
LECTURE 6-1

Self-Supervised Learning and Imaging Foundation Models

Learning from unlabeled biomedical data



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



Where We Are

Weeks 6 to 8: foundation models for biomedical data.



today

The Label Bottleneck

Unlabeled biomedical data is abundant. Expert labels are slow, expensive, and scarce.

Unlabeled

Every ECG, slide, scan, and note already stored in hospital systems



**How can unlabeled data
reduce the labels we need?**

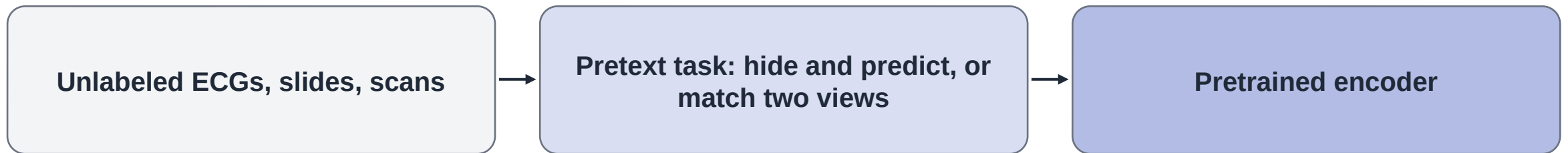
Labeled

What experts had time
to annotate

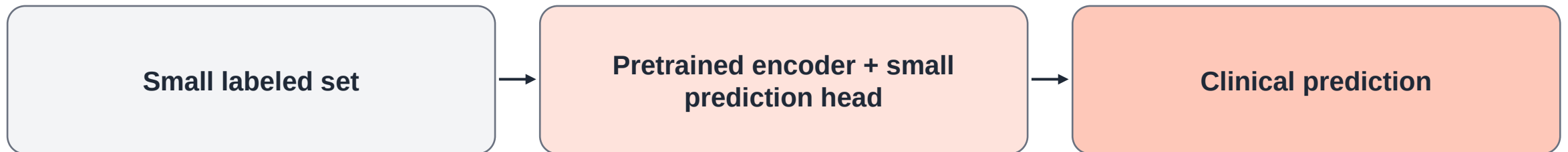
The Idea: Let the Data Supervise Itself

A pretext task creates labels from the data itself. Solving it forces the encoder to learn useful structure.

Stage 1: pretraining (no labels)



Stage 2: adaptation (few labels)



Three pretext families today: masked reconstruction, contrastive learning, and self-distillation.

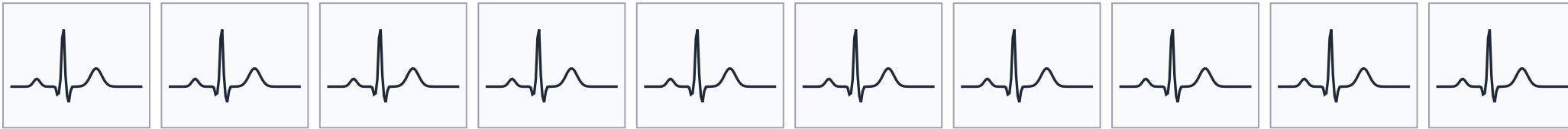


Masked Reconstruction

Step 1: Cut the Signal into Patches

A 10-second ECG becomes a sequence of patches: tokens, exactly like the image patches of a ViT in Week 3.

single lead · 10 seconds · no diagnosis



each patch → linear projection + position → token

Same idea as in Week 3: images → 16×16 patches; text → word pieces.

Step 2: Hide Most of the Patches

Randomly mask a large fraction of the patches. The model must infer what is missing from what remains.

