Causal Inference

Fadwa Idlahcen

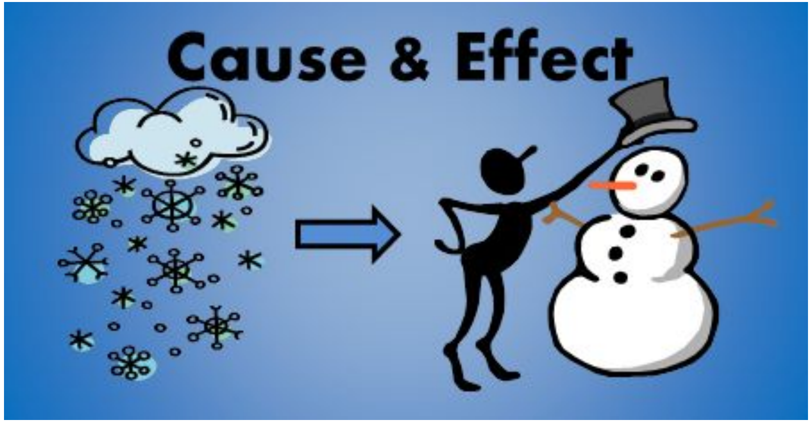
1/Intro - motivation
2/ Potential outcome and Structural Causal Model
3/ Applications of causal inference
4/ ML and causal inference

Intro

Causal inference (CI) aims to draw conclusions from data and correctly predict the causal effect of actions.

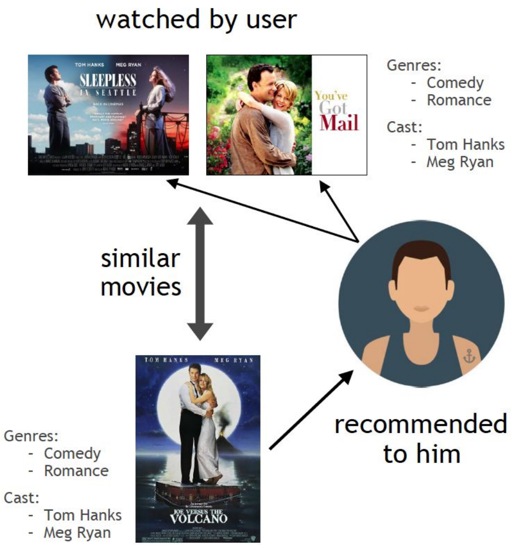
Traditional ML is more towards correlation not causation / Correlation doesn’t imply causation

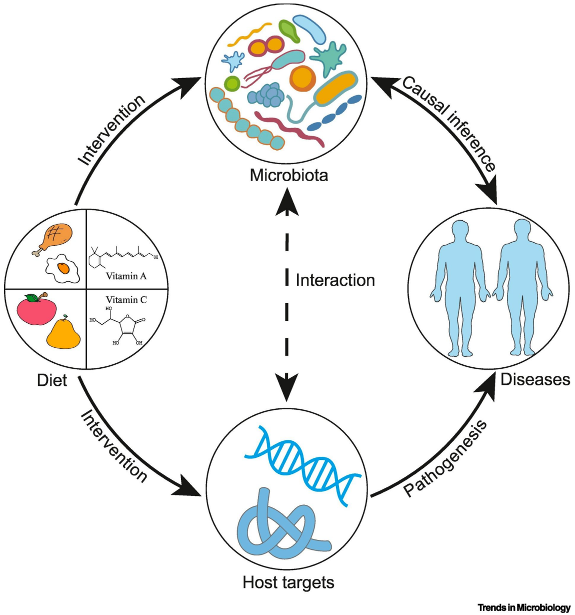
Correlation measures the tendency of two random quantities to move together.



The use of causal inference :

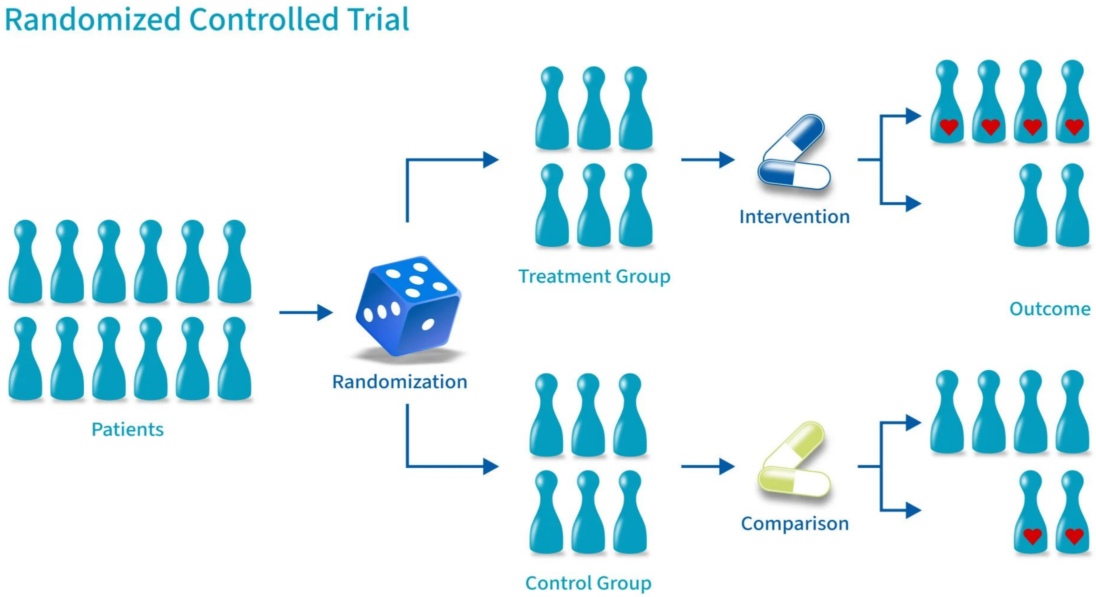
- recommender systems
- social media
- medicine
- education





Observational data

- Large amount of available data
- Less expensive than randomized controlled trials (RCT)

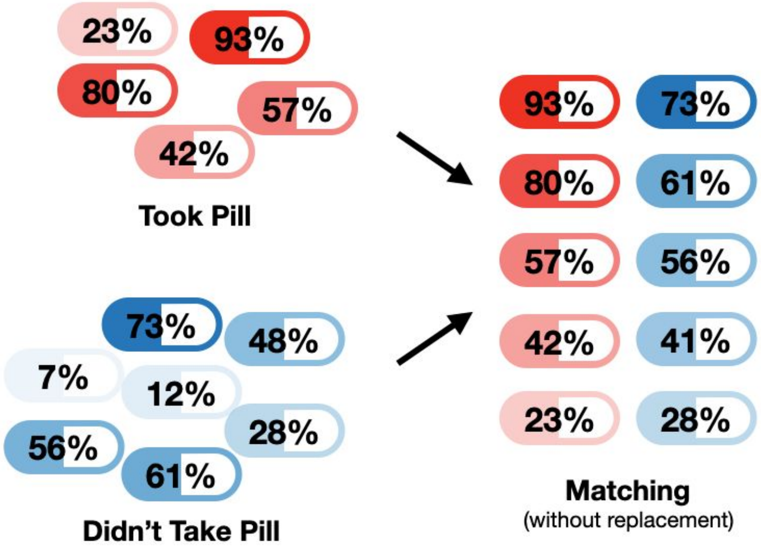


Estimating the causal effect

Counterfactuals : “Would this patient have different results if he received a different medication? “

Comparing results of a treatment in “almost” two identical worlds.

