

AI, ML, DL, GenAI, Oh my!

What is all this and how does it work?

Sean Davis, MD, PhD

University of Colorado Anschutz School of Medicine

October 14, 2026



What we will cover

- **AI, ML, DL, GenAI:** how the four fit inside each other
- **One model, start to finish:** the eGFR, fitting a model to data, adding a predictor
- **The fitting loop,** and two questions that sort any model
- **What goes wrong:** models that are too flexible, neural networks, and why a score on cases a model has already seen proves nothing
- **Back to the clinic:** what this changes on the wards

Questions along the way are live: answer on your phone or laptop. Thursday starts with the chatbot half of this story.

Speaker notes

Welcome. This session takes about sixty minutes live, or about the same on your own with the pauses.

Here is the plan. First, four words, AI, machine learning, deep learning, and generative AI, and how they fit inside each other. Then one model from start to finish: the eGFR on a chemistry panel, where its number comes from, how a model is fit to data, and what happens when you add a predictor. Then the fitting loop that every model comes from, and two questions that sort any model.

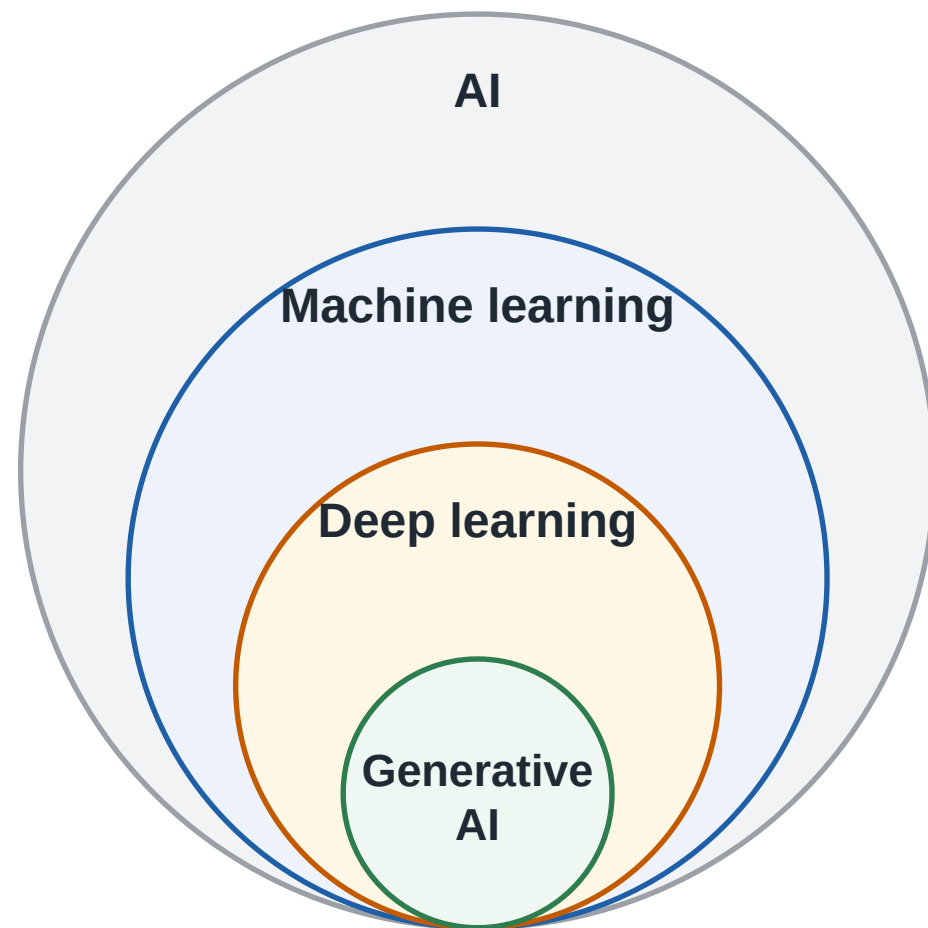
After that, what goes wrong. A model can be too flexible. A neural network is the same loop with far more dials. And a score on cases a model has already seen proves nothing. Then back to the clinic, and what this changes on the wards.

The questions along the way are live: you answer on your phone or laptop. Thursday starts with the chatbot, which is the same idea with words.

LIVE: Run sheet, with minutes per block, for about 60 minutes of talk and 3 for questions. Before class: open this deck once with ?presenter=YOUR-KEY so the live questions open. 0 to 9 (9 min): slides 1 to 5; slide 4 is the setup (target, model, predictors, error, loss), about a minute and a half, with one click for the loss box; slide 5 is height and weight, move the two sliders yourself for about two minutes, press “Show the best fit”, click once to name the fitting loop, then click again for “What is missing?” and take two or three answers aloud. 9 to 19 (10 min): slides 6 to 11, with the live question on slide 8 (about 45 seconds, then press A); on slide 10 the room has already fit a line, so move the sliders for only about a minute, then press “Show the best fit”; on slide 11 ask the room which way the average miss will move, then turn on “Add sex”. 19 to 35 (16 min): slides 12 to 20; slide 14 one minute, pointing at the padlock; slide 15 two minutes; slides 18 and 19 are live questions too, about 45 seconds each, then press A to reveal; slide 20 ninety seconds, pointing at the Type column and then at the last column. 35 to 52 (17 min): slides 21 to 27; on slide 21 move the slider yourself, about two minutes, and let the room call out where it goes wrong; slide 22 two minutes; slide 25 three minutes, as a callback to Tuesday’s talk; slide 27 is the bridge to Thursday, one minute. Keep about two minutes in this block for questions from the room. 52 to 60 (8 min): slides 28 to 31; read slide 30 in one minute; end on the Thursday line on slide 31. 60 to 63: questions, on slide 35, the last slide (press End from slide 31). Slides 32 and 33 are for students afterwards; slide 33 has the TensorFlow Playground exercise. Slide 34 is the reference list.

LIVE: Turn on Zoom live captions and paste the deck link into the chat at the start. Paragraphs in these notes that start with PAUSE are for the recorded version; skip them live. ChatGPT Edu and Microsoft Copilot are this course's institutional tools; swap in your school’s. Click outside a slider or button before using the arrow keys.

Four words, each inside the one before



- **Artificial intelligence**
A computer doing a task that seems to need judgment.
Example: MYCIN (1970s), rules written by hand.
- **Machine learning**
The rules are learned from data, not written by hand.
Example: the eGFR equation.
- **Deep learning**
Machine learning with many layers: a neural network.
Example: a model that reads chest X-rays.
- **Generative AI**
Deep learning that writes new text or makes images.
Example: a chatbot.

Each is one kind of the one around it. Yesterday we saw how we got here.
Today: how the models inside these circles are made.

Speaker notes

Yesterday’s talk was the history: how we got from rules written by hand to the chatbots of today. Here are four words from that history, and how they fit together. If you missed yesterday, nothing today depends on it.

Artificial intelligence is the largest circle. The sizes of the circles aren’t meant to show how big each field is. It means a computer doing a task that seems to need judgment, by any method. MYCIN, from the 1970s, advised on antibiotics with rules that experts wrote by hand.

Machine learning is inside it. The rules aren’t written by hand. They are learned from data. The eGFR equation on every chemistry panel is an example. A statistician would call it a regression. The line between statistics and machine learning is blurry, and for this course a model learned from data counts.

Deep learning is inside machine learning. It uses a neural network: many layers of simple units, each passing numbers to the next. Models that read images, such as chest X-rays, are usually built this way.

Generative AI is inside deep learning. These models write new text or make new images. A chatbot is the example you know. So a chatbot is generative AI, and it’s also deep learning, machine learning, and AI. The words overlap, but each one means something specific.

Today we start in the second circle, with the eGFR, and reach the third, deep learning. Thursday we reach the smallest circle, the chatbot. The same idea, learning from data, is under all of them.

Every model has the same setup

$$Y_i = f(X_i) + \epsilon_i$$

The diagram shows the equation $Y_i = f(X_i) + \epsilon_i$. Below each term is a label with an arrow pointing to it: 'Target' points to Y_i , 'Model' points to f , 'Predictors' points to X_i , and 'Error' points to ϵ_i .