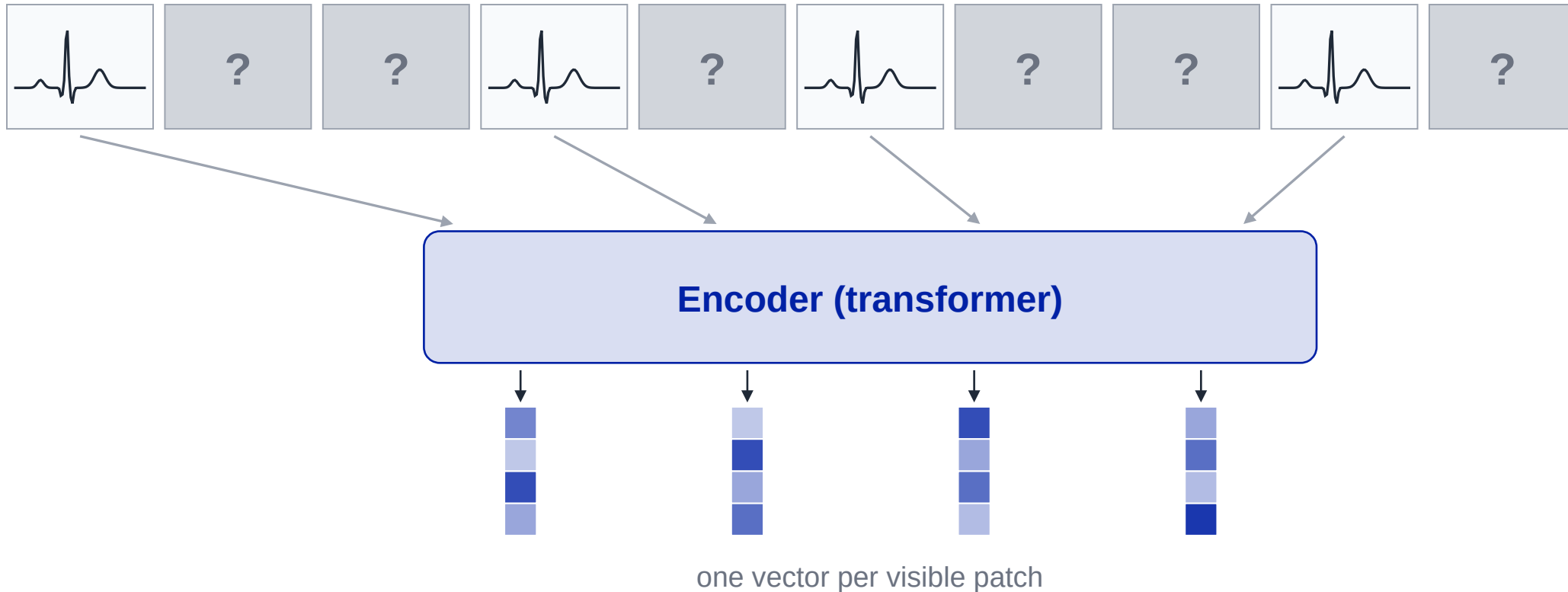


6 of 10 patches hidden (MAE hides 75% of image patches)

Hiding a lot makes the task hard enough that copying a neighbor is not enough.

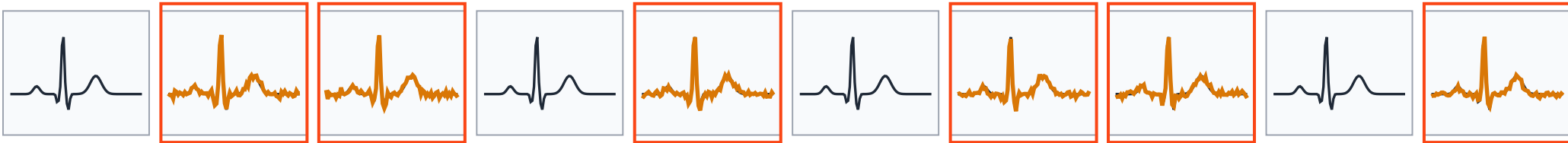
Step 3: Encode Only the Visible Patches

The encoder sees only the visible patches. This keeps pretraining fast and forces it to build a summary.



Step 4: Reconstruct and Compare

A small decoder predicts the hidden patches. The loss compares prediction and truth on the hidden patches only.

