Testing and Error Metrics

Machine Learning and Pattern Recognition

(Largely based on slides from Luis Serrano)

Prof. Sandra Avila
Institute of Computing (IC/Unicamp)

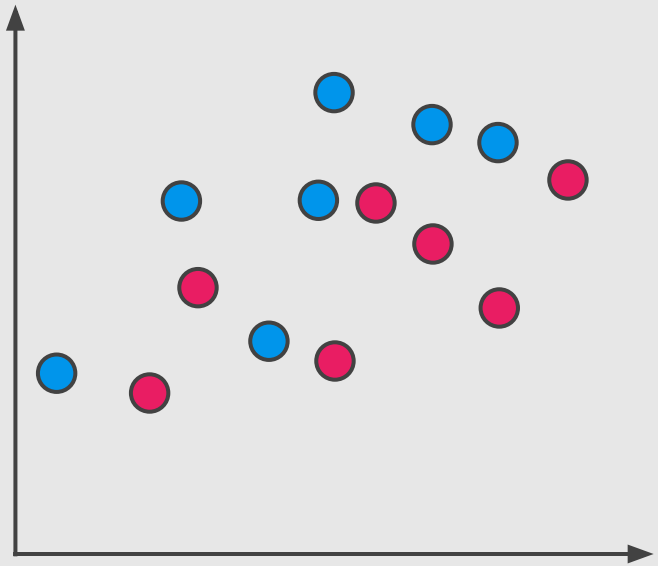
MC886/MO444, August 29, 2017

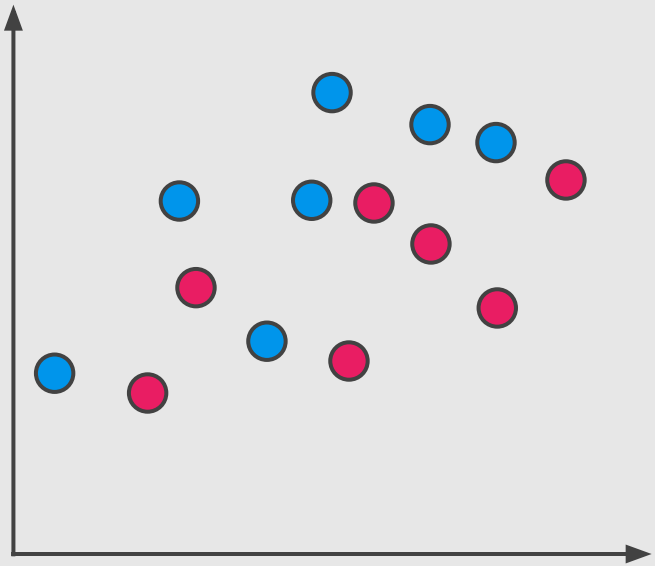
How well is my model doing?

Today's Agenda

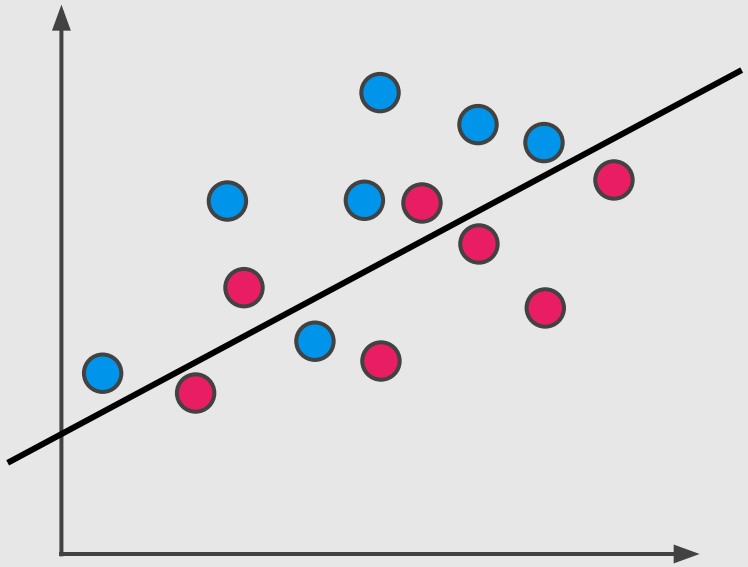
- Testing and Error Metrics
  - Training, Testing
  - Accuracy
  - Precision
  - Recall
  - F1 Score

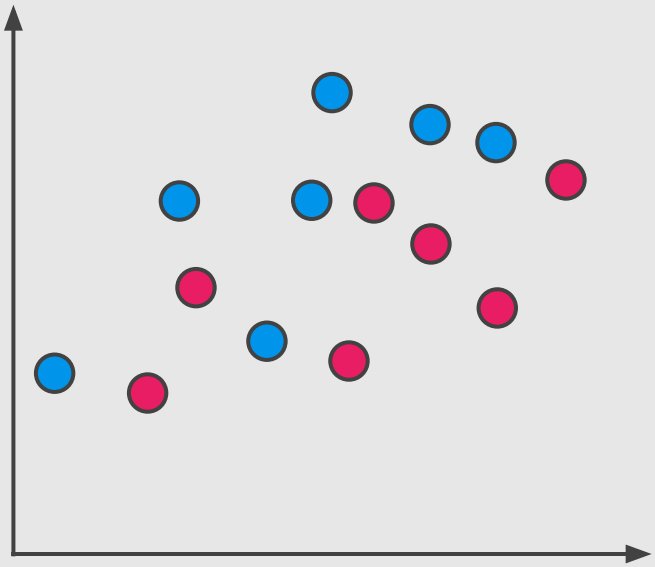
Which model is better?



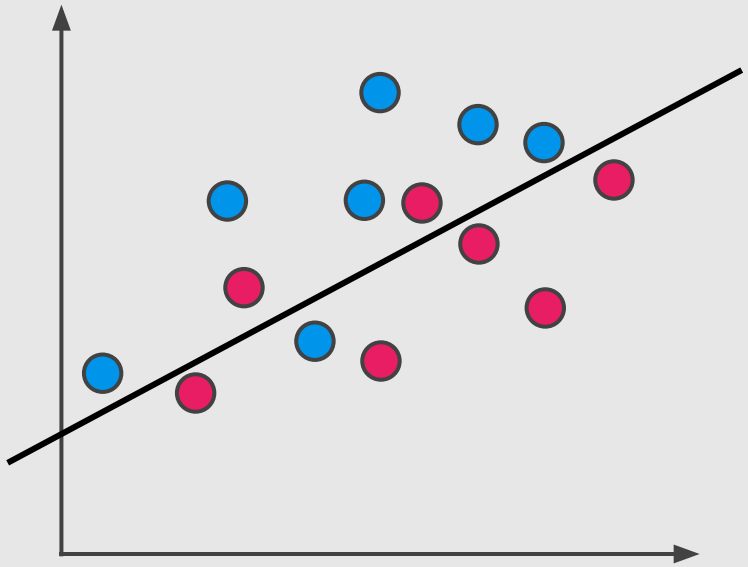


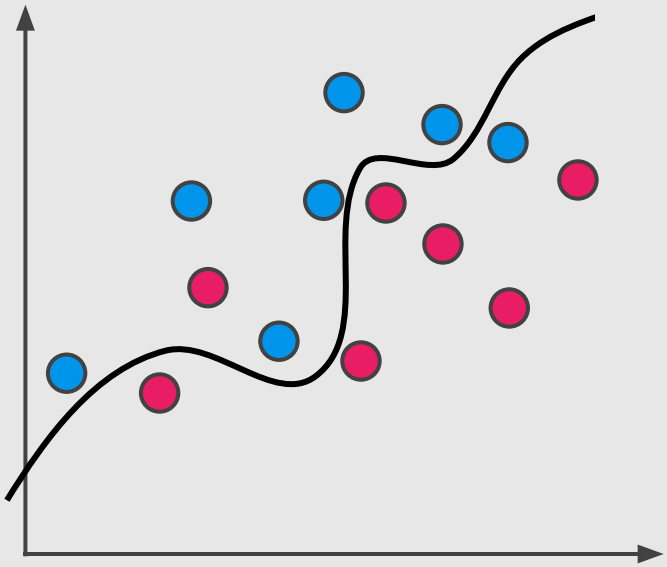
Which model is better?



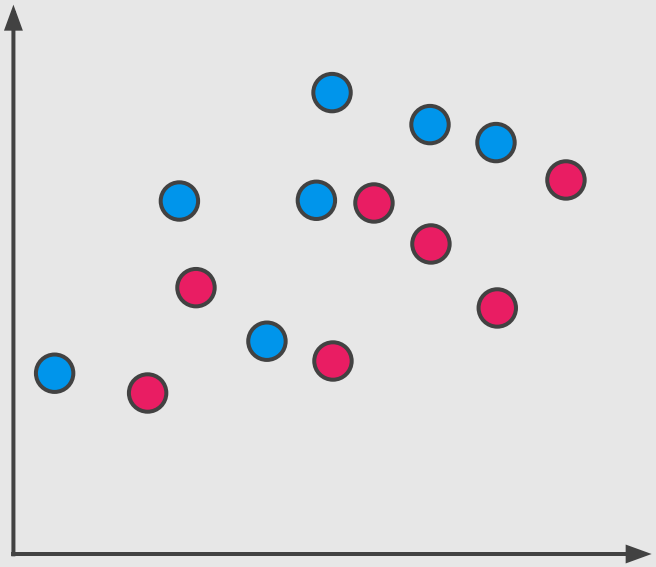


Which model is better?

