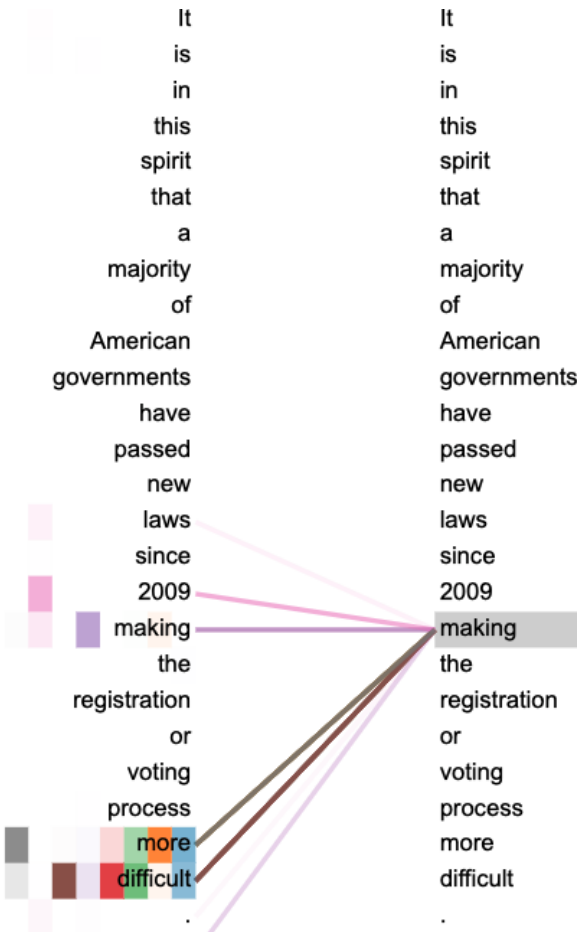


Visualization of self attentions

example sentence

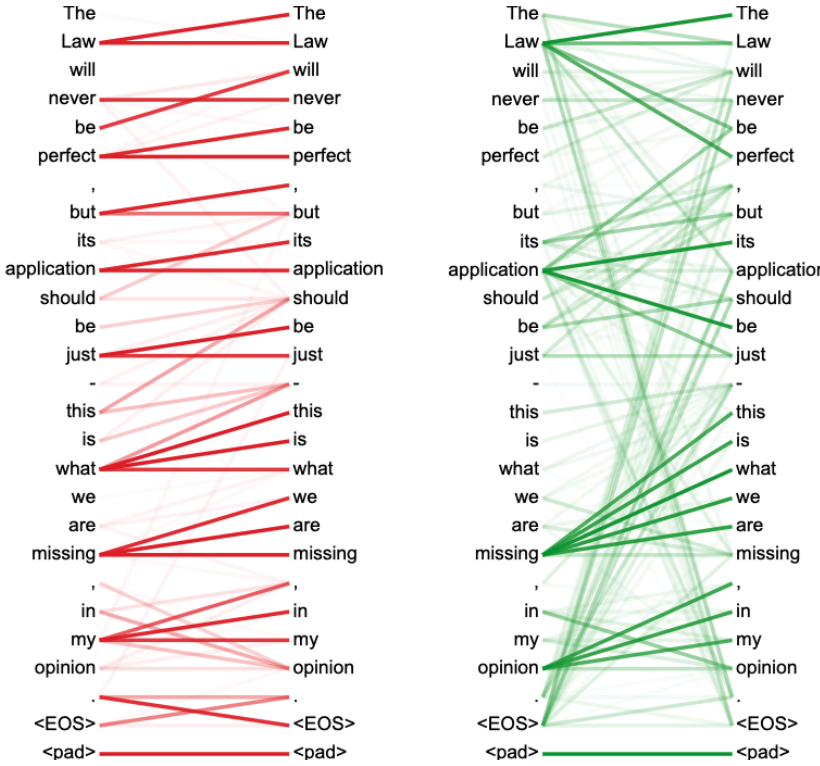
"It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult."

- self attention of encoder (of a layer)
  - right figure
    - show dependencies between "making" and other words
    - different columns of colors represent different heads
  - "making" has strong dependency to "2009", "more", and "difficult"

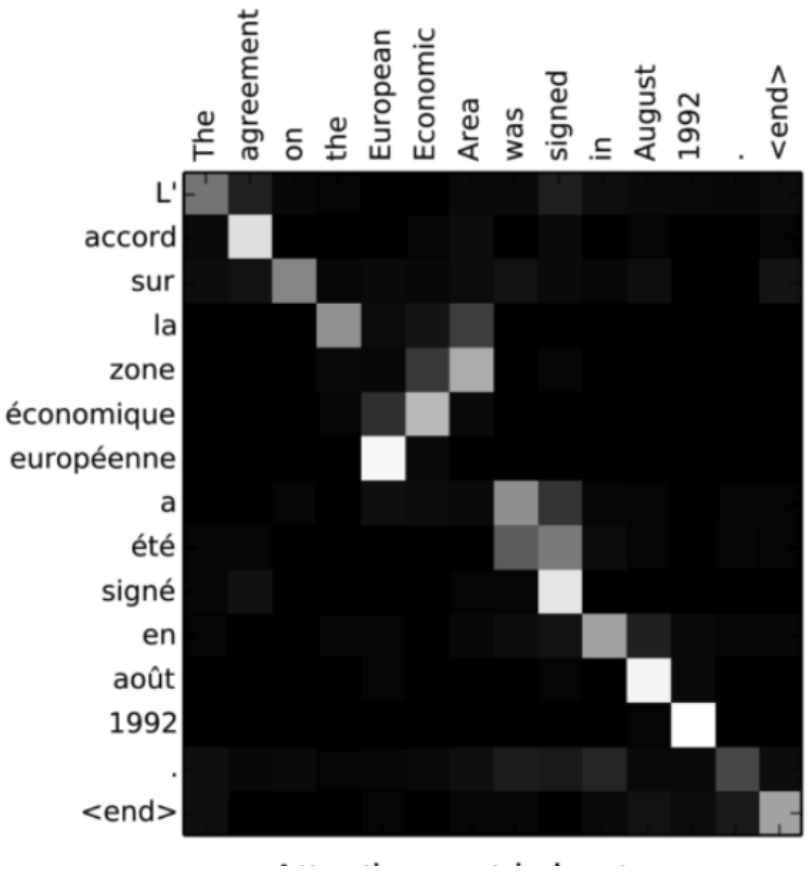


Visualization of multi-head self attentions

- self attentions of encoder for two heads (of a layer)
  - different heads represent different structures
    $\rightarrow$ advantages of multiple heads
  - multiple heads work together to colletively yield good results
  - dependencies *not* have absolute meanings (like embeddings in collaborative filtering)
  - randomness in resulting dependencies exists due to stochastic nature of ML training



Visualization of encoder-decoder attentions



- machine translation: English $\rightarrow$ French
  - input sentence: "The agreement on the European Economic Area was signed in August 1992."
  - output sentence: "L' accord sur la zone économique européenne a été signé en août 1992."
- encoder-decoder attention reveals relevance between
  - European $\leftrightarrow$ européenne
  - Economic $\leftrightarrow$ européconomique
  - Area $\leftrightarrow$ zone

Model complexity

- computational complexity
  - n: sequence length, d: embedding dimension
  - complexity per layer - self-attention: $\mathcal{O}(n^2 d)$, recurrent: $\mathcal{O}(1)$
  - sequential operations - self-attention: $\mathcal{O}(1)$, recurrent: $\mathcal{O}(n)$
  - maximum path length - self-attention: $\mathcal{O}(1)$, recurrent: $\mathcal{O}(n)$

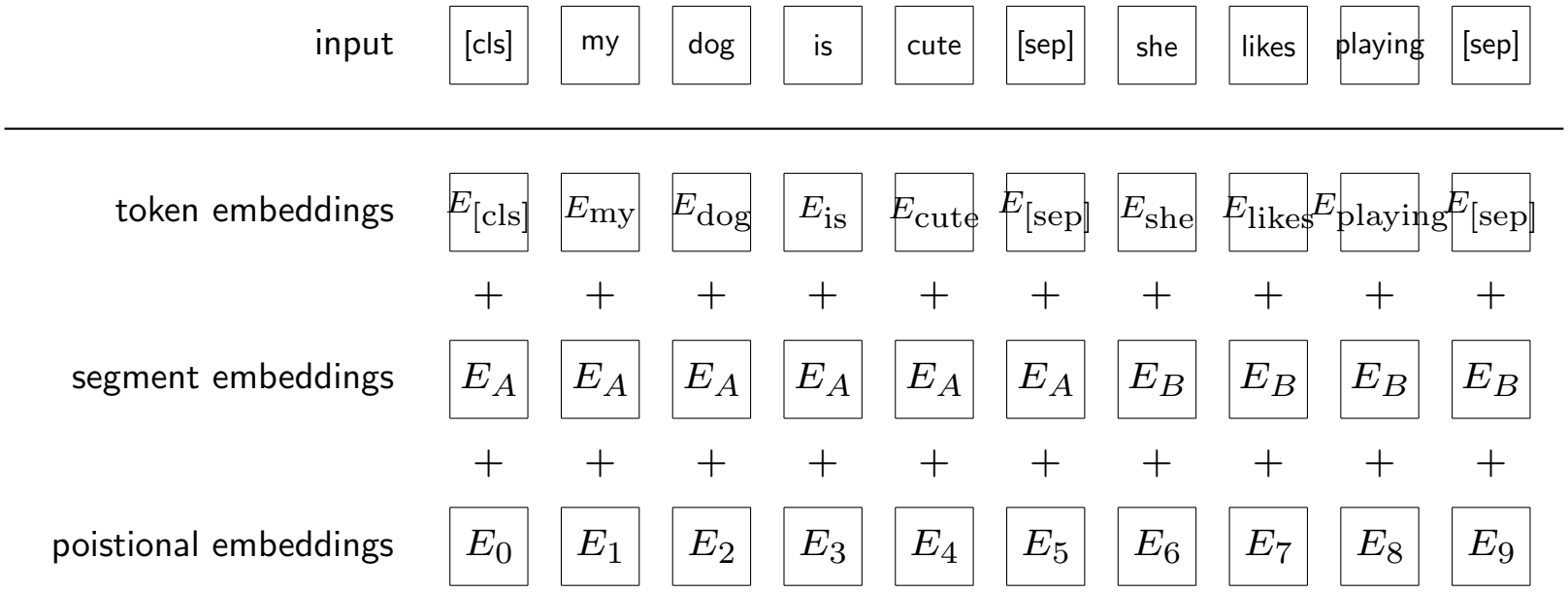
- *massive parallel processing, long context windows*

  $\longrightarrow$ *makes NVidia more competitive, hence profitable!*

  $\longrightarrow$ *makes SK Hynix prevail HBM market!*

Derivatives of Transformer - BERT

- Bidirectional Encoder Representations from Transformers [DCLT19]
- pre-train deep bidirectional representations from unlabeled text
- fine-tunable for multiple purposes



Implications & Challenges

Multimodal learning

- understand information from multiple modalities, $e.g.$, text, images, audio, and video

- representation learning
  - language representation + image / video / text / audio representation
  - learn multimodal representations together

- outputs
  - captions for images, videos with narration, musics with lyrics

- collaboration among different modalities
  - understand image world (open system) using language (closed system)

Implications of success of LLMs

- (very) many researchers change gears towards LLM
  - from computer vision (CV), speach, music, video, even reinforcement learning
- *LLM is not (only) about languages . . .*
  - humans have . . .
    - evolved and optimized (natural) language structures for eons
    - handed down knowledge using natural languages for thousands of years
  - natural language optimized (in human brains) through *thousands of generation by evolution*
  - *can connect non-linguistic world (open system) using language structures (closed system)*

Challenges in LLMs

- *hallucination - can give entirely plausible outcome that is false*
- data poison attack
- unethical or illegal content generation
- huge resource necessary for both training & inference
- model size - need compact models
- outdated knowledge - can be couple of years old
- lack of reproducibility
- *biases - more on this later . . .*

do not, though, focus on downsides but on *infinite possibilities!*

- it evolves like internet / mobile / electricity
- only "tip of the iceburg" found & releaved

genAI

Generative AI (genAI)

- definition of generative model

$$\boxed{\mathcal{Z}} \xrightarrow{g_\theta(z)} \boxed{\mathcal{X}}$$

- *generate samples in original space, $\mathcal{X}$, from samples in latent space, $\mathcal{Z}$*
- g_θ is parameterized model $e.g.$, CNN / RNN / Transformer / diffuction-based model
- training
  - finding θ that minimizes/maximizes some (statistical) loss/merit function so that $\{g_\theta(z)\}_{z\in\mathcal{Z}}$ generates plausiable point in $\mathcal{X}$
- inference
  - random samples z to generated target samples $x = g_\theta(z)$
  - $e.g.$, image, text, voice, music, video

VAE - early genAI model

- variational auto-encoder (VAE) [KW19]

$$\boxed{\mathcal{X}} \xrightarrow{q_\phi(z|x)} \boxed{\mathcal{Z}\text{o}} \xrightarrow{p_\theta(x|z)} \boxed{\mathcal{X}}$$

- log-likelihood & ELBO - for any $q_\phi(z|x)$

$$\begin{aligned} \log p_\theta(x) &= \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log p_\theta(x) = \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)} \cdot \frac{q_\phi(z|x)}{p_\theta(z|x)} \\ &= \mathcal{L}(\theta,\phi;x) + D_{KL}(q_\phi(z|x)\|p_\theta(z|x)) \geq \mathcal{L}(\theta,\phi;x) \end{aligned}$$

- (indirectly) maximize likelihood by maximizing evidence lower bound (ELBO)

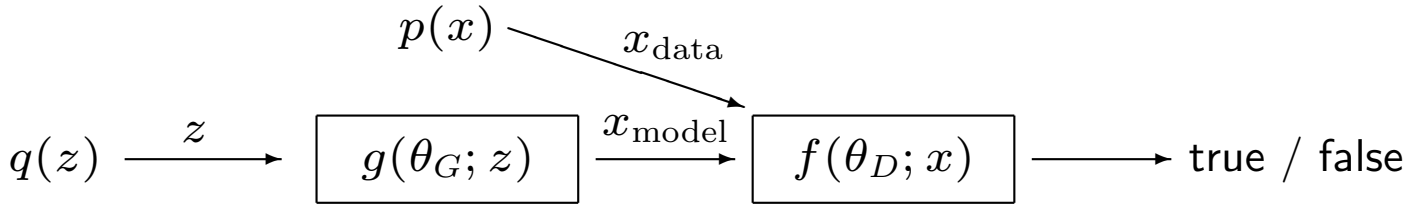
$$\mathcal{L}(\theta,\phi;x) = \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

- generative model

$$p_\theta(x|z)$$

GAN - early genAI model

- generative adversarial networks (GAN) [GPAM$^+$14]



  - value function

$$V(\theta_D, \theta_G) = \mathop{\mathbf{E}}_{x \sim p(x)} \log f(\theta_D; x)) + \mathop{\mathbf{E}}_{z \sim q(z)} \log(1 - f(\theta_D; g(\theta_G; z)))$$

  - modeling via playing min-max game

$$\min_{\theta_G} \max_{\theta_D} V(\theta_D, \theta_G)$$

  - generative model

$$g(\theta_G; z)$$

  - variants: conditional / cycle / style / Wasserstein GAN

genAI - LLM

- *maximize conditional probability*

$$\underset{\theta}{\text{maximize}} \; d\left(p_\theta(x_t|x_{t-1}, x_{t-2}, \ldots), p_{\text{data}}(x_t|x_{t-1}, x_{t-2}, \ldots)\right)$$

  where $d(\cdot, \cdot)$ distance measure between probability distributions

  - previous sequence: $x_{t-1}, x_{t-2}, \ldots$
  - next token: x_t

- p_θ represented by (extremely) complicated model
  - $e.g.$, containing multi-head & multi-layer Transformer architecture inside
- model parameters, $e.g.$, for Llama2

$$\theta \in \mathbf{R}^{70{,}000{,}000{,}000}$$

AI Applications

genAI applications

- ChatGPT, Cohere
- Anthropic, Dolly, Mosaic MPT
- Stable Diffusion
- Midjourney, DALL-E, LLaMA 2
- Mistral AI, Amazon Bedrock, and Falcon

ChatGPT & VR/AR

- new appropriately to teaching
- power of ChatGPT and VR/AR unlocks immersive learning
  - *learning language* - immersive VR environment provides immediate feedback, responding to inquiries & interactive discussions
  - *medical education* - experience diagnosing and treating patients in lifelike scenarios
  - *investing history & culture* - integration of ChatGPT into VR enables virtual visit to historical places and cultural landmarks
  - *development of soft skills* - practice and hone soft skills, $e.g.$, leadership, teamwork & communication through VR simulations augmented by ChatGPT
  - *extracurricular activities*
    - personalized learning, gamification of education, international cooperation, educator empowerment
- *VR & ChatGPT integration opens up new training and educational opportunities!*

AI Market

genAI products

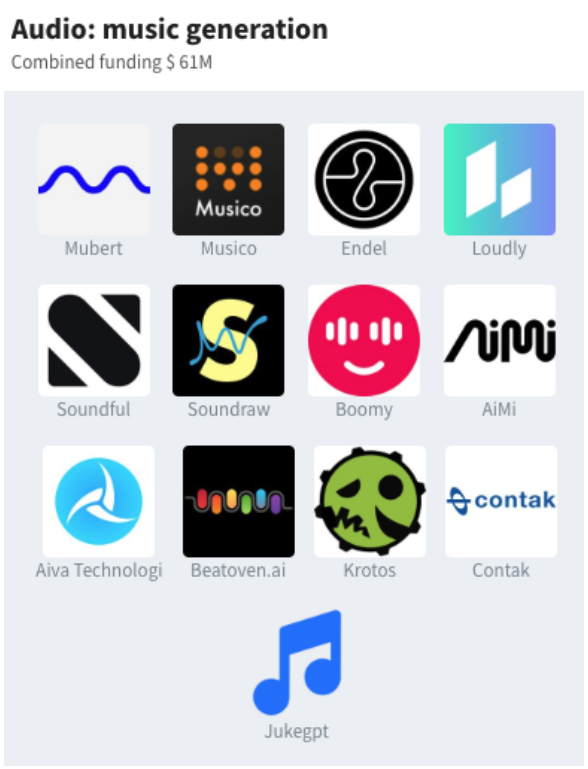
- DALL-E (OpenAI)
  - trained on a diverse range of images
  - *generate unique and detailed images based on textual descriptions*
  - understanding context and relationships between words

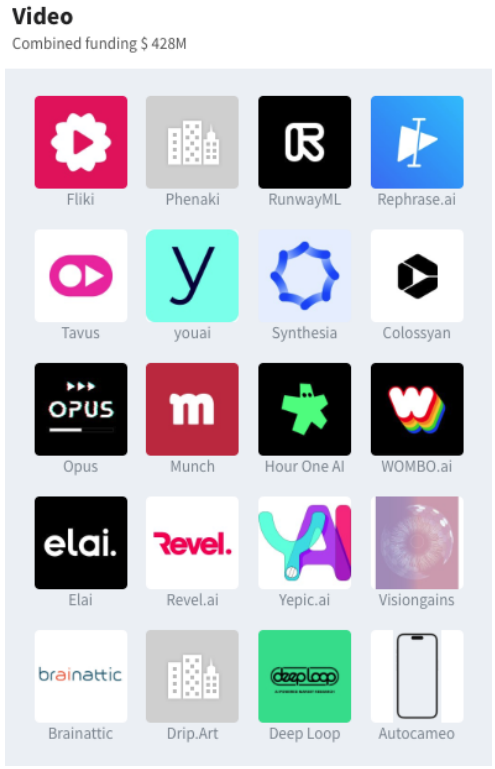
- Midjourney
  - let people *create imaginative artistic images*
  - can interactively guide the generative process, providing high-level directions

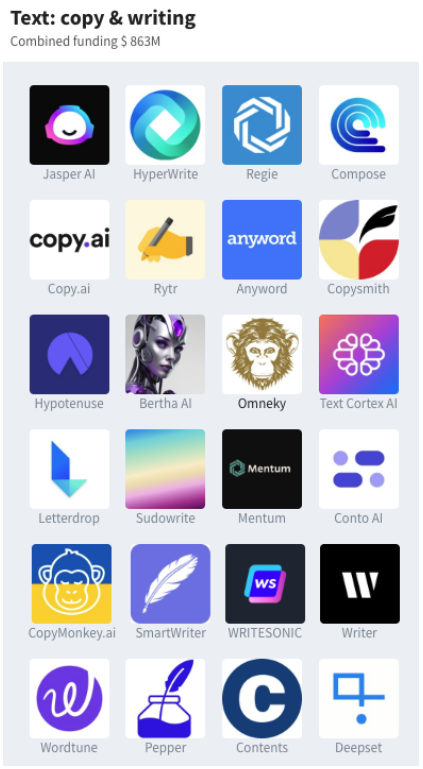
genAI products

- Dream Studio
  - enables people to create music
  - *analyze patterns in music data and generates novel compositions based on input and style*
  - *allows musicians to explore new ideas and enhance their* creative *processes*
  - offer open-source free version
- Runway
  - provide range of generative AI tools for creative professionals
  - *realistic images, manipulate photos, create 3D models, automate filmmaking, . . .*
  - "artificial intelligence brings automation at every scale, introducing dramatic changes in how we create"

