

# Introduction to Machine Learning

## Module 1

**What is Artificial Intelligence?**

**How does it relate to Machine Learning?**



# 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

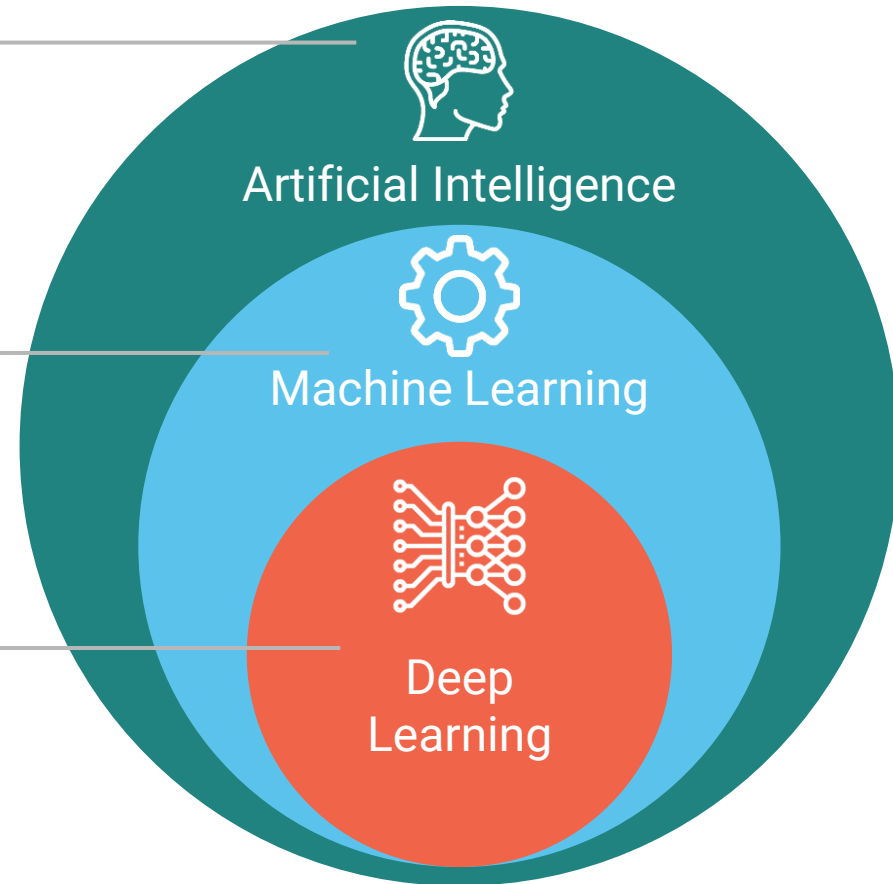
Techniques that enable computers to mimic human behaviour and intelligence. Could be using logic, if-then rules, machine learning, etc.

## Machine Learning

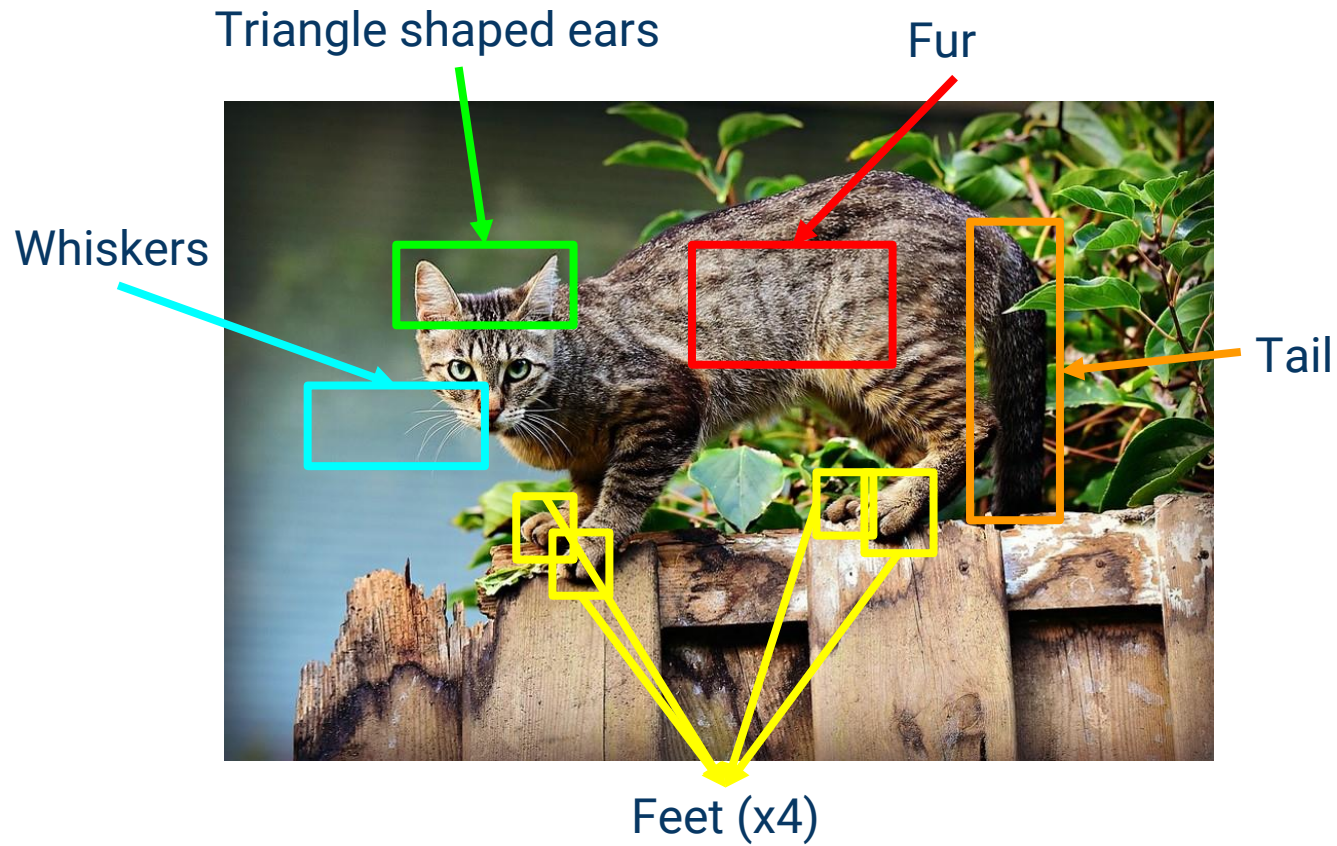
Subset of AI techniques using statistical methods that enable the systems to learn and improve with experience.

## Deep Learning

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

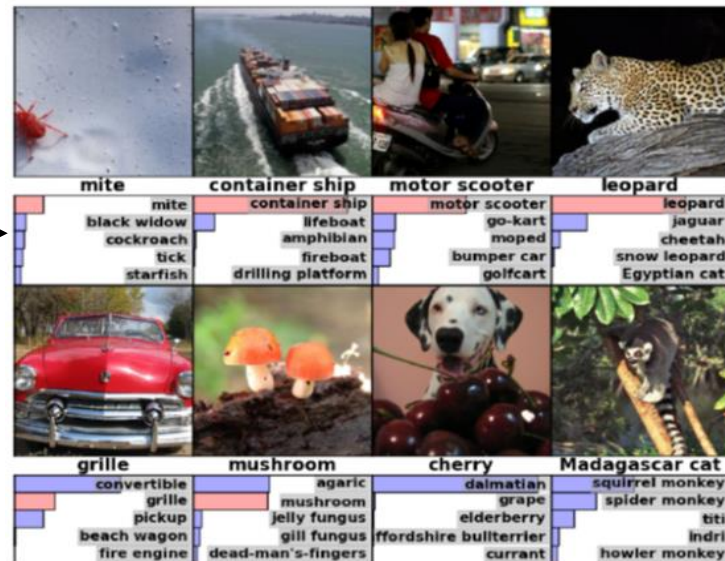






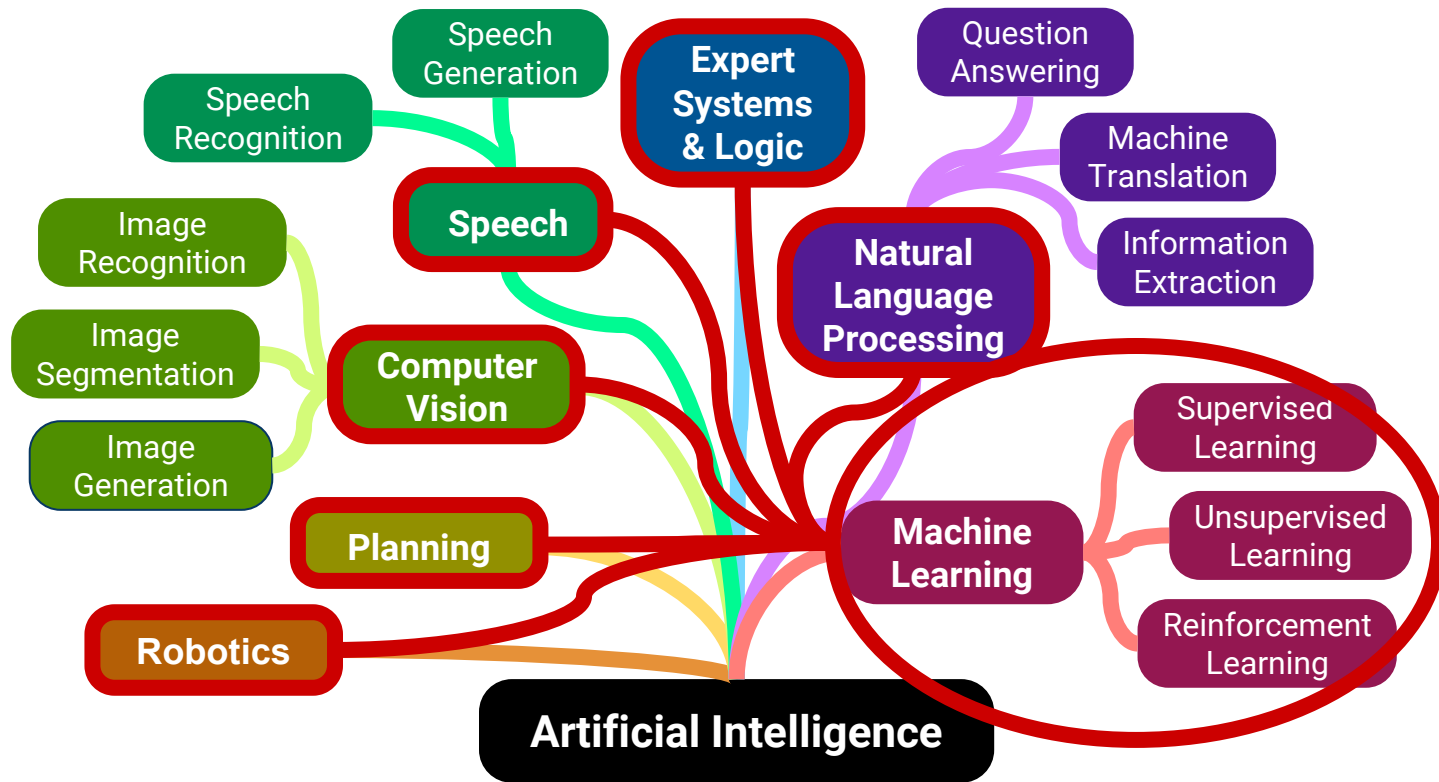


Monarch butterfly



Generalisation









# 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.

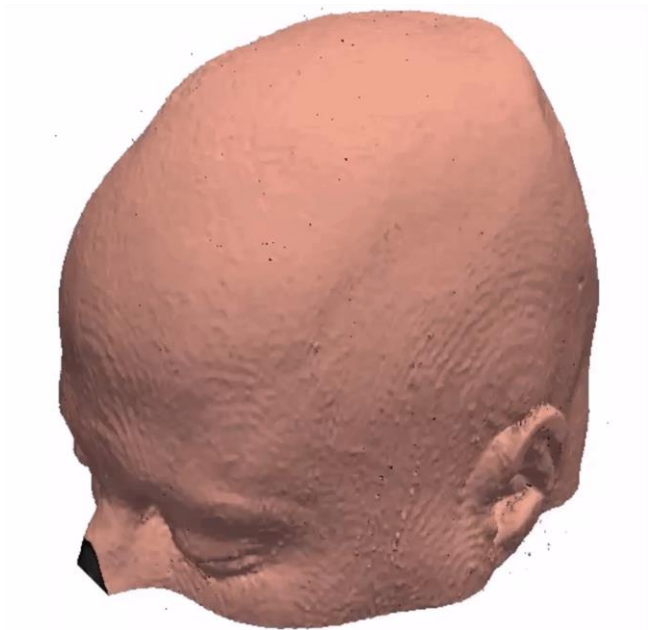
– **Dr Aldo Faisal**  
Study author

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.

