ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 13: Finding Causes

William Marsh

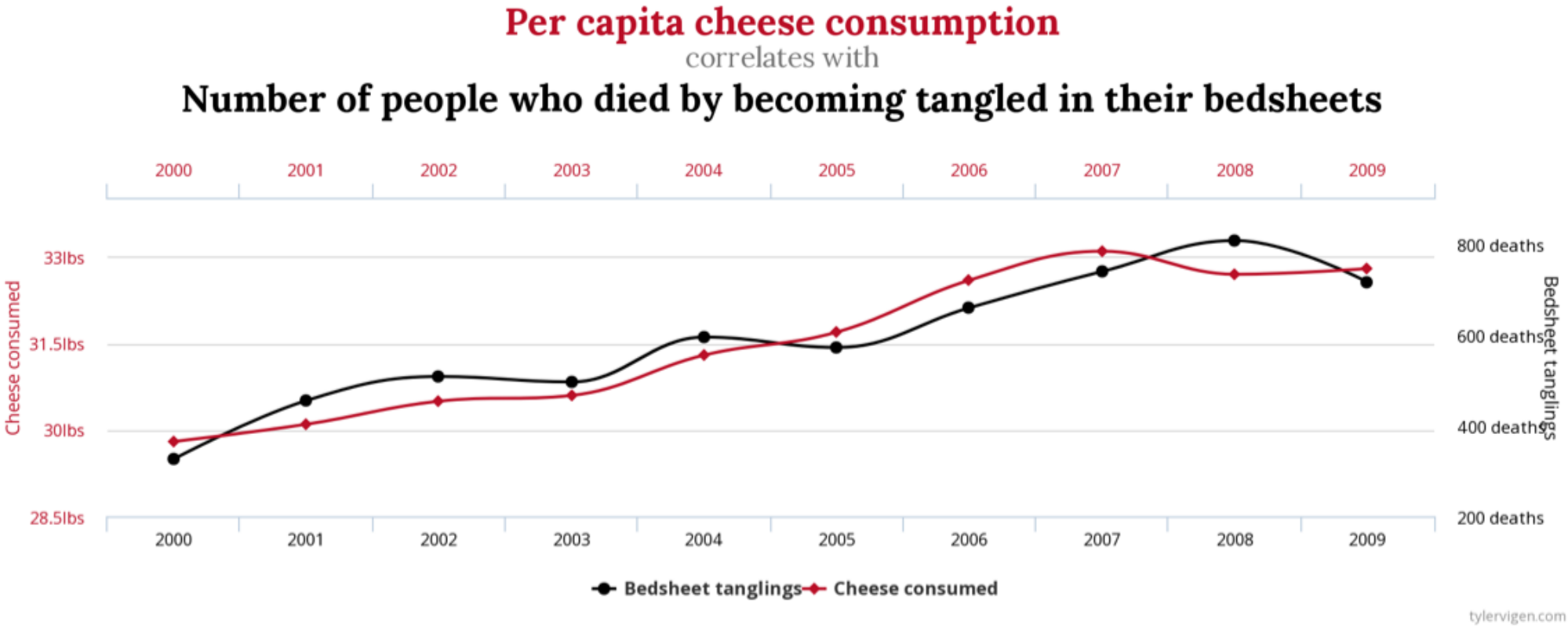
Outline

- Aim: Understand why experiments are used to find causes and introduce ideas of causal analysis

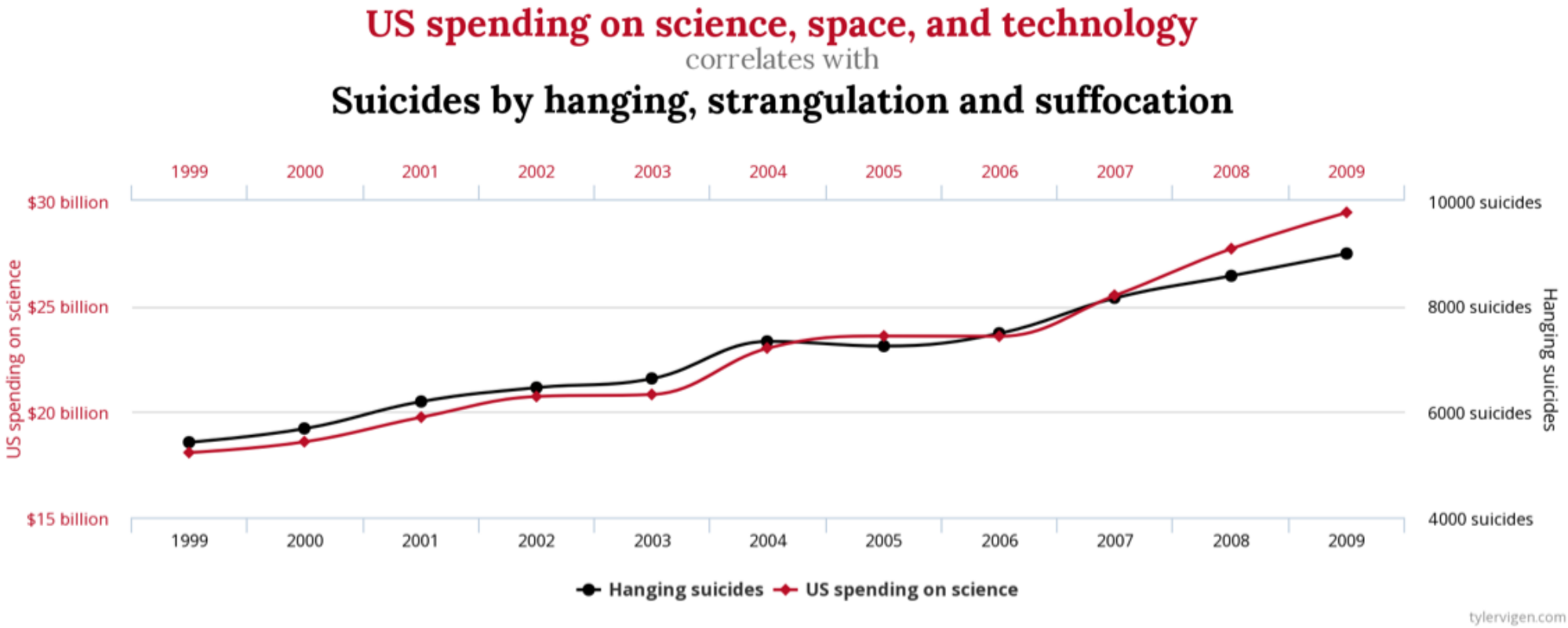
- Correlation and causation
- Randomised Controlled Trial (RCT)
- Causal discovery
  - Criteria
  - Causal maps and the ladder of causation

Correlation and Causation

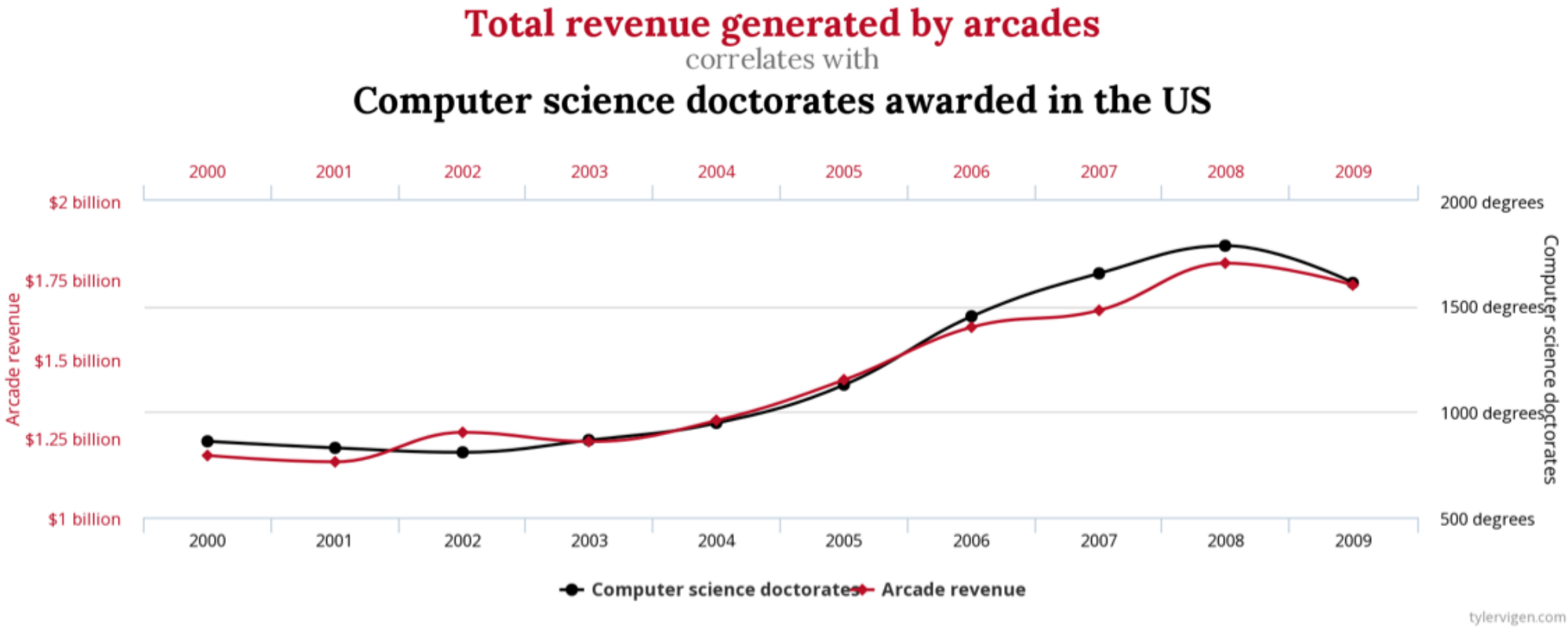
Correlation is not causation



Correlation is not causation



Correlation is not causation



What Do We Mean by Cause?

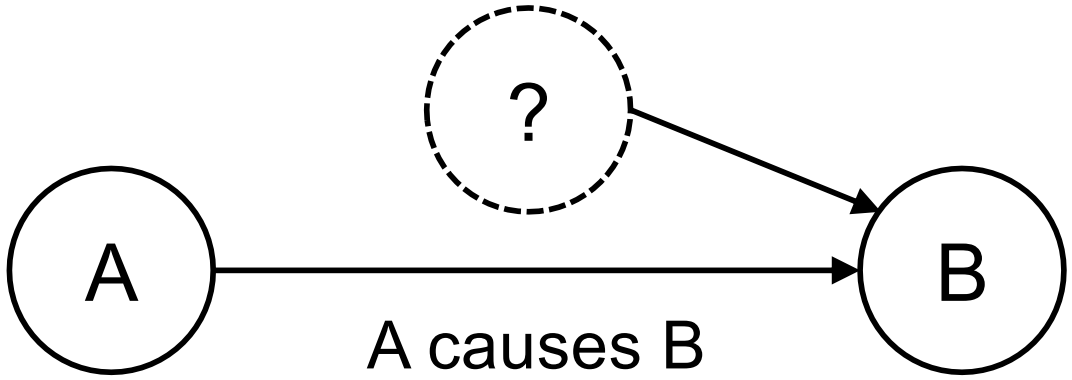
- Philosophers find this difficult!

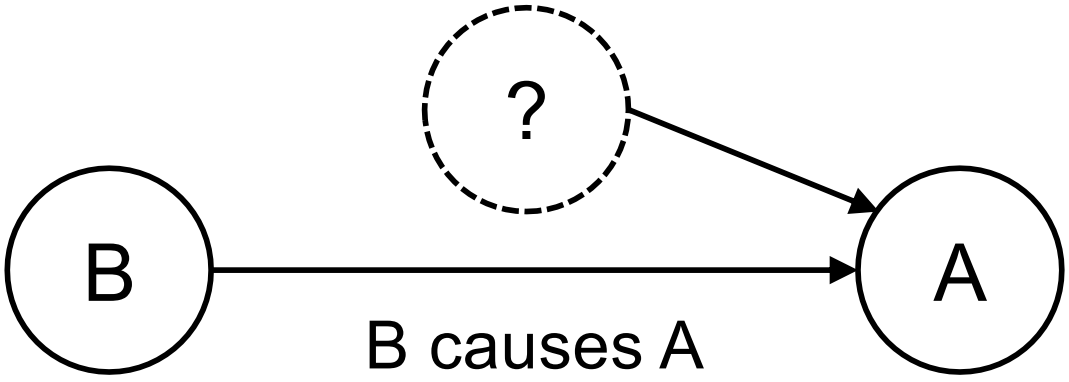
- A useful idea is
  - If a change to A leads to a change to B
  - ... then A is a cause of B
- Example: an infection (I) causes a fever (F) (a high temperature)
  - If I find a way to lower the temperature, this will not change the infection: F is not the cause of I
  - If I remove the infection, the fever will diminish: I is the cause of F

- This idea is useful but not without problems

How could A and B be Correlated?

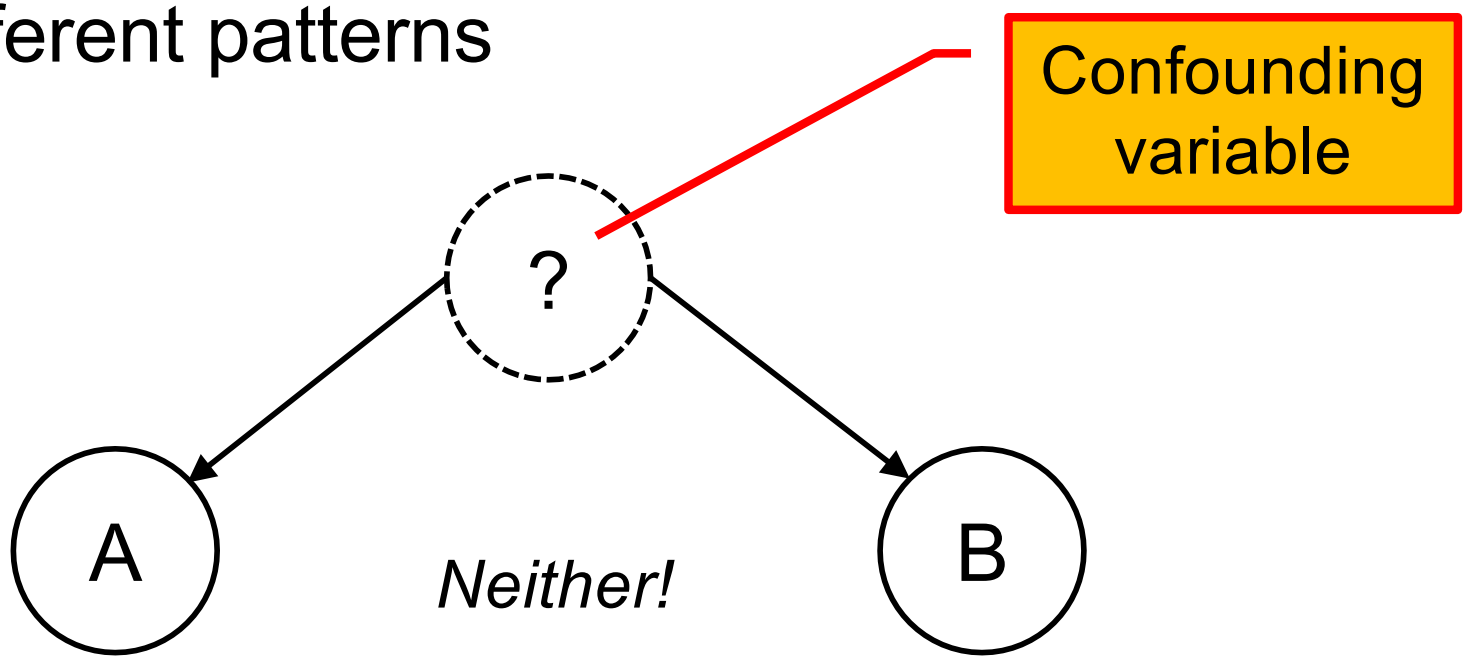
- Different patterns





How could A and B be Correlated?

- Different patterns



- Possibly unknown common cause
  - Could add to relationship between A and B
  - Could entirely explain the correlation

Randomised Controlled Trial

Confounding Variables

- Gather data and shows that people who play Beatles are more likely to have a heart attack that those who play (?? Kylie)

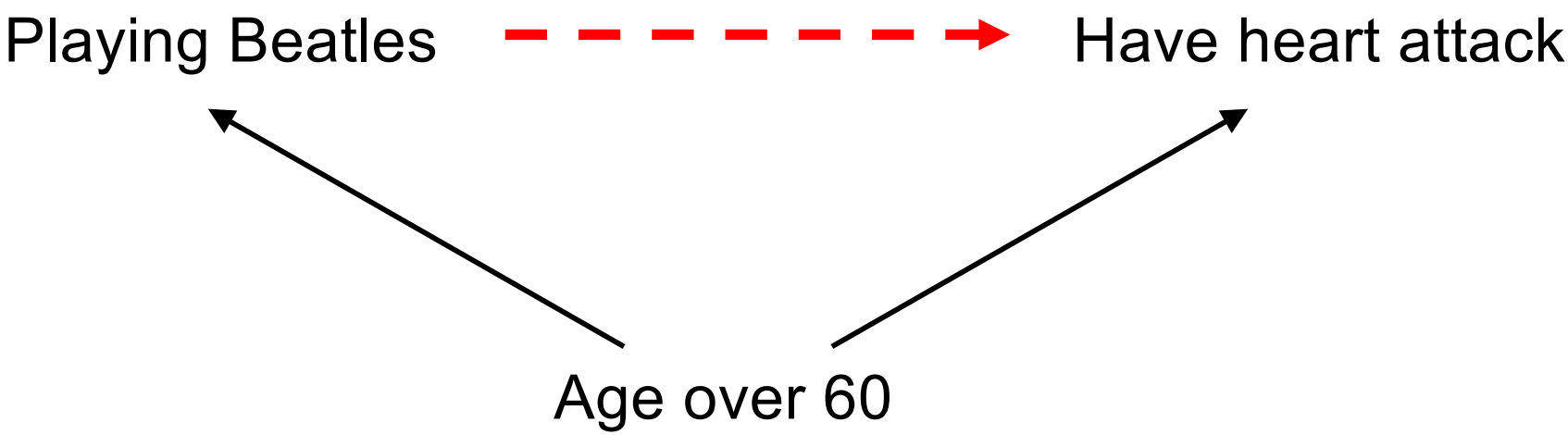
Is it a cause?

