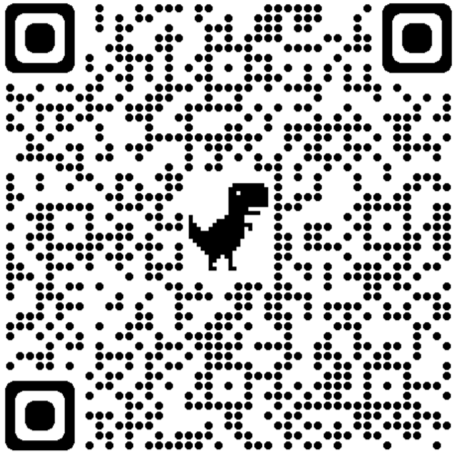
SPAI

An AI Singapore Student Chapter

Advanced ML Workshop

Day 2



Scan the QR code to mark your attendance

Attendance

Learning Objectives

☑ Train, Test and Validate

☑ Cross Validation

☑ Understanding Bias & Variance in ML models

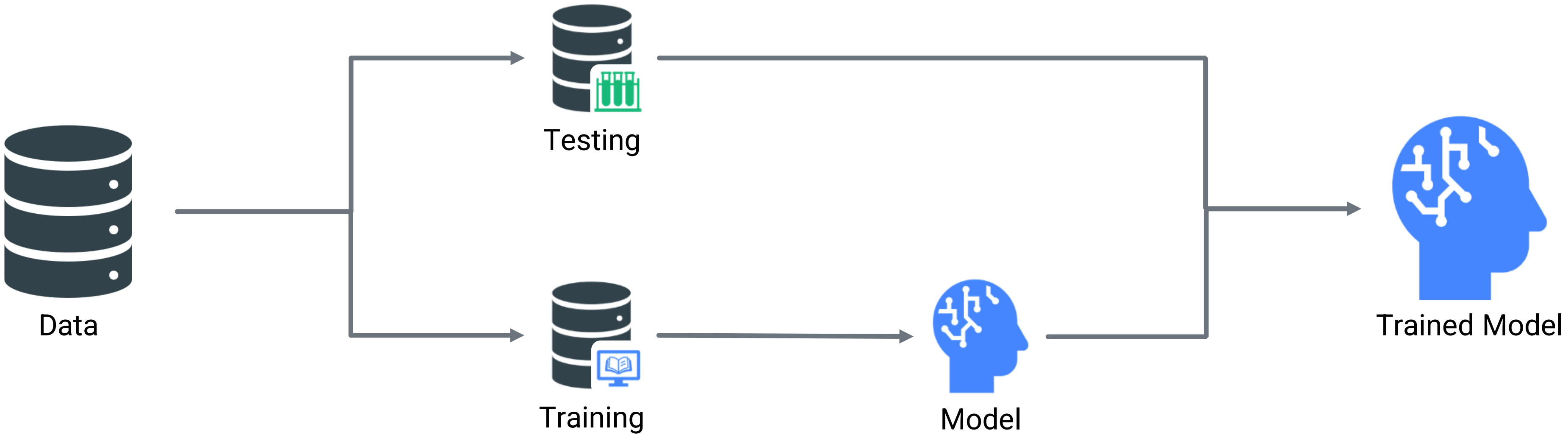
☑ Interpreting Model Complexity using learning curve

Predict if customers will switch to your telco based on certain features

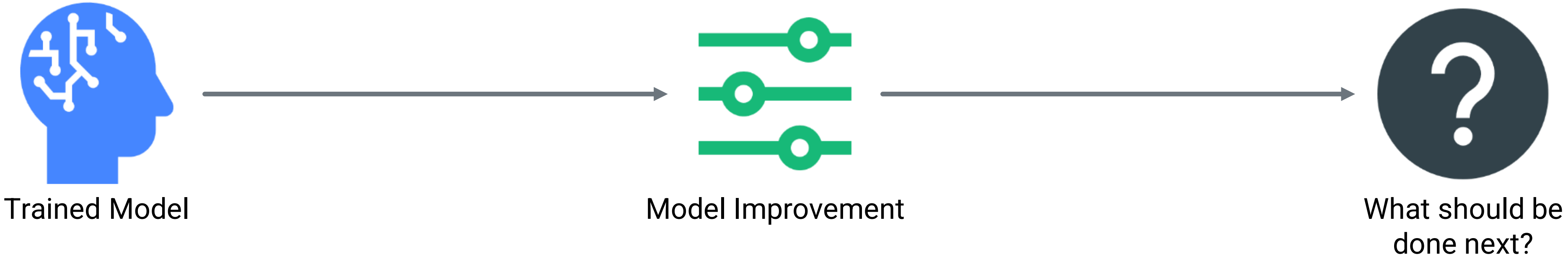
Scenario

SPAI

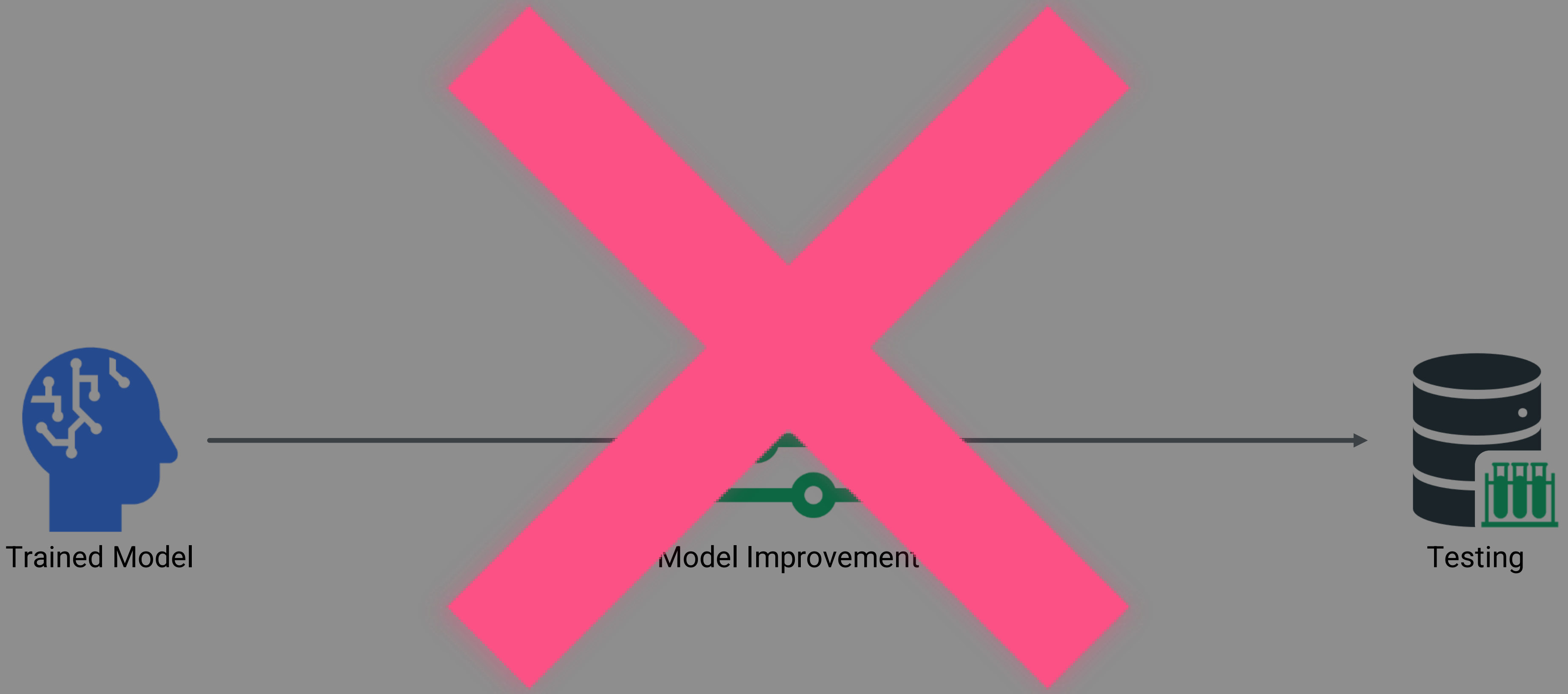
We learnt previously...



Improving Model



Improving Model



Trained Model

Model Improvement

Testing

Why can’t we use the testing set to evaluate the model?

The testing set allows us to get a realistic representation of the performance of the model

Recall

Why can’t we use the testing set

- ☑ Testing set allows us to test the model in an unbiased way
- ☑ However, we are improving our model based on the results of our testing set
- ☑ This can result in biases in our model, trying to “suit” our testing set

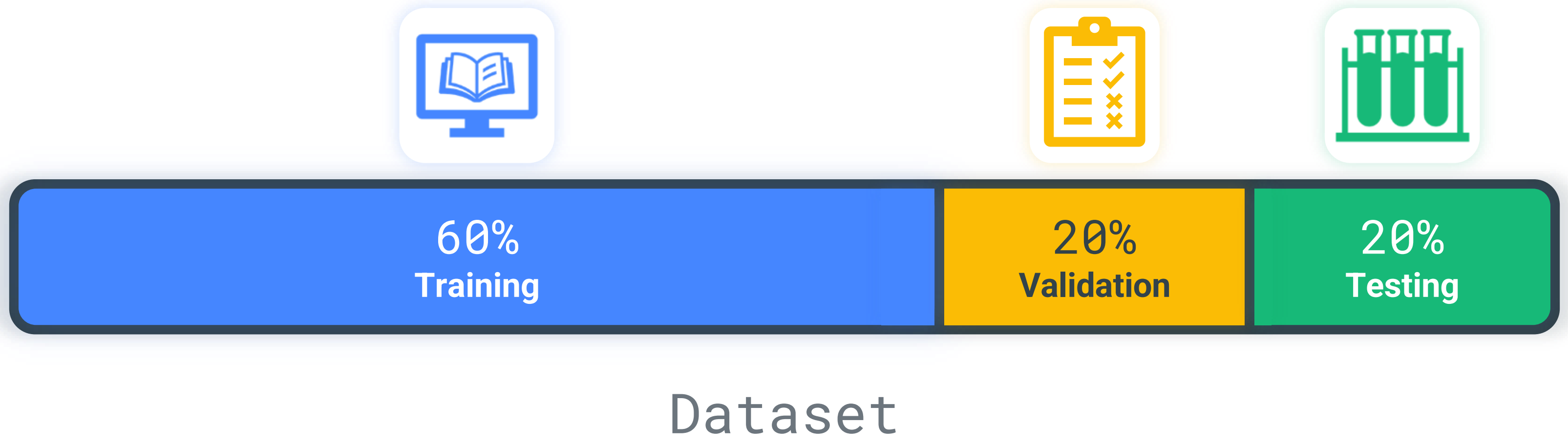
1

Train, Validate and Test

Using the same ideology of training and testing, we now further split our data into 3 sets

What is it?

What is Train Test Validate?