Target	Model	Predictors	Error
the value we want to predict	a function: turns the inputs into a prediction	the inputs we know, for person i	what the model misses: target minus prediction

LOSS FUNCTION

One number that adds up the errors across everyone: the sum of squared errors, or the average miss. Big errors, big loss. **Fitting** means choosing the model that makes the loss as small as it can get.

Every model in this talk has the same setup, and it's worth one slide. Person number i has a target, which is the thing we want to predict, and a set of predictors, which are the things we know about them. The model is a function: it takes the predictors and returns a prediction. The prediction is almost never exactly the target. What's left over is the error. So the target equals the model's prediction plus the error. The target can be a number, like weight, or the chance that a patient falls in the hospital. It can also be a category: benign or malignant, pneumonia or not.

The error isn't only that each person is a little different. It also holds any error in measuring the target, and in measuring the predictors. My PhD is in human genetics, and one of our constant problems was defining the disease. If you can't define it well, it's very hard to find the gene that goes with it. The best models start with the best measurements.

Those four words, target, model, predictors, error, will come back all session. The eGFR is a model. A chatbot is a model. The f can be as simple as a straight line or as complex as a neural network. [Click]

How do we choose the model? We need a way to score it. A loss function adds up the errors across everyone into one number. One common choice squares each error and adds them up, which is called the residual sum of squares. Another is the plain average miss. Big errors make a big loss. Fitting a model means choosing the version of the function that makes the loss as small as it can be. On the next slide, we do that by hand.

Predict weight from height

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Made-up data for 40 adults.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

$$\text{weight} = \mathbf{a} + \mathbf{b} \times \text{height} + \text{error}$$

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

What is missing? What else would help predict weight? In other words, what additional predictors

a: the intercept. **b**: the slope, kg per extra centimeter.

should we consider?

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

What you just did is the **fitting loop**: guess the dials, score the misses with the loss, adjust, repeat.

Speaker notes

Here is that setup with the simplest model there is. The target is weight. The one predictor is height. Each dot is one made-up adult. The model is a straight line with two numbers in it: a, the intercept, and b, the slope. Each thin line is that person’s error, how far their weight is from the line. The loss is the sum of the squared errors, and the average miss is beside it.

PAUSE: 2 min. Move the two sliders and make the loss as small as you can. Then press “Show the best fit”.

I could move these sliders all day long, but it turns out we have good ways of fitting a model. Fitting is what the computer does when you press that button: it finds the line with the smallest sum of squared errors. For a straight line, that’s the best you can do. This is called linear regression, and it’s machine learning in its simplest form. [Click]

What you just did by hand has a name: the fitting loop. Guess the dials, score the misses with the loss, adjust the dials, and repeat. When to stop is a question we come back to. For a straight line there’s a formula that jumps to where the loop would end, which is what the button uses. Bigger models, like neural networks, really do run the loop, lap after lap. [Click]

Even the best line leaves errors. Some tall people are light and some short people are heavy. So what’s missing? What else would help us predict weight? Age, sex, how much of the weight is muscle, waist size. Each one is another predictor, another column in X. The eGFR does this with creatinine, age, and sex.

LIVE: About two and a half minutes. Move the sliders yourself and ask the room which way to move b. Press the button. Click once to name the fitting loop, then again for the last question, and take two or three answers aloud.

PART 1

The number nobody measured

Speaker notes

Now to a clinic, and a model you already use.

Pre-op clinic, surgery sub-internship

A 62-year-old woman is booked for a laparoscopic cholecystectomy. Her basic metabolic panel is back.

The attending asks: **are her kidneys all right for the contrast CT?**

You say: *“eGFR 57, stage 3a, mildly reduced. Probably fine with hydration.”*

BASIC METABOLIC PANEL · COLLECTED 07:12			
Sodium	139		136–145
Potassium	4.1		3.5–5.1
Creatinine	1.10		0.50–1.10
eGFR (CKD-EPI 2021)	57	L	>60

Units: mmol/L; creatinine mg/dL; eGFR mL/min/1.73 m². Illustrative panel, made up for this course. Not a real patient.

Speaker notes

Pre-operative clinic, on your surgery sub-internship. A sixty-two-year-old woman is being worked up for a laparoscopic cholecystectomy. Her basic metabolic panel is back, and one line is flagged low: an eGFR of fifty-seven, where the lab’s reference is above sixty. [Click]

The attending asks whether her kidneys are all right for the contrast CT the surgeon wants. You answer the way everyone answers. eGFR fifty-seven, stage three a, mildly reduced, probably fine with hydration. The attending nods, and that’s the whole conversation. You’ll have it several hundred times in intern year.

The panel on the slide is made up for this course. It isn’t a real patient.

LIVE: Before the case, tell the room to get a phone nearby; the next slide is a live question.

Where does the 57 come from?

Her report says **eGFR 57**. Where does the 57 come from?

- The analyzer measured it in her blood sample
- It was measured from a timed urine collection
- An equation calculated it from her creatinine, age, and sex

Speaker notes

Before I go on, commit to an answer. Where does the fifty-seven come from? Did a machine measure it in her blood? Was it measured from a timed urine collection? Or did an equation calculate it from three other numbers?

LIVE: Open the poll. About 45 seconds, then press A to reveal. Add aloud: no urine was collected, and no blood test measures GFR itself.

PAUSE: 1 min. Pause here and pick one before you go on.

The answer is the equation. Nobody measured her kidney function. Three numbers went into an equation, and the equation produced fifty-seven. The point isn't knowing the answer. It's noticing that the number was calculated.

Nobody measured her GFR

- Measuring GFR means **infusing a marker** the kidney filters and does nothing else with, such as iohexol, and timing how fast it clears.
- That takes hours. It is **rarely done in routine care**.
- The 57 was **computed** from creatinine, age, and sex.
- The equation was **learned from data**, not from physiology. A few thousand people did have their GFR measured. A computer was given their creatinine, age, and sex, and one instruction: *find the formula that best predicts the measured number*.

The eGFR is a **model**. Remember her: 45 kg, and a year of muscle loss.

Speaker notes

Nobody measured her glomerular filtration rate. [Click] Measuring it means infusing a marker that the kidney filters and does nothing else with, such as iohexol, and timing how fast it clears. [Click] That takes hours, and it’s rarely done in routine care.

[Click] The fifty-seven was computed from three things the lab did have: her creatinine, her age, and her sex. Something turned those three numbers into a kidney function, and you trusted it enough to clear her for contrast.

[Click] So where did that equation come from? Not from the physiology of nephrons. It was learned from data. A few thousand people did have their GFR measured. Their creatinine, age, and sex were given to a computer with one instruction: find the formula that best predicts the measured number. [Click]

The eGFR is a model. It’s the model you already trust most, and you have probably never asked what it was trained on.

Remember this patient. She weighs forty-five kilograms and has lost a good deal of muscle over a year of illness. We’ll come back to her at the end. Nobody has told her that a model had a part in the fifty-seven.

Fit the curve yourself

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Illustrative data, not the CKD-EPI data. A simpler curve than the real eGFR.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

$$\text{GFR} = \mathbf{a} \times \text{creatinine}^{-\mathbf{b}}$$

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

a: the GFR when creatinine is 1.0. **b**: how fast GFR falls as creatinine rises.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Fitting means a computer turns these dials until the error is as small as it can get.

● one made-up person · thin line: how far the curve misses them · thick line: the curve your dials make. Generated in your browser.

Speaker notes

Here is fitting, done by hand. Each dot is one made-up person: creatinine across, measured GFR up. The thick curve is a simple equation with two numbers in it, a and b. Those two numbers are the dials. The number a sets how high the curve sits: it's the GFR the curve gives at a creatinine of one. The number b sets how fast GFR falls as creatinine rises. Each thin line is a miss: how far one person's measured GFR is from the curve. The number beside the sliders is the average miss.

PAUSE: 1 min. Move the two sliders and make the average miss as small as you can. Then press “Show the best fit”. [Click]

Fitting means a computer turns these dials until the error is as small as it can get. This is the same fitting loop you ran with height and weight. Here the computer tries every setting the sliders allow and keeps the best one. Real fitting uses faster methods, with the same goal. The real eGFR equation has a few more numbers in it, and it was fit to eight thousand two hundred fifty-four people whose GFR was measured, not to forty made-up dots. A chatbot is the same idea with billions of numbers.

