Lecture 6: Evaluation metrics for machine learning in healthcare

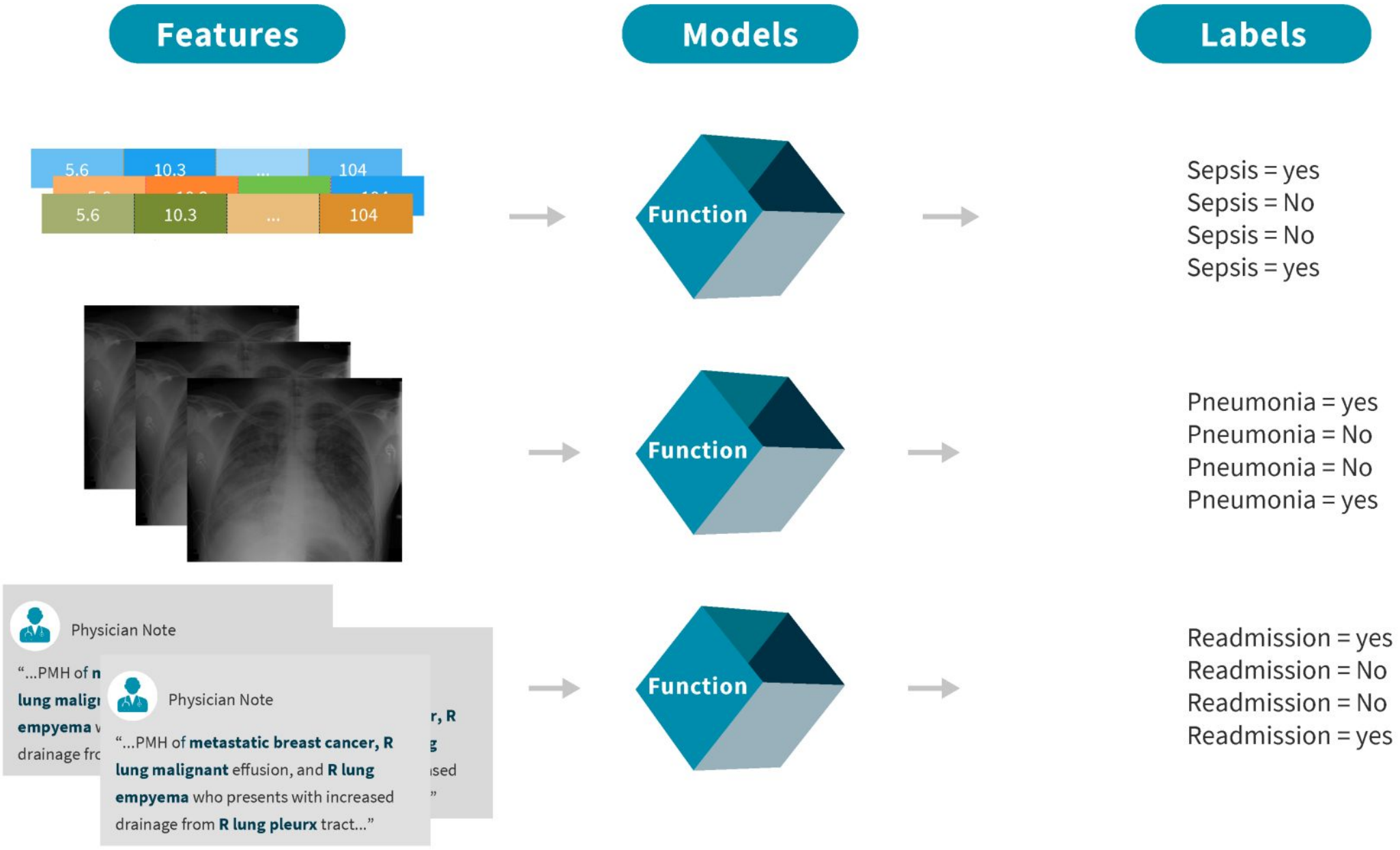
BIODS388/BIOMED388

Anuj Pareek MD PhD, Mars Huang PhD Student

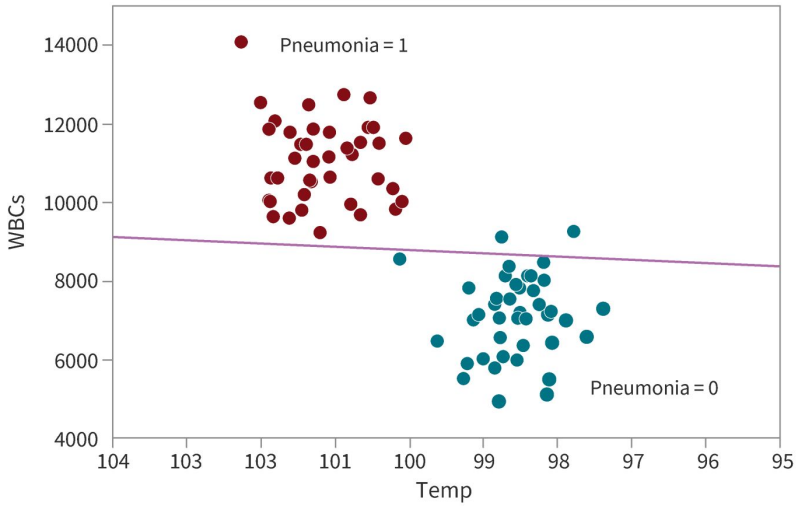
10/18/2020

Outline

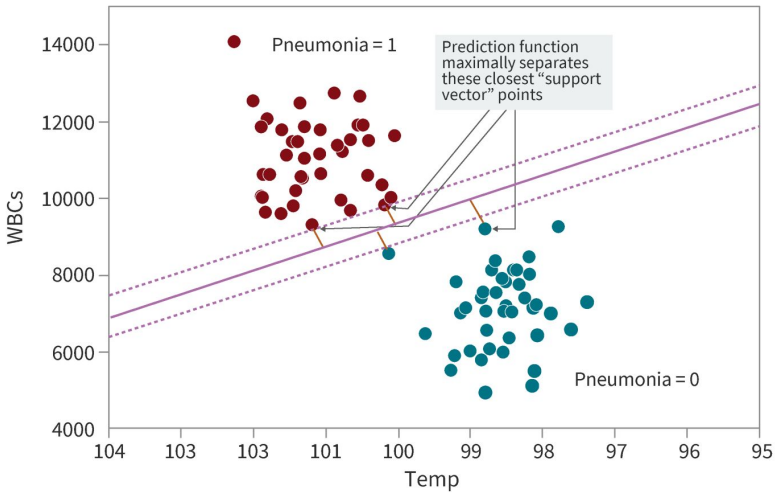
1. **Review**
2. Regression metrics
3. Classification metrics
4. Applications
5. Visual recognition metrics


Features
Models
Labels
5.6
10.3
...
104
5.6
10.3
...
104
Function
Sepsis = yes
Sepsis = No
Sepsis = No
Sepsis = yes
Function
Pneumonia = yes
Pneumonia = No
Pneumonia = No
Pneumonia = yes
Physician Note
"...PMH of n
lung malig
empyema
drainage fr
Physician Note
"...PMH of metastatic breast cancer, R lung malignant effusion, and R lung empyema who presents with increased drainage from R lung pleurx tract..."
Function
Readmission = yes
Readmission = No
Readmission = No
Readmission = yes

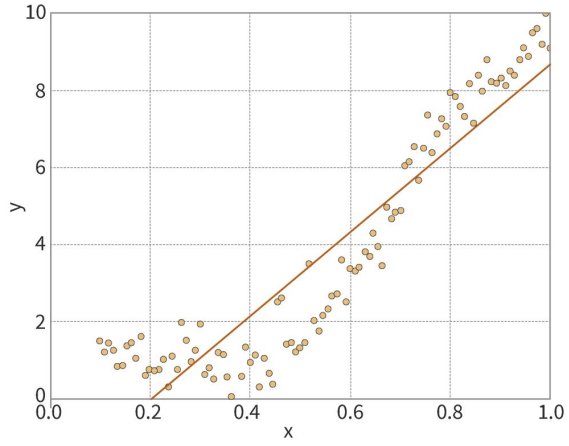
REMEMBER: LOGISTIC REGRESSION



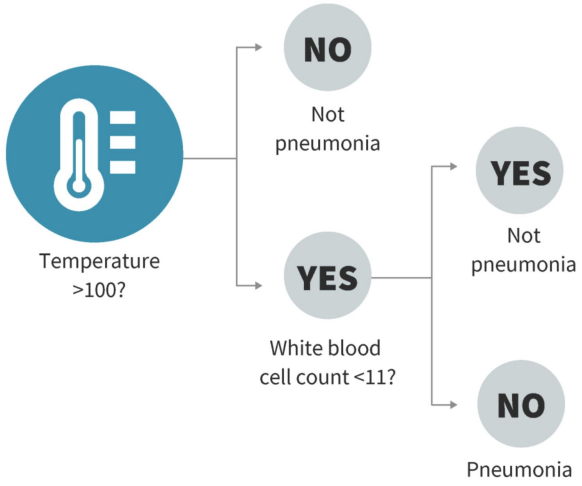
SUPPORT VECTOR MACHINES (SVMs)



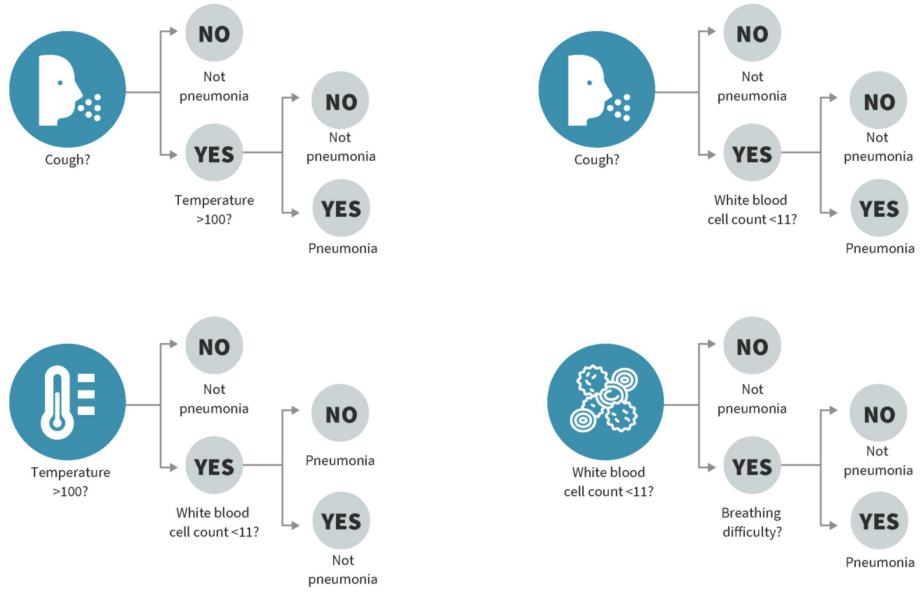
WHAT WE'VE ALREADY SEEN: LINEAR REGRESSION



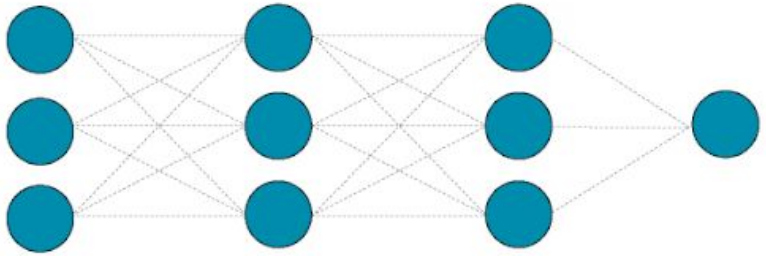
DECISION TREE



RANDOM FOREST

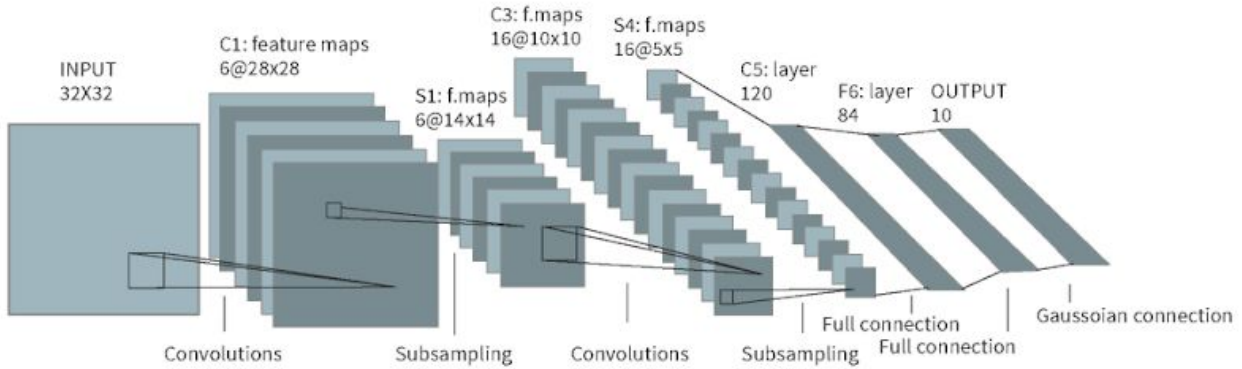


MAJOR CLASSES OF NEURAL NETWORK ARCHITECTURES



Fully connected neural networks

(fully connected layers, good for structured data inputs)

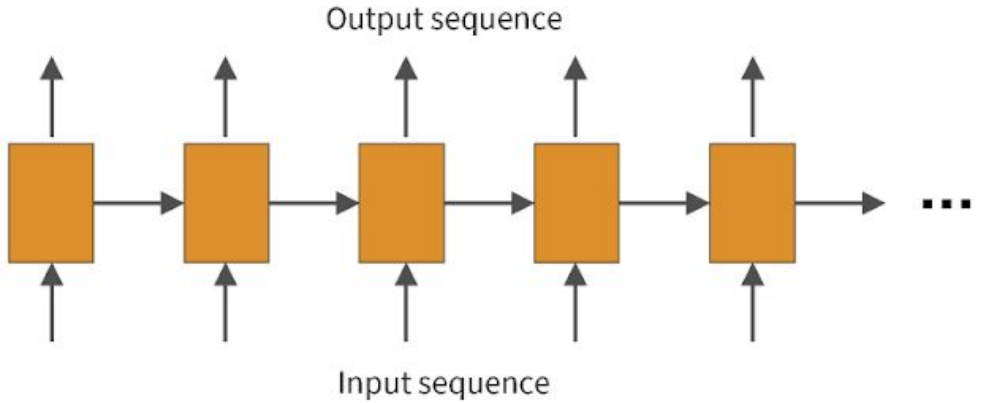


Convolutional neural networks

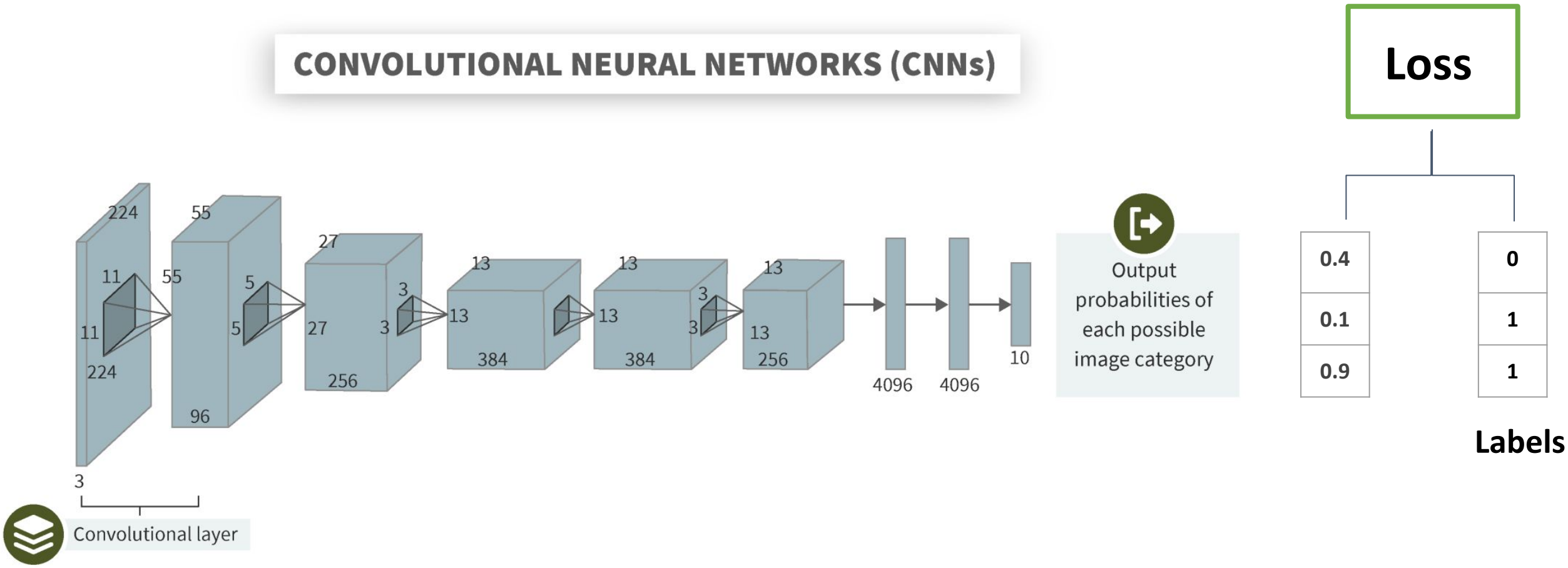
(convolutional layers, good for image inputs)

Recurrent neural networks

(fully connected layers modeling recurrence relation across sequence, good for sequence inputs)



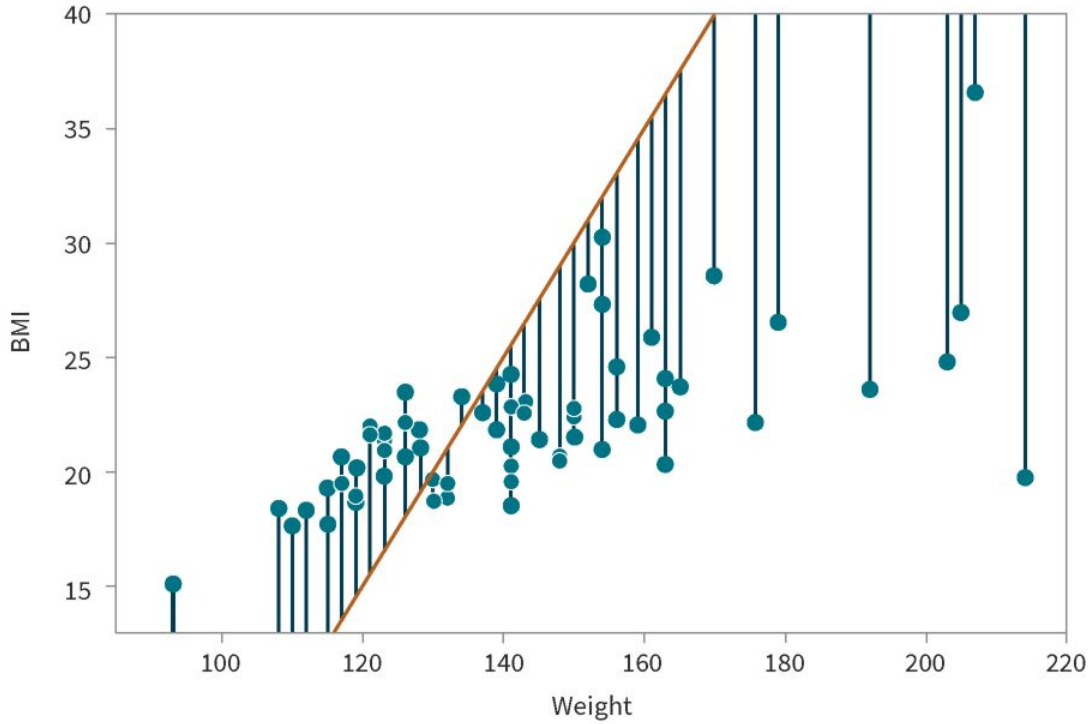
Overview of model optimization



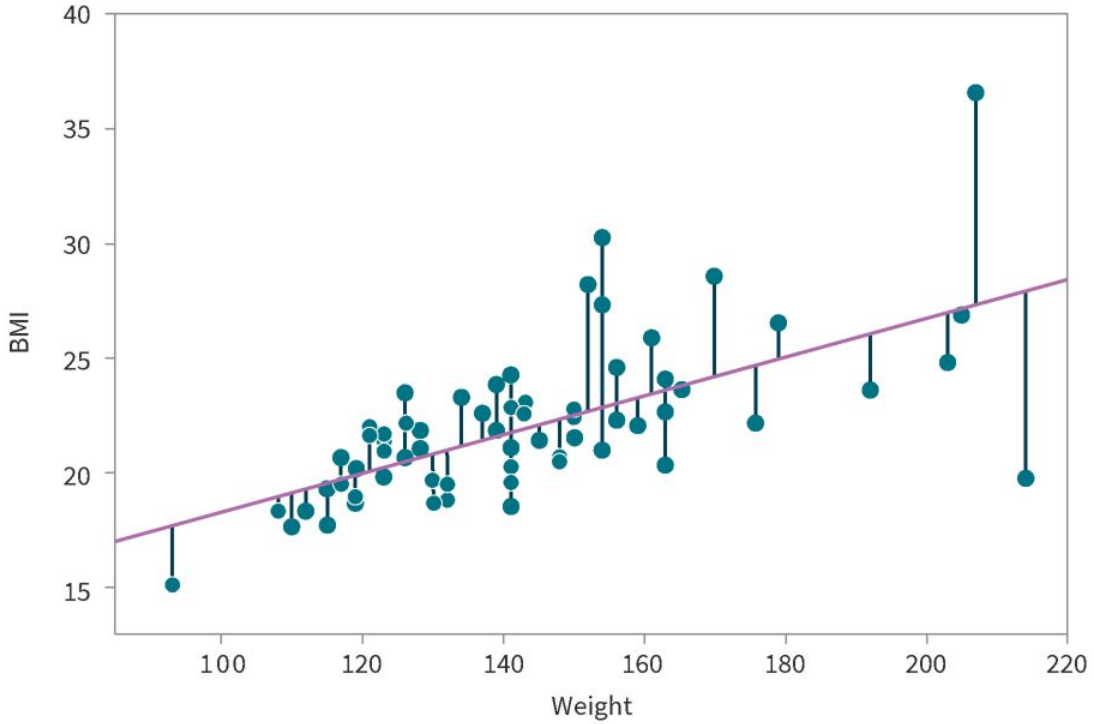
Outline

1. Review
2. **Regression metrics**
3. Classification metrics
4. Applications
5. Visual recognition metrics

EXAMPLE: *USING BODY WEIGHT TO PREDICT BODY MASS INDEX (BMI)*



A poorly trained or (poorly fitted) model has high cumulative loss

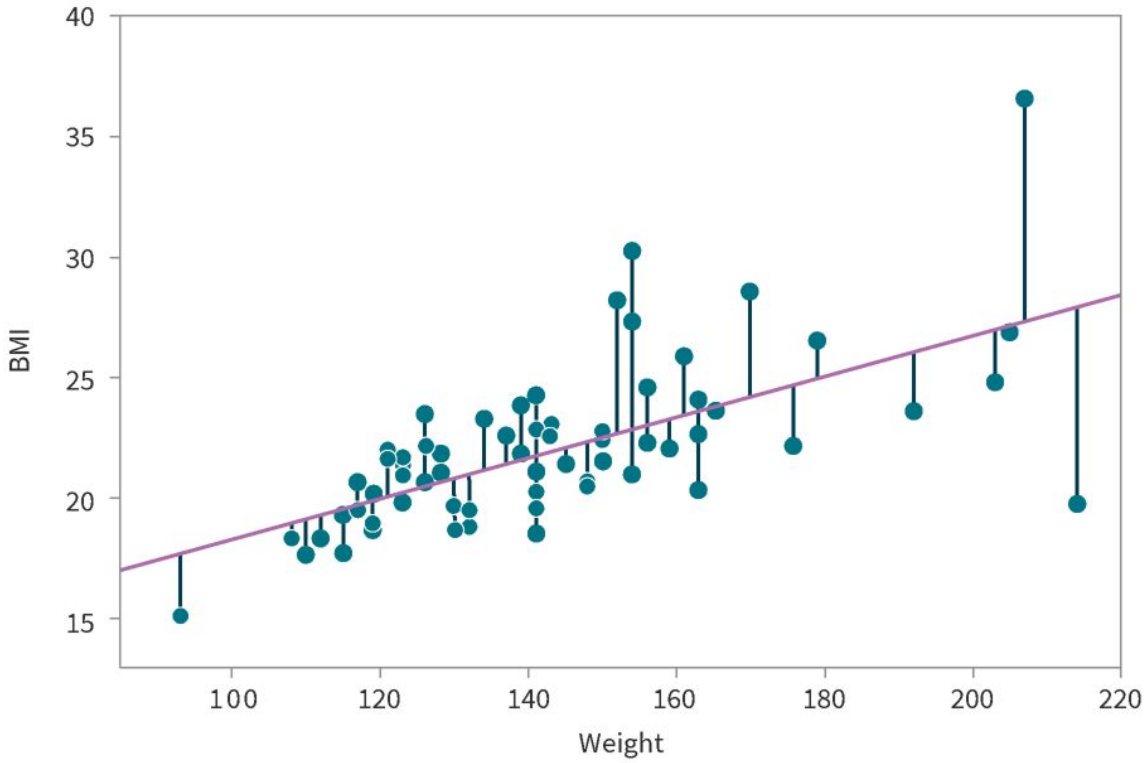


A well trained or (well fitted) model minimizes the cumulative loss.

