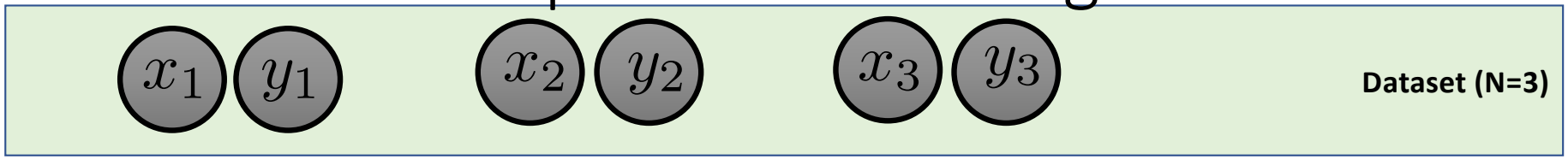
Topics in Machine Learning
Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Outline

- Machine learning thus far
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning
- Self-supervised learning
- Case study : Histopathology

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)

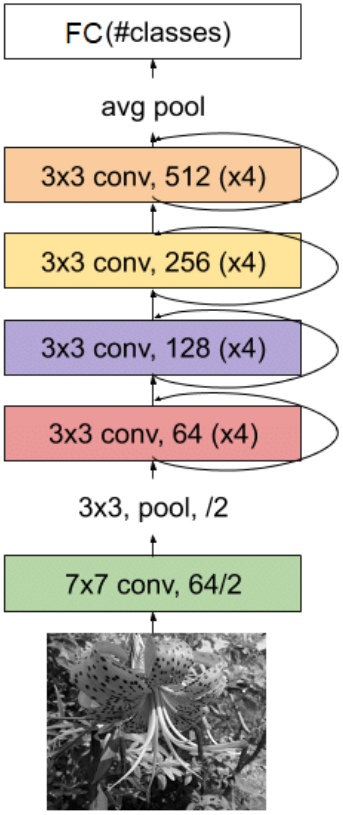
$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, x_i)$$

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

Deep residual neural networks



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

They added a skip-connections between layers:

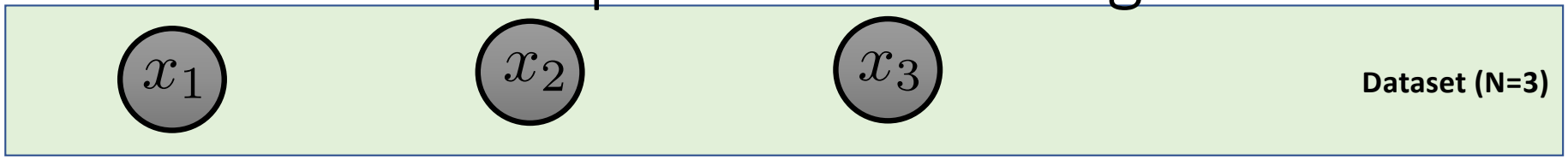
$$h_k = \phi(conv(h_{k-1})) + \sum_{j<k-1} h_{k-j}$$

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



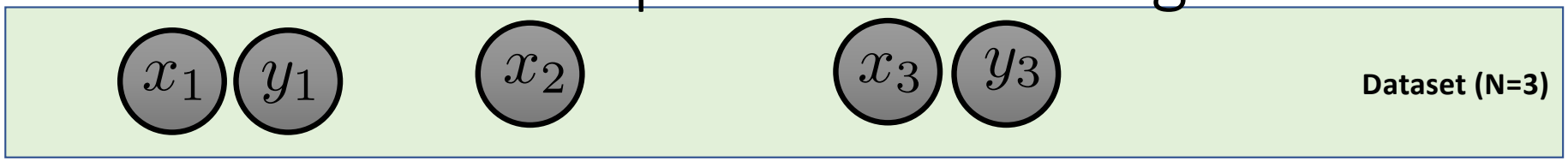
$$\mathcal{L}(x) = \log p(x;\theta)$$

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

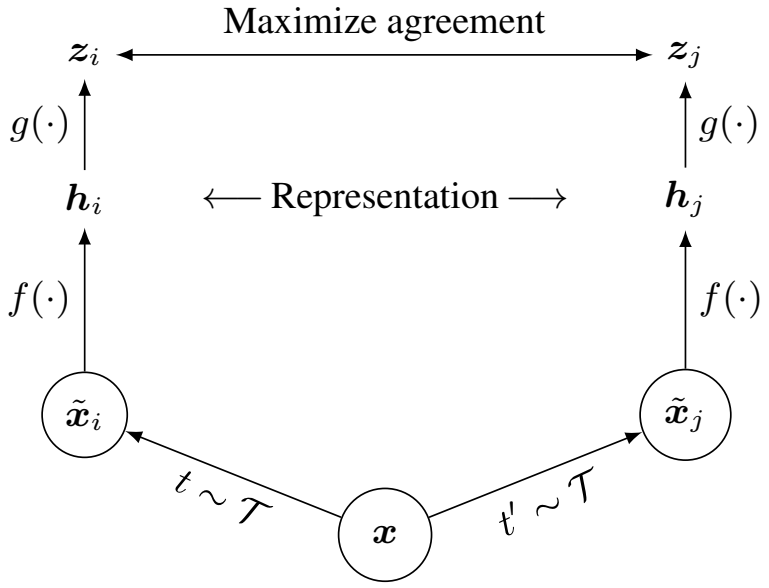
$\mathrm{sim}(k, k')$

- 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.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} , \qquad (1)$$

How good are the representations?

A Simple Framework for Contrastive Learning of Visual Representations

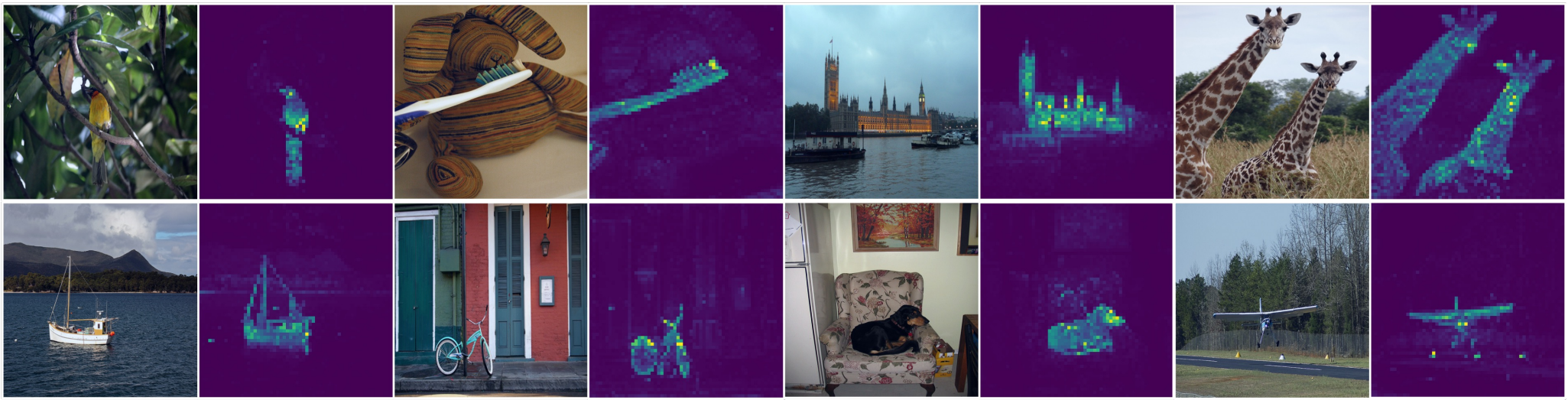
	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**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 1 - 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



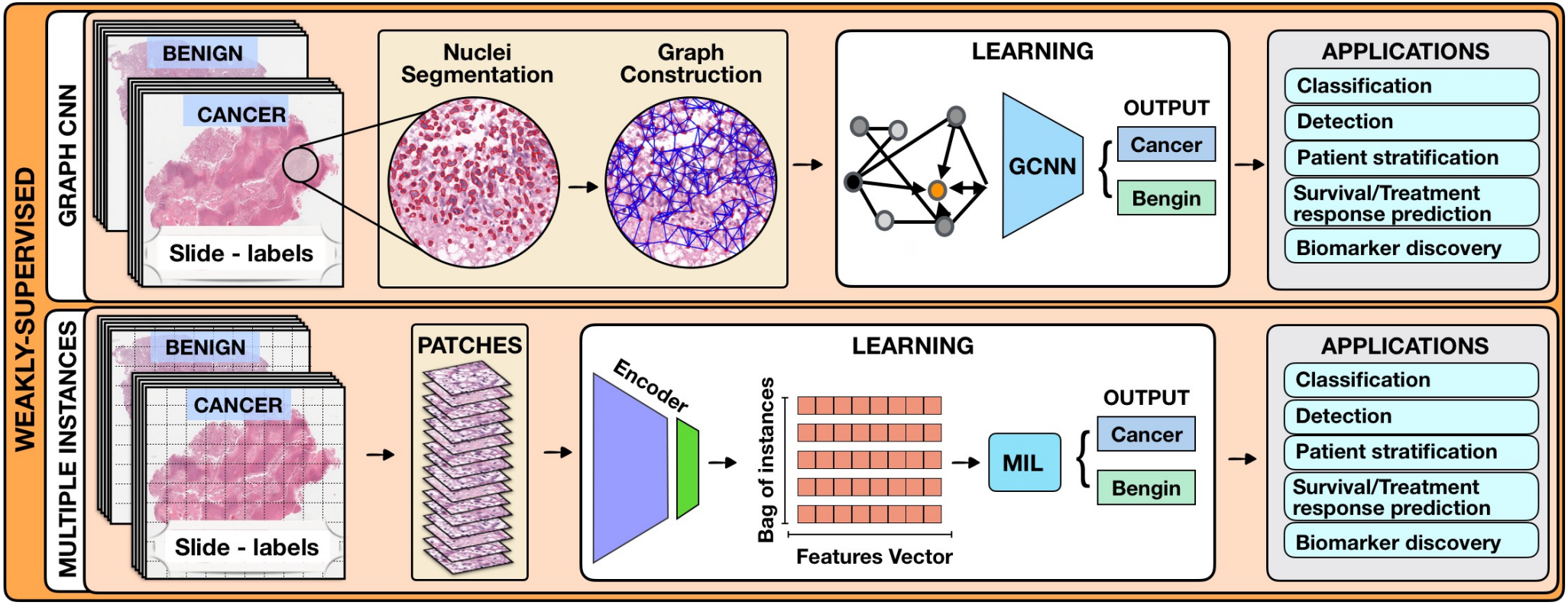
Case study : Deep learning for histopathological image data

- Research by Richard J. Chen
- 3rd year Ph.D. Candidate, Harvard University / BWH, Broad Institute

Histopathological images in the clinical workflow

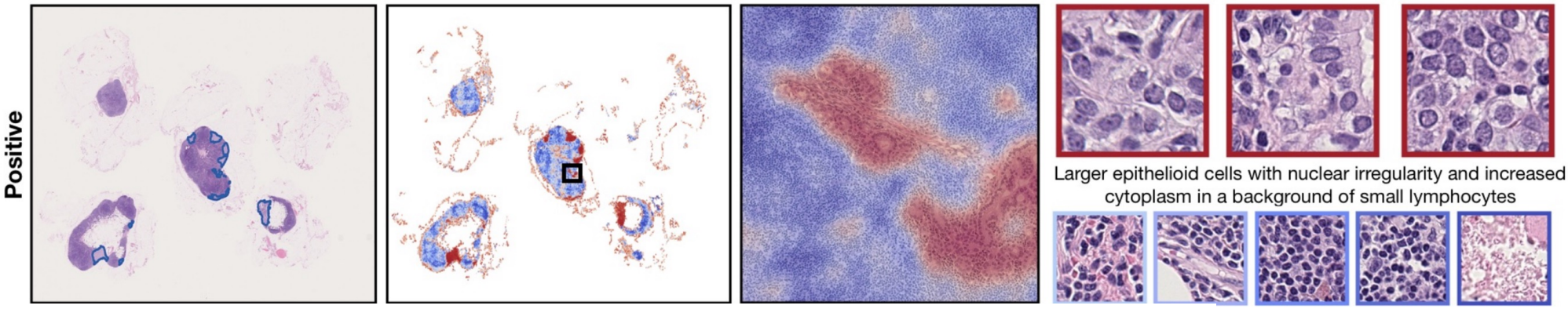
- Histopathology: Microscopic examination of tissue to study diseases and their different presentations,
- Pipeline:
  - Surgery, biopsy or autopsy for excision of tissue
  - Placed in a fixative to stabilize tissue
  - Investigated under a microscope
- Histopathological images are routinely used for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

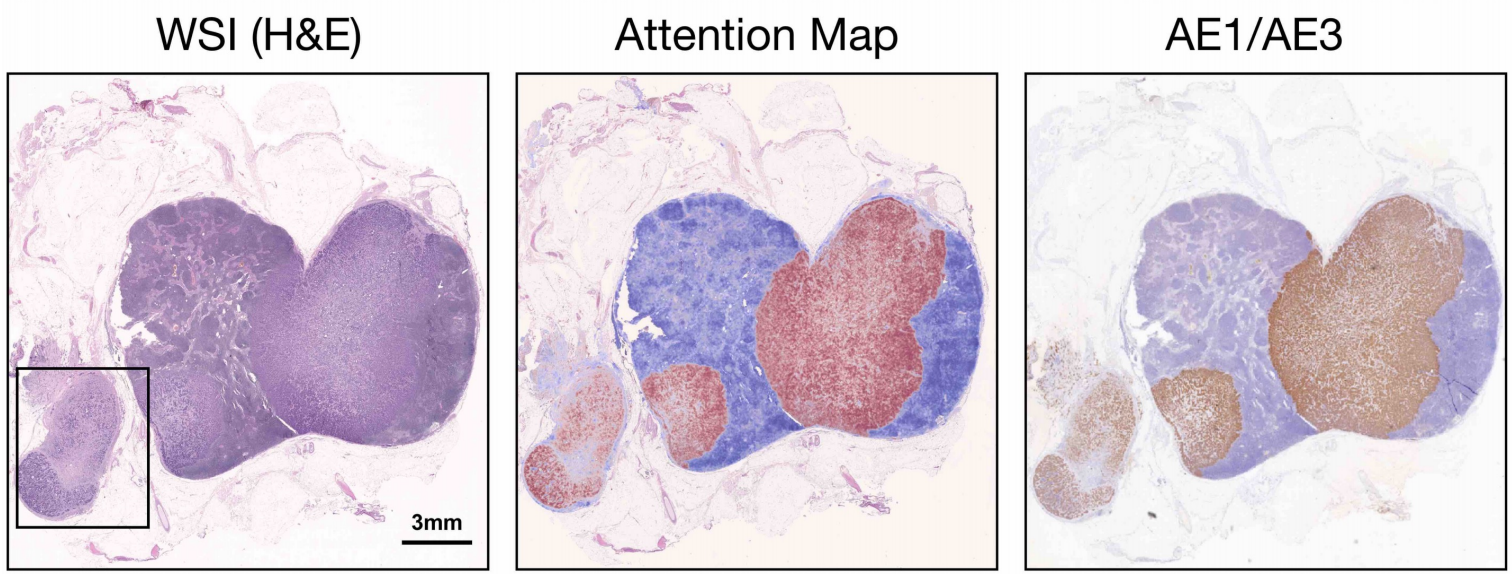
Slide-Level Supervised Learning (Weak Supervision)



