Science and Technology Facilities Council

Introduction to machine learning

Associated notebook: https://github.com/keeeto/reading-ml-chemistry/blob/master/01_classification_decision_tree.ipynb

Course overview

- Intro to machine learning
- Classical Machine Learning
- Deep Machine Learning
- Notebooks to work through the lectures

What you (don't) need

- Don't need
  - Strong mathematical background
  - Any particular computer programming experience
- Do need
  - A google account
  - Curiosity about the content

Overview Today(ish)

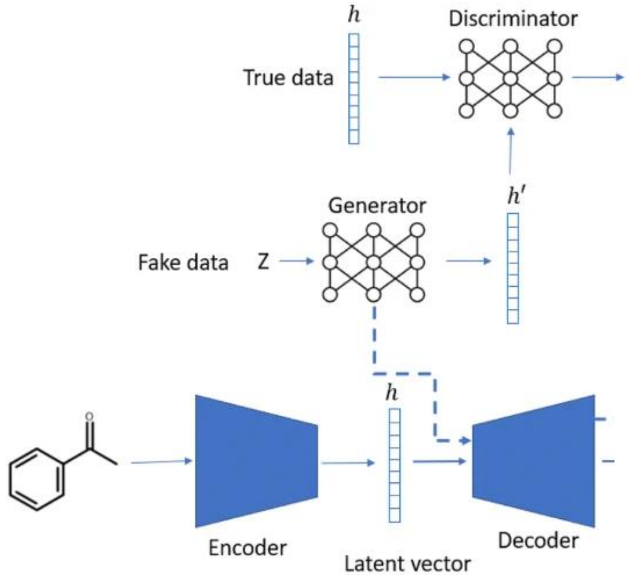
- Define ML
- Types of ML
- Parameters and hyperparameters
- Features
- Decision trees

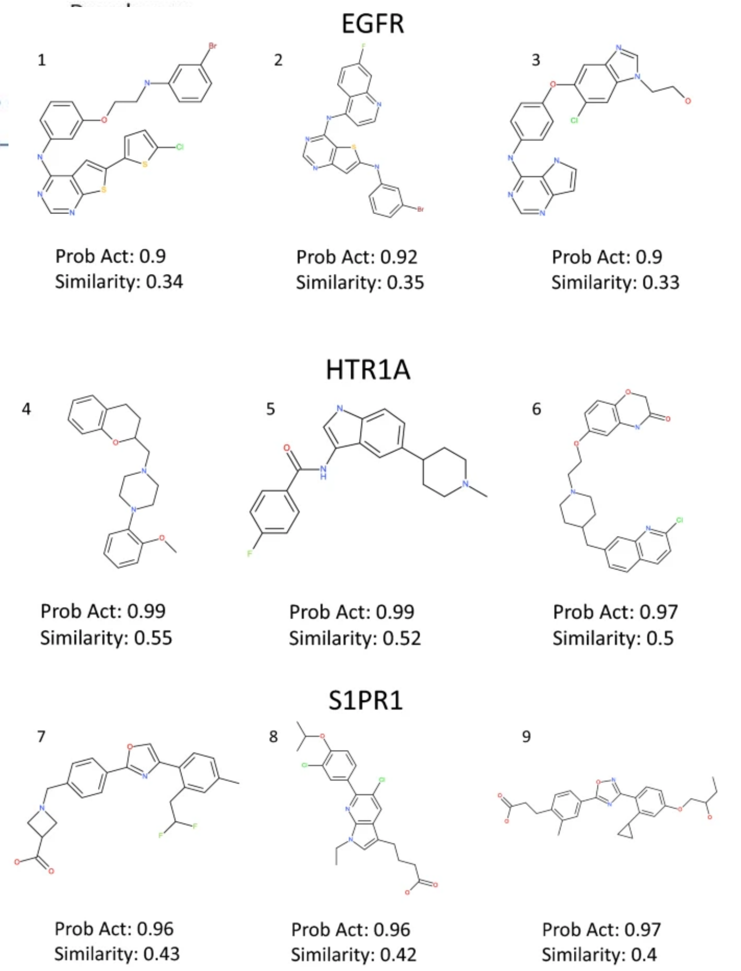
Setting up a notebook

- You will need a Google account to do this
- Go to https://colab.research.google.com/
- Search for https://github.com/keeeto/reading-ml-chemistry

