THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society

CDCS TRAINING PROGRAMME

AN INTRODUCTION TO MACHINE LEARNING.

ARRANGEMENTS FOR THE COURSE

Time	Session 1 Monday 14th April	Session 2 Monday 21st April
Topic A	Introduction to Machine Learning and Data Exploration	More Classification Models (Decision Trees and k-NN)
Topic B	Classification Basics and Logistic Regression	Regression and Practical Considerations in ML

WHO AM I?

Chris Oldnall

SESSIONS THROUGHOUT THE COURSE

Wrap-Up

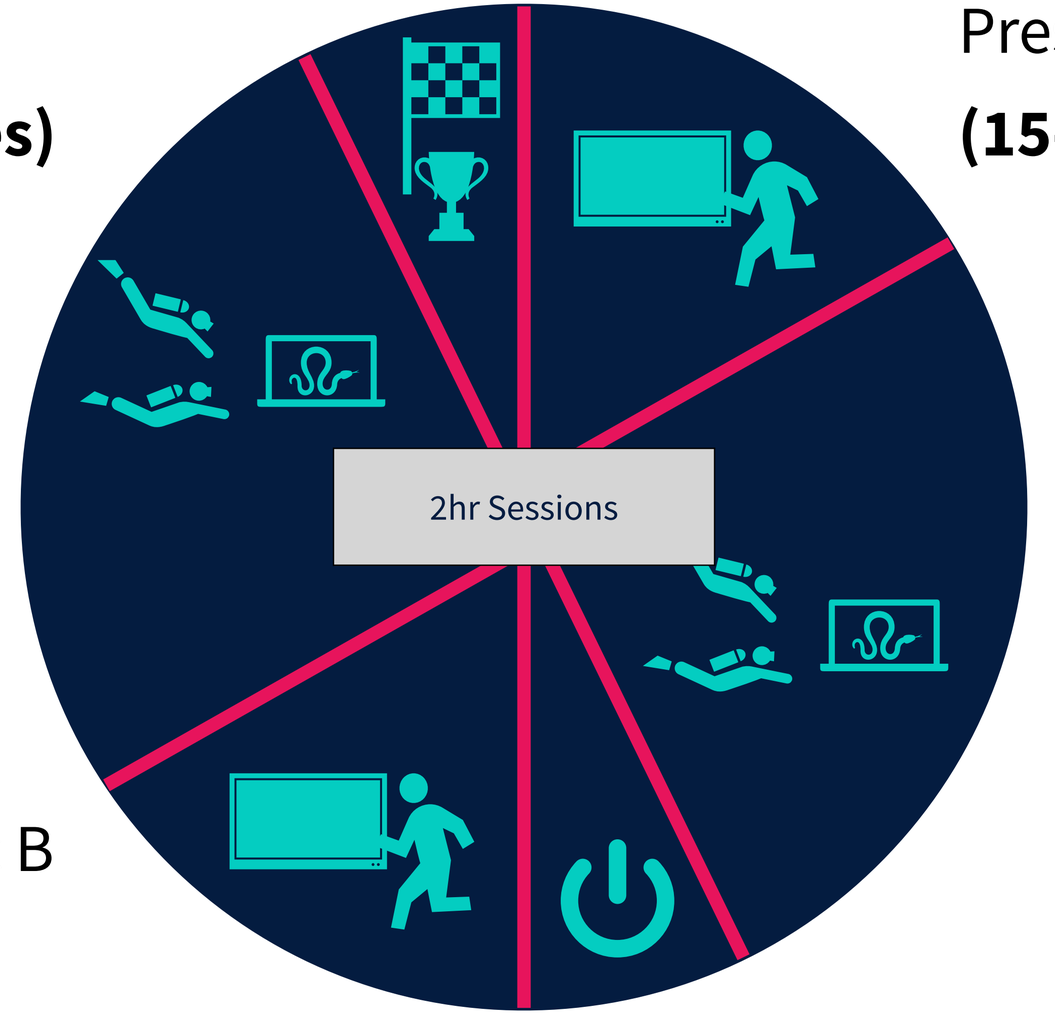
(10 minutes)

Presentation of Topic A

(15-20 minutes)

Paired Programming Practice B

(25-30 minutes)



Paired Programming Practice A

(25-30 minutes)