Can We Guess Your Age Based on Your Taste in Music?

- *web site offering a quiz*

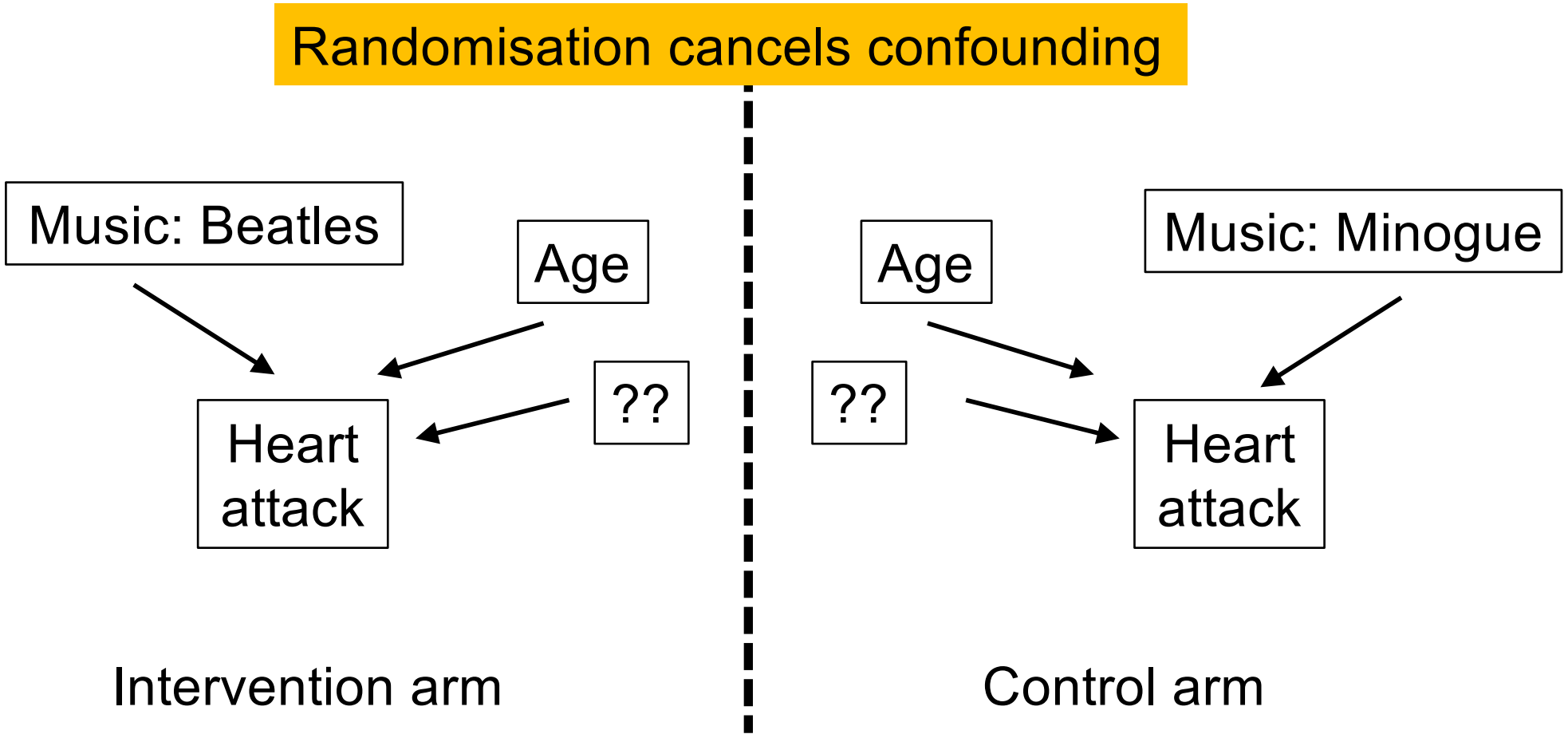
Confounding Variables

- Gather data and shows that people who play Beatles are more likely to have a heart attack that those who play (?? Kylie)



Randomised Trials

- Investigate effect of music on heart attacks
  - (Is this ethical?)



RCTs

- Best way to discover causes
  - Randomised for unknown (and known) confounders
  - Blinded: patient and investigator do not know arm
  - Stratify for known confounders

- Disadvantages
  - Expense and time taken
  - Ethics
  - Impossible to blind

Observational Data

- Gather data without a trial
  - Most “big data” like this
- Advantages
  - Much larger dataset
  - Potential cheaper
- Disadvantage
  - Confounding: is the association causative?
  - *For some purposes it does not matter*

Preview of C/W 2

Overview

Uses the Texas Bridge data (notebook 3)

- Can you:
  - Carry out an exploratory analysis of data with both continuous and categorical variables
  - Use regression modelling to look at the effect of predictor variables on a target variable.
- Scenario
  - Investigate the use of (given) variables to predict bridge condition

1. Age (from variable Year)
2. Average use (AverageDaily)
3. Trucks_percent
4. Material (from variable Material)
5. Design (variable Design)

Issues

- As before, writing to a ‘domain expert’
- Need for (limited) judgement
  - Decisions based on evidence
  - E.g.

