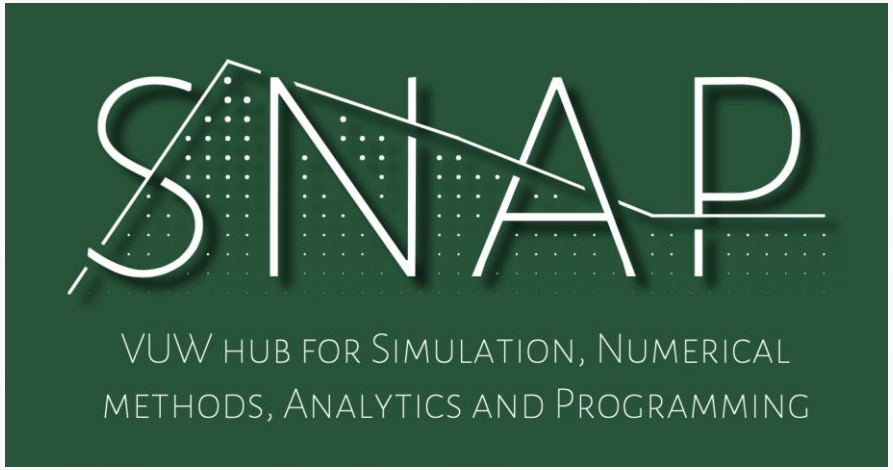
Introduction to AI & ML – Day 1



SNAP Workshop, 02/09/26

Emily O'Riordan
emily.o'riordan@vuw.ac.nz

What is VUW-SNAP?

VUW-SNAP a Network Hub for Simulation, Numerical methods, Analytics, and Programming

Purpose

- Bring computational researchers together, community building
- Improve computational, numerical skills of researchers PG, Post-Doc & Academics
- Promote our capabilities externally, raising our profile, facilitate external engagement (govt/industry/funding)

Website

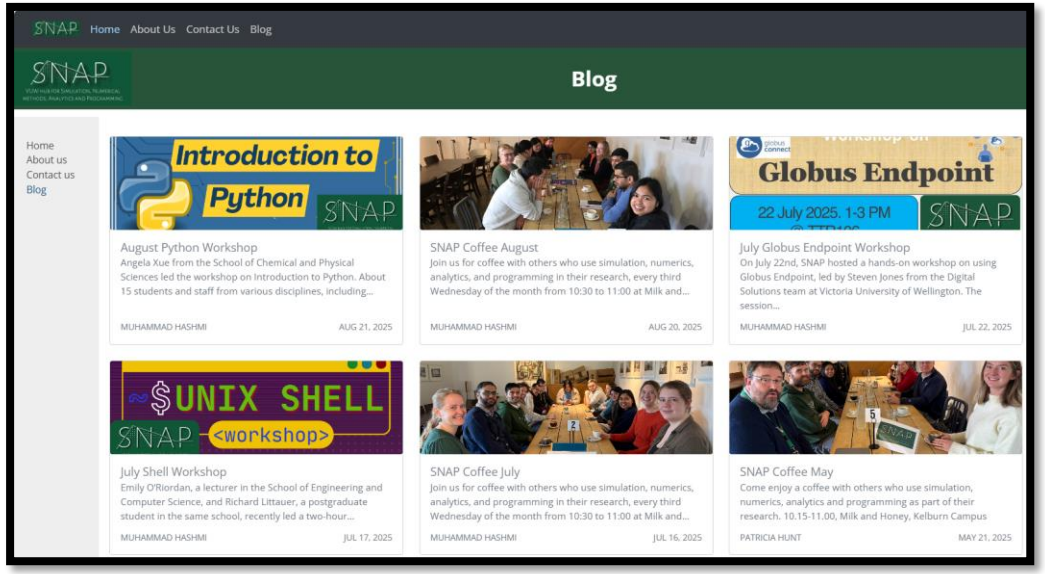
- https://vuw-snap.github.io/SNAP-decisions/

