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

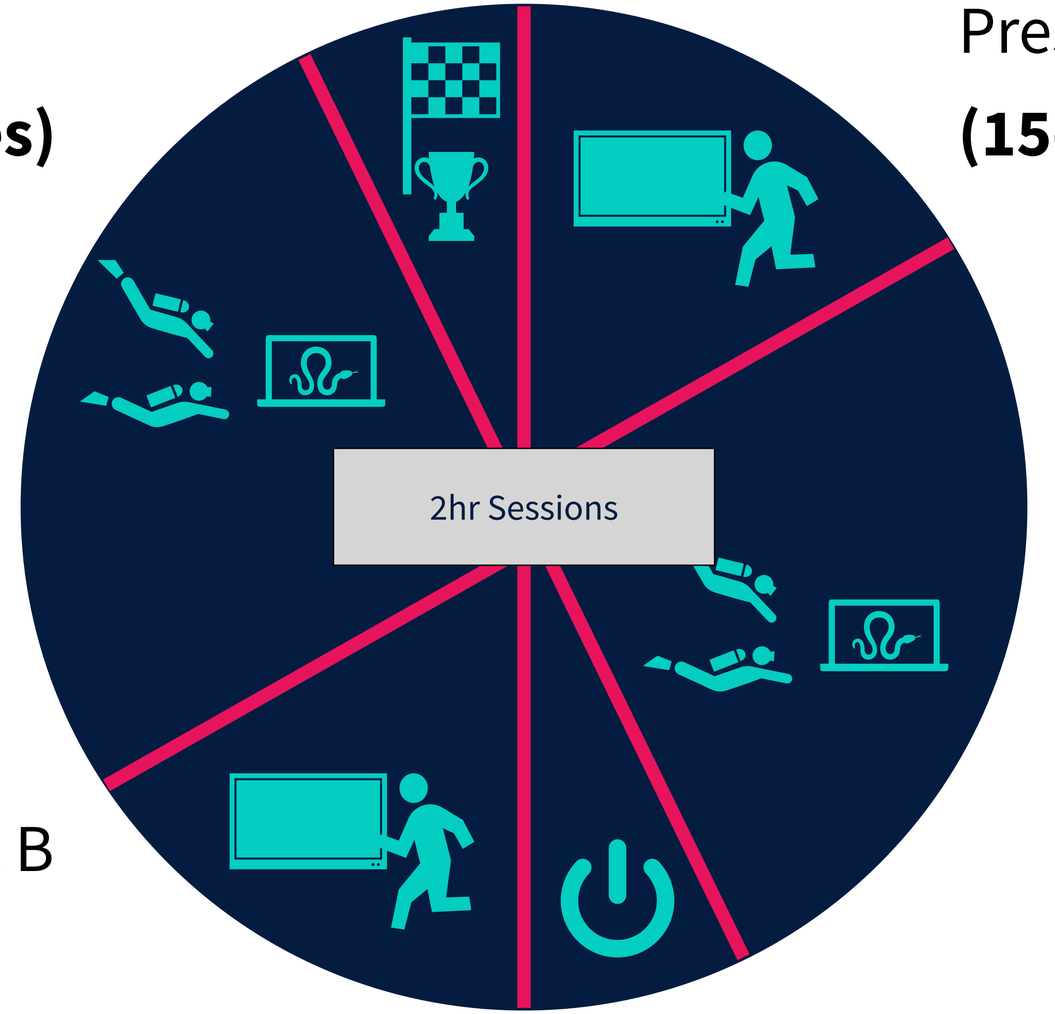
(10 minutes)

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

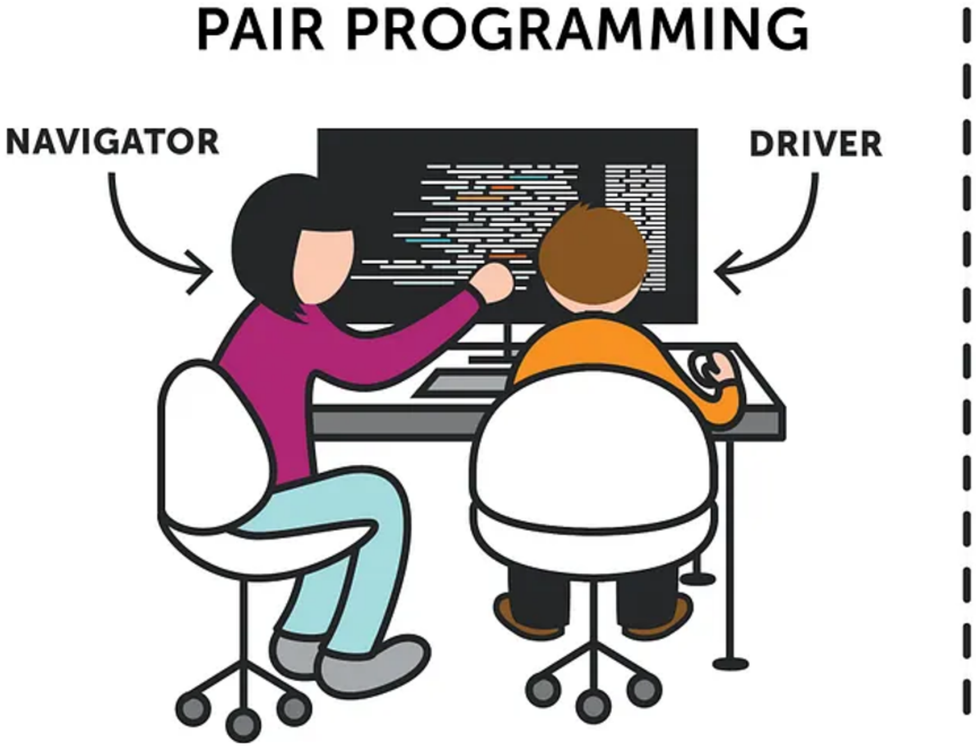


Image from https://medium.com/@tomspencer_uk/pair-programming-and-problem-solving-4531ef3bf171

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

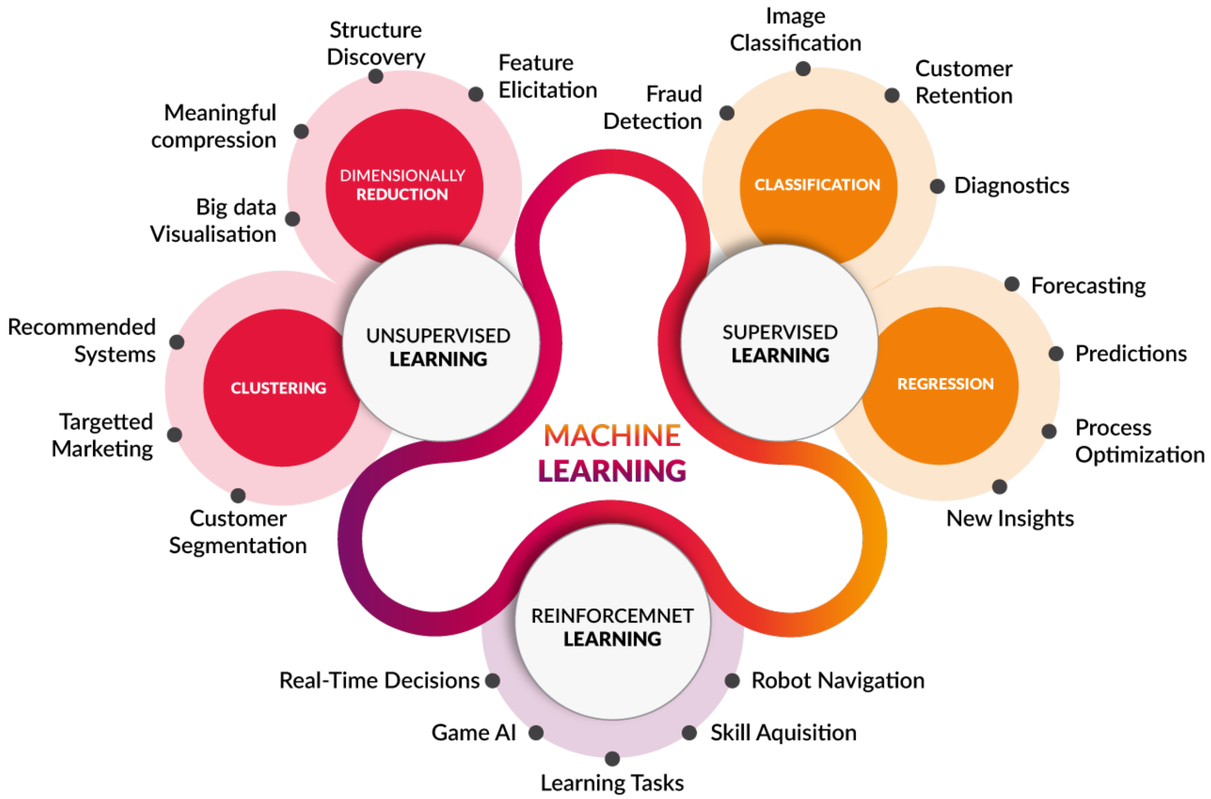
In the slides we may see text which is ‘pink’ in colour and a different font. This is to indicate it is a chunk of text, written in Python. The colour/font don’t matter just noticing it is code is important!

