GENERAL ASSEMBLY

MODEL EVALUATION II & UNBALANCED CLASSES

Matt Brems, Data Science Immersive

RECAP

- What is classification?
- What are some examples of classification problems?
- What are examples of classifiers (algorithms for classification)?
- How might you evaluate the performance of a classifier?

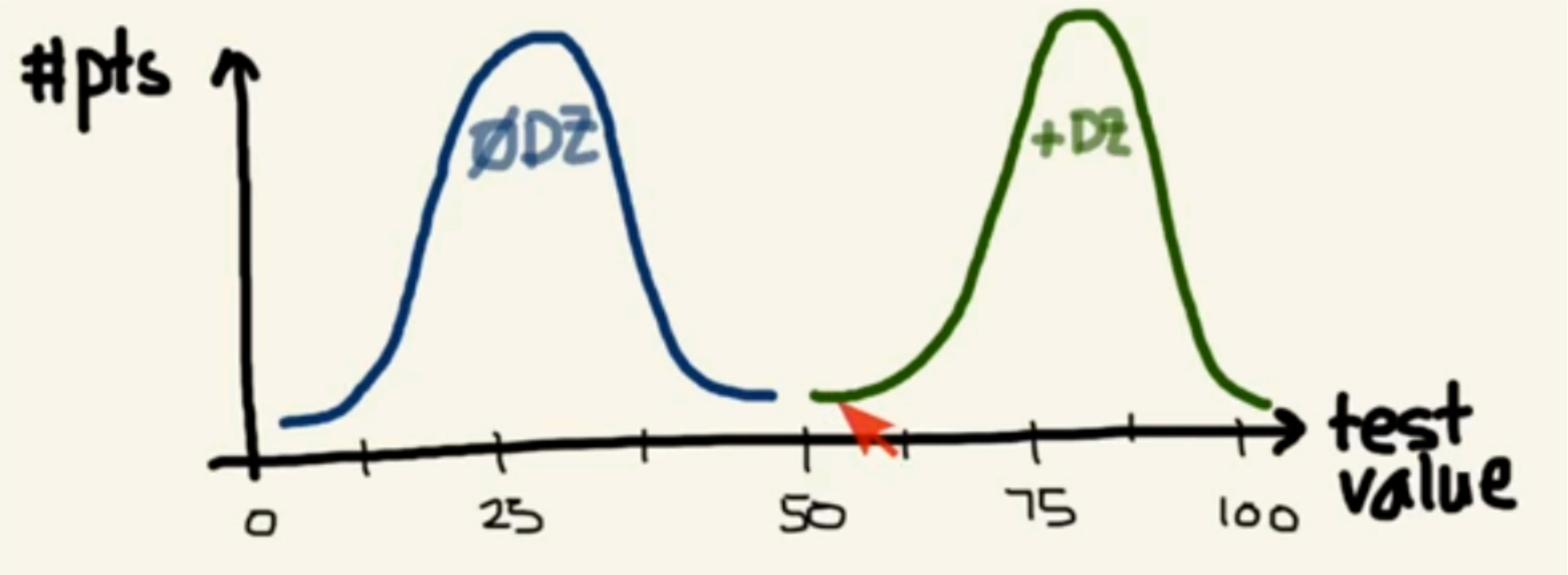
RECAP

- Suppose I want to predict fraudulent credit card transactions. (I'll say that fraudulent transactions are my success/"positive" event, despite the connotation.)