Email Addresses

- snap@vuw.ac.nz
- snap-students@vuw.ac.nz

Subscribe to SNAP mailing list

- https://lists.vuw.ac.nz/mailman/listinfo/snap



Day 1: Intro & Supervised Learning

13:00 – 13:45: Introductions & Intro to ML

13:45 - 14:30: Regression models

14:30 - 14:40: Break

14:40 - 16:00: Classification models

Day 2: Ethics & Unsupervised Learning

13:00 – 13:20: Ethical considerations

13:20 - 13:50: Clustering

14:00 - 14:30: Dimensionality Reduction

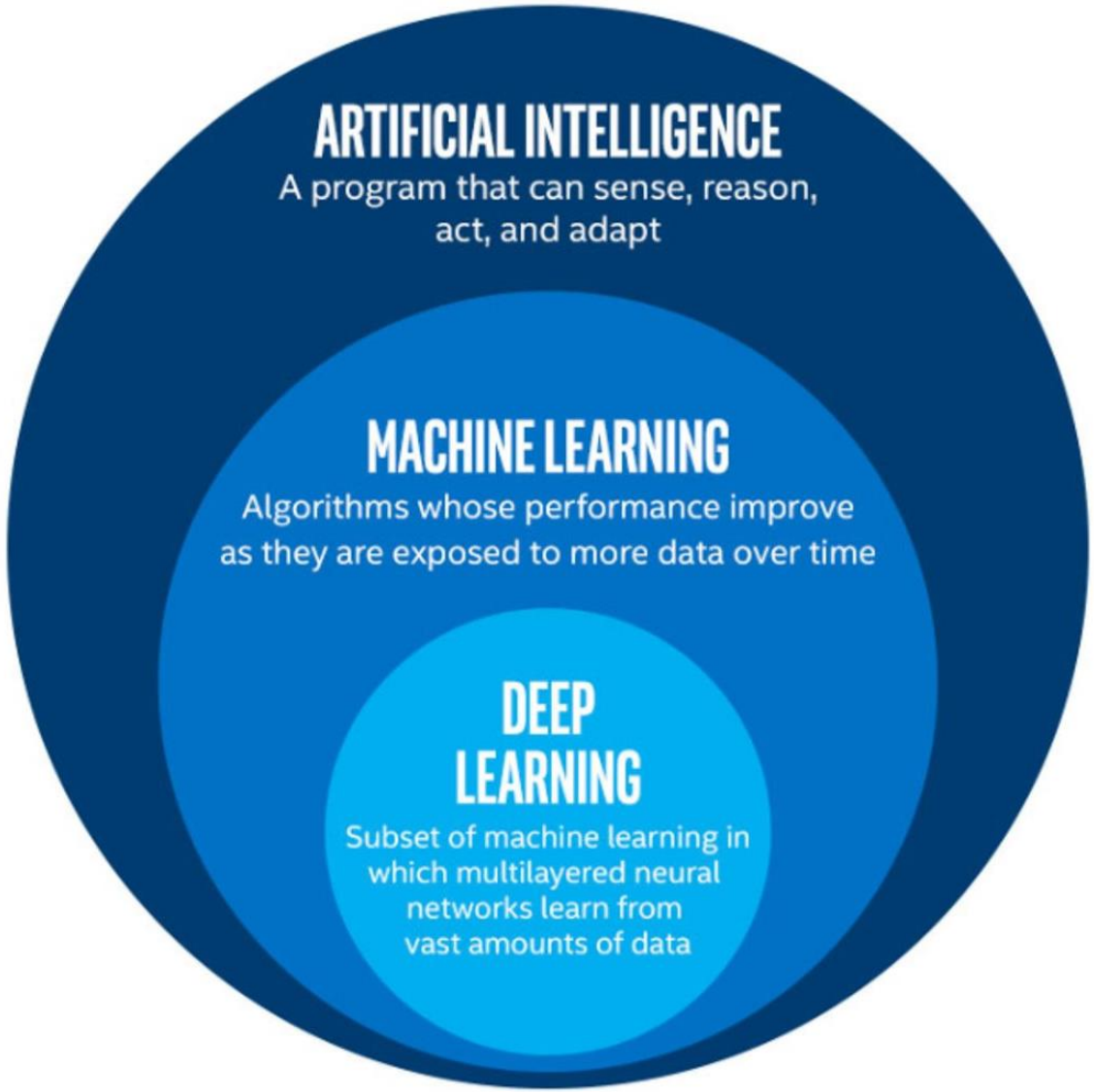
14:30 - 14:35: Break

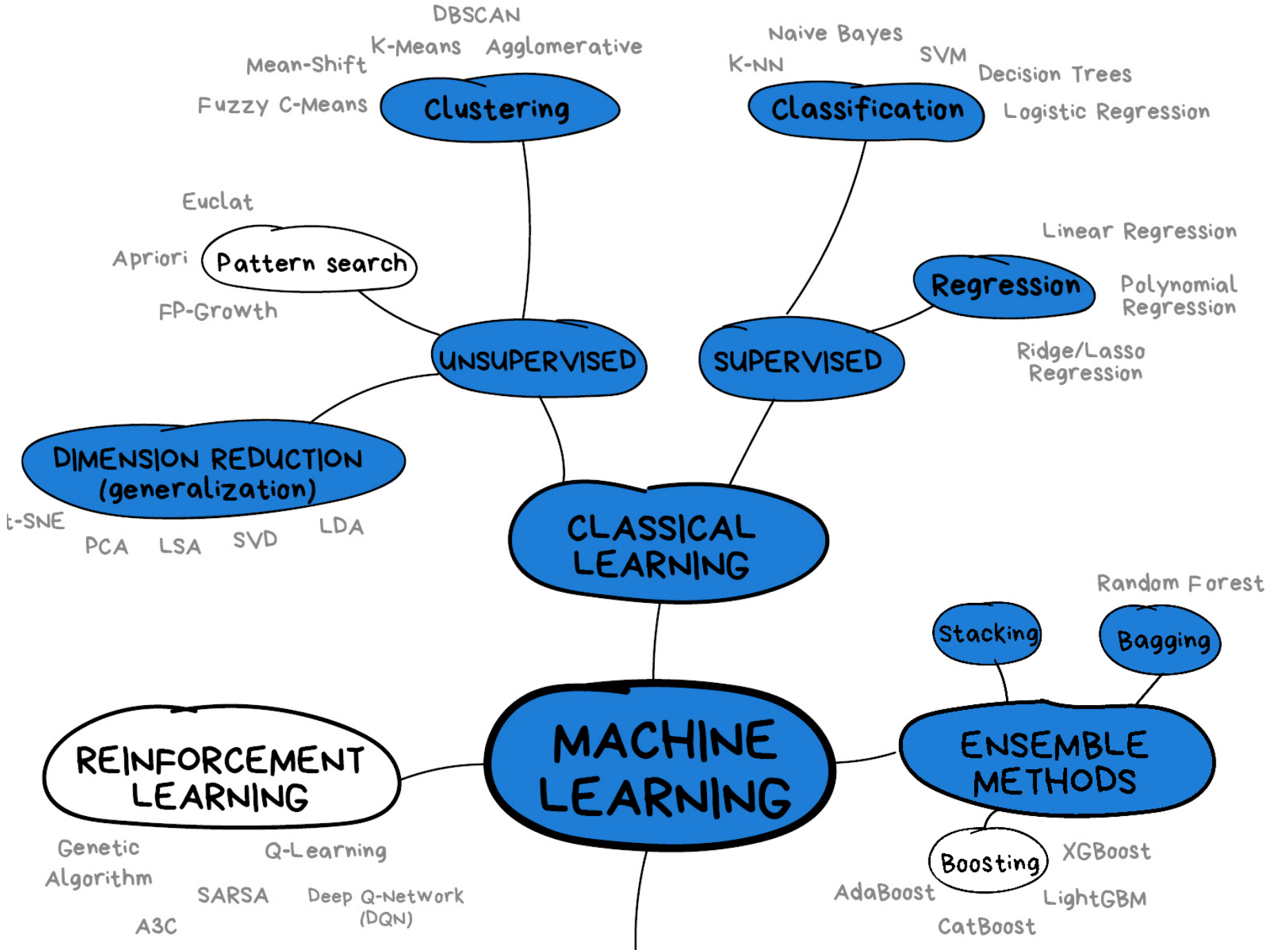
14:35 – 16:00: Neural Networks

Introductions

- Introduce yourself and your research
- How you're using AI day-to-day
- How you think you might be able to use AI in your research (if you aren't already!)

