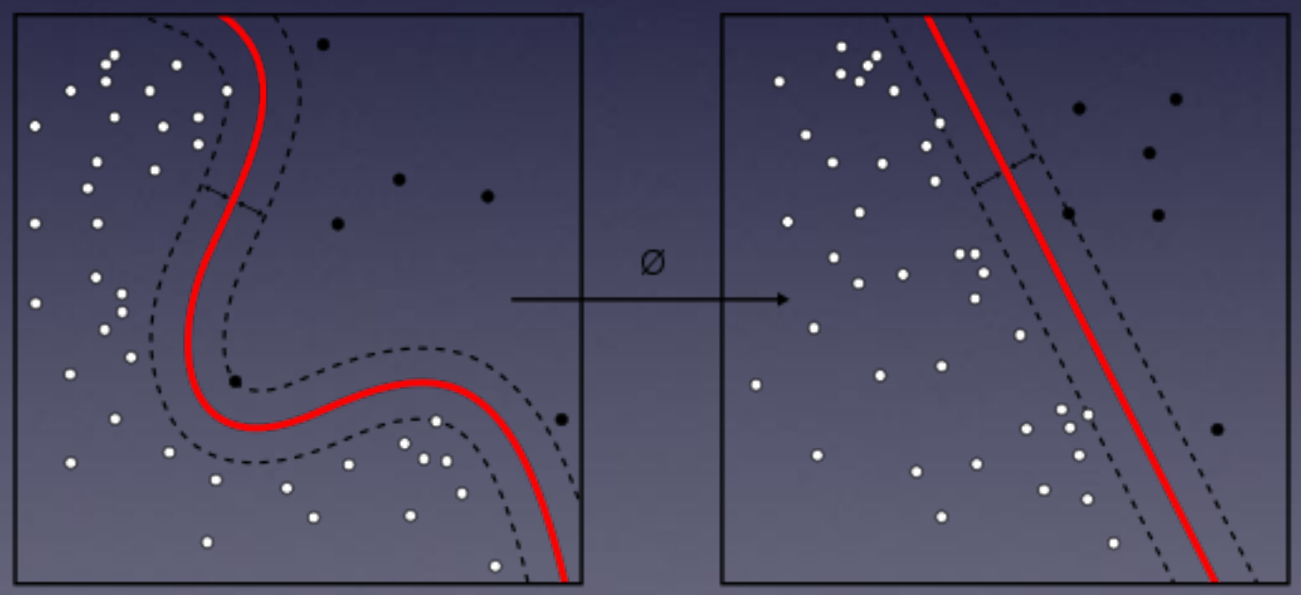
Data Science Workflow

Binary Classification

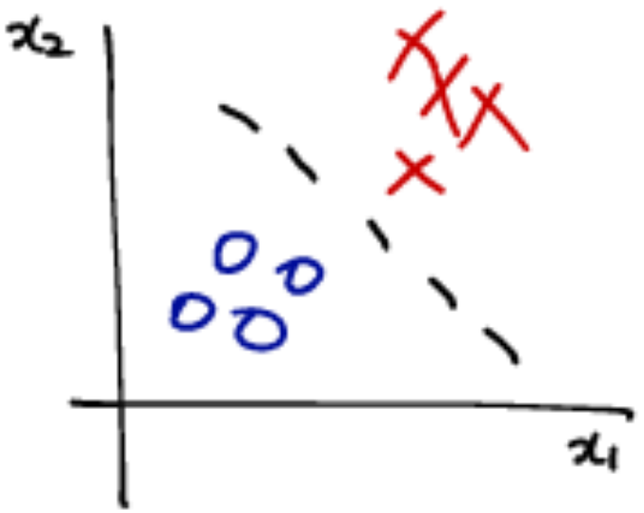
Parvin Shakibaei

What is a classifier

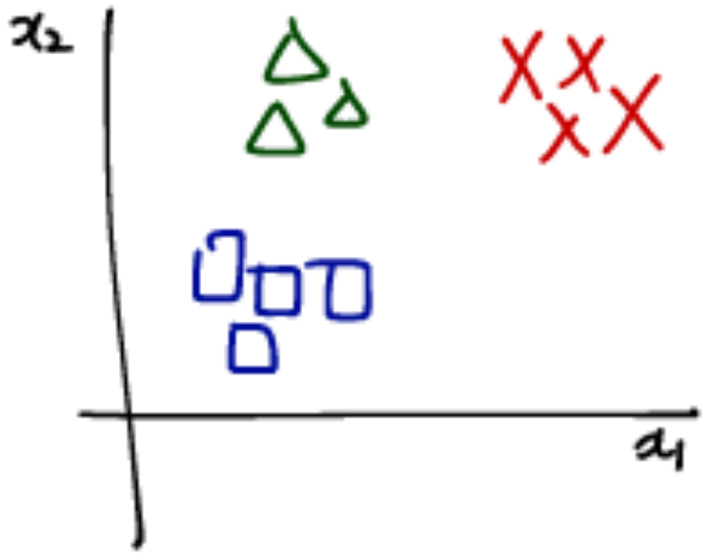
- In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-population) a new observation belongs



