The Alan Turing Institute

#TMCFTURING

Intro to Causal Inference + Counterfactual Prediction

PETER WG TENNANT PHD

ROMIN PAJOUHESHNIA PHD

www.turing.ac.uk

turinginst

theturinginst

#TMCFTURING

Intro to Causal Inference

Peter WG Tennant

University of Leeds, UK

PWGTennant adhd.scientist

The Alan Turing Institute

LEEDS Institute for Data Analytics

UNIVERSITY OF LEEDS

WHAT IS DATA SCIENCE?

#TMCFTURING

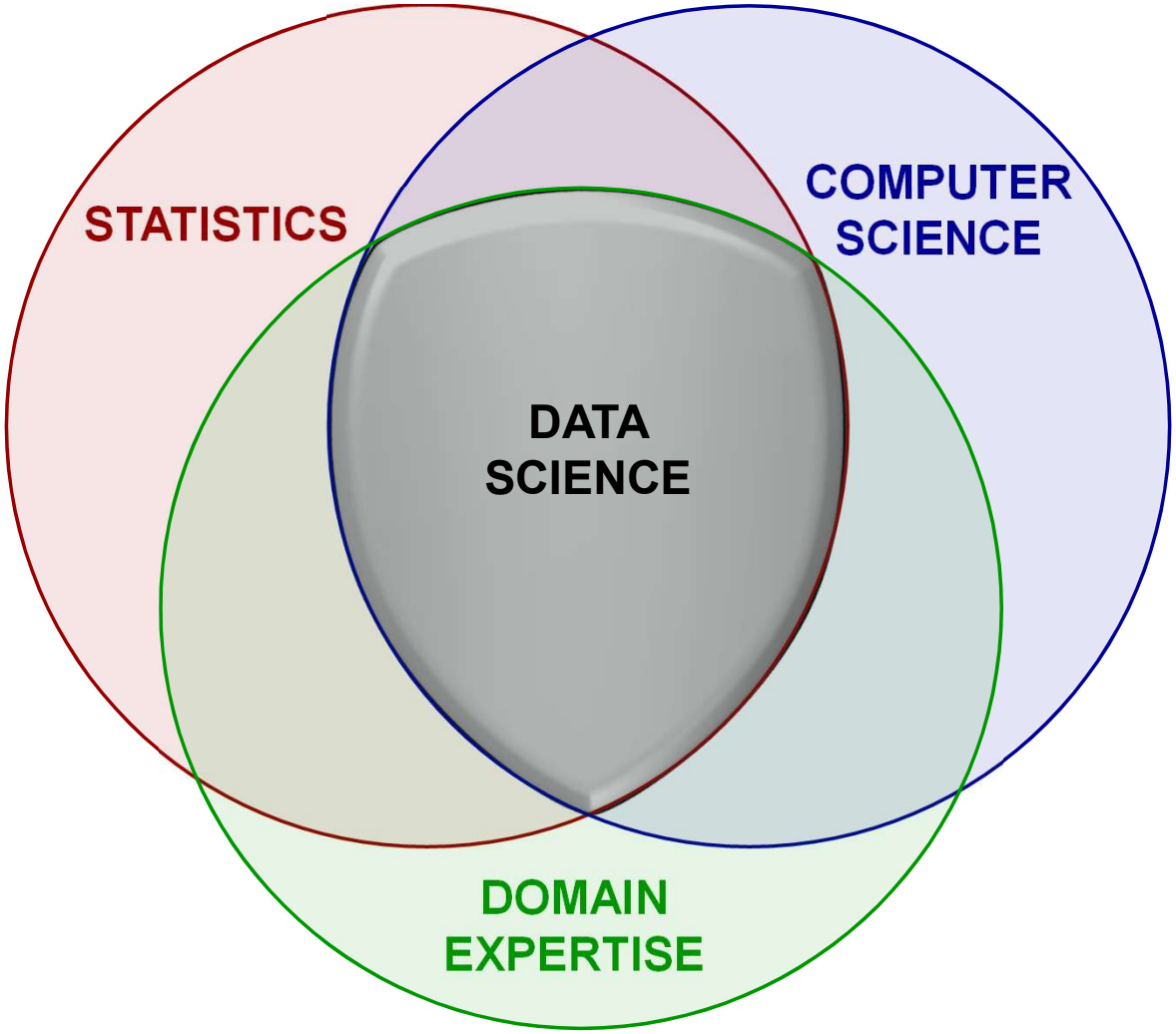
WHAT IS DATA SCIENCE?

The Alan Turing Institute

“I sit in a dark room and project code straight to my face while solving complicated problems. This helps me to immerse myself in it and "feel" the code. ”

WHAT IS DATA SCIENCE?

#TMCFTURING



The Alan Turing Institute

THE THREE TASKS OF DATA SCIENCE

#TMCFTURING

- Science is about building knowledge and understanding
- Data science is about '*gaining insights*' and '*extracting meaning*' from data
- Most data science activity can be divided into **three scientific tasks**, each with different methods **and philosophies**

Hernán et al 2019. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks, CHANCE, 32:1, 42-49, DOI: 10.1080/09332480.2019.1579578

The Alan Turing Institute

DESCRIPTION

#TMCFTURING

Description (& visualisation)

- Focussed on **summarising**, **describing**, &/or **visualising** features of interest
- Data driven - involves simple calculations & unsupervised learning

Questions

- What happened?
- Who was affected?
- What was occurrence of **Y** in people with **X**?

Example

- Occurrence and spread of COVID-19

The Alan Turing Institute

Prediction (AKA classification and regression)

- Focussed on **pattern recognition** and **forecasting**
- Data driven – involves statistical modelling and supervised learning

Questions

- What **will** happen?
- Who **will** be affected?
- Are people with **X** are more likely to have **Y**?

Examples

- Screening for COVID-19 with symptoms or CT scan
- Predicting prognosis or severity of infection

Causal inference (AKA counterfactual prediction)

- Focussed on understanding
- **NOT data driven** – involves fusion of external knowledge with statistical modelling and supervised learning

Questions

- What will happen **if**…?
- **Why** were they affected?
- If we **changed X**, how would it **change Y**?

Examples

- Effect of opening/closing schools on infection spread
- Risks / benefits of ventilation

PROGRAMMED FOR CAUSAL INFERENCE

#TMCFTURING

- We learn about cause-and-effect from the moment we're born
- From observing how things interact, and what happens when we *do* things

The Alan Turing Institute

PROGRAMMED FOR CAUSAL INFERENCE

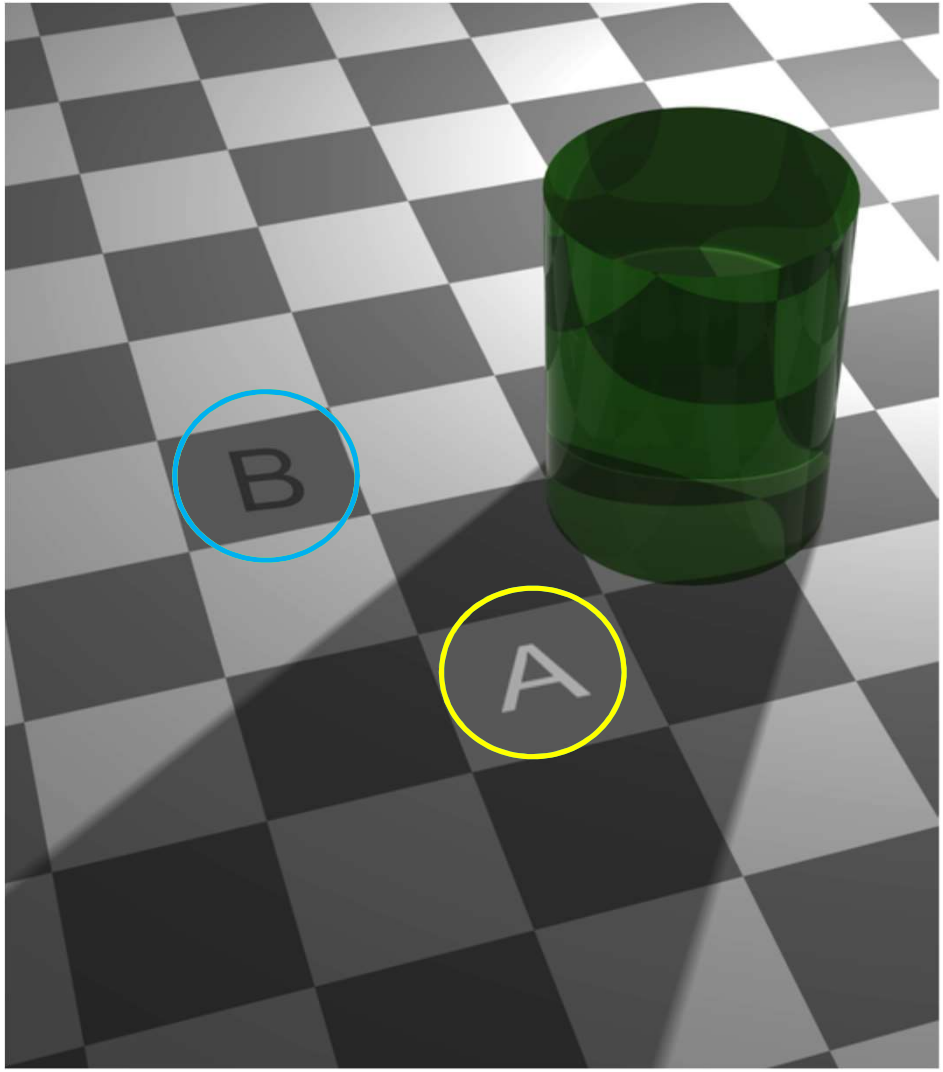
#TMCFTURING

- This extends beyond we can see & do, to whatever we can imagine!
- Our ability to ask '*what might have been*?' and '*how things could be different*?' has led to us changing our world beyond our ancestors imaginations

The Alan Turing Institute

EASILY FOOLED

#TMCFTURING



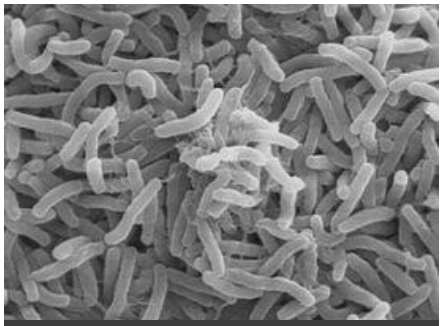
https://www.deviantart.com/butisit/art/Checker-shadow-illusion-263331875

https://www.newslinq.com/24-photos-need-really-look-understand-v1/

This is especially true once we're faced with questions and events that don't fit our childhood model of cause-and-effect…

HEALTHY

+



VIBRIO CHOLERAE

→

UNHEALTHY

✓ **Deterministic**

A MIND FOR DETERMINISM

#TMCFTURING

This is especially true once we're faced with questions and events that don't fit our childhood model of cause-and-effect…

HEALTHY

+

SMOKING

→

HEALTHY

× **Probabilistic**

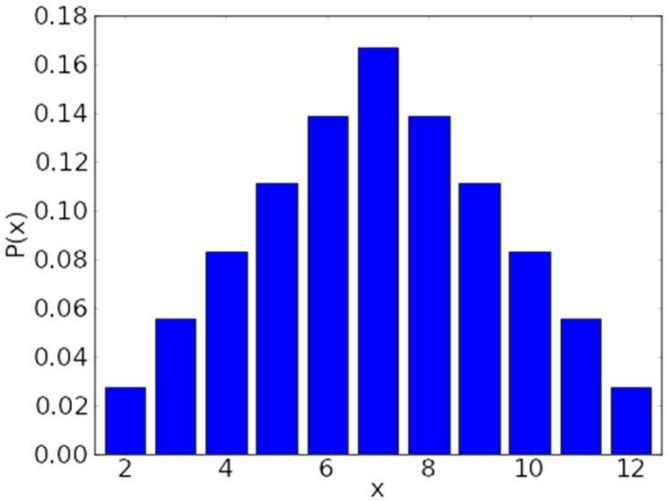
The Alan Turing Institute

PROBABILISTIC REASONING

#TMCFTURING

- Statistics & probability have resolved this problem
- What's **unknowable** for an individual… can be **predictable** for a group!

?

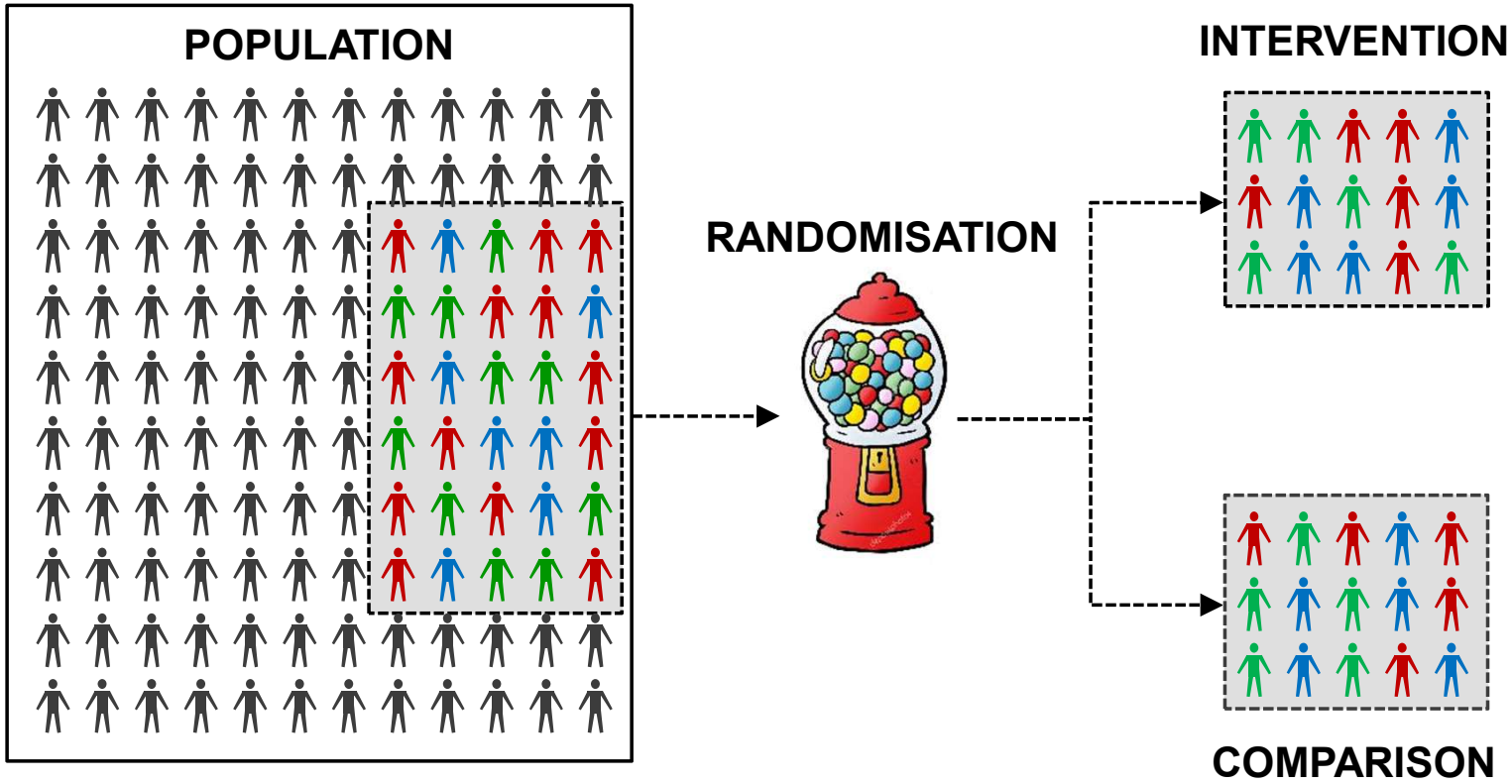


The Alan Turing Institute

RANDOMISATION

#TMCFTURING

With **randomisation,** this provides a potent way to estimate causal effects that we'd have little hope with experience and intuition alone