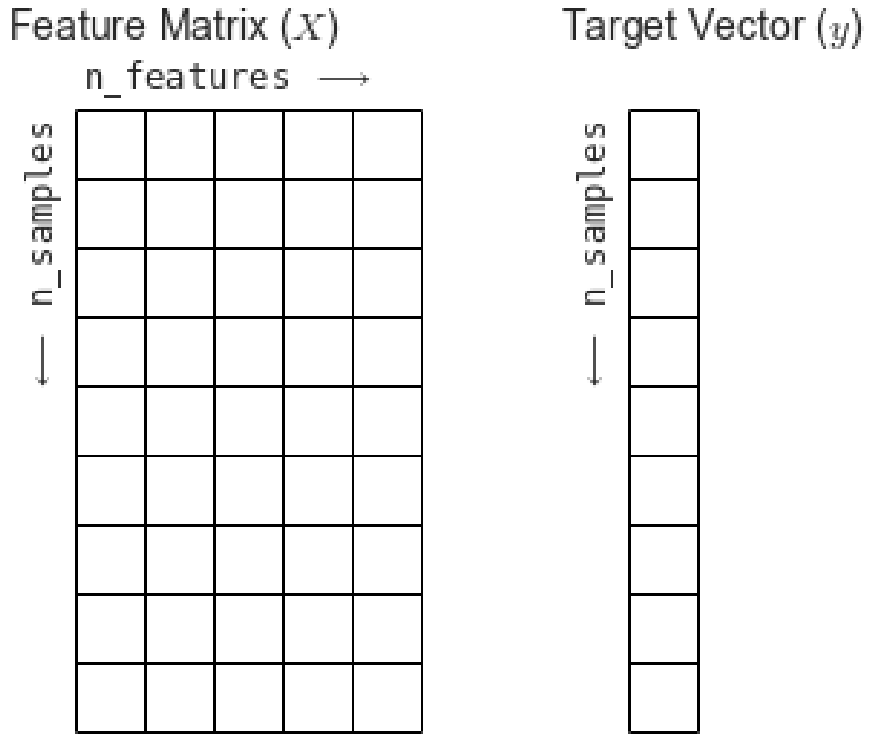
Introduction to Machine Learning




MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost

Data is 🔑

- Machine learning models are data-hungry: they need a lot of data!
- Data is usually stored as a set of samples, made up of features. *Sometimes* we include a target variable.
- A sample can be a row in a CSV file, a picture, a sound, an astronomical object… whatever you can describe with a fixed set of quantitative traits.
- Data quality is vital. “Garbage in, garbage out”



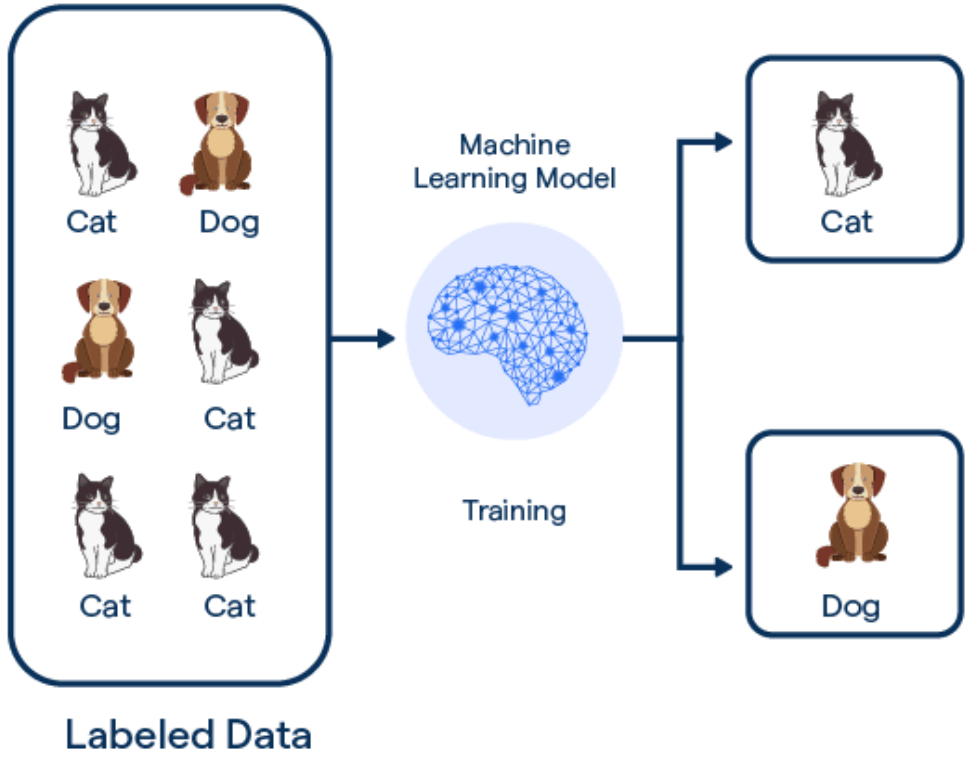
Python packages

- Scikit-learn
  - Open Source 🙌
  - Imported as sklearn
  - We'll be using this!
- Pytorch
  - Owned by Meta
  - Used in most ML research
- Tensorflow
  - Owned by Google
  - Used in industry

