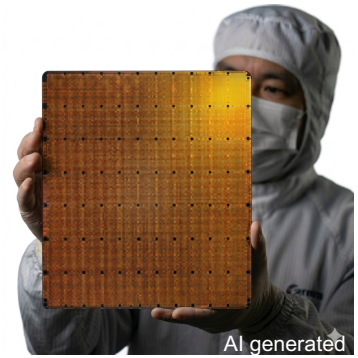
ECE8803-AI4
Lecture 1 – Introduction to AI for Semiconductors

Asif Khan
Associate Professor
School of Electrical and Computer Engineering
School of Materials Science and Engineering
Georgia Institute of Technology
akhan40@gatech.edu

Three Relentless Curves: Logic, DRAM & NAND

THE THREE PILLARS OF THE DIGITAL WORLD — EACH ITS OWN EXPONENTIAL

Logic



Cerebras WSE-3

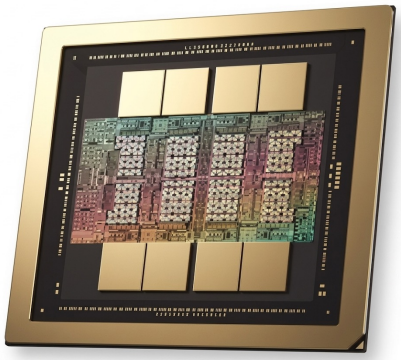
TRANSISTOR COUNT

1T+

BANDWIDTH

44 GB of on-chip Memory

DRAM and High Bandwidth Memory



NVIDIA Rubin

CAPACITY

288 GB

BANDWIDTH

22 TB/s

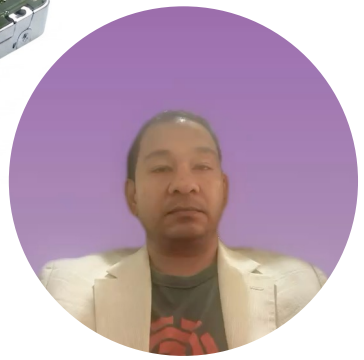
NAND Flash Storage

CAPACITY

30 – 245 TB

BANDWIDTH

~14 GB/s



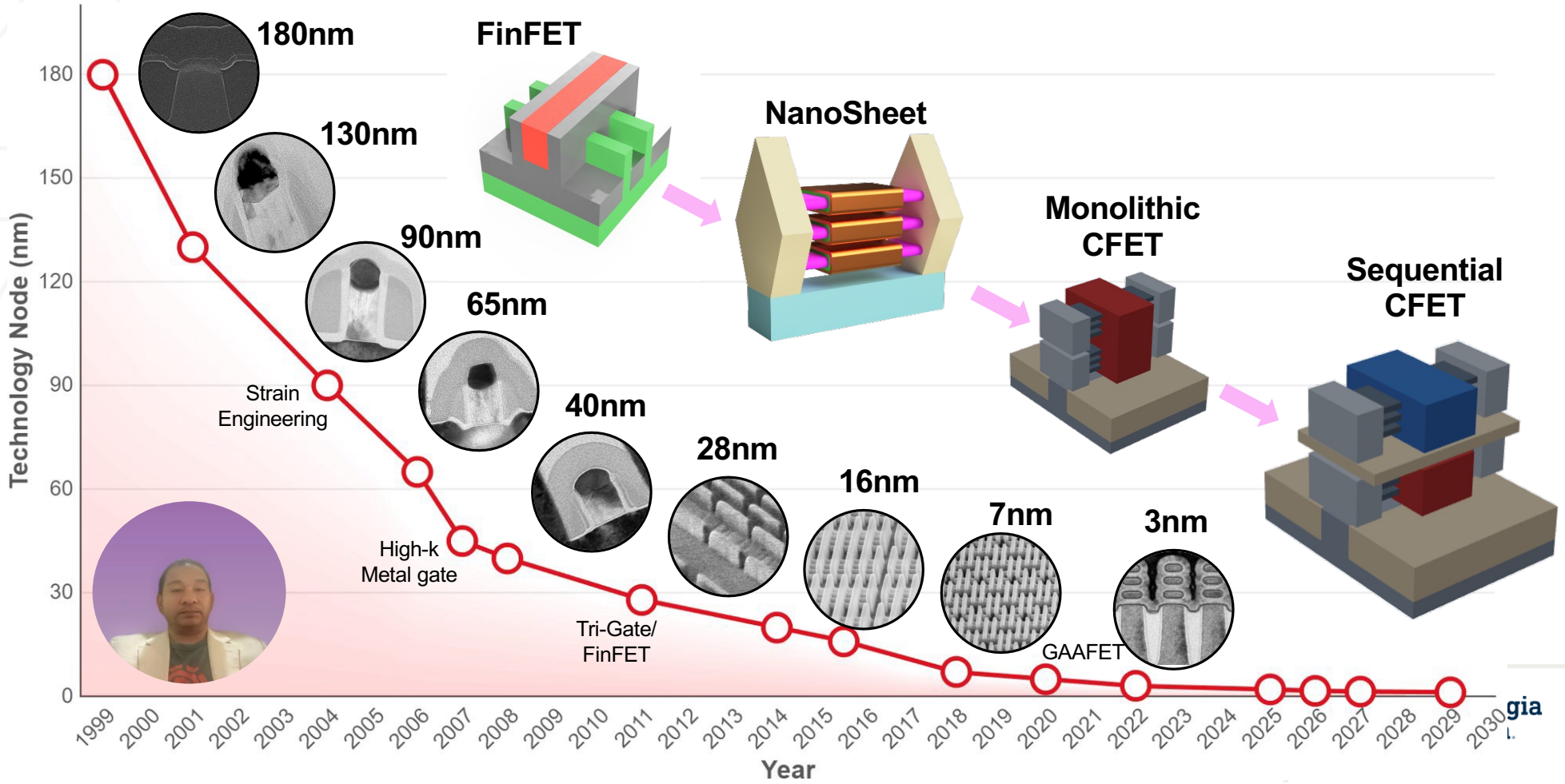
Three different physics, three separate exponentials — and every bend created a brand-new way to build and to measure. That is the industry this course is about.

Sources: NVIDIA (Blackwell, 208B); SK hynix (321-layer NAND, 2025); HBM3e (~1.2 TB/s per stack).