You are recommended to exclude very old bridges (possibility the historic ones)

Comparing an RCT with a Cohort Study: Example

Previous infection and vaccination to protect against COVID-19

Protecting Against COVID-19

Vaccination

- Randomised trial
  - Allocated randomly to intervention or control ‘arm’
- Blinded
  - x 2: patient & investigator
  - Everyone injected
- Surveillance
  - E.g. regular testing
- Result
  - Compare infections in intervention versus control

Previous Infection

- Prospective cohort study
- Non-random allocation
  - Previous infection
  - No previous infection
- Surveillance
  - E.g. regular testing
- Result
  - Compare infections in intervention versus control

Why not an RCT?

Common Issues

- Detailed protocol
  - Exactly how the study runs and results are collected
- Comparison: two (or more) ‘arms’
- Problem of low number of infections
  - Recruit large number
  - Wait!
- Detection of infection
  - Tests not perfect (accuracy only approximately known)

Study of Previous Infection

https://doi.org/10.1136/bmj.n124

volunteered

NHS staff who volunteered for the study were assigned to either the positive cohort (antibody positive or prior polymerase chain reaction (PCR) antibody test positive) or negative cohort (antibody negative, not previously known to be PCR antibody positive). They attended regular PCR and antibody testing (every two to four weeks) …

84% female and younger

Between 18 June and 9 November 2020, 20 787 staff (84% female, [illegible] median age 45.9) were included in this analysis, of whom 6614 (32%) were assigned to the positive cohort and 14 173 (68%) to the negative cohort.

non-compliance

By 24 November 2020, 409 new infections were detected in the negative cohort, of whom 249 (79%) were symptomatic at infection. Meanwhile, 40 (12%) were asymptomatic and 26 (9%) did not complete a questionnaire at the time of their symptoms.

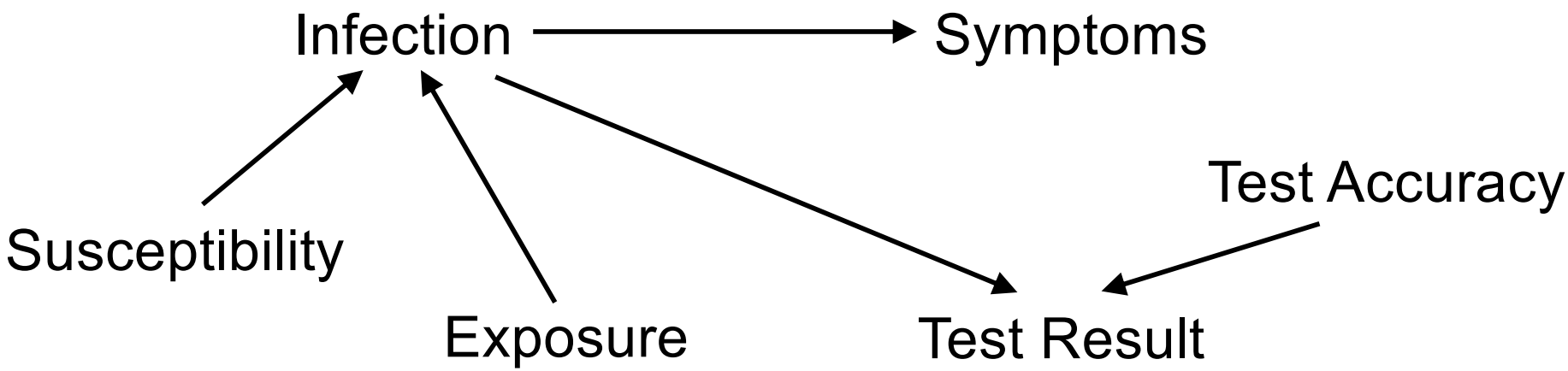
Uncertainty

The researchers detected 44 reinfections in the positive cohort, 15 (34%) of which were symptomatic. Some 42 were defined as possible (two positive PCR samples 90 or more days apart, or a new PCR positive at least four weeks after an antibody positive result), and two were defined as probable (additionally required quantitative serological data or supportive viral genomic data). This compares with 318 new PCR positive infections (249 symptomatic) and 94 antibody seroconversions in the negative cohort.

The researchers calculated that adjusted odds ratio was 0.17 for all reinfections (95% confidence interval 0.13 to 0.24) compared with PCR confirmed primary infections, equating to 83% protection. The median interval between primary infection and reinfection was over 160 days.