Supervised vs Unsupervised Learning

Supervised learning:

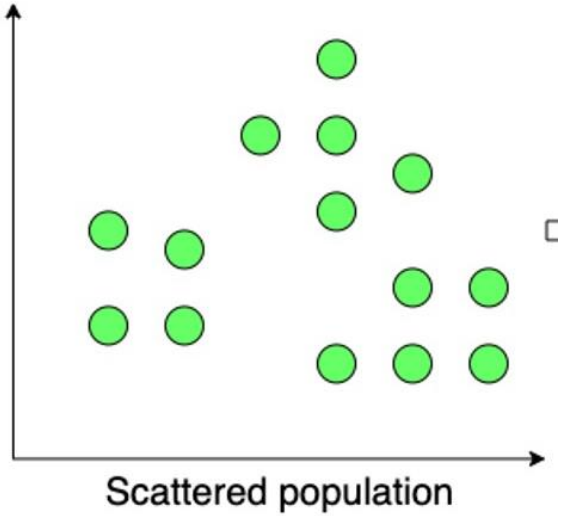
- Each sample in the dataset has a target value
- Supervised learning models aim to predict this target value for new samples
- Example: a model to distinguish between images of cats and dogs
- We provide the model with images that have already been labelled as “cat” or “dog” so that it can learn from these examples.

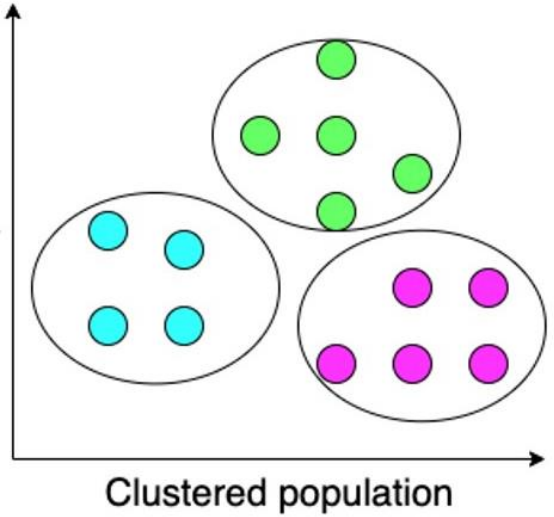


Supervised vs Unsupervised Learning

Unsupervised learning:

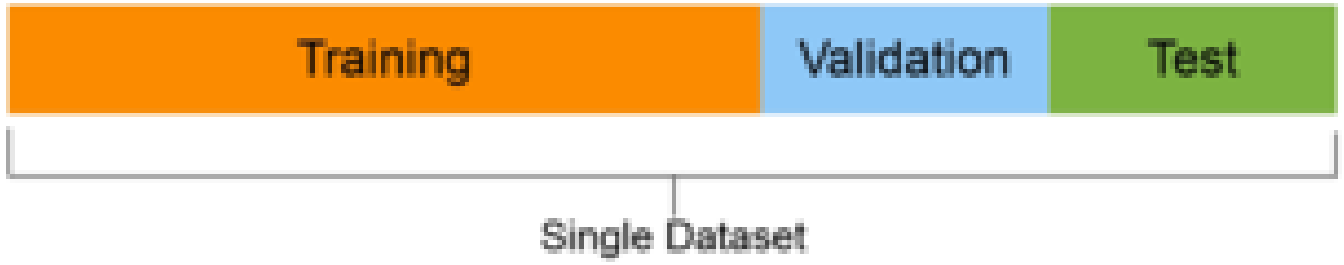
- Samples in the dataset do not have a target value
- Unsupervised learning models aim to find patterns in the dataset
- Example: a model to group customers based on their spending and frequency habits
- We provide the model with unlabelled data (e.g., customer purchase histories), and it learns to group or cluster similar samples without being told the “right” answer.





