Causal Data Science with Directed Acyclic Graphs (DAGs)

Section 1: Introduction

Dr. Paul Hünermund

Online Course at Udemy.com

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?

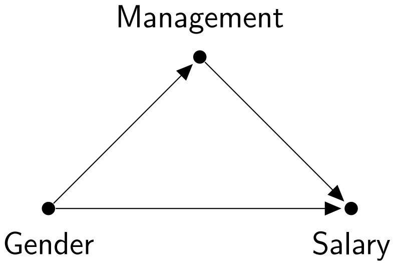
Simpson's Paradox (III)

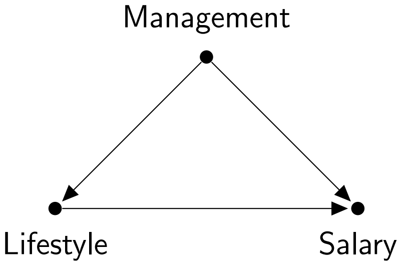
- The phenomenon that a statistical association, which holds in a population, can be reversed in every subpopulation is named after the British statistician Edward Simpson
- Simpson's paradox well-known, for example, in epidemiology and labor economics
- Here, the unadjusted gender pay (−$481) gap gives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

	Healthy Lifestyle	Unhealthy Lifestyle
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 a healthy lifestyle earn more on average ($181.74). What is the difference between the two examples?





Simpson’s Paradox (V)

- 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, life style affects salaries, but is itself affected by a person’s job position
- 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

Course Outline

Section 1: Introduction

Section 2: Structural Causal Models, Interventions, and Graphs

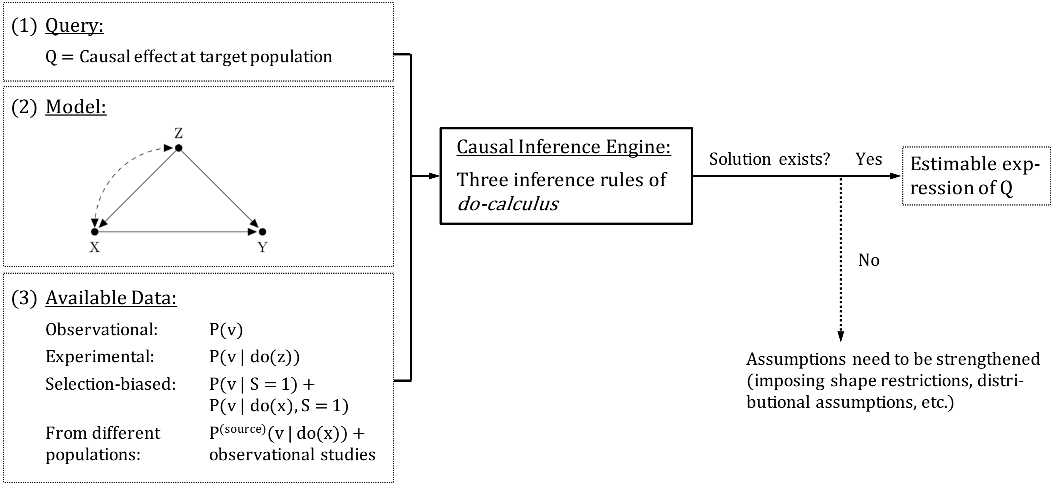
Section 3: Causal Discovery

Section 4: Confounding Bias and Surrogate Experiments

Section 5: Recovering from Selection Bias

Section 6: Transportability of Causal Knowledge Across Domains

The Causal Data Science Process



(1) Query:

Q = Causal effect at target population

(2) Model:

(3) Available Data:

Observational:	P(v)
Experimental:	P(v | do(z))
Selection-biased:	P(v | S = 1) + P(v | do(x), S = 1)
From different populations:	$P^{(source)}$(v | do(x)) + observational studies

Causal Inference Engine:

Three inference rules of *do-calculus*

Solution exists? Yes

Estimable expression of Q

No

Assumptions need to be strengthened (imposing shape restrictions, distributional assumptions, etc.)