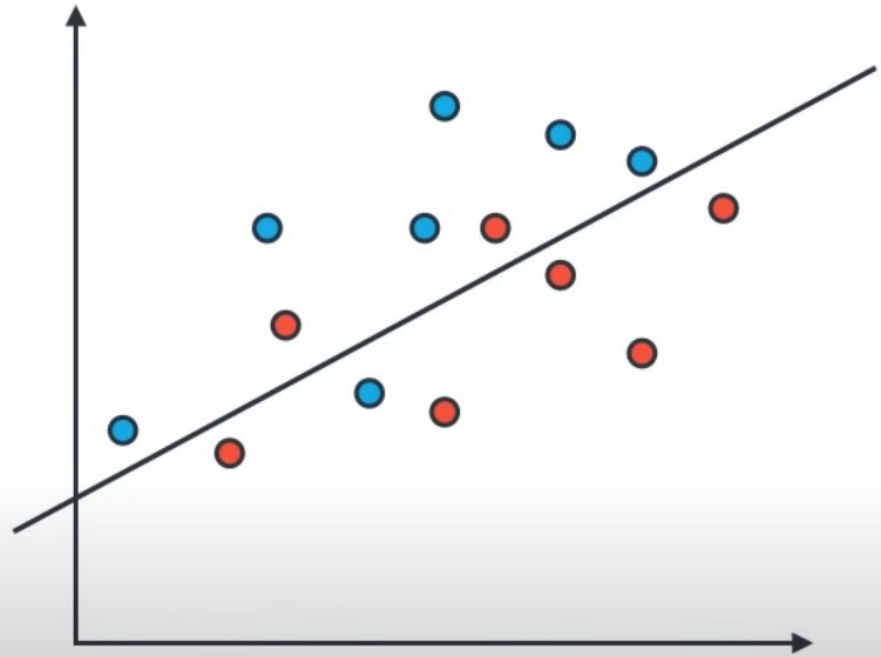
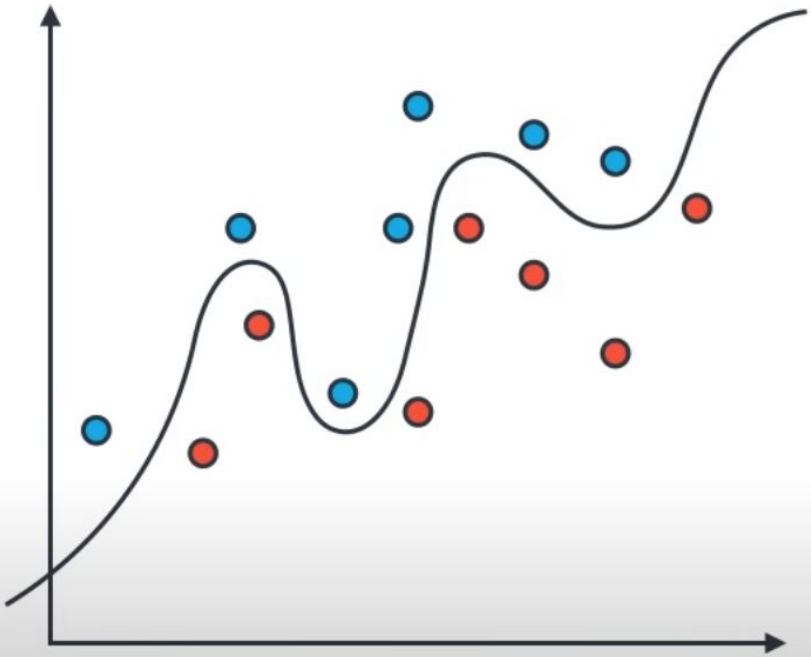
ML Model Testing

ModTest

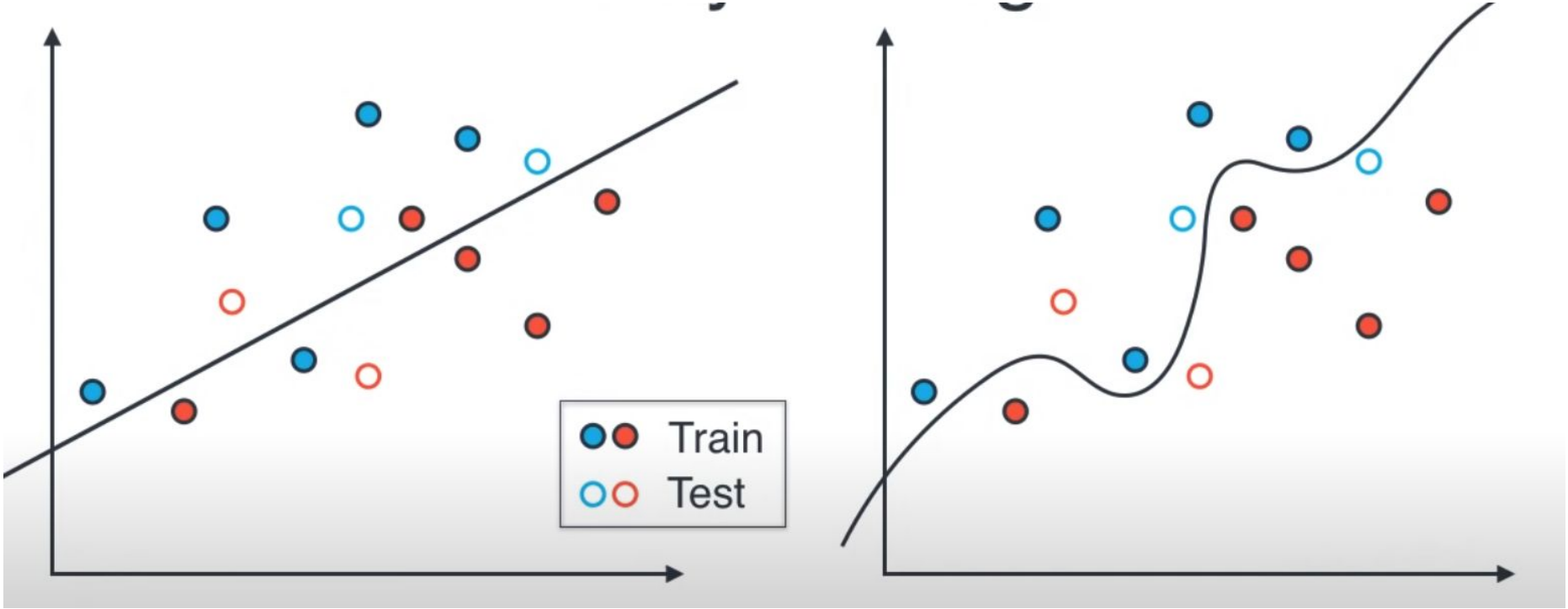
Prof. Dr. Fazlul Hasan Siddiqui
Office: DUET | **Education**: BSc: IUT; MSc: BUET; PhD: ANU
Source: https://youtu.be/aDW44NPhNw0 | Serrano Academy

Which model is better?

