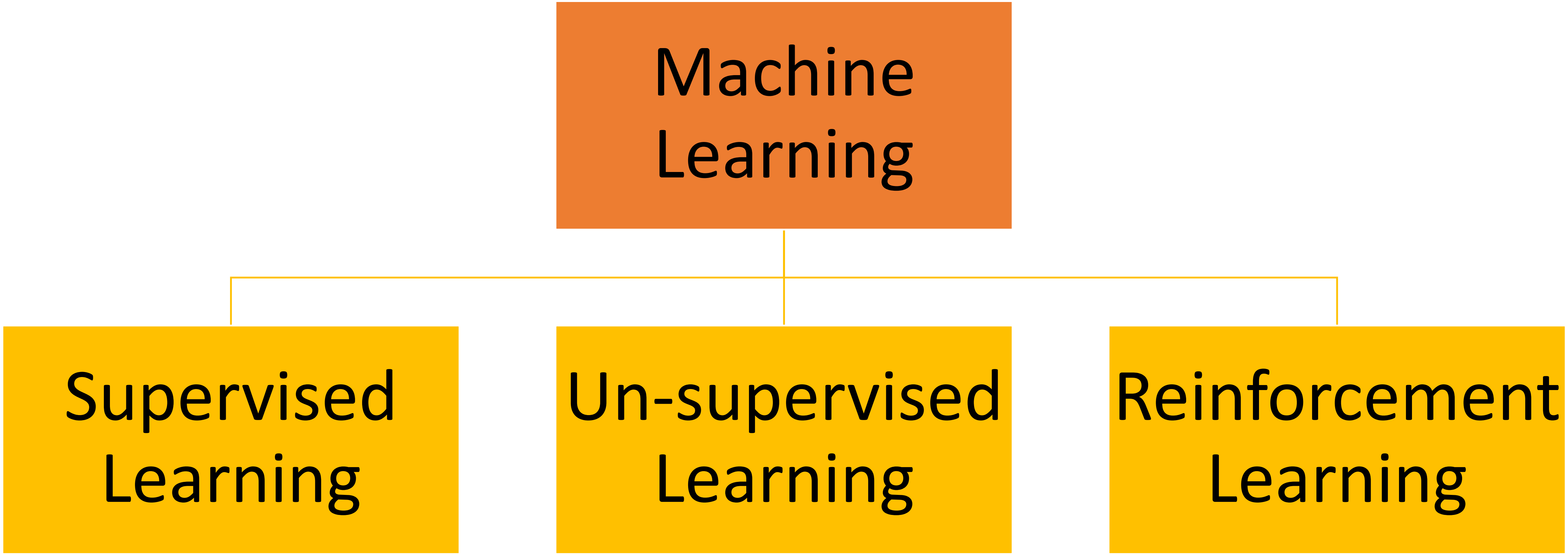
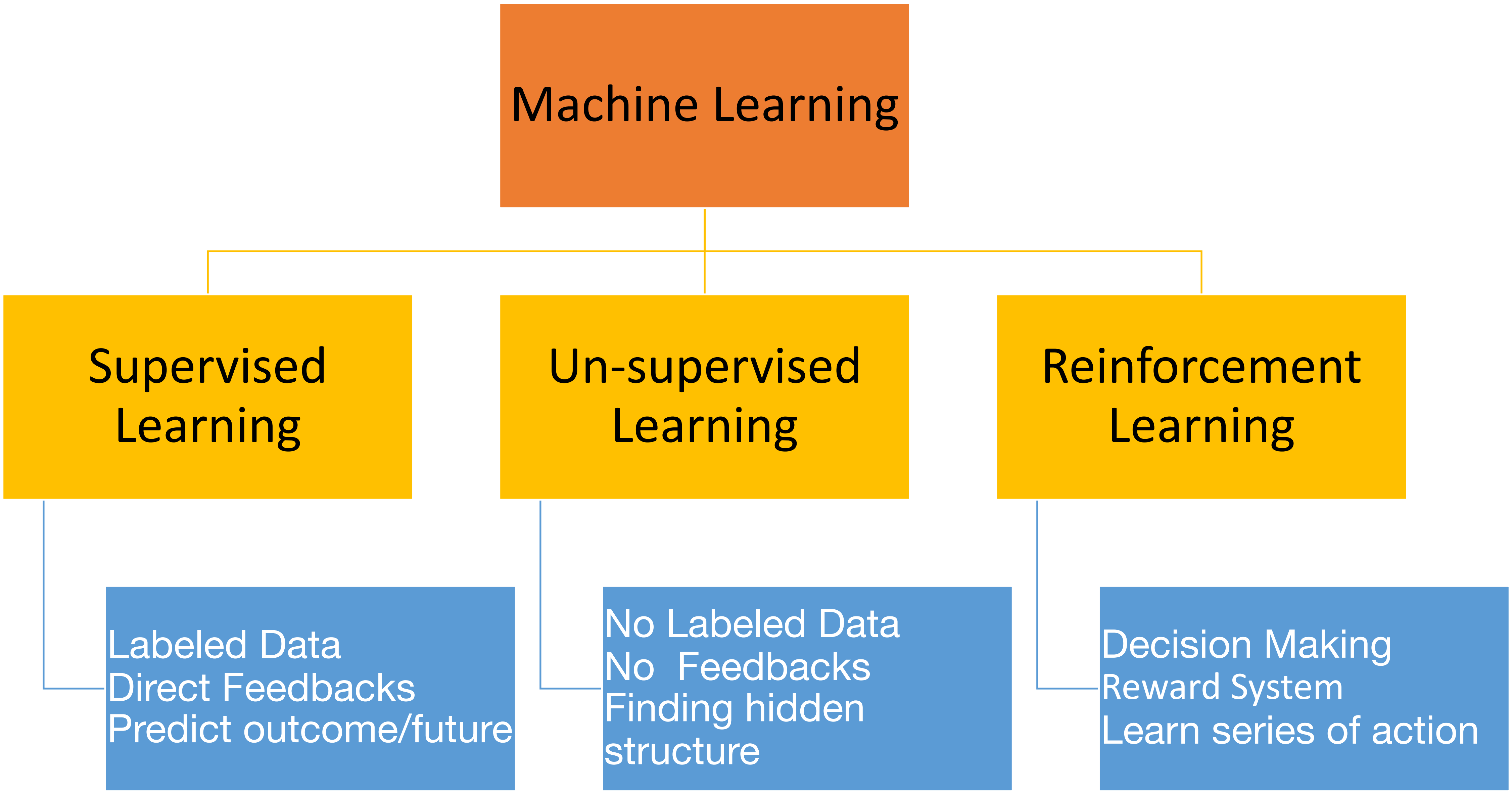
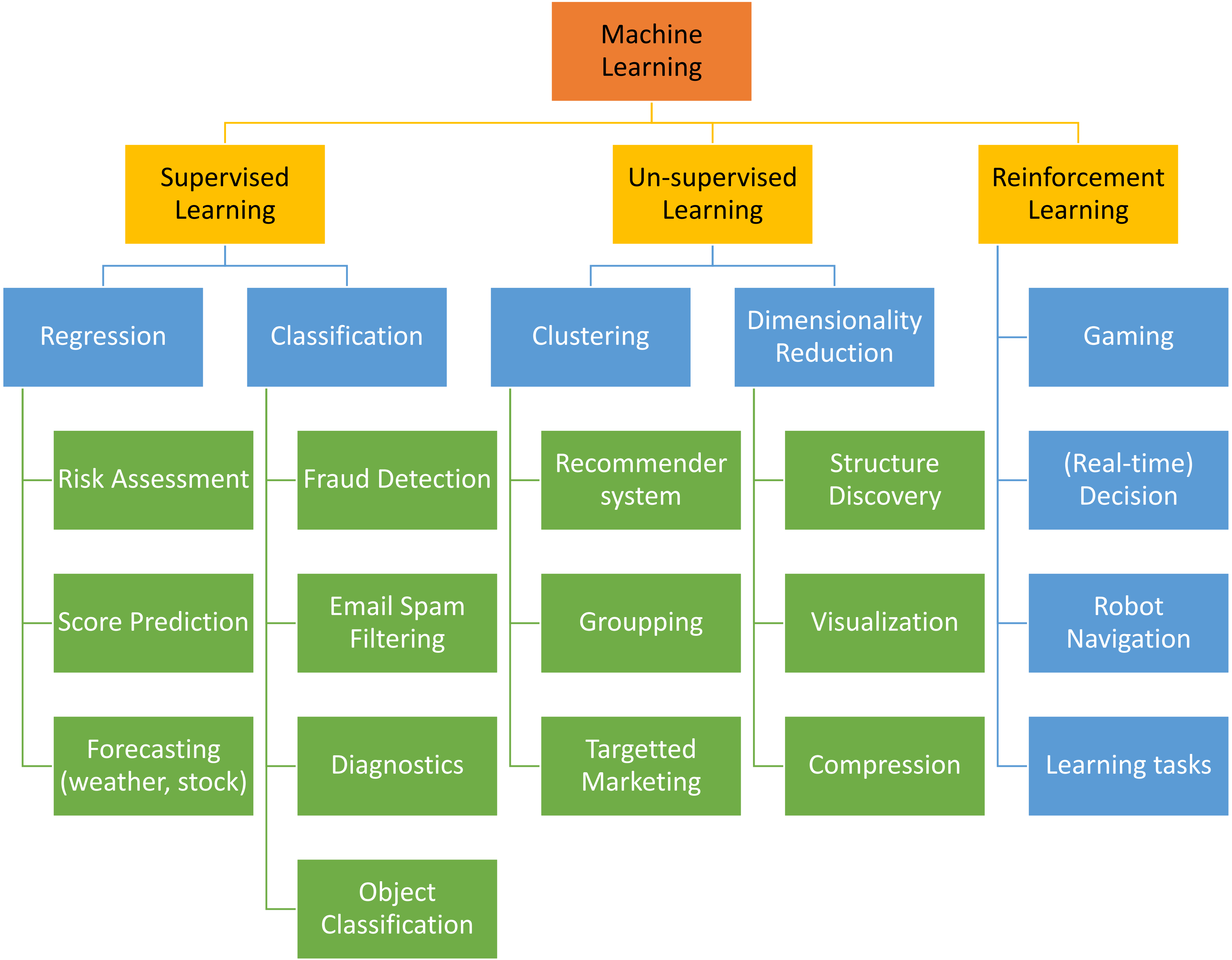
Applied Machine Learning Classification: Introduction