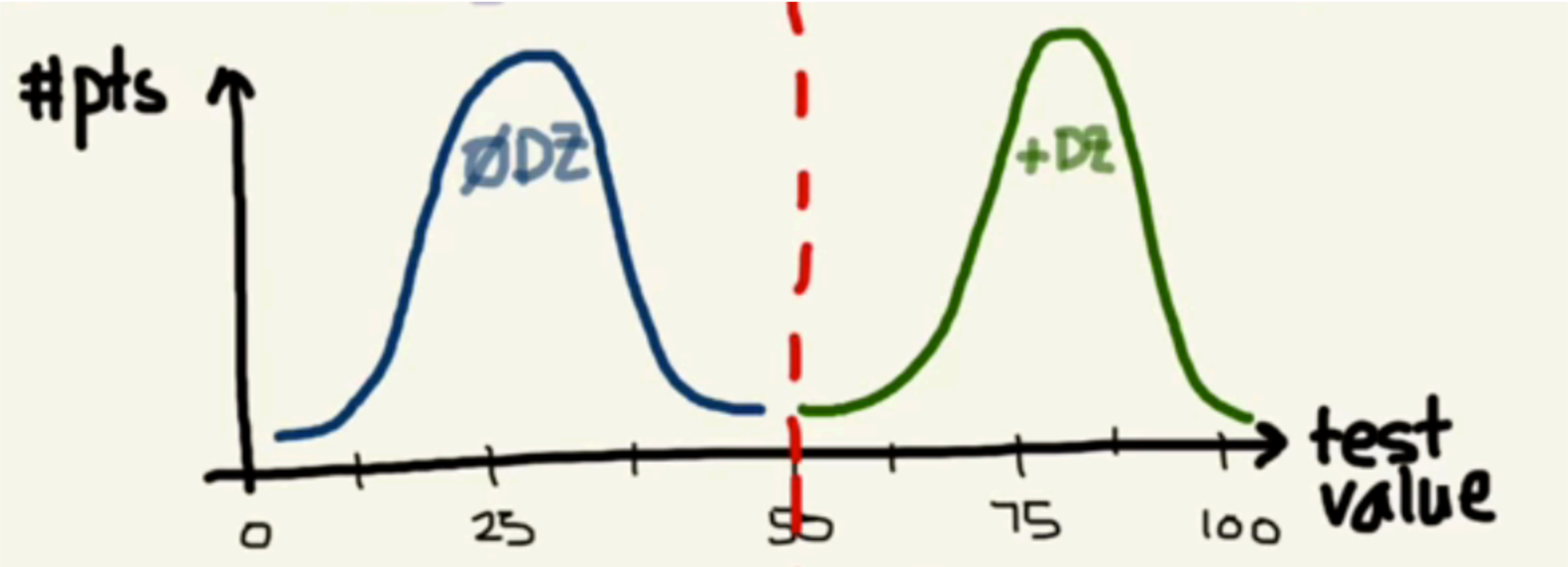
- I build a model and correctly predict 50 transactions as fraudulent. I correctly predict 900 transactions as not fraudulent. I incorrectly predict 40 transactions as fraudulent and incorrectly predict 10 transactions as not fraudulent.

- Let's build a confusion matrix.
- How many false positives do we have? How many false negatives?
- Find the sensitivity (also called recall), specificity, accuracy, and misclassification rate.

