

Slide 1



Data Design. Lesson 8 – Modelling and evaluation 2: Machine Learning, by Iain Rice.

**Contents**

---

**SECTION 1** Introduction

**SECTION 2** Regression and Model Trust

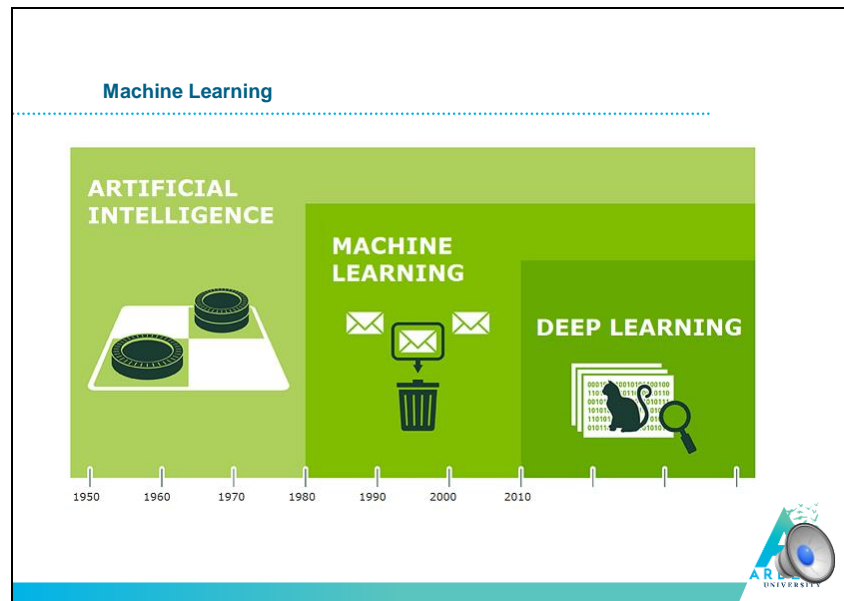
**SECTION 3** Wider Machine Learning



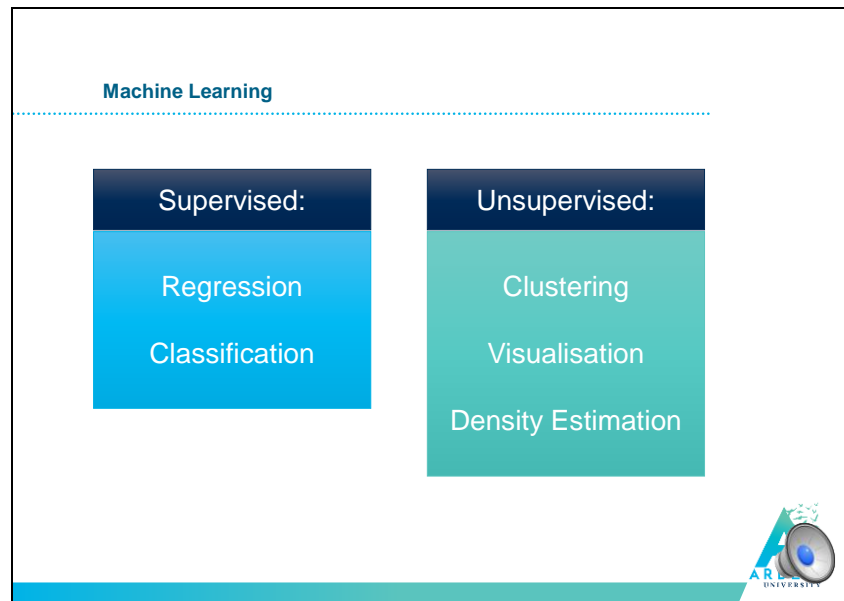
Slide 3



Section 1: Lesson introduction.



Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data and information. In machine learning, computers don't have to be explicitly programmed but can change and improve their algorithms by themselves. The field attracted a lot of attention from 1980 onwards but the origins of machine learning date back to the 1940s. Recently a sub-set of machine learning, known as deep learning, has been at the forefront of technology reporting in the media.