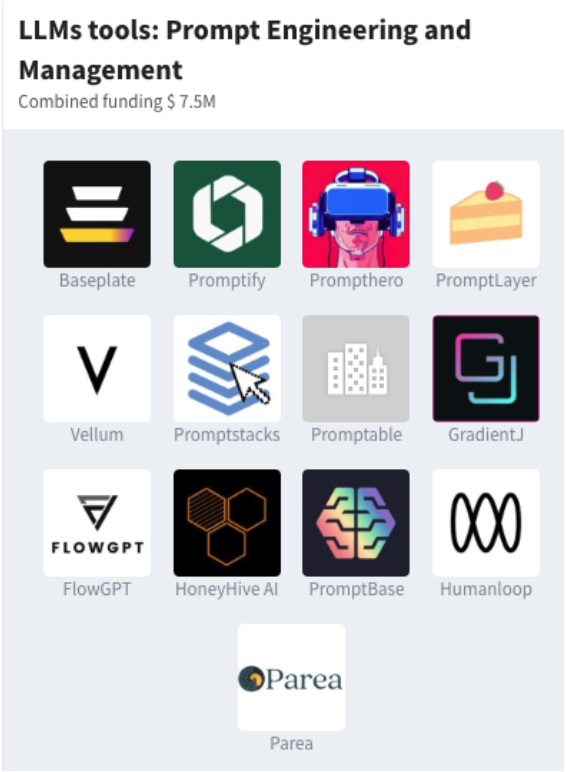
AI products

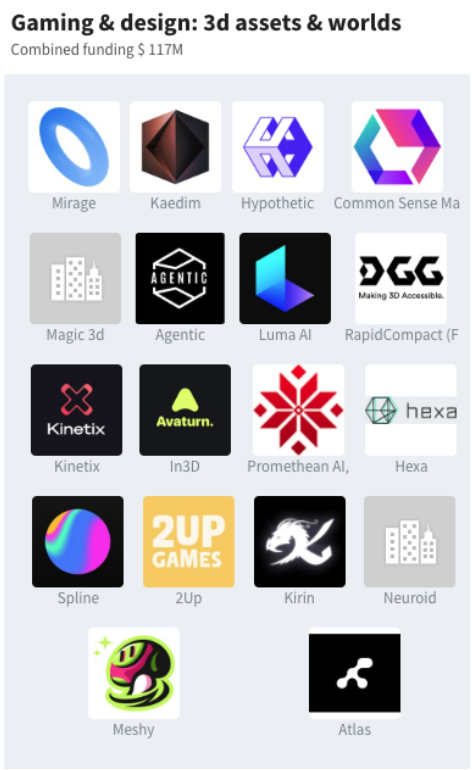

Audio: music generation
Combined funding $ 61M
Mubert
Musico
Endel
Loudly
Soundful
Soundraw
Boomy
AiMi
Aiva Technologi
Beatoven.ai
Krotos
Contak
Jukegpt

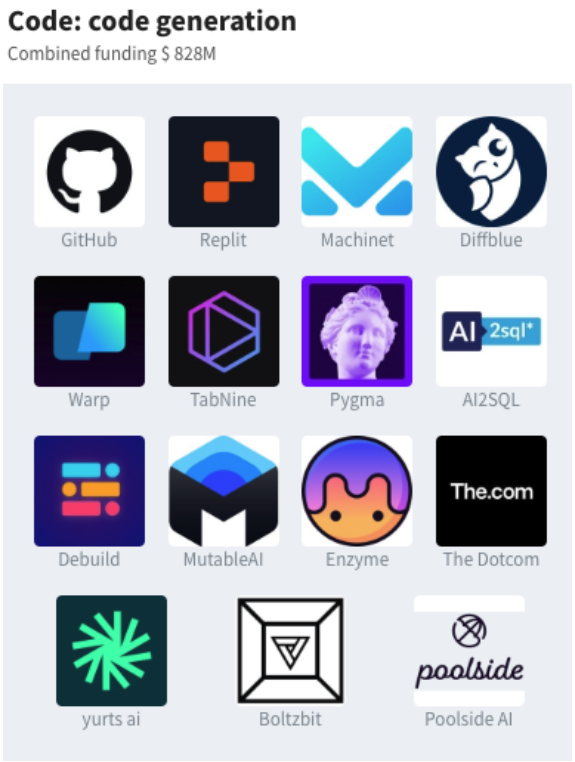

Video
Combined funding $ 428M
Fliki
Phenaki
RunwayML
Rephrase.ai
Tavus
youai
Synthesia
Colossyan
Opus
Munch
Hour One AI
WOMBO.ai
Elai
Revel.ai
Yepic.ai
Visiongains
Brainattic
Drip.Art
Deep Loop
Autocameo


Text: copy & writing
Combined funding $ 863M
Jasper AI
HyperWrite
Regie
Compose
Copy.ai
Rytr
Anyword
Copysmith
Hypotenuse
Bertha AI
Omneky
Text Cortex AI
Letterdrop
Sudowrite
Mentum
Conto AI
CopyMonkey.ai
SmartWriter
WRITESONIC
Writer
Wordtune
Pepper
Contents
Deepset

AI products


LLMs tools: Prompt Engineering and Management
Combined funding $ 7.5M
Baseplate
Promptify
Prompthero
PromptLayer
Vellum
Promptstacks
Promptable
GradientJ
FLOWGPT
FlowGPT
HoneyHive AI
PromptBase
Humanloop
Parea
Parea


Gaming & design: 3d assets & worlds
Combined funding $ 117M
Mirage
Kaedim
Hypothetic
Common Sense Ma
Magic 3d
AGENTIC
Agentic
Luma AI
RapidCompact (F
Kinetix
Kinetix
Avaturn.
In3D
Promethean AI,
hexa
Hexa
Spline
2UP GAMES
2Up
Kirin
Neuroid
Meshy
Atlas


Code: code generation
Combined funding $ 828M
GitHub
Replit
Machinet
Diffblue
Warp
TabNine
Pygma
AI2SQL
Debuild
MutableAI
Enzyme
The.com
The Dotcom
yurts ai
Boltzbit
poolside
Poolside AI

AI companies

- big tech companies
  - OpenAI, Microsoft, Google, Meta - foundation models
- small(er) players
  - Figure AI, Mistral AI
- AI hardware companies - benefiting from LLM and genAI market dominance
  - Nvidia, AMD, Samsung, SK hynix, Micron, Intel, TSMC (AI processors & memory chips)
- *tiny fraction of Silicon Valley startups gets majority of total funding*
  - Anthropic - $3.5B - large-scale AI systems - Claude
  - AssemblyAI - $58M - speech AI
  - Hugging Face - $400M - AI model/data platforms
  - Inflection AI - $1.5B - conversational AI - Pi

Opportunities among big tech's domination

- OpenAI/Microsoft, Meta, Google's races for foundation models heated up!
- no small players can compete with rare exceptions, $e.g.$, Mistral AI
- hyperscalers stand strong - AWS, Azure, and Google Cloud

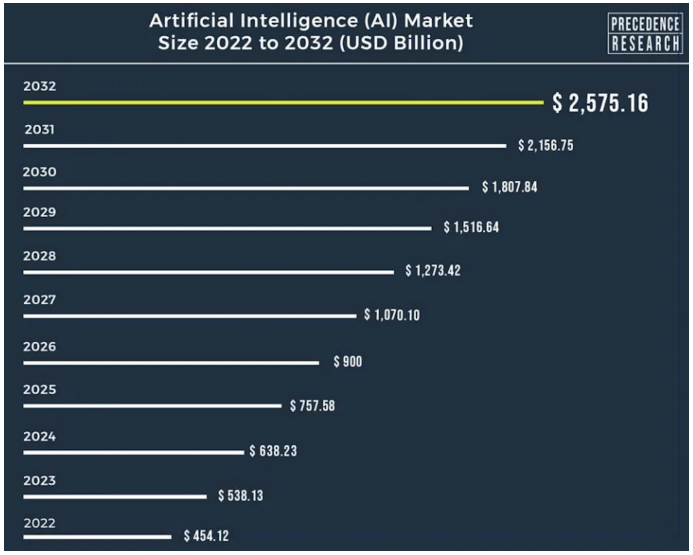
- *speaker's proposals for strategies*
  - accurately (or roughly) predict how far & up to where big players will reach
  - target niche markets
  - focus on (creative) downstream applications of LLMs and/or genAIs

AI market outlook in 2024

- global AI market expected to reach *USD 0.5T by 2024* (IDC @ Mar-2023) & expected to reach around *USD 2.5T by 2032* (Precedence Research @ Dec-2023) [P.R23]
  - was valued at USD 454B in 2022, expanding at *double-digit CAGR of 19%* from 2023 to 2032

- *AI funding soars to USD 17.9B for Q3 in 2023 in Silicon Valley while rest of tech slumps* (PitchBook data, Bloomberg @ Oct-2023) [Blo23]
  - multibillion-dollar investment in AI startups almost commonplace in Silicon Valley
  - genAI dazzles users and investors with photo-realistic images & human-sounding text

- genAI software sales could surge *18,647% by 2032*



Productivity, inflation & jobs

- Federal Reserve probes AI's impact on productivity, inflation & jobs - Jul-2024
  - feds acknowledging significant AI investments
  - Jerome Powell emphasizes uncertainties on whether AI will eliminate, augment, or create jobs - stating it's too early to predict
  - Powell acknowledges limited influence of central banks like the Fed on AI's technological shifts
  - fed actively researching various AI forms beyond genAI to understand potential economic impacts
  - IMF predicts AI (could) impact up to 60% of jobs in advanced economies potentially lowering labor demand and wages in sectors like finance and insurance

AI & global economy

- five ways AI is transforming global economy
  - reshape job markets, creating new roles while rendering some obsolete
  - enhance productivity across industries
  - contribute to global economy by optimizing processes and innovation
  - *may widen economic disparities if not managed inclusively*
  - *governments* has to develop policies to address AI's economic and social impacts

AI Industry

Heavy Lifting of LLMs

News - OpenAI's "$8.5B bills" report sparks bankruptcy speculation

- OpenAI's financial situation reflects its ambitious vision
  - projected $8.5B expenses vs $3.5–4.5B revenue in 2024 w/ massive investment in AI infrastructure and talent
- caused by Sam Altman's reckless & non-strategic commitment to AGI development
  - "Whether we burn $500M, $5B, or $50B a year, I don't care..." - prioritizing long-term impact over short-term profitability
- reflect broader AI industry trend of high burn rates
  - indicative of the resource-intensive nature of cutting-edge AI research

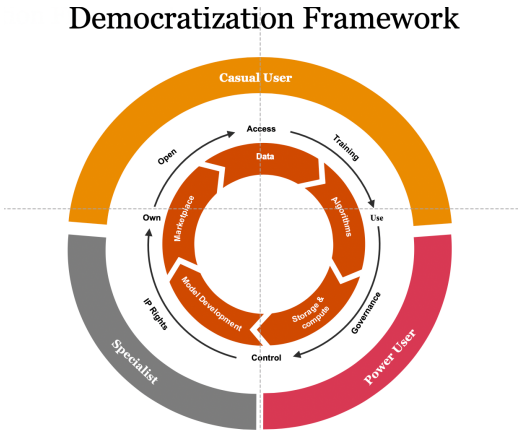
LLM - strategic challenges & industry dynamics

- evolving competitive landscape
  - threat from open-source models ($e.g.$, Meta's Llama 3.1) & potential commoditization of LLMs
- balancing act with Microsoft partnership
  - critical financial support vs maintaining independence - Microsoft's $\$13$B investment provides both opportunity and constraint
- sustainability of current business model
  - high costs of AI development vs monetization challenges
  - need for breakthrough applications or efficiency improvements
- ethical & regulatory considerations
  - balancing rapid development with responsible AI principles
  - potential impact of future AI regulations on operations and costs

Impacts of Open-source AI Models

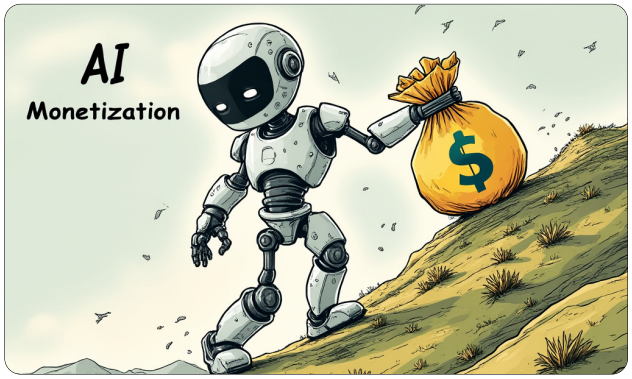
Industry disruption of open-source AI models on industry

- rise of open-source models such as Meta's Llama 3.1 reshaping the AI landscape
- industry disruption
  - AI democratization - open-source making advanced AI capabilities accessible to wider range of developers and companies
  - innovation acceleratation - collaborative improvement of open-source models could lead to faster progress
  - pressure on proprietary models - companies like OpenAI may need to offer significant advantages over free alternatives to justify their costs



Impact of open-source AI models on industry

- business model challenges
  - monetization difficulties - capable models becoming freely available
  - shift to services & applications - focus may move from selling access to models to providing *specialized services* or *applications built on top of them*
- ethical & security concerns
  - responsible AI - open-source models raise questions about control and responsible use
  - dual-use potential - wider access to powerful AI models could increase risks of misuse or malicious applications, $e.g.$, Deepfake



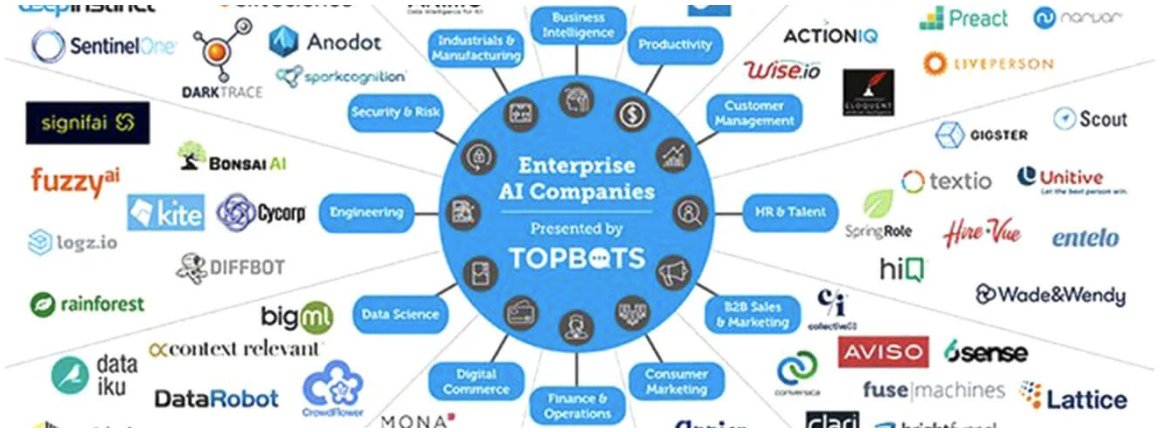
Tech Giants & AI Companies

Evolving relationship between tech giants & AI companies

- partnership between OpenAI & Microsoft exemplifies broader trend of collaboration & integration in AI industry
- symbiotic relationships
  - tech giants provide :esources & funding - AI companies research & innovation
  - provide AI companies w/ instant access to large user bases & distribution channels
- power dynamics
  - independence concerns - AI companies' risk of losing autonomy
  - tech giants' access to advanced AI potentially widening gap with smaller competitors

AI industry consolidation

- mergers & acquisitions
  - will see increased M&A activities as tech giants seek to bring AI capabilities in-house
- ecosystem development
  - tech giants creating AI-focused ecosystems, similar to cloud services, to attract and retain developers & businesses

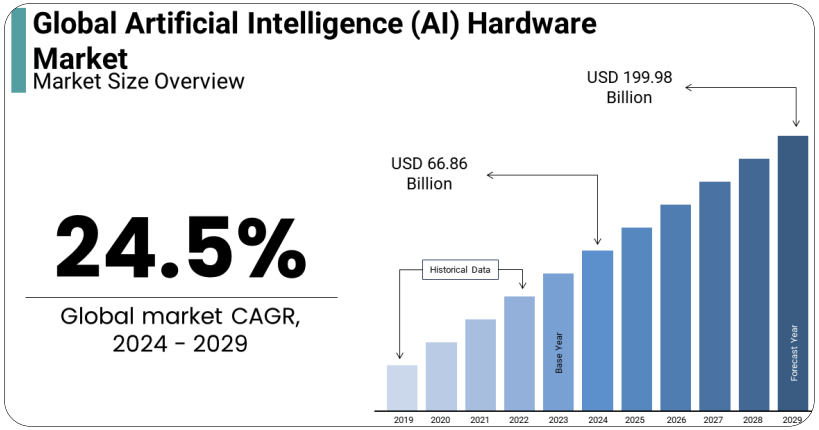


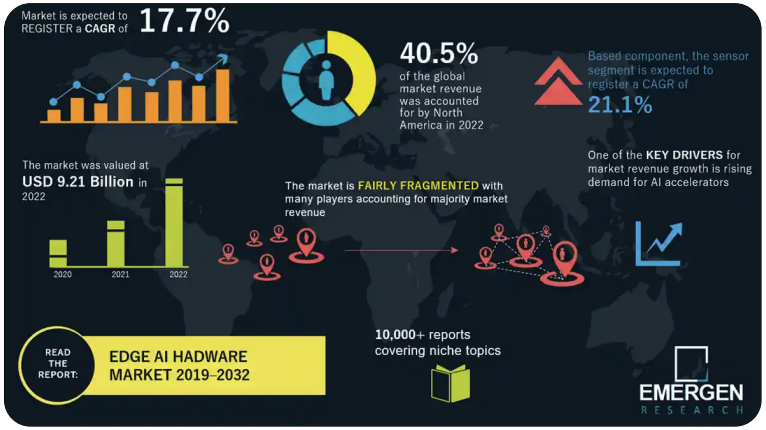
AI Hardware

AI Hardware Industry

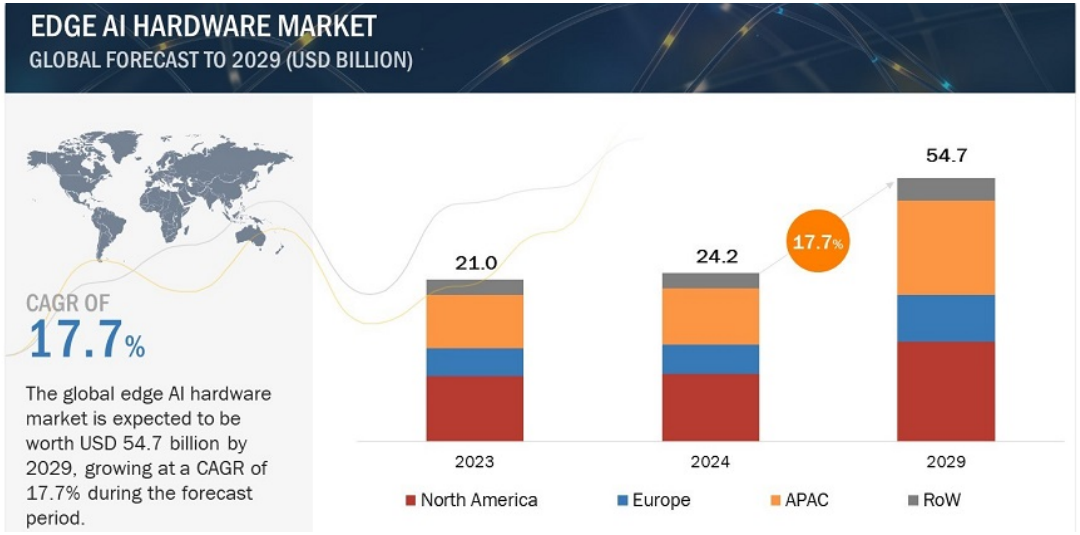
Landscape of AI hardware industry

- global AI hardware market valued at $66.96B in 2024, projected to grow significantly
- major companies - Nvidia, Intel, AMD, Qualcomm, and IBM w/ Nvidia holding substantial market share



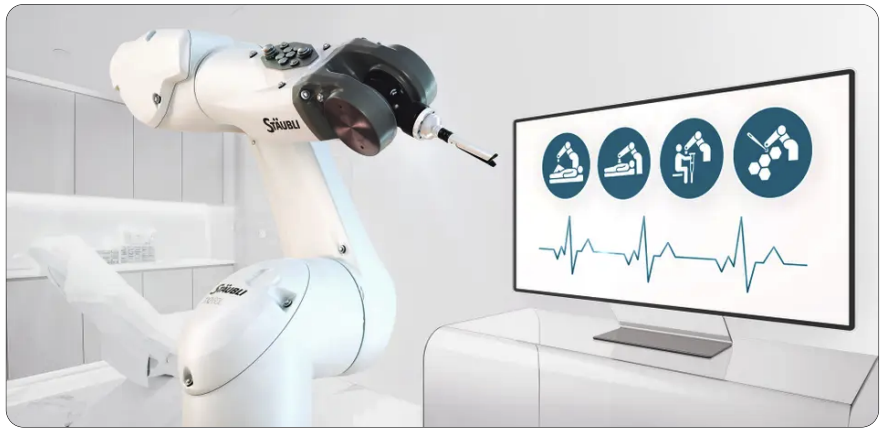


- North America leading market - high R&D investments & key industry players
- Asia Pacific rapidly expanding - strong semiconductor industries in South Korea, China & Japan
- demand for advanced processors such as GPUs, TPUs & AI accelerators rising due to complexity of AI algorithms & high computational power



Predictions for future of AI hardware market

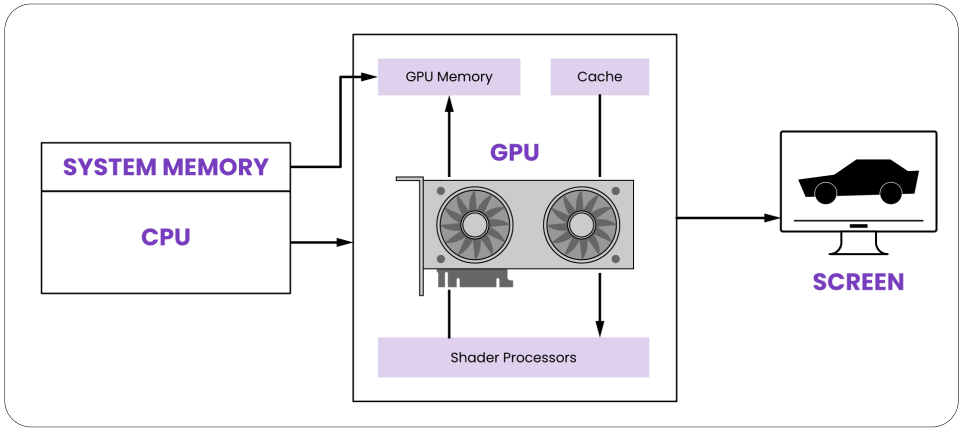
- AI hardware market expected to reach $382B by 2032 - significant growth in data center AI chips
- integration of AI w/ 5G & increased use of AI in edge computing anticipated to drive future demand
- AI hardware becoming crucial in sectors such as autonomous vehicles, robotics & medical devices
- need to address challenges such as heat and power management along with technical complexities



