[KAIST AI Seminar]
AI - Technology, Industry, Market & Hardware

Sunghee Yun

Co-founder / CTO - AI Technology @ Erudio Bio, Inc.

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*

Today

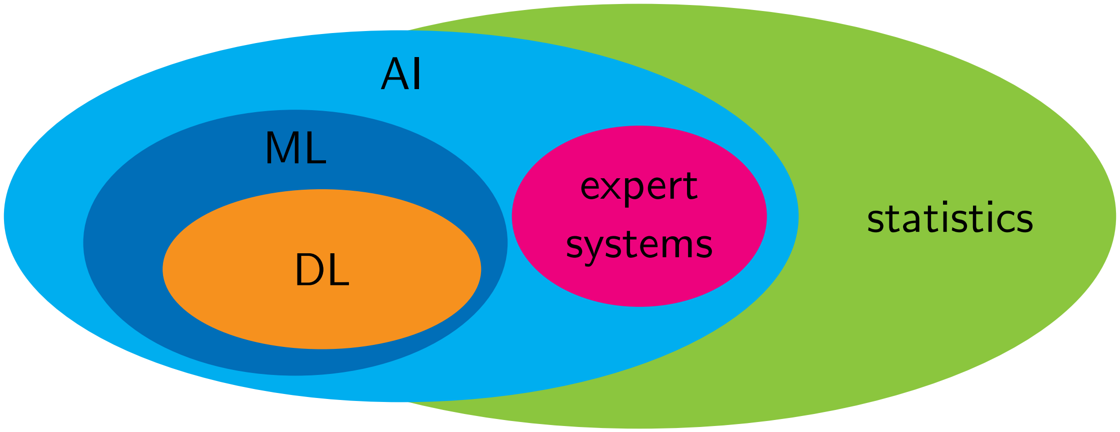
- Artificial Intelligence
  - history
  - AI achievement from 2014 to 2024
- AI research and development trend
- AI hardware
  - industry & startups
  - GPUs & AI accelerators
- global semiconductor industry
- appendices
  - some interesting and noteworthy recent AI development
  - AI products
  - AI & biotech

Artificial Intelligence

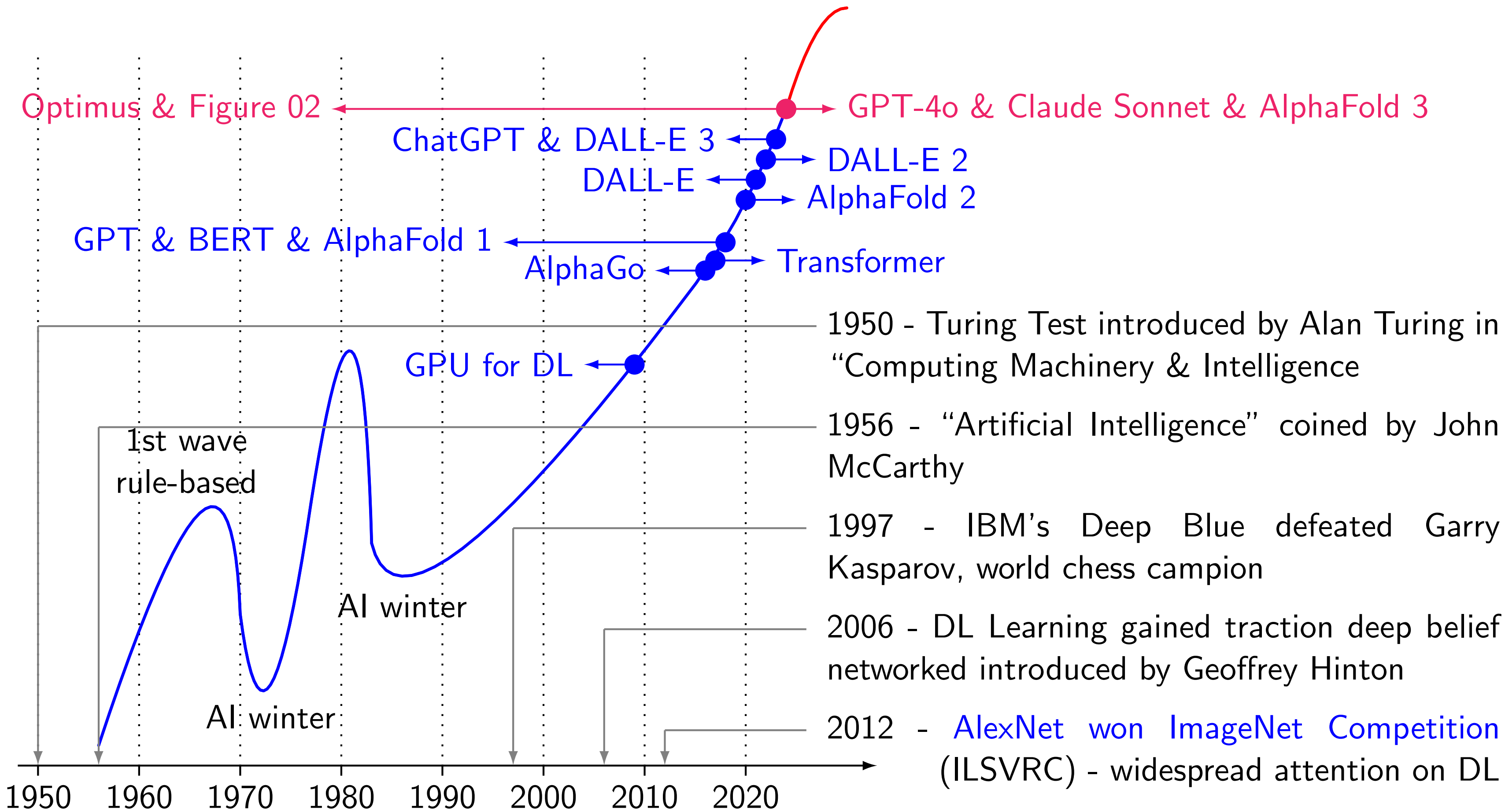
Definition and History

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

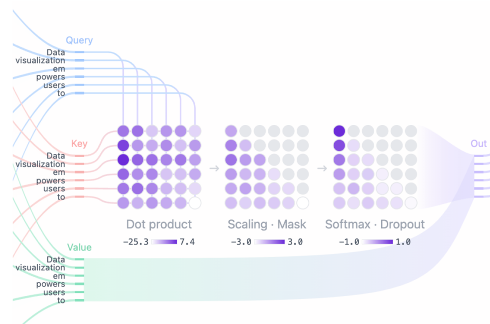
Deep learning revolution

- 2012 – 2015 - DL 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

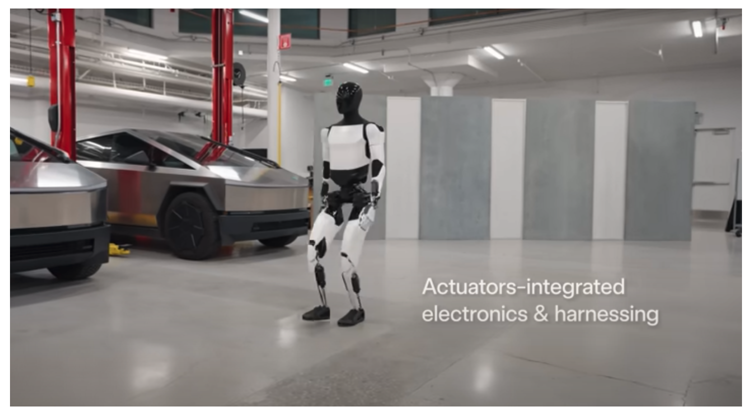
- 2017 – 2018 - Transformers & NLP breakthroughs[2]
  - *Transformer ($e.g.$, BERT & GPT) revolutionized NLP*
  - major advancements in, $e.g.$, machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting 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

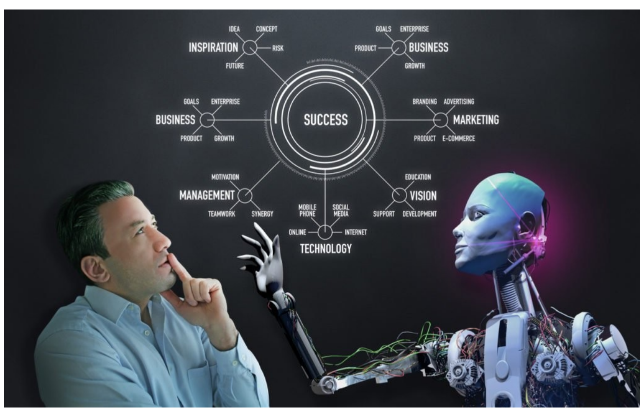
Lots of breakthroughs within 6 months in 2024

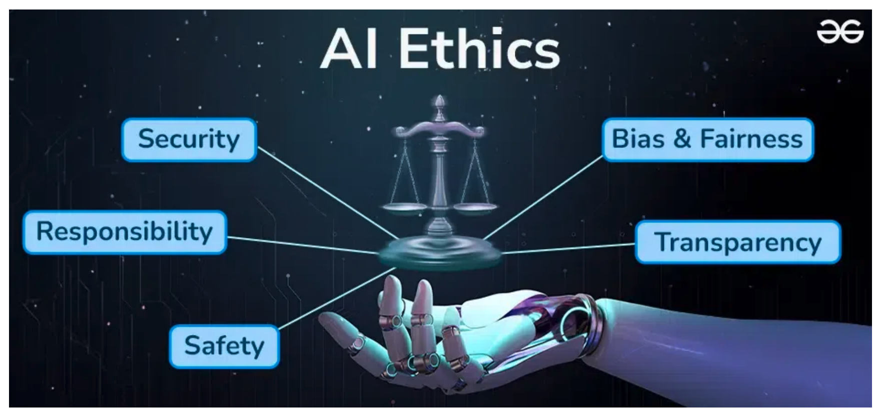
- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Llama 3, Sora
  - *transforming industries* such as content creation, customer service, education, *etc.*
- breakthroughs in specialized AI applications
  - Figure 02, Optimus, AlphaFold 3
  - driving unprecedented advancements in automation, 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*

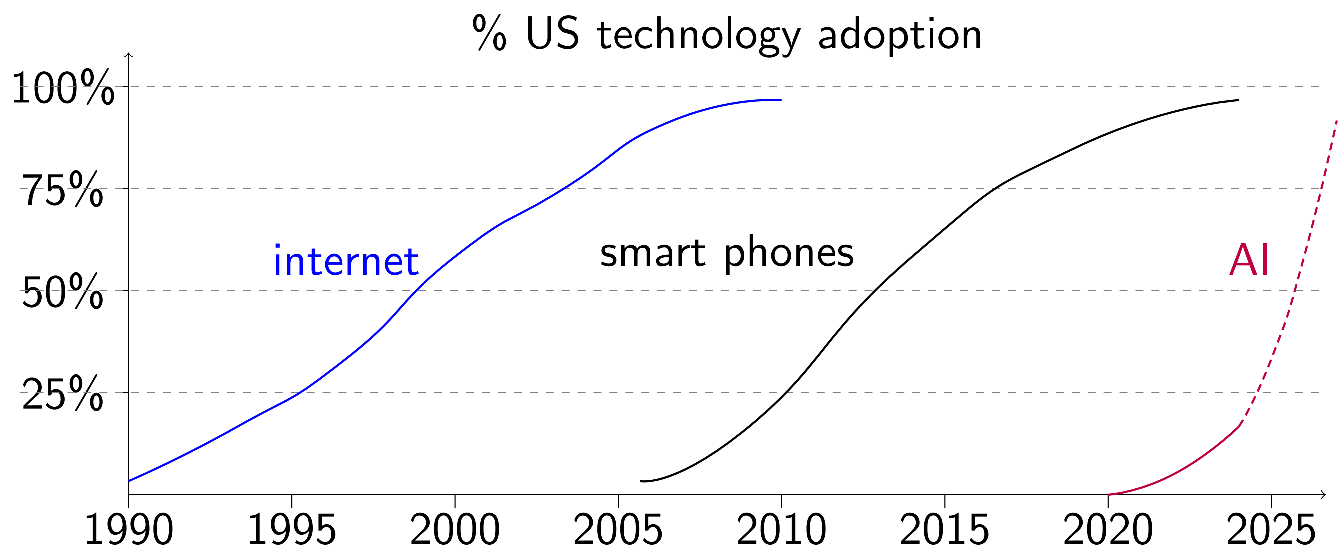




Recent Advances in AI

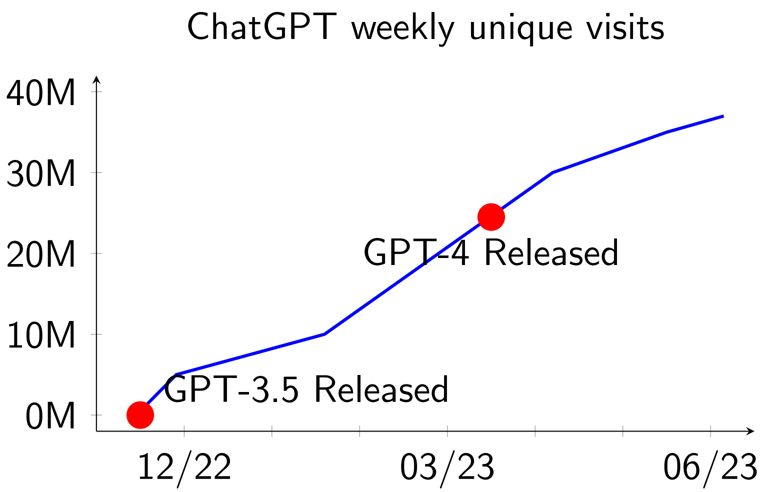
Where are we in AI today?