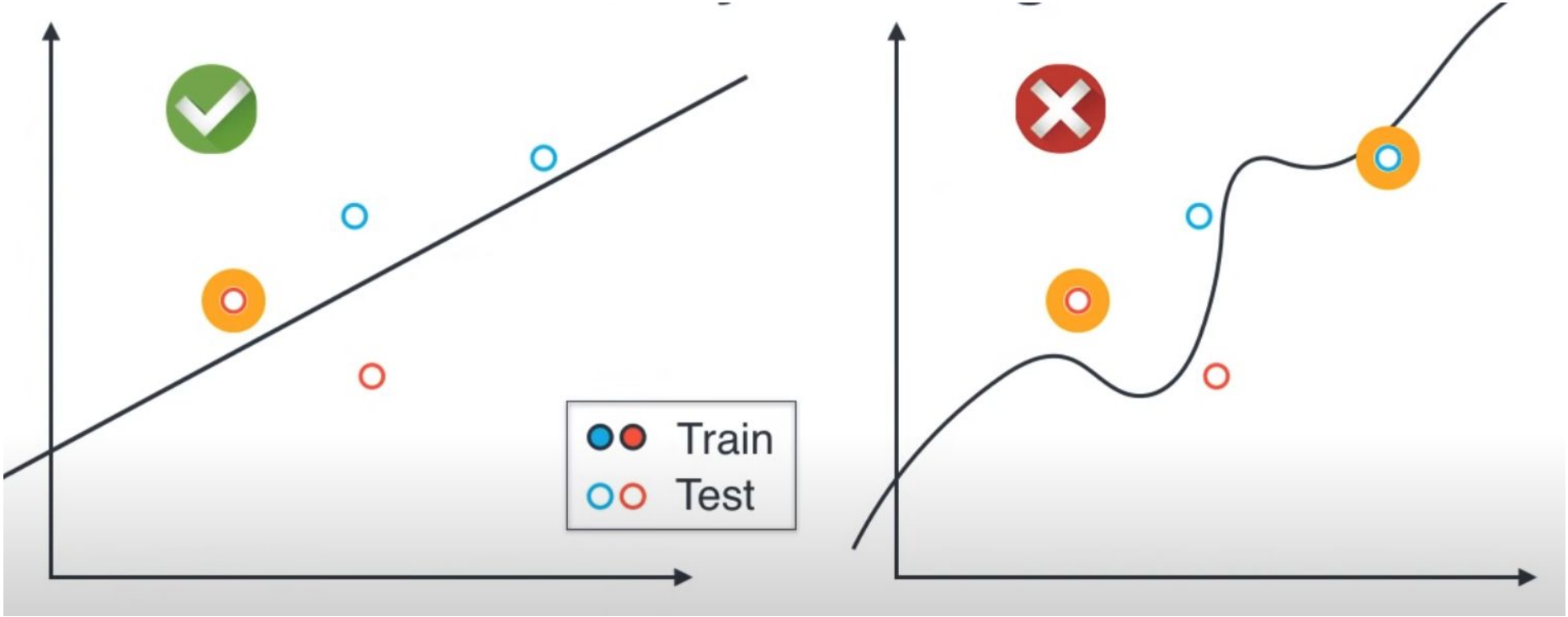




Which model is better? Model over train data

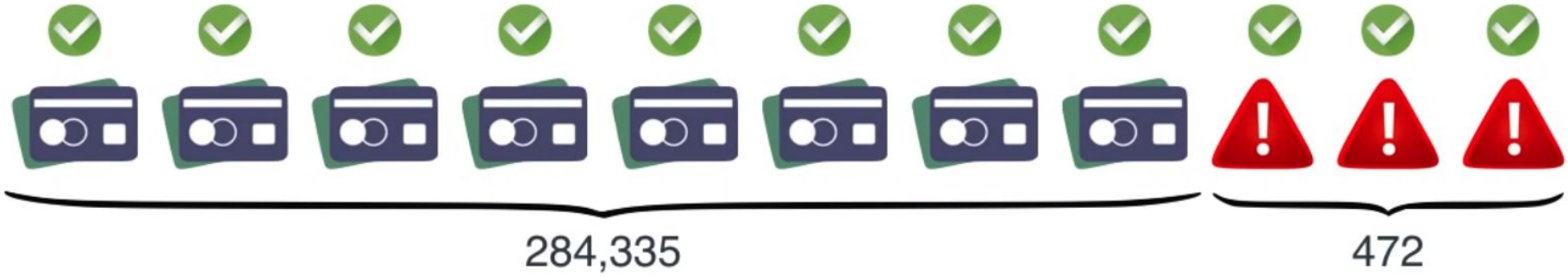


Which model is better? Test over "Test" data



K-Fold Cross Validation – Don’t loose training data

How well is my model? Credit Card Fraud



Model: All transactions are good.

$$\text{Correct} = \frac{284{,}335}{284{,}807} = 99.83\%$$

Problem: I'm not catching any of the bad ones!

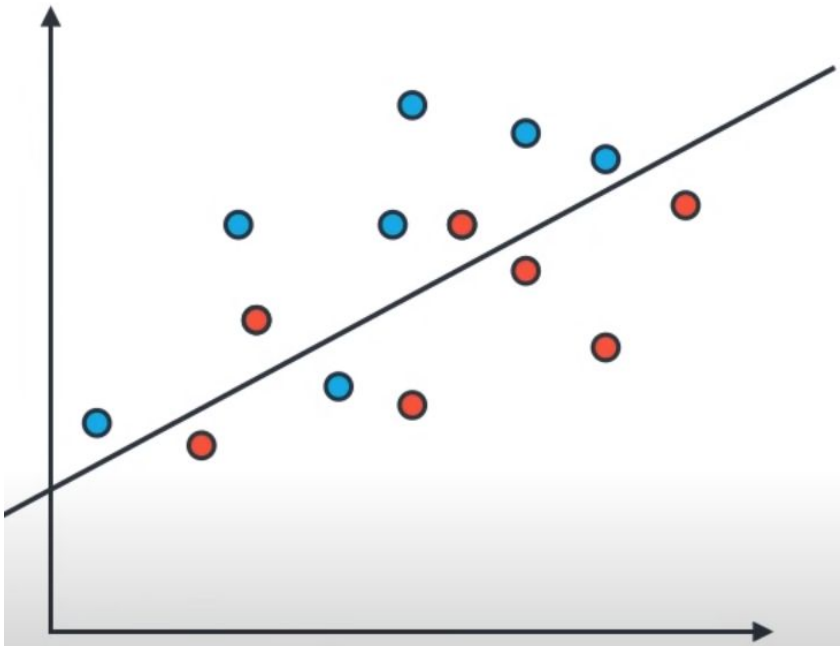
How well is my model? Credit Card Fraud

Model: All transactions are fraudulent.

Great! Now I'm catching *all* the bad transactions!

Problem: I'm accidentally catching all the good ones!

Model Evaluation | Confusion Matrix



Prediction

Data	Guessed Positive	Guessed Negative
Positive	6 True positives	1 False Negatives
Negative	2 False Positives	5 True Negatives

