[KAIST Mechanical Engineering Department AI Seminar]
AI in Action - Biotechnology and Industrial Applications

Sunghee Yun

Co-founder & CTO - AI Technology & Biz Dev @ Erudio Bio, Inc.

About Speaker

- *Co-founder & CTO @ Erudio Bio, San Jose & Novato, CA, USA*
- Advisor & Evangelist @ CryptoLab, Inc., San Jose, CA, USA
- Chief Business Development Officer @ WeStory.ai, Cupertino, CA, USA
- Advisory Professor, Electrical Engineering and Computer Science @ DGIST, Korea
- Adjunct Professor, Electronic Engineering Department @ Sogang University, Korea
- Global Advisory Board Member @ Innovative Future Brain-Inspired Intelligence System Semiconductor of Sogang University, Korea
- *KFAS-Salzburg Global Leadership Initiative Fellow @ Salzburg Global Seminar*, Salzburg, Austria
- Technology Consultant @ Gerson Lehrman Gruop (GLG), NY, 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 @ Mobile Shopping Team, Amazon.com, Inc., Vancouver, BC, Canada – 2020
- Principal Engineer @ Software R&D Center of DS Division, Samsung, Korea – 2017
- Principal Engineer @ Strategic Marketing & Sales Team, Samsung, Korea – 2016
- Principal Engineer @ DT Team of DRAM Development Lab, Samsung, Korea – 2015
- Senior Engineer @ CAE Team - Samsung, Korea – 2012
- MS & PhD - Electrical Engineering @ Stanford University, CA, USA – 2004
- Development Engineer @ Voyan, Santa Clara, CA, USA – 2001
- BS - Electrical Engineering @ Seoul National University, Seoul, Korea – 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 - time-series anomaly detection, deep reinforcement learning & recommender system
  - Jeff Bezos's project - increase sales by *$200M* via Amazon Mobile Shopping App
- Co-founder & CTO & Head of Global R&D & Chief Applied Scientist & Senior Fellow @ Gauss Labs, Inc.
- Co-founder & CTO - AI Technology & Business Development @ Erudio Bio, Inc.

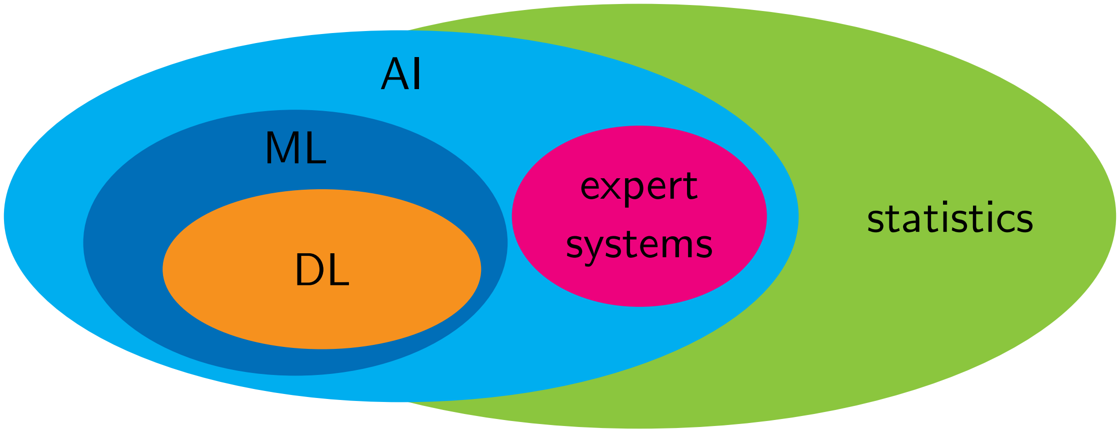
Today

Artificial Intelligence

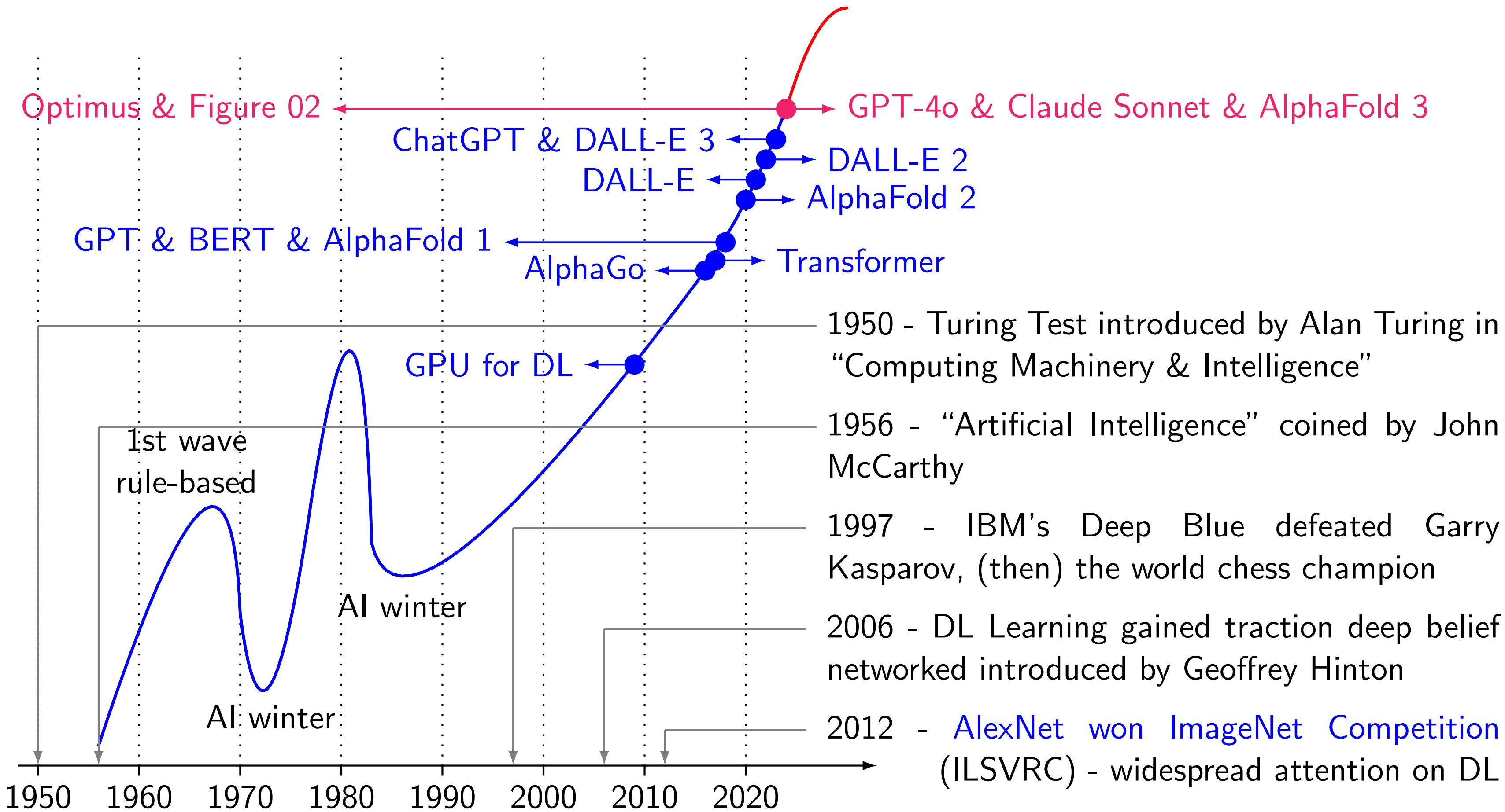
Definition and History

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

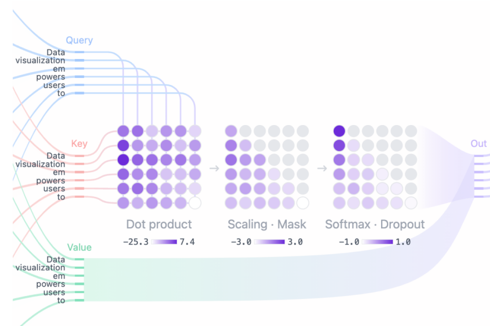
Deep learning revolution

- 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

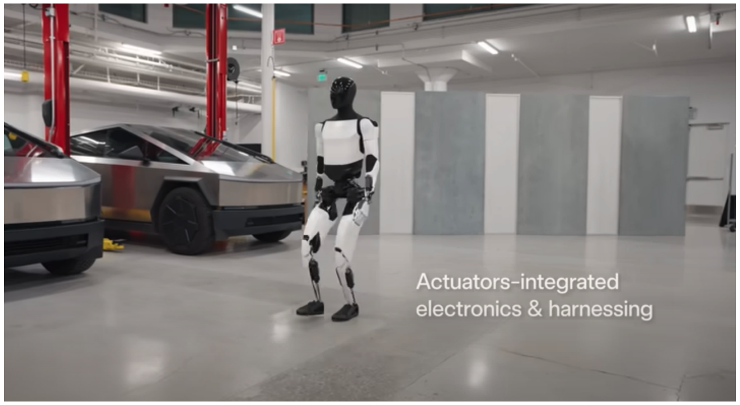
- 2017 – 2018 - Transformers & NLP breakthroughs[2]
  - *Transformer (e.g., BERT & GPT) revolutionized NLP*
  - major advancements in, *e.g.*, machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting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

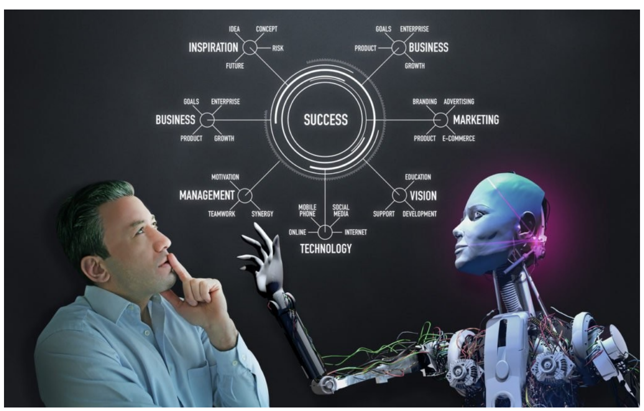
Lots of breakthroughs in AI technology and applications in 2024

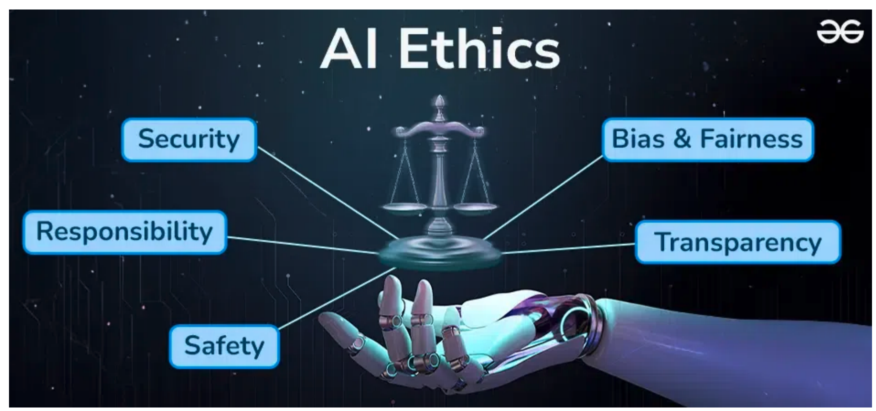
- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Llama 3, Sora
  - *transforming industries* such as content creation, customer service, education, *etc.*
- breakthroughs in specialized AI applications
  - Figure 02, Optimus, AlphaFold 3
  - driving unprecedented advancements in automation,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

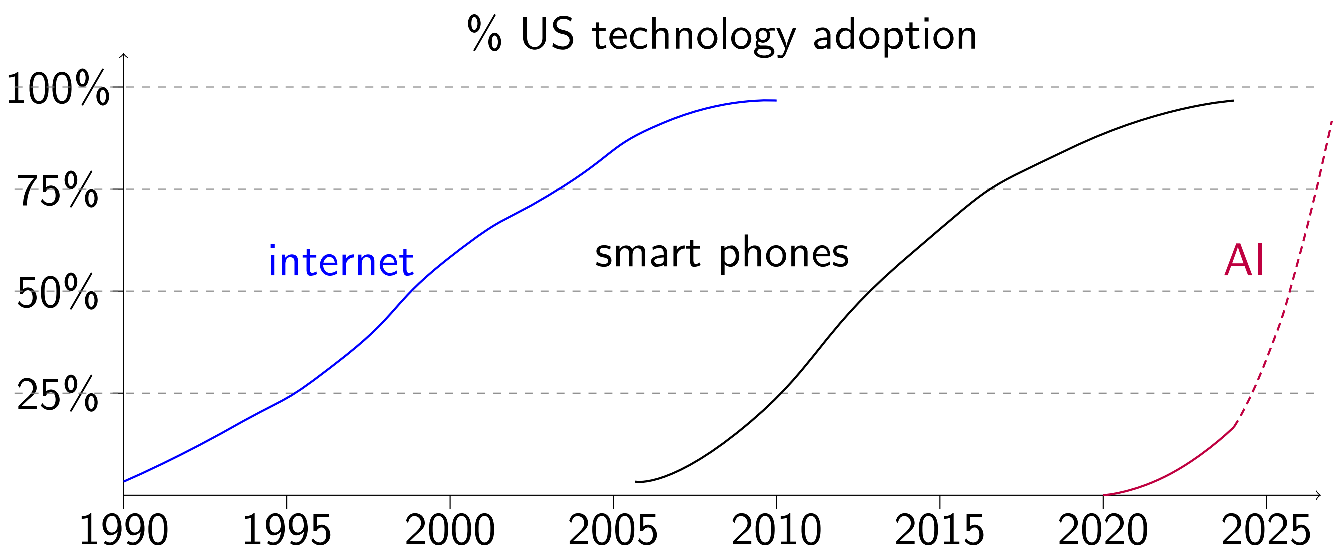




Recent Advances in AI

Where are we in AI today?