Presentation of Topic B

(15-20 minutes)

Mini-Break

(10 minutes)

PAIRED PROGRAMMING

DATA CULTURE SOCIETY
CDCS.ED.AC.UK

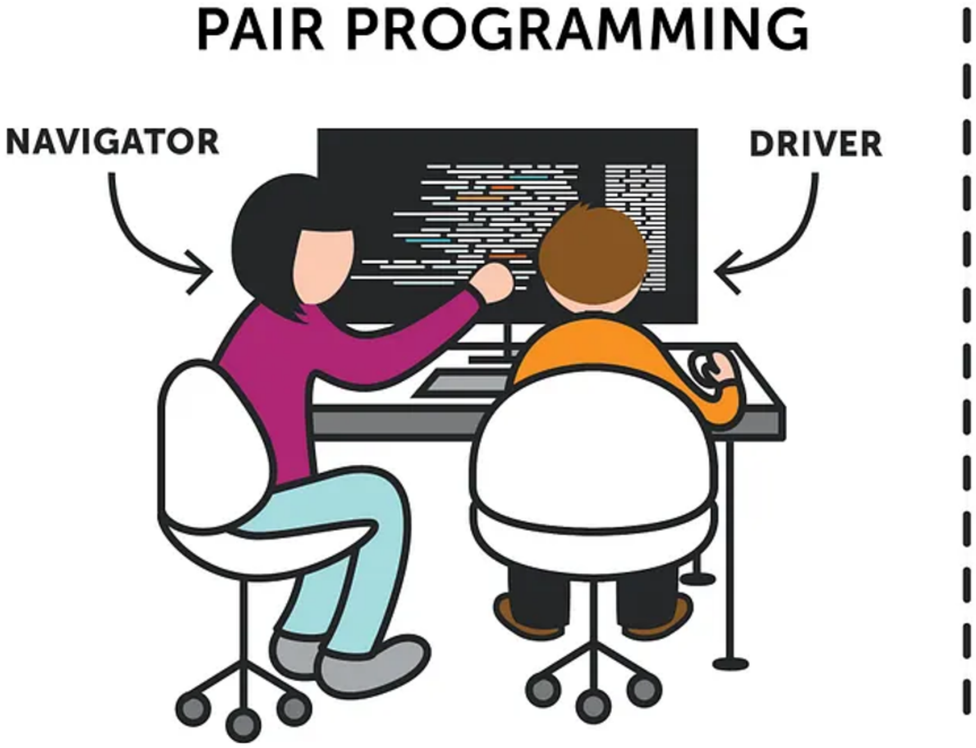


Image from https://medium.com/@tomspencer_uk/pair-programming-and-problem-solving-4531ef3bf171

OTHER THINGS YOU WILL SEE THROUGHOUT THE COURSE

DEMONSTRATIONS

Sometimes you might see the typewriter symbol. This means we are going to demonstrate something in Python/Noteable.

Bear with us if it takes a moment to switch windows.

```
variable_name = sensible
print(variable_name)

“sensible”
```