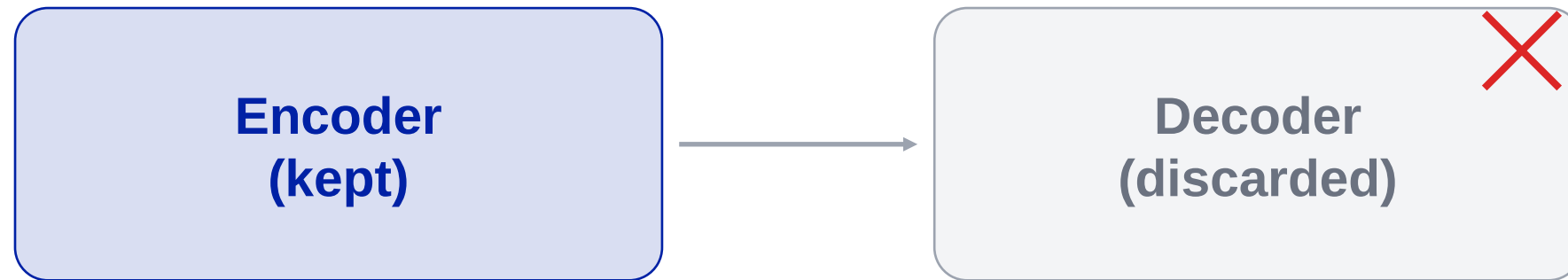


orange = prediction on hidden patches

loss = mean squared error, on hidden patches only

Step 5: Keep the Encoder, Discard the Decoder

After pretraining, only the encoder is kept. It is the pretrained model used for downstream tasks.



The pretrained encoder turns any new ECG into a useful vector, before seeing a single label.