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

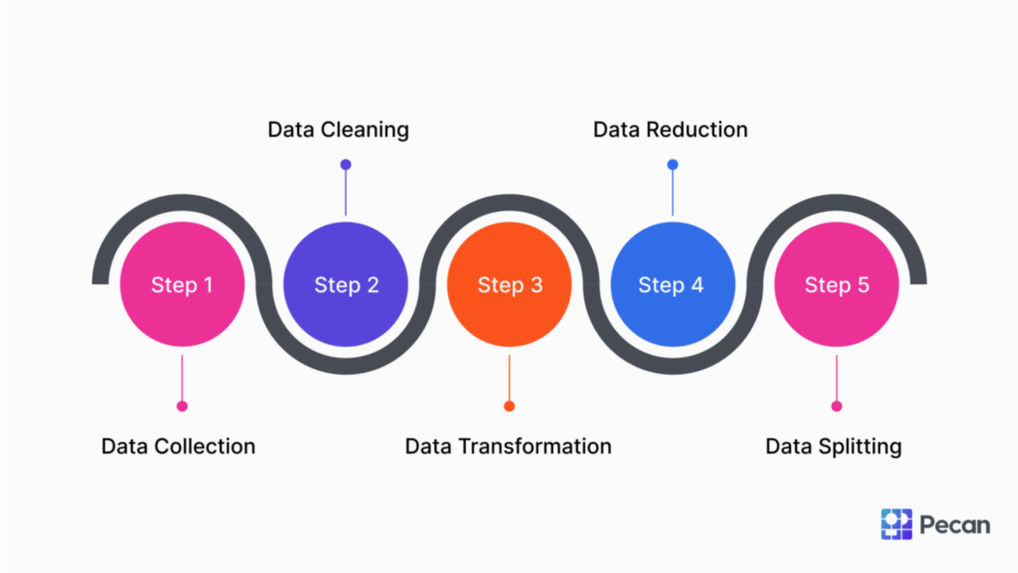


There are different types of machine learning branches and they are used for a range of different tasks in different areas.

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

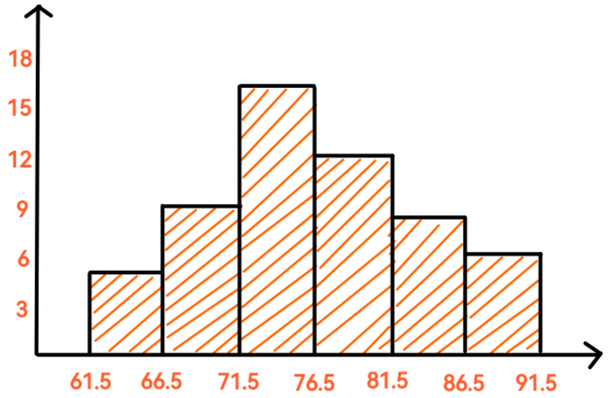
WHAT IS THE GOAL OF DATA PREPROCESSING?



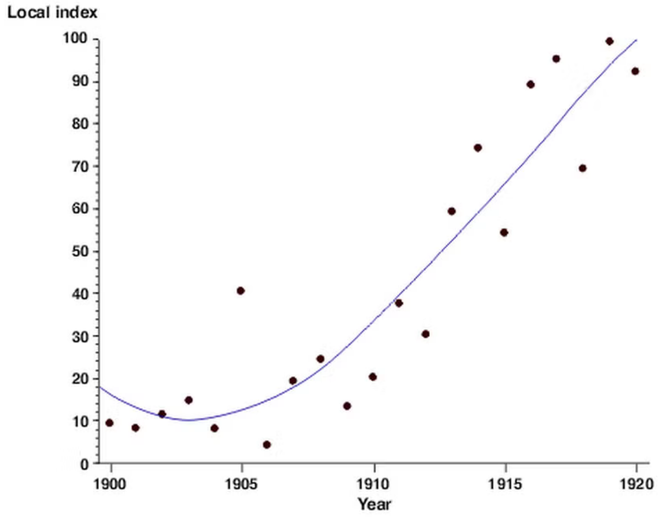
- Understand structure
- Spot issues
- Find patterns

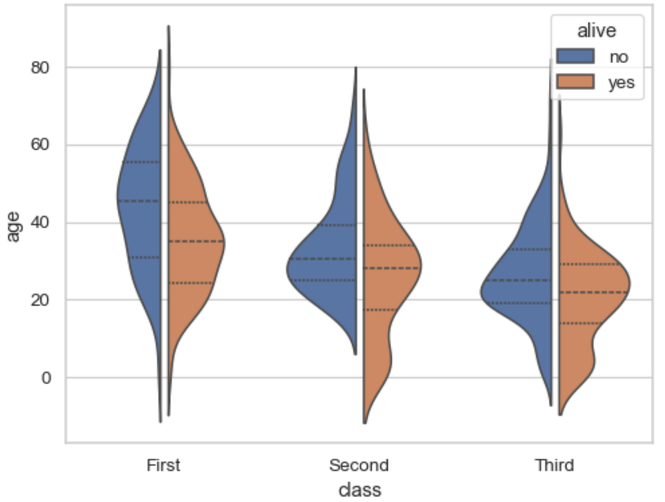
VISUAL EXPLORATION

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

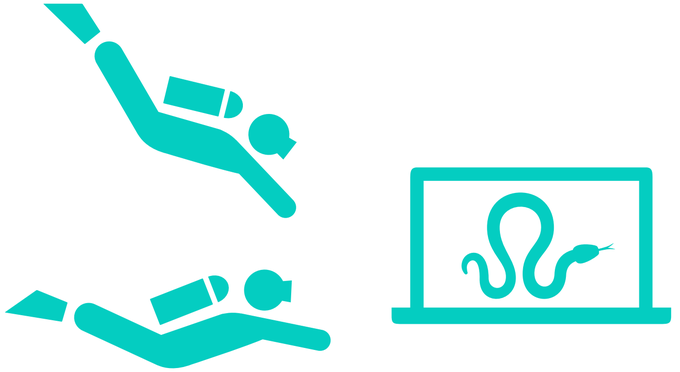
“Garbage in, garbage out.”