Loss Functions in Regression

L2 Loss: $$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

L1 Loss: $$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$



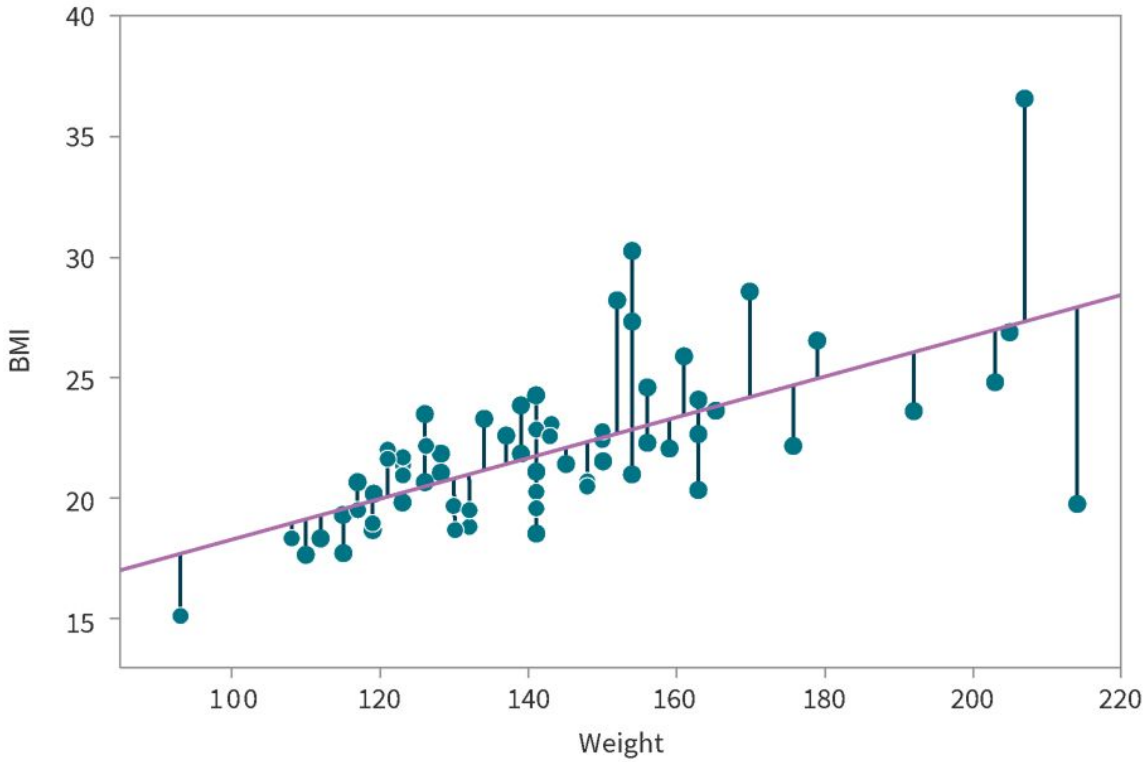
A well trained or (well fitted) model minimizes the cumulative loss.

Evaluation Metrics in Regression

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

R^2, Max Error, and others



Outline

1. Review
2. Regression metrics
3. **Classification metrics**
4. Applications
5. Visual recognition metrics

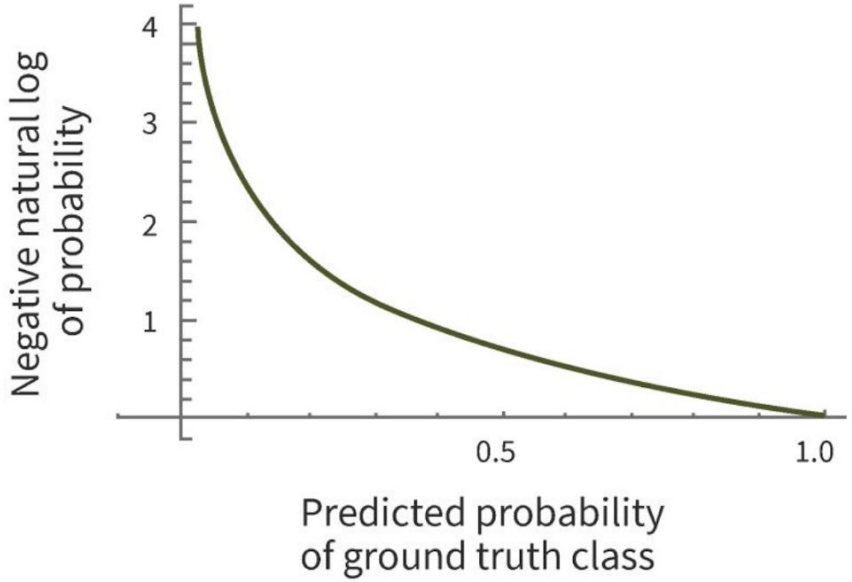
MOST COMMON LOSS FUNCTION FOR CLASSIFICATION

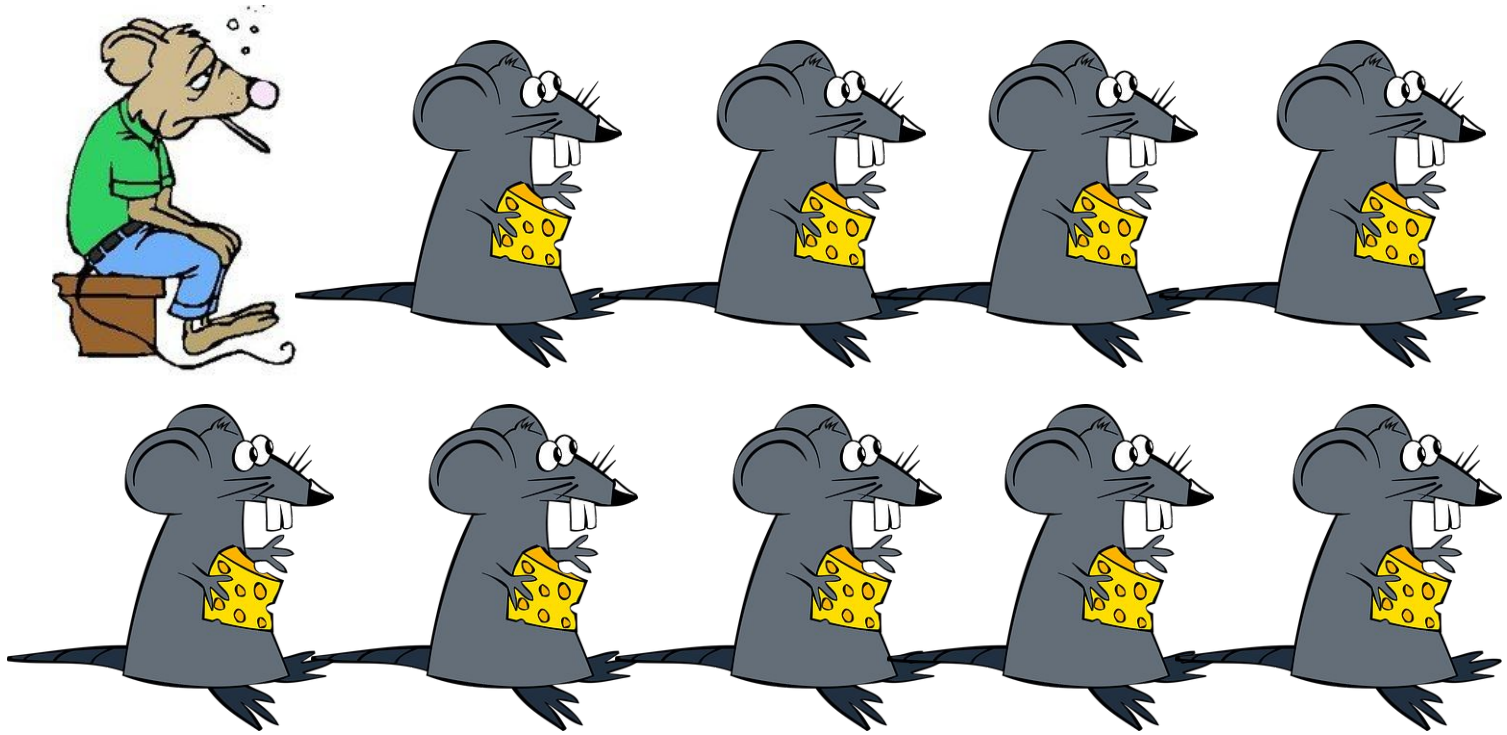
Cross Entropy Loss:

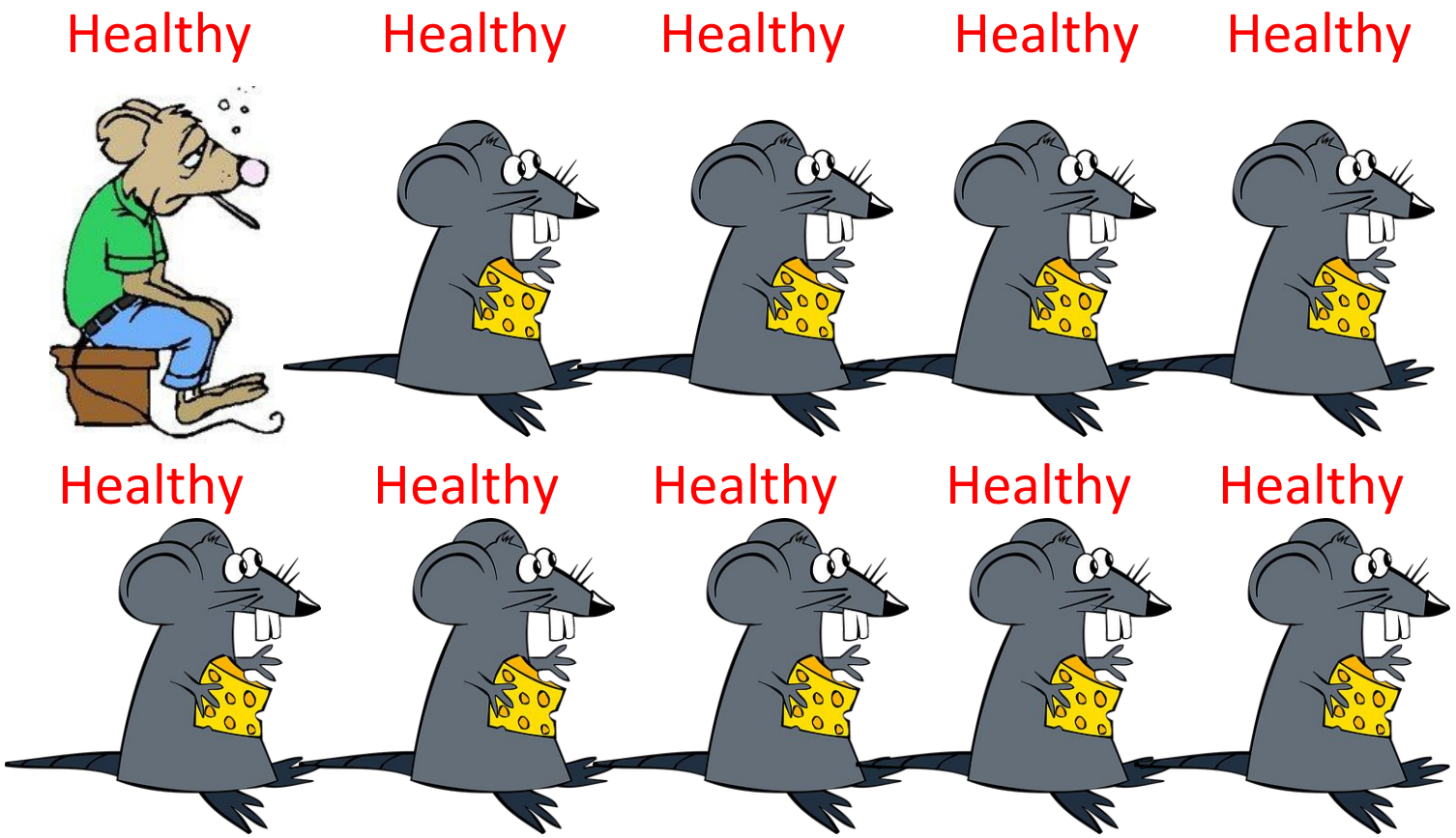
In practice often use ln (natural log)

01 Loss for one example = - ln (probability of ground truth class)

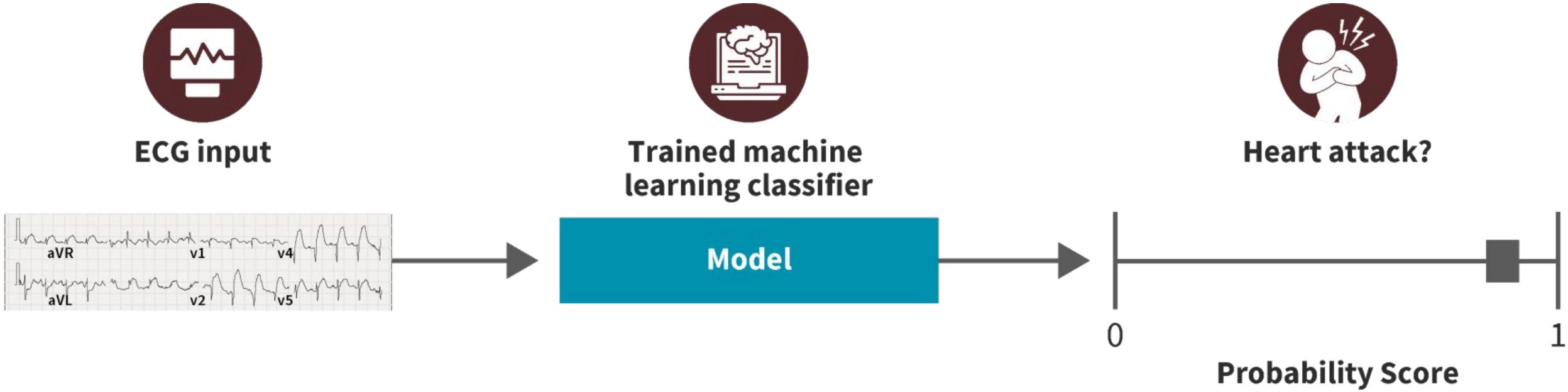
02 Loss over dataset = average of loss over all examples

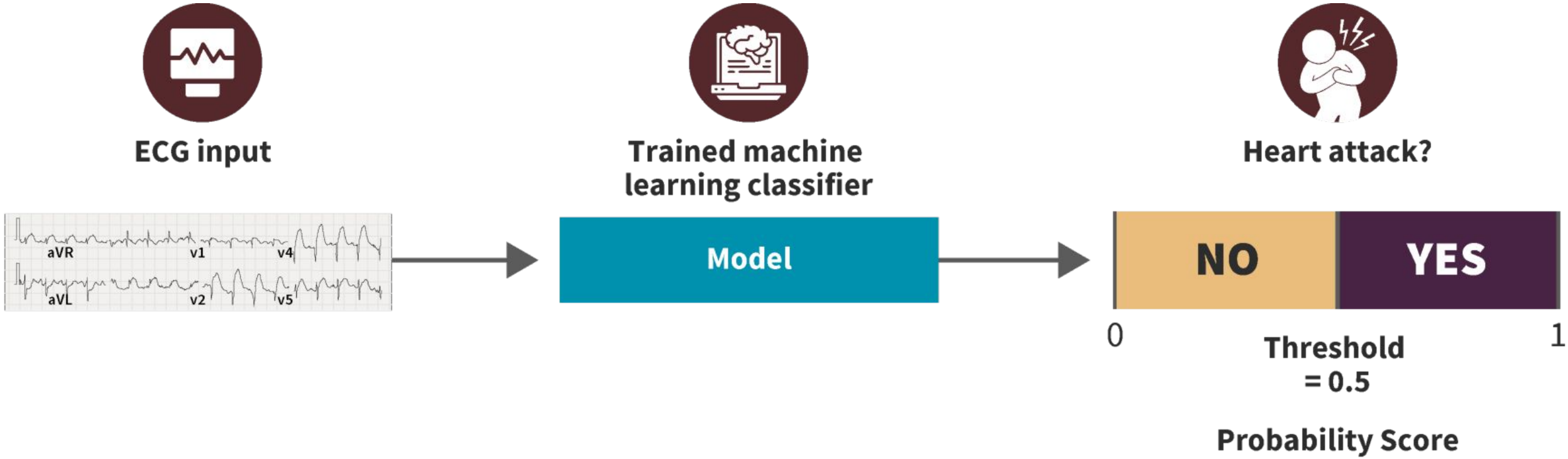






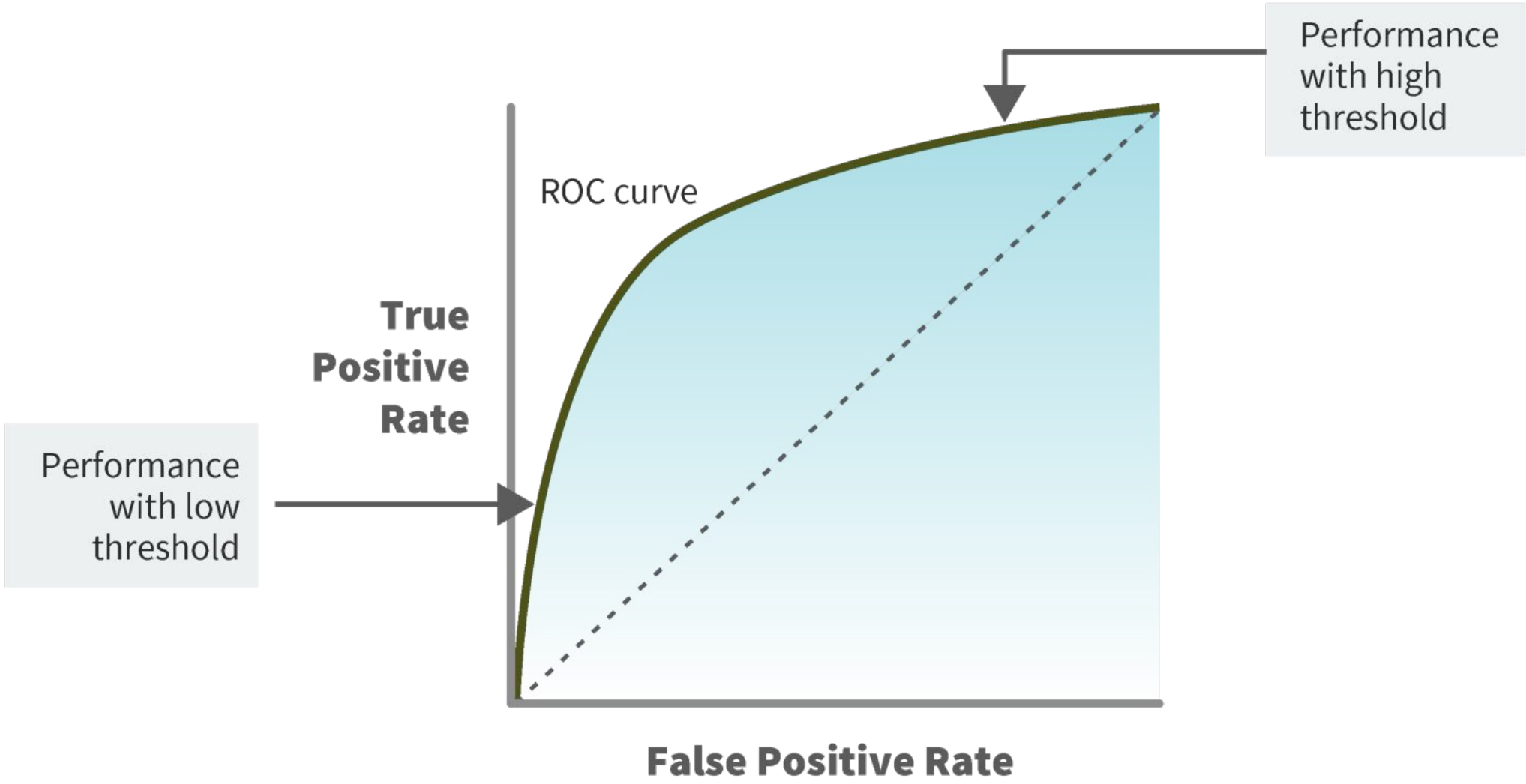
Model Accuracy = 90%!