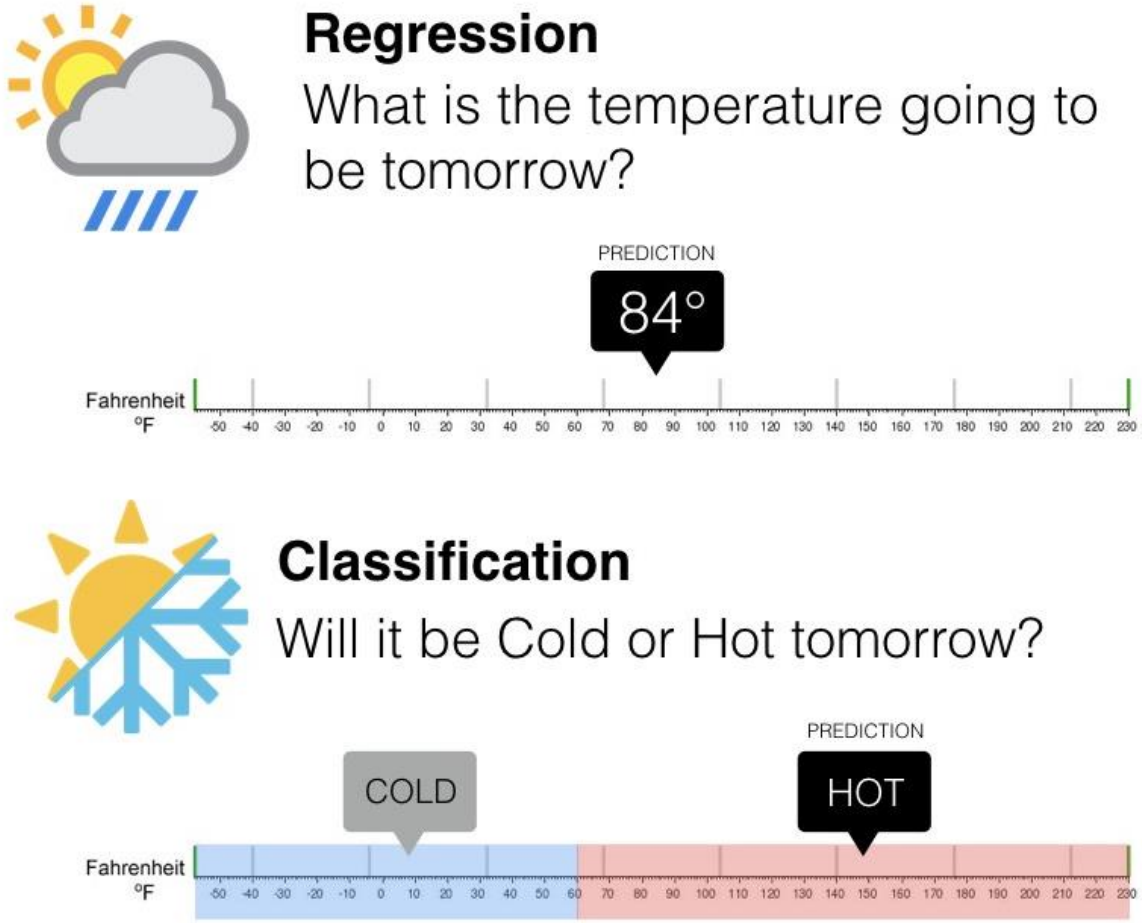
Training, Validation, Testing

- Supervised ML models require us to split our data into distinct groups:
  - Training set (~80%)
  - Test set (~20%)
- We do this to ensure our model can generalise beyond the data it is trained on
  - If it can't, we call this overfitting
- In practise, we also use a validation set too…