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



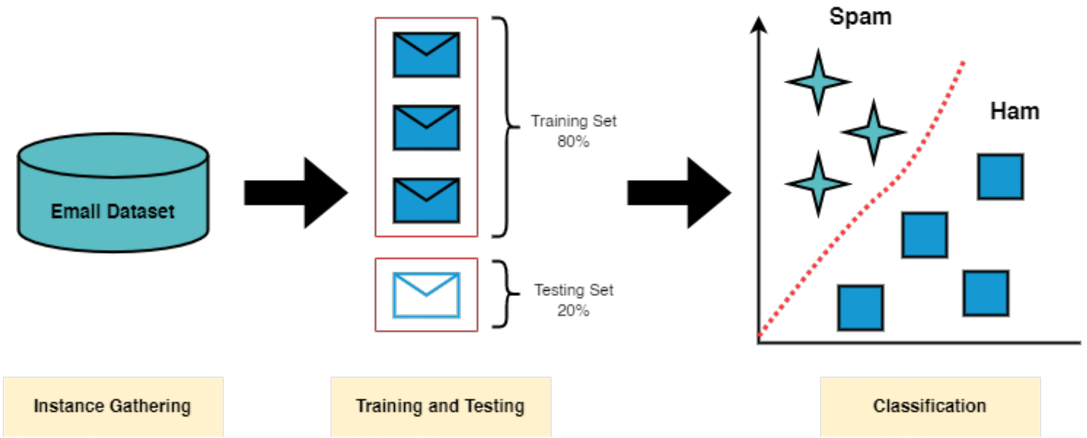
DATA CULTURE SOCIETY
CDCS.ED.AC.UK

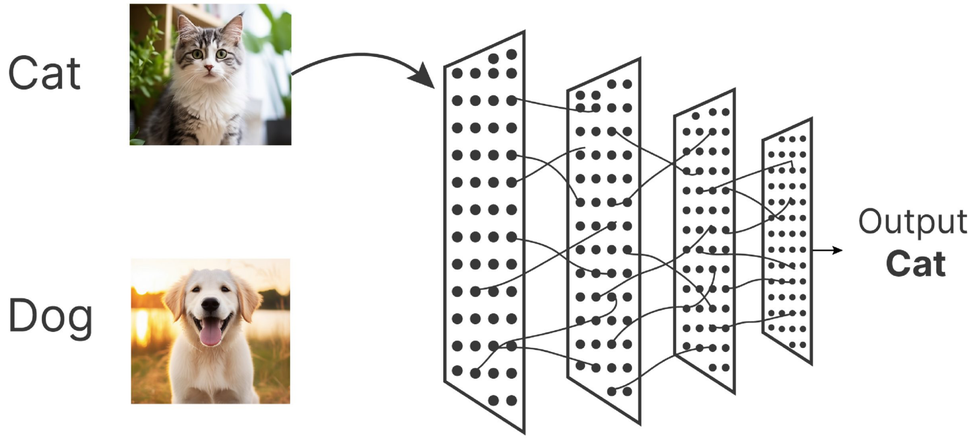
CLASSIFICATION BASICS AND LOGISTIC REGRESSION

WHAT IS CLASSIFICATION?

A supervised learning task

Goal: predict a category/label





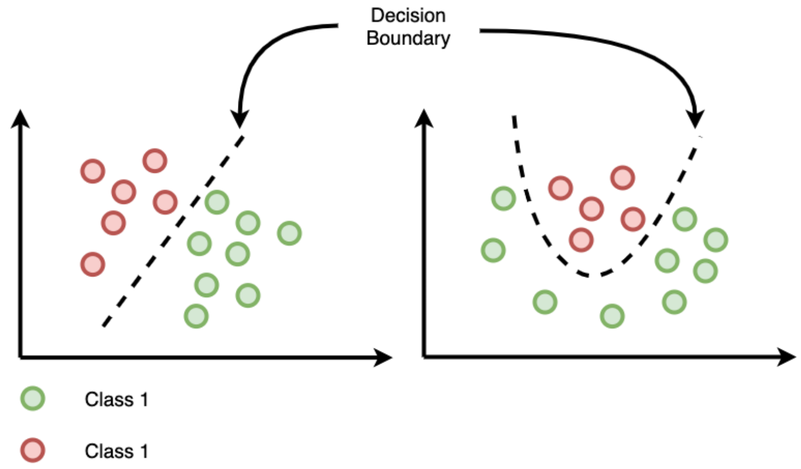
CLASSIFICATION VS REGRESSION

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

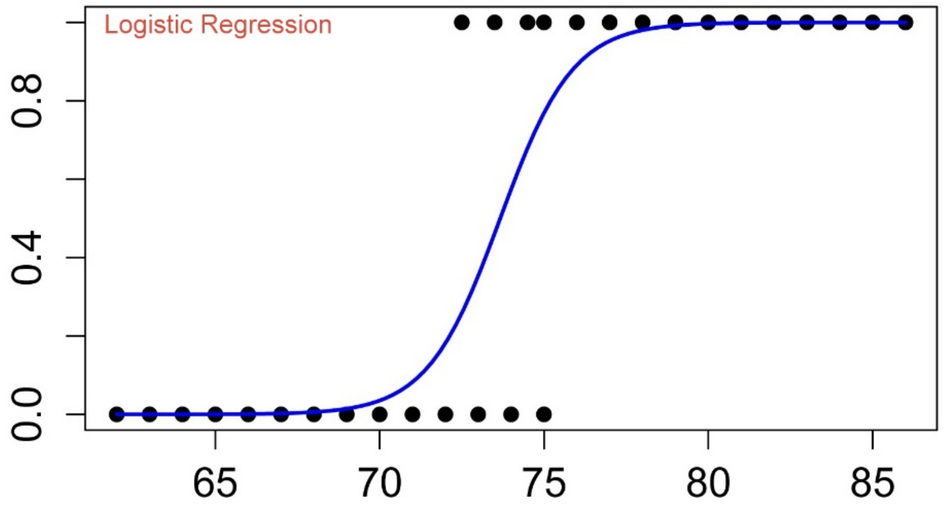
Decision Trees

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?

