COMP2200/COMP6200 Lecture 1b – What is Data Science?

Greg Baker

23rd February 2026

What is data science?

Applying the scientific method to commercial data

But what does that involve?

Data science is applied science: messy data, real decisions, and uncertainty

Data scientists often work with a lot of data

RECOVERY

`https://www.recoverytrial.net/`

- Trial with 50,000 patients on Covid treatments
- One of the largest medical trials in history

`https://www.nature.com/articles/nature11421`

- (2010) Ran a test with 61 million users to encourage voting
- Not even a particularly large data science project

Different scale of data: data scientists often need different skills and tools compared to typical statisticians

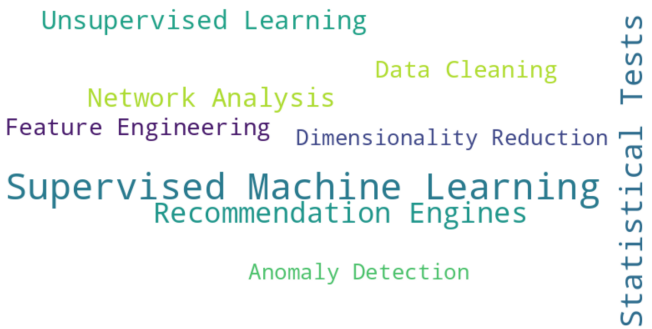
Four things data scientists do

1. Representation & modelling as compression
2. Generalisation discipline
3. Causality and intervention (when you're not just predicting)
4. Communication + governance

It's an end-to-end workflow, not just a bag of algorithms

1. Representation & modelling as compression

In data science, we create models of reality. We find ways of simplifying messy reality and turning it into something we can understand and talk about.



- Choose a representation of the world: bag-of-words, graphs, embeddings, neighbours
- Compress it into something you can reason with: trees, clustering, UMAP, linear models
- Trade off interpretability, accuracy, and cost

A model is a compression of reality that is useful for some purpose

2. Generalisation discipline

- Train/test splits and cross-validation (not "evaluate on the training set")
- Leakage: using information you won't have at deployment time
- Baselines, calibration, and class imbalance
- Drift: the world changes after you ship

Data science lives or dies by honest evaluation

3. Causality and intervention (when you're not just predicting)

- Prediction answers "what's likely next?"
- Causality answers "what happens if we do X?"
- In the workplace, this is where the money, ethics, and career-limiting mistakes live
- Best tool: experiments (A/B tests); otherwise be very careful

If you're going to act on a model, you need a causal story (or an experiment)

4. Communication + governance

- Explain what you built (model cards, data sheets, assumptions)
- Privacy, fairness, safety, and security are part of the job
- Monitoring, reproducibility, and "what would convince us we're wrong?"
- Good work is legible to other humans (including future you)

Data science is a social contract: document, monitor, and be ready to be wrong