Binary classification

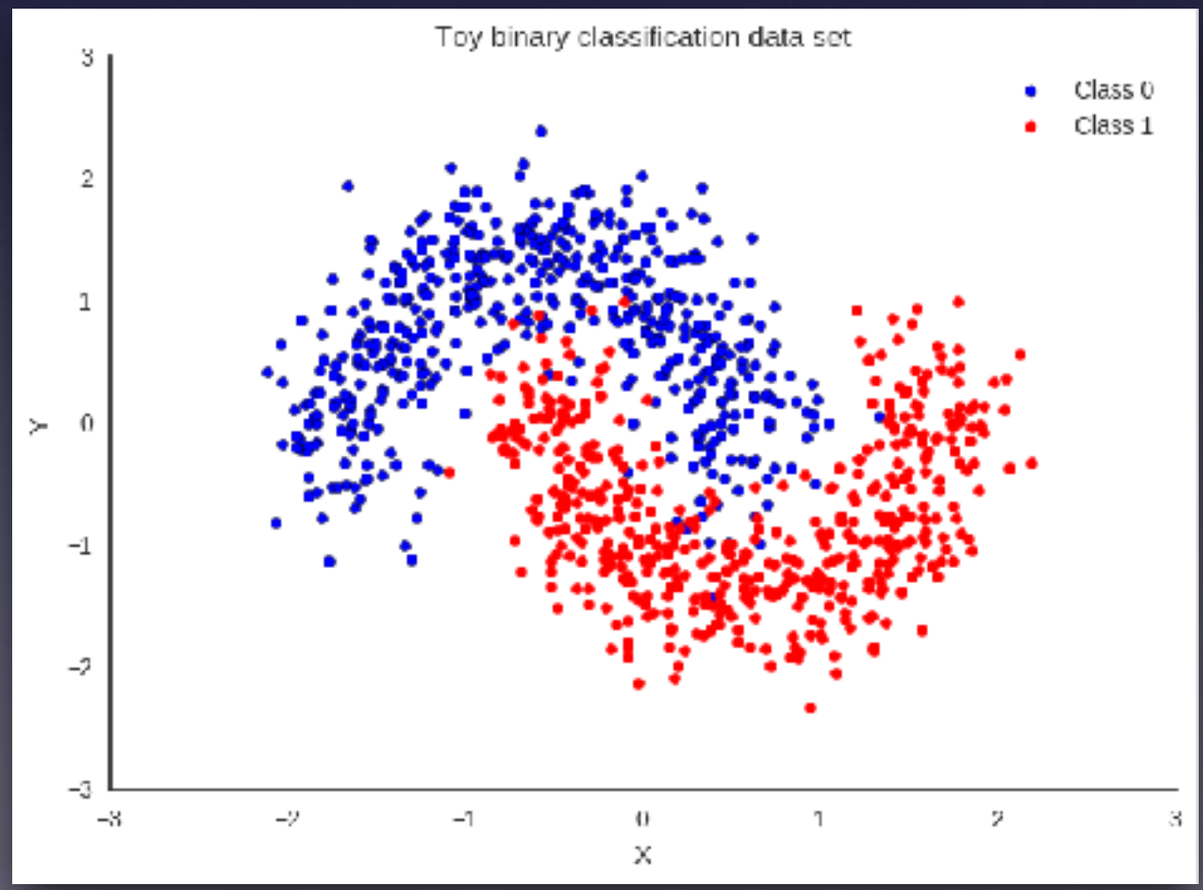


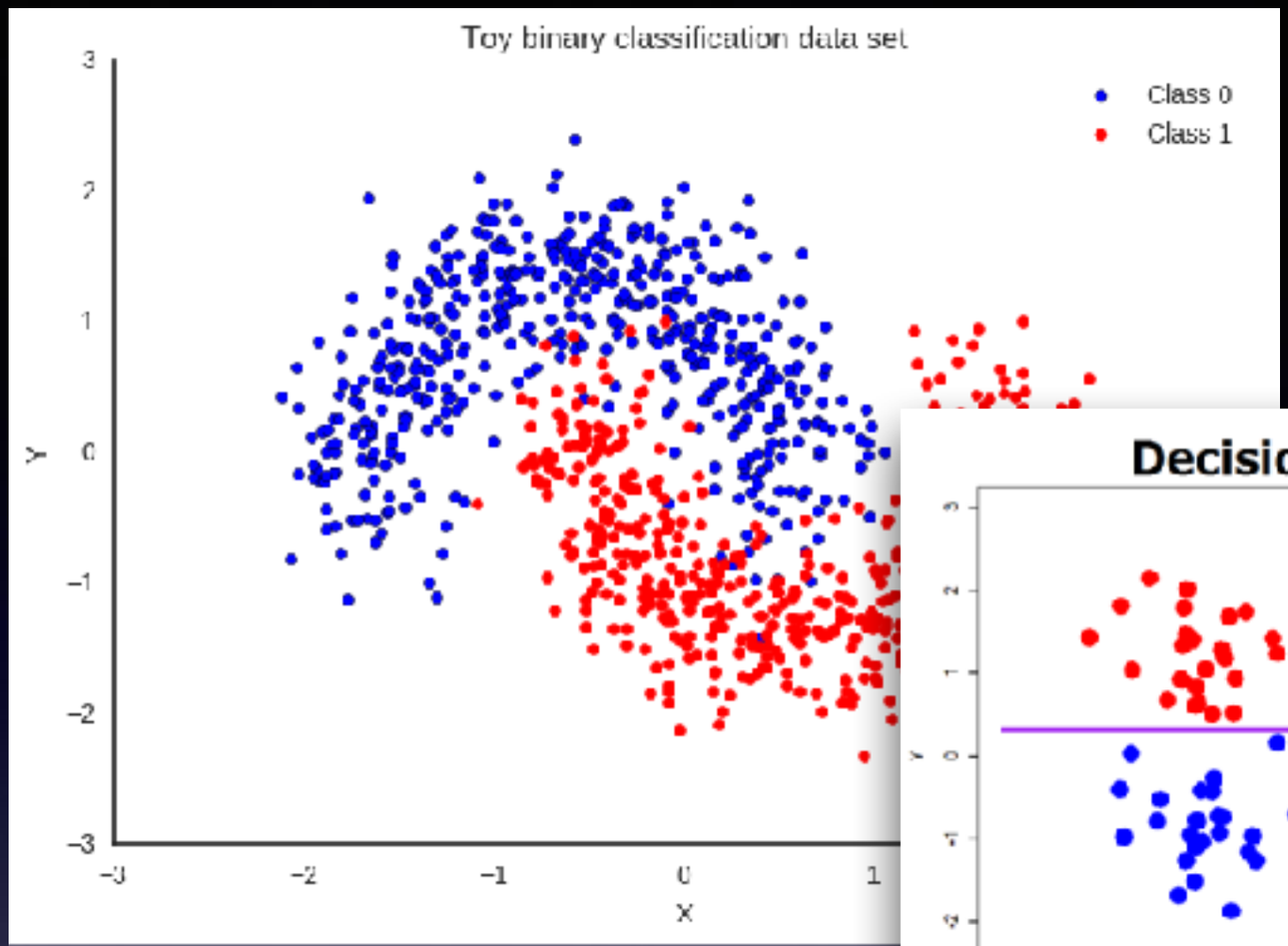
Multi-class classification

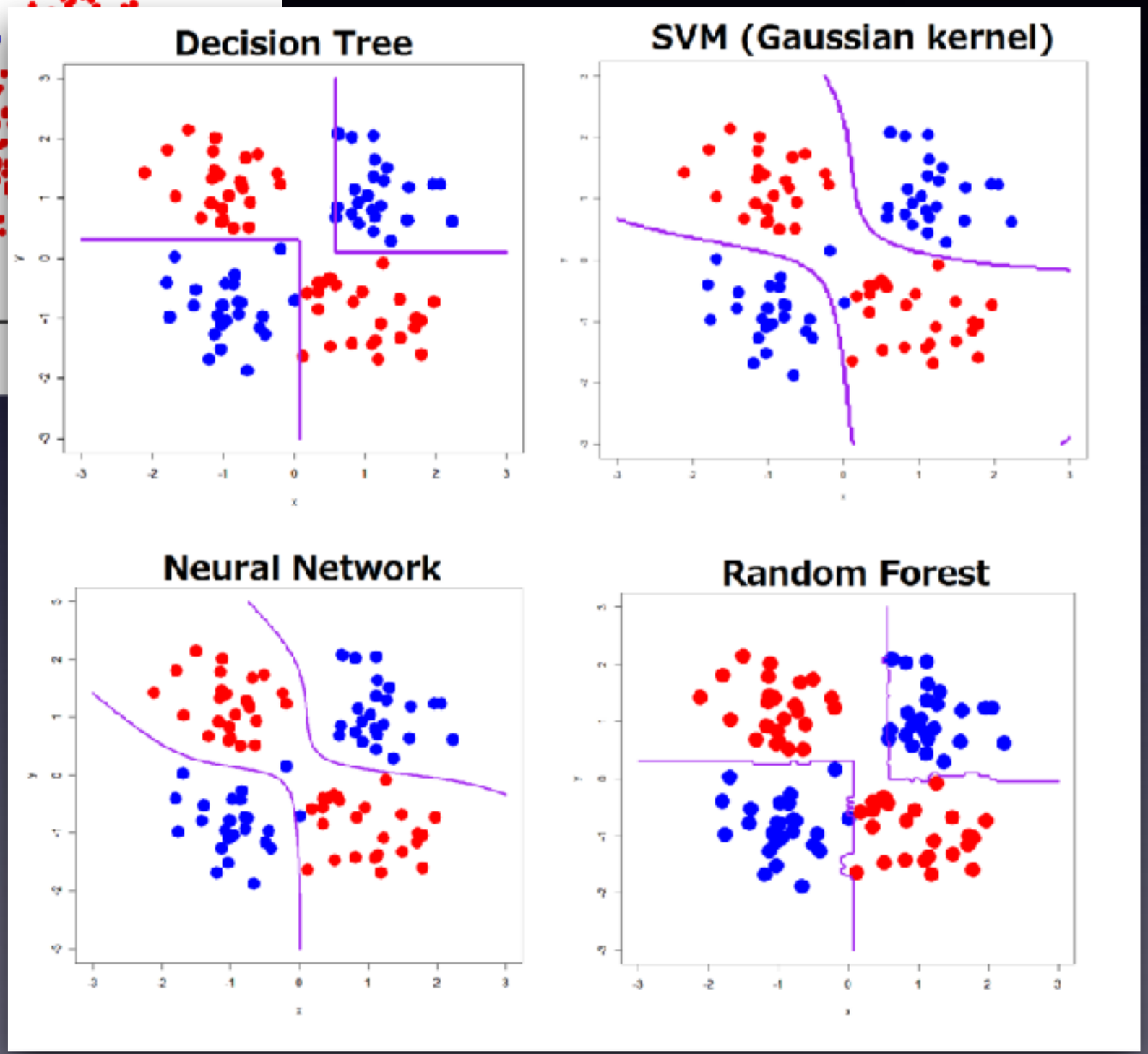


Binary Classifier

- Binary or binomial classification is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) on the basis of a classification rule.

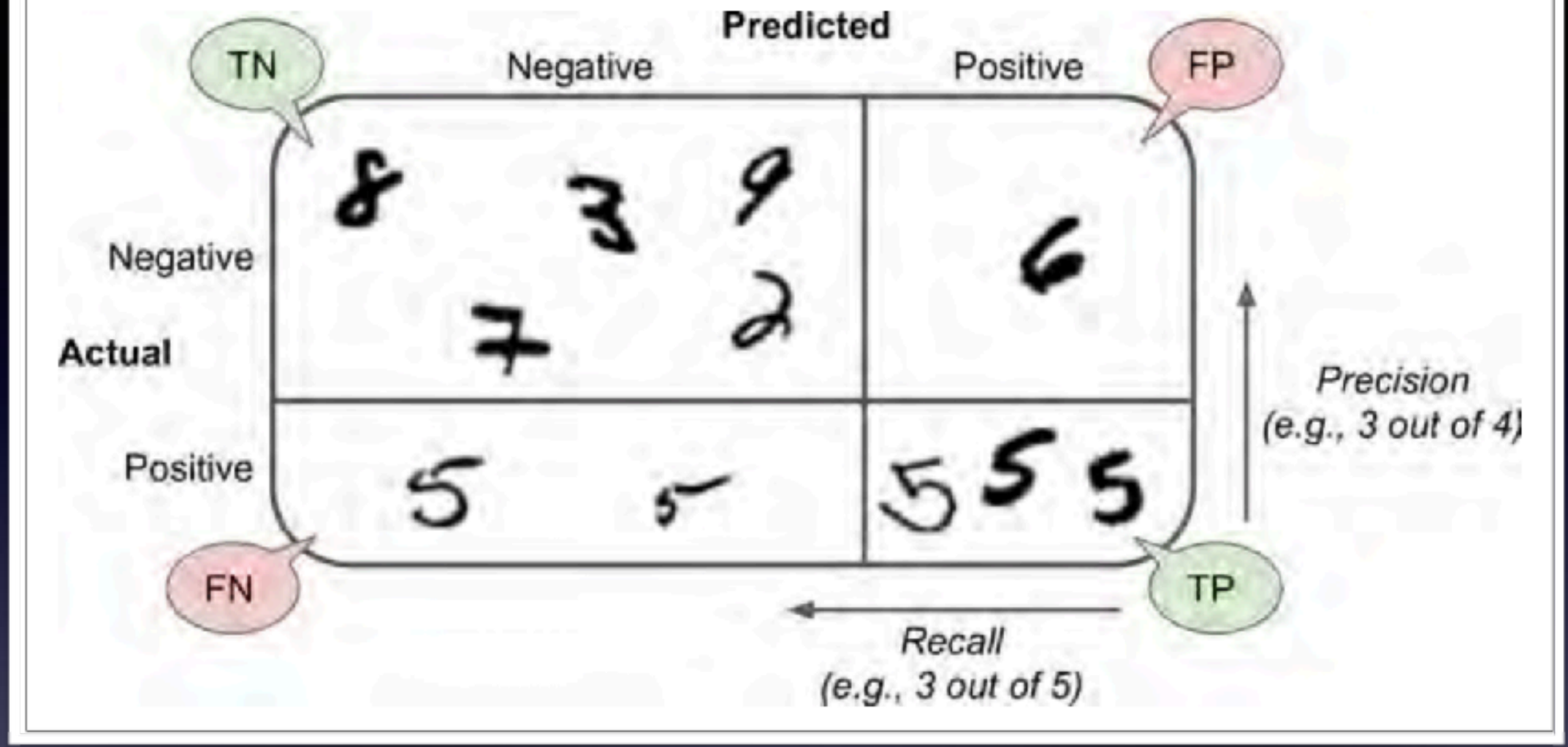



Toy binary classification data set
Class 0
Class 1
X
Y


Decision Tree
SVM (Gaussian kernel)
Neural Network
Random Forest

Confusion Matrix

		Actual Value (as confirmed by experiment) positives	negatives