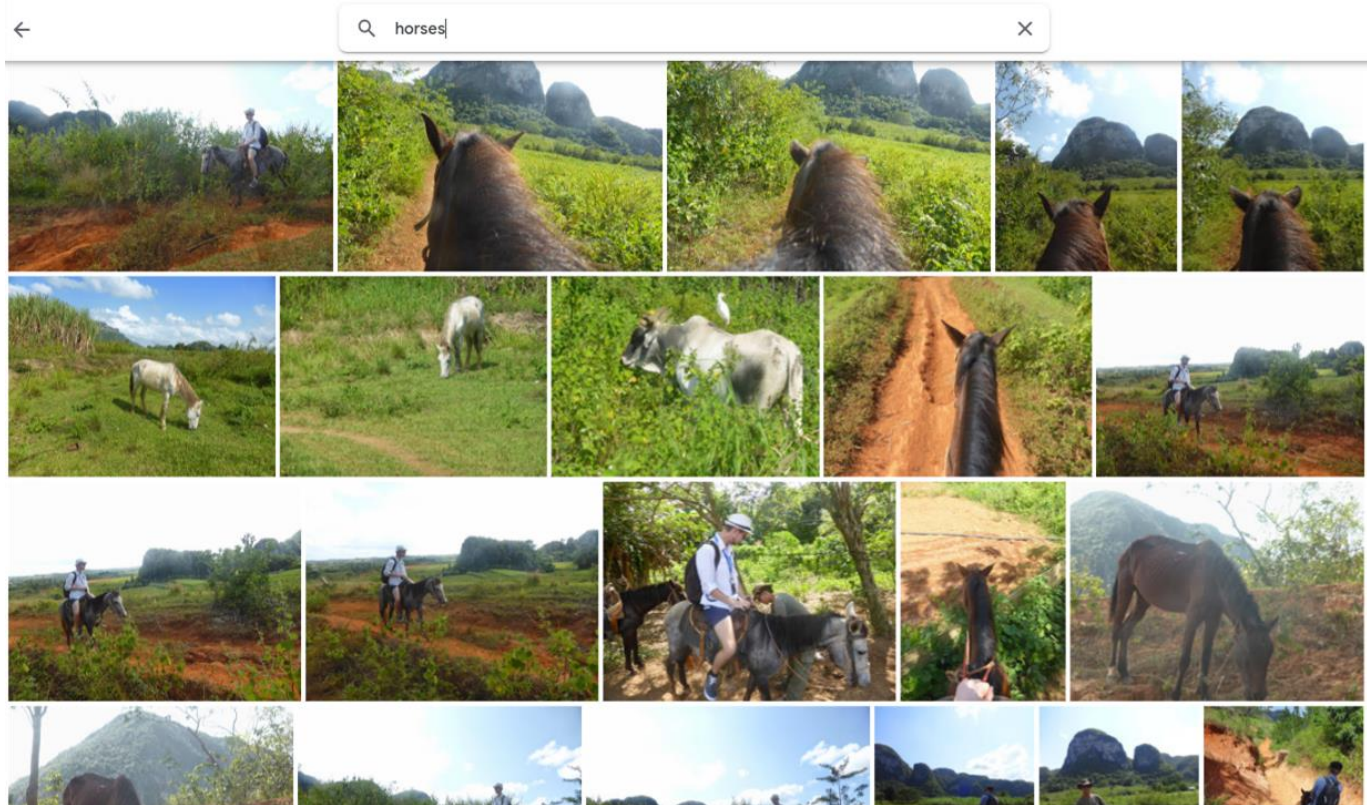
The team behind the technology

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

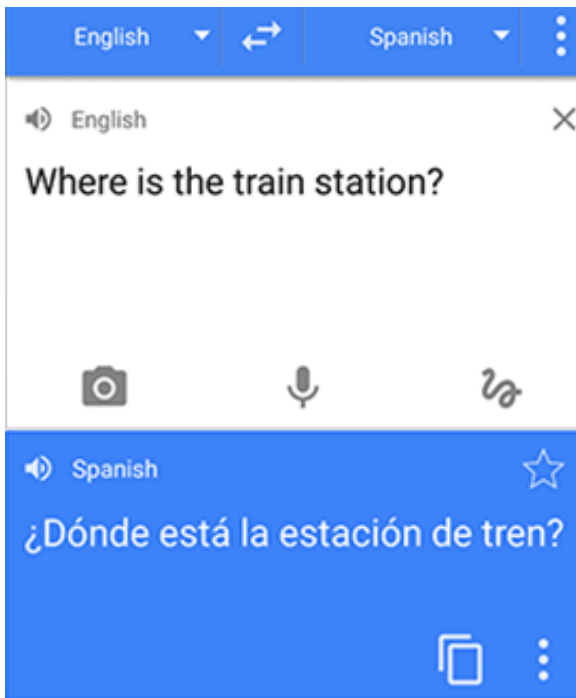


<http://wp.doc.ic.ac.uk/bglocker/>

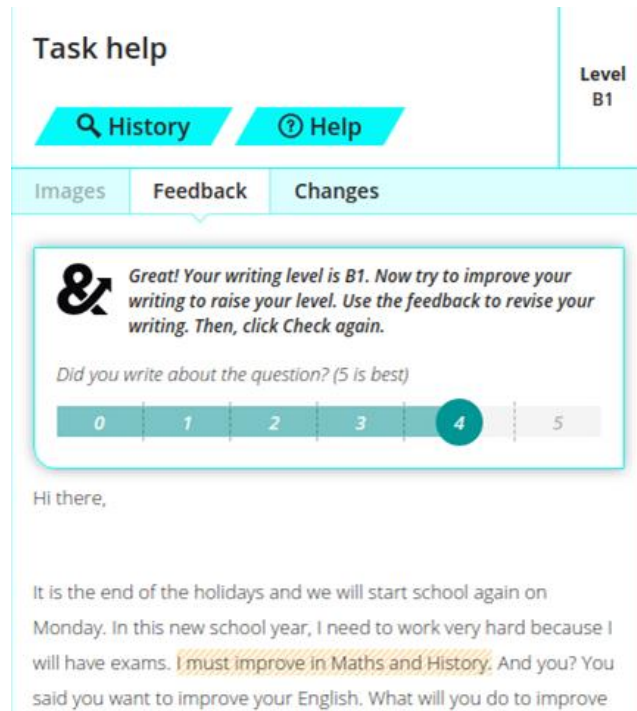
# Application: Analysing images



# Application: Working with language



<https://translate.google.com>



<https://writeandimprove.com>

# Application: Quick draw



Can a neural network learn to recognize doodling?

Help teach it by adding your drawings to the [world's largest doodling data set](#), shared publicly to help with machine learning research.

Let's Draw!

This is on  
A.I.  
Experiment

Made with  
some friends from  
Google

English

[Privacy & Terms](#)

<https://quickdraw.withgoogle.com>



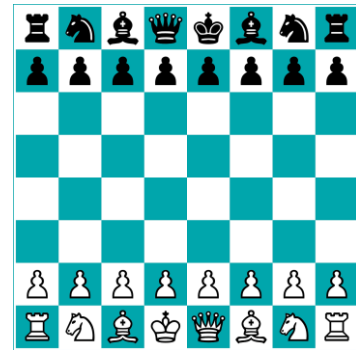
**So, what *exactly* is  
Machine Learning?**

*“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)

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

$$0.70 \times \text{Exam} + 0.30 \times \text{Coursework}$$

The Coursework component is further defined as:

$$0.40 \times \text{CW1} + 0.60 \times \text{CW2}$$

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

$\approx$

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

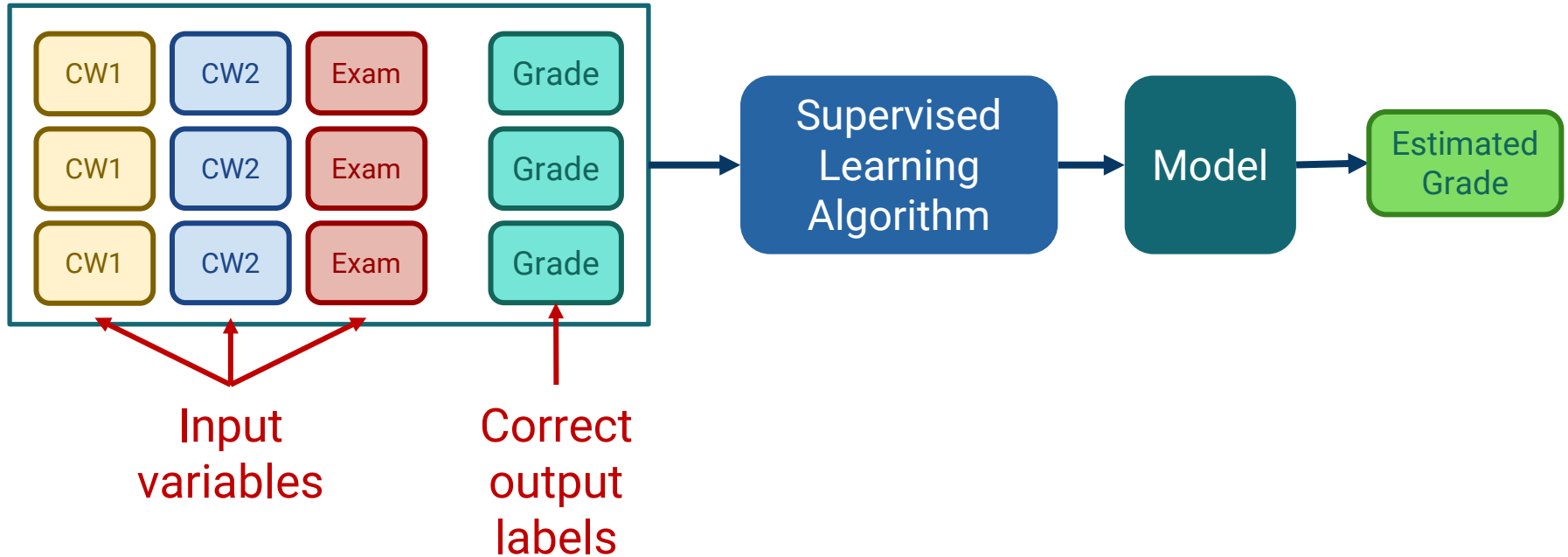
D

😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade

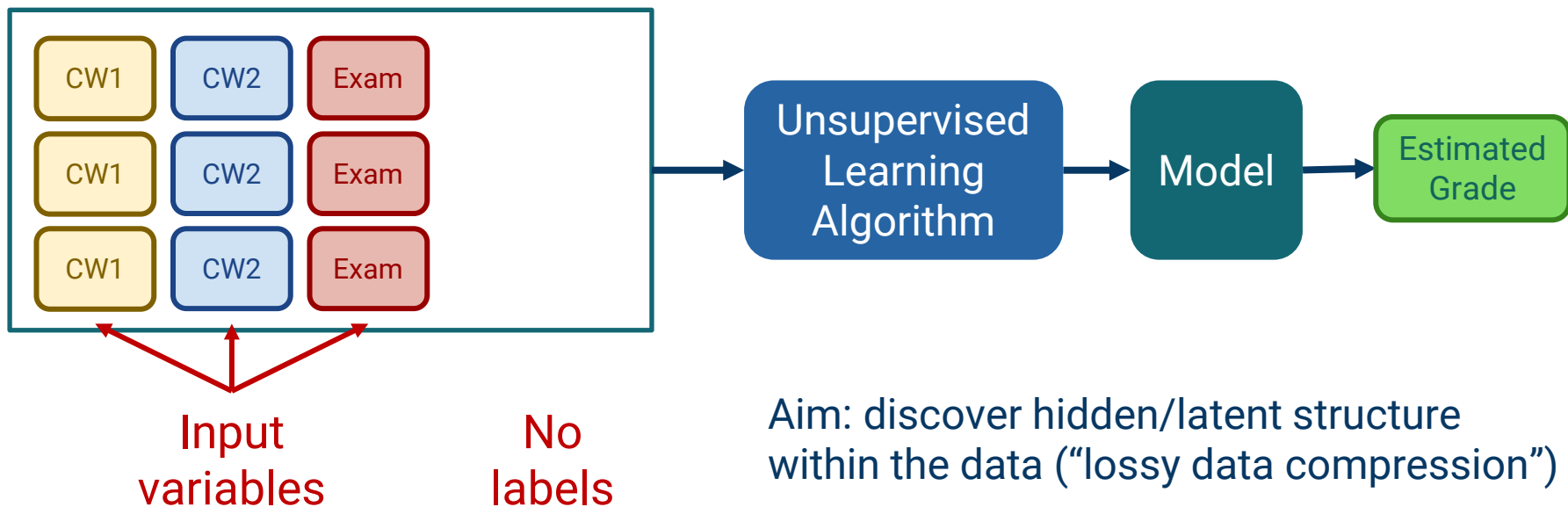
😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade

# Machine Learning settings

# Supervised learning



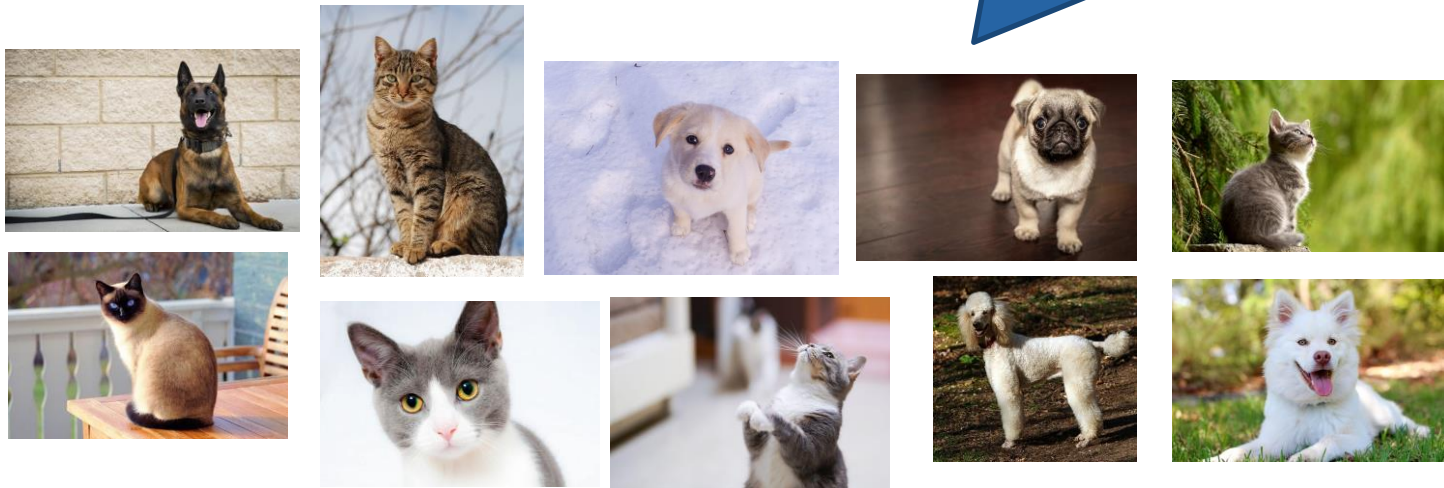
# Unsupervised learning





# Unsupervised learning

Divide us into two groups!