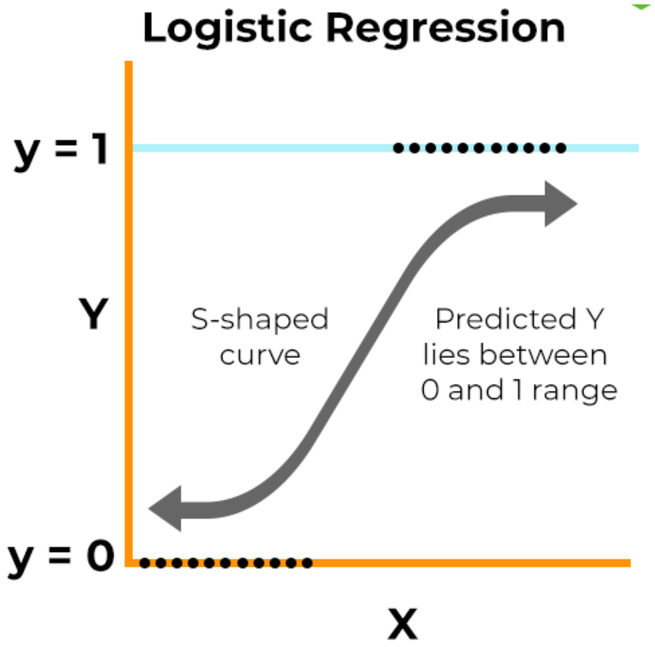
Finds the best-fit line/hyperplane

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

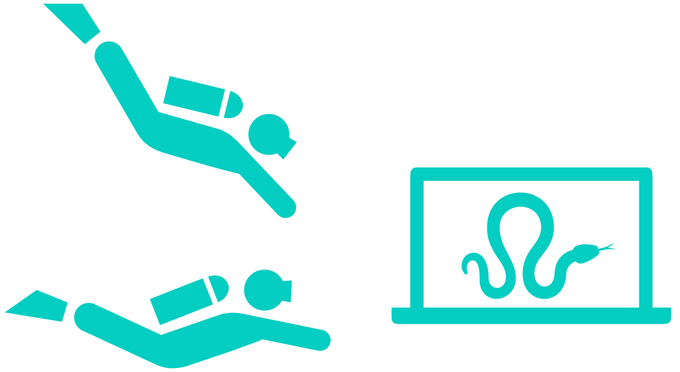
Solid baseline

Assumes linear separability

May underperform on complex data

LETS GET PROGRAMMING

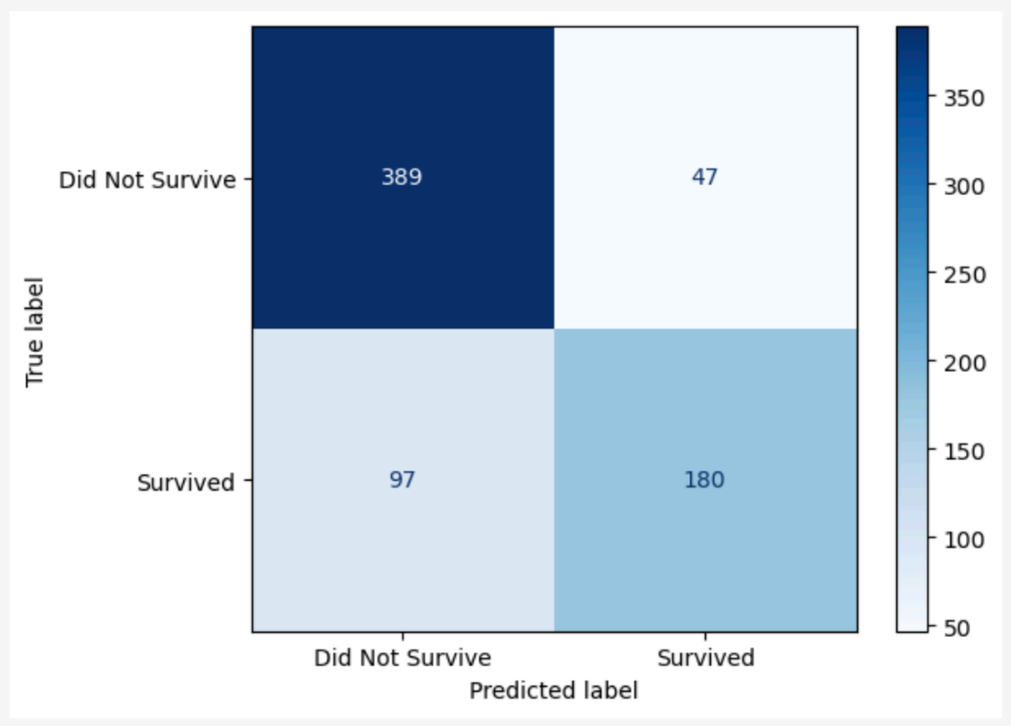
Session 1b



DATA CULTURE SOCIETY
CDCS.ED.AC.UK

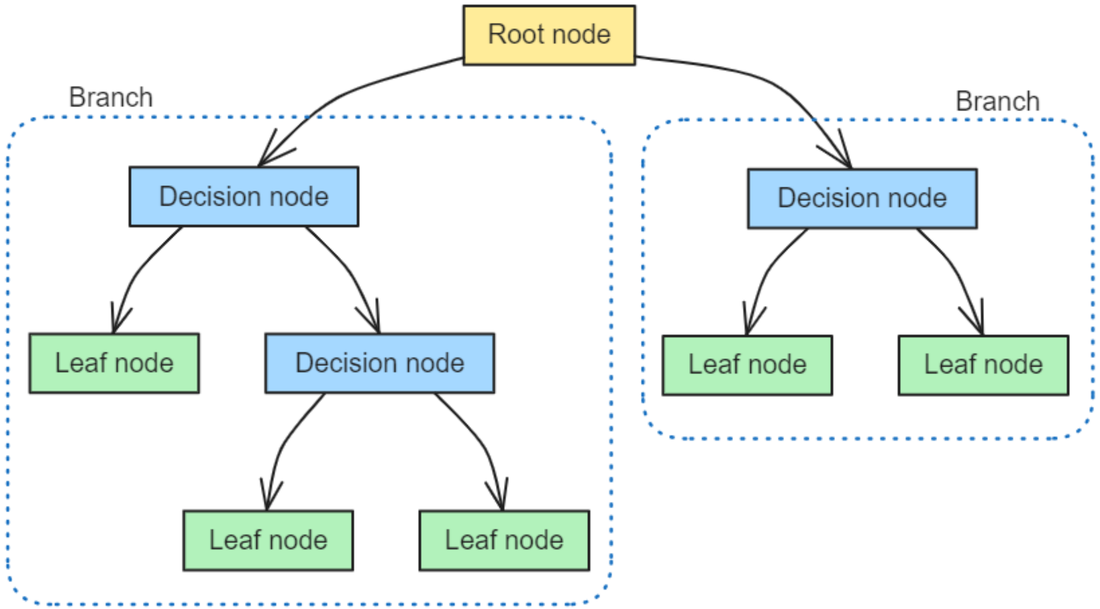
SESSION 1 WRAPUP

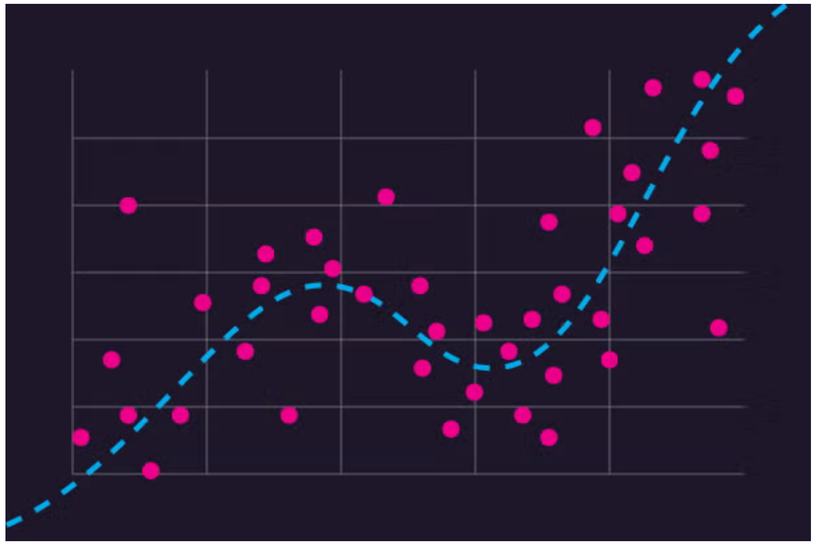
What can you practice before next week?



Can you think of any other interesting questions?

What does next week look like?