CODE CHUNK TEXT

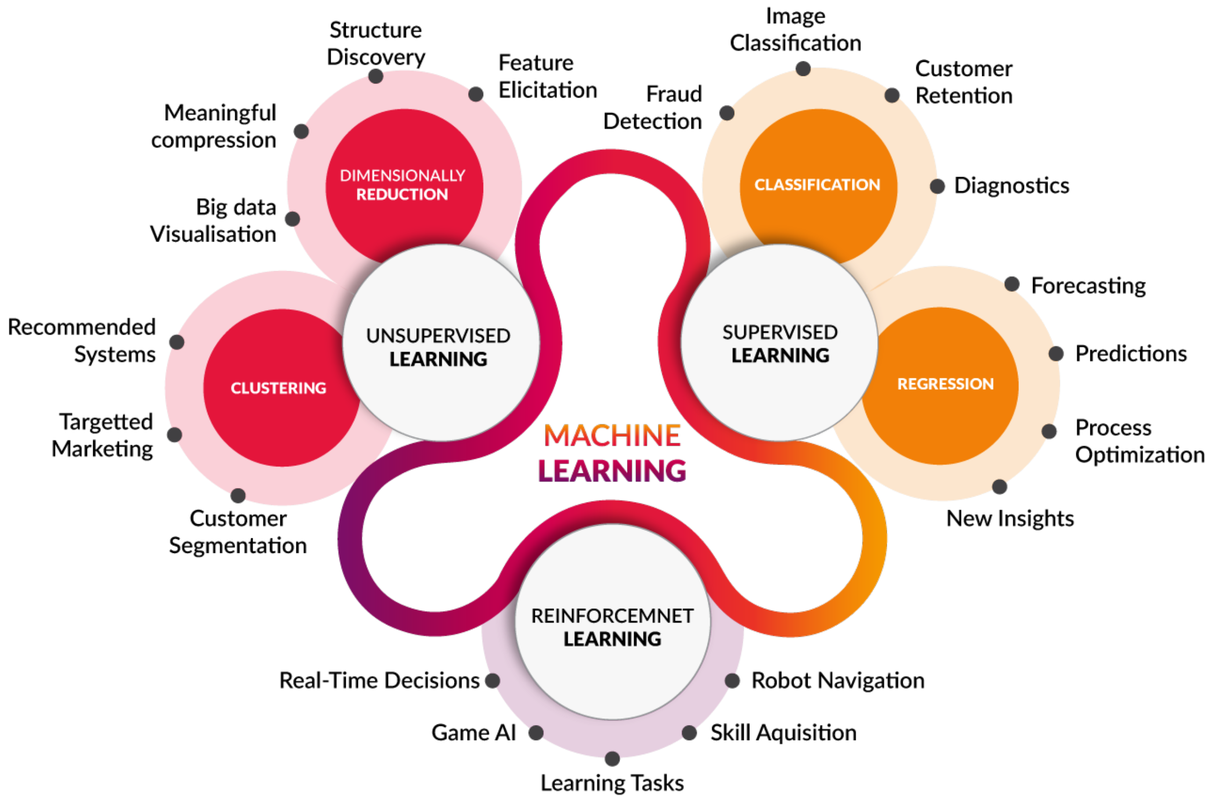
In the slides we may see text which is ‘pink’ in colour and a different font. This is to indicate it is a chunk of text, written in Python. The colour/font don’t matter just noticing it is code is important!

INTRODUCTION TO MACHINE LEARNING AND DATA EXPLORATION

Machine learning = Giving computers the ability to learn from data and make decisions without being explicitly programmed

To do this, we speak in a language they understand,
e.g. Python or R

BRANCHES OF MACHINE LEARNING

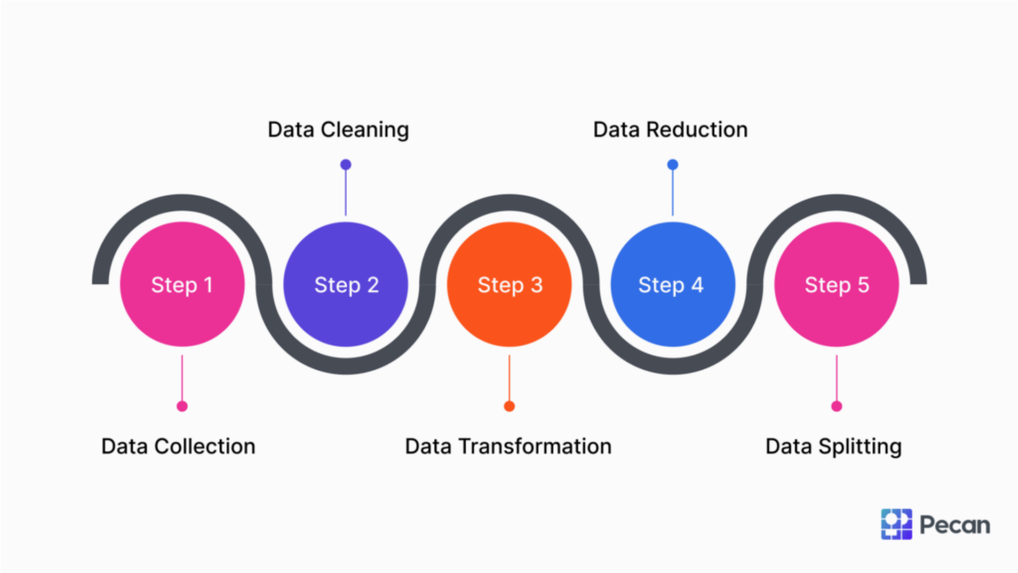


There are different types of machine learning branches and they are used for a range of different tasks in different areas.

