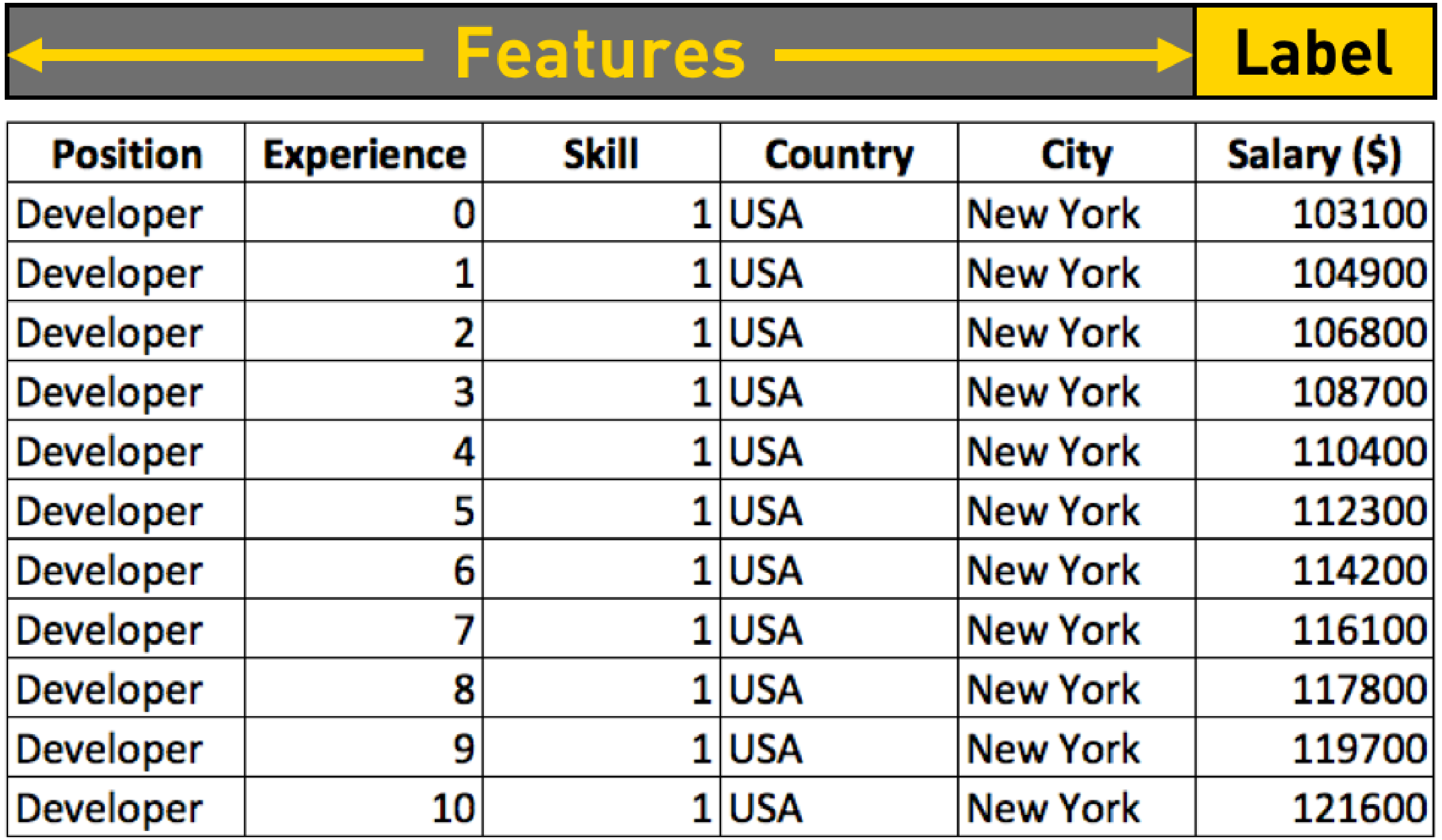
Basic ML Terminologies

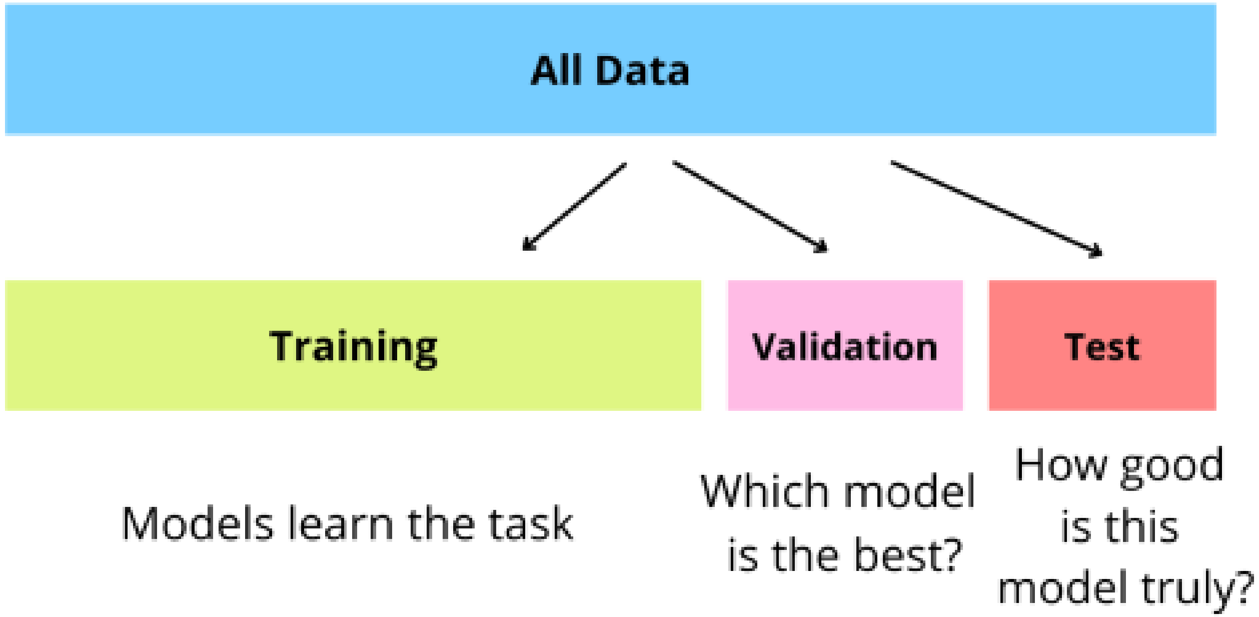
Week 03 4th-Sep-2024

Features and Labels



Features					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