Medical Image Analysis with CNNs

June 8, 2019

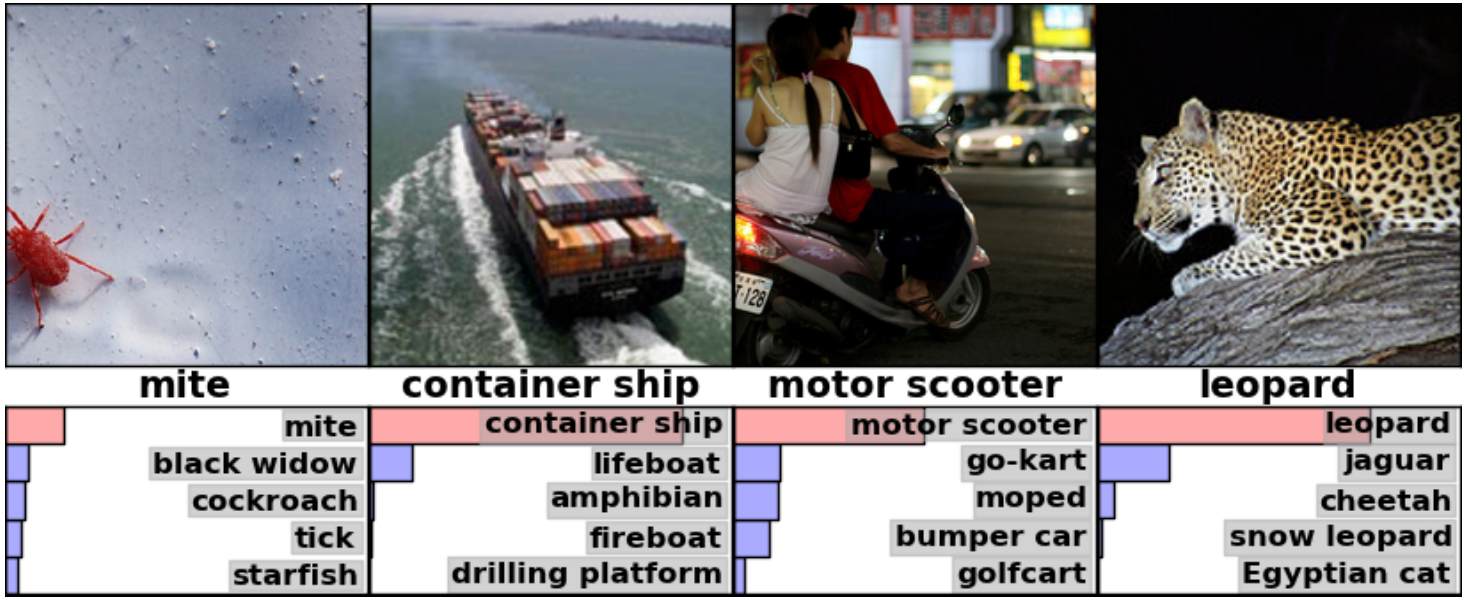
Block 2, Lecture 2
Applied Data Science
MMCi Term 4, 2019

Matthew Engelhard

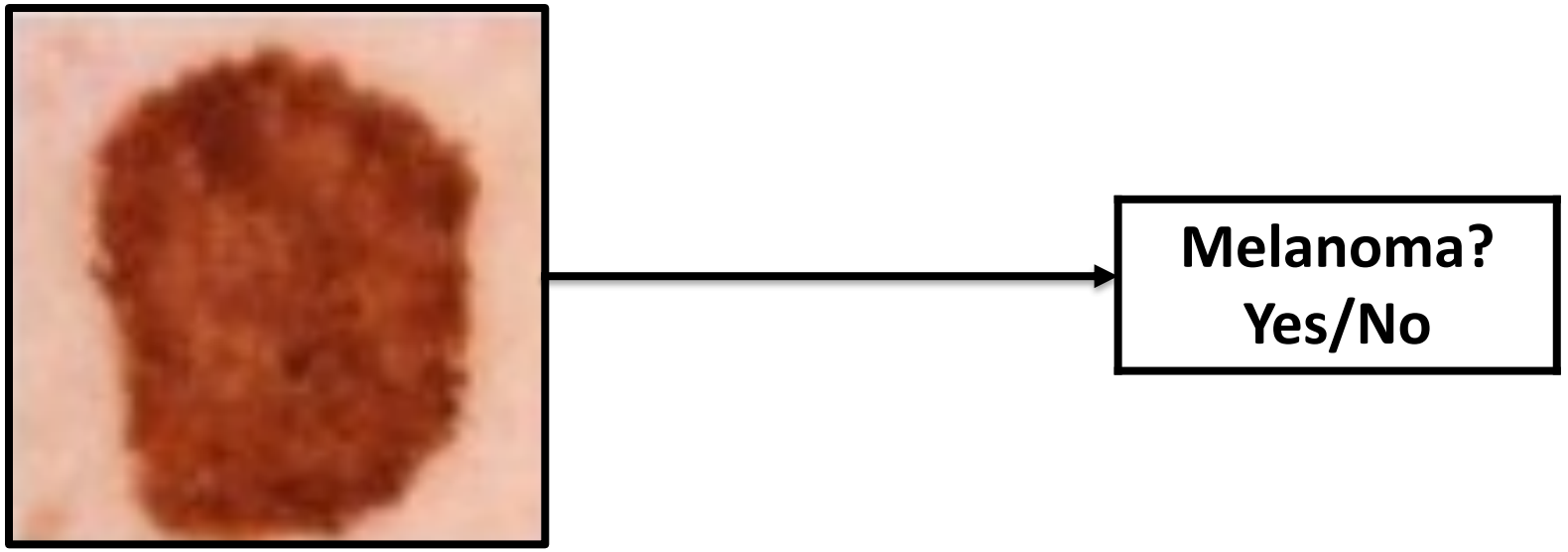
Identifying Diabetic Retinopathy and Skin Cancer

CLASSIFICATION

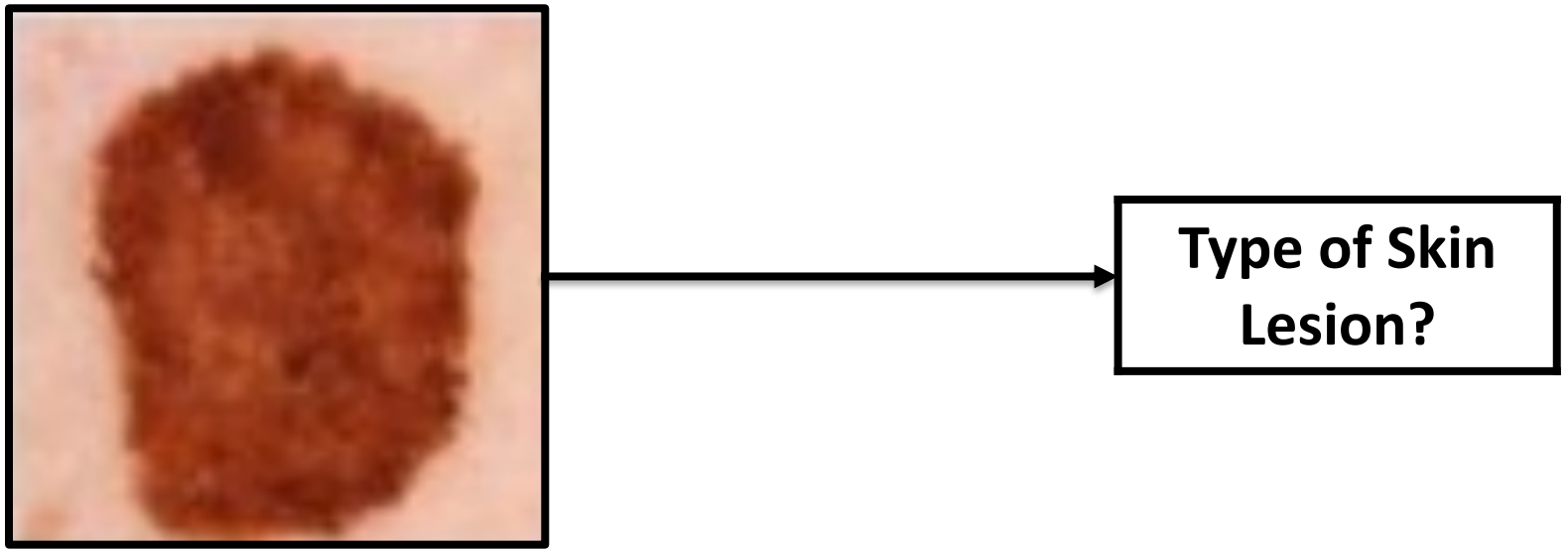
Classification:
predict the label associated with each image



Classifier Output: Two-Class (e.g. Yes/No)



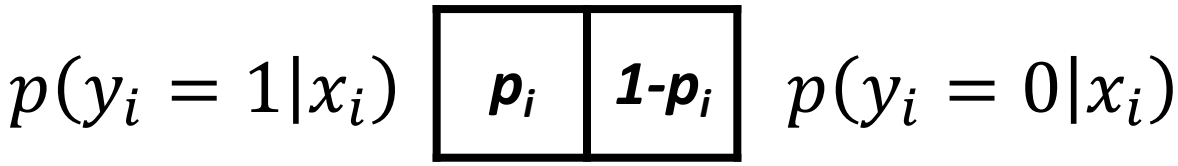
Classifier Output: Multi-Class (e.g. Lesion Type)



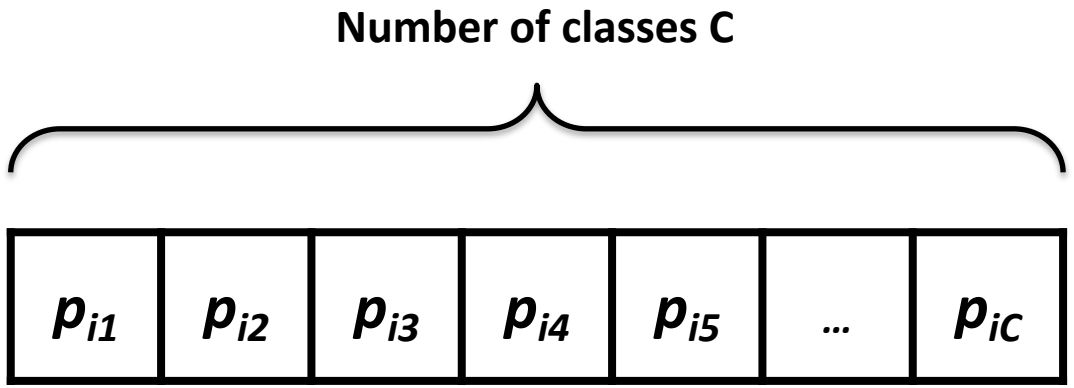
Classifier Output: Two-Class (e.g. Yes/No)

$$p(y_i = 1|x_i)$$

Classifier Output: Two-Class (e.g. Yes/No)



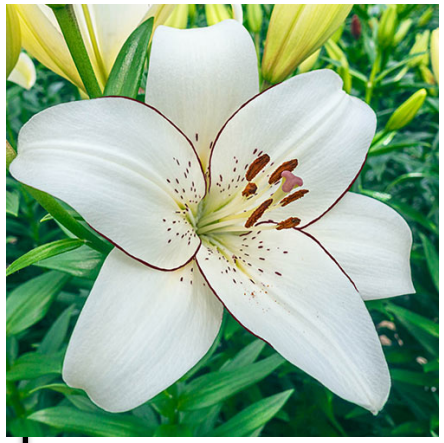
Classifier Output: Multi-Class (e.g. Lesion Type)

