An AI Singapore Student Chapter

Advanced ML Workshop

Day 0

Scan the QR code to mark your attendance

Attendance

Learning Objectives

- Exploratory Data Analysis
- Data Preprocessing
- Model Building and Validation
- Overfitting and Underfitting

Prerequisites

- Everything here is just a quick recap of what was taught during the beginner machine learning bootcamp
- Only important details will be gone through today
- You should have some exposure or knowledge of concepts taught today

A computer program which learns from **experience** (E), with respect to some class of **task** (T) and **performance** (P) measure. If its performance at **tasks** in T, as measured by P, improves with **experience** (E). – *Tom Mitchell*

What is ML?

SPAI

What is Machine Learning?

Performance

Experience

Task

SPAI

Task

Classification: Classify a data point into a category

Regression: Predict a numerical value given some input

Performance

Classification: Accuracy etc.

Regression: Mean Squared Error etc.

Evaluated on test set to see if the algorithm can generalise well

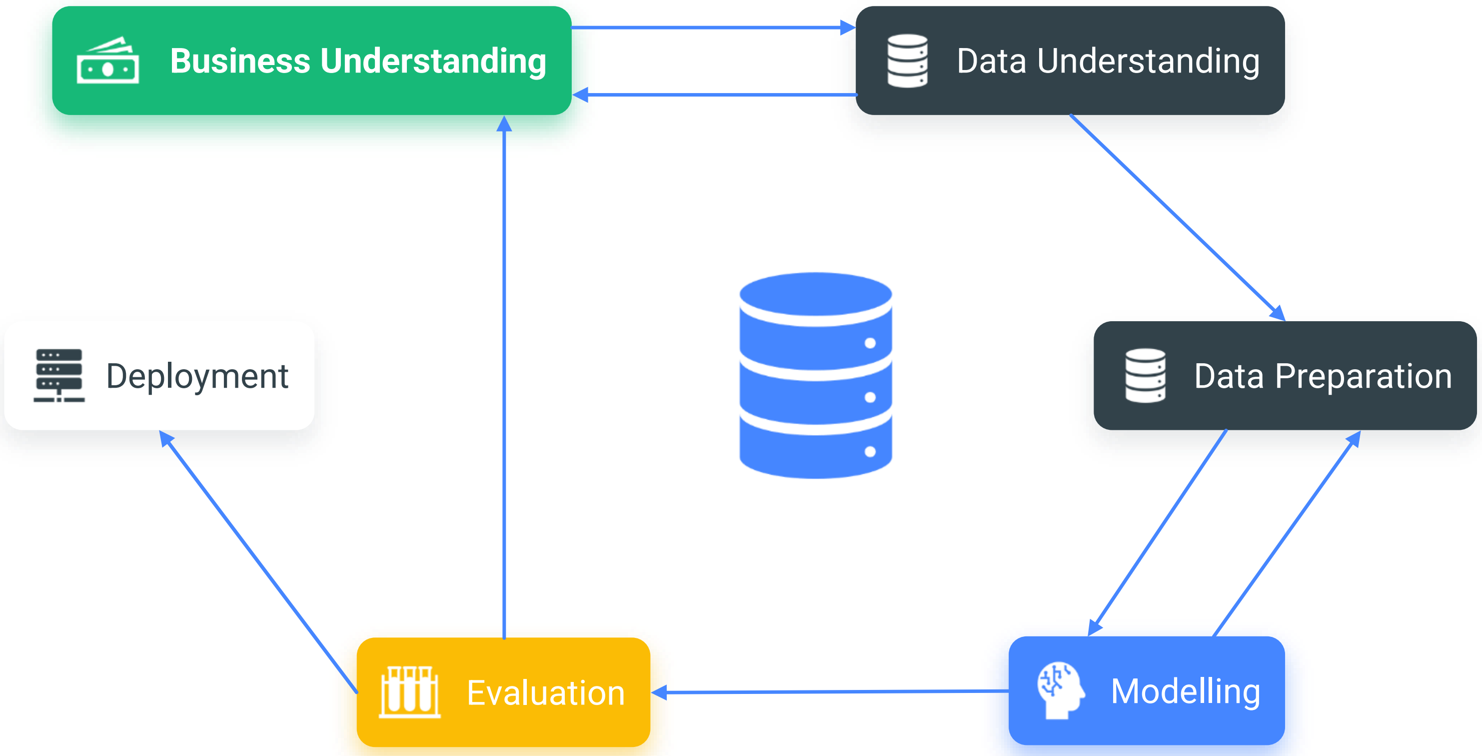
Experience

Algorithms are able to experience the dataset: Collection of many examples to learn from

