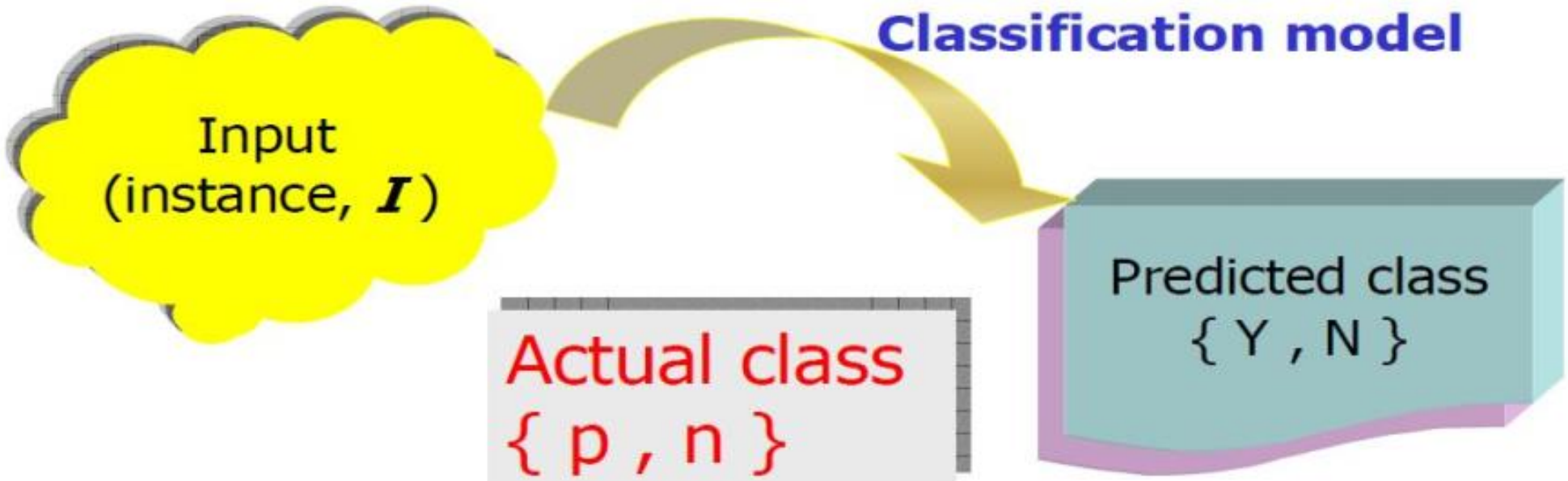
Model Evaluation

Classier Performance

- Problem: two classes classification



PS: actual class {p: positive class, n: negative class}

Confusion Matrix

- Given a classifier and an instance:

Classifier / Predicted class	TRUE CLASS	
	p (positive)	n (negative)
Y	True Positives	False Positives
N	False Negatives	True Negatives
Total	P	N

P = True Positives + False Negatives

Measuring Classifier Performance

→ If the instance is positive and it is classified as positive, it is counted as a **true positive**.
→ if the instance is positive and it is classified as negative, it is counted as a **false negative.**
→ If the instance is negative and it is classified as negative, it is counted as a **true negative.**
→ If the instance is negative and it is classified as positive, it is counted as a **false positive.**

Example- Confusion Matrix

Cancer Blood Test

True Positive

- Actually have it and test also says so

True Negative

- Don’t have it and test also says so

False Positive

- Have it but test doesn’t say so

False Negative

- Don’t have it but test says you have it.

Performance index

TP	FP
FN	TN

$$TPR = \frac{TP}{P} = Recall,\ FPR = \frac{FP}{N}$$

$$Precision = \frac{TP}{TP + FP},\ Accuracy = \frac{TP + TN}{P + N}$$

$$Sensitivity = Recall,\ Specificity = 1 - FPR$$

Performance Metrics

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N} \qquad \text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

$$\text{sensitivity} = \text{recall}$$

$$\text{specificity} = \frac{\text{True negatives}}{\text{False positives} + \text{True negatives}} = 1 - \textit{fp rate}$$

→ **tp (true positive) rate / Recall / Hit rate / Sensitivity:** The proportion of positives that are correctly identified.
→ **fp (false positive) rate:** The proportions of negatives that are incorrectly identified.
→ **Precision / Positive predicted value**: The proportion of positively identified that are correct.
→ **F-measure:** A measure that combines precision and recall. It is the harmonic mean of precision and recall.
→ **Specificity:** The proportion of negatives that are correctly identified.

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Test → 1000

Yes – 900

No – 100

Bad Model → positive

Actual \ Pred	1	0
1	TP	FN
0	FP	TN

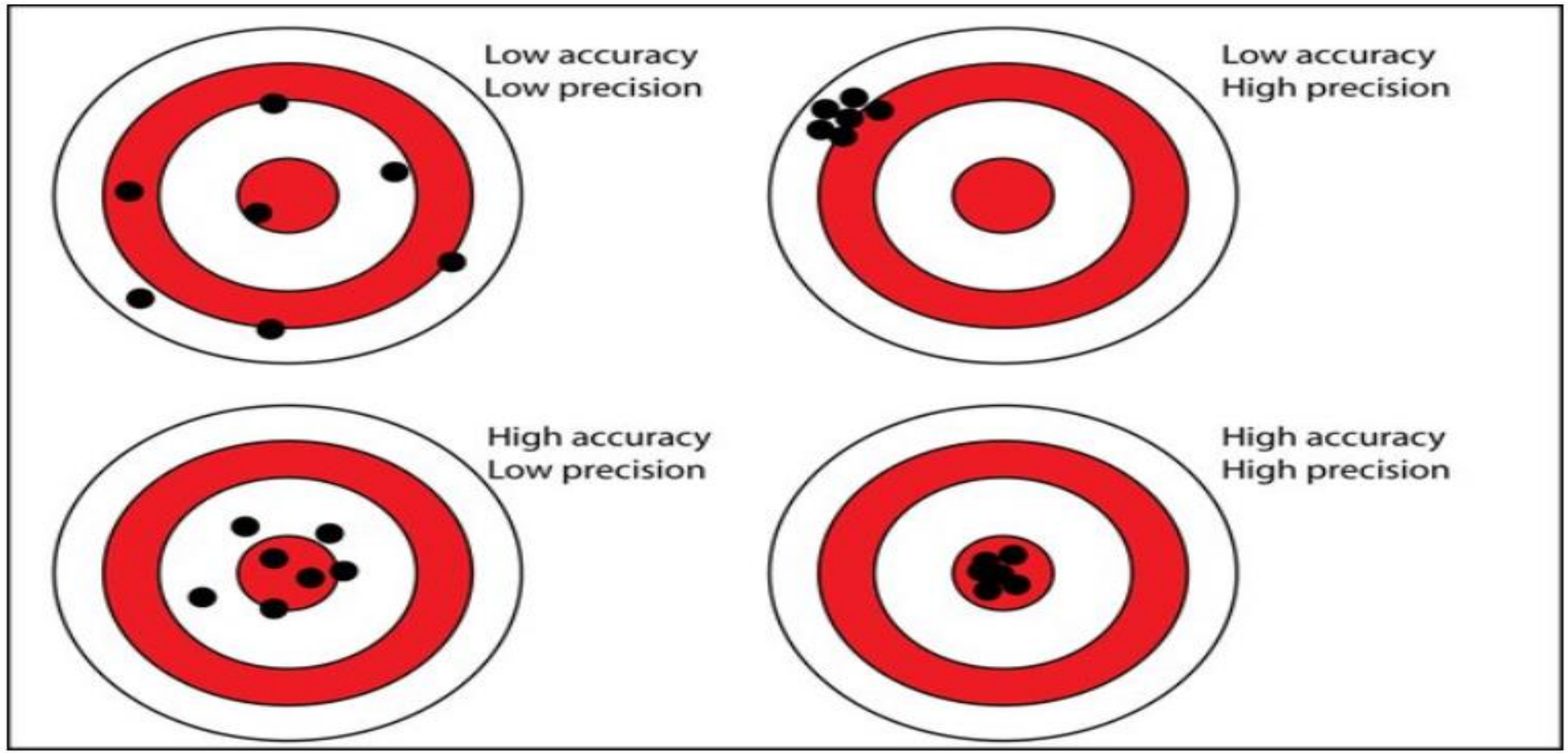
acc \ Pred	1	0
900	900	0
100	100	0

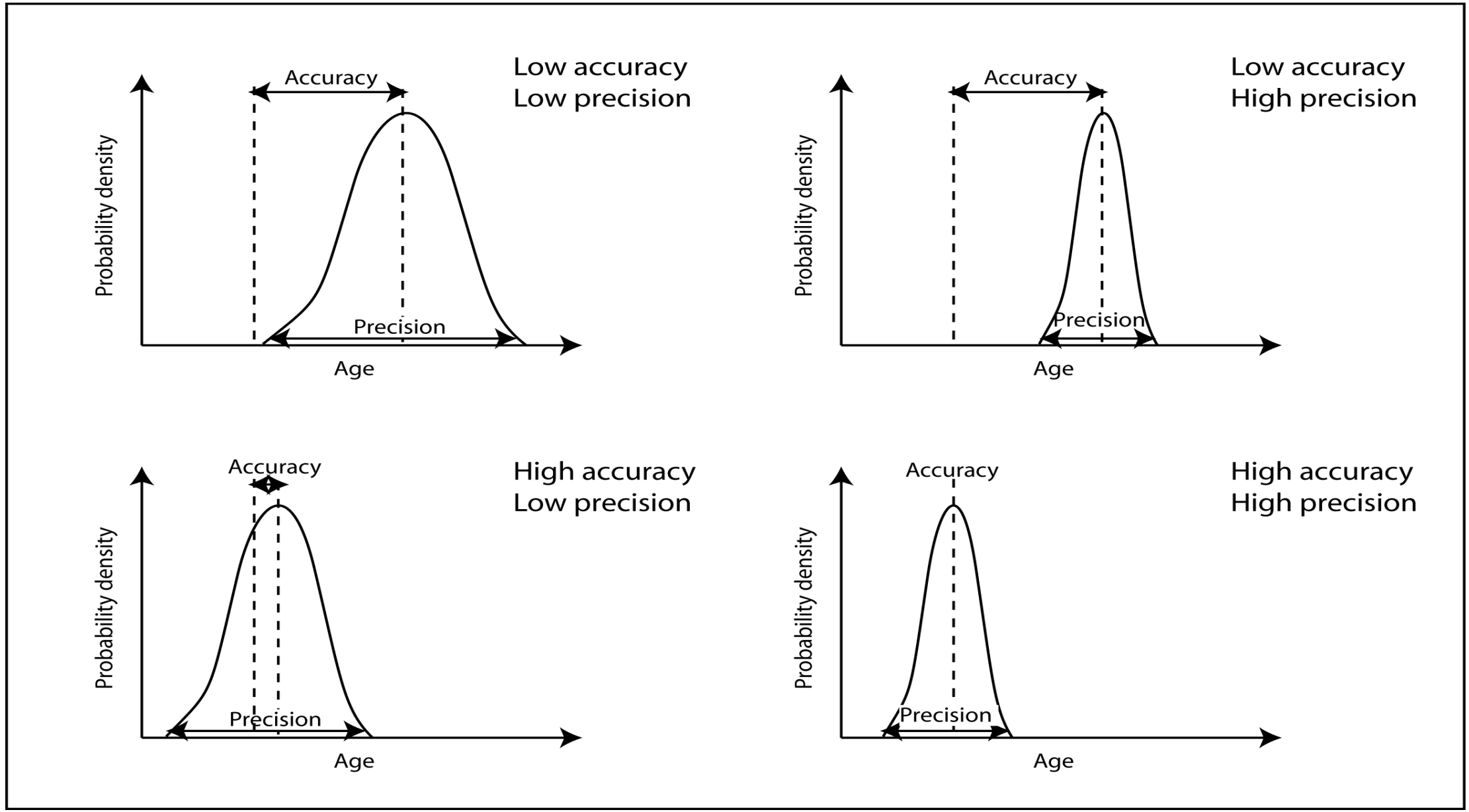
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$= \frac{TP + TN}{P + N}$$

$$= \frac{900 + 0}{900 + 100 + 0 + 0}$$

$$= \frac{90}{100} = 90\%$$


Low accuracy
Low precision
Low accuracy
High precision
High accuracy
Low precision
High accuracy
High precision


Low accuracy
Low precision
Accuracy
Probability density
Precision
Age
Low accuracy
High precision
Accuracy
Probability density
Precision
Age
High accuracy
Low precision
Accuracy
Probability density
Precision
Age
High accuracy
High precision
Accuracy
Probability density
Precision
Age

Truth										
Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog

Truth										
Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog


Truth
Prediction
Dog
Dog
Dog
No Dog
Dog
No Dog
Dog
No Dog
Dog
Dog


Truth
Prediction
Dog
Dog
Dog
No Dog
Dog
No Dog
Dog
No Dog
Dog
Dog

Truth

Prediction	Dog	Dog	Dog	No Dog	Dog	No Dog	Dog	No Dog	Dog	Dog
	✓	✗	✓		✓		✗		✓	✗

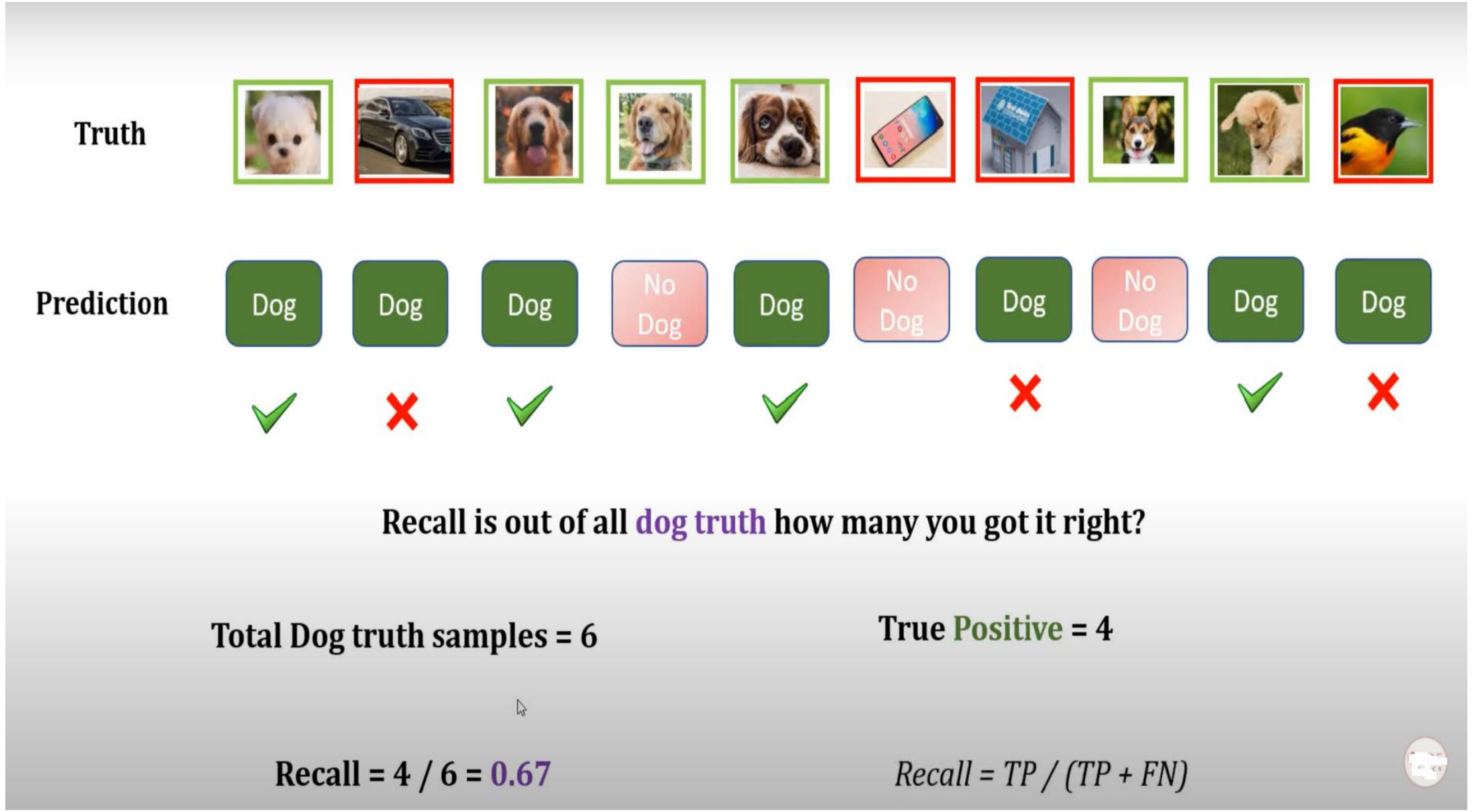
True Positive = 4

False Positive = 3

Precision is out of all dog predictions how many you got it right?

Precision = 4 / 7 = 0.57

Precision = TP / (TP + FP)


Truth
Prediction
Dog
Dog
Dog
No Dog
Dog
No Dog
Dog
No Dog
Dog
Dog
Recall is out of all dog truth how many you got it right?
Total Dog truth samples = 6
True Positive = 4
Recall = 4 / 6 = 0.67
Recall = TP / (TP + FN)

INTRODUCTION

- It is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.
- It is a table of two dimensions; Actual Value and Predicted Value.
- Confusion matrix, also known as an error matrix.