Problem : counterfactuals are not observable.



CI’s frameworks

Tools for formalizing implicit assumptions about causal mechanisms.

Two main frameworks : potential outcome and structural causal model.

Major challenges in causal inference :

- Confounders
- Selection bias
- Simpson’s paradox

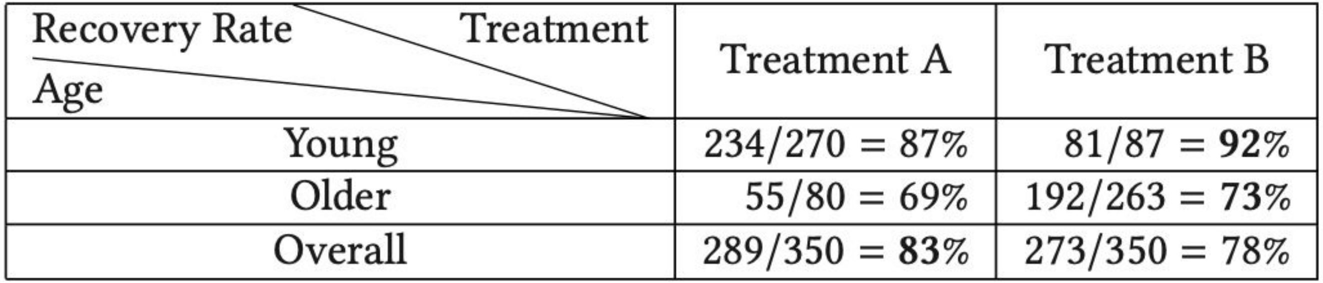
Confounders

Factors affecting both the assignment of the treatment and the outcome.

Creates a spurious effect

Example : **age** in medical treatments

Table 1. An Example to Show the Spurious Effect of Confounder Variable *Age*



Recovery Rate Treatment Age	Treatment A	Treatment B
Young	234/270 = 87%	81/87 = **92%**
Older	55/80 = 69%	192/263 = **73%**
Overall	289/350 = **83%**	273/350 = 78%

Selection bias

The distribution of the observed group is not representative to the group we are interested in.

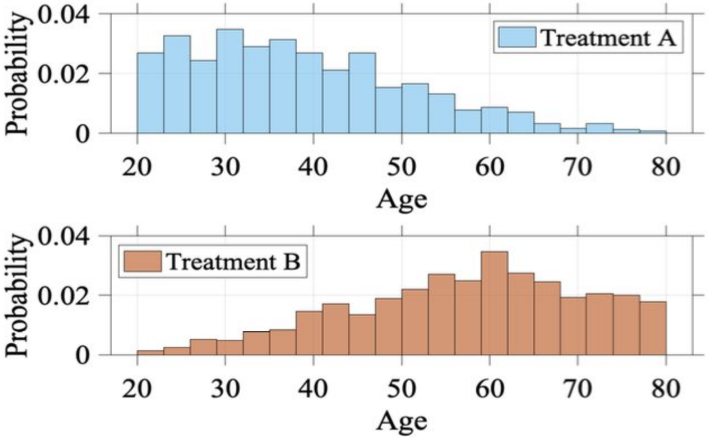


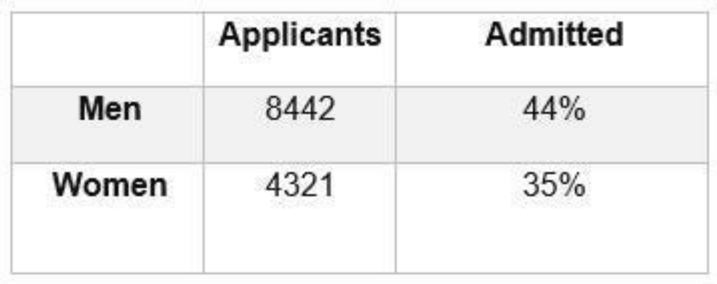
Fig. 1. An example to show the selection bias caused by confounder variable *Age*.

Simpson’s paradox

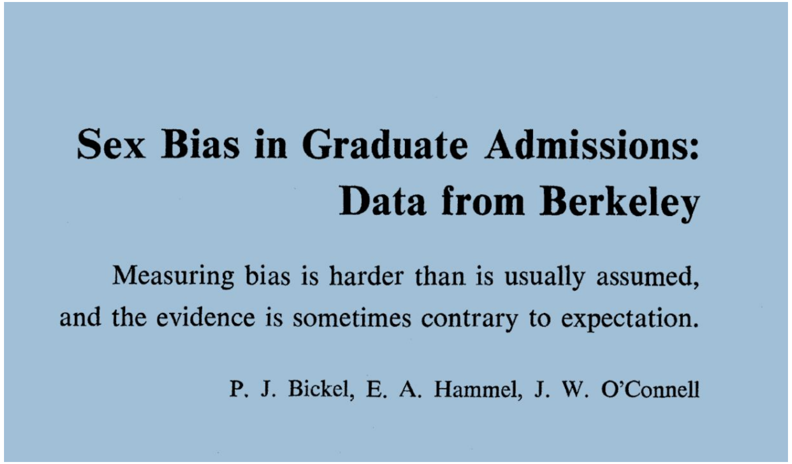
Young and Older patient groups: Medicine B > Medicine A; but when combining the groups, Medicine A is the one with a higher recovery rate.