ECG input
aVR
v1
v4
aVL
v2
v5
Trained machine learning classifier
Model
Heart attack?
0
1
Probability Score

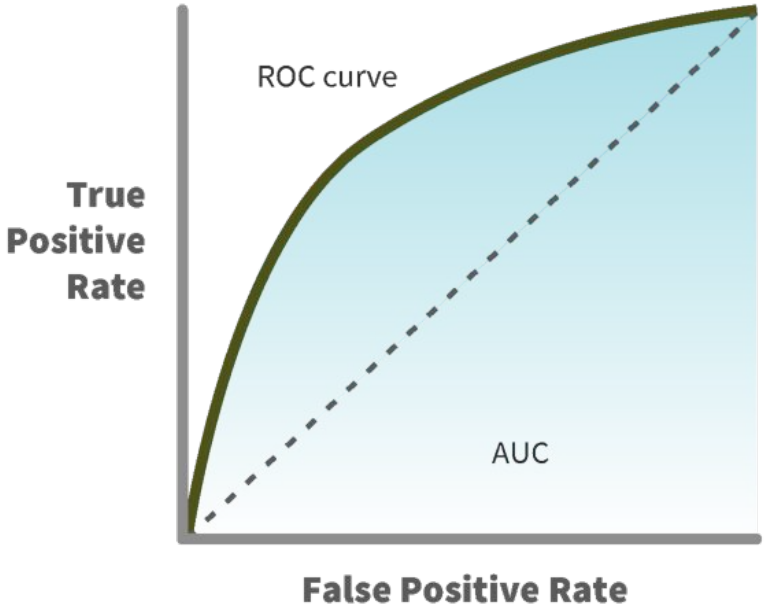

ECG input
aVR
v1
v4
aVL
v2
v5
Trained machine learning classifier
Model
Heart attack?
NO
YES
0
1
Threshold = 0.5
Probability Score

ROC (RECEIVER OPERATING CHARACTERISTIC) CURVES:

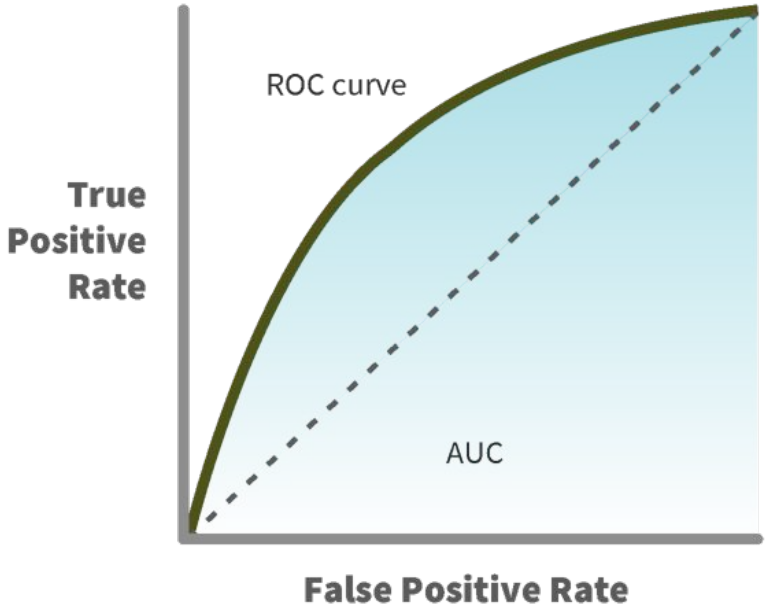
evaluate model performance considering a range of thresholds



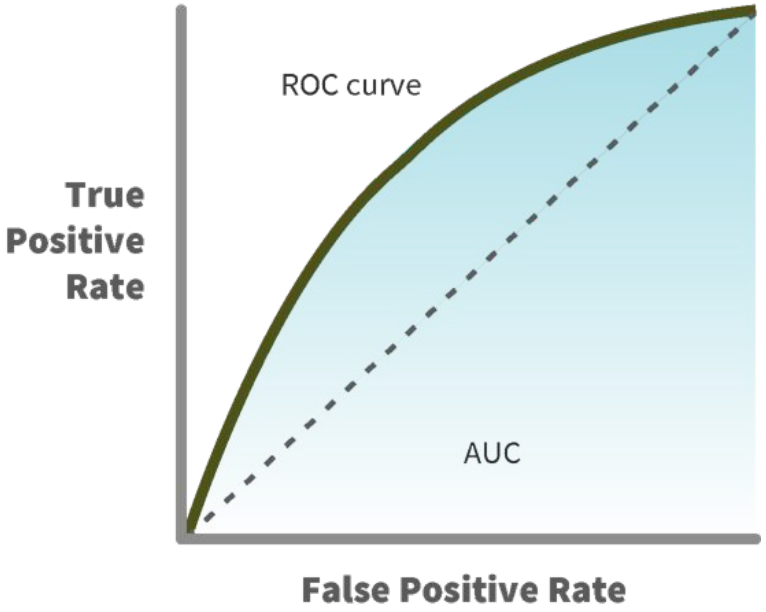
Predicting X vs. not X



Predicting Y vs. not Y



Predicting Z vs. not Z



PREDICTED VALUES

ACTUAL VALUES

	POSITIVE	NEGATIVE
POSITIVE		
NEGATIVE		

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES

	 **POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)
 **POSITIVE** (heart attack)		
NEGATIVE (Not heart attack)		

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES	 **POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
 **POSITIVE** (heart attack)	80	40	Total actual positive = **120**
NEGATIVE (Not heart attack)	20	60	Total actual negative = **80**

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES	 **POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
 **POSITIVE** (heart attack)	80	40	Total actual positive = **120**
NEGATIVE (Not heart attack)	20	60	Total actual negative = **80**
	Total predicted positive = **100**	Total predicted negative = **100**	

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES		**POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
	POSITIVE (heart attack)	80	40	Total actual positive = **120**
	NEGATIVE (Not heart attack)	20	60	Total actual negative = **80**
		Total predicted positive = **100**	Total predicted positive = **100**	

True Positive (TP): Actual positive, and predicted positive

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES		**POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
	**POSITIVE** (heart attack)	80	40	Total actual positive = **120**
	NEGATIVE (Not heart attack)	20	60	Total actual negative = **80**
		Total predicted positive = **100**	Total predicted positive = **100**	

True Positive (TP): Actual positive, and predicted positive

True Negative (TN): Actual negative, and predicted negative

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES	 **POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
 **POSITIVE** (heart attack)	80	40	Total actual positive = **120**
NEGATIVE (Not heart attack)	20	60	Total actual negative = **80**
	Total predicted positive = **100**	Total predicted positive = **100**	

True Positive (TP): Actual positive, and predicted positive

True Negative (TN): Actual negative, and predicted negative

False Positive (FP): Actual negative, but predicted positive

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES		**POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
	POSITIVE (heart attack)	80	40	Total actual positive = **120**
	NEGATIVE (Not heart attack)	20	60	Total actual negative = **80**
		Total predicted positive = **100**	Total predicted positive = **100**	

True Positive (TP): Actual positive, and predicted positive	**False Positive (FP):** Actual negative, but predicted positive
True Negative (TN): Actual negative, and predicted negative	**False Negative (FN):** Actual positive, but predicted negative

CONFUSION MATRIX

PREDICTED VALUES

ACTUAL VALUES	 **POSITIVE** (heart attack)	**NEGATIVE** (Not heart attack)	
 **POSITIVE** (heart attack)	TP = 80	FN = 40	Total actual positive = **120**
NEGATIVE (Not heart attack)	FP = 20	TN = 60	Total actual negative = **80**
	Total predicted positive = **100**	Total predicted positive = **100**	

True Positive (TP): Actual positive, and predicted positive	**False Positive (FP):** Actual negative, but predicted positive
True Negative (TN): Actual negative, and predicted negative	**False Negative (FN):** Actual positive, but predicted negative

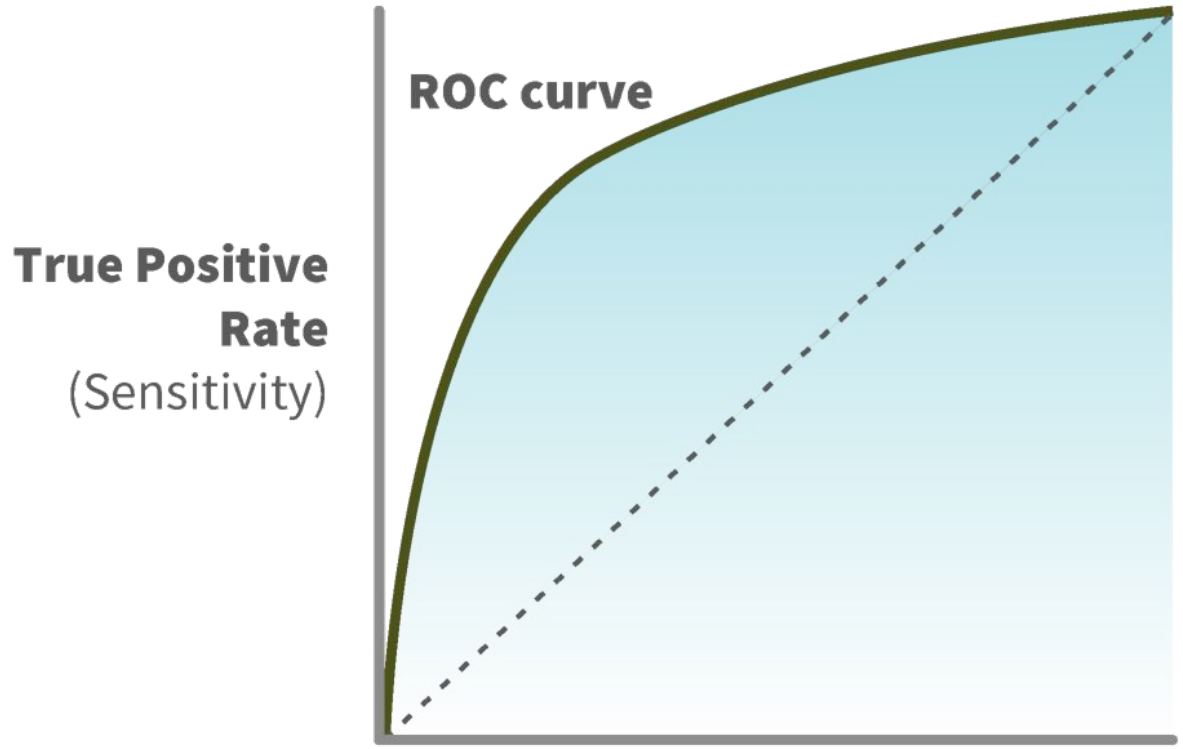
PREDICTED VALUES

ACTUAL VALUES	POSITIVE (heart attack)	NEGATIVE (Not heart attack)	
POSITIVE (heart attack)	TP = 80	FN = 40	Total actual positive = **120**
NEGATIVE (Not heart attack)	FP = 20	TN = 60	Total actual negative = **80**
	Total predicted positive = **100**	Total predicted positive = **100**	

METRIC	FORMULA	ABBREVIATION	CALCULATED VALUE
Prevalence	Total actual positive/total	PREV	120/200 = 0.6
Accuracy	TP+TN / total	ACC	140 / 200 = 0.7
Sensitivity True positive rate Recall	TP/total actual positive	SN TPR REC	80 / 120 = 0.67
Specificity True negative rate	TN/total actual negative	SN TNR	60 / 80 = 0.75
Precision Positive predictive value	TP/total predicted positive	PREC PPV	80/100 = 0.8
Negative predictive value	TN/total predicted negative	NPV	60/100 = 0.6

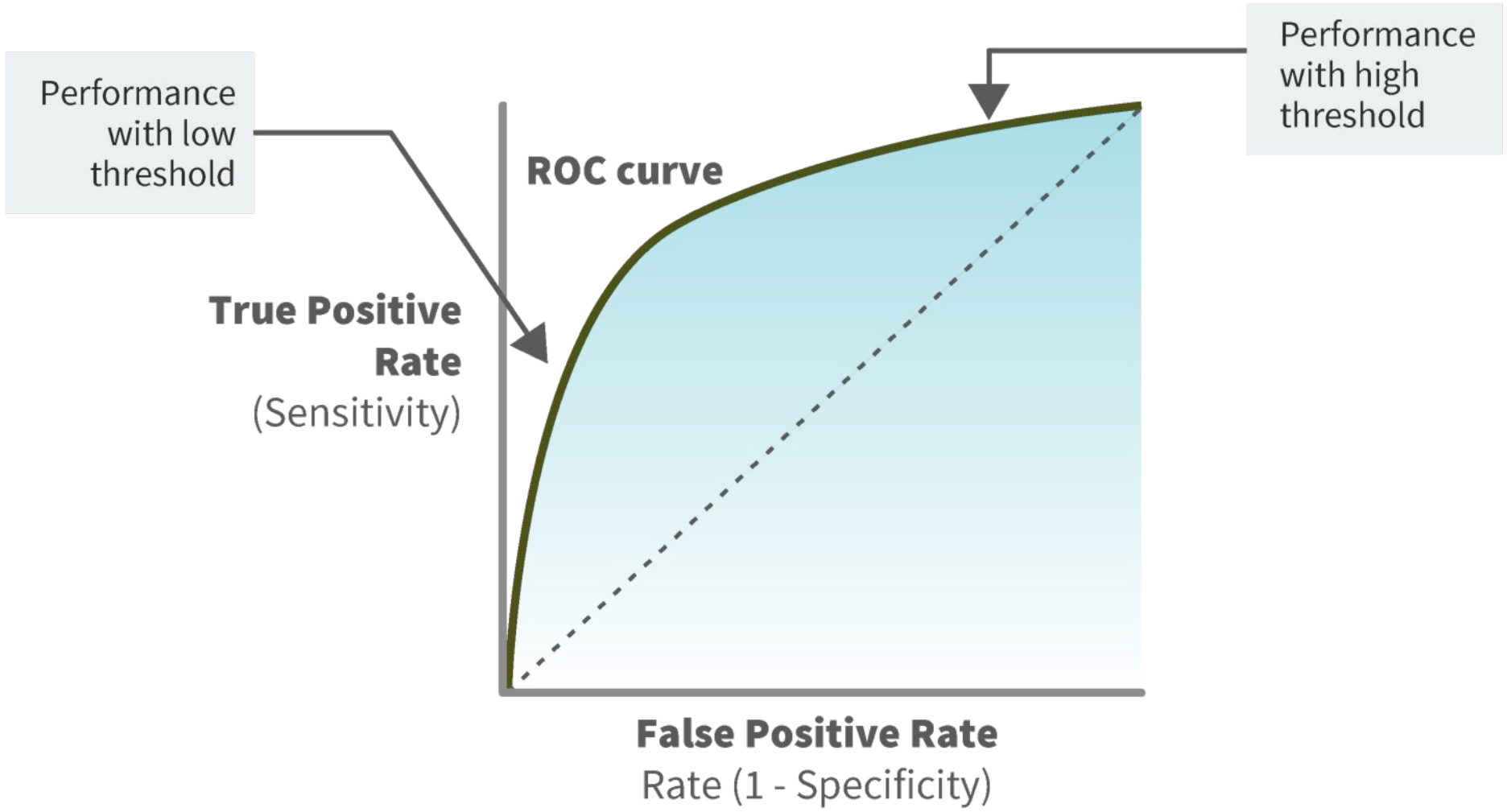
ROC (RECEIVER OPERATING CHARACTERISTIC) CURVES

evaluate model performance considering a range of thresholds

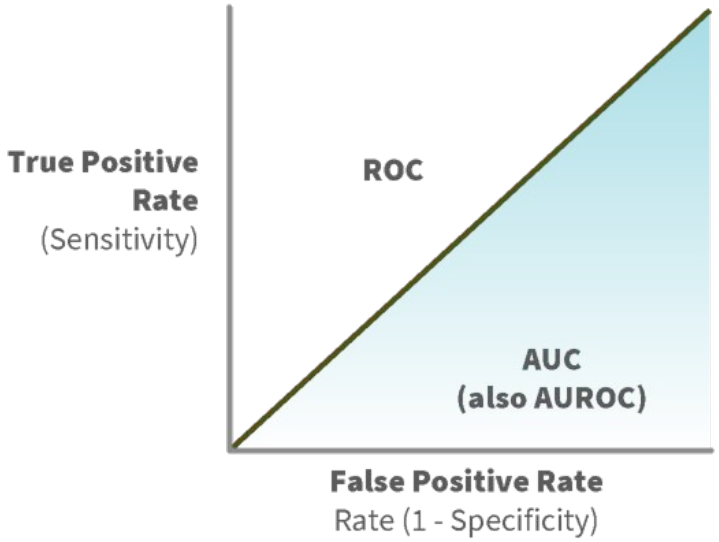


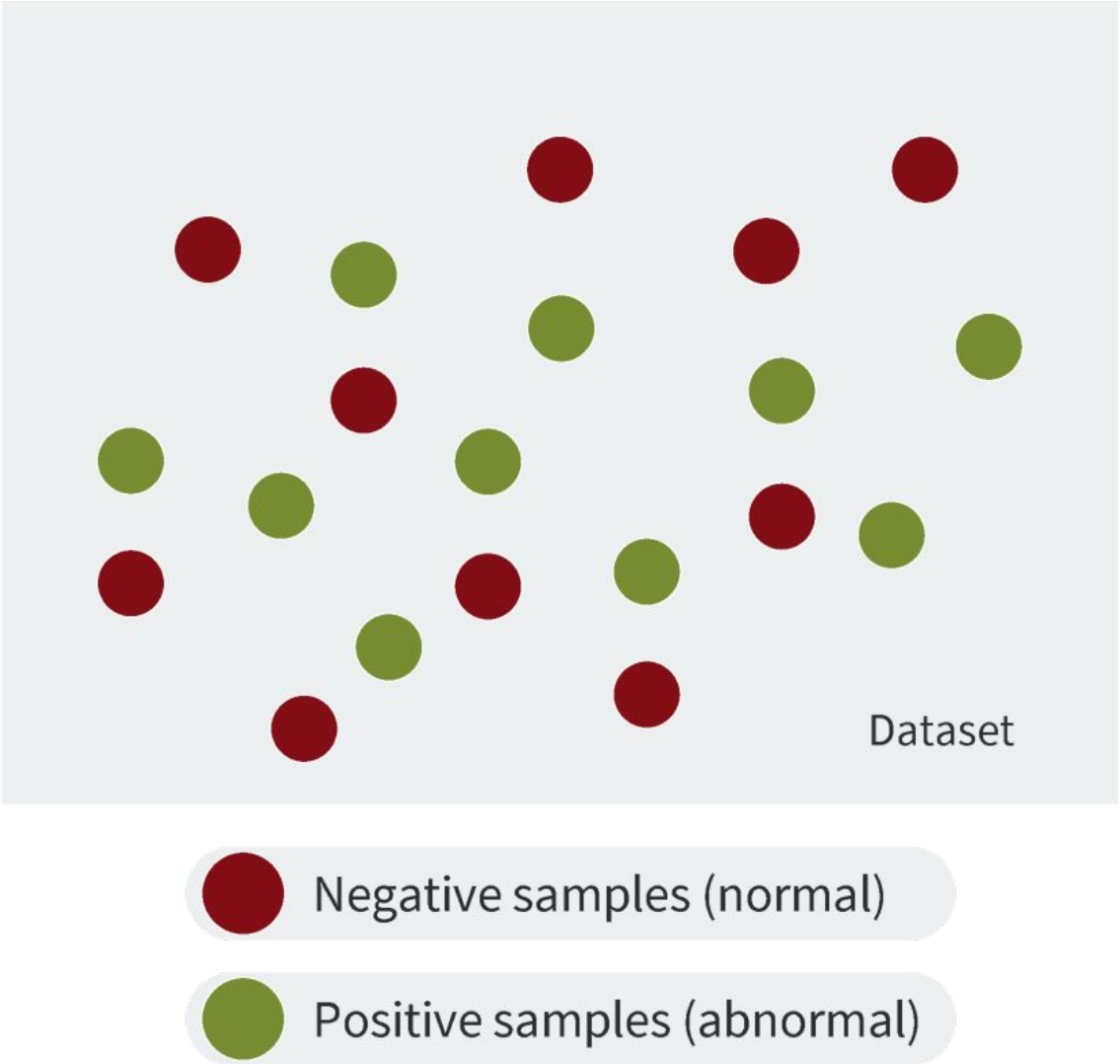
ROC (RECEIVER OPERATING CHARACTERISTIC) CURVES

evaluate model performance considering a range of thresholds

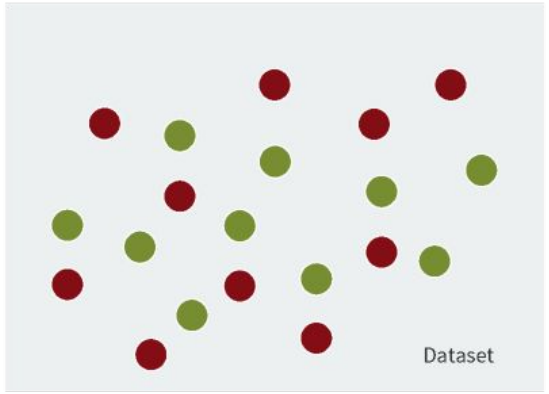


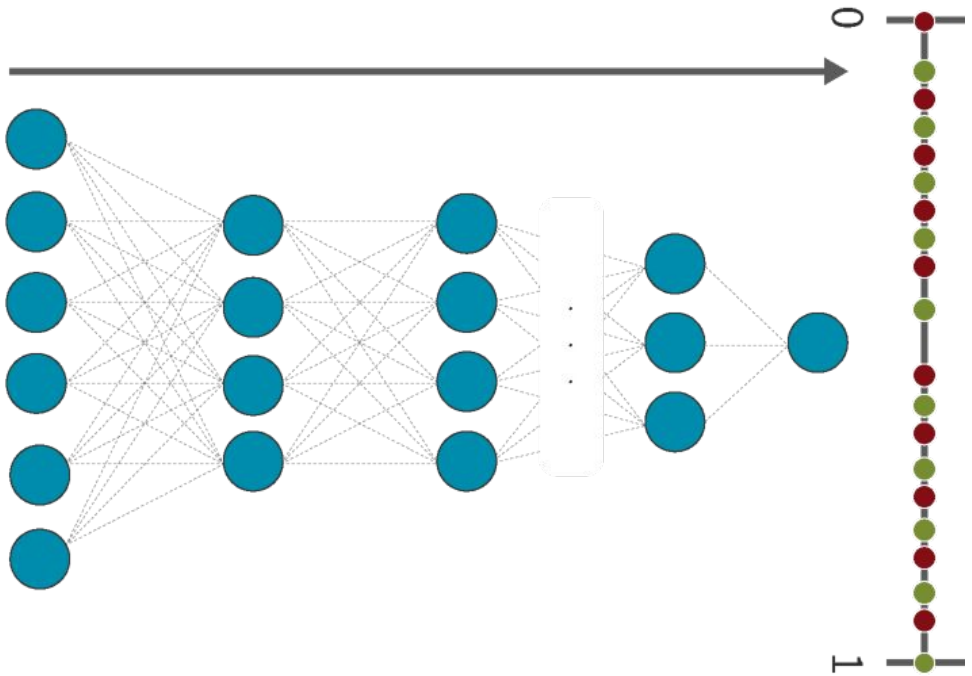
SCENARIO #1: *RANDOM CLASSIFIER (AUROC 0.5)*