Lipkova *et al.* 2021, In Review

Weakly-Supervised Learning: Finding Needles in Haystacks via Attention





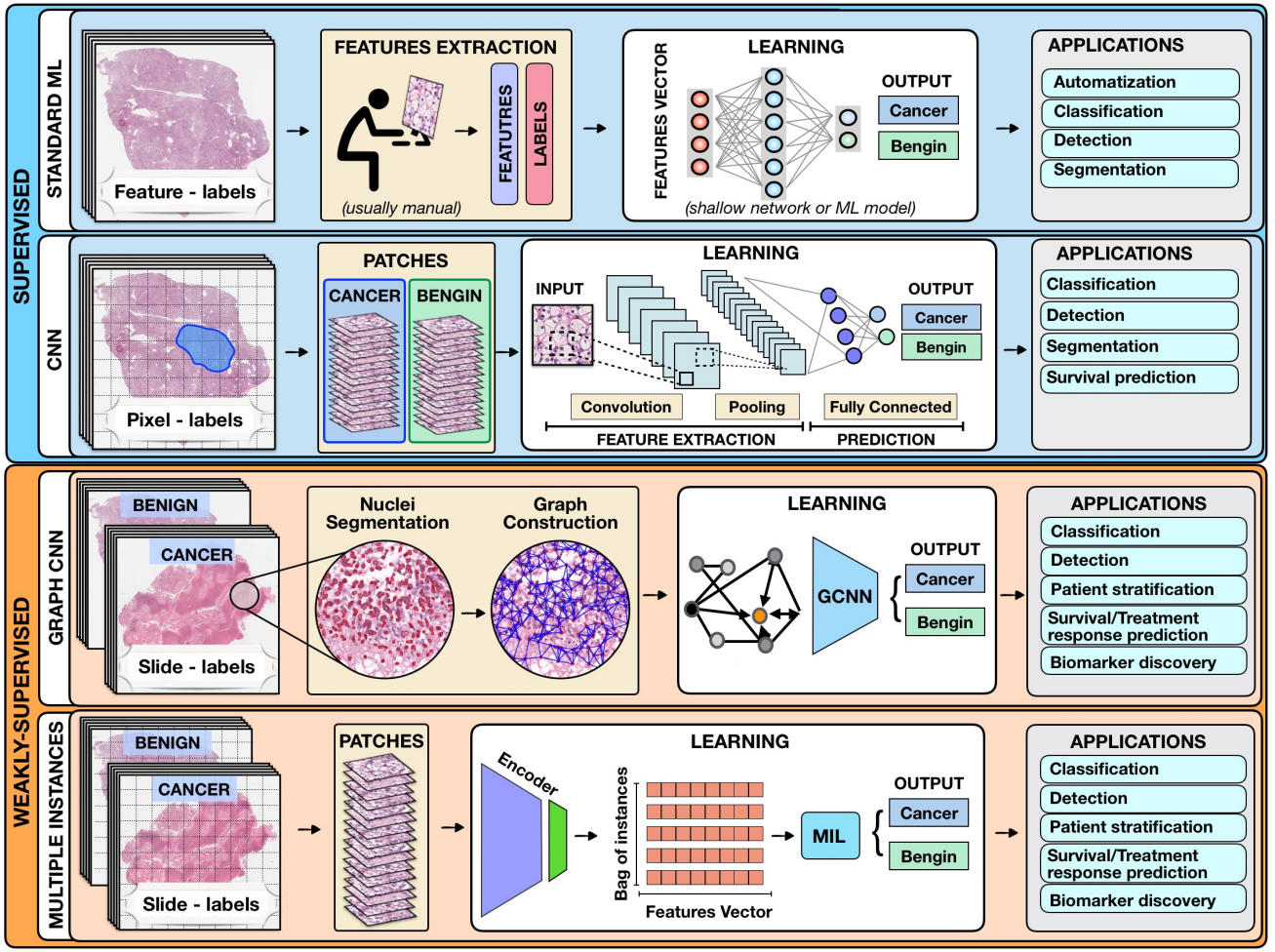
- Attention weights **saliently localize tumor regions** in binary classification tasks of benign / metastasis

Lu *et al.* 2021, Nature BME

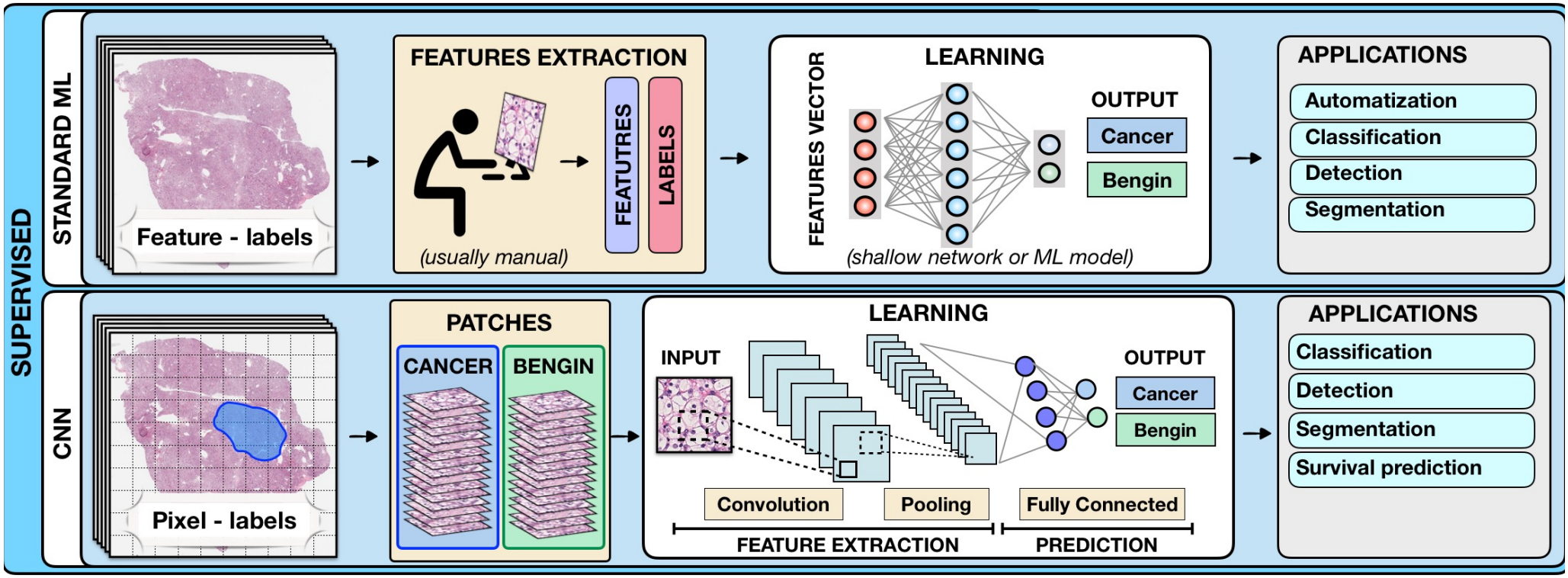
Current Paradigm is limited by: Clinical Domain Knowledge

- 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?