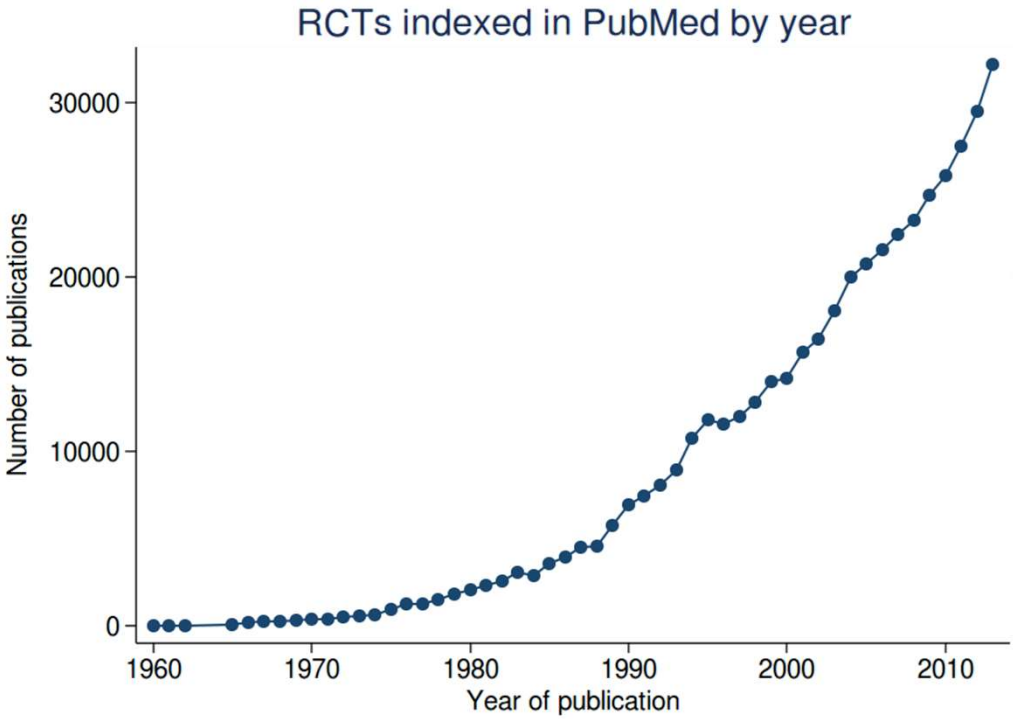


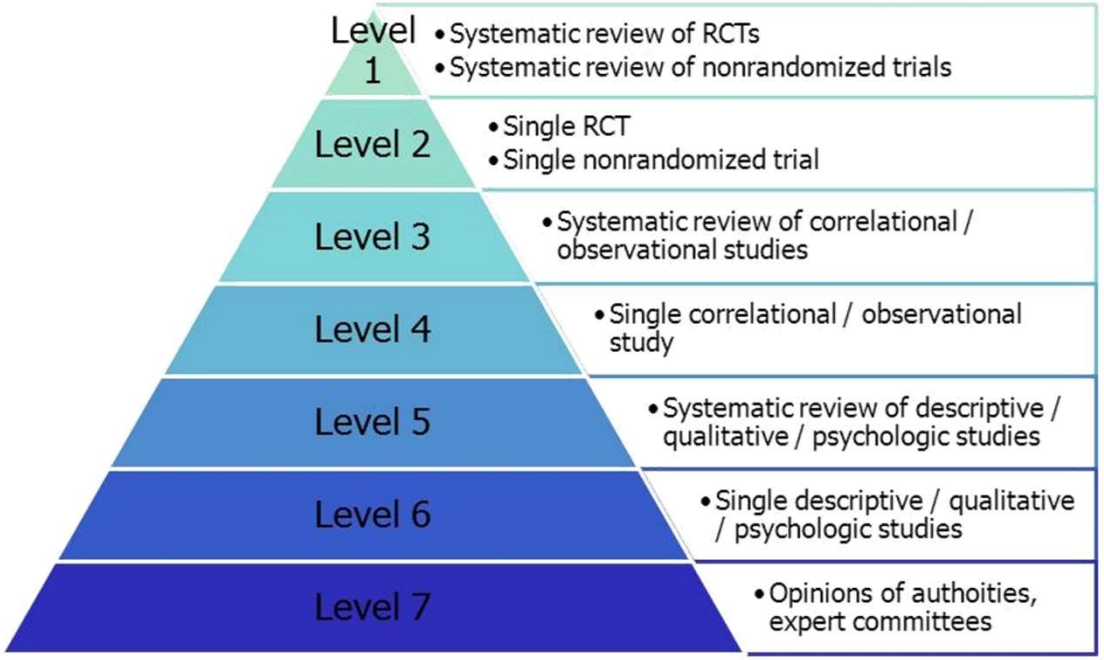
The Alan Turing Institute

RANDOMISATION

#TMCFTURING

In health and medical research, **randomised controlled trials** embraced with impassioned fervour, and canonised beyond all other forms of evidence





The Alan Turing Institute

IGNORANCE FROM NON-EXPERIMENTAL DATA #TMCFTURING

- Non-experimental studies have a poor record for causal inference
- **12 RCTs** studied **52 claims from non-experimental data**
  - **0** replicated
  - **5** found opposite!

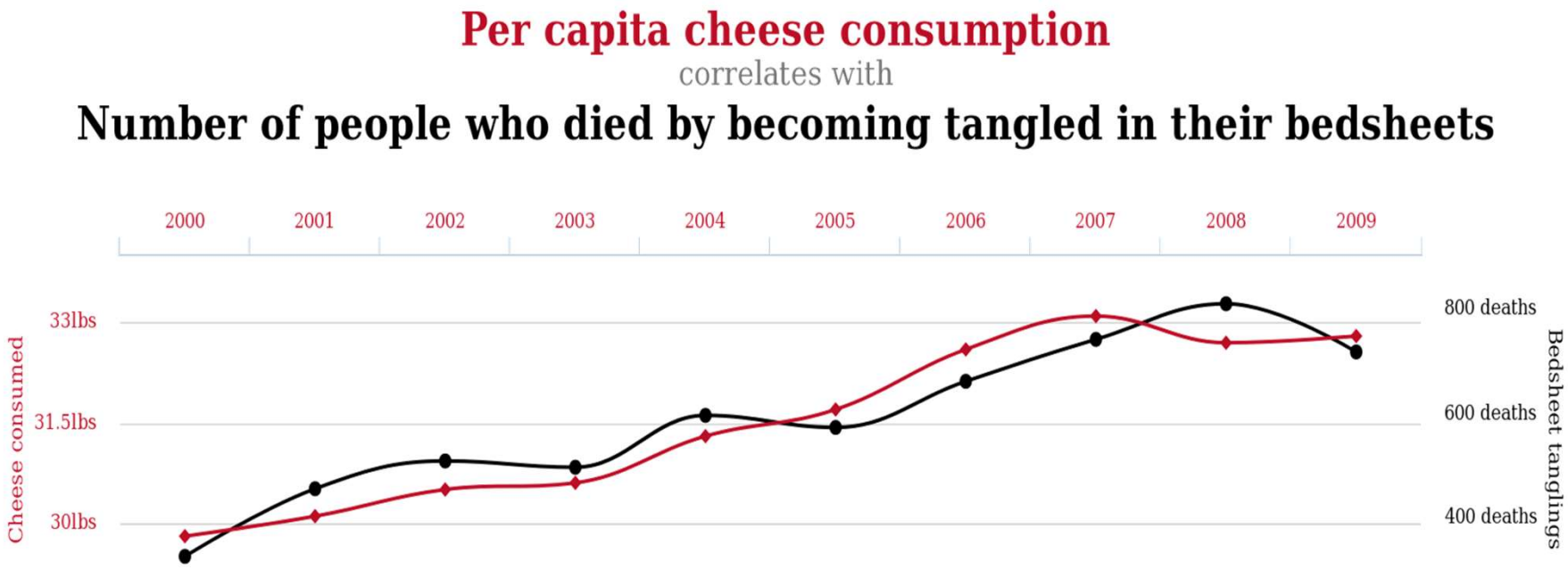
ID no.	*Pos.*	*Neg.*	*No. of claims*	*Treatment(s)*	*Reference*
1	0	1	3	Vit E, beta-carotene	*NEJM* 1994; **330**: 1029–1035
2	0	3	4	Hormone Replacement Ther.	*JAMA* 2003; **289**: 2651–2662, 2663–2672, 2673–2684
3	0	1	2	Vit E, beta-carotene	*JNCI* 2005; **97**: 481–488
4	0	0	3	Vit E	*JAMA* 2005; **293**: 1338–1347
5	0	0	3	Low Fat	*JAMA*. 2006; **295**: 655–666
6	0	0	3	Vit D, Calcium	*NEJM* 2006; **354**: 669–683
7	0	0	2	Folic acid, Vit B6, B12	*NEJM* 2006; **354**: 2764–2772
8	0	0	2	Low Fat	*JAMA* 2007; **298**: 289–298
9	0	0	12	Vit C, Vit E, beta-carotene	*Arch Intern Med* 2007; **167**: 1610–1618
10	0	0	12	Vit C, Vit E	*JAMA* 2008; **300**: 2123–2133
11	0	0	3	Vit E, Selenium	*JAMA* 2009; **301**: 39–51
12	0	0	3	HRT + Vitamins	*JAMA* 2002; **288**: 2431–2440
Totals	0	5	52		

Young & Karr 2011 Significance, 116:120, 2011

The Alan Turing Institute

DISTRUST NON-EXPERIMENTAL RESEARCH? #TMCFTURING

Teach: Correlation ≠ causation



https://www.tylervigen.com/spurious-correlations

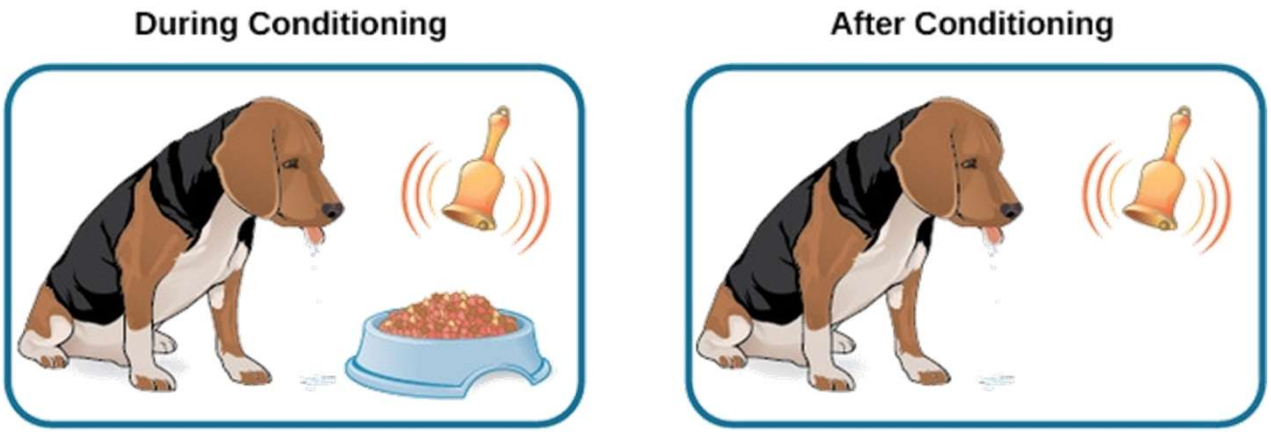
- **Discourage:** Causal inference from non-experimental data

The Alan Turing Institute

CAUSAL HABITS

#TMCFTURING

But: We are programmed to infer causality…



And: Spurious correlations have limited scientific and practical interest

So: People tend to infer causality regardless

The Alan Turing Institute

SCHRÖDINGER'S INFERENCE

#TMCFTURING

- To infer causality while claiming '*no causal inferences can be made*'

5. Discussion

Among 14-year olds living in the UK, we found an association between social media use and depressive symptoms and that this was stronger for girls than for boys. The magnitude of these associations reduced when potential explanatory factors were taken into account, suggesting that experiences of online harassment, poorer sleep quantity and quality, self-esteem and body image largely explain observed associations. There was no evidence of differences for girls and boys in hypothesised pathways between social media use and depressive symptoms. Findings are based largely on cross sectional data and thus causality cannot be inferred.

Our findings add weight to the growing evidence base on the potential pitfalls associated with lengthy time spent engaging on social media. These findings are highly relevant to current policy development on guidelines for the safe use of social media and calls on industry to more tightly regulate hours of social media use for young people [[10], [11]]. Clinical, educational and family settings are all potential points of contact whereby young people could be encouraged to reflect not only on their social media use but also other aspects of their lives including online experiences and their sleep patterns. For instance, in the home setting all family members could reflect on patterns of use and have in place limits for time online, curfews for use and the overnight removal of mobile

Kelly Y et al. Social media use and adolescent mental health: findings from the UK Millennium Cohort Study. EClinicalMedicine 2019.

The Alan Turing Institute

SCHRÖDINGER'S INFERENCE

#TMCFTURING

BMJ 2015;351:h4596 doi: 10.1136/bmj.h4596 (Published 5 September 2015)

ANALYSIS

Increased mortality associated with weekend hospital admission: a case for expanded seven day services?

Nick Freemantle and colleagues discuss the findings of their updated analysis of weekend admissions and the implications for service design

Nick Freemantle *professor of clinical epidemiology and biostatistics*[1 2], Daniel Ray *professor of health informatics*[2 3 4], David McNulty *medical statistician*[2 3], David Rosser *medical director*[5], Simon Bennett *director, clinical policy and professional standards*[6], Bruce E Keogh *national medical director*[6], Domenico Pagano *professor, cardiac surgery*[2 7]

[1]Department of Primary Care and Population Health, University College London, UK; [2]Quality and Outcomes Research Unit, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [3]Department of Informatics, University Hospitals Birmingham NHS Foundation Trust; [4]Farr Institute of Health Informatics Research, University College London; [5]University Hospitals Birmingham NHS Foundation Trust; [6]Medical Directorate, NHS England, London, UK; [7]Department of Cardiothoracic Surgery, University Hospitals Birmingham NHS Foundation Trust

Freemantle et al 2015
BMJ 5;351

The Alan Turing Institute

SCHRÖDINGER'S INFERENCE

#TMCFTURING

theguardian

home › politics election UK election world sport football opinion cult ≡ all

Health policy Michael White's political briefing

Quiet hospitals kill, but mindless union bashing can give us a nasty injury

Jeremy Hunt has every right to question NHS consultants' working practices but there are questions they should be allowed to ask him too

Hospital equipment is being underused. Photograph: Justin Paget/Justin Paget/Corbis

theguardian

home › politics election UK election world sport football opinion cult ≡ all

Health policy

Two deaths possibly linked to 'Hunt effect', study suggests

Research suggests some patients may be avoiding going to hospital at the weekend because of health secretary's statements about inadequate NHS staffing levels

Almost a third of the 40 patients studied suffered an increased chance of dying as a result of waiting to seek help. Photograph: Chris Radburn/PA

STOP NON-EXPERIMENTAL RESEARCH?