GPUs and AI Accelerators

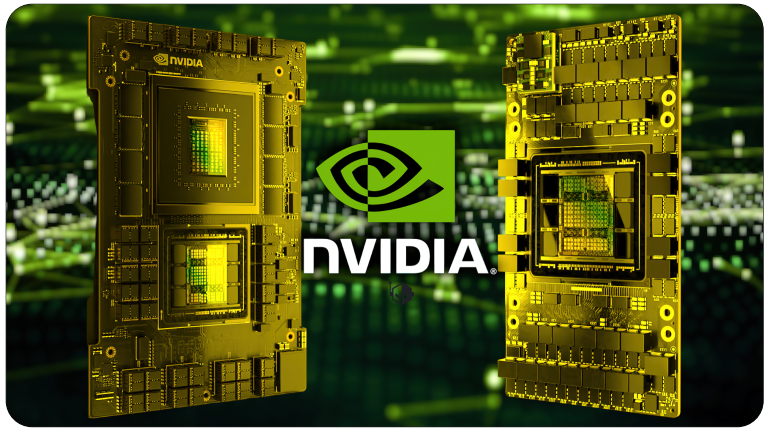
Technical challenges of GPUs & AI accelerators

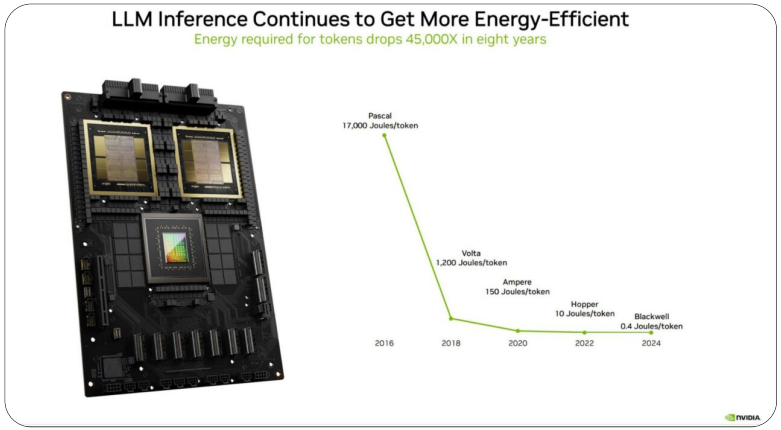
- facing challenges in scaling to handle increasingly large AI models and datasets - traditional architectures struggling w/ massive parallel processing demands of modern AI applications
- AI applications require extensive memory bandwidth often leading to bottlenecks - efficient memory management is crucial
- AI accelerators consume significant power - high operational costs and environmental concerns for both cloud-based & edge AI applications



Potential solutions for overcoming challenges

- development of AI-specific architectures such as tensor cores and custom ASICs to improve efficiency and performance - novel architectures like FPGAs for specific AI tasks, $e.g.$, for RAG & vectorDB
- implementing software optimizations to enhance hardware usability and performance - use of compilers and frameworks that maximize efficiency of existing hardware
- encouraging market competition to drive innovation and reduce monopolistic control - exploring alternative hardware solutions and improving energy efficiency standards





Big tech's in-house chip development

- shift towards in-house AI hardware - major tech companies increasingly developing their own AI chips - move to enhance AI capabilities and reduce dependence
- collaboration with specialized partners - partnering with specialized firms for manufacturing and technology blending in-house expertise with external innovation

	Microsoft	Google	Amazon	Meta
Chip	Maia 100	TPU v5e	Inferentia2	MTIA v1
Launch Date	November, 2023	August, 2023	Early 2023	2025
IP	ARM	ARM	ARM	RISC-V
Process Technology	TSMC 5nm	TSMC 5nm	TSMC 7nm	TSMC 7nm
Transistor Count	105 billion	-	-	-
INT8	-	393 TOPS	-	102.4 TOPS
FP16	-	-	-	51.2 TFLOPS
BF16	-	197 TFLOPS	-	-
Memory	-	-	-	LPDDR5
TDP	-	-	-	25W
Packaging Technology	CoWoS	CoWoS	CoWoS-S	2D
Collaborating Partners	Global Unichip Corp.	Broadcom	Alchip Technologies	Andes Technology
Application	Training/Inference	Inference	Inference	Training/Inference
LLM	GPT-3.5, GPT-4	BERT, PaLM, LaMDA	Titan FM	Llama, Llama2

AMD - Nvidia's new competitor

- key points
  - AMD launched new AI accelerator chip, *Instinct MI300X*, on Dec 6, 2023
  - CDNA 3 architecture, mix of 5nm and 6nm IPs, delivering 153B transistors
  - *outperforms Nvidia's H100 TensorRT-LLM* by 1.6X higher memory bandwidth and 1.3X FP16 TFLOPS
  - up to 40% faster vs Nvidia's Llama-2 70B model in 8x8 server configurations
- market impact
  - significant challenge to Nvidia's dominance in AI accelerator market
  - performance gains over Nvidia's offerings could drive *customer adoption and market share for AMD*
- future prediction
  - *AMD stocks soared* since launch indicating investor confidence in their competitiveness
  - Lisa Su, AMD's CEO, categorized Instinct MI300X as "next big thing" in tech industry
  - potential risks include need to *manage ROCm vs CUDA software ecosystem* & ensure rapid customer adoption and production coverage

AI Accelerator Startups - Market Impact & Prediction

AI accelerator startups

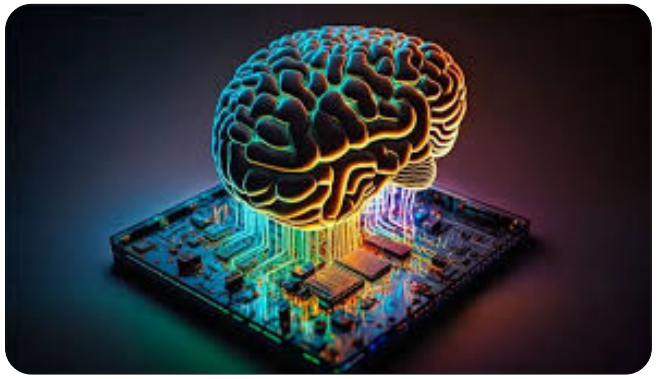
- innovative architectures - startups like Groq, SambaNova & Graphcore leading with *novel architectures designed to accelerate AI workloads*
  - *Groq* - tensor streaming processor (TSP) offering ultra-low latency & high throughput, high-performance AI inference chips enhancing speed & efficiency
  - *SambaNova* - reconfigurable dataflow architecture optimizing for various AI workloads
  - *Graphcore* - intelligence processing unit (IPU) tailored for graph-based computation excelling in sparse data processing
  - *Cerebras Systems* - develop wafer scale engine (WSE), largest chip built for AI workloads, unmatched computational power revolutionizing AI hardware capabilities
  - *Hailo* - specialize for edge devices optimizing AI processes for real-time applications, raised $120M emphasizing potential to disrupt traditional AI chip markets

Technological competitiveness

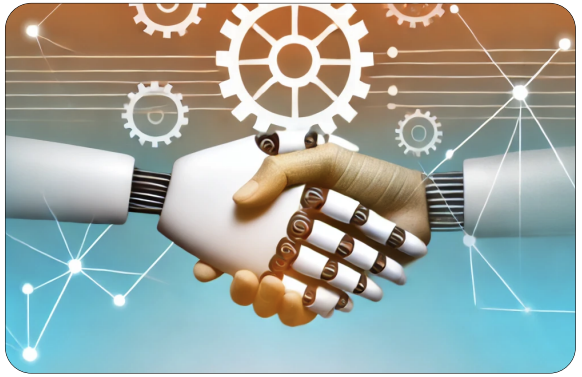
- energy efficiency
  - energy-efficient designs crucial for scalability in data centers and edge devices
  - startups developing solutions significantly reducing power consumption without compromising performance
- customization & flexibility
  - AI accelerators from startups often offer greater customization options for specific AI tasks compared to traditional GPUs
  - flexibility in hardware allows for tailored solutions that can outperform general-purpose accelerators in certain applications
- software integration
  - robust software ecosystems critical - startups investing in developing software stacks that optimize performance for their hardware
  - compatibility with existing AI frameworks is competitive advantage, $e.g.$, `TensorFlow` & `PyTorch`

Industry and market influence

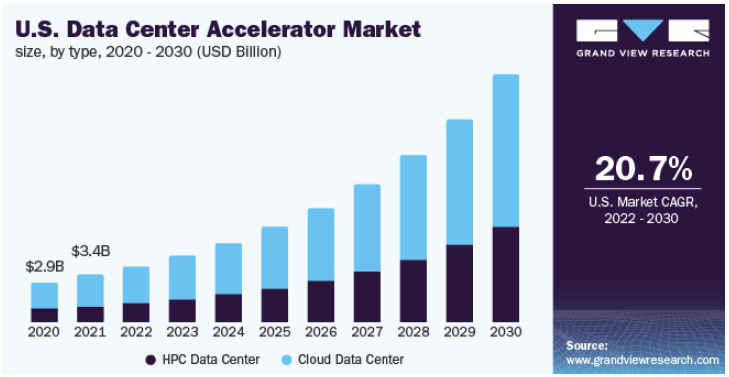
- disruption of traditional players
  - challenging dominance of established players like NVIDIA & Intel
  - unique architectures providing specialized solutions traditional GPUs and CPUs cannot efficiently handle
- driving down costs
  - offering competitive alternatives pushing down cost of AI computation
  - could lead to democratization of AI w/ more companies affording high-performance AI capabilities



- accelerating AI innovation
  - contributing to rapid innovation providing hardware that can handle emerging AI models & workloads
  - adaptability and specialization enable advancements in AI research & faster development cycles
- strategic partnerships & acquisitions
  - big techs increasingly forming strategic partnerships or acquiring startups to stay competitive
  - collaborations can speed up integration of advanced AI hardware into mainstream products



- market growth & opportunities
  - AI accelerator market expected to grow significantly driven by demand in data centers, edge computing & autonomous systems
  - startups well-positioned to capture significant share of growing market particularly in niche applications
- future outlook
  - dependency on Asia for fabrication might lead to strategic shifts in global tech policies and investments in local manufacturing
  - increasing demand for efficient AI processing on edge devices and in data center.



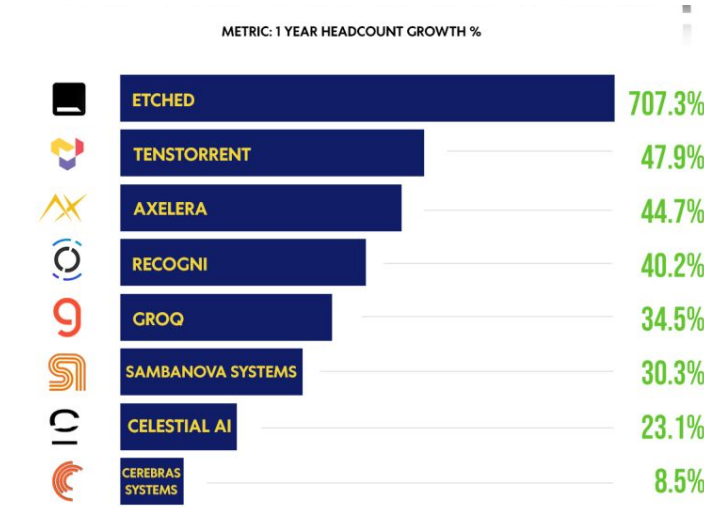
Silicon Valley Startups

Incubators

- Y Combinator
  - invests $500,000 in each startup receiving 7% equity
  - program culminates in "Demo Day" where startups present to investors
  - helped Airbnb, Dropbox, DoorDash, Reddit, Stripe, . . .
  - highly competitive - acceptance rate of 1.5% – 2% - 10k applicants per six months
  - significantly higher survival rate - around 18% valued at over $100M and 4% becoming unicorns—valued at over $1B
- XXX

We're not talking about AI chip startups enough

- *not* talking about AI chip startups enough
  - Etched - specializes in AI accelerator chips for LLMs - Sohu chip embeds transformer architecture into silicon
  - Tenstorrent - develops high-performance AI processors focused on efficient training and inference
  - Axelera AI - specializes in high-performance AI hardware for edge computing, excelling in CV applications
  - Recogni - develops high-performance AI chips for edge computing applications focusing on low-latency processing for autonomous vehicles
- *global AI chip market will gross $30B in 2024 and become $300B in next 10 years*



Global Semiconductor Business

Hard-to-predict AI hardware markets

- US
  - birthplace for modern semiconductor chips driving PC market, internet, multi-media, mobile phones, and AI . . .
    - Intel, Texas Instrument (TI), Global Foundry
  - traditionally strong with design houses - NVIDIA, AMD, Broadcom, Apple, . . .
  - threatened experiencing global chip shortage & vulnerable supply chain via COVID
  - national security concerns & economic competitiveness
- China
  - strong fast followers - SMIC[4], Huawei, Hua Hong Semiconductor (foundry)
- South Korea
  - best memory chip makers - Samsung, SK hynix
  - struggling with LSI and foundry business

[4] SMIC - Semiconductor Manufacturing International Corporation

Reshoring semiconductor manufacturing industry

- trade & semiconductor WAR between US and China
  - export controls on advanced chips and equipment to China
  - investment restrictions in both directions
- CHIPS & Science Act (Aug, 2022)
  - $52B in subsidies for domestic semiconductor production, 25% investment tax credit for chip plants
  - (coerce) world-best semiconductor companies build factories in US with support of government and states
    - GlobalFoundries - $1.5B @ Feb-2024 - Global Foundary
    - Intel - $8.5B @ Apr-2024 - Ohio - two fabs expandable to $100B
    - Samsung - $6.4B @ Apr-2024 - Talor, Texas - advanced logic chips
    - TSMC - $6.6B @ Apr-2024 - Phoenix, Arizona - two foundry fabs (3nm & 4nm)

Turmoils in global semiconductor business

- global context
  - EU Chips Act - €43B to boost European chip production
  - Japan & South Korea - significant investments in domestic capacity
- industry dynamics
  - Intel's foundry ambitions - targeting 50% global market share by 2030
  - TSMC expanding global footprint (US, Japan, possibly Germany)
- future outlook
  - projected shift in global semiconductor manufacturing landscape
  - increased geographical diversification of chip production

Export controls on US chip technology to China

- goal - limit China's access to advanced semiconductor tech to maintain US strategic advantage
- impacts on
  - China - advanced chips and equipment not allowed, domestic innovation increased
  - US - short-term - US lose market share and revenue in China
  - US - long-term - potential decline in US global competitiveness
- Chinese response - circumvent controls and adapt supply chains
- conclusion
  - US-China chip rivalry transforms global supply chains with deep implications for *security & industry*
  - US success hinges on better coordination and policy analysis

- reference - Balancing the Ledger - Center for Strategic & International Studies (CSIS)

China strikes back on US sanction

- Huawei's launch of Mate 60 Pro smartphone
  - these domestically produced chips represent major breakthrough against US sanctions
  - its success with *advanced 7nm Kirin 9000S chip* demonstrates significant progress in China's self-reliance in high-tech manufacturing - narrowing the technological gap with global leaders
- Huawei case highlights potential failure of US sanctions potentially leading to more aggressive US measures
  - US export controls on China's semiconductor industry are effective in the short term but insufficient to halt China's progress especially in legacy chip manufacturing
  - to maintain technological edge, US must balance further restrictions with supporting its semiconductor industry to avoid overreliance on export controls

Chinese semiconductor companies

- Chinese major semiconductor companies
  - SMIC - China's largest chip foundry, advancing 7nm technology
  - HiSilicon - Huawei's chip design arm, crucial for the Kirin processors
  - YMTC - leader in 3D NAND memory chip production
  - Huahong Group, CXMT, SMEE, GigaDevice, UniIC Semiconductors, ASMC, $etc.$
- *SMIC shows significant progress in producing 7nm chips* & YMTC leads memory chip manufacturer - both face challenges from US export controls
- industry faces internal challenges, $e.g.$, corruption & misallocation of resources
- but remains crucial to China's goal of technological self-reliance

Serendipities around AIs

Serendipity or inevitability

- What if Geoffrey Hinton had not been persistent researcher?
- What if symbolists won AI race over connectionists?
- What if attention mechanism did not perform well?
- What if Transfomer architecture did not perform super well?
- What if Jensen Hwang had not been crazy about making hardware for professional gamers?

- Is it like Alexander Fleming's Penicillin?
- Or more like Inevitability?

AI in Biotech

AI in biology

- AI has been used in biological sciences, and science in general
- AI's ability to process large amounts of raw, unstructured data ($e.g.$, DNA sequence data)
  - reduces time and cost to conduct experiments in biology
  - enables others types of experiments that previously were unattainable
  - contributes to broader field of engineering biology or biotechnology
- AI increases human ability to make direct changes at cellular level and create novel genetic material ($e.g.$, DNA and RNA) to obtain specific functions.

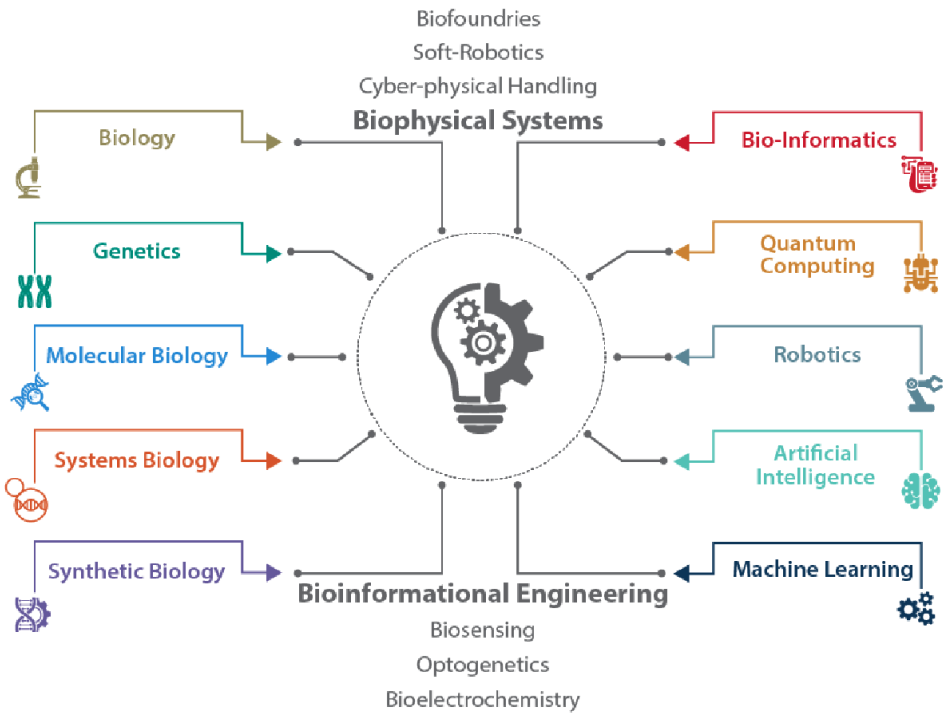
Biotech

Biotech

- biotechnology
  - is multidisciplinary field leveraging broad set of sciences and technologies
  - relies on and builds upon advances in other fields such as nanotechnology & robotics, and, increasingly, AI
  - enables researchers to read and write DNA
    - sequencing technologies "read" DNA while gene synthesis technologies takes sequence data and "write" DNA turning data into physical material
- 2018 National Defense Strategy & senior US defense and intelligence officials identified emerging technologies that could have disruptive impact on US national security [Say21]
  - artificial intelligence, lethal autonomous weapons, hypersonic weapons, directed energy weapons, *biotechnology,* quantum technology
- other names for biotechnology are engineering biology, synthetic biology, biological science (when discussed in context of AI)

biotech - multidisciplinary field

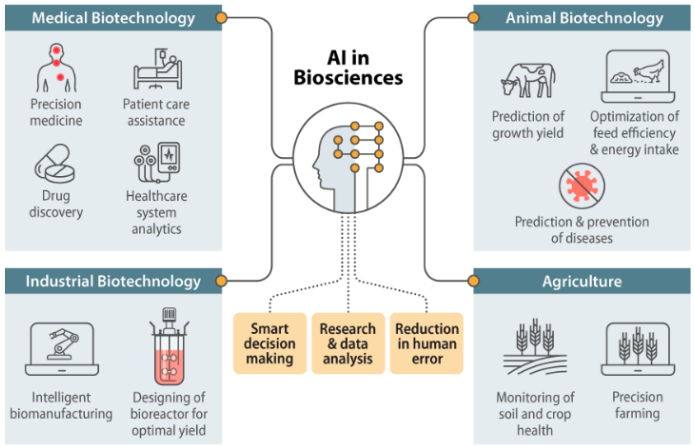
- sciences and technologies enabling biotechnology include, but not limited to,
  - (molecular) biology, genetics, systems biology, synthetic biology, bio-informatics, quantum computing, robotics [DFJ22]



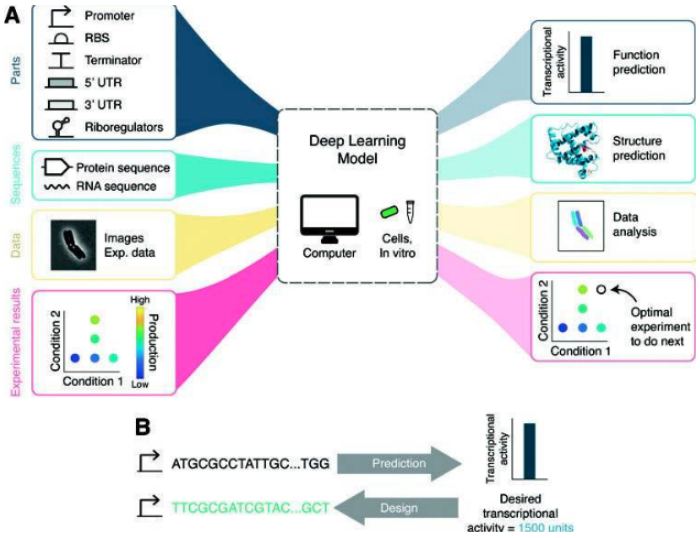
Convergence of AI and biological design

- both AI & biological sciences increasingly converging [BKP22]
  - each building upon the other's capabilities for new research and development across multiple areas
- Demo Hassabis, CEO & cofounder of DeepMind, said of biology [Toe23]