THE ML PIPELINE

1. Ask a question
2. Get the data
3. Explore and clean the data
4. Choose a model
5. Train the model
6. Evaluate
7. Deploy & monitor

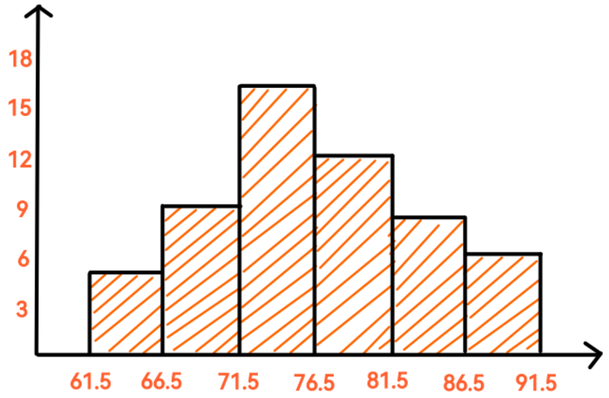
WHAT IS THE GOAL OF DATA PREPROCESSING?



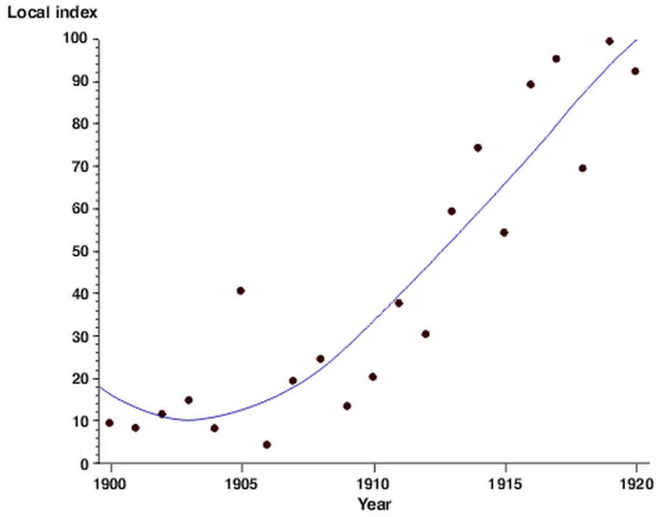
- Understand structure
- Spot issues
- Find patterns

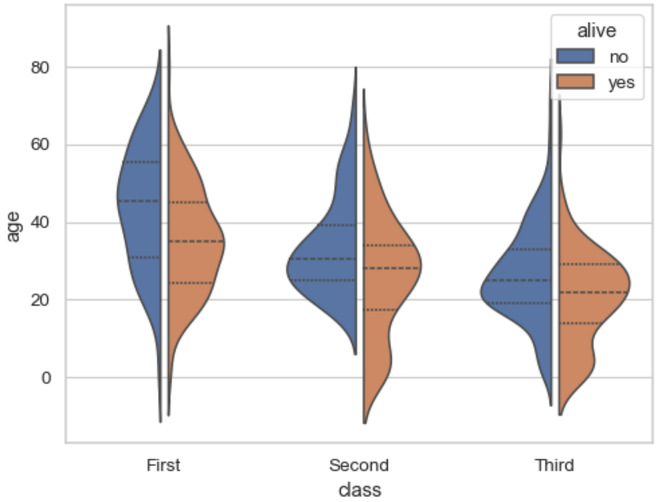
VISUAL EXPLORATION

Scatter plots for relationships



Histograms for distributions





Box plots for spotting outliers

WHY DOES DATA EXPLORATION MATTER?

- Prevents bad models
- Helps choose the right algorithm
- Improves interpretability

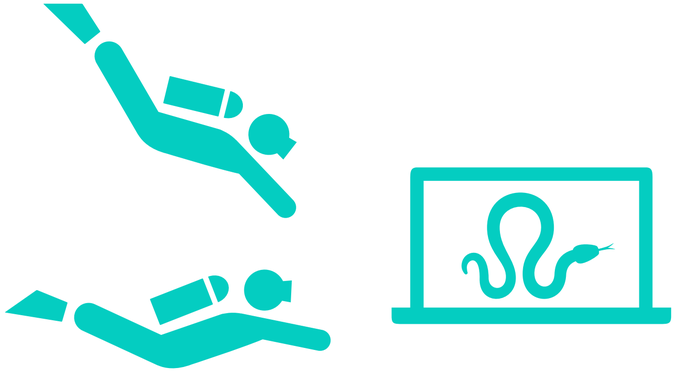
“Garbage in, garbage out.”

