

[Texas A&M Seminar / Intro to Semiconductor & Microelectronics]

Industrial AI - Technological Innovations and Impacts on Semiconductor Manufacturing

Sunghee Yun

Co-founder / CTO - AI Technology & Product Strategy

Erudio Bio, Inc.

Today

- Artificial Intelligence
 - history
 - AI achievement from 2014 to 2024
- AI research and development trend
- industrial AI (inAI)
 - definition and characteristics of inAI
 - computer vision (CV) and time-series (TS) MLs
 - industrial AI success story - virtual metrology (VM)
- conclusion & speaker's recommendations / advice

About Speaker

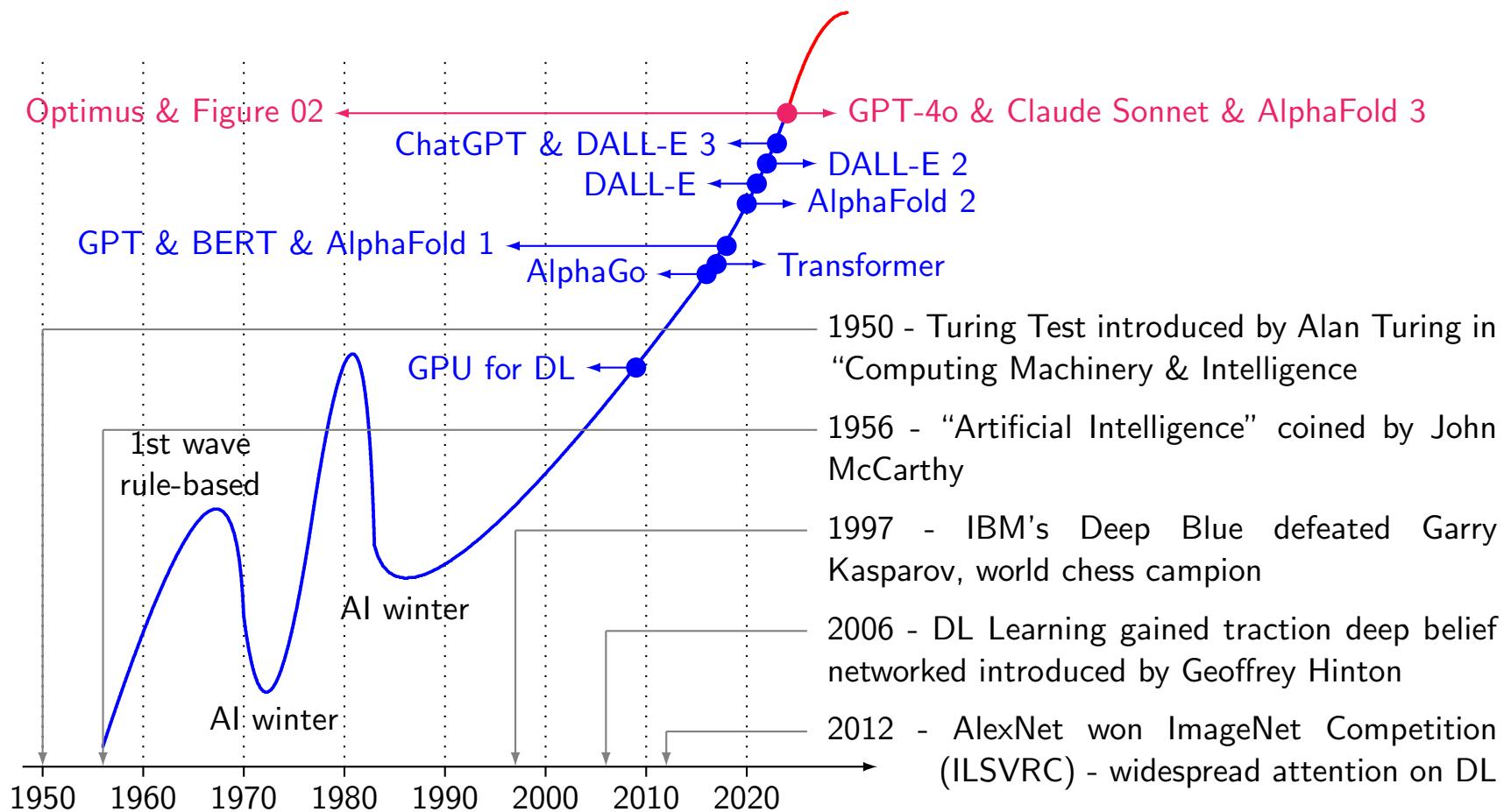
- *Co-founder / CTO - AI Technology & Product Strategy @ Erudio Bio, CA, USA*
- Advisory Professor, Electrical Engineering and Computer Science @ DGIST
- Adjunct Professor, Electronic Engineering Department @ Sogang University
- Technology Consultant @ Gerson Lehrman Group (GLG), NYC, USA
- *Co-founder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

Highlight of career journey

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

Artificial Intelligence

History of AI



Significant AI Achievements - 2014 – 2024

Deep learning revolution

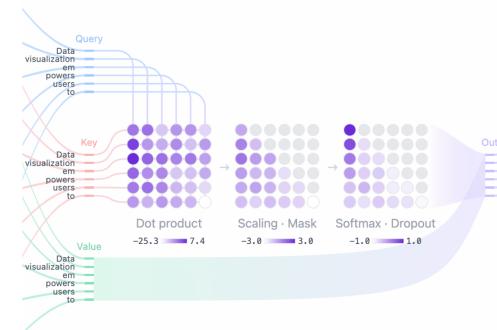
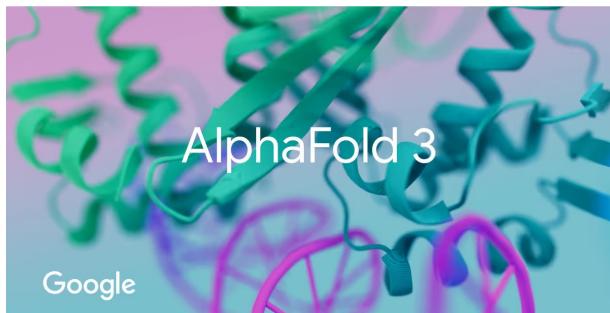
- 2012 – 2015 - Deep Learning Revolution¹
 - CNNs demonstrated exceptional performance in image recognition, e.g., *AlexNet's victory in ImageNet competition*
 - widespread adoption of DL learning in CV transforming industries
- 2016 - AlphaGo Defeats Human Go Champion
 - DeepMind's AlphaGo defeated 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



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

Transformer changes everything

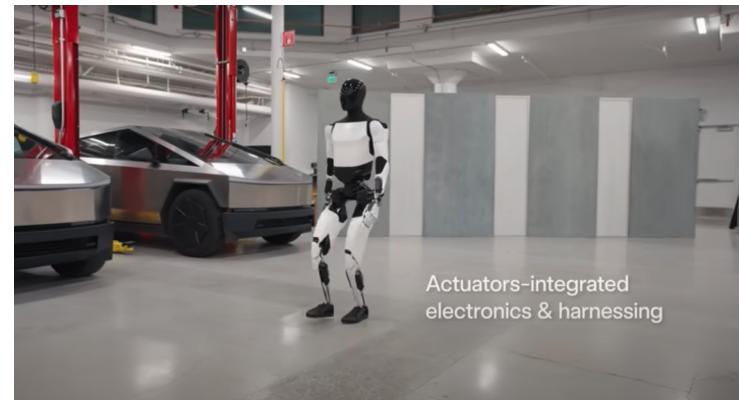
- 2017 – 2018 - Transformers and NLP breakthroughs²
 - *Transformer (e.g., BERT & GPT) revolutionized NLP*
 - major advancements in, *e.g.*, machine translation & chatbots
- 2020 - AI in Healthcare – AlphaFold and Beyond
 - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting 3D protein structures with remarkable accuracy
 - accelerates drug discovery and personalized medicine - offering new insights into diseases and potential treatments



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

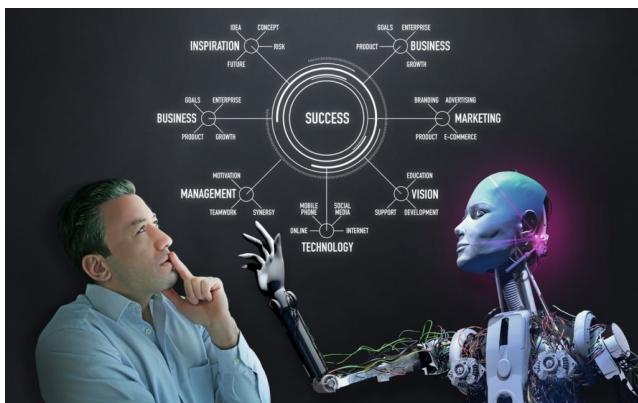
Lots of breakthroughs within 6 months in 2024

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



Transformative impact of AI - reshaping industries, work & society

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



AI Research

AI research race gets crazy

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



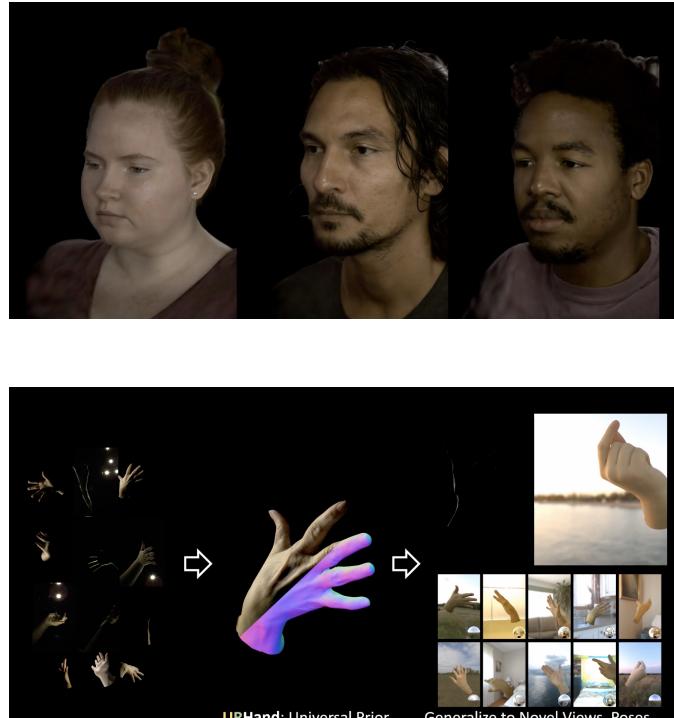
AI progress within a month - March, 2024

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

Papers of single company accepted by single conference

- CVPR 2024

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



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*
 - *semiconductor, steel, oil & gas, cement, and other various manufacturing industries* (unlike general AI, which is frontier research discipline striving to achieve human-level intelligence)



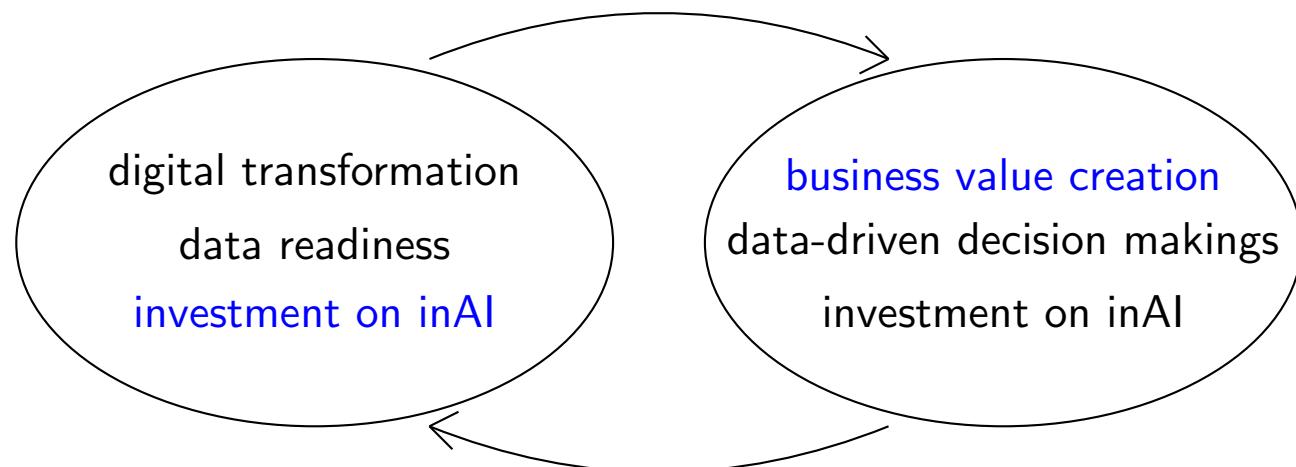
inAI fields

- 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

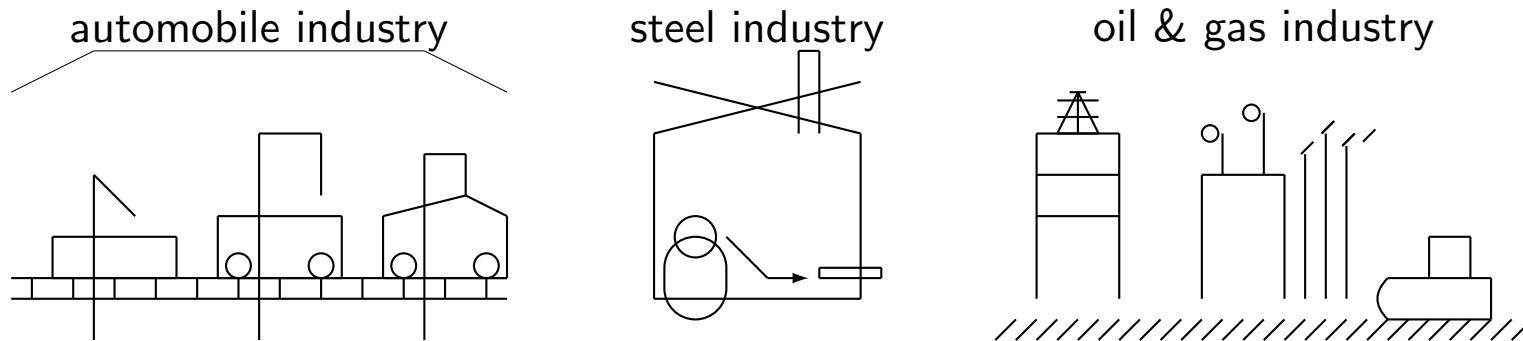
Vicious (or virtuous) cycle

- 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

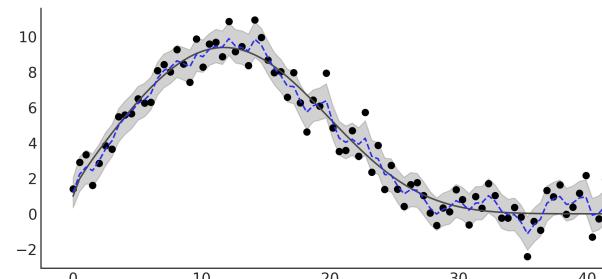
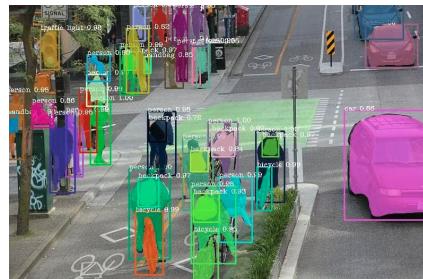
- huge volume
- data multi-modality
- high velocity requirement
- very fat data
- sever data shift & drift (in many cases)
- label imbalance
- data quality



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 maps³
→ 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 data⁴
→ regression, anomaly detection, semi-supervised learning, Bayesian inference



³SEM: scanning electron microscope, TEM: transmission electron microscope

⁴MES: 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

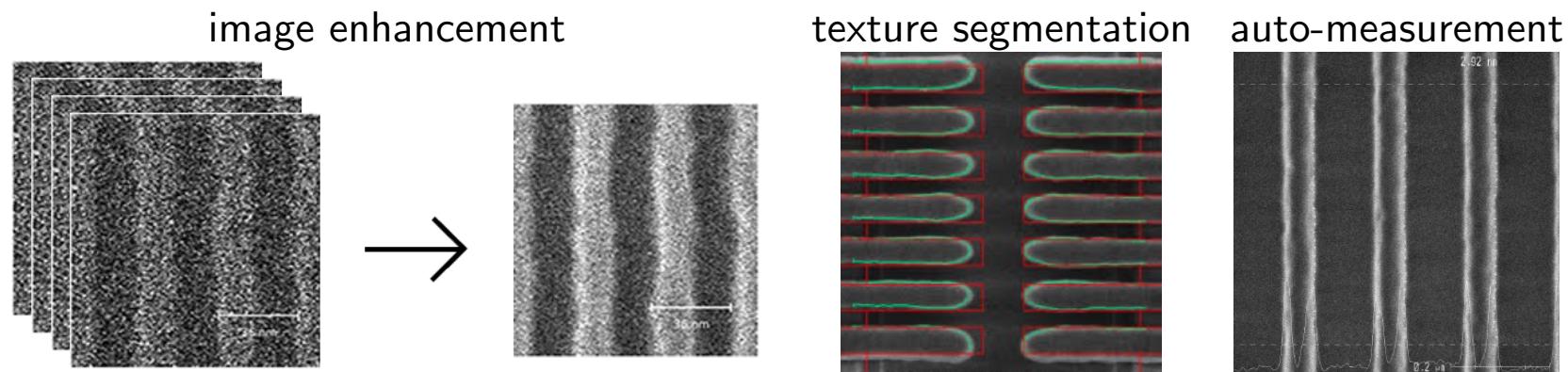


Image enhancement

- 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

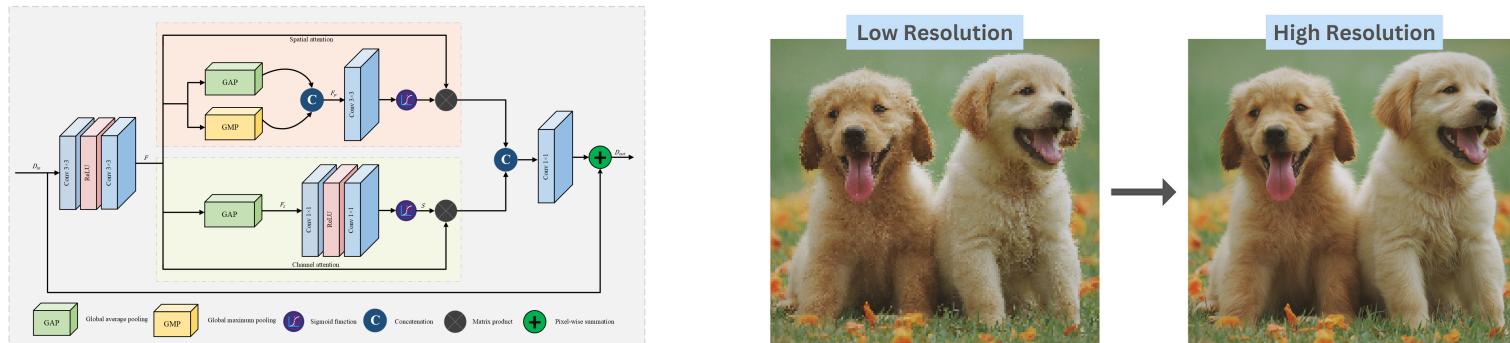
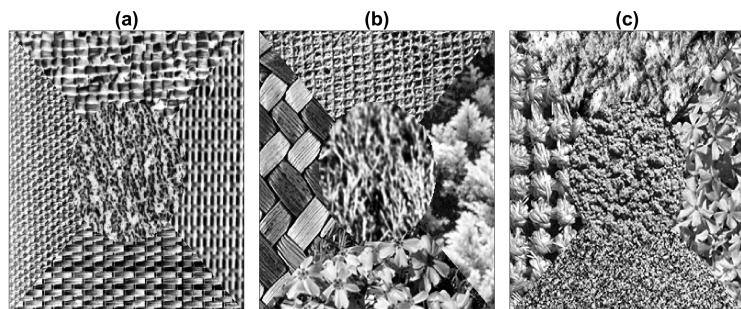


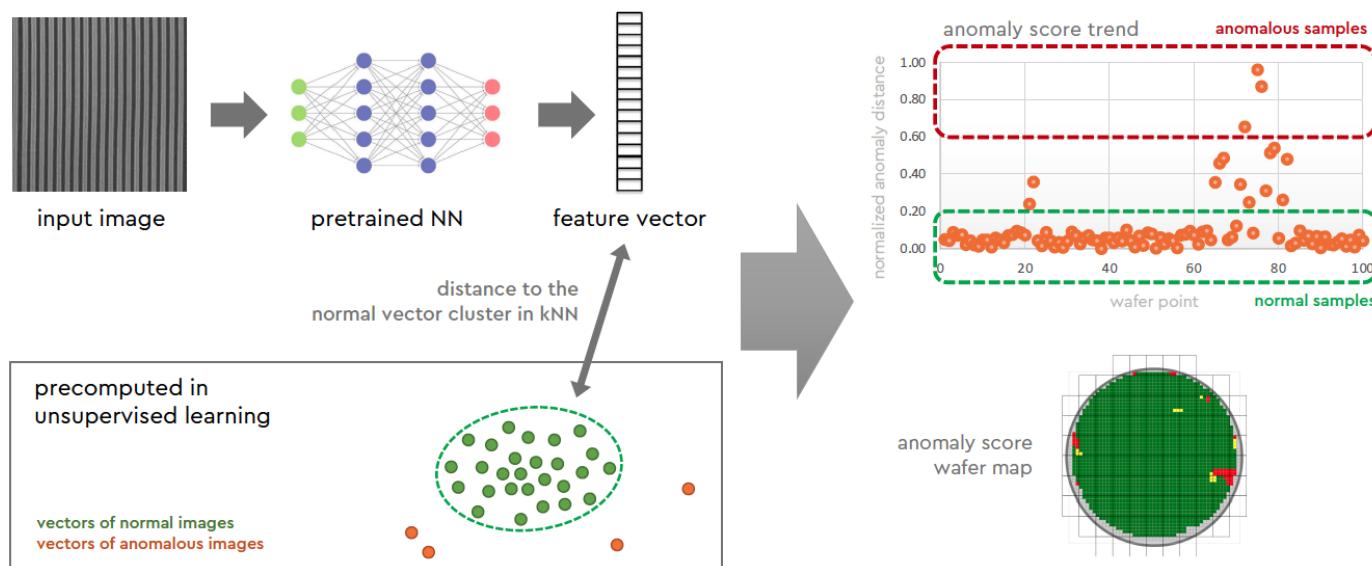
Image segmentation

- 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



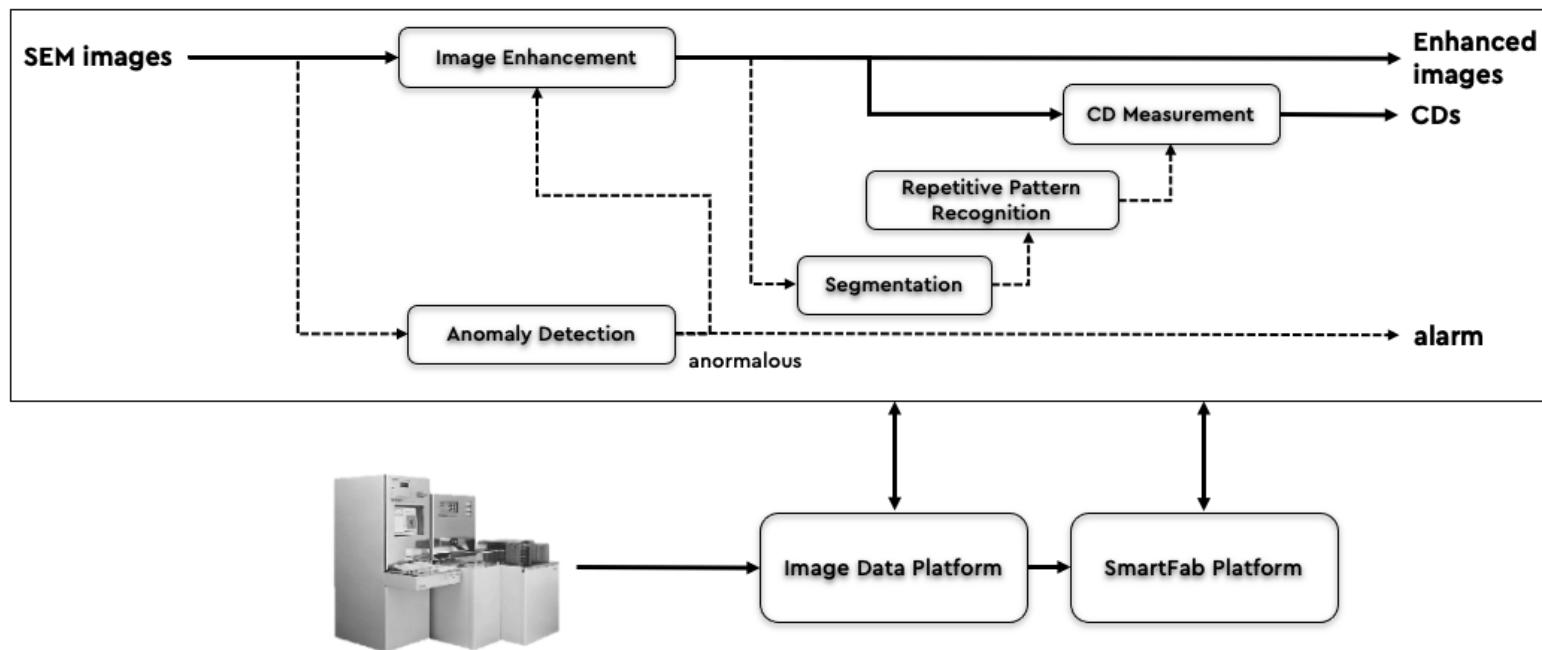
Anomaly detection using side product

- 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



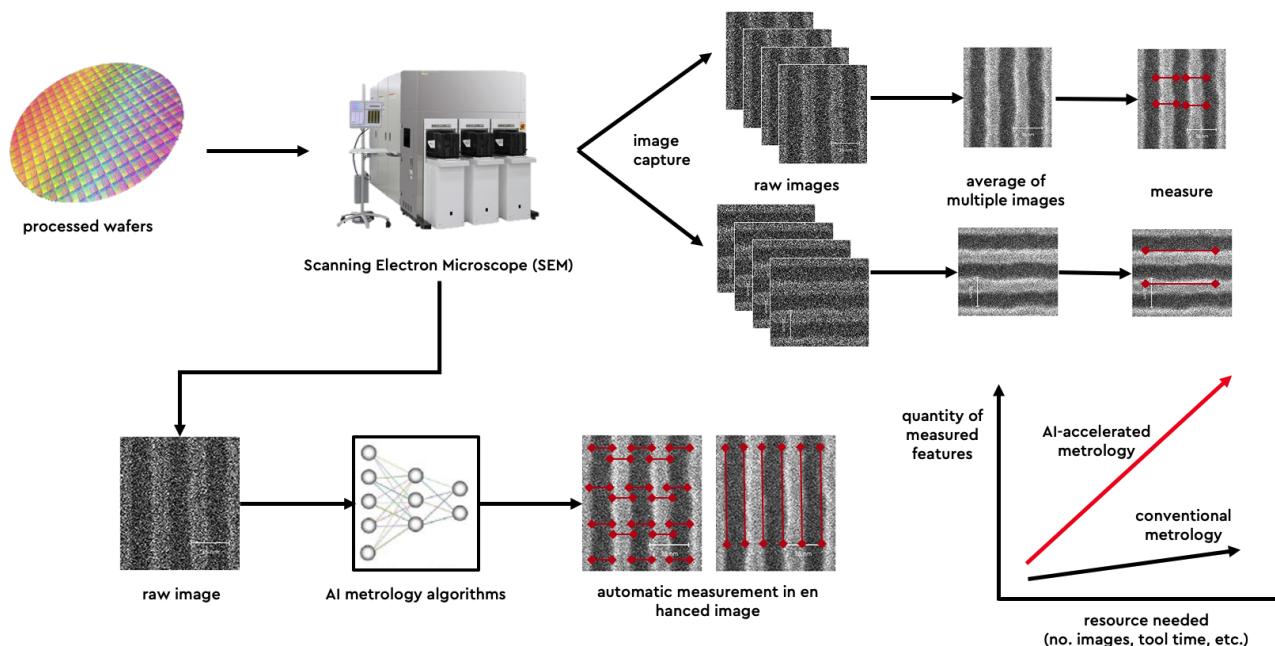
AI-enabled metrology system

- integration of separate components creates AI-enabled metrology system



Benefits of new system

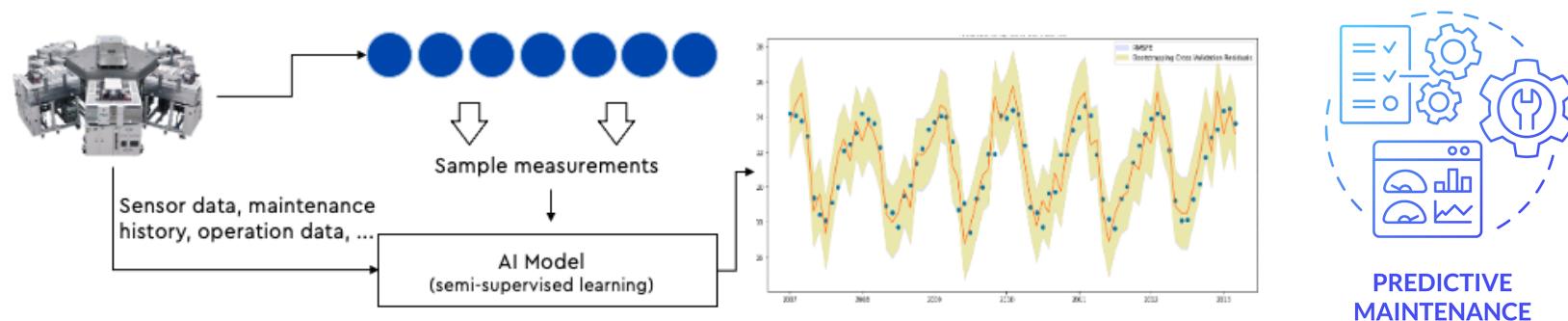
- new system provides
 - improved accuracy and reliability
 - improved throughput
 - savings on investment on measurement equipment



TS ML in manAI

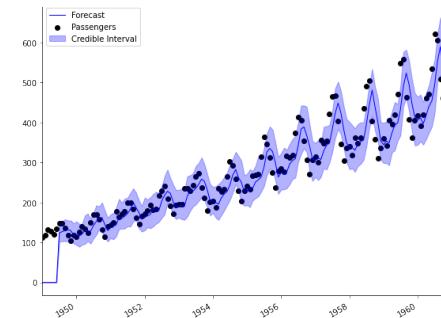
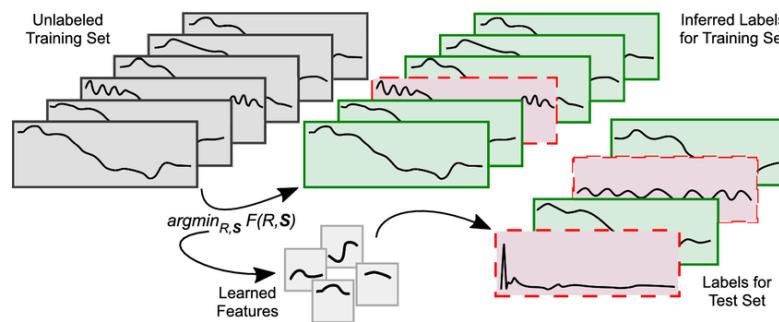
Time-series ML applications in manAI

- 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

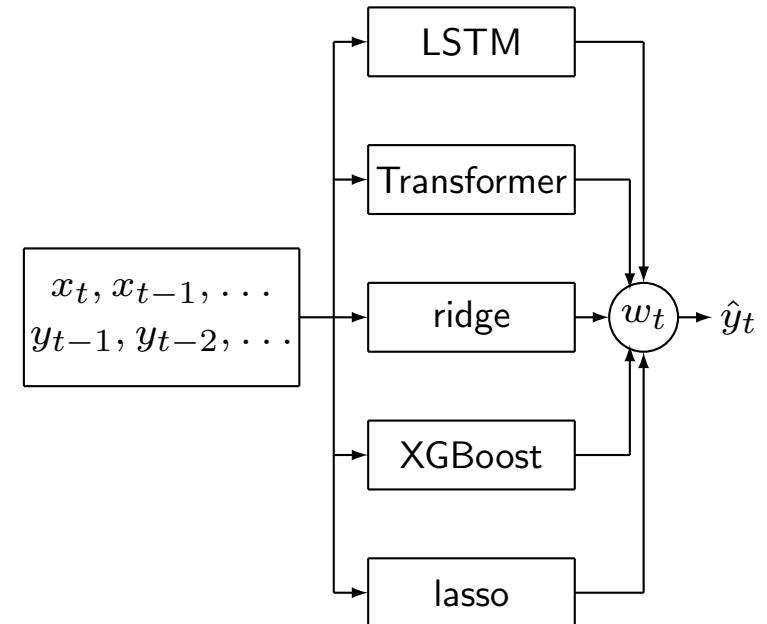


Difficulties with TS ML

- no definition exists for general TS data
- data drift & shift
 - $p(x_{t_k}, x_{t_{k-1}}, \dots)$ changes over time
 - $p(y_{t_k} | x_{t_k}, x_{t_{k-1}}, \dots, y_{t_{k-1}}, y_{t_{k-2}}, \dots)$ 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 multiple experts - $f_{1,k}, \dots, f_{p_k,k}$ 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 , $g_k : \mathbf{R}^n \rightarrow \mathbf{R}^m$ as weighted sum of experts



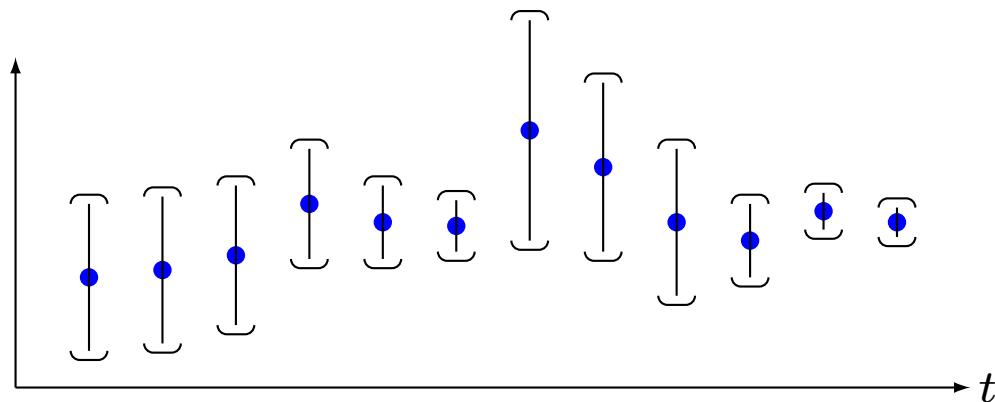
$$g_k = w_{1,k}f_{1,k} + w_{2,k}f_{2,k} + \cdots + w_{p_k,k}f_{p_k,k} = \sum_{i=1}^{p_k} w_{i,k}f_{i,k}$$

Credibility intervals

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

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

- 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 results are
 - sometimes *more crucial than algorithm accuracy*



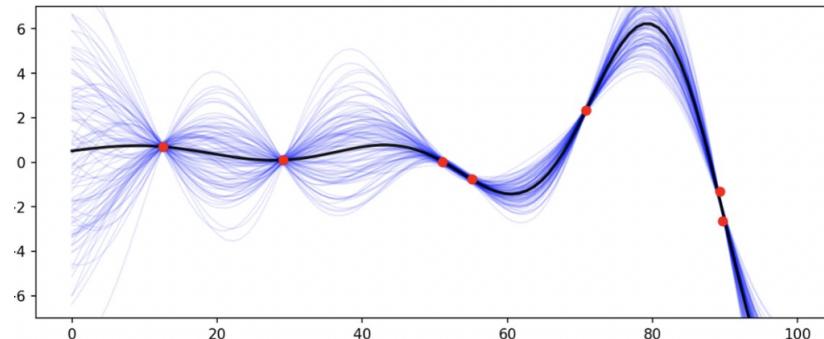
Bayesian approach for credibility interval evaluation

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

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

- 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

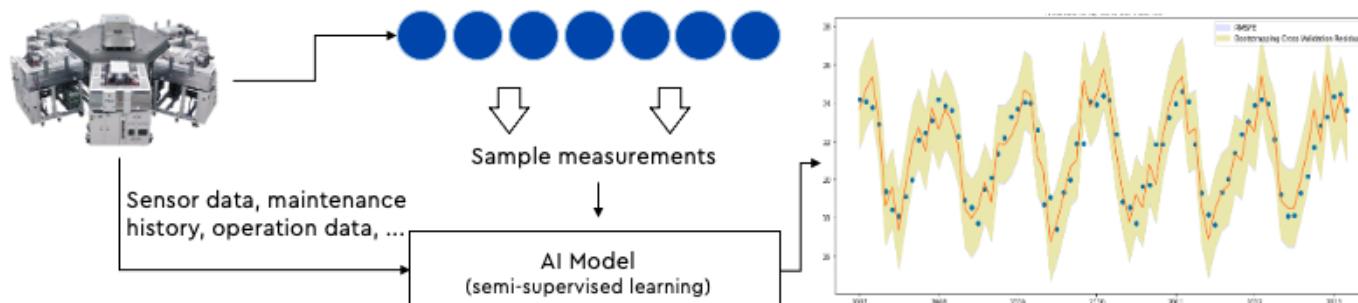
- problem description

(stochastically) predict y_{t_k}
 given $x_{t_k}, x_{t_{k-1}}, \dots, y_{t_{k-1}}, y_{t_{k-2}}, \dots$

- our problem formulation

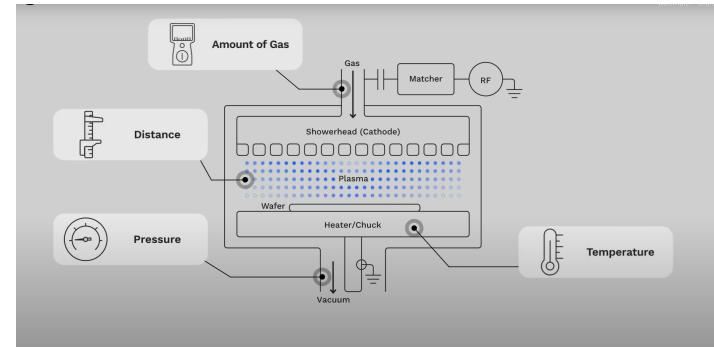
$$\begin{aligned} & \text{minimize} && \sum_{k=1}^K w_{k,K-k} l(y_{t_k}, \hat{y}_{t_k}) \\ & \text{subject to} && \hat{y}_{t_k} = g_k(x_{t_k}, x_{t_{k-1}}, \dots, y_{t_{k-1}}, y_{t_{k-2}}, \dots) \end{aligned}$$

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



VM - Gauss Labs' inAI success story

- Gauss Labs' ML solution & AI product
 - fully home-grown online TS adaptive 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



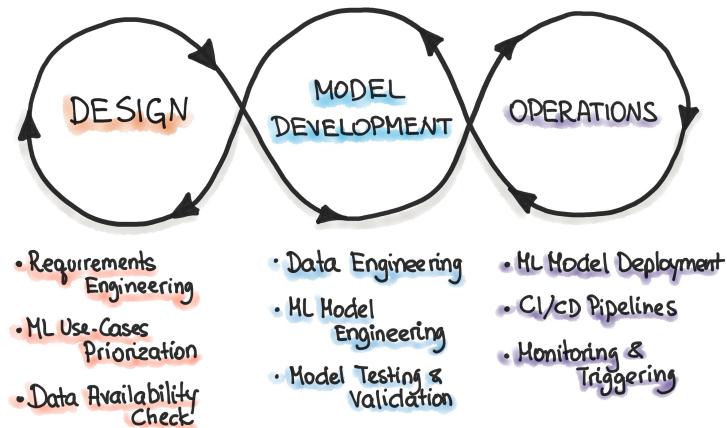
Manufacturing AI Productionization

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

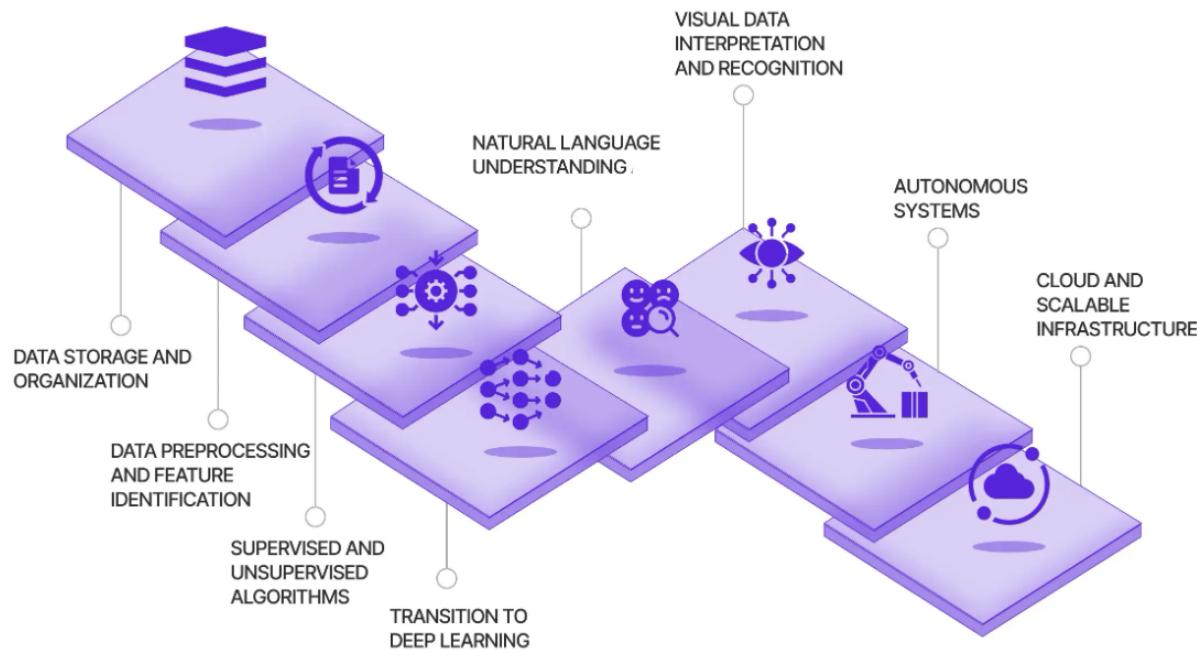
- environment for flexible and agile exploration - EDA⁵
- fast & efficient iteration of algorithm selection, experiments, & analysis
- correct training / validation / test data sets critical!
- seamless productionization from, e.g., Jupyter notebook to production-ready code
- monitoring, *right* metrics, notification, re-training



⁵EDA - exploratory data analysis

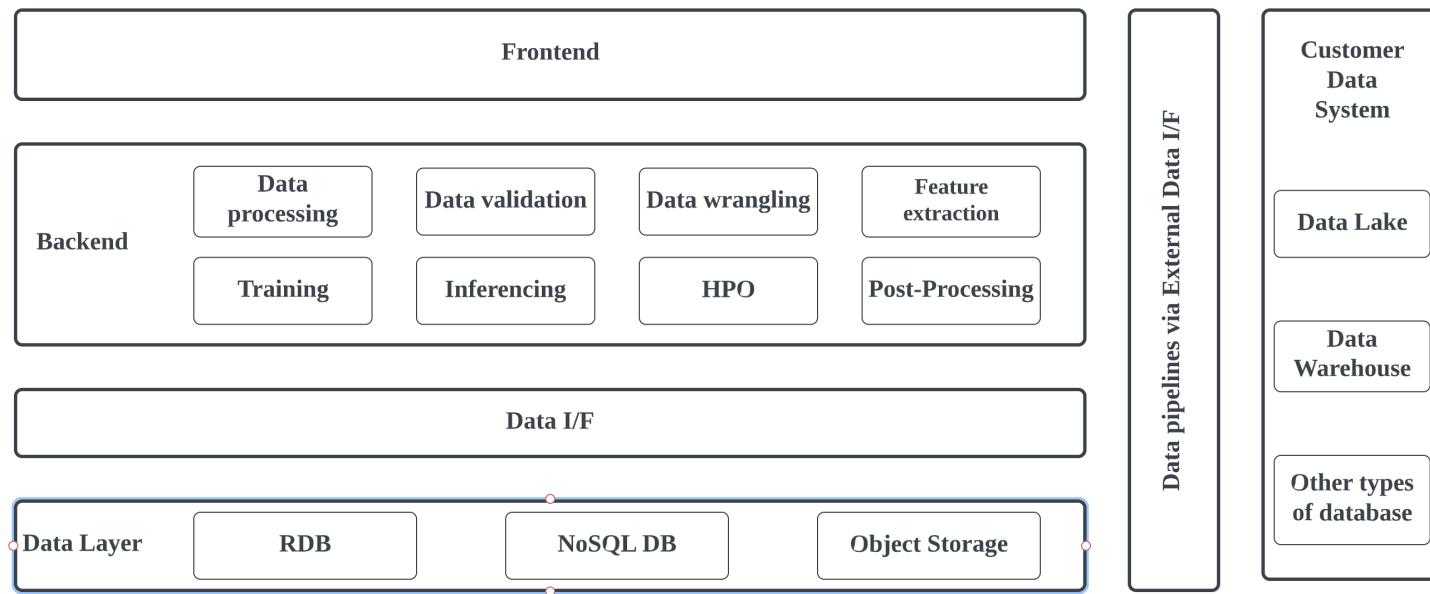
manAI software system

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



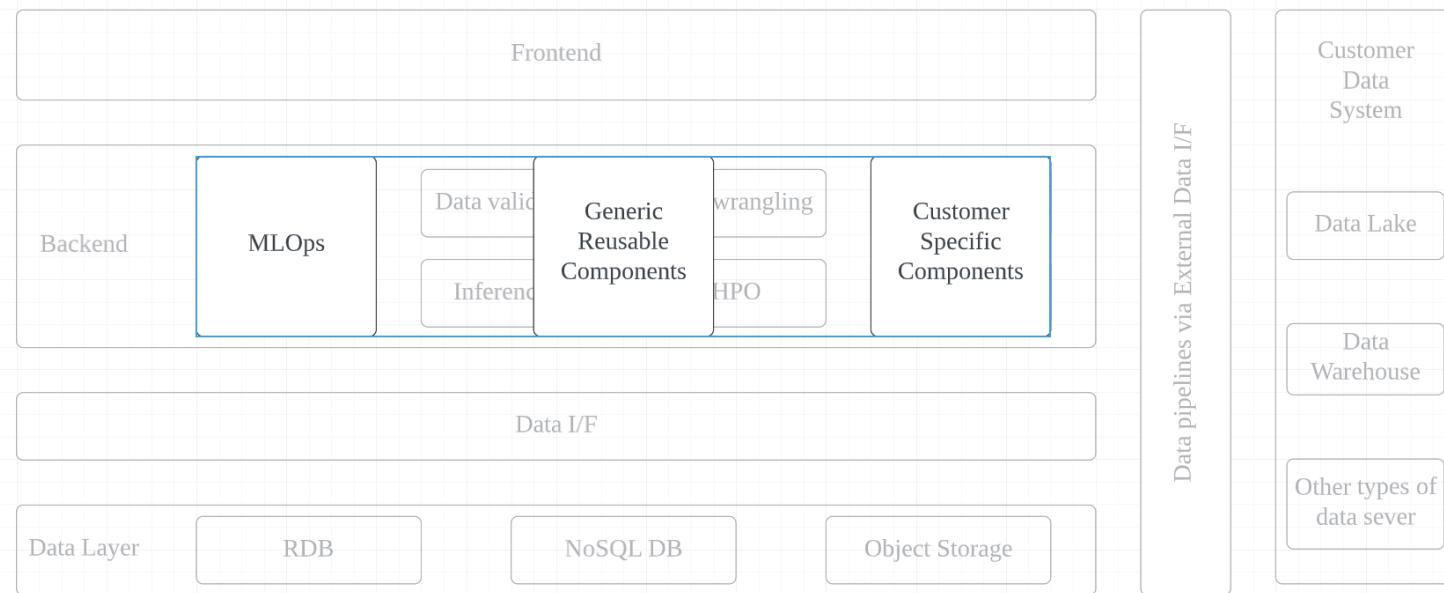
manAI system architecture

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



Reusable 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 crucial – 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, preprocessing, monitoring, notification, and retraining
 - data latency, availability, and reliability
 - excellency in software platform design and development using cloud services

Thank You