## Clustering

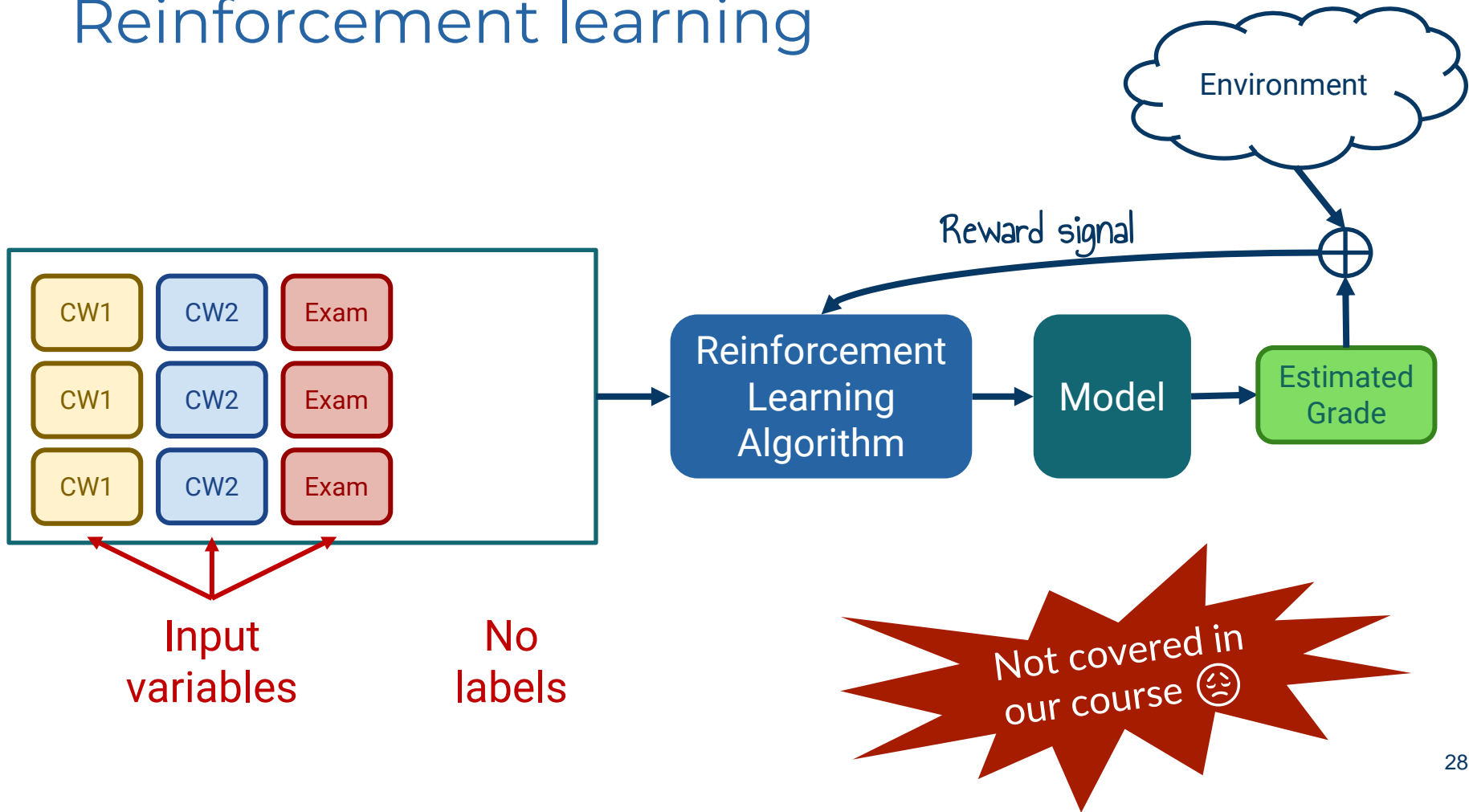
# Unsupervised learning



## Dimensionality reduction

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

# Reinforcement learning



# Reinforcement learning



DQN from DeepMind



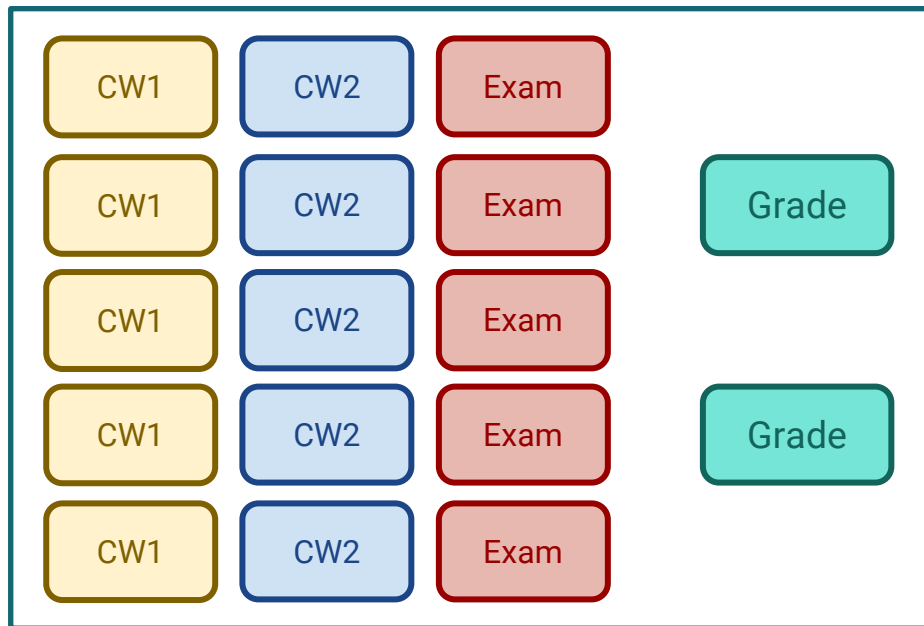
Kormushev et al., 2010

## Policy search

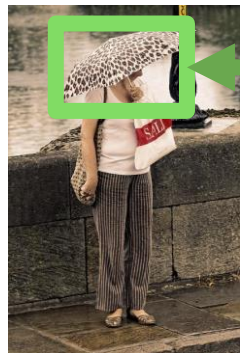
Find which action an agent should take, depending on its current state, to maximise the received rewards.

**Things are not always clear cut!**

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



- Weakly-supervised learning
  - Inexact output labels

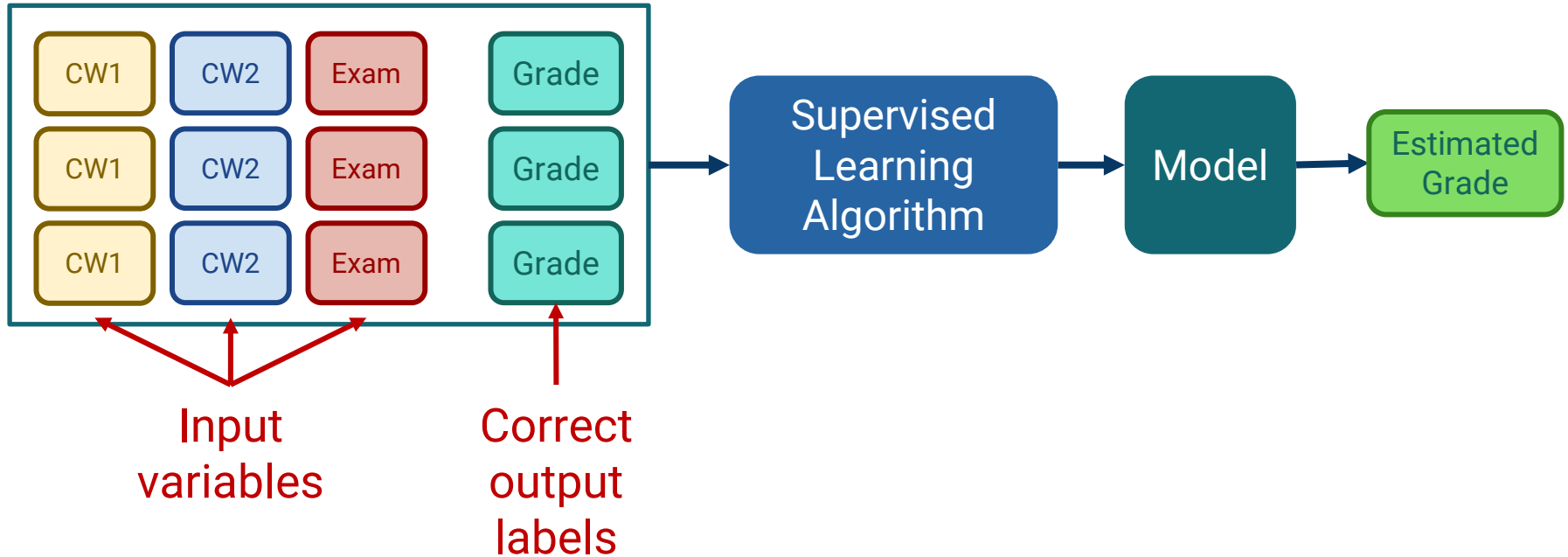


umbrella

umbrella

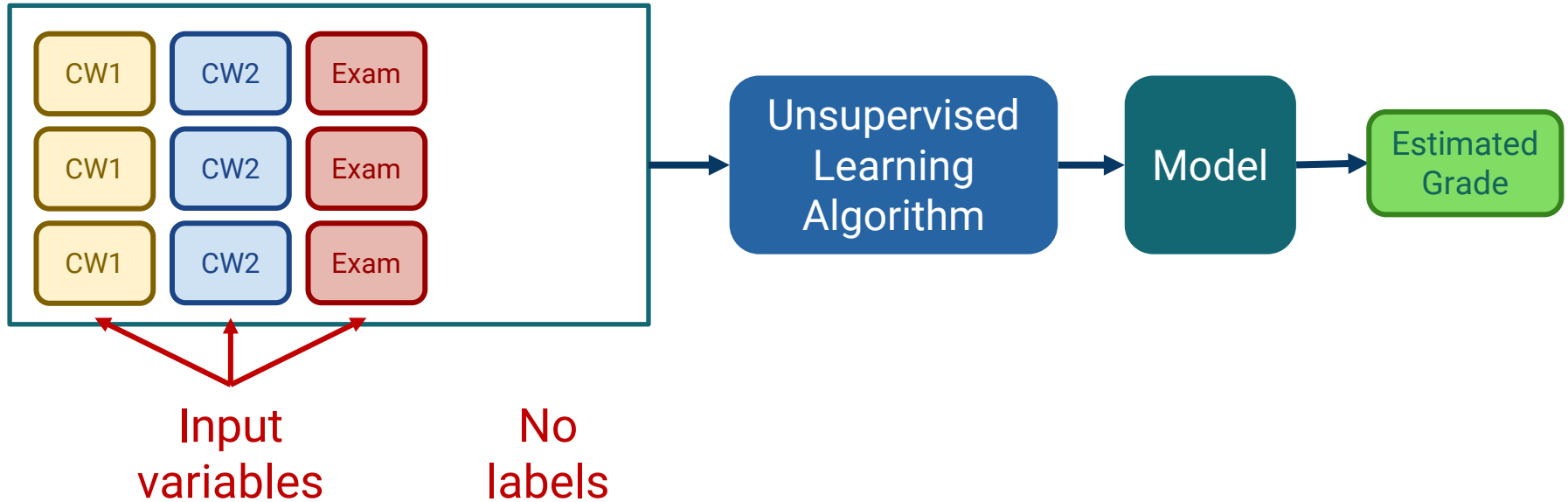


# Supervised learning

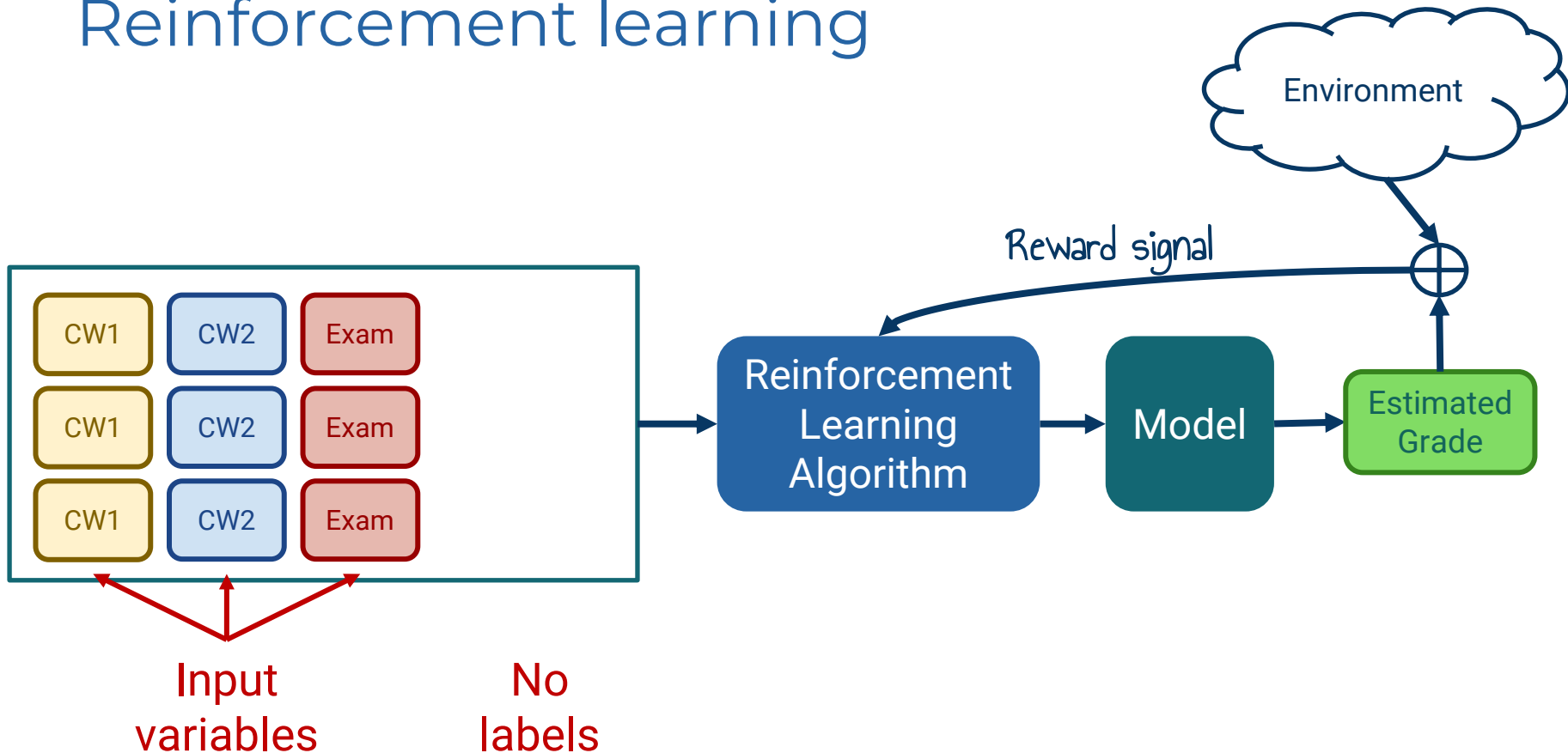




# Unsupervised learning



# Reinforcement learning



# **Classification and regression**

**The two most popular ML tasks**

# Classification

Most common!



$$f(\text{Image of a monarch butterfly}) = y$$

Discrete/  
categorical

# Classification



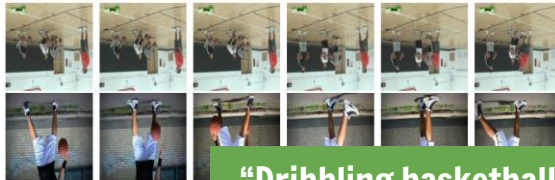
pneumonia



no pneumonia



Photoshopped



"Dribbling basketball"

**Claim** (by Minister Shailesh Vara)

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

**Claim is False**



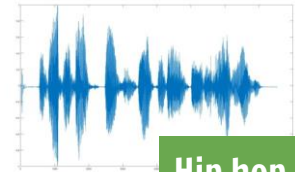
@ [redacted]

Lets kill [redacted] and kill them for fun

#kill [redacted]

7/20/14, 8:05 AM

**Hate speech**



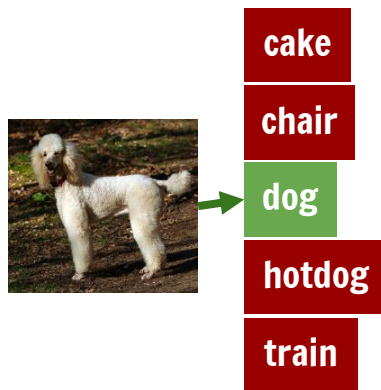
Hip hop

# Classification

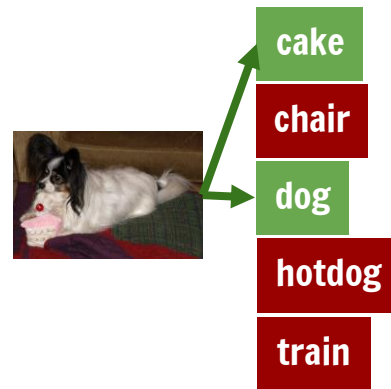
## Binary classification



## Multi-class classification



## Multi-label classification



# Regression



$$f(\text{Monarch Butterfly Image}) = y$$

Real-valued/  
continuous

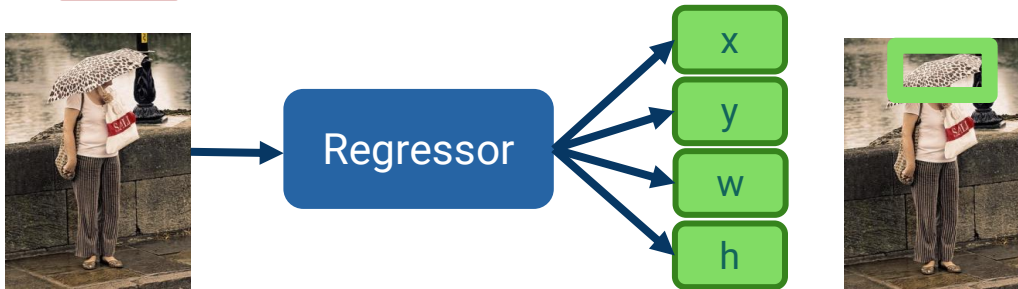
# Regression



**Simple  
regression**



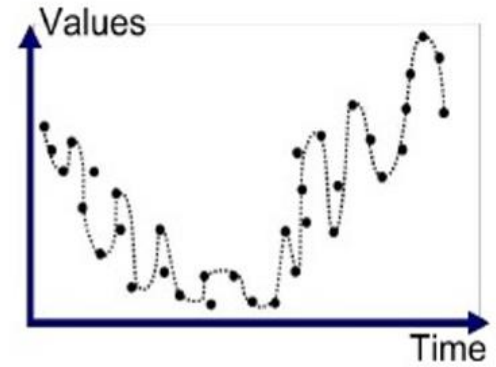
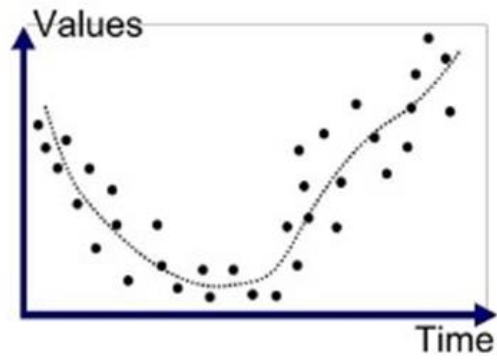
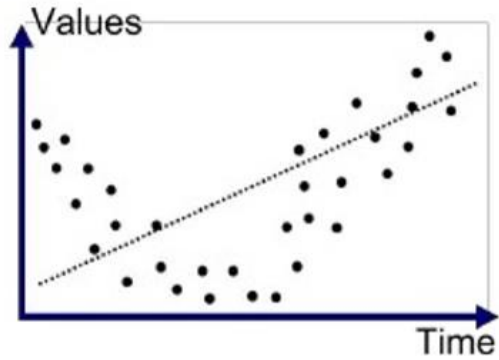
**Multiple  
regression**