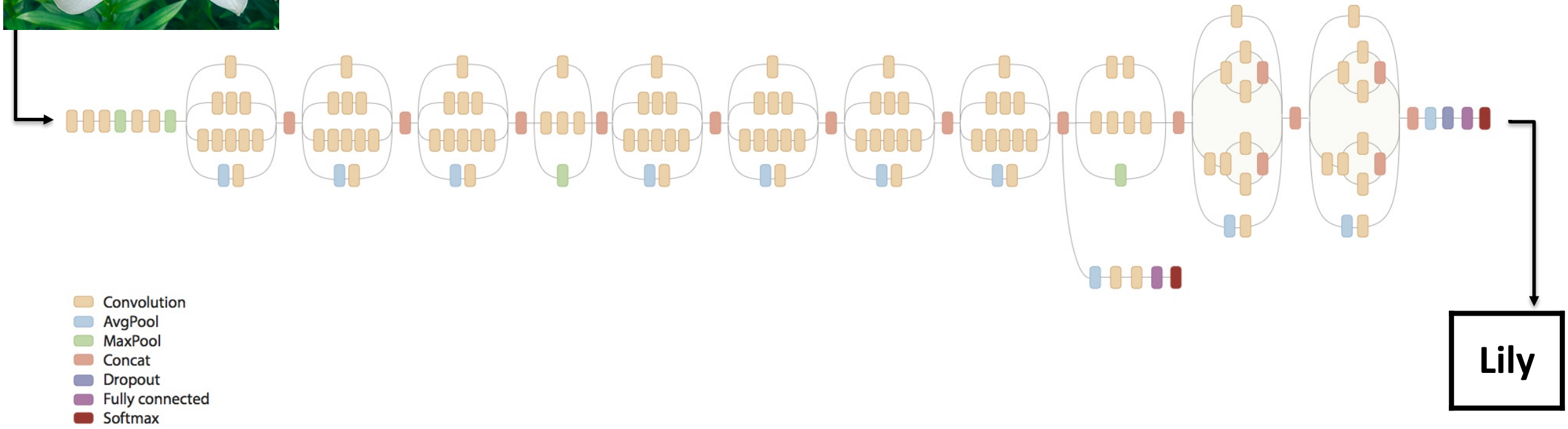


Values in all boxes sum to 1

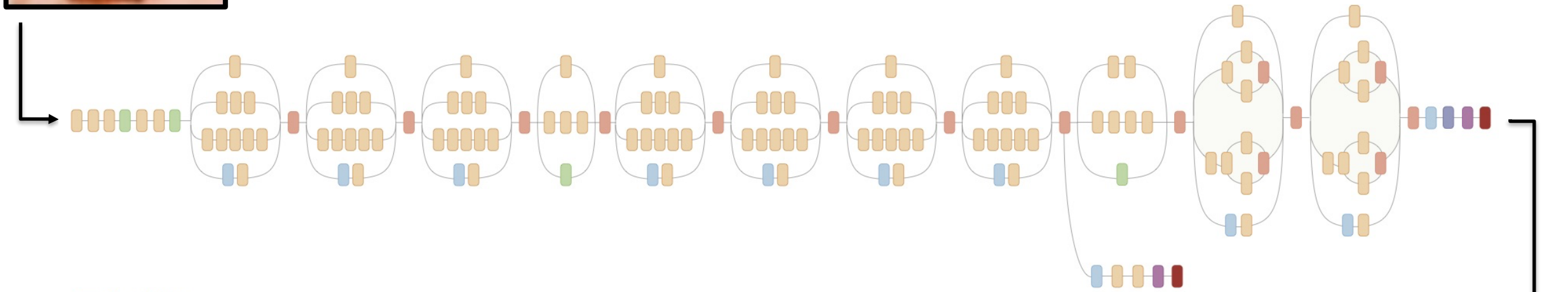
$$p_{ij} = p(y_i = j|x_i)$$

Take a model trained on naturalistic images...



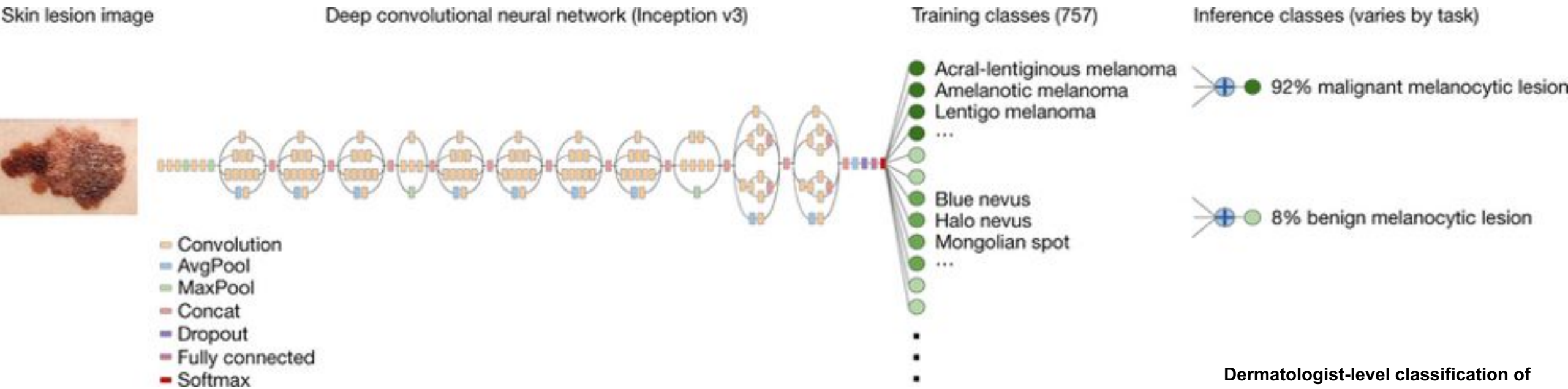


…and fine-tune it to evaluate medical images



Melanoma

…and fine-tune it to evaluate medical images



Dermatologist-level classification of skin cancer with deep neural networks
Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S
Nature, 2017 Feb;542(7639):115

Inception v3 and many other models are freely available

Pre-trained Models

Neural nets work best when they have many parameters, making them powerful function approximators. However, this means they must be trained on very large datasets. Because training models from scratch can be a very computationally intensive process requiring days or even weeks, we provide various pre-trained models, as listed below. These CNNs have been trained on the ILSVRC-2012-CLS image classification dataset.

In the table below, we list each model, the corresponding TensorFlow model file, the link to the model checkpoint, and the top 1 and top 5 accuracy (on the imagenet test set). Note that the VGG and ResNet V1 parameters have been converted from their original caffe formats (here and here), whereas the Inception and ResNet V2 parameters have been trained internally at Google. Also be aware that these accuracies were computed by evaluating using a single image crop. Some academic papers report higher accuracy by using multiple crops at multiple scales.

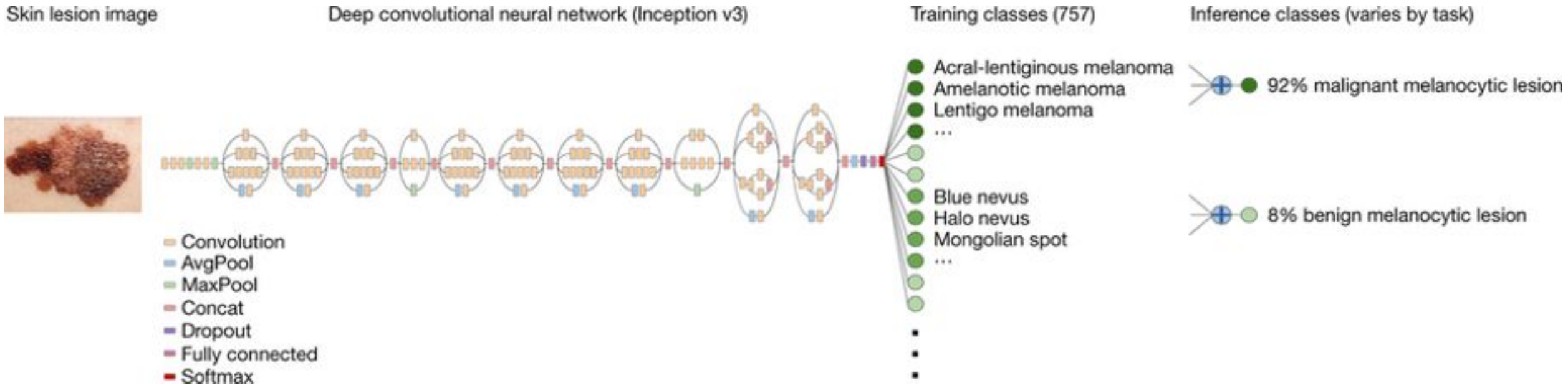
Model	TF-Slim File	Checkpoint	Top-1 Accuracy	Top-5 Accuracy
Inception V1	Code	inception_v1_2016_08_28.tar.gz	69.8	89.6
Inception V2	Code	inception_v2_2016_08_28.tar.gz	73.9	91.8
Inception V3	Code	inception_v3_2016_08_28.tar.gz	78.0	93.9
Inception V4	Code	inception_v4_2016_09_09.tar.gz	80.2	95.2

TF-Slim Code:
Defines the model architecture

Checkpoint File:
Trained model parameters

https://github.com/tensorflow/models/tree/master/research/slim#Pretrained

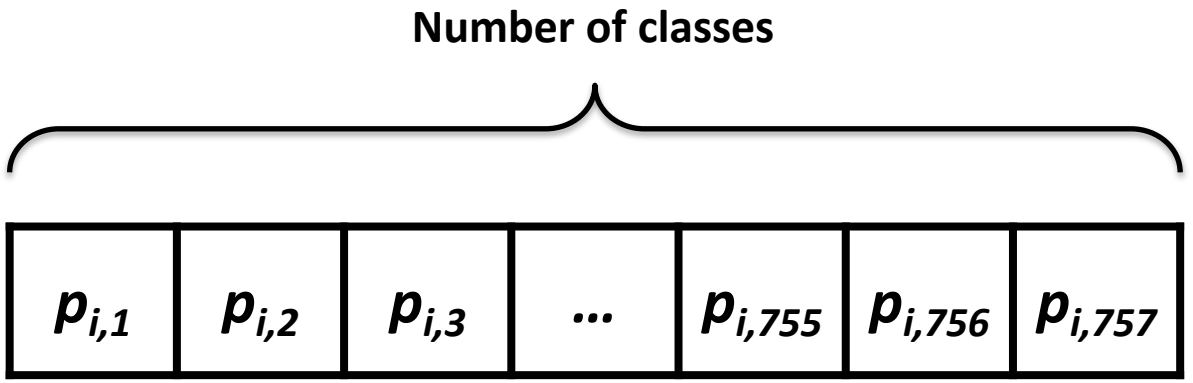
Retraining the Inception v3 CNN



- Begin with the parameters learned for the 2014 ImageNet challenge (to classify everyday images)
- Make only a very minor modification to the architecture
- Fine-tune ALL parameters using images of skin lesions

Modifying the Architecture:

1000 ImageNet Classes -> 757 Skin Lesion Classes

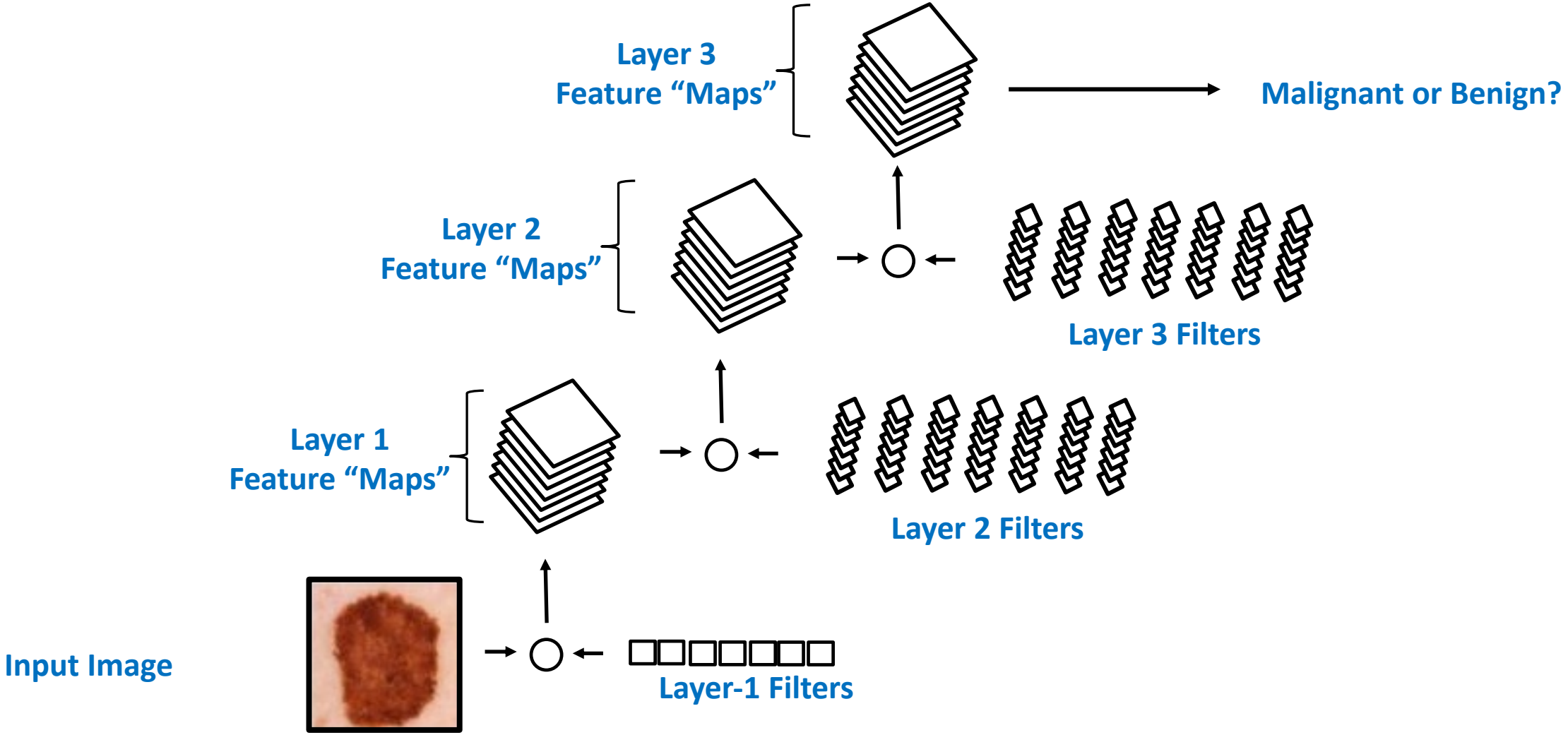


Values in all boxes sum to 1

$$p_{ij} = p(y_i = j|x_i)$$

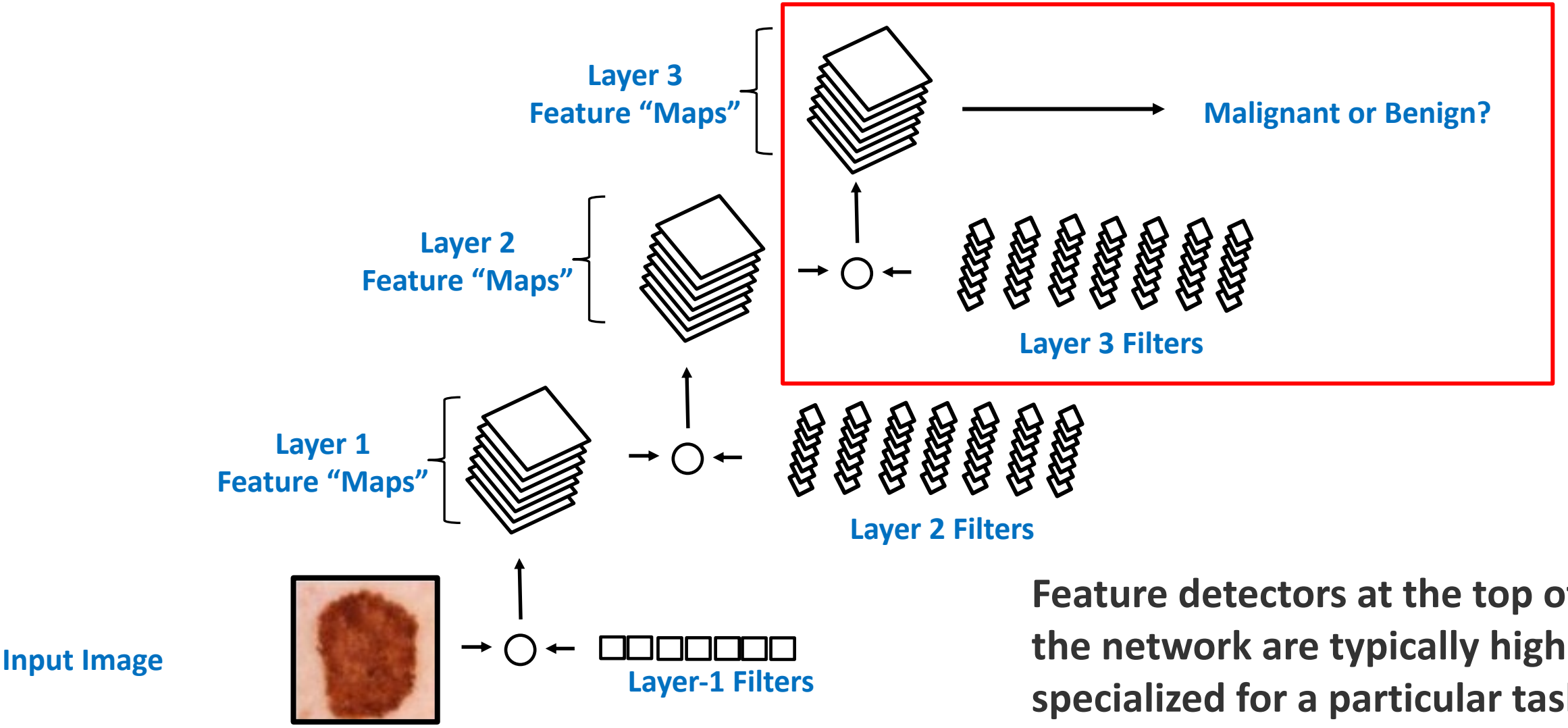
Fine-tuning the Parameters

”pre-training”, or “transfer learning”



Fine-tuning the Parameters

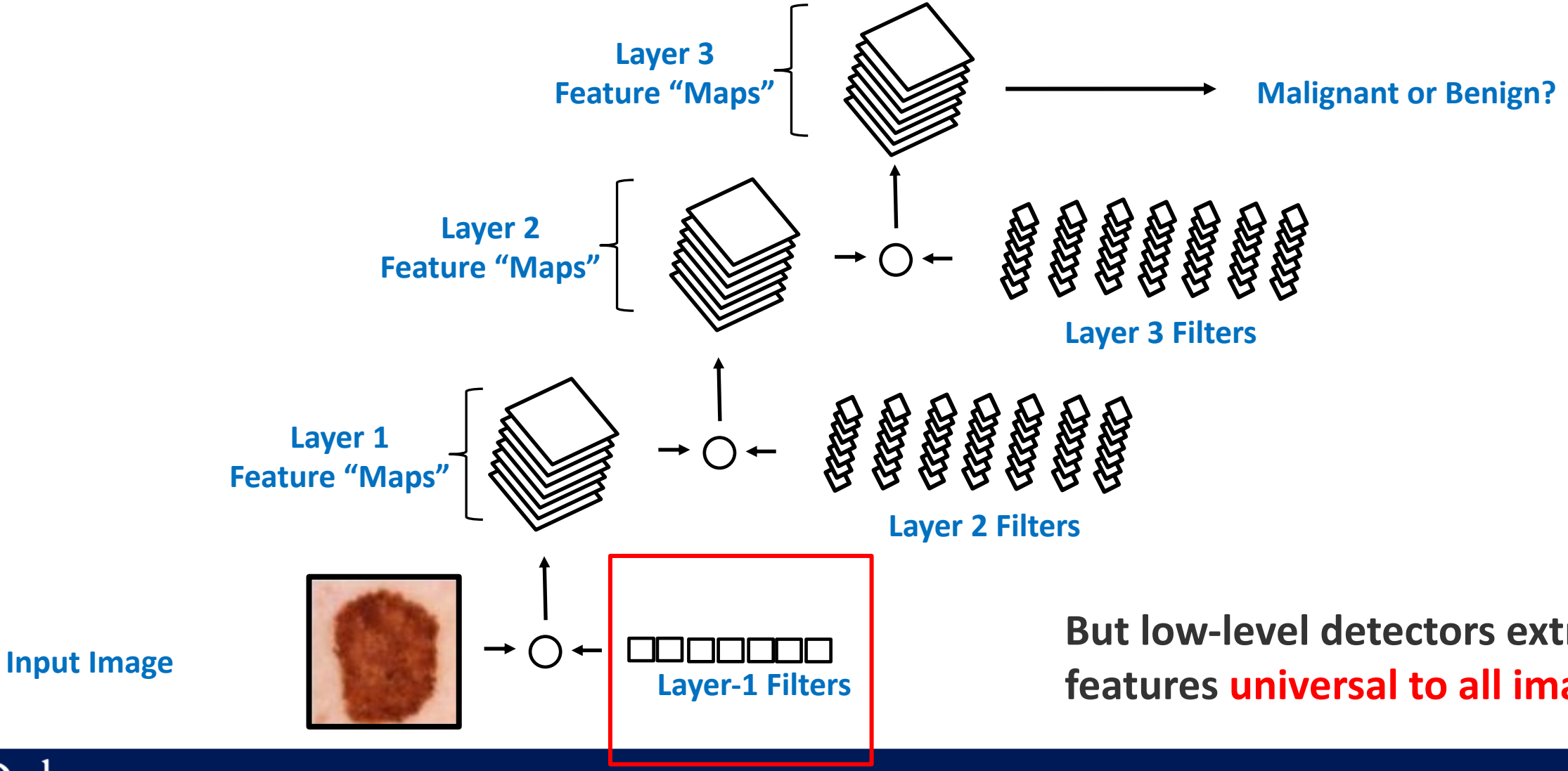
”pre-training”, or “transfer learning”



Feature detectors at the top of the network are typically highly specialized for a particular task

Fine-tuning the Parameters

”pre-training”, or “transfer learning”



But low-level detectors extract features universal to all images

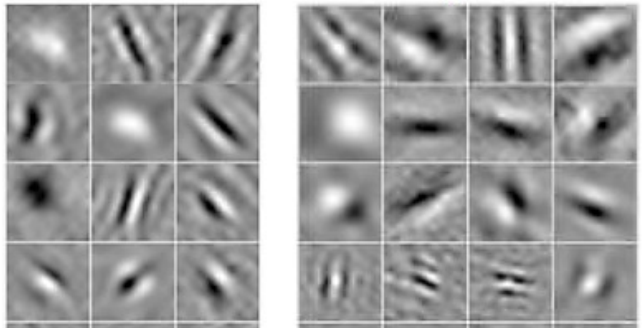
https://docs.gimp.org/en/plug-in-convmatrix.html

A filter that detects edges may be useful for many classification tasks.

Transfer Learning

Layer 1 Filters,
Convolutional Neural Network

Neuron Receptive Fields,
Macaque Visual Cortex



Benefits of Pre-training, in Brief

1) fine-tuning a pre-trained model is **at least as good as learning from scratch**
   (has been empirically demonstrated)

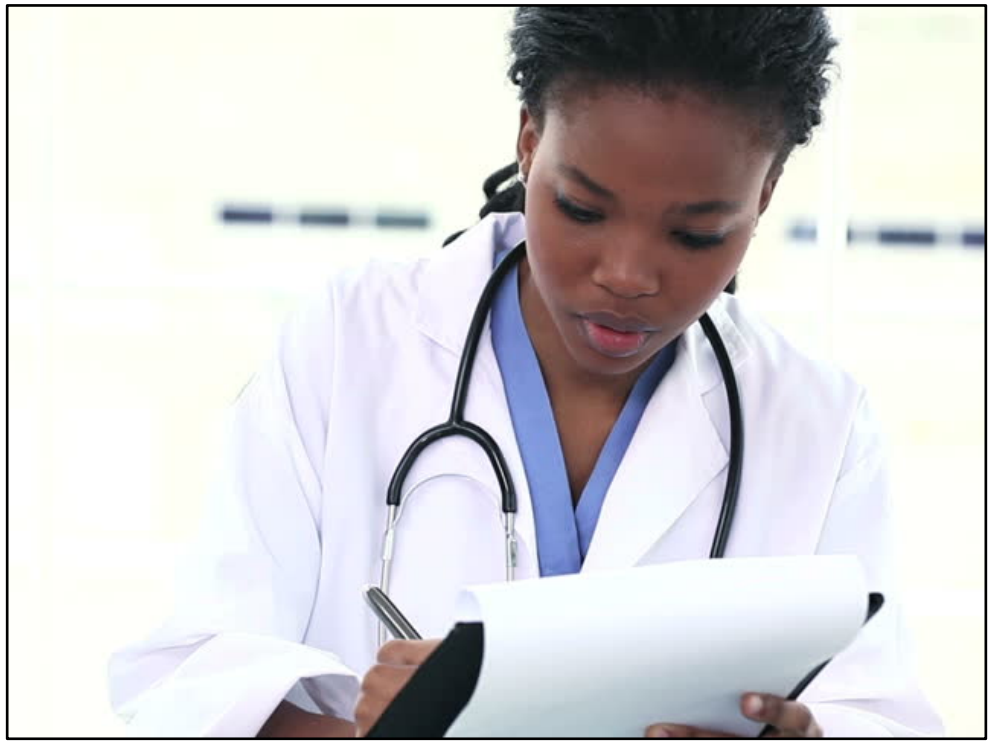
2) fine-tuning tends to work **better with smaller datasets**

3) best tuning “depth” depends on the application and size of dataset

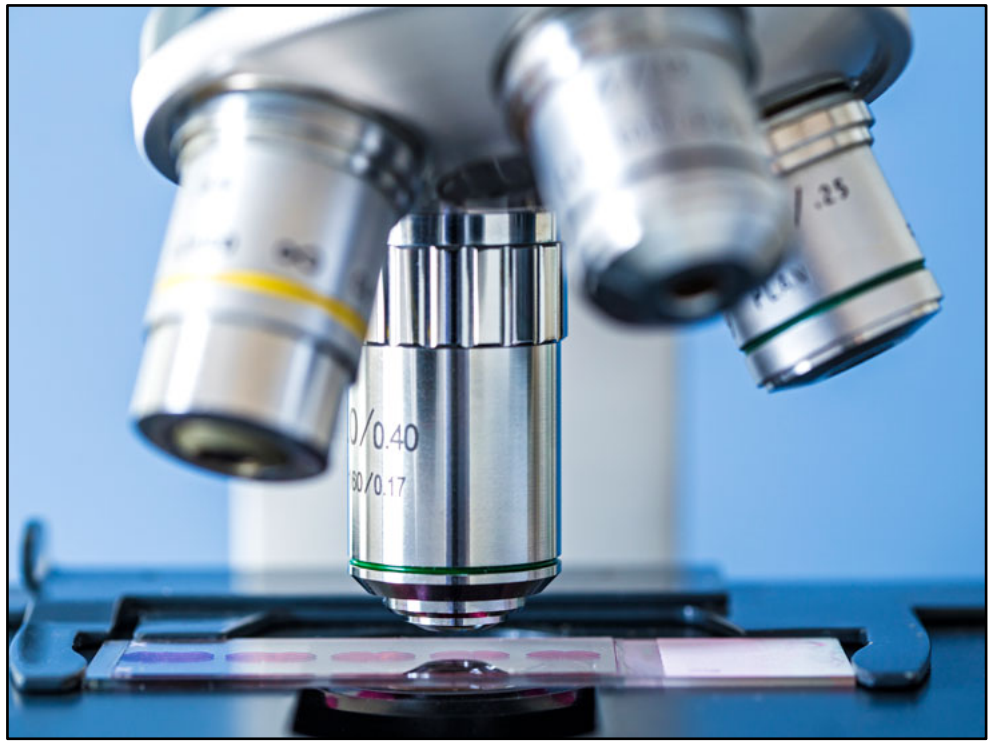
METHODS: ESTEVA ET AL.

Two Types of Labels

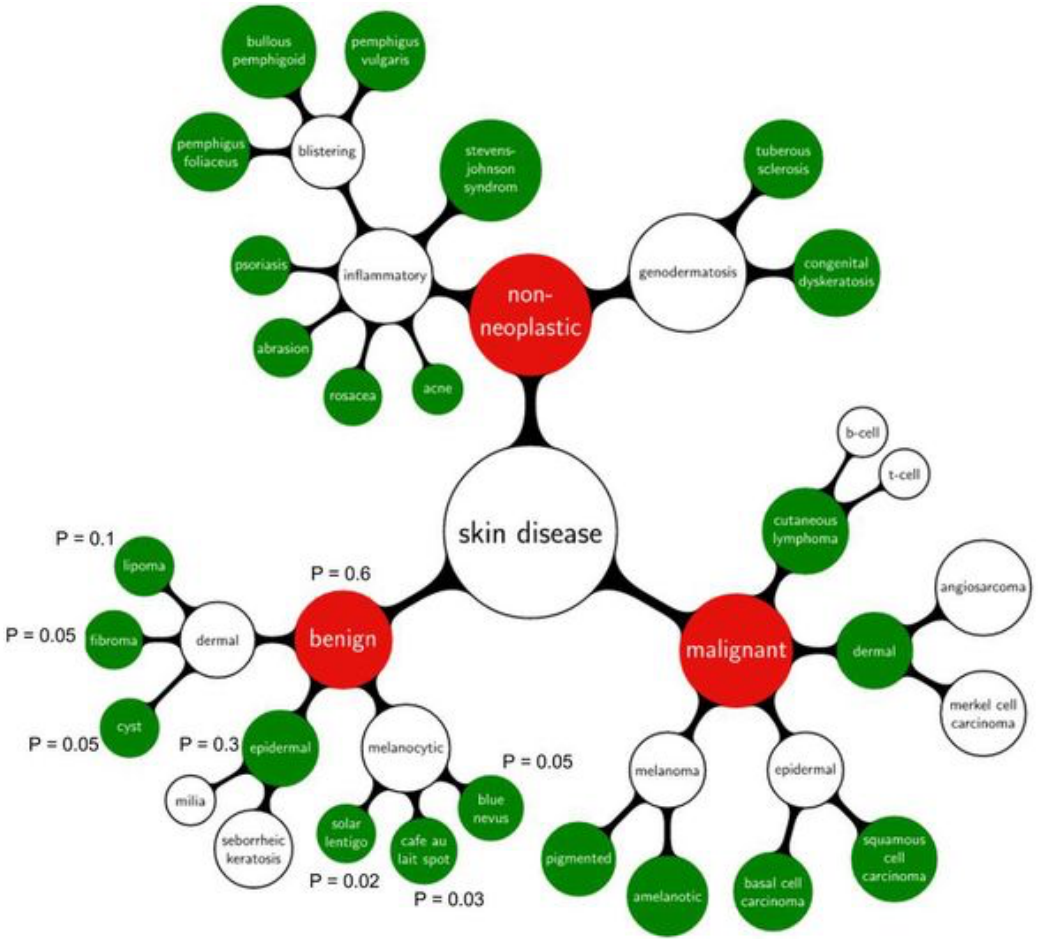
All images: dermatologists' annotations



Some images: biopsy results



Dermatologists' Taxonomy -> Training Classes



$$P[u] = \sum_{v \in C(u)} P[v]$$

Disease Partitioning Algorithm:

- Descend the tree until the current node contains <1000 images across all child nodes. Add these images as a distinct training class.
- Resulted in 757 training classes.
- However, validation was performed on higher-level nodes.

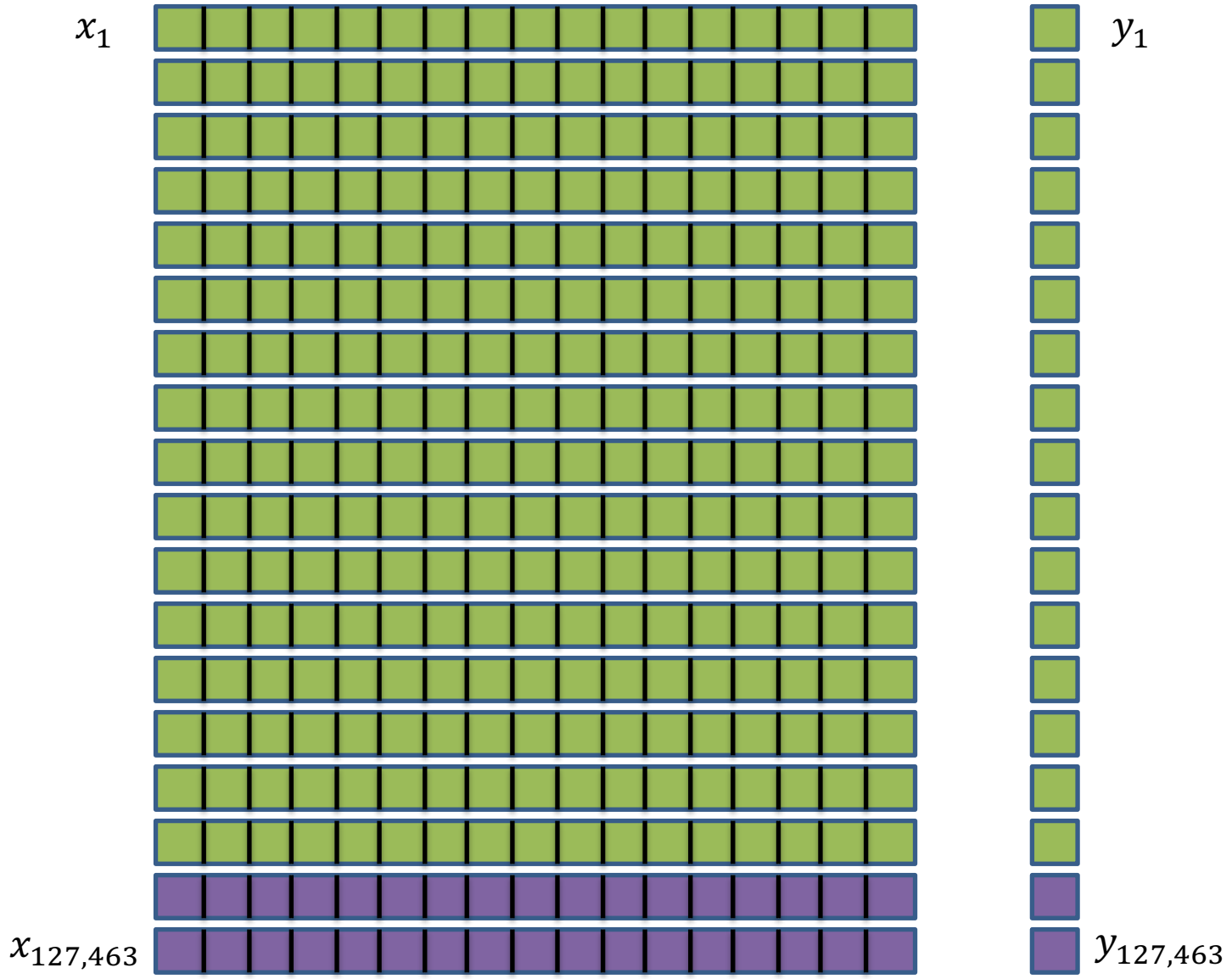
Multiple Classification Tasks

1. Match dermatologists' labeling:
   - Three-class disease partition
   - Nine-class disease partition

2. Biopsy-proven benign versus malignant lesions:
   - Keratinocyte carcinoma vs benign seborrheic keratosis
   - Malignant melanoma vs benign nevus
     - Standard images
     - Dermoscopy