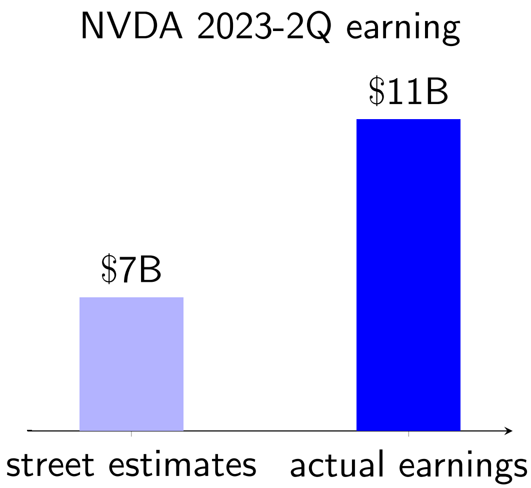
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



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]

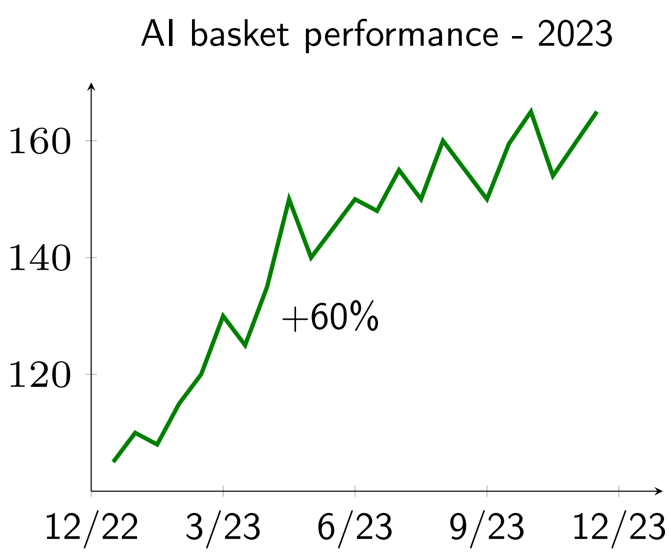


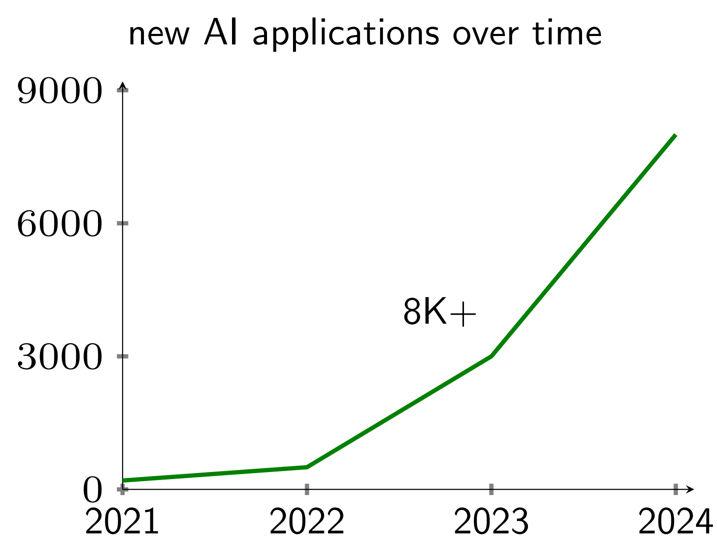


[3] source - Bloomberg

Explosion of AI ecosystems - AI stock market

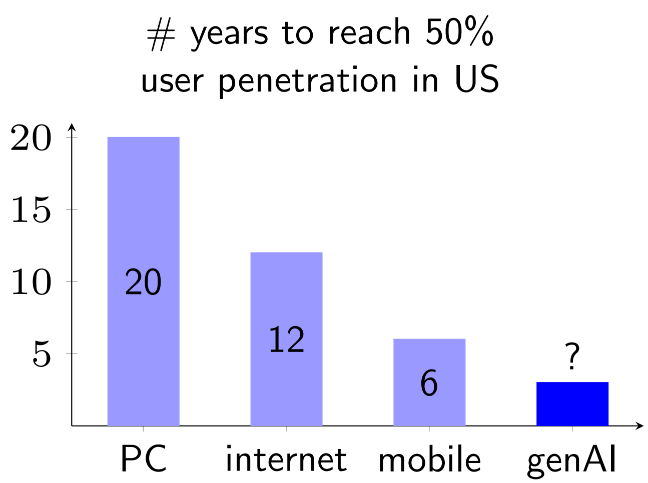
- *AI investment surge in 2023 - portfolio performance soars by 60%*
  - AI-focused stocks significantly outpaced traditional market indices
- *over 8,000 new AI applications* developed in last 3 years
  - applications span from healthcare and finance to manufacturing and entertainment

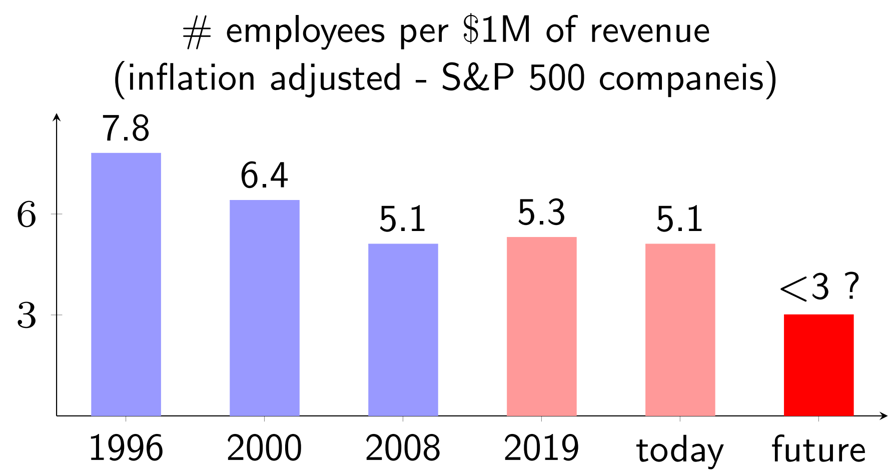




AI's transformative impact - adoption speed & economic potential

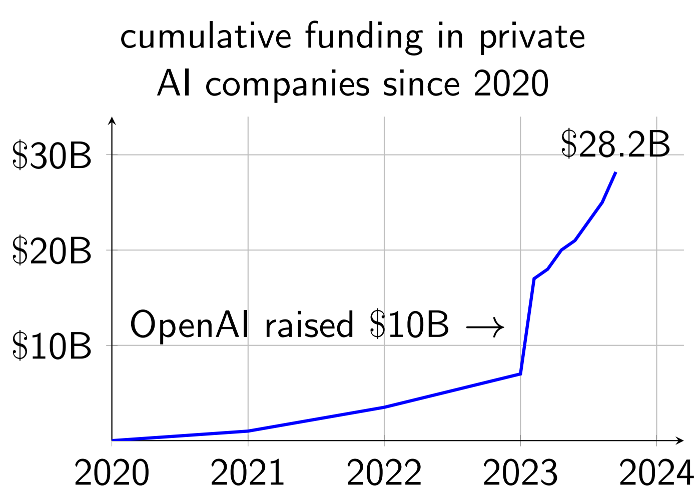
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI's potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater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

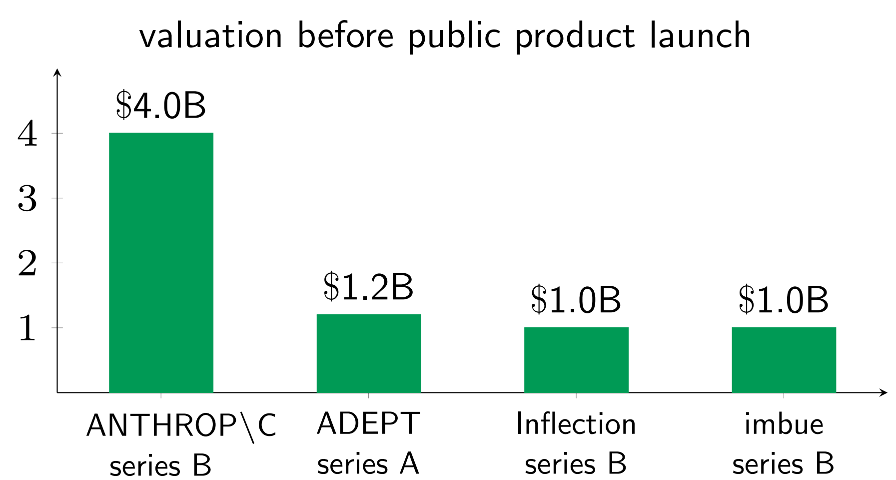




Massive investment in AI

- *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

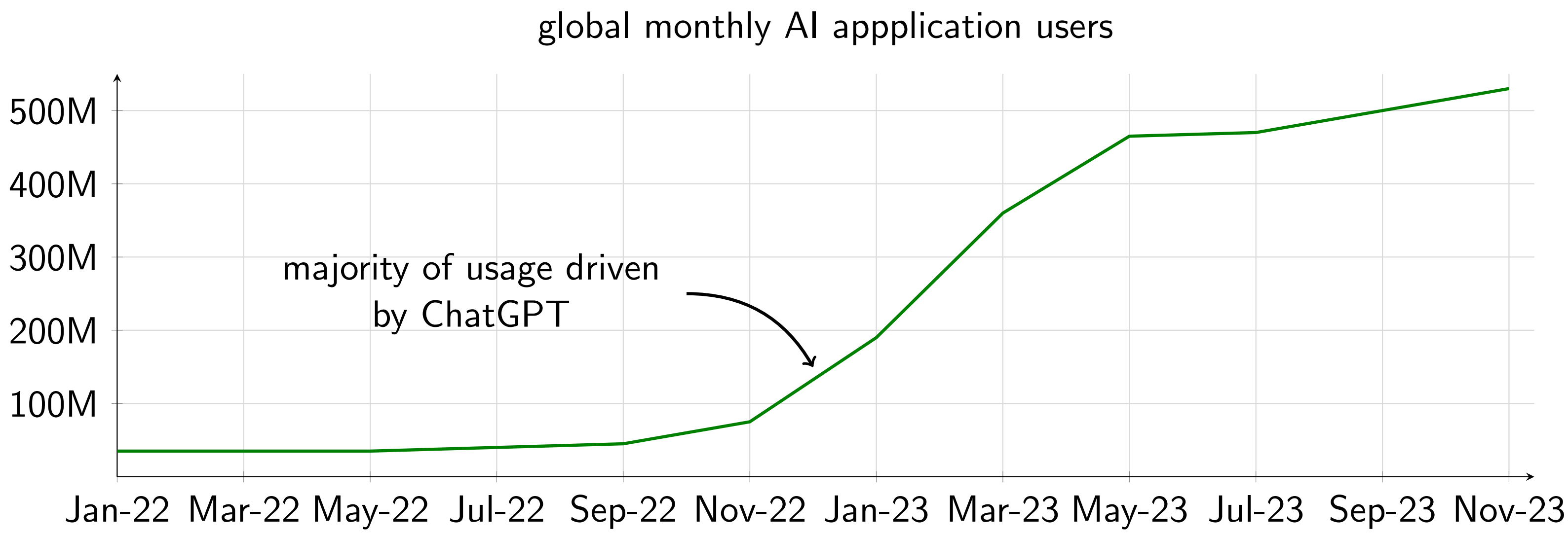
autonomy level	autonomous vehicles	genAI
L5	fullly autonomous	superhuman reasoning & perception
L4	highly autonomous	AI autopilot for complex tasks
L3	self-driving with light intervention	AI co-pilot for skilled labor
L2	Tesla autopilot	supporting humans with basic tasks
L1	cruise control	generating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