**Multivariate  
regression**



# Regression



## Classification

 $f($  $) = y$ 

Discrete/  
categorical

## Regression

 $f($  $) = y$ 

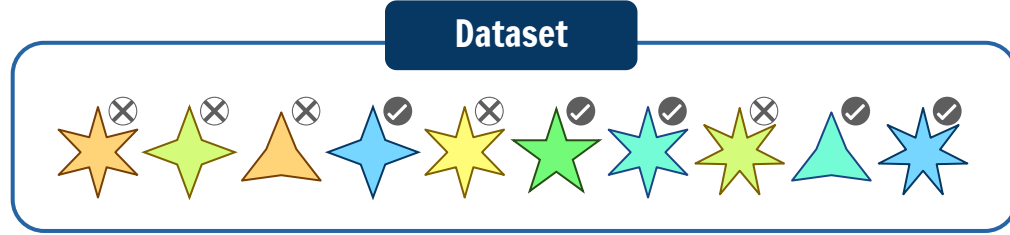
Real-valued/  
continuous

# **Supervised learning**

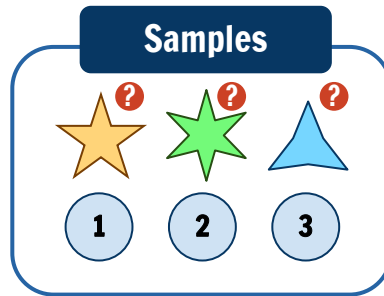
## **The pipeline**

# Are you a good binary classifier?



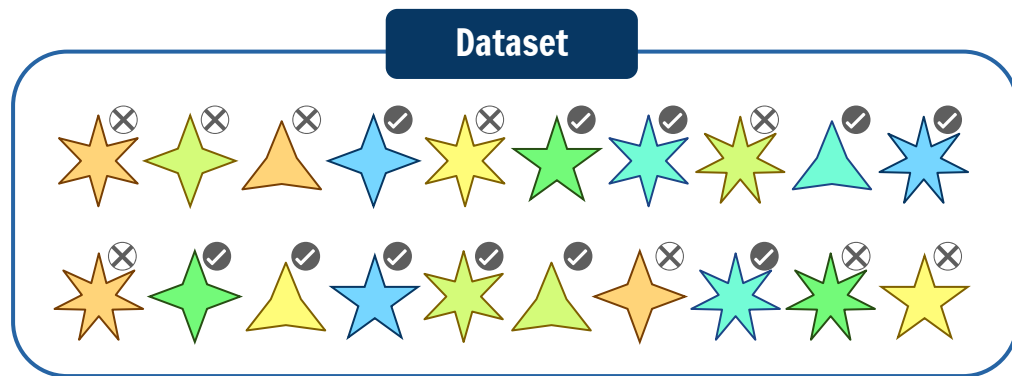


Try this!

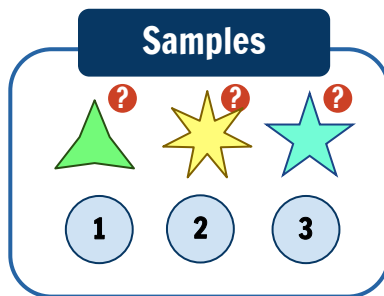


# Are you a good binary classifier?





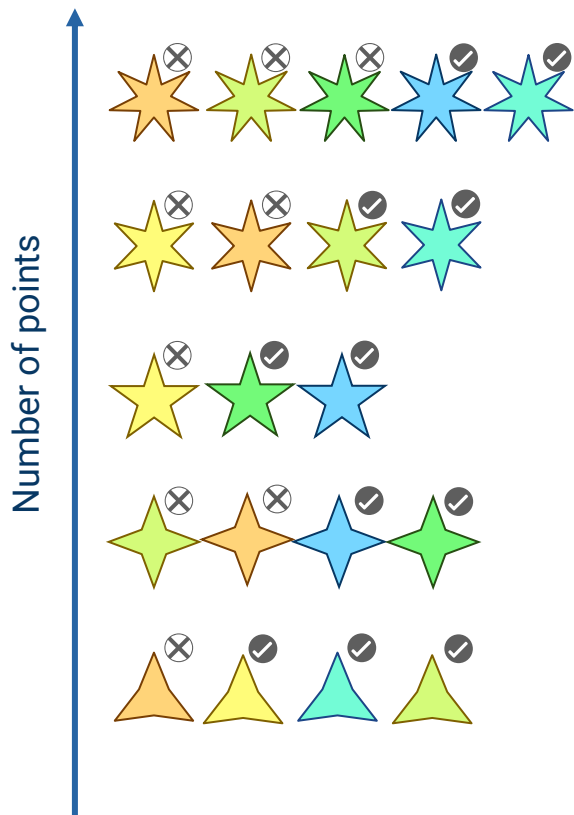
Try this!

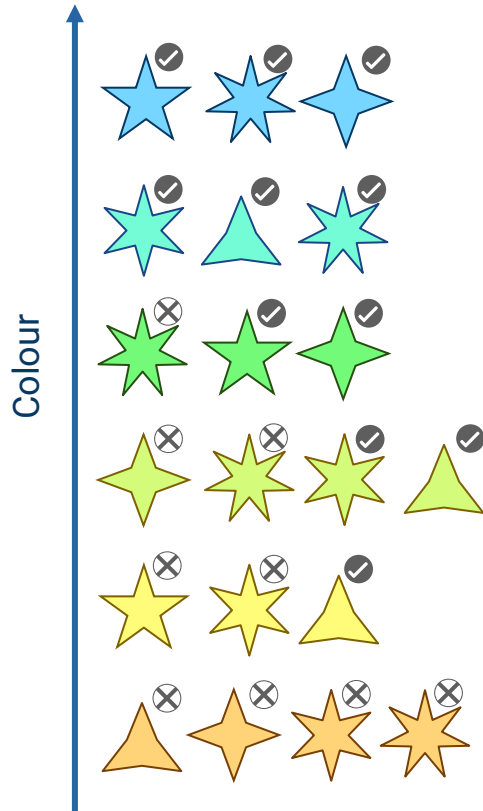


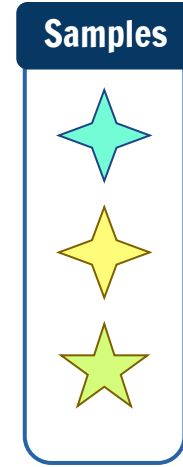
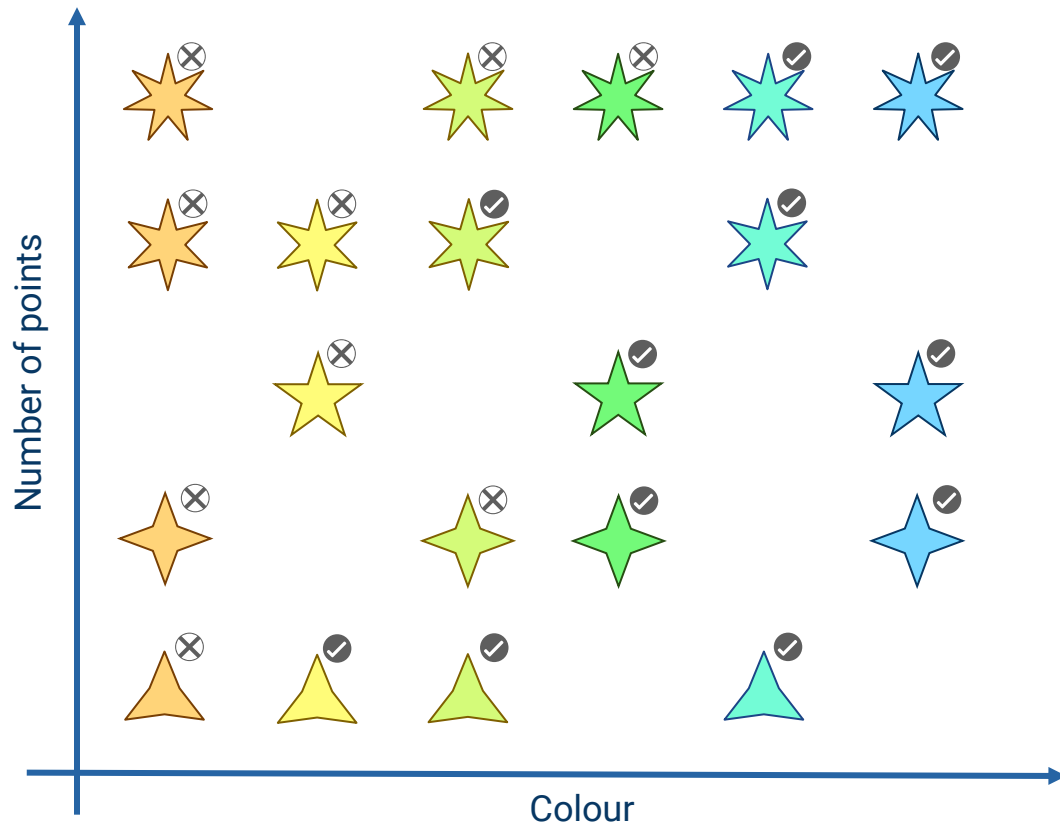
# Are you still a good binary classifier?



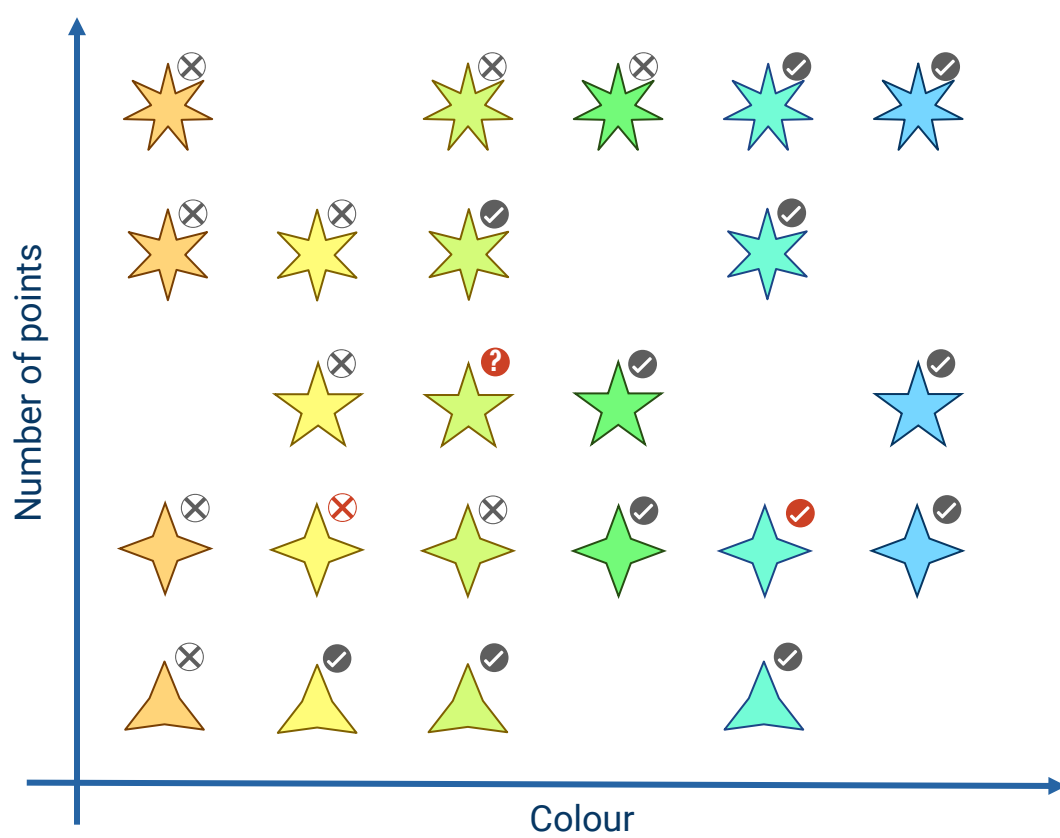






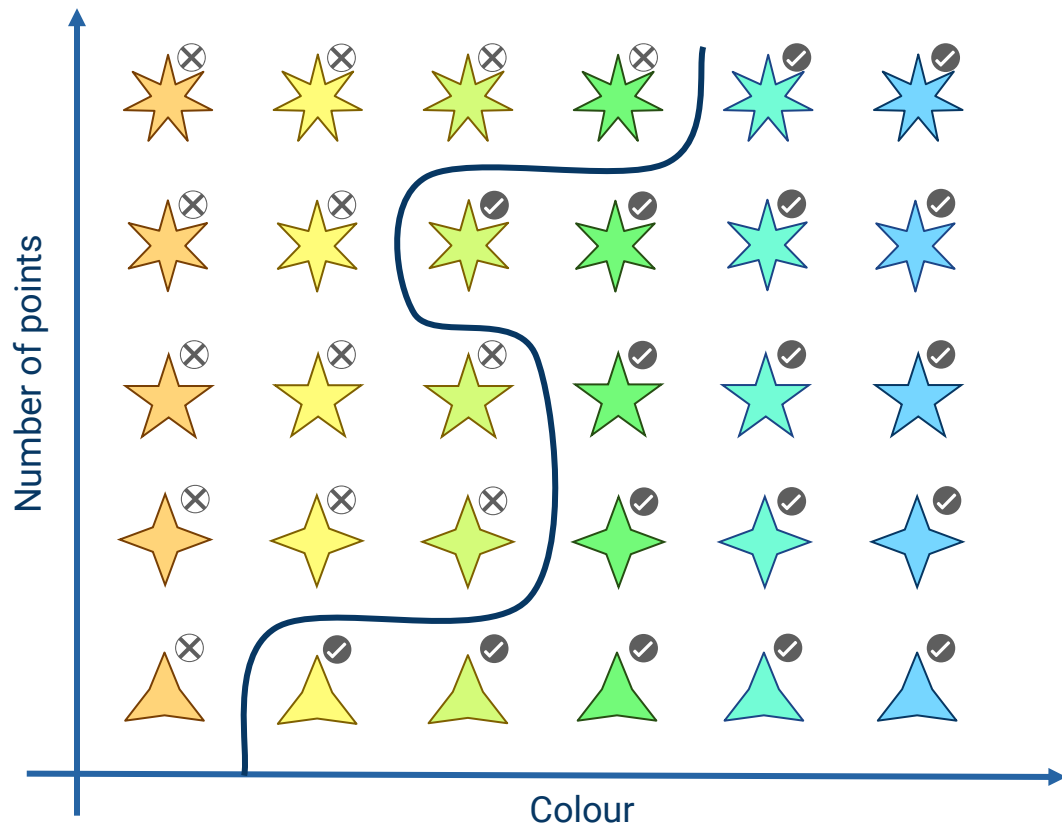


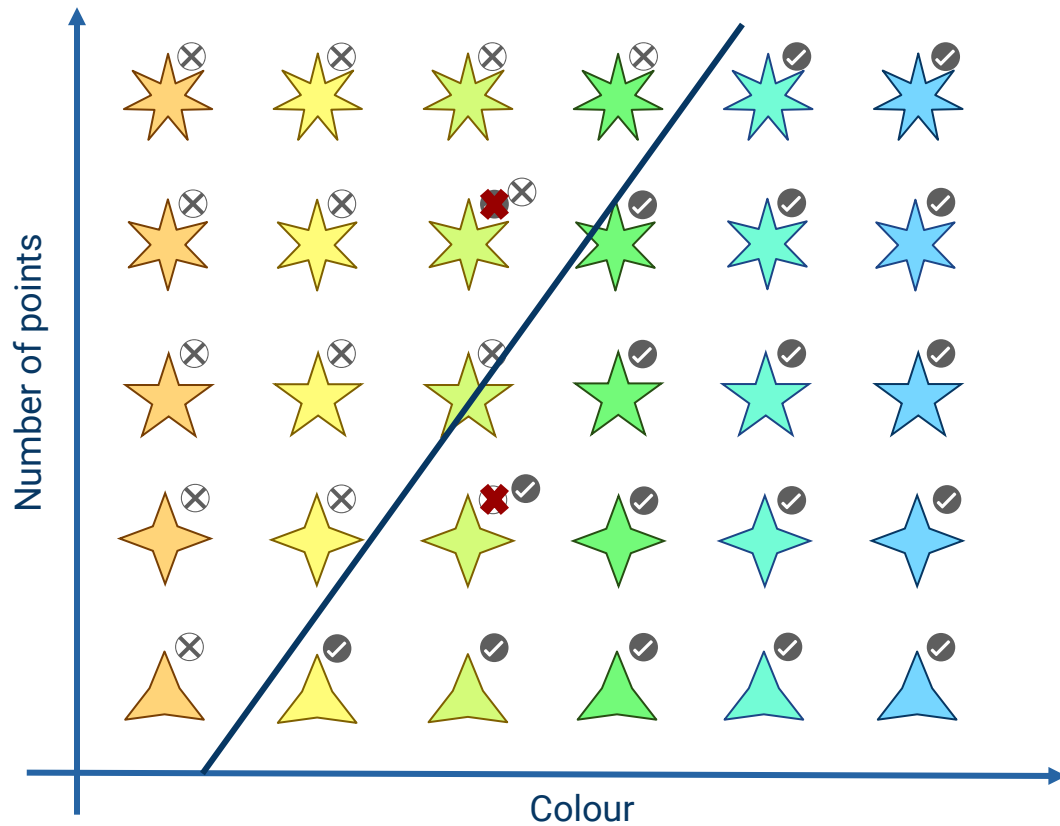
Try this!



**Samples**

- 
- 
-

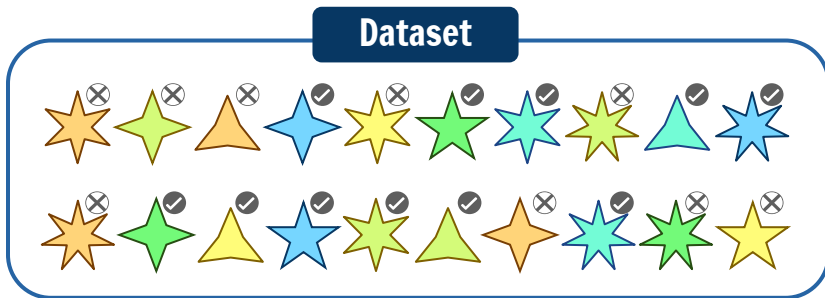
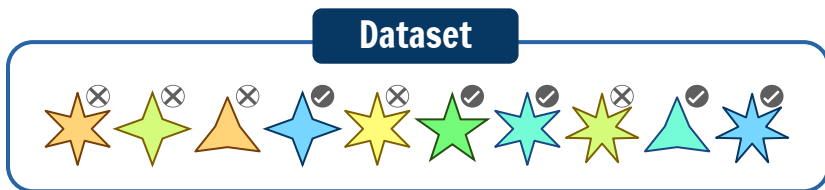




Linear  
classifier

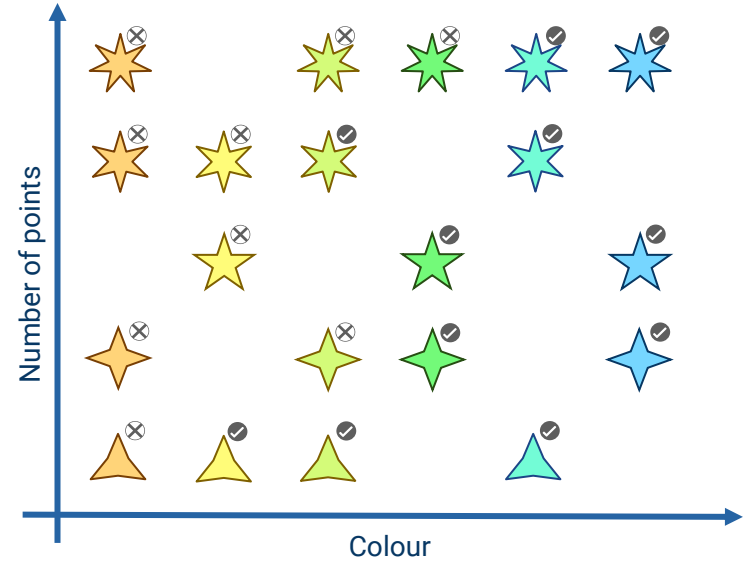
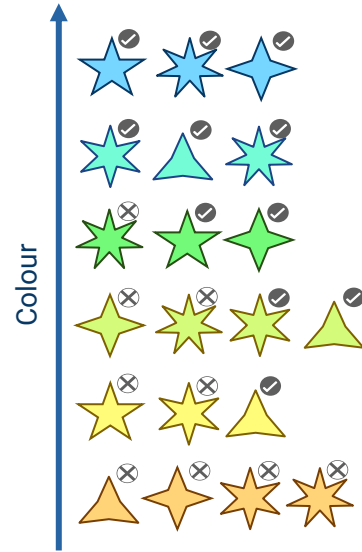
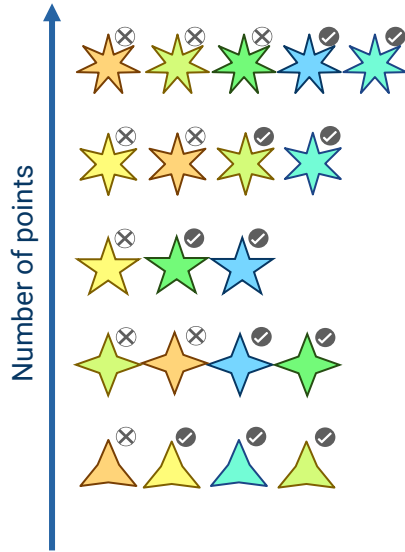
**What have we learnt?**

# More data == more accurate predictions

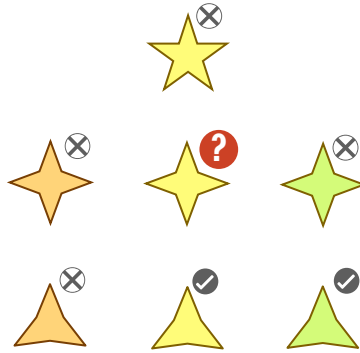




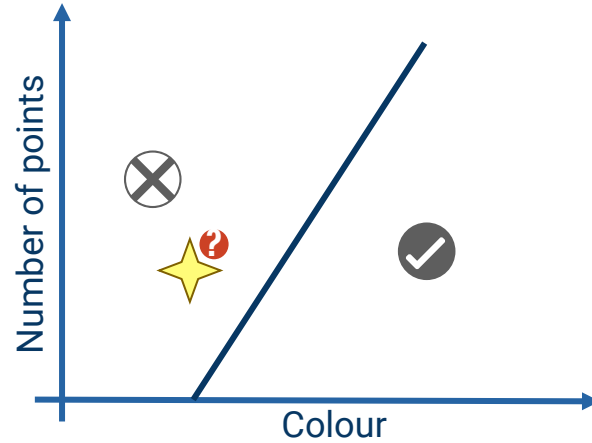
# Selecting good features is crucial!



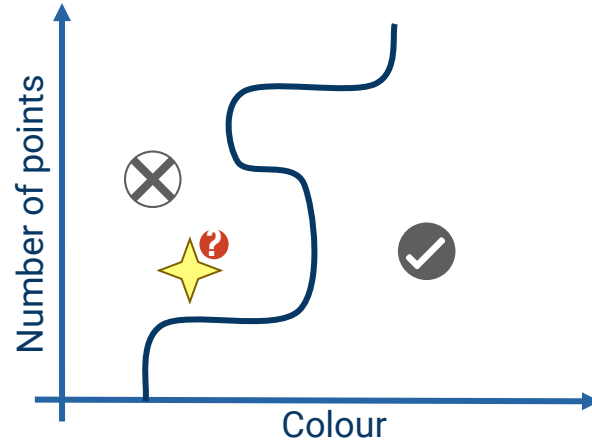
# Classifiers make predictions differently



# Classifiers make predictions differently



# Classifiers make predictions differently



# The supervised learning pipeline

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

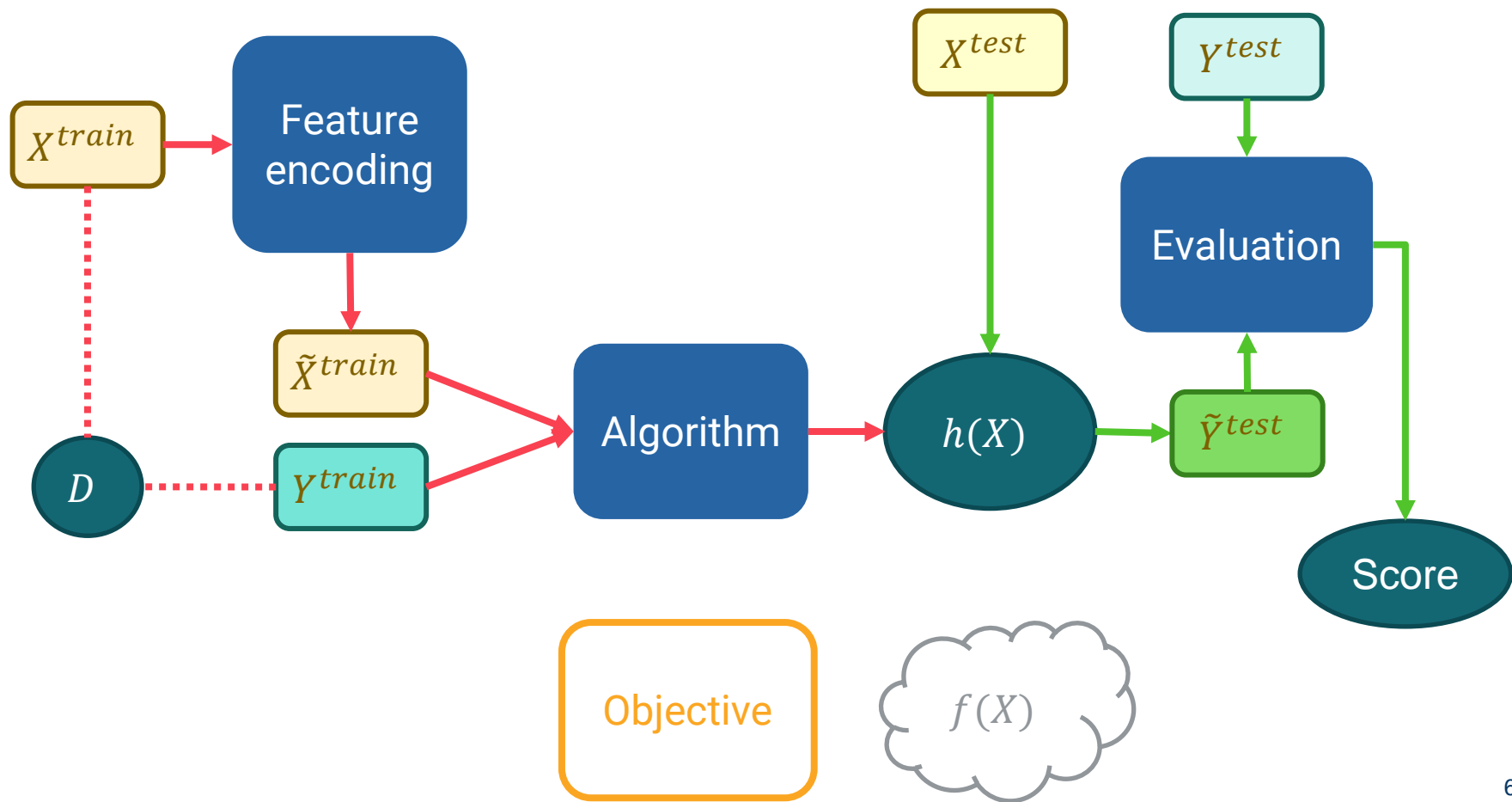
$\approx$

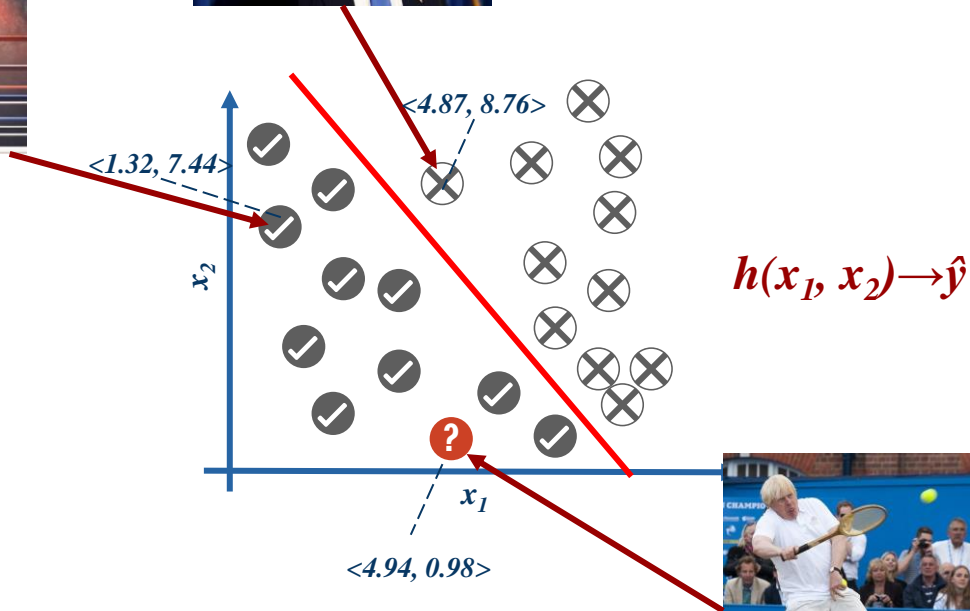
$$h(\text{CW1}, \text{CW2}, \text{Exam} \mid \text{D}) = \text{Estimated Grade}$$