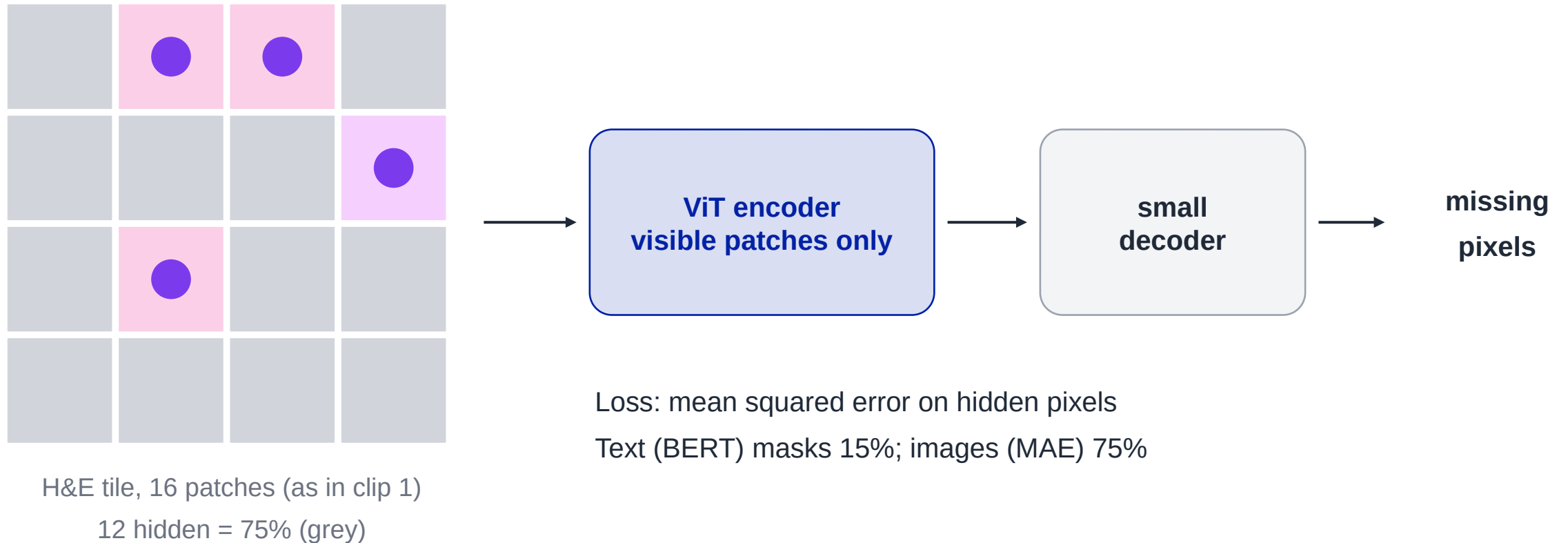
Animation: Masked Reconstruction

Animation clip: week06-clip1-masked-reconstruction.mp4

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

Same Recipe for Images: Masked Autoencoders

MAE cuts an image into 196 patches, hides 75%, encodes the visible 49, and reconstructs the missing pixels.

