LMP 1210 - Basic principles of machine learning in biomedical research

Rahul G. Krishnan

Canada CIFAR AI Chair

Tier II Canada Research Chair in Computational Medicine

Announcements

- Next week, we will start project presentations.
- 6 teams on March 28
  - Your project does not need to be 100% done, report what remains to be done for the upcoming week
- 7 teams on April 4
- Each team will have 15 minutes total – 12 minutes for presentation
- Remember:
  - Most of your grade is based on the presentation + project report.
  - Reports will be graded by me and the TAs, presentations by me
- Bring:
  - A *practiced* presentation
  - Laptop
  - HDMI Adaptor

Grading rubric

- Scope of the project
- How much of the initial project proposal you have completed
- How well you present the core idea:
  - What problem do you care about
  - Why do you care about solving it?
  - How did you solve it?
  - When you implemented your method, what happened?
- What have you learned through doing the project?

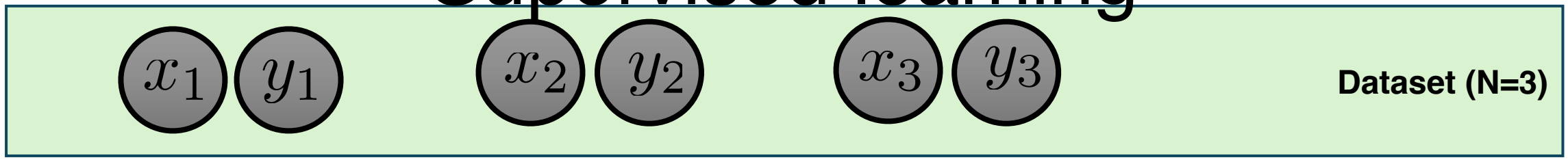
Overview

- Limitations of supervised learning
- Self supervised learning
- [Time permitting: Transformers]
- Imaging in computational histopathology

Imaging in medicine

- History of Medical Imaging, Bradley et. al, 2008
- Nuclear medicine: Using radiation to *see* inside the human body
  - X-ray discovered in 1895 (won the Nobel in 1901)
  - CT, PET discovered thereafter
- Magnetic resonance imaging: Mapping resonance in the body to images
- Ultrasound imaging: Mapping high-frequency sound waves to images
- Histopathological imaging: Images of stained tissue samples

Supervised learning



- Given a dataset, the model parameters are learned via **maximum likelihood estimation**

$$\mathcal{L}(y, x) = \log p(y|x; \theta)$$

Score function (high is good, low is bad)

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, x_i)$$

Solve this optimization problem to **learn** the model. Often formulated as a minimization of the negative of the log-likelihood function

Deep neural networks typically learned using tools that leverage automatic differentiation