Training Data

Usually takes up **60%** of the dataset

Dataset only used to train the model

NEVER use model score from this dataset to judge the performance of the model

Validation Data

Usually takes up to **20%** of the dataset

Dataset only used to evaluate model performance

NEVER use this dataset as a conclusion of the model performance

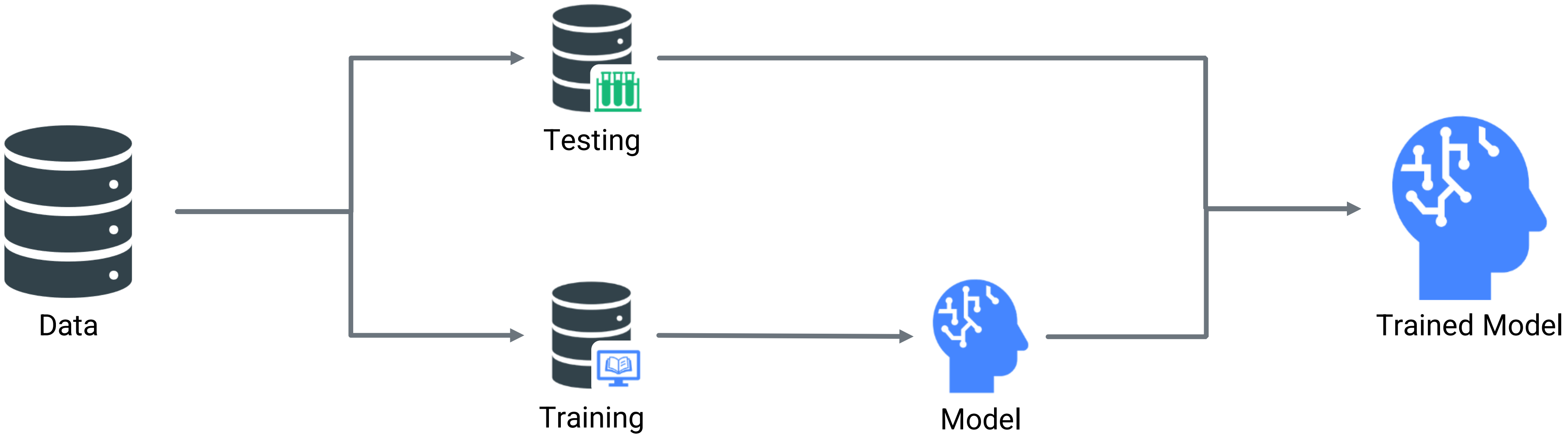
Testing Data

Usually takes up **20%** of the dataset

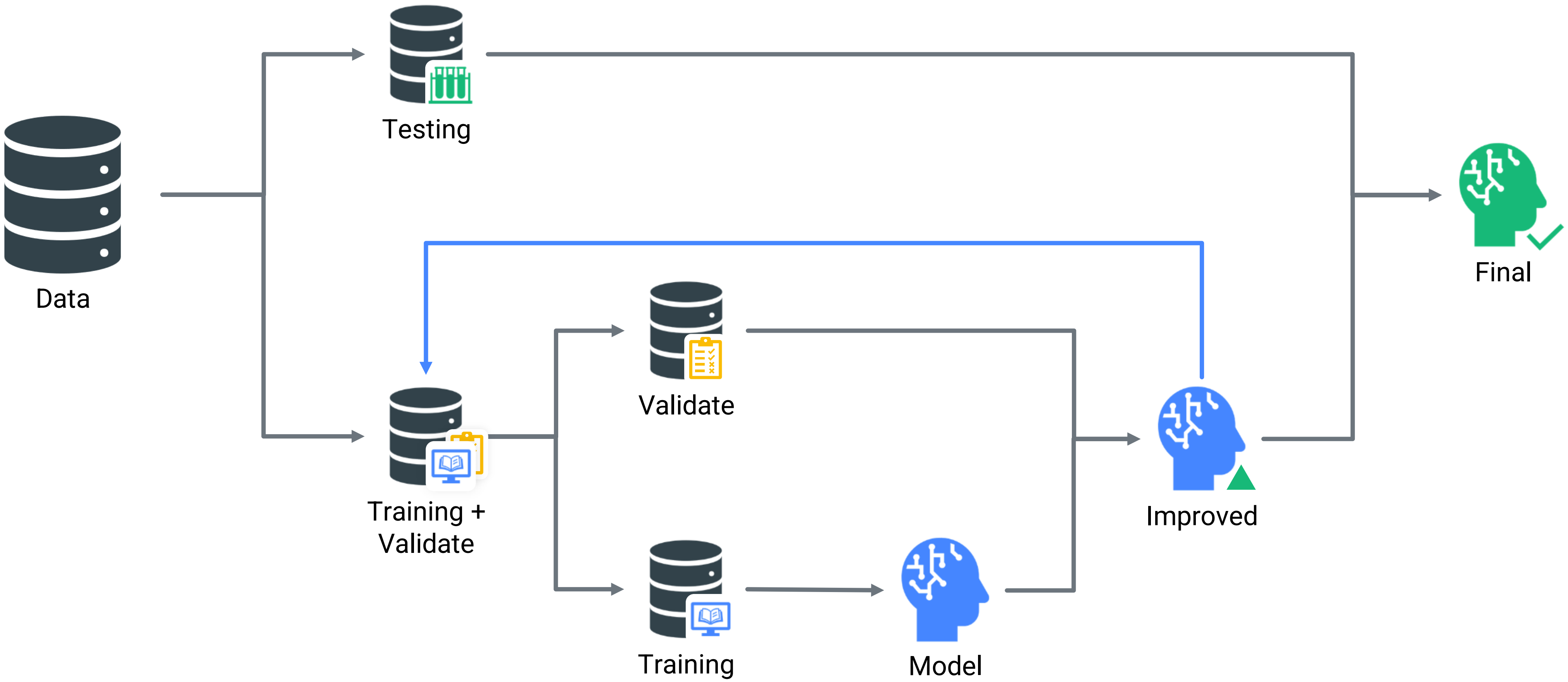
Dataset only used to evaluate final model performance

NEVER make changes on the model based of the performance from this dataset

We learnt previously…



Actual Diagram



Very waste of data.

Only 60% of data is used to train the model

Problem

2

Cross Validation (CV)

