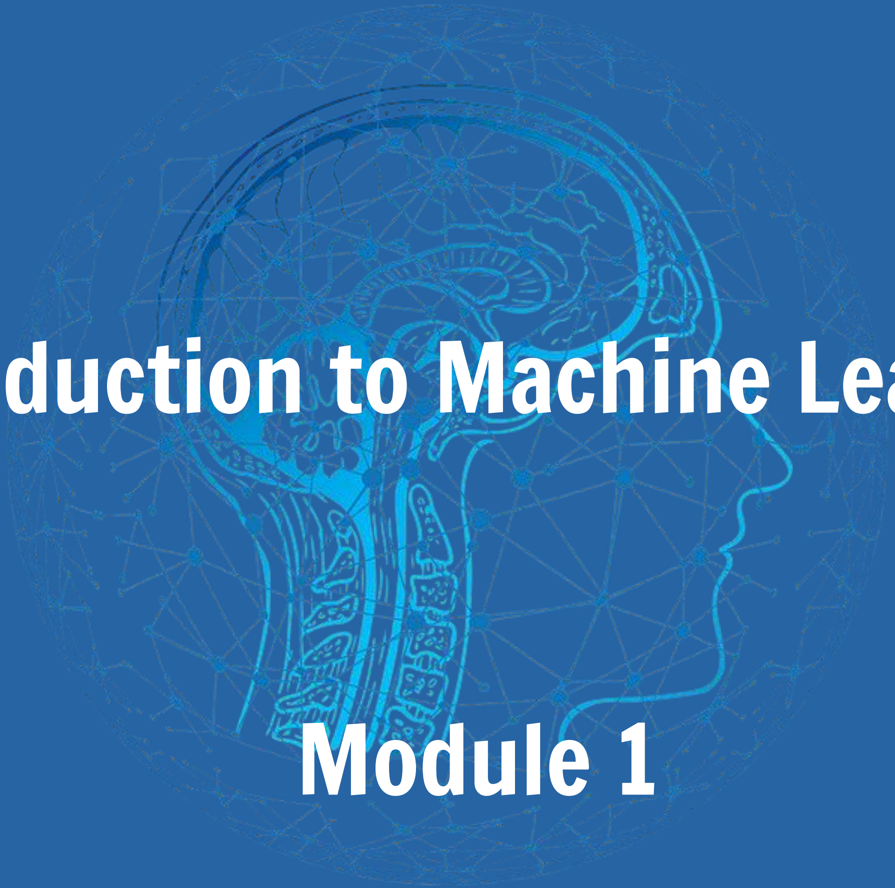
Imperial College London



Introduction to Machine Learning

Module 1

What is Artificial Intelligence?

How does it relate to Machine Learning?


WALL-E

Acting Humanly

"The art of creating machines that perform functions that require intelligence when performed by people."

(Kurzweil, 1990)

"The study of how to make computers do things at which, at the moment, people are better."

(Rich & Knight, 1991)

Turing test

Chinese room

Acting Rationally

"Computational Intelligence is the study of the design of intelligent agents."

(Poole et al.,1998)

"AI . . . is concerned with intelligent behavior in artifacts."

(Nilsson, 1998)

Thinking Rationally

"The study of mental faculties through the use of computational models."

(Charniak and McDermott, 1985)

"The study of the computations that make it possible to perceive, reason, and act."

(Winston, 1992)

Thinking Humanly

“The exciting new effort to make computers think ... machines with minds, in the full and literal sense.”

(Haugeland, 1986)

“[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning ...”

(Bellman, 1978)

Artificial Intelligence

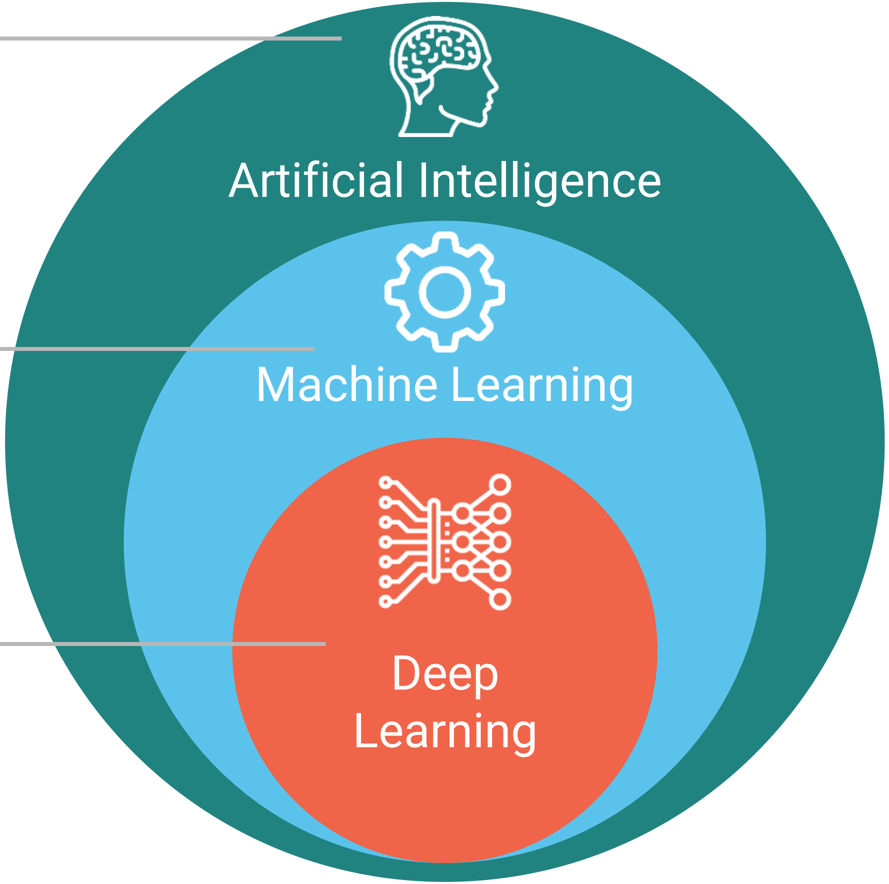
Techniques that enable computers to mimic human behaviour and intelligence. Could be using logic, if-then rules, machine learning, etc.

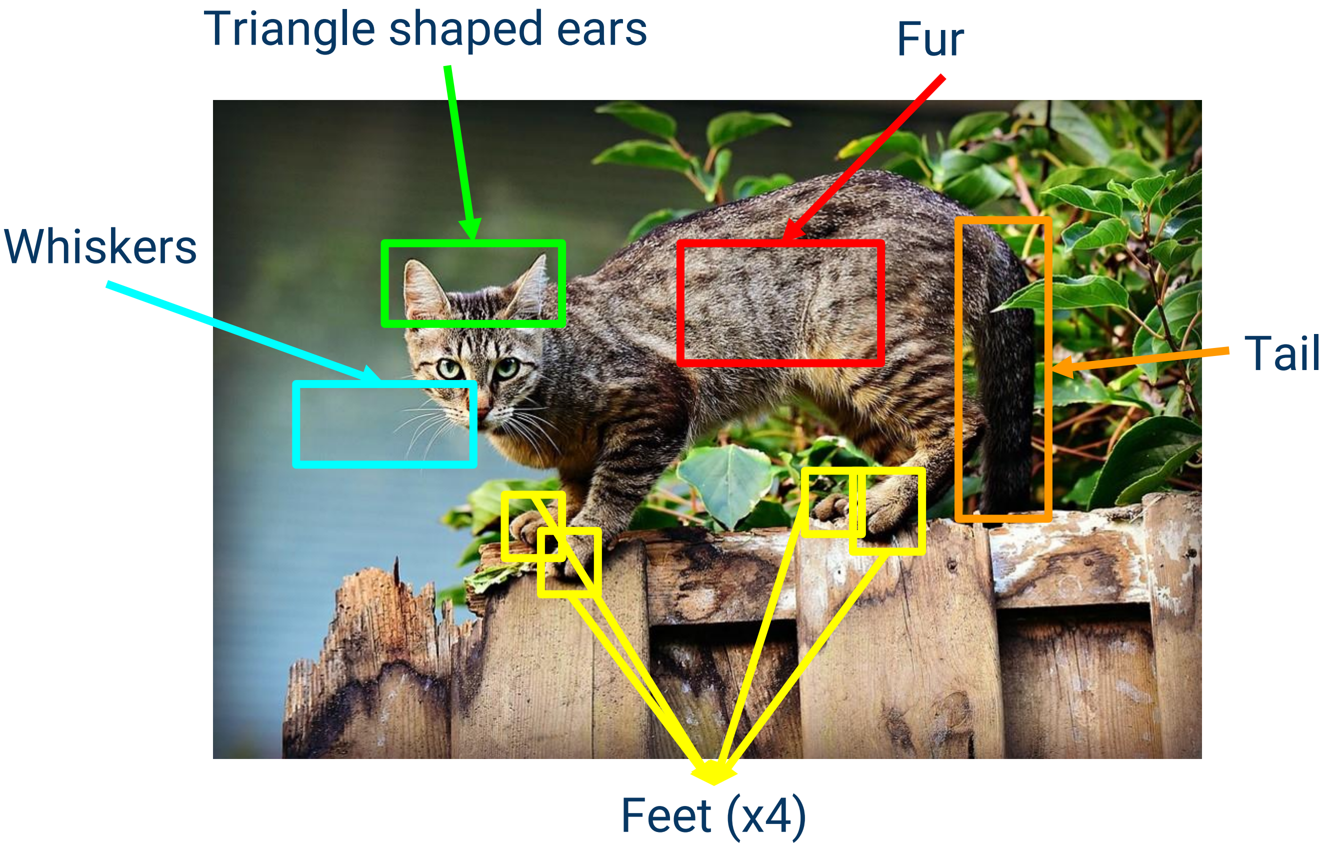
Machine Learning

Subset of AI techniques using statistical methods that enable the systems to learn and improve with experience.

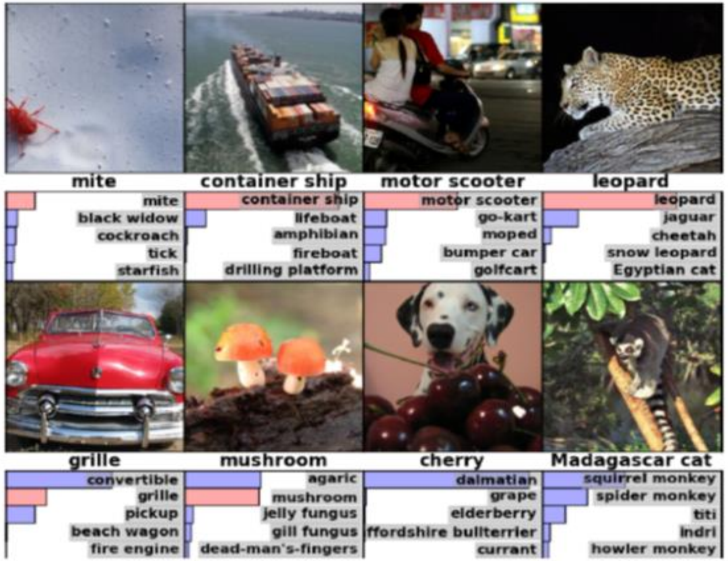
Deep Learning

Subset of machine learning techniques using multi-layer artificial neural networks and vast amounts of data for learning.

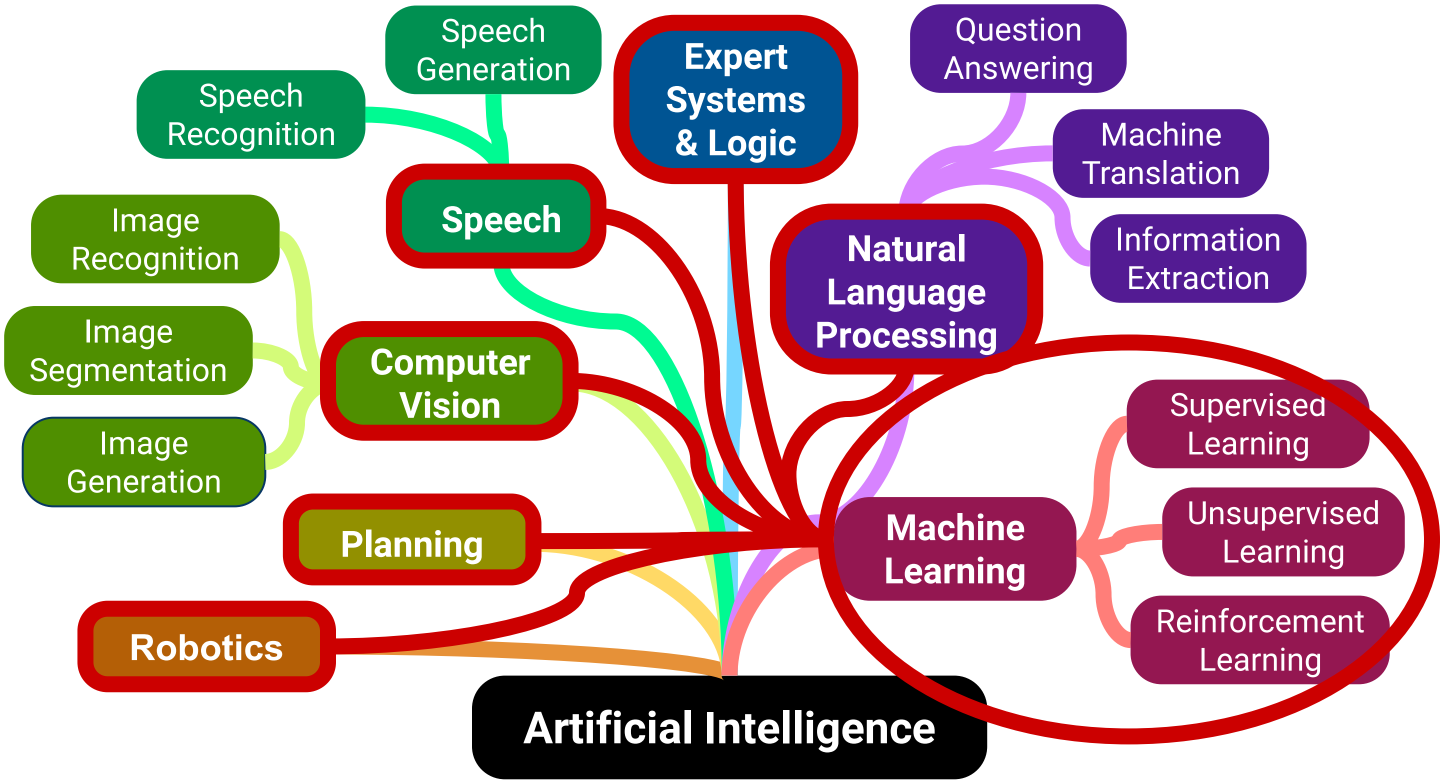



Triangle shaped ears
Fur
Whiskers
Tail
Feet (x4)

Monarch butterfly



Generalisation


Speech Generation
Speech Recognition
Expert Systems & Logic
Question Answering
Machine Translation
Information Extraction
Image Recognition
Speech
Natural Language Processing
Image Segmentation
Computer Vision
Supervised Learning
Image Generation
Unsupervised Learning
Planning
Machine Learning
Reinforcement Learning
Robotics
Artificial Intelligence


ACTIVE
10
Yielding to an
occluded pedestrian

Application: Helping doctors

AI doctor could boost chance of survival for sepsis patients

by Kate Wighton
22 October 2018

Scientists have created an artificial intelligence system that could help treat patients with sepsis.

The technology, developed by researchers from Imperial College London, was found to predict the best treatment strategy for patients.

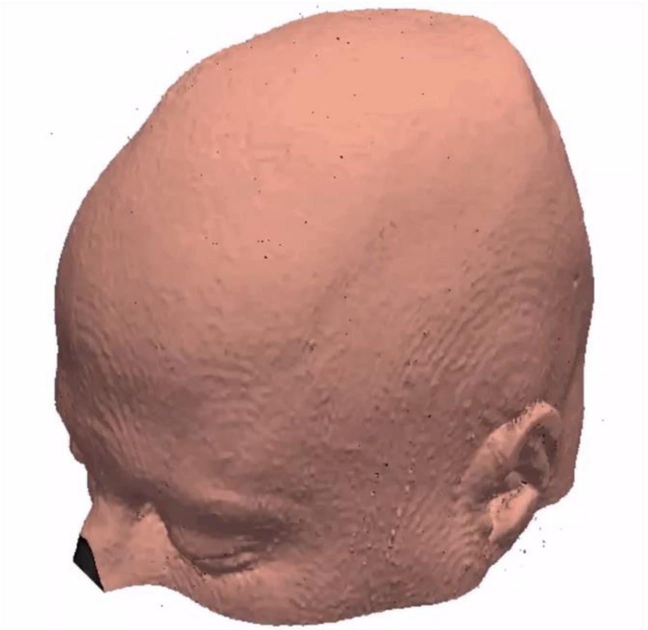
> Our new AI system was able to analyse a patient's data – such as blood pressure and heart rate – and decide the best treatment strategy.
>
> – ***Dr Aldo Faisal***
> *Study author*

The system 'learnt' the best treatment strategy for a patient by analysing the records of about 100,000 hospital patients in intensive care units and every single doctor's decisions affecting them.

The findings, published in the journal Nature Medicine, showed the AI system made more reliable treatment decisions than human doctors.