PyTorch

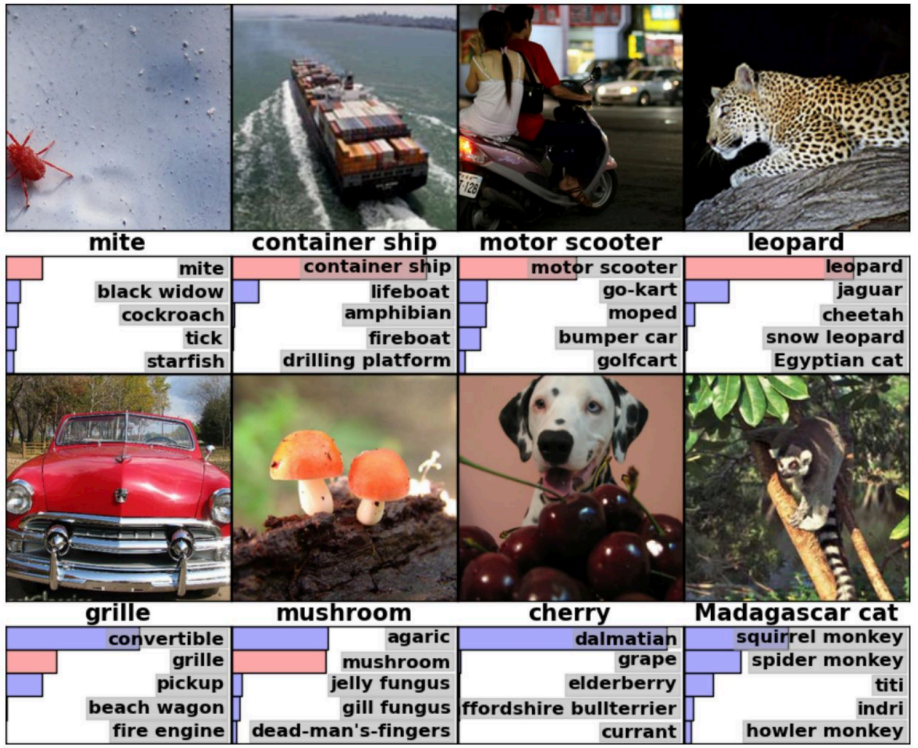
Computer vision

- Computer vision has had a front row seat to the advances in deep learning

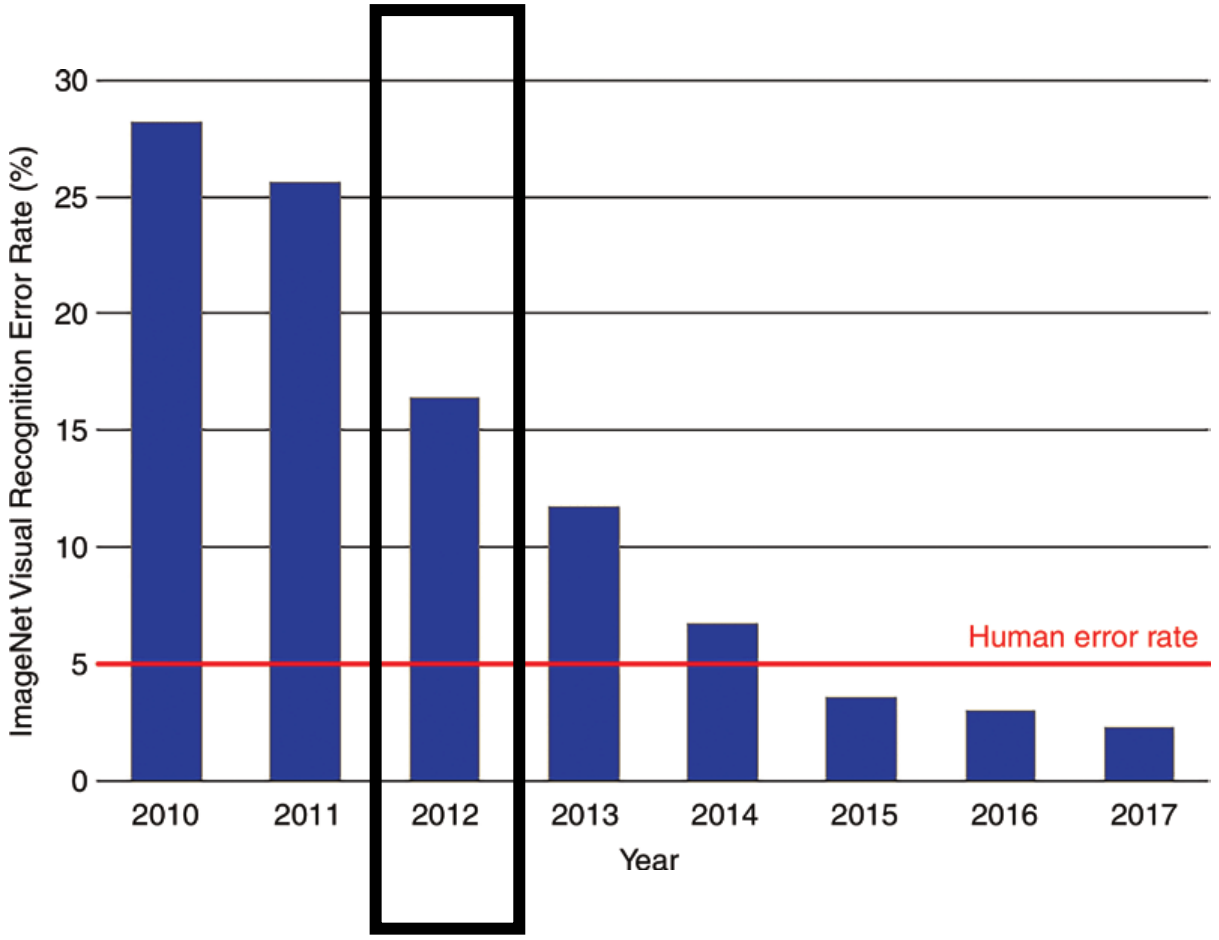
ImageNet Challenge

Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

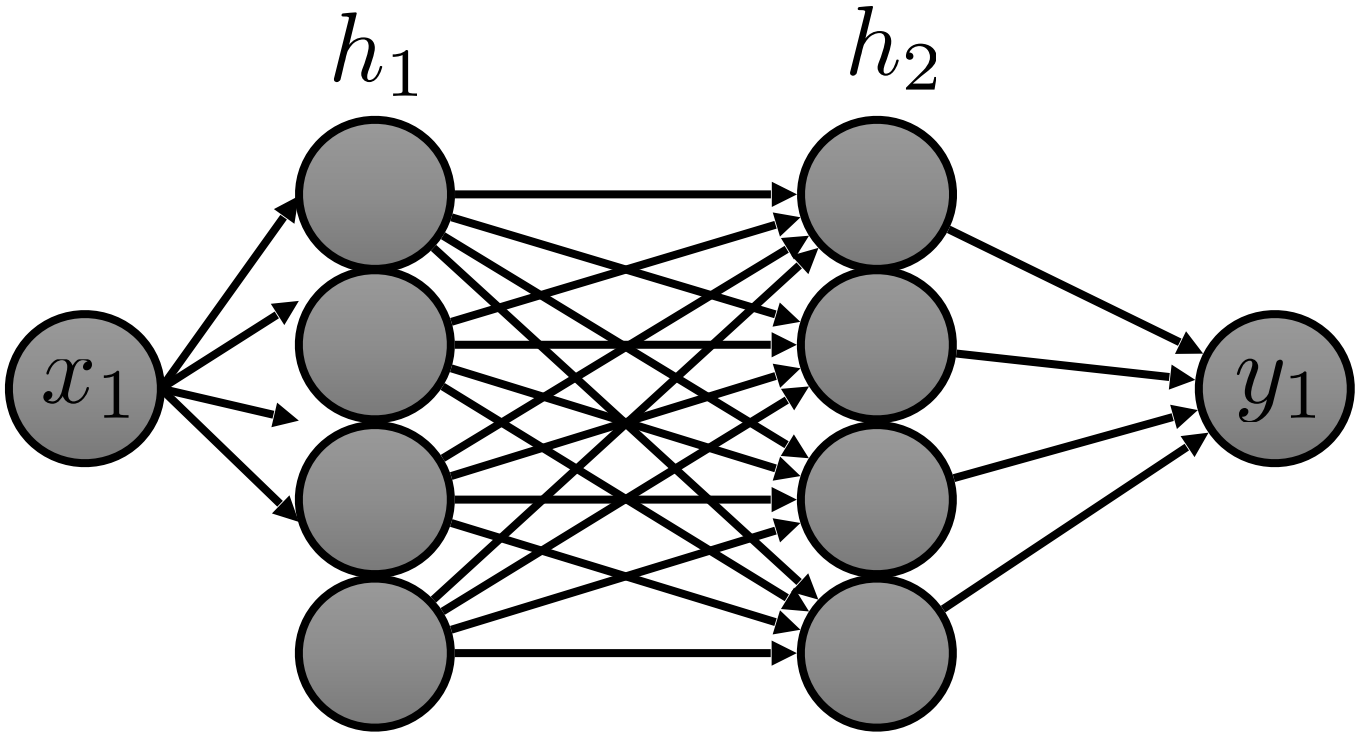


Error rates on Imagenet over time



Neural networks in a slide

- Simplest neural network is a multi-layer perceptron
- Neural networks are known to be universal function approximators



$$p(y|x)$$

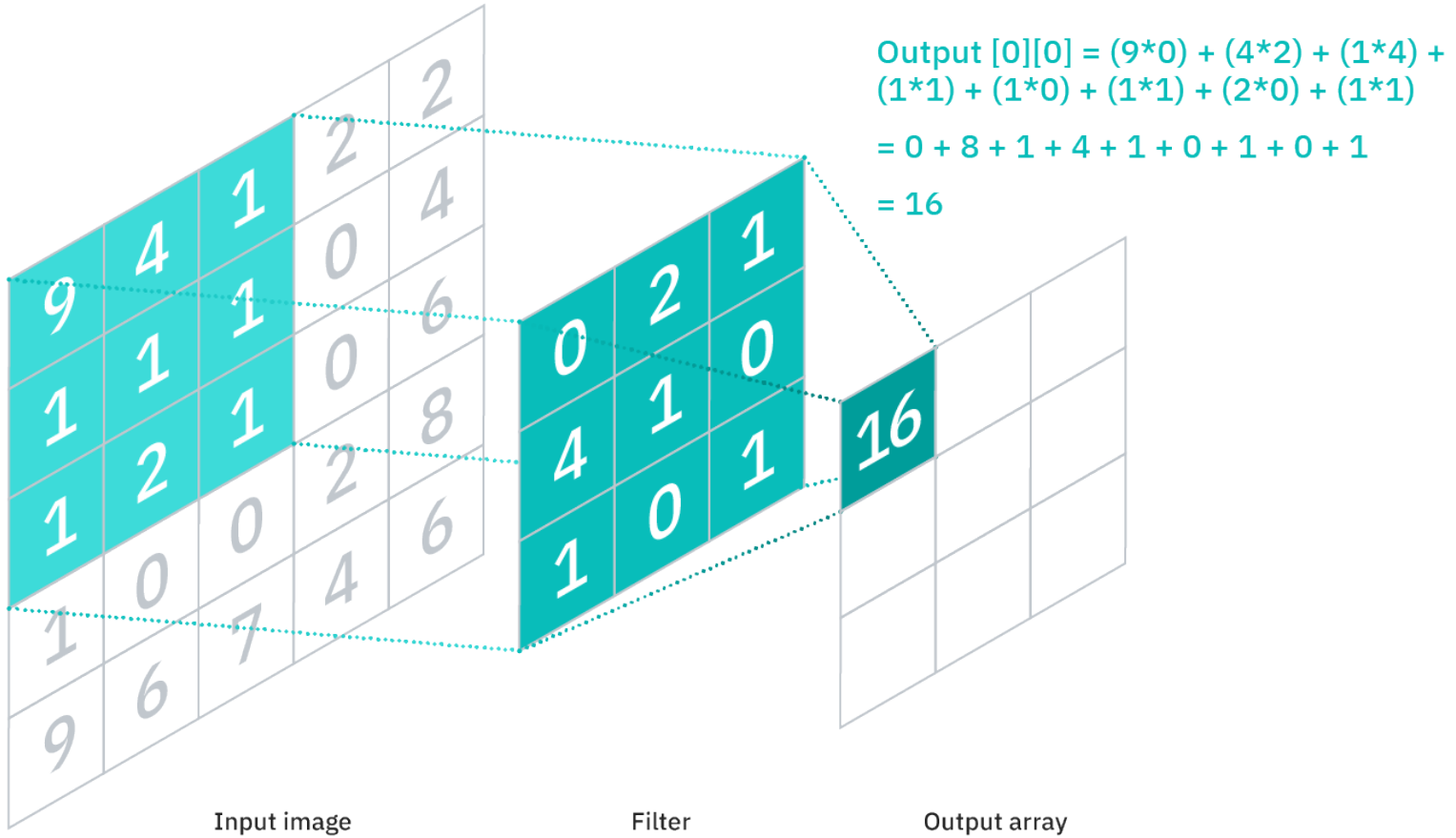
$$y = \phi(W^T h_2 + b)$$

$$h_2 = \phi(W^T h_1 + b)$$

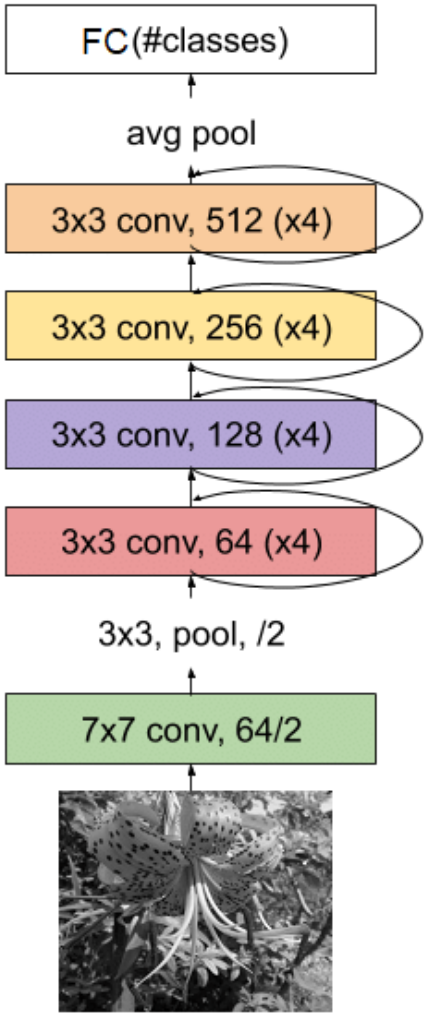
$$h_1 = \phi(W^T x + b)$$

Convolutional neural networks

Capture the fact that we may want representations that are spatially invariant



Deep residual neural networks



Researchers found that deep networks had a hard time learning the identity function.

They added a skip-connections between layers:

$$h_k = \phi(conv(h_{k-1})) + \sum_{j<k-1} h_{k-j}$$

Deep Residual Learning for Image Recognition, He et. al, 2015

Limitations of supervised learning

- Deep neural networks have proven very successful in learning useful representations of image data from large datasets
- Models like AlexNet, ResNet trained on imagenet capture features useful for multiple different tasks
- For a new task:
  - Need fine-grained labels associated with each example
  - Standard approach: Use a pre-trained imagenet model and fine-tune on new dataset
- Self-supervised learning:
  - What if we do not need labels to learn good representations?

Unsupervised learning

$$\mathcal{L}(x) = \log p(x;\theta)$$

Score function (high is good, low is bad)

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(x)$$

Solve this optimization problem to **learn** the model. Often formulated as a minimization of the negative of the log-likelihood function

Semi-supervised learning

$$\theta = \arg\max_{\theta} \sum_{i=1}^{3} \mathcal{L}(x;\theta) + \mathcal{L}(y_1|x_1;\theta_2) + \mathcal{L}(y_3|x_3;\theta_2)$$

- Have a combination of labelled and un-labelled data in your dataset

Unsupervised and semi-supervised learning of high-dimensional images is hard

- Even if there is a small space of concepts unsupervised models of image data are challenging to build
- Need a good model of each pixel in the image.
- Recently there has been a lot of work in leveraging generative adversarial networks for this problem
- Idea: Can we build representations without labels and without modeling each pixel as a random variable?