-Statistical phenomenon described by Edward Hugh Simpson in a technical paper in 1951

- A historical example : UC Berkeley’s suspected gender-bias 1970



	Applicants	Admitted
Men	8442	44%
Women	4321	35%



Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

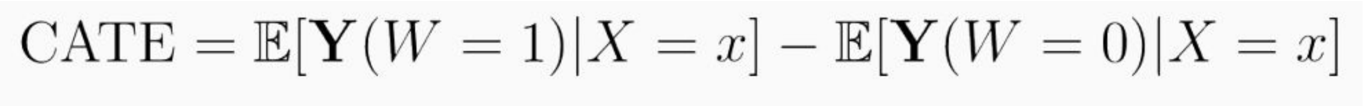
P. J. Bickel, E. A. Hammel, J. W. O’Connell

Potential Outcome

Terminology :

- Unit : physical object, a firm, a patient, an individual person …
- Treatment : action applied to the unit (W=1 treated group; W=0 control group)
- Outcome : results from the treatment/control
- Treatment effect : change of outcome when applying the different treatments on the units.

Estimating the treatment effect via CATE, ATE …



$$\mathrm{CATE} = \mathbb{E}[\mathbf{Y}(W = 1)|X = x] - \mathbb{E}[\mathbf{Y}(W = 0)|X = x]$$

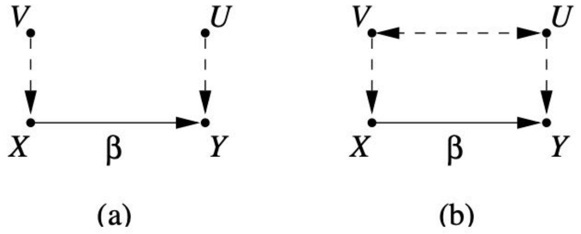
Structural Causal Models

Introduced by Judea Pearl but actually started with Wright (1921)

Notation: u are all the factors, y is the severity of the symptom and x is the severity of the disease

+ path diagram + nonlinear system of equations

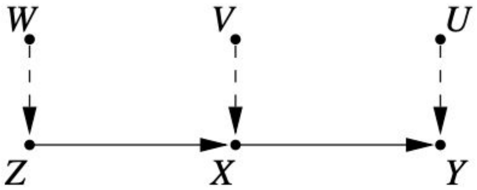
$$x = v$$
$$y = \beta x + u$$

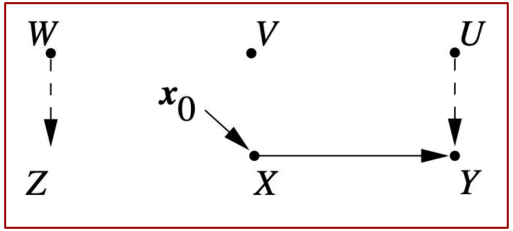


: A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

Structural Causal Models

Do operator : to represent interventions and identify the causal effect





$$\begin{aligned} z &= f_Z(w) \\ x &= f_X(z, v) \\ y &= f_Y(x, u) \end{aligned}$$

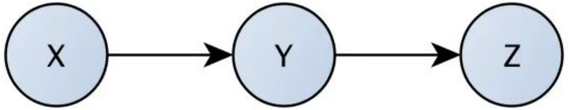
$$\boxed{\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x, u) \end{aligned}}$$

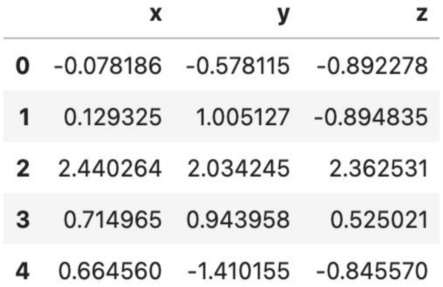
Why is identifiability important?

$$P(Y = y | do(X = x)) = \sum_t P(y|t, x) P(t)$$

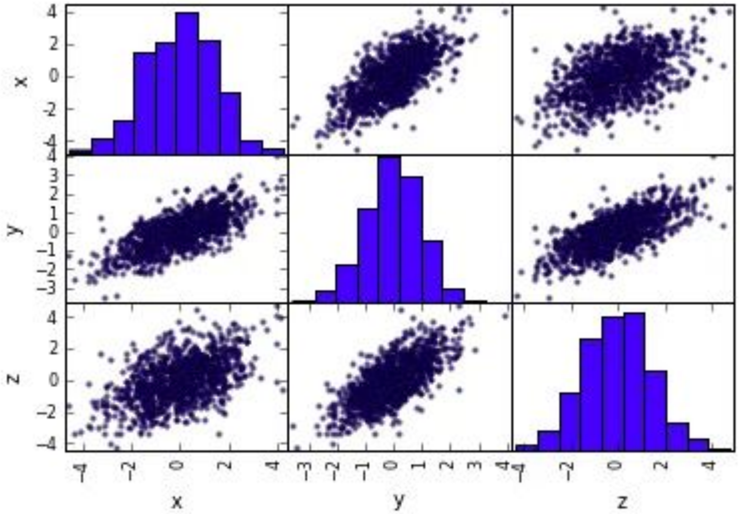
A general identification theorem is the following : “*A sufficient condition for identifying the causal effect P(y|do(x)) is that every path between X and any of its children traces at least one arrow emanating from a measured variable.*”

Example





	x	y	z
0	-0.078186	-0.578115	-0.892278
1	0.129325	1.005127	-0.894835
2	2.440264	2.034245	2.362531
3	0.714965	0.943958	0.525021
4	0.664560	-1.410155	-0.845570

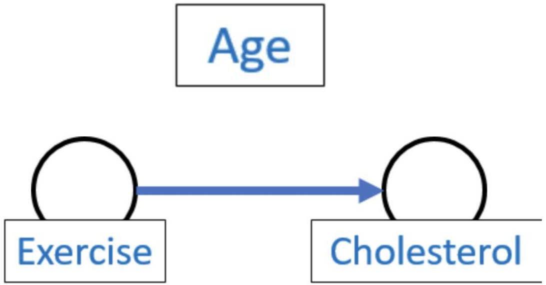


-Regression shows that Z is related to Y, Z is related to X but regressing Z on both Y and X, X's coefficient goes away!

-Direct effect vs indirect effect.

-you have to be very careful what you regress on . But with no graph, you don't what what you should regress on.

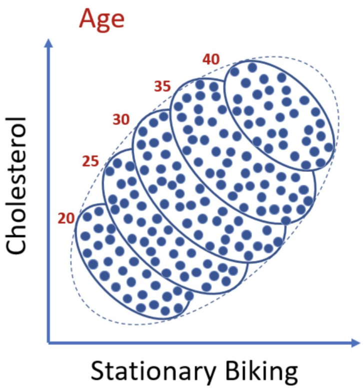
Conditioning

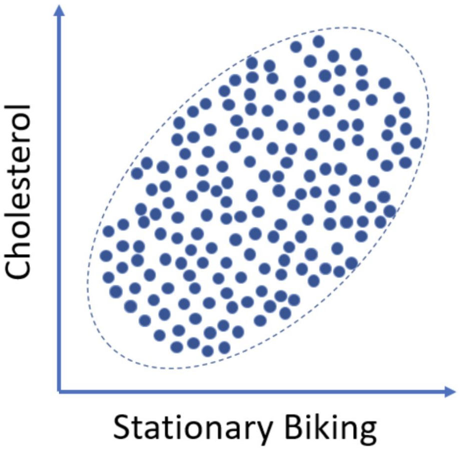


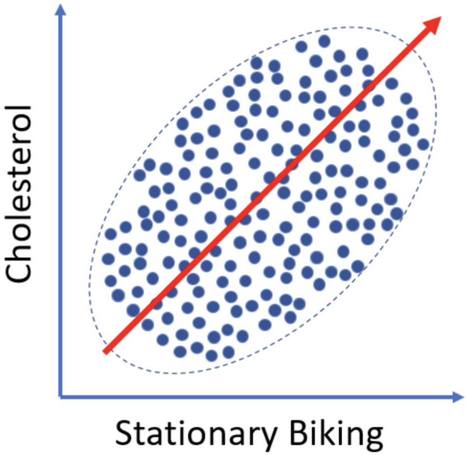
Age

Exercise

Cholesterol







Stationary biking causes cholesterol?

Causal inference assumptions

Assumptions :

- Stable Unit Treatment Value Assumption (SUTVA) :

  The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

- Ignorability :

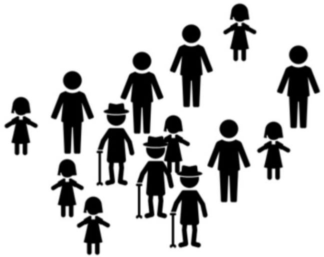
  Given the background variable, X , the treatment assignment W is independent to the potential outcomes

- Positivity :

  For any set of values of X, treatment assignment is not deterministic

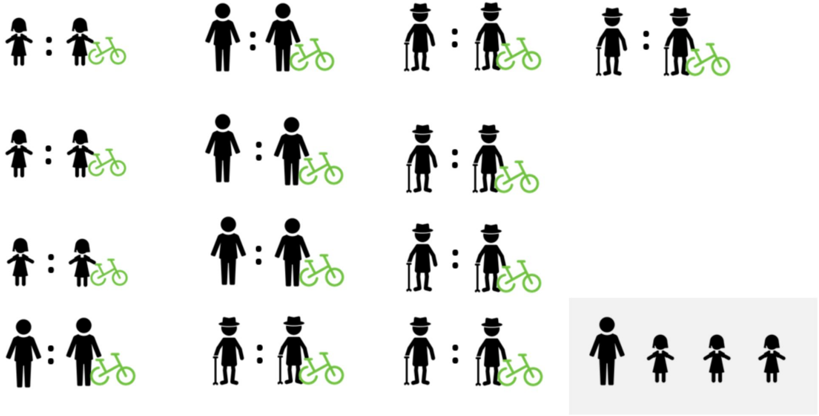
Propensity score based matching

Matching similar units



Avg Cholesterol = 200

Avg Cholesterol = 206

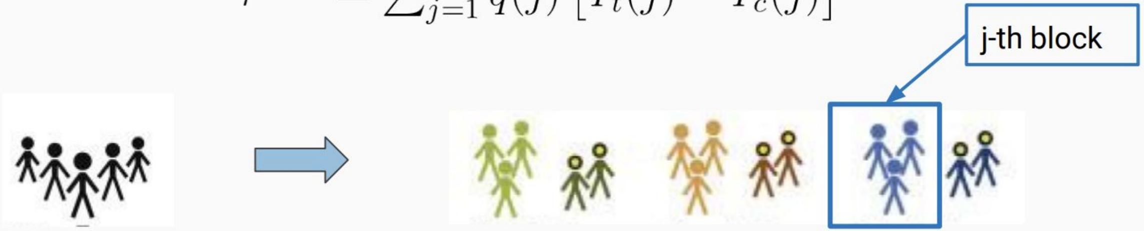


Stratification

-Pairing groups with similar covariates and different assignments.

-Should have enough data (i.e enough control and treated units per group) for each strata

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \left[\bar{Y}_t(j) - \bar{Y}_c(j)\right]$$



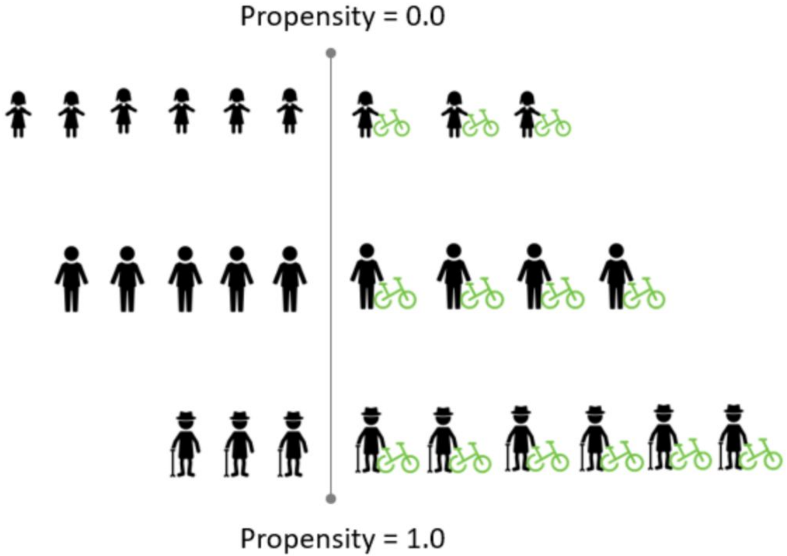
Example

$$ATT = \sum_{s \in strata} \frac{1}{N_{s,T=1}} \left(\bar{Y}_{s,T=1} - \bar{Y}_{s,T=0}\right)$$

where,

$\bar{Y}_{s,T}$ is the average outcome at strata s and treatment status T

And $N_{s,T=1}$ is the number of treated individuals in strata s

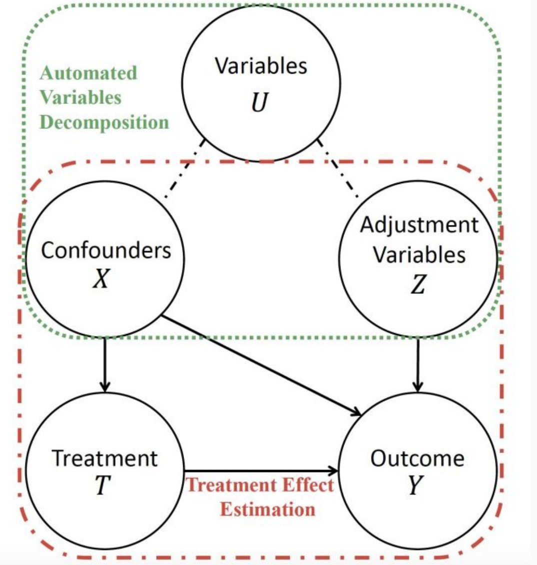


Re-weighting

Creating balanced dataset = breaking the dependence between treatment and covariates

$$ATE = \frac{1}{N_{T=1}} \sum_{i \in treated} w_i Y_i - \frac{1}{N_{T=0}} \sum_{j \in untreated} w_j Y_j$$

Data-Driven Variable Decomposition (D2VD) algorithm distinguishes the confounders and adjustment variables, while eliminating the irrelevant variables.



