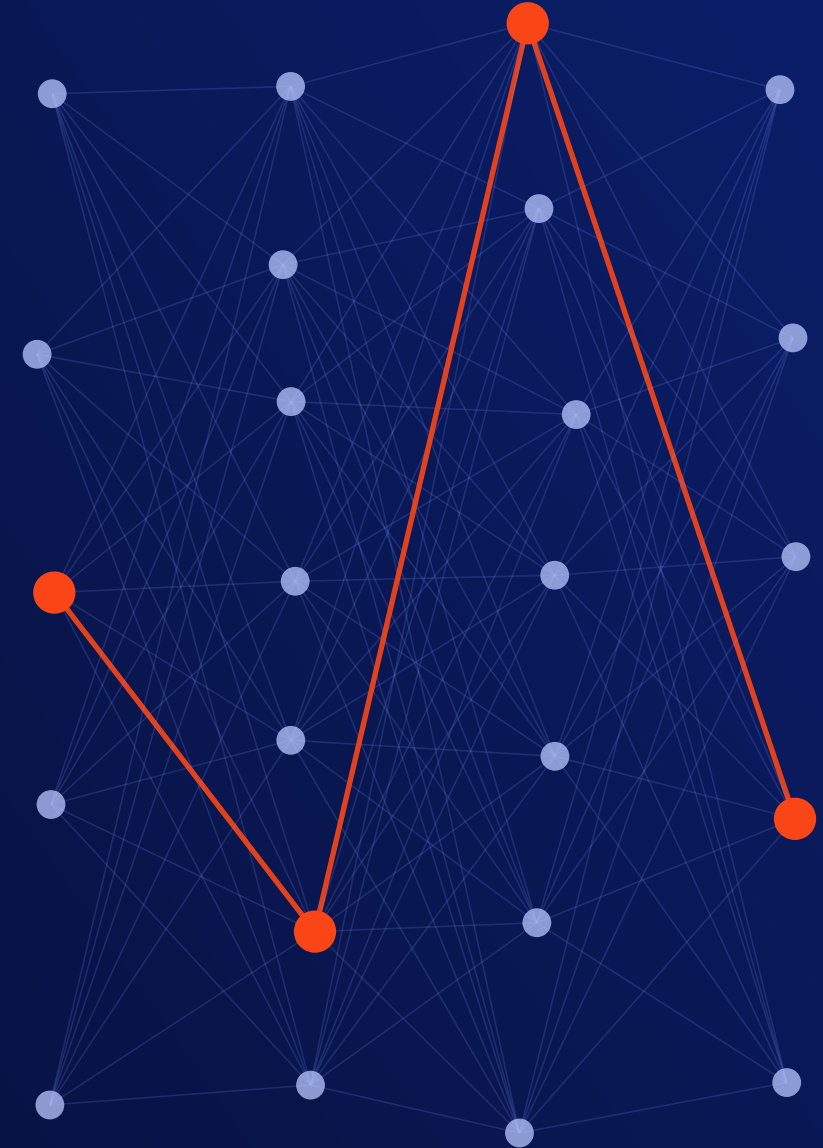

LECTURE 12-1

Evaluation, Robustness, and Fairness

What a score does and does not tell you



Parisa Rashidi · Herbert Wertheim College of Engineering, University of Florida



Where We Are

The last two weeks are about what happens after the model works.

Week 11

**Generative models and
AI for science**

Diffusion for structure and
design

Week 12: today

**Evaluation, robustness,
fairness**

What a score does and does
not tell you

Week 13

**Deployment and human-
AI teaming**

A model meets a hospital

Three Ways a Good Score Misleads

Each section of today answers one of them.

A benchmark is not a clinic

high score, no change in what clinicians do

A point is not a result

one number from one sample, no interval

An average hides

a shortcut the model reads, and the group it fails



**A Benchmark Score Is
Not a Clinical Effect**

The Trial

50 physicians, randomized, with and without GPT-4.