Supervised learning: each example comes with an associated correct answer

E.g. Learning mathematics while being guided on what to do

ML Process



SPAI

Tools for Data Science

Python: Primary Programming language for Data Science

NumPy: Library for numerical computation

Pandas: Library for manipulation of tabular data

Matplotlib/Seaborn: Libraries for generating graphs

Scikit-Learn: Library for implementing classical ML

1

Exploratory Data Analysis

Loading Data

- Pandas used to load tabular datasets

```
data = pd.read_csv("melb_data.csv")
```

- Tabular datasets: Excel, CSV, SQL Database Tables

Practice Time!

Exercise 1

5 minutes

Practice Time!

Exercise 2

5 minutes

It allows us to understand the data and summarise it's main characteristics

Why EDA?

EDA helps check...

Patterns

Outliers

EDA

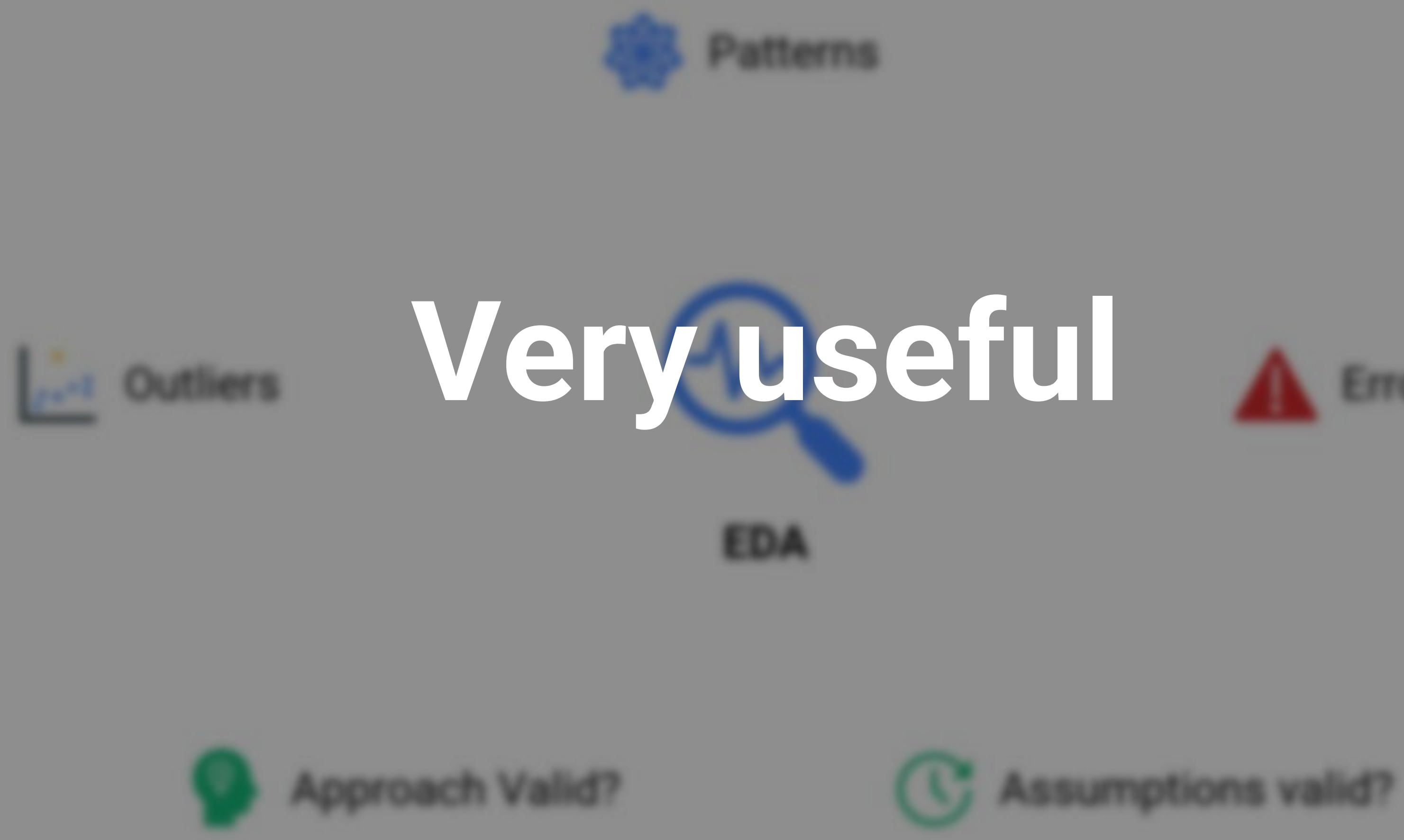
Errors

Approach Valid?

Assumptions valid?

EDA helps check…

Very useful



Steps for EDA

- Use seaborn pair plot to summarise data distribution and bi-variate relationships
- Compare how the distribution changes with different categorical levels

Knowledge Check

```
> Which of the following statements are NOT the motive of carrying out EDA

A. To detect potential errors in the dataset

A. To understand the semantic meaning behind each columns

A. To determine which model can give us the best performance

A. To determine what pre-processing steps could be undertaken before modelling
```

2

Data Pre-processing

If 80% of our work is data preparation, then ensuring data quality is the important work of the machine learning team

