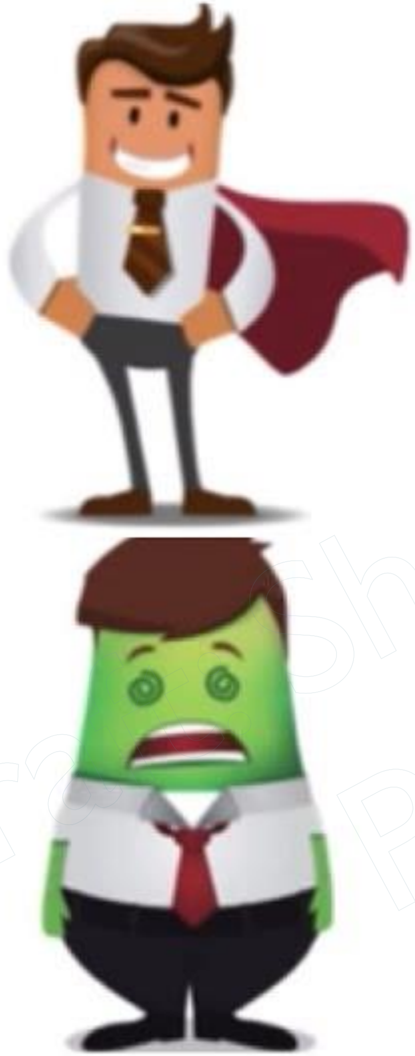
ML Testing and Error Metrics

False Positive

False Negative



Testing

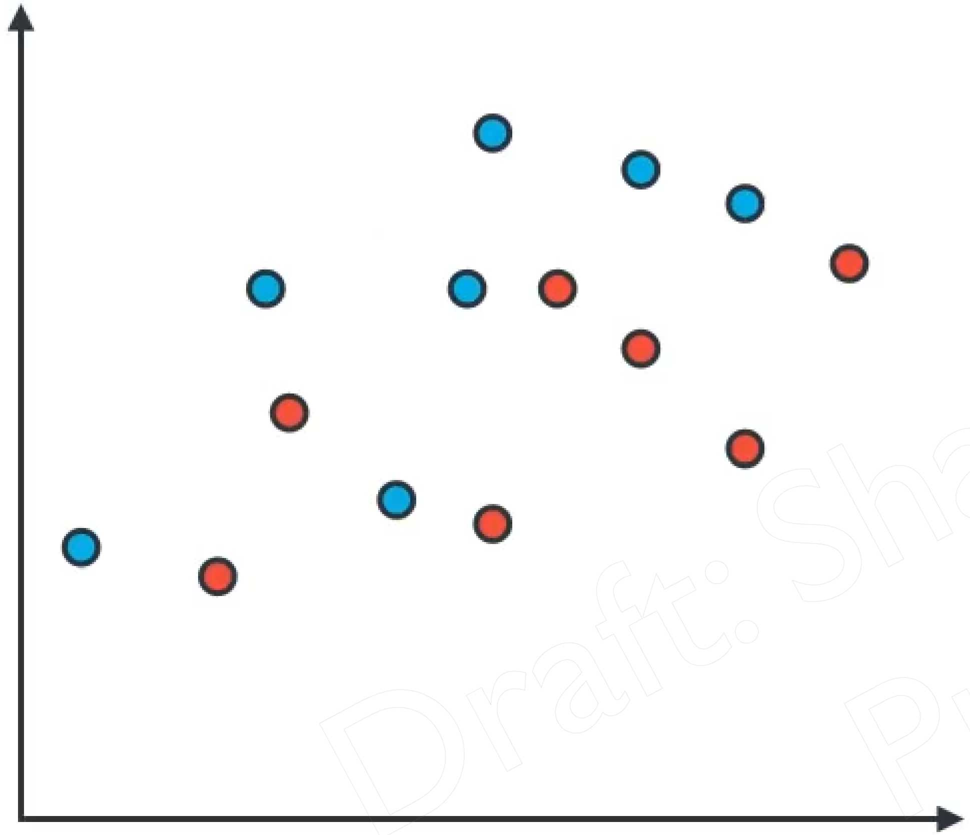
How well is my model doing?

Testing

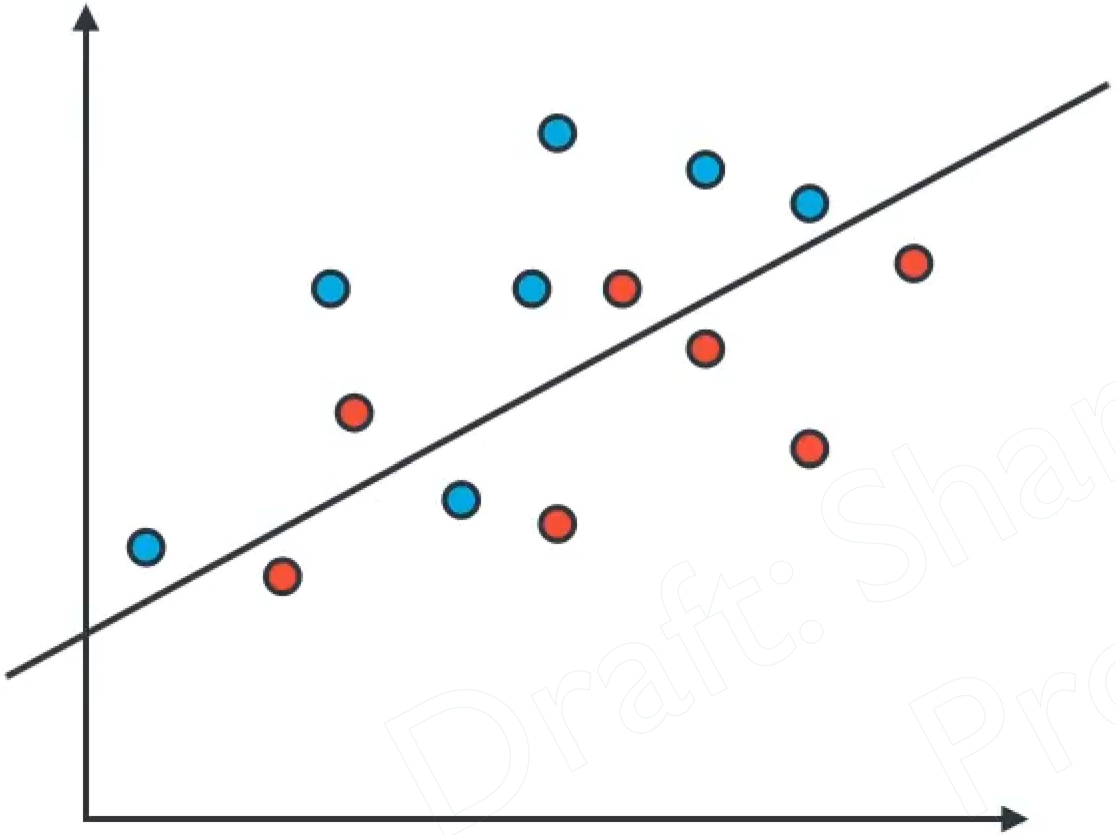
How well is my model doing?

How do I improve it?

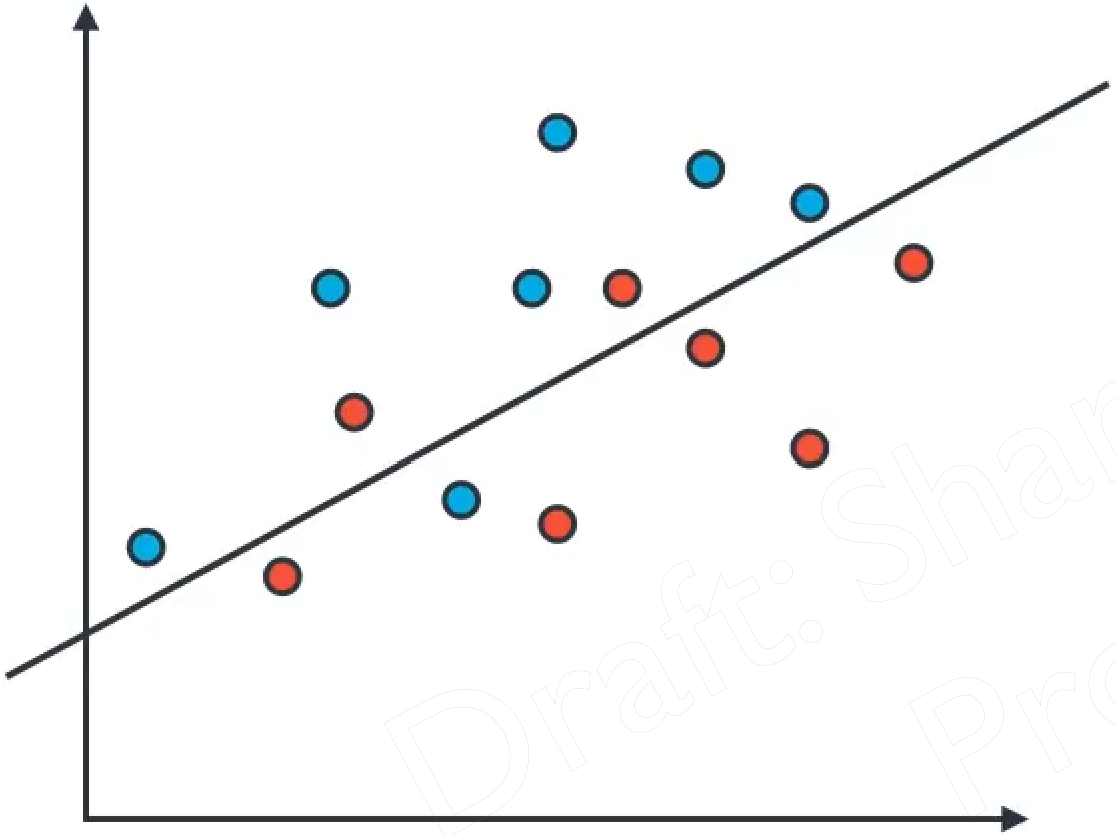
Which model is better

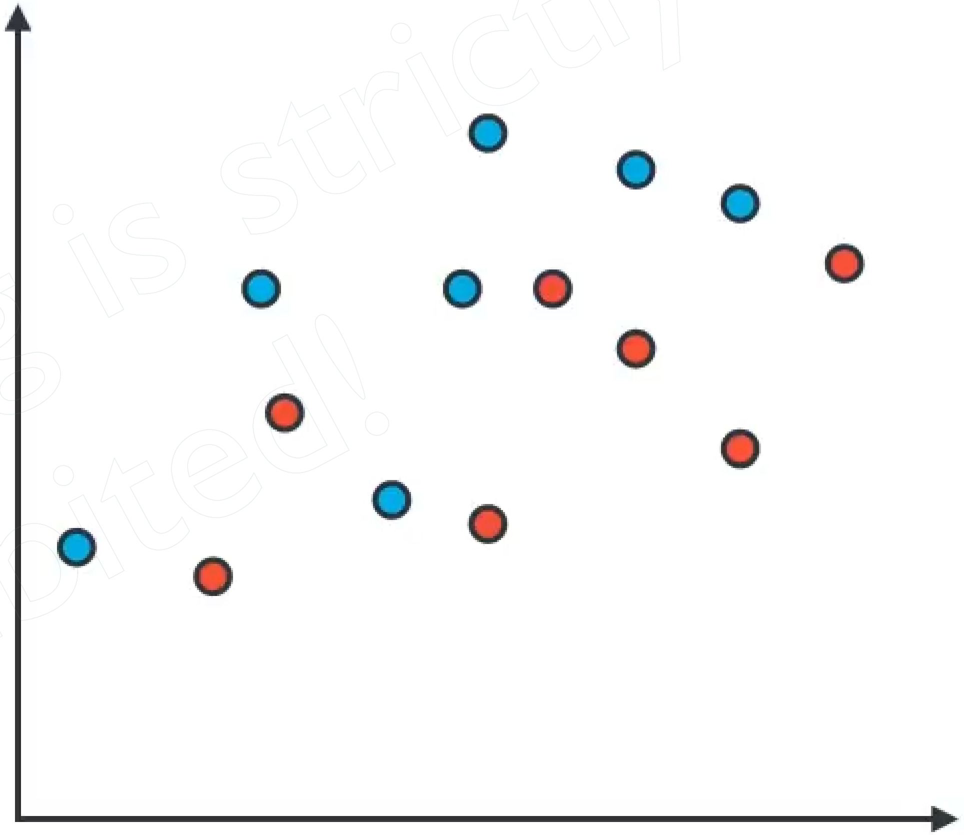


Which model is better

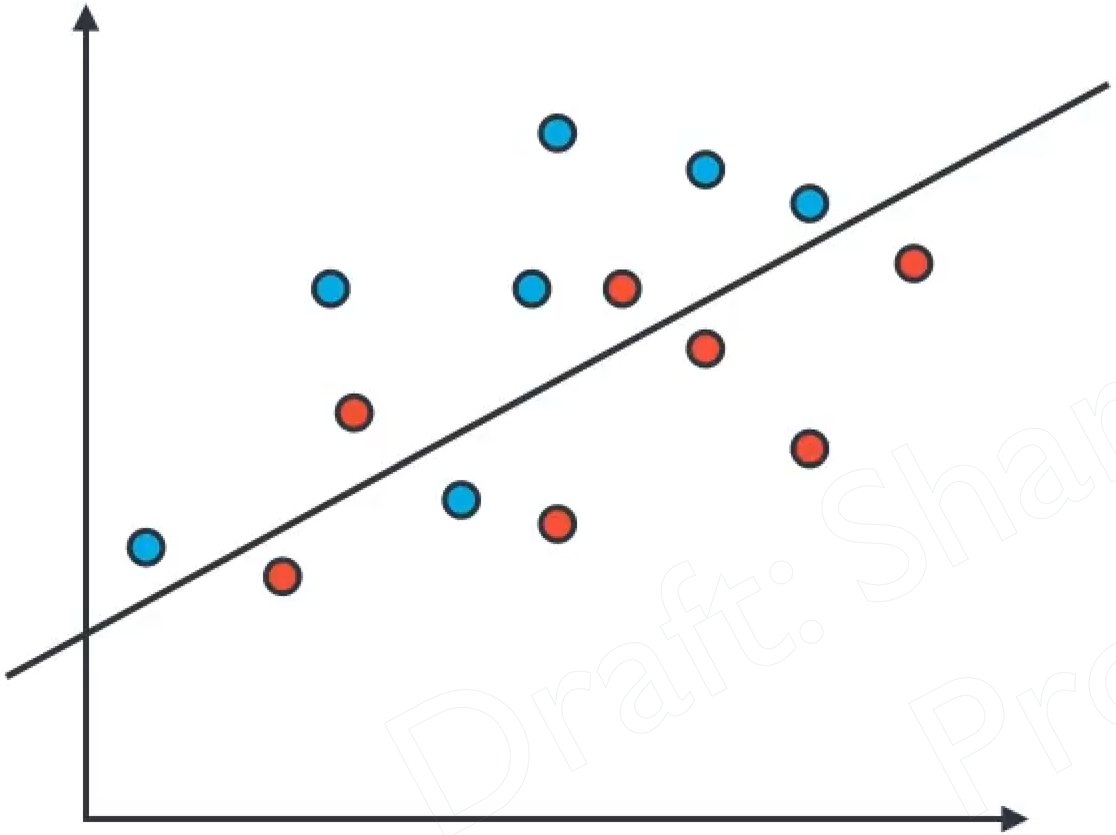


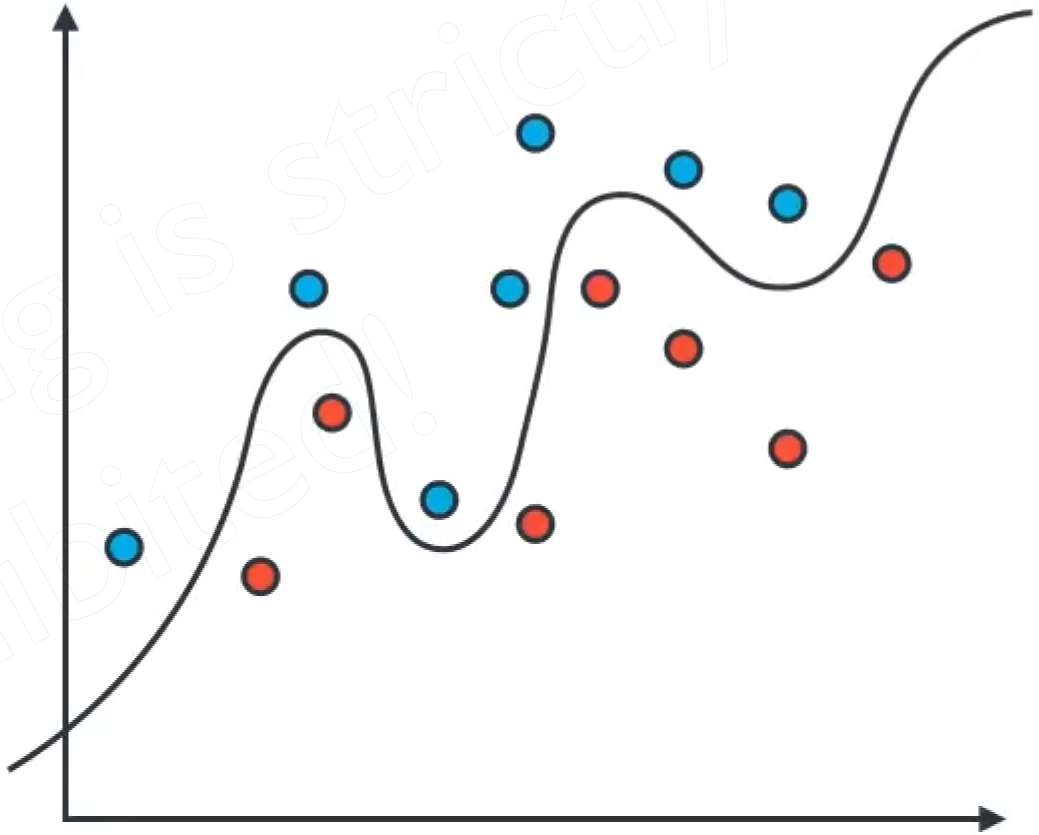
Which model is better



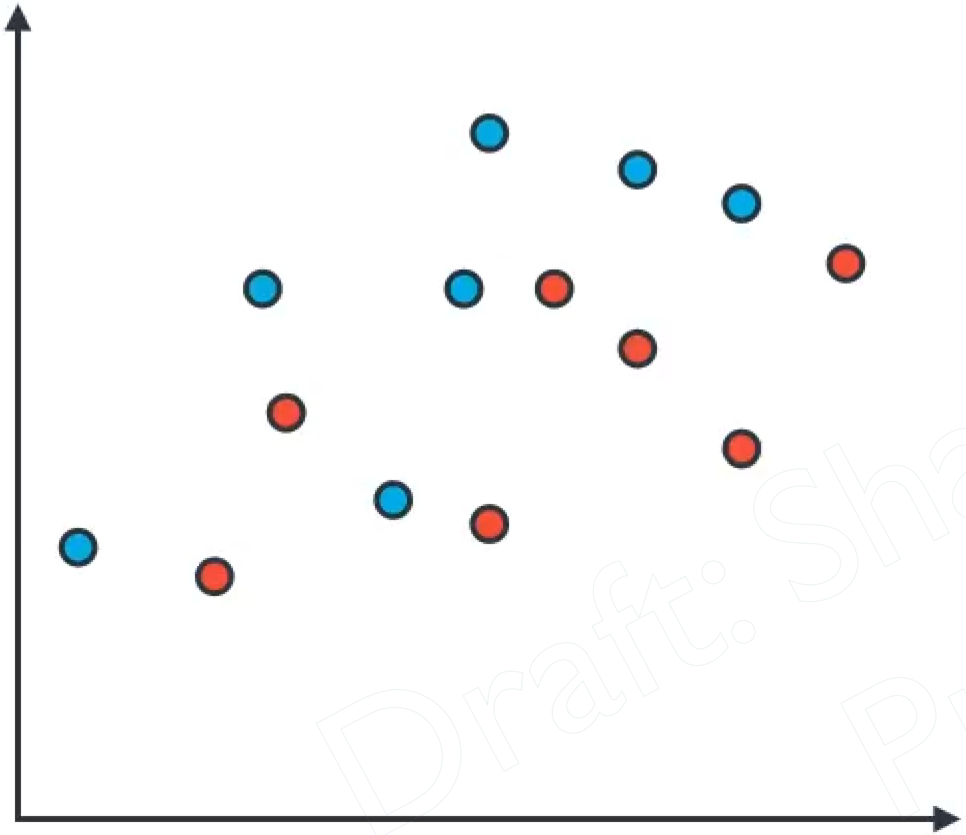


Which model is better

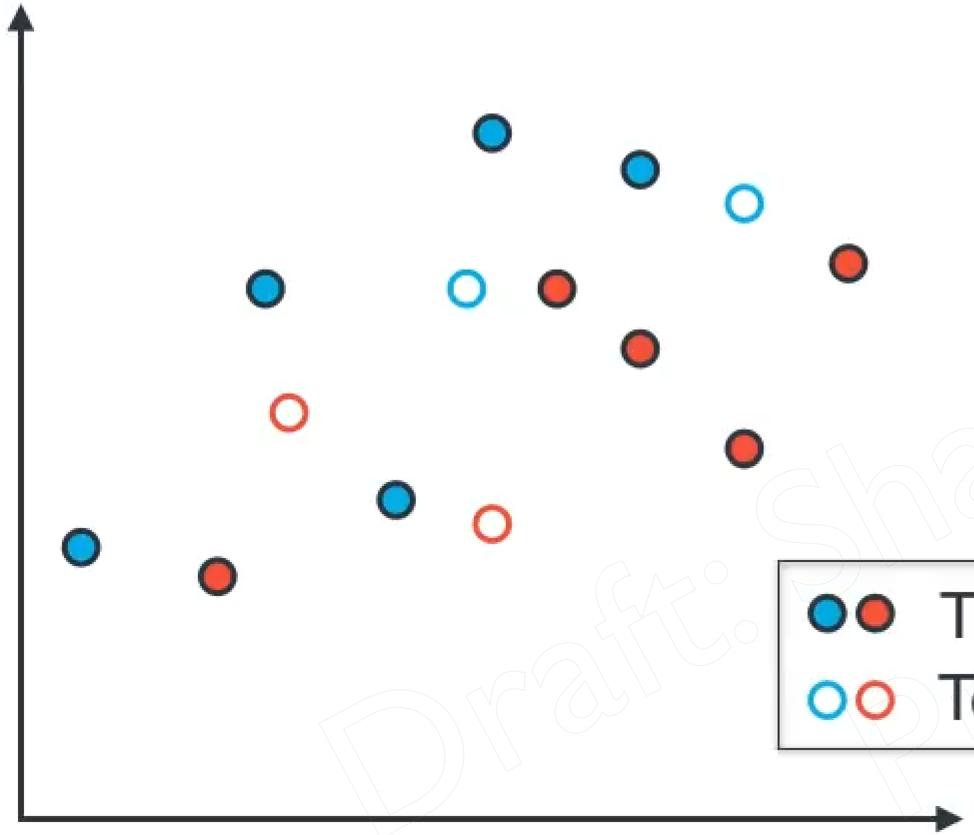




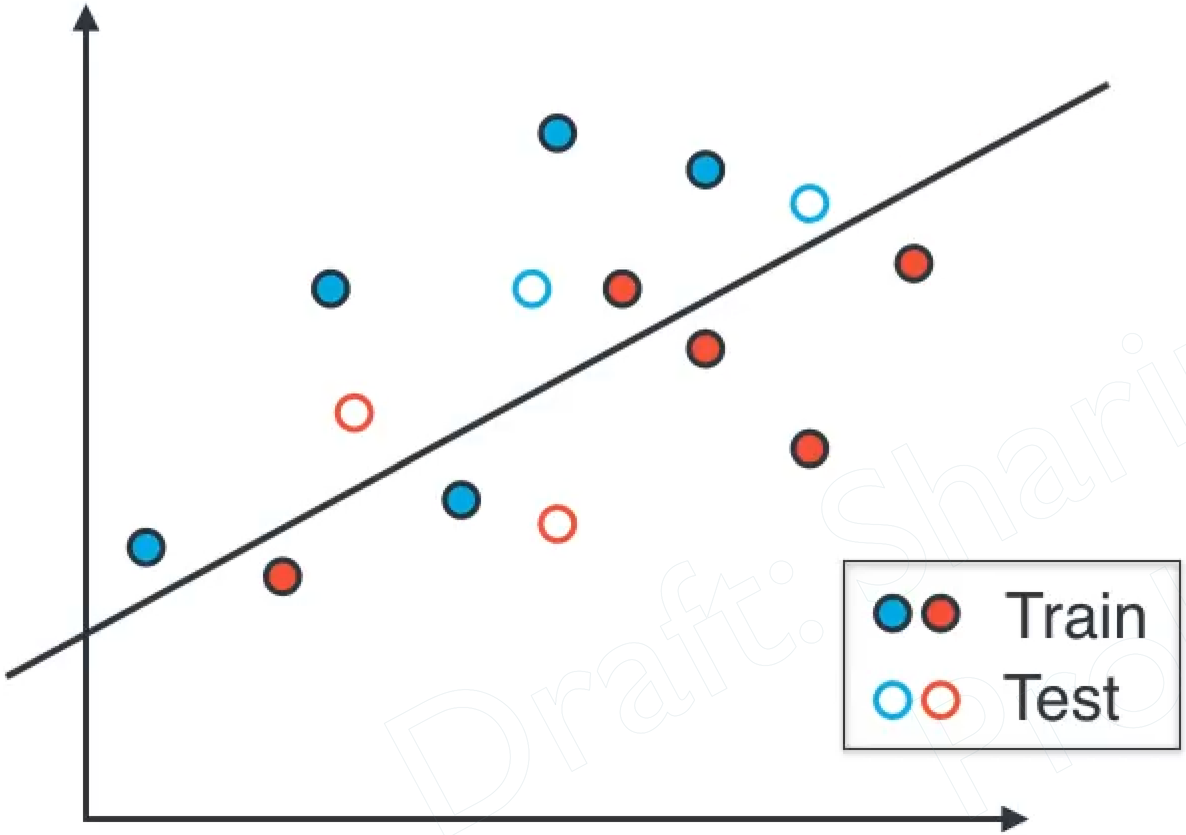
Why Testing?



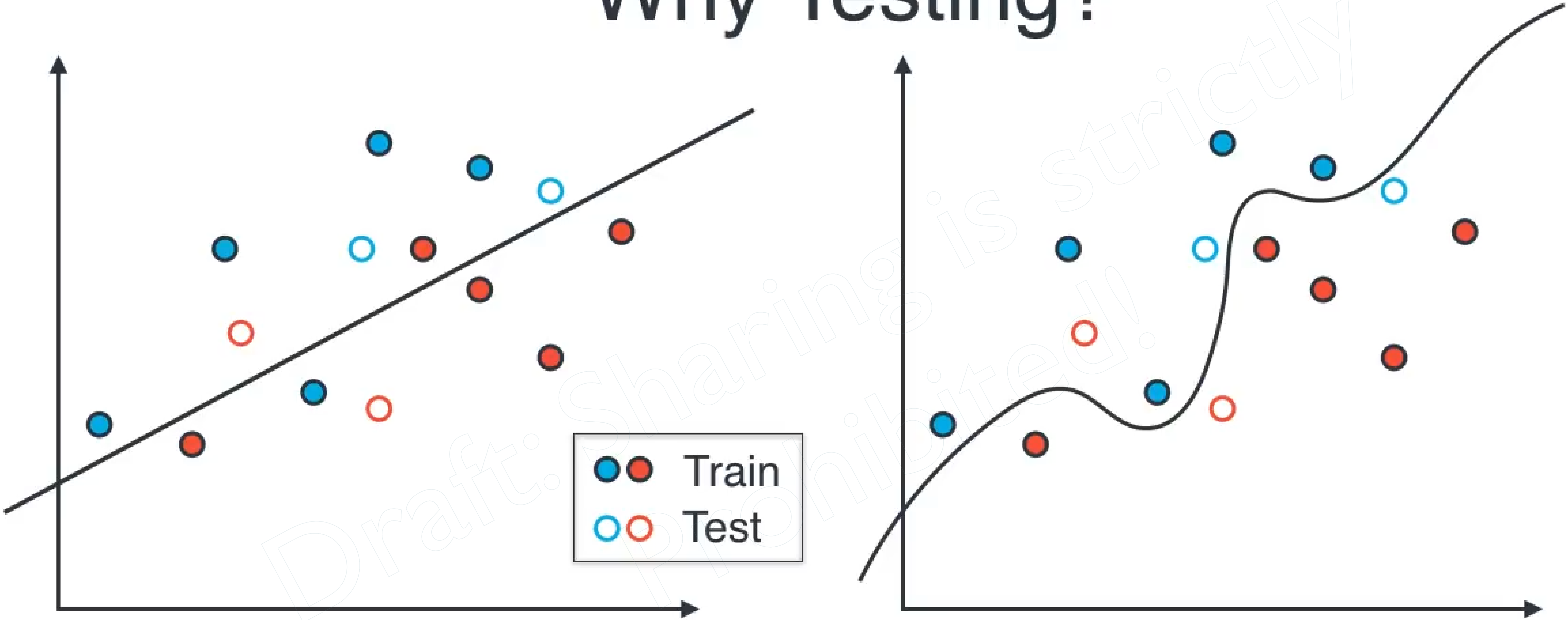
Why Testing?



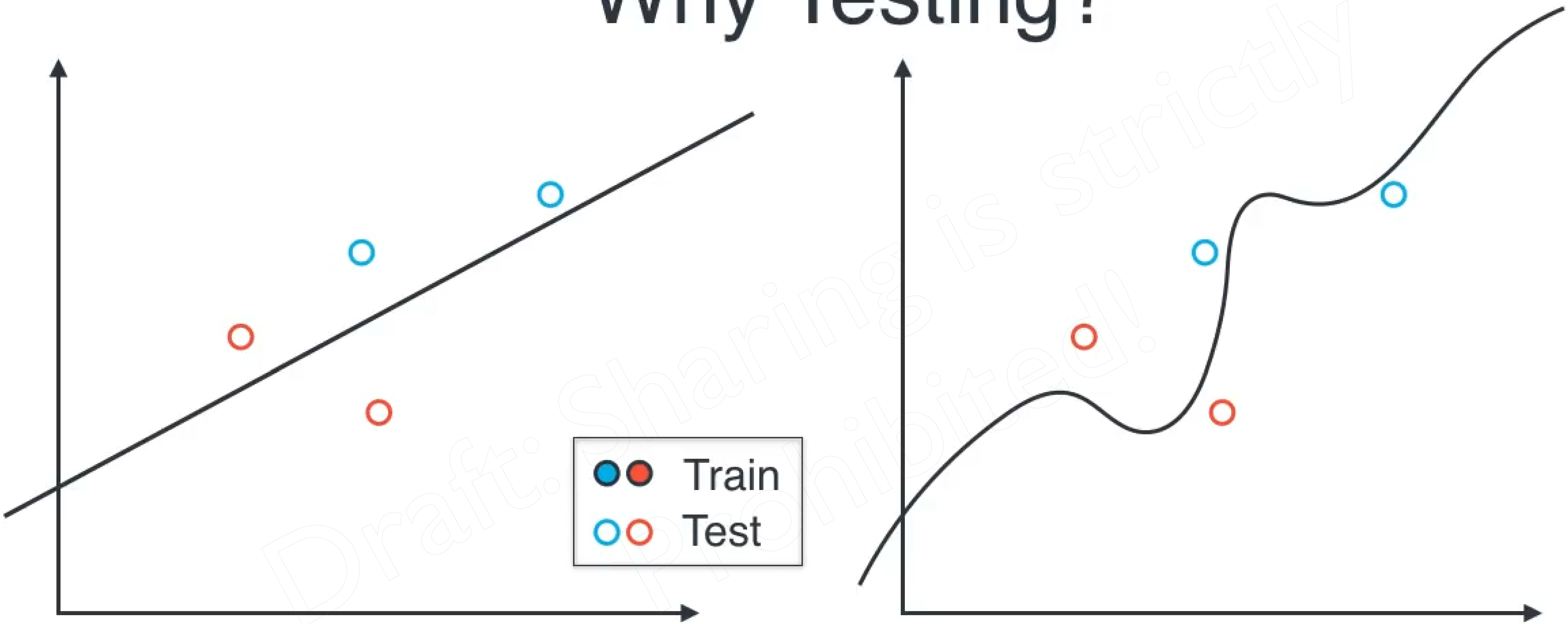
Why Testing?



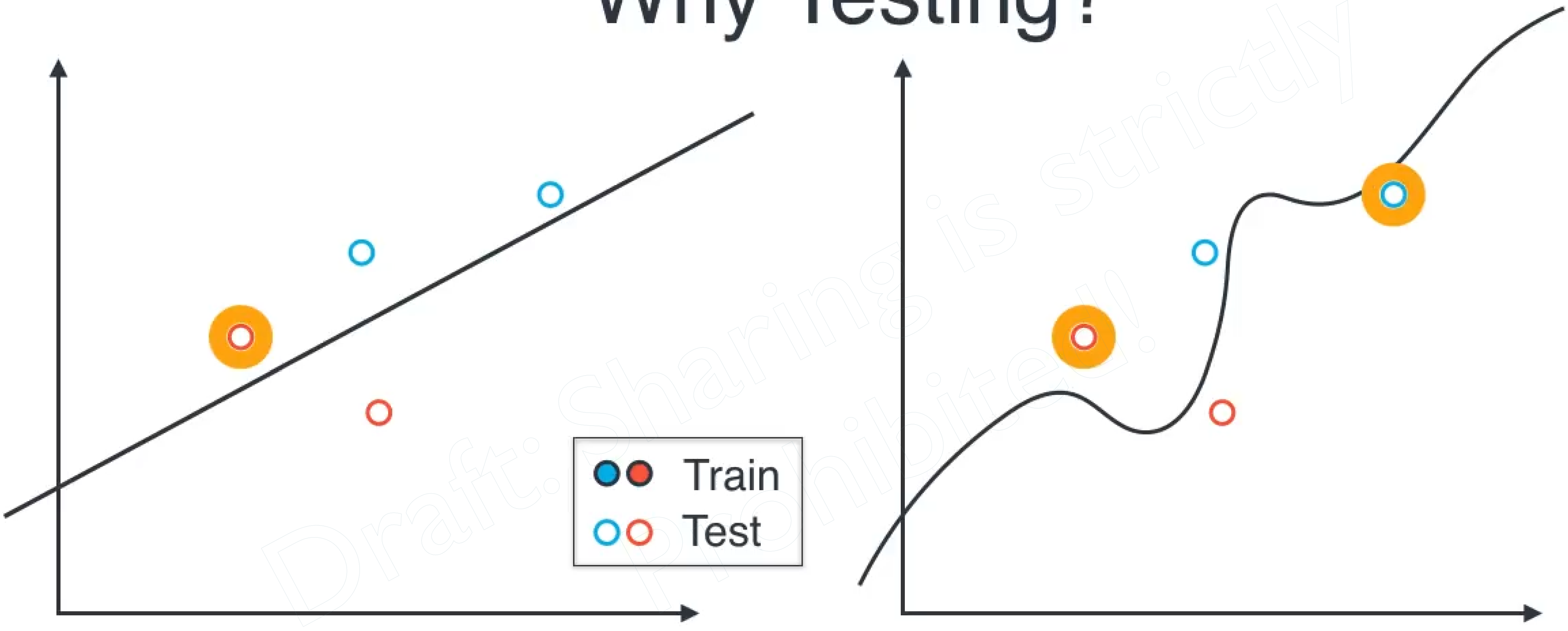
Why Testing?



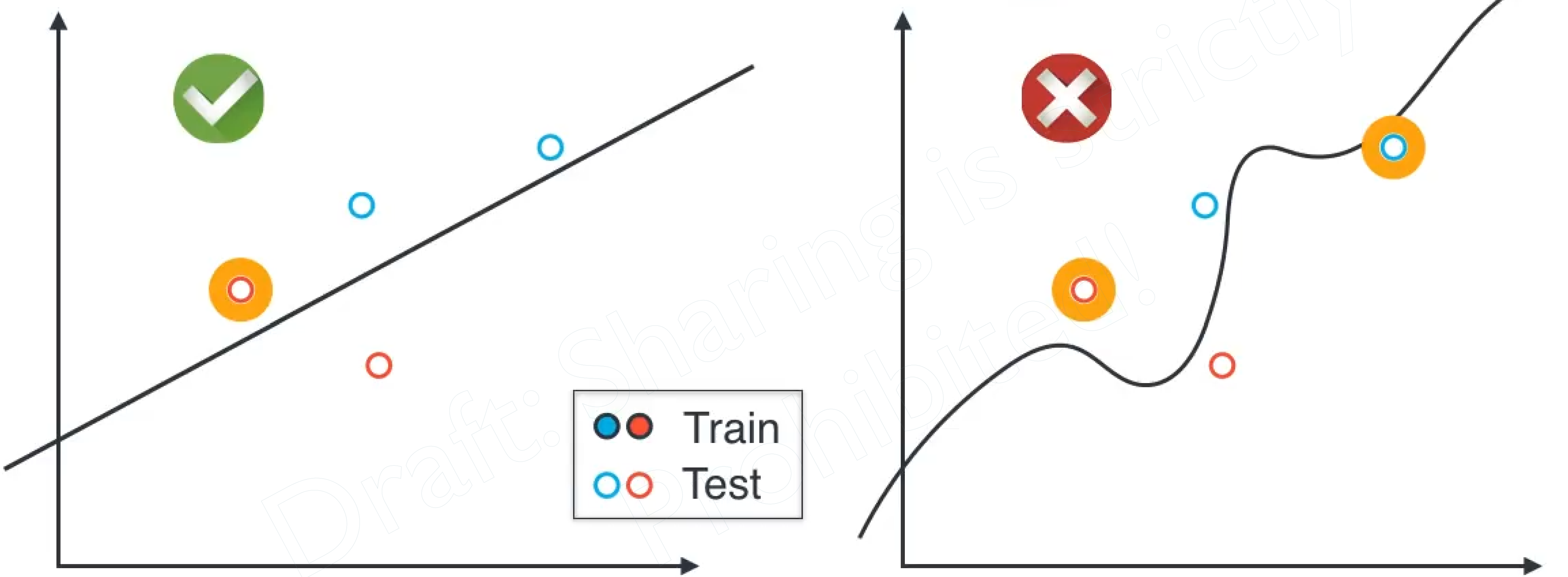
Why Testing?



Why Testing?



Why Testing?

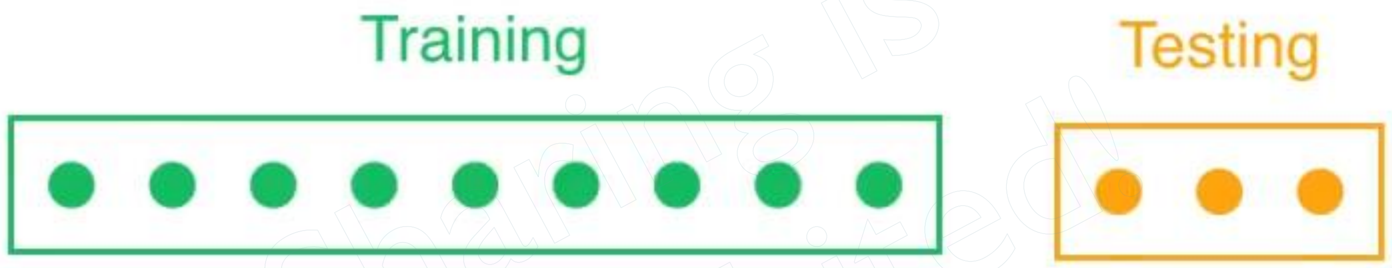


Golden Rule # 1

Golden Rule # 2

Golden Rule # 3

How do we not ‘lose’ the training data?

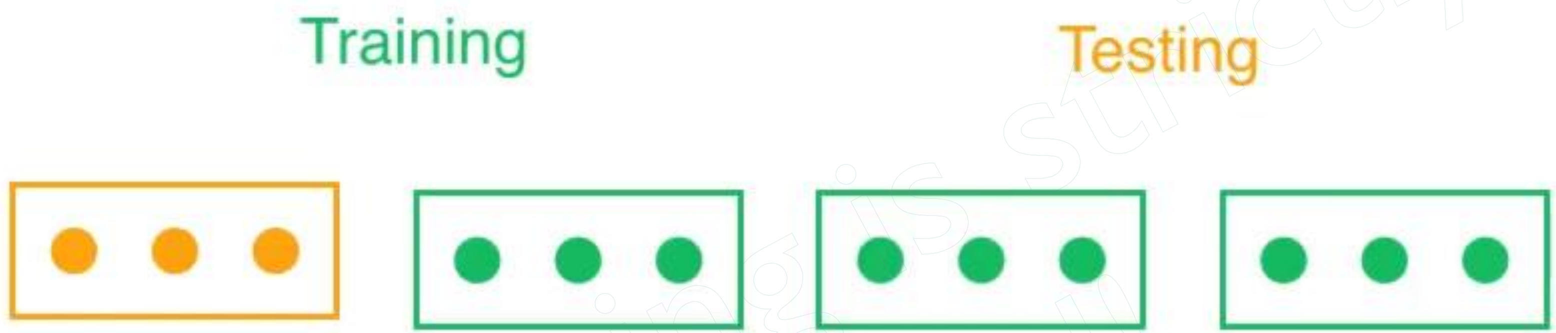


K-Fold Cross Validation

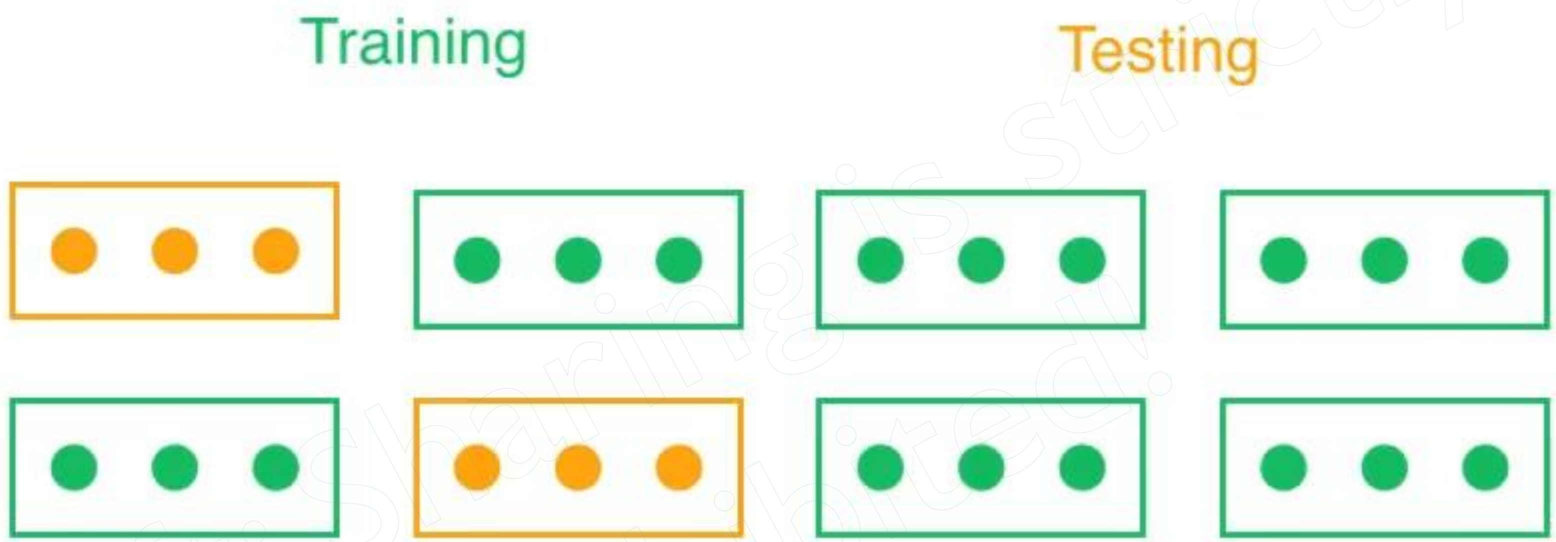
Training

Testing

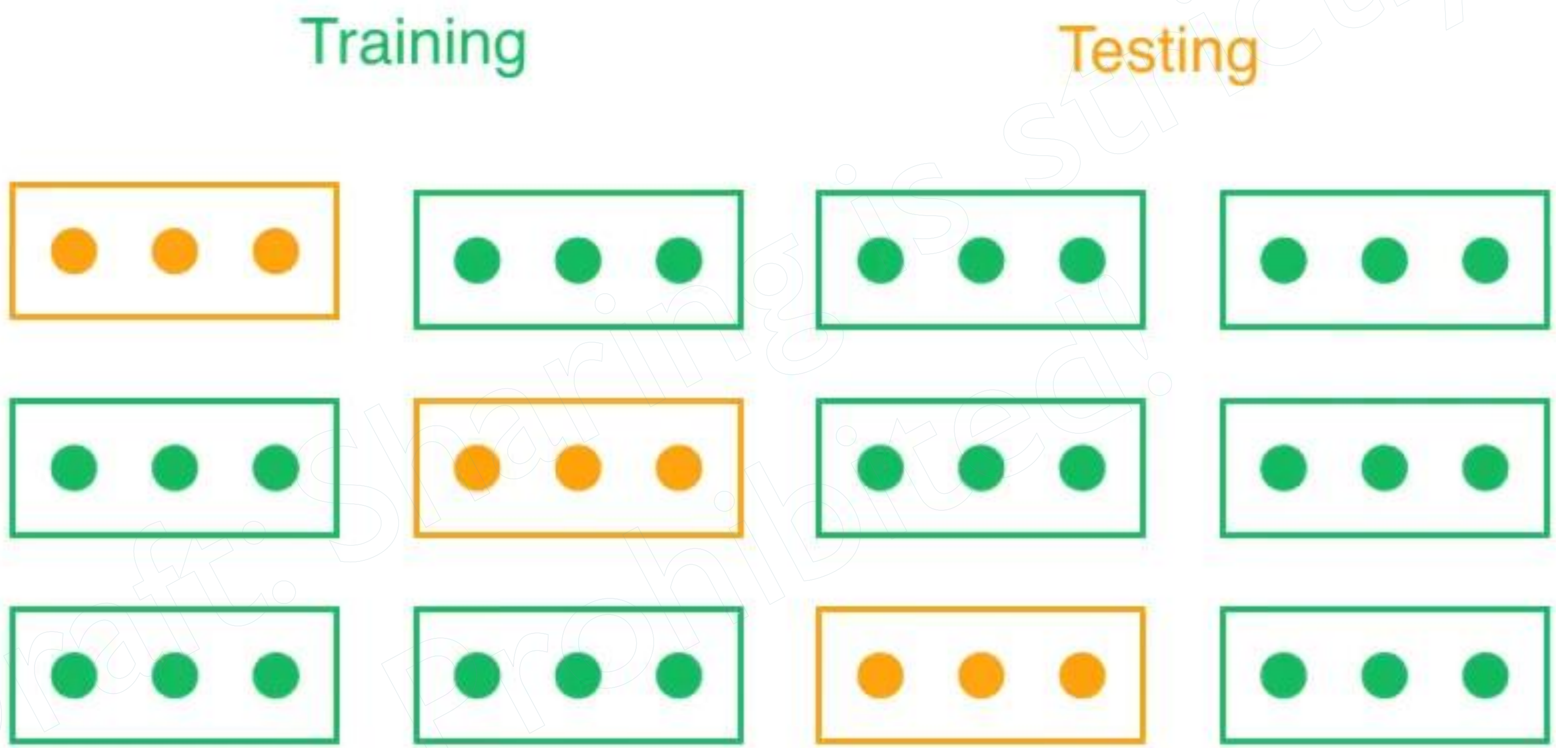
K-Fold Cross Validation



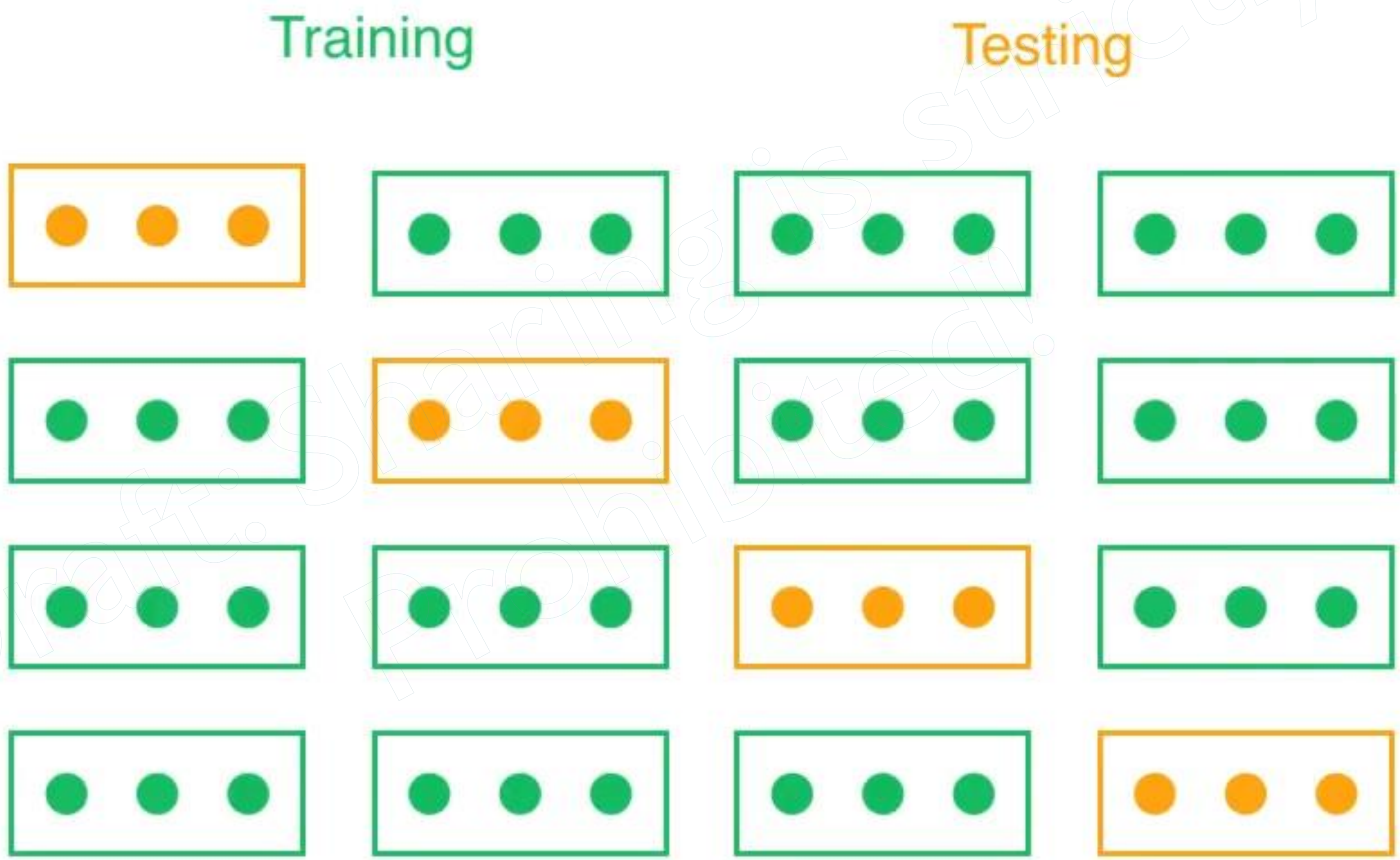
K-Fold Cross Validation



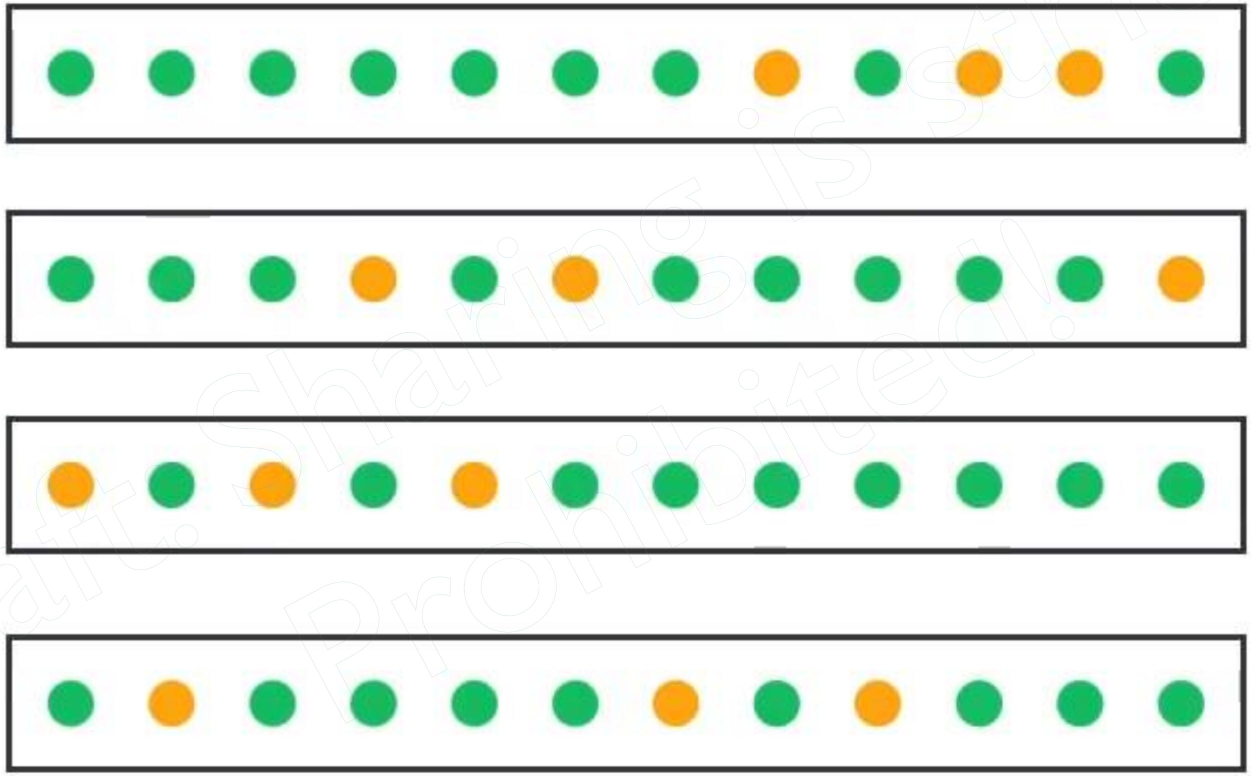
K-Fold Cross Validation



K-Fold Cross Validation



Randomizing in Cross Validation



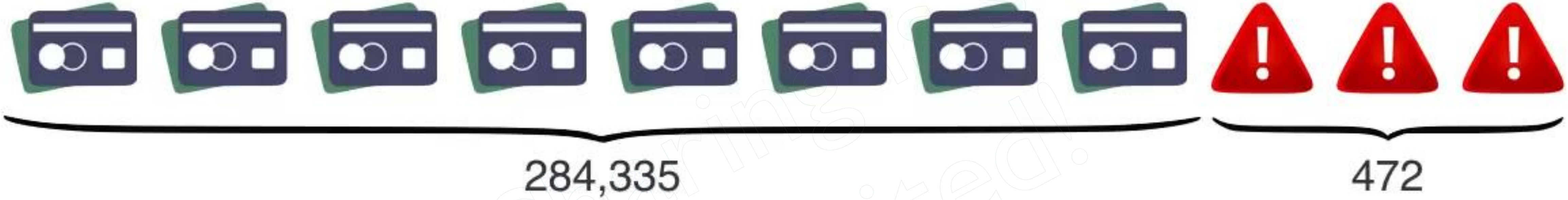
Evaluation Metrics

How well is my model doing?

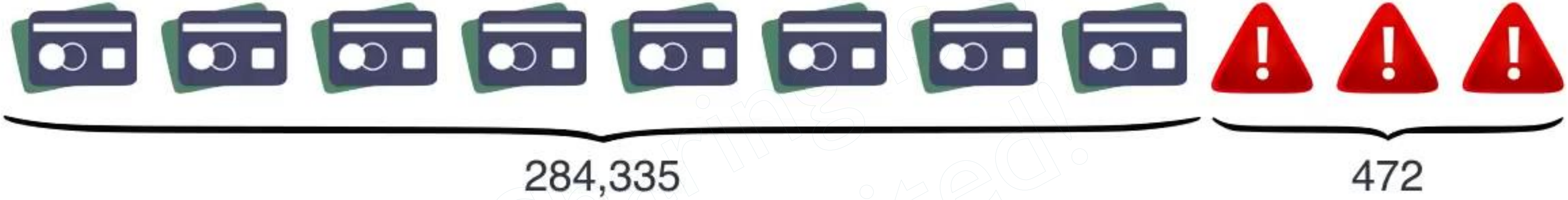
Credit Card Fraud

Credit Card Fraud

Credit Card Fraud



Credit Card Fraud



Model: All transactions are good.

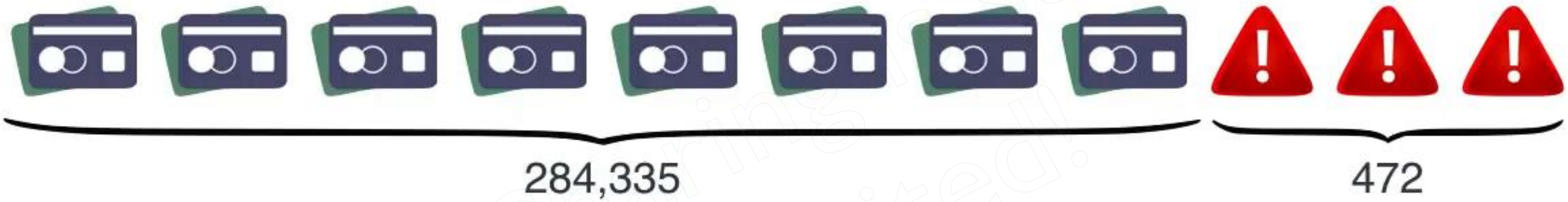
Credit Card Fraud

Model: All transactions are good.

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

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

Credit Card Fraud



Credit Card Fraud

Model: All transactions are fraudulent.

Credit Card Fraud

Model: All transactions are fraudulent.

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

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!

Medical Model

Case 1

Medical Model

Case 1

Healthy

Sick

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

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

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

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

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

Confusion Matrix

10,000
Patients

Diagnosis

Patients	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Healthy	800	8000

Confusion Matrix

10,000
Patients

Diagnosis

Patients	Diagnosed sick	Diagnosed Healthy
Sick	1000 True positives	200
Healthy	800	8000

Confusion Matrix

10,000
Patients

Diagnosis

Patients

	Diagnosed sick	Diagnosed Healthy
Sick	1000 True positives	200 False Negatives
Healthy	800	8000

Confusion Matrix

10,000
Patients

Diagnosis

		Diagnosed sick	Diagnosed Healthy
Patients	Sick	1000 True positives	200 False Negatives
	Healthy	800 False Positives	8000

Confusion Matrix

10,000
Patients

Diagnosis

Patients

	Diagnosed sick	Diagnosed Healthy
Sick	1000 True positives	200 False Negatives
Healthy	800 False Positives	8000 True Negatives

Spam Classifier Model

Case 2

Spam Classifier Model

Case 2

Not spam

Spam

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

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

Confusion Matrix

1,000
e-mails

Folder

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

Confusion Matrix

1,000
e-mails

Folder

E-mail

	Spam Folder	Inbox
Spam	100 True positives	170 False Negatives
Not spam	30	700

Confusion Matrix

1,000
e-mails

E-mail \ Folder	Spam Folder	Inbox
Spam	100 True positives	170 False Negatives
Not spam	30 False Positives	700

Confusion Matrix

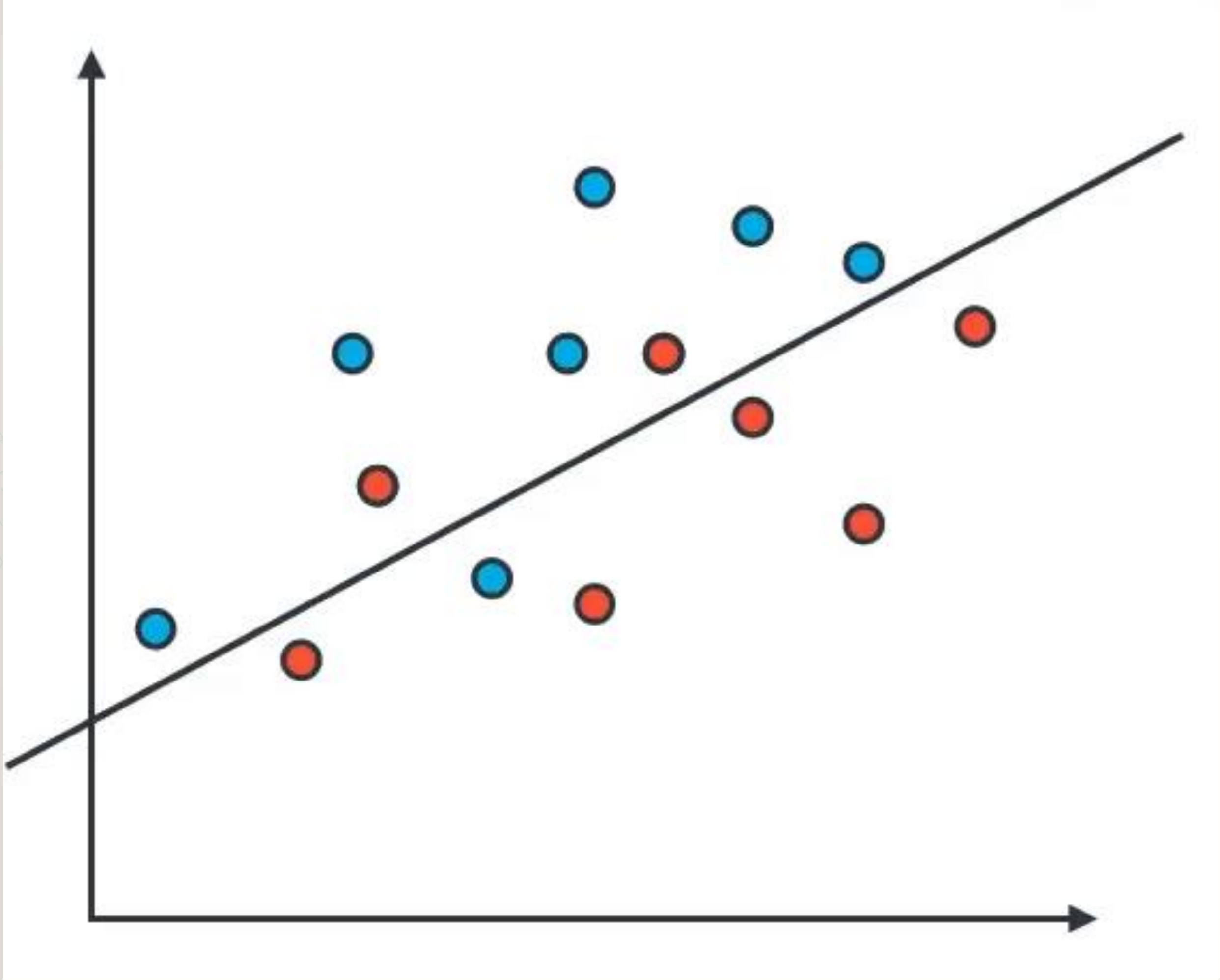
1,000
e-mails

Folder

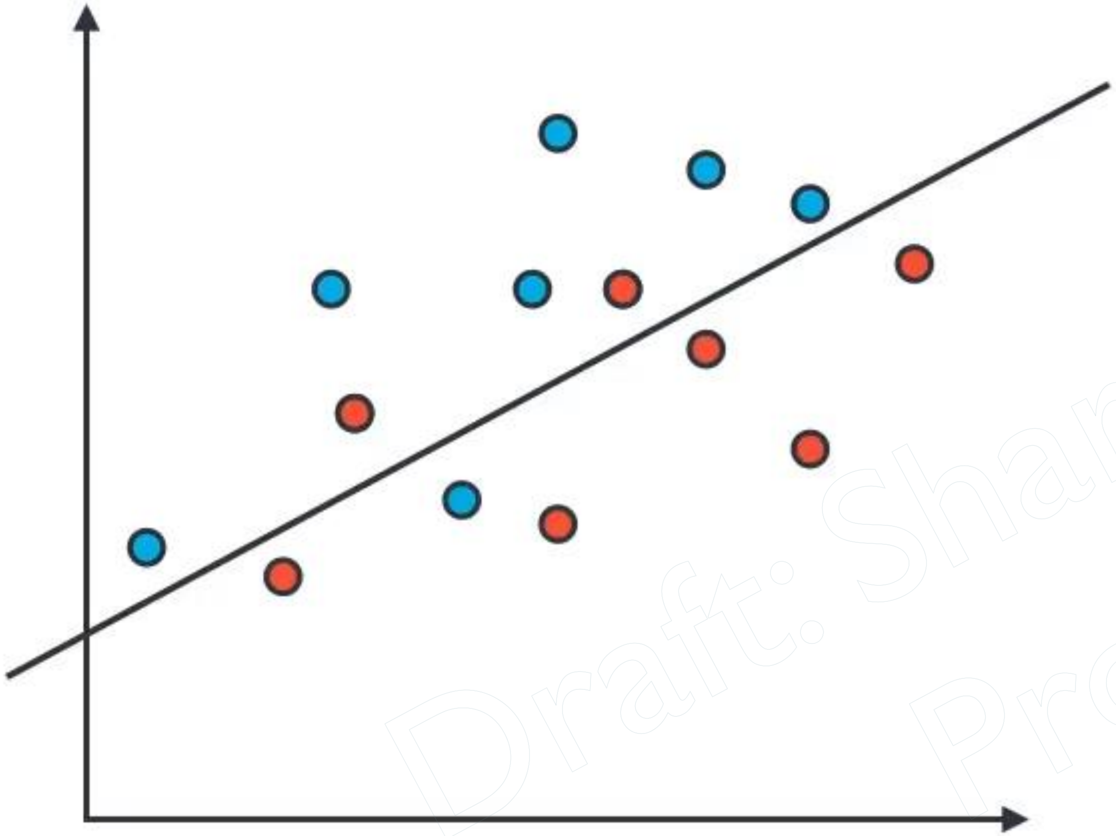
E-mail	Spam Folder	Inbox
Spam	100 True positives	170 False Negatives
Not spam	30 False Positives	700 True Negatives

Pass/Fail or Win/Loss

Case 3



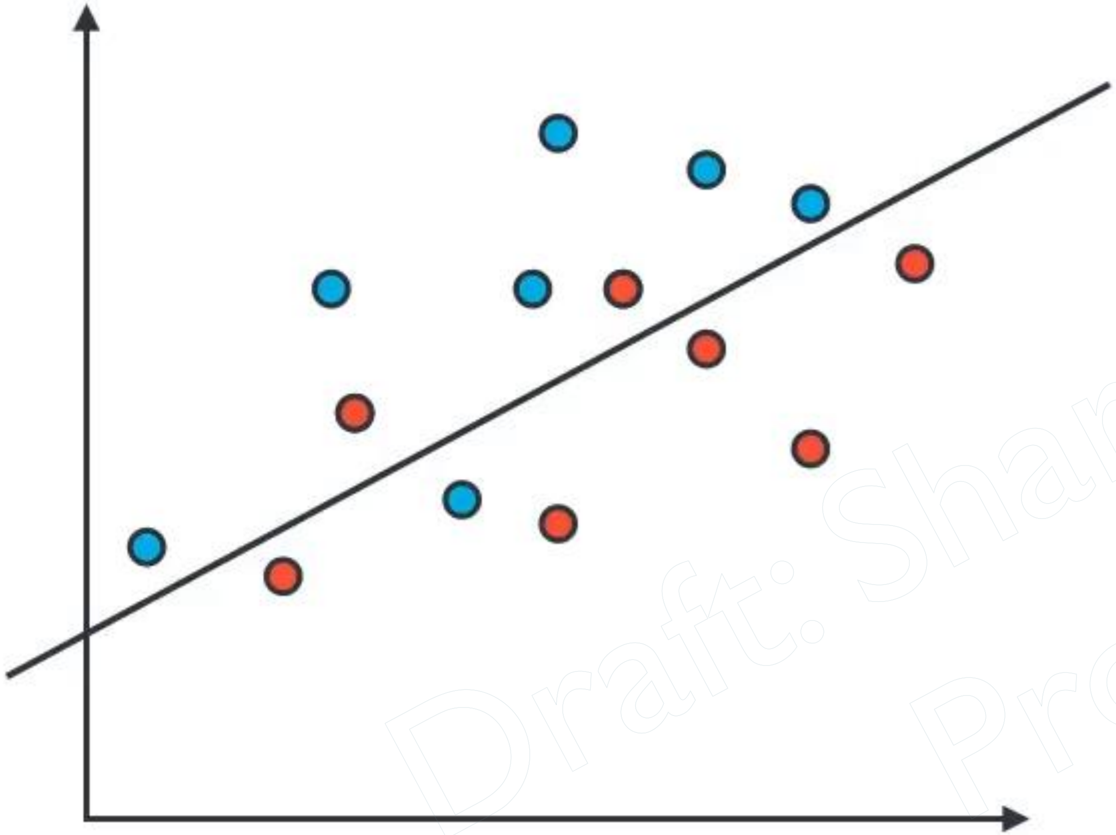
Confusion Matrix



Prediction

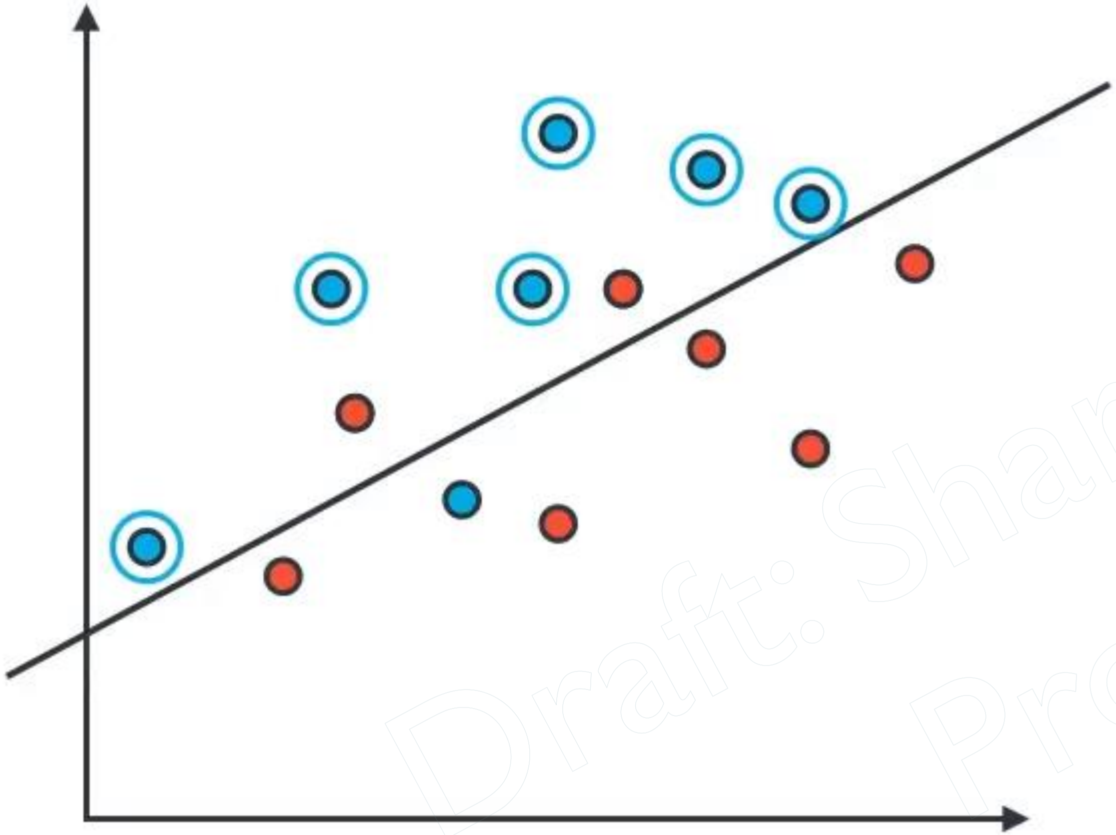
Data	Guessed Positive	Guessed Negative
Positive		
Negative		

Confusion Matrix



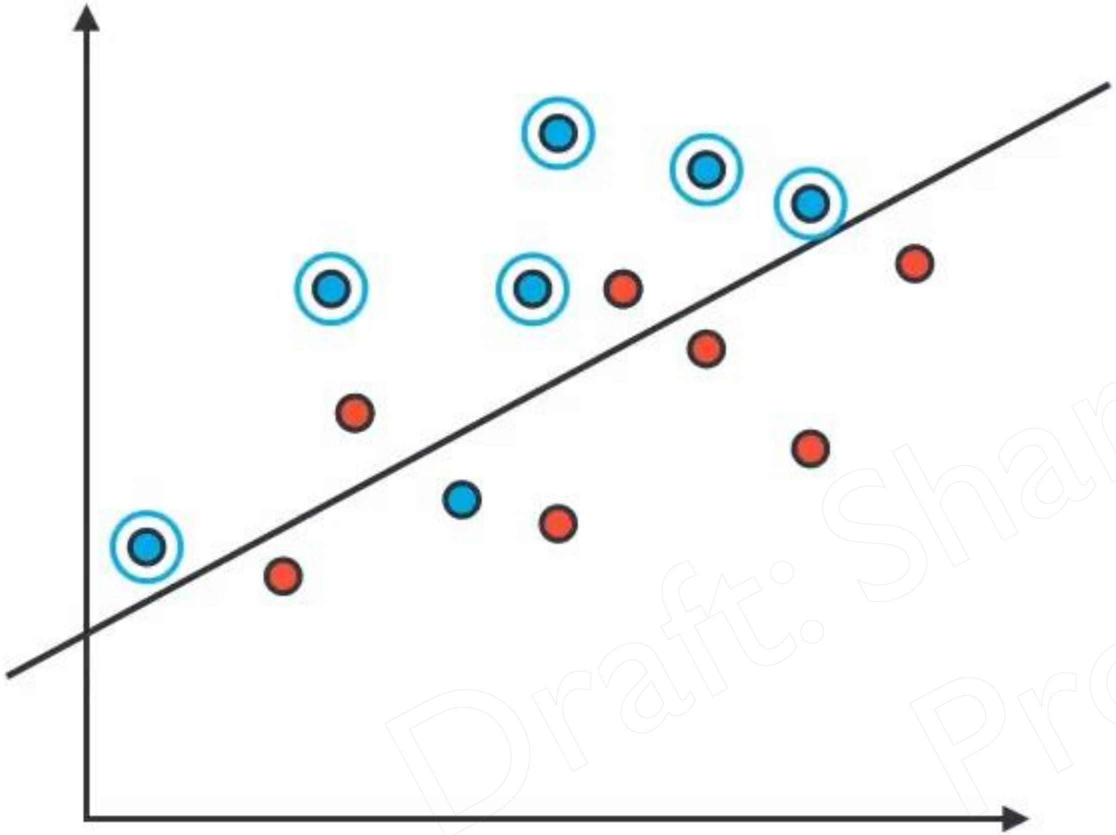
Data \ Prediction	Guessed Positive	Guessed Negative
Positive	True positives	
Negative		

Confusion Matrix



Data \ Prediction	Guessed Positive	Guessed Negative
Positive	True positives	
Negative		

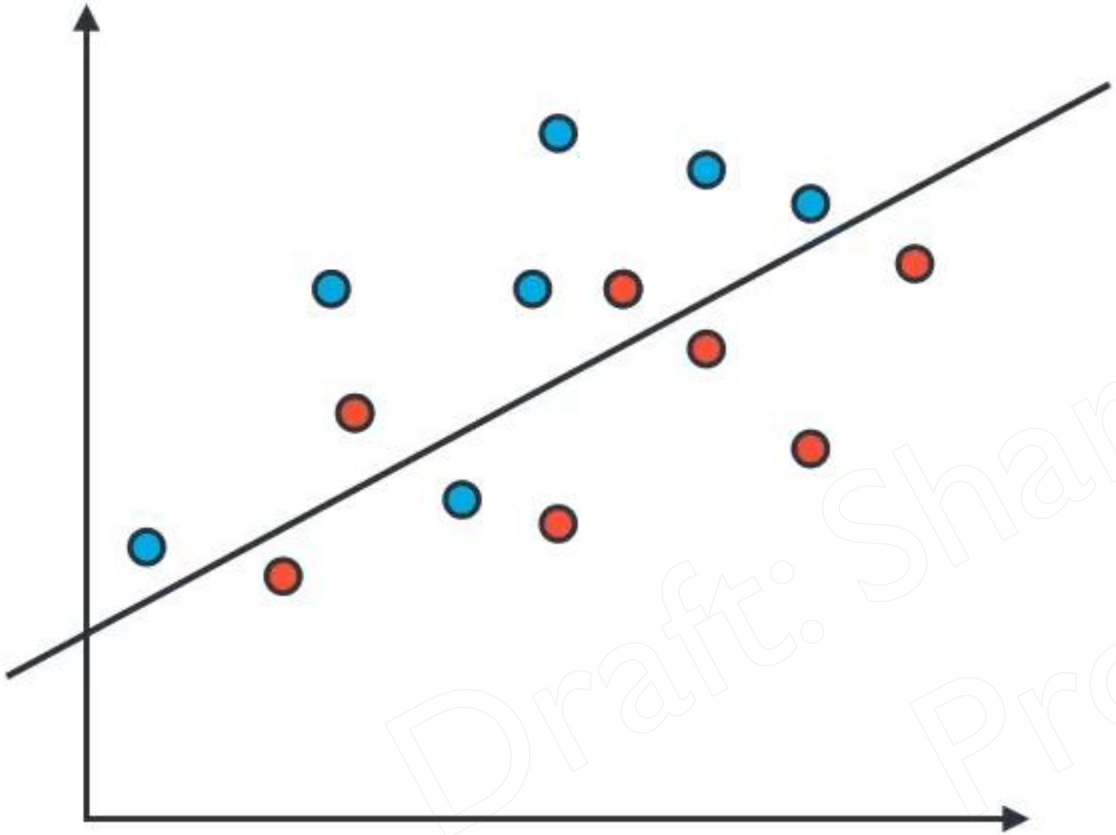
Confusion Matrix



Prediction

Data	Guessed Positive	Guessed Negative
Positive	6 True positives	
Negative		

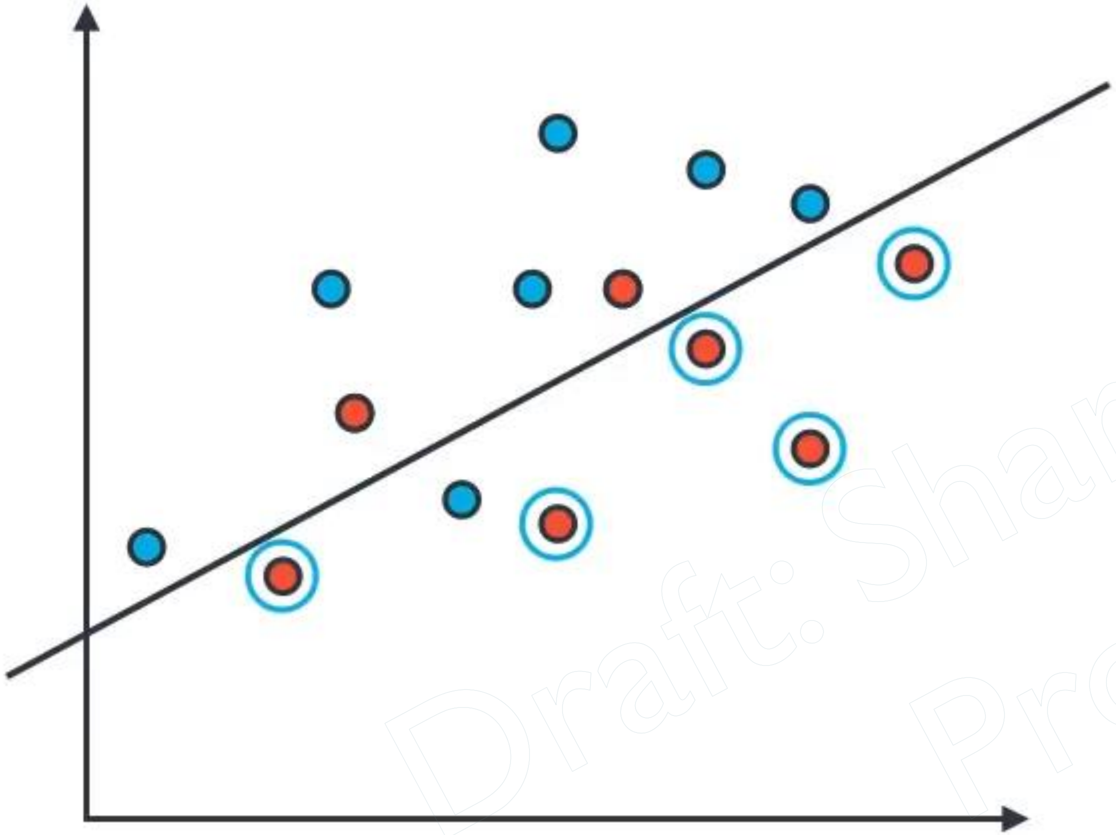
Confusion Matrix



Prediction

Data		Guessed Positive	Guessed Negative
	Positive	6 True positives	
	Negative		True Negatives

Confusion Matrix

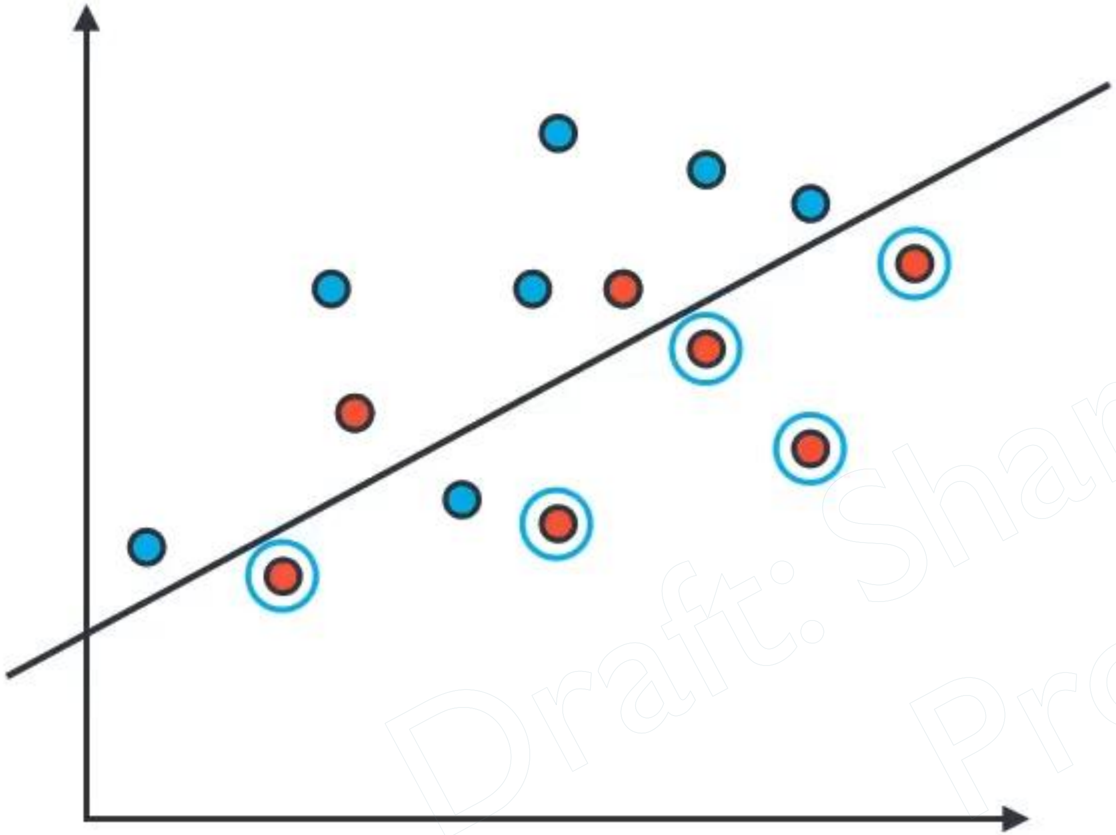


Prediction

Data

	Guessed Positive	Guessed Negative
Positive	6 True positives	
Negative		True Negatives

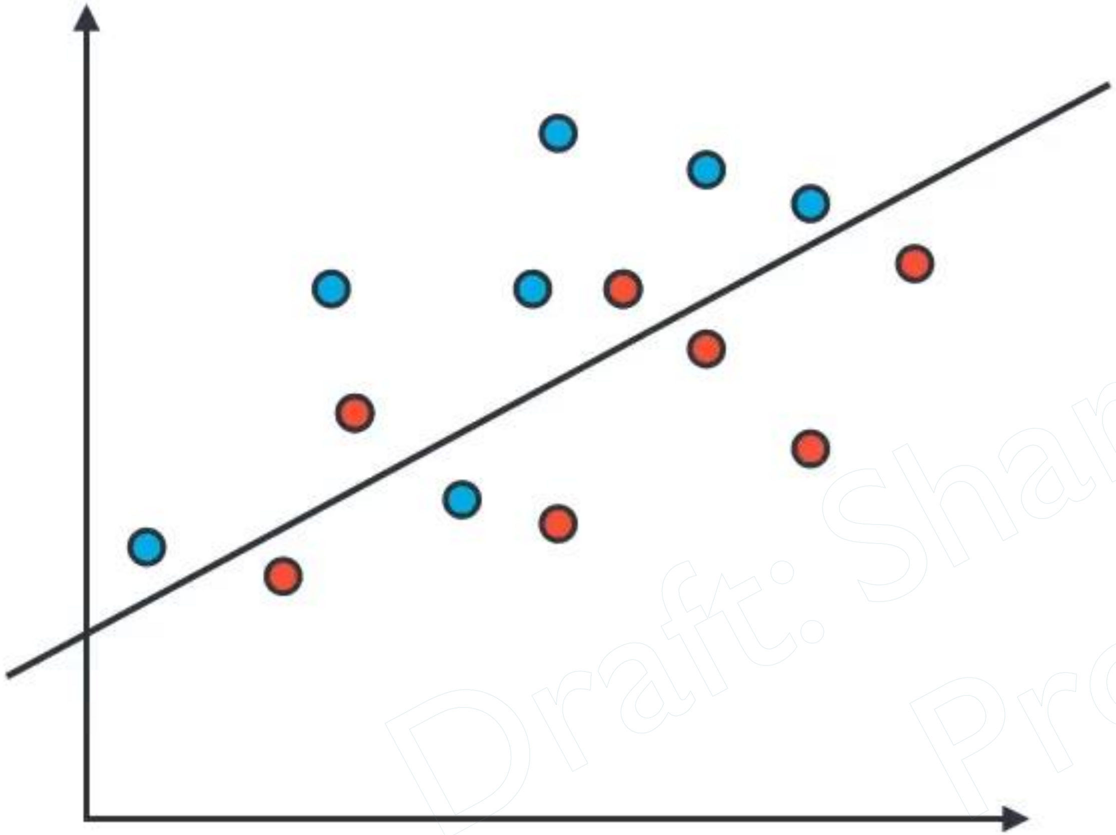
Confusion Matrix



Prediction

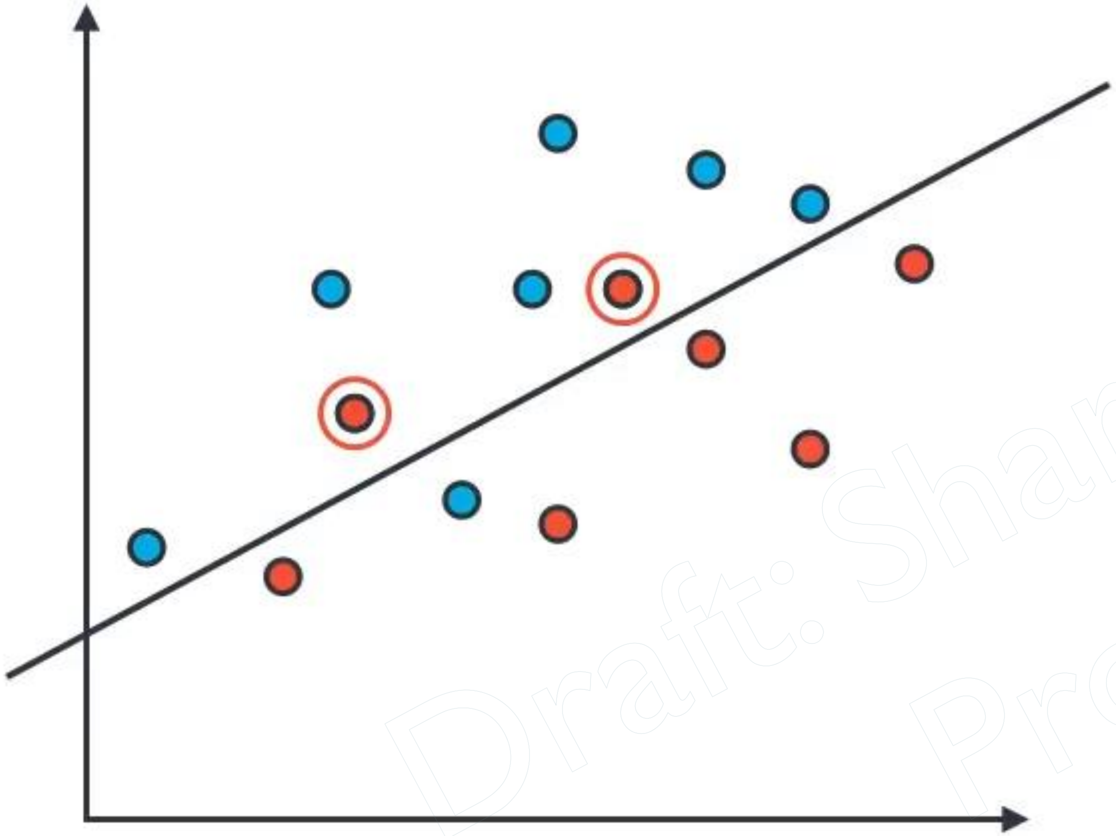
Data	Guessed Positive	Guessed Negative
Positive	6 True positives	
Negative		5 True Negatives

Confusion Matrix



		Prediction	
		Guessed Positive	Guessed Negative
Data	Positive	6 True positives	
	Negative	False Positives	5 True Negatives

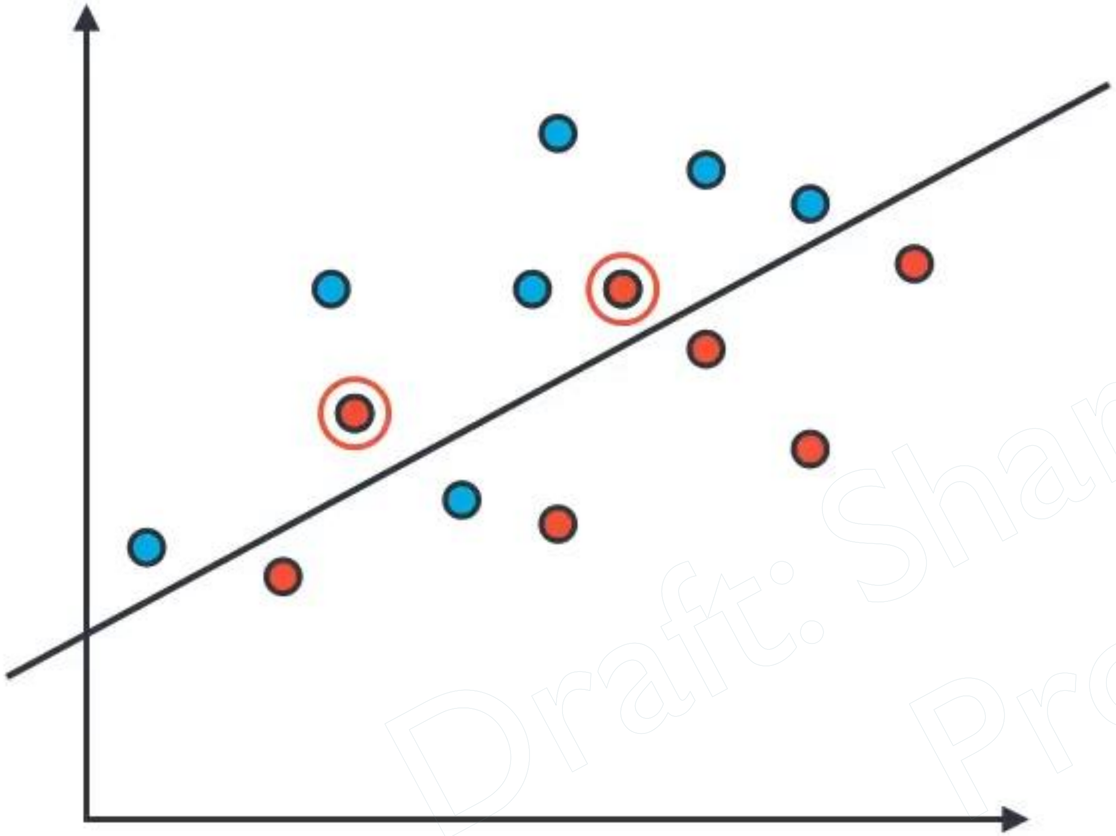
Confusion Matrix



Prediction

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

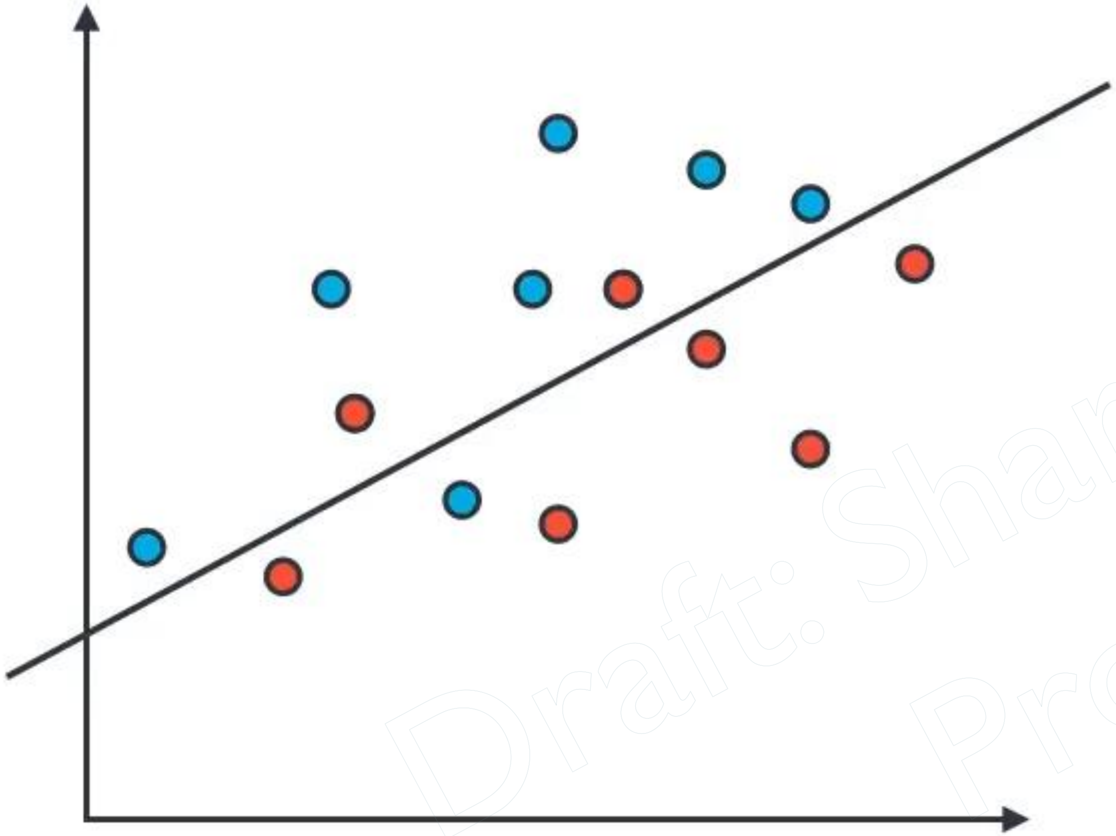
Confusion Matrix



Prediction

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

Confusion Matrix

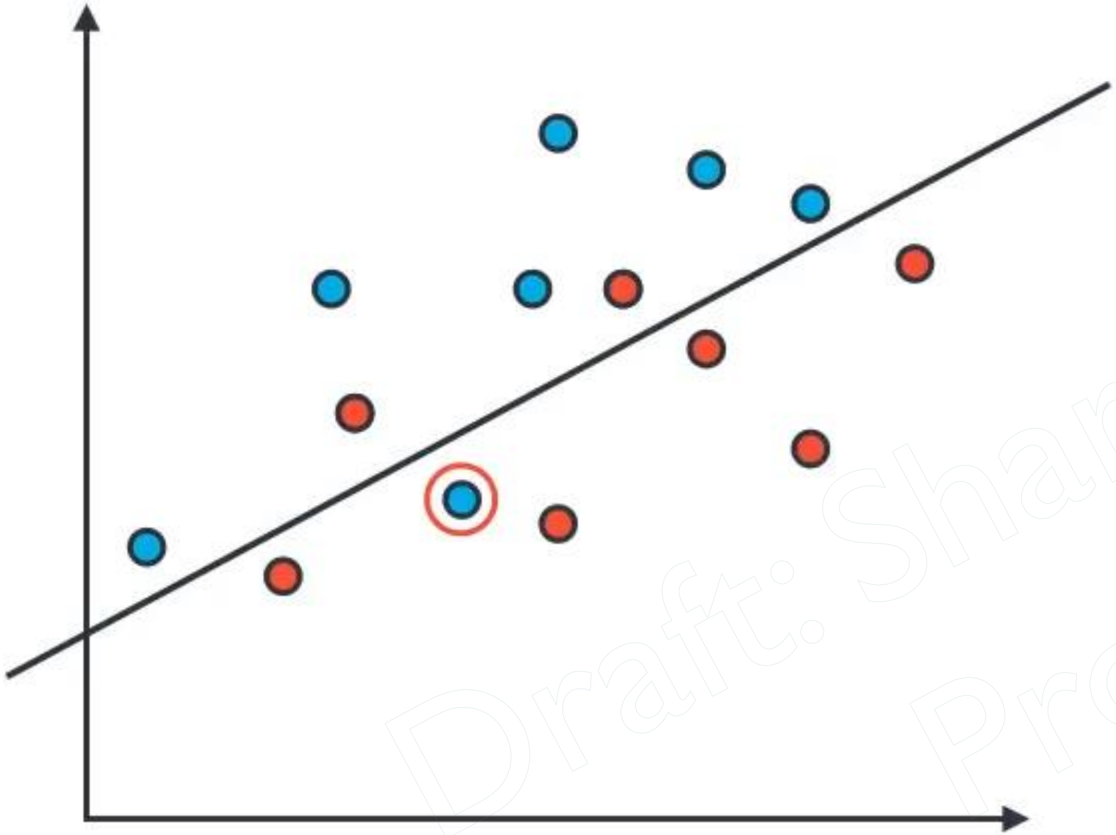


Prediction

Data

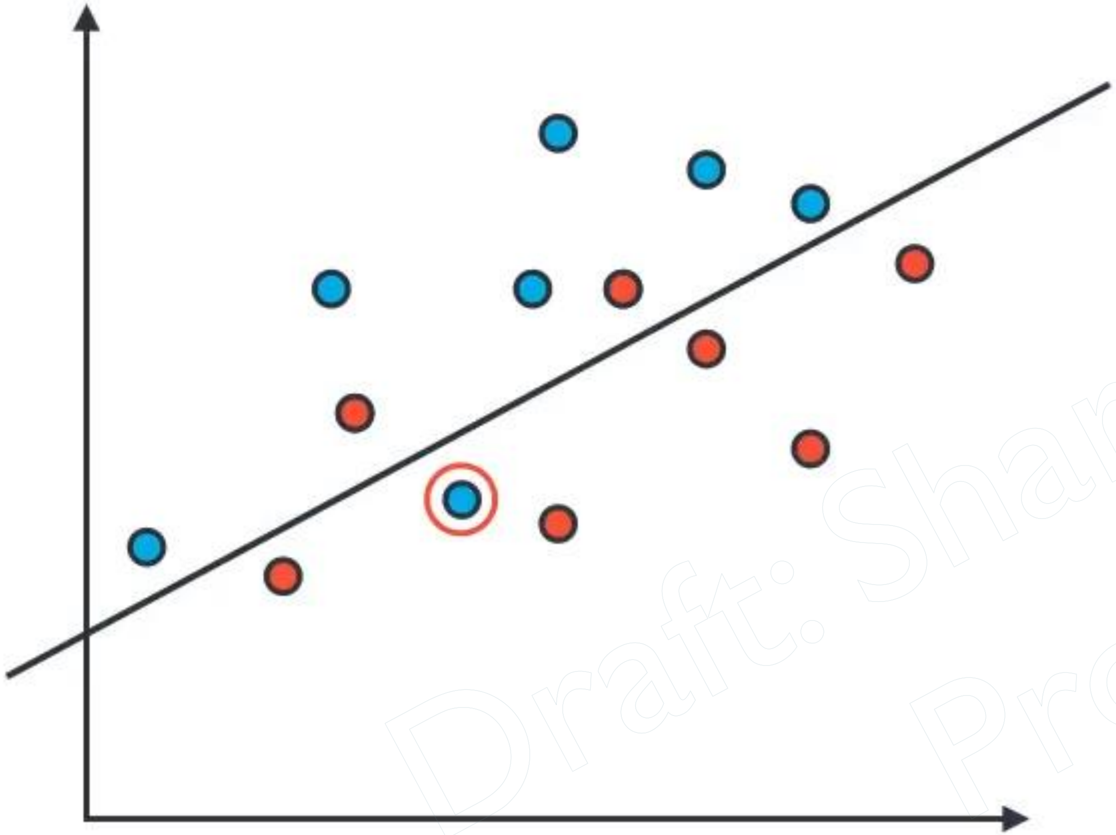
	Guessed Positive	Guessed Negative
Positive	6 True positives	False Negatives
Negative	2 False Positives	5 True Negatives

Confusion Matrix



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

Confusion Matrix



Prediction

Data

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

Accuracy

Diagnosis

Patients

	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Healthy	800	8000

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?

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?

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

Accuracy

		Diagnosis	
		Diagnosed sick	Diagnosed Healthy
Patients	Sick	1000	200
	Healthy	800	8000

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

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

Accuracy

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?

Accuracy =

Accuracy

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?

$$\text{Accuracy} = \frac{100 + 700}{\quad}$$

Accuracy

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?

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

EVALUATION METRICS

Medical Model

False positives ok
False negatives **NOT** ok

Spam Detector

False positives **NOT** ok
False negatives ok

EVALUATION METRICS

Medical Model

False positives ok
False negatives **NOT** ok

Find all the sick people
Ok if not all are sick

Spam Detector

False positives **NOT** ok
False negatives ok

EVALUATION METRICS

Medical Model

False positives ok
False negatives **NOT** ok

Find all the sick people
Ok if not all are sick

Spam Detector

False positives **NOT** ok
False negatives ok

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

EVALUATION METRICS

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

Precision

Diagnosis

Patients

	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Is Healthy	600	9000

Precision

Diagnosis

Patients

	Diagnosed sick	Diagnosed Healthy
Sick	1000	200
Is Healthy	600	9000

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

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?

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?

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

Precision

Folder

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

Precision

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?

Precision

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?

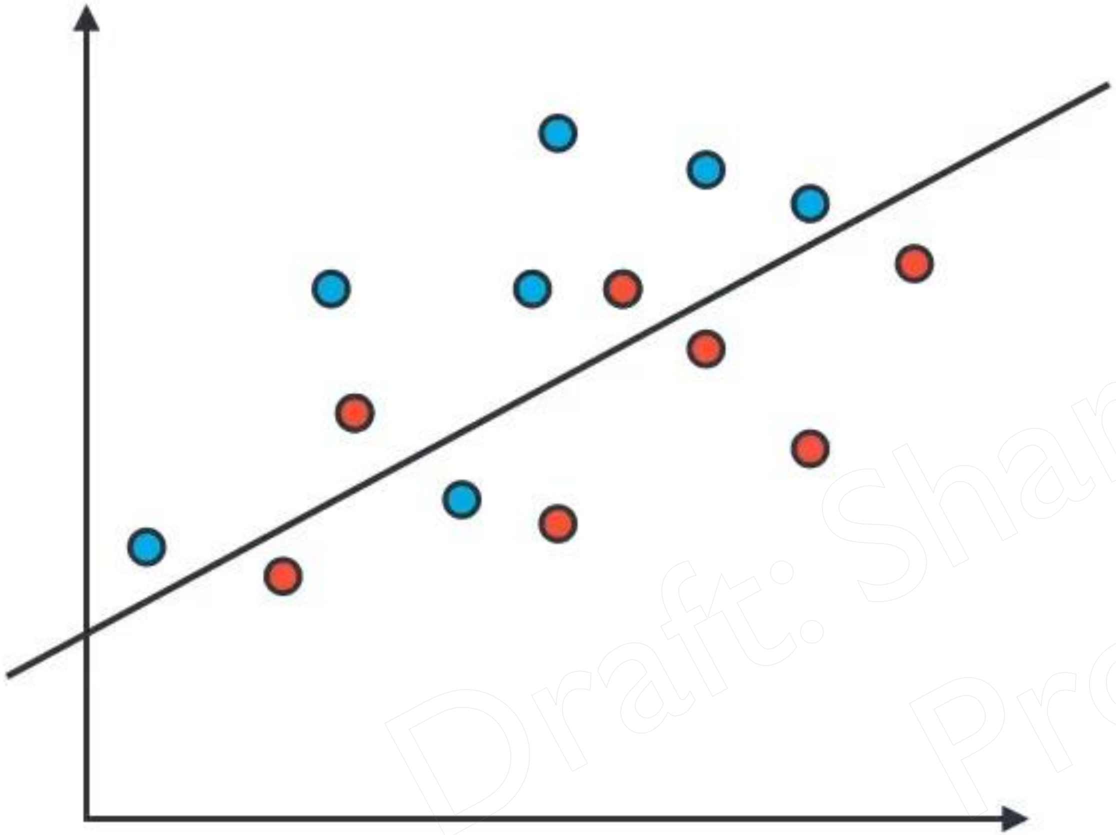
Precision

		Folder	
		Spam Folder	Inbox
E-mail	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?

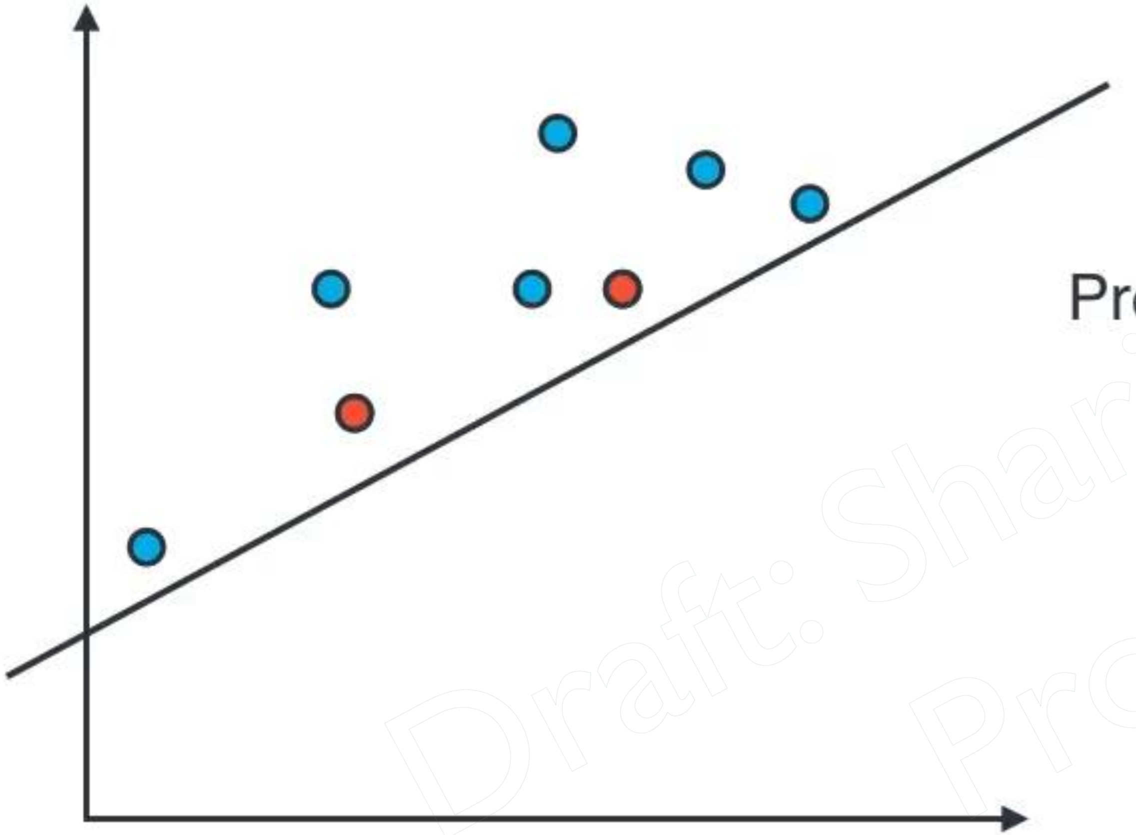
$$\text{Precision} = \frac{100}{100 + 30} = 76.9\%$$

Precision



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

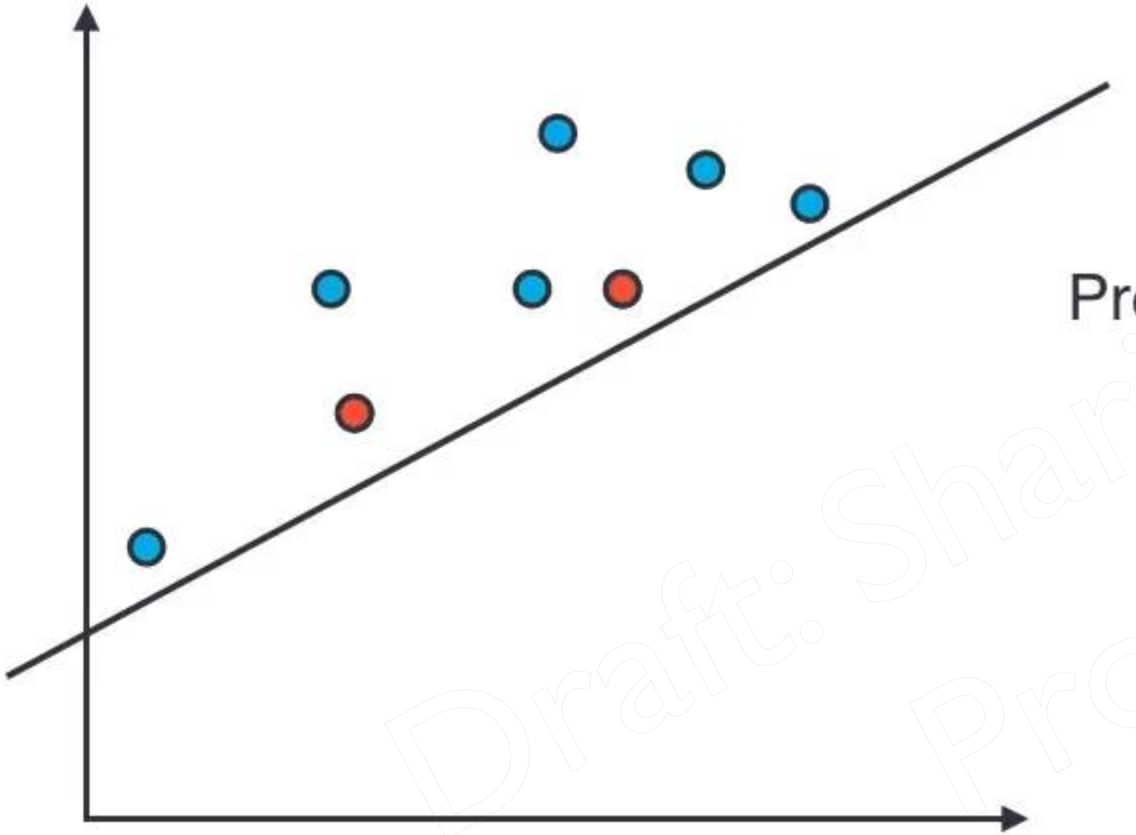
Precision



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

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False Positives}}$$

Precision



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

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False Positives}}$$

$$= \frac{6}{6 + 2}$$

$$= \frac{6}{8}$$

$$= 75\%$$

Recall

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

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?

Recall

Patients	Diagnosis: Diagnosed Sick	Diagnosis: Diagnosed Healthy
Sick	1000	200
Is Healthy	800	8000

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

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?

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

Recall

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?

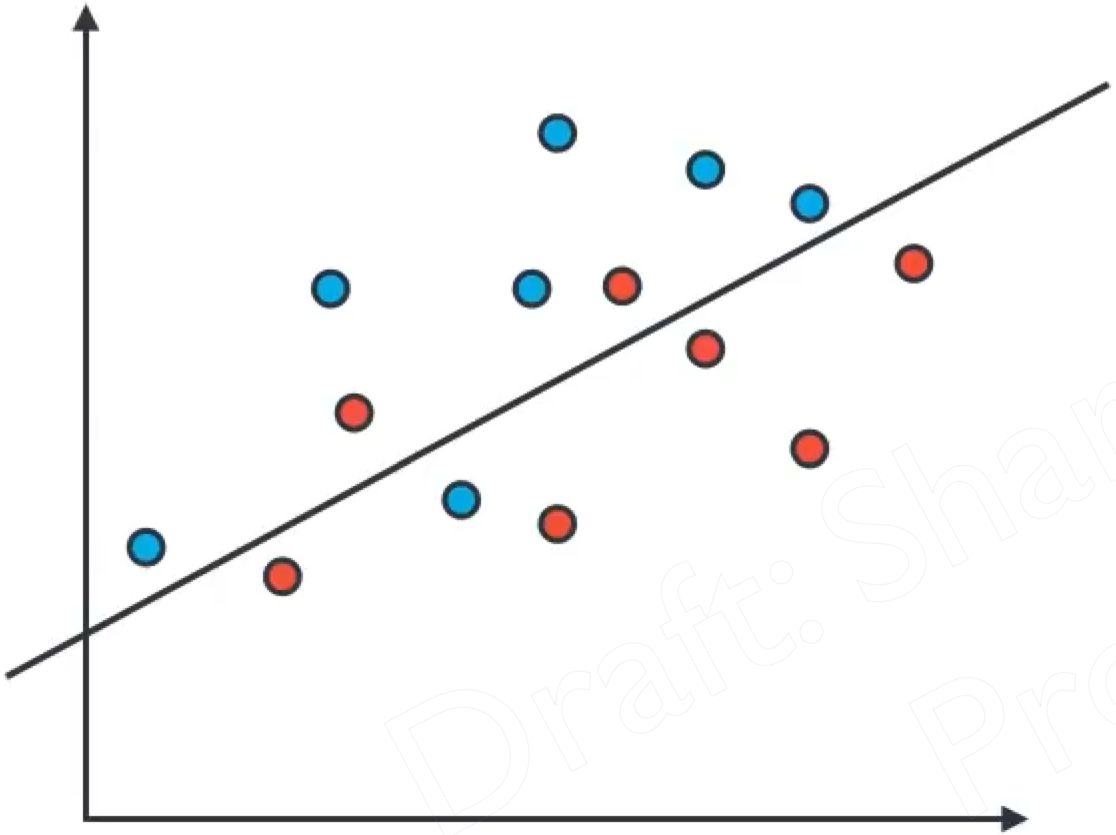
Recall

E-mail \ Folder	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?

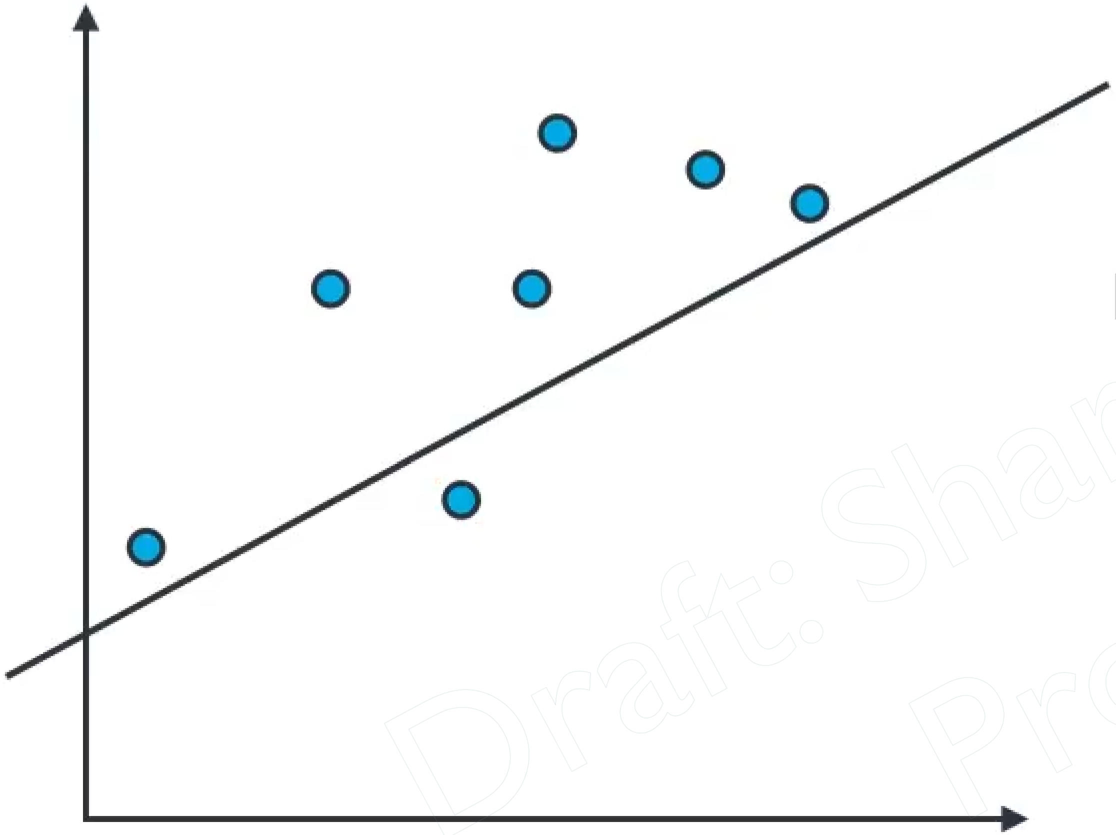
$$\text{Recall} = \frac{100}{100 + 170} = 37\%$$

Recall



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

Recall



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

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False Negatives}}$$

$$= \frac{6}{6 + 1}$$

$$= \frac{6}{7}$$

$$= 85.7\%$$

Precision and Recall

Medical Model

Precision: 55.7%
Recall: 83.3%

Spam Detector

Precision: 76.9%
Recall: 37%

Precision and Recall

Medical Model

Precision: 55.7%
Recall: 83.3%

One score?

Spam Detector

Precision: 76.9%
Recall: 37%

F1 Score

Medical Model

Precision: 55.7%
Recall: 83.3%

Average = 69.5%

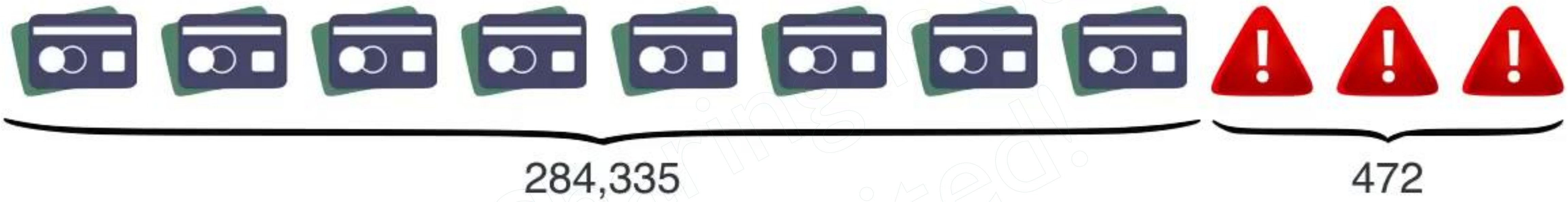
Spam Detector

Precision: 76.9%
Recall: 37%

Average = 56.95%

Credit Card Fraud

Credit Card Fraud



Credit Card Fraud

Model: All transactions are good.

Credit Card Fraud

284,335 472

Model: All transactions are good.

Precision = 100%

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

Credit Card Fraud

284,335

472

Model: All transactions are good.

Precision = 100%

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

Average = 50%

Credit Card Fraud

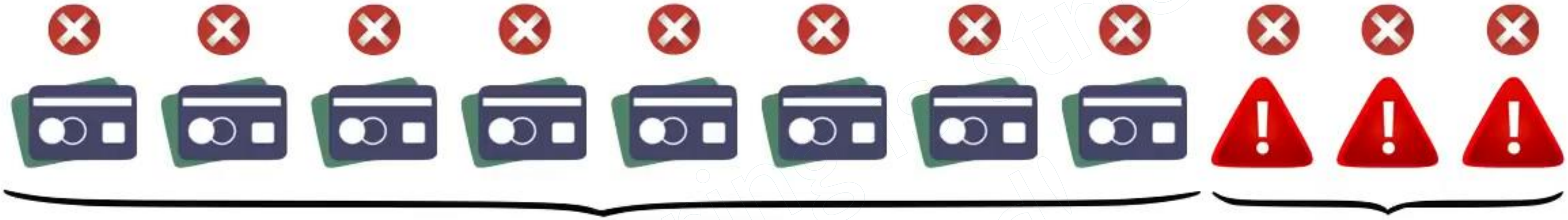
Credit Card Fraud

Model: All transactions are fraudulent.

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

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

Credit Card Fraud



Model: All transactions are fraudulent.

$$\text{Precision} = \frac{472}{284{,}807} = .016\% \qquad \text{Recall} = \frac{472}{472} = 100\%$$

Average = 50.008%

Harmonic mean

Harmonic mean

Arithmetic Mean = $\dfrac{x+y}{2}$

Harmonic mean

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

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

Harmonic mean

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

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

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

Harmonic mean

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

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

Harmonic mean

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

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)~~

Harmonic mean

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

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

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)

F1 Score

Medical Model

Precision = 55.7%

Recall = 83.3%

Average = 69.5%

F1 Score

Medical Model

Precision = 55.7%

Recall = 83.3%

Average = 69.5%

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

F1 Score

Spam Detector Model

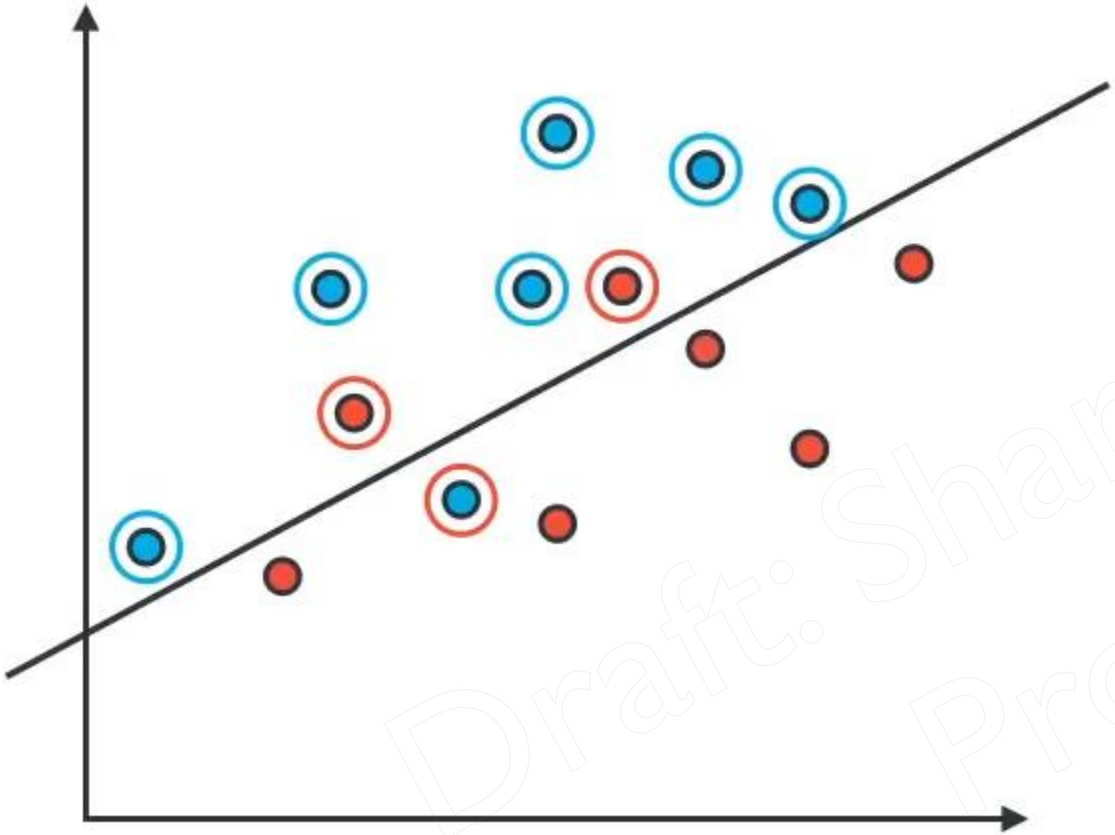
Precision = 76.9%

Recall = 37%

Average = 56.95%

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

F1 Score

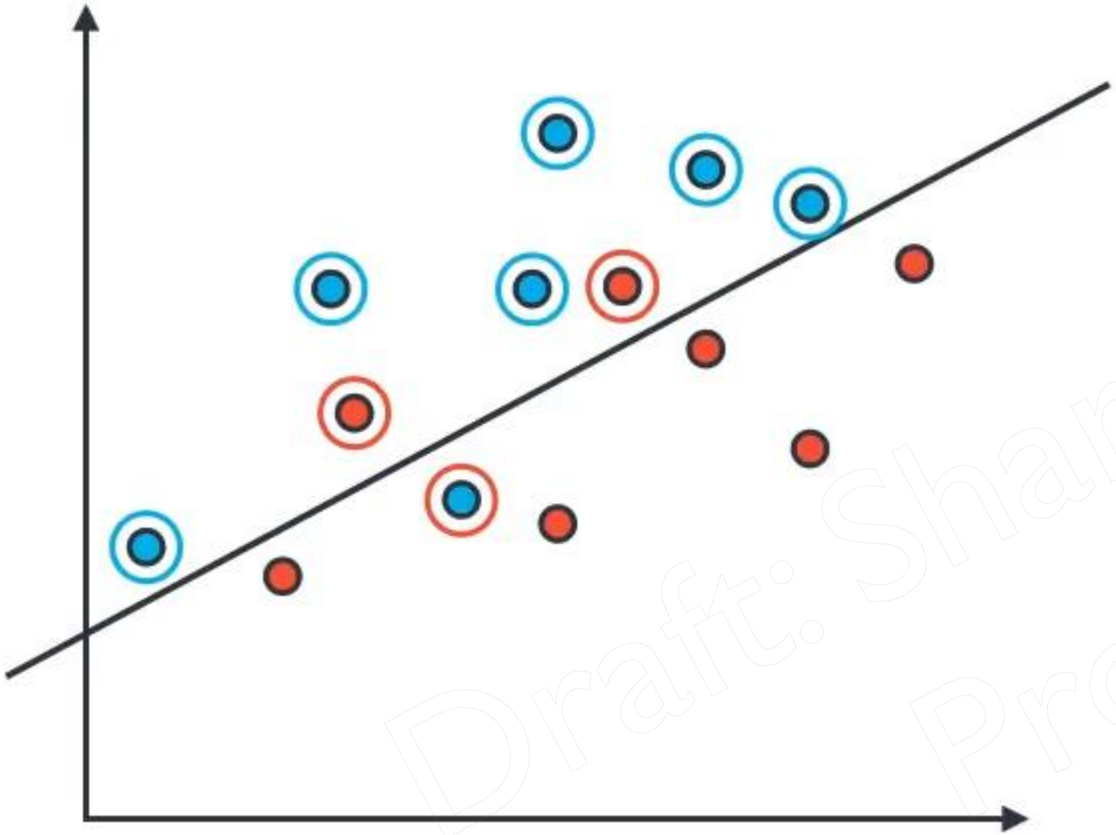


Precision = 75%

Recall = 85.7%

Average = 80.35

F1 Score



Precision = 75%

Recall = 85.7%

Average = 80.35

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

Types of Errors

Types of Errors

Types of Errors

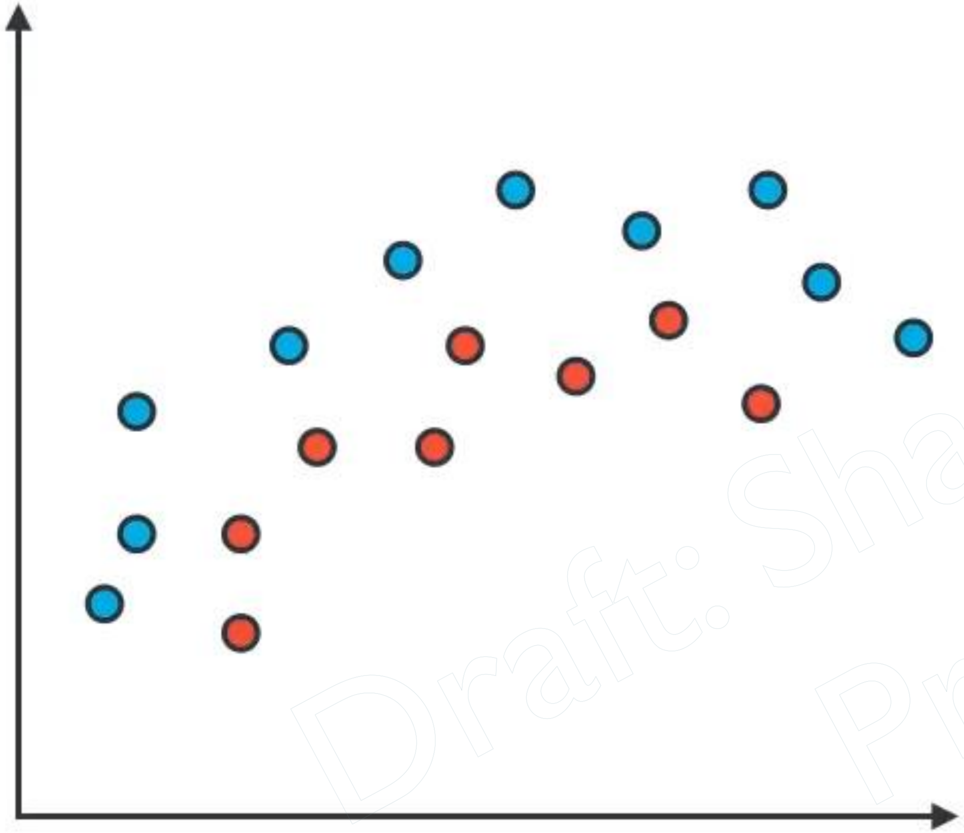
Underfitting

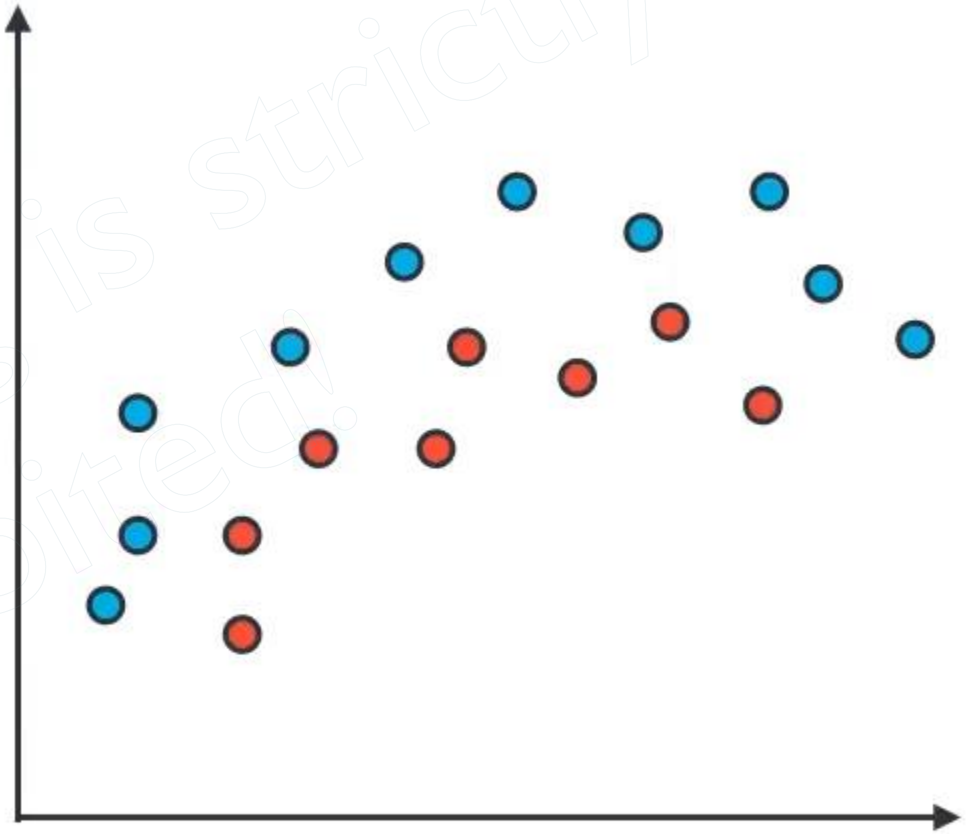
Types of Errors

Underfitting

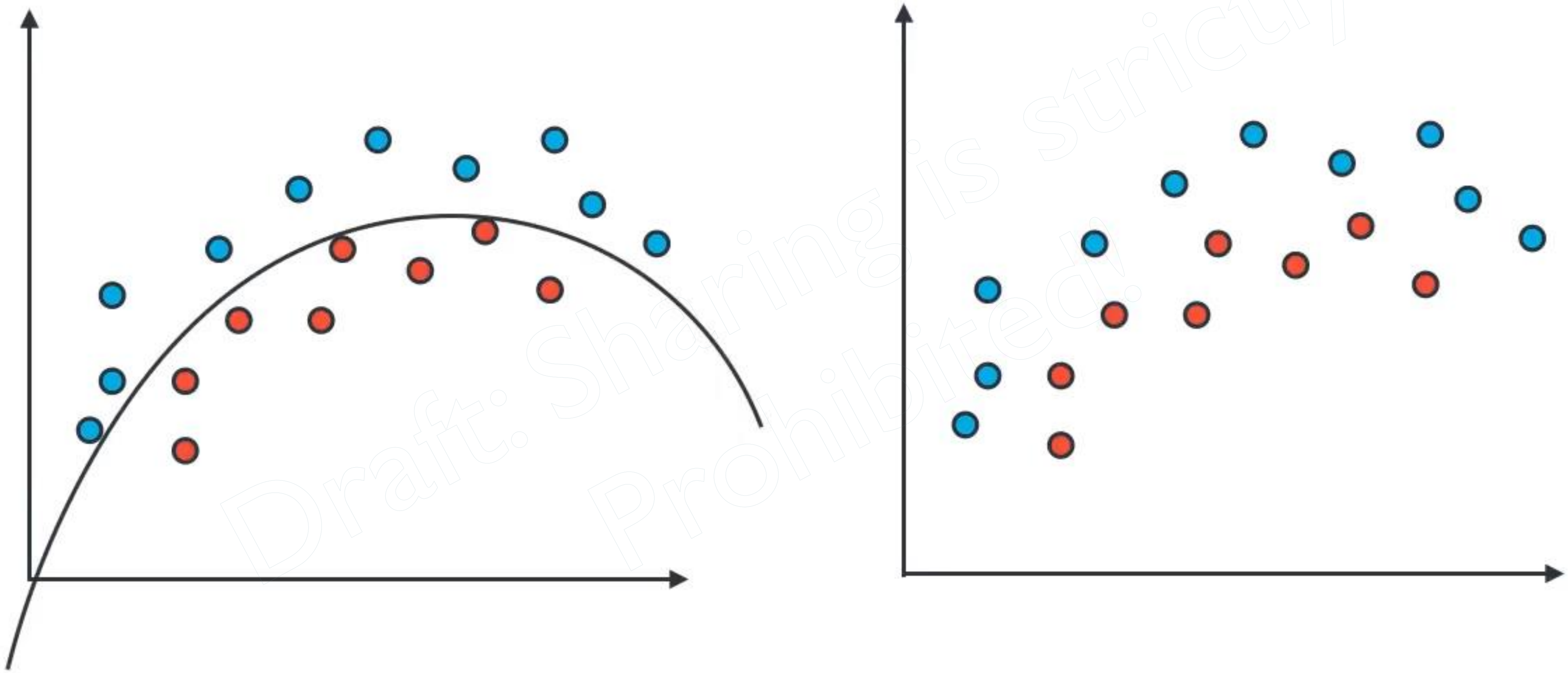
Overfitting

Error due to bias (underfitting)

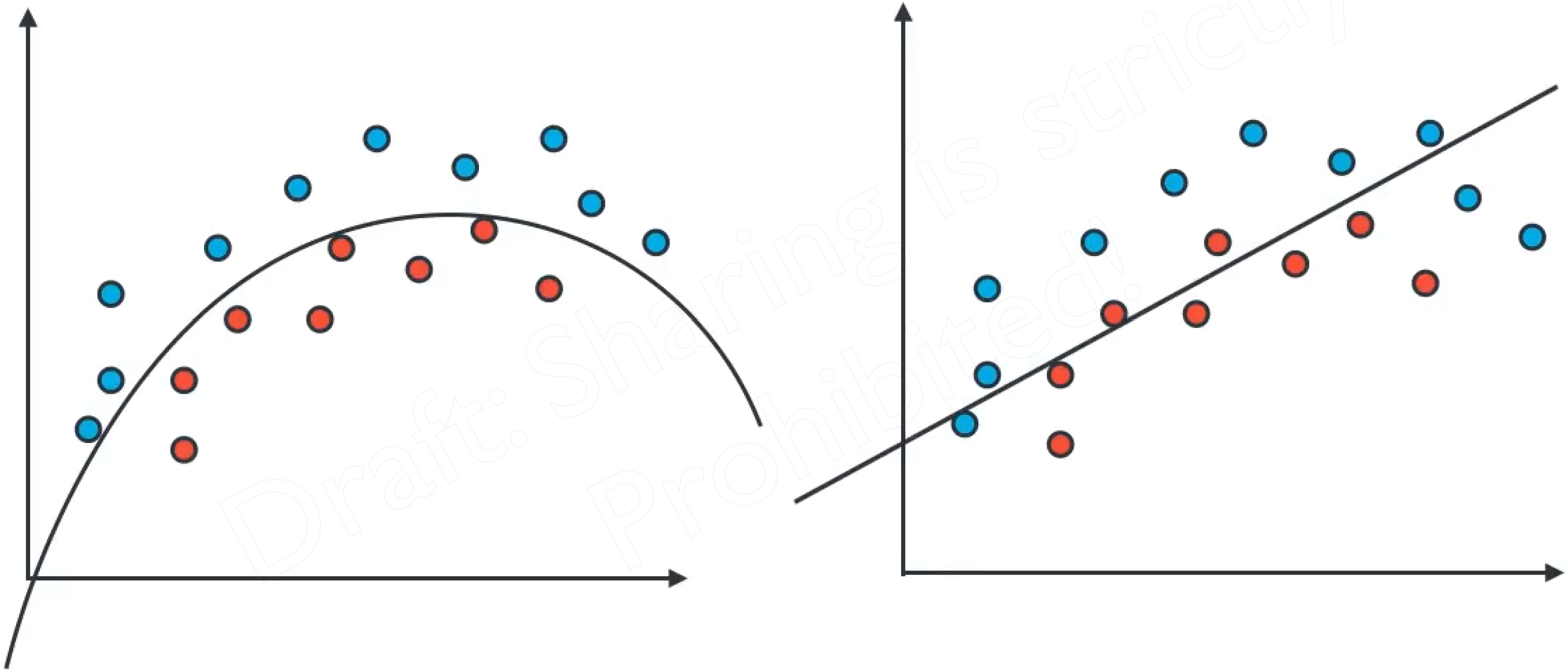




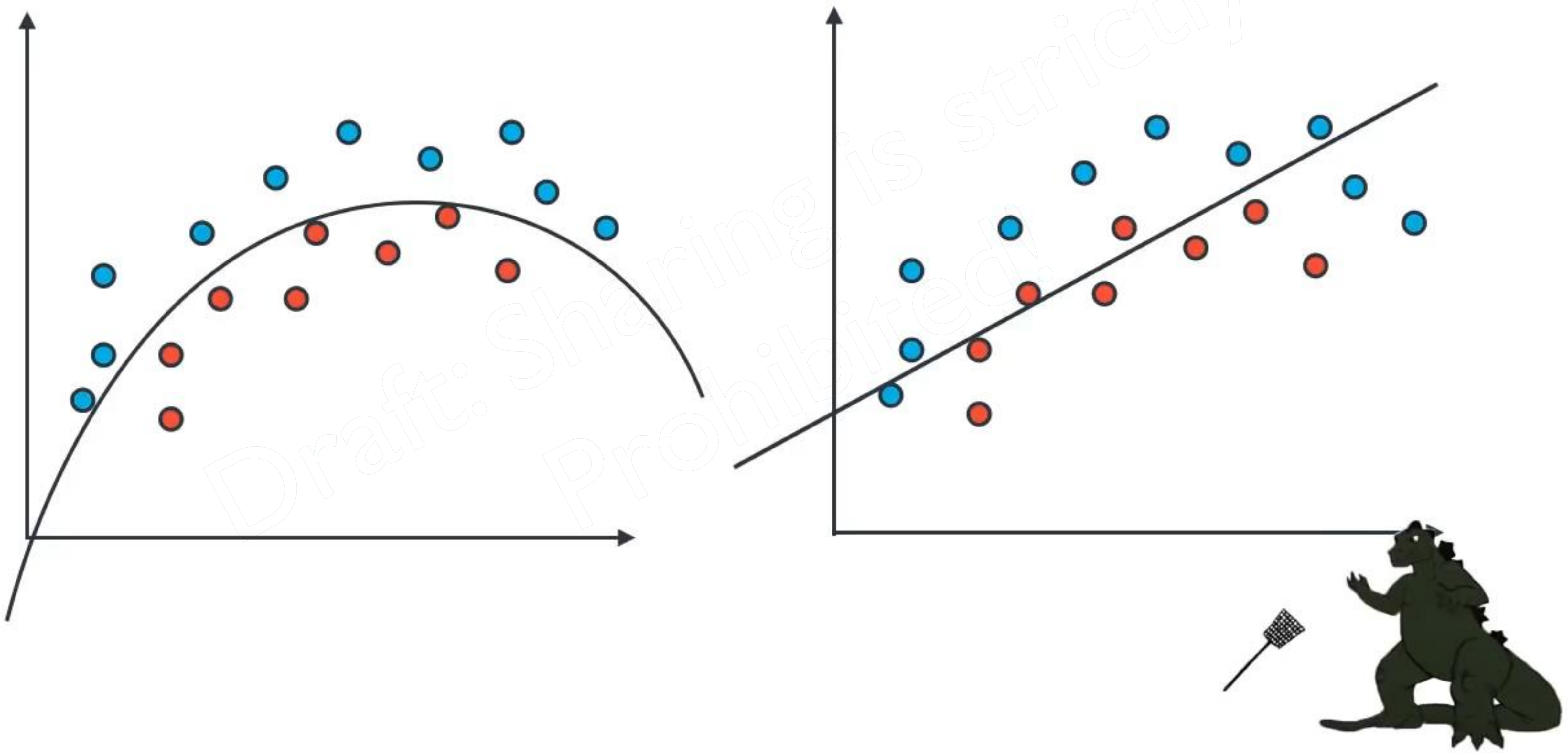
Error due to bias (underfitting)



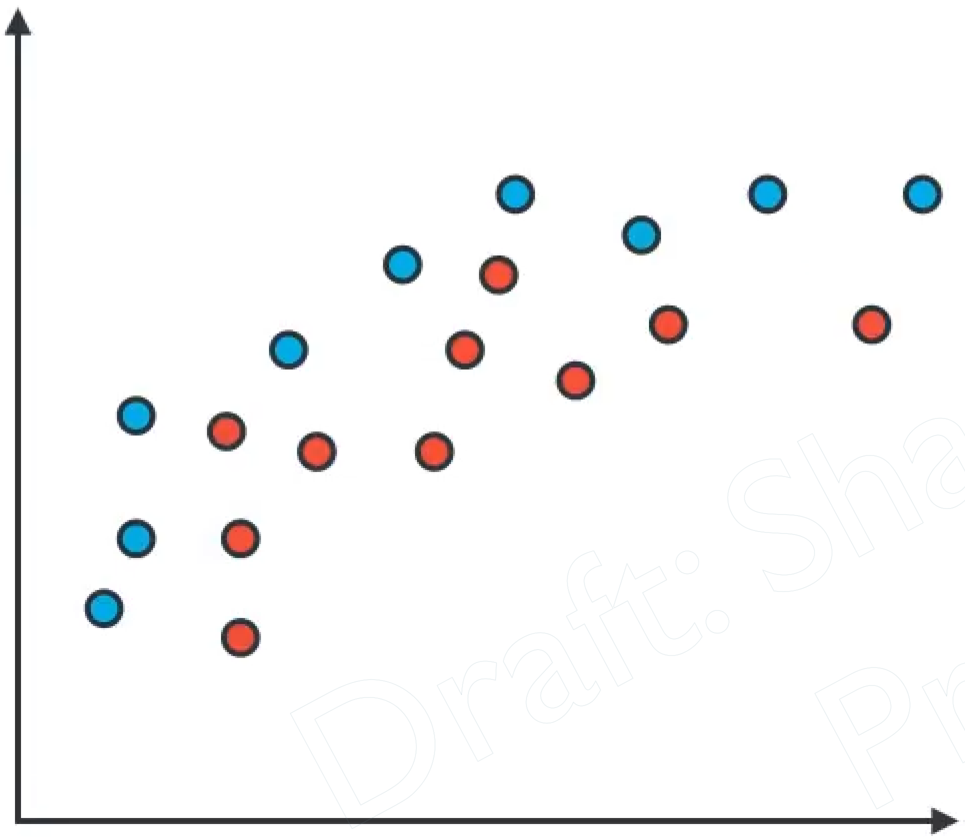
Error due to bias (underfitting)

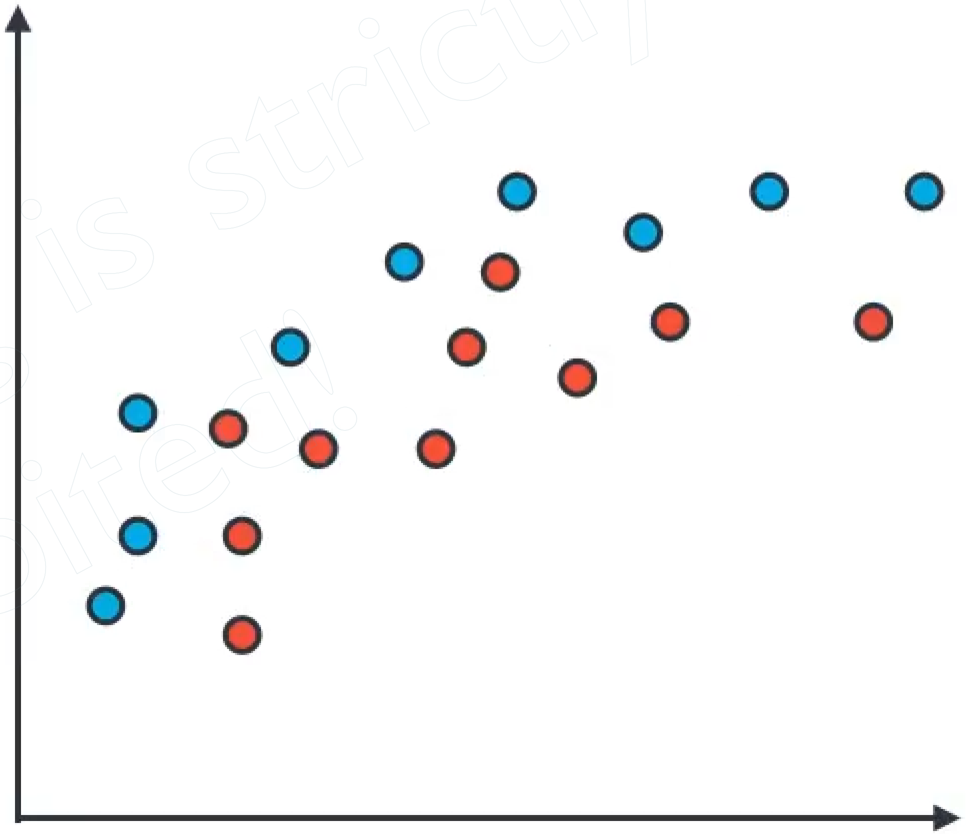


Error due to bias (underfitting)

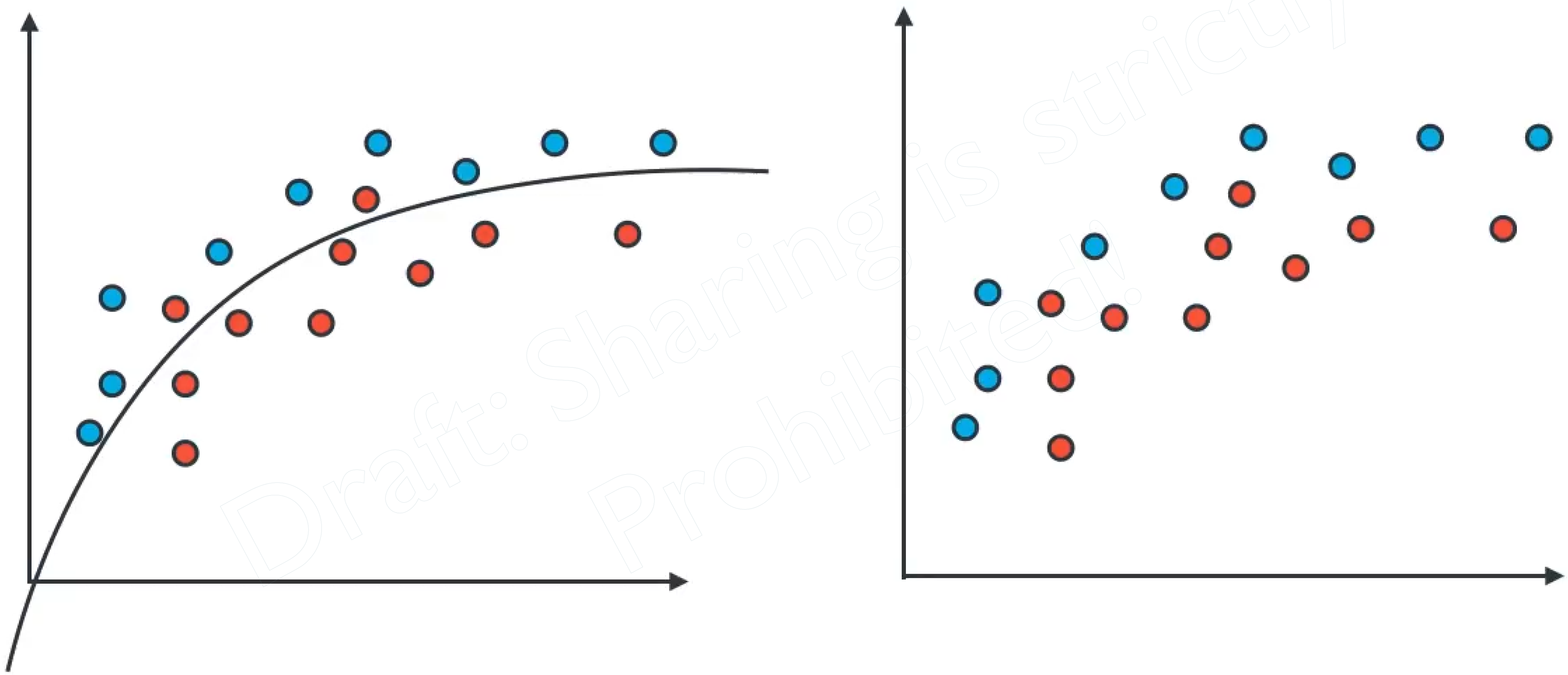


Error due to variance (overfitting)

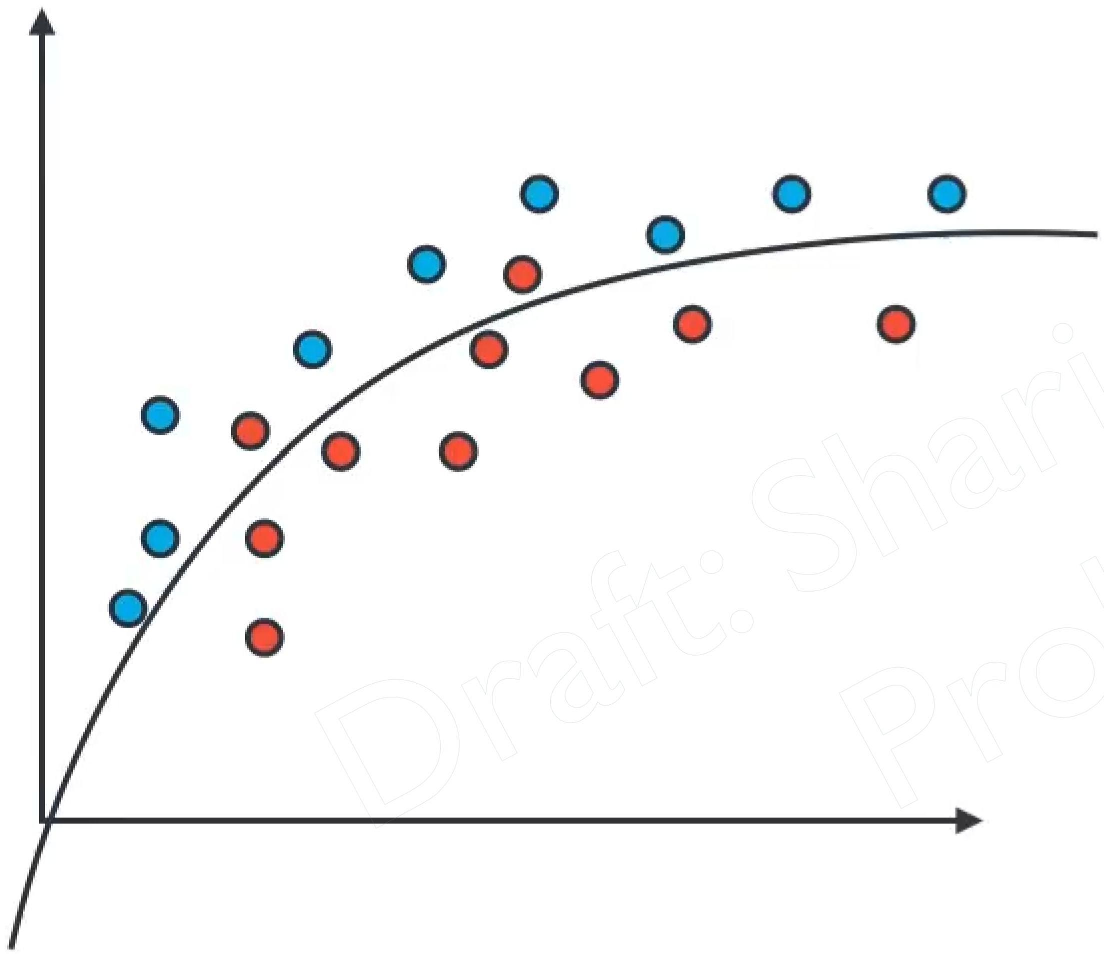


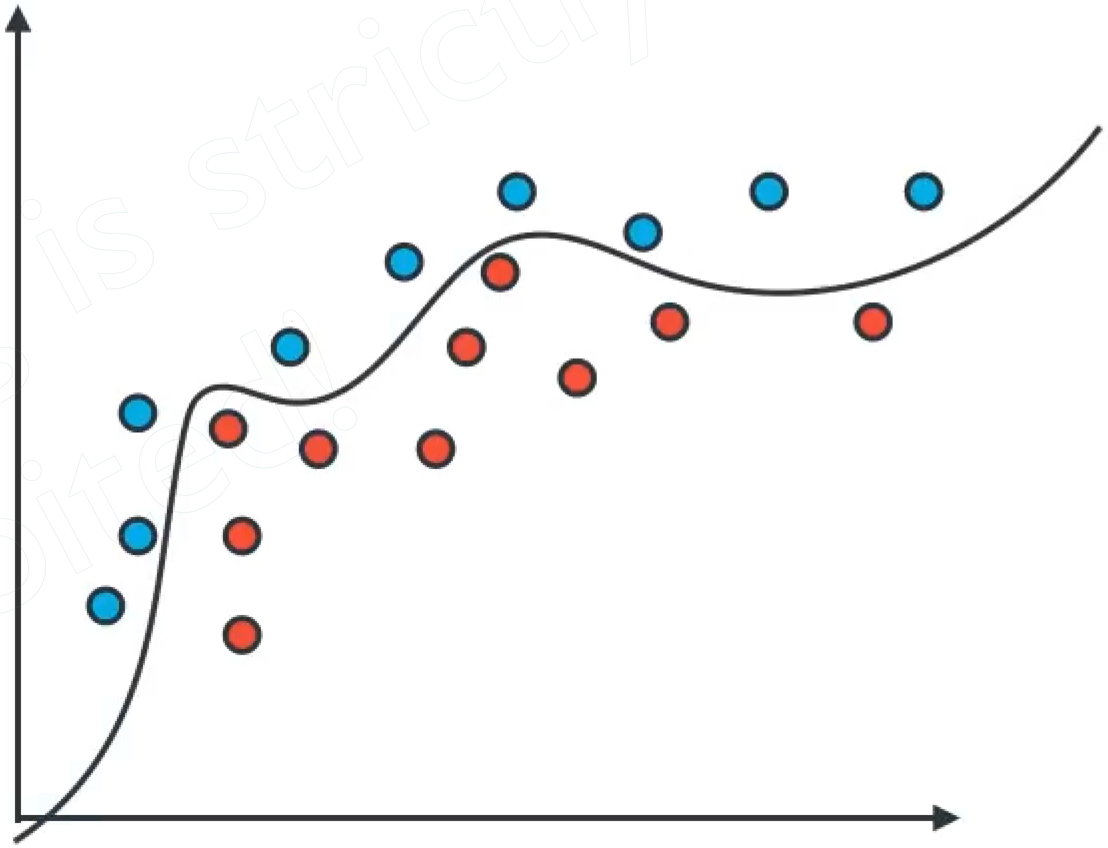


Error due to variance (overfitting)

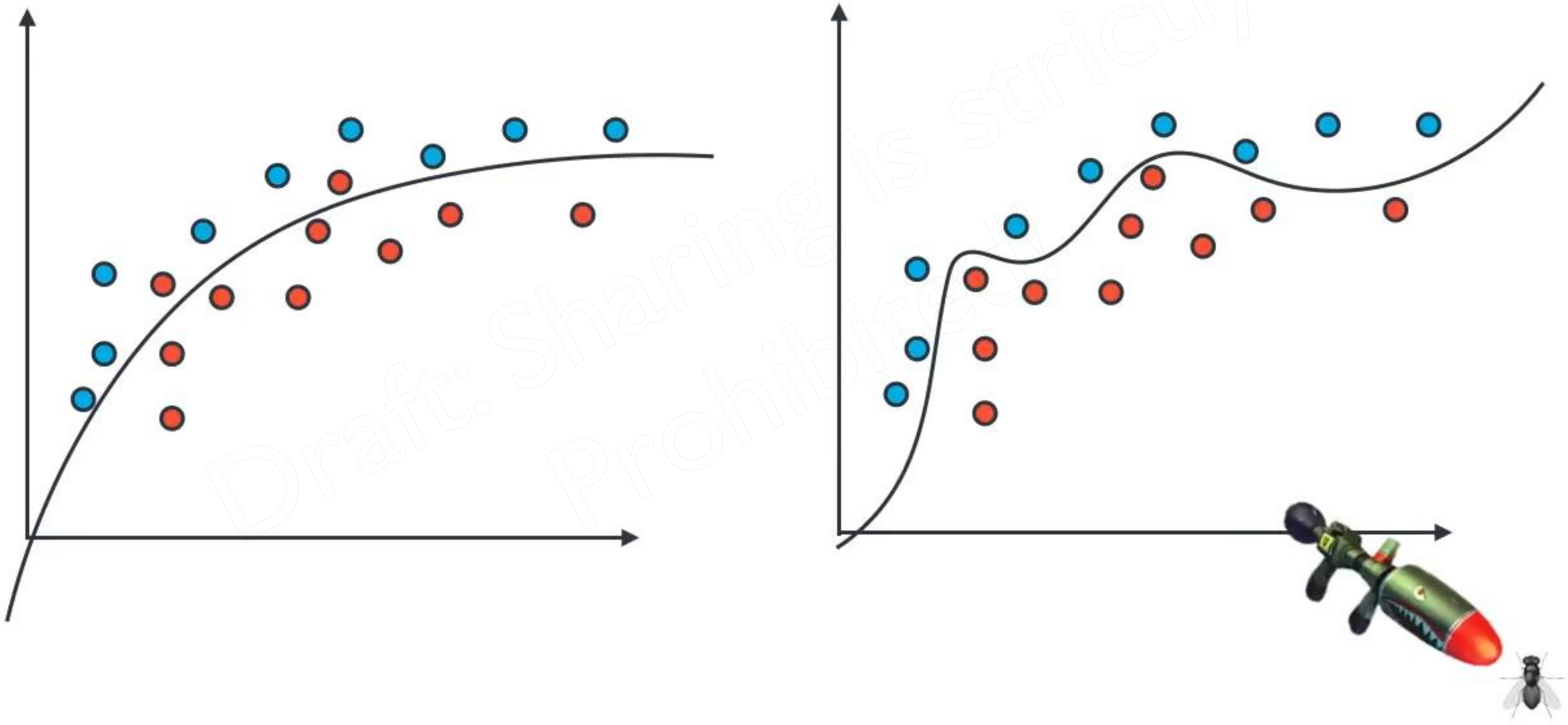


Error due to variance (overfitting)

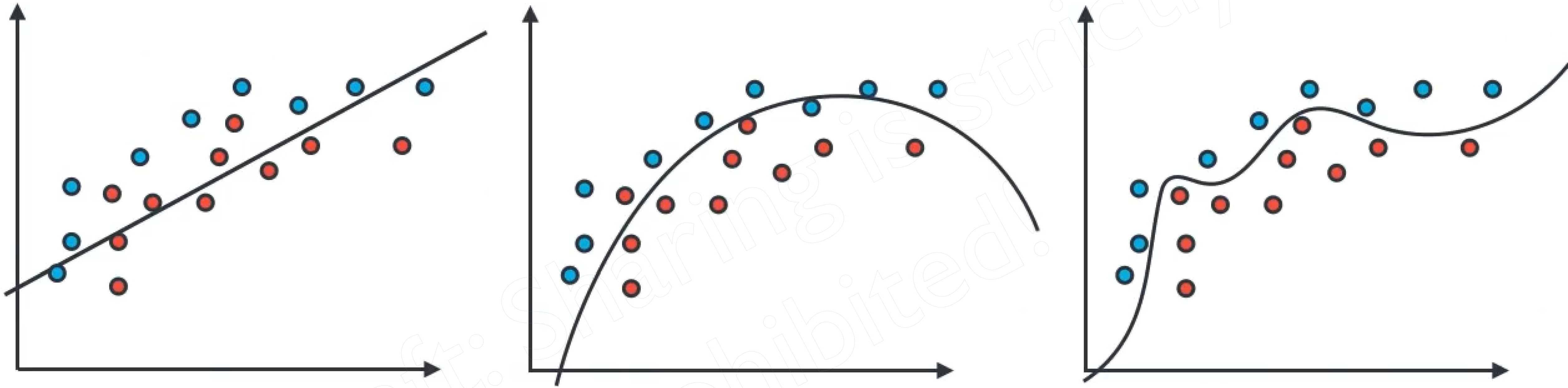




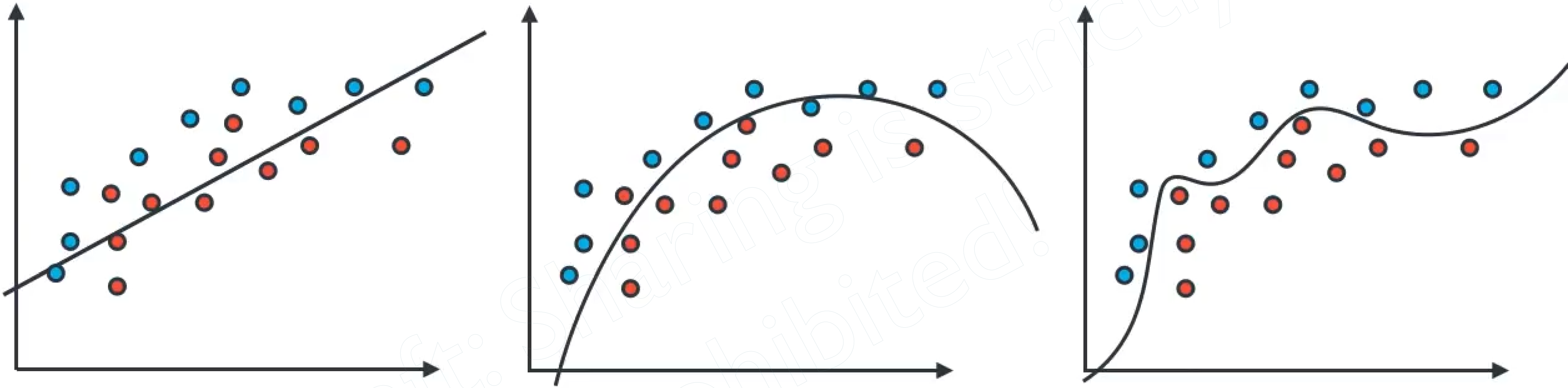
Error due to variance (overfitting)