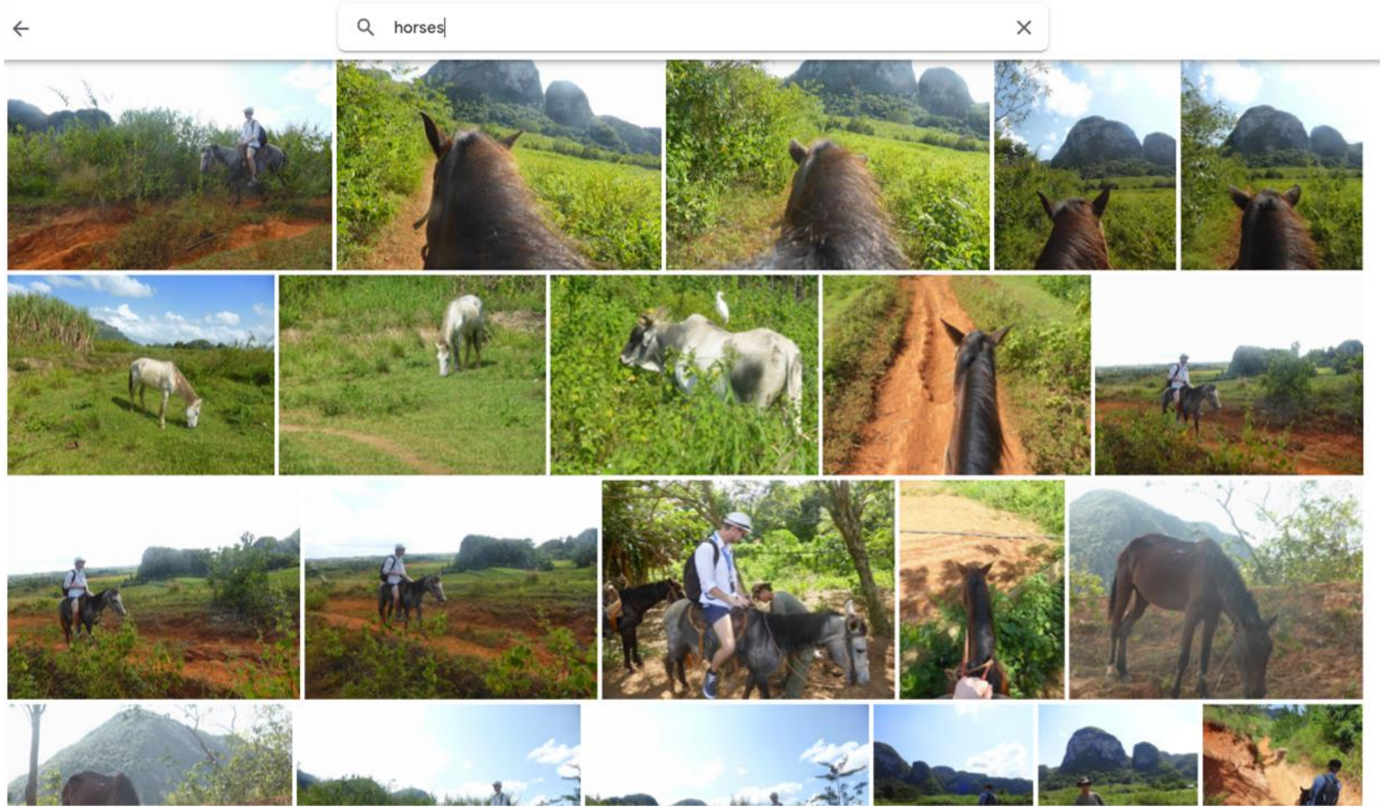
The team behind the technology

https://www.imperial.ac.uk/news/188705/ai-doctor-could-boost-chance-survival/

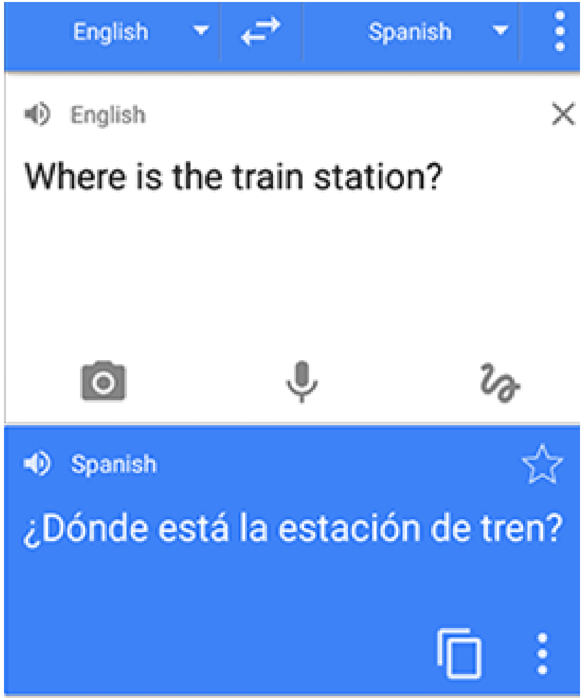


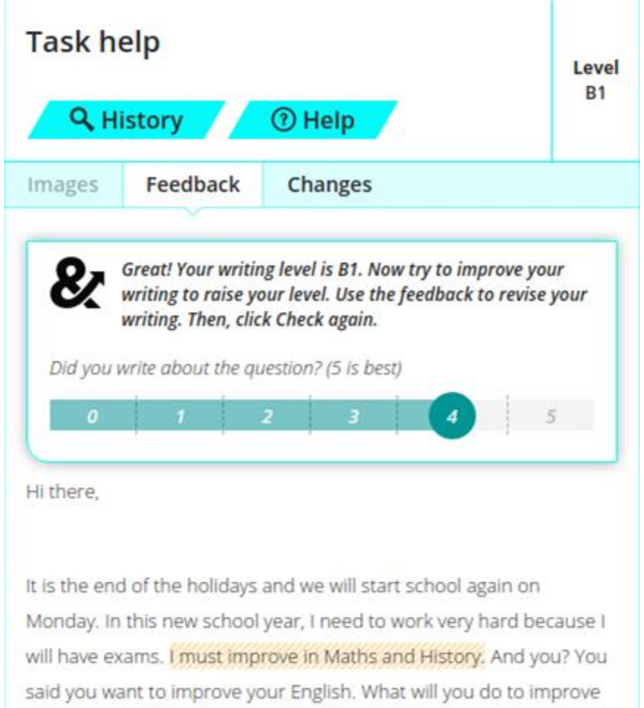
http://wp.doc.ic.ac.uk/bglocker/

Application: Analysing images



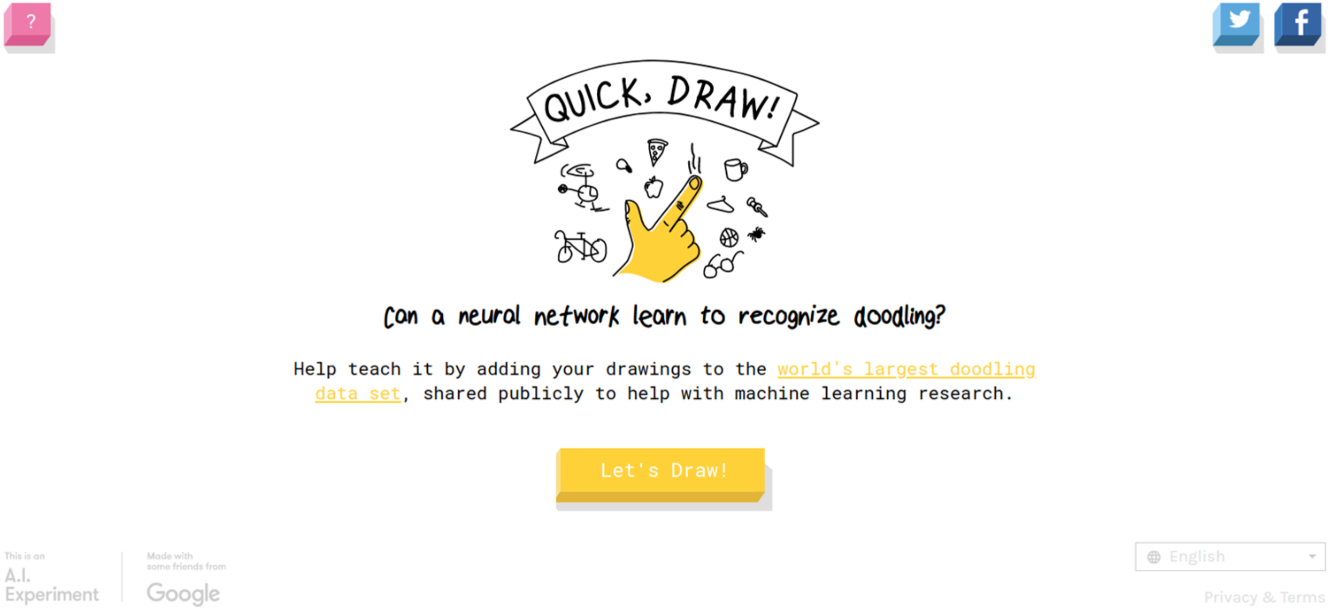
Application: Working with language





https://translate.google.com https://writeandimprove.com

Application: Quick draw



https://quickdraw.withgoogle.com

So, what *exactly* is Machine Learning?

“The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.”

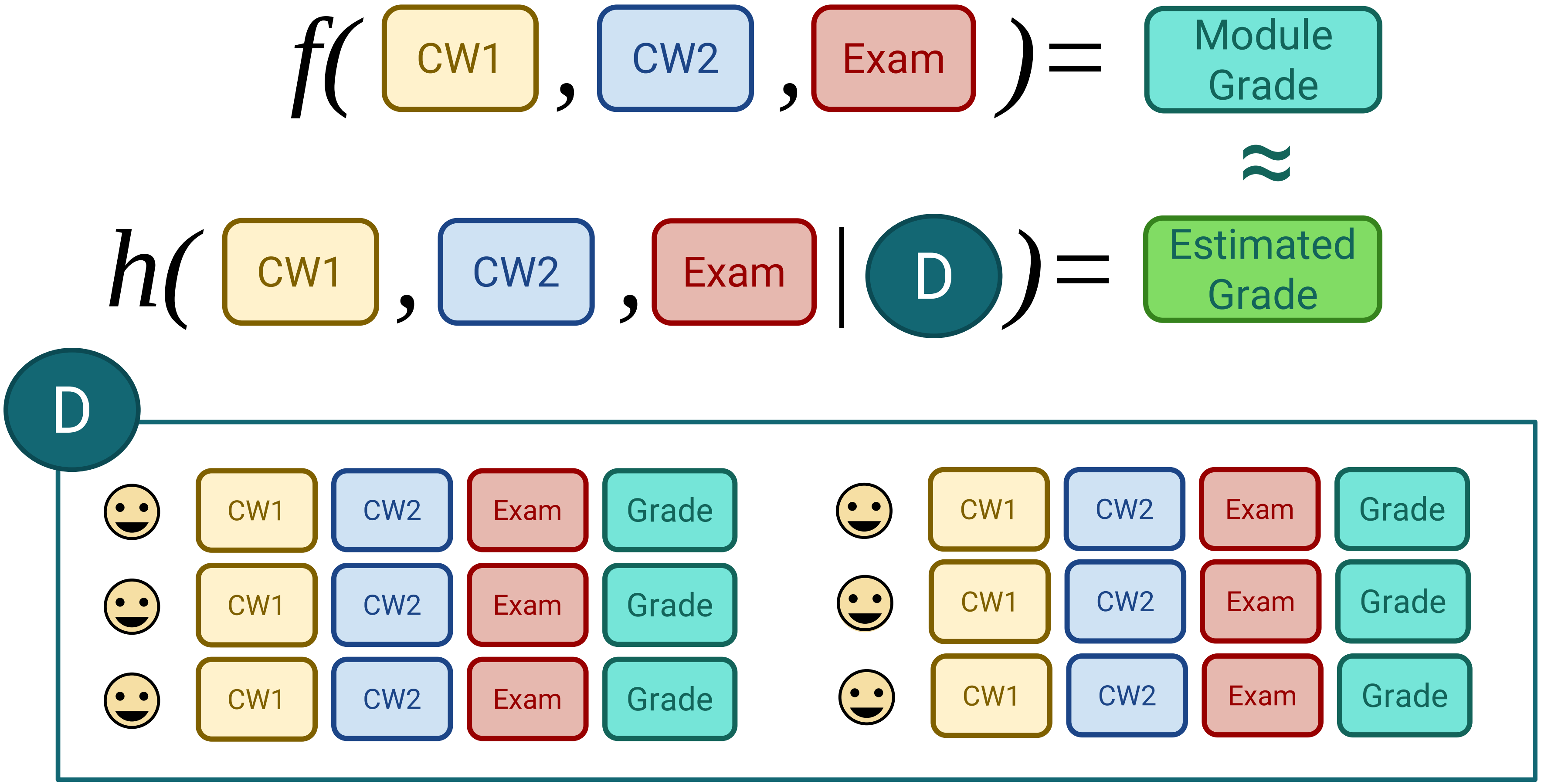
Tom Mitchell (1997)

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell (1997)

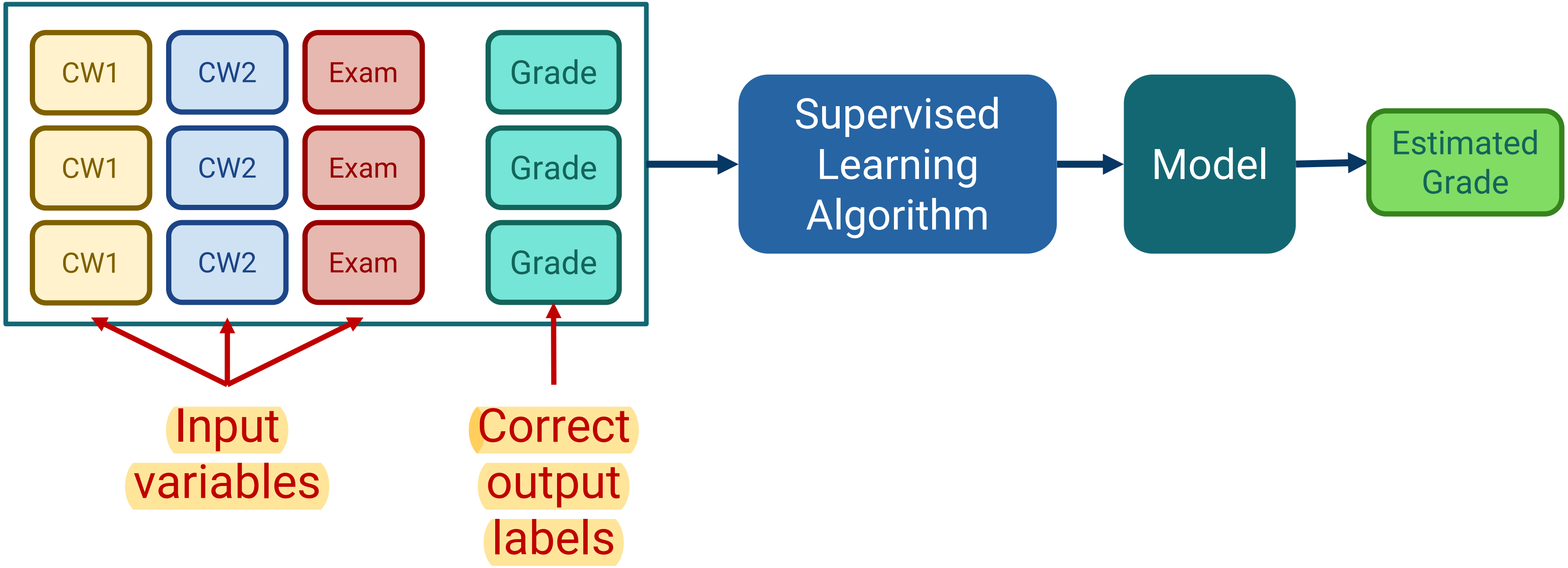
$$f(\boxed{\text{CW1}}, \boxed{\text{CW2}}, \boxed{\text{Exam}}) = \boxed{\text{Module Grade}}$$

$$0.70 \text{ x } \boxed{\text{Exam}} + 0.30 \text{ x } \boxed{\underset{\text{Coursework}}{\underbrace{0.40 \text{ x } \boxed{\text{CW1}} + 0.60 \text{ x } \boxed{\text{CW2}}}}}$$

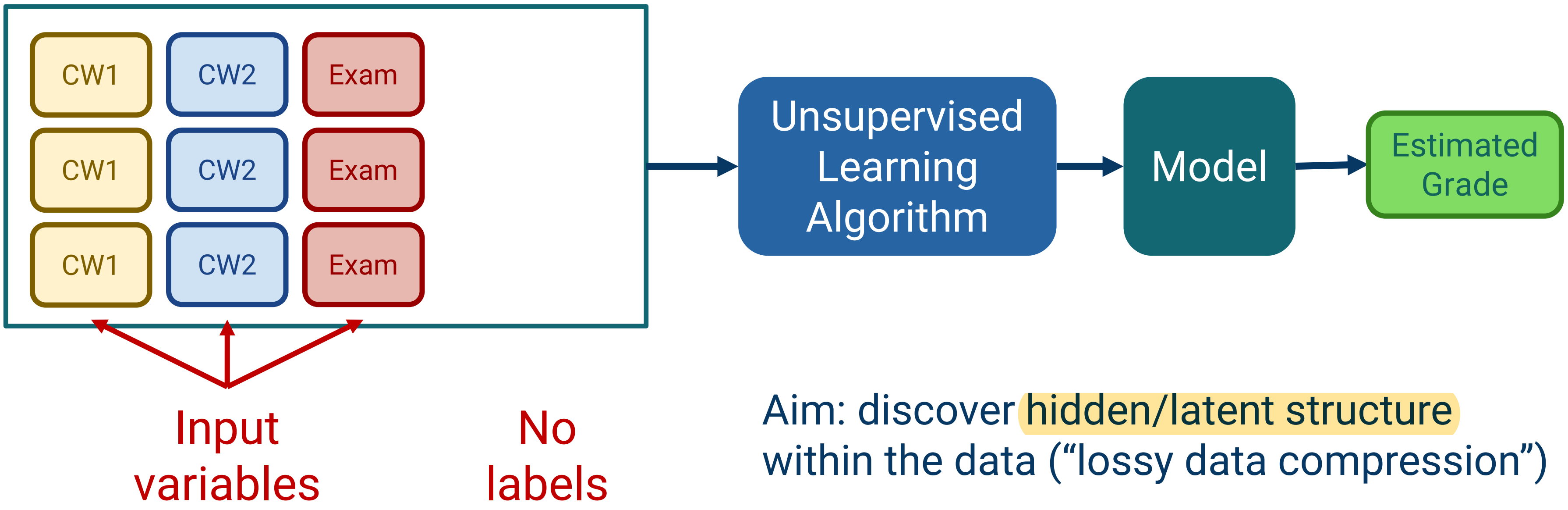

f(CW1 , CW2 , Exam) = Module Grade
≈
h(CW1 , CW2 , Exam | D) = Estimated Grade
D
CW1 CW2 Exam Grade
CW1 CW2 Exam Grade
CW1 CW2 Exam Grade
CW1 CW2 Exam Grade
CW1 CW2 Exam Grade
CW1 CW2 Exam Grade

Machine Learning settings

Supervised learning



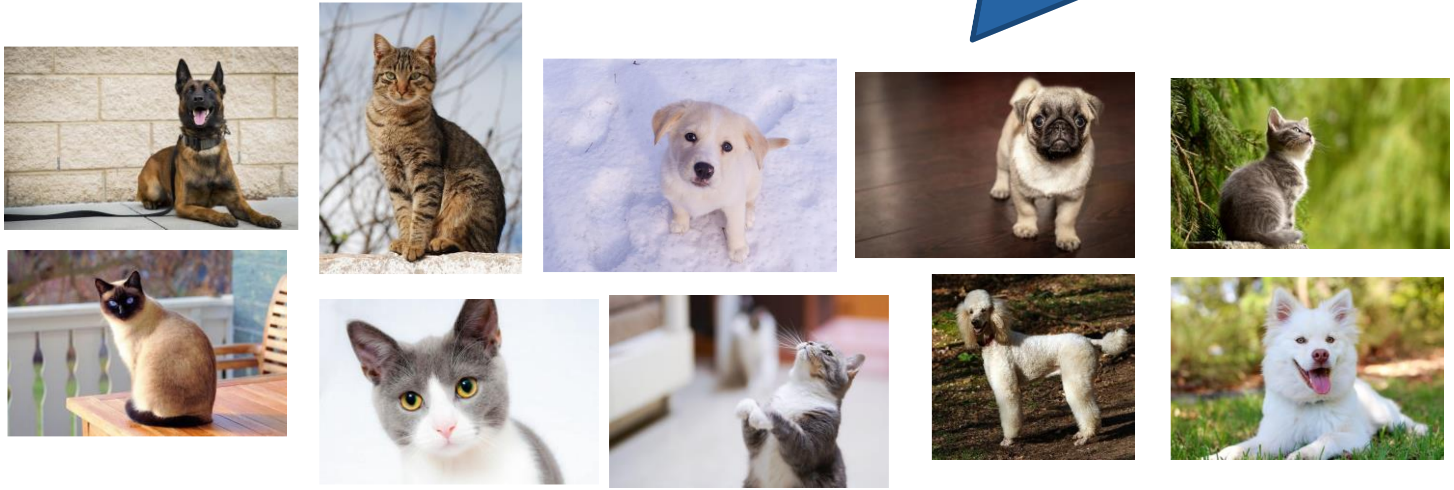
Unsupervised learning



Aim: discover hidden/latent structure within the data (“lossy data compression”)

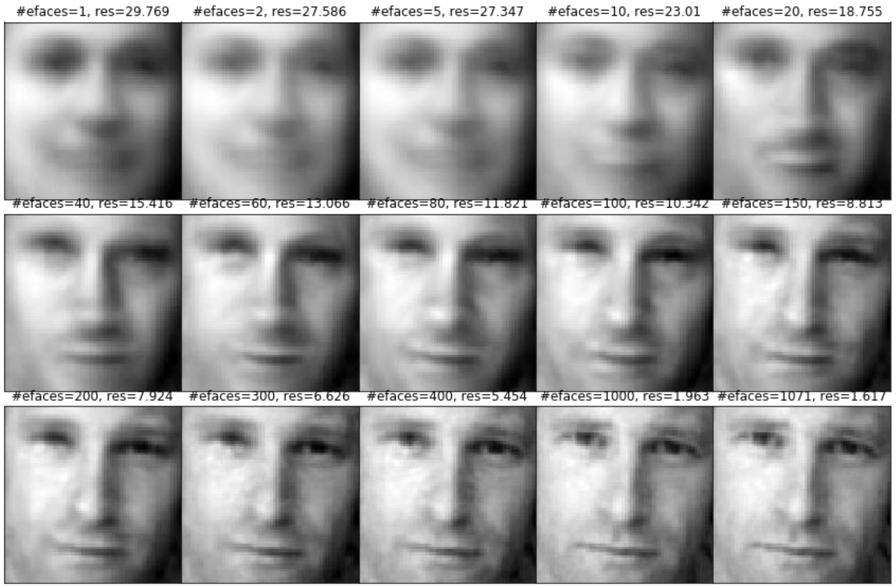
Unsupervised learning

Divide us into two groups!