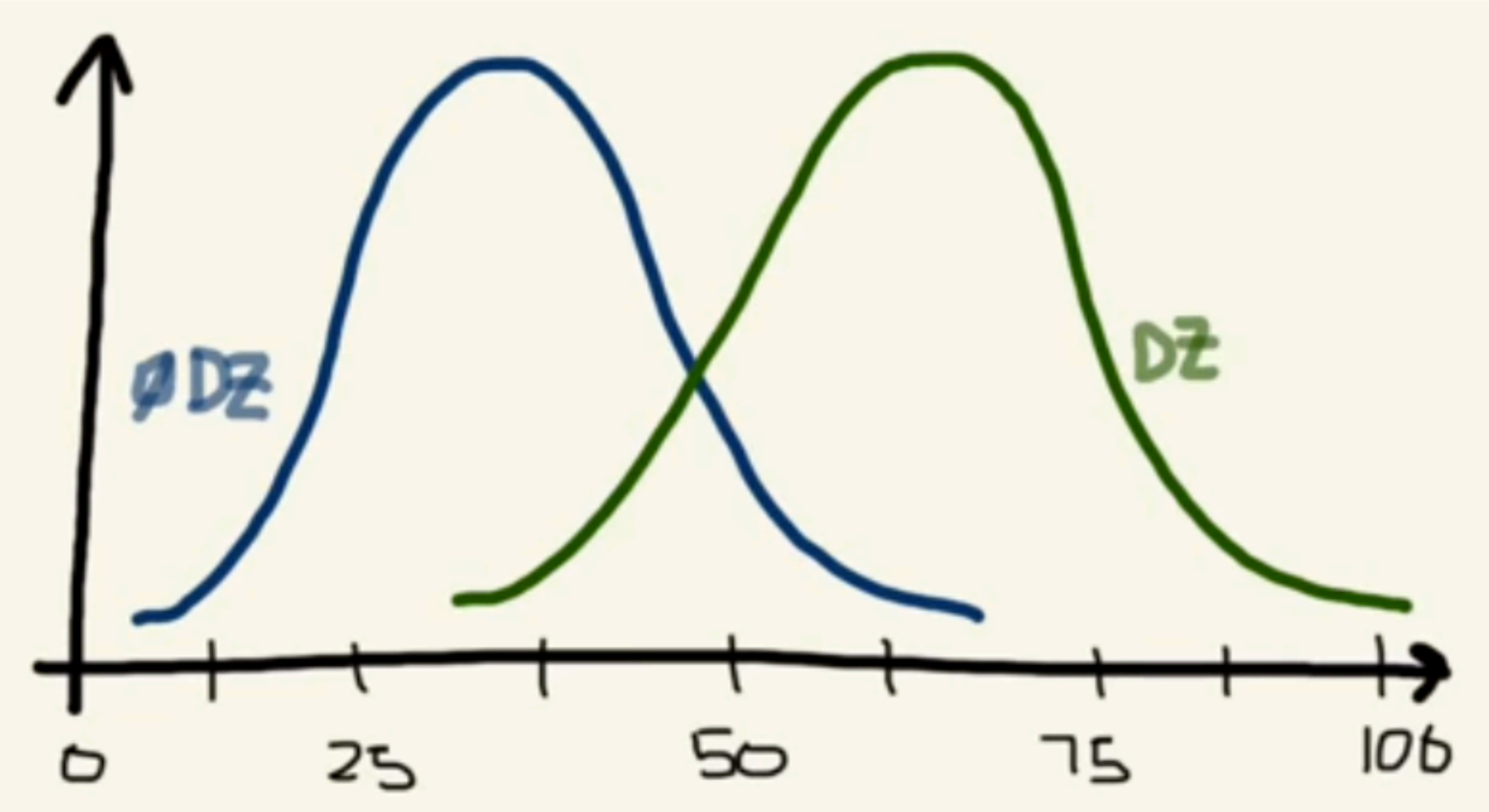
SENSITIVITY/SPECIFICITY TRADE OFF



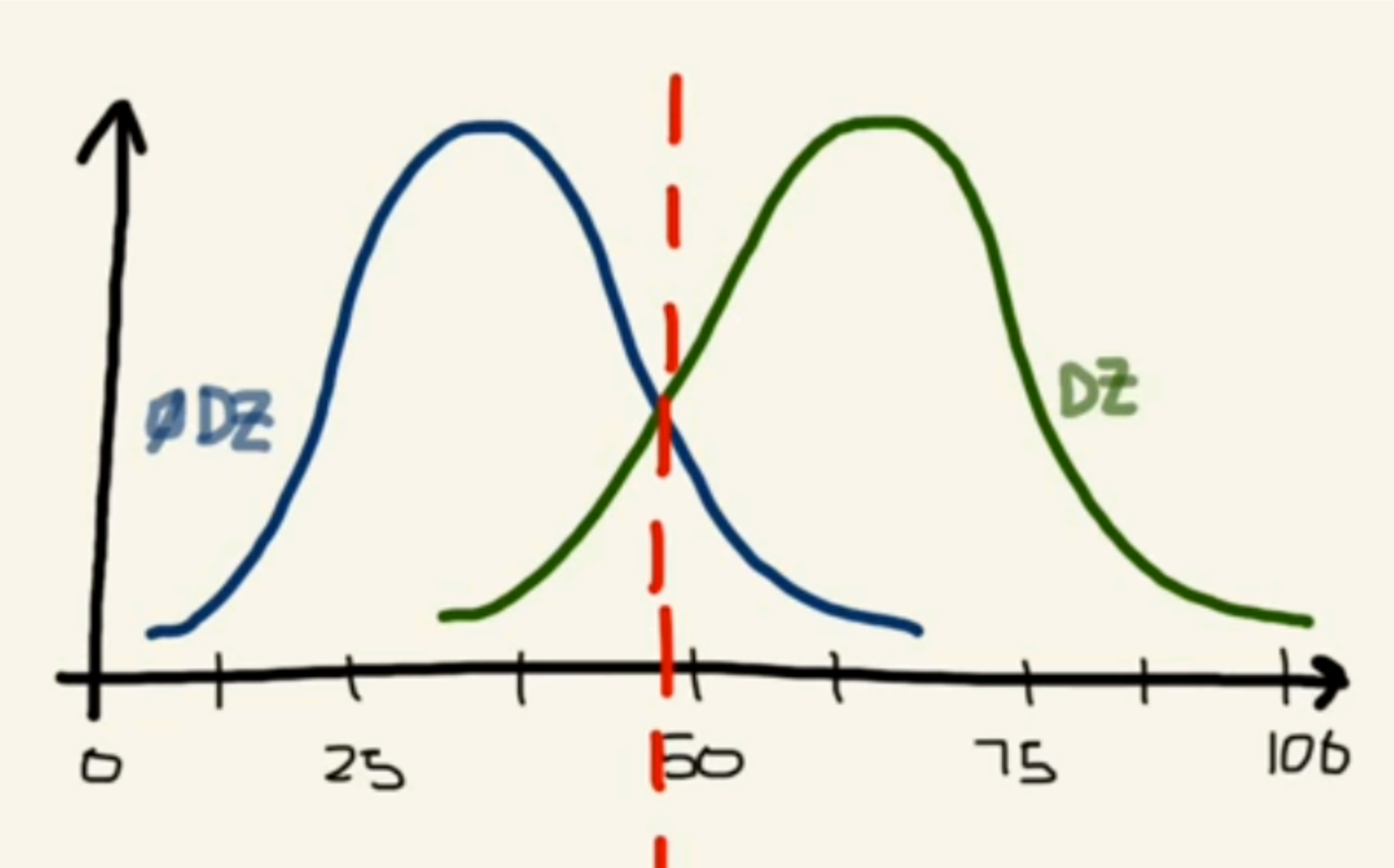
SENSITIVITY/SPECIFICITY TRADE OFF



