GENERAL ASSEMBLY

EVALUATING MODEL FIT: ROC-AUC

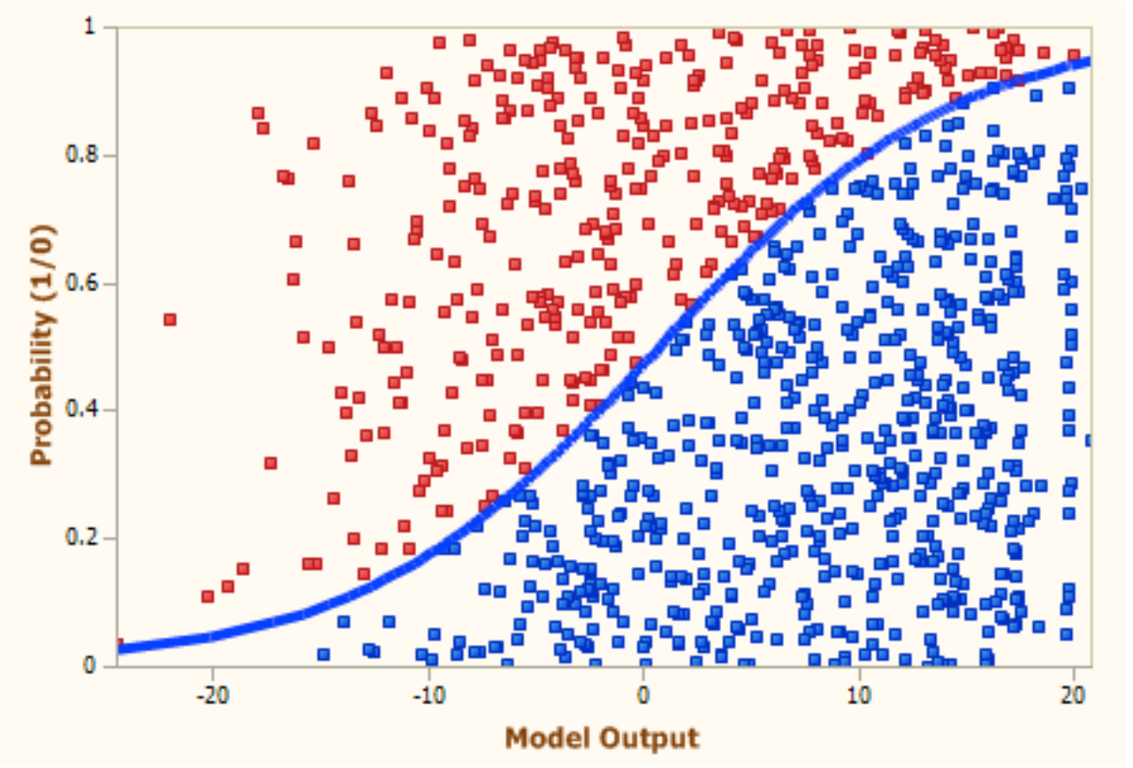
Joseph Nelson, Data Science Immersive

AGENDA

- Logistic Regression Quick Review
- Confusion Matrix
- Sensitivity & Specificity Tradeoff
- AUC and ROC Graphs
- Coding Implementation

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



The Logistic Function

$$\text{Log}\left[\frac{Y}{(1-Y)}\right] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$$

Log(Likelihood)

diet score (0-15)

age group (0/1)

sex (0/1)