Model Complexity Graph



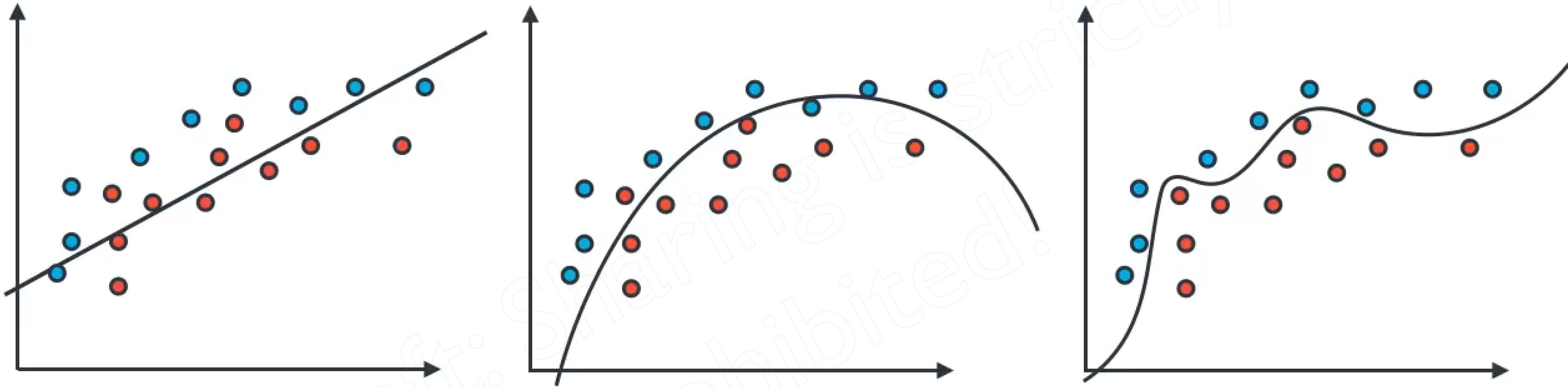
Model Complexity Graph



High Bias

degree = 1

Model Complexity Graph



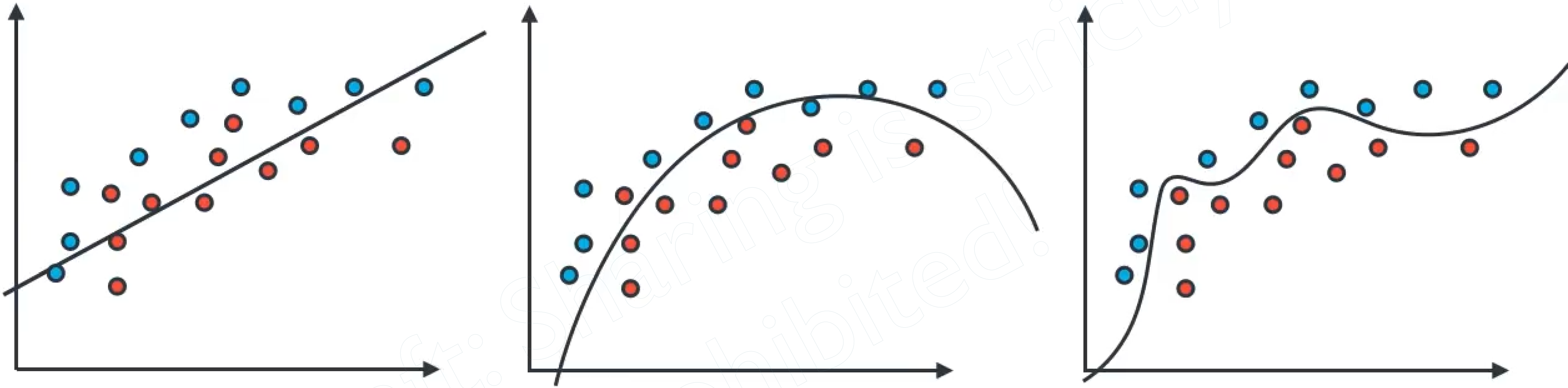
High Bias

degree = 1

High Variance

degree = 6

Model Complexity Graph



High Bias

degree = 1

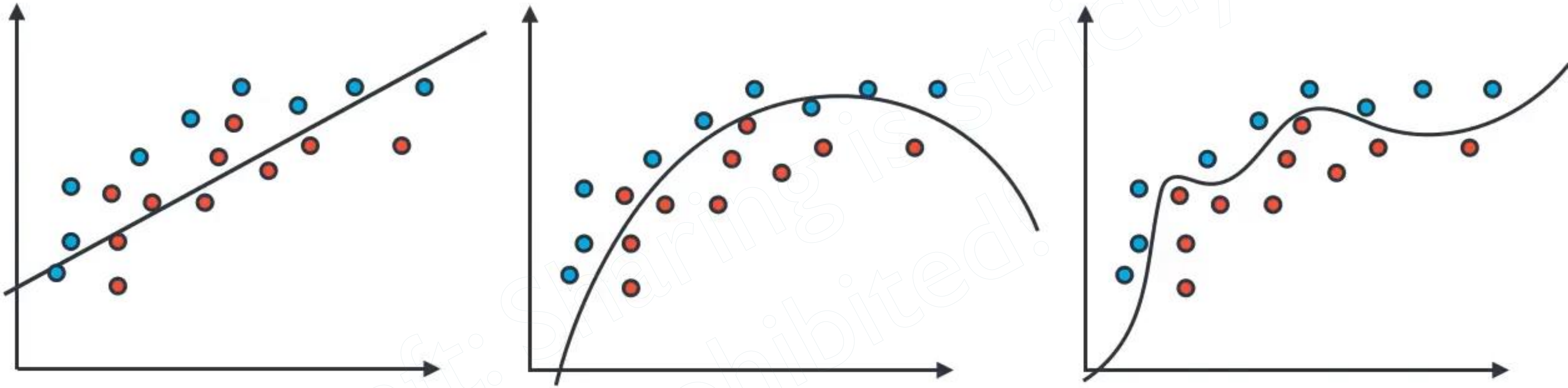
Just Right

degree = 2

High Variance

degree = 6

Model Complexity Graph



High Bias

degree = 1

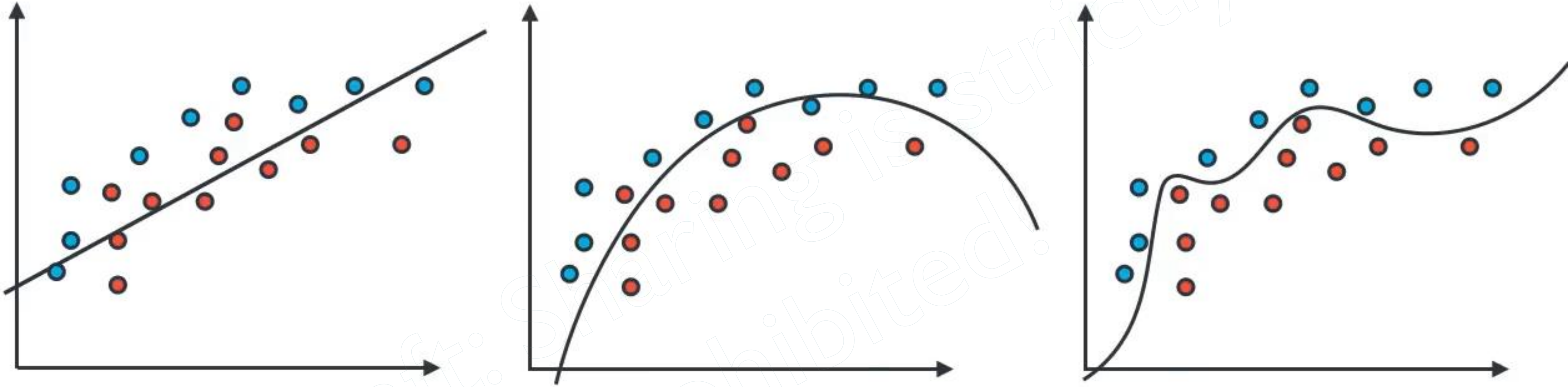
Just Right

degree = 2

High Variance

degree = 6

Model Complexity Graph



High Bias

degree = 1

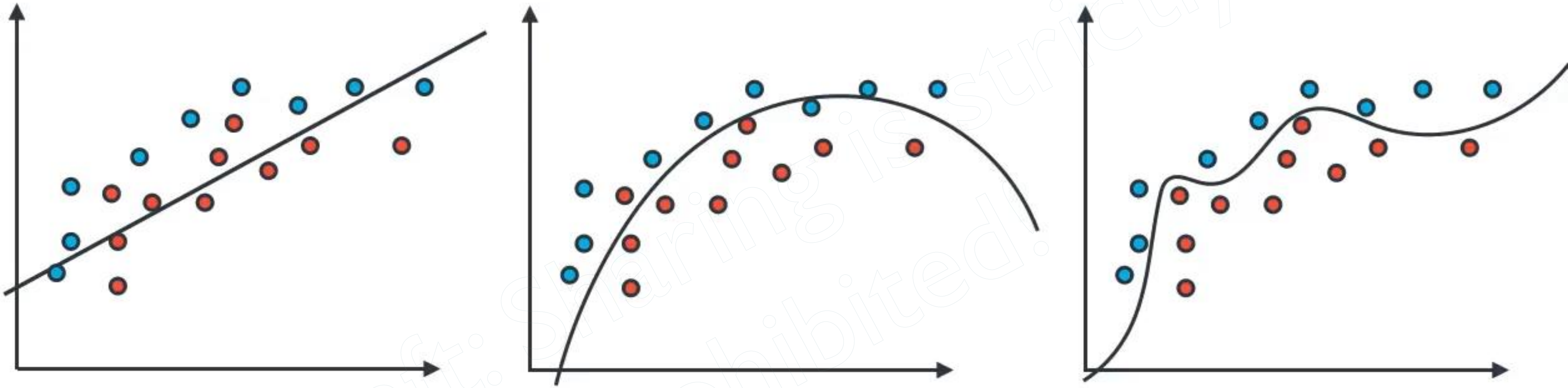
Just Right

degree = 2

High Variance

degree = 6

Model Complexity Graph



High Bias

degree = 1

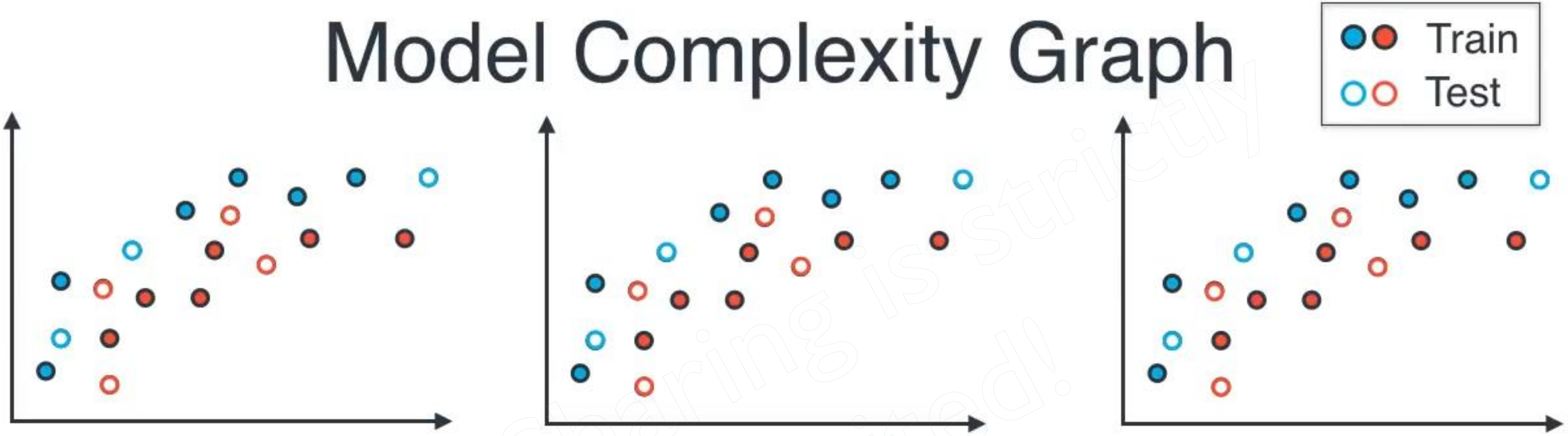
Just Right

degree = 2

High Variance

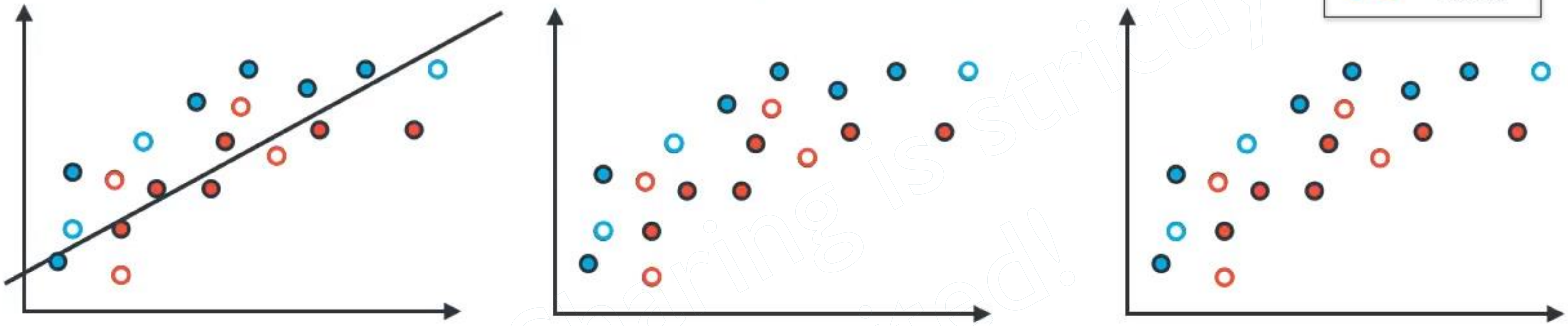
degree = 6

Model Complexity Graph



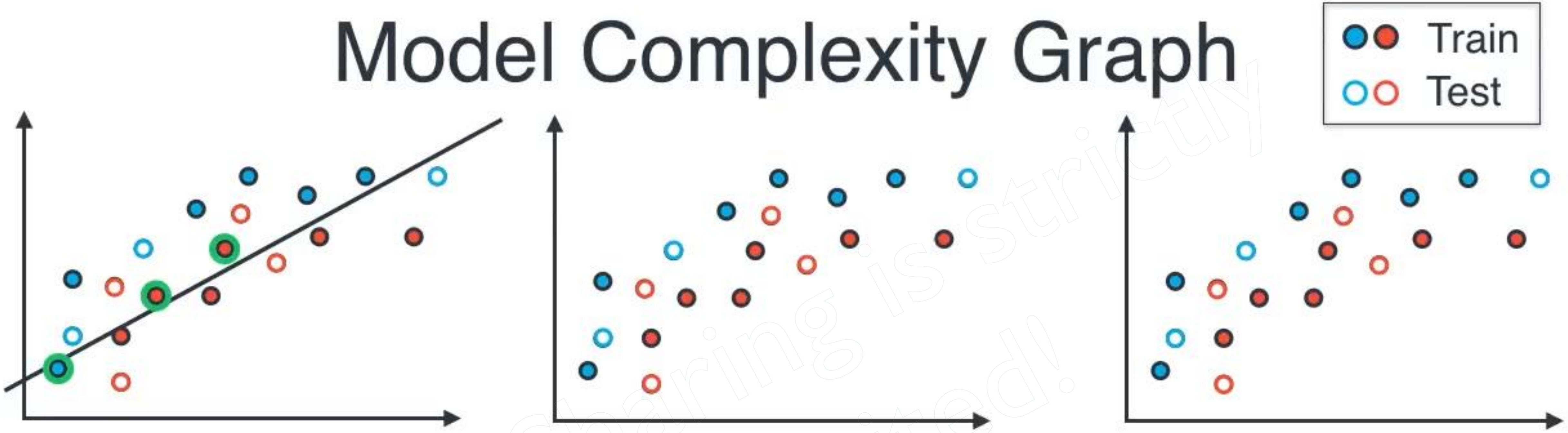
Degree 1 | Degree 2 | Degree 6

Model Complexity Graph



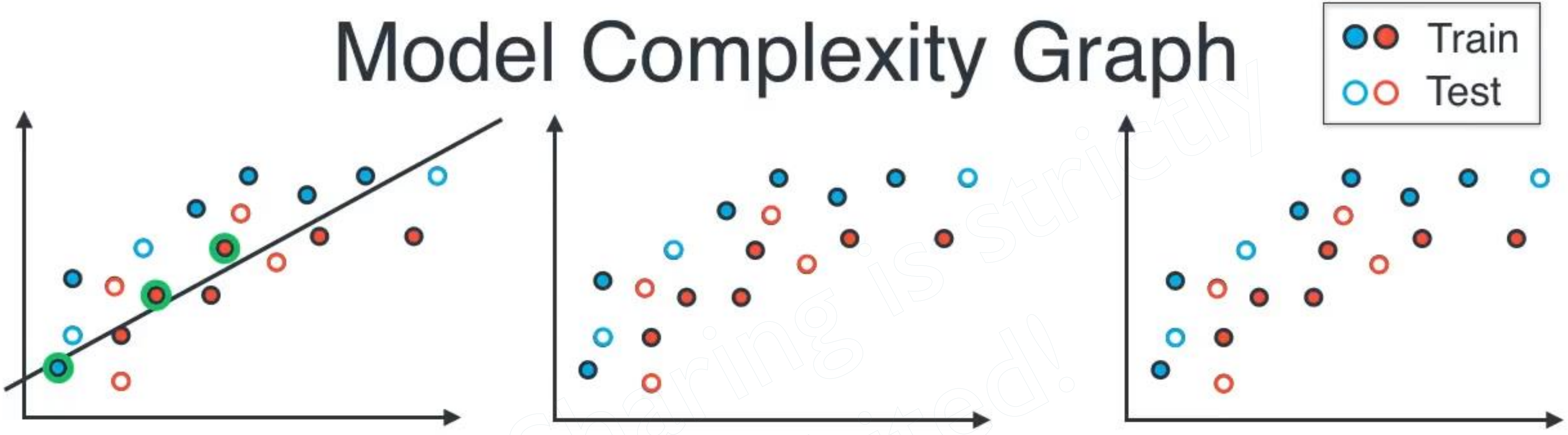
Degree 1 | Degree 2 | Degree 6

Model Complexity Graph



Degree 1 Degree 2 Degree 6

Model Complexity Graph

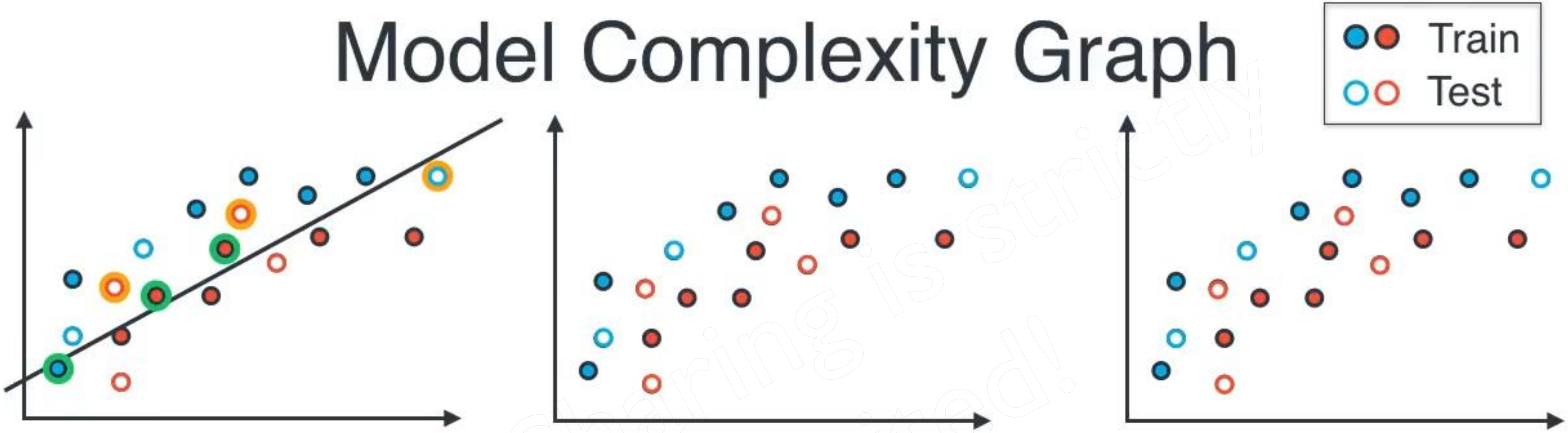


Degree 1
Training Error: 3

Degree 2

Degree 6

Model Complexity Graph



Degree 1

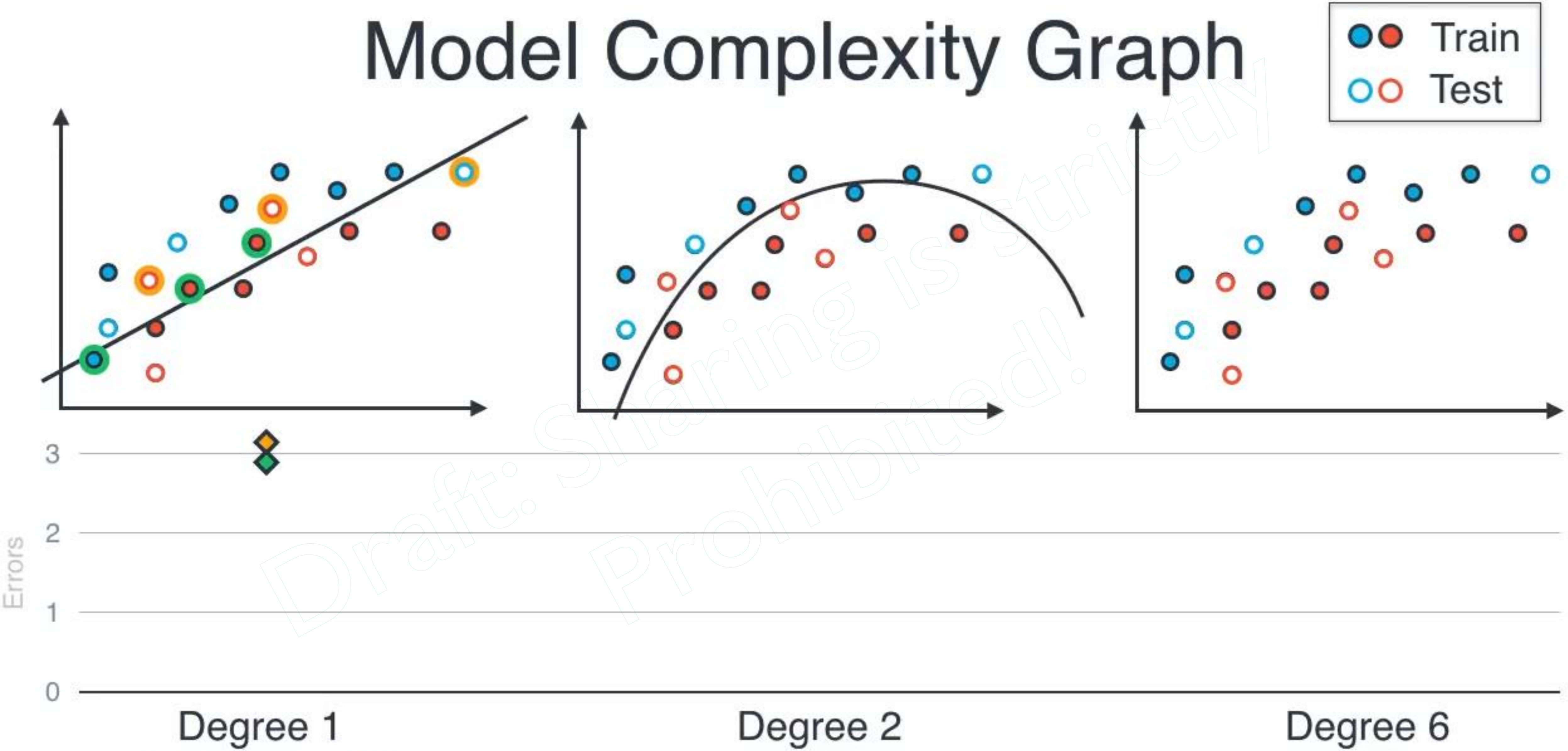
Training Error: 3

Testing Error: 3

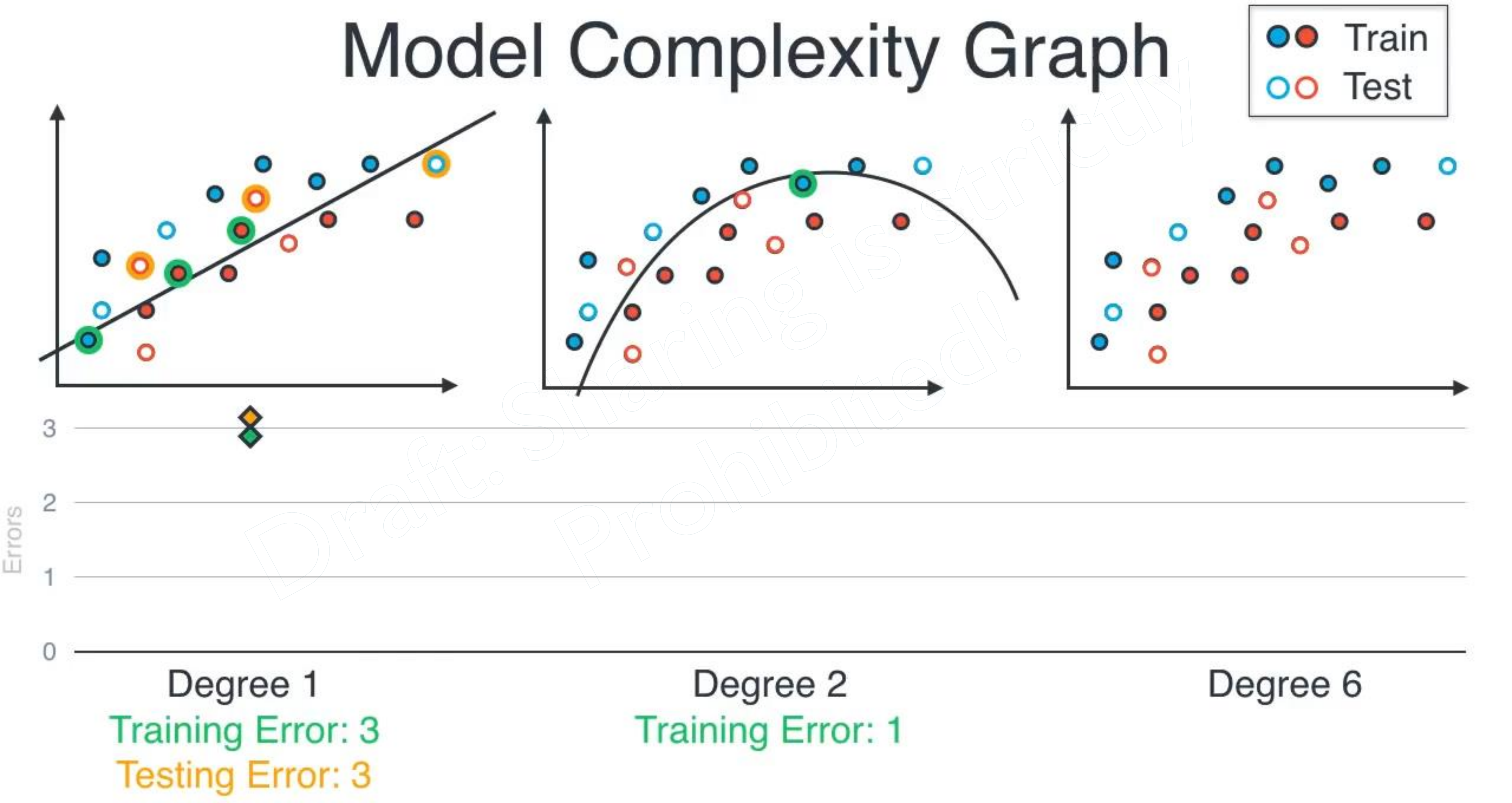
Degree 2

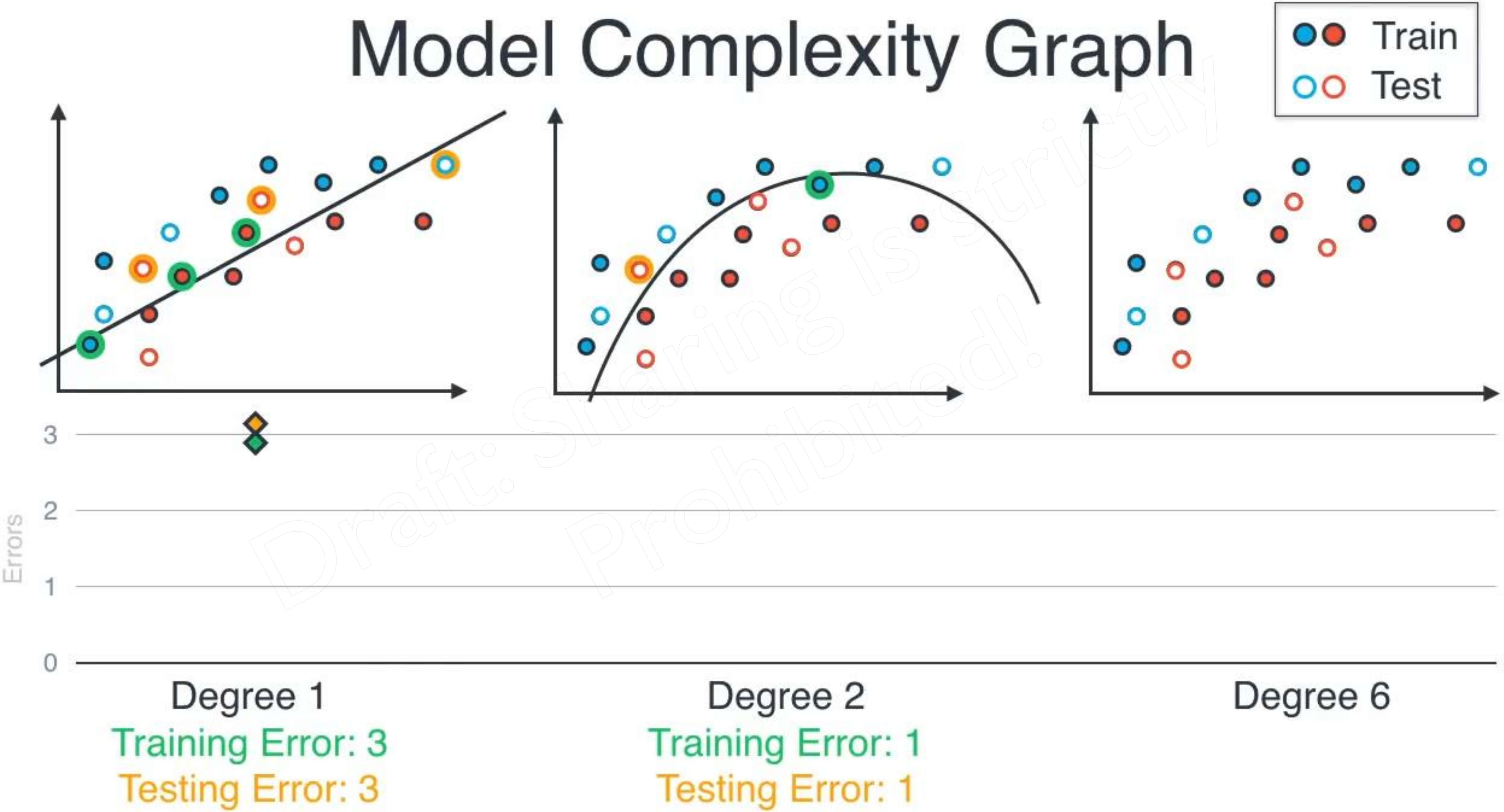
Degree 6

Model Complexity Graph

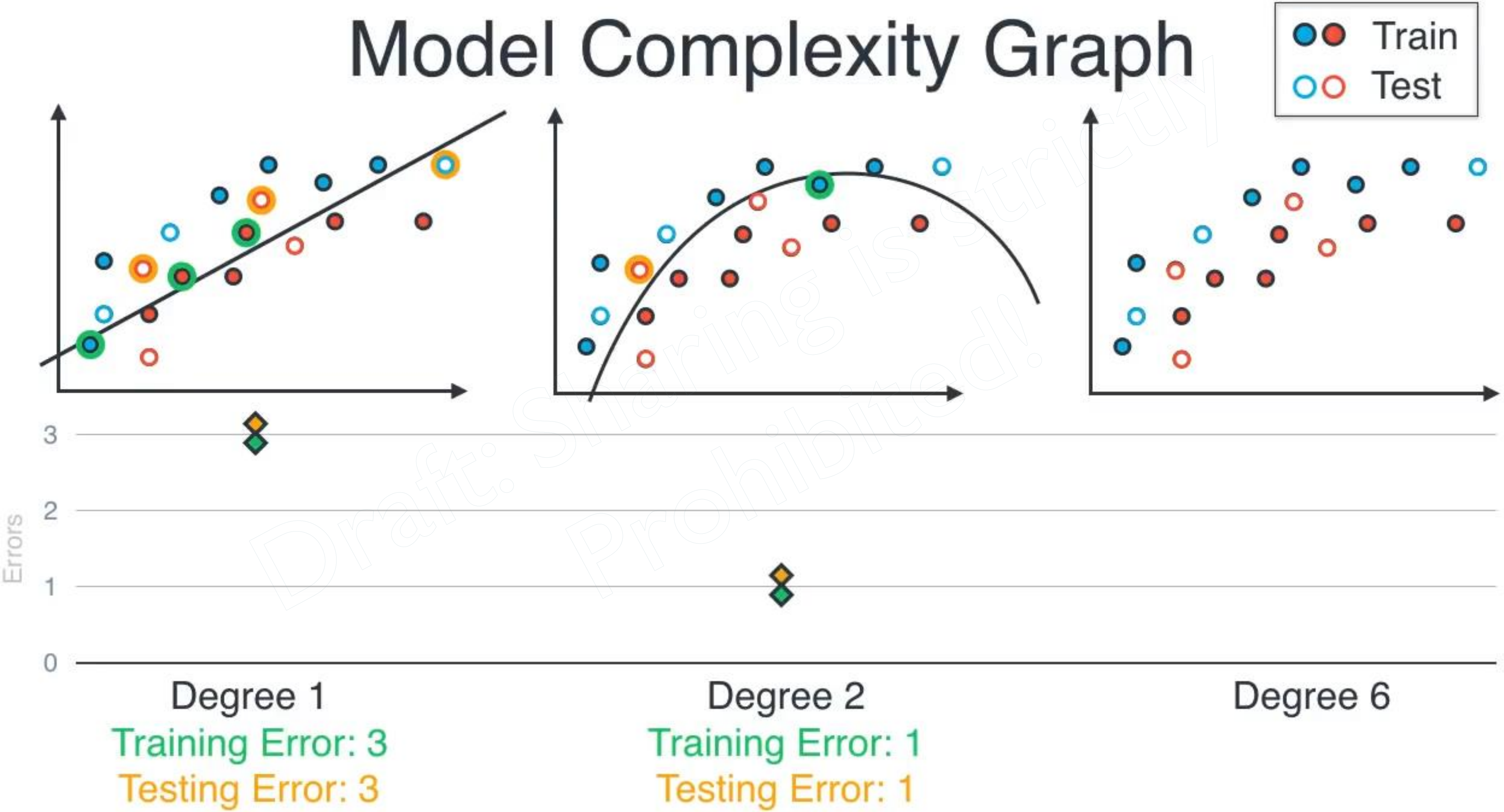


Training Error: 3
Testing Error: 3

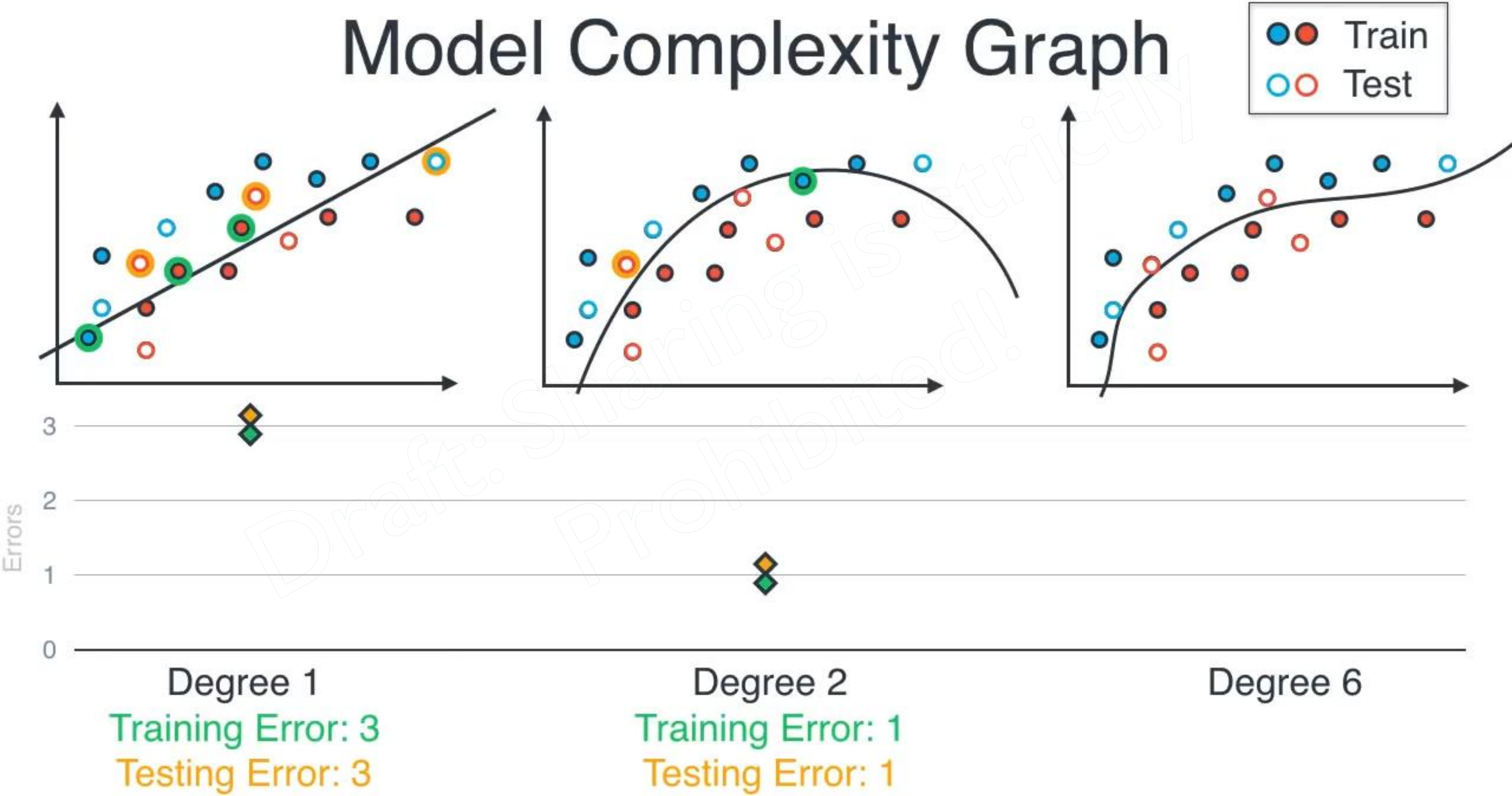

Model Complexity Graph
Train
Test
Errors
3
2
1
0
Degree 1
Training Error: 3
Testing Error: 3
Degree 2
Training Error: 1
Degree 6

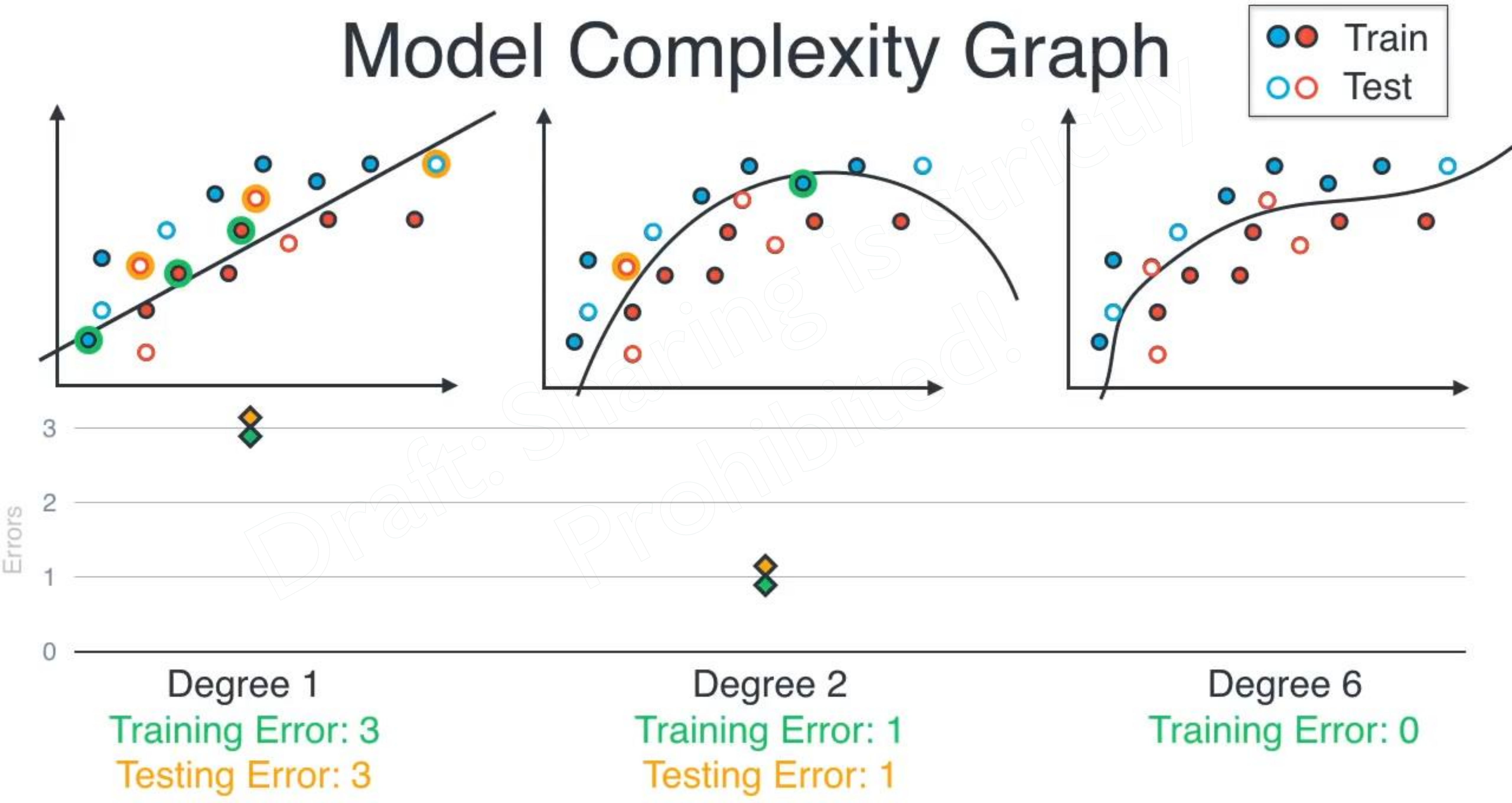

Model Complexity Graph
Train
Test
3
2
1
0
Errors
Degree 1
Training Error: 3
Testing Error: 3
Degree 2
Training Error: 1
Testing Error: 1
Degree 6