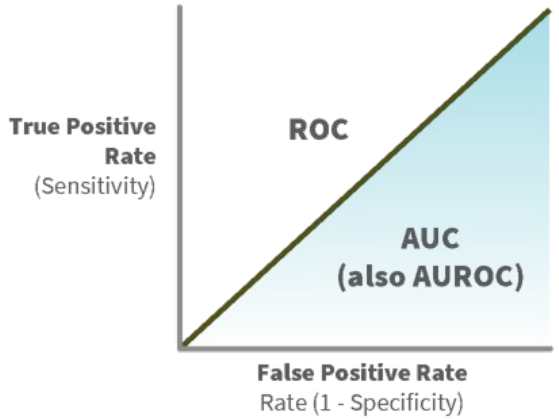




SCENARIO #1: *RANDOM CLASSIFIER (AUROC 0.5)*

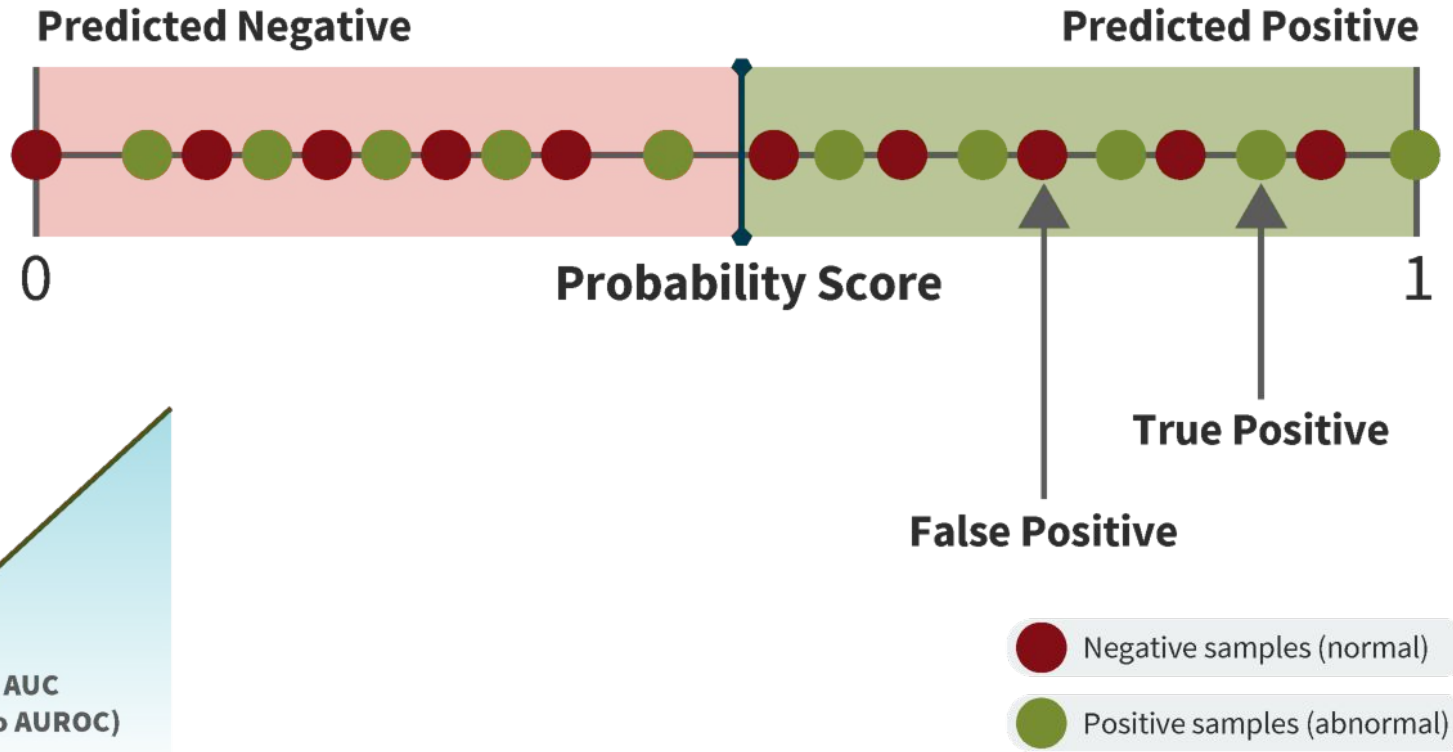


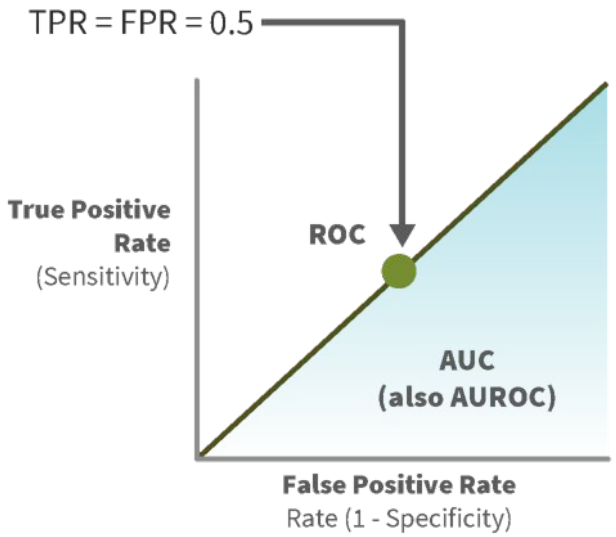




SCENARIO #1: *RANDOM CLASSIFIER (AUROC 0.5)*

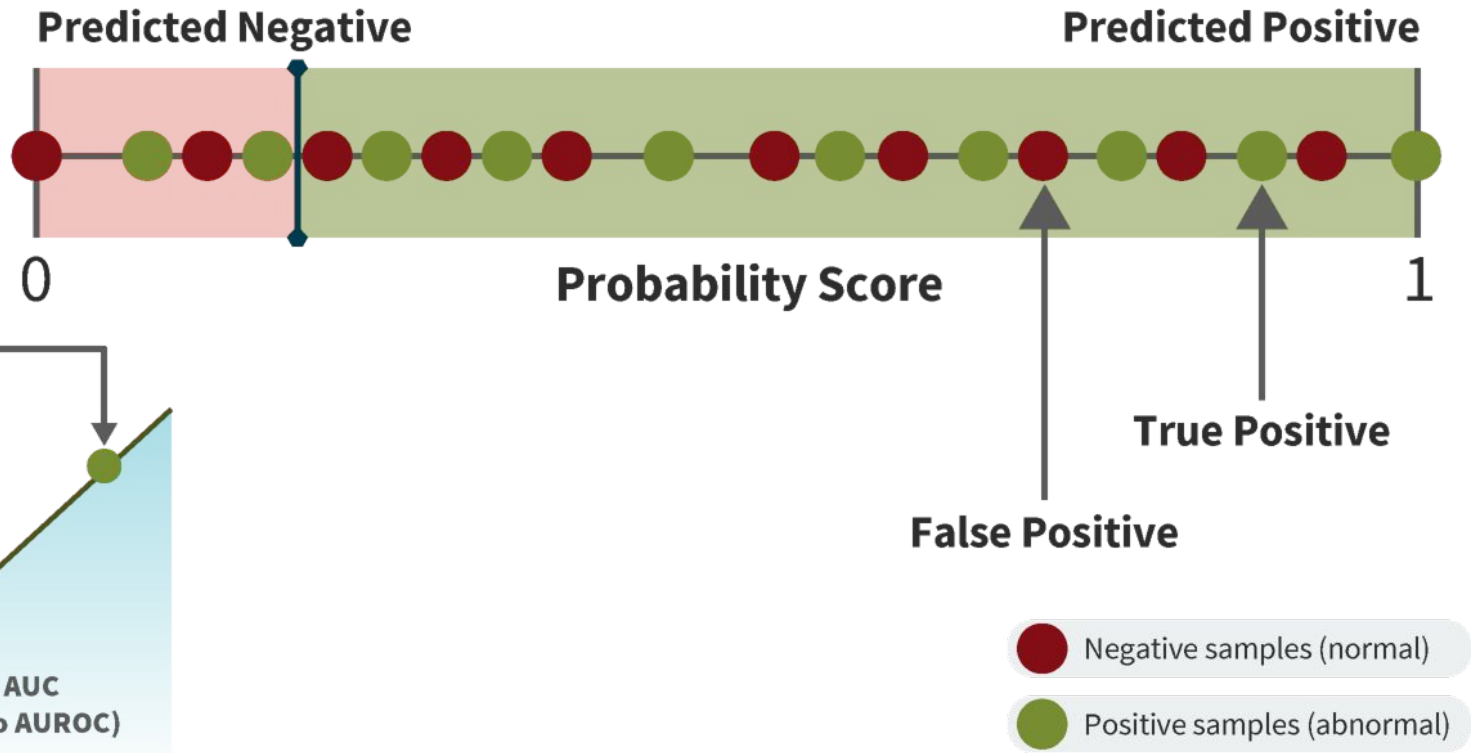
Threshold: 0.5
TPR: 0.5
FPR: 0.5

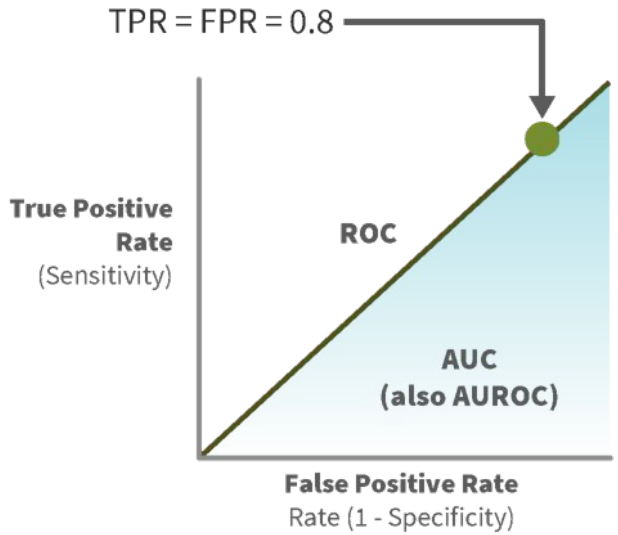




SCENARIO #1: *RANDOM CLASSIFIER (AUROC 0.5)*

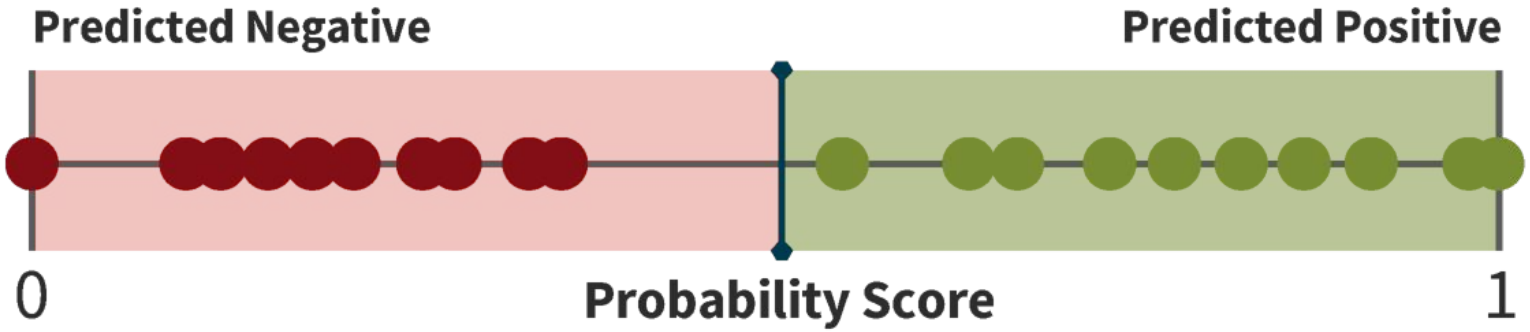
Threshold: 0.25
TPR: 0.8
FPR: 0.8

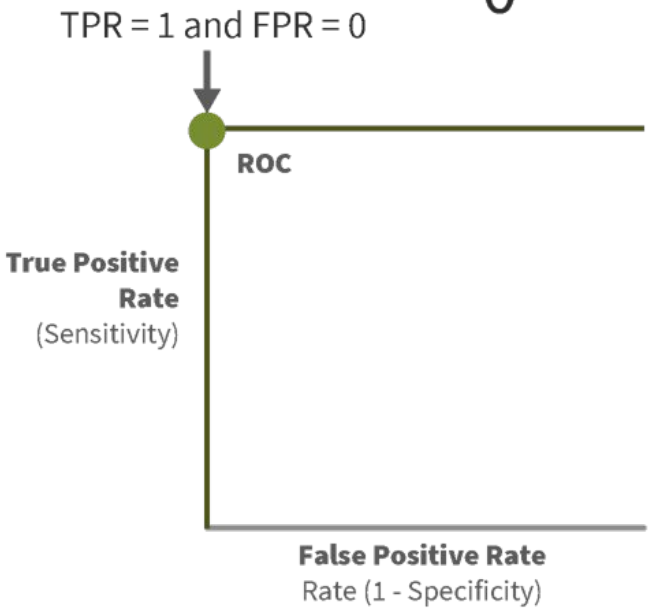




SCENARIO #2: *Perfect Classifier (AUROC 1.0)*

Threshold: 0.5
TPR: 1.0
FPR: 0.5



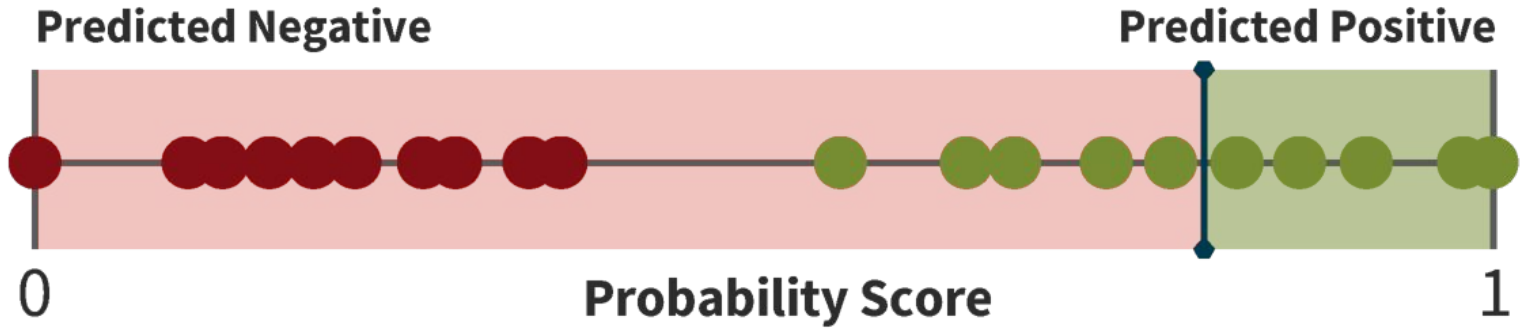


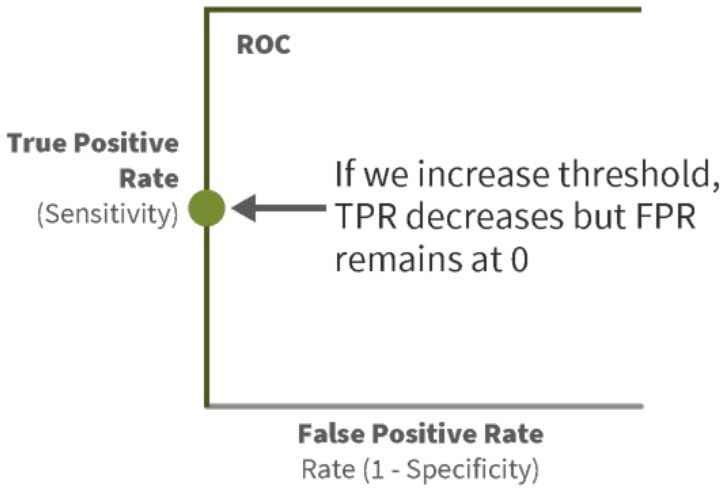
Negative samples (normal)

Positive samples (abnormal)

SCENARIO #2: *Perfect Classifier (AUROC 1.0)*

Threshold: 0.8
TPR: 0.5
FPR: 0.0



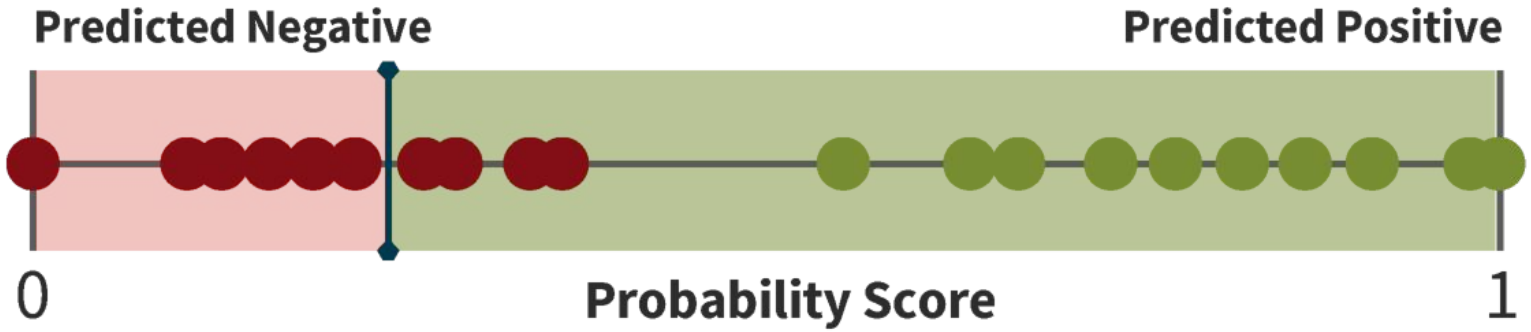


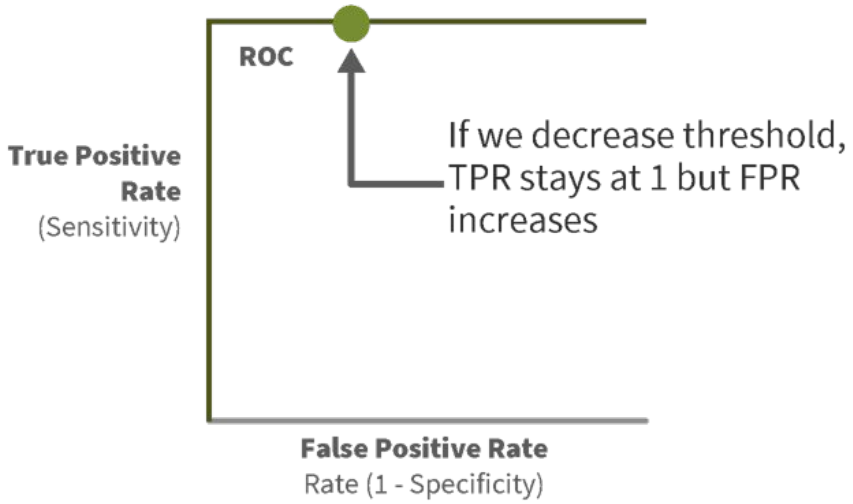
Negative samples (normal)

Positive samples (abnormal)

SCENARIO #2: *Perfect Classifier (AUROC 1.0)*

Threshold: 0.2
TPR: 1.0
FPR: 0.4

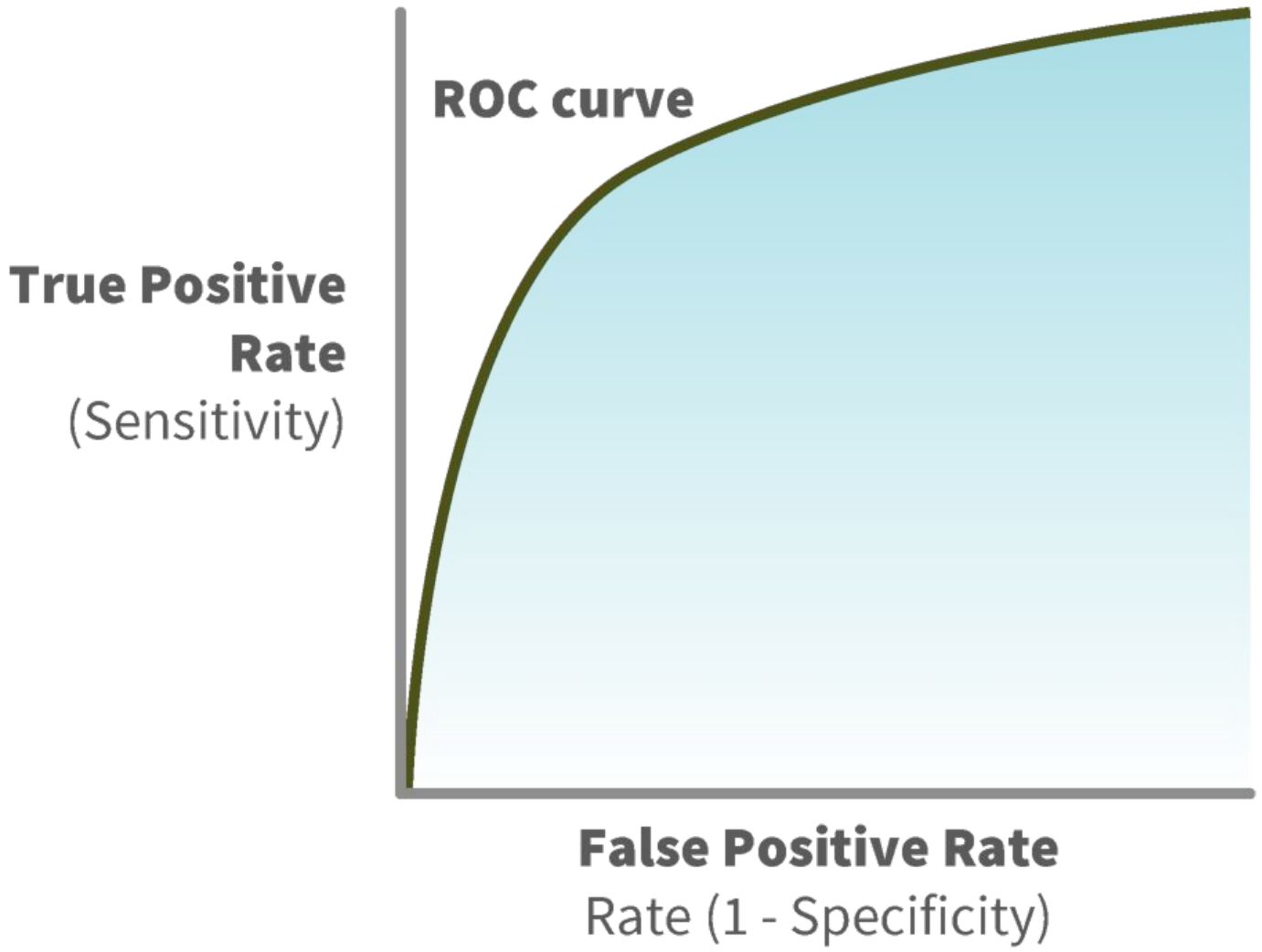




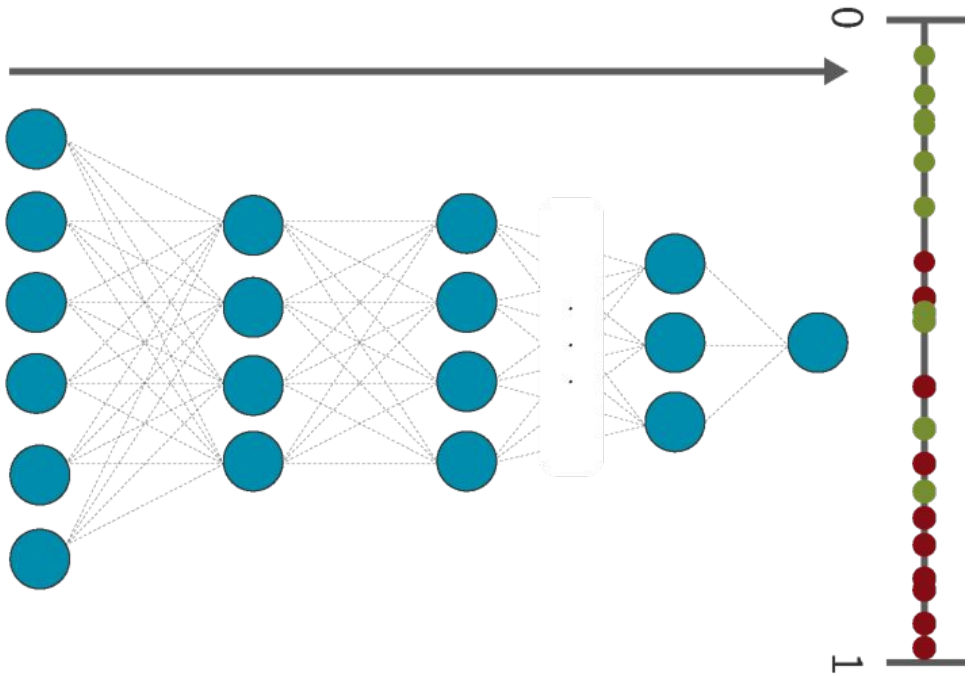
Negative samples (normal)

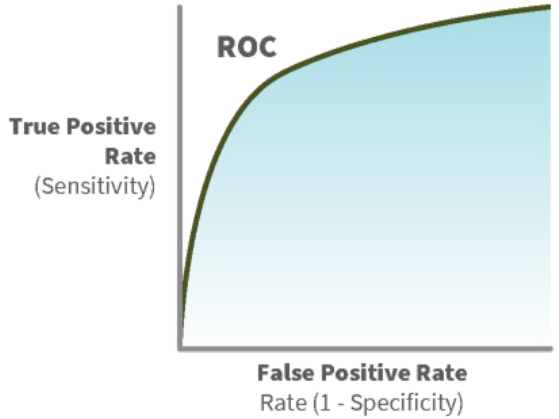
Positive samples (abnormal)

SCENARIO #3: *"Good" Classifier (AUROC 0.9)*



SCENARIO #3: *"Good" Classifier (AUROC 0.9)*

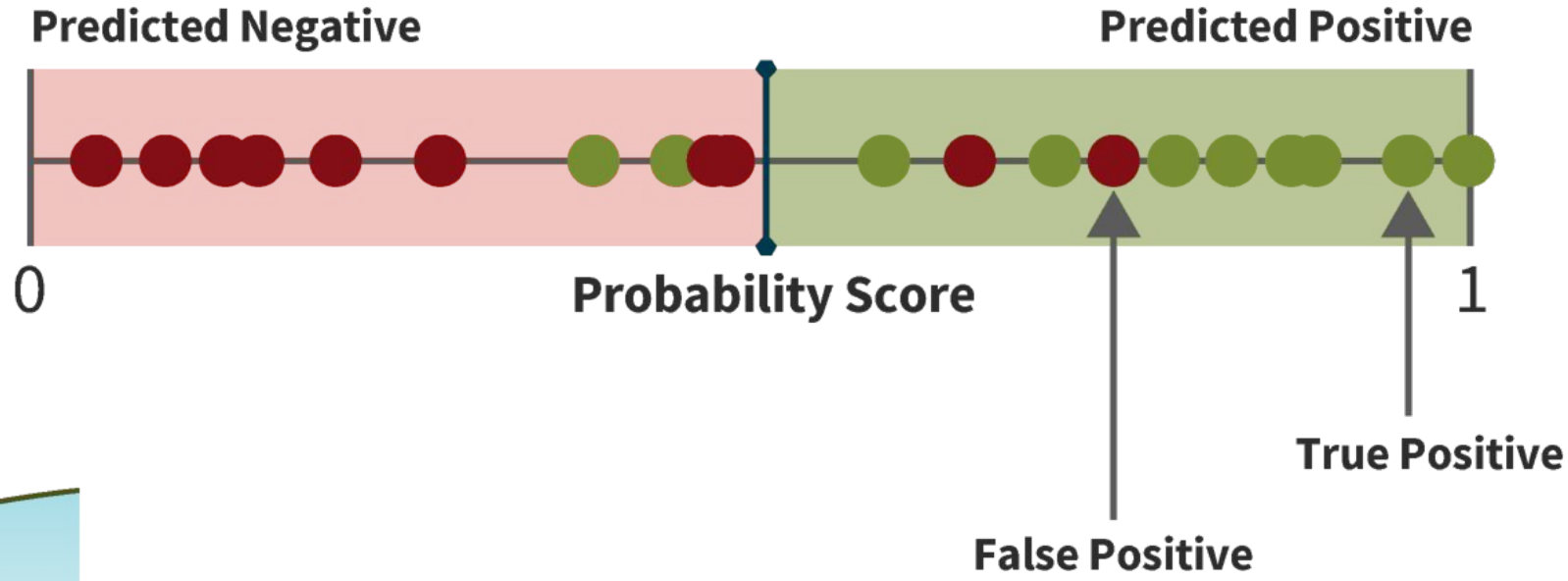


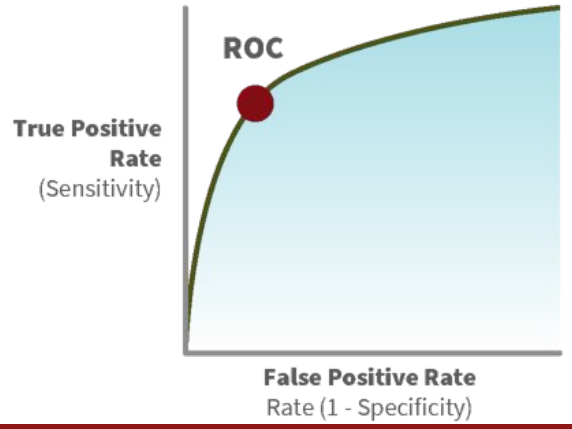


The model cannot perfectly discriminate between the two classes, but the samples are largely separable.

