[SNU ECS Four-Departments Joint Seminar] AI Pulse Check - What Silicon Valley Is Actually Building Right Now

Sunghee Yun

Co-Founder & CTO @ Erudio Bio, Inc.
Co-Founder & CEO @ Erudio Bio Korea, Inc.
Adjunct Professor & Advisory Professor @ Sogang Univ. & DGIST
Global Leadership Initiative Fellow @ Salzburg Global Seminar

About Speaker

- *Co-Founder & CTO @ Erudio Bio, Inc., San Jose & Novato, CA, USA*
- *Co-Founder & CEO @ Erudio Bio Korea, Inc., Korea*
- *KFAS-Salzburg Global Leadership Initiative Fellow @ Salzburg Global Seminar, Austria*
- *Adjunct Professor, Electronic Engineering Department @ Sogang University, Korea*
- *Advisory Professor, Electrical Engineering and Computer Science @ DGIST, Korea*
- *Global Advisory Board Member @ Innovative Future Brain-Inspired Intelligence System Semiconductor of Sogang University, Korea*
- Technology Consultant @ Gerson Lehrman Gruop (GLG), NY, USA
- Advisor & Evangelist @ CryptoLab, Inc., San Jose, CA, USA
- Chief Business Development Officer @ WeStory.ai, Cupertino, CA, USA
- *Co-Founder & CTO / Head of Global R&D & Chief Applied Scientist / Senior Fellow @ Gauss Labs, Inc., Palo Alto, CA, USA* *2020 ∼ 2023*

- Senior Applied Scientist @ Amazon.com, Inc., Vancouver, BC, Canada $\sim$ 2020
- Principal Engineer @ Software R&D Center, DS Division, Samsung, Korea $\sim$ 2017
- Principal Engineer @ Strategic Marketing & Sales Team, Samsung, Korea $\sim$ 2016
- Principal Engineer @ DT Team, DRAM Development Lab, Samsung, Korea $\sim$ 2015
- Senior Engineer @ CAE Team, Samsung, Korea $\sim$ 2012
- PhD - Electrical Engineering @ Stanford University, CA, USA $\sim$ 2004
- Development Engineer @ Voyan, Santa Clara, CA, USA $\sim$ 2001
- MS - Electrical Engineering @ Stanford University, CA, USA $\sim$ 1999
- BS - Electrical & Computer Engineering @ Seoul National University 1994 $\sim$ 1998

Highlight of Career Journey

- BS in EE @ SNU, MS & PhD in EE @ Stanford University
  - *Convex Optimization - Theory, Algorithms & Software*
  - advised by *Prof. Stephen P. Boyd*
- Principal Engineer @ Samsung Semiconductor, Inc.
  - AI & Convex Optimization
  - collaboration with *DRAM/NAND Design/Manufacturing/Test Teams*
- Senior Applied Scientist @ Amazon.com, Inc.
  - e-Commerce AIs - anomaly detection, deep RL, and recommender system
  - *Jeff Bezos's project - drove $200M in additional sales via Amazon Mobile Shopping App*
- *Co-Founder & CTO* / Global R&D Head & Chief Applied Scientist @ Gauss Labs, Inc.
- *Co-Founder & CTO* - AI Technology & Business Development @ Erudio Bio, Inc.
- *Co-Founder & CEO* - AI Technology & Business Development @ Erudio Bio Korea, Inc.

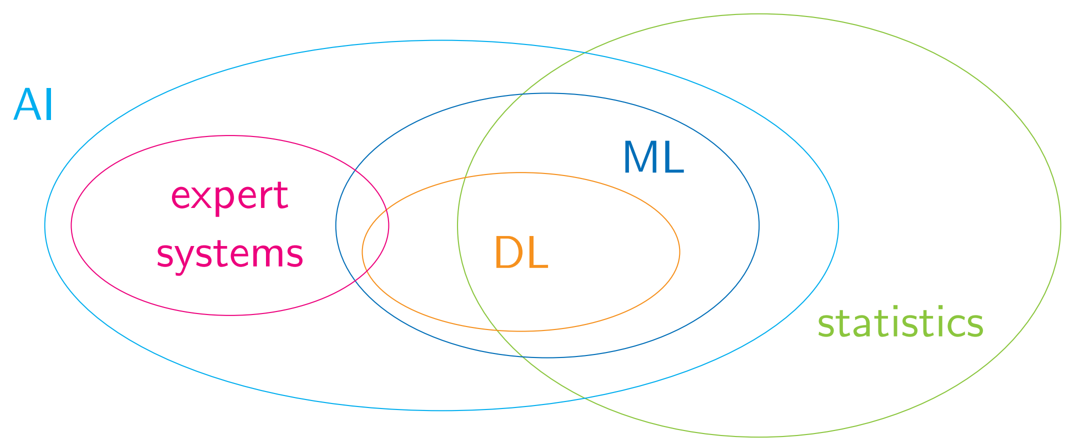
Today

Artificial Intelligence

Definition and History

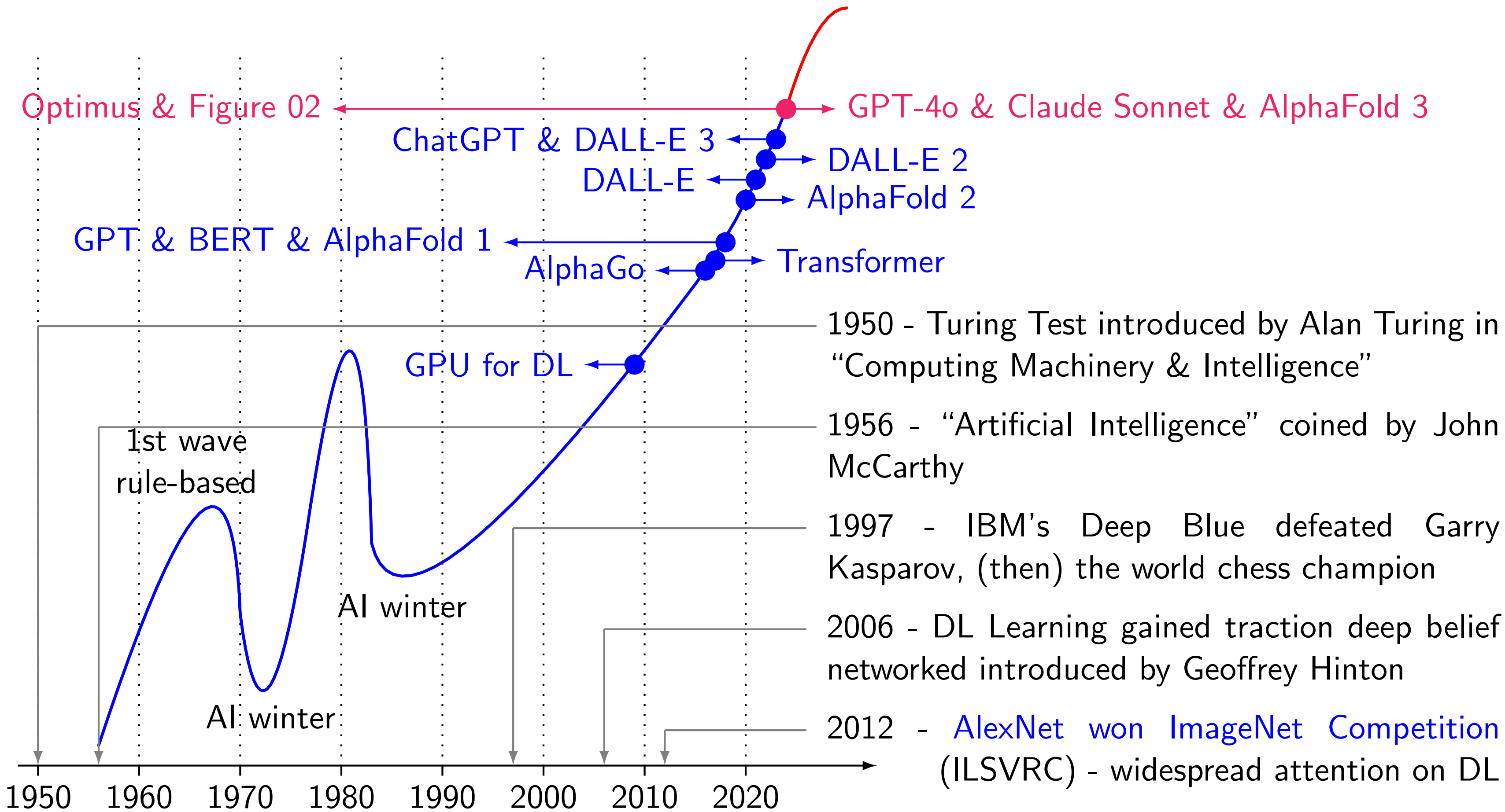
Definition & relation to other technologies

- AI
  - is technology doing tasks requiring human intelligence, such as learning, problem-solving, decision-making & language understanding
  - encompasses *range of technologies, methodologies, applications & products*
- AI, ML, DL, statistics & expert system[1] [HGH+22]



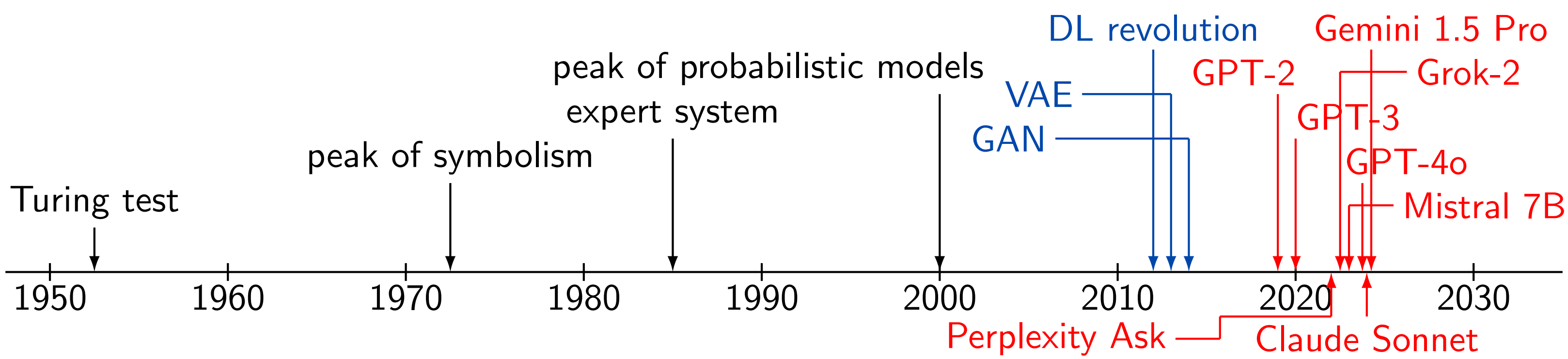
[1] ML: machine learning & DL: deep learning

History


Optimus & Figure 02
GPT-4o & Claude Sonnet & AlphaFold 3
ChatGPT & DALL-E 3
DALL-E 2
DALL-E
AlphaFold 2
GPT & BERT & AlphaFold 1
AlphaGo
Transformer
GPU for DL
1st wave
rule-based
AI winter
AI winter
1950 - Turing Test introduced by Alan Turing in "Computing Machinery & Intelligence"
1956 - "Artificial Intelligence" coined by John McCarthy
1997 - IBM's Deep Blue defeated Garry Kasparov, (then) the world chess champion
2006 - DL Learning gained traction deep belief networked introduced by Geoffrey Hinton
2012 - AlexNet won ImageNet Competition (ILSVRC) - widespread attention on DL
1950
1960
1970
1980
1990
2000
2010
2020

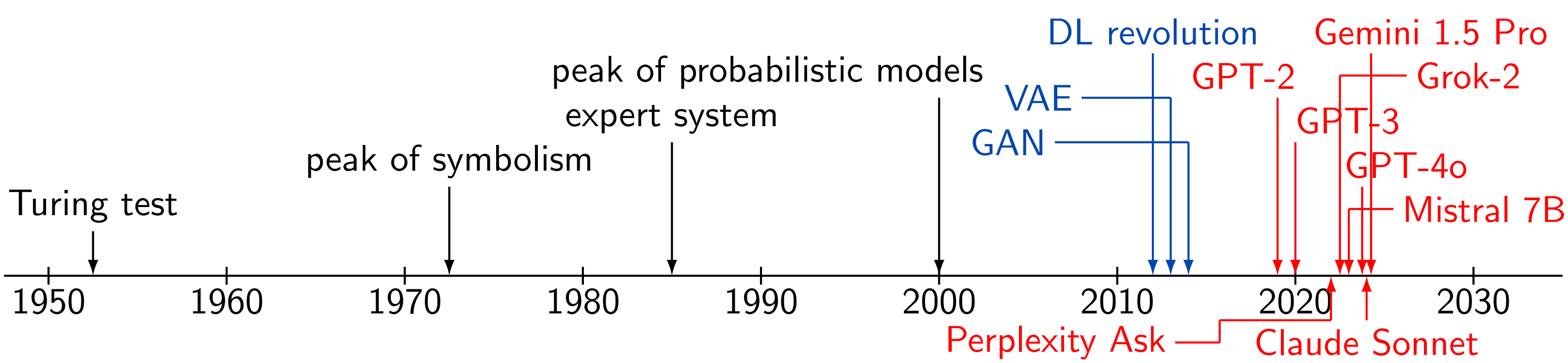
Birth of AI - early foundations & precursor technologies

- 1950s $\sim$ 1970s
  - Alan Turing - concept of *"thinking machine" & Turing test* to evaluate machine intelligence (1950s)
  - *symbolists* (as opposed to connectionists) - early AI focused on symbolic reasoning, logic & problem-solving - Dartmouth Conference in 1956 by *John McCarthy, Marvin Minsky, Allen Newell & Herbert A. Simon*
  - precursor technologies - genetic algorithms (GAs), Markov chains & *hidden Markov models (HMMs)* - laying foundation for generative processes (1970s $\sim$)



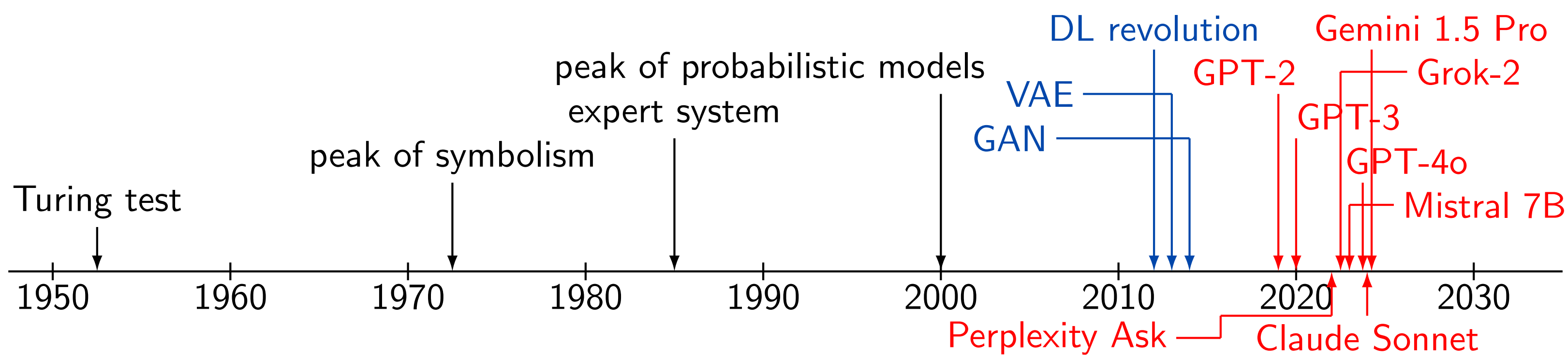
Rule-based systems & probabilistic models

- 1980s ∼ early 2000s
  - *expert systems* (1980s) - AI systems designed to mimic human decision-making in specific domains
  - development of neural networks (NN) w/ backpropagation *training multi-layered networks* - setting stage for way more complex generative models
  - *probabilistic models* (including network models, $i.e.$, Bayesian networks) & Markov models - laying groundwork for data generation & pattern prediction



