GENERAL ASSEMBLY

EVALUATING MODEL FIT: ROC-AUC

Matt Speck, Data Science Immersive h/t Joseph Nelson

AGENDA

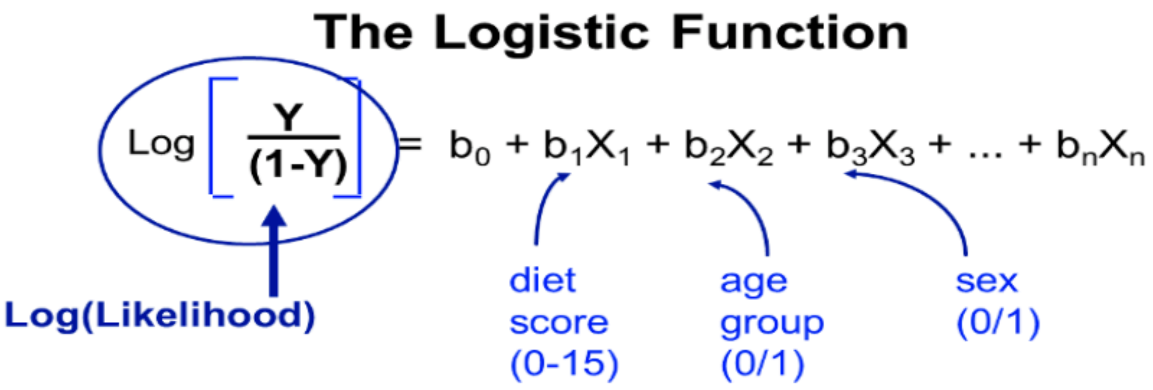
- Logistic Regression Quick Review
- Confusion Matrix
- Sensitivity & Specificity Tradeoff
- ROC Curve

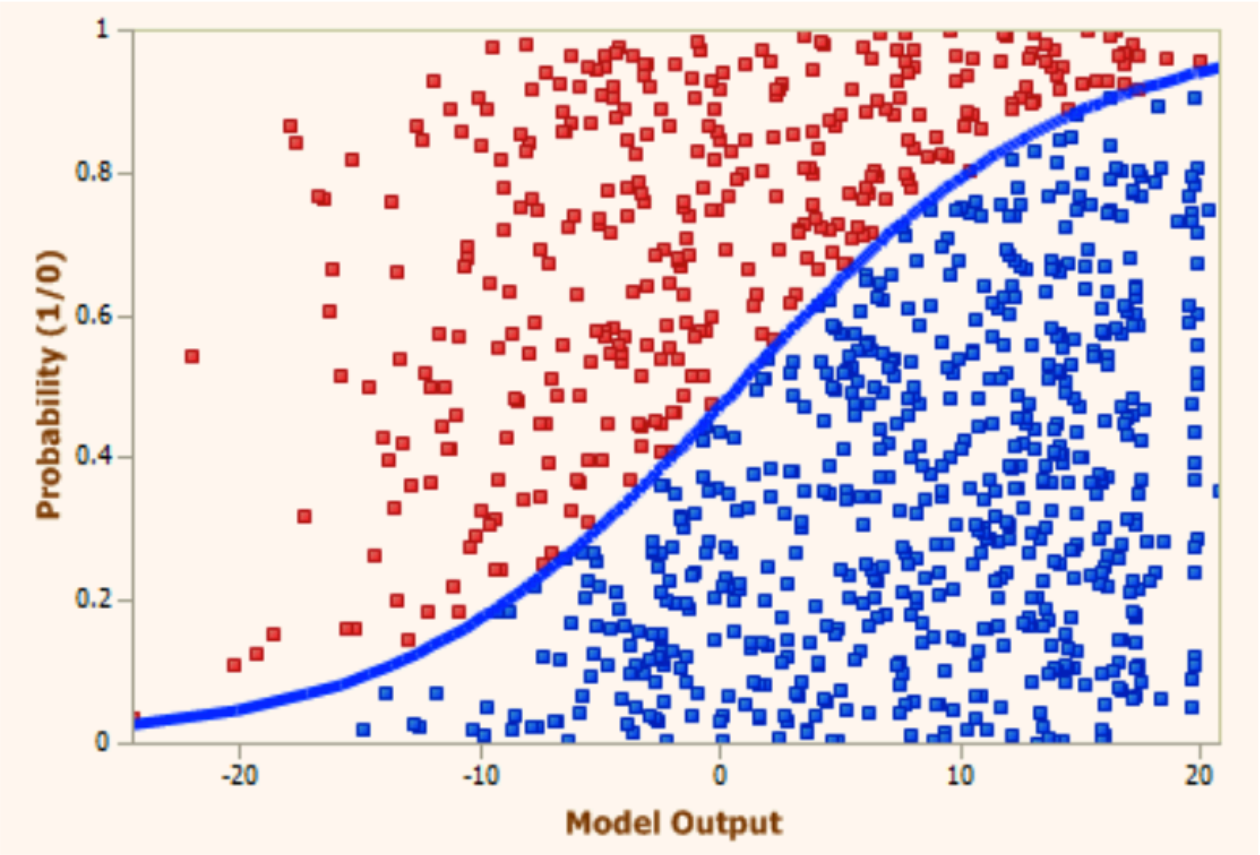
QUICK REVIEW: LOGISTIC REGRESSION

- What is a logistic regression?

QUICK REVIEW: LOGISTIC REGRESSION

- Logistic regression is a modeling tactic where our dependent variable is bound by [0,1] used for class predictions
- If a value exceeds some threshold, we can say the outputted response is of class=1, or of class=0 if we're below some threshold





FROM ACCURACY ONWARDS

- So far we've been evaluating logistic regression by looking at accuracy. But there's a lot more we can do…

CONFUSION MATRIX

- A confusion matrix is a table of how we plot the output of our classifier.
- Example: A confusion matrix showing the output of a classifier to predict presence of a disease

- How many classes are there?
- How many patients?
- How many times is a disease predicted?
- How many patients actually have the disease?

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

CONFUSION MATRIX

- A confusion matrix is a table of how we plot the output of our classifier.
- Example: A confusion matrix showing the output of a classifier to predict presence of a disease

- True Positives?
- True Negatives?
- False Positives?
- False Negatives?
- Accuracy? (How often was I right?)
- Misclassification? (How often was I wrong?)

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

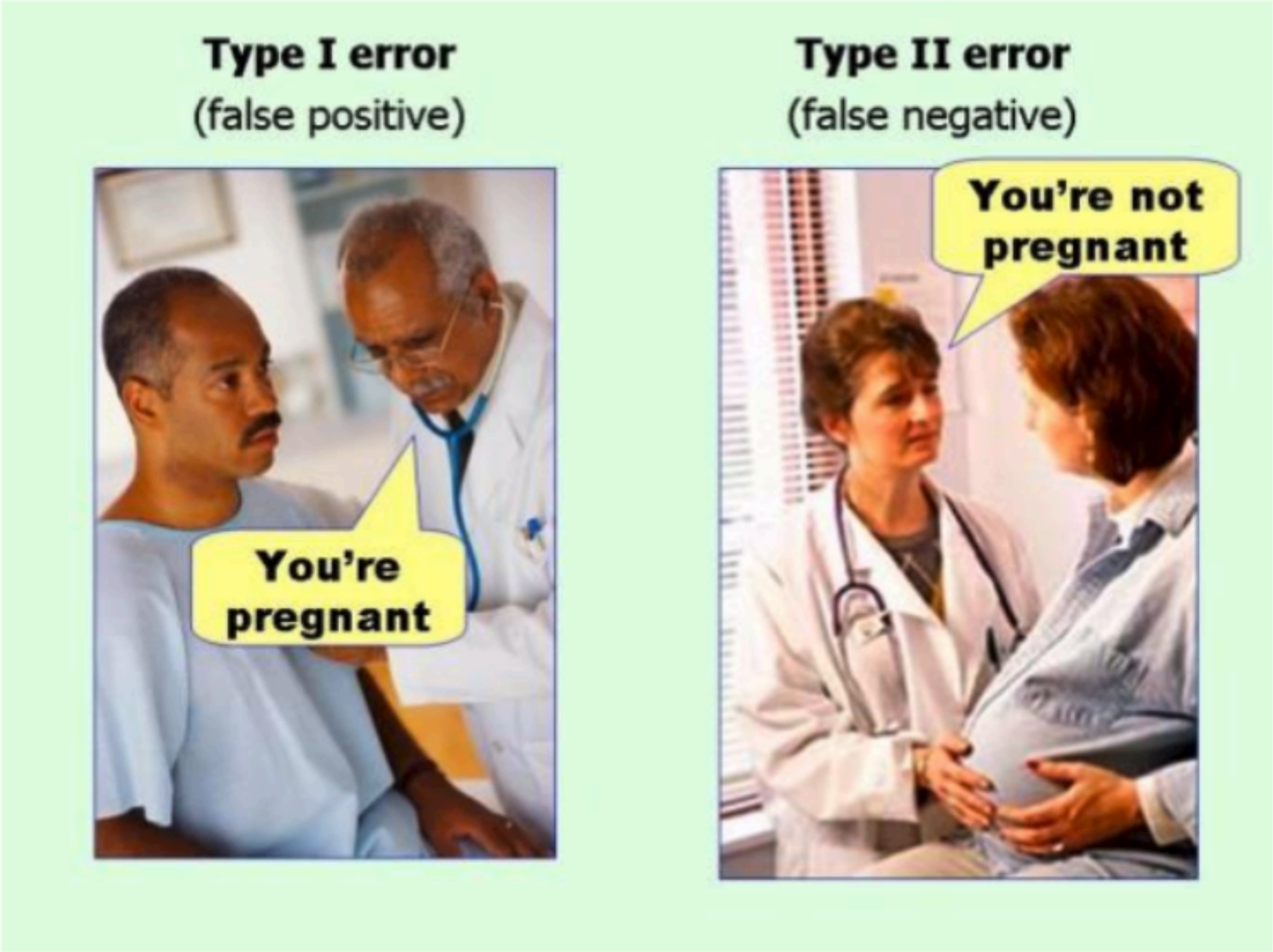
CONFUSION MATRIX

- Accuracy?

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

CONFUSION MATRIX



ALL OF THE TERMS

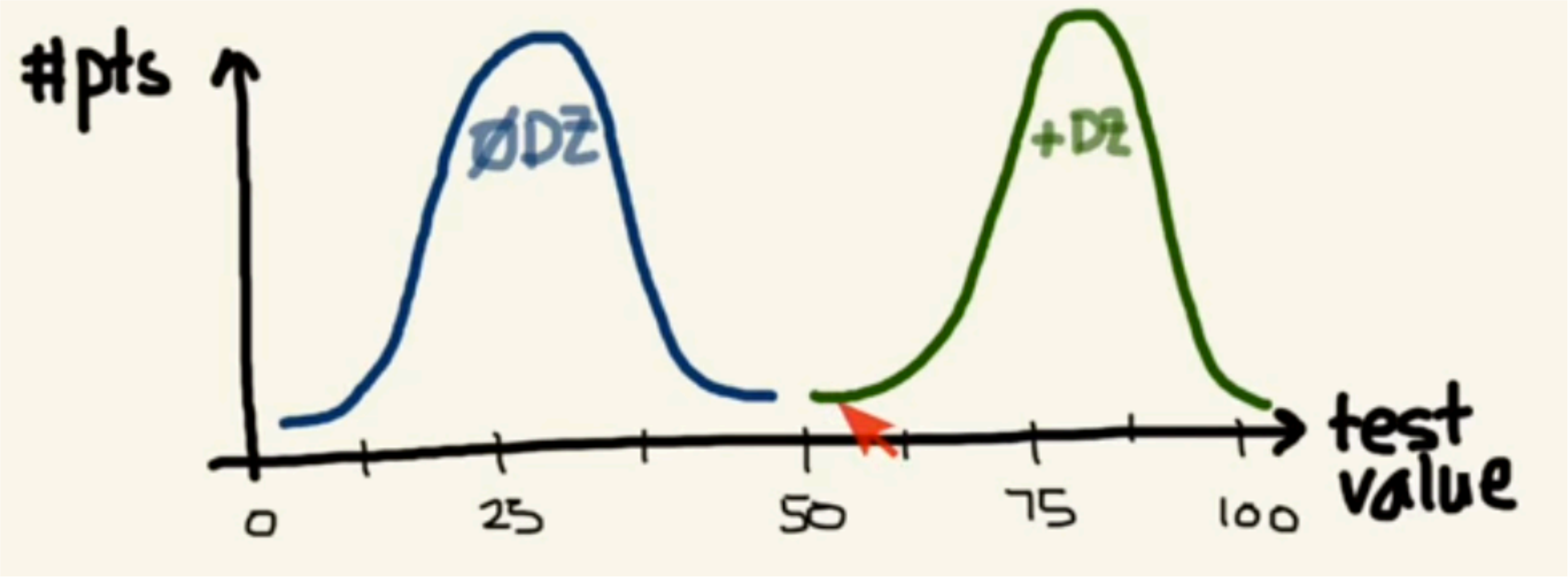
- True Positives
- True Negatives
- False Positives
- False Negatives
- **Sensitivity: TP/Actual Yes**
- Aka Recall aka TPR
- **Specificity: TN/Actual No**
- Aka TNR
- **Precision: (TP/(TP+FP)**

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

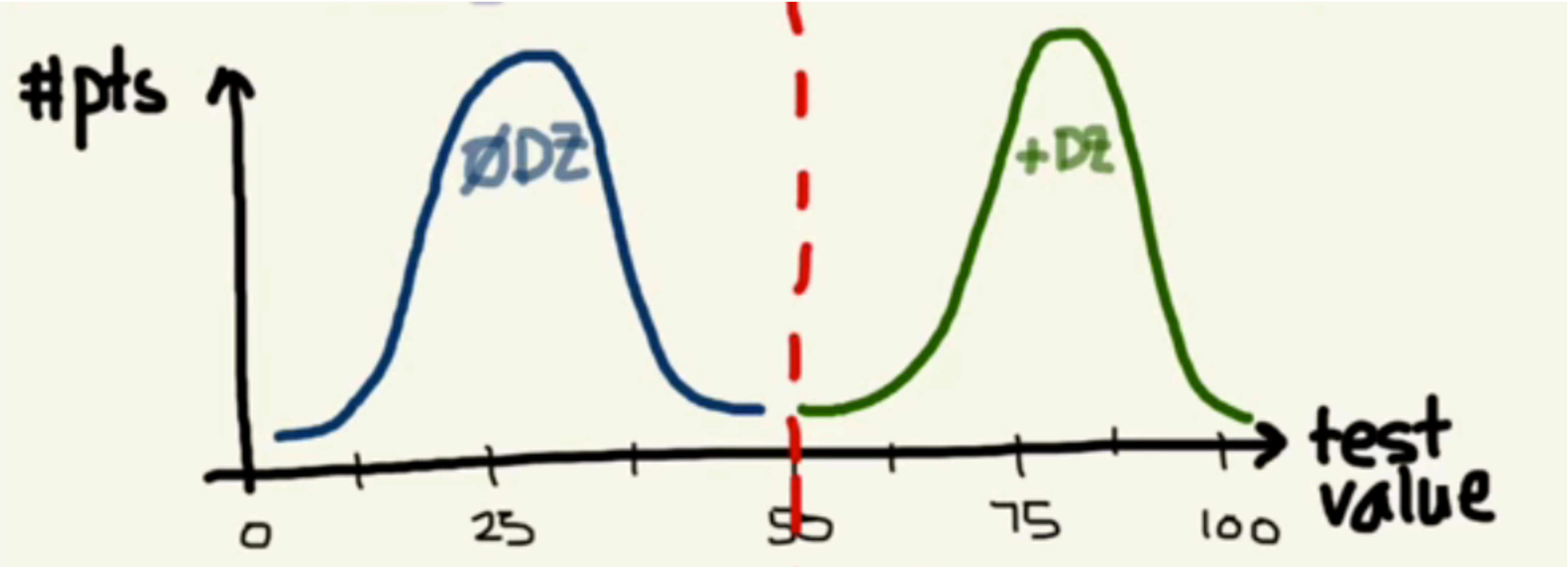
CONFUSION MATRIX

- **Check:** We built a model to predict whether a person would vote in the upcoming election. Out of 220 people, we correctly predicted that 130 would vote, incorrectly predicted that 40 would vote, correctly predicted that 20 would not vote, and incorrectly predicted that 30 would not vote.
- Fill out a confusion matrix and calculate Sensitivity, Specificity, and Precision
- Sensitivity: (TP/(TP+FN)
- Specificity: (TN/(TN+FP)
- Precision: (TP/(TP+FP)

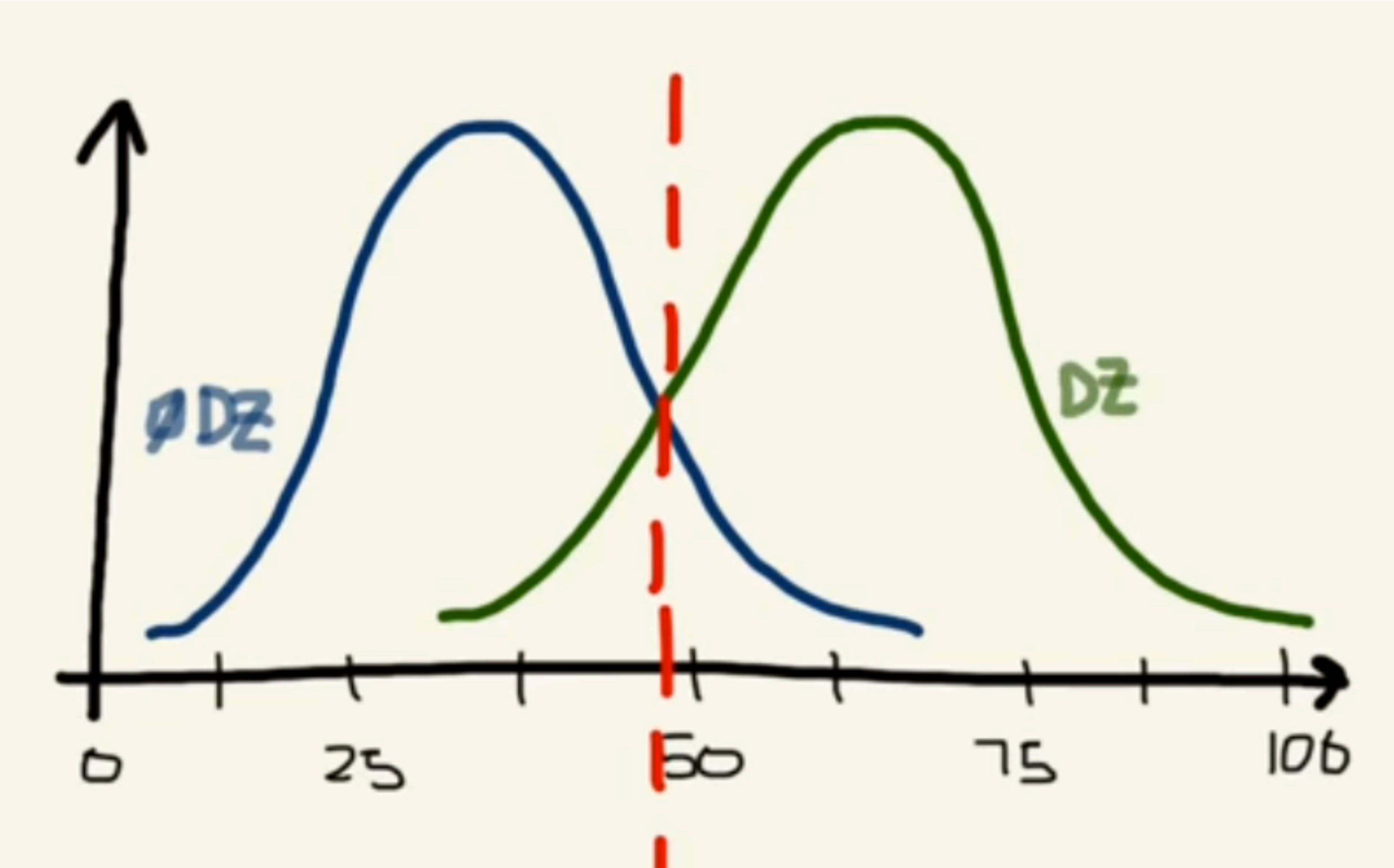
SENSITIVITY/SPECIFICITY TRADE OFF



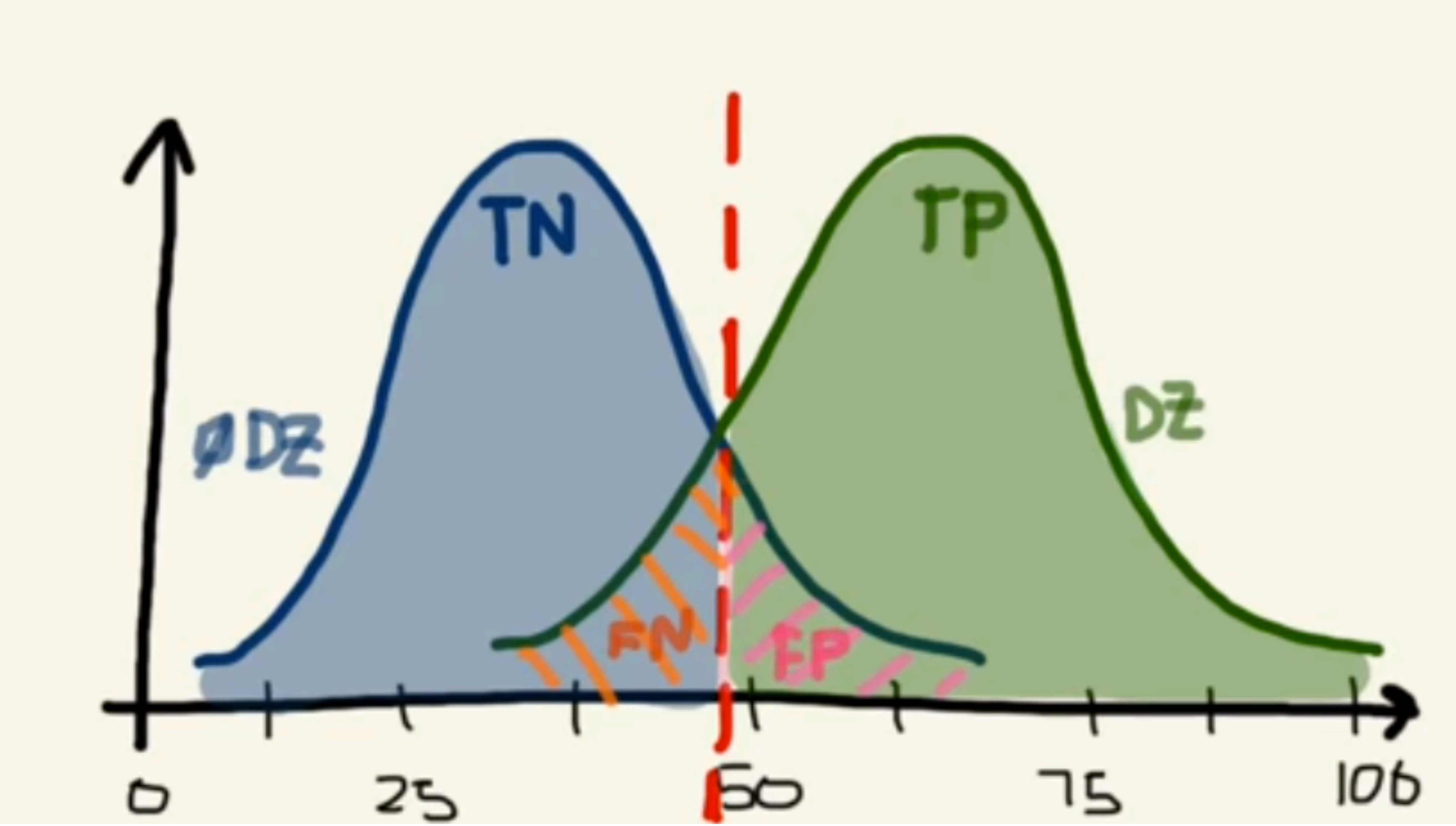
SENSITIVITY/SPECIFICITY TRADE OFF



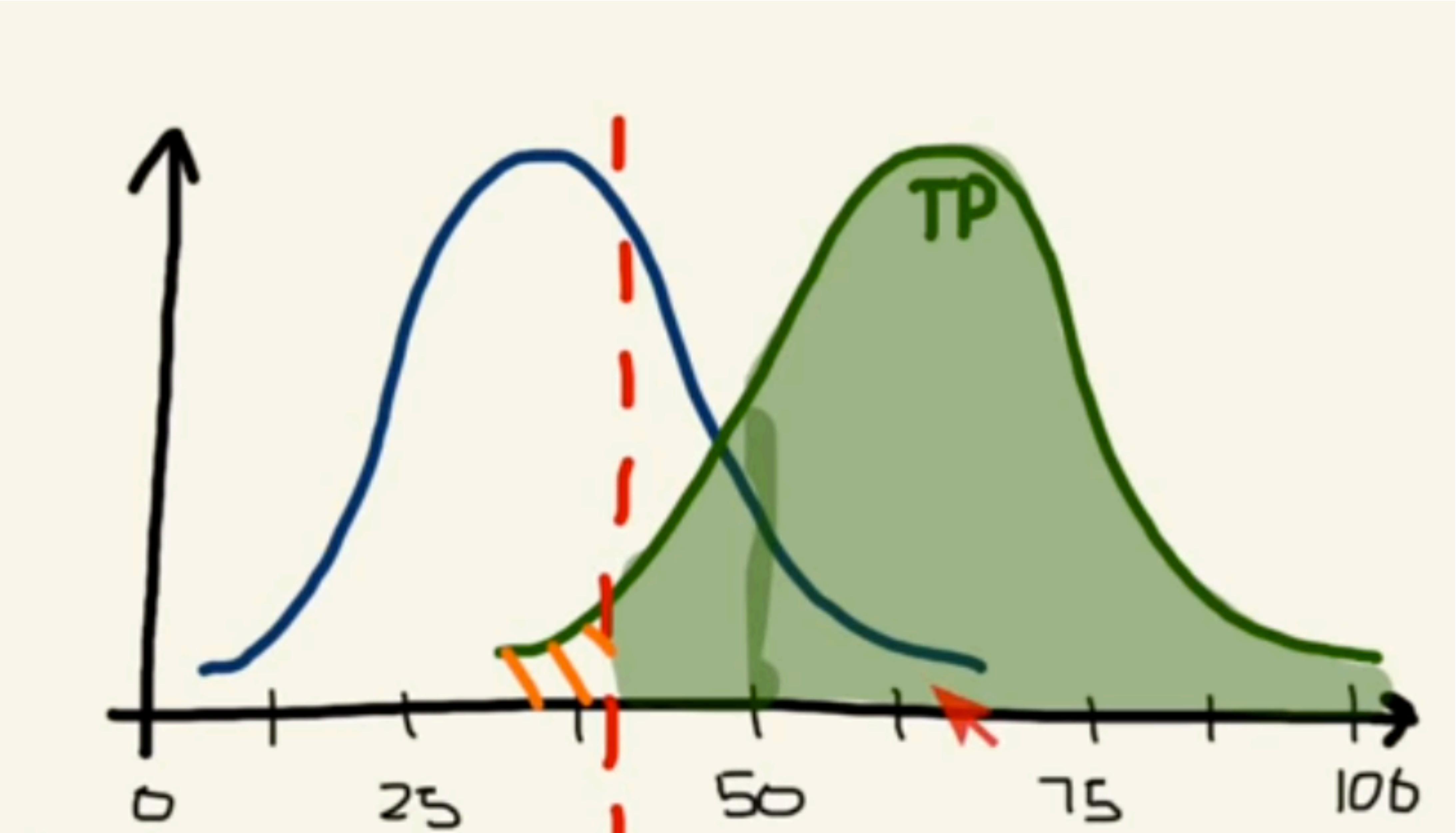
SENSITIVITY/SPECIFICITY TRADE OFF



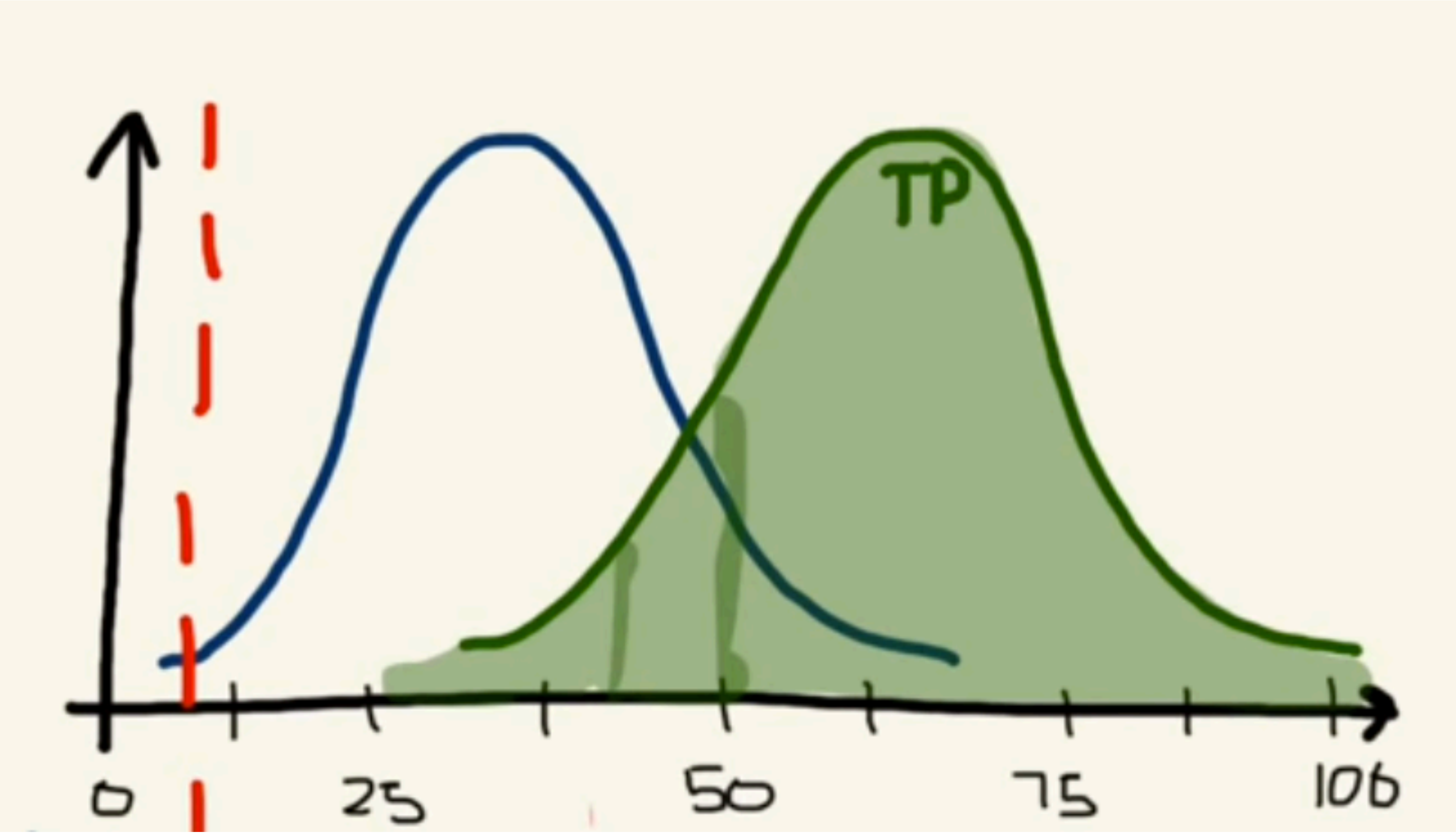
SENSITIVITY/SPECIFICITY TRADE OFF



SENSITIVITY/SPECIFICITY TRADE OFF

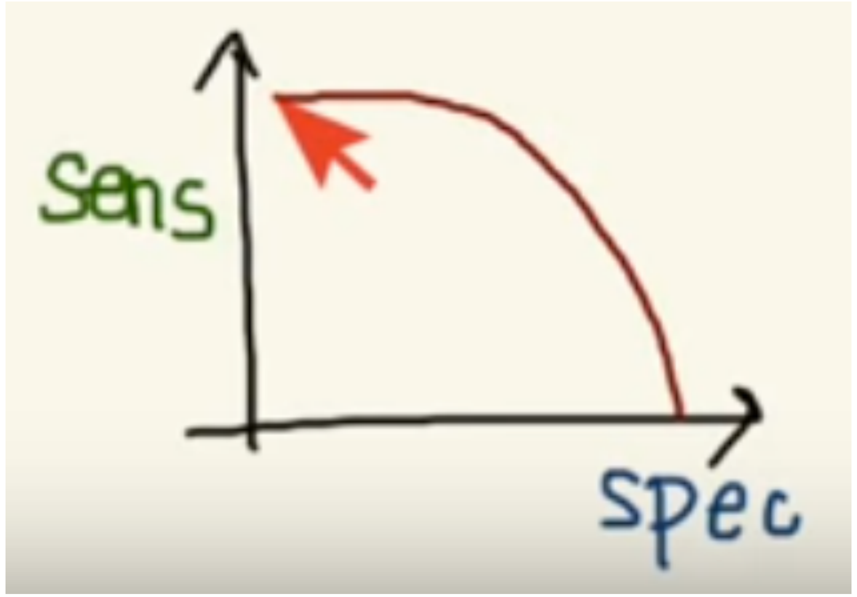


SENSITIVITY/SPECIFICITY TRADE OFF



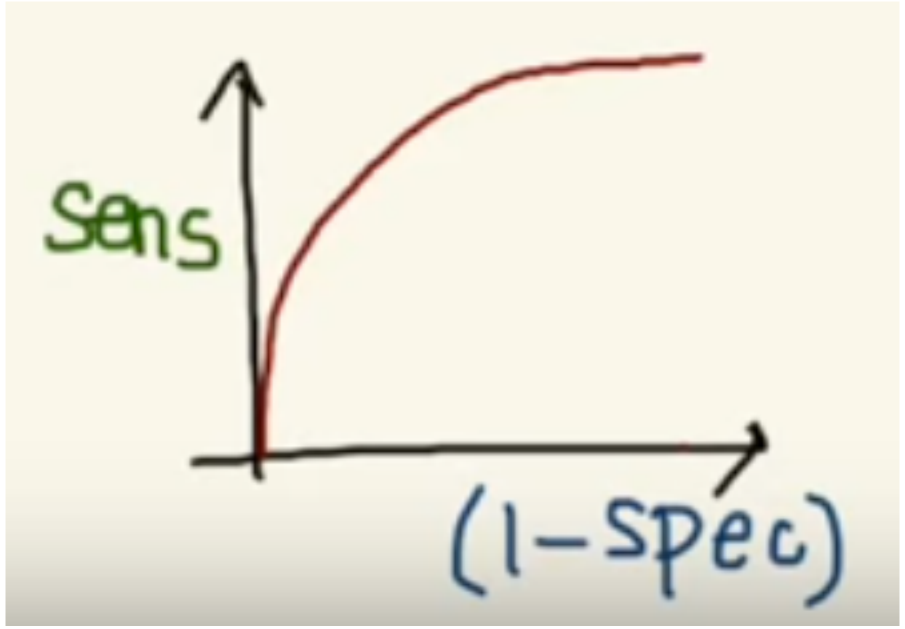
ROC-AUC

- Sensitivity and Specificity move in opposite directions – but there is an optimum value to be found
- Area under the curve – plotting the sensitivity and specificity against one another yields the strength of