D

😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade

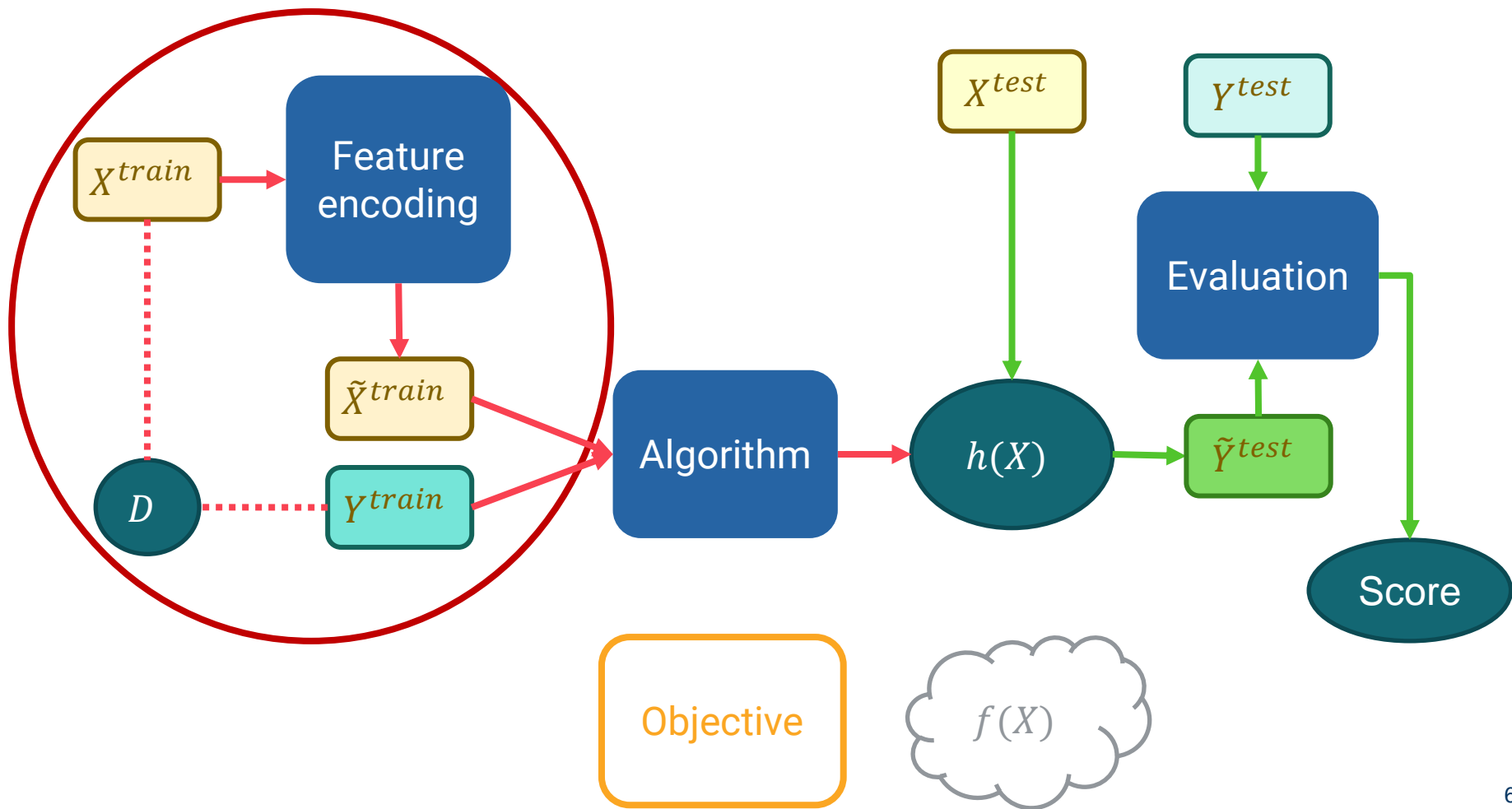
😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade
😊	CW1	CW2	Exam	Grade





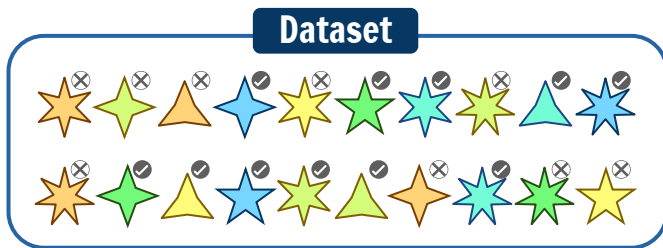


# Feature encoding



# Understanding your data

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

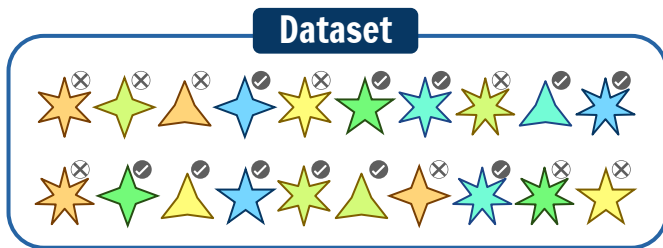


# Understanding your data

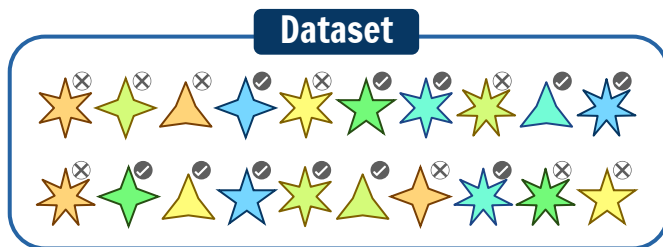
- Always examine your data!



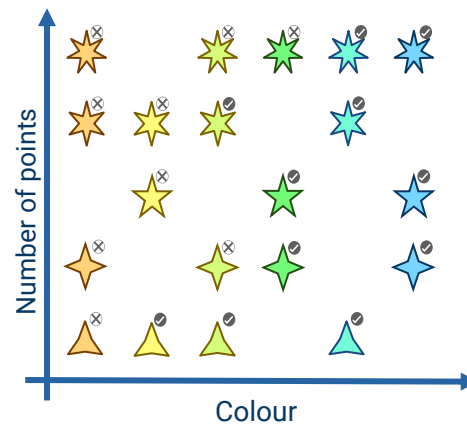
- Clues to help you design your classifier
- Distribution of class labels?
  - Balanced?
  - Imbalanced?



## Measurement space



## Feature space



# Understanding your features

features  attributes

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

**N instances**

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

**K-dimensional features**

Categorical

Integers

Real numbers

**Normalisation**

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

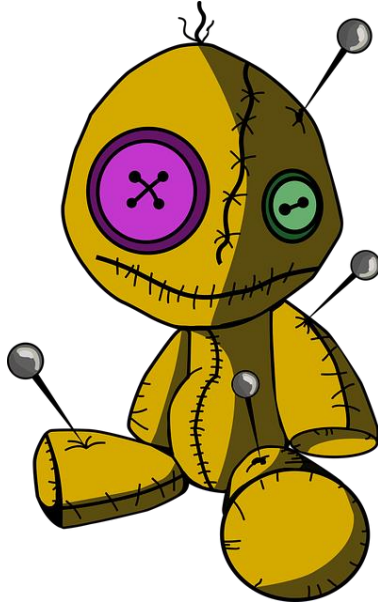
This form of normalisation is called **standardisation**



**More features == better?**

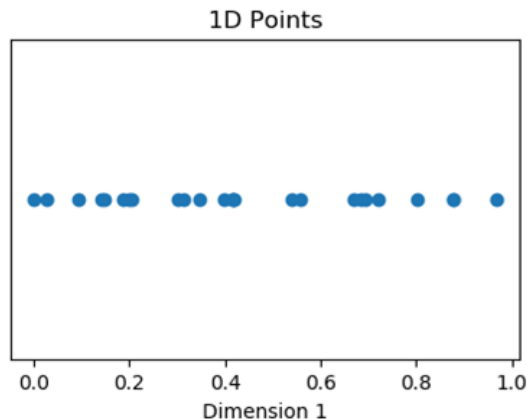
**Only up to a certain point!**

# The curse of dimensionality

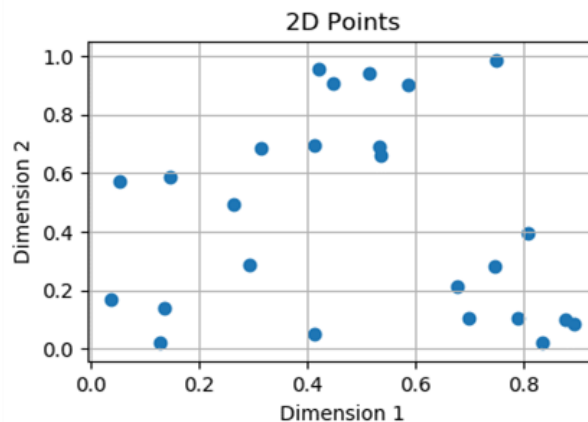




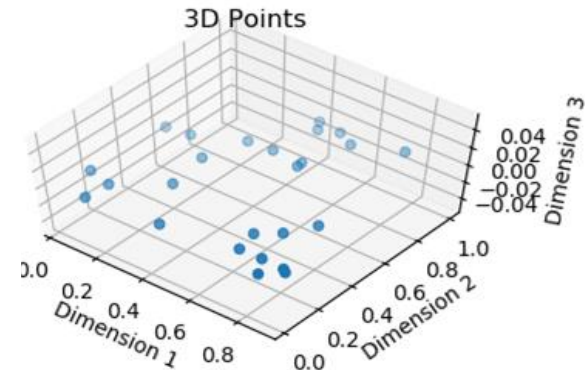
# The curse of dimensionality



**Increased  
computational  
complexity**



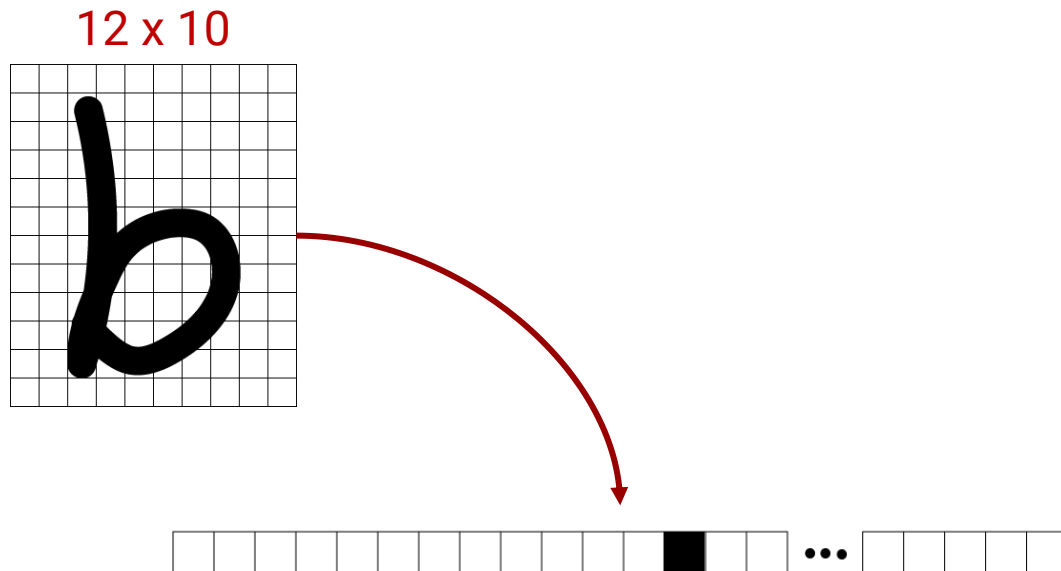
**Data  
sparsity**



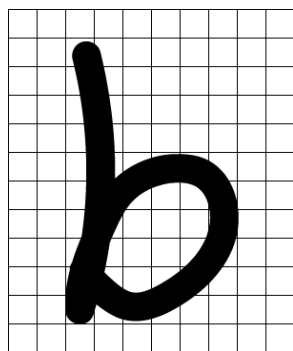
**Overfitting**

# Mapping data to features

- Raw measurements: sometimes inefficient



# Mapping data to features



**feature  
selection**  
(subset)



Width	6
Height	10
# "on" pixels	27
Ratio	0.6

**feature  
extraction**  
(new)

Manual  
(feature engineering)

Automatic

# 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.

We
are
going
to
level
up
and
unite
...
our
amazing
NHS
.

# Mapping data to features



**Boris Johnson** ✓ @BorisJohnson · 31 Dec 2019

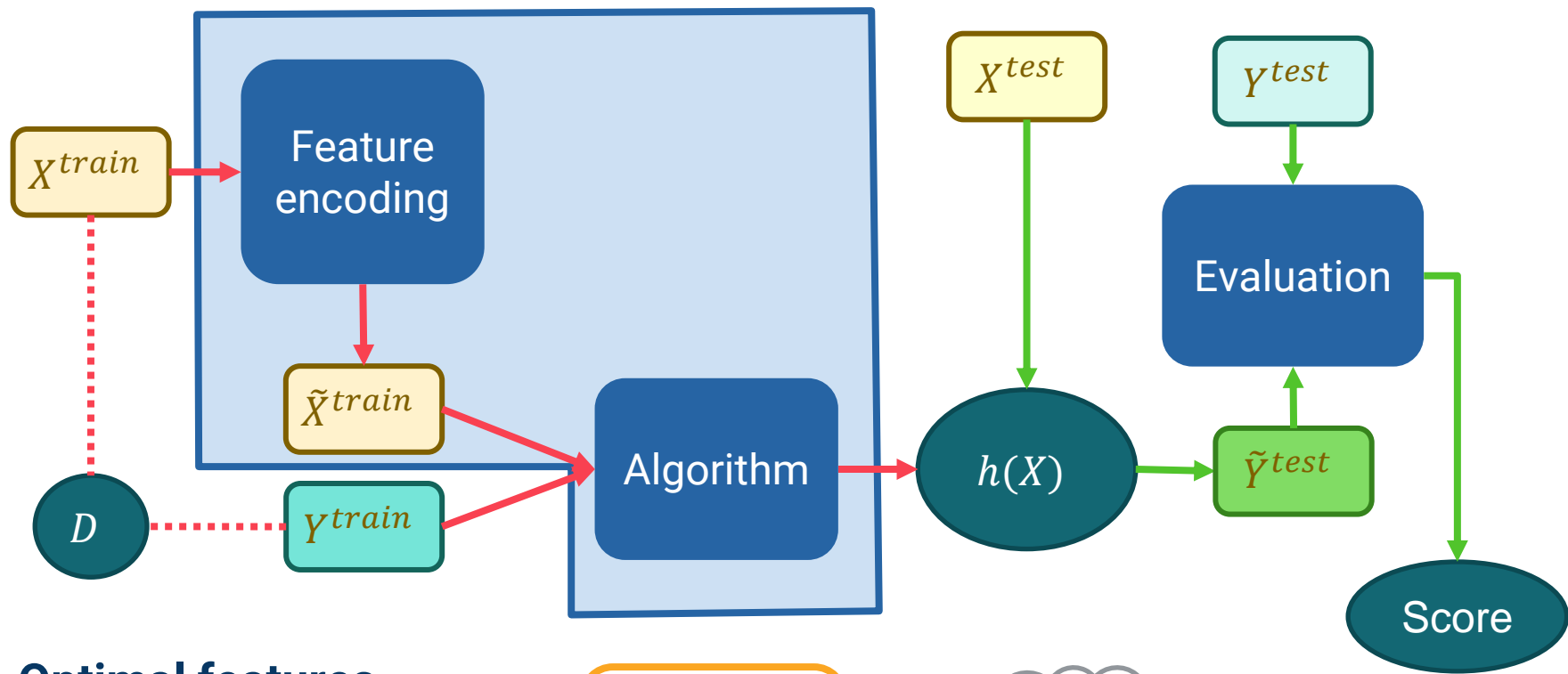
We are going to level up and unite our country. We are going to get Brexit done and deliver the change that people voted for. And we are going to protect and invest in our amazing NHS.