Self-supervised learning

- Recent (last 4-5 years) development in machine learning
- **Principle**: Leverage domain knowledge about what kinds of information the representation should contain when building it
- Learn about self-supervised learning by examples

Notation

ϕ

- Feature function [Resnet]

$\mathcal{T} : x \rightarrow \tilde{x}$

- Transformation of an image [random crop, rotation, jittering, color normalization]
  - Preserves the identity of the image

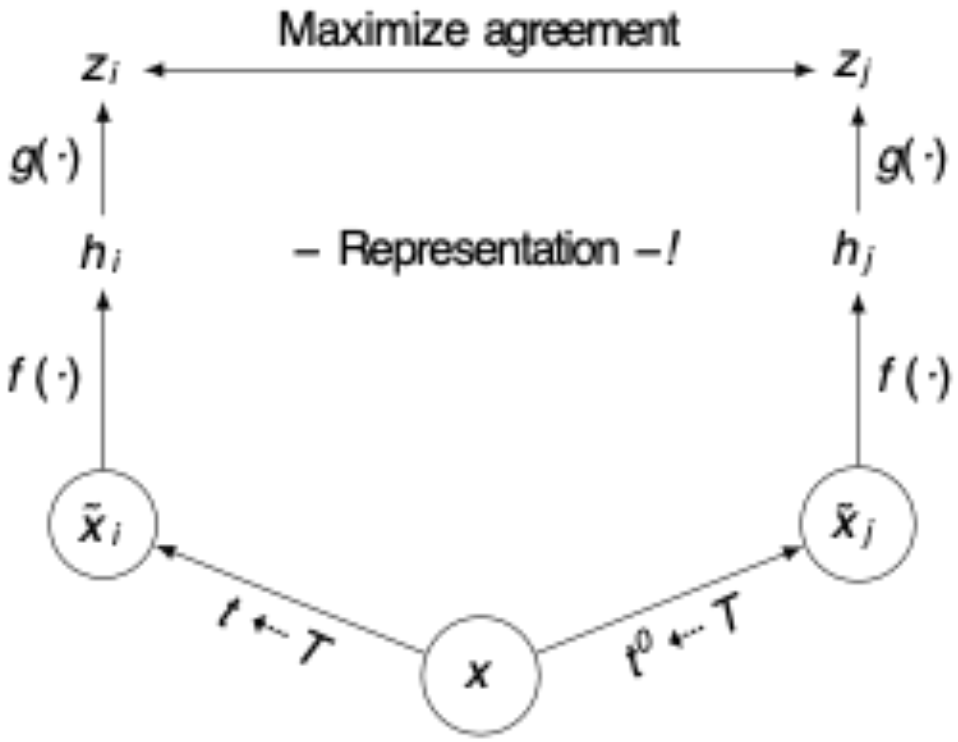
$\mathrm{sim}(k, k')$

- Similarity function
  - Measure of similarity of two vectors
  - Mean squared error, cosine similarity

SSL 1 - Learning with contrastive examples

- A Simple Framework for Contrastive Learning of Visual Representations, Chen et. al, ICML 2020
- **Builds upon earlier work:** Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA, Hyvarinen et. al

SIMCLR: Self-supervised learning with contrastive examples



Randomly sample a mini-batch of datapoints.

Minimize loss below

Goal: Learn representations that recognize that the class of transformations in T preserve identity.

Note: No labels used.

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}, \qquad (1)$$

How good are the representations?

A Simple Framework for Contrastive Learning of Visual Representations

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.8 for experimental details and results with standard ResNet-50.

SSL - Learning without contrastive examples

- In the above examples, the quality of representations will depend on the choice of negative examples used.
- Can we learn without negative examples?
- DINO: Emerging Properties in Self-Supervised Vision Transformers, Caron et. al, 2021
  - Key idea: Instead of comparing the representations with respect to random negative examples, compare the representation to a different crop of itself

DINO

Decision making with images

- Ultrasound:
  - Echocardiograms
    - Visualize beating of the heart to assess normal function
  - Abdominal ultrasounds
    - Assess healthy function of abdominal organs
- X-rays:
  - Breast cancer screening
  - Guiding surgery to remove blood clots, insert catheters
  - Friday: Hear from Ruizhi Liao on combining text and chest x-ray data

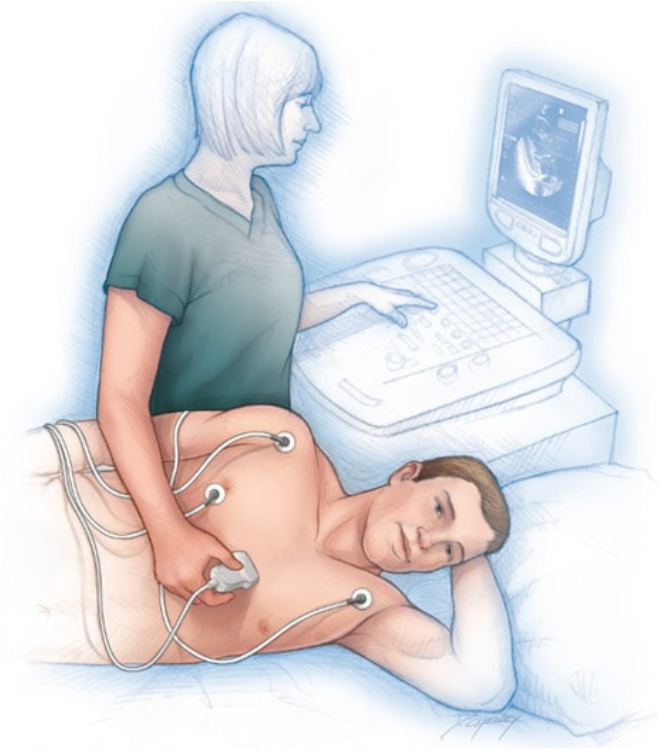
Technical issues in machine learning for medical imaging

- The general setup is almost always as follows:
  - Collect a large set of images [X]
  - Use notes/clinical variables/expert annotation to come up with labels [Y]
  - Use a deep learning model predict Y from X

- Fairness:
  - Reading Race: AI Recognises Patient's Racial Identity In Medical Images, Banerjee et. al, 2021
- Selection bias:
  - Causality matters in medical imaging, Castro et. al, 2019

Case study 1: Deep learning for echocardiograms

- Sound waves to image the heart
- Why:
  - Check for problems with your valves or chambers
  - Check if heart problems are causing shortness of breath
  - Assess congenital heart defects



A taxonomy of echocardiograms

- Most common: Transthoracic echocardiogram
- Transesophageal echocardiogram
  - Transducer guided down patient's throat
  - Records sound waves bouncing off the heart pumping and interprets them as images
- Doppler echocardiogram
  - Used to assess bloodflow
- Stress echocardiogram
  - Ultrasound after excercise

Case study 1: Predicting cardiac amyloidosis

- Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms, Goto et. Al, Nature Communications, 2021
- Cardiac amyloidosis
  - deposition of protein in the heart muscle, can result in heart failure
  - believed to be rare but likely underdiagnosed
  - manifests in both ECGs and echo-cardiography but features are not highly specific and difficult to spot
  - Gold standard: biopsy (costly and risky to patient)

Where machine learning can help

- How can we design a method that:
  - Fits into the clinical workflow for cardiac patients
  - If used, improve underdiagnosis of disease?
- Key-idea: Two-stage approach
  - Step 1: Build ML models from ECG data (readily available at most care providers)
    - Finding: Models have decent accuracy but not enough for conclusive diagnosis
  - Step 2: Build ML models from echocardiogram data
    - Finding: Models outperform human experts
  - Use step 1 to decide which patients should undergo an echocardiogram and apply model from step 2

A multi-center study

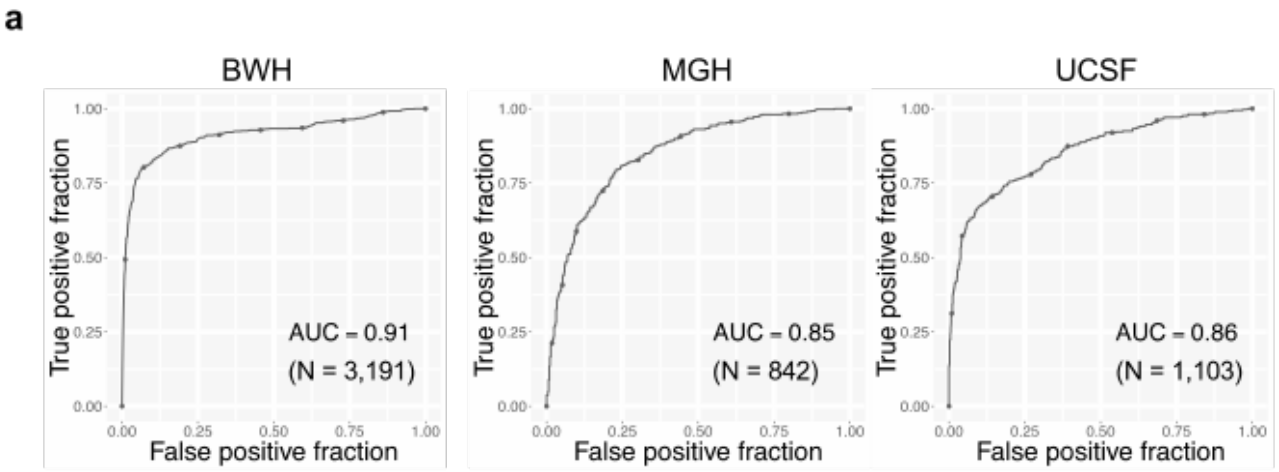
Table 1 Study-level demographic information (ECG cohort).