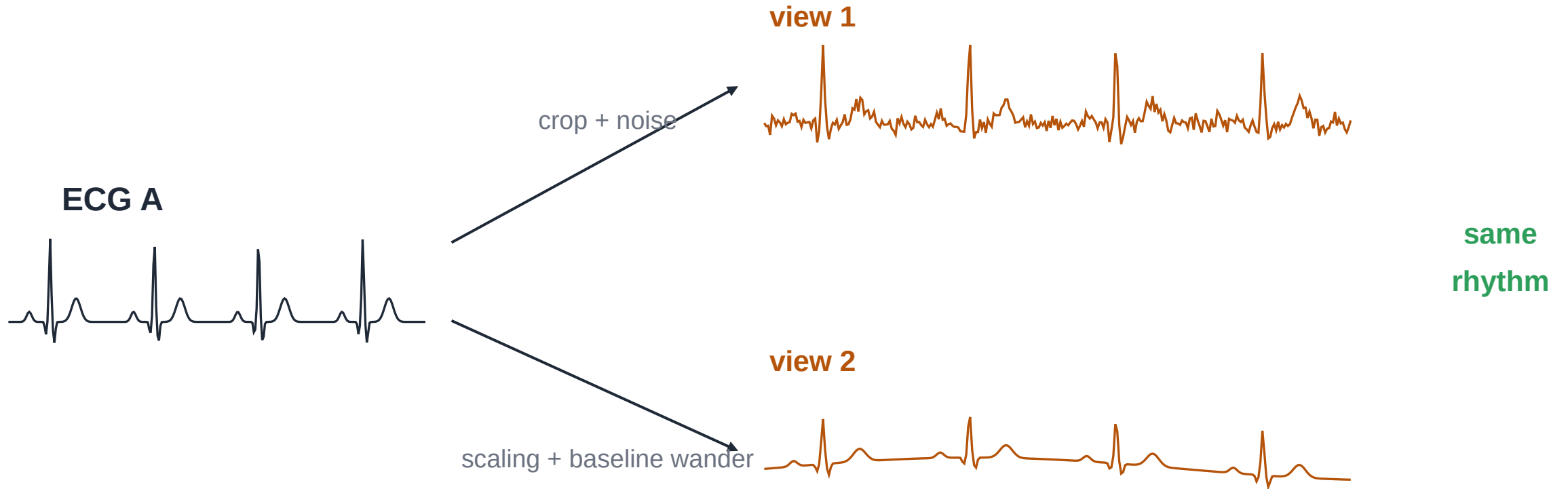




Contrastive Learning

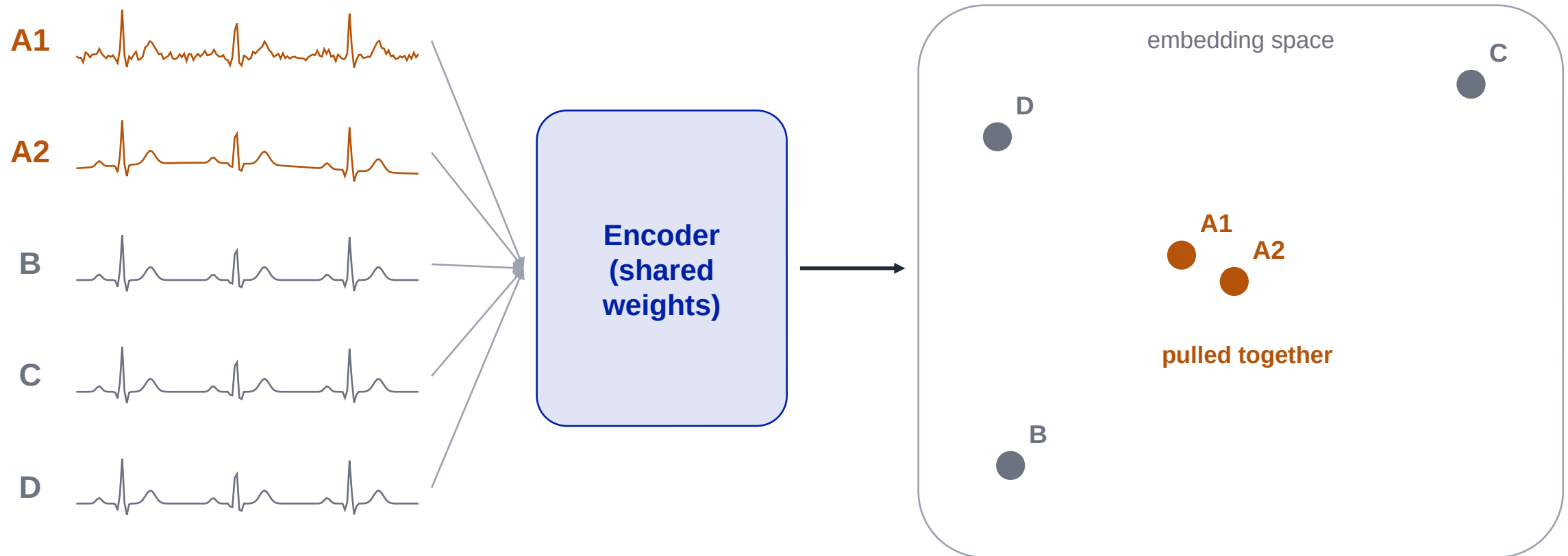
Two Views of the Same Recording

Augmentations change how the recording looks, not what it shows. Both views should map to the same embedding.