Participants

50 physicians: 26 attendings, 24 residents

Randomized

25 with GPT-4 and usual resources; 25 usual resources only

Task

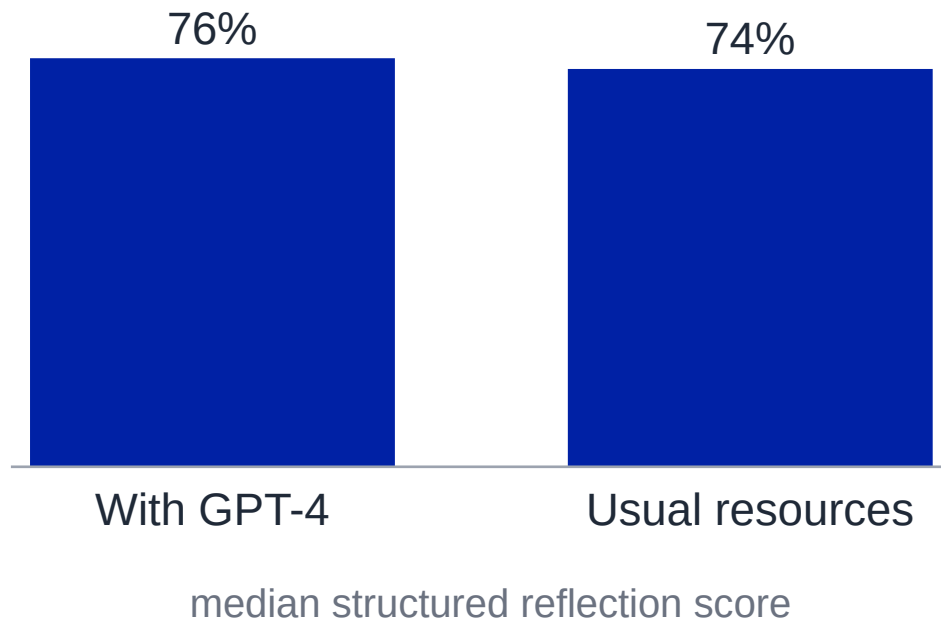
up to 6 clinical vignettes in 60 minutes; 244 cases completed

Outcome

structured reflection score, graded by blinded raters

The Primary Result

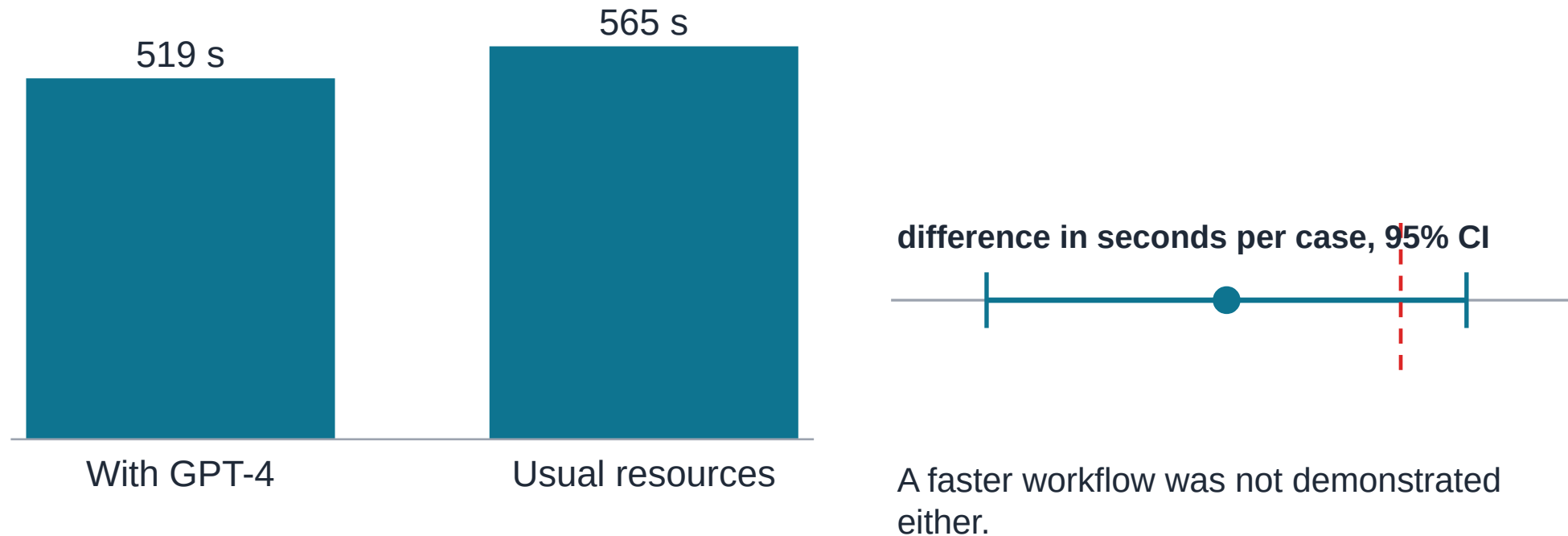
Median 76% with the model, 74% without. The difference: 2 points, 95% CI -4 to 8.