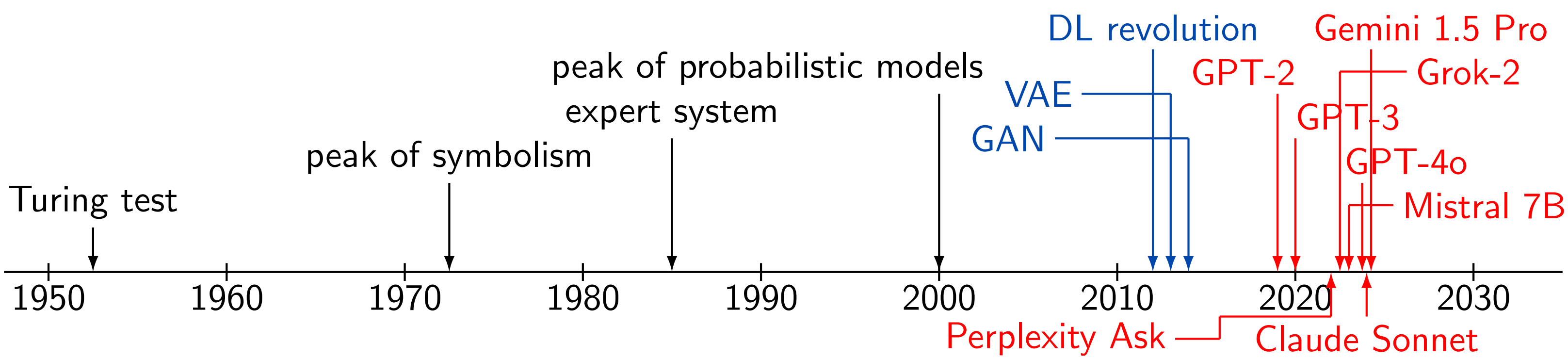
Rise of deep learning & generative models

- 2010s - breakthrough in genAI
  - *deep learning (DL) revolution* - advances in GPU computing and data availability led to the rapid development of deep neural networks.
  - *variational autoencoder (VAE)* (2013) - by Kingma and Welling - learns mappings between input and latent spaces
  - *generative adversarial network (GAN)* (2014) - by Ian Goodfellow - game-changer in generative modeling where two NNs compete each other to create realistic data
    - widely used in image generation & creative tasks



Transformer models & multimodal AI

- late 2010s $\sim$ Present
  - Transformer architecture (2017) - by Vaswani et al.
    - *revolutionized NLP,* $e.g.$, LLM & various genAI models
  - GPT series - generative pre-trained transformer
    - GPT-2 (2019) - generating human-like texts - *marking leap in language models*
    - GPT-3 (2020) - 175B params - set *new standards for LLM*
  - multimodal systems - DALL-E & CLIP (2021) - *linking text and visual data*
  - emergence of diffusion models (2020s) - new approach for generating high-quality images - progressively "denoising" random noise (DALL-E 2 & Stable Diffusion)



Significant AI Achievements - 2014 – 2025

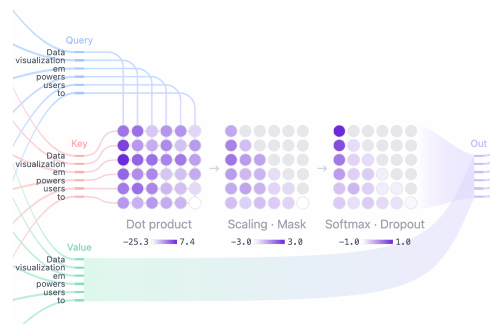
Deep learning revolution

- 2012 – 2015 - DL revolution[2]
  - 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

[2] CV: computer vision, NN: neural network, CNN: convolutional NN, RL: reinforcement learning

Transformer changes everything

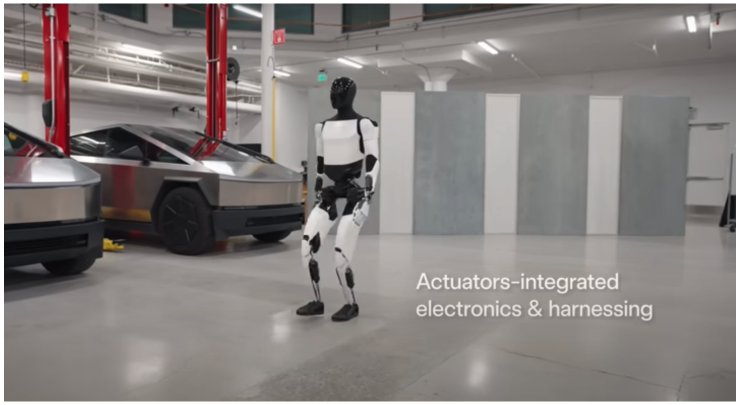
- 2017 – 2018 - Transformers & NLP breakthroughs[3]
  - *Transformer ($e.g.$, BERT & GPT) revolutionized NLP*
  - major advancements in, $e.g.$, machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting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



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

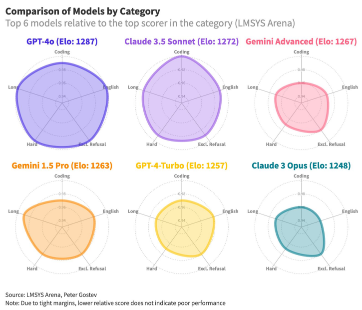
Lots of breakthroughs in AI technology and applications in 2024

- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Claude 3 series, Llama 3, Sora, Gemini
  - *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*



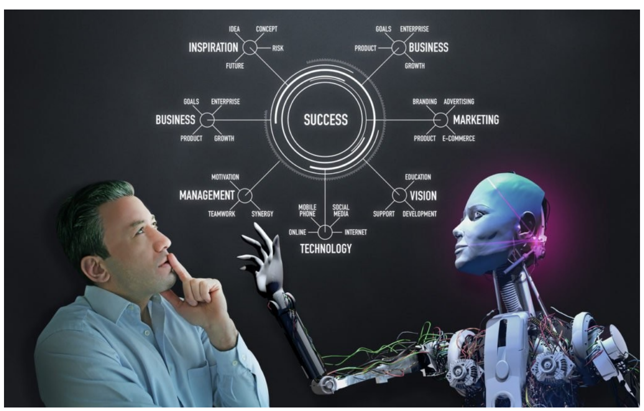
Major AI Breakthroughs in 2025

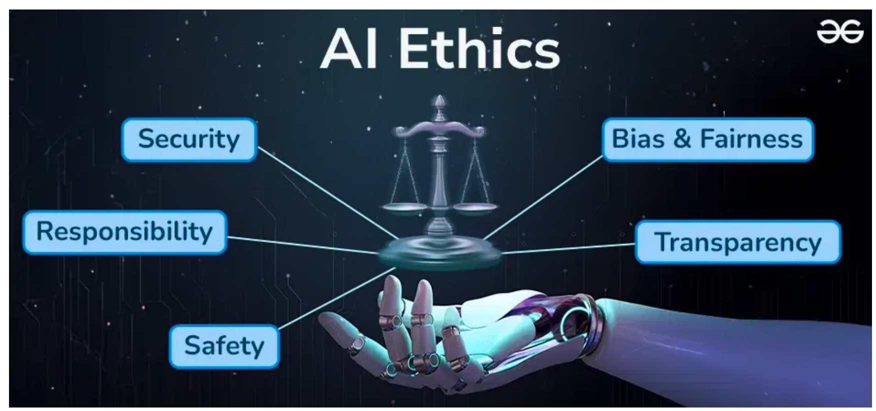
- next-generation foundation models
  - GPT-5 and Claude 4 demonstrate emergent reasoning abilities
  - open-source models achieving parity with leading commercial systems from 2024
- hardware innovations
  - NVIDIA's Blackwell successor architecture delivering 3-4x performance improvement
  - AMD's MI350 accelerators challenging NVIDIA's market dominance
- AI-human collaboration systems
  - seamless multimodal interfaces enabling natural human-AI collaboration
  - AI systems effectively explaining reasoning and recommendations
  - augmented reality interfaces providing real-time AI assistance in professional contexts



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*

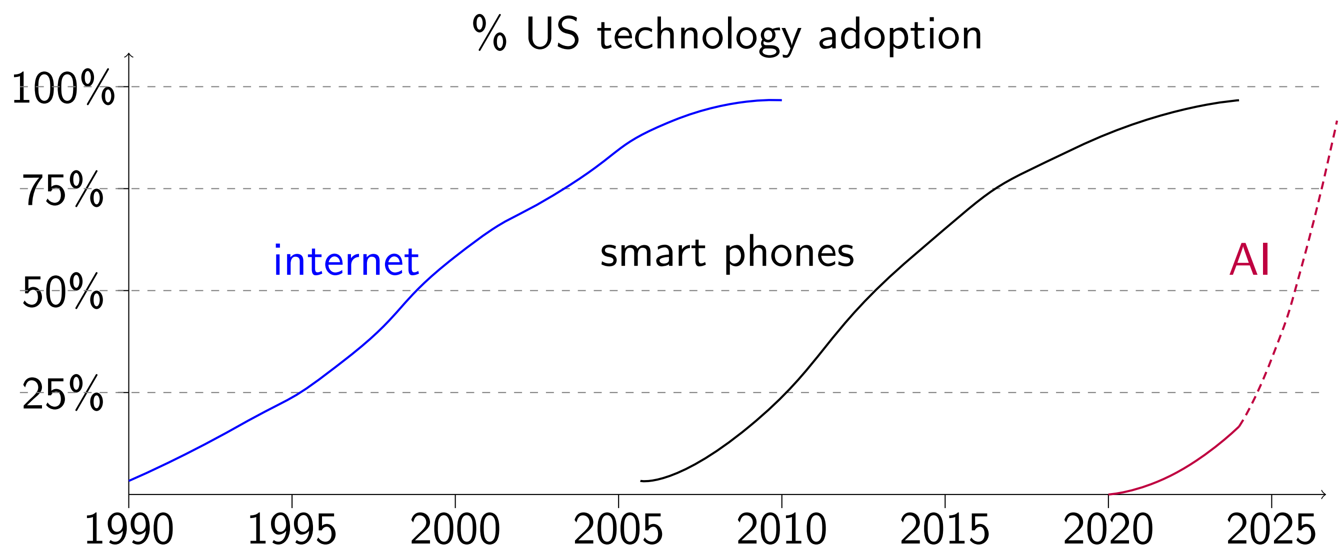




Measuring AI's Ascent

Where are we in AI today?

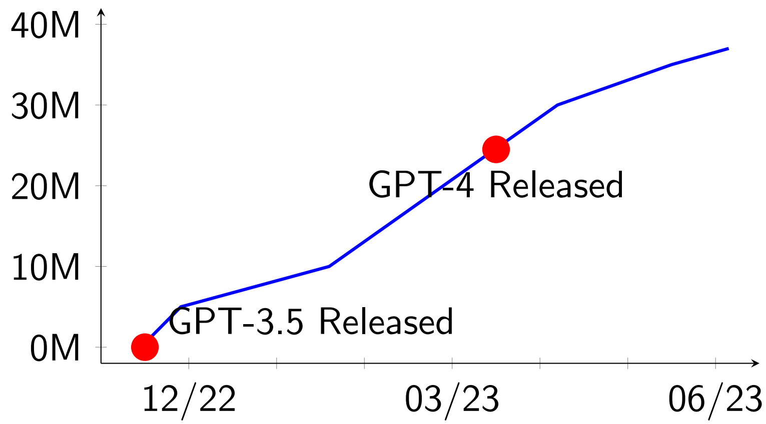
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



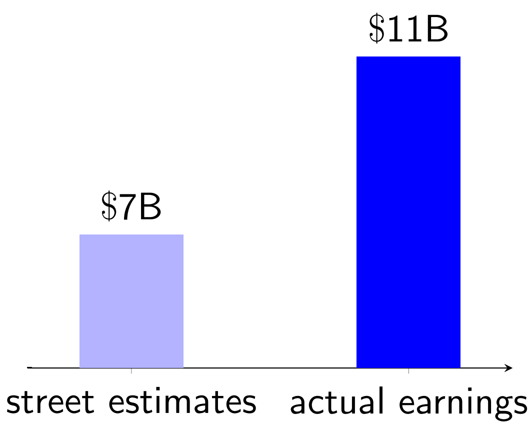
Explosion of AI ecosystems - ChatGPT & NVIDIA

- 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[4]

ChatGPT weekly unique visits



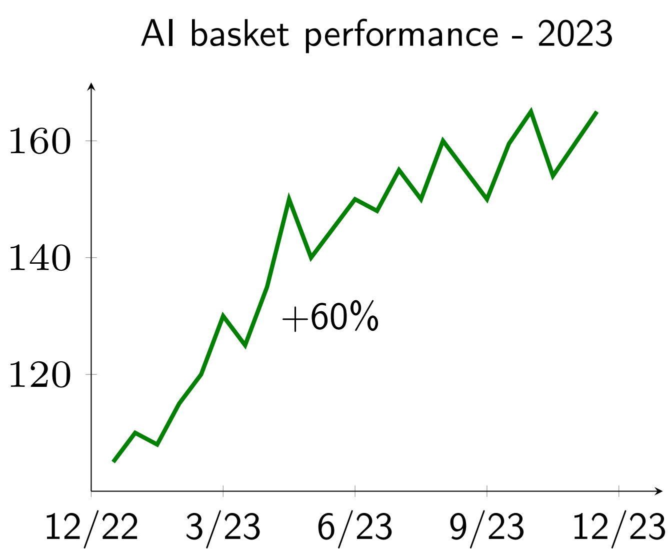
NVDA 2023-2Q earning

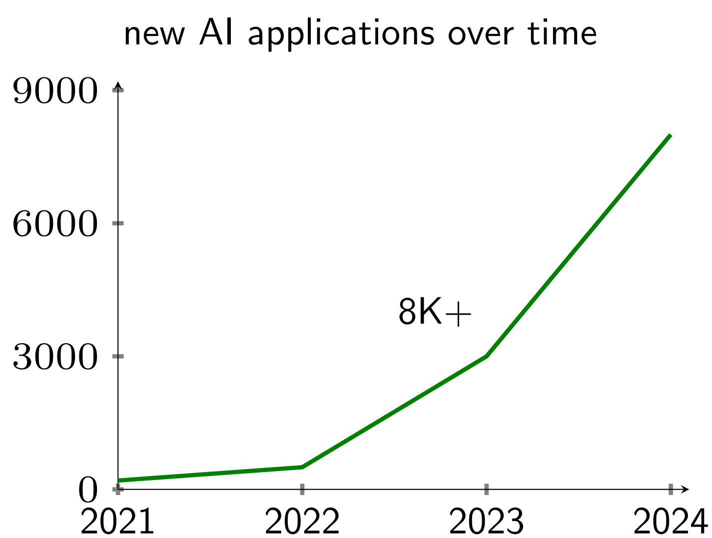


[4] source - Bloomberg

Explosion of AI ecosystems - AI stock market

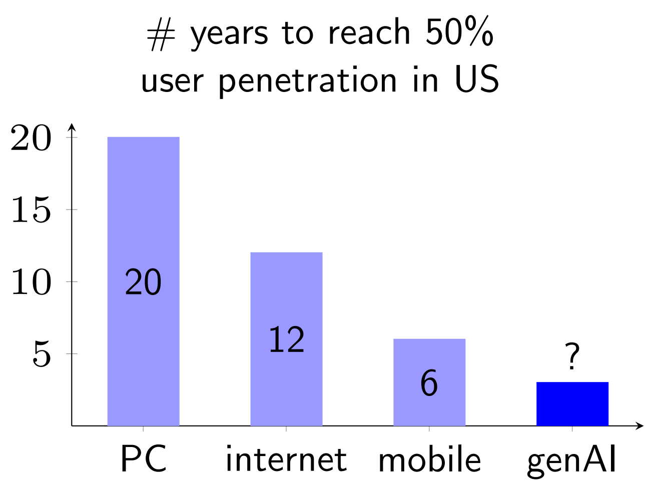
- *AI investment surge in 2023 - portfolio performance soars by 60%*
  - AI-focused stocks significantly outpaced traditional market indices
- *over 8,000 new AI applications* developed in last 3 years
  - applications span from healthcare and finance to manufacturing and entertainment