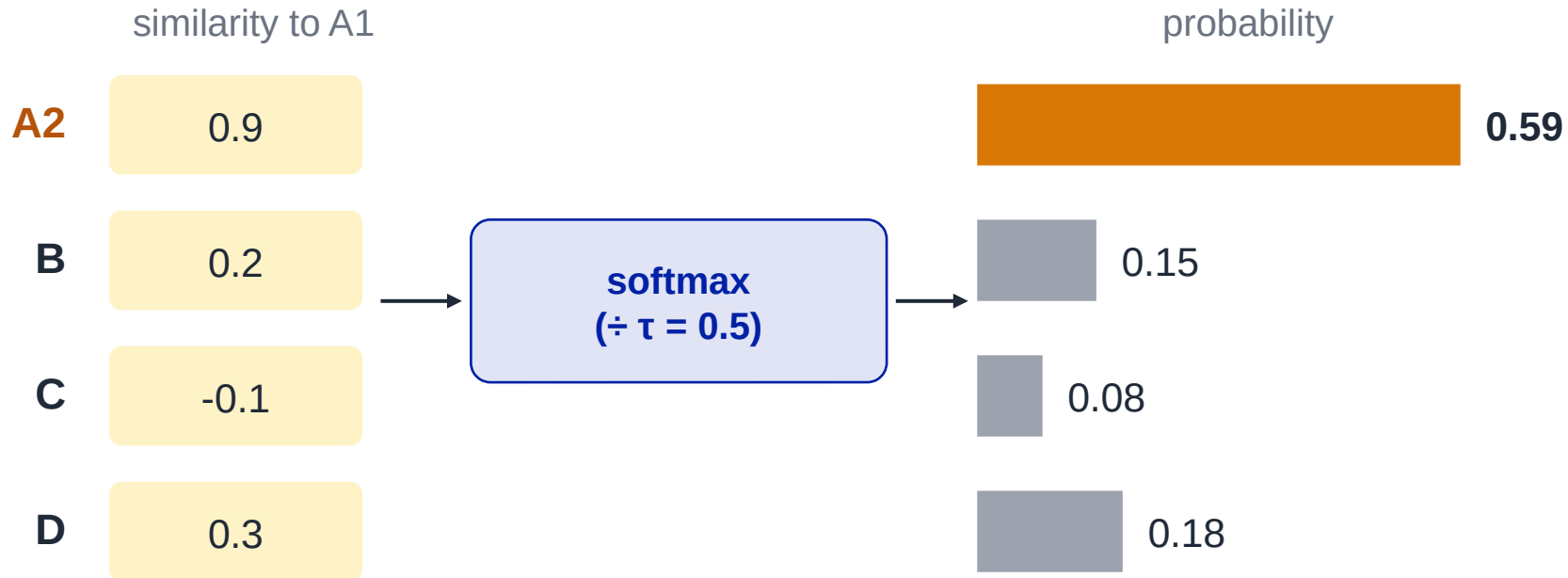
Pull Together, Push Apart

In a batch, each view's partner is the positive. Every other recording in the batch is a negative.



The InfoNCE Loss

Turn similarities into probabilities with a softmax. The loss is low when the positive gets the highest probability.



$$\text{loss for A1} = -\log p(\text{A2}) = -\log 0.59 = 0.52$$

Augmentations Encode Your Assumptions

An augmentation says 'this change does not matter'. In medicine, the wrong one can erase the signal.

Data	Usually safe	Use with care
ECG	noise, amplitude scaling, baseline wander	short crops (can miss intermittent events)
Pathology	rotation, flips, blur	strong color jitter (stain carries information)
Chest X-ray	small crops, contrast	horizontal flip (moves the heart)

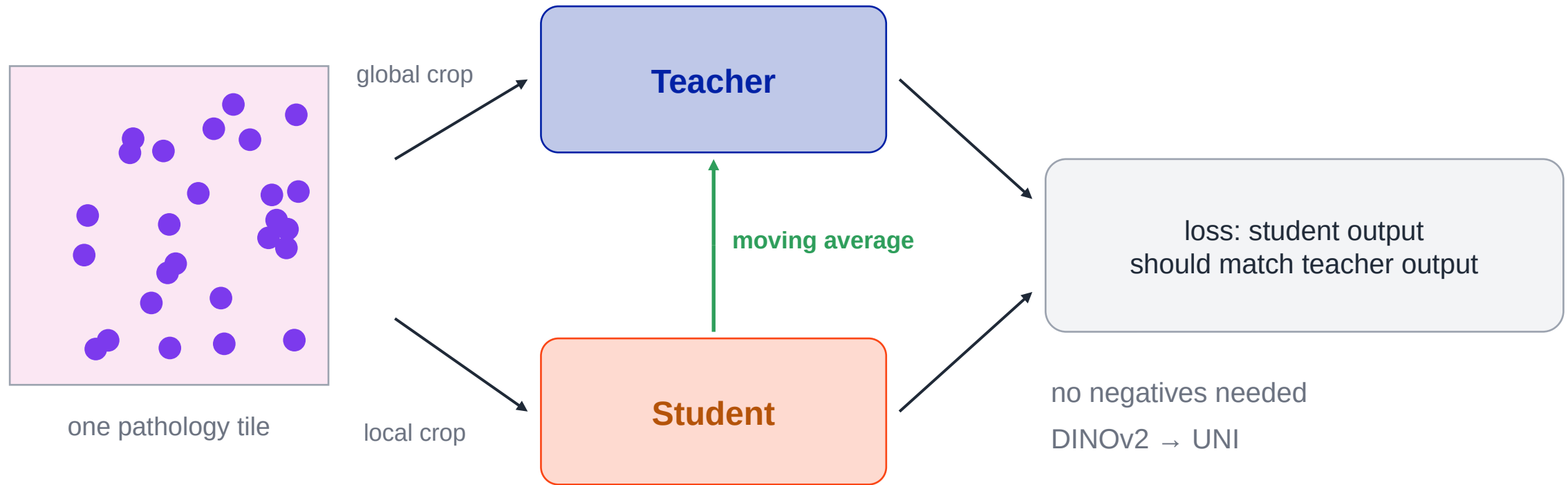
Animation: Contrastive Learning

Animation clip: week06-clip2-contrastive-learning.mp4

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

Without Negatives: Self-Distillation

A student network learns to match a teacher on different crops. The teacher is a slow-moving average of the student.