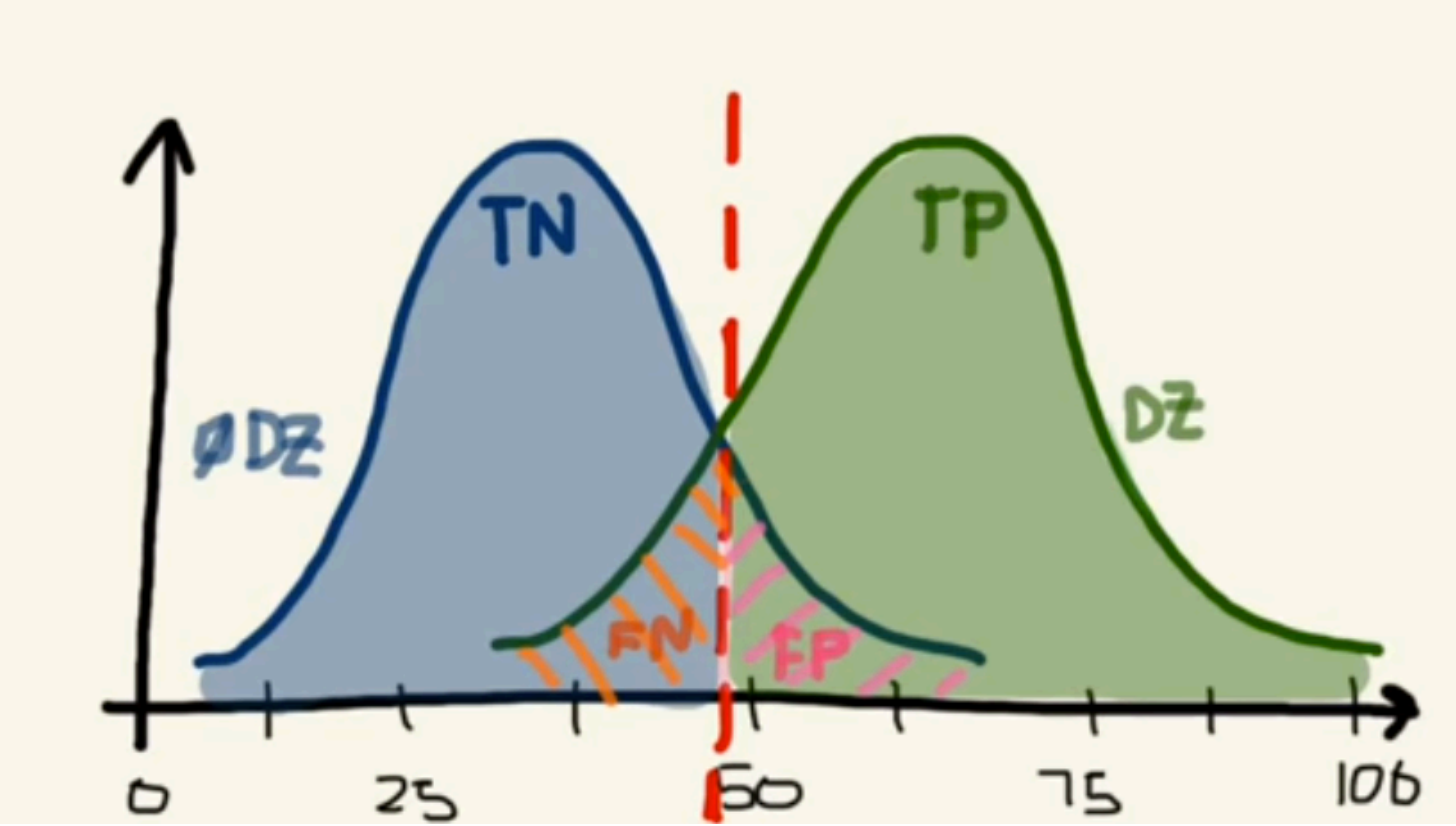
SENSITIVITY/SPECIFICITY TRADE OFF



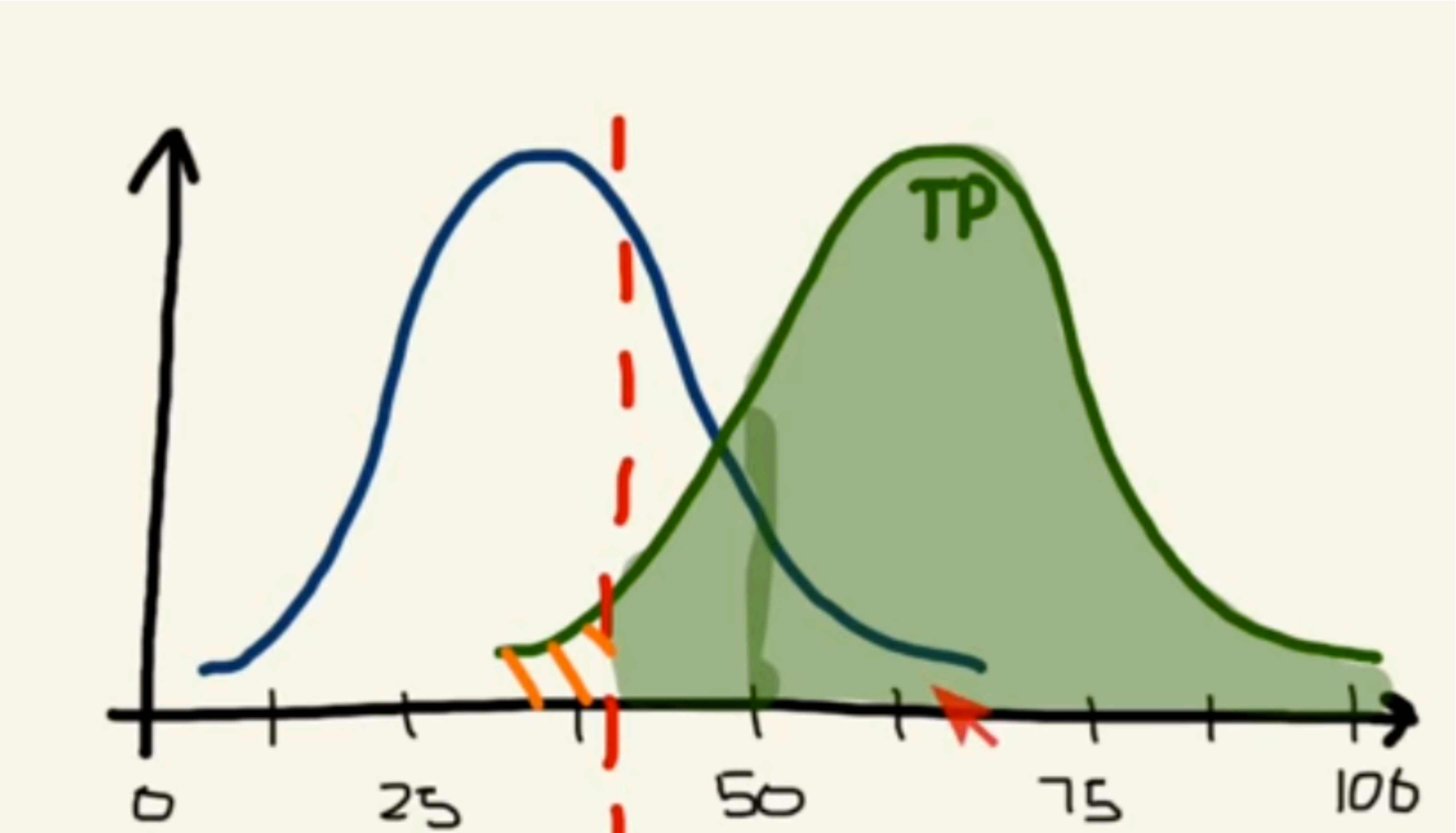
SENSITIVITY/SPECIFICITY TRADE OFF