The interval contains 0, and also contains 8: this does not show the model helps, and does not show it does not.

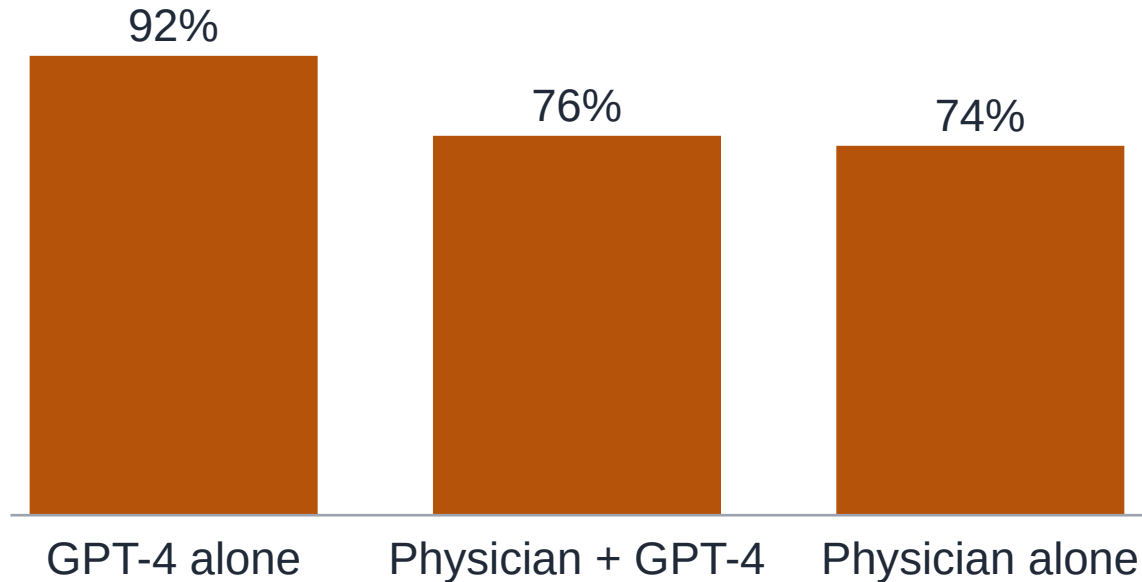
Nor Was It Faster

519 seconds per case against 565: a difference of -82 s, 95% CI -195 to 31 .



The Model Alone Scored 92%

And that is the number to read most carefully.



The gap is about use, not about the model.