15 yrs

5 yrs

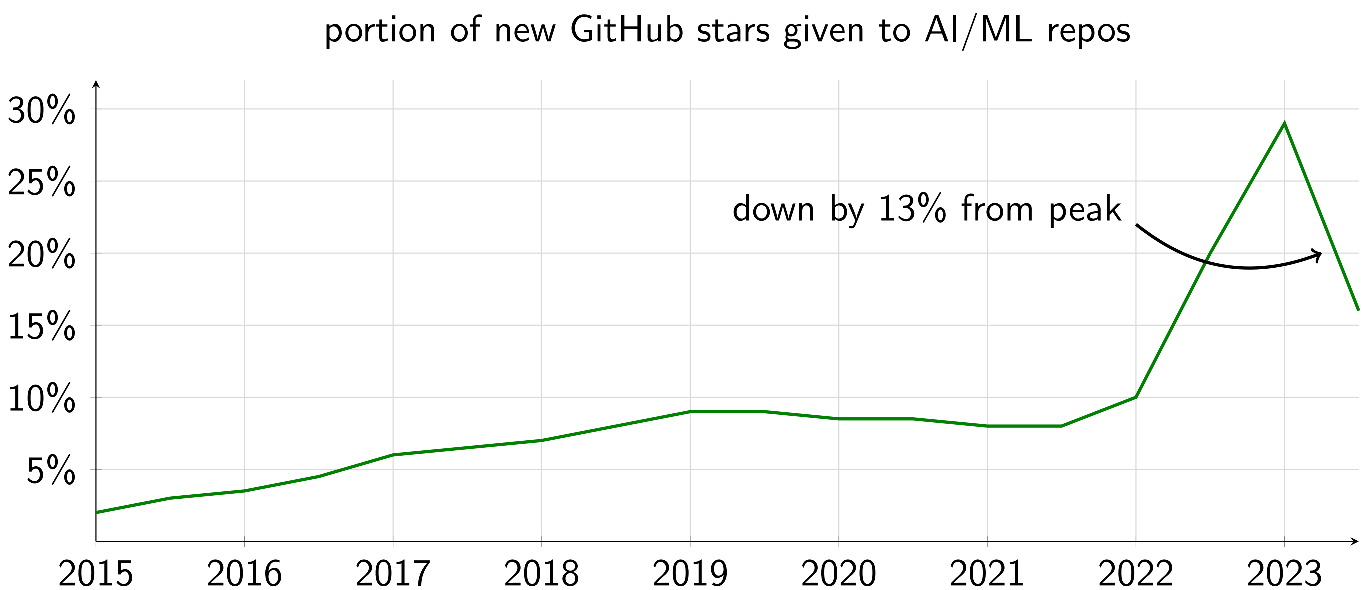
AI interest of users

- 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

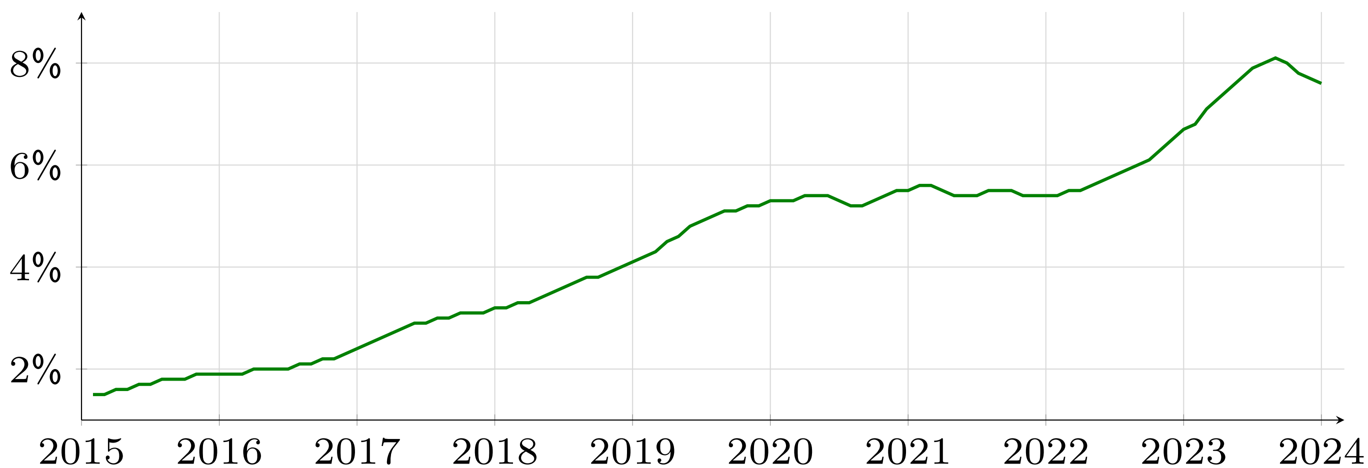
- rising popularity - portion of new GitHub stars given to AI/ML repositories steadily increased from 2015 to 2022
- excitement waning & washing out AI "tourists" - decline of 13% from peak in 2022
- could indicate potential factors such as market saturation, economic conditions, or shifts in developer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



Developers' contribution to software packages

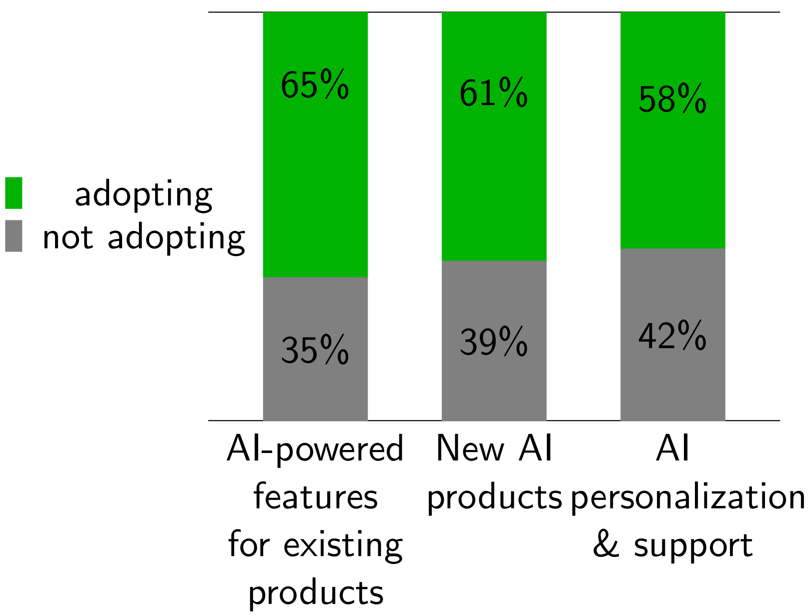
- steep acceleration from 2022 to 2024 correlates with explosion of LLMs & genAI
- suggesting transformative shift in AI landscape beyond gradual growth
- AI/ML still represents relatively small portion (less than 10%)
- indicating significant room for growth and mainstream adoption across various software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

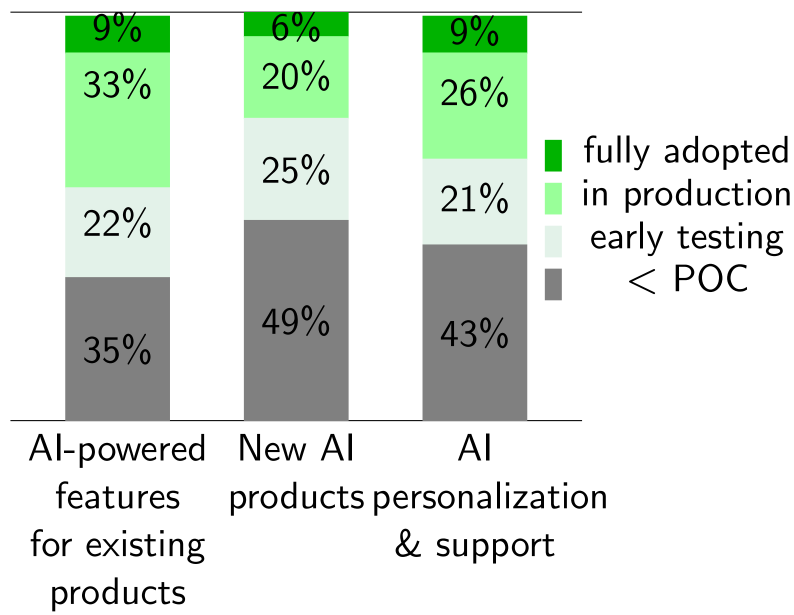
portion of AI/ML GitHub repo commits



Enterprises adoptiong AI

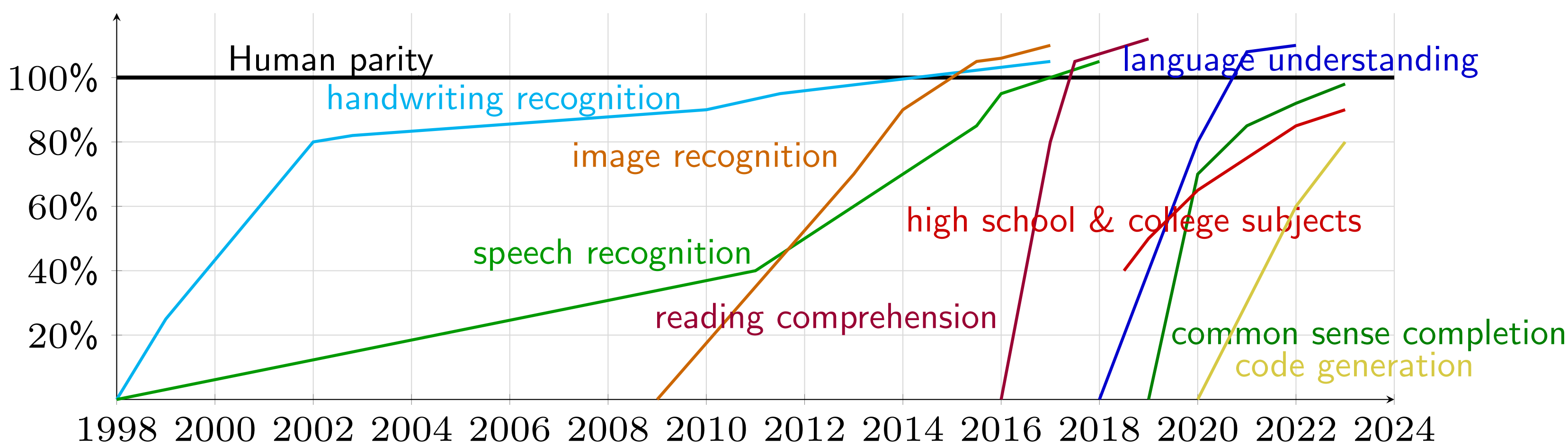
- more than 60% of enterprises planning to adopt AI
- full adoption rate is less than 10% - will take long time





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

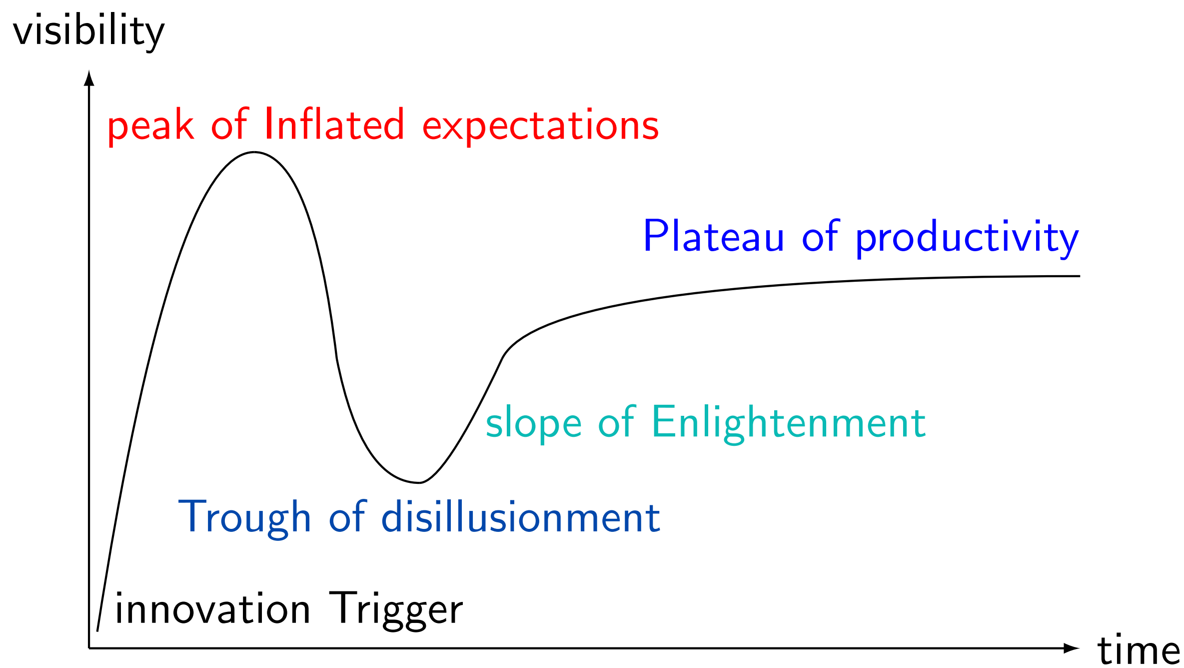


AI delivers game-changing values

- 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

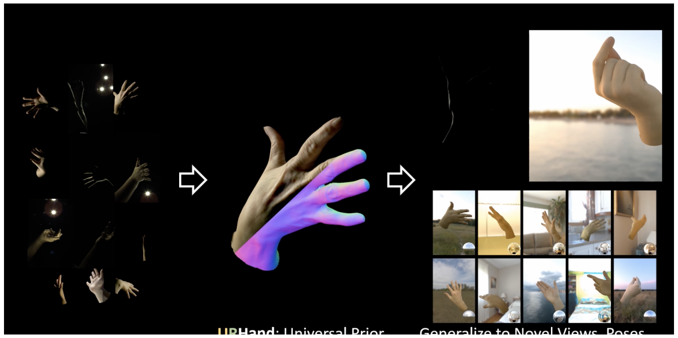
AI research race gets crazy

- 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



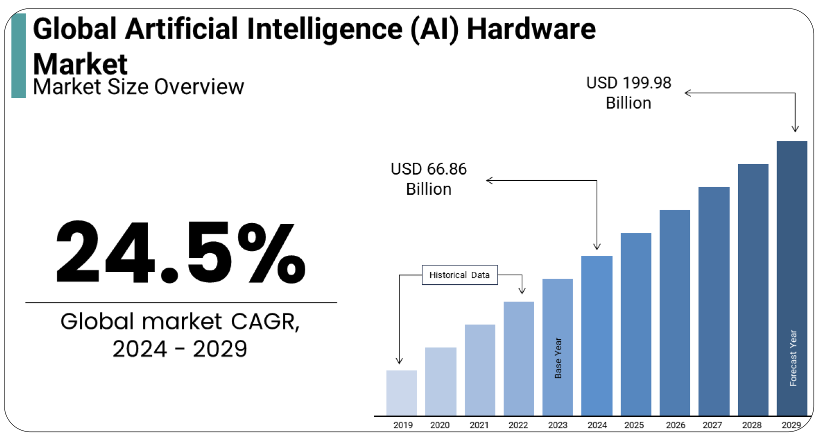
- CVPR 2024
  - PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar - MIT, Codec Avatars Lab, & Meta [KXS+24]
    - 3D reconstruction from single-view
  - Nymeria Dataset
    - large-scale multimodal egocentric dataset for full-body motion understanding
  - Relightable Gaussian 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