Why Cross Validate

☑ Helps us make full use of our dataset

☑ Training model is a random process

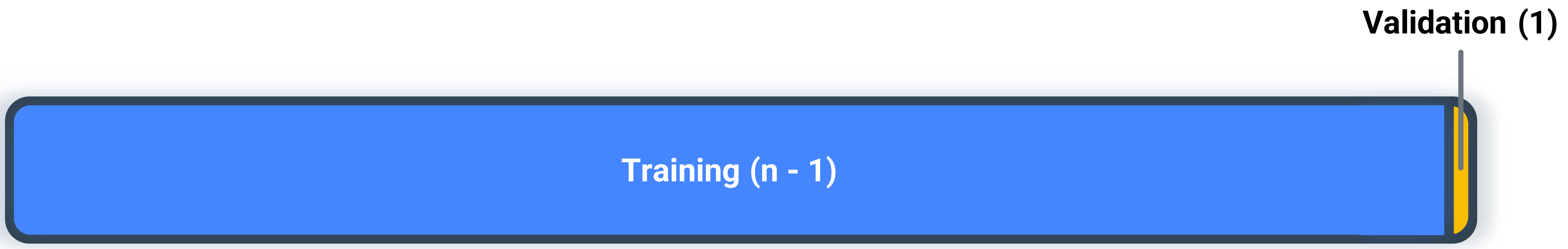
Types of CV

LOO

K-fold

Stratified

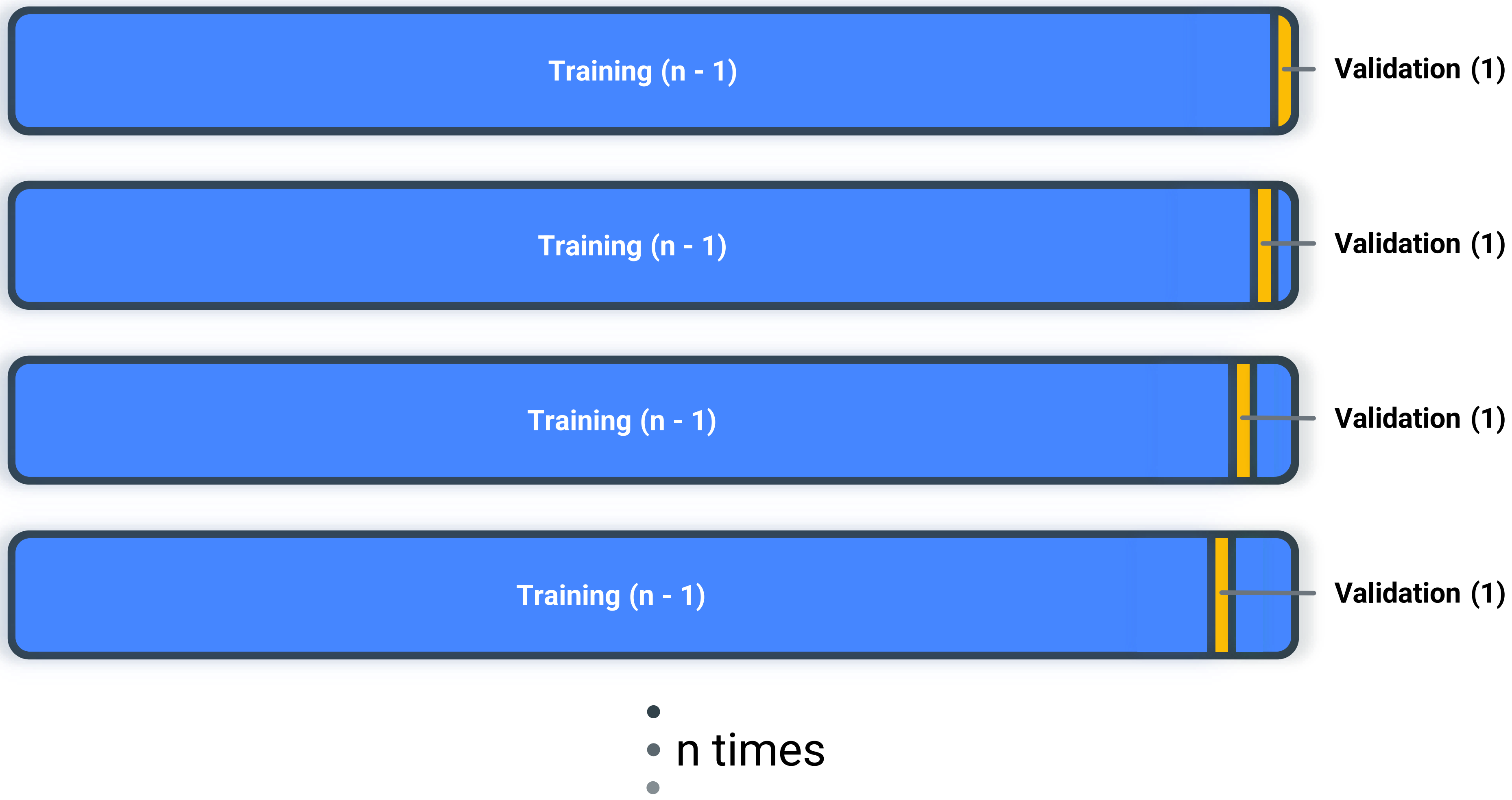
Leave one out CV



Training &
Validation Set

Where n is the number of rows of data

Leave one out CV



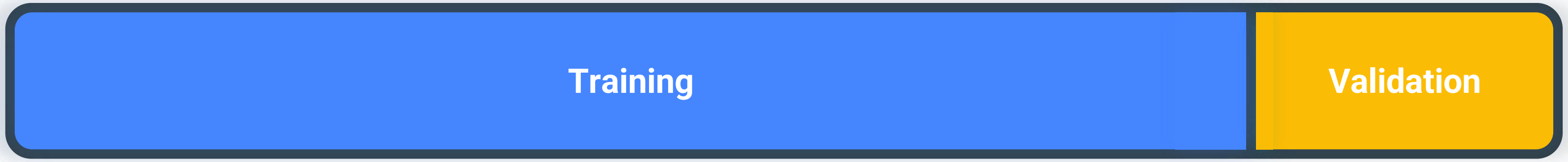
LOO

Best use of data as guarantees 100% use of it

Very time consuming and resource intensive

Not used often

K-Fold CV

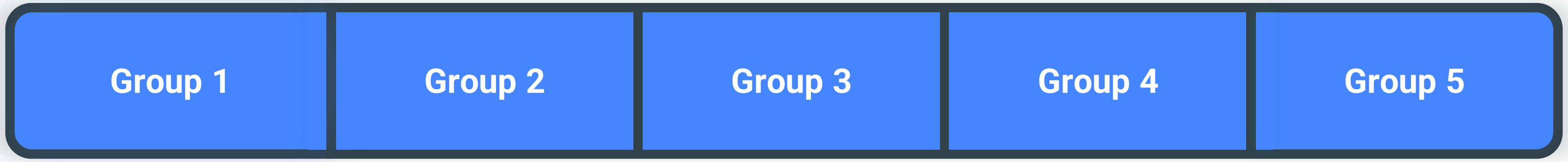


Training

Validation

Training &
Validation Set

K-Fold CV



Group 1

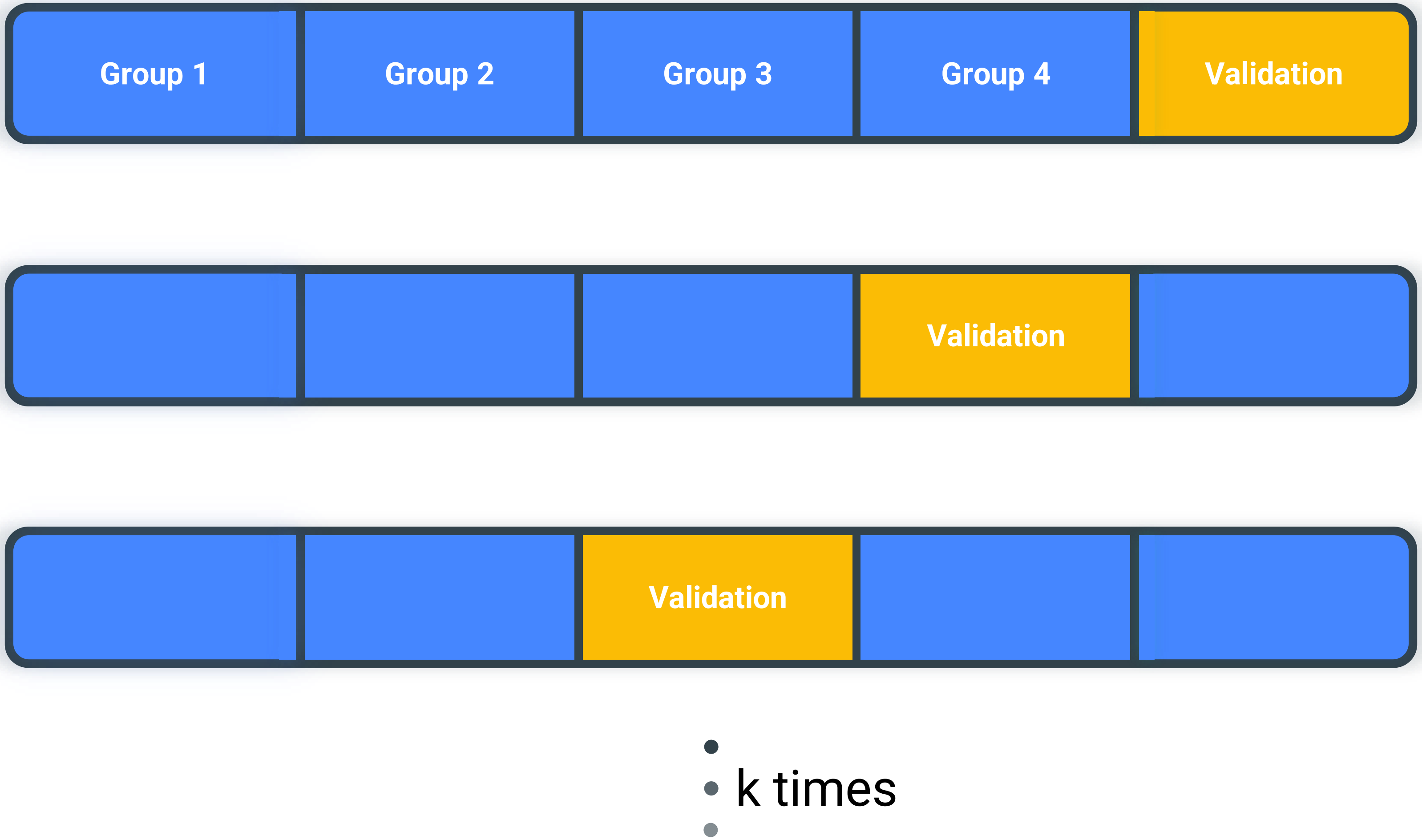
Group 2

Group 3

Group 4

Group 5

K-Fold CV



Where k = 5

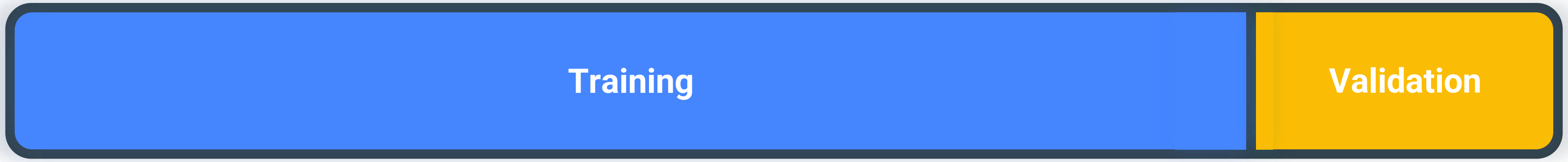
K-Fold

Good use of data as guarantees 100% use of it

Very efficient and fast process

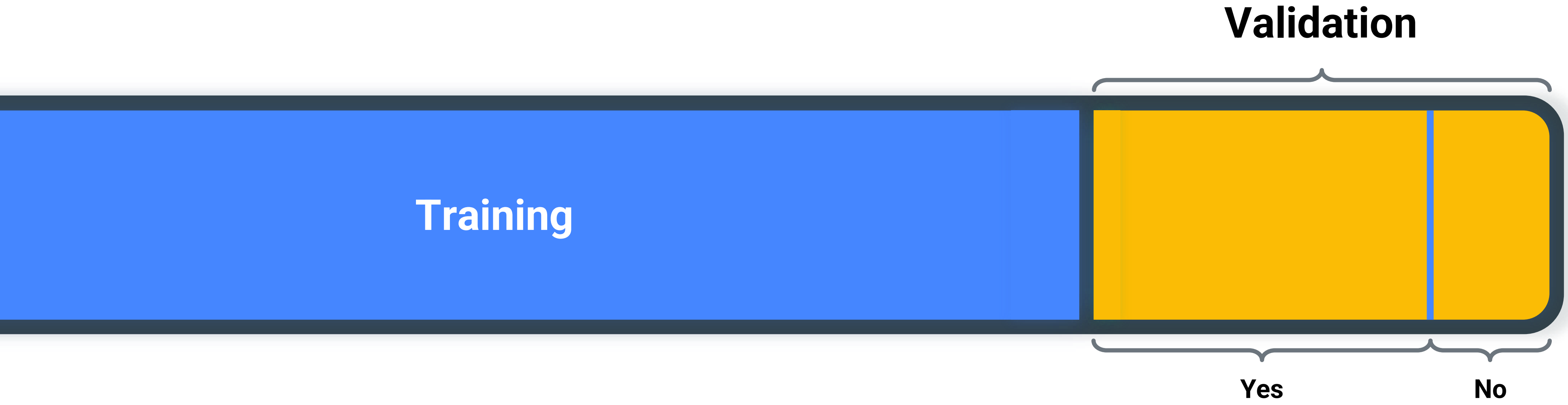
Most used method of cross validation

Stratified CV



Training &
Validation Set

Stratified CV



Where ratio of classes is 5:2 and there are only 2 classes

◐

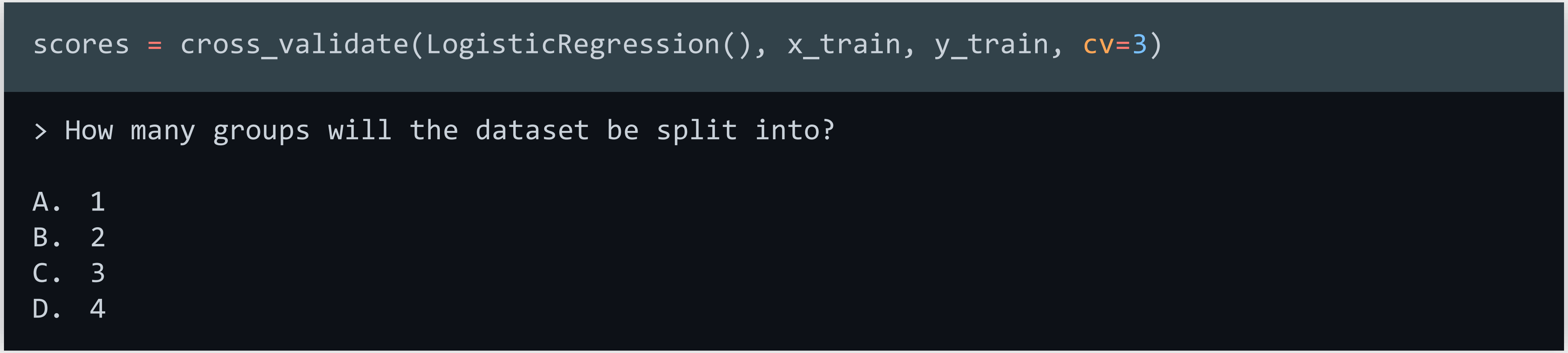
Stratified

Good use of data as guarantees 100% use of it

Slightly slower than K-fold but still highly efficient