Read with care

median of 3 runs

16 points over control,

95% CI 2 to 30

exploratory, not the

randomized comparison

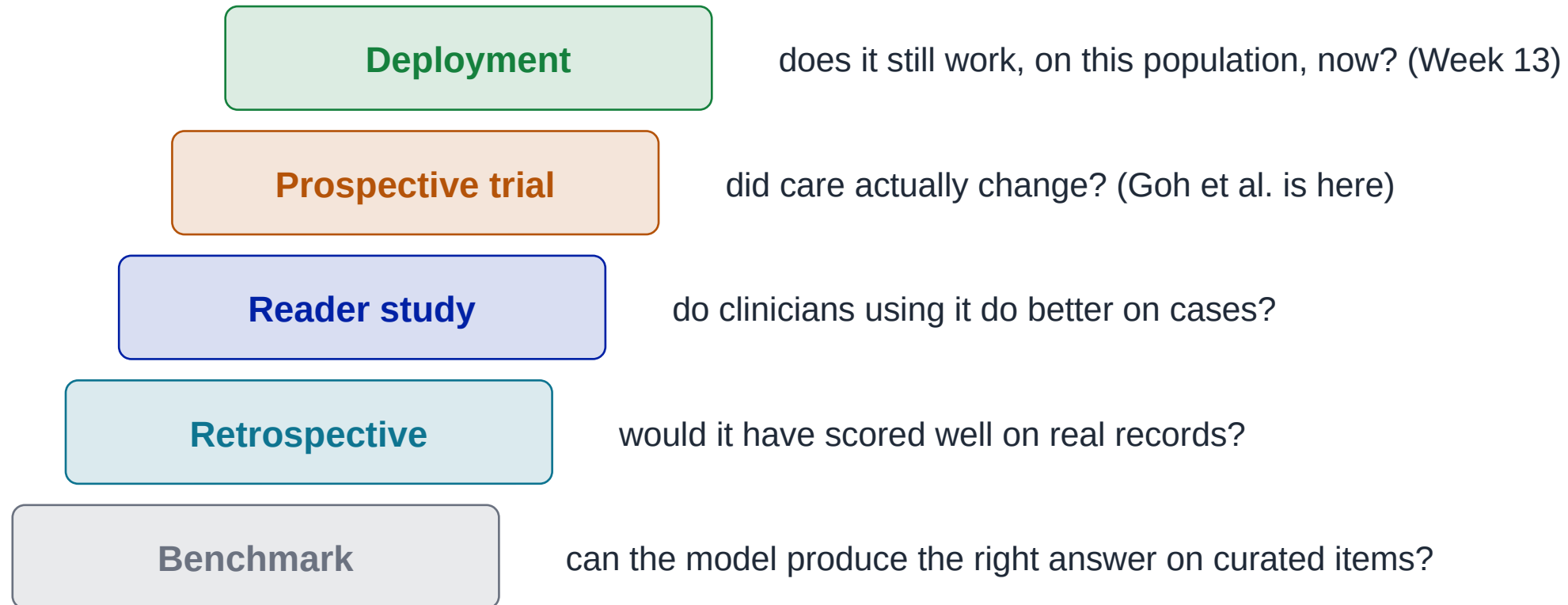
Animation: A Benchmark Score Is Not a Clinical Effect

Animation clip: week12-clip1.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

The Evaluation Ladder

Each rung answers a question the one below it cannot.





How Sure Are We?

Two Models on 200 Patients

A is right on 170, B on 164. Is A better?

Model A

$$170 / 200 = 85\%$$

Model B

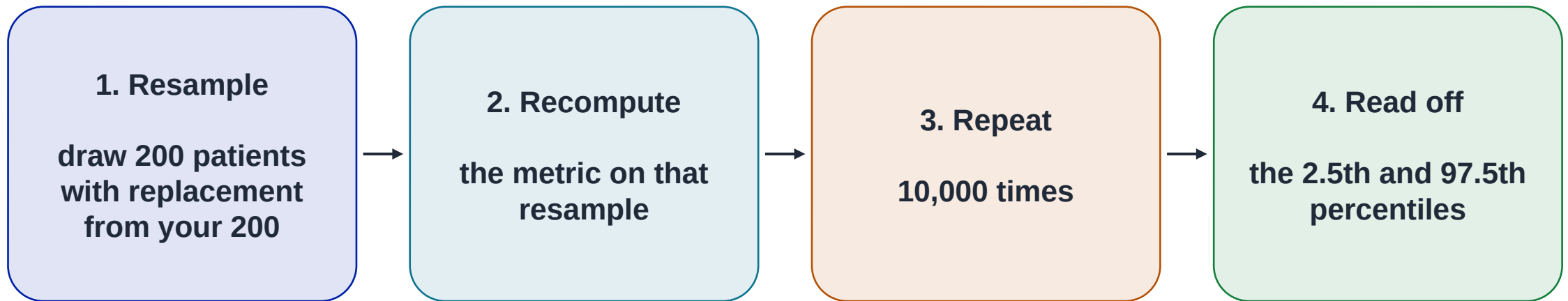
$$164 / 200 = 82\%$$

Difference: 3 percentage points. On 200 patients, is that a real difference?