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**

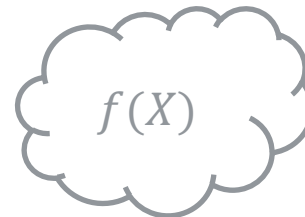
**The deep learning era**

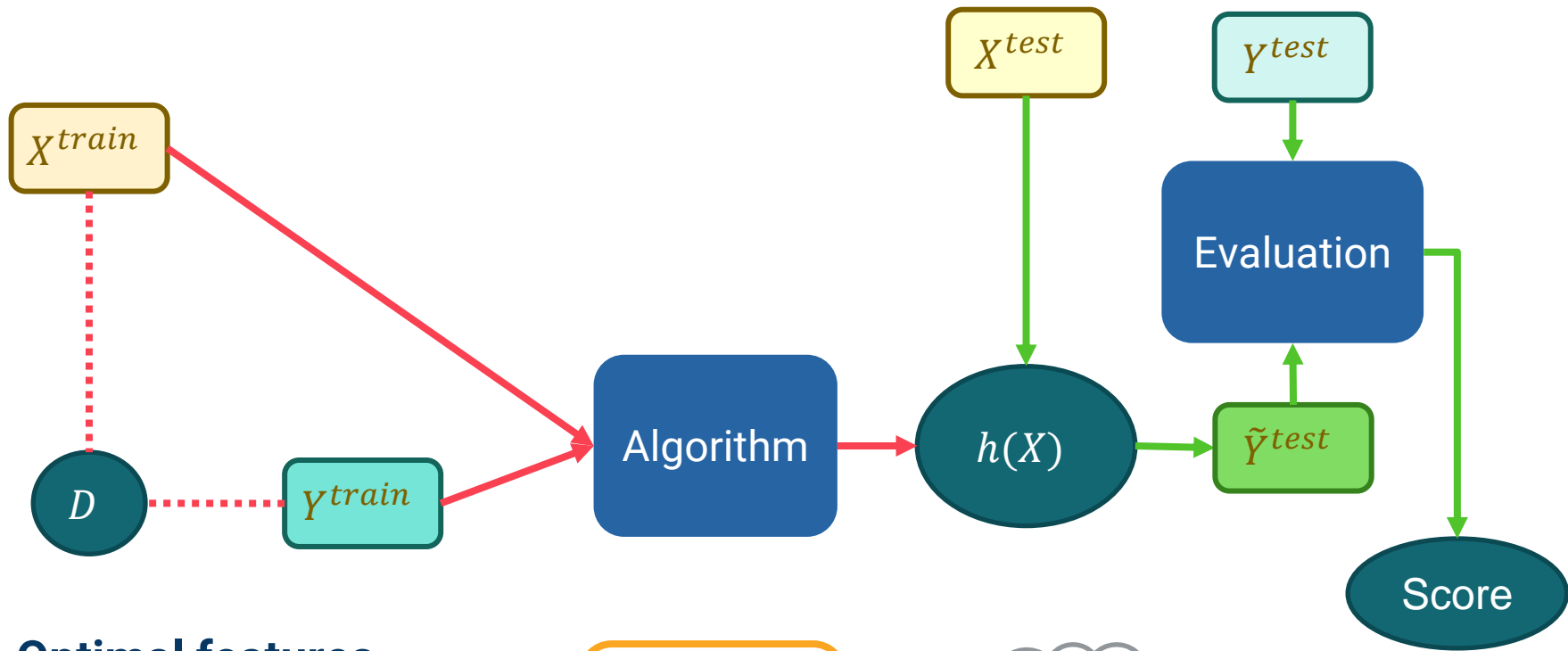


**Optimal features  
automatically learnt**

**Optimised for task**

Objective





**Optimal features  
automatically learnt**

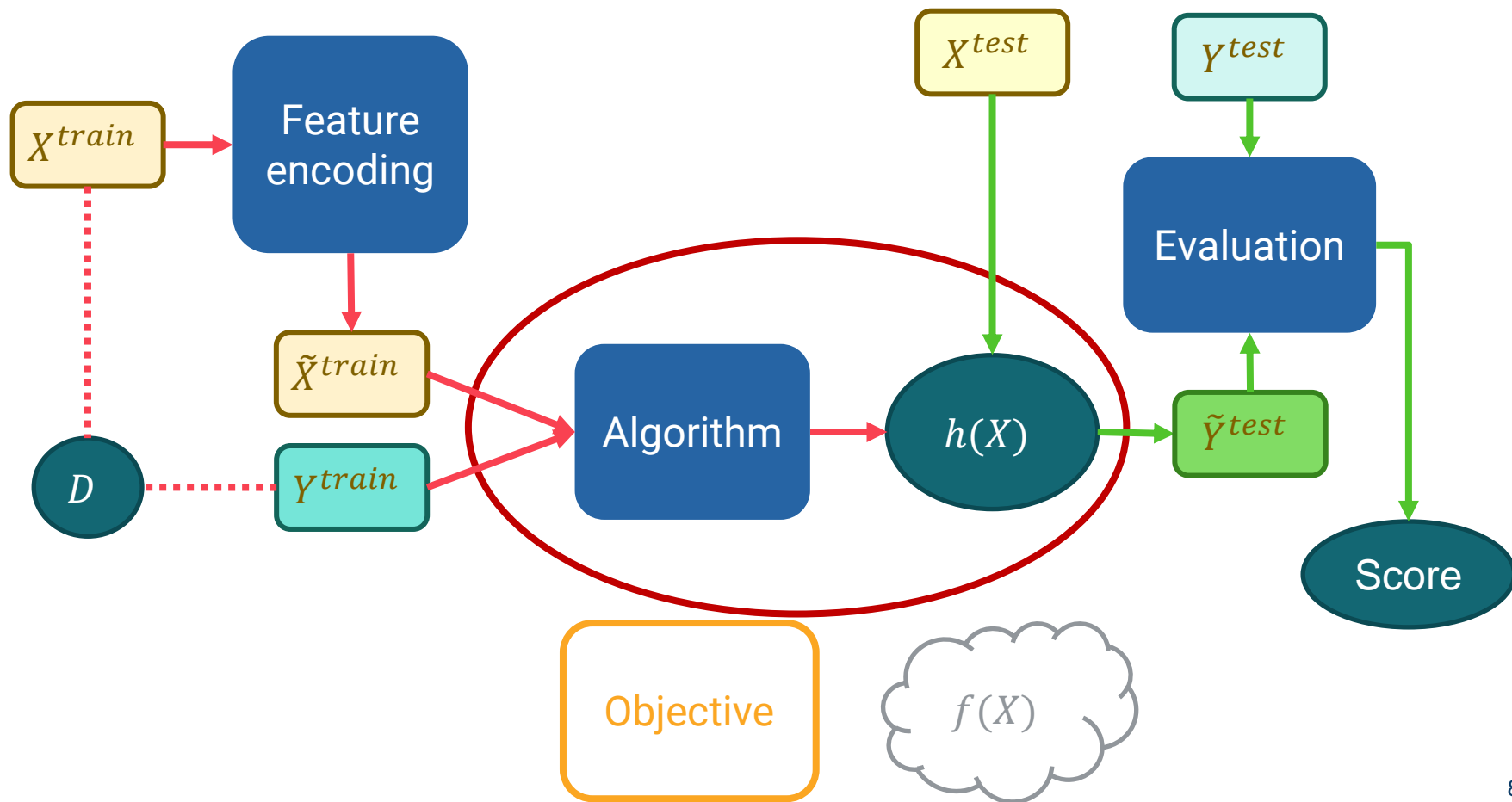
**Optimised for task**

Objective

$f(X)$



# Machine learning algorithms



# Lazy vs. Eager Learning

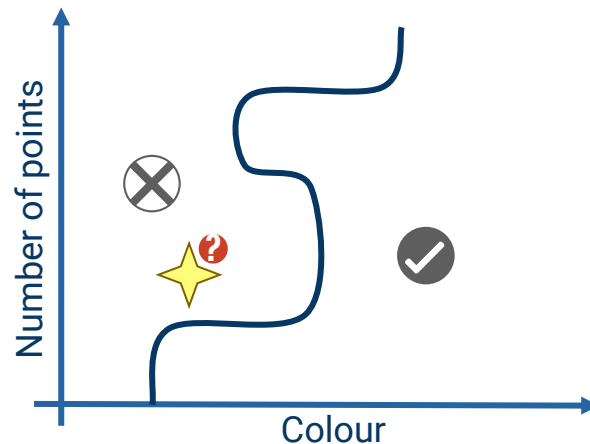
## Lazy Learner

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



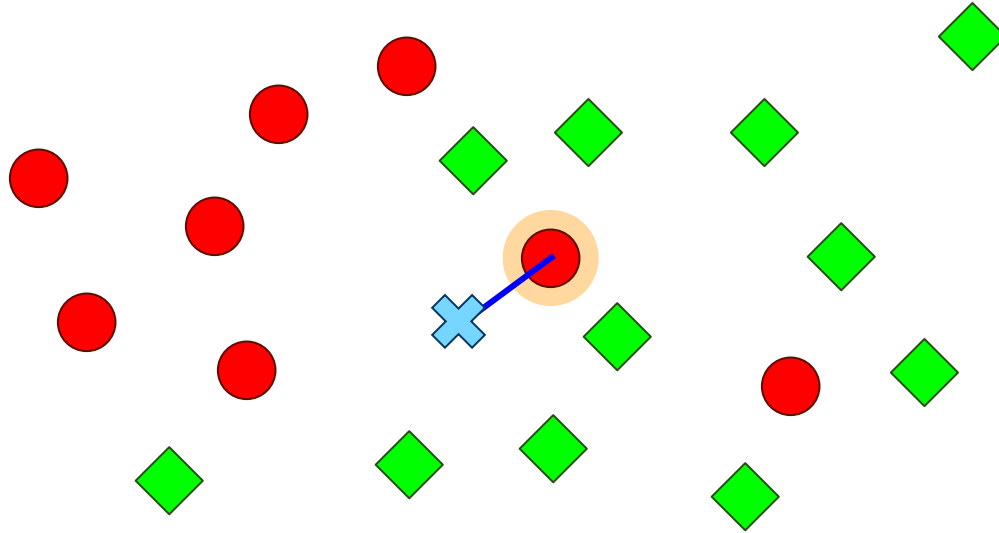
## Eager Learner

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



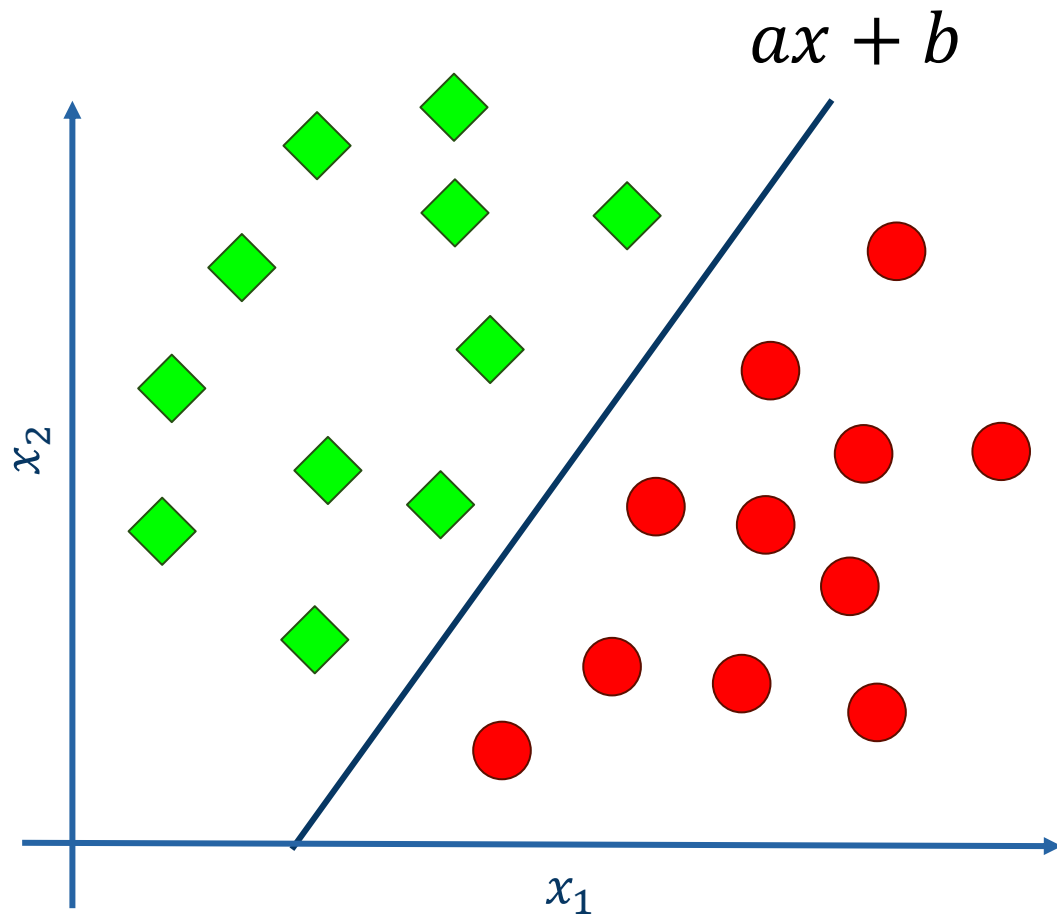
# Non-parametric model

## Nearest neighbour



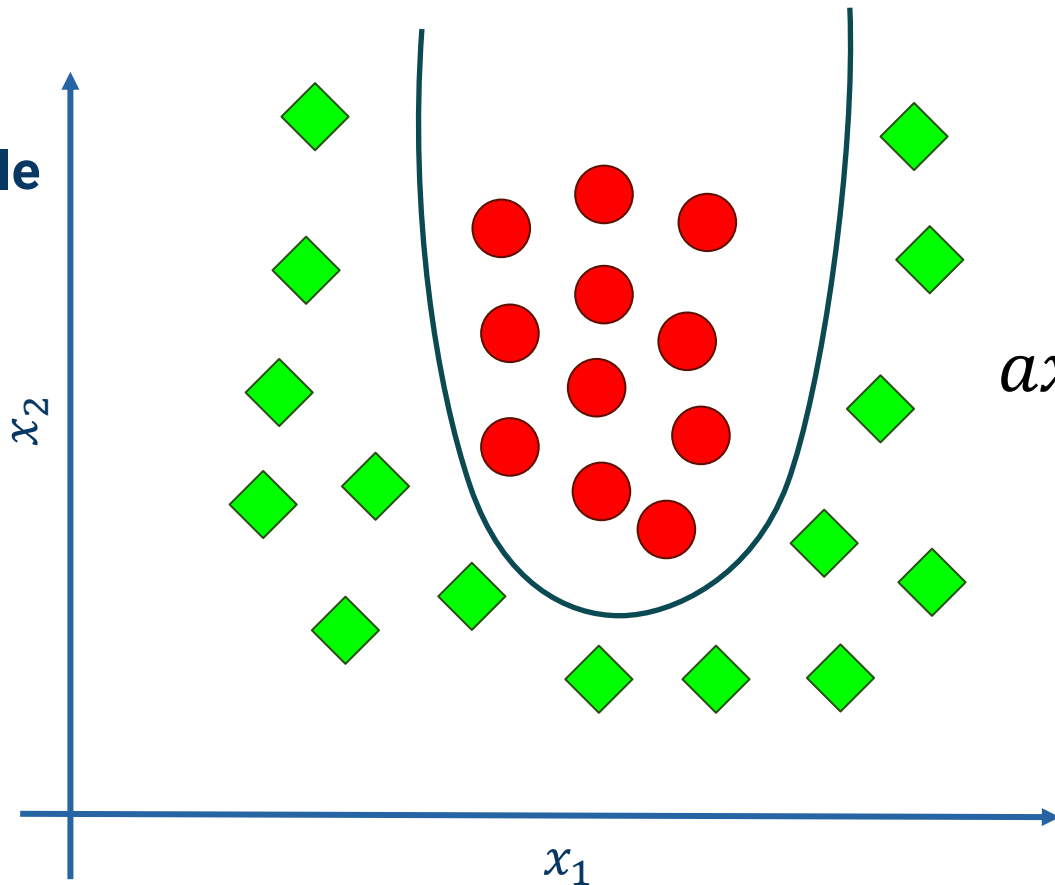
# Linear model

**Linearly  
separable**



# Nonlinear model

**Linearly  
non-separable**

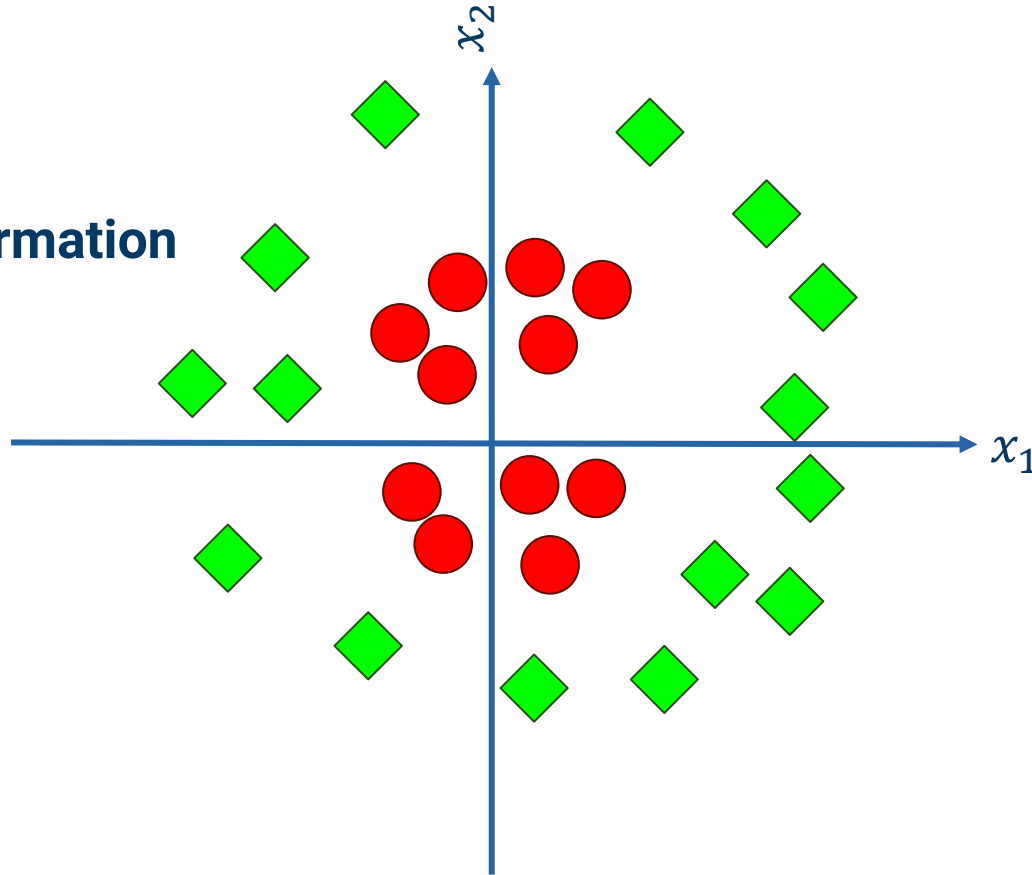


~~$a$~~   ~~$b$~~

$$ax^2 + bx + c$$

# Nonlinear model

**Feature  
space  
transformation**