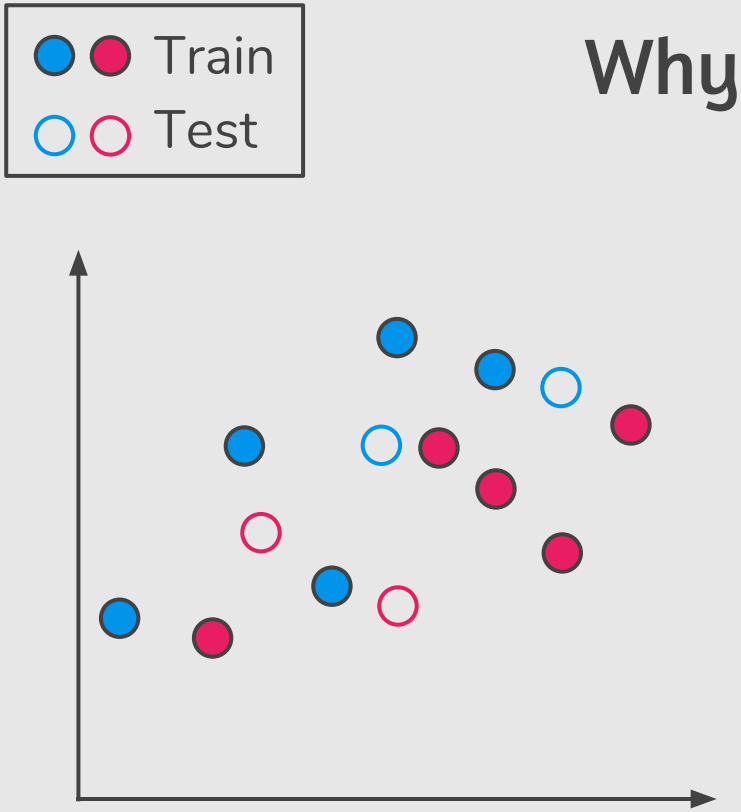




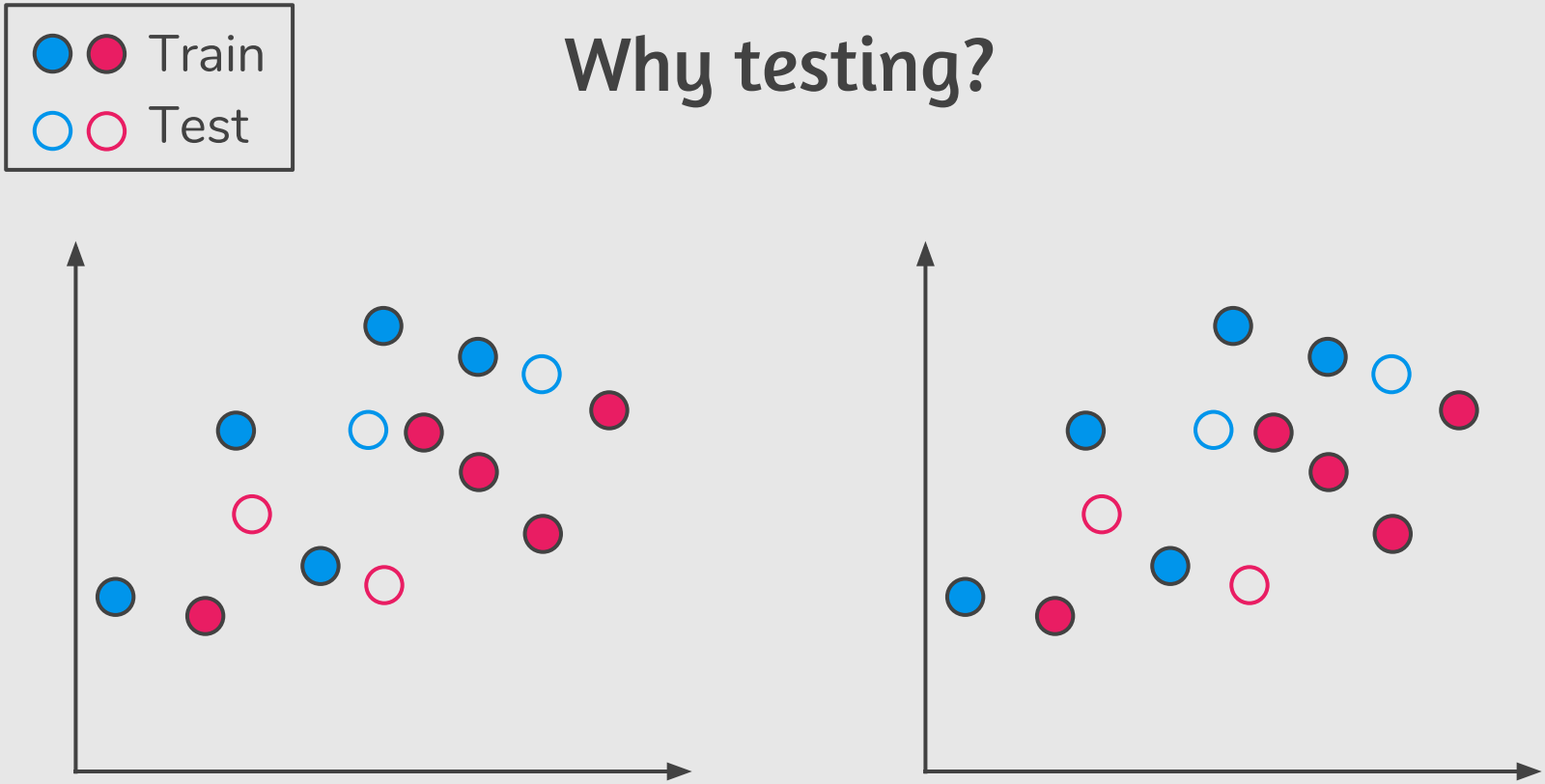
Why testing?



Why testing?

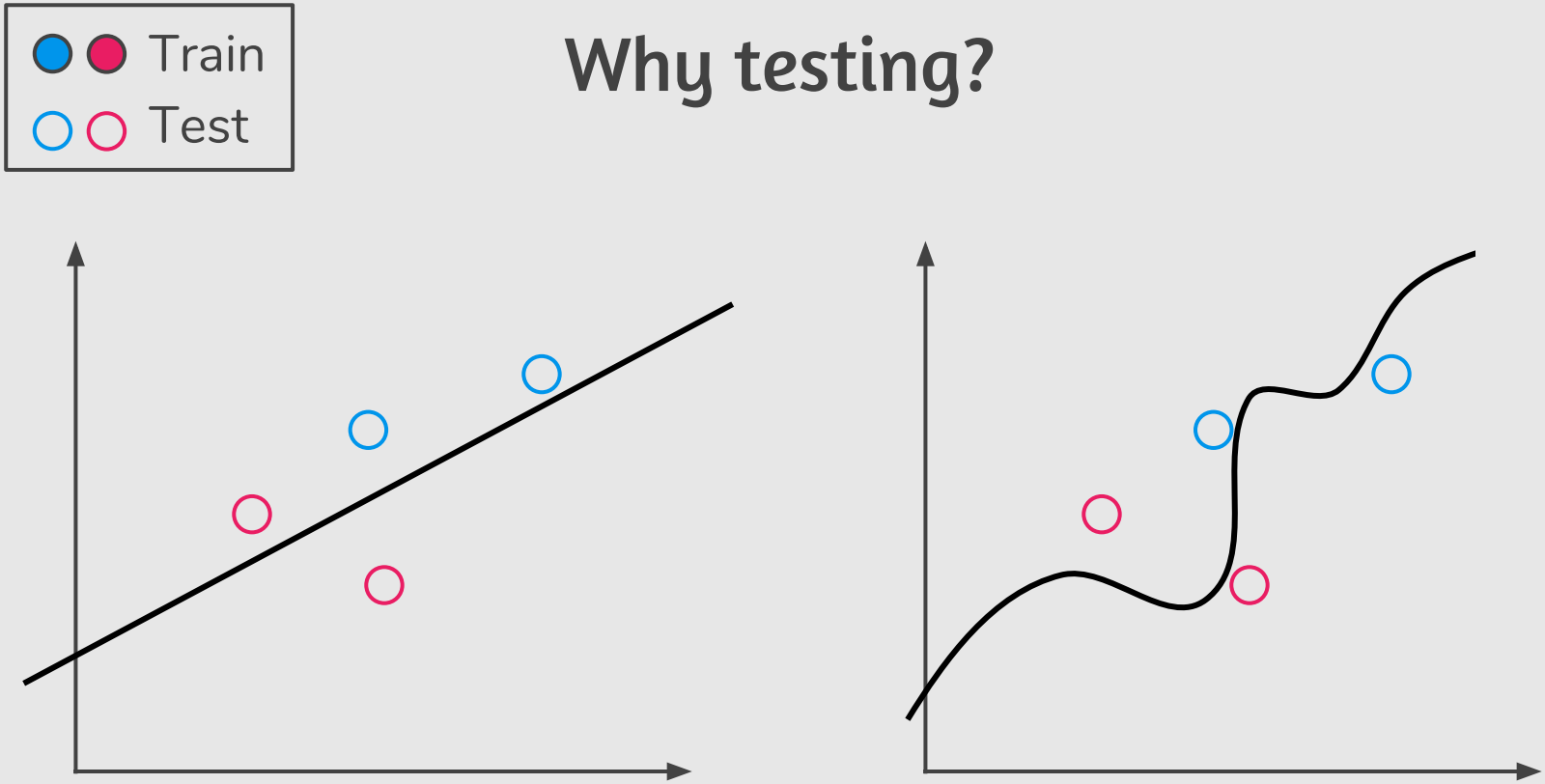


Why testing?

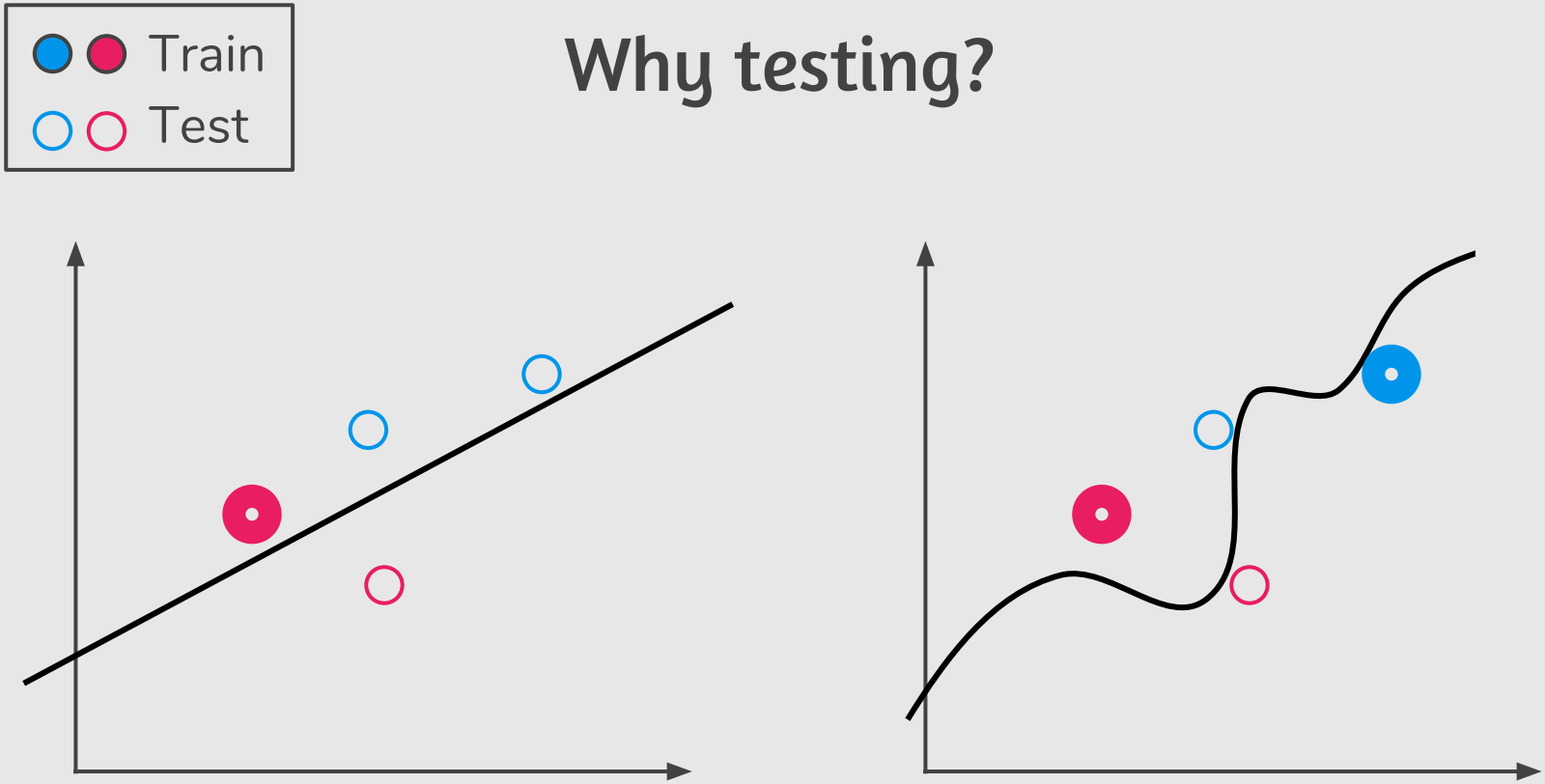


Why testing?

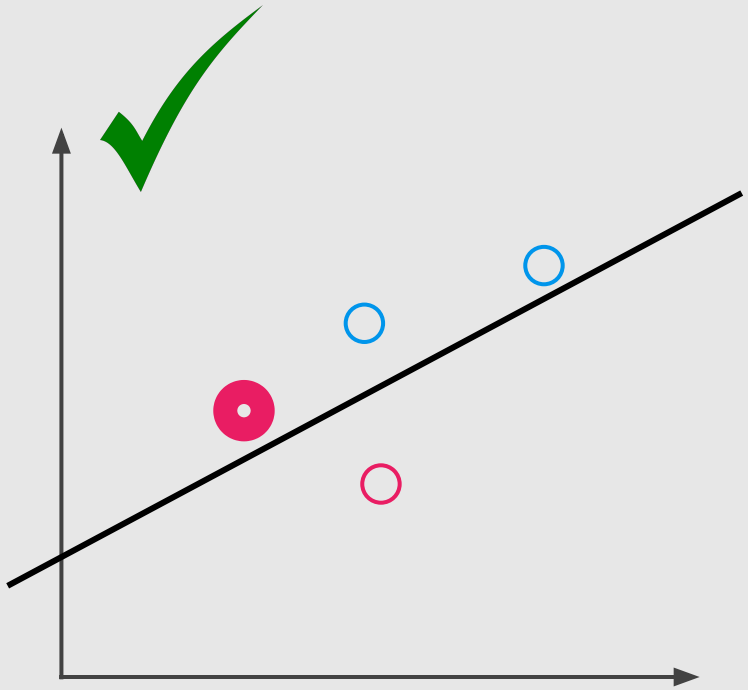
Why testing?