CONFUSION MATRIX

- A confusion matrix is a table of how we plot the output of our classifier

- 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

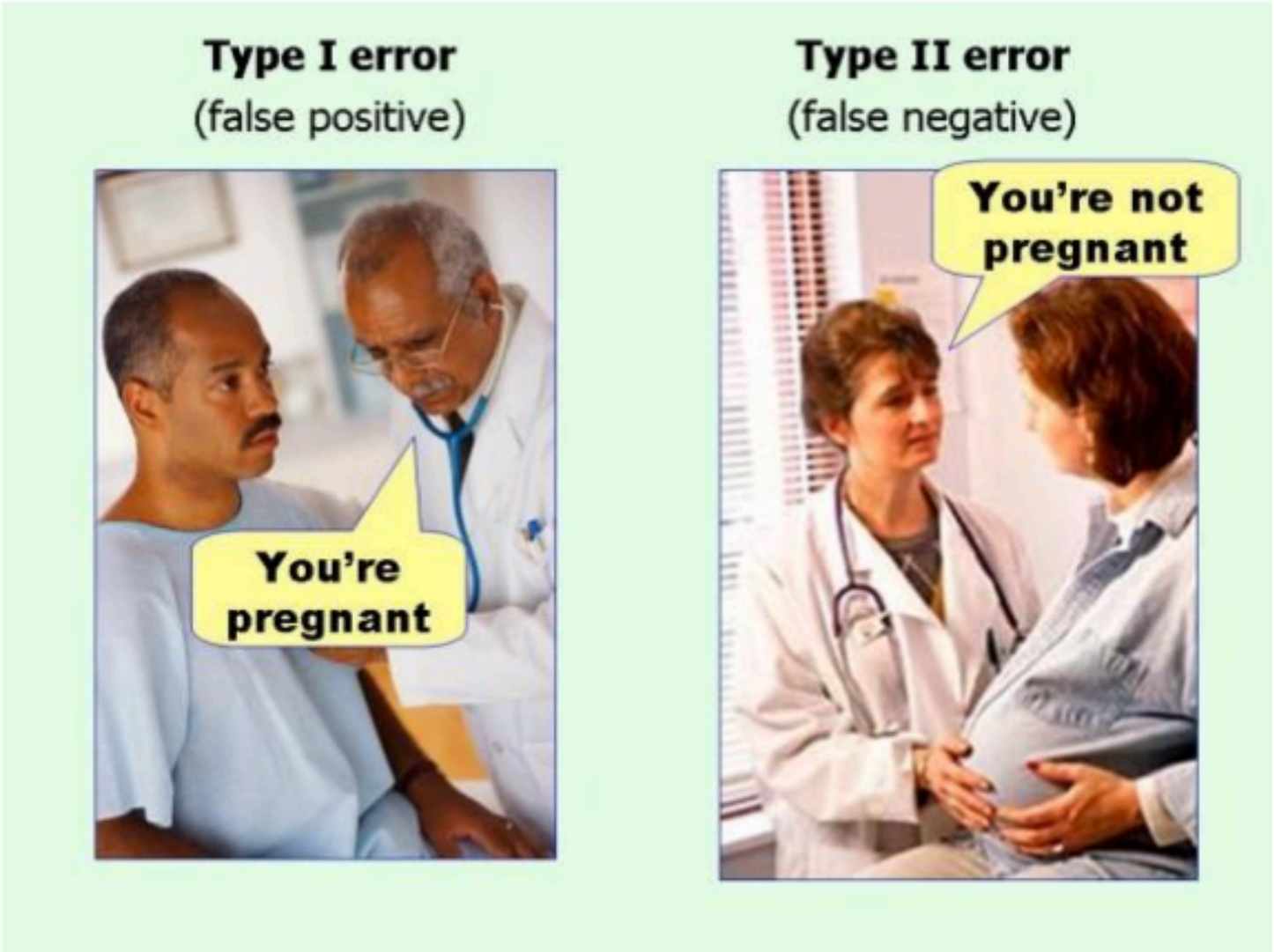
CONFUSION MATRIX

- A confusion matrix is a table of how we plot the output of our classifier
- True Positives
- True Negatives
- False Positives
- False Negatives

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

- Accuracy: Overall, how often is this correct?
- Misclassification: Overall, how often is this wrong?

