Introduction to Causal Inference

Evgeniy Riabenko
Facebook, Core Data Science
riabenko.e@gmail.com

Machine Learning in High Energy Physics Summer School
June 6, 2019

Why bother?

Predictive models are great, why do we need causal inference?

- in real life today's train could differ from tomorrow's test
- especially if we want to act on the results of the predictions!
- causal mechanisms are more stable than correlations

What is causality?

Lewis D. (1973) *Causation*. The journal of philosophy: 556-567: causation is "something that makes a difference, and the difference it makes must be a difference from what would have happened without it".

The "interventionis" definition: T causes Y iff changing T leads to a change in Y, *keeping everything else constant*.
The causal effect is the magnitude by which Y is changed by a unit change in T.

Keeping everything else constant: parallel, counterfactual reality.
Causal questions are weird!

The Three Layer Causal Hierarchy

Level	Typical Activity	Typical Question	Examples
1. Association $\mathbf{P}(y \mid x)$	Seeing	What is?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $\mathbf{P}(y \mid do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactual $\mathbf{P}(y_x \mid x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? What I had not been smoking the past 2 years?

Pearl J. *Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution*. arXiv:1801.04016v1, 2018

Potential outcomes framework

Y_{1i} — the outcome for unit i that would be observed in condition $T = 1$ (“treatment“),
Y_{0i} — the outcome that would be observed, all else held constant, in condition $T = 0$ (“control”).
Causal effect of treatment on Y:

$$\tau_i = Y_{1i} - Y_{0i}$$

.

Fundamental problem of causal inference: only one outcome is observed for each unit
$\Rightarrow$ causal effect cannot be measured.
Solution — estimate something else, e.g. average causal effect:

$$\text{ATE} = \mathbb{E}(\tau_i) = \mathbb{E}\left(Y_{1i} - Y_{0i}\right) = \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{0i})$$

(population) average treatment effect.

Randomized experiment

- A large population of experimental units
- Treatment T with support $\{0, 1\}$
- Each unit in $i \in U$ has potential outcomes Y_{0i}, Y_{1i}
- Population average treatment effect:

$$\mathrm{ATE} = \mathbb{E}\left(Y_1 - Y_0\right)$$

- Random sample of size N from the population
- Sample average treatment effect — an estimate of ATE:

$$\mathrm{SATE} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_{1i} - Y_{0i}\right)$$

- Randomly assign N_1 units to treatment ($T_i = 1$) and $N_0 = N - N_1$ to control ($T_i = 0$)
- Observe $Y_i = T_i Y_{1i} + \left(1 - T_i\right) Y_{0i}$
- Because treatment assignment is random,

$$\widehat{\mathrm{SATE}} = \frac{1}{N_1} \sum_{i:\ T_i=1} Y_i - \frac{1}{N_0} \sum_{j:\ T_j=1} Y_j = \bar{Y}_1 - \bar{Y}_0$$

is an unbiased estimate of SATE (and ATE)

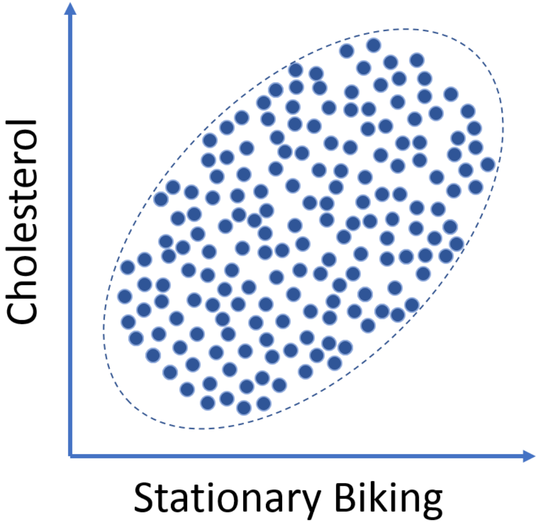
Randomized experiment

Experiment is not always feasible:

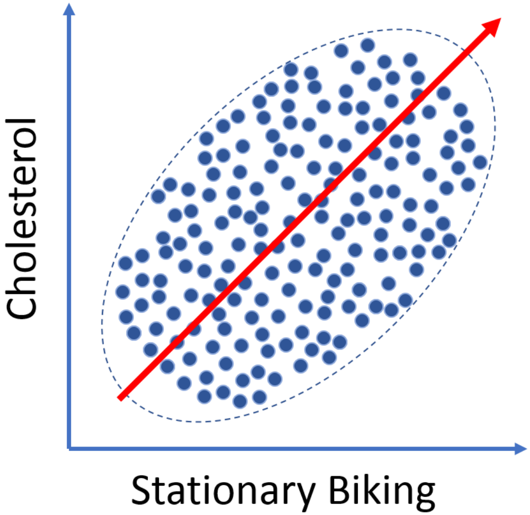
- thunderstorms $\to$ forest fires — we cannot manipulate the treatment
- TV violence $\to$ cruelty — treatment is difficult to fix, response is difficult to measure in a lab
- alcohol consumption $\to$ performance in school — unethical

In such cases we have to resort to observational data.

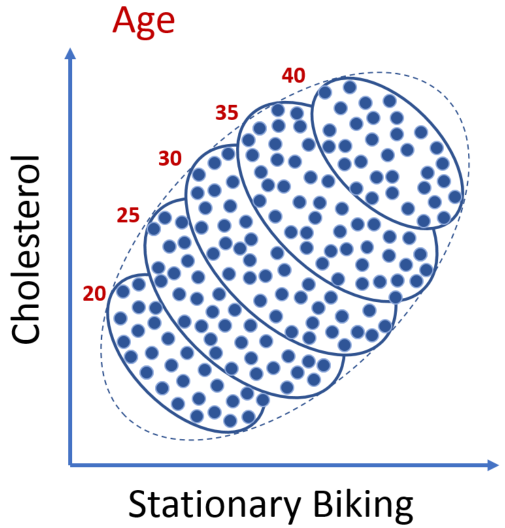
Cholesterol and exercise



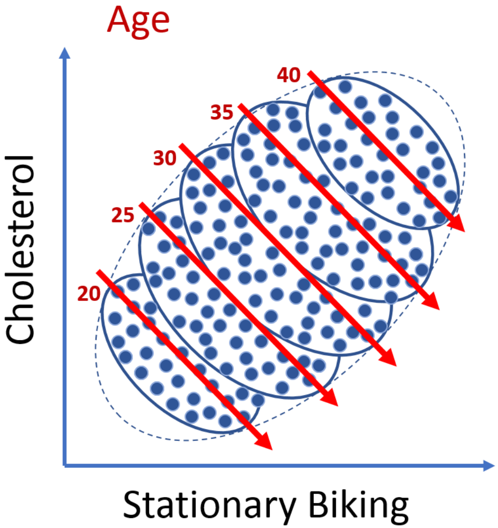
Cholesterol and exercise



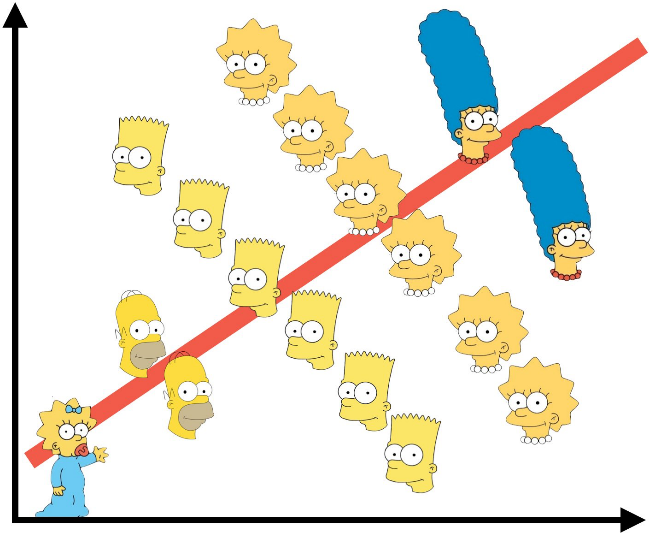
Cholesterol and exercise



Cholesterol and exercise



Simpson's paradox



Simpson's paradox

Example 1:

Σ	Recovered	Not recovered	Recovery rate
Drug	273	77	78%
Placebo	289	61	83%

Placebo is 5% more effective

Men	Recovered	Not recovered	Recovery rate
Drug	81	6	93%
Placebo	234	36	87%

Drug is 5% more effective

Women	Recovered	Not recovered	Recovery rate
Drug	192	71	73%
Placebo	55	25	69%

Drug is 4% more effective

Simpson's paradox

Does the drug increases chance to recover compared to placebo?