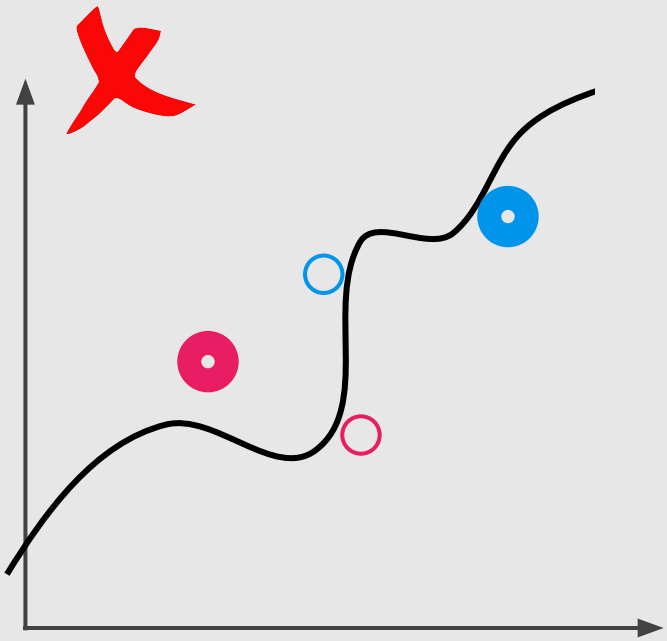


Why testing?

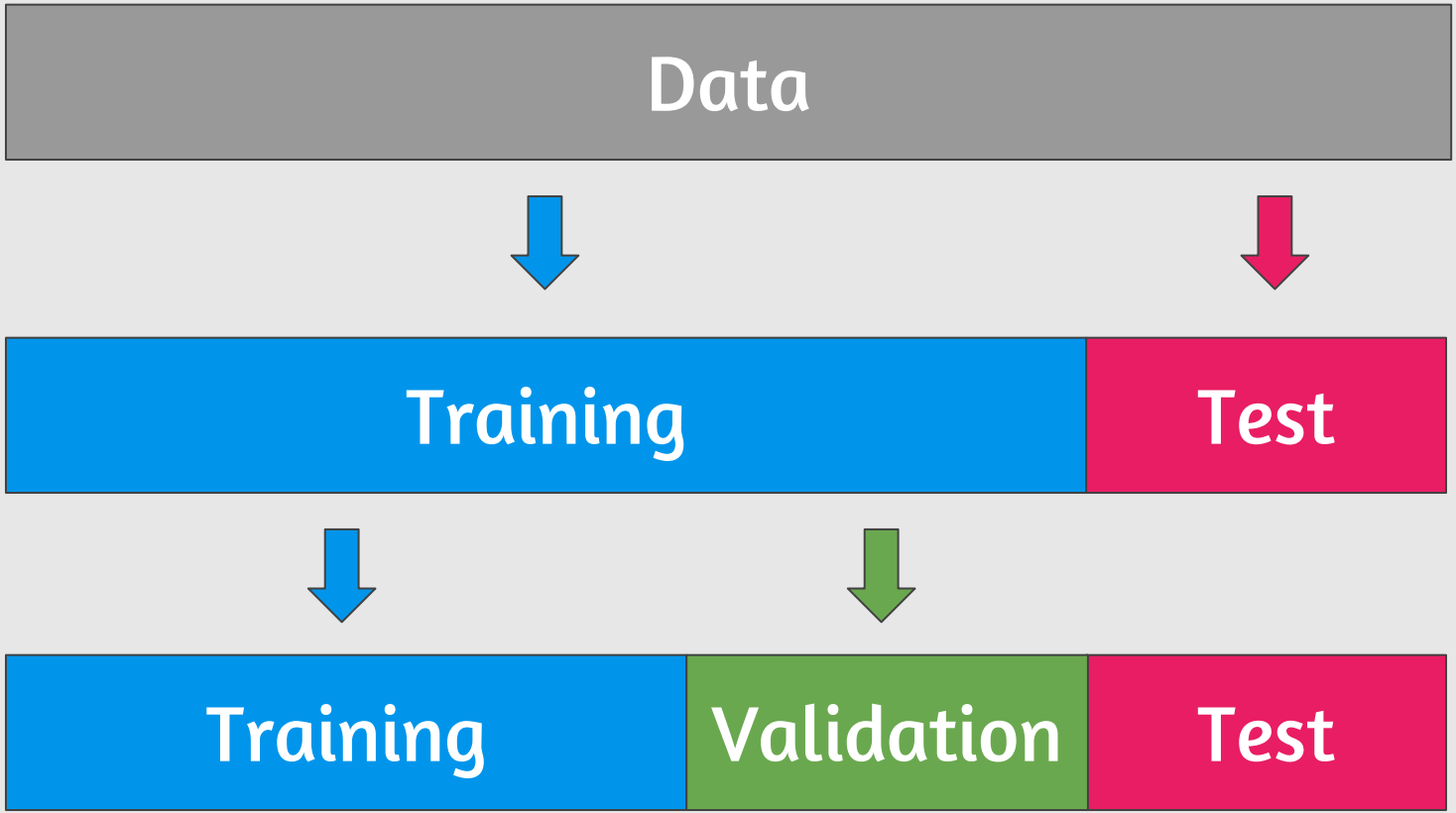


Why testing?

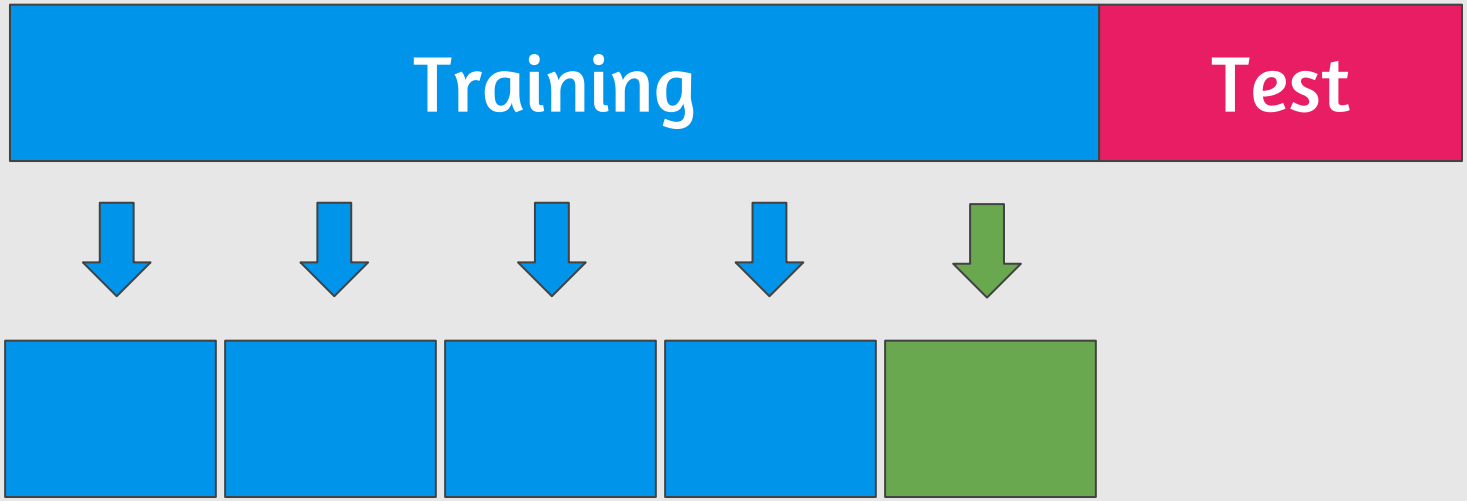




Friends don’t let friends
use testing data
for training


Data
Training
Test
Training
Validation
Test

k-fold Cross Validation

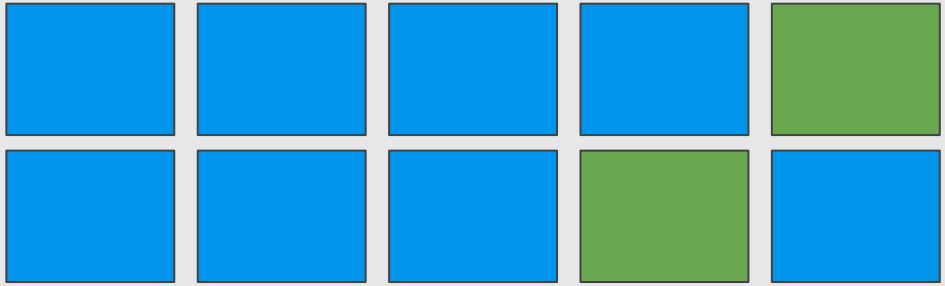


k-fold Cross Validation

$k = 5$

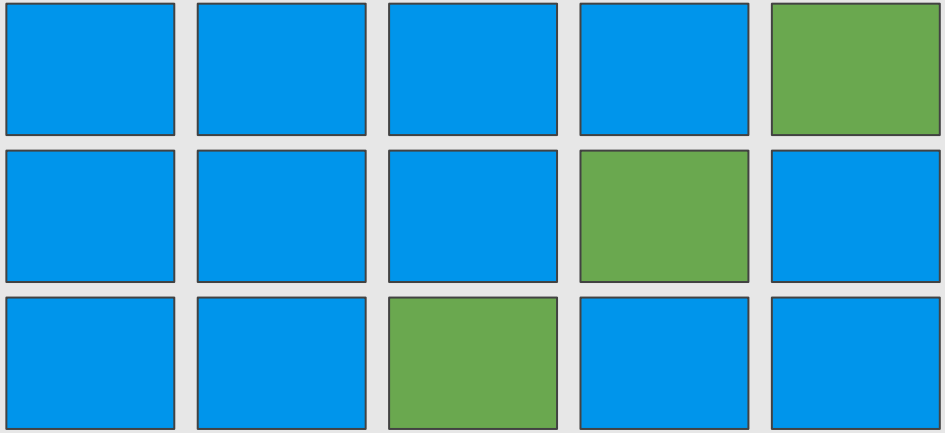
k-fold Cross Validation

$k = 5$



k-fold Cross Validation

$k = 5$

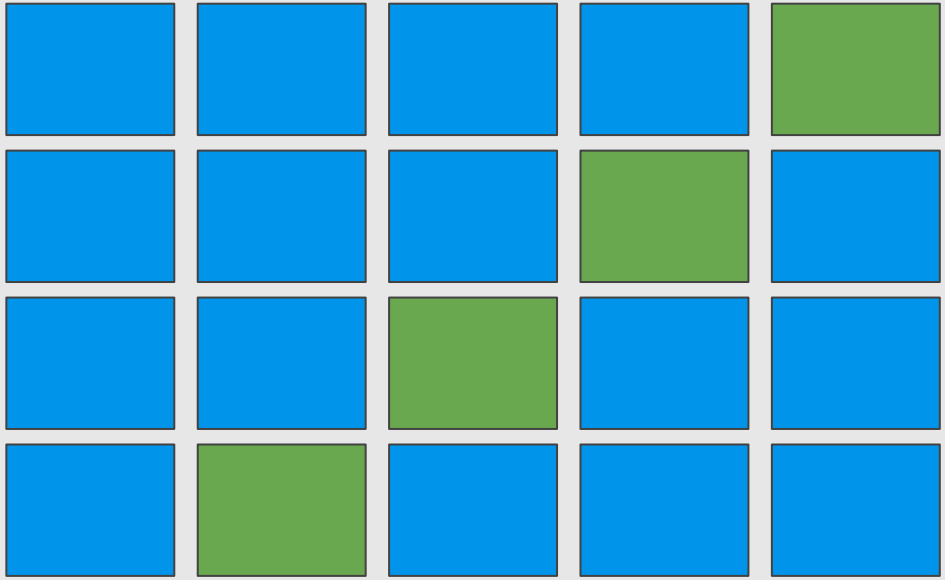


k-fold Cross Validation

Training

Validation

k = 5

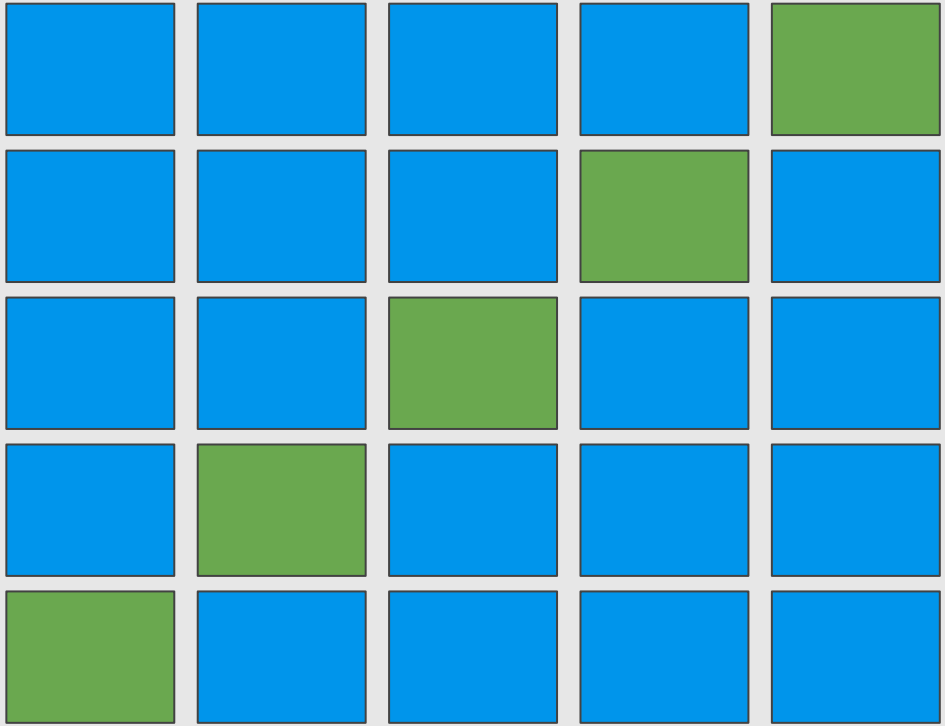


k-fold Cross Validation

Training

Validation

k = 5

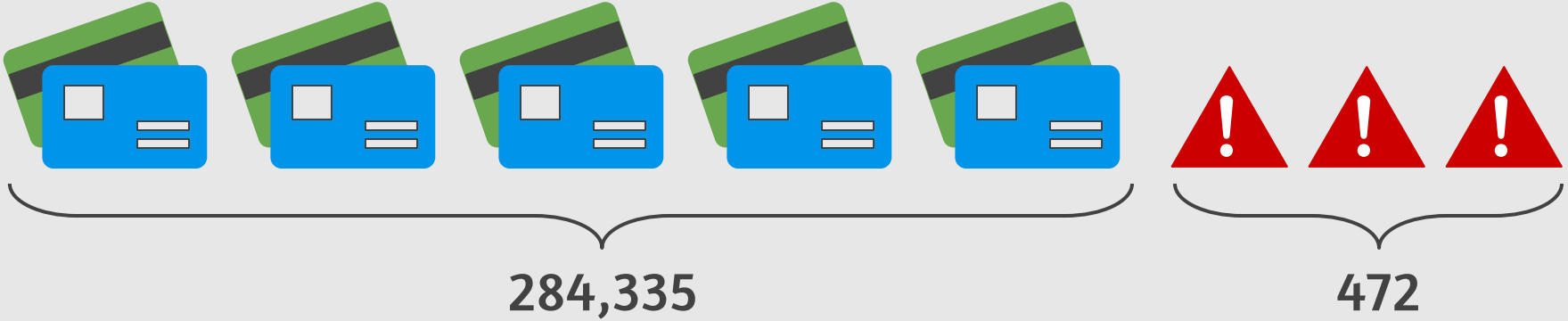


Evaluation Metrics

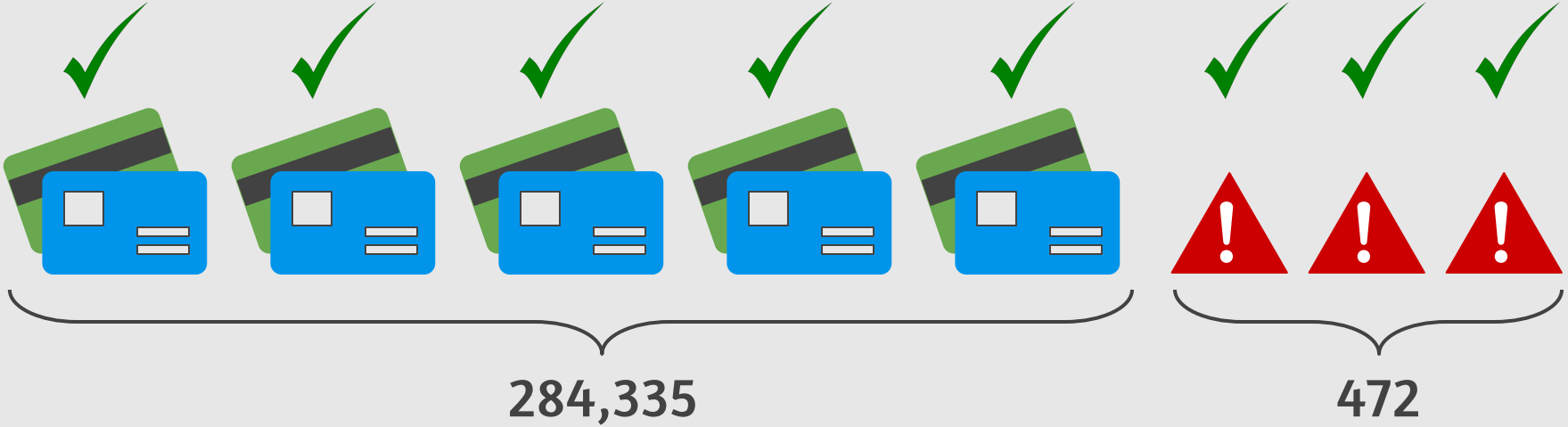
How well is my model doing?

Credit Card Fraud

Credit Card Fraud

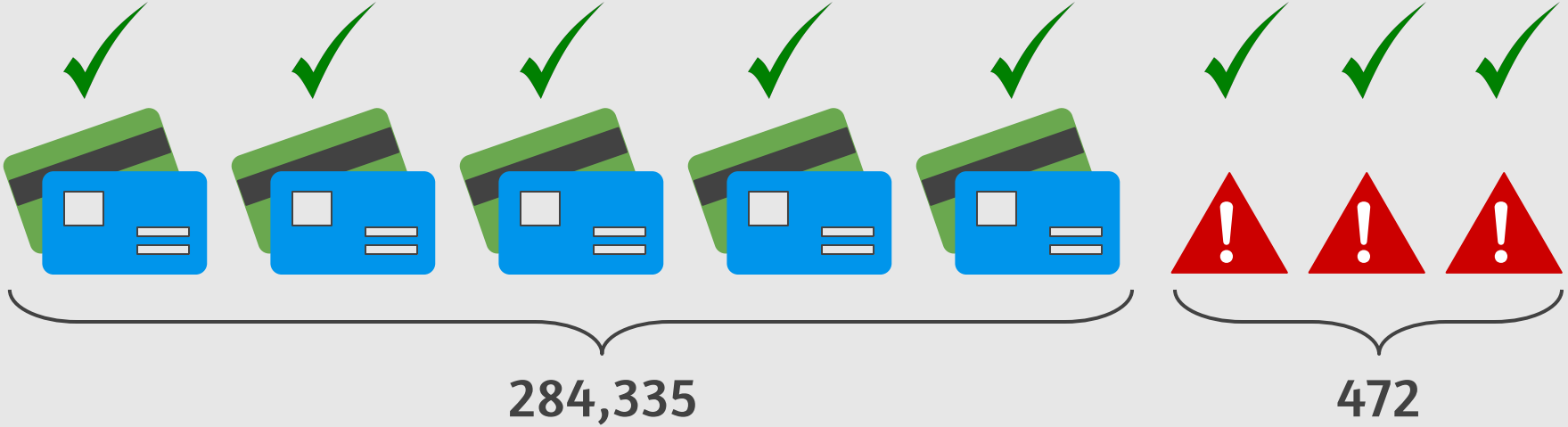


Credit Card Fraud



Model: All transactions are good.

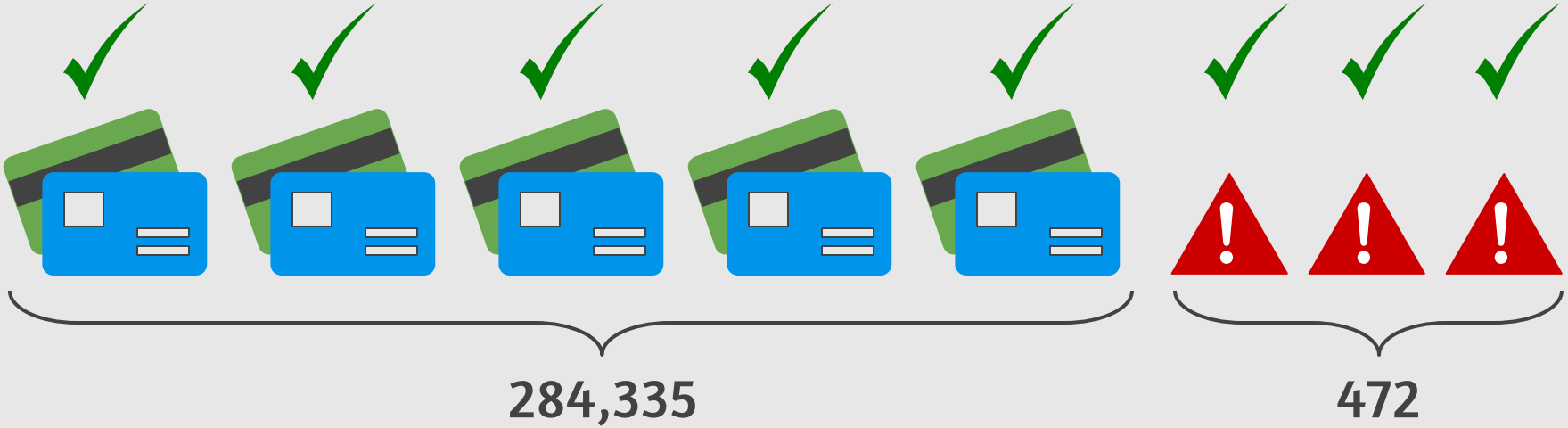
Credit Card Fraud



Model: All transactions are good.

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

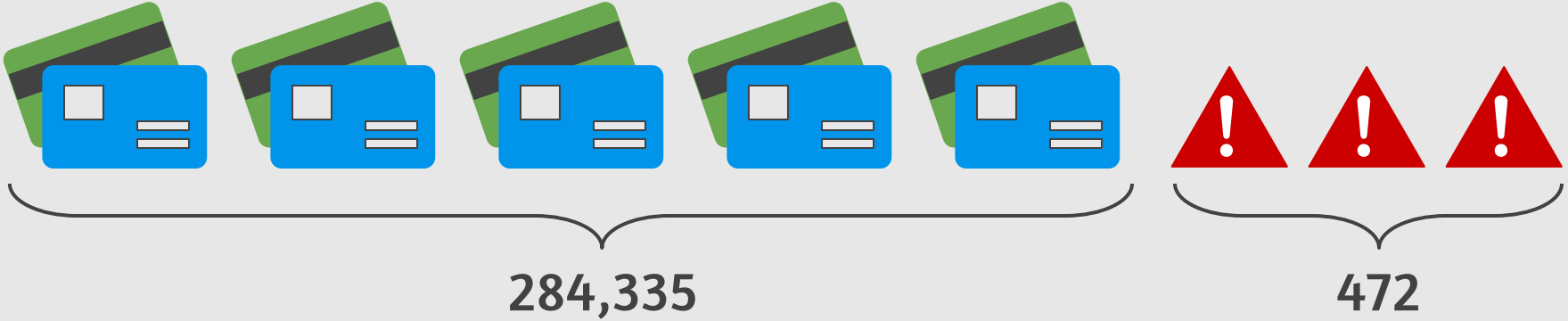
Credit Card Fraud



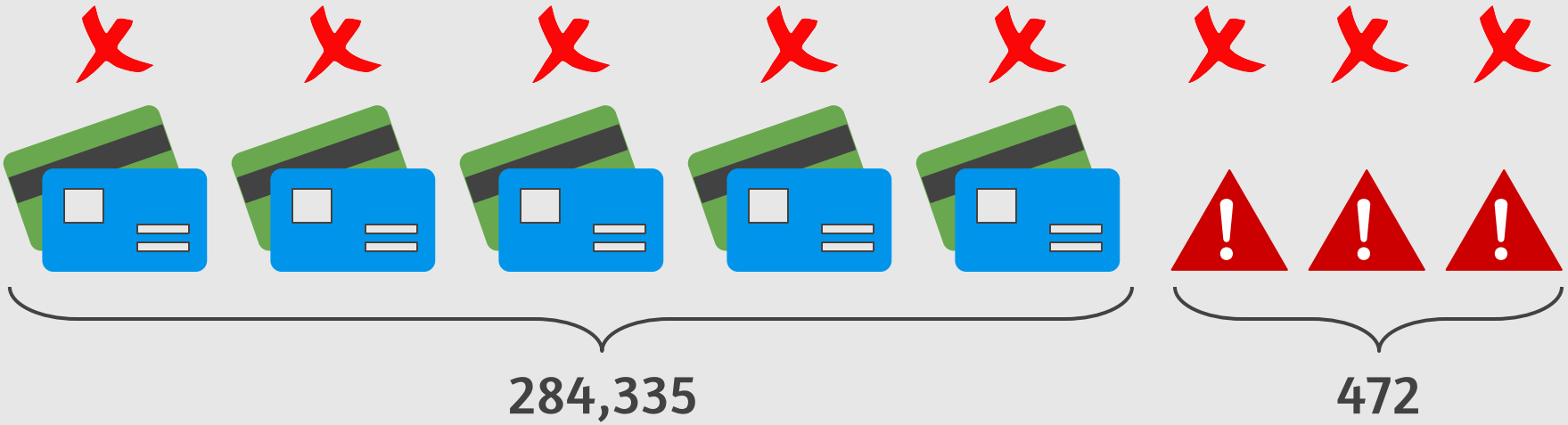
Model: All transactions are good.

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

Credit Card Fraud

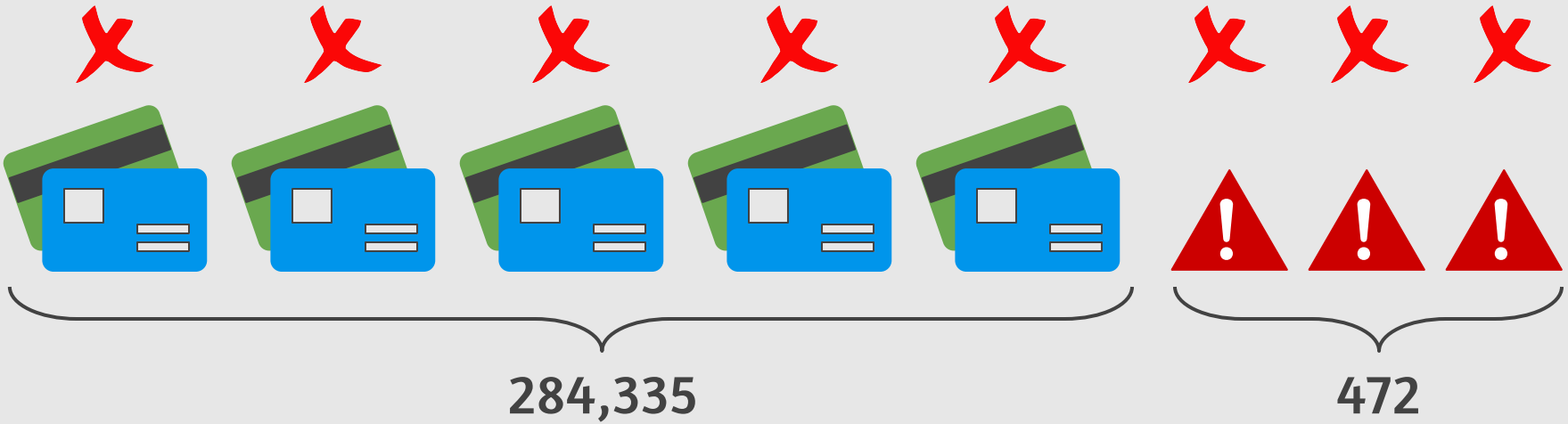


Credit Card Fraud



Model: All transactions are fraudulent.

Credit Card Fraud



Model: All transactions are fraudulent.
Problem: I'm accidently catching all the good ones!

Medical Model

Health

Sick

Spam Classifier Model

Not Spam

Spam

Confusion Matrix Table

	Diagnosed Sick	Diagnosed Healthy
Sick		
Healthy		

Confusion Matrix Table

	Diagnosed Sick	Diagnosed Healthy
Sick	True Positive 	
Healthy		

Confusion Matrix Table

	Diagnosed Sick	Diagnosed Healthy
Sick	True Positive 	
Healthy		True Negative 

Confusion Matrix Table

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

Confusion Matrix Table

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

Confusion Matrix Table

10,000 patients

Diagnosis

		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1000	200
	Healthy	800	8000

Confusion Matrix Table

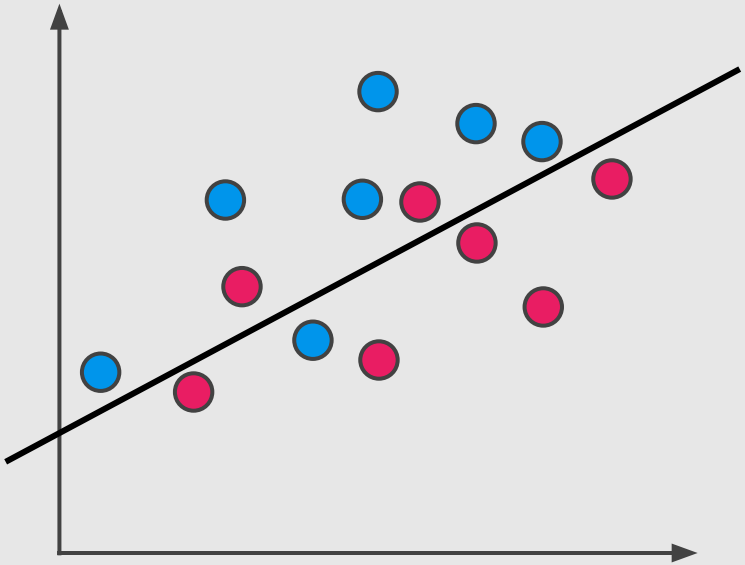
	Sent to Spam Folder	Sent to Inbox
Spam	True Positive 	False Negative 
Not Spam	False Positive 	True Negative 

Confusion Matrix Table

1,000 emails

Email \ Folder	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

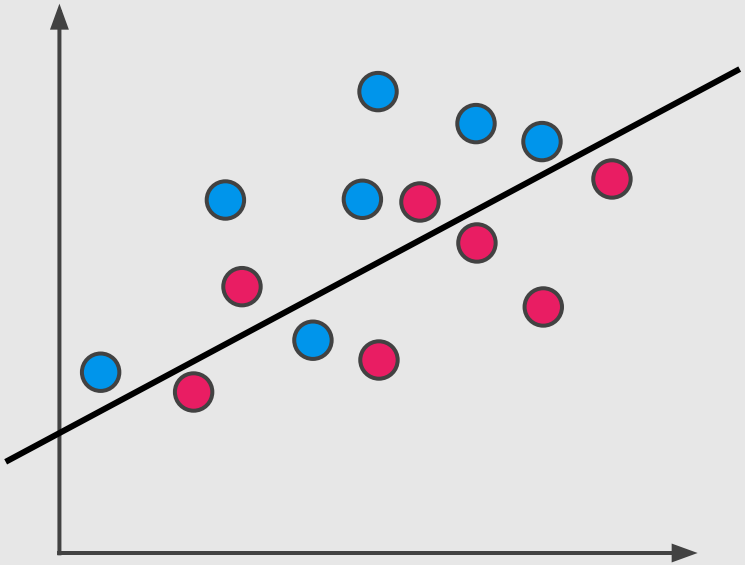
Confusion Matrix Table



Prediction

		Guessed Positive	Guessed Negative
Data	Positive		
	Negative		

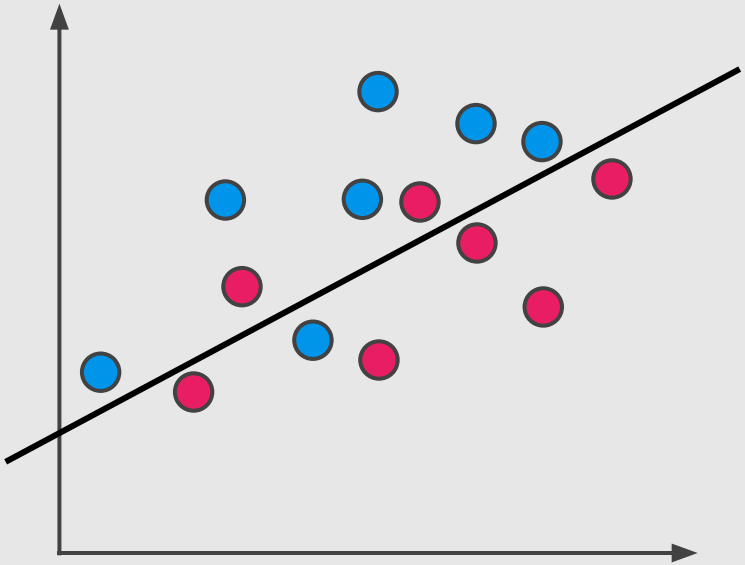
Confusion Matrix Table



Prediction

		Guessed Positive	Guessed Negative
Data	Positive	6 True positives	
	Negative		

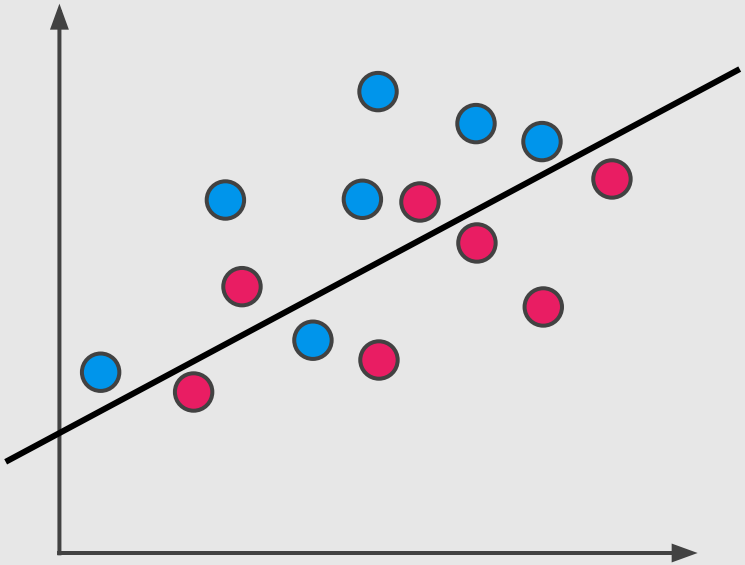
Confusion Matrix Table



Prediction

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

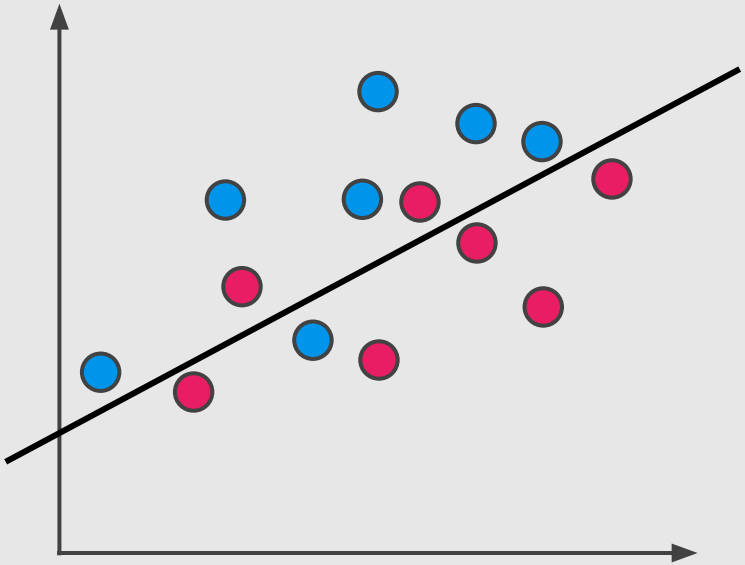
Confusion Matrix Table



Prediction

Data	Guessed Positive	Guessed Negative
Positive	6 True positives	
Negative	2 False positives	5 True negatives

Confusion Matrix Table



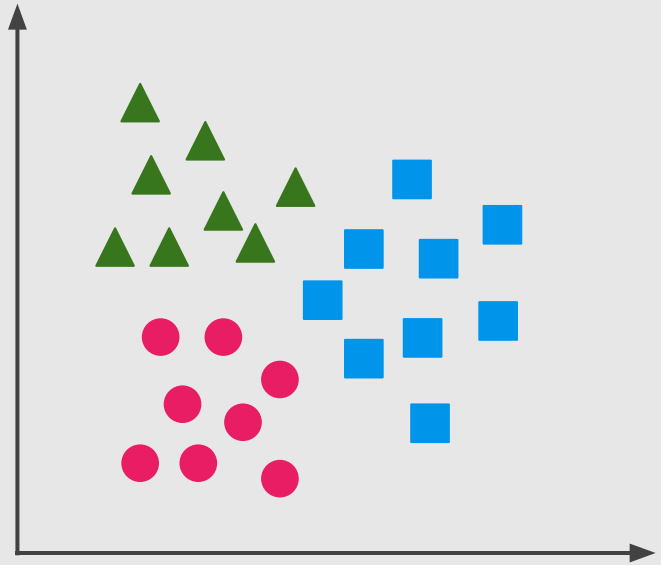
Prediction

Data

	Guessed Positive	Guessed Negative
Positive	6 True positives	1 False negative
Negative	2 False positives	5 True negatives

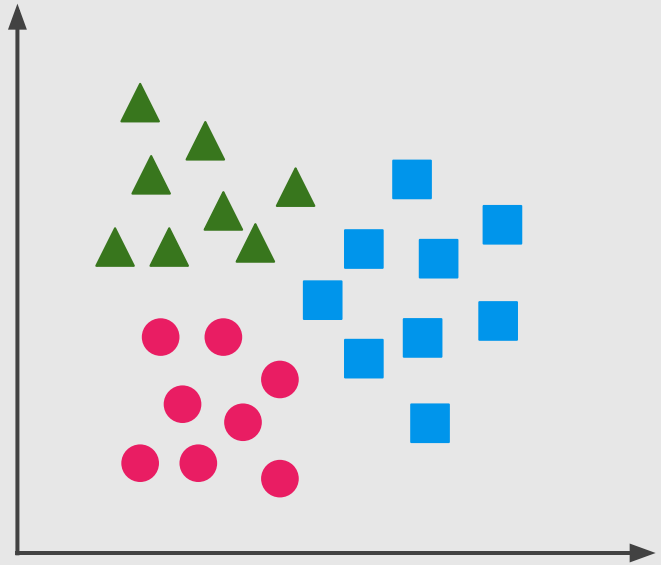
Confusion Matrix Table (n classes)

Class 1: ▲
Class 2: ■
Class 3: ●



Confusion Matrix Table (n classes)

Class 1: ▲
Class 2: ■
Class 3: ●



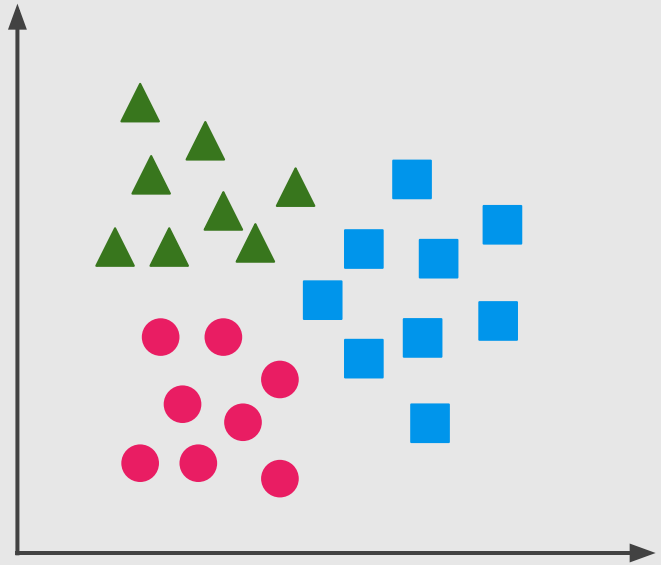
Predicted Class

True Class

	Guessed Class 1	Guessed Class 2	Guessed Class 3
Class 1			
Class 2			
Class 3			

Confusion Matrix Table (n classes)

Class 1: ▲
Class 2: ■
Class 3: ●

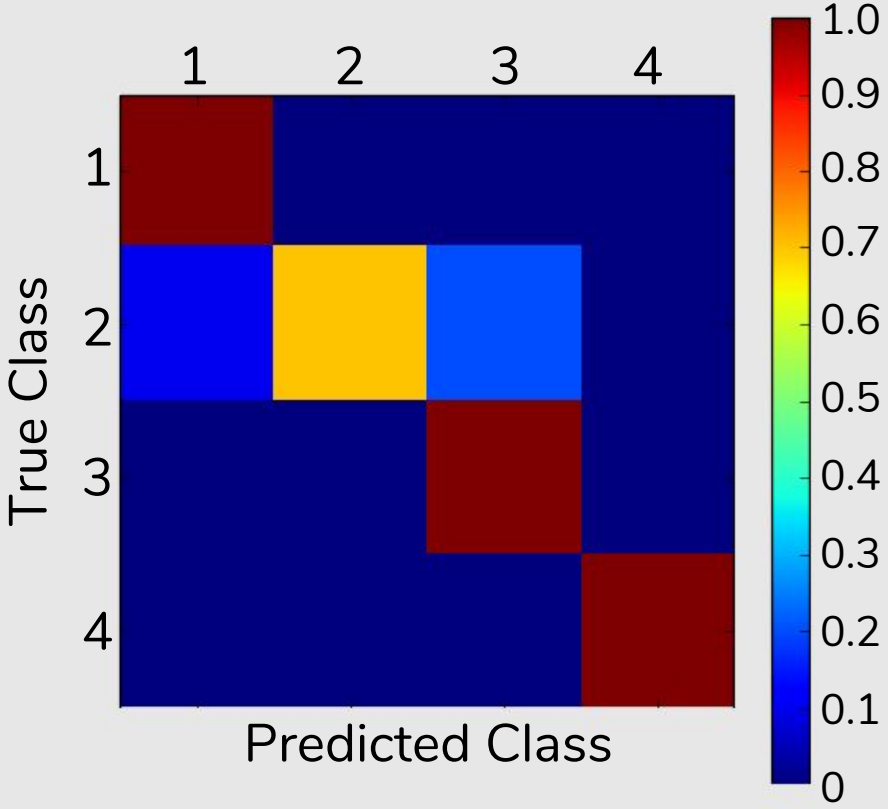


Predicted Class

True Class

	Guessed Class 1	Guessed Class 2	Guessed Class 3
Class 1	5	2	1
Class 2	3	6	0
Class 3	0	1	7

Confusion Matrix Table (n classes)



Accuracy

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Accuracy

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

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

Accuracy

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

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

Accuracy =

$$\frac{1{,}000 + 8{,}000}{}$$

Accuracy

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

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

Accuracy =

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

Accuracy

Email \ Folder	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

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

Accuracy

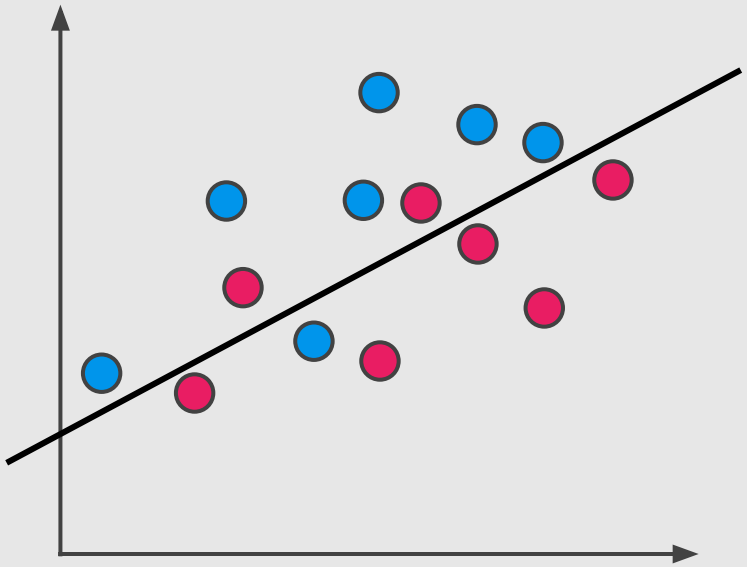
Email \ Folder	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

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

Accuracy =

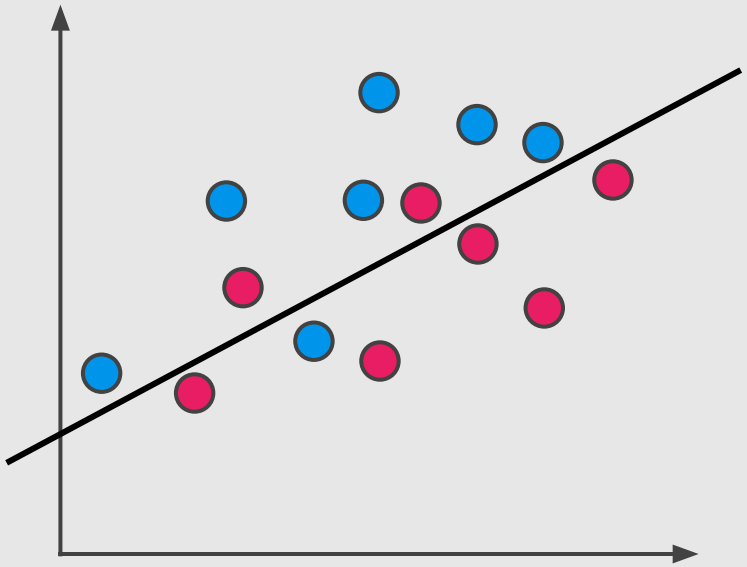
$$\frac{100 + 700}{1{,}000} = 80\%$$

Accuracy



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

Accuracy

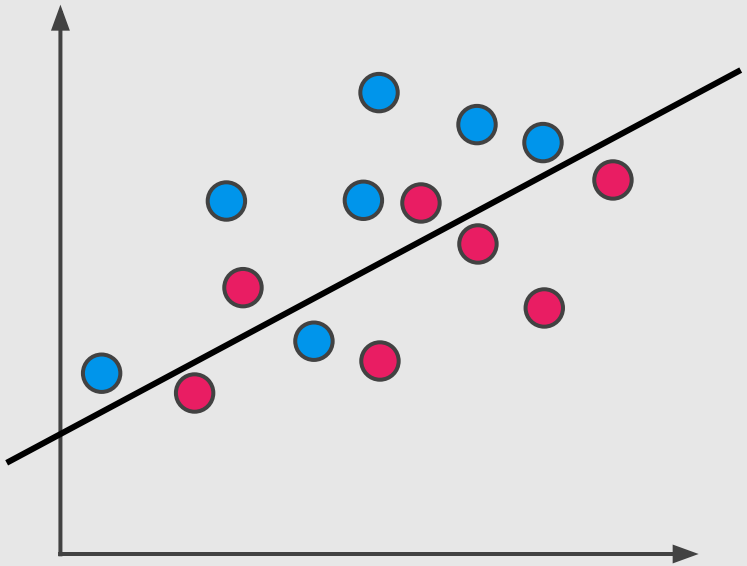


Accuracy:
Out of all the **data**, how many points did we classify correctly?