PREDICTED

ACTUAL	185	NO	YES	
	NO	55 [TN]	15 [FP]	70
	Yes	10 [FN]	105 [TP]	115
		65	120	

Matrix Terms

- **True Positives (TP)** − It is the case when both actual class & predicted class of data point is 1.
- **True Negatives (TN)** − It is the case when both actual class & predicted class of data point is 0.
- **False Positives (FP)** − It is the case when actual class of data point is 0 & predicted class of data point is 1.
- **False Negatives (FN)** − It is the case when actual class of data point is 1 & predicted class of data point is 0.

Measure Terms

- **Accuracy:**
  - It is how close a measured value to the actual (True) value.

Accuracy = (TP + TN) /Total

= (55+105)/185

= 0.86

Cont...

- **Precision:**
  - It is how close the measured values are to each other.

Precision = TP / Predicted Yes

= 105 / 120

= 0.87

Recall

- **Recall:**
  - It is the ratio of all correctly predicted positive predictions

    Recall = TP / Actual Yes

    = 105 / 115

    = 0.91

Cont…

- **Error Rate:**
  - It is calculated as the number of all incorrect predictions divided by the total number of the datasets.
  - The best error rate is 0.0
  - The worst error rate is 1.0.

Error Rate = 1 - Accuracy = (FN + FP) / Total

= 1 - 0.86 = (15 +10) /185

= 0.14

Analysis with Performance Measurement Metrics

- Based on the various performance metrics, we can characterize a classifier.
- We do it in terms of TPR, FPR, Precision and Recall and Accuracy
- **Case 1: Perfect Classifier**

When every instance is correctly classified, it is called the perfect classifier. In this case, $TP = P$, $TN = N$ and CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \frac{P}{P} = 1$$

$$F_1\ Score = \frac{2 \times 1}{1+1} = 1$$

$$Accuracy = \frac{P+N}{P+N} = 1$$

		Predicted Class	
		+	-
Actual class	+	P	0
	-	0	N

Analysis with Performance Measurement Metrics

- **Case 2: Worst Classifier**

  When every instance is wrongly classified, it is called the worst classifier. In this case, $TP = 0$, $TN = 0$ and the CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{0}{N} = 0$$

F_1 *Score* = Not applicable as *Recall* + *Precision* = 0

$$\text{Accuracy} = \frac{0}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	0	P
	-	N	0

Analysis with Performance Measurement Metrics

- **Case 3: Ultra-Liberal Classifier**

The classifier always predicts the + class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{P}{P+N}$$

$$F_1\ Score = \frac{2P}{2P+N}$$

$$\text{Accuracy} = \frac{P}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	P	0
	-	N	0

Analysis with Performance Measurement Metrics

- **Case 4: Ultra-Conservative Classifier**

  This classifier always predicts the - class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \text{Not applicable}$$

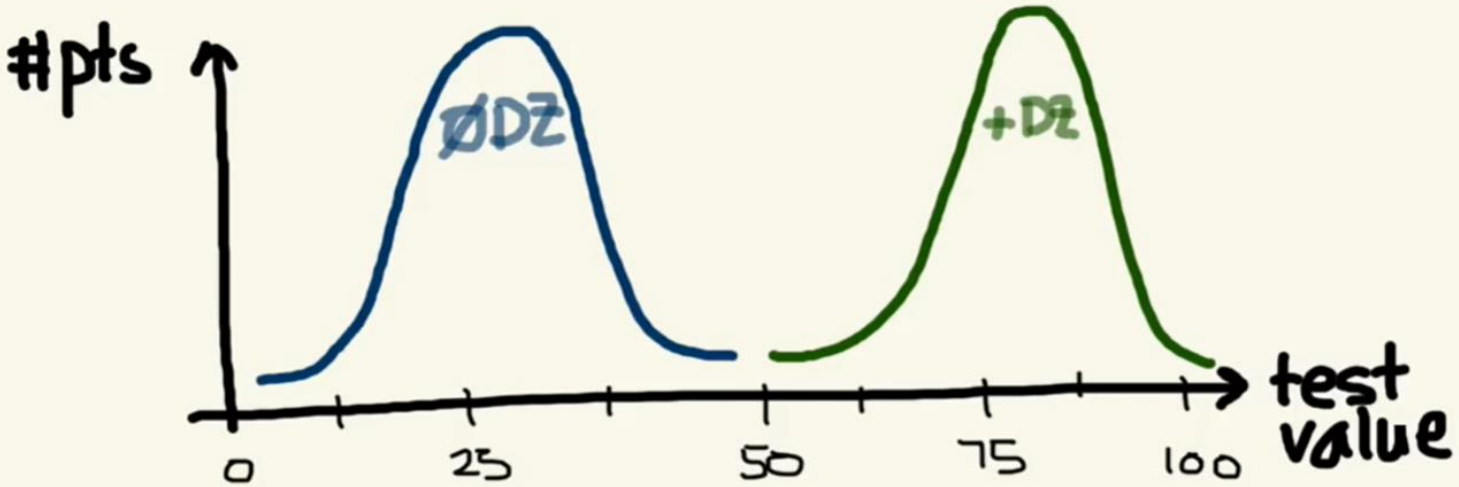
(as $TP + FP = 0$)

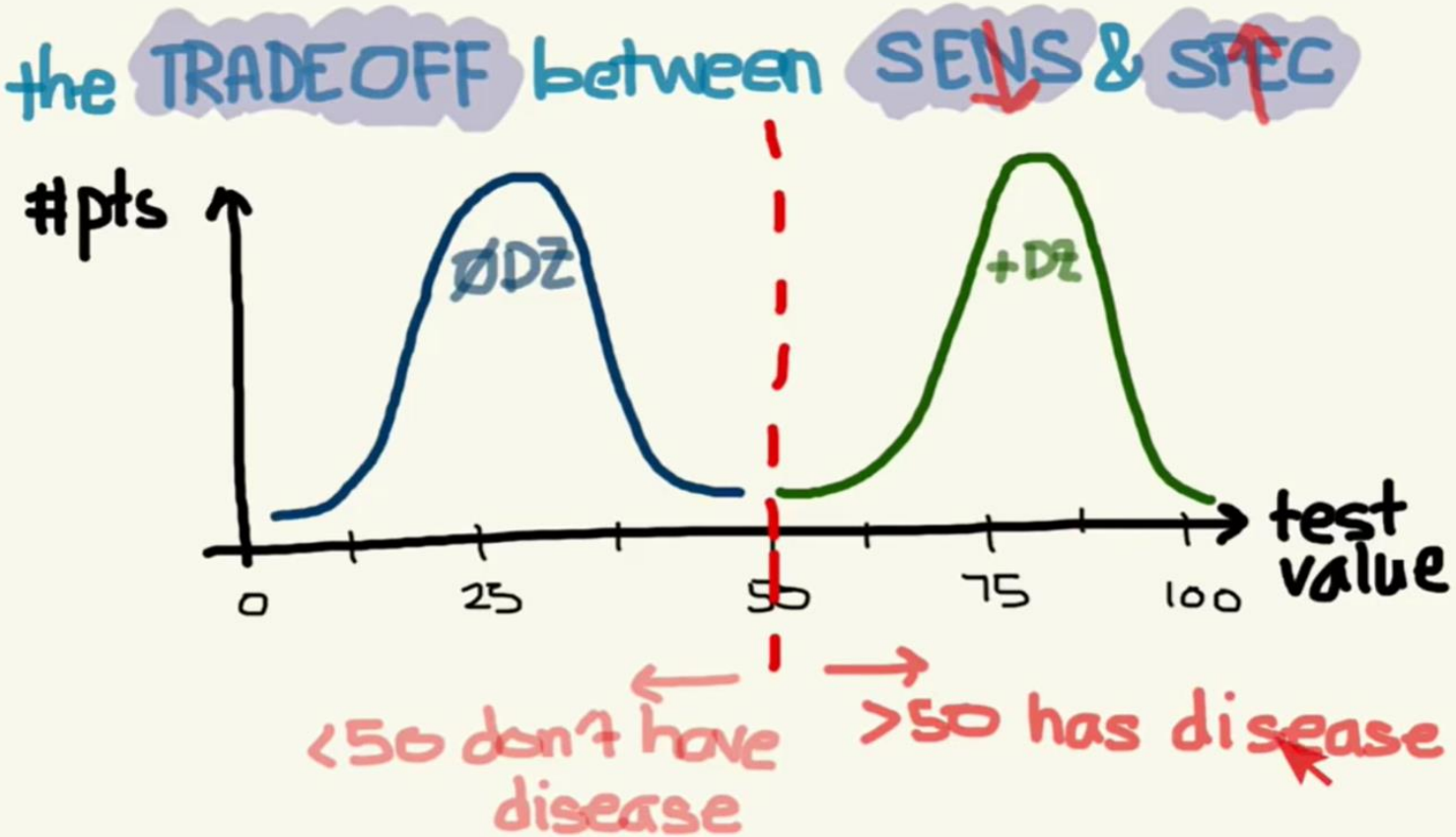
$$F_1\ Score = \text{Not applicable}$$

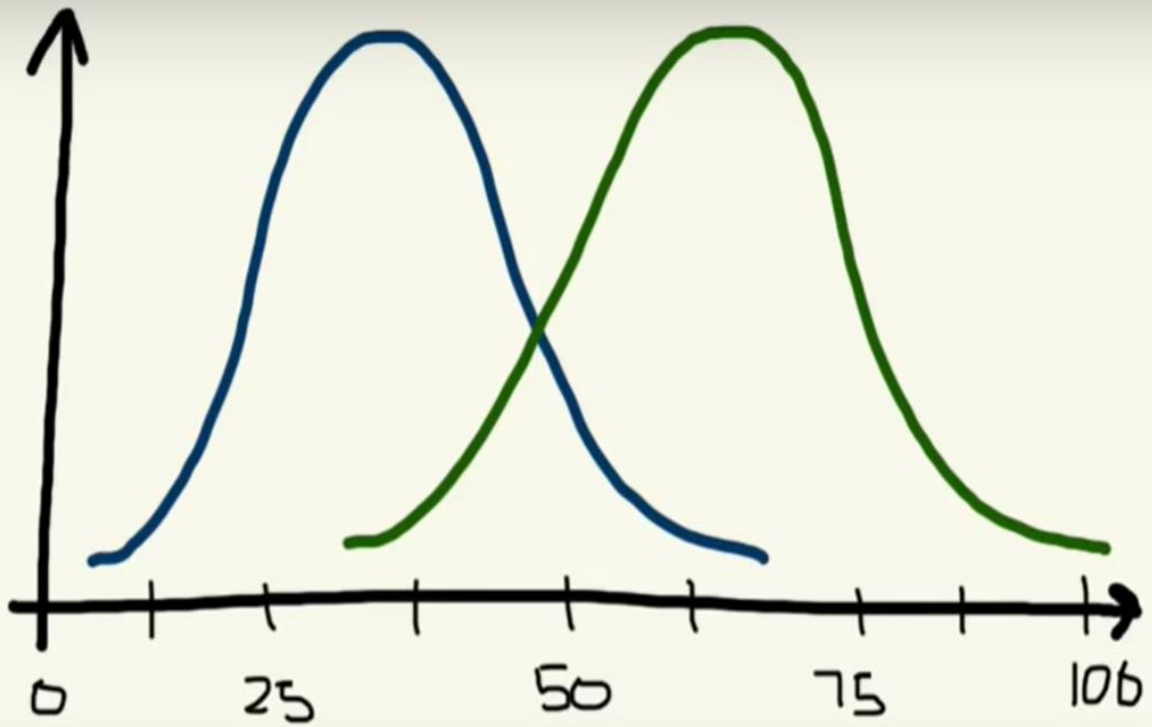
$$\text{Accuracy} = \frac{N}{P+N} = 0$$

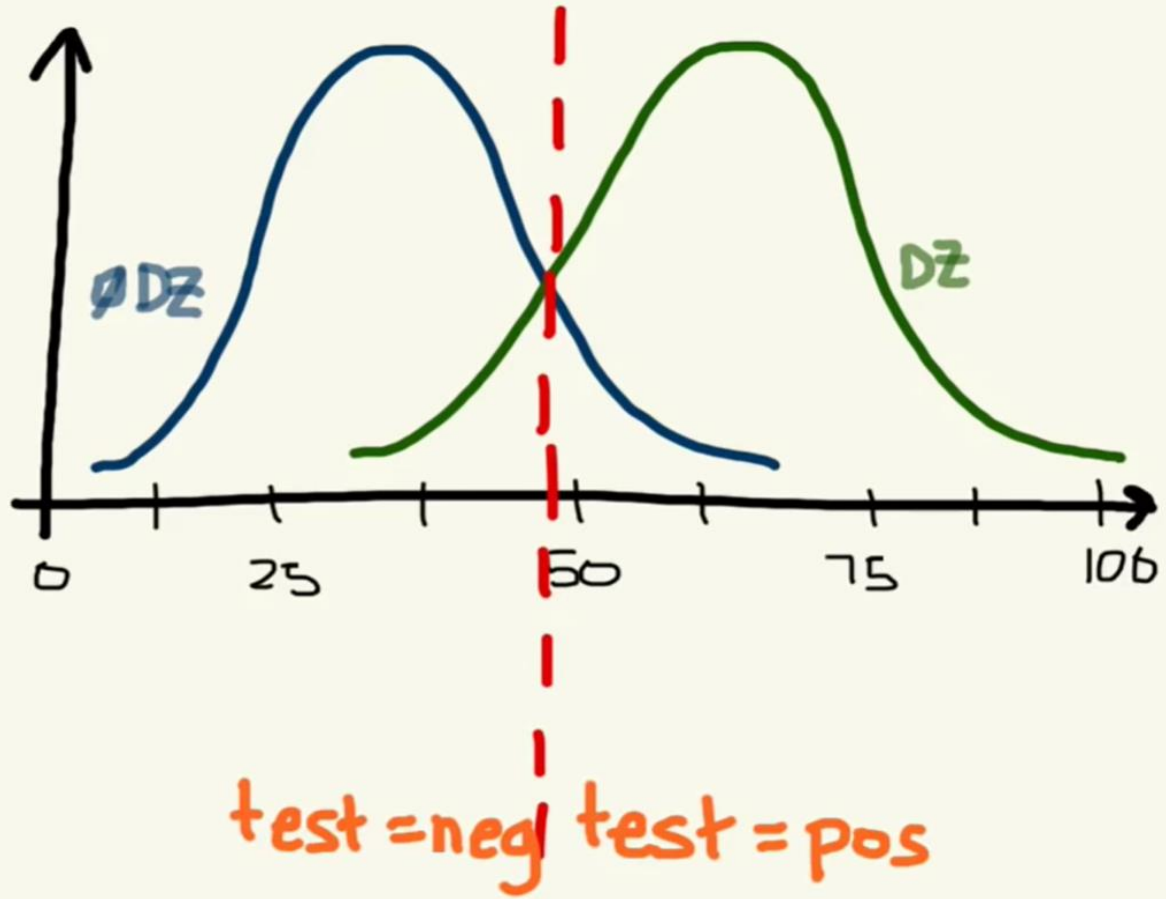
		Predicted Class	
		+	-
Actual class	+	0	p
	-	0	N