#TMCFTURING

Physical (E.g. Repetitive strain)

Chemical (E.g. Asbestos)

Biological (E.g. Influenza)

Cultural (E.g. Childhood diet)

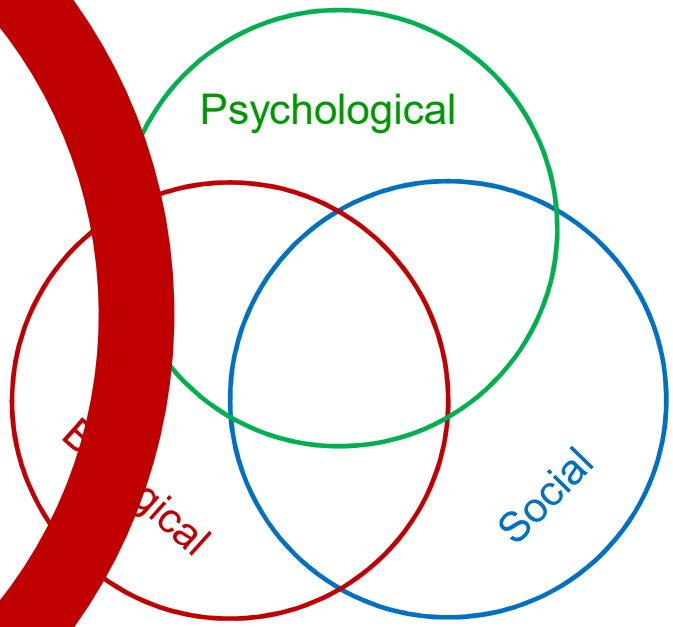
Economic (E.g. Poverty)

Environmental (E.g. Pollution)

Political (E.g. Austerity)

Behavioural (E.g. Smoking)

Personal (E.g. Self-esteem)



The Alan Turing Institute

CHANGE NON-EXPERIMENTAL RESEARCH! #TMCFTURING

- **Accept** & **admit** our causal ambitions

The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data

Causal inference is a core task of science. However, authors and editors often refrain from explicitly acknowledging the causal goal of research projects; they refer to causal effect estimates as a estimates.

This commentary using the term "cau sary to improve th observational rese

Miguel A. Hernán, MD, DrPH

See also Galea and Vaughan, p. 602; Begg and March, p. 620; Ahern, p. 621; Chiolero, Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

Data Are Not Enough—Hurray For Causality!

See also Galea and Vaughan, p. 602; Hernán, p. 616; Begg and March, p. 620; Ahern, p. 621; Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

Causal inference is of major importance in epidemiology and public health because the determination that an association between an exposure and a health outcome is causal indicates a potential for intervention to im-

had neither been formalized nor taught in epidemiology. Indeed, training epidemiologists, health scientists, and public health practitioners in causality has often been limited to the study of complex epistemological and

statistica properl plicitly tional epidemiology will help

analyses in comp because the con predicti plied to weak as ance for causal e findings public h

Start With the "C-Word," Follow the Roadmap for Causal Inference

See also Galea and Vaughan, p. 602; Hernán, p. 616; Begg and March, p. 620; Chiolero, p. 622; Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

In this issue of *AJPH*, Hernán (p. 616) argues that when the aim of an investigation is to estimate a causal effect, researchers should be allowed to say so. I agree, and extend this conversation to argue

In following the roadmap, the causal question starts the investigation, consistent with Hernán's call. In the subsequent step of identifiability, the specific ways in which a particular appli-

maximum likelihood estimation (TMLE; bit.ly/2FVfdfj).[4,5] As these newer statistical methods have become more widely used, traditional analysis approaches such as regression have begun to seem that they must fall short when it comes to causality.

This is one common conflation

The C-Word: The More We Discuss It, the Less Dirty It Sounds

See also Galea and Vaughan, p. 602; Hernán, p. 616; Begg and March, p. 620; Ahern, p. 621; Chiolero, p. 622; Glymour and Hamad, p. 623; and Jones and Schooling, p. 624.

I thank Chiolero (p. 622), Ahern (p. 621), Glymour and Hamad (p. 623), Jones and Schooling (p. 624), and Begg and

in social epidemiology when the goal is translating causal inferences into action.

One way of performing this

appropriate incorporation of expert knowledge in the process. Indeed, incorrect causal inferences from observational data are often the result of a flawed emulation of the basic design of the target trial (e.g., choice of time zero and classification of

Again, the process of specifying and emulating a target trial helps by providing a systematic way to explore each type of bias and its potential influence on the effect estimate. The Cochrane tool has adopted this target trial–based approach to assess the risk of bias of nonrandomized studies.[3]

American Journal of Public Health - May 2018

Alan Turing Institute

CHANGE NON-EXPERIMENTAL RESEARCH! #TMCFTURING

- Upgrade our **epistemological (philosophical) machinery**

"How can we infer causal relations from observations?"

$$p(\mathrm{X} = \mathrm{x}) = \prod_i p(X_i = x_i | \mathrm{PA_i} = \mathrm{pa_i})$$

Philosophy

Maths

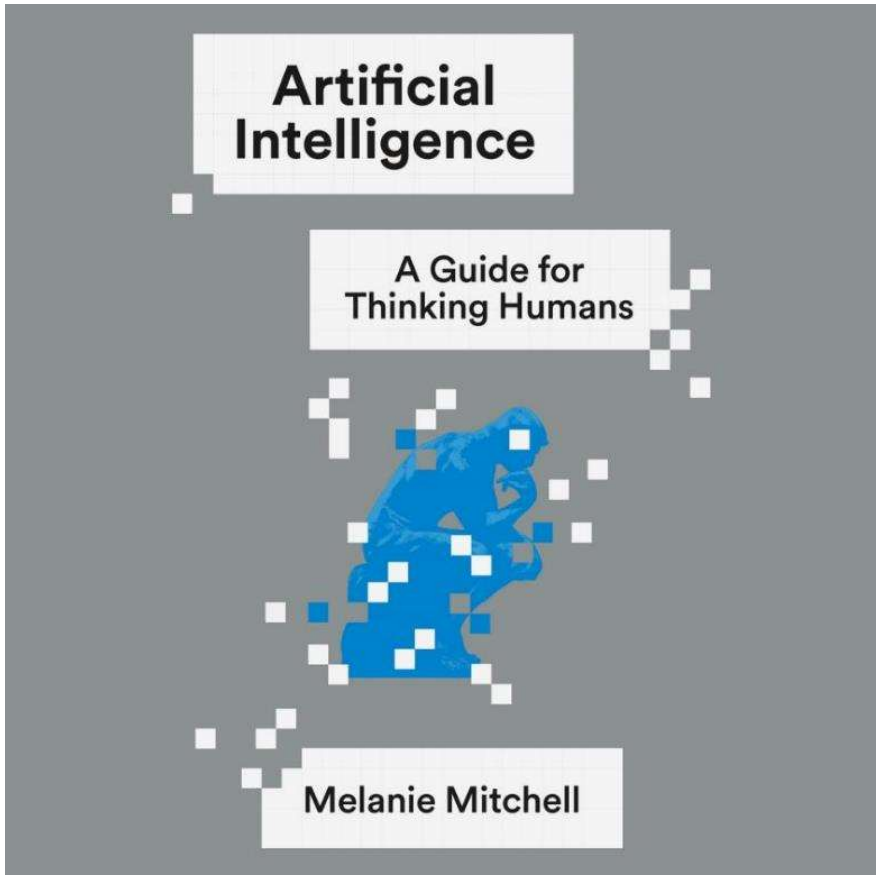
Data science

The Alan Turing Institute

WHAT THE MACHINE CANNOT LEARN

#TMCFTURING

- Machine learning is excellent at finding patterns in complex data > **prediction**
- Causal inference requires identifying and estimating **counterfactuals** (i.e. events that did **NOT** happen).
- In non-experimental data, machine learning can't do **causal inference** because you **CAN'T** learn what's not there!
- Need external knowledge of '**data generating process**'; the story behind how the data came into being

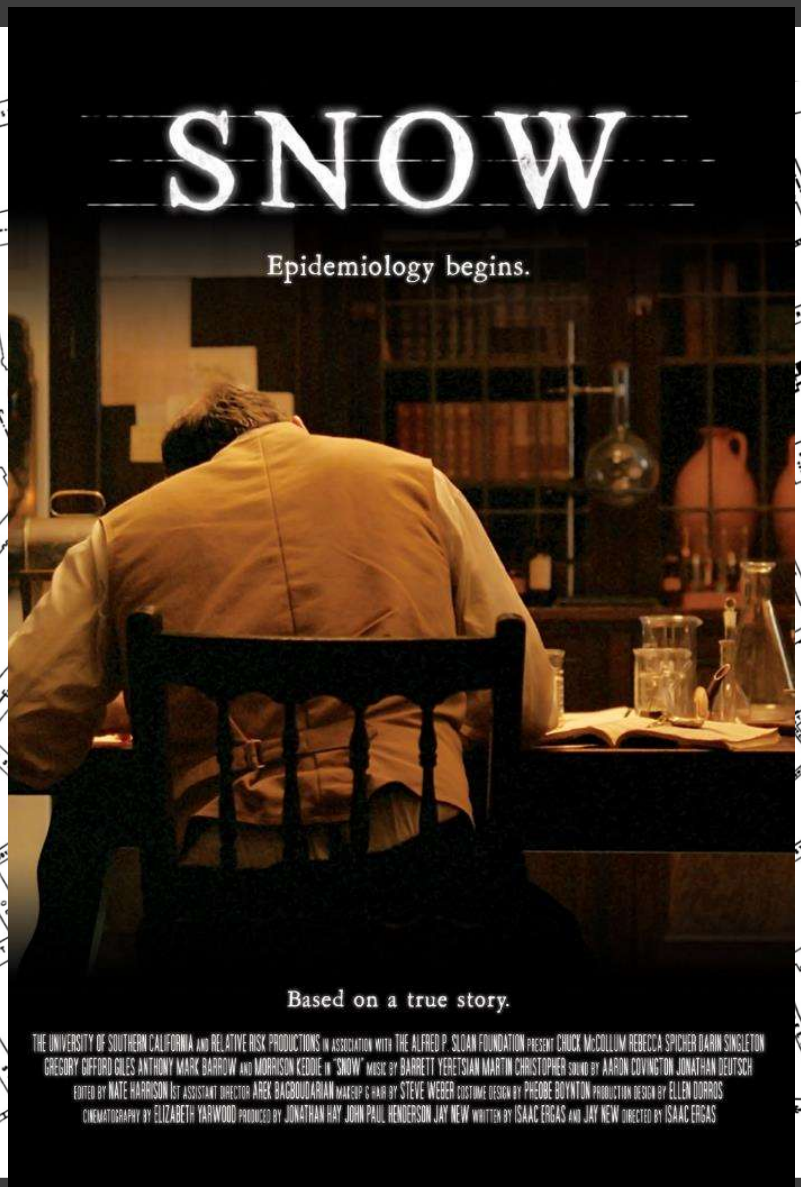


The Alan Turing Institute

LEGENDS AND MYTHS

#TMCFTURING

JOHN SNOW, LEGEND, DATA SCIENTIST



'BROAD STREET' PUMP

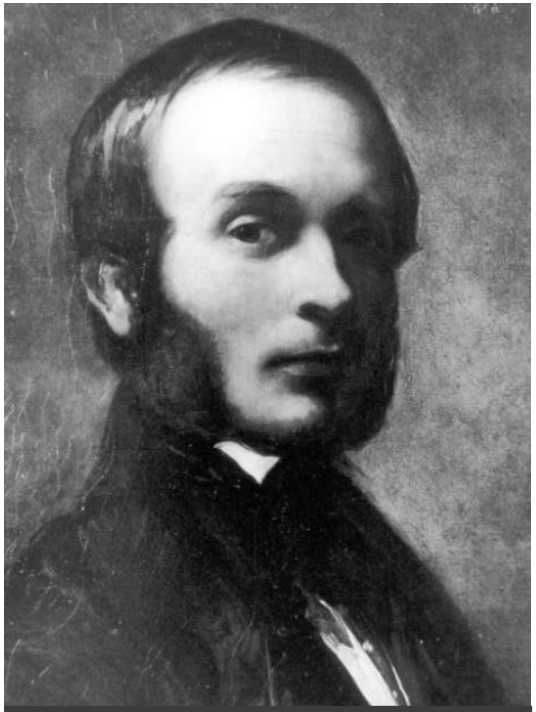
The Alan Turing Institute

1854

31 Aug	Cholera ou in Soho
5 Sep	Snow m s where deaths e occurring
6 Sep	83% d drank from Broad eet pump
7 Sep	Snow n ets p is guardia o argue for pump clo e
8 Sep	Snow remove ndle > Cholera outbreak