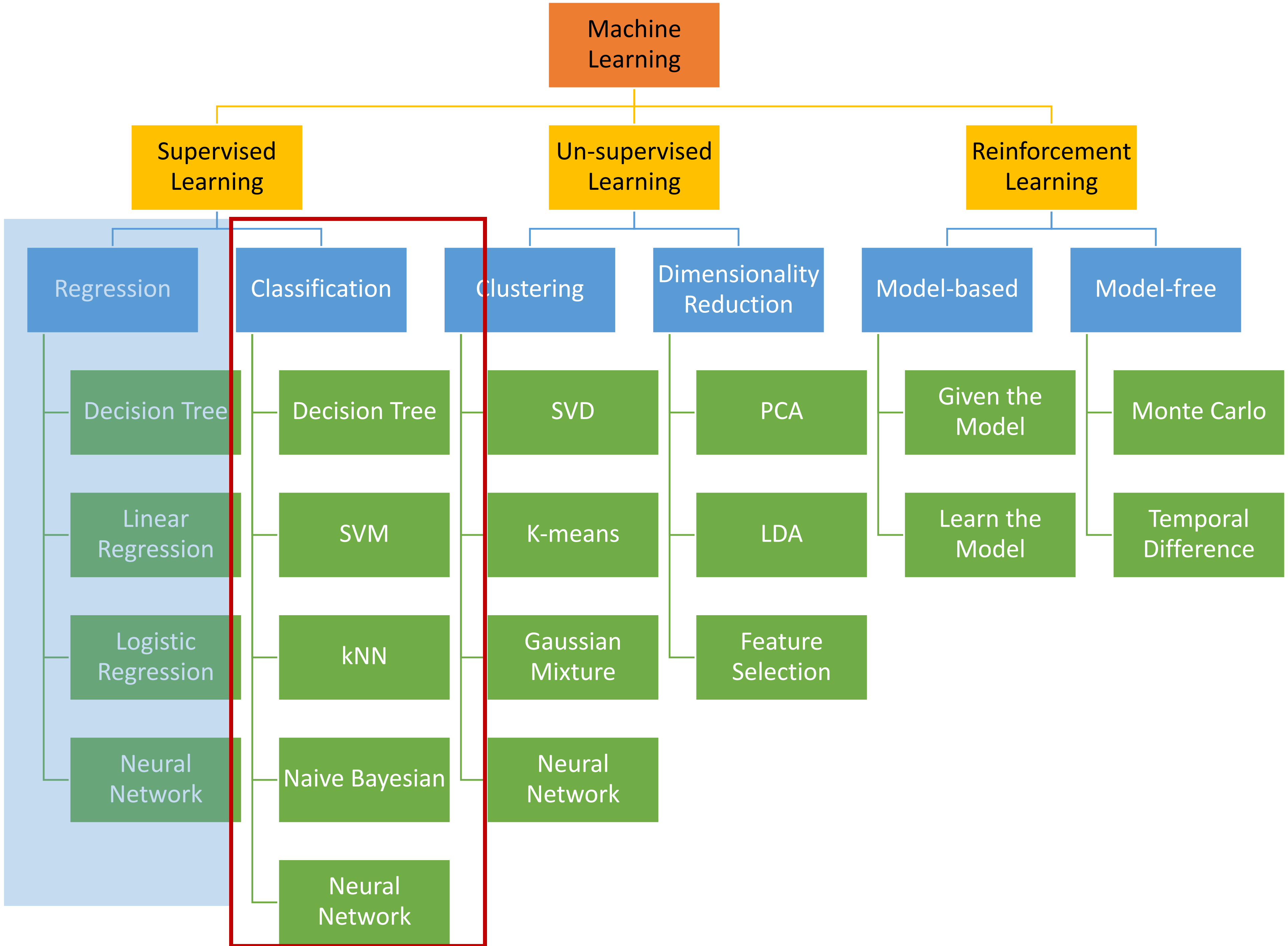
Ngan Le

thile@uark.edu


Machine Learning
Supervised Learning
Un-supervised Learning
Reinforcement Learning


Machine Learning
Supervised Learning
Un-supervised Learning
Reinforcement Learning
Labeled Data
Direct Feedbacks
Predict outcome/future
No Labeled Data
No Feedbacks
Finding hidden structure
Decision Making
Reward System
Learn series of action


Machine Learning
Supervised Learning
Un-supervised Learning
Reinforcement Learning
Regression
Classification
Clustering
Dimensionality Reduction
Gaming
Risk Assessment
Fraud Detection
Recommender system
Structure Discovery
(Real-time) Decision
Score Prediction
Email Spam Filtering
Groupping
Visualization
Robot Navigation
Forecasting (weather, stock)
Diagnostics
Targetted Marketing
Compression
Learning tasks
Object Classification


Machine Learning
Supervised Learning
Un-supervised Learning
Reinforcement Learning
Regression
Classification
Clustering
Dimensionality Reduction
Model-based
Model-free
Decision Tree
Linear Regression
Logistic Regression
Neural Network
Decision Tree
SVM
kNN
Naive Bayesian
Neural Network
SVD
K-means
Gaussian Mixture
Neural Network
PCA
LDA
Feature Selection
Given the Model
Learn the Model
Monte Carlo
Temporal Difference