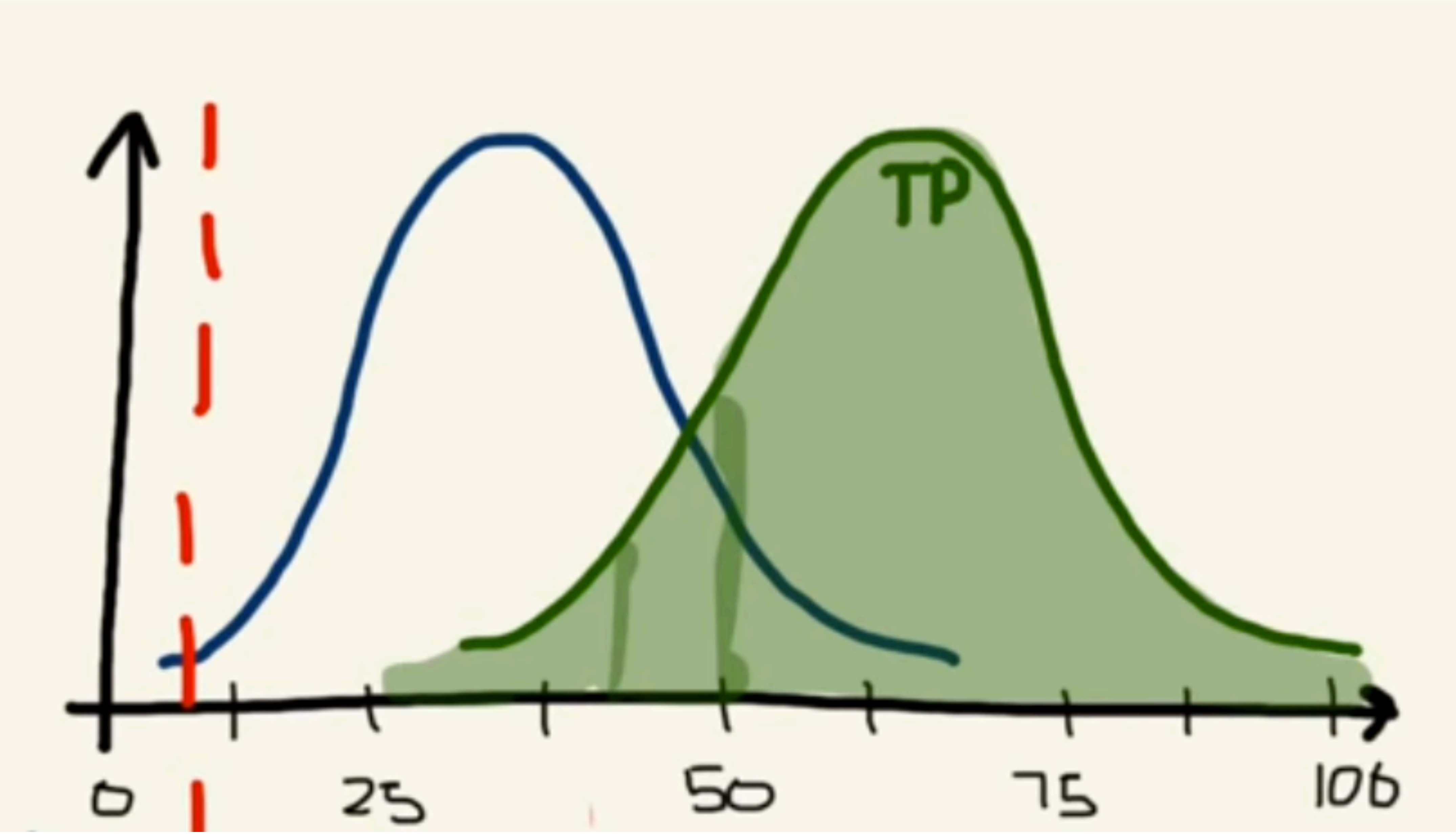
SENSITIVITY/SPECIFICITY TRADE OFF



SENSITIVITY/SPECIFICITY TRADE OFF



SENSITIVITY/SPECIFICITY TRADE OFF

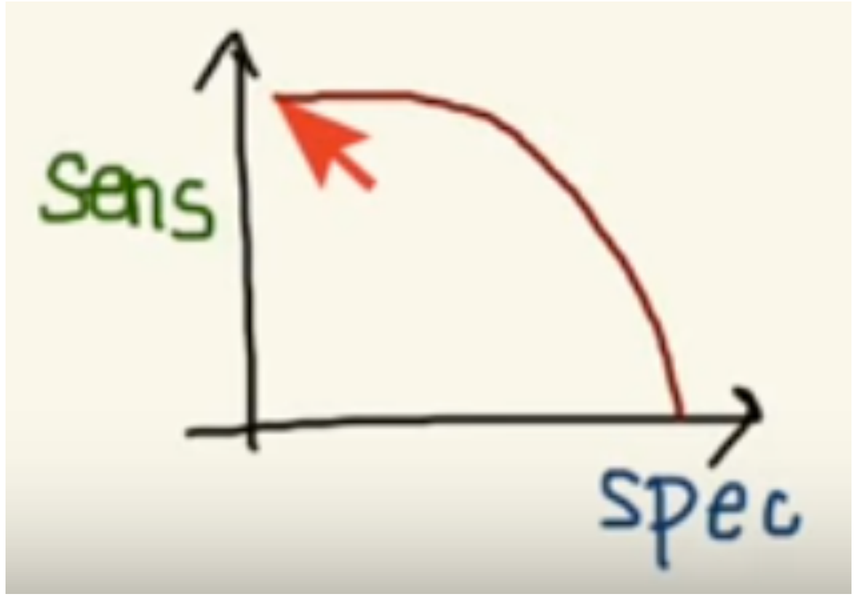


AUC AND ROC CURVES

- Sensitivity and specificity move in opposite directions, but we'd like to identify an "optimal" combination of the two.