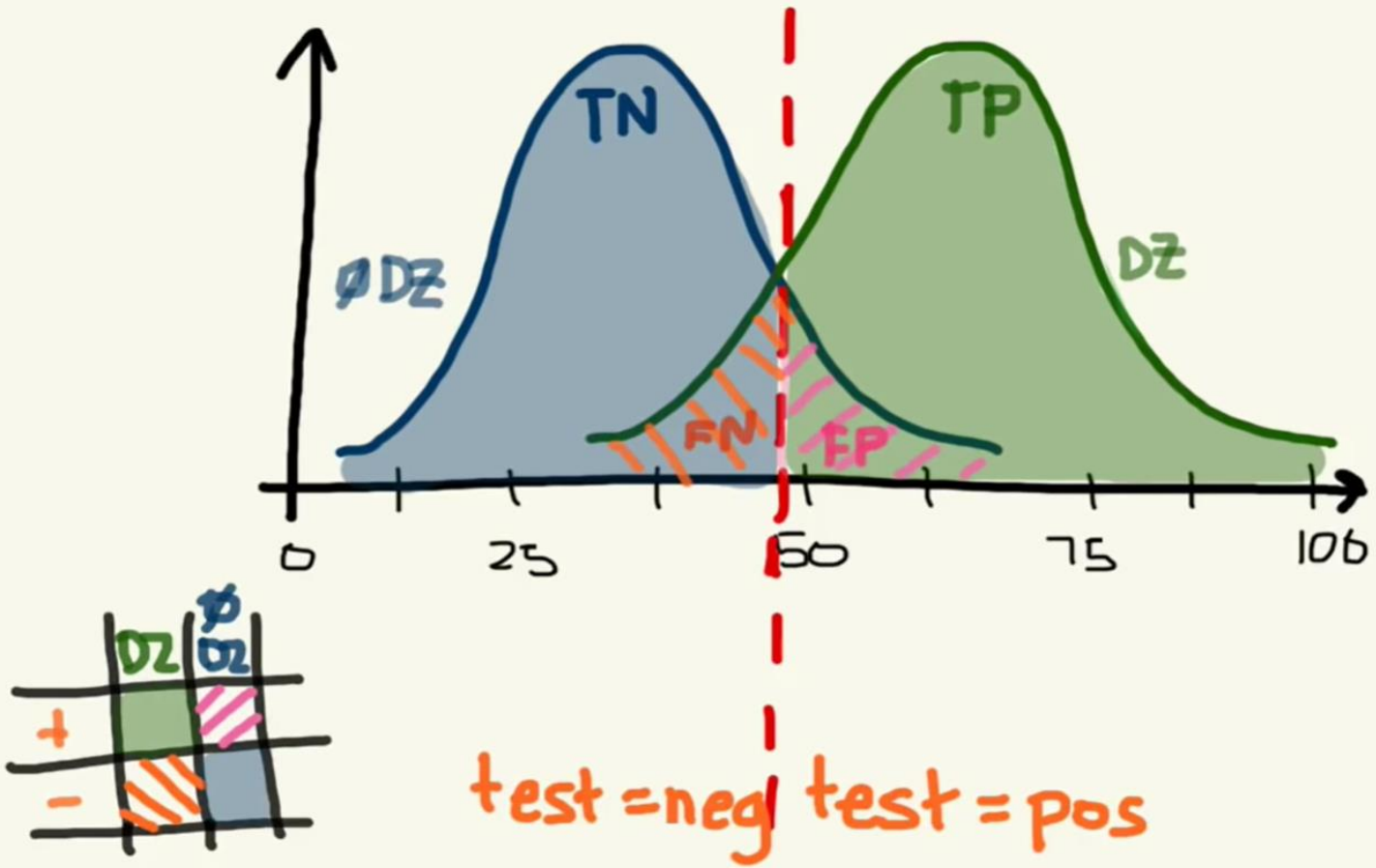
the TRADEOFF between SENS & SPEC

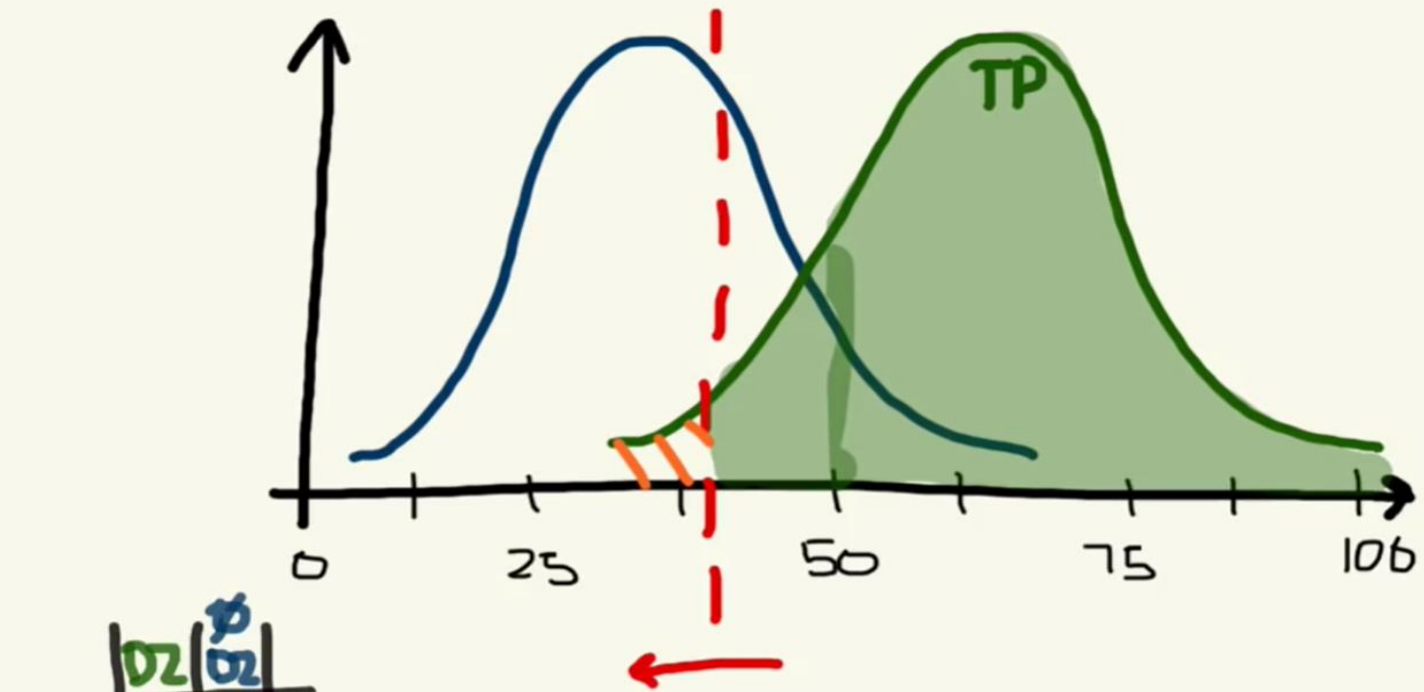


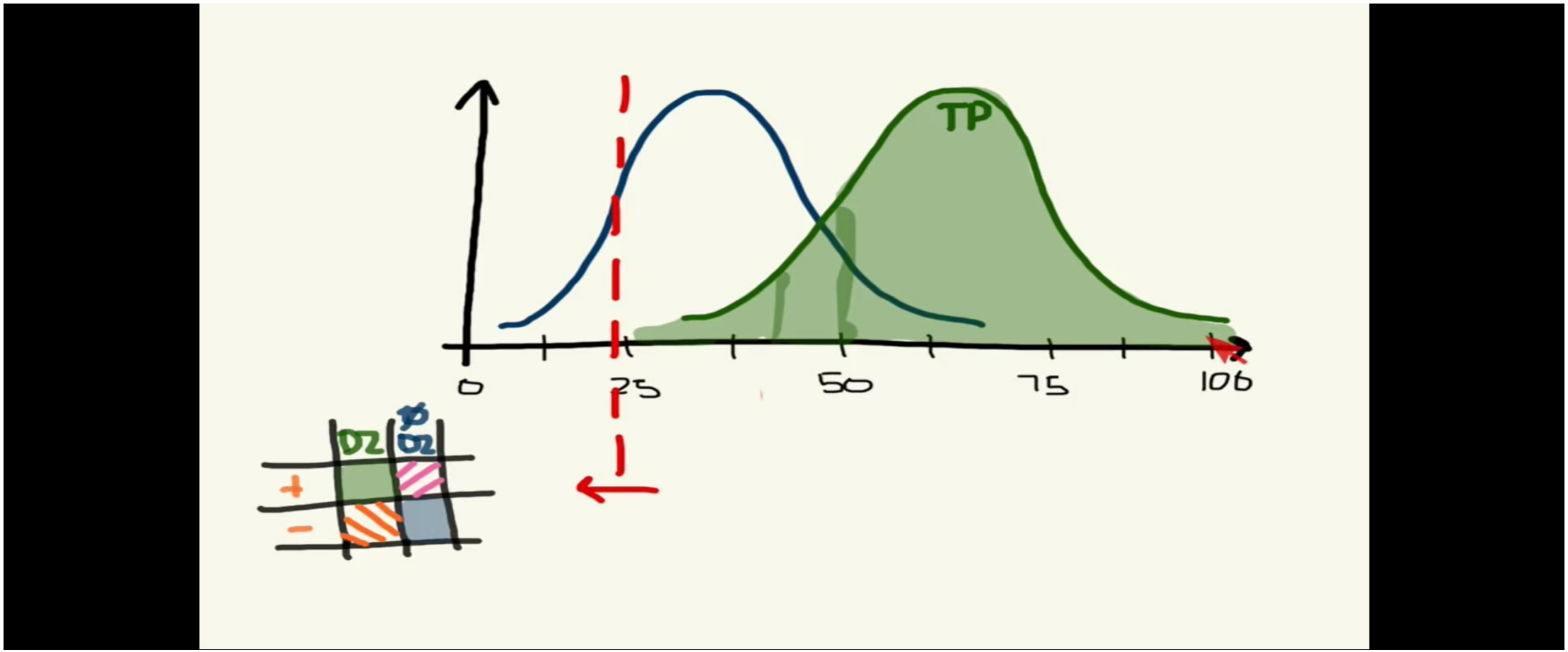

the TRADEOFF between SENS & SPEC
#pts
ØDZ
+Dz
test value
0
25
50
75
100
<50 don't have disease
>50 has disease

