Classification as a Machine Learning Problem

Overview

Classification is a canonical problem in Machine Learning

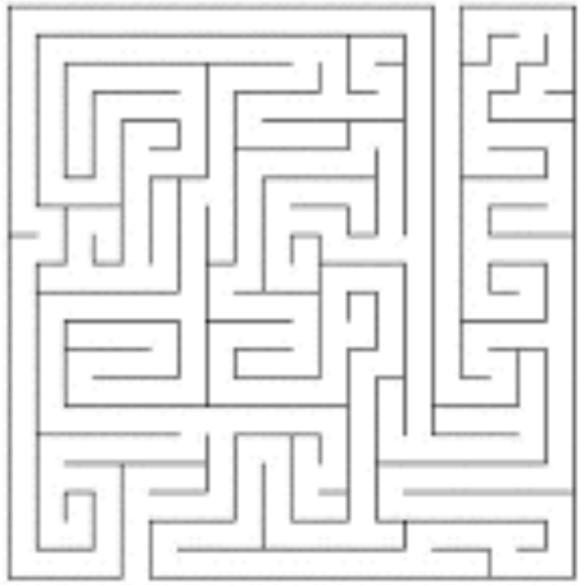
Classifiers can be measured using accuracy, precision and recall

Traditional ML models for classification include SVM and Naive Bayes

Neural networks perform very well on classification problems

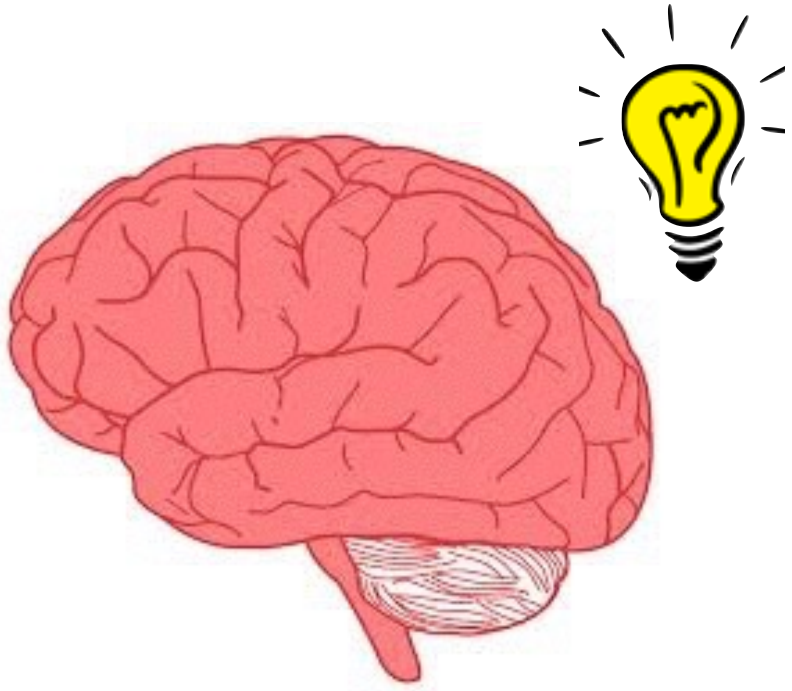
Classification and Classifiers

Machine Learning



Work with a huge maze of data

Find patterns



Make intelligent decisions

Machine Learning

Emails on a server

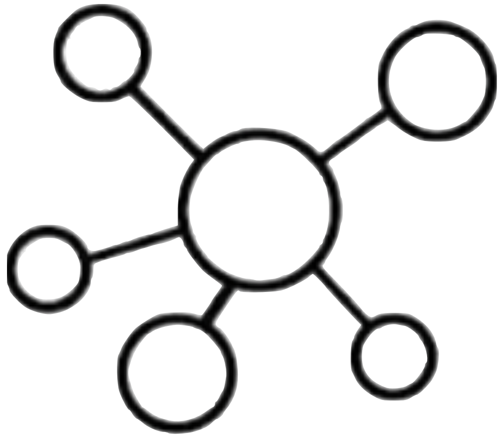
Spam or Ham?

Trash or Inbox

Types of Machine Learning Problems

Classification

Regression



Clustering

Rule-extraction

Types of Machine Learning Problems

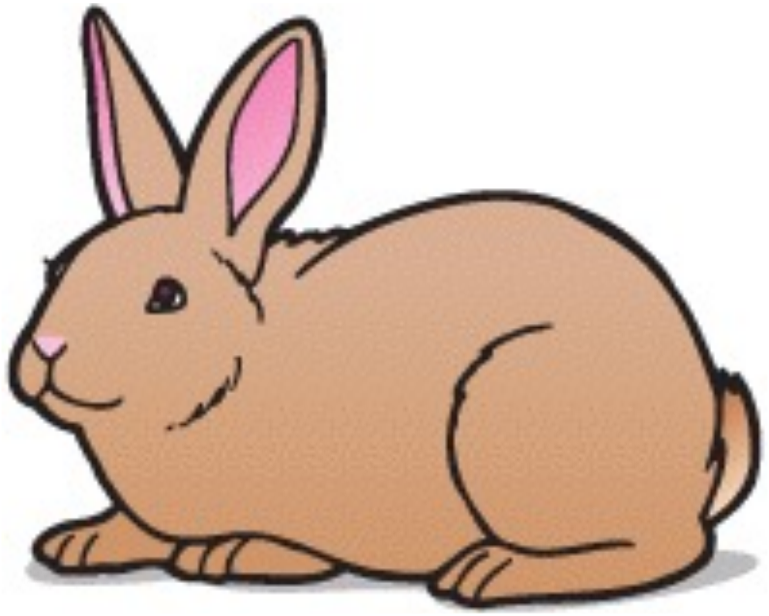
Classification

Regression

Clustering

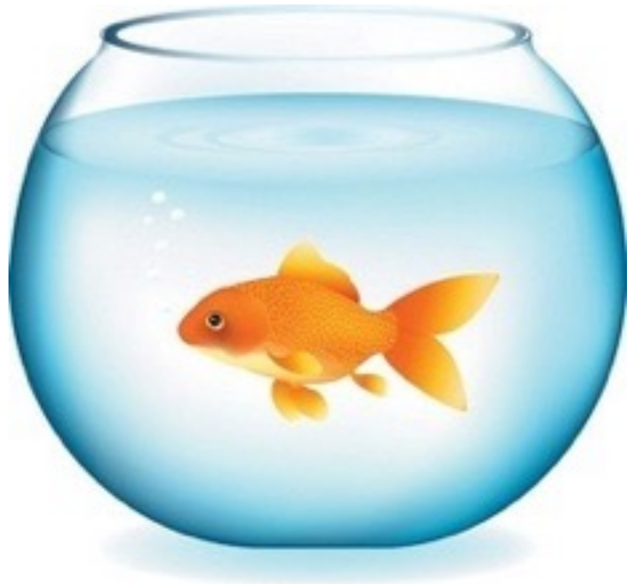
Rule-extraction

Whales: Fish or Mammals?



Mammals

Members of the infraorder Cetacea



Fish

Look like fish, swim like fish, move with fish

Whales: Fish or Mammals?

ML-based Classifier

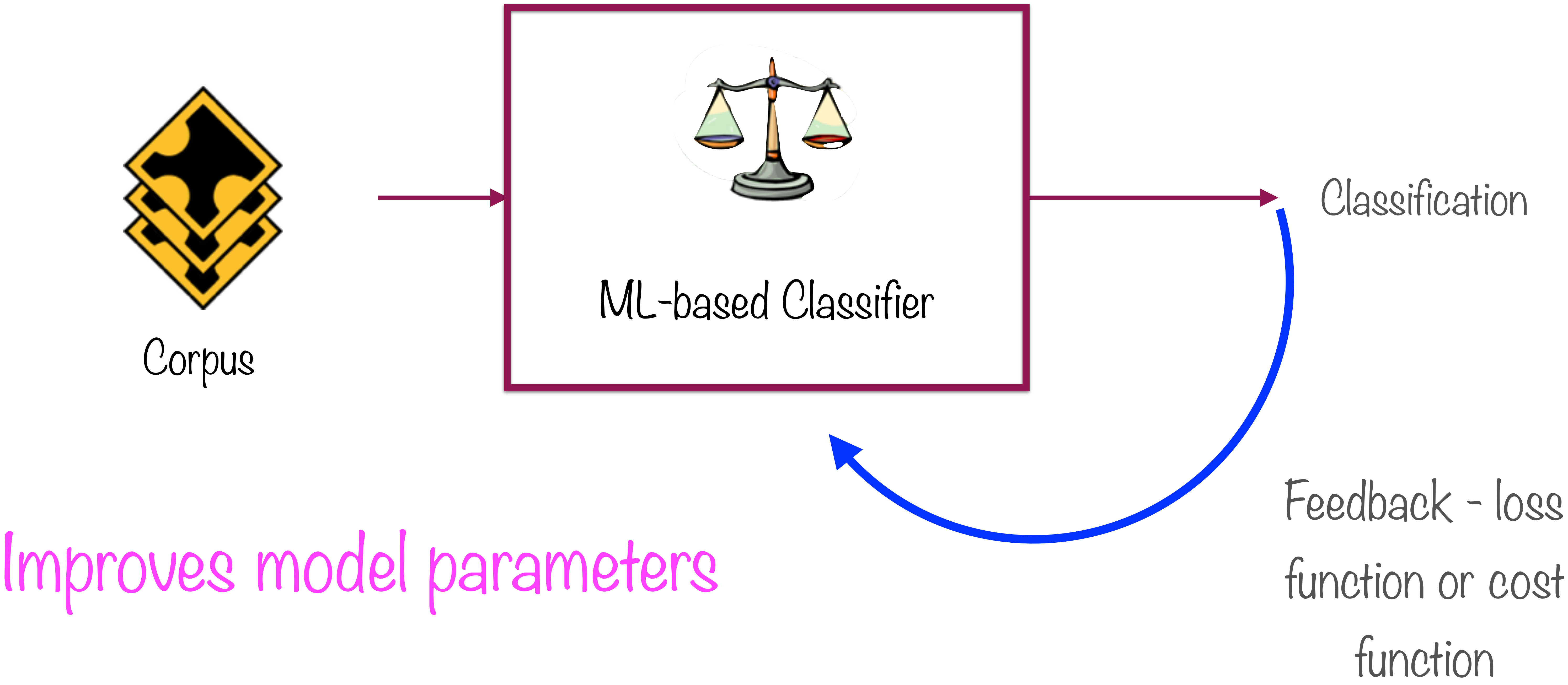
Training

Feed in a large corpus of data classified correctly

Prediction

Use it to classify new instances which it has not seen before

Training the ML-based Classifier



An algorithm might have high accuracy but still be a poor machine learning model

Its predictions are useless

Accuracy, Precision, Recall

All-is-well Binary Classifier

Medical reports

Always classify as "normal"

No Cancer

Here, accuracy for rare cancer may be 99.9999%, but...

Accuracy

Some labels maybe much more common/rare than others

Such a dataset is said to be skewed

Accuracy is a poor evaluation metric here

Confusion Matrix

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 instances	4 instances
No Cancer	5 instances	1000 instances

Confusion Matrix

Predicted Labels

Actual Label	Cancer	No Cancer
Cancer	10	4
No Cancer	5	1000

True Positive

Actual Label	Predicted Labels: Cancer	Predicted Labels: No Cancer
Cancer	10	4
No Cancer	5	1000

Actual Label = Predicted Label

True Positive

		Predicted Labels	
		Cancer	No Cancer
Actual Label	Cancer	10 TP	4
	No Cancer	5	1000

Actual Label = Predicted Label

False Positive

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10	4
No Cancer	5	1000

Actual Label $\neq$ Predicted Label

False Positive

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10	4
No Cancer	5 FP	1000

Actual Label ≠ Predicted Label

True Positive

Predicted Labels

Actual Label	Cancer	No Cancer
Cancer	10	4
No Cancer	5	1000

Actual Label = Predicted Label

True Negative

Predicted Labels

Actual Label	Cancer	No Cancer
Cancer	10	4
No Cancer	5	1000 TN

Actual Label = Predicted Label

False Negative

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10	4
No Cancer	5	1000

Actual Label $\neq$ Predicted Label

False Negative

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10	4 FN
No Cancer	5	1000

Actual Label ≠ Predicted Label

Confusion Matrix

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Accuracy

Actual Label \ Predicted Labels	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Accuracy

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Actual Label = Predicted Label

Accuracy

Predicted Labels

Actual Label	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

$$\text{Accuracy} = \frac{TP + TN}{\text{Num Instances}} = \frac{1010}{1019} = 99.12\%$$

Accuracy

Accuracy = 99.12%

Classifier gets it right 99.12% of the time

But…

Accuracy

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

People on chemotherapy, radiation when not required

Accuracy

Predicted Labels

Actual Label	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Cancer not detected, no treatment prescribed

Accuracy is not a good metric to evaluate whether this model performs well

Precision

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Precision

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Precision = Accuracy when classifier flags cancer