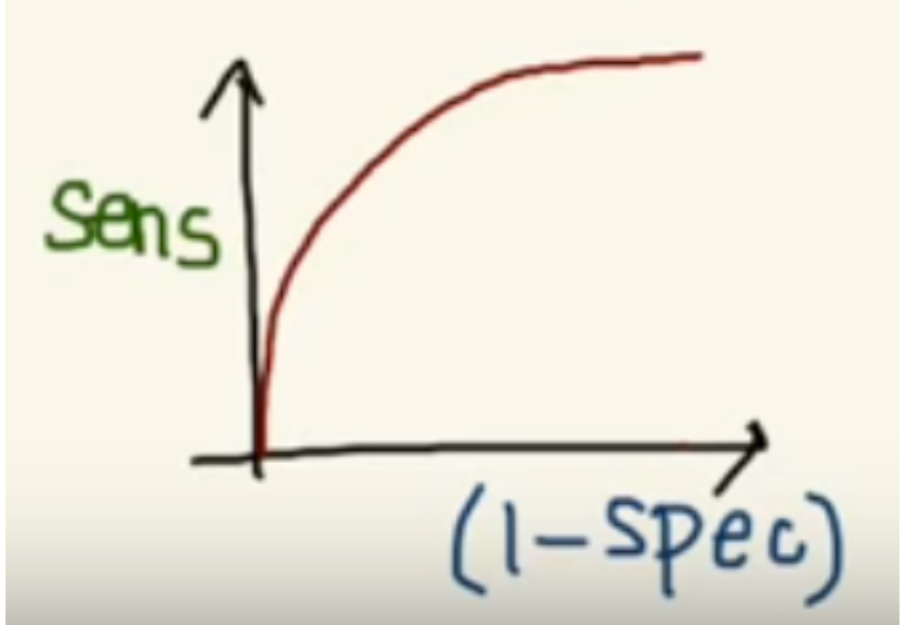
- We generate the ROC by plotting the sensitivity and specificity as we move our "classification threshold" from 0 to 1.

- We measure the strength of our classifier by taking the area under the curve. The acronym AUC-ROC refers to the "Area Under the Receiver Operating Characteristic curve."

