AI for Fusion Biweekly Seminar: Industrial AI & Biotechnology - Technology, Market, and Future

Sunghee Yun

Co-founder / CTO - AI Technology & Product Strategy

Erudio Bio, Inc.

Today

- industrial AI
  - why industrial AI?
  - computer vision (CV) and time-series (TS) AI in manufacturing
  - challenges for manufacturing AI
  - industrial AI success story - virtual metrology
- biotechnology
  - AI & bio
  - biotechnology - multidisciplinary field
  - bio data and processing cost
  - emerging trends in biotech
- AI industry
  - heavy lifting of LLMs
  - tech giants & AI companies

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)

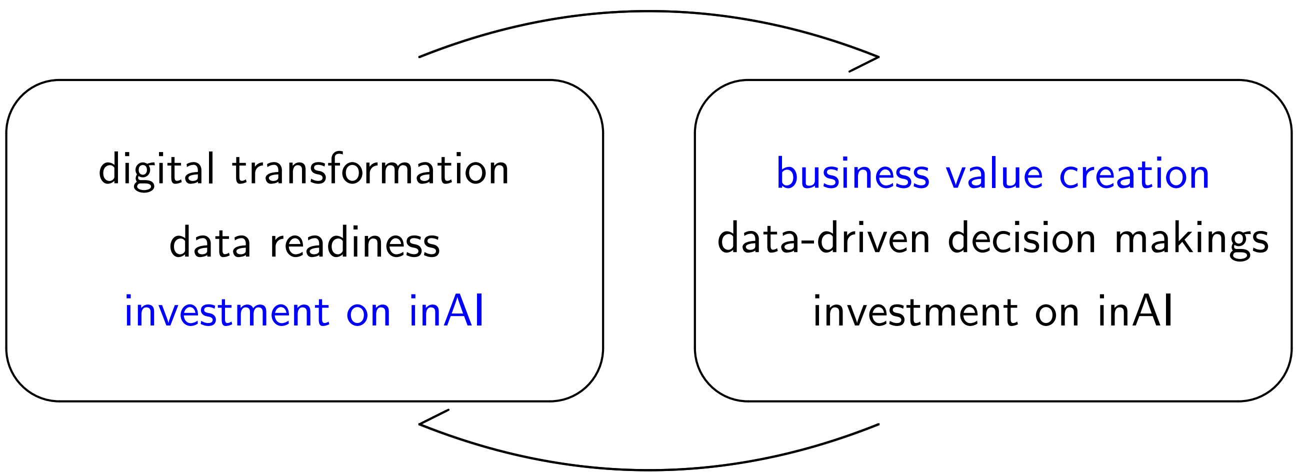
inAI fields

- 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!

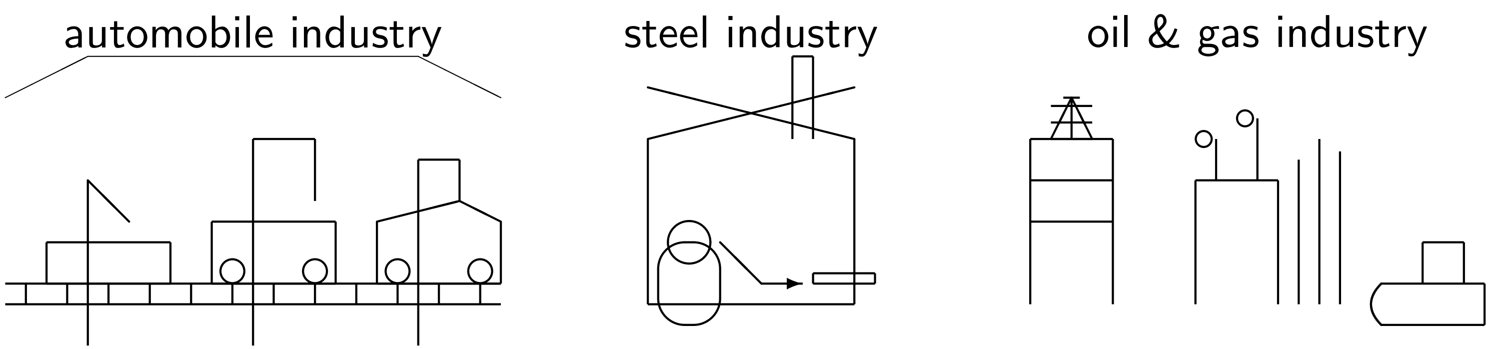


Data-centric AI

- 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



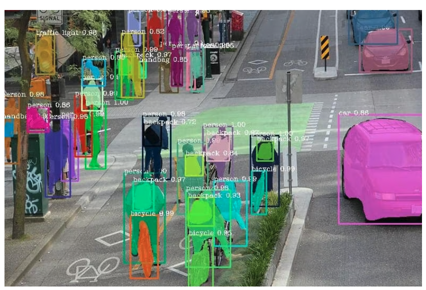
Challenging data characteristics

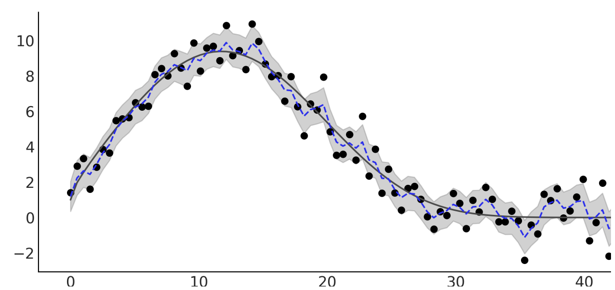
- huge volume
- data multi-modality
- high velocity requirement
- very fat data
- sever data shift & drift (in many cases)
- label imbalance
- data quality

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 [1]
  - → 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 [2]
  - → regression, anomaly detection, semi-supervised learning, Bayesian inference





[1] SEM: scanning electron miscroscope, TEM: transmission electron miscroscope
[2] MES: manufacturing execution system

CV ML in manAI

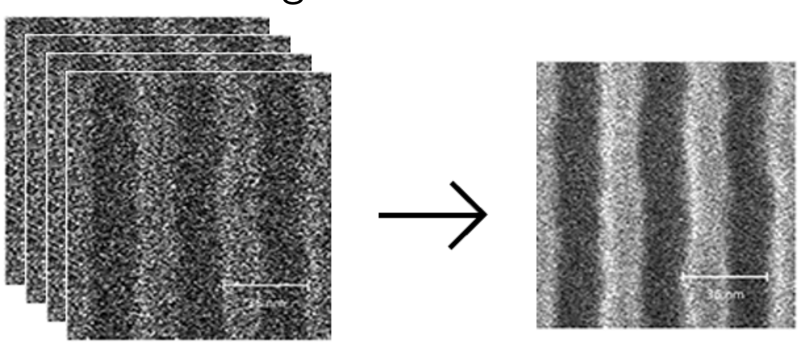
Computer vision ML in manAI

- 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

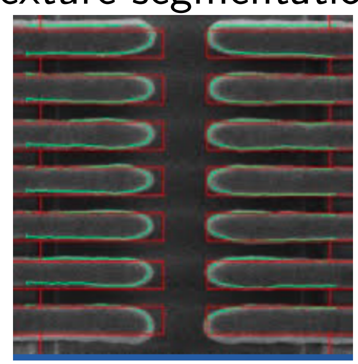
Automatic feature measurement

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

image enhancement



texture segmentation



auto-measurement

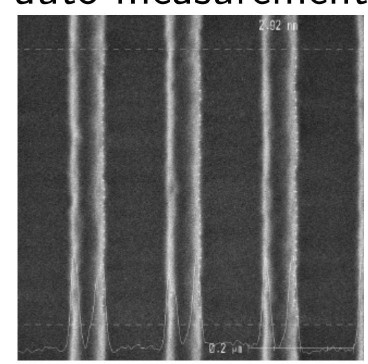
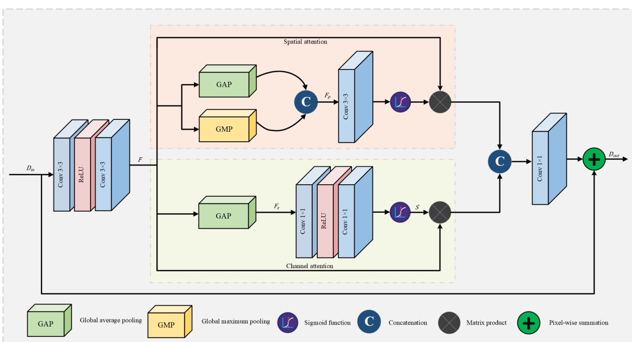