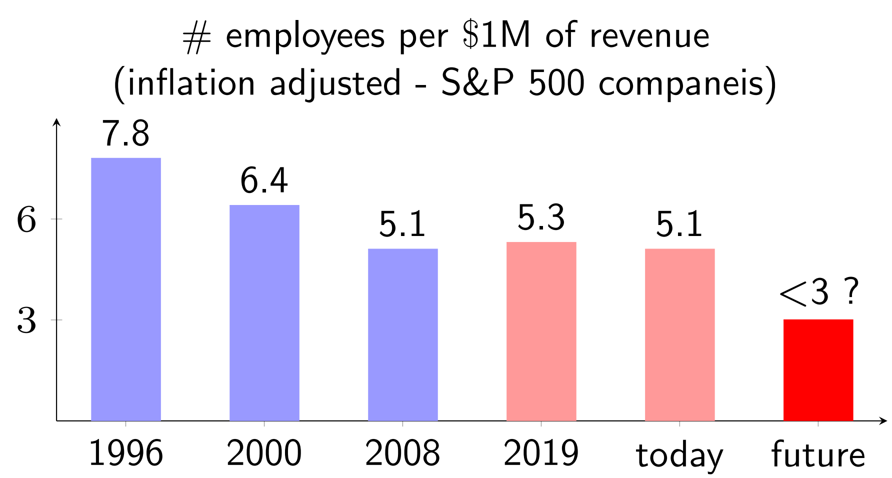




AI's transformative impact - adoption speed & economic potential

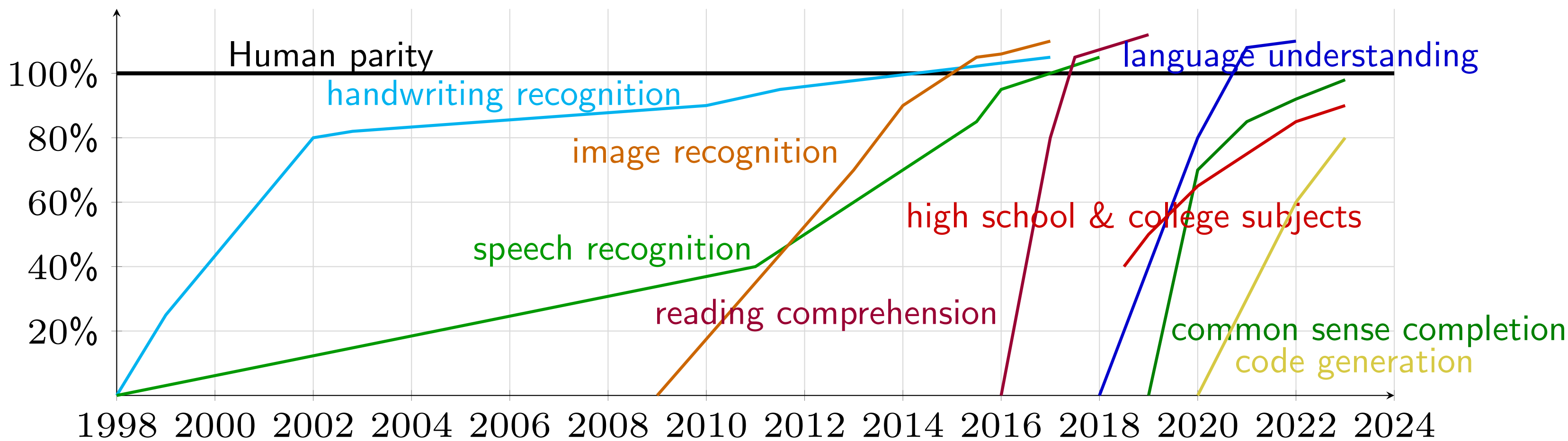
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI's potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater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





AI getting more & more faster

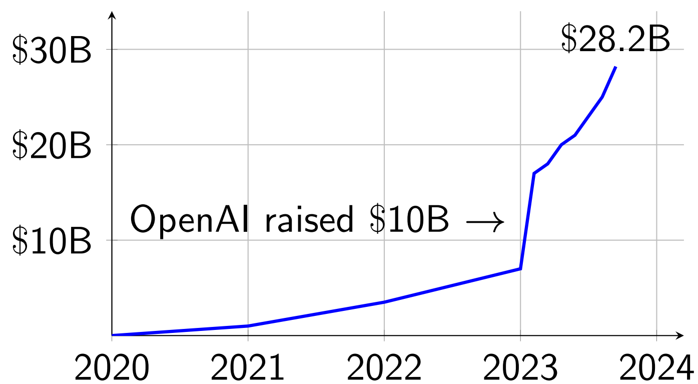
- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached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



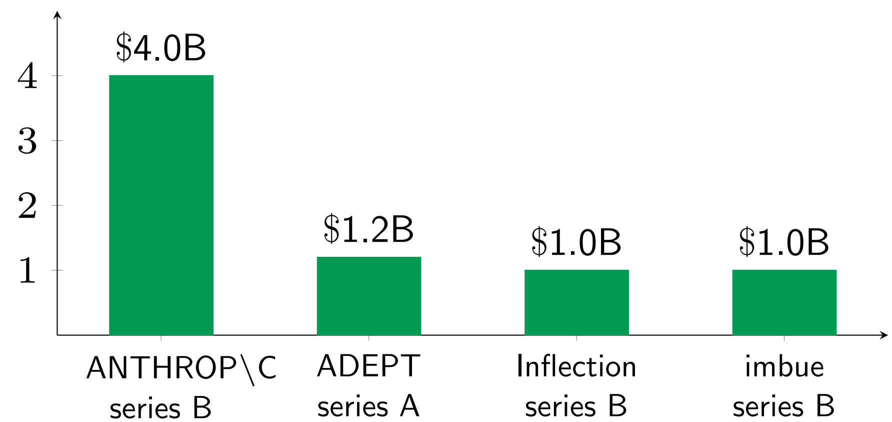
Massive investment in AI

- *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

cumulative funding in private
AI companies since 2020

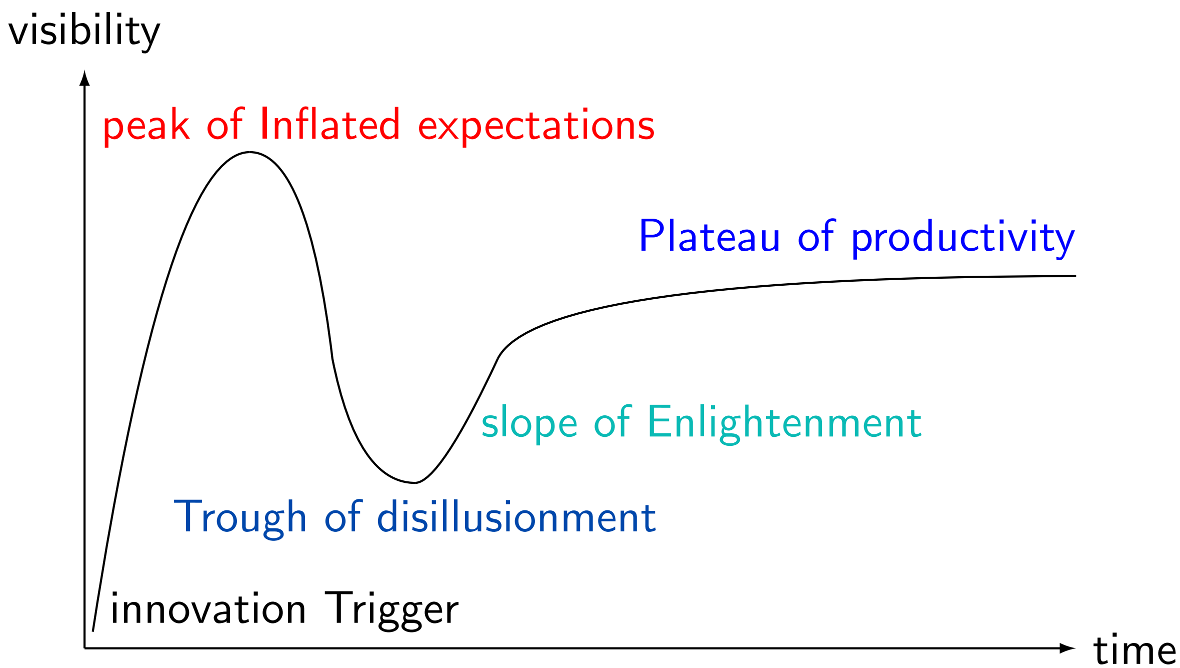


valuation before public product launch



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

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

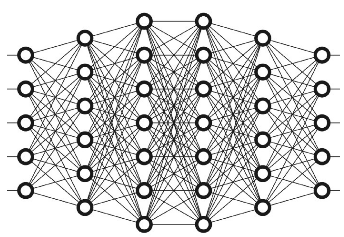
Yes & No

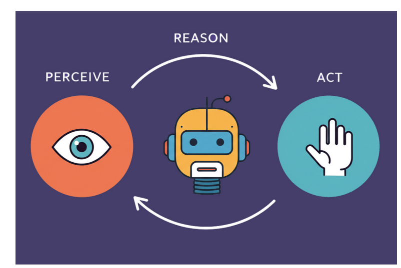
characteristics of hype cycles	speaker's views
value accrual misaligned with investment	• OpenAI still operating at a loss; business model *still* not clear • gradual value creation across broad range of industries and technologies (*e.g.*, CV, LLMs, RL) unlike fiber optic bubble in 1990s
overestimating 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 Agents

AI progress in 21st century in keywords

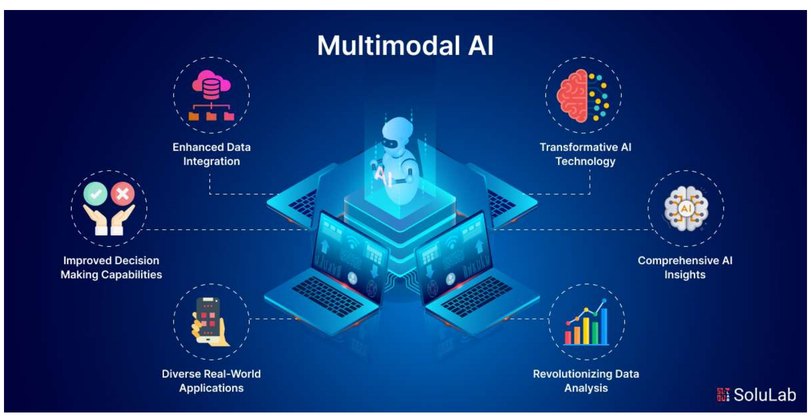
- 2010 $\sim$ Big Data
- 2012 $\sim$ Deep Learning
- 2017 $\sim$ Transformer - Attention is All you need!
- 2022 $\sim$ LLM & genAI
- 2024 $\sim$ AI Agent (Agentic AI)

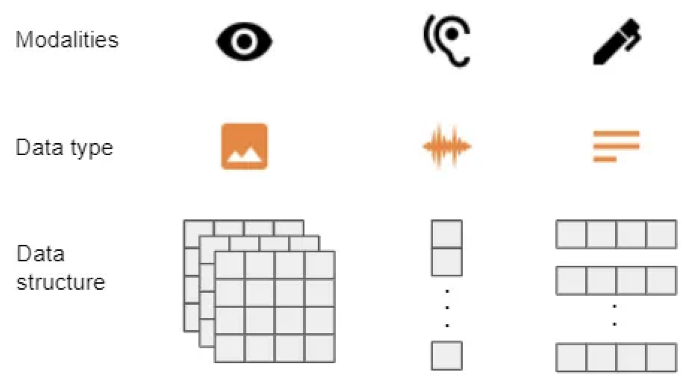




Multimodal learning

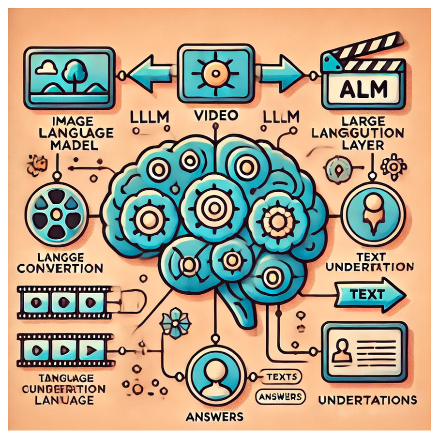
- understand information from multiple modalities, *e.g.*, text, images, audio, video
- representation learning methods
  - combine multiple representations or learn multimodal representations simultaneously
- applications
  - images from text prompt, videos with narration, musics with lyrics
- collaboration among different modalities
  - understand image world (open system) using language (closed system)

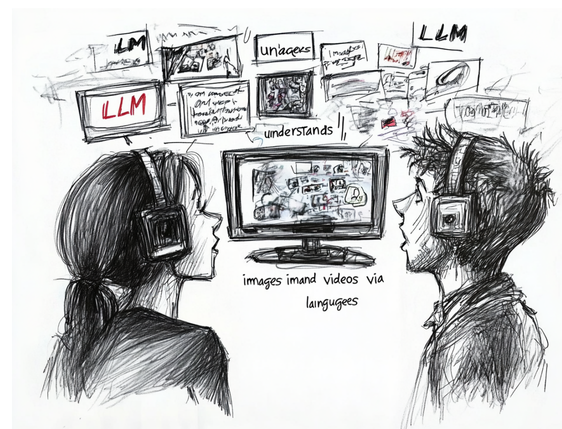




Implications of success of LLMs

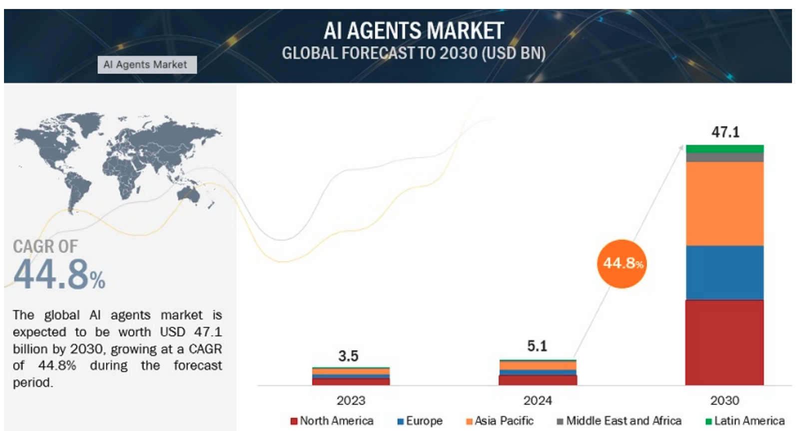
- many researchers change gears towards LLM
  – from computer vision (CV), speach, music, video, even reinforcement learning
- *LLM is not only about NLP . . .* humans have . . .
  - evolved to optimize natural language structures for eons
  - handed down knowledge using *this natural languages* for thousands of years
  – internal structure (or equivalently, representation) of natural languages optimized via *thousands of generation by evolution*
- LLM *connects non-linguistic world (open system) via natural languages (closed system)*

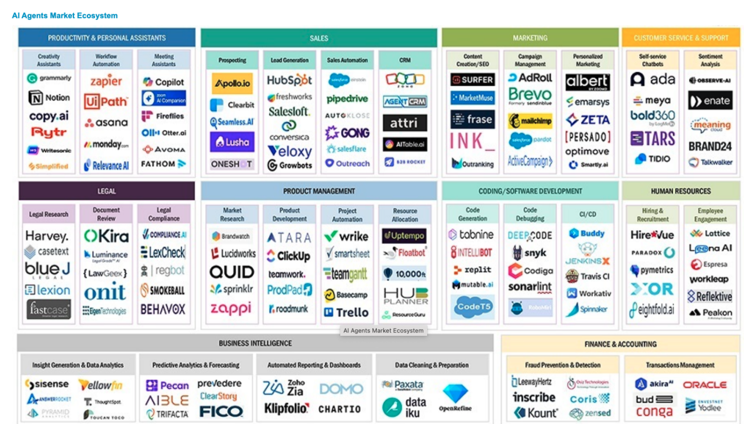




Multimodal AI (mmAI)

- mmAI - systems processing & integrating data from multiple sources & modalities, to generate unified response / decision
- 1990s – 2000s - early systems - initial research combining basic text & image data
- 2010s - CNNs & RNNs enabling more sophisticated handling of multimodality
- 2020s - modern multimodal models - Transformer-based architectures handling complex multi-source data at highly advanced level
- mmAI *mimics human cognitive ability* to interpret and integrate information from various sources, leading to holistic decision-making