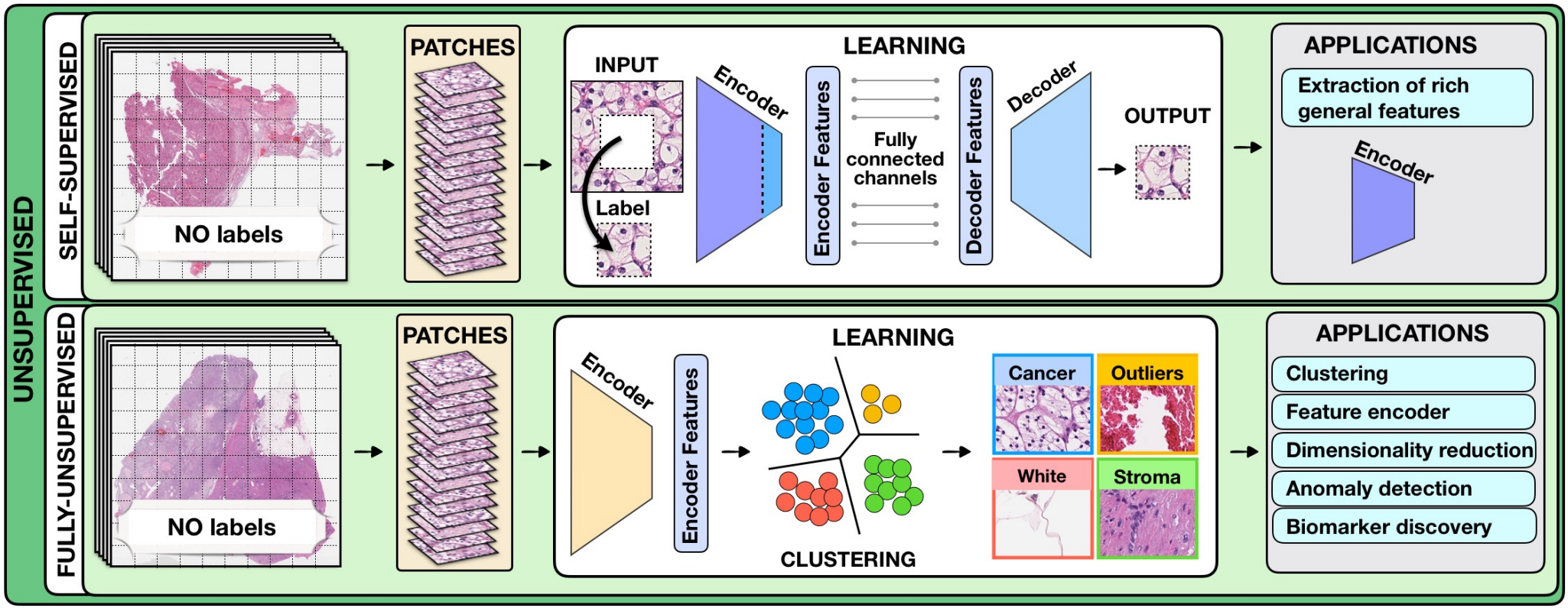


Current Paradigm is limited by: Clinical Domain Knowledge



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

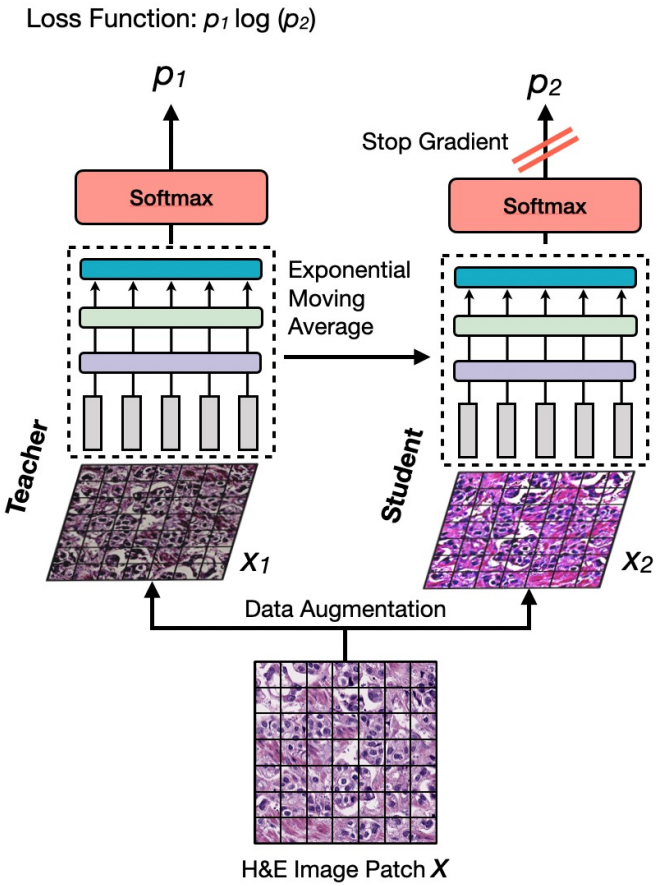
Self-Supervised Learning: Pixel-Level Annotations are Not Needed!