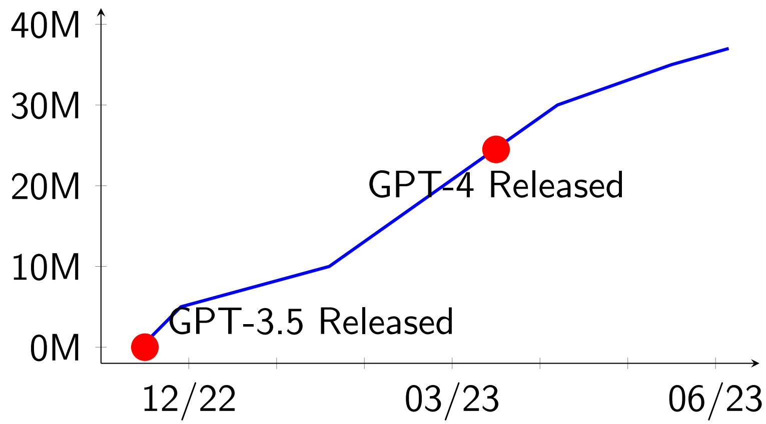
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



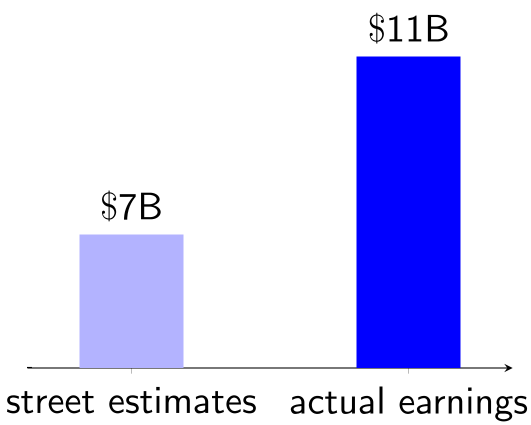
Explosion of AI ecosystems - ChatGPT & NVIDIA

- 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]

ChatGPT weekly unique visits



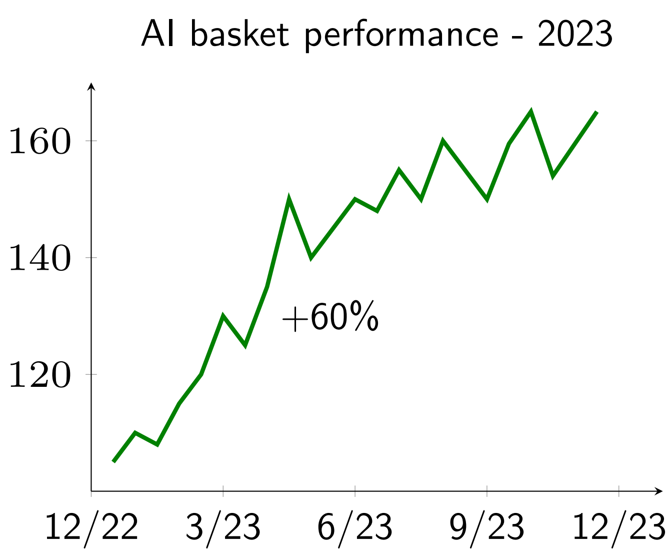
NVDA 2023-2Q earning

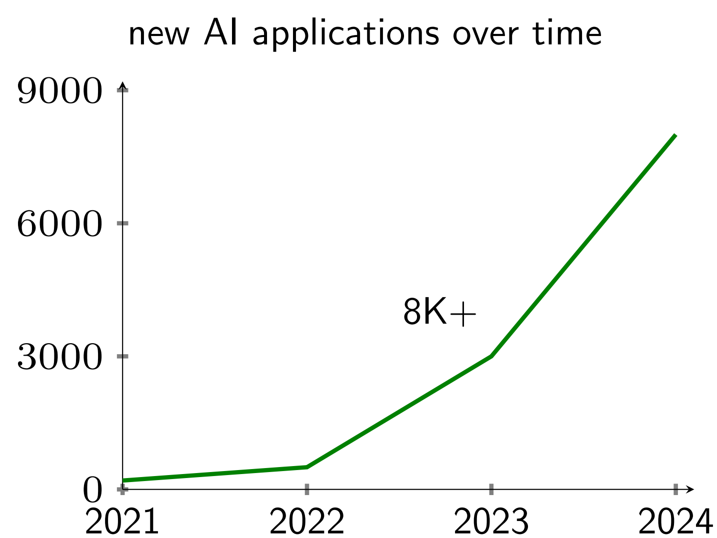


[3] source - Bloomberg

Explosion of AI ecosystems - AI stock market

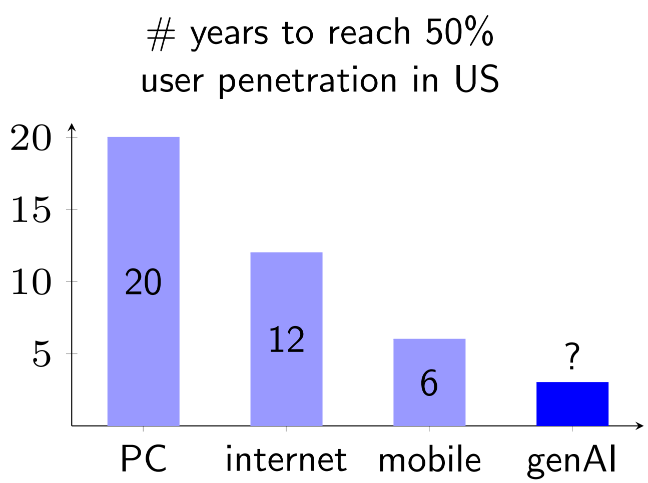
- *AI investment surge in 2023 - portfolio performance soars by 60%*
  - AI-focused stocks significantly outpaced traditional market indices
- *over 8,000 new AI applications* developed in last 3 years
  - applications span from healthcare and finance to manufacturing and entertainment

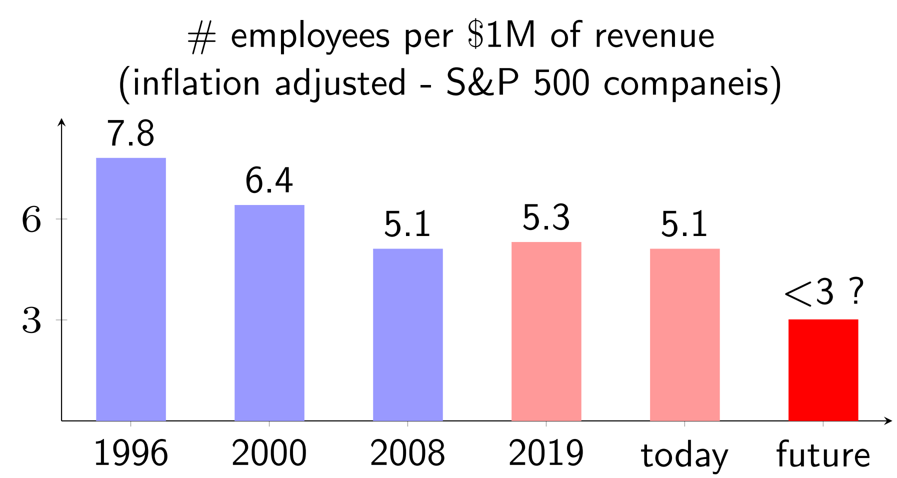




AI’s transformative impact - adoption speed & economic potential

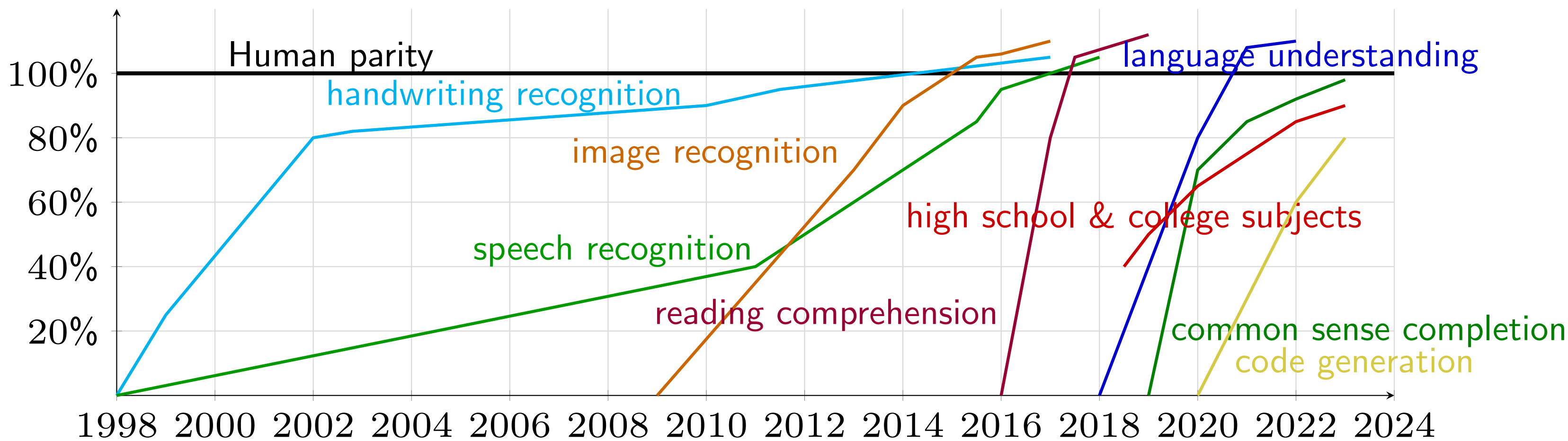
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI’s potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater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





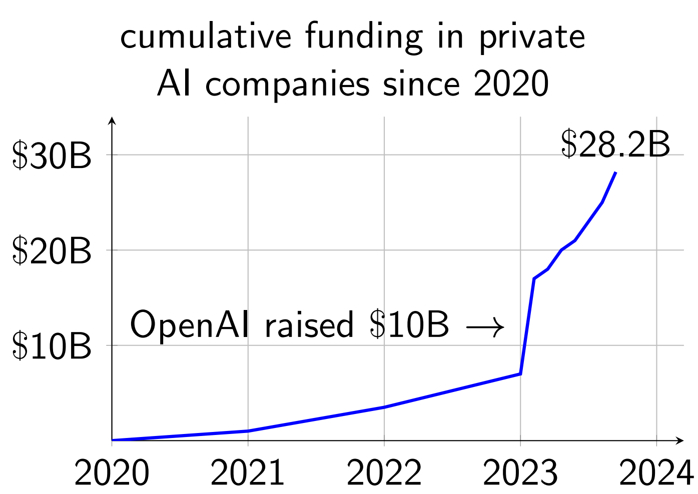
AI getting more & more faster

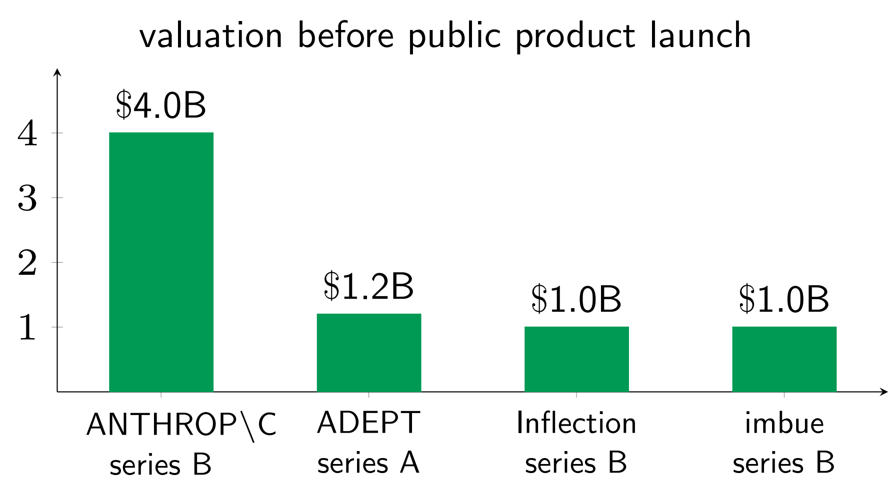
- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached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