DATA CULTURE SOCIETY
CDCS.ED.AC.UK

MORE CLASSIFICATION MODELS
DECISION TREES AND K-NN

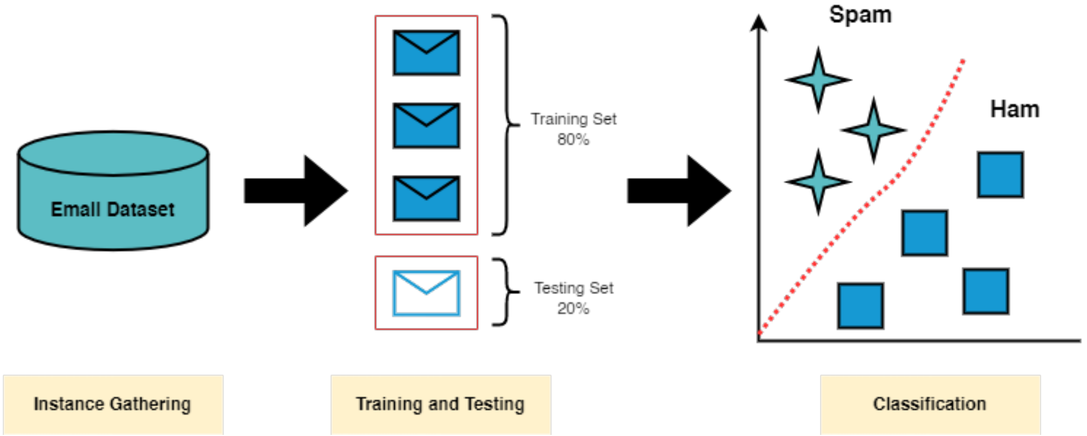
ESTIMATION VS PREDICTION

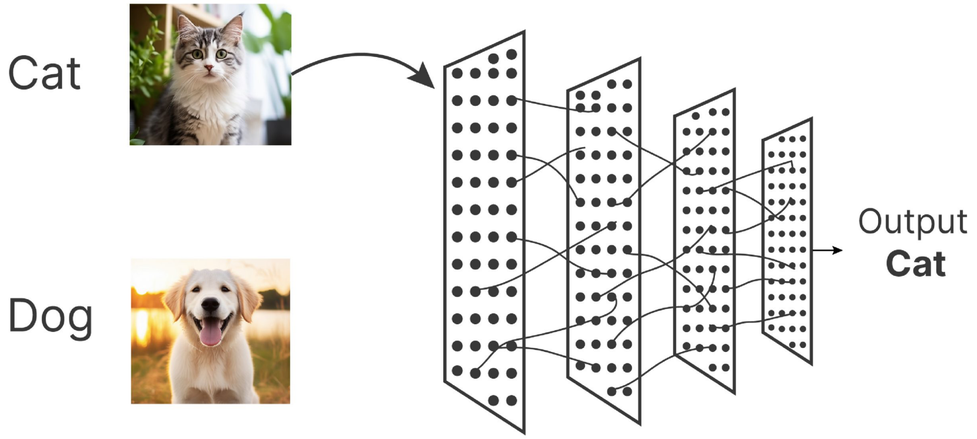
Aspect	Estimation	Prediction
Purpose	Understand which features are most associated with a class	Accurately assign the correct class label
Focus	Feature importance, effect direction (e.g. odds ratios)	Classification accuracy, precision, recall [How well does it do].
Interpretation	What increases the chance of spam?’	Is this new email spam or not?’

A SHORT RECAP ON CLASSIFICATION

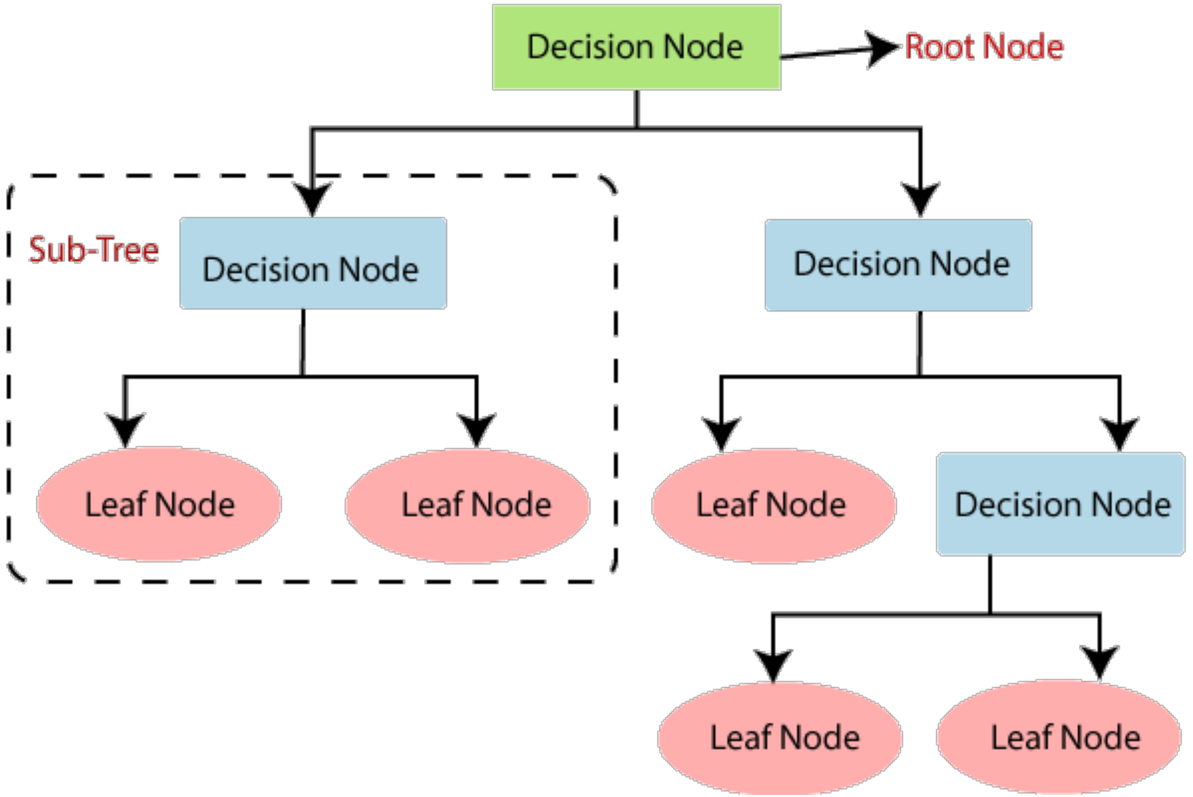
A supervised learning task

Goal: predict a category/label





DECISION TREES



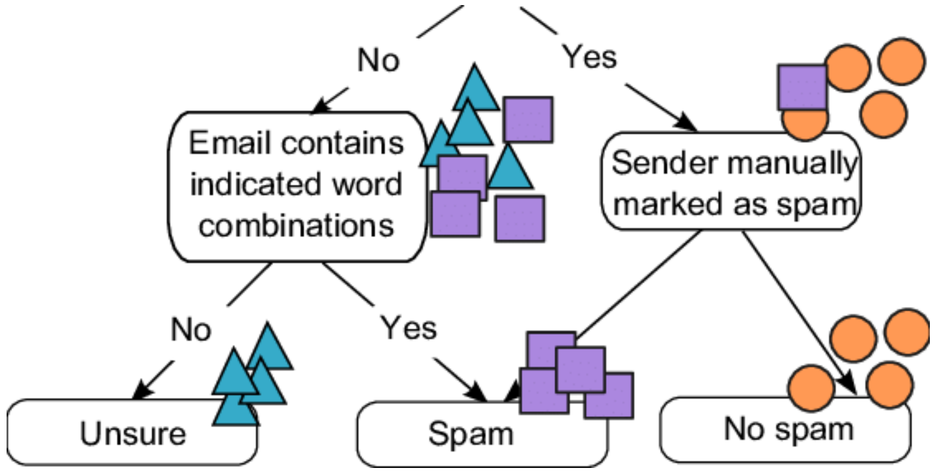
HOW DOES A DECISION TREE MAKE A DECISION?

Picks the best split at each node

Keeps splitting until:

- Max depth reached
- No more gain from splitting

Assigns class based on majority vote in leaf nodes



PROS AND CONS OF DECISION TREES

Easy to understand

Can handle mixed data types

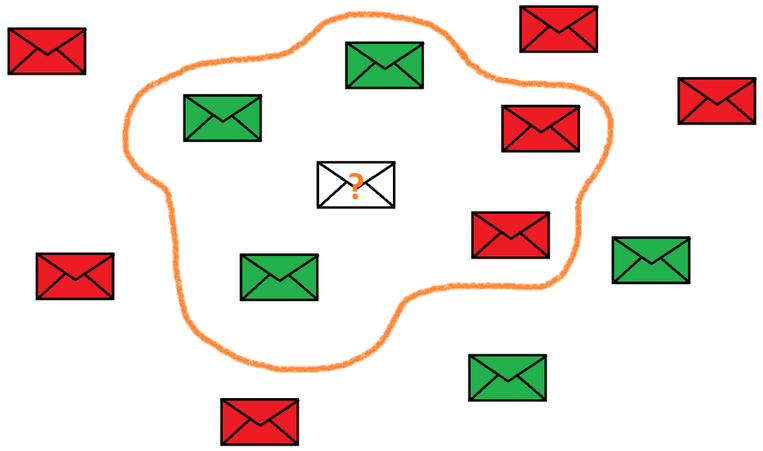
No need to scale features

Can overfit

Unstable to small data changes

Doesn’t always generalise well

🤝 K-NEAREST NEIGHBOURS



Instance-based learning: remembers all training data

Predicts based on closest k neighbours

Simple, but computationally heavy for large datasets

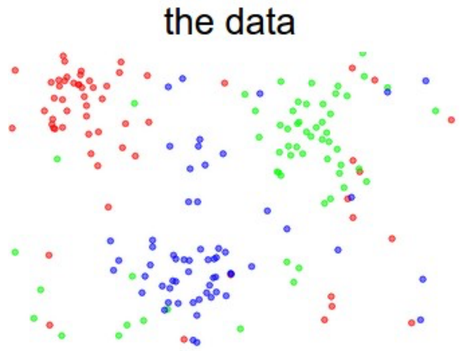
HOW DOES THE K-NN ALGORITHM MAKE A DECISION?