Synthetic example, as in clip 2

The Bootstrap

Resample the patients you have, and watch the number move.



No formula and no distributional assumption: it works for AUC, F1, or anything else you can recompute.

Overlapping Intervals Are Not a Test

Both intervals are wide, and they overlap. That settles nothing.

Model A accuracy, 95% CI 0.800 to 0.900



Model B accuracy, 95% CI 0.765 to 0.870



0.74

0.93

Overlapping intervals are not a test of the difference.

Pair the Comparison

The two models were graded on the same patients, so compare them that way.

	B right	B wrong
A right	156	14
A wrong	8	22

Where the answer lives

178 patients agree:
they carry no information

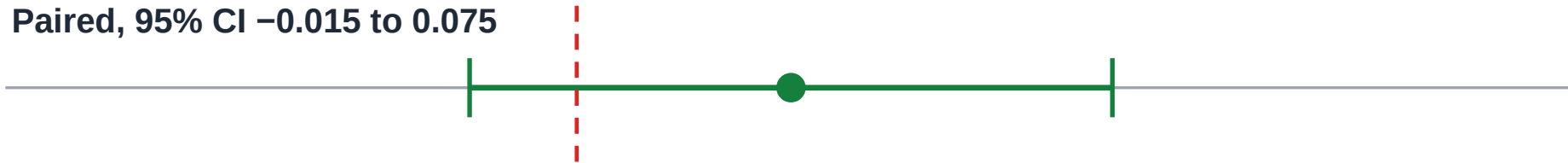
22 disagree: 14 to A, 8 to B

Paired difference: 0.03, 95% CI -0.015 to 0.075. The interval includes 0.

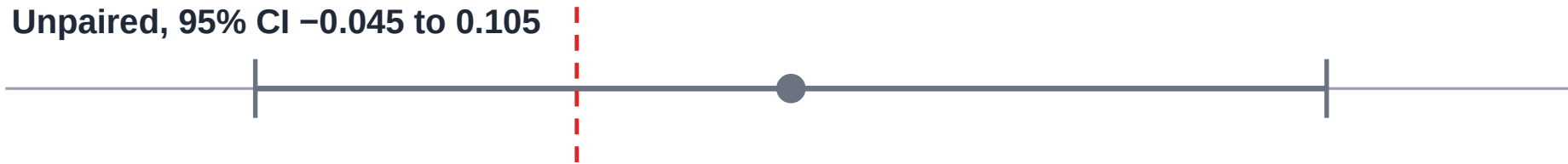
Pairing Is Worth Two Thirds of the Width

Same data, same difference, a much narrower interval.

Paired, 95% CI -0.015 to 0.075



Unpaired, 95% CI -0.045 to 0.105



dashed line: no difference

Animation: How Sure Are We? The Bootstrap

Animation clip: week12-clip2.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

In the Notebook: Two Extras

A pair that looks identical and is not, and the sample size you would need.

	Our pair	Hypothetical pair
Accuracy A, B	85%, 82%	85%, 82%
Marginal intervals	identical	identical
Disagreements	22	6
Paired 95% CI	-0.015 to 0.075	0.010 to 0.055
Conclusion	not shown	A is better

About 466 patients would be needed to resolve a 3-point difference of this kind.



Shortcuts and Hidden Subgroups

Two Datasets, One of Them Fair

The second dataset was the favorable case, and it failed anyway.

Dataset I: the worst case

positives: GitHub-COVID

negatives: NIH ChestX-ray14

source and label almost coincide

Dataset II: the favorable case

both classes from one region
and a shared research team

shortcut learning happened anyway

What the Models Were Reading

Saliency maps and generative counterfactuals, not guesswork.

