Python for Data Analysis
Introduction to Modeling in Python
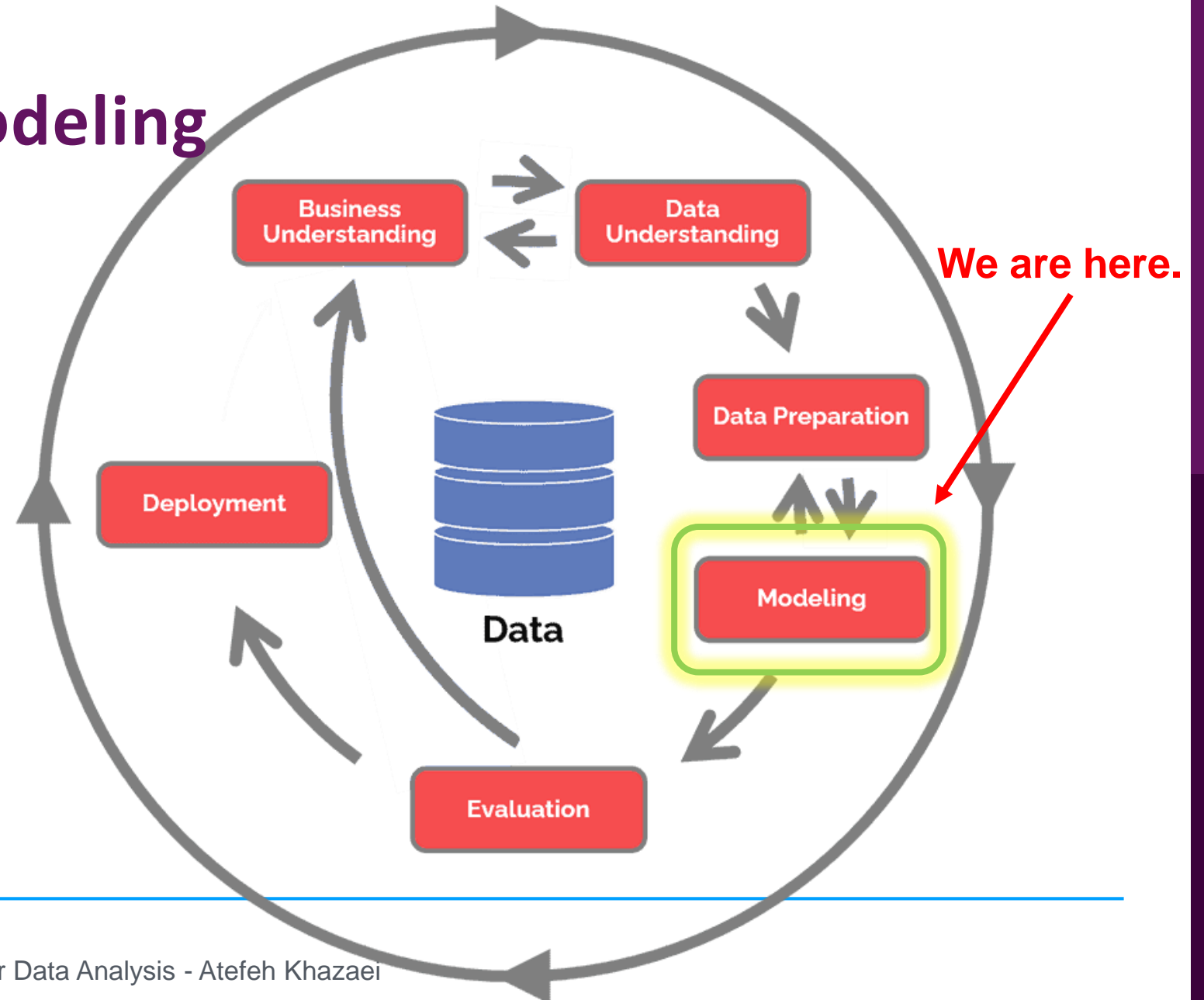
(Week 10)