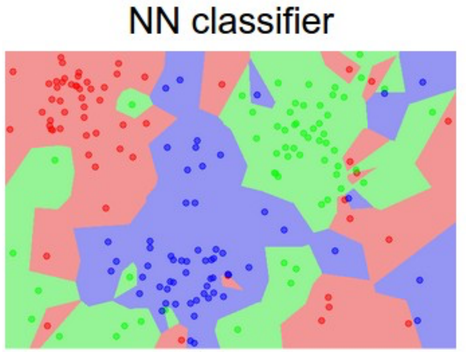
1. Compute distance between test point and all training points

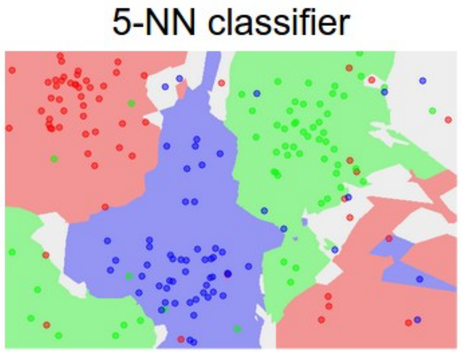
2. Pick the k closest ones

3. Use majority vote to assign class label

4. Sensitive to k and scaling of features







PROS AND CONS OF K-NN

Very intuitive

No training phase

Works with any number of classes

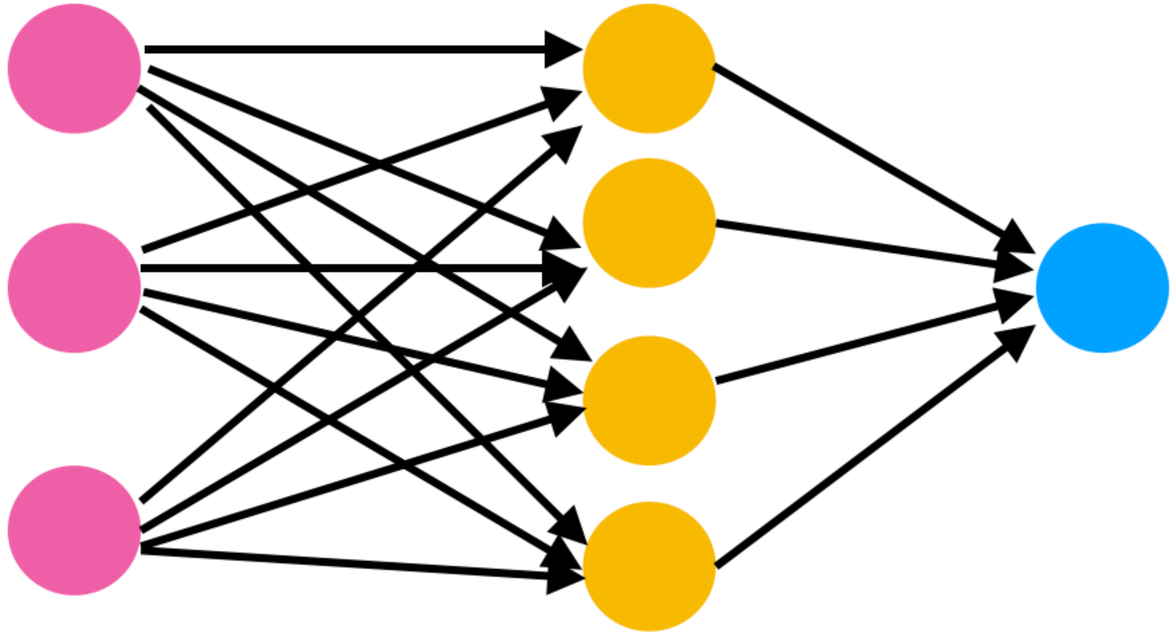
Slower with large datasets

Sensitive to feature scaling

Needs careful choice of ‘k’

A SHORT NOTE ON NEURAL NETWORKS

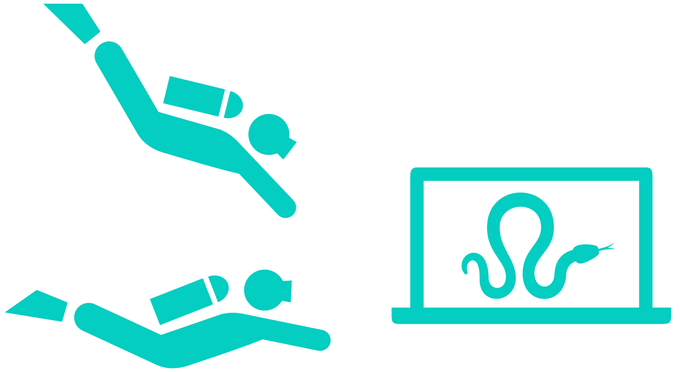
- Inspired by the brain, Layers of connected “neurons”.
- Great for complex, high-dimensional problems but require more data & compute
- Typically used when simpler models don’t perform well



Method	When?	What to worry about.
Logistic Regression	Quick baseline, linearly separable data	Doesn’t capture complex boundaries
Decision Trees	Interpretability, non-linear relationships	Prone to overfitting
K-NN	Simple, low-dimensional problems	Slow with big data, needs scaling
Neural Networks	Large datasets with complex patterns	Requires tuning, hard to interpret