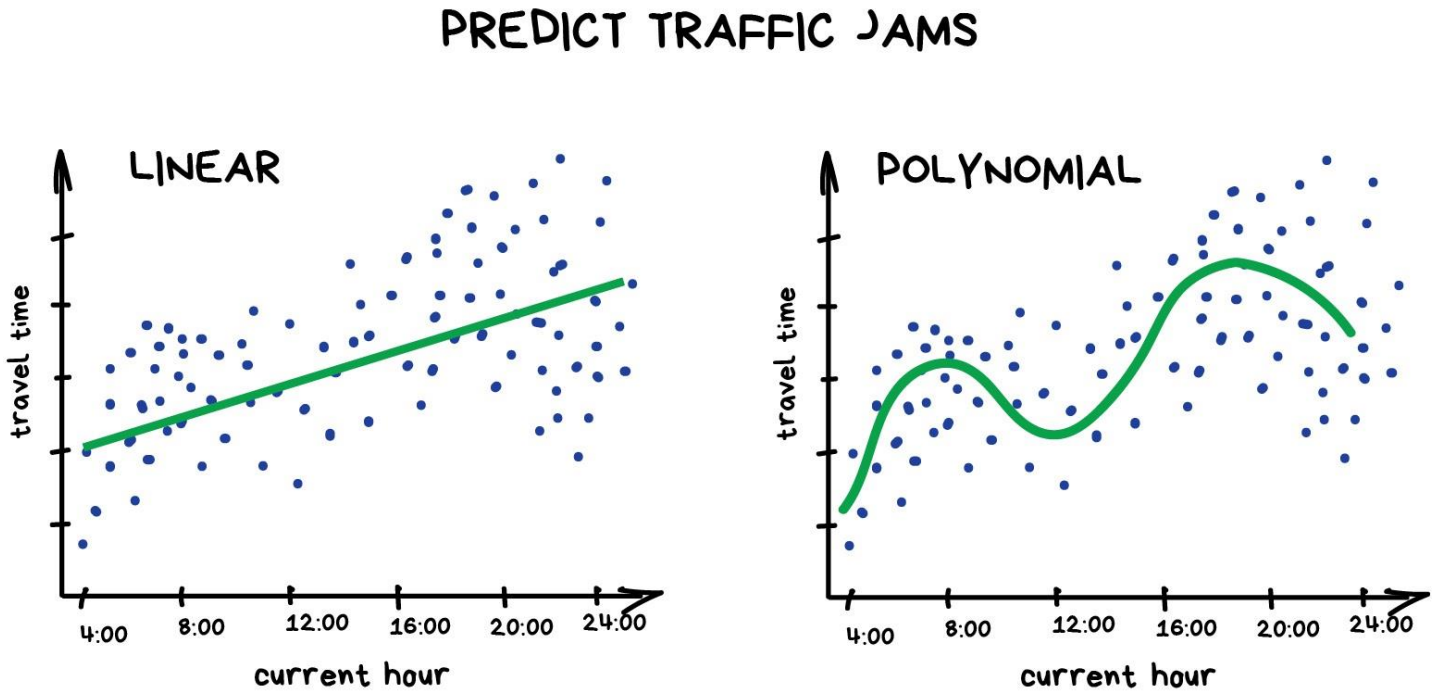
Supervised Learning: Regression vs Classification

- We use regression algorithms when the labels are continuous
  - The house is $500,000
  - An exam score, 83%
  - Blood pressure is 128 mmHg
- We use classification algorithms when the labels are discrete
  - The house is $, $$, $$$
  - An exam grade, A, B, C…
  - Blood pressure levels are normal, elevated, or high



Regression

- Aim: learn a line of best fit
- Linear regression: $y = mx + c$
- Polynomial regression: $y = c_0 + c_1x + c_2x^2 + \ \ldots$