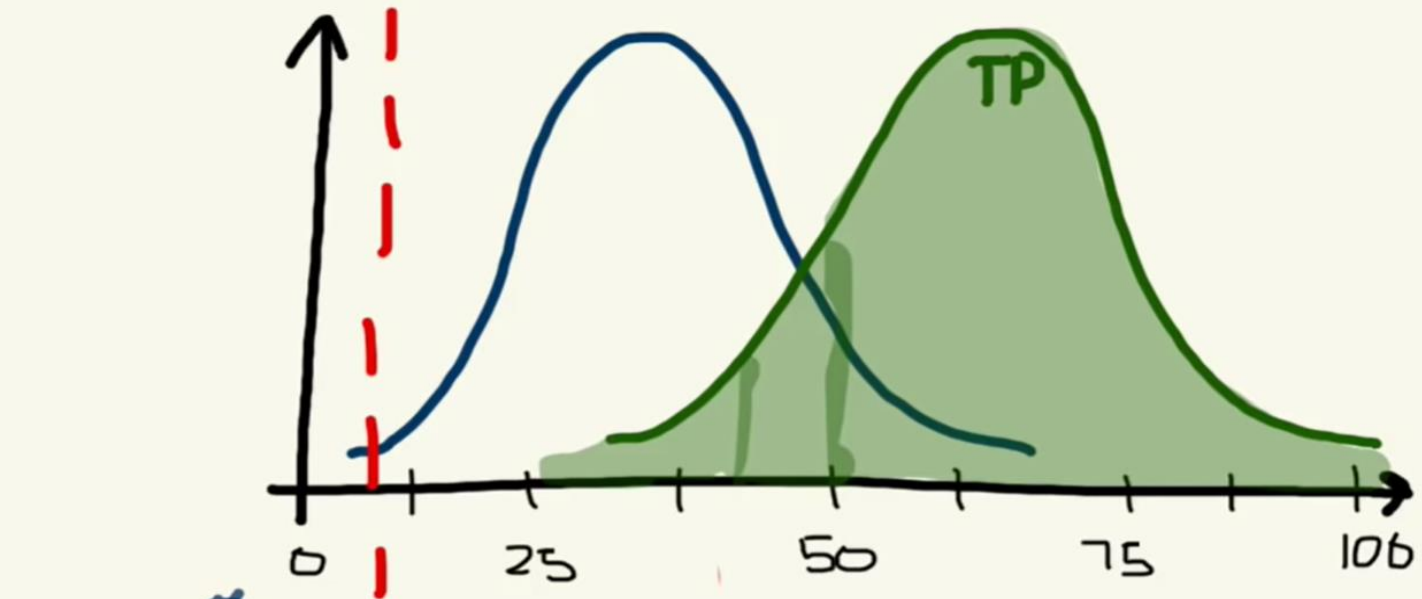

0
25
50
75
106


ØDZ
DZ
0
25
50
75
106
test = neg
test = pos

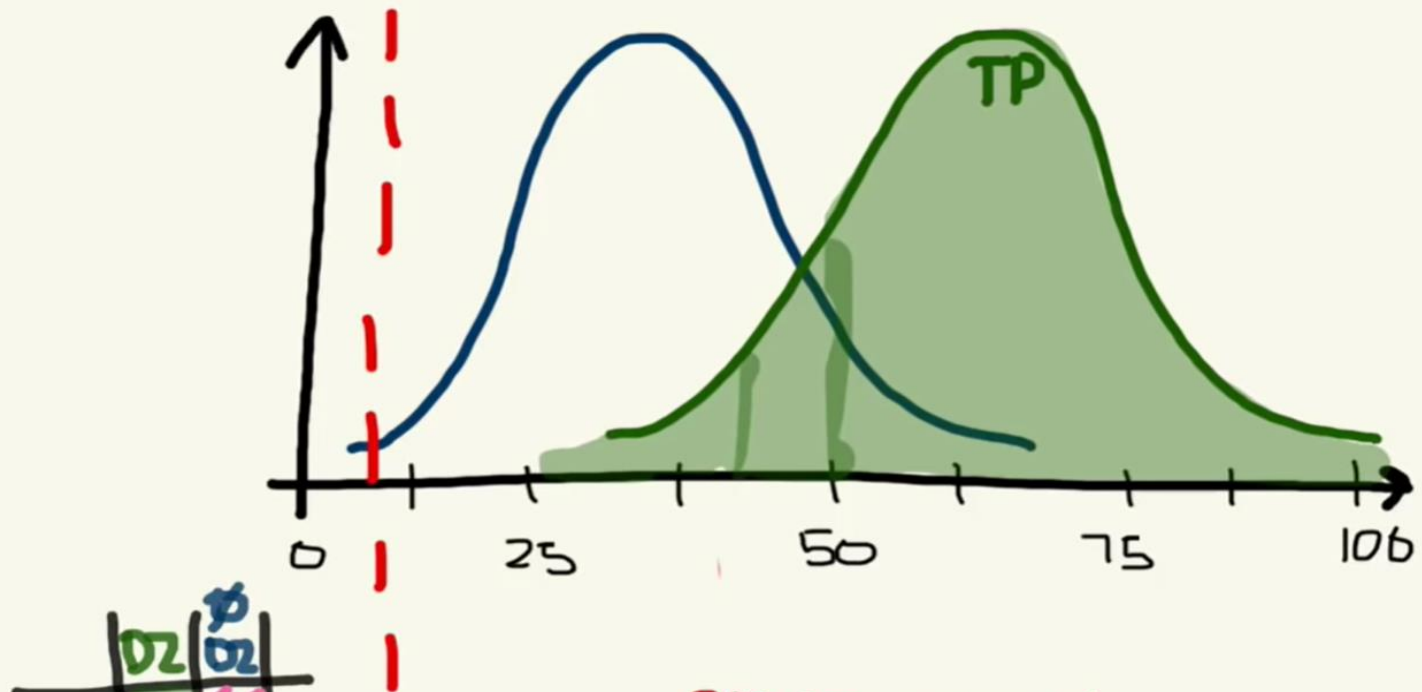

TN
TP
∅ DZ
DZ
FN
FP
0
25
50
75
100
∅ DZ
DZ
+
-
test = neg
test = pos


TP
0
25
50
75
106
+
-


TP
0
25
50
75
106
D2
D2
+
−

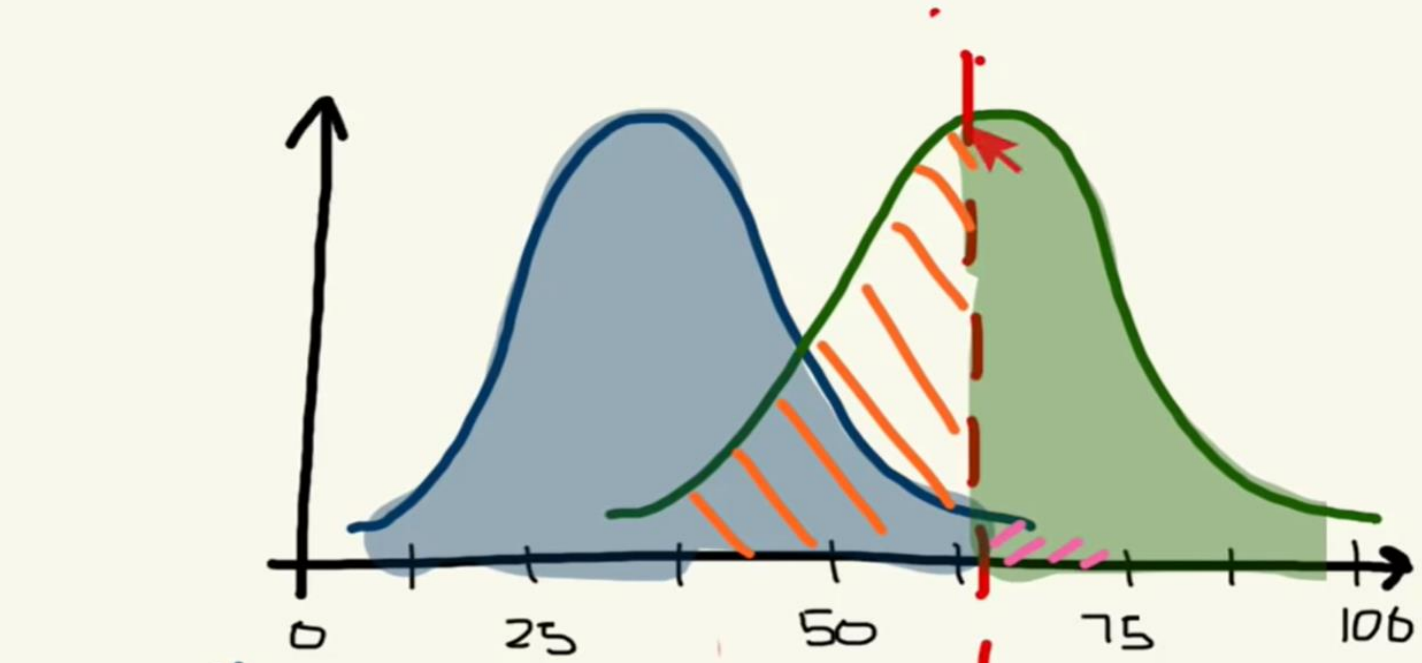


SENS 100%

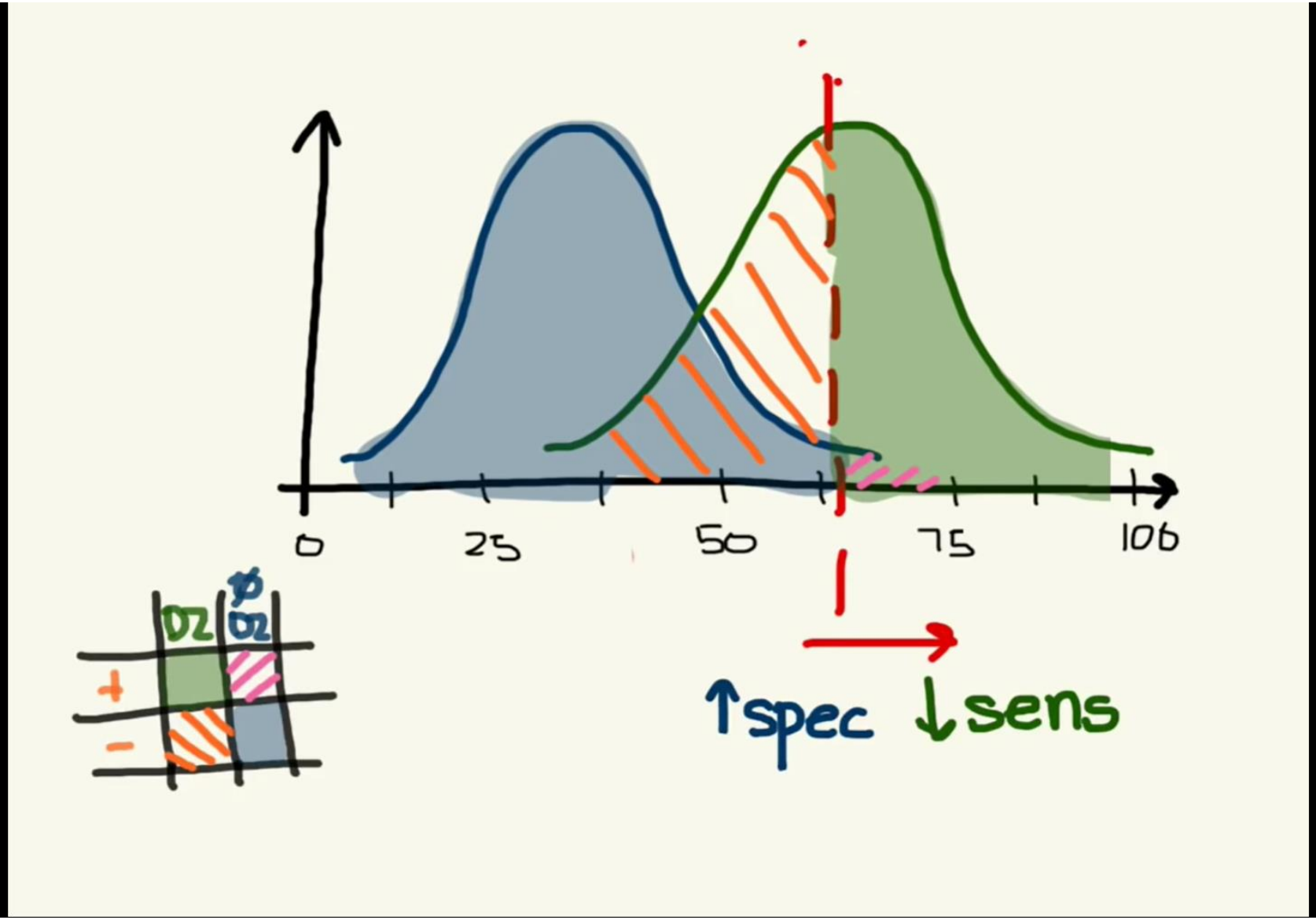


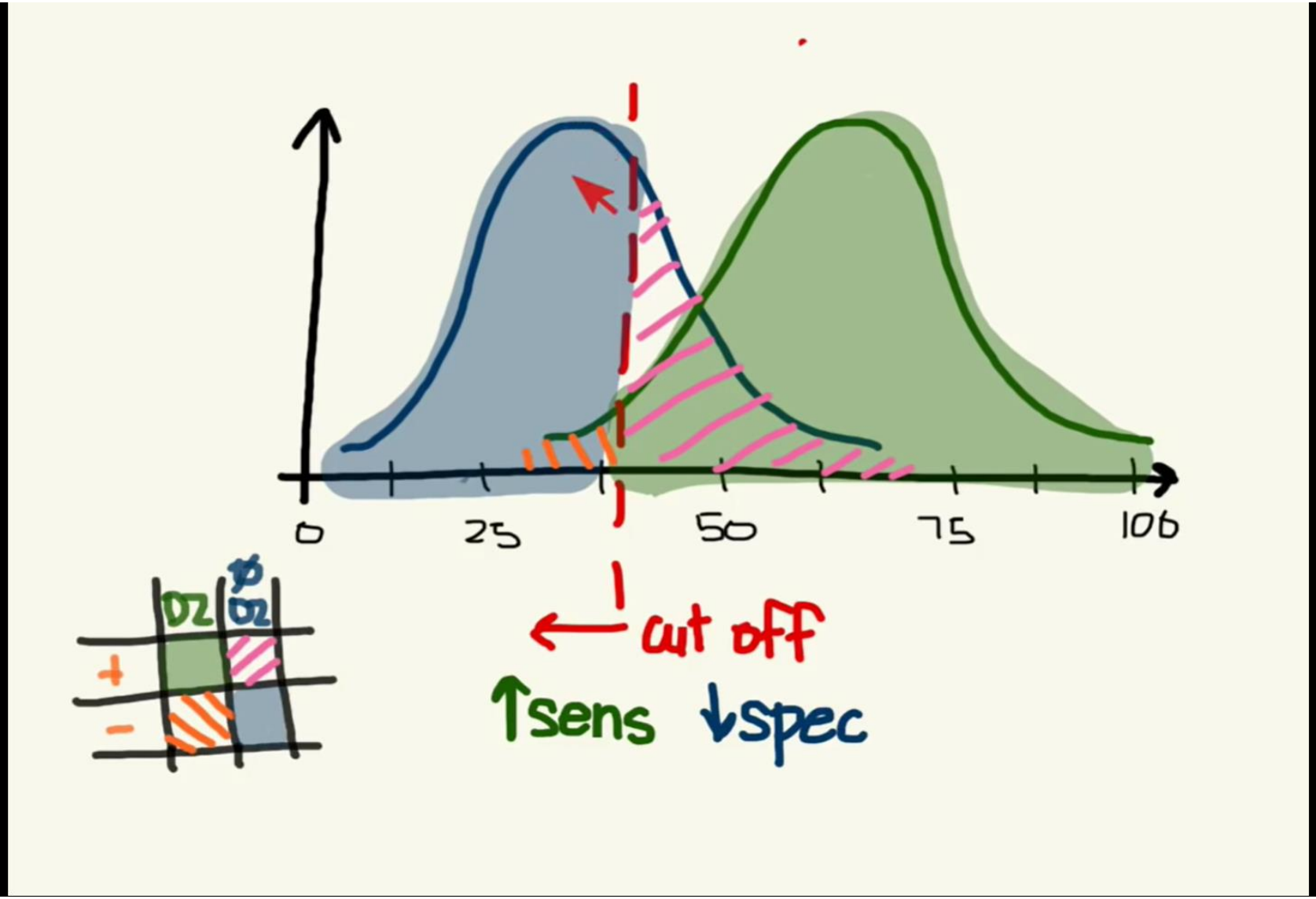
SENS 100%

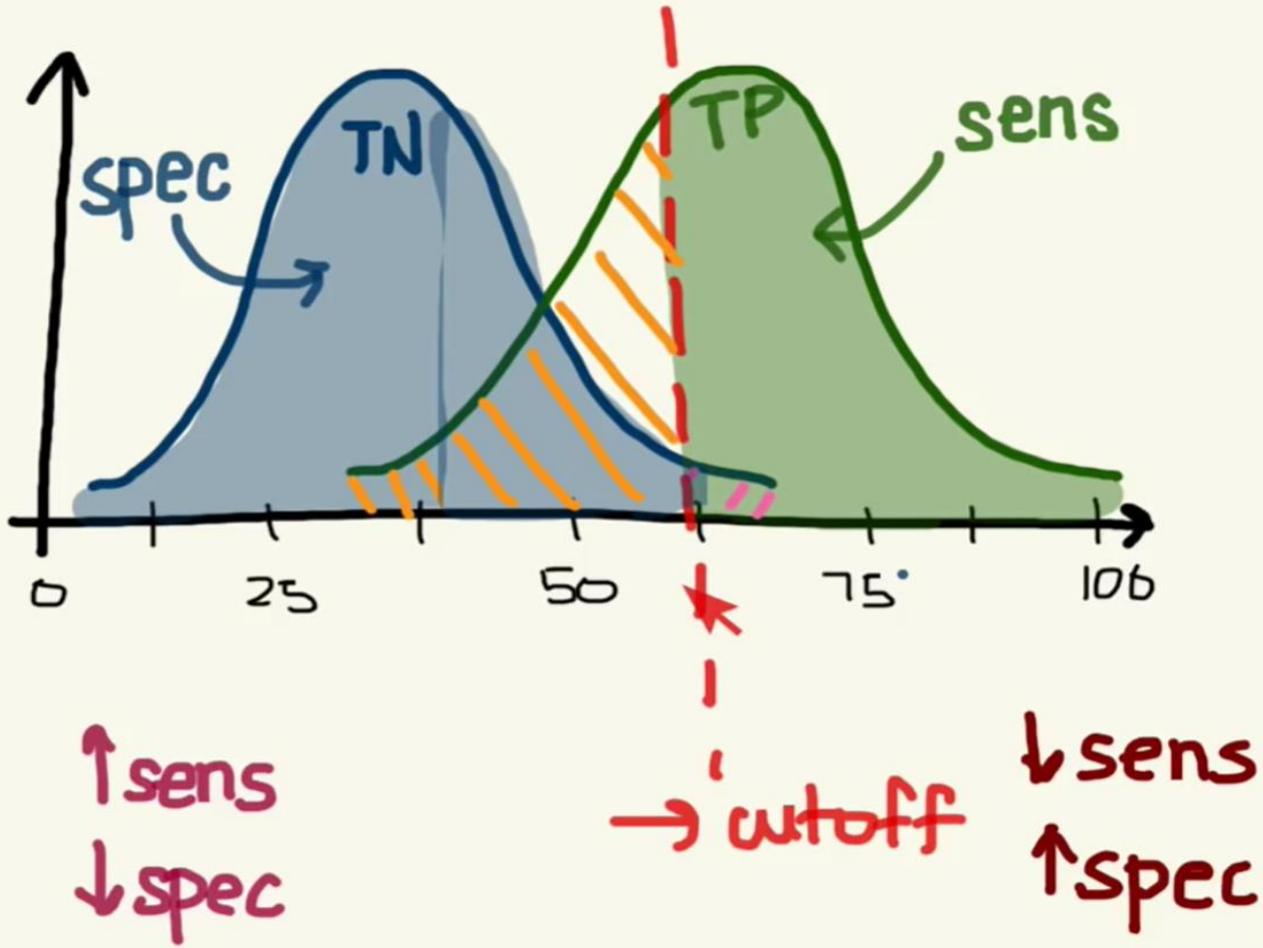
"everyone is positive!!"