Conclusion 1: drug is 5% worse than placebo.

$$\widehat{\text{ATE}} = \mathbf{P}(recovery\,|drug\,) - \mathbf{P}(recovery\,|placebo\,)$$

Conclusion 2: drug is 4.51% better than placebo (assuming patients are 49% women).

$$\widehat{\text{ATE}} = \sum_{sex_i} \left(\mathbf{P}(recovery\,|drug, sex_i\,) - \mathbf{P}(recovery\,|placebo, sex_i\,)\right) \mathbf{P}(sex_i)$$

Which one is correct?
What would happen if we intervene?

Simpson's paradox

Example 2:

Σ	Recovered	Not recovered	Recovery rate	
Drug	273	77	78%	Placebo is 5%
Placebo	289	61	83%	more effective

Low pressure by the end of treatment	Recovered	Not recovered	Recovery rate	
Drug	81	6	93%	Drug is 5%
Placebo	234	36	87%	more effective

High pressure by the end of treatment	Recovered	Not recovered	Recovery rate	
Drug	192	71	73%	Drug is 4%
Placebo	55	25	69%	more effective

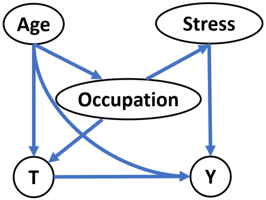
Simpson's paradox

In example 1, conclusion 2 is correct, in example 2 — conclusion 1.

Everything depends on the directions of causal relationships between a feature determining subgroups and the rest of features.

Causal graphs

Causal relationships could be represented on graphs where variables are vertices and directed edges are causal relationships.



Edges — direct causes, directed paths — indirect causes.
Graph encodes all causal assumptions:

- occupation does affect outcome Y
- age does not affect stress
- stress does not affect occupation
- treatment does not affect stress
- ...

Elements of causal graph

$$A \to B \to C \text{ — } \textbf{chain}$$

B — mediator

Example:

- A — school budget
- B — average score of graduates
- C — proportion of students admitted to college

Properties:

1. A and B, B and C are dependent:
   $\exists a, b : \mathbf{P}(B = b \,|\, A = a\,) \neq \mathbf{P}(B = b)$
   $\exists b, c : \mathbf{P}(C = c \,|\, B = b\,) \neq \mathbf{P}(C = c)$
2. C and A are likely dependent
3. $C \perp A|B$ conditionally independent: $\forall a, b, c$

$$\mathbf{P}(C = c \,|\, A = a, B = b\,) = \mathbf{P}(C = c \,|\, B = b\,)$$

   (if B is fixed, then A and C are independent)

Elements of causal graph

$$B \leftarrow A \rightarrow C$$ — **fork**

A — confounder

Example:

- A — average daily temperature
- B — ice cream sales
- C — number of violent crimes per day

Properties:

1. A and B, A and C are dependent
2. B and C are likely dependent
3. $B \perp C | A$ conditionally independent

Elements of causal graph

$$B \to A \leftarrow C$$ — **collider**

A — also collider

Example (Monty Hall problem):

- A — choice of the game host
- B — choice of the player
- C — position of the prize

Properties:

1. B and A, C and A are dependent
2. B and C are independent
3. $B \not\perp C | A$ conditionally dependent

Intervention

We need to use observational data to estimate the effect of **intervention**: what would happen with Y if we set the value of T equal to t?
Notation: $do(T = t)$.