Showcase

- A model is trained to generate new drug molecules from scratch
- Trained to design molecules which are drug-like but have not been tested before
- The method can develop previously un-explored molecules
- Probable activities are promising



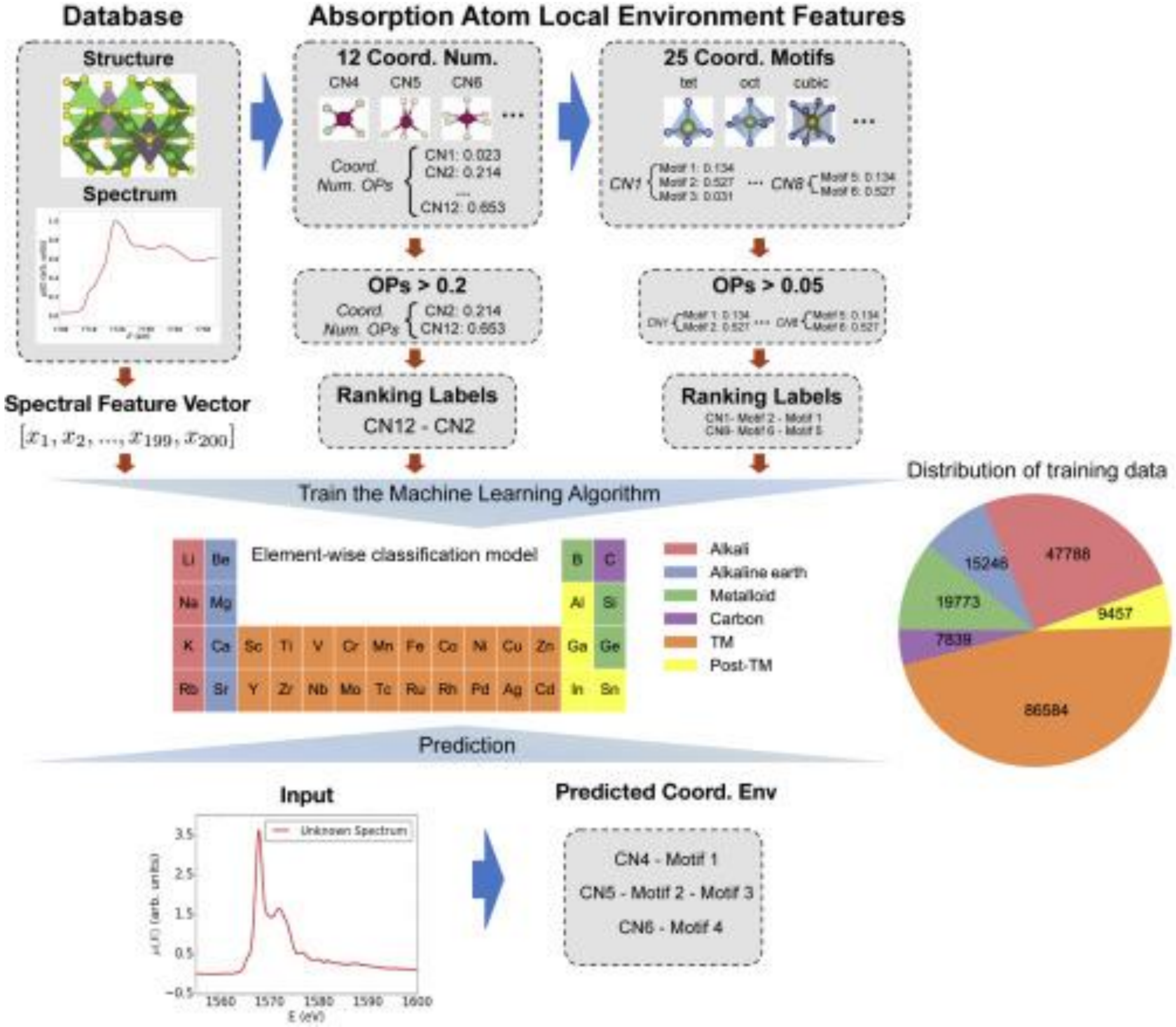


Science and Technology Facilities Council

Journal of Cheminformatics **volume 11**, Article number: 74 (2019)

Showcase

- Directly predicts the atomic environment labels from the X-ray absorption near-edge structure
- Accuracy exceeding 80%
- Accelerated or even on-the-fly interpretation of spectra directly from experiments