CONFUSION MATRIX



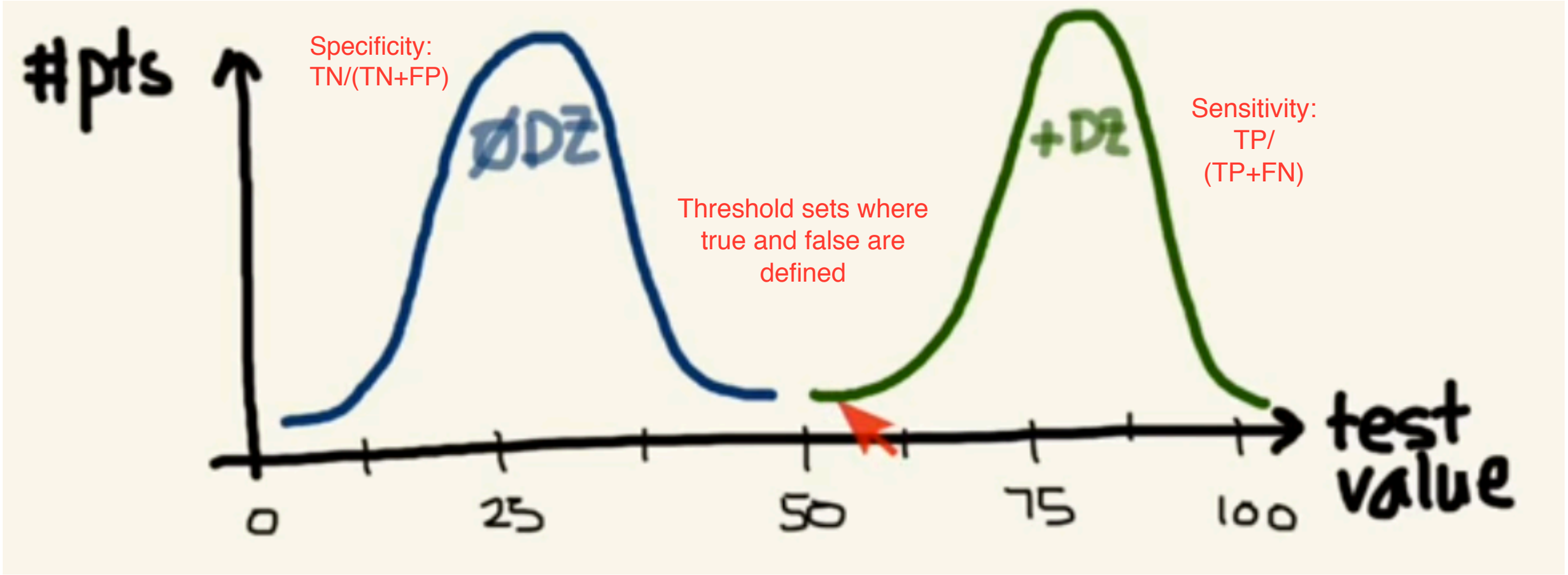
CONFUSION MATRIX

- **Sensitivity**: when actual value is positive, how often is our prediction correct?
- (True Positive/Recall)
- **Specificity**: when actual value is negative, how often is our prediction correct?
- **False Positive Rate:** When actual value is negative, how often is our prediction wrong?

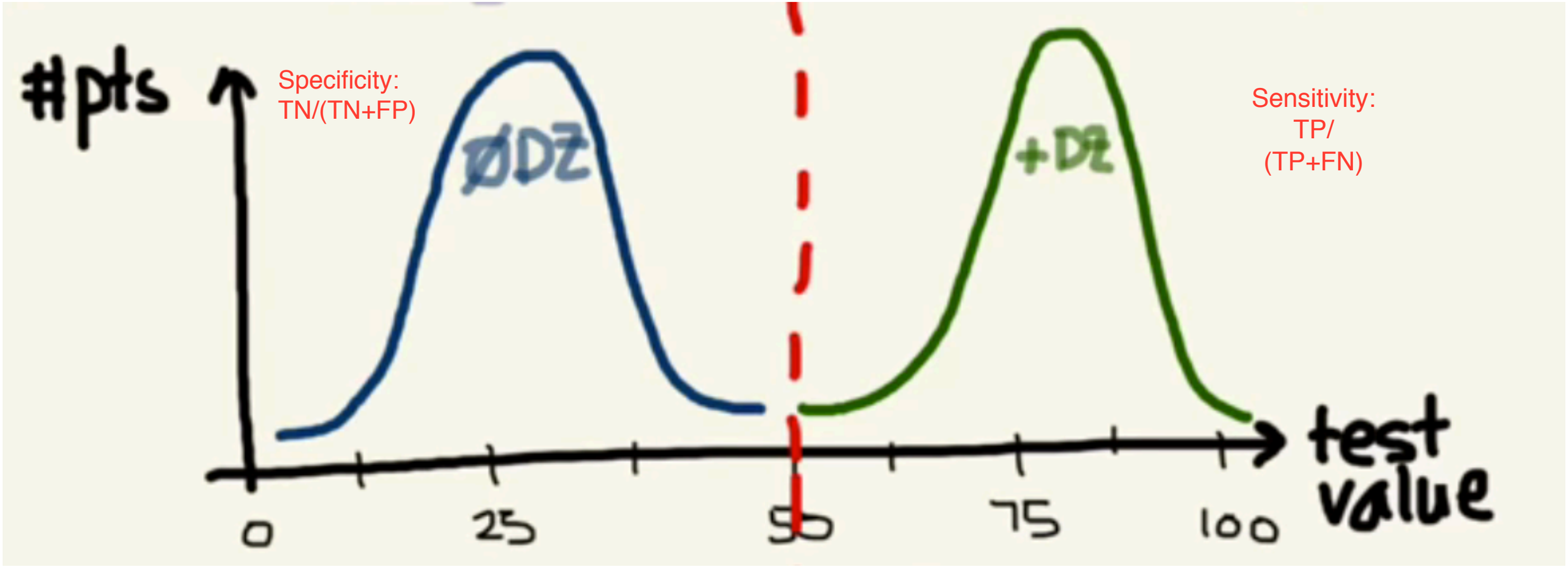
n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