Useful for imbalanced classes

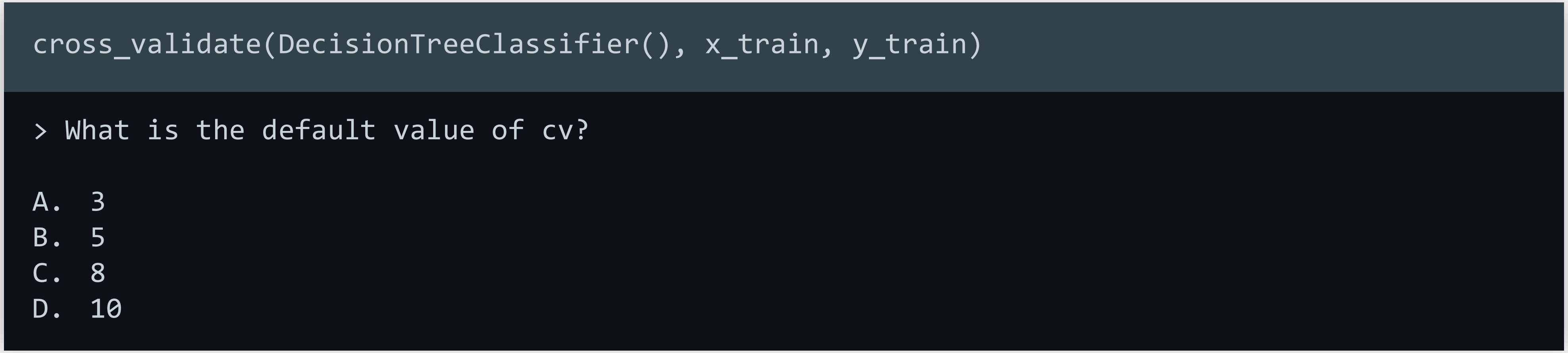
Knowledge Check



```
scores = cross_validate(LogisticRegression(), x_train, y_train, cv=3)
```

```
> How many groups will the dataset be split into?

A.  1
B.  2
C.  3
D.  4
```

Knowledge Check



```
cross_validate(DecisionTreeClassifier(), x_train, y_train)
```

```
> What is the default value of cv?

A.  3
B.  5
C.  8
D.  10
```

Practice Time!

10 Minutes

Please attempt exercise 1

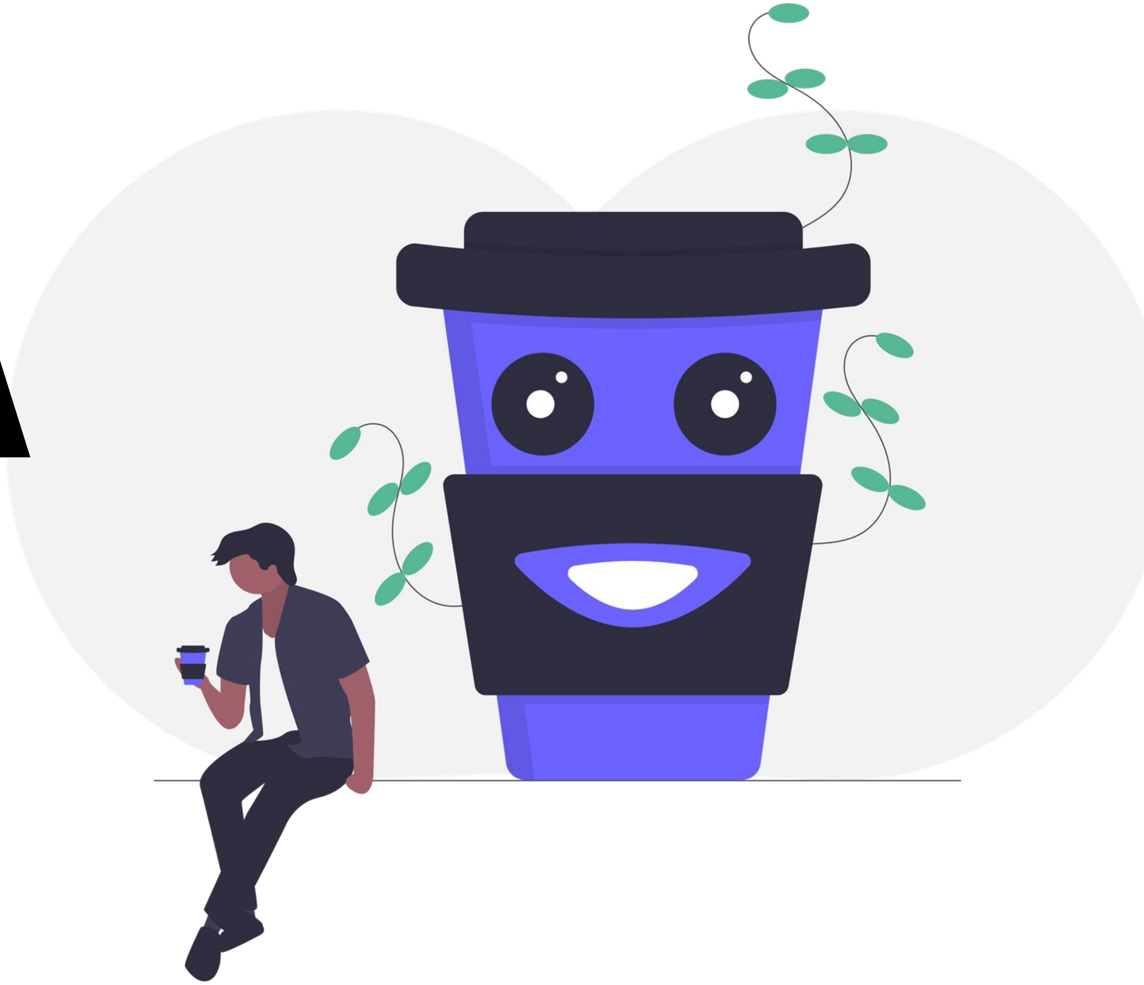
We will go through the exercises later

Times up

We will now go through the exercises

Break & QnA

10 Minutes



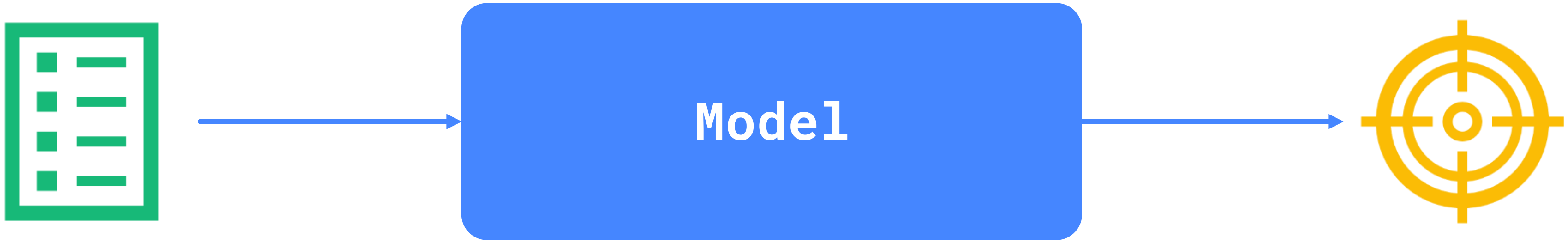
3

Bias and Variance

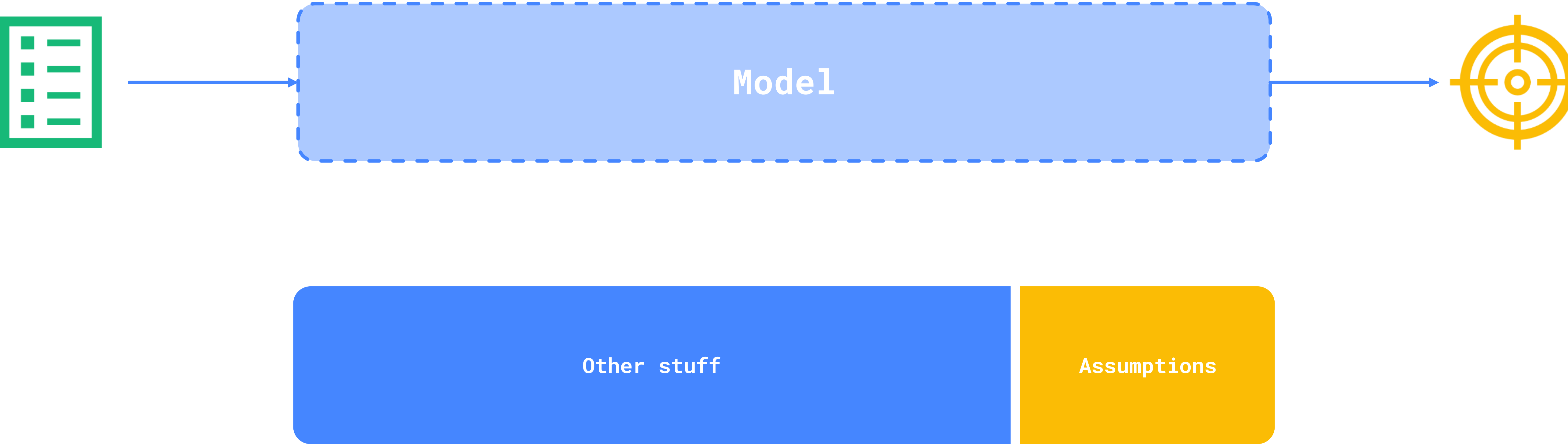
Amount of assumptions made by a model to make the target function easier to learn

What is Bias

What is Bias?



What is Bias?



Why model make assumptions?

- ☑ Makes it easier to learn and predict
- ☑ Results in faster learning speed

Bias

Low **Bias:** Less assumptions made

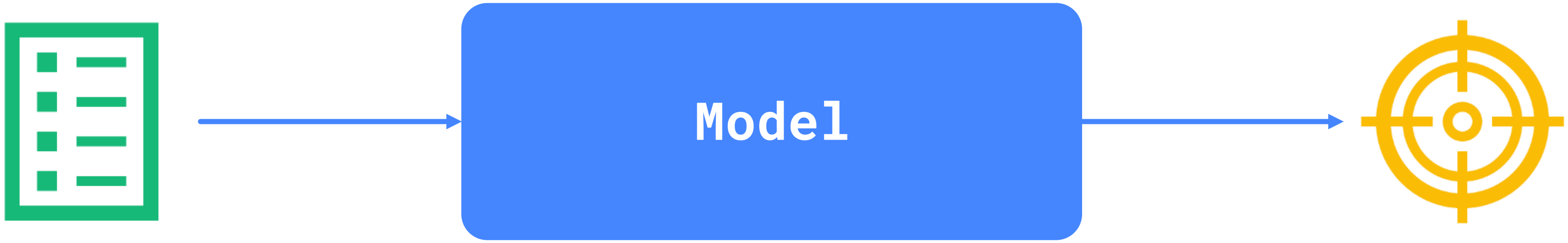
High **Bias:** More assumptions made

Amount changes to the estimate of the target function if different training data was used

