Causal Data Science for Business Decision Making

Introduction

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

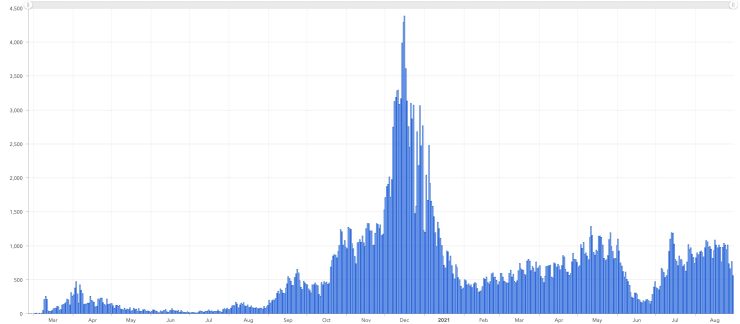
ASSOCIATION AMBA ACCREDITED

CEMS

PIM

COVID-19 & Safety

- To secure a safe learning environment and make sure that we can stay on campus until the end of the semester, let's agree on a few rules
  - Let's keep the windows open to ensure proper ventilation
  - Try to maintain proper distancing
  - If you want to hang out in groups, please do so outside
  - If you're sick, please stay at home (and get tested)



A few words about myself...

- Assistant professor at Strategy & Innovation (SI)
- Joined in September 2021
  - Before 3 years in the Netherlands at Maastricht University
- M.Sc. in economics from Mannheim University and Ph.D. in business economics from KU Leuven
- Research interests:
  - Innovation
  - Science, technology & innovation policy
  - Causal inference & applied econometrics
- Associate editor at the Journal of Causal Inference
- Teaching: causal inference, digital economics, industrial organization, innovation strategy

Why this course?

- Causal data science is becoming a more and more important topic in industry
  - Not least because of the "Book of Why" which was published in 2018
- Standard econometrics and statistics courses teach causal inference in a very process-oriented way
- A conceptual framework for causality is often missing
- On the other hand, data science and machine learning courses teach data analytics as something purely data-driven
- Importance of theory and background knowledge for causal inference
  - Bring together people skilled in data science techniques and business domain experts for effective business decision making
  - This course is meant to be a step in this direction

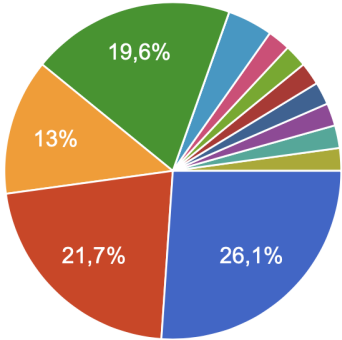
What kind of course is this?

- A little bit of everything
  - Management & Strategy
  - Economics (& Finance)
  - Computer science and AI
  - Statistics
  - Mathematics
  - Philosophy of science
- Focus lies on conceptual understanding of causality and how to infer it from data
- This will involve some statistics and math, but is also **no rocket science**...
- We will see many examples and cases where these concepts are very relevant for decision-making and strategy formulation

You are quite a heterogeneous group...

What is your main background of studies?

46 Antworten

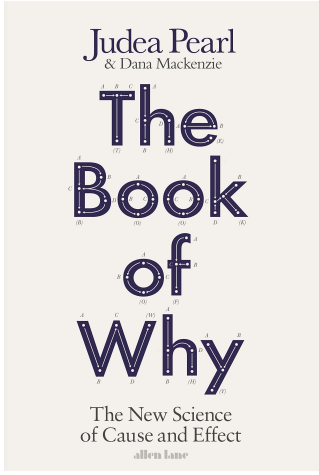


Economics
Management
Marketing
Finance
Sociology
Data Science
IT/IS and Economics
American studies