Clustering

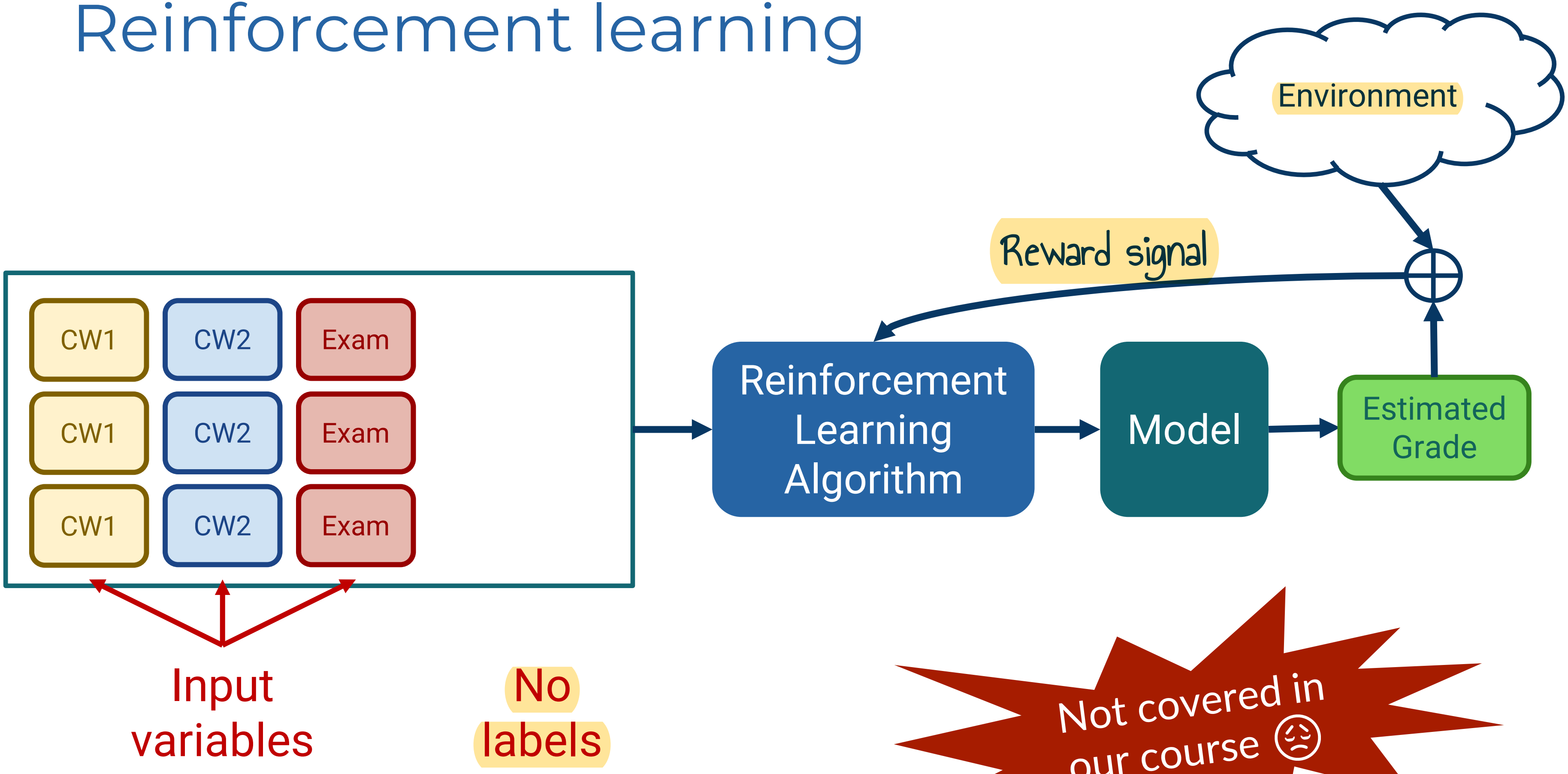
Unsupervised learning



Dimensionality reduction

Figure from https://sandipanweb.wordpress.com/2018/01/06/eigenfaces-and-a-simple-face-detector-with-pca-svd-in-python/

Reinforcement learning



Not covered in our course 😟

Reinforcement learning

DQN from DeepMind

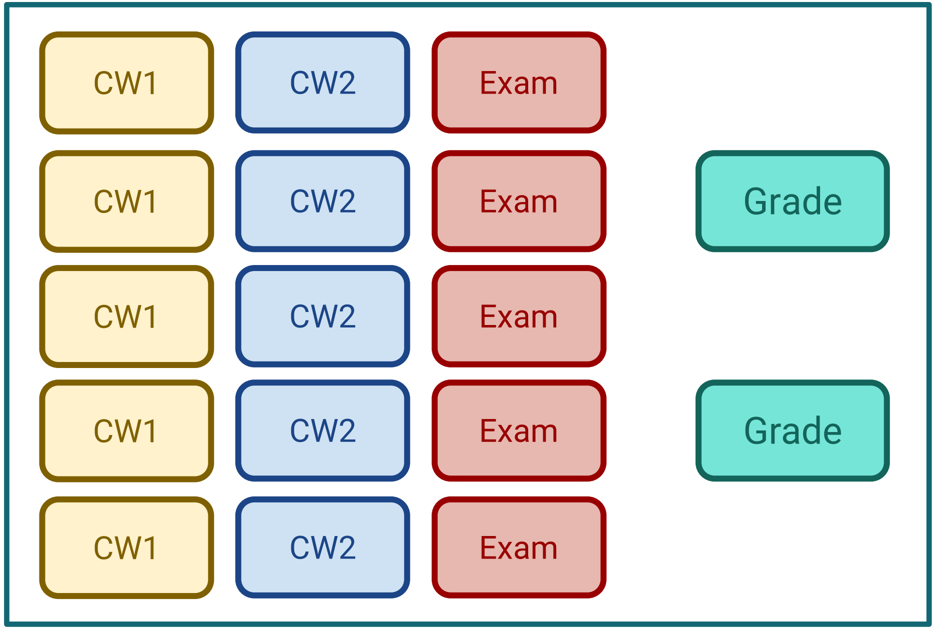
Kormushev et al., 2010

Policy search

Find which action an agent should take, depending on its current state, to maximise the received rewards.

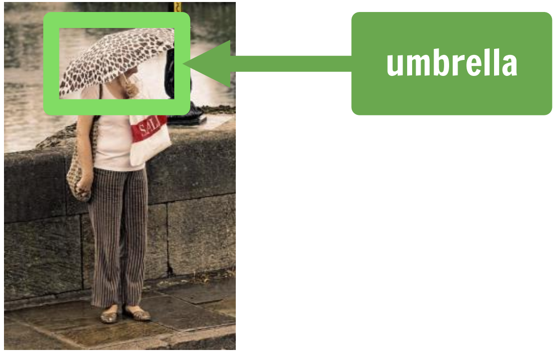
Things are not always clear cut!

- Semi-supervised learning
  - Some data have labels, some do not

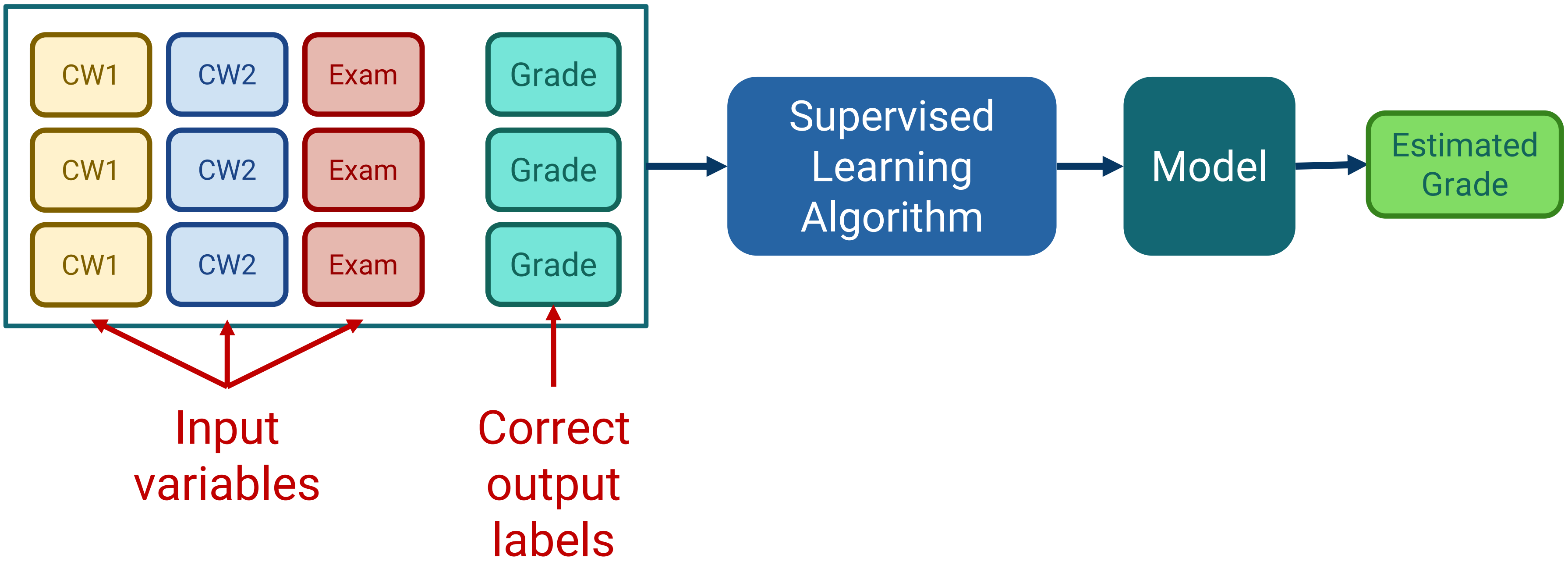


- Weakly-supervised learning
  - Inexact output labels

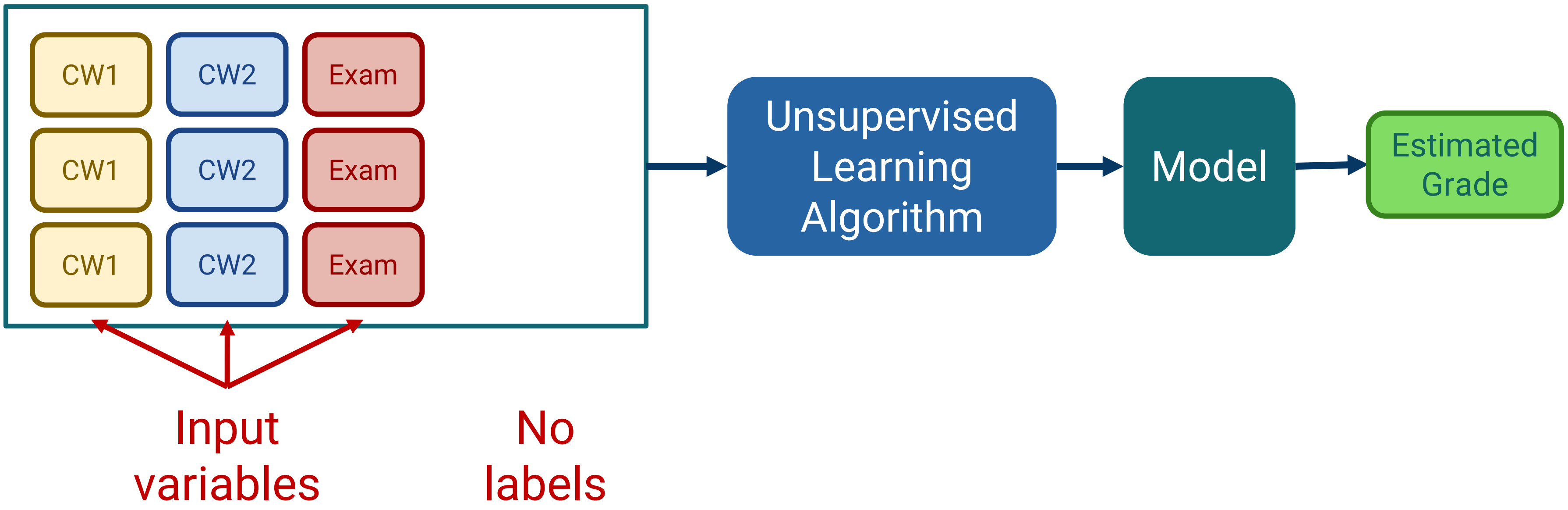
Without giving the actual location of umbrella in the picture.



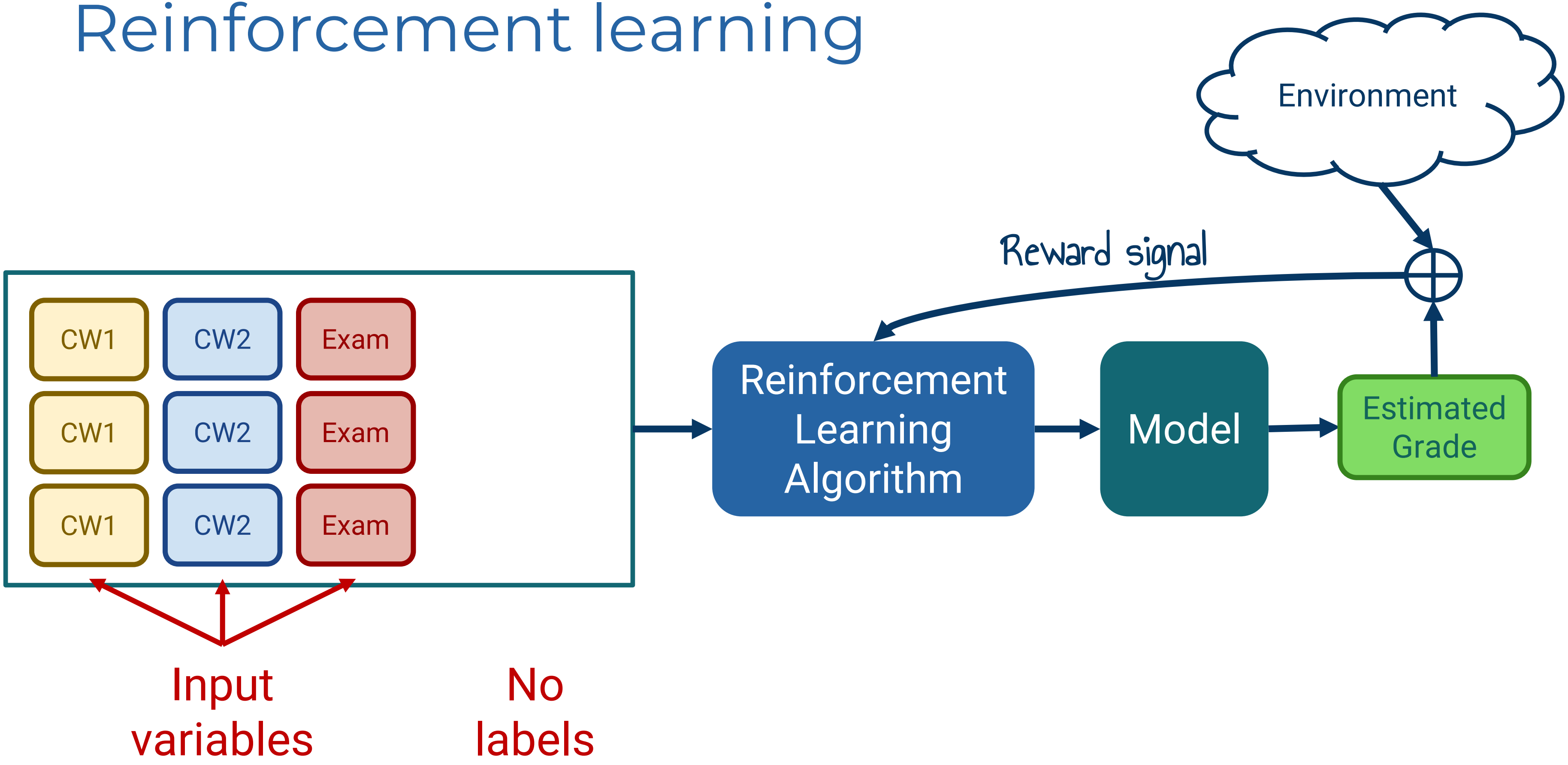
Supervised learning



Unsupervised learning



Reinforcement learning



Classification and regression

The two most popular ML tasks

Classification

Most common!

Classifier

butterfly

$f($  $) = y$

Discrete/ categorical

Classification

pneumonia

no pneumonia

Claim *(by Minister Shailesh Vara)*

"The average criminal bar barrister working full-time is earning some £84,000."