Classification V.S Regression

Regression

predict a continuous value

Classification

predict the "class" of a data point

Classification

Blue

Binary Classification

Yellow

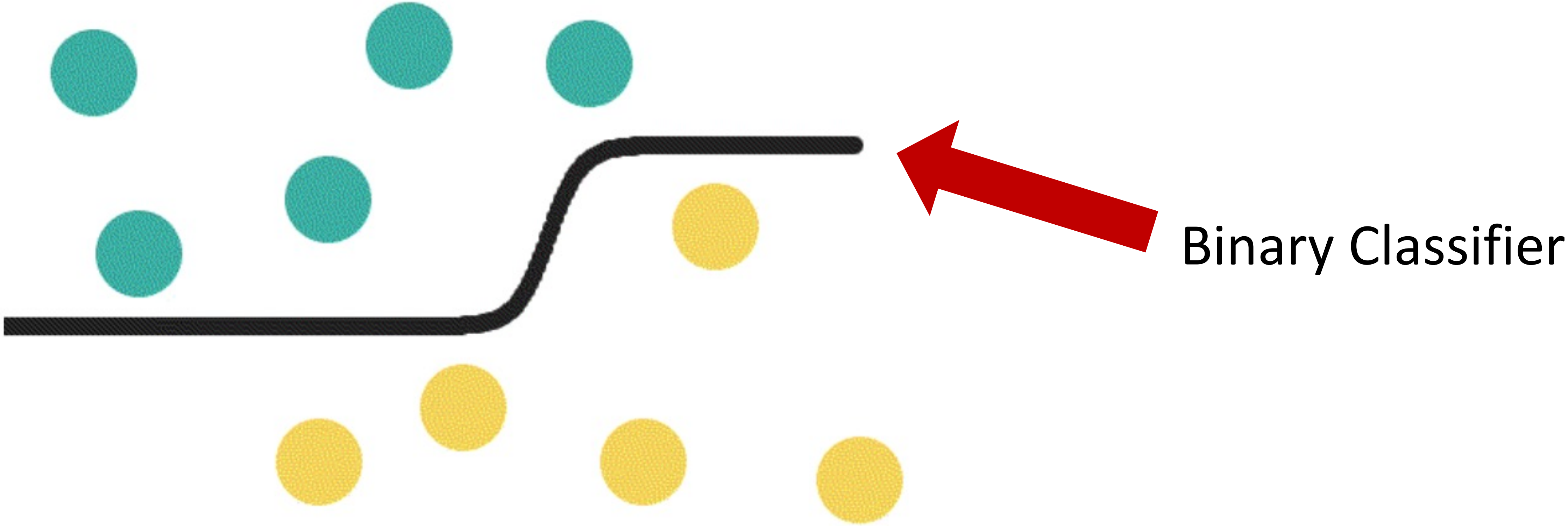
Classification

Spam

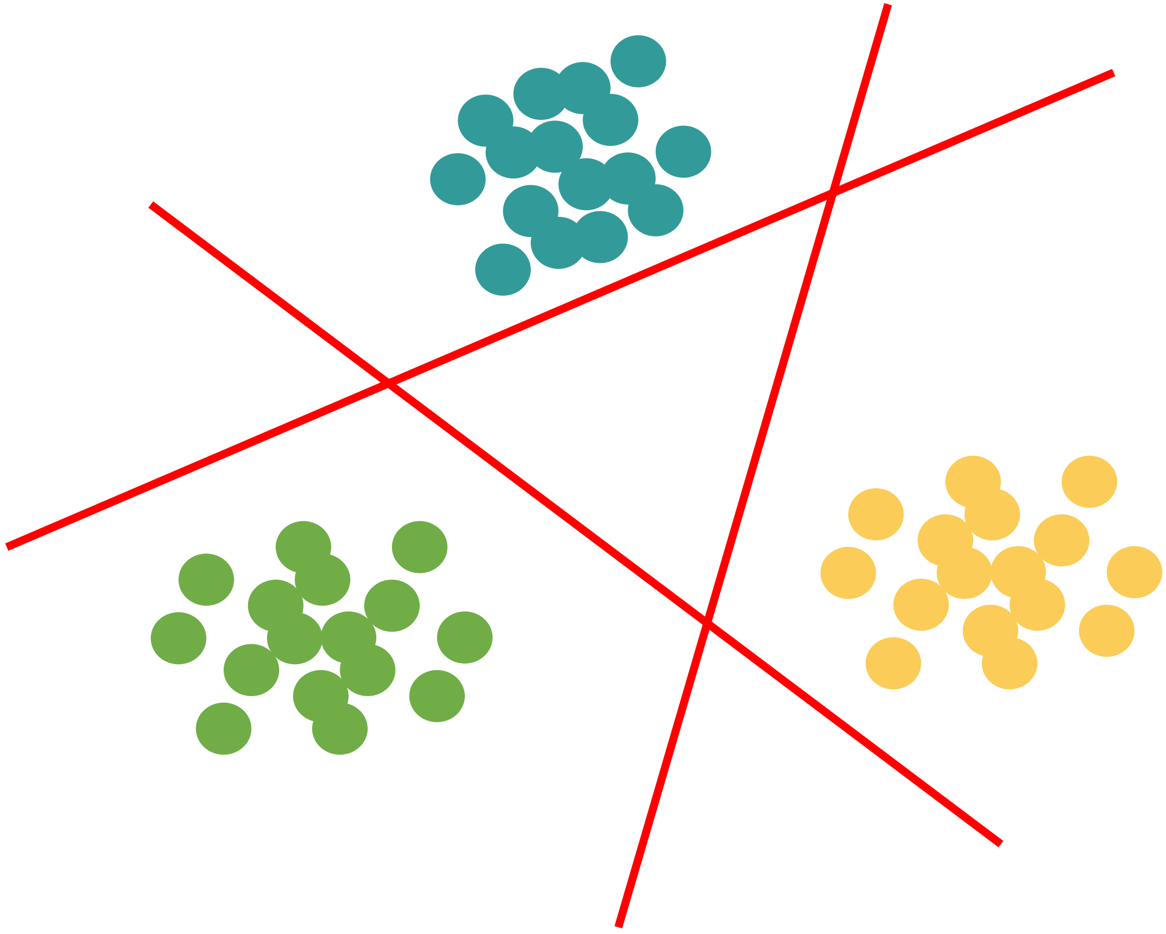
Binary Classification

Not Spam

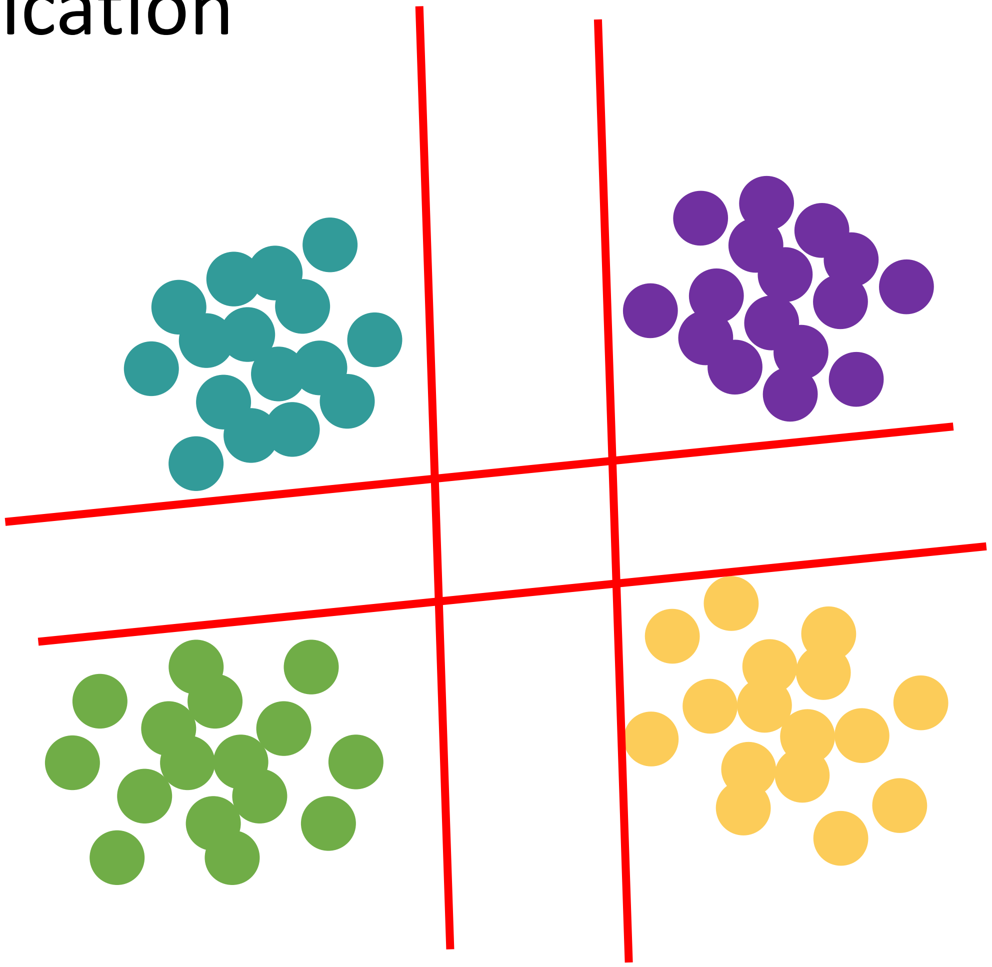
Classification



Classification

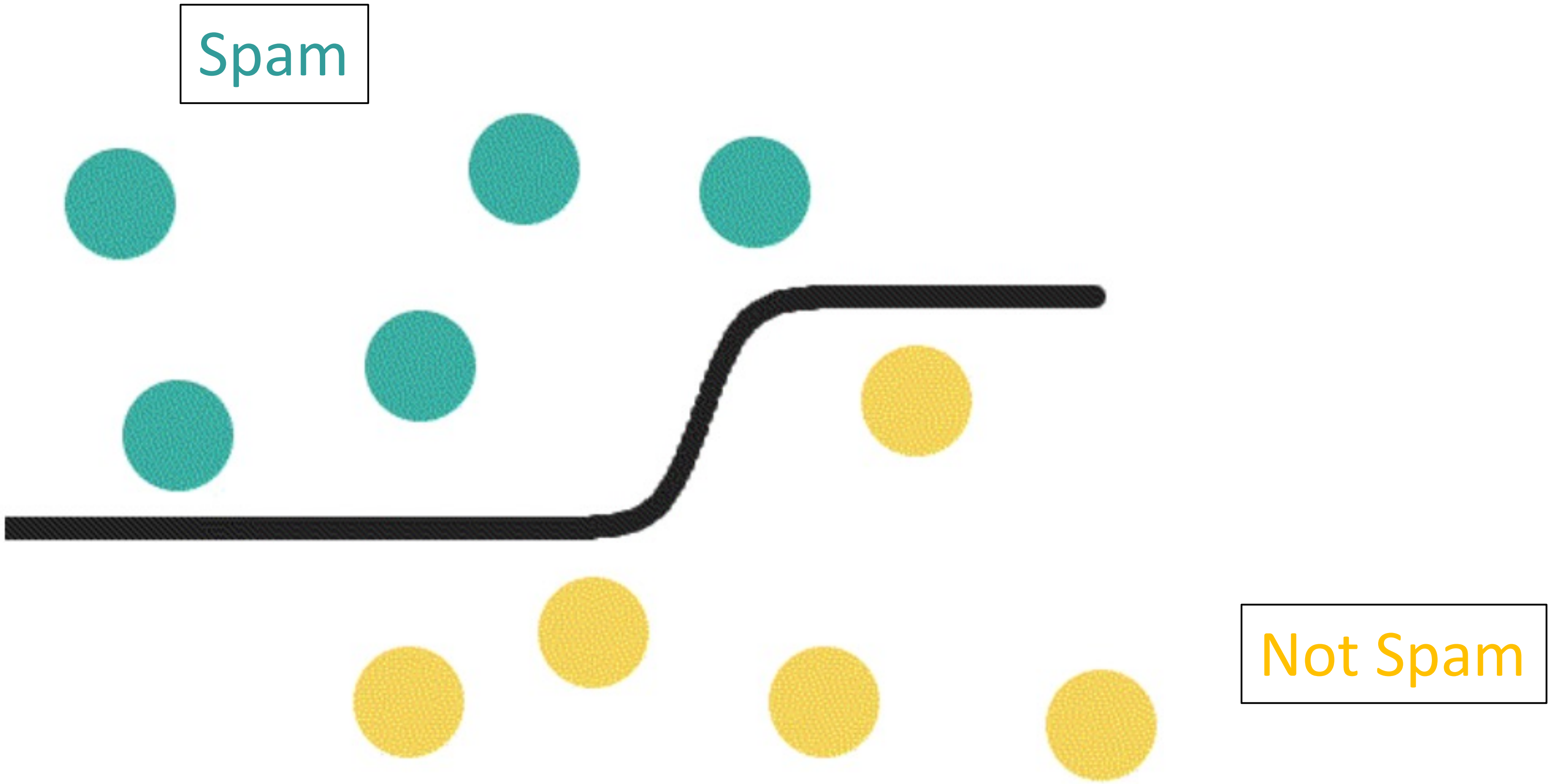


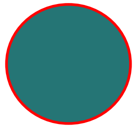
Multi- Classification



What Does It Mean to Classify?