– Andrew Ng

SPAI

General Steps

Missing Values

Categorical Variables

Feature Scaling

Missing Values

Datasets often come with missing values

Most ML algorithms cannot handle missing values

Solution: Drop or Impute missing values

Dropping Missing Values

Pro: Simply and Easy to implement

Con: Leads to loss of info

```
df = df.dropna()
```

Imputing Missing Values

Pro: Distribution remains the same, with more training data

Con: In cases where data is missing systematically, imputation may affect relationships between variables

Numerical Data: Mean or Median
Categorical Data: Mode (most common value)

Practice Time!

Exercises 3-4

10 minutes

SPAI

Categorical Variables

ML algorithms only work with numerical data

How do we represent categorical data

Types: Ordinal and Nominal

Ordinal Data

- If data is already in numerical format, nothing needs to be done
- If data is a string, we use `OrdinalEncoder` to create a mapping between each category to a number

Nominal Data

- Replacing each category with a number does not preserve nominal nature of data
- One common method is to create separate binary variable for each possible categorical value
- Approaches: `pd.get_dummies()` or `OneHotEncoder`

Feature Scaling

ML perform best when data are equally scaled

We can convert numeric variables to similar scales

Approach: Z-Score Standardization (`StandardScaler`)

What is Z-Score

- It is calculated using the (feature – mean) / SD
- Tells us how far a data point is away from the mean

Practice Time!