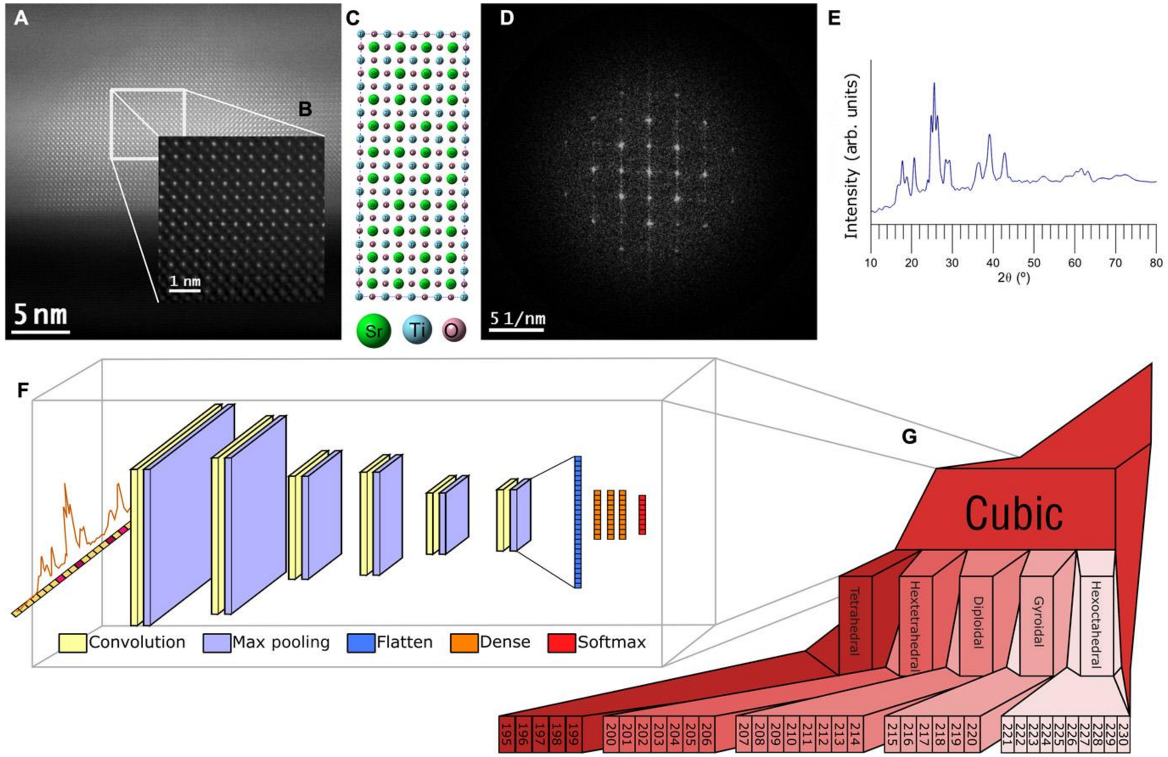
Science and Technology Facilities Council