Markers and text

laterality markers; arrows and annotations unique to one source

Image borders

the radiopacity of the margins of the film

Positioning

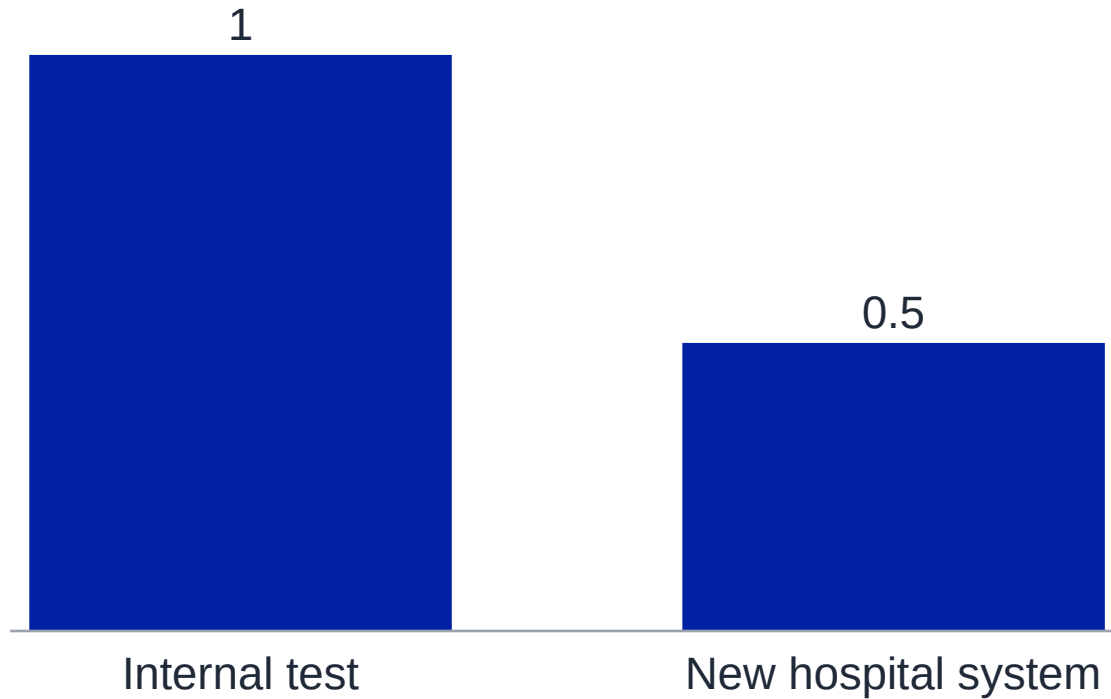
shoulders and clavicles, diaphragm, cardiac silhouette, AP versus PA

And some signal

the counterfactuals also changed lung opacity

Half the Performance Was the Shortcut

On radiographs from a new hospital system, half of it is gone.



predictive performance, internal set = 1

In the paper

"half of the model's predictive performance is lost when testing on Dataset II"

and only a minority was genuine pathology

Some Shortcuts Travel

Which is why external validation is necessary and not sufficient.

Shortcut that does not travel

a marker only one source uses

external validation catches it

Shortcut that travels

patient positioning, AP versus PA

external validation passes,
and the model still reads the margins

External validation is necessary, not sufficient.

Three Checks That Do Not Need a New Site

From the Week 12 notebook, with its measured numbers.

Check	What the notebook found
Paint out the marker	site C falls from 1.000 to 0.500
Stamp a marker on negatives	mean predicted probability rises from 0.001 to 0.696
Probe the features for the site	site predicted at 0.695, against 0.50 by chance
Train with no marker anywhere	0.947, 0.925, 0.895 across the three sites

The toy is deliberately more extreme than the paper: it lands at chance, the real classifiers lost about half.