Training/Validation/Testing sets



All Data

Training

Validation

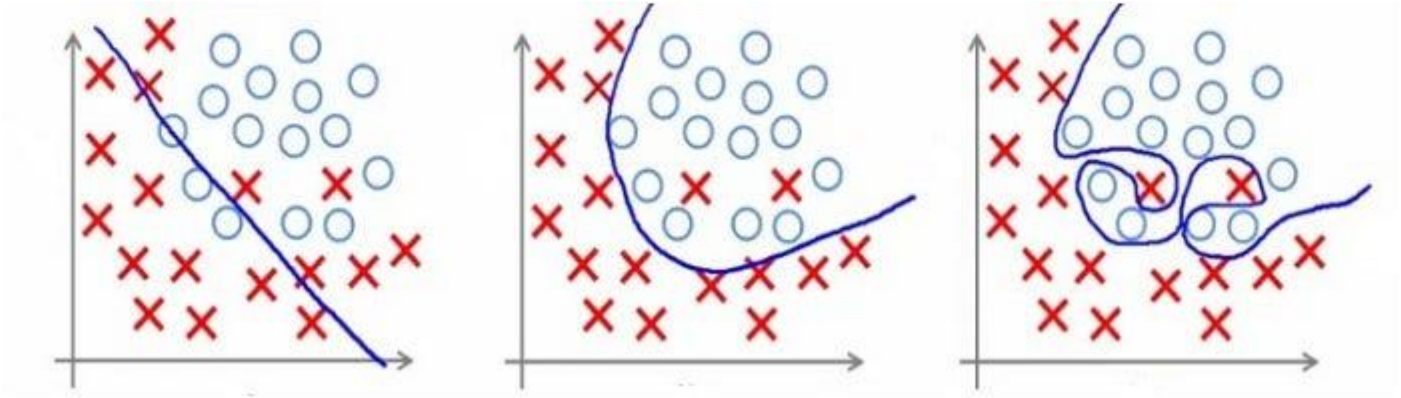
Test

Models learn the task

Which model is the best?

How good is this model truly?