Georgia Tech

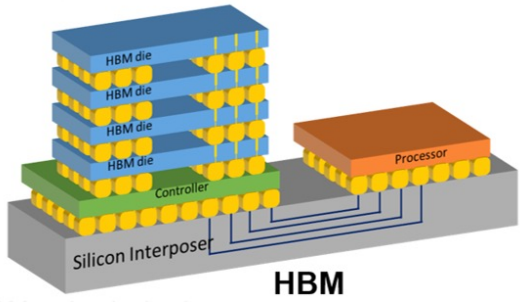
Courtesy: Taeyoung Song

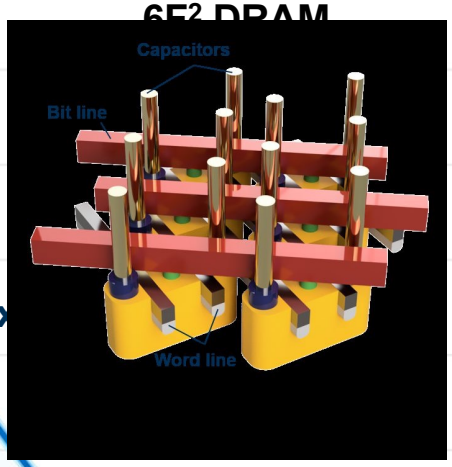
Logic Technology Roadmap

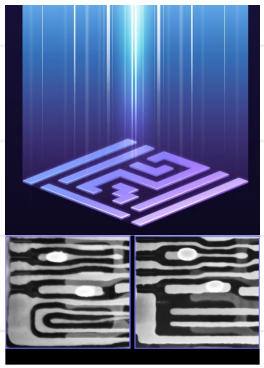


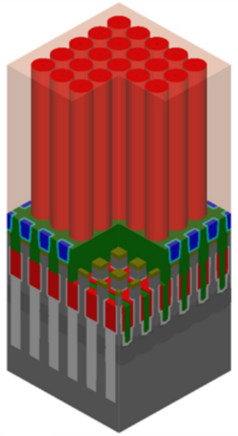
DRAM Technology Roadmap

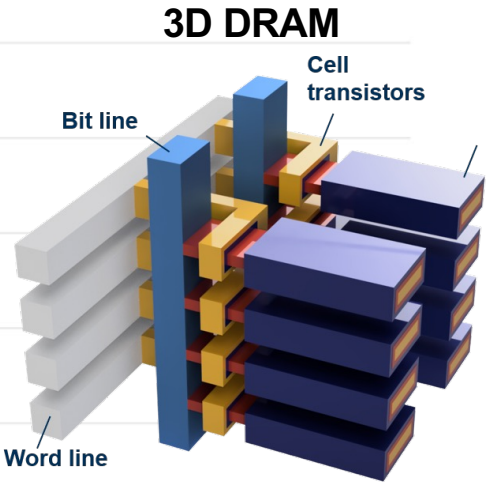
Courtesy: Taeyoung Song







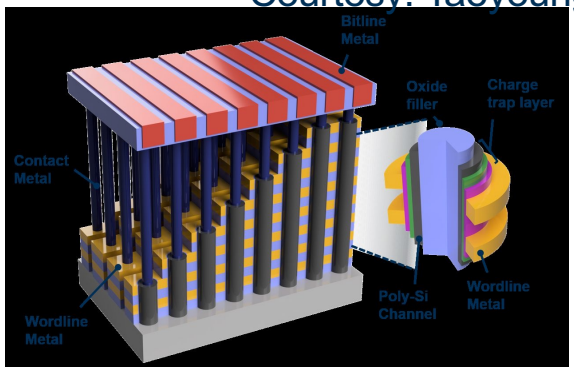


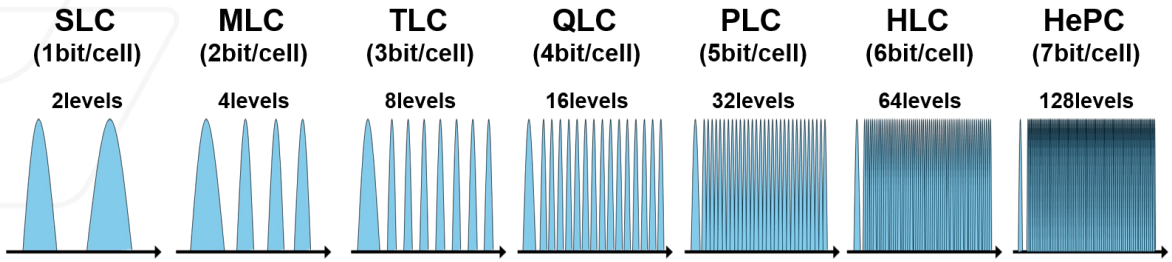


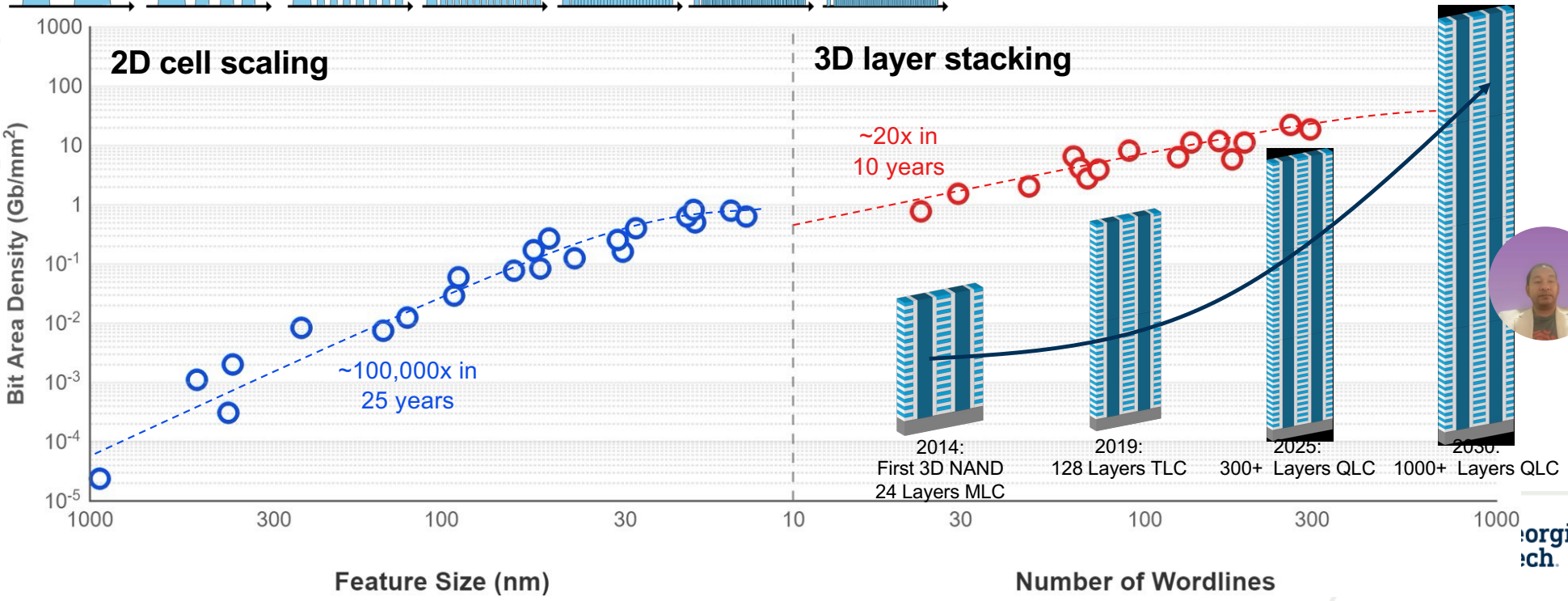
eorgia
ech

NAND Technology Roadmap

Courtesy: Taeyoung Song

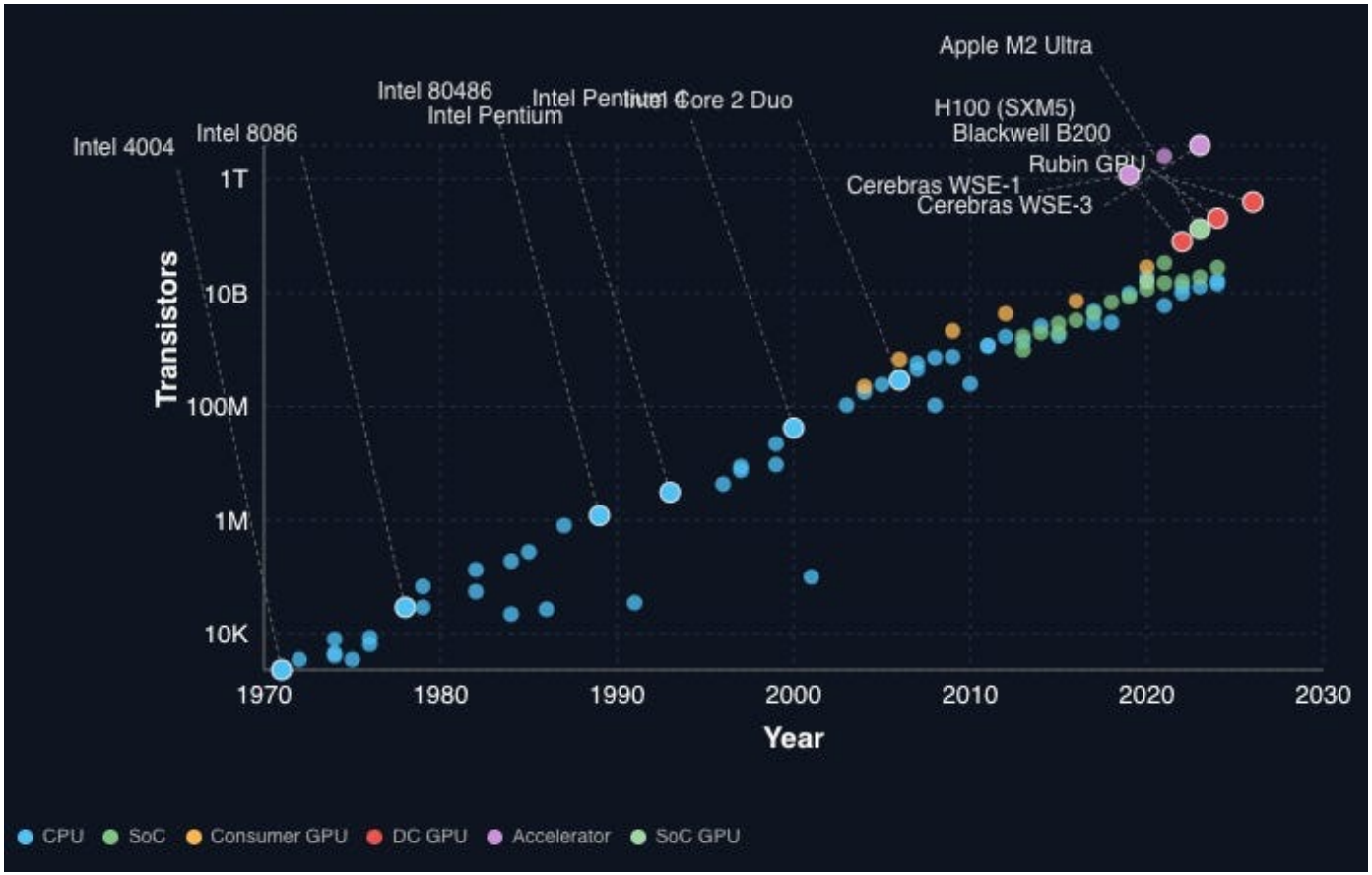






The Greatest Exponential in Industrial History

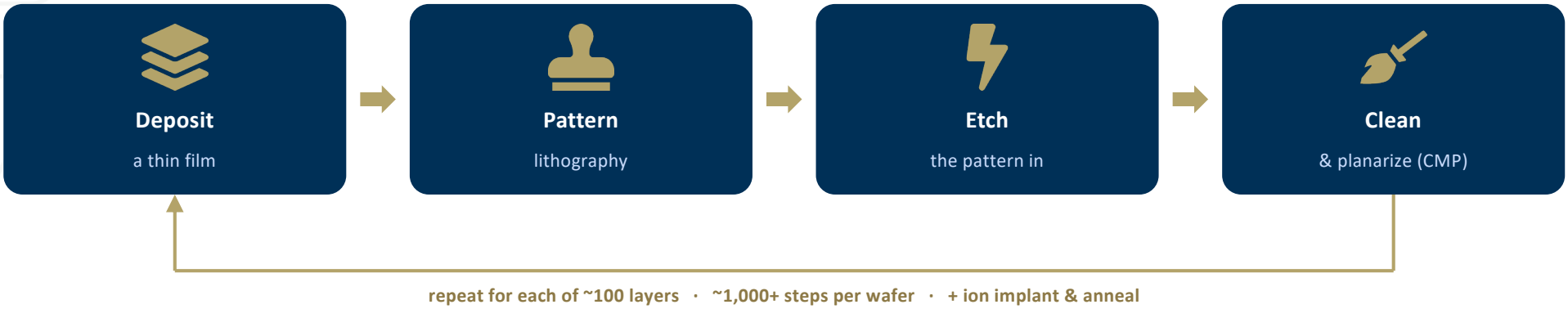
FIFTY YEARS OF SHRINKING — MICRONS TO ÅNGSTRÖMS



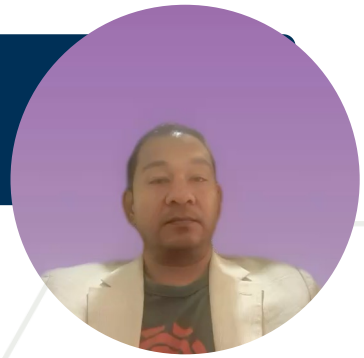
Building the Chip: Deposit, Pattern, Etch — Repeat

THE OTHER HALF OF THE FAB — THE PROCESSES THAT ACTUALLY BUILD IT

A chip is built upward in ~100 layers. Each layer is one turn of the same loop — **and that loop runs over a thousand times per wafer.**

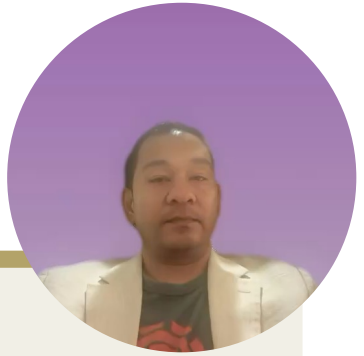


Etch & deposition are ~half of all steps — plasma processes so nonlinear that no fast first-principles model exists.
So they are developed and tuned from wafers, empirically — and that **data-driven process modeling is where our story begins.**



Semiconductor front-end fabrication — the layer-build loop.

The Processes You'll Meet This Semester



WHAT WE'LL MODEL — THE STEPS THAT BUILD THE CHIP

Plasma etch

Carve the pattern by removing material with a reactive plasma. Our running example all term.

Deposition (CVD · PVD · ALD)

Grow ultrathin films — sometimes one atomic layer at a time.

Lithography

Print the circuit pattern with light — now EUV at 13.5 nm.

CMP

Polish each layer perfectly flat before the next one is added.

Ion implant & anneal

Fire in dopant atoms to control where the silicon conducts.

Each runs hundreds of times per wafer and floods out sensor data — the raw material the machine-learning models in this course learn from.

Georgia Tech

How We See and Measure It

METROLOGY & INSPECTION — THE TOOLS WE'LL USE

CD-SEM

An electron microscope that measures feature width — the “critical dimension” — at the nanometer scale.

AFM

Atomic-force microscopy drags a tiny sharp tip across the surface to map its 3-D shape.

Scatterometry (OCD)

Shines light on a pattern and infers its shape from the reflection — fast and non-destructive.

Optical & e-beam inspection

Scan the whole wafer to find defects, wherever they are.

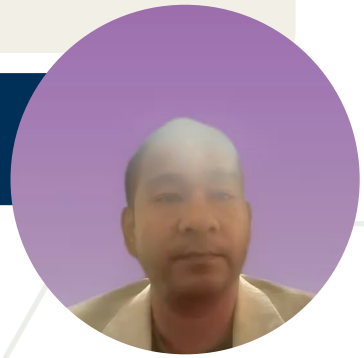
Ellipsometry

Measures film thickness and optical properties from how light’s polarization changes.

Virtual metrology

Machine learning predicts the measurement from sensor data when measuring every wafer is too slow.

You’ll build models on exactly this data — SEM images, CD measurements, and equipment sensor traces.



Four Decades of Making — and Measuring

AFTER IMEC · FAB METROLOGY & INSPECTION, 1986 → TODAY

1986–2005

"Happy scaling" — optical microscopy & ellipsometry; KrF 248 nm

2005–2011

Device innovation — strain, high-k / metal gate; X-ray (XRD, XPS); ArF 193 nm

2011–2019

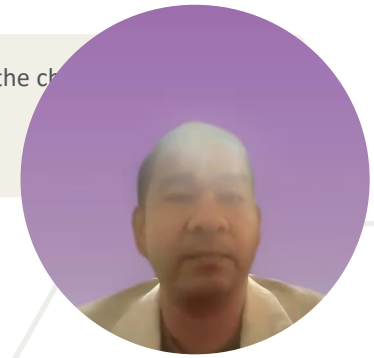
FinFET era — first 3D transistor; multipatterning; edge-placement error

2019–now

EUV era — 13.5 nm; stochastic defects; CD-SEM, AFM; AI-denoising

road ahead

Making and measuring advanced together: every new device (immersion, FinFET, EUV) demanded a new way to see it. The tools that build the ch
that inspect it are twins.