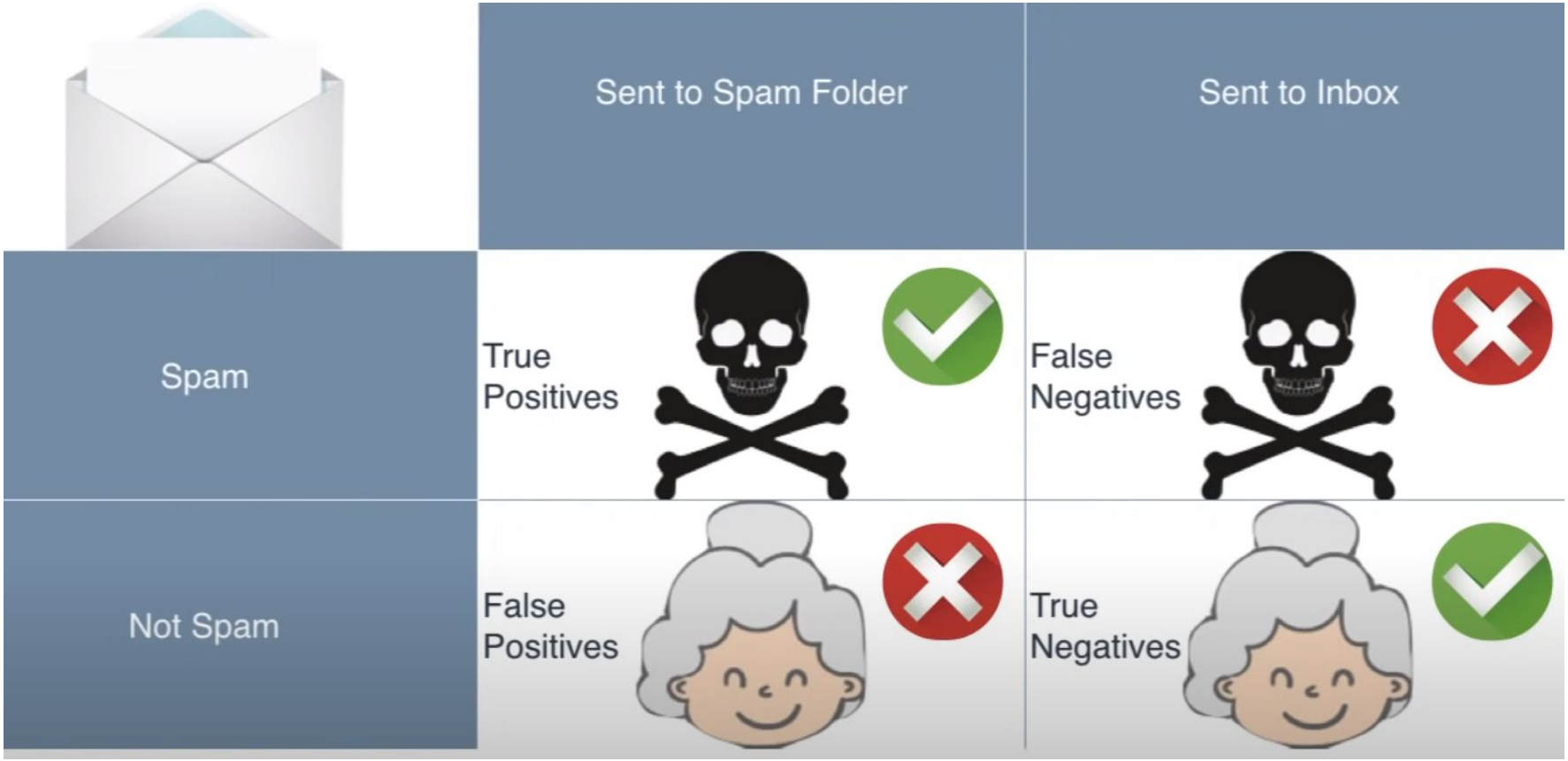
Model Evaluation | Confusion Matrix | Medical Diagnosis

	Diagnosed Sick	Diagnosed Healthy
Sick	True positive  	False Negative  
Healthy	False Positive  	True Negative  

Model Evaluation | Confusion Matrix | Medical Diagnosis

	Diagnosed Sick	Diagnosed Healthy
Sick		False Negative 
Healthy	False Positive 	

Model Evaluation | Confusion Matrix | Spam Detect



Model Evaluation | Confusion Matrix | Spam Detect

	Sent to Spam Folder	Sent to Inbox
Spam		False Negatives  
Not Spam	False Positives 	

Model Evaluation | Confusion Matrix | Accuracy

Diagnosis

Patients	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Healthy	800	8000

Accuracy: Out of the all the patients, how many did we classify correctly?

$$\text{Accuracy} = \frac{1{,}000 + 8{,}000}{10{,}000} = 90\%$$

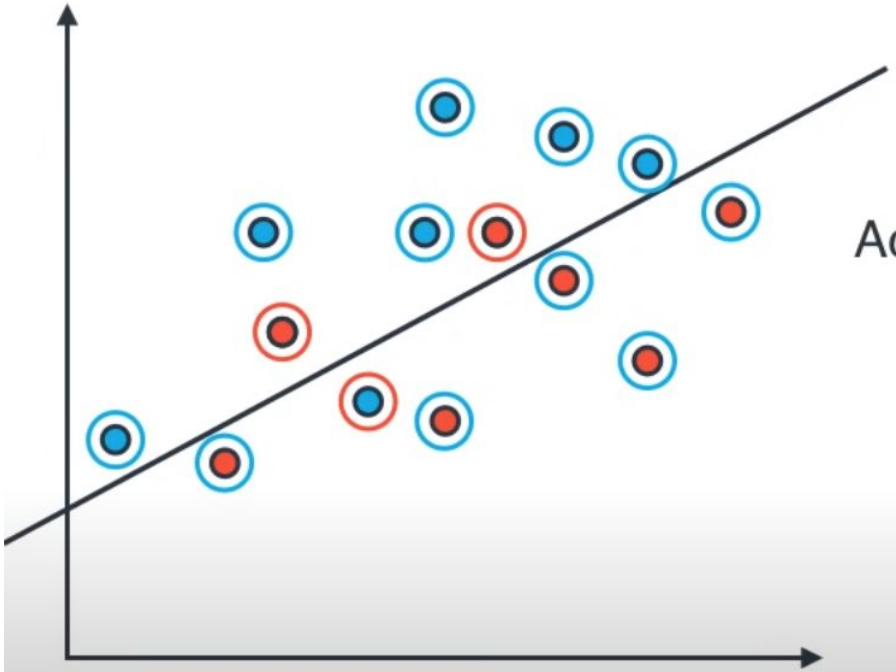
Folder

E-mail	Spam Folder	Inbox
Spam	100	170
Not spam	30	700

Accuracy: Out of the all the e-mails, how many did we classify correctly?

$$\text{Accuracy} = \frac{100 + 700}{1000} = 80\%$$

Model Evaluation | Confusion Matrix | Accuracy



Precision: Out of all the data, how many points did we classify correctly?

$$\text{Accuracy} = \frac{\text{Correctly Classified points}}{\text{All points}}$$

$$= \frac{11}{11 + 3}$$

$$= \frac{11}{14}$$

$$= 78.57\%$$

Model Evaluation | Confusion Matrix | Recall & Precision

Medical Model

False positives ok
False negatives **NOT** ok

Find all the sick people
Ok if not all are sick

High Recall

Spam Detector

False positives **NOT** ok
False negatives ok

You don't necessarily need to find all spam
But they better all be spam

High Precision

Model Evaluation | Confusion Matrix | Precision

Diagnosis

Patients	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Healthy	800	8000

Precision: Out of the patients we diagnosed with an illness, how many did we classify correctly?

$$\text{Precision} = \frac{1{,}000}{1{,}000 + 800} = 55.7\%$$

Folder

E-mail	Spam Folder	Inbox
Spam	100	170
Not spam	30	700

Precision: Out of the all the e-mails, sent to the spam inbox, how many were actually spam?

$$\text{Precision} = \frac{100}{100 + 30} = 76.9\%$$

Model Evaluation | Confusion Matrix | Recall

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1000	200 ✗
Is Healthy	800	8000

Recall: Out of the sick patients, how many did we correctly diagnose as sick?

$$\text{Recall} = \frac{1{,}000}{1{,}000 + 200} = 83.3\%$$

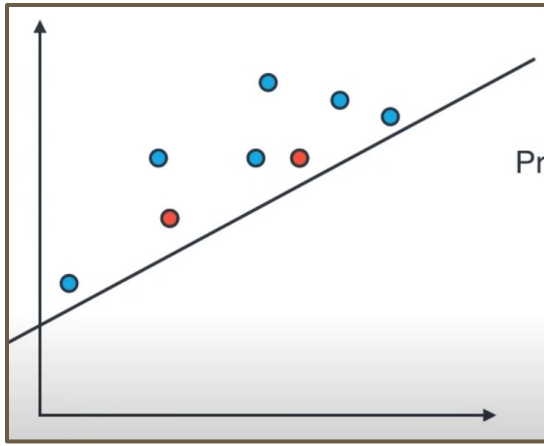
Folder

E-mail	Spam Folder	Inbox
Spam	100	170
Not spam	30 ✗	700

Recall: Out of the all the spam e-mails, how many were correctly sent to the spam folder?

$$\text{Recall} = \frac{100}{100 + 170} = 37\%$$

Model Evaluation | Precision & Recall



Precision: Out of the points we've predicted to be positive, how many are correct?