Real world situations

SUTVA :

- non i.i.d samples : presence of both unobserved confounding and data dependence .

Some solutions : Graph Convolutional Network (for unobserved confounders), segregated graphs (for data dependence), using a classifier instead of regression models (for time series data type), deconfounder (for time series with hidden confounders).

- there exists more than one version for each treatment : e.g with dosage parameters, many version for each treatment will be obtained, a solution is to consider each treatment with its specific dosage as a new treatment.

Ignorability :

In real-world situations, collecting all background variables is not possible.

Some solutions : Variational autoencoders, using instrumental variables that only affect the treatment assignment but not the outcome variable

High dimensional data

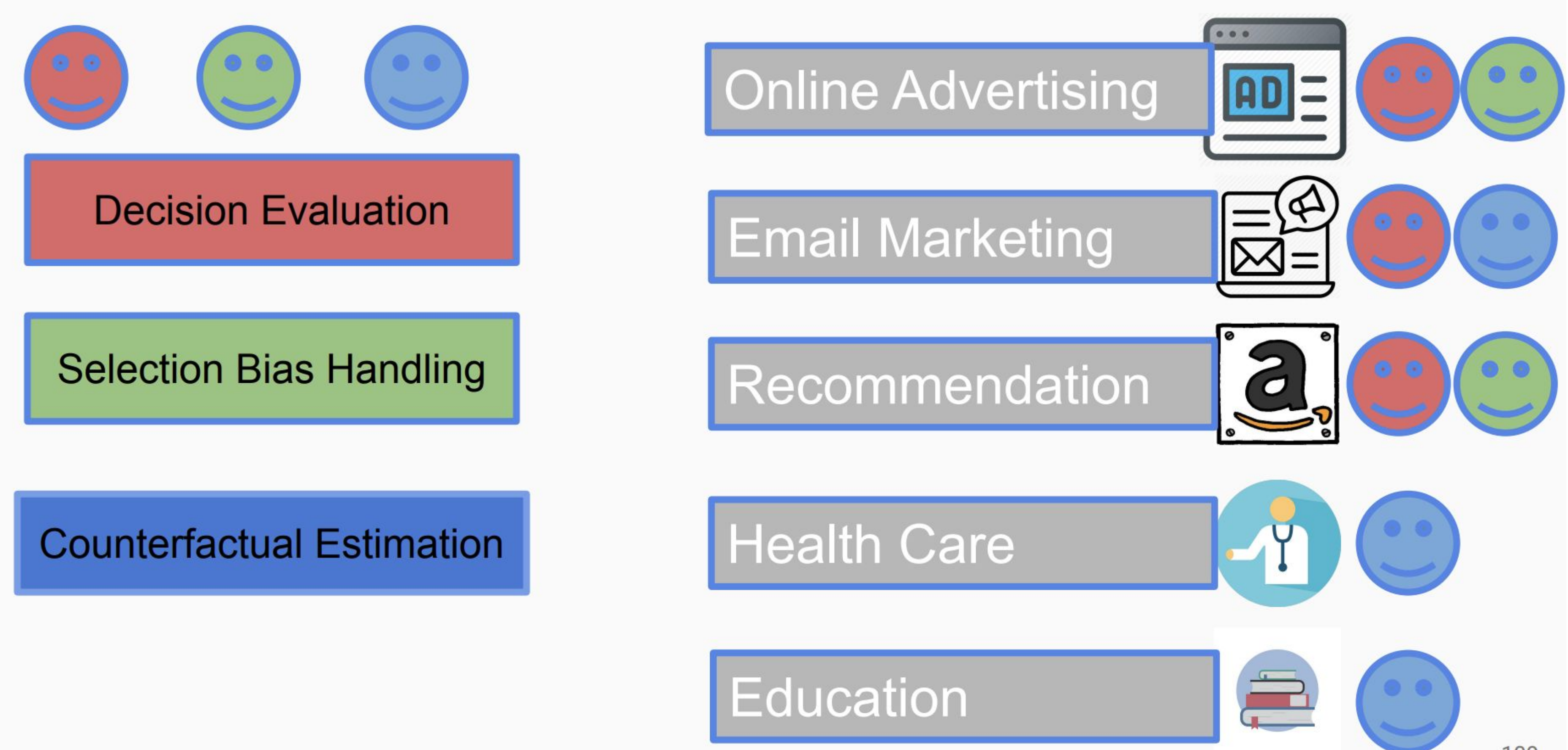
Positivity : it’s hard to satisfy this assumption in high dimensional datasets.

-Data is sparse (e.g not all tests are given to all patients)

-Some solutions : dimensionality reduction, regularized models, transforming input space.

Applications of causal inference

- Decision evaluation
- Counterfactual estimation
- Dealing with selection bias



🔴 Decision Evaluation

🟢 Selection Bias Handling

🔵 Counterfactual Estimation

Online Advertising 🔴 🟢

Email Marketing 🔴 🔵

Recommendation 🔴 🟢

Health Care 🔵

Education 🔵

Online advertising

Will the ad attract user clicks?

Will a campaign increase sales?

Randomized experiments such as A/B testing?

Time-consuming and Expensive

👍 **Estimating the ad effect from observational data!**

Online Advertising as Causal Inference:

Estimating the ad effect from observational data

Observational data → Logged feedback records under **current** advertising system's policy

Treatment W

Ads

Outcome Y

Click

Variable X

Ad content

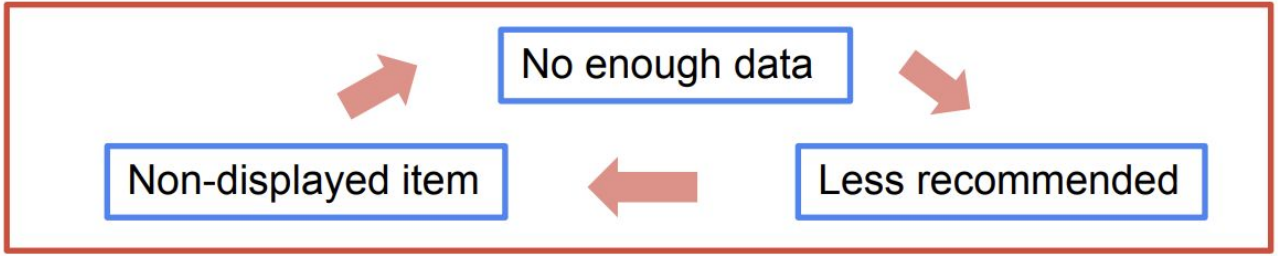
Recommendation

Selection bias:

❏ Users tend to rate the items that they like:

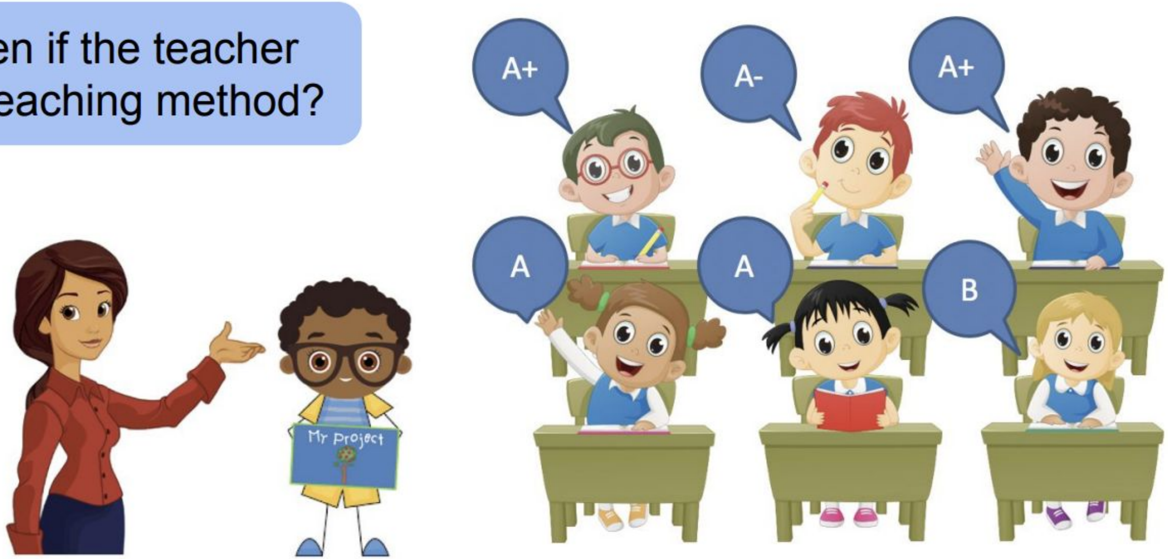
❏ The horror movie ratings are mostly made by horror movie fans and less by romantics movie fans.

❏ The records in the datasets are not representative of the whole population.



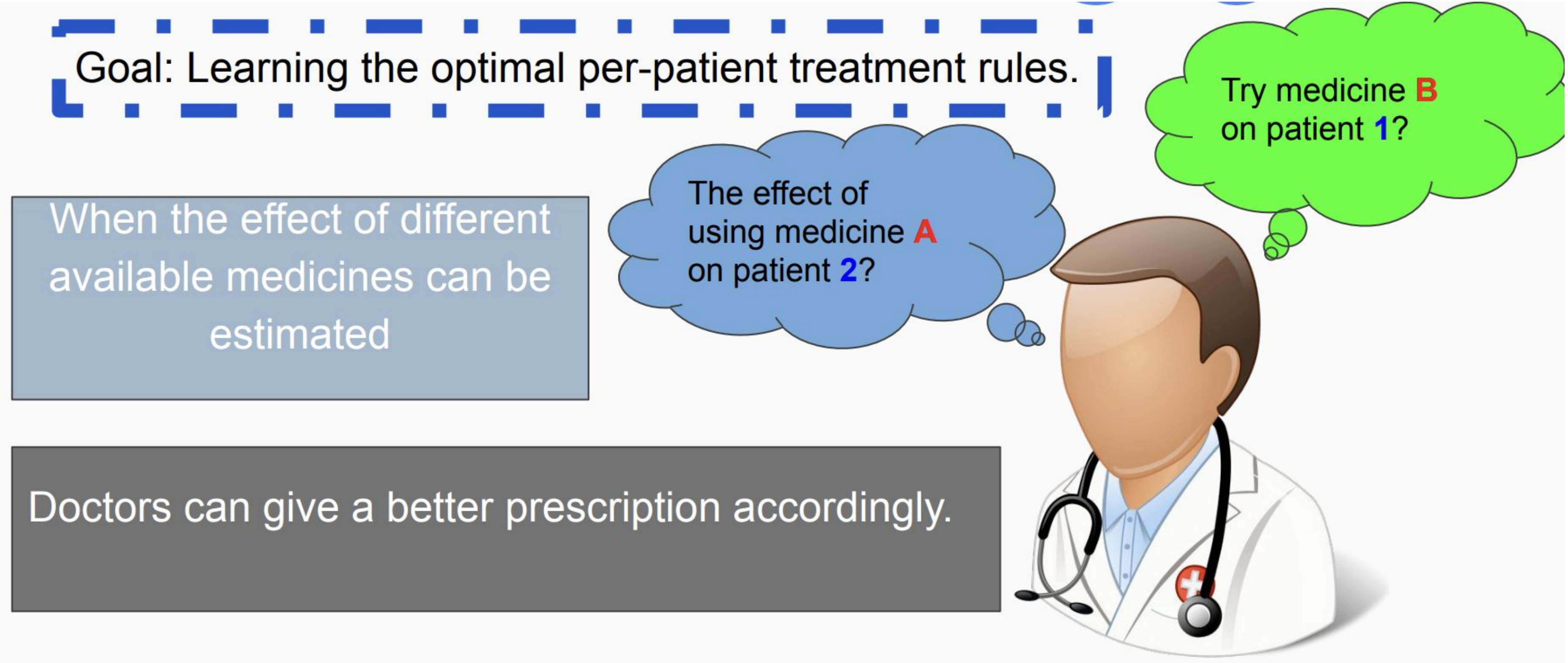
Education

What would happen if the teacher adopted another teaching method?



👍 Teachers can find the best teaching method for each individual!

Healthcare



Goal: Learning the optimal per-patient treatment rules.

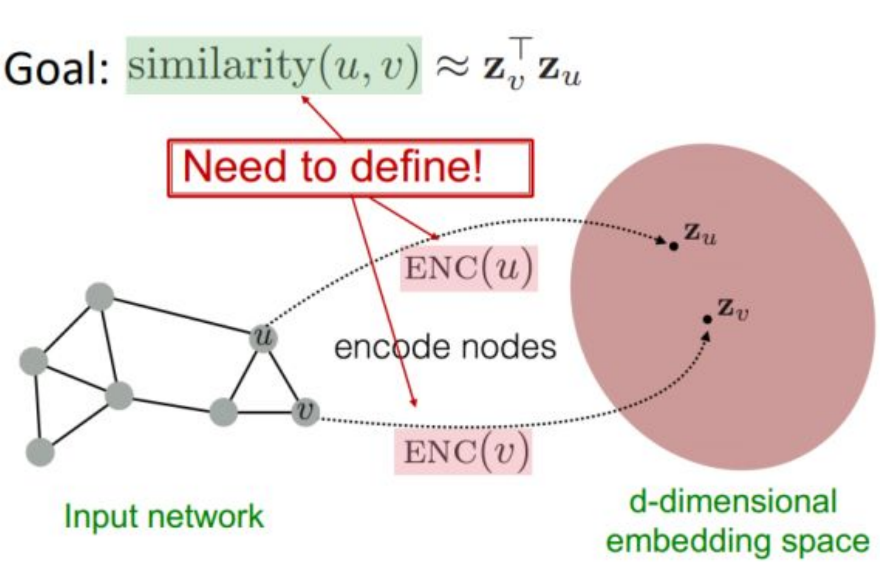
When the effect of different available medicines can be estimated

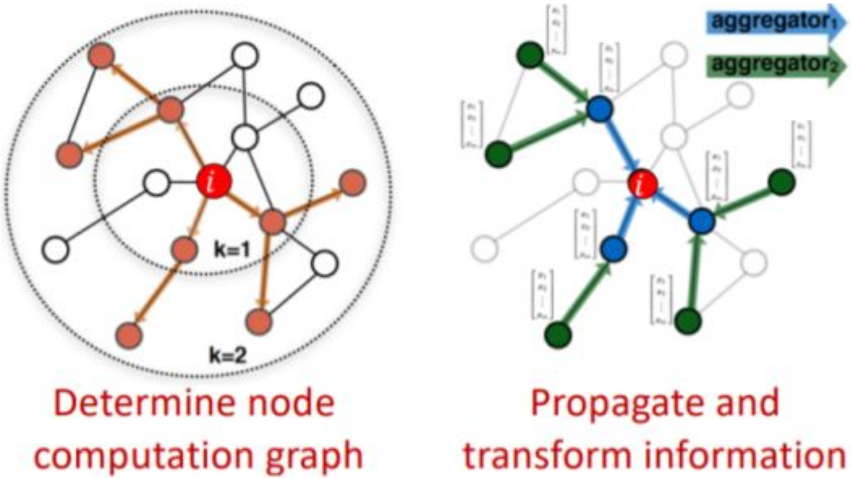
Doctors can give a better prescription accordingly.

The effect of using medicine A on patient 2?
Try medicine B on patient 1?

Machine learning for causal inference

-Graph Neural Network : a class of deep learning methods designed to perform inference on data described by graphs.





aggregator₁
aggregator₂
i
k=1
k=2
Determine node computation graph
Propagate and transform information

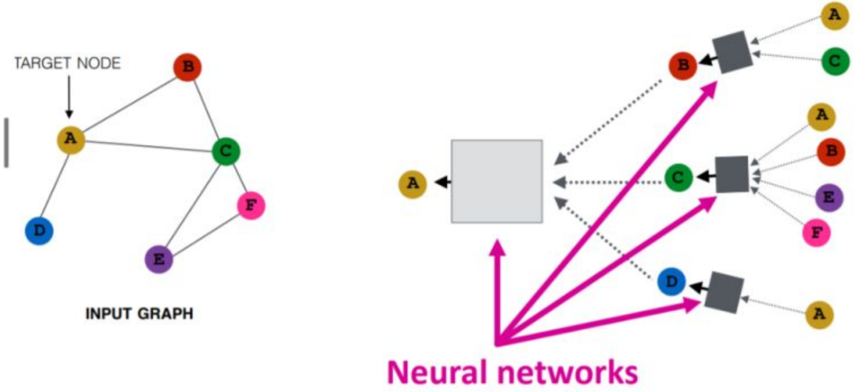
Locality (local network neighborhoods) in the encoder

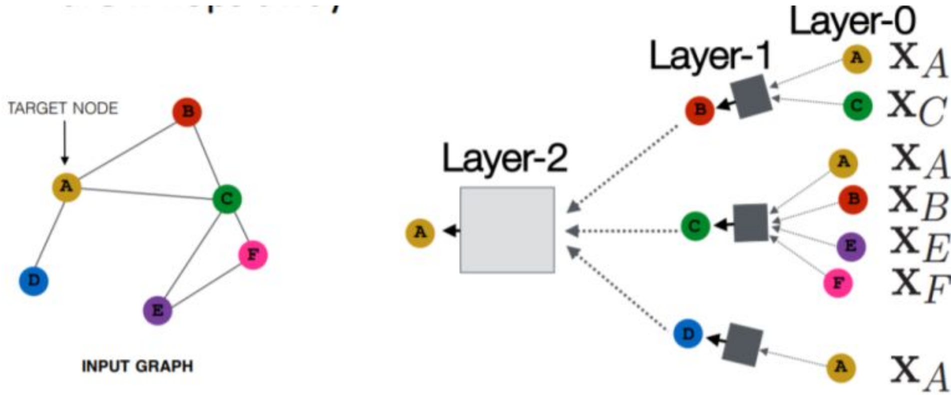
GNN

Aggregation and forward propagation rule

- **Simple neighborhood aggregation:**

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u\in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1}\right)$$





$h_v^0 = X_v\ (feature\ vector)$

$W_k \sum \frac{h_u^{k-1}}{|\ N(v)|}$

$B_k h_v^{k-1}$

$z_v = h_v^K$

Graph Convolutional Network layers

The simplest GCN has only three different operators:

-Graph convolution

-Linear layer

-Nonlinear activation

In practice : COVID

Treatment: Whether a certain policy is in effect (1 or 0) in different counties.

Outcome: The number of confirmed cases and death cases in different counties.

