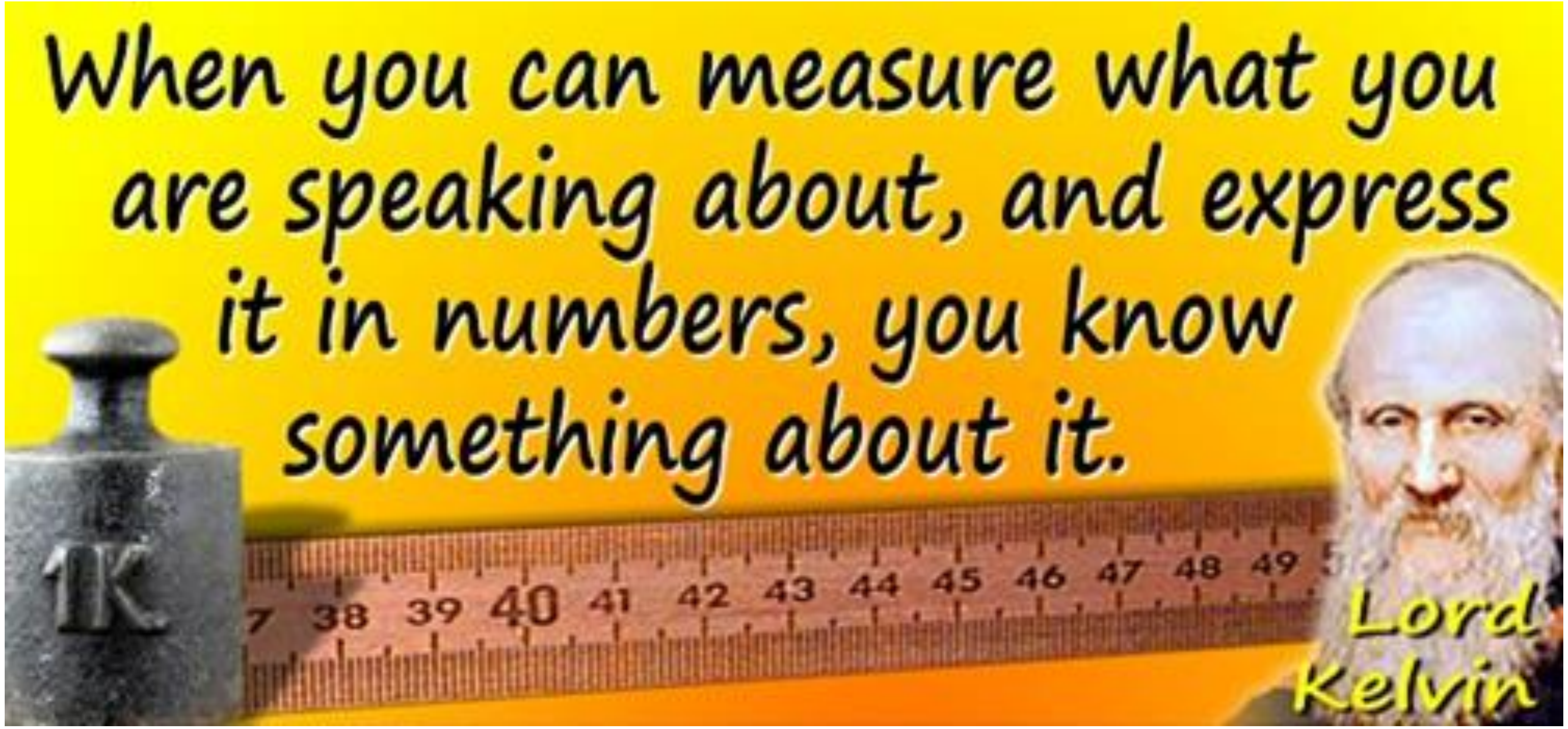
Focus for this lecture

Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- Performance Evaluation Metrics
  - Accuracy Metrics
  - Precision and Recall
  - F-measure (F1 Score)
- ROC and PR curves

Accuracy, Precision And Recall

Performance Evaluation Metrics


When you can measure what you are speaking about, and express it in numbers, you know something about it.
1K
Lord Kelvin

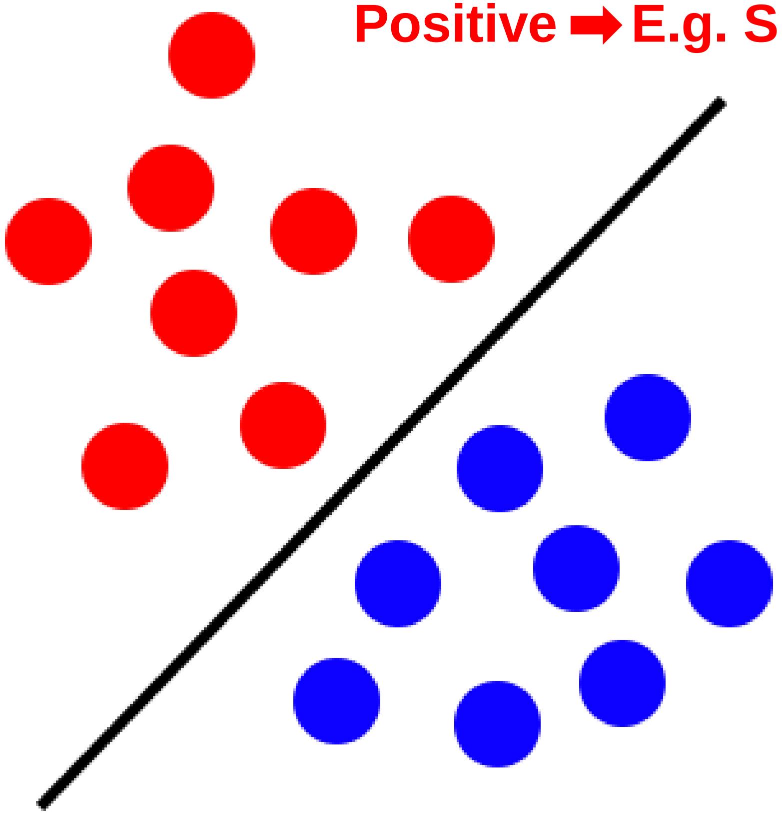
Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- Accuracy Metrics

Accuracy Metrics

Revisiting Binary case...

Positive ➡ E.g. Spam = YES



Negative➡E.g. Spam = NO

Revisiting Binary case...

$$Accuracy = \frac{(100+50)}{165} = 0.91$$

$$Misclassification = \frac{(10+5)}{165} = 0.09$$

$$TruePositiveRate(TP) = \frac{(100)}{105} = 0.95$$

$$FalsePositiveRate(FP) = \frac{(10)}{60} = 0.17$$

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

Revisiting Binary case...

$$TrueNegativeRate(TN) = \frac{(50)}{60} = 0.833$$

$$FalseNegativeRate(FN) = \frac{(5)}{105} = 0.048$$

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

Key accuracy measures and terminologies

- Classification Error = $\dfrac{errors}{total}$

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

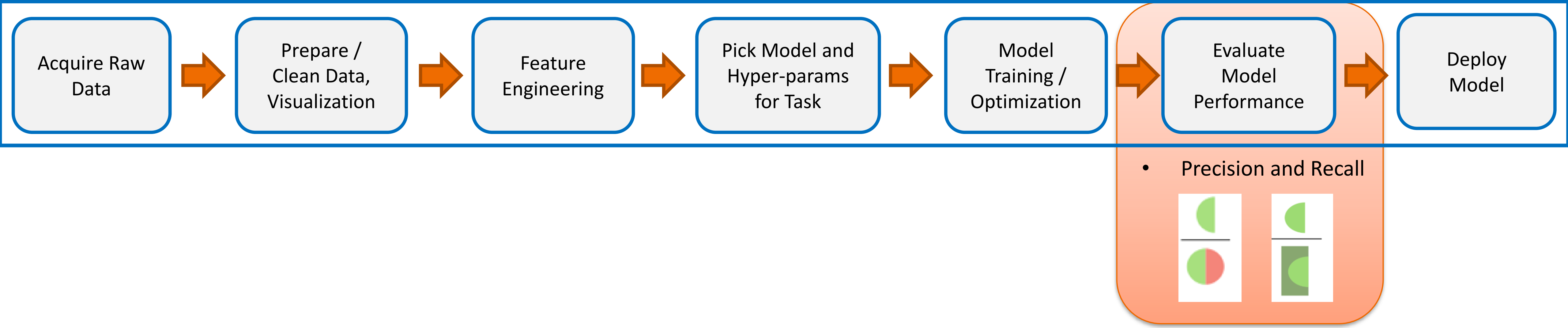
- Accuracy = 1 - Error = $\dfrac{correct}{total}$

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

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

Revisiting scenarios where metrics are appropriate

- When you do cancer screening what do you care?
  - High TP and Low FN
- When you classify between “apple” and “orange”
  - High Accuracy
- Automatic Firing on detecting a violation.
  - Very low FP



Precision and Recall

Precision and Recall

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

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

Precision and Recall – examples

- **Cancer-Prediction System**
- **Pool of 100 patients' data**
- 3 patients are selected for chemotherapy ; Rest (100-3=97) are declared healthy !
- 1 year later …
- 1 of them did not actually have cancer ! (FP)
- Precision = 2/(2+1) = 67%
- 3 from the 97 healthy declared ones have cancer (FN)
- Recall = 2/(2+3) = 40%
- Accuracy = (94+2)/100 = 96%

Key accuracy measures and terminologies

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

- n = # of patients who underwent a new cancer screening test
- Recall = Probability of the test result being +ve given that only cancer patients are examined

$$\frac{TP}{TP + FN}$$

- Precision = Probability of actually having cancer given the test result is +ve

$$\frac{TP}{TP + FP}$$

Precision and Recall – examples

- A system which needs to launch a missile at a terrorist hideout located in a dense urban area.
- Precision not 100% ➡ civilian casualties
- A system which needs to identify cancer-risk patients
- Recall not 100% ➡ some patients will die of cancer

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

Numerical problem: Precision and recall

- Suppose there are 6000 images of Amitabh Bachchan, ever, on the web. Suppose you fire an image search which is programmed to return 4000 images. Out of this you find 3000 are indeed Amitabh’s images. What the precision and recall in this case?

Solution: Precision and recall

$$Precision = \frac{TP}{(TP + FP)} \qquad Recall = \frac{TP}{(TP + FN)}$$

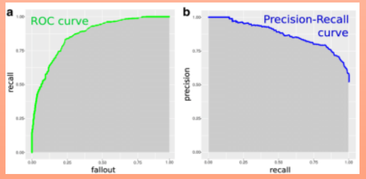
- Total images returned = 4000
- TP= All the images of Amitabh successfully returned =3000
- FP = Images returned that are not Amitabh = 4000-3000=1000
- FN =All the images of Amitabh not returned = 6000-3000 = 3000

$$Precision = \frac{3000}{3000 + 1000} = 0.75 \qquad Recall = \frac{3000}{3000 + 3000} = 0.5$$

Classifier Evaluation

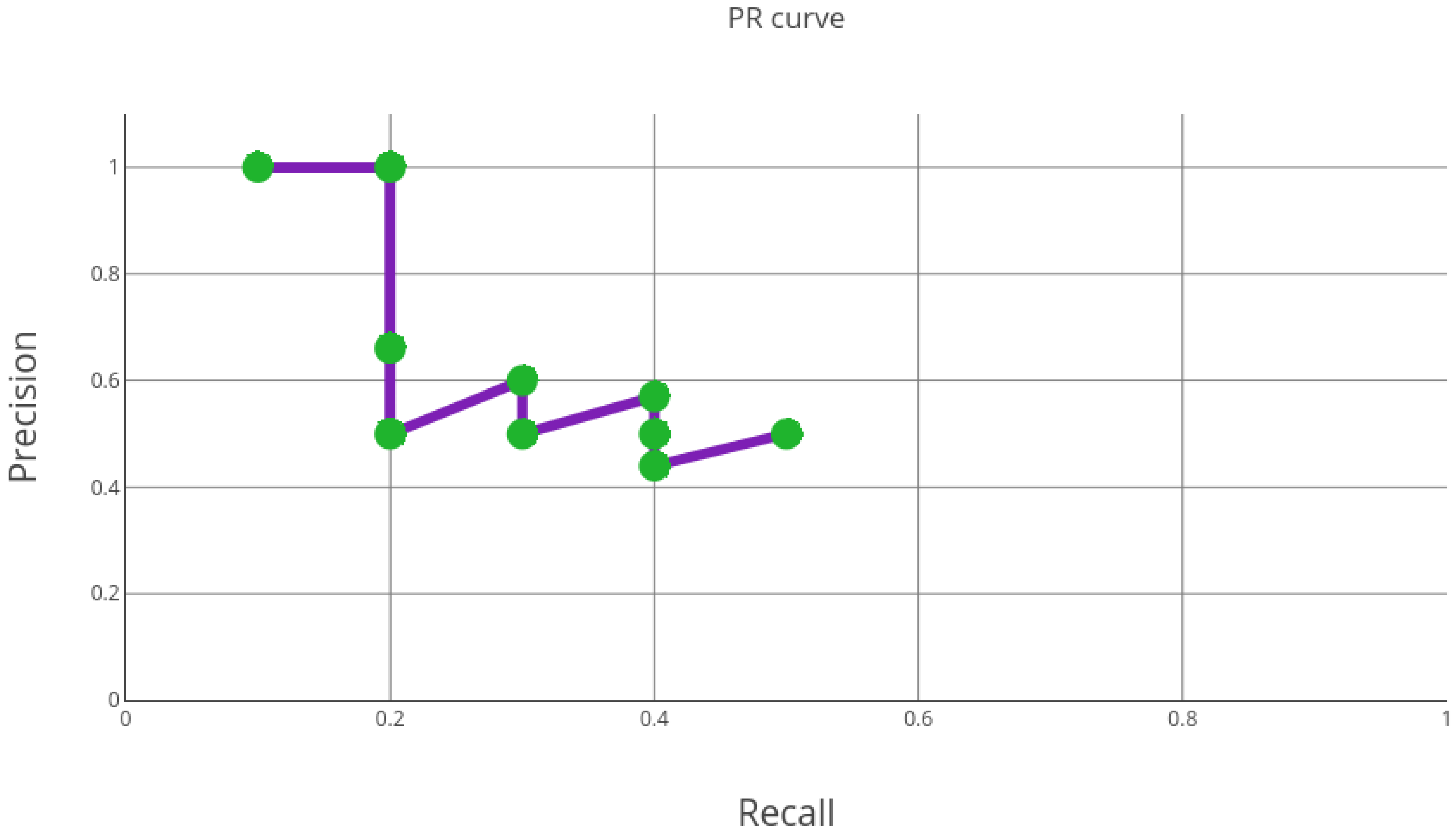
Acquire Raw Data → Prepare / Clean Data, Visualization → Feature Engineering → Pick Model and Hyper-params for Task → Model Training / Optimization → Evaluate Model Performance → Deploy Model

- ✓ Performance Evaluation Metrics
  - ✓ Accuracy Metrics
  - ✓ Precision and Recall
  - F-measure (F1 Score)
- ROC and PR curves

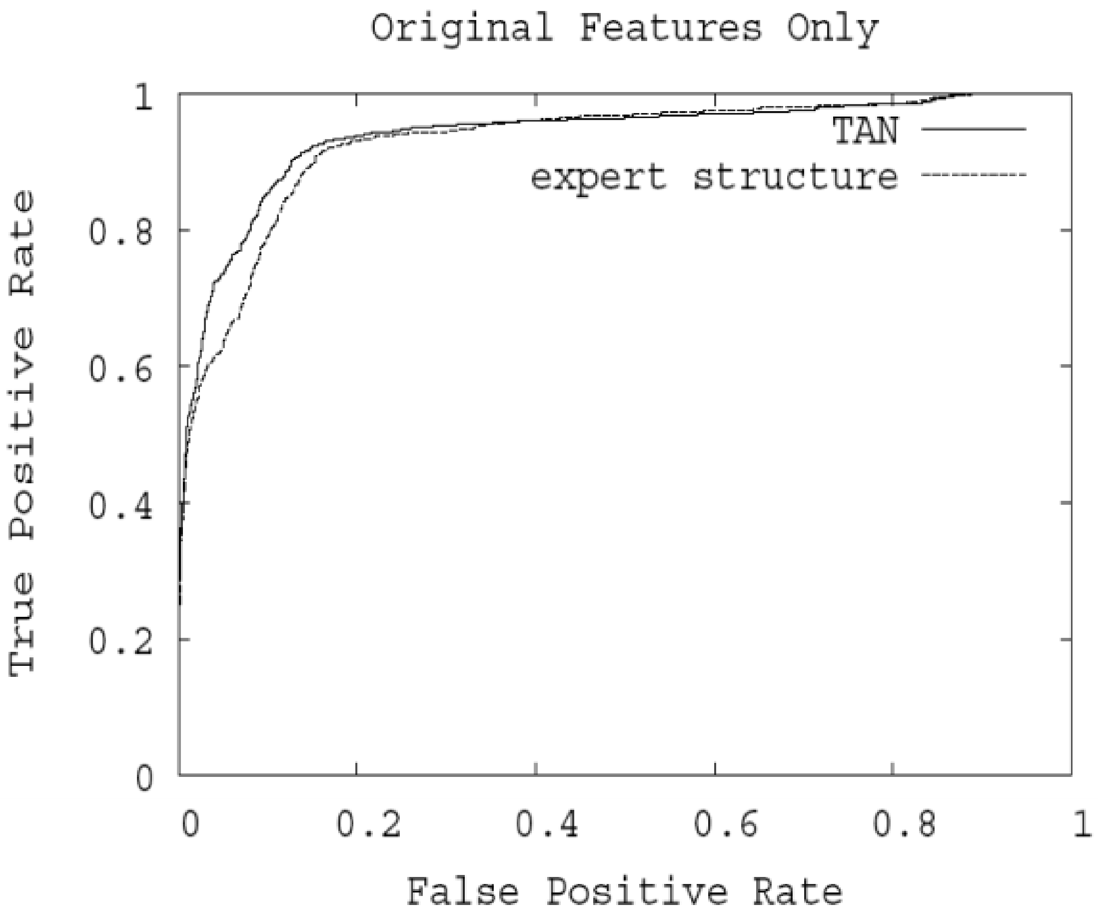


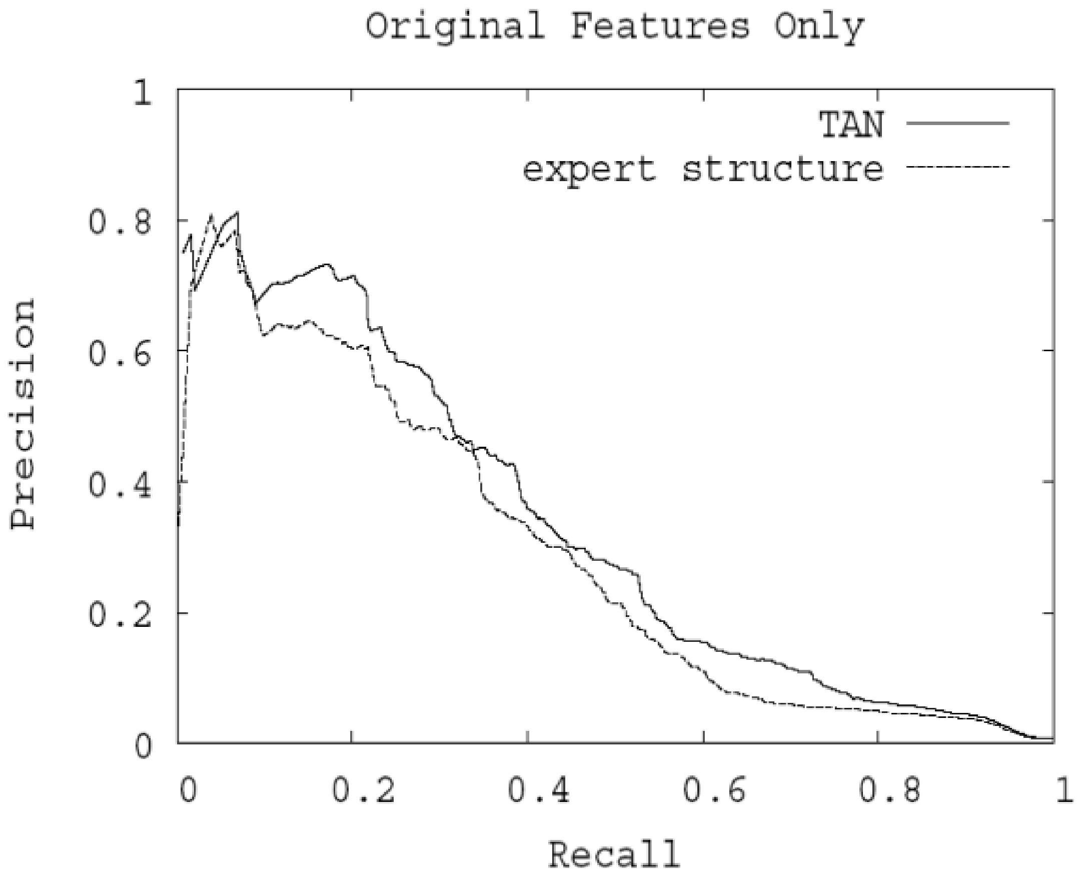
PR curves

Precision and Recall



ROC + PR Curves Example

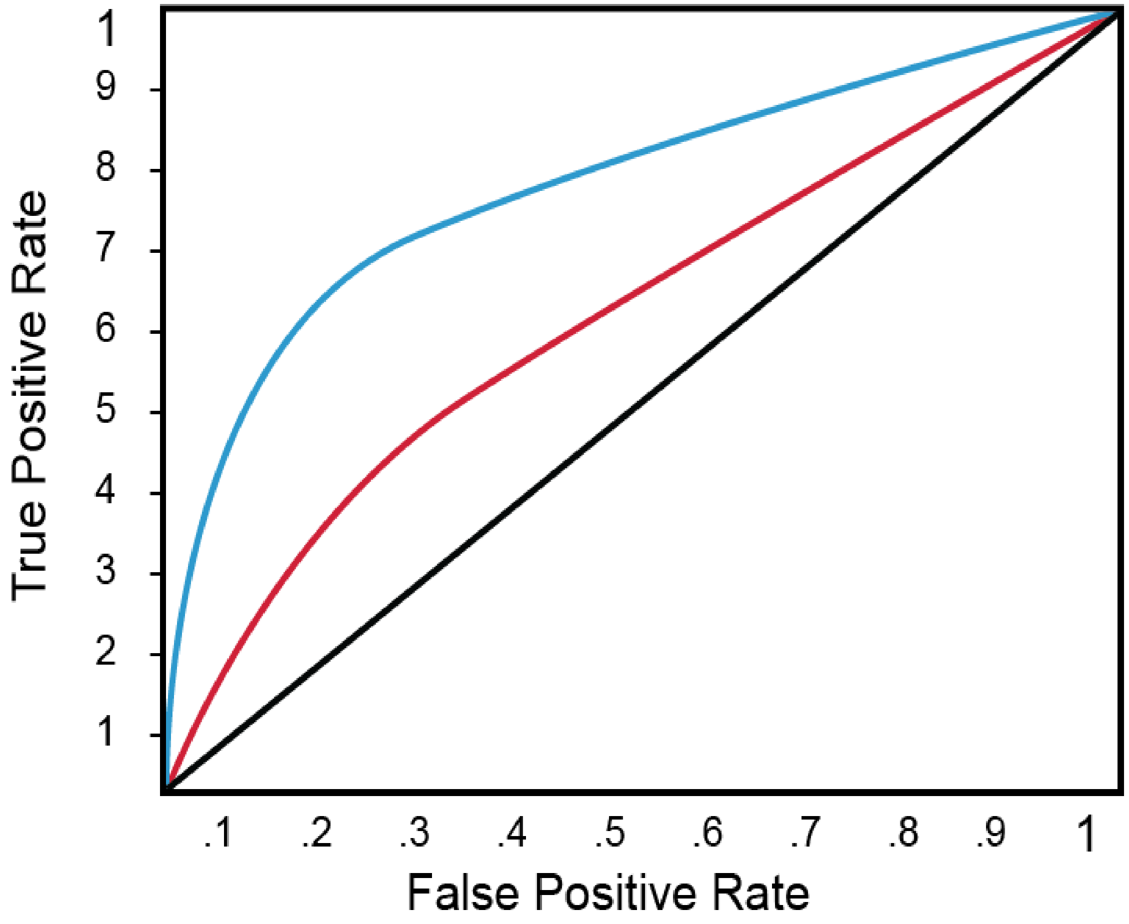




Trade Off...

- To compare two screening tests, at ROC(Receiver Operating Characteristics):

- The higher the Curve, the better.



F-measure: Combines Precision and Recall

- What to do when one classifier has better Precision but worse Recall, while other classifier behaves exactly opposite?
  - F-measure (Information Retrieval)

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

F-measure

$$Precision = \frac{TP}{(TP + FP)} \qquad Recall = \frac{TP}{(TP + FN)}$$

- What to do when one classifier has better Precision but worse Recall, while other classifier behaves exactly opposite?
  – F-measure (Information Retrieval)

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

- F1 measure punishes extreme values more !
- Definition of Recall and Precision have same numerator, different denominators. A sensible way to combine them is harmonic mean.

F-measure

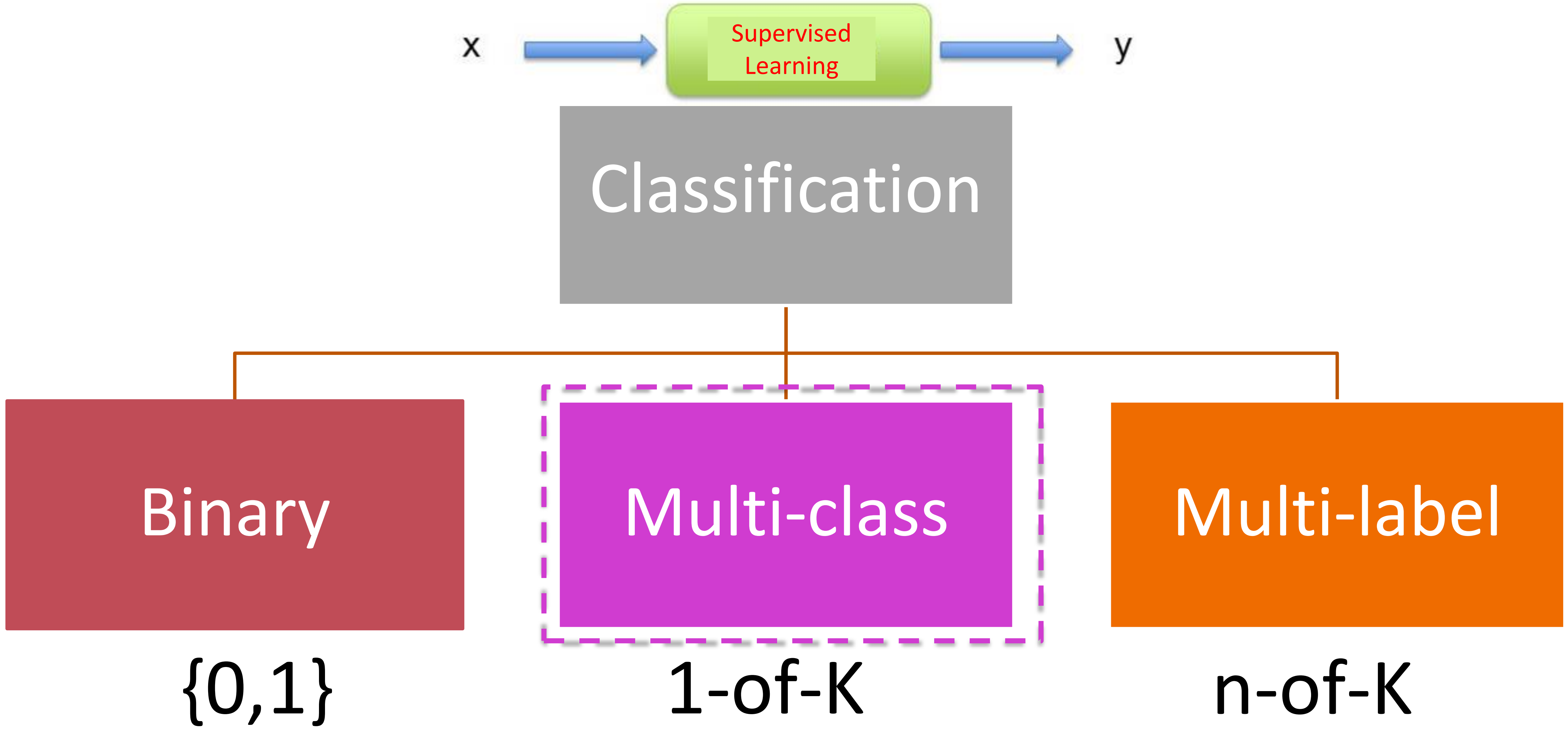
- Use when
  - FP and FN are ‘equally costly’
  - You don’t expect results to change when more data is added
  - TN is high (e.g. face detector)

$$Precision = \frac{TP}{(TP + FP)} \qquad Recall = \frac{TP}{(TP + FN)}$$

$$\frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

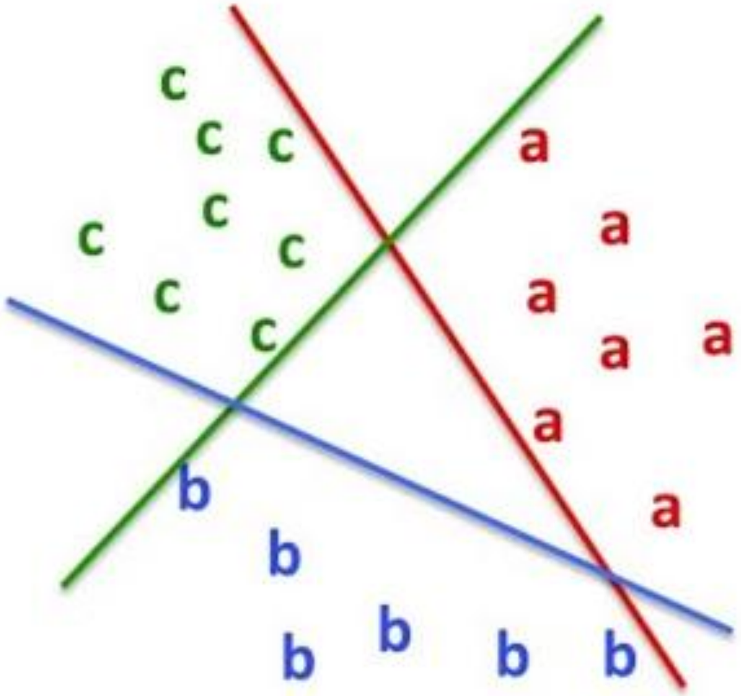
Utility and Cost

- Sometimes, there is a cost for each error
  - E.g. Earthquake prediction
    - False positive: Cost of preventive measures
    - False negative: Cost of recovery
- Detection Cost (Event detection) -Can be applied to example above
  - Cost = C_{FP} * FP + C_{FN} * FN


x
Supervised Learning
y
Classification
Binary
Multi-class
Multi-label
{0,1}
1-of-K
n-of-K

How to use 2-class measures for multi-class ?

- Convert into 2-class problem(s) !



Multi-class problems - Confusion matrix

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

activity recognition from video

actual class

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100	0	0	0	0	0	0	0	0	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0	0	89	0	0	0	11	0	0	0
pjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	89	0	11	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	0	0	100	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	67	33
wave2	0	0	0	0	0	0	0	0	0	100

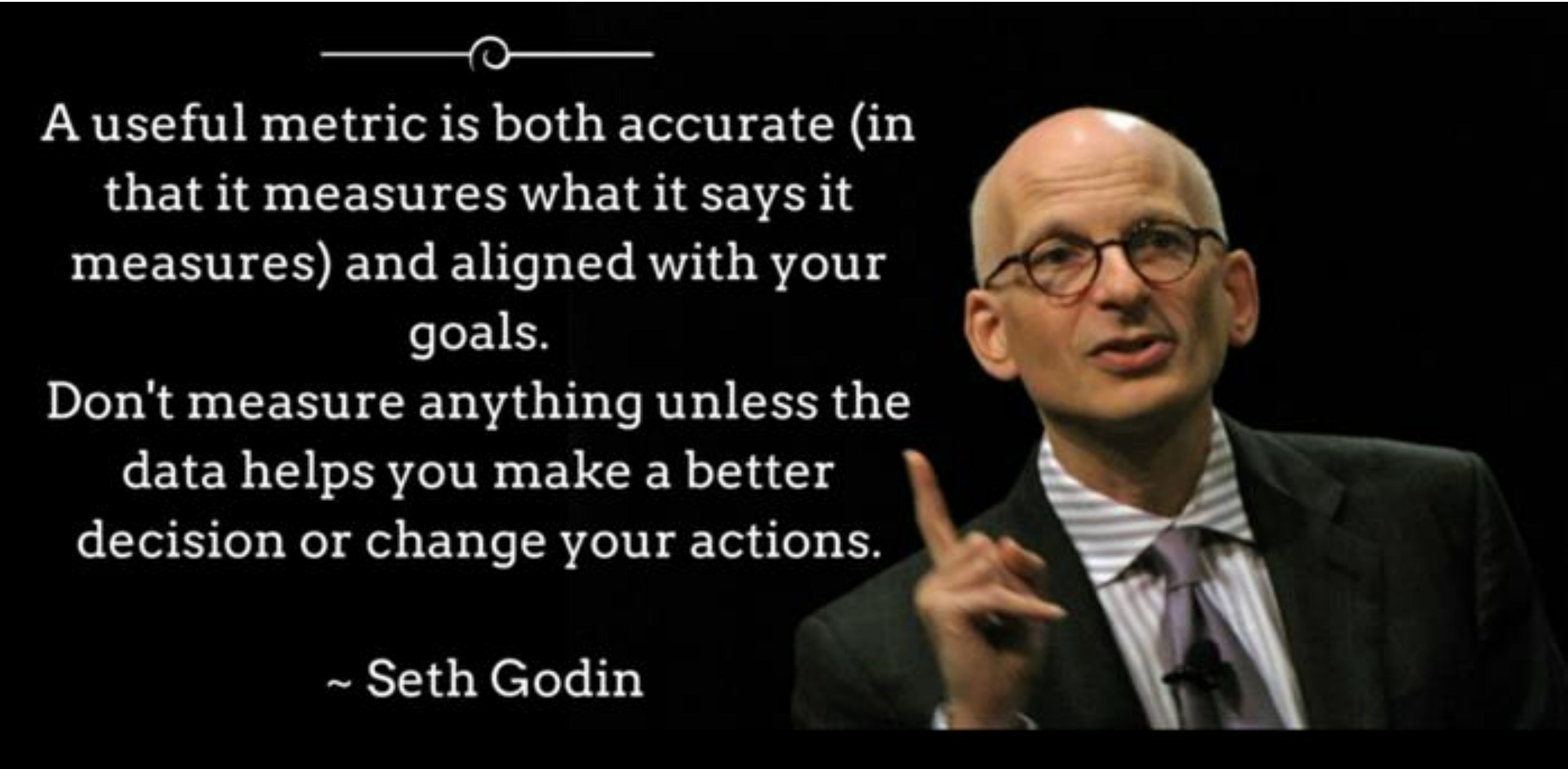
predicted class

Courtesy: vision.jhu.edu

Avg. accuracy may not be very meaningful with imbalanced class label distribution

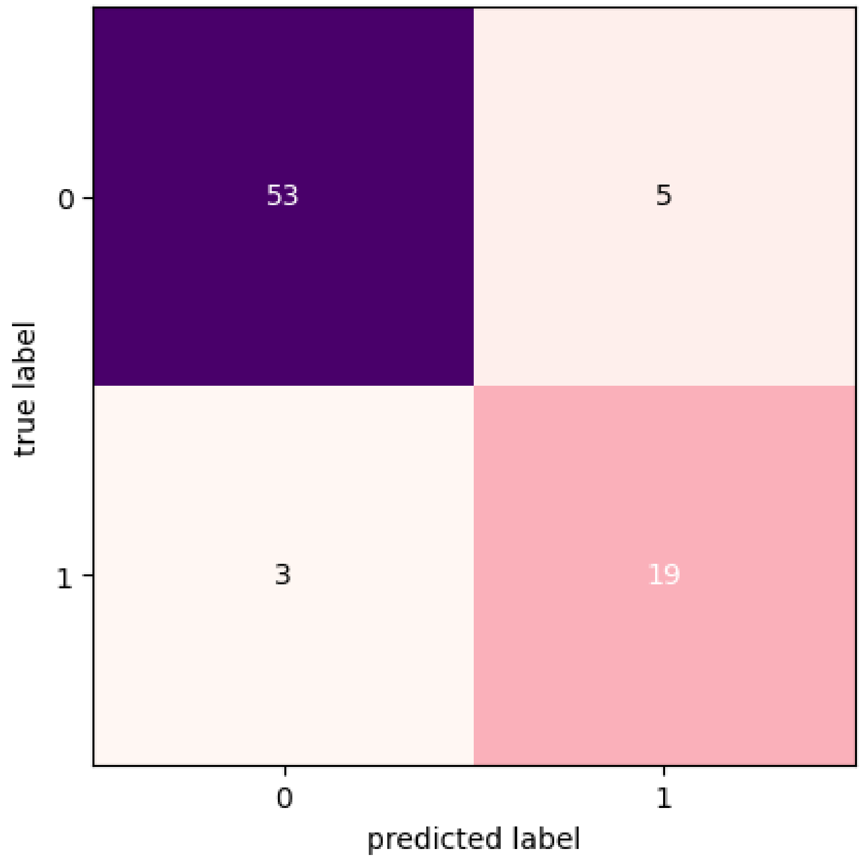
Summary

- Many metrics:
  - Accuracy, TP, FP, AUC, Precision, Recall, AP/mAP
  - Class imbalance and decision-cost imbalance must be taken into account
- Confusion Matrix: Important to analyse and refine solution.
- Curves provide “Trade off” and help compare classifiers / retrieval systems


A useful metric is both accurate (in that it measures what it says it measures) and aligned with your goals.
Don't measure anything unless the data helps you make a better decision or change your actions.
~ Seth Godin

Think ?

- The image given below shows a sample confusion matrix plot for a binary image classification problem. While predicting an input image using the trained model, the predicted label would be either 0 or 1. What percentage of samples are predicted correctly out of all the samples that are predicted as the class '0'?



Options:

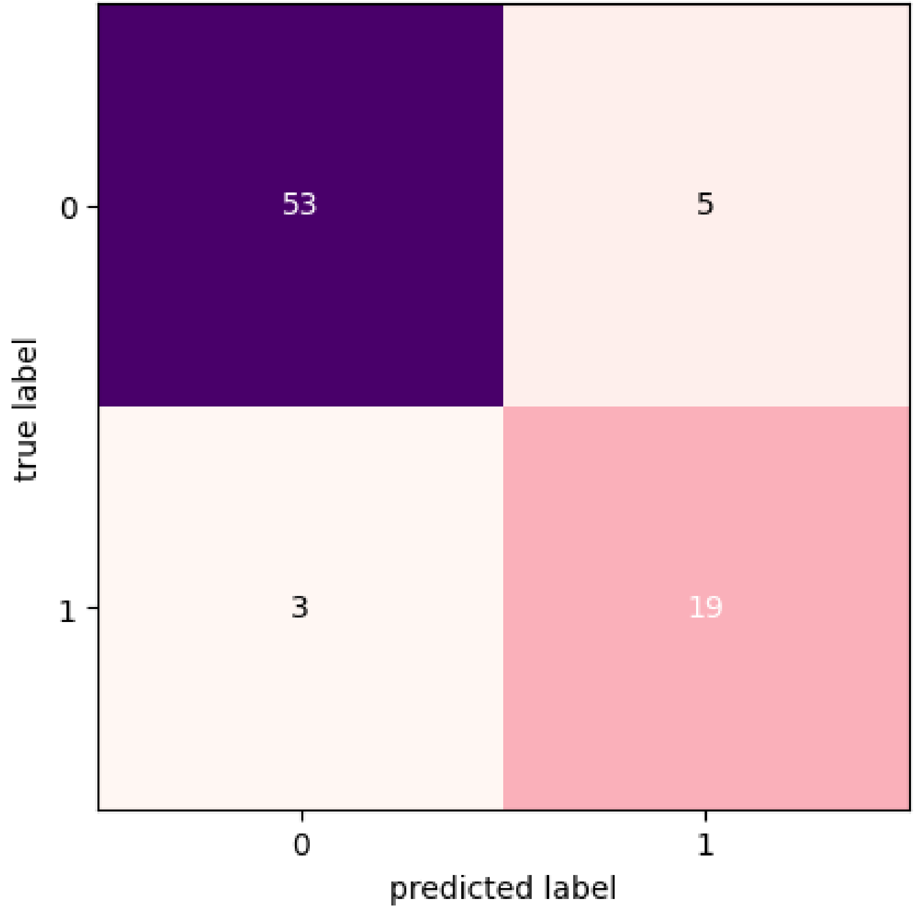
1.79%
2.95%
3.86%
4.91%

Answer : B. 95%

Explanation: The question is asking precision for negative class indirectly.

Expression for Precision: $\text{Precision} = \frac{\textbf{TruePositive}}{\textbf{TruePositive+FalsePositive}}$

And for Recall: $\text{Recall} = \frac{\textbf{TruePositive}}{\textbf{TruePositive+FalseNegative}}$



From the confusion matrix:

True positive count = 19

False positive count =5

False Negative count=3

True negative =53

Precision = 19/(19+5) → 79.1% ~ 79% for class 1 (Positive)

Precision = 53/(53+3) → 94.6% ~ 95% for class 0 (Negative)

Recall =19/(19+3) → 86.3% ~ 86 % for class 1 (Positive)

Recall =53/(53+5) → 91.3% ~91% for class 0 (Negative)

Thanks!!

Questions?