Precision

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{10}{15} = 66.67\%$$

Precision

Precision = 66.67%

1 in 3 cancer diagnoses is incorrect

Recall

Predicted Labels

Actual Label	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Recall

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

Recall = Accuracy when cancer actually present

Recall

Predicted Labels

Actual Label

	Cancer	No Cancer
Cancer	10 TP	4 FN
No Cancer	5 FP	1000 TN

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{10}{14} = 71.42\%$$

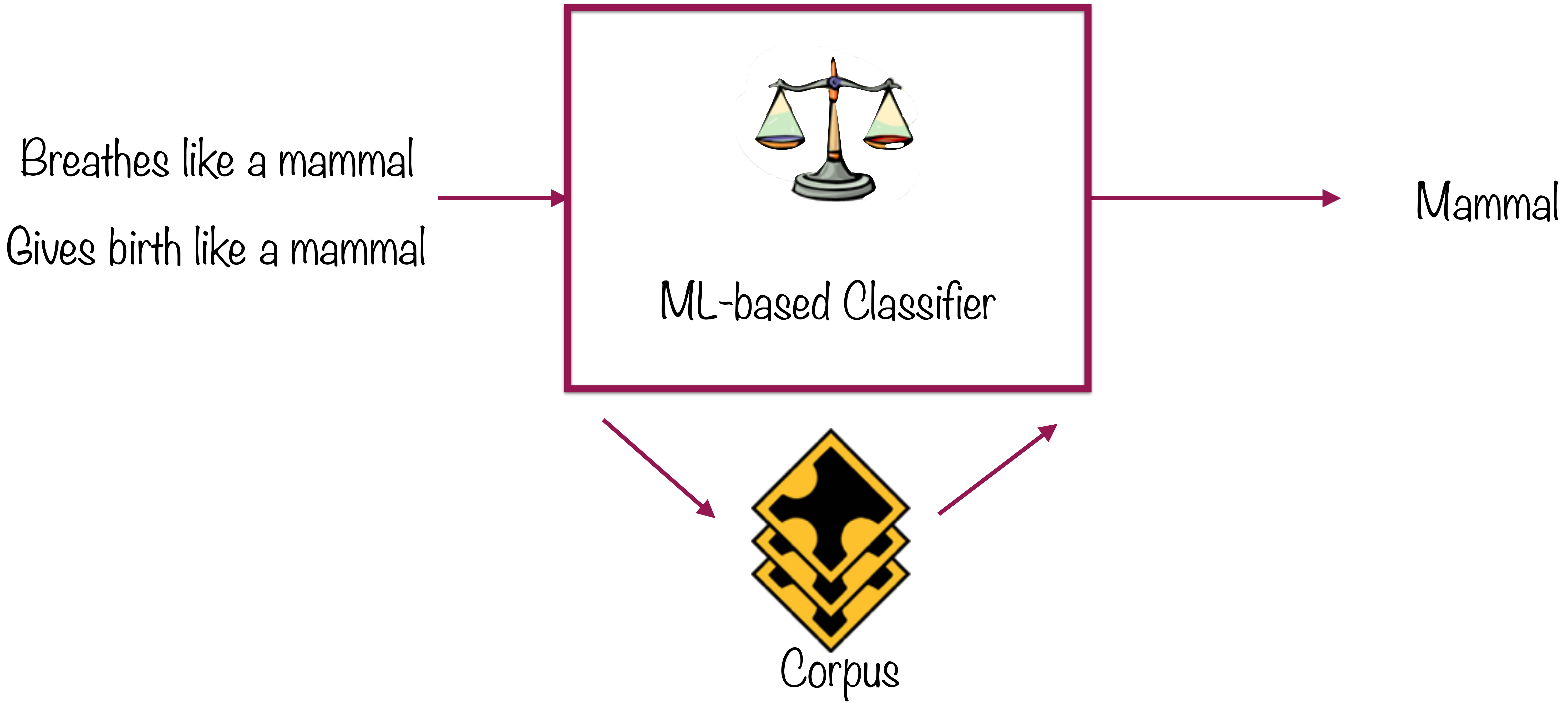
Recall

Recall = 71.42%

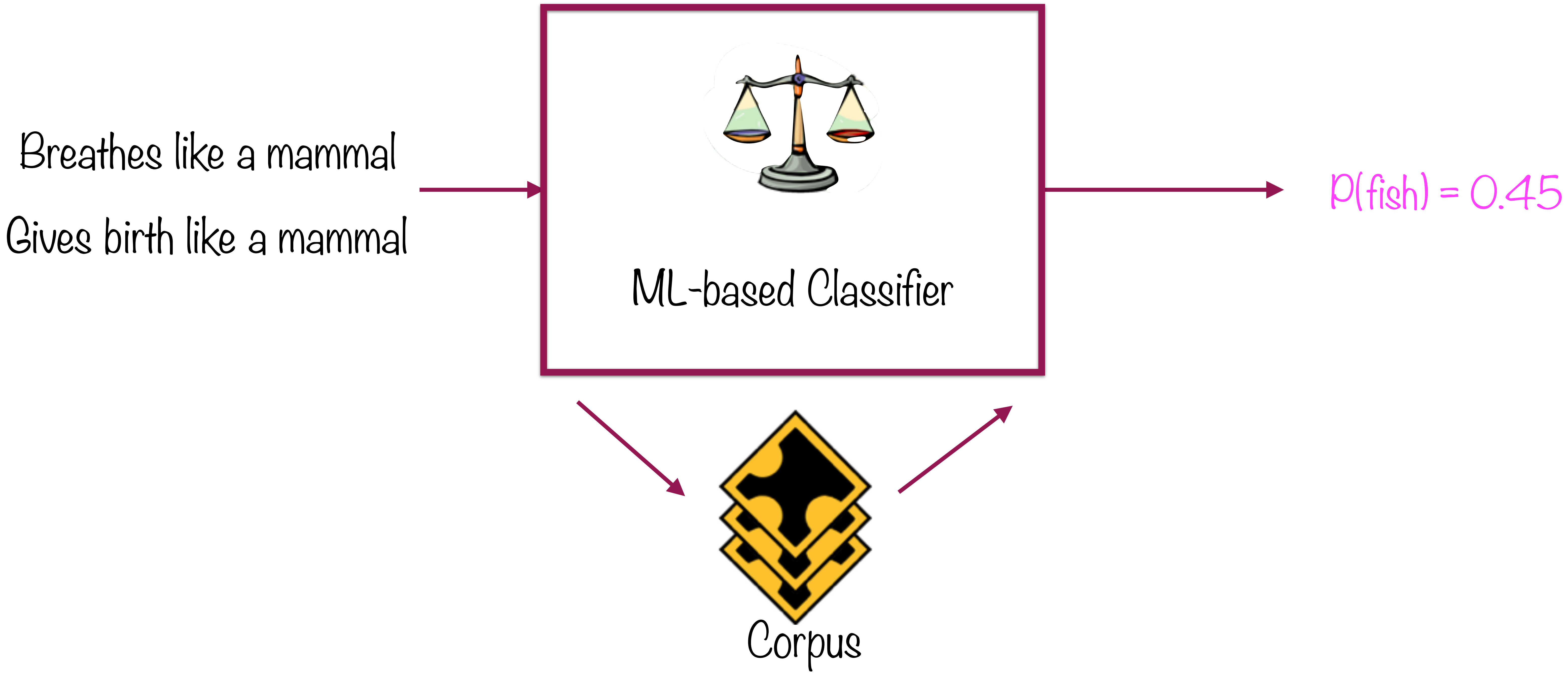
2 in 7 cancer cases missed

Choosing a Machine Learning Model

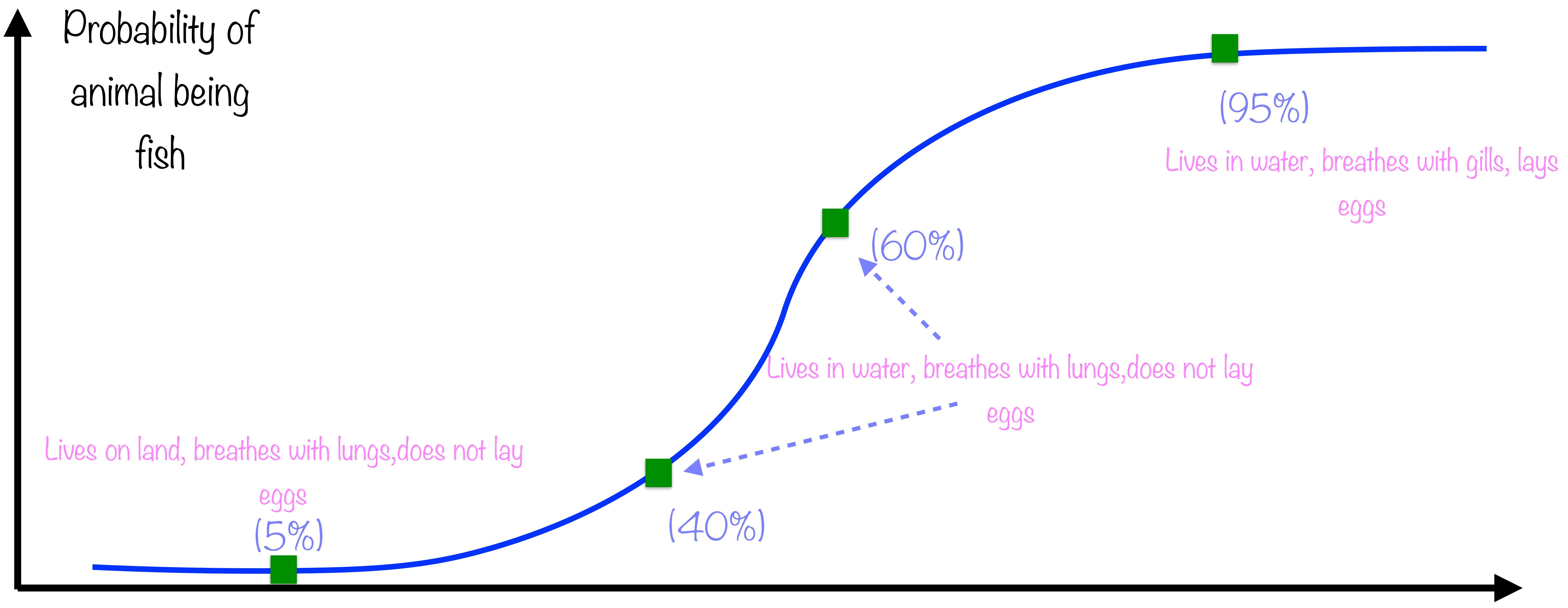
ML-based Binary Classifier



ML-based Binary Classifier

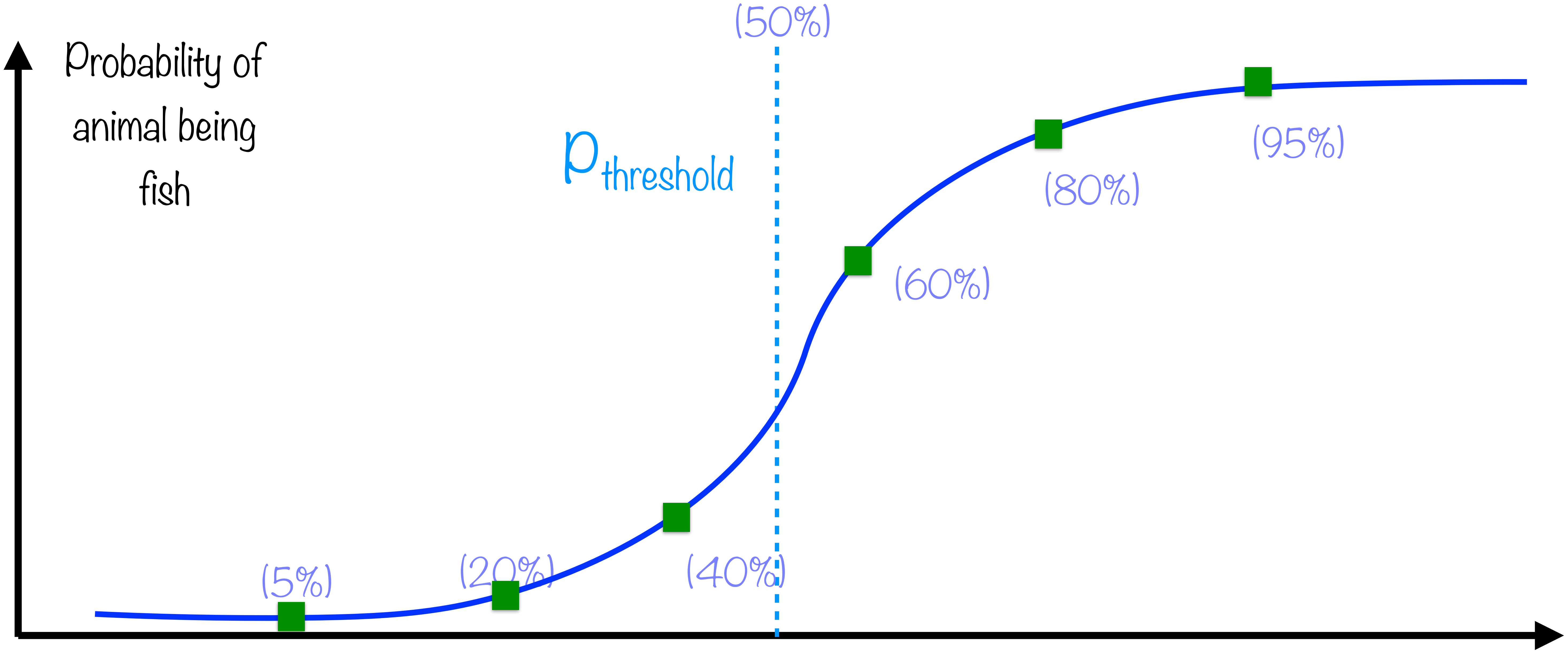


Applying Logistic Regression

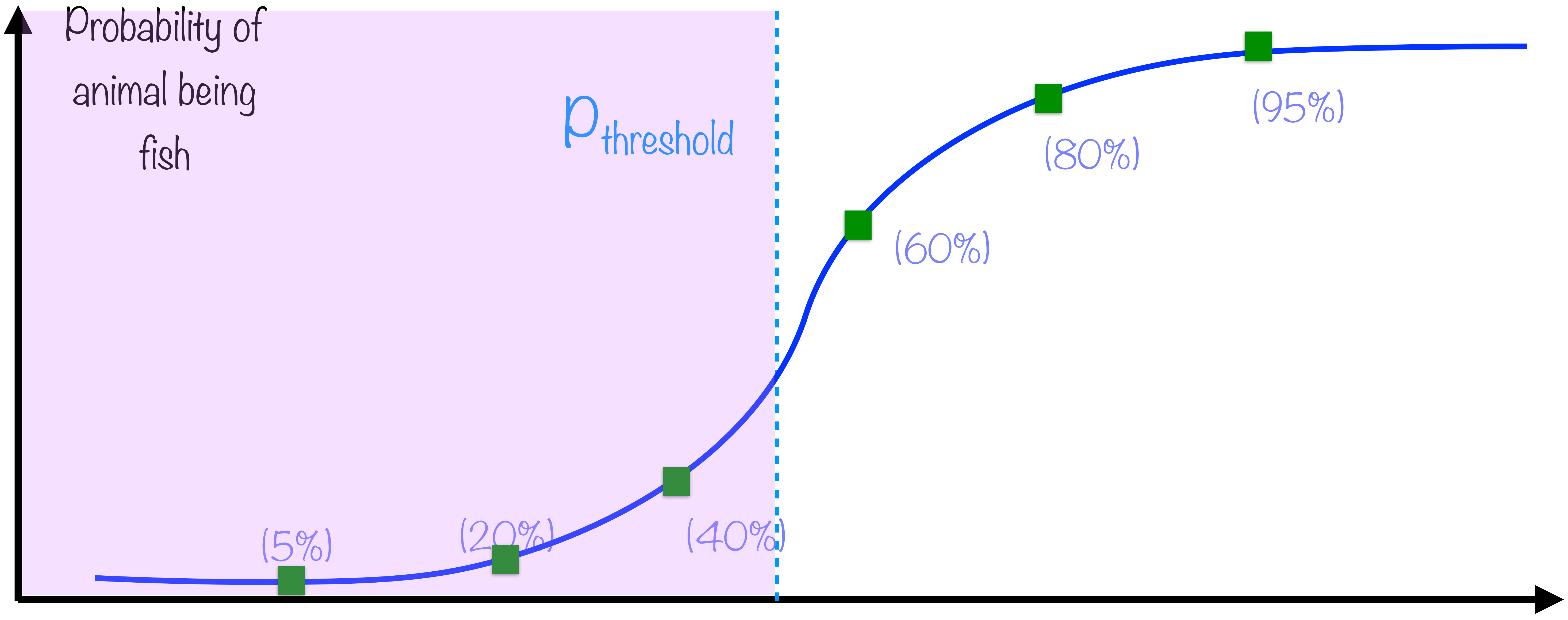


Whales: Fish or Mammals?

Choosing Decision Threshold

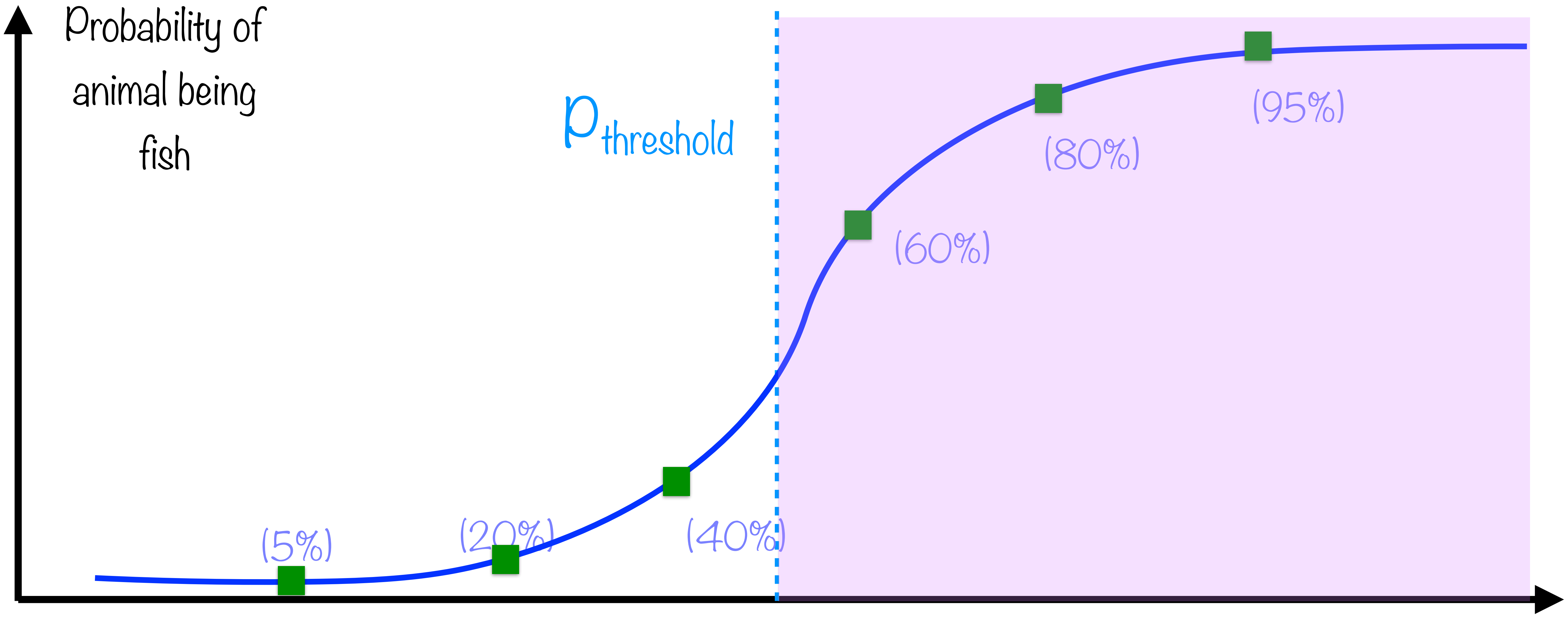


Choosing Decision Threshold



If probability < $p_{threshold}$, it's a mammal

Applying Logistic Regression



If probability > $p_{threshold}$, it's a fish

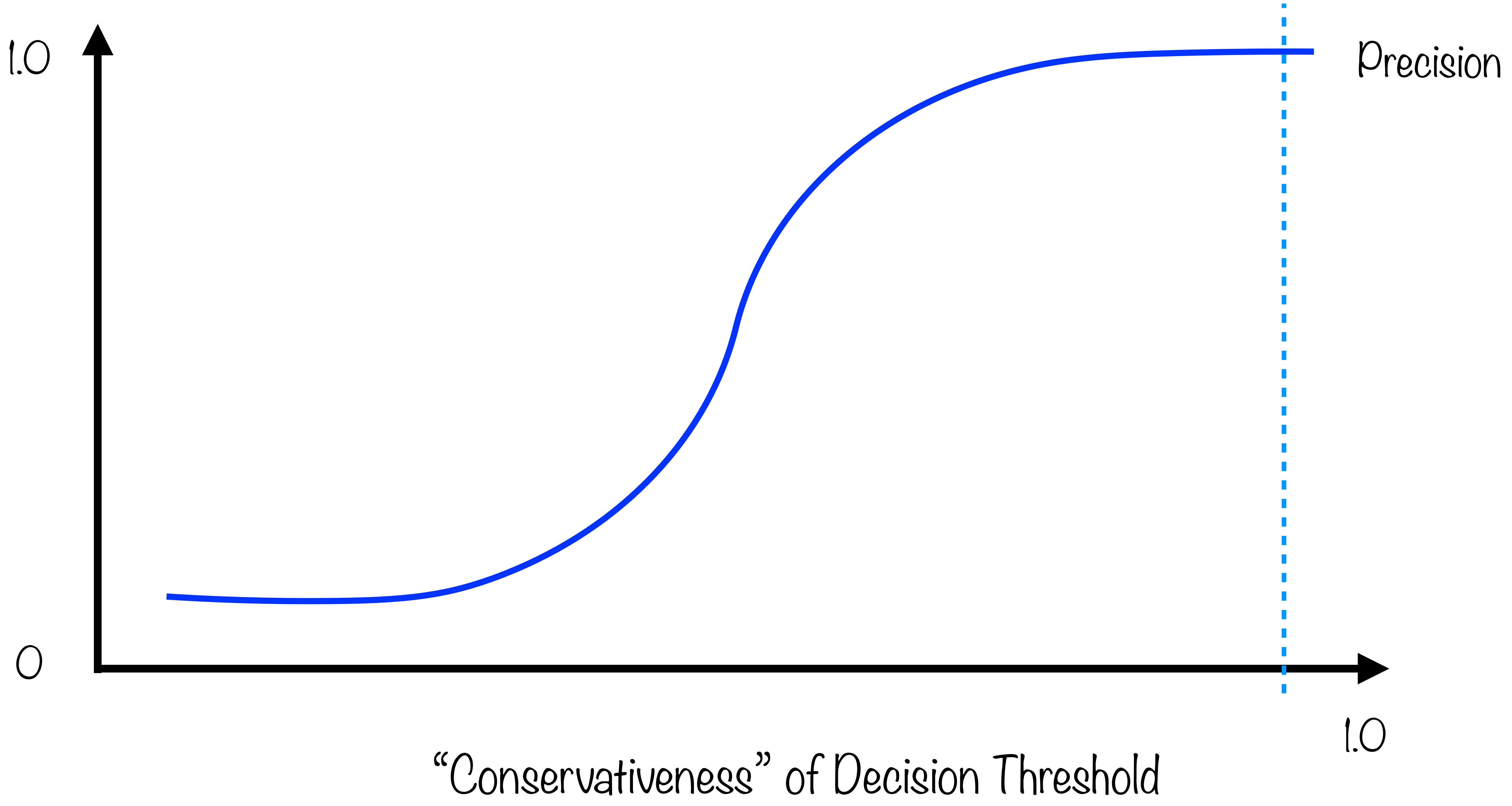
"Always Negative"

$p_{threshold} = 1$

Actual \ Predicted	Cancer	No Cancer
Cancer	TP 0	FN 14
No Cancer	FP 0	TN 1005

- Recall = 0%
- Precision = Infinite
- Classifier too conservative

Precision vs. "Conservativeness"



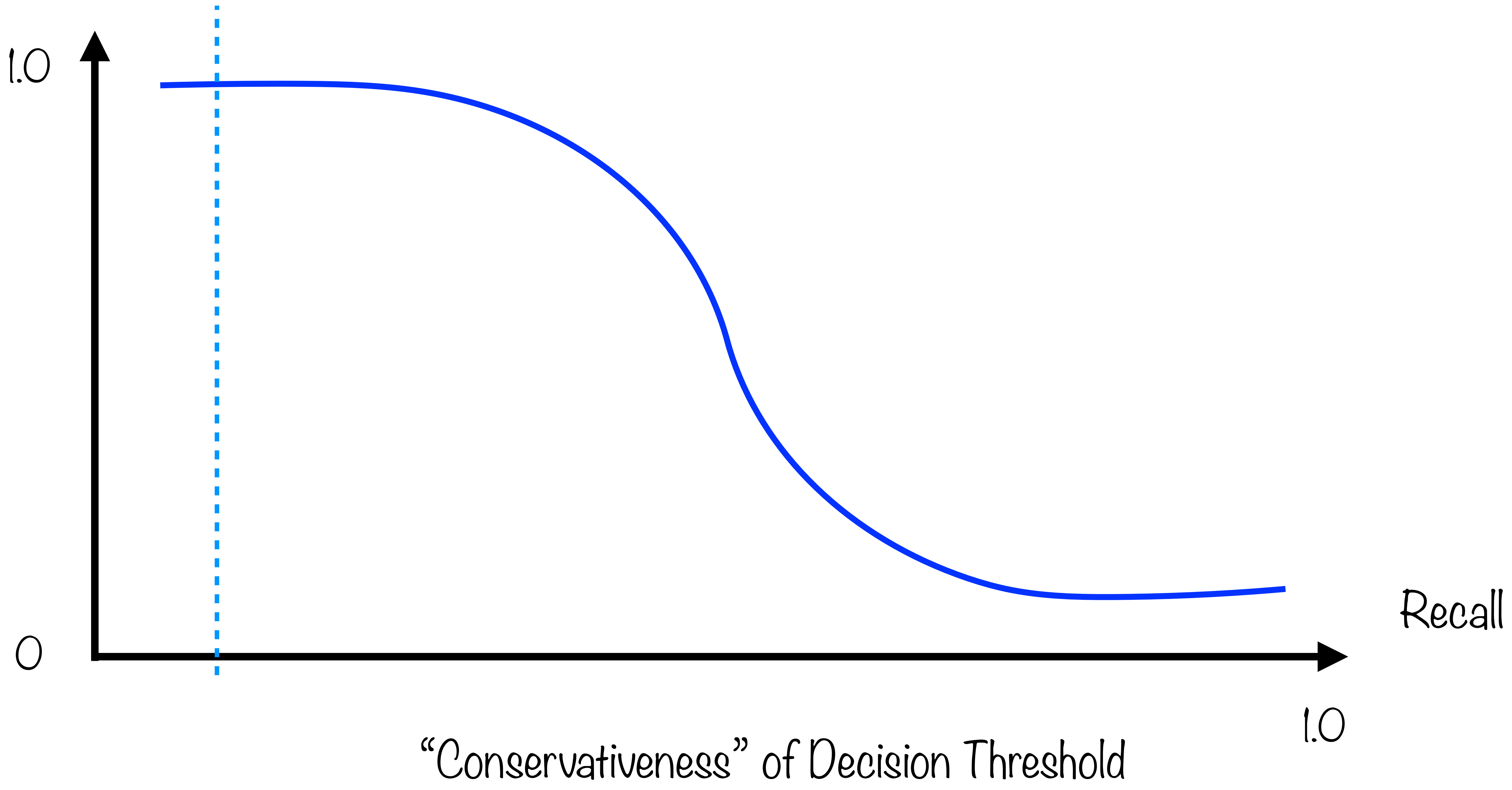
"Always Positive"

$p_{threshold} = 0$

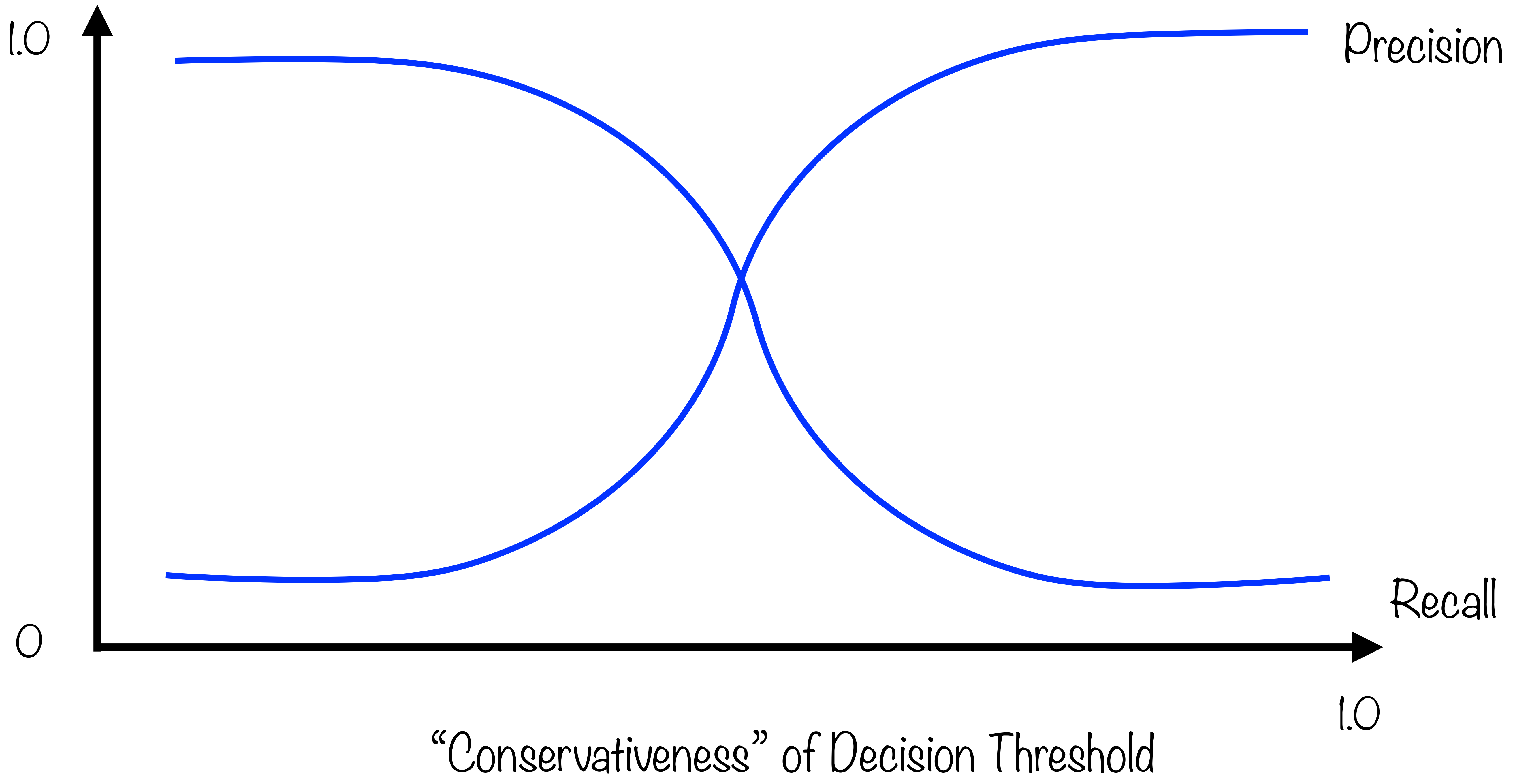
		Predicted: Cancer	Predicted: No Cancer
Actual	Cancer	TP 14	FN 0
	No Cancer	FP 1005	TN 0

- Recall = 100%
- Precision = 14/1019 = 13.7%
- Classifier not conservative enough

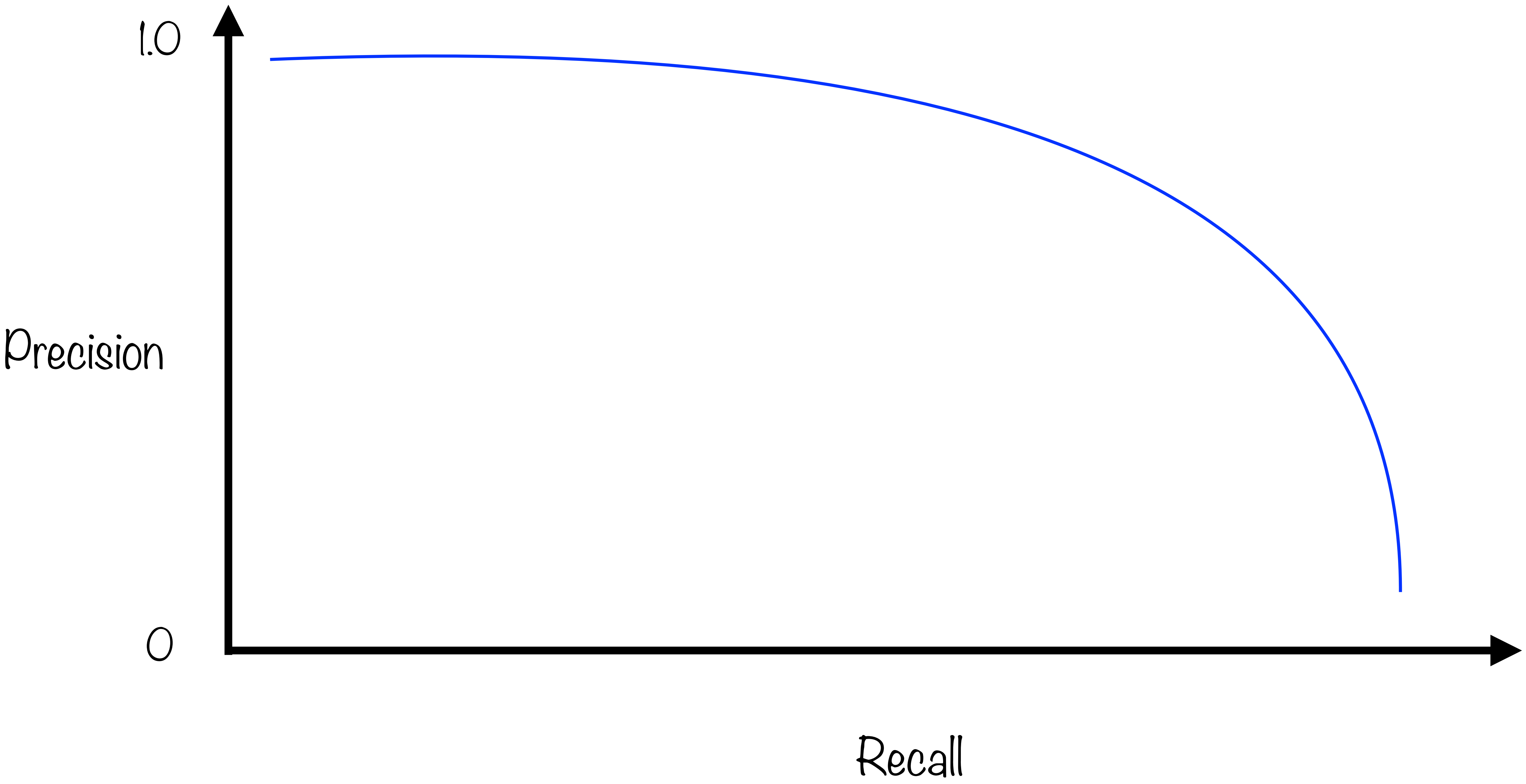
Recall vs."Conservativeness"



Precision-Recall Tradeoff



Precision-Recall Tradeoff



Heuristics to Choose a Model

F1 Score

Harmonic mean of precision and recall

ROC Curve

Plot a curve to maximize true positives, minimize false positives

Heuristics to Choose a Model

F1 Score

Harmonic mean of precision and recall

ROC Curve

Plot a curve to maximize true positives, minimize false positives

F_1 Score

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Harmonic mean of precision, recall
- Closer to lower of two
- Favors even tradeoff

Choosing $P_{threshold}$

Tweak threshold values

Run training by changing threshold values for each execution

Calculate precision, recall

Find values for each training run

Calculate F1 Score

Each training run produces a model, calculate F1 score for each model

High F1 score better

Choose threshold which results in the highest F1 score

Heuristics to Choose a Model

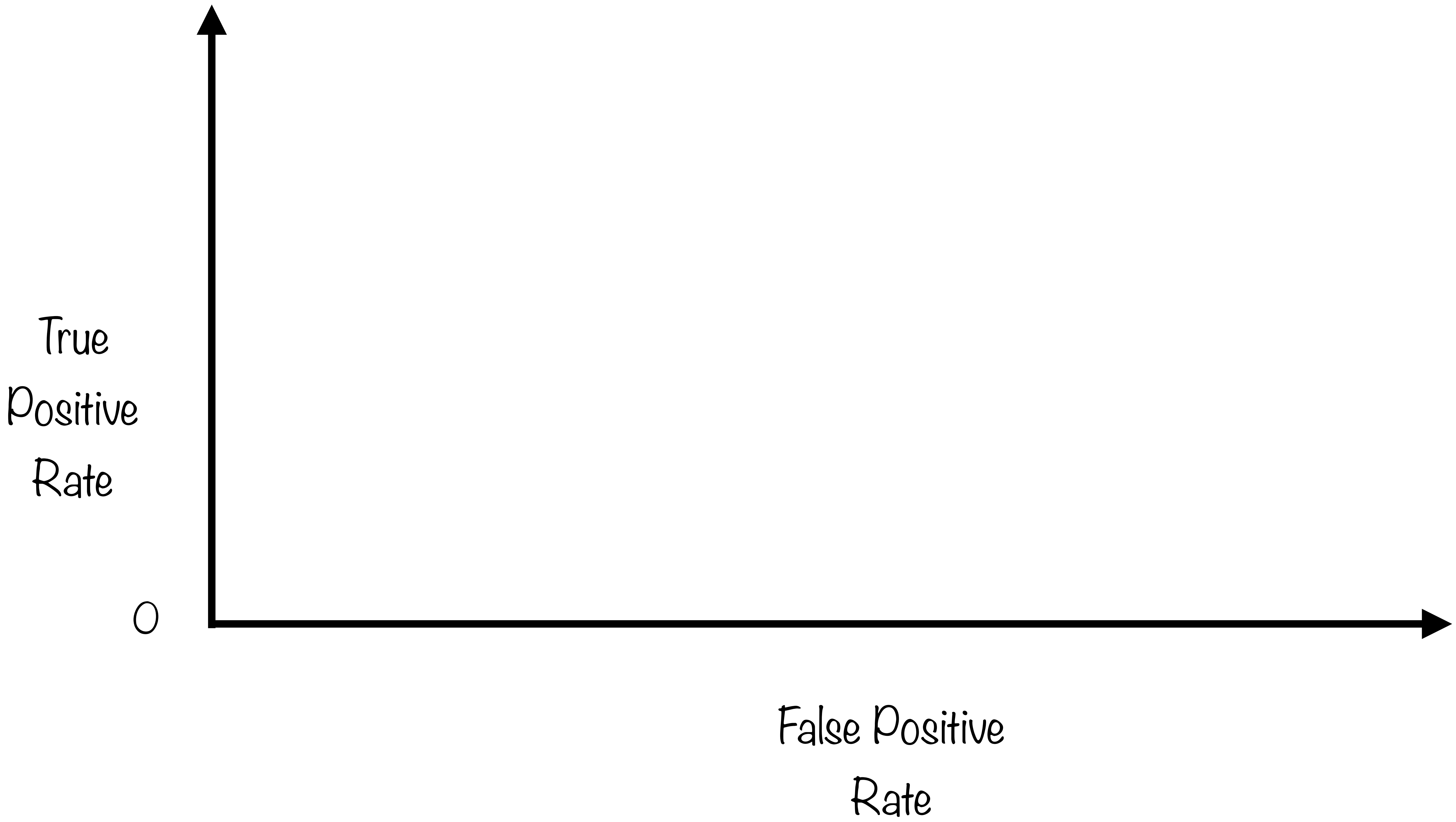
F1 Score

Harmonic mean of precision and recall

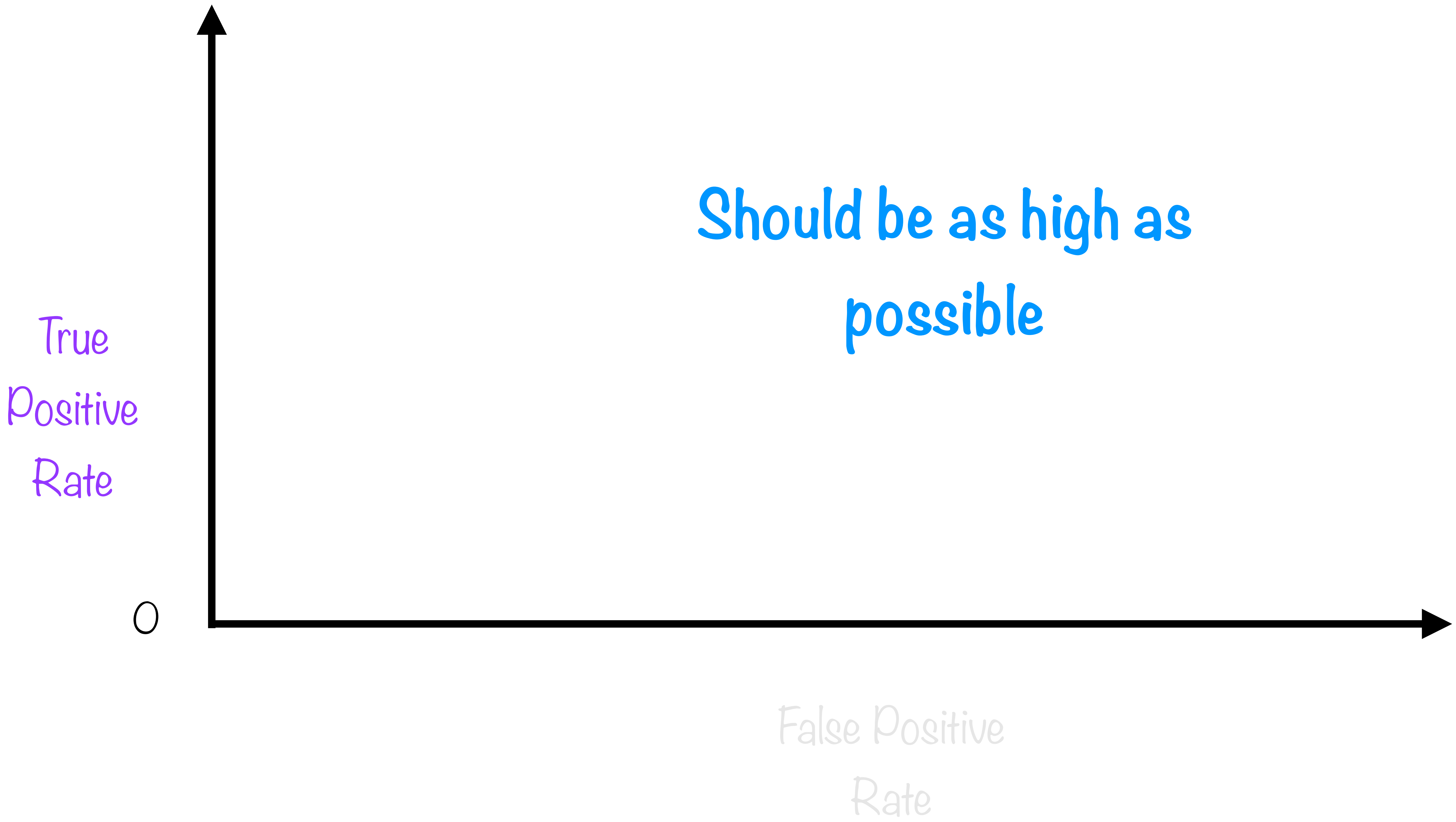
ROC Curve

Plot a curve to maximize true positives, minimize false positives

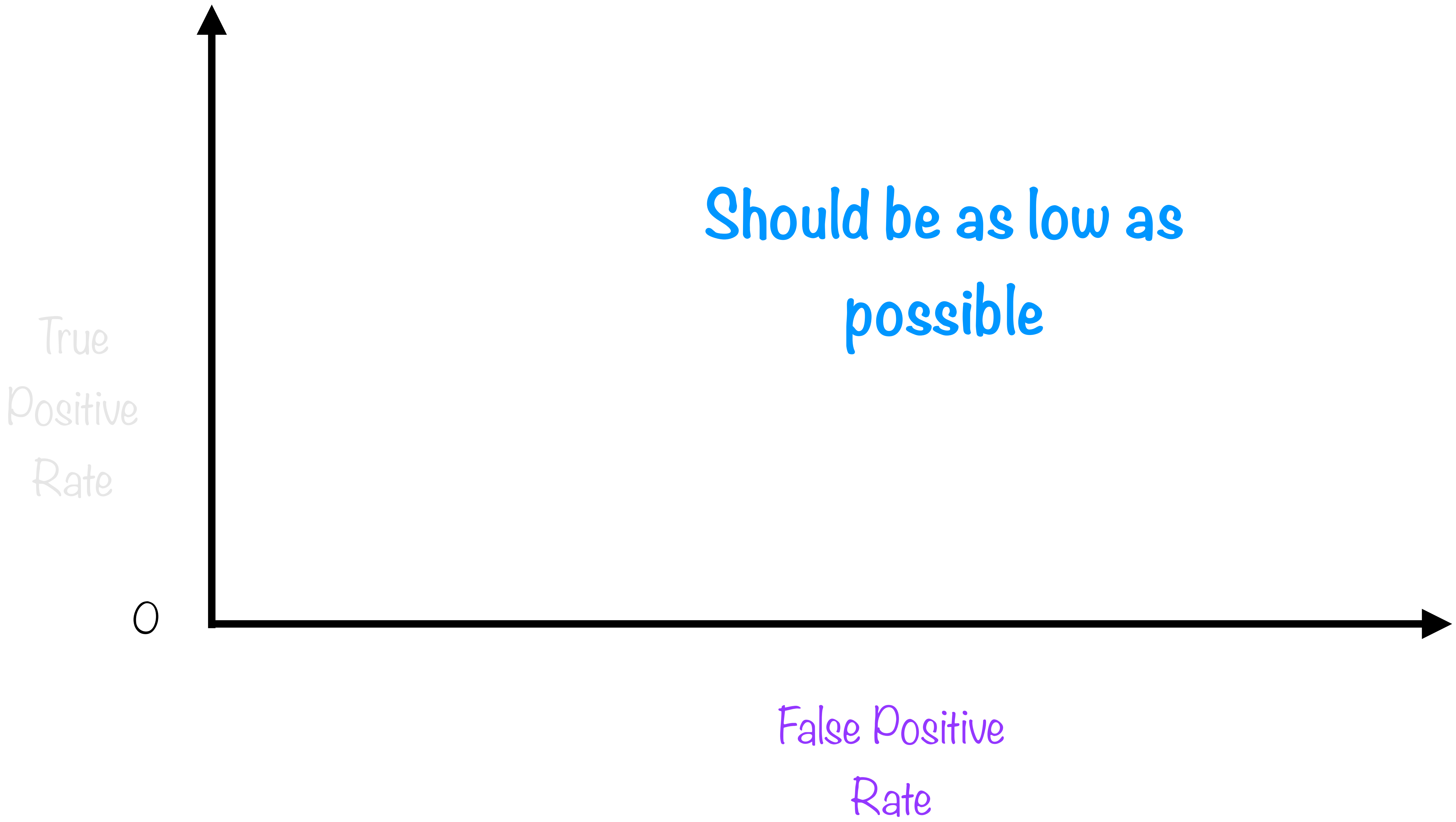
Choosing $P_{threshold}$



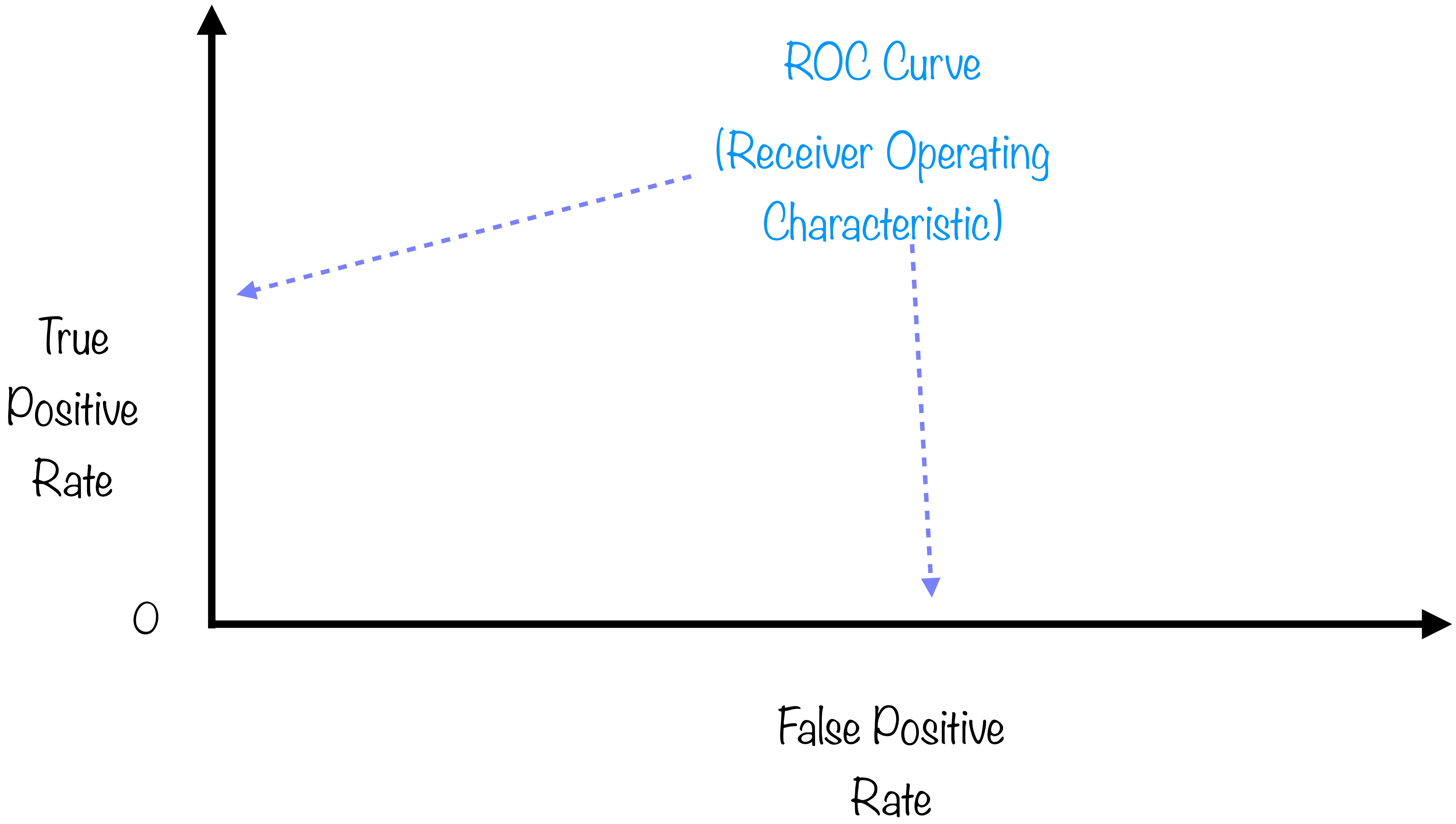
Choosing $P_{threshold}$



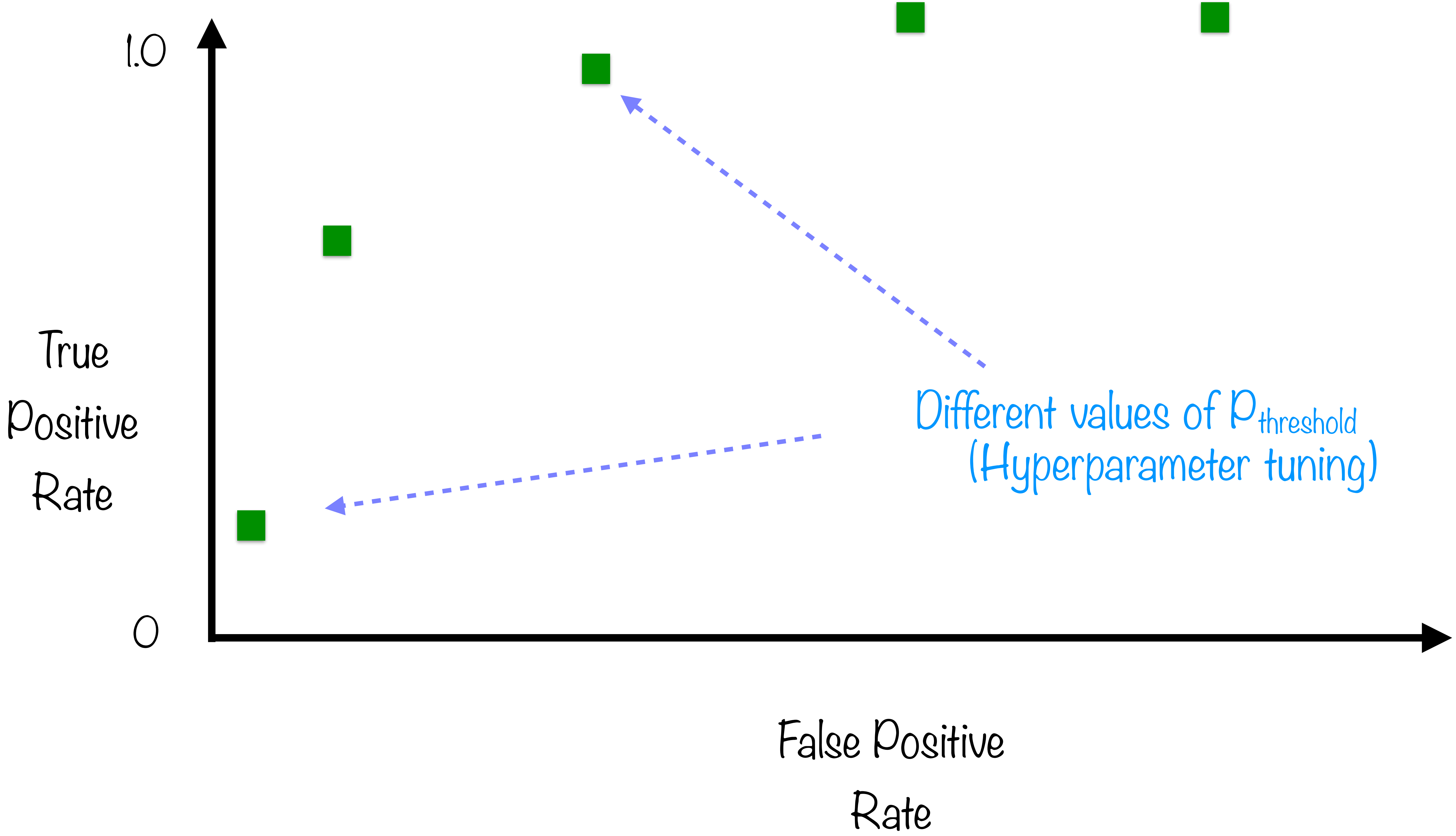
Choosing $P_{threshold}$



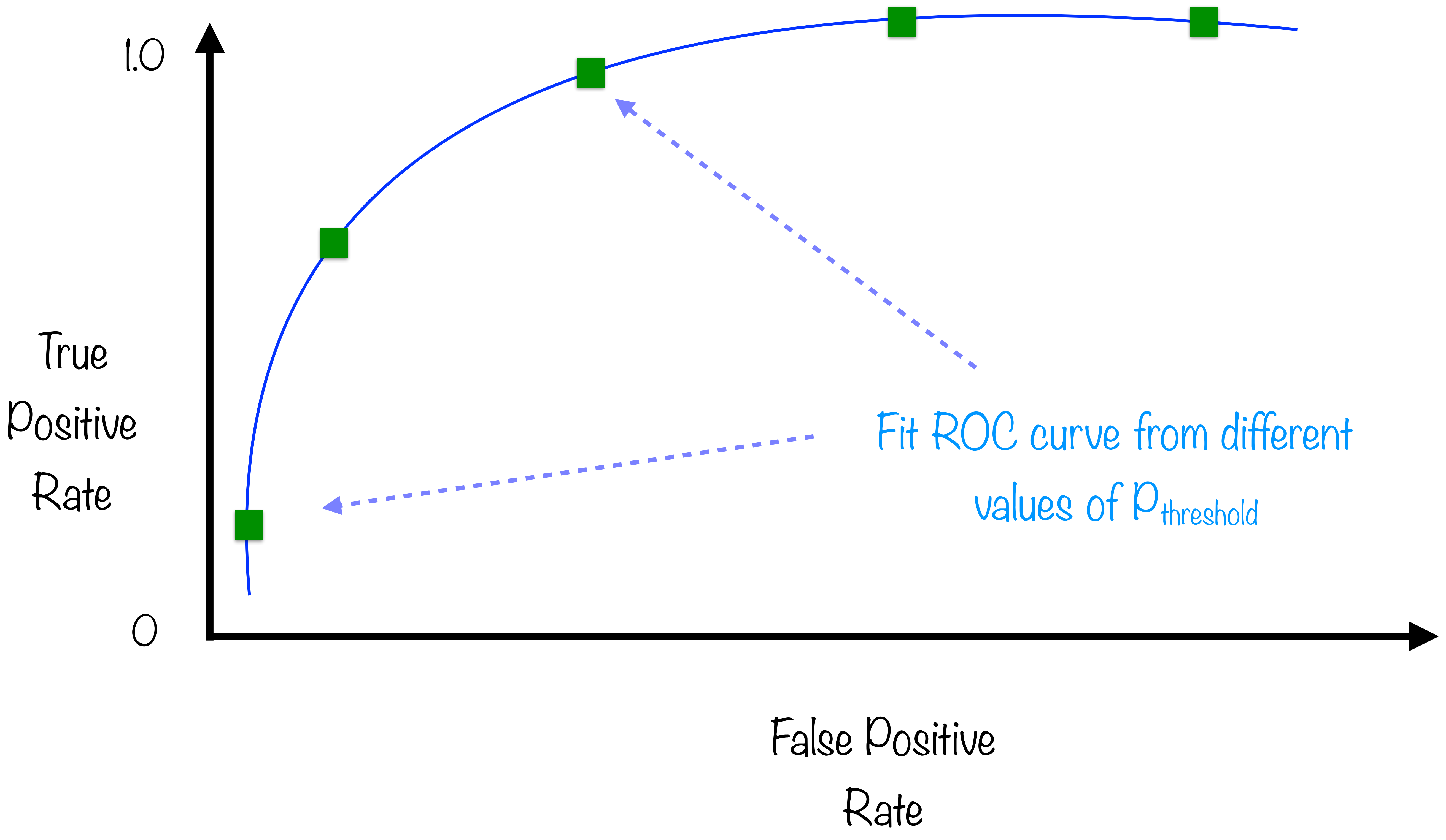
Choosing $P_{threshold}$



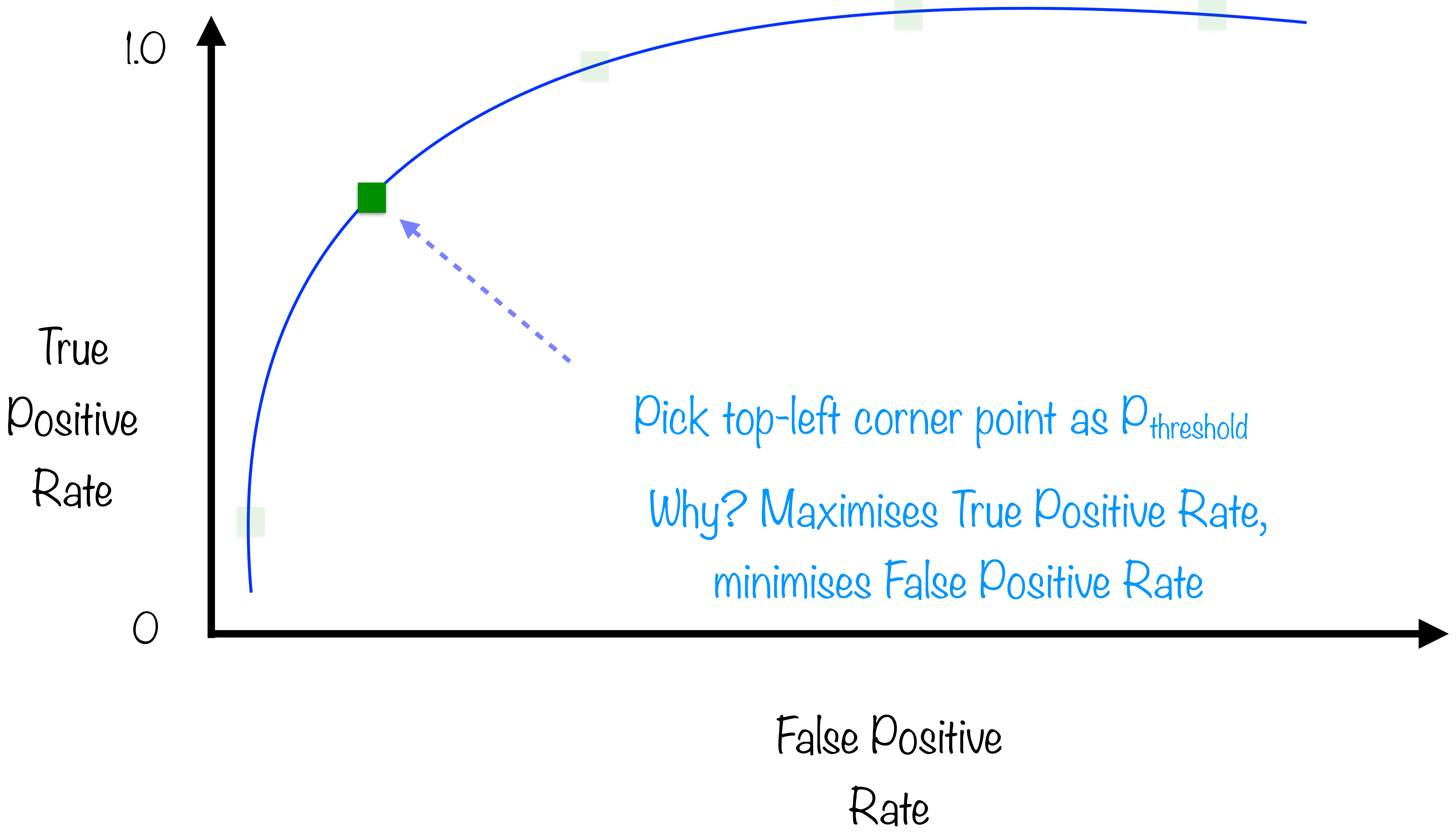
Choosing $P_{threshold}$



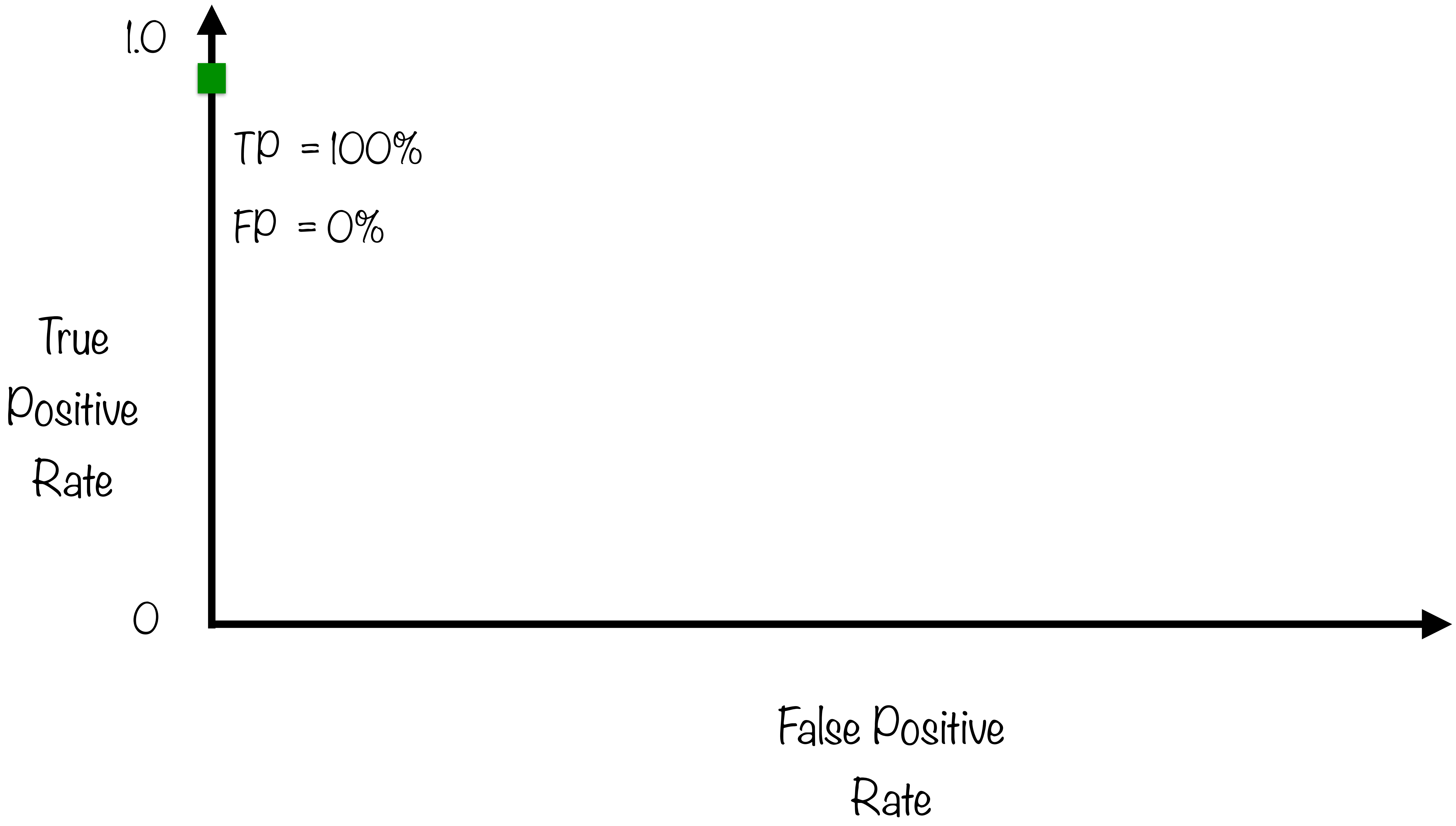
Choosing $P_{threshold}$



ROC Curve


1.0
0
True Positive Rate
False Positive Rate
Pick top-left corner point as $p_{threshold}$
Why? Maximises True Positive Rate,
minimises False Positive Rate

ROC of Perfect Classifier



ROC of Random Classifier