  ". . . biology can be thought of as information processing system, albeit extraordinarily complex and dynamic one . . . just as mathematics turned out to be the right description language for physics, biology may turn out to be *the perfect type of regime for the application of AI!*"
- Both AI & biotech rely on and build upon advances in other scientific disciplines and technology fields, such as nanotechnology, robotics, and increasingly big data (*e.g.*, genetic sequence data)
  - each of these fields itself convergence of multiple sciences and technologies
- so *their impacts can combine to create new capabilities*



Multi-source genetic sequence data



- AI is essential to analyzing exponential growth of genetic sequence data

  > "AI will be essential to fully understanding how genetic code interacts with biological processes"
  > - US National Security Commission on Artificial Intelligence (NSCAI)

  – process huge amounts of biological data, *e.g.*, genetic sequence data, coming from different biological sources for understanding complex biological systems

    - sequence data, molecular structure data, image data, time-series, omics data

- *e.g.*, analyze genomic data sets to determine the genetic basis of particular trait and potentially uncover genetic markers linked with that trait

Quality & quantity of biological data

- limiting factor, however, is quality and quantity of the biological data, $e.g.$, DNA sequences, that AI is trained on
  - $e.g.$, accurate identification of particular species based on DNA requires reference sequences of *sufficient quality* to exist and be available
- databases have varying standards - access, type and quality of information
- design, management, quality standards, and data protocols for reference databases can affect utility of particular DNA sequence

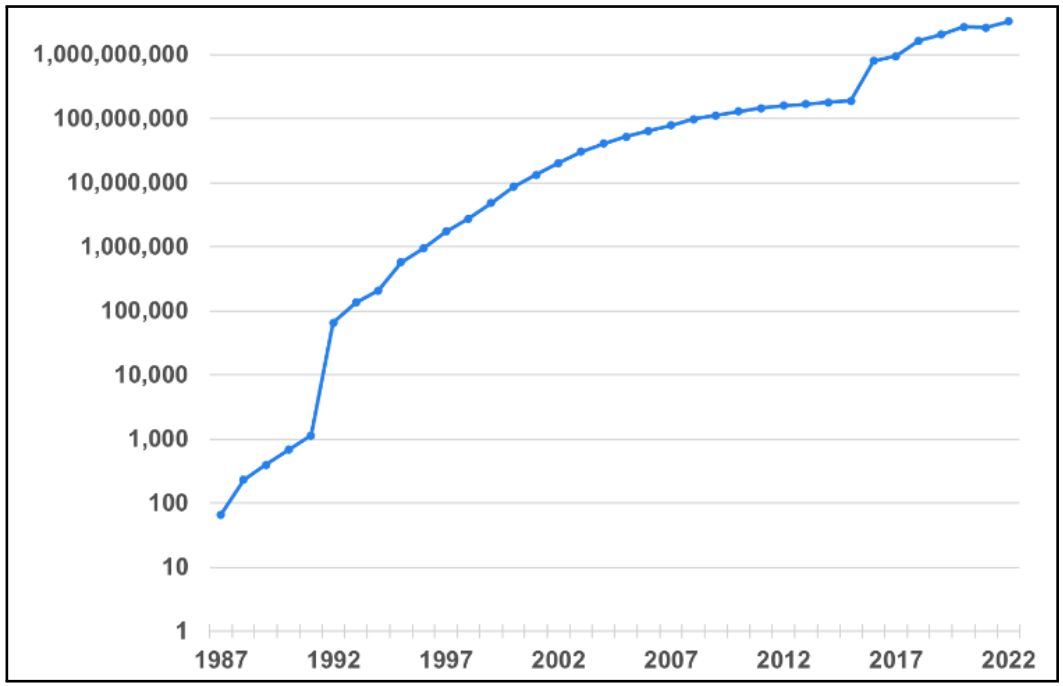
Rapid growth of biological data

- volume of genetic sequence data grown exponentially as sequencing technology has evolved
- more than 1,700 databases incorporating data on genomics, protein sequences, protein structures, plants, metabolic pathways, $etc.$, $e.g.$
  - open-source public database
    - Protein Data Bank, US-funded data center, contains more than *terabyte of three-dimensional structure data* for biological molecules, including proteins, DNA, and RNA
  - proprietary database
    - Gingko Bioworks - possesses more than *2B protein sequences*
  - public research groups
    - Broad Institute - produces roughly *500 terabases of genomic data per month*
- great potential value in aggregate volume of genetic datasets that can be collectively mined to discover and characterize relationships among genes

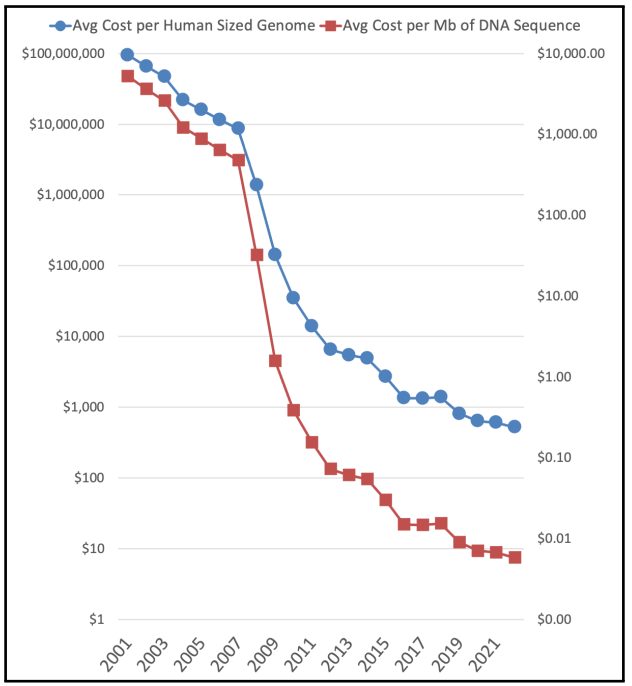
Volume and sequencing cost of DNA over time

- volume of DNA sequences & DNA sequencing cost
  - data source: National Human Genome Research Institute (NHGRI) [Wet23] & International Nucleotide Sequence Database Collaboration (INSDC)

sequences in INSDC



DNA sequencing cost



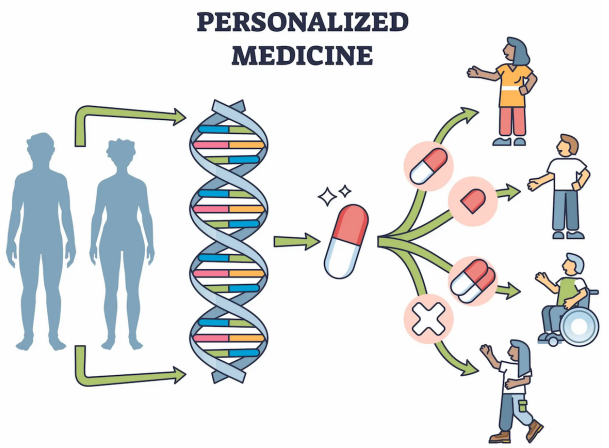
Bio data availability and bias

- US National Security Commission on Artificial Intelligence (NSCAI) recommends
  - US fund and prioritize development of a biobank containing *"wide range of high-quality biological and genetic data sets securely accessible by researchers"*
  - establishment of database of broad range of human, animal, and plant genomes would
    - *enhance and democratize biotechnology innovations*
    - *facilitate new levels of AI-enabled analysis of genetic data*
- bias - availability of genetic data & decisions about selection of genetic data can introduce bias, $e.g.$
  - training AI model on datasets emphasizing or omitting certain genetic traits can affect how information is used and types of applications developed - *potentially privileging or disadvantaging certain populations*
  - access to data and to AI models themselves may impact communities of differing socioeconomic status or other factors unequally

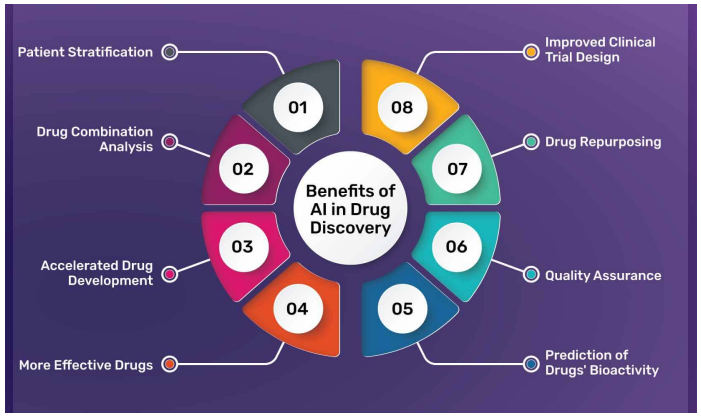
Emerging Trends in Biotech

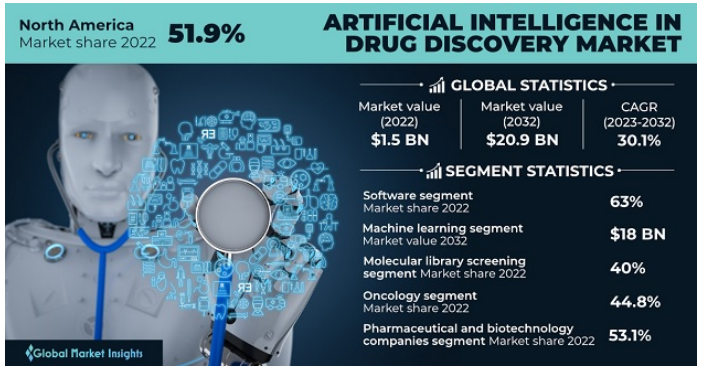
Personalized medicine

- *shift from one-size-fits-all approach to tailored treatments*
- based on individual genetic profiles, lifestyles & environments
- AI enables analysis of vast data to predict patient responses to treatments, thus enhancing efficacy and reducing adverse effects
- $e.g.$, custom cancer therapies, personalized treatment plans for rare diseases & precision pharmacogenomics.
- companies - Tempus, Foundation Medicine, $etc.$



AI-driven drug discovery

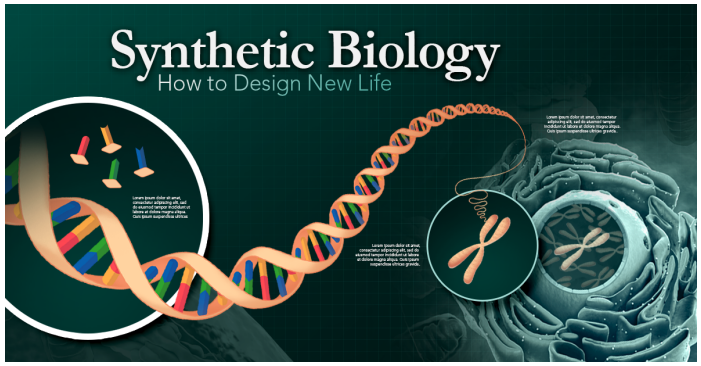


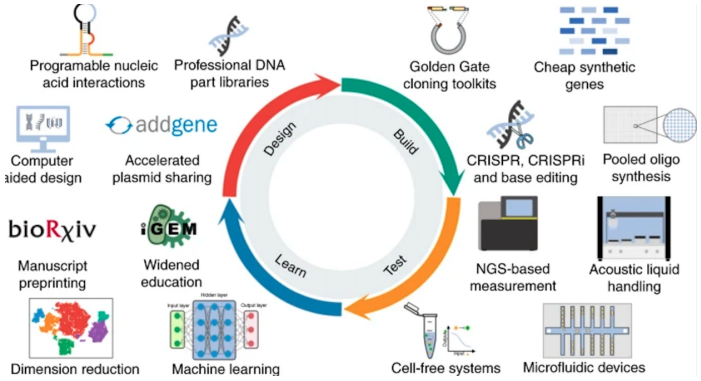


- traditional drug discovery process - time-consuming and costly often taking decades and billions of dollars
- AI streamlines this process by predicting the efficacy and safety of potential compounds with more speed and accuracy
- AI models analyze chemical databases to identify new drug candidates or repurpose existing drugs for new therapeutic uses
- companies - Insilco Medicine, Atomwise.

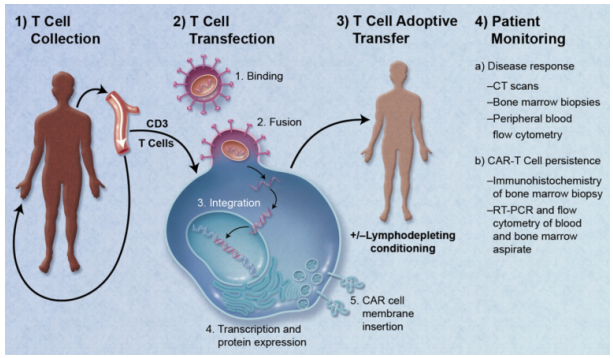
Synthetic biology

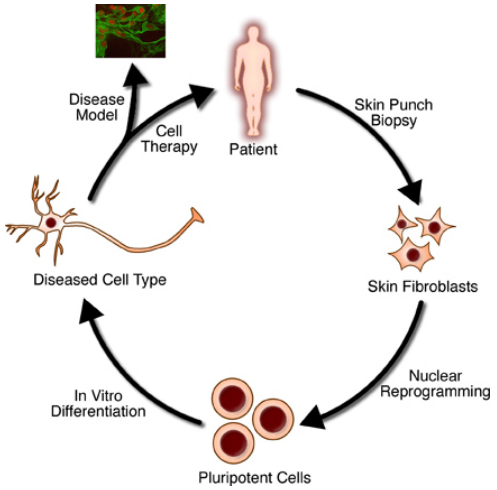
- use AI for gene editing, biomaterial production and synthetic pathways
- combine principles of biology and engineering to design and construct new biological entities
- AI optimizes synthetic biology processes from designing genetic circuits to scaling up production
- company - Ginkgo Bioworks uses AI to design custom microorganisms for applications ranging from pharmaceuticals to industrial chemicals





Regenerative medicine

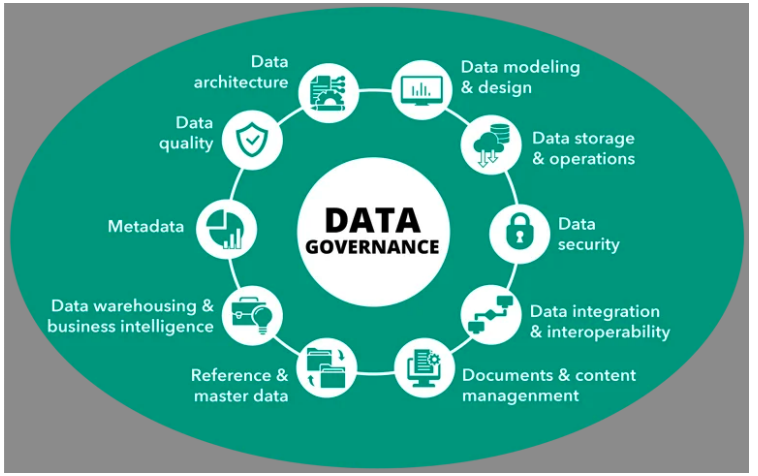




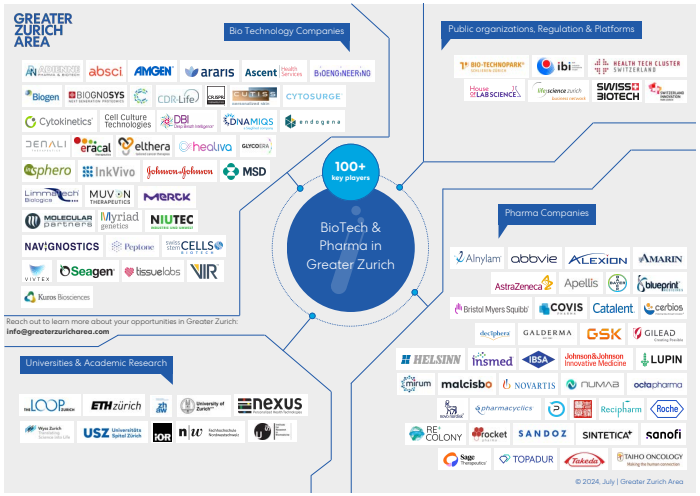
- AI advances development of stem cell therapies & tissue engineering
- AI algorithms assist in identifying optimal cell types, predicting cell behavior & personalized treatments
- particularly for conditions such as neurodegenerative diseases, heart failure and orthopedic injuries
- company - Organovo leverages AI to potentially improve the efficacy and scalability of regenerative therapies, developing next-generation treatments

Bio data integration

- integration of disparate data sources, including genomic, proteomic & clinical data - one of biggest challenges in biotech & healthcare
- AI delivers meaningful insights *only when* seamless data integration and interoperability realized
- developing platforms facilitating comprehensive, longitudinal patient data analysis - vital enablers of AI in biotech
- company - Flatiron Health working on integrating diverse datasets to provide holistic view of patient health



Biotech companies



- Atomwise - small molecule drug discovery
- Cradle - protein design
- Exscientia - precision medicine
- Iktos - small molecule drug discovery and design
- Insilico Medicine - full-stack drug discovery system
- Schrödinger, Inc. - use physics-based models to find best possible molecule
- Absci Corporation - antibody design, creating new from scratch antibodies, $i.e.$, "de novo antibodies", and testing them in laboratories

Industrial AI

Industrial AI (inAI)

- inAI (collectively) refers to AI technology & software and their products developed for
  - *customer values creation, productivity improvement, cost reduction, production optimization, predictive analysis, insight discovery*

  in industries such as
  - *semiconductor, steel, oil & gas, cement, and other various manufacturing industries*

  (unlike general AI, which is frontier research discipline striving to achieve human-level intelligence)

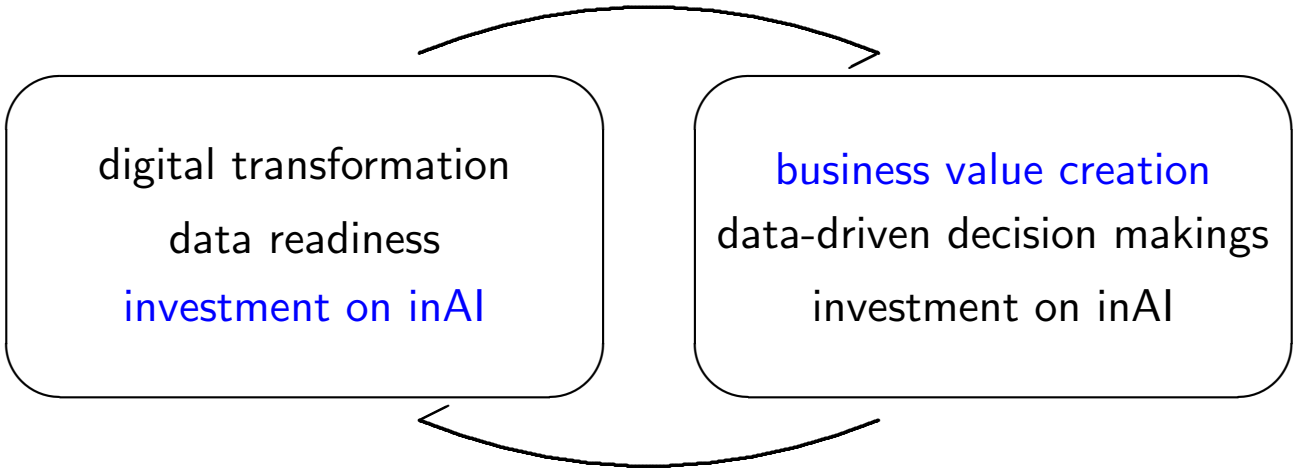
inAI fields

- product
  - product design & innovation, adaptability & advancement, product quality & validation, design for reusability & recyclability, performance optimization
- production process
  - *production quality,* process management, inter-process relations, process routing & scheduling, process design & innovation, *traceability, predictive process control*
- machinery & equipment
  - *predictive maintenance, monitoring & diagnosis,* component development, *ramp-up optimization,* material consumption prediction
- supply chain
  - supply chain monitoring, material requirements planning, customer management, supplier management, logistics, reusability & recyclability

Characteristics of inAI

Vicious (or virtuous) cycle

- integration of inAI with customers' business creates monetary values and encourages data-driven decisions
- however, to do so, digital transformation with data-readiness is MUST-have
- created values, in turn, can be invested into infrastructure required for digital transformation and success of inAI!