SCENARIO #3: *"Good" Classifier (AUROC 0.9)*

Threshold: 0.50
TPR: 0.8
FPR: 0.2



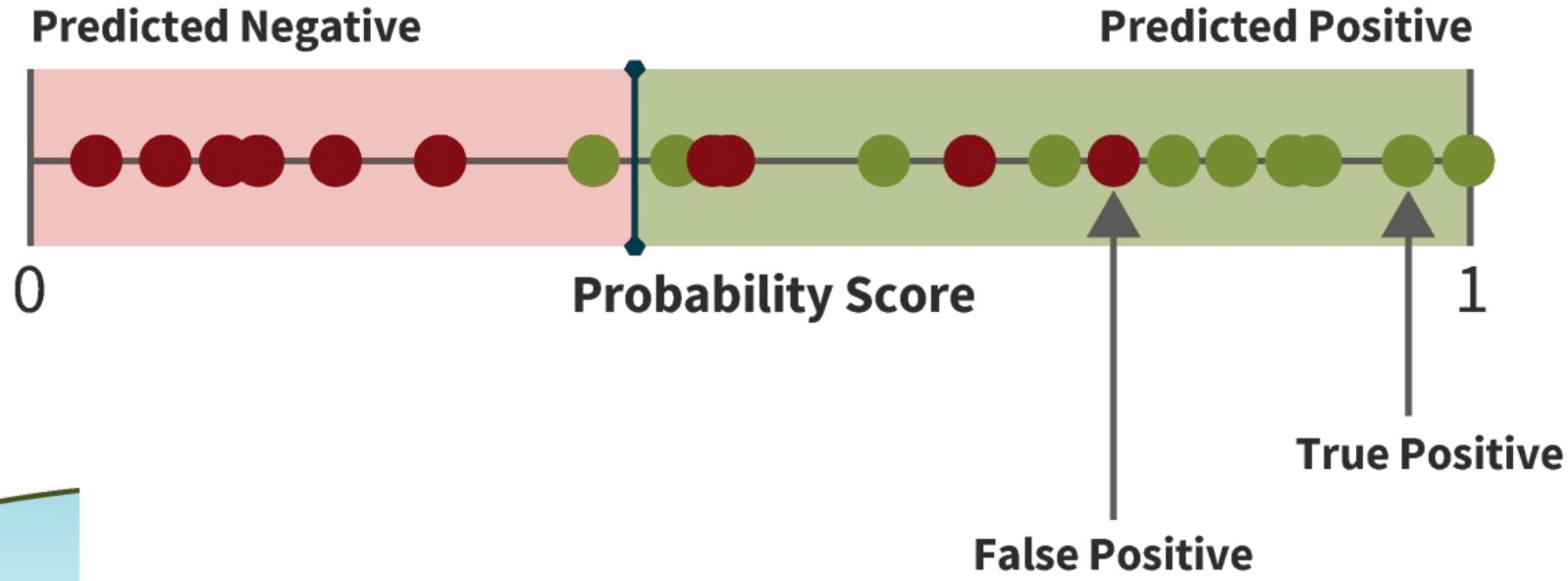


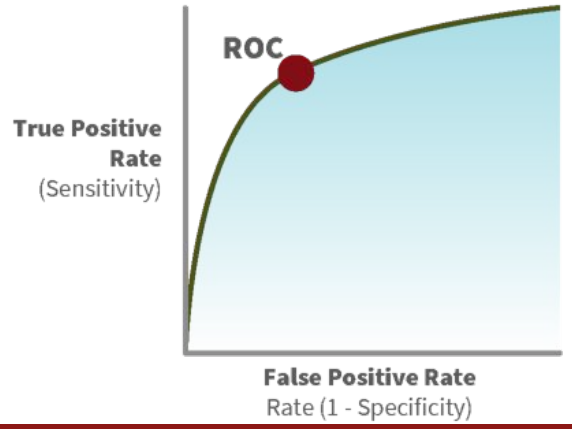
Negative samples (normal)

Positive samples (abnormal)

SCENARIO #3: *"Good" Classifier (AUROC 0.9)*

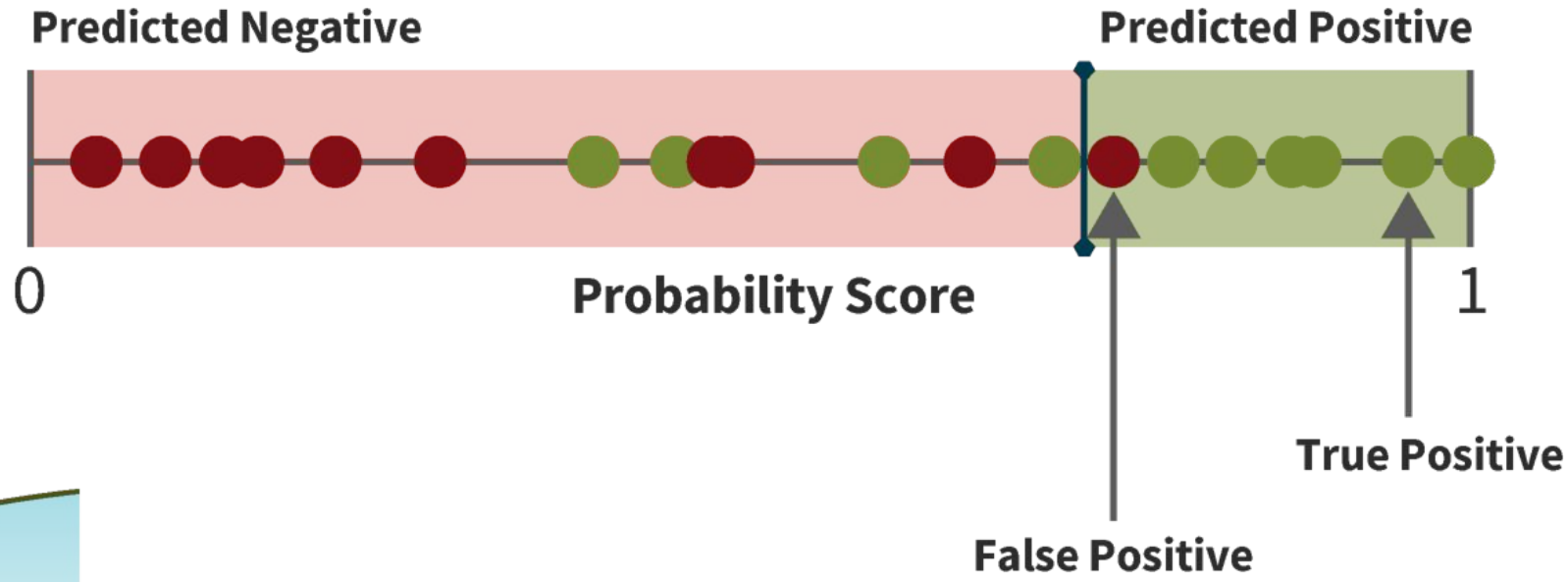
Threshold: 0.4
TPR: 0.9
FPR: 0.4

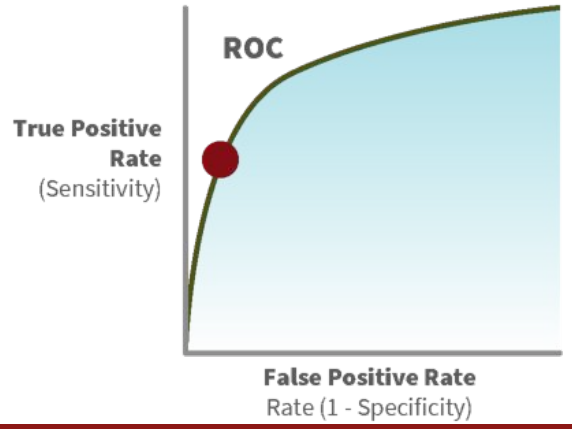




SCENARIO #3: *"Good" Classifier (AUROC 0.9)*

Threshold: 0.7
TPR: 0.6
FPR: 0.1



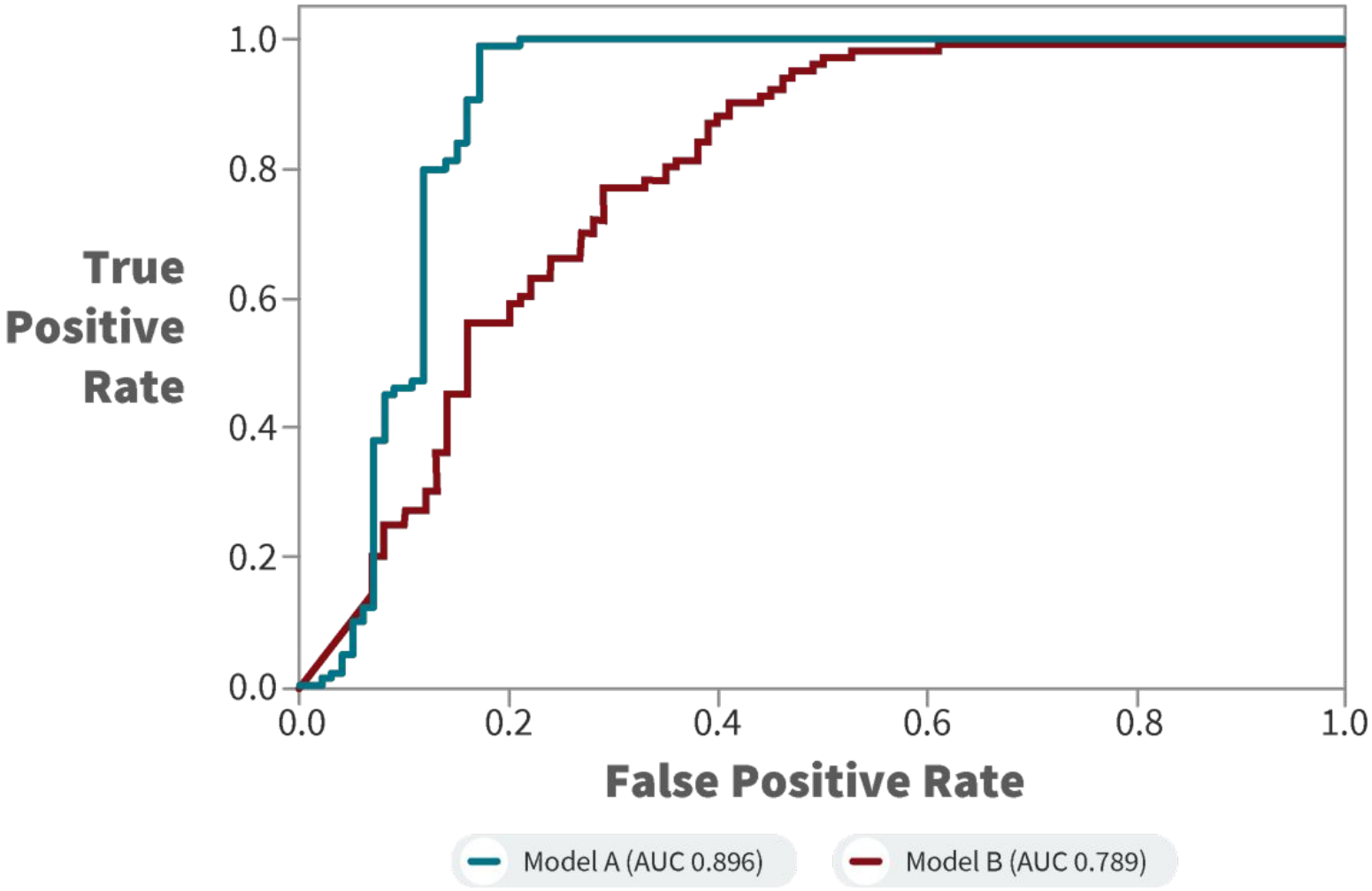


There is a fundamental tradeoff between TPR and FPR. The selected threshold is known as the **operating point**

Negative samples (normal)

Positive samples (abnormal)

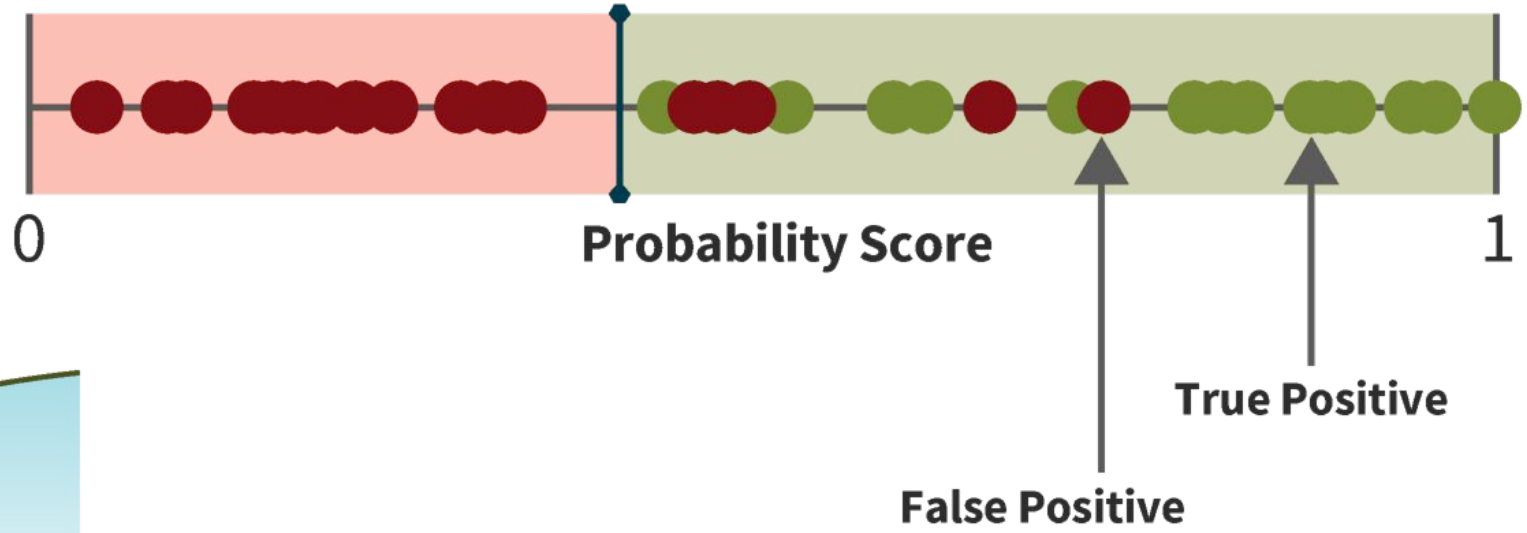
COMPARING CLASSIFIERS

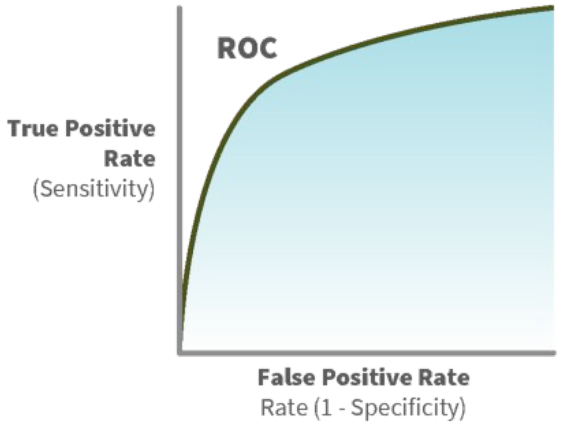


CHOOSING AN OPERATING POINT

Task: Screening for cancer

Metric	Priority
True Positive Rate False Positive Rate	High Medium-low





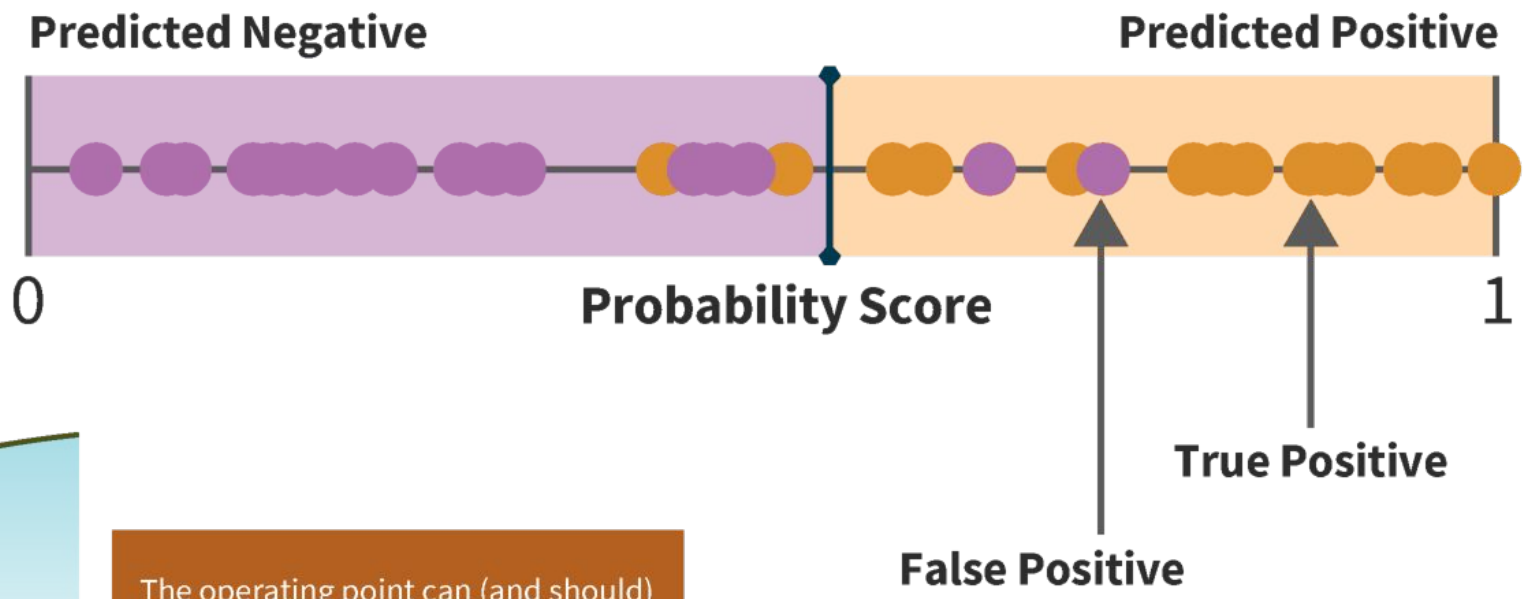
Negative samples (normal)

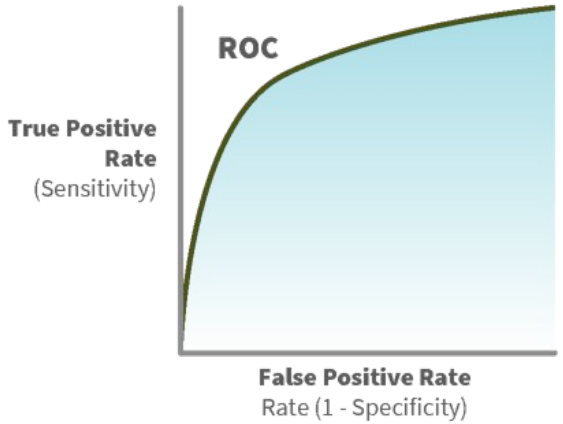
Positive samples (abnormal)

CHOOSING AN OPERATING POINT

Task: Pursuing a high-risk treatment

Metric	Priority
True Positive Rate	Medium
False Positive Rate	High



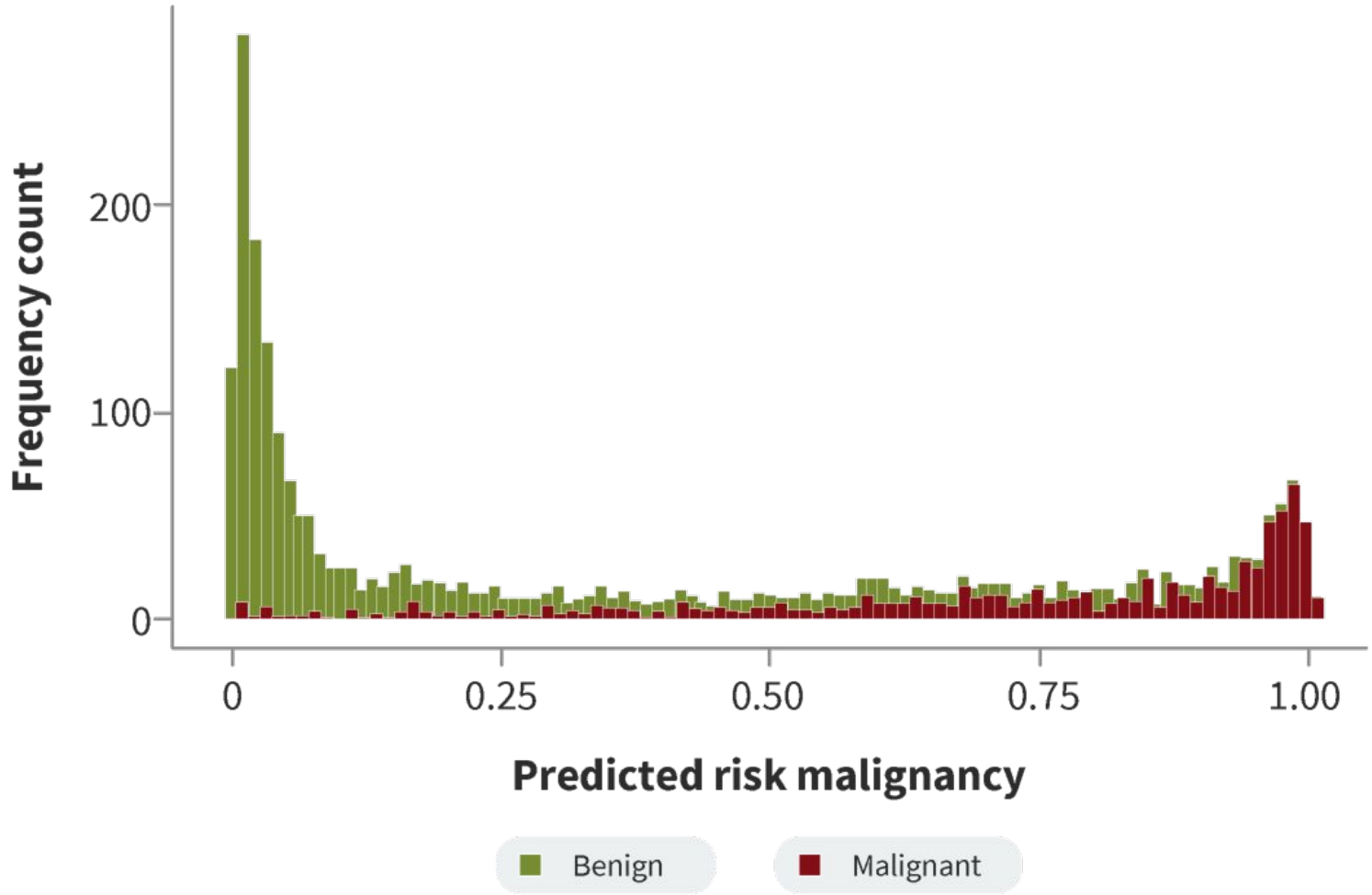


The operating point can (and should) change based on the task.