Image enhancement

- 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



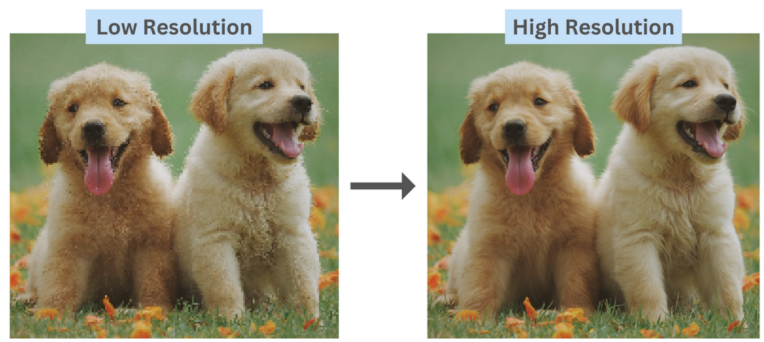
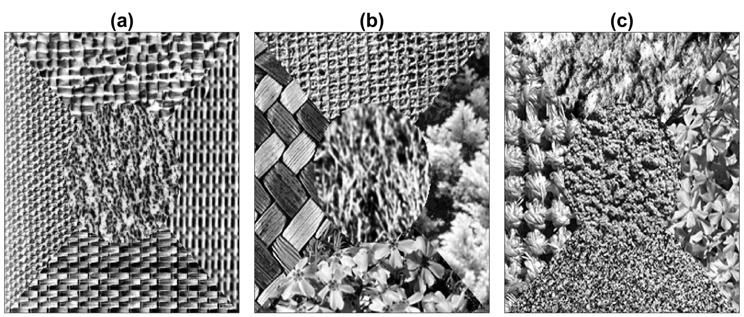


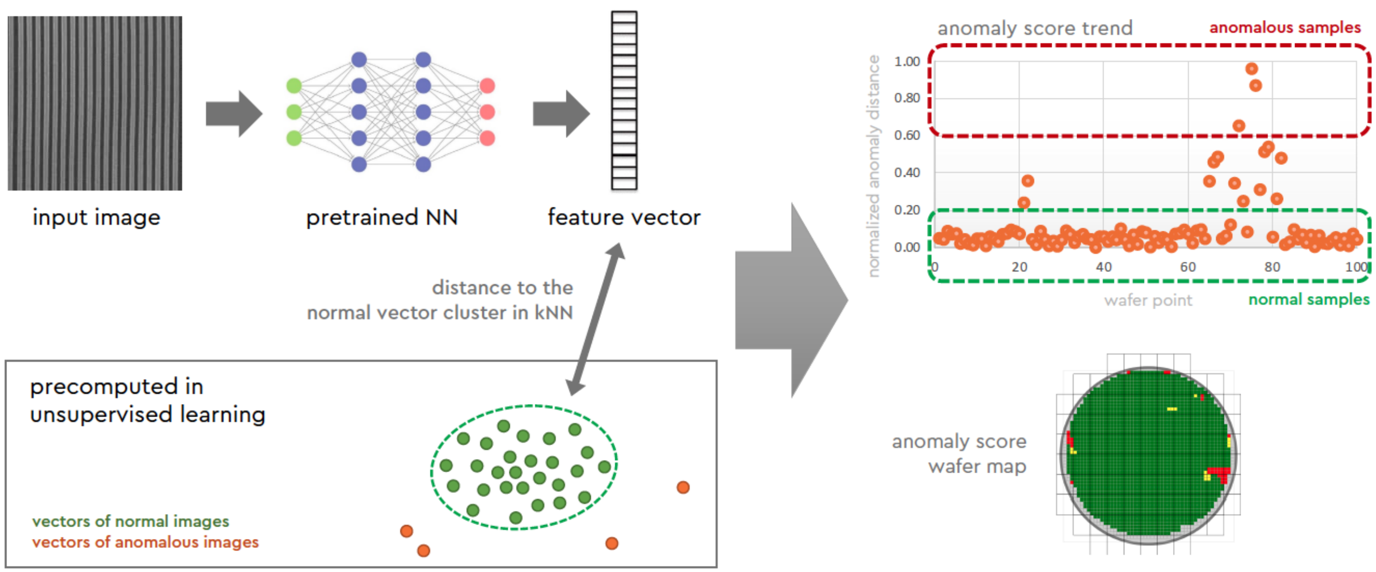
Image segmentation

- 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



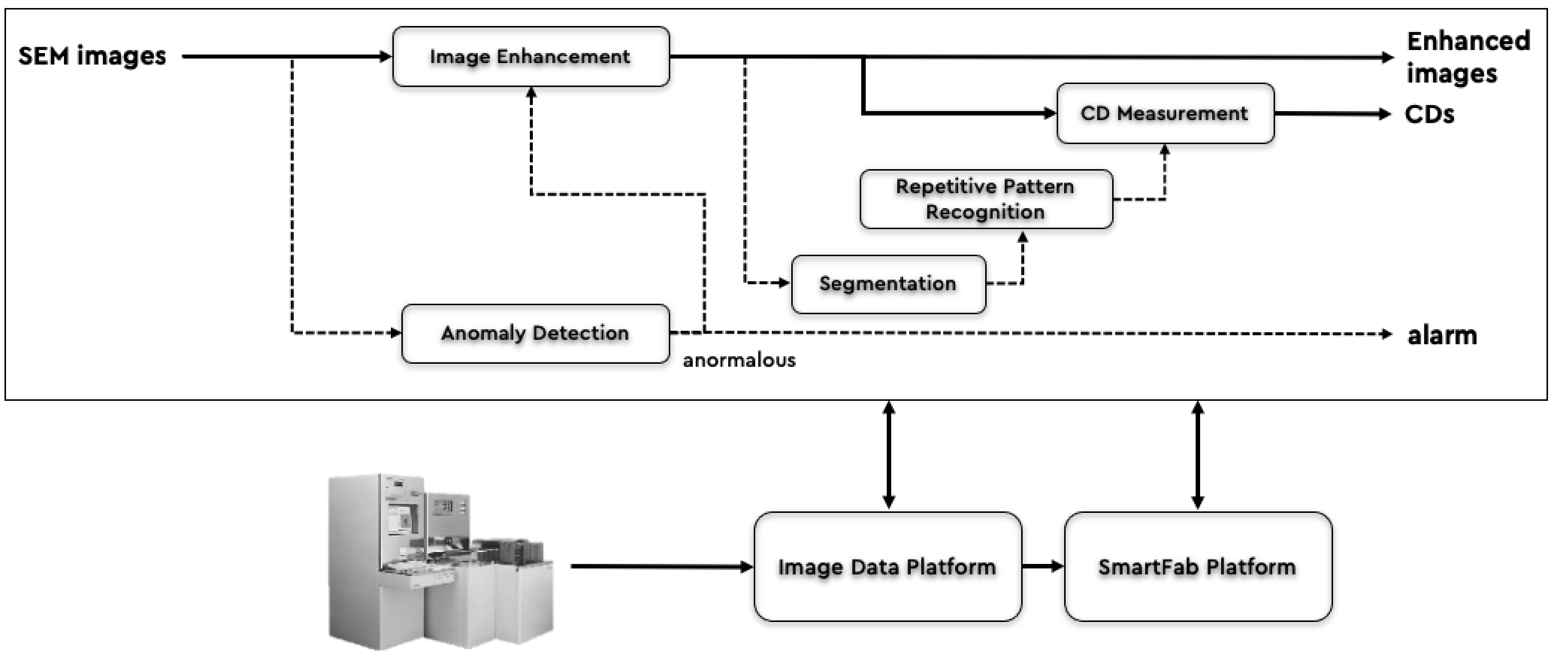
Anomaly detection using side product

- 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



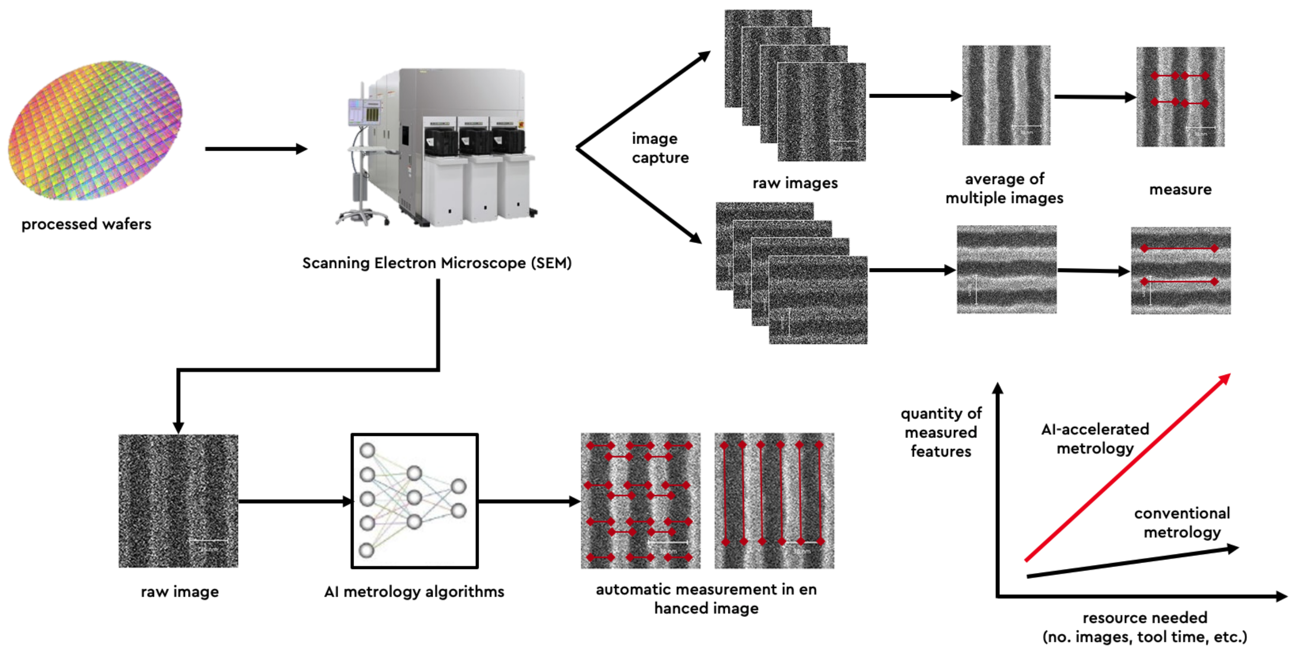
AI-enabled metrology system

- integration of separate components creates AI-enabled metrology system



Benefits of new system

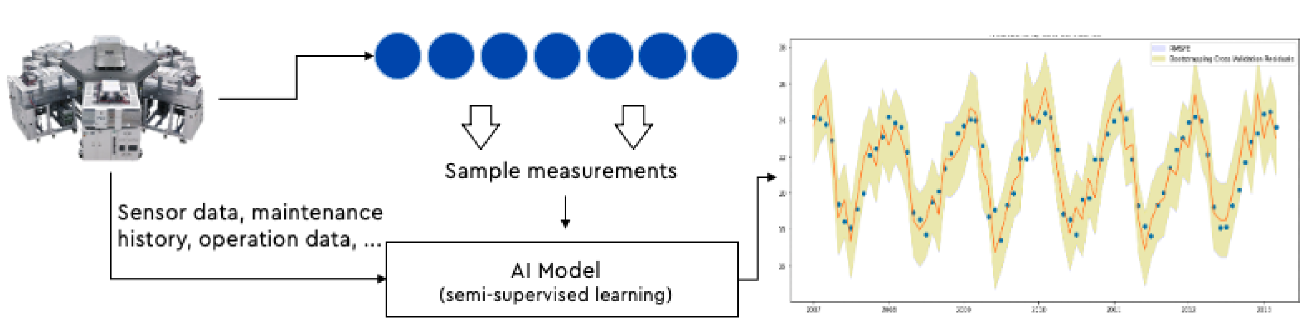
- new system provides
  - improved accuracy and reliability
  - improved throughput
  - savings on investment on measurement equipment



TS ML in manAI

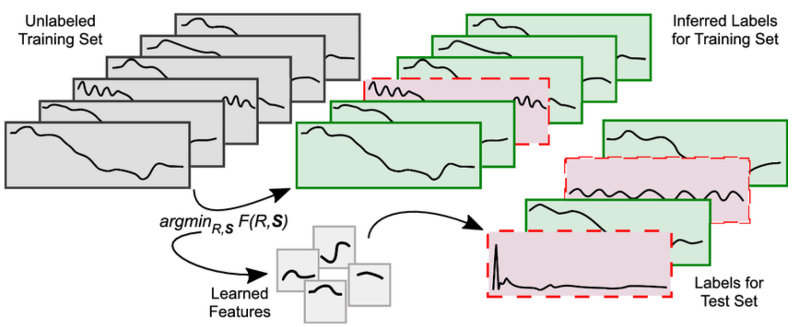
Time-series ML applications in manAI

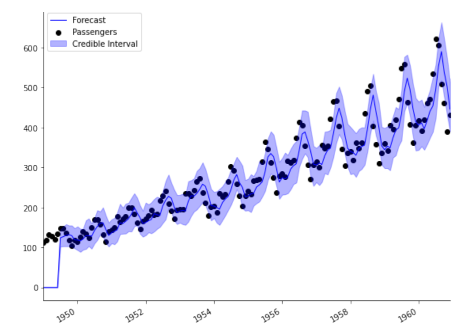
- estimation of TS values
  - virtual metrology - estimate measurement without physically measuring things
- anomaly detection on TS
  - predictive maintenance - predict maintenance times ahead
- multi-modal ML using LLM & genAI
  - root cause analysis and recommendation system