SENSITIVITY/SPECIFICITY TRADE OFF

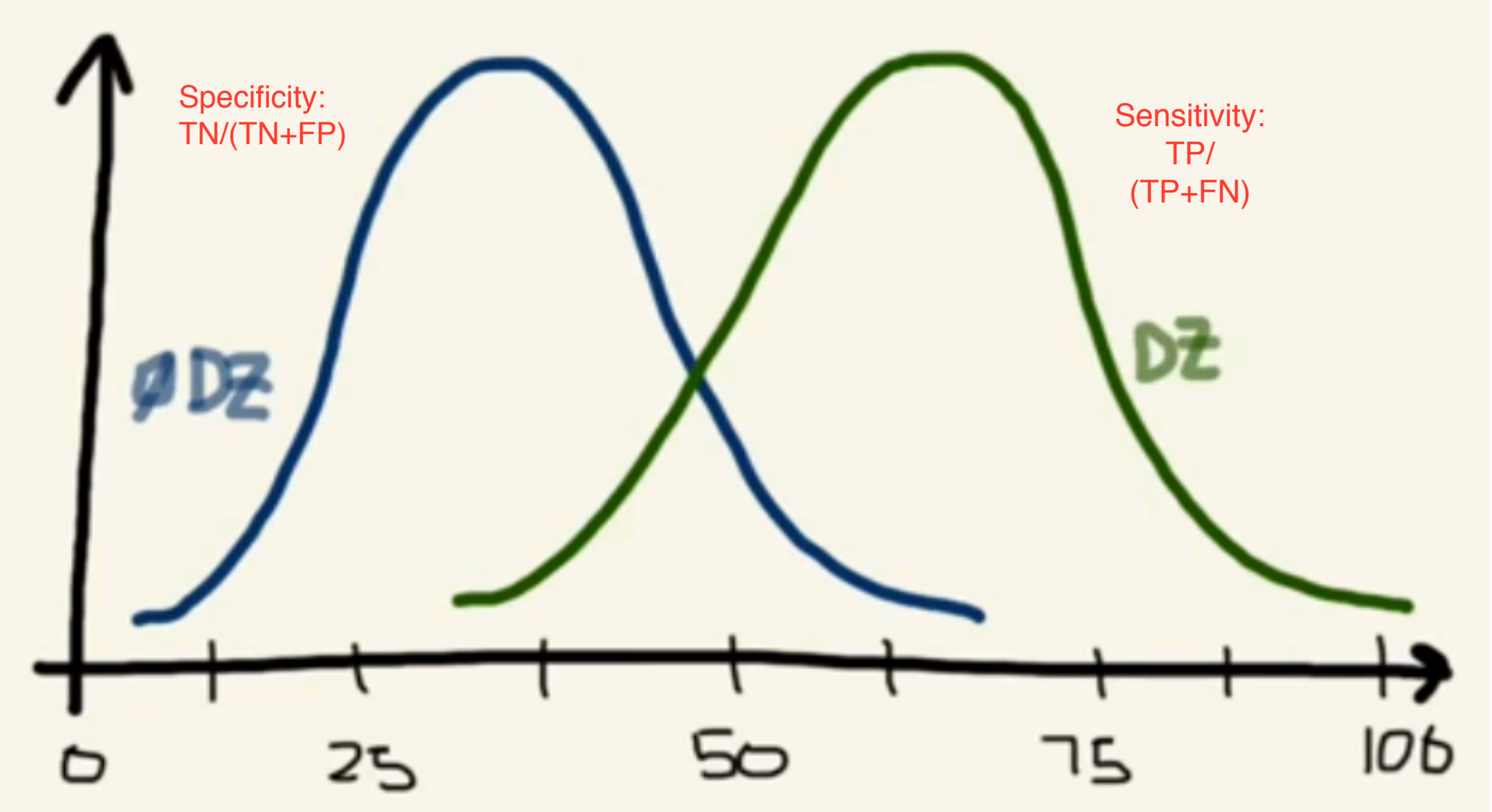
Probability density function for ‘true’ and ‘false’ cases of logistic regression