Massive investment in AI

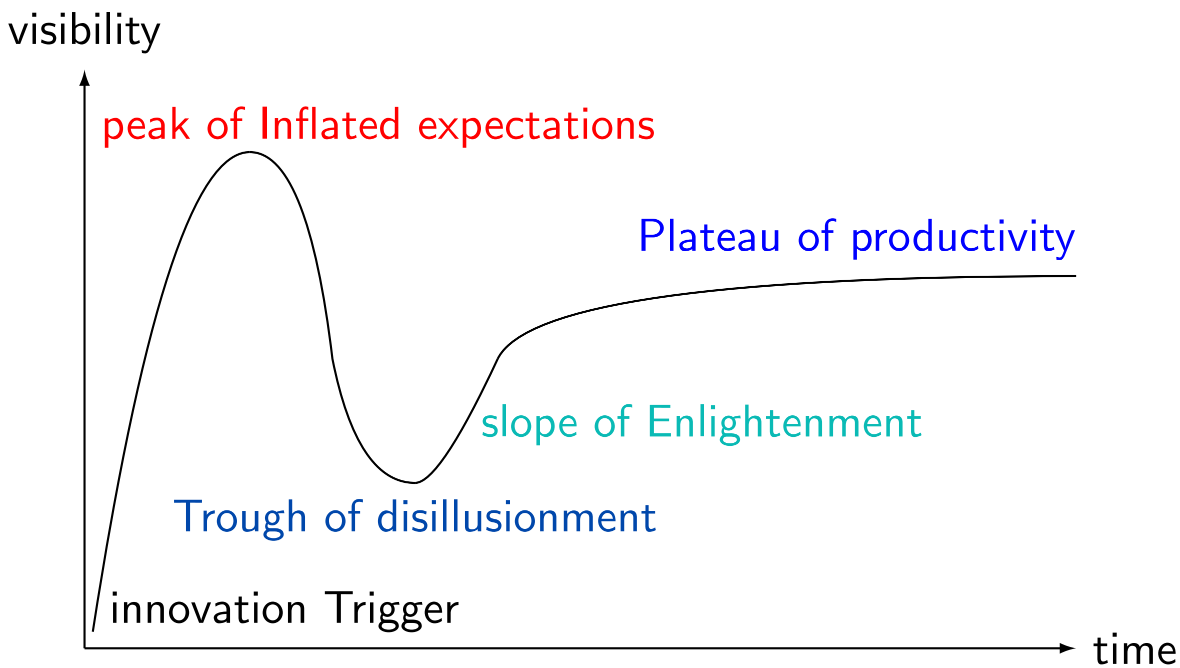
- *explosive growth* - cumulative funding skyrocketed reaching staggering $28.2B
- OpenAI - significant fundraising (=$10B) fueled rapid growth
- *valuation surge* - substantial valuations even before public products for stella companies
- *fierce competition for capital* among AI startups driving innovation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





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

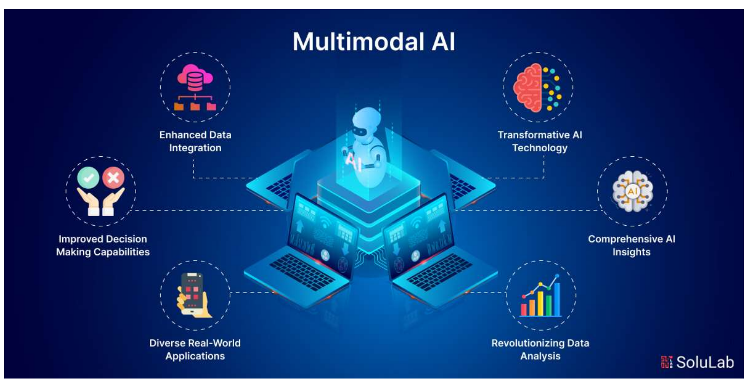
Yes & No

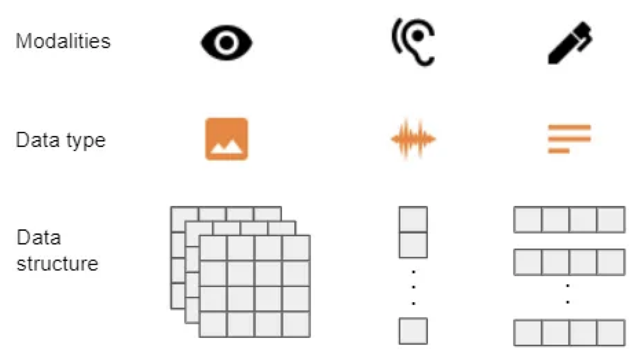
characteristics of hype cycles	speaker's views
value accrual misaligned with investment	• OpenAI still operating at a loss; business model *still* not clear • gradual value creation across broad range of industries and technologies (*e.g.*, CV, LLMs, RL) unlike fiber optic bubble in 1990s
overestimating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

Multimodal AI Agents

Multimodal learning

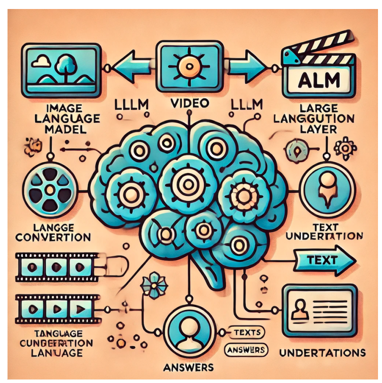
- understand information from multiple modalities, *e.g.*, text, images, audio, video
- representation learning methods
  - combine two representations or learn multimodal representations simultaneously
- applications
  - images from text prompt, videos with narration, musics with lyrics
- collaboration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





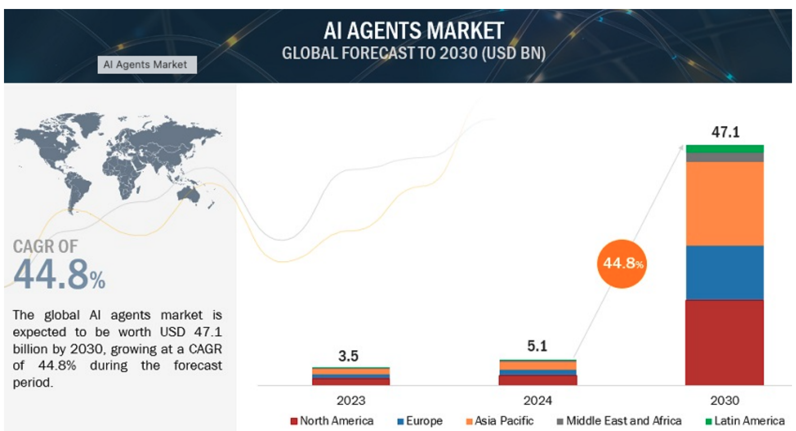
Implications of success of LLMs

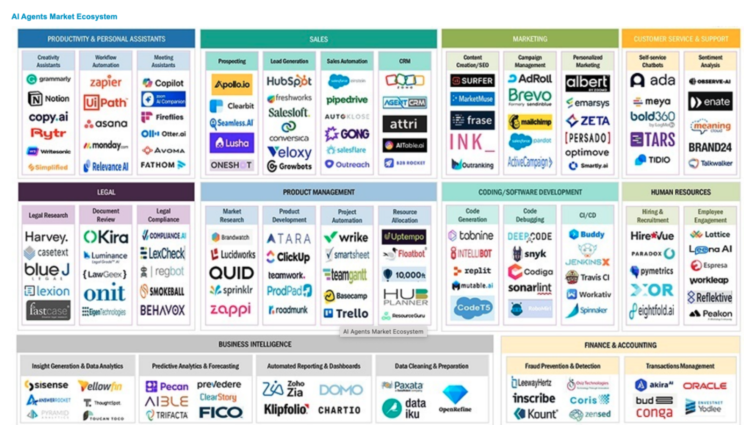
- many researchers change gears towards LLM
  - from computer vision (CV), speach, music, video, even reinforcement learning
- *LLM is not only about NLP . . .* humans have . . .
  - evolved and optimized natural language structures for eons
  - handed down knowledge using this natural languages for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



Multimodal AI (mmAI) - definition & history

- 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



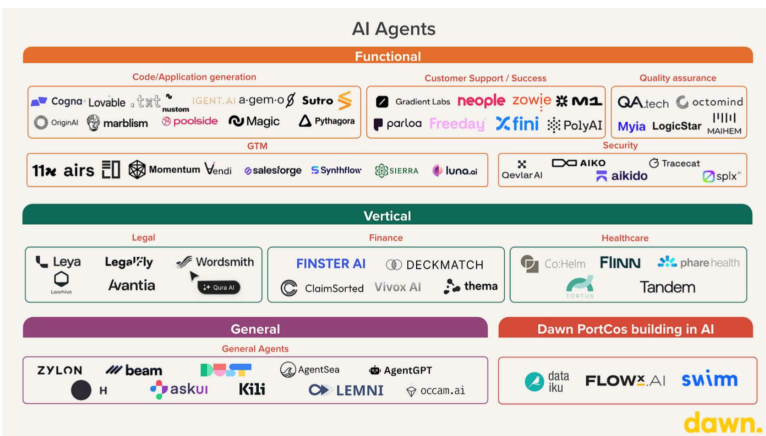


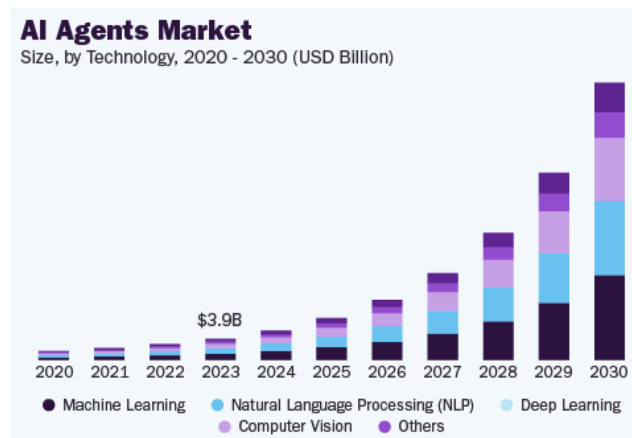
mmAI Technology

- 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

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.

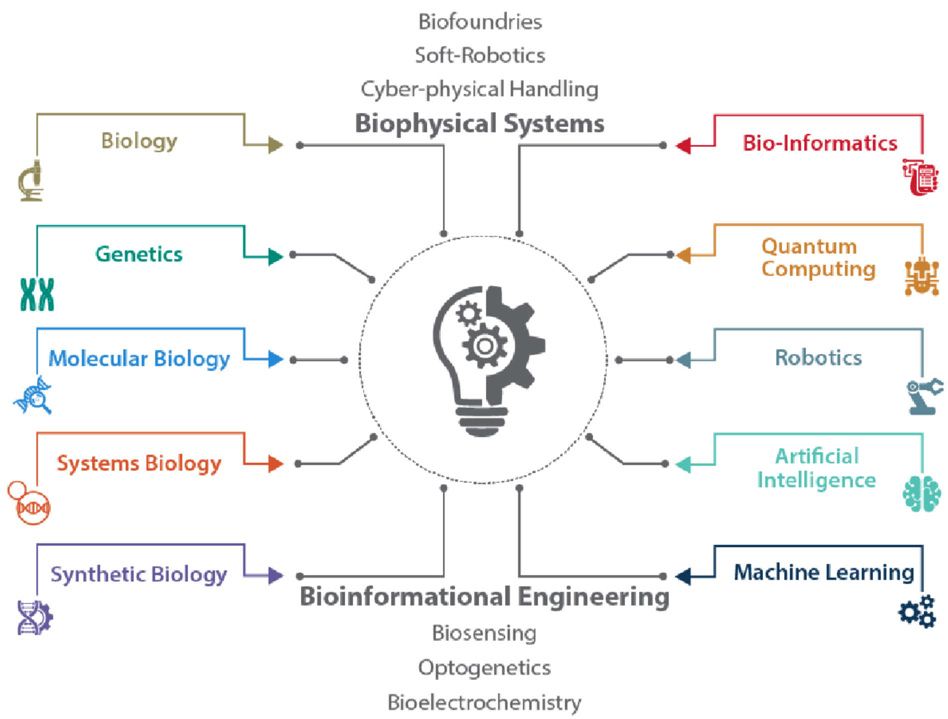
Biotech

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]