The data on this slide are made up, and the curve is simpler than the real equation. Its numbers aren't the real eGFR's.

LIVE: About one minute; the room fit a line on slide 5. Move the sliders briefly, press the button, then click for the last line.

Add a predictor

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Illustrative data, made up. Women are built to have a lower GFR at the same creatinine.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Dots: teal men, terracotta women. Thin lines: the miss for each person. The fit makes the RSS as small as it can, on a log scale.

Fitting with a second predictor is the same job: find the numbers that make the misses smallest.

Speaker notes

Now add a second predictor. The real eGFR equation uses sex as well as creatinine and age, so we do the same. Same made-up people, same idea, but now the equation can have a different level for women and men.

PAUSE: 1 min. Turn on “Add sex” and watch the average miss and the RSS.

The fit now gives each sex its own curve, and both numbers fall. In this made-up data, women have a lower GFR at the same creatinine, so a model that knows the sex can follow both groups. Statisticians call the sum of squared misses the RSS. The fit makes it as small as it can, so the RSS can only go down when a predictor is added. Whether the drop is real, or only a better fit to these forty people, is a different question, and we come back to it with the slider in Part 2. [Click]

Fitting with a second predictor is the same job as before: find the numbers that make the misses smallest.

The data are made up. The real equation has more terms, and it was fit to eight thousand two hundred fifty-four people whose GFR was measured.

LIVE: About one minute. Ask the room which way the average miss will move, then turn it on.

PART 1, CONTINUED

The loop every model comes from

Speaker notes

Next: six words, the fitting loop, and two questions. Together they sort every model you'll meet in residency.

Six words for the rest of the course

Word	Meaning	In the eGFR
Predictors	the inputs	creatinine, age, sex
Response	what is being predicted	measured GFR
Training set	the cases used to fit the equation	8,254 adults in 10 studies, mostly research participants; 31.5% Black
Validation set	cases held back <i>during</i> fitting, to choose between versions	<i>see the note below</i>
Test set	cases held back to check it once, at the end	4,050 people in 12 other studies; the paper calls them its “validation data set”
Model	the equation once fit	$\text{eGFR} = f(\text{creatinine}, \text{age}, \text{sex})$

f stands for **function**: a rule that turns inputs into an output. **Fit** means a computer tried numbers in the equation until its predictions were as close to the measured values as they could get (Inker et al., 2021).

Words vary by community. Statistics says *response* or *outcome*; machine learning says *target* or *label* (slide 4 used *target*). Medical papers often say *validation set* for what machine learning calls the test set. Ask what the held-back cases were used for, not what they are called.

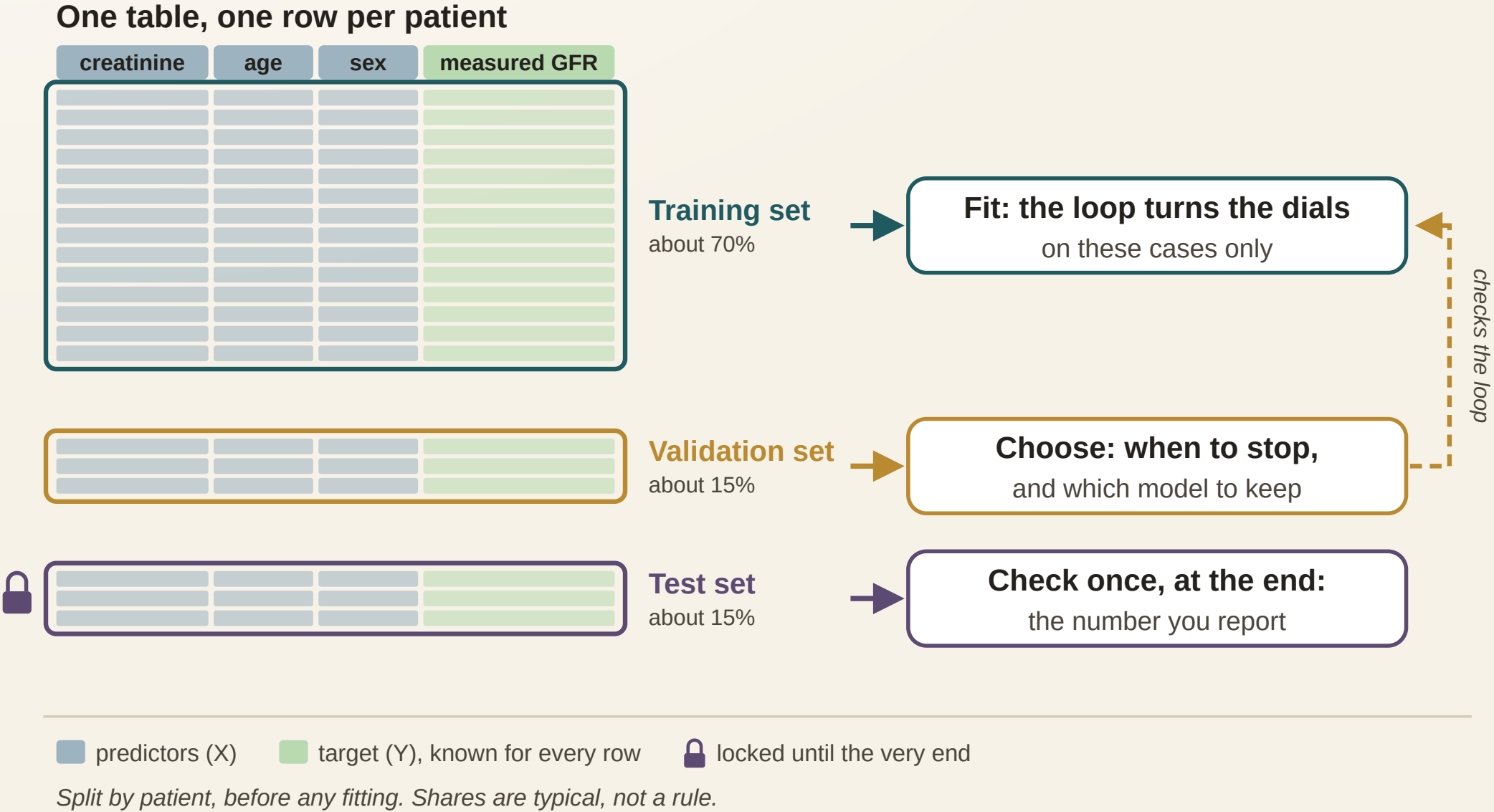
The eGFR story contains almost every word you need for the rest of the course.

Predictors are the inputs: creatinine, age, and sex. The response is the thing being predicted: the measured GFR. The training set is the eight thousand two hundred fifty-four adults whose numbers were used to fit the equation. They came from ten studies and were mostly research participants. About a third were Black. The validation set is a second group held back while the model is being fit. It’s used to choose between versions: when to stop, or which of two models to keep. The test set is checked once, at the end. For the eGFR, that was four thousand fifty people from twelve other studies that the equation had never seen.

Watch the names, because they vary by community. The response is what statisticians call it; machine learning says target, which is the word we used on the setup slide, or label. Same idea. And the eGFR paper calls those four thousand fifty people its validation data set, and many medical papers use the word that way. In machine learning, a group held back and checked once at the end is the test set. When you read a paper, ask what the held-back cases were used for, not what they are called. And the model is the equation itself, once it has been fit.

The letter f stands for function, which just means a rule that turns inputs into an output. Fit means that a computer tried numbers in the equation until its predictions were as close to the measured values as they could get. For the eGFR, the model is a handful of numbers. For a chatbot, it’s billions. The role is the same.

Split the data before you fit



The eGFR’s test set wasn’t a random slice of the same table: it was 4,050 people from 12 other studies. That’s a harder test, and a better one.

Speaker notes

Here are those three sets as a picture. Start with one table, one row per patient. For the eGFR that’s creatinine, age, and sex, the predictors, and the measured GFR, the target. Every row has the answer.

Before any fitting, you split the rows. Most go into the training set, and that’s what the fitting loop turns the dials on. A smaller slice is the validation set. The loop gets to look at it, to decide when to stop and which model to keep, so think of it as a pseudo test set. And the last slice is the test set. It’s locked. Nobody looks at it while the model is being built, and you check it once, at the very end. That’s the number you report.