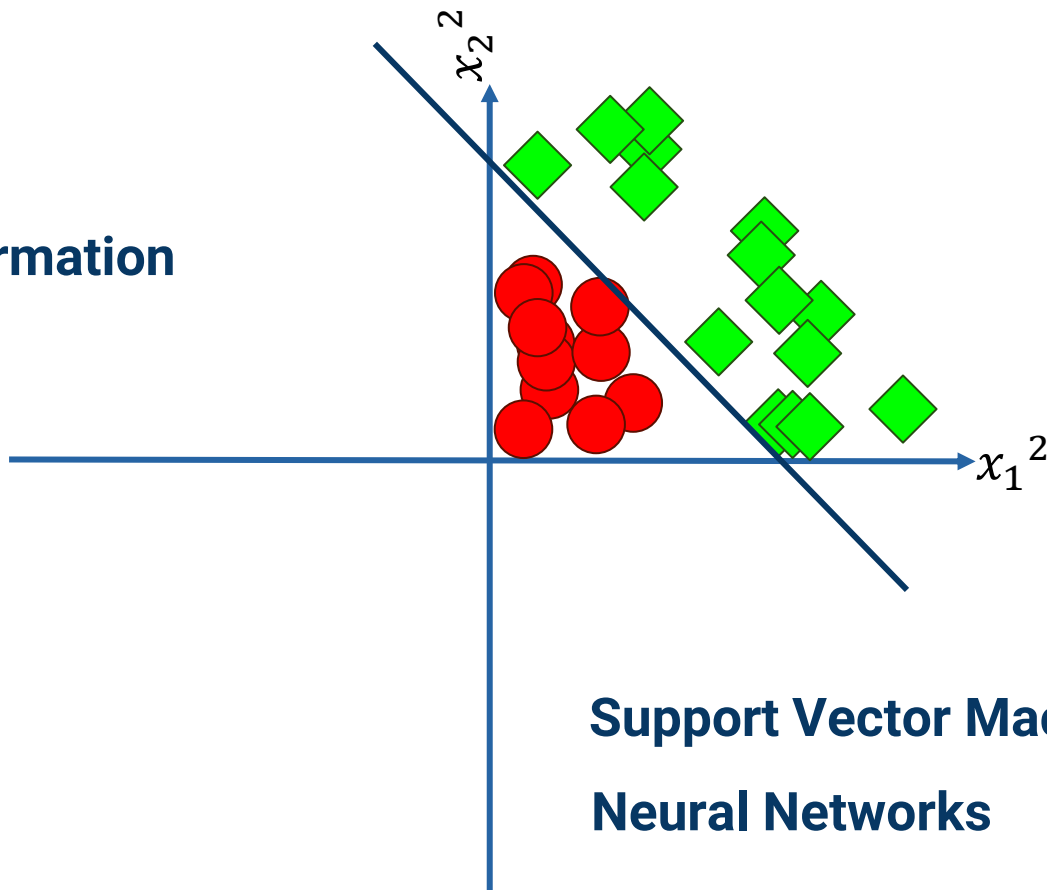
$$(x_1, x_2)$$

↓

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

# Nonlinear model

**Feature  
space  
transformation**



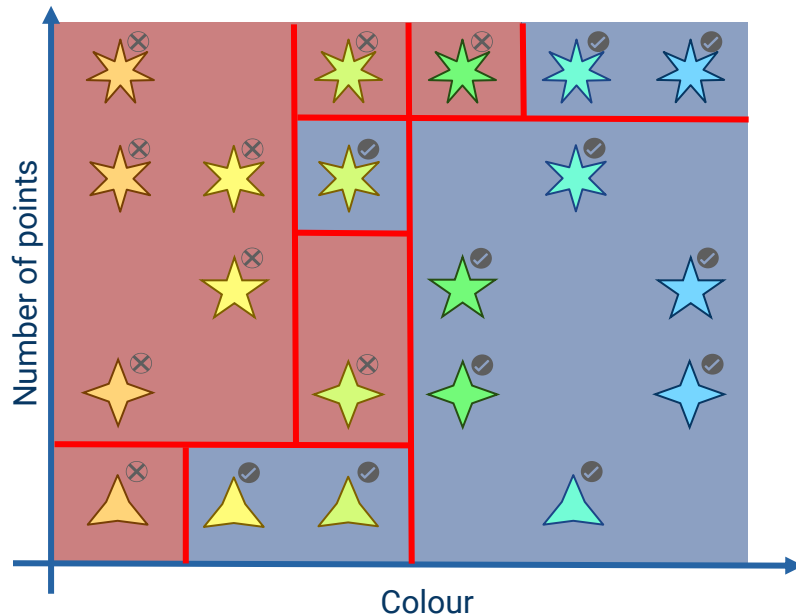
**Support Vector Machines (SVM)**

**Neural Networks**



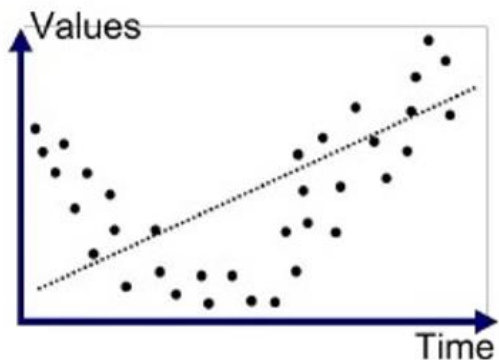
## Combine multiple simple classifiers

# Decision Trees

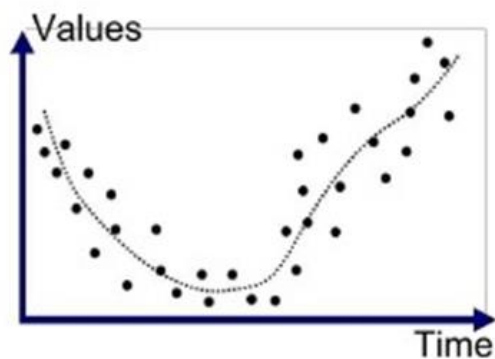


# **Bias-variance trade-off**

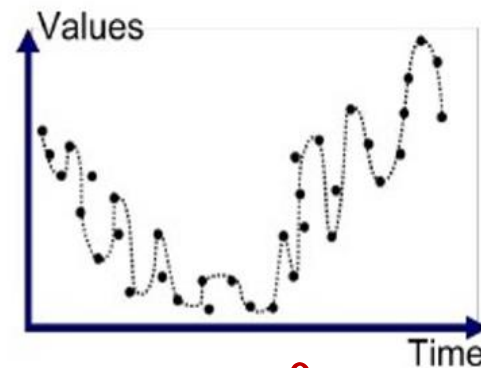
**One of the most important ML concepts!**



Underfit



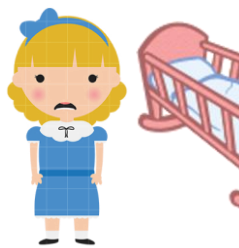
Robust, good fit



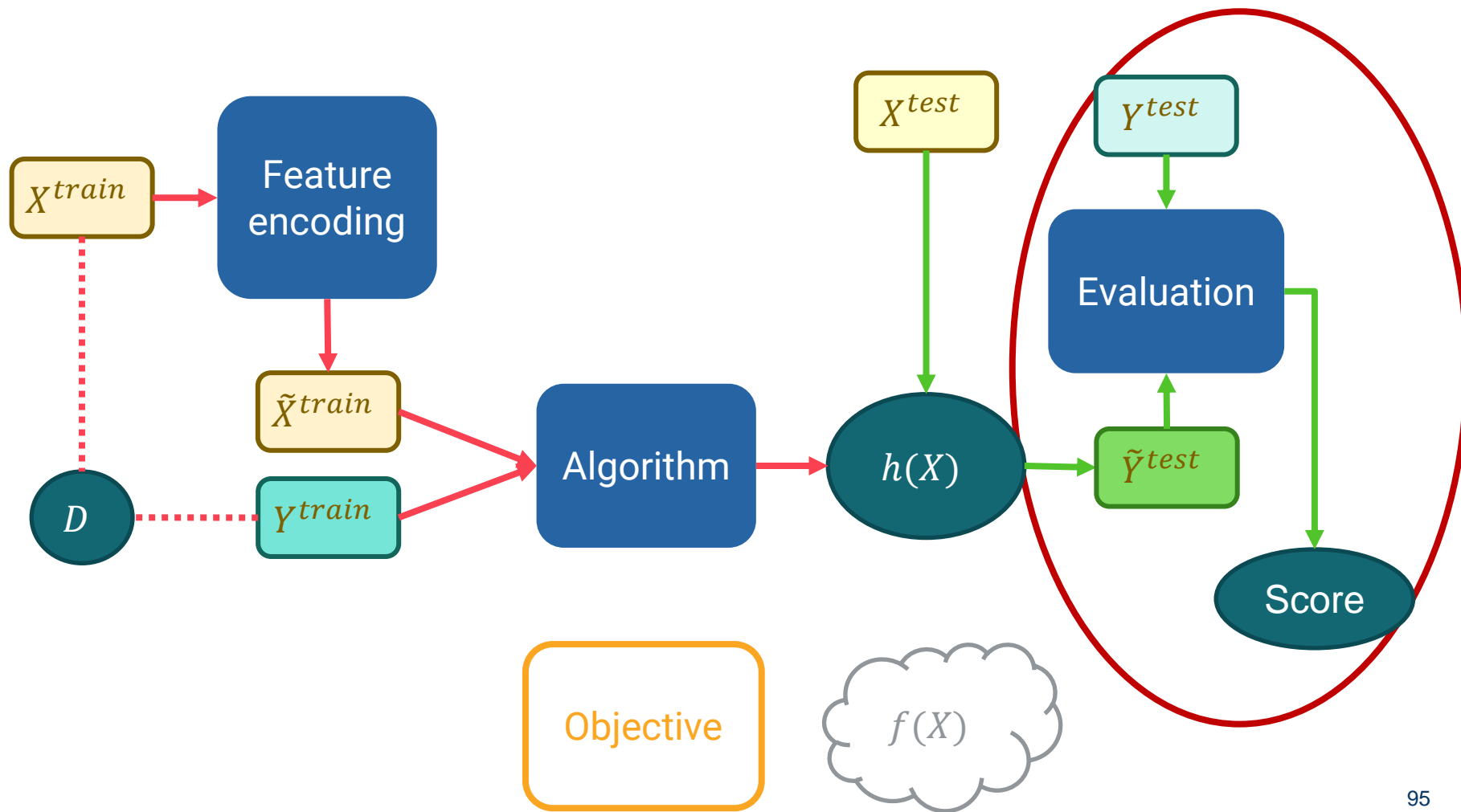
Overfit



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



# Evaluation



# How well does the model perform?

## Evaluation metric/measure



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

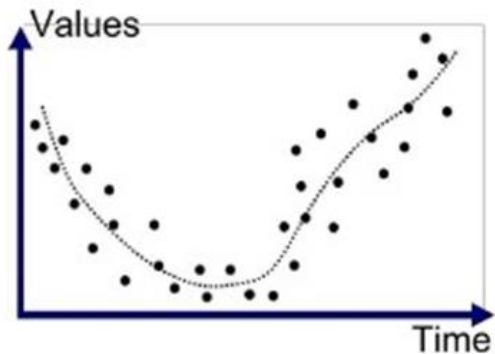
# How well does the model perform?

## Evaluation metric/measure

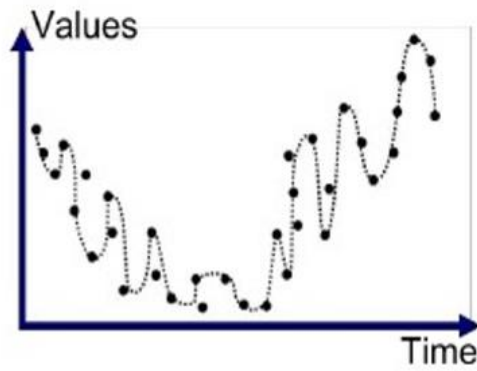


Regressor

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