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False Positives}} = \frac{6}{6 + 2} = \frac{6}{8} = 75\%$$

Spam Detector

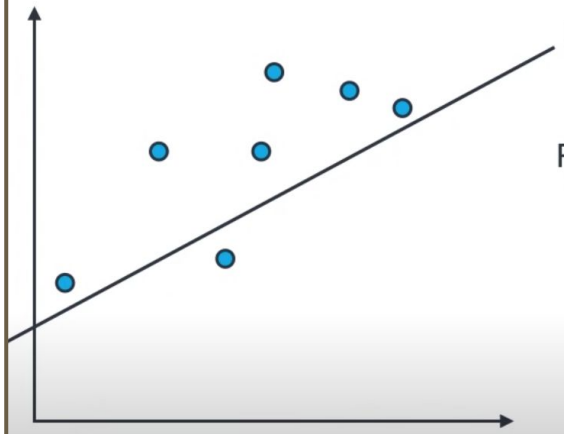
Precision: 76.9%

Recall: 37%

Medical Model

Precision: 55.7%

Recall: 83.3%



Recall: Out of the points labelled positive, how many did we correctly predict?

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False Negatives}} = \frac{6}{6 + 1} = \frac{6}{7} = 85.7\%$$

Model Evaluation | One Score – Average !!

Medical Model

Precision: 55.7%
Recall: 83.3%

Average = 69.5%

Spam Detector

Precision: 76.9%
Recall: 37%

Average = 56.95%

Model Evaluation | One Score – Average !!

Model: All transactions are good.

Precision = 100%

$$\text{Recall} = \frac{0}{472} = 0\%$$

Average = 50%

Model Evaluation | One Score – Average !!

Model: All transactions are fraudulent.

$$\text{Precision} = \frac{472}{284{,}807} = .016\%$$

$$\text{Recall} = \frac{472}{472} = 100\%$$

Average = 50.008%

Model Evaluation | One Score – F1 Score

$$\text{Arithmetic Mean} = \frac{x+y}{2}$$

$$\text{Harmonic Mean} = \frac{2xy}{x+y}$$

Precision = 1
Recall = 0
Average = 0.5
Harmonic Mean = 0

Precision = 0.2
Recall = 0.8
Average = 0.5
Harmonic Mean = 0.32

~~Arithmetic Mean(Precision, Recall)~~

F1 Score = Harmonic Mean(Precision, Recall)

Model Evaluation | One Score – F1 Score

Medical

Precision = 55.7%

Recall = 83.3%

Average = 69.5%

$$\text{F1 Score} = \frac{2 \times 55.7 \times 83.3}{55.7 + 83.3} = 66.76\%$$

Spam

Precision = 76.9%

Recall = 37%

Average = 56.95%

$$\text{F1 Score} = \frac{2 \times 76.9 \times 37}{76.9 + 37} = 49.96\%$$

Linear

Precision = 75%

Recall = 85.7%

Average = 80.35

$$\text{F1 Score} = \frac{2 \times 75 \times 85.7}{75 + 85.7} = 80\%$$

Model Evaluation | One Score – F1 Score

Model: All transactions are good.