The shares on the slide are typical, not a rule. Two details matter more than the percentages. Split by patient, so the same person never lands in two sets. And split before you fit, because once the test cases have influenced a choice, they’re no longer a test.

The eGFR researchers did one better. Their test set wasn’t a random slice of the same table. It was four thousand fifty people from twelve other studies. A test on different people, from different studies, is harder to pass, and it tells you more.

One equation, three generations

Equation	Learned from	Where it failed, and for whom
MDRD, 1999	1,070 patients who already had kidney disease (558 more held back to test it)	Values too low at higher GFR, so healthy people were labelled as diseased (Levey et al., 1999; Levey et al., 2009)
CKD-EPI, 2009	8,254 people in 10 studies	Included a race term. Even with it, the equation overestimated measured GFR in Black adults (median 3.7 mL/min/1.73 m ²). An overestimate makes kidney disease look milder than it is (Inker et al., 2021)
CKD-EPI, 2021	the same 8,254, refit without race	Now a few mL/min too low in Black adults and too high in others. A version that adds cystatin C, a second blood marker that does not depend on muscle, is more accurate (Delgado et al., 2021; Inker et al., 2021)

The training population is part of the model,
as much as the equation is.

Speaker notes

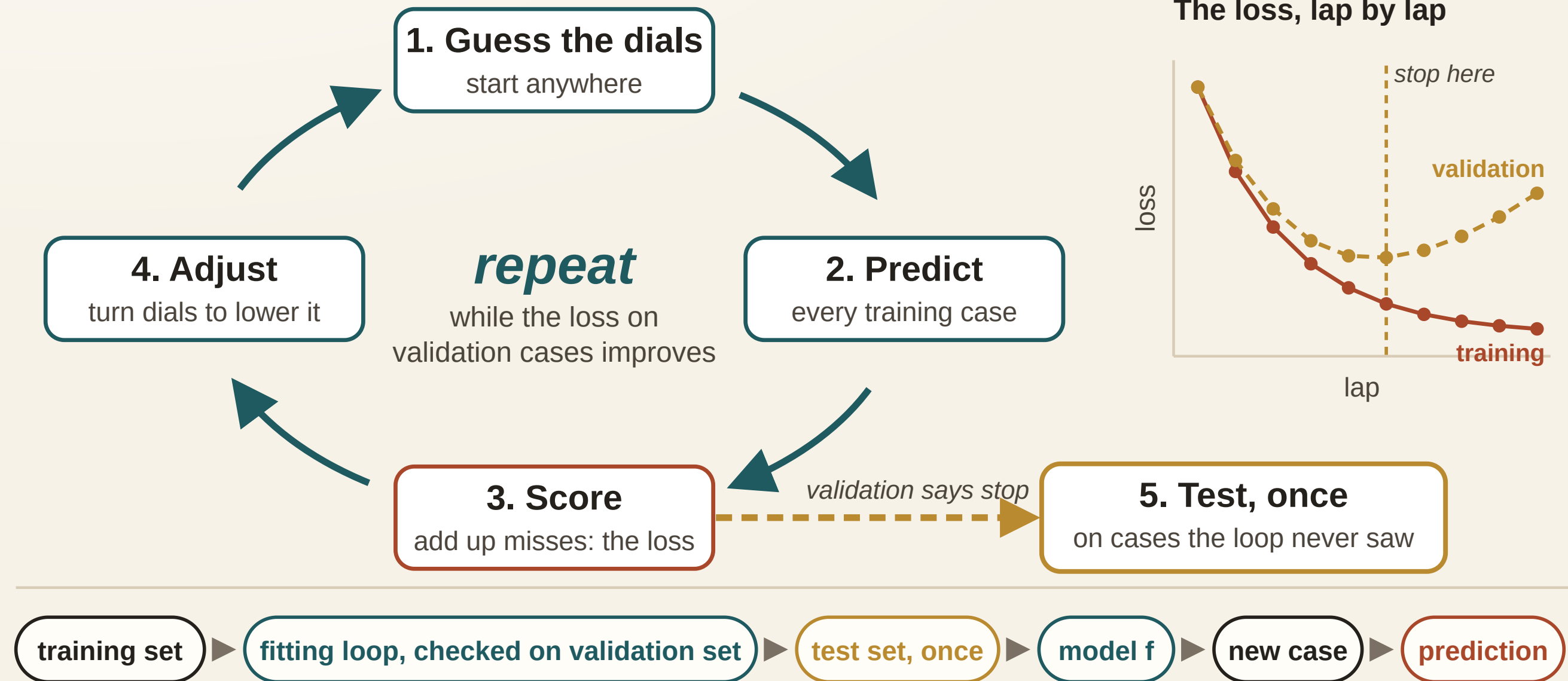
Here is the history of one equation, rebuilt from the numbers in the papers. Read the last column as a pattern, not as nephrology.

MDRD was fit to one thousand and seventy people who already had kidney disease, and tested on five hundred fifty-eight more. In healthy people it gave values that were too low, and healthy people were labelled as diseased. The 2009 equation was more accurate at higher GFR, and it included a term that adjusted the result for Black race. Even with that term, the 2021 study found it overestimated measured GFR in Black adults. An overestimate makes kidney disease look milder than it is. The 2021 refit removed race, and it's still a few points off in both directions. A version that adds cystatin C, a second blood marker that doesn't depend on muscle, is more accurate. [Click]

Every row says the same thing. The model was fit to these people, and here is what happened to people unlike them. The training population is part of the model, as much as the equation is. That's a take-home message, and we'll come back to it. Why the race term was there, and who was harmed by it, is the subject of the bias and equity session later in the course.

LIVE: If asked why the race term was added: in the 2009 data, Black participants on average had a higher measured GFR at the same creatinine, and the term adjusted for that. The 2021 paper states that race in these equations is a social and not a biologic construct. In 2021 a national task force recommended the refit without race for all US laboratories, and wider use of cystatin C to confirm the estimate. What the term did to Black patients is the bias and equity chapter's subject.

The fitting loop



You ran this loop by hand with the height and weight sliders. Every model in this course is fit this way: only the data and the number of dials change.

Speaker notes

Here is the fitting loop as a picture. You ran it by hand with the height and weight sliders. Step one, guess the dials; anywhere will do. Step two, use them to predict every training case. Step three, score the predictions: add up the misses into one number, the loss. Step four, adjust the dials in the direction that lowers the loss. Then go around again.

When do you stop? Not when the training loss stops shrinking. Look at the chart. The red line is the loss on the training cases, and it keeps falling: with enough laps, a model can fit the cases it learns from almost perfectly. The gold line is the loss on a second set of cases the loop never fits to, the validation set. It falls, bottoms out, and then rises. From there on, the model is memorizing the training cases, noise included, instead of learning the pattern. That's overfitting. So the loop stops where the validation loss is lowest. The validation set is also how you choose between models, for example a simple one and a flexible one. You'll see that happen with the slider in a few minutes. Think of it as a pseudo test set: the loop gets to look at it, so it can't be the final word.

Then you leave the loop, and there's one more step. Step five: test the model once, on a third set of cases, the test set, that neither the loop nor the choices ever saw. The loss inside the loop only tells you how well the model fits the cases it was built from. The test set tells you how it does on new ones. We'll see what happens when that step is skipped or done badly.