Predicted Value (predicted by the test)	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative



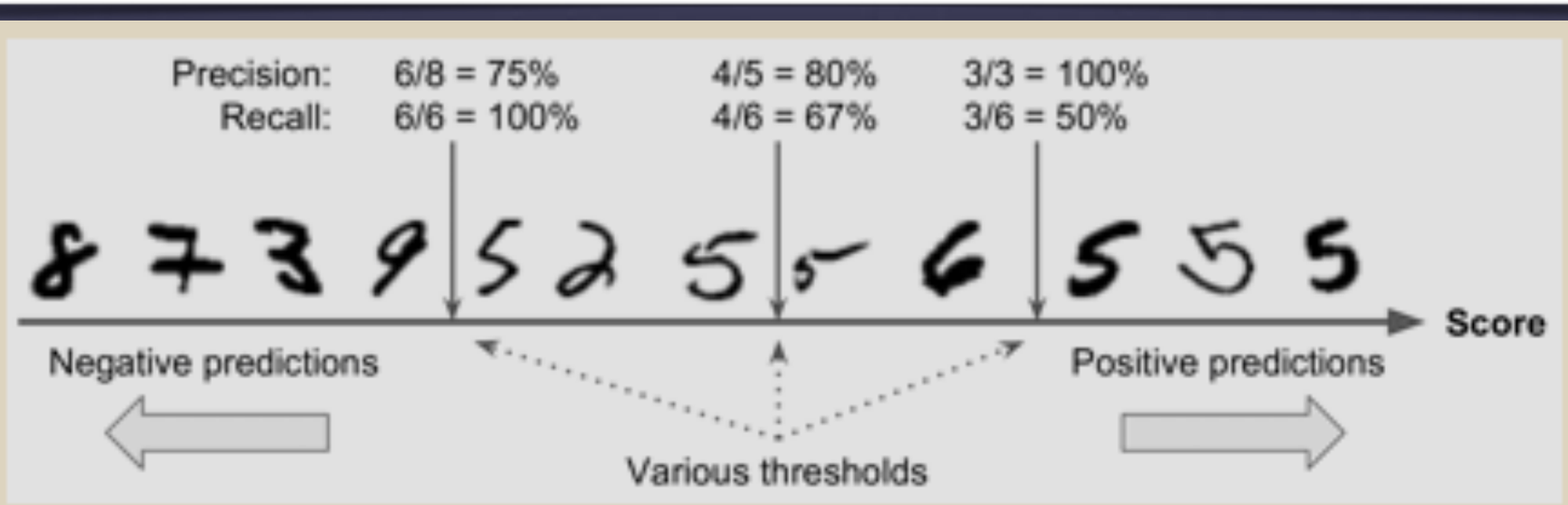
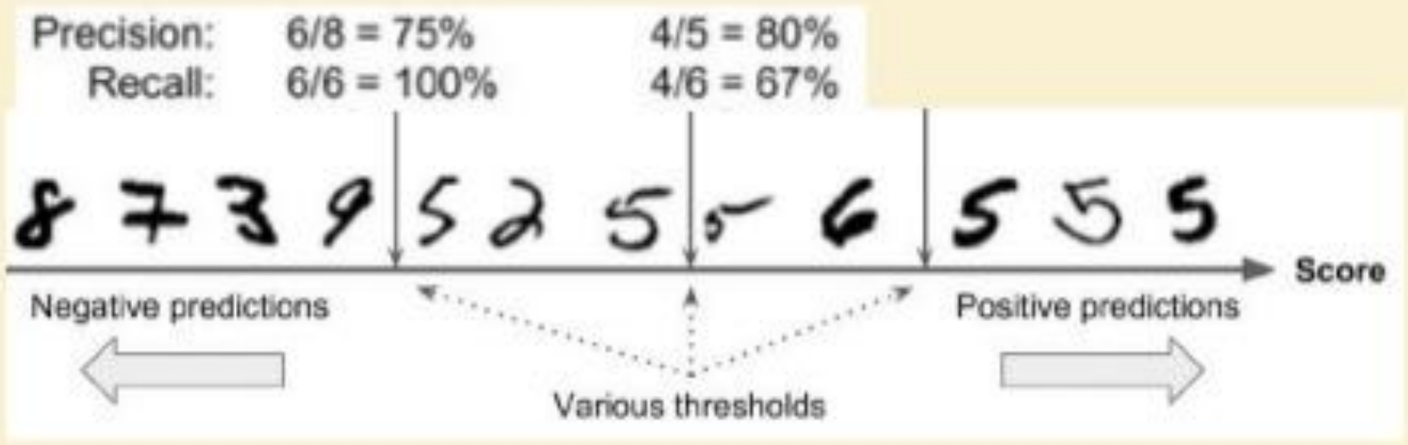