Overfitting and Underfitting



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

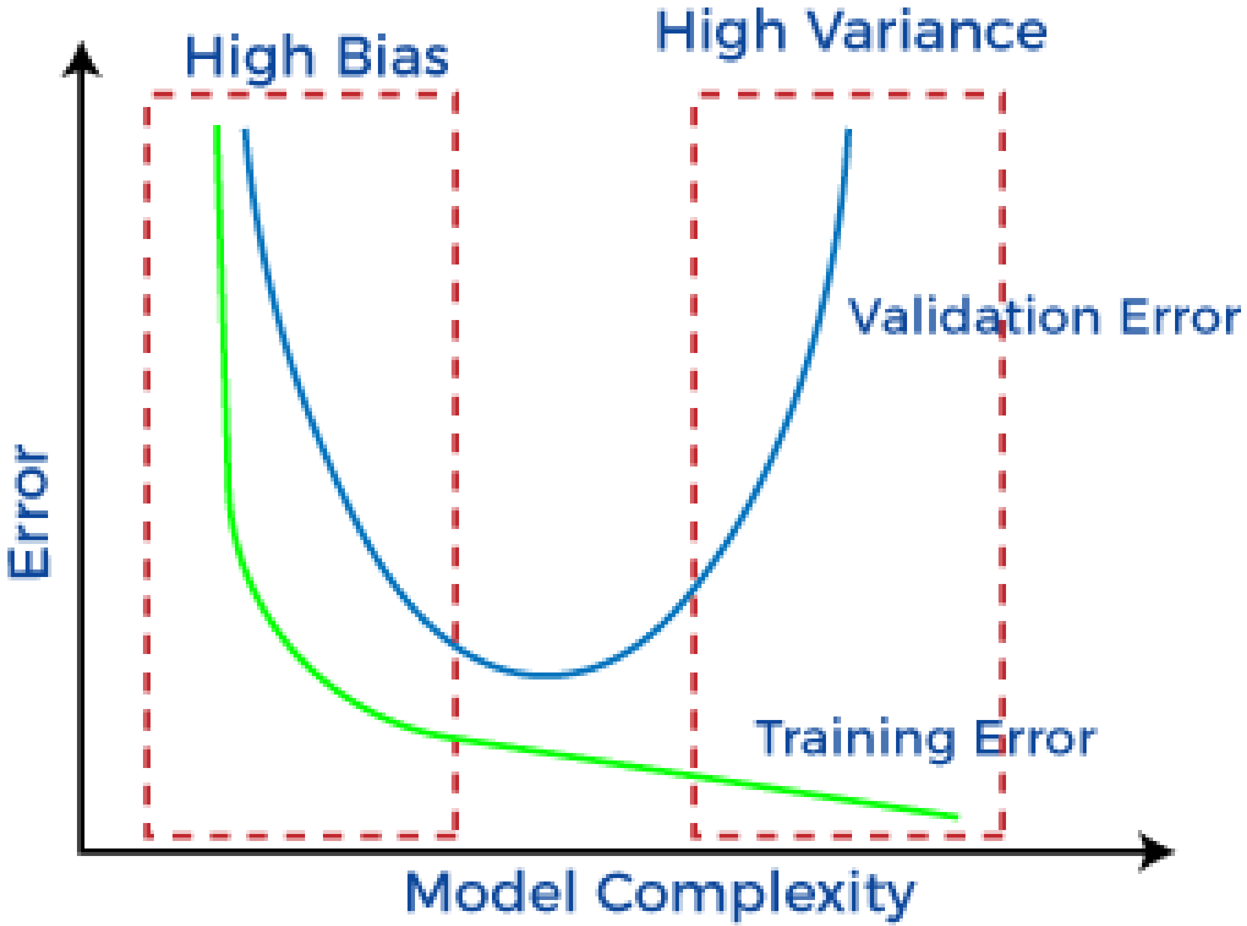
(forcefitting -- too good to be true)