WHICH DATA WILL WE LOOK AT?

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

LETS GET PROGRAMMING

Session 1a



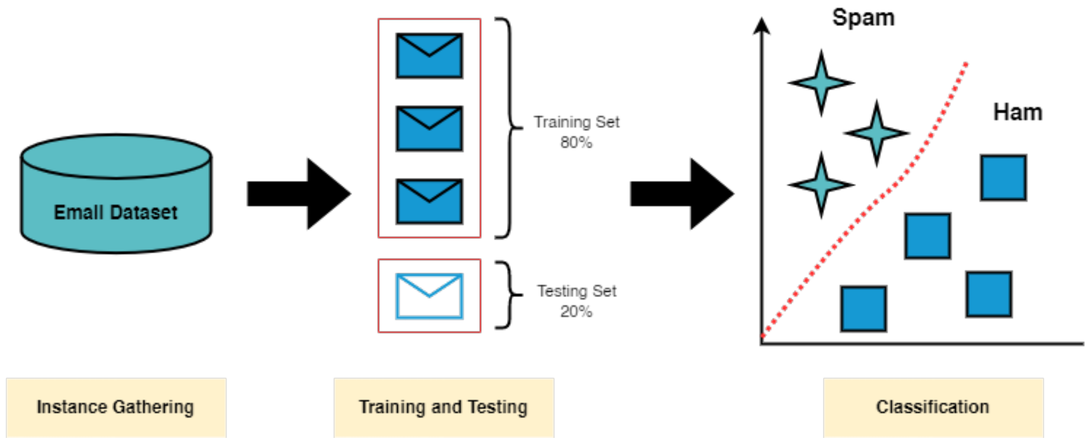
DATA CULTURE SOCIETY
CDCS.ED.AC.UK

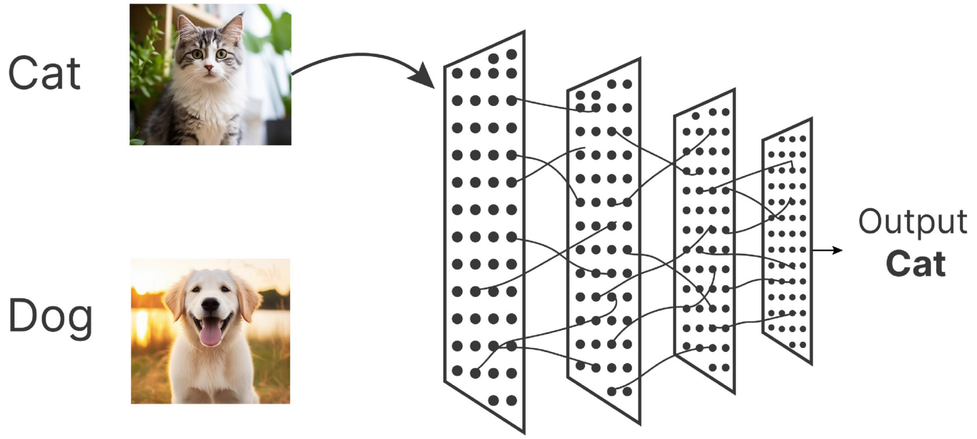
CLASSIFICATION BASICS AND LOGISTIC REGRESSION

WHAT IS CLASSIFICATION?

A supervised learning task

Goal: predict a category/label





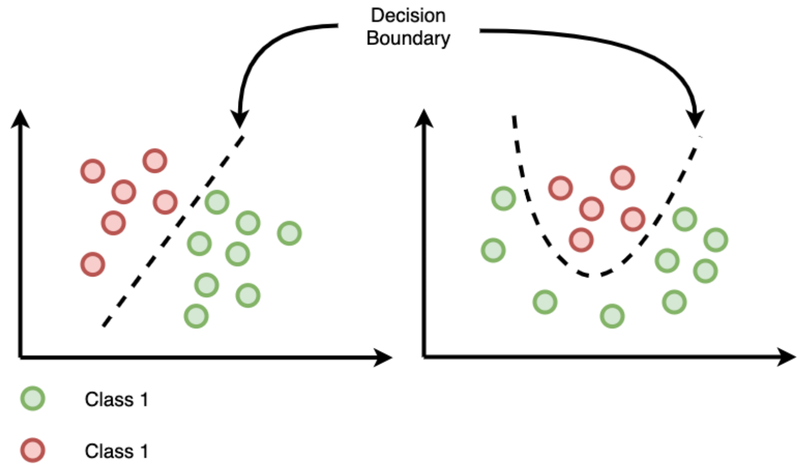
CLASSIFICATION VS REGRESSION

Feature	Classification	Regression
Output type	Categories (labels)	Continuous numbers
Example	Spam vs Not Spam	House price prediction

HOW DOES A CLASSIFIER WORK?

Looks for patterns in labelled data

Finds boundaries between classes



Common classifiers:

Logistic Regression

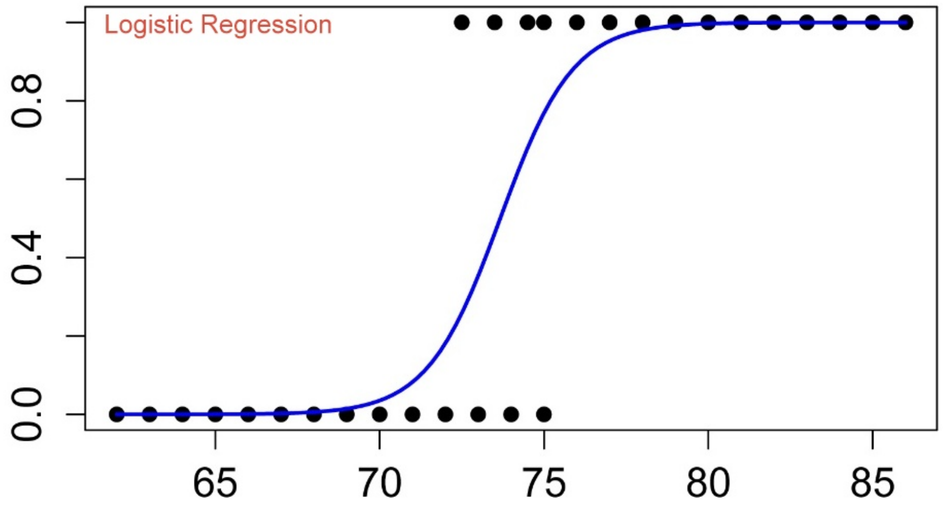
Decision Trees

k-NN

SVM

Neural Networks (deep learning)

INTRODUCING LOGISTIC REGRESSION



Used for binary classification

Predicts probability between 0 and 1

Based on the logistic (sigmoid) function

S-shape curve

Compresses values to [0, 1]

Probability threshold typically set at 0.5

HOW DOES A LOGISTIC REGRESSION MAKE A DECISION?

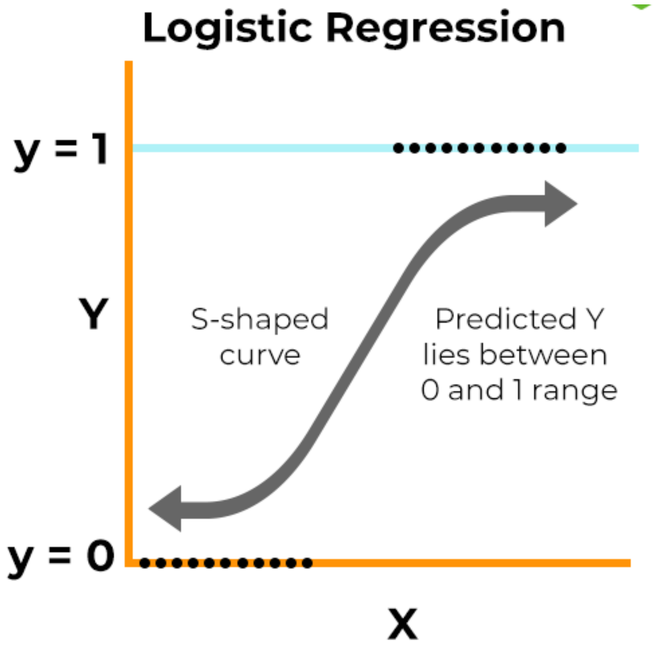
Finds the best-fit line/hyperplane

Predicts probability of outcome = 1

If…

Probability > 0.5 → Class 1

else → Class 0



HOW DO WE KNOW IF A CLASSIFIER DID WELL?

Accuracy: The proportion of **total predictions** the model got right.

(TP + TN) / total

Precision: The proportion of **positive predictions** that were actually correct.

TP / (TP + FP)

Recall: The proportion of **actual positives** that the model correctly identified.

TP / (TP + FN)

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

PROS AND CONS OF LOGISTIC REGRESSION

Fast

Interpretable

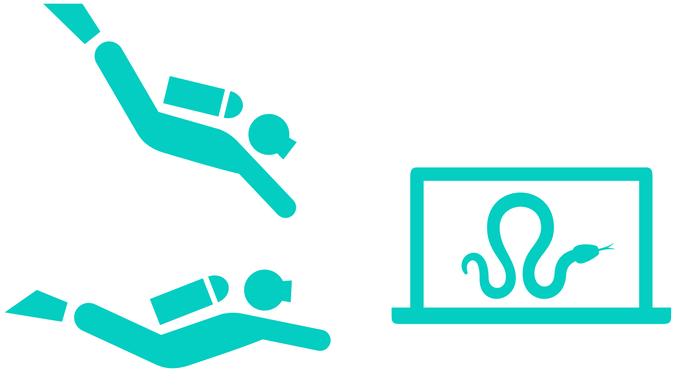
Solid baseline

Assumes linear separability

May underperform on complex data

LETS GET PROGRAMMING

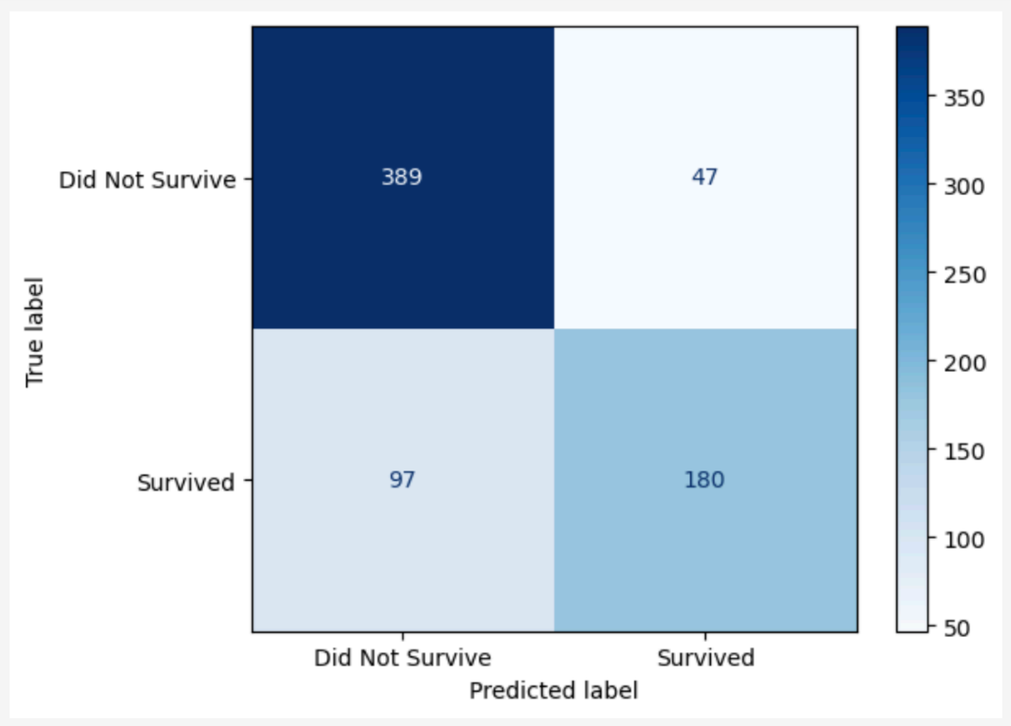
Session 1b



DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SESSION 1 WRAPUP

What can you practice before next week?



Can you think of any other interesting questions?

What does next week look like?