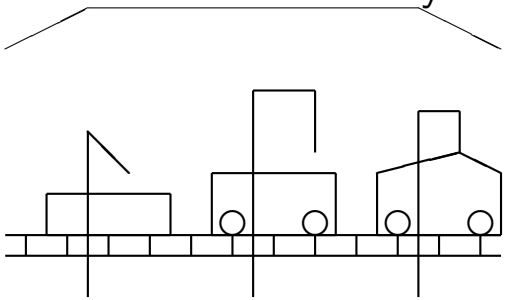
Data-centric AI

- unlike many ML disciplines where foundation models do generic representation learning, $i.e.$, learn universal features
- each equipment has (gradually) different data characteristics, hence need data-centric AI

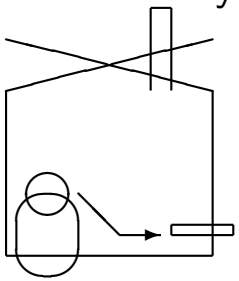
  ". . . need 1,000 models for 1,000 problems" - Andrew Ng

  – data-centric AI - discipline of systematically engineering the data used to build AI system

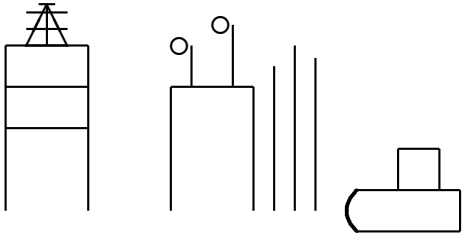
automobile industry



steel industry



oil & gas industry



Challenging data characteristics

- huge volume
- data multi-modality
- high velocity requirement
- very fat data
- sever data shift & drift (in many cases)
- label imbalance
- data quality

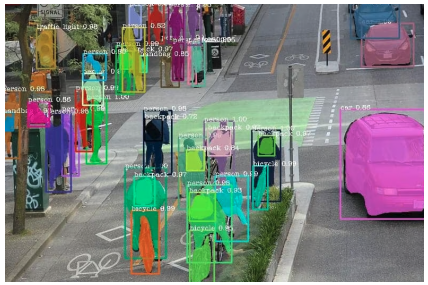
Manufacturing AI

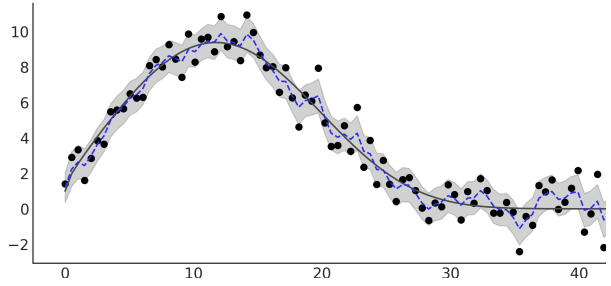
MLs in manufacturing AI (manAI)

- *image data* - huge amount of image data measured and inspected
  - SEM/TEM images, wafer defect maps, test failure pattern maps [5]

  $\rightarrow$ semantic segmentation, defect inspection, anomaly detection
- *time-series (TS) data* - *all the data* coming out of manufacturing is TS
  - equipment sensor data, process times, various measurements, MES data [6]

  $\rightarrow$ regression, anomaly detection, semi-supervised learning, Bayesian inference





[5] SEM: scanning electron miscroscope, TEM: transmission electron miscroscope

[6] MES: manufacturing execution system

CV ML in manAI

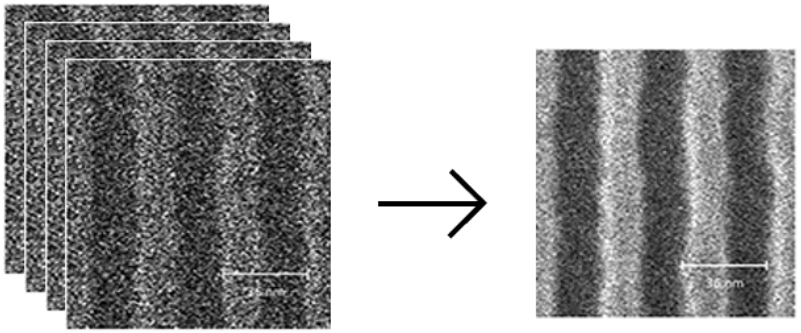
Computer vision ML in manAI

- measurement and inspection (MI)
  - metrology - measurement of critical features
  - inspection - defect inspection, defect localization, defect classification
  - failure pattern analysis
- applications
  - automatic feature measurement
  - anomaly detection
  - defect inspection

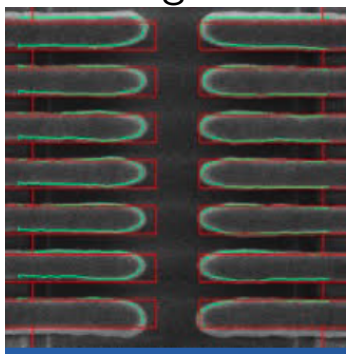
Automatic feature measurement

- ML techniques
  - image enhancement (denoising)
  - texture segmentation
  - repetitive pattern recognition
  - automatic measurement

image enhancement



texture segmentation



auto-measurement

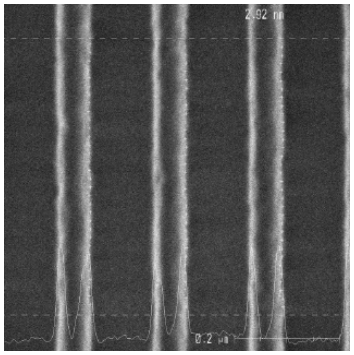
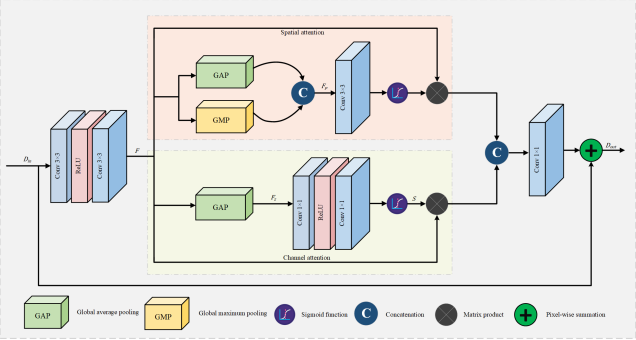


Image enhancement

- image enhancement techniques
  - general supervised denoising using DL
  - blind denoising using DL - remove noise without prior knowledge of noise adapting to various noise types
  - super-resolution - upscale low-resolution images, add realistic details for sharper & higher-quality images



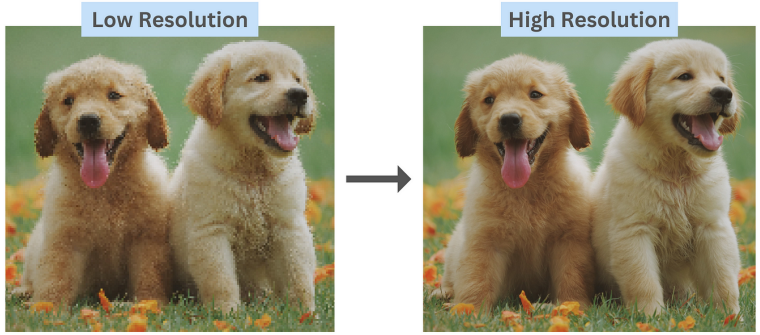
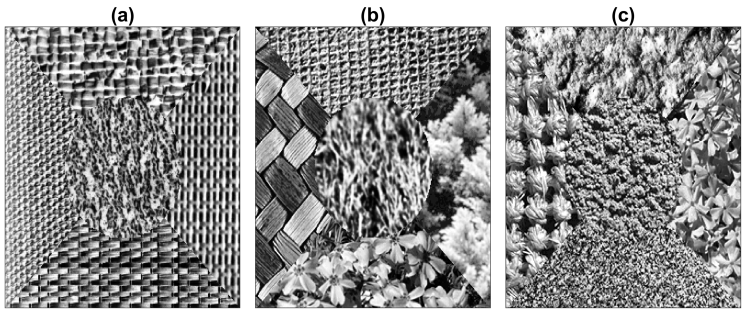


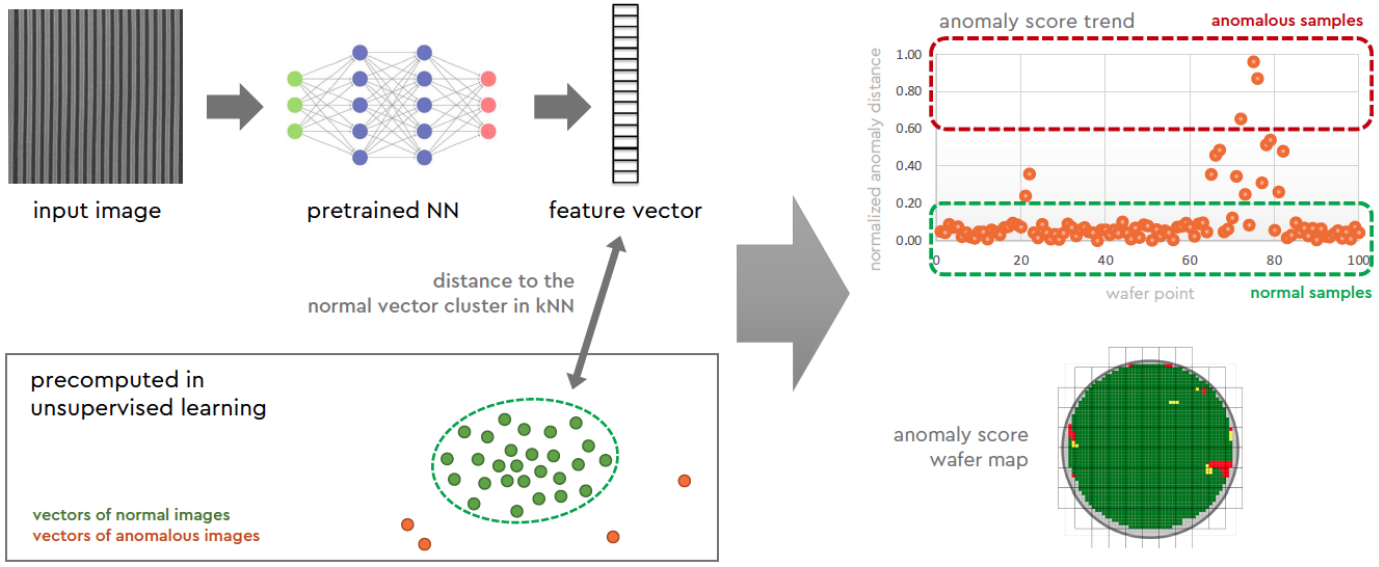
Image segmentation

- texture segmentation
  - distinguish areas based on texture patterns - identifying regions with similar textural features - used for material classification, surface defect detection, medical imaging
  - methods - Gabor filters, wavelet transforms, DL
- semantic segmentation
  - assign class labels to every pixel - enabling precise object and region identification - used for autonomous driving, scene understanding, medical diagnostics
  - methods - fully convolutional network (FCN), U-net, DeepLab



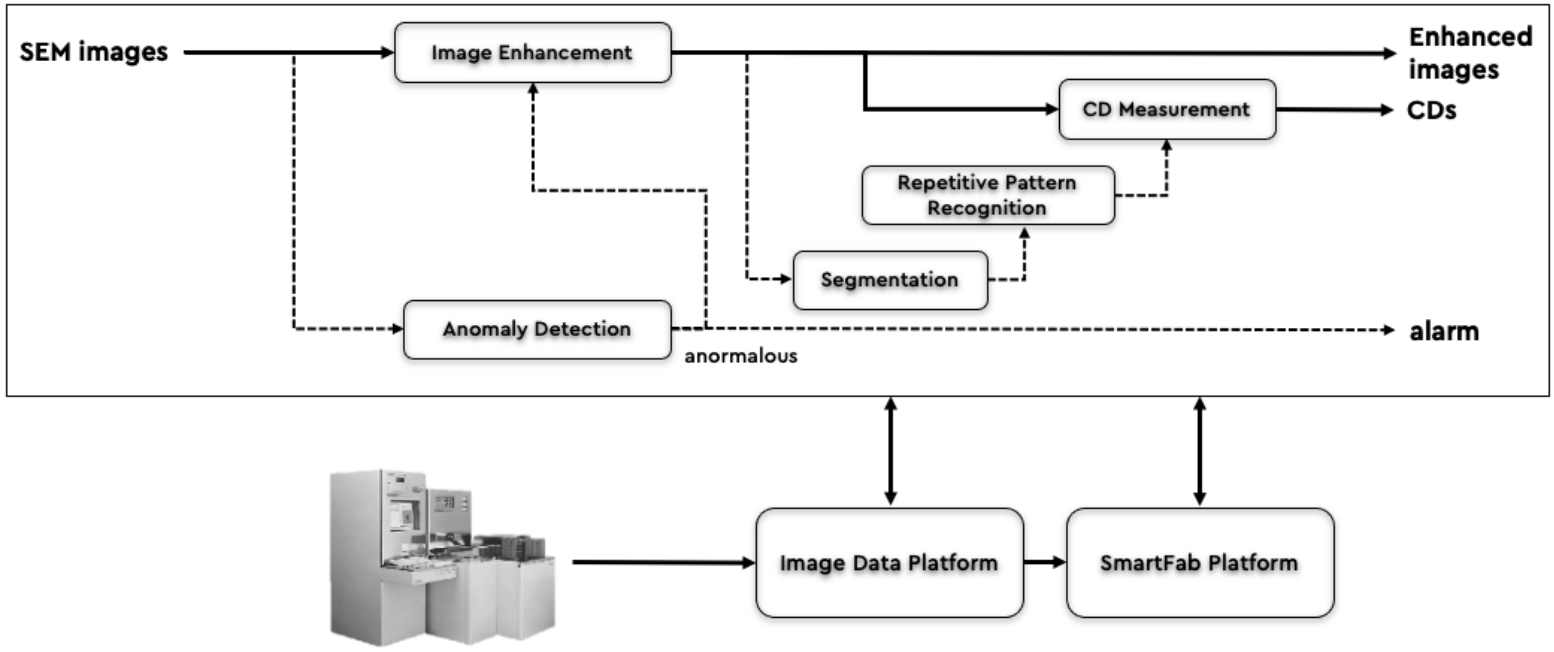
Anomaly detection using side product

- representation in embedding space obtained as side product from previous processes
- distance from normal clusters used for anomaly detection
- can be used for yield drop prediction and analysis



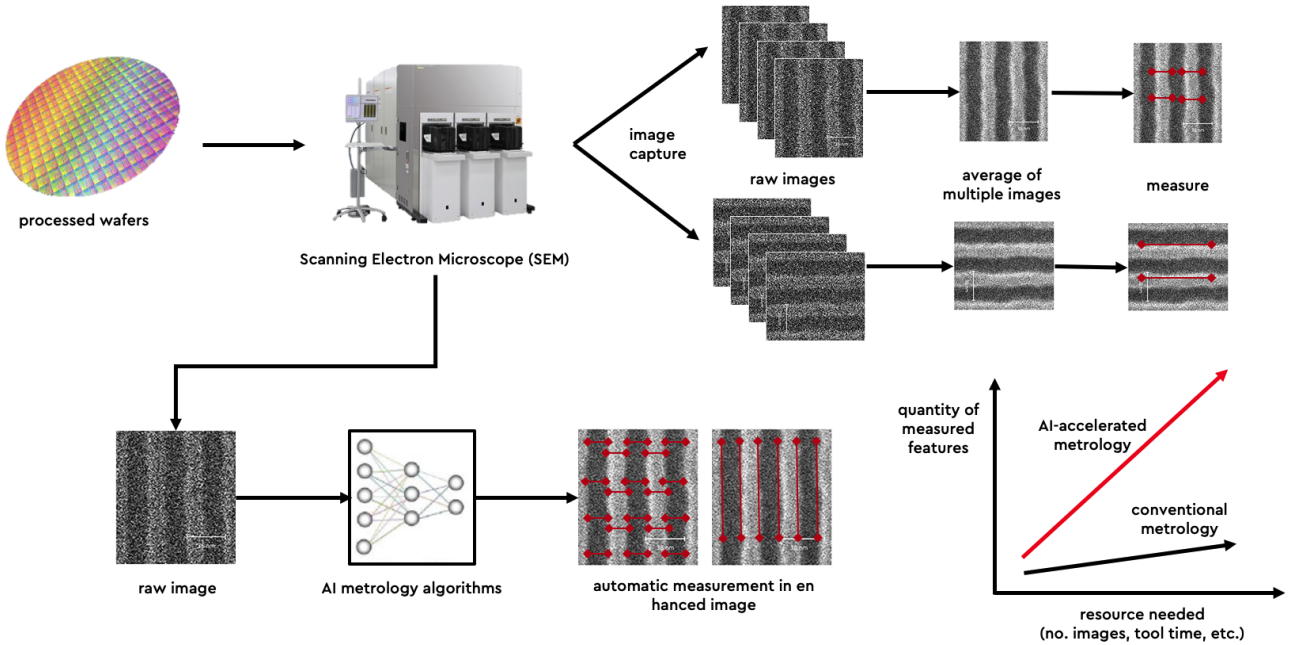
AI-enabled metrology system

- integration of separate components creates AI-enabled metrology system



Benefits of new system

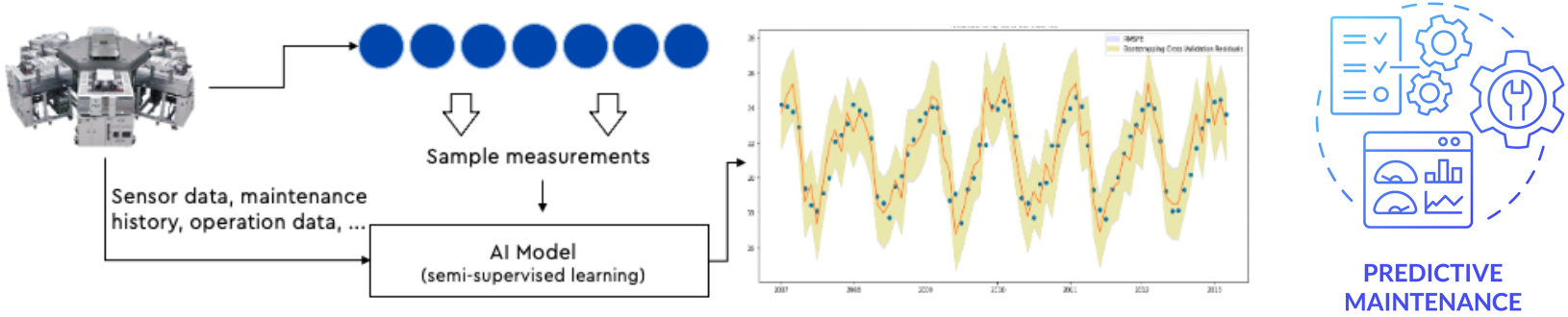
- new system provides
  - improved accuracy and reliability
  - improved throughput
  - savings on investment on measurement equipment



TS ML in manAI

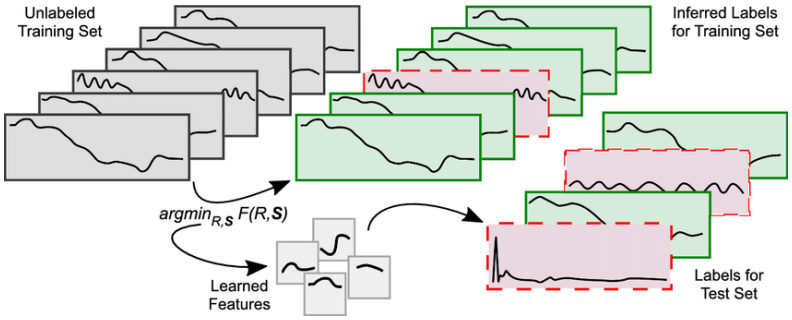
Time-series ML applications in manAI

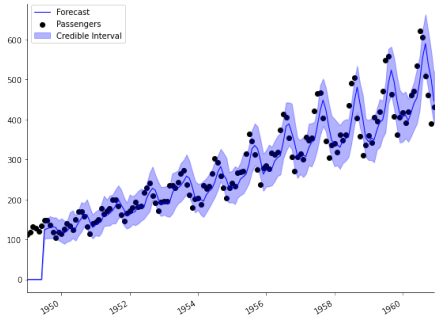
- estimation of TS values
  - virtual metrology - estimate measurement without physically measuring things
- anomaly detection on TS
  - predictive maintenance - predict maintenance times ahead
- multi-modal ML using LLM & genAI
  - root cause analysis and recommendation system



TS MLs in manAI

- TS regression/prediction/estimation
  - LSTM, GRU, attention-based models, Transformer-based architecture for capturing long-term dependencies and patterns
- anomaly detection
  - isolation forest, autoencoders, one-class SVM
- TS regression providing credibility intervals
  - Bayesian-based approaches offering uncertainty estimation alongside predictions



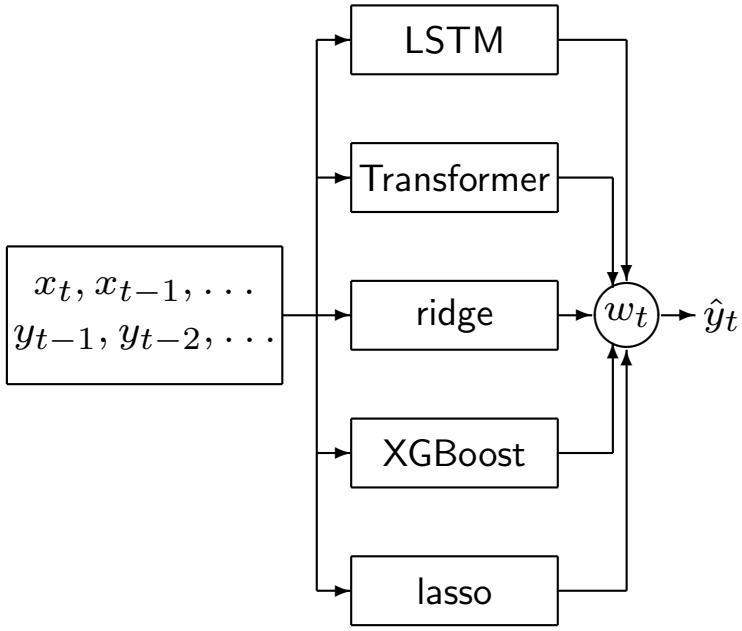


Difficulties with TS ML

- no definition exists for general TS data
- data drift & shift
  - $p(x_{t_k}, x_{t_{k-1}}, \ldots)$ changes over time
  - $p(y_{t_k}|x_{t_k}, x_{t_{k-1}}, \ldots, y_{t_{k-1}}, y_{t_{k-2}}, \ldots)$ changes over time
- (extremely) fat data, poor data quality, huge volume of data to process
- not many research results available
- none of algorithms in academic papers work / no off-the-shelf algorithms work

Online learning for TS regression

- use multiplie experts - $f_{1,k}, \ldots, f_{p_k,k}$ for each time step $t = t_k$ where $f_{i,k}$ can be any of following
  - seq2seq models ($e.g.$, LSTM, Transformer-based models)
  - non-DL statistical learning models ($e.g.$, online ridge regression)
- model predictor for t_k, $g_k : \mathbf{R}^n \to \mathbf{R}^m$ as weighted sum of experts