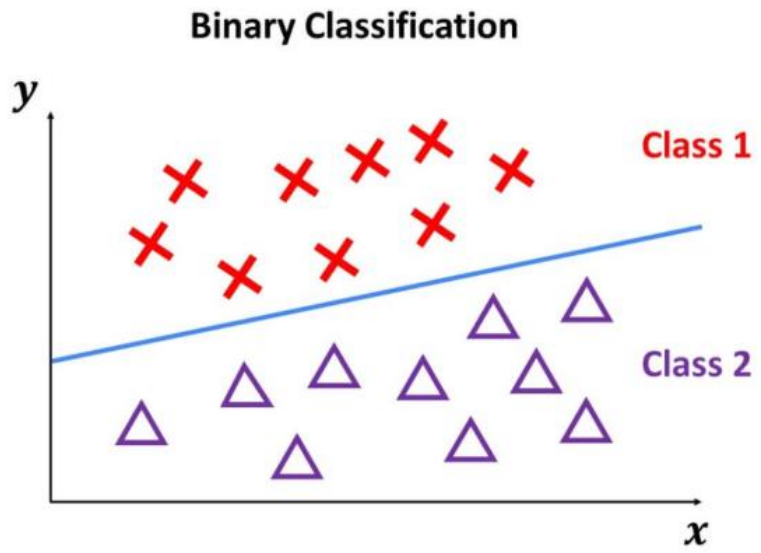
Google Colab

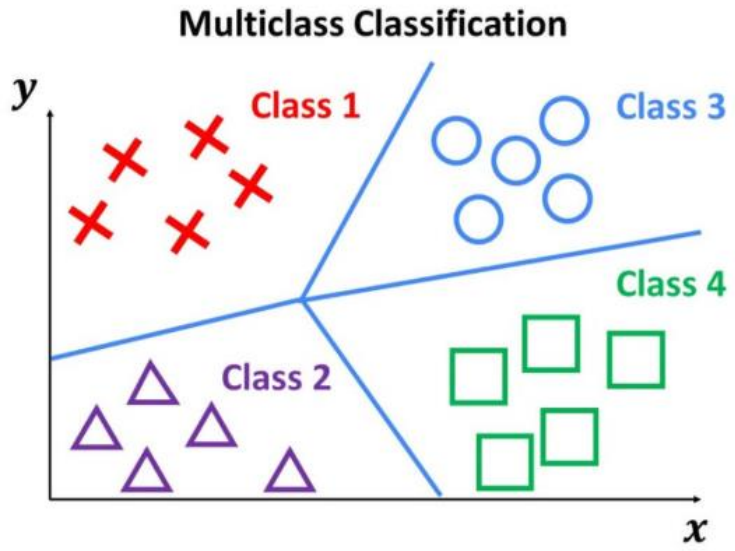
Go to https://colab.google/ and click new notebook

A complete version is here: https://tinyurl.com/VUW-SNAP-AI-1

Classification

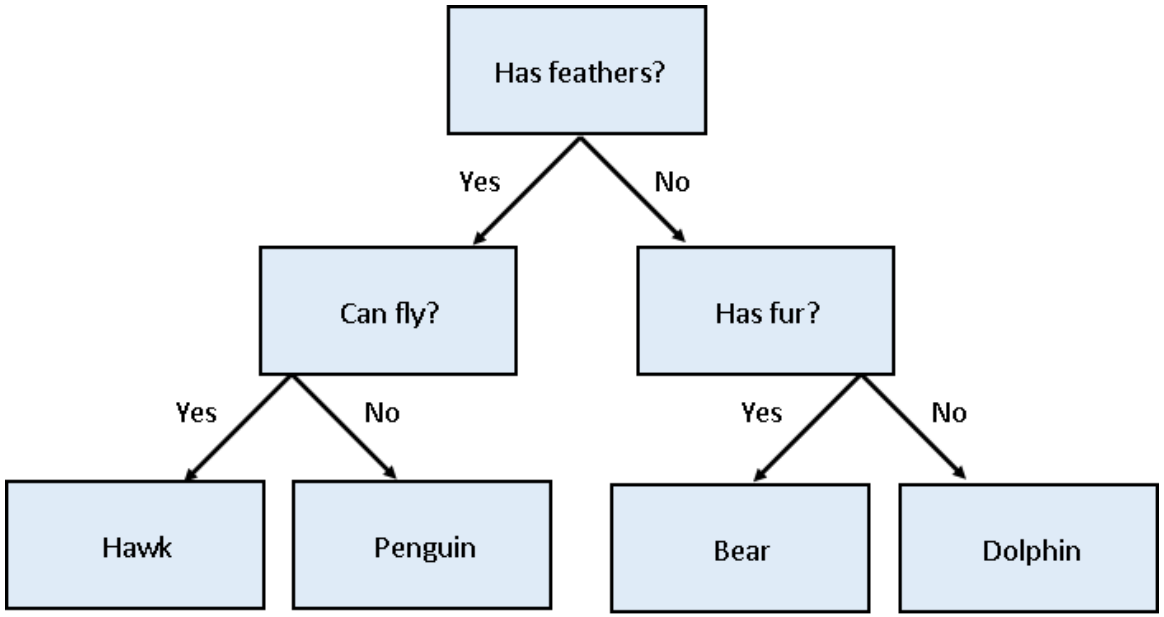
- Aim of classification: group data into labelled categories (classes)
- Binary classification: two possible classes
  - An image is a cat or dog
  - Cell is cancerous or not
- Multi-class classification: > 2 possible classes
  - A galaxy is spiral, elliptical, or irregular
  - A rock is igneous, sedimentary, or metamorphic





Classification Models

- Decision Trees: similar to a flow diagram. Split classification problem into a binary tree of comparisons.



- *k*-Nearest Neighbours (kNN): Classify a data point based on the class of its *k* nearest neighbours.