Evaluation, Part 1:

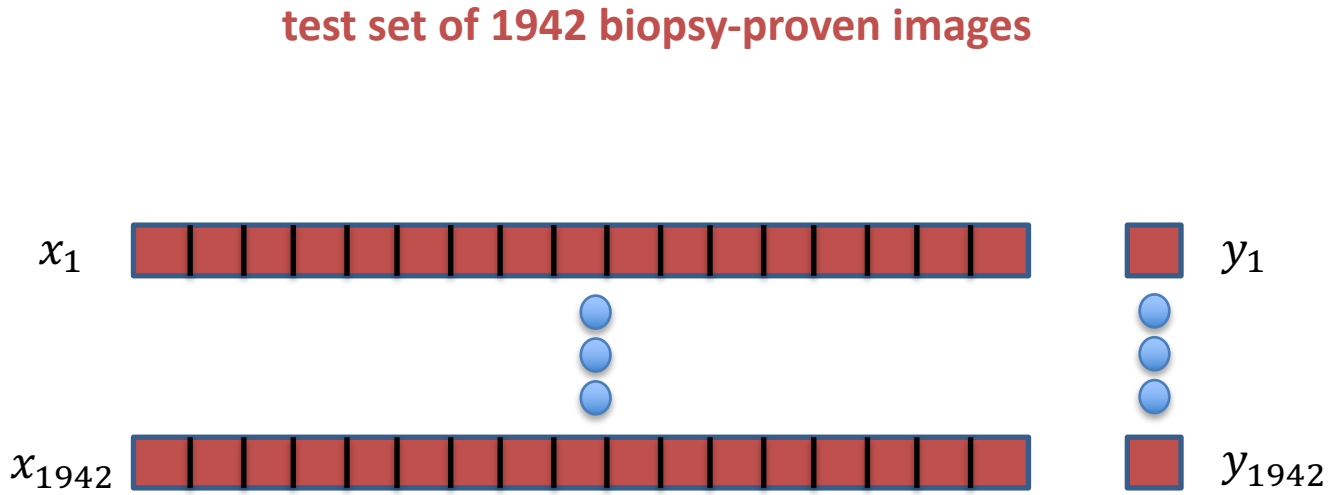
- 9-fold cross-validation
  - 757 training classes derived from dermatologists' annotations
  - 3 and 9-class validation partitions
  - two dermatologists
- Not clear what data were used to decide when to stop training



training set **validation set**

Evaluation, Part 2:

Model trained during Part 1 is compared to 21 dermatologists on a test set of biopsy-proven images



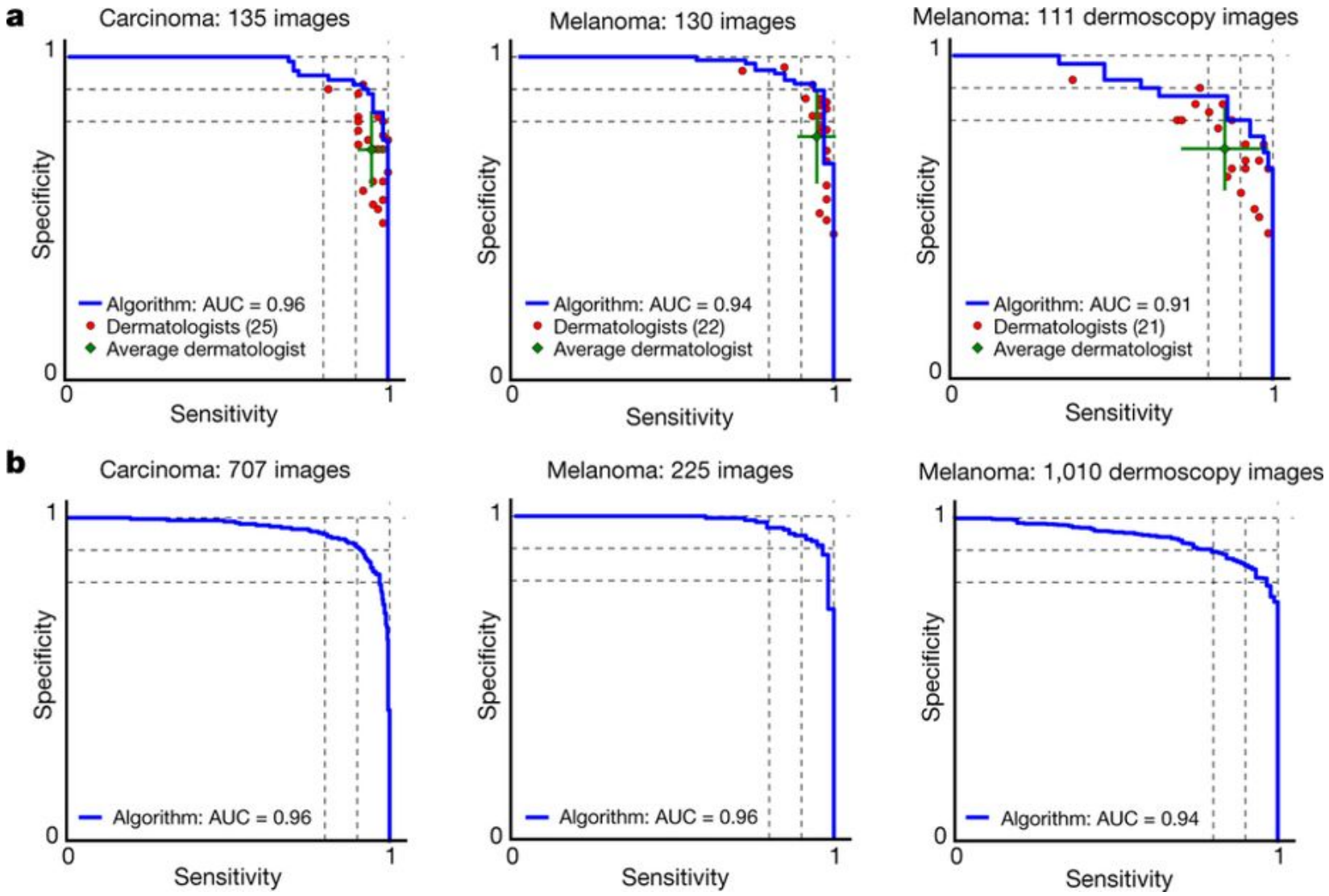
Q: In the “data preparation” section, authors emphasize that images of the same lesion (from the same person) were not split between the training and validation sets. Why is this important?

“Our dataset contains sets of images corresponding to the same lesion but from multiple viewpoints, or multiple images of similar lesions on the same person. While this is useful training data, extensive care was taken to ensure that these sets were not split between the training and validation sets.”

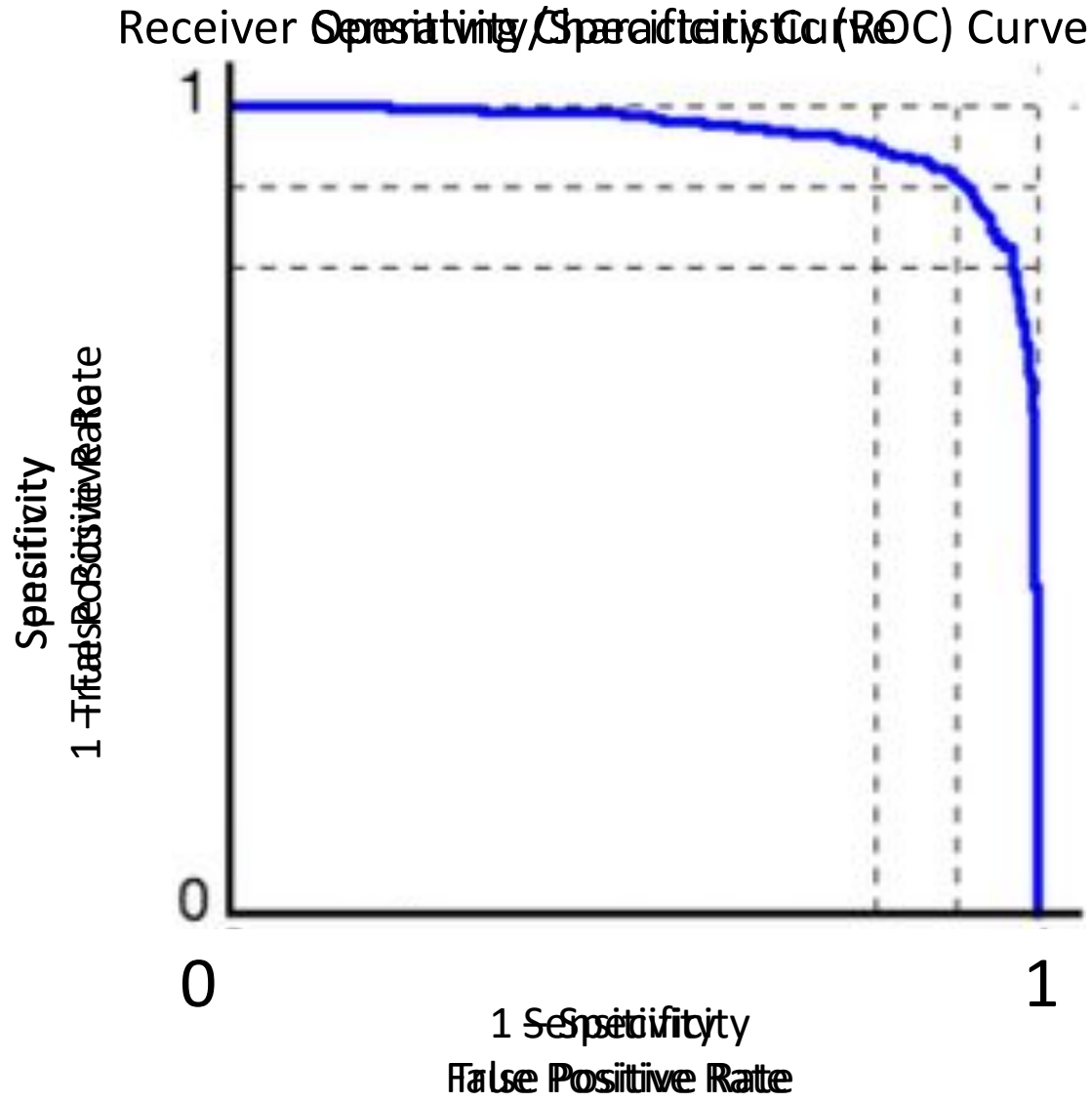
Interpreting the ROC Curve

CLASSIFICATION RESULTS

Results: CNN Performance vs Dermatologists



Evaluation Measures: Classification



Sensitivity, or True Positive Rate:

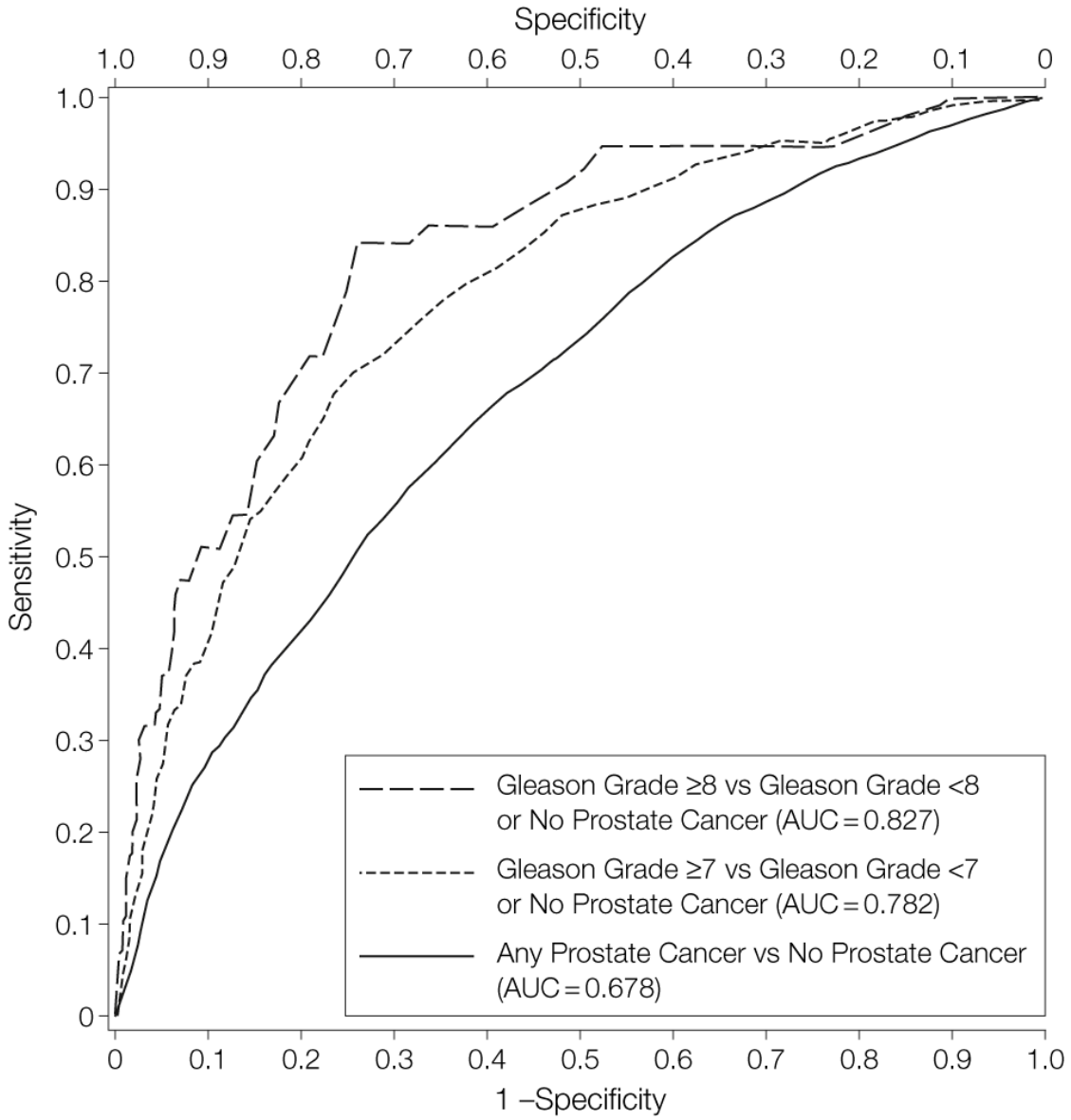
$$\frac{\text{true positives}}{\text{all condition positives}}$$

Specificity, or (1 – False Positive Rate):

$$\frac{\text{true negatives}}{\text{all condition negatives}}$$

Accuracy:

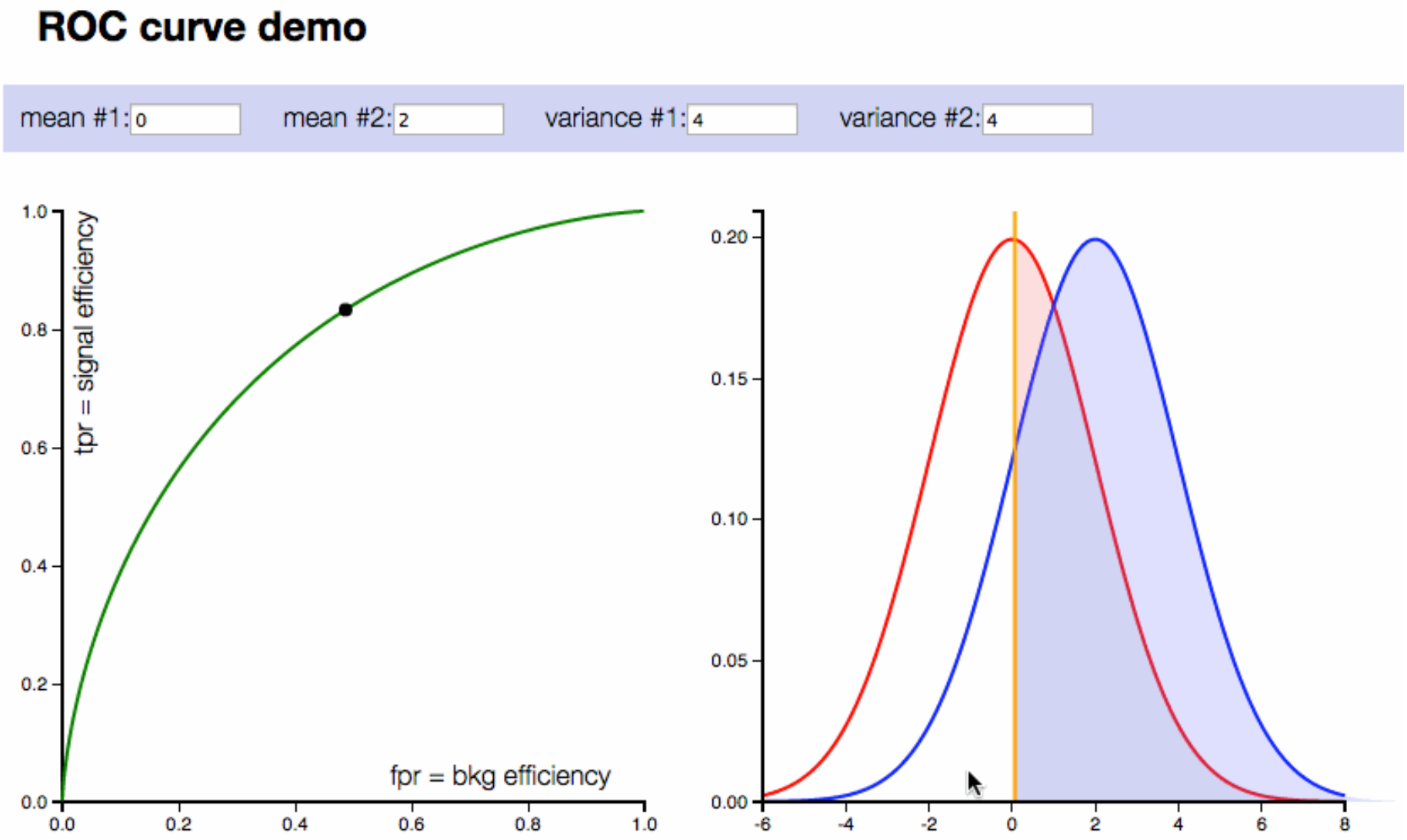
$$\frac{\text{true positives} + \text{true negatives}}{\text{total cases}}$$



Receiver Operating Characteristic Curve for Prostate-Specific Antigen (PSA)

Thompson IM, Ankerst DP, Chi C, et al. Operating Characteristics of Prostate-Specific Antigen in Men With an Initial PSA Level of 3.0 ng/mL or Lower. *JAMA.* 2005;294(1):66–70. doi:10.1001/jama.294.1.66

Set a “classification threshold” to distinguish between groups



http://arogozhnikov.github.io/2015/10/05/roc-curve.html

Once a threshold is set, we get a “confusion matrix”

	Condition Positive	**Condition Negative**
Prediction Positive	True Positive	False Positive
Prediction Negative	False Negative	True Negative

Sensitivity, or True Positive Rate:

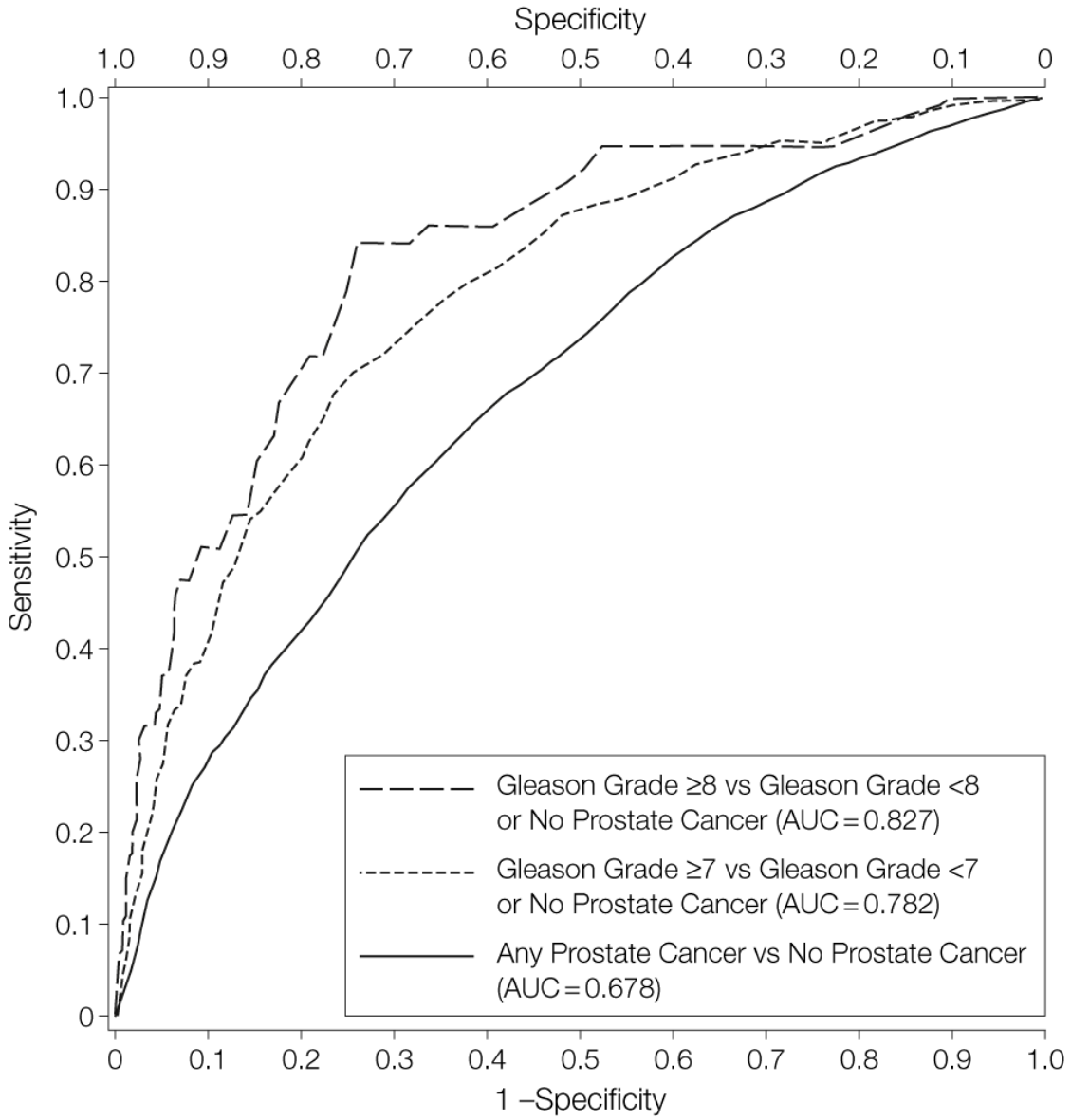
$$\frac{\text{true positives}}{\text{all condition positives}}$$