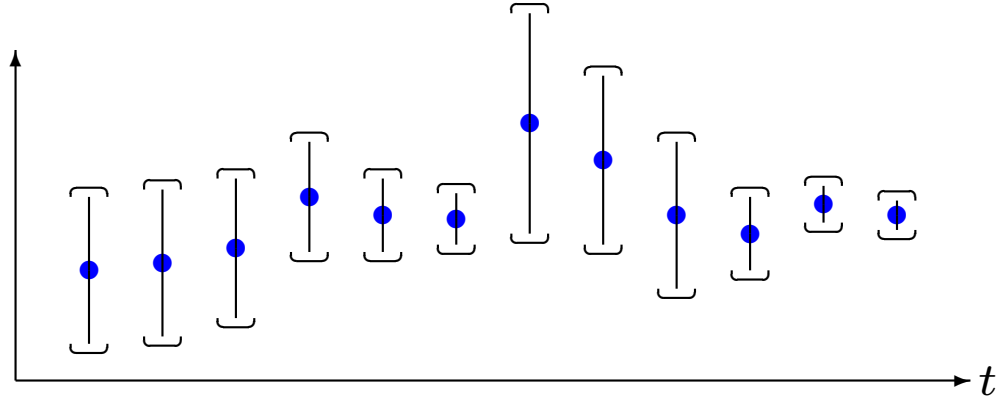
$$g_k = w_{1,k} f_{1,k} + w_{2,k} f_{2,k} + \cdots + w_{p_k,k} f_{p_k,k} = \sum_{i=1}^{p_k} w_{i,k} f_{i,k}$$

Credibility intervals

- every point prediction is wrong, $i.e.$

$$\mathbf{Prob}(\hat{y}_t = y_t) = 0$$

- reliability of prediction matters, however, *none* literature deals with this (properly)
- critical for our customers, $i.e.$, *such information is critical for downstream applications*
  - $e.g.$, when used for feedback control, need to know how reliable prediction resuilts are
  - sometimes *more crucial than algorithm accuracy*



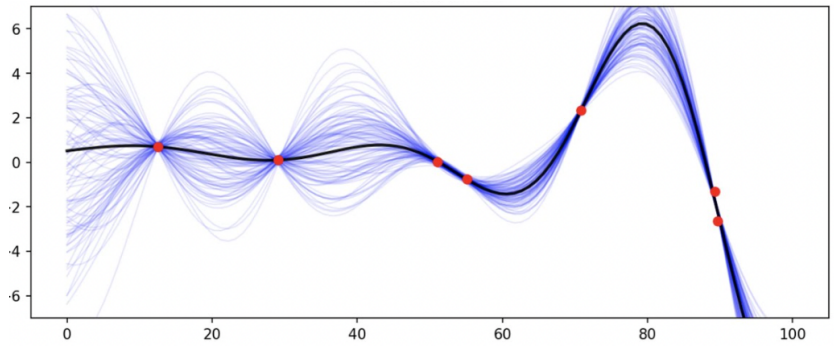
Bayesian approach for credibility interval evaluation

- assume conditional distribution ith predictor parameterized by $\theta_{i,k} \in \Theta$

$$p_{i,k}(y(t_k)|x_{t_k}, x_{t_{k-1}}, \ldots, y(t_{k-1}), y(t_{k-2}), \ldots) = p_{i,k}(y(t_k); x_{t_k}, \theta_{i,k})$$

  - depends on prior & current input, *i.e.*, $\theta_{i,k}$ & x_{t_k}
- update $\theta_{i,k+1}$ from $\theta_{i,k}$ after observing true $y(t_k)$ using Bayesian rule

$$p(w; \theta_{i,k+1}) := p(w|y(t_k); x_{t_k}, \theta_{i,k}) = \frac{p(y(t_k)|w, x_{t_k})p(w; \theta_{i,k})}{\int p(y(t_k)|w, x_{t_k})p(w; \theta_{i,k})dw}$$



Virtual Metrology

VM

- background
  - every process engineer wants to (so badly) measure every material processed - make sure process done as desired
    - $e.g.$, in semiconductor manufacturing, photolithography engineer wants to make sure diameter of holes or line spacing on wafers done correctly to satisfy specification for GPU or memory chips
  - however, various constraints prevent them from doing it, $e.g.$, in semiconductor manufacturing
    - measurement equipment requires investment
    - incur intolerable throughput
    - fab space does not allow
- GOAL - *measure every processed material without physically measuring them*

VM - problem formulation

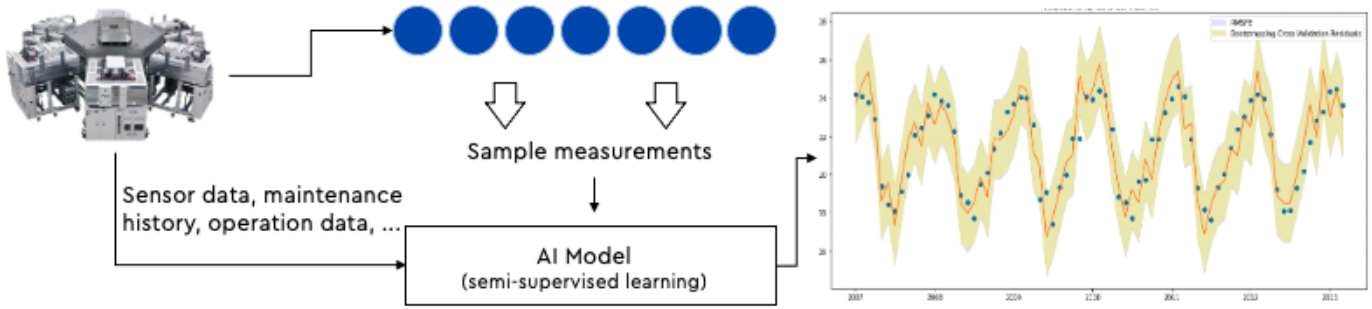
- problem description

$$\begin{array}{rl} \text{(stochastically) predict} & y_{t_k} \\ \text{given} & x_{t_k}, x_{t_{k-1}}, \ldots, y_{t_{k-1}}, y_{t_{k-2}}, \ldots \end{array}$$

- our problem formulation

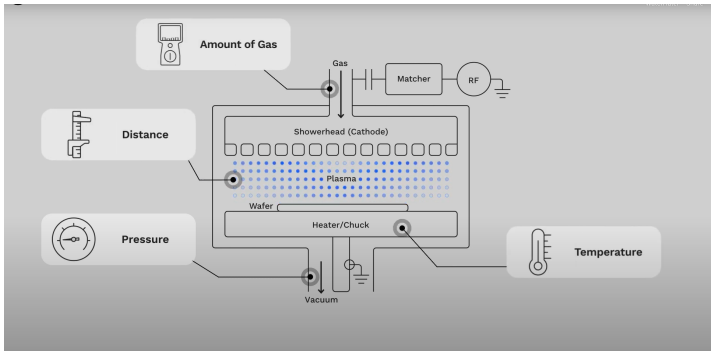
$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{K} w_{k,K-k}\, l(y_{t_k}, \hat{y}_{t_k}) \\ \text{subject to} & \hat{y}_{t_k} = g_k(x_{t_k}, x_{t_{k-1}}, \ldots, y_{t_{k-1}}, y_{t_{k-2}}, \ldots) \end{array}$$

where optimization variables - $g_1, g_2, \ldots : \mathcal{D} \to \mathbf{R}^m$



VM - Gauss Labs' inAI success story

- Gauss Labs' ML solution & AI product
  - fully home-grown online TS adpative ensemble learning method
  - outperform competitors and customer inhouse tools, *e.g.*, *Samsung, Intel, Lam Research*
  - published & patented in US, Europe, and Korea
- business impacts
  - improve process quality - reduction of process variation by tens of percents
  - (indirectly) contribute to better product quality and yield
  - Gauss Labs' main revenue source



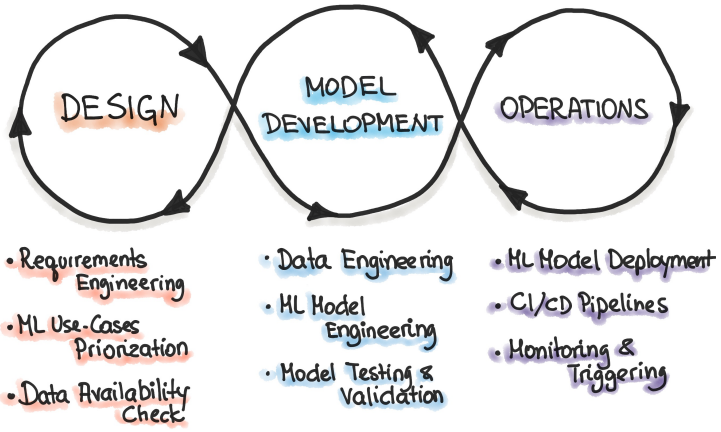
Manufacturing AI Productionization

Minimally required efforts for manAI

- MLOps - for CI/CD
- data preprocessing - missing values, inconsistent names, difference among different systems
- feature extraction & selection
- monitoring & retraining
- notification, via messengers or emails
- mainline merge approvals by humans
- data latency, data reliability, & data availability

MLOps for manAI

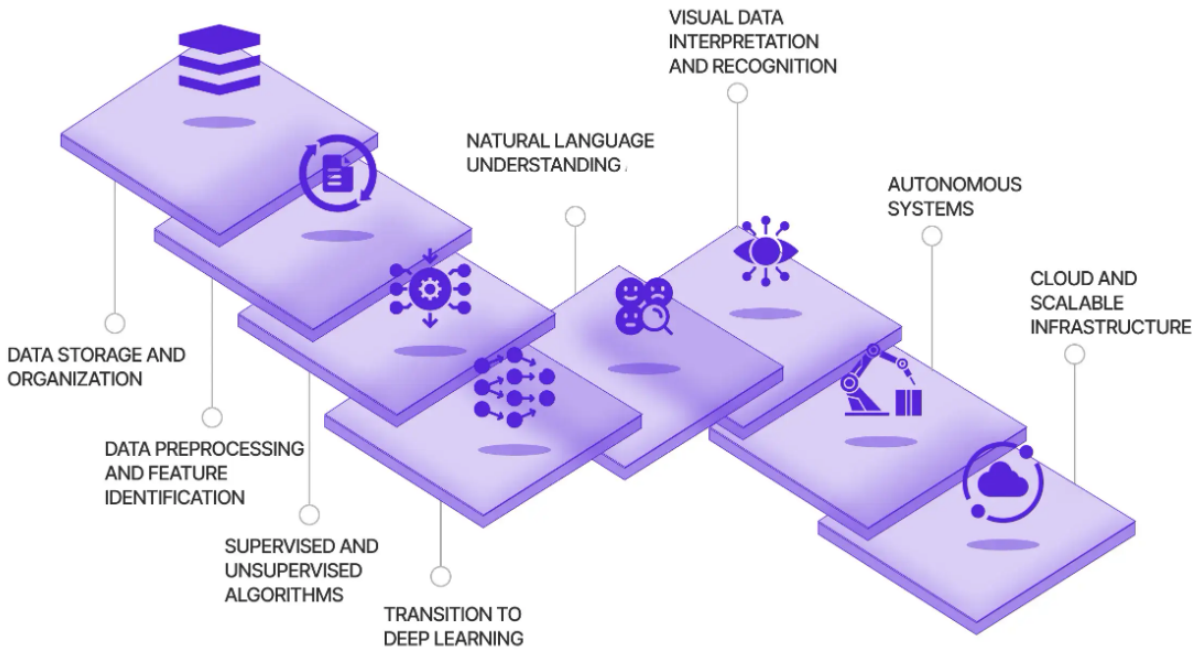
- environment for flexible and agile exploration - EDA[7]
- fast & efficient iteration of algorithm selection, experiements, & analysis
- correct training / validation / test data sets critical!
- seamless productionization from, $e.g.$, Jupyter notebook to production-ready code
- monitorning, *right* metrics, notification, re-training



[7] EDA - exploratory data analysis

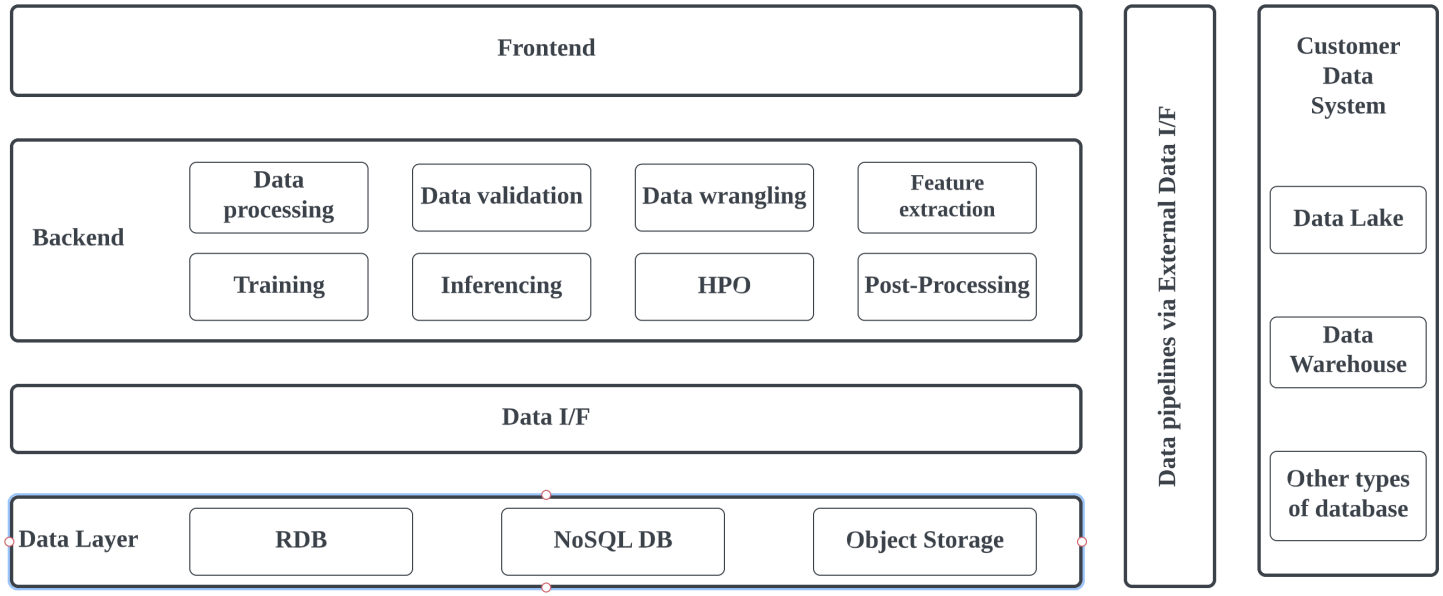
manAI software system

- data, data, data! – store, persist, retrieve, data quality
- seamless pipeline for development, testing, running deployed services
- development envinroment should be built separately



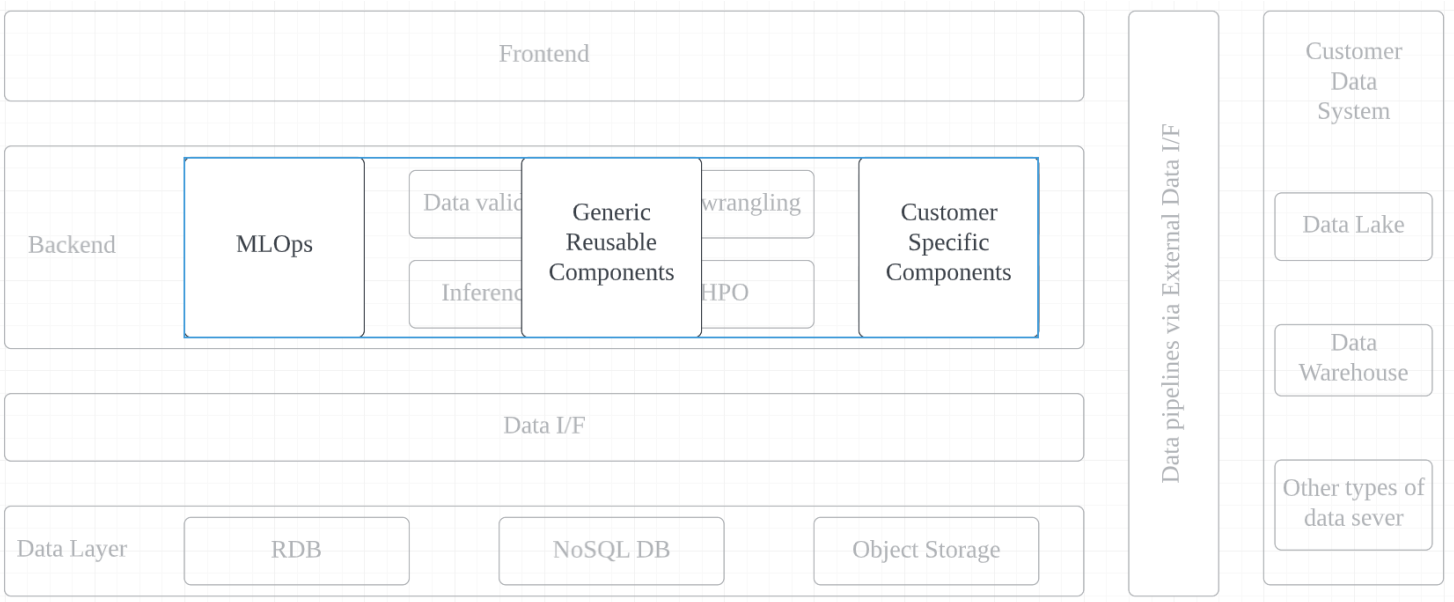
manAI system architecture

- frontend / backend / data I/F / data layer
- efficient and effective MLOps in backend or development environment



Reusuable components vs customer specific components

- make sure to build two components separate - generic reusable and customer specific
- generic models should be tuned for each use case
- generic model library grows as interacting with more and more customers



My Two Cents

Recommendations for maximum impact via inAI

- concrete goals of projects
  - north star – yield improvement, process quality, making engineers' lives easier
  - hard problem – scheduling and optimization
- be strategic!
  - learn from others – lots of successes & failures of inAI
  - ball park estimation for ROI cricial – efforts, time, expertise, data
  - utilities vs technical excellency / uniqueness vs common technology
  - home-grown vs off-the-shelf

Remember . . .

- data, data, data! – readiness, quality, procurement, pre-processing, DB
- *never* underestimate domain knowledge & expertise – data do NOT tell you everything
- EDA
- do *not* over-optimize your algorithms – ML is all about trials-&-errors
- overfitting, generalization, concept drift/shift - way more important than you could ever imagine
- devOps, MLOps, agile dev, software development & engineering

Conclusion

Conclusion

- various CV MLs used for inAI applications
- TS ML applications found in every place in manufacturing
- drift/shift & data noise make TS MLs very challenging, but working solutions found
- in reality, crucial bottlenecks are
  - data quality, prepocessing, monitoring, notification, and retraining
  - data latency, avaiability, and reliability
  - excellency in software platform design and development using cloud services

Some Important Questions

Some important questions around AI

- why human-level AI in the first place?
- what lies in very core of DL architecture? what makes it work amazingly well?
- biases that can hurt judgement, decision making, social good?
- ethical and legal issues
- consciousness, knowledge, belief, reasoning
- future of AI

Human-level AI?

Why human-level in the first place?

- lots of times, when we measure AI performance, we say
  - how can we achieve human-level performance, $e.g.$, CV models?
- why human-level?
  - are all human traits desirable? are humans flawless?
  - aren't humans still evolving?
- advantage of AI over humans
  - $e.g.$, self-driving cars can use extra eyes, GPS, computer network
  - $e.g.$, recommendation system runs for hundreds of millions of people overnight
  - AI is available 24 / 7 while humans cannot
    . . . critical advanages for medical assitance, emergency handling
  - AI does not make more mistakes because task is repetative and tedius
  - AI does not request salary raise or go on strike

What makes DL so successful?

Factors constributing to astonishing success of DL

- analysis based on speaker's mathematical, numerical algorithmic & statistical perspectives considering hardware innovations

  30% universal approximation theorem? - (partially) yes! but that's not all

  - function space of neural network is *dense* (math theory), $i.e.$, for every $f : \mathbf{R}^n \to \mathbf{R}^m$, exists $\langle f_n \rangle$ such that $\lim_{n\to\infty} f_n = f$

  25% architectures/algorithms tailored for each class of applications, $e.g.$, CNN, RNN, Transformer, NeRF, diffusion, GAN, VAE, . . .

  20% data labeling - expensive, data availability - unlimited web text corpus

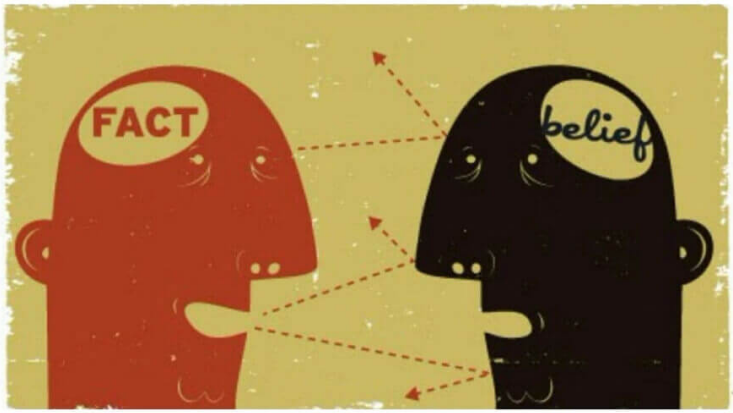
  15% computation power/parallelism - AI accelerators, $e.g.$, GPU, TPU & NPU

  10% rest - Python, open source software, cloud computing, MLOps, . . .