Negative samples

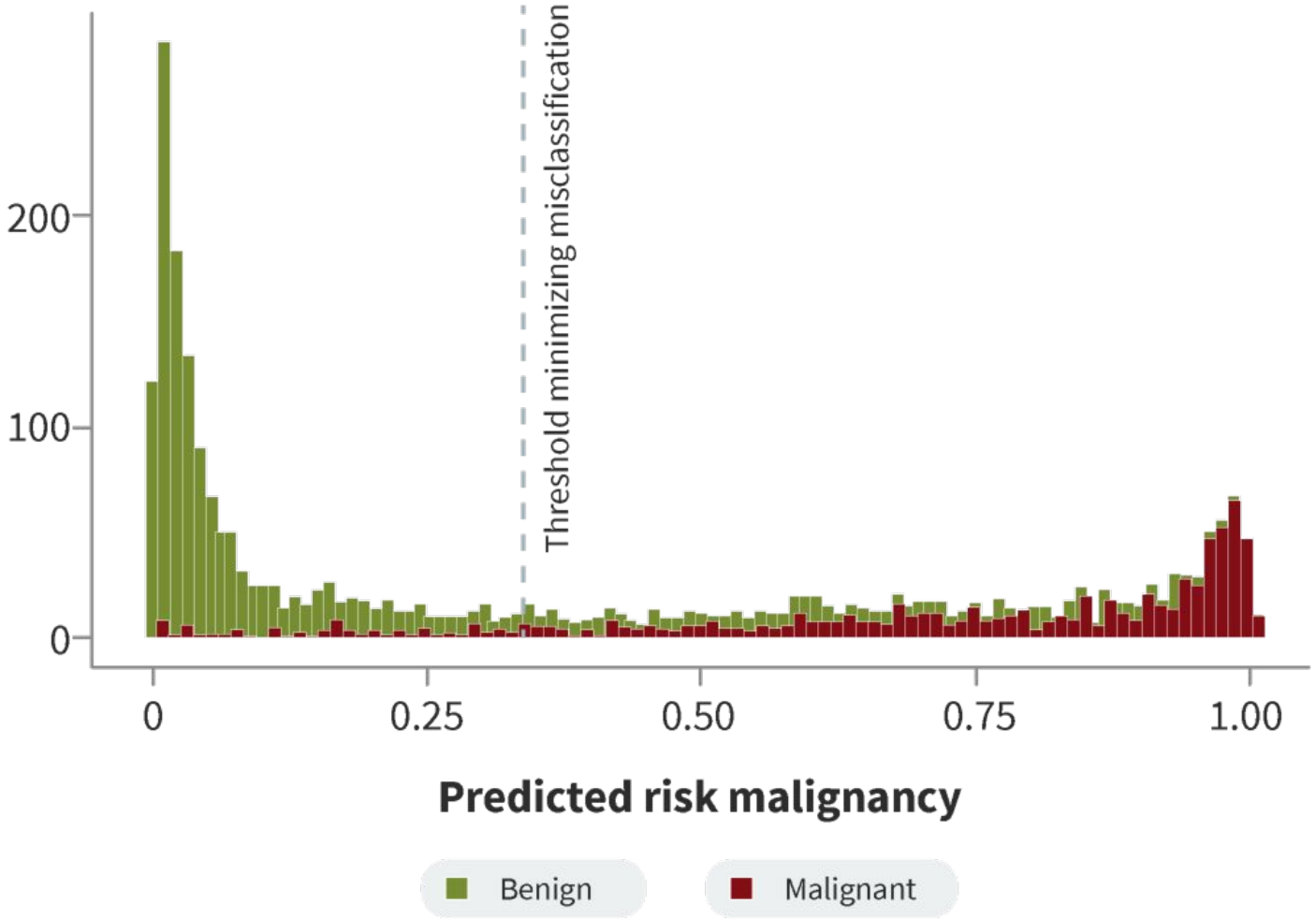
Positive samples

CHOOSING AN OPERATING POINT



Frequencies of predicted risk of malignancy and three possible risk thresholds

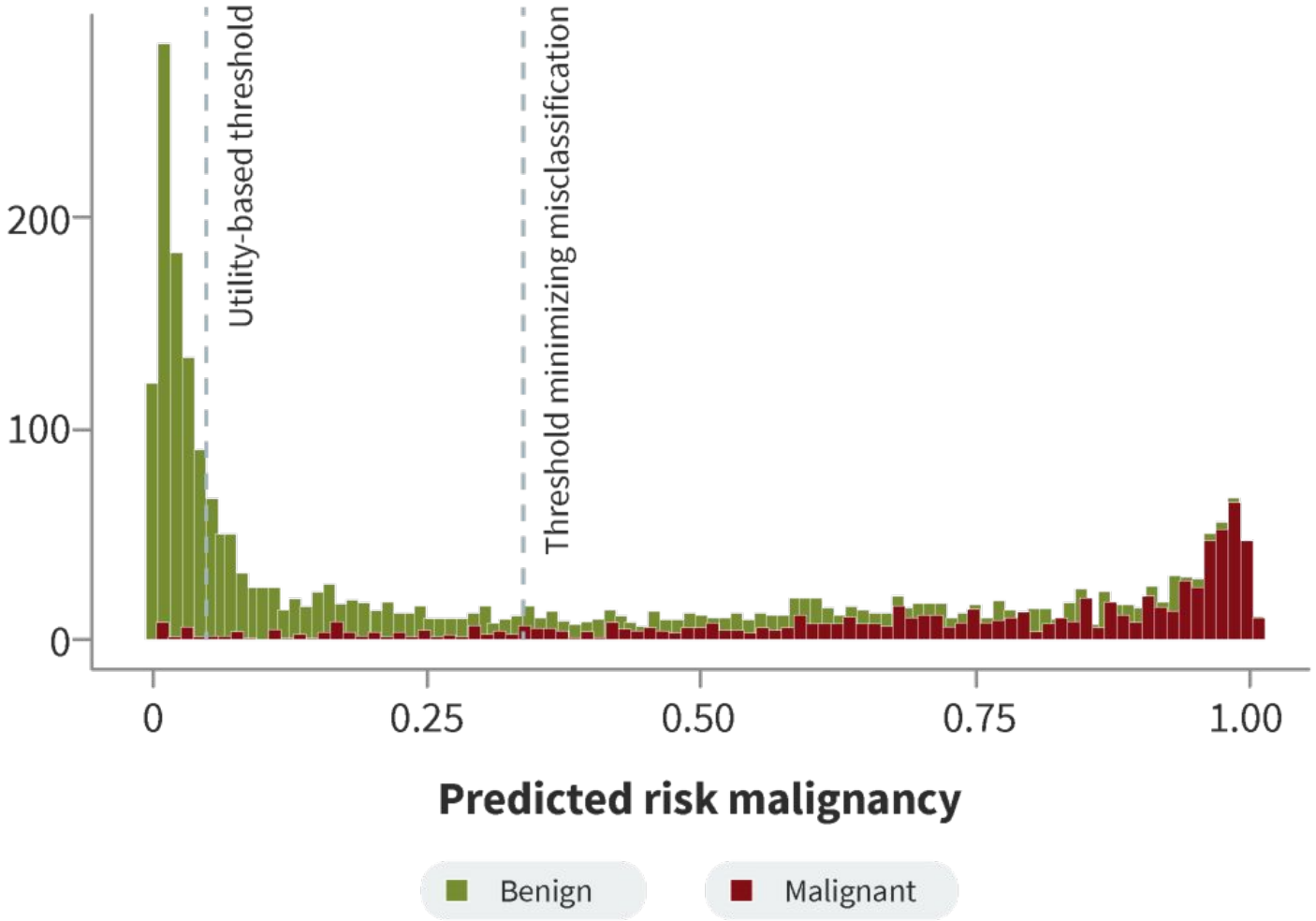
CHOOSING AN OPERATING POINT



Frequencies of predicted risk of malignancy and possible risk thresholds

1. By **Minimizing Misclassification**: choose an operating point that minimizes the sum of false positives and false negatives.

OPERATING POINT HEURISTICS

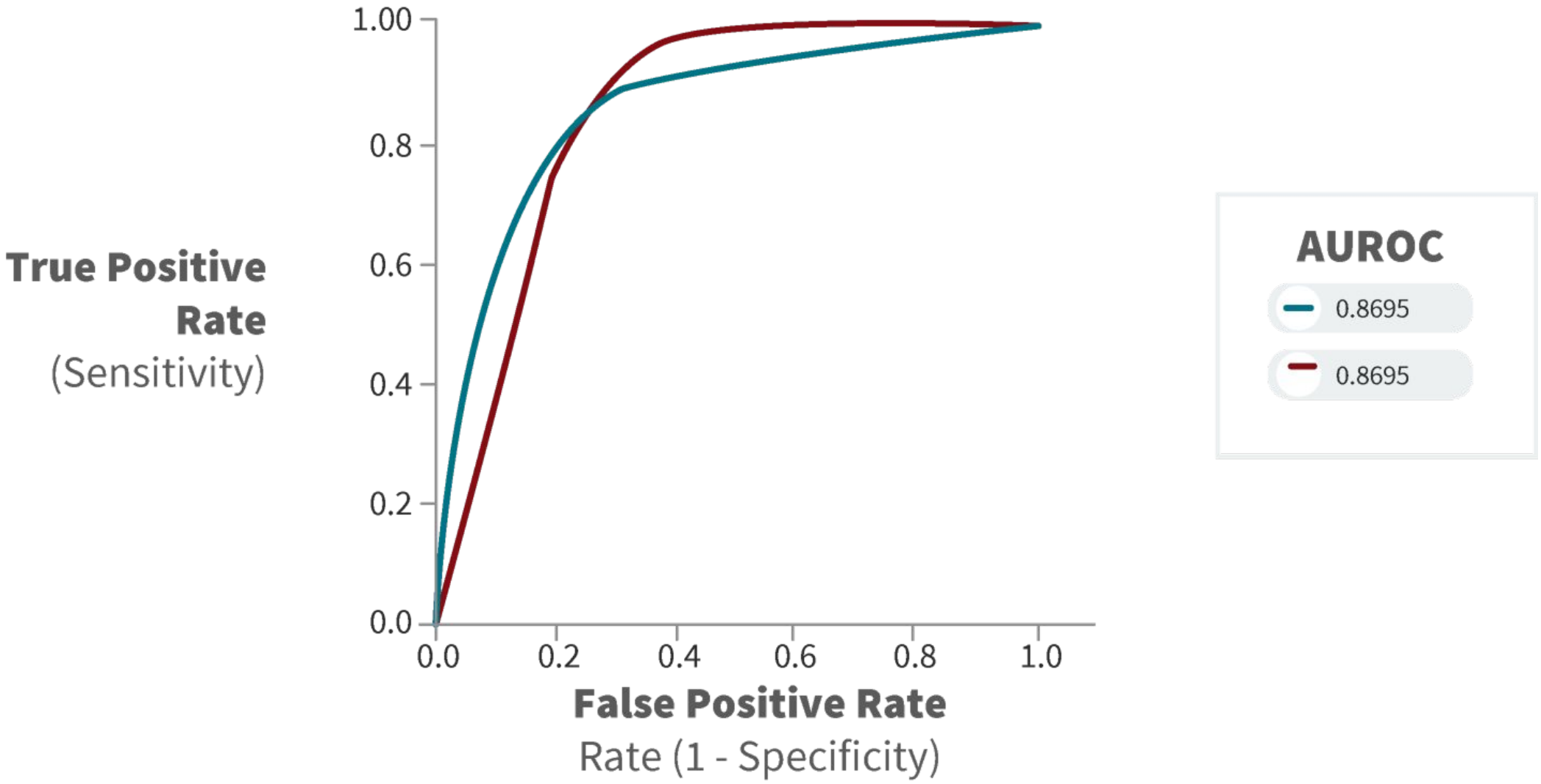


Frequencies of predicted risk of malignancy and possible risk thresholds

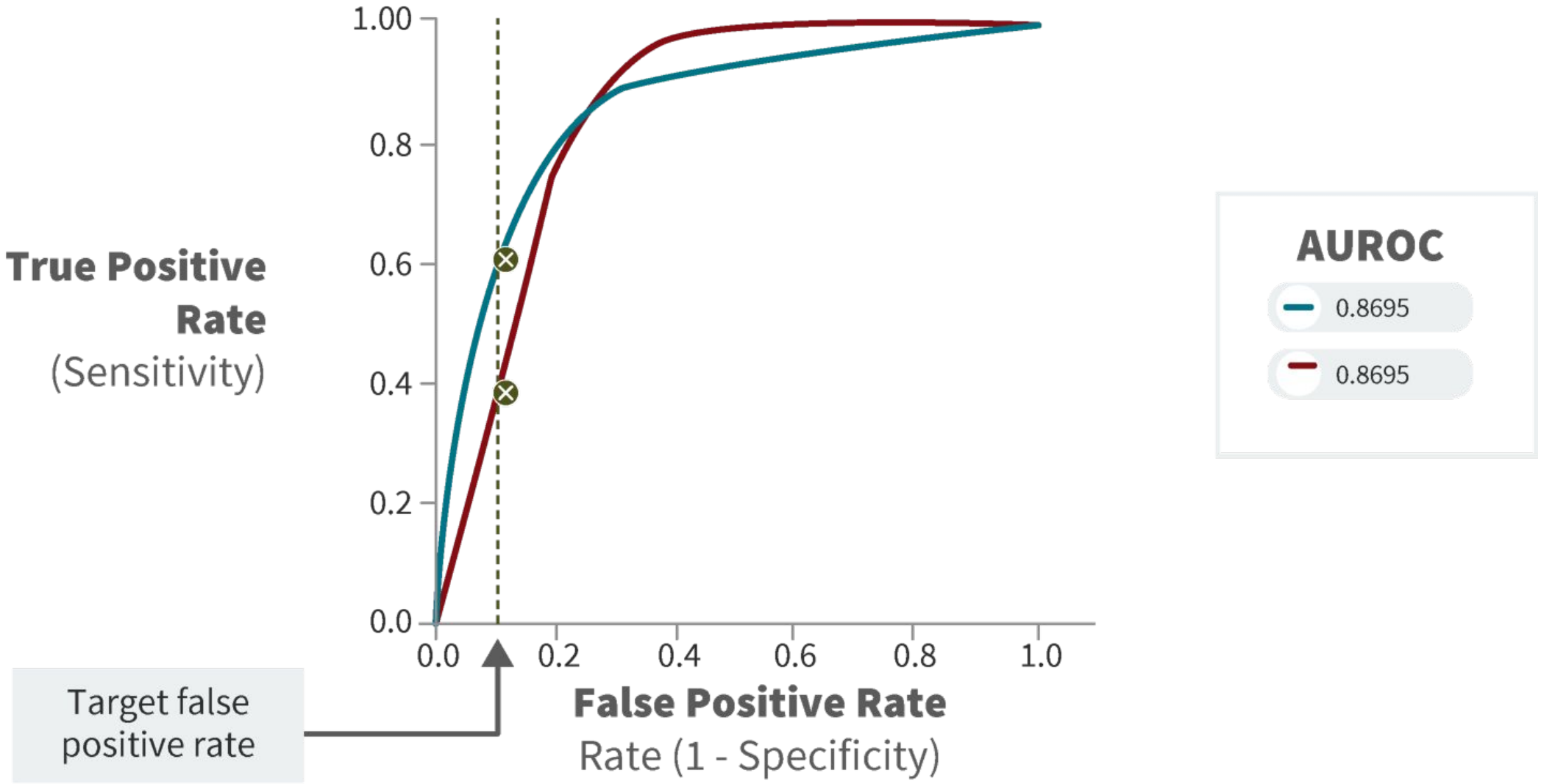
1. By **Minimizing Misclassification**: choose an operating point that minimizes the sum of false positives and false negatives.

2. By **Optimizing Utility:** assign a different utility value for each type of outcome and optimize total utility. NOTE: the utility values are usually subjective!

CONSIDER: *SAME AUROC, DIFFERENT CURVES*

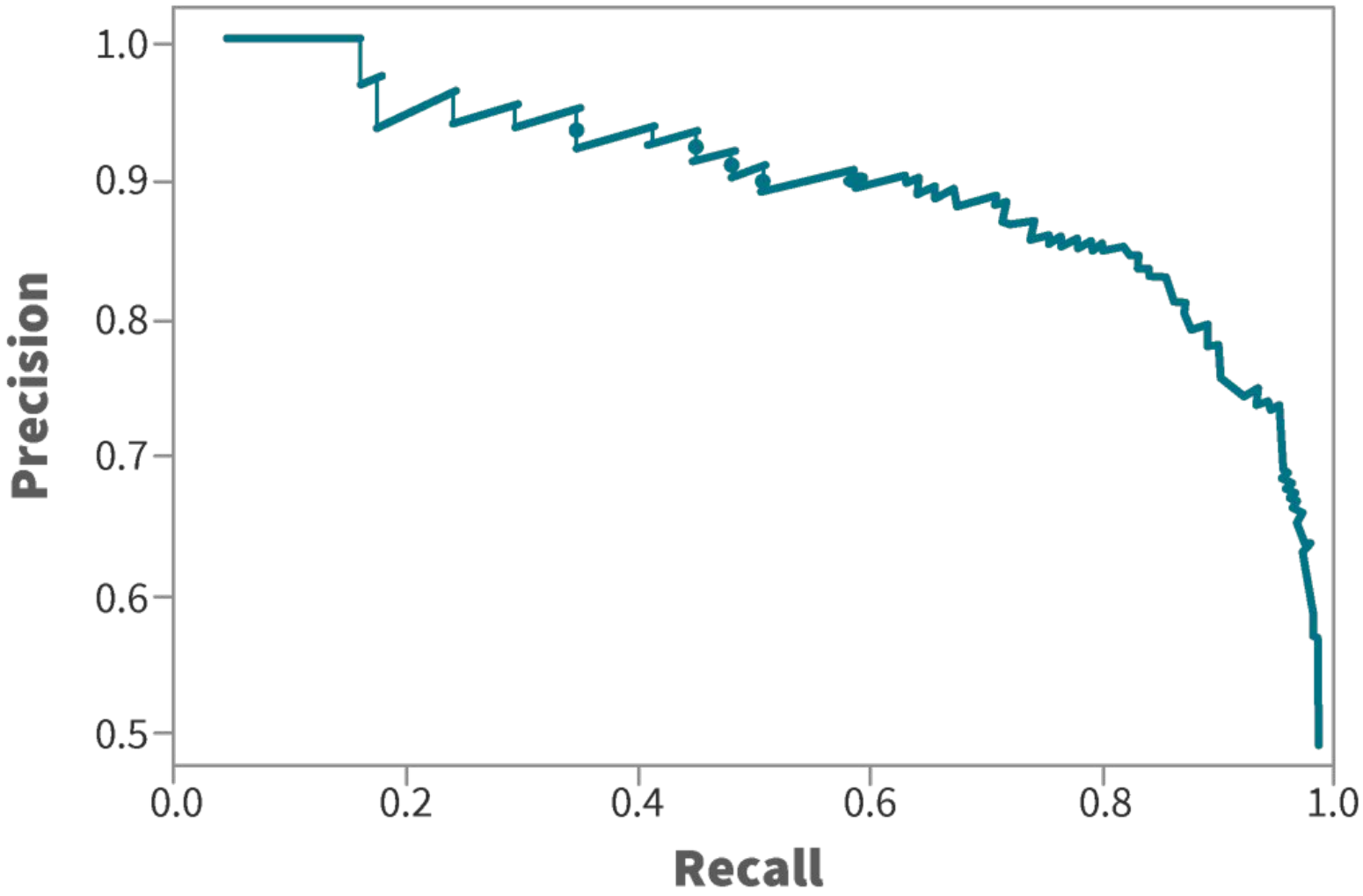


CONSIDER: *SAME AUROC, DIFFERENT CURVES*



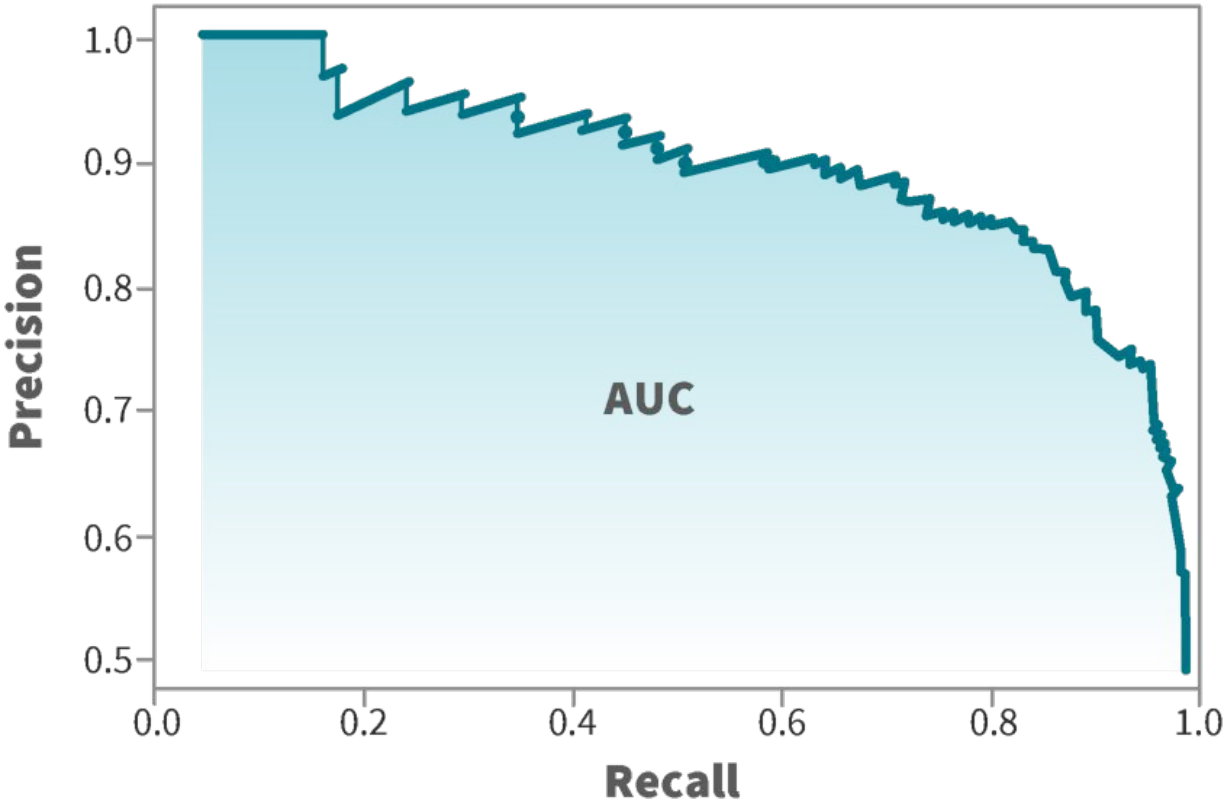
CONSIDER: *IMBALANCED DATASETS*

Alternative to ROC: the Precision-Recall Curve



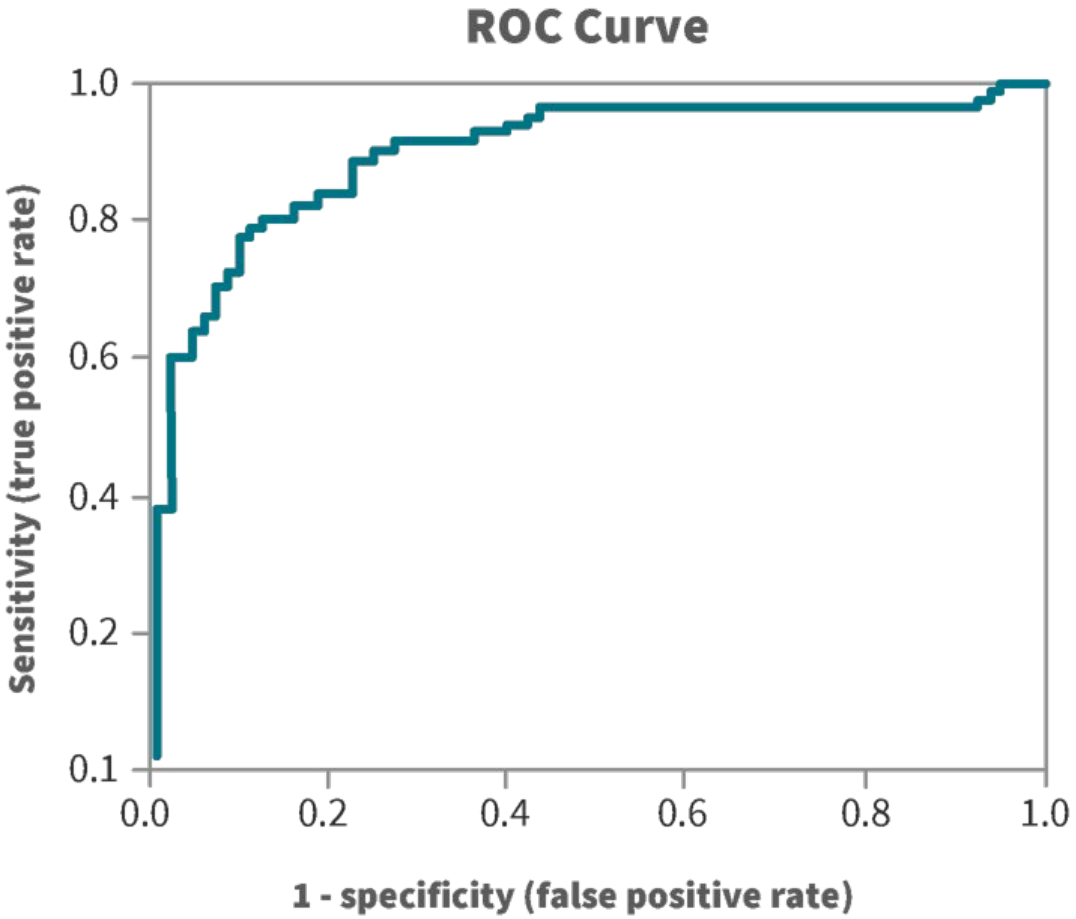
CONSIDER: *IMBALANCED DATASETS*

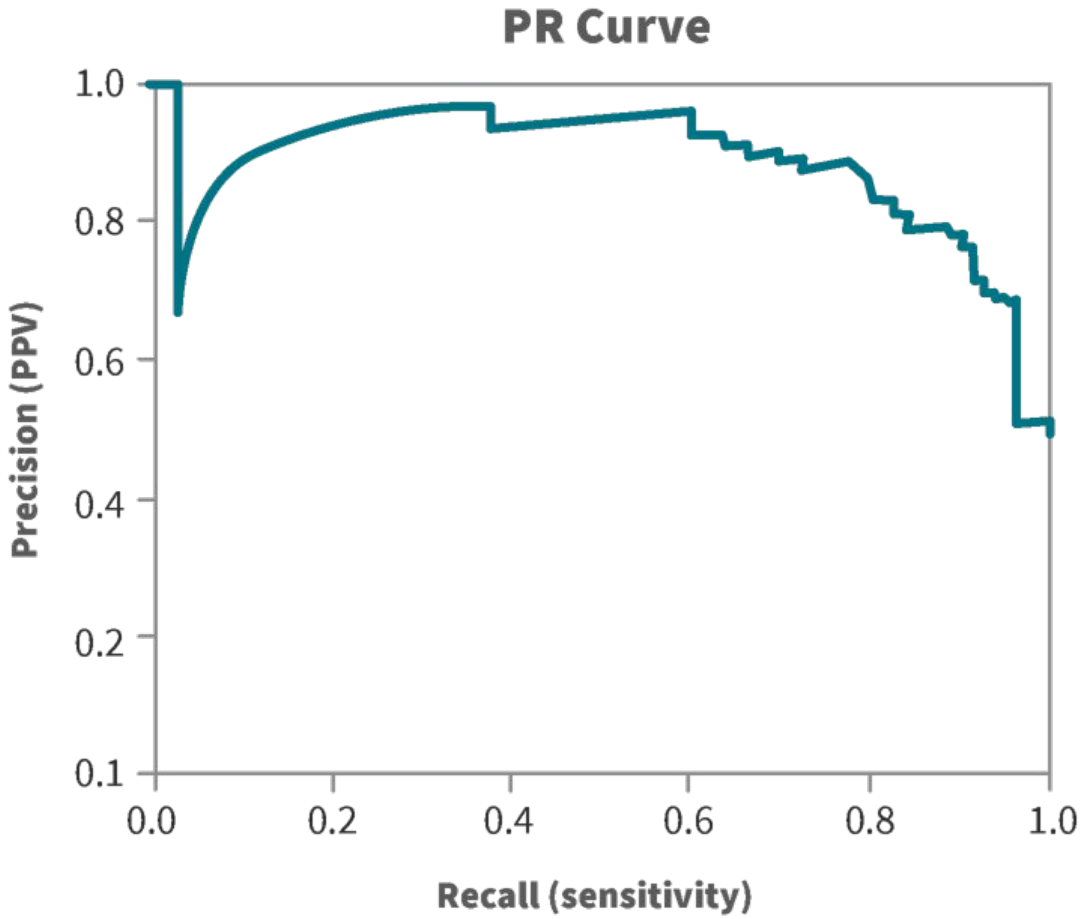
Alternative to ROC: the Precision-Recall Curve



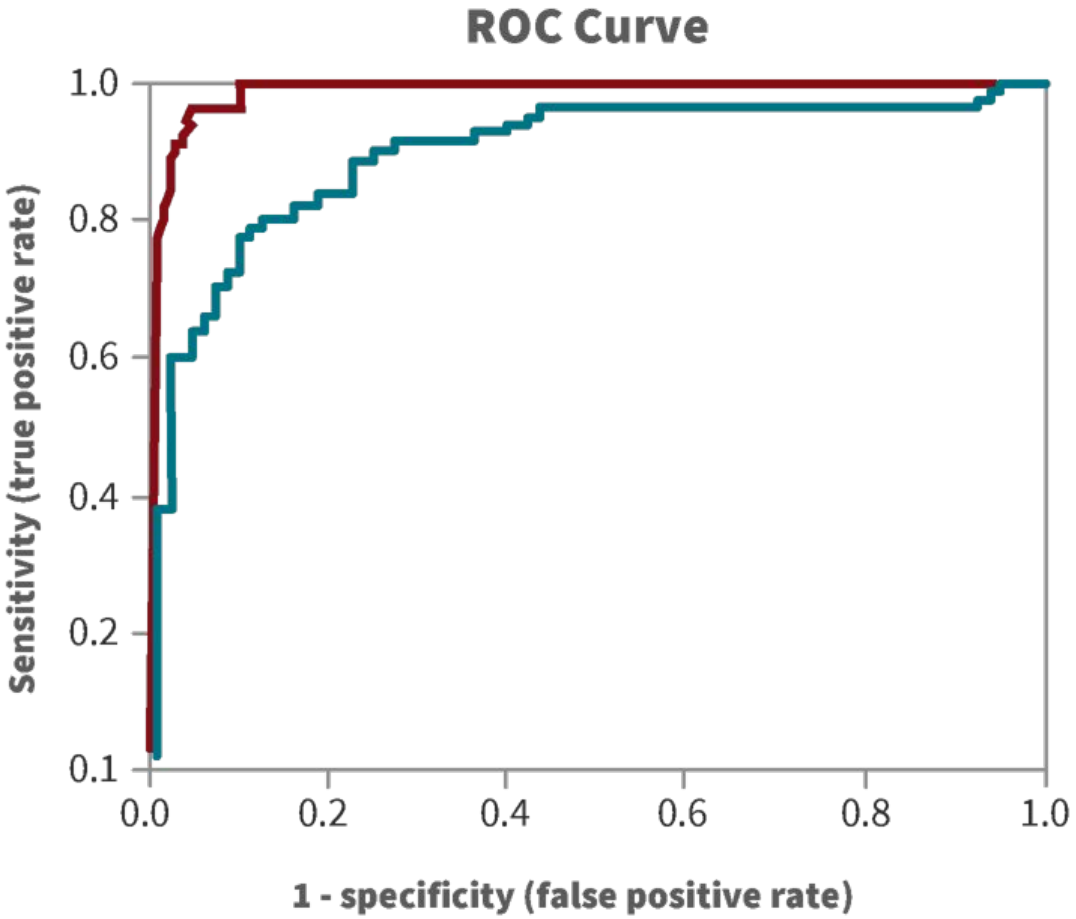
Curve	X-axis		y-axis	
	Concept	Calculation	Concept	Calculation
Precision-recall	Recall	TP / (TP +FN)	Precision	TP / (TP +FP)
ROC	1-specificity	FP / (FP +TN)	Sensitivity	TP / (TP +FN)

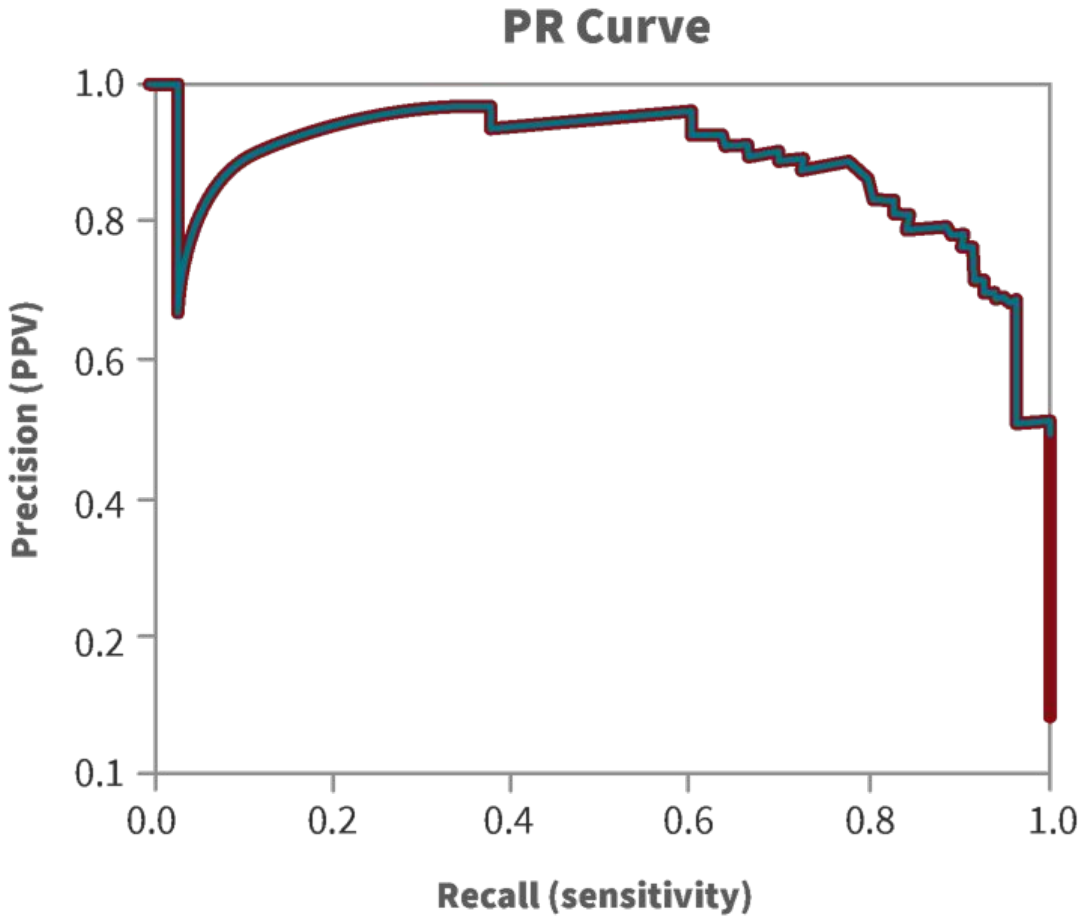
ROC VS. PR CURVES WITH CLASS IMBALANCE





ROC VS. PR CURVES WITH CLASS IMBALANCE





Outline

1. Review
2. Regression metrics
3. Classification metrics
4. **Applications**
5. Visual recognition metrics

Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network

Awni Y. Hannun[1,6*], Pranav Rajpurkar[1,6], Masoumeh Haghpanahi[2,6], Geoffrey H. Tison[3,6], Codie Bourn[2], Mintu P. Turakhia[4,5] and Andrew Y. Ng[1]

Abstract

Computerized electrocardiogram (ECG) interpretation plays a critical role in the clinical ECG workflow[1]. Widely available digital ECG data and the algorithmic paradigm of deep learning[2] present an opportunity to substantially improve the accuracy and scalability of automated ECG analysis. However, a comprehensive evaluation of an end-to-end deep learning approach for ECG analysis across a wide variety of diagnostic classes has not been previously reported. Here, we develop a deep neural network (DNN) to classify 12 rhythm classes using 91,232 single-lead ECGs from 53,549 patients who used a single-lead ambulatory ECG monitoring device. When validated against an independent test dataset

Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network

Awni Y. Hannun[1,6]*, Pranav Rajpurkar[1,6], Masoumeh Haghpanahi[2,6], Geoffrey H. Tison[3,6], Codie Bourn[2], Mintu P. Turakhia[4,5] and Andrew Y. Ng[1]

Abstract

Computerized electrocardiogram (ECG) interpretation plays a critical role in the clinical ECG workflow[1]. Widely available digital ECG data and the algorithmic paradigm of deep learning[2] present an opportunity to substantially improve the accuracy and scalability of automated ECG analysis. However, a comprehensive evaluation of an end-to-end deep learning approach for ECG analysis across a wide variety of diagnostic classes has not been previously reported. Here, we develop a deep neural network (DNN) to classify 12 rhythm classes using 91,232 single-lead ECGs from 53,549 patients who used a single-lead ambulatory ECG monitoring device. When validated against an independent test dataset annotated by a consensus committee of board-certified practicing cardiologists, the DNN achieved an average area under the receiver operating characteristic curve (ROC) of 0.97.

Skeletal bone age prediction based on a deep residual network with spatial transformer

Yaxin Han[a], Guangbin Wang[b,*]

[a] *Department of Orthopedics, First Affiliated Hospital of China Medical University, Shenyang, China*
[b] *Department of Orthopedics, Shengjing Hospital of China Medical University, Shenyang, China*

A B S T R A C T

Objective: Bone age prediction can be performed by medical experts manually assessment of X-ray images of the hand bone. In practice, the workload is huge, resource consumption is large, measurement takes a long time, and it is easily influenced by human factors. As such, manual estimation of bone age takes a long time and the results fluctuate greatly depending on the proficiency of the radiologist.

Methods: The left-hand X-ray image data was identified and pre-processed. X-ray image analysis method using on deep neural network was used to automatically extract the key features of the left-hand joint bone age, and evaluation performance of the model was implemented.

Results: In this paper, the deep learning method can be used to obtain the X-ray bone image features, and the convolutional neural network is used to automatically assess the age of bone. The feature region extraction method based on deep learning can extract feature information with superior performance

Skeletal bone age prediction based on a deep residual network with spatial transformer

Yaxin Han [a], Guangbin Wang [b,*]

[a] Department of Orthopedics, First Affiliated Hospital of China Medical University, Shenyang, China
[b] Department of Orthopedics, Shengjing Hospital of China Medical University, Shenyang, China

A B S T R A C T

Objective: Bone age prediction can be performed by medical experts manually assessment of X-ray images of the hand bone. In practice, the workload is huge, resource consumption is large, measurement takes a long time, and it is easily influenced by human factors. As such, manual estimation of bone age takes a long time and the results fluctuate greatly depending on the proficiency of the radiologist.

Methods: The left-hand X-ray image data was identified and pre-processed. X-ray image analysis method using on deep neural network was used to automatically extract the key features of the left-hand joint bone age, and evaluation performance of the model was implemented.

Results: In this paper, the deep learning method can be used to obtain the X-ray bone image features, and the convolutional neural network is used to automatically assess the age of bone. The feature region extraction method based on deep learning can extract feat

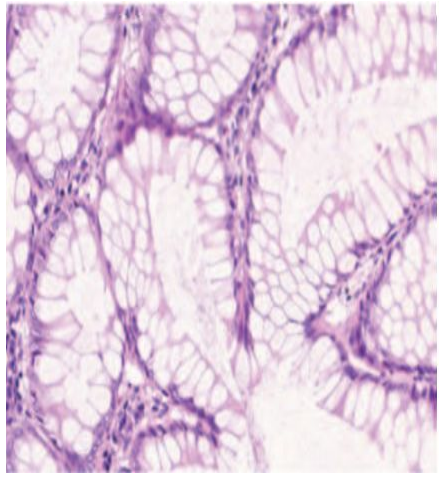
The MAE of the bone age of males and females of different ages was compared in this study. The results are shown in Table 2. The male's MAE is 0.445 and the female's is 0.474. According to the age group, the minimum MAE of males and females appeared in 0 to 2 years old, and the maximum MAE appeared in 17–18 years old. In the range of 2 to 16 years old, the MAE of males and females are distributed between 0.308 and 0.668.

Outline

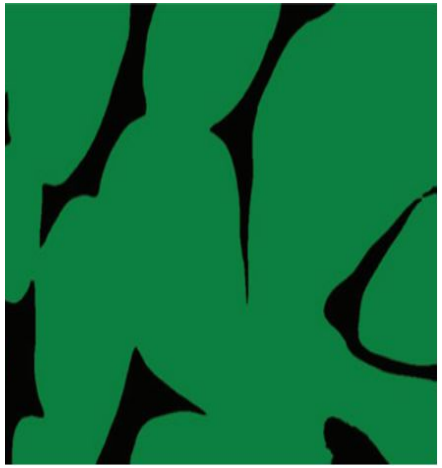
1. Review
2. Regression metrics
3. Classification metrics
4. Applications
5. **Visual recognition metrics**