❑ To control for unobserved confounders, we collect

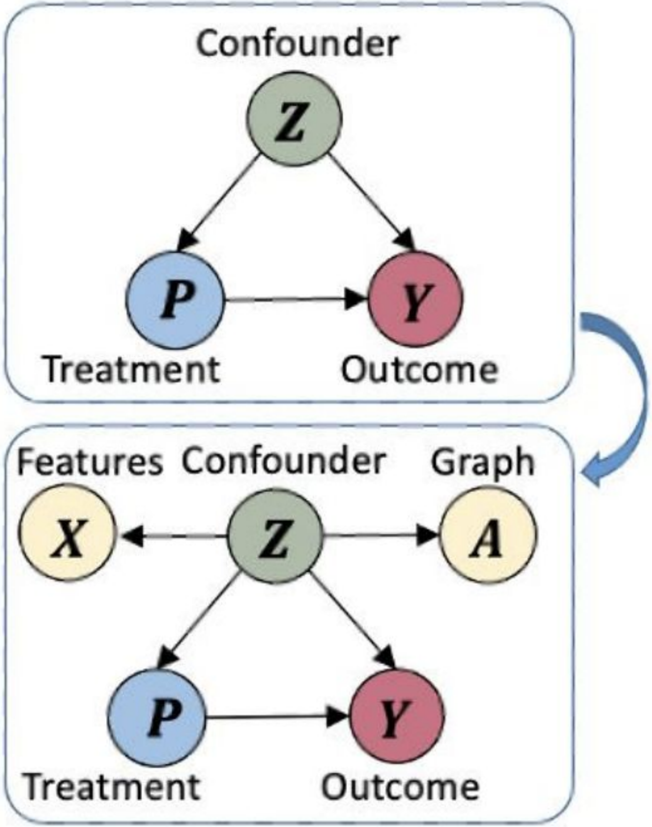
■ Features (covariates): data that reflect confounders (e.g., residents' vigilance) in counties – web searches

■ Graphs: relational information among counties, e.g., distance network/mobility flow

■ We assume these features and networks are correlated with the unobserved confounders

Graph neural network History embedding Graph structure

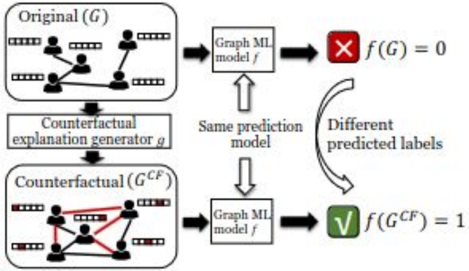
$$\boldsymbol{Z}_i^t = g(([\boldsymbol{X}^t, \widetilde{H}^{t-1}])_i, \boldsymbol{A}^t)$$



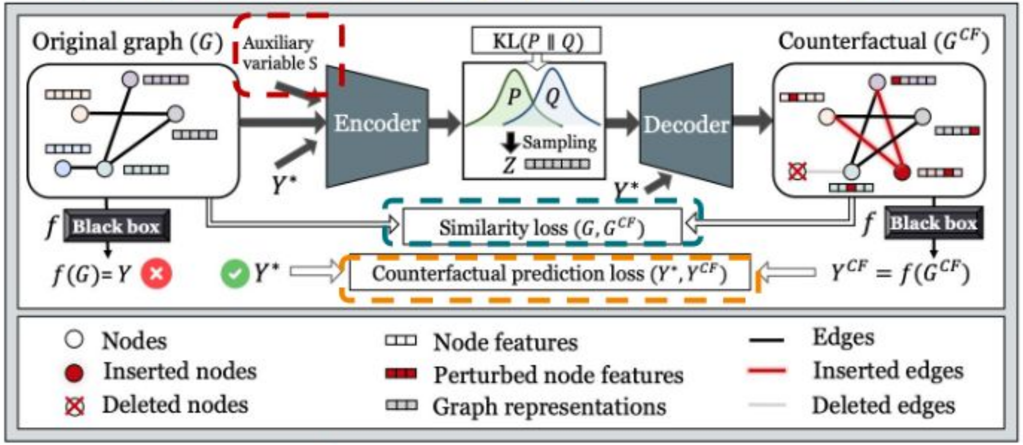
Explainability in machine learning

- "how should an input instance be perturbed to obtain a desired predicted label?" = counterfactual explanation

- An example for graphs : CLEAR



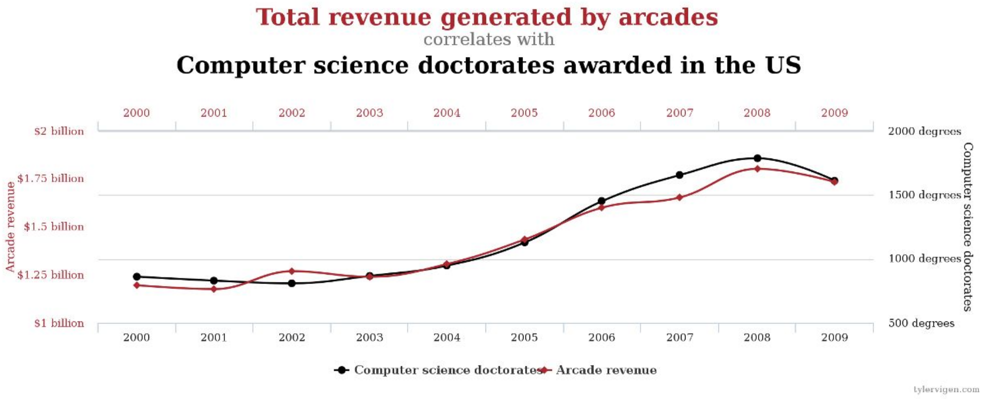
$$\mathcal{L} = \mathbb{E}_Q[d(G, G^{CF}) + \alpha\, l(f(G^{CF}), Y^*)] + \mathrm{KL}(Q(Z|G, S, Y^*)\|P(Z|G, S, Y^*))$$



Conclusion

Causal inference is tricky

Correlations are seldom enough. And sometimes horribly misleading.



Always be skeptical of causal claims from ~~observational~~ any data.

More data does not automatically lead to better causal estimates.

http://tylervigen.com/spurious-correlations

Try at least two methods with different assumptions: Higher confidence in estimate if both methods agree.

Connections between traditional machine learning problems and causal inference problems : missing data, high dimensionality …

“Machine learning for causal inference” and “Causal inference for machine learning”.

References

https://causalinference.gitlab.io/kdd-tutorial/

https://aaai23causalinference.github.io/causal_inference_slide.pdf

https://cobweb.cs.uga.edu/~shengli/Docs/AAAI-20-Causal-Inference-Tutorial.pdf

https://slideslive.com/38927861/time-series-deconfounder-estimating-treatment-effects-over-time-in-the-presence-of-hidden-confounders

https://neptune.ai/blog/graph-neural-network-and-some-of-gnn-applications#:~:text=Graph%20Neural%20Networks%20(GNNs)%20are,and%20graph%2Dlevel%20prediction%20tasks.

https://arxiv.org/pdf/2210.08443.pdf

https://www.youtube.com/watch?v=zCEYiCxrL_0&ab_channel=MicrosoftResearch

Thank you