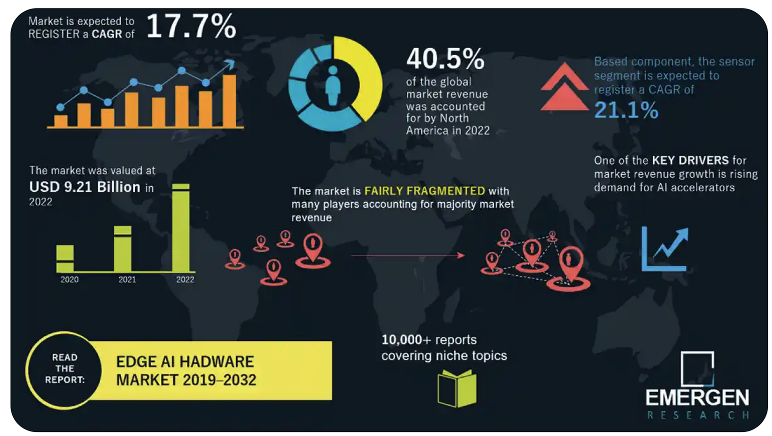
AI Hardware

AI Hardware Industry

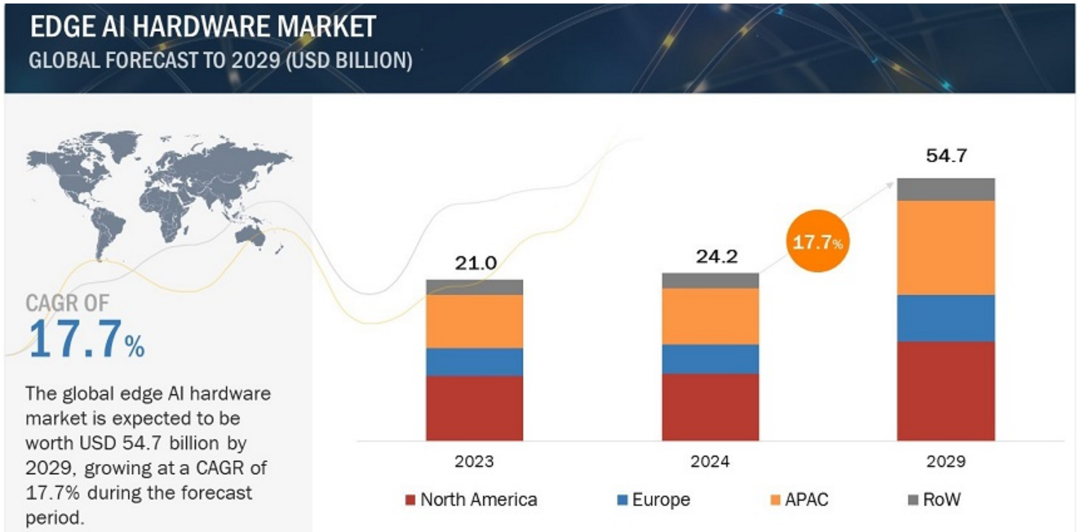
Landscape of AI hardware industry

- 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



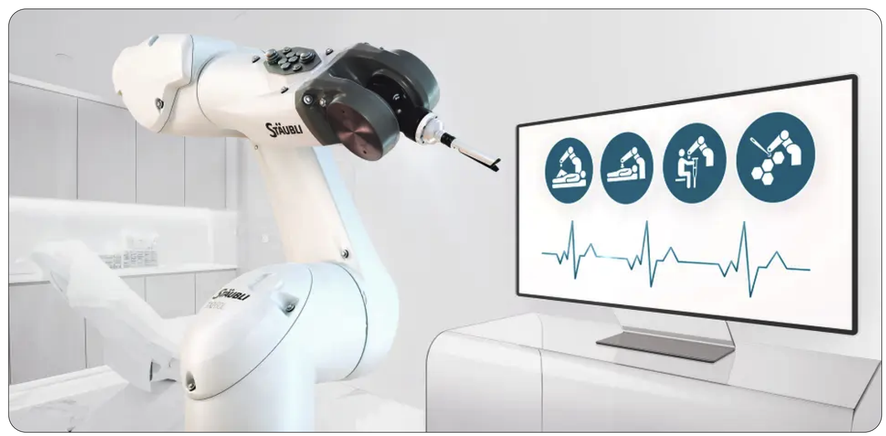


- North America leading market - high R&D investments & key industry players
- Asia Pacific rapidly expanding - strong semiconductor industries in South Korea, China & Japan
- demand for advanced processors such as GPUs, TPUs & AI accelerators rising due to complexity of AI algorithms & high computational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



Predictions for future of AI hardware market

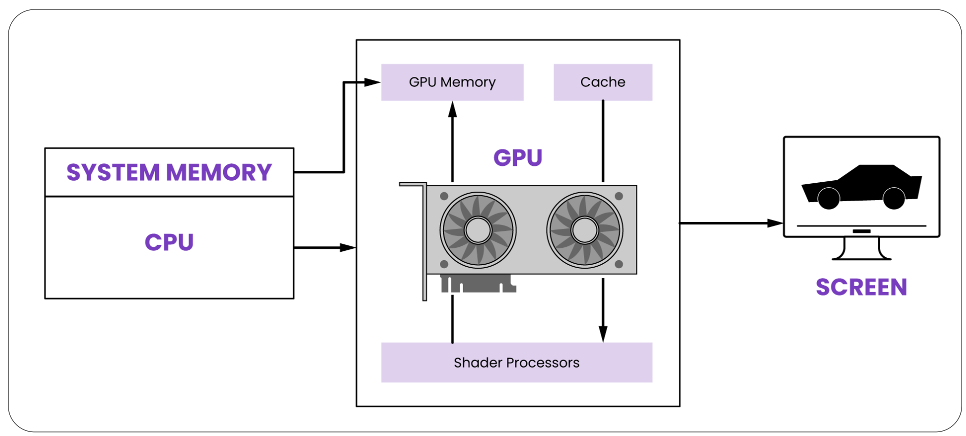
- AI hardware market expected to reach \$382B by 2032 - significant growth in data center AI chips
- integration of AI w/ 5G & increased use of AI in edge computing anticipated to drive future demand
- AI hardware becoming crucial in sectors such as autonomous vehicles, robotics & medical devices
- need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



GPUs and AI Accelerators

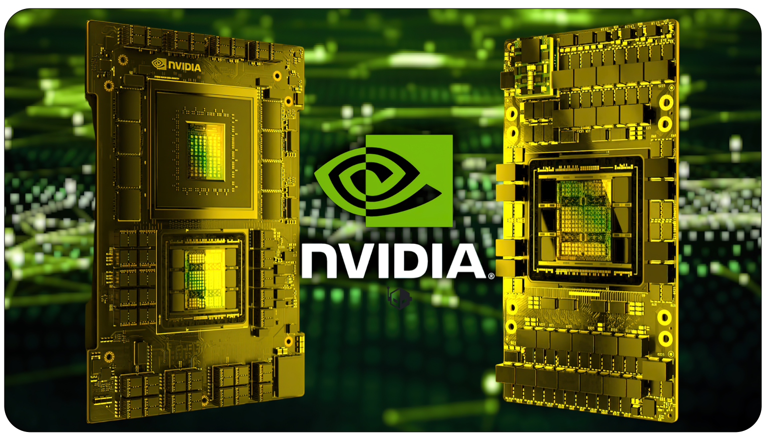
Technical challenges of GPUs & AI accelerators

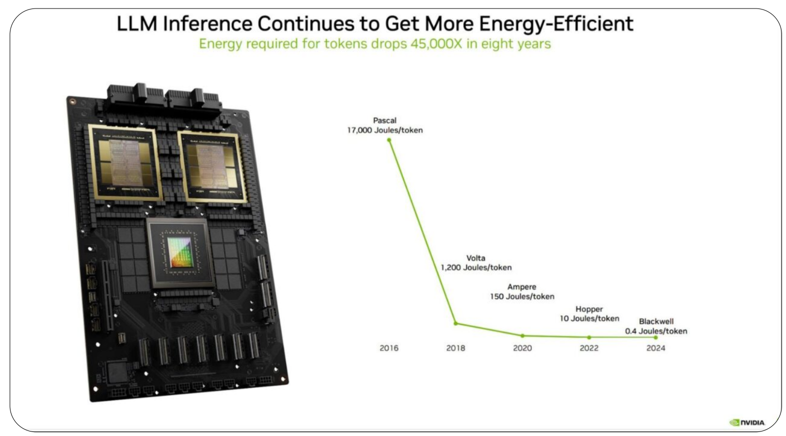
- facing challenges in scaling to handle increasingly large AI models and datasets - traditional architectures struggling w/ massive parallel processing demands of modern AI applications
- AI applications require extensive memory bandwidth often leading to bottlenecks - efficient memory management is 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

AI accelerator startups

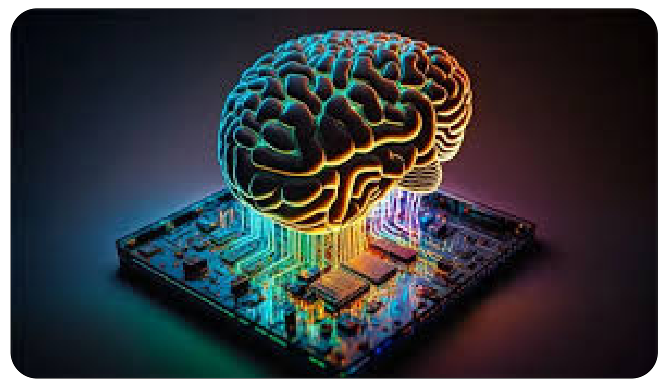
- innovative architectures - startups like Groq, SambaNova & Graphcore leading with *novel architectures designed to accelerate AI workloads*
  - *Groq* - tensor streaming processor (TSP) offering ultra-low latency & high throughput, high-performance AI inference chips enhancing speed & efficiency
  - *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

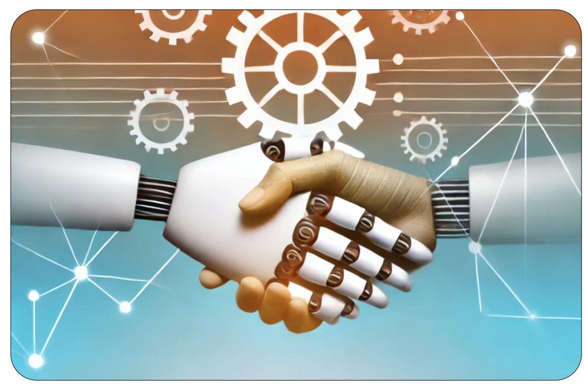
- energy efficiency
  - energy-efficient designs crucial for scalability in data centers and edge devices
  - startups developing solutions significantly reducing power consumption without compromising performance
- customization & flexibility
  - AI accelerators from startups often offer greater customization 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

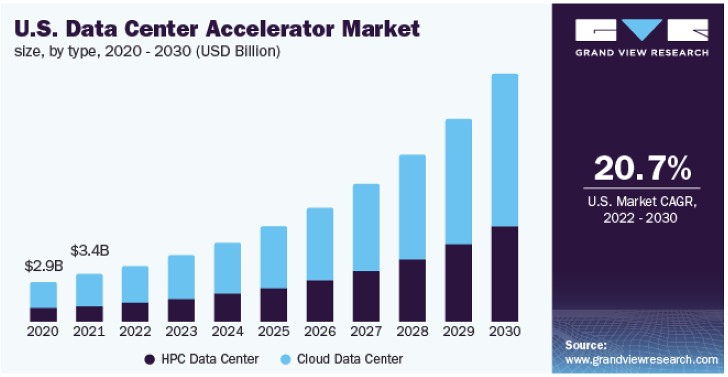
- disruption of traditional players
  - challenging dominance of established players like NVIDIA & Intel
  - unique architectures providing specialized solutions traditional GPUs and CPUs cannot efficiently handle
- driving down costs
  - offering competitive alternatives pushing down cost of AI computation
  - 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.