Atefeh Khazaei

atefeh.khazaei@port.ac.uk
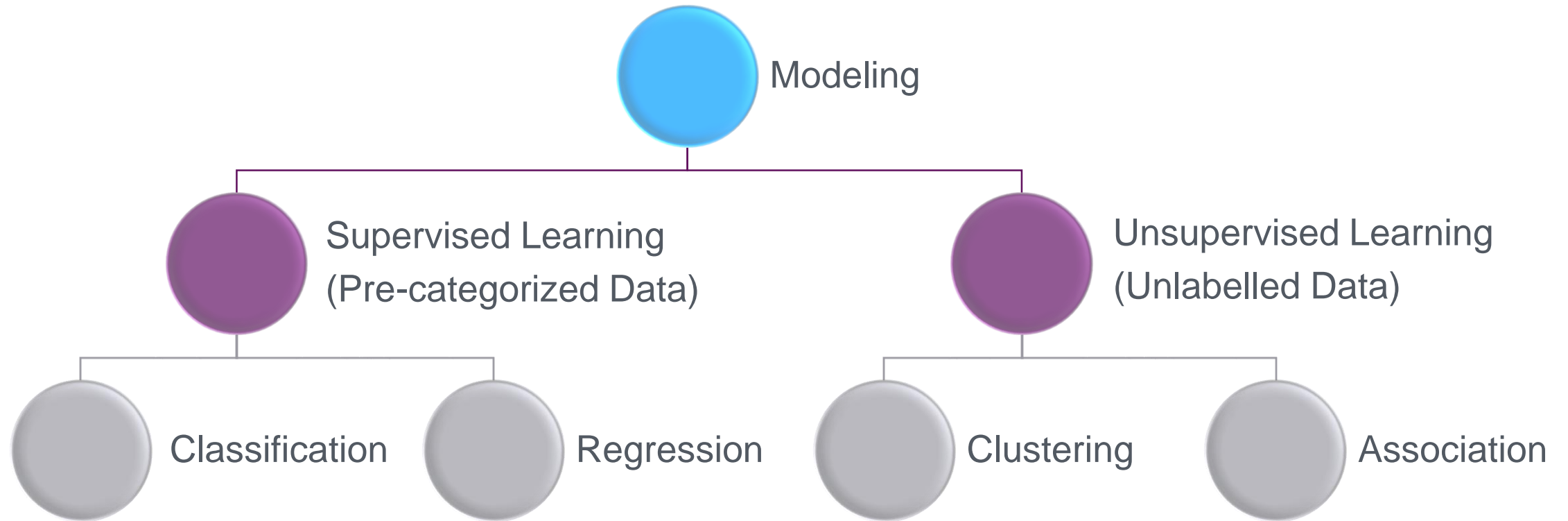
What we will learn this week?

❑ Supervised vs. Unsupervised Learning

❑ Classification vs. Regression

❑ Introduction to Modelling and Evaluation

CRISP-DM
Introduction to Modeling

We are here.

UNIVERSITY OF PORTSMOUTH

Supervised vs. Unsupervised Learning

UNIVERSITY OF PORTSMOUTH

Classification vs. Regression

- A marketing manager would like to know whether a given costumer will buy a product or not