Exercises 4-6

15 minutes

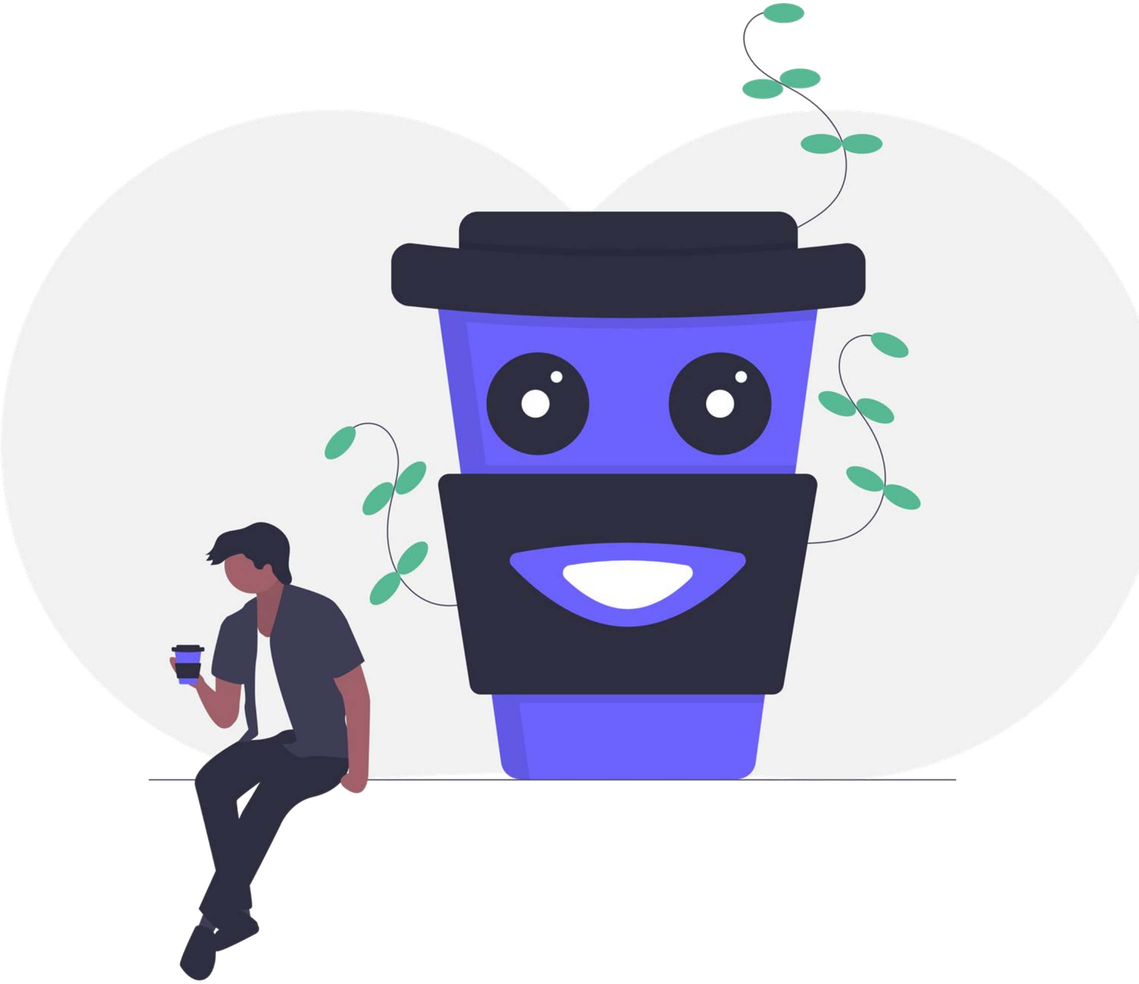
Knowledge Check

```
RangeIndex: 100 entries, 0 to 99
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   columnA  35 non-null     float64
 1   columnB  100 non-null    int64
 2   columnC  100 non-null    int64
dtypes: float64(1), int64(2)
memory usage: 2.5 KB
```