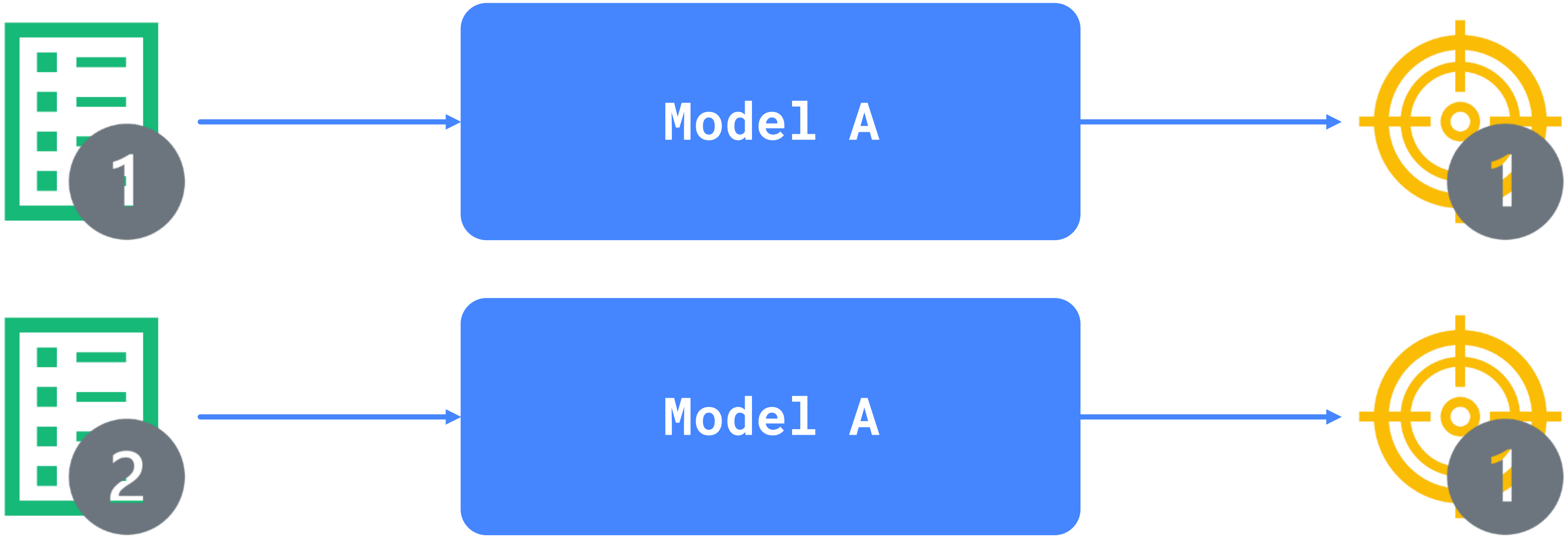
What is Variance

SPAI

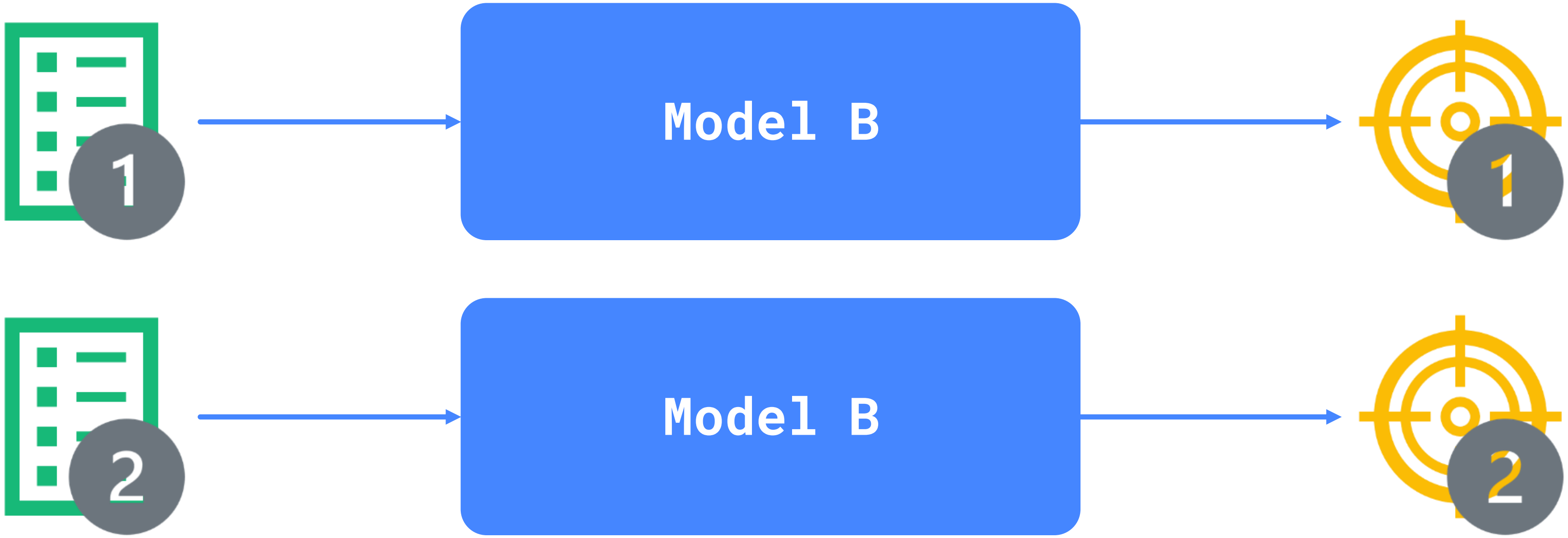
What is Variance?



What is Variance?



What is Variance?

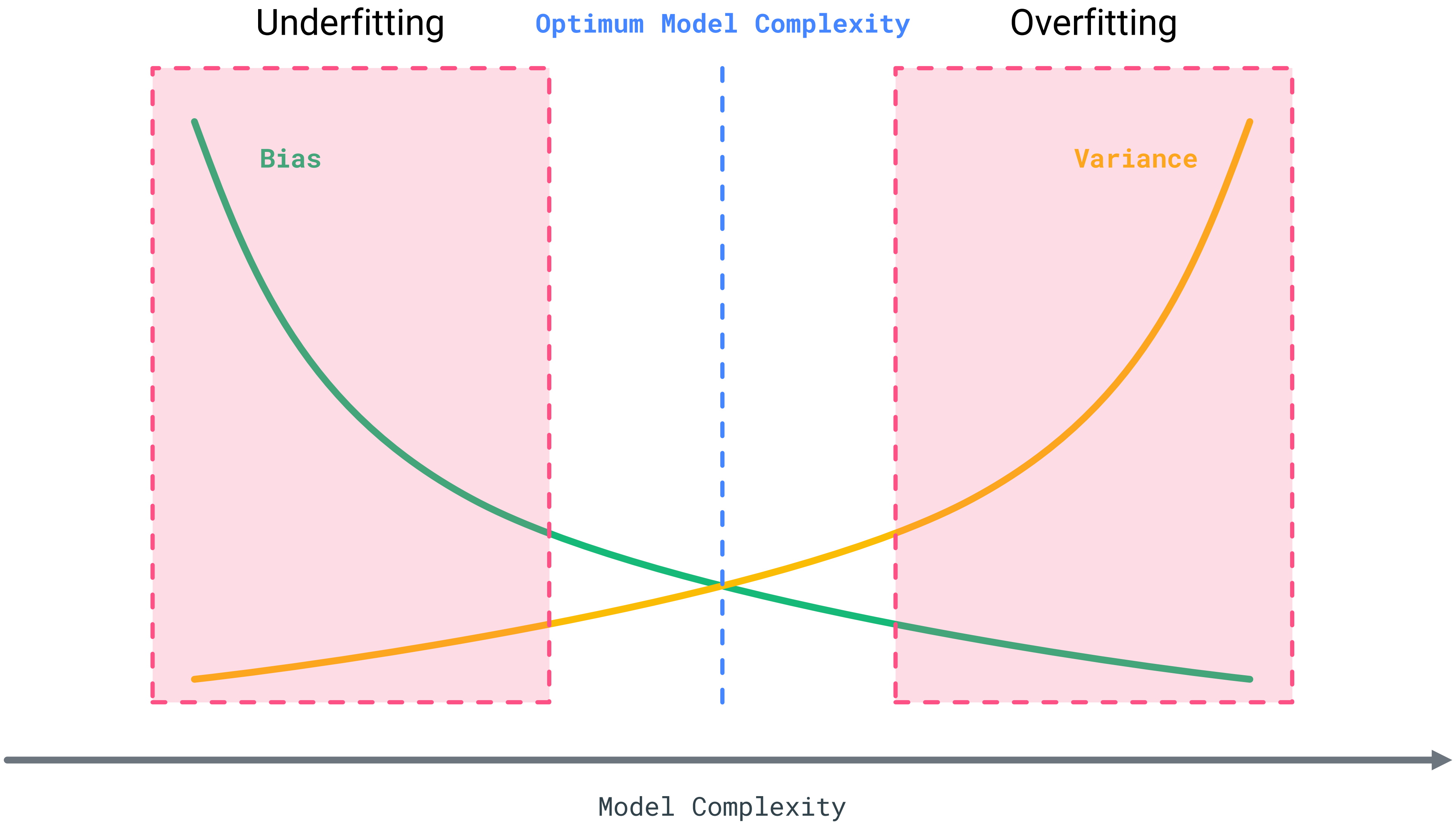


Variance

Low Variance: Small changes to training dataset results in small changes to prediction

High Variance: Small Changes to training dataset results in large changes to prediction

Bias & Variance together

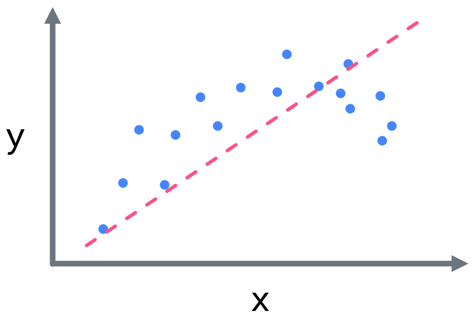


SPAI

Under/Over fitting

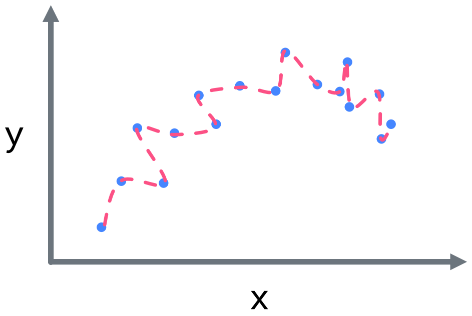
Underfitting

- High Bias
- Low Variance

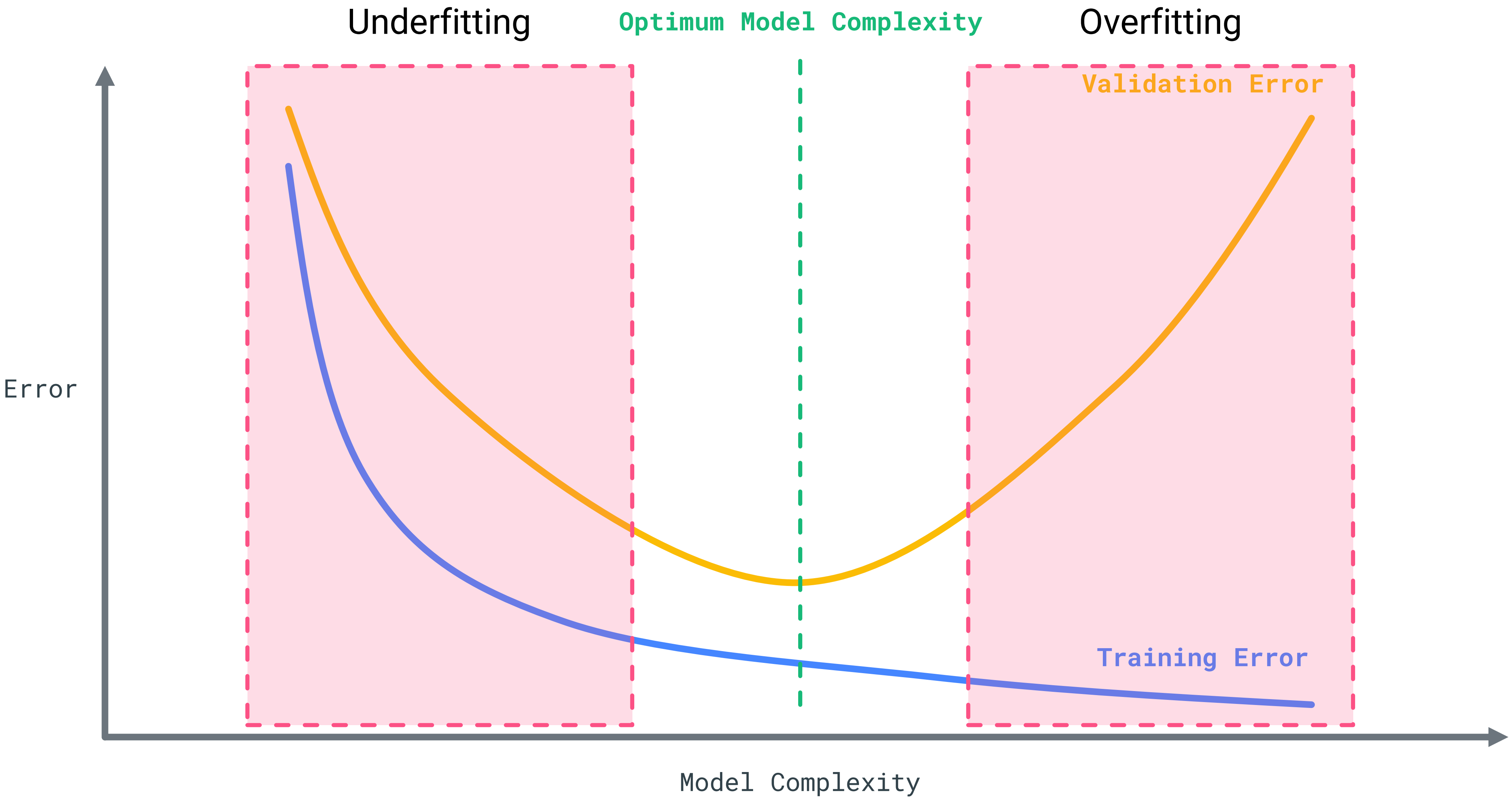


Overfitting

- Low Bias
- High Variance



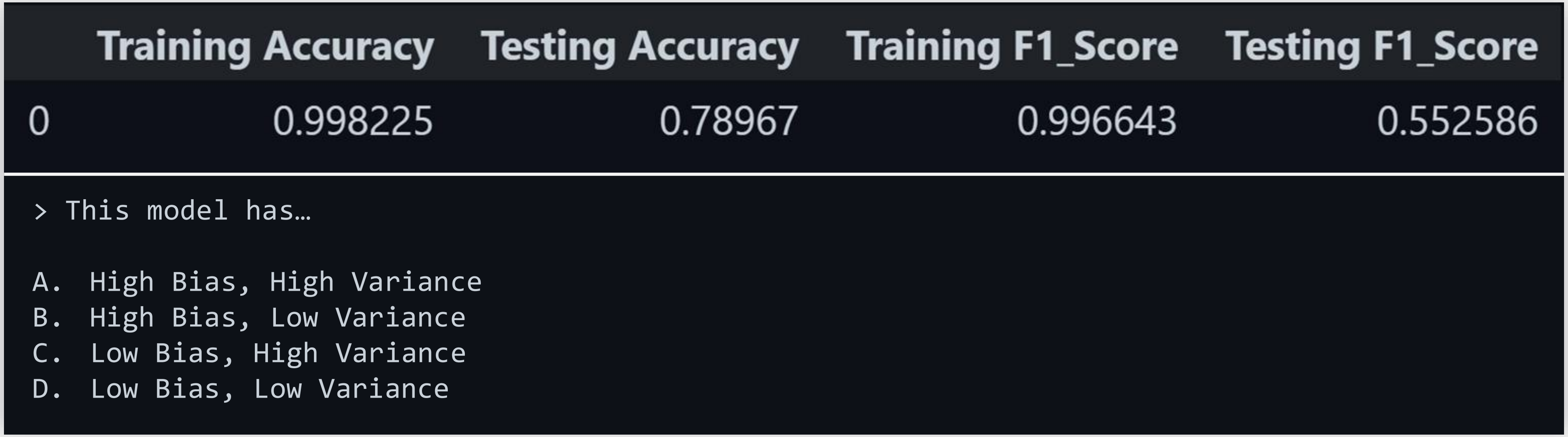
Underfitting & Overfitting



Knowledge Check

```
> Which of the following statements are True

A.  A model with low bias and high variance is an underfitted model

A.  When a model changes drastically with small changes on its training set, it is said to
    have high bias

A.  When a model changes drastically with small changes on its training set, it is said to
    have high variance

A.  It is best when model have high bias and high variance
```

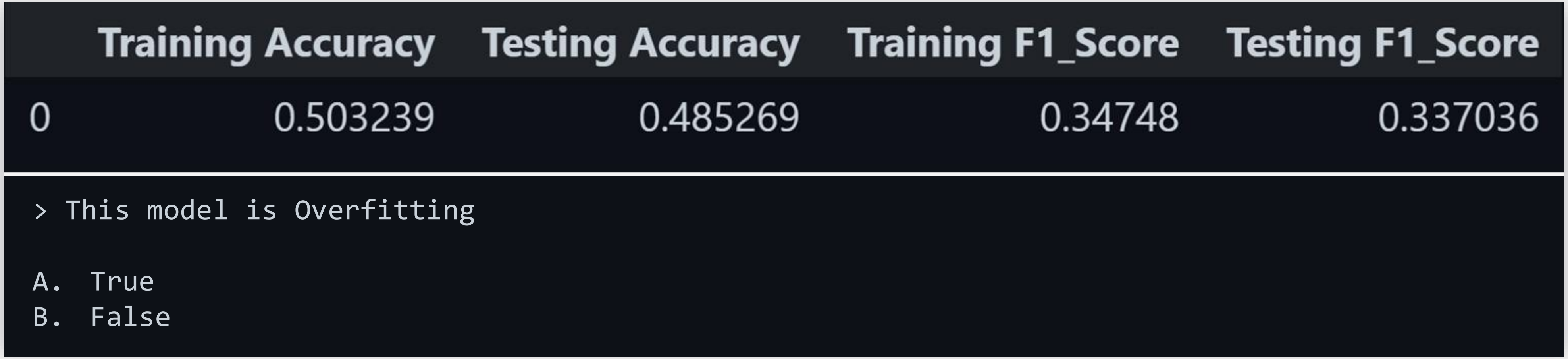
Knowledge Check



	Training Accuracy	Testing Accuracy	Training F1_Score	Testing F1_Score
0	0.998225	0.78967	0.996643	0.552586

```
> This model has…

A.  High Bias, High Variance
B.  High Bias, Low Variance
C.  Low Bias, High Variance
D.  Low Bias, Low Variance
```

Knowledge Check



	Training Accuracy	Testing Accuracy	Training F1_Score	Testing F1_Score
0	0.503239	0.485269	0.34748	0.337036

```
> This model is Overfitting

A.  True
B.  False
```

Purpose: Allows us to understand why a model performs a certain way

4

Model Learning Curves

A learning curve is the correlation between a model's score against the amount of data it is given

What is a Learning Curve?

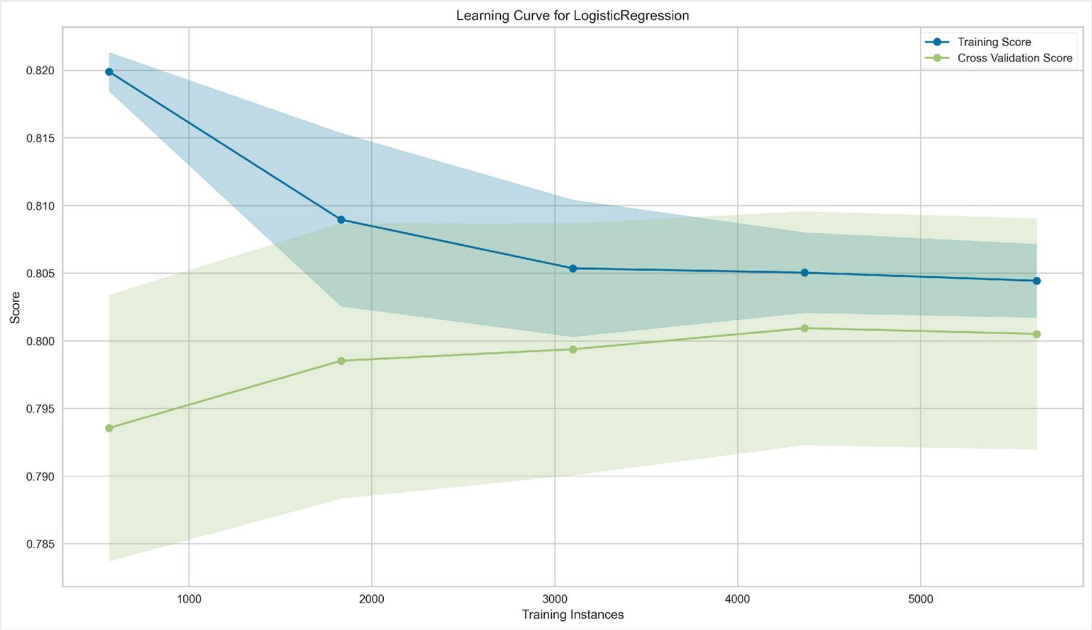
SPAI

Why plot learning curves?