Figure 3-3. Decision threshold and precision/recall tradeoff

Precision / Recall Tradeoff



TN - TN - TN - TN - FN - TN - FN - TP - FP - TP - TP - TP

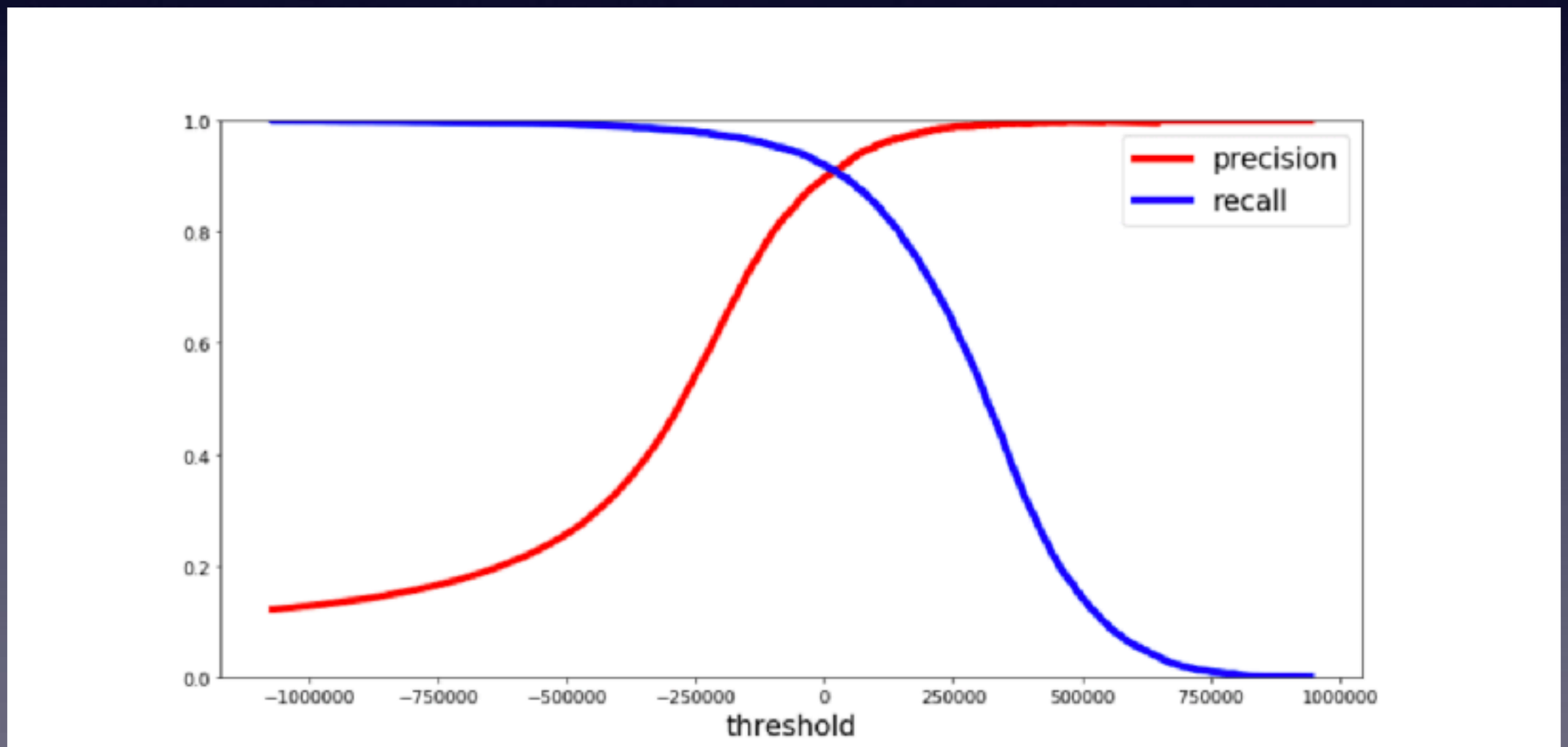
TN = 5, TP = 4, FN = 2, FP = 1

Precision = 4/ (4+1) = 80%
Recall = 4/ (4+2) = 67%

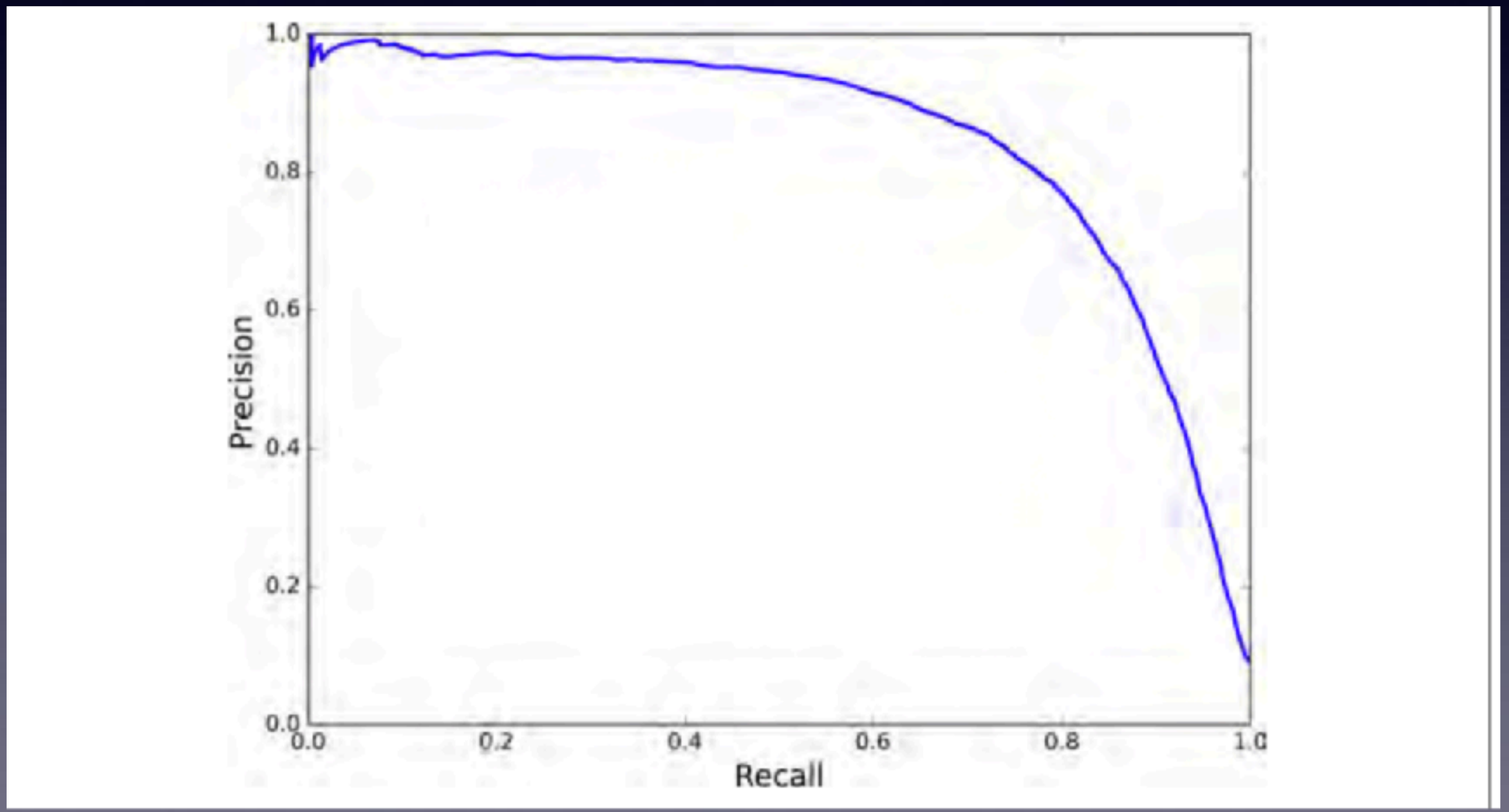
$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Precision and recall versus the decision threshold