Bias-Variance Tradeoff

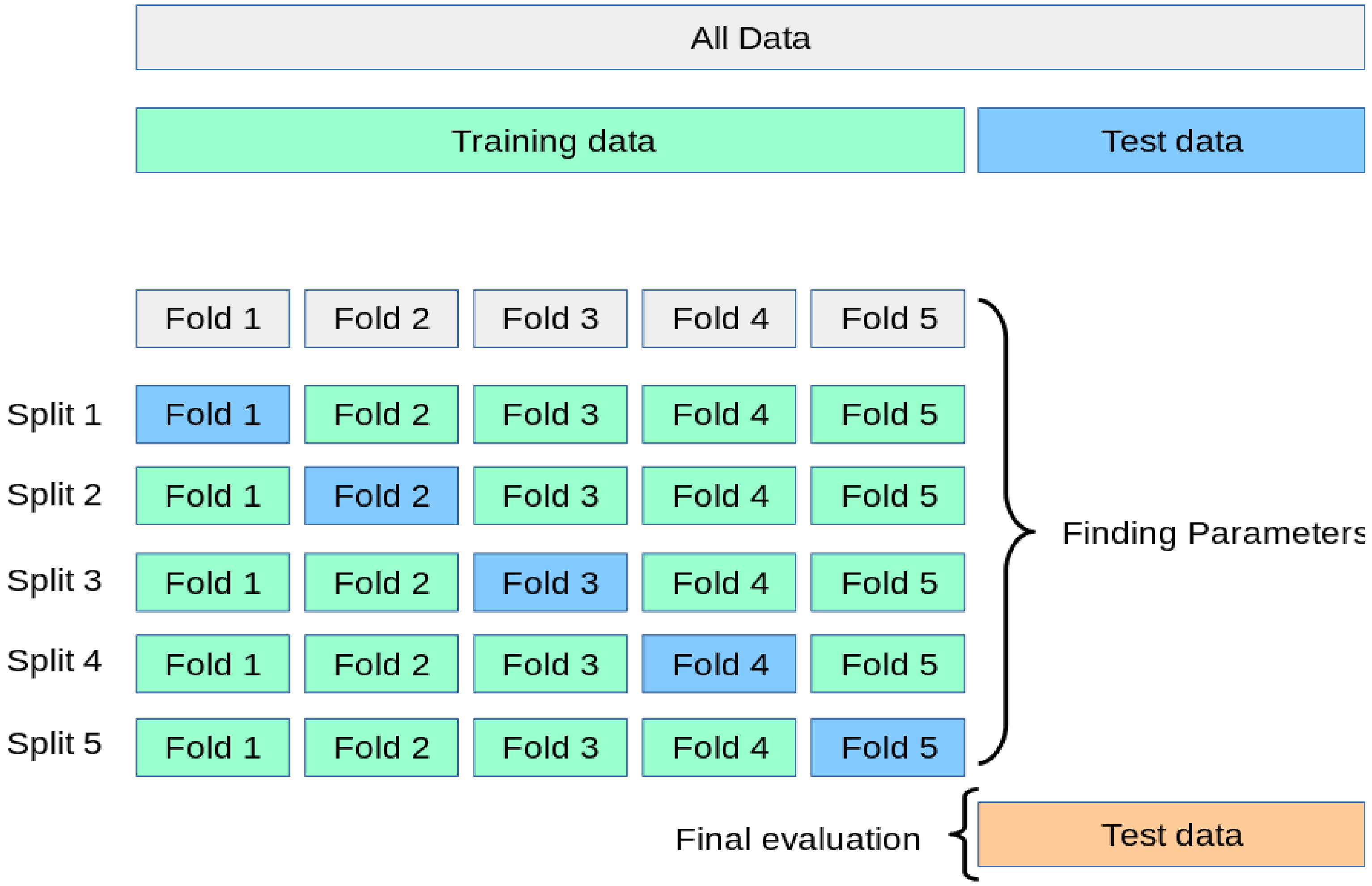
- ***Bias:*** The error due to overly simplistic assumptions in the learning algorithm. High bias can cause the model to underfit the data.

- ***Variance:*** The error due to excessive sensitivity to small fluctuations in the training data. High variance can cause the model to overfit the data.

- ***Tradeoff:*** There’s a balance between bias and variance. Reducing one typically increases the other, and the goal is to find the right balance for the best model performance.



