CPH 100: Foundations for CPH

Building a Clinical Prediction Model

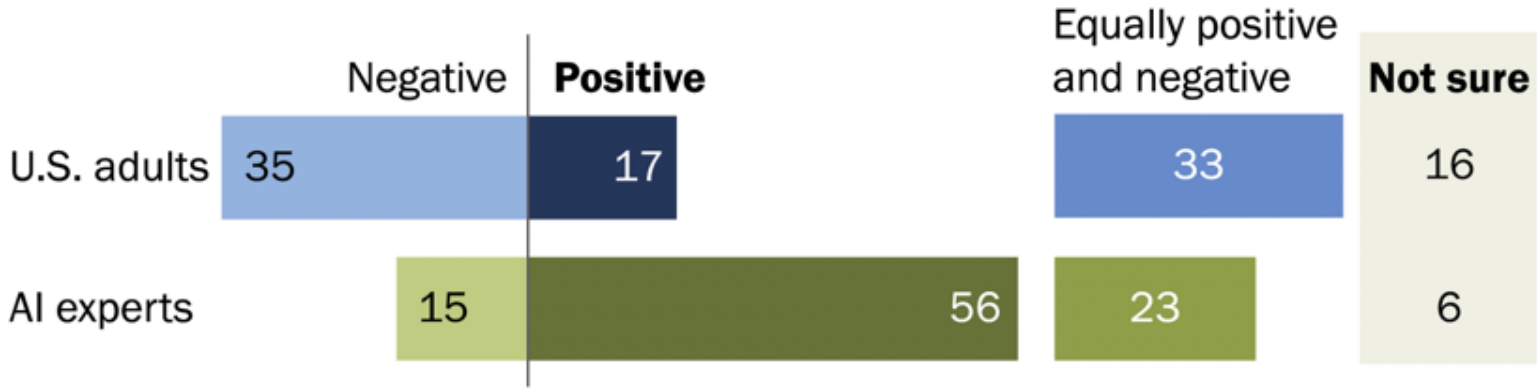
Irene Y. Chen

Here early? Take this quick survey.

https://forms.gle/7R4rorzTgc4FUpcZA

AI experts more likely than the public to say AI will have a positive effect on the U.S. over next 20 years

% who say they think the impact of artificial intelligence (AI) on the U.S. over the next 20 years will be ...



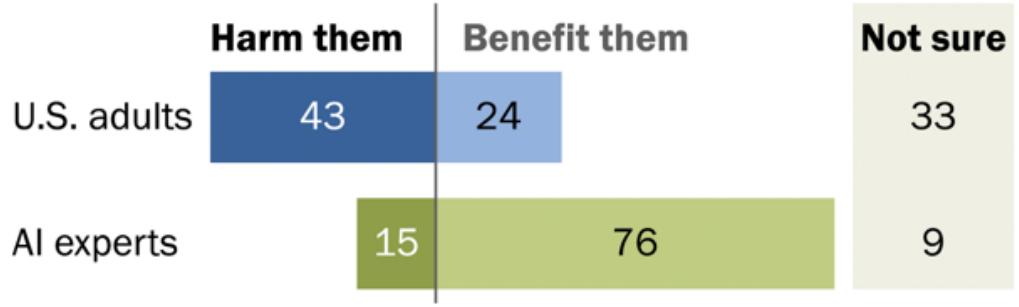
Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. "Very/somewhat positive" and "very/somewhat negative" are combined. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.
"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/

About three-quarters of AI experts think AI will likely benefit them personally, much higher than share of U.S. adults

% who say they think the increased use of artificial intelligence (AI) is more likely to ...



Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. For full question wording, refer to the topline. Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.
"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/

Who should we screen for lung cancer?

The ML pipeline appears simple

1. Define task
2. Collect data
3. Represent features + labels
4. Train model
5. Evaluate on held-out data
6. Use prediction

Healthcare affects every step

1. *Define task* → Prediction or intervention?
2. *Collect data* → Generated for care not for analysis
3. *Represent features* + *labels* → Often proxies or delayed
4. *Train model* → Missingness can encode clinician behavior
5. *Evaluate on held-out data* → Average performance can conceal subgroup performance
6. *Use prediction* → Practice and populations change

Our patient is 68 years old and has 40 pack years.

How do we estimate risk?