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



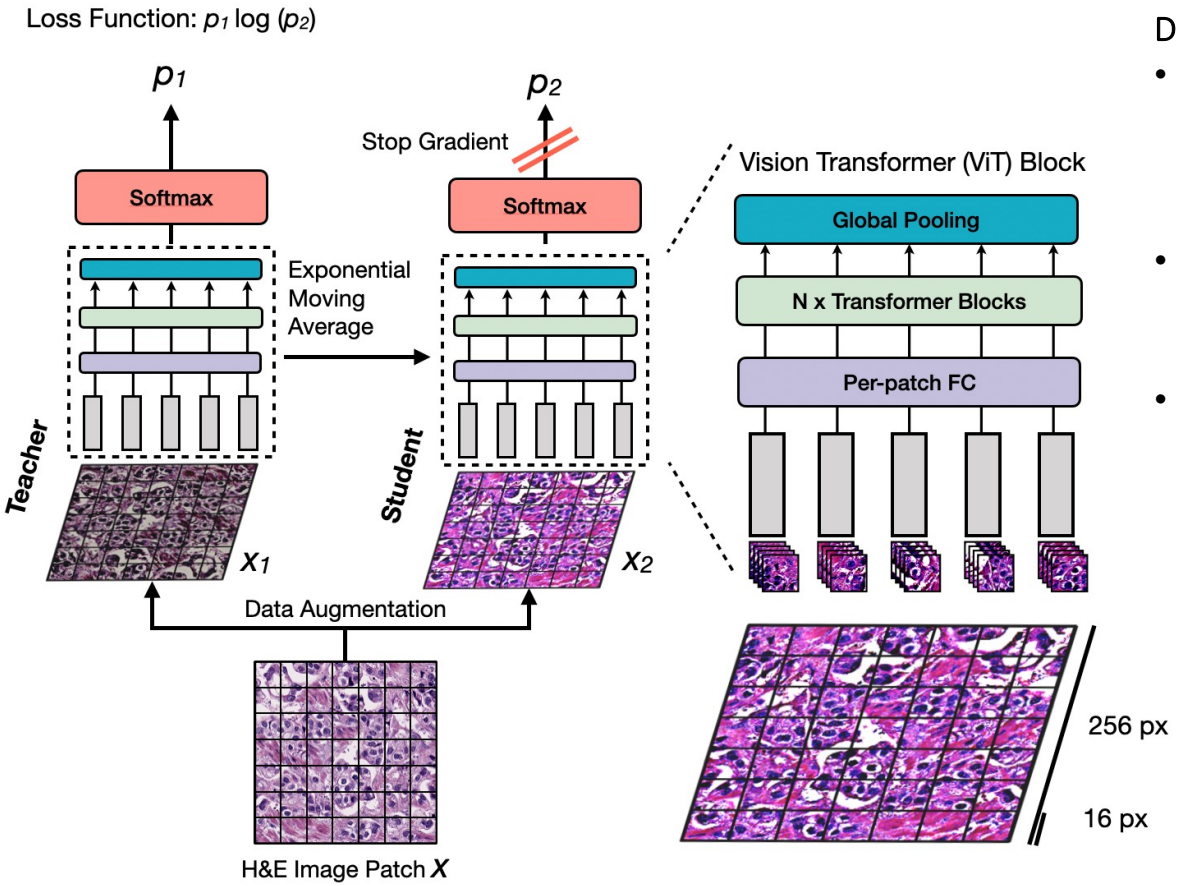
Chen *et al.* 2021, In Preparation

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, 2021

DINO-based Knowledge Distillation for Patch-based Representations

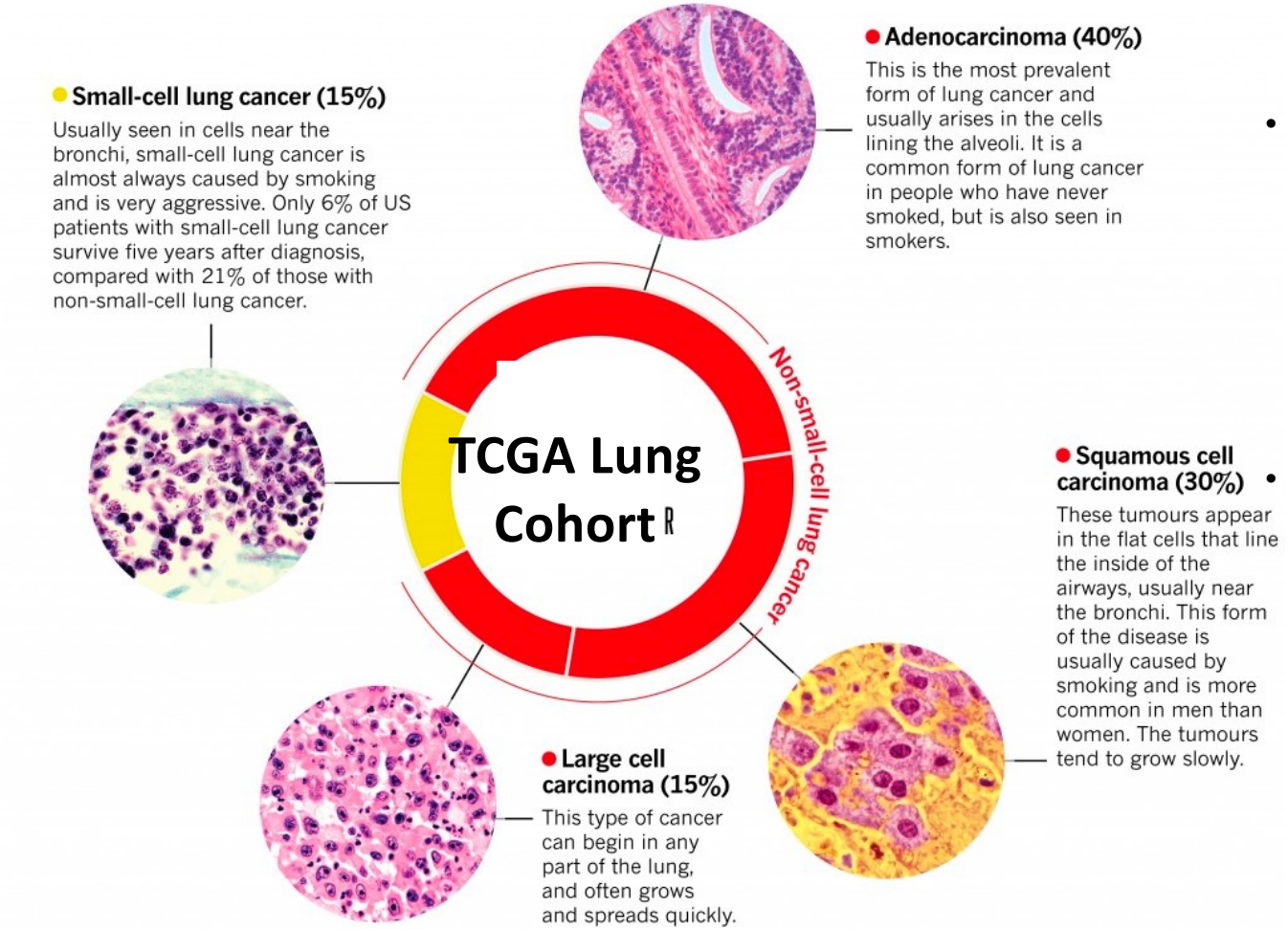


Chen *et al.* 2021, In Preparation

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

Study Design

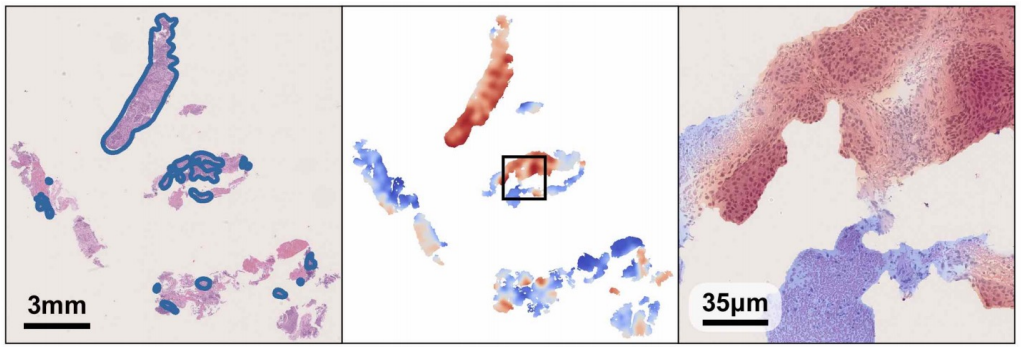


- Experiments:
  - Organ-specific vs. pan-cancer training
    - TCGA Lung (1033 WSIs) vs Entire TCGA (~8788 WSIs)
  - Comparisons with SOTA methods
    - SimCLR, SimSiam

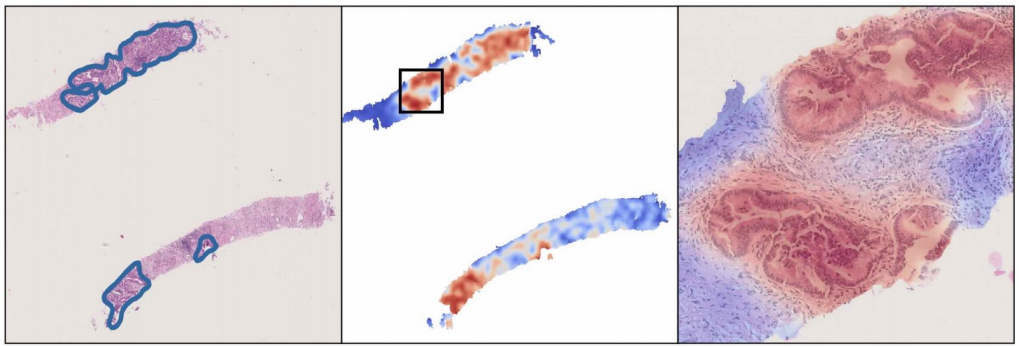