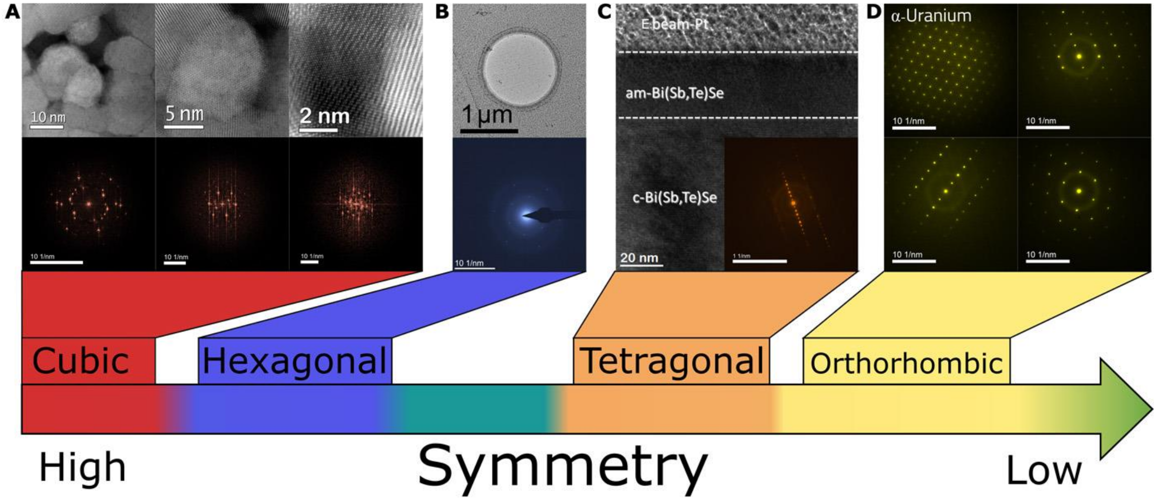
Patterns, 2020, 1, 100013

Showcase

- Interpreting electron diffraction crystallography using supervised ML
- Models trained on simulated diffraction patterns
- Can narrow down possible spacegroups to the top two with 95% confidence
- Even peaks in low signal-to-noise images can be potentially used

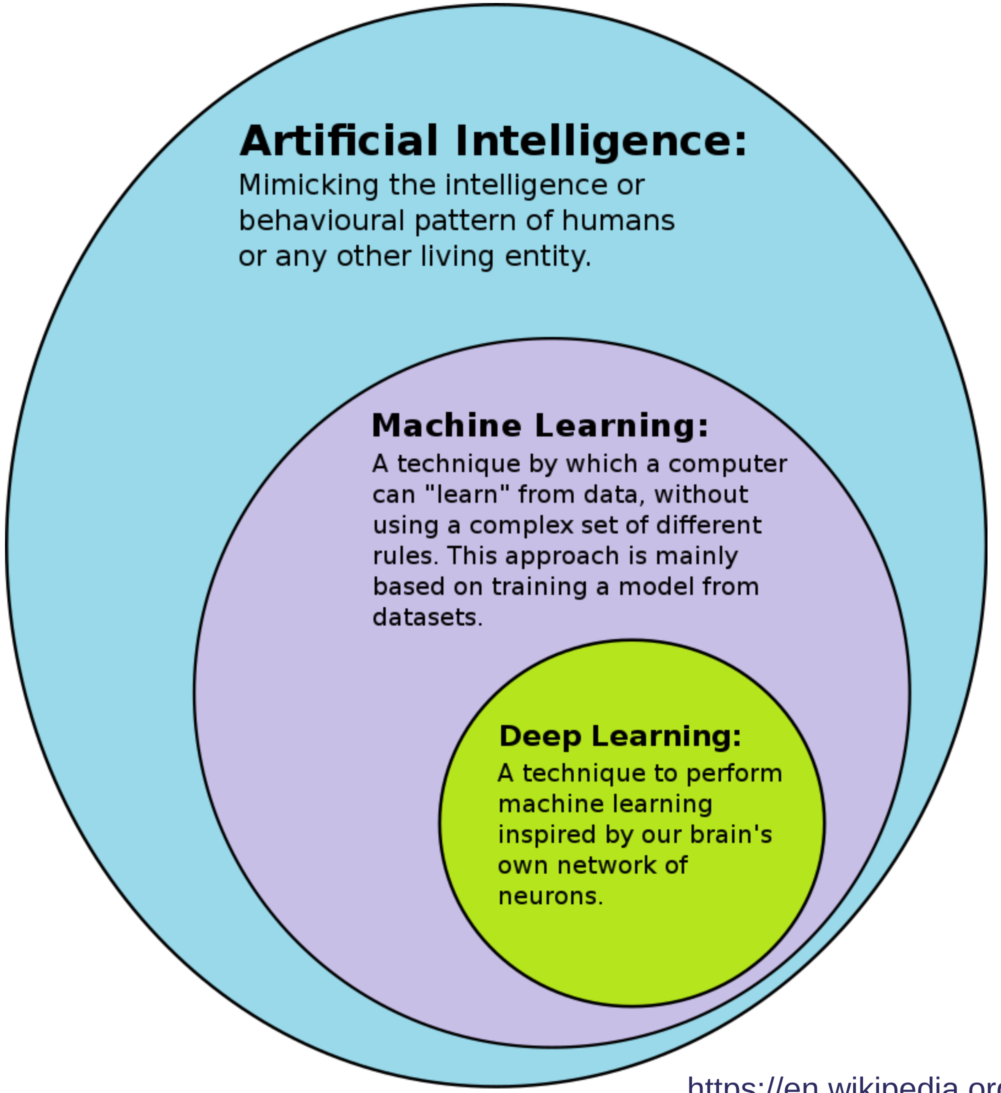
Science and Technology Facilities Council





Science Advances 2019, 5, eaaw1949

What is machine learning?