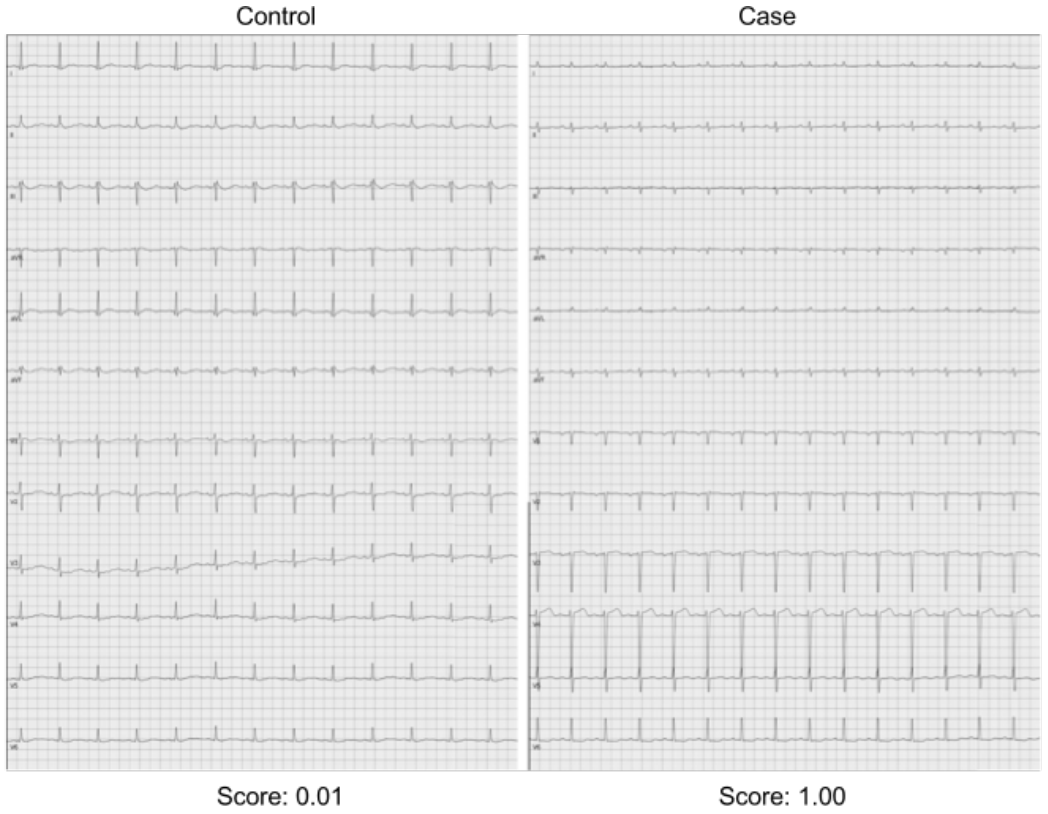
	BWH		MGH		UCSF	
	Case	**Control**	**Case**	**Control**	**Case**	**Control**
Number of studies	2249	8684	405	437	372	731
Age, years ± SD	69.9 ± 10.4	62.3 ± 13.2	72.9 ± 9.0	73.8 ± 8.8	67.7 ± 12.9	67.5 ± 11.7
Age Groups						
≤30, *n* (%)	2 (0.1)	97 (1.1)	1 (0.2)	1 (0.2)	2 (0.5)	0 (0.0)
30–50, *n* (%)	78 (3.5)	1,370 (15.8)	7 (1.7)	6 (1.4)	36 (9.7)	69 (9,4))
50–70, *n* (%)	901 (40.1)	4548 (52.4)	143 (35.3)	135 (30.9)	136 (36.6)	278 (38.0)
70–90, *n* (%)	1242 (55.2)	2606 (30.0)	254 (62.7)	295 (67.5)	198 (53.2)	384 (52.5)
>90, *n* (%)	26 (1.2)	63 (0.7)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
HR, bpm ± SD	76.4 ± 16.7	75.9 ± 18.5	78.6 ± 16.6	75.1 ± 19.8	79.6 ± 18.7	72.2 ± 16.3
Sinus rhythm, *n* (%)	1,736 (77.2)	8,072 (93.0)	283 (69.9)	371 (84.9)	365 (98.1)	729 (99.7)

HR heart rate, *BWH* Brigham and Women's Hospital, *MGH* Massachusetts General Hospital, *UCSF* University of California San Francisco. *N* represents the number of studies.

Step 1: ECG model

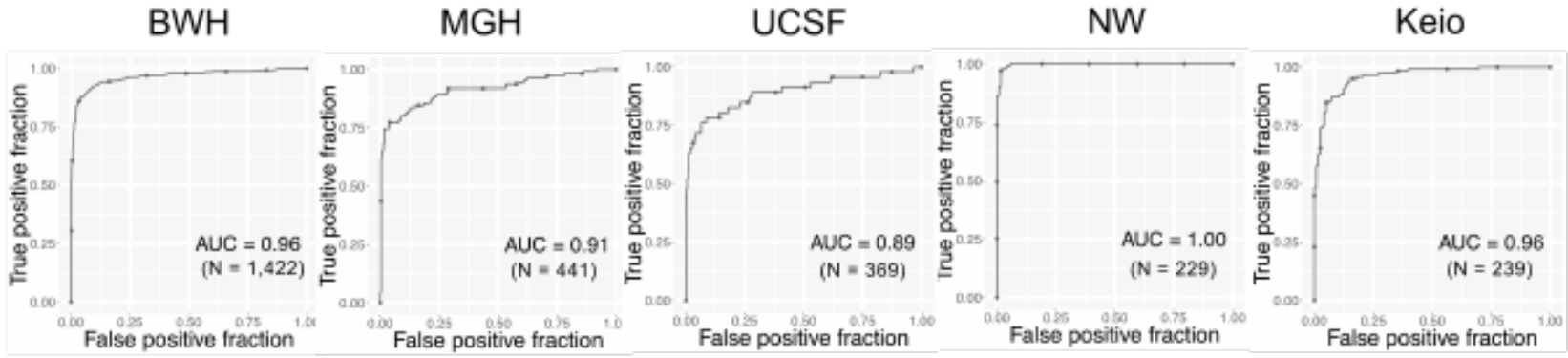
- Results Ok but not considered good enough for evaluating interventions for a rare diagnosis since it will result in a large number of false positives

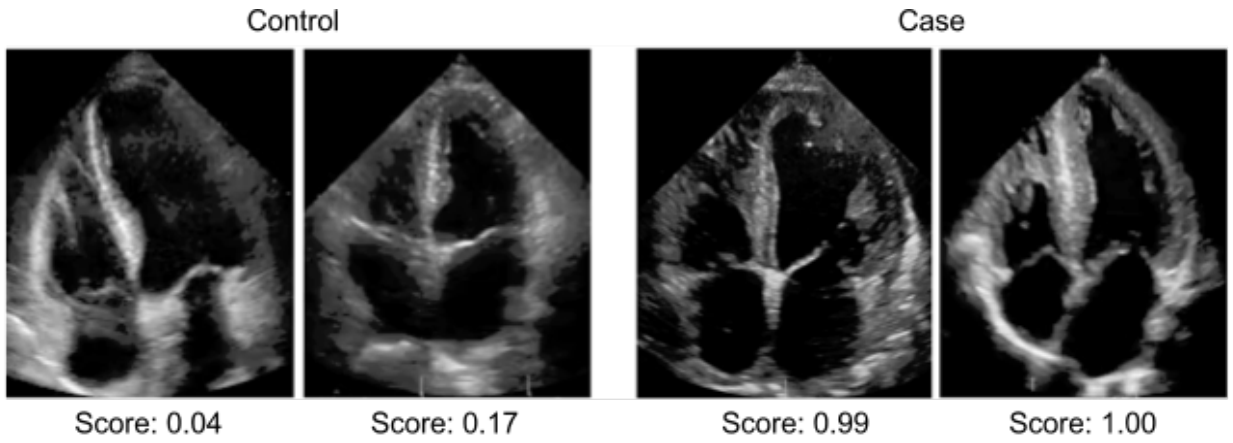




Step 2: Echocardiogram model

- Performance significantly better when using a richer (but more expensive) data modality





Recall: Metrics

- Positive Predictive Value (PPV): TP/(TP+FP)
  - A high PPV will indicate that a positive result is likely correct
- Sensitivity: TP/(TP+FN)
  - A highly sensitive test will have few-false negatives

Analyzing the combined approach

- ECG model:
  - MGH: PPV 3.9% with Sensitivity 71%
  - BWH: PPV 3.4% with Sensitivity 52.4%
- Echo model:
  - MGH: PPV 32.7% with Sensitivity 66.9%
  - BWH: PPV: 33.4% with Sensitivity 67%
- Combined:
  - MGH: PPV: 76.6% with Sensitivity 47.5%
  - BWH: PPV: 73.9% with Sensitivity 34.8%

Transformers

- Built on the attention mechanism
- Many modern tools with deep learning are based off the transformer architecture (ChatGPT, Claude, Vision Transformers, OpenAI Whisper)
- General purpose neural network that works well on images, speech, text….
- Resources:
  - Roger and Jimmy Ba's lecture notes
  - https://nlp.seas.harvard.edu/2018/04/03/attention.html
  - https://lilianweng.github.io/posts/2020-04-07-the-transformer-family/