Classification

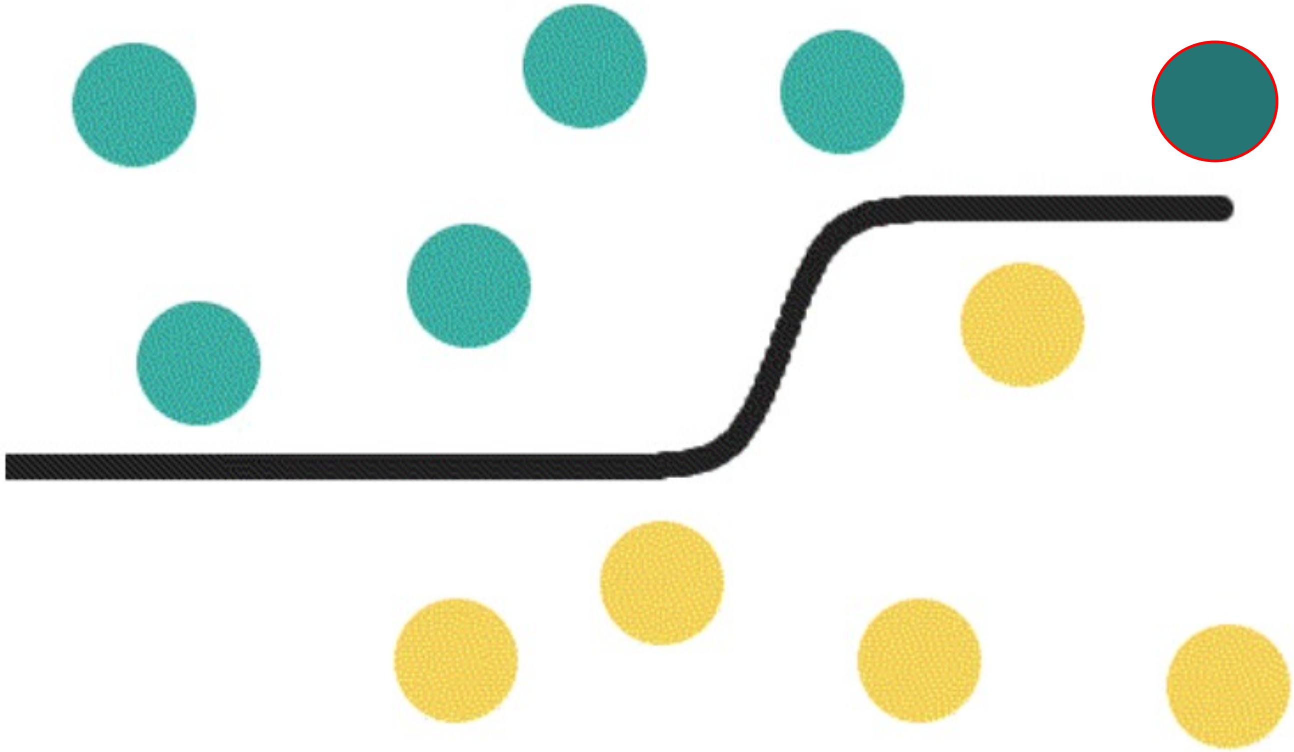




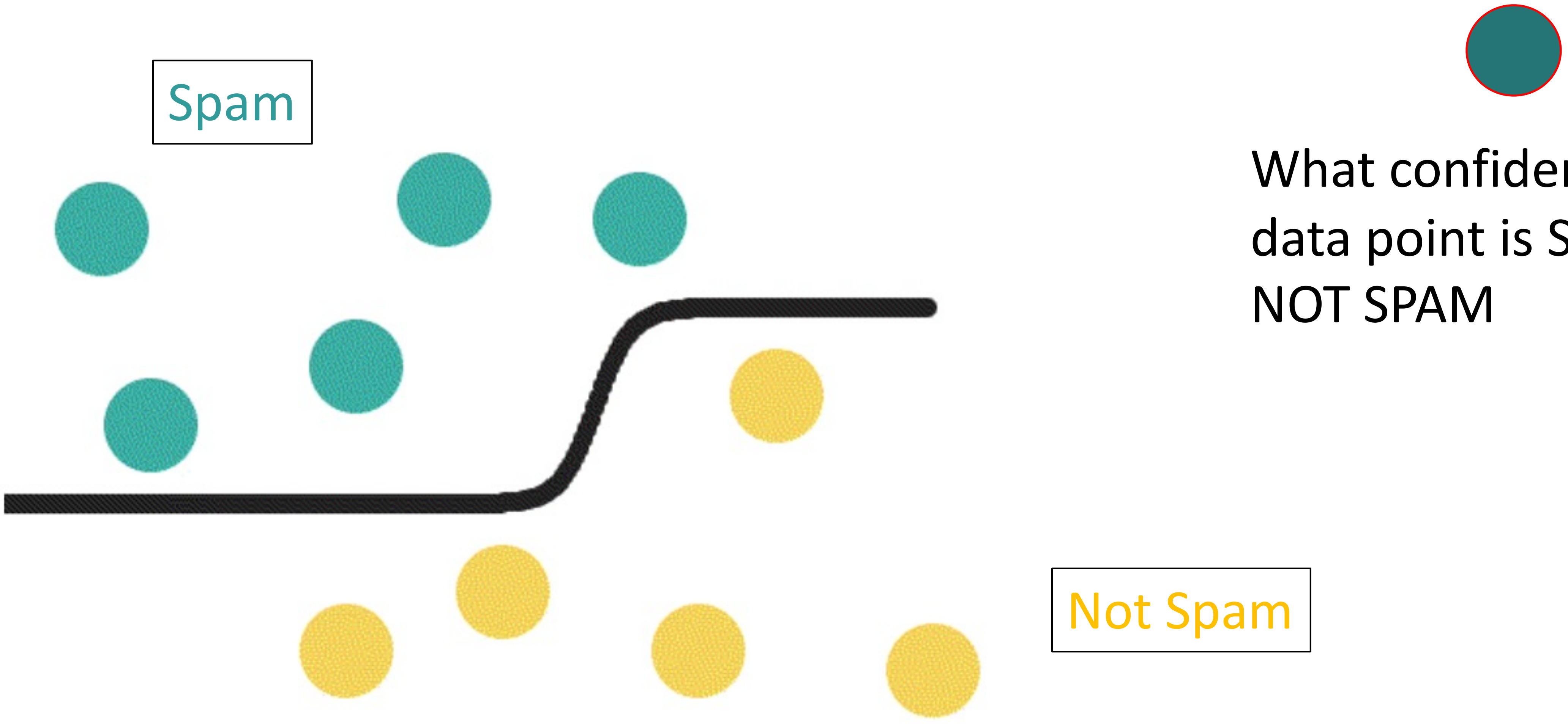
Is this spam?