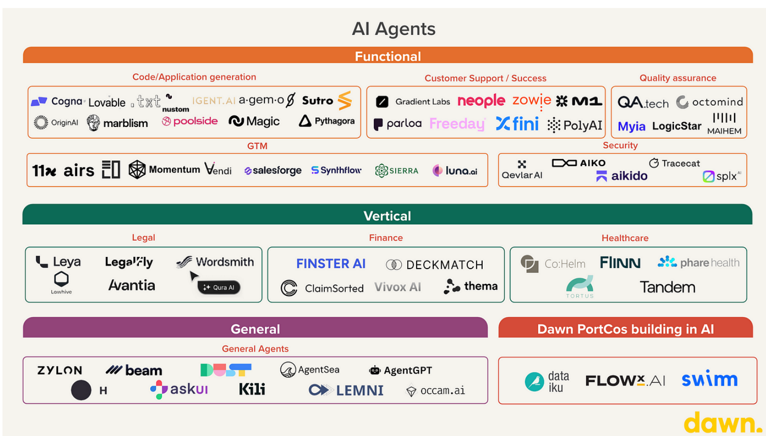


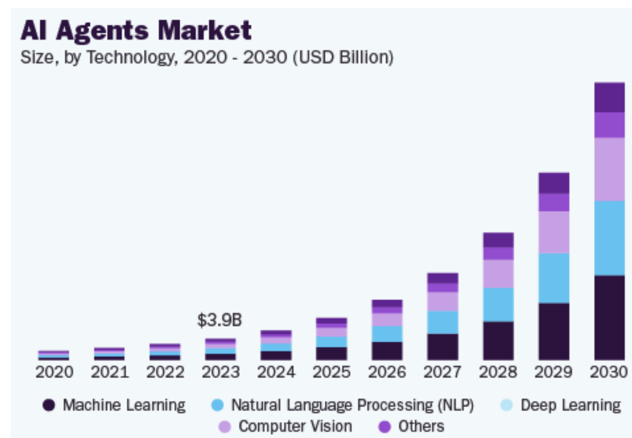
mmAI Technology

- core components
  - data preprocessing - images, text, audio & video
  - architectures - unified Transformer-based ($e.g.$, ViT) & cross-attention mechanisms / hybrid architectures ($e.g.$, CNNs + LLMs)
  - integration layers - fusion methods for combining data representations from different modalities
- technical challenges
  - data alignment - accurate alignment of multimodal data
  - computational demand - high-resource requirements for training and inferencing
  - diverse data quality - manage variations in data quality across modalities
- advancements
  - multimodal embeddings - shared feature spaces interaction between modalities
  - self-supervised learning - leverage unlabeled data to learn representations across modalities

AI agents powered by multimodal LLMs

- foundation
  - integrate multimodal AI capabilities for enhanced interaction & decision-making
- components
  - perceive environment through multiple modalities (visual, audio, text), process using LLM technology, generate contextual responses & take actions
- capabilities
  - understand complex environments, reason across modalities, engage in natural interactions, adapt behavior based on context & feedback





AI agents - Present & Future

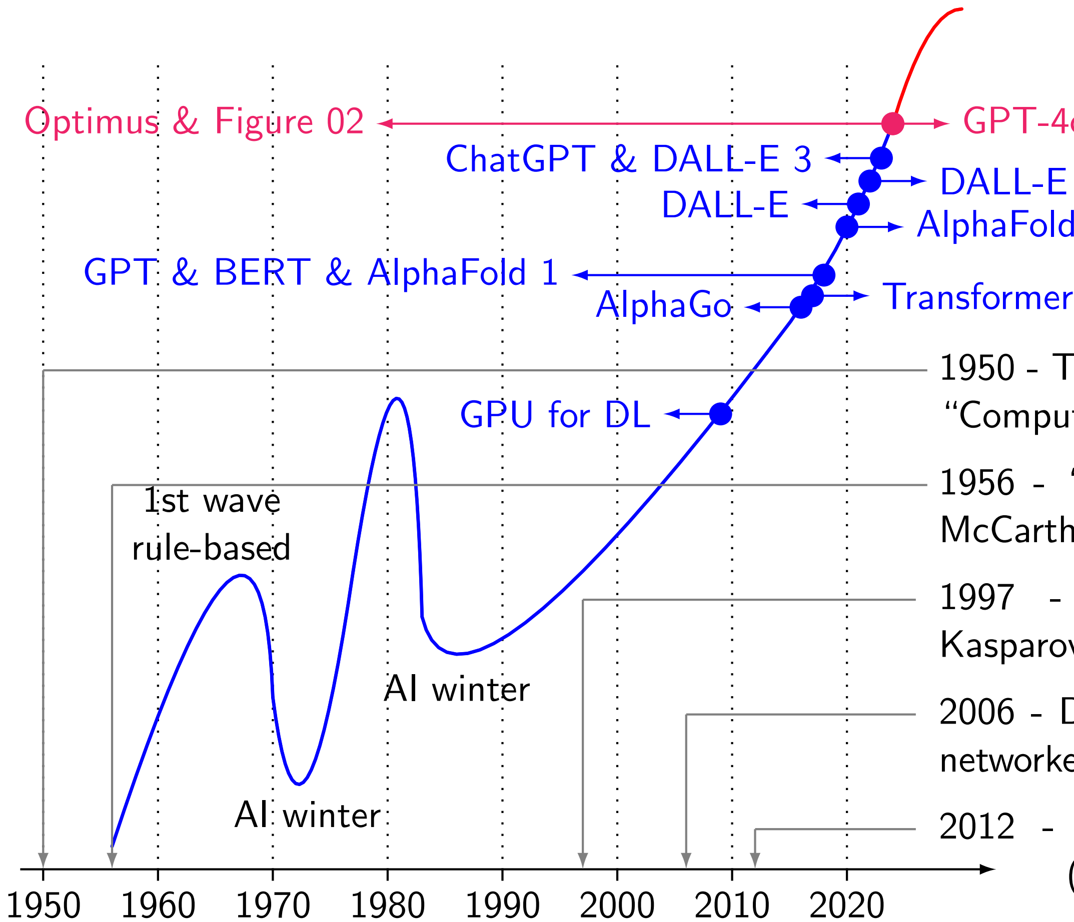
- emerging applications
  - scientific research - agents analyzing & running experiments & generating hypotheses
  - creative collaboration - AI partners in design & art combining multiple mediums
  - environmental monitoring - processing satellite sensor data for climate analysis
  - healthcare - enhanced diagnostic combining imaging, *e.g.*, MRI, with patient history
  - customer experience - virtual assistants understanding spoken language & visual cues
  - autonomous vehicles - integration of visual, radar & audio data
- future
  - ubiquitous AI agents - seamless integration into everyday devices
  - highly tailored personalized experience - in education, entertainment & healthcare

Silicon Valley's Cultural Engine of Innovation and Disruption

My journey from Samsung & Amazon to Gauss Labs & Erudio Bio

- Samsung Semiconductor, Inc.
  - inception into industry from academia, the world's best memory chip maker!
- Amazon.com, Inc.
  - experience so-called Silicon Valley big tech culture and technology
  - set tone for my future career trajectory!
- Gauss Labs, Inc.
  - found & operate AI startup, shaping corporate culture & spearheading R&D as CTO
  - inherent challenges of Korean conglomerate spin-off startup - cultural constraints, over-capitalization, and leadership limitations
- Erudio Bio, Inc.
  - concrete & tangible bio-technology in addition to AI
  - great decisions regarding business development; business models, market fit, go-to-market (GTM) strategies based on lessons learned *in a hard way* ☺

Joining Amazon.com, Inc. at the inflection point of AI



1950 - Turing Test introduced by Alan Turing in "Computing Machinery & Intelligence"

1956 - "Artificial Intelligence" coined by John McCarthy

1997 - IBM's Deep Blue defeated Garry Kasparov, (then) the world chess champion

2006 - DL Learning gained traction deep belief networked introduced by Geoffrey Hinton

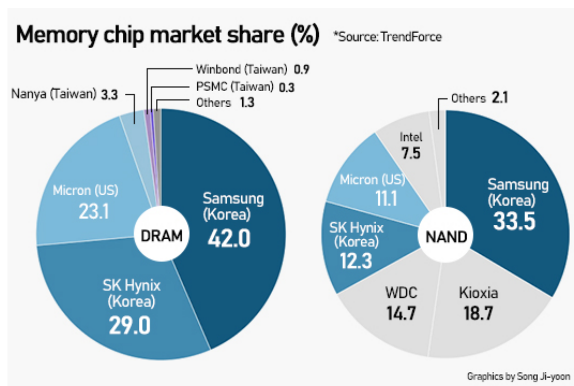
2012 - AlexNet won ImageNet Competition (ILSVRC) - widespread attention on DL

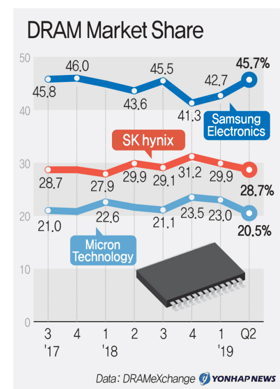
Innovation ecosystem of Silicon Valley

- key characteristics
  - risk-taking culture, *trust* in technology → *genuine* respect for engineers and scientists
  - easy access to huge capital - VCs, angel investors alike
  - talent density - engineers, researchers, scientists, entrepreneurs, PMs, TPMs, . . .
  - diversity, "collision density" of ideas
  - ecosystem of collaboration and competition - startups, academia, industry leaders
- what they mean for global big tech
  - set trends in AI, software & hardware (and or hence) product & industry innovation
  - act as testing ground for disruptive ideas

Case study: Amazon - amazing differentiators of big techs

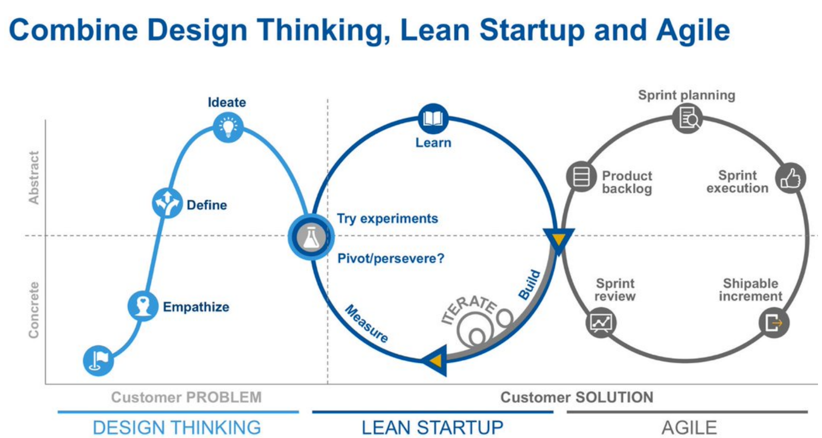
- Amazon's culture & leadership principles
  - customer obsession as driver of innovation
  - high standards & ownership culture, disagree & commit
  - bias for action and long-term thinking - sounds contradictory?
  - mechanisms like "two-pizza teams" & "Day One" for (or rather despite) scalability
- lessons for Korean corporations
  - applying customer-centric innovation in hardware & AI, *e.g.*, on-device AI
  - balancing agility with long-term R&D
  - *build / adapt / apply on the core strength of Samsung that no other company has!*

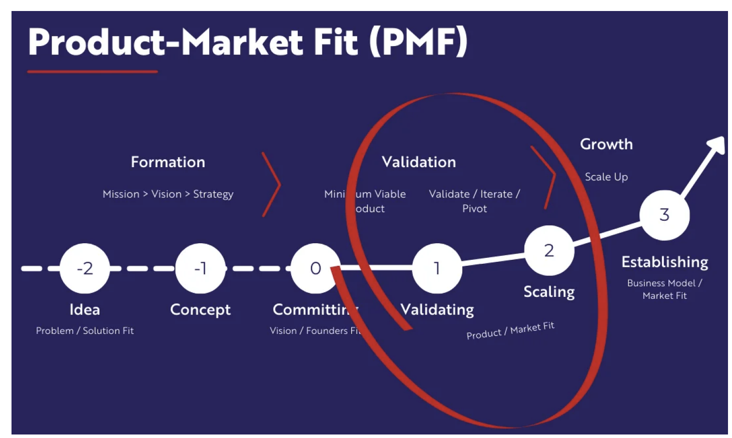




Founding and scaling startups

- challenges
  - competence of and chemistry among co-founders crucial
  - technology & great team are *necessary*, but *not sufficient (at all!)* for success
  - business models, market fit, timing, agility, flexibility for pivoting / perseverance
- insight
  - importance of domain expertise in addition to AI
  - balancing innovation with good business decisions





Bridging Silicon Valley & Korea

- cultural differences
  - risk appetite & failure tolerance
  - decision-making speed vs hierarchy
  - innovation vs execution focus
- opportunities for collaboration
  - leveraging Korea's manufacturing expertise with Silicon Valley's software/AI strengths
  - building global teams with diverse perspectives

To be successful . . .

- embrace customer/market-centric mindset in innovation and for business decisions
- balance agility with long-term vision
- foster cross-cultural collaboration for global impact
- ((very) strategically and carefully) leverage AI to solve real-world industrial challenges

Appendix

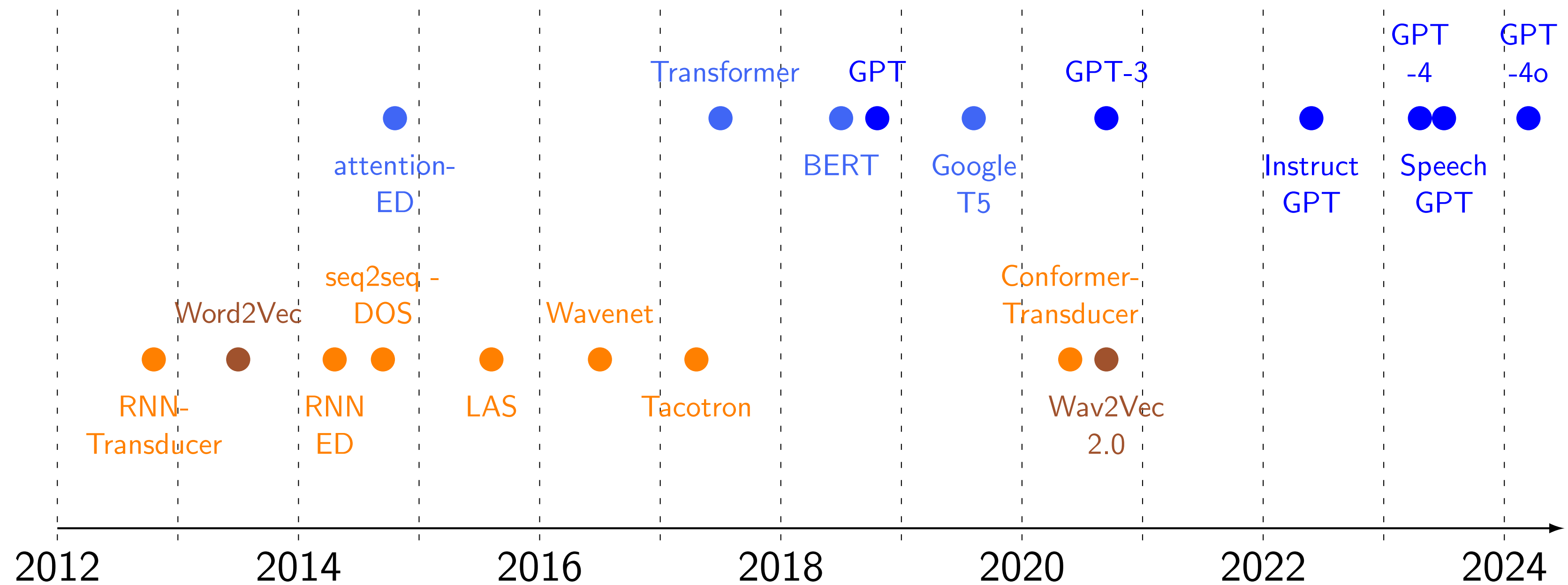
LLM

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

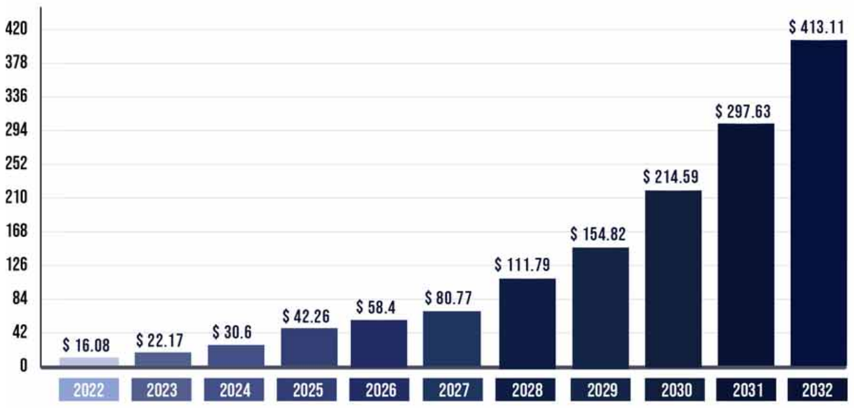
voice recognition	NL generation	summarization	spelling correction	action planning
voice synthesis	translation	topic model	topic model	
	Q&A	NLU		

NLP Market

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

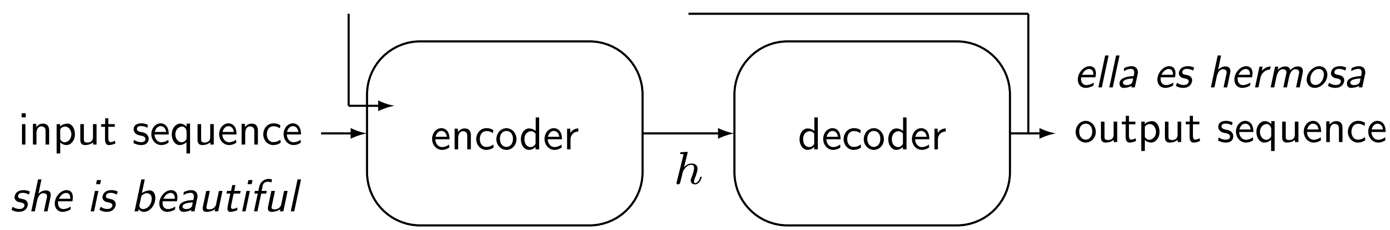
- source - Precedence Research



Sequence-to-Sequence Models

Sequence-to-sequence (seq2seq) model

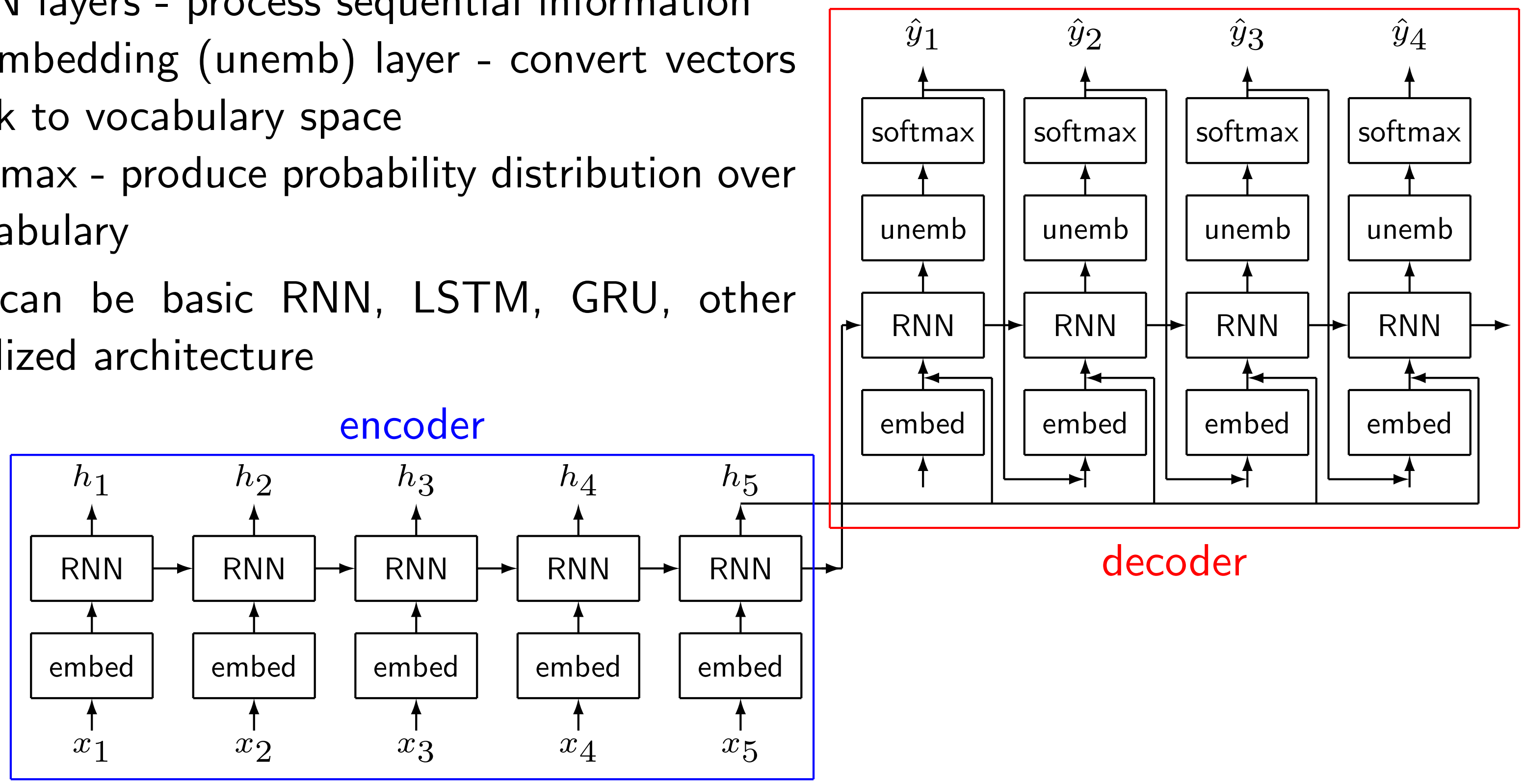
- seq2seq - take sequences as inputs and spit out sequences
- encoder-decoder architecture



  - encoder & decoder can be RNN-type models
  - $h \in \mathbf{R}^n$ - 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

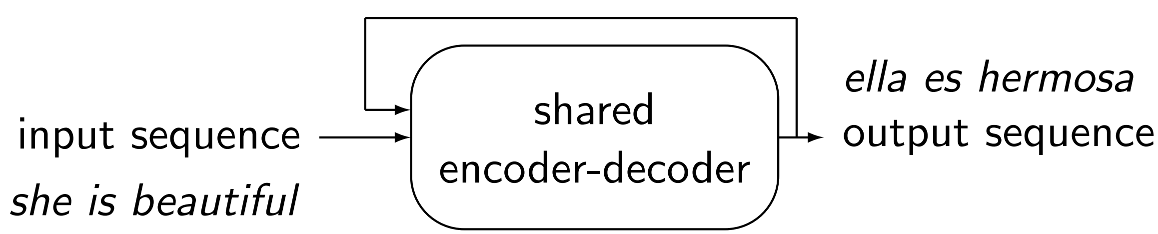
RNN-type encoder-decoder architecture

- components
  - embedding layer - convert input tokens to vector representations
  - RNN layers - process sequential information
  - unembedding (unemb) layer - convert vectors back to vocabulary space
  - softmax - produce probability distribution over vocabulary
- RNN can be basic RNN, LSTM, GRU, other specialized architecture



Shared encoder-decoder model

- single neural network structure can handle both encoding & decoding tasks
  - efficient architecture reducing model complexity
  - allow for better parameter sharing across tasks
- widely used in modern LLMs to process & generate text sequences
  - applications - machine translation, text summarization, question answering
- advantages
  - efficient use of parameters, versatile for multiple NLP tasks



Large Language Models

LLM

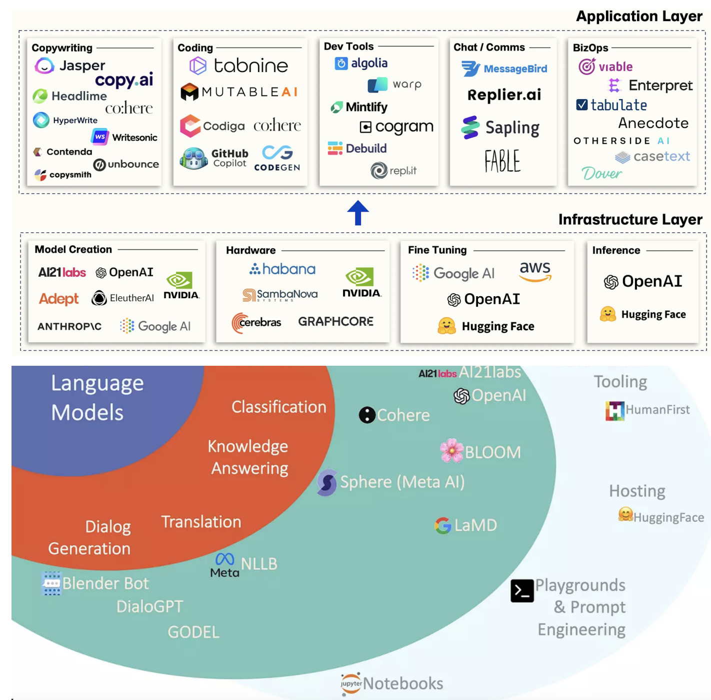
- LLM
  - type of AI aimed for NLP trained on massive corpus of texts & programming code
  - allow learn statistical relationships between words & phrases, $i.e.$, conditional probabilities
  - *amazing performance shocked everyone - unreasonable effectiveness of data (Halevry et al., 2009)*

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

  + virtually infinite # of applications

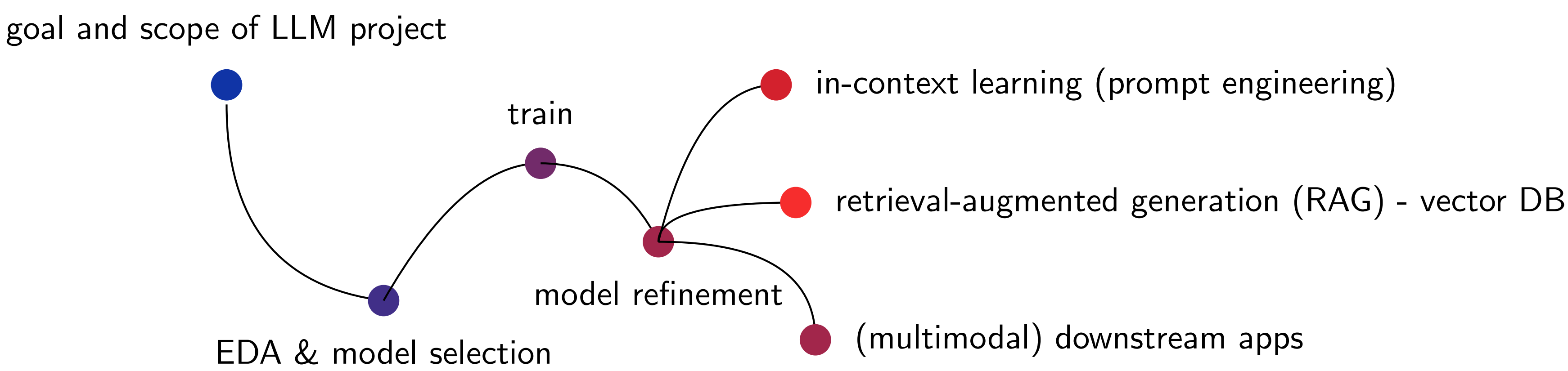
LLMs

- 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

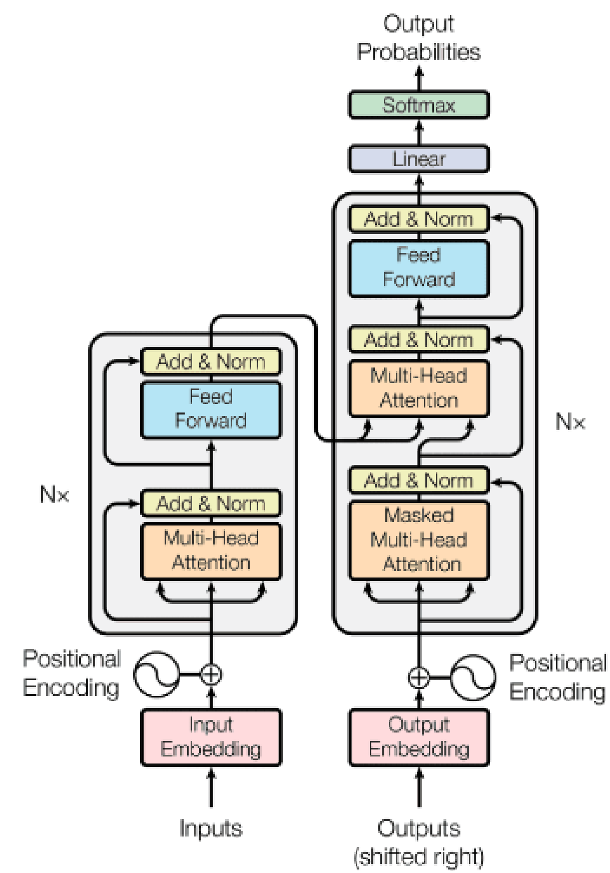
LLM architectural secret (or known) sauce

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)

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

- *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

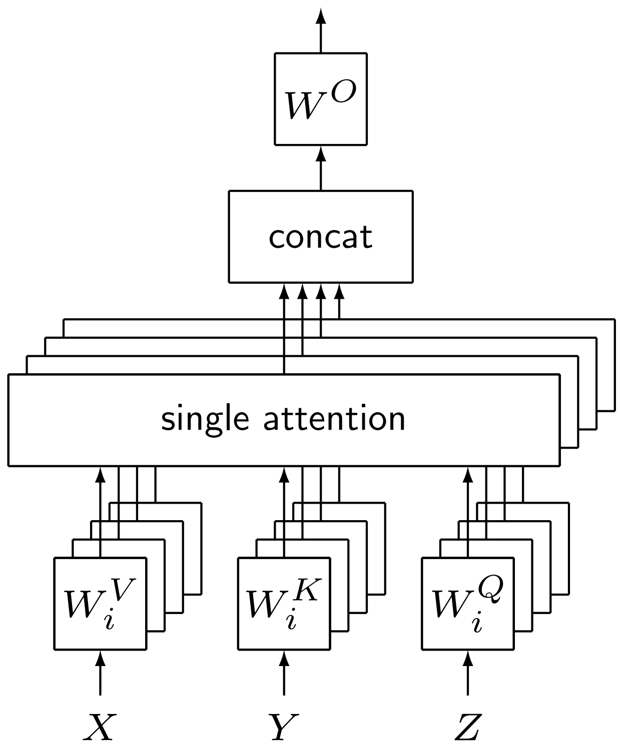
- evaluate h single-head attentions (in parallel)
- d_e: dimension for embeddings
- embeddings

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

  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!*

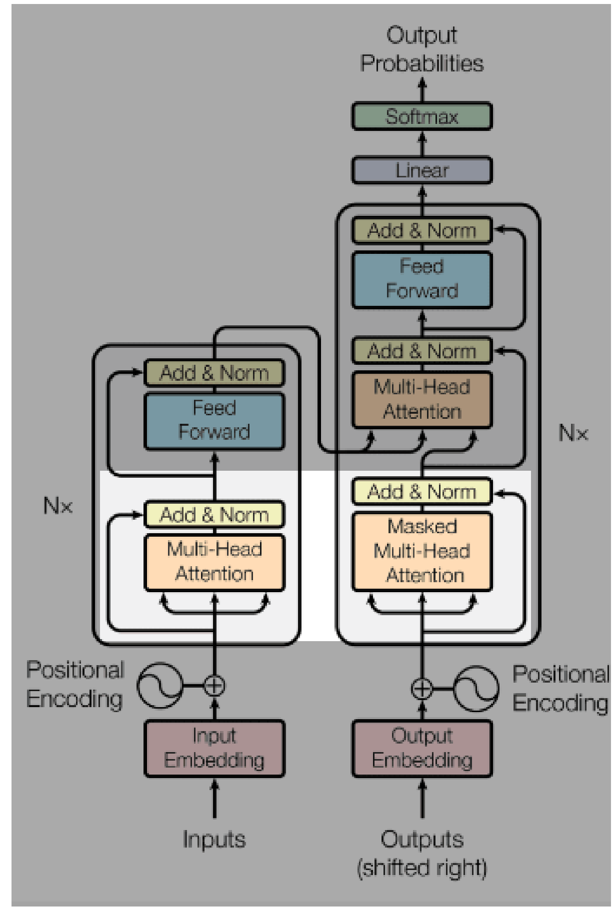
$$W^O \begin{bmatrix} A_1 \\ \vdots \\ A_h \end{bmatrix} \in \mathbf{R}^{d_e \times n},$$

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



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

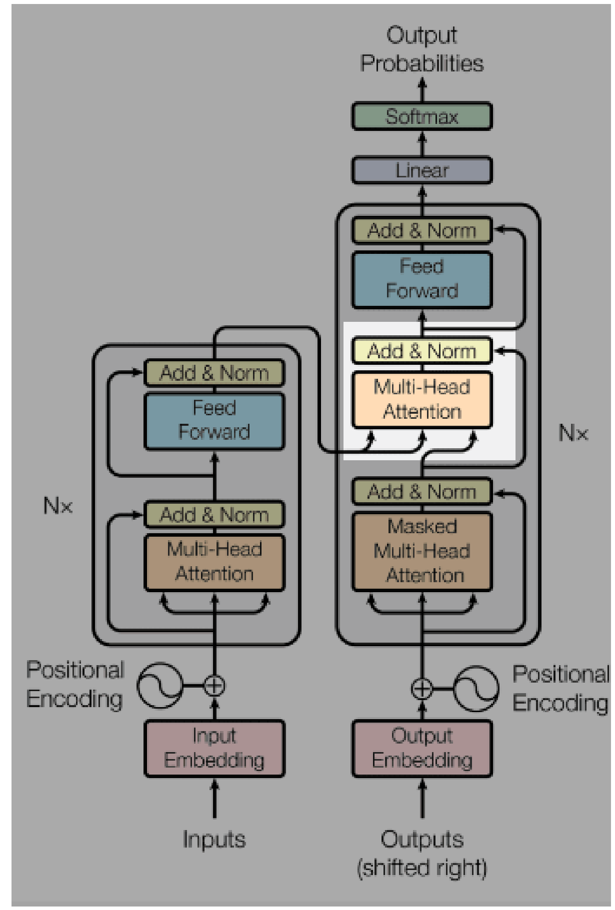
- m: length of input sequence

- 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

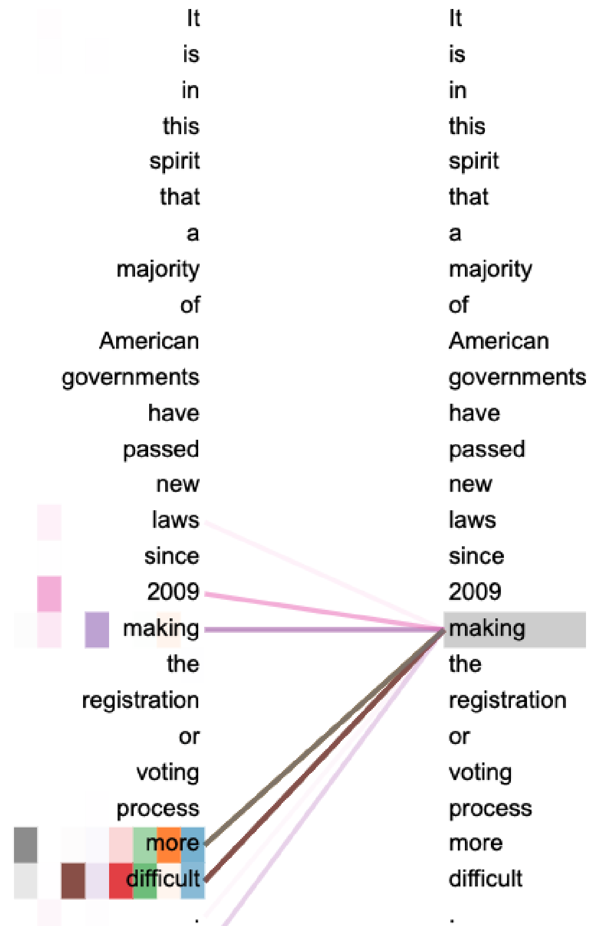


Visualization of self attentions

example sentence

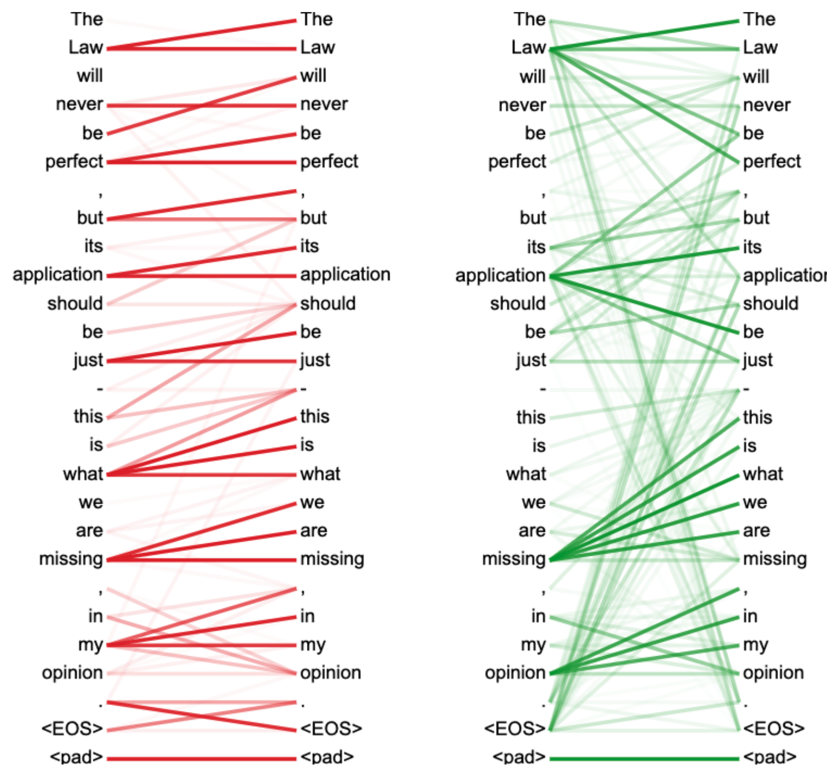
"It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult."

- 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"

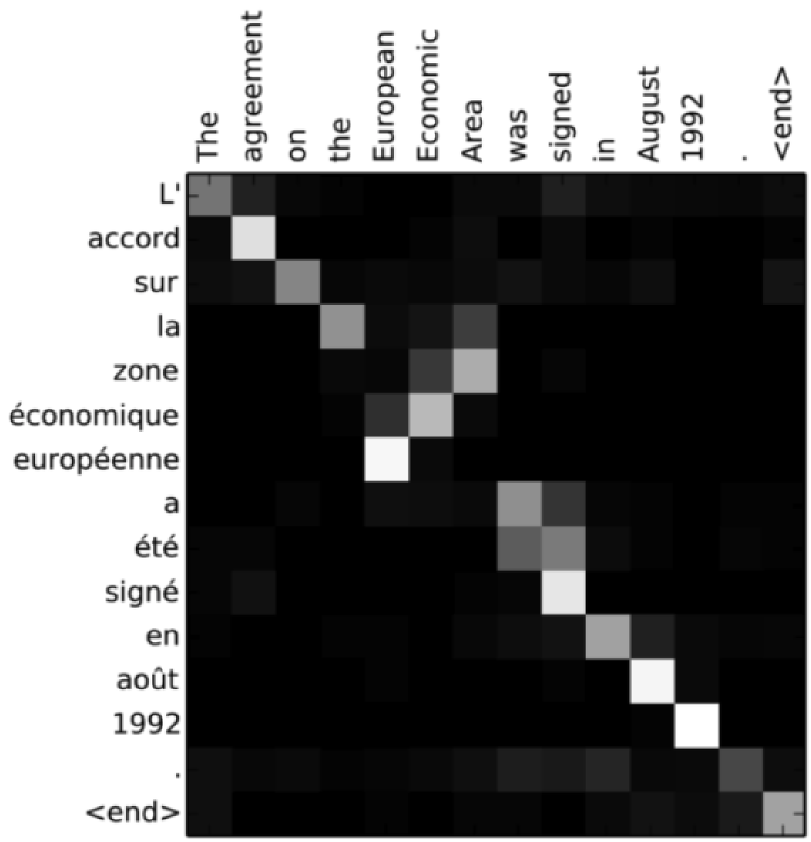


Visualization of multi-head self attentions

- 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!*

Variants of Transformer

Bidirectional encoder representations from transformers (BERT)

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

input	[cls]	my	dog	is	cute	[sep]	she	likes	playing	[sep]
token embeddings	$E_{\text{[cls]}}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{\text{[sep]}}$	E_{she}	E_{likes}	E_{playing}	$E_{\text{[sep]}}$
	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
segment embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B
	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
poistional embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9

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

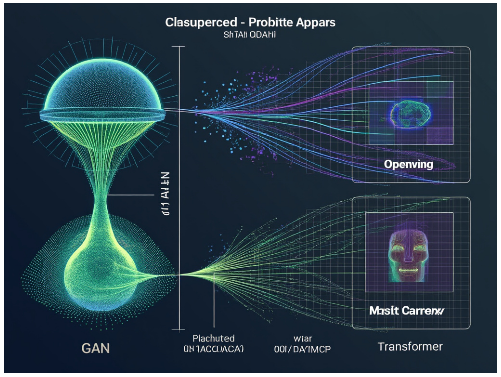
genAI

Definition of genAI

Generative AI

- genAI refers to systems capable of producing new (& original) contents based on patterns learned from training data (representation learning)
  - as opposed to discriminative models for, *e.g.*, classification, prediction & regression
  - here content can be text, images, audio, video, *etc.* - what about smell & taste?
- genAI model examples
  - generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, Transformers

by Midjourney



by Grok 2 mini

by Generative AI Lab

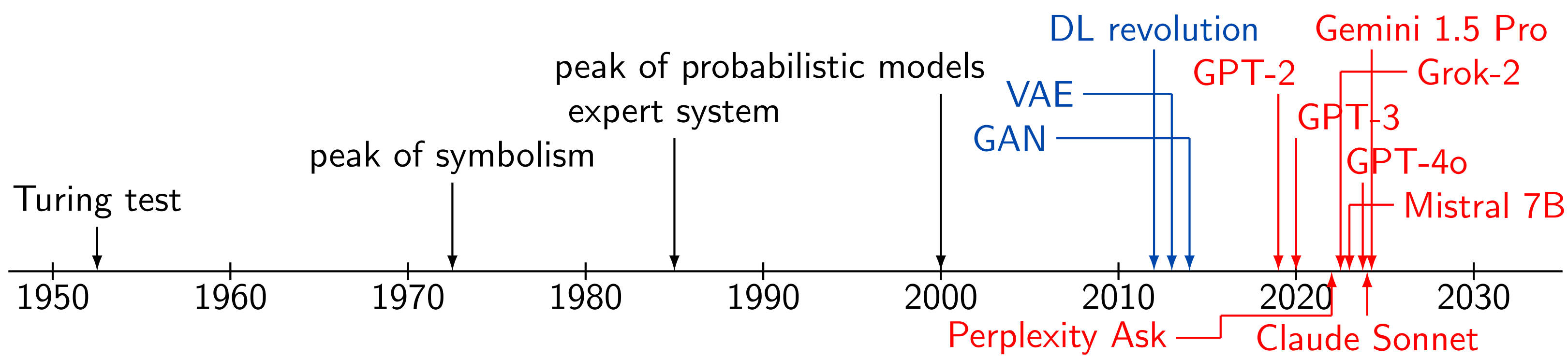
Examples of genAI in action

- text generation
  - Claude, ChatGPT, Mistral, Perplexity, Gemini, Grok
  - conversational agent writing articles, code & even poetry
- image generation
  - DALL-E - creates images based on textual descriptions
  - Stable Diffusion - uses diffusion process to generate high-quality images from text prompts (by denoising random noise)
  - MidJourney - art and visual designs generated through deep learning
- music generation
  - Amper Music - generates unique music compositions
- code generation
  - GitHub Copilot - generates code snippets based on natural language prompts

History of genAI

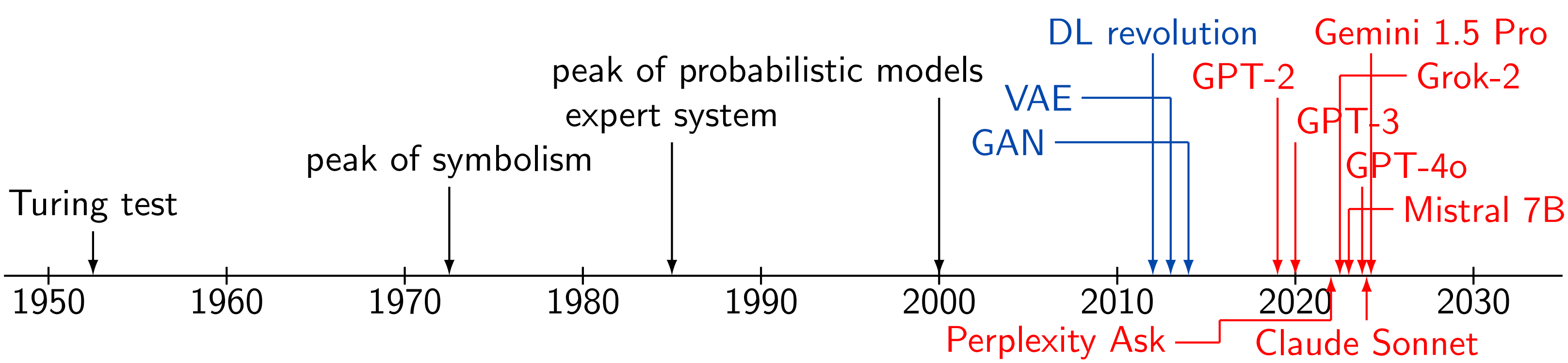
Birth of AI - early foundations & precursor technologies

- 1950s $\sim$ 1970s
  - Alan Turing - concept of *"thinking machine" & Turing test* to evaluate machine intelligence (1950s)
  - *symbolists* (as opposed to connectionists) - early AI focused on symbolic reasoning, logic & problem-solving - Dartmouth Conference in 1956 by *John McCarthy, Marvin Minsky, Allen Newell & Herbert A. Simon*
  - precursor technologies - genetic algorithms (GAs), Markov chains & *hidden Markov models (HMMs)* - laying foundation for generative processes (1970s $\sim$)



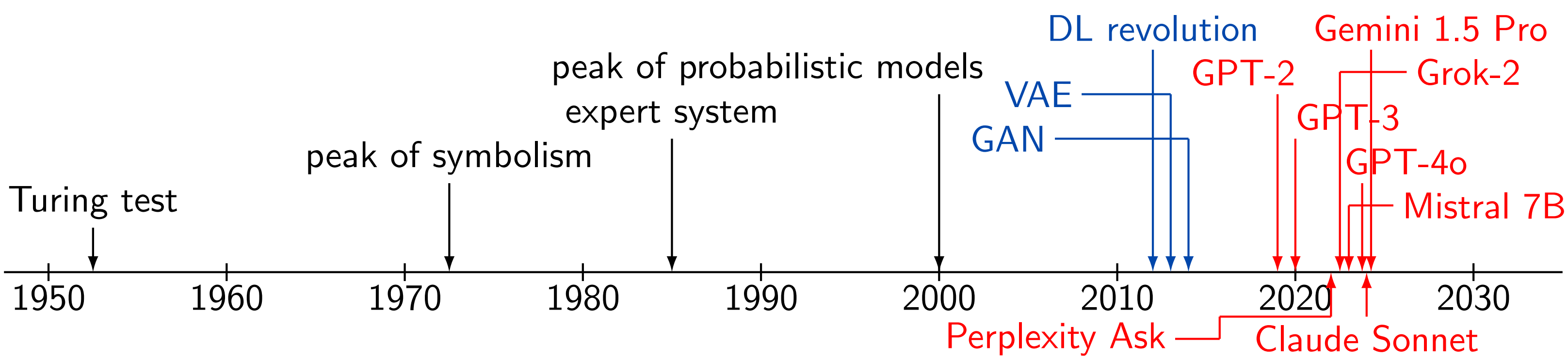
Rule-based systems & probabilistic models

- 1980s $\sim$ early 2000s
  - *expert systems* (1980s) - AI systems designed to mimic human decision-making in specific domains
  - development of neural networks (NN) w/ backpropagation *training multi-layered networks* - setting stage for way more complex generative models
  - *probabilistic models* (including network models, $i.e.$, Bayesian networks) & Markov models - laying groundwork for data generation & pattern prediction



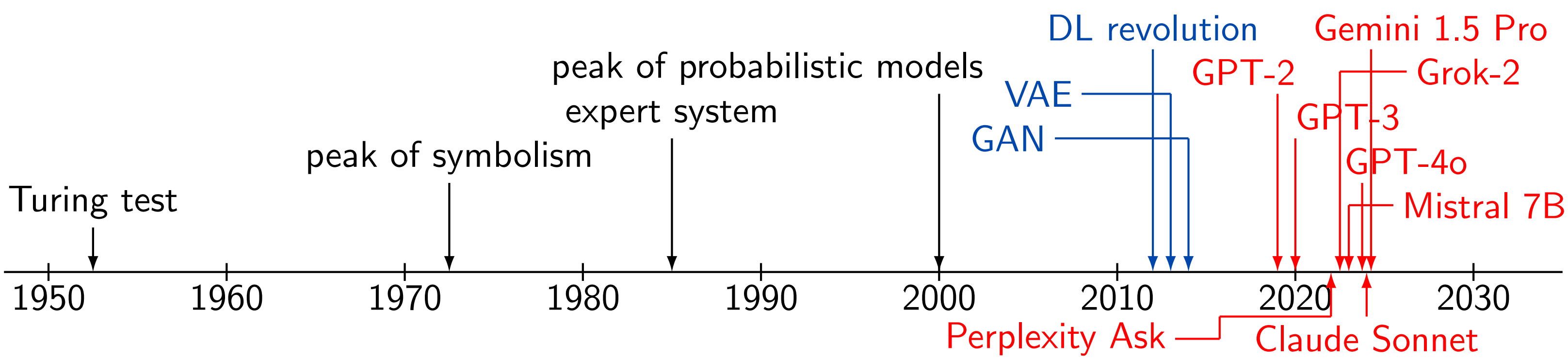
Rise of deep learning & generative models

- 2010s - breakthrough in genAI
  - *deep learning (DL) revolution* - advances in GPU computing and data availability led to the rapid development of deep neural networks.
  - *variational autoencoder (VAE)* (2013) - by Kingma and Welling - learns mappings between input and latent spaces
  - *generative adversarial network (GAN)* (2014) - by Ian Goodfellow - game-changer in generative modeling where two NNs compete each other to create realistic data
    - widely used in image generation & creative tasks



Transformer models & multimodal AI

- late 2010s $\sim$ Present
  - Transformer architecture (2017) - by Vaswani et al.
    - *revolutionized NLP,* $e.g.$, LLM & various genAI models
  - GPT series - generative pre-trained transformer
    - GPT-2 (2019) - generating human-like texts - *marking leap in language models*
    - GPT-3 (2020) - 175B params - set *new standards for LLM*
  - multimodal systems - DALL-E & CLIP (2021) - *linking text and visual data*
  - emergence of diffusion models (2020s) - new approach for generating high-quality images - progressively "denoising" random noise (DALL-E 2 & Stable Diffusion)



Mathy Views on genAI

genAI models

- 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]

$$\begin{tikzcd} & & p(x) \arrow[dr, "x_{\mathrm{data}}"] & & \\ q(z) \arrow[r, "z"] & \boxed{g(\theta_G; z)} \arrow[rr, "x_{\mathrm{model}}"] & & \boxed{f(\theta_D; x)} \arrow[r] & \text{true / false} \end{tikzcd}$$

  – 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*

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

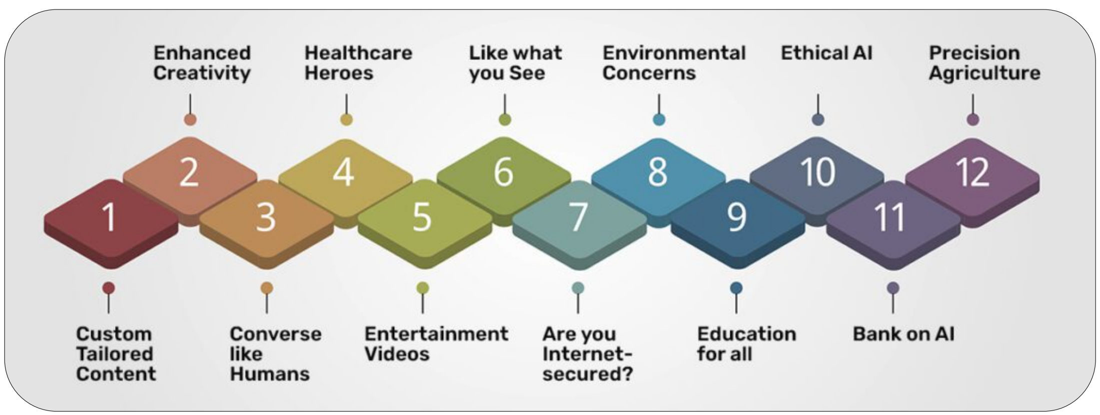
  where $d(\cdot,\cdot)$ distance measure between probability distributions
  - previous sequence: $x_{t-1}, x_{t-2}, \ldots$
  - next token: x_t
- p_θ represented by (extremely) complicated model
  - $e.g.$, containing multi-head & multi-layer Transformer architecture inside
- model parameters, $e.g.$, for Llama2

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

Current Trend & Future Perspectives

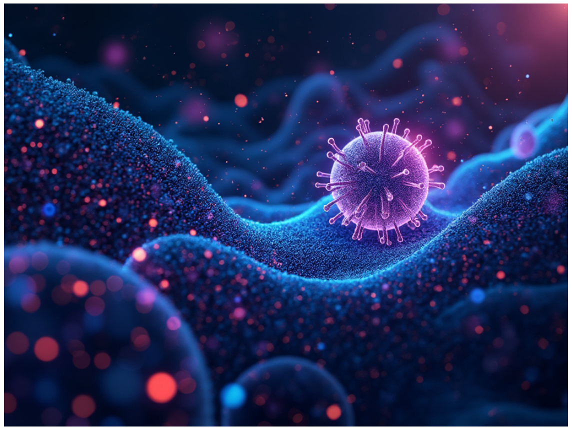
Current trend of genAI

- rapid advancement in language models & multimodal AI capabilities
- rise of AI-assisted creativity & productivity tools
- growing adoption across industries
  - creative industries - design, entertainment, marketing, software development
  - life sciences - healthcare, medical, biotech
- infrastructure & accessibility, *e.g.*, Hugging Face democratizes AI development
- integration with cloud platforms & enterprise-level tools
- increased focus on AI ethics & responsible development



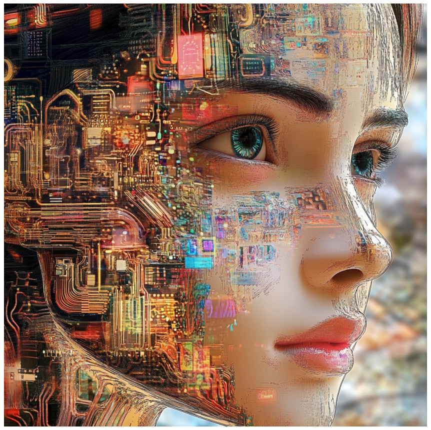
Industry & business impacts

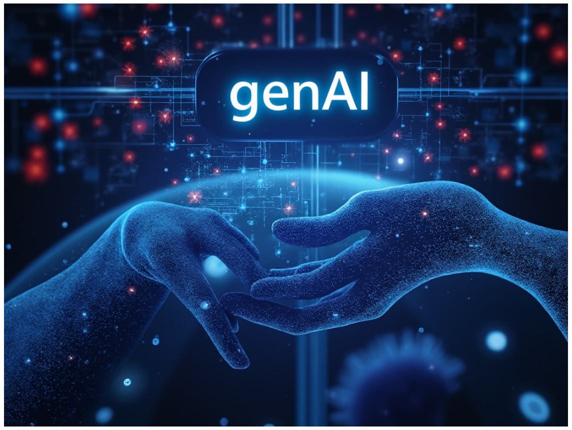
- how genAI is transforming industries
  - creative industries - content creation - advertising, gaming, film
  - life science - enhance research, drug discovery & personalized treatments
  - finance - automating document generation, risk modeling & fraud detection
  - manufacturing & Design - rapid prototyping, 3D modeling & optimization
  - business operations - automate routine tasks to boost productivity



Future perspectives of genAI

- hyper-personalization - highly personalized content for individual users - music, products & services
- AI ethics & governance - concerns over deepfakes, misinformation & bias
- interdisciplinary synergies - integration with other fields such as quantum computing, neuroscience & robotics
- human-AI collaboration - augment human creativity rather than replace it
- energy efficiency - have to figure out how to dramatically reduce power consumption

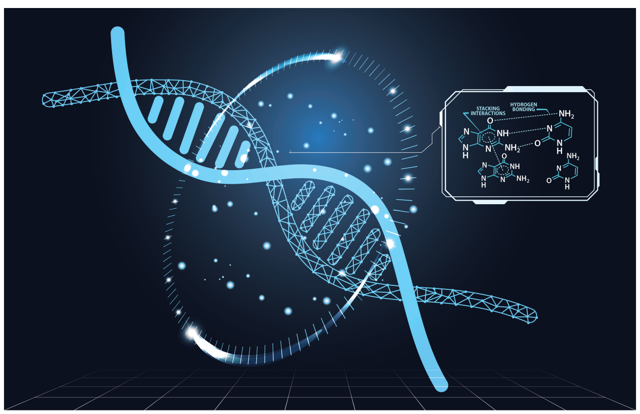


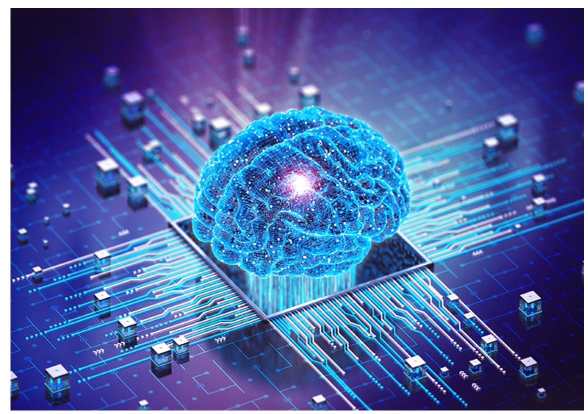


AI & Biotech

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





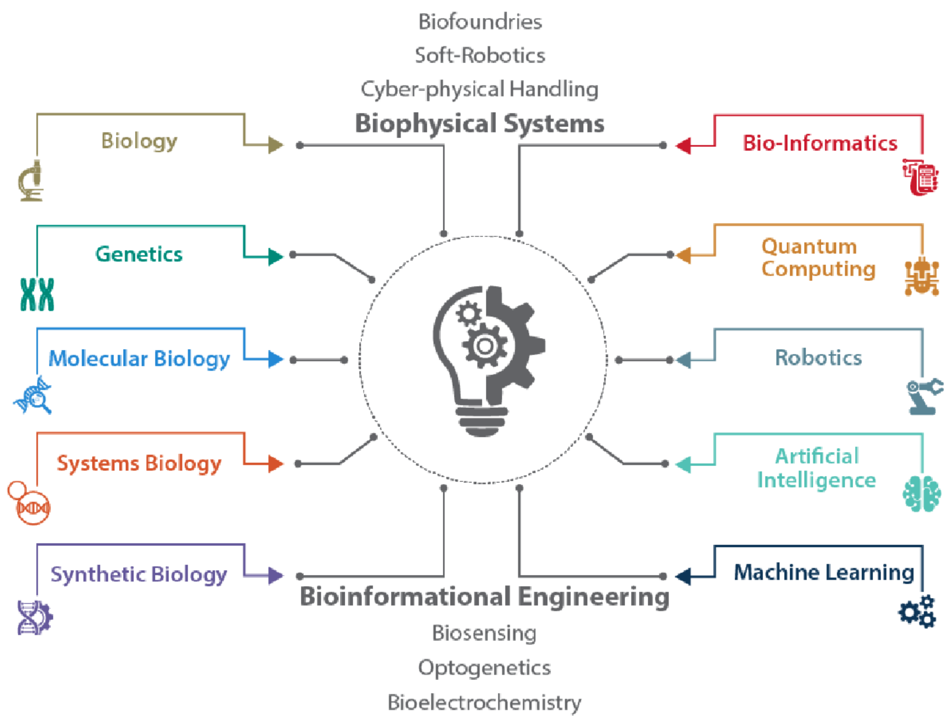
Biotech

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 take 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]
  - *AI*, 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]

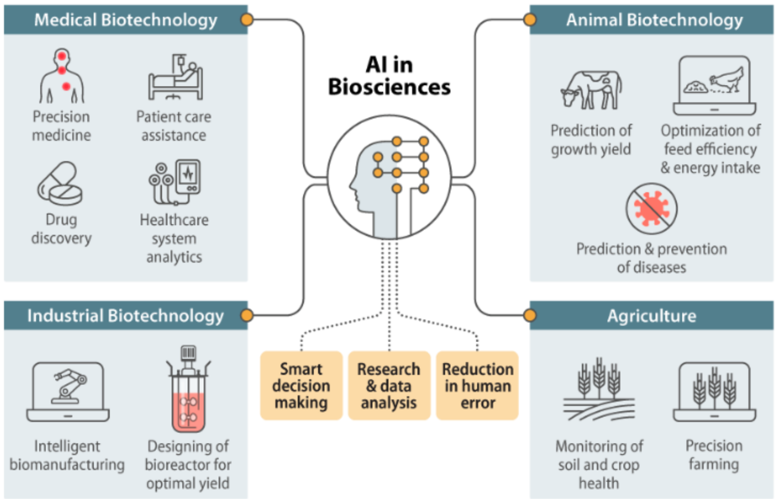


Convergence of AI and biological design

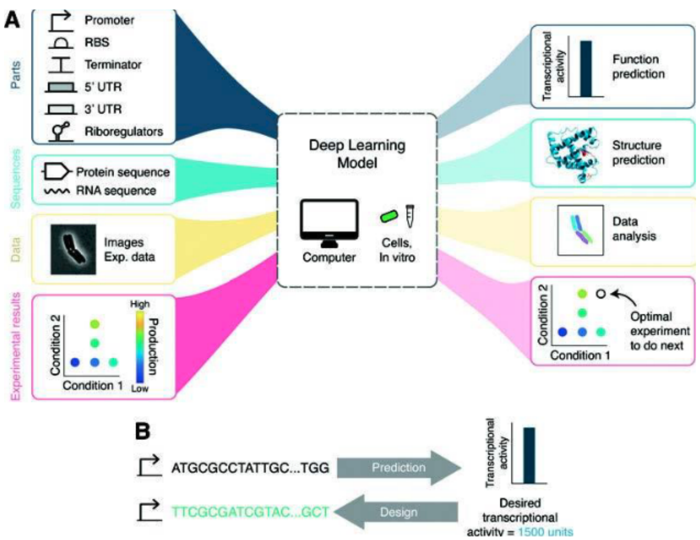
- AI & biological sciences converging [BKP22]
  - each building upon the other's capabilities for new research and development across multiple areas
- Demis Hassabis, CEO & cofounder of DeepMind, said of biology [Toe23]

  > ". . . 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



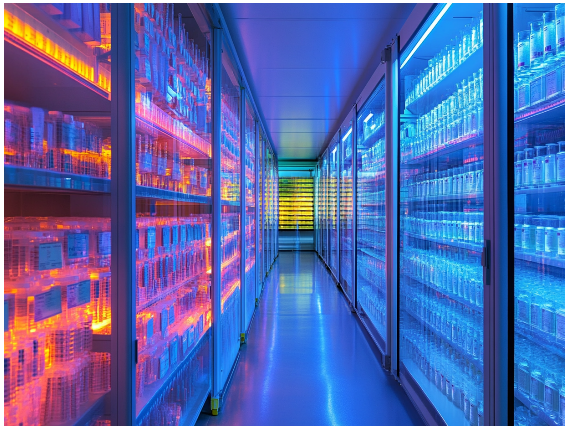
- AI, essential to analyzing exponential growth of genetic sequence data