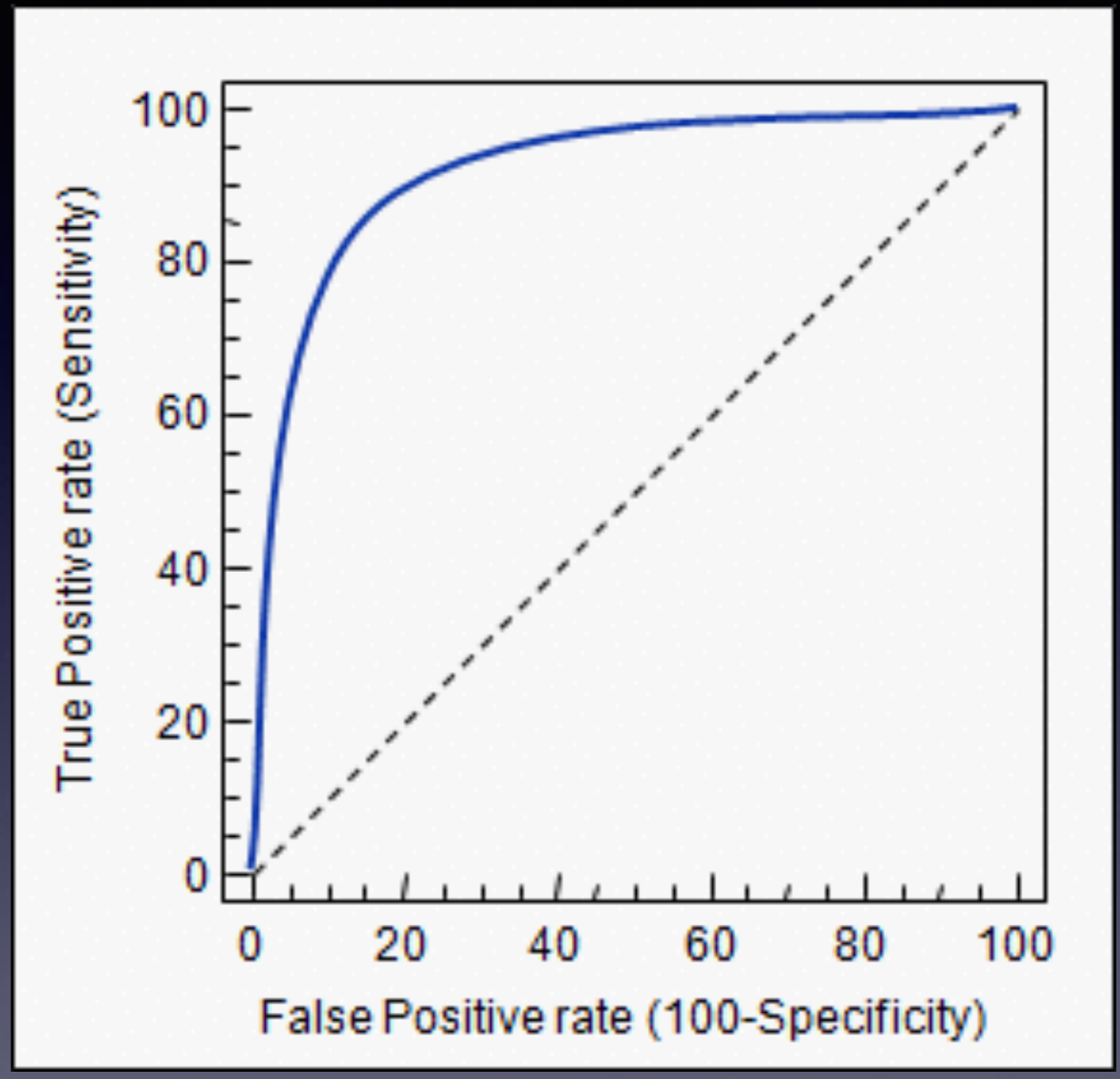


Precision versus recall

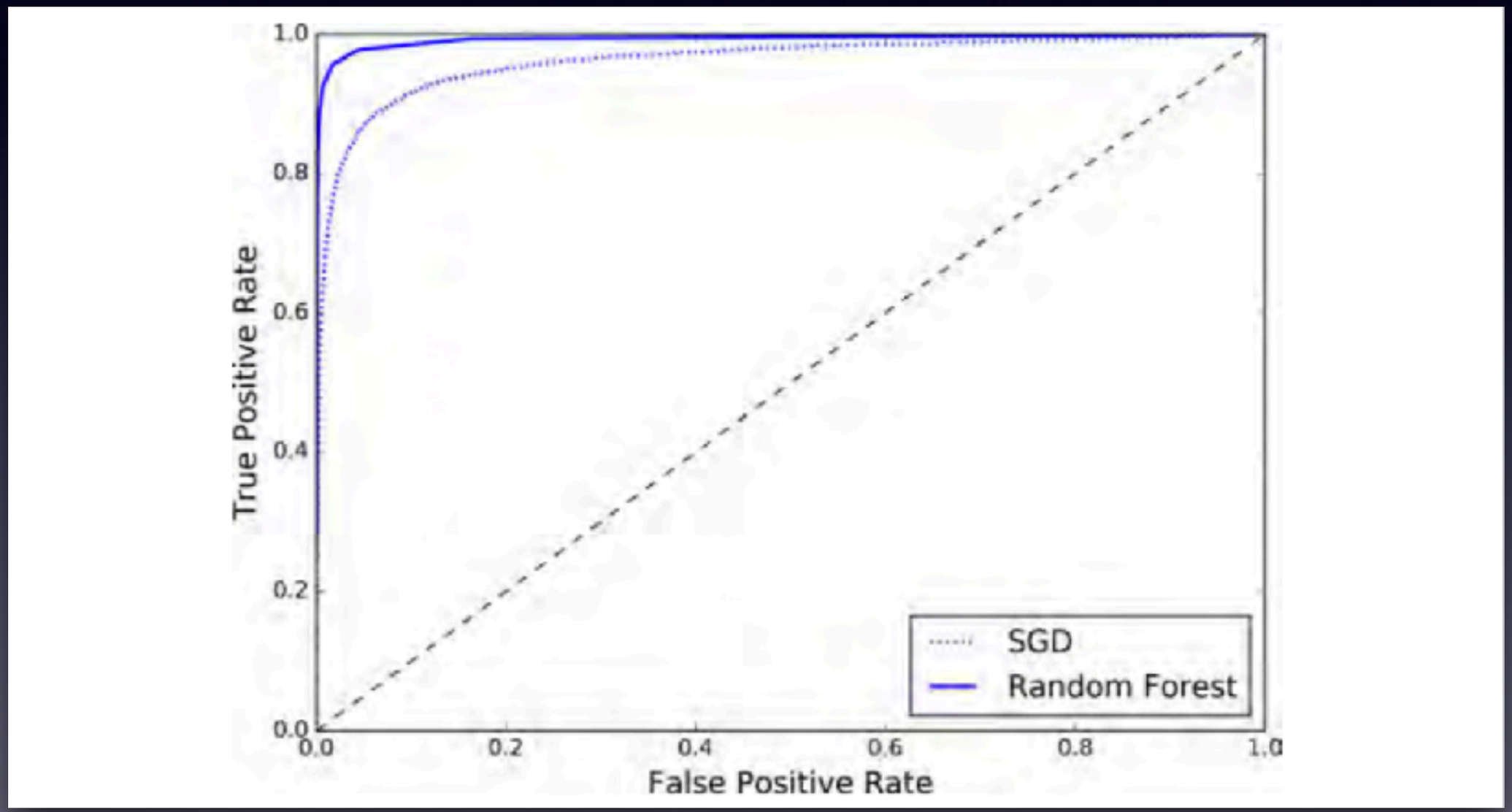


ROC curve

In Statistics, a receiver operating characteristic curve, i.e. ROC curve is a plot that illustrates the diagnostic ability of binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection.