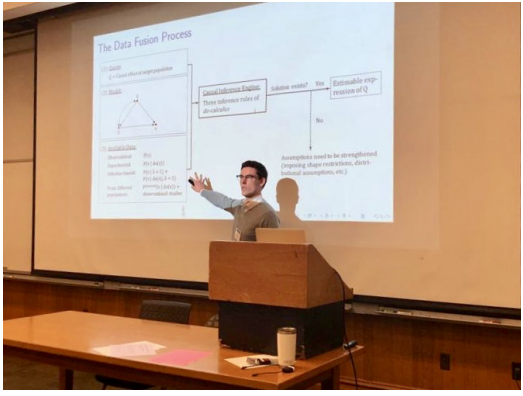
1/2

Course readings

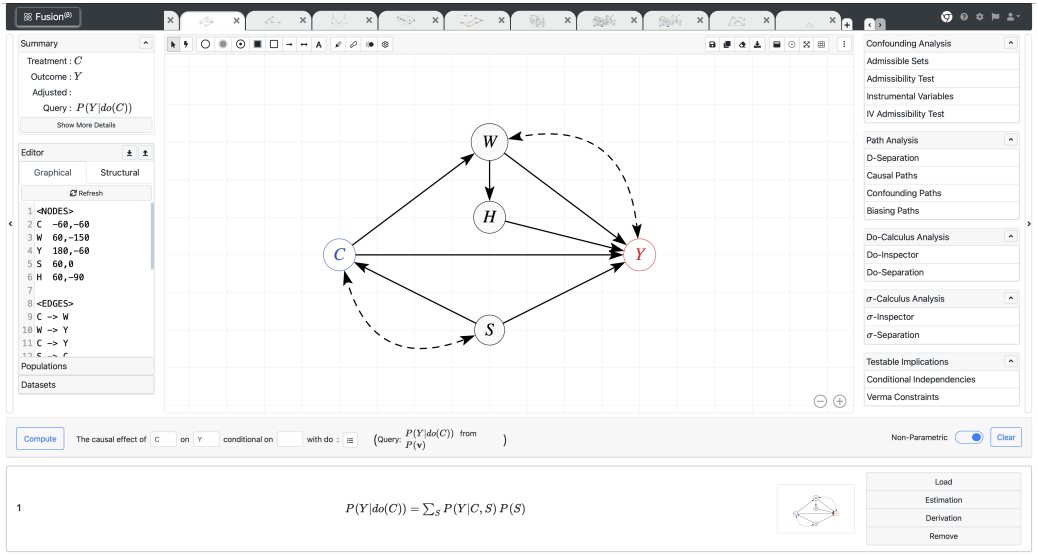
- Mandatory readings
  - The Book of Why (Judea Pearl & Dana Mackenzie)
  - Accompanying papers
  - Blog posts and online material, if it fits
- Additional references
  - Additional sources to dive deeper into a specific topic (and find relevant references there)
  - Classic papers in that area
  - "Causal Inference in Statistics – A Primer" (technical compendium to BoW, full ebook access at CBS library)



WHY-19, AAAI Spring Symposium, Stanford, CA



Causal Fusion



Examination

- Individual written product (**max.** 15 pages)
- Two weeks in December
- Conceptual "essay-style" piece
- Focus on highlighting the relevance of the concepts discussed in course for strategic and business decision making
- Software exercises with Fusion will be relevant for the exam
  - Problem set as preparation (part of self-paced module in week 37)
- More info to come...

Writing Challenge

- Work in teams of 3–5 students
- Write a 800–1500 words blog post about a topic of your choice
  - E.g., pick out a technique or topic we have discussed in class and illustrate its relevance for industry
  - Try to find practically relevant examples
  - Format similar to articles on `www.medium.com` or `www.towardsdatascience.com`
- Feedback opportunity for you!!
- The best three submissions will be published on `www.causalscience.org`
- Deadline: November 28, 2021

Motivating Example: How to Estimate the Gender Pay Gap?

- The New York Times reported in March 2019:
  - *"When Google conducted a study recently to determine whether the company was underpaying women and members of minority groups, it found, to the surprise of just about everyone, that men were paid less money than women* for doing similar work.*"*

    https://www.nytimes.com/2019/03/04/technology/google-gender-pay-gap.html
- The study led Google to increase the pay of its male employees to fight this blatant discrimination of men
- What's going on here? Wasn't Google just recently accused of discriminating against women, not men?
  - *"Department of Labor claims that Google systematically underpays its female employees"*

    https://www.theverge.com/2017/4/8/15229688/department-of-labor-google-gender-pay-gap

Simpson's Paradox

- Suppose we collected data on wages payed to 100 women and 100 men in company X. We observe the following distribution of average monthly salaries for women and men in management and non-management positions (case numbers in parentheses). And our goal is to estimate the magnitude of the gender pay gap in company X. How should we tackle this problem?

	Female	Male
Non-management	\$3163.30 (87)	\$3015.18 (59)
Management	\$5592.44 (13)	\$5319.82 (41)

Simpson's Paradox (II)

- On average, women earn less in this example

$$\left(\frac{87}{100}\cdot \$3163.30 + \frac{13}{100}\cdot \$5592.44\right) - \left(\frac{59}{100}\cdot \$3015.18 + \frac{41}{100}\cdot \$5319.82\right)$$
$$\approx -\$481$$

- But in each subcategory women actually have higher salaries?
  - Non-management: $\$3163.30 - \$3015.18 = \$148.12$
  - Management: $\$5592.44 - \$5319.82 = \$272.62$
- Conditioning on job position gives adjusted gender pay gap

$$\frac{87+59}{200}\cdot \$148.12 + \frac{13+41}{200}\cdot \$272.62 \approx \$181.74$$

- Which estimate gives us a more accurate picture of the gender pay gap?