Science and Technology Facilities Council

https://en.wikipedia.org/wiki/Deep_learning#/media/File:AI-ML-DL.svg

A working definition

ML = Representation + Evaluation + Optimisation

DOI:10.1145/2347736.2347755

Tapping into the "folk knowledge" needed to advance machine learning applications.

BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning

Science and Technology Facilities Council

"A Few Useful Things to Know About Machine Learning" by Pedro Domingos

A working definition

ML = Representation + Evaluation + Optimisation

- Representation
    - How we represent the knowledge.
    - What type of model do you use.
    - What data do you use in what format
    - Hypothesis space.
    - Eg. Neural network, decision tree …

A working definition

ML = Representation + Evaluation + Optimisation

- Evaluation
  - Objective function or scoring function.
  - Distinguish good from bad models.

A working definition

ML = Representation + Evaluation + Optimisation

- Optimisation
  - Searches between models.
  - Updates the parameters of a model to improve performance.
  - Identifies the highest-scoring one.
  - Determines the efficiency of a learner.

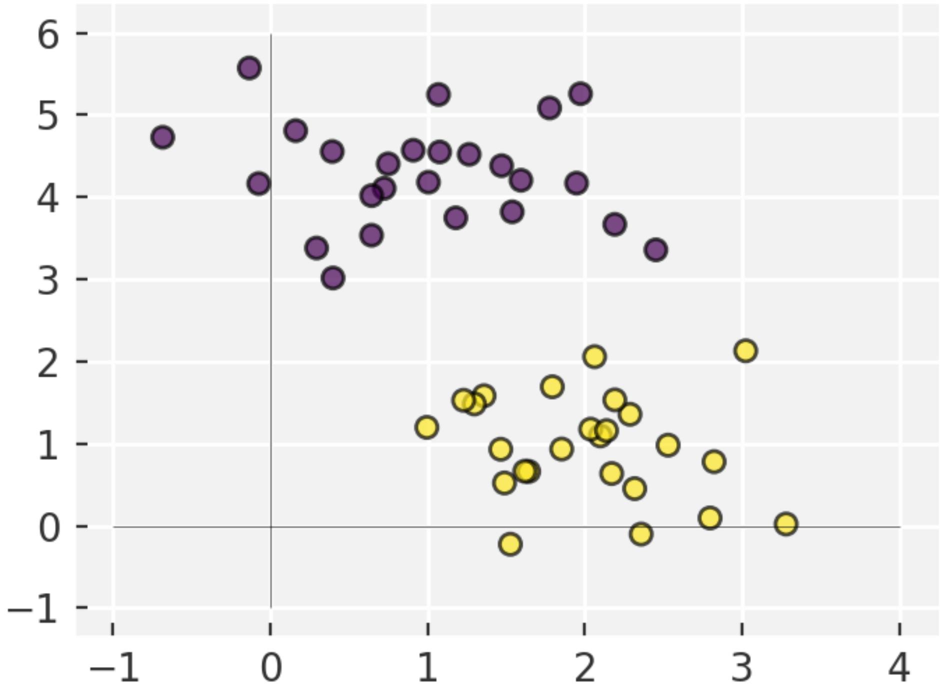
A working definition

ML = Representation + Evaluation + Optimisation

- Representation
  - Choice of model
  - Choice of hyper-parameters
  - Choice of features

Supervised ML

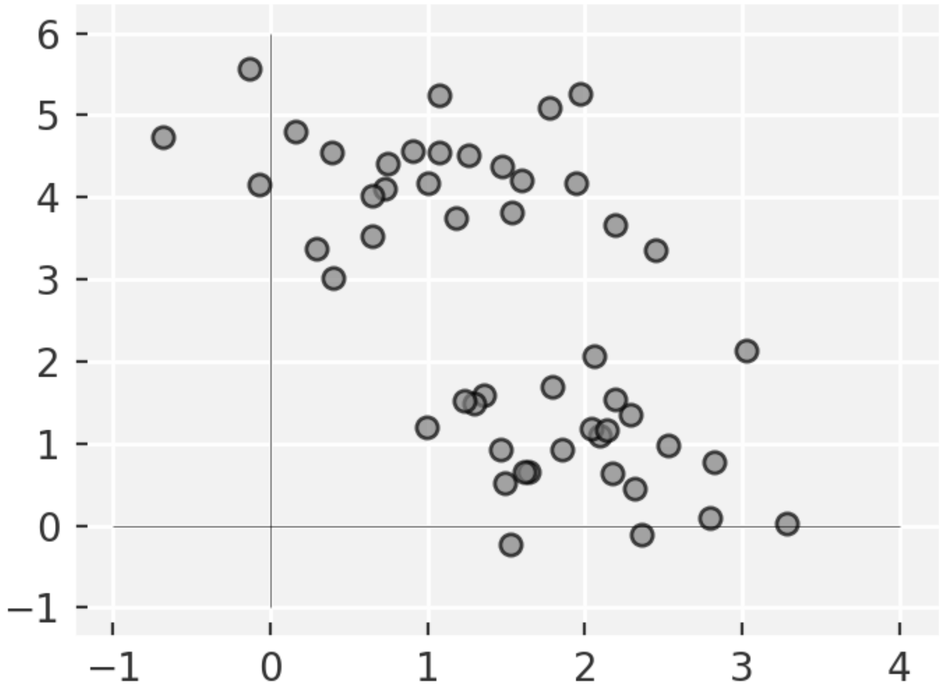
- Data plus labels
- **Learning** a function that maps an input to an output based on example input-output pairs.



Science and Technology Facilities Council

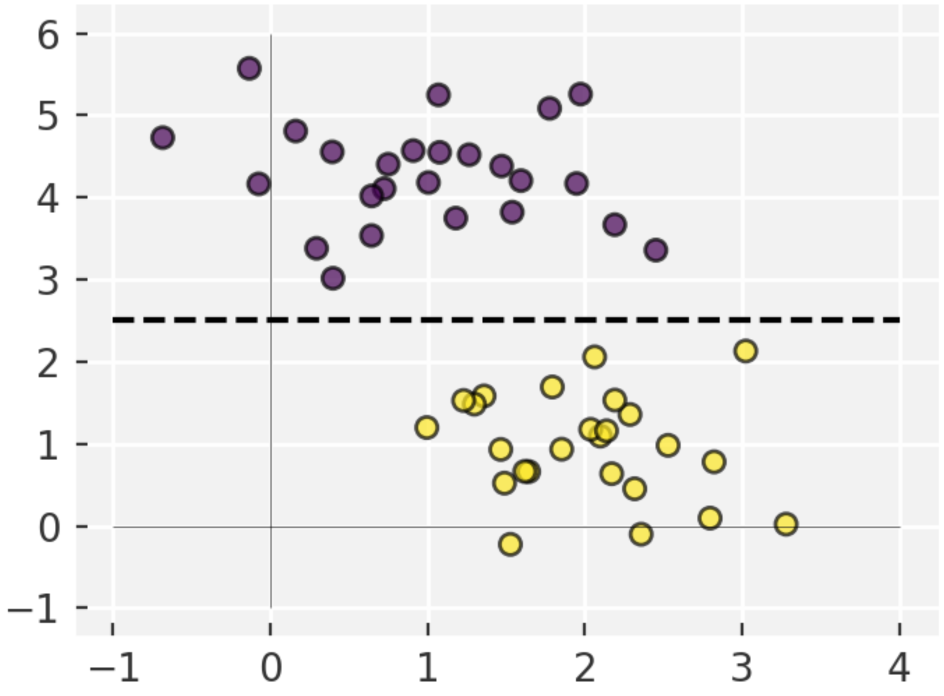
Unsupervised ML

- Data do not have labels
- Identifying trends in unlabelled datasets
- E.g. cluster analysis, is used for exploratory data analysis to find hidden patterns or grouping in data



Classification

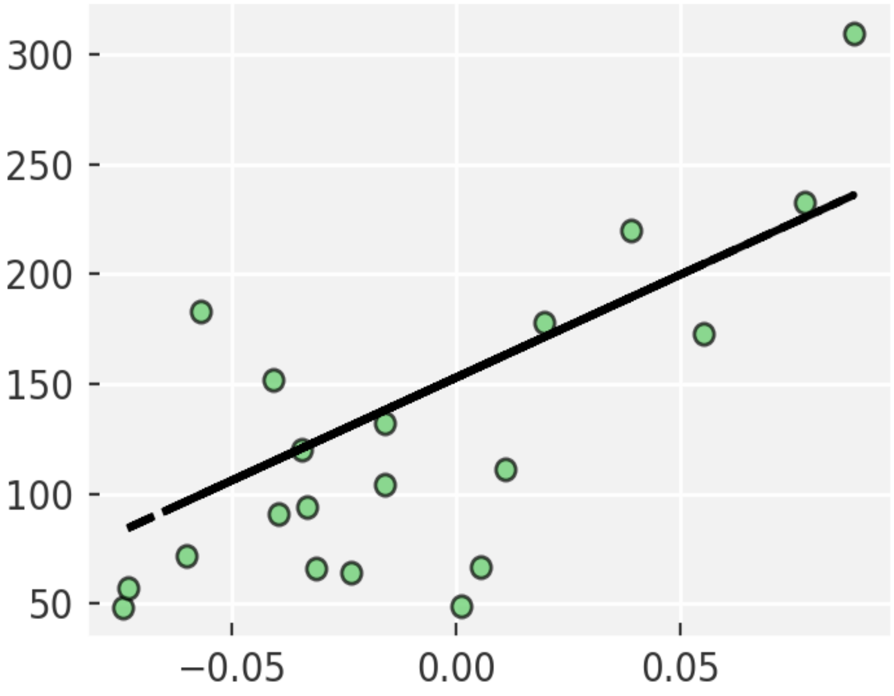
- Identifying to which of a set of categories a new sample belongs, on the basis of a training set
- E.g. spam filter or which crystal structure gives a certain pattern



Science and Technology Facilities Council

Regression

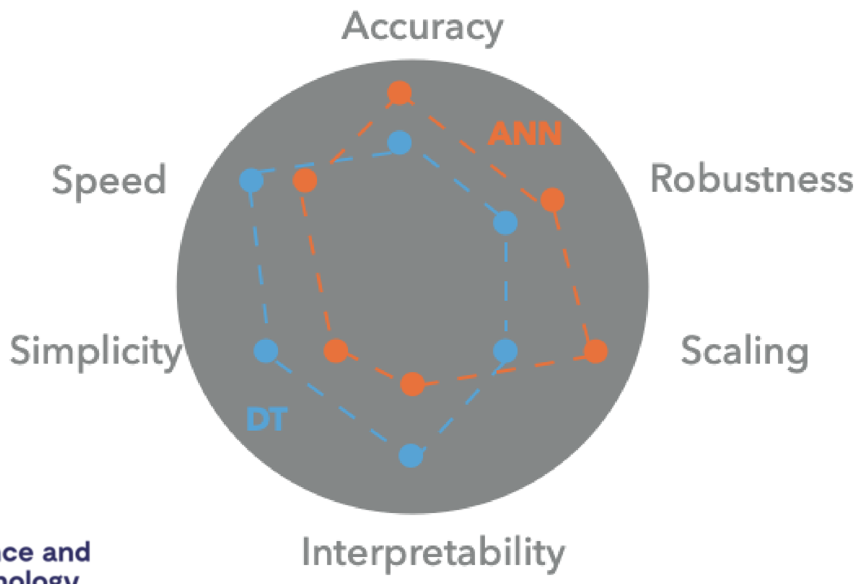
- Models a target prediction value based on independent variables

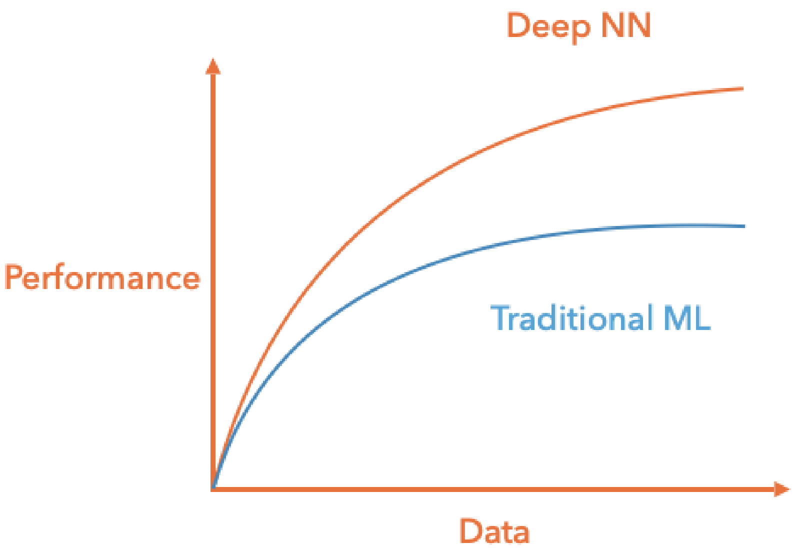


Science and Technology Facilities Council

Classical/deep methods

- Classical: linear regression, trees etc..
- Deep: neural network type models





Parameters and hyper-parameters

- Parameters – properties of the model that are modified during training
- Hyperparameters – set of values that define the model and how it trains. Do not update during training
  - E.g. loss function, learning rate, number of parameters

Features

- In ML approaches the data will typically consist of several or more features
- Features are simply the input variables for the model – x in f(x) = y

> "…some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used."

Science and Technology Facilities Council

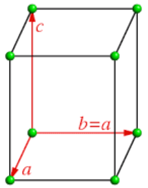
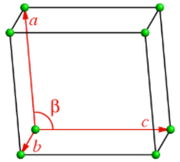
"A Few Useful Things to Know About Machine Learning" by Pedro Domingos

Feature engineering

- Transforming raw data into features that better represent the underlying problem
- Make inputs into things that an algorithm can understand
- E.g. Convert a date-time stamp into something more useful 2014-09-20T20:45:40Z -> Day: Tuesday; Year: 2014; Month: Sept
- Note that ‘Tuesday’ and ‘Sept’ are not particularly algorithm ready – how can we convert them to something more useful?

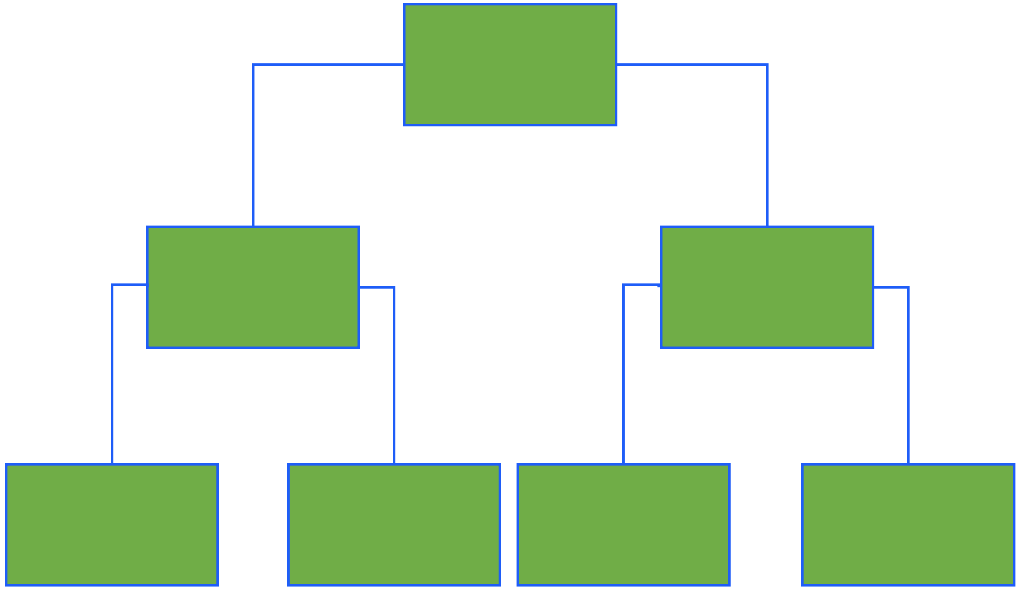
One hot encoding

- Classification problems
- Vector of length = number of categories
- Each element is the probability that the data represents a given class

Material	Ortho	Rhomb
	1	0
	0	1

Science and Technology Facilities Council

Decision trees



Data is split by features. E.g. brightness of a pixel
Splits are arranged such that the data splits as evenly as possible at each point.

Decision trees

$$Q_{left}(\theta) = (x, y) | x_f \leq t_j$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$

Data is split according to a threshold value tj.

$$C(Q, \theta) = \frac{n_{left}}{N_j} H(Q_{left}(\theta)) + \frac{n_{right}}{N_j} H(Q_{right}(\theta))$$

The cost of the split is calculated based on some impurity function H() e.g. RMSD of the data.

$$\theta^* = \underset{\theta}{\operatorname{argmin}} C(Q, \theta)$$

The splitting parameters are chosen to minimise C at each split.

Science and Technology Facilities Council

Go to notebook

Concept checklist

- Supervised/unsupervised machine learning
- Classical machine learning/deep learning
- Parameters/hyperparameters
- Features and feature engineering
- Decision trees