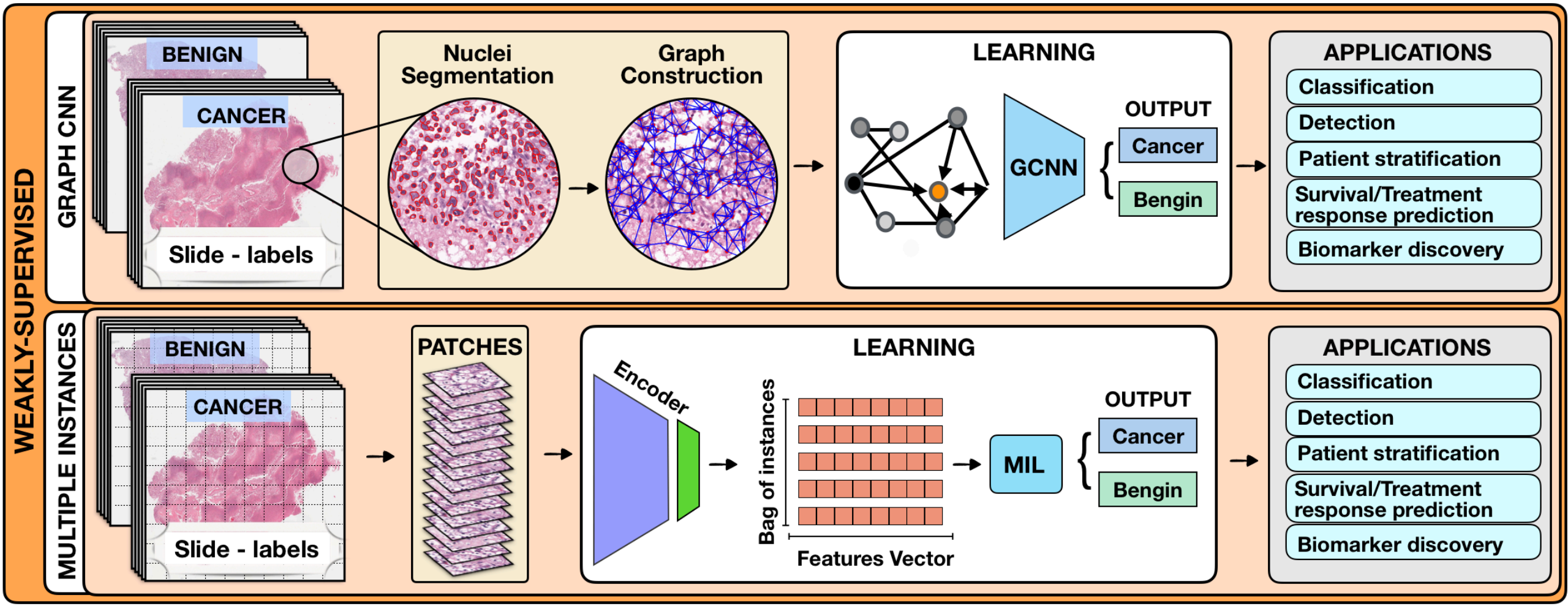
Case study 2: Deep learning for histopathological image data

- Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning, Chen et. al, CVPR 2022

Histopathological images in the clinical workflow

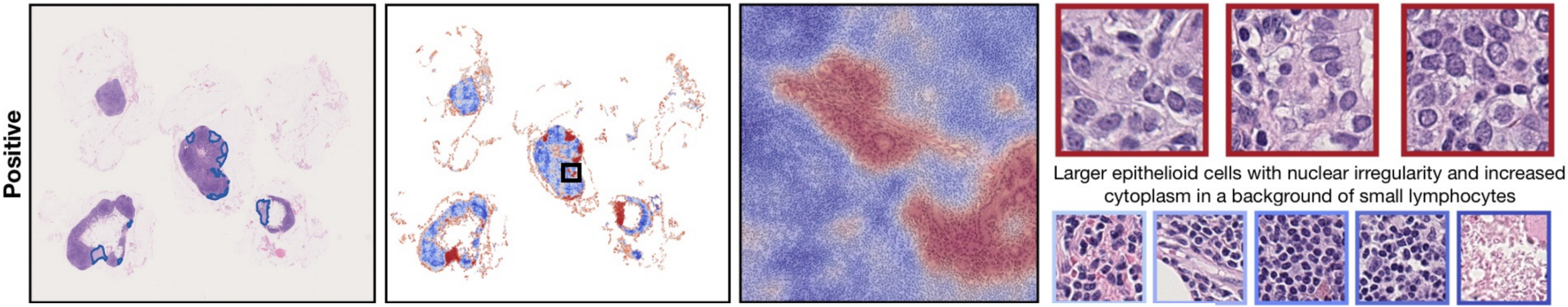
- Histopathology: Microscopic examination of tissue to study diseases and their different presentations,
- Pipeline:
  - Surgery, biopsy or autopsy for excision of tissue
  - Placed in a fixative to stabilize tissue
  - Investigated under a microscope
- Histopathological images are routinely used for clinical diagnoses of cancer
  - **Key question: How can we use machine learning to build representations of histopathological image data?**

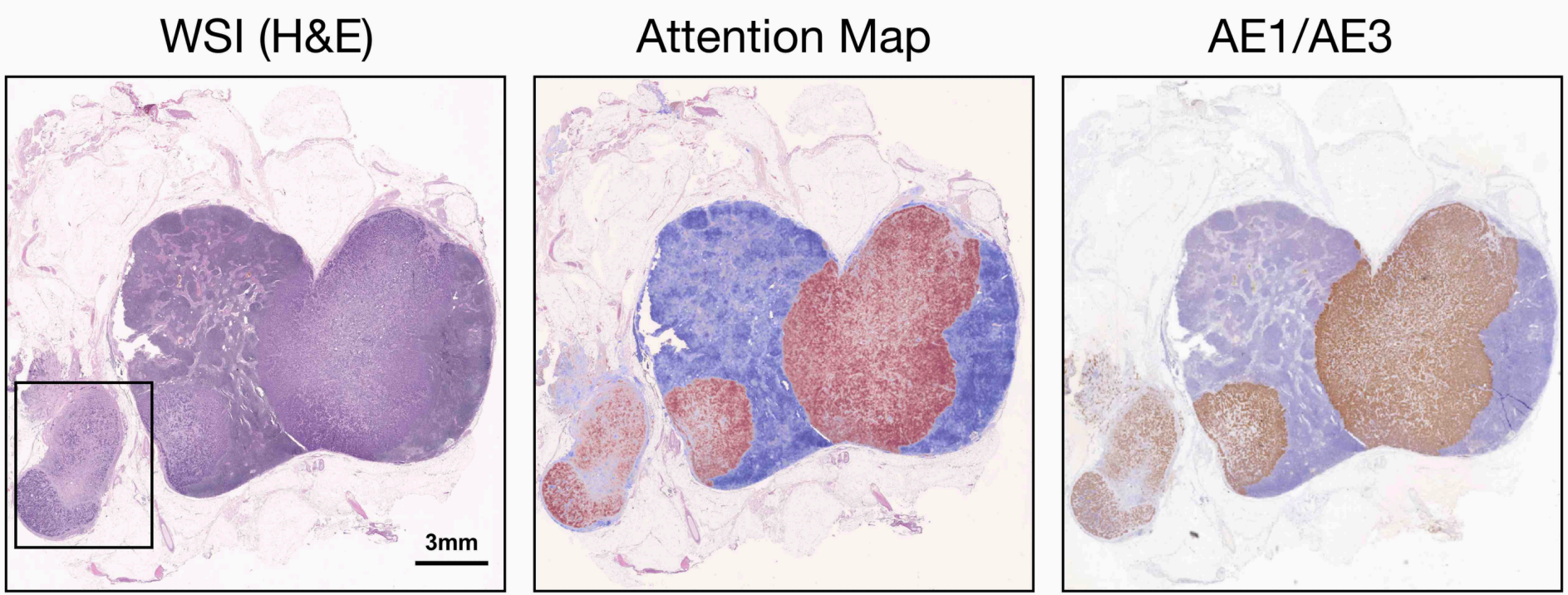
Slide-Level Supervised Learning (Weak Supervision)



Lipkova *et al.* 2021, In Review

Weakly-Supervised Learning: Finding Needles in Haystacks via Attention





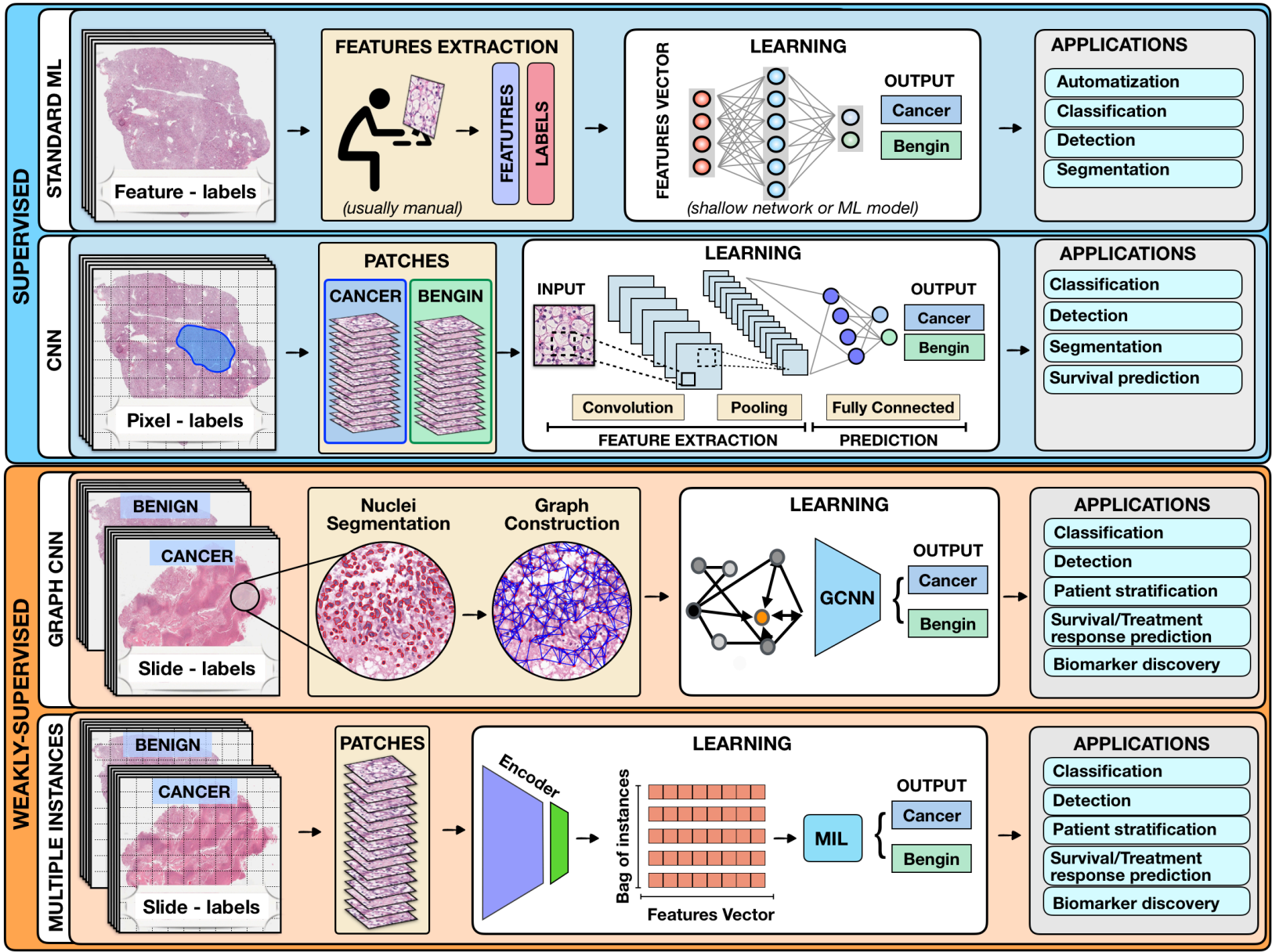
- Attention weights saliently localize tumor regions in binary classification tasks of benign / metastasis