Potential outcomes are outcomes under intervention:

$$Y_{1i} = Y_i \,|do(T = 1)\;, Y_{0i} = Y_i \,|do(T = 0)$$

Hence, causal effect could be represented through intervention:

$$\mathrm{ATE} = \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{0i}) = \mathbb{E}\left(Y_i \,|do(T = 1)\,\right) - \mathbb{E}\left(Y_i \,|do(T = 0)\,\right)$$

Intervention

$$\begin{tikzcd}
 & X \text{ (gender)} \arrow[dl] \arrow[dr] & \\
T \text{ (treatment)} \arrow[rr] & & Y \text{ (outcome)}
\end{tikzcd}$$

Drug effect in terms of interventions:

$$\begin{aligned}
\mathrm{ATE} = & \mathbf{P}(Y = \text{recovery} \,|do\,(T = \text{drug})\,) - \\
& - \mathbf{P}(Y = \text{recovery} \,|do\,(T = \text{placebo})\,)\,.
\end{aligned}$$

Graph surgery

Graph surgery — removal of all edges directed into treatment variable X.

Example 1, original graph G:

$$\begin{tikzcd}
& X \text{ (gender)} \arrow[dl] \arrow[dr] & \\
T \text{ (treatment)} \arrow[rr] & & Y \text{ (outcome)}
\end{tikzcd}$$

Modified graph G_m:

$$\begin{tikzcd}
& X \text{ (gender)} \arrow[dr] & \\
T \text{ (treatment)} \arrow[rr] & & Y \text{ (outcome)}
\end{tikzcd}$$

$$\mathbf{P}(Y = y \,|do\,(X = x)\,) = \mathbf{P}_m(\,Y = y|\,X = x)$$

Graph surgery

In the modified graph:

$$\mathbf{P}_m(X = x) = \mathbf{P}(X = x)\,,$$
$$\mathbf{P}_m(\,Y = y|\,T = t, X = x) = \mathbf{P}(Y = y\,|T = t, X = x\,)\,,$$

because the edges pointing to T and Y did not change;

$$\mathbf{P}_m(\,X = x|\,T = t) = \mathbf{P}_m(X = x)\,,$$

because T and X are independent;

$$\begin{aligned}\mathbf{P}(Y = y\,|do\,(T = t)\,) &= \mathbf{P}_m(\,Y = y|\,T = t) = \\ &= \sum_x \mathbf{P}_m(Y = y|\,T = t, X = x)\,\mathbf{P}_m(\,X = x|\,T = t) = \\ &= \sum_x \mathbf{P}_m(Y = y|\,T = t, X = x)\,\mathbf{P}_m(X = x) = \\ &= \sum_x \mathbf{P}(Y = y\,|T = t, X = x\,)\,\mathbf{P}(X = x)\,.\end{aligned}$$

Graph surgery

In example 1:
$\mathbf{P}(Y = \text{recovery} \,|do\,(T = \text{drug})\,) = 0.832$,
$\mathbf{P}(Y = \text{recovery} \,|do\,(T = \text{placebo})\,) = 0.7818$
$\Rightarrow \text{ATE} = 0.05.$

In example 2 $G = G_m$:

$$\begin{tikzcd}
& X \text{ (pressure)} \arrow[dr] & \\
T \text{ (treatment)} \arrow[ur] \arrow[rr] & & Y \text{ (outcome)}
\end{tikzcd}$$

Therefore,
$\mathbf{P}(Y = y \,|do\,(T = t)\,) = \mathbf{P}_m(Y = y|\, T = t) = \mathbf{P}(Y = y \,|T = t\,)$
$\mathbf{P}(Y = \text{recovery} \,|do\,(T = \text{drug})\,) = 0.78$,
$\mathbf{P}(Y = \text{recovery} \,|do\,(T = \text{placebo})\,) = 0.83$
$\Rightarrow \text{ATE} = -0.05.$

Adjustment formula

Adjustment formula allows to calculate the effect of an intervention by conditioning on the vertices of X:

$$\mathbf{P}(Y = y\,|do\,(T = t)\,) = \sum_x \mathbf{P}(Y = y\,|T = t, X = x\,)\,\mathbf{P}(X = x)\,.$$

What is X?

Causal effect formula:

$$\mathbf{P}(Y = y\,|do\,(T = t)\,) = \sum_x \mathbf{P}(Y = y\,|T = t, PA = x\,)\,\mathbf{P}(PA = x)\,,$$

where PA — parents of T.