Specificity, or (1 – False Positive Rate):

$$\frac{\text{true negatives}}{\text{all condition negatives}}$$

Accuracy:

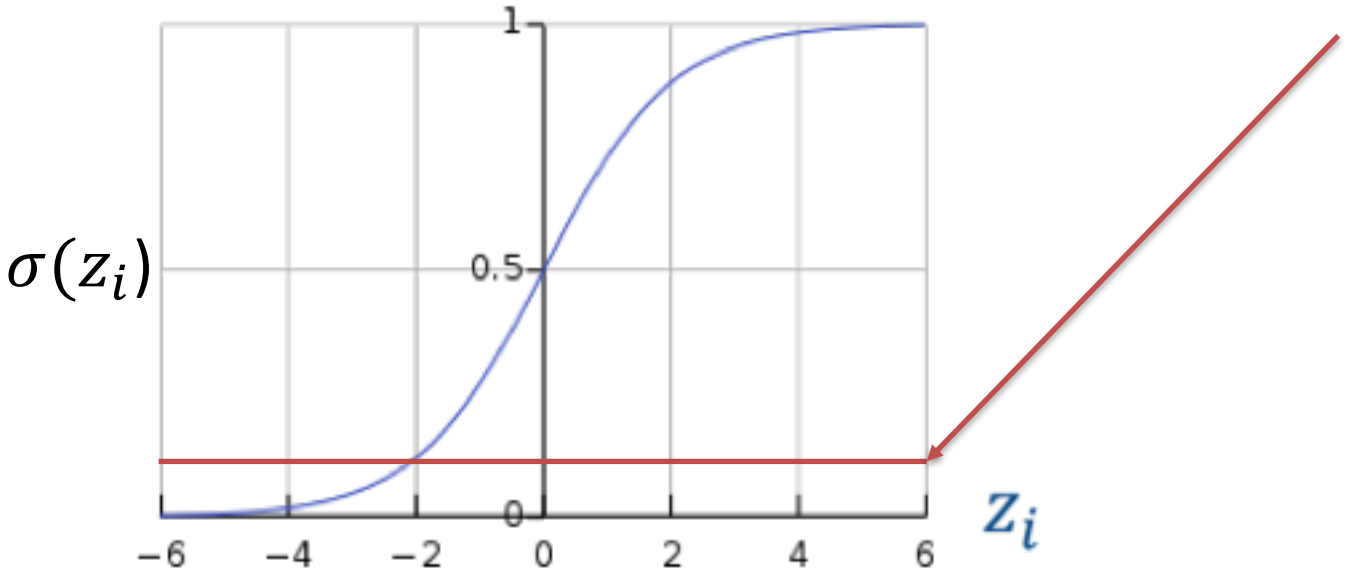
$$\frac{\text{true positives} + \text{true negatives}}{\text{total cases}}$$



1) Set a threshold on PSA
2) Make predictions:
   - Above threshold: cancer-positive
   - Below threshold: cancer-negative
3) Count true positives, true negatives, false positives, and false negatives
4) Calculate sensitivity and specificity
5) Plot point and repeat

A probabilistic classifier is like a PSA level: we can set a threshold on its predictions

$$p(y_i = 1|x_i) = \sigma(z_i)$$

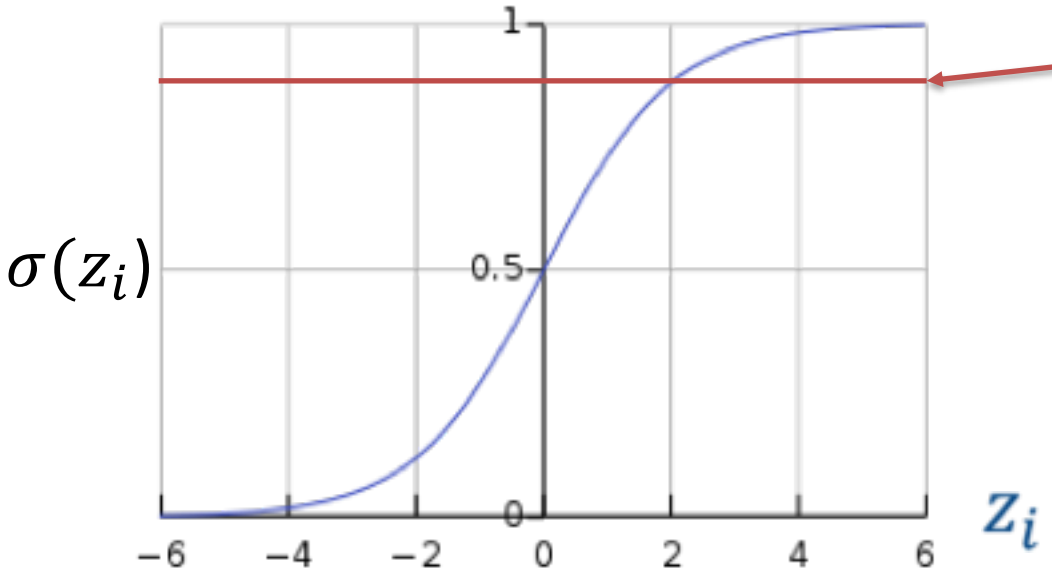


classification threshold

A low threshold favors sensitivity, because more points are classified as condition positive (i.e. 1s)

A probabilistic classifier is like a PSA level: we can set a threshold on its predictions

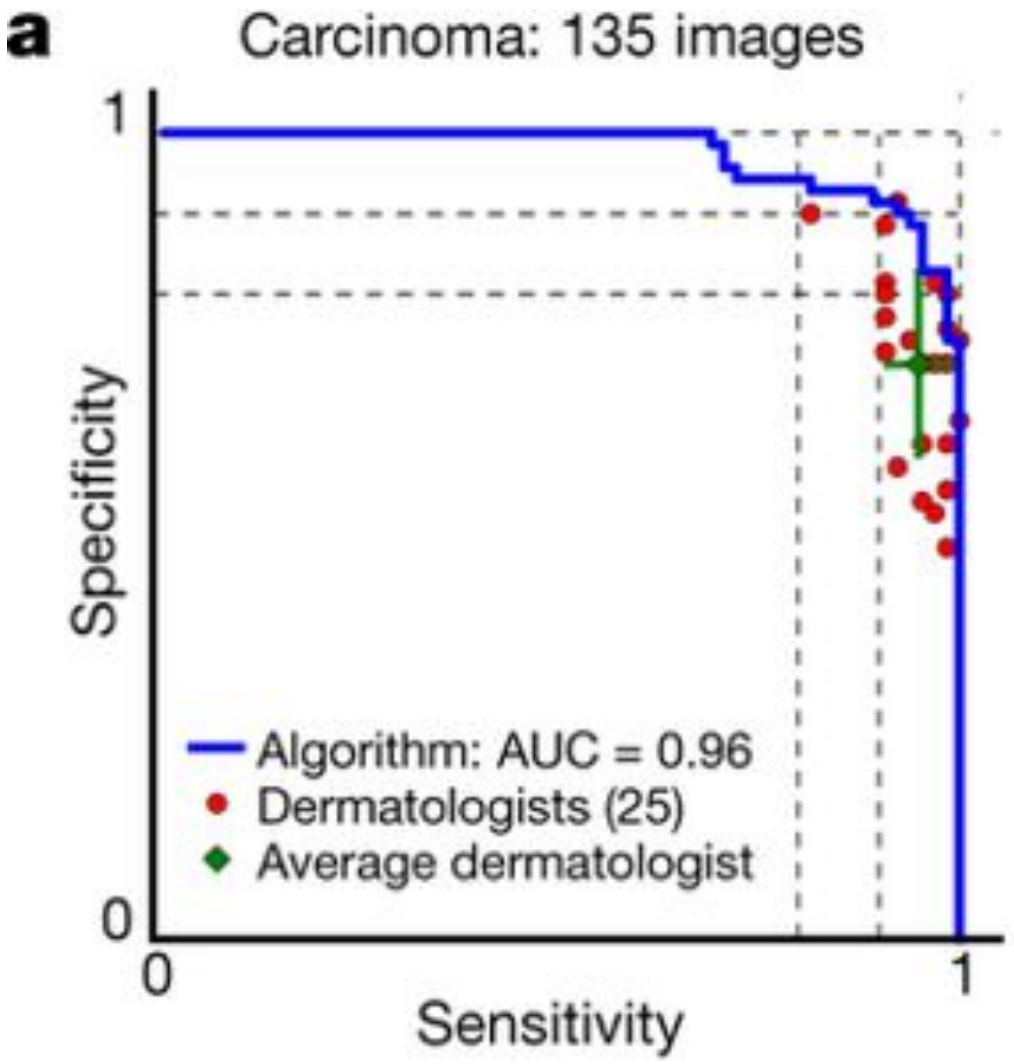
$$p(y_i = 1|x_i) = \sigma(z_i)$$



classification threshold

A high threshold favors specificity, because more points are classified as condition negative (i.e. 0s)

Results: CNN Performance vs Dermatologists



High Sensitivity ≠ High PPV

	Condition Positive	**Condition Negative**
Prediction Positive	True Positive	False Positive
Prediction Negative	False Negative	True Negative

Sensitivity, or True Positive Rate:

$$\frac{\text{true positives}}{\text{all condition positives}}$$

Positive Predictive Value, or Precision:

$$\frac{\text{true positives}}{\text{all prediction positives}}$$

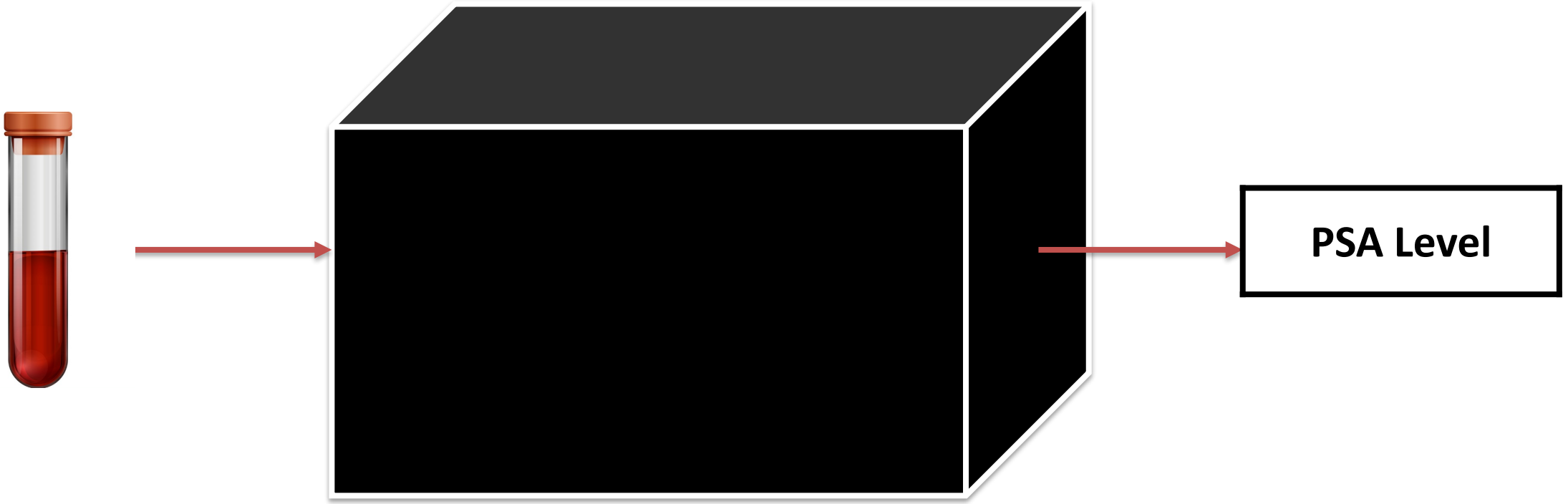
Baseline prevalence is critical

How do the authors attempt to look inside the “black box”?