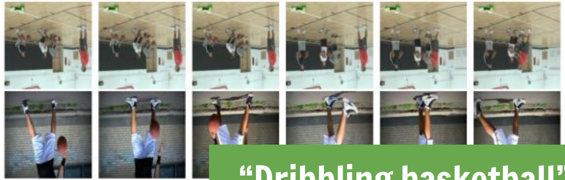
Claim is False

[redacted]
@[redacted]

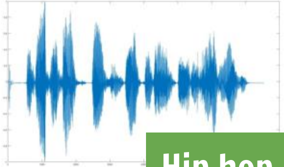
Lets kill [redacted] and kill them for fun
#kill[redacted]
7/20/14, 8:05 AM

Hate speech

Photoshopped



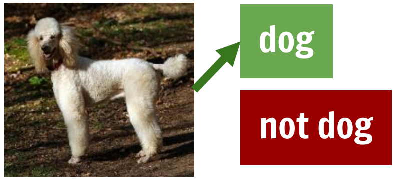
"Dribbling basketball"



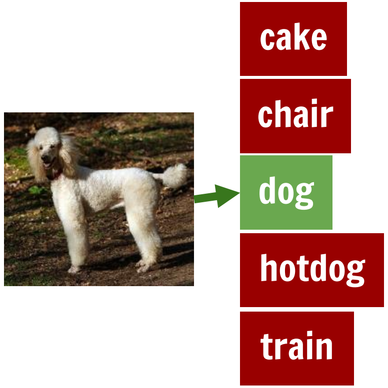
Hip hop

Classification

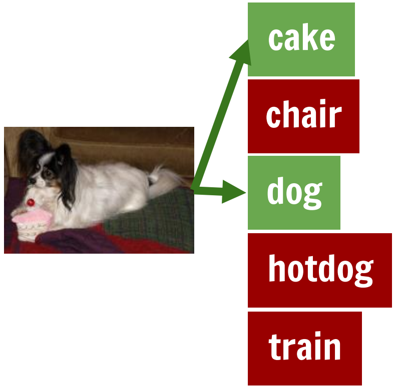
Binary classification



Multi-class classification



Multi-label classification



Regression

$f($  $) = y$ Real-valued/ continuous

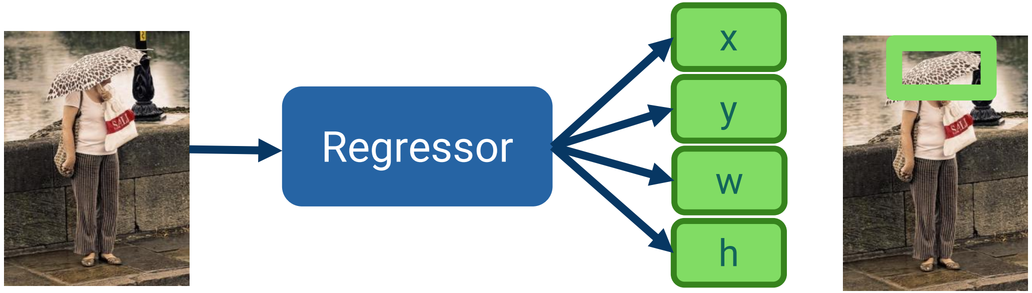
Regression

Given House Size => Predict House Price

Simple regression

Given Multiple Inputs => Predict Single Variable

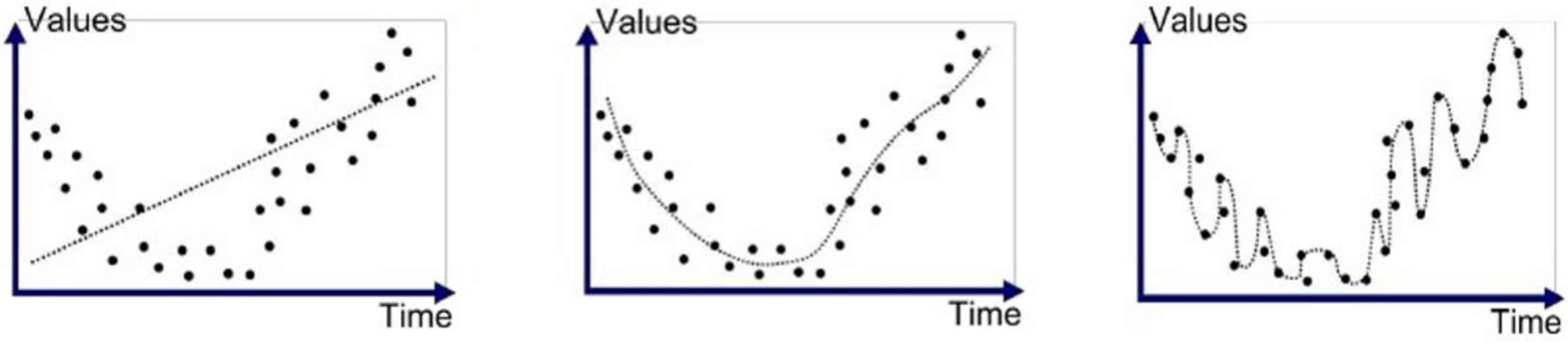
Multiple regression



Multivariate regression

Given multiple variables => Predict multiple variables

Regression



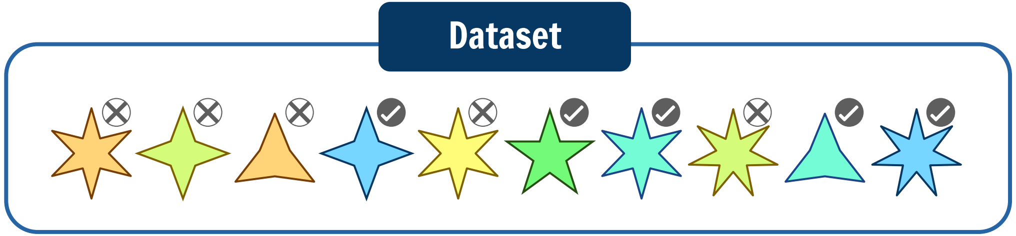
Classification $f($  $) = y$ Discrete/ categorical

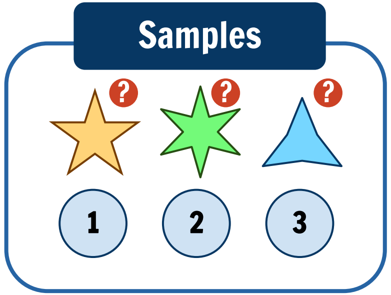
Regression $f($  $) = y$ Real-valued/ continuous

Supervised learning

The pipeline

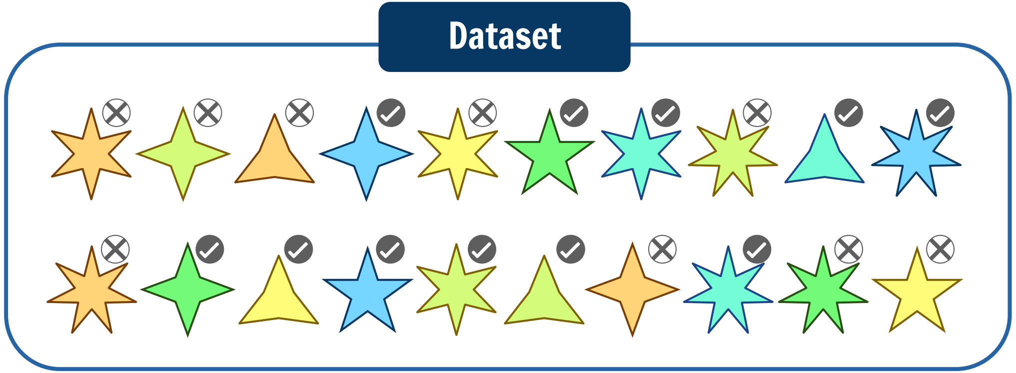
Are you a good binary classifier?

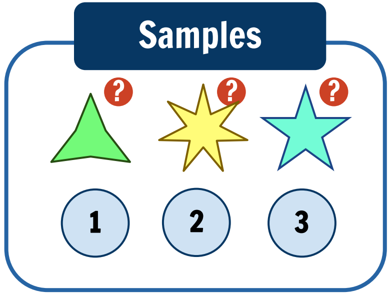




Try this!

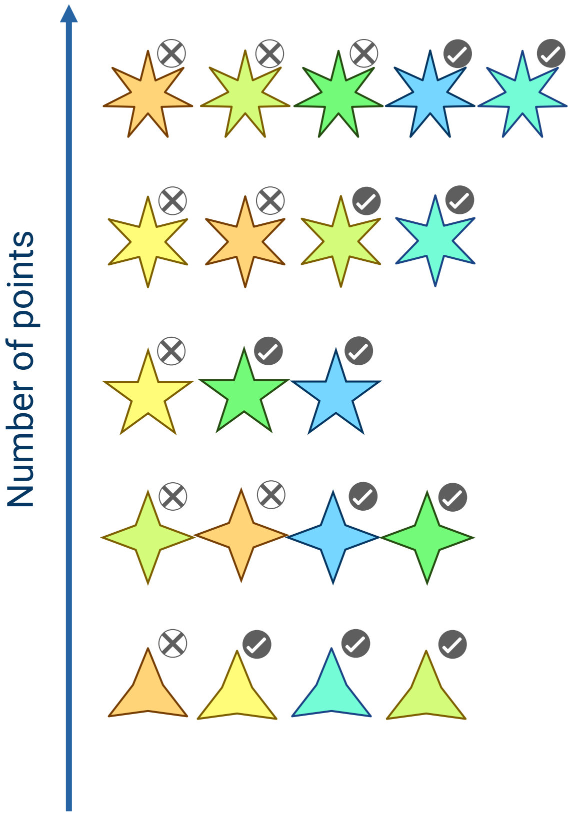
Are you a good binary classifier?

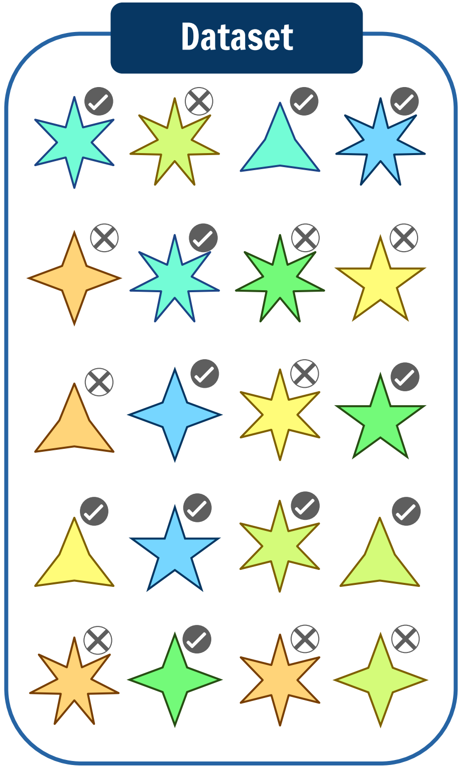


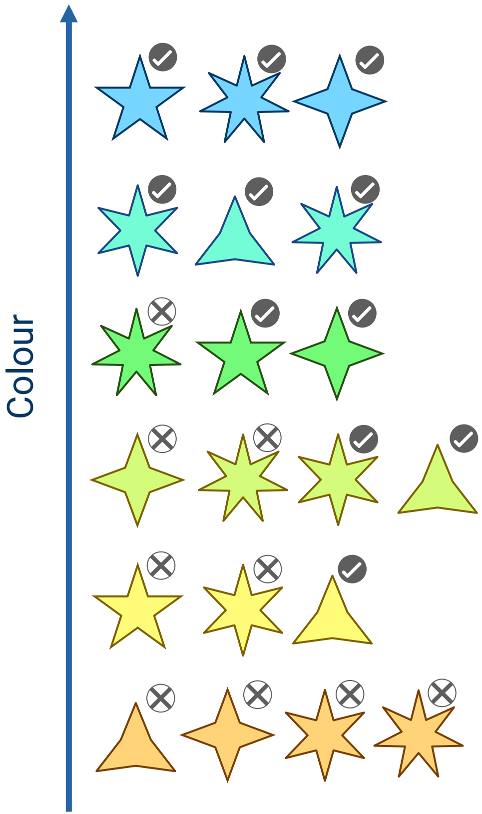


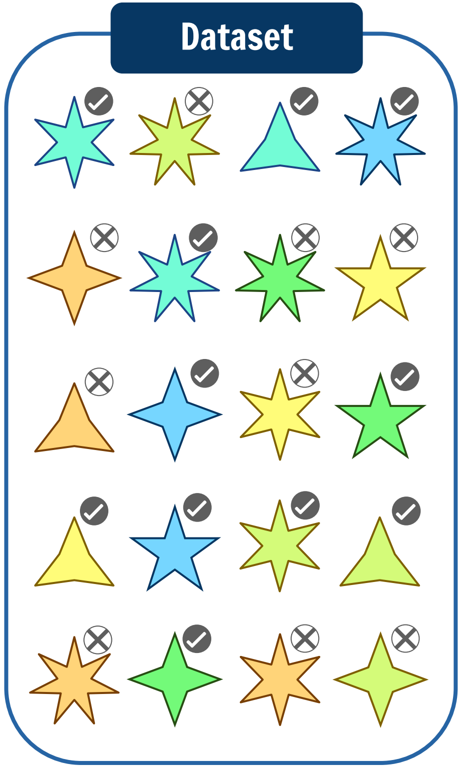
Try this!

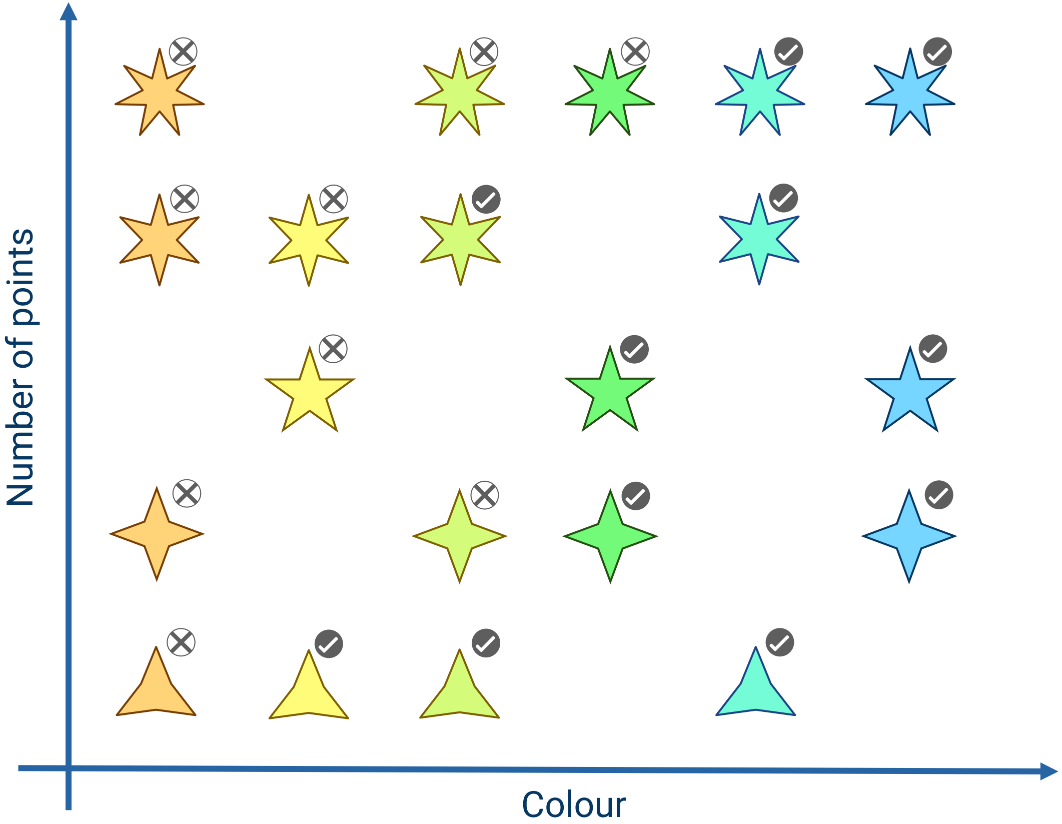
Are you still a good binary classifier?

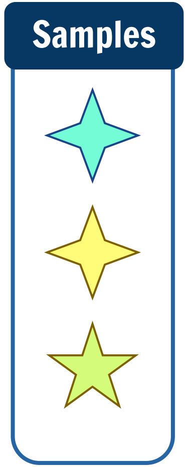

Number of points


Dataset

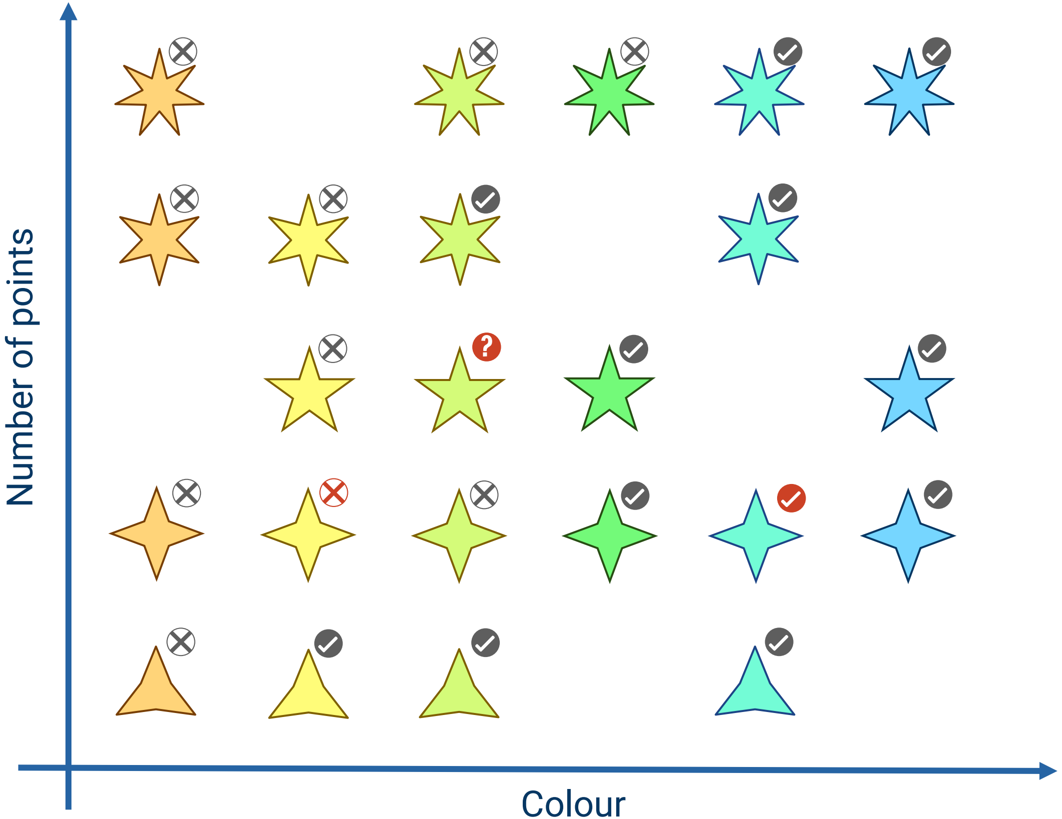

Colour


Dataset

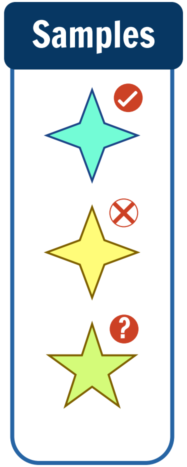




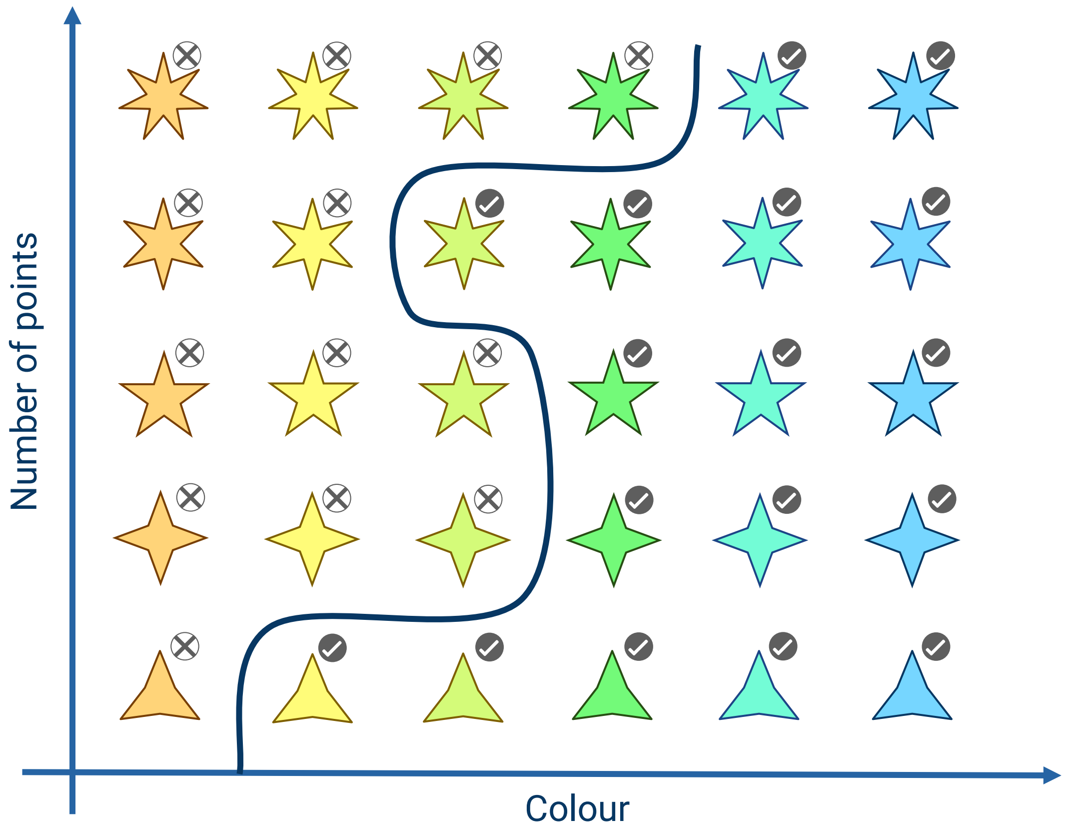
Try this!