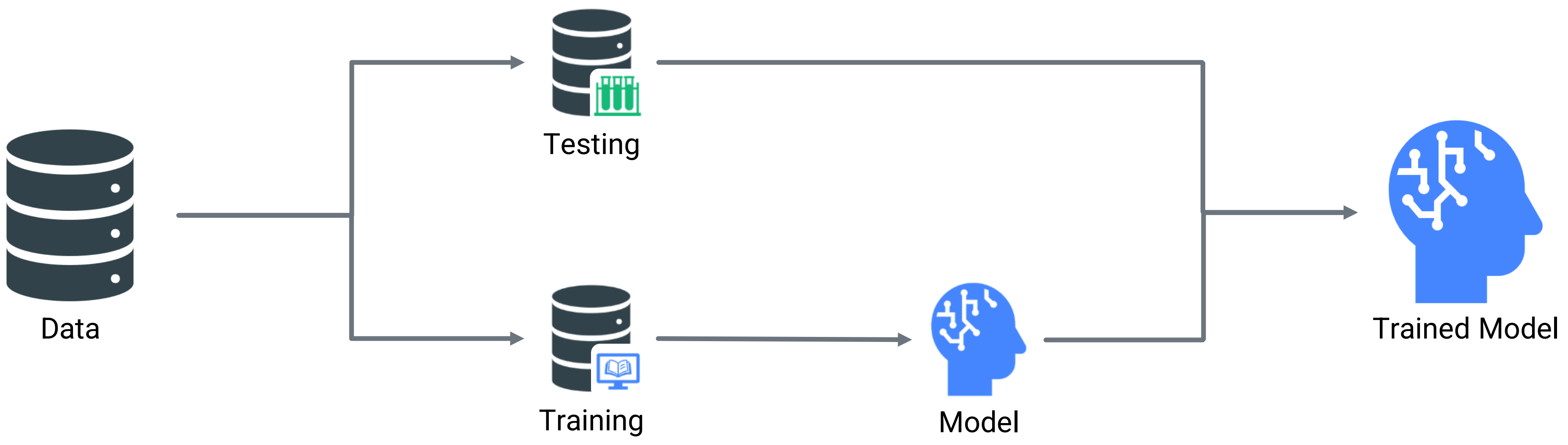
```
> Which of the following pre-processing steps should we take

A.  Standard Scaling - columnA

A.  One-Hot Encoding - columnA

A.  Impute with mean - columnA

A.  Drop - columnA
```