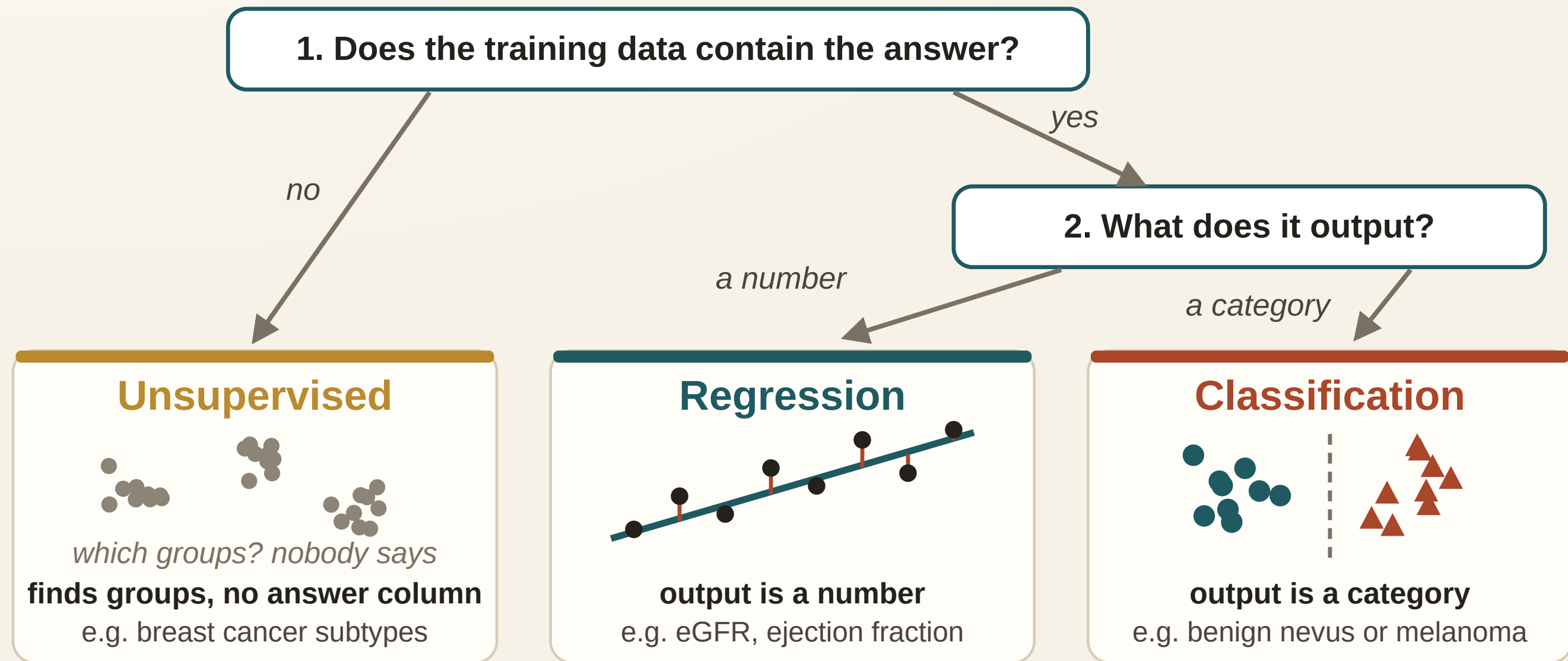
For a straight line, and for the eGFR equation, there's a formula that lands where the loop would end, so nobody needs to go around. For a neural network there's no such shortcut, and the computer really does run this loop, many times.

The strip at the bottom is the whole path. Fitting needs training cases with both the predictors and the truth. The loop is checked on the validation set as it runs. The model it gives, f , is tested once on the test set. Then a new patient arrives with only the predictors, and f turns them into a prediction.

The eGFR, a sepsis alert in the EHR, and a chatbot differ in the data and in how many dials they have. The loop is the same.

So there are two take-home messages here. There's a loop. And there's such a thing as overfitting: a model can do very well on the cases it was fit to, and poorly on new ones.

Two questions sort any model



A regression is wrong **by an amount**. A classification is wrong **by a kind**: a miss, or a false alarm. A miss costs the patient who was missed. A false alarm costs the staff who answer it, and the patient worked up for nothing.

Two questions sort any model in about ten seconds. Start at the top.

First: does the training data contain the answer? For the eGFR, yes. Someone measured every one of those GFRs. That’s supervised learning. If there’s no answer column, only patterns to find, it’s unsupervised learning. The example is breast cancer subtypes, which were found by grouping tumors that share gene activity, with nobody telling the computer the groups (Perou et al., 2000). We won’t say more about unsupervised learning today, but it’s worth knowing it exists.

Second, and only for a supervised model: what does it output? A number makes it a regression. The eGFR is a regression, and so is a model that reports an ejection fraction. In the picture, each red line is how far a prediction misses. A category makes it a classification: benign or melanoma, sepsis or no sepsis. Here the model draws a boundary, and a case on the wrong side is an error. An unsupervised tool that sorts cases into groups isn’t a classification, because nobody told it what the groups were. [Click]

The difference matters for how you judge a model. A regression is wrong by an amount. A classification is wrong by a kind, a miss or a false alarm. A miss costs the patient who was missed. A false alarm costs the staff who answer it, and the patient who is worked up for nothing.

Sometimes it isn’t obvious which kind a model is. Sorting that out is a good first step to understanding how it was built and how it’ll be used.

A model watches a heart ultrasound

A model watches a heart ultrasound video and reports **an ejection fraction of 55%**. What kind of model is it?

- Regression: it gives a number
- Classification: it gives a category
- Unsupervised: it finds groups, with no answer column

Speaker notes

Now you try, with the two questions from the last slide. First: does the training data contain the answer? Second: is the output a category or a number?

A model watches a heart ultrasound video and reports an ejection fraction of fifty-five percent. What kind of model is it?

LIVE: Open the poll. Give the room about 45 seconds, then press A to reveal. Regression: the training data contain the answer, a measured ejection fraction, and the output is a number.

PAUSE: 1 min. Pause here and pick an answer before you go on.

This is the EchoNet-Dynamic model (Ouyang et al., 2020). The same model also says whether the heart failure has reduced ejection fraction, which is a yes or no. One system can give a number and a category. The type of problem depends on the output you ask for.

A model looks at a skin photo

A model looks at a photo of a skin lesion and answers: **benign nevus or melanoma**. What kind of model is it?

- Regression: it gives a number
- Classification: it gives a category
- Unsupervised: it finds groups, with no answer column

Speaker notes

Same two questions. A model looks at a photo of a skin lesion and answers: benign nevus or melanoma. What kind of model is it?

LIVE: Open the poll. About 30 seconds, then press A. Classification: every training image carried a diagnosis, and the output is one of two categories.

PAUSE: 1 min. Pause here and pick an answer before you go on.

This is the dermatology model from Stanford (Esteva et al., 2017). A classification is wrong by a kind: it can call a melanoma benign, or a nevus malignant. Which of those costs more is a clinical question, and it's yours to answer.

Four fields, the same two questions

Field	The model reads	It gives	Type	What the authors reported
Cardiology	a heart ultrasound video	ejection fraction	Regression	average error 4.1 points at the same hospital, 6.0 at another health system (Ouyang et al., 2020)
Pediatrics	a hand X-ray	bone age in months	Regression	the best of 105 entries was off by 4.2 months on average, on 200 test X-rays (Halabi et al., 2018)
Dermatology	a skin photo	benign or malignant	Classification	on par with 21 dermatologists, on biopsy-proven images (Esteva et al., 2017)
Pathology	a lymph node slide	cancer present or absent	Classification	best entry AUC 0.994; the entries ranged from 0.556 to 0.994 (Ehteshami Bejnordi et al., 2017)

More of the same: an **ECG** rhythm classifier that matched cardiologists (Hannun et al., 2019); **breast tumors grouped by gene expression**, with no diagnoses attached, which is unsupervised (Perou et al., 2000); and the diabetic retinopathy system from Tuesday (Abràmoff et al., 2018).

Speaker notes

Four fields, the same two questions. Read the Type column, then the last column.

In cardiology, a model reads ultrasound video and gives a number, the ejection fraction. In pediatrics, a model reads a hand X-ray and gives a bone age in months. Both are regressions: they are wrong by an amount. In dermatology and pathology, a model reads an image and gives a category. Both are classifications: they are wrong by a kind.

Now look at the last column, because the same lesson is in it. The cardiology model was off by four point one points at the hospital where it was built, and by six point zero at another health system. In the pathology challenge, the best entry scored zero point nine nine four on the area under the curve, and the worst scored zero point five five six, barely better than a coin flip. Results on the authors’ own test data are the optimistic number. Ask where it was tested, and by whom.

For the record, three more examples are listed under the table: an ECG rhythm classifier, tumor subtypes found without any diagnoses, and the diabetic retinopathy system from Tuesday. They are here so you can find the papers.

LIVE: About one minute and a half. Don’t read the table. Point at the Type column and then at the last column.

Too simple, too flexible

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

A model **overfits** when a simpler model would do worse on the training cases but better on new cases from the same population (Mitchell, 1997).

● learns from · ○ held back for testing · dashed: the real pattern. Made-up data. **Bias**: error of a too-simple model. **Variance**: error of a too-flexible one.

Speaker notes

Move the slider and watch. I simulated these data, so I know exactly where they came from: the dashed line is the real pattern, which the model never gets to see. The filled dots are the cases the model learns from. The hollow dots were held back for testing, and while we fit, we pretend they don't exist. Under the charts, the two numbers are the error on each set.

At the far left the model is too simple. It's a flat or straight line through a curve, and it's wrong everywhere, on both sets. Statisticians call that kind of error bias, in a technical sense that has nothing to do with the social meaning.

Now let the model bend. As you slide right, the training error goes down, and for a while the test error goes down with it. Then, somewhere in the middle, something happens. The training error keeps falling, but the test error starts to climb. Keep going and it gets kind of absurd: the curve bends through every filled dot, the training error drops to zero, and the test error goes off the chart. The model has learned the random scatter in these twelve cases along with the real pattern. That error is called variance.

That's overfitting, and here's the real definition, from a standard machine learning textbook. A model overfits when a simpler model would do worse on the training cases but better on new cases from the same population. Two things in that definition matter. First, it compares two models. Test error is almost always a little higher than training error, and that alone isn't overfitting; the sign is that more flexibility keeps lowering the training error while the error on held-back cases rises. Second, it's about new cases from the same population. A model that fails at a different hospital, on different patients, has a different problem, and we'll see it on the sepsis slide.

If you were only ever shown the training error, you'd pick the worst model. This is the job of the validation set in the fitting loop: choose the setting where the error on held-back cases is lowest, and keep a separate test set for the end. And if you find you've overfit, the fix is to go back and make the model simpler.

PAUSE: 2 min. Pause here and move the slider from one end to the other.

A neural network: the same loop, more dials

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

Illustrative. Every line is a dial. This toy network has

Error

This document uses OJS, which requires JavaScript features disabled when running in file:///... URLs. In order for these features to work, run this document in a web server.

; the networks behind the examples have far more.

- The **inputs** can be the raw pixels of an image. Nobody has to write down what to look for.
- Each unit adds up what it receives, with its own dials, and passes the result on.
- **Fitting** is the same loop: turn every dial until the misses on the training cases get smaller.

Speaker notes

Here is a neural network. It’s the fitting loop again, with a far bigger equation. [Click]

On the left are the inputs. For an image, these can be the raw pixel values, so nobody has to decide in advance what to look for. [Click] In the middle are layers of simple units. Each unit adds up what it receives, with its own dials, and passes the result to the next layer. On the right is the output: a number, such as an ejection fraction, or a category, such as melanoma. You don’t have to tell it what to look for, but you do still need the answer for every training case. [Click]

Every line in this picture is a dial. This toy network has fifty. The networks behind the four examples have far more. GPT-3, an early ancestor of ChatGPT, had a hundred and seventy-five billion (Brown et al., 2020). Turning that many dials takes an enormous amount of computing time and energy, which is a big part of why a new chatbot model costs so much to build. Fitting is the loop you already know: turn every dial until the misses on the training cases get smaller. The word deep in deep learning means many layers.

Everything from the slider still applies. A network with that many dials can memorize its training cases, so the test set matters even more. And what it learned depends on the people in its training data.

LIVE: About one minute. Don’t explain the arithmetic. The picture is for the idea of many dials, and the next section uses it for words.

Simple models are easier to check

SIMPLE: THE EGFR EQUATION

- A **handful of numbers**. You can read each one.
- You can check them against what you know: higher creatinine, lower eGFR.
- A few thousand people were enough to fit it.
- It can only follow the shape its equation allows.

FLEXIBLE: A NEURAL NETWORK

- **Thousands to billions of numbers**. Nobody reads them one by one.
- It can fit complex patterns: an image, a sentence.
- It is harder to explain why it gave one answer.
- It **needs much more data**, or it fits the noise, as in the slider.

More flexible: more patterns it can fit,
less you can explain, more data it needs.

| A model is **interpretable** when a person can read why it gave its answer.

Speaker notes

The slider showed one trade-off. Here is a second one that comes with it.

The eGFR equation is a handful of numbers. You can read each one and check it against what you know. Higher creatinine gives a lower eGFR, older age gives a lower eGFR, and you could work it out by hand. A few thousand people were enough to fit it. The cost is that it can only follow the shape its equation allows.

A neural network has thousands to billions of numbers. It can fit patterns no short equation can, such as what a pneumothorax looks like on a film, or how a sentence goes on. But nobody can read those numbers one by one, so it’s harder to explain why it gave a particular answer. And it needs far more data, or it fits the noise, as the slider showed at the far right. Turn that around: with only a few hundred patients, you can’t fit a very complex model. That’s one reason the internet and the electronic health record mattered so much in yesterday’s history. Machine learning needs a lot of data. [Click]

So the more flexible the model, the more patterns it can fit, the less you can explain, and the more data it needs. A model is called interpretable when a person can read why it gave its answer. Yesterday’s expert systems were very interpretable: they could tell you the rules they followed. Neither kind is better in general. For a lab value, a simple model you can check is often the right choice. Images and language need the flexible kind, and then the test set matters even more.

LIVE: One minute. Read the last line and move on.

Cases it has already seen prove nothing

Fit a formula to 8,254 people, then report how well it predicts **those same people**, and you have measured how well it **memorized** them.

A **separate test set** is the least you should accept.

A number with no test set is not a number yet.

The usual rule is three sets: a **training set** to fit the model, a **validation set** to choose between models, and a **test set**, used once, at the end.

Speaker notes

Here is the mistake the eGFR researchers avoided, and the mistake many vendor slides make.

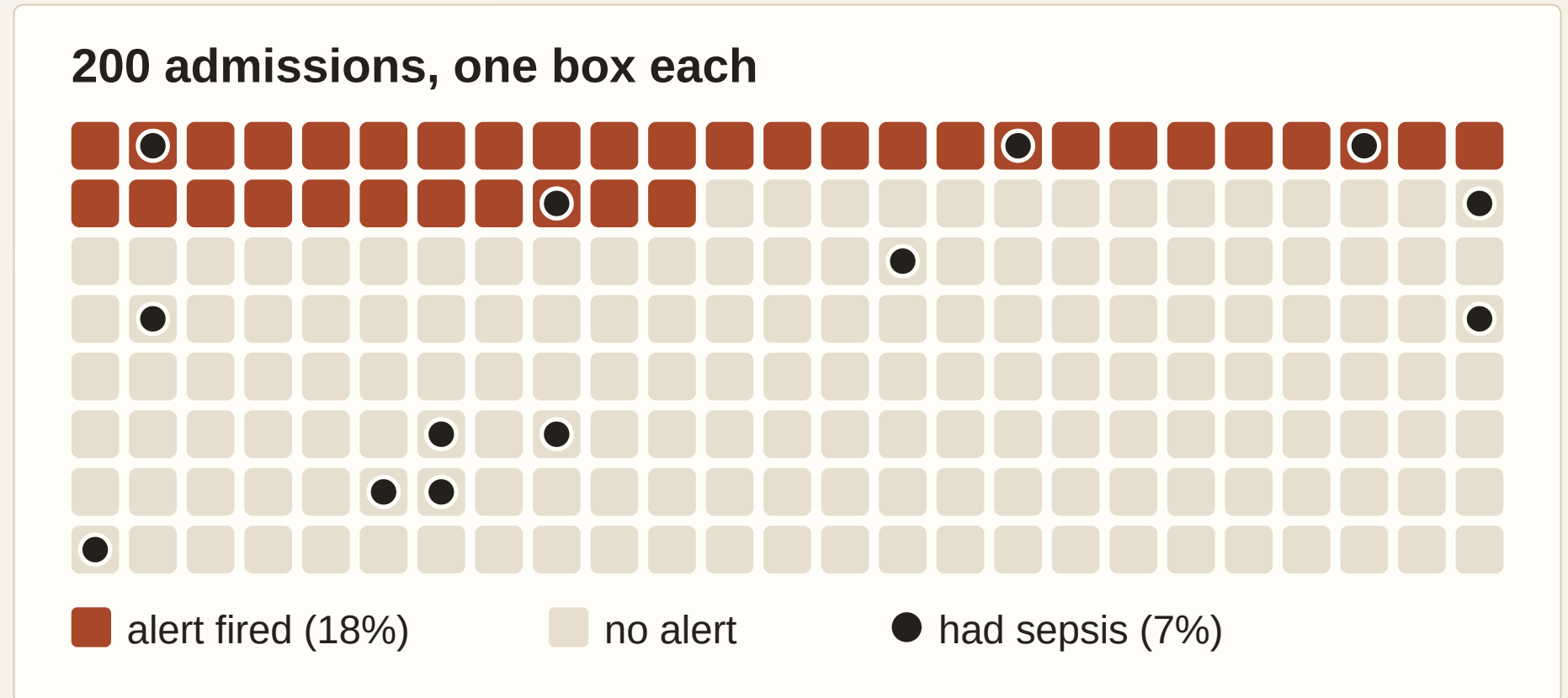
Suppose you fit a formula to eight thousand people, and then you report how well it predicts those same eight thousand people. You have measured how well the formula memorized them, not how well it will work on the next patient. A flexible enough model can fit its training data perfectly and still be useless. [Click]