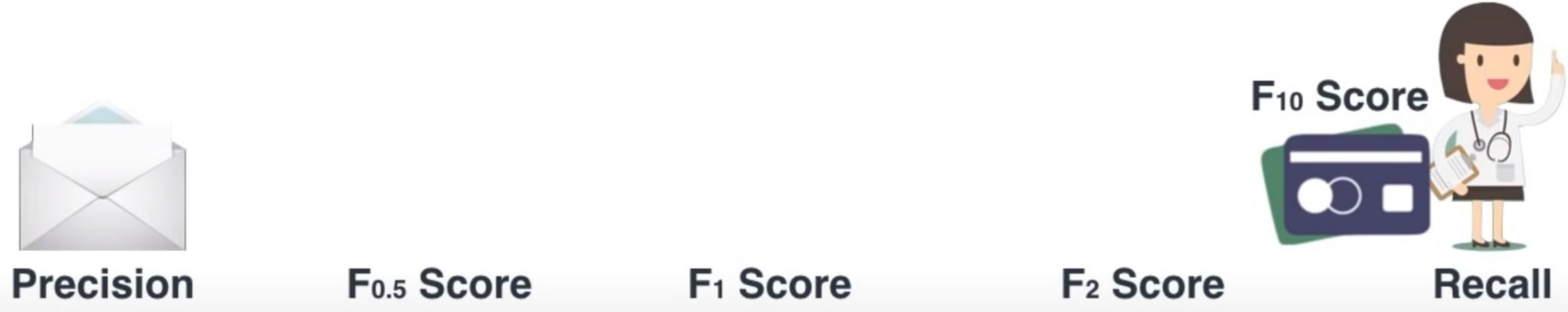
Precision = 100%

$$\text{Recall} = \frac{0}{472} = 0\%$$

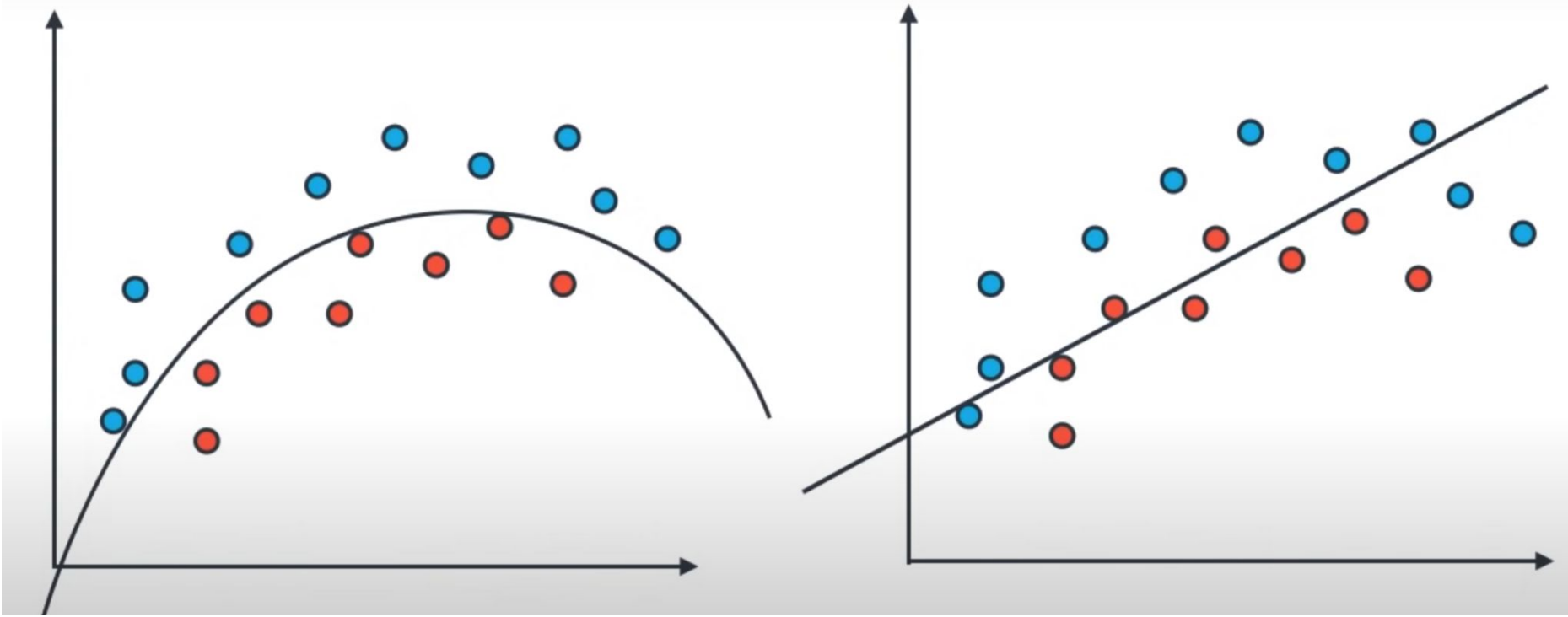
Average = 50%

F1 Score = 0

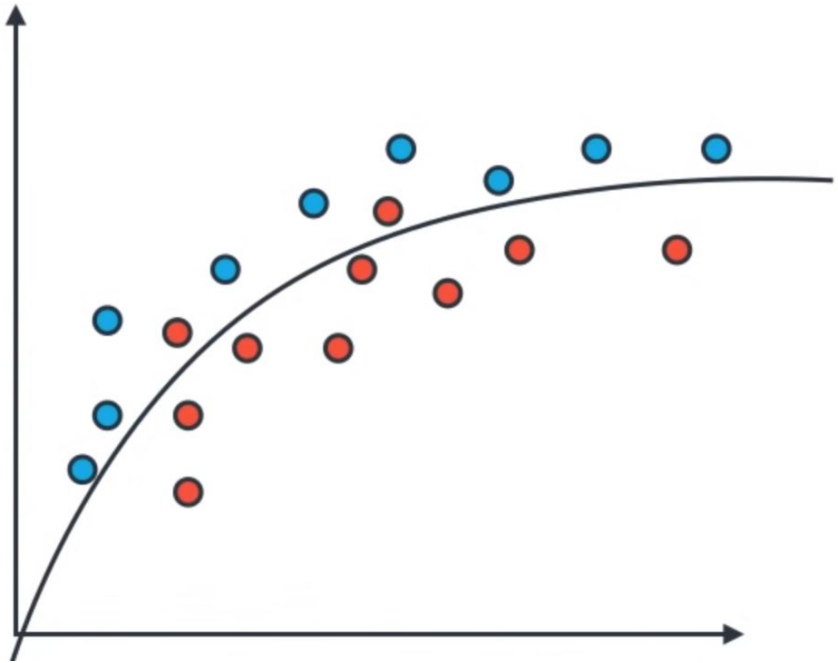
Model Evaluation | One Score – F_w Score

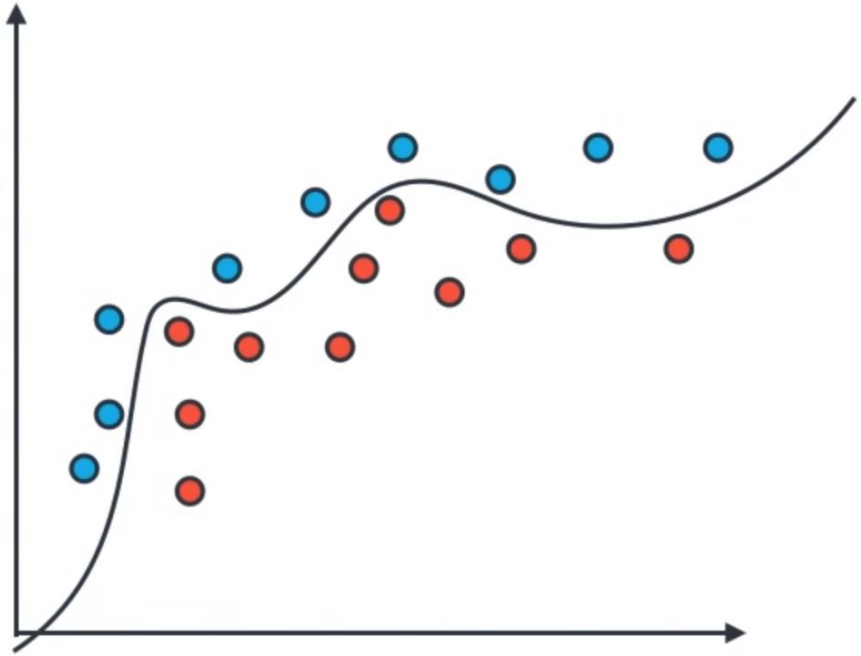


Model Evaluation | Error for Bias (Underfitting)

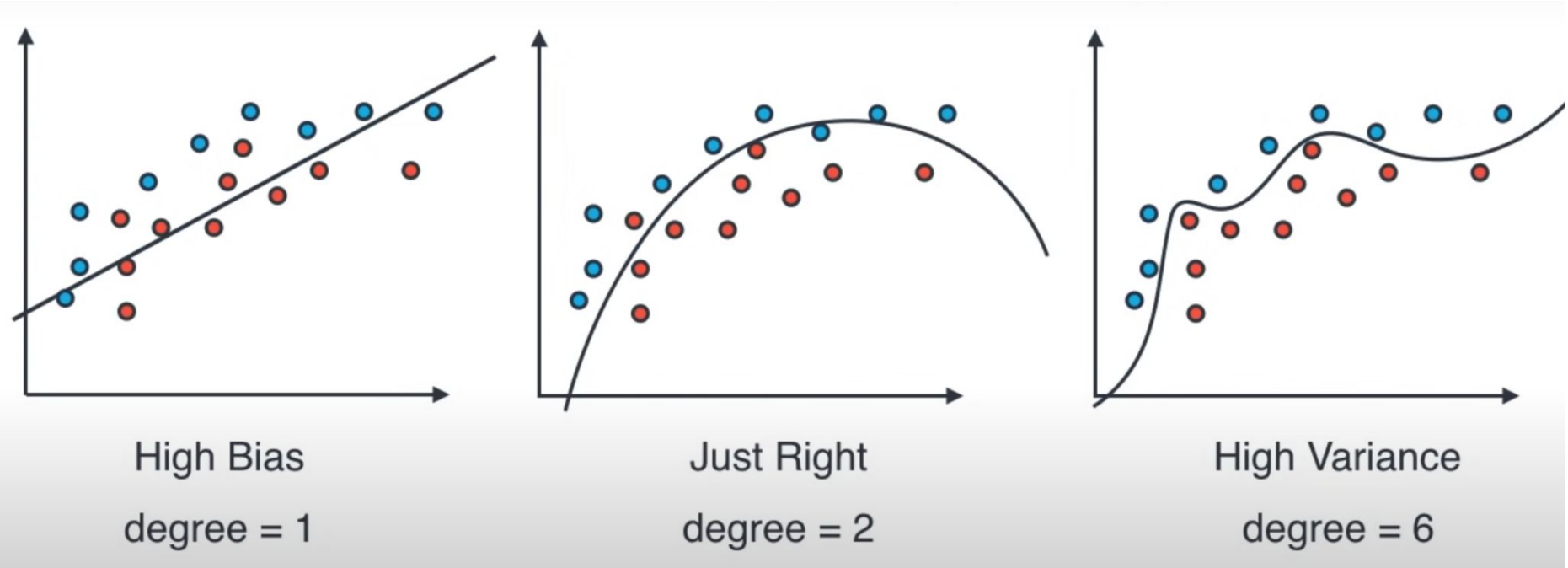


Model Evaluation | Error for Variance (Overfitting)