Break Time

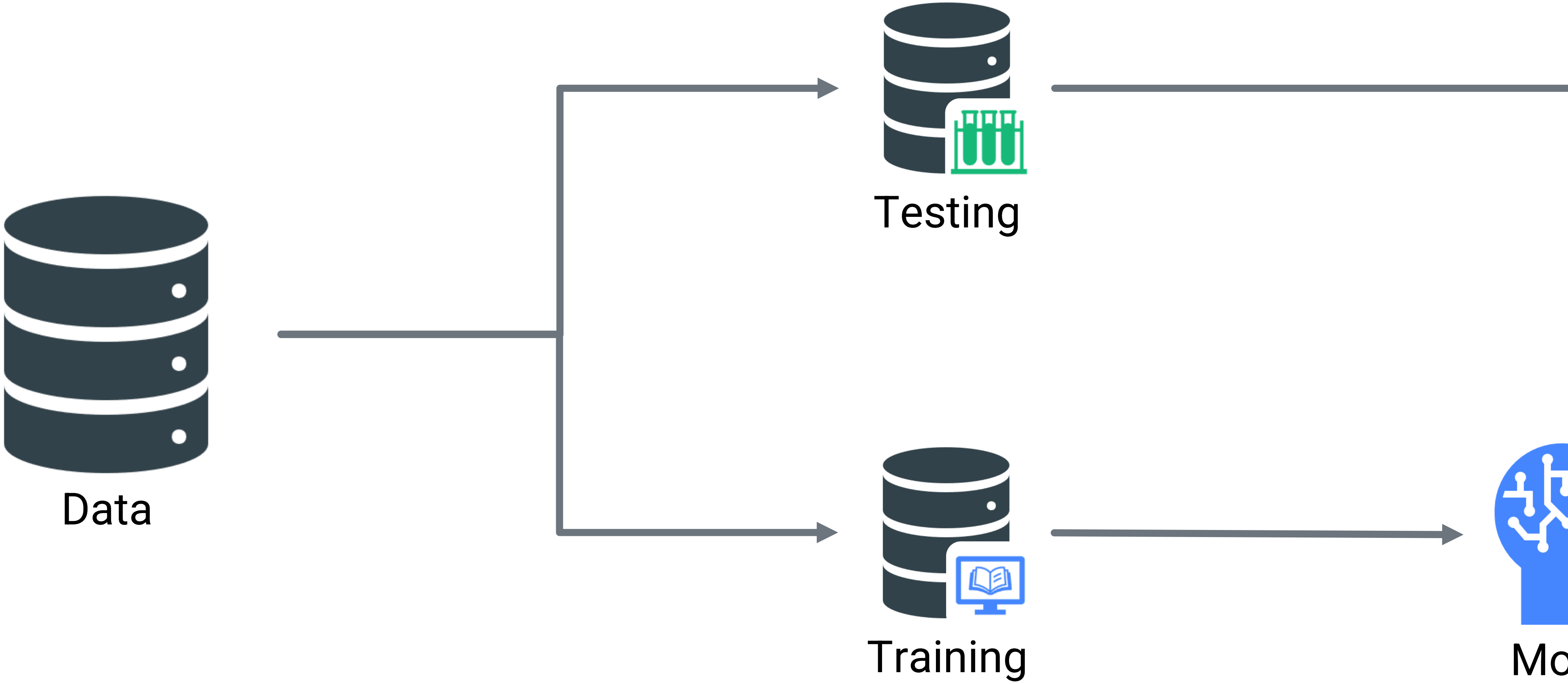
10 minutes



Model Building Process



Model Building Process



SPAI

We need an **unbiased** way to tell how well the model understands the task

Why split?

SPAI

Analogy

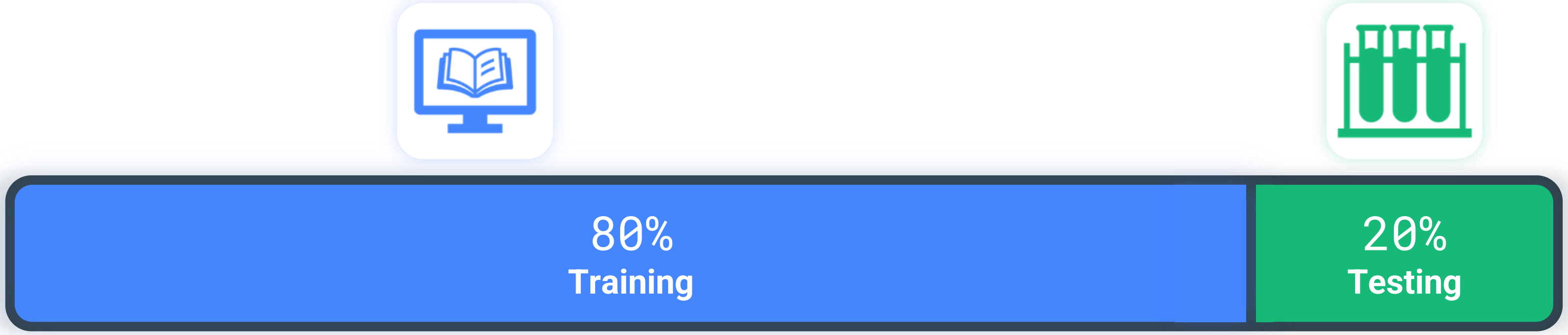
AI Model

- Seen data used to test model
- Model would score better

Student

- Mock paper given to students are tested in actual test
- Student would score better

What is Train Test Split?



Dataset

**70:30 ratio is used too*

Data leakage would result in unreliability in the test scores

Importance

Data Leakage

Data Leakage can occur when:

- Missing values are imputed with mean of Entire dataset before splitting
- Standardization is applied on the entire dataset before splitting

Analogy

AI Model

- Model learns key traits (mean) and patterns (standardization)

Student

- Students are given tips and hints before taking the test

This is why it is important to split your data before pre-processing it

SPAI

3

Model Building & Validation

Modelling Key Steps

1. Create

```
model = ModelName()
```

2. Fit

```
trained_model = model.fit(x_train, y_train)
```

3. Predict

```
y_pred = trained_model.predict(x_test)
```

4. Evaluate

```
score = metrics(y_pred, y_test)
```

ML Models

- Linear Model (e.g. Linear Regression, Logistic Regression
- Tree Based Models (e.g. Decision Trees)
- Nearest Neighbours Models (e.g. K-Nearest Neighbours)
- Support Vector Machines
- Ensemble Models (e.g. Random Forests)