Accuracy =

$$\frac{\text{Correctly Classified Points}}{\text{All points}}$$

Accuracy



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

Accuracy =

$$\frac{\text{Correctly Classified Points}}{\text{All points}}$$

$$\frac{11}{11 + 3} = 78.57\%$$

Accuracy

Transactions \ Prediction	Fraudulent	Not Fraudulent
Fraudulent	0	472
Not Fraudulent	0	284,335

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

Accuracy =

$$\frac{0 + 284{,}335}{284{,}807} = 99.83\%$$

Normalized Accuracy

Transactions \ Prediction	Fraudulent	Not Fraudulent
Fraudulent	0	472
Not Fraudulent	0	284,335

Normalized Accuracy

Transactions \ Prediction	Fraudulent	Not Fraudulent
Fraudulent	0	472
Not Fraudulent	0	284,335

Normalized Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

Normalized Accuracy

Transactions \ Prediction	Fraudulent	Not Fraudulent
Fraudulent	0	472
Not Fraudulent	0	284,335

Normalized Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

$$\frac{\frac{0}{0 + 472} + \frac{284{,}335}{284{,}335 + 0}}{2} =$$

Normalized Accuracy

Transactions	Prediction: Fraudulent	Prediction: Not Fraudulent
Fraudulent	0	472
Not Fraudulent	0	284,335

Normalized Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

$$\frac{\frac{0}{0 + 472} + \frac{284{,}335}{284{,}335 + 0}}{2} =$$

$$\frac{0 + 100}{2} = 50\%$$

Normalized Accuracy

Accuracy = 80%

Folder

Email	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

Normalized Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

$$\frac{\frac{100}{100 + 170} + \frac{700}{700 + 30}}{2} =$$

$$\frac{37.0 + 95.9}{2} = 66.5\%$$

Normalized Accuracy

Accuracy = 90%

Diagnosis

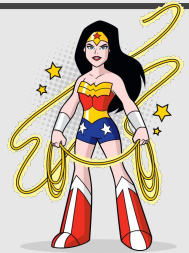
Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Normalized Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

$$\frac{\frac{1000}{1000 + 200} + \frac{8000}{8000 + 800}}{2} =$$

$$\frac{83.3 + 90.9}{2} = 87.1\%$$

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

	Diagnosed Sick	Diagnosed Healthy
Sick		False Negative 
Healthy	False Positive 	

	Sent to Spam Folder	Sent to Inbox
Spam	True Positive 	False Negative 
Not Spam	False Positive 	True Negative 

	Sent to Spam Folder	Sent to Inbox
Spam		False Negative 
Not Spam	False Positive 	

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

High Recall

Spam Detector

False positives **NOT** ok
False negatives ok

High Precision

Precision

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Precision

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Precision:
Out of all the patients we diagnosed with illness, how many were actually sick?

Precision

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Precision:
Out of all the patients we diagnosed with illness, how many were actually sick?

Precision

Patients \ Diagnosis	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Precision:
Out of all the patients we diagnosed with illness, how many were actually sick?

Precision =

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

Precision

Folder

Email

	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

Precision:
Out of all the emails sent to the spam inbox, how many did were actually spam?

Precision

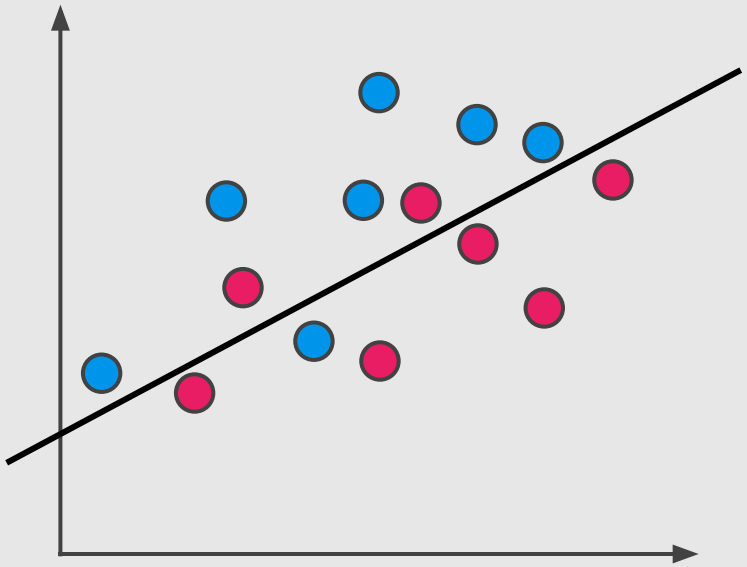
Email \ Folder	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

Precision:
Out of all the emails sent to the spam inbox, how many did were actually spam?

Precision =

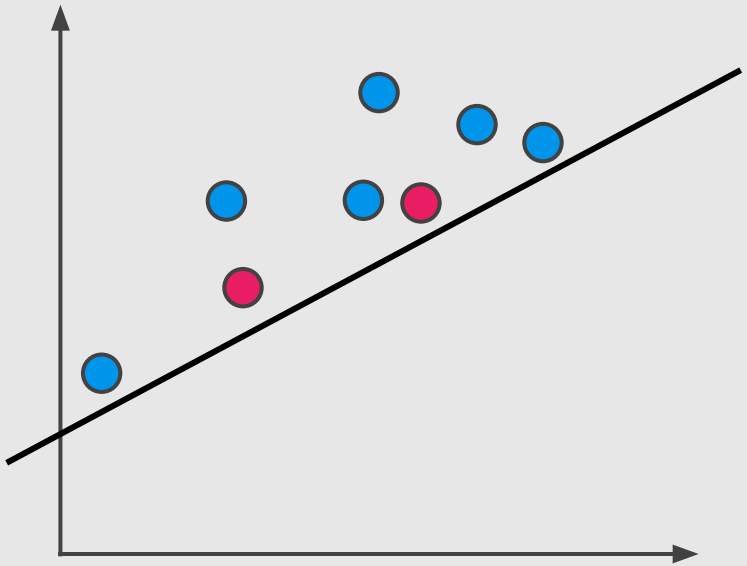
$$\frac{100}{100 + 300} = 76.9\%$$

Precision



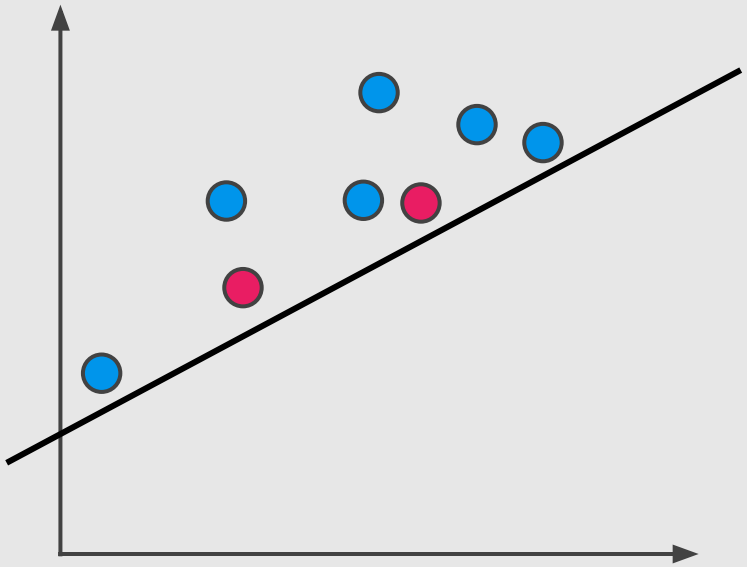
Precision:
Out of all the points we've predicted to be positive, how many are correct?

Precision



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

Precision



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

Precision =

$$\frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Recall

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Recall

Diagnosis

Patients	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

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

Recall

Patients \ Diagnosis	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

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

Recall

Patients \ Diagnosis	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

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

Recall =

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

Recall

Email \ Folder	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

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

Recall

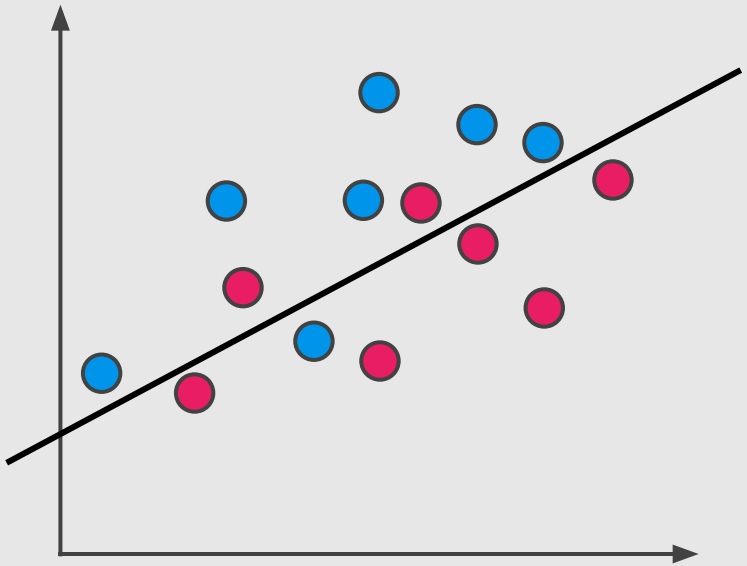
Email \ Folder	Spam Folder	Inbox
Spam	100	170
Not Spam	30	700

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

Recall =

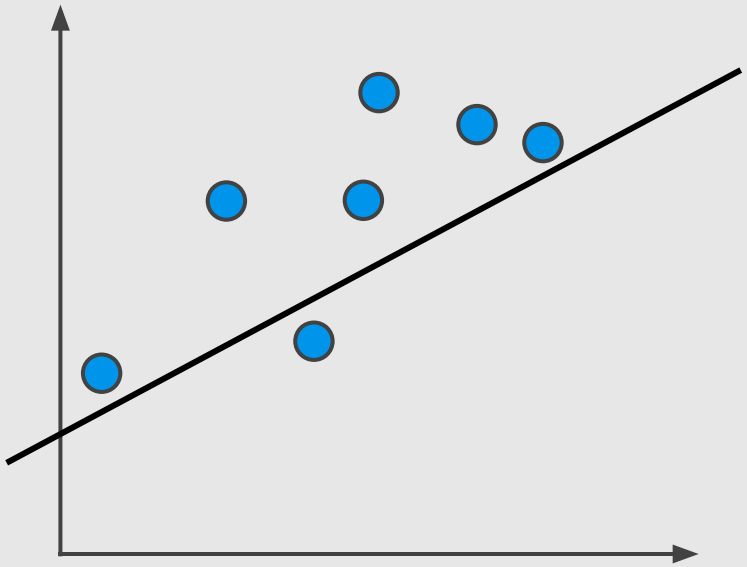
$$\frac{100}{100 + 170} = 37\%$$

Recall



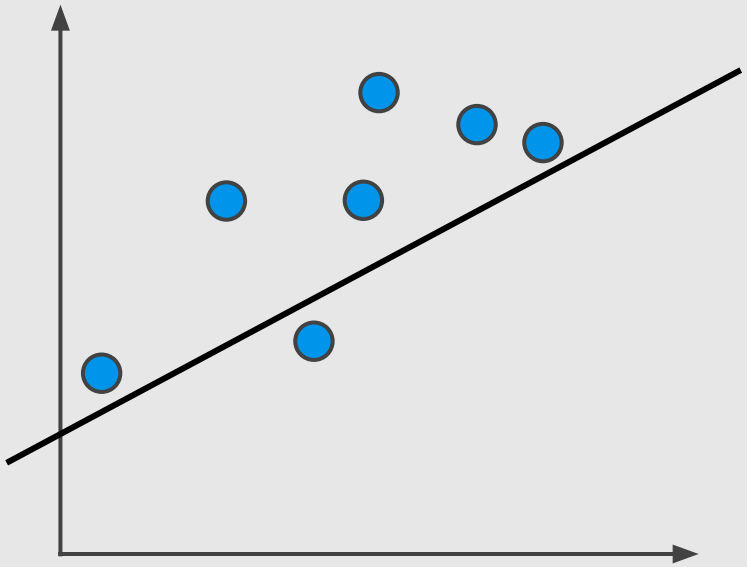
Recall:
Out of all the points labelled positive, how many did we correctly predict?