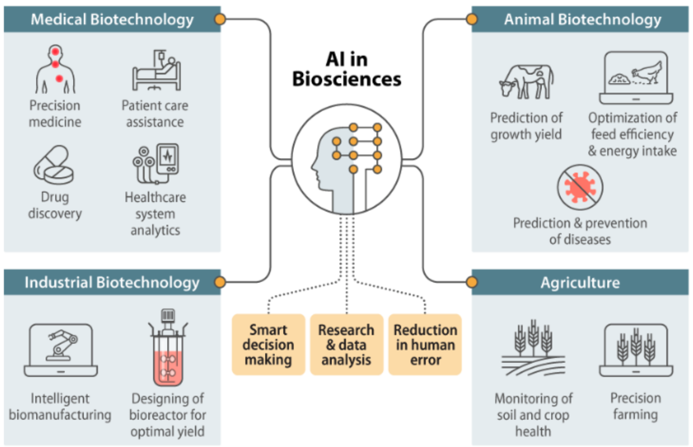


Convergence of AI and biological design

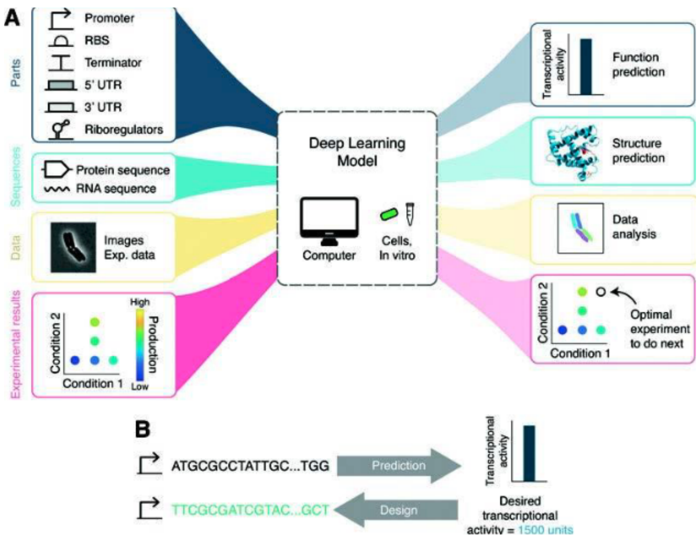
- both AI & biological sciences increasingly converging [BKP22]
  - each building upon the other's capabilities for new research and development across multiple areas
- Demo Hassabis, CEO & cofounder of DeepMind, said of biology [Toe23]

  "... 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

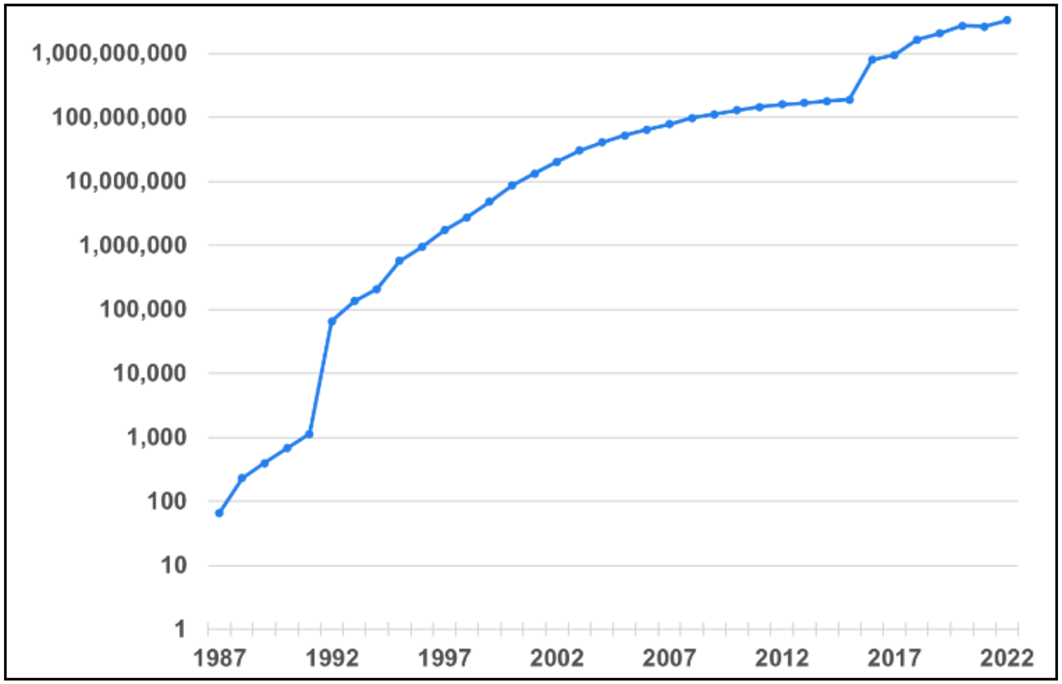
Rapid growth of biological data

- 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

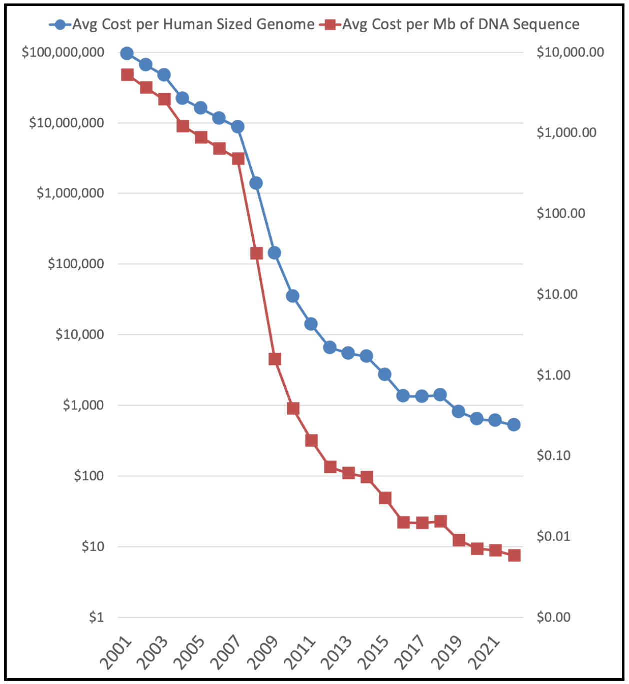
Volume and sequencing cost of DNA over time

- 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



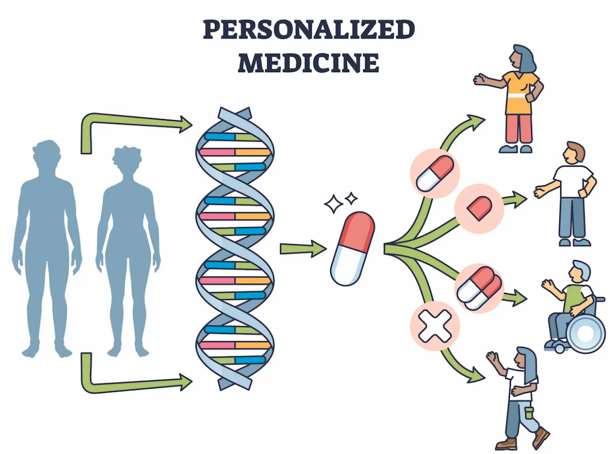
Bio data availability and bias

- 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

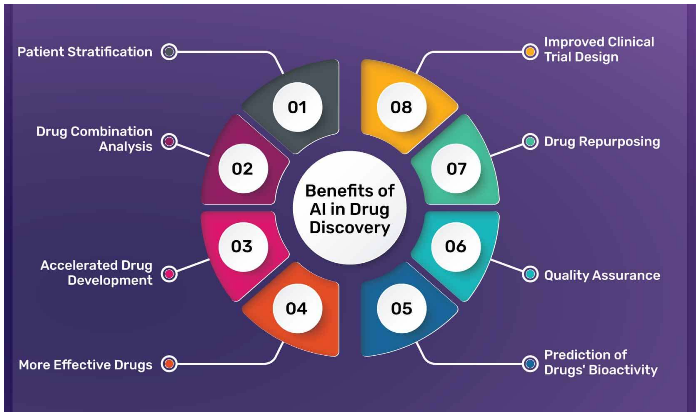
Emerging Trends in Biotech

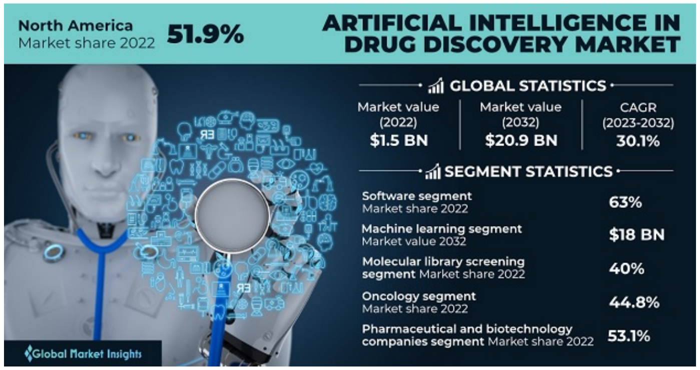
Personalized medicine

- *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

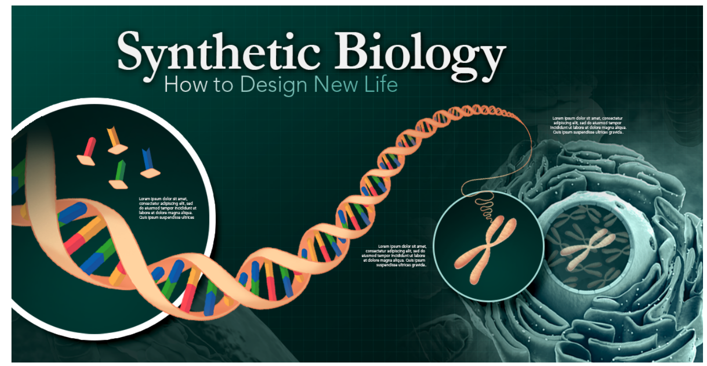


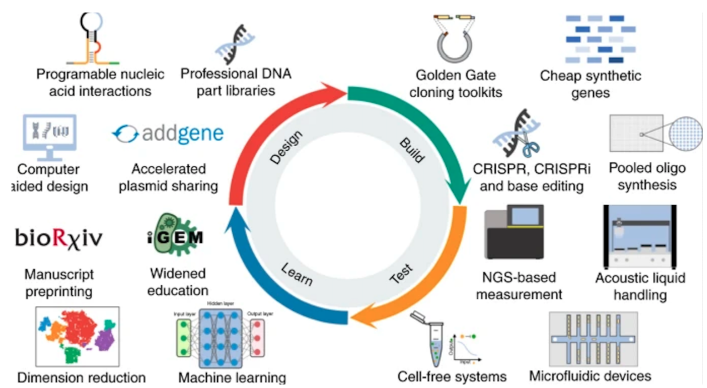


- traditional drug discovery process - time-consuming and costly often taking decades and billions of dollars
- AI streamlines this process by predicting the efficacy and safety of potential compounds with more speed and accuracy
- AI models analyze chemical databases to identify new drug candidates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

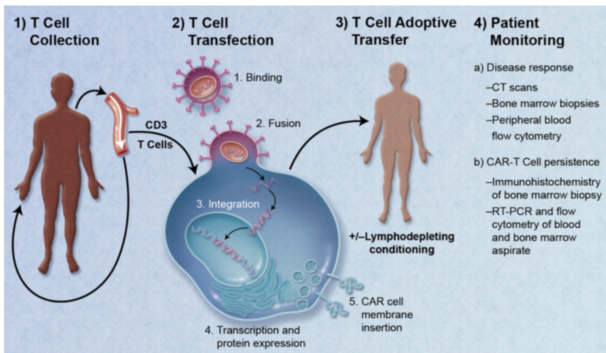
Synthetic biology

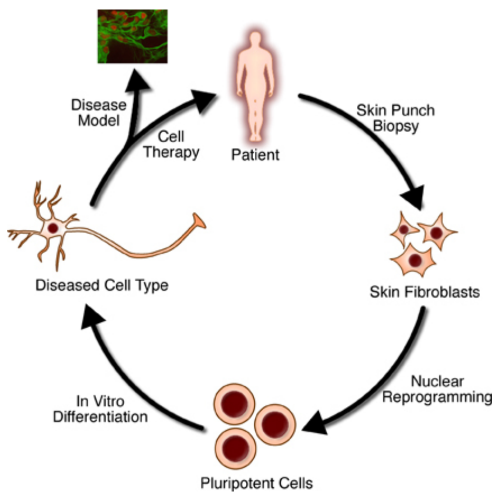
- use AI for gene editing, biomaterial production and synthetic pathways
- combine principles of biology and engineering to design and construct new biological entities
- AI optimizes synthetic biology processes from designing genetic circuits to scaling up production
- company - Ginkgo Bioworks uses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