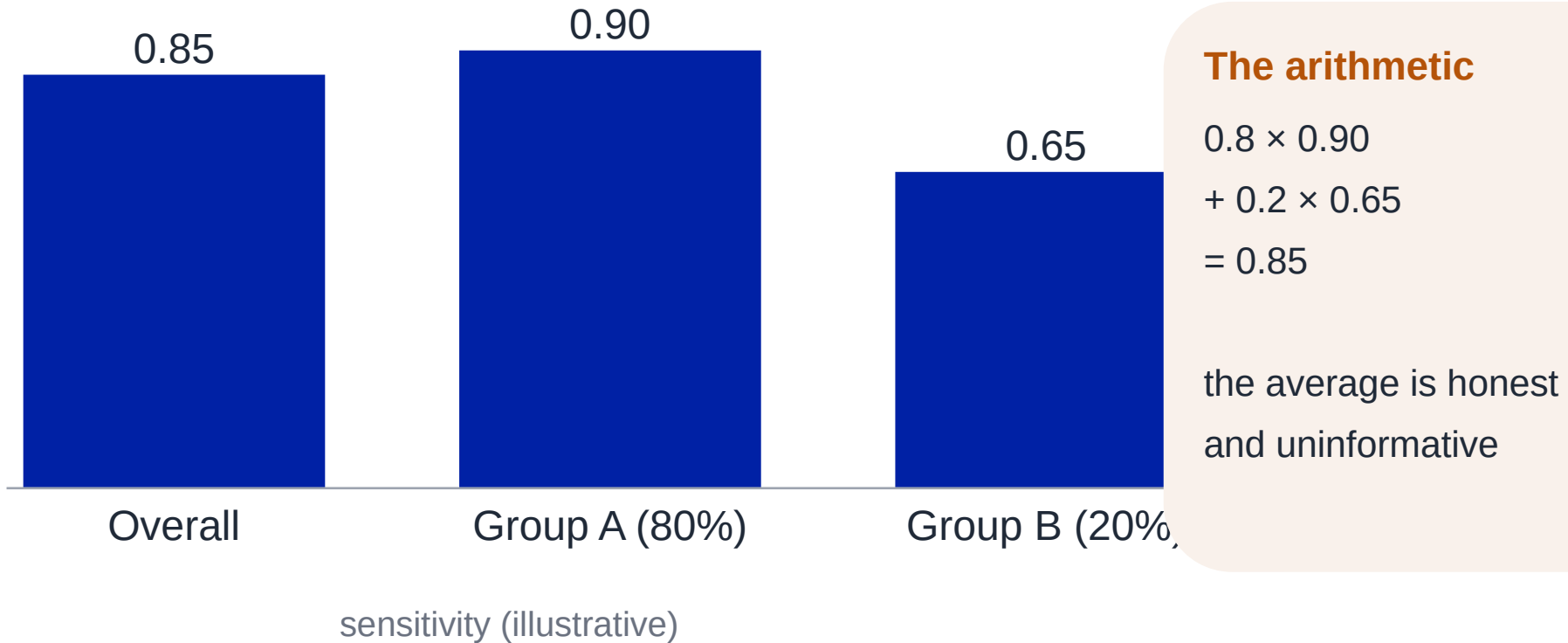
Animation: Shortcuts and Hidden Subgroups

Animation clip: week12-clip3.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

An Average Hides the Group It Fails

Sensitivity 0.85 overall: 0.90 in 80% of patients, 0.65 in the other 20%.



A Subgroup Number Needs an Interval

In the notebook's draw of 400, group B had 74 patients.

Group B sensitivity: estimate 0.743, 95% CI 0.635 to 0.838



dashed line: the true value, 0.65

The estimate missed by nearly 10 points on 74 patients.

The interval is what kept the truth in view.

Recap: What a Score Does Not Tell You

Effect

A benchmark is not a clinic. Goh et al.: 2 points, 95% CI -4 to 8.

Interval

Bootstrap it. Pair the comparison: the width falls by two thirds.

Mechanism

Ask what the model reads. External validation is necessary, not sufficient.

Subgroups

0.85 can be 0.90 and 0.65. Report the group, with its interval.

Before Thursday

Read

Goh et al., LLM influence on diagnostic reasoning: an RCT, and DeGrave et al., COVID-19 shortcuts.

Write

Reading responses on Canvas before 1:00 PM Thursday.

Explore

Week 12 reference notebook (ungraded): the paired bootstrap, a shortcut you can catch three ways, subgroups.

Next week

Deployment, human-AI teaming, and ambient intelligence: the last content week.