Using a Pretrained Encoder

Three Ways to Use a Pretrained Encoder

They differ in how much is trained and how many labels are needed.

Linear probe

- Encoder frozen
- Train one linear layer
- Few labels; cheap
- Measures representation quality

Fine-tune

- Encoder updated too
- Usually most accurate
- More labels and compute
- Can overfit small sets

Few-shot / kNN

- No training
- Compare embeddings
- A handful of labels
- Quick baseline

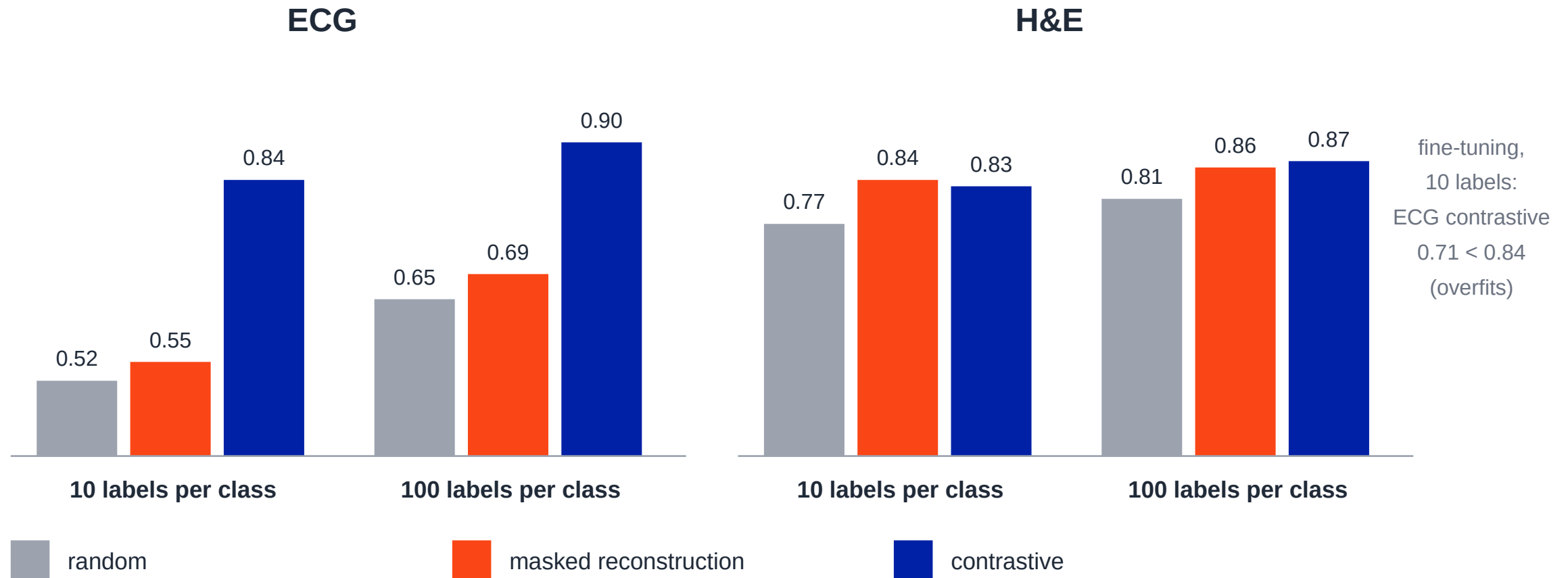
Animation: Linear Probing

Animation clip: week06-clip3-linear-probing.mp4

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

Label Efficiency in the Notebook

Linear probes with 10 and 100 labels per class, each encoder against a random one of the same architecture.