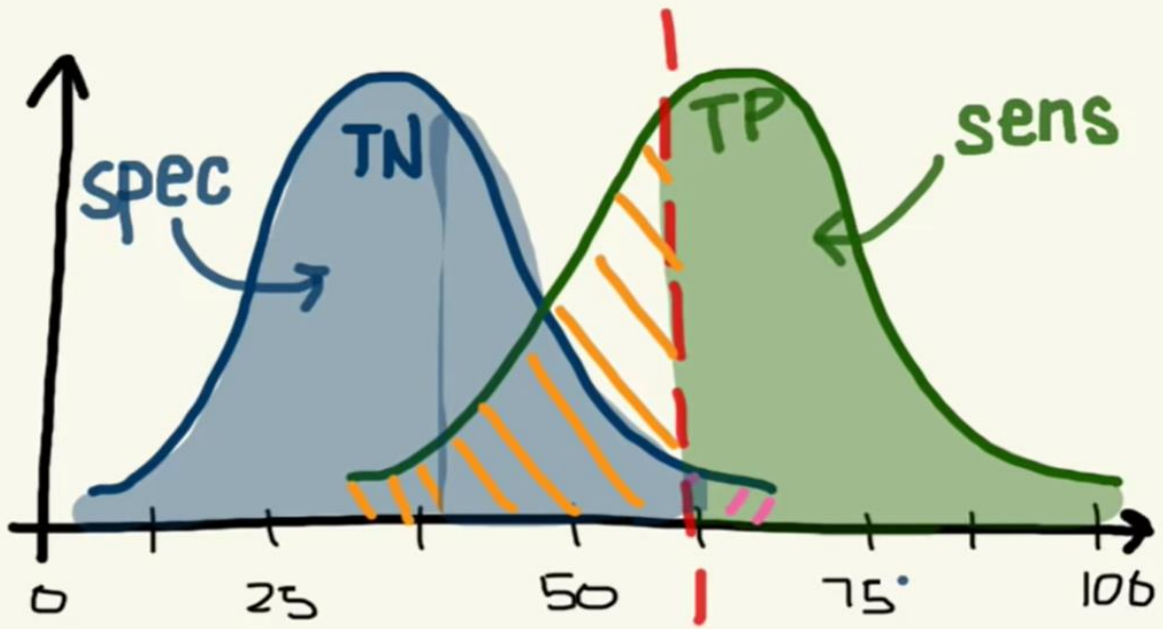

0
25
50
75
106


+
−

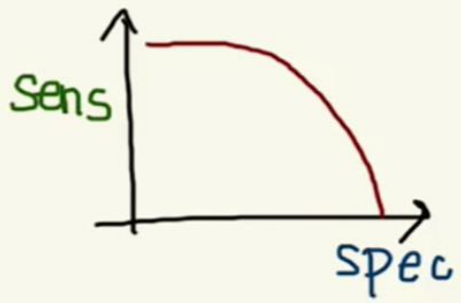

0
25
50
75
106
D2
D2
+
-
↑spec ↓sens


0
25
50
75
106
cut off
↑sens ↓spec
+
-

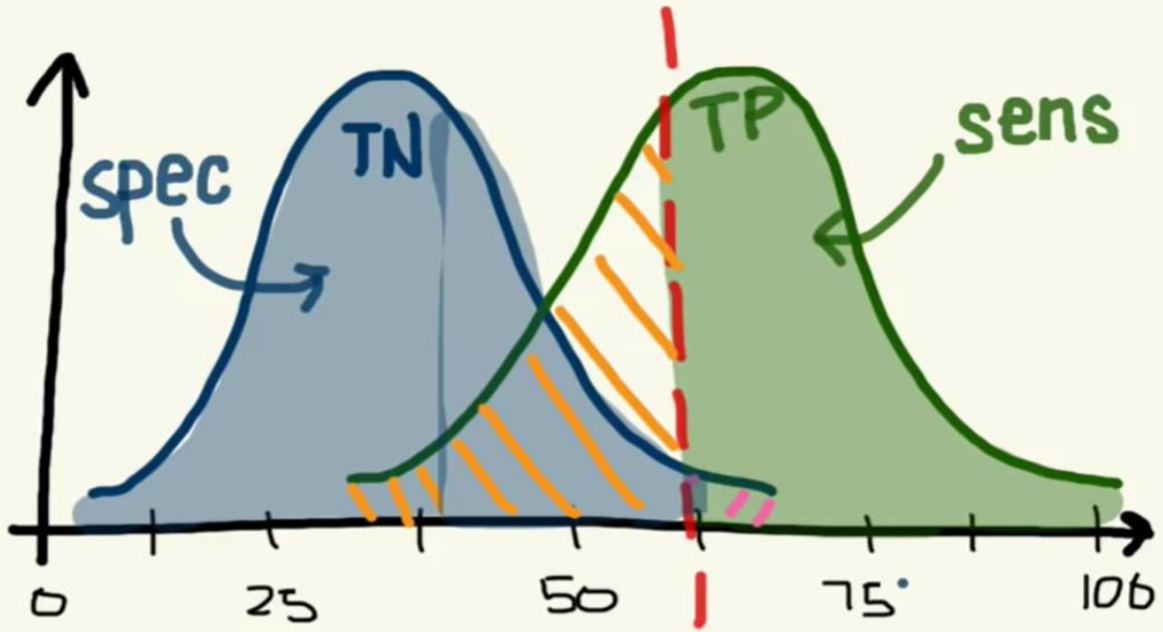

spec
TN
TP
sens
0
25
50
75
106
↑sens
↓spec
→ cutoff
↓sens
↑spec



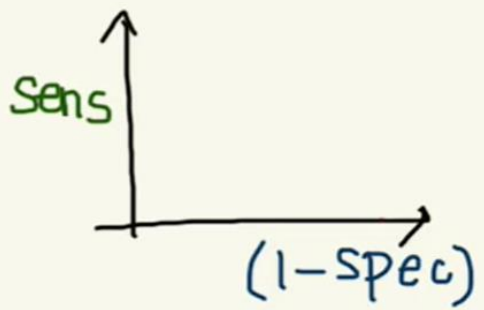
↑sens
↓spec



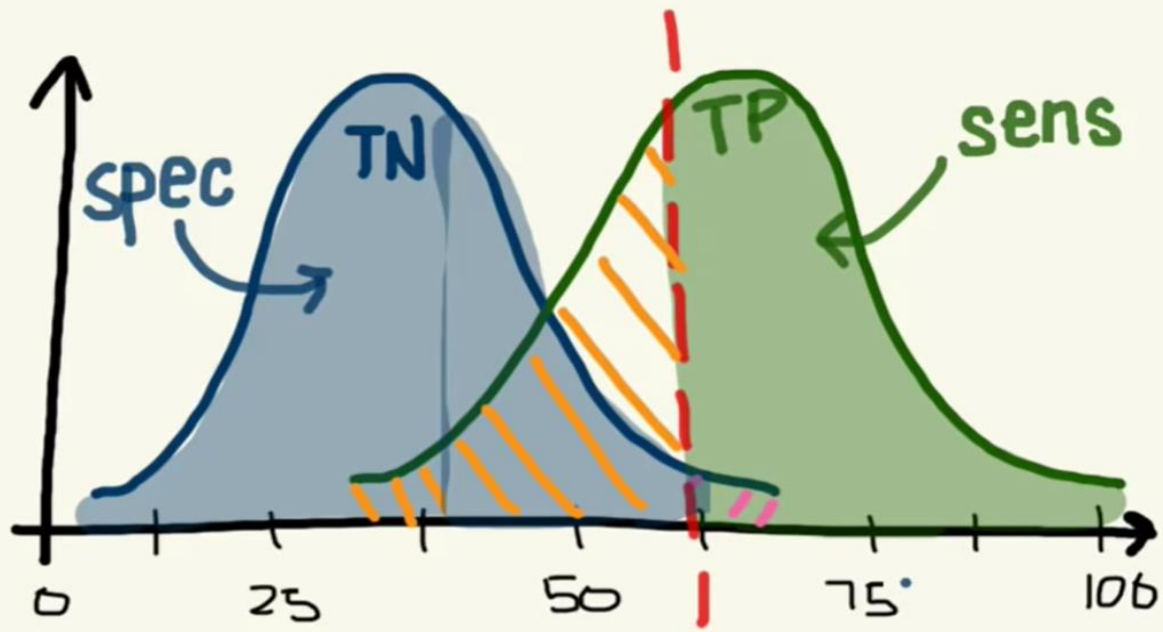
↓sens
↑spec



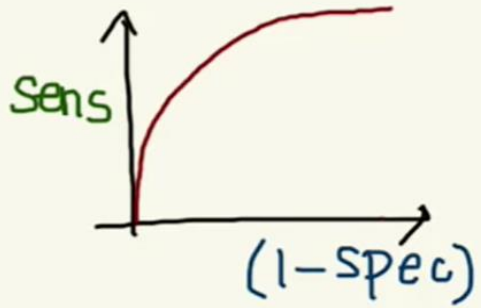
↑sens
↓spec



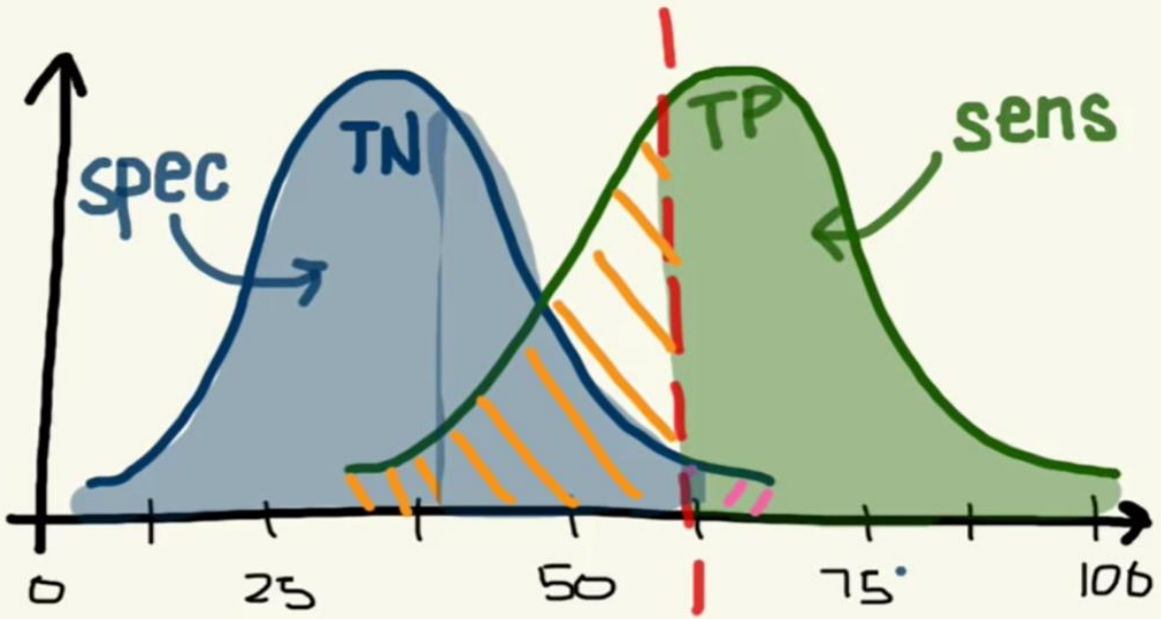
↓sens
↑spec



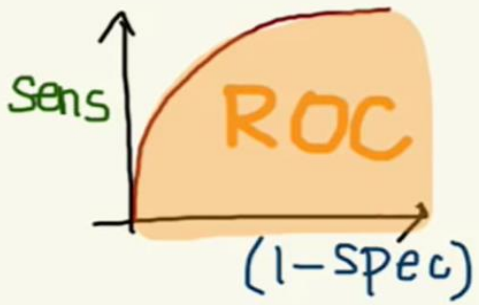
↑sens
↓spec



↓sens
↑spec



↑sens
↓spec



↓sens
↑spec

Receiver Operating Characteristic :

- is a graph showing the performance of a classification model at all classification thresholds.
- This curve plots two parameters:

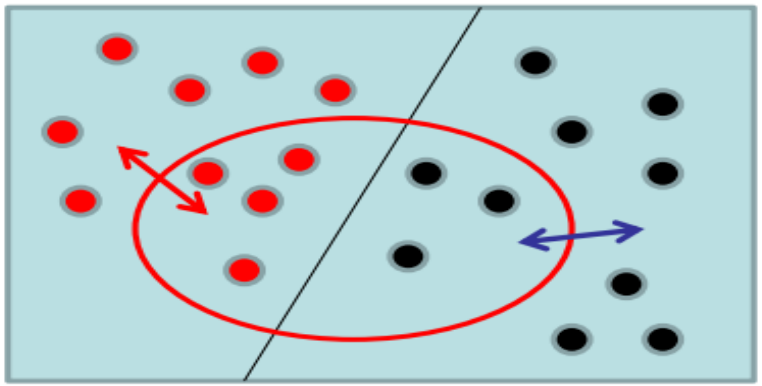
- True Positive Rate
- False Positive Rate

- An ROC curve plots TPR vs. FPR at different classification thresholds.
- Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

True Positive Rate (Fraction) and False Positive Rate (Fraction)

		Predicted class		
		+	-	Total
Actual class	+	True positive	False negative	Positives
	-	False positive	True negative	Negatives

Confusion matrix

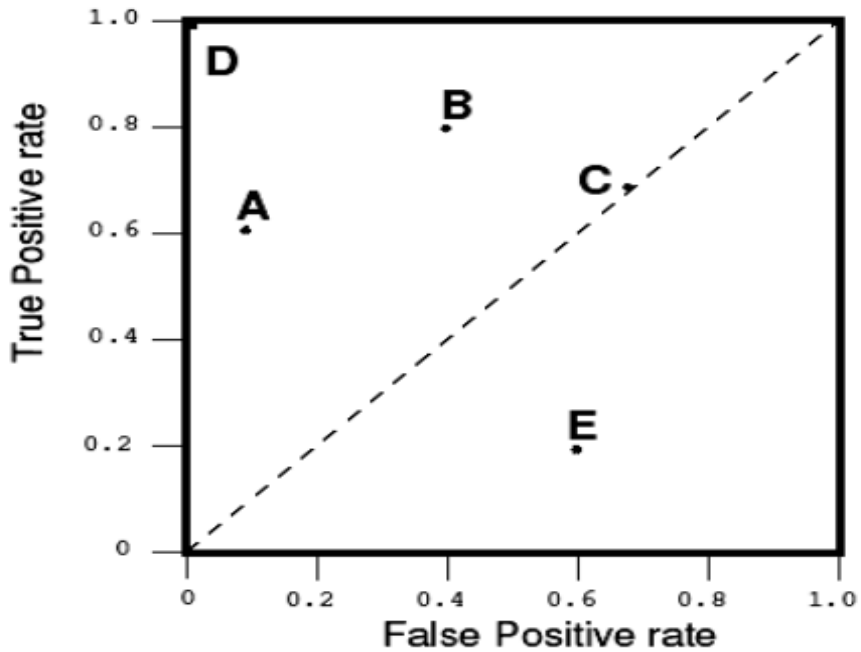


TPF= P correctly classified as P / P

FPF= N incorrectly classified as P / N

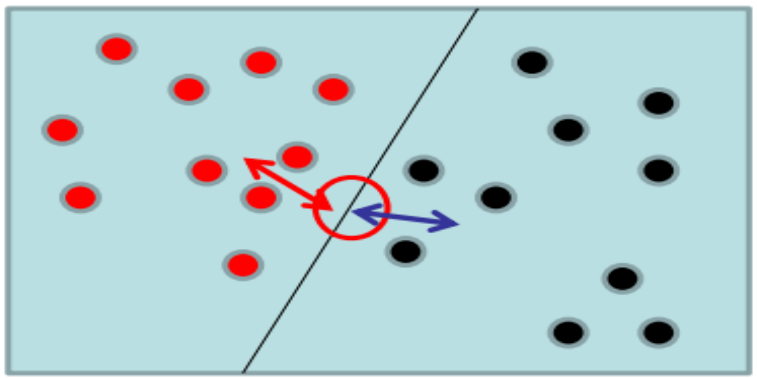
ROC Space

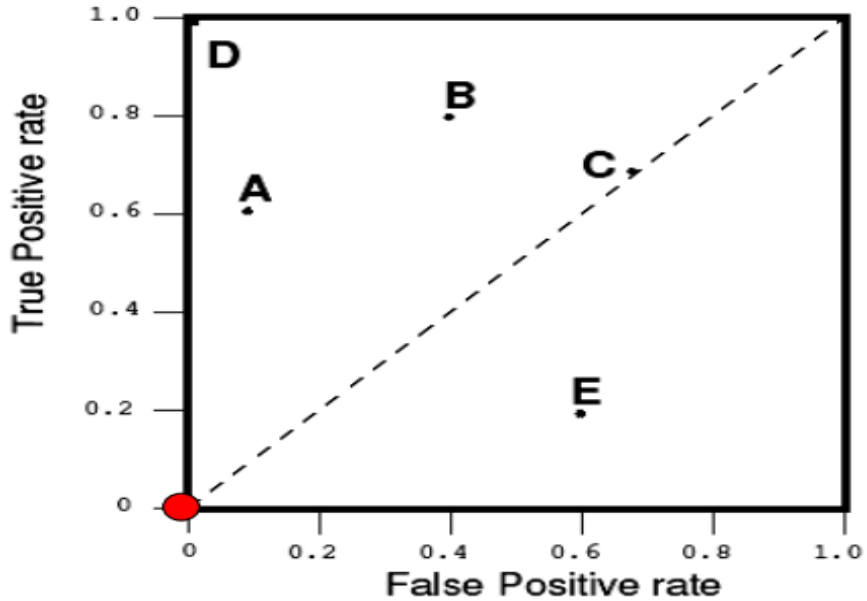
- Figure shows an ROC graph with five classifiers labeled A through E.
- A discrete classier is one that outputs only a class label.
- Each discrete classier produces a confusion matrix (fp rate, tp rate pair) corresponding to a single point in ROC space.
- Classifiers in figure are all discrete classifiers.



Special Points in ROC Space

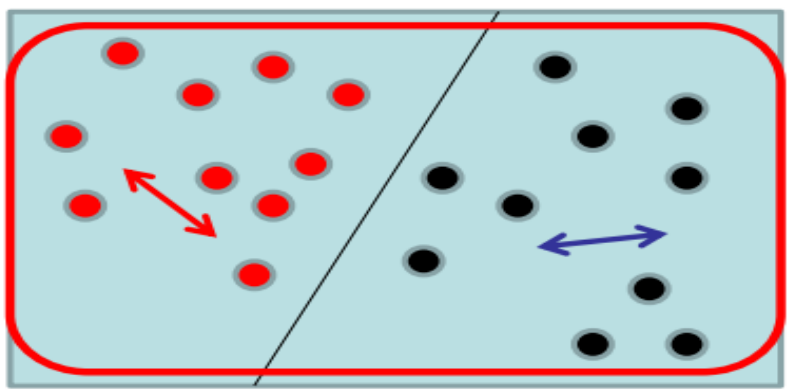
- Lower left point (0, 0) represents the strategy of never issuing a positive classification;
  - such a classier commits no false positive errors but also gains no true positives.

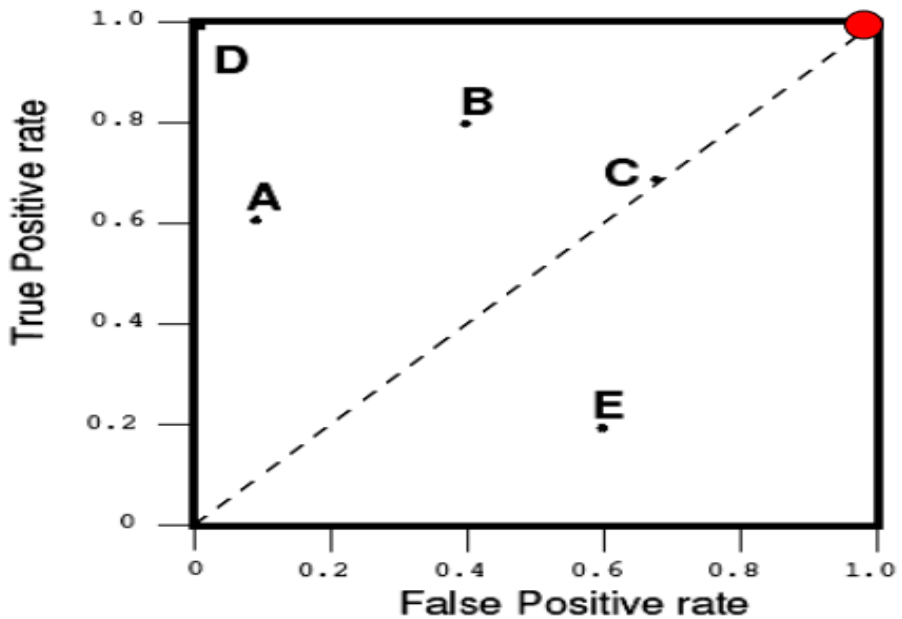




Special Points in ROC Space

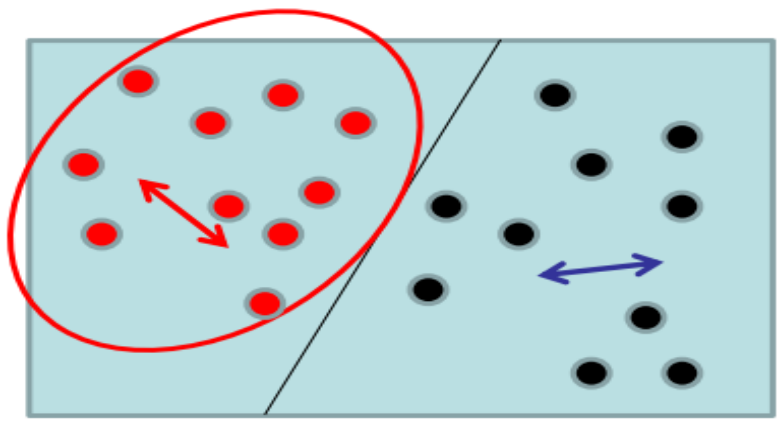
- Upper right corner (1, 1) represents the opposite strategy, of unconditionally issuing positive classifications.

