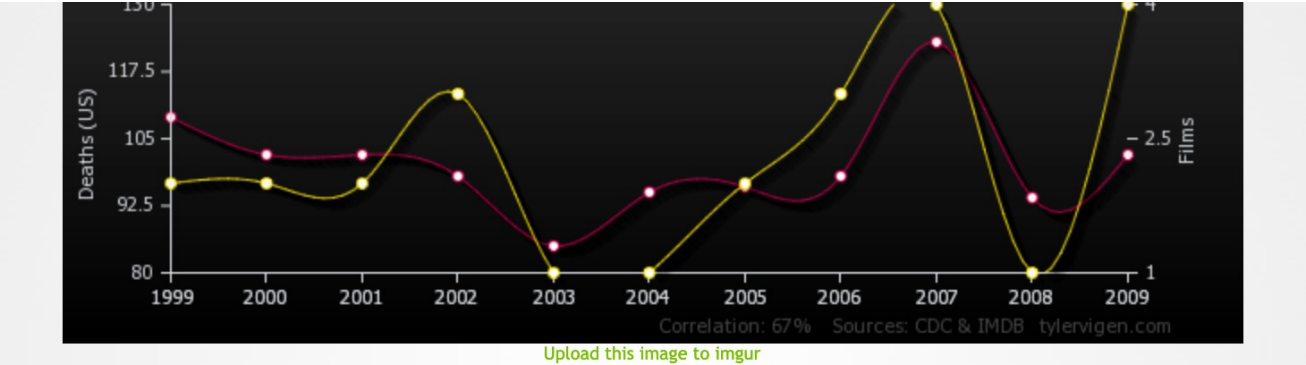
Imperial College London

Explainability

- The right to an explanation
- Classical Interpretability and Partial Dependence Plots
- An overview of XAI techniques
- **Are all explanations causal?**

Imperial College London

Spurious correlations



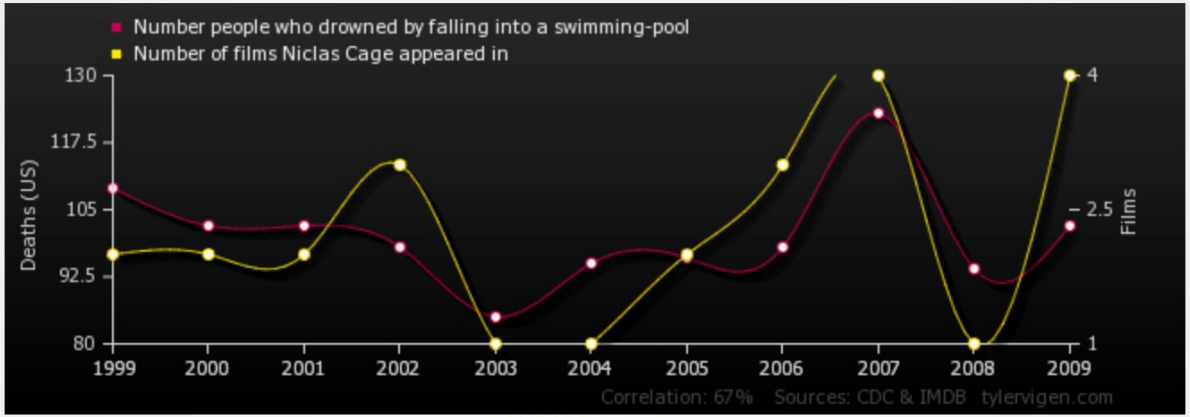
Imperial College London

Spurious correlations

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in



Upload this image to imgur

Imperial College London

Simpson's paradox

- A group of patients are given the option to try a new drug. 350 choose to, and 350 do not.
- Among those who took the drug, a smaller percentage recovered than in the placebo.
- However, among men, those that took the treatment did better. Among women, the same.
- So the drug helps men. It helps women. But it hurts the population as a whole.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Imperial College London

Simpson's paradox

- A group of patients are given the option to try a new drug. 350 choose to, and 350 do not.
- Among those who took the drug, a smaller percentage recovered than in the placebo.
- However, among men, those that took the treatment did better. Among women, the same.
- So the drug helps men. It helps women. But it hurts the population as a whole.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Imperial College London

Simpson's paradox

- A group of patients are given the option to try a new drug. 350 choose to, and 350 do not.
- Among those who took the drug, a smaller percentage recovered than in the placebo.
- However, among men, those that took the treatment did better. Among women, the same.
- So the drug helps men. It helps women. But it hurts the population as a whole.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- Note that women are less likely to recover than men, regardless of the drug.
- Note also that women are more likely to choose to take the drug.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Imperial College London

Simpson's paradox

- Assume now you record the blood pressure post treatment instead.
- For patients with high BP post-treatment, the drug makes things worse. For patients with low BP post-treatment, the drug makes things worse. But it helps the overall population (83 vs 78).
- Should we still recommend that the drug should be prescribed?

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- Assume moreover that we suspect the drug works by lowering blood pressure.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Imperial College London

Simpson's paradox

- Within each age stratum, exercise decreases cholesterol significantly.
- But it turns out that older patients exercise more. So the marginal correlation between exercise and cholesterol is positive: people that exercise more, as a general rule, have higher cholesterol (because they are older).
- This is a common problem, known as a confounding effect. It is also a common way in which insights can be biased.

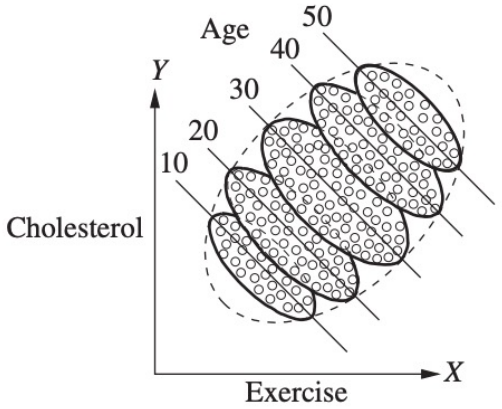


Figure 1.1 Results of the exercise–cholesterol study, segregated by age

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Imperial College London

Drivers and actionable insights

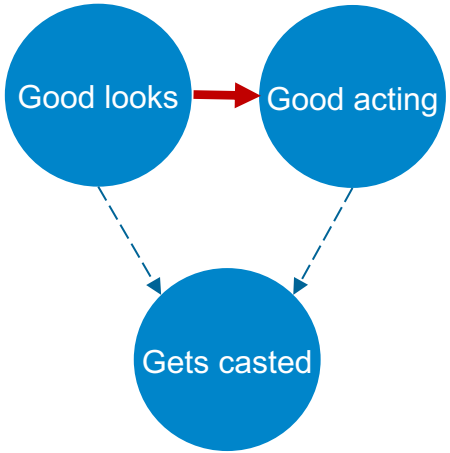
- In predictive modelling, this problem is not so important, as long as the model is applied on exactly the same process that generated the observational data, and not used to *intervene.*
- Much of data science is about *actionable insights,* or discovering *drivers* of performance. In both cases the implication is that you could intervene to change the value of a certain variable, and achieve the value the model predicts counterfactually.
- Unfortunately, counterfactual explanations are not real counterfactuals: they represent what happened in the data when the value was observed, not what would happen if you act.

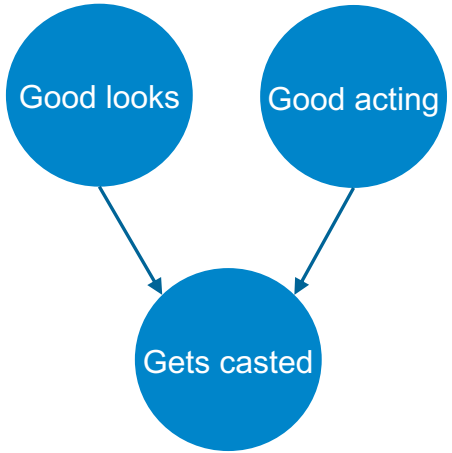
- Fairness by unawareness (hiding attributes) does not help when race can be proxied.
- Controlling for obvious confounders (like socioeconomic status when predicting probability of defaulting as a function of postcode which in turn proxies race) can help, but adding more variables is not always the right thing to do, as seen by earlier examples on Simpson's paradox.
- Counterfactual fairness is a growing field we will revisit later in the course.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Imperial College London

Which variables should I add in the model?





Berkson's paradox, or selection / collider bias

- Add all confounders. But remove all colliders. What about parents of confounders? Children of colliders?

Imperial College London

Summary

- Humans naturally interpret patterns as being causal. Moreover, data science pipelines are sometimes specifically tasked with producing actionable insights and identified drivers of performance.
- Extracting causal insights from purely observational data is impossible without additional causal assumptions. These can be hard to state properly without significant extra work. Graphs help.
- In general, many paradoxes are the result of causal confusion, such as Simpson's or Berkson's paradox.
- There are promising approaches to fairness and explainability that rely on causal explanations.