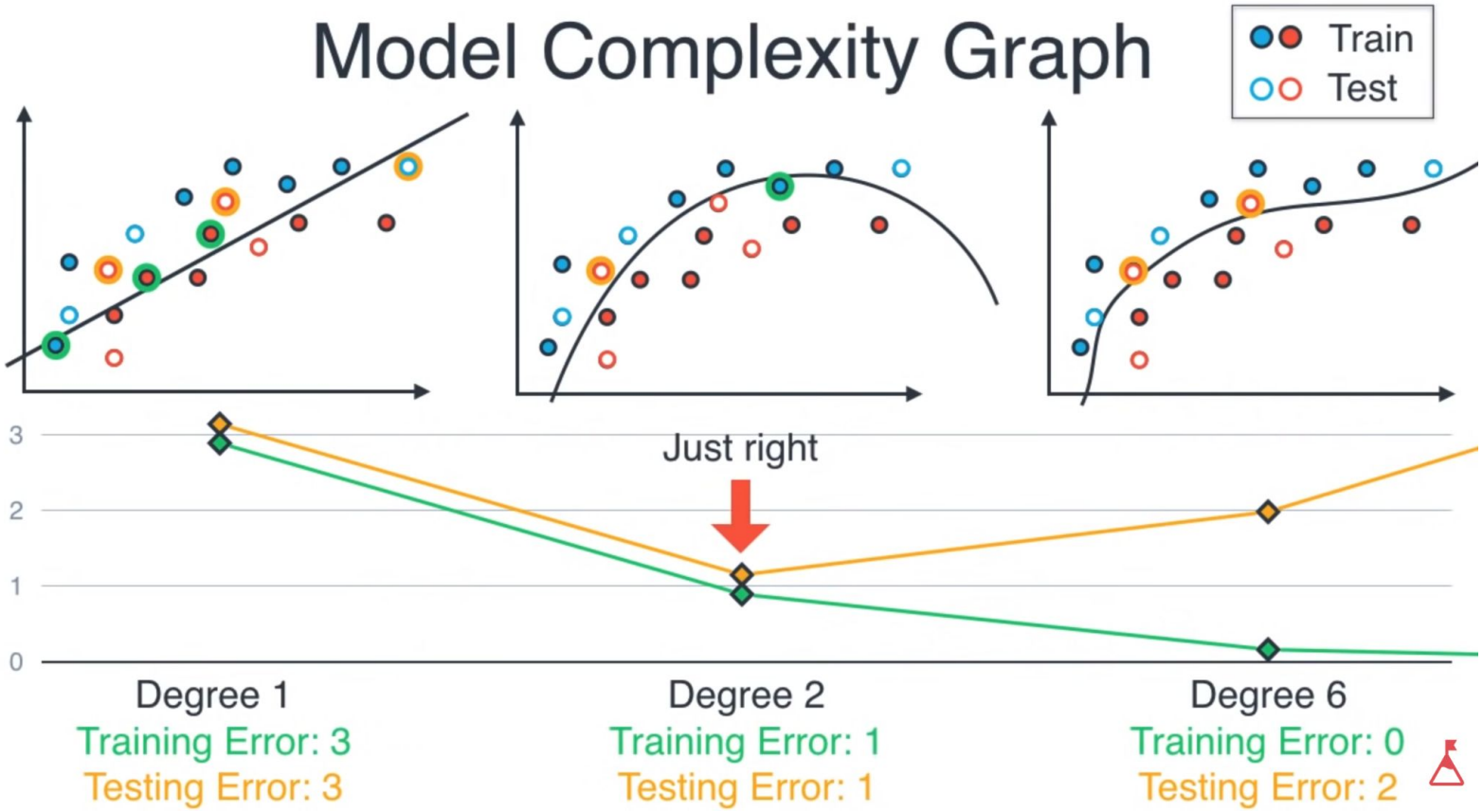




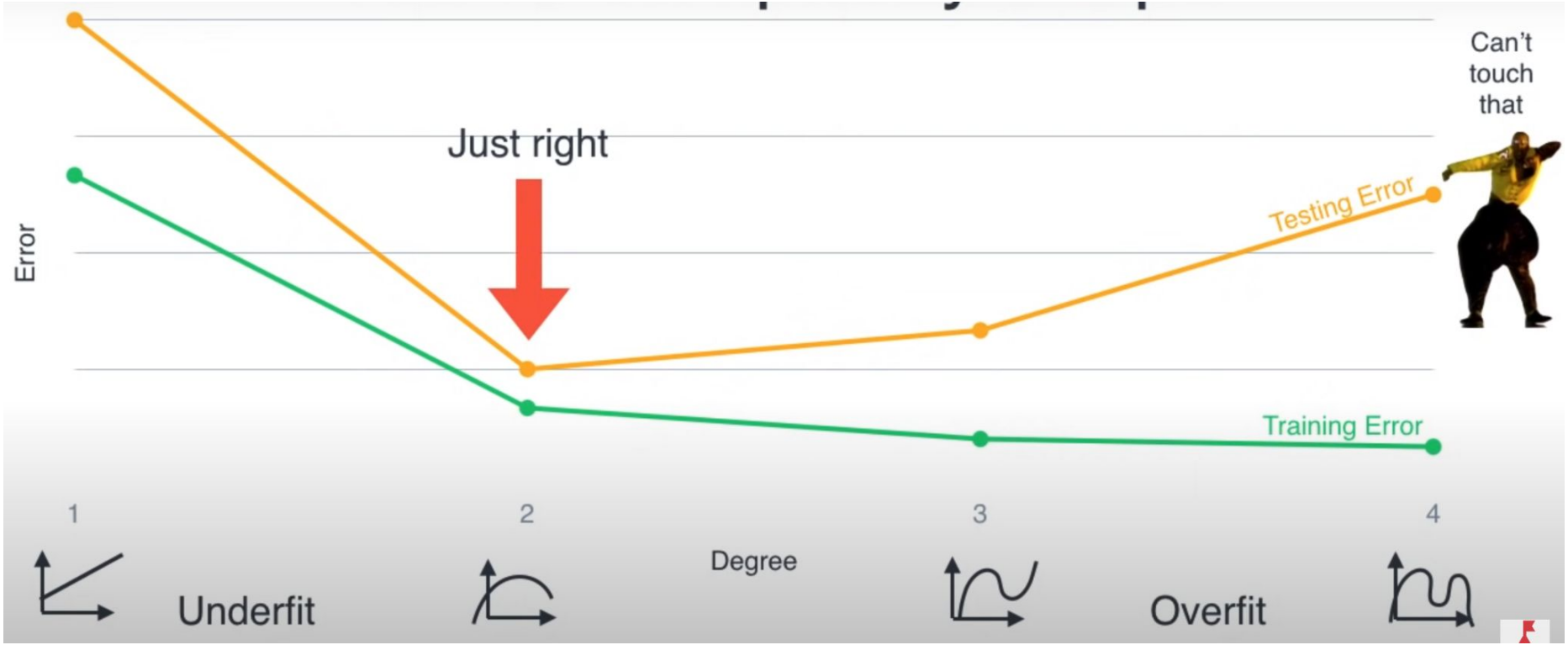
Model Evaluation | Model Complexity Graph