Outline

- **Intro** (5 mins)
- **Course logistics** (5 mins)
- **ML pipeline** (40 mins)
- **Break** (10 mins)
- **Real World Applications of Generative AI in Liver Diseases and Transplantation** (50 mins)

Learning Objective: Prepare class for Problem Set 1

Class Schedule

Note: tentative schedule is subject to change.

Lecture Number	Date	Topic and Readings	Lecturer
1	Aug 28	Introduction to CPH [slides] • Artificial intelligence in healthcare (Yu et al, 2018) • Do no harm: a roadmap for responsible machine learning for health care (Wiens et al, 2019)	Prof. Irene Chen
2	Aug 28	Case Study: AI and Radiology [slides] • CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning (Rajpurkar et al, 2017) • Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study (Zech et al, 2018)	Prof. Irene Chen
3	Sep 4	What makes healthcare unique? • Machine learning in medicine (Rajkomar et al, 2019) • A review of challenges and opportunities in machine learning for health (Ghassemi et al, 2020)	Prof. Irene Chen
4	Sep 4	Real World Applications of Generative AI in Liver Diseases and Transplantation • Development of a liver disease-specific large language model chat interface using retrieval-augmented generation (Ge et al, 2023) • Hepatology e-consult responses generated by artificial intelligence demonstrate accuracy but require human oversight (Huang et al, 2026)	Dr. Jin Ge (UCSF)

irenechen.net/cph100 has been updated with readings!

Course Logistics

- Office hours in Gateway 1110A on Thursday 9-10am
- Problem Set 1 has been released!
- Please submit DSP accommodations soon
- Meet our new CPH100 Reader Crystal Zhang

Step 1: Define the task

- I claim that we can define our problem as:
  - **X**: PLCO questionnaire information at the given prediction time
  - **Y**: Whether the patient develops lung cancer in the given prediction window
  - **S = f(Y)**: Risk score to define low-dose CT eligibility

Prediction is not causation

Risk Model

$$P(Y = 1 \mid X = x)$$

Treatment-effect Model

$$\mathbb{E}[Y(1) - Y(0) \mid X = x]$$

$Y(1)$ = lung-cancer if patient smokes

$Y(0)$ = lung-cancer if patient does not smoke

Prediction is not causation

Risk Model

$$P(Y = 1 \mid X = x)$$

- Which patient will get readmitted?
- Does this image contain a tumor?

Treatment-effect Model

$$\mathbb{E}[Y(1) - Y(0) \mid X = x]$$

- Should we give this patient this drug?
- Will surgery be better than chemo?

$Y(1)$ = lung-cancer if patient smokes

$Y(0)$ = lung-cancer if patient does not smoke

Step 2: Represent the patient

- Our goal is to turn messy clinical variables into a feature vector that our model can use

$$\begin{bmatrix} 1.1 \\ 0.5 \\ 2.1 \\ \vdots \\ 3.3 \end{bmatrix}$$

Who should we screen for lung cancer?

- Our patient is **68 years old** and has **40 pack years**
- What if we just used patient age?
- First, we normalize:
  - $x_{norm} = \frac{(x - \mu_{train})}{\sigma_{train}}$

Clinical data can be heterogenous

- Numerical:
- Categorical:
- Ordinal:
- Missing:

Clinical data can be heterogenous

- *Numerical*: age, pack years → normalize
- *Categorical*: sex, race → one-hot encoding
- *Ordinal*: education → ordered encoding
- *Missing*: unrecorded values → impute and/or indicate missing

Missingness is often information

MCAR

Missingness is *unrelated* to observed or unobserved features

MAR

Missingness depends on *observed* information

MNAR

Missingness depends on *observed and unobserved* information

Missingness is often information

MCAR

Missingness is *unrelated* to observed or unobserved features

Example: Smoking history forms were lost during digitizing records process

MAR

Missingness depends on *observed* information

Example: Pack-years is missing for younger patients, and you observe age

MNAR

Missingness depends on *observed and unobserved* information

Example: Heavy smokers are less likely to report how much they smoke

Step 3: Make a prediction

- How do we choose Y?

- We are really asking: “Did this person get lung cancer diagnosed within the window of time we are measuring?”
  - Did they get diagnosed?
  - Did that get recorded in the study?

Mystery: Performance is too high

- You build a model and evaluate it on held-out data (more later). You find out the AUC is 0.999 to predict lung cancer.

- What is happening here?

Label leakage can be subtle

- When the features contain information about the outcome that would not be available at prediction time

- Example: Patient has received a lung biopsy → could indicate lung cancer diagnosis that hasn't been recorded yet

Let’s make a prediction

- We assume a linear score:

$$z = \theta^T x + b$$

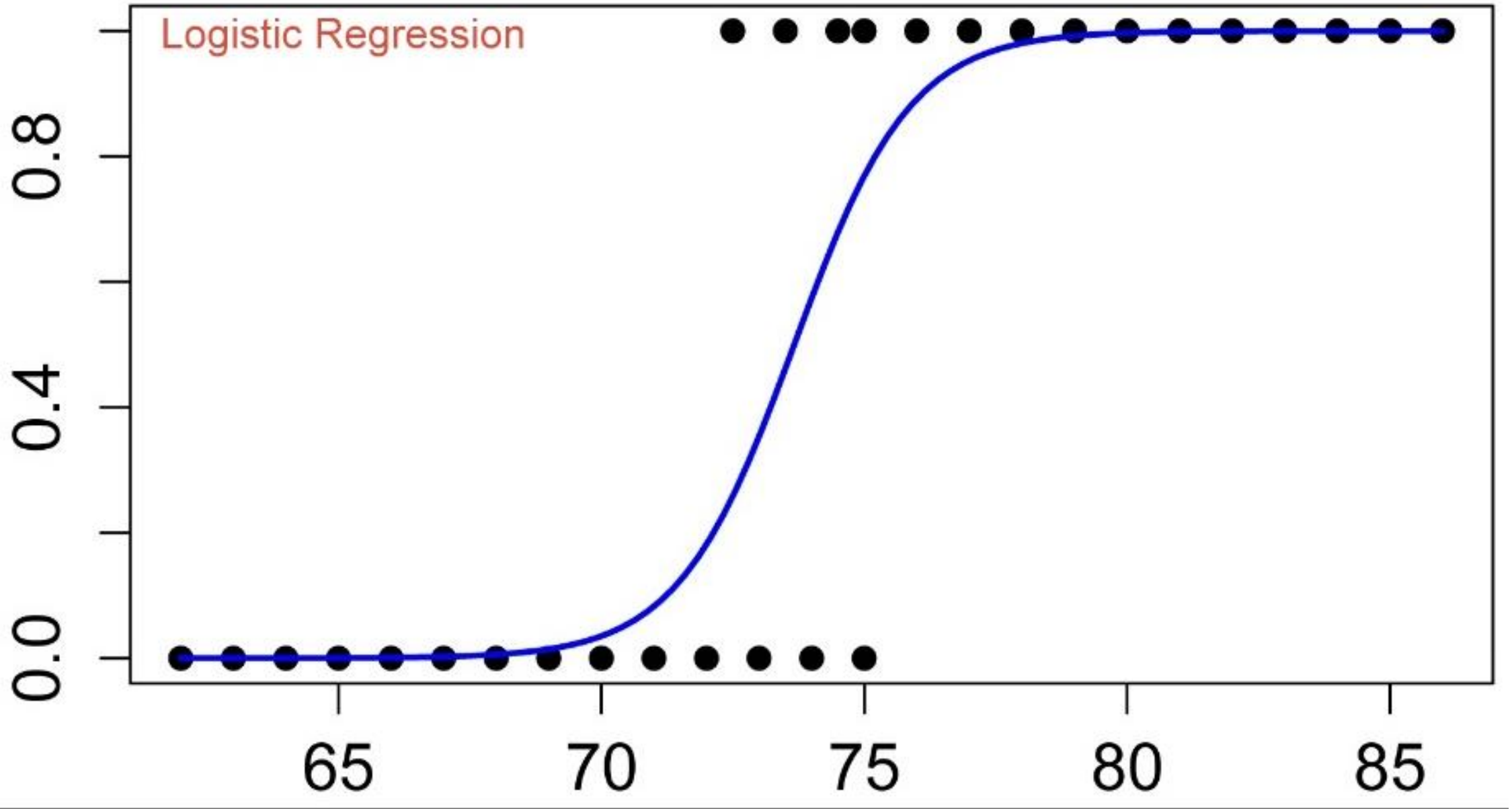
Let’s make a prediction

- We assume a linear score:

$$z = \theta^T x + b$$

- We can then convert that score into a probability

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$


Logistic Regression
0.0
0.4
0.8
65
70
75
80
85

Let's make a prediction

- Our patient is **68 years old** and has **40 pack years**

- Normalized **x = [1.2 , 1.4]**
- **theta = [0.8 , 0.5] , b = -1**

- What is this patient's risk?
  - $z = \theta^T x + b$
  - $\sigma(z) = \frac{1}{1+e^{-z}}$

Let's make a prediction

- Our patient is **68 years old** and has **40 pack years**

- Normalized **x = [1.2 , 1.4]**
- **theta = [0.8 , 0.5] , b = -1**

- What is this patient's risk?
  - $z = 0.8 * 1.2 + 0.5 * 1.4 - 1 \approx -1.16$
  - $\sigma(z) = \frac{1}{1+e^{1.16}} \approx \mathbf{0.76}$

Let's make a prediction

- Our patient is **68 years old** and has **40 pack years**

- Normalized **x = [1.2 , 1.4]**
- **theta = [0.8 , 0.5] , b = -1**

- What is this patient's risk?
  - $z = 0.8 * 1.2 + 0.5 * 1.4 - 1 \approx -1.16$
  - $\sigma(z) = \frac{1}{1+e^{1.16}} \approx 0.76$

But! This patient did **NOT** develop lung cancer. How should we update the weights?

Step 4: Learn from mistakes

- Binary Cross-Entropy Loss

$$L(y, p) = -[y \log p + (1 - y) \log 1 - p]$$

- For a given *p*, what happens if *y* = 1? What happens if *y* = 0?

Regularization penalizes large coefficients

- We can add a new term to our BCE loss:

$$L_{total} = L(y, p) + \left(\frac{\lambda}{2}\right)||\theta||^2$$

Regularization penalizes large coefficients

- We can add a new term to our BCE loss:

$$L_{total} = L(y, p) + \left(\frac{\lambda}{2}\right)||\theta||^2$$

- Why would we want this?

Gradients for Problem Set

$$\frac{\partial L}{\partial \theta} = (p - y)x + \lambda\theta$$

$$\frac{\partial L}{\partial b} = p - y$$

- Sanity check: If y = 1 and p is too small, what direction should the update move?

Stochastic Gradient Descent

$$\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta}$$

- If η is too large, learning is unstable.
- If η is too small, learning is too slow

Stability Trick

- Probabilities too close to 0 or 1 can make log(p) or log (1-p) numerically unstable

$$p \leftarrow clip(p, \epsilon, 1 - \epsilon)$$

Step 5: Choose a model

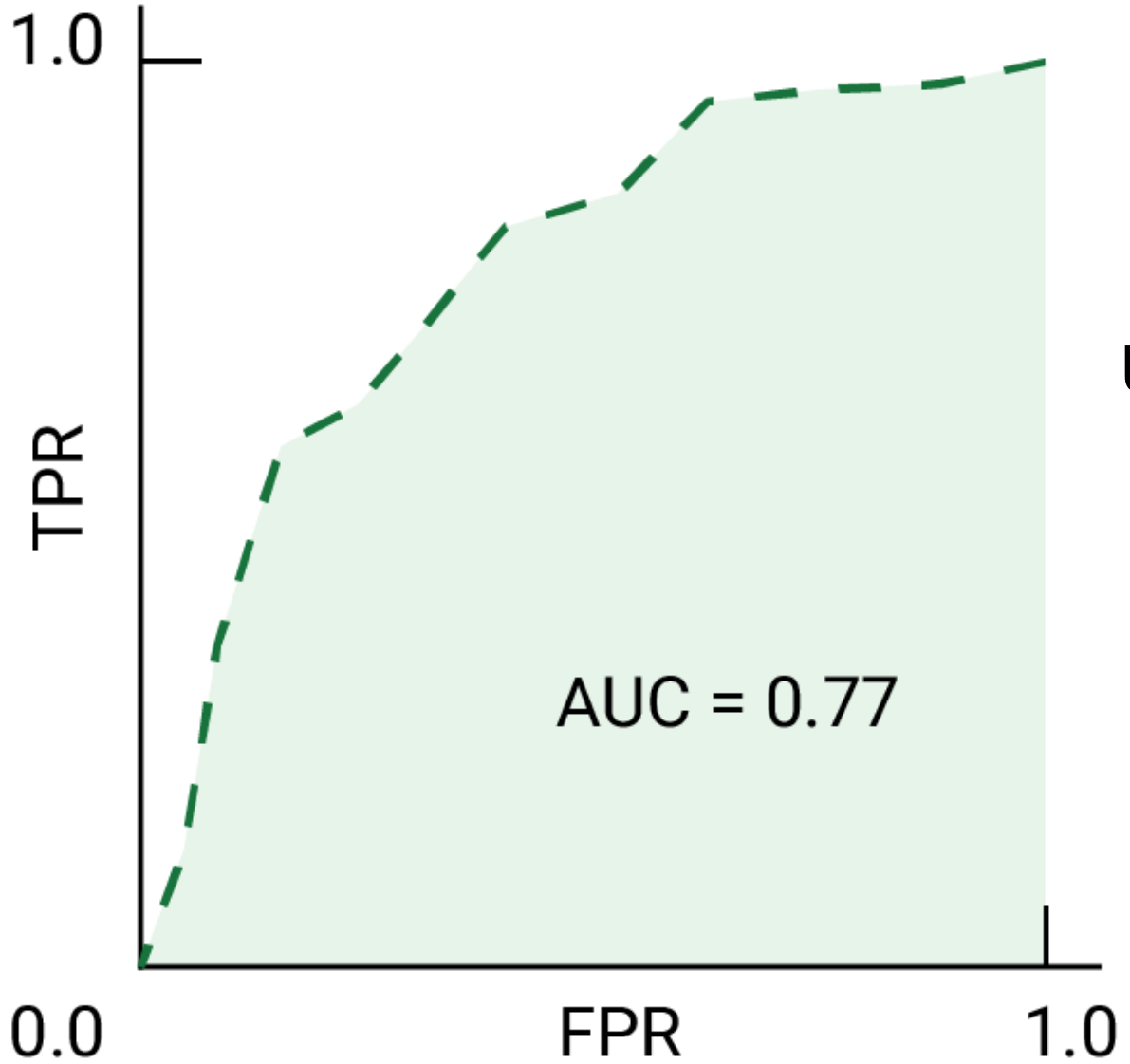
- **Train**: Fit θ and b
- **Validate**: Choose feature and hyperparameters
- **Test**: Estimate performance on held-out data

Hyperparameters are choices outside of the fitted coefficients

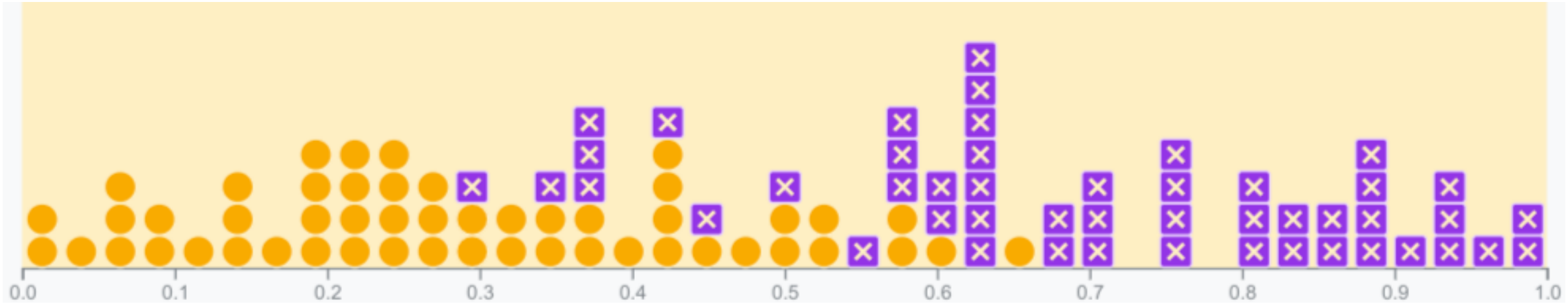
- **Learning rate** η : How large of a step to take in SGD?
- **Regularization** λ : How much to squeeze coefficients?
- **Epochs**: How long to let training go on for?

Step 6: Evaluate the chosen model

- **AUC**: Measures ranking of probabilities
- **ROC-curve and precision-recall curve**: trade-off sensitivity and specificity

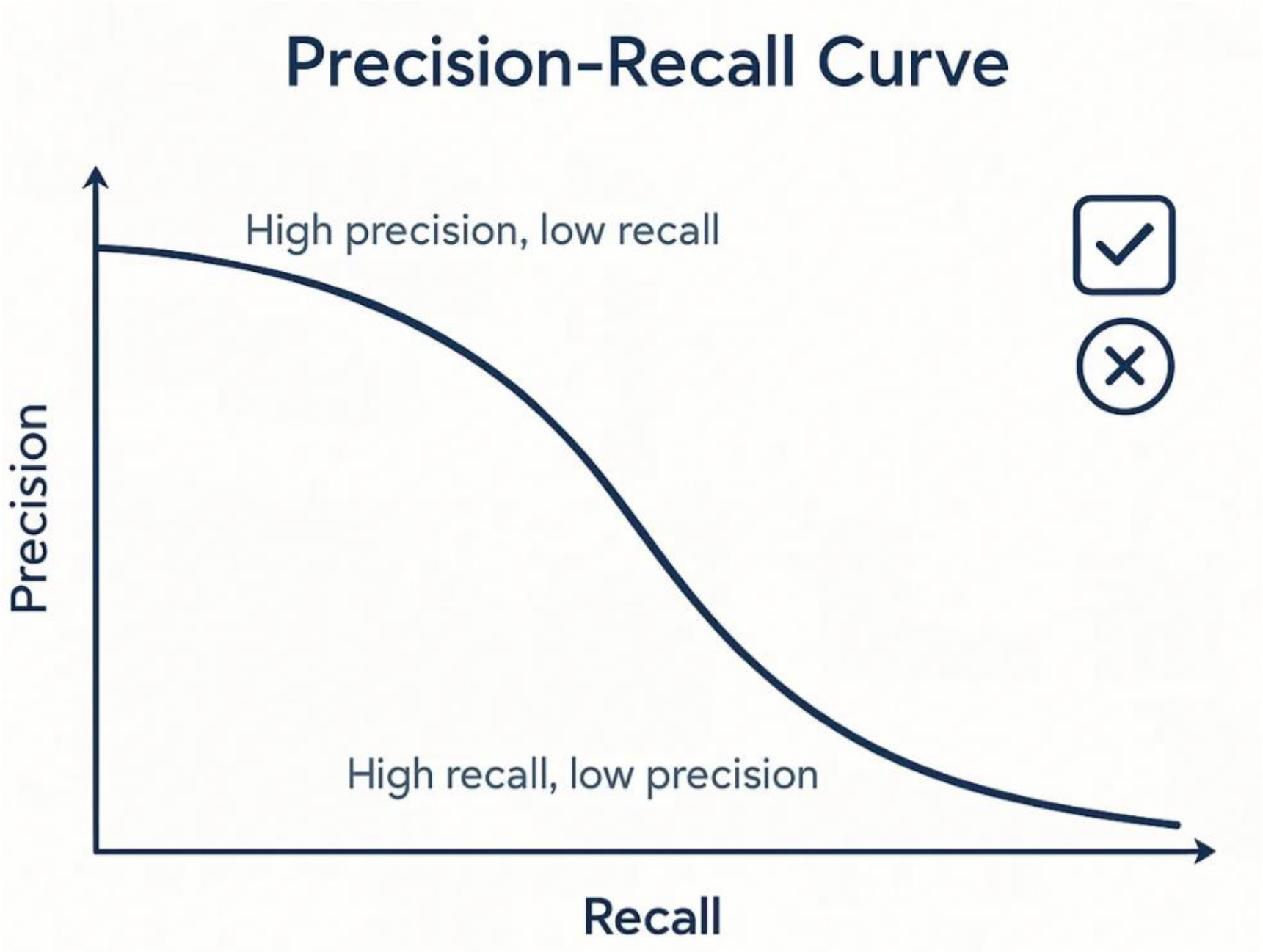


AUC = area under the ROC curve



AUC = probability that a random square is to the left of a random circle

https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc


Precision-Recall Curve
High precision, low recall
High recall, low precision
Precision
Recall

Make decisions after setting a threshold

- Given some score P(Y), we can set a threshold to determine action
  - If score > τ, we will do the low-dose CT
  - Otherwise, no screen
- If we set τ too high, we could have a lot of false negatives.
- If we set τ too low, we could have a lot of false positives.

What makes healthcare different?

What makes healthcare different?

EHR data is flawed	→	Created for billing, assumes access to care
Different protocols	→	Measurement and practice vary
Changing populations	→	Deployment differs from training
Evolving knowledge	→	Changing clinical practice

Summary

- ✓ Course logistics
- ✓ ML pipeline

How can we make Data 146 better for you?

Next class: Risk stratification and supervised learning