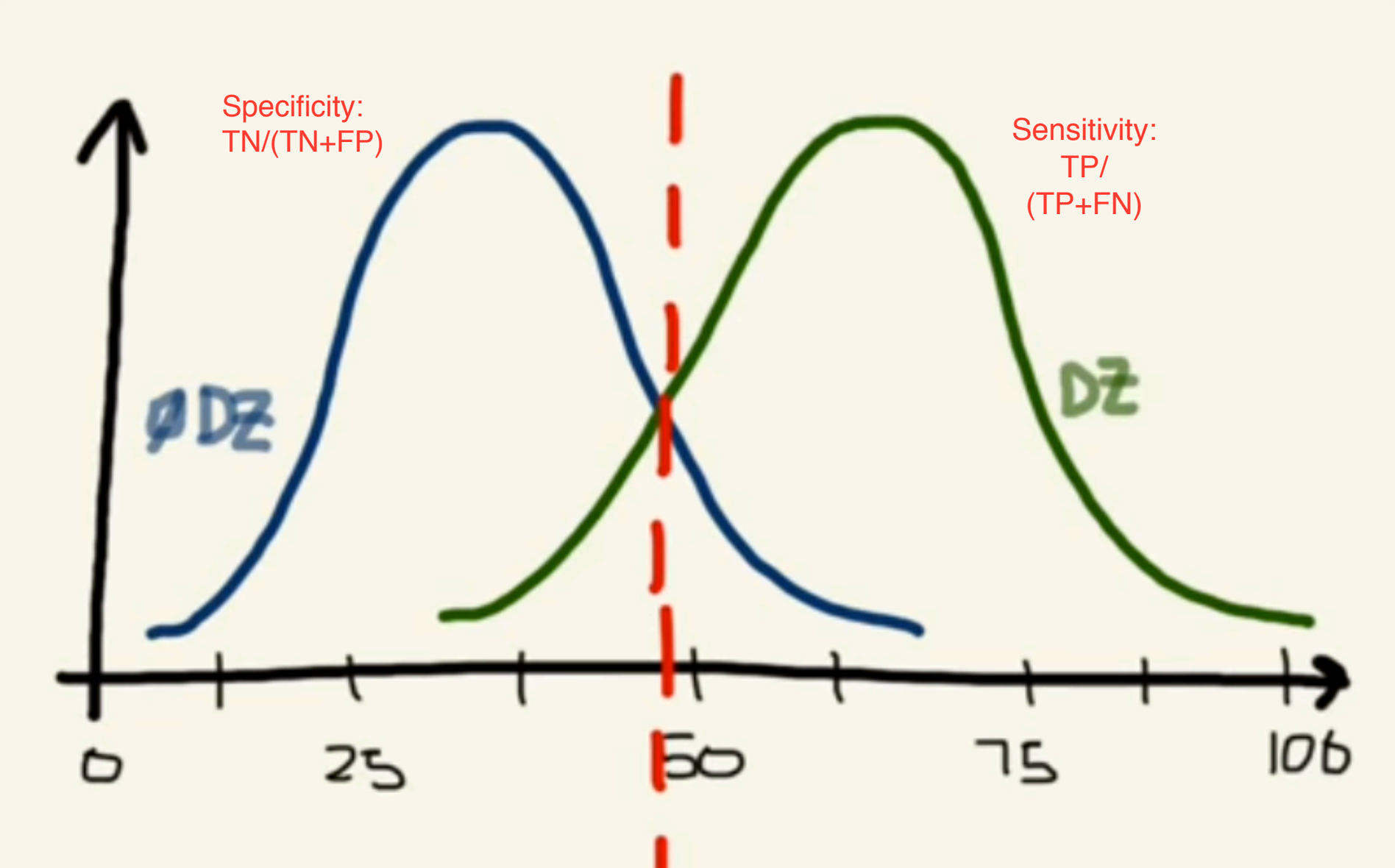
SENSITIVITY/SPECIFICITY TRADE OFF



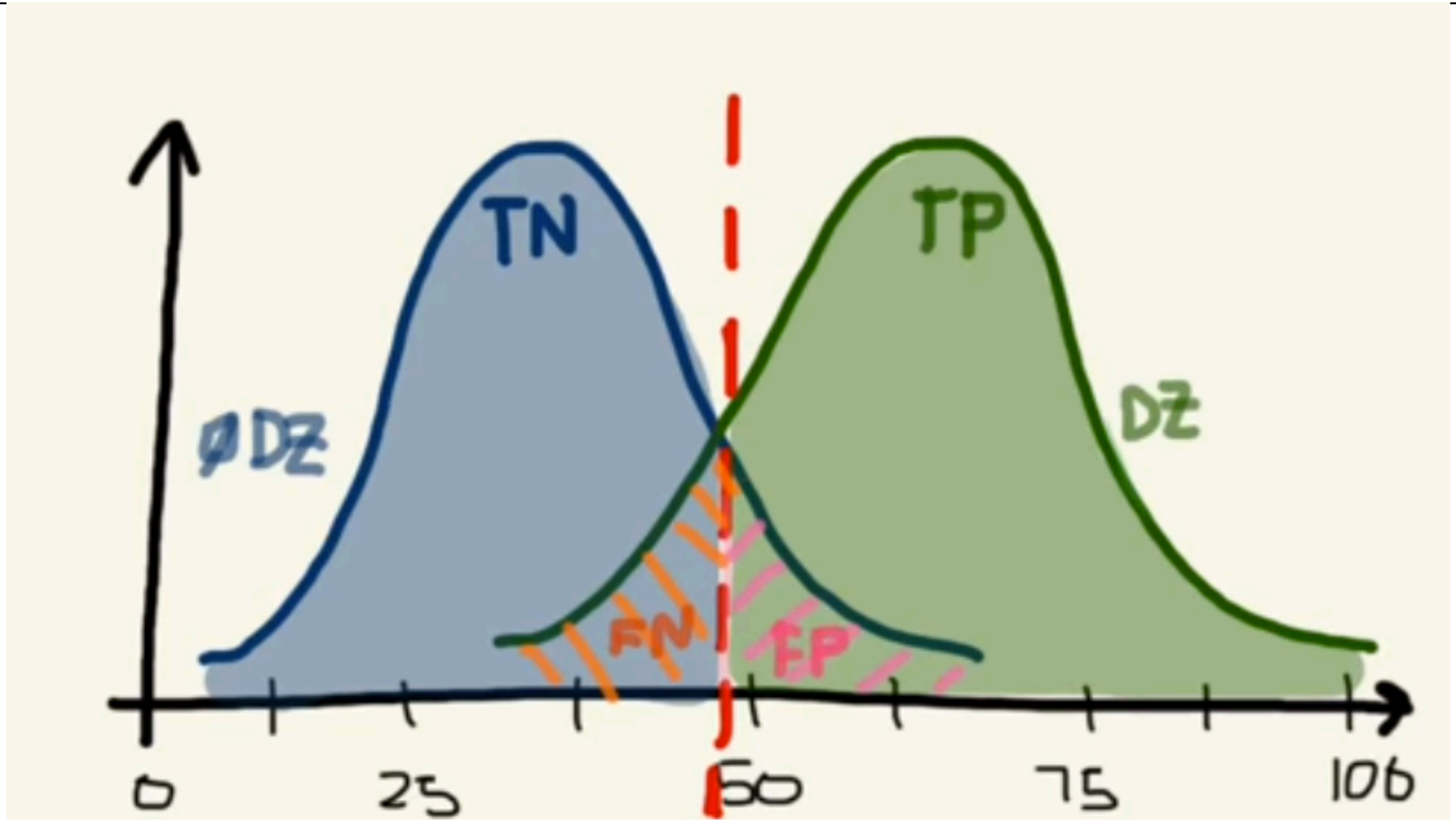
SENSITIVITY/SPECIFICITY TRADE OFF



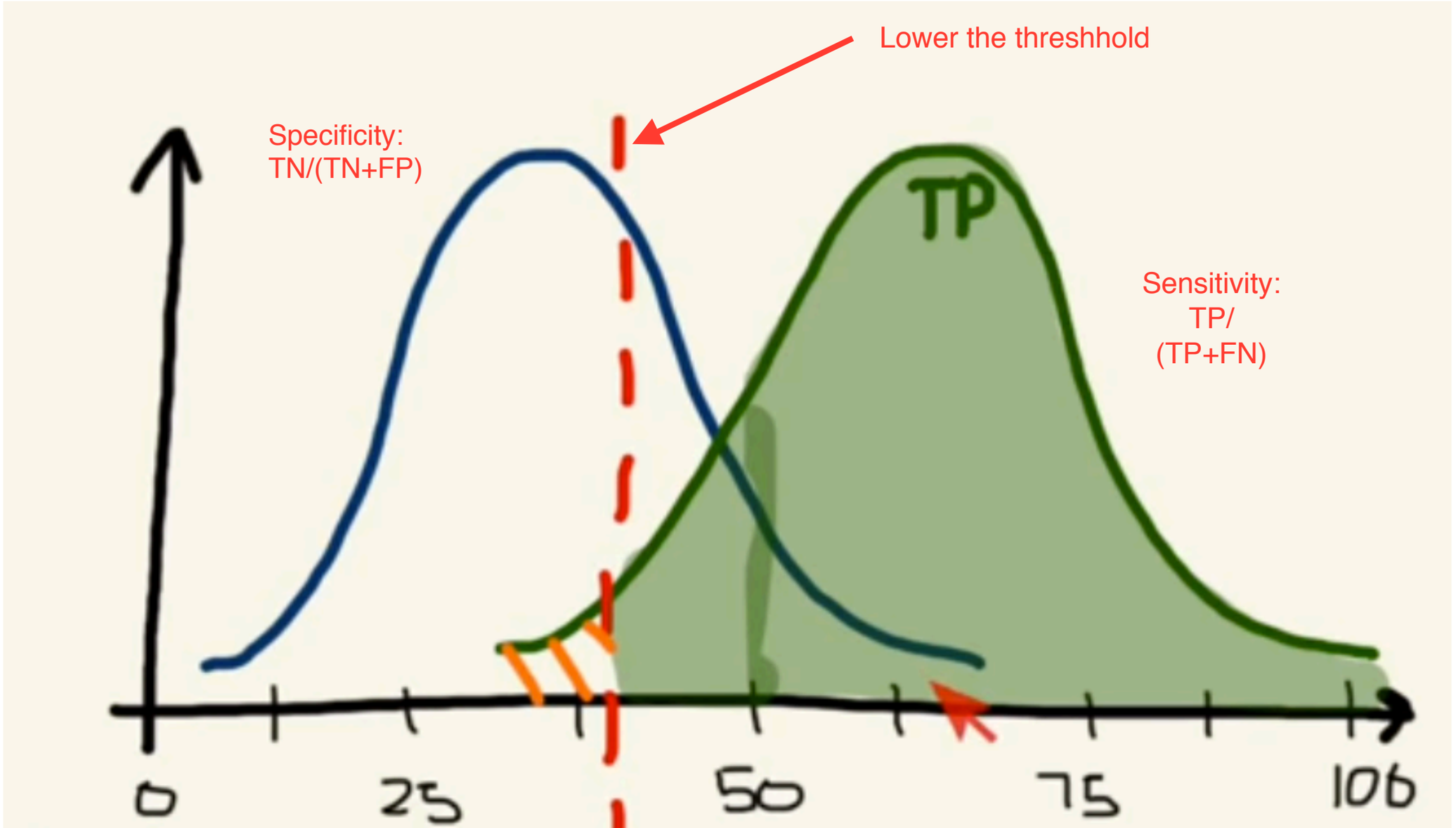
SENSITIVITY/SPECIFICITY TRADE OFF



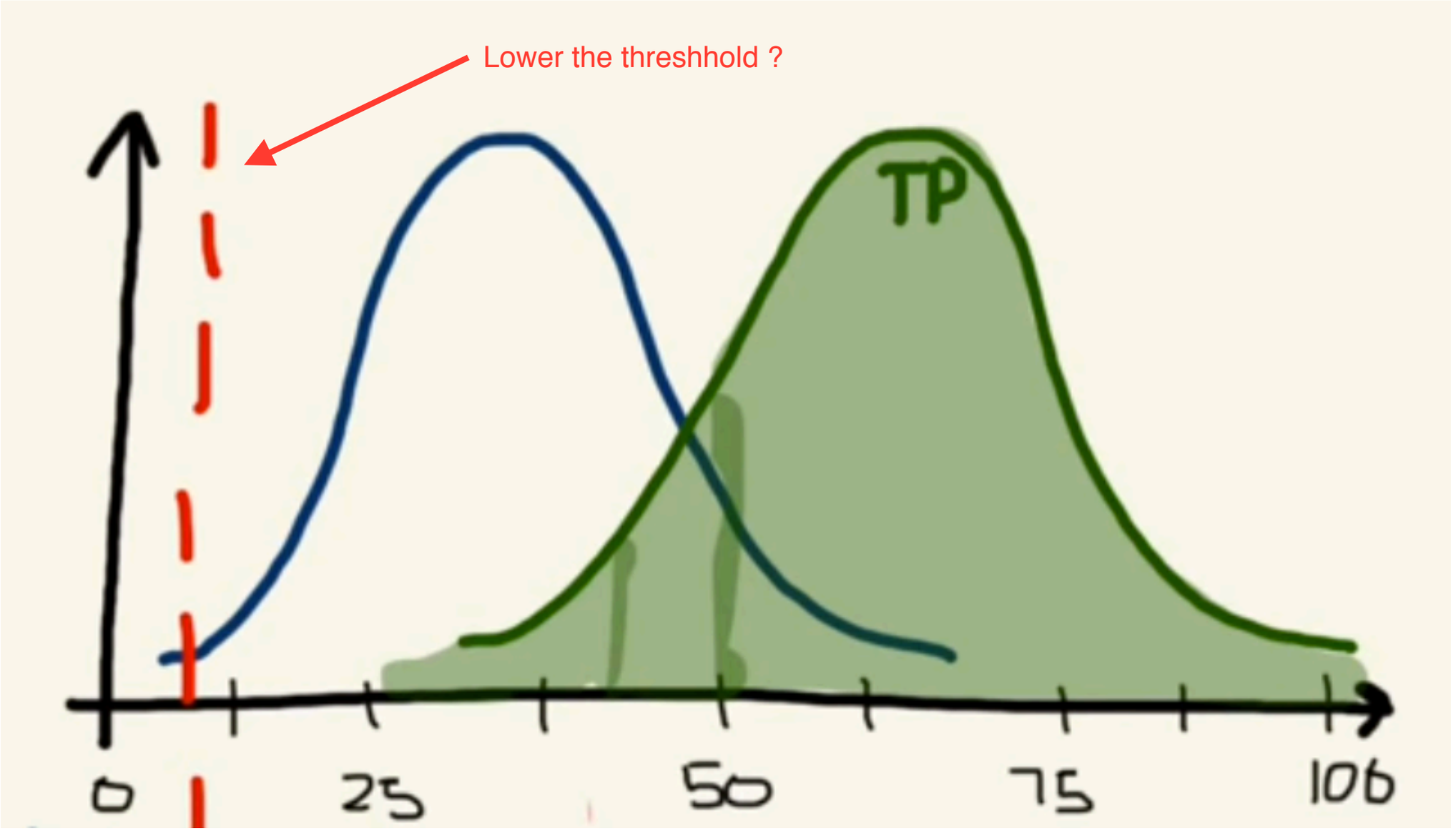
SENSITIVITY/SPECIFICITY TRADE OFF