☑ Allows us to see if a model is over/underfitting

☑ From there, we can make useful decisions on how we can improve our model

Good Fit Characteristics

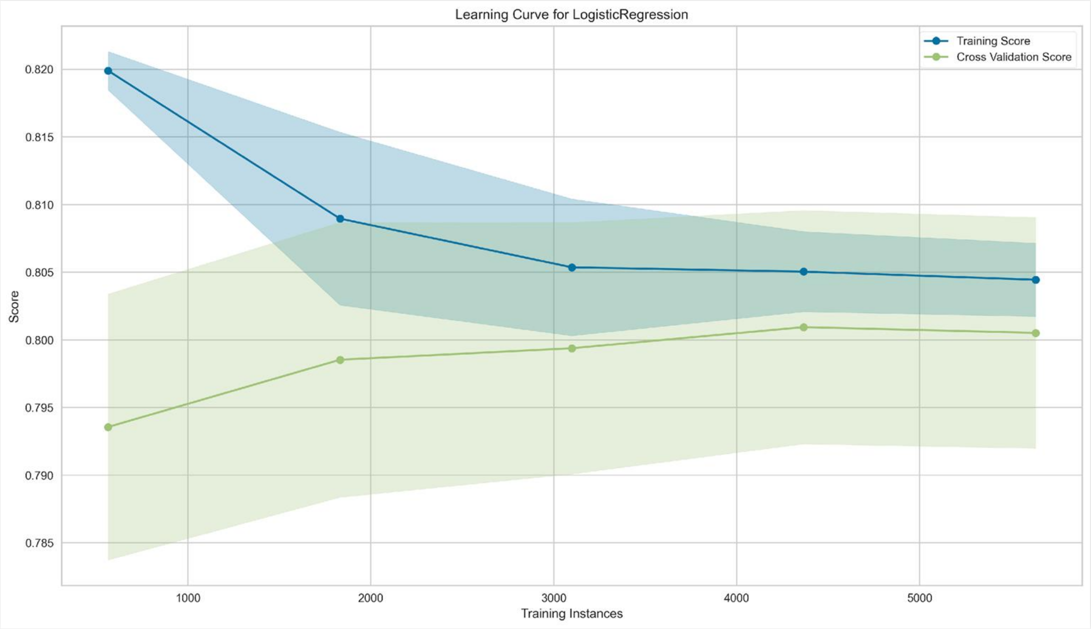


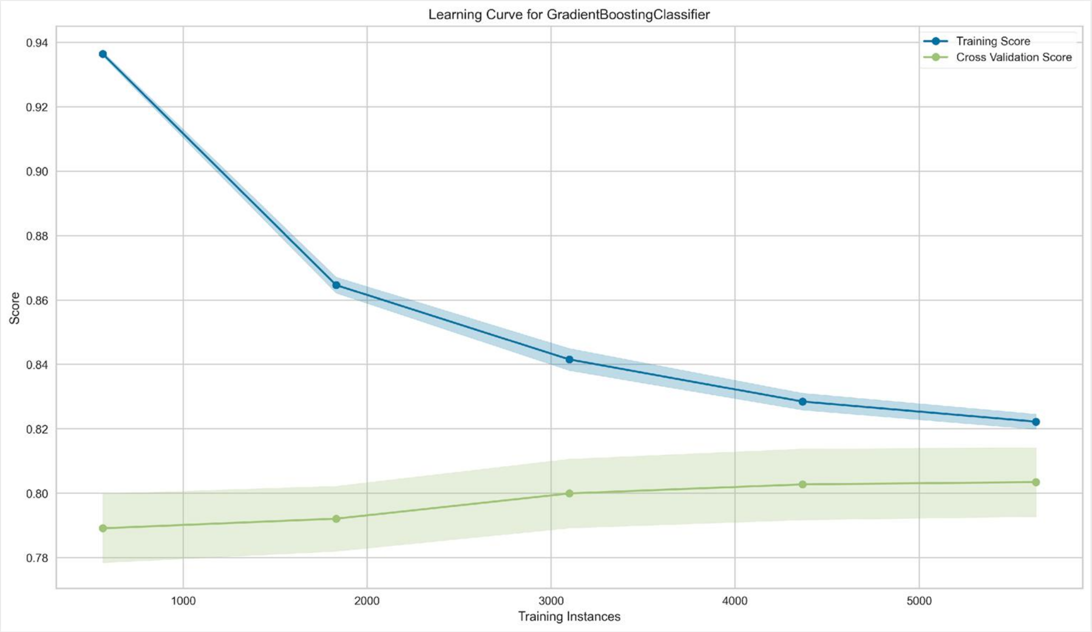
Lines moves towards each other

Lines maintain small space between each other

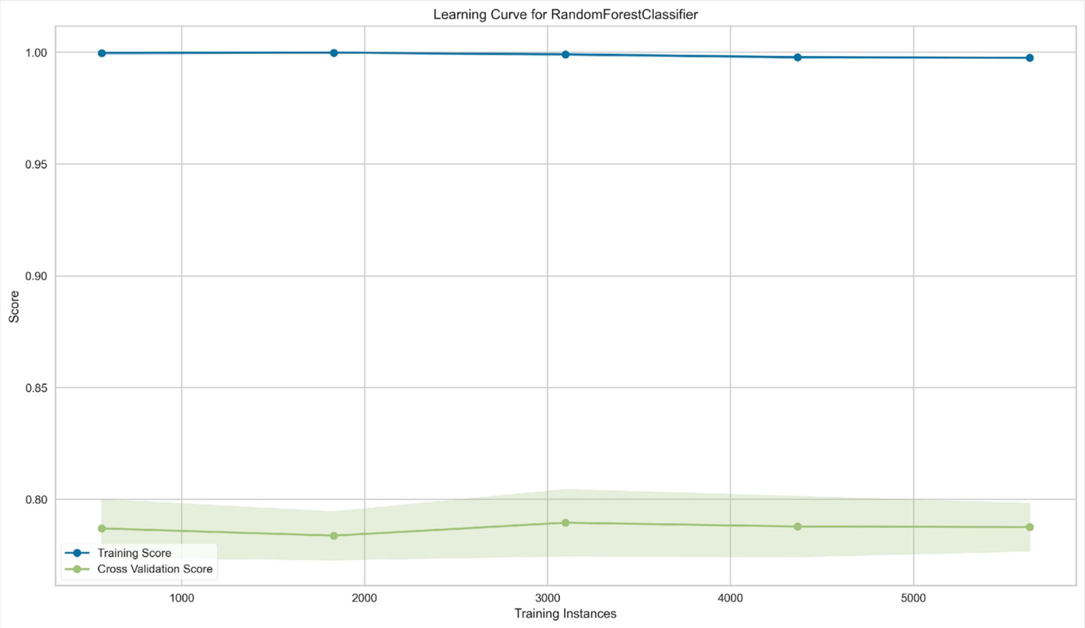
Score for BOTH are generally high

Good Fit Characteristics





Overfit Characteristics

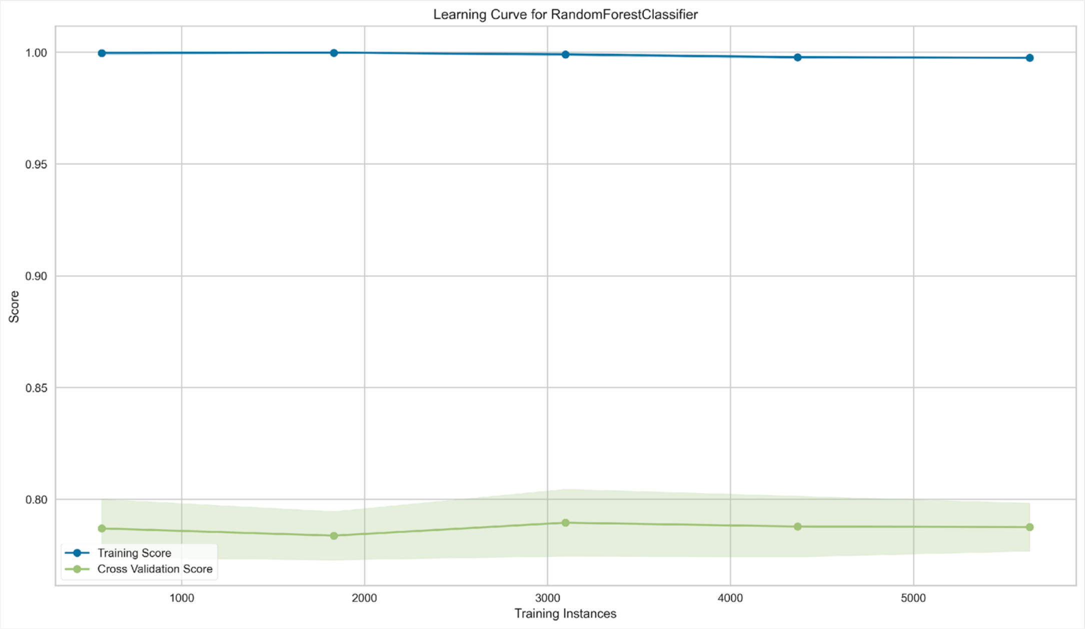


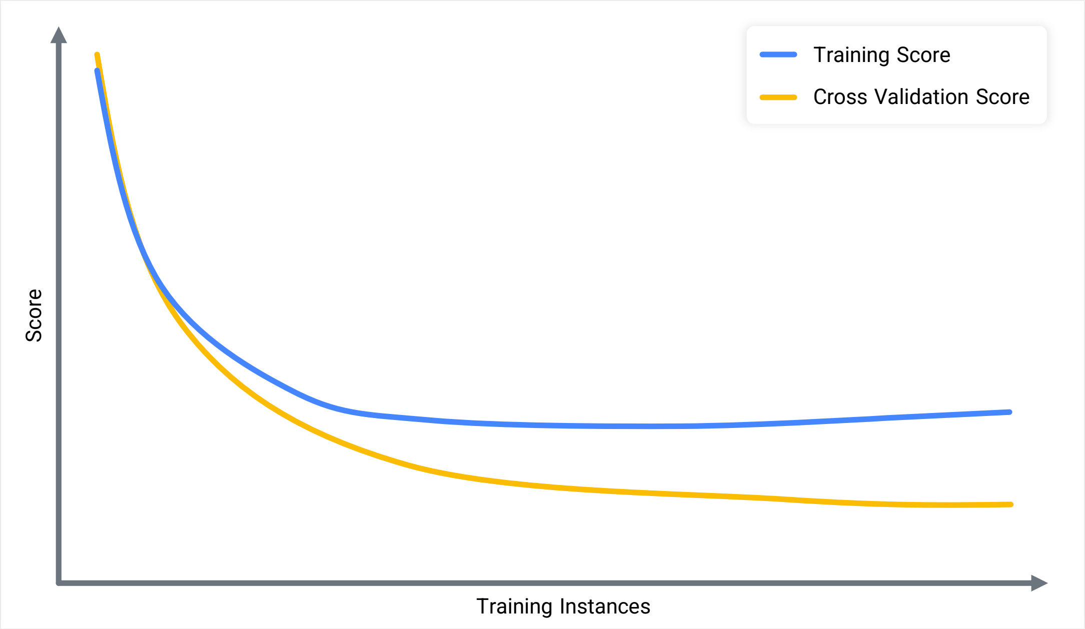
Lines are very far apart from one another