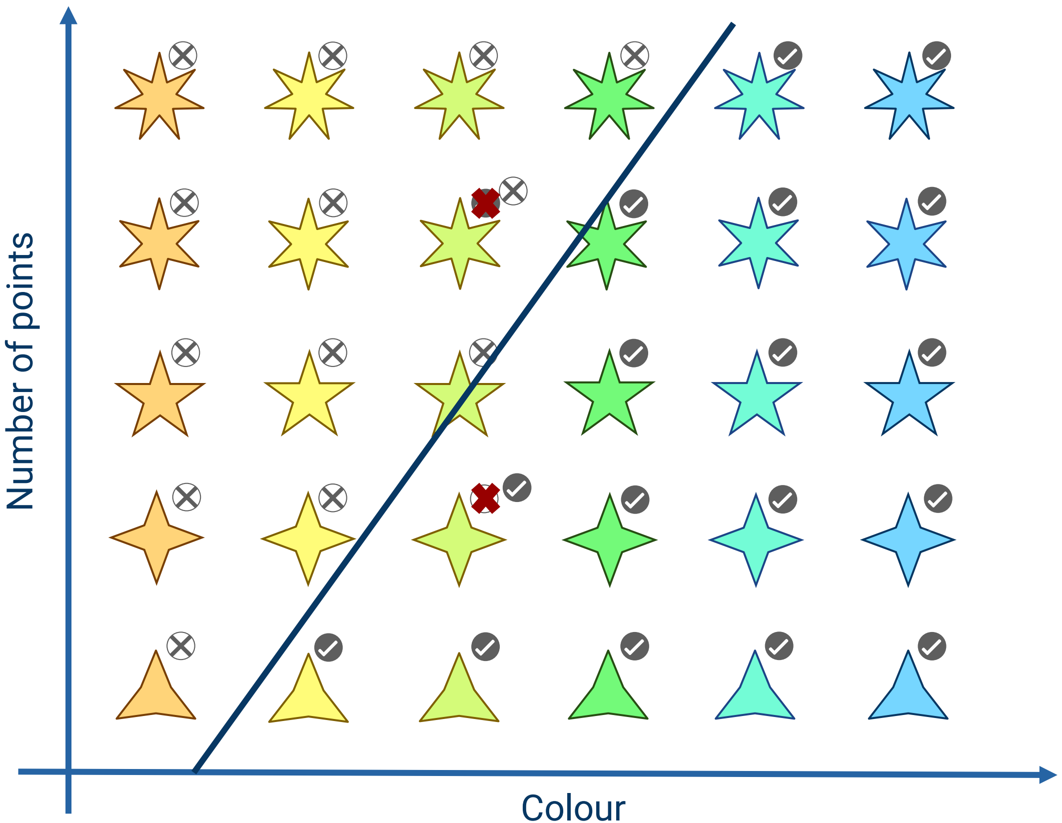


Number of points
Colour



Samples

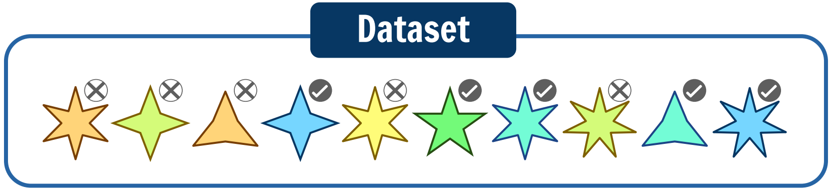

Number of points
Colour

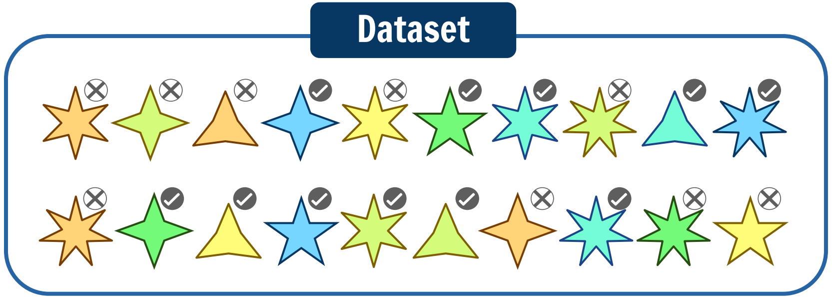


Linear classifier

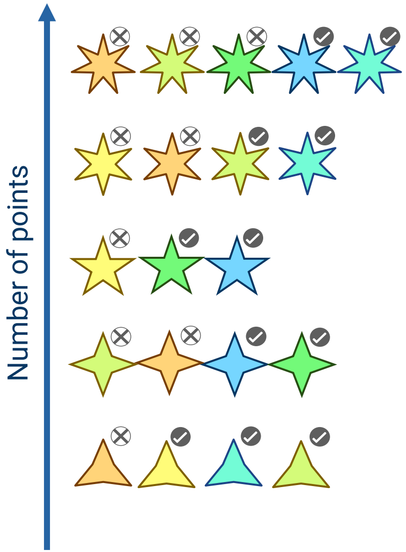
What have we learnt?

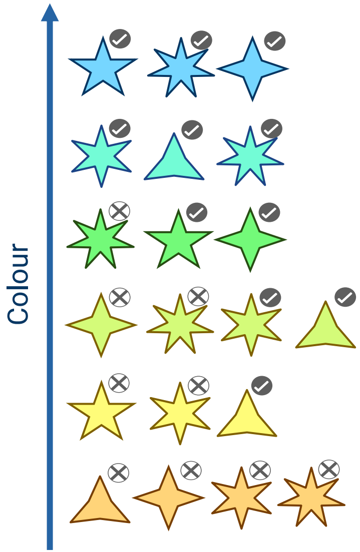
More data == more accurate predictions

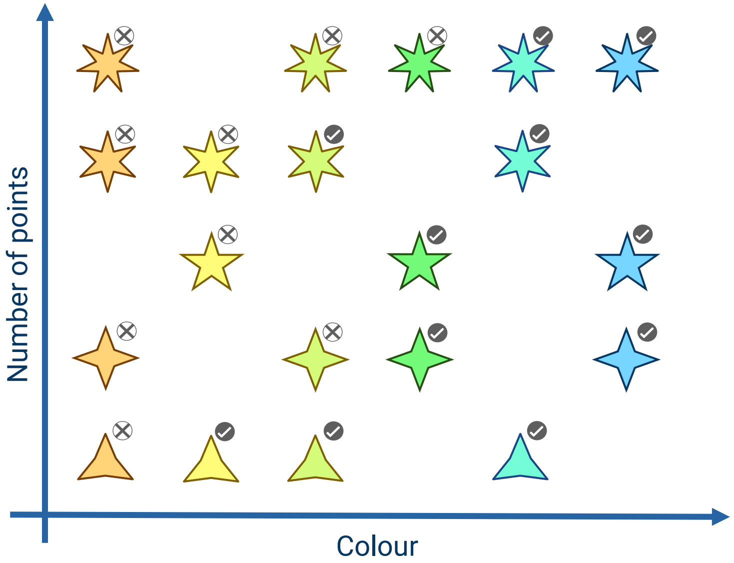




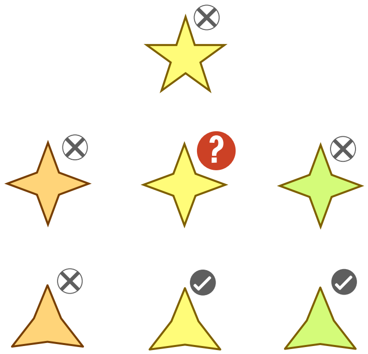
Selecting good features is crucial!



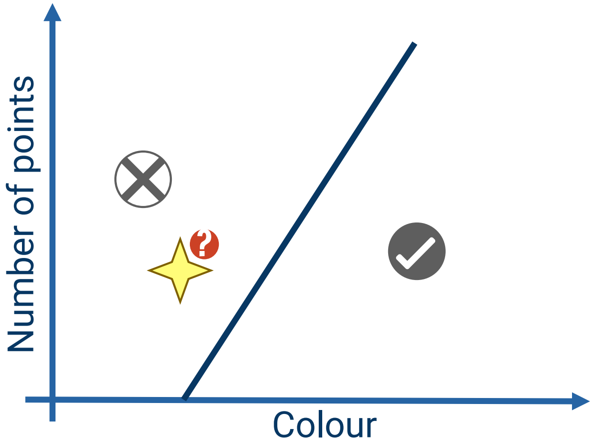




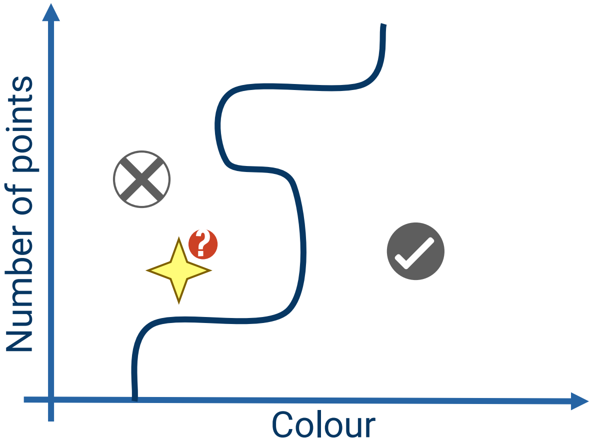
Classifiers make predictions differently



Classifiers make predictions differently



Classifiers make predictions differently

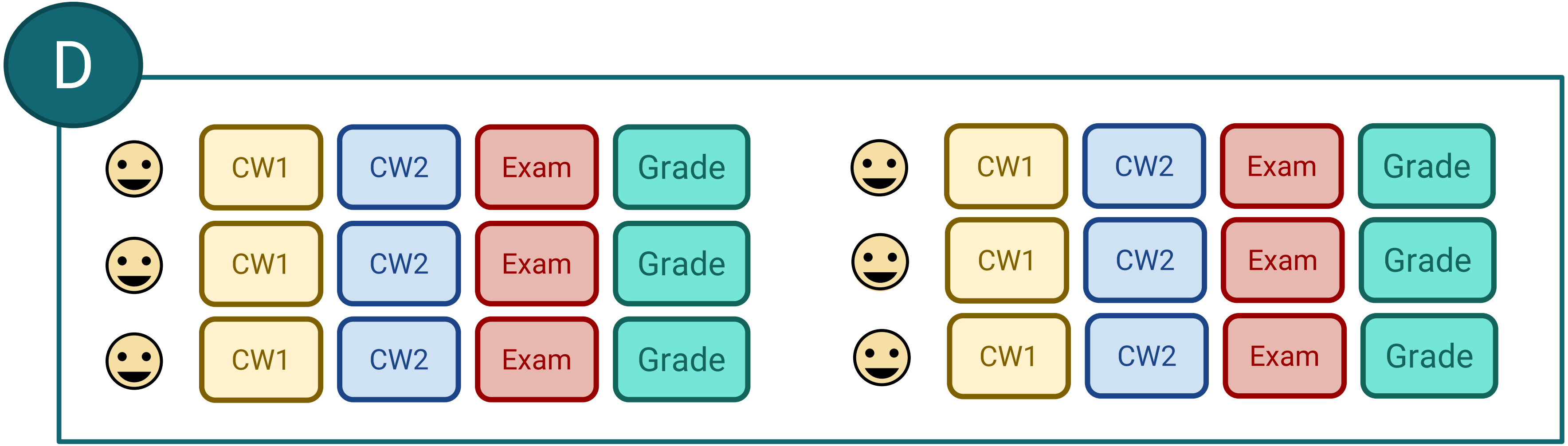


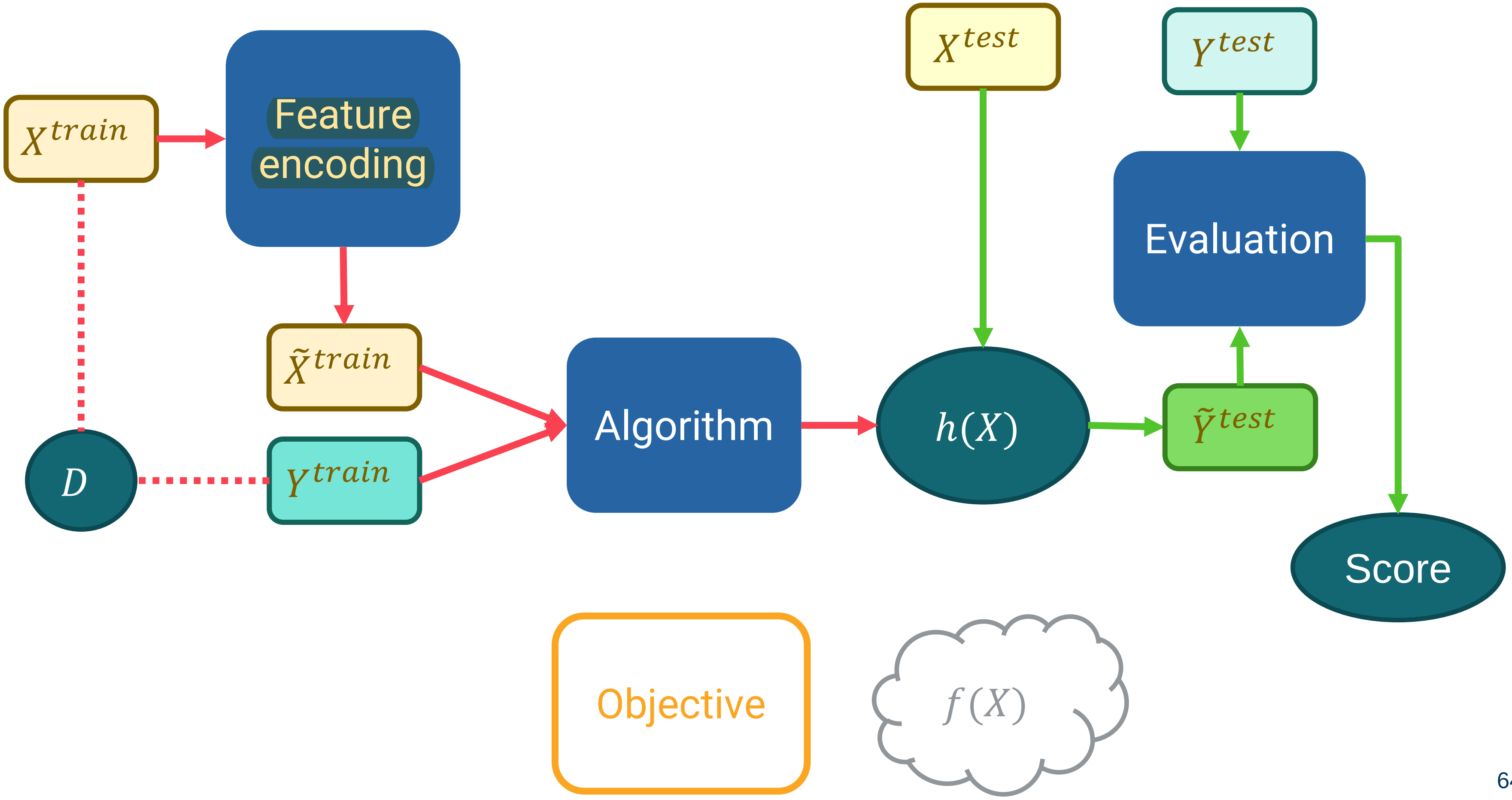
The supervised learning pipeline

$$f(\boxed{\text{CW1}}, \boxed{\text{CW2}}, \boxed{\text{Exam}}) = \boxed{\text{Module Grade}}$$

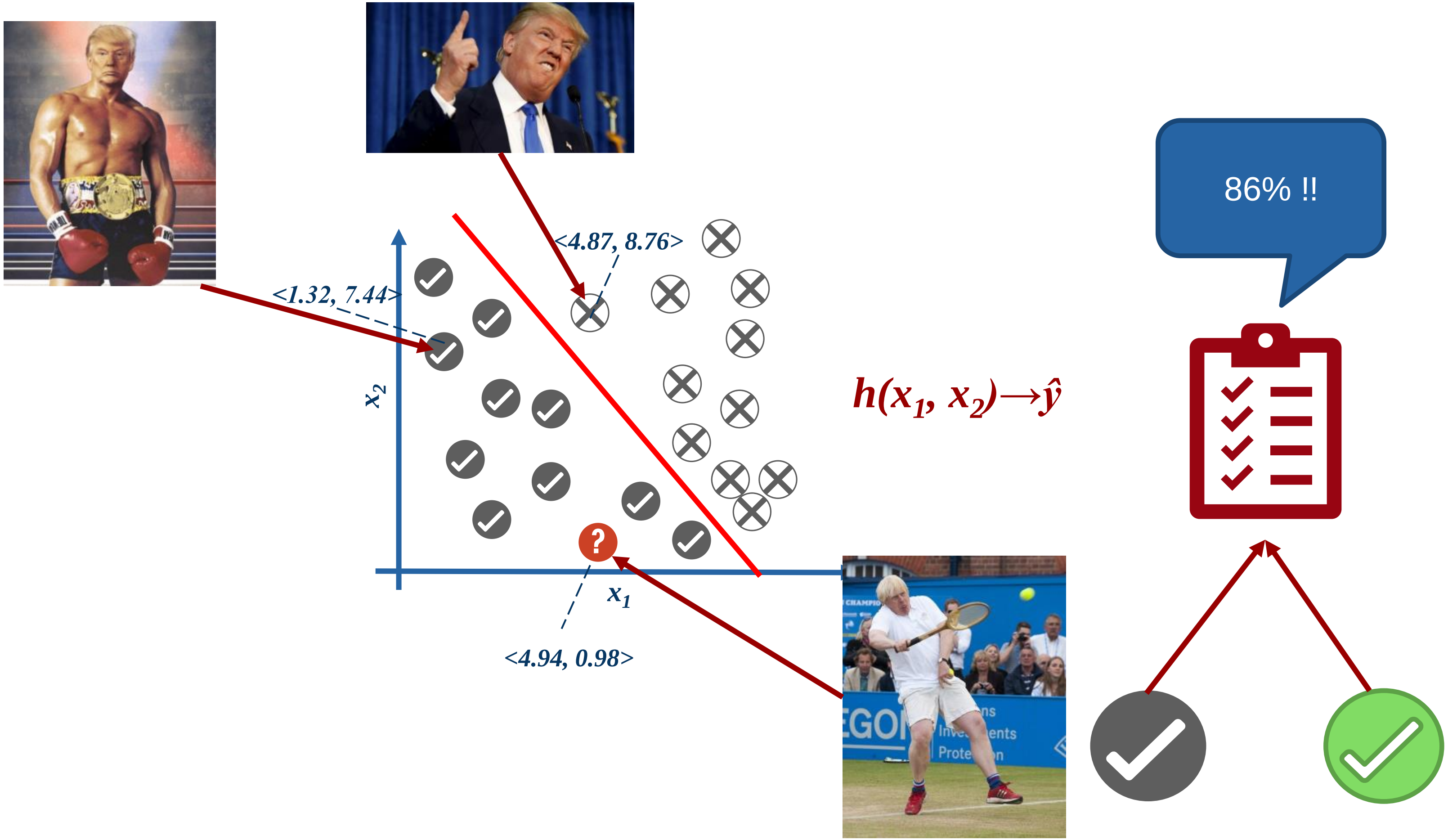
$$\approx$$

$$h(\boxed{\text{CW1}}, \boxed{\text{CW2}}, \boxed{\text{Exam}} \mid D) = \boxed{\text{Estimated Grade}}$$



