Measurement Systems, Fab Case Studies, and SPC inside Advanced Process Control

Statistical Process Control in Semiconductor Manufacturing

ECE 8803 · AI for Semiconductor Manufacturing

Asif Khan

Associate Professor, School of Electrical and Computer Engineering

School of Materials Science and Engineering · Georgia Institute of Technology

akhan40@gatech.edu

Georgia Tech

What you should be able to do by the end of today

1. Run and interpret a gauge R&R study, and separate measurement variation from process variation
2. Correct a capability index for gauge error — and say when the correction changes the decision
3. Chart the physics rather than the raw number: residuals, coefficients, and signatures
4. Explain how SPC and run-to-run control interact, and what a controller hides
5. Describe how T^2, SPE and autoencoder reconstruction error are the same idea at different scales

Georgia Tech

Two lectures in, one assumption unexamined

What we have built

Lecture 1: variation, σ, the z-score, and two kinds of error.

Lecture 2: $\bar{X}$–R, I–MR and p-charts, run rules, and Cp / Cpk.

You can now take a stream of measurements and decide, with a stated false-alarm rate, whether the process changed.

What we assumed without saying so

That the measurements are the truth.

Every σ in the last two lectures — every control limit, every capability index, every ppm figure — was computed from numbers produced by a metrology tool that has its own variation.

Today we take that apart.

And then we look at what happens when the process is being actively controlled, and when there are three hundred sensors instead of one.

Georgia Tech

First, a refresher

Four ideas from Lecture 2 that today leans on

The p-chart

For data that is a count. Every die either passes or fails, and you chart the fraction that failed.

Cp

Asks whether the process is narrow enough to fit inside the specification.

Cpk

Asks the same question, and adds whether the process sits in the middle.

Pp and Ppk

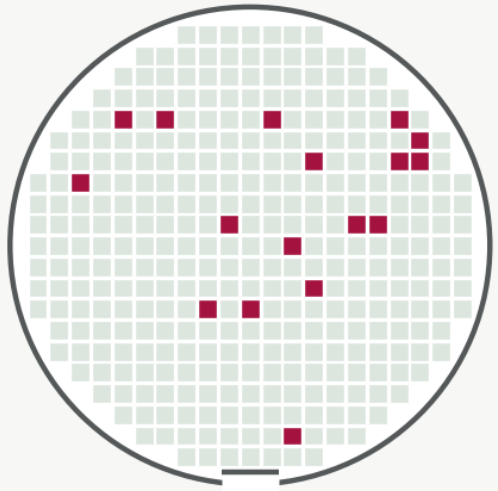
The same two questions asked over a month instead of a morning.

Ten slides. Then we spend the rest of today asking whether the measurements behind all four numbers can be trusted.

Georgia Tech

Counting good and bad: what p means

Some data is a measurement. Some data is a count.



one wafer · 349 die · 17 failed at probe

THE FRACTION

p = 17 / 349

= 0.049

the proportion that failed

WHAT GOES ON THE CHART

one p per lot

a p-chart is a run chart of that fraction

Every die is either good or bad. Count the bad ones, divide by how many you tested.

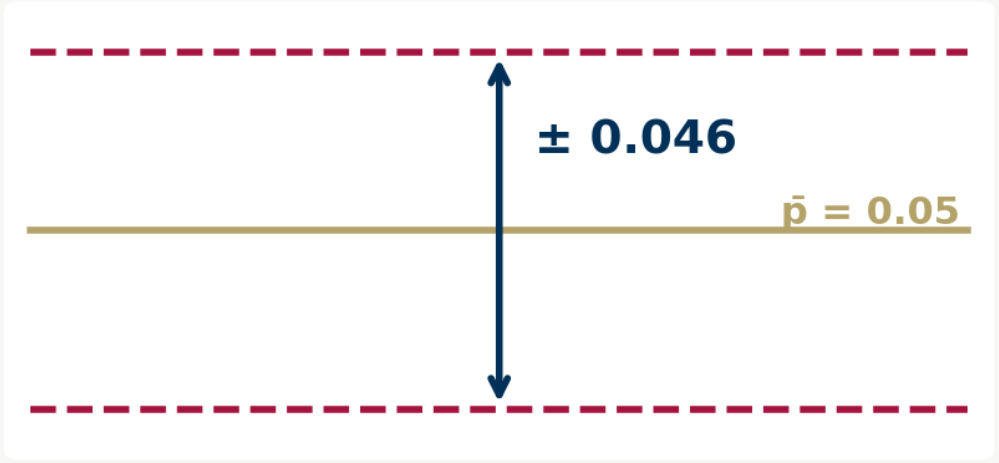
p is a fraction between 0 and 1. One p per lot, plotted in time order, is a p-chart.

Georgia Tech

Why a p-chart's limits move up and down

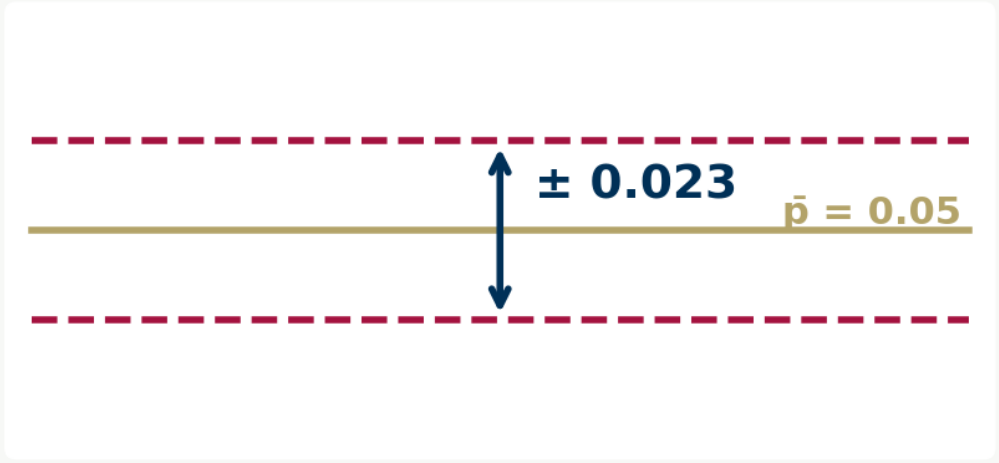
The band width depends on how many die you tested

A SMALL LOT · n = 200 die tested



wide band — one lot tells you less

A BIG LOT · n = 800 die tested

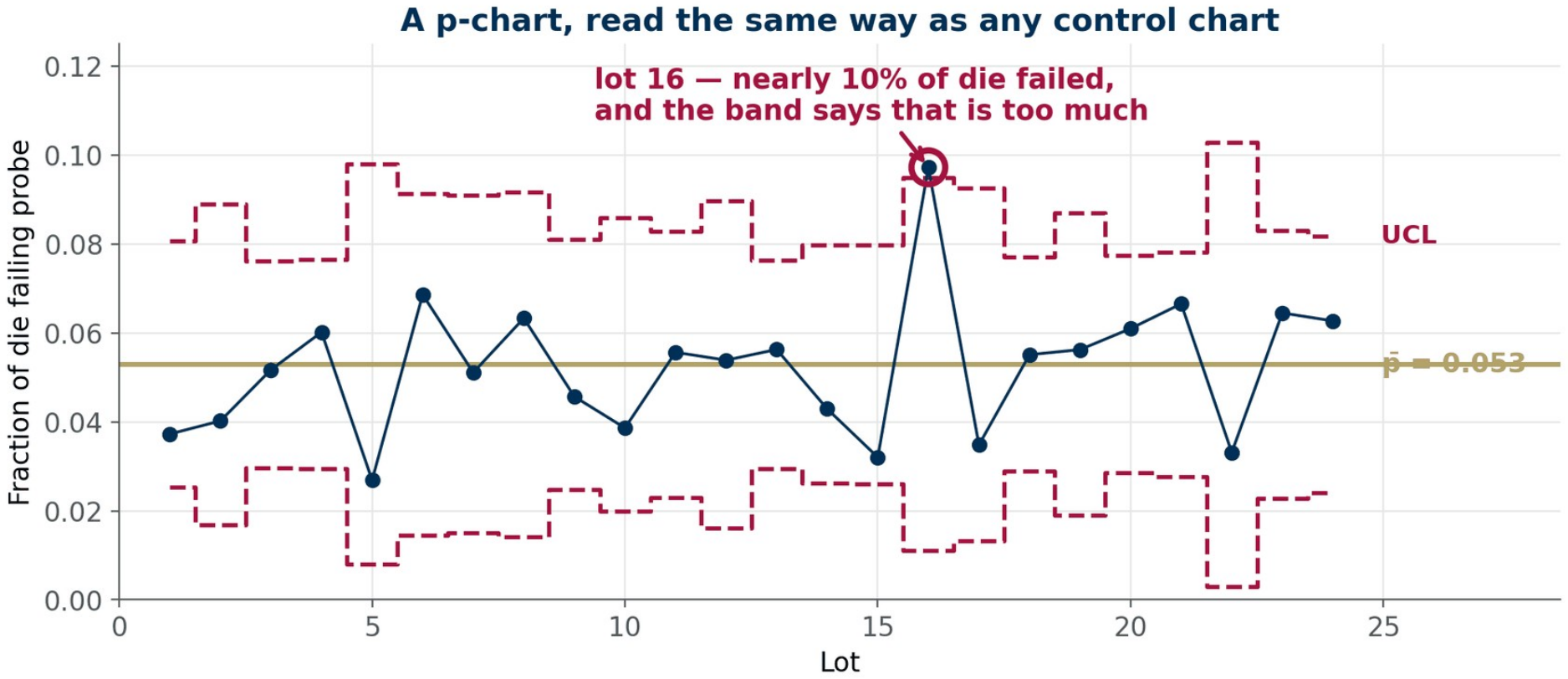


tight band — one lot tells you more

The more die you test, the sharper the statement you can make about that lot.

Georgia Tech

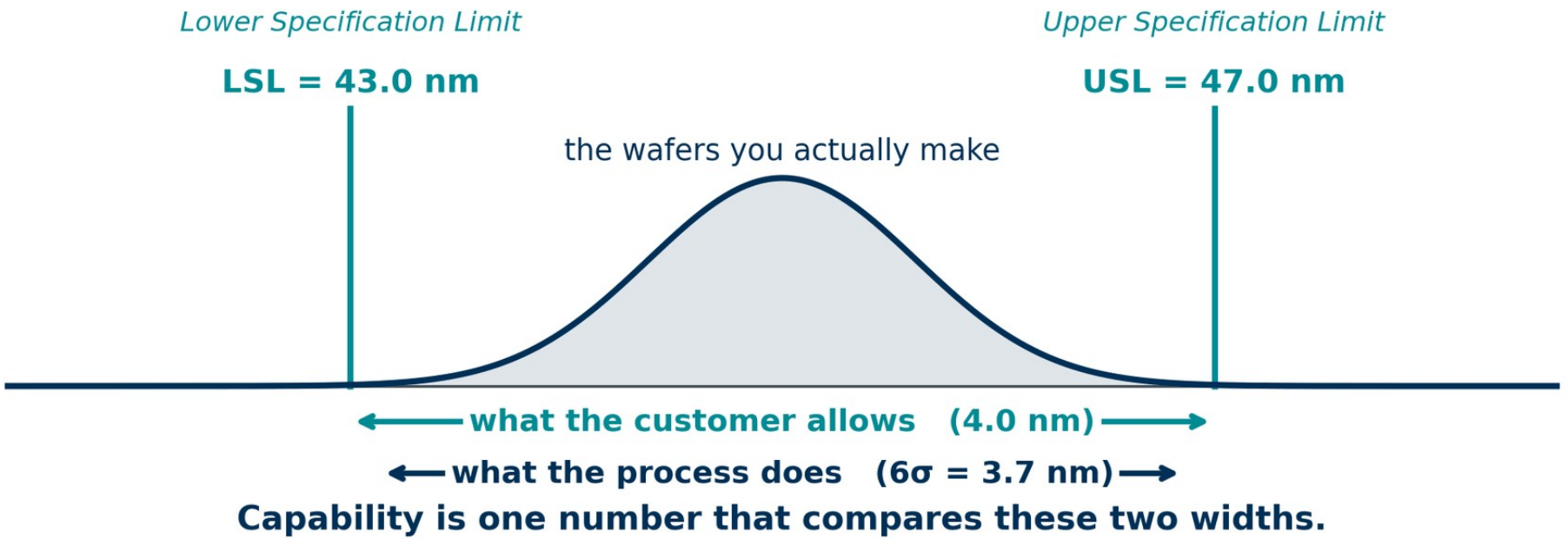
Reading a p-chart



Centre line, limits, points in time order — the same three things you read on every chart in Lecture 2.

Capability compares two widths

LSL and USL are the Lower and Upper Specification Limits — the window the designer allows



LSL — Lower Specification Limit

The smallest value the design will accept. Below it, the transistor does not work.

USL — Upper Specification Limit

The largest value the design will accept. Both limits come from design, not the tool.

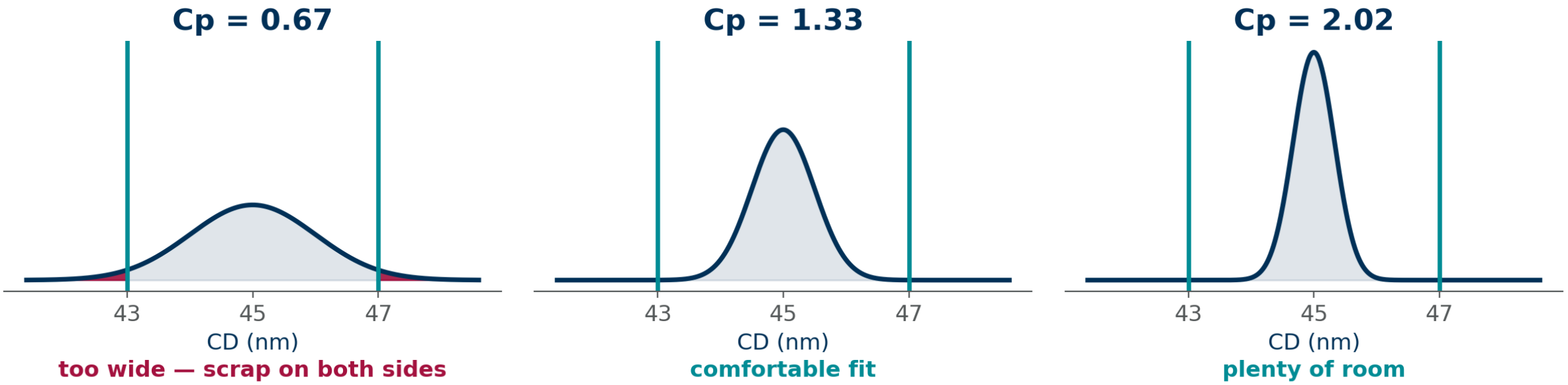
Georgia Tech

Cp — is the process narrow enough?

$$Cp = (USL - LSL) / 6\sigma$$

how wide you are allowed to be, divided by how wide you actually are

Cp asks one thing: is the process narrow enough to fit between the two spec lines?



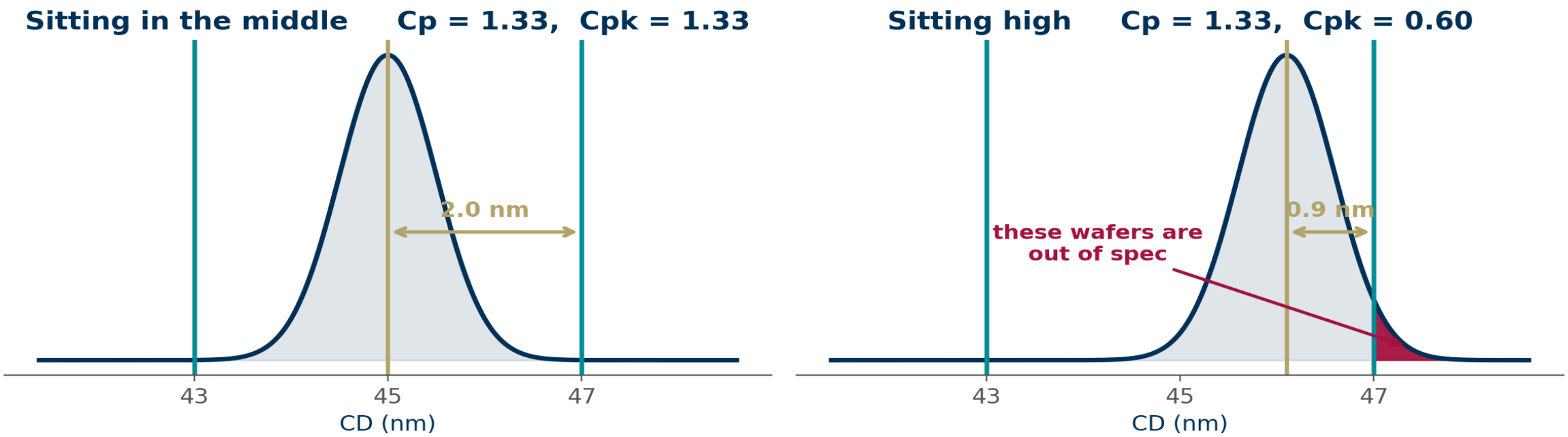
Above 1.33 is comfortable. Below 1.00 the process produces scrap even when it is perfectly centred, because it simply does not fit.

Georgia Tech

Cpk — and is it in the right place?

Cpk = min(USL − μ, μ − LSL) / 3σ — distance from the mean to the nearer spec line, divided by 3σ

Same width. Cpk adds the second question: how close is the process to the nearer line?



μ is the process mean. Only one side is measured, so the divisor is 3σ rather than 6σ. A centred process gives Cpk = Cp.

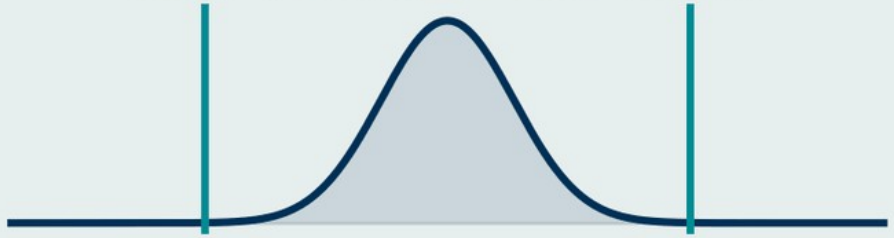
Georgia Tech

Reading Cp and Cpk together

The two numbers together tell you which problem you have

Cp ≈ Cpk

the process sits in the middle

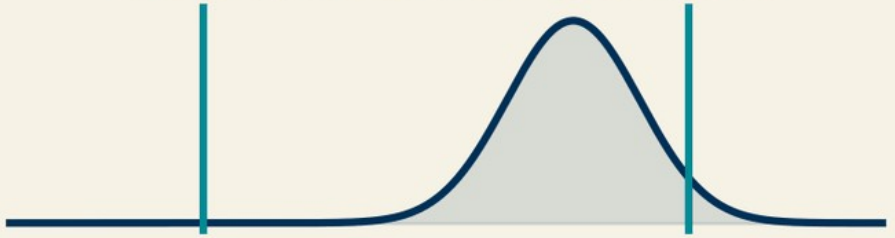


**If capability is still low,
the process is too WIDE.**

A wide process needs a better
tool, recipe or gauge — months.

Cp ≫ Cpk

the process sits off to one side



**The width is fine.
The process is OFF CENTRE.**

Off centre is usually one
recipe offset — an afternoon.

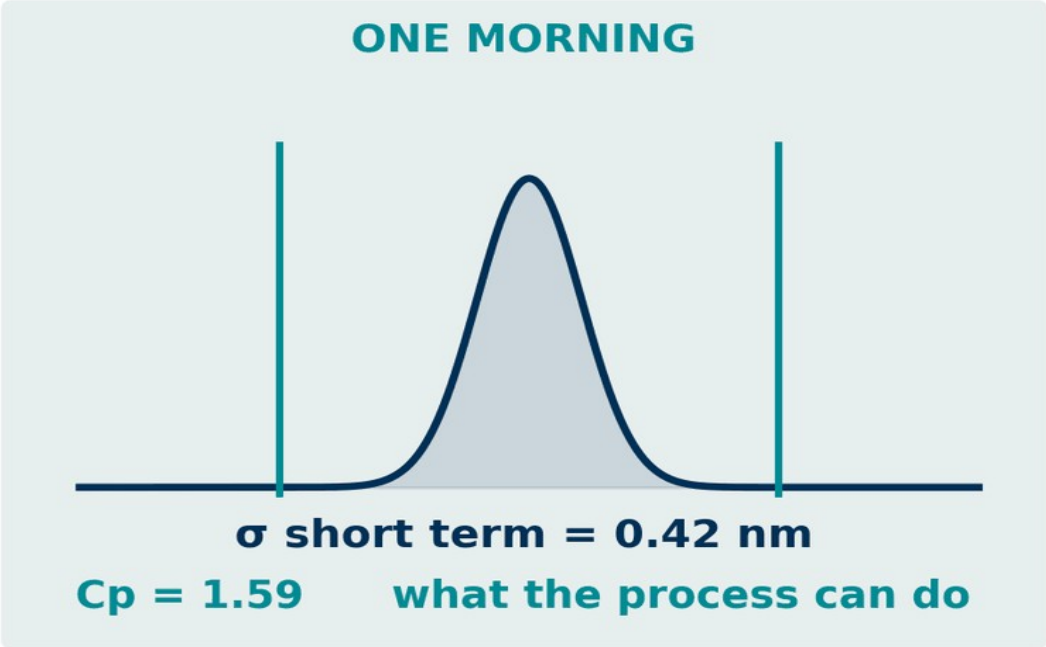
Report both numbers. The gap between them tells you which problem you have.

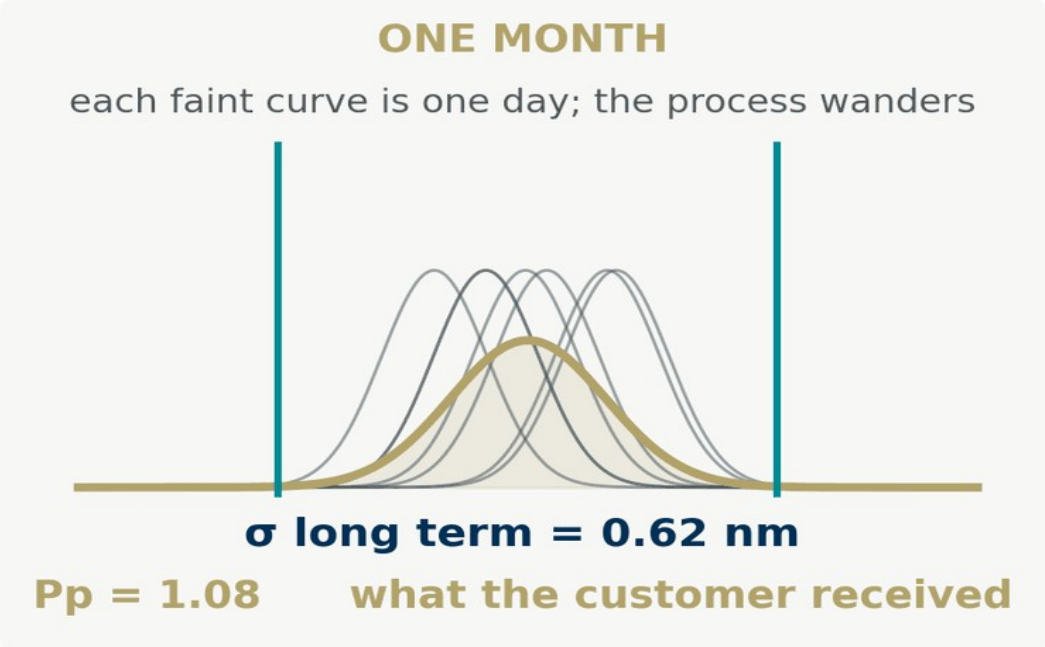
Georgia Tech

Pp and Ppk — the same questions, over a longer window

One morning's spread, and a whole month's

Pp, Ppk = the same two formulas, with σ computed from every measurement over the long run





Cp and Cpk use one morning's spread. Pp and Ppk use everything, including the wandering.

Georgia Tech

The four numbers, side by side

Cp	short term	**Narrow enough?**
Cpk	short term	**Narrow enough, and centred?**
Pp	long term	**Narrow enough all month?**
Ppk	long term	**Narrow and centred all month?**

WHAT THE NUMBER BUYS YOU

Cpk	bad wafers
1.00	1 350 ppm
1.33	32 ppm
1.67	0.3 ppm
2.00	0.002 ppm

out of every million features made

Every one of these is built on σ. Today we ask how much of that σ belongs to the metrology tool.

Georgia Tech

The uncomfortable question

Do you believe your own numbers?

A CD-SEM reports 45.3 nanometres.

How much of that number is the wafer?

It is never all of it

Every measurement is the true value plus measurement error. The only question is how much error, and whether you have quantified it.

It is not a small effect

On tightly toleranced parameters, measurement error is routinely 20 to 40 percent of the total observed variance.

And it is measurable

This is the good news. Measurement variation can be isolated with a designed experiment that takes an afternoon.

Georgia Tech

The equation this whole segment rests on

$$\sigma^2_{observed} = \sigma^2_{process} + \sigma^2_{measurement}$$

It is Lecture 1's variance-addition equation, with two terms.

Everything you have charted is the left-hand side

You never observe σ_process. You observe the sum, and you have been treating it as if it were the process.

Which has two consequences, both bad

Your control limits are too wide, so you detect less.

Your capability index is too low, so a good process reports as a poor one.

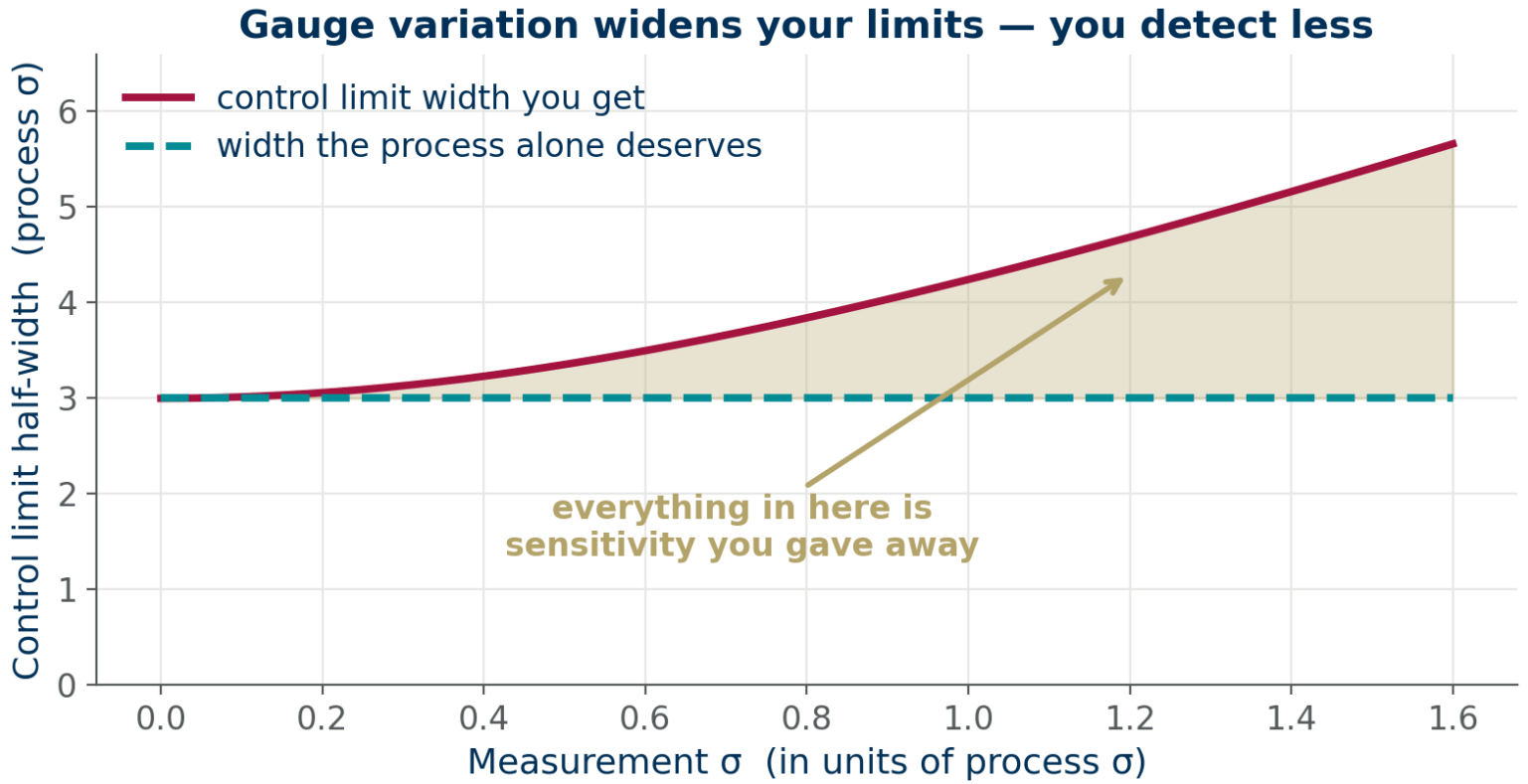
gauge — the whole measurement system: the tool, the operator, the recipe, the fixture, the sample prep.

gauge variation ($\sigma_{measurement}$) — the spread you get when you measure the same wafer over and over.

Georgia Tech

Consequence one: wider limits

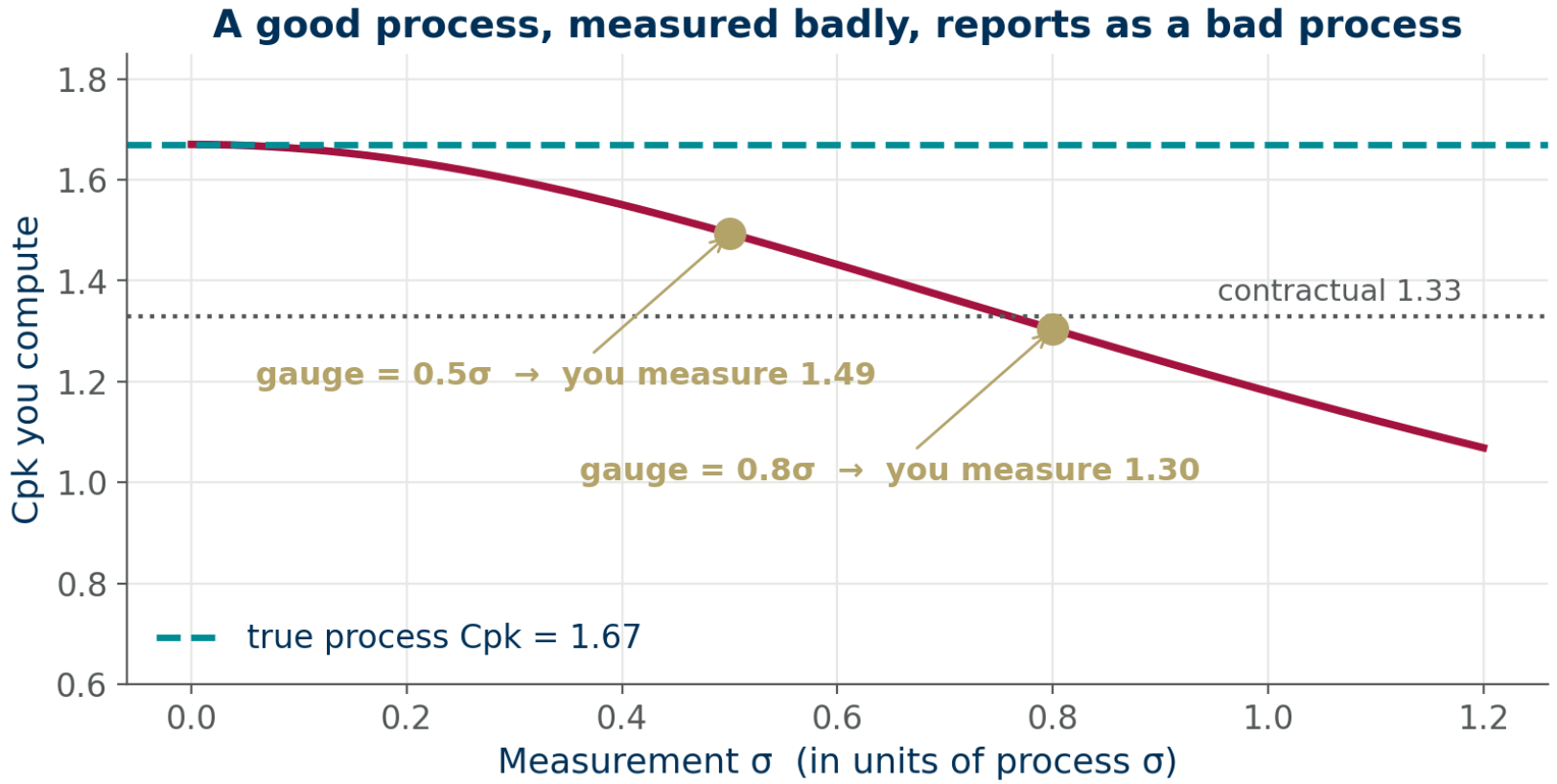
Gauge variation is $\sigma_{measurement}$ — the part of every number you chart that belongs to the tool, not the wafer



A gauge equal to the process σ widens your limits by 41% — and every bit of that is detection you gave away.

Georgia Tech

Consequence two: depressed capability



A process with a true Cpk of 1.67, measured with a gauge at 0.8σ, reports 1.30 — and fails a contract it actually meets.

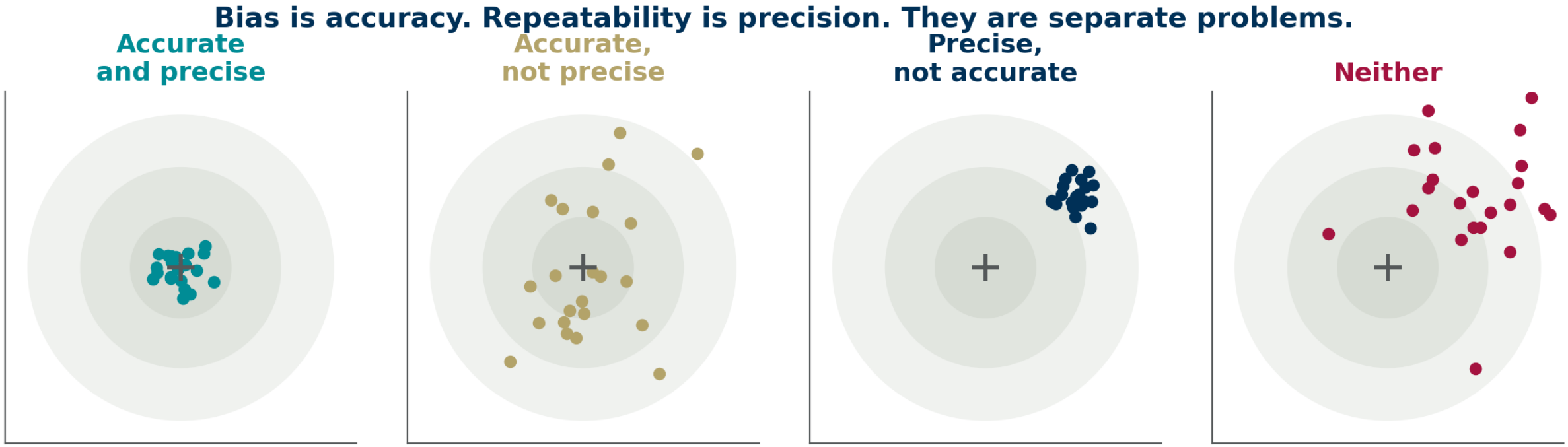
MEASUREMENT SYSTEM ANALYSIS

Your Cpk is fiction if the gauge eats your tolerance.

And you cannot tell from the number itself. It looks exactly as authoritative either way.

Georgia Tech

Two separate problems



Accuracy is whether you are on target. Precision is whether you can repeat. A gauge can fail at either independently.

Georgia Tech

The accuracy family

Systematic errors — they shift, they do not scatter

Bias

A constant offset from the true value. Your tool reads 0.4 nm high on everything.

Found by measuring a certified reference. Corrected by recalibration — and it is the easy one.

Linearity

Bias that changes across the range. Accurate at 45 nm, reads high at 60 nm.

Matters when one recipe spans several nodes on the same tool.

Stability

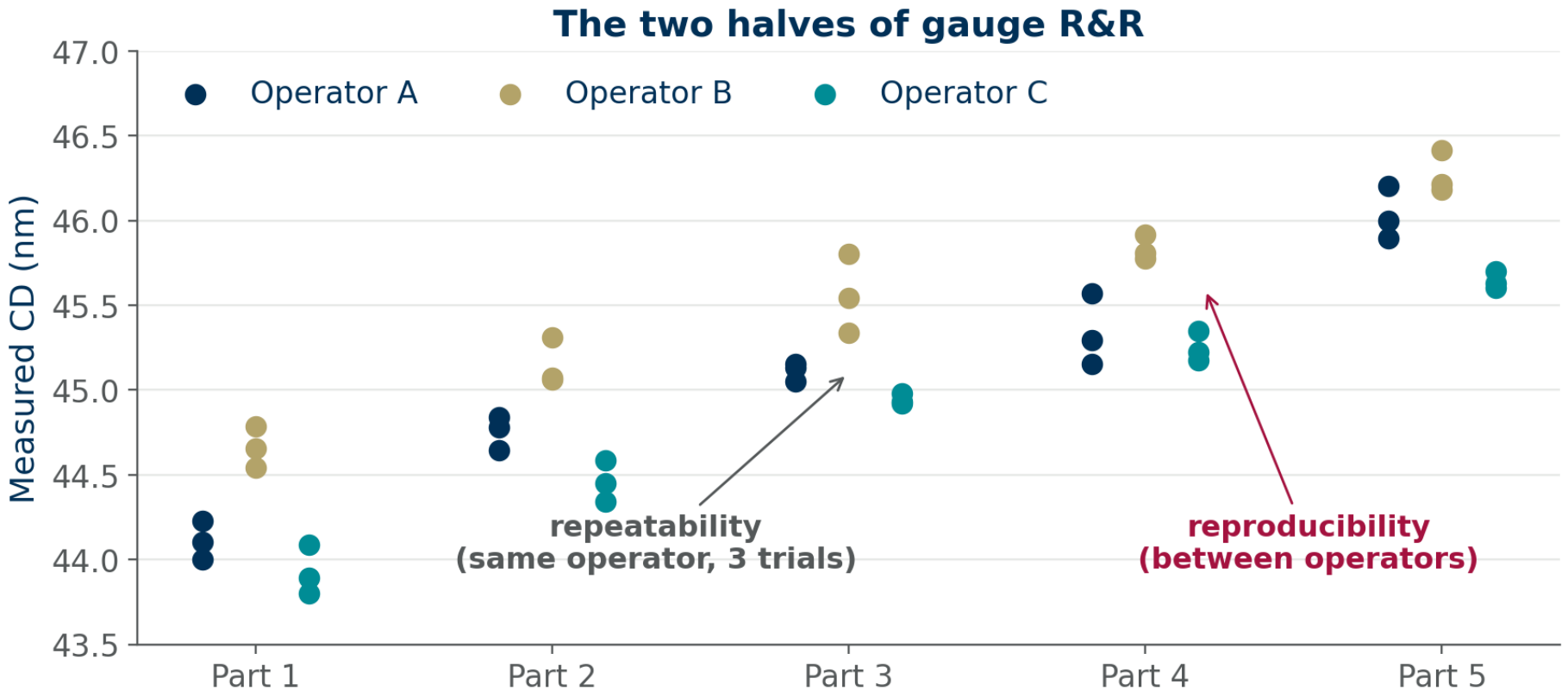
Bias that changes over time — drift between calibrations.

This is why fabs run daily monitor wafers on metrology tools. It is SPC applied to the gauge itself.

Bias does not affect your control limits at all — but it moves your Cpk, because Cpk depends on distance to spec.

Georgia Tech

The precision family: repeatability and reproducibility



Repeatability is the same operator on the same part. Reproducibility is the difference between operators, tools, or days.

Georgia Tech

Designing the study

The standard gauge R&R experiment

The classic design

10 parts × 3 operators × 2 trials = 60 measurements.

Parts should span the range you actually see in production — not ten parts from one good lot.

Randomise the order, and the operators must not see previous readings.

What each factor buys you

Repeated trials on the same part by the same operator → repeatability.

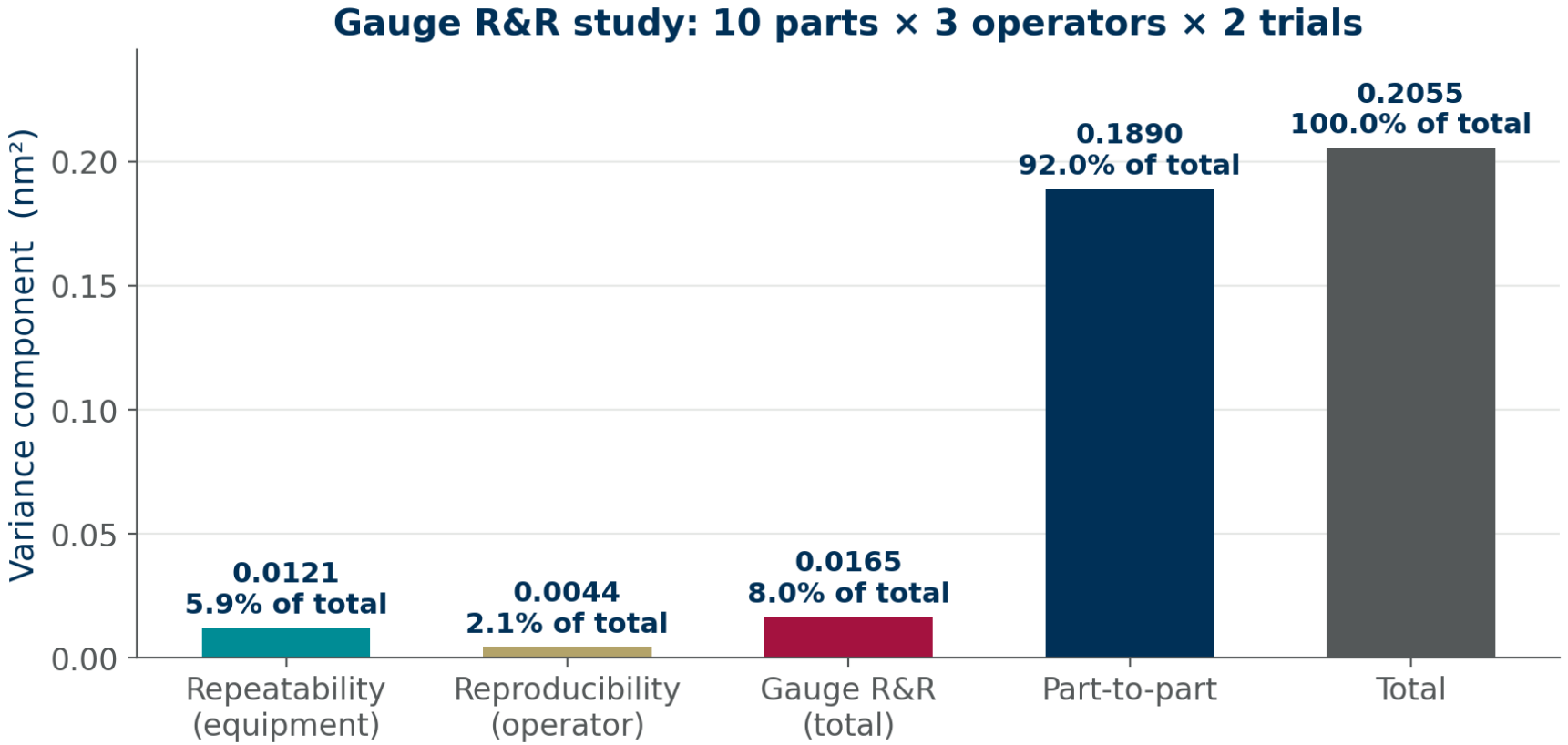
Different operators on the same parts → reproducibility.

Different parts → the part-to-part variation you are comparing against.

In a fab, 'operator' is usually 'metrology tool' or 'chamber' — the design is the same, the factor is different.

Georgia Tech

A worked study



Gauge R&R is 8% of total variance. Part-to-part is 92%. This is a healthy measurement system.

Georgia Tech

Two different percentages

And people quote whichever flatters them

%R&R against total variation

σ_gauge / σ_total.

Answers: can this gauge distinguish one part from another?

Depends on how variable your parts happen to be — so it changes when the process improves, without the gauge changing at all.

%R&R against tolerance

6σ_gauge / (USL − LSL).

Answers: can this gauge decide whether a part passes?

Independent of the process. This is the one that matters for disposition, and the one to insist on.

A gauge can look excellent on one and poor on the other. Always ask which.

Georgia Tech

Acceptance criteria

The AIAG convention — and what a fab really does

%R&R	Verdict	What it means in practice
under 10%	Acceptable	Use it. The gauge is not your problem.
10 – 30%	Conditional	Acceptable if the parameter is not critical, or if the cost of a better gauge is prohibitive. Document the decision.
over 30%	Unacceptable	The gauge is a substantial part of what you are charting. Fix it before you interpret any capability number.
ndc under 5	Unacceptable	Cannot resolve enough distinct levels to support a control chart at all.

These are conventions, not laws — exactly like 3σ. A 25% gauge on a non-critical layer may be a perfectly rational choice.

Georgia Tech

Number of distinct categories

A more intuitive way to say the same thing

$$ndc = 1.41 \times (\sigma\_part / \sigma\_gauge)$$

What it means

How many distinct levels the gauge can actually resolve across the range of parts.

An ndc of 2 means your expensive CD-SEM is effectively a pass/fail gate.

Why 5 is the threshold

Below 5 categories you cannot reliably estimate the within-subgroup range, so the control chart itself becomes unreliable — not just the capability number.

The intuition

It is the resolution of your measurement, expressed in units of the thing you are trying to see. A ruler marked only in metres cannot chart a person's height.

Georgia Tech

What makes this hard in a fab

Four problems the textbook does not have

CD-SEM measurement damages the resist

The electron beam shrinks the feature. So your second 'repeat' measurement is of a different, smaller feature. Repeatability is confounded with the act of measuring.

Overlay has tool-induced shift

TIS: the metrology tool contributes an apparent overlay of its own. It is measured by rotating the wafer 180° and re-measuring — a calibration step, every time.

Scatterometry is model-based

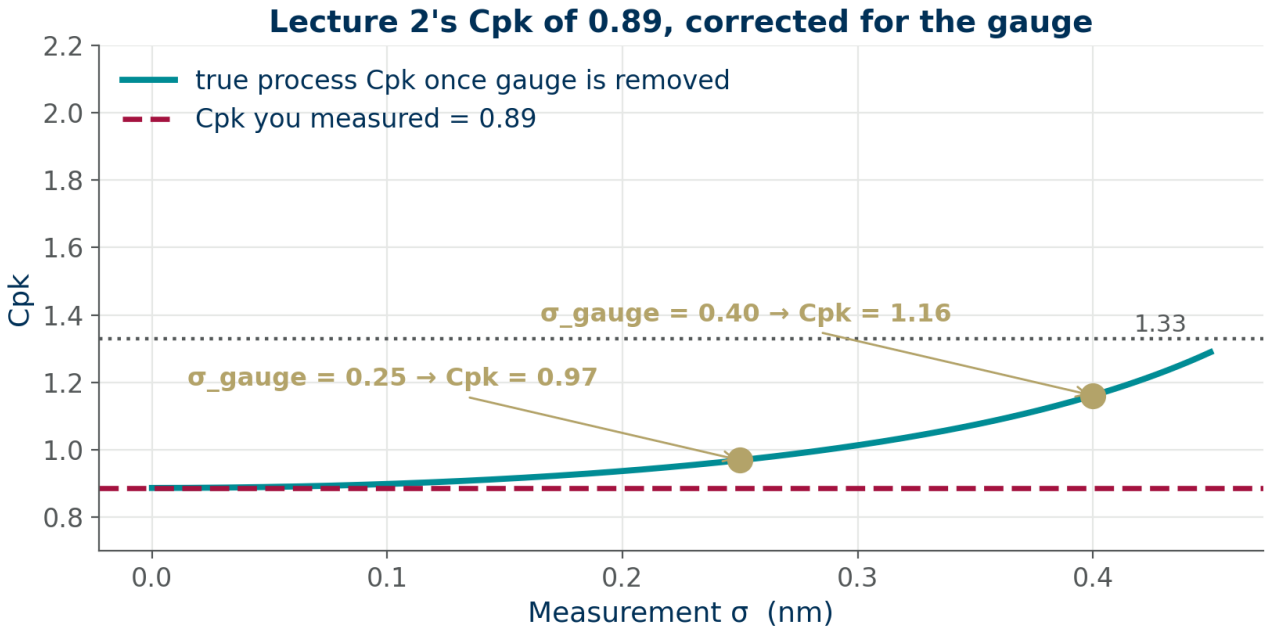
You do not measure CD. You measure a spectrum and fit an optical model. Model error is a bias that no amount of repeat measurement will reveal.

Many measurements are destructive

Cross-sections and SIMS give you one number per sample. Classical repeatability is impossible; you need nested designs on near-identical parts.

Georgia Tech

Now go back and fix Lecture 2's number



- We computed Cpk = 0.89 on the gate CD data, from $\hat{\sigma}$ = 0.62 nm, and called it unacceptable.
- **But that $\hat{\sigma}$ included the CD-SEM.**
- With a gauge σ of 0.25 nm, the true process σ is $\sqrt{(0.62^2 - 0.25^2)} = 0.57$, and Cpk rises to 0.97.
- With a gauge σ of 0.40 nm, the true process σ is 0.47, and Cpk is 1.16.
- **Same wafers. Same data.** The engineering conclusion changes from 'the process is broken' to 'the metrology is half the problem.'

Before you fund a tool upgrade, find out how much of the variance you are paying to fix belongs to the gauge.

Georgia Tech

Three parameters, three different problems

Everything so far has assumed one number per lot. Reality is messier.

Etch rate

A process that is supposed to drift.

The naive chart alarms on the design intent. The fix is to chart what the physics does not explain.

Critical dimension

A process with structure inside the wafer.

One number per wafer throws away the spatial signature — which is where the diagnosis lives.

Overlay

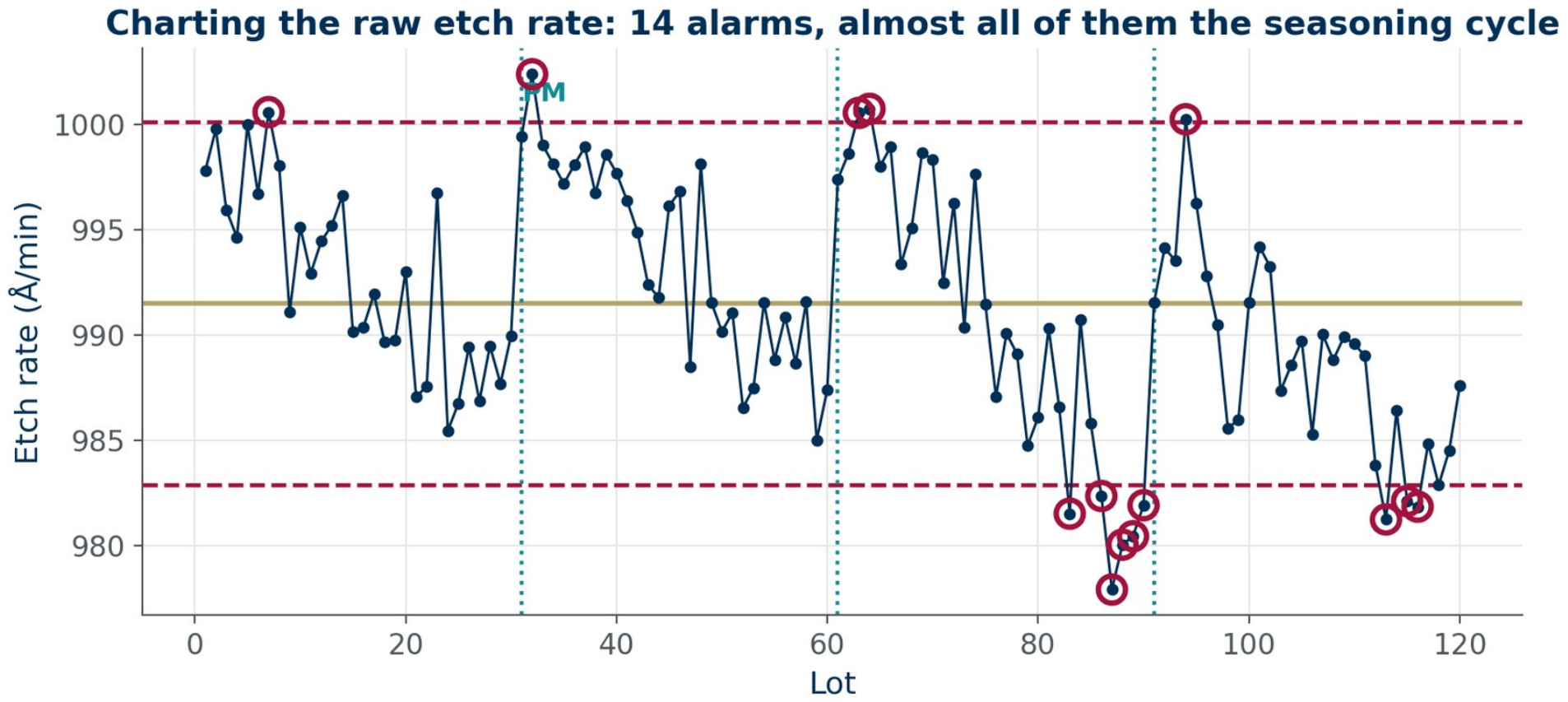
A process that is inherently multivariate.

Translation, rotation and magnification are different faults. A single magnitude cannot separate them.

In all three cases the fix is the same idea: chart the physics, not the raw number.

Georgia Tech

Case 1 — etch rate, charted naively



Fourteen alarms in a hundred and twenty lots. Almost every one of them is the seasoning cycle doing exactly what it was designed to do.

Georgia Tech

The problem with charting a process that is meant to drift

What the chart assumes

A stable mean, with random scatter around it. Any sustained movement is a signal.

What the process does

Etch rate falls predictably as the chamber seasons, then resets at every preventive maintenance. The sawtooth is the design, not a fault.

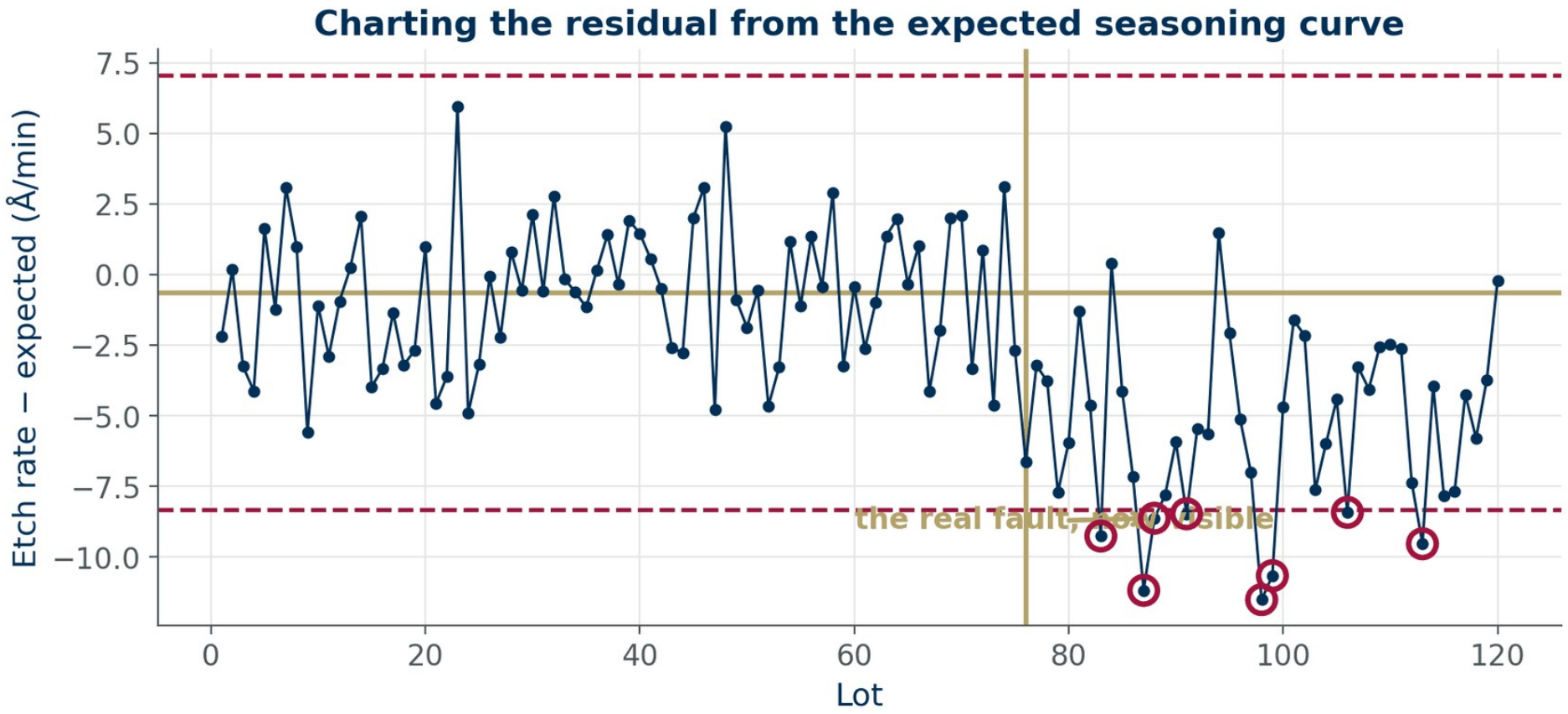
So the chart is right and useless at the same time.

It is correctly reporting that the mean is not constant. That is true, it is known, and it is not actionable — and a chart that reports known behaviour as an alarm gets switched off within a month.

Whenever a chart alarms on something you already understand, the chart is modelling the wrong quantity.

Georgia Tech

Chart the residual instead



Subtract the expected seasoning curve, chart what is left. The seasoning alarms vanish — and a real fault at lot 76 becomes obvious.

Georgia Tech

Residual charting, generalised

One of the most useful ideas in the course

Build a model of what you expect

It can be as simple as an average seasoning curve indexed by wafers-since-PM, or as complex as a regression on chamber state.

Chart the residual

Observed minus expected. If your model is right, the residual is noise — which is exactly what a Shewhart chart wants.

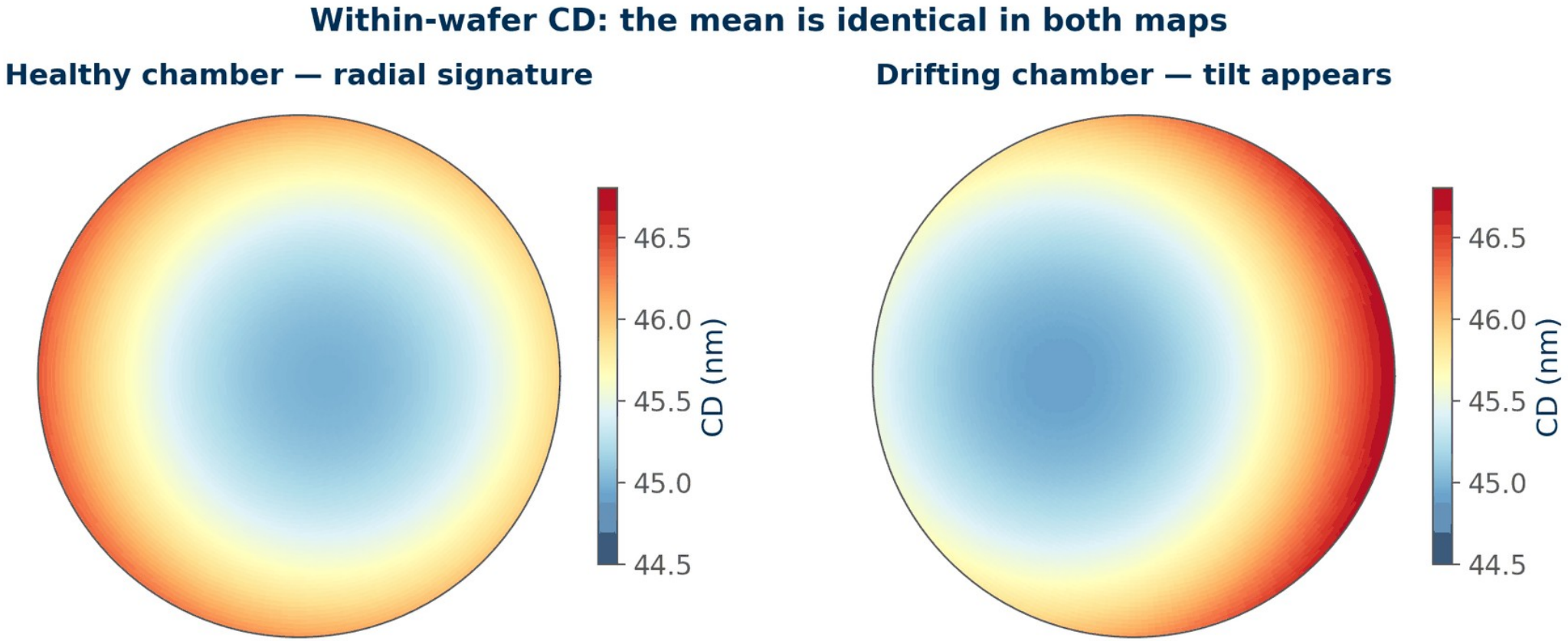
Now the chart means something

An alarm says 'the process departed from what we understand about it', which is far more actionable than 'the number moved'.

This is also the bridge to the machine learning half of the course: a model that predicts, and a chart on what it gets wrong.

Georgia Tech

Case 2 — critical dimension has structure inside the wafer



Identical wafer mean. Identical wafer standard deviation. Completely different physical state.

What one number per wafer throws away

The chart sees

Mean CD: 45.6 nm on both wafers.

Within-wafer σ: 0.42 nm on both.

Both wafers pass. Both plot at the same place. No rule fires.

The physics is

Wafer A: a symmetric radial signature — normal for this chamber.

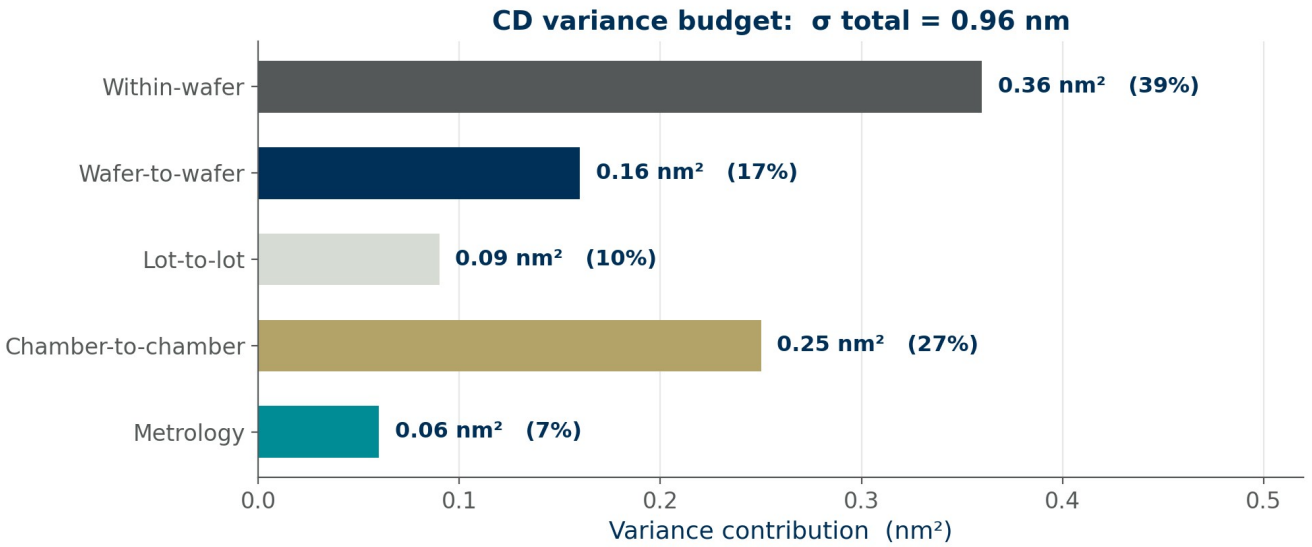
Wafer B: a tilt has appeared across the wafer. Something is asymmetric — gas flow, electrode wear, a lift-pin issue.

So chart the signature, not just the summary.

Fit a simple spatial model — mean, radial term, tilt in x, tilt in y — and put a control chart on each coefficient. Four charts instead of one, and each has a physical meaning.

Georgia Tech

Where the CD variance actually lives



- Lecture 1's equation, applied to a real fleet.
- **Within-wafer is the largest single term at 39%.**
- Chamber-to-chamber is 27% — that is a matching problem, and it is often the cheapest one to fix.
- Metrology is 7%. Worth knowing, and not where the leverage is here.
- **Attack within-wafer and chamber matching.** Eliminating lot-to-lot entirely would take total σ from 0.96 to 0.91 nm — almost nothing for a great deal of work.

The variance budget is how you decide where to spend. Compute it before the project, not after.

Georgia Tech

Case 3 — overlay is multivariate by construction

What overlay is

The misregistration between the layer you are printing and the layer underneath, measured at many points across each field and each wafer.

It is a vector field, not a number.

Why that matters

Different physical faults produce different fields. A stage calibration error produces translation. A thermal problem produces magnification. A rotational misalignment produces rotation.

The magnitude can be identical in all three.

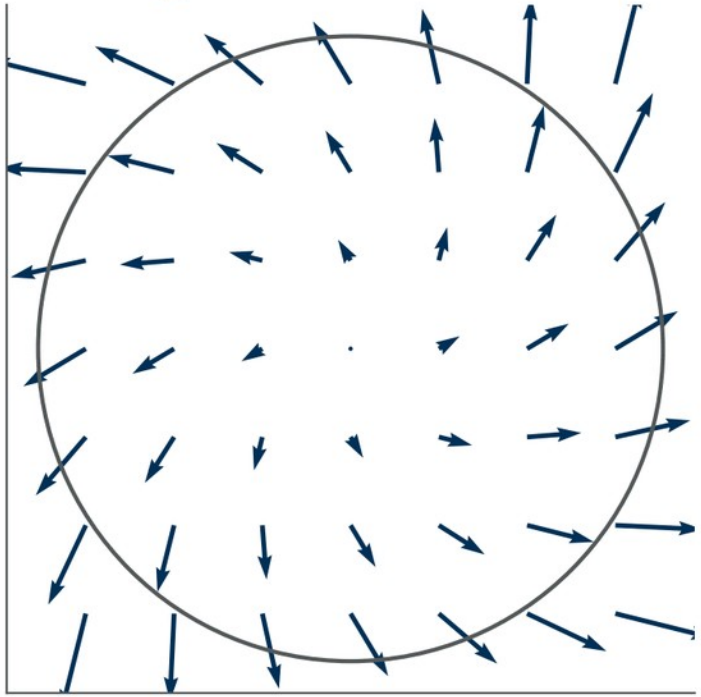
Averaging the magnitude destroys exactly the information you need to fix it.

Georgia Tech

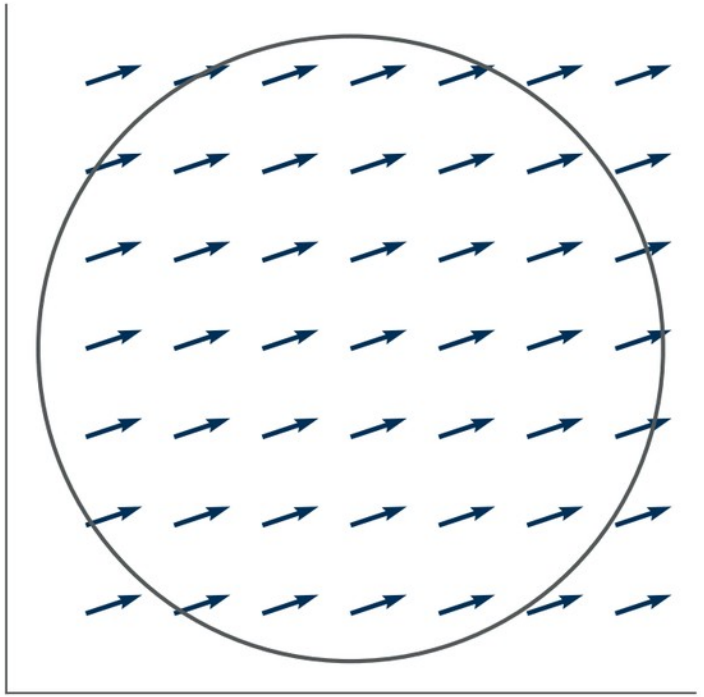
Same magnitude, different fault

Both wafers have the same mean overlay magnitude. One univariate number cannot tell them apart.

Magnification + rotation



Pure translation



Both wafers average about 0.1 nm of overlay. One needs a stage calibration; the other needs a thermal investigation.

Georgia Tech

So fit a model and chart its coefficients

The standard linear overlay model

$$dx = Tx + M \cdot x - R \cdot y + \text{higher-order terms}$$
$$dy = Ty + M \cdot y + R \cdot x + \text{higher-order terms}$$

Translation Tx, Ty

The whole field is shifted. Stage calibration, alignment mark detection.

Magnification M

The field is scaled. Thermal expansion, lens heating, wafer bow.

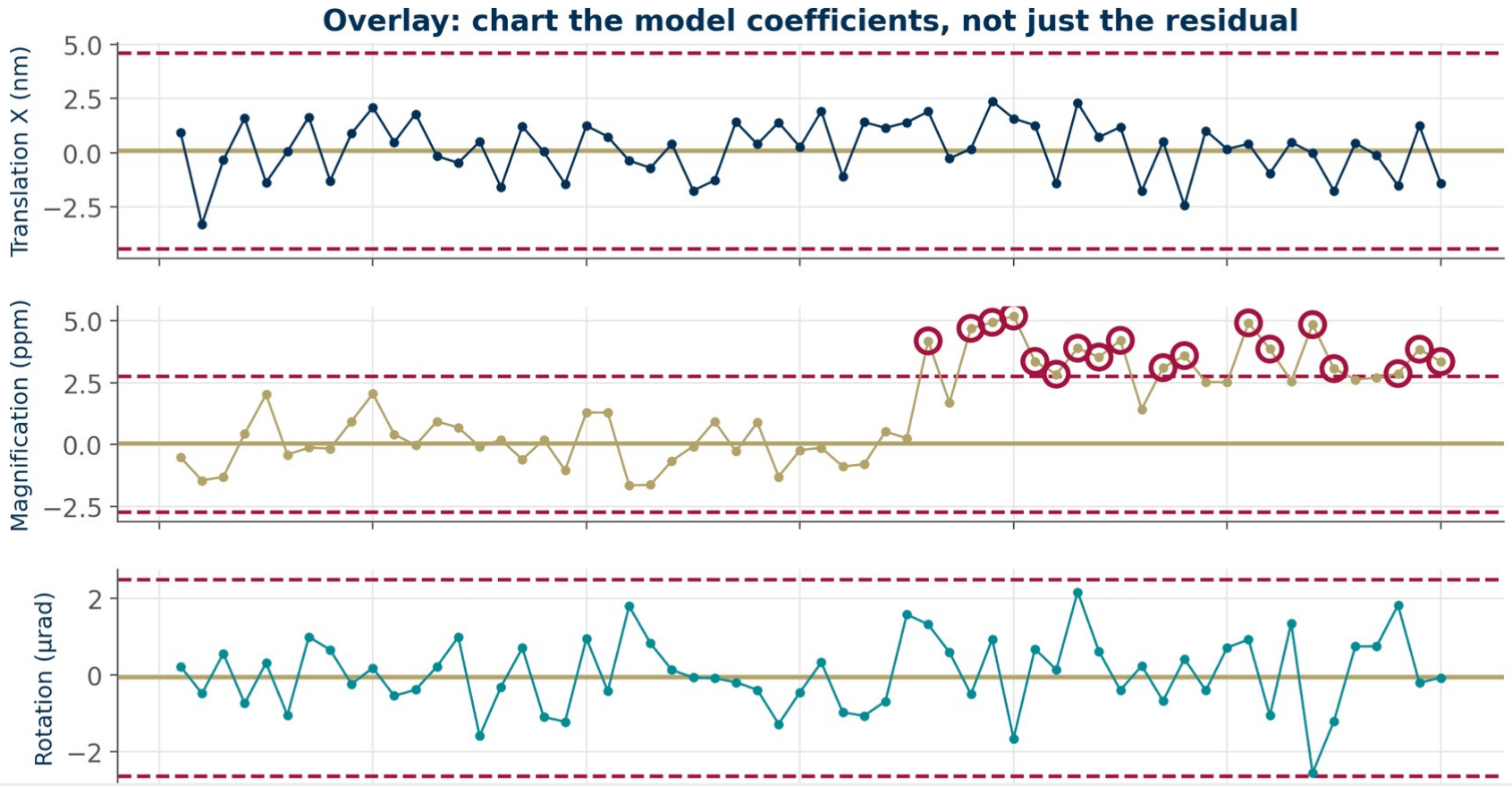
Rotation R

The field is rotated. Wafer pre-alignment, chuck rotation.

Each coefficient gets its own control chart — and each alarm points at a different piece of hardware.

Georgia Tech

Charting the coefficients



Magnification goes out of control at lot 36. Translation and rotation are fine. That is a diagnosis, not just an alarm.

Georgia Tech

What the three cases have in common

The raw number is the wrong thing to chart

Etch rate drifts by design. CD has spatial structure. Overlay is a vector field. In all three, the obvious summary destroys the signal.

A model turns it into something chartable

A seasoning curve, a spatial fit, a linear overlay model. Each one converts physics you understand into an expectation you can subtract.

And then Shewhart still works

On the residual, or on the coefficients, the 1924 machinery applies unchanged. You have not replaced SPC — you have given it better inputs.

Almost every advanced monitoring method in this course is this pattern: model what you can explain, chart what you cannot.

Georgia Tech

THE UNIFYING IDEA

Chart the physics, not the number.

A model of what you expect, and a control chart on the difference. That structure scales all the way to deep learning.

Georgia Tech

SPC detects. APC compensates.

Two different jobs that are constantly confused

Statistical process control

Watches. Says nothing until something changes, then raises a signal for a human.

Goal: detect a real change quickly, without reacting to noise.

It never touches the process.

Advanced process control

Acts. Adjusts recipe parameters lot to lot, automatically, to hold the output on target.

Goal: keep the output where you want it despite drift.

It changes the process continuously.

They interact — and the interaction is where the interesting failures live.

Georgia Tech

Three loops around the same process

Feedforward

Measure something upstream, predict its effect, and pre-compensate before the step runs.

Incoming film thickness → adjust etch time. Resist CD → adjust exposure dose.

Feedback (run-to-run)

Measure the output, and adjust the next lot's recipe by some fraction of the error.

The workhorse of modern fabs — and an EWMA in disguise.

Monitoring (SPC)

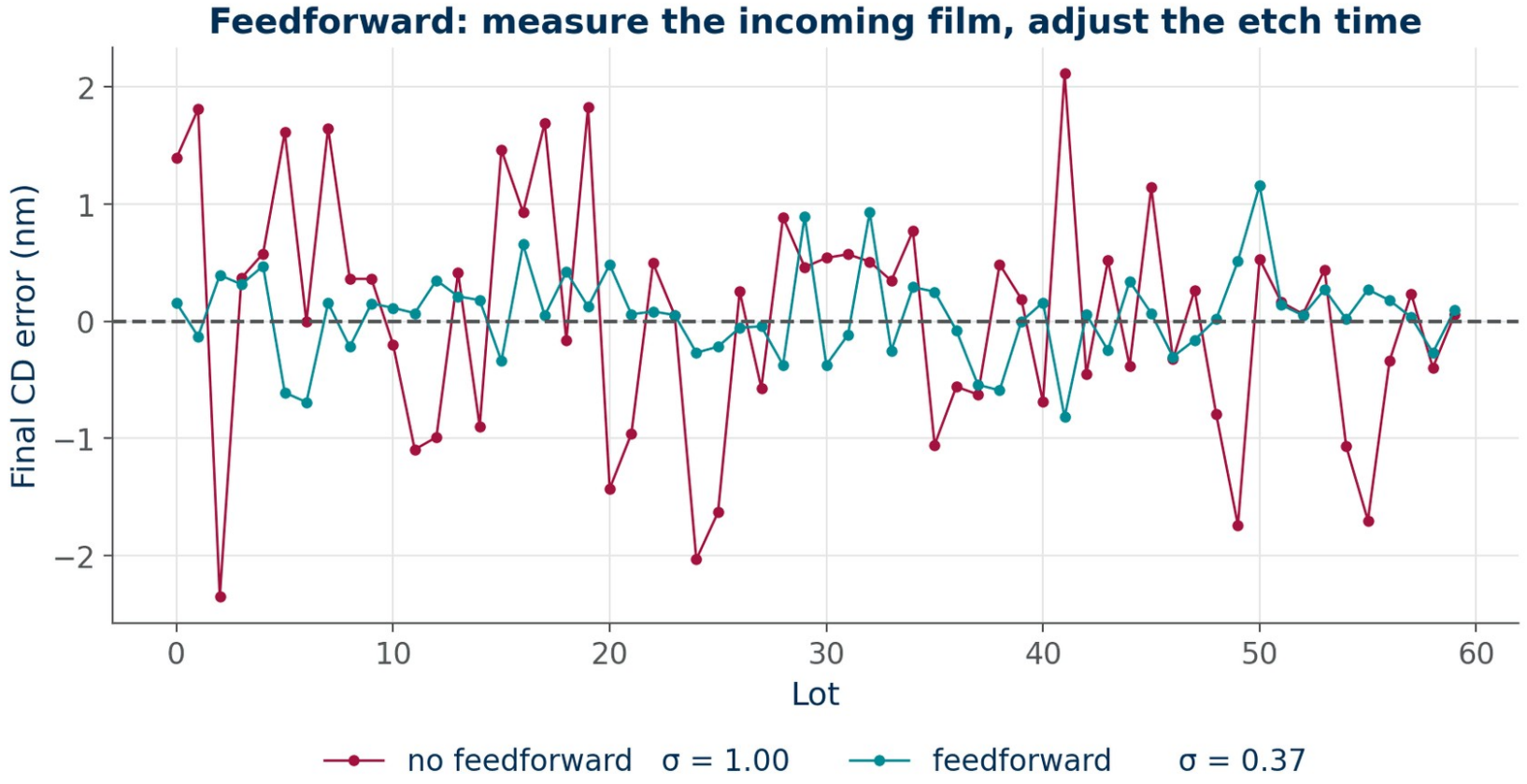
Watch everything, change nothing, and raise a signal when the pattern breaks.

Still necessary — arguably more necessary — when the other two loops are running.

The third loop is the one that gets neglected once the first two are working. That is a mistake, and the next few slides show why.

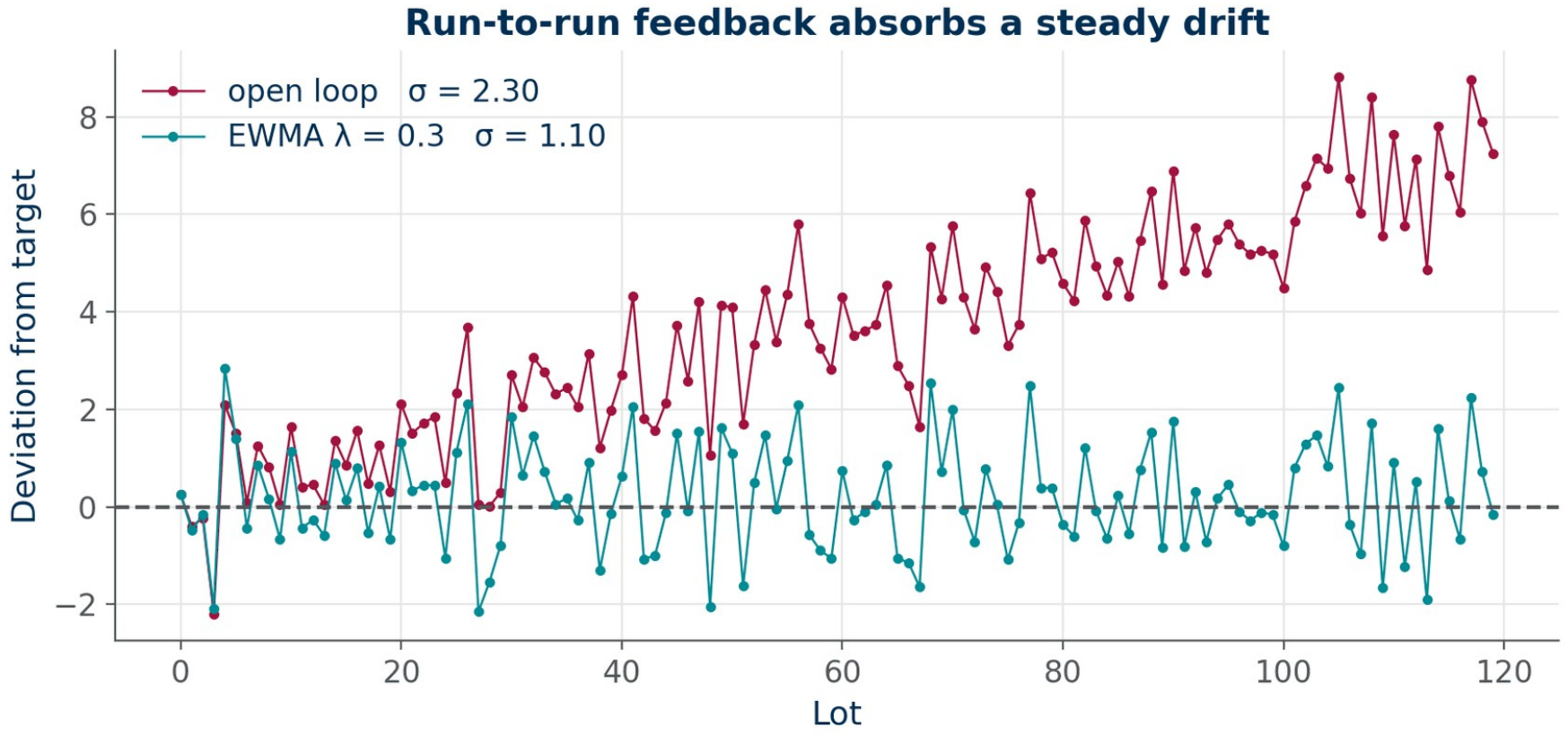
Georgia Tech

Feedforward



Incoming variation you can measure is variation you can cancel. σ drops from 1.00 to 0.37 nm.

Feedback: run-to-run control



A steady drift that would have walked the process off target is held flat, lot by lot, with no human involved.

Georgia Tech

The EWMA controller

Almost every R2R controller in production is this

$$\text{offset}_{n+1} = \text{offset}_n - \lambda \cdot \text{error}_n$$

λ is the whole design

The fraction of the observed error you correct on the next lot. Everything about the controller's behaviour follows from it.

Small λ

Sluggish. Ignores noise beautifully, and lets a real drift accumulate for many lots before catching up.

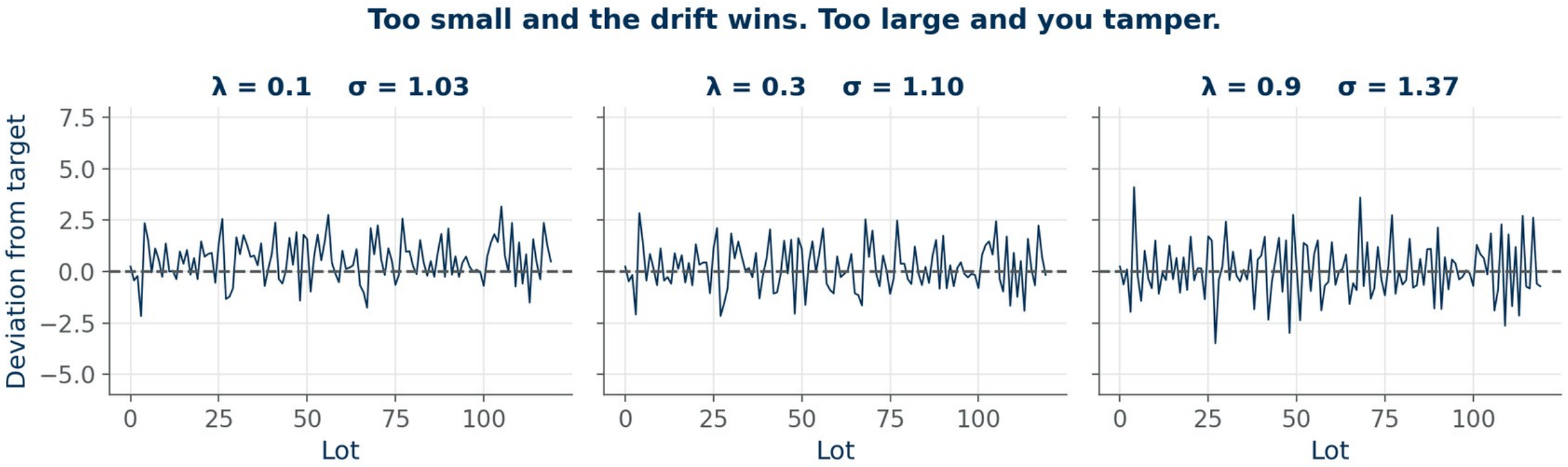
Large λ

Twitchy. Tracks a real shift almost immediately, and chases every random wobble — which is tampering.

λ = 1 is Deming's funnel exactly: correct the full error every time, and make the variance worse.

Georgia Tech

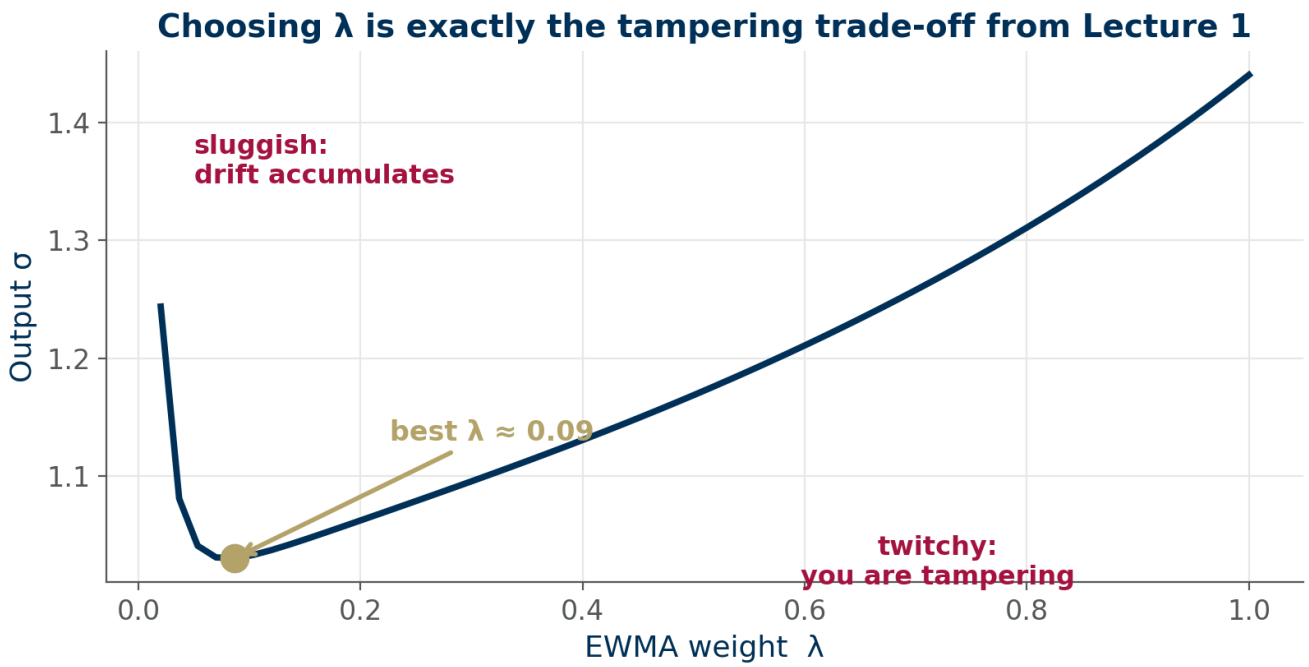
Choosing λ



Three values, same process. There is a genuine optimum, and it depends on how fast your process drifts relative to its noise.

Georgia Tech

It is Lecture 1's trade-off, again



- Too small and the drift wins. Too large and you are tampering.
- **The minimum is where drift rate meets noise level.**
- Which means the optimal λ is a property of the process, not a universal constant — and it should be re-tuned when the process changes.
- **Note the shape:** the penalty for being too twitchy grows without bound, while the penalty for being sluggish is bounded. When in doubt, err small.
- In practice λ between 0.1 and 0.3 covers most fab applications.

Every controller has a tampering knob. Know which parameter it is.

Georgia Tech

Threading

One controller per what, exactly?

The problem

A single offset for a process step is wrong the moment two products, two chambers, or two incoming layers behave differently.

Correct for product A's error and you have just mis-set the recipe for product B.

The solution: threads

Maintain a separate offset for each combination that behaves differently — typically product × tool × chamber, sometimes × incoming layer.

Each thread is its own little EWMA controller with its own state.

And now you have a new problem: sparse threads.

A high-mix fab can have thousands of threads, many of which see a lot once a month. An EWMA with three observations is not a controller — it is a rumour.

Georgia Tech

Deadbands and guardrails

Engineering that surrounds the equation

Deadband

Do not correct at all unless the error exceeds a threshold. Directly prevents tampering on noise, at the cost of a small standing offset.

Maximum move

Cap how far the recipe can change in one step. Protects against a single bad measurement driving a large, wrong correction.

Sanity limits

Absolute bounds on the recipe parameter. The controller must never be able to command something physically unreasonable.

Measurement validation

Reject the update if the metrology reading fails its own checks. A controller fed a bad number will faithfully act on it.

Most real R2R failures are not control theory failures. They are a bad measurement, acted on with confidence.

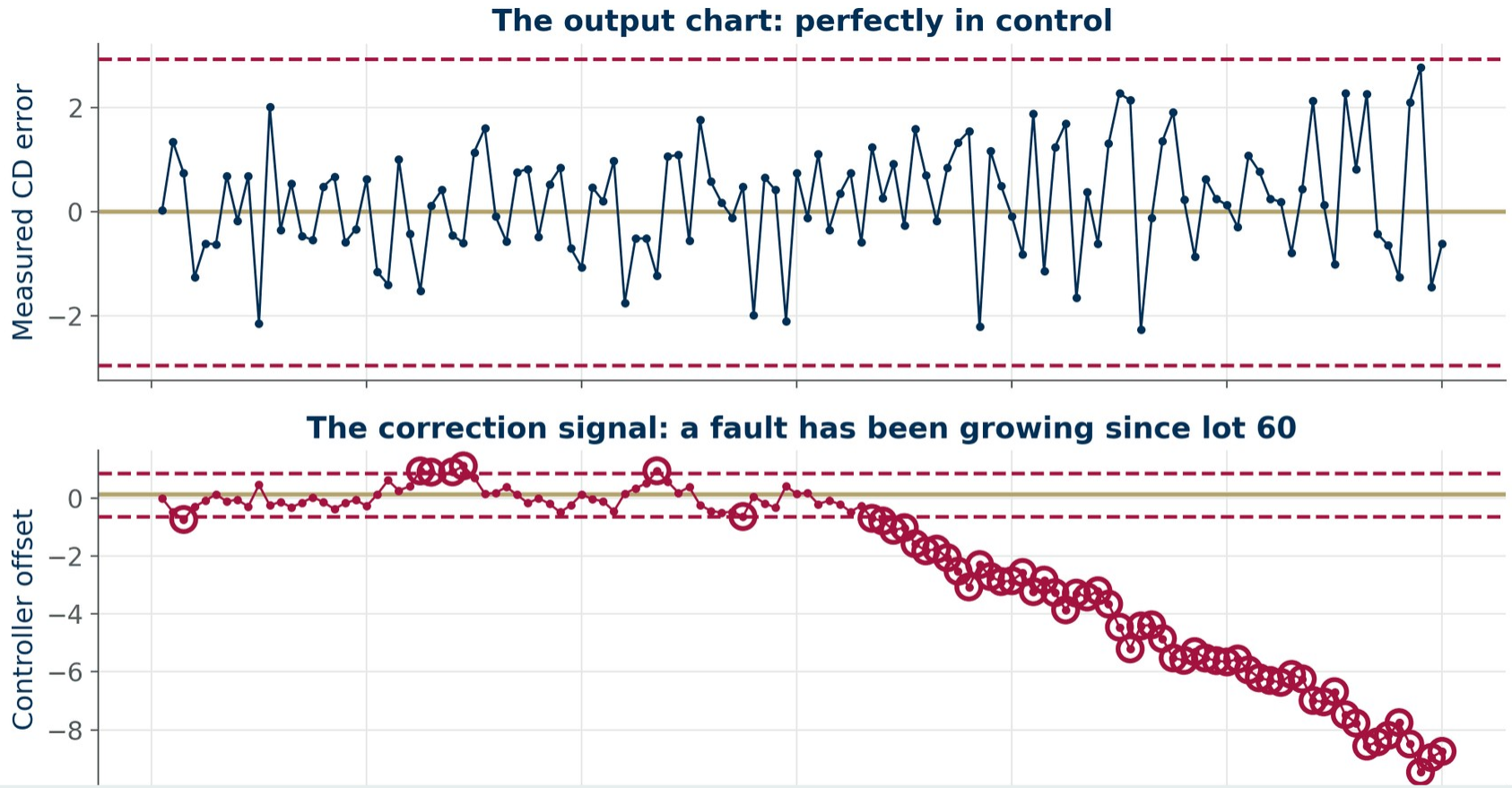
Georgia Tech

THE TRAP

A working controller hides the fault from you.

This is the single most counterintuitive interaction in the course, and it is worth an exam question.

The output looks perfect



Top: the CD error chart, beautifully in control. Bottom: the controller has been winding in a growing correction since lot 60.

Georgia Tech

Why this happens, and what to do about it

The mechanism

The controller's job is to cancel the output error. It does that job perfectly well whether the error comes from ordinary drift or from a developing hardware fault.

So the fault shows up in the correction, not in the output. The output chart is measuring the controller's competence, not the tool's health.

The fix

Chart the controller's correction signal as a process variable in its own right.

It has a normal range. A sustained trend in the offset means something is being compensated for — and eventually the controller will run out of range.

In a controlled loop, the output tells you whether control is working. The correction tells you what it is working against.

Georgia Tech

What to chart when a loop is closed

A checklist

1. **The output, as always**

   Confirms the loop is doing its job. Necessary, and no longer sufficient.

2. **The correction signal**

   Trends here are the fault the controller is hiding. Chart it with the same run rules.

3. **Headroom remaining**

   How close is the offset to its guardrail? A controller at 90% of its range is a scheduled surprise.

4. **Frequency of large moves**

   A rising rate of near-maximum corrections means the process is becoming harder to hold.

Georgia Tech

And chart the tool, not just the wafer

Fault detection and classification

Why bother

Metrology tells you about the wafer, after the fact, on a sample. The tool's own sensors tell you about the process, in real time, on every wafer.

FDC catches faults before metrology ever sees them.

What you get

Pressure, RF power and reflected power, gas flows, valve positions, temperatures, endpoint signals — hundreds of channels at 1 to 10 Hz.

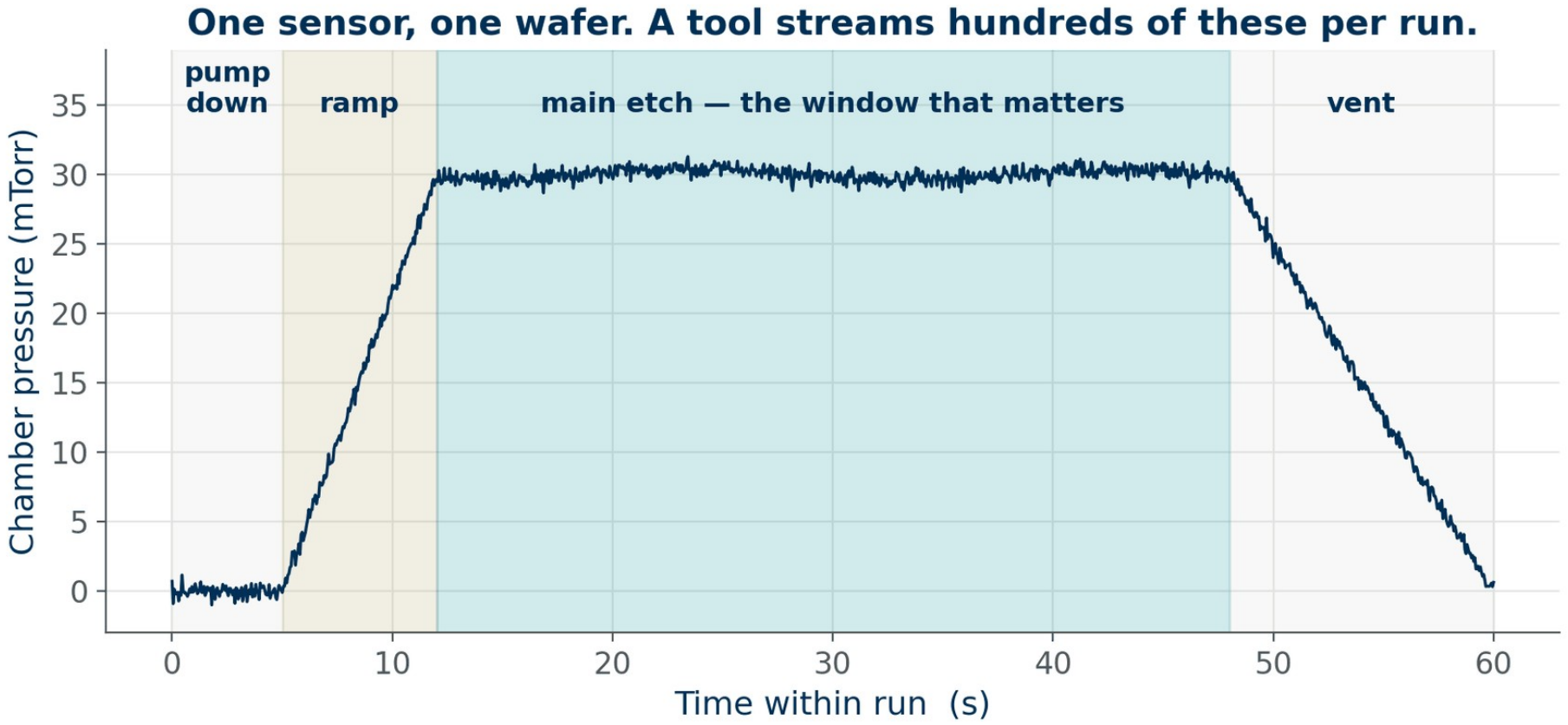
For every wafer, not every lot.

Which is a wonderful problem to have, and an entirely new statistical one.

Three hundred correlated channels, sampled continuously, on a process with legitimate step structure.

Georgia Tech

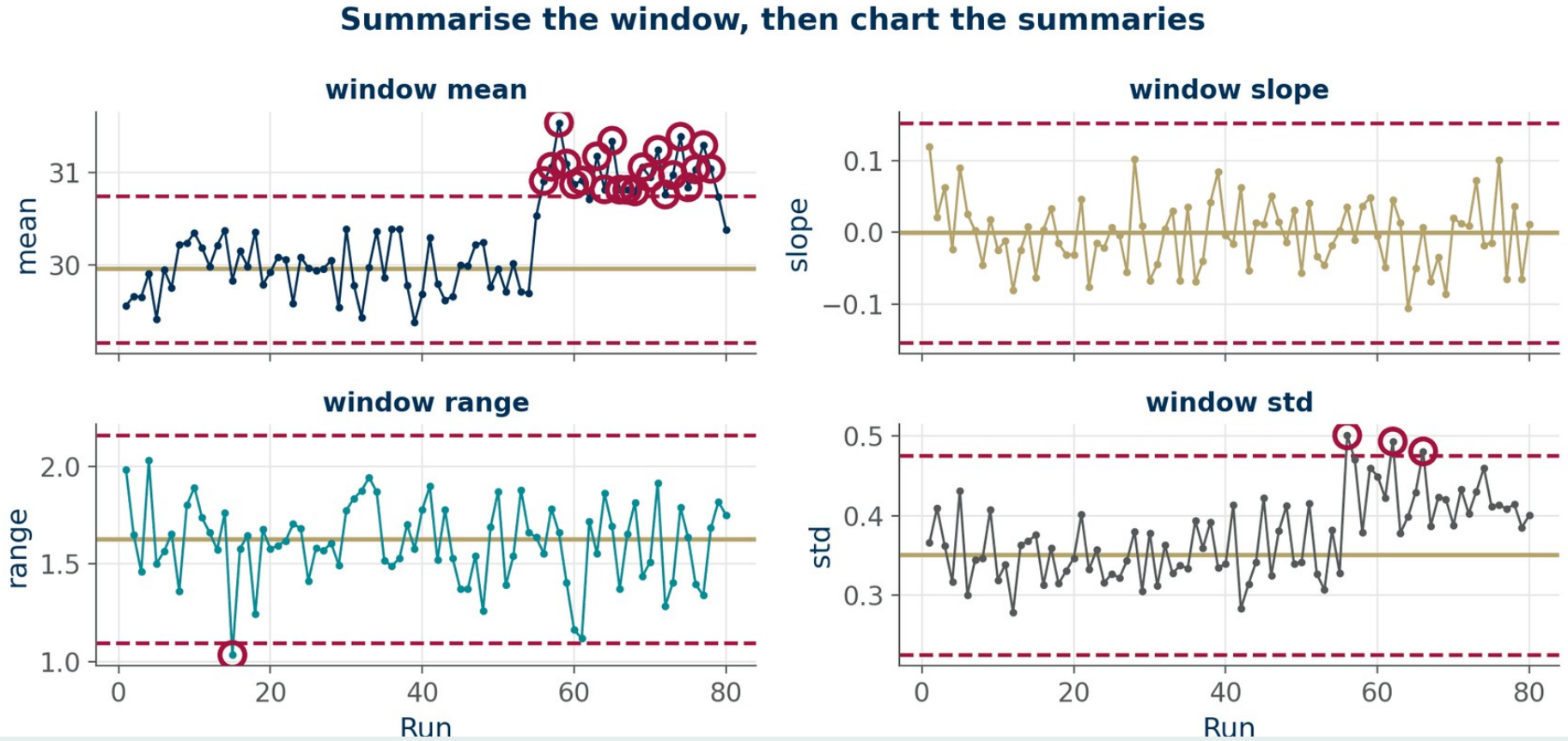
What a trace actually looks like



One sensor, one wafer, sixty seconds. Note the step structure — the recipe has phases, and only some of them matter.

Georgia Tech

From traces to something chartable

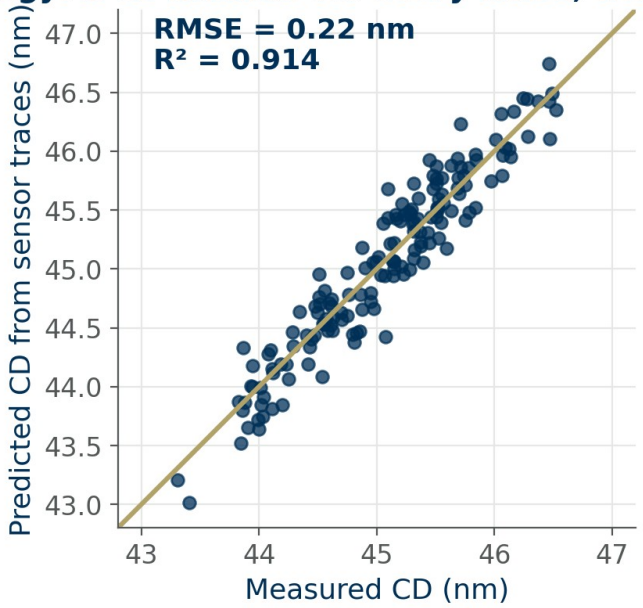


Segment by recipe step, summarise each window — mean, slope, range, standard deviation — and chart the summaries.

Georgia Tech

Virtual metrology

Virtual metrology: a CD number for every wafer, with no metrology time



- If sensor traces predict the measurement, predict it for every wafer.
- **You get a CD number on all 25 wafers instead of 3, at no metrology cost.**
- That means tighter subgroups, faster detection, and feedback for the R2R controller on lots that were never measured.
- It is also the bridge to the machine learning half of this course — this is a supervised regression problem with a real business case.
- **And it is genuinely deployed.** VM for etch depth and CD is production technology, not research.

Virtual metrology does not replace metrology. It fills in the wafers you could not afford to measure.

Georgia Tech

The danger with virtual metrology

You are charting a model output

If the model degrades, the chart moves — and nothing about the process changed. You now have two things that can drift.

Models degrade quietly

A chamber is cleaned, a sensor is replaced, a recipe is tweaked. The relationship the model learned no longer holds, and it keeps predicting confidently.

So chart the model too

Keep measuring a small sample of real wafers and chart the prediction error. That residual chart is your model-health monitor.

Every deployed model needs a control chart on its own residual. That is the single most transferable idea from this course into your projects.

Georgia Tech

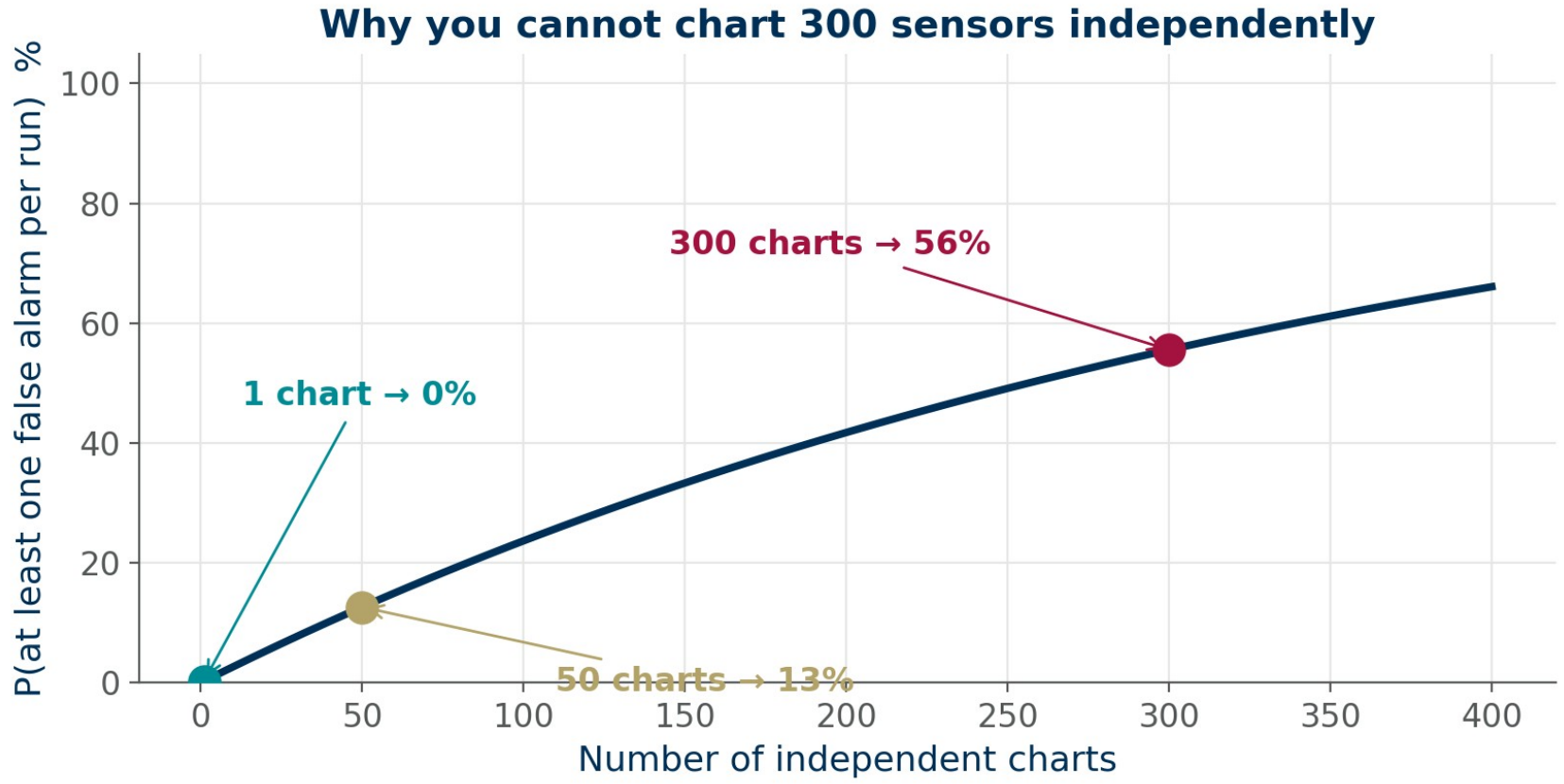
THE MULTIVARIATE PROBLEM

Three hundred sensors.
Three hundred charts. One alarm per run.

Everything we have built assumes you are watching one thing at a time.

Georgia Tech

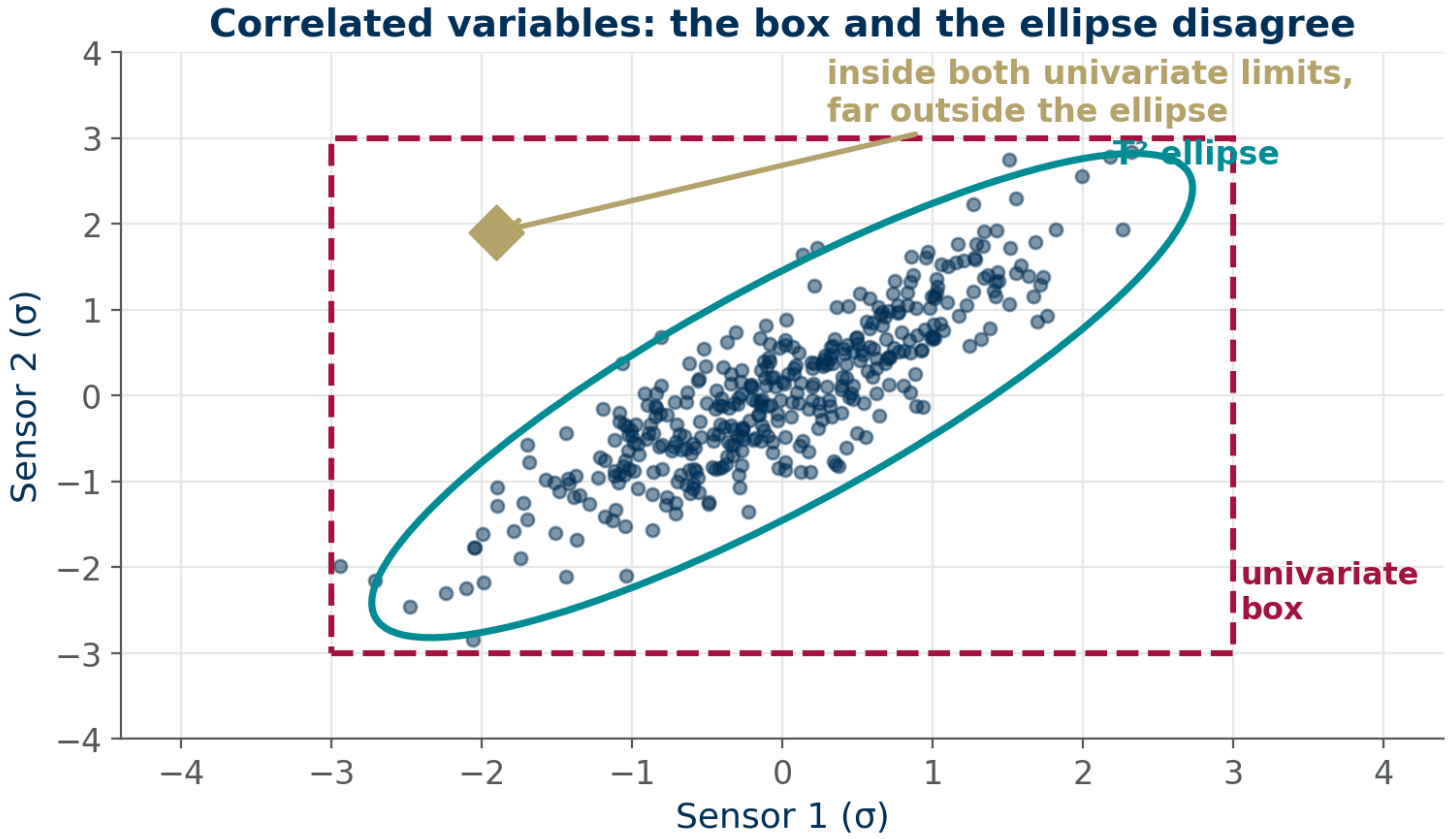
The multiplicity problem



At a 0.27% false alarm rate per chart, 300 independent charts give you a 56% chance of at least one false alarm — every single run.

Georgia Tech

And independence is not the only issue



Sensors are correlated. A point can sit inside every univariate limit and be far outside the joint distribution.

Georgia Tech

Hotelling's T^2

The multivariate generalisation of the z-score

$$T^2 = (x - \mu)^T \, S^{-1} \, (x - \mu)$$

It is a squared z-score

For one variable this reduces exactly to $((x - \mu)/\sigma)^2$. The inverse covariance matrix generalises 'divide by sigma' to many correlated dimensions.

It measures distance correctly

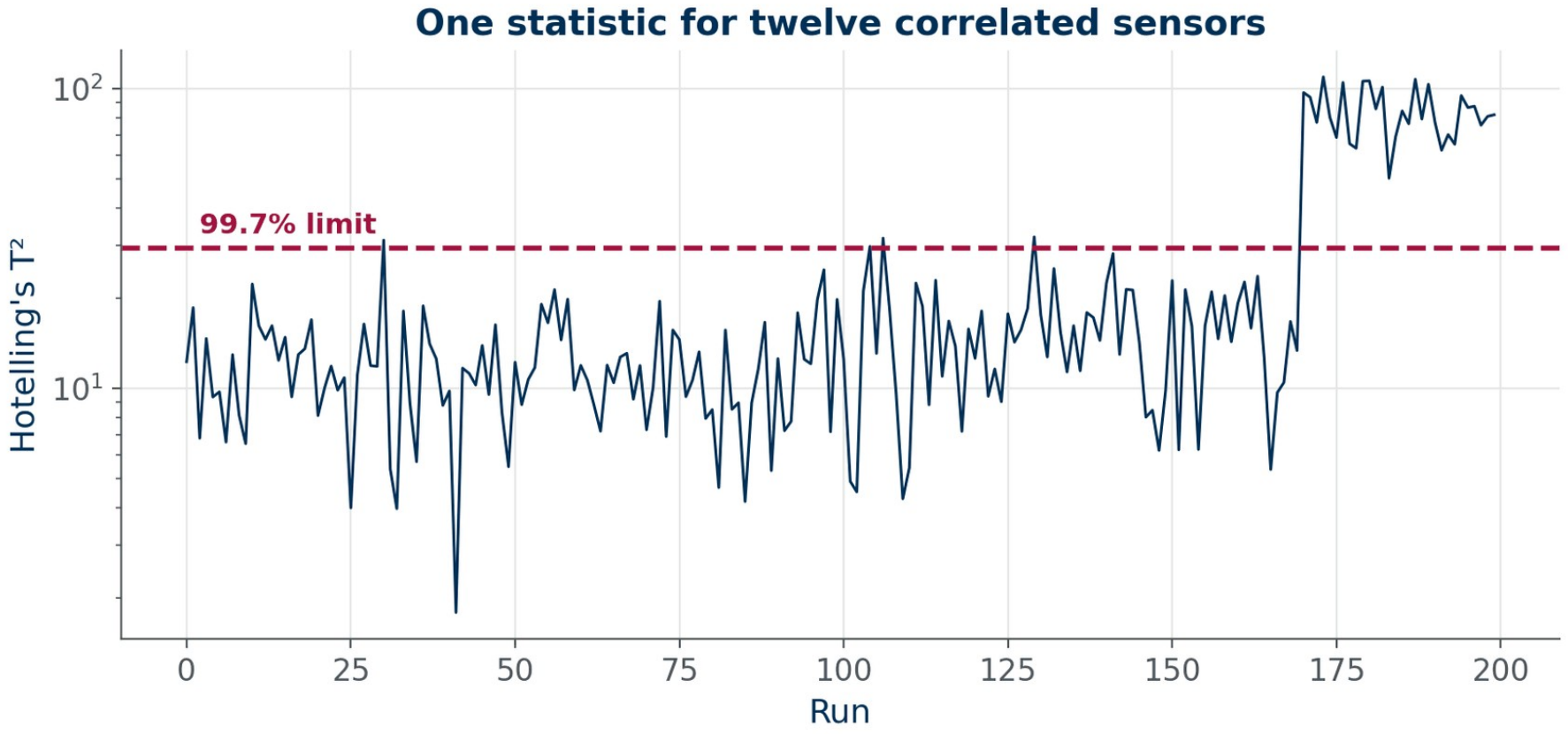
Distance in units of how much the data actually varies in that direction — which is why it draws an ellipse rather than a box.

One chart, many sensors

You plot a single statistic per run and compare it to one limit. The multiplicity problem disappears.

Georgia Tech

A T² chart



Twelve correlated sensors, one line, one limit. The excursion after run 140 is a coordinated shift no single sensor would have flagged.

Georgia Tech

What T^2 cannot do

It does not tell you which sensor

One number for everything means the alarm carries no diagnostic information on its own. You need contribution analysis.

It struggles when p is large

With hundreds of variables and limited history, the covariance matrix is poorly estimated or singular. You cannot invert what you cannot estimate.

It only sees inside the model

T^2 measures distance within the space your training data spanned. A genuinely new kind of behaviour may not register at all.

The first two are why PCA is almost always used with T^2. The third is why SPE exists.

Georgia Tech

PCA in one slide

The move that makes everything else possible

The observation

Three hundred sensors on an etch tool are not three hundred independent things. Pressure, flow and throttle position move together, because they are physically coupled.

The real dimensionality is far lower — often five to twenty.

What PCA does

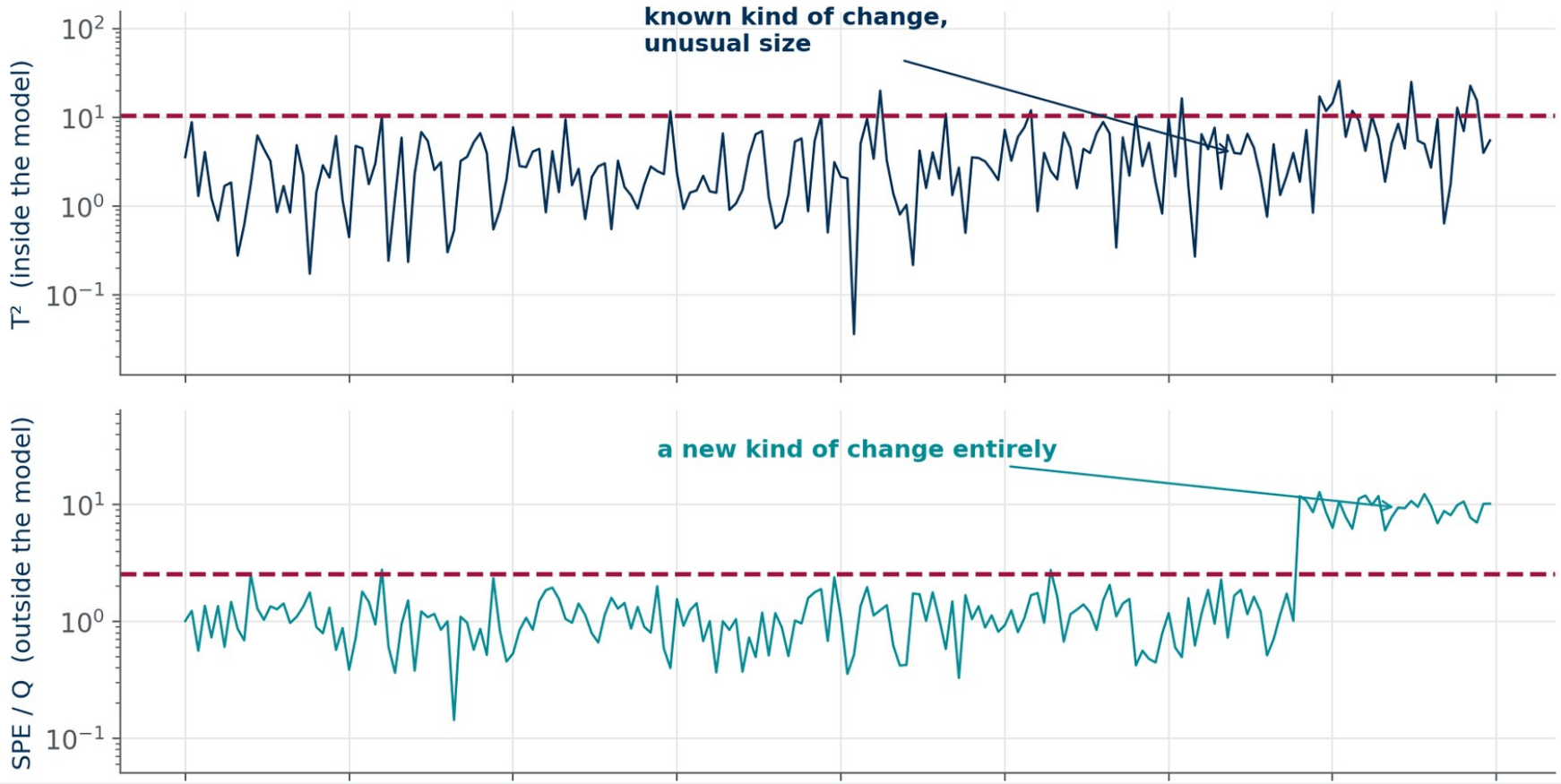
Finds the directions in which the data actually varies, ordered by how much. Keep the first few and you have captured most of the behaviour in a handful of numbers.

Now compute T^2 on those few, where the covariance is well estimated.

And the part you discarded is not garbage. It is a second statistic.

Georgia Tech

Two statistics, two kinds of fault



T² asks: is this an unusual amount of the normal kind of variation? SPE asks: is this a kind of variation I have never seen?

Georgia Tech

T^2 and SPE, side by side

You need both, and they mean different things

T^2 — inside the model

The run projects onto the normal directions, but unusually far along them.

Physical reading: the tool is doing a familiar thing to an unfamiliar degree. A drift, a shifted setpoint, a known mode taken too far.

SPE / Q — outside the model

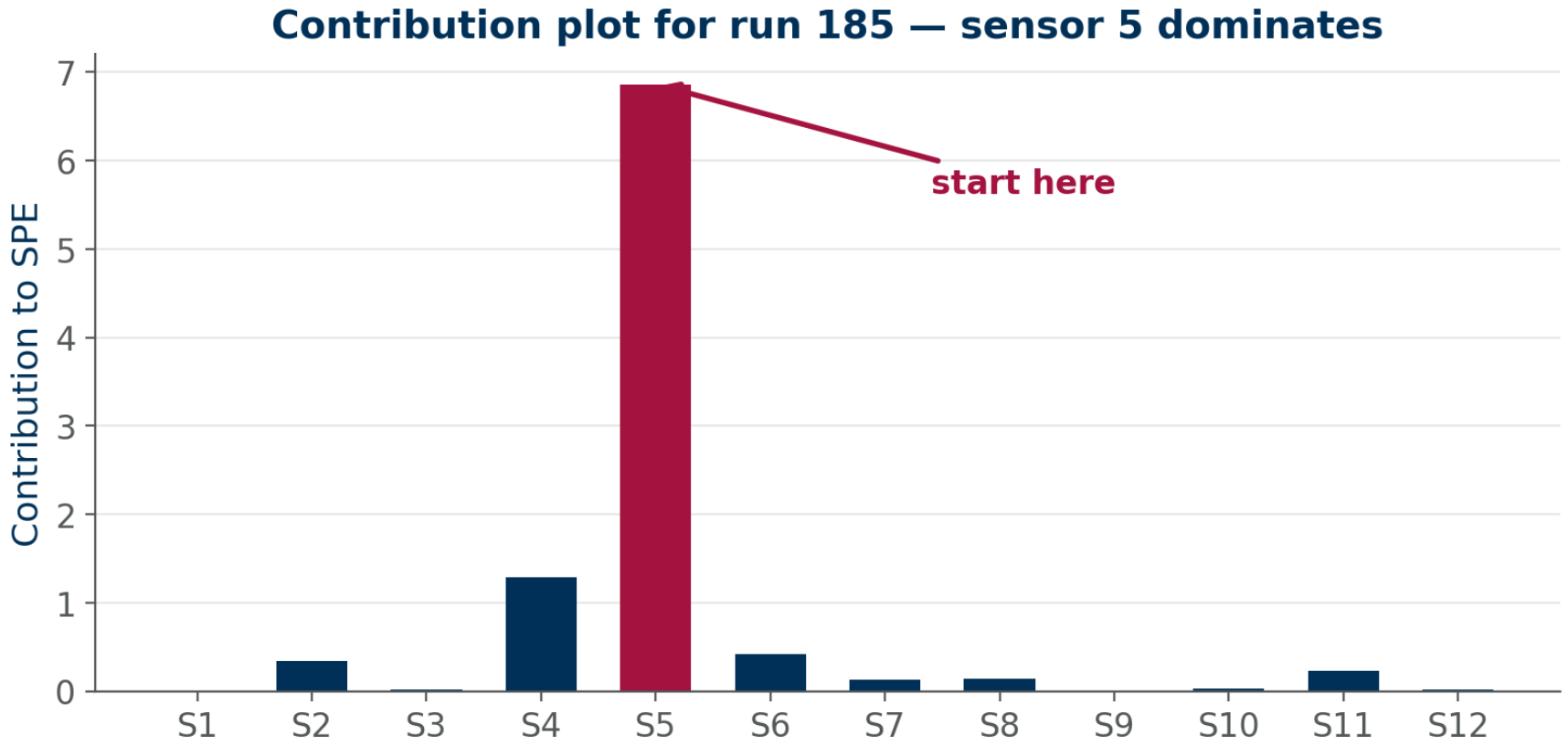
The run does not lie in the normal subspace at all. The correlation structure itself broke.

Physical reading: something new. A sensor failed, a valve stuck, a relationship that has always held stopped holding.

SPE usually fires first on genuinely novel faults — and it is the direct ancestor of autoencoder reconstruction error.

Georgia Tech

Diagnosis: contribution plots



Decompose the statistic back onto the original sensors. Sensor 5 is carrying the excursion — start there.

Georgia Tech

The honest limitation: smearing

What the contribution plot promises

'Sensor 5 caused this.' Clean, actionable, and what everyone wants from a multivariate alarm.

What it actually delivers

Because the sensors are correlated, a fault in one variable spreads its contribution across all the variables it is correlated with. The true culprit may not have the largest bar.

This is called smearing, and it is a well-documented limitation.

So read it as a ranked list of where to look — not as an answer.

The same caveat applies to SHAP values on a neural network, for exactly the same mathematical reason. Attribution among correlated inputs is genuinely ambiguous.

Georgia Tech

Beyond linear

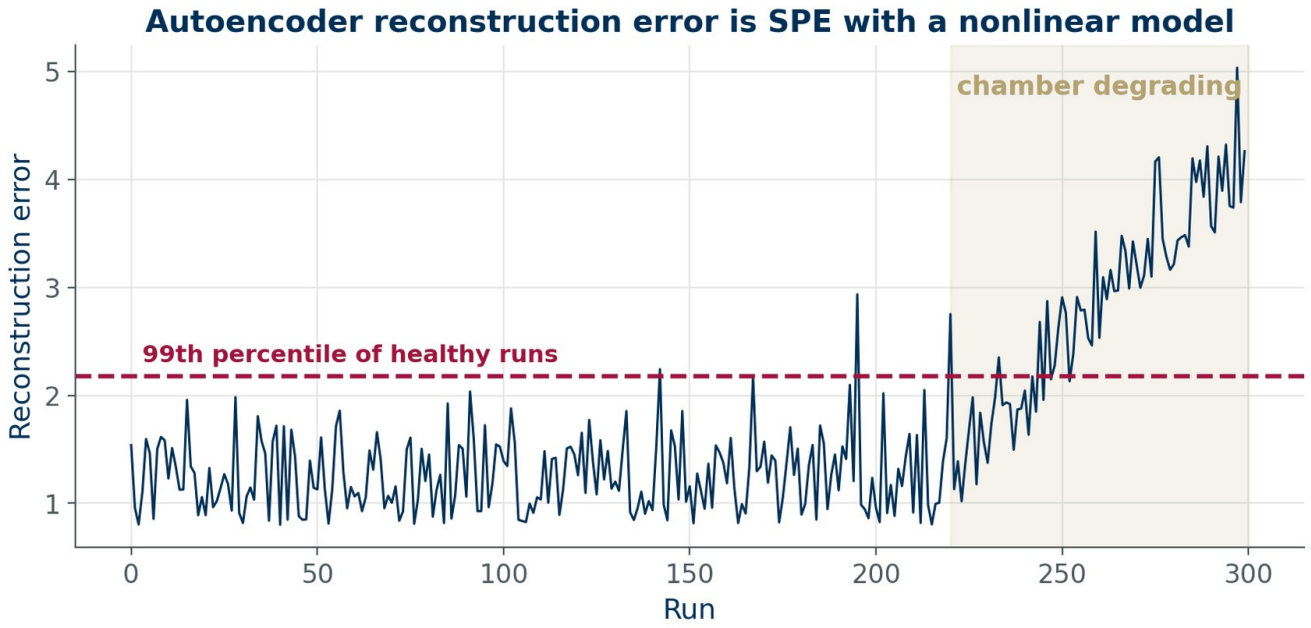
Where the machine learning starts

Method	What it generalises	When it earns its complexity
Kernel PCA	PCA, with a nonlinear feature map	Curved normal-operating regions
Dynamic PCA	PCA on time-lagged copies of the data	Autocorrelated processes — the I–MR problem from Lecture 2
One-class SVM	A learned boundary around normal, with no distribution assumed	Non-elliptical normal regions
Isolation forest	Anomaly score from how easily a point is separated	High dimension, mixed variable types, minimal tuning
Autoencoder	SPE, with a nonlinear encoder and decoder	Complex, nonlinear normal behaviour and plenty of healthy data

Every row is answering Shewhart's question. They differ in what shape they allow 'normal' to be.

Georgia Tech

The autoencoder is SPE with a nonlinear model



- Train a network to compress each run and reconstruct it, using only healthy runs.
- **Reconstruction error is the residual. It is SPE.**
- A healthy run reconstructs well, because it resembles the training data. A novel fault reconstructs badly, because the network never learned that pattern.
- The limit is set the same way: a high percentile of the healthy distribution.
- **Structurally identical to PCA-based SPE.** The only change is that the model is nonlinear, so 'normal' can be a curved manifold instead of a flat subspace.

If you understood SPE, you already understand this. That was the point of doing PCA first.

Georgia Tech

Modelling the trace directly

Skipping the feature engineering

1-D CNN

Learns local shape features from the raw trace — ramps, overshoots, settling behaviour — instead of you specifying windowed statistics.

LSTM / GRU

Handles variable-length runs and long-range dependence naturally. Older, still competitive on modest datasets.

Transformer

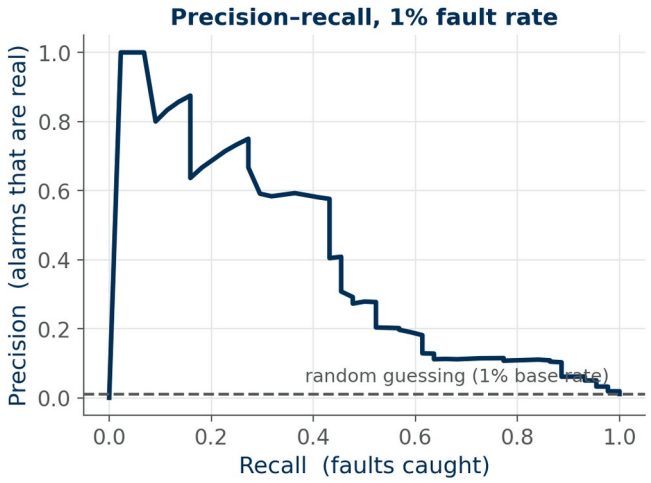
Attention over the whole run. Strong when you have a great deal of data, and increasingly the default for pretraining on unlabelled traces.

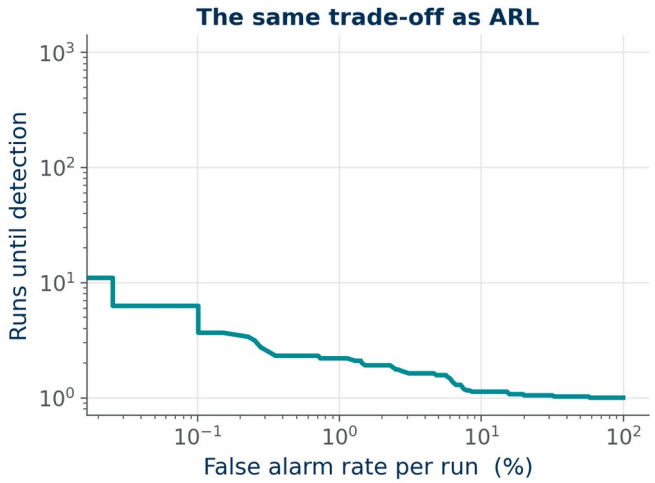
But label scarcity is the binding constraint, not model capacity.

A fab has millions of unlabelled runs and perhaps a few hundred confirmed faults. Self-supervised pretraining plus anomaly detection usually beats a supervised classifier.

Georgia Tech

Evaluating a detector honestly





- Faults are rare — often well under one percent of runs.
- **So accuracy is meaningless. A model that predicts 'healthy' always scores 99%.**
- Use precision and recall, and report the precision–recall curve rather than a single operating point.
- **And the right-hand plot should look familiar.**
- Detection delay against false alarm rate is exactly the ARL trade-off from Lecture 1, drawn in machine learning vocabulary.

You are still choosing a point on the same curve Shewhart chose in 1924.

Georgia Tech

The vocabulary translates

Same concepts, different community

Statistical process control	Machine learning	The shared idea
Type I error / false alarm	False positive	You stopped a healthy tool
Type II error / missed shift	False negative	You kept making bad wafers
ARL_0 (points between false alarms)	1 / false positive rate	How often you cry wolf
ARL_1 (points until detection)	Detection delay	How long the damage runs
Control limit	Decision threshold	The knob you choose
Contribution plot	Feature attribution / SHAP	Where do I look first?

If you can move between these two columns fluently, you can read both literatures.

Georgia Tech

Deployment is where these projects die

Four problems, all of them real

Concept drift

PMs, chamber rebuilds, recipe changes and new products all move the definition of normal. A model trained in March alarms constantly by August.

Tool and chamber matching

A model trained on chamber A rarely transfers to chamber B without adaptation. You may need one model per chamber, or a transfer-learning scheme.

Label scarcity

Confirmed root causes are rare and inconsistently recorded. The OCAP log from Lecture 2 is the most valuable dataset you are not collecting.

Alarm fatigue, again

A model with better recall and worse precision than the existing chart will be switched off, regardless of its AUC.

The winning system is rarely the most accurate one. It is the one the engineers still trust after six months.

Georgia Tech

A hundred years, one question

Era	Method	What is 'normal'	What is charted
1924	Shewhart chart	A mean and a σ	One measurement
1930s	Run rules	A mean, a σ, and patterns	Sequences of points
1950s	EWMA / CUSUM	Plus a memory of the past	Accumulated evidence
1980s	Run-to-run control	A target you actively hold	The correction signal
1990s	PCA · T^2 and SPE	A linear subspace	Distance in and out of it
2010s	Autoencoders	A nonlinear manifold	Reconstruction error
Now	Sequence models	A learned distribution over traces	Likelihood or residual

The models get more flexible. The question — has this process changed? — has not moved since 1924.

Georgia Tech

Homework and project prompt

Due as posted · datasets on the course site

1. Gauge R&R: run the supplied 10 × 3 × 2 study. Report repeatability, reproducibility, %R&R against both tolerance and total variation, and ndc. State whether you would accept the gauge and why.

2. Take your Cpk from Lecture 2 and correct it using the gauge σ from problem 1. State whether the engineering conclusion changes, and what you would do differently.

3. Residual charting: fit a seasoning model to the etch dataset, chart the residual, and compare the alarms to the naive chart. Report both false alarms avoided and real signals gained.

4. Multivariate: build a PCA model on the healthy portion of the supplied FDC dataset. Chart T^2 and SPE, and use contribution plots to identify the faulty sensor. Then discuss how you would validate that identification given smearing.

Georgia Tech

References, and where this goes next

Core reading

Montgomery Ch. 8.7 (gauge capability) and Ch. 11 (multivariate SPC).

May & Spanos Ch. 6–7 — semiconductor SPC and run-to-run control.

Qin, 'Survey on data-driven industrial process monitoring and diagnosis' — the standard review of T^2, SPE and their extensions.

Chiang, Russell & Braatz — fault detection and diagnosis, the reference text.

Where the course goes from here

Everything after this unit is the rightmost column of the last table.

Virtual metrology as supervised regression. Fault detection on trace data. Self-supervised pretraining. Transfer between chambers.

Benchmarks to know: SECOM for high-dimensional imbalanced fab data, and Tennessee Eastman as the canonical process-monitoring testbed.

Every model you build for the rest of this course needs a control chart on its own residual.

That is the through-line from Shewhart to your project.

Georgia Tech