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



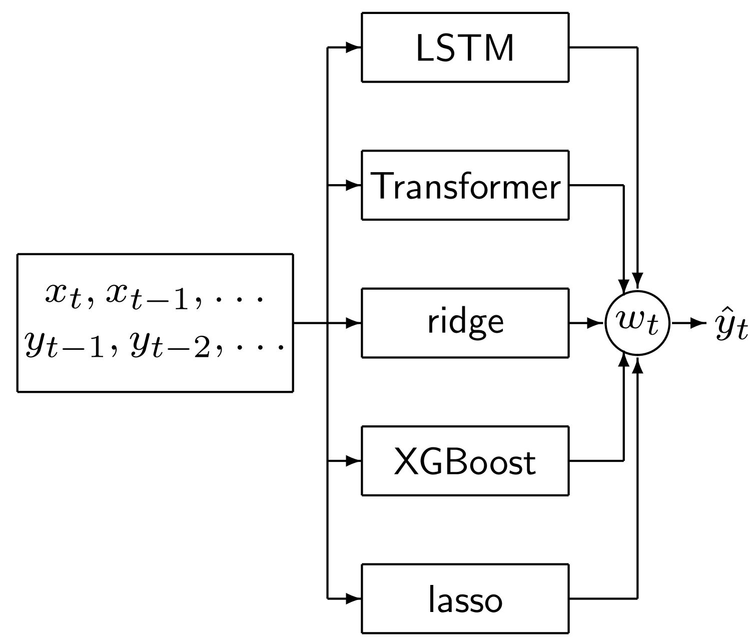


Difficulties with TS ML

- 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



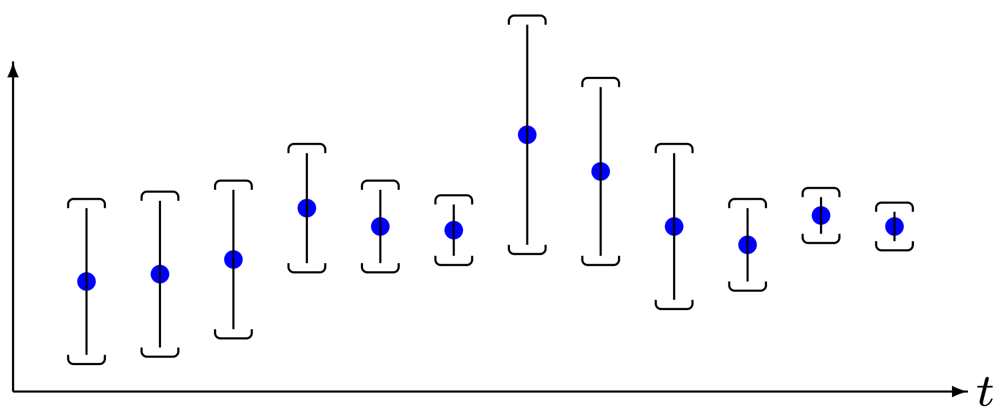
$$g_k = w_{1,k} f_{1,k} + w_{2,k} f_{2,k} + \cdots + w_{p_k,k} f_{p_k,k} = \sum_{i=1}^{p_k} w_{i,k} f_{i,k}$$

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*



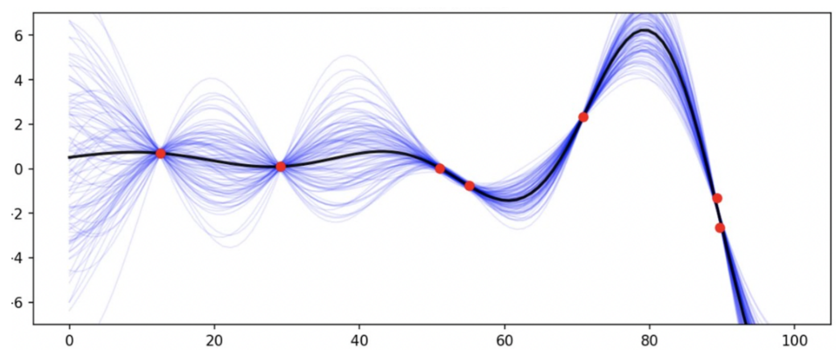
Bayesian approach for credibility interval evaluation

- 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

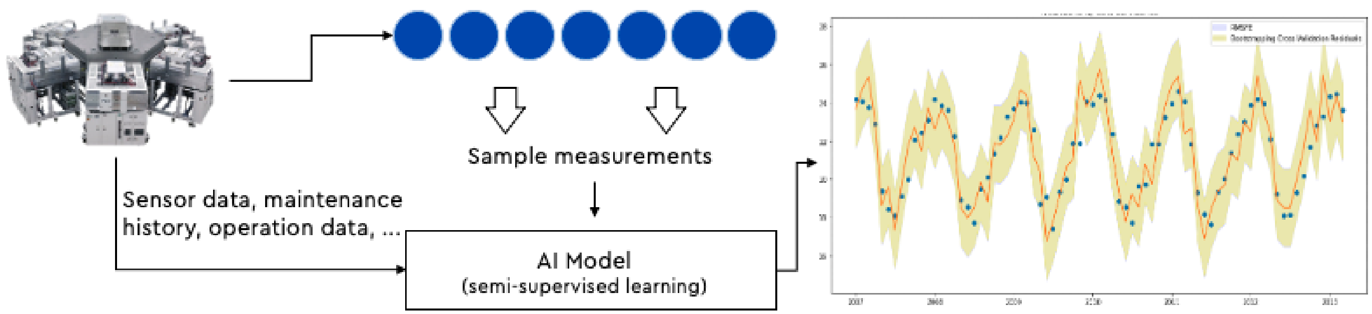
- problem description

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

- our problem formulation

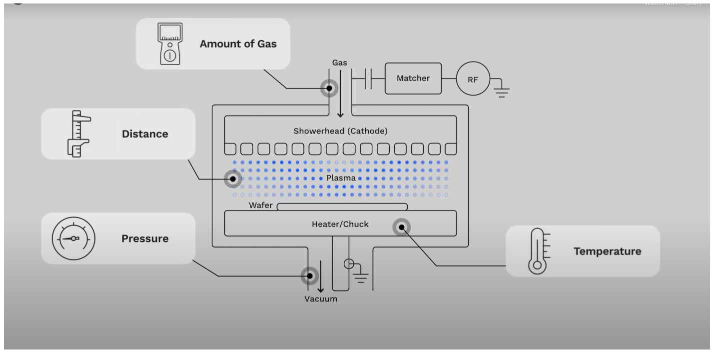
$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{K} w_{k,K-k}\, l(y_{t_k}, \hat{y}_{t_k}) \\ \text{subject to} & \hat{y}_{t_k} = g_k(x_{t_k}, x_{t_{k-1}}, \ldots, y_{t_{k-1}}, y_{t_{k-2}}, \ldots) \end{array}$$