- Slide-Level Tasks:
  - LUAD vs. LUSC Subtyping
  - LUAD + LUSC Survival Analysis
  - TP53 + KRAS Mutation Prediction

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

Lung Adenocarcinoma (LUAD)



Lung Squamous Cell Carcinoma (LUSC)

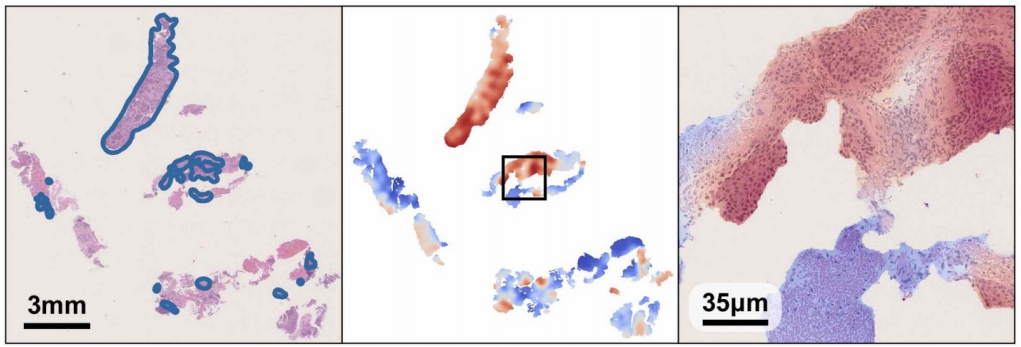


Method	Model Architecture	Training Source	Epochs	100%	75%	50%	25%
ImageNet Transfer	ResNet-50	ImageNet	100	0.945 ± 0.018	0.943 ± 0.019	0.917 ± 0.024	0.888 ± 0.031
SimCLR	ResNet-50	Lung Only	100	0.950 ± 0.026	0.947 ± 0.017	0.934 ± 0.025	0.897 ± 0.028
SimSiam	ResNet-50	Lung Only	100	0.952 ± 0.017	0.944 ± 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**