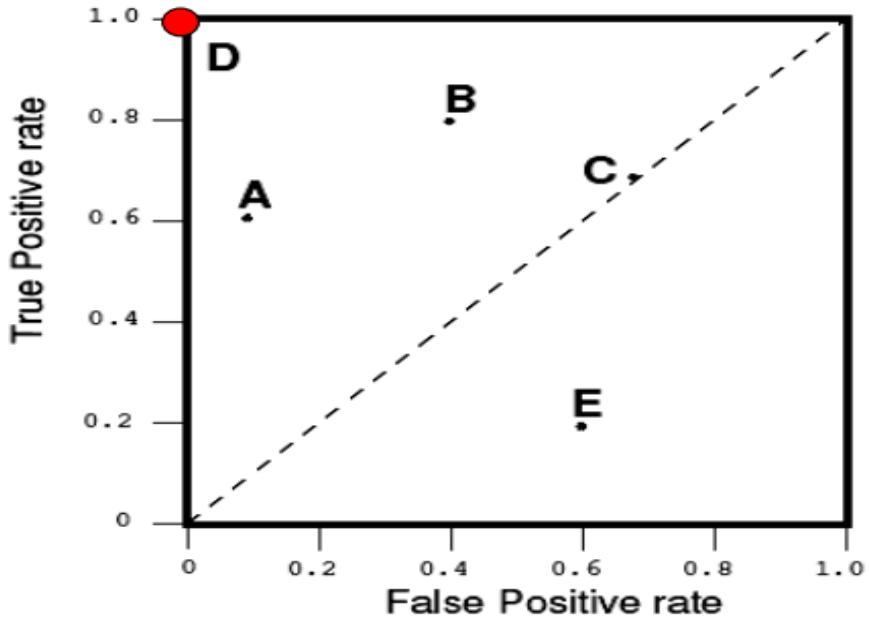




Special Points in ROC Space

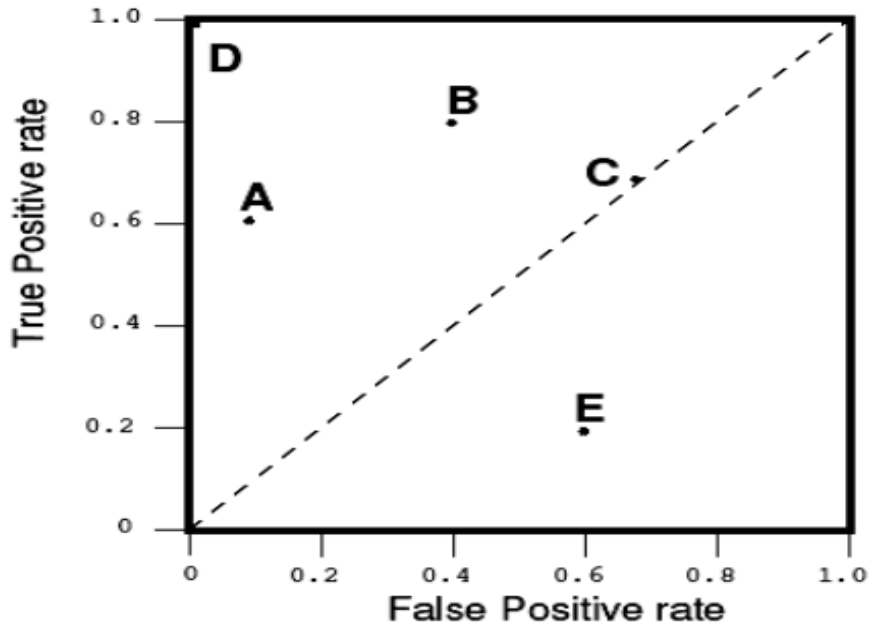
- Point (0, 1) represents perfect classification.
  - D's performance is perfect as shown.





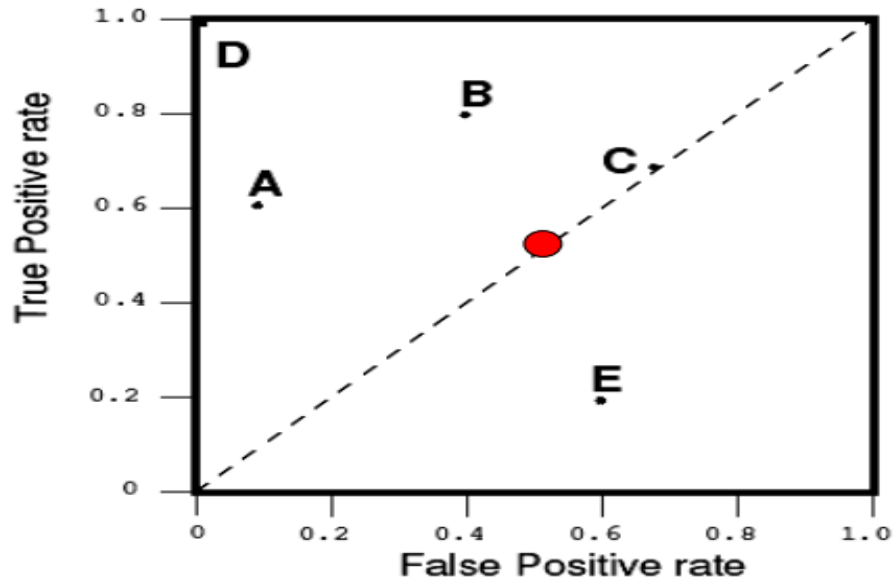
Point in ROC space

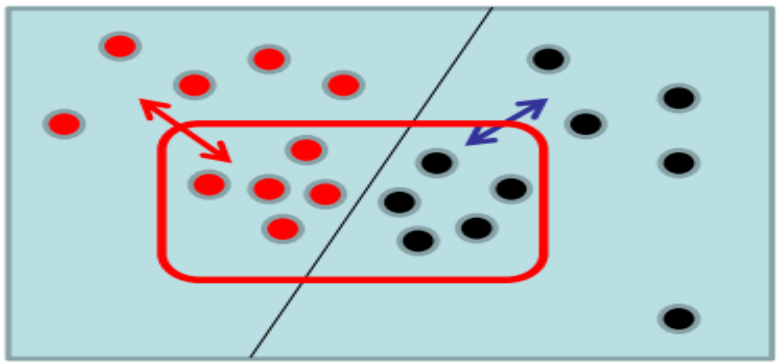
- Informally, one point in ROC space is better than another if it is to the northwest of the first
  - tp rate is higher, fp rate is lower, or both.



Random Classifiers

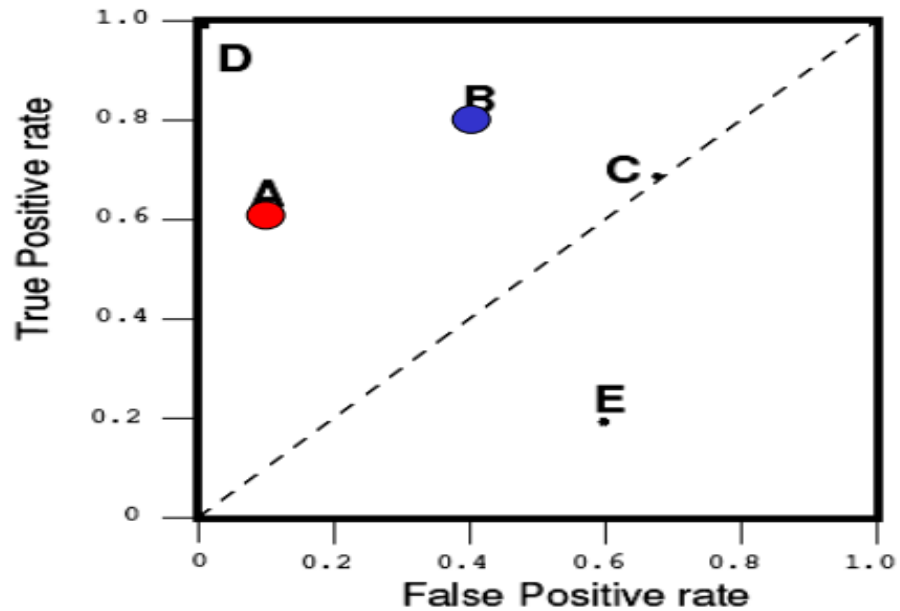
- The diagonal line y = x represents the strategy of randomly guessing a class.
- For example, if a classier randomly says “Positive” half the time (regardless of the instance provided), it can be expected to get half the positives and half the negatives correct;
  - this yields the point (0.5; 0.5)





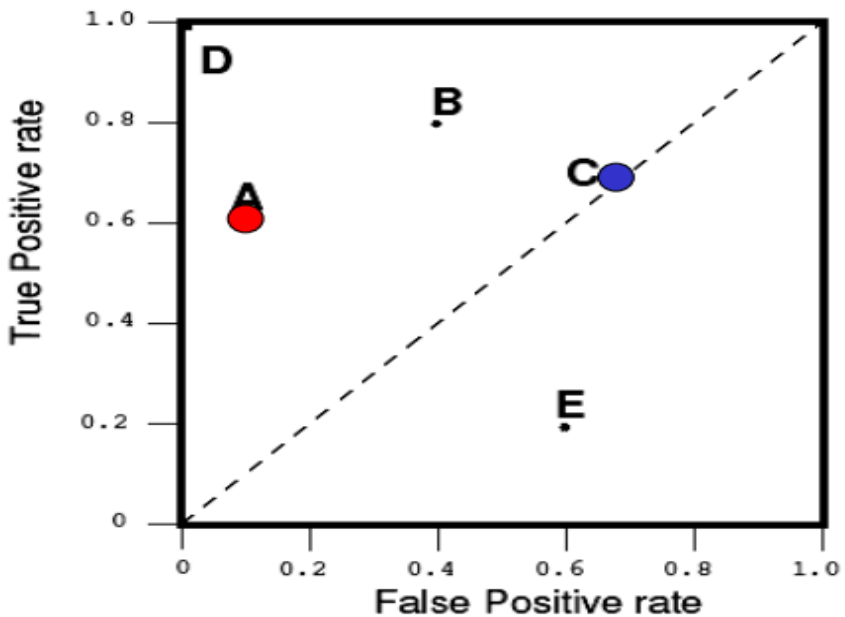
“Conservative” vs. “Liberal”

- Classifiers appearing on the left hand-side of an ROC graph, near the Y axis, may be thought of as “conservative”
  - they make positive classifications only with strong evidence so they make few false positive errors,
  - but they often have low true positive rates as well.
- In figure, A is more conservative than B.



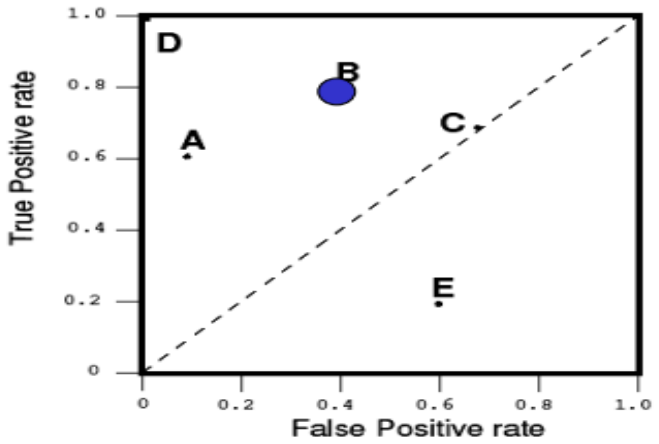
“Conservative” vs. “Liberal”

- Classifiers on the upper right-hand side of an ROC graph may be thought of as “liberal”
  - they make positive classifications with weak evidence so they classify nearly all positives correctly,
  - but they often have high false positive rates.
- In figure, C is more liberal than A.



Curves and points in ROC space

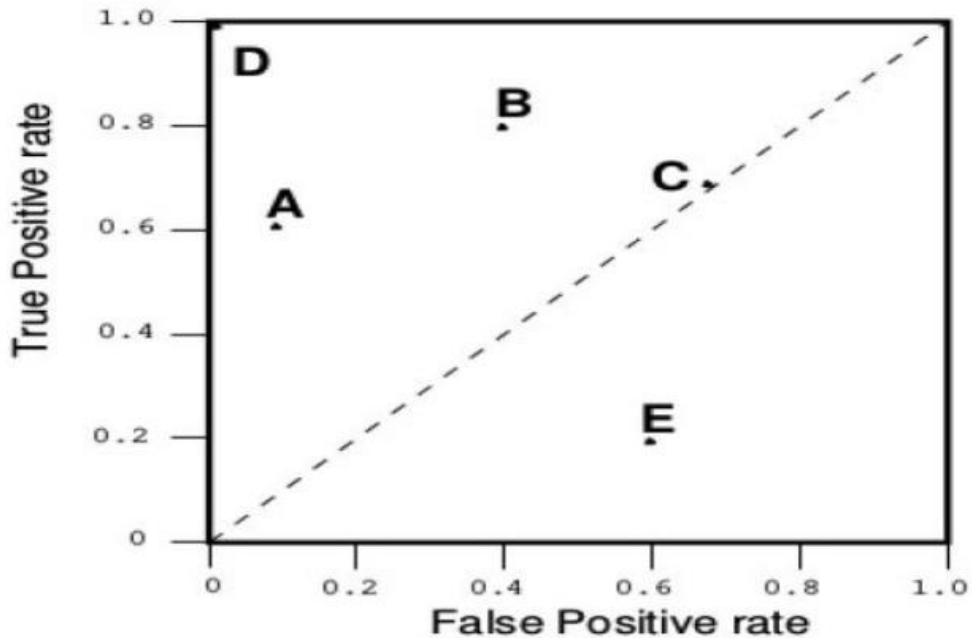
- Many classifiers, such as decision trees or rule learners, are designed to produce only a class decision, i.e., a Y or N on each instance.
  - When such a discrete classier is applied to a test set, it yields a single confusion matrix, which in turn corresponds to one ROC point.
  - Thus, a discrete classifier produces only a single point in ROC space.



		Predicted class	
		Class +	Class -
Actual class	Class +	80	20
	Class -	40	60

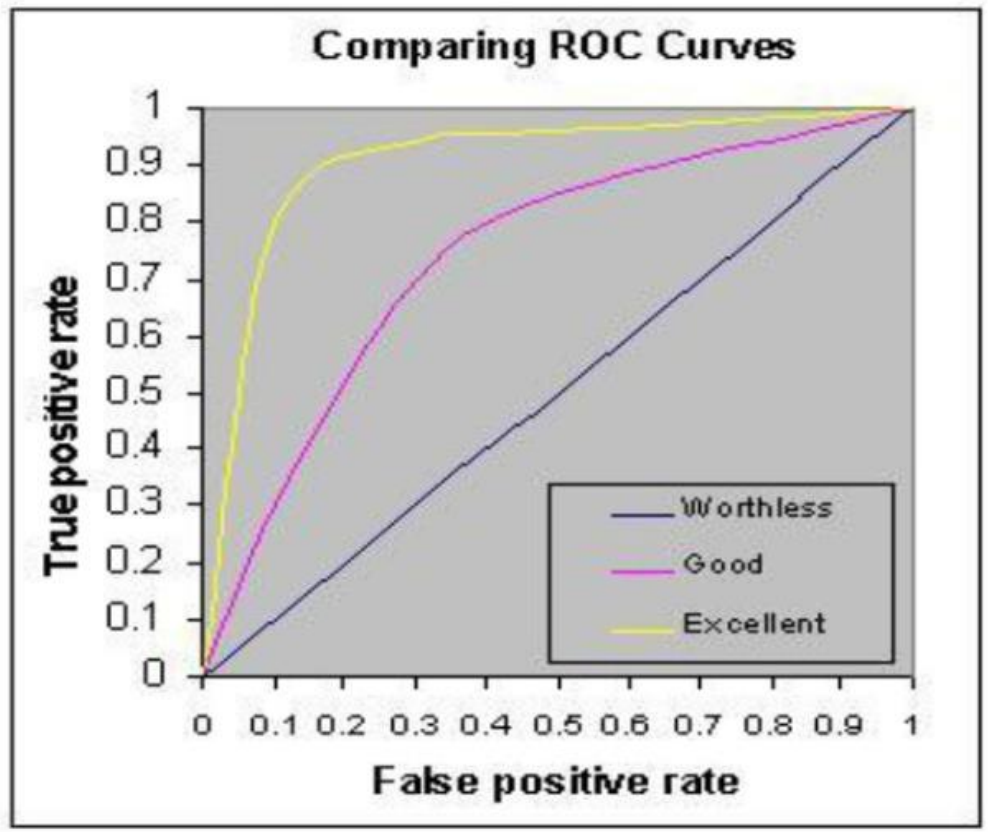
Confusion matrix for classifier B

ROC Graph



→ One point in ROC space is better than another if it is to the northwest (tp rate is higher, fp rate is lower, or both) of the first. (D is better than all other points)
→ Classifiers in the lower left part are known as conservative.
→ Classifiers in the upper right part are known as liberals.
→ the diagonal y=x line represents a random classifier.
→ Classifiers in the lower right triangle performs worse than random classifier.

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)



Algorithm

- Exploit monotonicity of thresholded classifications:
  - Any instance that is classified positive with respect to a given threshold will be classified positive for all lower thresholds as well.

- Therefore, we can simply:
  - sort the test instances decreasing by their scores and
  - move down the list, processing one instance at a time and
  - update TP and FP as we go.