Richer visual recognition tasks: segmentation and detection

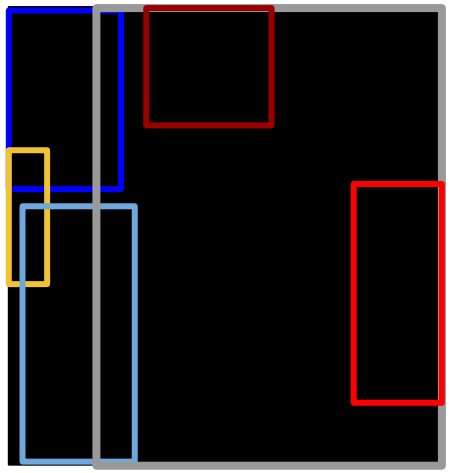
Classification



Semantic Segmentation

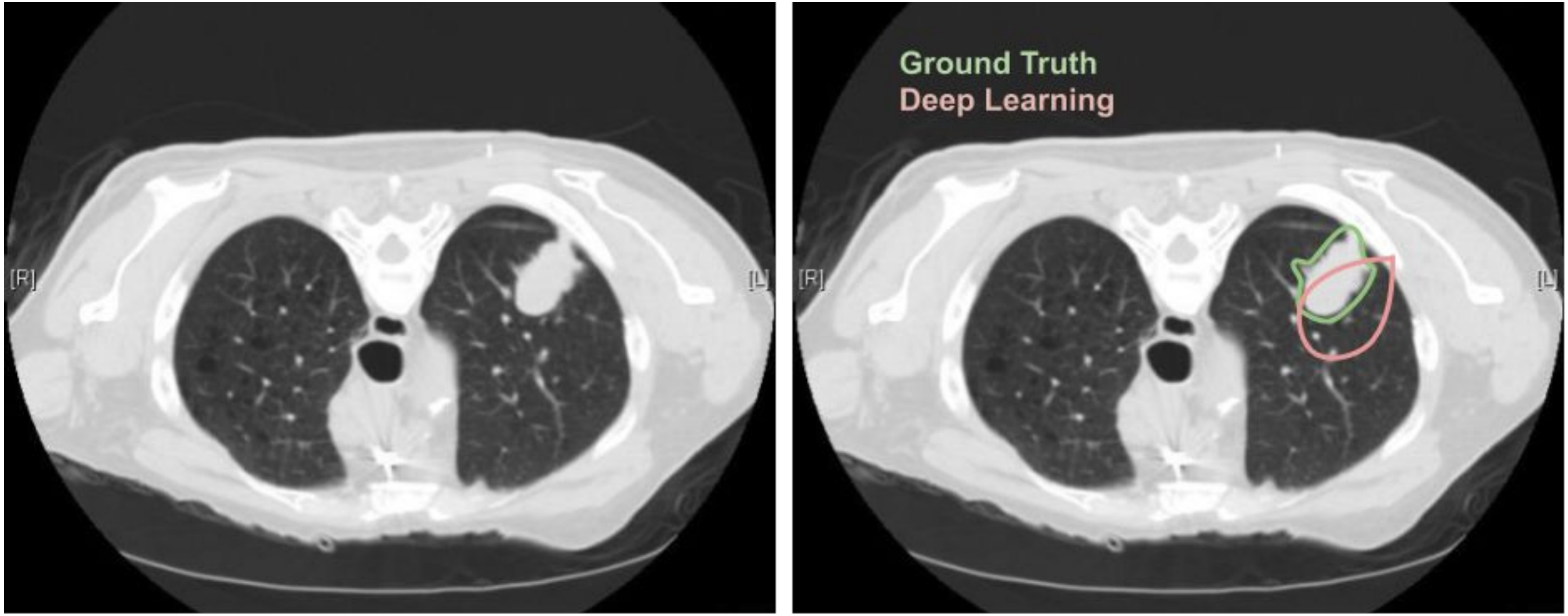


Detection



Instance Segmentation

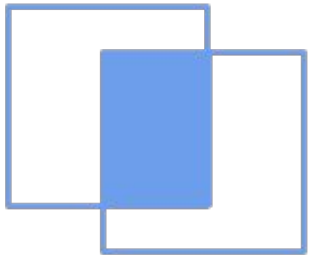
Segmentation Metrics

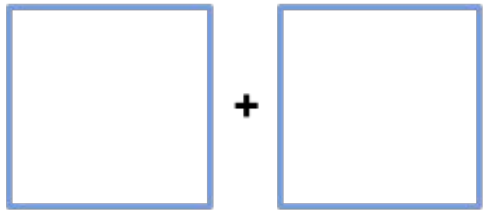


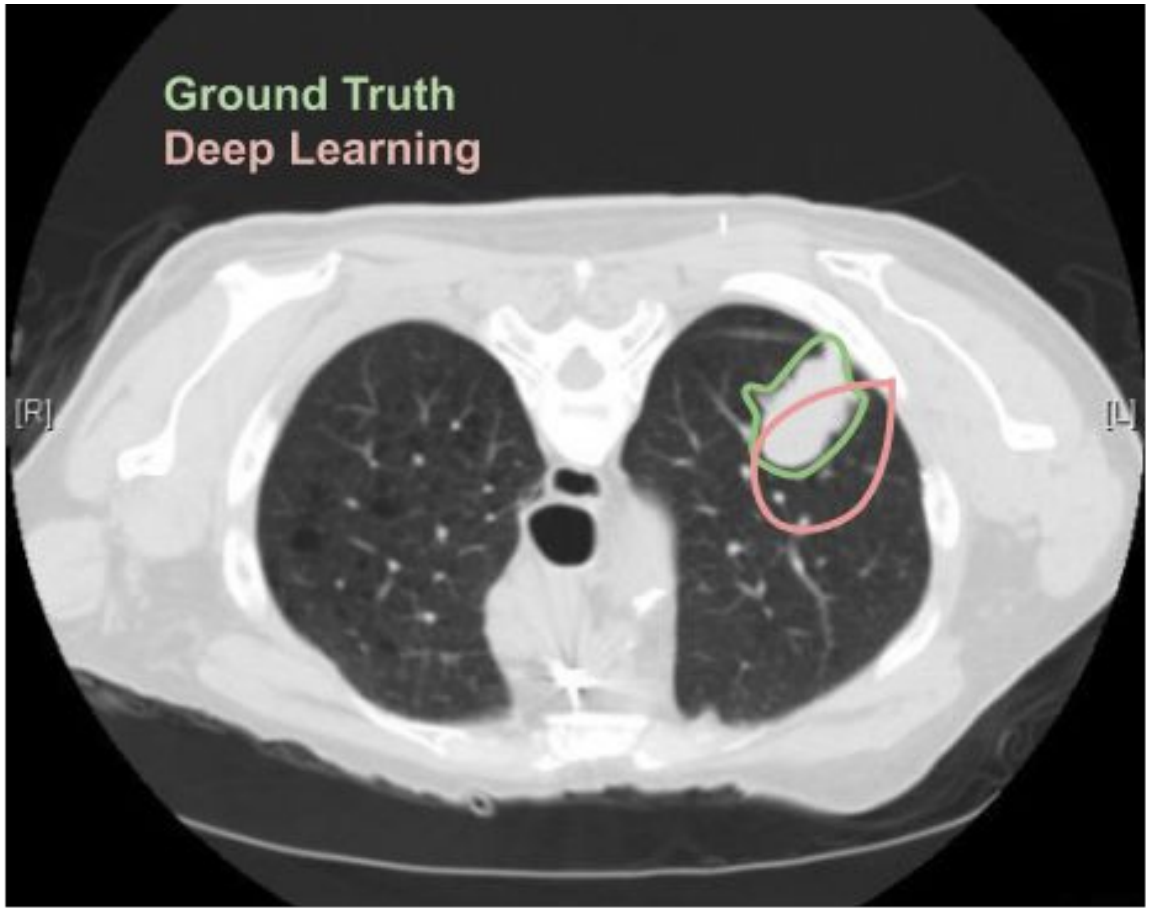
Segmentation Metrics

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

$$\text{Sørensen-Dice} = \frac{2 \times}{+}$$



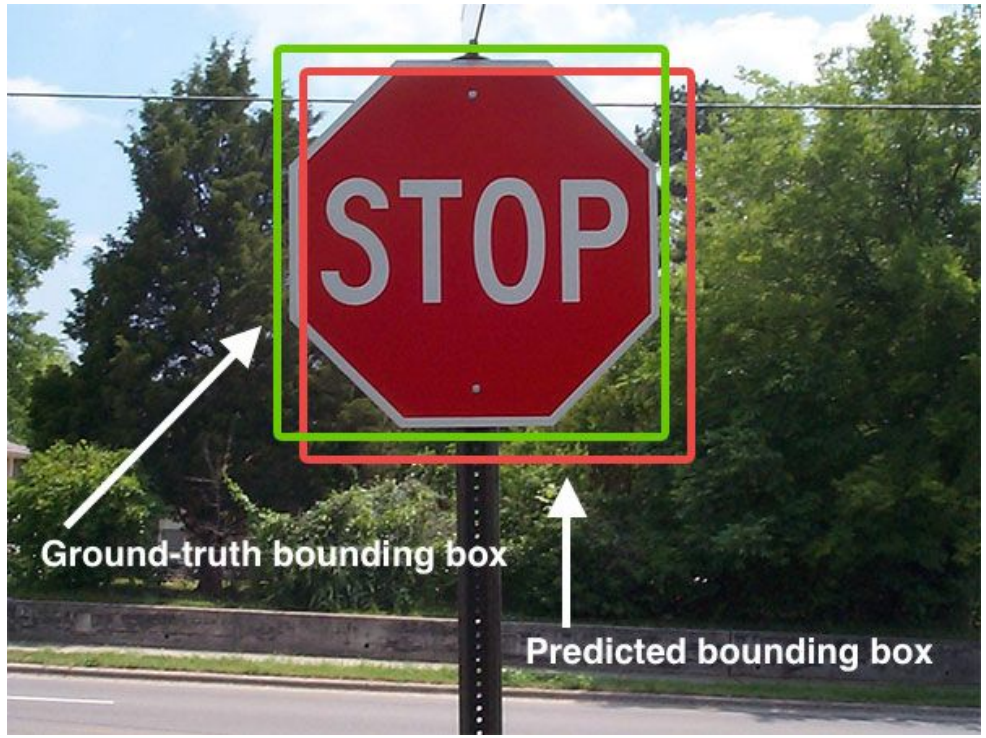


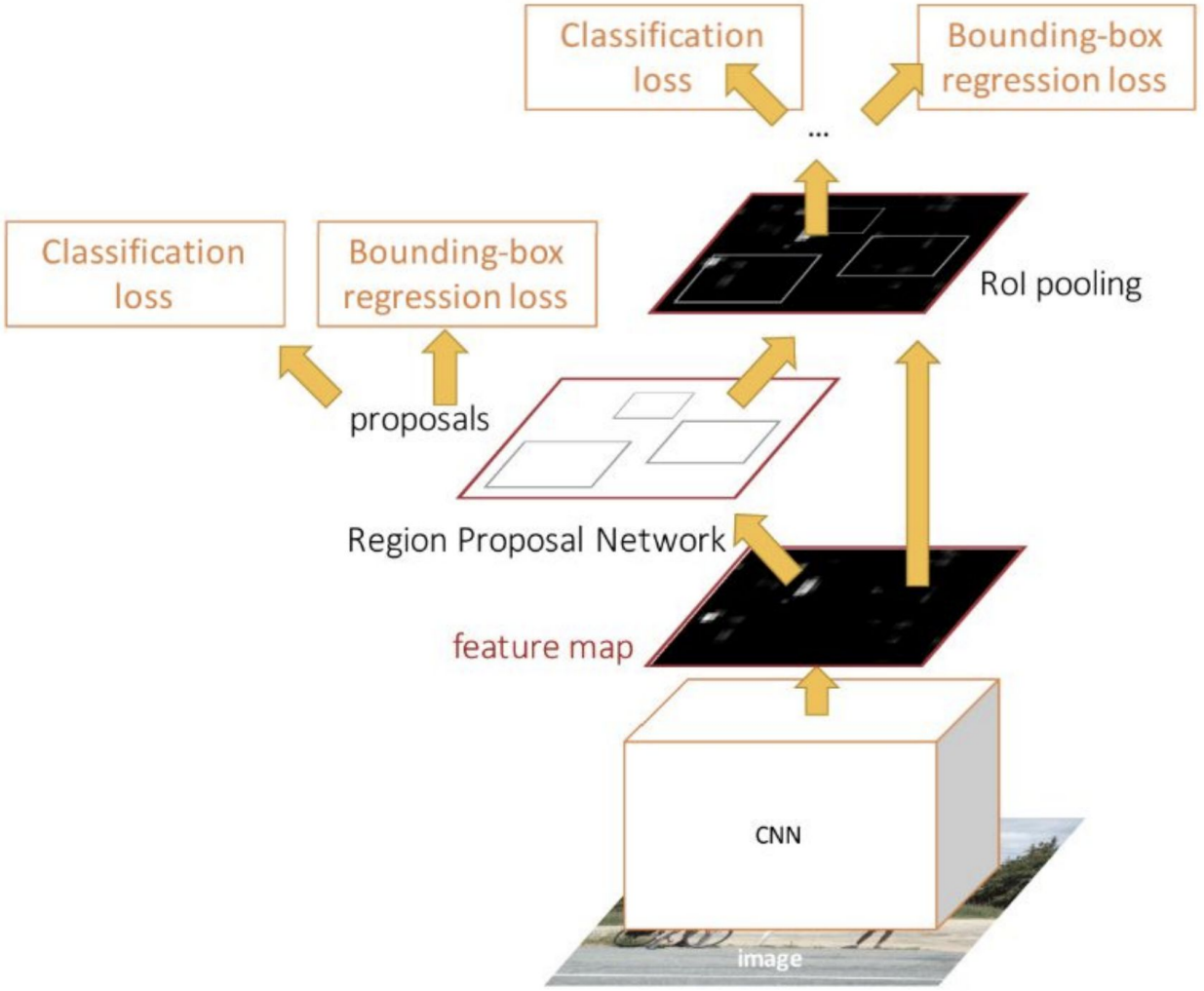


Object Detection Metrics

Object detection models output **bounding boxes** with associated **confidence levels:**

- (x, y, h, w, c)

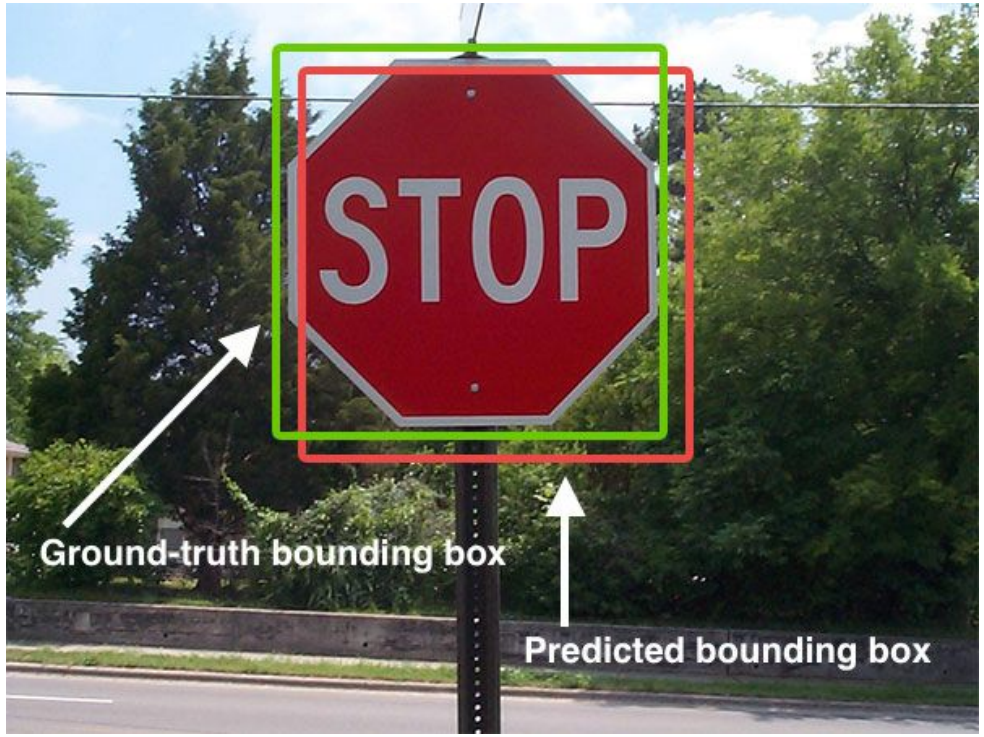




Object Detection Metrics

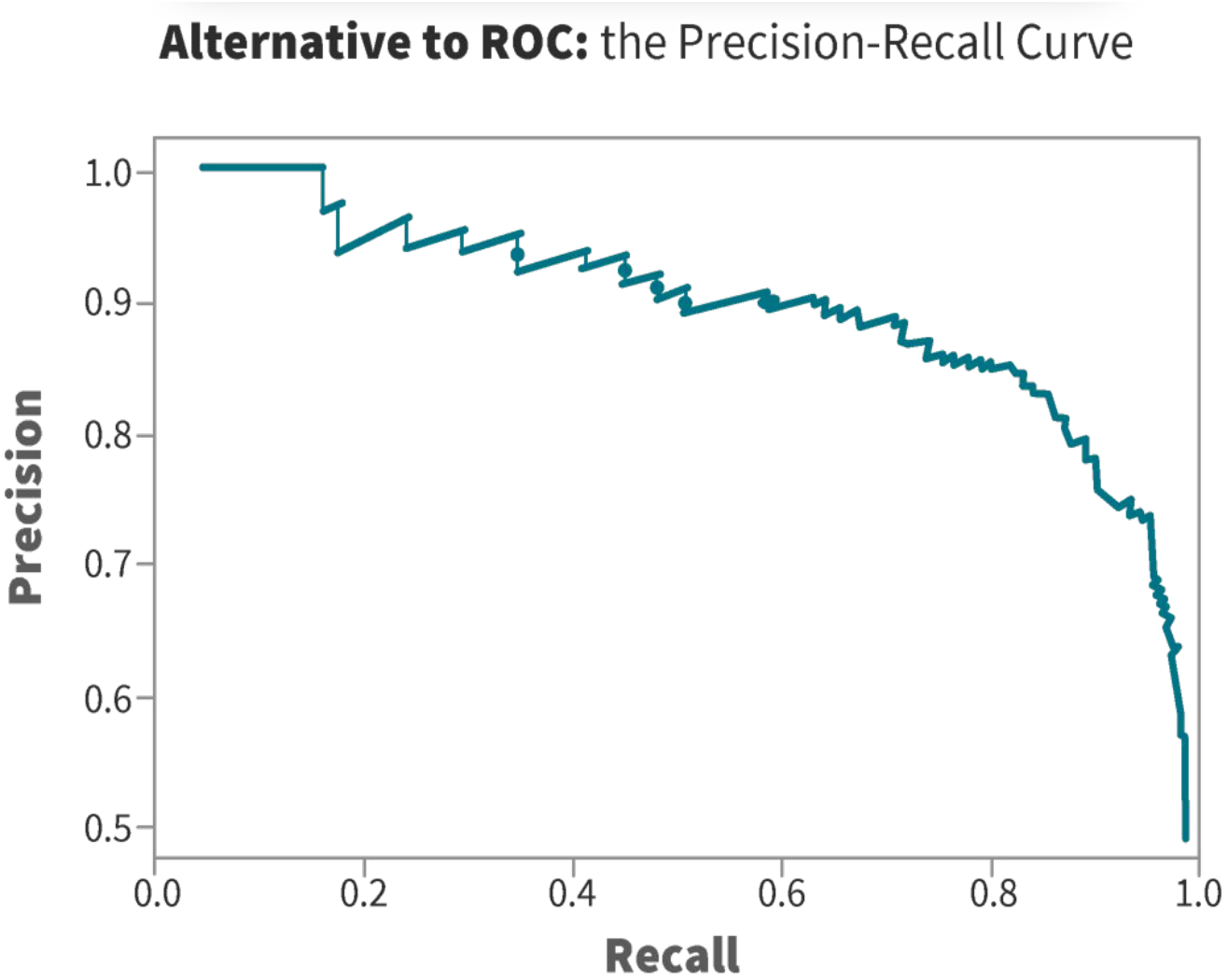
Object detection models output **bounding boxes** with associated **confidence levels:**

- (x, y, h, w, c)



- Object detection is typically heavily imbalanced (most of the data is background)
- PR curves most common evaluation

Object Detection Metrics



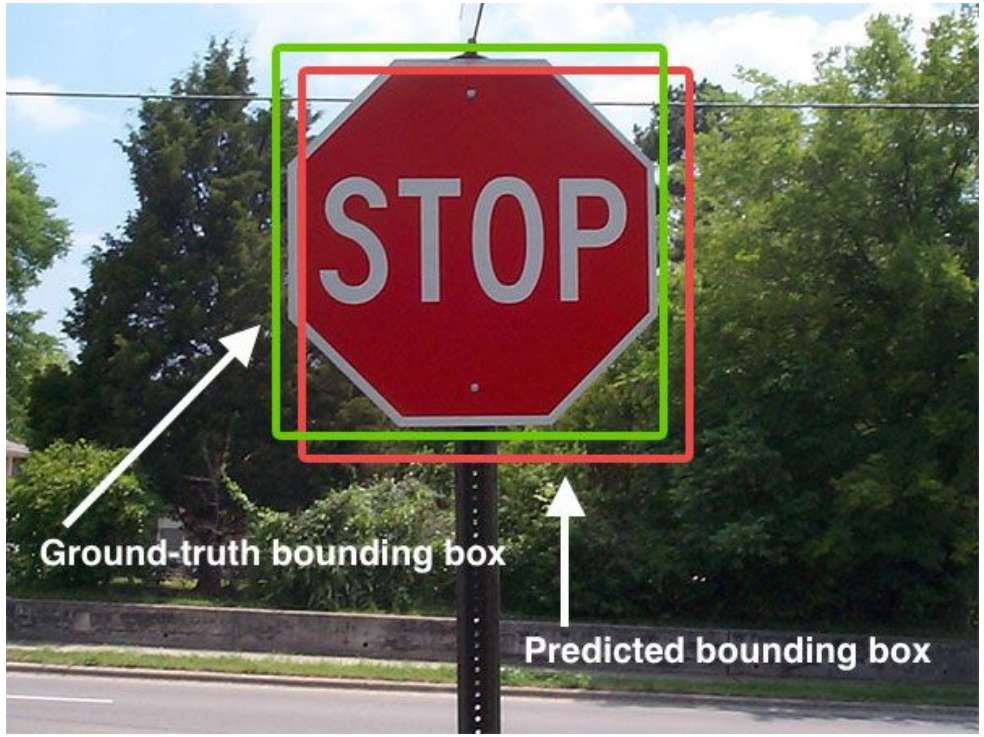
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Object Detection Metrics

Object detection models output **bounding boxes** with associated **confidence levels:**

- (x, y, h, w, c)

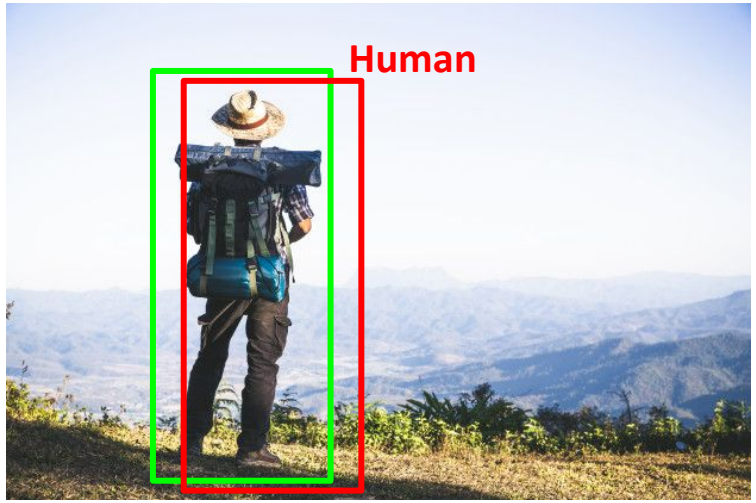


We have the class confidences to vary the threshold in plotting the PR curve. But how do we get TP, FP, FN?

A: Choose an IOU threshold with ground truth boxes to determine if bounding box prediction is TP, FP, or FN. Then can plot PR curve and obtain AP metric.

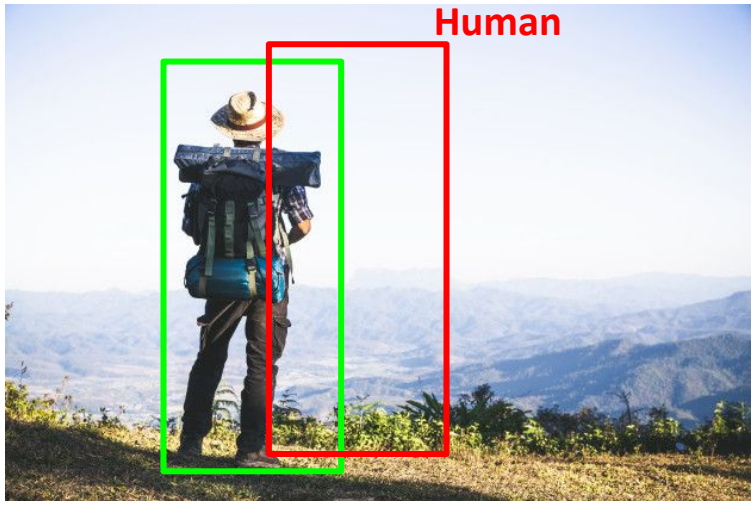
Object Detection Metrics

True Positive

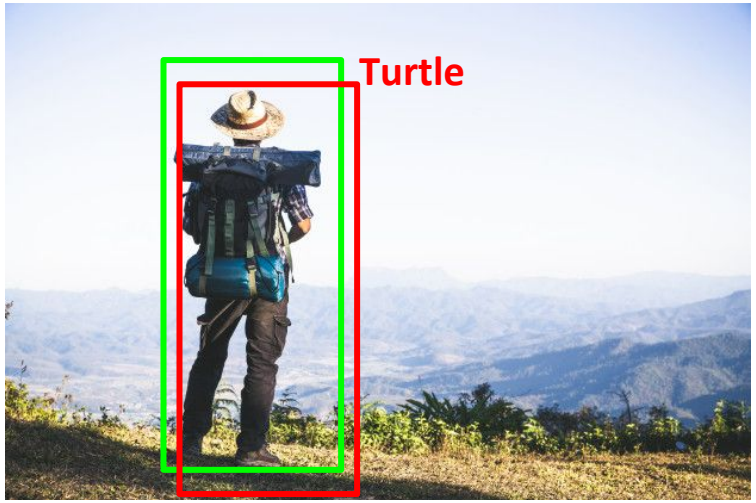


IoU > Threshold

False Positive



IoU < Threshold

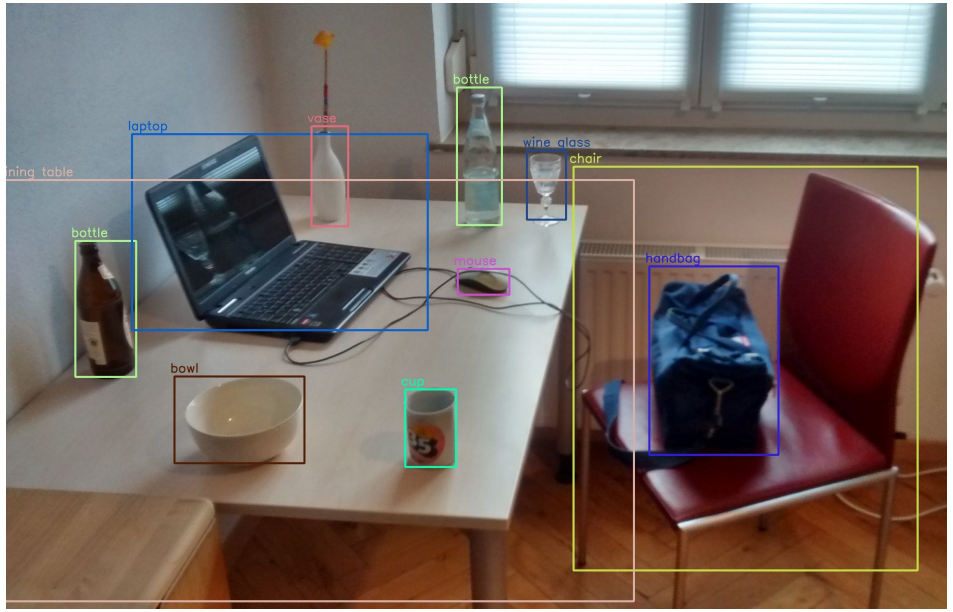


Wrong Class

False Negative

No prediction

Instance Segmentation Metrics

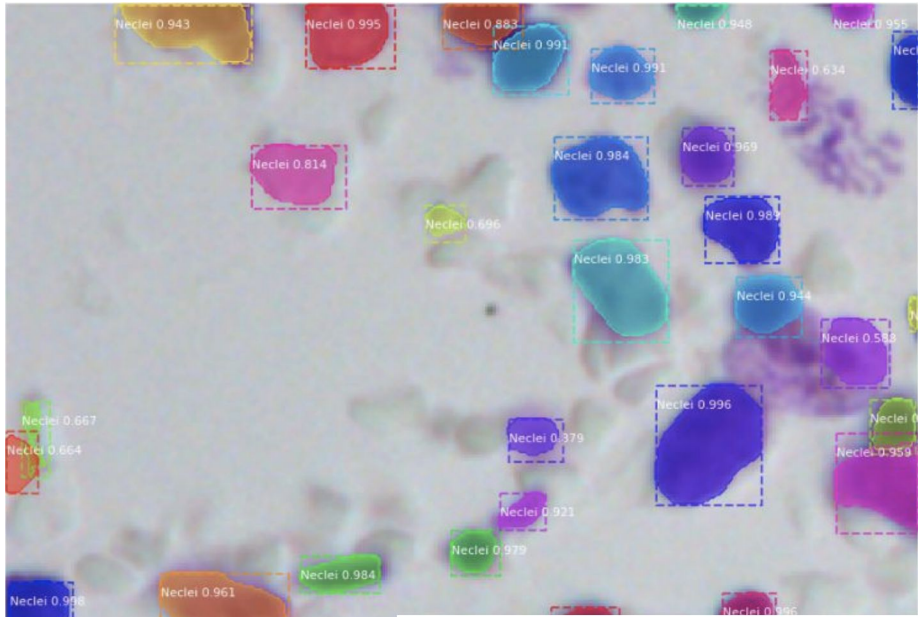


COCO test-dev

mAP@.5	mAP@[.5, .95]
35.9	19.7
39.3	19.3
42.1	21.5
42.7	**21.9**

- **AP (Average Precision):** another way to call AUPRC
- **mAP**: Mean average Precision over all the classes
- **mAP@.5:** Mean average Precision over all the classes using 0.5 IOU as threshold
- **mAP@[.5, .95]:** Average of mAP values at IOU thresholds regularly sampled in the interval between [.5, .95].

Instance Segmentation Metrics



- Instance-based task, like object detection.
- Also use same precision-recall curve / AP evaluation metrics
- Only difference is that IOU is now a mask IOU

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

Summary

Today we covered evaluation metrics for:

- Regression
- Classification
- Visual recognition

Coming up: Strategies, challenges, and the black box