Lu *et al.* 2021, Nature BME

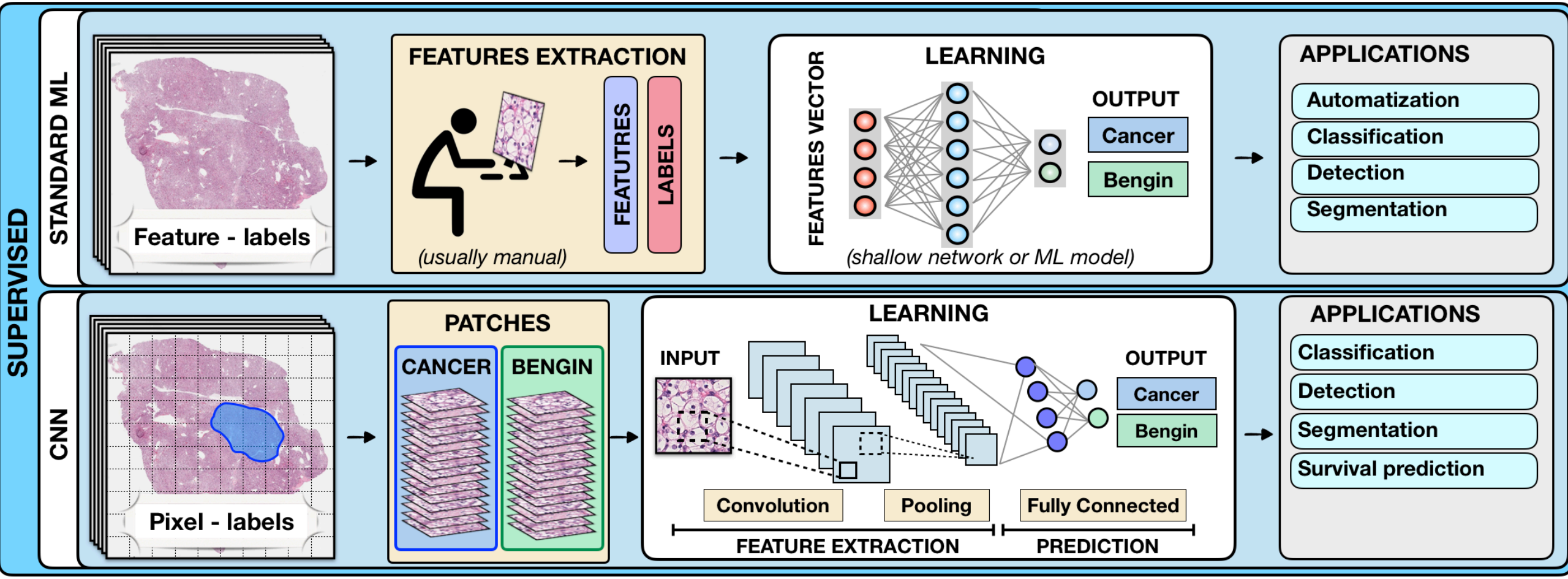
Current Paradigm is limited by: Clinical Domain Knowledge

- Requires clinical domain knowledge to:
  1. label image regions in WSIs with known morphological phenotypes (patch-level tasks)
  2. Make prognostic decisions from subjective interpretation of the entire WSI (slide-level tasks)

- How can we identify new phenotypic biomarkers?
- What are we missing in current decision-making that can guide prognosis?

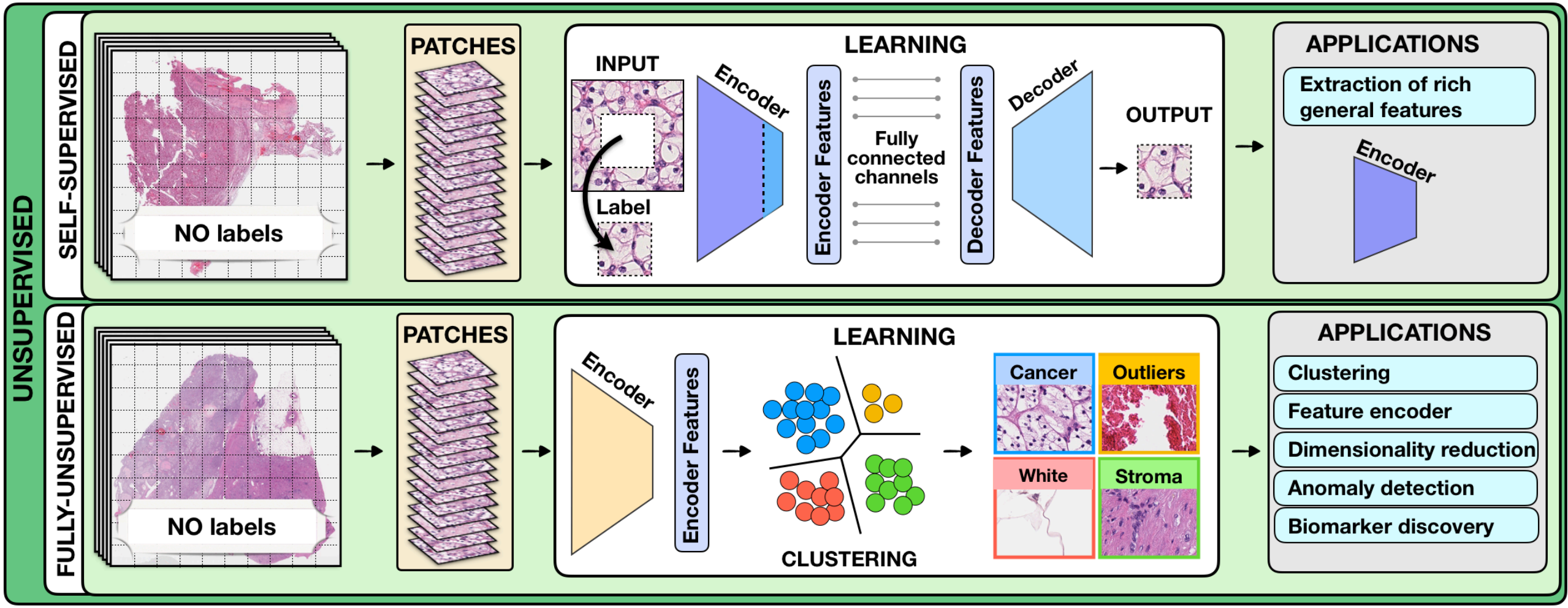


Current Paradigm is limited by: Clinical Domain Knowledge



Current pipelines for creating representations of whole slide images make use of ResNet50 architectures pretrained on imagenet.

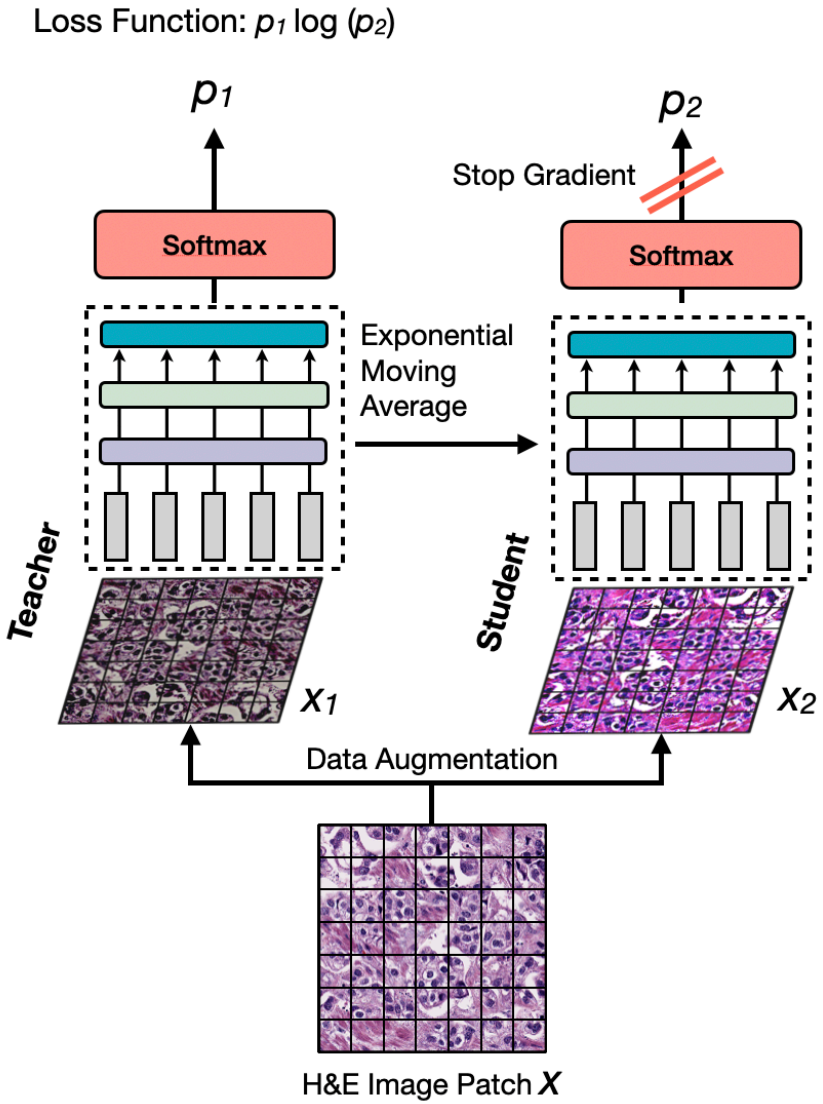
Self-Supervised Learning: Pixel-Level Annotations are Not Needed!