Model Complexity Graph

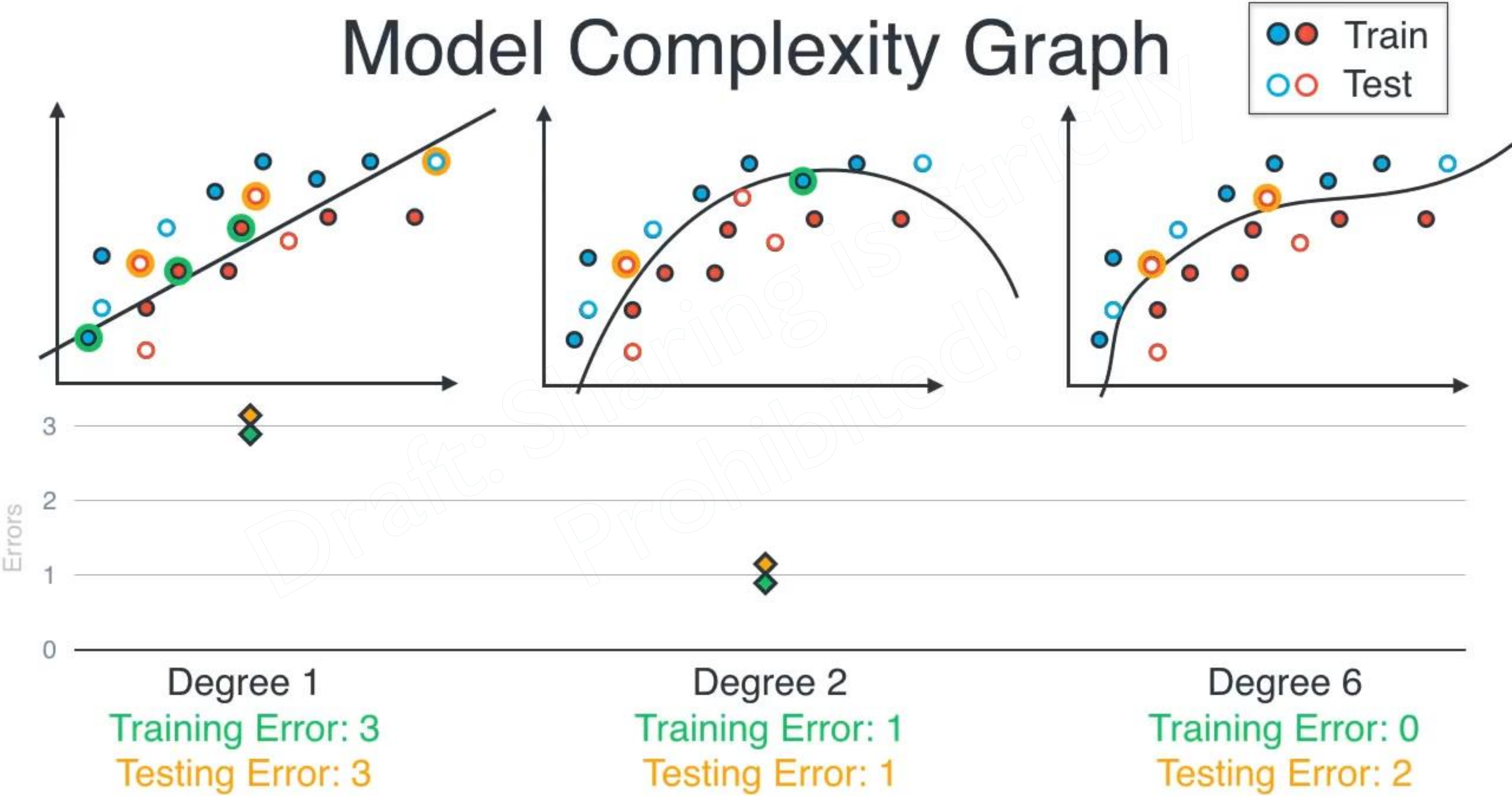


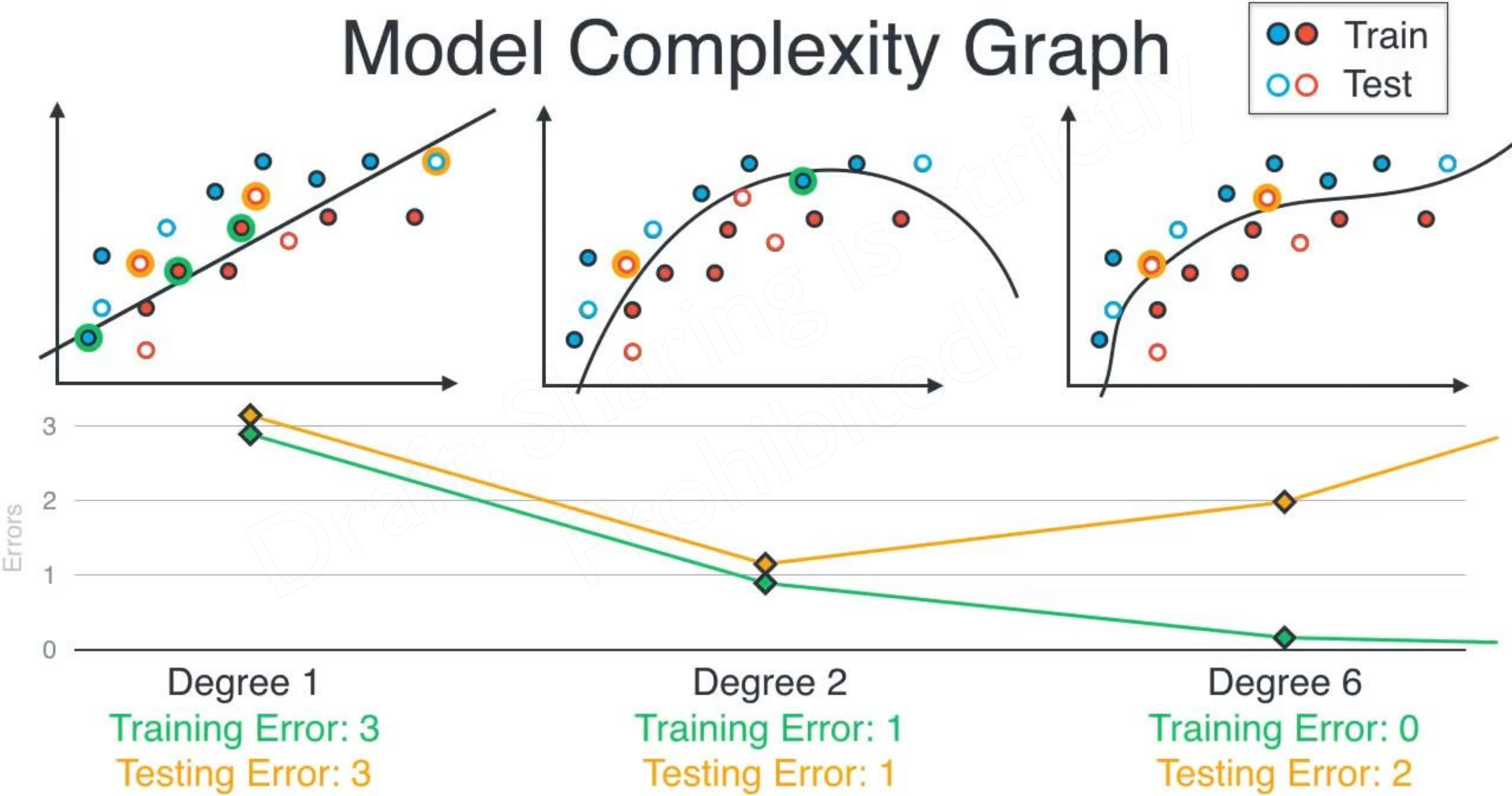
Model Complexity Graph

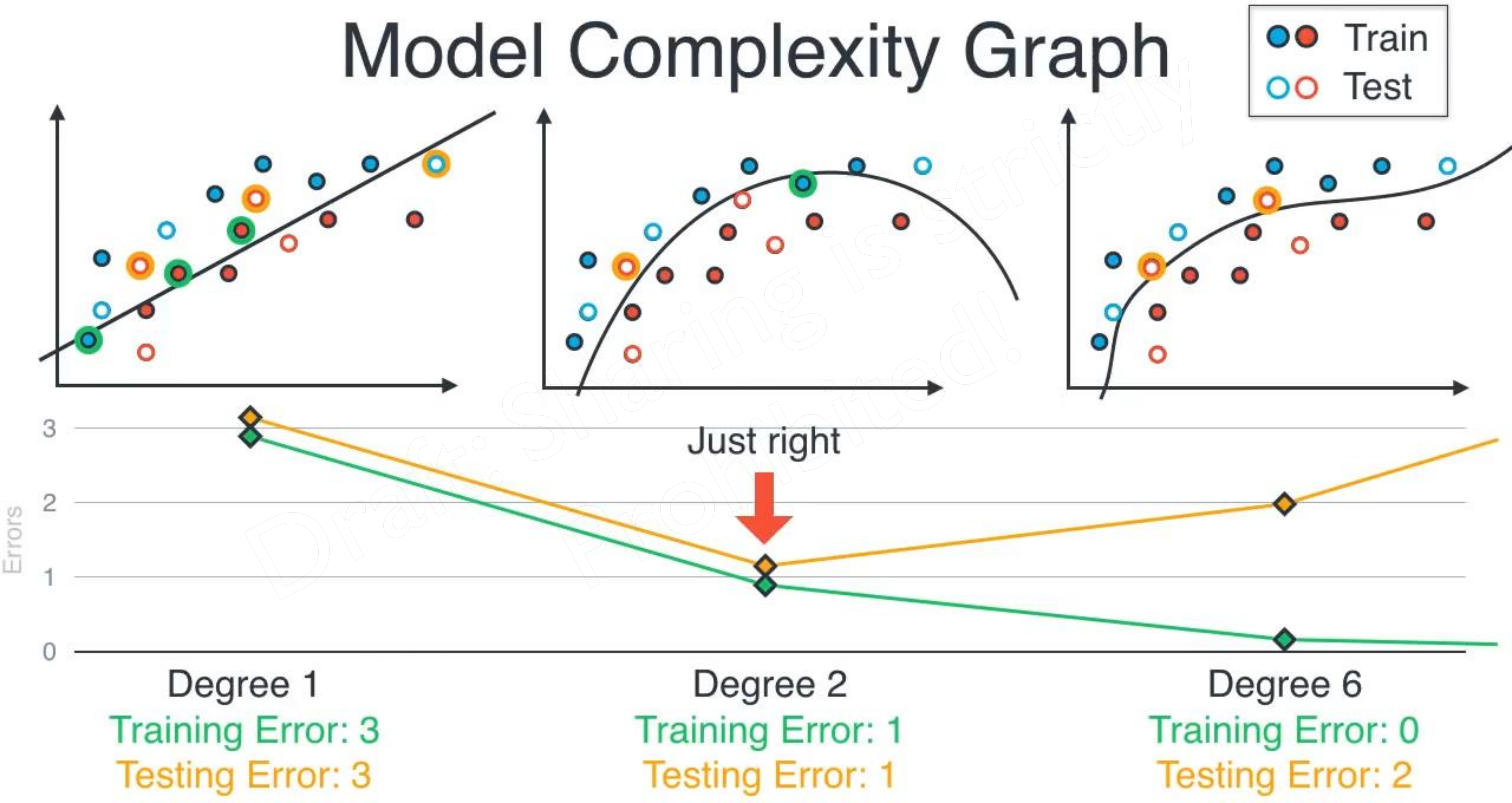



Model Complexity Graph
Train
Test
Errors
3
2
1
0
Degree 1
Training Error: 3
Testing Error: 3
Degree 2
Training Error: 1
Testing Error: 1
Degree 6
Training Error: 0

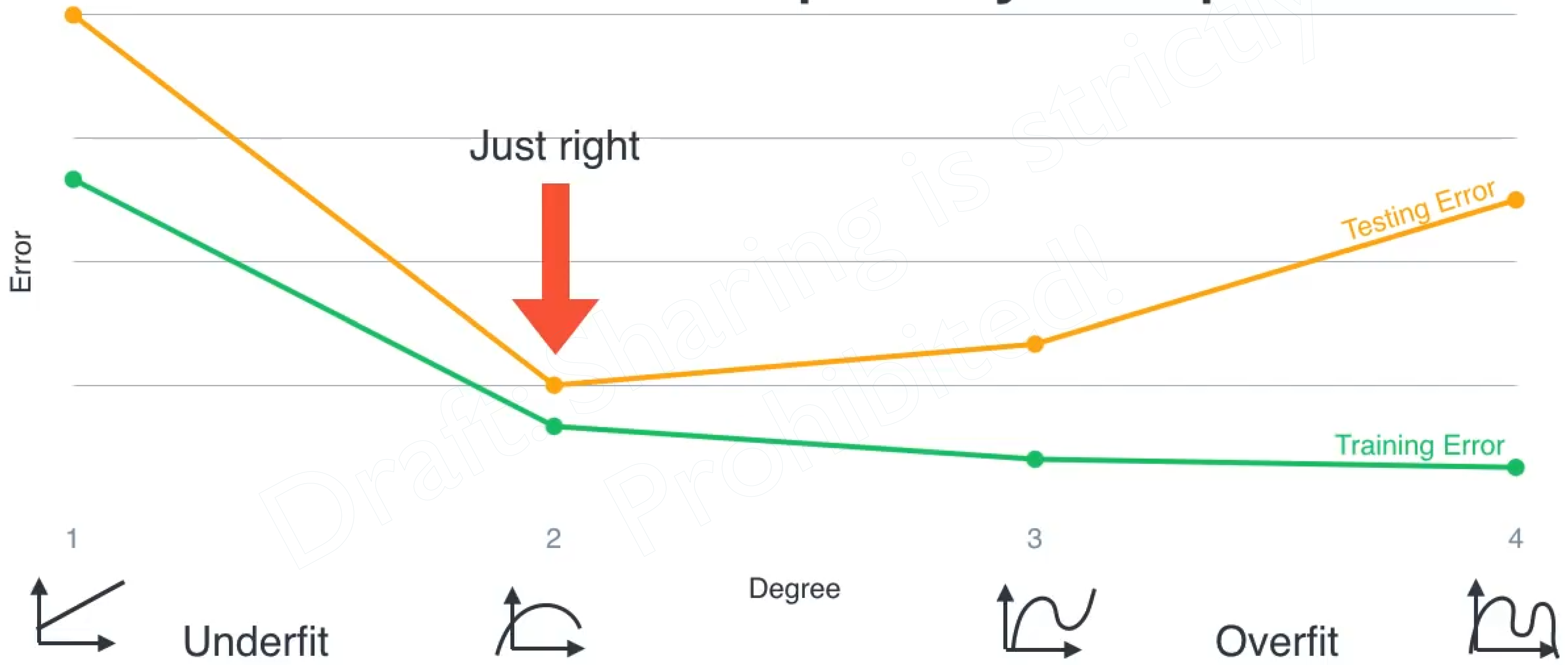
Model Complexity Graph




Model Complexity Graph
Train
Test
Errors
0
1
2
3
Degree 1
Training Error: 3
Testing Error: 3
Degree 2
Training Error: 1
Testing Error: 1
Degree 6
Training Error: 0
Testing Error: 2

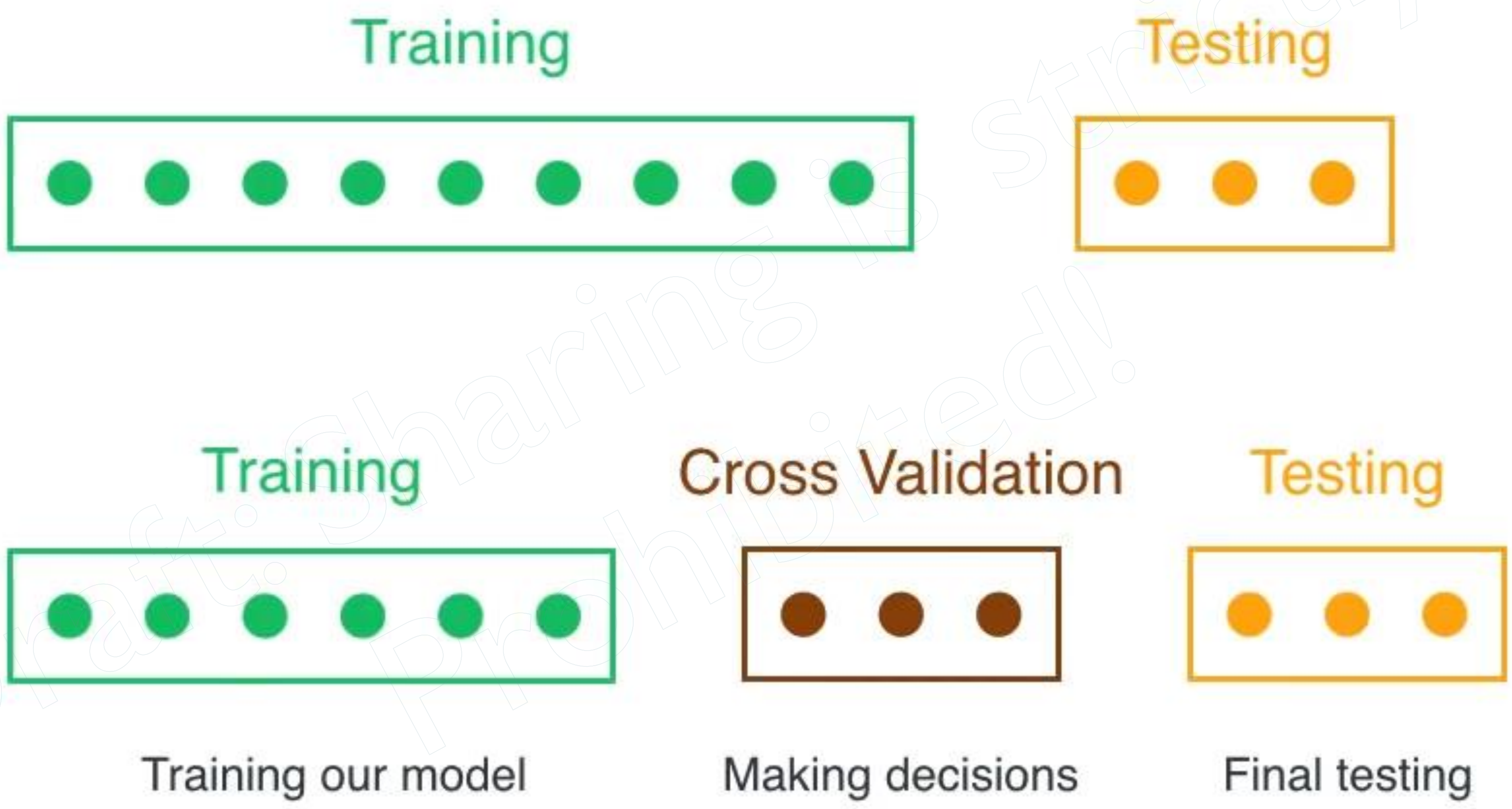

Model Complexity Graph
Train
Test
3
2
1
0
Errors
Just right
Degree 1
Training Error: 3
Testing Error: 3
Degree 2
Training Error: 1
Testing Error: 1
Degree 6
Training Error: 0
Testing Error: 2

Model Complexity Graph

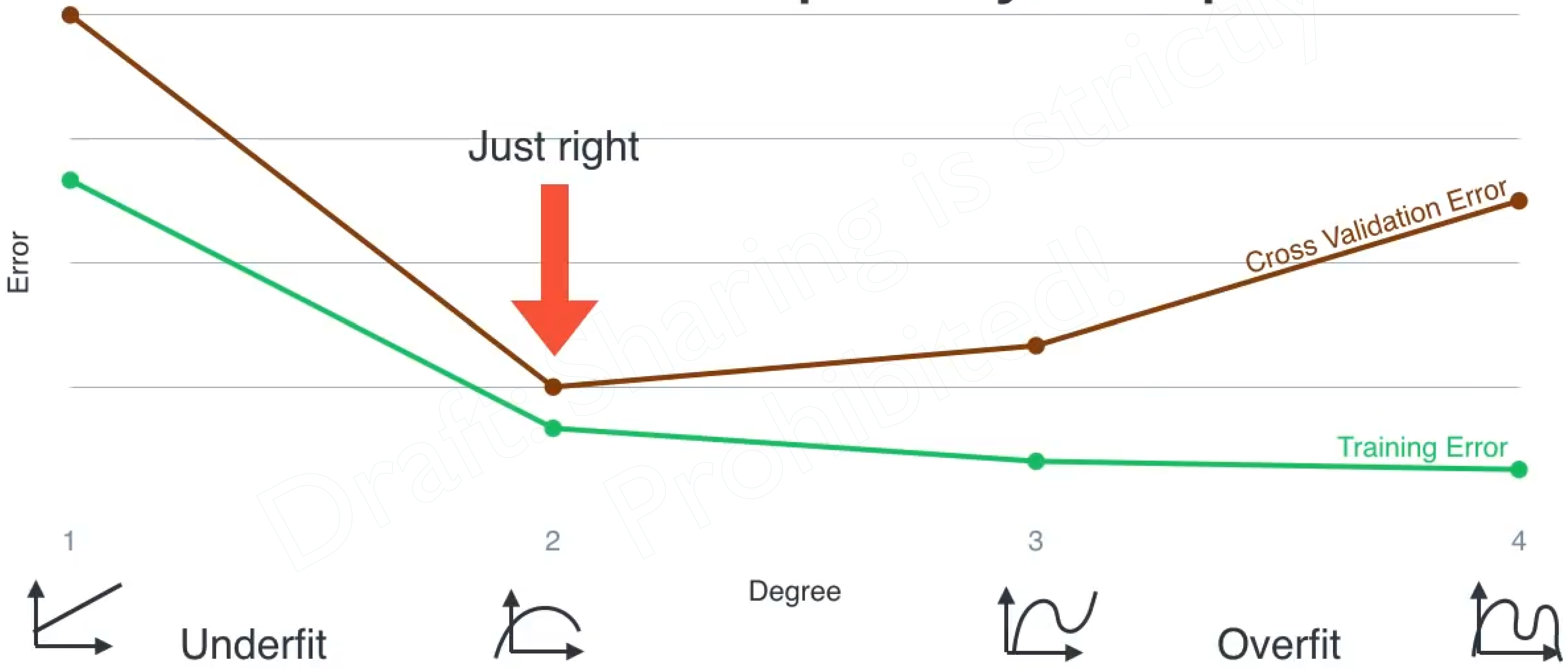


Model Complexity Graph

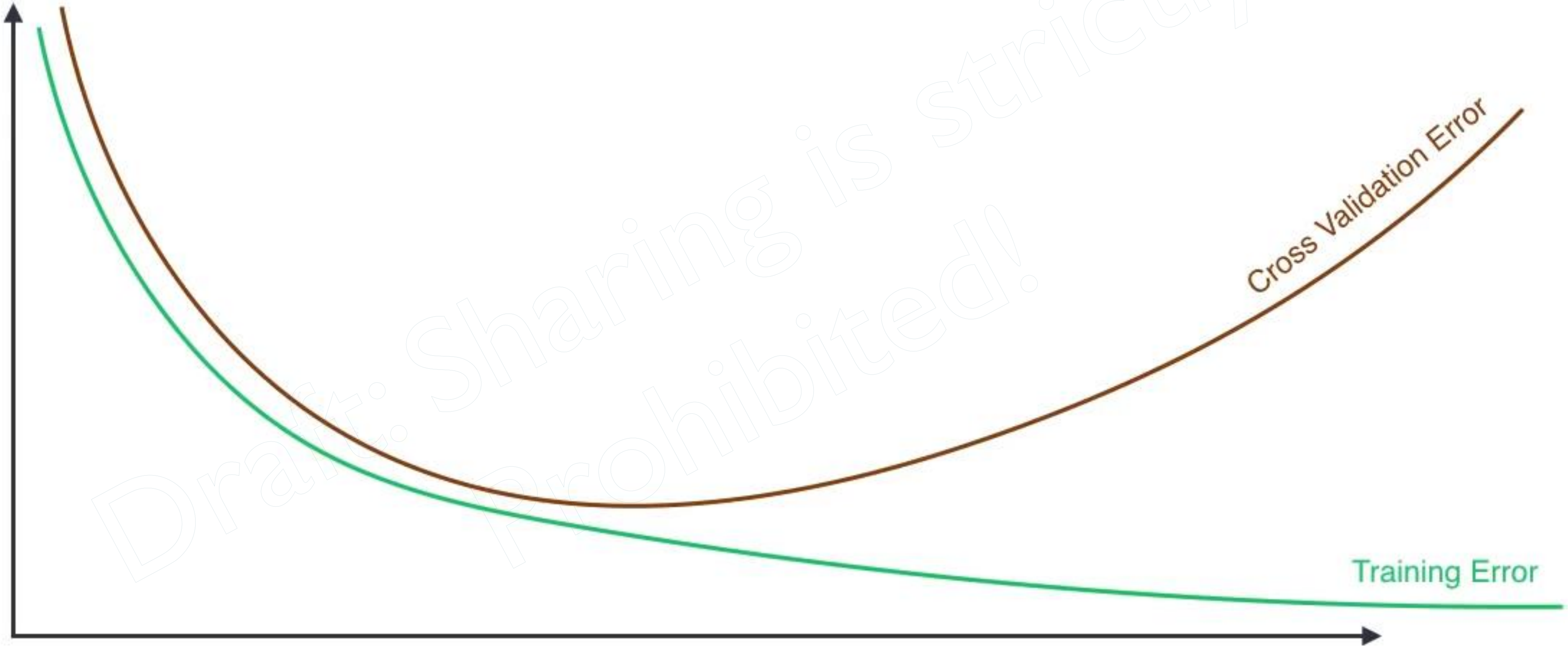
Solution: Cross Validation



Model Complexity Graph

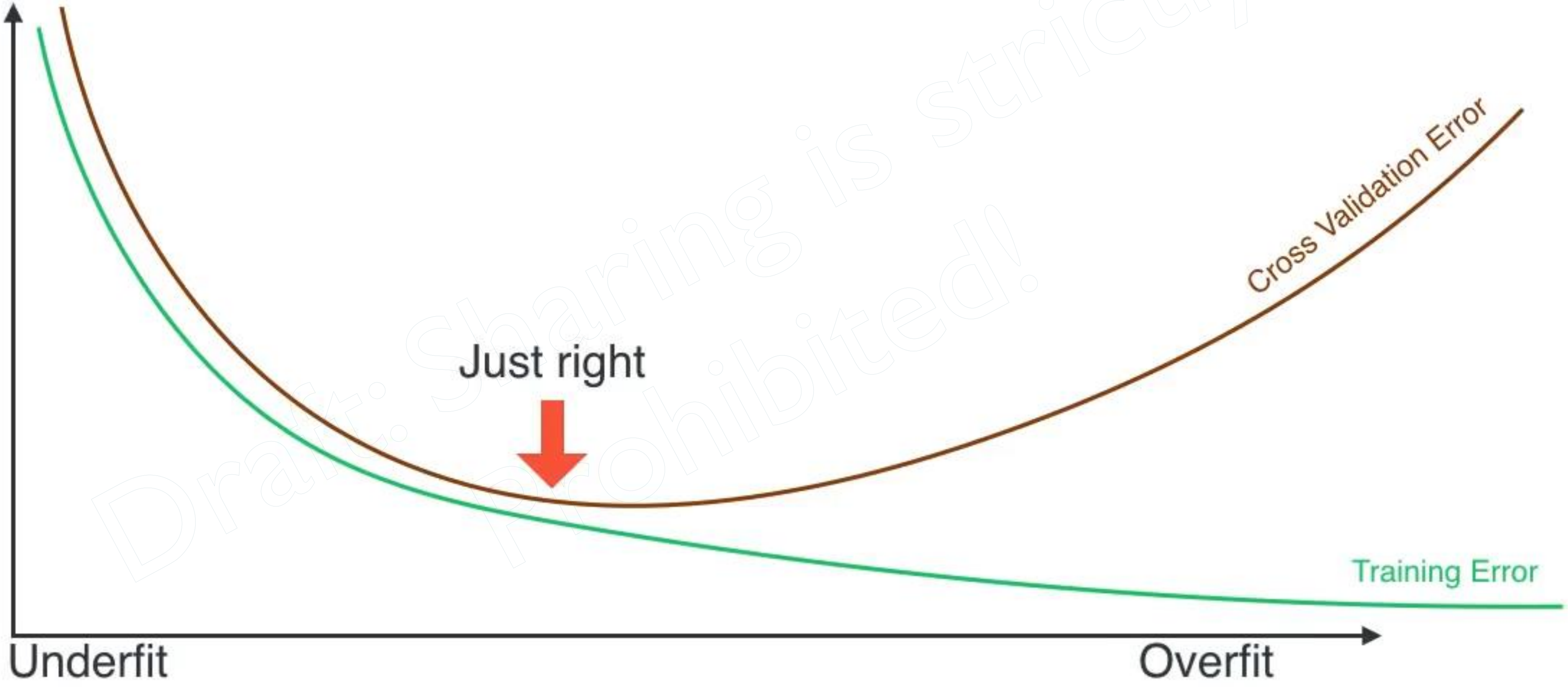


Model Complexity Graph



Model Complexity Graph

Model Complexity Graph



Summary

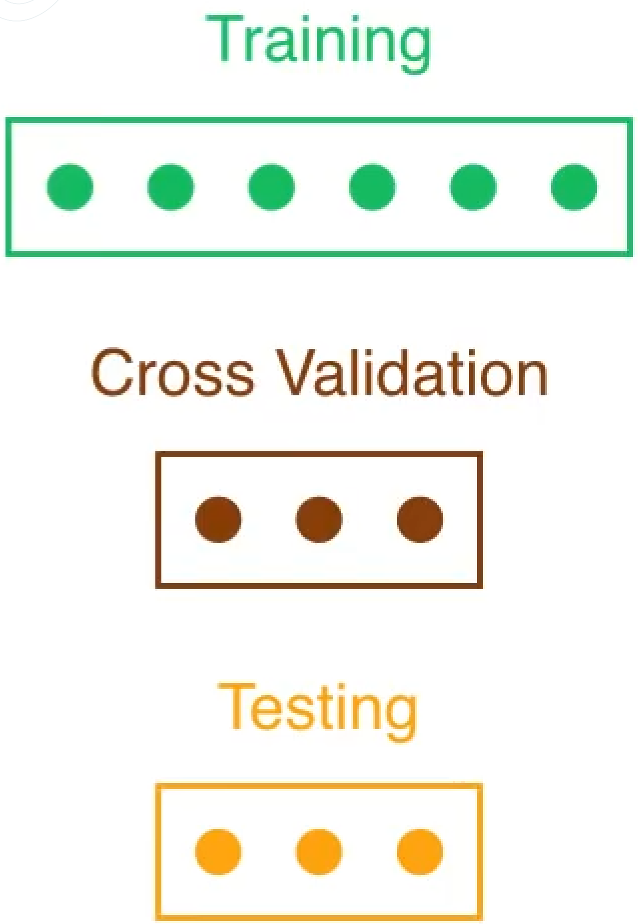
Training data: Train a bunch of models

Summary

Training data: Train a bunch of models

Cross validation data: Pick the best one of the models

Training a Logistic Regression Model



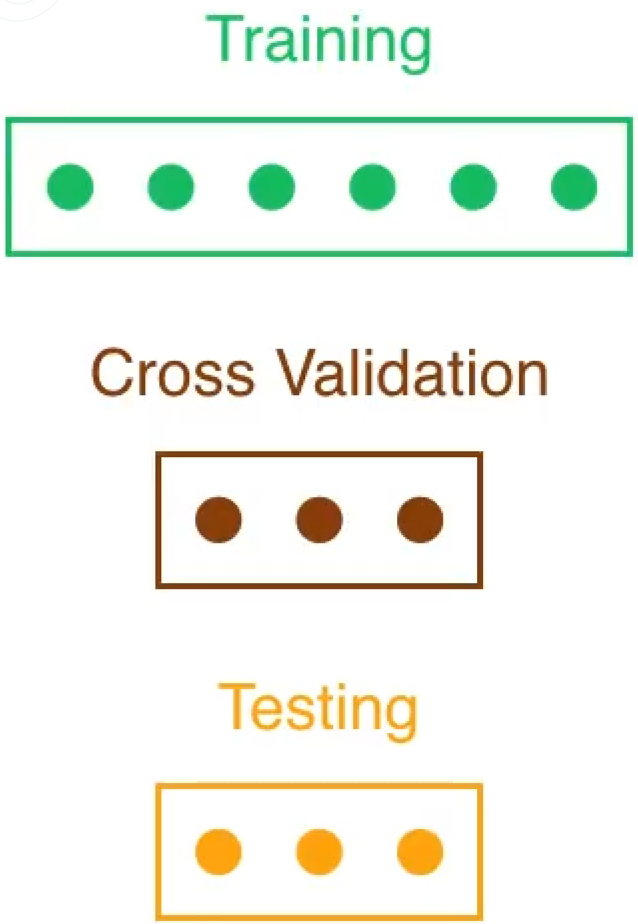
Training a Logistic Regression Model

Degree = 1

Degree = 2

Degree = 3

Degree = 4



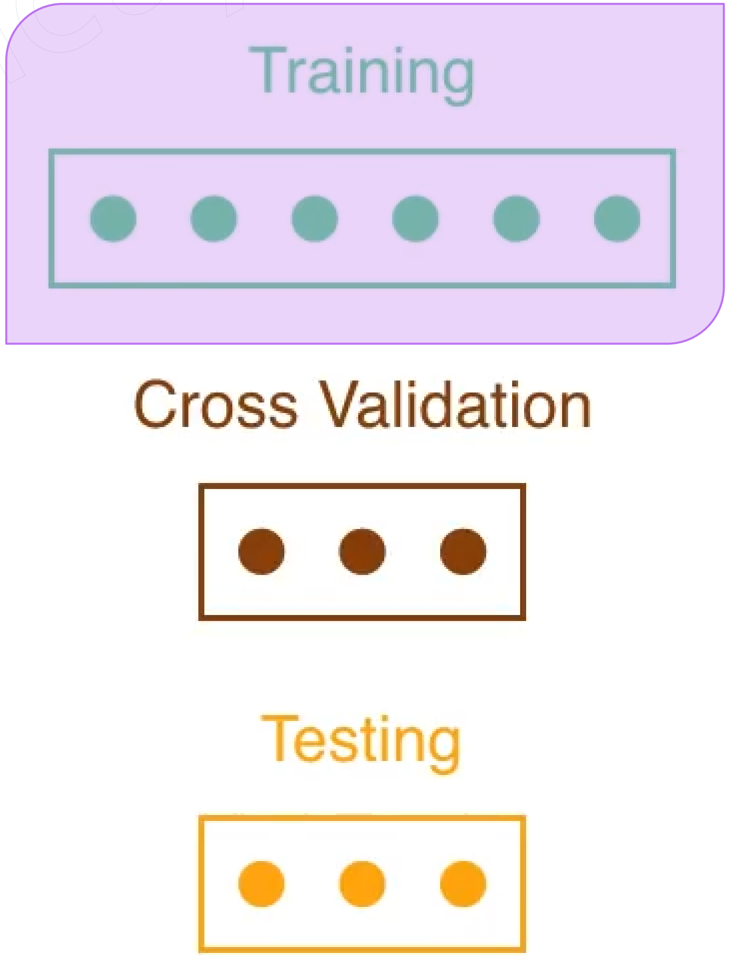
Training a Logistic Regression Model

Degree = 1

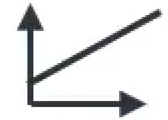
Degree = 2

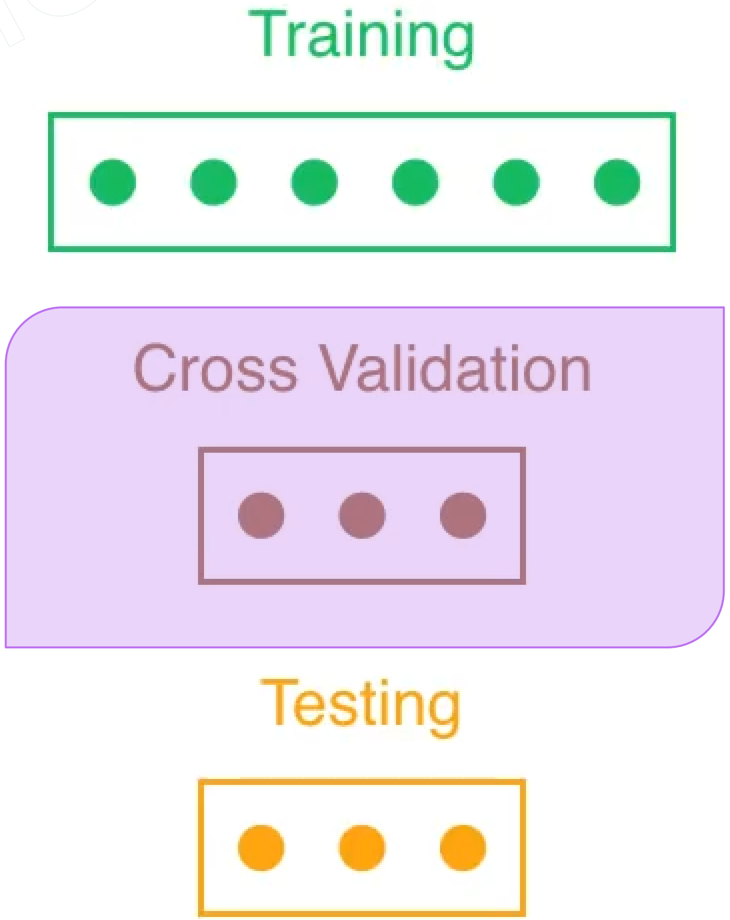
Degree = 3

Degree = 4

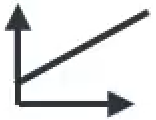


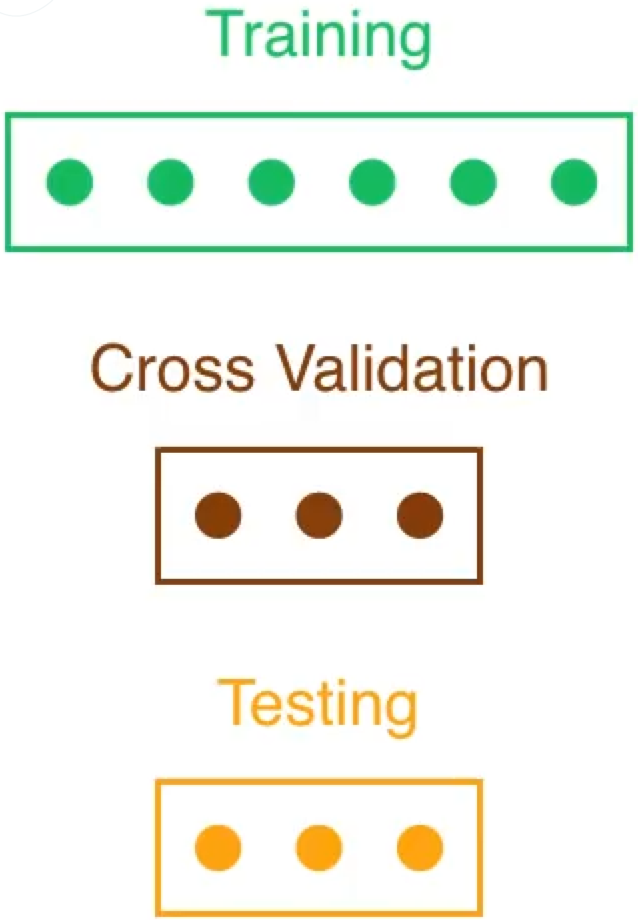
Training a Logistic Regression Model

		F1 Score
Degree = 1		0.5
Degree = 2		0.8
Degree = 3		0.4
Degree = 4		0.2



Training a Logistic Regression Model

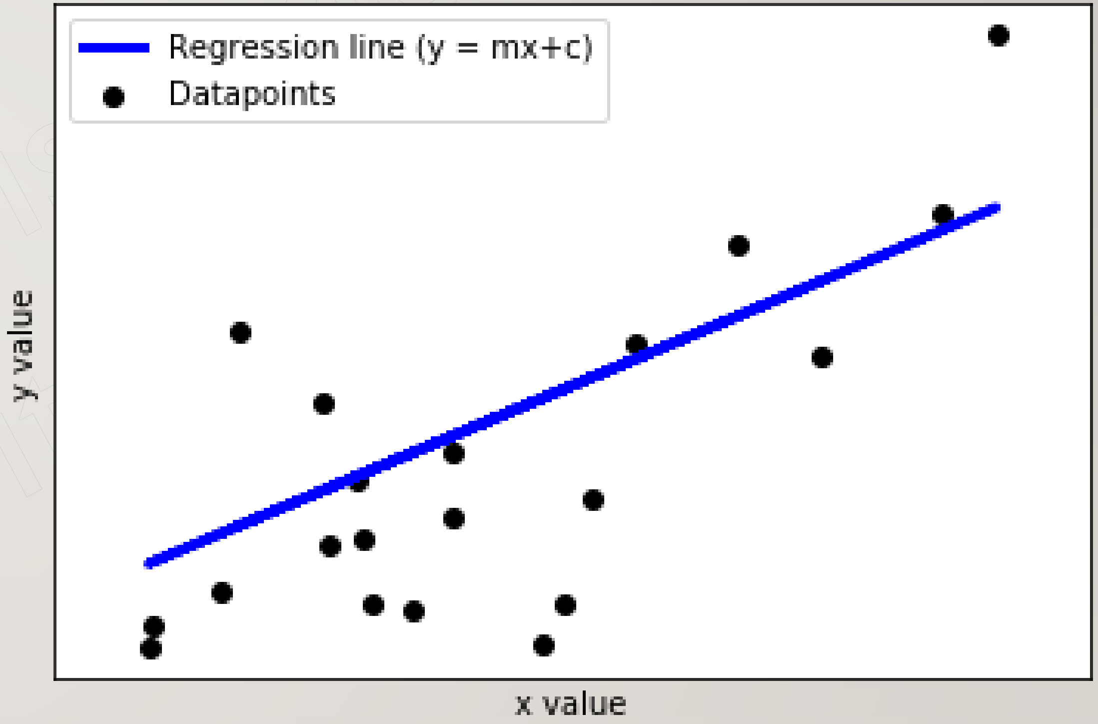
Hyperparameters	Parameters	F1 Score
Degree = 1		0.5
Degree = 2		0.8
Degree = 3		0.4
Degree = 4		0.2



Parameter vs Hyperparameter

What is Model Parameter?
A model parameter is a variable whose value is estimated from the dataset. Parameters are the values learned during training from the historical data sets.

