1 What makes ML?

Uses the available data to predict unknown events in the future

Predictive Analytics

Identifying patterns in data and grouping them. For example identifying useful and "spam" email

Pattern Recognition

Summarise the data in ways in which patterns may emerge

Descriptive Analytics

The ability of a system to classify a sample based on samples it has been shown before

Computational Learning

ML

Build on predictive analytics – to not only predict future events but explain why they may occur and provide options (as well as their implications)

Prescriptive Analytics

2 When should I use it?

What is it good for?
When the answer to your question could be affected by a large number of factors, or require fine tuning it can be too complex to code simple rules. It can also be used when the size of the problem makes it impractical to use traditional methods.

What is it not good for?
If you can reliably answer your question with simple coded rules or steps you should not use ML –

5 Know your data

- Using a numeric score on qualitative data does not make it quantitative!! (E.g. rate on a scale of 1 to 5 how much you like this poster)
- Independence is important, the previous coin toss has no bearing on the result of the next.
- Correlation is not causation – just because the lines on the chart line up you need to be certain there isn't a confounding (unknown) variable skewing the results

3 How do I get my model?

ML Errors:
- **Type 1 are false positives**
- **Type 2 are false negatives**

The hardest part of the process – choosing which attributes of the data you will use in your algorithm and describing them in the model

Split your data for "training" your model and testing your model. Generally 70 / 30

Your model has been tested and you are confident in its accuracy

Data → Feature Engineering → Choose Algorithm → Split Data → Scored Model → Publish

Make your model "live"!

Data is either quantitative (measurable distance between two points) or qualitative (what is your favourite colour?)
- Categorical data (e.g. eye colour) the difference has little import but it can be simply grouped.
- Ordinal data - whilst the difference still does not matter (Monday vs. Thursday) there is an implied order (Thursday after Monday)
- Interval – differences between samples has an impact e.g., 2,4,8. each number is twice the preceding one
- Ratio – Similar to intervals however they have a known origin – e.g., Temperature cannot fall below Absolute Zero

Level of constraint: Low → High

Supervised – models are "trained" on known data.
Types of algorithm include:
- Classification - Assigning the sample to a category
- Regression - Predicts the value of an item
- Anomaly detection - Identify what is "normal" and highlight unusual items

Unsupervised – models have no labelled data here we use:
- Clustering - Finding commonality between items and grouping them together

More data trumps better algorithms!

4 Is your model useful?

Your model can be biased towards:
- ***Precision – out of the returned results how many were correct. If your application cannot afford to be wrong bias for precision***
- ***Recall – out of the available items how many were returned. If your application cannot afford to miss anything bias for recall***

In your workflow you could combine multiple models to build a quorum

How accurate is your model?

Accuracy is the error rate of your predictions e.g., 1 in 5 will be incorrect.

You are extremely unlikely to get 100% accuracy 80%+ is considered v. good

Beware just because ML gives you results unless you can articulate what they are & why you cannot rely on them – it won't tell you when you have taught it wrong! You must understand your data