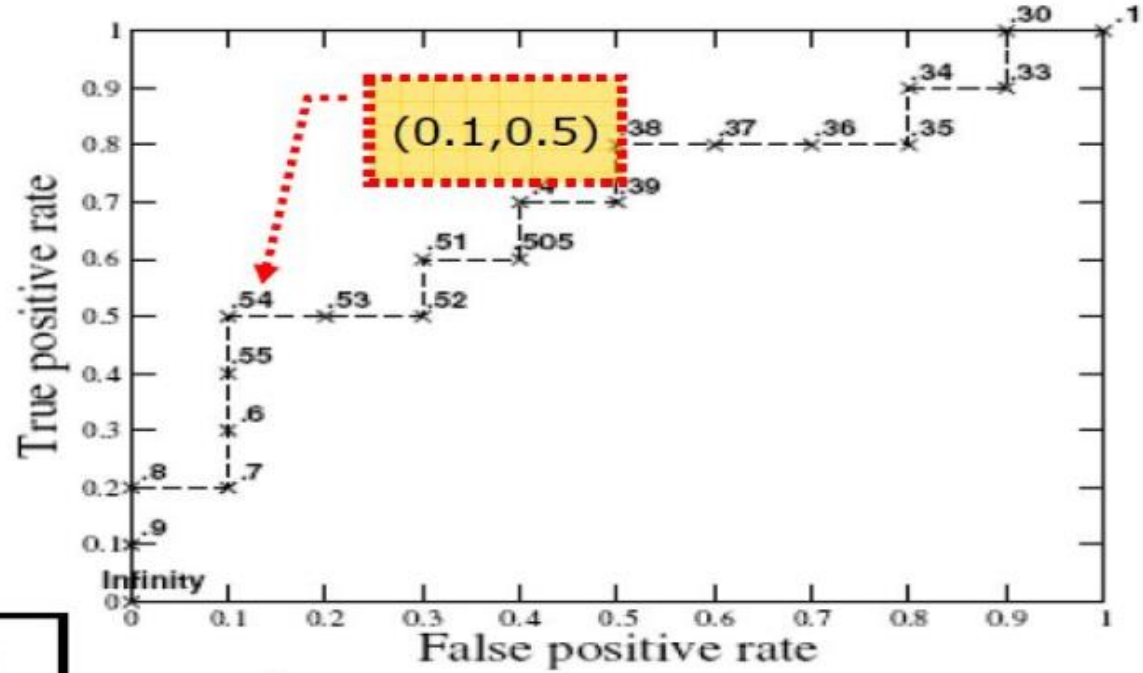
- In this way, an ROC graph can be created from a linear scan.

Use thresholds to create ROC curve

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

If threshold =0.54 →
Numbers of Score $\geqq$0.54 → 6

5	1	6
5	9	14
10	10	20



$$x : \frac{1}{10} = 0.1$$

$$y : \frac{5}{10} = 0.5$$

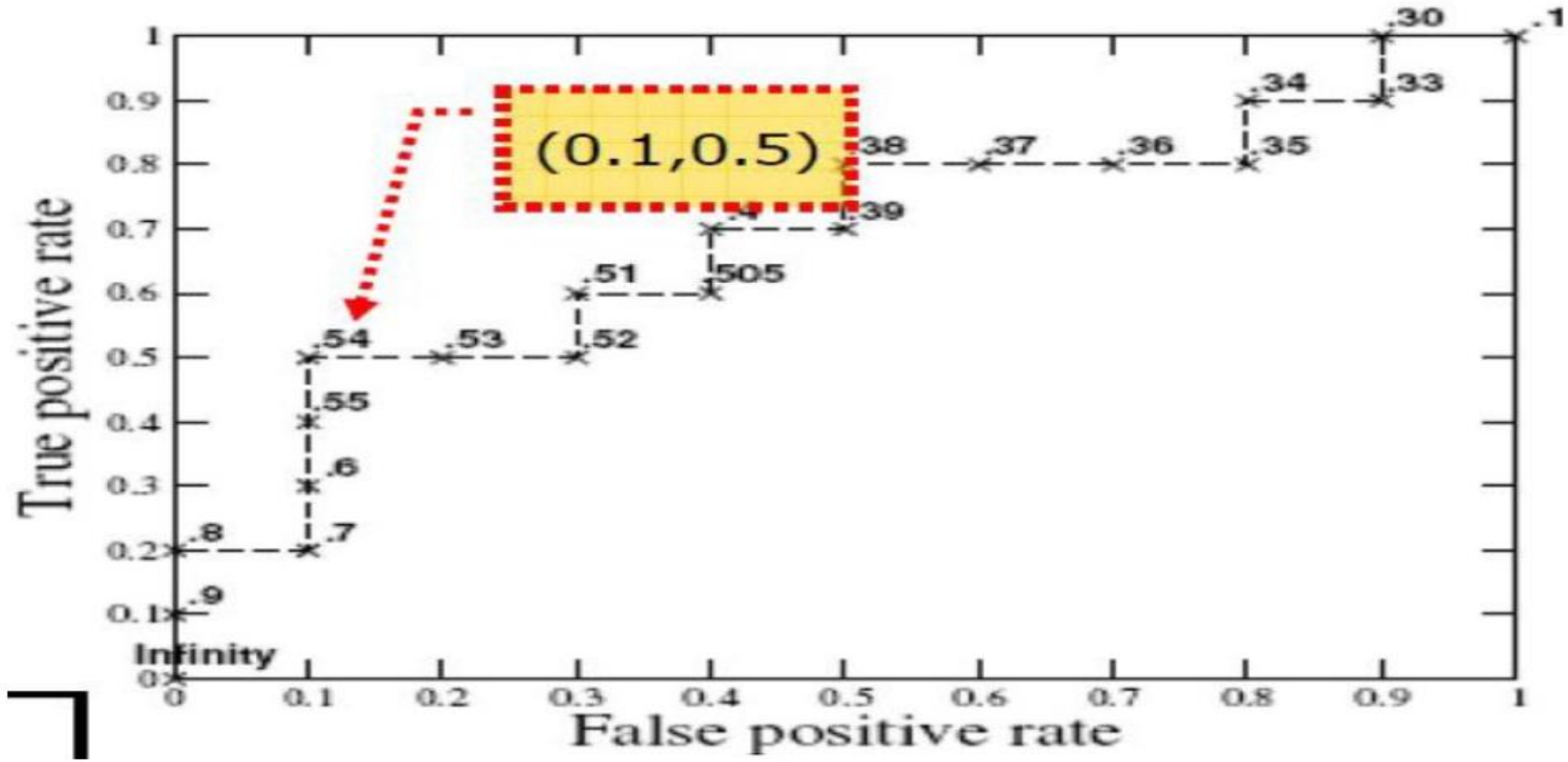
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

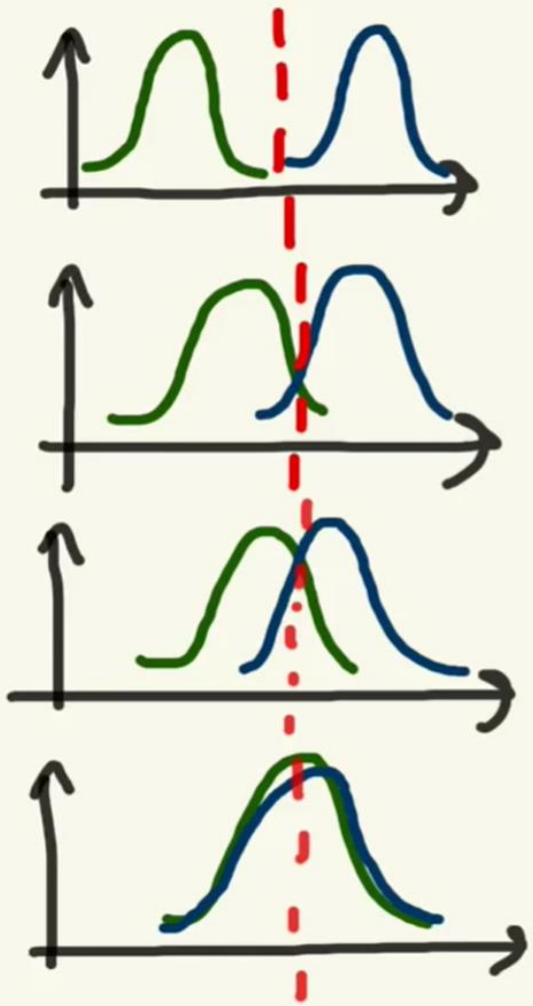
Threshold - 0.54

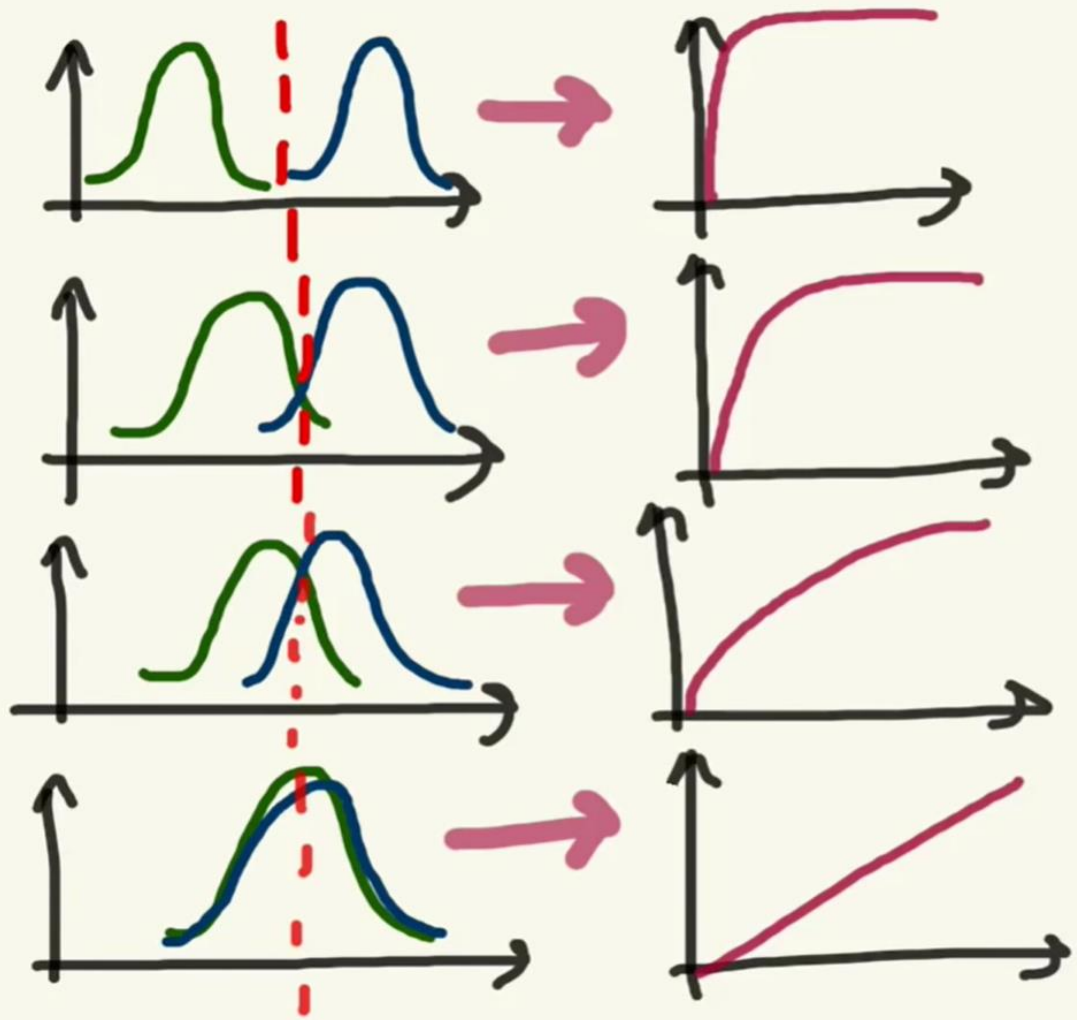
	p	n	
Threshold	5	1	6
Not Threshold	5	9	14
	10	10	20

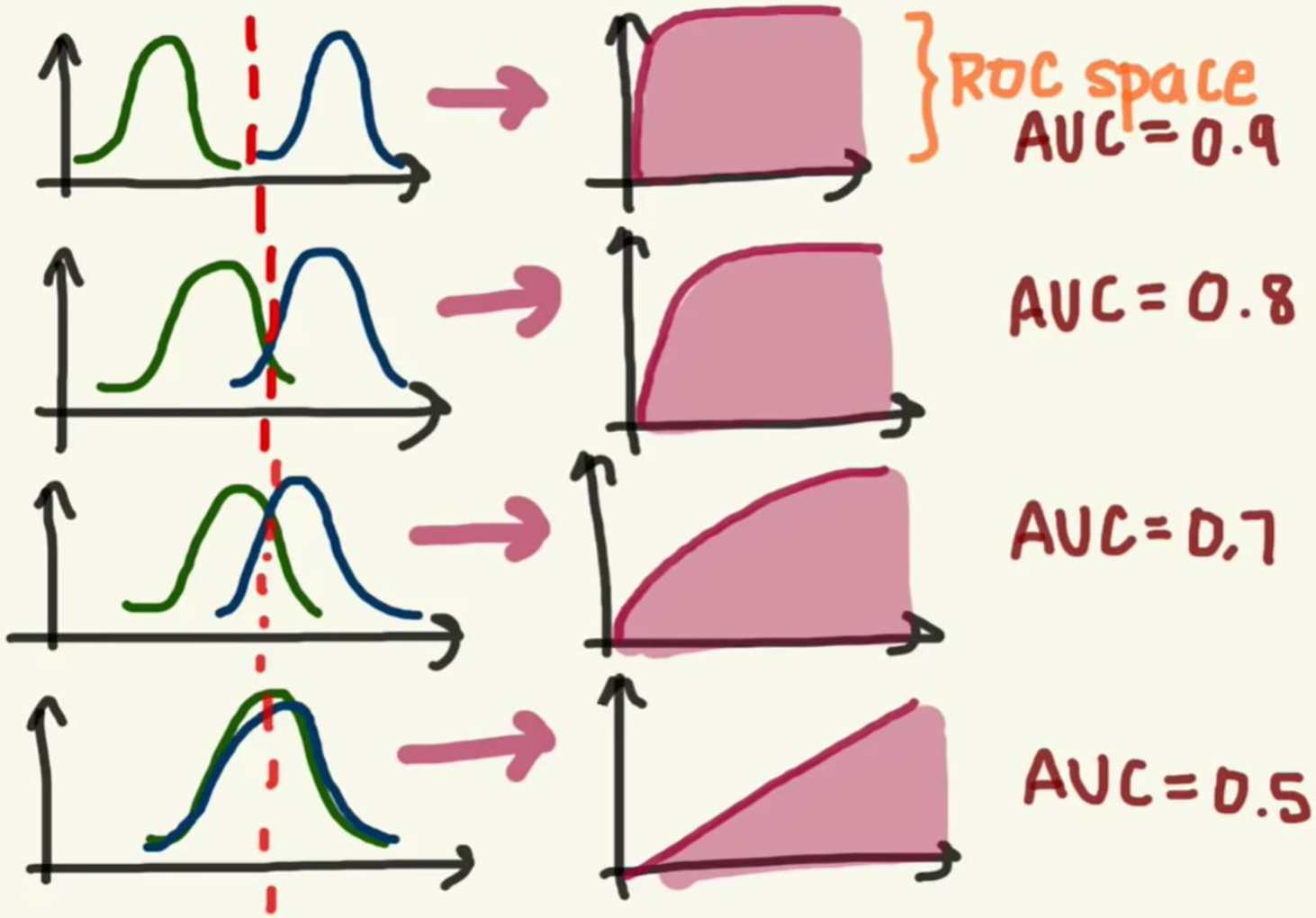
FPR = 1/10 = 0.1

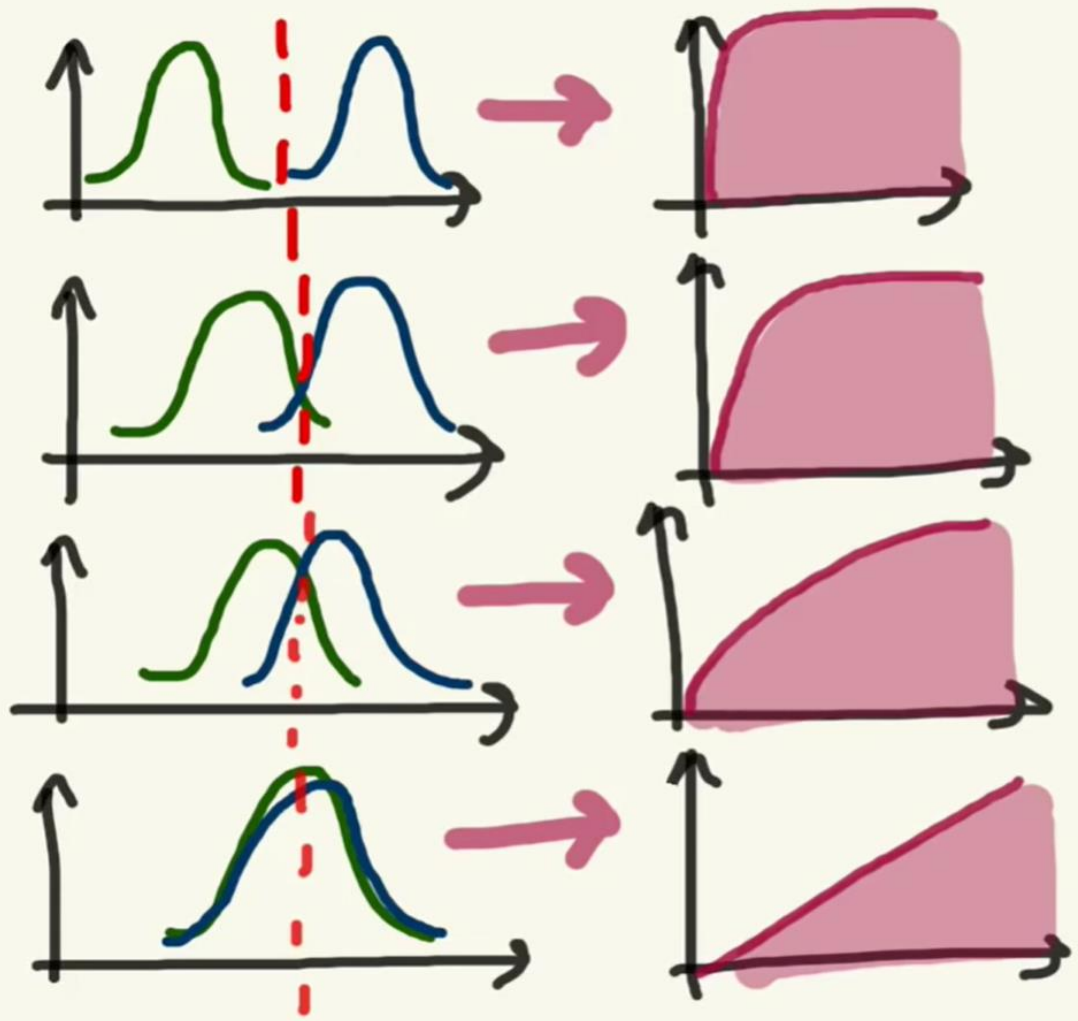
TPR = 5/10 = 0.5


(0.1,0.5)
True positive rate
False positive rate
Infinity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
.1
.2
.3
.4
.5
.6
.7
.8
.9
.30
.33
.34
.35
.36
.37
.38
.39
.4
.505
.51
.52
.53
.54
.55