Classification Metrics

- Commonly used metrics includes: Accuracy, precision, Recall and F1 Score
- Can all be described using a confusion matrix
- Metrics can all be extended for multi-class classification

Accuracy

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

Described as "How many were classified correctly"

Easy to understand and interpret

Potentially misleading when there is class imbalance

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Described as “How many were correctly Identified”

Tells us how confident we can be that our model prediction is correct

Does not tell us anything about how our model handles negative classes

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Described as “Out of all actual positive examples, how many were correctly identified”

Gives us an idea how good our model is at picking positive classes

Model predicting everything as positive would have a perfect recall

F1 Score

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Harmonic Mean of Precision and Recall

Useful in cases where there are imbalance classes

Can be misleading since is ignores true negatives

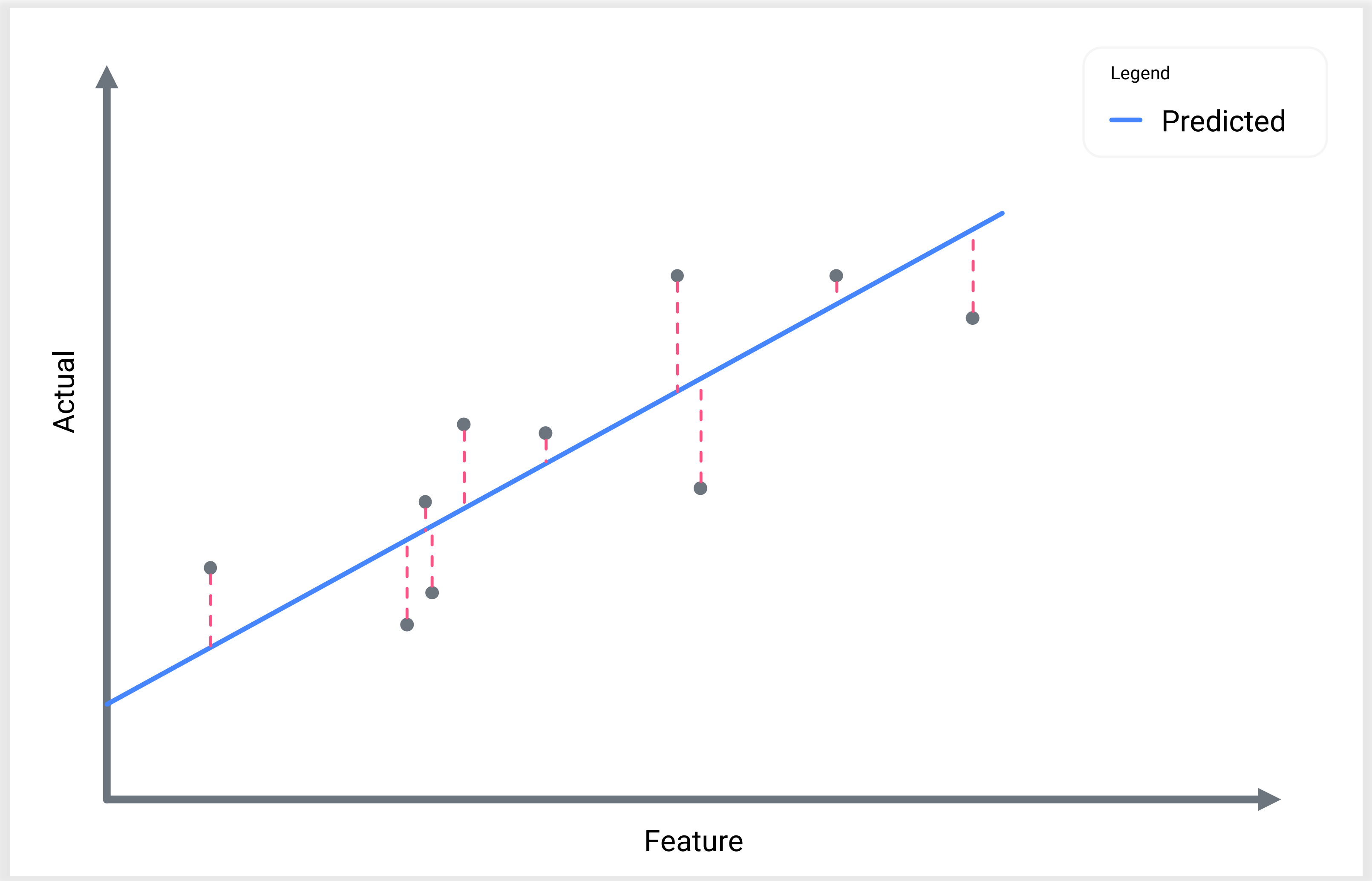
Gives equal importance to precision and recall