- A marketing manager would like to predict how much a given costumer will spend during a sale
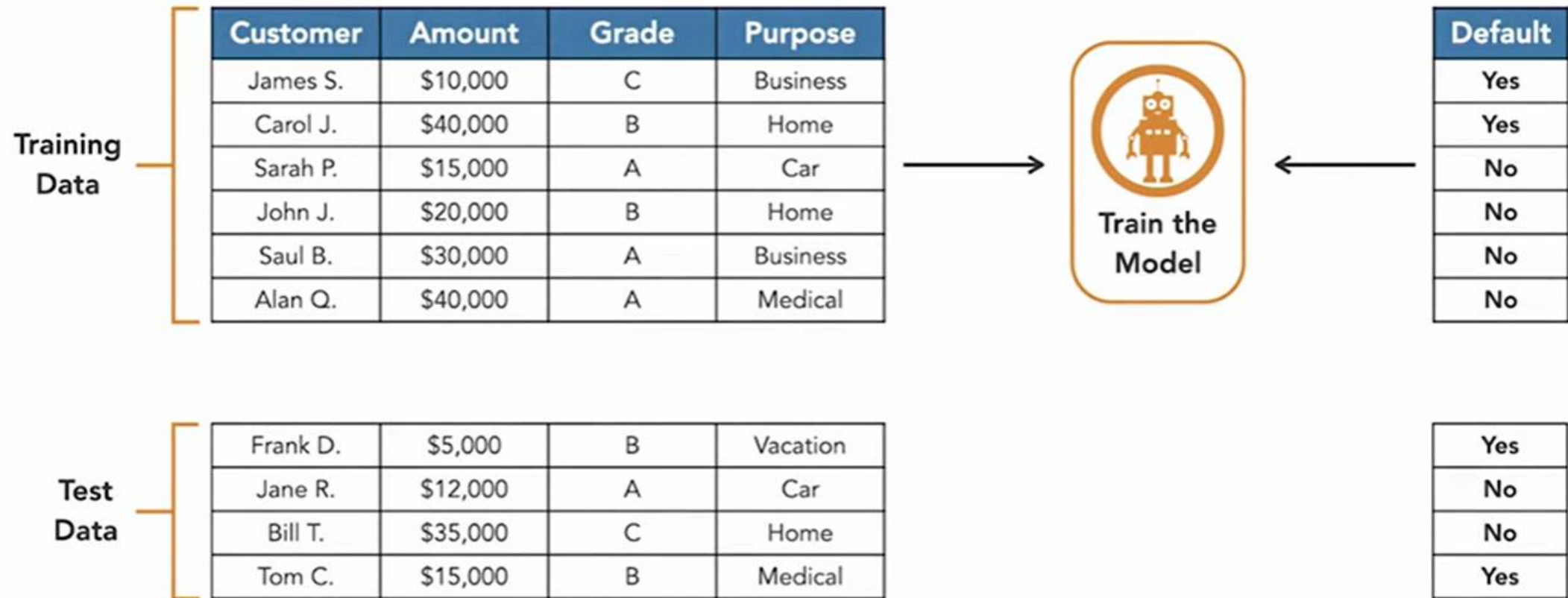
UNIVERSITY OF PORTSMOUTH
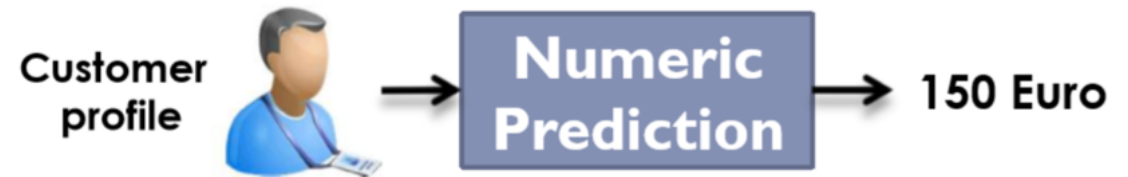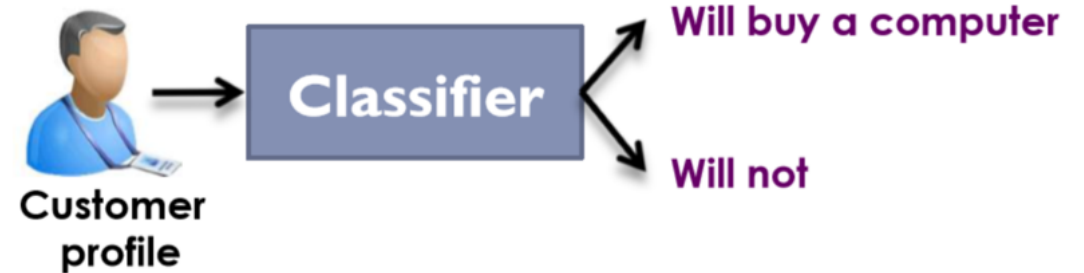
Our Plan