our classifier (we want to bring this value to one)
- The most popular AUC is the Receiver Operating Characteristic (ROC) Curve

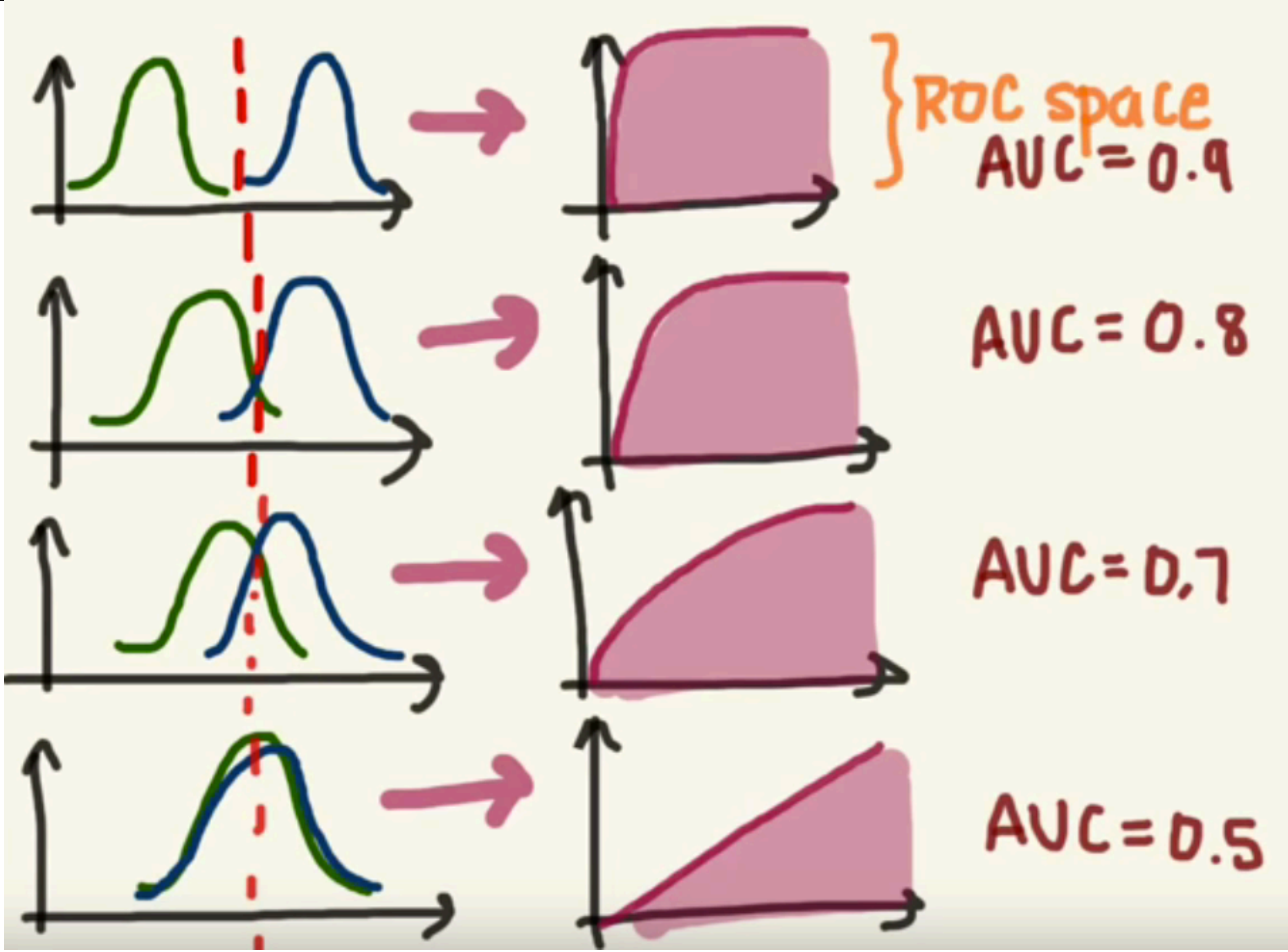


ROC-AUC

- We plot Sensitivity vs 1-Specificity so that the two move in the same direction
- The ROC curve compares the true positive rate against the false positive rate. It is unaffected by the distribution of class labels since it is only comparing the correct vs. incorrect label assignments for one class.
- The ROC curve models what sensitivity and specificity will be at various thresholds



ROC-AUC



ROC-AUC

- Let's look at an applet to better understand how the ROC-AUC works: http://www.navan.name/roc/