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



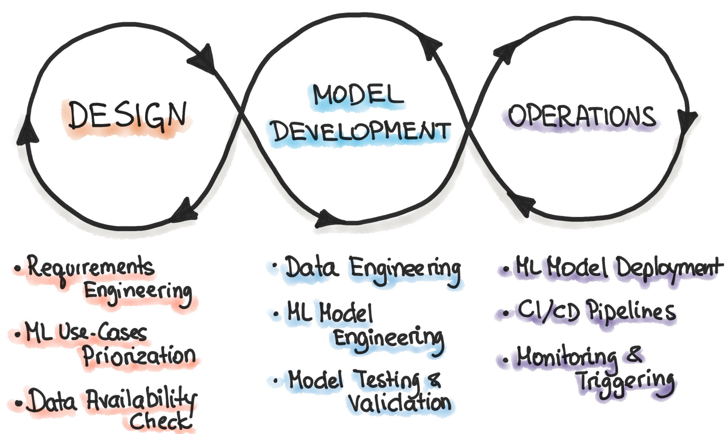
Manufacturing AI Productionization

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

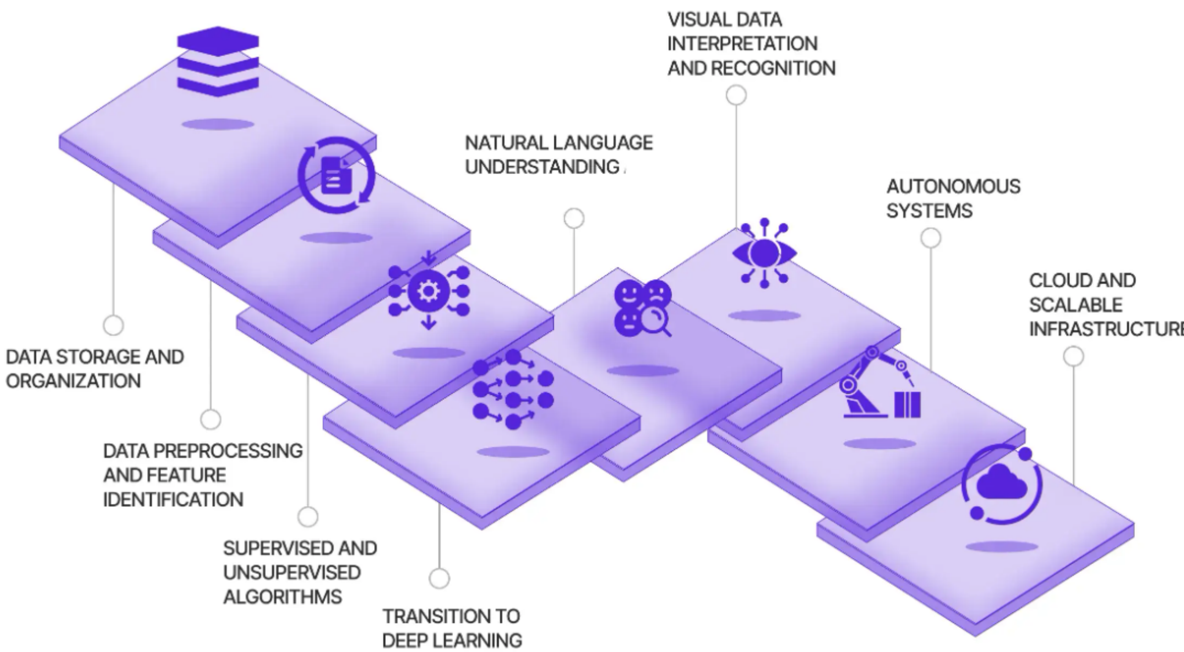
- environment for flexible and agile exploration - EDA[3]
- fast & efficient iteration of algorithm selection, experiements, & analysis
- correct training / validation / test data sets critical!
- seamless productionization from, *e.g.*, Jupyter notebook to production-ready code
- monitorning, *right* metrics, notification, re-training



[3] EDA - exploratory data analysis

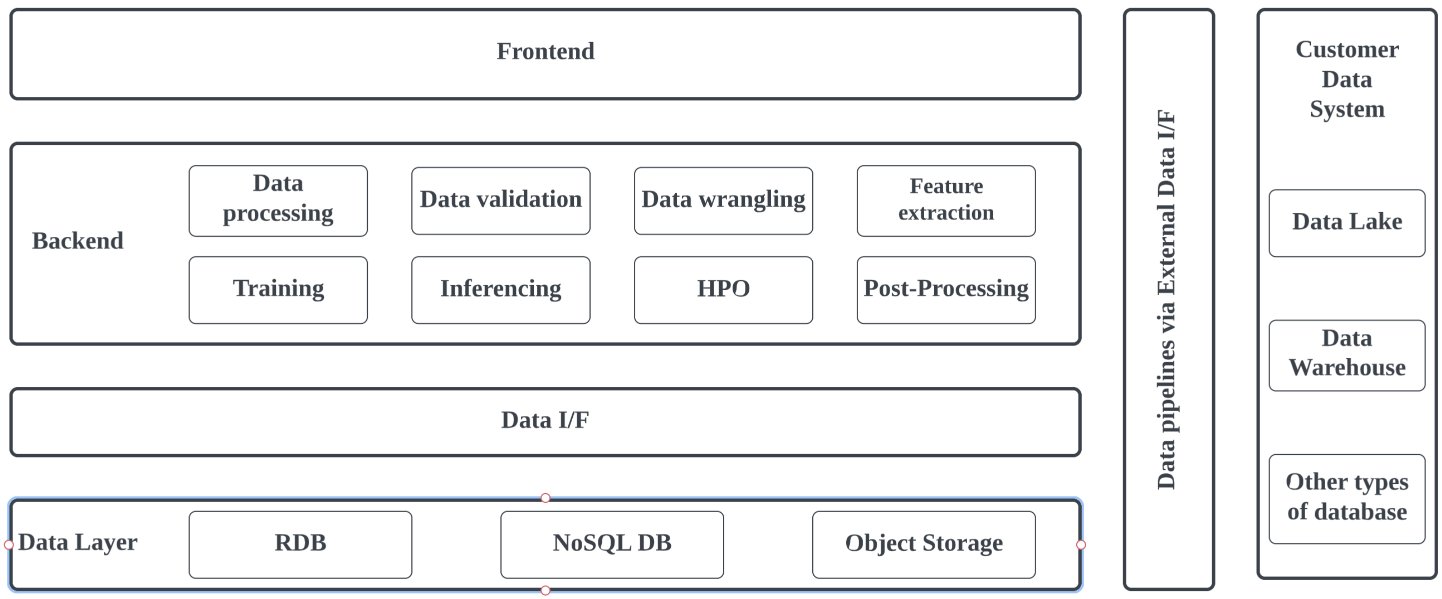
manAI software system

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



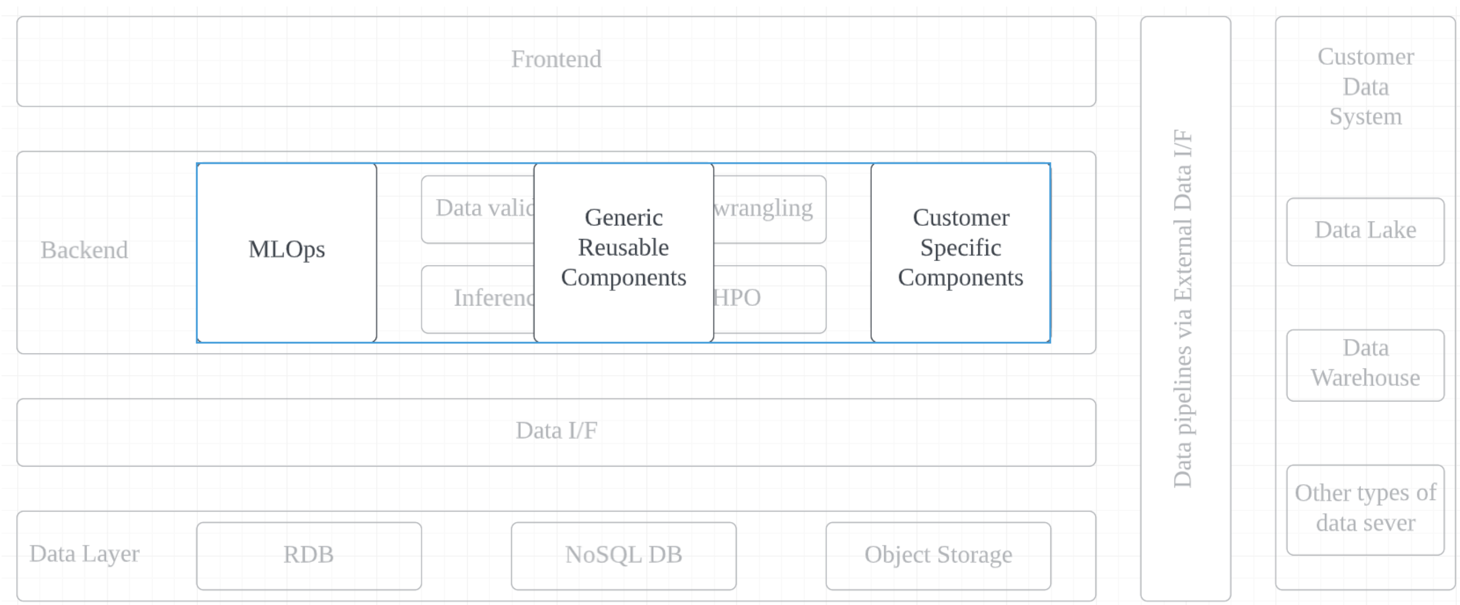
manAI system architecture

- 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

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.

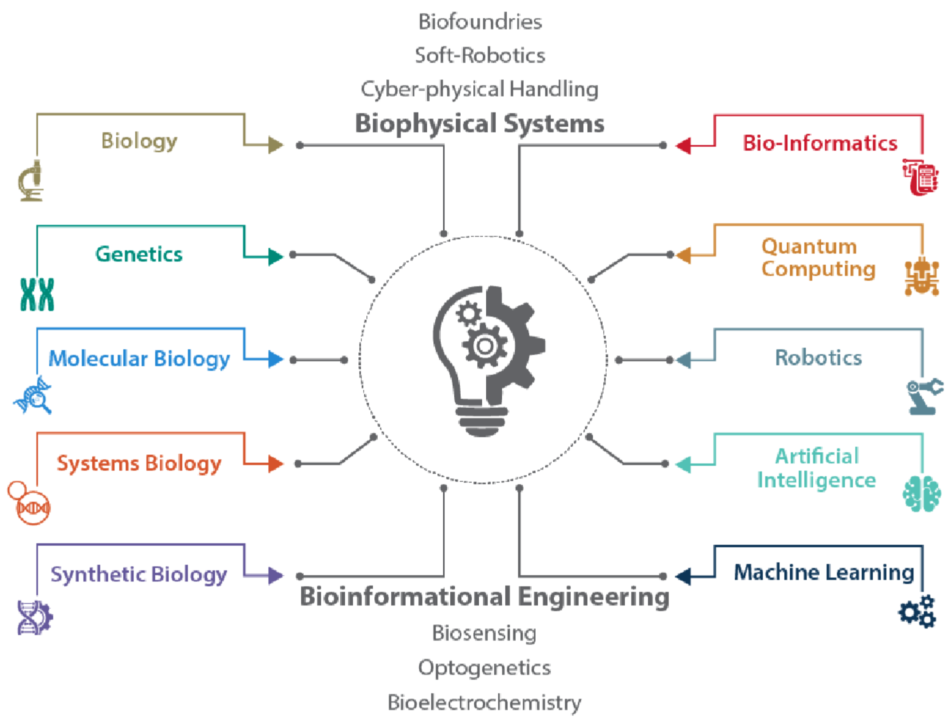
Biotech

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]

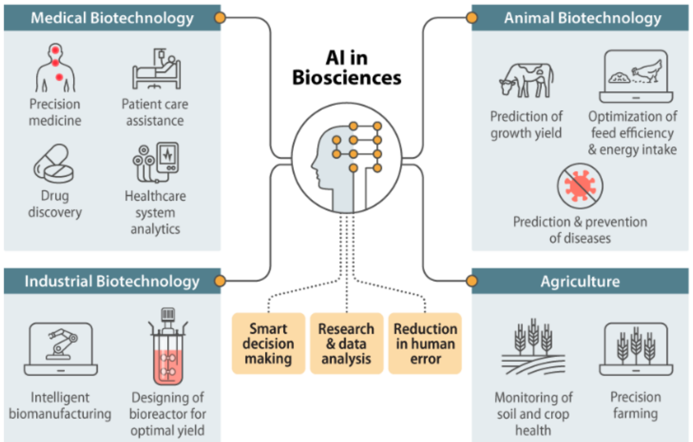


Convergence of AI and biological design

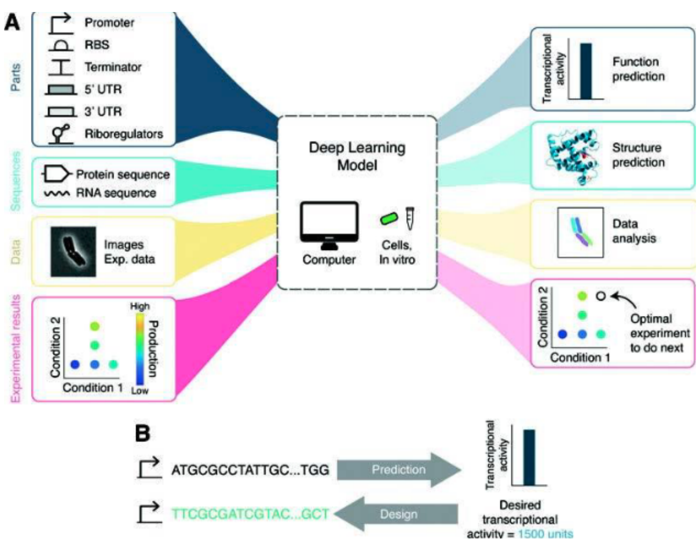
- both AI & biological sciences increasingly converging [BKP22]
  - each building upon the other's capabilities for new research and development across multiple areas
- Demo Hassabis, CEO & cofounder of DeepMind, said of biology [Toe23]

  > ". . . 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

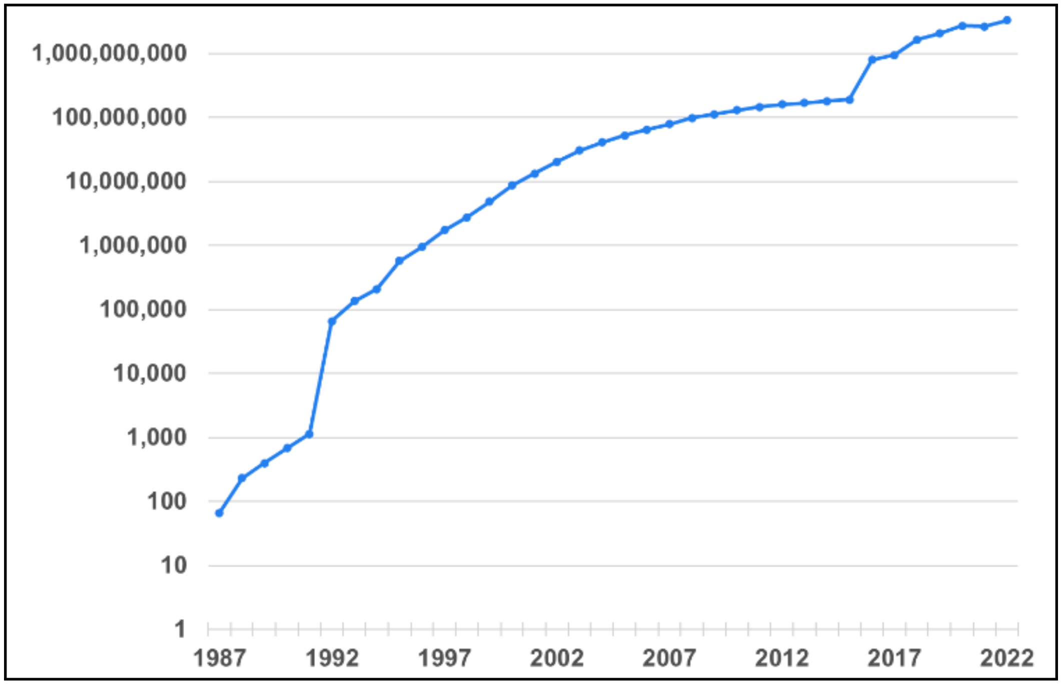
Rapid growth of biological data

- 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

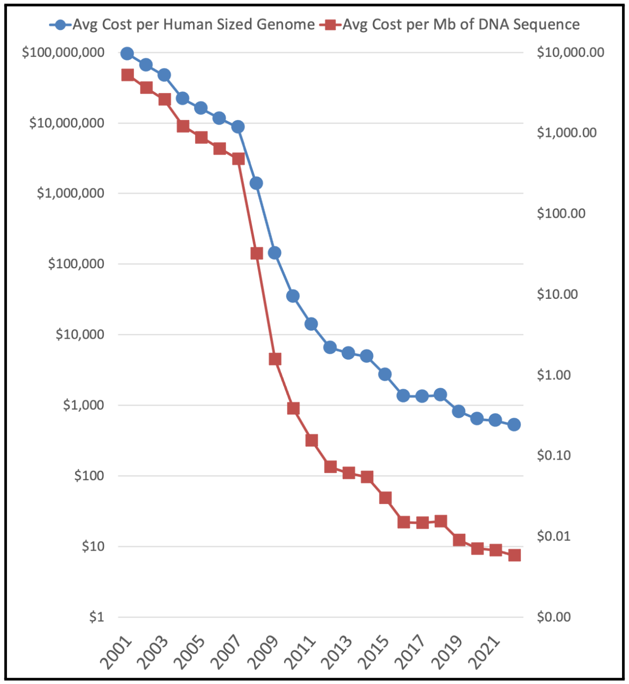
Volume and sequencing cost of DNA over time