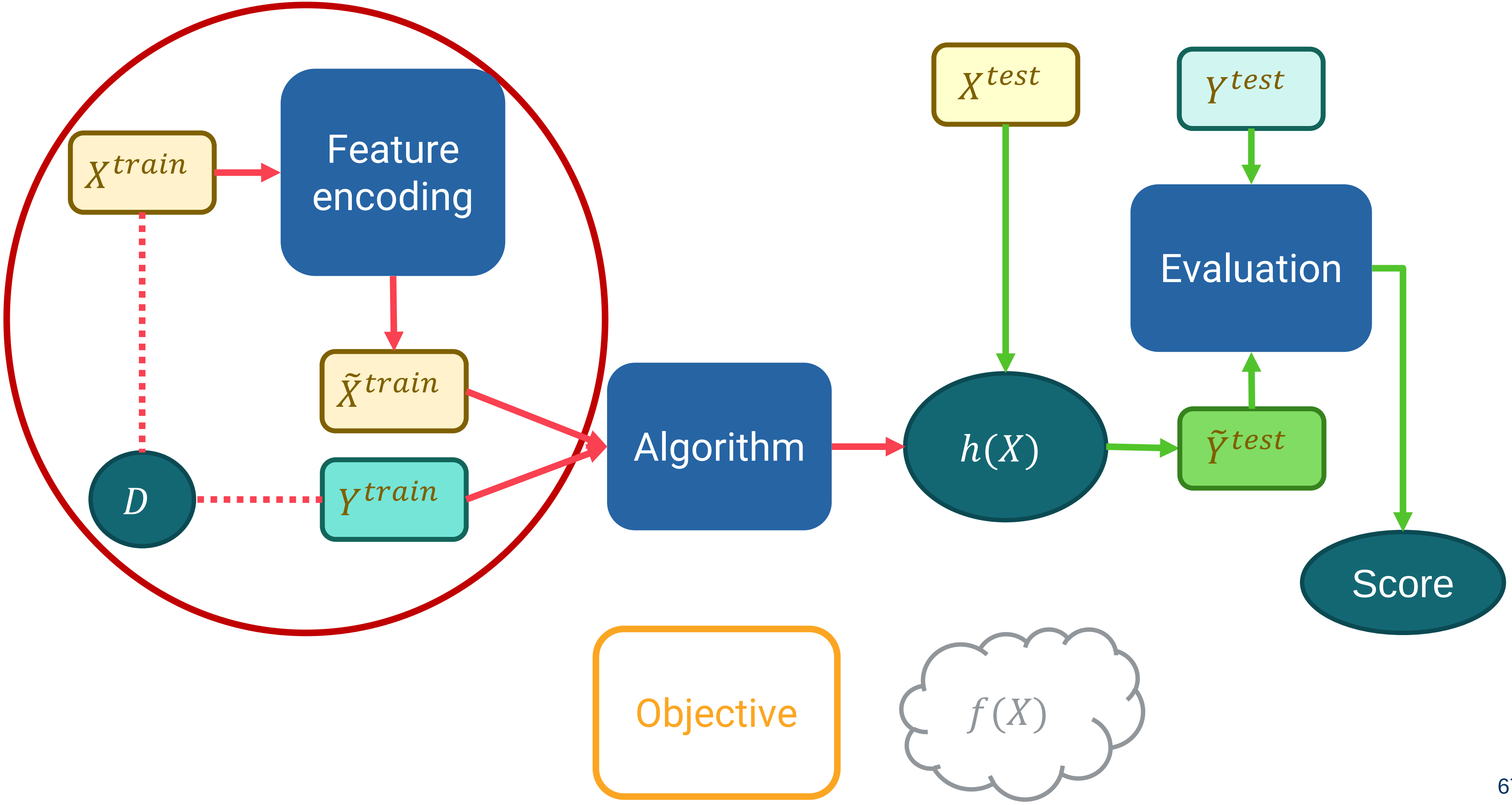


X^{train}
Feature encoding
$\tilde{X}^{train}$
D
Y^{train}
Algorithm
X^{test}
$h(X)$
Y^{test}
Evaluation
$\tilde{Y}^{test}$
Score
Objective
$f(X)$



<4.87, 8.76>
<1.32, 7.44>
x_2
x_1
<4.94, 0.98>
$h(x_1, x_2) \rightarrow \hat{y}$
86% !!

Feature encoding

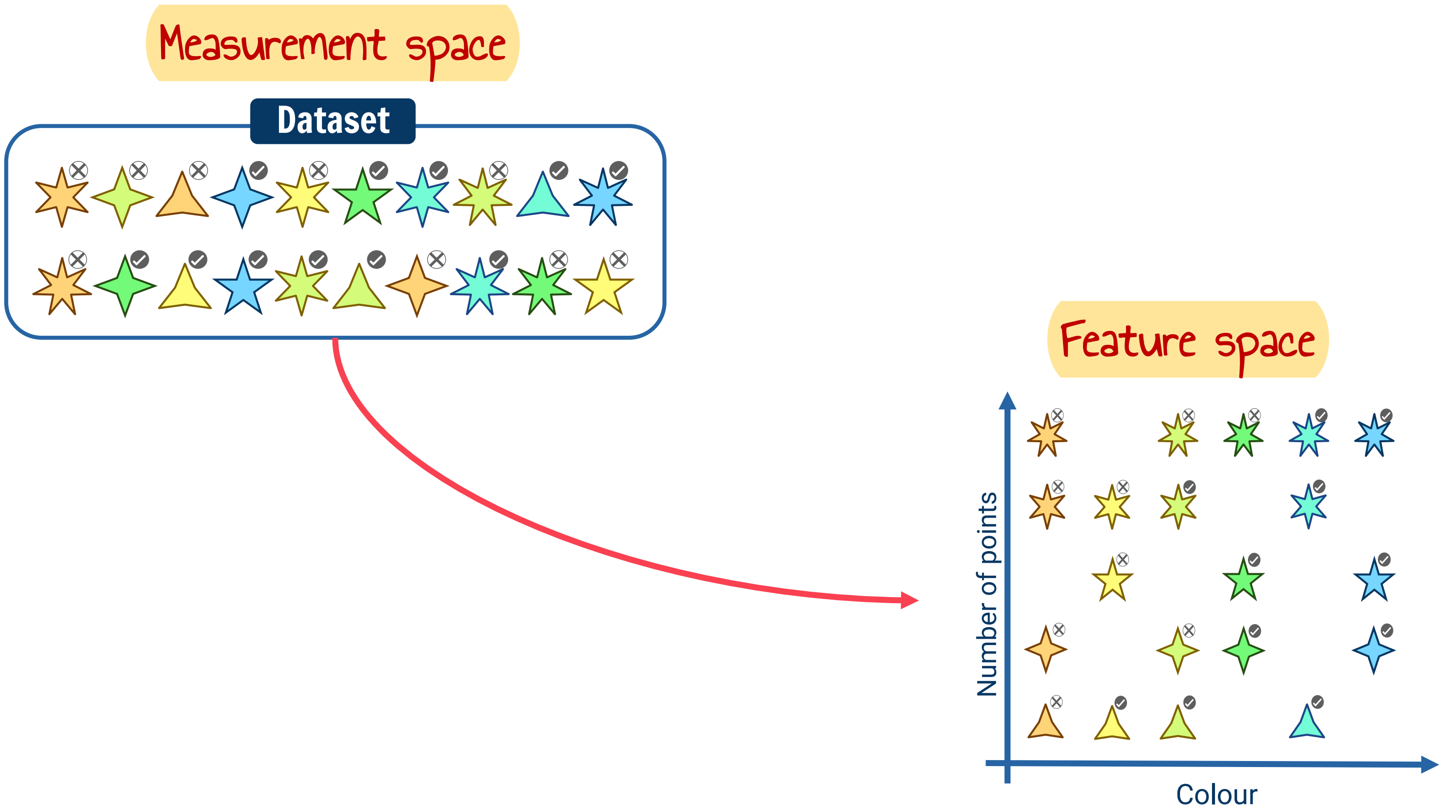

X^{train}
Feature encoding
$\tilde{X}^{train}$
D
Y^{train}
Algorithm
X^{test}
$h(X)$
Y^{test}
Evaluation
$\tilde{Y}^{test}$
Score
Objective
$f(X)$

Understanding your data

- Sometimes given as raw measurements
  - an image, a news article, a tweet, a graph, a time series, a molecular shape, etc.

Understanding your data

- Always examine your data! 
  - Clues to help you design your classifier
  - Distribution of class labels?
    - Balanced?
    - Imbalanced?



Measurement space
Dataset
Feature space
Number of points
Colour

Understanding your features

features $\tilde{X}$ attributes

$$\left\{\tilde{X}^{(i)}\right\}^{N}$$

N **instances**

$$\left\langle \tilde{x}_1{}^{(i)}, \tilde{x}_2{}^{(i)}, \tilde{x}_3{}^{(i)}, \ldots, \tilde{x}_K{}^{(i)} \right\rangle$$

The K in here means # features

***K*-dimensional features**

Categorical **Integers** **Real numbers**

Normalisation

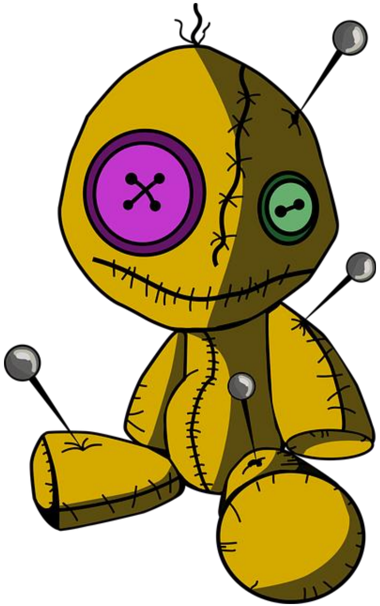
$$\tilde{x}_k{}^{(i)} = \frac{x_k{}^{(i)} - \mu_k}{\sigma_k}$$

Normalize a feature's value into a fixed range.

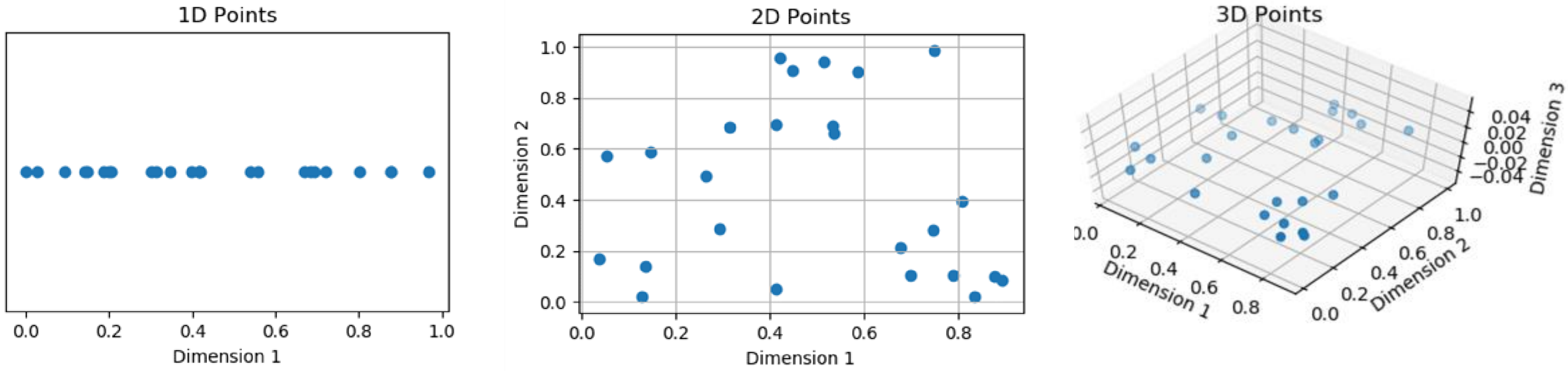
More features == better?

Only up to a certain point!

The curse of dimensionality



The curse of dimensionality



Increased computational complexity

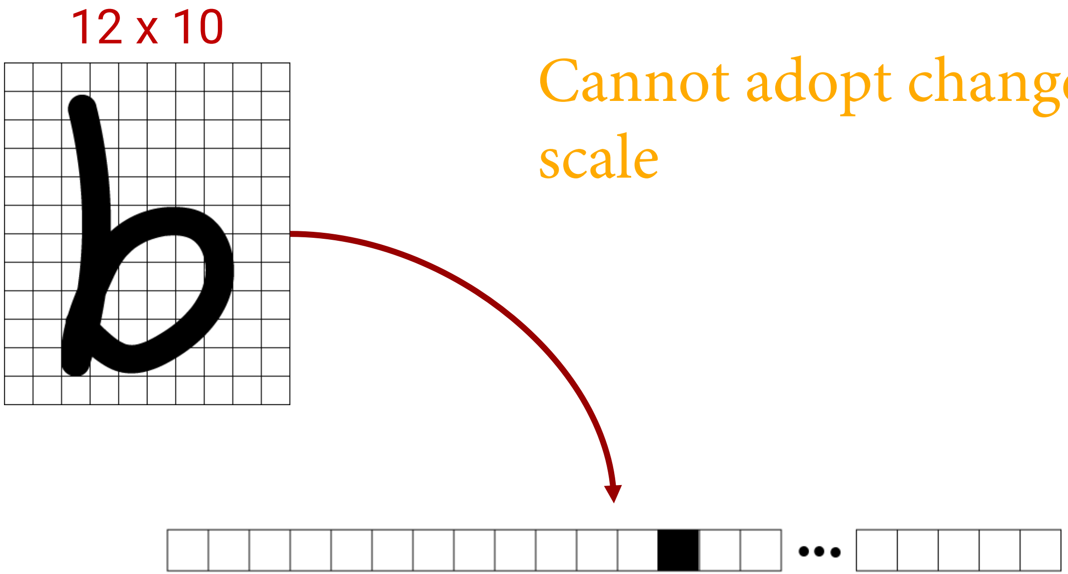
Data sparsity

Overfitting

Figure from: https://aiaspirant.com/curse-of-dimensionality/

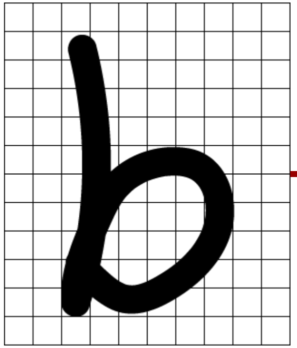
Mapping data to features

- Raw measurements: sometimes inefficient



Cannot adopt changes in position / scale

Mapping data to features



Width	6
Height	10
# “on” pixels	27
Ratio	0.6

Manual
(feature engineering)

Automatic

select central pixels only

feature selection

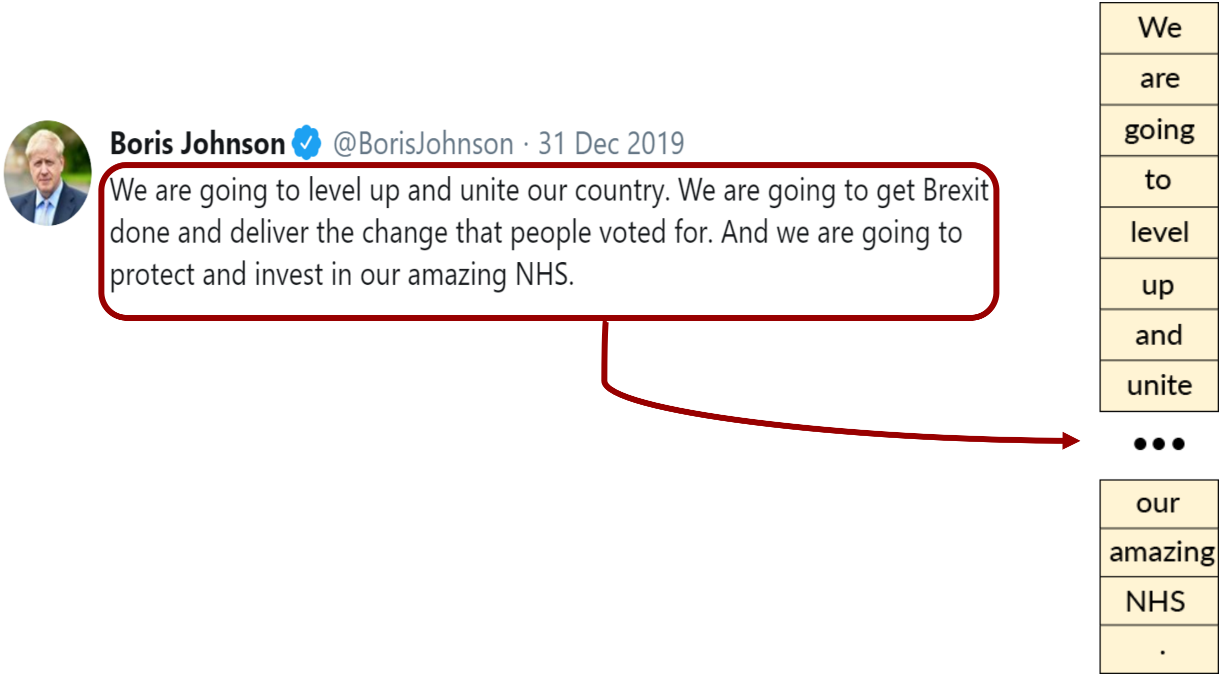
(subset)

feature extraction

(new)

Extract new features
"on" pixels

Mapping data to features



Mapping data to features

Boris Johnson @BorisJohnson · 31 Dec 2019

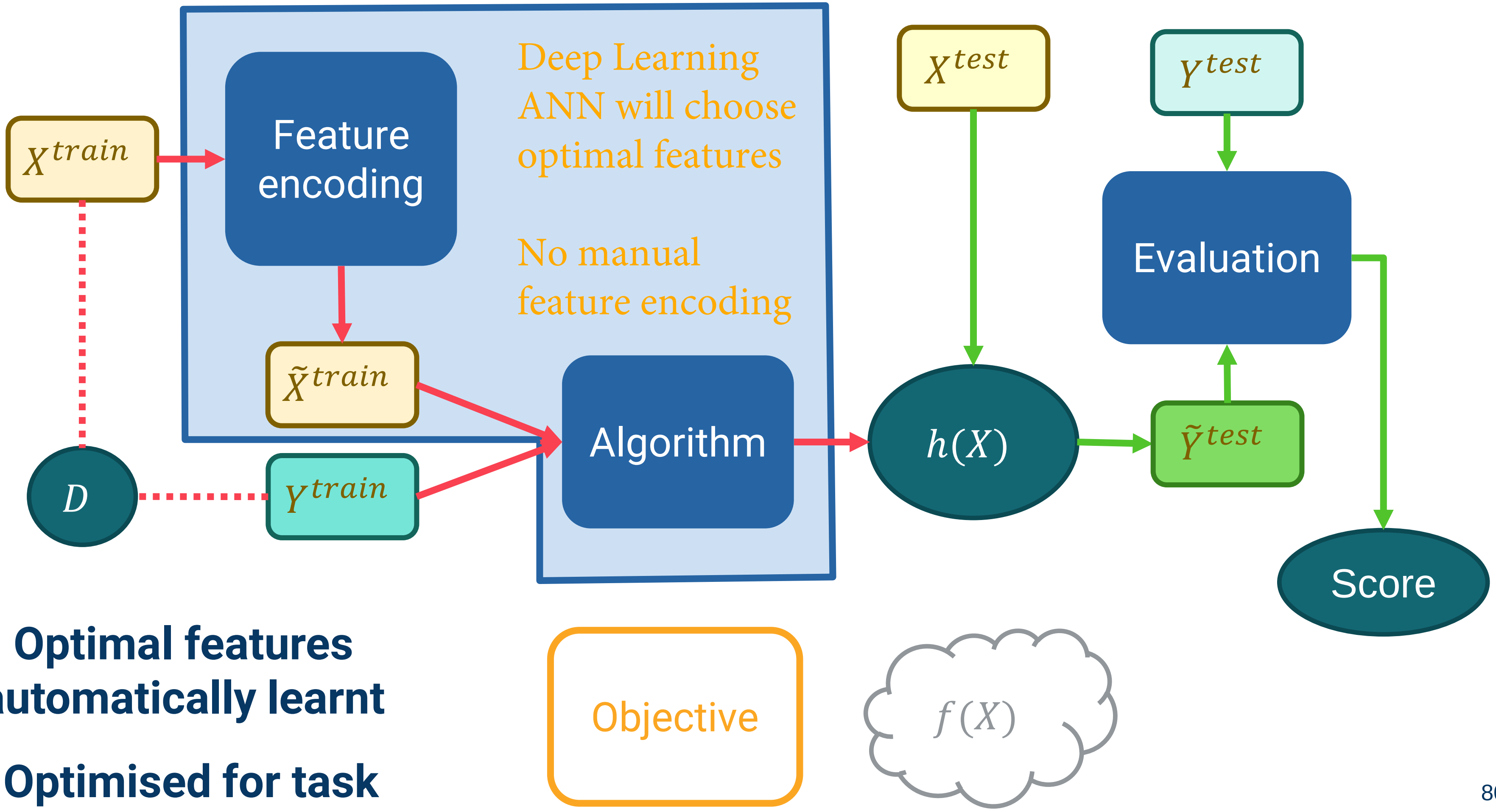
We are going to level up and unite our country. We are going to get Brexit done and deliver the change that people voted for. And we are going to protect and invest in our amazing NHS.