LETS GET PROGRAMMING

Session 2a

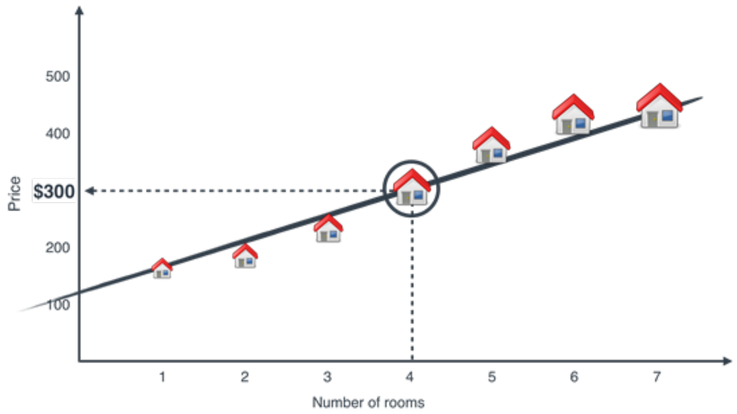


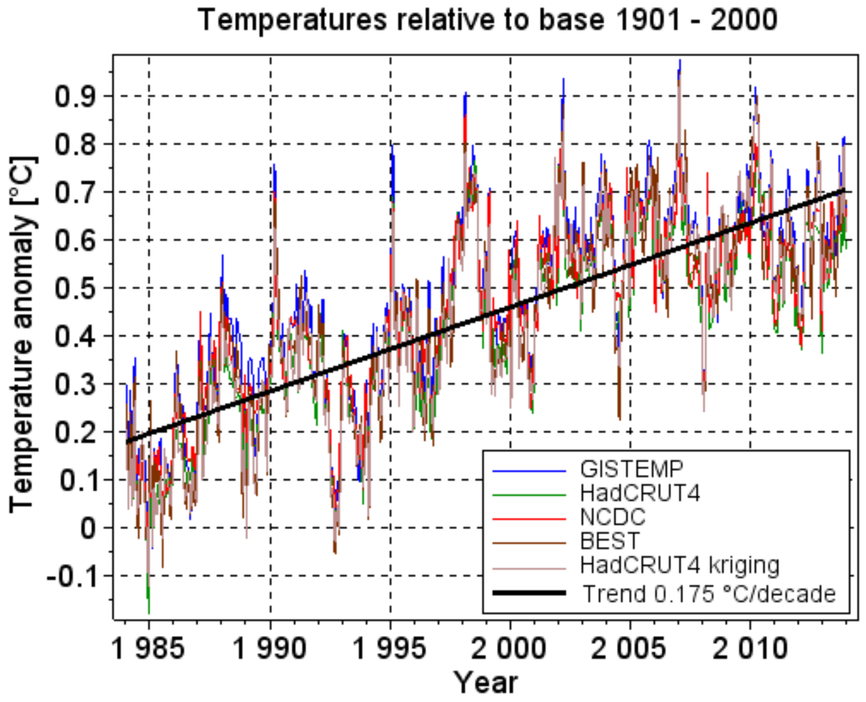
REGRESSION AND PRACTICAL CONSIDERATIONS IN ML

WHAT IS REGRESSION?

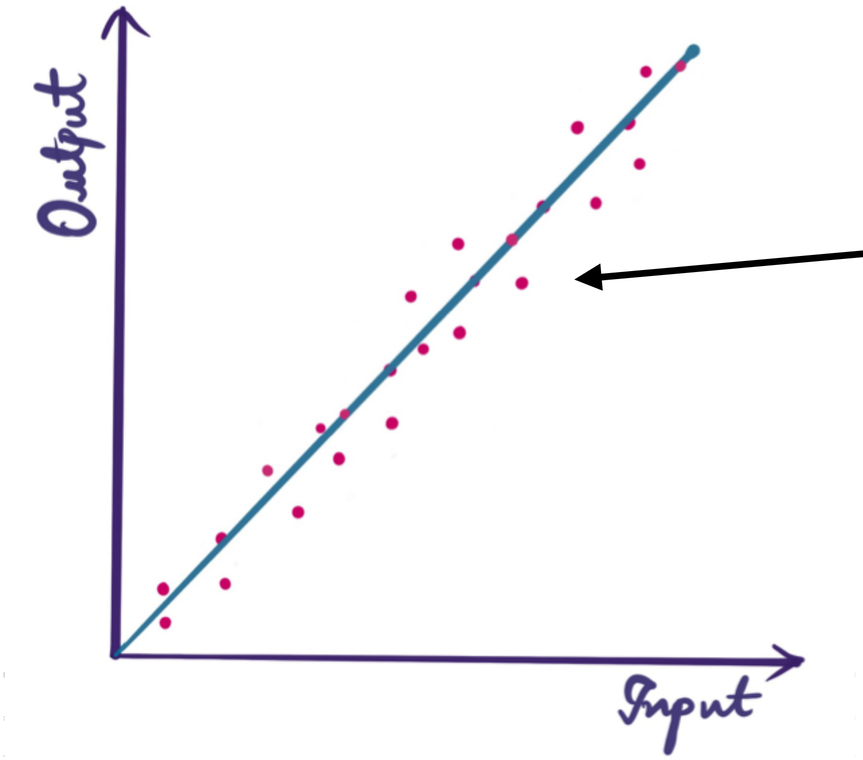
A supervised learning task

Goal: Predicting continuous outcomes





HOW DOES REGRESSION WORK?



Common regressors:

Linear regression

Ridge/Lasso

Decision Tree Regressor

Neural networks

WHAT IS LINEAR REGRESSION IN AN ML CONTEXT.

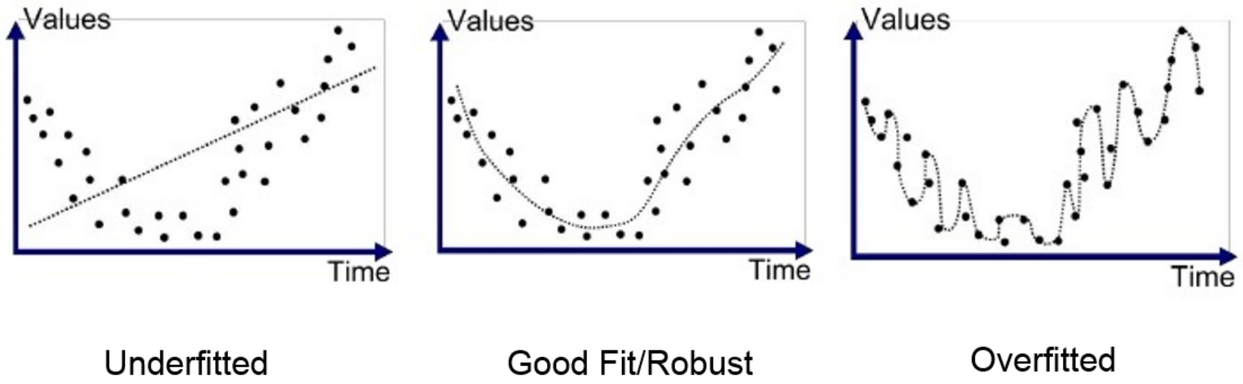
Traditional (modelling-as-analysis)	Machine Learning Approach
Fit the model once	Train/test split (or cross-validation)
Interpret coefficients	Prioritise predictive performance
Check p-values, confidence intervals	Track metrics like MAE, MSE, RÂ²
Choose features based on theory	Use feature selection, regularisation
Assume model is correct	Evaluate on unseen data
Focus on statistical assumptions	Focus on robustness and generalisation

HOW DO WE KNOW IF A REGRESSOR DID WELL?

Metric	What is it?	What's the formula?	Good Value?
MAE **Mean Absolute Error**	Average absolute difference	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$	Lower
MSE **Mean Square Error**	Squares large errors	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	Lower
R^2	Proportion of variance explained	$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$	Closer to 1

OVERALL PRACTICAL CONSIDERATIONS

Overfitting / Underfitting



Underfitted | Good Fit/Robust | Overfitted

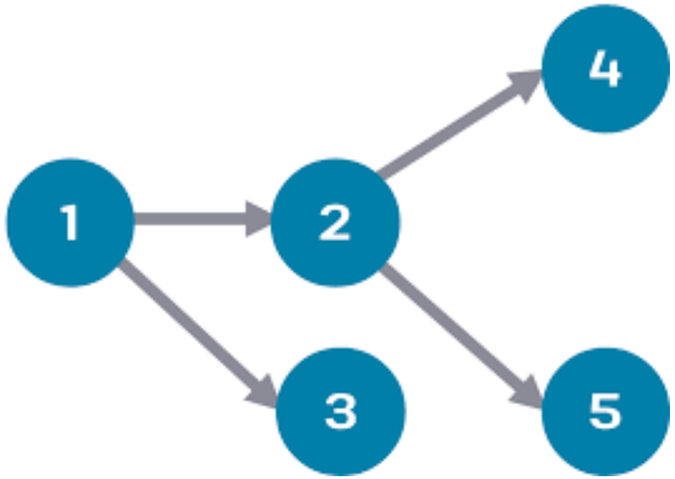
- Model performs well on training data but poorly on new data [OVER-FITTING]

- Model is too simple to capture patterns [UNDER-FITTING]