Regenerative medicine

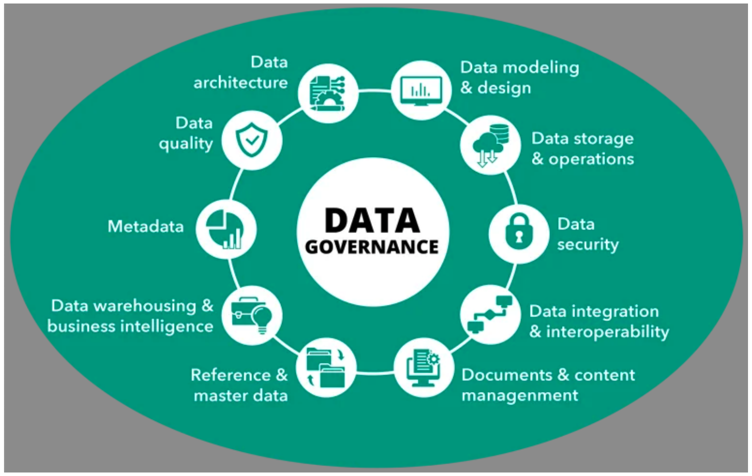




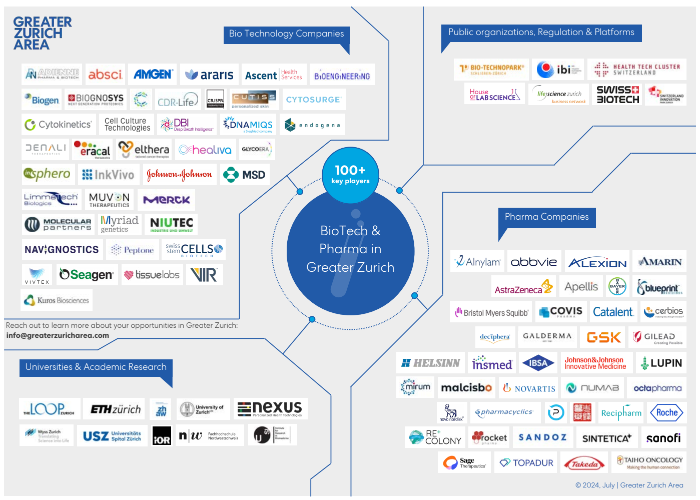
- AI advances development of stem cell therapies & tissue engineering
- AI algorithms assist in identifying optimal cell types, predicting cell behavior & personalized treatments
- particularly for conditions such as neurodegenerative diseases, heart failure and orthopedic injuries
- company - Organovo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

Industrial AI

Industrial AI (inAI)

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

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

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

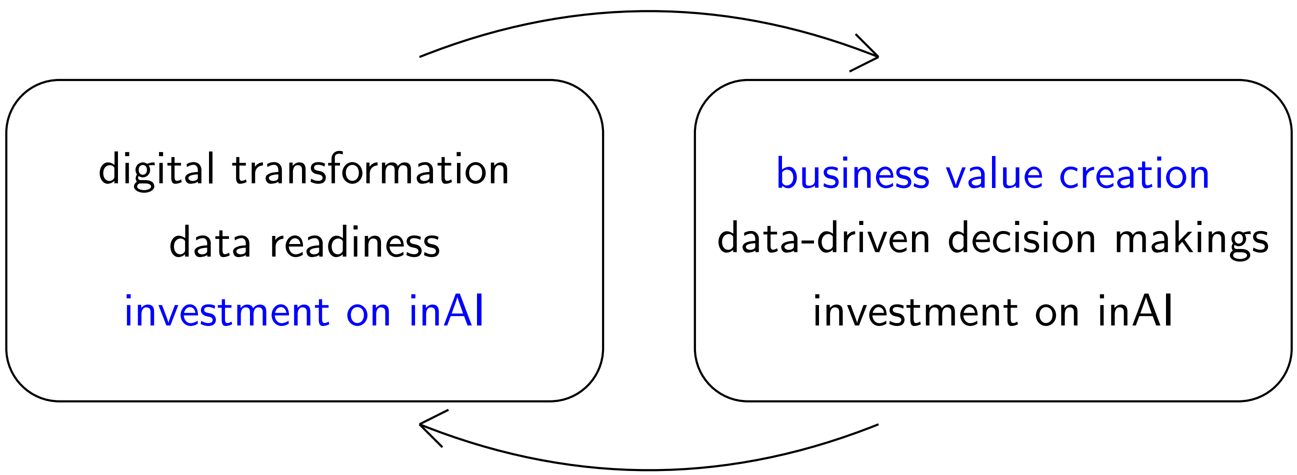
inAI fields

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

Characteristics of inAI

Vicious (or virtuous) cycle

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

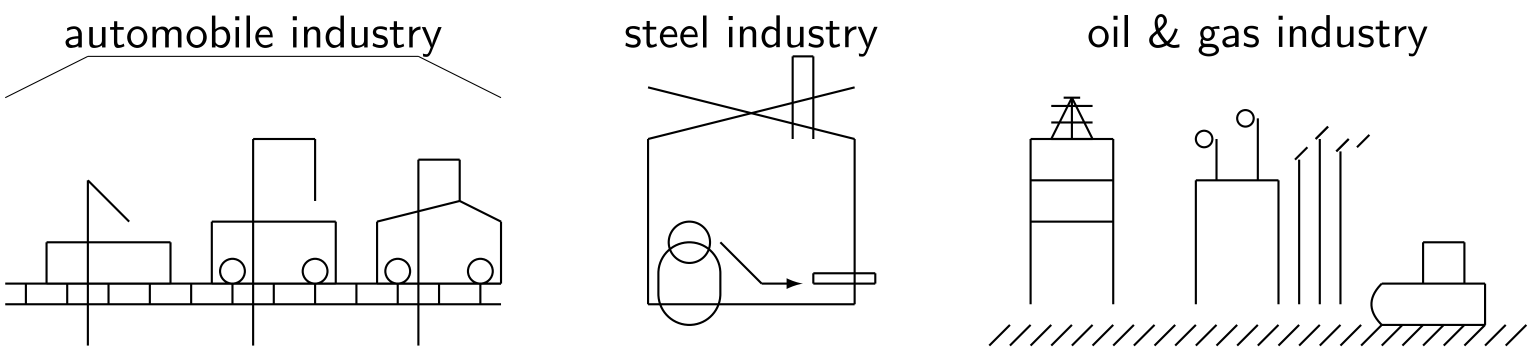


Data-centric AI

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

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

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



Challenging data characteristics

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

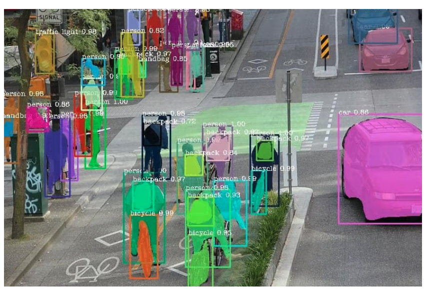
Manufacturing AI

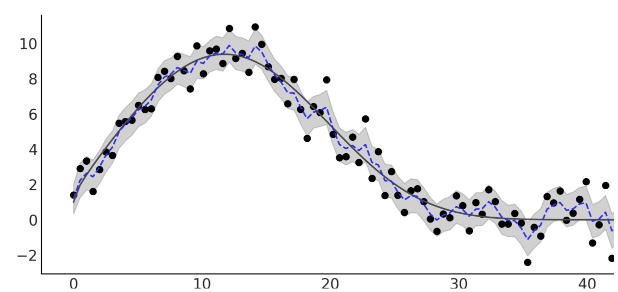
MLs in manufacturing AI (manAI)

- *image data* - huge amount of image data measured and inspected
  - SEM/TEM images, wafer defect maps, test failure pattern maps [4]

  → semantic segmentation, defect inspection, anomaly detection
- *time-series (TS) data* - *all the data* coming out of manufacturing is TS
  - equipment sensor data, process times, various measurements, MES data [5]

  → regression, anomaly detection, semi-supervised learning, Bayesian inference





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

[5] MES: manufacturing execution system

CV ML in manAI

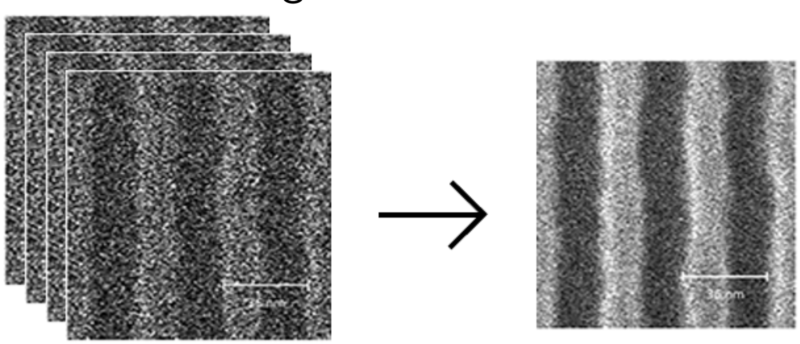
Computer vision ML in manAI

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

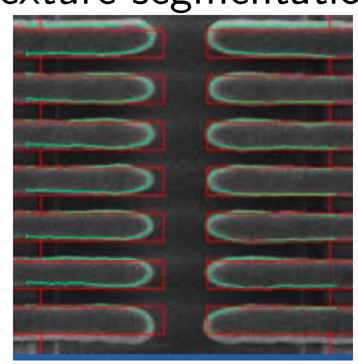
Automatic feature measurement

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

image enhancement



texture segmentation



auto-measurement

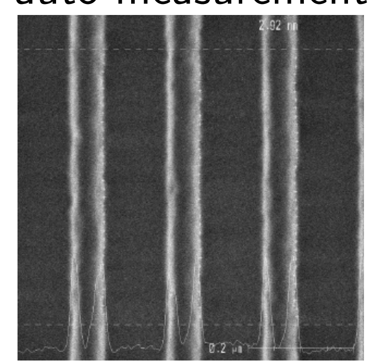
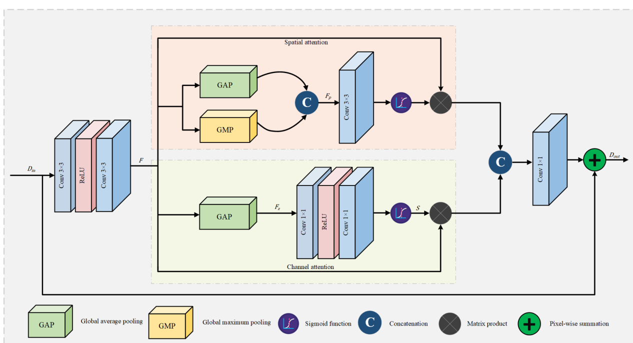


Image enhancement

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



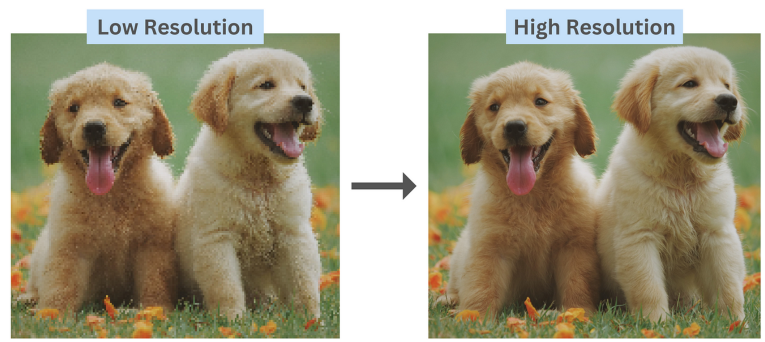
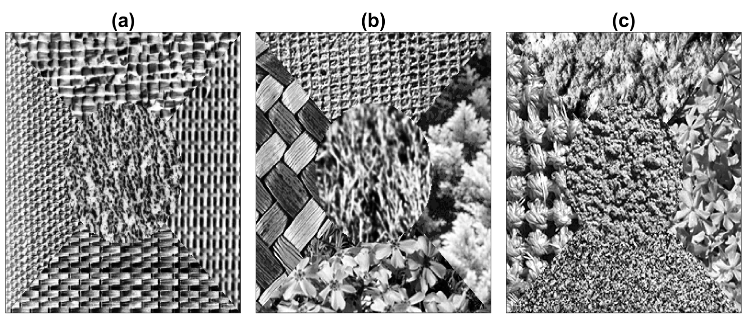