Lines would sometimes cross each other

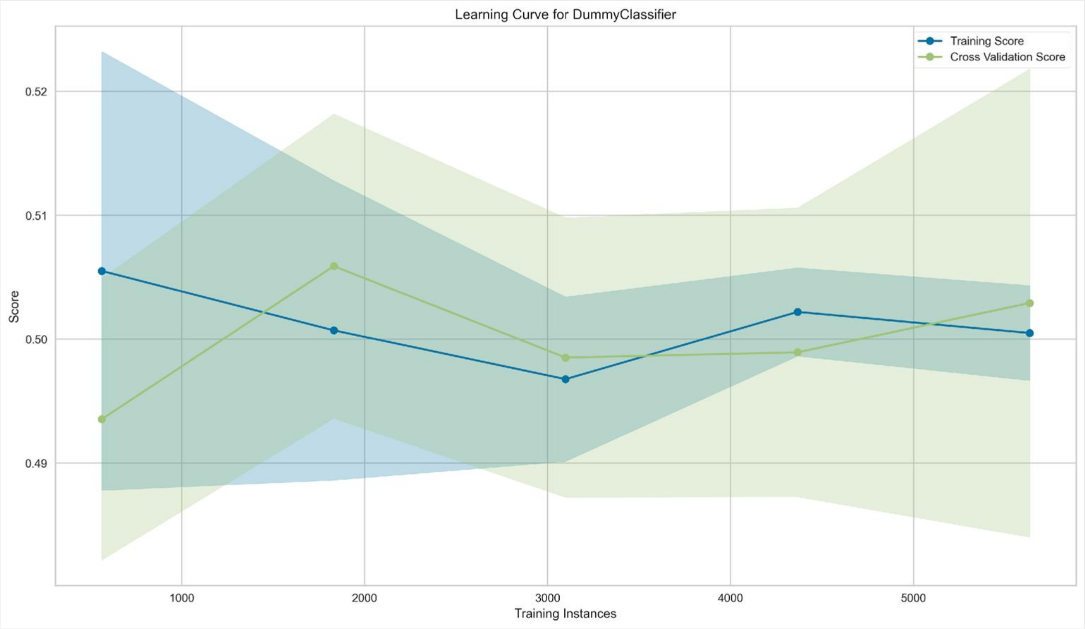
Scores for training set would be significantly higher than training set

Overfit Characteristics





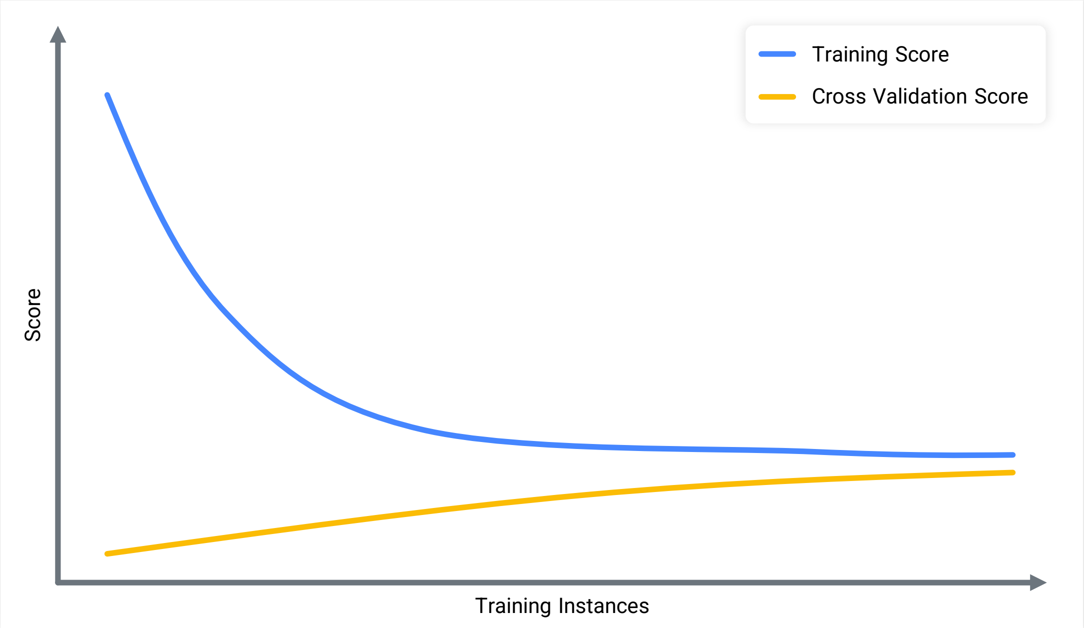
Underfit Characteristics



Lines are relatively close to each other

Scores for validation training set are both low

Knowledge Check



```
> This model has…

A.  High Bias, High Variance
B.  High Bias, Low Variance
C.  Low Bias, High Variance
D.  Low Bias, Low Variance
```

Practice Time!

5 Minutes

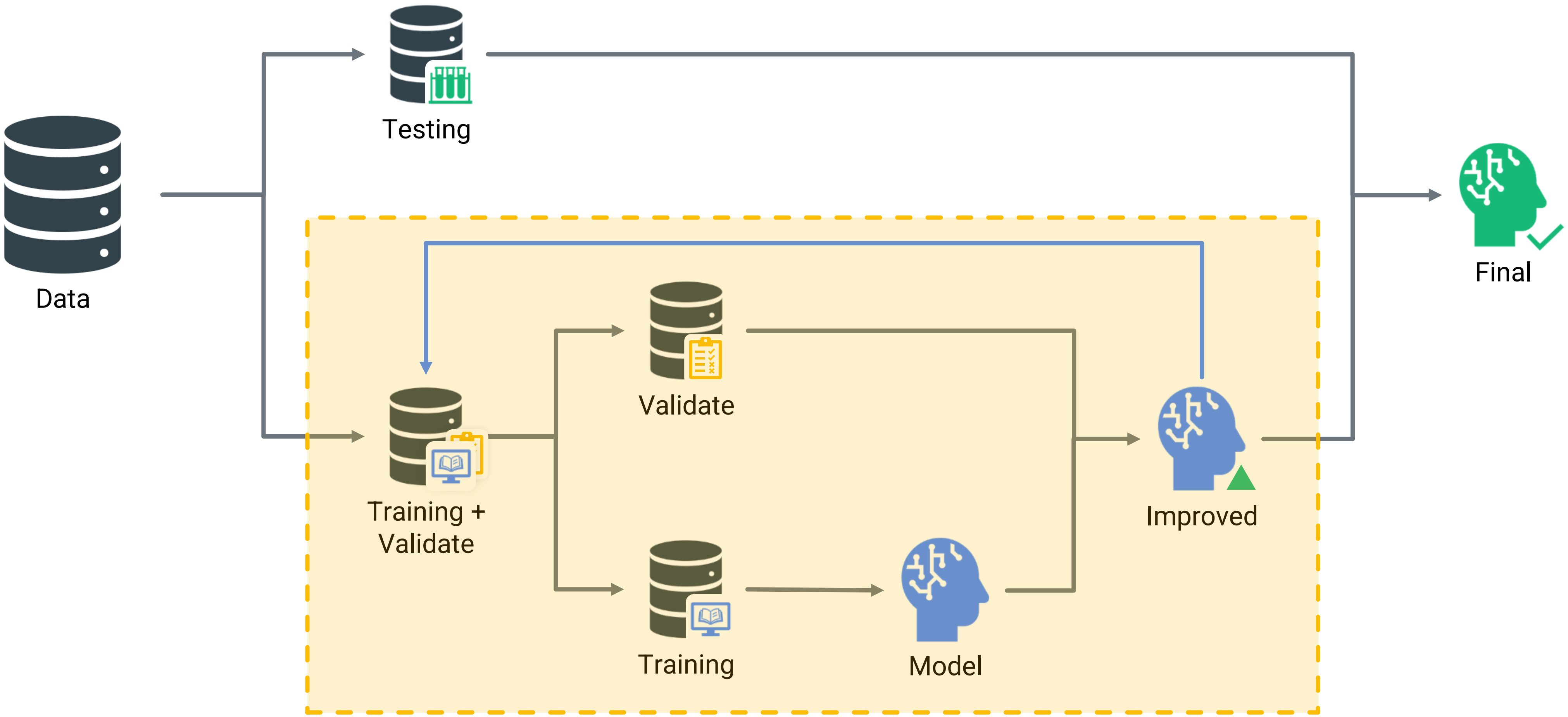
Please attempt exercise 2

We will go through the exercise later

Times up

We will now go through the exercises

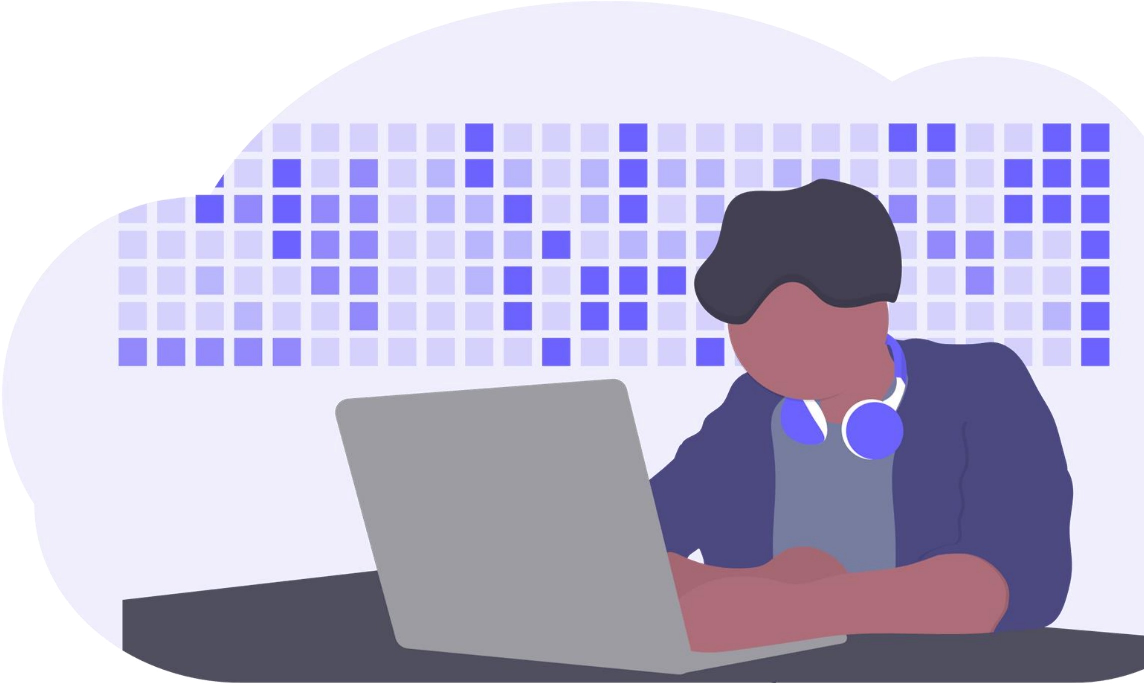
Actual Diagram

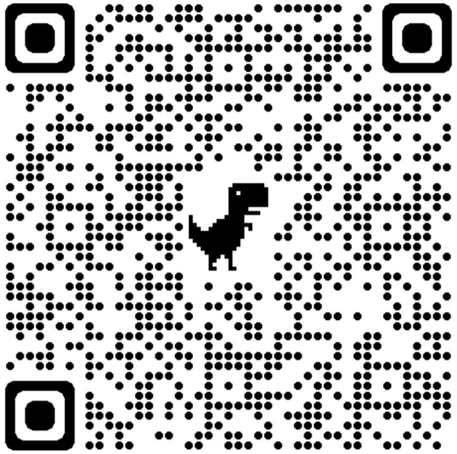


SPAI

An AI Singapore Student Chapter

Thank You





Scan the QR code to mark your attendance

Attendance