That's why a separate test set is the least you should accept. Fit on one set of cases, and report performance on a different set the model never saw. If someone tells you their model's error is near zero, your first question is: on the training data or the test data? Did you try it on new data, and what was the error there? The same goes for a model trained on synthetic patients. Synthetic training data can help, but the test still has to be on real ones. When someone gives you one number and doesn't say which set it came from, you do not have a number yet. [Click]

The usual rule uses three sets, the same three as in the fitting loop. The training set is used to fit the model. The validation set is used to choose between models, for example between a simpler one and a more flexible one. The test set is kept aside and used once, at the end. If the test set is used to choose, it becomes a second validation set, and its number is optimistic too. The next slide shows why this matters.

You saw this sepsis model yesterday

Here is why it belongs in today's talk: it is the same failure, **on new patients**. A vendor sepsis model, in use at hundreds of US hospitals in 2021, was tested at one academic center on **38,455 hospitalizations** (2018 to 2019), by people who had not built it (Wong et al., 2021).



Redrawn from the counts reported by Wong et al. (2021), scaled to 200 admissions.

Testing on patients from a different setting than the model was built on is **external validation**. Of the 13 with sepsis, 9 sat in a no-alert box. **No alert does not mean no sepsis.**

Speaker notes

You saw this model yesterday. Here is why it belongs in today’s talk: it’s the same failure as the slider, in the clinic. One EHR vendor’s sepsis model, the Epic Sepsis Model, was in use at hundreds of US hospitals in 2021. One academic center tested it on thirty-eight thousand four hundred fifty-five of its own hospitalizations. A test on patients from a different setting than the ones the model was built on is called external validation.

The area under the curve was zero point six three, where zero point five is no better than chance. Of every hundred patients who had sepsis, the model missed sixty-seven. And since it alerted on eighteen of every hundred admissions, when only about seven had sepsis, most alerts were for patients who didn’t.

The grid shows the same counts, scaled to two hundred admissions. Each box is one admission. The colored boxes are the thirty-six where the alert fired. The dots are the thirteen patients who had sepsis. Only four of the dots sit in a colored box. The other nine had sepsis and no alert.

The study doesn’t show why. A common reason is that the patients a model meets in use differ from the patients it was built on. The name for that is dataset shift.

For the wards: no alert does not mean no sepsis. And if you silence an alert, you should be able to say why.

LIVE: About three minutes. They saw this study yesterday. Walk the grid (colored boxes, then dots), then name the link to the test set.

The published number is the optimistic one

A model's performance on the cases it was built from is **usually the optimistic number**.

Ask what it was **trained on**.

Ask what it was **tested on**, and by whom.

If the training set and the test set are the same cases, the score tells you nothing yet.

Better still, a hospital tests the model on its own past patients before go-live (**localization**), then keeps checking that it still works (**monitoring**).

That gives you another of the course’s recurring lessons: the published number is the optimistic one.

A model’s performance on the cases it was built from is usually the best it will look. On new patients, especially from another hospital, it’s usually lower. [Click] So ask three things of any model. What was it trained on? What was it tested on, and by whom? If the training set and the test set are the same cases, the score tells you nothing yet, because the model has already seen the answers.

Why ask what it was trained on? Because the training data are part of the model. A model built on patients in one country, or at one kind of hospital, may not work the same way somewhere else.

Often you won’t get the details: companies frequently don’t publish what a commercial model was trained or tested on. Ask anyway. But there’s a second answer that doesn’t depend on them, which is to test it on your own patients. I sit on the committee that vets machine learning models for our health system, and one of our first questions about any model is what it will cost to localize. Localizing means running the model on our own past patients, say the last six months or the last few years, and scoring it against what we know happened. Take a fall-risk model. Of the patients who fell, how many did the alert fire for? And how often did it fire for patients who didn’t fall? Sometimes the next step is to tune the model to fit our patients better.

And it doesn’t stop at go-live. Patients change over time, and so do the records; even the codes used for diagnoses have changed their definitions over the years. So someone has to keep monitoring the model to make sure it still works.

A language model is the same loop, with a different table

the patient was started on empiric vancomycin

Predictors: the words so far	Truth: the next word
the	patient
the patient	was
the patient was	started
the patient was started	on
the patient was started on	empiric
the patient was started on empiric	vancomycin

Nobody labels anything. **Every sentence ever written is already a table like this.**

Speaker notes

Take any sentence from the internet. The patient was started on empiric vancomycin.

Now turn it into a table, the same kind of table as the eGFR. The predictors are the words so far. The truth is the very next word. “The” is followed by “patient”. “The patient” is followed by “was”. And so on, until “the patient was started on empiric” is followed by “vancomycin”. You might ask: lots of sentences start with “The”, and many different words can come next. That’s why the model uses all the words so far, not just the last one. The more of the sentence it sees, the better it can guess. [Click]

Nobody had to label anything, because the next word is already in the text. Every sentence ever written becomes training cases with no labelling work. That’s why these models could be trained on a very large fraction of the written internet.

Thursday starts here: a chatbot is machine learning on words, and the same test-set questions apply.

LIVE: About one minute. Stop here. This is the bridge to Thursday.

PART 3

What this changes for you

Speaker notes

Last few minutes. Back to the clinic.

Back to the pre-op clinic

- She weighs **45 kg** and has lost muscle over a year of illness.
 - Creatinine comes from **muscle**.
 - Her creatinine of 1.10 is a normal number for a normal amount of muscle. She does not have a normal amount of muscle.
 - So her true GFR is probably **lower than 57**.
- Name her model's predictors, response, and training population.
- What do you now say to the attending?**
- "The 57 is estimated from her creatinine. With her low muscle mass it probably overstates her kidney function. If the decision depends on it, I would ask for a cystatin C-based eGFR."*
- It matters most for a renally dosed drug, or if her true value could be near your radiology department's contrast threshold. A national task force recommends wider use of cystatin C, especially to confirm the estimate (Delgado et al., 2021).

Nobody told her a model was involved. Whether someone should have is a question for the patient communication session.

Speaker notes

Go back to the clinic. The sixty-two-year-old woman is still on the schedule.

[Click] She weighs forty-five kilograms and has lost muscle over a year of illness. [Click] Creatinine comes from muscle. [Click] Her creatinine of one point one is a normal number for a normal amount of muscle, and she doesn't have a normal amount of muscle. [Click] So the same creatinine probably means a lower true GFR than the fifty-seven the equation returned. Her low muscle mass changes the predictor without changing the truth, and nothing in the equation can tell. [Click]

Name her model's predictors, response, and training population. And what do you now say to the attending about the number on the panel?

PAUSE: 2 min. Pause here and answer out loud or in writing.