❑ Today and Next Week:

   ❑ An **introduction** to Supervised Learning (Linear Regression Alg.) and Evaluation

❑ Next Teaching Block:

   ❑ Supervised Modeling Algorithms in Python

      ❑ KNN

      ❑ Decision Tree

      ❑ Random Forest

      ❑ Linear Regression

   ❑ Clustering in Python

   ❑ Model Evaluation in Python

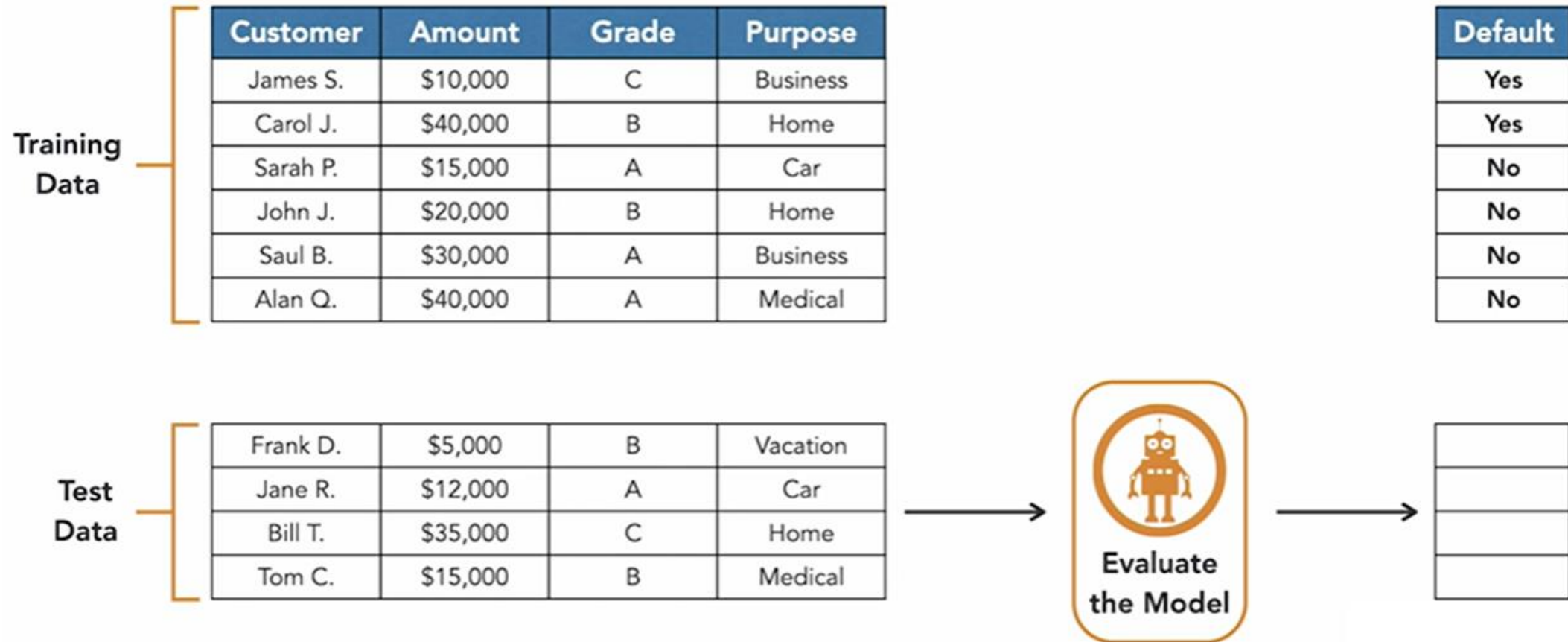# Introduction to Linear Regression

❑ In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables.

❑ The model is a formula like this:

$$❑ \quad Z = w_0 + w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$$

❑ In other words, the goal is finding the best values for *Ws*.