Global Semiconductor Industry

Hard-to-predict AI hardware markets

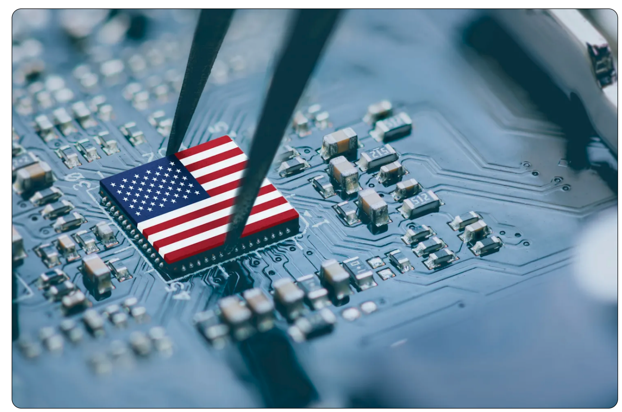
- US
  - birthplace for modern semiconductor chips driving PC market, internet, multi-media, mobile phones, and AI . . .
    - Intel, Texas Instrument (TI), Global Foundry
  - traditionally strong with design houses - NVIDIA, AMD, Broadcom, Apple, . . .
  - threatened experiencing global chip shortage & 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 & China
  - export controls on advanced chips and equipment
- CHIPS & Science Act (Aug, 2022)
  - \$52B in subsidies for domestic production, 25% investment tax credit for chip plants
  - (coerce) world-best semiconductor manufacturers build factories in US with support
    - GlobalFoundries - \$1.5B @ Feb-2024
    - Intel - \$8.5B @ Apr-2024 - Ohio - two fabs expandable to \$100B
    - Samsung - \$6.4B @ Apr-2024 - Talor, Texas
    - 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

Appendix

Recent AI Development

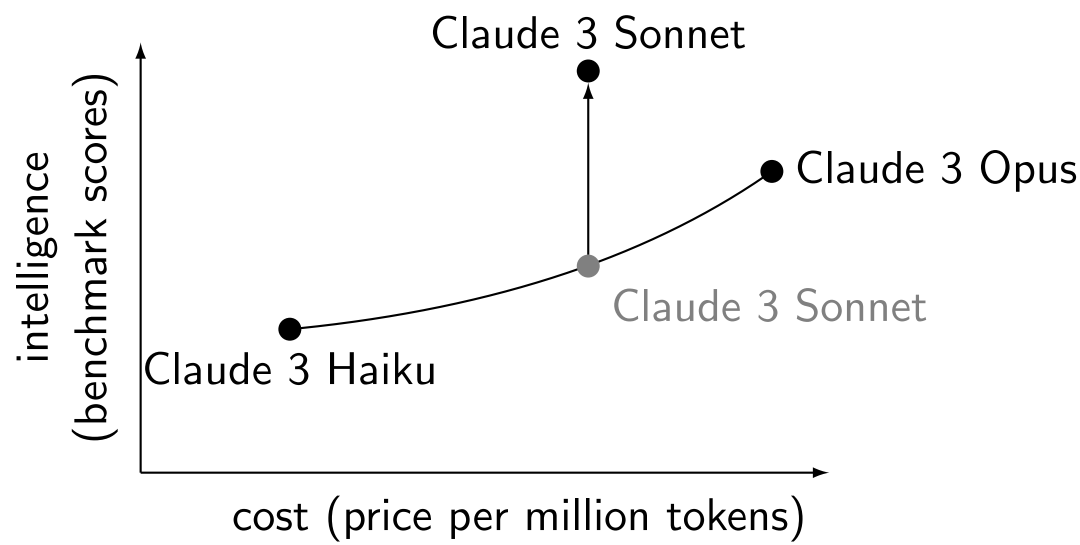
Notable recent AI research and new development

- 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*

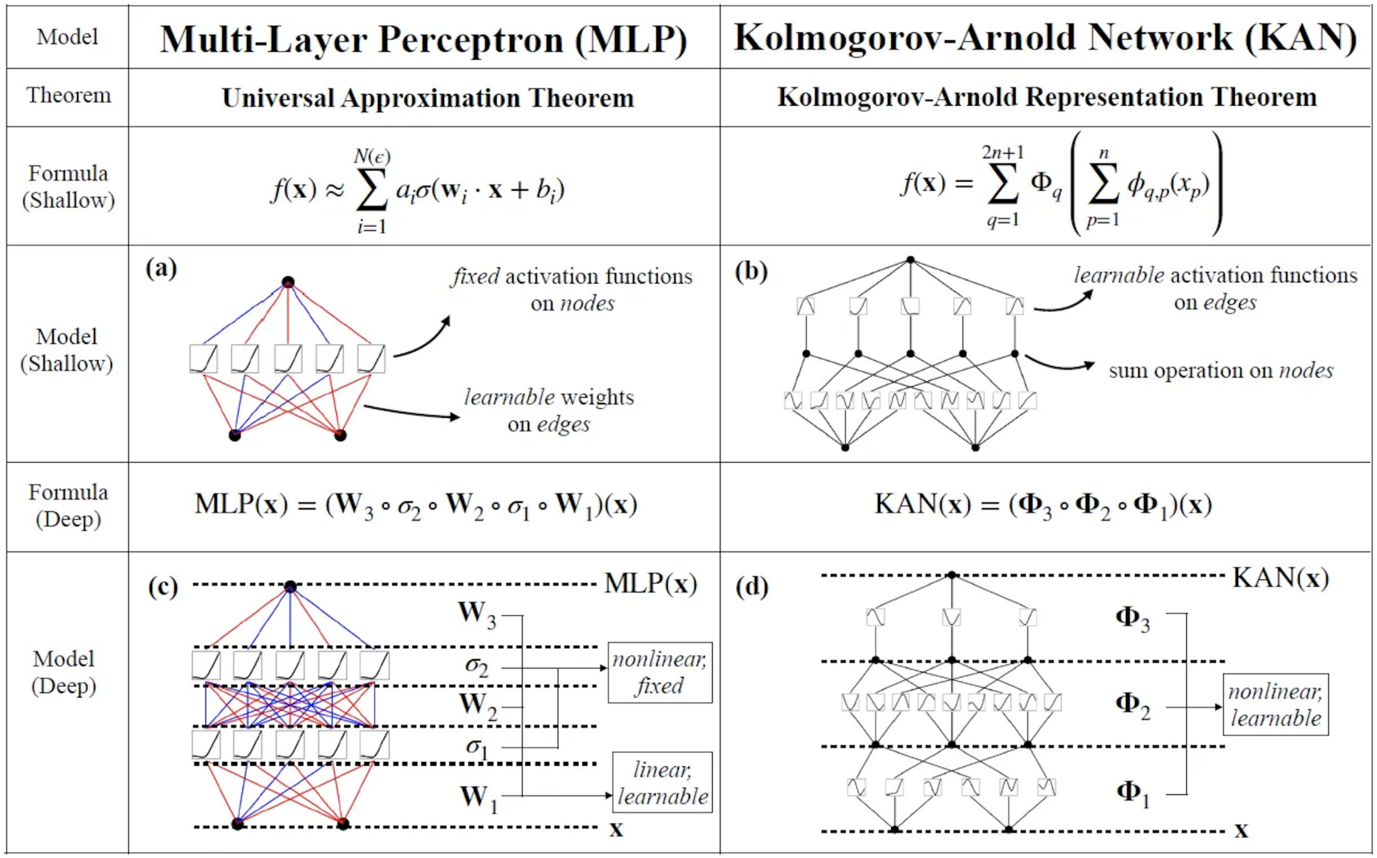
	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
visual math reasoning	67.7%	50.5%	63.8%	63.9%
science diagrams	94.7%	88.1%	94.2%	94.4%
visual question answering	68.3%	59.4%	69.1%	62.2%
chart Q&A	90.8%	80.8%	85.7%	87.2%
document visual Q&A	95.2%	89.3%	92.8%	93.1%

KAN

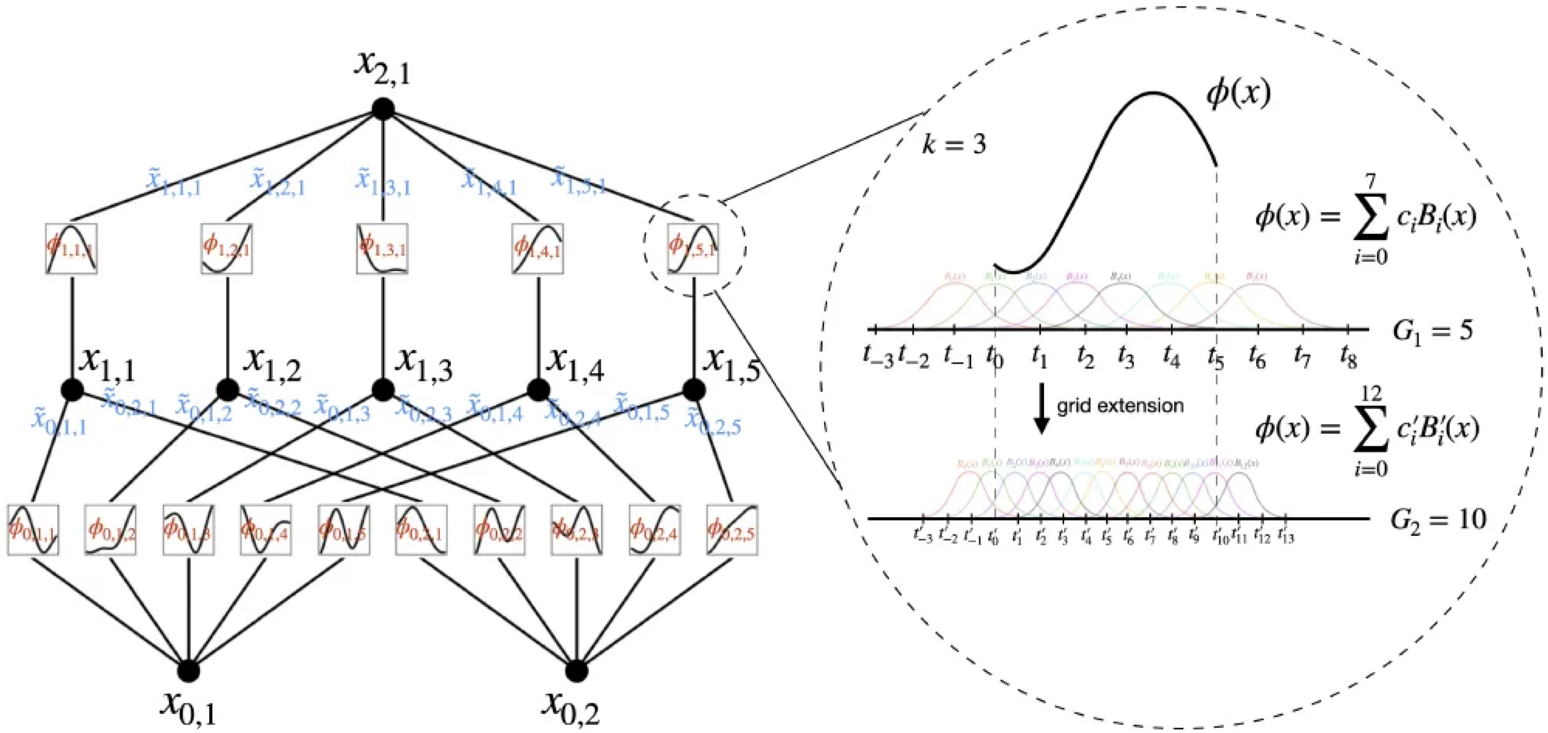
Kolmogorov–Arnold networks (KAN)

- 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

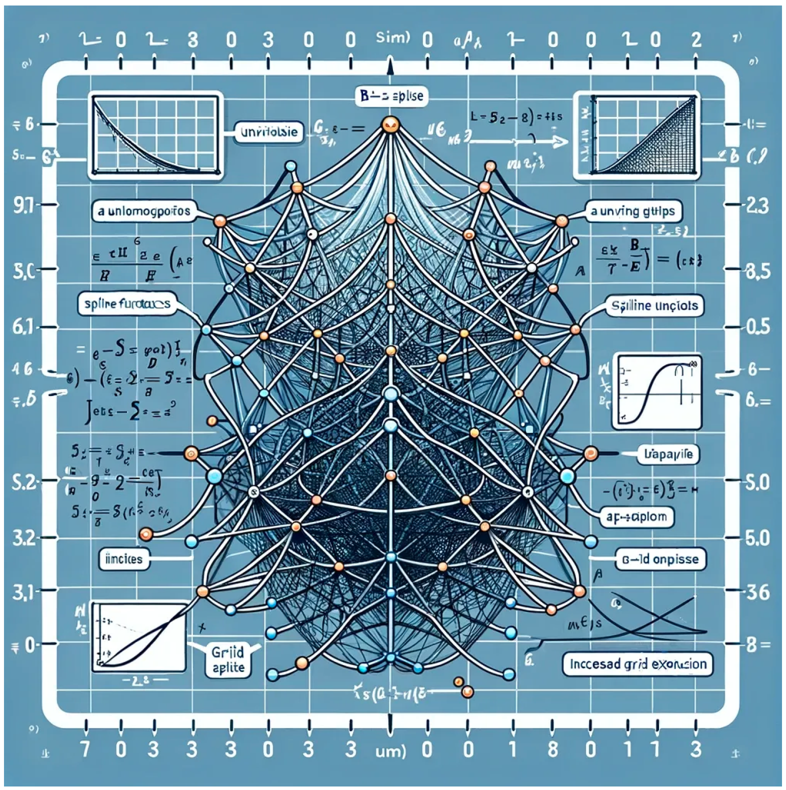
MLP vs KAN


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)
(a)
fixed activation functions on *nodes*
learnable weights on *edges*
(b)
learnable activation functions on *edges*
sum operation on *nodes*
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)
(c)
MLP(x)
$\mathbf{W}_3$
σ_2
$\mathbf{W}_2$
σ_1
$\mathbf{W}_1$
x
nonlinear, fixed
linear, learnable
(d)
KAN(x)
$\mathbf{\Phi}_3$
$\mathbf{\Phi}_2$
$\mathbf{\Phi}_1$
x
nonlinear, learnable

KAN architecture with spline parametrization unit layer



Future work on KAN



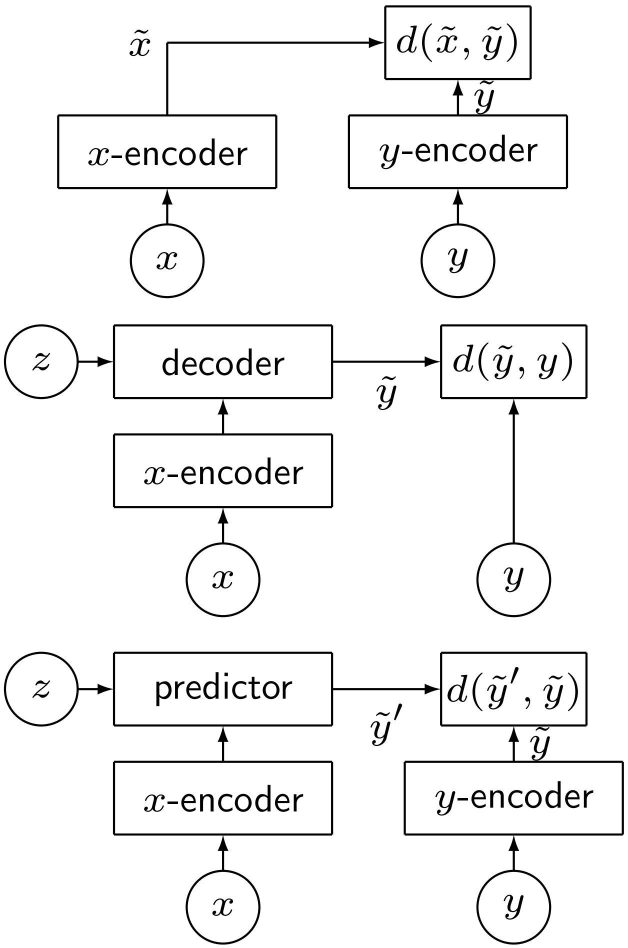
- natural question is
  - what if use both MLP and KAN?
  - what if use other types of splines?
  - how to control forgetfulness of continual learning?
  - why functions of one varible? possible to use functions of two variables?