SENSITIVITY/SPECIFICITY TRADE OFF



SENSITIVITY/SPECIFICITY TRADE OFF



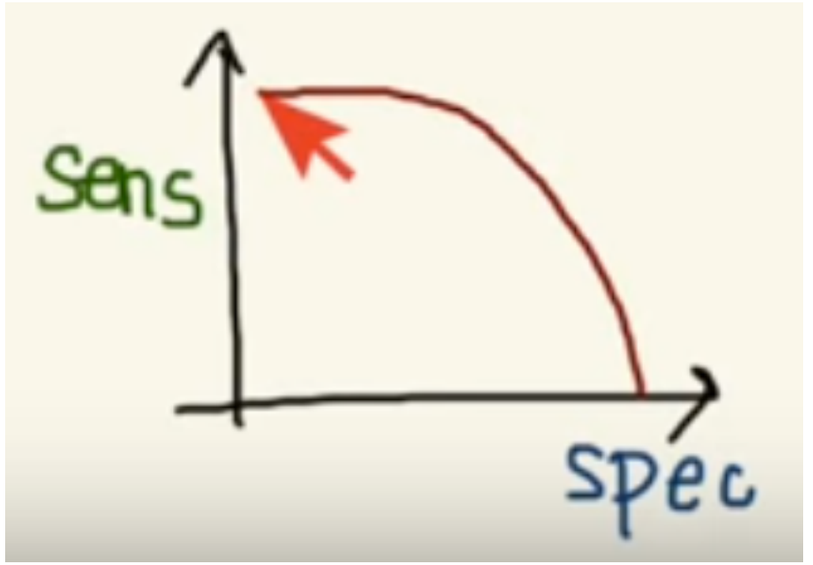
AUC AND ROC CURVES

AUC = curve created by Sensitivity versus Specificity

- 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

Sensitivity:
TP/
(TP+FN)

Specificity:
TN/(TN+FP)