Assumptions of conditioning on X

Ignorability (no unmeasured confounders)
Under random experiments, $T \perp X$ for both observed and unobserved covariates.
But conditioning and related techniques can only construct $T \perp X$ for observed covariates.
So assume that after conditioning on observed covariates, any unmeasured covariates are irrelevant:

$$\mathbf{P}(Y_T \,|X\,) = \mathbf{P}(Y_T \,|X, T\,)$$

Stable Unit Treatment Value (SUTVA) (no spillover)
The effect of treatment on an individual is independent of whether or not others are treated:

$$\mathbf{P}(Y_i \,|do(T_i, T_j)\,) = \mathbf{P}(Y_i \,|do(T_i)\,)$$

Overlap (common support)
There should be overlap on observed covariates between treated and untreated individuals:

$$0 < \mathbf{P}(T = 1 \,|X = x\,) < 1$$

Unknown parents

$$\begin{tikzcd}
S \text{ (socioeconomical status)} \arrow[r] \arrow[d] & W \text{ (weight)} \arrow[d] \\
T \text{ (treatment)} \arrow[r] & Y \text{ (outcome)}
\end{tikzcd}$$

Socioeconomical status — unobservable variable; how can we estimate the effect of intervention on T?

More definitions

Path — a sequence of vertices where each vertex is connected to the next one with an edge.
Directed path — a path where all edges have the same direction.
Backdoor path from A to B starts with $A \leftarrow$ and ends with $\rightarrow B$.

A path P is **blocked** by variable X, if:

1. P contains $A \rightarrow B \rightarrow C$, $A \leftarrow B \rightarrow C$, $B \in X$
2. P contains $A \rightarrow B \leftarrow C$, $B \notin X$ and all the descendants of $B \notin X$

Backdoor criterion

For an ordered pair of vertices (A, B) in acyclic graph G a set of vertices X satisfies **backdoor criterion**, if it:

- X does not contain the descendants of A
- X blocks all backdoor paths from A to B

If X satisfies backdoor criterion for (T, Y), then

$$\mathbf{P}(Y = y \,|do\,(T = t)\,) = \sum_x \mathbf{P}(Y = y \,|T = t, X = x\,)\,\mathbf{P}(X = x)$$

(backdoor formula).

Backdoor criterion

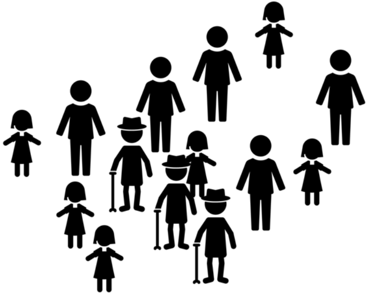
To calculate less conditional probabilities, backdoor formula could be simplified:

$$\mathbf{P}(Y = y \,|do\,(T = t)\,) = \sum_x \mathbf{P}(Y = y \,|T = t, X = x\,)\,\mathbf{P}(X = x) =$$
$$= \sum_x \frac{\mathbf{P}(Y = y, T = t, X = x)}{\mathbf{P}(T = t \,|X = x\,)}$$

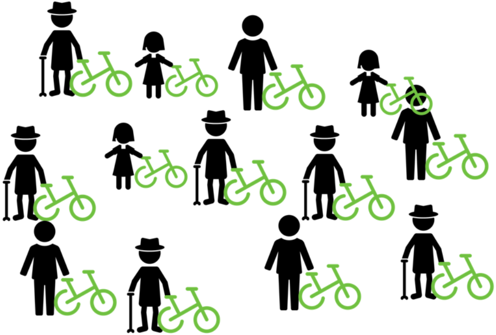
This way

- the method is called **inverse probability weighting**
- denominator $e_i = \mathbf{P}(T = t \,|X = x\,)$ — propensity score.

Biking vs Cholecterol



Avg Cholesterol = 200



Avg Cholesterol = 206

Regression

Model Y as a function of T and X:

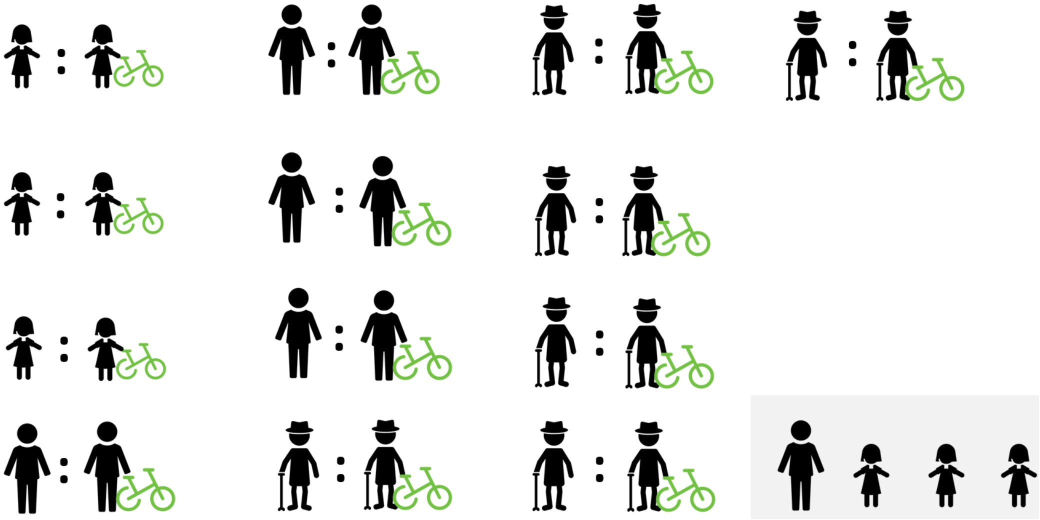
$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k + \alpha T + \varepsilon,$$

i.e., $Cholesterol = \beta_0 + \beta_1 \cdot Age + \alpha \cdot Excercise + \varepsilon$.
$\hat{\alpha}$ — an estimate of the average effect of changing T from 0 to 1, **if** among $X_1, \dots, X_k$ there are:

- all the parents of T, or a set of variables that satisfies backdoor criterion for (T, Y)
- no colliders of T and Y

Also, the model must be true.

Matching



Matching

- Paired individuals provide the counterfactual estimate for each other
- Reduces sample size
- Could be approximate:
  - on distances in X space
  - on propensity scores $e_i = \mathbf{P}(T = 1 \,|\, X = x\,)$

Stratification

Stratification

- Many:many matching
- Stratum sizes — bias-variance tradoeff
- You can stratify on binned propensity scores! But they must be well-calibrated.

Weighting

Propensity scores could be used as weights:

$$\widehat{\mathrm{ATE}} = \frac{1}{N_1} \sum_{i:\ T_i=1} w_i Y_i - \frac{1}{N_0} \sum_{j:\ T_j=1} w_j Y_j,$$

$$w_i = \frac{T}{e_i} + \frac{1-T}{1-e_i}$$

Inverse Probability of Treatment Weighting (IPTW).

- High variance when e_i close to 0 or 1 (could be stabilized heuristically)
- Assumes propensity score model is correctly specified

Doubly robust

Combines models $\hat{Y}_{T=t}$ and propensity scores $\hat{e}$:

$$DR_1 = \begin{cases} \frac{Y}{\hat{e}} - \frac{\hat{Y}_{T=1}(1-\hat{e})}{\hat{e}}, & T = 1, \\ \hat{Y}_{T=1}, & T = 0; \end{cases}$$

$$DR_0 = \begin{cases} \hat{Y}_{T=0}, & T = 1, \\ \frac{Y}{1 \hat{-} e} - \frac{\hat{Y}_{T=1}\hat{e}}{1-\hat{e}}, & T = 0 \end{cases}$$

Causal effect on T — difference between mean DR_1 and DR_0.

- Works if at least one of two is correctly specified
- But if both propensity score or regression are slightly incorrect, may become very biased

Causal analysis simple checks

- Adding random covariates should not change the analysis
- AA-test: randomizing the treatment should turn causal effect into 0
- Subampling should not change the conclusions

References

- theory:
  - Pearl J., Glymour M., Jewell N.P. *Causal Inference in Statistics: A Primer*, 2016
  - Pearl J., Mackenzie D. *The Book of Why: The New Science of Cause and Effect*, 2018
  - Morgan S.L., Winship C. *Counterfactuals and Causal Inference* (2015, 2nd ed)
- good introduction: `https://causalinference.gitlab.io/kdd-tutorial/`
- implementations:
  - `http://www.bnlearn.com/` (R)
  - `https://github.com/microsoft/dowhy` (Python)