Lipkova *et al.* 2021, In Review, Ciga et. al

We build upon recent work [Resource and data efficient self supervised learning, Ciga et. al, 2021] who show that self-supervision yields general purpose representations of histopathological images

DINO-based Knowledge Distillation for Patch-based Representations



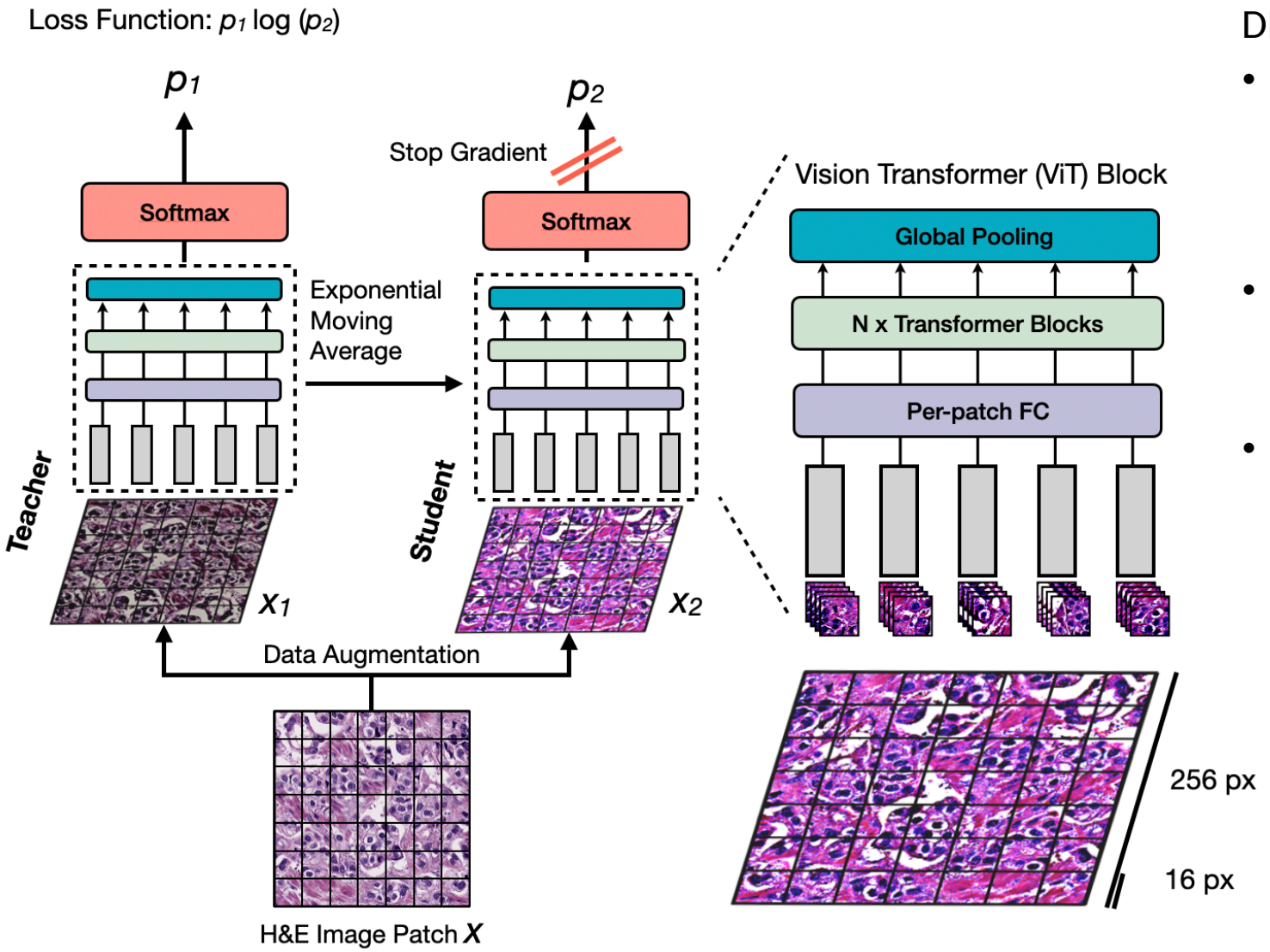
Chen *et al.* 2021, In Preparation

DINO

- We wanted to study the use of non-contrastive self-supervised learning for creating representations
- Input:
  - Two crops with color contrasts from the same image
- Goal of self-supervised learning:
  - Teach the network that these two crops are from the same image
  - Output of student network is trained to match the distribution of teacher network via minimizing cross-entropy loss
  - Avoid network collapse by having two networks
    - Train the student via gradient descent
    - Teacher is not trained, weights are updated via exponential moving average from students
- Does not require negative samples
  - Data inductive biases in natural images may not hold in H&E pathology slides

DINO: Emerging properties in self-supervised vision transformers, Caron et. Al,

DINO-based Knowledge Distillation for Patch-based Representations

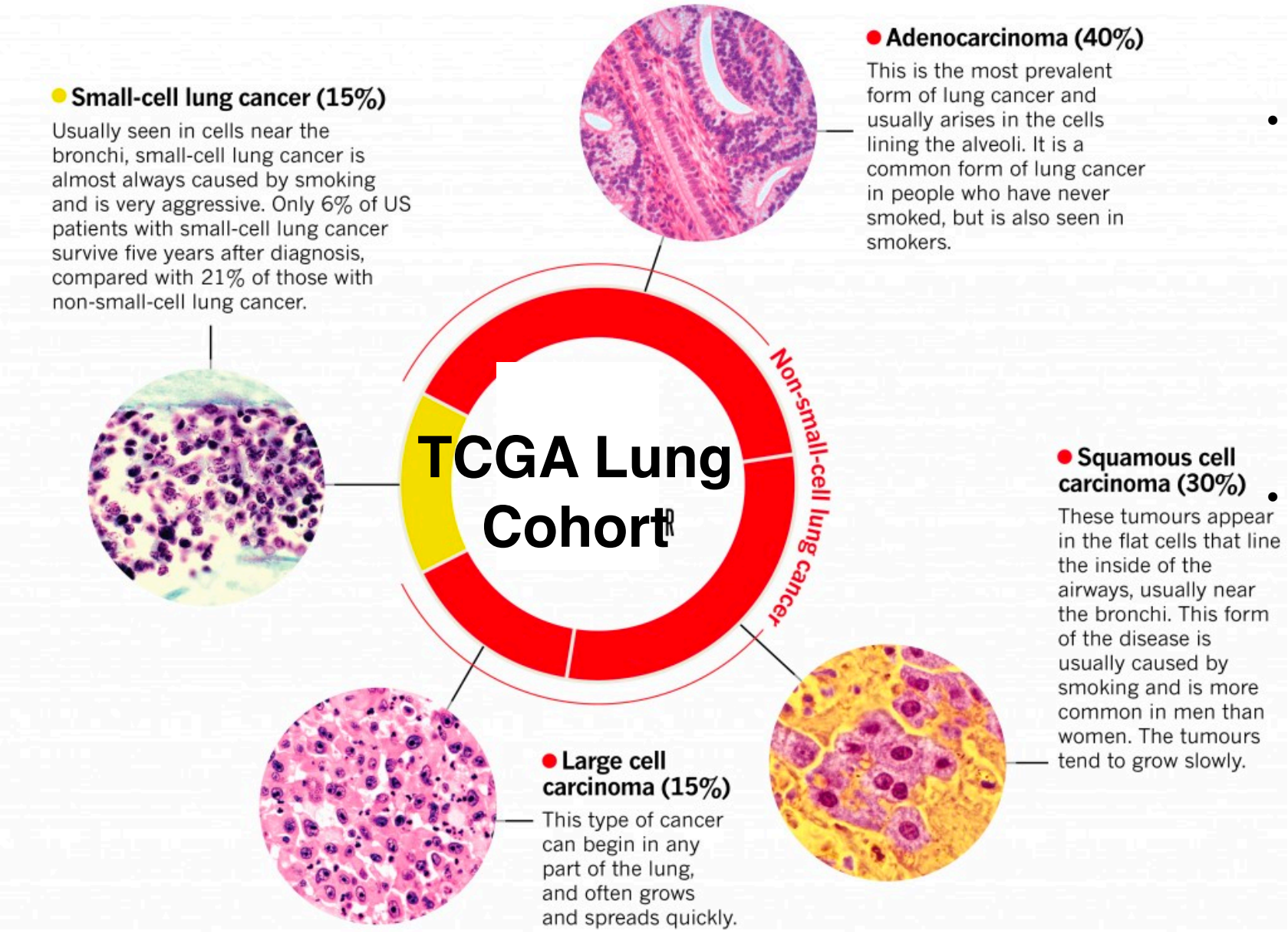


DINO

- Output of student network is trained to match the distribution of teacher network via:
  - minimizing cross-entropy loss
  - EMA to update teacher network
- Does not require negative samples
  - Data inductive biases in natural images may not hold in H&E pathology slides
- Vision Transformer (ViT) used as encoder
  - 256 x 256 H&E tissue patches are further patched as 16 x 16 patch embeddings