MODEL INTERPRETATION

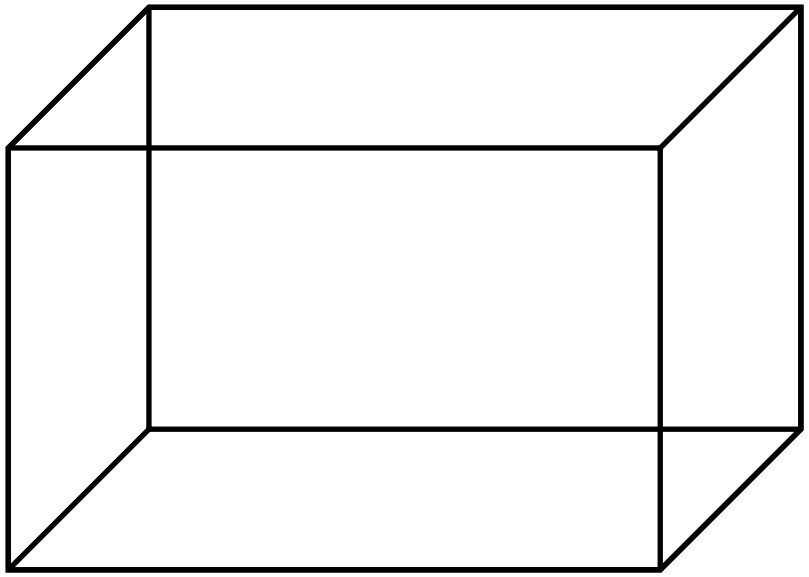
Machine Learning: A Black Box?

Prostate-specific antigen measurement: A Black Box?

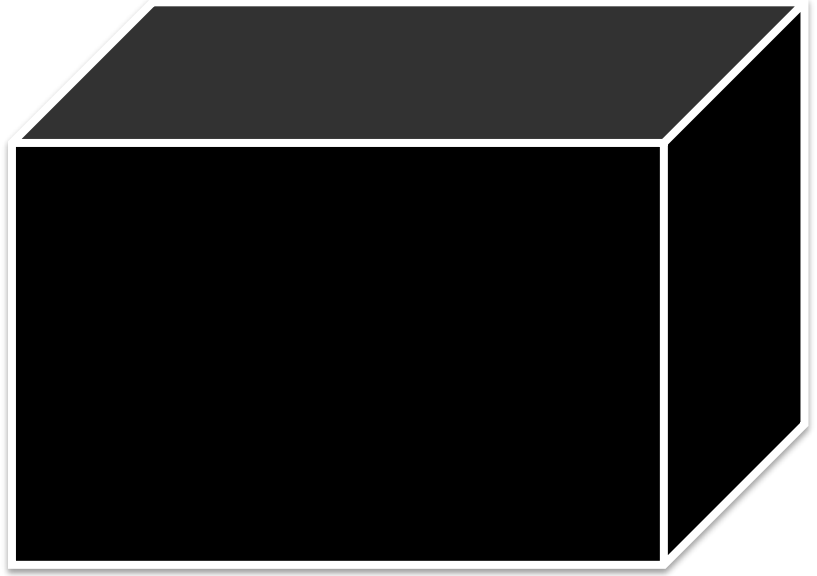


Two competing perspectives

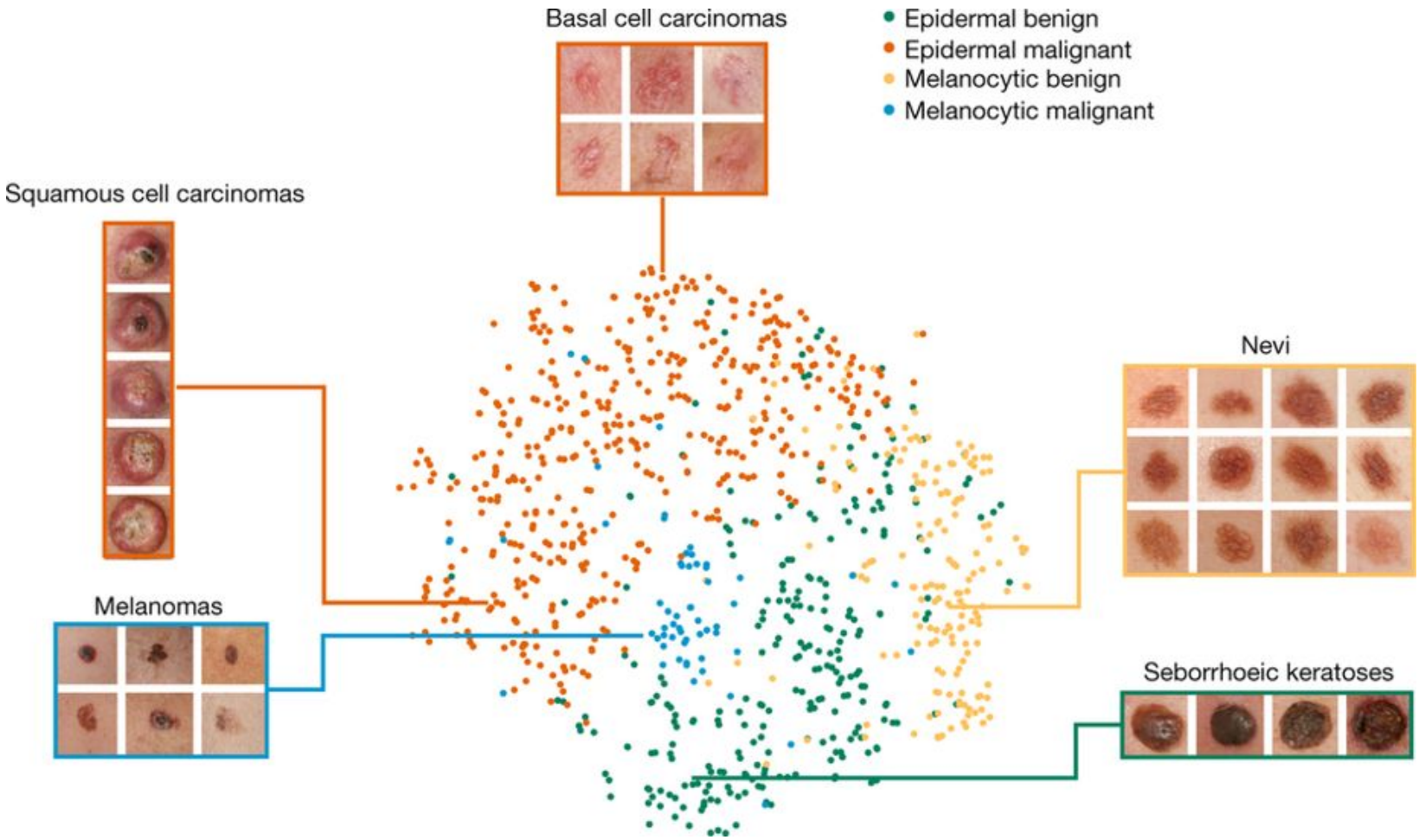
Clinicians must fully understand how their diagnostic tools work



Clinicians must be sure these tools are *valid* and *reliable*



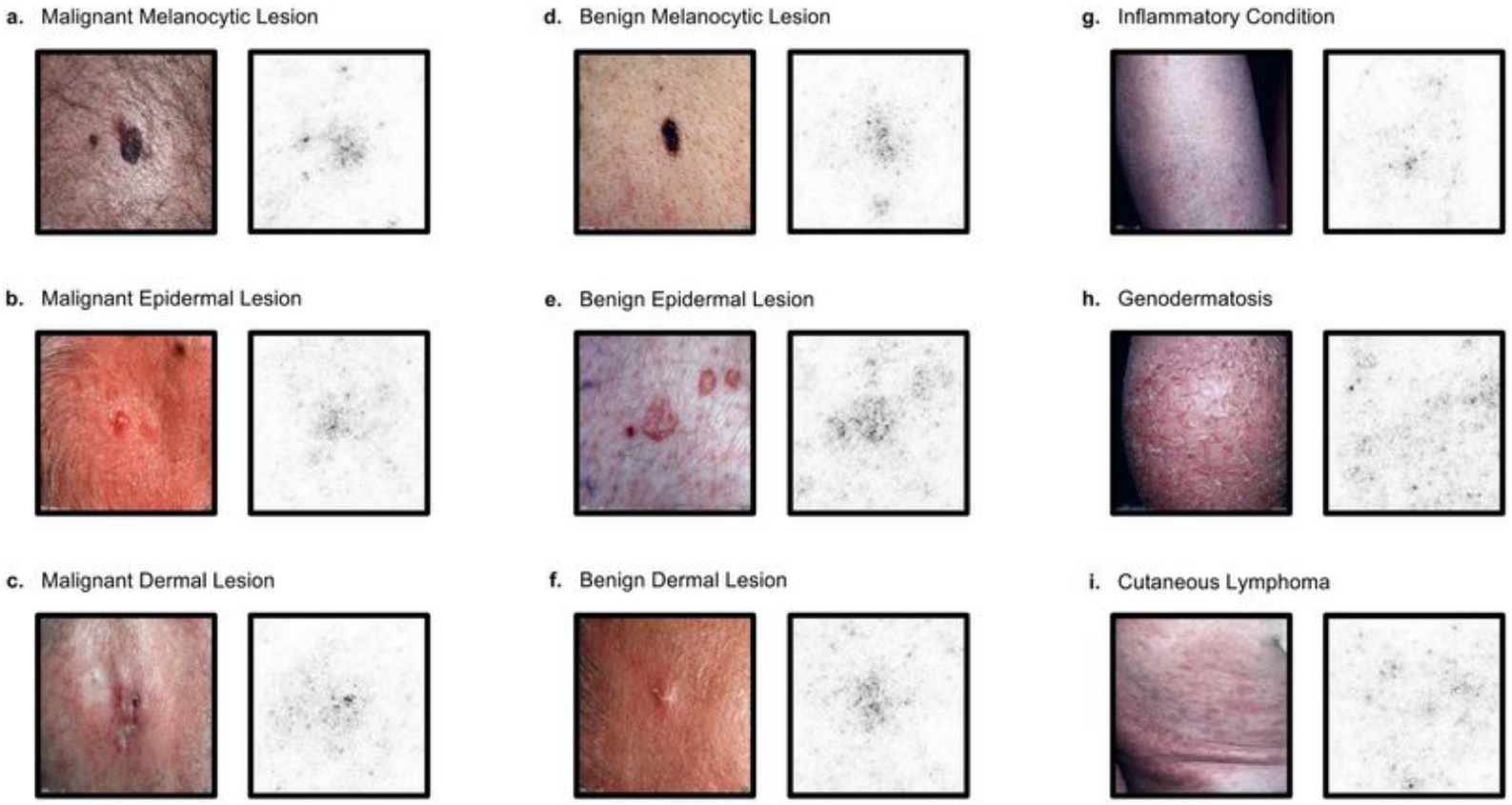
t-SNE visualization of last hidden layer



t-distributed stochastic neighbor embedding (t-SNE) maps high-dimensional points to two-dimensional points such that similarity between pairs of points is (approximately) preserved

Q: How much does this visualization help us understand the model?

Saliency maps for example images



Saliency maps show gradients for each pixel with respect to the CNN's loss function. Darker pixels represent those with more influence.

Q: How much does this visualization help us understand the model?