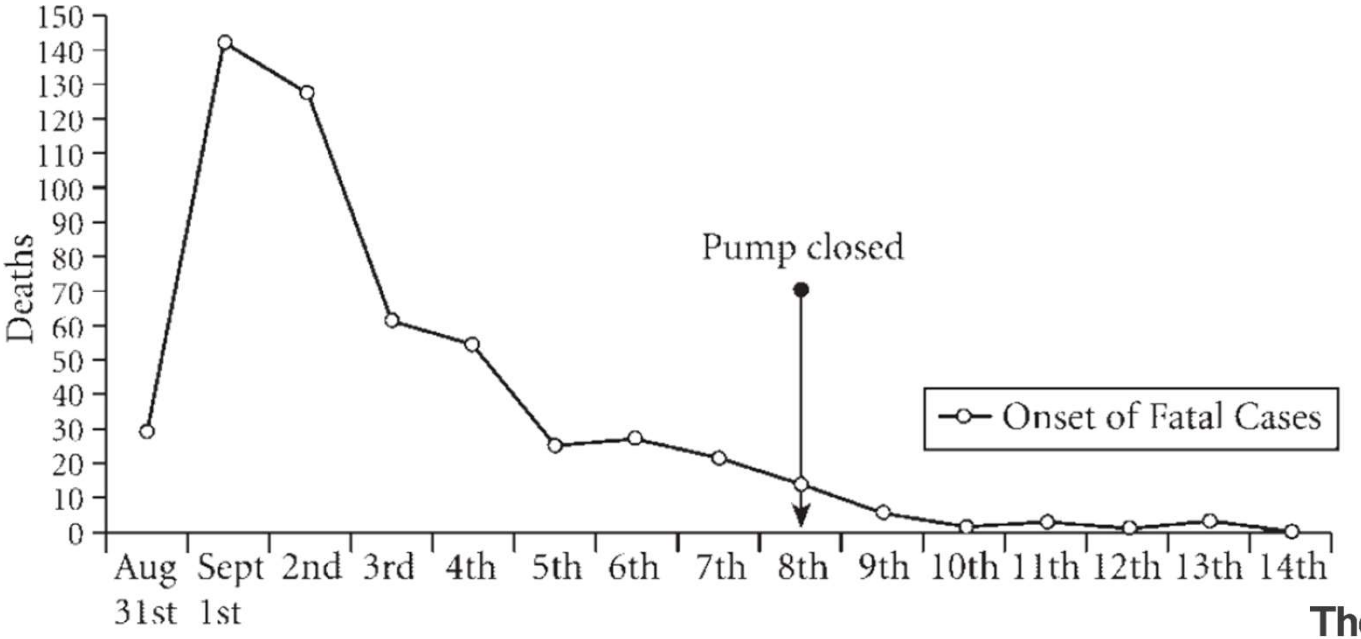
LEGENDS AND MYTHS

#TMCFTURING



JOHN SNOW, LEGEND, DATA SCIENTIST

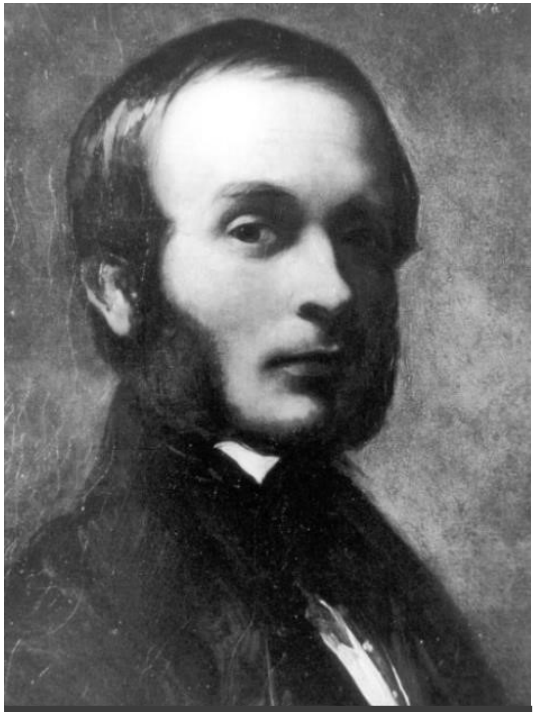
Snow observed the **outcome** of what happened when the pump was ‘closed’ (**factual**)



The Alan Turing Institute

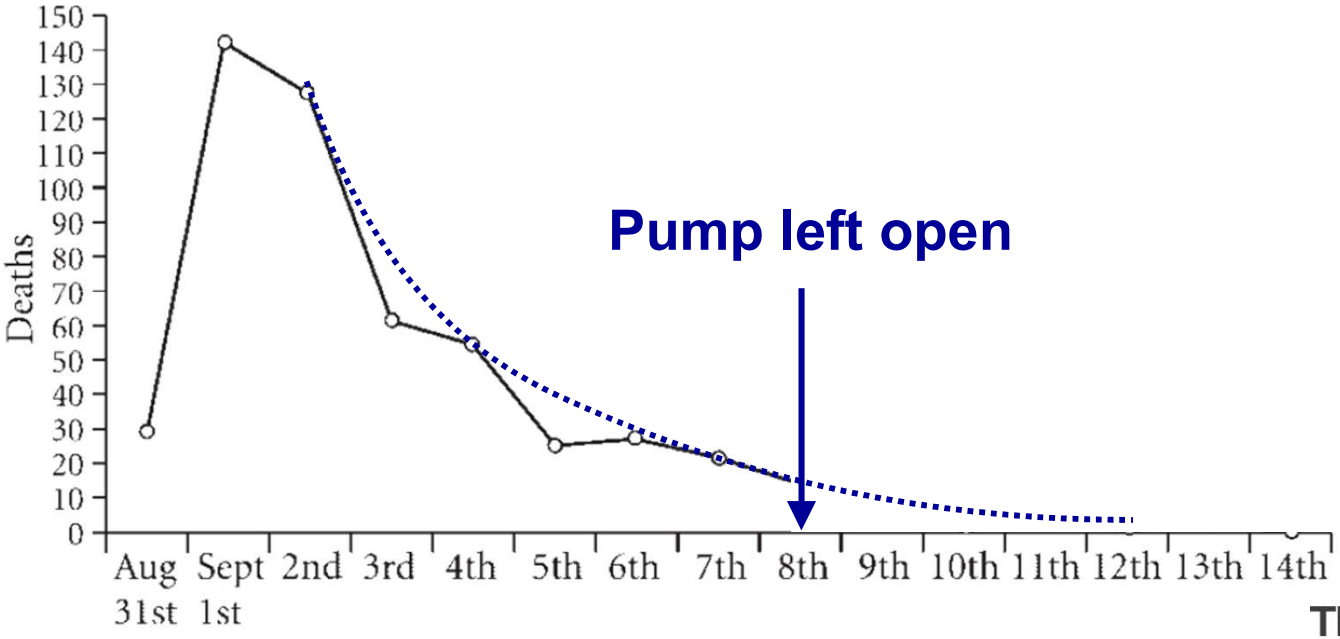
FUNDAMENTAL PROBLEMS

#TMCFTURING



JOHN SNOW, LEGEND, DATA SCIENTIST

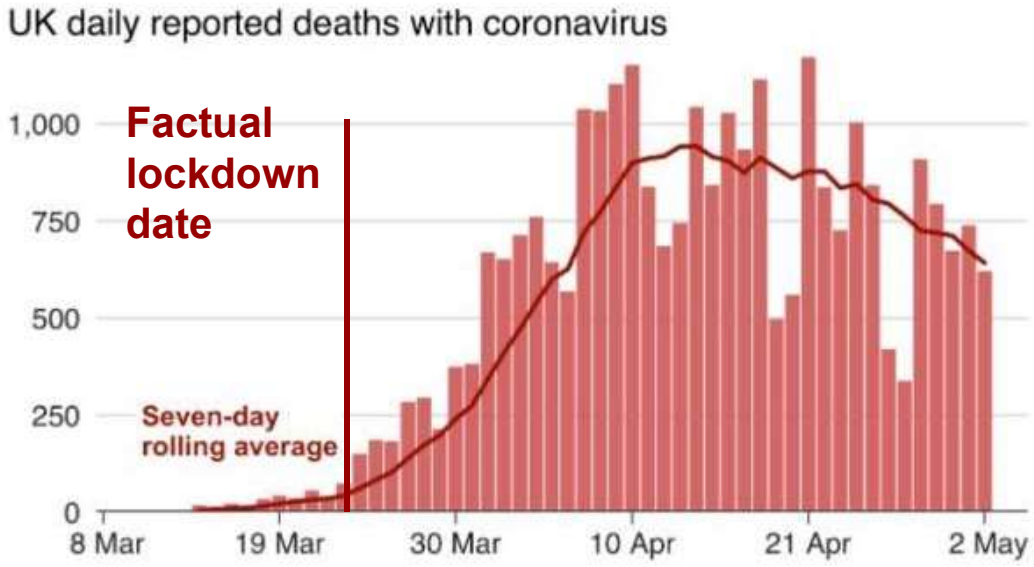
Didn't observe the **potential outcome** that *would have happened* if the pump had been – counter to fact - left 'open' (**counterfactual**)

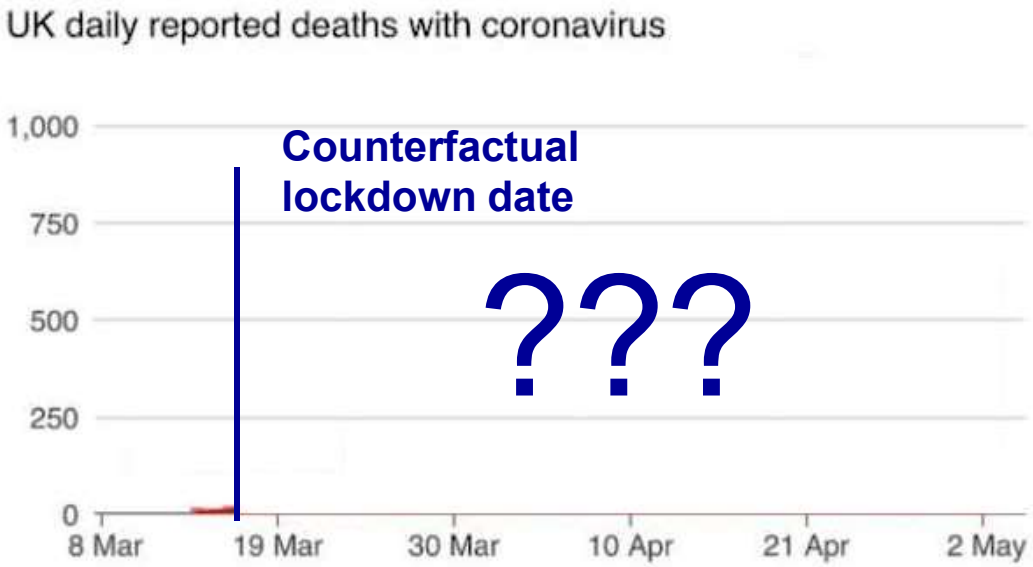


The Alan Turing Institute

FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE #TMCFTURING

- We can **never** know the **potential outcome** for a **counterfactual exposure**!
- For each '**unit of analysis**' we can only observe one potential outcome
- This is known as the **fundamental problem of causal inference**





The Alan Turing Institute

ESTIMATING POTENTIAL OUTCOMES

#TMCFTURING

Instead we must *estimate* the **potential outcome** for the **counterfactual exposure** from **exchangeable units of analysis**

The Alan Turing Institute

ESTIMATING POTENTIAL OUTCOMES

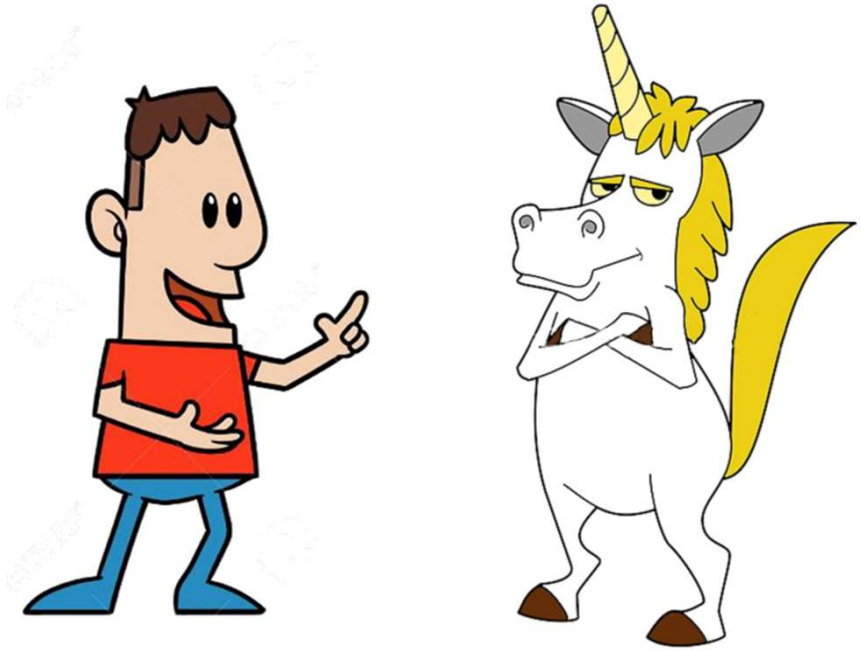
#TMCFTURING

GIVE MAGIC BEANS

The Alan Turing Institute

ESTIMATING POTENTIAL OUTCOMES

#TMCFTURING



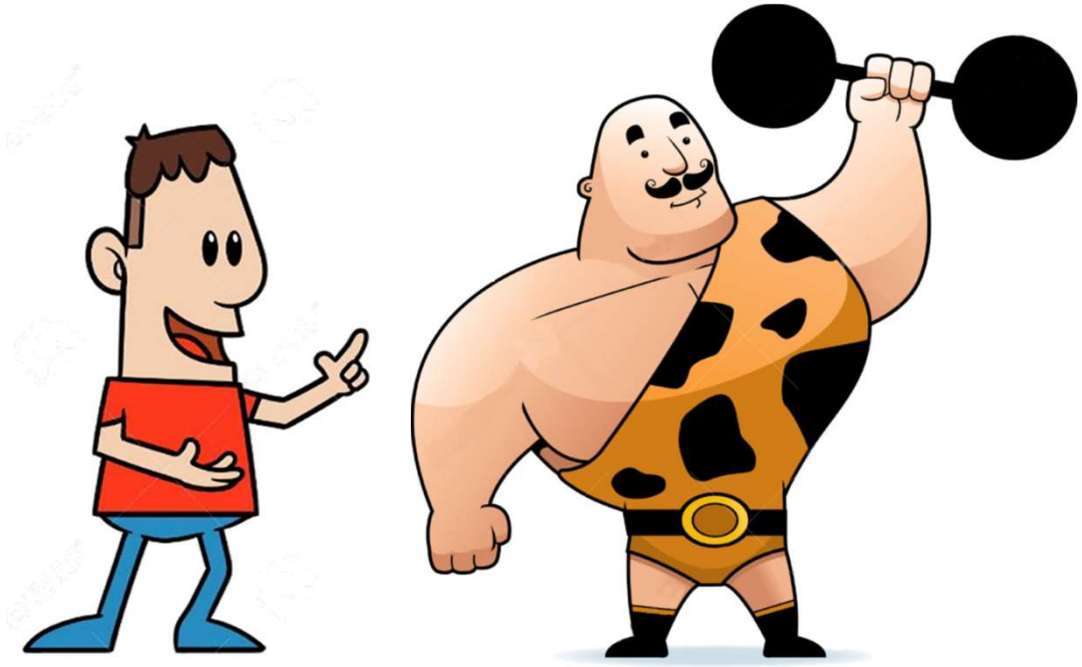
The Alan Turing Institute

POOR COUNTERFACTUALS

#TMCFTURING

Problem:

- Units (e.g. people) are **very different**
- Even the **same units** can respond **differently** at **different times**



Not **exchangeable** units of analysis!