Area under the curve shows how far you are from the baseline

baseline AUC = 0.5

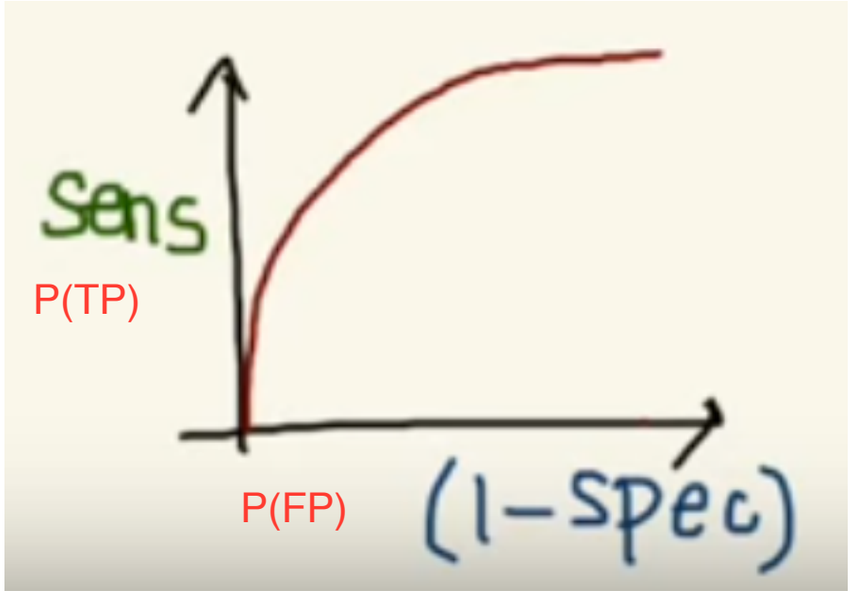
Perfect AUC = 1.0

AUC AND ROC CURVES

ROC = curve created by FP (x) versus TP (y)

- 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.

Threshold sets where you land on this curve



Shape of ROC determined by overlap in distributions (our zero and one in logistic regression)

AUC AND ROC CURVES

ROC AUC used in machine learning to compare models: how well can your model seperate examples and create a threshold?

Area under the curve shows how far you are from the baseline

baseline AUC = 0.5

Perfect AUC = 1.0