Classification

YES/NO

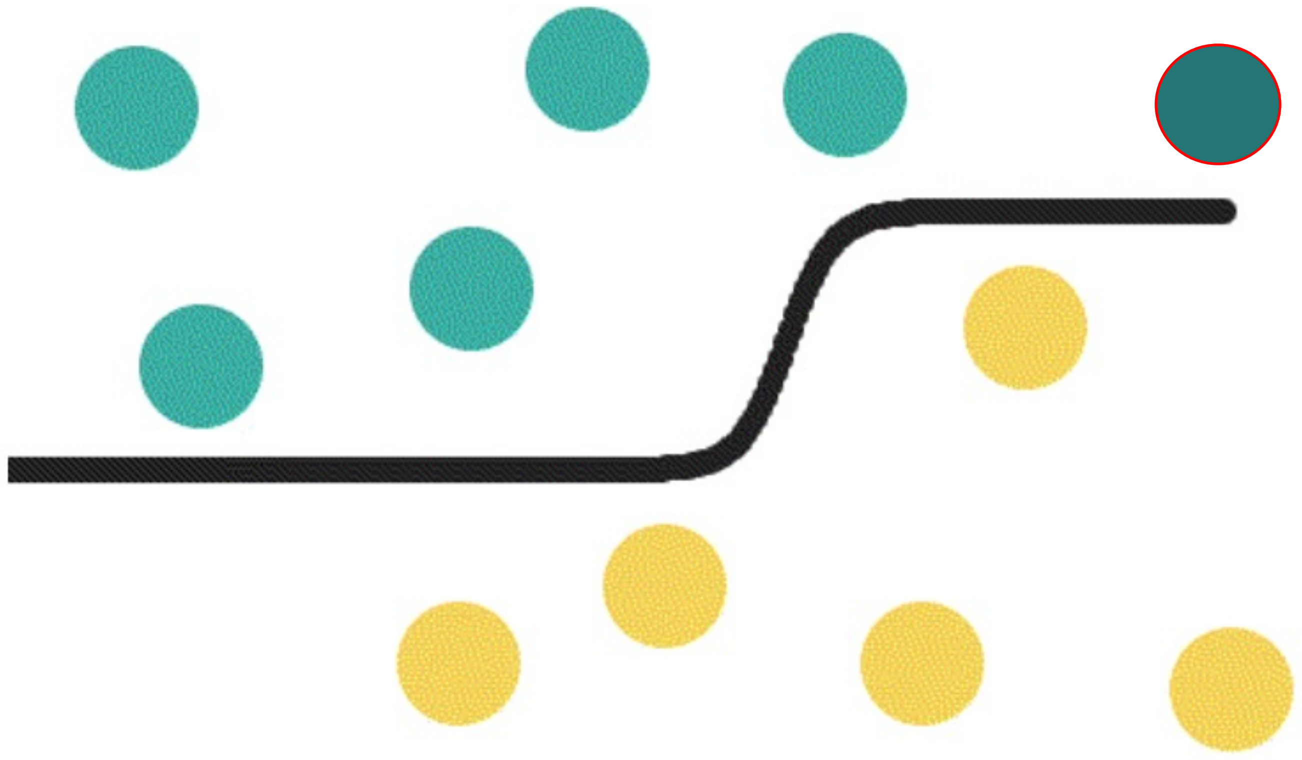


Classification



What confidences that data point is SPAM and NOT SPAM

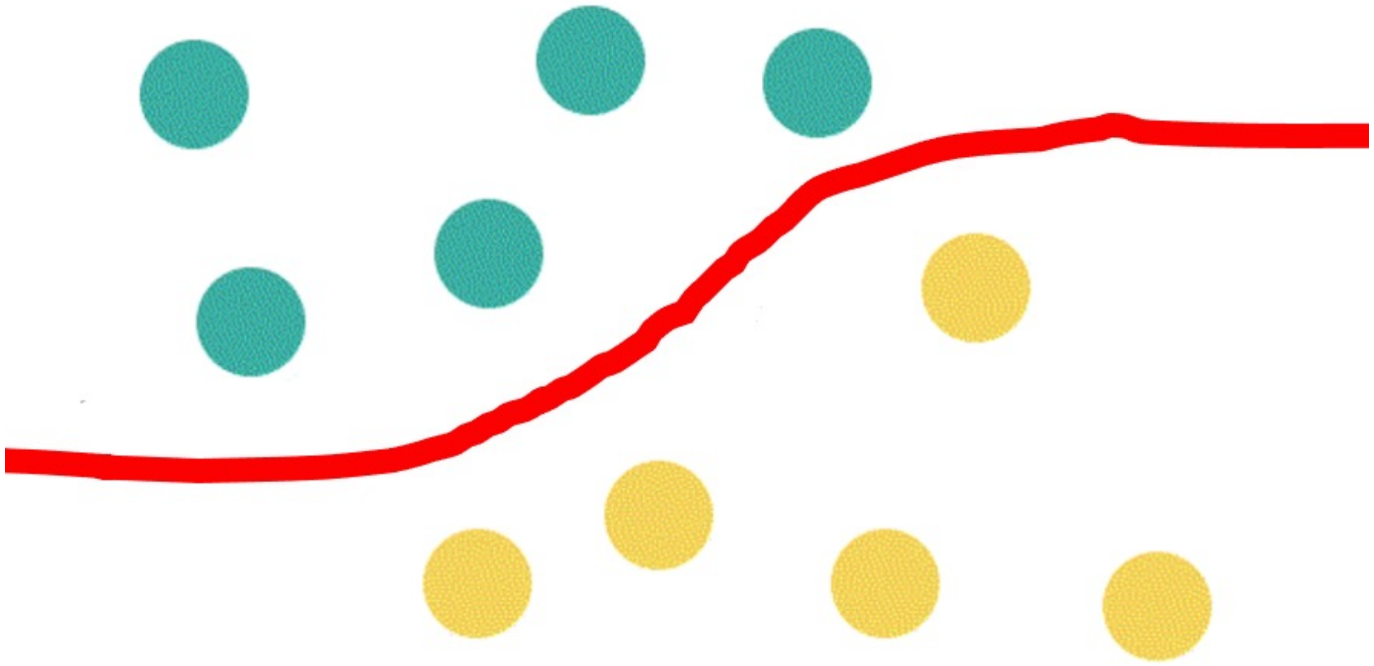
Classification



SPAM: 95%

NOT SPAM: 25%

Classification

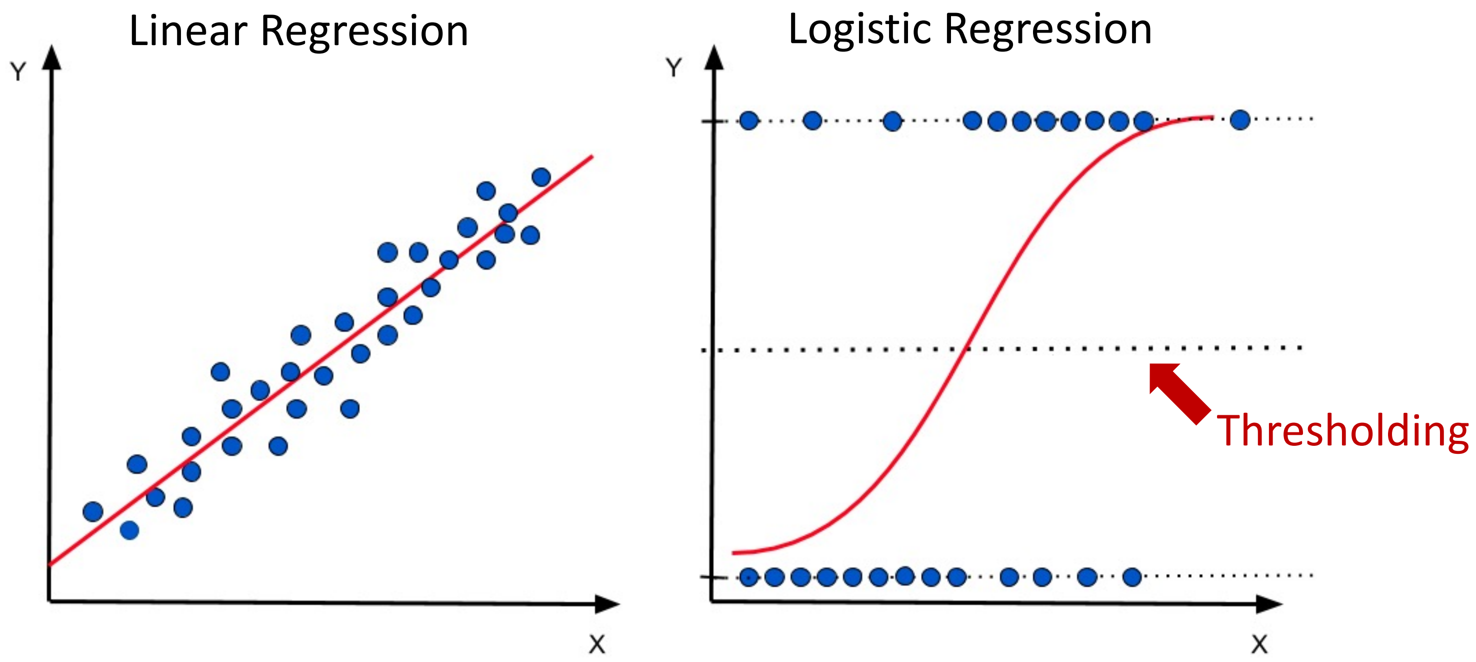


- Logistic regression
- Nearest Neighbors
- Decision trees
- Random forests
- SVM
- Naive Bayes
- Deep Learning (Neural Network)

Classification

- **Logistic regression**

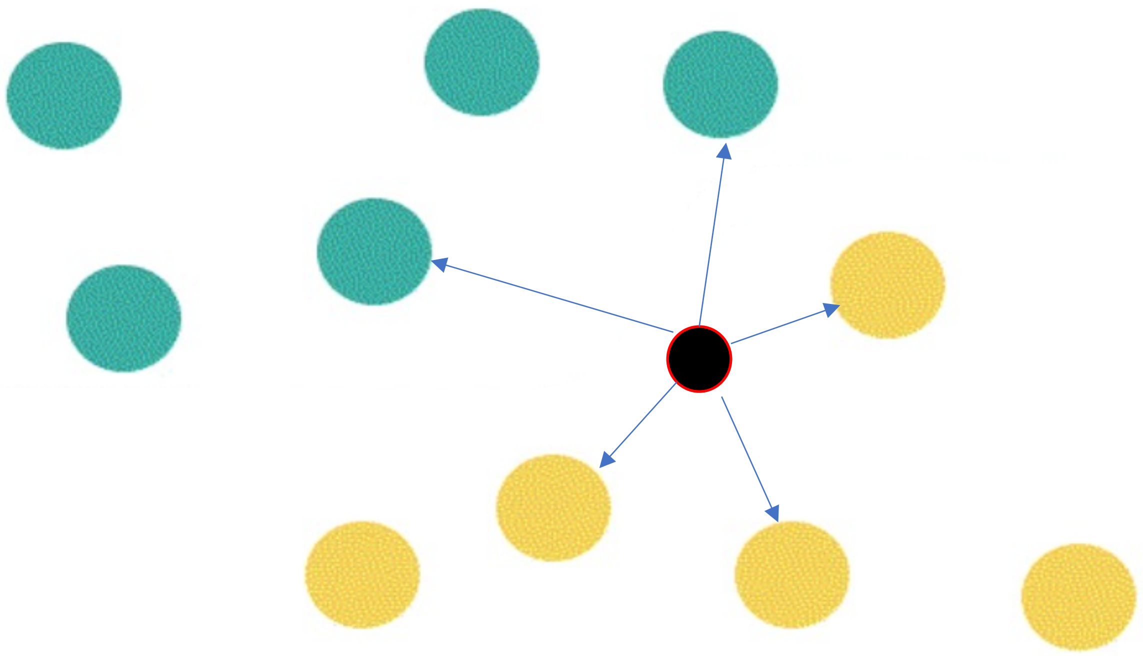
A variation of linear regression that performs a regression, then uses some threshold to make a classification decision



Classification

- **Nearest Neighbors**

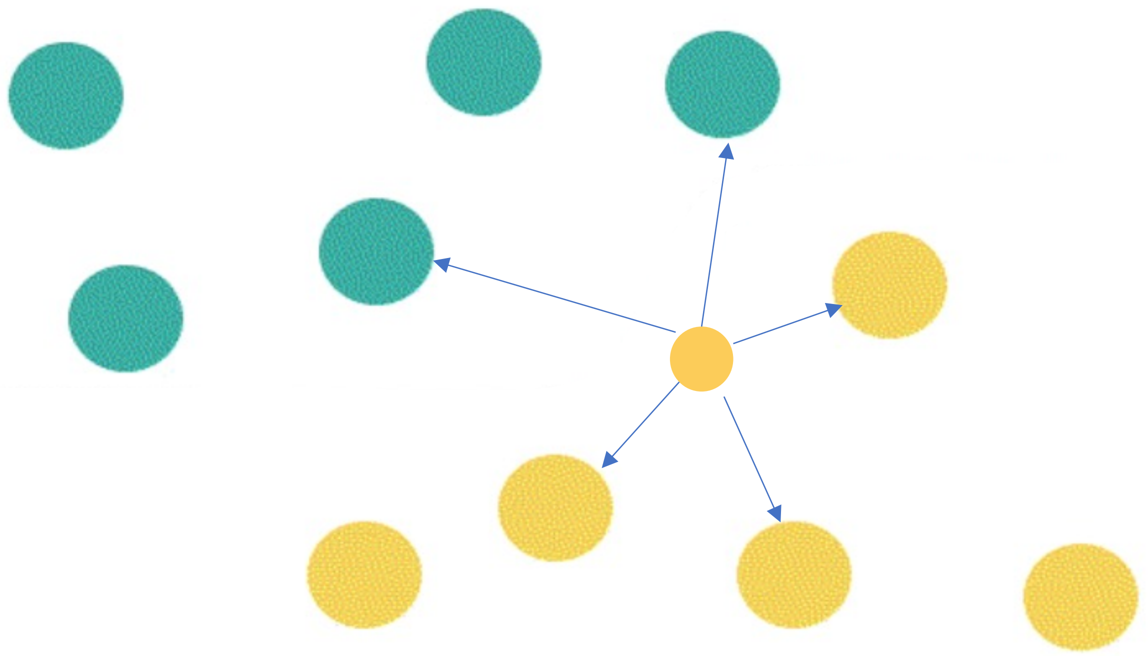
a distance measure is used to find the neighbors of a datapoint, and classification decisions are made from those



Classification

- **Nearest Neighbors**

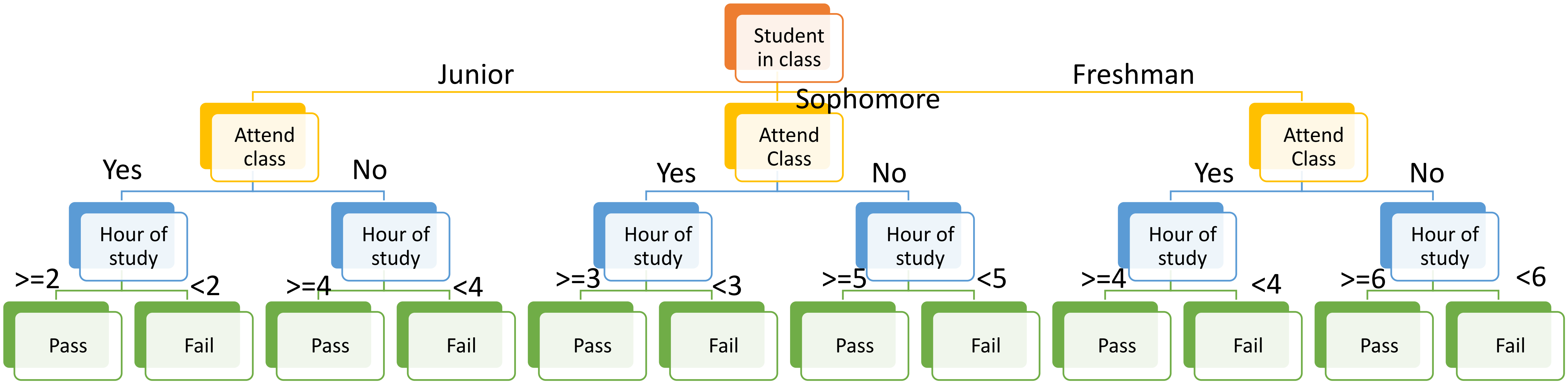
a distance measure is used to find the neighbors of a datapoint, and classification decisions are made from those



Classification

- **Decision trees**

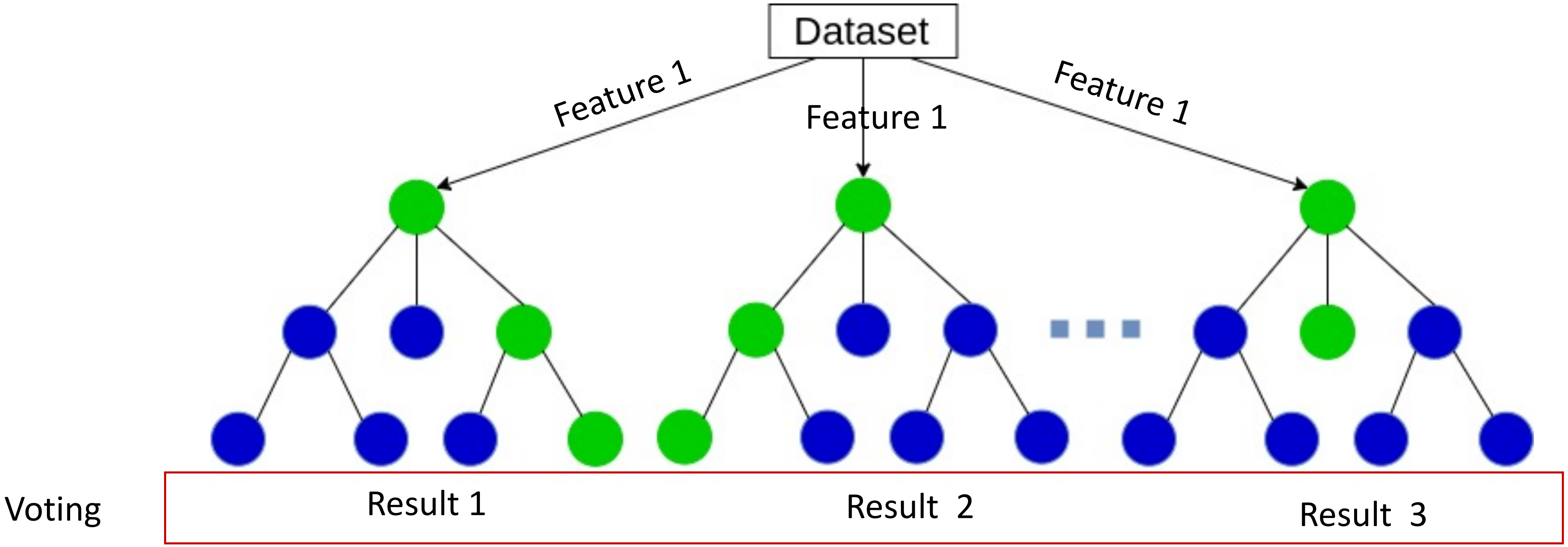
a tree structure where a classification is made through a series of small decisions that ultimately lead to the leaf of a tree



Classification

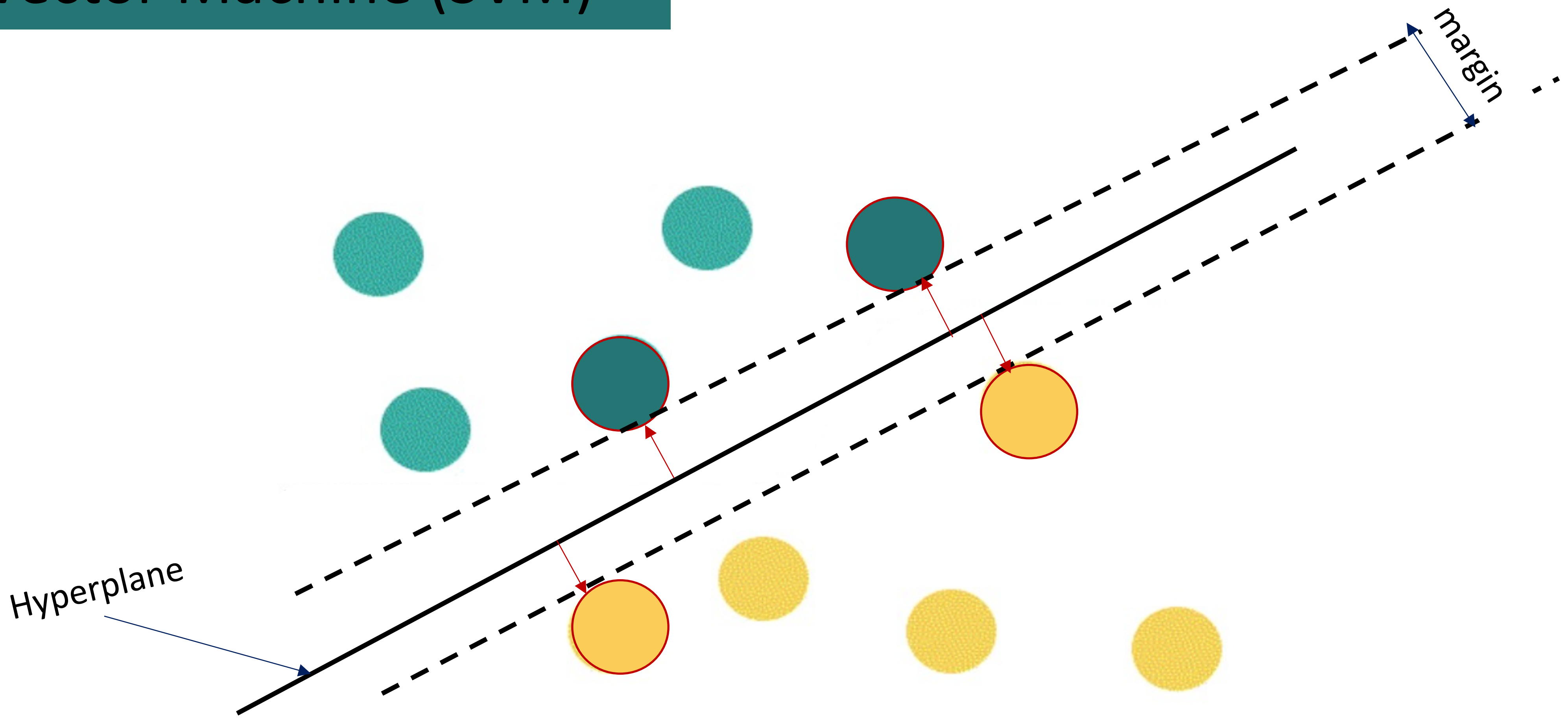
- Random forests

a group of trees, each with a random part of the training data, are queried and a consensus classification decision is made



Classification

- Support Vector Machine (SVM)



Classification

- **Naïve Bayes**

Bayesian statistics applied to data to make a classification decision

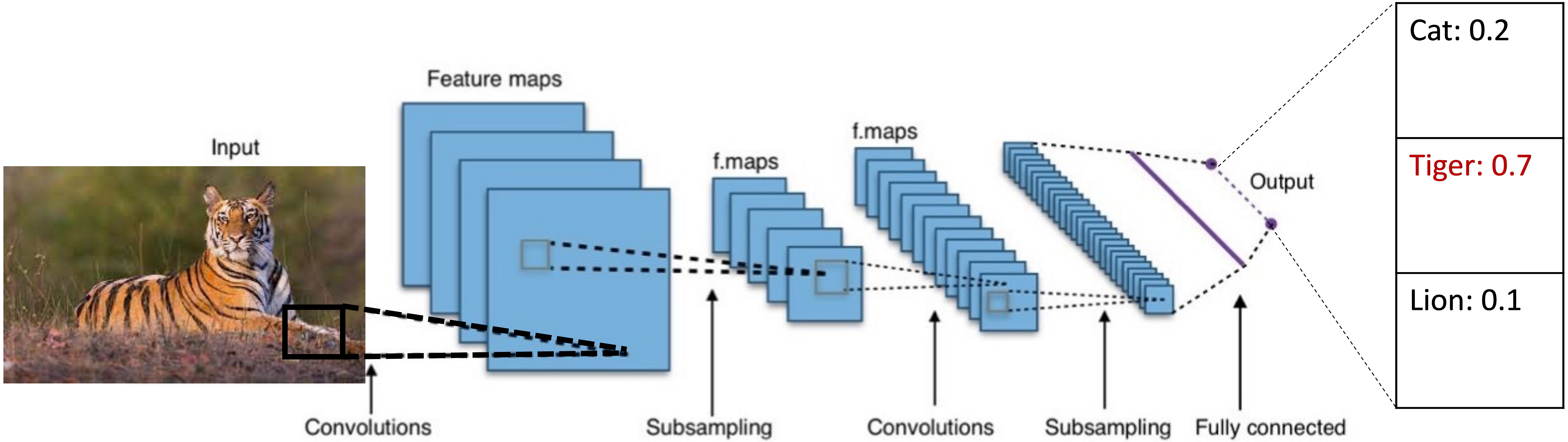
Chills	Cough	Headache	Fever	Covid-19
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Milk	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Chills	Cough	Headache	Fever	Covid-19
Y	N	Mild	Y	?

Classification

- **Deep Learning**

Neural networks trained to make classification decisions

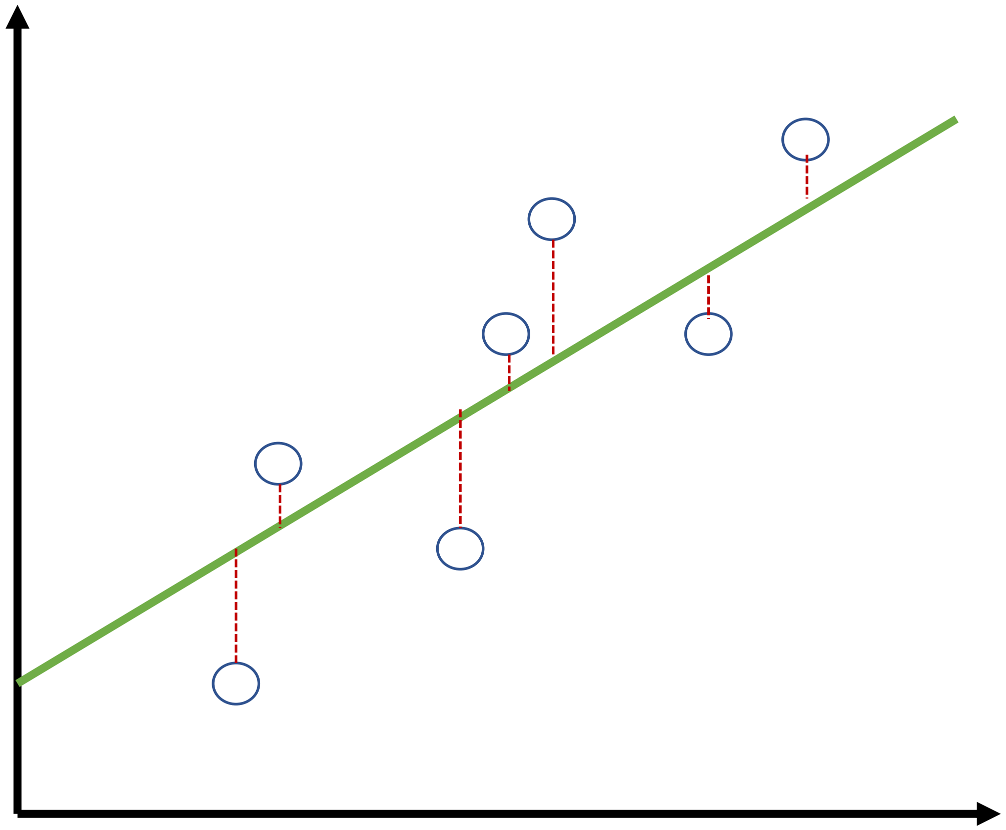


Convolution Neural Network

Source: Wikipedia

Model Performance

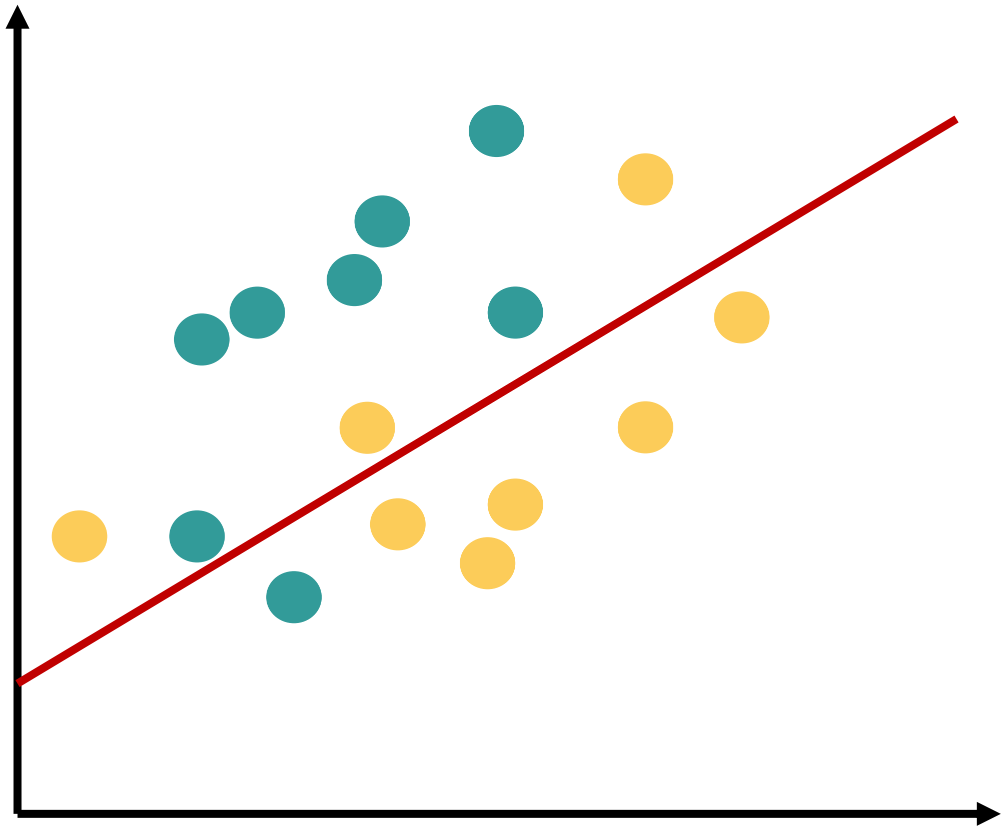
Regression



measuring the distance
between continuous values

Model Performance

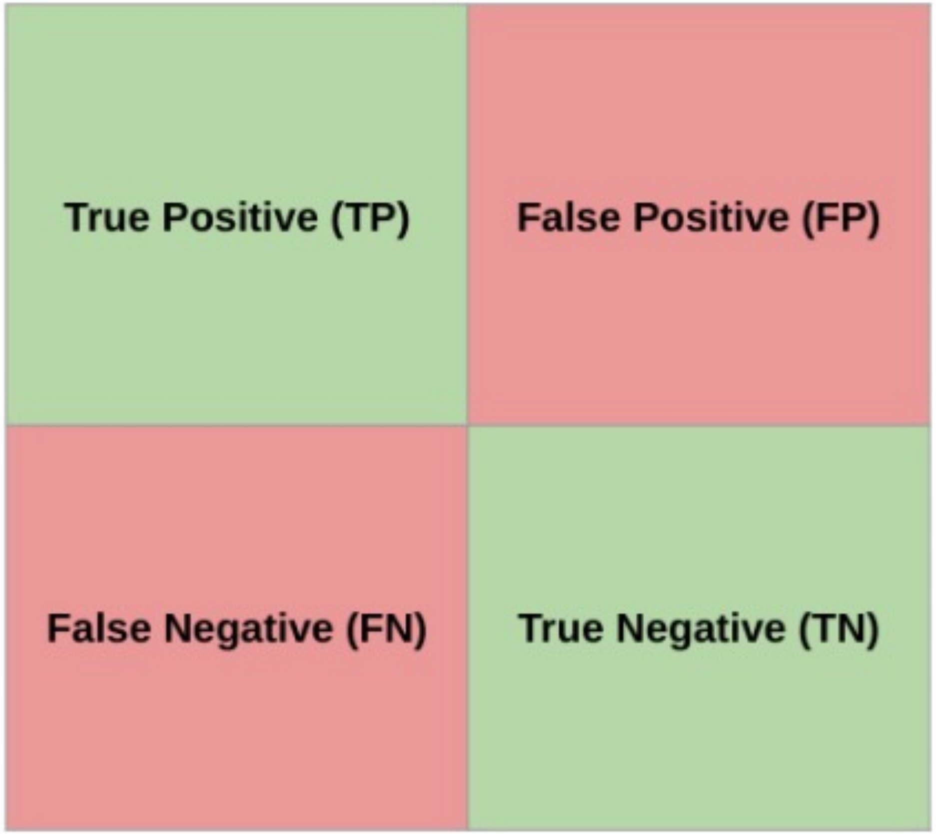
Classification



number of predictions that the model got **correct** and the number that were **incorrect**

Model Performance

Confusion Matrix



model predicted ”**Positive**"
correct class is ”**Positive**

model predicted ”**Positive**"
correct class is ”**Negative”**

model predicted ”**Negative**"
correct class is ”**Positive”**

model predicted ”**Negative**"
correct class is ”**Negative”**

Model Performance - <u>**Accuracy**</u>

The fraction of predictions that a classification model got right

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

number of predictions that the classifier got **correct**

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

the total number of predictions made

But,

- When the model predicted positive, how often was it right?
- What is the probability that a tumor is actually malignant, given that our model classified it as malignant ?

Model Performance - <u>Precision</u>

The fraction of prediction that a classification model got right when predicting positive cases

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

True positive:
correct positive case prediction

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

all positive case predictions

Higher Precision?

Precision = 1.0 : ?

But,

- This says nothing about how many malignant tumors our model is missing.
- Out of all possible positives, how many did the model correctly identify?
- What is the probability that our model will classify a tumor as malignant, given that it actually is malignant

Model Performance - <u>Recall</u>

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

Higher Recall

Recall = 1.0 : ?

$$\frac{\overbrace{TP}^{\text{True positive: correct positive case prediction}}}{\underbrace{TP + FN}_{\text{all actual positive}}}$$

Balancing precision and recall is a tug-of-war between the metrics.

Precision

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

Recall

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

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

If we want to increase **recall**, we should predict positive more often.

If we want to increase **precision**, we should only predict positive when we're absolutely sure

In general, raising the classification threshold reduces false positives, thus raising precision.

What is a good way to determine
if precision and recall are balanced?

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

F1: computes the harmonic mean for the values.

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} \qquad \frac{TP}{TP + \frac{FN + FP}{2}}$$

high F1 score helps keep both precision and recall high.

Classification Performance: Recap

model predicted "**Positive**" correct class is "**Positive**	True Positive (TP)	False Positive (FP)	model predicted "**Positive**" correct class is "**Negative**"
model predicted "**Negative**" correct class is "**Positive**"	False Negative (FN)	True Negative (TN)	model predicted "**Negative**" correct class is "**Negative**"

Which do I use?

Accuracy	Precision	Recall	F1
number of predictions that the classifier got **correct** $\frac{TP+TN}{TP+TN+FP+FN}$ the total number of predictions made	True positive: correct positive case prediction $\frac{TP}{TP+FP}$ all positive case predictions	True positive: correct positive case prediction $\frac{TP}{TP+FN}$ all actual positive	$\frac{TP}{TP+\frac{FN+FP}{2}}$

Example – Model to predict a tumor is malignant

A false alarm scenario, also called Type I error

True Positive (TP) Reality: Malignant Prediction: Malignant	**False Positive (FP)** Reality: Benign Prediction: Malignant
False Negative (FN) Reality: Malignant Prediction: Benign	**True Negative (TN)** Reality: Benign Prediction: Benign

A miss scenario, also called Type II error

Example – Model to predict a tumor is malignant

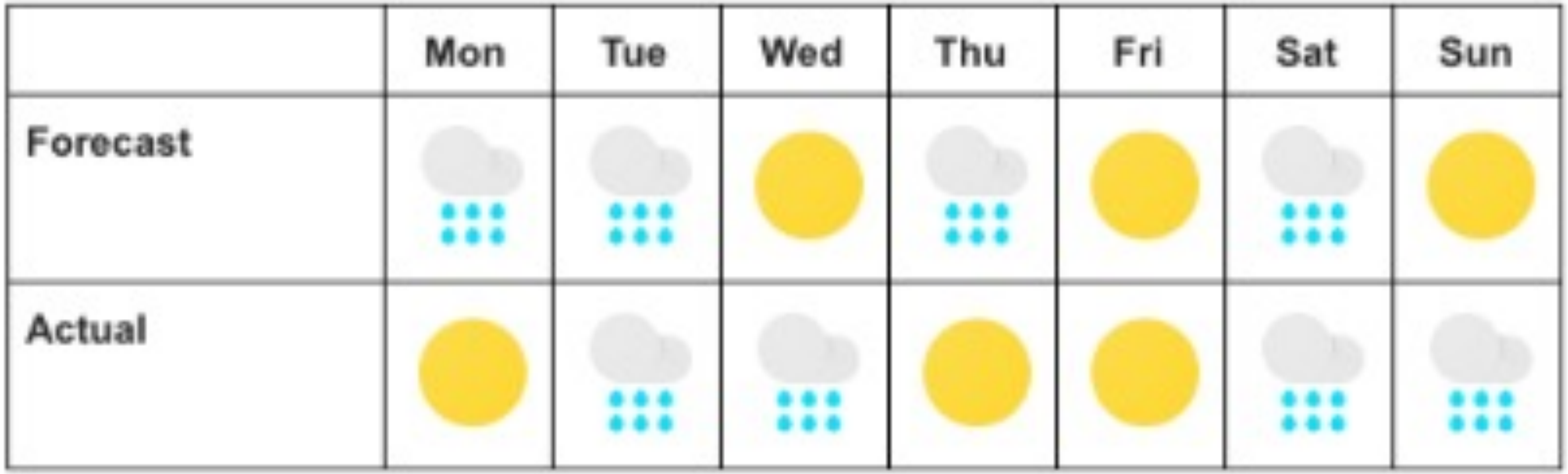
The total number of predictions is 100 counts

1 count of TP
1 count of FP
8 counts of FN
90 counts of TN

True Positive (TP): 1 count Reality: Malignant Prediction: Malignant	False Positive (FP): 1 count Reality: Benign Prediction: Malignant
False Negative (FN): 8 counts Reality: Malignant Prediction: Benign	True Negative (TN): 90 counts Reality: Benign Prediction: Benign

Example – Model to predict a tumor is malignant

Model to predict if it is going to **rain**

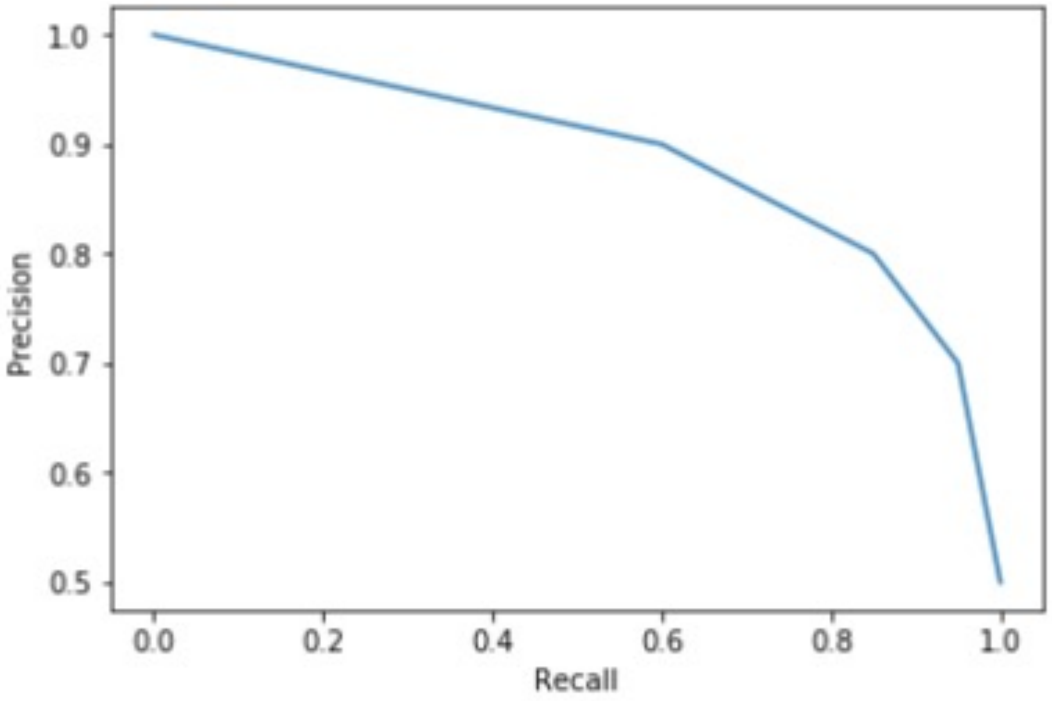


Accuracy	Precision	Recall	F1
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+\frac{FN+FP}{2}}$

Graphical Measurements

Precision VS. Recall Curve

Varying the threshold value for a positive prediction

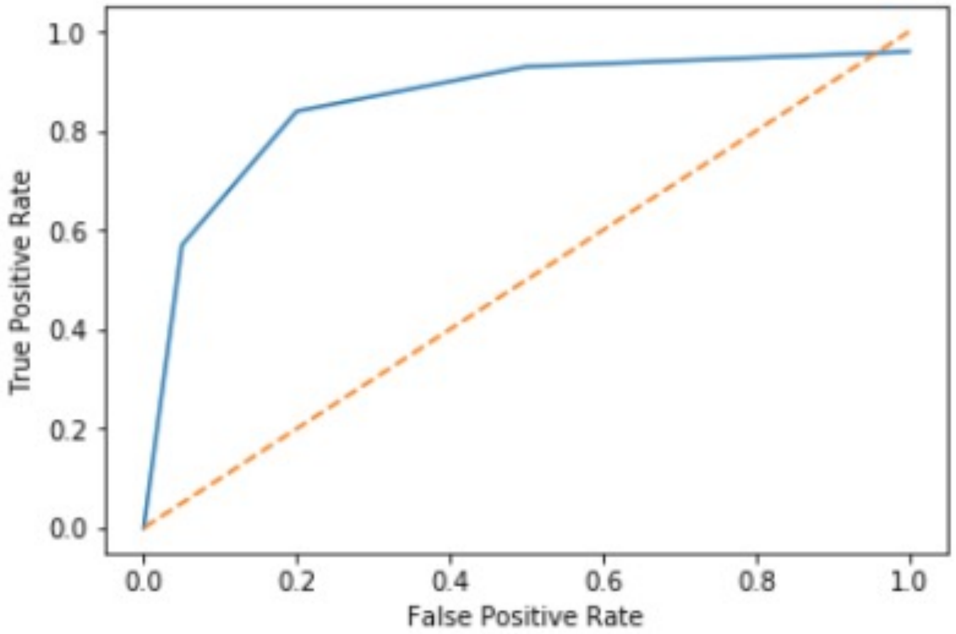


Besides F1 score, detailed plots of precision vs. recall can also be used to pick where to find a balance

Receiver Operating Characteristics (ROC) Curve

True Positive Rate (TPR) (recall)

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



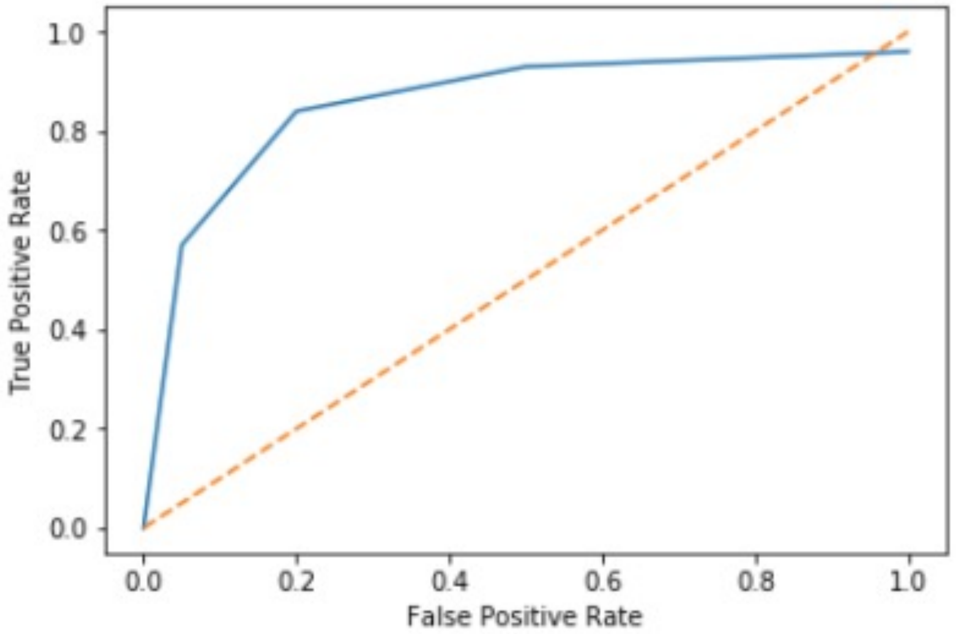
1 - specificity

$$\text{specificity} = \frac{TN}{TN + FP}$$

False Positive Rate (FPR) = 1 - true negative rate = 1 - specificity

Receiver Operating Characteristics (ROC) Curve

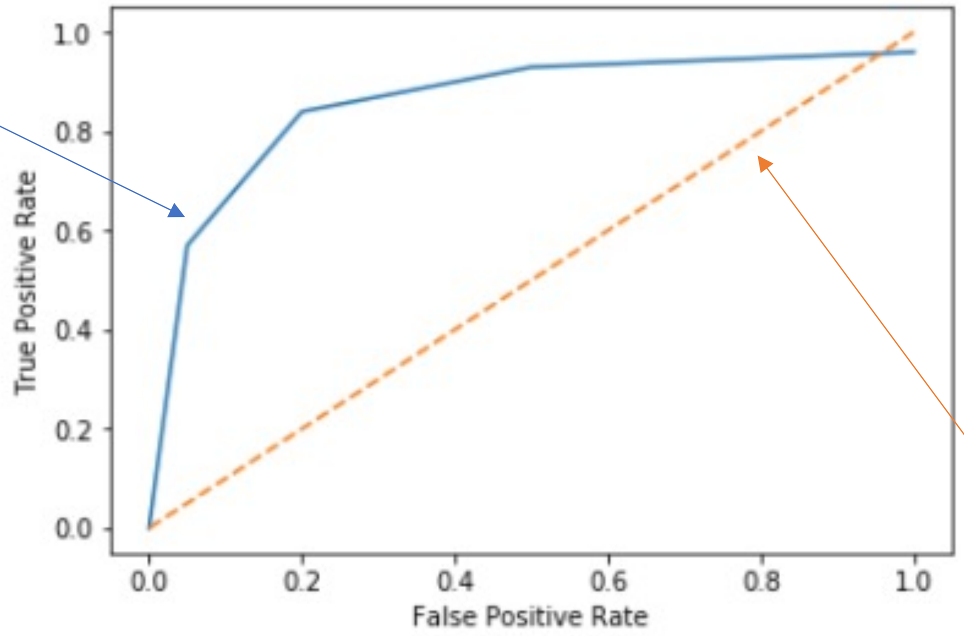
it is the proportion of correctly classified malignant tumors



it is the proportion of incorrectly classified benign tumors (negative samples falsely predicted as positive).

Receiver Operating Characteristics (ROC) Curve

TPR> FPR : probability that you correctly classify malignant tumors is greater than the probability of incorrectly classifying benign tumors. You want this!



TPR = FPR : Probability that you correctly classify a malignant tumor is equal to the probability that you incorrectly classify a benign tumor, i.e., given any sample, malignant or benign, the model has an equal probability of classifying them as malignant.

Classification Performance: Recap

Accuracy	Precision	Recall	F1
number of predictions that the classifier got **correct** $\frac{TP+TN}{TP+TN+FP+FN}$ the total number of predictions made	True positive: correct positive case prediction $\frac{TP}{TP+FP}$ all positive case predictions	True positive: correct positive case prediction $\frac{TP}{TP+FN}$ all actual positive	$\frac{TP}{TP+\frac{FN+FP}{2}}$