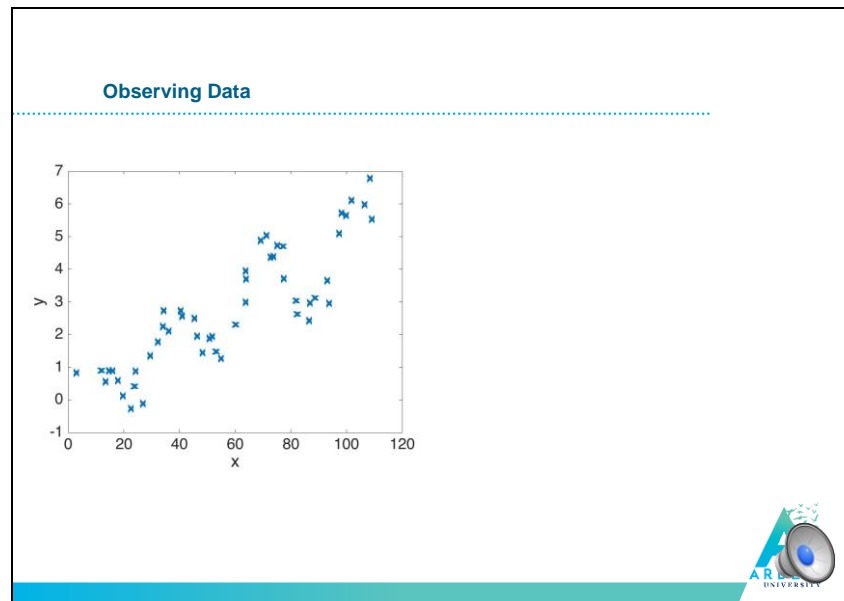
There are, broadly speaking, two types of machine learning: supervised methods such as regression and classification and unsupervised methods such as clustering, visualisation and density estimation. Supervised methods consider the task of modelling observations and making predictions about data based on existing variables, for instance trying to predict future stock prices or performing speech recognition. Unsupervised methods are designed to uncover structure and descriptors where the desired outputs are not known such as estimating a probability distribution or clustering a dataset into distinct groups.

Slide 6

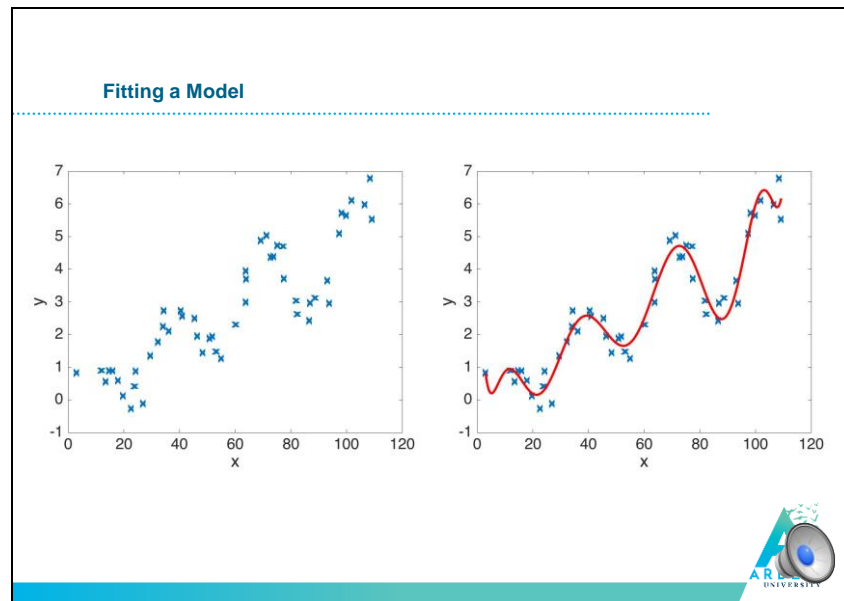


Section 2: Regression and model trust.

## Slide 7

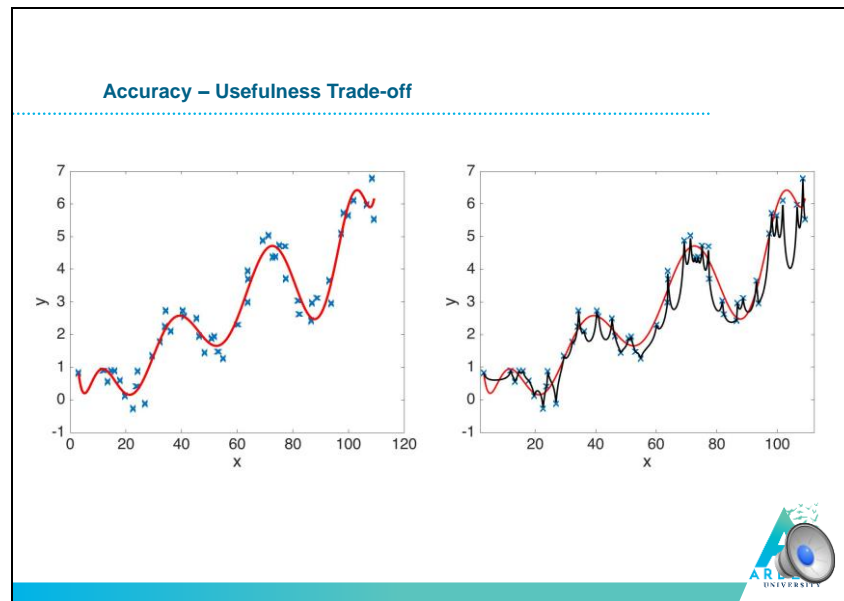


The most basic regression scenarios concern the modelling of observed data outputs,  $y$ , from inputs,  $x$ . Here there are 50 input-output datapoints and the task is to generate a model which given an input value,  $x$ , can predict what the observed output value,  $y$ , would be.

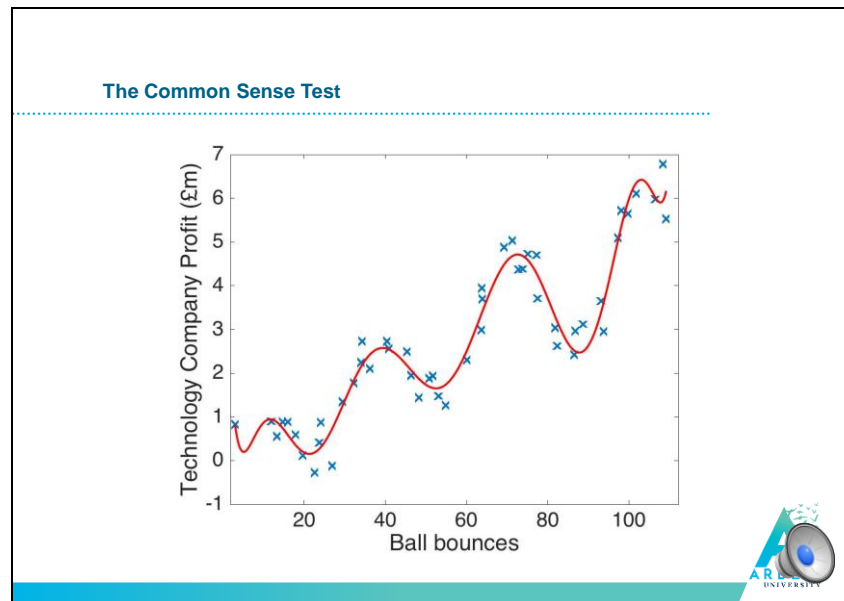


The figure on the right shows such a nonlinear model – the red curve – which predicts with sufficient accuracy what the output values,  $y$ , would be for each input,  $x$ . The statement ‘sufficiently accurate’ is ambiguous, for instance if the  $y$ -axis units are millions of pounds spent on projects with different labour costs,  $x$ , then a modelling inaccuracy of half a million pounds for  $x$  equal to 110 may have significant budgeting implications. This is however a toy example. It may seem like the ‘best’ red curve model should actually run through each of the observed  $y$ -values showing perfect prediction; however, in nearly all cases this will result in poor models.










On the right figure an overly complex non-linear model is used which perfectly predicts all of the observed  $y$ -values for each input; however, in the regions between these observations, the model behaves unexpectedly, seemingly without reason. The black curve may have a higher prediction accuracy based on this set of datapoints, but if a new 'test' dataset were to be predicted and the model validated against the true test  $y$ -values it would obviously perform poorly. There are tools to avoid this 'overfitting' phenomenon such as restricting the amount of data used to construct a model and ensuring a test set is available for validating which model is better.




It should also be noted that just because a model can accurately predict output values from inputs,  $x$ , that there is not necessarily a connection between the variables, even if they are highly correlated. For instance, if I were to bounce a ball and catch it each day, recording the number of times I can bounce it and catch it and plot that against the profit for a technology company, the values may appear connected. As time goes on my skill at catching the ball should improve and likewise the technology company should be increasing profits each day, but there is obviously no causal relation between these two variables. The easiest way to guard against issues such as this is to make sure modelling assumptions are not void of common sense.

**Modelling Capabilities**

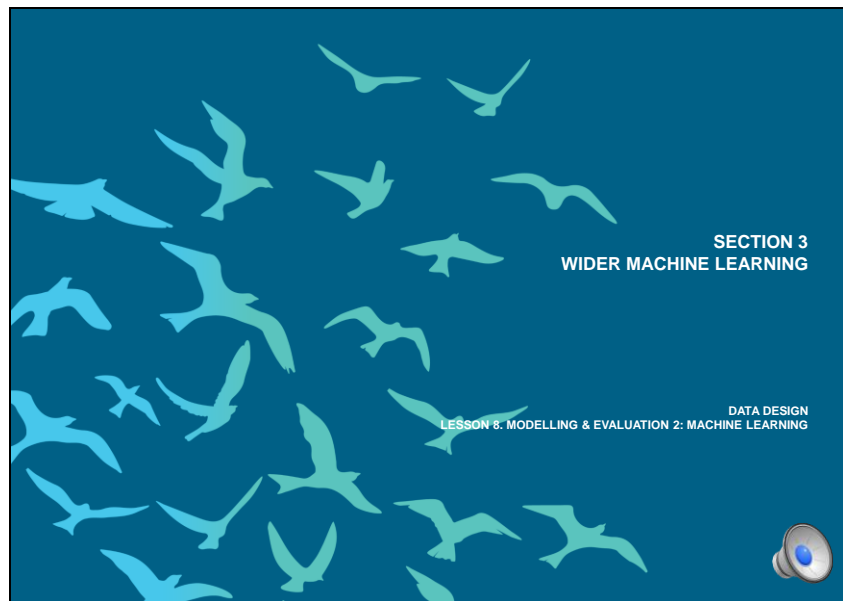
- Regression is for interpolation not extrapolation:  
Ball bounces: 1000  
Predicted profit (£m): 1.0313e+15
- Models have to give an output regardless of the input:

5	0	4	1	1
				

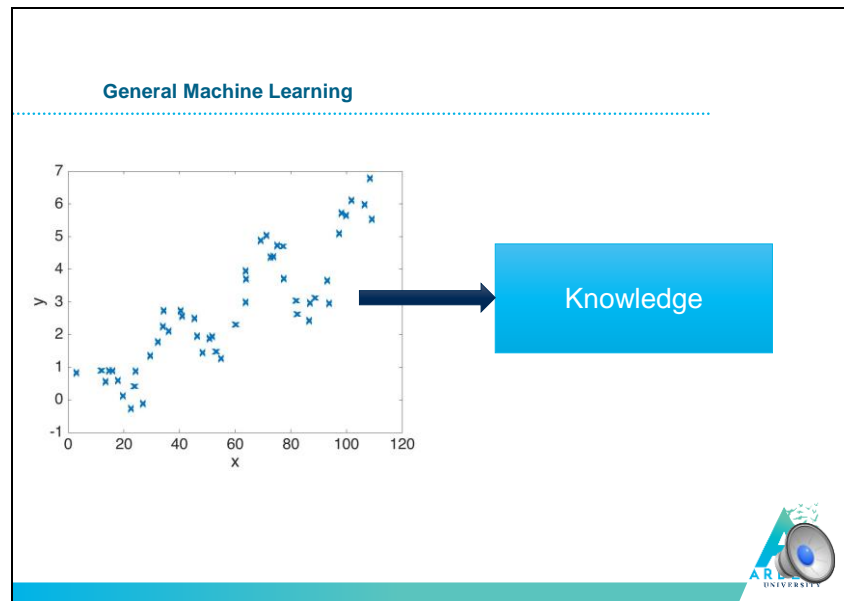


Complex machine learning models are capable of learning complex connections within data; however, these models cannot be blindly trusted. Regression models are interpolators between datapoints and should not be used for extrapolation. For instance, based on the red curve in the previous example, the predicted profit on a day where I bounce a ball 1,000 times is 1 times Euler's number (approximately 2.7) to the power of 15 million pounds, more money than there exists in total in the world. Furthermore, models have to generate an output for each input datapoint. Take for instance a classifier which is presented images of handwritten digits and has to classify which number they represent, as is used in the UK postal service. The model can correctly classify each of the four observed images, but will also classify the random image at the end as a 1 because it only knows how to classify handwritten digits and has to output a label based on this training process. Models are only fit for the purpose for which they are designed.

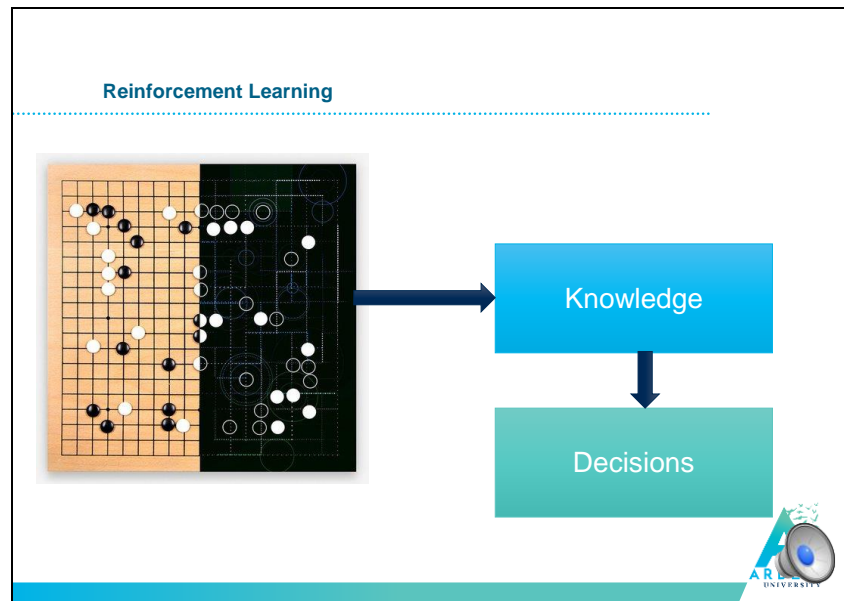
Slide 12



Section 3: Wider machine learning.

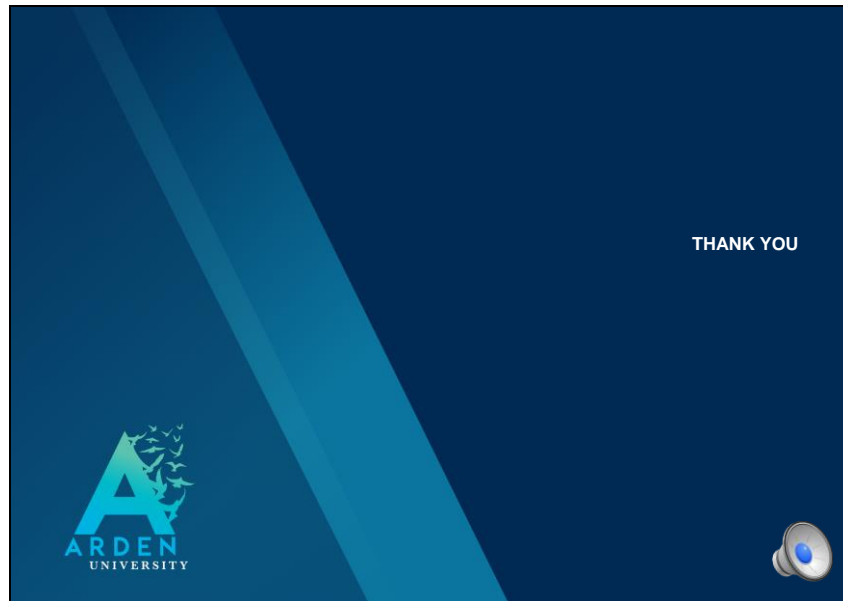


The standard machine learning scenario involves generating knowledge from data to perform tasks and inform decision-making processes. However, machine learning is not limited to cases where a clear transformation between data and information exists.



In 2016, a deep learning machine created by DeepMind, a company owned by Google, learned how to play the ancient game of Go, a more challenging board game than traditional chess. The machine was only told the rules governing movement of pieces and given a penalty system to encourage it to win games. After a significant amount of training, the deep learning machine managed to beat the international grandmaster in a best-of-five challenge. The game itself is not a series of datapoints as in the regression scenario, but a complicated game with numerous strategies. This has shown that machine learning tools can be applied to a wide variety of problems which they are not typically associated with, such as auctions and business strategy development for instance.

Slide 15



Thank you for listening.