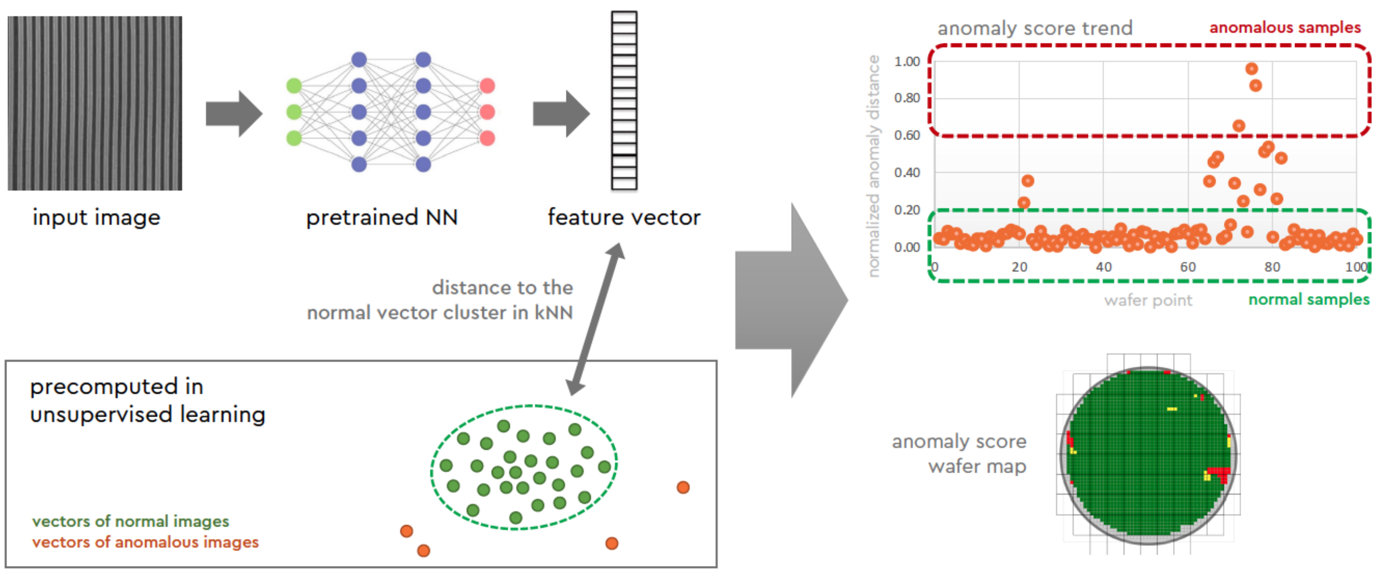
Image segmentation

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



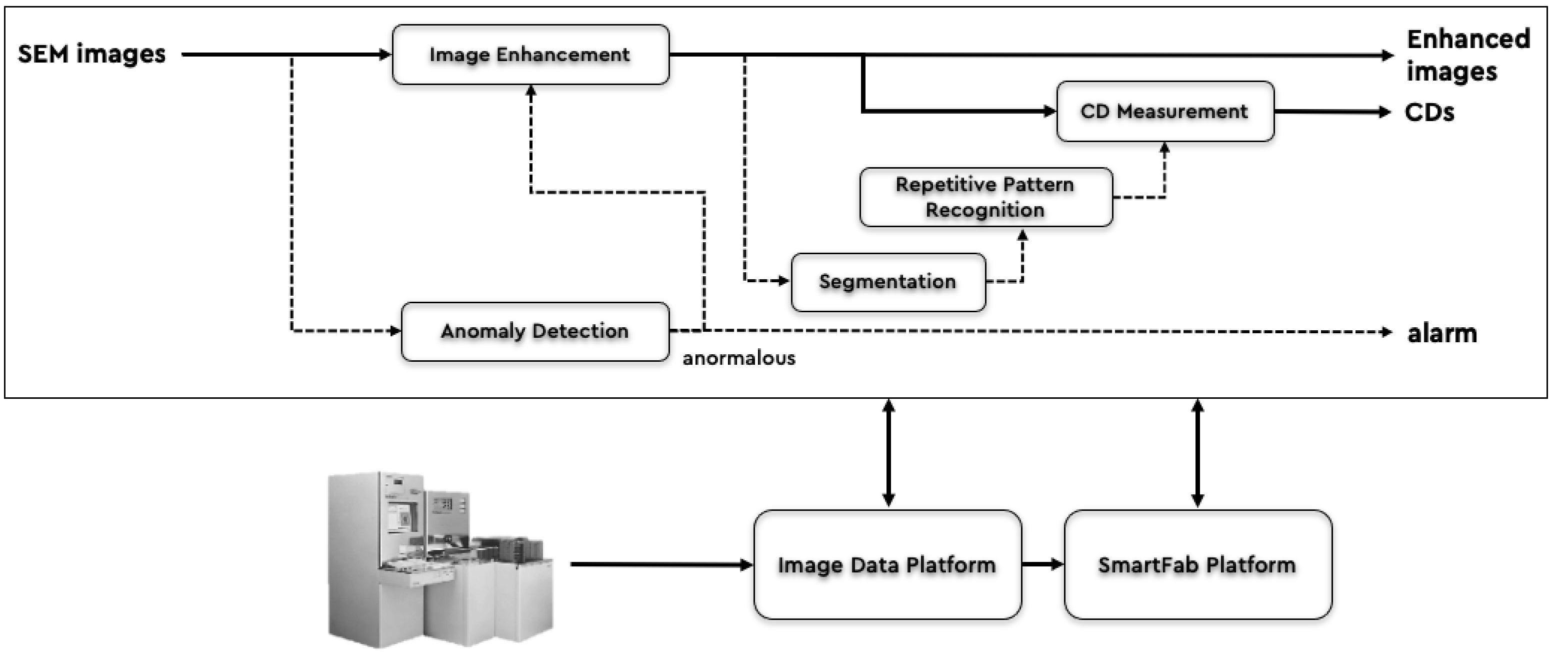
Anomaly detection using side product

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



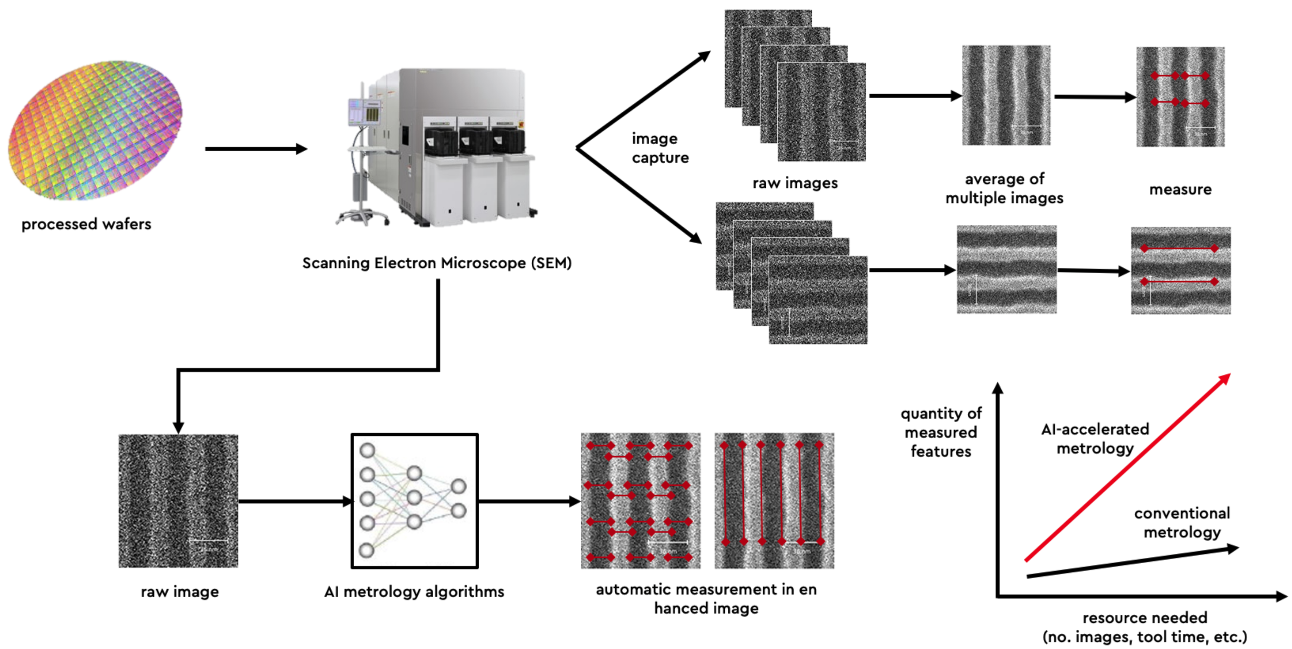
AI-enabled metrology system

- integration of separate components creates AI-enabled metrology system



Benefits of new system

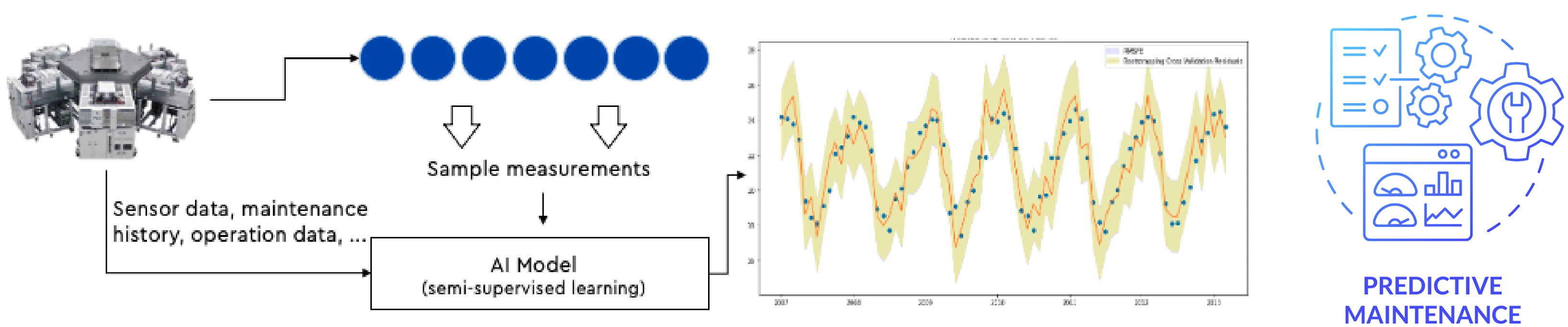
- new system provides
  - improved accuracy and reliability
  - improved throughput
  - savings on investment on measurement equipment



TS ML in manAI

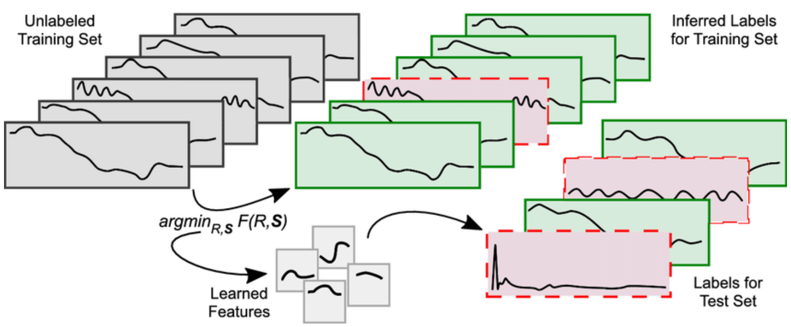
Time-series ML applications in manAI

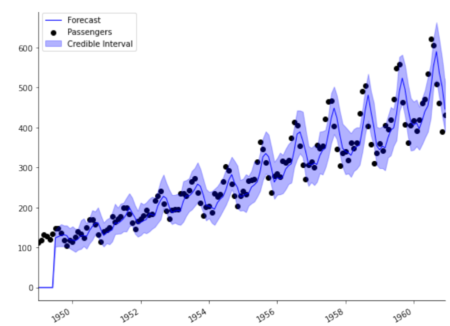
- estimation of TS values
  - virtual metrology - estimate measurement without physically measuring things
- anomaly detection on TS
  - predictive maintenance - predict maintenance times ahead
- multi-modal ML using LLM & genAI
  - root cause analysis and recommendation system



TS MLs in manAI

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



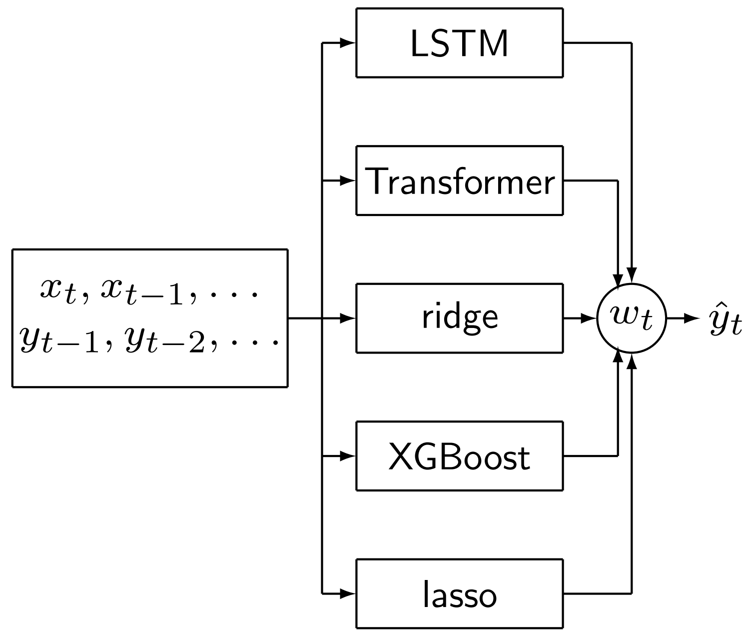


Difficulties with TS ML

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

Online learning for TS regression

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



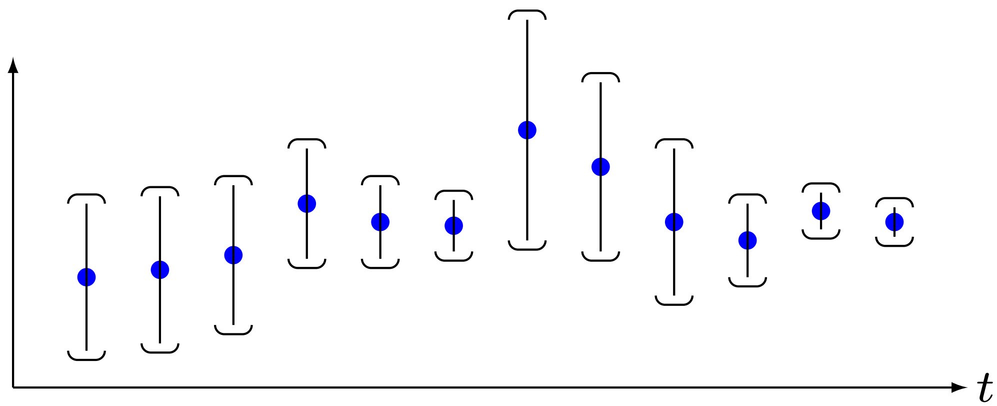
$$g_k = w_{1,k} f_{1,k} + w_{2,k} f_{2,k} + \cdots + w_{p_k,k} f_{p_k,k} = \sum_{i=1}^{p_k} w_{i,k} f_{i,k}$$

Credibility intervals

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

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

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



Bayesian approach for credibility interval evaluation

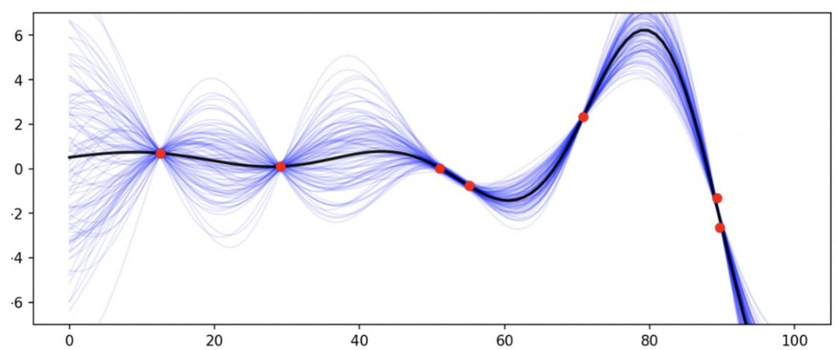
- assume conditional distribution ith predictor parameterized by $\theta_{i,k} \in \Theta$

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

  - depends on prior & current input, *i.e.*, $\theta_{i,k}$ & x_{t_k}

- update $\theta_{i,k+1}$ from $\theta_{i,k}$ after observing true $y(t_k)$ using Bayesian rule

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



Virtual Metrology

VM

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

VM - problem formulation

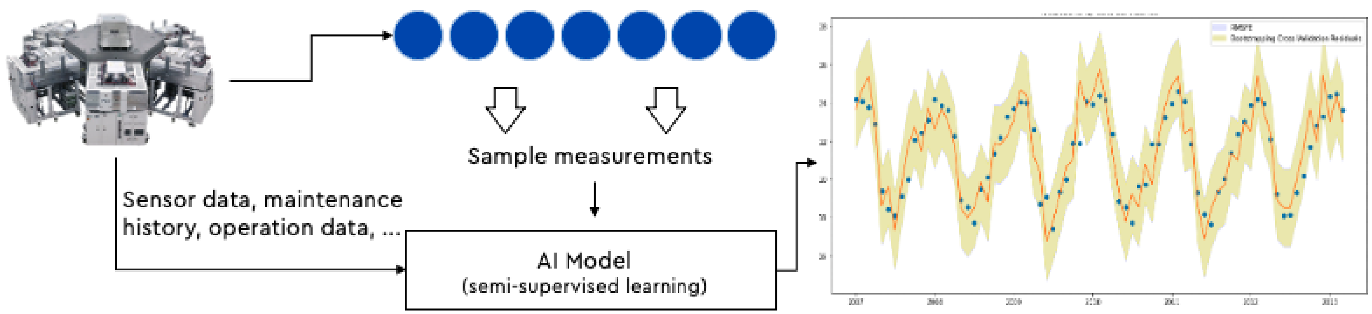
- problem description

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

- our problem formulation

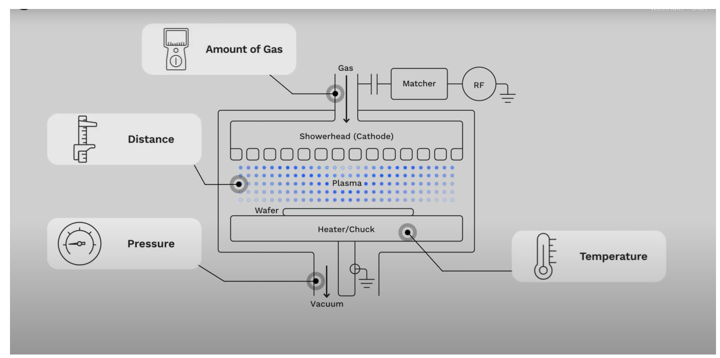
$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{K} w_{k,K-k}\, l(y_{t_k}, \hat{y}_{t_k}) \\ \text{subject to} & \hat{y}_{t_k} = g_k(x_{t_k}, x_{t_{k-1}}, \ldots, y_{t_{k-1}}, y_{t_{k-2}}, \ldots) \end{array}$$

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



VM - Gauss Labs' inAI success story

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



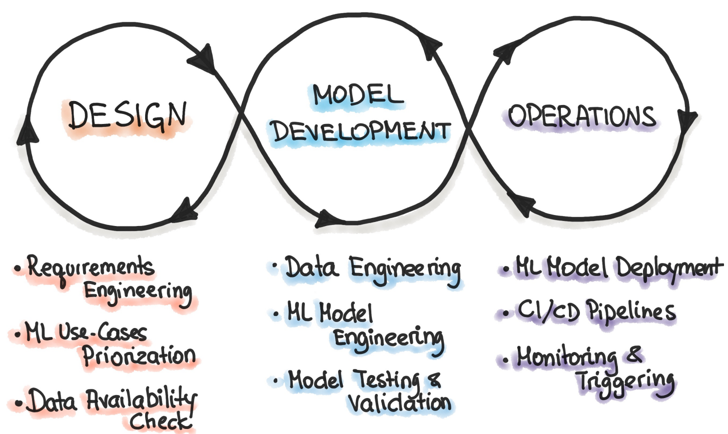
Manufacturing AI Productionization

Minimally required efforts for manAI

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

MLOps for manAI

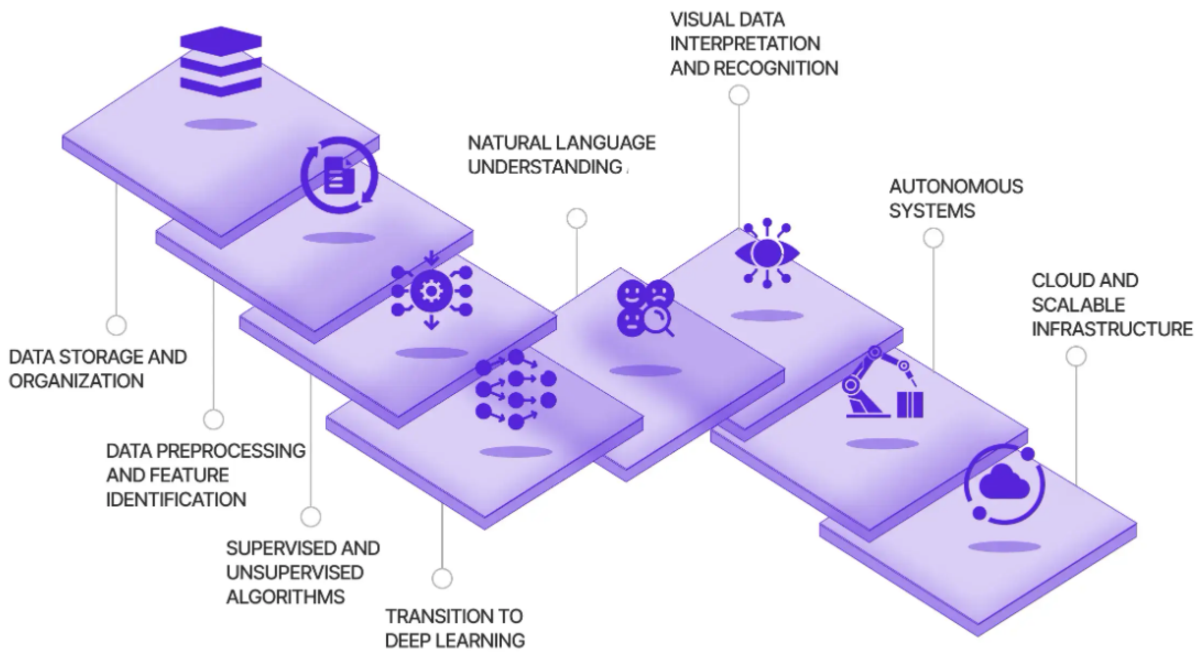
- environment for flexible and agile exploration - EDA[6]
- fast & efficient iteration of algorithm selection, experiements, & analysis
- correct training / validation / test data sets critical!
- seamless productionization from, *e.g.*, Jupyter notebook to production-ready code
- monitorning, *right* metrics, notification, re-training



[6] EDA - exploratory data analysis

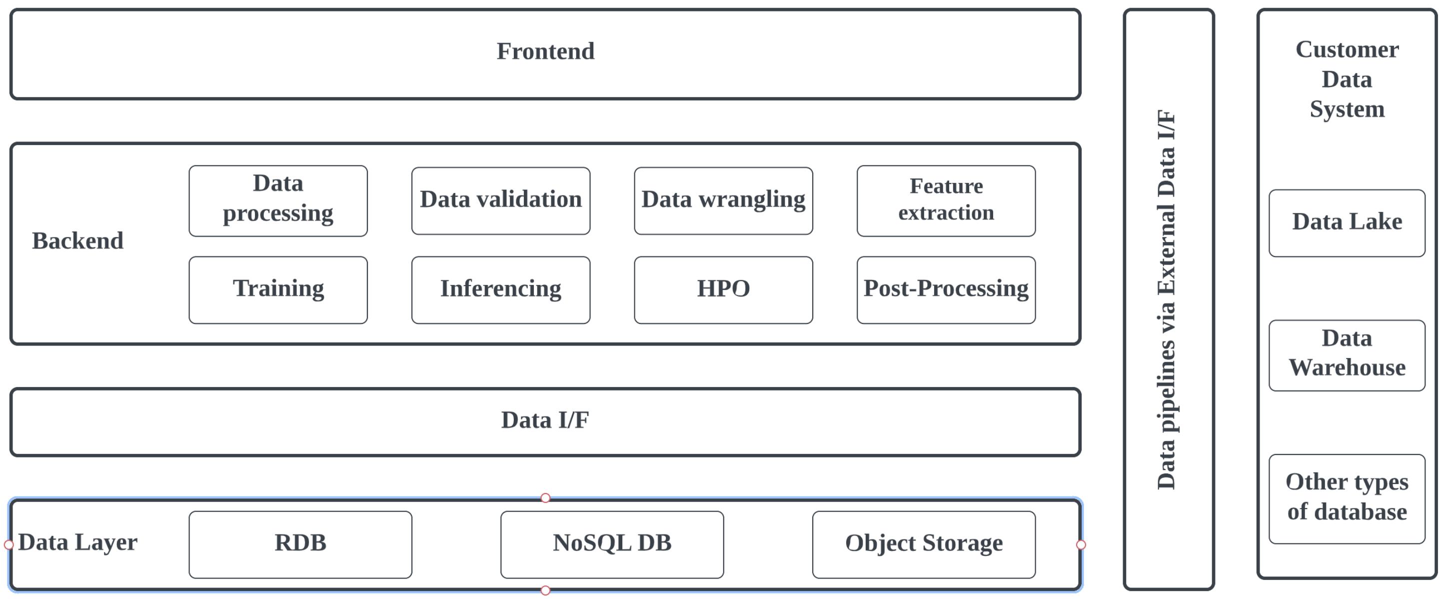
manAI software system

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



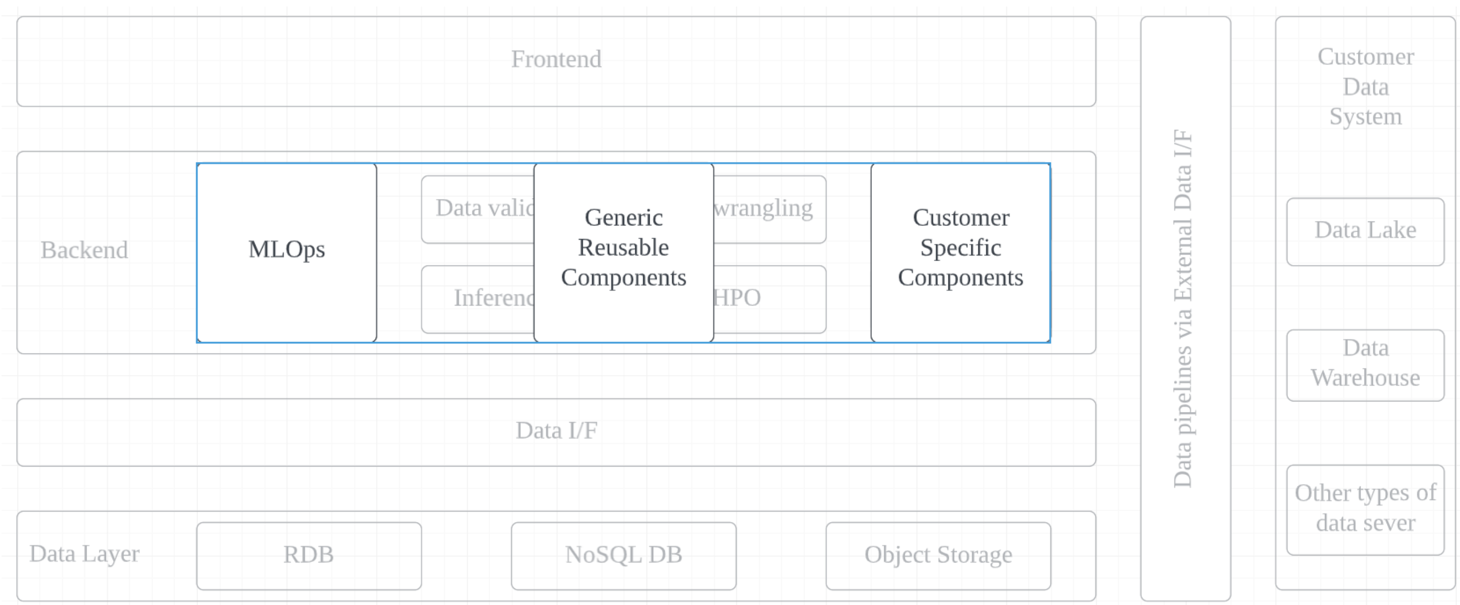
manAI system architecture

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



Reusuable components vs customer specific components

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



My Two Cents

Recommendations for maximum impact via inAI

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

Remember . . .

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

Conclusion

Conclusion

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

Appendices

Serendipities around AIs

Serendipity or inevitability?

- What if Geoffrey Hinton had not been a persistent researcher?
- What if symbolists won AI race over connectionists?
- What if attention mechanism did not perform well?
- What if Transformer architecture did not perform super well?
- What if OpenAI had not been successful with ChatGPT in 2022?
- What if Jensen Huang had not been crazy about making hardware for professional gamers?

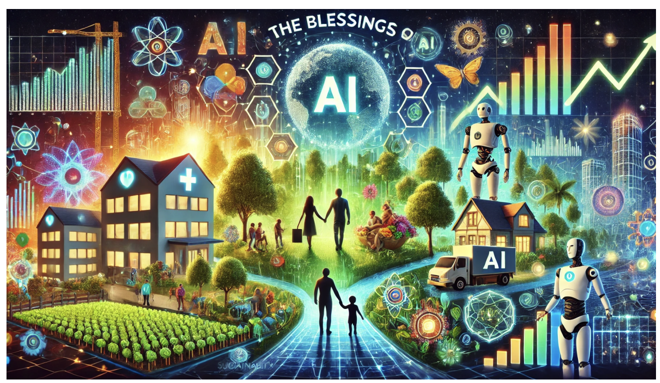
- Is it like Alexander Fleming's Penicillin?
- Or more like Inevitability?

Empowering Humanity for Future Enriched by AI

Blessings & Curses of AI

Blessings

- advancements in healthcare & improved quality of life
  - much faster & more accurate diagnosis, far superior personalized medicine, accelerated drug discovery, assistive technologies
- economic growth & efficiency
  - automation to increase productivity and reduce cost, far superior decision-making
- environmental solutions
  - climate change prediction, global warming effect mitigation, solutions for sustainability
- safety & security
  - natural disaster prediction & relief, cybersecurity



Curses

- job displacement & overall impacts on labor market
  - millions of jobs threatened, wealth gap widened
- bias & inequality, misinformation & manipulation
  - existing human biases, both conscious and unconscious, perpetuated through AIs, asymmetric accessibility to advanced AI technologies by nations & corporations
- ethical dilemmas
  - infringing privacy & human rights, accountability for weapon uses and damages by AI
- environmental costs
  - significant energy for training AI models, waste generated by obsolescent AI hardware

Salzburg Global Seminar

KFAS-Salzburg Global Leadership Initiative

- "Uncertain Futures and Connections Reimagined: Connecting Technologies" - 41 global leaders convened from 4-Dec to 8-Dec, 2024 @ Schloss Leopoldskron in Salzburg, Austria
- My working group was "Technology, Growth, and Inequality: The Case of AI"
  - International Cooperation Officer (Portugal)
  - Gender Equality, Disability Inclusion Consultant, UN Women (Lithuania)
  - Assistant Professor @ Lincoln Alexander School of Law (Canada)
  - Research Associate @ Luxembourg Institute of Socio-Economic Research
  - Policy Officer & Delegation of the EU Union (India)
- blog: Bridging Technology & Humanity - Reflections from Lyon, Salzburg, and München

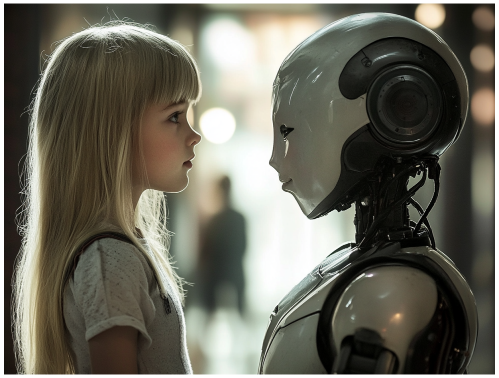
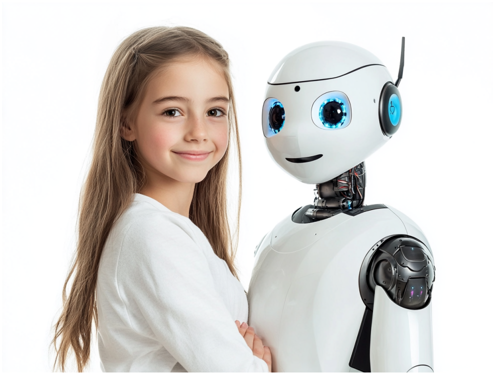
KFAS-Salzburg Global Leadership Initiative

Salzburg Global photo collections

Empowering Humanity

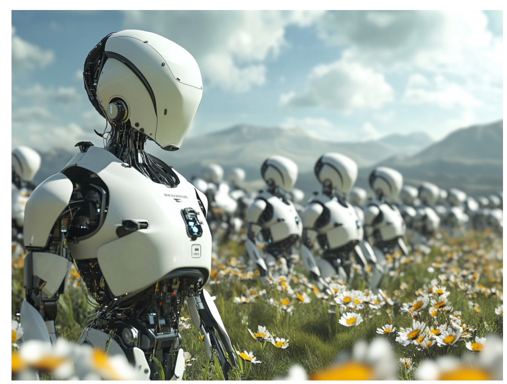
AI capacity building - scientists, engineers & practitioners

- ethics and responsible AI education or campaign via interdisciplinary collaboration
  - foster continuous learning programs on AI risks, bias & societal impacts
- bias detection & mitigation
  - bias-detection tools to identify & reduce discrimination in data & models
  - regular fairness audits
- transparency & explainability
  - explainable AI (xAI) techniques, frameworks like Model Cards for transparency
- environmental impact awareness
  - reduce AI's carbon footprint, advocate for sustainable AI development practices

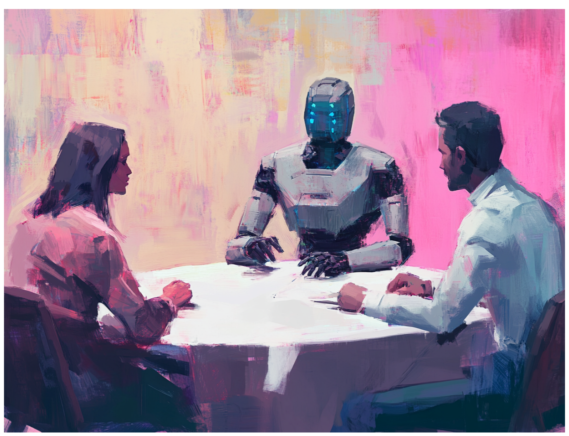

AI capacity building - lawmakers & policy makers

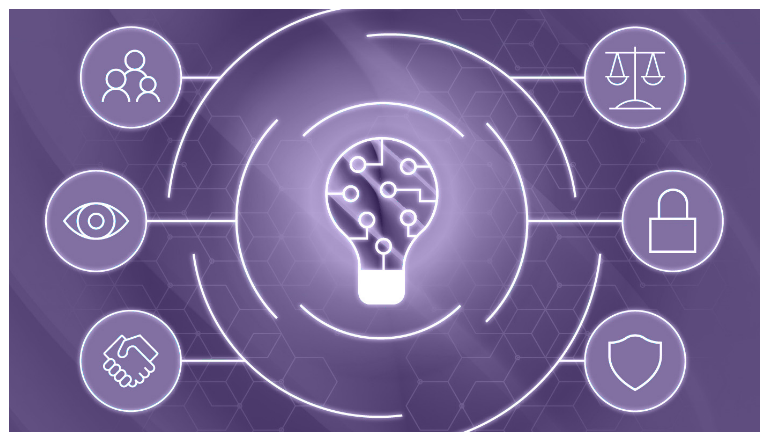
- problems
  - difficulties in understanding of rapidly evolving AI technologies
  - lead to reactive or insufficient regulation
- proposed solutions
  - develop comprehensive regulatory frameworks addressing transparency, bias & privacy concerns
    - gender bias, racial bias, hallucinations
  - foster public debates on ethical AI use & societal implications
  - introduce policies to limit spread of AI-generated misinformation,



Participatory social agreements

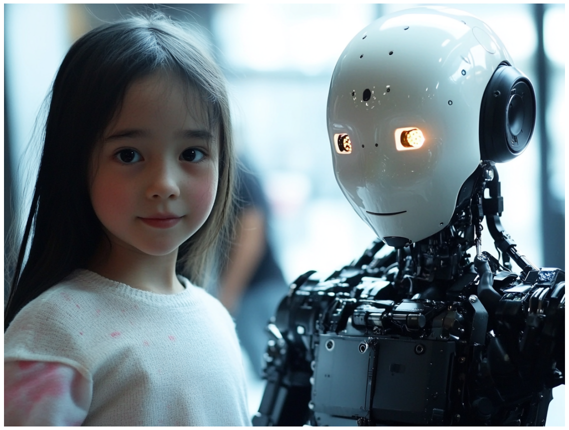
- open data frameworks including data sovereignty, regulation of data transfer, storage & localization
- corporate social responsibility, extra-territorial obligations & environmental protection
  - including outside the jurisdiction of the country
- labour and employment displacements, tax cuts & algorithmic impact assessments
  - including remedies for AI harms and enforcements





Reclaiming technology for Humanity

- strategic approach to AI development
  - *leverage very technologies alienating humans to strengthen human connection*
  - transform automation from replacement to *enhancement of human capabilities*
  - leverage technological scale to address fundamental human needs
- *paradigm shift* in technological implementation
  - recognize the duality of advanced technologies
  - *systematically channel AI capabilities toward human-centric solutions*
  - convert technological challenges into opportunities for human advancement



Selected References & Sources

Selected references & sources

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

References

References

[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.

[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.

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

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

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

[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