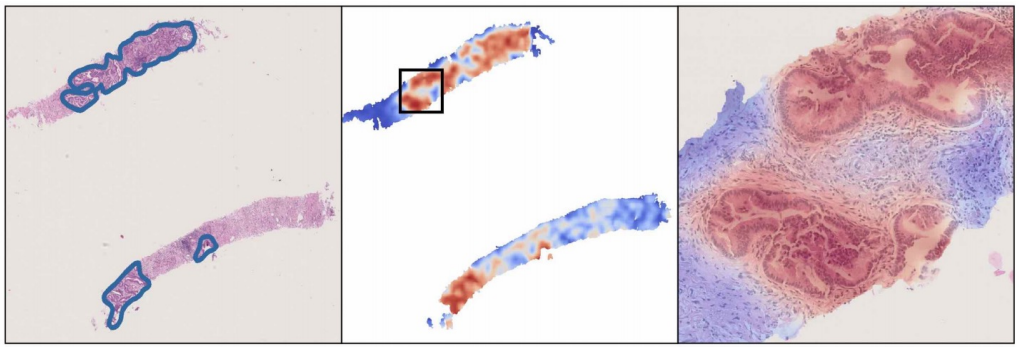
- Self-supervised feature extractors from DINO are more data-efficient than pretrained ResNet-50 on ImageNet for subtyping

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

Lung Adenocarcinoma (LUAD)



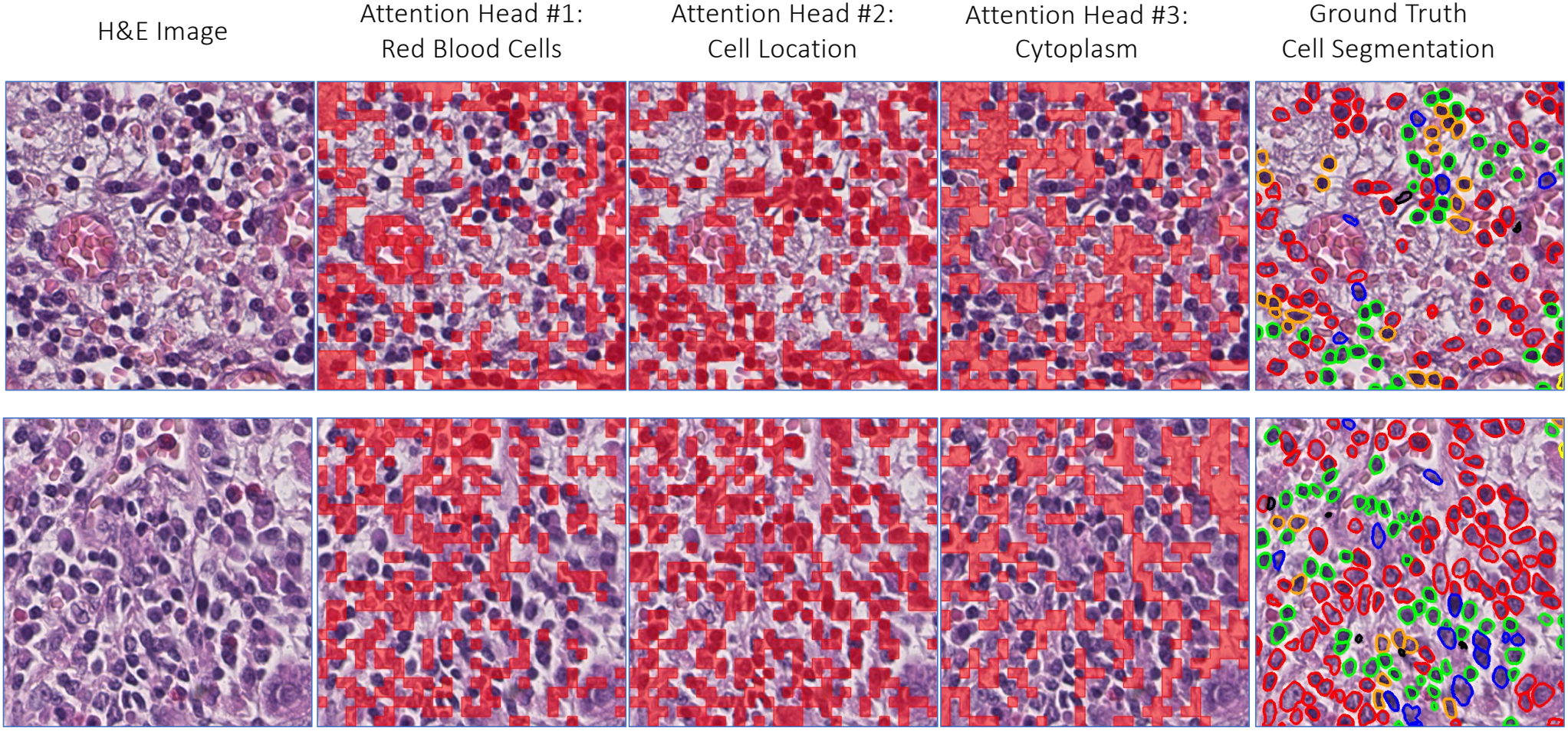
Lung Squamous Cell Carcinoma (LUSC)



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



Questions?