❑ *Xs* are explanatory variables (features / attributes)

Introduction to Evaluation

❑ The Goal: Assess how well our model works.

Customer	Amount	Grade	Purpose		Default
James S.	$10,000	C	Business		Yes
Carol J.	$40,000	B	Home		Yes
Sarah P.	$15,000	A	Car		No
John J.	$20,000	B	Home		No
Saul B.	$30,000	A	Business		No
Alan Q.	$40,000	A	Medical		No

Training Data → Train the Model ←

Frank D.	$5,000	B	Vacation		Yes
Jane R.	$12,000	A	Car		No
Bill T.	$35,000	C	Home		No
Tom C.	$15,000	B	Medical		Yes

Test Data

Introduction to Evaluation (Cont.)

	Customer	Amount	Grade	Purpose		Default
Training Data	James S.	$10,000	C	Business		Yes
	Carol J.	$40,000	B	Home		Yes
	Sarah P.	$15,000	A	Car		No
	John J.	$20,000	B	Home		No
	Saul B.	$30,000	A	Business		No
	Alan Q.	$40,000	A	Medical		No

	Customer	Amount	Grade	Purpose
Test Data	Frank D.	$5,000	B	Vacation
	Jane R.	$12,000	A	Car
	Bill T.	$35,000	C	Home
	Tom C.	$15,000	B	Medical

Evaluate the Model
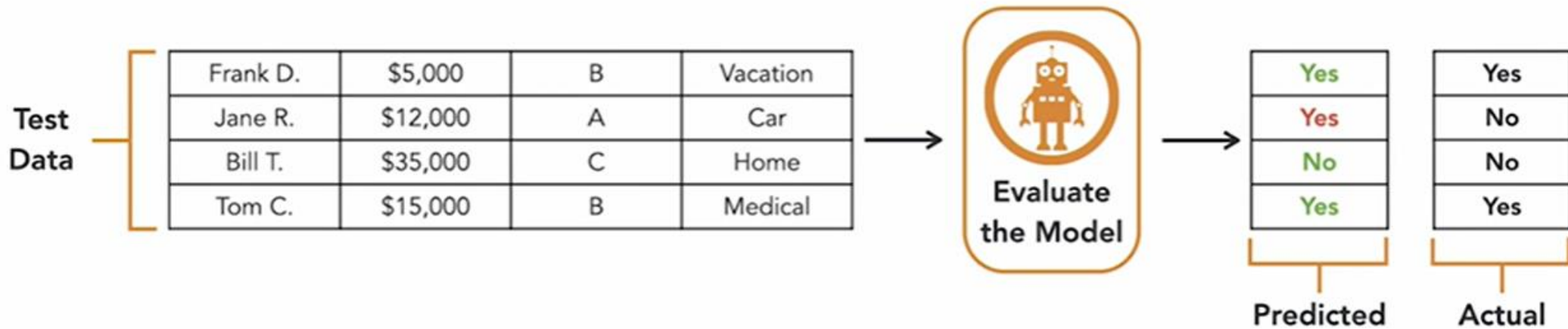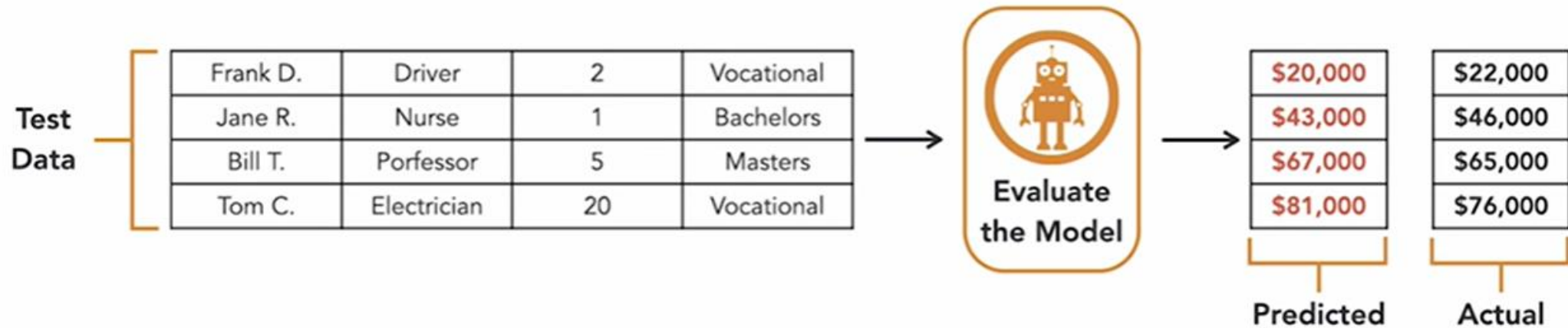
Introduction to Evaluation (Cont.)
Accuracy