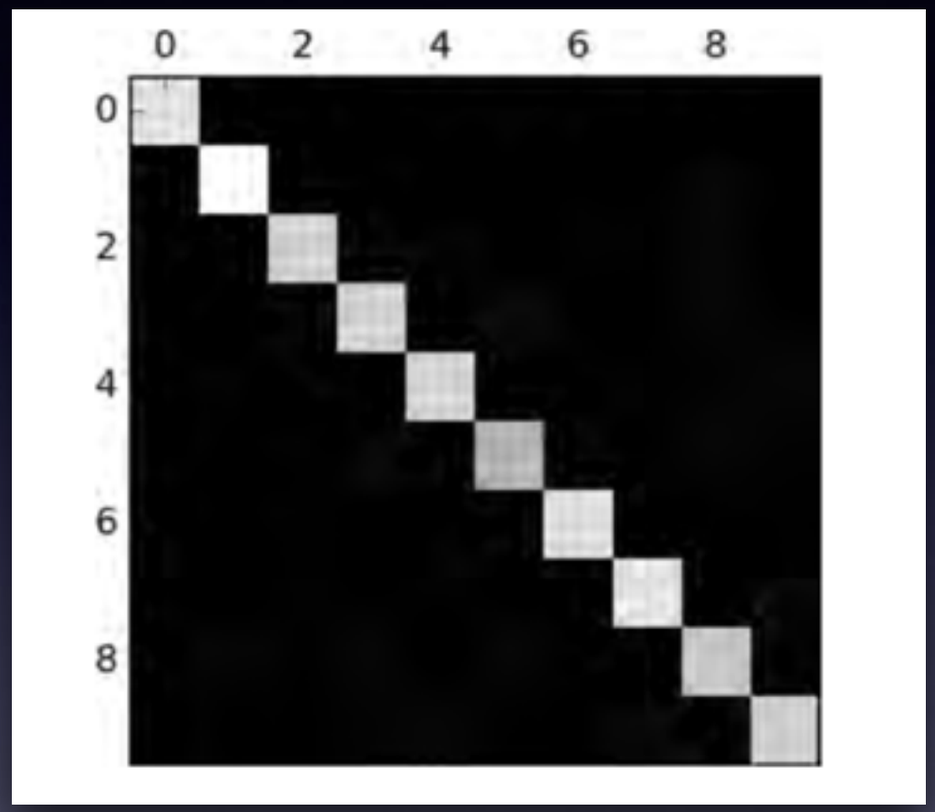
Comparing ROC curves

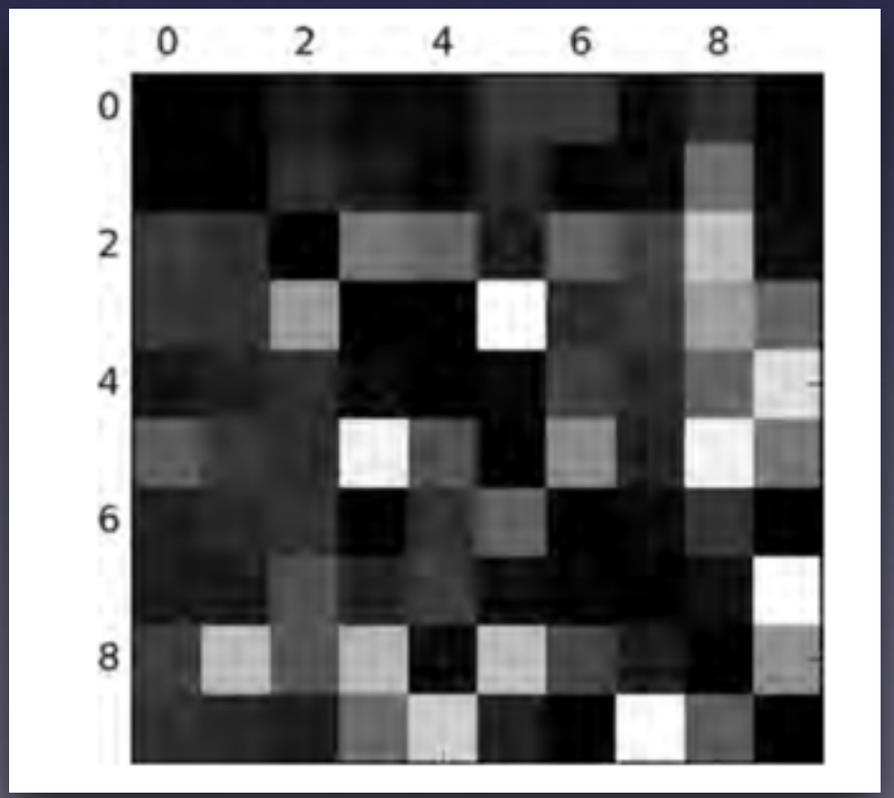


Multiclass Classification

Whereas binary classifiers distinguish between two classes, multi class classifiers can distinguish between more than two classes.

Error Analysis





Multilabel classification