Chen *et al.* 2021, In Preparation

Study Design



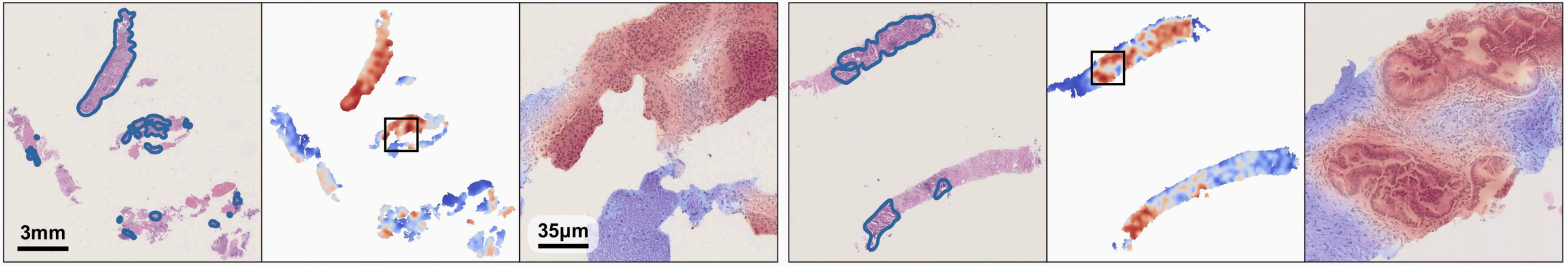
- Experiments:
  - Organ-specific vs. pan-cancer training
    - TCGA Lung (1033 WSIs) vs Entire TCGA (~8788 WSIs)
  - Comparisons with SOTA methods
    - SimCLR, SimSiam

- Slide-Level Tasks:
  - LUAD vs. LUSC Subtyping
  - LUAD + LUSC Survival Analysis
  - TP53 + KRAS Mutation Prediction

Results [1]: TCGA Lung Subtyping (LUAD vs. LUSC)

Lung Adenocarcinoma (LUAD)

Lung Squamous Cell Carcinoma (LUSC)



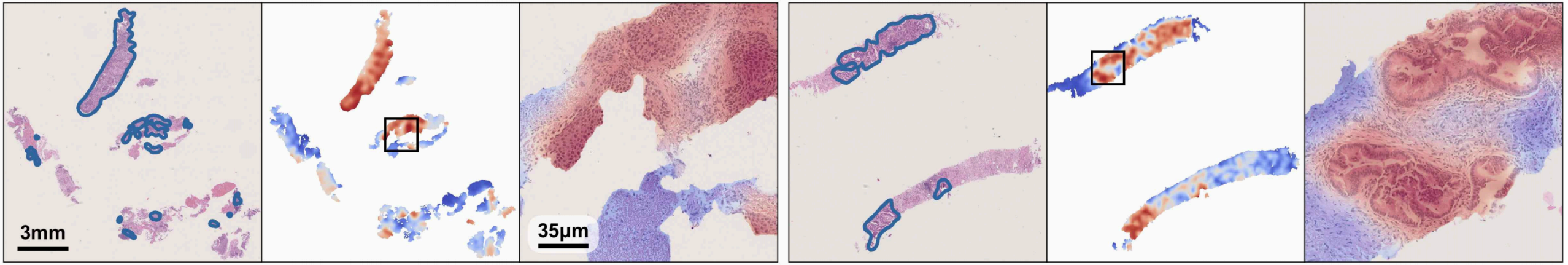
Method	Model Architecture	Training Source	Epochs	100%	75%	50%	25%
ImageNet Transfer	ResNet-50	ImageNet	100	0.945 ± 0.018	0.943 ± 0.019	0.917 ± 0.024	0.888 ± 0.031
SimCLR	ResNet-50	Lung Only	100	0.950 ± 0.026	0.947 ± 0.017	0.934 ± 0.025	0.897 ± 0.028
SimSiam	ResNet-50	Lung Only	100	0.952 ± 0.017	0.944 ± 0.018	0.935 ± 0.026	0.897 ± 0.029
DINO	ViT	Lung Only	100	0.948 ± 0.021	0.942 ± 0.019	0.937 ± 0.021	0.928 ± 0.024
SimCLR	ResNet-50	Pan-Cancer	100	0.951 ± 0.016	0.948 ± 0.017	0.930 ± 0.023	0.898 ± 0.026
SimSiam	ResNet-50	Pan-Cancer	100	0.493 ± 0.085	0.534 ± 0.072	0.508 ± 0.085	0.603 ± 0.040
DINO	ViT	Pan-Cancer	100	**0.957 ± 0.019**	0.949 ± 0.019	**0.941 ± 0.022**	**0.931 ± 0.024**

- Self-supervised feature extractors from DINO are more data-efficient than pretrained ResNet-50 on ImageNet for subtyping

Results [2]: TCGA Lung Subtyping (LUAD vs. LUSC) + Mutation Prediction (TP53 + KRAS)

Lung Adenocarcinoma (LUAD)

Lung Squamous Cell Carcinoma (LUSC)



Method	Model Architecture	Training Source	Epochs	100%	75%	50%	25%	TP53	KRAS
ImageNet Transfer	ResNet-50	ImageNet	100	0.945 ± 0.018	0.943 ± 0.019	0.917 ± 0.024	0.888 ± 0.031	**0.756 ± 0.053**	0.761 ± 0.073
SimCLR	ResNet-50	Lung Only	100	0.950 ± 0.026	0.947 ± 0.017	0.934 ± 0.025	0.897 ± 0.028	0.694 ± 0.073	0.737 ± 0.044
SimSiam	ResNet-50	Lung Only	100	0.952 ± 0.017	0.944 ± 0.018	0.935 ± 0.026	0.897 ± 0.029	0.698 ± 0.084	0.681 ± 0.117
DINO	ViT	Lung Only	100	0.948 ± 0.021	0.942 ± 0.019	0.937 ± 0.021	0.928 ± 0.024	0.751 ± 0.041	**0.771 ± 0.059**
SimCLR	ResNet-50	Pan-Cancer	100	0.951 ± 0.016	0.948 ± 0.017	0.930 ± 0.023	0.898 ± 0.026	0.687 ± 0.100	0.711 ± 0.127
SimSiam	ResNet-50	Pan-Cancer	100	0.493 ± 0.085	0.534 ± 0.072	0.508 ± 0.085	0.603 ± 0.040	0.516 ± 0.073	0.612 ± 0.051
DINO	ViT	Pan-Cancer	100	**0.957 ± 0.019**	0.949 ± 0.019	**0.941 ± 0.022**	**0.931 ± 0.024**	0.746 ± 0.051	0.740 ± 0.052

- Self-supervised feature extractors from DINO are more data-efficient than pretrained ResNet-50 on ImageNet for subtyping
- No difference found in gene mutation prediction

Results [3]: DINO Attentions to Cellular Identities

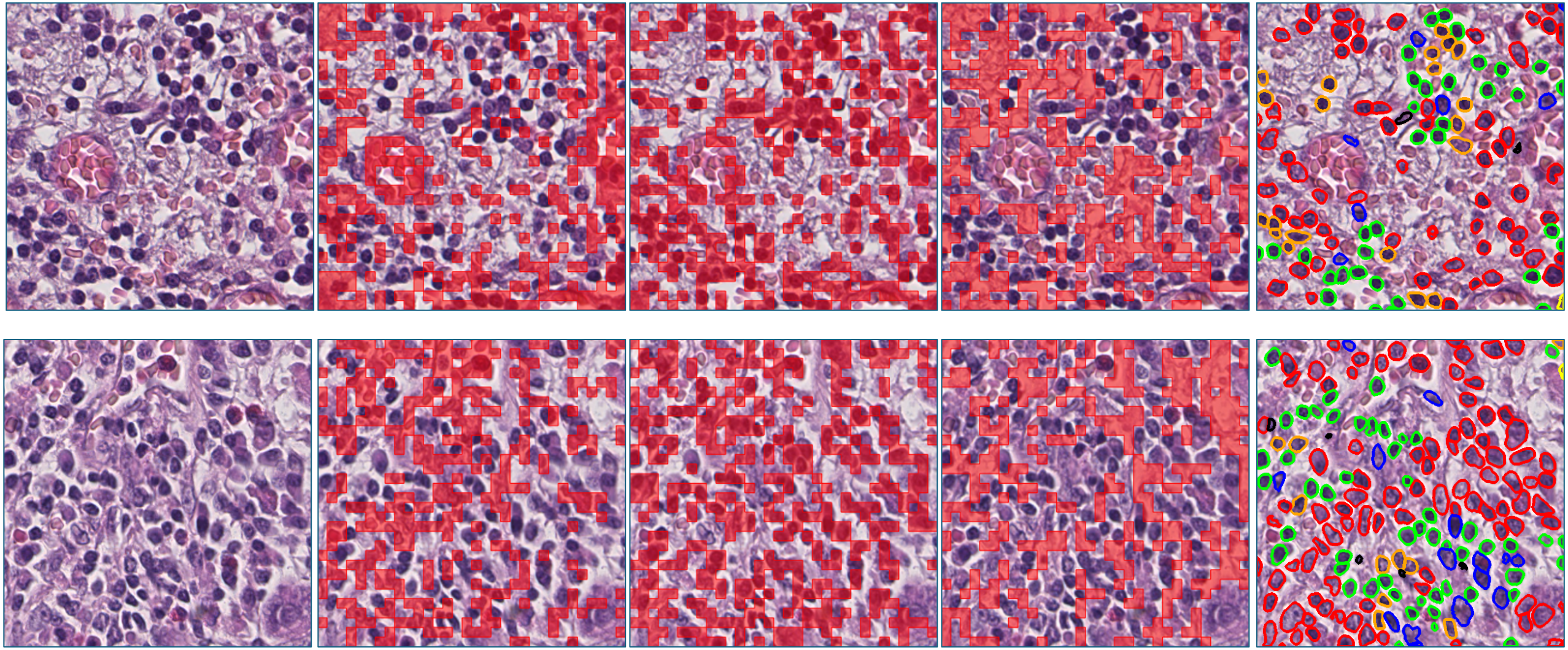
H&E Image

Attention Head #1: Red Blood Cells

Attention Head #2: Cell Location

Attention Head #3: Cytoplasm

Ground Truth Cell Segmentation



Questions?