$$\text{Predictive Accuracy} = \frac{\text{Correct Predictions}}{\text{Number of Test Instances}} = \frac{3}{4} = 0.75$$

Introduction to Evaluation (Cont.)
Mean Absolute Error (MAE)

$$MAE = \frac{\sum | \, Predicted - Actual \, |}{Number\ of\ Test\ Instances} = \frac{2000 + 3000 + 2000 + 5000}{4} = 3000$$

A Simple Modeling Example

❑ Week10_A Simple Model.ipynb

References & More Resources

❑ References:

   ❑ McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2012.

❑ More Resources:

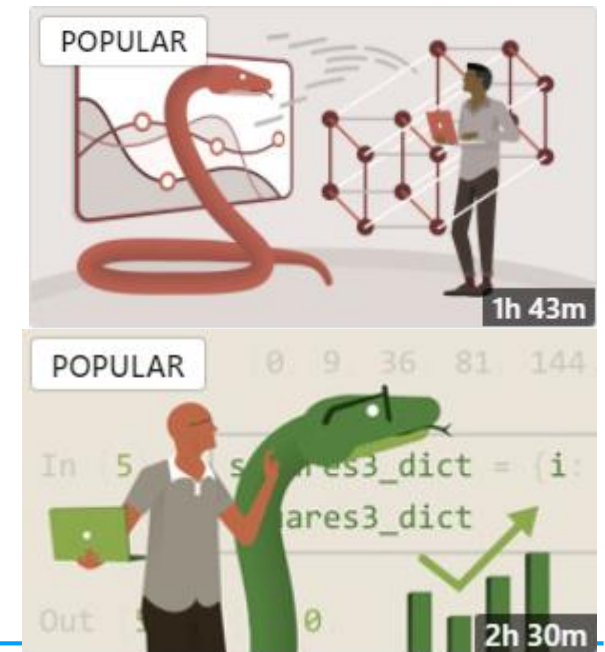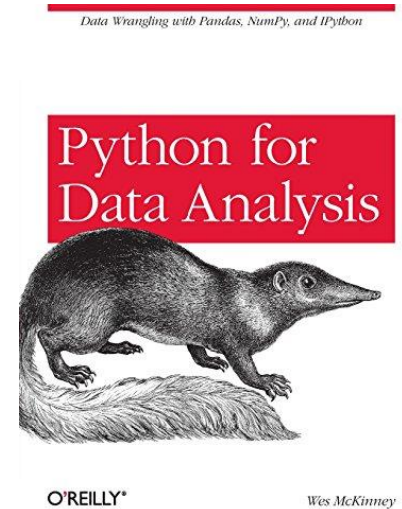   ❑ Machine Learning with Python: Foundations:

https://www.linkedin.com/learning/machine-learning-with-python-foundations

   ❑ Python Data Analysis on Linkedin Learning:

https://www.linkedin.com/learning/python-data-analysis-2

      ❑ To use Linkedin Learning, you can log in with your university account:

        https://myport.port.ac.uk/study-skills/linkedin-learning

Practical Session

❑ Please download Week10_A-Simple-Model.ipynb file, and run it to learn new points.

❑ Please read the practical sheet (Week10_Practicals.pdf) and do the exercise.