(figure created by DALLE-3)

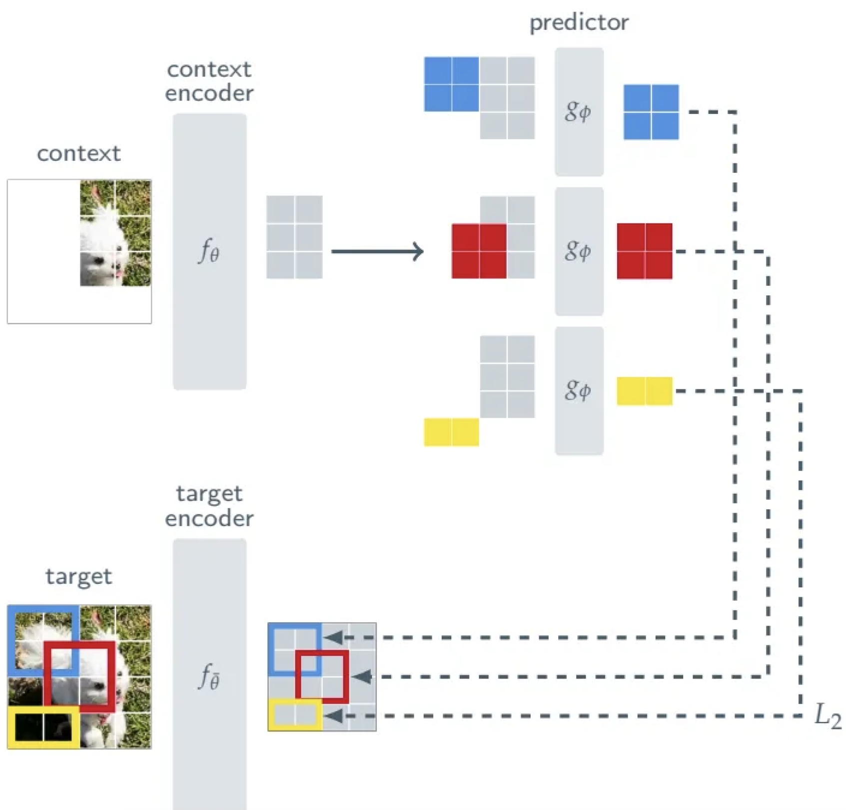
JEPA

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

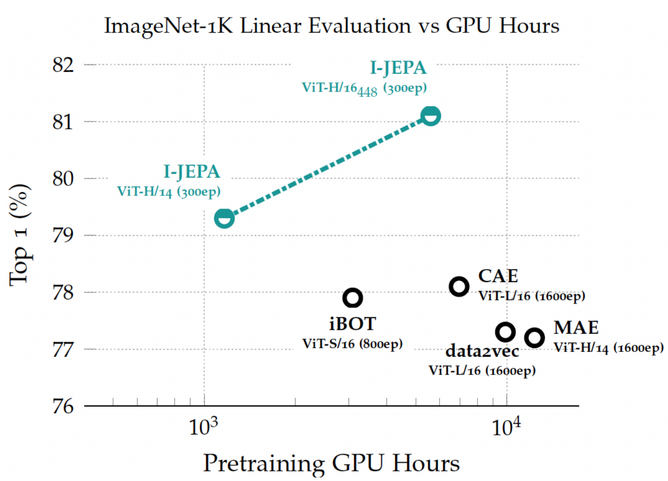


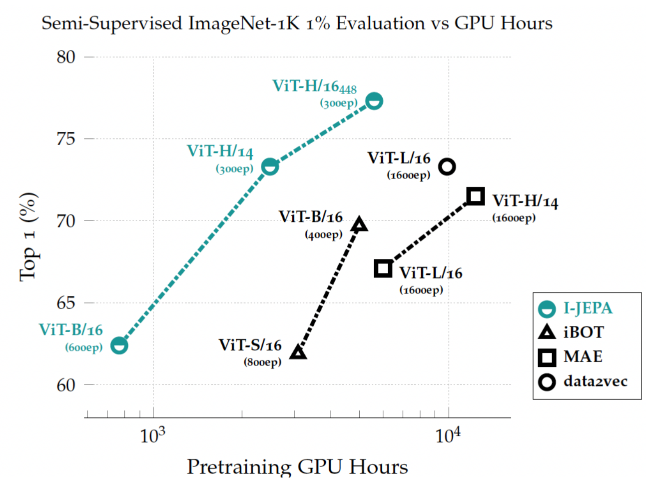
- I-JEPA
  - predicts missing information in *abstract representation space*
    - *e.g.*, given single context block (unmasked part of the image), predict representations of various target blocks (masked regions of same image) where target representations computed by learned target-encoder
  - *generates 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**

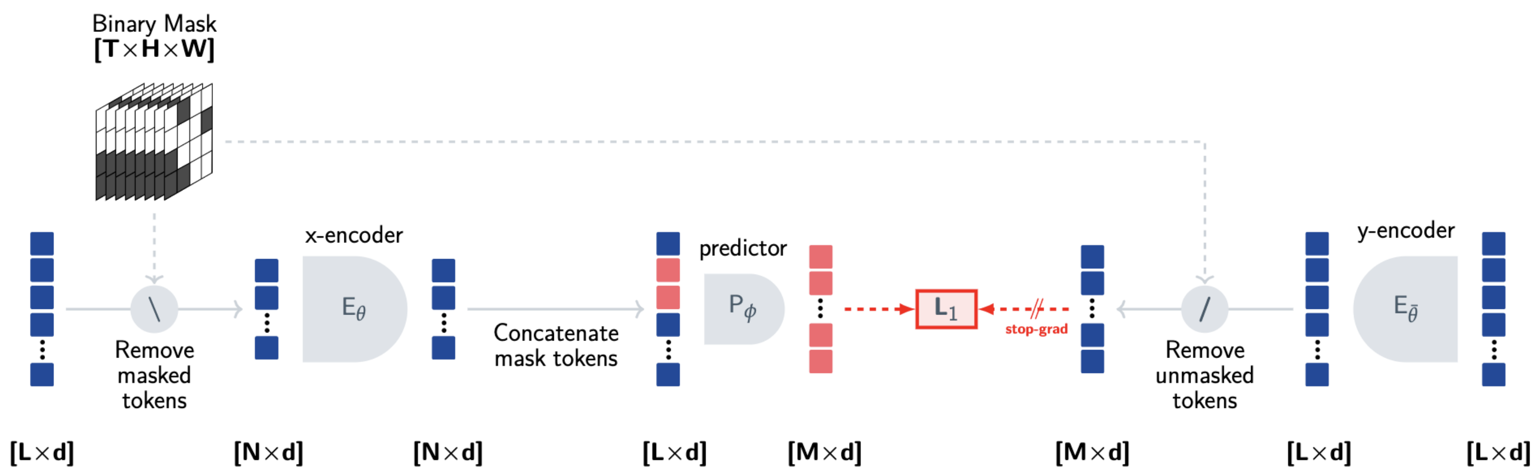
Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
Methods using extra data augmentations			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8





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

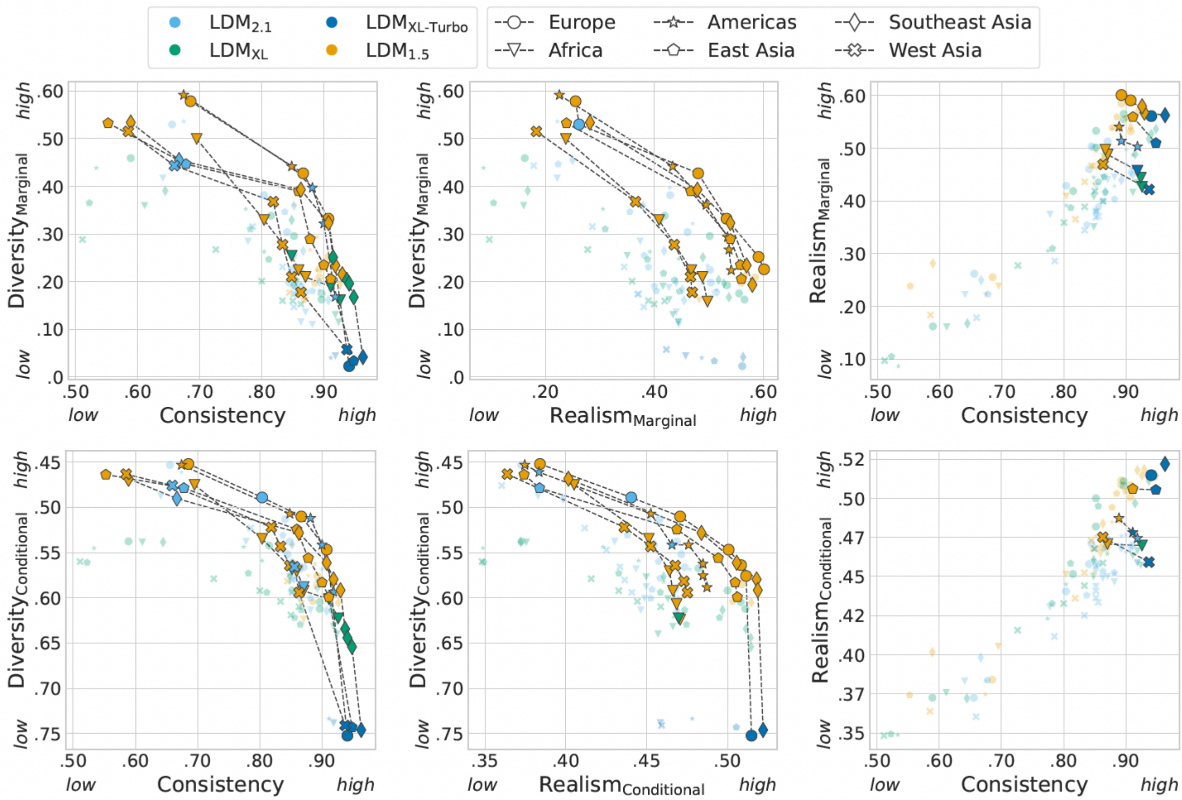
- Consistency-diversity-realism Pareto fronts of conditional image generative models - FAIR at Meta - Montreal, Paris & New York City labs, McGill University, Mila, Quebec AI institute, Canada CIFAR AI - Jun-2024

- 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



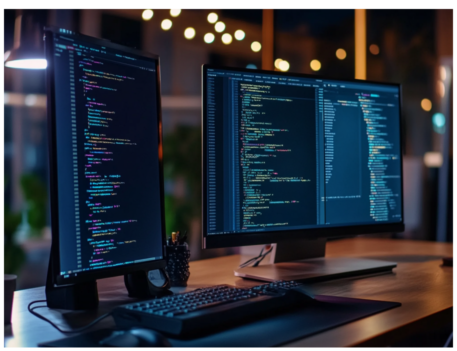
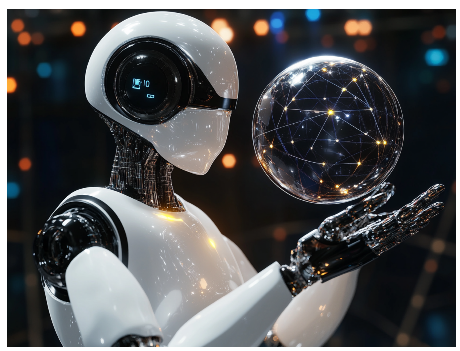
AI Products

AI product development - trend and characteristics

- *rapid pace* of innovation - new AI models & products being released at unprecedented rate, improvements coming in weeks or months (rather than years)
- *LLMs dominating* - models like GPT-4 & Claude pushing boundaries in NLP & genAI
- *multimodal AI* gaining traction - models processing & generating text, images & even video becoming more common, $e.g.$, Grok, GPT-4, Gemini w/ vision capabilities
- *open-source* AI movement - growing trend of open-source AI models and tools, challenging dominance of proprietary systems
- *AI integration in everyday products* - from smartphones to home appliances, AI being integrated into wide array of consumer products

AI product development - trend and characteristics

- *ethical AI & regulatory focus* - increased attention on ethical implications of AI & calls for regulation of AI development and deployment
- AI in enterprise - businesses across industries rapidly adopting AI for various applications
- *specialized AI models* - development of AI models tailored for specific industries or tasks, *e.g.*, healthcare, biotech, financial analysis
- AI-assisted *coding and development* - help software developers write code more efficiently & tools becoming increasingly sophisticated
- *concerns about AI safety & existential risk* - growing debate about potential short & long-term risks of advanced AI