- 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



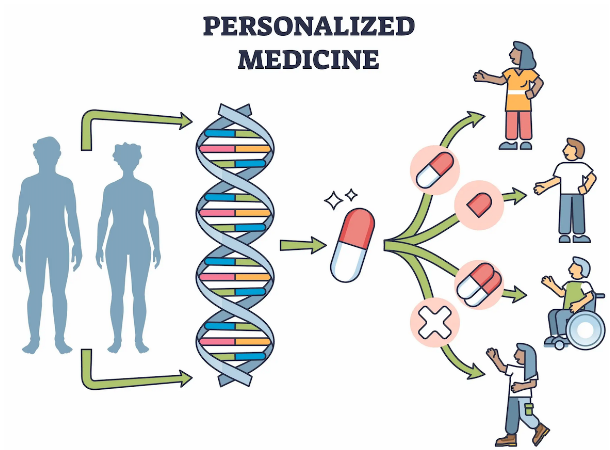
Bio data availability and bias

- 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

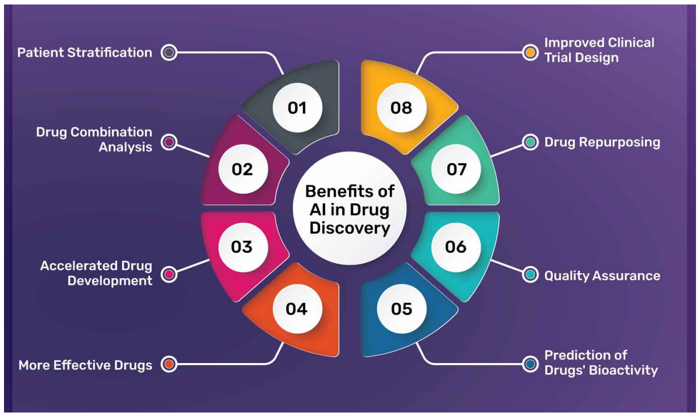
Emerging Trends in Biotech

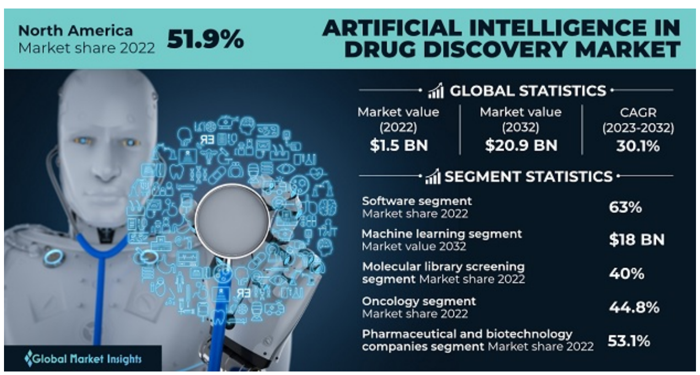
Personalized medicine

- *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

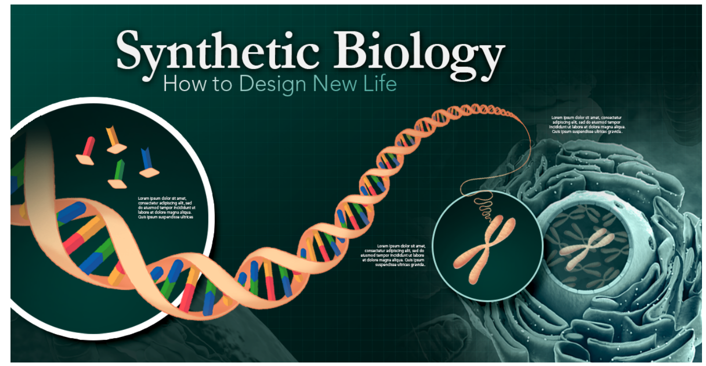


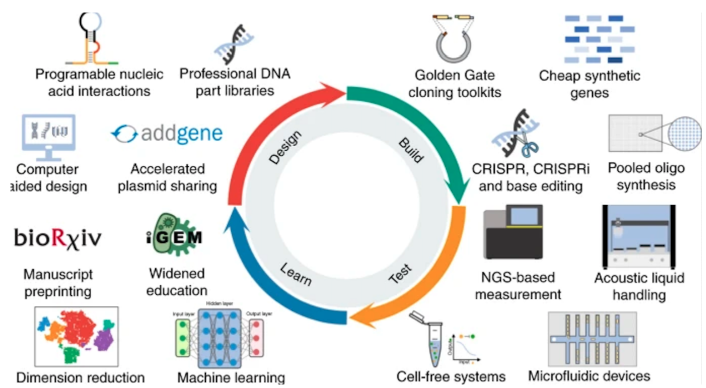


- traditional drug discovery process - time-consuming and costly often taking decades and billions of dollars
- AI streamlines this process by predicting the efficacy and safety of potential compounds with more speed and accuracy
- AI models analyze chemical databases to identify new drug candidates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

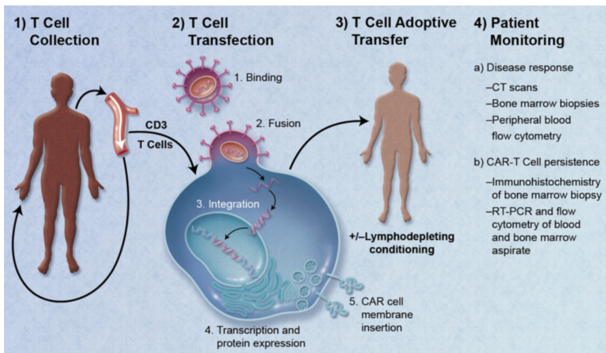
Synthetic biology

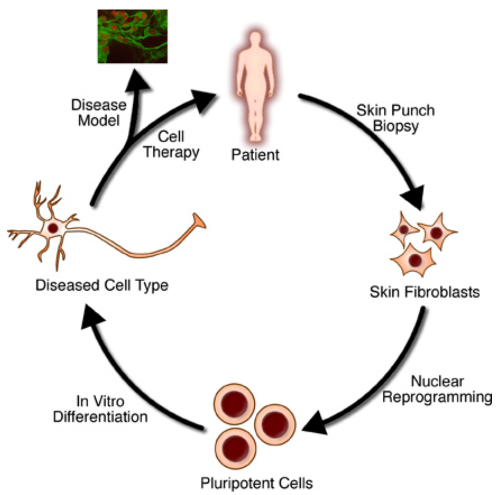
- use AI for gene editing, biomaterial production and synthetic pathways
- combine principles of biology and engineering to design and construct new biological entities
- AI optimizes synthetic biology processes from designing genetic circuits to scaling up production
- company - Ginkgo Bioworks uses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