Biases - by Humans & Machines

Cognitive biases

- cognitive biases [Kah11]
  - confirmation bias, availability bias
  - hindsight bias, confidence bias, optimistic bias
  - anchoring bias, halo effect, framing effect, outcome bias
  - belief bias, negativity bias, false consensus,



LLM biases

- plausible with LLM
  - availability bias - baised by imbalancedly available information
    - LLM trained by imbalanced # articles for specific topics
  - belief bias - derive conclusion not by reasoning, but by what it saw
    - LLM eaisly inferencing what it saw, $i.e.$, data it trained on
  - halo effect - overemphasize on what prestigious figures say
    - LLM trained by imbalanced # reports about prestigious figures
- similar facts true for other types of ML models,
  - $e.g.$, video caption, text summarization, sentiment analysis
- cognitive biases only human represent
  - confirmation bias, hindsight bias, confidence bias, optimistic bias, anchoring bias, negativity bias, framing effect

Ethical and Legal Issues

Ethics - possibilities & questions

- AI can be exploited by those who have bad intention to
  - manupilate / deceive people - using manupilated data corpus for training
    - $e.g.$, spread false facts
  - induce unfair social resource allocation
    - $e.g.$, medical insurance, taxation
  - exploit advantageous social and economic power
    - $e.g.$, unfair wealth allocation, mislead public opinion
- AI for Good - advocated by Andrew Ng
  - $e.g.$, public health, climate change, disaster management
- should scientists and engineers be morally & politically conscious?
  - $e.g.$, Manhattan project

Ethically controversial issues

- AI girlfriends
  - lots of AI girlfriend apps already developed
  - ethical considerations and provisions for user privacy with AI partners imperative - as with every technology involving personal data and emotional interaction
  - prospect of developing lifelike digital companions will grow better with evolution of AI
  - perhaps changing ways relationships and companionship perceived in digital age one day
  - why not many AI boyfriend apps? is this sexual discrimination issue (at all)?

Legal issues with ethical consideration - (hypothetical) scenarios

- scenario 1: full self-driving algorithm causes traffic accident killing people
  - who is responsible? - car maker, algorithm developer, driver, algorithm itself?
- scenario 2: self-driving cars kill less people than human drivers
  - $e.g.$, human drivers kill 1.5 people for 100,000 miles & self-driving cars kill 0.2 people for 100,000 miles
  - how should law makers make regulations?
  - utilitarian & humanistic perspectives
- scenario 3: someone is not happy with their data being used for training
  - "The Times sues OpenAI and Microsoft over AI use of copyrighted work" (Dec. 2023)
  - "Newspaper publishers in California, Colorado, Illinois, Florida, Minnesota and New York said Microsoft and OpenAI used millions of articles without payment or permission to develop ChatGPT and other products" (Apr. 2024)

Consciousness

Consciousness

- what is consciousness, anyway?
  - recognizes itself as independent, autonomous, valuable entity?
  - recognizes itself as living being, unchangeable entity?
  - will to survive?
- no agreed definition on consciousness exists yet
  . . . and will be so forever
- can it be seperated from fact that humans are biological living being?
  - (speaker) doesn't think so . . .
- is SKYNET ever plausible (without someone's intention)?
  - can AI have *desire* to survive (or save earth)?

Utopia or dystopia

- not important questions (speaker thinks)
  - what we should worry about is not doomday or destoying humankind
- but rather we should focus on
  - our limit in controlling or unintended consequences of AI
  - misuse by those possessing social, economic, political power
  - social good and welfair imparied by (exploting of) AI
  - choice among utilitarianism / humanism / justice / equity
  - handle ethical and legal issues

Knowledge, Belief, and Reasoning of AI

Does LLM (or AI) have knowledge or belief? Can it reason?

What categories of questions should they be?

Philosophical? Cognitive scientific?

Three surprises of LLM

- LLM is very different sort of animal . . . except that it is *not* an animal!
- *unreasonable* effectiveness of data [HNF09]
  - *performance scales with size of training data*
  - *qualitative leaps* in capability as models scale
  - tasks demanding human intelligence *reduced to next token prediction*
- focus on third surprise

  "conditional probability model looks like human with intelligence"

  - making vulnerable to anthropomorphism
- examine it by throwing questions
  - *"does LLM have knowledge and belief?"*
  - *"can it reason?"*

Knowledge, belief & reasoning around LLM

- *not* easy topic to discuss, or even impossible because
  - we do *not* have agreed definition of these terms especially in context of being asked questions like

    does ChatGPT have belief?

    or

    do humans have knowledge?

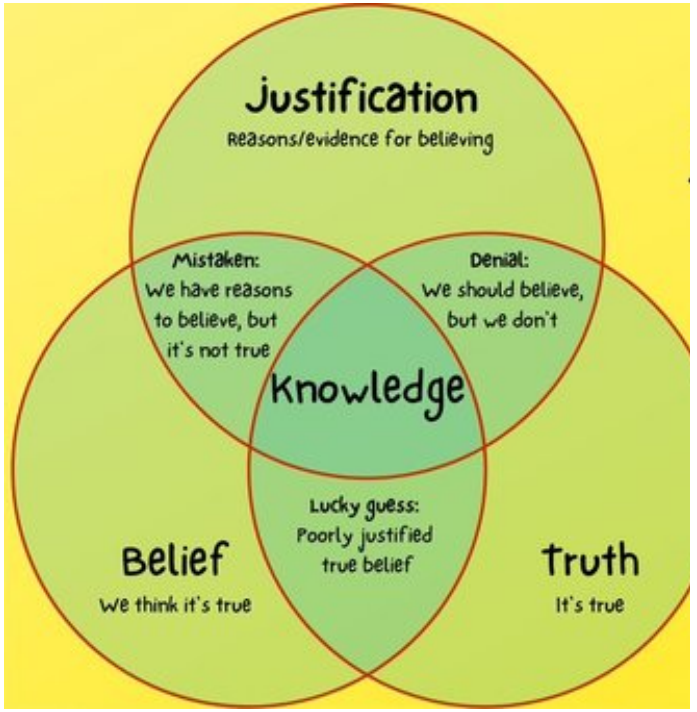
- let us discuss them in two different perspectives
  - laymen's perspective
  - cognitive scientific perspective

Laymen's perspective on knowledge, belief & reasoning

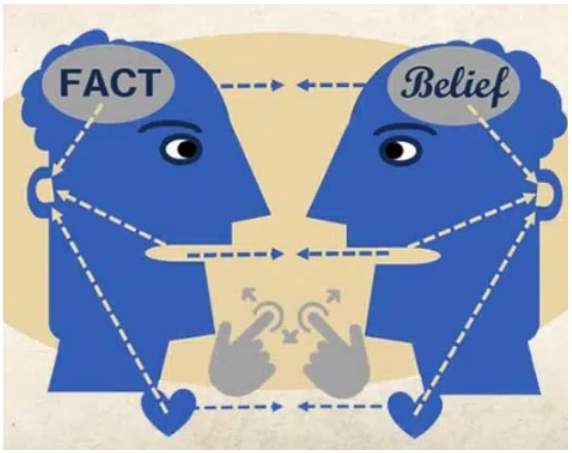
- does (good) LLM have knowledge?
  - Grandmother - looks like it cuz when instructed *"explaing big bang"*, it says
    " The Big Bang theory is prevailing cosmological model that explains the origin and evolution of the universe. . . . 13.8 billion years ago . . . "
- does it have belief?
  - Grandmother: I don't think so, $e.g.$, it does not believe in God.
- can it reason?
  - Grandmother: seems like it! $e.g.$, when asked *"Sunghee is a superset of Alice and Beth is a superset of Sunghee. is Beth a superset of Alice?"*, it says
    " Yes, based on information provided, if Sunghee is a superset of Alice and Beth is a superset of Sunghee, then Beth is indeed a superset of Alice . . . "
- can it reason to prove theorem whose inferential structure is more complicated?
  - Grandmother: I'm not sure. - actually, I don't know what you're talking about!

Cognitive scientific perspective on knowledge

- does LLM have knowledge?
  - Speaker: I don't think so.
- why?
  - Speaker: we say we have "knowledge" when
    "we do so against ground of various human capacities that we all take for granted when we engage in everyday conversation with each other."
    — LLM *cannot* do this.
  - Speaker: also when asked *"who is Tom Cruise's mother?"*, it says *"Tom Cruise's mother is Mary Lee Pfeiffer."* However, this is nothing but
    "guessing" by conditional probability model the most likely following words after "Tom Cruise's mother is."
  - Speaker: so *we cannot say it really knows the fact!*



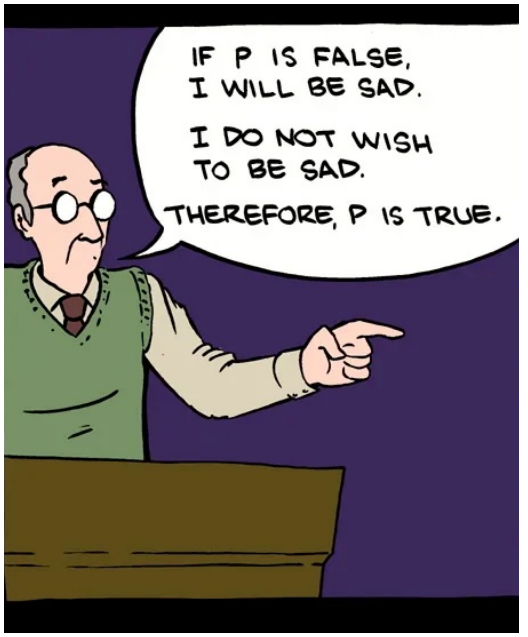
Cognitive scientific perspective on belief



- for the discussion
  - we do not concern *any specific belief*
  - we concern prerequisites for ascribing any beliefs to AI system
- so does it have belief?
  - Speaker: nothing can count as belief about the world we share unless
    > *it is against ground of the ability to update beliefs appropriately in light of evidence from that world, an essential aspect of the capacity to distinguish truth from falsehood.*
  - Speaker: when a human being takes to Wikipedia and confirms some fact, what happens is not her language model update, but
    > *reflection of her nature as language-using animal inhabiting shared world with a community of other language-users.*
  - Speaker: LLM does not have this ground, an essential consideration when deciding whether it *really* had beliefs.
  - Speaker: so *no, LLM cannot have belief!*

Cognitive scientific perspective on reasoning

- note reasoning is *content neutral*
  - *e.g.*, following logic is perfect regardless of truth of premises
    if Socrates is a human and humen are immortal, then Socrates would have survived today.
- Speaker: when asked *"if humans are immortal, would Socrates have survived today?"*, LLM says
  " . . . it's logical to conclude that Socrates would likely still be alive today. . . . "
  - however, remember, once again, what we just asked it to do is *not* "deductive inference", but
    given the statistical distribution of words in public corpus, what words are likely to follow the sequence, "humans are immortal and Socreates is human therefore."
- Speaker: so LLM *cannot* or rather *does not* reason
- however, LLM can *mimic even multi-step reasoning whose inferencing structure is complicated* using *in-context learning* or *few-short prompting*!



A simple example supporting reasoning incapability

- You

  "Who is Tom Cruise's mother?"

- ChatGPT

  "Tom Cruise's mother is Mary Lee Pfeiffer. She was born Mary Lee South. . . . Information about his family, including his parents, has been publicly available, *. . . "*

- You

  "Who is Mary Lee Pfeiffer's son?"

- ChatGPT

  "As of my last knowledge update in January 2022, I don't have specific information about Mary Lee Pfeiffer or her family, including her son. *. . . "*

Future of AI

Aschenbrenner's essay

- Leopold Aschenbrenner, who left OpenAI shwoing concerns about safety, wrote *epic 165-page treatise* - Jun-2024
  - rapid progress
    - AI development (is) accelerating at unprecedented rate, predicting by 2027, AI models lead to intelligence explosion surpassing human intelligence
  - economic and security implications
    - trillions of dollars being invested into infrastructure supporting AI systems
    - critical need for securing technologies to prevent misuse, $e.g.$, by state actors like Chinese Communist Party (CCP)
  - technical and ethical challenges
    - significant challenges in controlling AI (smarter than humans), $i.e.$, "superalignment" problem, to prevent catastrophic outcomes
  - predictions and societal impact
    - few people truly understand scale of change by AI
    - potential for AI to reshape industries, enhance national security
    - pose new ethical and governance challenges

More about Aschenbrenner's essay

- AGI by 2027
  - seen AI advancing from preschool-level to high-schooler abilities in 4 years highlighting rapid progress from GPT-2 to GPT-4
- superintelligence following AGI - post AGI
  - rapid advancement from human-level to superhuman capabilities
- G-dollar investment on AI clusters
- national & global security dynamics
  - may lead to all-out war, $e.g.$, with China, if not managed properly
- superalignment challenges
  - keeping superintelligent AI aligned with human values and interests - "one of the most critical predictions"
- societal and economic transformations, project involvement by US government, technological mobilization

Moral

Moral

- AI, $e.g.$, LLM, shows incredible utility and commercial potentials, hence we should
  - make informed decisions about trustworthiness and safety
  - avoid ascribing capacities they lack

- today's AI is so powerful, so (seemingly) convincingly intelligent
  - obfuscate mechanism
  - actively encourage *anthropomorphism* with philosophically loaded words like "believe" and "think"
  - easily mislead people about character and capabilities of AI

- matters not only to scientists, engineers, developers, and entrepreneurs, but also
  - *general public, policy makers, media people*

Recent AI Development

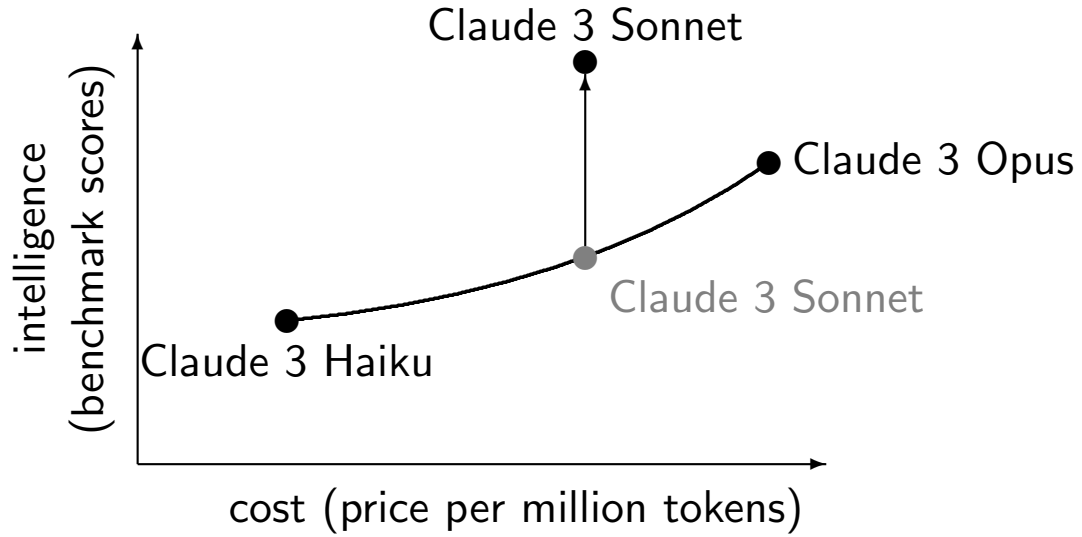
Notable recent AI research and new development

- Claude 3.5 Sonnet
- Kolmogorov–Arnold networks (KAN)
- JEPA ($e.g.$, I-JEPA & V-JEPA) & consistency-diversity-realism trade-off

Claude 3.5 Sonnet

Claude 3.5 Sonnet

- Anthropic
  - releases Claude 3.5 Sonnet (Jul-2024)
    - when! GPT-4o accepted to be default best model for many tasks, $e.g.$, reasoning & summarization
  - claims Claude 3.5 Sonnet sets *new industry standard for intelligence*



Main features & performance

- Claude 3.5 Sonnet shows off
  - improved vision tasks, 2x speed (compared to GPT-4o), artifacts - new UIs for, $e.g.$, code generation & animation
- with GPT-4o, Claude 3.5 Sonnet
  - wins at code generation
  - on par for logical reasoning
  - loses at logical reasoning
  - *wins at generation speed*

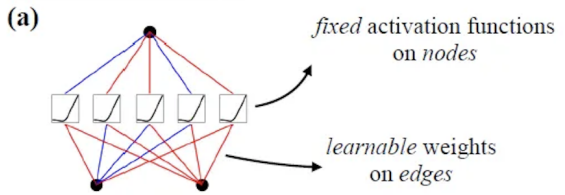
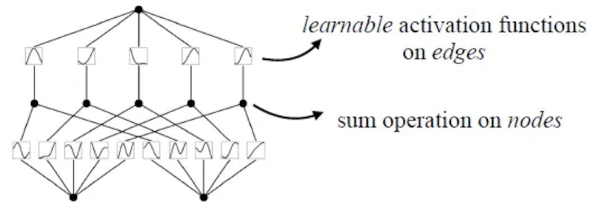
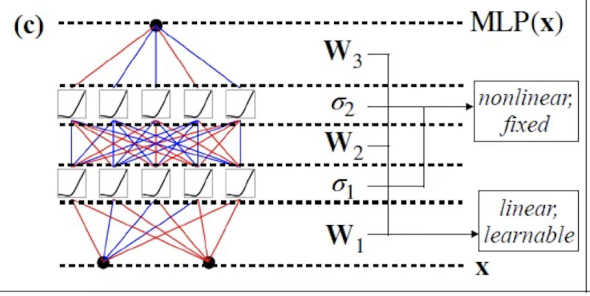
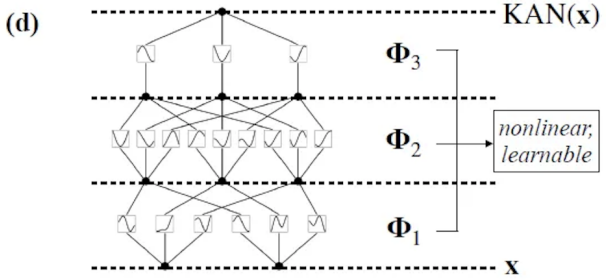
	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
visual math reasoning	67.7%	50.5%	63.8%	63.9%
science diagrams	94.7%	88.1%	94.2%	94.4%
visual question answering	68.3%	59.4%	69.1%	62.2%
chart Q&A	90.8%	80.8%	85.7%	87.2%
document visual Q&A	95.2%	89.3%	92.8%	93.1%

KAN

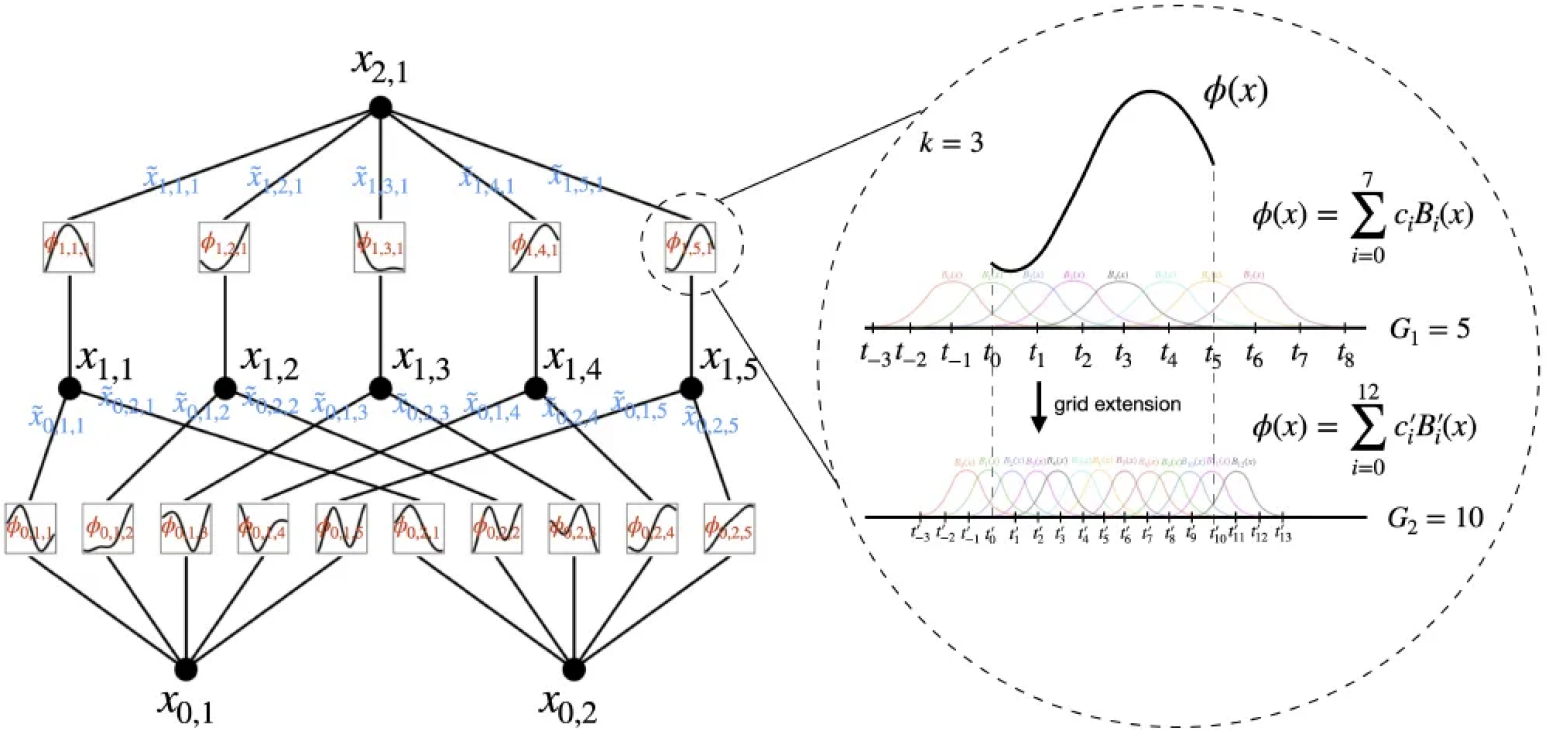
Kolmogorov–Arnold networks (KAN)