LLM products

- OpenAI - ChatGPT 4o, GPT-4 Turbo Canvas
- Anthropic - Claude 3.5 Sonnet (with Artifacts), Claude 3 Opus, Claude 3 Haiku
- Mistral AI - Mistral 7B, Mistral Large 2, Mistral Small `xx.xx`, Mistral Nemo (12B)
- Google - Gemini (w/ 1.5 Flash), Gemini Advanced (w/ 1.5 Pro)
- X - Grok [mini] [w/ Fun Mode]
- Perplexity AI - Perplexity [Pro] - combines GPT-4, Claude 3.5, and Llama 3
- Liquid AI - Liquid-40B, Liquid-3B (running on small devices)

flying cats generated by Grok, ChatGPT 4o & Gemini

Comparison of LLMs & LLM products

model	developer	training data	# params	strength	weakness
GPT-4	OpenAI	web & books	170B	advanced reasoning & multimodal capabilities	high computational resources
LLaMA-2	Meta	public info & research articles	7∼70B	open access & good performance for different sizes	not powerful for complex tasks
Claude	Anthropic	mix of high-quality datasets	not disclosed	safety-first approach avoiding harmful responses	limited in publicly available details
PaLM 2	Google	multilingual text corpus	540B	high multilingual comprehension supporting various downstream apps	significant resources & not versatile in some contexts

Comparison of LLMs & LLM products

model	developer	training data	# params	strength	weakness
BLOOM	BigScience Community	diverse multilingual corpus (46 languages)	176B	open & support multiple languages	resource-intensive & lower performance
Mistral[5]	Mistral AI	public web data	7∼13B	lower parameter count	limited scalability for specialized apps
Liquid Foundation Model (LFM)	Liquid AI	adaptive datasets	adaptive & dynamic parameters	modular & support more specialized fine-tuning for niche use-cases & adaptable in deployment	complexity in design and implementation

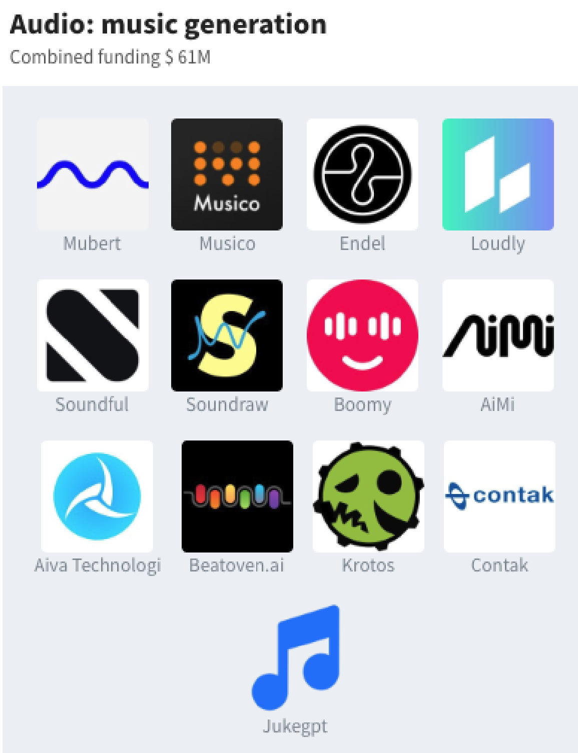
Multimodal genAI products

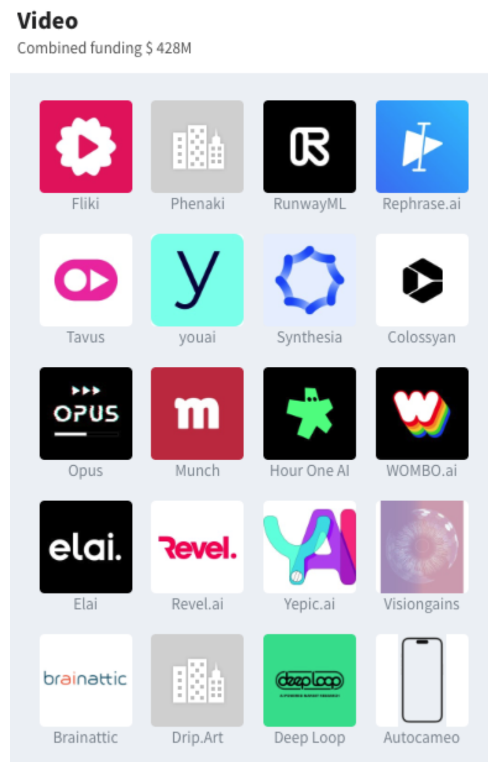
- DALL-E by OpenAI
  - *generate unique and detailed images based on textual descriptions*
  - understanding context and relationships between words
- Midjourney by Midjourney
  - let people *create imaginative artistic images*
  - can interactively guide the generative process, providing high-level directions

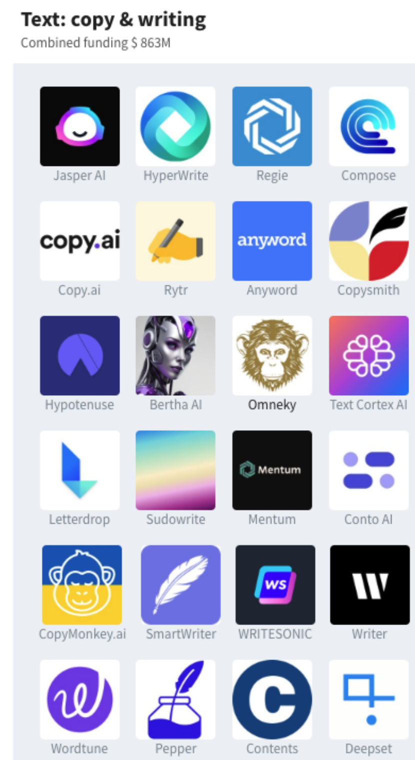
Multimodal genAI products

- Dream Studio by Stability AI
  - *analyze patterns in music data & generates novel compositions*
  - musicians can explore new ideas and enhance their *creative* processes
- Runway by Runway AI
  - *realistic images, manipulate photos, create 3D models & automate filmmaking*

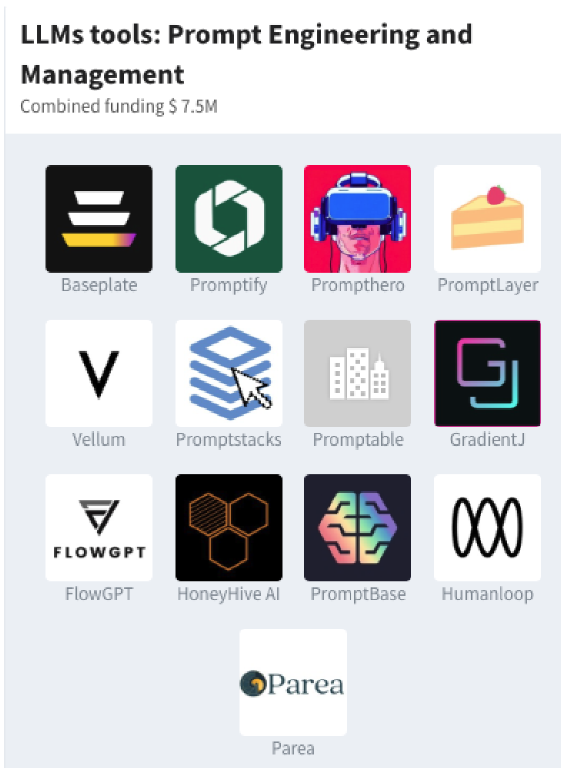
Diverse AI products

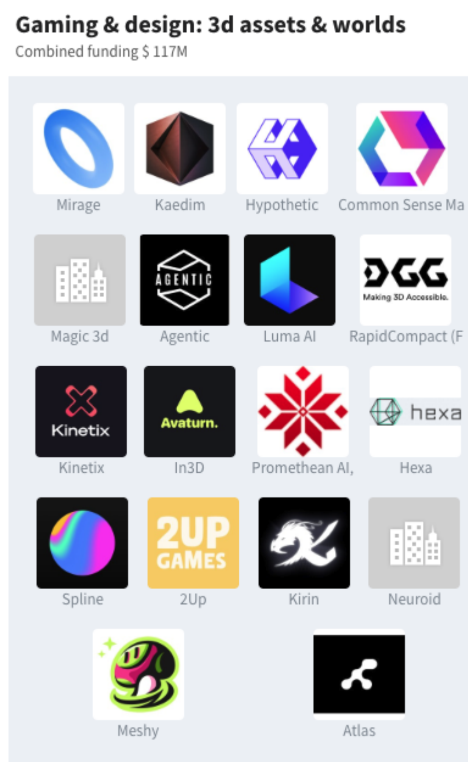


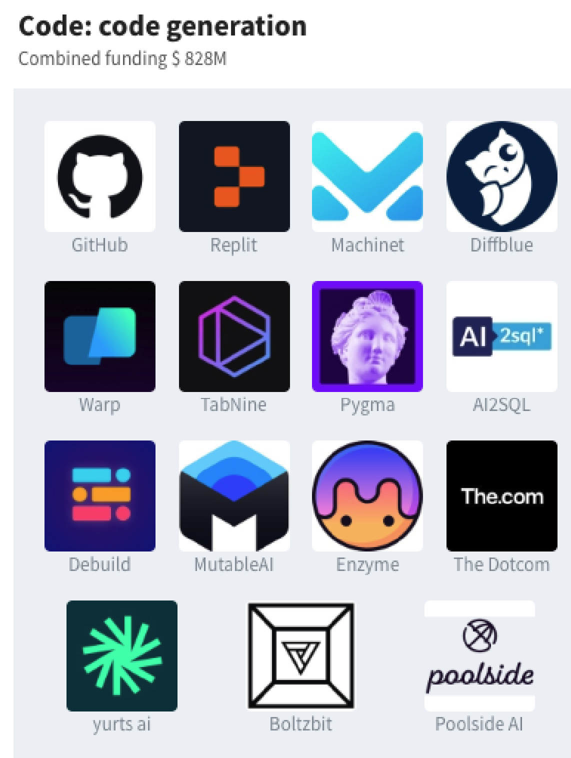




Diverse AI products







AI & 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.

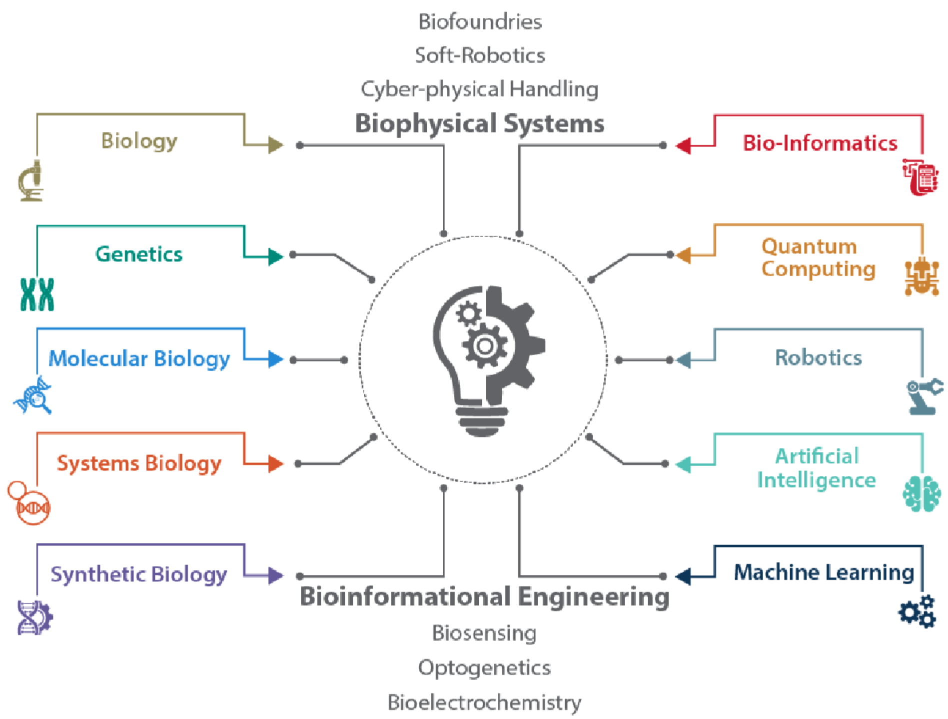
Biotech

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]

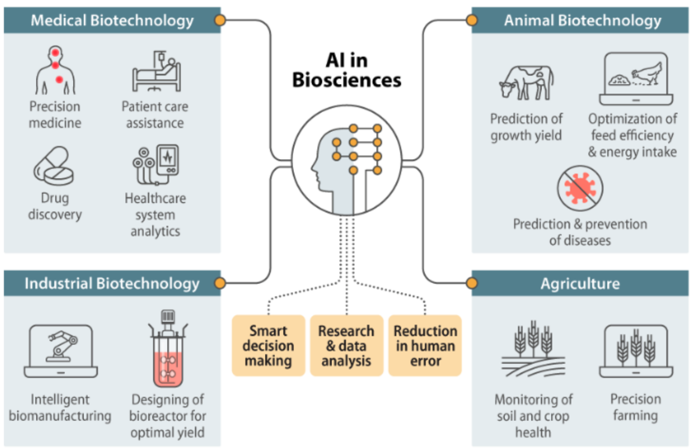


Convergence of AI and biological design

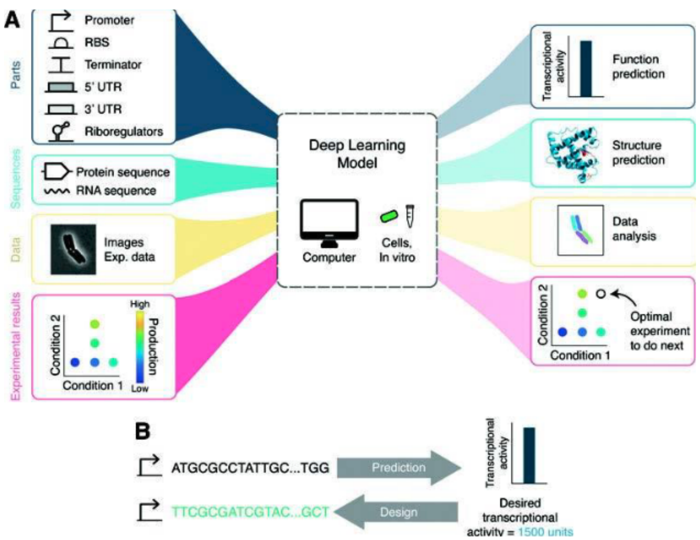
- both AI & biological sciences increasingly converging [BKP22]
  - each building upon the other's capabilities for new research and development across multiple areas
- Demo Hassabis, CEO & cofounder of DeepMind, said of biology [Toe23]