a	0
are	3
and	2
car	0
change	1
country	1
•••	
protect	1
queen	0
stand	0
unite	1
voted	1
zoo	0

Bag of words

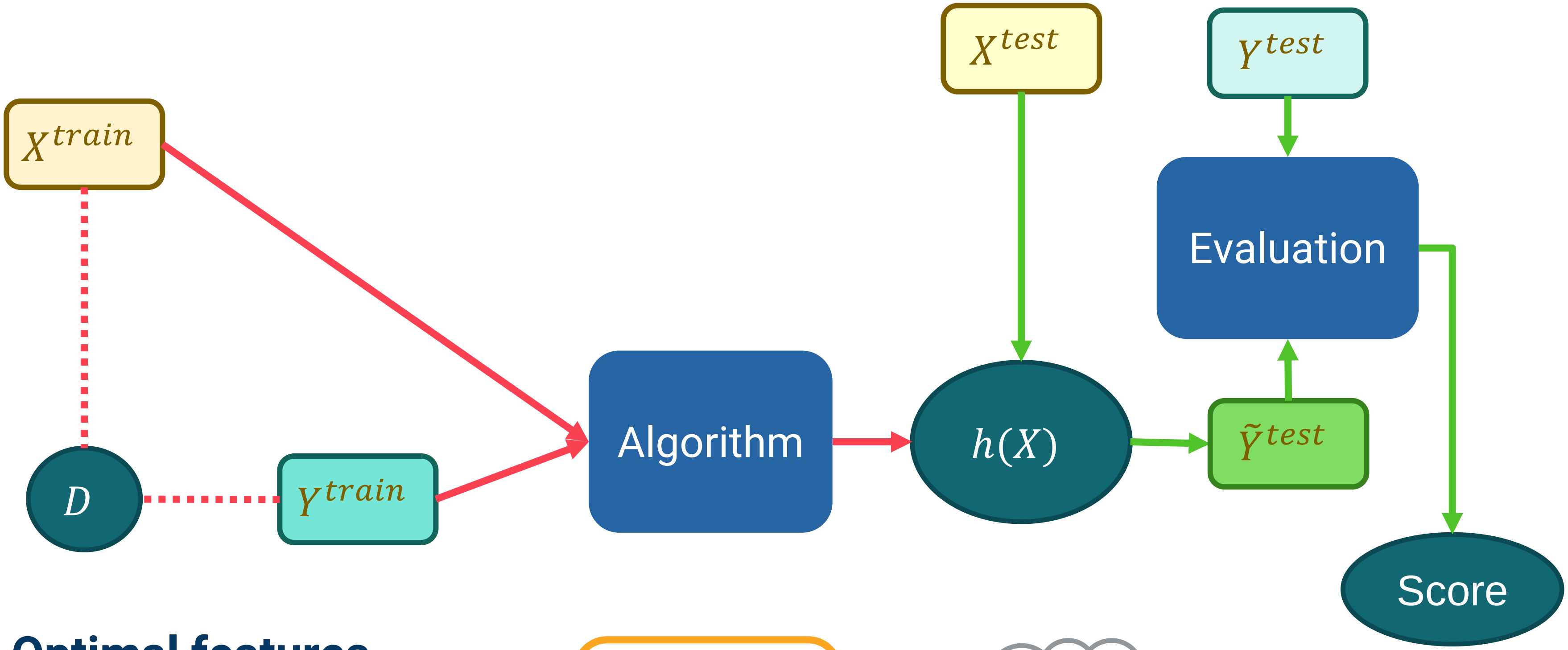
Feature spaces

The deep learning era



Optimal features automatically learnt

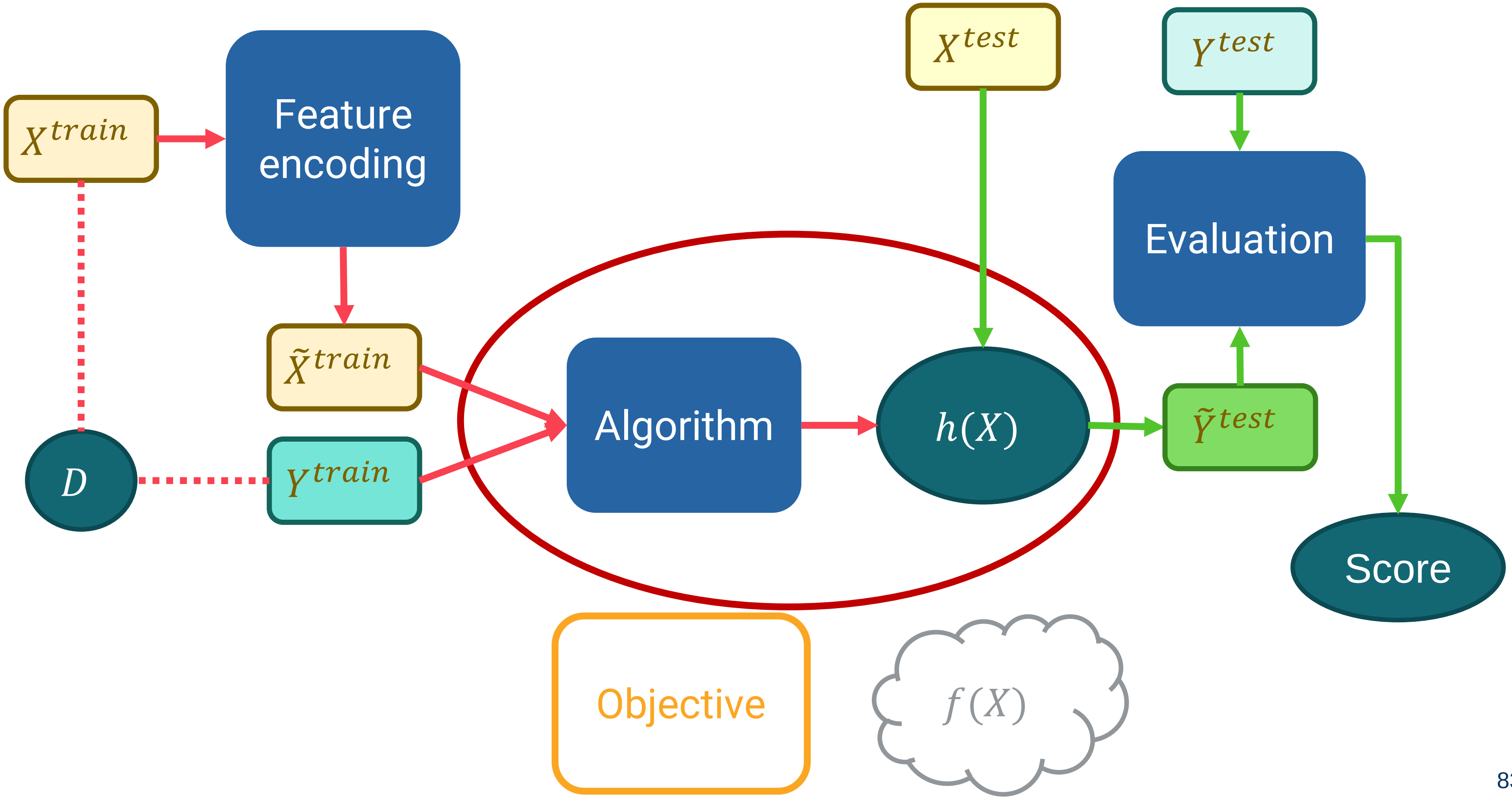
Optimised for task



Optimal features automatically learnt

Optimised for task

Machine learning algorithms

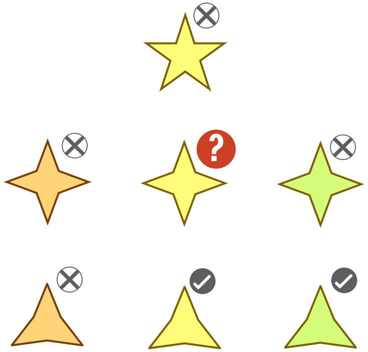

X^{train}
Feature encoding
$\tilde{X}^{train}$
D
Y^{train}
Algorithm
X^{test}
$h(X)$
Y^{test}
Evaluation
$\tilde{Y}^{test}$
Score
Objective
$f(X)$

Lazy vs. Eager Learning

Lazy Learner

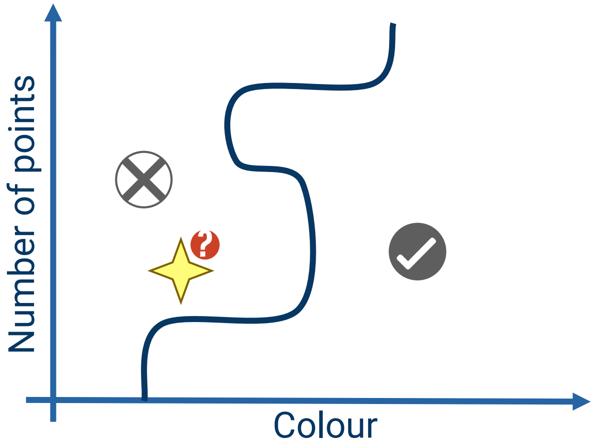
Stores the training examples and postpones generalising beyond these data until an explicit request is made at test time.

Learn/predict at run-time



Eager Learner

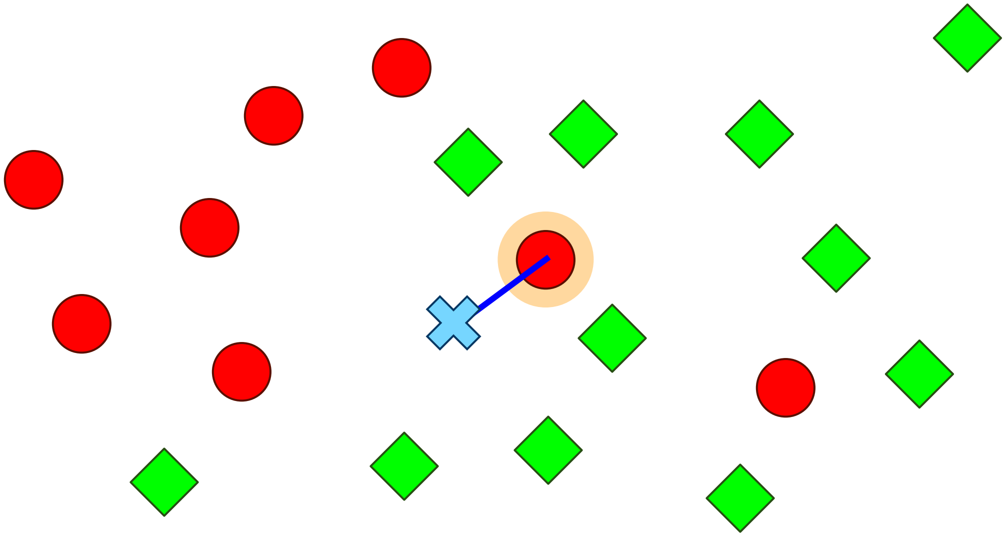
Constructs a general, explicit description of the target function based on the provided training examples.