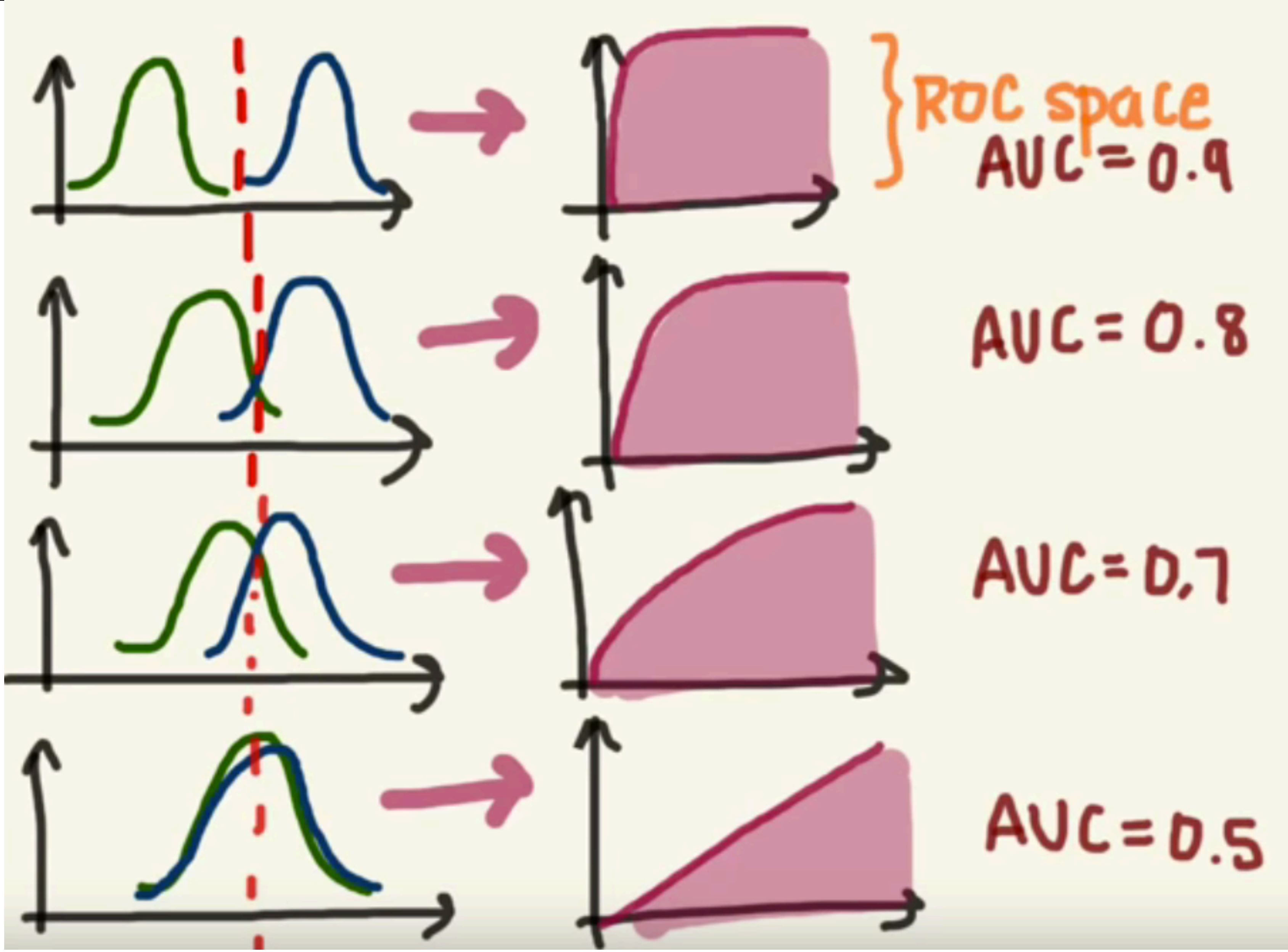


AUC AND ROC CURVES

- We plot Sensitivity vs. 1 – Specificity so that the two move in the same direction.
- Sensitivity: recall, true positive rate
- 1 – Specificity: false positive rate
- The ROC curve, therefore, compares the true positive rate against the false positive rate as we move our “threshold” from 0 to 1.



AUC AND ROC CURVES



RECEIVER OPERATING CHARACTERISTIC CURVE

- We generate **one** ROC curve for a classifier. The ROC curve is generated by varying our threshold from 0 to 1. (Therefore, changing our threshold for classification doesn’t affect our AUC ROC score!)

- We often use the ROC curve to identify an optimal threshold for our classifier by finding where we’re comfortable balancing sensitivity and 1 – specificity.

- We may also use the AUC-ROC score to evaluate the performance of our classifier.

BALANCED CLASSES

- In classification problems, methods generally work well when we have roughly equally-sized classes. (i.e. 50% in the positive class and 50% in the negative class for binary classification problems)

- However, there are many cases where this isn't true.

- One example of poor performance with unbalanced classes: logistic regression.
- If $Y = 1$ is a rare event, logistic regression will underestimate $P(Y = 1)$ and thus overestimate $P(Y = 0)$.

METHODS

- Bias correction.
- Oversampling/undersampling.
- Weighting observations. (i.e. weighted least squares)
- Stratified cross-validation.
- Changing threshold for classification.
- Purposefully optimizing evaluation metrics.

BIAS CORRECTION

- Because logistic regression will naturally underestimate the proportion of "successes" when successes are rare, we say that $E\left[\hat{P}(Y=1)\right] < P(Y=1)$.

- Gary King proposed methods for correcting for this bias in his paper (https://gking.harvard.edu/files/gking/files/0s.pdf) that include ways to counter this bias.

- While this is both theoretically rigorous and empirically shown to provide good results, data scientists often prefer "easier" methods of addressing bias.

OVERSAMPLING / UNDERSAMPLING

- In unbalanced classes, one class will be (by definition) larger than the other.
- We might bootstrap the minority class so that we artificially balance the classes when fitting our model.
- We might randomly sample the majority class so that we artificially balance the classes when fitting our model.
- **NOTE: WE ALWAYS EVALUATE OUR MODEL ON THE REAL DATA.**

WEIGHTING OBSERVATIONS

- We might prefer to “weight” our observations so that the minority and majority classes are more equally represented, then model with the weighted observations.

- This can run into issues with increasing variance, but also isn’t “generating” or “dropping” data at random.

- The choice of weight is usually arbitrary – so be sure you can defend why you made the decision that you did!

- **NOTE: WE ALWAYS EVALUATE OUR MODEL ON THE REAL DATA.**

STRATIFIED CROSS-VALIDATION

- If we use k-fold cross-validation entirely randomly, we may run into issues where some of our folds have no observations from the minority class.

- By stratifying on our output variable with unbalanced classes during cross-validation, we protect ourselves from this situation and ensure that our estimate of our model performance has lower variance.

CHANGING CLASSIFICATION THRESHOLD

- As we classify observations into classes, we usually defer to a 50% threshold when separating observations.

- However, by adjusting our classification threshold, we might find a better fit for our particular use-case.

OPTIMIZING SPECIFIC EVALUATION METRICS

- We have lots of evaluation metrics available! Look back on the confusion matrix from our first set of model evaluation slides.
- In cases where false positives incur a different cost than false negatives, we may want our model to more rigorously classify in a certain direction.
- We may choose to optimize for certain evaluation metrics because we'd like to maximize or minimize some particular metric or measure.
- This often is accompanied by adjusting the classification threshold.

METHODS

- Bias correction.
- Oversampling/undersampling.
- Weighting observations. (i.e. weighted least squares)
- Stratified cross-validation.
- Changing threshold for classification.
- Purposefully optimizing evaluation metrics.