  > "... 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

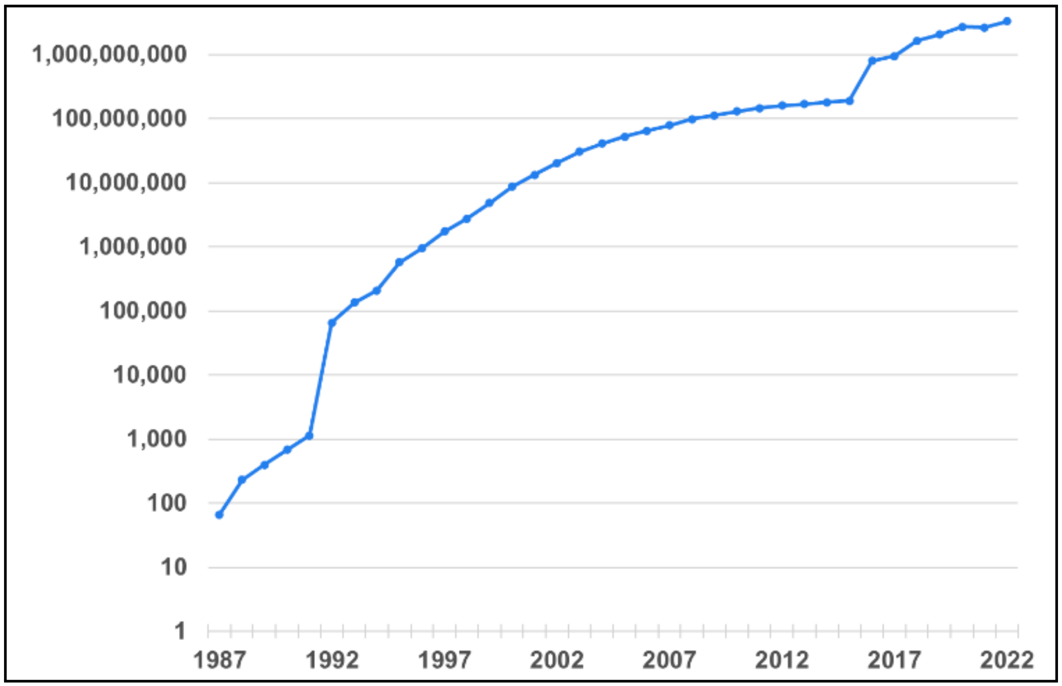
Rapid growth of biological data

- 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

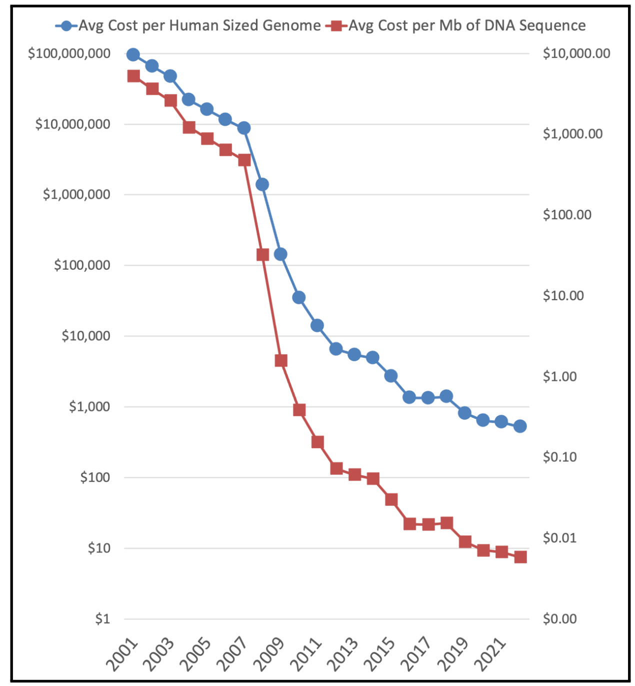
Volume and sequencing cost of DNA over time

- 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



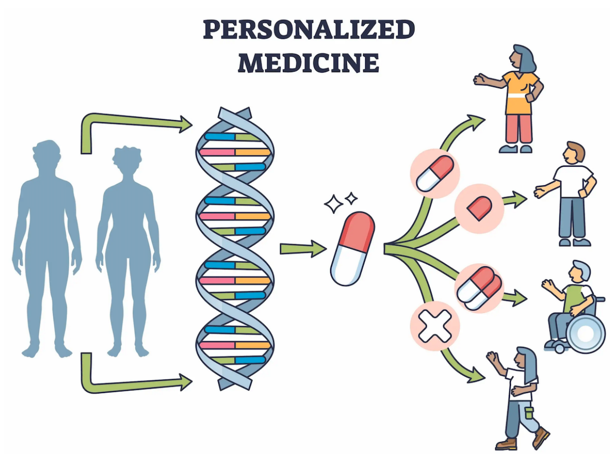
Bio data availability and bias

- 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

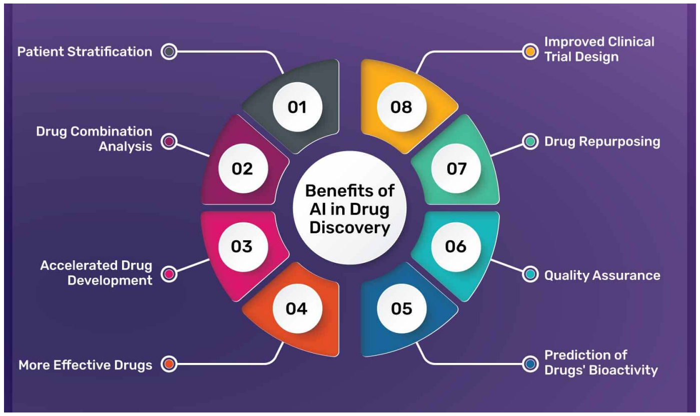
Emerging Trends in Biotech

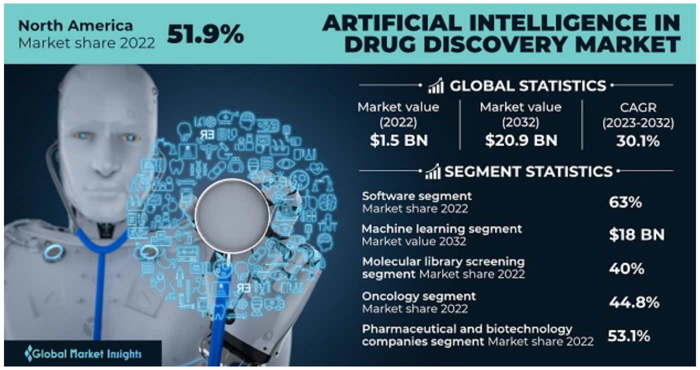
Personalized medicine

- *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

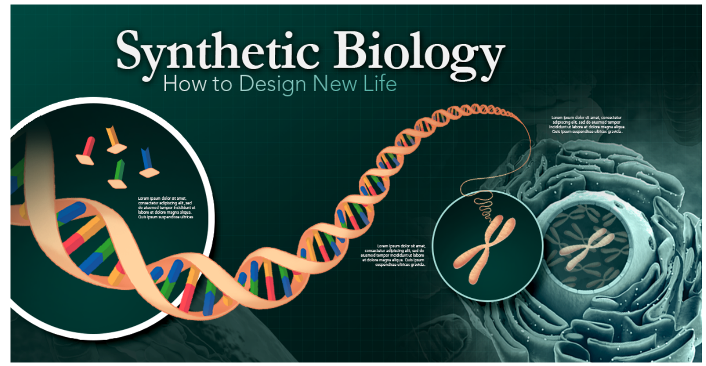


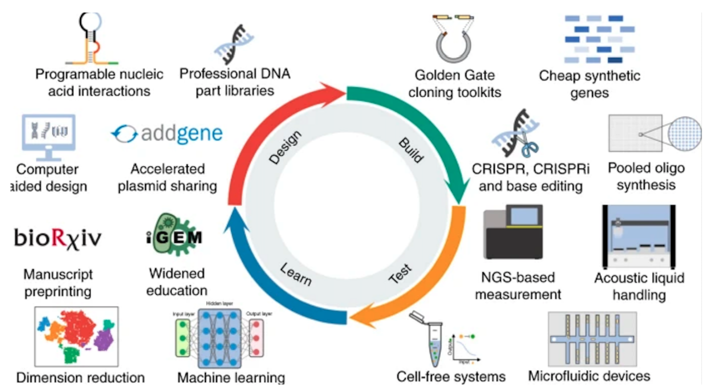


- traditional drug discovery process - time-consuming and costly often taking decades and billions of dollars
- AI streamlines this process by predicting the efficacy and safety of potential compounds with more speed and accuracy
- AI models analyze chemical databases to identify new drug candidates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

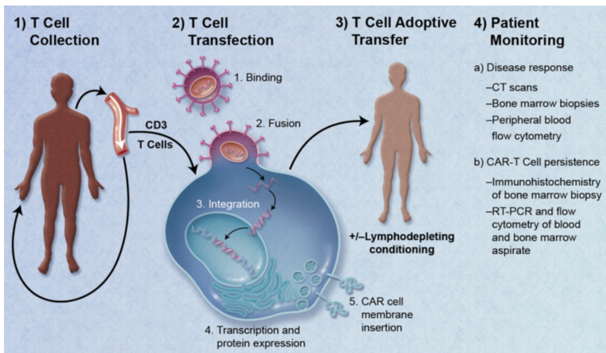
Synthetic biology

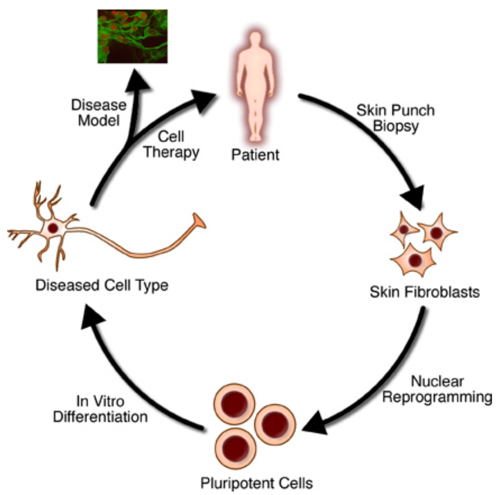
- use AI for gene editing, biomaterial production and synthetic pathways
- combine principles of biology and engineering to design and construct new biological entities
- AI optimizes synthetic biology processes from designing genetic circuits to scaling up production
- company - Ginkgo Bioworks uses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





Regenerative medicine

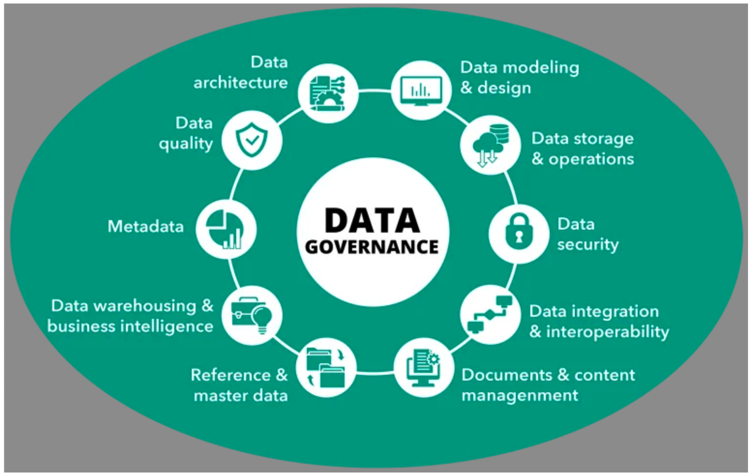




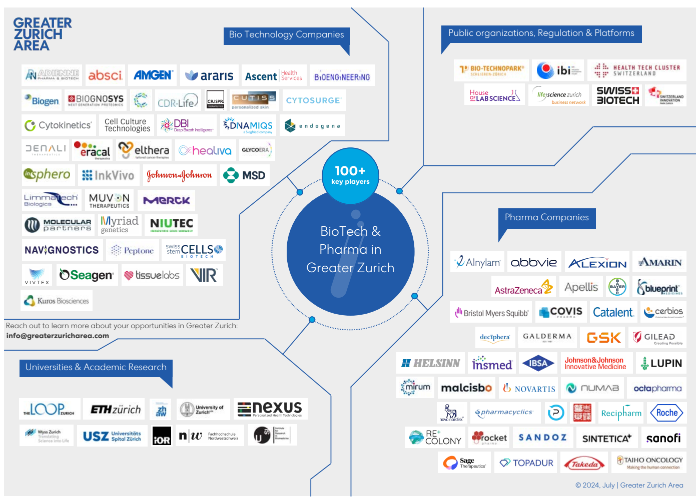
- AI advances development of stem cell therapies & tissue engineering
- AI algorithms assist in identifying optimal cell types, predicting cell behavior & personalized treatments
- particularly for conditions such as neurodegenerative diseases, heart failure and orthopedic injuries
- company - Organovo 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

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.

[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.

[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.

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

[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.

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

[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.

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

[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.

Thank You