Cross Validation



All Data

Training data | Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

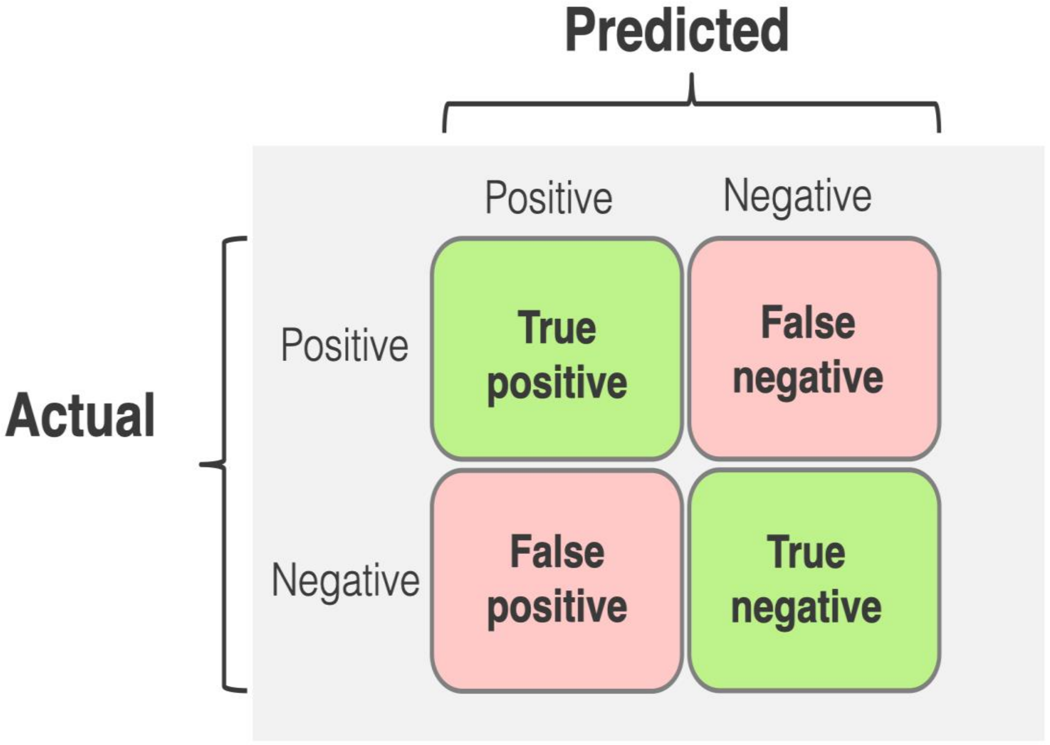
Finding Parameters

Final evaluation

Test data

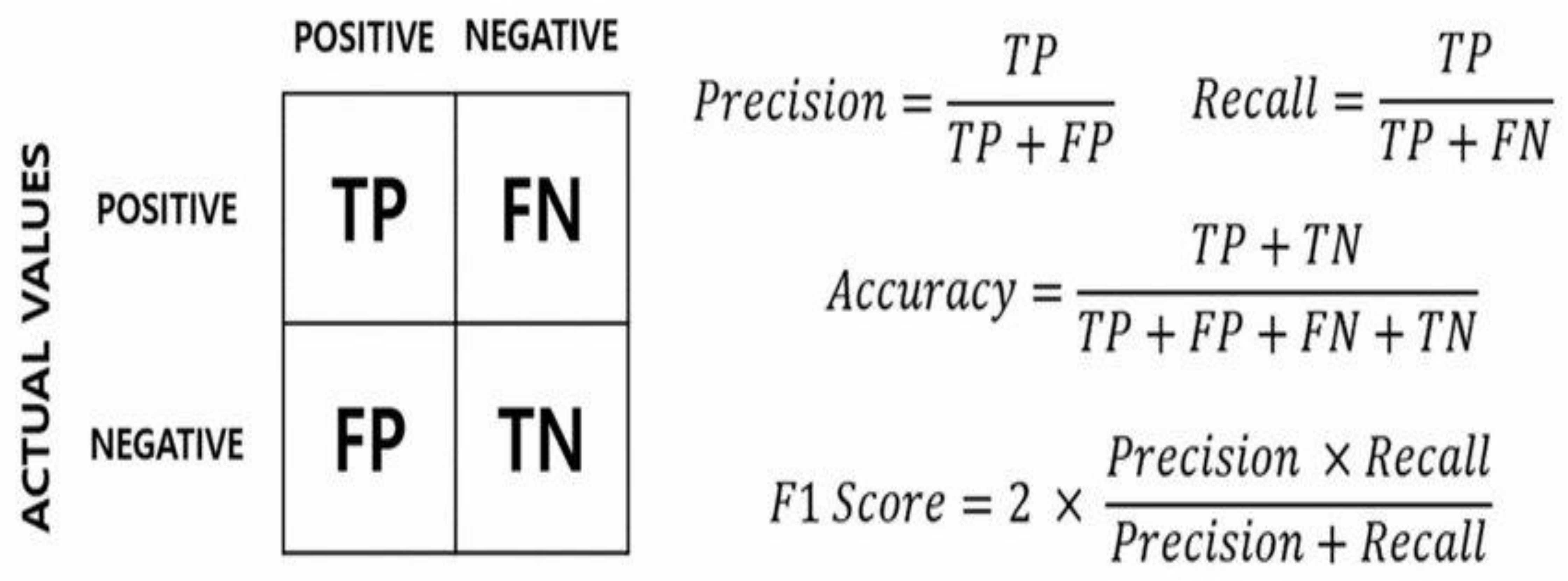
Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It provides insight into the true positives, false positives, true negatives, and false negatives made by the model, helping to calculate metrics like accuracy, precision, recall, and F1 score.



		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

Accuracy, Precision, Recall, and F1 Score



		POSITIVE	NEGATIVE
ACTUAL VALUES	POSITIVE	TP	FN
	NEGATIVE	FP	TN

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

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

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$