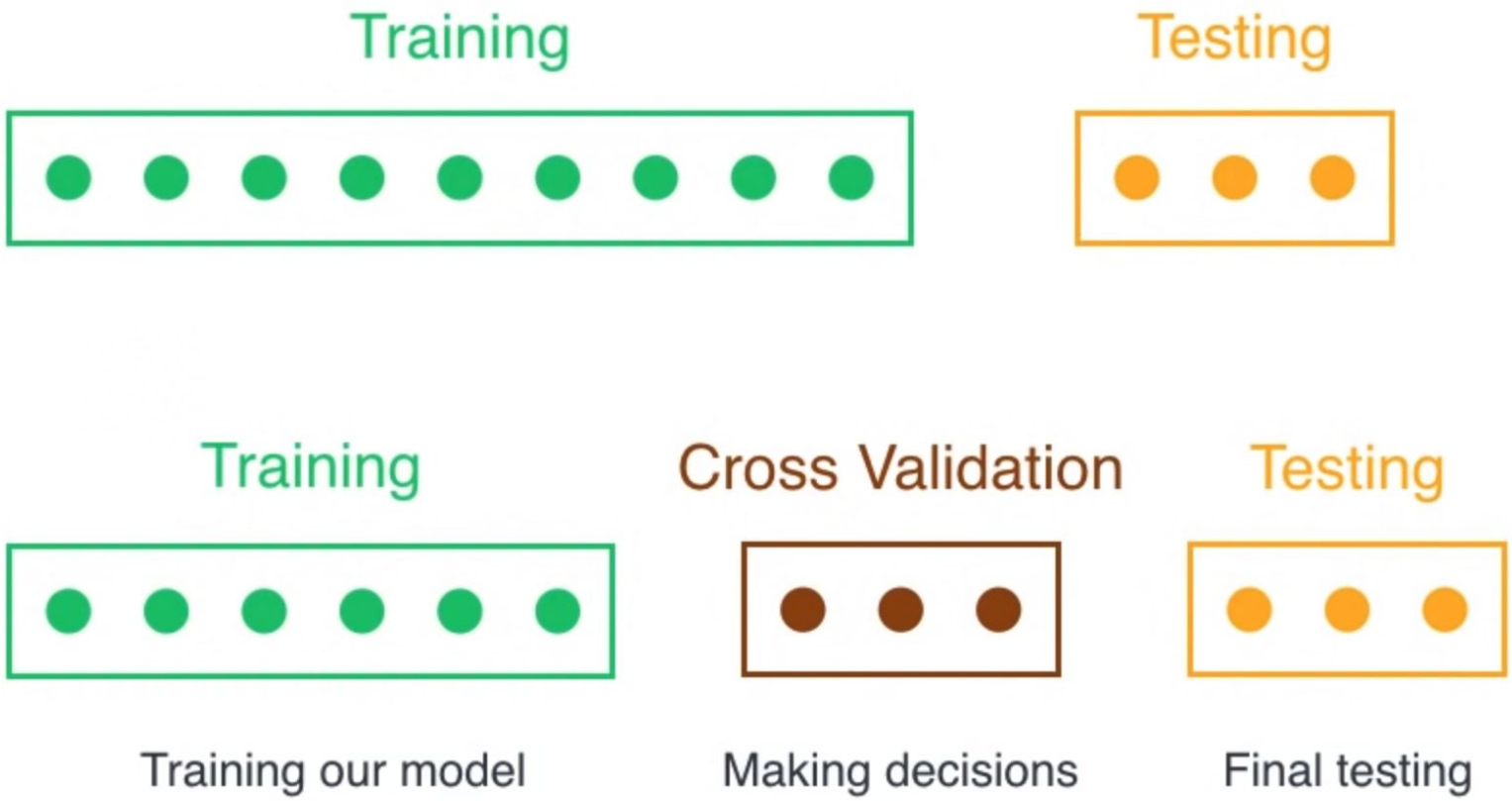
Model Evaluation



Model Evaluation | Model Complexity Graph

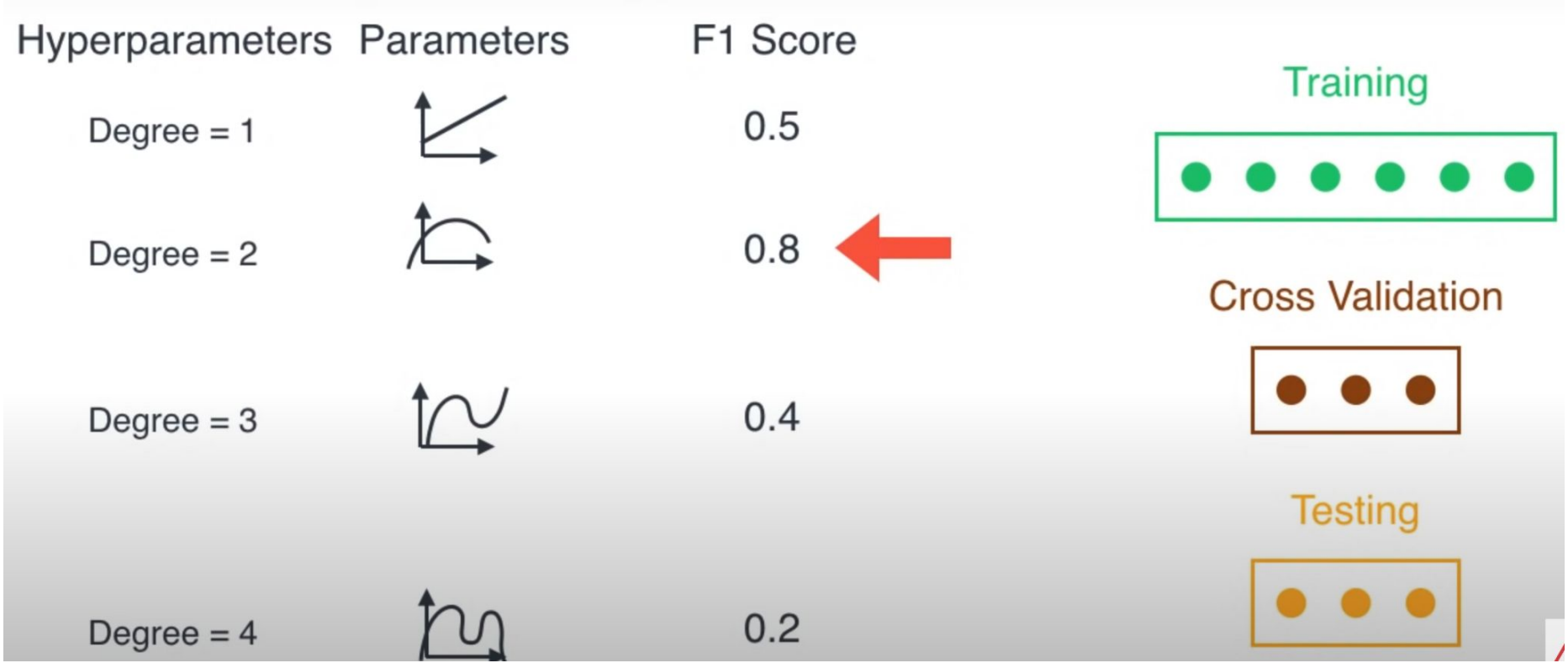


Model Evaluation | Complexity Graph | Cross validation

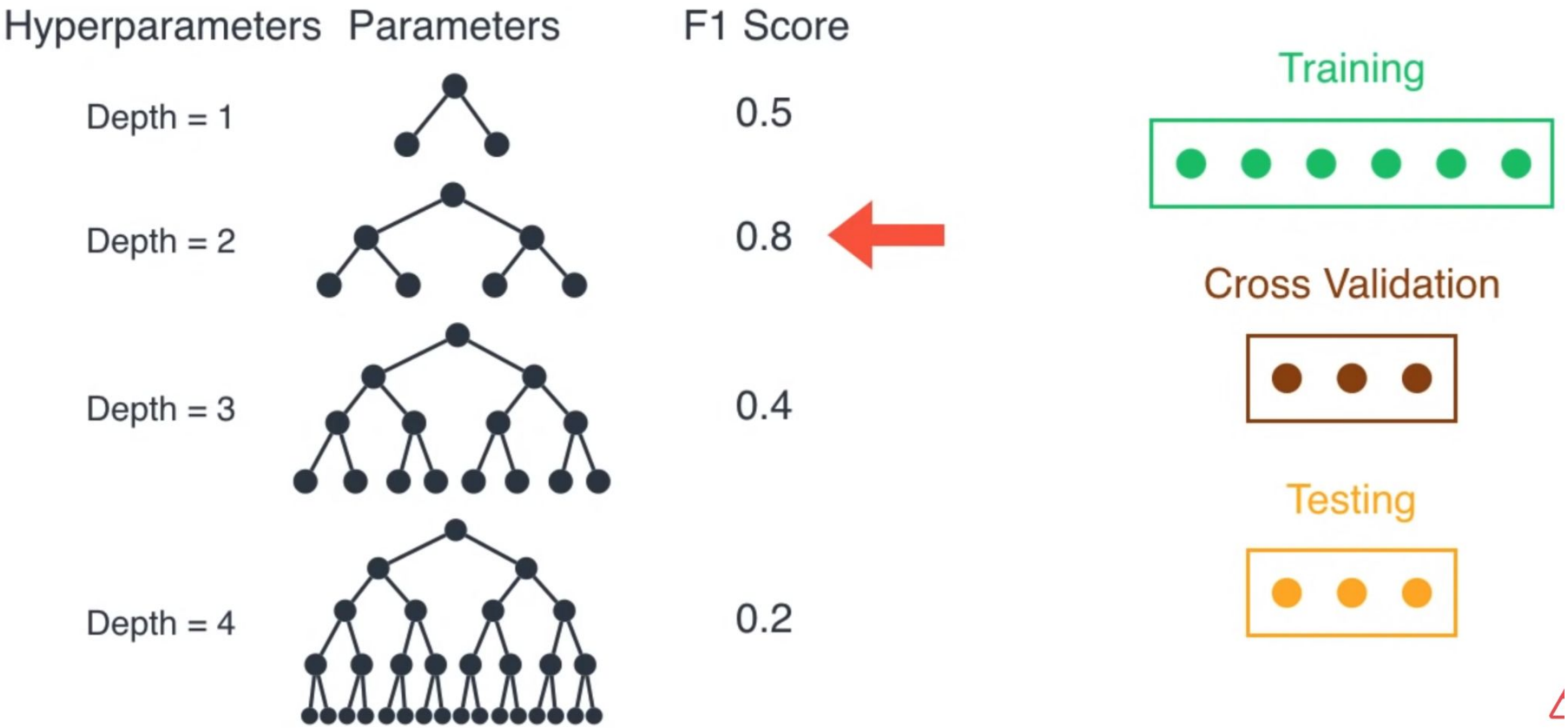


Model Evaluation | Complexity Graph | Cross validation

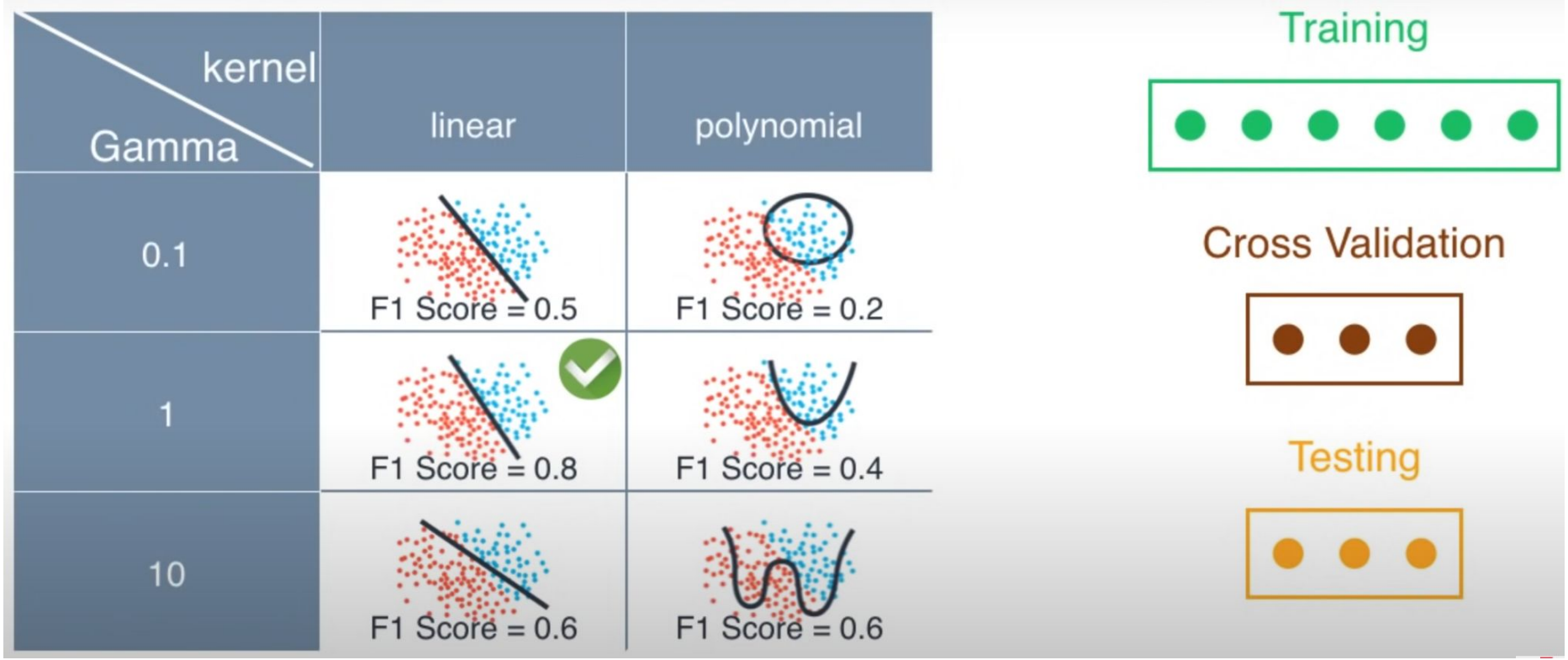
Model Complexity Graph | Logistic Regression



Model Complexity Graph | Decision Tree



Model Complexity Graph | Support Vector Machines



Model Evaluation

Algorithm	Parameters	Hyperparameters