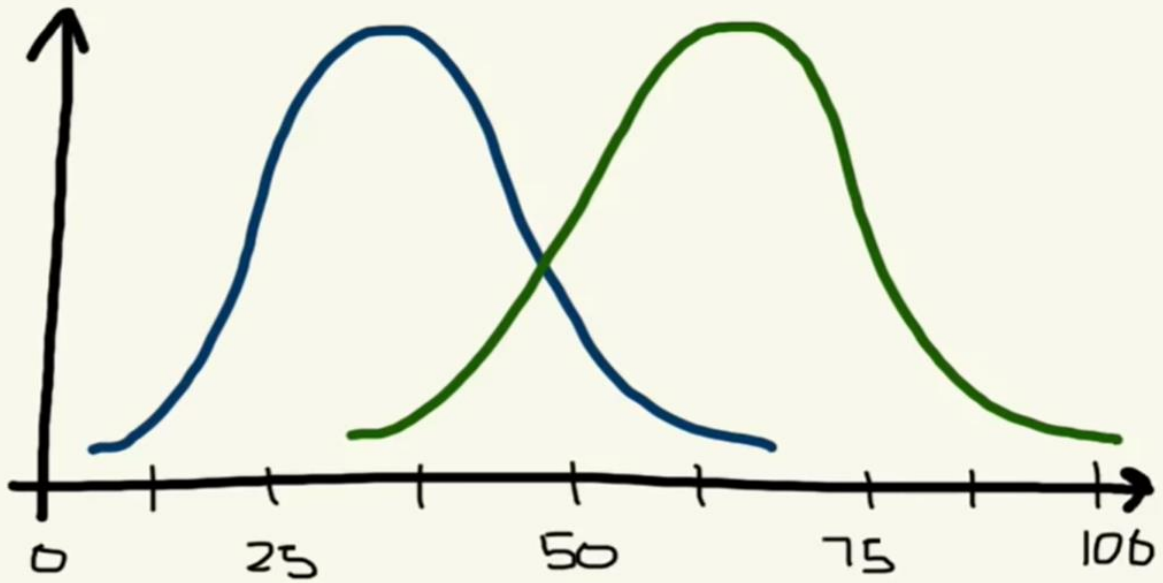




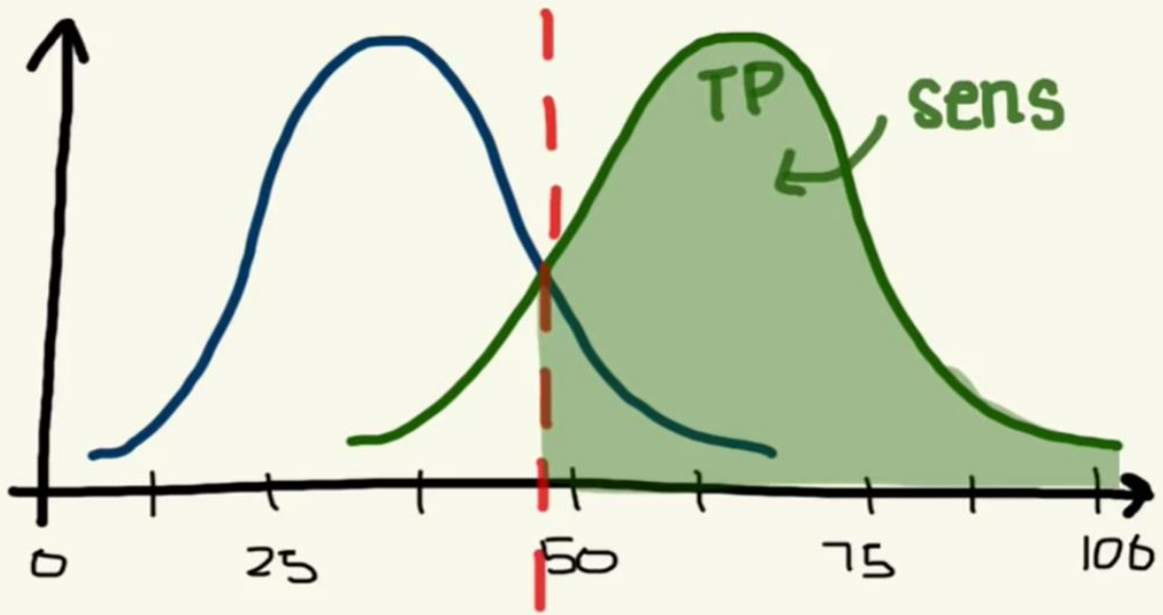

ROC space
AUC = 0.9
AUC = 0.8
AUC = 0.7
AUC = 0.5



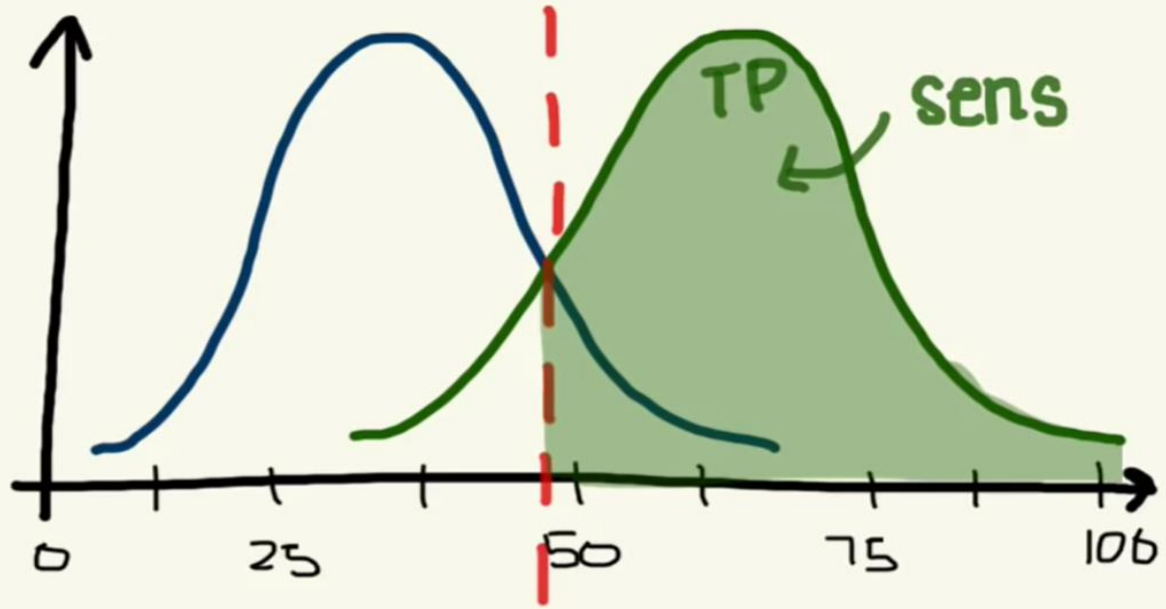
AUC	Quality Of Test
0.9-1	Excellent
0.8-0.9	Good
0.7-0.8	Fair
0.6-0.7	Poor
0.5-0.6	Fail



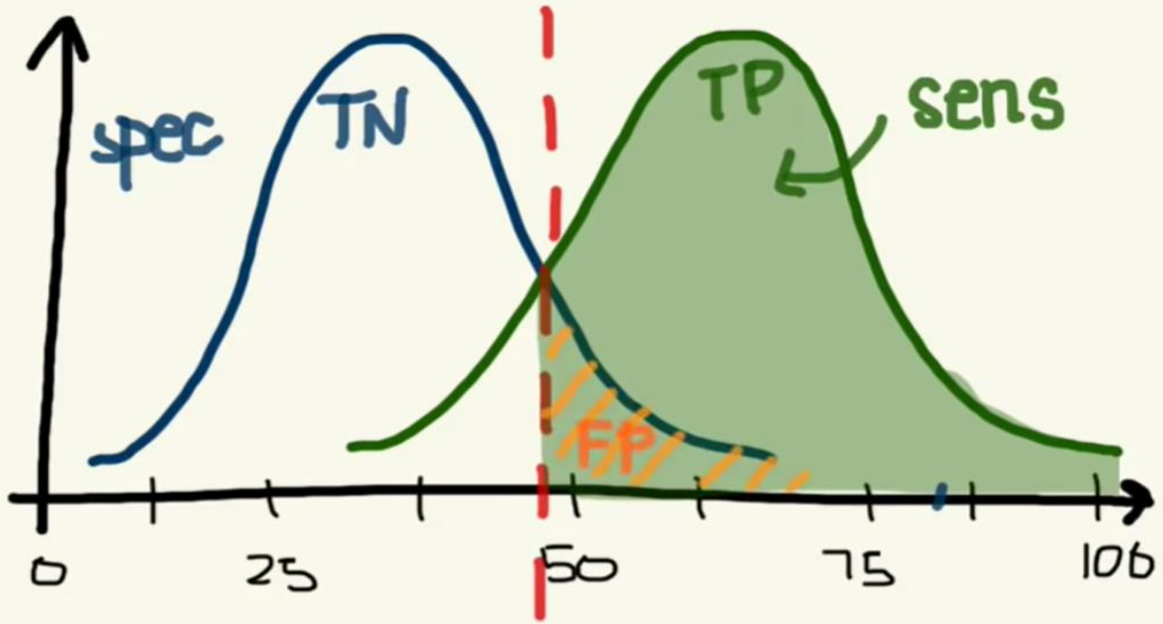
"Why (1-spec)?"



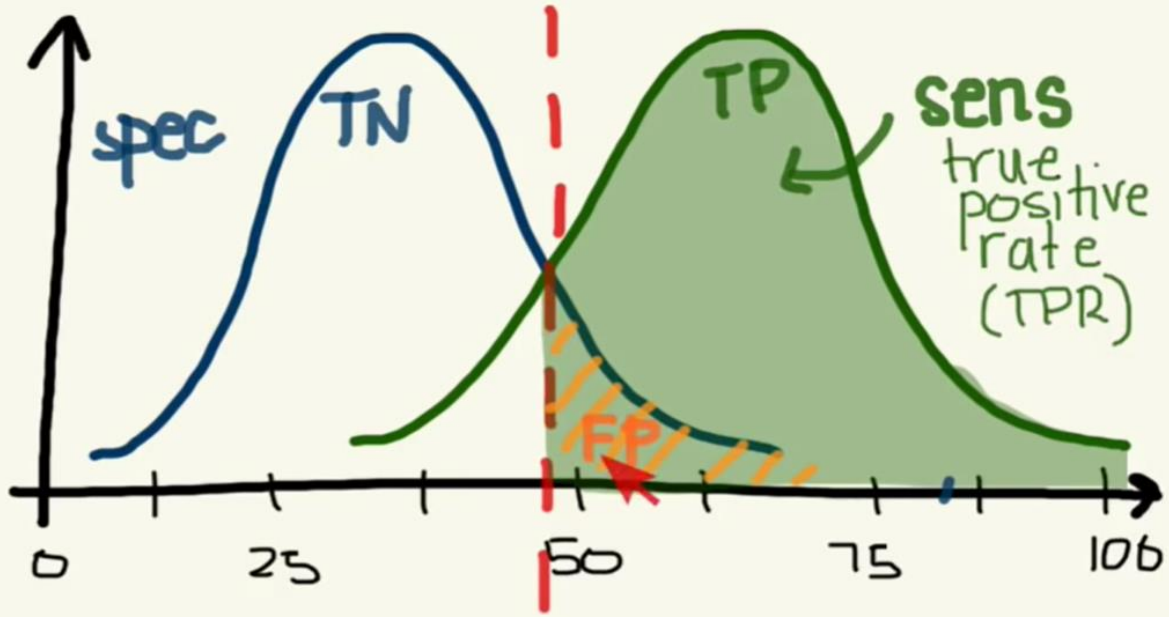
"why (1-spec)?"



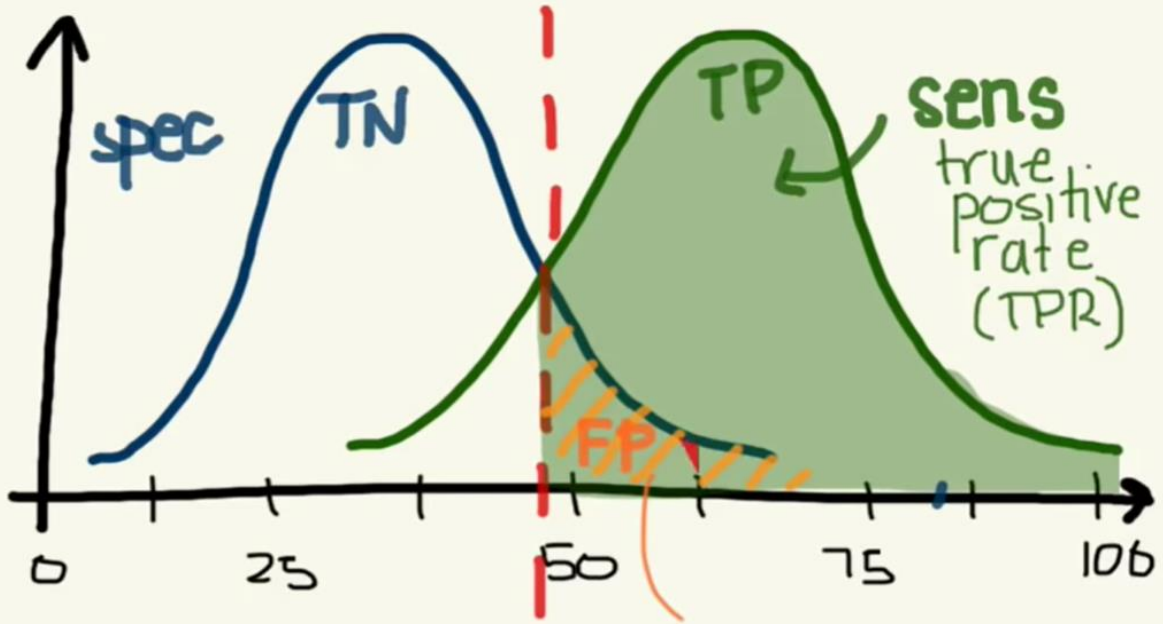
"why (1-spec)?"



"Why (1-spec)?"

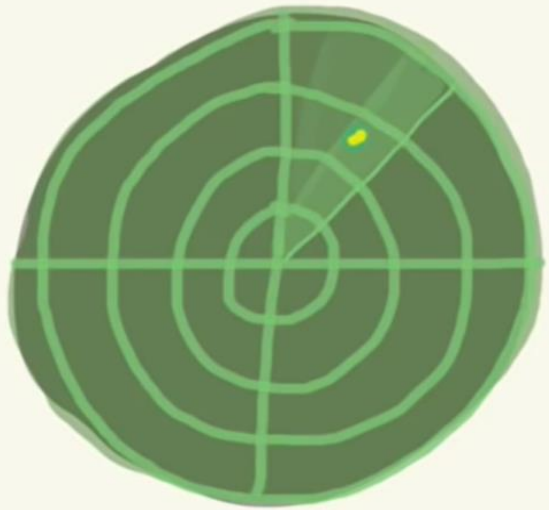


"Why (1-spec)?"


spec
TN
TP
sens
true positive rate (TPR)
FP
0
25
50
75
106
(1-spec)
false positive rate

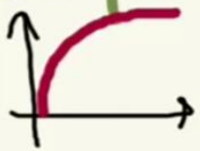
"Why (1-spec)?"

WW II - Signal Detection Theory



true positive rate

false positive rate



ROC

radar receiver

receiver operator