- Fix with simpler/more complex models, regularisation, or **more data**

OVERALL PRACTICAL CONSIDERATIONS

Feature selection

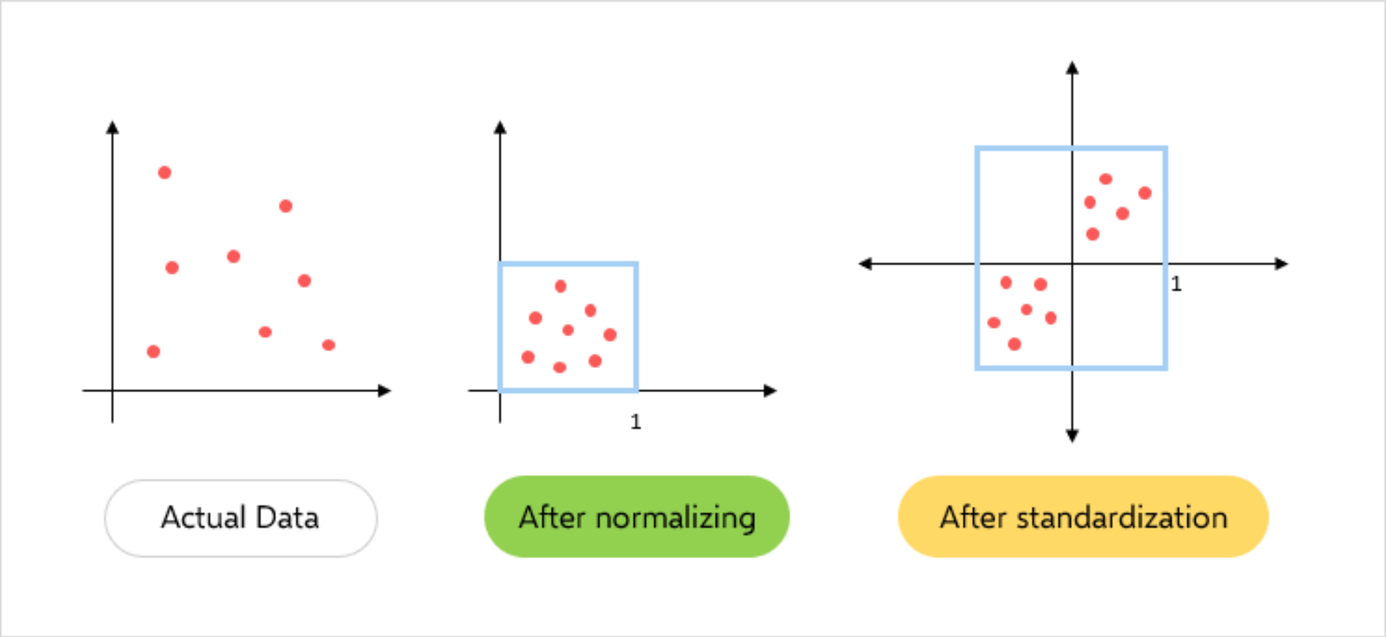


– Some features may be noisy or irrelevant.

– Use causal inference techniques to establish a directed acyclic graph first when considering features.

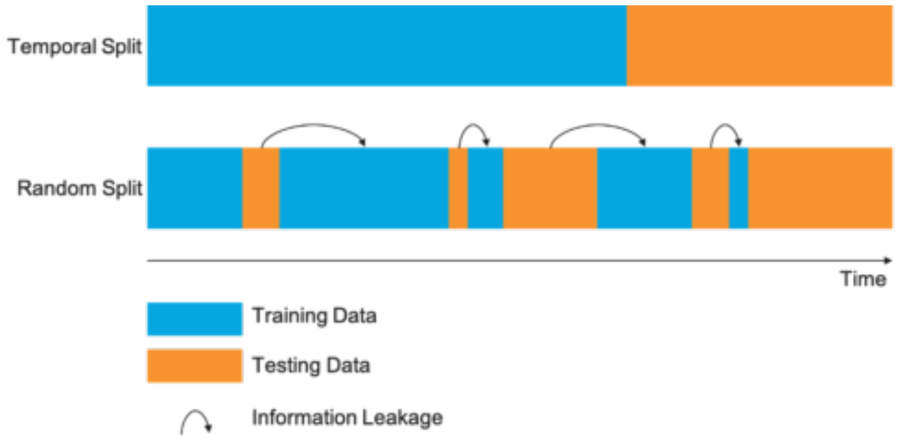
OVERALL PRACTICAL CONSIDERATIONS

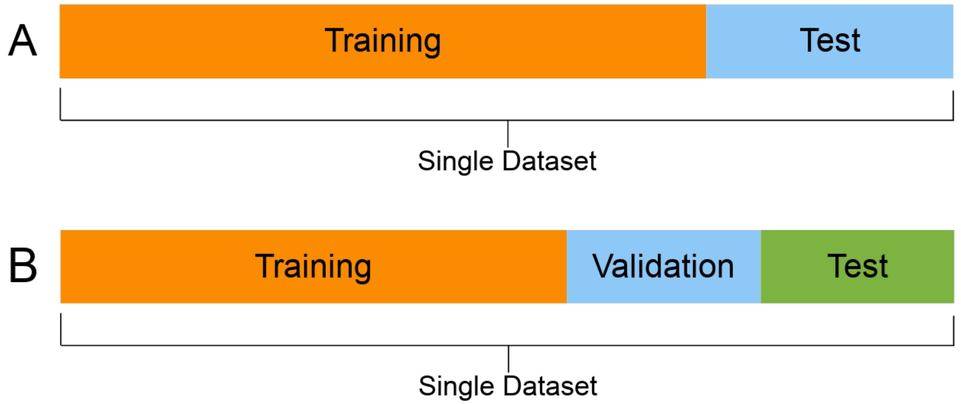
Scaling…some models (e.g. k-NN, linear regression) need scaled inputs



OVERALL PRACTICAL CONSIDERATIONS

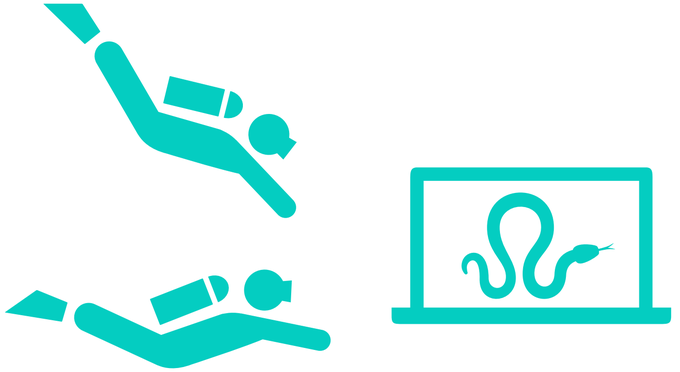
Data leakage…when information from outside the training set is used improperly.





LETS GET PROGRAMMING

Session 2b



DATA CULTURE SOCIETY
CDCS.ED.AC.UK

SESSION 2 WRAPUP

Additional Resources

- Introduction to Machine Learning with Python: A Guide for Data Scientists
  by Sebastian Raschka and Vahid Mirjalili

- Machine Learning with R
  by Brett Lantz

- Artificial Unintelligence: How Computers Misunderstand the World
  by Meredith Broussard

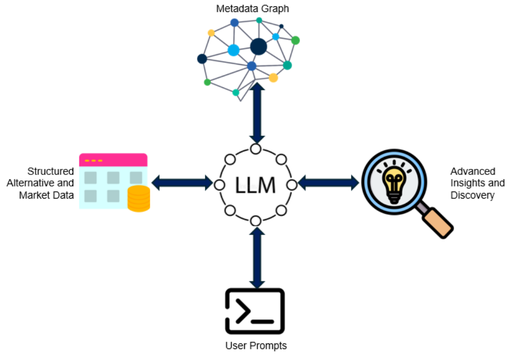
- The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking
  by Shannon Vallor

Feedback for us...

- We hope you've enjoyed the course as much as we did.
- It is really useful for us to hear your feedback

https://forms.office.com/r/YYNrqvuNr8
Should be really quick and only take 5 mins (maximum!)

- **Advanced Uses of LLMs**
  28th April, 5th and 12th May
  with Martin Disley



- **Regression and Mixed Effects Modelling**
  Sessions: 29th April, 6th May, 13th May
  Drop-Ins: 9th May and 16th May
  with Fang Yang and Aislinn Keogh