[Click] Here is one good answer. "The fifty-seven is estimated from her creatinine. With her low muscle mass it probably overstates her kidney function. If the decision depends on it, I would ask for a cystatin C-based eGFR." Whether it changes the contrast decision depends on how close her true value could be to your department's threshold. It matters more for a drug dosed by kidney function. The training population was eight thousand two hundred fifty-four adult research participants. Nothing in the equation knows how much muscle she has.

Notice also that nobody told her a model had a part in her care. Whether someone should have is a question for the patient communication session.

LIVE: Ask for answers in the chat, sent on your count. Cystatin C is the answer you're looking for, with the caveat that cystatin C has limits of its own and illness can affect it. A measured clearance only if the decision really depends on it.

On the wards

- **An alert fires, or stays quiet.** No alert does not mean no sepsis. If you silence one, be able to say why.
- **A tool comes with an accuracy number.** Ask what it was tested on, and by whom. A score on the cases it was built from is the optimistic one.
- **A tool drafted your note or message.** Read all of it before you sign. You are responsible for what it says.
- **Patient information.** Not in any chatbot, unless your hospital's privacy office has approved that tool in writing.
- **Documentation.** What to record when a tool helped depends on your site's policy. Ask your supervisor.

Ask what it was tested on,
and by whom.

Speaker notes

Here is what this means on the wards tomorrow.

When an alert fires, or stays quiet, remember the sepsis model. No alert does not mean no sepsis. If you silence one, you should be able to say why.

When a tool comes with an accuracy number, ask what it was tested on, and by whom. A score on the cases the model was built from is the optimistic number.

When a tool has drafted your note or message, read all of it before you sign. You’re responsible for what it says. What to document when a tool helped depends on your site’s policy, and the ethics and regulation session covers it.

And patient information goes into no chatbot, unless your hospital’s privacy office has approved that tool in writing. [Click]

If you tell a co-intern one sentence, make it this one. Ask what it was tested on, and by whom.

The standing question

What does this change about how you'll practice?

Write it down before you close this deck. Then one more: the next time someone shows you a model's accuracy number, **what will you ask in the next two minutes?** For example: *what was it tested on, and by whom?*

Thursday: we start with the chatbot, the same loop with words. Then we open ChatGPT Edu together and use it to set up your own learning goals and journal.

Speaker notes

Every session ends on the same question. What does this change about how you'll practice? [Click] Write your answer down before you close this deck. It is not optional, and one word isn't an answer.

PAUSE: 2 min. Pause here and write it.

Then one more. The next time someone shows you a model's accuracy number, what will you ask in the next two minutes? For example: what was it tested on, and by whom? Some of you'll be in exactly that spot. A company shows up at lunch with a model that predicts which of your clinic patients will end up in the emergency department, and asks, don't you want to catch those patients? There are a lot of questions to ask before you say yes, and I hope today's slides help with some of them. [Click]

Thursday: we start with the chatbot, the same loop with words. Then we open ChatGPT Edu together and use it to set up your own learning goals and journal.

Summary

- **AI** contains **machine learning**, which contains **deep learning**, which contains **generative AI**.
- The eGFR is a **model learned from data**: creatinine, age, and sex in; a stand-in for measured GFR out; 8,254 adults as the population it was fit to. **Fitting** chose its numbers to make the error as small as possible.
- **Supervised** has the answer in the training data; **unsupervised** does not. **Classification** gives a category; **regression** gives a number.
- A model that is too flexible fits the noise: it **overfits**. A **neural network** is the same loop with far more dials, so it needs more data, and a test set matters even more.
- A score on the training data describes the fit. **Ask for the test set, and who ran it.**
- **Thursday**: a language model is the same loop, with words.

Speaker notes

To sum up. AI contains machine learning, which contains deep learning, which contains generative AI.

The eGFR is a model learned from data, with predictors, a response, and a population it was fit to. Fitting chose its numbers to make the error as small as possible. Every learned model has those three parts, and you can ask for them.

Supervised learning has the answer in the training data; unsupervised doesn't. Classification gives a category; regression gives a number.

A model that's too flexible fits the noise. That's overfitting: a simpler model would have done better on new patients. A neural network is the same loop with far more dials, so it needs more data, and a test set matters even more.

A score on the training data describes the fit. It is not evidence. Ask for the test set, and who ran it.

Thursday, the same loop with words.

The fifty-seven on the panel came from somewhere. So does everything else.

LIVE: Skip in a live session.

Further exploration

TRY: OVERFITTING YOU CAN SEE

TensorFlow Playground, 10 minutes, no account. Open [step 1](#) (too simple), [step 2](#) (in between) and [step 3](#) (too flexible). At about 500 epochs, compare **training loss** with **test loss**.

READ: WHEN THE PATIENTS CHANGE

The clinician and dataset shift in AI, NEJM, 2021. Why models fail on new kinds of patients.

Speaker notes

For students who want more, two things to follow by interest.

Try: the TensorFlow Playground exercise. Each link opens the site already set up, with the same noisy spiral and only one dot in five used for training. Press play, pause at about five hundred epochs, and compare training loss with test loss. One run gave: step one, both near 0.5; step two, training near 0.09 and test near 0.27; step three, training near 0.1 but test near 0.55. That's overfitting: training loss makes steps two and three look alike, but on new dots step three is the worst. Don't press regenerate, which changes the data.

Read: a two-page paper on dataset shift, written for clinicians.

LIVE: Skip in a live session.

References

- Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1, 39. <https://doi.org/10.1038/s41746-018-0040-6>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. <https://arxiv.org/abs/2005.14165>
- Delgado, C., Baweja, M., Crews, D. C., Eneanya, N. D., Gadegbeku, C. A., Inker, L. A., Mendu, M. L., Miller, W. G., Moxey-Mims, M. M., Roberts, G. V., St Peter, W. L., Warfield, C., & Powe, N. R. (2021). A Unifying Approach for GFR Estimation: Recommendations of the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. *American Journal of Kidney Diseases : The Official Journal of the National Kidney Foundation*, 79(2), 268–288.e1. <https://doi.org/10.1053/j.ajkd.2021.08.003>
- Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., Ginneken, B. van, Karssemeijer, N., Litjens, G., Laak, J. A. W. M. van der, Hermsen, M. and, Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., Dijk, M. C. van, Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., ... Venâncio, R. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22), 2199–2210. <https://doi.org/10.1001/jama.2017.14585>
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
- Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The Clinician and Dataset Shift in Artificial Intelligence. *The New England Journal of Medicine*, 385(3), 283–286. <https://doi.org/10.1056/nejmc2104626>
- Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M., Pan, I., Pereira, L. A., Sousa, R. T., Abdala, N., Kitamura, F. C., Thodberg, H. H., Chen, L., Shih, G., Andriole, K., Kohli, M. D., Erickson, B. J., & Flanders, A. E. (2018). The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology*, 290(2), 498–503. <https://doi.org/10.1148/radiol.2018180736>
- Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65–69. <https://doi.org/10.1038/s41591-018-0268-3>
- Inker, L. A., Eneanya, N. D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., Crews, D. C., Doria, A., Estrella, M. M., Froissart, M., Grams, M. E., Greene, T., Grubb, A., Gudnason, V., Gutiérrez, O. M., Kalil, R., Karger, A. B., Mauer, M., Navis, G., ... Levey, A. S. and. (2021). New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. *The New England Journal of Medicine*, 385(19), 1737–1749. <https://doi.org/10.1056/nejmoa2102953>
- Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N., & Roth, D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of Internal Medicine*, 130(6), 461–470. <https://doi.org/10.7326/0003-4819-130-6-199903160-00002>
- Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T., & Coresh, J. and. (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, 150(9), 604–612. <https://doi.org/10.7326/0003-4819-150-9-200905050-00006>
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., & Zou, J. Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802), 252–256. <https://doi.org/10.1038/s41586-020-2145-8>
- Perou, C. M., Sørlie, T., Eisen, M. B., Rijn, M. van de, Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797), 747–752. <https://doi.org/10.1038/35021093>
- Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestruie, J., Phillips, M., Konye, J., Penozza, C., Ghous, M., & Singh, K. (2021). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8), 1065–1070. <https://doi.org/10.1001/jamainternmed.2021.2626>

Questions

Speaker notes

Let's take questions.

LIVE: Jump here from slide 31 with the End key; this is the last slide. Thursday starts with the chatbot half of this story, and then everyone opens ChatGPT Edu.