The Alan Turing Institute

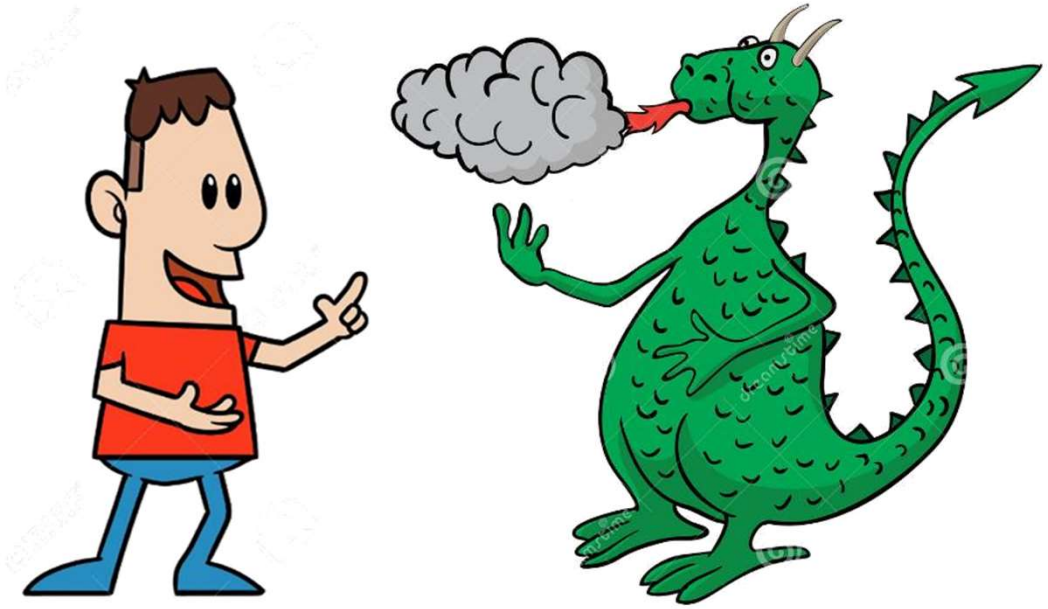
POOR COUNTERFACTUALS

#TMCFTURING

The Alan Turing Institute

POOR COUNTERFACTUALS

#TMCFTURING



Poor estimate of **potential outcome**!

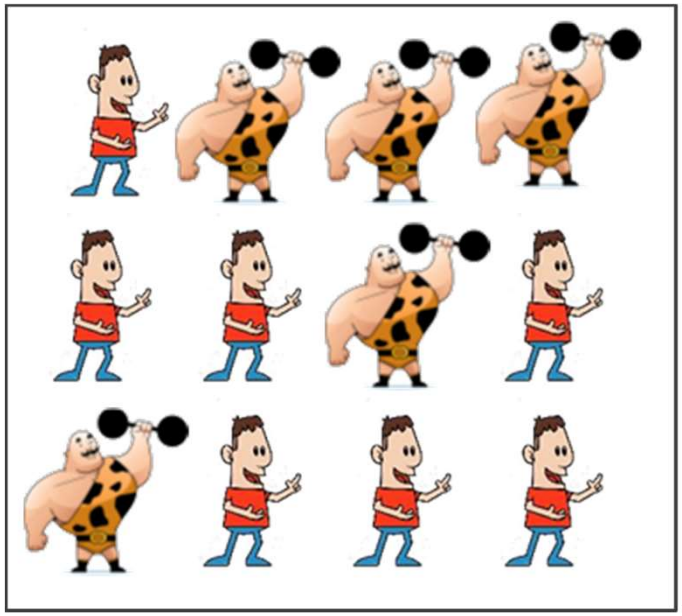
The Alan Turing Institute

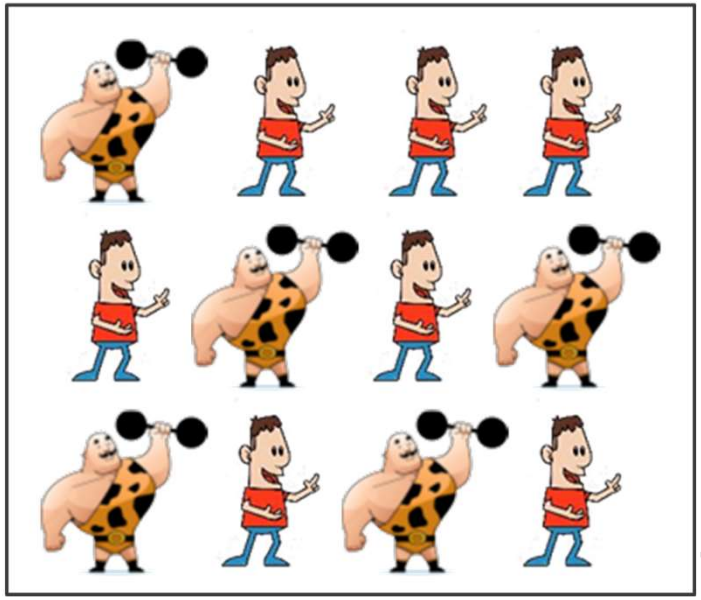
AVERAGE CAUSAL EFFECTS

#TMCFTURING

- We therefore have to **identify** *(sub)groups* of units that were **exchangeable** at the time of exposure

- We can then estimate the **average causal effect** by **comparing** the **outcomes** between these (sub)groups

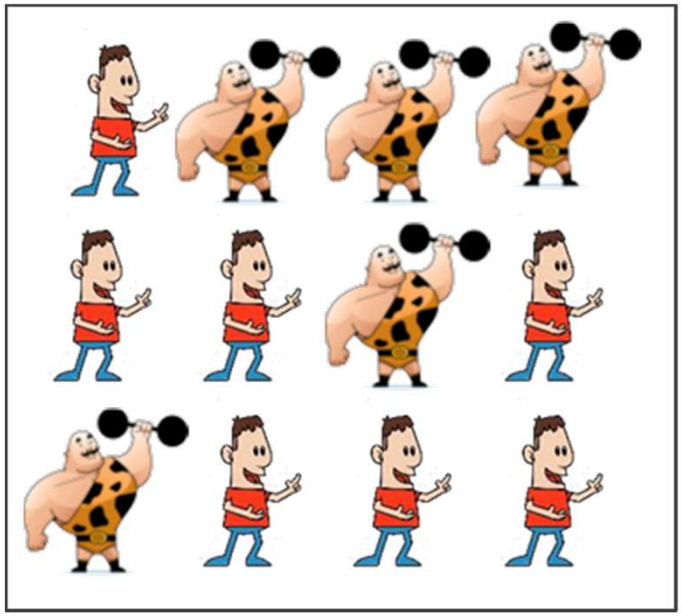




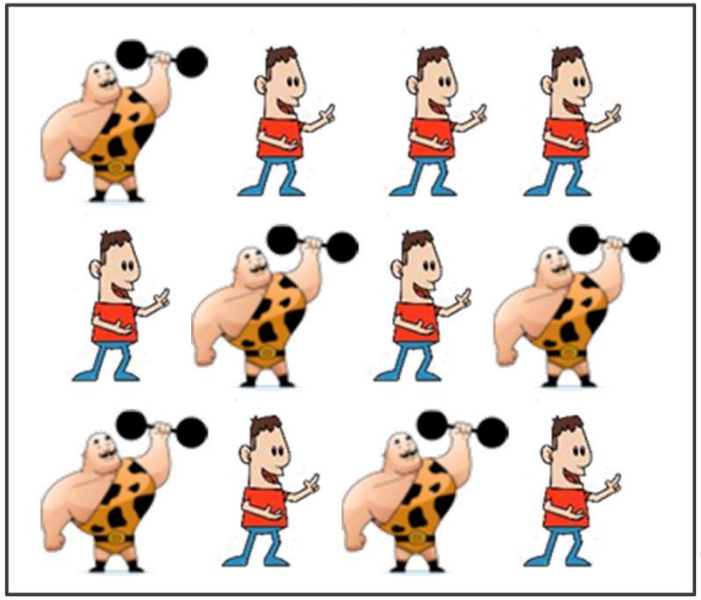
The Alan Turing Institute

AVERAGE CAUSAL EFFECTS

#TMCFTURING



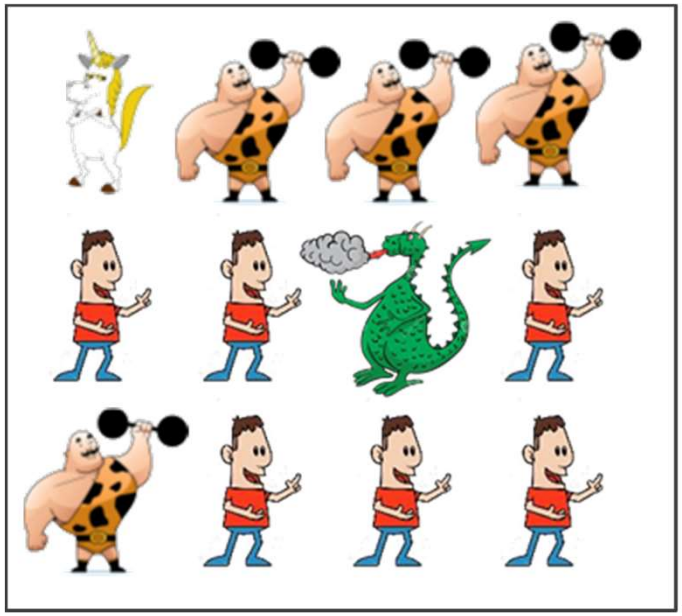
GIVE MAGIC BEANS

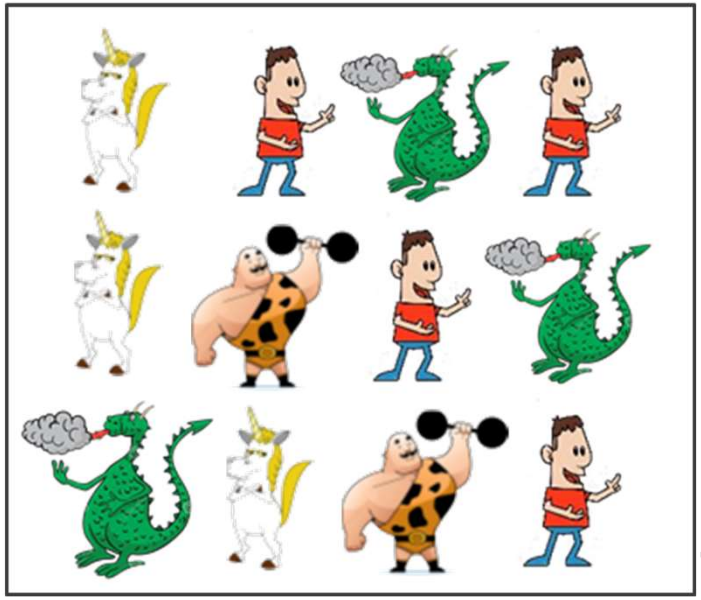


The Alan Turing Institute

AVERAGE CAUSAL EFFECTS

#TMCFTURING



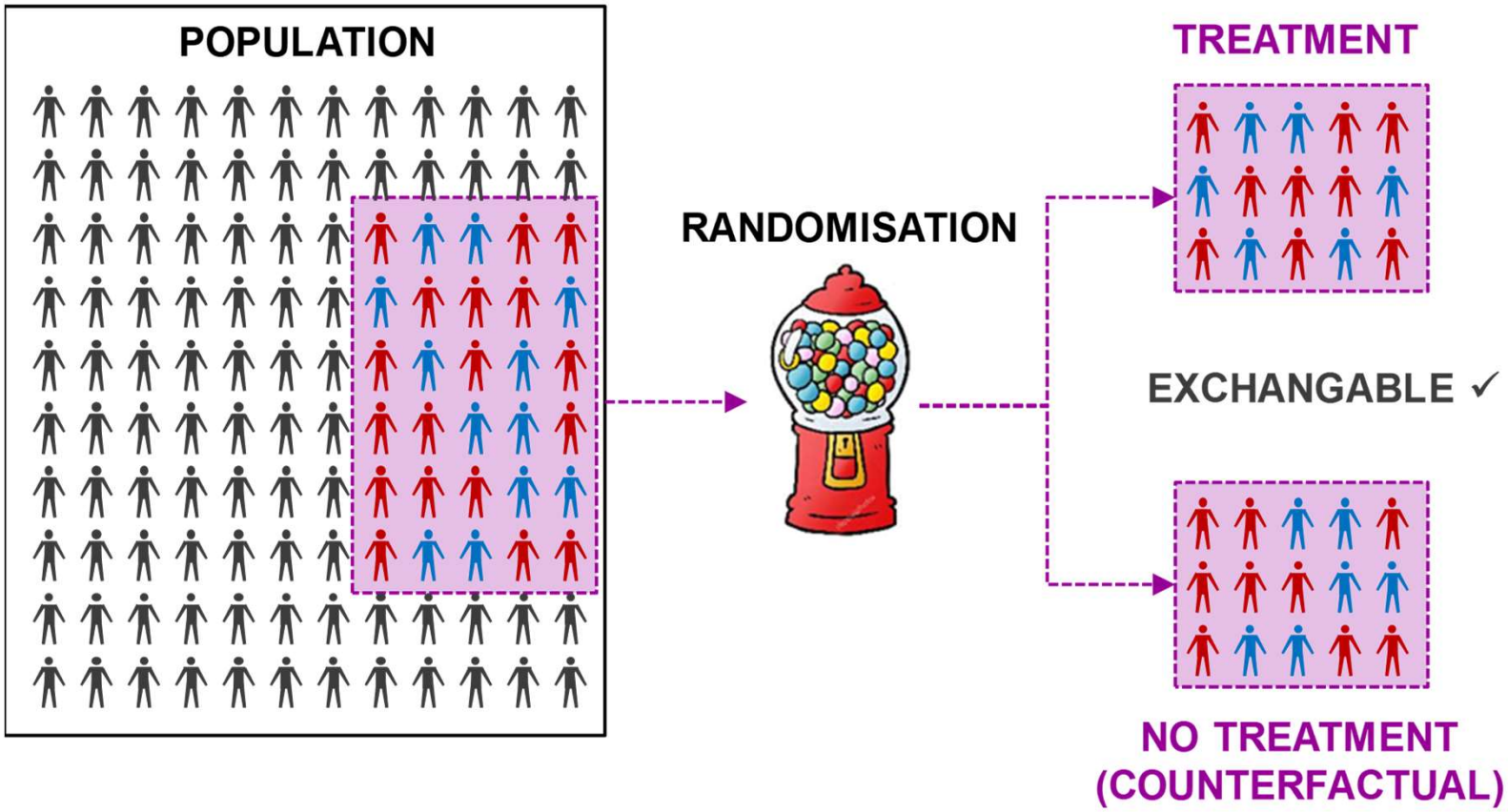


The Alan Turing Institute

UNCONDITIONAL EXCHANGEABILITY

#TMCFTURING

- The easiest way to do this is through randomisation

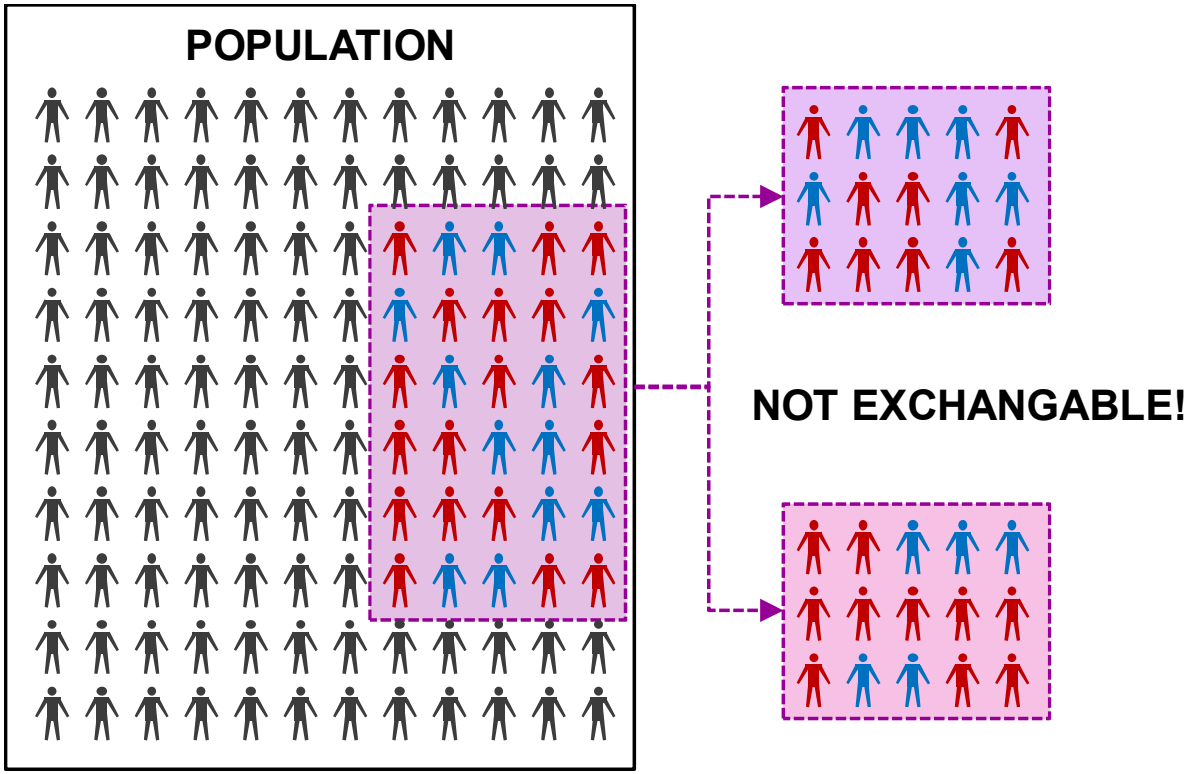


The Alan Turing Institute

CONDITIONAL EXCHANGEABILITY

#TMCFTURING

- Without randomisation, you have to identify and compare like-for-like subgroups