Source: imec, "Four decades of fab metrology and inspection: pivotal milestones and the road ahead" (imec-int.com).

The Road Ahead — and Now, AI

2026 AND BEYOND

WHAT'S COMING (THE PHYSICS)

- **High-NA EUV** (0.55 NA) for sub-2 nm nodes
- **Gate-all-around → CFET** transistors
- **3D DRAM & 3D NAND** — stacking upward
- **Backside power delivery** & curvilinear masks
- **High-aspect-ratio & multilayer** structures to measure

THE NEW TOOL IN THE BOX

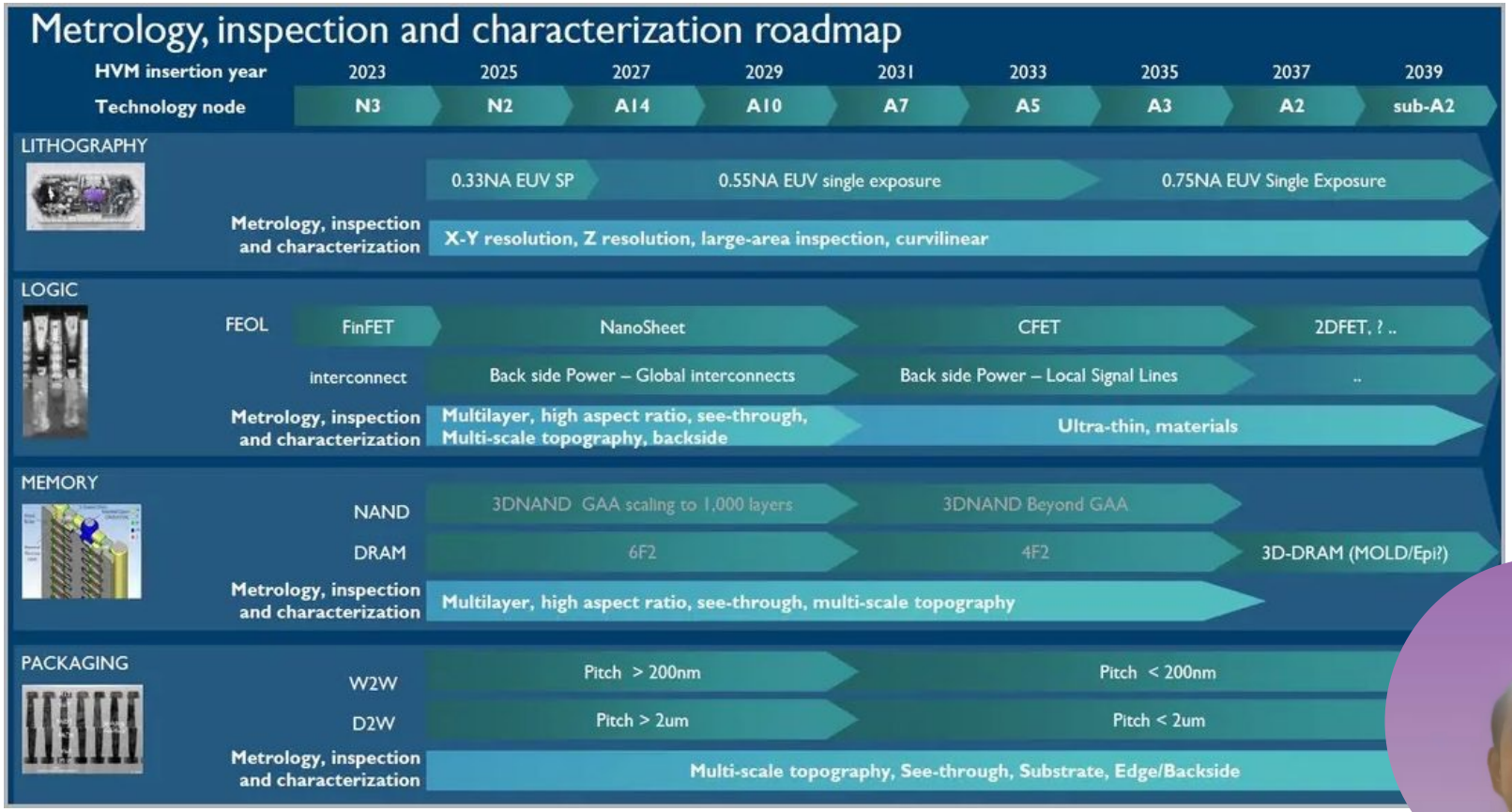
AI is now entering the fab itself

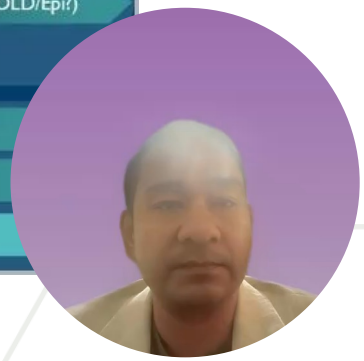
It already sits inside metrology and inspection — **AI-denoising, virtual metrology** — and, increasingly, across the whole manufacturing flow.

Which is exactly where this course begins.

Source: imec, "Four decades of fab metrology and inspection: pivotal milestones and the road ahead."

The Metrology & Inspection Roadmap — to 2039

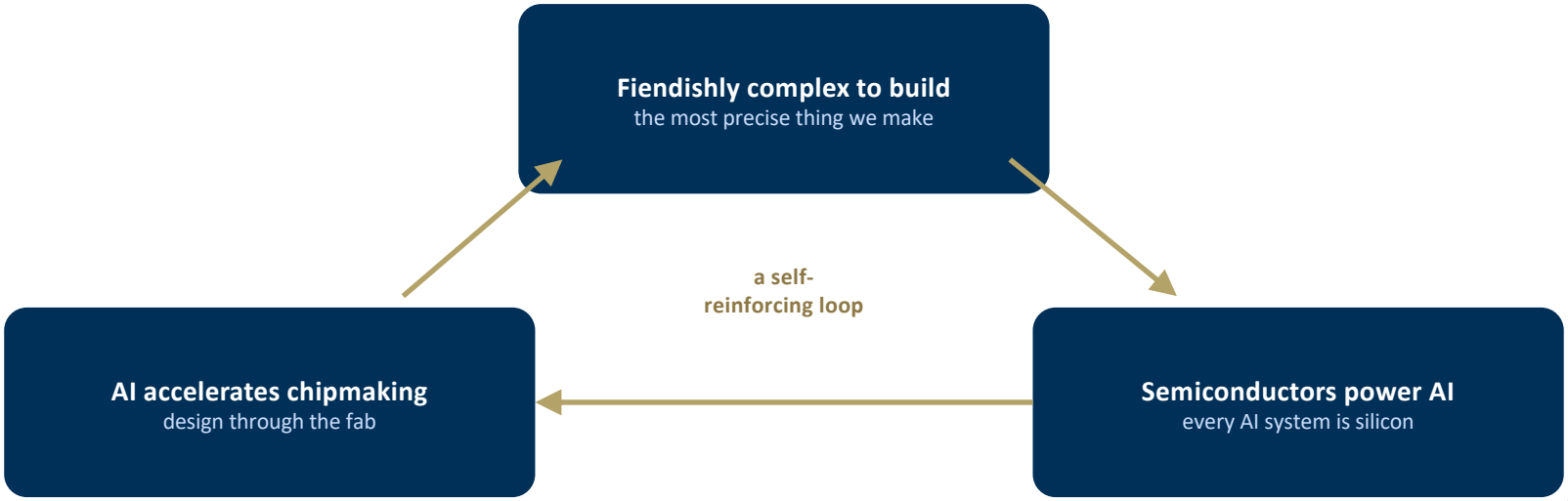




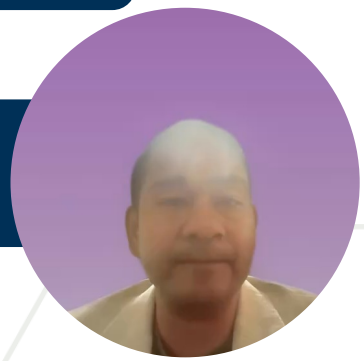
Source: imec, “Four decades of fab metrology and inspection: pivotal milestones and the road ahead” (imec-int.com).

Semiconductors & AI: A Two-Way Street

THE FRAME FOR EVERYTHING THAT FOLLOWS



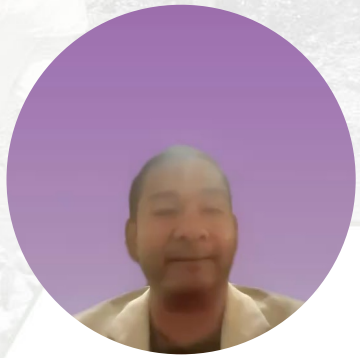
Today — Part A: where AI enters the fab (six trends + a Nature spotlight) · Part B: has the industry done AI all along?



Where AI Enters the Fab

A tour of six trends transforming semiconductor manufacturing

ECE 8803-AI4 · Lecture 1 · Section 5



Where AI Plugs Into the Manufacturing Flow

4 @ Lithography
AI process control & EUV defect prediction

1·2 @ Metrology / Inspection
Defect detection & virtual metrology

2 @ Test & Yield
Yield optimization & prediction

3·5·6

Cross-cutting: digital twins · predictive maintenance · AI-orchestrated "smart fab"
span every tool and step, not just one - the backbone of the modern fab

Six Trends: Where AI Enters the Fab

1 Defect detection & wafer inspection

Deep-learning vision flags defects in wafer maps and SEM / optical images at scale.

2 Virtual metrology & yield optimization

Predict every wafer's measurements from sensor data; tighten run-to-run control.

3 Digital twins of tools & fabs

Virtual replicas simulate tools and whole fabs for tuning and fleet matching.

4 AI process control in lithography (EUV)

ML for EUV defect prediction, optical proximity correction, overlay / CD control.

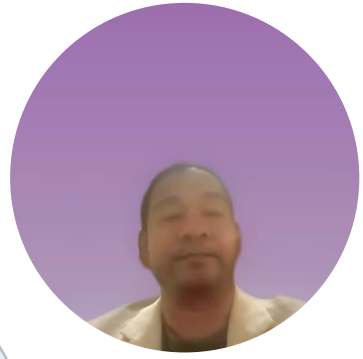
5 Predictive maintenance

Forecast tool failures from telemetry before an idle tool gates the line.

6 AI factories / smart manufacturing

Fully instrumented, AI-orchestrated fabs that tie all of the trends together.

Trend 1 · Defect Detection & Wafer Inspection



THE TASK — AUTOMATED DEFECT CLASSIFICATION (ADC)

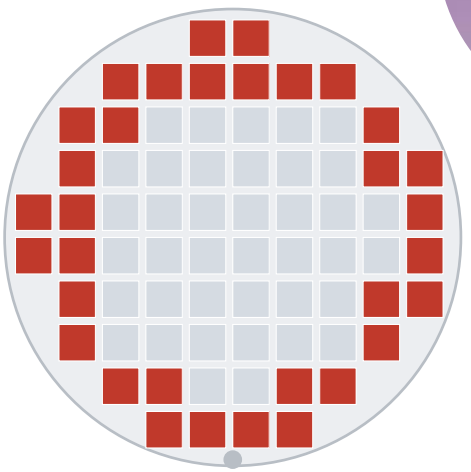
Inspect **every wafer** for defects — from wafer **bin maps** (die pass/fail) to **optical & SEM images**.

Detect and classify recurring **spatial defect patterns**, then trace each back to the process step that caused it.

WHY IT'S HARD

- Defects are tiny and rare → severe **class imbalance**
- Millions of features per wafer → 100% manual inspection is infeasible
- Human classification is slow and **varies operator-to-operator**

WAFER BIN MAP — "EDGE-RING" DEFECT PATTERN



Six canonical pattern classes (WM-811K benchmark):

- **Center**
- **Donut**
- **Edge-ring**
- **Scratch**
- **Random**
- **None**

Sources: link.springer.com/article/10.1007/s10845-026-02845-z robovision.ai/blog/ai-based-wafer-defect-inspection-an-accurracy-and-efficiency-boost

Georgia Tech

Trend 1 · Methods & Impact

FROM RULES TO DEEP LEARNING

1. **Image processing & statistics**
   Morphological / thresholding rules
2. **Classical machine learning**
   SVM, decision trees, k-means, DBSCAN
3. **Deep learning (today)**
   CNNs, transformers, GANs, attention

Coping with scarce, imbalanced defect data:
data augmentation & GAN-synthesized defects; focal / triplet loss; transfer learning.

WHY IT MATTERS

dozens of images / second

vs **~1 / second** for a skilled human inspector

- **Consistent** — removes operator-to-operator variation
- **Scalable** — enables near-100% in-line inspection
- **Faster root-cause** — patterns map to a process step

Open problems: unseen / unknown defect types and mixed-type patterns (several defects on one wafer).

Sources: link.springer.com/article/10.1007/s10845-026-02845-z robovision.ai/blog/ai-based-wafer-defect-inspection-an-accurracy-and-efficiency-boost

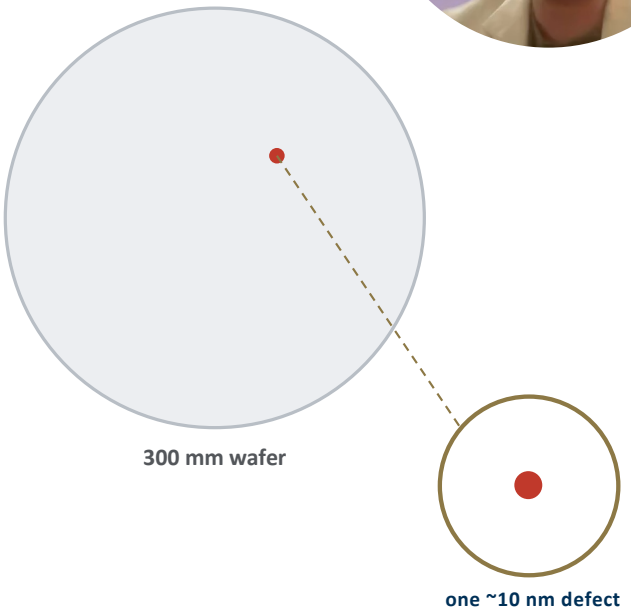
To Build at the Atomic Scale, You Must See It

METROLOGY & INSPECTION — THE UNSUNG HALF OF THE FAB

Every feature is measured and every wafer inspected — **far below the wavelength of visible light.**

The needle-in-a-continent problem

Finding one defect ~10 nm wide, anywhere on a 300 mm wafer, is like spotting a single **dinner plate somewhere across the entire United States.**



The toolbox: CD-SEM · atomic-force microscopy · scatterometry · e-beam · optical — increasingly combined (hybrid metrology).

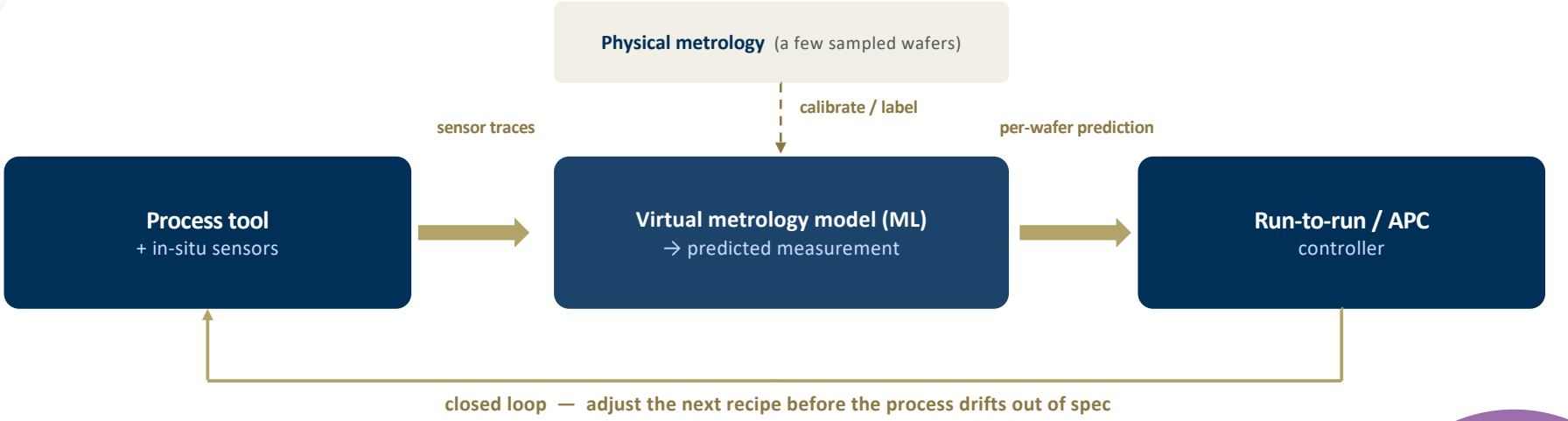
Source: After imec, "Four decades of fab metrology and inspection: pivotal milestones and the road ahead."

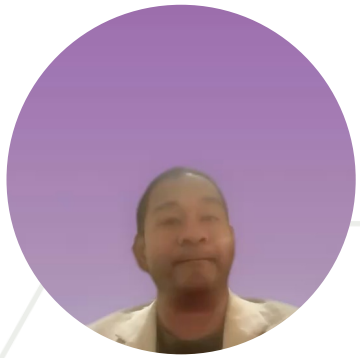
Georgia Tech

Trend 2 · Virtual Metrology & Yield Optimization

THE IDEA — MEASURE EVERY WAFER, VIRTUALLY

Physical metrology is slow and costly, so fabs measure only a **sampled fraction** of wafers. **Virtual metrology (VM)** trains an ML model to **predict each wafer's measurement** (film thickness, CD, overlay) directly from in-situ equipment sensor data — then feeds it into closed-loop control.





Sources: sciencedirect.com/science/article/abs/pii/S0957417410008365 sciencedirect.com/science/article/abs/pii/S0360835222003151

Trend 2 · Impact on Yield

WHY IT MATTERS — FROM MEASUREMENT TO YIELD

- **Every wafer measured (virtually)** → process drift is caught within one run, not after a sampling delay.
- **Tighter run-to-run control** → less variability, fewer out-of-spec wafers and less scrap.
- **Faster yield ramp** on new nodes → excursions are flagged before whole lots are lost.

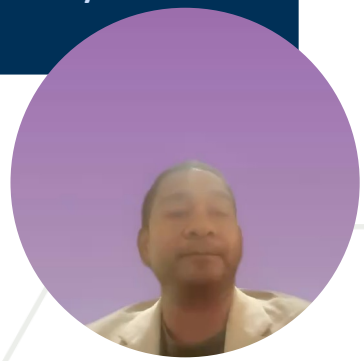
Caveat: a VM model is only as good as its sensor coverage — models drift and must be periodically recalibrated against real physical metrology.

~20% higher overall chip yield

40% fewer defects on advanced nodes

95% defect-detection accuracy

Vendor-reported case study (TSMC, via Indium).
Treat industry figures as indicative, not peer-reviewed.

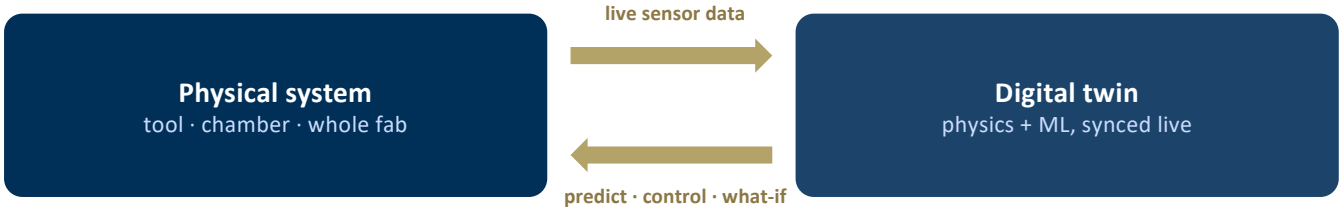


Sources: indium.tech/ai-advantage-semiconductor-fabrication-defect-detection-yield-optimization sciencedirect.com/science/article/abs/pii/S0360835222003151

Trend 3 · Digital Twins of Tools & Fabs

WHAT IS A DIGITAL TWIN?

A **live virtual replica** of a tool, process, or whole fab — a physics- and data-driven model kept **in sync with sensor data** and used to simulate, predict, and optimize before or alongside the real system.



Tool / equipment twin

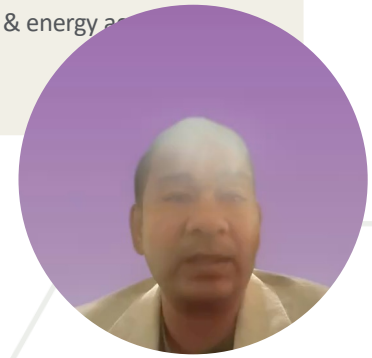
A single chamber or tool — e.g., a self-tuning etch chamber.

Process twin

Virtual fabrication of the process stack — e.g., SEMulator3D.

Fab / operations twin

Material flow, scheduling & energy a
whole line.



Sources: synopsys.com/blogs/chip-design/digital-twins-semiconductor-industry.html
semi.org/en/blogs/technology-and-trends/digital-twins-in-semiconductor-operations-insights-from-semi-workshop

Trend 3 · Digital Twins in Practice

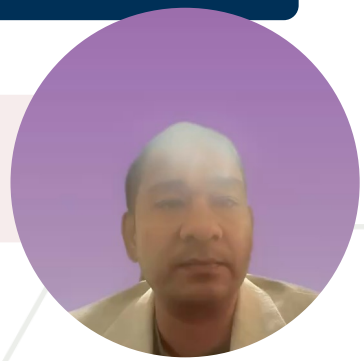
INDUSTRY EXAMPLES & PAYOFF

- **Lam Research** — Sense.i etch self-monitors and auto-adjusts via ML; virtual tool builds; SEMulator3D process twin.
- **Applied Materials** — EcoTwin — virtual equipment replicas for process development and sustainability analysis.
- **Intel** — AFS high-speed simulators for planning and scheduling across multiple fabs.
- **Bosch (Dresden fab)** — Fab twin simulates process changes and renovations without disrupting production.

4 months	faster tool development (Lam virtual build)
~30%	higher material-handling (AMHS) utilization
~50%	lower development cost (human-AI hybrid)

Figures reported by Lam Research and SEMI.

The hard parts (SEMI): standardized twin architecture · tool-to-tool interoperability · data security · messy, uneven factory data.



Sources: newsroom.lamresearch.com/how-digital-twins-can-revolutionize-chipmaking
semi.org/en/blogs/technology-and-trends/digital-twins-in-semiconductor-operations-insights-from-semi-workshop

Spotlight: Can AI Develop a Plasma Process?

KANARIK ET AL. · LAM RESEARCH · NATURE 616, 2023 — A HUMAN-VS-MACHINE BENCHMARK

Lam Research built a virtual **"process game"** — a calibrated plasma-etch simulator mapping a recipe (pressure, powers, gas flows, pulsing, temperature) to etch metrics and a cross-sectional profile — to benchmark **human engineers against Bayesian-optimization algorithms** on cost-to-target.

The obstacle it names: plasma processes are still developed **manually**, and algorithms struggle because experimental data is **scarce and expensive** — every wafer costs money.

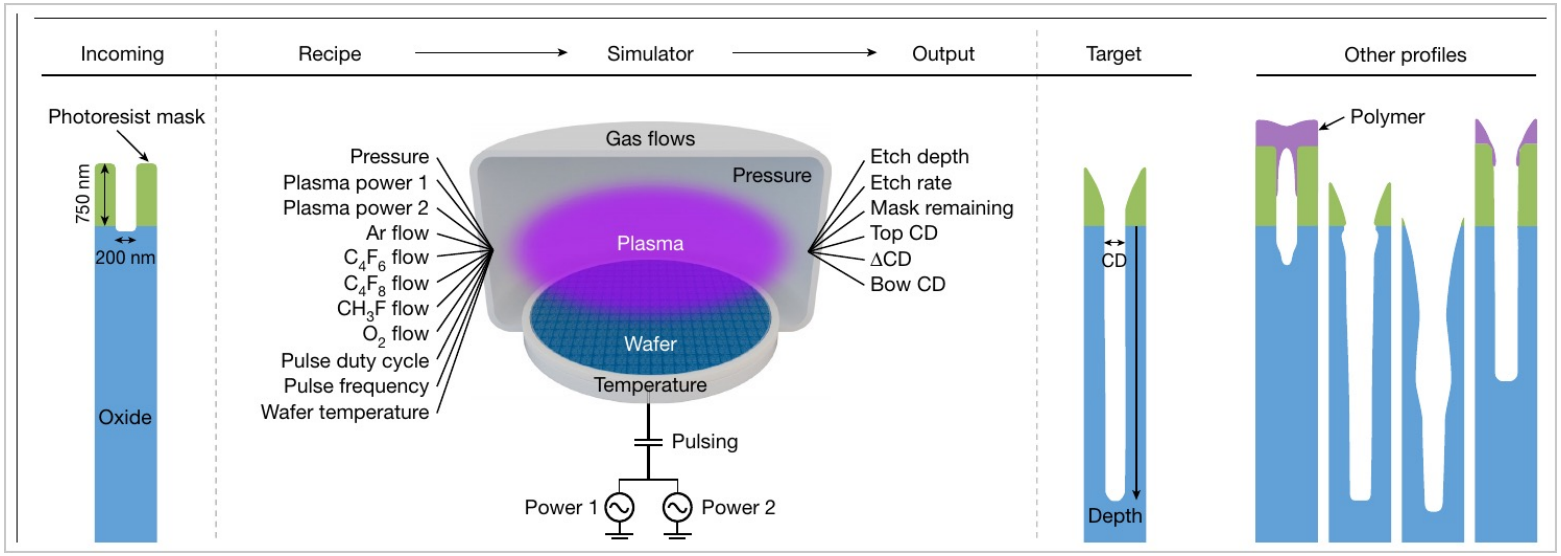
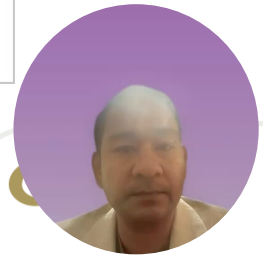


Fig. 1 — the virtual process game: recipe → plasma simulator → etched profile vs. target (Kanarik et al., 2023).



Source: K. J. Kanarik et al., "Human–machine collaboration for improving semiconductor process development," Nature 616, 2023.

The Result: Human-First, Computer-Last

HUMANS EXCEL EARLY · ALGORITHMS EXCEL NEAR THE TARGET

$105k → $52k

median cost-to-target: expert alone → human-first, computer-last

½

the cost-to-target of expert engineers working alone

- **Humans win early —** algorithms alone fail, wasting experiments exploring a huge space with no prior knowledge.
- **Algorithms win late —** they are far more cost-efficient near the tight tolerances of the target.
- **Hand off after rough tuning —** expert first, algorithm last (HF–CL) roughly halves cost-to-target.
- **Beyond the numbers —** the paper flags real cultural challenges in partnering humans with computers.

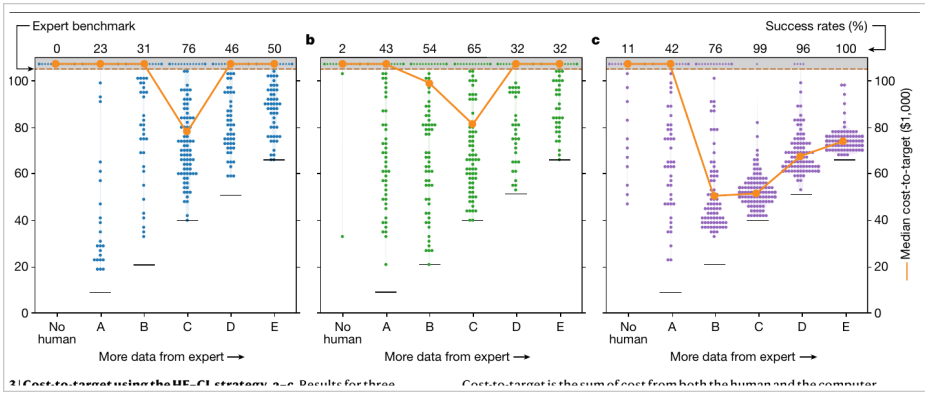
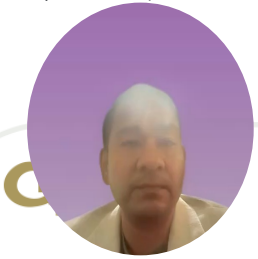


Fig. 3 — cost-to-target vs. how much rough-tuning the expert does first. The orange median dips to ~$52k (half the $105k expert benchmark).

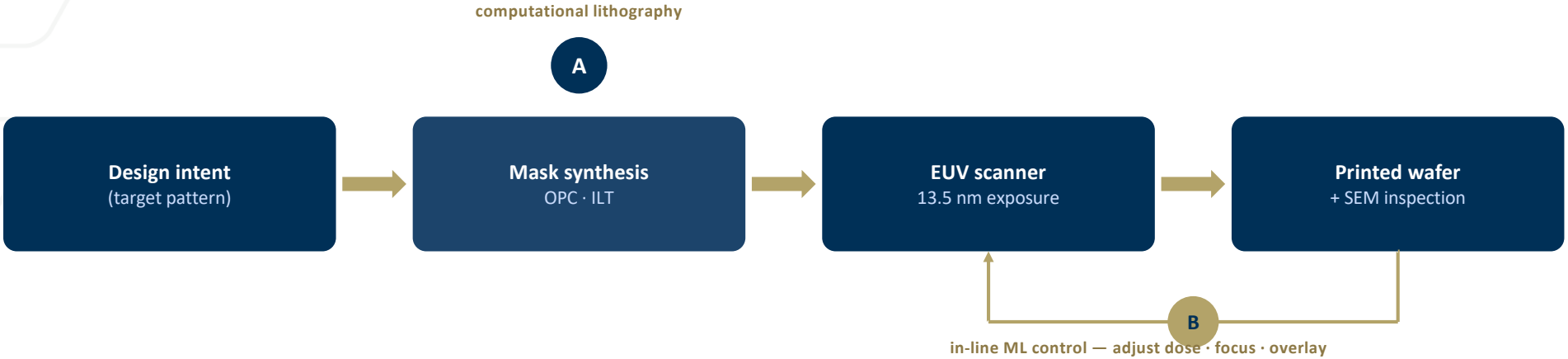


Source: K. J. Kanarik et al., Nature 616, 2023 (Fig. 3); Lam explainer: newsroom.lamresearch.com/humans-vs-artificial-intelligence

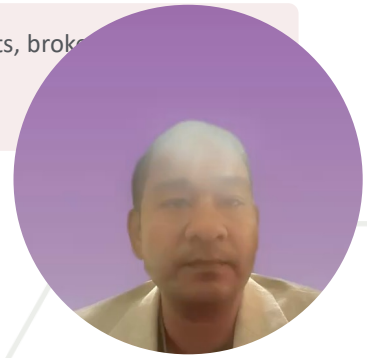
Trend 4 · AI Process Control in Lithography (EUV)

TWO PLACES AI ENTERS LITHOGRAPHY

Lithography prints the circuit pattern — the hardest, costliest step. At the leading edge (**EUV, 13.5 nm**), AI enters in two places:



Why EUV is hard — stochastic effects: random photon absorption & resist chemistry cause ppm-level random defects (bridges, missing contacts, broke
very tight process window.



Sources: semiengineering.com/finding-predicting-euv-stochastic-defects/ nature.com/articles/s41377-025-01923-w

Trend 4 · Computational Lithography & Defect Prediction

IN PRACTICE — TWO AI FRONTS, TODAY

A · Computational lithography at scale

NVIDIA cuLitho GPU-accelerates OPC / ILT mask computation; in production at TSMC & Synopsys, with ASML; generative AI now explored to go further.

B · EUV stochastic-defect prediction

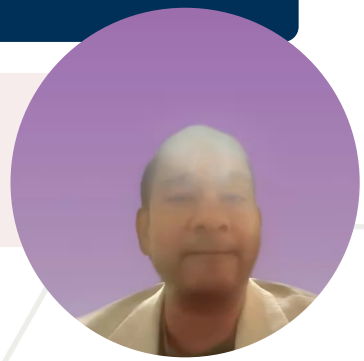
ML fuses SEM images, optical inspection & real-time scanner sensor data (supervised + anomaly detection) to flag ppm-level defects and drive real-time parameter correction.

40–60× faster mask computation (NVIDIA cuLitho)

In production TSMC + Synopsys, with ASML (2024)

ppm-level stochastic EUV defects targeted by ML

Caveat: mask compute is huge; stochastic defects are random, so prediction is probabilistic and needs large labeled SEM datasets.



Sources: nvidianews.nvidia.com/news/nvidia-asml-tsmc-and-synopsys-set-foundation-for-next-generation-chip-manufacturing
semiengineering.com/finding-predicting-euv-stochastic-defects/

Trend 5 · Predictive Maintenance

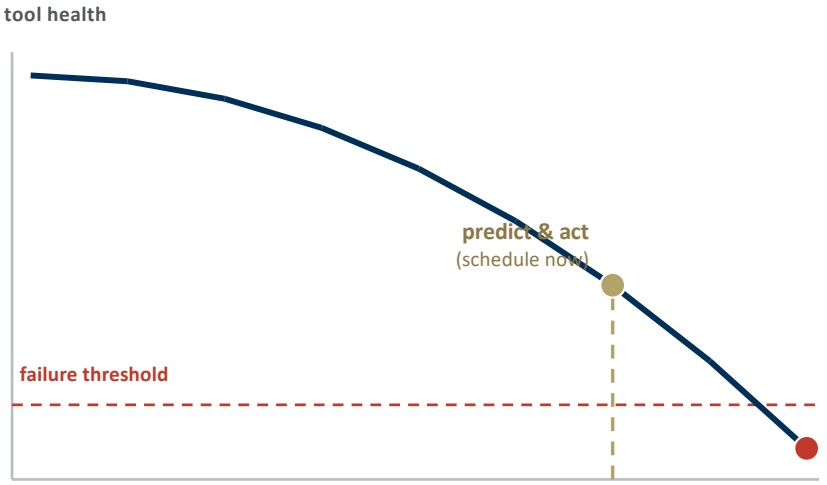
FROM FIX-IT-WHEN-IT-BREAKS TO PREDICT-AND-SCHEDULE

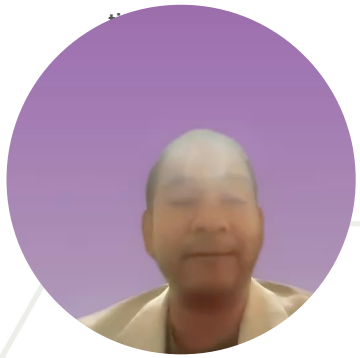
A single down tool can stall a whole line. ML reads **thousands of tool sensors** (RF power, gas flow, pressure, vibration) to predict failure and **remaining useful life**.

Reactive: fix after it breaks — worst downtime

Preventive: fixed calendar cycles — often too early or too late

Predictive: act exactly when data says failure is near





Sources: semiengineering.com/using-predictive-maintenance-to-boost-ic-manufacturing-efficiency/
sciencedirect.com/science/article/abs/pii/S0360835225003572

Trend 5 · Why It Pays

THE ECONOMICS OF UPTIME

3–4 hrs	**of unplanned downtime avoided per 1 hr of planned maintenance (McKinsey)**
>15%	**higher equipment availability → 70–80% OEE at bottleneck tools**
~$21M / yr	**saved per preventive-maintenance cycle avoided in an advanced logic fab**

Historical arc

Multi-chamber tools ran at **~30% OEE** 30 years ago; modern fabs exceed **80%** — data-driven maintenance is a big reason.

Caveat

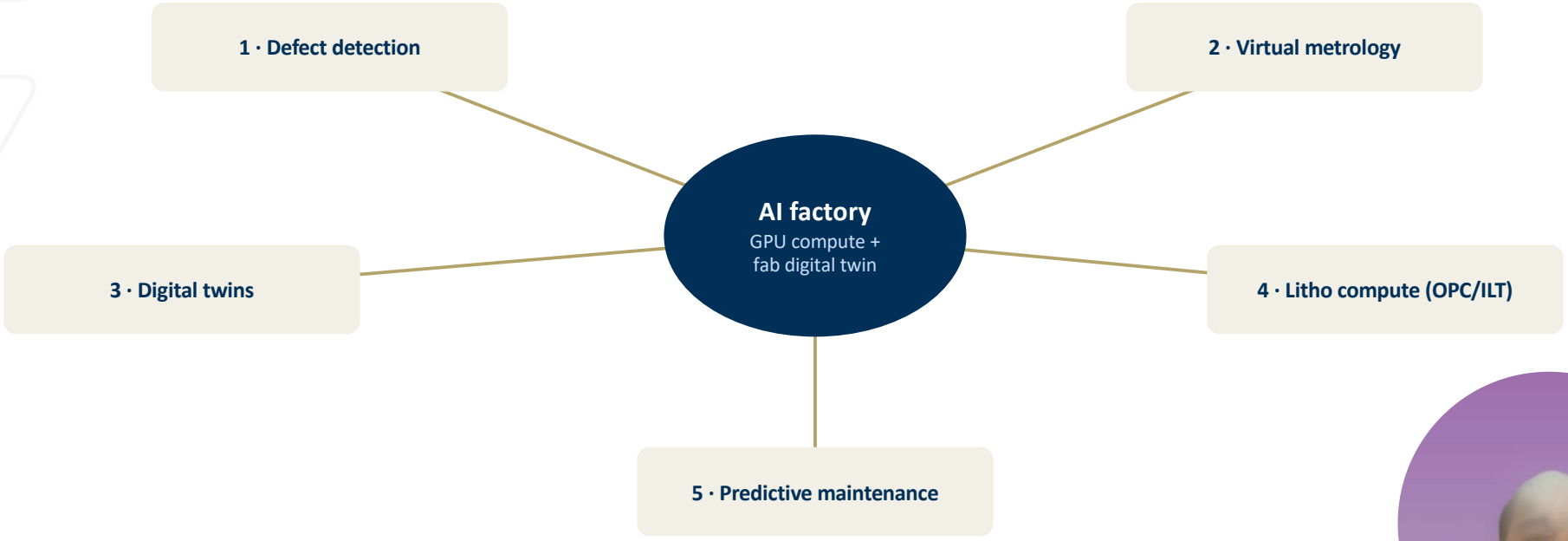
Needs rich sensor histories and labeled failures; every tool type needs its own model; and there is a real tradeoff between false alarms and missed failures.

Sources: semiengineering.com/using-predictive-maintenance-to-boost-ic-manufacturing-efficiency/
sciencedirect.com/science/article/abs/pii/S0360835225003572

Trend 6 · AI Factories & Smart Manufacturing

THE INTEGRATING VISION

An **"AI factory"** is the compute + software layer — big GPU infrastructure plus a whole-fab digital twin — that **hosts and connects every trend in this section**.



Sources: powerelectronicsnews.com/ai-driven-smart-manufacturing-in-the-semiconductor-industry/
nvidianews.nvidia.com/news/samsung-ai-factory

Trend 6 · In Practice — The AI Megafactory

NVIDIA × SAMSUNG, AND A BROADER SHIFT

NVIDIA × Samsung “AI megafactory”

50,000+ GPUs; NVIDIA Omniverse digital twins across global fabs for design-to-operations, predictive maintenance, real-time decisions and factory automation.

cuLitho inside Samsung’s OPC platform

Computational lithography — the fab’s heaviest workload — accelerated ~20×; plus GPU speedups for EDA (Synopsys, Cadence, Siemens).

Not just Samsung

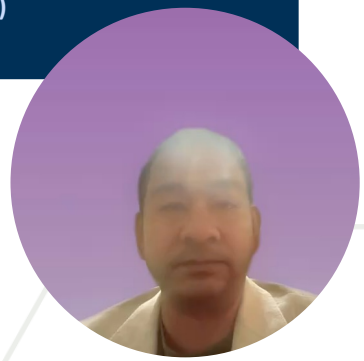
TSMC and SK hynix (via SK Telecom) are also building Omniverse fab twins — an industry-wide move.

50,000+ NVIDIA GPUs in Samsung’s AI megafactory

20× faster computational lithography (cuLitho in OPC)

Global fabs run as Omniverse digital twins (design → ops)

Vendor-reported (NVIDIA/Samsung, 2025).
Enormous capital & data-infrastructure required.

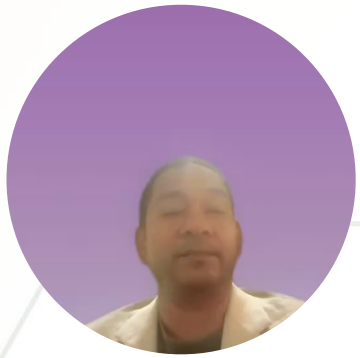


Sources: nvidianews.nvidia.com/news/samsung-ai-factory
rdworldonline.com/sk-telecom-puts-sk-hynix-fabs-into-an-nvidia-omniverse-twin-following-samsung-and-tsmc/

Has the Industry Been Doing AI All Along?

A short history of machine learning in semiconductor manufacturing

ECE 8803-AI4 · Lecture 1 · The claim is largely right — which is exactly why it needs unpacking.



Thirty Years of Shipping ML Into Fabs

NONE OF THIS IS NEW — PRODUCTION DEPLOYMENT LONG PREDATES THE CURRENT WAVE

Late 1980s–90s

Neural nets for process modeling, plasma-etch & fault diagnosis

1990s–2000s

Run-to-run control & APC — feedback loops with a learned model

2000s

Fault detection & classification; virtual metrology infers wafer state

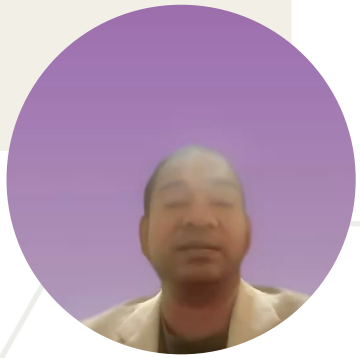
2000s–2010s

Computational lithography: OPC, ILT, source-mask optimization

Mid-2010s

CNNs for wafer-map defect classification at production volume

The through-line: for decades, "AI in the fab" meant deployed inference sitting inside a control loop — a learned model that replaced a slow simulation or a hand-built statistical model, one process step at a time. Real, valuable, and narrow.



Source(s): After A. Khan, ECE 8803-AI4 opening deck; canonical early work from G. S. May and co-workers (Georgia Tech).

The Earliest Reports: A Hybrid Model, 1991

NADI, AGOGINO & HODGES · UC BERKELEY · IEEE TRANS. SEMICONDUCTOR MFG. 4(1), 1991

The process — LPCVD of undoped polysilicon: predict **deposition rate, film stress and thickness** from pressure, temperature, flow-rate, wafer position and time.

THE ARCHITECTURE — A HYBRID OF TWO IDEAS

- **Influence diagram —** an expert encodes which inputs drive which outputs, splitting the process into small single-output subproblems. This is what lets it learn from less data.
- **Neural networks —** a feed-forward back-prop net learns each subproblem (deposition-rate net 4-5-1, stress net 5-5-1) plus a 7-7-7 associative-memory net holding expert recipe choices.
- **Synthesis —** run it backward with a stochastic optimizer (ALOPEX, ~simulated annealing) to generate new recipes.

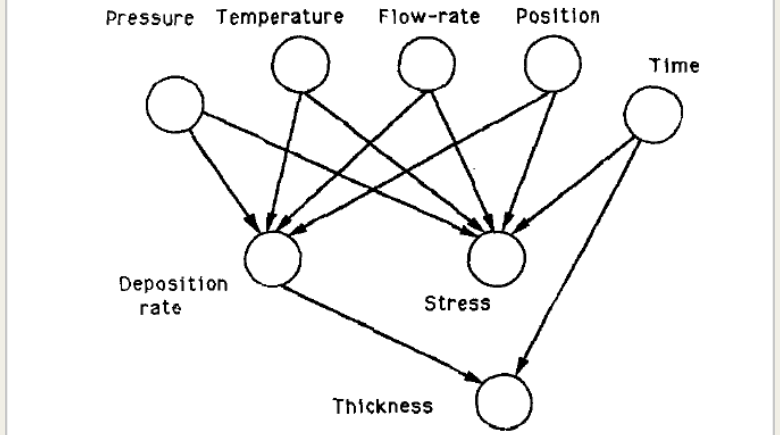
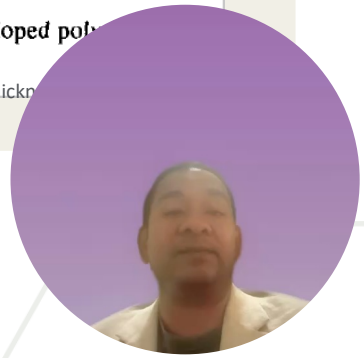


Fig. 2. Influence diagram for LPCVD of undoped pol

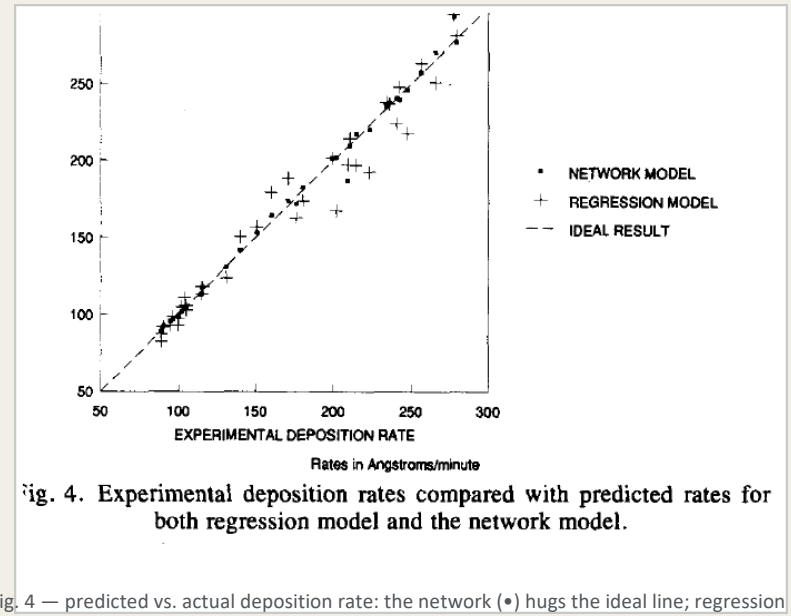
Fig. 2 — influence diagram: inputs → deposition rate, stress, thickn



Source(s): F. Nadi, A. M. Agogino & D. A. Hodges, IEEE Trans. Semiconductor Manufacturing, 4(1), 52–58, 1991 (Fig. 2).

Nadi et al., 1991: What They Showed

SAME DATA, HALF THE TRAINING SET — AND MORE ACCURATE



ig. 4. Experimental deposition rates compared with predicted rates for both regression model and the network model.

Fig. 4 — predicted vs. actual deposition rate: the network (•) hugs the ideal line; regression (+) scatters.

The baseline it beat: conventional models on the same data — regression + process physics for deposition rate; **pure statistical regression for stress** (no first-principle stress model existed).

1.4% vs 5.1%

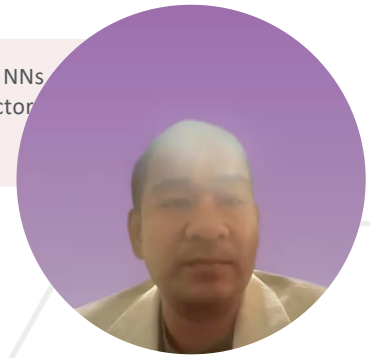
avg. error — deposition rate (network vs regression → ~3.6× lower)

13.2% vs 28.2%

avg. error — film stress (network vs regression → ~2.1× lower)

Plus — recipe synthesis: generated a zero-stress film recipe and a uniform-deposition-rate recipe (via a deliberately non-uniform temperature profile).

Their own caveat: based on only 12 runs; the authors stress NNs statistical models — those still bound variability and rank factor



Source(s): F. Nadi, A. M. Agogino & D. A. Hodges, IEEE Trans. Semiconductor Manufacturing, 4(1), 52–58, 1991 (Table I, Fig. 4).

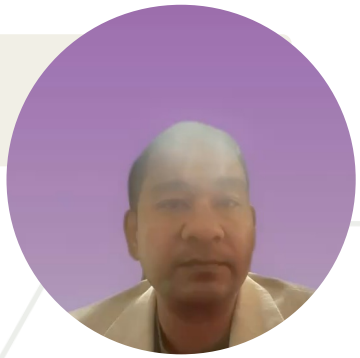
The Canonical Study: Georgia Tech, 1993

A NEURAL NETWORK PUT HEAD-TO-HEAD AGAINST THE METHOD IT WAS MEANT TO REPLACE

Himmel & May, "Advantages of plasma etch modeling using neural networks over statistical techniques," IEEE Trans. Semiconductor Mfg., 6(2), 1993.

- **The process —** removal of polysilicon films by CCl_4/He plasma etching (one chamber).
- **The baseline —** the same process had already been characterized with a designed experiment and an empirical response-surface model.
- **The test —** train a feedforward neural network on the same kind of data; compare the two directly.
- **The finding —** the neural network was more accurate and needed fewer training experiments — and wafers are the budget in a fab.

What it did not do: one process, one chamber, a handful of input knobs — the accuracy win did not survive contact with the next chamber.



Source(s): C. D. Himmel & G. S. May, IEEE Trans. Semiconductor Manufacturing, 6(2), 103–111, 1993.

The Problem, Stated Plainly

SIX KNOBS IN, FOUR NUMBERS OUT — AND NOTHING IN BETWEEN IS WRITTEN DOWN

SIX KNOBS YOU SET

- RF power
- Chamber pressure
- Electrode spacing
- CCl_4 flow
- He flow
- O_2 flow

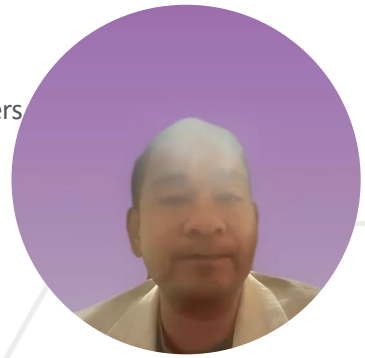
CCl_4 plasma etch
one chamber

strongly nonlinear · every knob interacts
no fast first-principles model exists

FOUR NUMBERS YOU NEED

- Etch rate
- Uniformity
- Selectivity
- Anisotropy

The deliverable is a map from recipe → outcome that runs in milliseconds. Theory can't supply it, so it must be **fitted from wafers** — and wafers



Source(s): Process characterization: G. S. May, J. Huang & C. J. Spanos, IEEE Trans. Semiconductor Mfg., 1991.

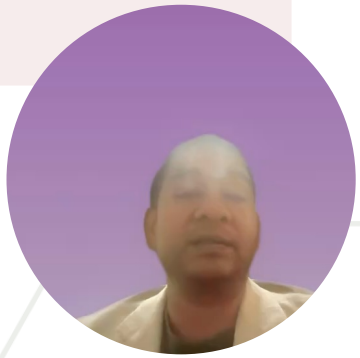
Before ML: Response-Surface Methodology

THE INCUMBENT — CHOOSE THE MODEL'S SHAPE FIRST, THEN RUN THE EXPERIMENT

1. Fix the six factors and their ranges before running anything.
2. Lay out a designed experiment: a 2^{6-1} fractional factorial, augmented with a Box–Wilson design to capture curvature.
3. Run the wafers; measure etch rate, uniformity, selectivity, anisotropy.
4. Fit a second-order polynomial in the six factors — one per response.

Where it runs out of room

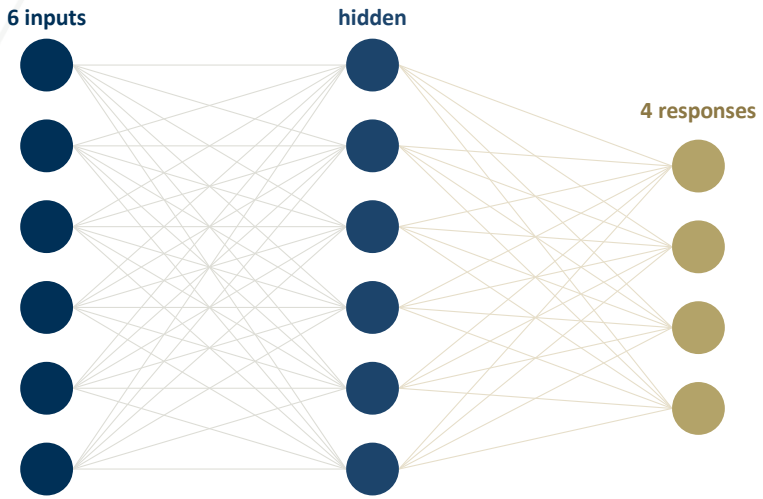
- The functional form is chosen before any data — quadratic terms plus a few interactions is the whole hypothesis space.
- Saturation and sharp thresholds become bias you can't remove by collecting more wafers.
- Widening a factor range means a whole new design and a new experiment.



Source(s): G. S. May, J. Huang & C. J. Spanos, "Statistical experimental design in plasma etch modeling," IEEE TSM, 1991.

After ML: A Back-Propagation Network

SAME WAFERS, SAME FOUR RESPONSES — BUT NOBODY WRITES DOWN THE SHAPE



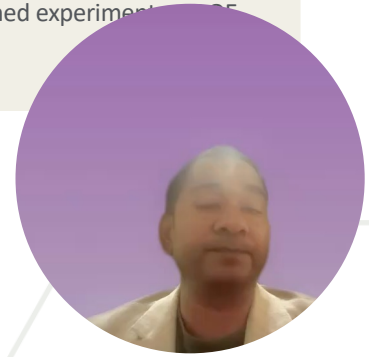
6-6-4 feed-forward, error back-propagation

WHAT CHANGED

- **No functional form is assumed —** interactions live in the weights, not in hand-written cross-terms.
- **Trained by back-propagation** on the same designed-experiment data.
- **The expensive resource shifts** from wafers to compute — and by 1993 that trade already paid.

The new problem it created: the network has its own knobs (learning rate, momentum, architecture) with no theory to set them.

Kim & May (1994) attacked that with a D-optimal designed experiment [illegible] turned on the model instead of the process.



Source(s): C. D. Himmel & G. S. May (1993); B. Kim & G. S. May, IEEE TSM (1994).

How Much Did It Actually Buy?

TWO CLAIMS FROM THE 1993 PAPER — ONE MATTERED FAR MORE COMMERCIALLY

More accurate

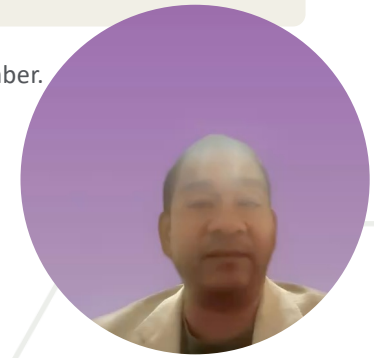
The neural models predicted all four etch responses better than the response-surface models built on the same data.

Fewer experiments

It reached that accuracy on **18 fewer runs** than the 53 used for RSM — in a fab, a run is a wafer plus tool time.

≈ 35 vs 53 experimental runs — a lower-error polysilicon etch-rate model on 18 fewer wafers. The RMS errors per response are tabulated in the paper.

Note the caveat that closes the paper: the accuracy win was real and the wafer saving was real — and neither survived contact with the next chamber.



Source(s): C. D. Himmel & G. S. May, IEEE Trans. Semiconductor Manufacturing, 6(2), 103–111, 1993 (see Table III).

What That “AI” Actually Was — and Its Limits

THE TRACK RECORD IS REAL; THE REGIME IT CAME FROM HAS HARD STRUCTURAL LIMITS

THE CLASSICAL REGIME

- Narrow supervised learning on well-instrumented, high-SNR sensor data.
- One model per chamber, per recipe, per process step.
- Tight closed loops against tight tolerances — the loop, not the model, carried the value.
- Retrained whenever the tool drifted or the campaign changed.

Why it never compounded

- New tool, new node, new fab → start over from scratch.
- Nothing transfers: no shared representation across steps or sites.
- Nothing composes: models can’t be queried, chained, or reasoned over.
- Process knowledge stays locked in artifacts no engineer can interrogate.

Deployment Era: Closing the Loop & Inferring State

1990s–2000s — LEARNED MODELS BECOME ROUTINE FAB INFRASTRUCTURE

Run-to-run & APC

1990s–2000s

A learned model sits inside a feedback loop and adjusts the next run's recipe. Advanced process control becomes standard practice — this is Trend 2's ancestor.

Fault detection & classification

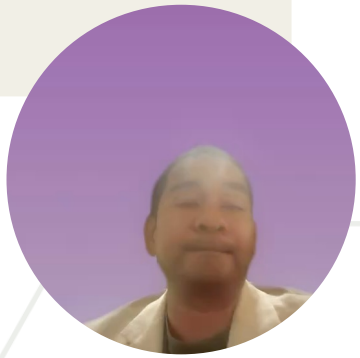
2000s

ML on tool sensor traces (RF power, gas flow, pressure) flags and classifies equipment faults — the direct ancestor of today's predictive maintenance (Trend 5).

Virtual metrology

2000s

Infer wafer state from equipment signals where you can't afford to measure every wafer — exactly the idea we saw in Trend 2, deployed twenty years ago.



Source(s): After A. Khan, ECE 8803-AI4 opening deck.

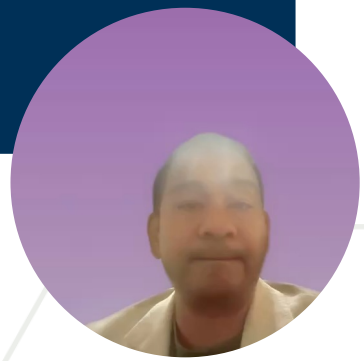
Deployment Era: Computational Lithography

2000s–2010s — ENORMOUS OPTIMIZED MODELS IN THE CRITICAL PATH TO TAPEOUT

Long before "generative AI," lithography already depended on massive learned and optimized computation to make a mask that prints what you drew. **This is the deepest, most mission-critical deployment of the classical era.**

- **Optical proximity correction (OPC) —** pre-distort the mask to counter diffraction.
- **Inverse lithography (ILT) —** solve for the mask that best prints the target pattern.
- **Source-mask optimization (SMO) —** co-optimize illumination and mask together.

The straight line to today: these optimizations were always compute-bound — which is exactly why GPU acceleration (NVIDIA cuLitho, Trend 4) landed here first, reporting ~20–60× speedups on the same problem.



Source(s): After A. Khan, ECE 8803-AI4 opening deck; connects to Trend 4 (computational lithography).

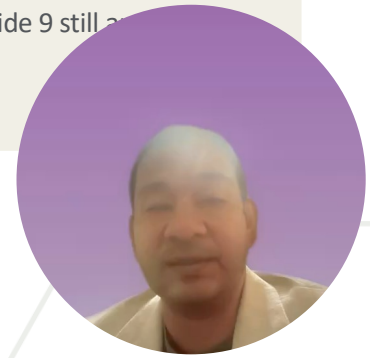
Deployment Era: Deep Learning Reaches Inspection

MID-2010s — CNNs ENTER PRODUCTION DEFECT REVIEW

When convolutional networks broke through on natural images, the fab had an obvious, data-rich target waiting: **wafer maps and inspection images**. This is where Trend 1 begins.

- **CNNs for wafer-map defect classification —** recognize spatial signatures (center, edge-ring, scratch…) at production volume.
- **Automated defect review (ADR/ADC) —** deep models triage SEM/optical images faster and more consistently than human classifiers.
- **A shift in kind —** from modeling one process to perceiving outcomes across many — but still narrow, supervised, per-fab.

Continuity, not rupture: more capable models on richer data — yet still one model per fab, retrained on drift. The regime's limits from slide 9 still a



Source(s): After A. Khan, ECE 8803-AI4 opening deck; connects to Trend 1 (defect detection).

So What Actually Changed?

“WE'VE DONE THIS FOREVER” CAN FLATTEN A REAL DISCONTINUITY — FOUR PLACES THE BREAK IS REAL

1 Surrogate & generative models

Learned models of process and device behavior that stand in for simulation — not just a regressor bolted onto one step.

2 Pretraining & transfer

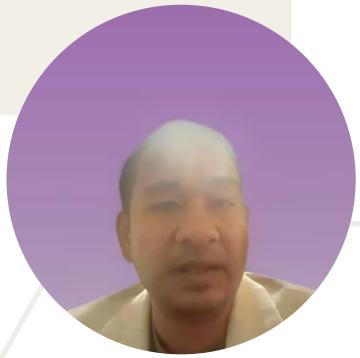
Representations learned once and adapted across tools, chambers, and nodes — the thing the classical regime could never do.

3 RL inside the design loop

DSO.ai (2020), Cerebrus and successors put reinforcement learning into place-and-route — a real break from heuristic optimization.

4 Agentic & language interfaces

Multi-step workflows over fab and design data, queried in natural language, chaining tools instead of returning a single number.



Source(s): After A. Khan, ECE 8803-AI4 opening deck; DSO.ai — Synopsys (2020).

The Bottleneck Was Never the Algorithm

FABS ARE THE MOST INSTRUMENTED FACTORIES ON EARTH — THE DATA IS STILL BARELY USABLE

Siloed

Design, litho, etch, metrology, test and yield each sit in separate systems that were never meant to join.

Proprietary

Tool vendors own the richest sensor streams. Access is contractual before it is technical.

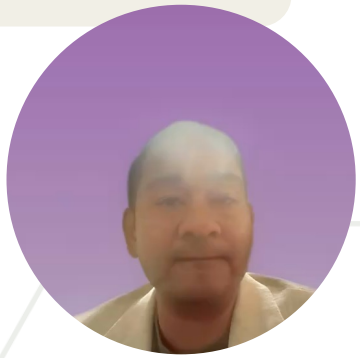
Unlabeled

Ground truth arrives weeks later at test, sparsely, and rarely maps back to the step that caused it.

Non-i.i.d.

Distributions shift by chamber, campaign, seasoning state, and site. Yesterday's training set is a different process.

Abundant and nearly unusable. That gap — not model architecture — is what "AI for semiconductors" has actually been running into.



Source(s): After A. Khan, ECE 8803-AI4 opening deck.

Two Things to Hold at Once

HOW TO READ THE CLAIM-MAKING YOU'LL MEET THIS TERM

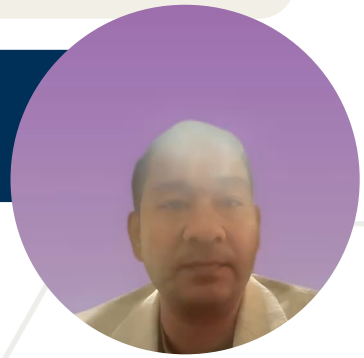
1 **“The industry has used AI for decades.”**

True, and a useful corrective to hype. But it can also be used deflationarily — to argue that nothing now is new.

2 **“The bottleneck is data infrastructure.”**

Also largely true. It is also, precisely, the commercial pitch of the vendors who most often make it.

Ask of every claim you read this term: what does this claim license, and who benefits if I accept it?



Where This Is Heading: From Copilots to Agents

THE FRONTIER — AND WHY IT STALLS

Copilots
an assistant automates one slice · human in the loop

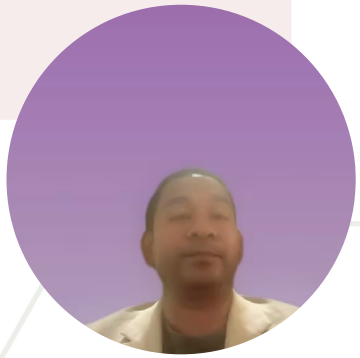
→

Agentic frameworks
many agents under an orchestrator · decide autonomously

It stalls because chip manufacturing spans **1,000+ process steps** across domains — and each step drops its data into its own silo:

EDA	Inspection / Metrology	MES	Test	Engineering notes & logs

The tax of fragmentation: an engineer can spend a **full quarter just assembling a dataset** — before building anything.



Source: Dr. Janhavi Giri (Principal Architect, NetApp) — talk on autonomous fabs & the data fabric, 2026 (youtu.be/ucSY4kK2kNw).

The Data Deluge — Abundant, and Unusable

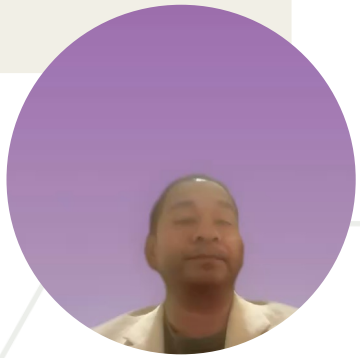
A SINGLE GIGAFAB, EVERY MINUTE

~95 GB	Petabytes	75,000	15,000	360,000
equipment data / minute	per day, per gigafab	wafer-move events / min	sensor readings / min	recipe events / min

AND THE PRESSURE AROUND IT

- **$133 B → $151 B** fab-equipment spend (2026 → 2027)
- **1 M+ skilled workers** needed by 2030
- **20–25% of cost** now in back-end assembly / packaging / test
- **water & sustainability** limits tightening

The data isn't missing — it's fragmented. Abundant and nearly unusable, exactly the gap Part B kept running into.



Source: Dr. Janhavi Giri (Principal Architect, NetApp) — talk on autonomous fabs & the data fabric, 2026 (youtu.be/ucSY4kK2kNw).

The Fix: A Unified Data Fabric

A SHARED LAYER CONNECTING DESIGN · MANUFACTURING · TEST · PACKAGING, IN REAL TIME

1

Unified access

Low-latency, interoperable data — at exabyte scale.

2

Standardized metadata

End-to-end traceability & root-cause analysis.

3

Data / compute locality

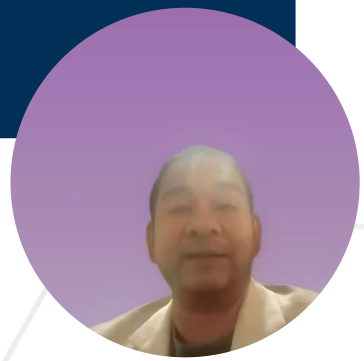
Keep compute near data — minimize movement.

4

Governance & security

Trust, access control, and IP protection.

The scalability-vs-IP-security paradox: semiconductor firms are cloud-averse, yet can't meet compute needs on-prem. The answer is to shift from **cloud-first to data-first** — so security travels with the data, not the other way around.



Source: Dr. Janhavi Giri (Principal Architect, NetApp) — talk on autonomous fabs & the data fabric, 2026 (youtu.be/ucSY4kK2kNw).

Terafab — and the Payoff of a Data Plane

ONE ROOF, ONE CONTINUOUS DATASET

THE ILLUSTRATION: "TERAFAB"

Musk's one-roof vision — design, fab, test, packaging, logic + memory all together — read as fundamentally a **data-continuity argument.**

One continuous dataset enables a **fast recursive improvement loop**. Only a greenfield fab can realistically do it — making it a blueprint for future fabs.

WHAT A SHARED DATA PLANE UNLOCKS

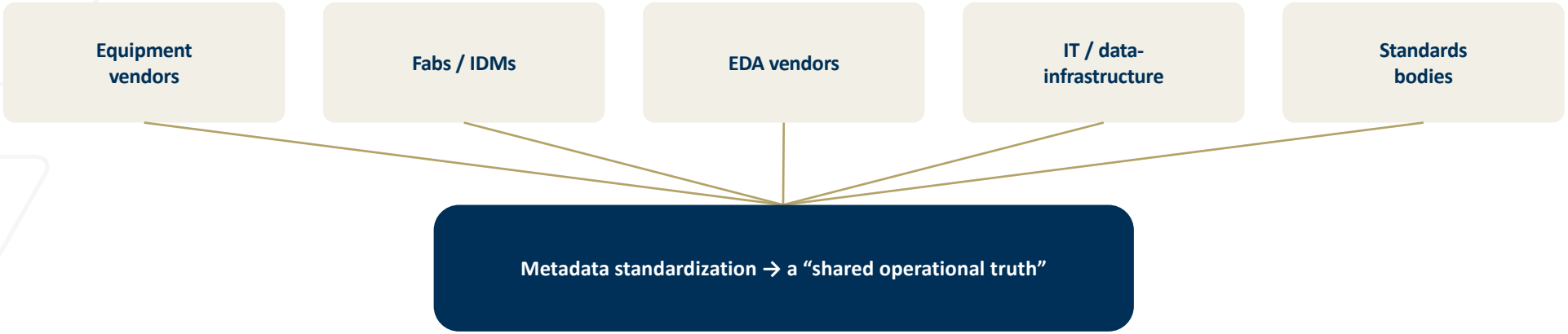
- ✓ Automated advanced process control (APC)
- ✓ Predictive maintenance
- ✓ Virtual metrology
- ✓ Dynamic scheduling & dispatch
- ✓ Faster defect resolution
- ✓ Back-end optimization
- ✓ Higher overall equipment effectiveness (OEE)

— i.e., today's six trends, unlocked at fab scale.

Source: Dr. Janhavi Giri (Principal Architect, NetApp) — talk on autonomous fabs & the data fabric, 2026 (youtu.be/ucSY4kK2kNw).

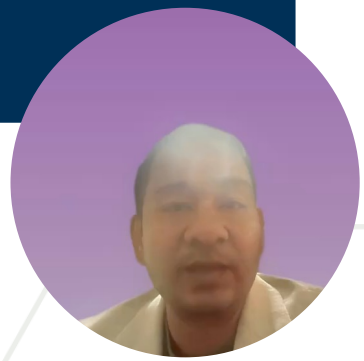
No Common Data Plane, No Autonomous Manufacturing

THE CALL TO ACTION — A SHARED OPERATIONAL TRUTH



“No common data plane, no autonomous manufacturing.”

The next chapter of the two-way street — and where your semester project lives.



Source: Dr. Janhavi Giri (Principal Architect, NetApp) — talk on autonomous fabs & the data fabric, 2026 (youtu.be/ucSY4kK2kNw).