What is Hyperparameter?
A hyperparameter is a configuration variable that is external to the model. It is defined manually before the training of the model with the historical dataset. Its value cannot be evaluated from the datasets.



Parameter: m and c
Hyperparameter: Degree of polynomials, number of iterations, acceptable error thresholds, etc.

Training a Decision Tree

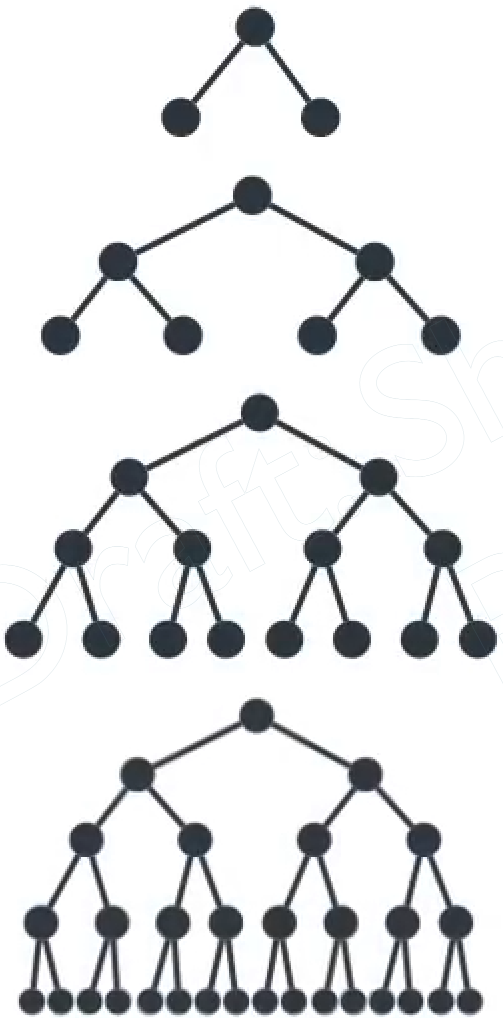
Hyperparameters Parameters

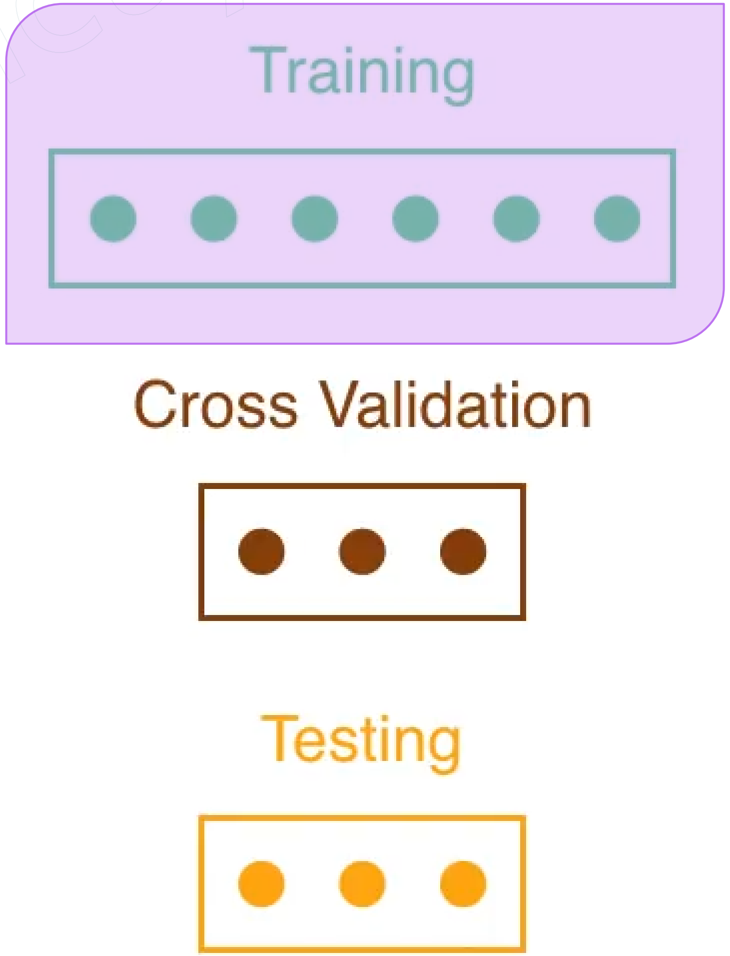
Depth = 1

Depth = 2

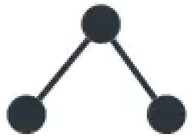
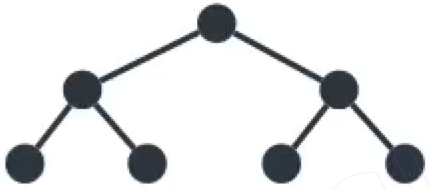
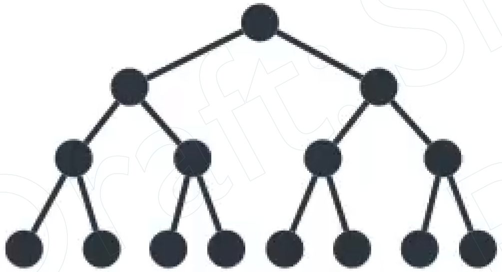
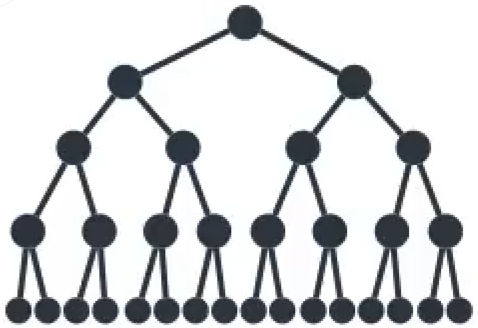
Depth = 3

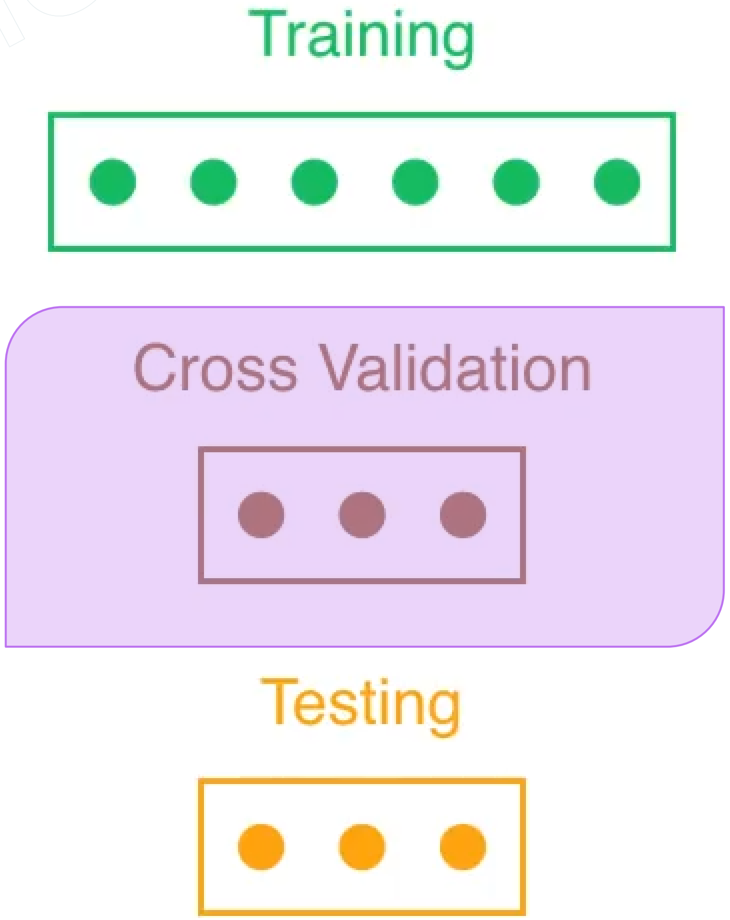
Depth = 4



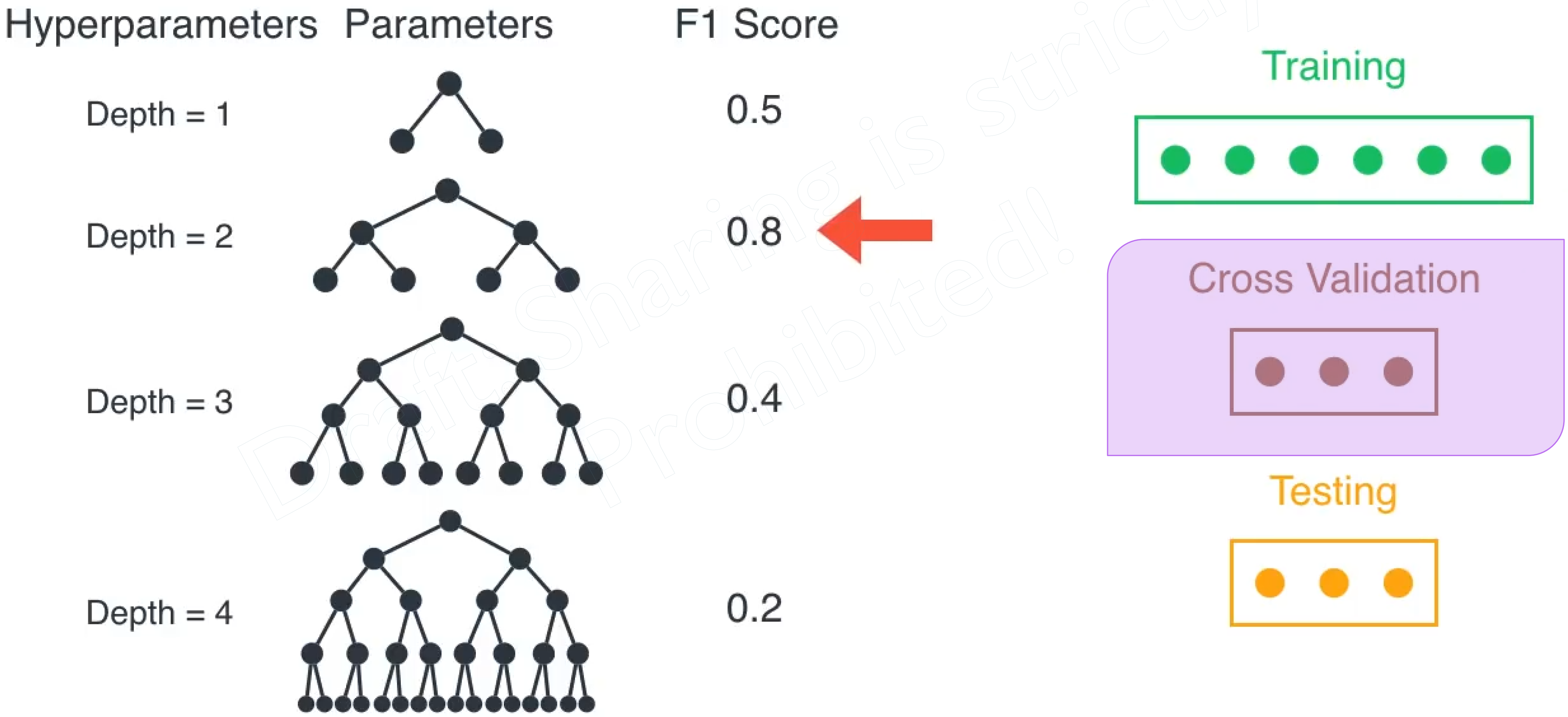


Training a Decision Tree

Hyperparameters	Parameters	F1 Score
Depth = 1		0.5
Depth = 2		0.8
Depth = 3		0.4
Depth = 4		0.2



Training a Decision Tree



How to solve a problem

How to solve a problem

Problem

How to solve a problem

Problem

Tools

How to solve a problem

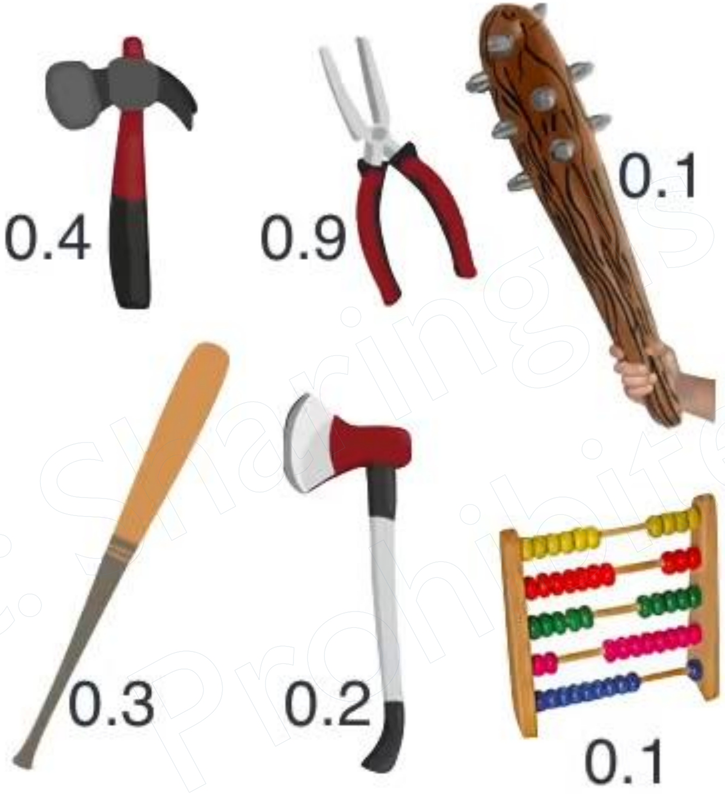
Problem

Tools

Measurement Tools

How to solve a problem

Problem



Tools

Measurement Tools

Measure each tool's performance

Pick the best tool

How to solve a problem

Problem

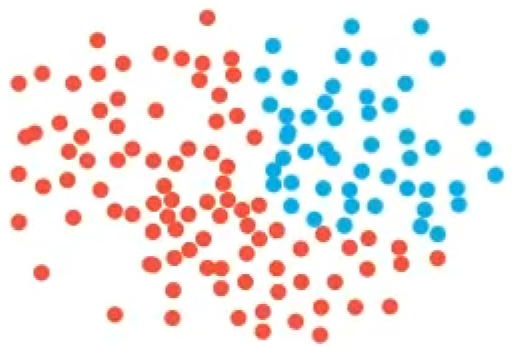
Tools

Measurement Tools

Measure each tool's performance

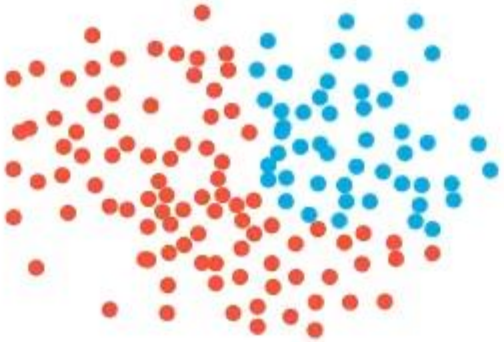
Pick the best tool

How to use machine learning

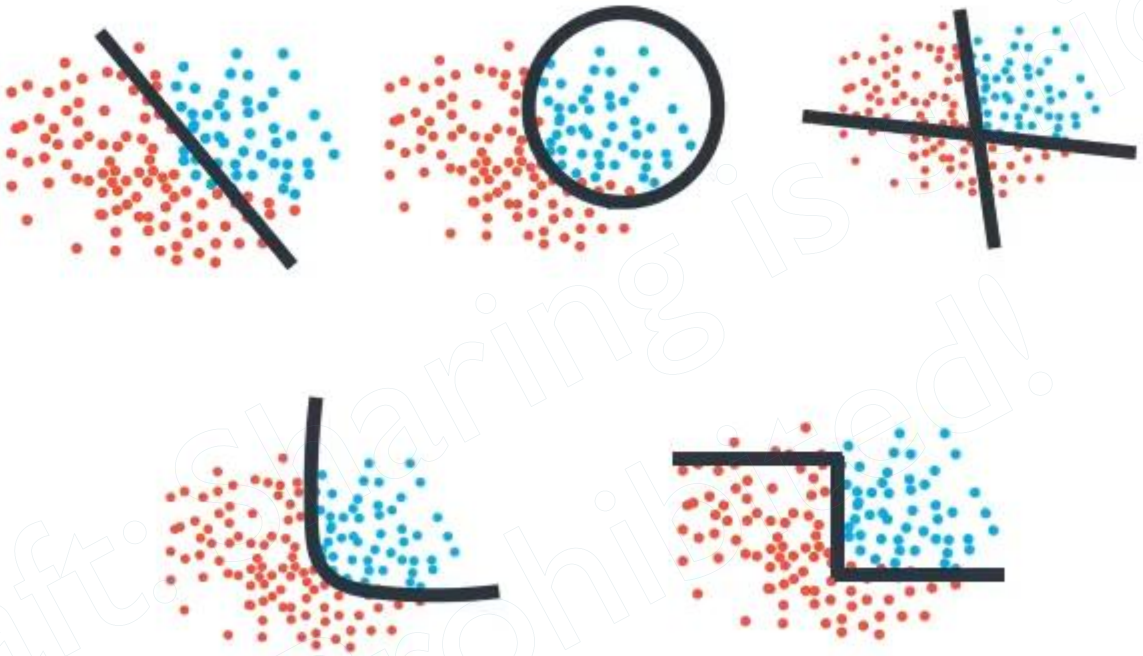


Data

How to use machine learning

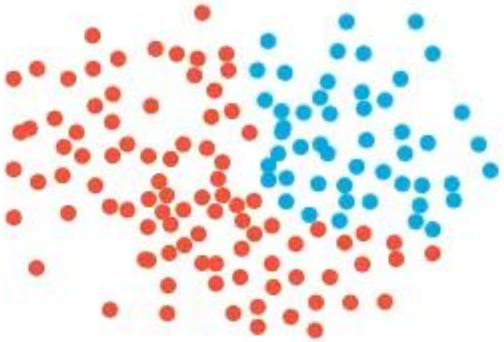


Data

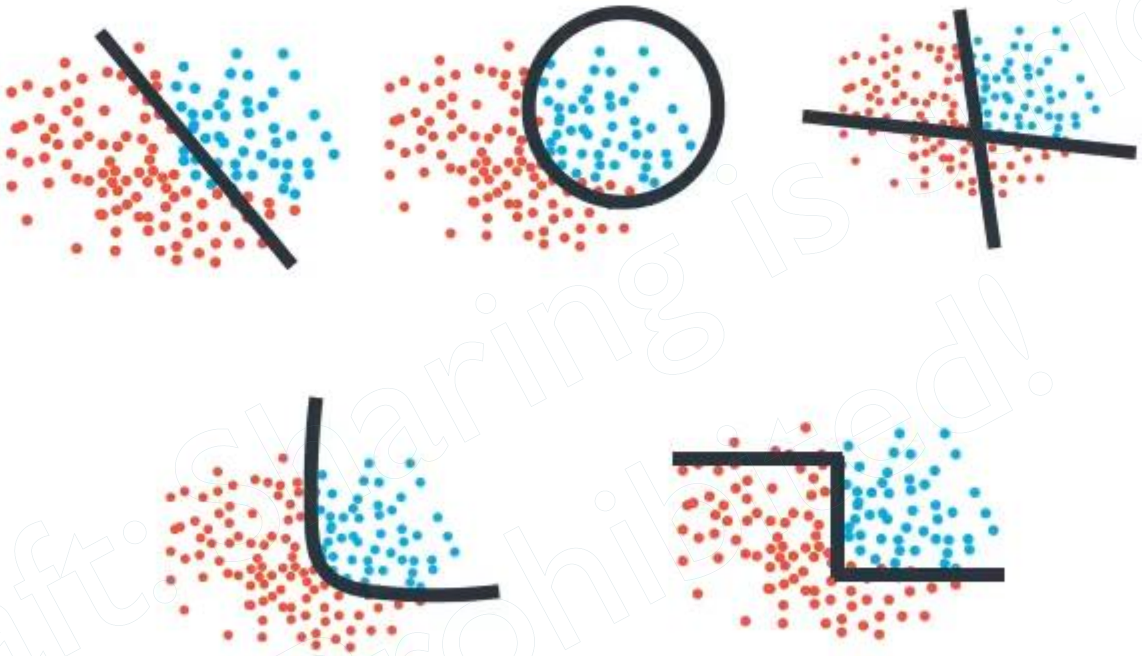


Algorithms

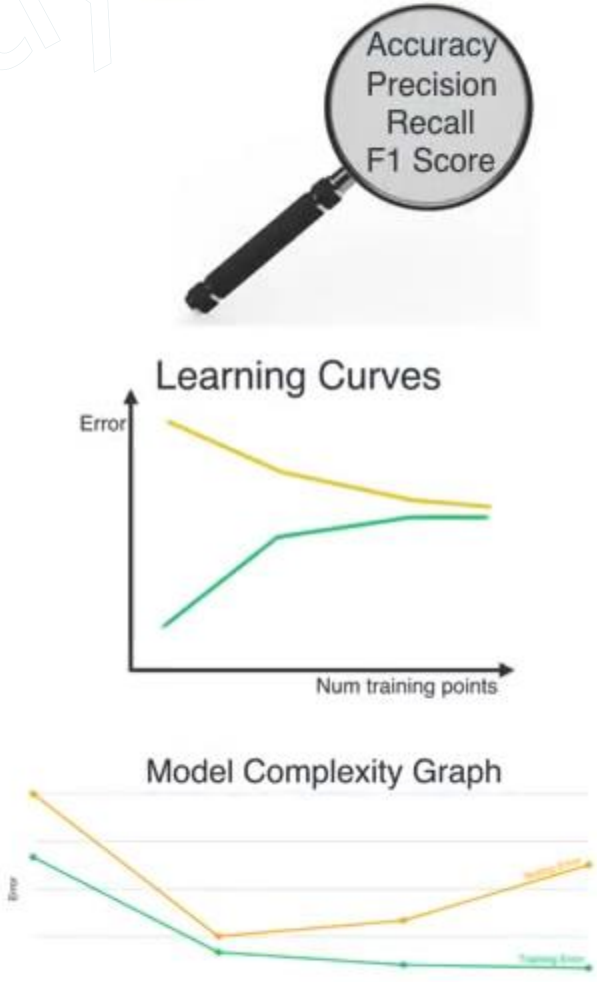
How to use machine learning



Data

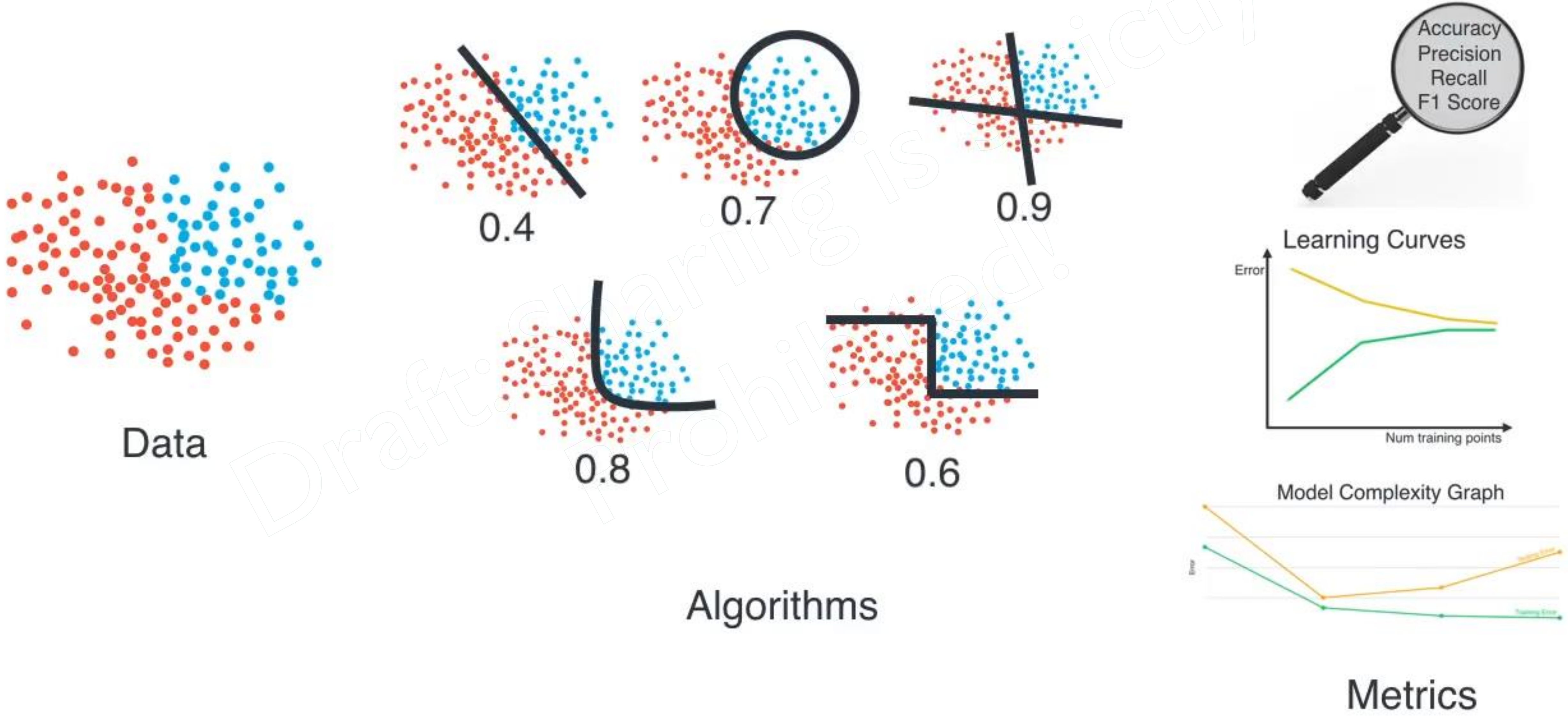


Algorithms

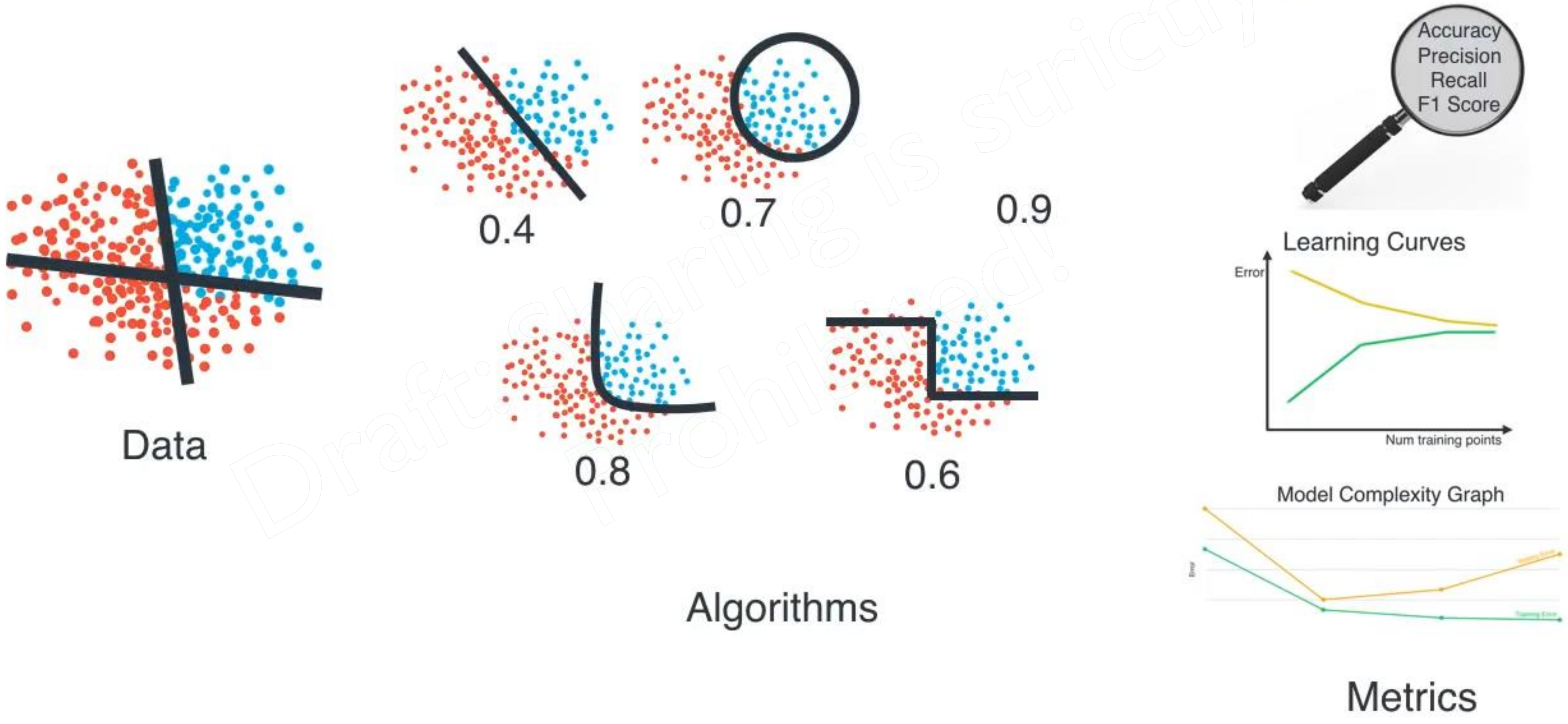


Metrics

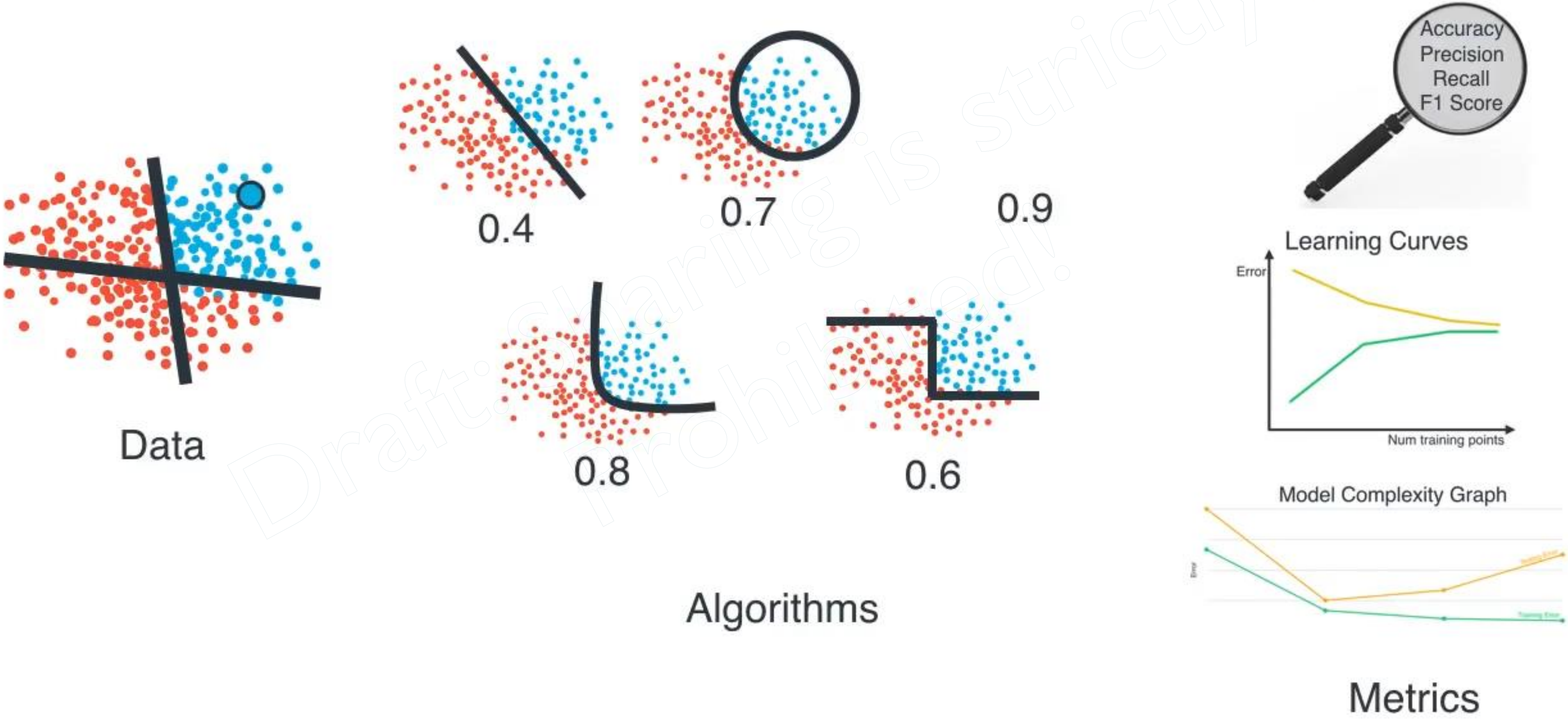
How to use machine learning



How to use machine learning



How to use machine learning



THANK YOU!