  "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 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



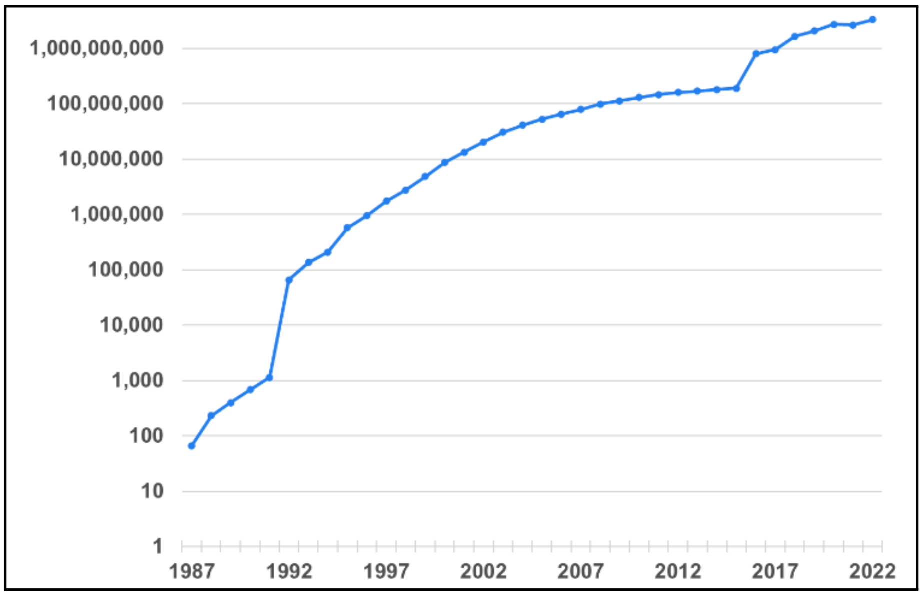
Rapid growth of biological data

- volume of genetic sequence data grown exponentially as sequencing technology 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 - more than *terabyte of three-dimensional structure data* for biological molecules, *e.g.*, proteins, DNA, RNA
  - proprietary database
    - Gingko Bioworks - 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

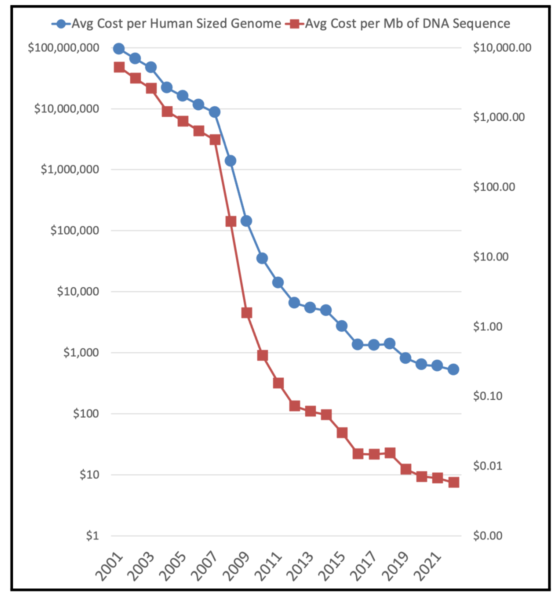
Volume and sequencing cost of DNA over time

- volume of DNA sequences & DNA sequencing cost
  - data source: National Human Genome Research Institute (NHGRI) [Wet23] & International Nucleotide Sequence Database Collaboration (INSDC)
- *more dramatic than Moore's law!*

\# sequences in INSDC



DNA sequencing cost



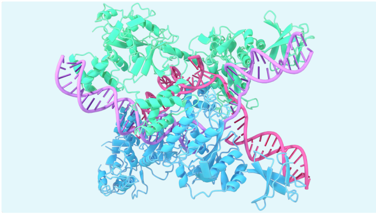
Bio data availability and bias

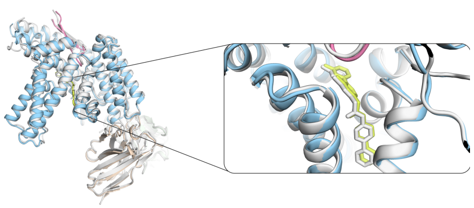
- US National Security Commission on Artificial Intelligence (NSCAI) recommends
  - US fund and prioritize development of a biobank containing *"wide range of high-quality biological and genetic data sets securely accessible by researchers"*
  - establishment of database of broad range of human, animal, 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

AlphaFold

- solving 50-year-old protein folding problem, *"one of biology's grand challenges"*
  - definition - given amino acid sequence, predict how it folds into a 3D structure
  - proteins fold in microseconds, but predicting computationally nearly impossible
- AlphaFold 1 (2018) - DL + physics-based energy functions → AlphaFold 2 (2020) - attention-based NN solving protein folding "in principle" → AlphaFold 3 (2024) - diffusion-based DL, drug-protein interactions, protein complexes
- AlphaFold protein structure database
  - >200MM protein structures - nearly every known protein, used by >2MM researchers
- Applications & implications
  - drug discovery - target identification, lead optimization, side effect prediction
  - enzyme engineering, agriculture, environmental, vaccine development



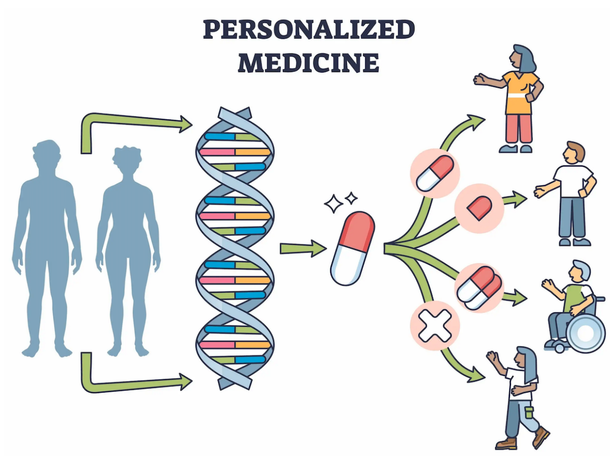


AlphaGo

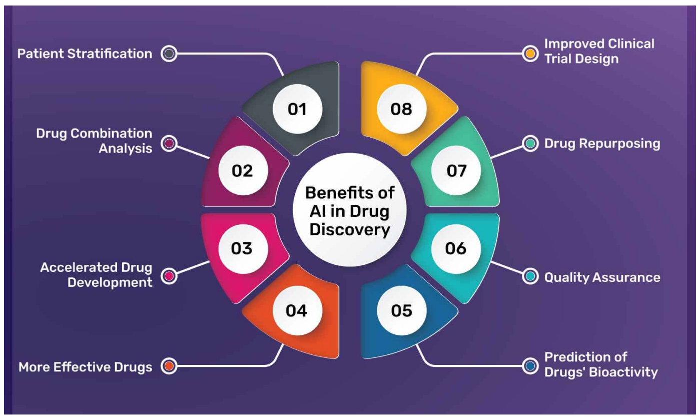
- Go board marked with 19×19 grid - $(19 \times 19)! = 361! \approx 1.44 \times 10^{768}$
  1437923258884890654832362511499863354754907538644755876127282765299227795534389618856841908003141196
  0714137944348905859683839682333043216077138088370565578796691924861827097800358990211005794501073330
  5079262777172275041226808677528136885057526541812043502150623466302643442673632627092764643302557772
  2695595343233942204301825548143785112222186834487969871267194205609533306413935710635197200721473378
  7338269803085351043174203653673779887217565513450041291061650506154496265581102824241428406627054585
  5623101563752892899924857388316647687165212001536218913733713768261861456295440900774337589490771443
  9917299937133680728459000034496420337066440853337001284286412654394495050773954560000000000000000000
  000000000000000000000000000000000000000000000000000000000000000000000
- deep reinforcement learning with Monte Carlo tree search
  - trained on thousands of years of Go game history
  - AlphaGo Zero learns by playing against itself
- development experience, insight, knowledge, know-how transferred to AlphaFold

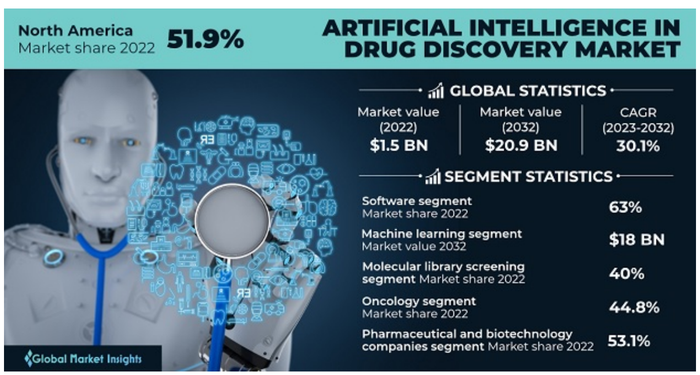
Personalized medicine

- *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

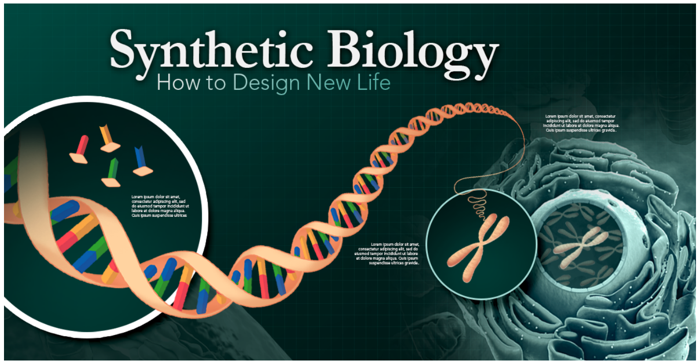


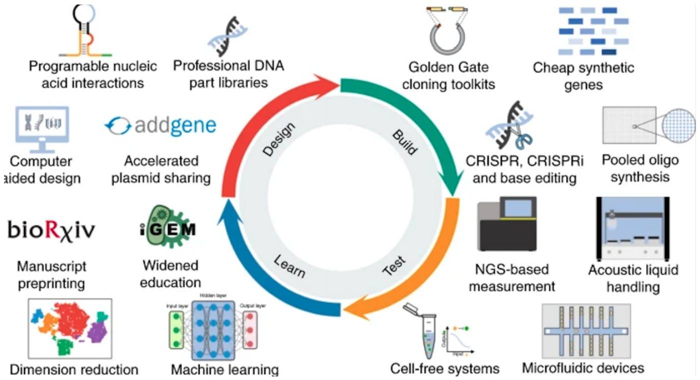


- traditional drug discovery process - time-consuming and costly often taking decades and billions of dollars
- AI streamlines this process by predicting the efficacy and safety of potential compounds with more speed and accuracy
- AI models analyze chemical databases to identify new drug candidates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

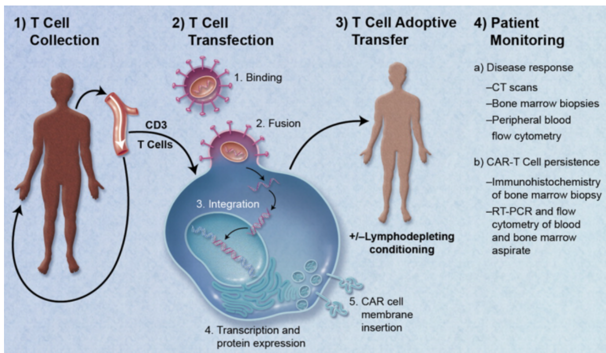
Synthetic biology

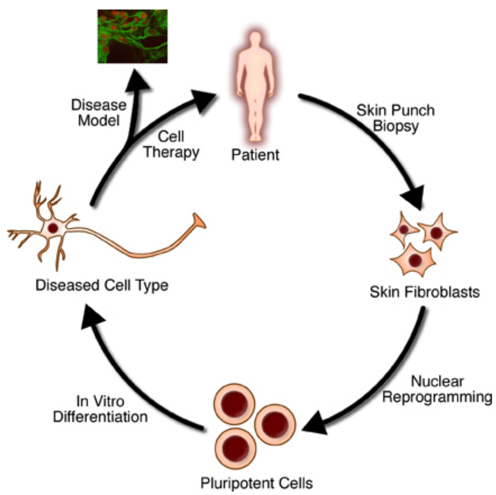
- use AI for gene editing, biomaterial production and synthetic pathways
- combine principles of biology and engineering to design and construct new biological entities
- AI optimizes synthetic biology processes from designing genetic circuits to scaling up production
- company - Ginkgo Bioworks uses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





Regenerative medicine

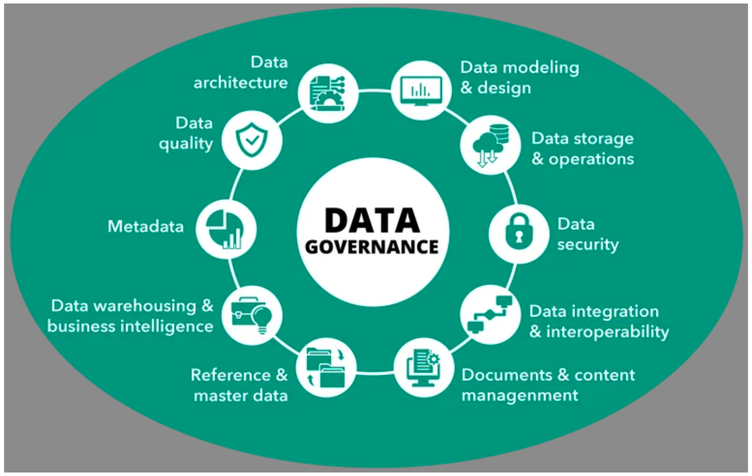




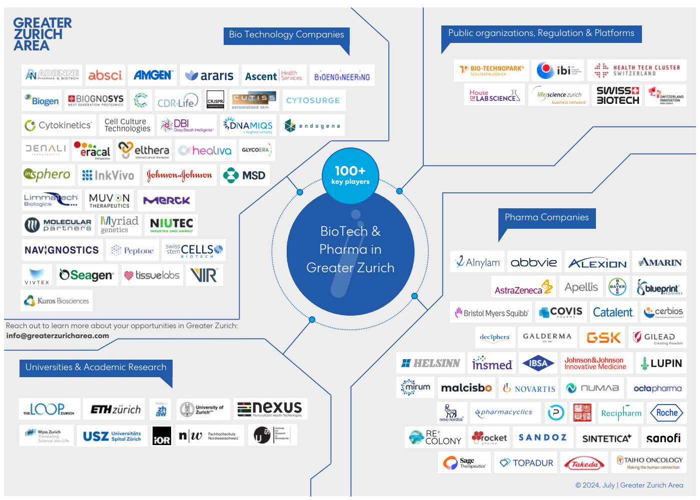
- AI advances development of stem cell therapies & tissue engineering
- AI algorithms assist in identifying optimal cell types, predicting cell behavior & personalized treatments
- particularly for conditions such as neurodegenerative diseases, heart failure and orthopedic injuries
- company - Organovo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

Selected References & Sources

Selected references & sources

- Robert H. Kane "Quest for Meaning: Values, Ethics, and the Modern Experience" 2013
- Michael J. Sandel "Justice: What's the Right Thing to Do?" 2009
- Daniel Kahneman "Thinking, Fast and Slow" 2011
- Yuval Noah Harari "Sapiens: A Brief History of Humankind" 2014
- M. Shanahan "Talking About Large Language Models" 2022
- A.Y. Halevry, P. Norvig, and F. Pereira "Unreasonable Effectiveness of Data" 2009
- A. Vaswani, et al. "Attention is all you need" @ NeurIPS 2017
- S. Yin, et. al. "A Survey on Multimodal LLMs" 2023
- Chris Miller "Chip War: The Fight for the World's Most Critical Technology" 2022

- CEOs, CTOs, CFOs, COOs, CMOs & CCOs @ startup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California, USA

References

References

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

[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.

[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.

[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.

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

[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.

Thank You