Non-parametric model

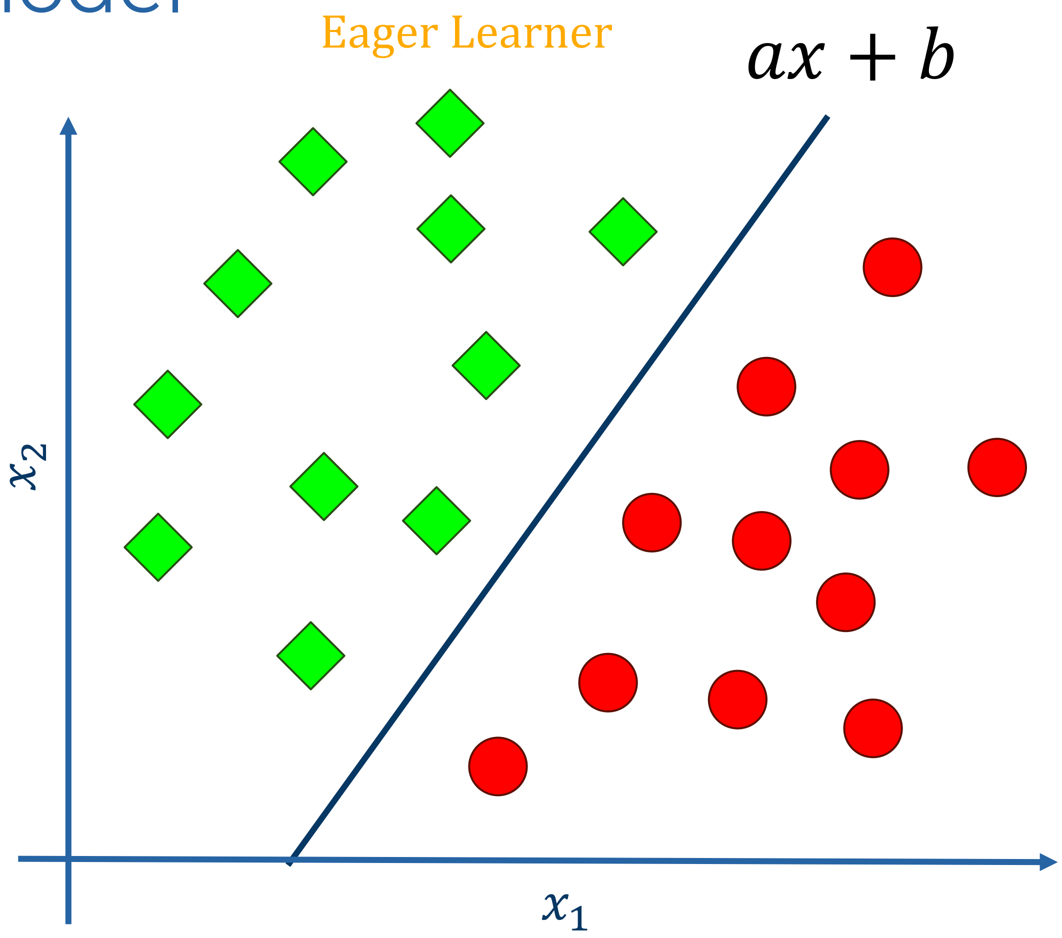
Nearest neighbour

Lazy Learner (when test data came, find the nearest available component



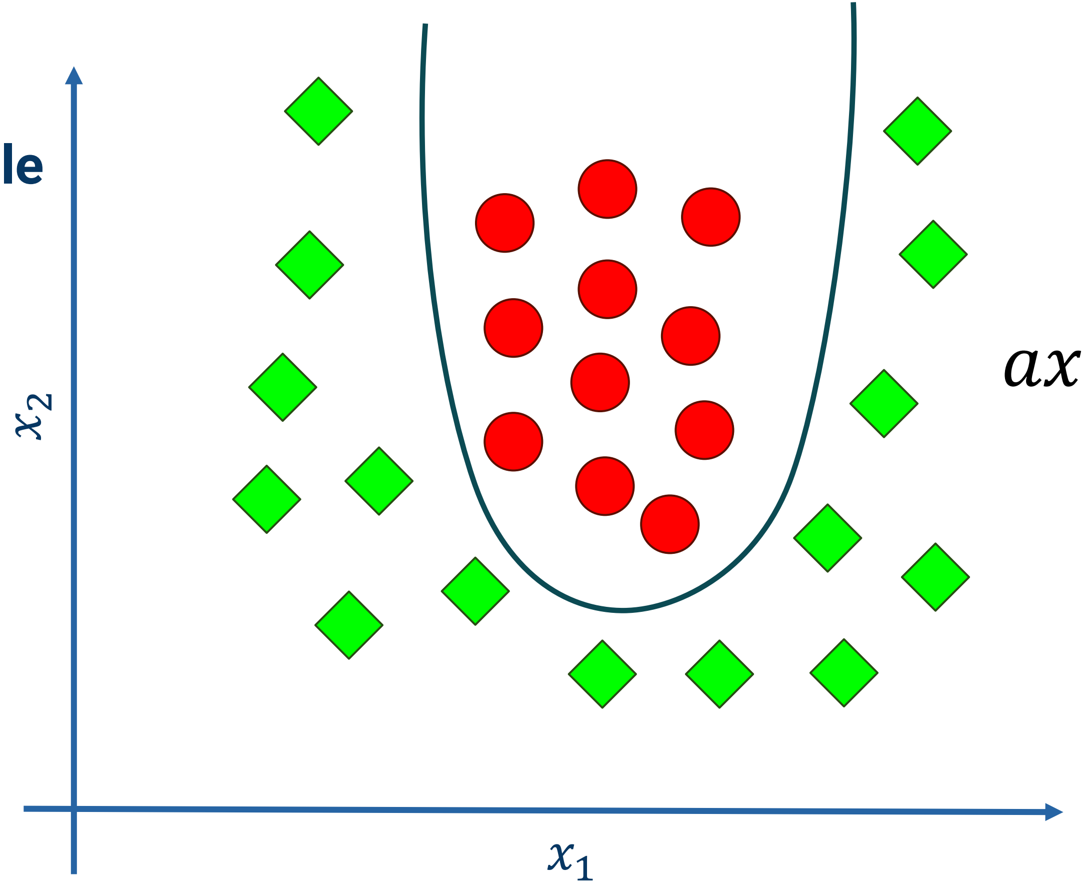
Linear model

Linearly separable



Nonlinear model

Linearly non-separable

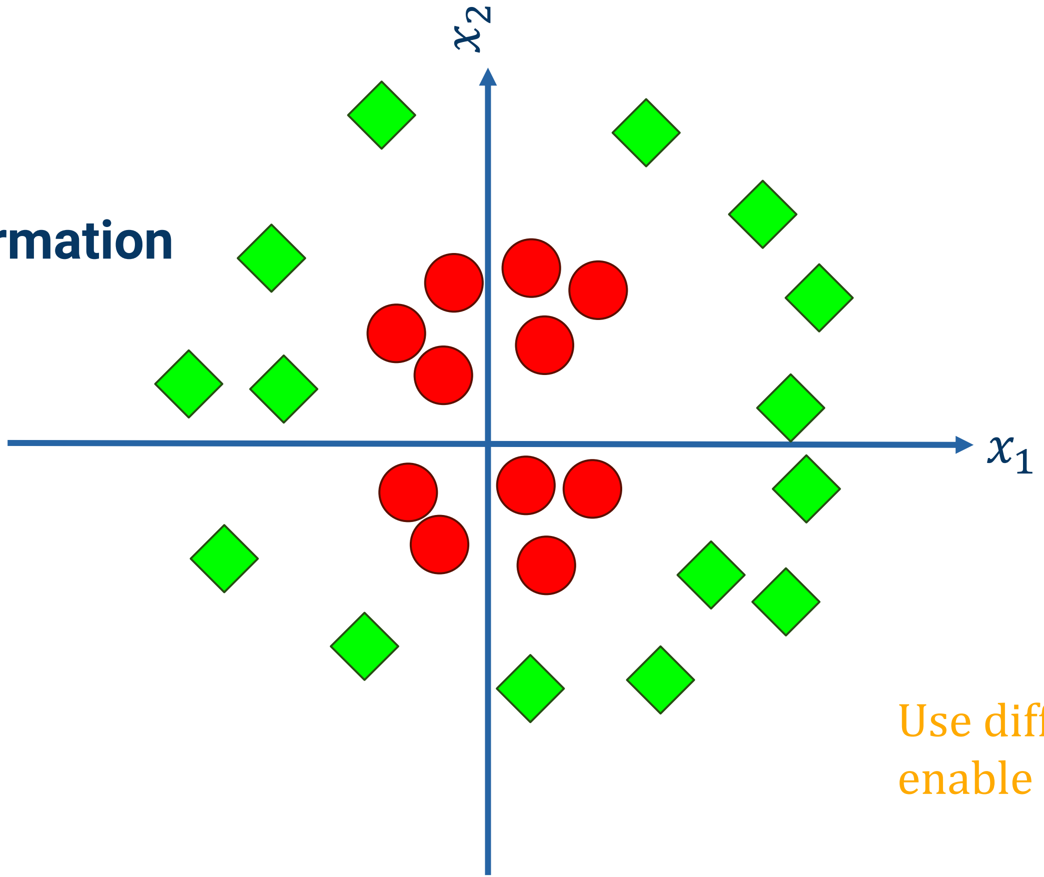


$\cancel{a \quad b}$

$ax^2 + bx + c$

Nonlinear model

Feature space transformation



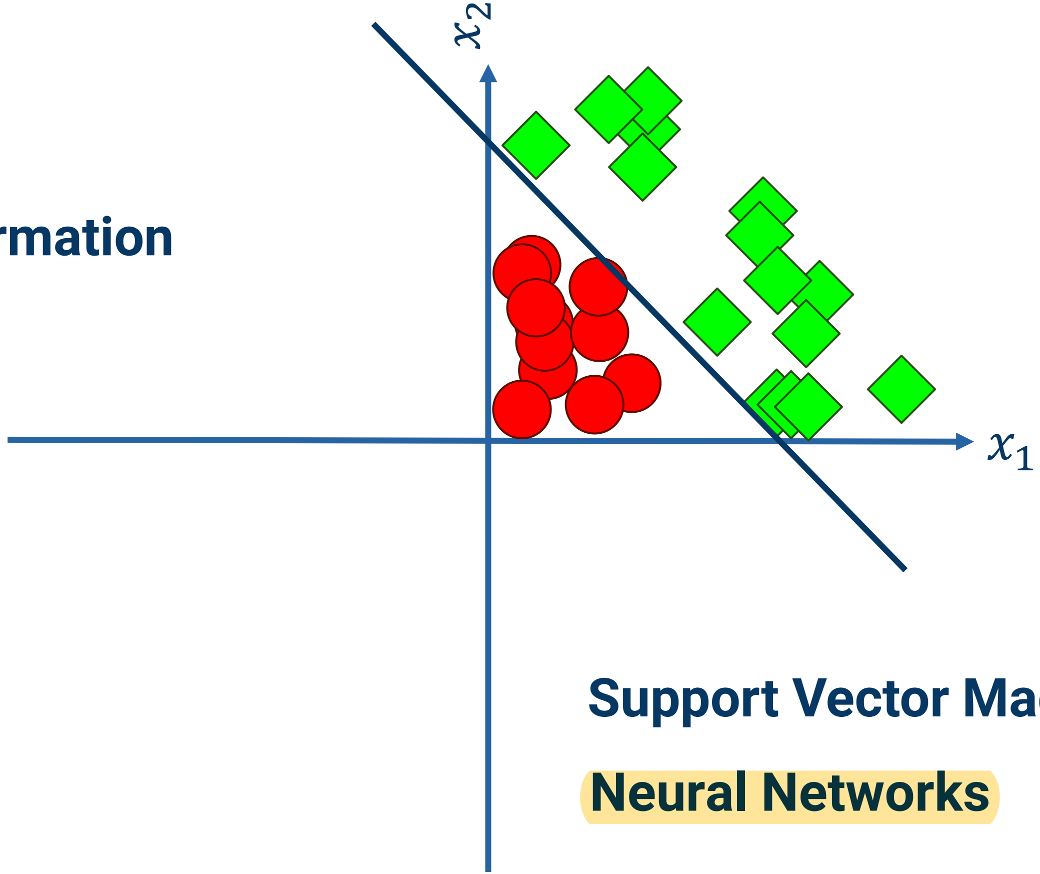
$$(x_1, x_2)$$

$$({x_1}^2, {x_2}^2)$$

Use different representation to enable linear separable.

Nonlinear model

Feature space transformation



$$(x_1, x_2)$$

$$\downarrow$$

$$({x_1}^2, {x_2}^2)$$

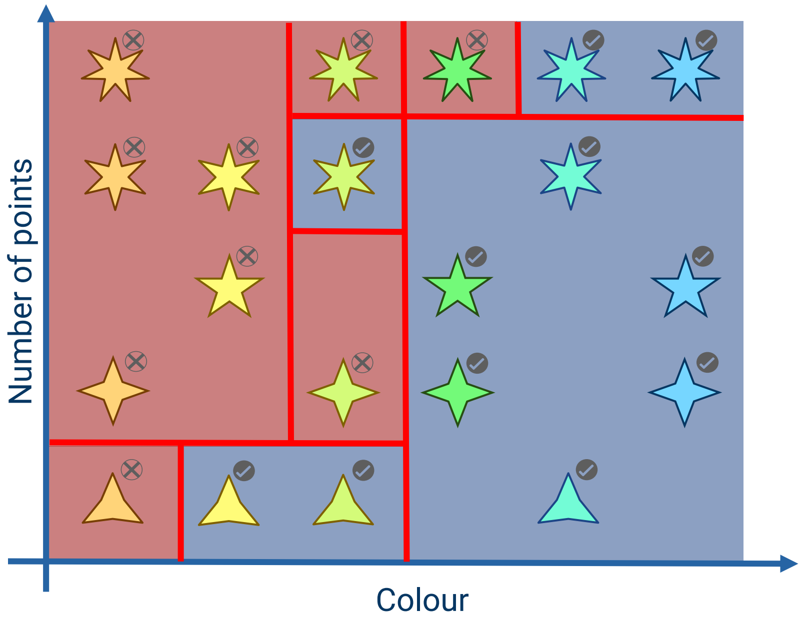
Support Vector Machines (SVM) Kernel

Neural Networks

Nonlinear model

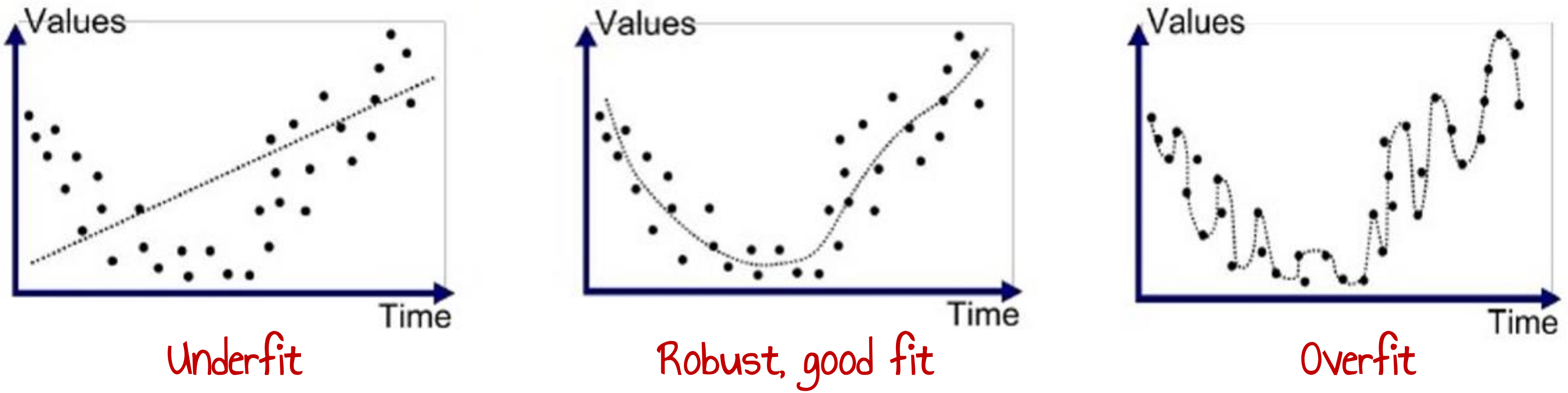
Combine multiple simple classifiers

Decision Trees



Bias-variance trade-off

One of the most important ML concepts!



Low variance

High bias

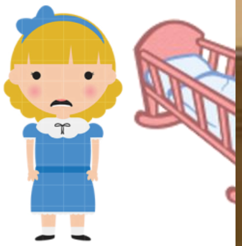
Bias-variance trade-off

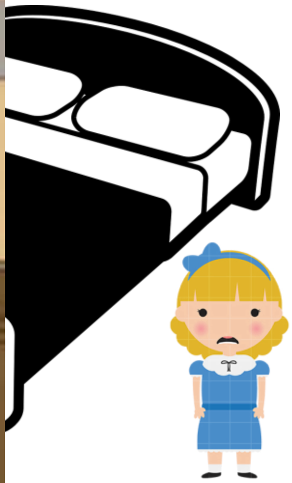
High variance

Low bias

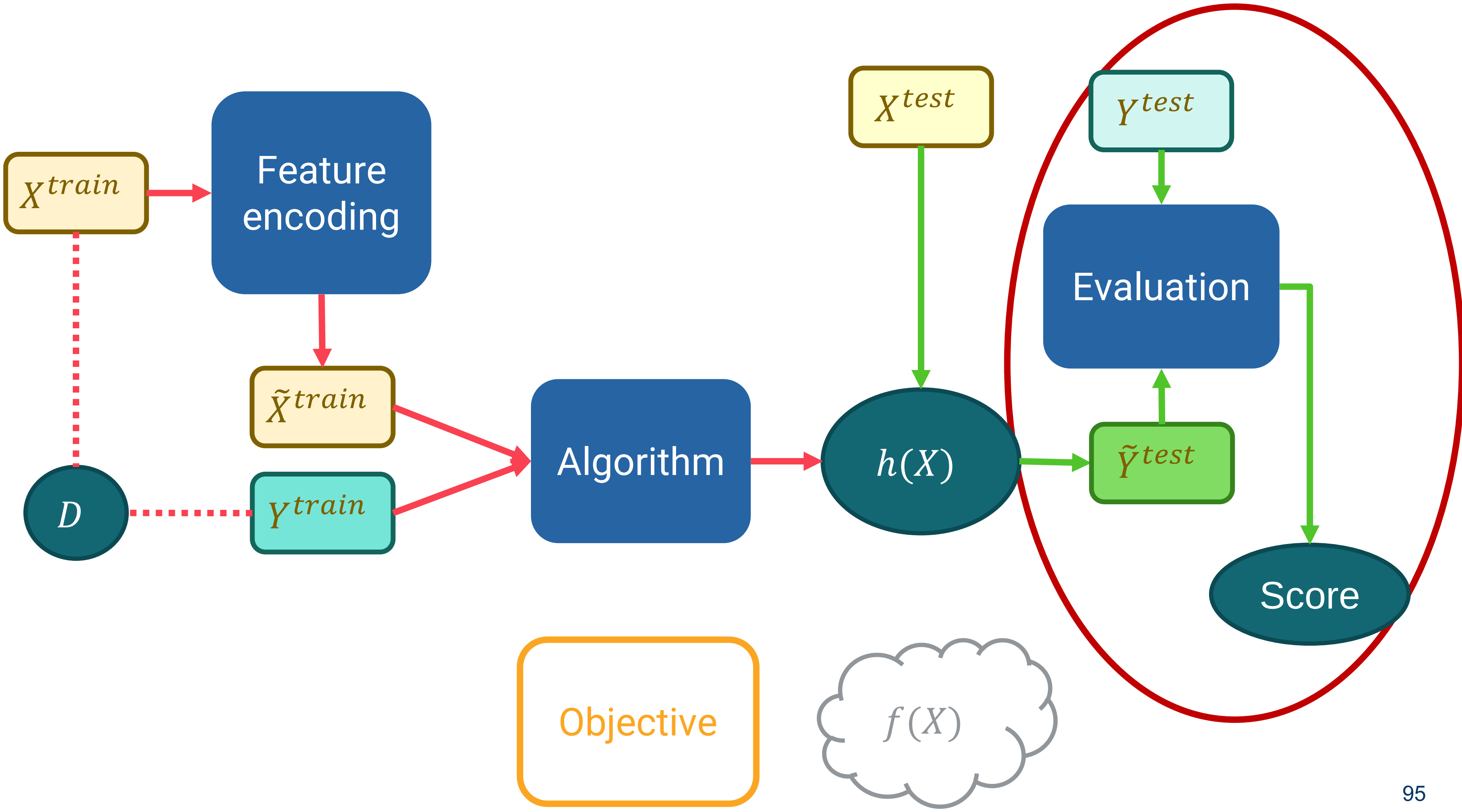
***Occam's razor**: More things should not be used than are necessary.*

Bias => Over-simplifed Assumption leads incorrect result





Evaluation



X^{train}
Feature encoding
$\tilde{X}^{train}$
D
Y^{train}
Algorithm
X^{test}
$h(X)$
Y^{test}
Evaluation
$\tilde{Y}^{test}$
Score
Objective
$f(X)$

How well does the model perform?

Evaluation metric/measure

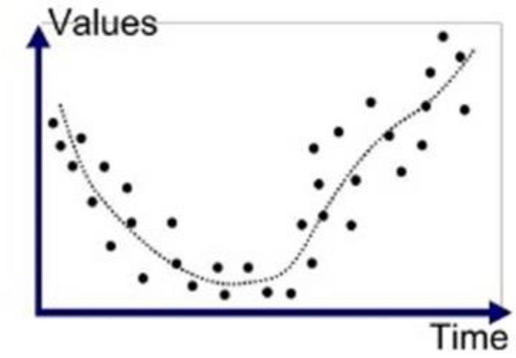
$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ test\ instances}$$

How well does the model perform?

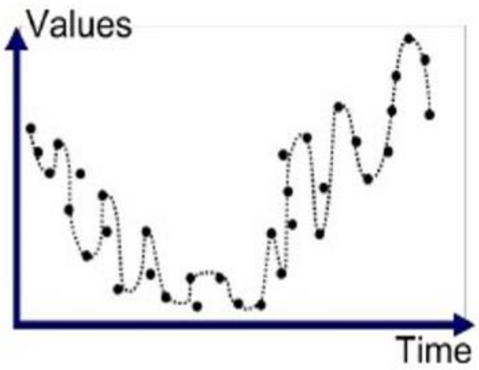
Evaluation metric/measure

Regressor

0.4682



$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$



Is 85% accuracy any good?

"It is all relative..."

Baseline

Chance/random performance [lower bound]

Is there a stronger baseline?

The base performance before any improvements

Is 85% accuracy any good?

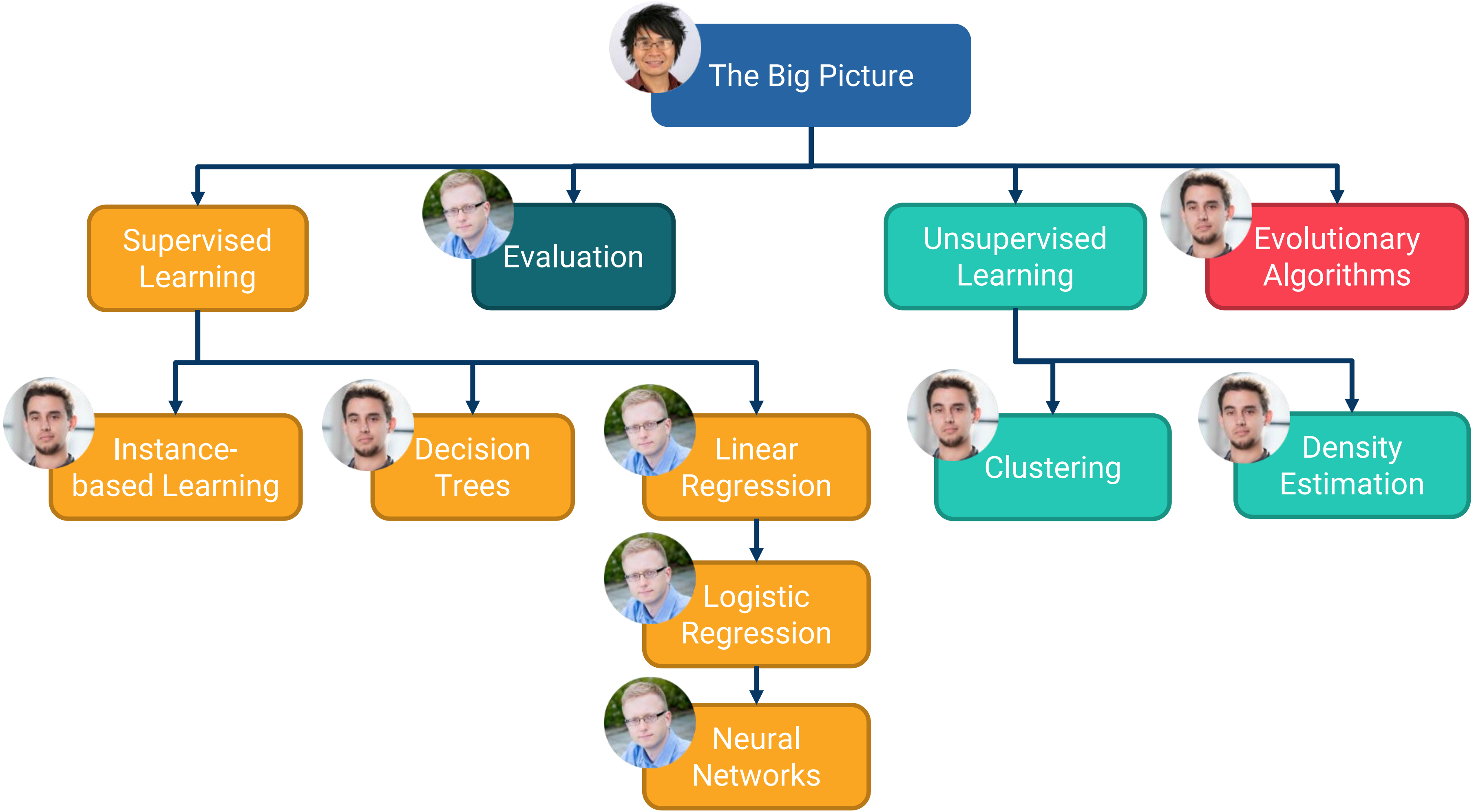
"It is all relative..."

Upper bound

The best case (usually comparison to human)

"Superhuman performance" ?

Course Roadmap


The Big Picture
Supervised Learning
Evaluation
Unsupervised Learning
Evolutionary Algorithms
Instance-based Learning
Decision Trees
Linear Regression
Logistic Regression
Neural Networks
Clustering
Density Estimation