- KAN: Kolmogorov-Arnold Networks - MIT, CalTech, Northeastern Univ. & IAIFI
- techniques
  - inspired by Kolmogorov-Arnold representation theorem - every $f : \mathbf{R}^n \to \mathbf{R}$ can be written as finite composition of continuous functions of single variable, $i.e.$
    $$f(x) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$$
    where $\phi_{q,p} : [0, 1] \to \mathbf{R}$ & $\Phi_q : \mathbf{R} \to \mathbf{R}$
  - replace (fixed) activation functions with learnable functions
  - use B-splines for learnable (uni-variate) functions - for flexibility & adaptability
- advantages
  - benefits structure of MLP on outside & splines on inside
  - reduce complexity and # parameters to achieve accurate modeling
  - *interpretable* by its nature
  - *better continual learning* - adapt to new data without forgetting thanks to local nature of spline functions

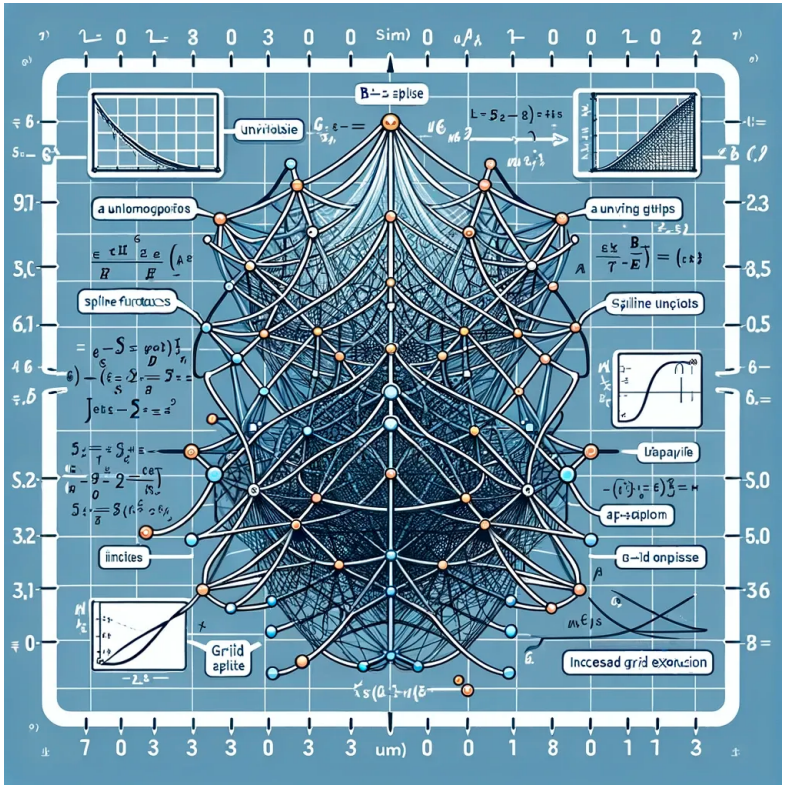
MLP vs KAN

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$
Model (Shallow)		
Formula (Deep)	$\mathrm{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\mathrm{KAN}(\mathbf{x}) = (\mathbf{\Phi}_3 \circ \mathbf{\Phi}_2 \circ \mathbf{\Phi}_1)(\mathbf{x})$
Model (Deep)		

KAN architecture with spline parametrization unit layer



Future work on KAN



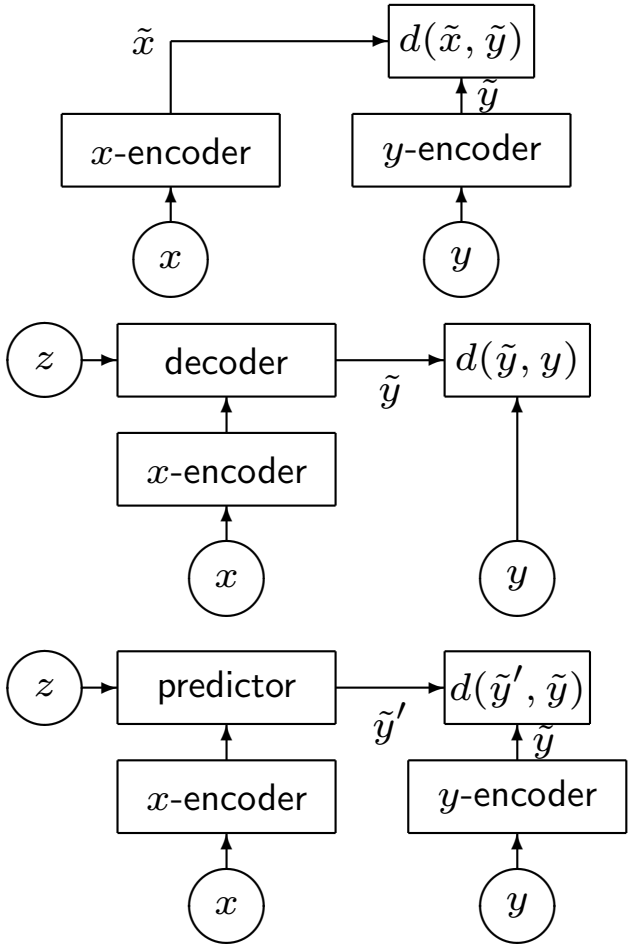
- natural question is
  - what if use both MLP and KAN?
  - what if use other types of splines?
  - how to control forgetfulness of continual learning?
  - why functions of one varible? possible to use functions of two variables?

(figure created by DALLE-3)

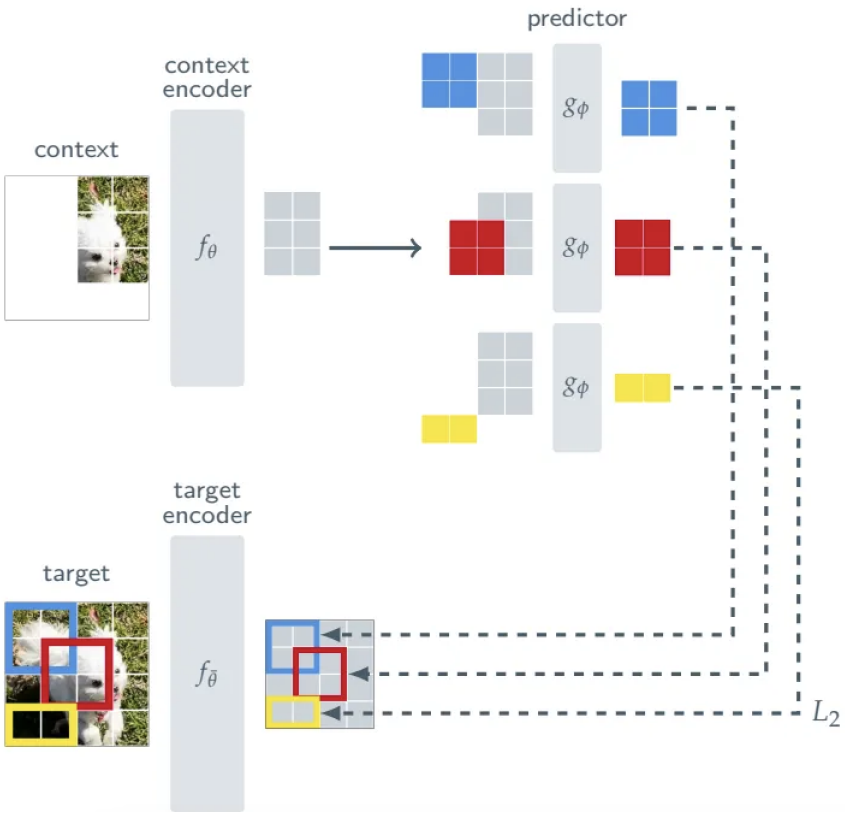
JEPA

Joint-Embedding Predictive Architecture (JEPA)

- Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (JEPA) - Yann LeCun et al. - Jan-2023
  - joint-embedding architecture (JEA)
    - output similar embeddings for compatible inputs x, y and dissimilar embeddings for incompatible inputs
  - generative architecture
    - directly reconstruct signal y from compatible signal x using decoder network conditioned on additional variables z *to facilitate reconstruction*
  - joint-embedding predictive architecture (JEPA)
    - similar to generative architecture, but comparison is done in embedding space
    - $e.g.$, I-JEPA learns y (masked portion) from x (unmasked portion) conditioned on z (position of mask)



Learning semantic representation better

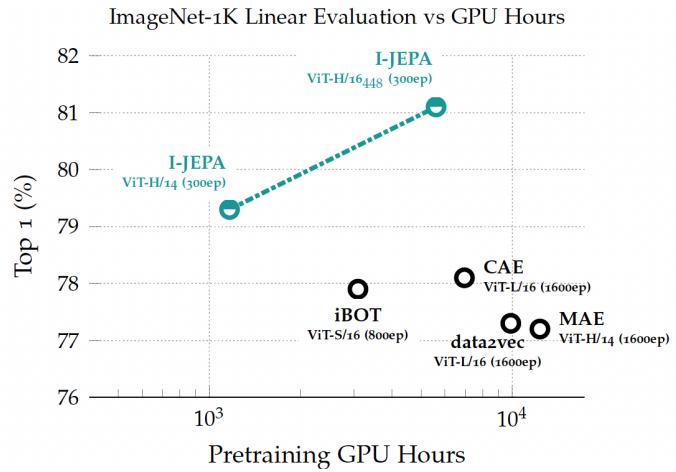


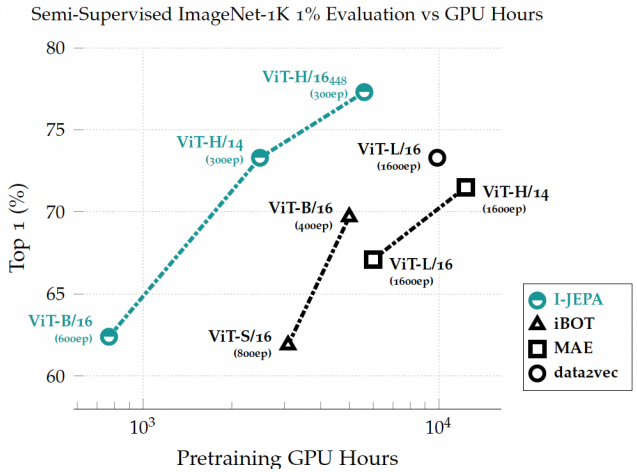
- I-JEPA
  - predicts missing information in *abstract representation space*
    - $e.g.$, given single context block (unmasked part of the image), predict representations of various target blocks (masked regions of same image) where target representations computed by learned target-encoder
  - *generates semantic representations* (not pixel-wise information) potentially eliminating unnecessary pixel-level details & allowing model to concentrate on learning more semantic features

I-JEPA outperforms other algorithms

Method	Arch.	CIFAR100	Places205	iNat18
Methods without view data augmentations				
data2vec [8]	ViT-L/16	81.6	54.6	28.1
MAE [36]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	**87.5**	**58.4**	**47.6**
Methods using extra view data augmentations				
DINO [18]	ViT-B/8	84.9	57.9	55.9
iBOT [79]	ViT-L/16	**88.3**	**60.4**	**57.3**

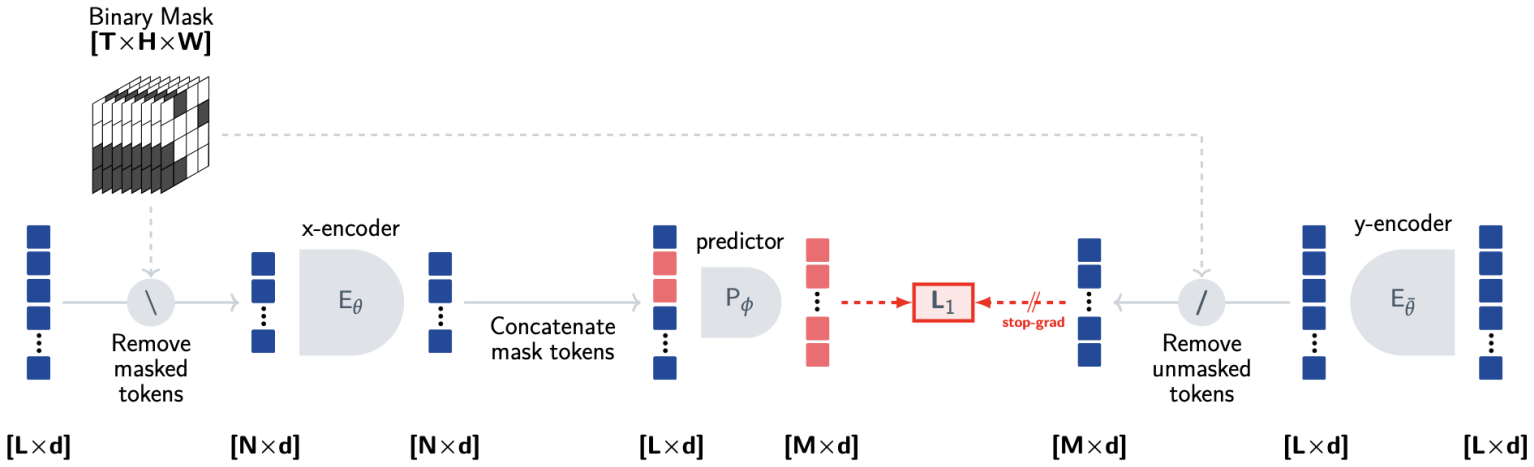
Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
Methods using extra data augmentations			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8





V-JEPA

- Revisiting Feature Prediction for Learning Visual Representations from Video - Yann LeCun et al. - Feb-2024
  - essentially same ideas of JEPA - loss function is calculated in embedding space - for better semantic representation learning (rather than pixel-wise learning)



More realistic generative model becomes, less diverse it becomes

- Consistency-diversity-realism Pareto fronts of conditional image generative models - FAIR at Meta - Montreal, Paris & New York City labs, McGill University, Mila, Quebec AI institute, Canada CIFAR AI - Jun-2024

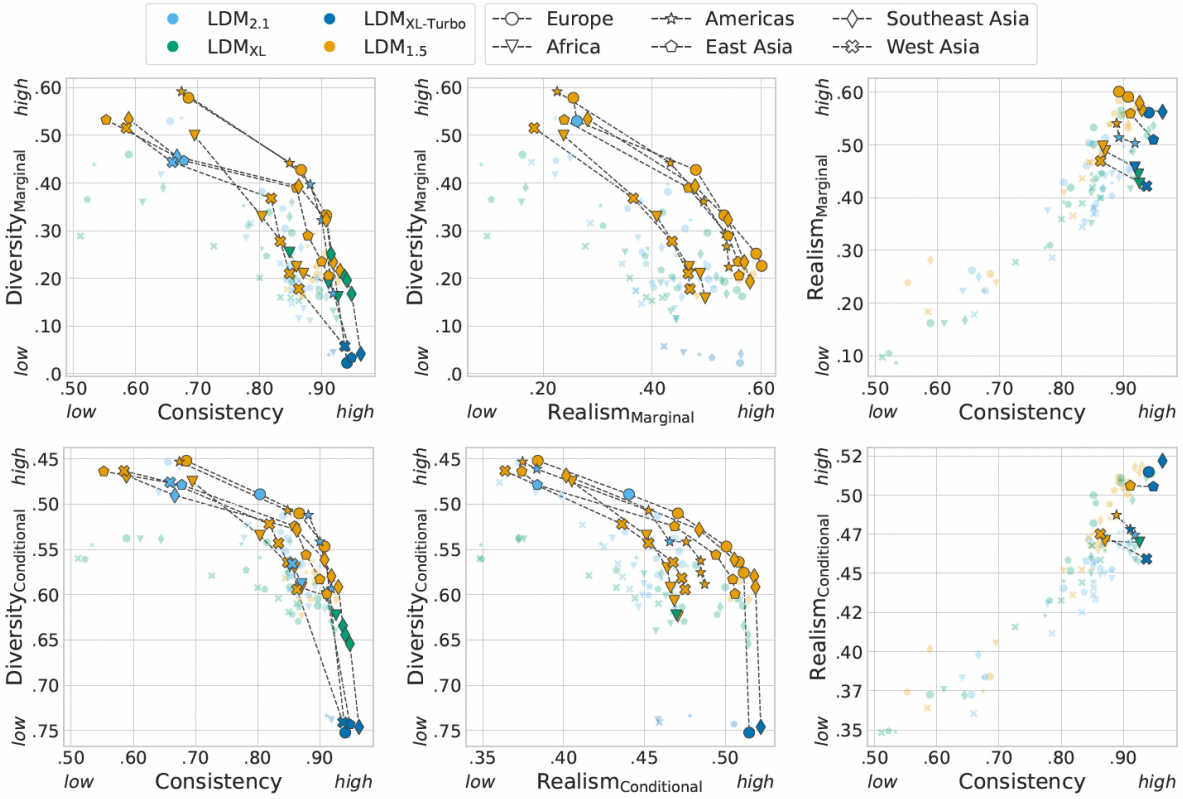
- realism comes at the cost of coverage, $i.e.$, *the most realistic systems are mode-collapsed!*

- intuition (or hunch)

  – world models should *not* be generative - should make predictions in representation space - in representation space, unpredictable or irrelevant information is absent

$\rightarrow$ main argument in favor of JEPA

Consistency-diversity-realism trade-off



Learning ML & AI

Best ways to learn ML & AI

- first, learn basics - college classes, online courses, (easy) books
  - not need to understand every mathematical details, but should know rough ideas!
- hands-on is MUST!
  - learn and practice coding - `Python` is MUST; do not do only `R`
  - learn `git` - know how to develop efficiently, plus `import` others' work
- *(I think) online cources are blessing to mankind!*
  - you *can't* say "I can't do it because resource is not available or classes of good schools are not available" because . . . they are available! :)
  - getting (expensive) certificates is good idea because . . . otherwise you wouldn't finish it! :) plus you can post it on your LinkedIn
- would be best if your task at work is related to ML
- however, even if that's not the case or can't be the case, can always do your own personal projects – or contribute to public projects (on `github`)!

Andrew Ng!

- Andrew Ng
  - (co-)founder of "Deep Learning.AI" and "Coursera", prominent figure in ML & AI
  - his courses highly regarded because well-structured and provide insights

- latest Andrew Ng courses
  - AI Agents in LangGraph
  - AI Agentic Design Patterns with AutoGen
  - Introduction to On-device AI
  - Multi AI Agent Systems with Crew AI
  - Building Multimodal Search and RAG - contrastive learning, multimodality to RAG
  - Building Agentic RAG with LlamaIndex
  - Quantisation In Depth
  - In Prompt Engineering for Vision Models
  - Getting Started with Mistral - open-source models (Mistral 7B, Mixtral 8x7B)
  - Preprocessing Unstructured Data for LLM

Selected References & Sources

Selected references & sources

- Daniel Kahneman, Thinking, Fast and Slow, 2011
- T. Kuiken, Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight, 2023
- S. Yin, et. al., A Survey on Multimodal LLMs, 2023
- M. Shanahan, Talking About Large Language Models, 2022
- A. Vaswani, et al., Attention is all you need, NeurIPS, 2017
- I.J. Goodfellow, . . . , Y. Bengio, Generative adversarial networks (GAN), 2014
- A.Y. Halevry, P. Norvig, and F. Pereira. Unreasonable Effectiveness of Data, 2009
- Stanford Vecture Investment Groups
- CEOs & CTOs @ starup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California

References

References

[ACH+24] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models, 2024.

[ADM+23] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

[BGP+24] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024.

[BKP22] Abhaya Bhardwaj, Shristi Kishore, and Dhananjay K. Pandey. Artificial intelligence in biological sciences. *Life*, 12(1430), 2022.

[Blo23] Bloomberg. Bloomberg, 2023.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[DFJ22] Thomas A. Dixon, Paul S. Freemont, and Richard A. Johnson. A global forum on synthetic biology: The need for international engagement. *Nature Communications*, 13(3516), 2022.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[HGH+22] Sue Ellen Haupt, David John Gagne, William W. Hsieh, Vladimir Krasnopolsky, Amy McGovern, Caren Marzban, William Moninger, Valliappa Lakshmanan, Philippe Tissot, and John K. Williams. The history and practice of AI in the environmental sciences. *Bulletin of the American Meteorological Society*, 103(5):E1351 – E1370, 2022.

[HM24] Guadalupe Hayes-Mota. Emerging trends in AI in biotech. *Forbes*, June 2024.

[HNF09] Alon Halevy, Peter Norvig, and Nanediri Fernando. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24:8 – 12, 05 2009.

[Kah11] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.

[Kui23] Todd Kuiken. Artificial intelligence in the biological sciences: Uses, safety, security, and oversight. *Congressional Research Service*, Nov 2023.

[KW19] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[KXS+24] Tzofi Klinghoffer, Xiaoyu Xiang, Siddharth Somasundaram, Yuchen Fan, Christian Richardt, Ramesh Raskar, and Rakesh Ranjan. Platonerf: 3D reconstruction in Plato's cave via single-view two-bounce lidar, 2024.

[LWV+24] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-arnold networks, 2024.

[Mil22] Chris Miller. *Chip war: fight for the world's most critical technology*. New York: Scribner, 2022.

[MLZ22] Louis-Philippe Morency, Paul Pu Liang, and Amir Zadeh. Tutorial on multimodal machine learning. In Miguel Ballesteros, Yulia Tsvetkov, and

Cecilia O. Alm, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 33–38, Seattle, United States, July 2022. Association for Computational Linguistics.

[P.R23] P.R. Precedence research, 2023.

[RAB+23] Ziaur Rahman, Muhammad Aamir, Jameel Ahmed Bhutto, Zhihua Hu, and Yurong Guan. Innovative dual-stage blind noise reduction in real-world images using multi-scale convolutions and dual attention mechanisms. *Symmetry*, 15(11), 2023.

[Say21] Kelley M. Sayler. Defense primer: Emerging technologies. *Congressional Research Service*, 2021.

[Sha23] Murray Shanahan. Talking about large language models, 2023.

[SSS+24] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable Gaussian codec avatars, 2024.

[Toe23] Rob Toews. The next frontier for large language models is biology. *Forbes*, July 2023.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[Wet23] Kris A. Wetterstrand. Dna sequencing costs: Data, 2023.

[YFZ+24] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[ZBX+24] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion, 2024.