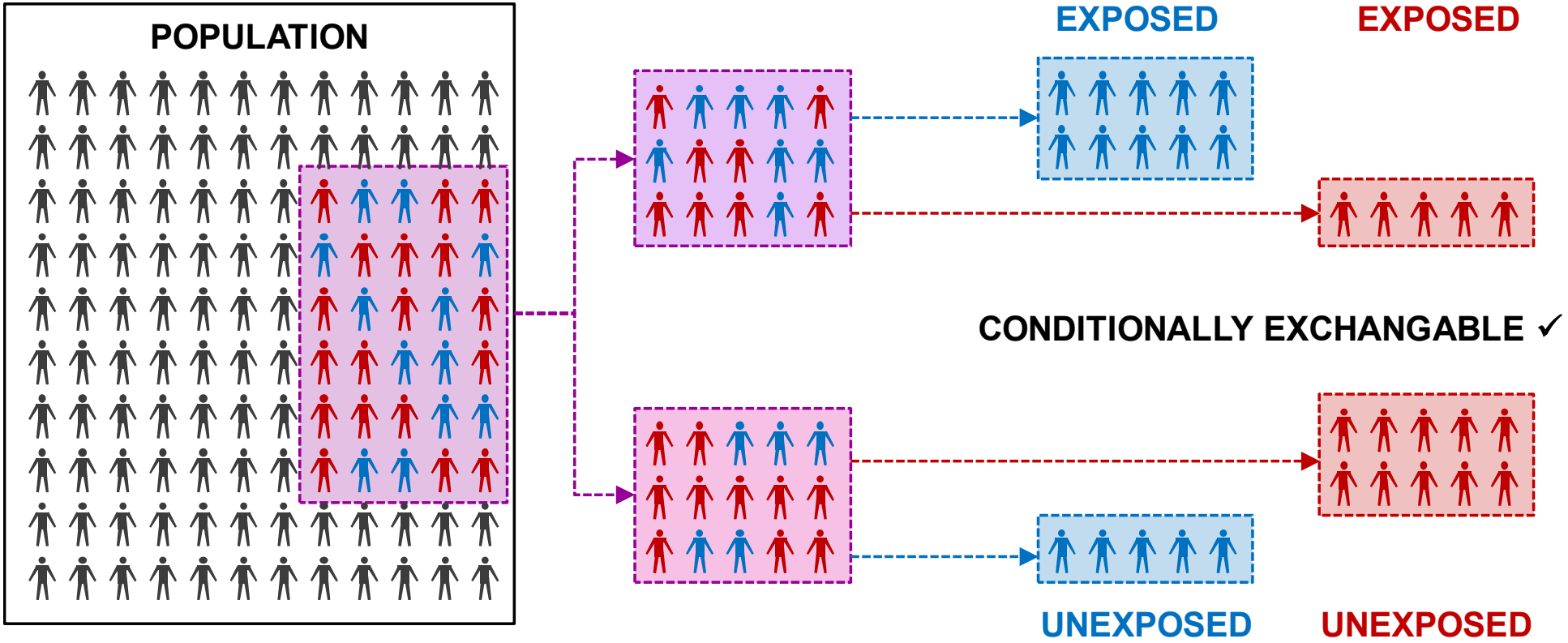


The Alan Turing Institute

CONDITIONAL EXCHANGEABILITY

#TMCFTURING

- Without randomisation, you have to identify and compare like-for-like subgroups

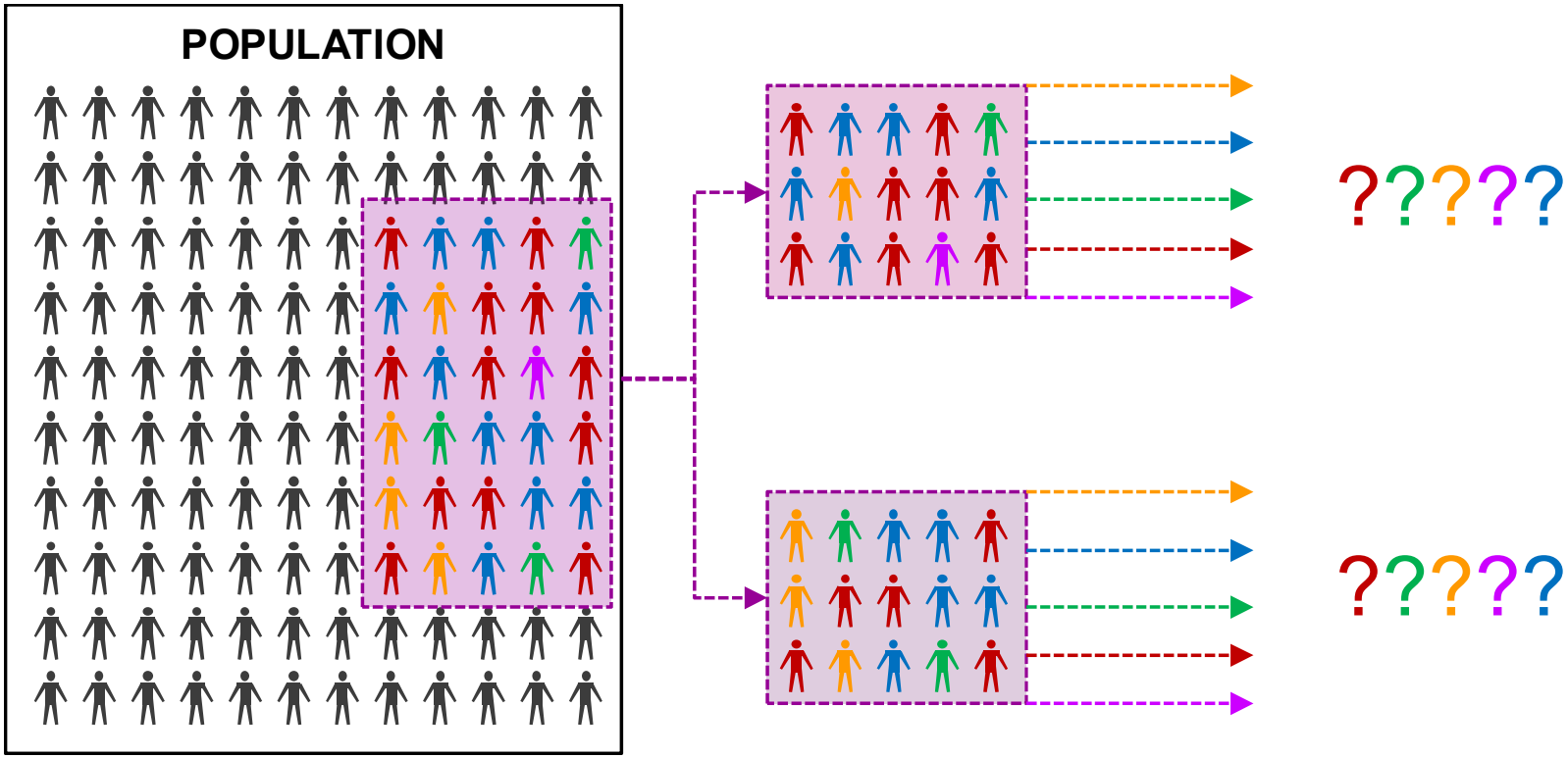


The Alan Turing Institute

CONDITIONAL EXCHANGEABILITY

#TMCFTURING

- Without randomisation, you have to identify and compare like-for-like subgroups

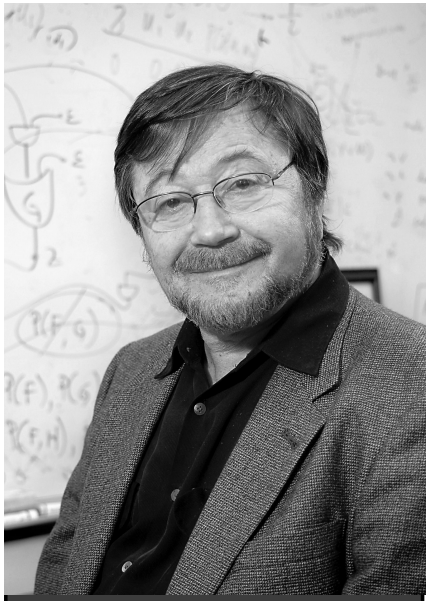


- Need to identify, measure, & control for *every* variable that (potentially) *causes* both the exposure and outcome (**confounder**)

The Alan Turing Institute

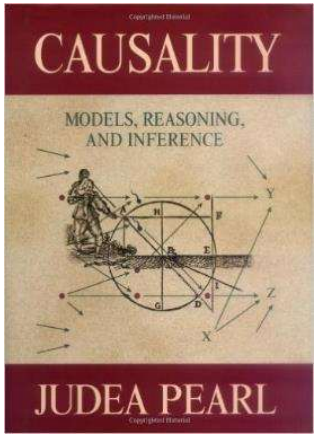
CAUSAL INFERENCE METHODS

#TMCFTURING

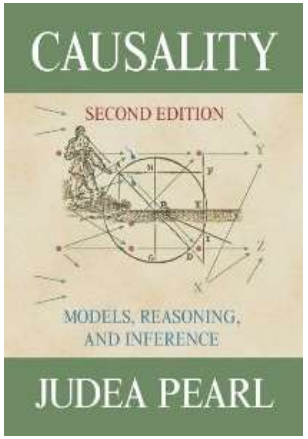


Judea Pearl, Computer scientist, philosopher, Turning Prize winner

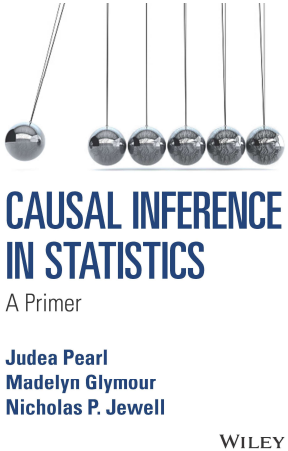
- **Causal inference methods –** like **Judea Pearl's 'Structural Causal Model' –** provide a mathematical and philosophical framework for considering causal effects drawing on:
  - **Probability theory**
  - **Counterfactual reasoning**
  - **Graphical model theory**



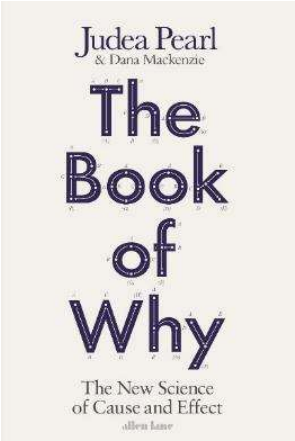
2000



2009



2016



2018

The Alan Turing Institute

IDENTIFICATION BEFORE ESTIMATION

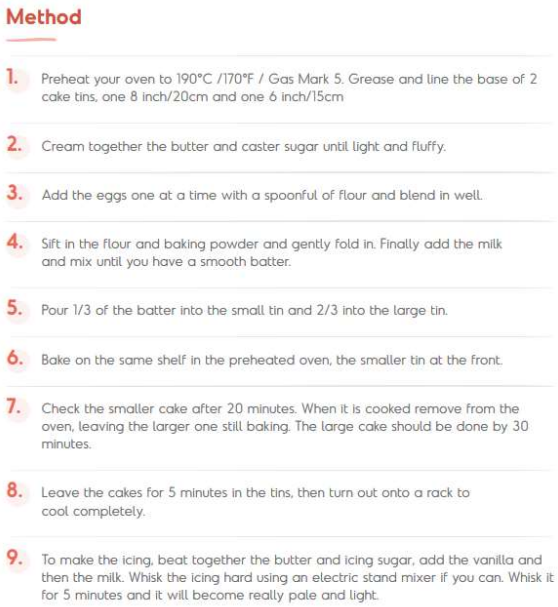
#TMCFTURING

- **Key feature:** Formally **identifying** your effect <u>before</u> analysis!

ESTIMAND
What you seek

E.g. The true difference in Y due to exposure

ESTIMATOR
How you will get there



Method

1. Preheat your oven to 190°C /170°F / Gas Mark 5. Grease and line the base of 2 cake tins, one 8 inch/20cm and one 6 inch/15cm
2. Cream together the butter and caster sugar until light and fluffy.
3. Add the eggs one at a time with a spoonful of flour and blend in well.
4. Sift in the flour and baking powder and gently fold in. Finally add the milk and mix until you have a smooth batter.
5. Pour 1/3 of the batter into the small tin and 2/3 into the large tin.
6. Bake on the same shelf in the preheated oven, the smaller tin at the front.
7. Check the smaller cake after 20 minutes. When it is cooked remove from the oven, leaving the larger one still baking. The large cake should be done by 30 minutes.
8. Leave the cakes for 5 minutes in the tins, then turn out onto a rack to cool completely.
9. To make the icing, beat together the butter and icing sugar, add the vanilla and then the milk. Whisk the icing hard using an electric stand mixer if you can. Whisk it for 5 minutes and it will become really pale and light.

E.g. Your regression model

ESTIMATE
What you get

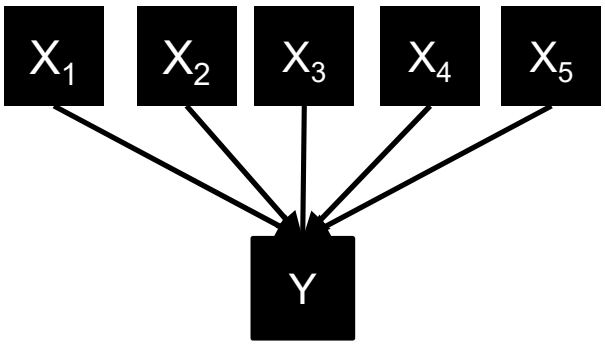
E.g. the estimated difference in Y from model coefficient

The Alan Turing Institute

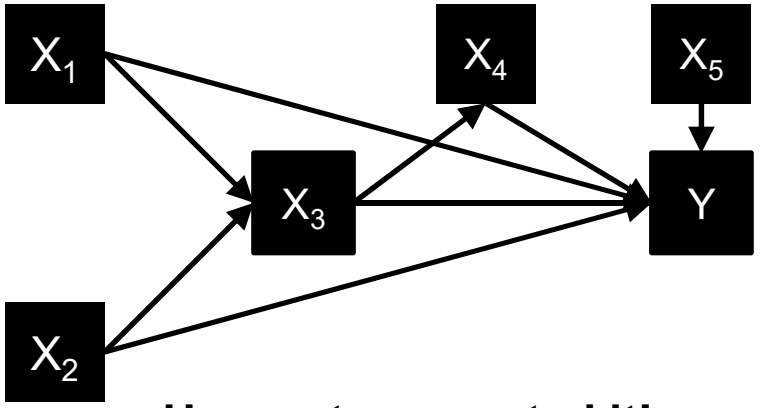
CAUSAL DIAGRAMS

#TMCFTURING

- **Causal diagrams** – such as **directed acyclic graphs** – encode counterfactual & probabilistic reasoning into diagrammatic form
- **First benefit**: helps to identify which variables are **confounders** and need controlling to provide exchangeable comparisons
- We must do this with our **outside knowledge**, the machine/software cannot!



How the machine/software sees it!

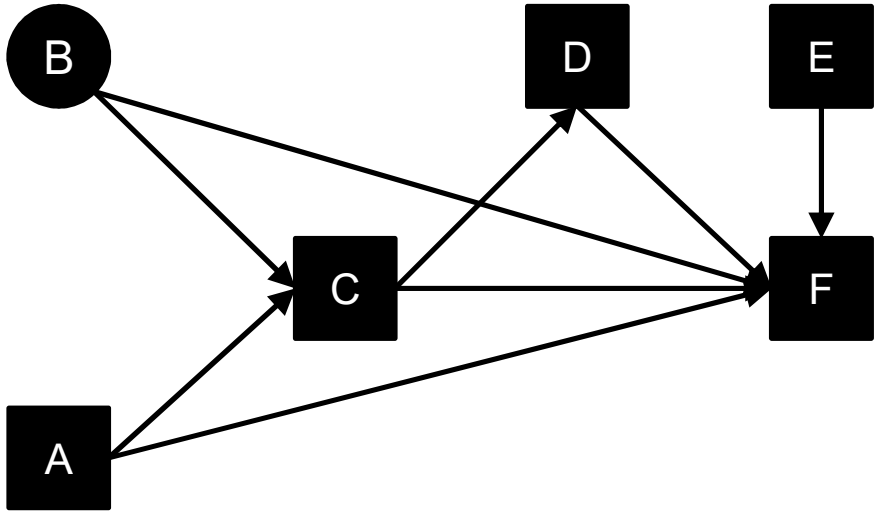


How nature created it!

The Alan Turing Institute

DIRECTED ACYCLIC GRAPHS

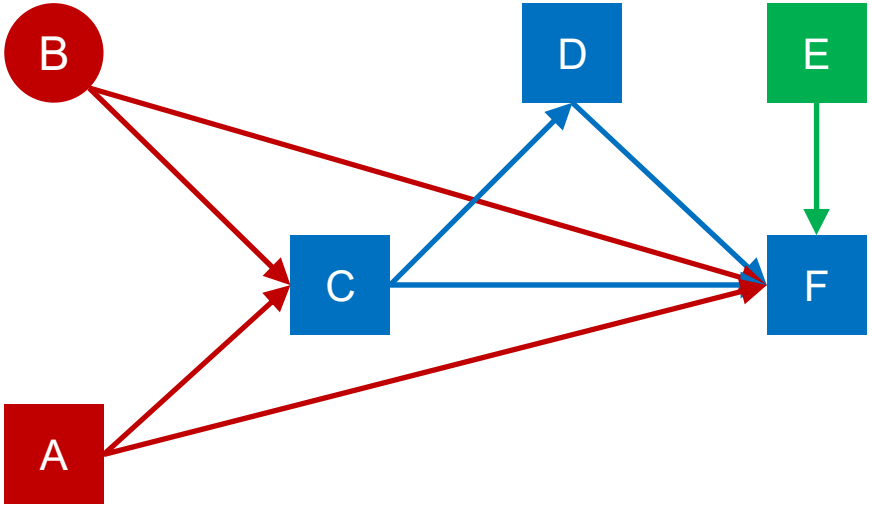
- **Directed Acyclic Graphs (DAGs)** are nonparametric representations of the (hypothesised) causal relationships between variables
  - Relationships between variables ('**nodes**') are represented by arrows ('**arcs**') creating **paths** between them
  - Paths can be **causal** or **confounding**; **open** or **closed**.
  - Open paths transmit correlations, closed paths do not



DIRECTED ACYCLIC GRAPHS

#TMCFTURING

- To estimate the causal effect of **C** on **F** (our '**focal relationship**'):
  - We want all **causal paths** open
  - And all **confounded paths** closed

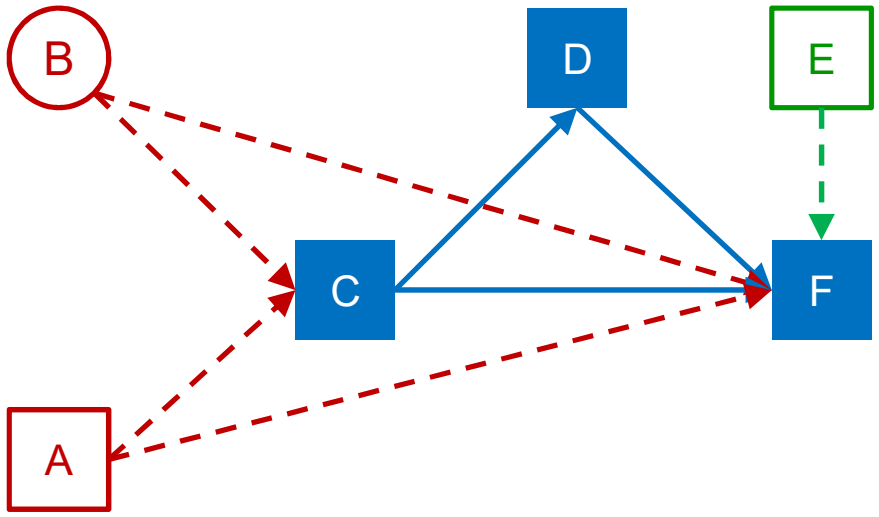


The Alan Turing Institute

DIRECTED ACYCLIC GRAPHS

#TMCFTURING

- To estimate the causal effect of **C** on **F** (the '**focal relationship**'):
  - We want all **causal paths** open
  - And all **confounded paths** closed
  - This means **controlling for** all **confounders** but no **mediators**



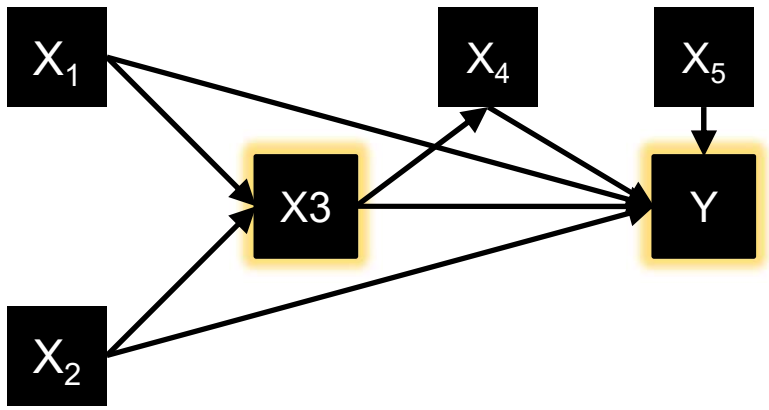
The Alan Turing Institute

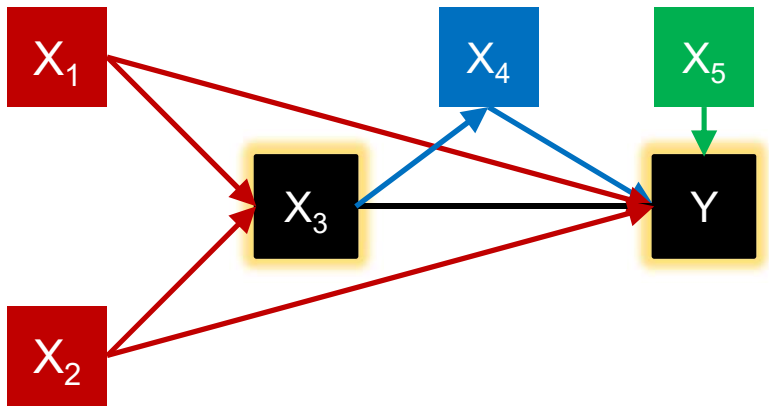
ONE EFFECT PER MODEL

#TMCFTURING

Example: Total causal effect of $\mathbf{X_3}$ on **Y**:

- Model should include **confounders** ($\mathbf{X_1}$, $\mathbf{X_2}$) [and **competing exposures** ($\mathbf{X_5}$)]
- $\mathbf{Y \sim X_3 + X_1 + X_2}$ [+ $\mathbf{X_5}$]





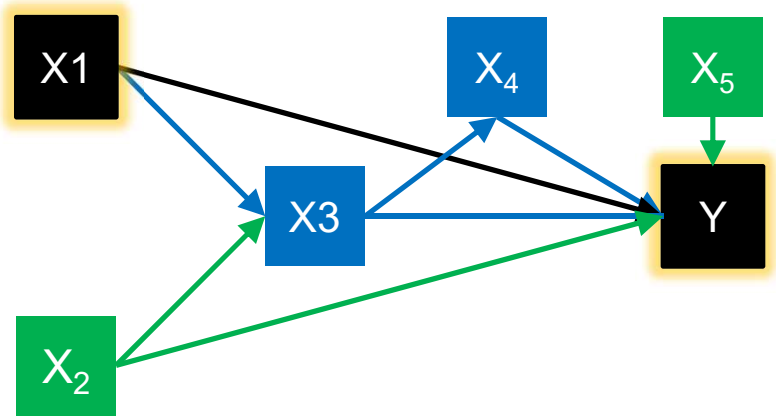
The Alan Turing Institute

ONE EFFECT PER MODEL

#TMCFTURING

- This model was constructed to estimate the effect of $\mathbf{X_3}$ on **Y**
- We should not interpret the other covariates ($\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_5}$), because they will not be appropriate controlled
- E.g. The coefficient on $\mathbf{X_1}$ is **NOT** the total causal effect of $\mathbf{X_1}$ on **Y**, because the inclusion of $\mathbf{X_3}$ means the causal path $\mathbf{X_1 \rightarrow X_3 \rightarrow Y}$ will be closed.

$$\mathbf{Y \sim X_3 + X_1 + X_2 + X_5}$$



The Alan Turing Institute

TABLE 2 FALLACY

#TMCFTURING

- The tradition of including all 'predictors' of our outcome (Y) in a **single model**, and interpreting the coefficients ($\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, $\mathbf{X_4}$ $\mathbf{X_5}$) as '**mutually adjusted**' effects has been dubbed the '**Table 2 Fallacy**'

The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients

Daniel Westreich* and Sander Greenland

* Correspondence to Dr. Daniel Westreich, Department of Obstetrics and Gynecology, Duke Global Health Institute, Duke University, DUMC 3967, Durham, NC 27710 (e-mail: daniel.westreich@duke.edu).

Initially submitted January 13, 2012; accepted for publication October 11, 2012.

It is common to present multiple adjusted effect estimates from a single model in a single table. For example, a table might show odds ratios for one or more exposures and also for several confounders from a single logistic regression. This can lead to mistaken interpretations of these estimates. We use causal diagrams to display the sources of the problems. Presentation of exposure and confounder effect estimates from a single model may lead to several interpretative difficulties, inviting confusion of direct-effect estimates with total-effect estimates for covariates in the model. These effect estimates may also be confounded even though the effect estimate for the main exposure is not confounded. Interpretation of these effect estimates is further complicated by heterogeneity (variation, modification) of the exposure effect measure across covariate levels. We offer suggestions to limit potential misunderstandings when multiple effect estimates are presented, including precise distinction between total and direct effect measures from a single model, and use of multiple models tailored to yield total-effect estimates for covariates.

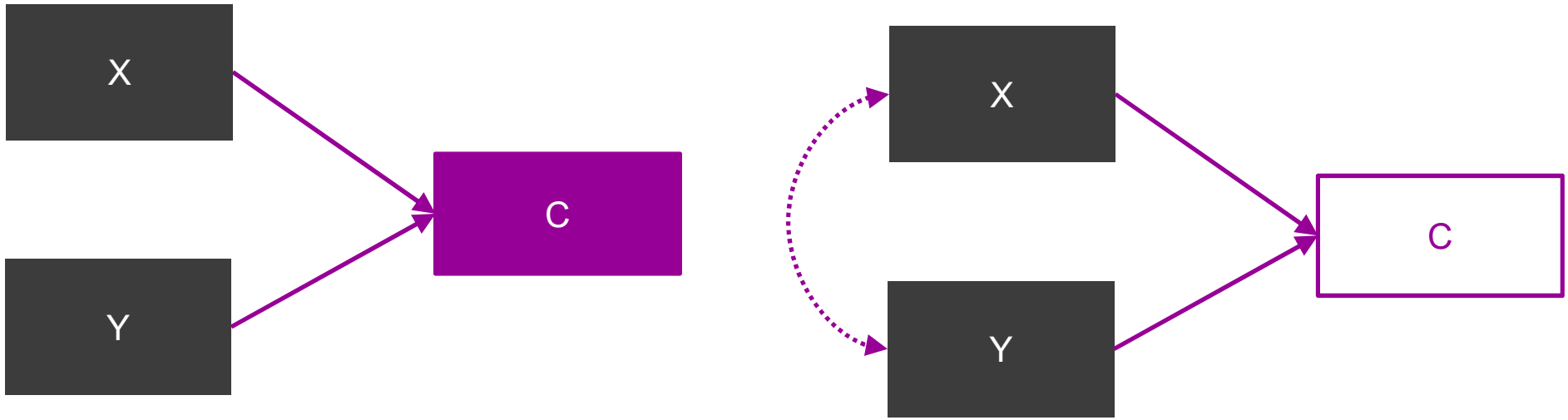
Westreich & Greenland. *Am J Epidemiol*. 2013;177(4):292–298.

The Alan Turing Institute

COLLIDER BIAS

#TMCFTURING

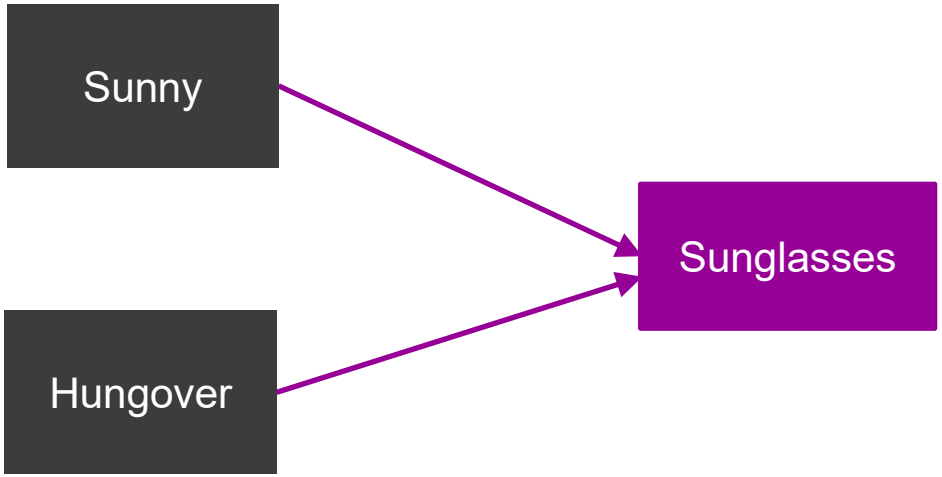
- DAGs have also been revolutionary in helping us understand **conditional dependencies** – non-causal associations introduced by conditioning



- Conditioning on **C** would transmit a **non-causal association** between **X** and **Y**

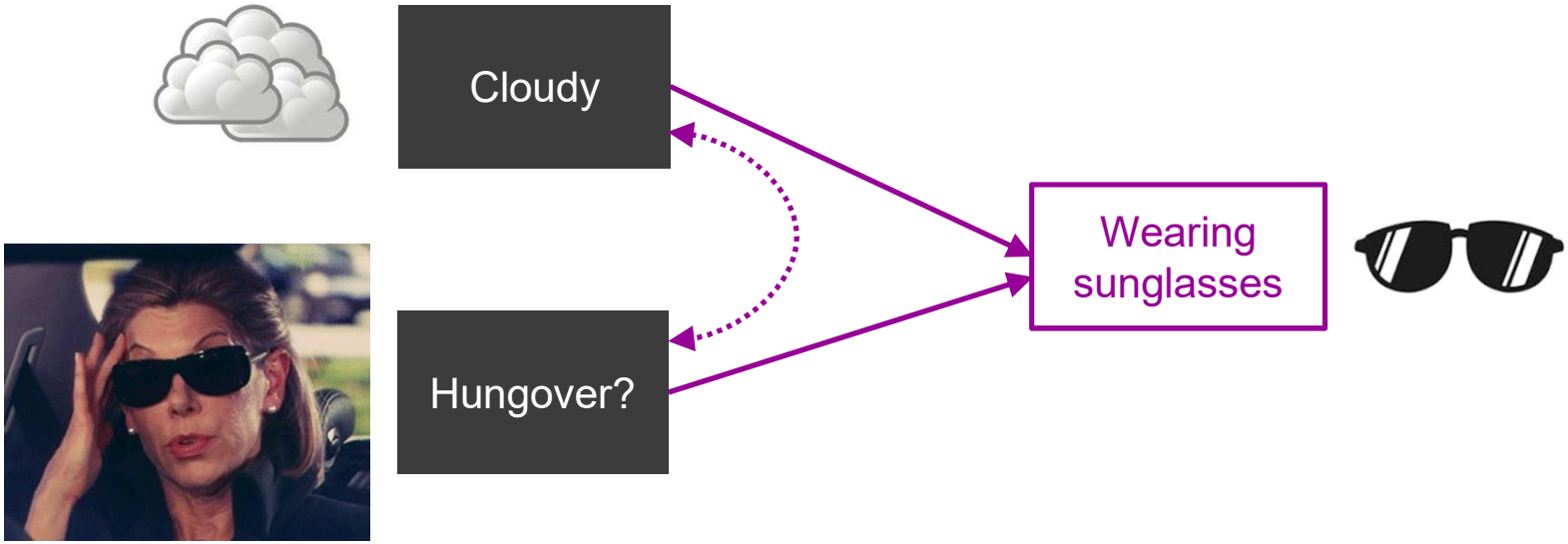
The Alan Turing Institute

- **Example:** The Sunglasses Revelation
  - British people typically wear sunglasses when the **sun is shining**.
  - They sometimes also wear them when they are **very hungover**



  - On average, suppose hangovers occur equally all year round, regardless of weather

- **Example:** The Sunglasses Revelation
  - You see someone **wearing sunglasses** on a **cloudy day**



  - What does this tell you about their likelihood of being **hungover**?

THE BIRTHWEIGHT PARADOX

#TMCFTURING

American Journal of Epidemiology

Vol. 164, No. 11
DOI: 10.1093/aje/kwj275
Advance Access publication August 24, 2006

Practice of Epidemiology

The Birth Weight "Paradox" Uncovered?

Sonia Hernández-Díaz[1,2], Enrique F. Schisterman[3], and Miguel A. Hernán[1]

[1] Department of Epidemiology, Harvard School of Public Health, Boston, MA.
[2] Slone Epidemiology Center, Boston University, Boston, MA.
[3] Epidemiology Branch, National Institute of Child Health and Human Development, Bethesda, MD.

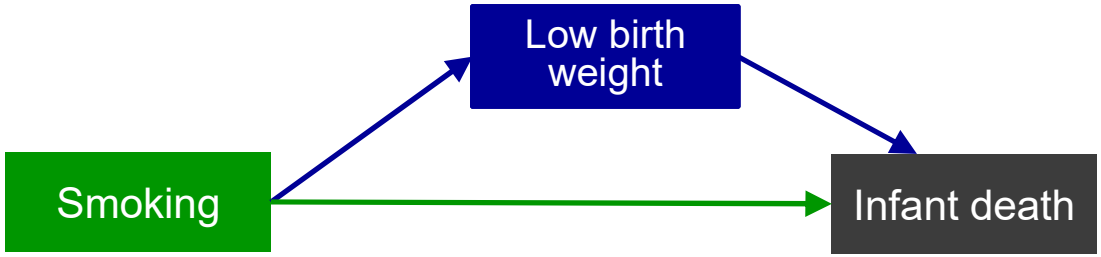
Received for publication February 7, 2005; accepted for publication January 23, 2006.

Hernández-Diaz et al. Am J Epidemiol. 2006;164(11):1115-20

The Alan Turing Institute

THE BIRTHWEIGHT PARADOX

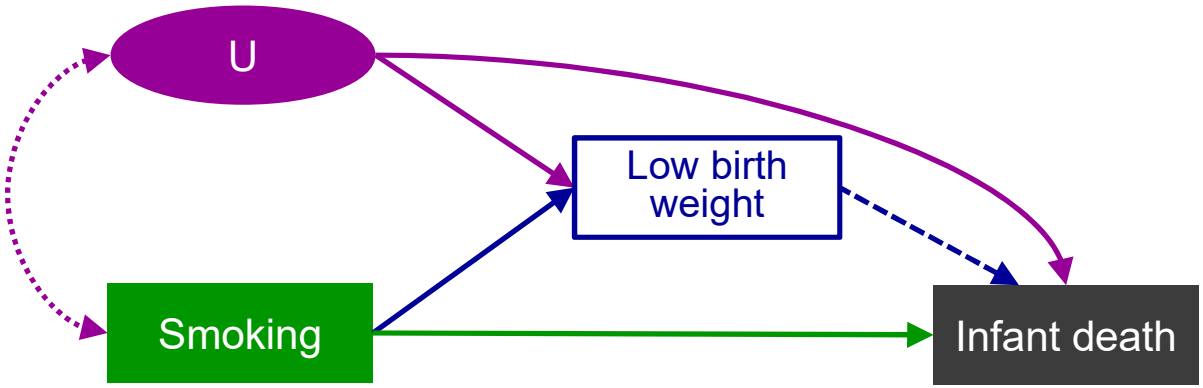
#TMCFTURING



The Alan Turing Institute

THE BIRTHWEIGHT PARADOX

#TMCFTURING



The Alan Turing Institute

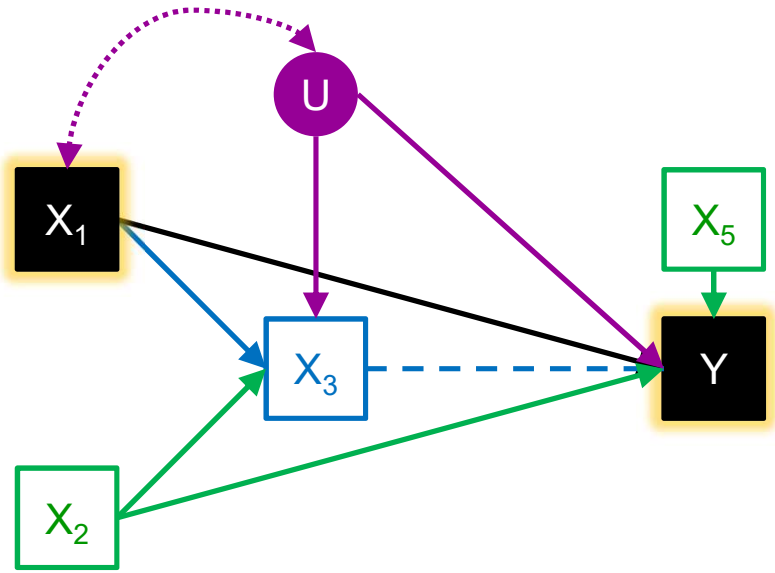
ONE EFFECT PER MODEL

#TMCFTURING

- In our model, we recognised the coefficient on $\mathbf{X_1}$ was **NOT** the total causal effect of $\mathbf{X_1}$ on **Y**, but we might think it gave a meaningful (direct) causal effect.

- In fact, the apparent effect will be **collider-biased** by all other mutual causes of $\mathbf{X_3}$ and **Y** (**mediator-outcome confounders**)

$$\mathbf{Y} \sim \mathbf{X_3} + \mathbf{X_1} + \mathbf{X_2} + \mathbf{X_5}$$



- This can **<u>reverse</u>** the apparent direction of effect (**reversal paradox**)!

The Alan Turing Institute

SUMMARY

#TMCFTURING

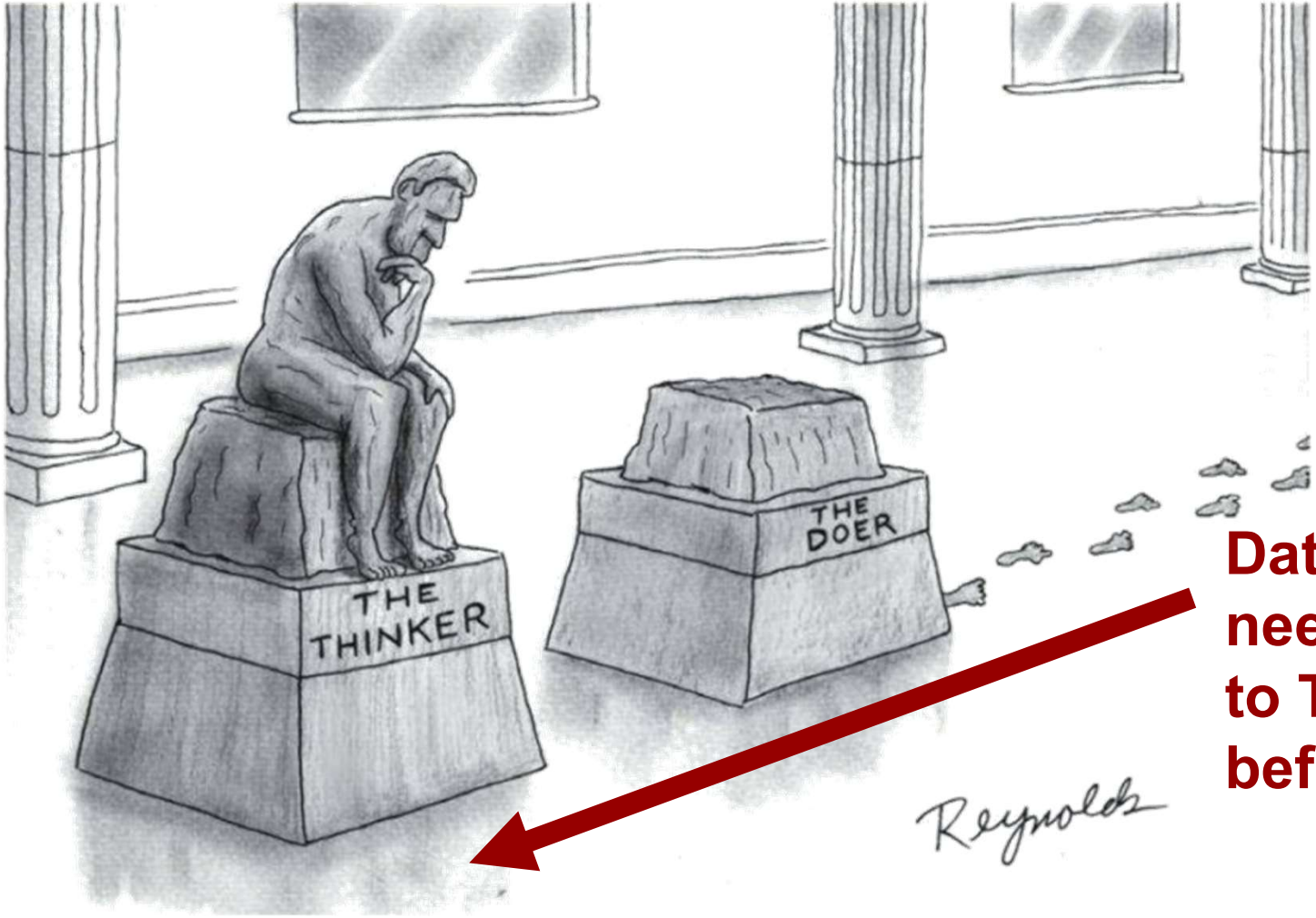
- **Causal inference** is one of the three core tasks of data science
- Unlike **description** and **prediction**, it requires external contextual knowledge
- Traditional data-driven approaches have a poor record at causal inference
- Joining the **causal revolution** means embracing our **causal ambitions**, & recognising the need for a completely different approach
- This starts with explicitly **identifying** what we want to know before **estimating**
- **Causal diagrams** are a simple & transparent aid to causal reasoning, & are helpful for understanding many issues in non-experimental data

The Alan Turing Institute

CAUSAL INFERENCE APPROACH

#TMCFTURING

- Causal inference requires external knowledge → careful thought



Data scientists need more time to THINK before they DO!

The Alan Turing Institute

FURTHER LEARNING

#TMCFTURING

- **Differences between prediction and causal inference**
  - Arnold KF *et al*. Generalised linear models for prognosis and intervention: Theory, practice, and implications for machine learning. ArXiv 2019. https://arxiv.org/abs/1906.01461v2
- **Using DAGs in your research**
  - Tennant PWG *et al*. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. MedArXiv 2019. https://doi.org/10.1101/2019.12.20.19015511
- **Causal Inference Book**
  - Hernán MA & Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

The Alan Turing Institute