Recall



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

Recall

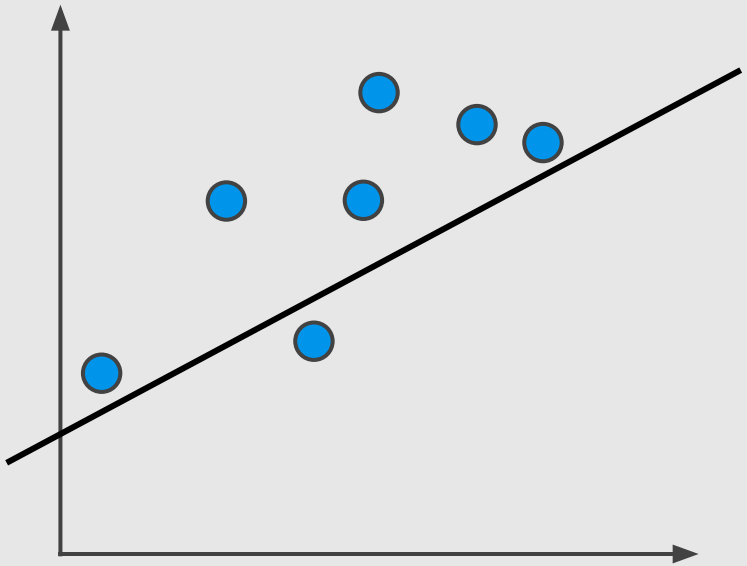


Recall:
Out of all the points labelled positive, how many did we correctly predict?

Recall =

$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

Recall



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

Recall =

$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\frac{6}{6 + 1} = 85.7\%$$

Precision and Recall

Medical Model

Precision: 55.7%

Recall: 83.3%

Spam Detector

Precision: 76.9%

Recall: 37%

One Score?

Medical Model

Precision: 55.7%

Recall: 83.3%

Average = 69.5%

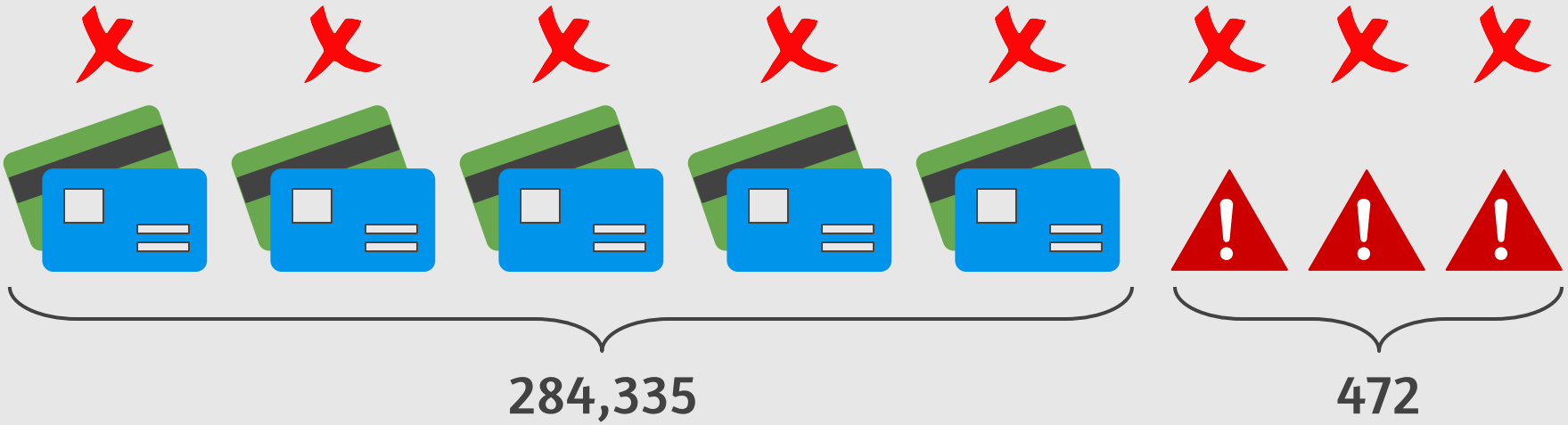
Spam Detector

Precision: 76.9%

Recall: 37%

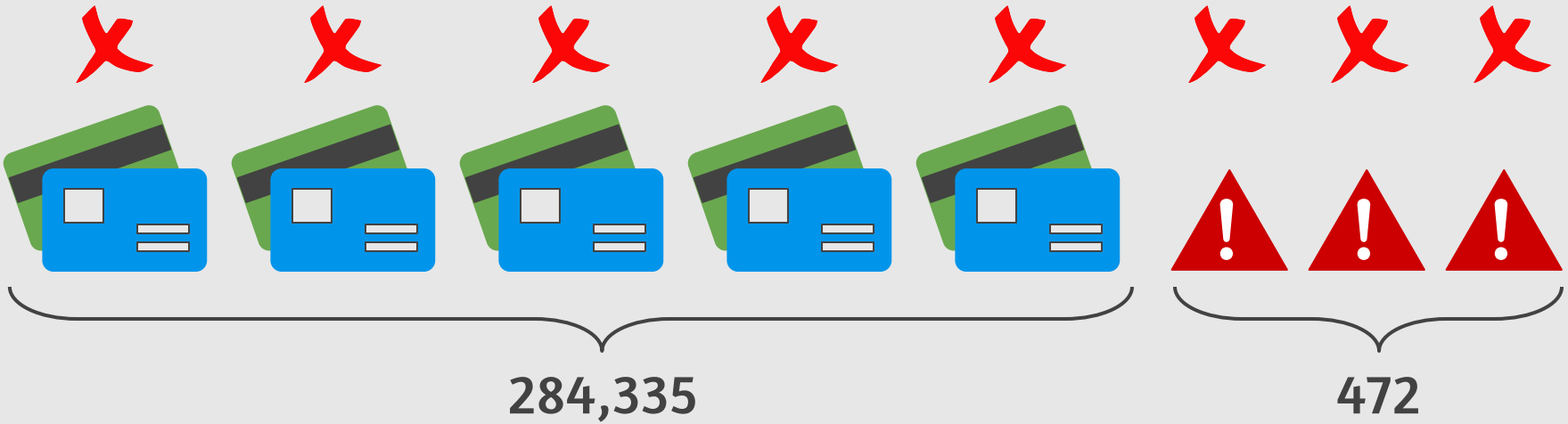
Average = 56.95%

Credit Card Fraud



Model: All transactions are fraudulent.

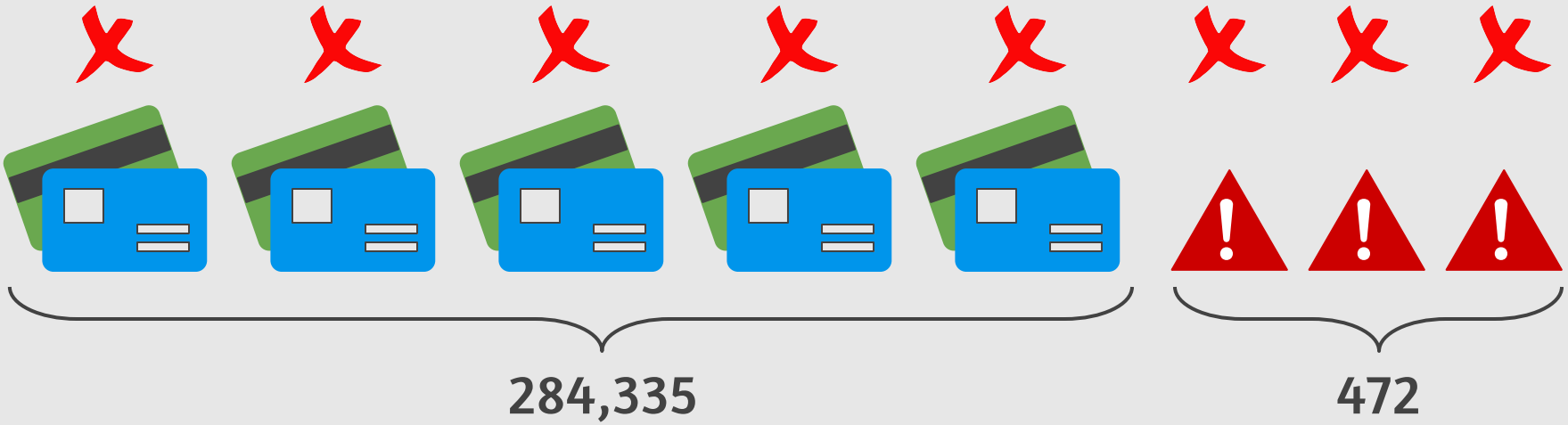
Credit Card Fraud



Model: All transactions are fraudulent.

$$\text{Precision} = \frac{472}{284{,}807} = 0.016\%$$

Credit Card Fraud



Model: All transactions are fraudulent.

$$\text{Precision} = \frac{472}{284{,}807} = 0.016\%$$

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

Harmonic Mean

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

Harmonic Mean

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

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

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

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

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

Precision: 1

Recall: 0

Average = 0.5

Harmonic Mean = 0

Precision: 0.2

Recall: 0.8

Average = 0.5

Harmonic Mean = 0.32

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

F1 Score = Harmonic Mean (Precision, Recall)

F1 Score

Medical Model

Precision: 55.7%

Recall: 83.3%

Average = 69.5%

F1 Score = 66.76%

F1 Score

Spam Detector

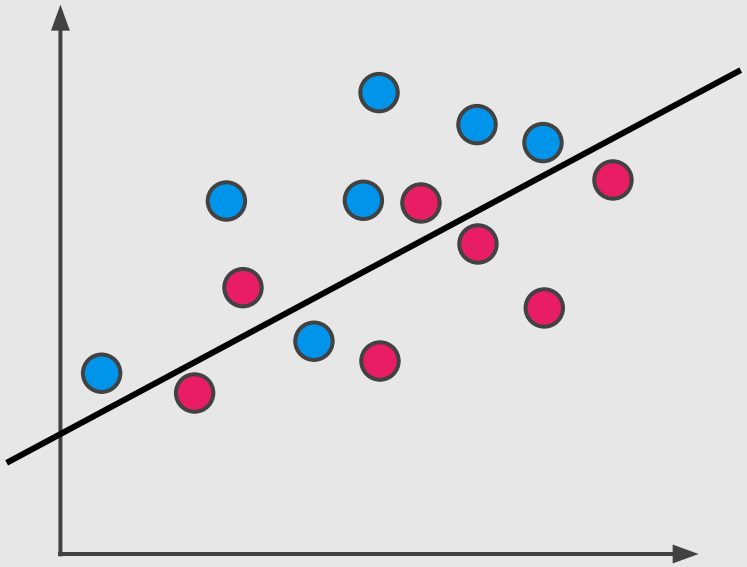
Precision: 76.9%

Recall: 37%

Average = 56.95%

F1 Score = 49.96%

F1 Score



Precision: 75%

Recall: 85.7%

Average = 80.35%

F1 Score = 80%

F_β Score

F_β Score

Precision

Recall

F_β Score

Precision

F0.5 Score

F1 Score

F2 Score

Recall

F_β Score

?

Precision F0.5 Score F1 Score F2 Score Recall

F_β Score

F10 Score

Precision F0.5 Score F1 Score F2 Score Recall

F_β Score

F10 Score

Precision F0.5 Score F1 Score F2 Score Recall

F_β Score

F1 Score = Harmonic Mean (Precision, Recall)

F_β Score

F1 Score = Harmonic Mean (Precision, Recall)

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

F_β Score

F1 Score = Harmonic Mean (Precision, Recall)

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

$$F_1 = 2\frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2\frac{\text{precison}\cdot\text{recall}}{\text{precision} + \text{recall}}$$

F_β Score

$$F_1 = 2\,\frac{\text{precison}\cdot\text{recall}}{\text{precision} + \text{recall}}$$

$$F_\beta = (1 + \beta^2)\,\frac{\text{precison}\cdot\text{recall}}{(\beta^2\cdot\text{precision}) + \text{recall}}$$

References

- https://en.wikipedia.org/wiki/Precision_and_recall
- https://en.wikipedia.org/wiki/Binary_classification
- https://en.wikipedia.org/wiki/F1_score
- https://www.quora.com/What-is-an-intuitive-explanation-of-F-score

Machine Learning Courses

- Luis Serrano: https://www.youtube.com/watch?v=aDW44NPhNw0