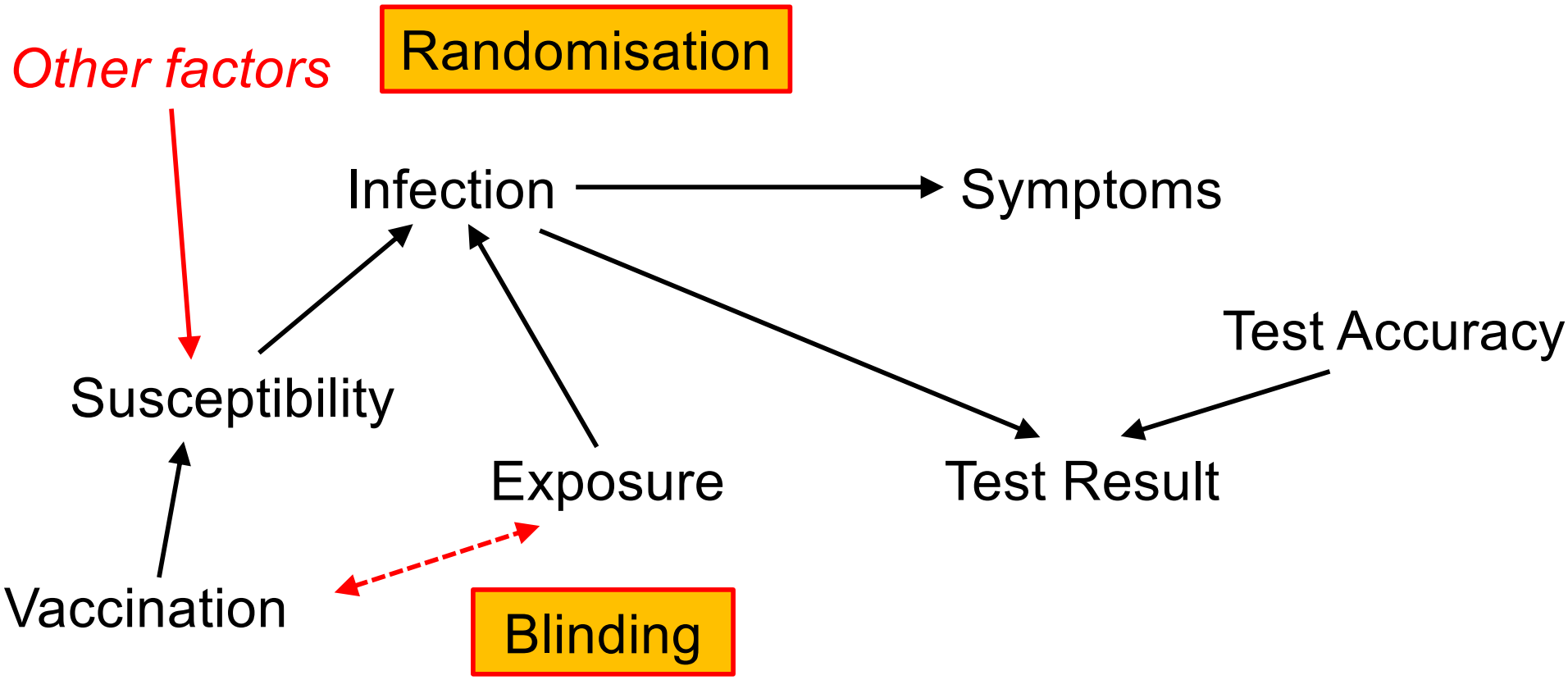
When looking at only symptomatic cases supported by positive PCR results, previous infection reduced the odds of reinfection by at least 90% (adjusted odds ratio 0.06 with 95% CI 0.03 to 0.09).

Causal Map

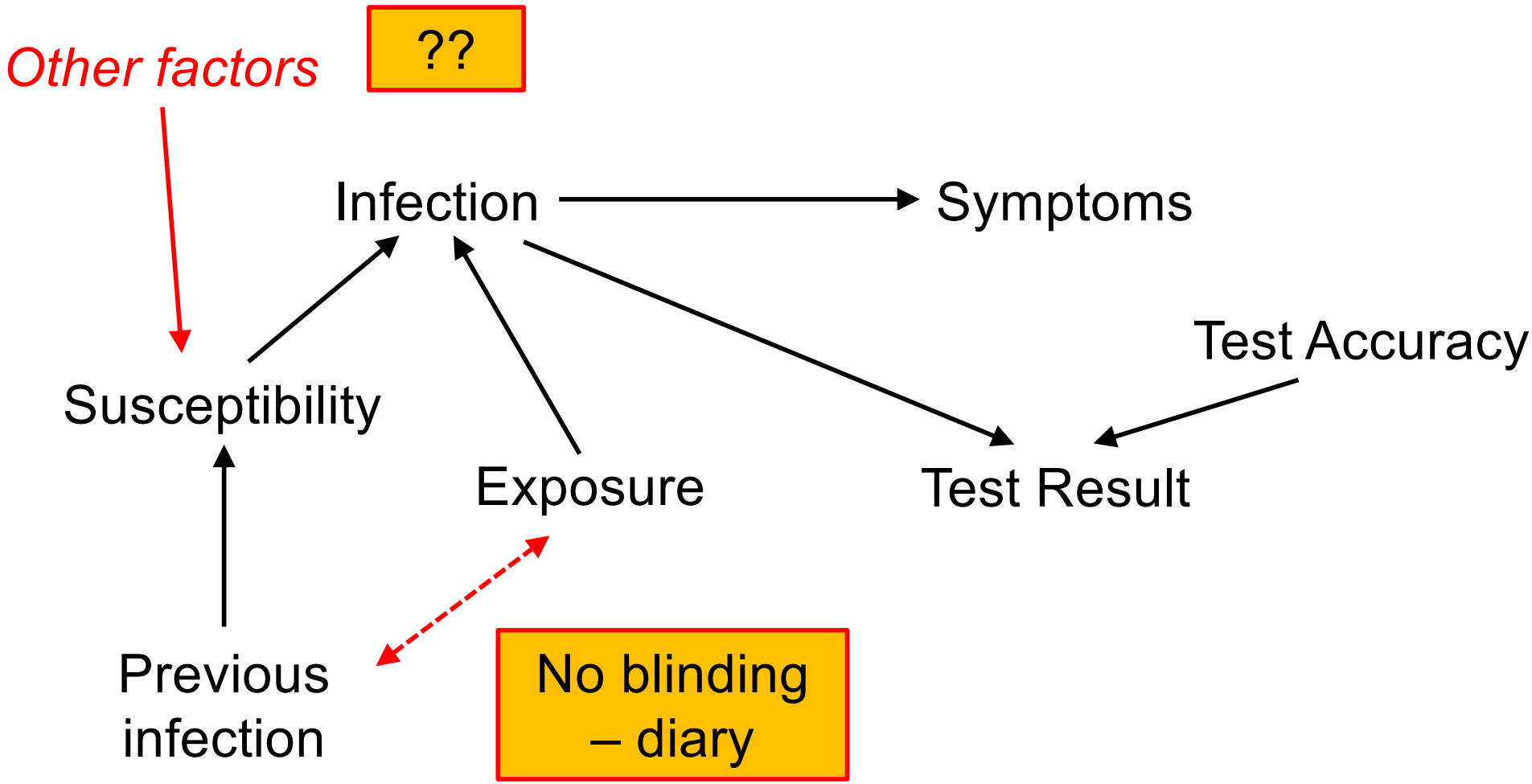


Invented by me: illustrative only

Causal Map: Vaccination



Causal Map: Previous Infection



What the Researchers Said

Study: 18 June and 24 November 2020

The research team warned, however, that early evidence from the next stage of the study suggested that some people who are themselves protected by antibodies still carry high levels of virus and could continue to infect others.

"*We now know that most of those who have had the virus, and developed antibodies, are protected from reinfection, but this is not total and we do not yet know how long protection lasts. Crucially, we believe people may still be able to pass the virus on*."

The study team stressed that these results give no insight into the effects of vaccines or the new more transmissible variant in the UK, because of the time period analysed. These factors will be considered in later stages of analysis.

Quiz

Epidemiology: Discovery of Causes

Bradford-Hill Criteria I

- Strength (effect size)
  - A larger association is more likely to be causal
- Consistency (reproducibility):
  - Observed by different persons in different places
- Specificity:
  - Specific population and site
- Temporality:
  - Effect occurs after the cause

Sir Austin Bradford Hill CBE FRS was an English epidemiologist and statistician, pioneered the randomised clinical trial and, together with Richard Doll, demonstrated the connection between cigarette smoking and lung cancer.

Bradford-Hill Criteria II

- Biological gradient (dose-response relationship):
  - Greater exposure increases incidence
- Plausibility:
  - A plausible mechanism
- Coherence:
  - Epidemiological and laboratory findings coher
- Experiment:
  - Occasionally!
- Analogy:
  - Similarities with other associations.

Does Big Data Change This?

Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology

https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC4589117/

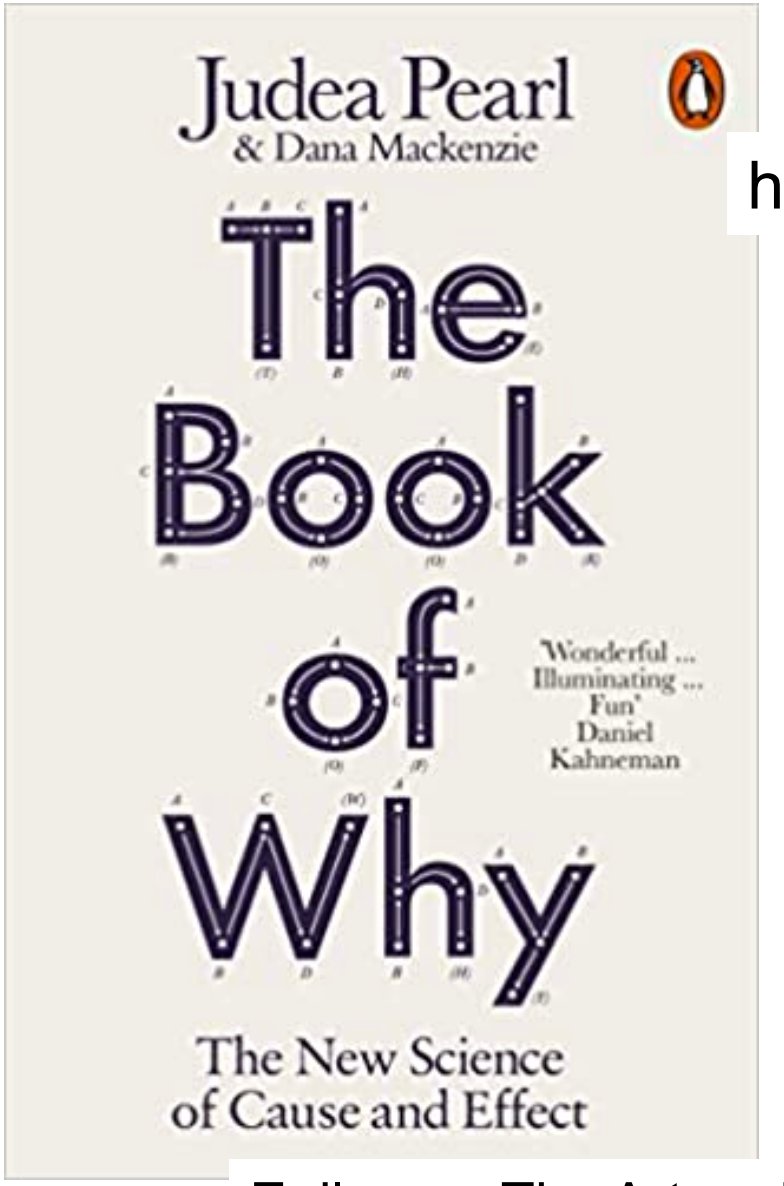
- Integrated approach: different types of data
- Danger of ‘false discovery’
  - Association and ‘statistical significance’ even less reliable

Ladder of Causation and Causal Models

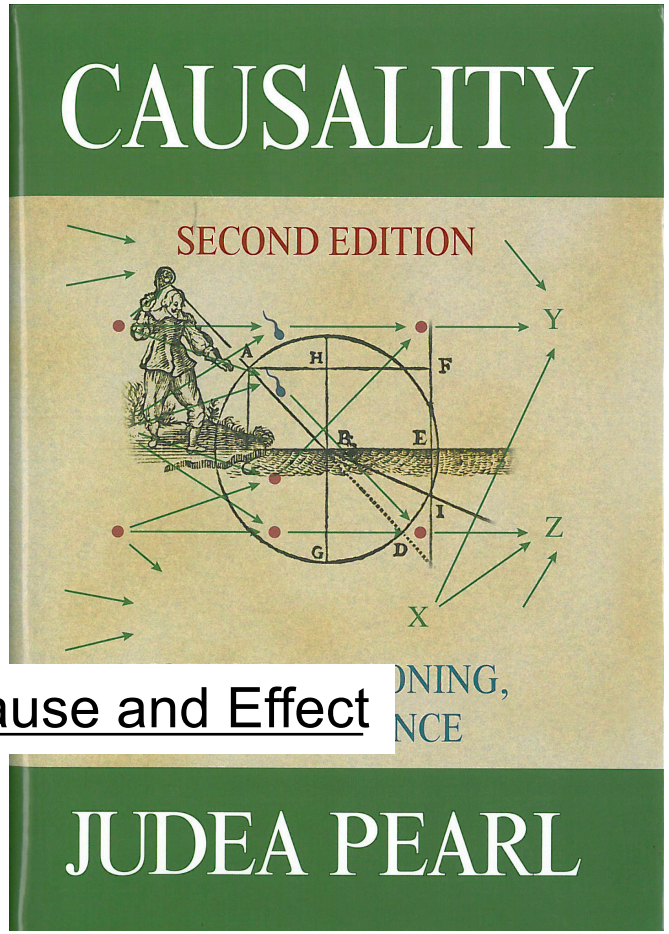
New Work on Causation

- Particularly associated with an academic from UCLA called Judea Pearl
  - Winner of the Turing prize

- Main argument
  - Distinguish between statistical and causal questions
  - Reason about causal questions using causal models
    - Data
    - Causal assumptions

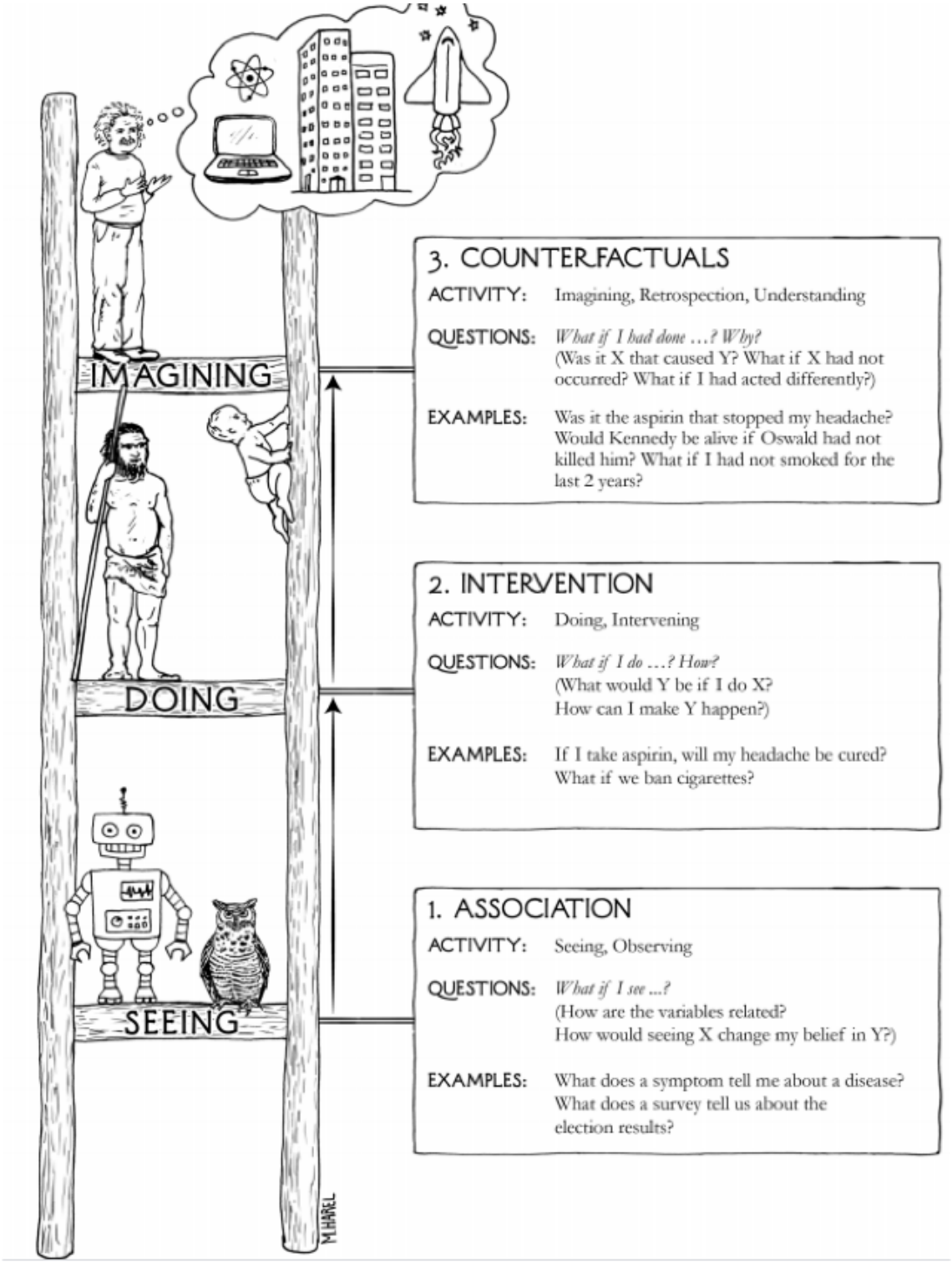


http://bayes.cs.ucla.edu/WHY/



Epilogue: The Art and Science of Cause and Effect

http://bayes.cs.ucla.edu/BOOK-2K/



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?* (How are the variables related? How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

If I hadn't eaten chocolate would I have this headache?

Does eating chocolate cause a headache?

Does eating chocolate go with headaches?

Statistics Answers “Association” Questions

- Does eating chocolate go together with headaches?
  - Can be answered with probability and statistics

- In ML, many ‘prediction’ problems are purely associational
  - E.g. is this transaction a fraud?

“Intervention” Level

- Running a trial can answer a causal question
- Without a trial we need causal assumptions (knowledge)
  - No amount of data avoids this
- Pearl has shown how to translate
  - Questions about causes
  - Assumptions about causal relationships
  - Data

} into probability question

Forms of Causal Model (many and developing)

- Structural equation models
  - Extension of regression
- Bayesian networks
  - Conditional probability relationships

- Possibility of causal discovery
  - From observational data + assumptions
  - Review of Causal Discovery Methods Based on Graphical Models, https://doi.org/10.3389/fgene.2019.00524
  - Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology https://www.nature.com/articles/s41598-020-59669-x

Summary

- Be aware that correlation is not causation
  - Statistician wary of causation

- New perspective on the discovery of causes