SPAI

Regression Metrics

- Regression evaluation is different from classification
- We measure a regression model performance based on error/residuals
- Error = Actual value – Predicted Value

Understanding Error



SPAI

Regression Metrics

Metric

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)

Formulae

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

SPAI

Practice Time!

Exercise 7

10 minutes

Conclusion

- Load data and perform EDA
- Split data to get an independent test set
- Perform data pre-processing
- Fit the model to training set and evaluate model using testing set

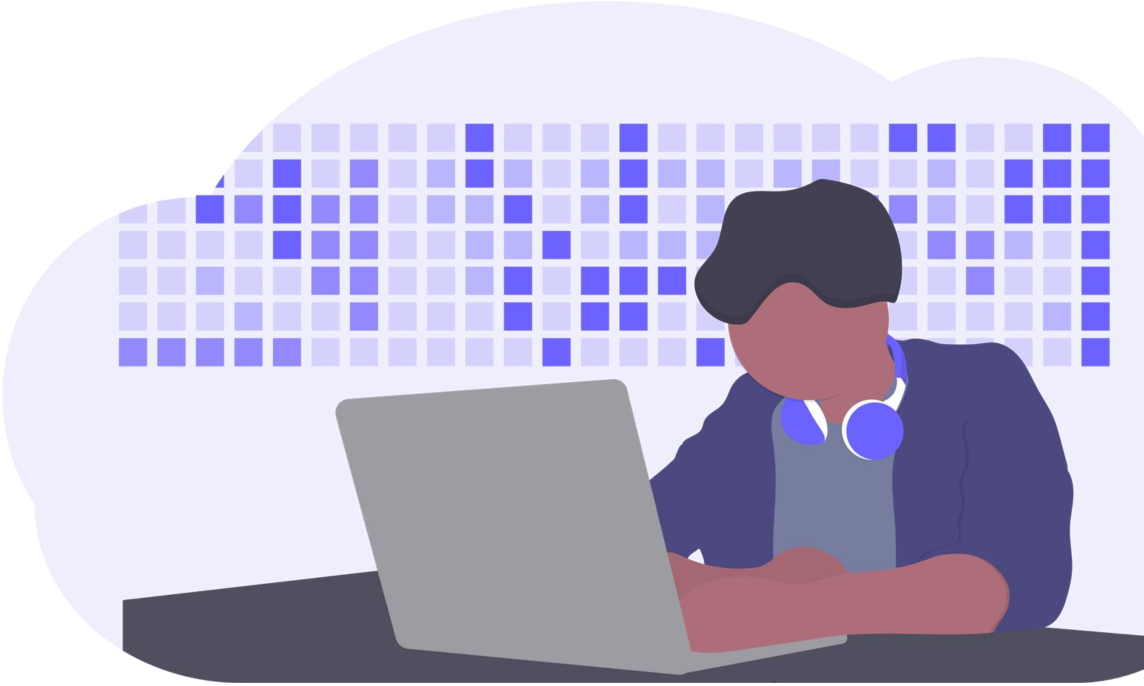
Knowledge Check

```
> Which of the following statements is True about Evaluation Metrics?

A.  You should pick whatever metrics that give you the best score

A.  You should pick your primary metrics before building your model

A.  You should always stick to Accuracy score for all classification problems

A.  You can use a Regression metrics for a Classification problem
```

SPAI

An AI Singapore Student Chapter

Thank You



Scan the QR code to mark your attendance

Attendance