Random Forest	Features Thresholds	Number of trees Depth
Logistic Regression	Coefficients of the polynomial	Degree of the polynomial
Support Vector Machines	Coefficients	Kernel Gamma C
Neural Networks	Coefficients	Number of layers Size of layers Activation function

How to solve a problem

Problem

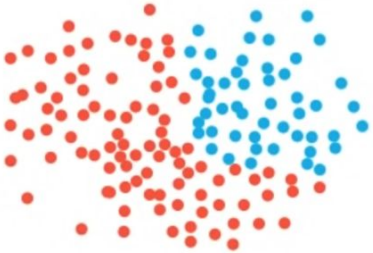
Tools

Measurement Tools

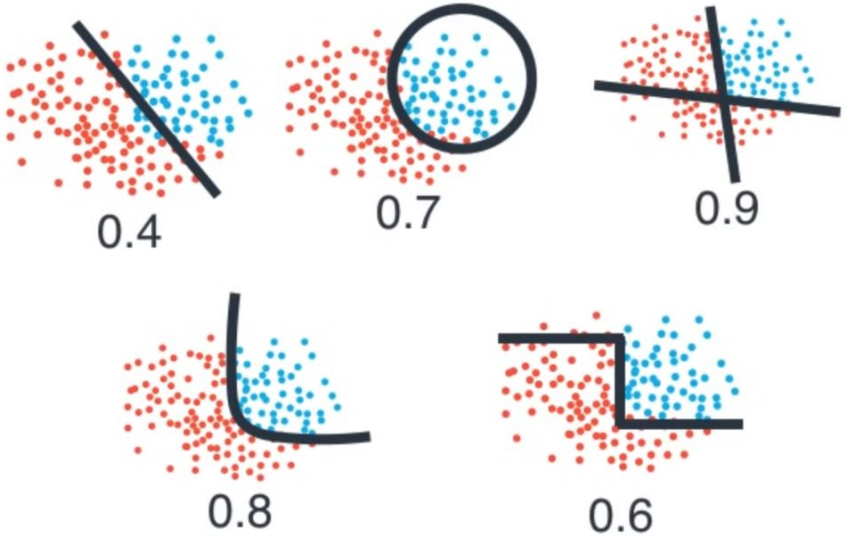
Measure each tool's performance

Pick the best tool

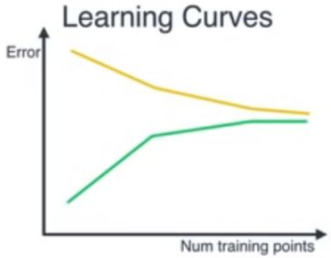
How to use ML to solve a problem



Data



Algorithms



Metrics