Regenerative medicine

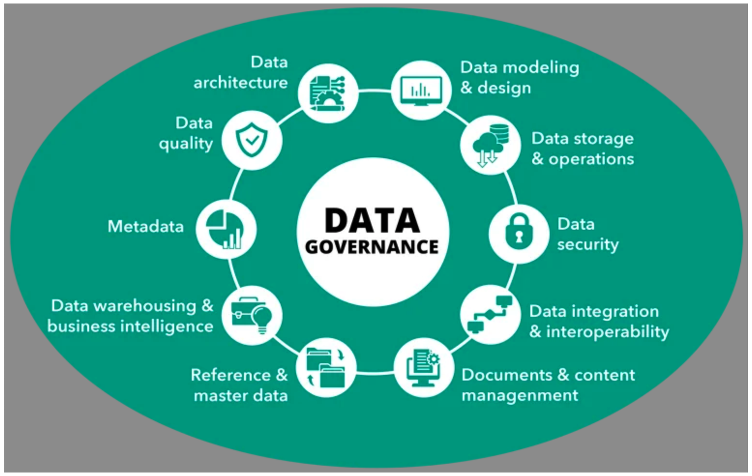




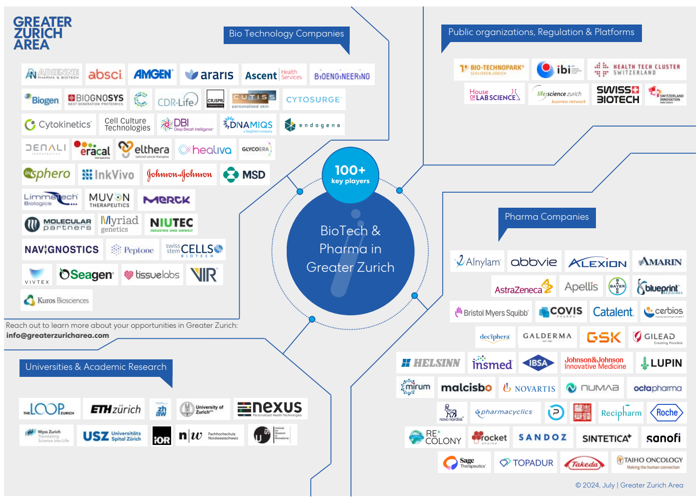
- AI advances development of stem cell therapies & tissue engineering
- AI algorithms assist in identifying optimal cell types, predicting cell behavior & personalized treatments
- particularly for conditions such as neurodegenerative diseases, heart failure and orthopedic injuries
- company - Organovo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

AI Industry

Heavy Lifting of LLMs

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

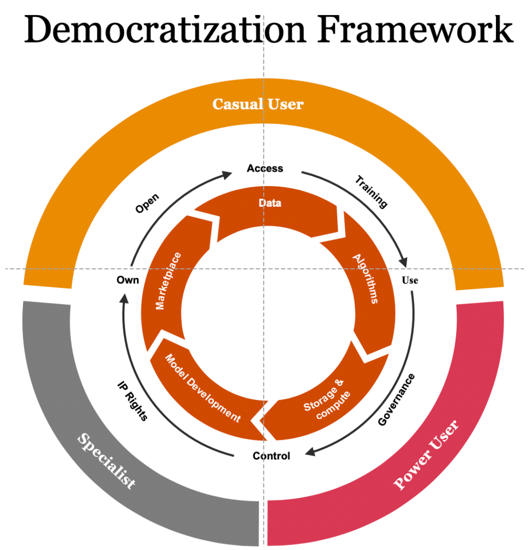
- OpenAI's financial situation reflects its ambitious vision
  - projected $8.5B expenses vs $3.5–4.5B revenue in 2024 w/ massive investment in AI infrastructure and talent
- caused by Sam Altman's reckless & non-strategic commitment to AGI development
  - "Whether we burn $500M, $5B, or $50B a year,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

LLM - strategic challenges & industry dynamics

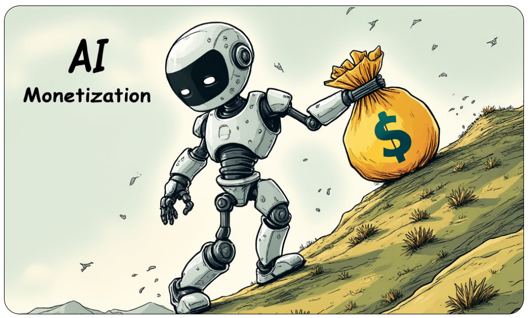
- evolving competitive landscape
  - threat from open-source models ($e.g.$, Meta's Llama 3.1) & potential commoditization of LLMs
- balancing act with Microsoft partnership
  - critical financial support vs maintaining independence - Microsoft's $13B investment provides both opportunity and constraint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

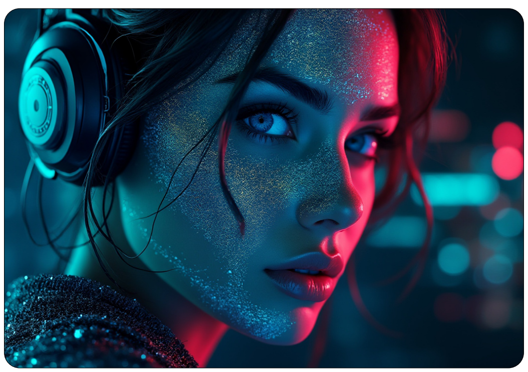
Industry disruption of open-source AI models on industry

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



Impact of open-source AI models on industry



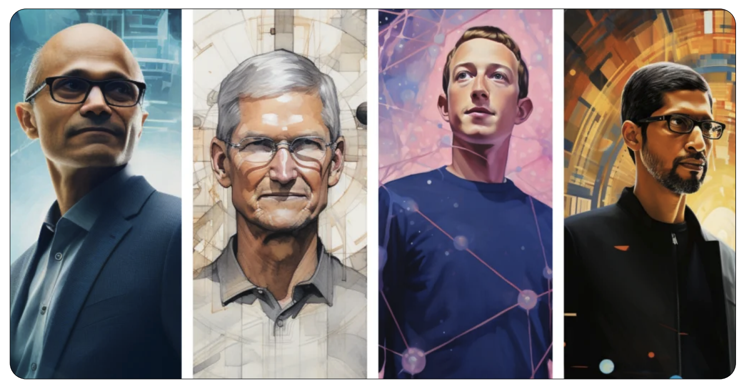


- business model challenges
  - monetization difficulties - capable models becoming freely available
  - shift to services & applications - focus may move from selling access to models to providing *specialized services* or *applications built on top of them*
- ethical & security concerns
  - responsible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

Tech Giants & AI Companies

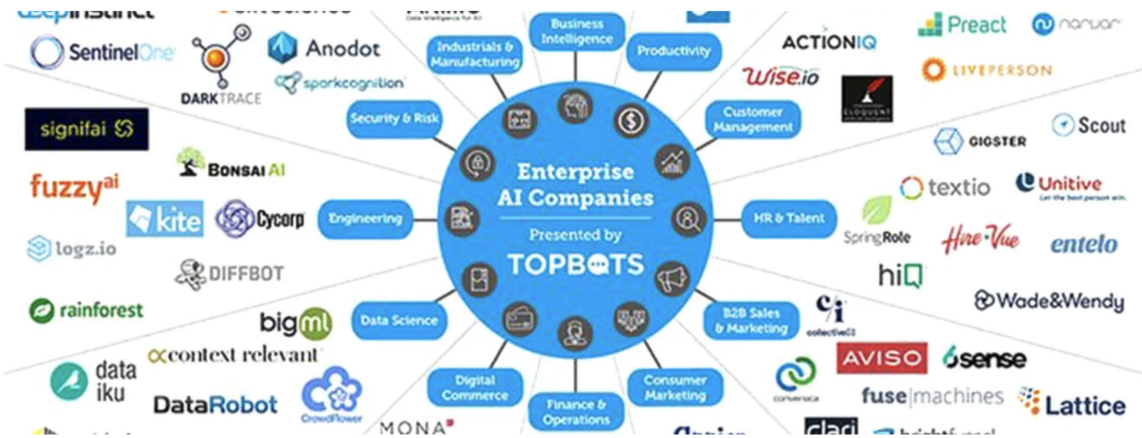
Evolving relationship between tech giants & AI companies

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



AI industry consolidation

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



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.

[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.

[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.

Thank You