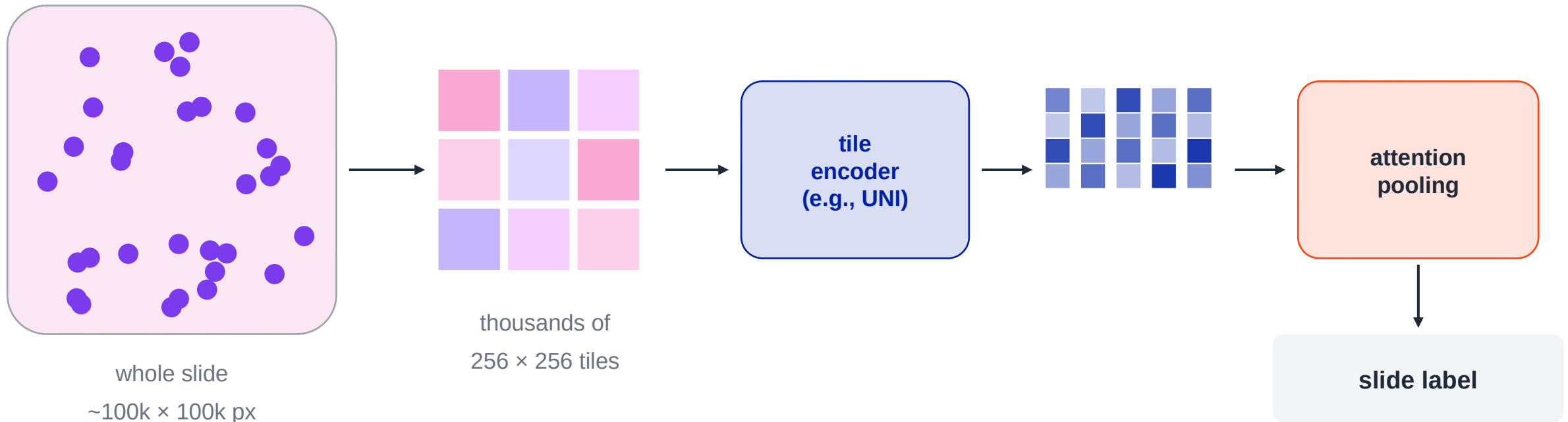




Imaging Foundation Models

Pathology: Gigapixel Slides as Tiles

A whole-slide image is far too large for one model input. Cut it into tiles, embed each tile, then aggregate.



Imaging Foundation Models

The same few recipes, scaled to millions of medical images.

Model	Data	Self-supervised method	Scale
RETFound	Retinal images	MAE	~1.6 M images
UNI	Pathology (H&E)	DINOv2, ViT-L	>100 M tiles, >100k slides
Virchow	Pathology (H&E)	DINOv2	~1.5 M slides
Prov-GigaPath	Pathology (H&E)	DINOv2 tiles + slide encoder	1.3 B tiles, ~171k slides

Thursday: He et al. (MAE) and Chen et al. (UNI)

What Can Go Wrong

Pretraining on millions of images does not guarantee the right features.

Site and stain shortcuts

Encoders can learn which hospital or scanner, not the disease (Week 12)

Benchmark overlap

Evaluation slides may appear in pretraining data

Unsafe augmentations

A transformation can remove the clinical signal

Compute and access

Few groups can pretrain; many models are gated

Recap

Masked reconstruction

Hide most patches; encode the visible ones; reconstruct the rest (MAE, RETFound)

Contrastive learning

Two views of one input should match; other inputs are negatives (InfoNCE)

Self-distillation

Student matches a moving-average teacher; no negatives (DINOv2, UNI)

Use the encoder

Linear probe, fine-tune, or kNN; far fewer labels needed

Before Thursday

Read

He et al., Masked Autoencoders Are Scalable Vision Learners (CVPR 2022) and Chen et al., UNI (Nature Medicine 2024).

Write

Reading responses on Canvas before 1:00 PM Thursday.

Explore

Week 6 reference notebook (ungraded): synthetic ECG and Kather tiles, with probing and fine-tuning against random baselines.

Next week

Vision-language and multimodal models.