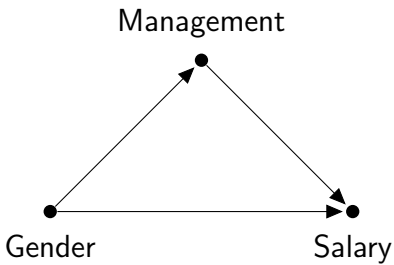
Simpson's Paradox (III)

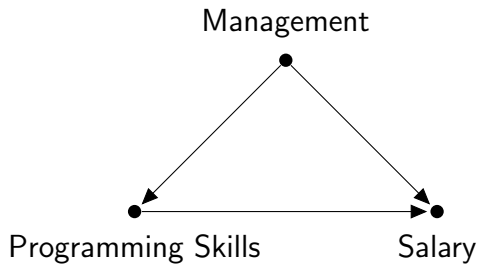
- The phenomenon that a statistical association, which holds in a population, can be reversed in every subpopulation is named after the British statistician Edward Simpson
- Simpson's paradox well-known, for example, in epidemiology and labor economics
- Here, the unadjusted gender pay ($-\$481$) gap gives the right answer
- But what about this example?

	Programming Skills	No Programming Skills
Non-management	\$3163.30 (87)	\$3015.18 (59)
Management	\$5592.44 (13)	\$5319.82 (41)

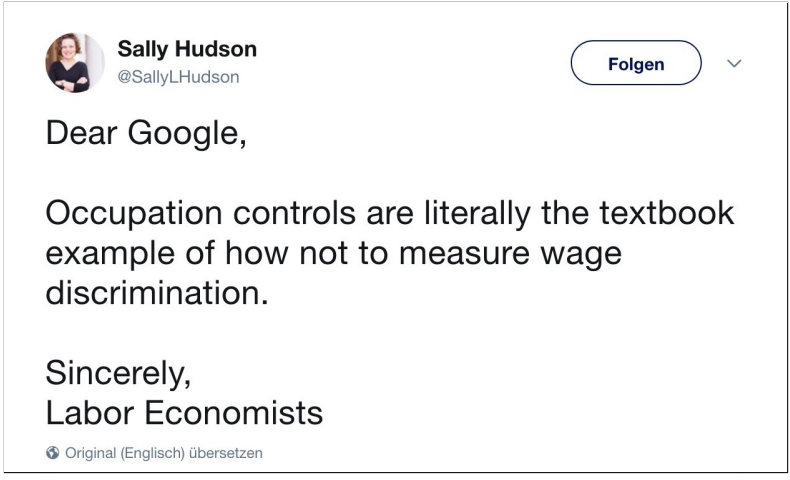
Simpson's Paradox (IV)

- Here we would correctly infer that people with programming skills earn more on average ($181.74). What is the difference between the two examples?





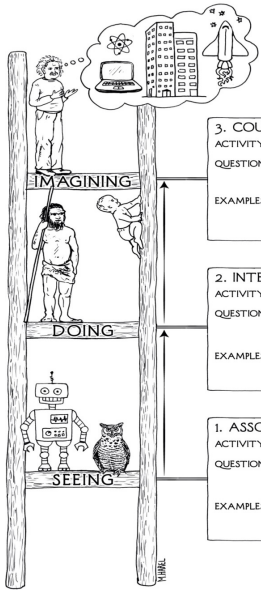
Simpson's Paradox (V)



Simpson's Paradox (VI)

- Statistics alone doesn't help us to answer this question
- Note that the joint distribution of salaries is the same in both cases
- Both problems are thus identical from a statistical point of view
- Instead, we need to make causal assumptions in order to come to a conclusion here
  - Gender affects both a person's salary level and job position
  - Whereas, programming skills increase salaries but persons in high-ranking positions usually have less of it
- After the course you will know how to incorporate this kind of causal knowledge in your analysis in order to solve all sorts of practical problems of causal inference

The Ladder of Causation



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?

Simpson's Paradox and Covid-19 Vaccination

Age	Population (%)		Severe Cases (per 100k)		Efficacy
	Not Vax	Fully Vax	Not Vax	Fully Vax	
All ages	18.2%	78.7%	16.4	5.3	67.5%
<50	23.3%	73.0%	3.9	0.3	91.8%
>50	7.9%	90.4%	91.9	13.6	85.2%

- Vaccine effectiveness defined as $1 - V/N$ (e.g., $1 - 5.3/16.4 = 0.675$)
- Vaccine effectiveness is higher in every age group than in the general population. How can that be?
- Vaccination status and risk of severe disease are systematically higher in the older age group $\Rightarrow$ Simpson's Paradox (full story here)
- Lesson: **Get vaccinated!!**

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk