Short Course
Pacific Causal Inference Conference

Xiao-Hua Andrew Zhou, Yuhao Deng, Chunyuan Zheng

2026/07/17 Tianjin, China

Lecturer: Xiao-Hua Andrew Zhou

- Distinguished Professor of the Overseas High-Level Talent Recruitment Programs (Innovative Long-term Project) of the Organization Department of the CPC Central Committee, Doctoral Supervisor
- Chair of the Department of Biostatistics of Peking University, PKU Distinguished Chair Professor of Beijing International Center for Mathematical Research, Vice Dean of National Institute for Regulatory Science of Drug and Medical Device of Peking University, Vice Dean of Peking University Chongqing Research Institute for Big Data
- Chair of International Biometric Society - China, Chair of the Medical Mathematics Professional Committee of the Chinese Mathematical Society, Fellow of the American Association for the Advancement of Science (AAAS), Fellow of the American Statistical Association (ASA), Fellow of the Institute of Mathematical Statistics (IMS), etc.

Lecturer: Yuhao Deng

- Education
  - Ph.D. in Statistics, School of Mathematical Sciences, Peking University
- Work Experience
  - Postdoc, Department of Biostatistics, University of Michigan
  - Postdoc, Vaccine and Infectious Disease Division, Fred Hutch Cancer Center
- Research Interests
  - Causal inference; Survival analysis; Semiparametric modeling; Econometrics
- Academic Service
  - Peer review 80+ times for 20+ journals
  - Three translated books

Lecturer: Chunyuan Zheng

- Education
  - Ph.D. student in Statistics, School of Mathematical Sciences, Peking University
- Research Interests
  - Causal inference; Large language model; Data mining
- Honors and Awards
  - National Scholarship for Doctoral Students, 2025
  - Peking University President Scholarship for Ph.D. Students, 2025
  - Top Award, Challenge Cup May Fourth Youth Science Award, 2025
  - Star of General Intelligence, 2025 — one of 10 awardees
  - CCF Ph.D. Student Funding Program, 2026 — one of 10 awardees
- Academic Service
  - Co-organizer, ICDM 2025 Workshop on User Modeling and Recommendation (UMRec)
  - Co-organizer, AAAI-25 Workshop on Artificial Intelligence with Causal Techniques (AICT)
  - Area Chair, NeurIPS 2026

Outline

- 1. Randomized Experiments
- 2. Observational Studies: Ignorability
- 3. Causal Graphs
- 4. Mediation Analysis
- 5. Principal Stratification
- 6. Application of Causal Inference in Medical Research: Intercurrent Events
- 7. Observational Studies: Non-Ignorability
- 8. Attribution
- 9. Deep Learning for Causal Effect Estimation in Complex Scenarios
- 10. Applications of Causal Methods in Real-World Scenarios

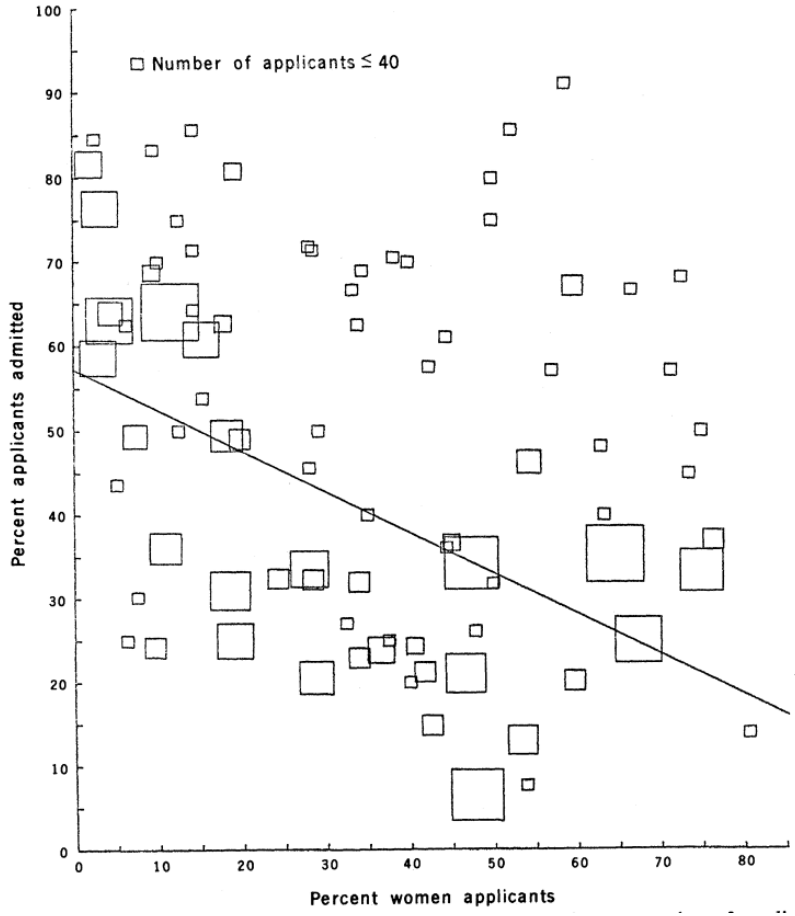
1. Randomized Experiments

Gender bias in graduate admissions

- UC Berkeley admissions in 1973 (Bickel et al, 1975)
- Admission rate lower for female than male

Male		Female	
Applicants	Admission rate	Applicants	Admission rate
8442	44%	4321	35%

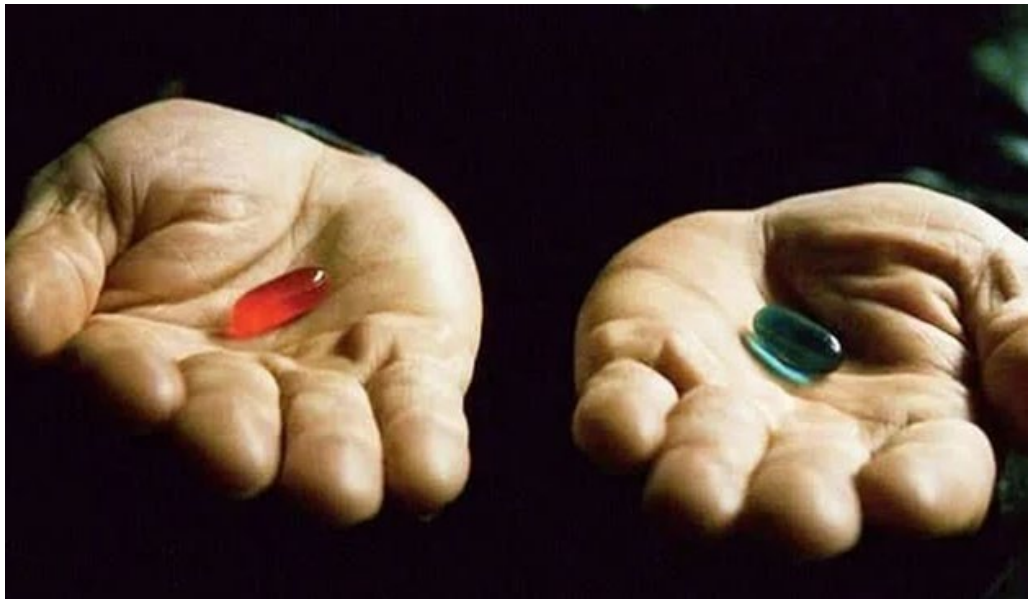
- Female applicants tend to apply for more competitive majors (confounder)



Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, *187*(4175), 398-404.

Potential outcome

- Elements of **Rubin causal model**: population, treatment, outcome, assignment mechanism
- Treatment: Z_i
- Potential outcome: $Y_i(1)$, $Y_i(0)$
- “No causation without manipulation”
- Individual treatment effect: $Y_i(1)-Y_i(0)$
- Individual treatment effect is not identifiable



Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Randomization

- Basic assumptions
- Stable unite treatment value assumption (SUTVA):
  - Only one version of treatment – $Y_i(1)$, $Y_i(0)$ are well-defined
  - No interference – $Y_i(z_1,\ldots,z_n) = Y_i(z_i)$
- Consistency: $Y_i = Y_i(Z_i) = Z_iY_i(1) + (1-Z_i)Y_i(0)$
- Randomization: $Z_i \perp (Y_i(1), Y_i(0))$
- “Golden standard” of evaluating treatment effects – other factors than treatment are balanced

Finite sample

- N individuals
- Finite-sample ATE:

$$\tau_{fs} = \bar{Y}(1) - \bar{Y}(0) = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}$$

- An estimator (difference-in-means):

$$\hat{\tau}_{fs} = \bar{Y}_t^{obs} - \bar{Y}_c^{obs} = \frac{1}{N_t}\sum_{i:Z_i=1} Y_i - \frac{1}{N_c}\sum_{i:Z_i=0} Y_i$$

Difference-in-means estimator

- Potential outcomes ($Y_i(1)$, $Y_i(0)$) are "fixed"; the only randomness comes from treatment assignment Z_i
- Unbiasedness:

$$E\left(\hat{\tau}_{fs}|\boldsymbol{Y}(1),\boldsymbol{Y}(0)\right)=\tau$$

- Variance:

$$\mathrm{var}\left(\hat{\tau}_{fs}\middle|\boldsymbol{Y}(1),\boldsymbol{Y}(0)\right)=\frac{S_t^2}{N_t}+\frac{S_c^2}{N_c}-\frac{S_{tc}^2}{N}$$

- S_t^2 is the variance of $Y_i(1)$, S_c^2 is the variance of $Y_i(0)$, estimable
- S_{tc}^2 is the variance of $Y_i(1)$-$Y_i(0)$, not estimable

Neyman variance

- If the individual treatment effect is constant, $S_{tc}^2 = 0$
- Neyman Variance estimate for $\hat{\tau}_{fs}$:

$$\hat{V} = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$$

- If the individual treatment effect is not constant, the Neyman variance is conservative

Super-population

- Assume the N units are a random sample from a super-population of size N_{sp}
- Super-population ATE:

$$\tau_{sp} = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} \{Y_i(1) - Y_i(0)\} = E\{Y_i(1) - Y_i(0)\}$$

- Due to random sampling,

$$E\left(\tau_{fs} | \boldsymbol{Y}(1), \boldsymbol{Y}(0)\right) = \frac{1}{N} \sum_{i=1}^{N_{sp}} E(R_i)\{Y_i(1) - Y_i(0)\} = \tau_{sp}$$

Difference-in-means estimator

- Unbiasedness:

$$E\left(\hat{\tau}_{fs}\middle|\boldsymbol{Y}(1),\boldsymbol{Y}(0)\right)=\tau_{sp}$$

- Another source of randomness: random sampling
- Variance of $\hat{\tau}_{fs}$:

$$\mathrm{var}\left(\hat{\tau}_{fs}\middle|\boldsymbol{Y}(1),\boldsymbol{Y}(0)\right)=\frac{\sigma_t^2}{N_t}+\frac{\sigma_c^2}{N_c}$$

- σ_t^2 is the variance of $Y_i(1)$, σ_c^2 is the variance of $Y_i(0)$, estimable
- The Neyman variance estimator is unbiased in super population

Regression estimator

- In regression analysis, we start from an infinite super population
- Potential outcomes are random in repeated sampling rom the super population
- Linear regression model for observed outcomes:

$$Y_i = \alpha + \tau Z_i + \epsilon_i$$

- Parameters themselves, which are model-dependent, do not necessarily have causal interpretations
- Parameters estimated by ordinary least squares (OLS)

Regression estimator

- Define $\alpha = E\{Y_i(0)\}, \tau = E\{Y_i(1) - Y_i(0)\}$
- The Gauss-Markov condition is satisfied under randomization:
  - $E(\epsilon_i | Z_i = 0) = E(Y_i(0) - \alpha | Z_i = 0) = 0$
  - $E(\epsilon_i | Z_i = 1) = E\big(Y_i(1) - \alpha - \tau_{sp} | Z_i = 1\big) = 0$
- Therefore, the OLS estimator $\hat{\tau}$ is an unbiased estimator for τ
- The robust variance estimator for $\hat{\tau}$ is identical to the Neyman variance estimator,

$$\hat{V} = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2 (Z_i - \bar{Z})^2}{\sum_{i=1}^{N} (Z_i - \bar{Z})^2}$$

Regression estimator

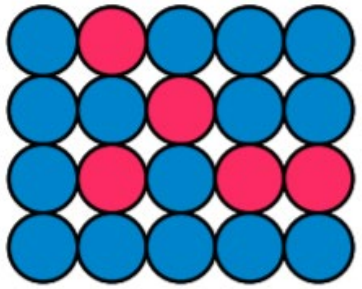
- Covariate adjustment:

$$Y_i = \alpha + \tau Z_i + \beta X_i + \epsilon_i$$

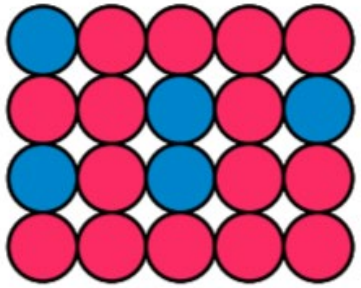
- The consistency of the regression estimator does not rely on correct model specification, because the population-level correlation between Z_i and X_i is zero due to randomization
- In a finite sample, the correlation between Z_i and X_i may not be zero
- Covariate adjustment can reduce finite-sample bias and variance

Re-randomization

- In a single trial, covariates may not be balanced between groups
- The more covariates, the more probability of imbalance
- Increasing the sample size does not help, because both the covariates distance and estimated ATE converge at the root-n rate



5 Females, 15 Males



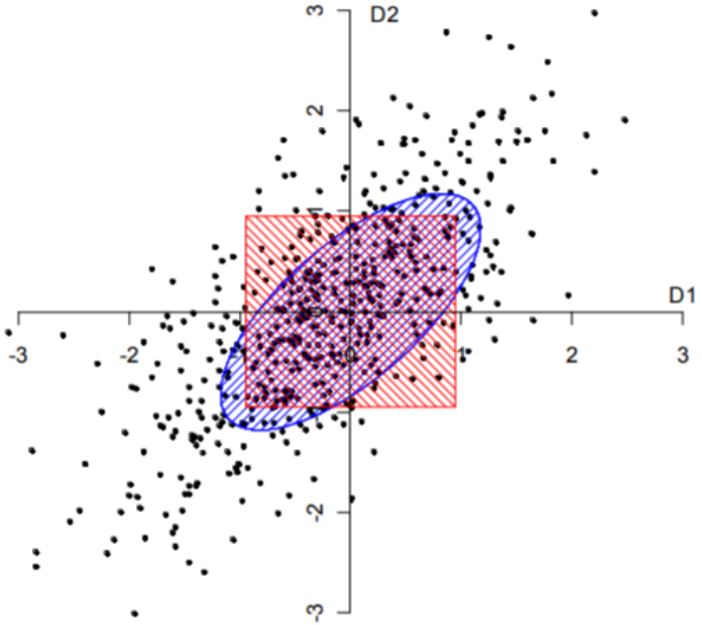
15 Females, 5 Males

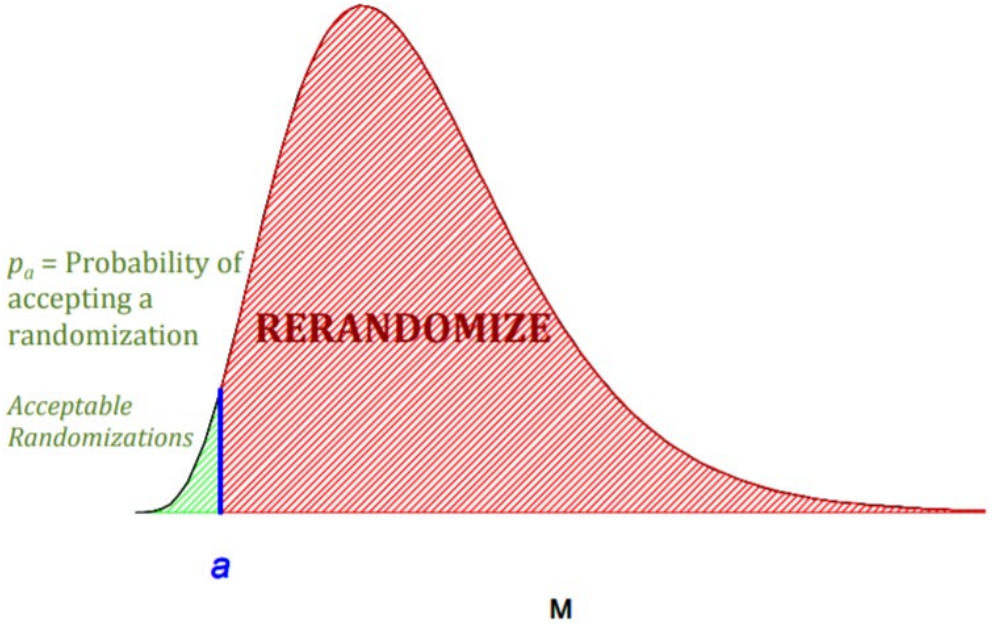
Re-randomization

- Mahalanobis distance of covariates

$$M = (\bar{X}_t - \bar{X}_c)' cov(X)^{-1} (\bar{X}_t - \bar{X}_c)$$

- Accept the randomization if M < a





Re-randomization

- Re-randomization reduces the variance of $\bar{X}_t - \bar{X}_c$ (Morgan and Rubin, 2012)
- Re-randomization reduces the variance of the ATE estimator
- The ATE estimator is not normal, but a mixture of normal and truncated-normal distributions
- Testing the ATE
  - Re-randomization test is computation demanding
  - Regression test is accurate
  - t-test is conservative

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2). 1263-1282

2. Observational Studies: Ignorability

Unconfounded treatment assignment

- Unconfoundedness (ignorability): $Z \perp (Y(1),Y(0)) \mid X$
- Consistency: $Y = Y(Z)$
- Propensity score: $e(x) = P(Z=1 \mid X=x)$
- Positivity: $0 < e(x) < 1$

- ATE is identifiable:
- $\tau = E\{Y(1) - Y(0)\} = E[E\{Y(1) \mid X\}] - E[E\{Y(0) \mid X\}]$
  $= E\{E(Y \mid Z=1, X)\} - E\{E(Y \mid Z=0, X)\}$

Identifiability

- Z: Treatment
- Y: Survival
- X: Sex
- $\hat{\tau} = \hat{E}\{\hat{E}(Y|Z=1,X)\} - \hat{E}\{\hat{E}(Y|Z=0,X)\}$

$$= \begin{bmatrix} \hat{P}(X=1)\hat{E}(Y|Z=1,X=1) \\ +\hat{P}(X=0)\hat{E}(Y|Z=1,X=0) \end{bmatrix} - \begin{bmatrix} \hat{P}(X=1)\hat{E}(Y|Z=1,X=1) \\ +\hat{P}(X=0)\hat{E}(Y|Z=1,X=0) \end{bmatrix}$$

$$= [0.5 \times 60\% + 0.5 \times 20\%] - [0.5 \times 70\% + 0.5 \times 30\%]$$

$$= -10\%$$

	Survived	Dead	Survival rate
Total			
Treated	20	20	50%
Control	16	24	40%
Male (X = 1)			
Treated	18	12	60%
Control	7	3	70%
Female (X = 0)			
Treated	2	8	20%
Control	9	21	30%

Inverse probability weighting

- Propensity score is a balancing score: Z ⊥ (Y(1),Y(0)) | e(X)
- Conditioning on e(X) generates a (stratified) randomized experiment
- Inverse probability weighting (IPW):

$$E\left[\frac{ZY}{e(X)}\right] = E\{Y(1)\}, \qquad E\left[\frac{(1-Z)Y}{1-e(X)}\right] = E\{Y(0)\}$$

- The propensity score needs to be estimated if unknown
- Horvitz-Thompson estimator:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}\right]$$

Inverse probability weighting

- e(1) = P(Z = 1 | X = 1) = 0.75
- e(0) = P(Z = 1 | X = 0) = 0.25
- $\hat{\tau} = \hat{E}\left[\frac{ZY}{\hat{e}(X)}\right] - \hat{E}\left[\frac{(1-Z)Y}{1-\hat{e}(X)}\right]$

$$= \frac{1}{80}\left[\left(\frac{18}{0.75} + \frac{2}{0.25}\right) - \left(\frac{7}{0.25} + \frac{9}{0.75}\right)\right]$$

$$= -10\%$$

	Survived	Dead	Survival rate
Total			
Treated	20	20	50%
Control	16	24	40%
Male (X = 1)			e(X) = 0.75
Treated	18	12	60%
Control	7	3	70%
Female (X = 0)			e(X) = 0.25
Treated	2	8	20%
Control	9	21	30%

Inverse probability weighting

- Super efficiency: using estimated propensity score can be more efficient than using the true propensity score (Hirano et al, 2003)
- Hájek estimator using stabilized propensity score:

$$\hat{\tau} = \sum_{i=1}^{N} \frac{Z_i Y_i}{\hat{e}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{Z_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{1-Z_i}{1-\hat{e}(X_i)}$$

- If Y is binary, the Hájek estimator is always between 0 and 1

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.

Variants of inverse probability weighting

- Subclassification estimator
  - Divide the sample into some strata based on $\hat{e}(X)$, each of which is like a randomized experiment
  - Average stratum-specific ATE: constant propensity score in each stratum
  - Covariate adjustment in each stratum
- Trimming
  - Discard units whose estimated propensity are close to 0 or 1
  - Change the estimand

Weighted linear regression

- A weighted estimator can be obtained from linear regression

$$Y_i = \alpha + \tau Z_i + \epsilon_i$$

- Weight:

$$w_i = \frac{Z_i}{\hat{e}(X_i)} + \frac{1 - Z_i}{1 - \hat{e}(X_i)}$$

- Covariate adjustment in the linear regression:

$$Y_i = \alpha + \tau Z_i + \beta X_i + \epsilon_i$$

- Doubly robust: If either the propensity score of linear model is correctly specified, then the weighted estimator is consistent

Doubly robust estimator

- Augment inverse probability weighting (AIPW) estimator:
- $\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Z_i Y_i}{\hat{e}(X_i)} + \left\{1 - \frac{Z_i}{\hat{e}(X_i)}\right\}\hat{\mu}_1(X_i)\right]$
- $\hat{\mu}_0 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} + \left\{1 - \frac{1-Z_i}{1-\hat{e}(X_i)}\right\}\hat{\mu}_0(X_i)\right]$
- $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$
- Doubly robust: consistent if either the propensity score $\hat{e}(x)$ or the outcome regression $\hat{\mu}_z(x)$ is correctly specified

Doubly robust estimator

- Assume models are correctly specified
- $\hat{e}(x)$ and $\hat{\mu}_z(x)$ can be estimated by machine-learning methods, with convergence rates $o_p(n^{-1/4})$
- Efficiency: least asymptotic variance among all regular and asymptotically linear estimators (under some regularity conditions)
- The “influence function” is the efficient influence function (EIF)
- Double machine learning: sample splitting and cross-fitting
  - Fit nuisance models
  - Calculate the EIF

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1). 1-68.

Target minimum loss estimation (TMLE)

- Outcome regression:

$$Y_i = \alpha + \tau Z_i + \beta X_i + \epsilon_i$$

- Regress residuals $\hat{\epsilon}_i$ on a "clever covariate" (propensity scores):

$$\hat{\epsilon}_i = \nu \left[\frac{Z_i}{\hat{e}(X_i)} + \frac{1 - Z_i}{1 - \hat{e}(X_i)} \right] + u_i$$

- Adjusted model: the score function of ν solves the EIF

$$\tilde{\mu}_z(X_i) = \hat{\alpha} + \hat{\tau} z + \hat{\beta} X_i + \hat{\nu} \left[\frac{z}{\hat{e}(X_i)} + \frac{1 - z}{1 - \hat{e}(X_i)} \right]$$

- TMLE estimator: $\tilde{\tau} = \frac{1}{N} \sum_{i=1}^{N} [\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)]$

van der Laan, M. J., & Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, *2*(1).

Target minimum loss estimation (TMLE)

- TMLE is a “substitution” estimator, preserving model boundaries
- Particularly appealing for binary outcomes
- Logistic regression:

$$P(Y_i = 1|Z_i, X_i) = \text{expit}\{\alpha + \tau Z_i + \beta X_i + \epsilon_i\}$$

- Fit another logistic regression for Y_i on $\left[\frac{Z_i}{\hat{e}(X_i)} + \frac{1-Z_i}{1-\hat{e}(X_i)}\right]$ with $\hat{P}(Y_i = 1|Z_i, X_i)$ as offset
- Adjusted model: $\tilde{P}(Y_i = 1|z, X_i)$
- TMLE estimator: $\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}[\tilde{P}(Y_i = 1|1, X_i) - \tilde{P}(Y_i = 1|0, X_i)]$

G-estimation

- Motivation: rank preservation, $Y_i(z) - Y_i(0) = \beta_1 z + \beta_2 z X_i$
- Under consistency, $Y_i(0) = Y_i - \beta_1 Z_i - \beta_2 Z_i X_i$
- Under ignorability, $Z_i \perp Y_i(0) | X_i$
- G-estimation (rank preservation not required):

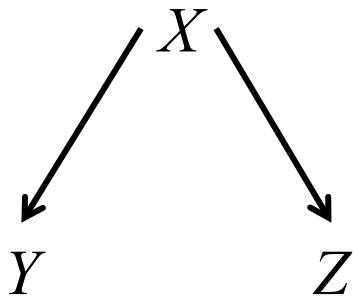
$$\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} c_1(X_i) \\ c_2(X_i) \end{pmatrix}\{Z_i - e(X_i)\}\{Y_i - \beta_1 Z_i - \beta_2 Z_i X_i\} = 0$$

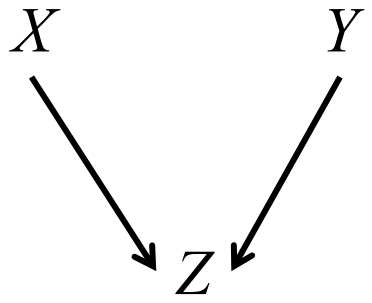
- If the "blip function" is correctly specified, then the parameter estimators are consistent (so that ATE)

3. Causal Graphs

Atom structures

- Chain
- Fork
- Collider

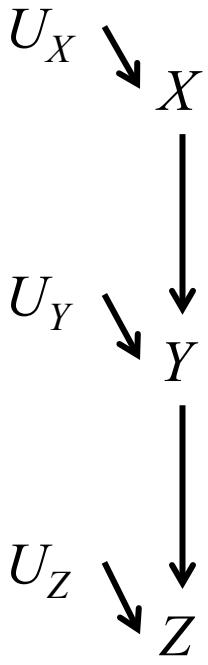




Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

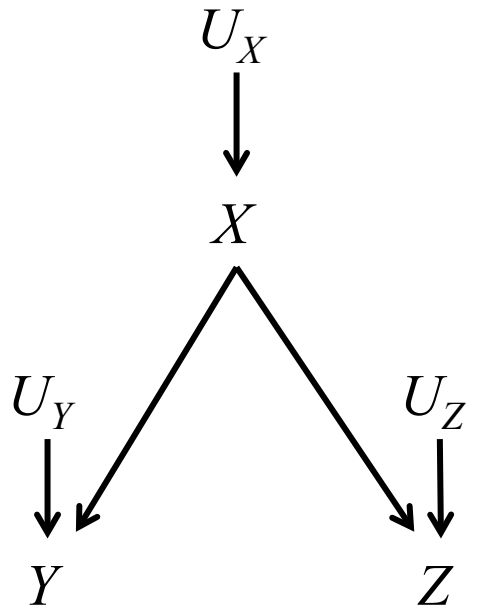
Chain

- Example
  - X: funding
  - Y: GPA
  - Z: collage admission
- Y is not independent of X
- Z is not independent of Y
- Z is likely not independent of X
- Conditional on Y, Z is independent of X



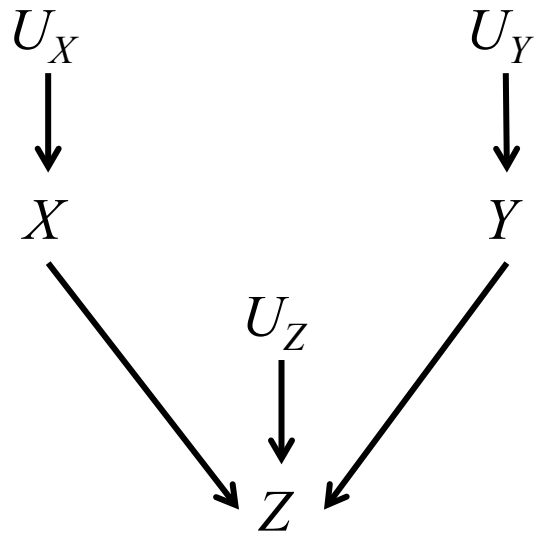
Fork

- Example
  - X: temperature
  - Y: sell of ice cream
  - Z: crime rate
- Y is not independent of X
- Z is not independent of X
- Z is likely not independent of Y
- Conditional on X, Z is independent of Y
- Yule-Simpson paradox



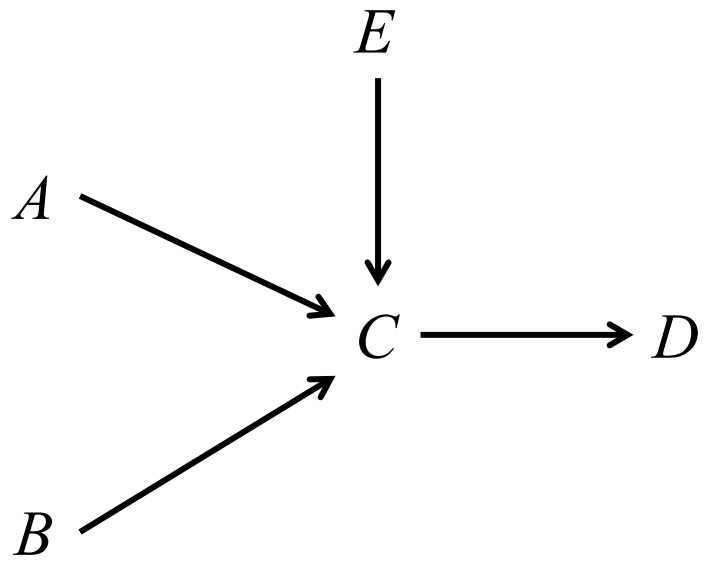
Collider

- Example
  - X: music talent
  - Y: GPA
  - Z: scholarship
- Z is not independent of X
- Z is not independent of Y
- X is independent of Y
- Conditional on Z, X is independent of Y
- Berskon paradox; selection bias; Monty Hall's problem



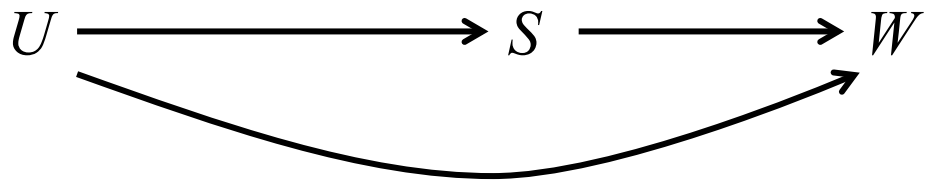
Children of a collider

- Conditioning on C introduces correlation between A and B
- D carries information of C, so controlling D partially controls C, thereby introducing correlation between A and B
- C carries information of E, but controlling E does not introduce correlation between A and B



Example

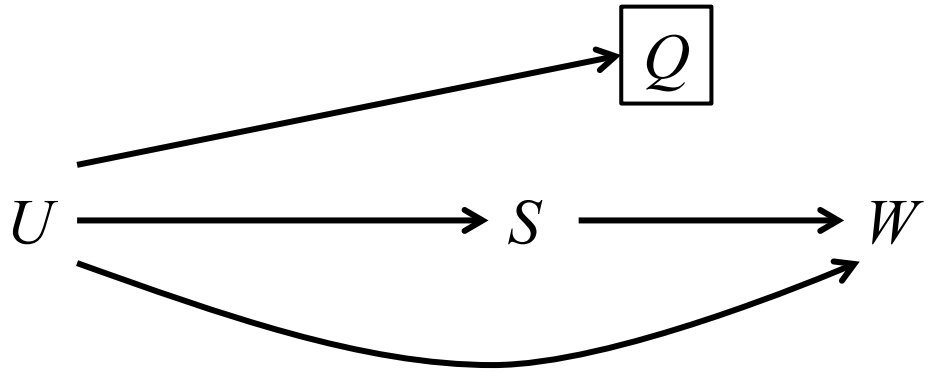
- S: education
- W: wages
- U: ability (unobserved)
- Aim: causal effect of S on Y



- Due to the confounder, the correlation between S and W is not causal

Example

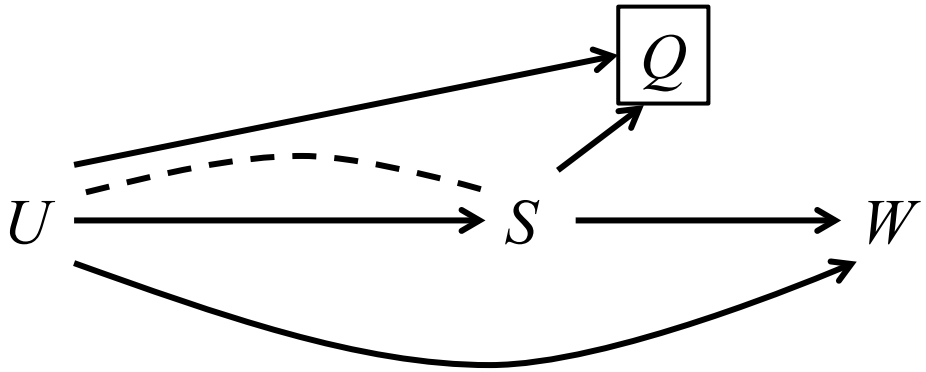
- S: education
- W: wages
- U: ability (unobserved)
- Q: IQ
- Aim: causal effect of S on Y



- We can observe a proxy of U
- Controlling Q partially reduces the confounding bias

Example

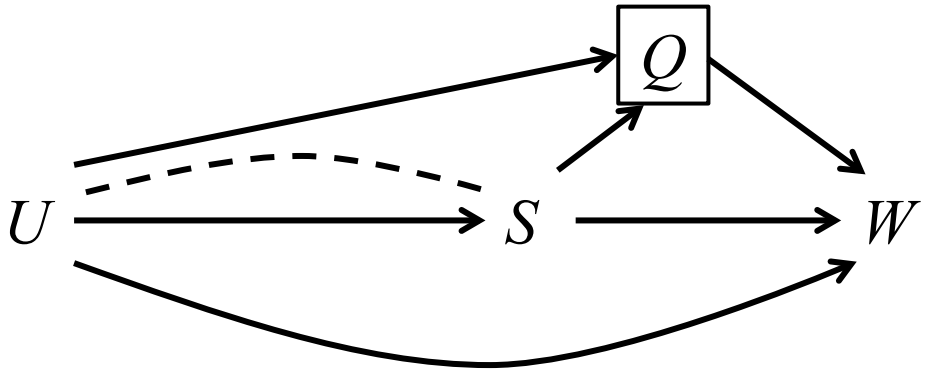
- S: education
- W: wages
- U: ability (unobserved)
- Q: IQ
- Aim: causal effect of S on Y



- If S influences Q, then controlling Q introduces a collider bias

Example

- S: education
- W: wages
- U: ability (unobserved)
- Q: IQ
- Aim: causal effect of S on Y



- If Q influences W, then controlling Q blocks part of S→W causal effect

d-separation

- If X and Y are d-separated by Z, then X is independent of Y in all probability distributions that are admissible with the causal graph
- How to achieve d-separation:
  - Conditioning on Z in the chain: X→Z→Y
  - Conditioning on Z in the fork: X←Z→Y
  - Not conditioning on Z in the collider: X→Z←Y

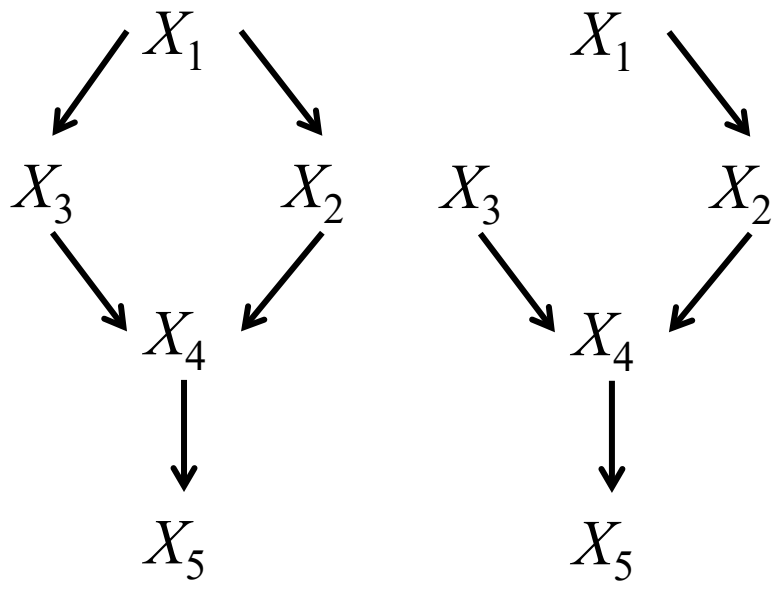
Directed acyclic graph (DAG)

- Truncated law:

$$P(x_1, x_2, x_3, x_4, x_5)$$
$$= P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2, x_3)P(x_5 \mid x_4)$$
$$= \prod_{j=1}^{5} P(x_j \mid pa_j)$$

- Intervention in $X_3 = 1$:

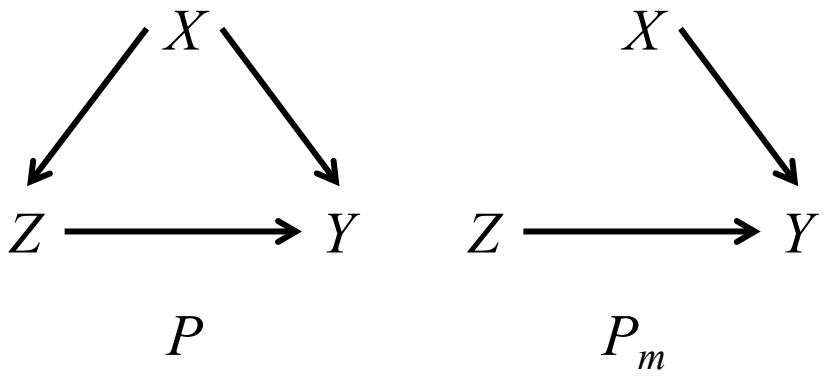
$$P_{X_3=1}(x_1, x_2, x_3, x_4, x_5)$$
$$= P(x_1)P(x_2 \mid x_1)P(x_4 \mid x_2, x_3 = 1)P(x_5 \mid x_4)$$



Calculating the causal effect

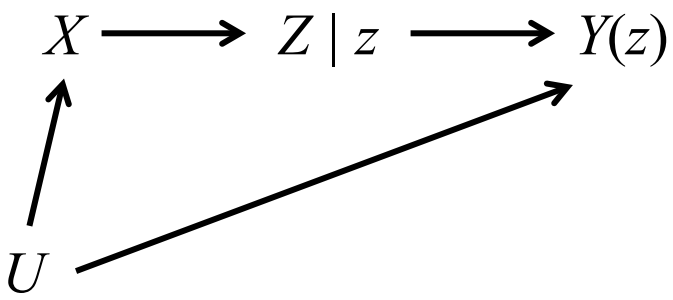
- To identify the causal effect of Z on Y, we make an intervention in Z
- In potential outcomes, Y(z) = Y|do(Z=z)

$$P(Y = y \mid do(Z = z)) = P_m(Y = y \mid Z = z)$$
$$= \sum_x P_m(Y = y \mid Z = z, X = x) P_m(X = x \mid Z = z)$$
$$= \sum_x P_m(Y = y \mid Z = z, X = x) P_m(X = x)$$
$$= \sum_x P(Y = y \mid Z = z, X = x) P(X = x)$$



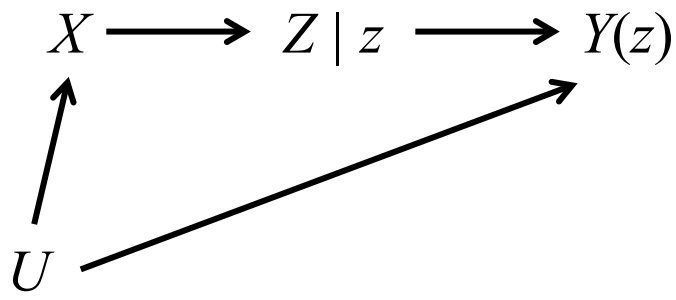
Single-world intervention graph (SWIG)

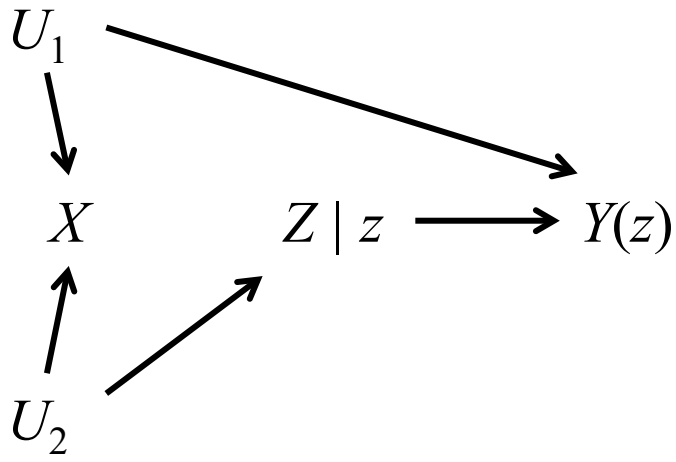
- Intervention induces a new node
- Is the causal effect of Z on Y identifiable?
- Is Z d-separated with Y(z)?
- Conditioning on X leads to d-separation
- Z→Y effect is identifiable by controlling for X



Single-world intervention graph (SWIG)

- Intervention induces a new node
- Is the causal effect of Z on Y identifiable?
  - Unconfoundedness Z ⊥ Y(z) | X
  - Y(z) is d-separated with Z by X
- 1. Conditioning on X leads to d-separation
- Z→Y effect is identifiable by controlling for X
- 2. Z is d-separated with Y(z)
- Z→Y effect is identifiable by not controlling for X





4. Mediation Analysis

Mediation analysis

- The treatment has two causal pathways to the outcome
  - 1. Direct pathway: Z to Y
  - 2. Indirect pathway: Z to M to Y
- Example: stem cell transplantation to cure leukemia
  - Z: transplant modality
  - M: leukemia relapse
  - Y: death
- Goal: Distinguish the direct/indirect from the total effect

Mediation analysis

- The classical mediation analysis (Baron and Kenny, 1986) was proposed in terms of linear models:

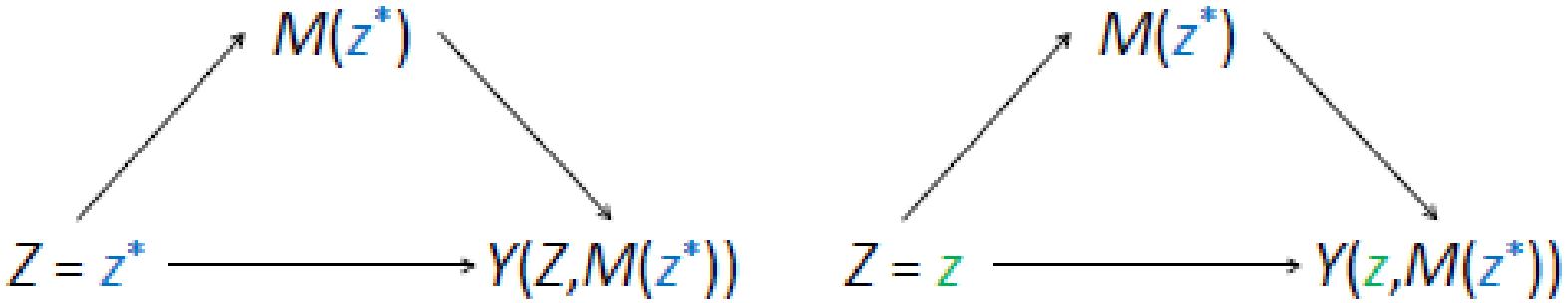
$$\begin{cases} M = \alpha Z + e \\ Y = \beta Z + \gamma M + u \end{cases}$$

- Direct effect β, indirect effect $\alpha\gamma$
- In potential outcomes, Y(z,m)
- Total effect τ = E{Y(1) – Y(0)} = E{Y(1,M(1)) – Y(0,M(0))}
- Controlled effect: E{Y(1,m) – Y(0,m)}

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Natural effects

- Natural direct effect E{Y(1,M(0)) – Y(0,M(0))}
- Natural indirect effect E{Y(1,M(1)) – Y(1,M(0))}
- Total effect = direct effect + indirect effect



Identification of natural effects

- Ignorability: Z ⊥ {M(z), Y(z*,m)} | X
- Sequential ignorability: M(z) ⊥ Y(z*,m) | Z=z, X
  - The mediator is as if randomized – no unmeasured confounding between Z, M, and Y
- Other assumptions: positivity, consistency
- Identification (Imai et al., 2010):

$$E\{Y(z, M(z^*))|X\} = \int_m E(Y|z, m, X) f(m|z^*, X) dm$$
$$= E_X[E_{M|z^*,X}\{\mu(z, M, X)\}]$$

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, *25*(1), 51-71.

Estimation of natural effects

- How to estimate E{Y(z,M(z*))}?
- Regression estimation
  - Estimate $\mu(z, m, x)$
  - Estimate (like imputation) $\nu(x) = \int \mu(z, m, x) f(m|z^*, x) dm$ (e.g., by Monte Carlo)
  - Average $\nu(x)$ over the sample
- Weighted estimation
  - Sample average of $\frac{I(Z=z) f(M|z^*, X)}{f(M|z, X) P(z|X)} Y$

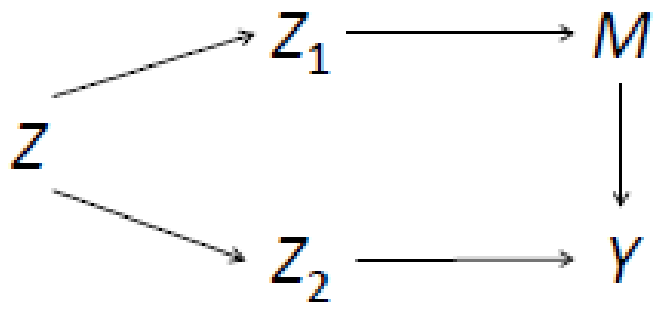
Multiply robust estimator

- Combine three models
  - Propensity score $P(z|x)$
  - Mediator density $f(m|z,x)$
  - Outcome regression $\mu(z,m,x)$
- Multiple robustness: consistent if either two of the three working models are correctly specified
- The estimator that solves the efficient influence function (EIF)
- Semiparametric efficiency (Tchetgen Tchetgen and Shpitser, 2012)

Tchetgen, E. J. T., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, *40*(3), 1816.

Separable effects

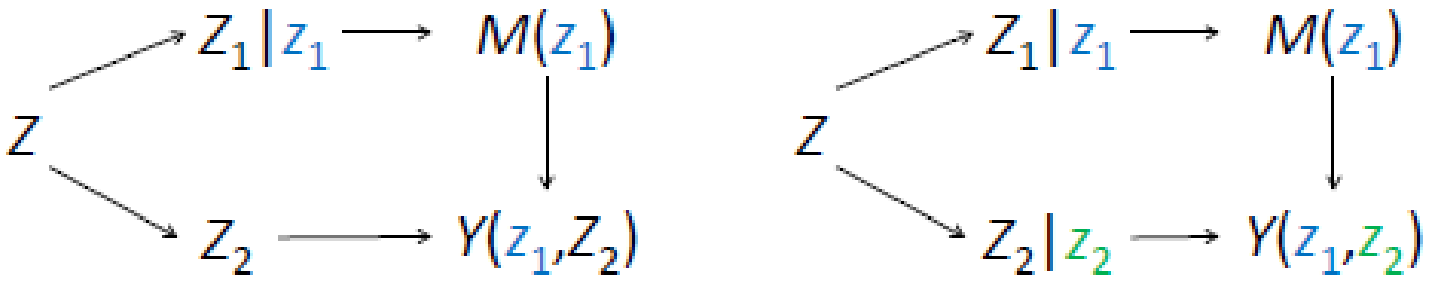
- Sequential ignorability is a “cross-world” assumption – it involves both the world of z and z* in a single independency
- Interventionist approach: separable effects (Stensrud et al., 2022)
- The treatment Z has two components Z_M and Z_Y
  - $Z = Z_M = Z_Y$ in a realized trial
- Potential outcomes: $M(z_M,z_Y)$, $Y(z_M,z_Y)$



Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., & Hernán, M. A. (2022). Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, *117*(537), 175-183.

Separable effects

- Total effect τ = E{Y(1)-Y(0)} = E{Y(1,1) – Y(0,0)}
- Separable direct effect E{Y(0,1) – Y(0,0)}
- Separable indirect effect E{Y(1,1) – Y(0,1)}
- Total effect = direct effect + indirect effect



Identification of separable effects

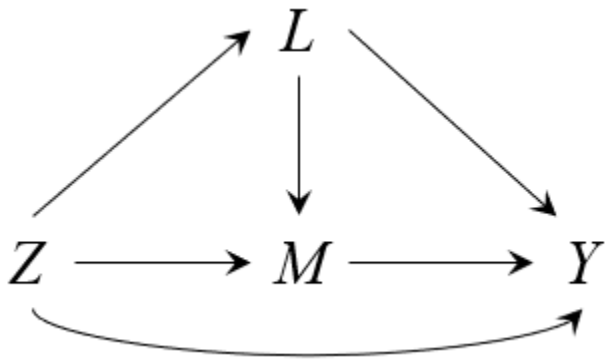
- Ignorability: $Z \perp \{M(z_M,z_Y), Y(z_M,z_Y)\} \mid X$
- Dismissible components condition:
  - $M(z_M,z_Y) = M(z_M,z_M)$
  - $E\{Y(z_M,z_Y) \mid M(z_M,z_Y), X\} = E\{Y(z_Y,z_Y) \mid M(z_Y,z_Y), X\}$
- A graphical justification of dismissible components condition: partial isolation – no confounding between M and Y
- Not testable using observed data, but it will be testable "in future trials" if the treatment components can be separated
- Other assumptions: positivity and consistency

Identification of separable effects

- Identification formula is identical with natural effects
- $E\{Y(z_M, z_Y)|X\} = \int_m E(Y|z_Y, m, X) f(m|z_M, X) dm$
  $= E_X[E_{M|z_M,X}\{\mu(z_Y, M, X)\}]$
- But different interpretations
- Separable effects may target on clinical interest directly
- Caveat: Can we find “treatment components”?

Randomized interventional effects

- If there is treatment-induced confounding, sequential ignorability fails
- Dismissible components condition is also hard to justify
- Randomized intervention：
  - Set Z at z
  - Replace M with an external variable $G(z^*) \sim M(z^*) \mid X$, marginalizing over L



VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(3), 917-938.

Randomized interventional effects

- Interventional direct effect: E{Y(1,G(0)) – Y(0,G(0))}
- Interventional indirect effect: E{Y(1,G(1)) – Y(1,G(0))}
- Direct effect + indirect effect ≠ total effect
- Assume sequential ignorability conditional on treatment-induced confounding

$$E\{Y(z, G(z^*))|X\} = \int\int E(Y|z,l,m,X)f(m|z^*,x)f(l|z,x)dmdl$$

- Limitation: interventional effect is not sharp
- Interventional direct effect = 0 does not imply no Z → Y effect

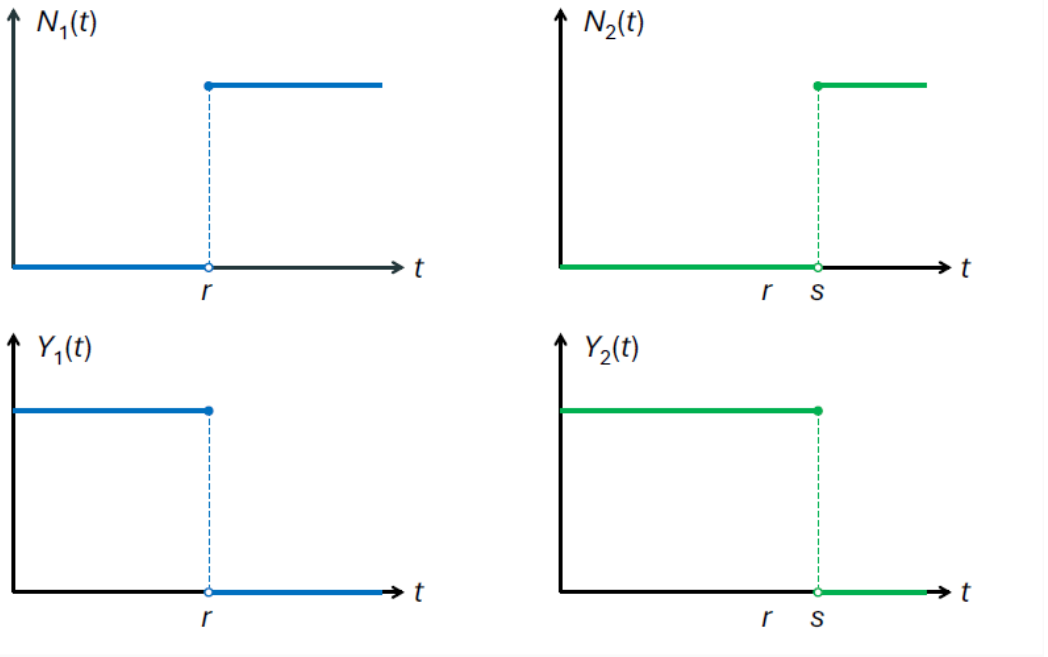
Mediation analysis for survival outcomes

- Semi-competing risks
- The mediator and outcomes are processes (Huang, 2021)
- Treatment: transplant modality
- A terminal (primary) event: death
- An intermediate event: relapse
- At every time, the intermediate event process 0,1; the terminal event process 0,1

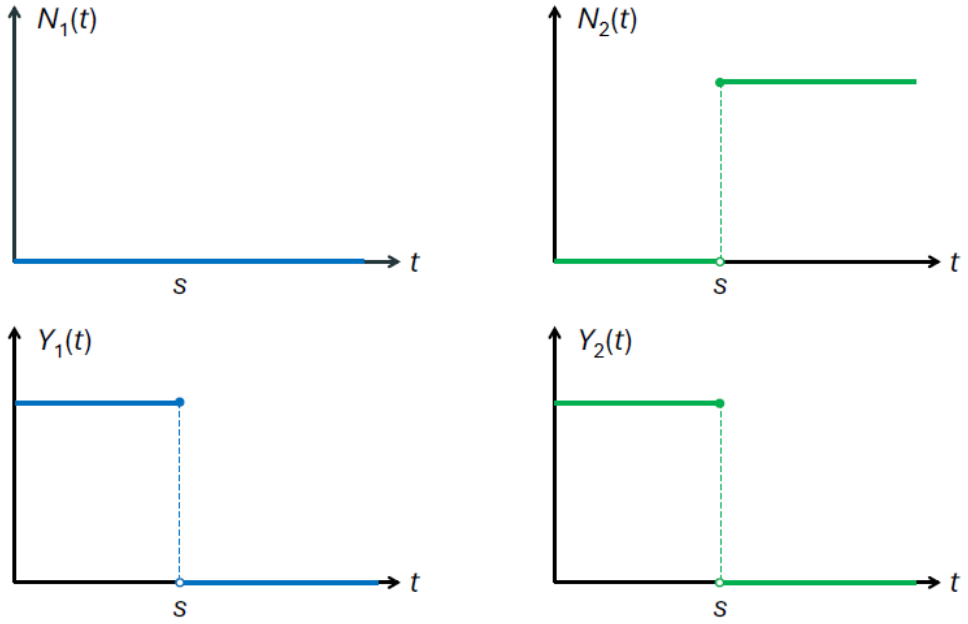
Huang, Y. T. (2021). Causal mediation of semicompeting risks. *Biometrics*, *77*(4), 1143-1154.

Counting processes

- Intermediate event occurs



- Intermediate event does not occur



Potential outcomes on counting processes

- Jump of the potential counting process
  - Intermediate event: $dN_1(t;z)$
  - Terminal event: $dN_2(t;z,\bar{n}_1(t))$
- Potential counting process
  - Intermediate event: $N_1(t;z) = \int_0^t dN_1(s;z)$
  - Terminal event: $N_2\big(t;z,\bar{n}_1(t)\big) = \int_0^t dN_2\big(s;z,\bar{n}_1(s)\big)$
- Natural counting process of the terminal event: $N_2(t;z,\bar{N}_1(t;z))$

Natural effects

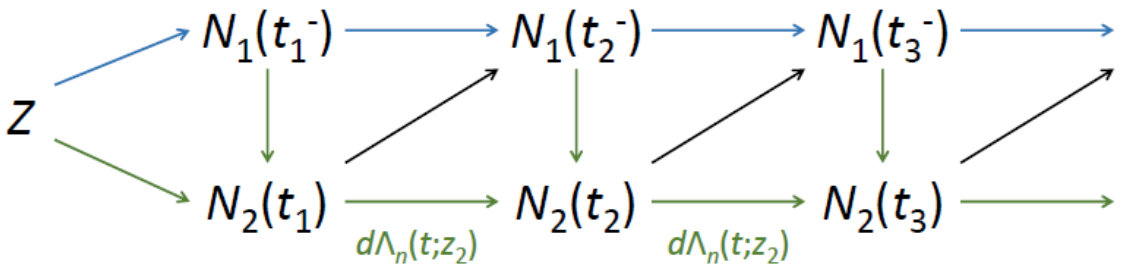
- Assign different treatments for N_1 and N_2 – potential counting process of the terminal event $N_2(t; z_2, \bar{N}_1(t; z_1))$
- Cumulative incidence function (CIF) of the potential terminal event:

$$F_2(t; z_1, z_2) = P\{N_2(t; z_2, \bar{N}_1(t; z_1)) = 1\}$$

- Total effect: $F_2(t; 1,1) - F_2(t; 0,0)$
- Natural direct effect: $F_2(t; 0,1) - F_2(t; 0,0)$
- Natural indirect effect: $F_2(t; 1,1) - F_2(t; 0,1)$
- Total effect = direct effect + indirect effect

Identification assumptions

- Ignorability: $Z \perp \{N_1(t; z_1), N_2(t; z_2, \bar{n}_1(t; z_1))\}|X$
- Consistency: $N_1(t) = N_1(t; Z)$, $N_2(t) = N_2(t; Z, \bar{N}_1(t; Z))$
- Random censoring
- Positivity: sufficient data in each group
- Under these assumptions, $F_2(t; z, z)$ is identifiable
- $F_2(t; z_1, z_2)$ is not identifiable if $z_1 \neq z_2$



Identification of natural effects

- Sequential ignorability
  - $dN_1(t; z_1)$ is as-if randomized conditional on history
  - $dN_2\big(t; z_2, n_1(t)\big)$ is as-if randomized conditional on history
- No unmeasured confounding between the intermediate event and the terminal event
- Cause-specific hazards are identifiable
  - Intermediate event: $d\Lambda_1(t; z_1, z_2, x) = d\Lambda_1(t; z_1, x)$
  - Terminal event: $d\Lambda_2(t; z_1, z_2, \bar{n}_1(t), x) = d\Lambda_2(t; z_2, \bar{n}_1(t), x)$
- Counterfactual CIF is identifiable by Kolmogorov forward equation in the illness-death model

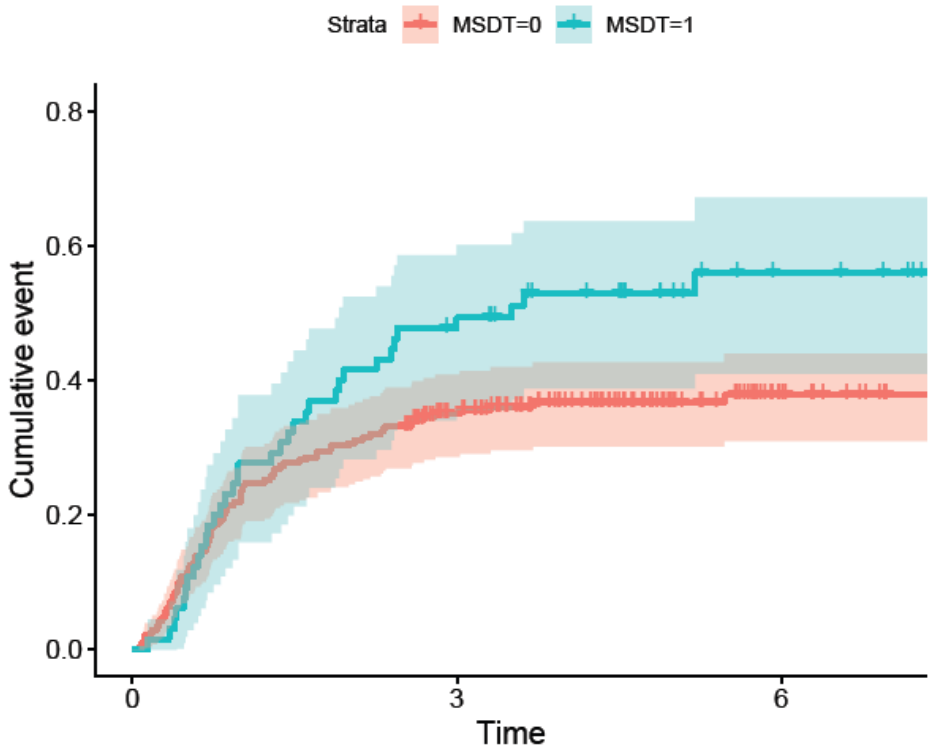
Deng, Y., Wang, Y., & Zhou, X. H. (2024). Direct and indirect treatment effects in the presence of semicompeting risks. *Biometrics*, *80*(2), ujae032.

Estimation

- A practical working assumption:
- Markovness $d\Lambda_2(t; z_2, \bar{n}_1(t), x) = d\Lambda_2(t; z_2, n_1(t), x)$
- Nonparametric estimation (without covariates):
  - Nelson-Aalen-Johansen estimator for three hazards
  - Kolmogorov forward equation in the illness-death model
- Semiparametrically efficient estimation (with covariates)
  - Semiparametric efficiency
  - Multiple robustness (propensity score, censoring hazard, cause-specific hazards)

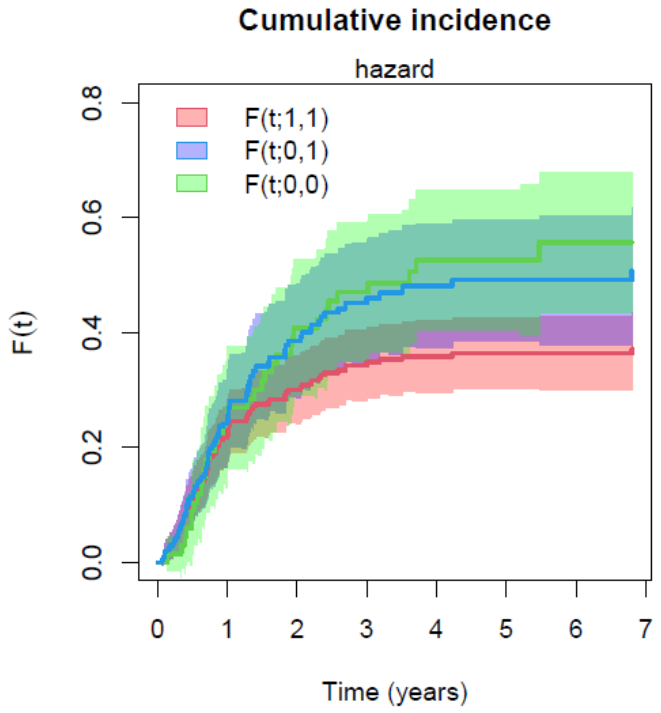
Application to stem cell transplantation

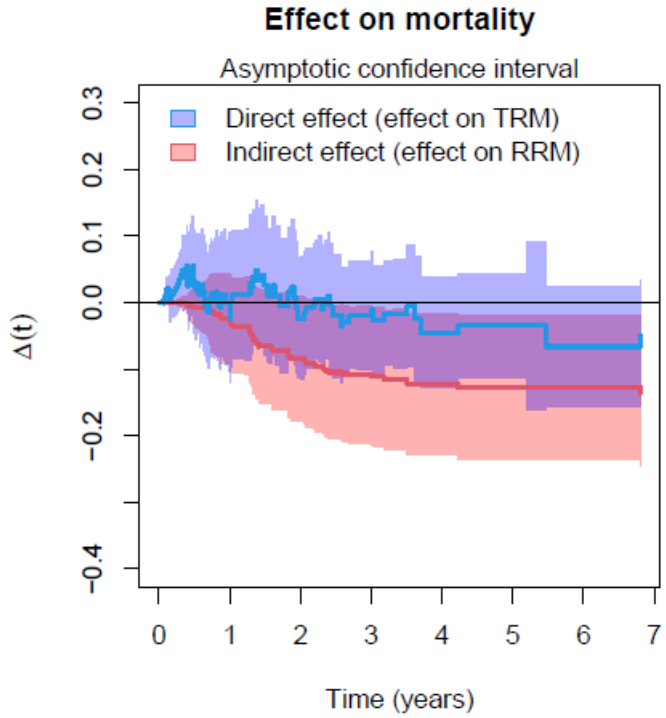
- Treatment: transplant modality (MSDT vs Haplo)
- Intermediate event: relapse
- Terminal event: death
- Why does Haplo show lower mortality?
  - Direct effect on death
  - Indirect effect via relapse to death



Application to stem cell transplantation

- Haplo has no direct effect – No effect on transplant-related mortality
- Haplo has an indirect effect – An effect on relapse-related mortality





Separable effects

- Natural effects may be hard to interpret
- If the treatment have two components
  - z_1 influences the hazard of the intermediate event
  - z_2 influences the hazard of the terminal event
- Potential counting processes: $N_1(t; z_1, z_2)$, $N_2(t; z_1, z_2)$
- Counterfactual CIF if treatment components are assigned as (z_1,z_2):

$$F_2(t; z_1, z_2) = P\{N_2(t; z_1, z_2) = 1\}$$

Breum, M. S., Munch, A., Gerds, T. A., & Martinussen, T. (2024). Estimation of separable direct and indirect effects in a continuous-time illness-death model. *Lifetime Data Analysis*, *30*(1), 143-180.

Identification of separable effects

- Ignorability
- Dismissible components condition
  - Terminal event hazard relies only on z_2
  - Intermediate event hazard relies only on z_1
- Essentially excluding treatment-induced or unmeasured confounding between counting processes
- Random censoring, positivity, consistency
- Identification formula for $F_2(t; z_1, z_2)$ identical with natural effects
- Other framework: randomized intervention

5. Principal Stratification

Principal stratum

- Consider a binary treatment Z, an outcome Y and an intermediate variable S
- Four classes of units:
  - S(1)=0, S(0)=0
  - S(1)=1, S(0)=0
  - S(1)=0, S(0)=1
  - S(1)=1, S(0)=1
- Principal stratum defined by (S(1),S(0)) – a specific class of units
- Principal stratum not observable

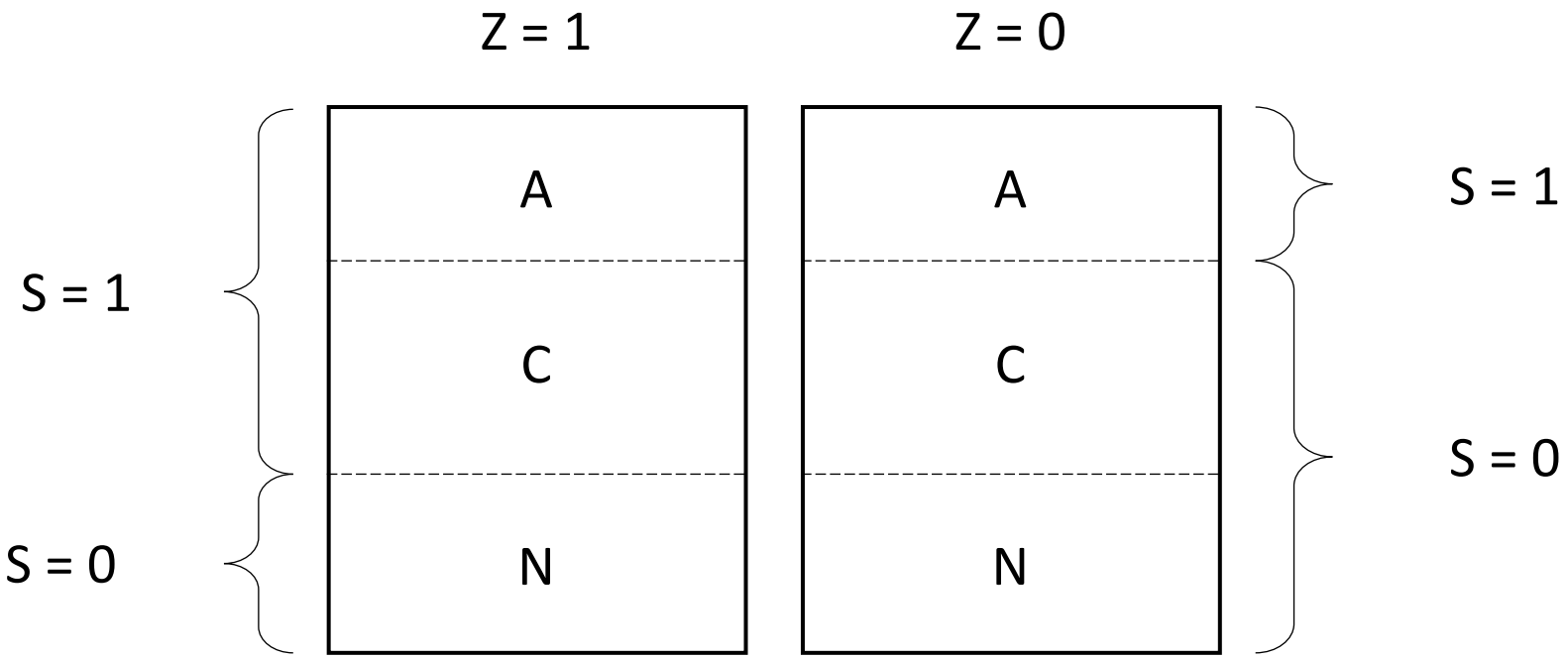
Non-compliance

- Treatment Z: assignment
- Intermediate variable S(z): actual treatment
- Four types of units (Angrist, Imbens, Rubin, 1996):
  - Complier: S(z) = z
  - Always-taker: S(z) = 1
  - Never-taker: S(z) = 0
  - Defier: S(z) = 1-z (usually excluded by monotonicity)
- Treatment effect is meaningful only in compliers
- CACE = E{Y(1) – Y(0) | G=c}

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444-455.

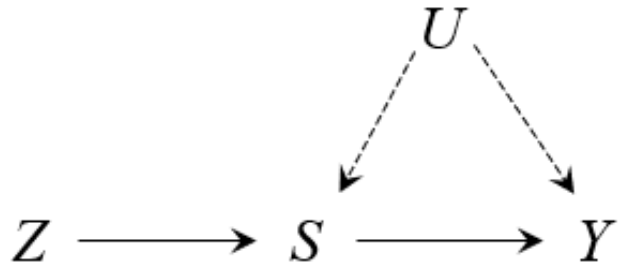
Non-compliance

- Under monotonicity, there are three principal strata
- The proportion of each principal stratum is identifiable



Complier average causal effect

- Ignorability: Z ⊥ {S(z), Y(z,s)} | X
- Relevance: E{S(1) – S(0)|X} ≠ 0
- Exclusion restriction: Y(1,s) = Y(0,s)
- Consistency: S = S(Z), Y = Y(Z) = Y(Z,S(Z))
- Monotonicity: S(1) ≥ S(0)
- Positivity
- Conditional CACE is identifiable



$$CACE(X) = \frac{E(Y|Z=1,X) - E(Y|Z=0,X)}{E(S|Z=1,X) - E(S|Z=0,X)}$$

Complier average causal effect

- CACE defined in compliers is identifiable

$$CACE = \int CACE(x)P(x|c)dx$$
$$= \frac{E_X\{E(Y|Z=1,X) - E(Y|Z=0,X)\}}{E_X\{E(S|Z=1,X) - E(S|Z=0,X)\}}$$

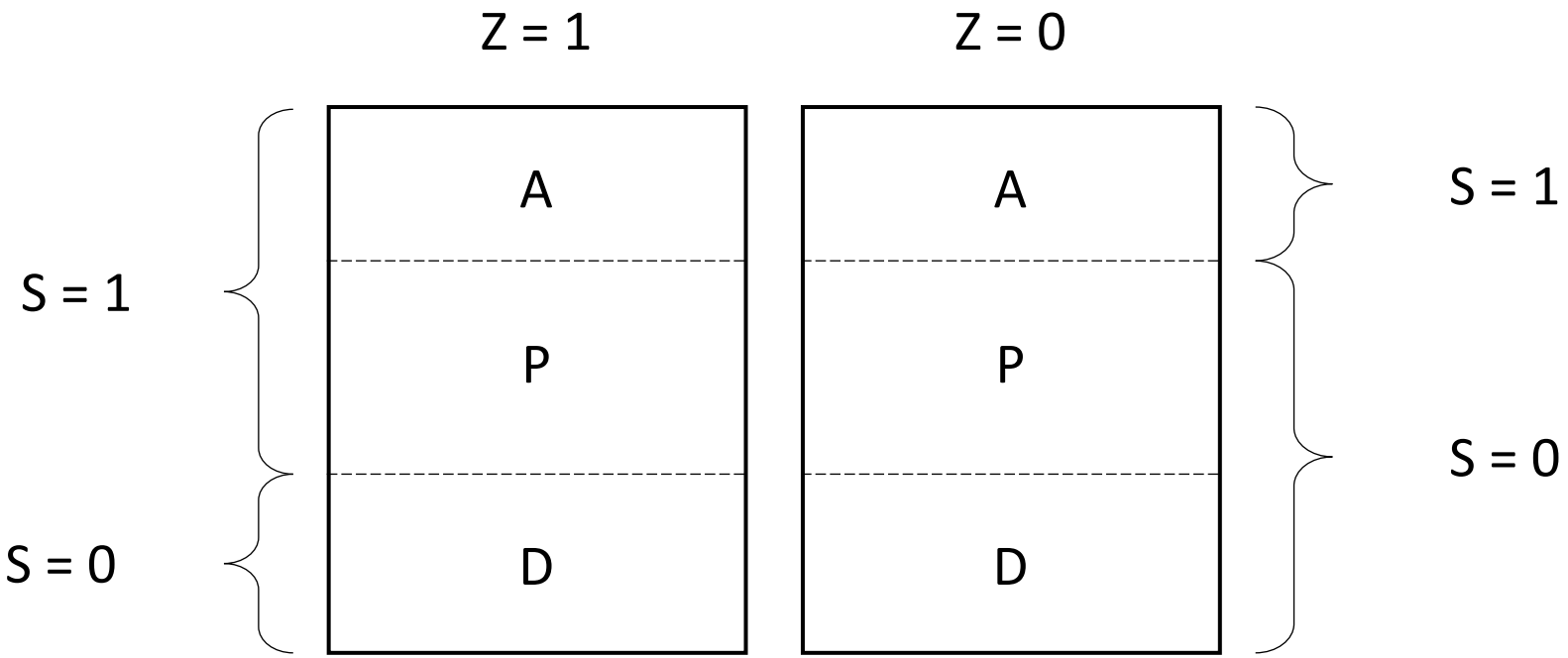
- Estimation:
  - Regression estimator (plug-in)
  - AIPW for the numerator and denominator
  - Local average response function (LARF)
  - Semiparametric efficient estimator

Truncation by death

- Treatment Z: assignment
- Intermediate variable S(z): survival indicator
- Four types of units:
  - Always-survivor: S(z) = 1
  - Protected: S(z) = z
  - Doomed: S(z) = 0
  - Harmed: S(z) = 1-z (usually excluded by monotonicity)
- Potential outcomes (Y(1),Y(0)) only well-defined in always-survivors
- SACE = E{Y(1) – Y(0) | G=a}

Truncation by death

- Under monotonicity, there are three principal strata
- The proportion of each principal stratum is identifiable



Survivor average causal effect

- Ignorability: $Z \perp S(z) \mid X$
- Latent ignorability: $Z \perp Y(z) \mid S(1)=S(0)=1, X$
- Principal ignorability: $E\{Y(1) \mid S(1)=1, S(0)=1, X\} = E\{Y(1) \mid S(1)=1, S(0)=0, X\}$
  - $E\{Y(1)\}$ is identical in always-survivors and protected groups
- Consistency: $S = S(Z)$, $Y = Y(Z)$ if $S(Z) = 1$
- Monotonicity: $S(1) \geq S(0)$
- Positivity

Ding, P., & Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(3), 757-777.

Survivor average causal effect

- Conditional SACE is identifiable by the "survivor-case" analysis
- $SACE(X) = E(Y|Z = 1, S = 1, X) - E(Y|Z = 0, S = 1, X)$
- SACE defined in always-survivors is identifiable

$$SACE = \int SACE(x)p(x|a)dx$$
$$= \frac{E_X\{SACE(X)[E(S|Z = 1, X) - E(S|Z = 0, X)]\}}{E_X\{E(S|Z = 1, X) - E(S|Z = 0, X)\}}$$

- If principal ignorability fails, use a substitutional variable to identify SACE (Wang, Zhou, Richardson, 2017)

Wang, L., Zhou, X. H., & Richardson, T. S. (2017). Identification and estimation of causal effects with outcomes truncated by death. *Biometrika*, *104*(3), 597-612.

6. Intercurrent Events

Intercurrent events

- Non-compliance
  - The actual treatment is different from the assignment
- Truncation by death
  - The primary outcome does not exist
- Intermediate events
  - Rescue mediation, treatment switch; the primary outcome is not interpretable

How to address non-compliance

- Treatment assignment Z, actual treatment D, outcome Y
- Intention-to treat: $E(Y|Z=1) – E(Y|Z=0)$
- $E\{Y(1) – Y(0)\}$ by treatment Y as a potential outcome of Z
- Randomization is the “golden standard” to evaluate the treatment effect – the treatment assignment is randomized
- Not aligned with primary interest

How to address non-compliance

- As-treated: E(Y|D=1) – E(Y|D=0)
- Per-protocol: E(Y|Z=D=1) – E(Y|Z=D=0)
- Problem: D is not randomized; unmeasured confounding between D and Y; not comparable population
- Principal stratum estimand: complier average causal effect
- E{Y(1)-Y(0) | D(1)>D(0)}
- Problem: the target population is not observable

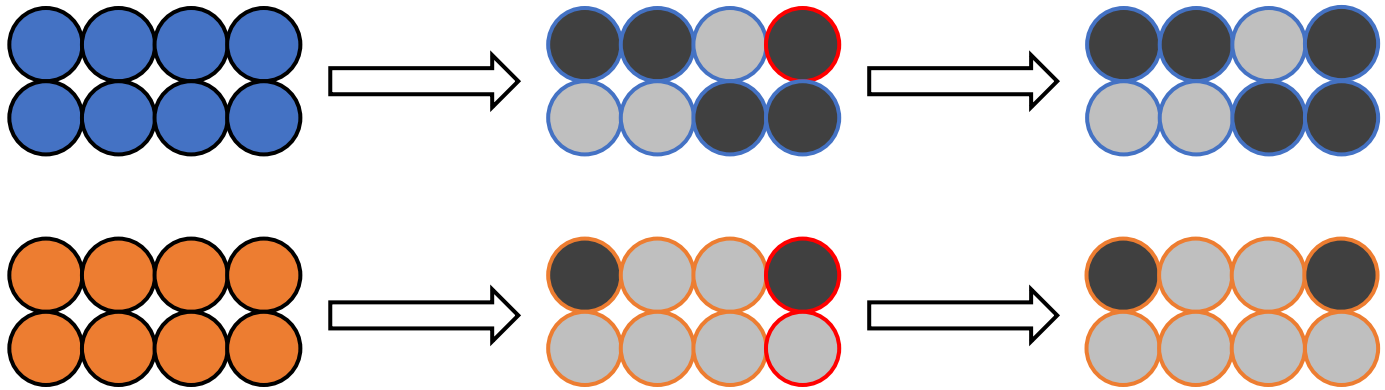
How to address intercurrent events

- Intercurrent events can either render the primary outcome not to exist or not interpretable
- ICH E9 (R1): an estimand framework
- Five strategies:
  - Treatment policy strategy
  - Composite variable strategy
  - While on treatment strategy
  - Hypothetical strategy
  - Principal stratum strategy

Han, S., & Zhou, X. H. (2023). Defining estimands in clinical trials: A unified procedure. *Statistics in Medicine*, *42*(12), 1869-1887.

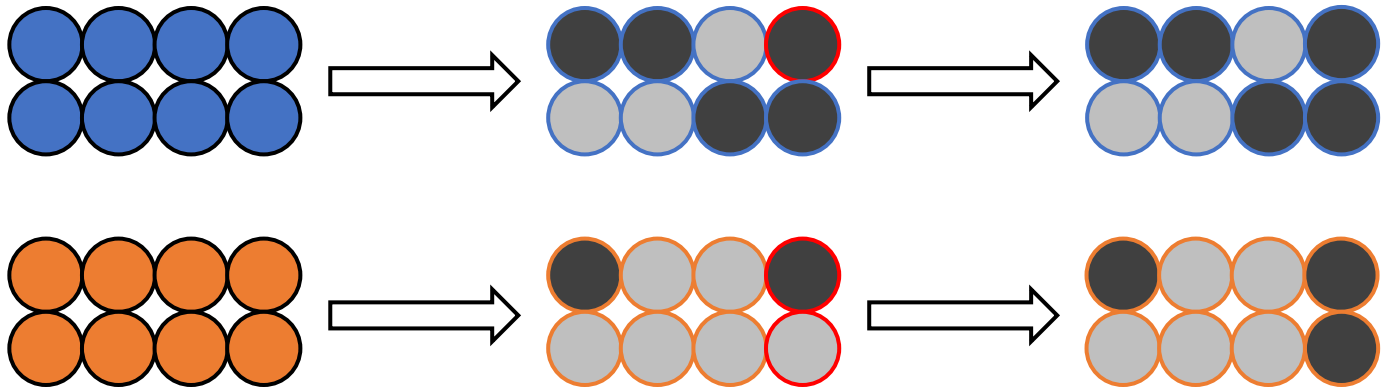
Treatment policy strategy

- Treatment policy strategy is closely related to intention-to-treat (ITT) analysis
- The occurrence of intercurrent events is treated as part of the treatment
- E{Y(1)-Y(0)} = E{Y(1,S(1))-Y(0,S(0))}



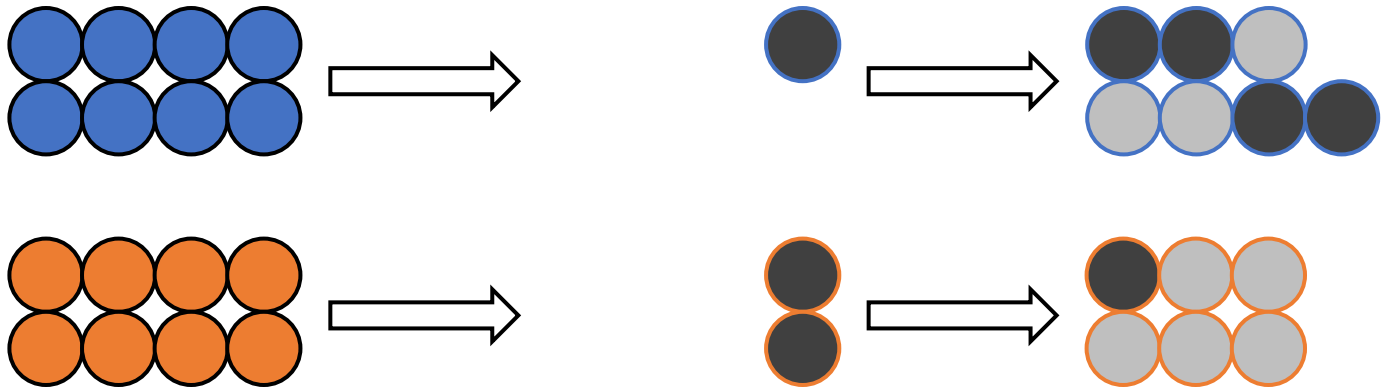
Composite variable strategy

- The occurrence of intercurrent events is treated as part of the outcome
- Typical composite outcomes: f(S,Y) = S*Y, aS+bY
  - Binary S and Y: occurrence of intercurrent events indicates treatment failure
- E{f(S(1),Y(1))-f(S(0),Y(0))}



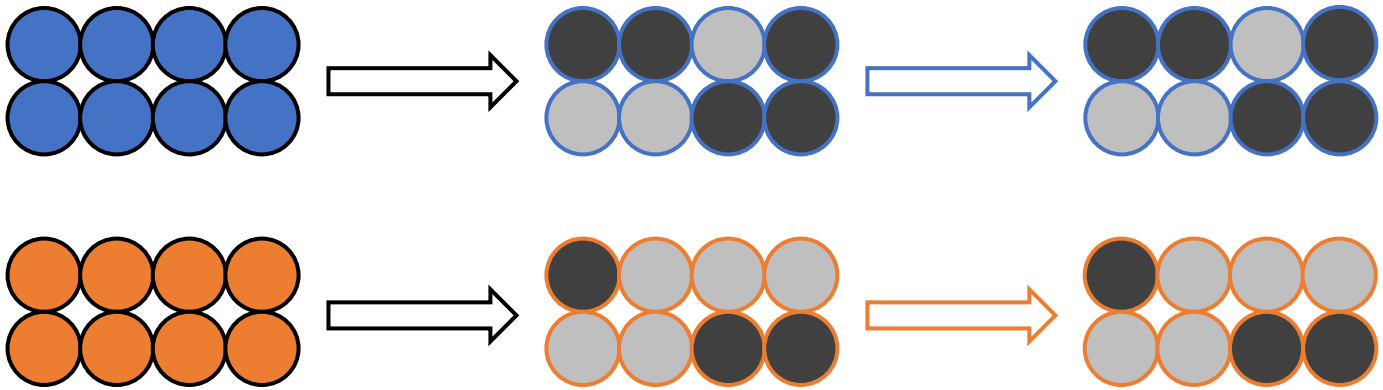
While on treatment strategy

- While on treatment strategy requires longitudinal observations
- Use the last observable outcome before intercurrent events
- E{Y(1;T(1))-Y(0;T(0))}
- Not reliable if there is systematic risk difference treatment groups



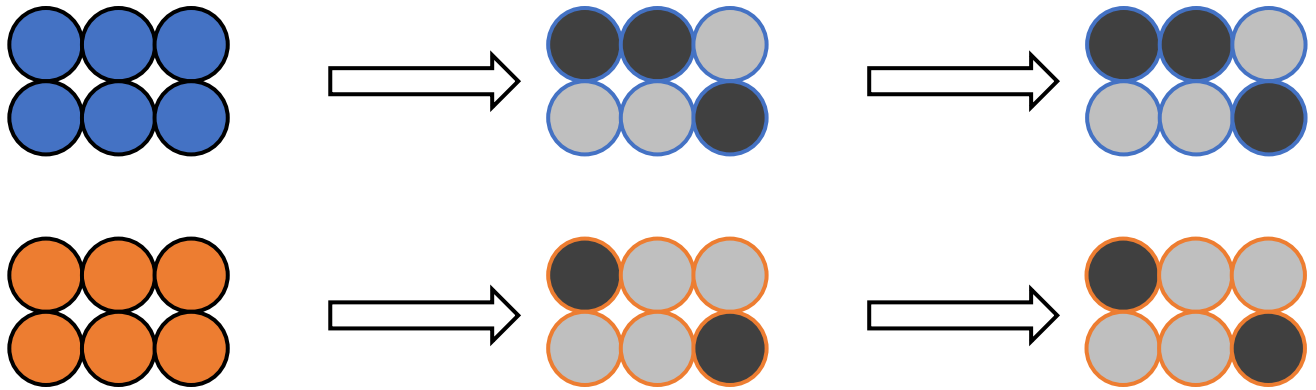
Hypothetical strategy

- To envision a hypothetical scenario where intercurrent events do not occur
- Hypothetical strategy is closely related to mediation analysis
- E{Y*(1)-Y*(0)} = E{Y(1,s=0)-Y(0,s=0)}
- Additional assumptions are required to identify this estimand



Principal stratum strategy

- The target population is restricted to a principal stratum
- The principal stratum is not observable, so interpretation can be hard
- Example: complier/survivor average causal effect
- Additional assumptions are required to identify this estimand



Time-to-event endpoint

- Now suppose the endpoint is time-to-event
- Data structures
  - Semi-competing risks: the primary outcome is not prevented by intercurrent events
  - Competing risks: the primary outcome is prevented by intercurrent events
- Potential outcomes
  - Time to primary outcome: T(z)
  - Time to (first) intercurrent event: R(z)
- Estimand: cumulative incidence function (CIF)

Deng, Y., Han, S., & Zhou, X. H. (2025). Inference for Cumulative Incidences and Treatment Effects in Randomized Controlled Trials With Time-to-Event Outcomes Under ICH E9 (R1). *Statistics in Medicine*, *44*(10-12), e70091.

Treatment policy strategy

- Comparing the CIF of the primary event as recorded
- P{T(1)<t} – P{T(0)<t} = P{T(1,R(1))<t} – P{T(0,R(0))<t}
- Only applicable to semi-competing risks data (otherwise potential outcomes are not well defined)
- Required observed data: T, Δ_T
- $F(t;z) = 1 - \exp\{-\Lambda(t;z)\}$

Composite variable strategy

- The composite outcome is the first occurrence of any event: f(T,R) = min(T,R)
  - Other form: quality adjusted survival time
- P{T(1)<t, R(1)<t} – P{T(0)<t, R(0)<t}
- Applicable to both semi-competing and competing risks data
- Required observed data: min(T,R), max(Δ_T,Δ_R)
- $F(t;z) = 1 - \exp\{-\Lambda_1(t;z) - \Lambda_2(t;z)\}$

While on treatment strategy

- Primary outcome after intercurrent events are not counted – leads to the subdistribution function of the primary outcome
- P{T(1)<t, R(1)>t} – P{T(0)<t, R(0)>t}
- Required observed data: min(T,R), Δ_T, Δ_R
- $F(t;z) = \int_0^t \exp\{-\Lambda_1(t;z) - \Lambda_2(t;z)\} d\Lambda_1(t;z)$

Hypothetical strategy

- Scenario I: Intercurrent events allowed only if they occur under control
- P{T(1,R(0))<t} – P{T(0,R(0))<t}
- $F(t;z) = \int_0^t \exp\{-\Lambda_1(t;z) - \Lambda_2(t;0)\} d\Lambda_1(t;z)$
- Required observed data: min(T,R), Δ_T, Δ_R
- Related to natural direct effect
- Sequential ignorability for causal interpretation

Hypothetical strategy

- Scenario II: Intercurrent events not allowed
- P{T(1,∞)<t} – P{T(0,∞)<t}
- $F(t;z) = \int_0^t \exp\{-\Lambda_1(t;z)\} d\Lambda_1(t;z) = 1 - \exp\{-\Lambda_1(t;z)\}$
- Required observed data: min(T,R), $\Delta_T(1-\Delta_R)$
- Related to controlled direct effect
- Sequential ignorability for causal interpretation

Principal stratum strategy

- The target population is restricted to a subpopulation who would never experience intercurrent events
- P{T(1)<t | R(1)=R(0)=∞} – P{T(0)<t | R(1)=R(0)=∞}
- For identification, use the potential occurrence at the end of the study t*: P{T(1)<t | R(1)>t*, R(0)>t*} – P{T(0)<t | R(1)>t*, R(0)>t*}
- $F(t;z) = \frac{\int_0^t \exp\{-\Lambda_1(t;z)-\Lambda_2(t;0)\}d\Lambda_1(t;z)}{1-\int_0^{t^*} \exp\{-\Lambda_1(t;z)-\Lambda_2(t;0)\}d\Lambda_2(t;z)}$
- Required observed data: min(T,R), Δ_T, Δ_R
- Principal ignorability for identification

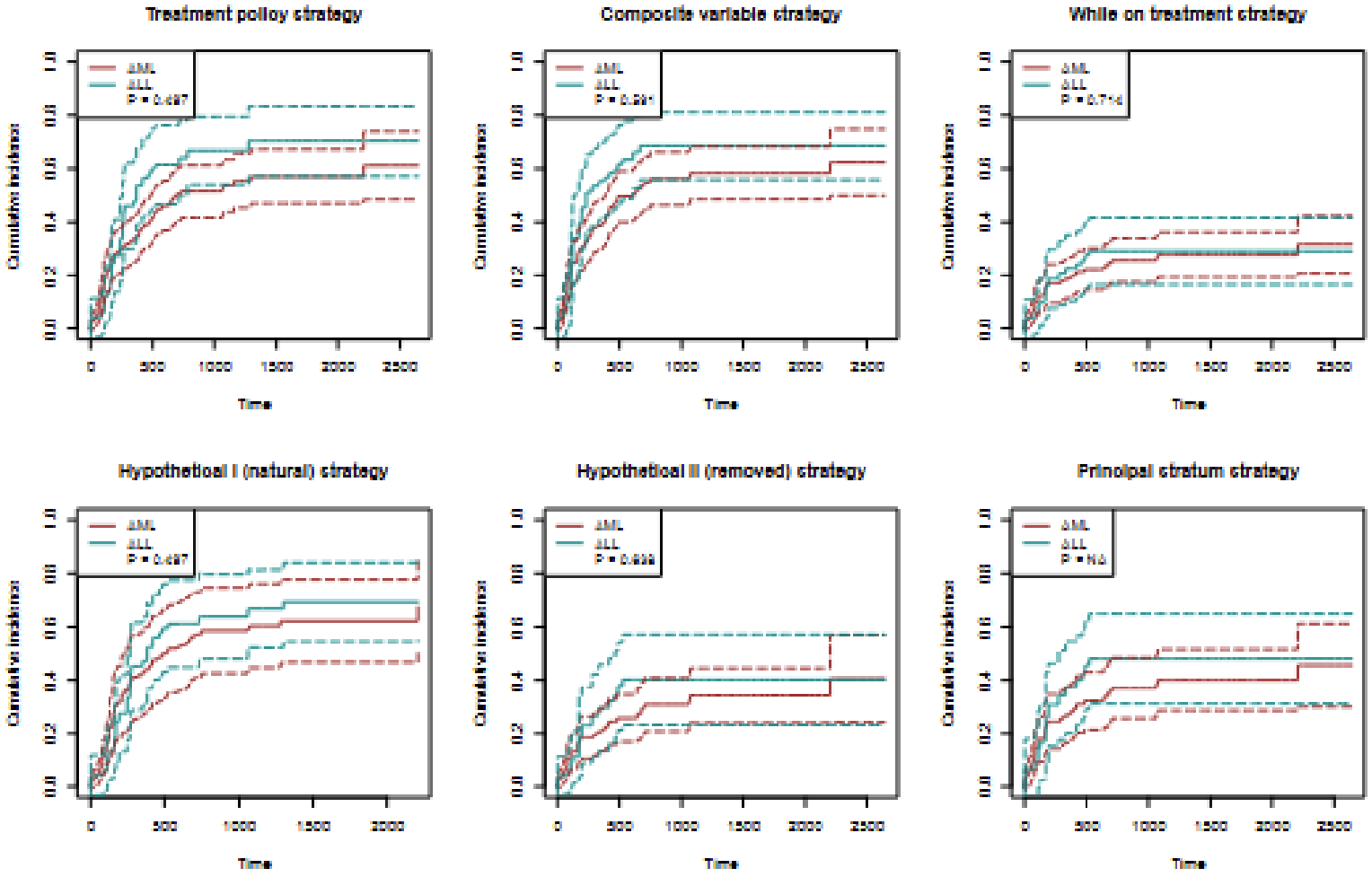
Observational studies

- Nonparametric estimation in randomized controlled trials
  - Nelson-Aalen-Johansen estimator for hazards
  - Markovness required for natural effects with semi-competing risks data
- Inverse probability weighting
  - To weight event and at-risk processes by the inverse of propensity score
  - Uncertainty of the propensity score not accounted for
- Semiparametric efficient estimation
  - To derive the efficient influence function
  - Double robustness
  - Standard error based on influence functions

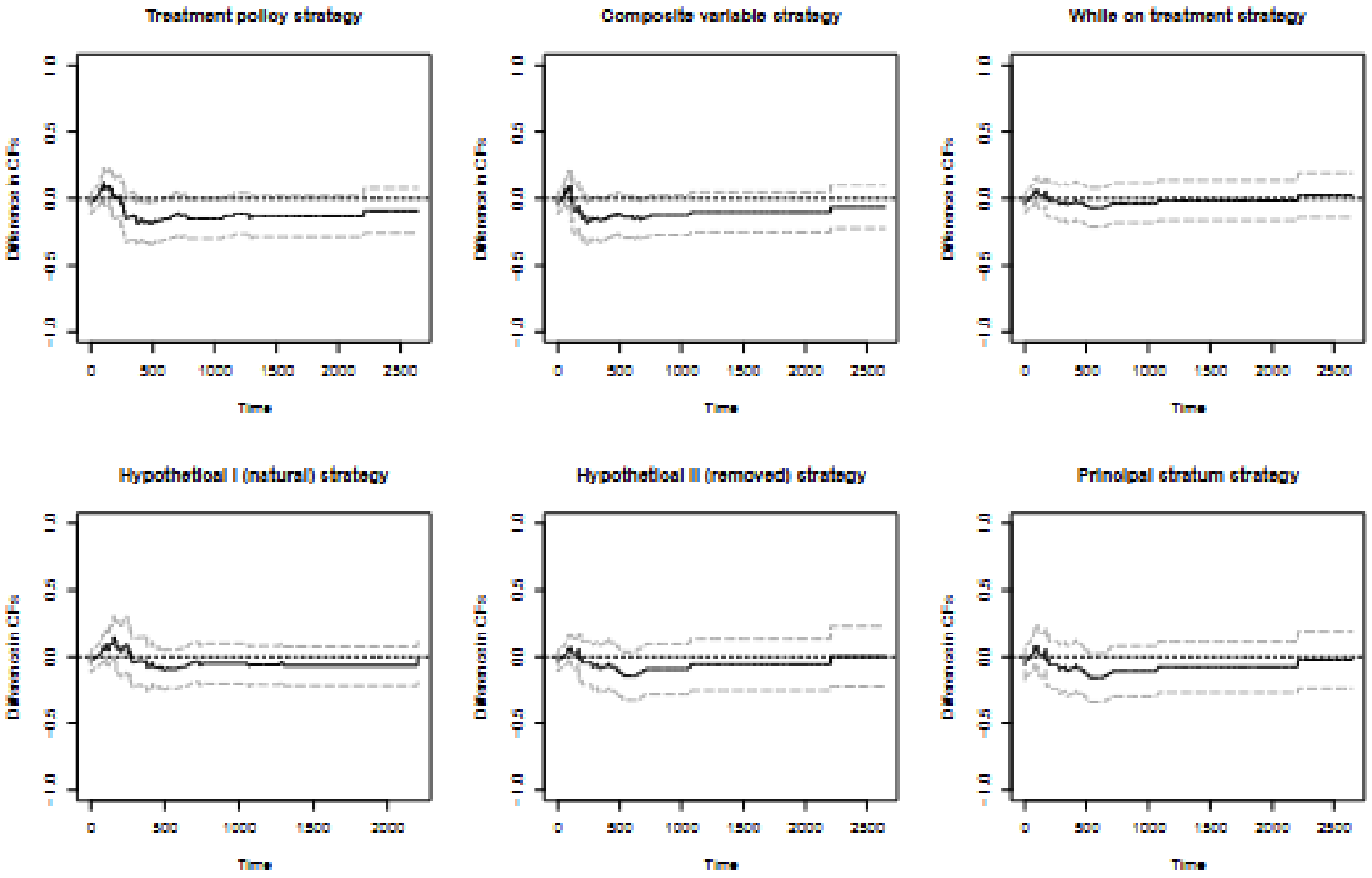
Example: Bone marrow transplantation data

- Data source: Klein and Moeschberger (1997)
- Sample size: 137
- Treatment: AML (Z=1) vs ALL (Z=0)
- Primary outcome event: death
- Intercurrent event: relapse
- Data structure: semi-competing risks
- Covariates: sex, age, Cy status
- R package: tteICE

Results: counterfactual CIFs



Results: treatment effects



7. Observational Studies: Non-Ignorability

Sensitivity analysis

- If ignorability fails, E{Y(z) | Z=1, X} ≠ E{Y(z) | Z=0, X}
- Suppose there is an unmeasured confounder U and Y(z) ⊥ Z | X, U
- Observed propensity score: $\hat{e}_i = P(Z_i = 1|X_i)$
- True propensity score: $\hat{p}_i = P(Z_i = 1|X_i, U_i)$
- Rosenbaum condition:

$$\left|\log\left(\frac{\hat{e}_i}{1-\hat{e}_i}\right) - \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right| \leq \Gamma, \qquad \forall i$$

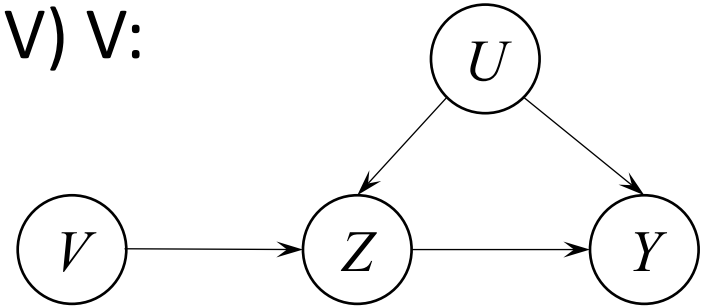
- Bounds for the P-value under a given Γ
- Design sensitivity $\tilde{\Gamma}$: for $\Gamma < \tilde{\Gamma}$, power goes to 1; otherwise 0

Instrumental variables

- How to estimate the effect of Z on Y?
- Suppose a linear model, homogeneous effect:

$$Y_i = \alpha + \tau Z_i + U_i$$

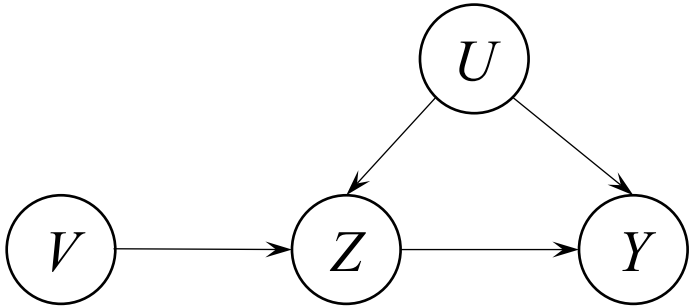
- Three conditions of an instrumental variable (IV) V:
  - Randomization: cov(V,U|X) = 0
  - Relevance: cov(V,Z|X) ≠ 0
  - Exclusion restriction: no effect of V on Y



- Two stage least squares:
  - Regress Z on V: $\hat{Z}$ removes the effect of U on V
  - Regress Y on $\hat{Z}$: the coefficient is τ

Instrumental variables

- In potential outcomes framework, suppose V is binary
- Three conditions for an IV:
  - Randomization: $V \perp (Z(v), Z(v), Y(v,d)) \mid X$
  - Relevance: $E\{Z(1)-Z(0)|X\} \neq 0$
  - Exclusion restriction: $Y(0,z) = Y(1,z) =: Y(z)$
- Other assumptions: positivity, consistency
- Example: encouragement design



The fourth condition for an IV

- A “homogeneity” condition of the treatment effect
  - Constant effect: $Y(1) - Y(0) = \tau$
  - $E\{Y(1) - Y(0)|V, Z = z, X\}$ does not depend on V
  - $E\{Y(1) - Y(0)|U, X\}$ does not depend on U
  - $E\{Z(1) - Z(0)|U, X\}$ does not depend on U
- Under the homogeneity condition, conditional ATE is identifiable

$$E\{Y(1) - Y(0)|X\} = \frac{E(Y|V = 1, X) - E(Y|V = 0, X)}{E(Z|V = 1, X) - E(Z|V = 0, X)}$$

- Numerator: intention-to-treat effect

The fourth condition for an IV

- An alternative to homogeneity condition
  - Monotonicity: $Z(1) \geq Z(0)$
  - Example: patients in the control group do not have access to the treatment
- Three principal strata under monotonicity:
  - Always-taker: $Z(1) = Z(0) = 1$
  - Complier: $Z(1) = 1, Z(0) = 0$
  - Never-taker: $Z(1) = 0, Z(0) = 0$
- Treatment effect not identifiable in always-takers and never-takers

Local average treatment effect

- Under monotonicity, conditional effect in compliers

$$E\{Y(1) - Y(0)|Z(1) > Z(0), X\} = \frac{E(Y|V = 1, X) - E(Y|V = 0, X)}{E(Z|V = 1, X) - E(Z|V = 0, X)}$$

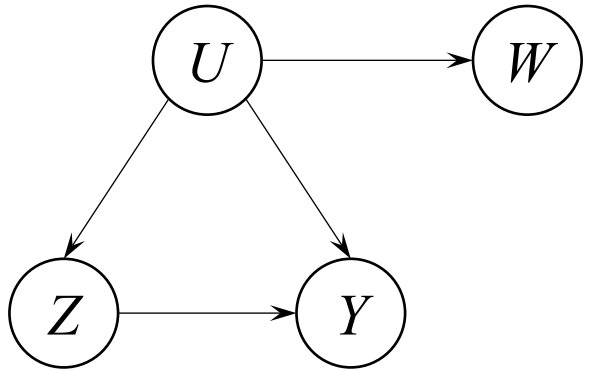
- Local average treatment effect (LATE)

$$E\{Y(1) - Y(0)|Z(1) > Z(0)\} = \frac{E\{E(Y|V = 1, X) - E(Y|V = 0, X)\}}{E\{E(Z|V = 1, X) - E(Z|V = 0, X)\}}$$

- Compliers are not observable

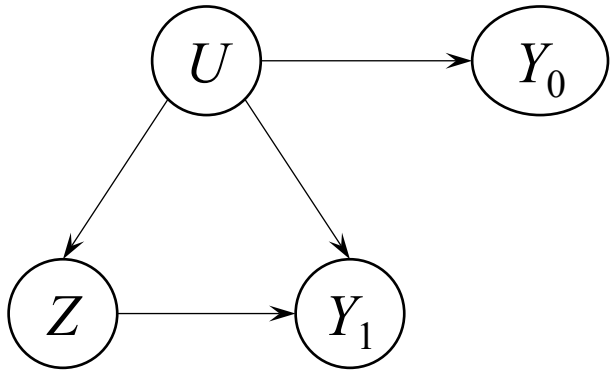
Negative control variables

- Negative control outcome (NCO) W
  - Latent unconfoundedness: Y(z) ⊥ Z |U, X
  - No effect of Z on W: Z ⊥ W | U, X
  - Known U-W and U-Y relation
- Example
  - Z: COVID-19 vaccine
  - Y: COVID-19 infection
  - W: flu infection
- In a linear model, $\beta_{ZY} = \beta_{Y|Z,X} - \frac{\alpha_{UY}}{\alpha_{UW}}\beta_{W|Z,X}$



Shi, X., Miao, W., & Tchetgen, E. T. (2020). A selective review of negative control methods in epidemiology. *Current epidemiology reports*, *7*(4), 190-202.

Difference-in-differences



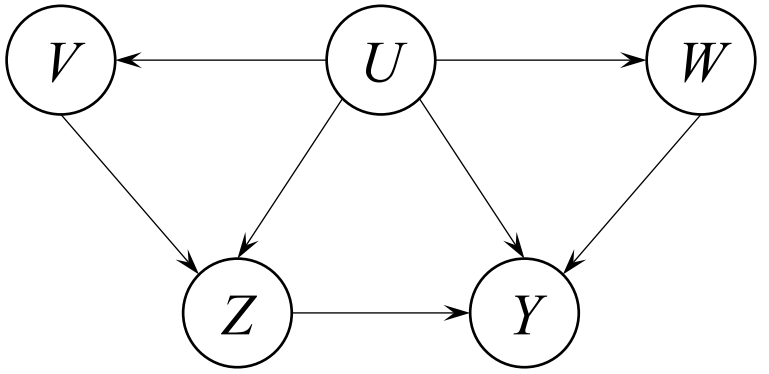
- Pre-treatment outcome Y_0 is an NCO
- No anticipation: $Y_0(1) = Y_0(0)$
- Parallel trends: $E\{Y_1(0)-Y_0(0) \mid Z=1, X\} = E\{Y_1(0)-Y_0(0) \mid Z=0, X\}$
- Other assumptions: Positivity, consistency
- Average treatment effect on the treated (ATT)

$$E\{Y_1(1) - Y_1(0)|Z = 1\}$$
$$= E(Y_1 - Y_0|Z = 1) - E\{E(Y_1 - Y_0|Z = 0, X)|Z = 1\}$$

- Doubly robust and efficient estimator available

Proximal causal inference

- Latent ignorability: $Z \perp Y(z) \mid U, X$
- Negative control exposure (NCE) V: $V \perp (Y, W) \mid Z, U, X$
  - No effect of V on (Y, W)
- Negative control outcome (NCO) W: $W \perp Z \mid U, X$
  - No effect of Z on W



Cui, Y., Pu, H., Shi, X., Miao, W., & Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, *119*(546), 1348-1359.

Proximal causal inference

- Completeness: E{g(U)|Z,V,X} = 0 implies g(U) = 0 almost surely
  - V contains all variation in U
- Outcome bridge function: $E\{q(z, W, X)|v, z, X\} = E(Y|v, z, x)$
- If U is discrete, then V and W should have more levels than U (Miao et al., 2018)
- ATE is identifiable,

$$E\{Y(z)\} = E\{q(z, W, X)|Z = z, X\}$$

- Also identifiable through a treatment bridge function
- So doubly robust estimation is possible

Miao, W., Geng, Z., & Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, *105*(4), 987-993.

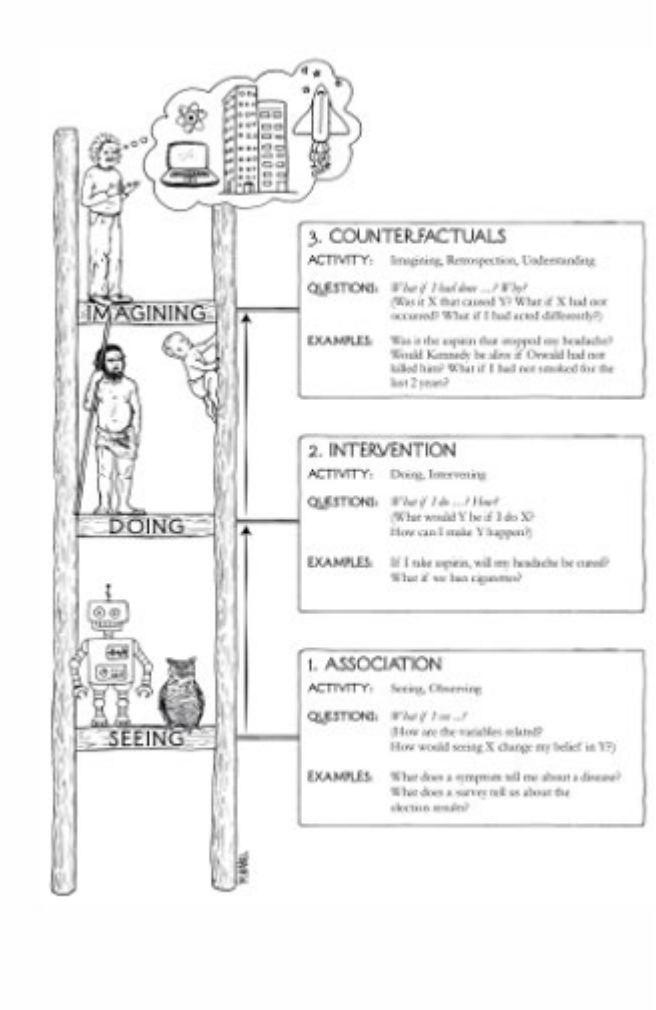
Outline

- Causal Attribution in Complex Scenarios
- Estimating Causal Effects with Deep Learning
- Application in Real-World Scenarios

Outline

- Causal Attribution
- Three levels of Causal Inference
- Basic Attribution Probabilities, Identification
- Modern Attribution Methods

Three Levels of Causal Inference



Three levels of causal reasoning:

1 Association — What do we observe?

2 Intervention — What changes if we act?

3 Counterfactuals — What would have happened otherwise?

Why Attribution Is a Level-3 Question

Association: Areas with higher vaccination rates may also show more severe outbreaks.

Intervention: Would vaccination reduce infections?

Counterfactual: For an infected person, would the outcome have changed under another vaccination decision?

Attribution asks the counterfactual question for the same observed case.

Outline

- Causal Attribution
- Three levels of Causal Inference
- Basic Attribution Methods, Probabilities, Identification
- Modern Attribution Methods

EoC vs. CoE

- "Effect of Cause" vs. "Cause of Effect"

'Effect of Cause (EoC)' or 'Cause of Effect (CoE)'

EoC: Does smoking harm lung cancer?

Question of intervention effect – Level 2

- CoE: Did smoking cause Zhang San's lung cancer? – Problem of Attribution.

Question of attribution – Level 3

- CoE cannot be directly inferred from population statistical evidence,
- as it involves individual counterfactuals.

Attribution Methods

- CoE: Did smoking cause Zhang San's lung cancer? – Problem of Attribution.
- Question of attribution – Level 3

1. Bayesian Method: Calculate conditional probability $P(X_k = 1 \mid Y = 1)$ using prior probabilities
2. Infer causes based on causal effects
3. Infer causes based on causal paths
4. Posterior causal effects

Use Bayesian Method to Infer CoE

- Probability of disease $X_{\mathrm{k}} = 1$ given symptom $Y = 1$:

$$P(X_{\mathrm{k}} = 1|Y = 1) = \frac{P(X_{\mathrm{k}} = 1)P(Y = 1|X_{\mathrm{k}} = 1)}{P(Y = 1)}$$

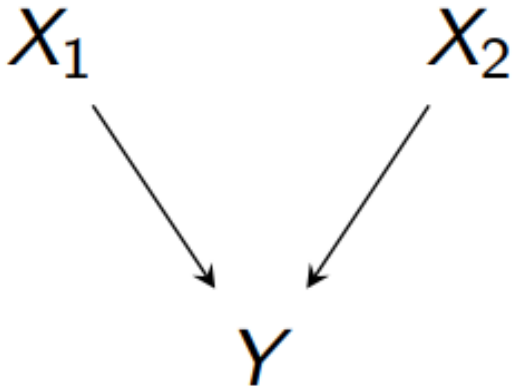
It depends on prior probability $P(X_{\mathrm{k}} = 1)$

- When $P(X_{\mathrm{k}} = 1 \mid Y = 1)$ is large, $P(X_{\mathrm{k}} = 1 \mid Y = 0)$ may also be large. $ACE(X_{\mathrm{k}} \rightarrow Y)$ could be tiny or even equal to 0, which means disease $X_{\mathrm{k}} = 1$ is not the cause of symptom $Y = 1$.
- In fact, $P(X_{\mathrm{k}} = 1|Y = 1)$ is within level 1, cannot be directed applied for Attribution.

Infer causes based on causal effects

Average Causal Effect: $ACE(X \to Y) = E(Y_{x=1} - Y_{x=0})$

- X_1 = poison, X_2 = food; both cause outcome Y (Survival status)
- $ACE(X_1 \to Y) \gg ACE(X_2 \to Y)$. However, if a person dies ($Y = 1$) with no evidence of poison intake, death cannot be attributed to poison X_1, as the prior probability of taking poison is extremely low.
- ACE is within Level 2, cannot actually represents CoE.
- Causal effects cannot be directly applied for attribution.



Level 3: Counterfactual Reasoning

Fact: We observe:

Zhang San smoked $(T = 1)$ and developed lung cancer $(Y = 1)$.

Counterfactual: If Zhang San did not smoke $(T = 0)$,

would he have developed cancer? What is Y_0?

Note: Zhang San being a non-smoker is different from “people who are non-smokers”, because those "non-smoking people" might have avoided cancer even if they smoked!

Probability of Causation (PC)

- $Y_1 = 1$: Lung cancer develops under smoking
- $Y_0 = 0$: No lung cancer without smoking

- Dawid et al. (2016): Probability of Causation (PC)

$$PC = P(Y_0 = 0 \mid Y_1 = 1)$$

Dawid, A. P., Musio, M., & Fienberg, S. E. (2016). From statistical evidence to evidence of causality. Bayesian Analysis, 11(3), 725-752.

Probability of Causation (PC)

- Core challenge: How to identify the joint distribution of Y_1 and Y_0 from observed data $P(Y_1 = y, Y_0 = y')$
- Level 2: population-level inference while the second tier targets population-level inference, needs only marginal distributions $P(y_1)$ and $P(y_0)$.
- Level 3: individual-level inference, requires joint distribution $P(y_1, y_0)$, generally harder than level 2.

Dawid, A. P., Musio, M., & Fienberg, S. E. (2016). From statistical evidence to evidence of causality. Bayesian Analysis, 11(3), 725-752.

Probability of Causation (PC)

- PC is hard to identify precisely, since counterfactuals are unobservable.
- Hence, attribution conclusions usually fall in an interval.
- E.g. Dawid et al. (2016)

$$PC = \frac{P(Y_0 = 0, Y_1 = 1)}{P(Y_1 = 1)} = \frac{P(Y_1 = 1) - P(Y_0 = 1, Y_1 = 1)}{P(Y_1 = 1)}$$
$$\geq \frac{P(Y_1 = 1) - P(Y_0 = 1)}{P(Y_1 = 1)} = 1 - \frac{1}{CRR},$$

- Here $CRR = \frac{P(Y_1=1)}{P(Y_0=1)}$
- They also derived proper upper bound.

Dawid, A. P., Musio, M., & Fienberg, S. E. (2016). From statistical evidence to evidence of causality. Bayesian Analysis, 11(3), 725-752.

Probability of Causation (PC)

- California silicone breast implant liability litigation (C.D. Cal, 2004): When analyzing causation, the relative risk (CRR) must exceed 2.0 to be admissible to juries.

Other Attribution Probabilities

- Probability of Necessity

$$PN = P(Y_{x=0} = 0 | Y = 1, X = 1)$$

- Probability of Sufficiency

$$PS = P(Y_{x=1} = 1 \mid X = 0, Y = 0)$$

- Probability of Necessity and Sufficiency

$$PNS = P(Y_{x=0} = 0, Y_{x=1} = 1)$$

- Probability of Disablement:

$$PD = P(Y_{x=0} = 0 | Y = 1)$$

- Probability of Enablement:

$$PE = P(Y_{x=1} = 1 | Y = 0)$$

Pearl, Judea. “Probabilities of Causation: Three Counterfactual Interpretations and Their Identification.” Synthese, vol. 121, no. 1, 1999, pp. 93–149.

Identifying PN

- Under Ignorability Assumption ($Y_x \perp\!\!\!\perp X$) and Monotonicity Assumption ($Y_1 \geq Y_0$), PN = $P(Y_{x=0} = 0 | Y = 1, X = 1)$ can be identified as

$$
\begin{aligned}
PN &= \frac{P(Y_0 = 0, Y_1 = 1 | X = 1)}{P(Y = 1 | X = 1)} \\
&= \frac{P(Y_1 = 1 | X = 1) - P(Y_0 = 1, Y_1 = 1 | X = 1)}{P(Y = 1 | X = 1)} \\
&= \frac{P(Y = 1 | X = 1) - P(Y_0 = 1 | X = 1)}{P(Y = 1 | X = 1)} \\
&= \frac{P(Y = 1 | X = 1) - P(Y = 1 | X = 0)}{P(Y = 1 | X = 1)} \\
&\quad + \frac{P(Y = 1 | X = 0) - P(Y_0 = 1 | X = 1)}{P(Y = 1 | X = 1)}
\end{aligned}
$$

Identification of PS and PNS

- Under Ignorability Assumption ($Y_x \perp\!\!\!\perp X$) and Monotonicity Assumption ($Y_1 \geq Y_0$), PS can be identified as

$$PS = \frac{P(Y = 1|X = 1) - P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)}$$

- Similarly, PNS can be identified as

$$PNS = P(Y = 1|X = 1) - P(Y = 1|X = 0)$$

- They also convert joint attribution probabilities into estimable marginal probabilities.

Tian, Jin, and Judea Pearl. “Probabilities of Causation: Bounds and Identification.” Annals of Mathematics and Artificial Intelligence, vol. 28, no. 1–4, 2000, pp. 287–313.

Conditional PN, PS and PNS

- With background variable H, Tian et al. (2000) further defines conditional PN, PS and PNS:

$$PN(X \Rightarrow Y|H = h) = P(Y_{x=0} = 0|Y_{x=1} = 1, X = 1, H = h)$$

$$PS(X \Rightarrow Y|H = h) = P(Y_{x=1} = 1|Y_{x=0} = 0, X = 0, H = h)$$

$$PNS(X \Rightarrow Y|H = h) = P(Y_{x=0} = 0, Y_{x=1} = 1|H = h)$$

Tian, Jin, and Judea Pearl. “Probabilities of Causation: Bounds and Identification.” Annals of Mathematics and Artificial Intelligence, vol. 28, no. 1–4, 2000, pp. 287–313.

Identification of Conditional PS and PNS

- Assumptions: monotonicity ($Y_{x=0} \leq Y_{x=1}$), conditional unconfoundedness ($Y_x \perp\!\!\!\perp X \mid H$)
- Identification:

$$PN(X \Rightarrow Y|H = h) = 1 - \frac{P(Y = 1|X = 0, H = h)}{P(Y = 1|X = 1, H = h)}$$

$$PS(X \Rightarrow Y \mid H = h) = 1 - \frac{1 - P(Y = 1 \mid X = 1, H = h)}{1 - P(Y = 1 \mid X = 0, H = h)}$$

$$PNS(X \Rightarrow Y \mid H = h) = P(Y = 1 \mid X = 1, H = h) - P(Y = 1 \mid X = 0, H = h)$$

Tian, Jin, and Judea Pearl. “Probabilities of Causation: Bounds and Identification.” Annals of Mathematics and Artificial Intelligence, vol. 28, no. 1–4, 2000, pp. 287–313.

Outline

- Causal Attribution
- Three Levels of Causal Inference
- Overview Attribution Methods, Probabilities, Identification
- Modern Attribution Methods

Path Attribution

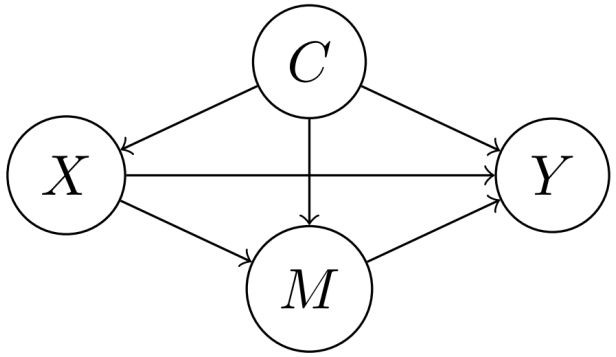


Figure 1: A causal graph representing SCM $\mathcal{M}$.

- Standard Attribution asks: Does X cause Y?
- Yet mediators widely exist in practice:

  $X \rightarrow M \rightarrow Y$
- Kawakami et al. (2024) integrates mediation analysis into Attribution.
- Goal: Distinguish whether the treatment yields the outcome directly or via a mediator.

Kawakami, Yuta, and Tian, Jin. “Mediation Analysis for Probabilities of Causation.” arXiv preprint arXiv:2412.14491, 2024.

Path Attribution

Similar to NDE, NIE in Mediator Analysis, Kawakami et at. (2024) defines:

ND-PNS: Attribution effect along the direct path

NI-PNS: Attribution effect along the mediating path

T-PNS: Total Probability of Necessity and Sufficiency (total attribution probability)

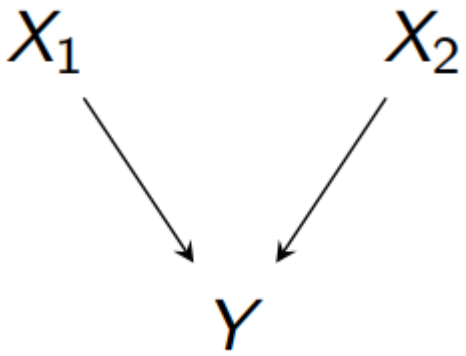
- We have T-PNS = ND-PNS + NI-PNS

- It further decomposes the overall causal contribution into direct-path and indirect-path attribution components, and propose identification conditions.

Kawakami, Yuta, and Tian, Jin. “Mediation Analysis for Probabilities of Causation.” arXiv preprint arXiv:2412.14491, 2024.

Attribution with Two Causes

- **Classic Methods' Limitation**: Pearl's PN/PS/PNS models a single binary exposure/treatment.
- **Interactive Exposures**: Real-world outcomes often arise from double interactive exposures (e.g., Smoking X_1+ Habit X_2 $\rightarrow$ Lung Cancer Y, the effect of $X_1$and the effect X_2 are not independent.)



Luo, Shanshan, Wei Li, Xueli Wang, Shaojie Wei, and Zhi Geng. "Assessing Interactive Causes of an Occurred Outcome Due to Two Binary Exposures." Biometrics, vol. 82, no. 2, 2026.

Attribution with Multiple Causes

- **Interactive Causal Types**: Under monotonicity, 16 counterfactual types reduce to 6 biologically plausible types.

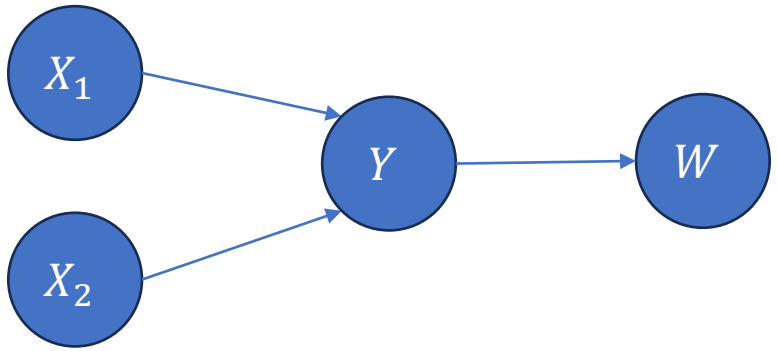
- **Under-Identification**: 6 latent types (5 free parameters) yield only 4 observable marginal probabilities. Point identification remains impossible even in RCTs.

A Downstream Outcome Adds Information

Core idea: observe a secondary outcome W after the primary outcome Y.

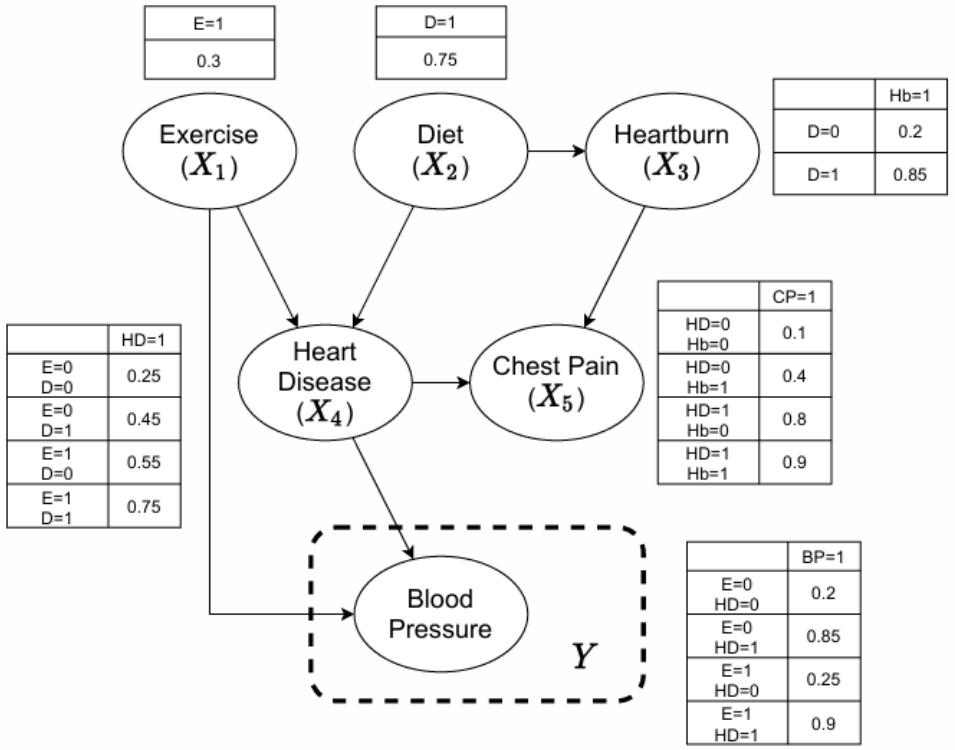
Examples: a post-cancer biomarker or a downstream clinical symptom.

W provides extra observed categories that help reconstruct the joint counterfactual distribution.



Luo, Shanshan, Wei Li, Xueli Wang, Shaojie Wei, and Zhi Geng. “Assessing Interactive Causes of an Occurred Outcome Due to Two Binary Exposures.” Biometrics, vol. 82, no. 2, 2026.

Attribution With Multiple Causes



- In practice, outcomes are usually jointly affected by multiple risk factors: $X = (X_1, \ldots, X_p)$

- Lu et al. (2023) introduces a path attribution method, considering PostTCE and PostDCE with multiple risk factors.

An outcome may stem from multiple causes. PostTCE and PostDCE distinguish which factor contributes more and which acts as the direct cause.

Lu, Zitong, et al. “Evaluating Causes of Effects by Posterior Effects of Causes.” Biometrika, vol. 110, no. 2, 2023, pp. 449–465.

Evaluating CoE via Posterior Causal Effects

- Posterior Total Causal Effect: Given Evidence: $E = e$

$$PostTCE(X \Rightarrow Y|e) = E(Y_{x=1} - Y_{x=0}|e)$$

- PN is a special case of PostTCE, where the evidence $e = (X = 1, Y = 1)$

- Posterior Intervention Causal Effect: Given Evidence: $E = e$

$$PostICE(X \Rightarrow Y|E = e) = E(Y - Y_{x=0}|E = e)$$

- PD is a special case of PostICE, where the evidence $e = (Y = 1)$

Lu, Zitong, et al. “Evaluating Causes of Effects by Posterior Effects of Causes.” Biometrika, vol. 110, no. 2, 2023, pp. 449–465.

Attribution with Multiple Outcome Variables

- **Lu et al. (2023) Limitation**: Evaluates posterior causes based on a single effect/outcome variable Y.
- **Multiple Outcome Variables**: In medical diagnosis and fault attribution, a cause typically triggers multiple outcome variables simultaneously (e.g., positive X-ray Y_1 and dyspnoea Y_2 caused by lung cancer X).

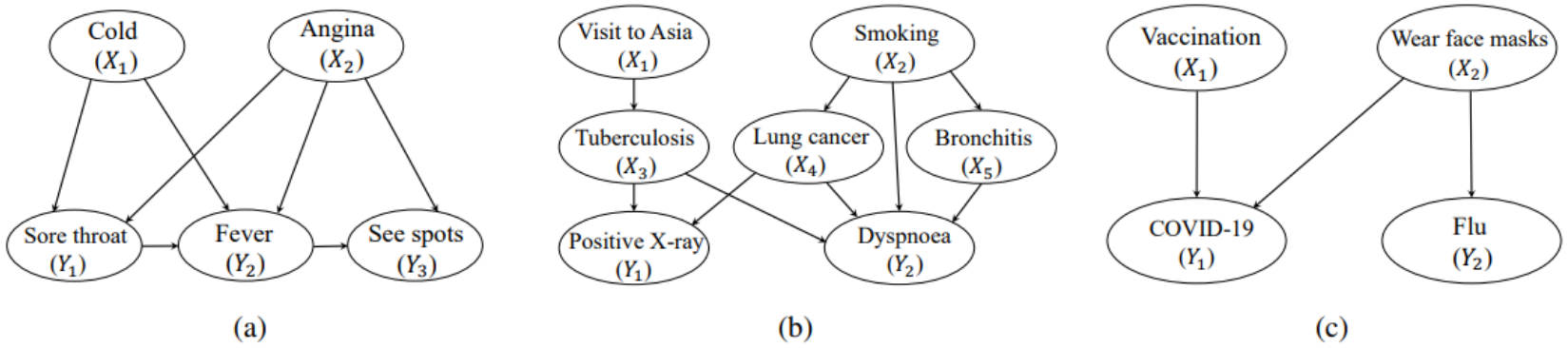


Fig. 1: (a) A modified causal network for the cold or angina example; (b) A Bayesian network modified from Lauritzen & Spiegelhalter (1988); (c) A causal network for COVID-19 and flu.

Li, Wei, Zitong Lu, et al. "Retrospective causal inference with multiple effect variables." Biometrika, 111(2), 2024, pp. 573–589.

Three Multivariate Posterior Causal Effects

All three estimands are case-specific: they condition on the observed evidence O = o.

Notation

$X_k \in \{0,1\}$	candidate cause or treatment
$\mathbf{Y}=(Y_1,\ldots,Y_m)$	vector of outcome variables
$O=o$	observed evidence for this case
$g(\mathbf{Y})$	scalar summary or weighting function of $\mathbf{Y}$
$\mathbf{Y}_a$	potential outcome vector under intervention a
$\mathbf{D}_k$	downstream variables or mediators fixed in DCE
d_k^*	chosen reference value assigned to $\mathbf{D}_k$

Example: $g(\mathbf{Y})=w_1Y_1+\cdots+w_mY_m$ combines several symptoms into one severity score.

MulPostTCE — How much did X_k contribute in total?

$$\text{MulPostTCE}(X_k \Rightarrow \mathbf{Y} \mid O=o) = E\{ g(\mathbf{Y}_{(X_{k=1})}) - g(\mathbf{Y}_{(X_{k=0})}) \mid O=o \}$$

Total effect: includes every direct and indirect pathway from X_k to $\mathbf{Y}$.

MulPostICE — What would improve if X_k were removed?

$$\text{MulPostICE}(X_k \Rightarrow \mathbf{Y} \mid O=o) = E\{ g(\mathbf{Y}) - g(\mathbf{Y}_{(X_{k=0})}) \mid O=o \}$$

Intervention effect: compares the observed outcome with removing X_k.

MulPostDCE — How much did X_k contribute directly?

$$\text{MulPostDCE}(X_k \Rightarrow \mathbf{Y}_{(\mathbf{D}_k=d_k^*)} \mid O=o) = E\{ g(\mathbf{Y}_{(\mathbf{D}_k=d_k^*)}) - g(\mathbf{Y}_{(X_{k=o},\, \mathbf{D}_k=d_k^*)}) \mid O=o \}$$

Direct effect: fixes $\mathbf{D}_k=d_k^*$ and therefore blocks pathways through $\mathbf{D}_k$.

TOTAL = all paths | INTERVENTION = remove X_k | DIRECT = block paths through $\mathbf{D}_k$

Li et al. (2024), Biometrika 111(2): 573–589. DOI: 10.1093/biomet/asad056

Identification Requires Two Explicit Assumptions

The counterfactual contrasts on the previous slide are point identified only when assignment is comparable and response directions are restricted.

1 Strong ignorability

$$\{ \mathbf{Y}(X_{k=1}) , \mathbf{Y}(X_{k=0}) \} \perp X_k \mid C, \quad 0 < P(X_k=1 \mid C) < 1$$

After adjustment for covariates C, exposure assignment is independent of the relevant potential-outcome vectors; both exposure levels remain possible.

What it provides: observed $X_k=1$ and $X_k=0$ groups can represent the missing counterfactuals.

2 Multivariate monotonicity

$$Y_j(X_k=1) \geq Y_j(X_k=0), \quad j=1,\ldots,m$$

For a harmful X_k and adversely coded outcomes, no component of $\mathbf{Y}$ can move in the beneficial direction. Reverse the inequality for a beneficial treatment.

What it provides: rules out counterfactual response types that observational data cannot distinguish.

Strong ignorability + positivity + multivariate monotonicity ⇒ **MulPostTCE, MulPostICE and MulPostDCE are point identified**

Consistency is also assumed. Identification means uniqueness from the observed-data distribution—not that the assumptions are empirically testable.

Li et al. (2024), Biometrika 111(2): 573–589. DOI: 10.1093/biomet/asad056

Attribution With Multi-level Variables

- Classic PoC is mainly designed for binary treatments and outcomes.

- In practice, multi-valued scenarios are prevalent:

Drug dosage: Low / Medium / High

Intervention intensity: None / Weak / Strong

Risk grade: Low / Medium / High

Recommendation policy: A / B / C / D

- Objective: Derive closed-form bounds for PoC under multi-valued treatments and multi-valued outcomes.

Shu, Xin, et al. “Identification of Probabilities of Causation: From Recursive to Closed-Form Bounds.” arXiv preprint arXiv:2505.15274, 2026.

Attribution with Ordinal Outcomes

- **Unstructured vs. Ordinal Outcomes**: Classic PoC assumes binary or unstructured multi-valued outcomes. However, many real-world outcomes are ordinal (e.g., disease severity stages, customer satisfaction).

Table 1: The Lalonde data with both the experimental source and the matched observational source

	experimental data			observational data		
	$Y = 0$	$Y = 1$	$Y = 2$	$Y = 0$	$Y = 1$	$Y = 2$
$Z = 1$	45	32	108	90	64	216
$Z = 0$	92	33	135	115	50	205

Zhang, Chao, Zhi Geng, Wei Li, and Peng Ding. "Identifying and bounding the probability of necessity for causes of effects with ordinal outcomes." Biometrika, 112(3), 2025.

Attribution with Continuous Variables

- Classic PN, PS and PNS apply mostly to binary variables: treatment status, disease status.

- In reality, many treatments and outcomes are continuous:

drug dosage,

blood glucose level,

blood pressure,

test score,

income level.

Kawakami, Yuta, et al. “Probabilities of Causation for Continuous and Vector Variables.” arXiv preprint arXiv:2405.20487, 2024.

Attribution with Continuous Variatbles

- Kawakami et al. (2024) extends PNS to continuous treatments, continuous outcomes, and multiple treatments & outcomes.

$$PNS(y; x_0, x_1) = P(Y_{x_0} < y \leq Y_{x_1})$$

- Focus on whether the cause pushes the outcome across the threshold.

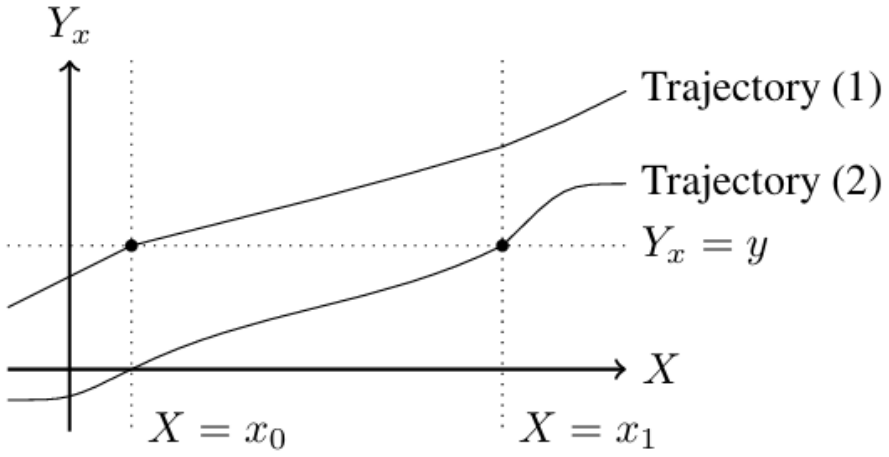


Figure 1: Trajectories for (1) $Y_x(u_{\rho(y;x_0)})$ and (2) $Y_x(u_{\rho(y;x_1)})$.

Kawakami, Yuta, et al. “Probabilities of Causation for Continuous and Vector Variables.” arXiv preprint arXiv:2405.20487, 2024.

SCM-Based RCA Models Anomalies Locally

Objective: identify the source variable responsible for an observed anomaly at a target node.

Anomaly model: a sudden mechanism shift or a local noise spike at one node of the structural causal model.

$X_i = f_i(pa(X_i), \varepsilon_i) + \delta$

RCA Attributes Anomalies Along Causal Paths

Given an anomaly score at the target node:

1 Trace upstream causal paths.

2 Decompose the target anomaly score.

3 Assign abnormal contributions to the source nodes' local noise.

The result explains which upstream disturbance generated the observed failure.

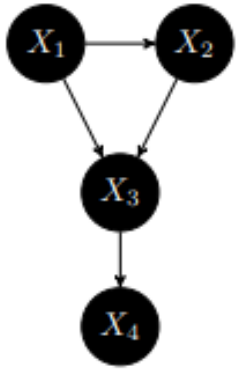
If X -> Y -> Z and we observe Z is abnormal, it may be caused by an abnormal Y/X.

The Full Causal DAG Is Rarely Known

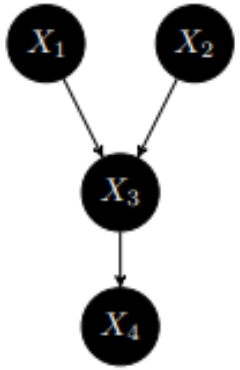
Prior causal RCA methods usually assume that the true causal DAG is fully known.

That assumption is unrealistic in large cloud services, where dependencies are incomplete or changing.

A practical method must work with weaker structural knowledge.



(a) General DAG



(b) Polytree

Orchard, William Roy, et al. "Root Cause Analysis of Outliers with Missing Structural Knowledge." Advances in Neural Information Processing Systems (NeurIPS), 2025.

Outline

- Causal Attribution in Complex Scenarios
- Estimating Causal Effects with Deep Learning
- Application in Real-World Scenarios

Outline

- Estimating Causal Effects with Deep Learning

- Deep Learning meets Causal Effect Estimation

- Causal DL under Potential Outcome Framework

Direct Neural PO Estimators

Balanced Representation Learning

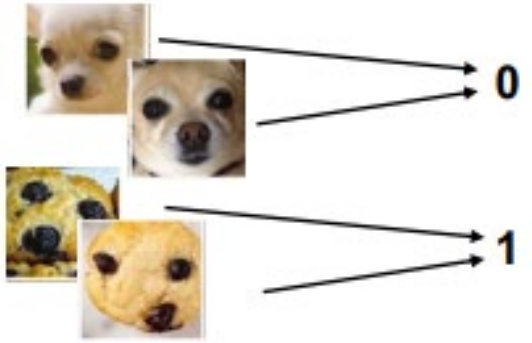
Generative Models for Latent Confounders

Instrumental Variable (IV)-based Methods

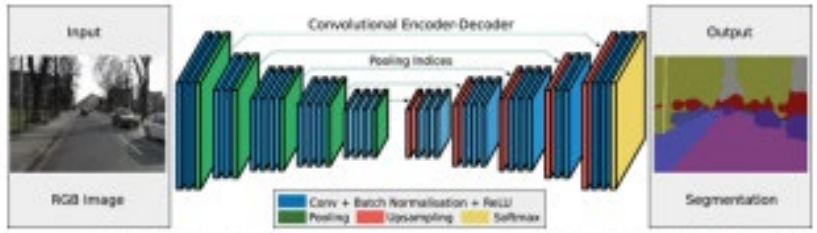
- - Disentangled Representation

How Deep Learning Helps Causal Inference

- Fit really complicated functions and distribution.



- Retrieve semantically meaningful latent features.



Vijay Badrinarayanan et. al 2017 "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation"

- Modern NNs Can Learn to Sample from Complex Data Distributions

Karras et al. *A Style-Based Generator Architecture for Generative Adversarial Networks* CVPR 2019.

Outline

- Estimating Causal Effects with Deep Learning

- Deep Learning (DL) meets Causal Effect Estimation

- Causal DL under Potential Outcome Framework

Direct Neural PO Estimators

Balanced Representation Learning

Generative Models for Latent Confounders

Instrumental Variable (IV)-based Methods

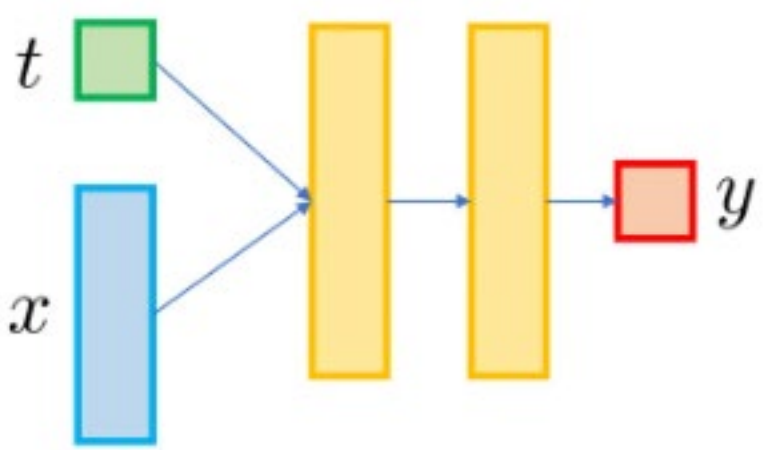
- - Disentangled Representation

Direct Neural PO Estimators

- **Goal**: Directly estimate the Conditional Average Treatment Effect (CATE) $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ from observational data.

- **Examples:** S-Learner, T-Learner, X-Learner, IPS-Learner, DR-Learner, DML, etc. These methods are simple and direct, serving as base methods for many advanced deep learning methods.

S-learner



Concatenate t as a regular feature with x to train a unified regression model $f : \mathcal{X} \times \{0, 1\} \to \mathbb{Y}$:

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i, t_i))^2$$

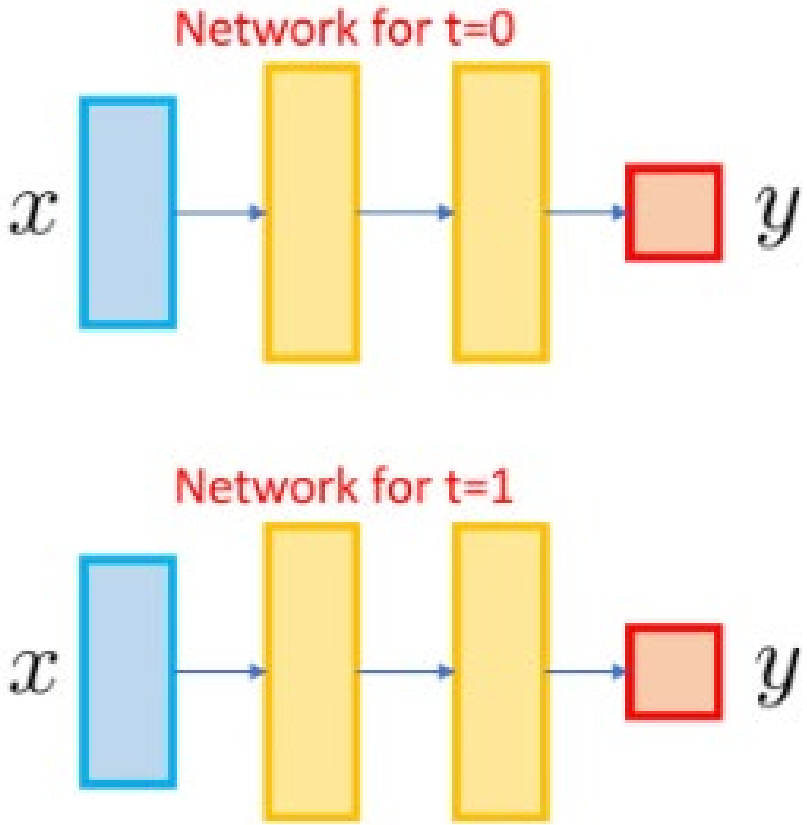
The CATE estimator is:

$$\hat{\tau}_S(x) = f(x, 1) - f(x, 0)$$

It has a simple structure and fully utilizes the data, but may dilute the effect of treatment t when x is high-dimensional.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10), 4156-4165.

T-learner



Split the dataset by t into the treatment group D_1 and control group D_0. Train two models independently:

$$\hat{\mu}_1 = \arg\min_{\mu_1} \sum_{i \in \mathcal{D}_1} (y_i - \mu_1(x_i))^2,$$

$$\hat{\mu}_0 = \arg\min_{\mu_0} \sum_{i \in \mathcal{D}_0} (y_i - \mu_0(x_i))^2$$

The CATE estimator is:

$$\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

Address the treatment effect dilution problem in S-Learner, but can only leverage subsample to train each model, sensitive to unbalanced sample size.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10), 4156-4165.

DR-learner

- Estimate the propensity score model $\hat{e}(x)$,and the conditional outcome models $\widehat{\mu_1}(x), \widehat{\mu_0}(x)$.
- Compute pseudo-outcomes:

$$Y_i^{DR}(1) = \hat{\mu}_1(x_i) + \frac{t_i(y_i - \hat{\mu}_1(x_i))}{\hat{e}(x_i)}$$

$$Y_i^{DR}(0) = \hat{\mu}_0(x_i) + \frac{(1 - t_i)(y_i - \hat{\mu}_0(x_i))}{1 - \hat{e}(x_i)}$$

- Compute pseudo-effect: $\tilde{Y}_i^{DR} = Y_i^{DR}(1) - Y_i^{DR}(0)$
- Fit the pseudo-effect: $\hat{\tau}_{DR} = \arg\min_{\tau} \sum_{i=1}^{n} \left(\tilde{Y}_i^{DR} - \tau(x_i)\right)^2$
- Double robustness: $\hat{\tau}_{DR}$ is consistent if either the propensity model or the outcome model is consistent.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics, 17(2), 3008-3049.

Outline

- Estimating Causal Effects with Deep Learning

- Deep Learning (DL) meets Causal Effect Estimation

- Causal DL under Potential Outcome Framework

Direct Neural PO Estimators

Balanced Representation Learning

Generative Models for Latent Confounders

Instrumental Variable (IV)-based Methods

- - Disentangled Representation

From Covariates to Representations

- $Y(t)$ is independent from T given X. Adjustment with X is sufficient.
- However, high-dimensional confounder may be hard to condition on.

Learnt representation is a child of confounder X in the true DAG.

A Good Representation W:

Sufficient: $Y \perp X|W, p(Y|W, X = x) = p(Y|W, X = x')$

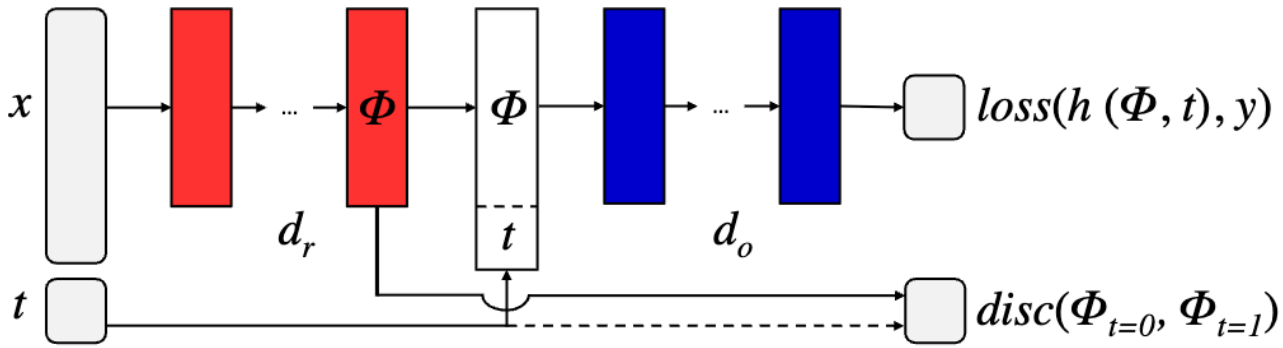
Balanced: $W \perp T, p(W|T = 0) = p(W|T = 1)$

Trade-off:

$$\min_{\Phi,h} \underbrace{L_{\text{factual}}}_{} + \alpha \underbrace{D(P(\Phi(X) \mid T = 1), P(\Phi(X) \mid T = 0))}_{}$$

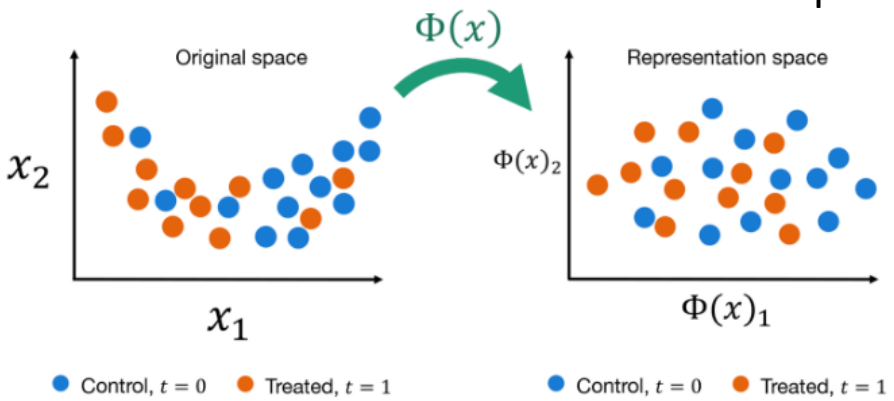
Johansson, Fredrik, Uri Shalit, and David Sontag. "Learning representations for counterfactual inference." *International conference on machine learning*. PMLR, 2016.

Balanced Representations Learning



BLR framework

Domain Adaptive



Matching: make Y^{CF} estimate "close" to the most similar observed (X, T) pair.

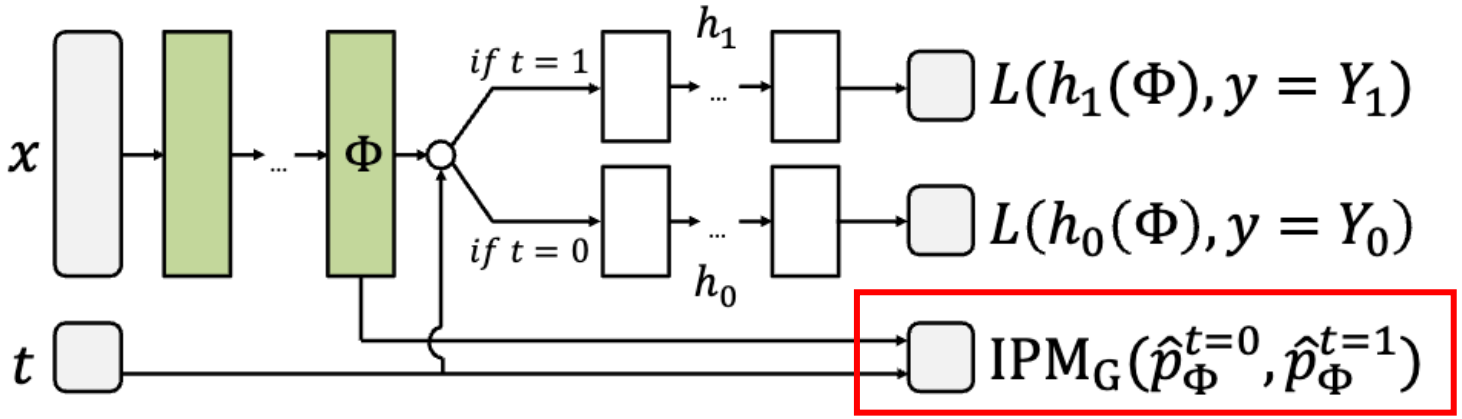
BLR: Use a NN to learn a representation $W = \Phi(X)$ that are balanced across Treatment, i.e. $P(\Phi(X) \mid T = 1) \approx P(\Phi(X) \mid T = 0)$

$$B_{\mathcal{H},\alpha,\gamma}(\Phi, h) = \frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F| + \qquad (2)$$

$$\alpha \operatorname{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}) + \frac{\gamma}{n}\sum_{i=1}^{n} |h(\Phi(x_i), 1 - t_i) - y_{j(i)}^F| \,,$$

Johansson, Fredrik, Uri Shalit, and David Sontag. "Learning representations for counterfactual inference." *International conference on machine learning*. PMLR, 2016.

CFRNet



CFRNet: Minimize to force the distribution of $\Phi(x)$ of treatment and control groups to be similar.

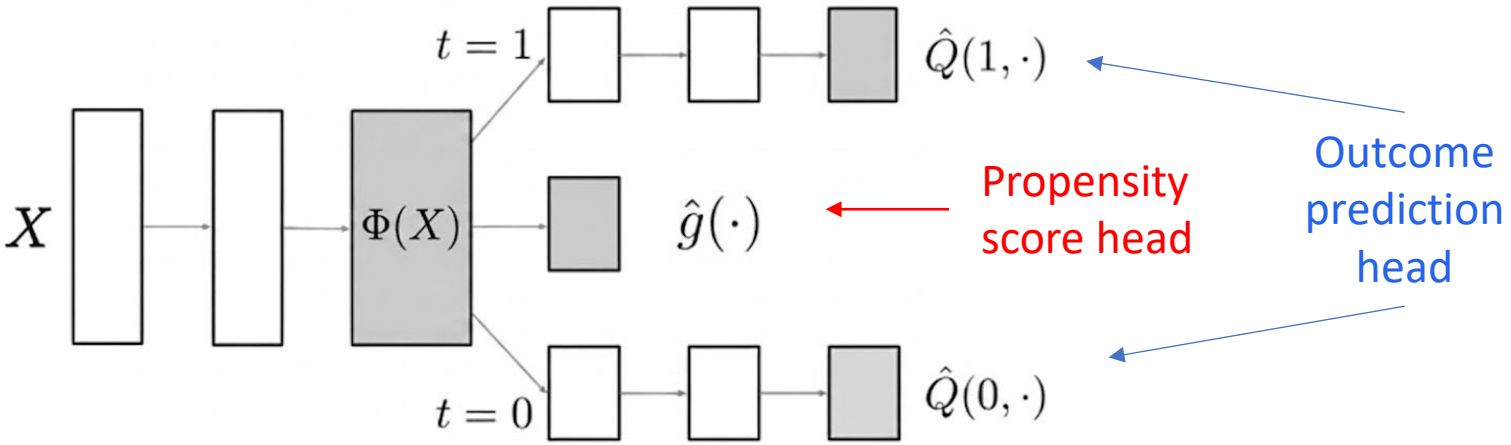
Integral Probability Metric (IPM): measures the distributional distance between the two groups.

$$\mathrm{IPM}_G(p, q) := \sup_{g \in G} \left| \int_S g(s)(p(s) - q(s))\, ds \right|$$

$$\min_{h,\Phi,IPM} \frac{1}{N} \sum_{i=1}^{N} \underbrace{MSE(Y_i, h(\Phi(X_i), T_i))}_{\text{Outcome Loss}} + \lambda \underbrace{IPM(\Phi(X|T=1), \Phi(X|T=0))}_{\text{Dist. b/w T \& C covar. distributions}} + \alpha \underbrace{\mathcal{R}(h)}_{L_2}$$

Shalit, U., Johansson, F. D., & Sontag, D. (2017, July). Estimating individual treatment effect: generalization bounds and algorithms. In International conference on machine learning (pp. 3076-3085). PMLR.

DragonNet



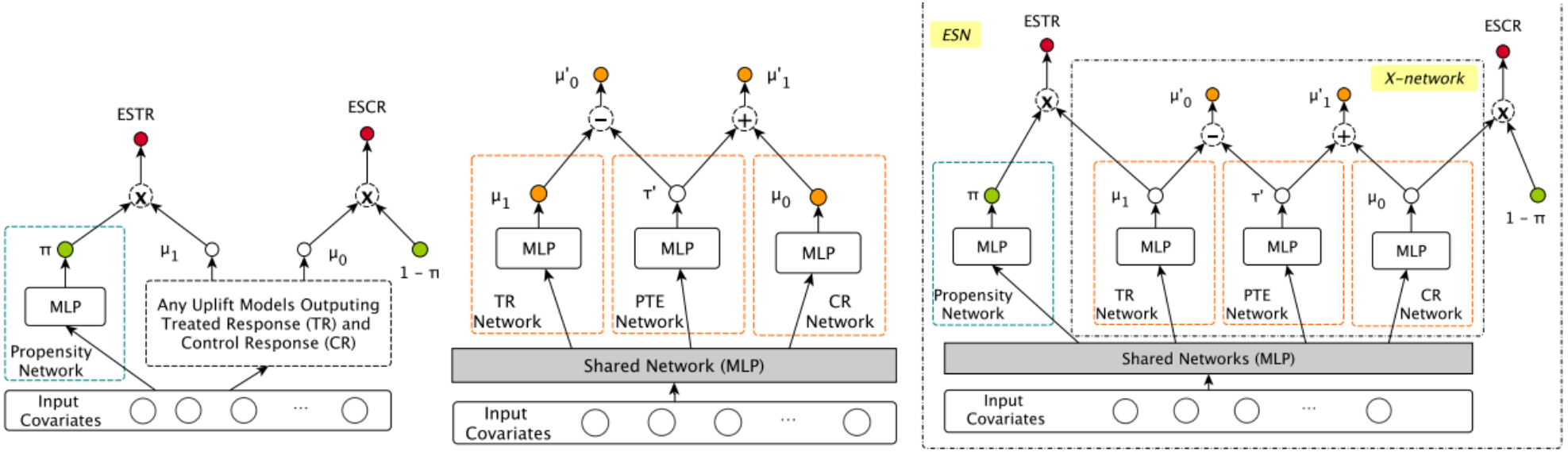
Find a good representation Z to estimate **both** propensity score **AND** conditional outcome.

- BLR, TARNet and CFRNet Learn Representation to serve only the prediction of potential outcome.

- DragonNet forces the Representation Learning to serve not only the outcome prediction, but also propensity score prediction. It trains Potential Outcome Model $h_1(\Phi(X)), h_0(\Phi(X))$ and PS model $\pi(\Phi(X))$ based on the same representation $\Phi(X)$.

Shi, Claudia, David Blei, and Victor Veitch. "Adapting neural networks for the estimation of treatment effects." *Advances in neural information processing systems* 32 (2019).

DESCN: Entire-Space Cross Learning



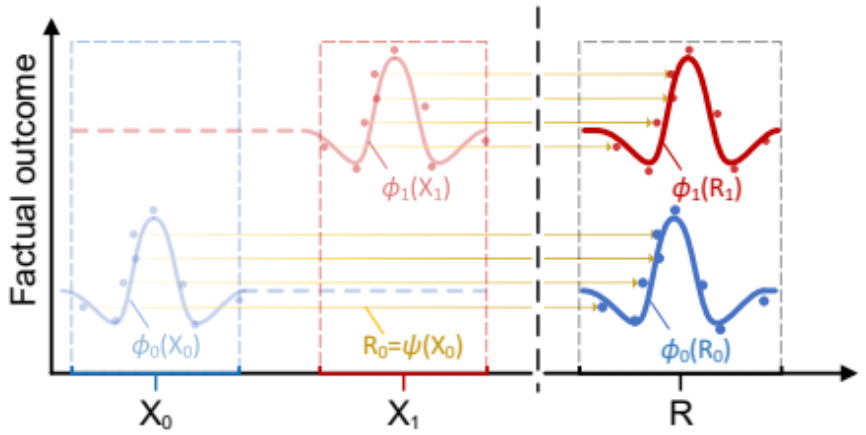
(a) Entire Space Network (ESN) (b) X-network (c) Deep Entire Space Cross Networks (DESCN)

Entire-space learning: jointly models treatment propensity and both response functions using all samples.

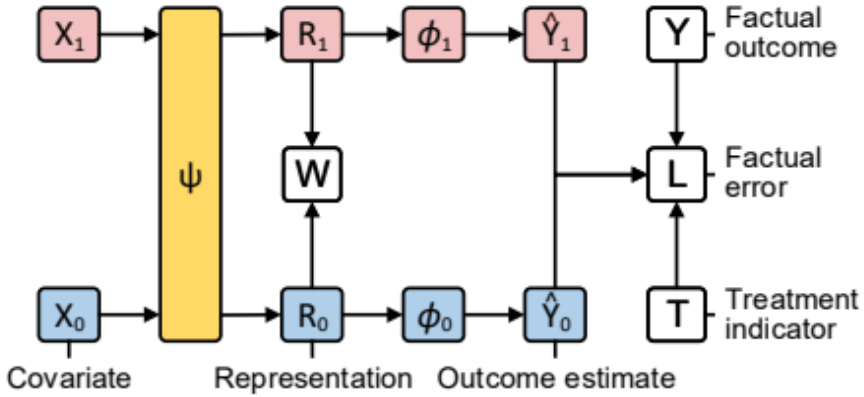
Cross-network: transfers signal through a pseudo treatment effect, helping when one arm is much smaller.

Zhong, K., et al. (2022). DESCN: Deep Entire Space Cross Networks for Individual Treatment Effect Estimation. KDD 2022.

ESCFR: Optimal-Transport Balancing



(a) Handling treatment selection bias with $\psi(\cdot)$.



(b) Architecture of ESCFR.

optimal transport: aligns treated and control representations with a generalized Sinkhorn discrepancy

Wang, H., et al. (2023). Optimal Transport for Treatment Effect Estimation. NeurIPS 2023.

Outline

- Estimating Causal Effects with Deep Learning

- Deep Learning (DL) meets Causal Effect Estimation

- Causal DL under Potential Outcome Framework

Direct Neural PO Estimators

Balanced Representation Learning

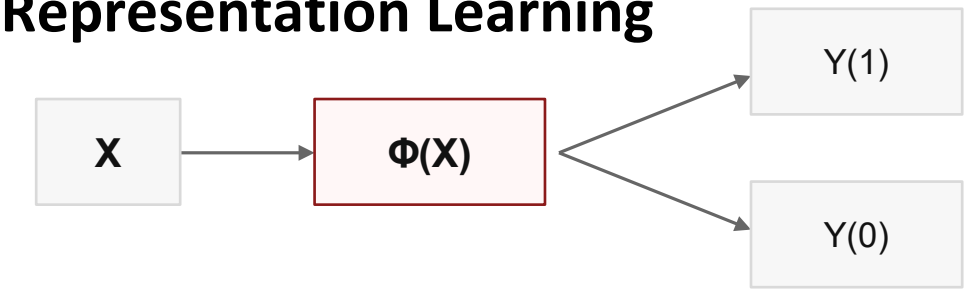
Generative Models for Latent Confounders

Instrumental Variable (IV)-based Methods

- - Disentangled Representation

Representation vs. Generative Modeling

Representation Learning



Goal: Learn a balancing mapping Φ: $X \rightarrow Z$ to reduce extrapolation in X-space.

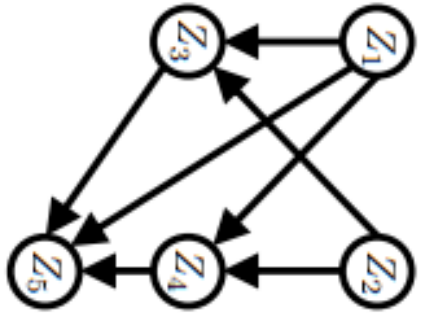
latent causal variables;

causal relationships;

invariant representations.

What are the causal variables behind complex observations?

Generative Modeling



Goal: Learn the entire joint distribution $P(X, Y, Z)$ by data-generating process.

structural mechanisms;

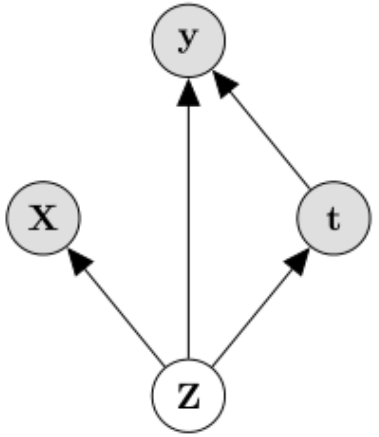
latent noise;

intervention effects;

counterfactual worlds.

How does the causal system generate observations under interventions?

Latent Feature Construction

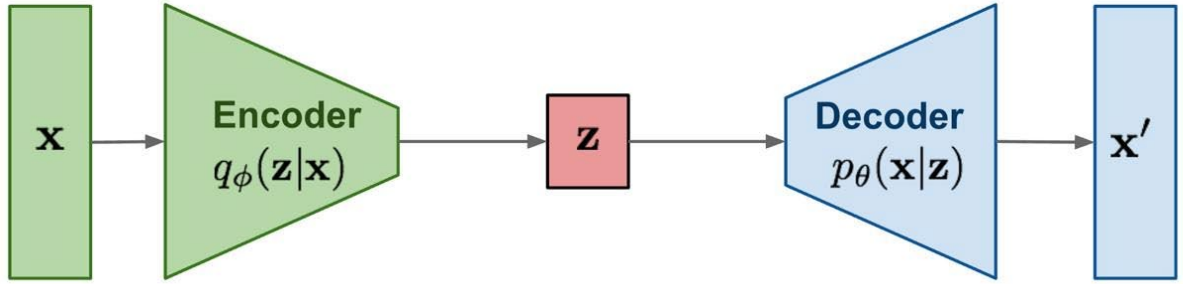


- Unobserved confounder Z prevent directadjustment.
- Sometime we can Observe variables X as proxy measurements of unobserved Z,
  e.g. Genetic factors → Blood sugar

Use VAE to recover latents and use them for adjustment.

Variational autoencoder (VAE) can be used to disentangle independent factors from observations.

VAE: maximize variational lower bound



Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." Advances in neural information processing systems 30 (2017).

CEVAE

Task: Estimate CATE when the true confounder Z is unobserved, but we have multiple noisy proxies X (e.g., medical test results proxying overall health)

Data Generative Process:

$$p(z, x, t, y) = p(z)\, p(x \mid z)\, p(t \mid z)\, p(y \mid t, z)$$

The objective from original VAE – ELBO:

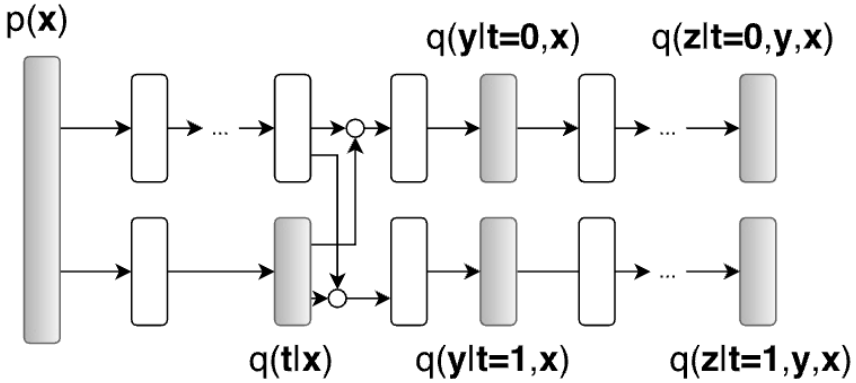
$$\mathcal{L}_{ELBO} = \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i,t_i,y_i)}[log\, p(\mathbf{x}_i, t_i \mid \mathbf{z}_i) + log\, p(y_i \mid t_i, \mathbf{z}_i) + log\, p(\mathbf{z}_i) - log\, q(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i)]$$

Two additional terms

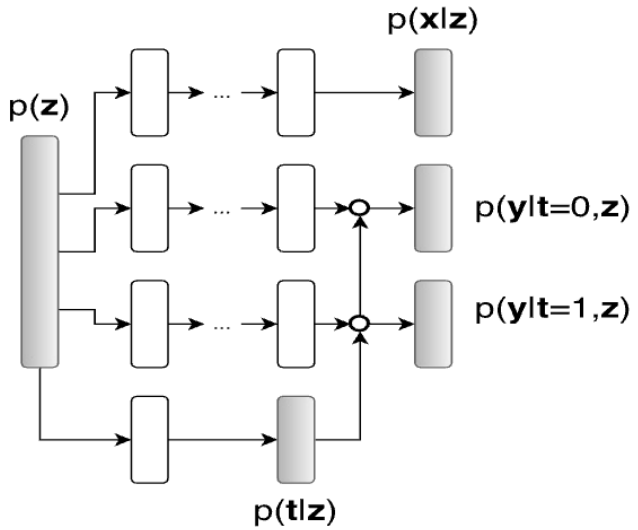
$$\mathcal{F}_{\text{CEVAE}} = \mathcal{L} + \sum_{i=1}^{N} \left(\log q(t_i = t_i^* | \mathbf{x}_i^*) + \log q(y_i = y_i^* | \mathbf{x}_i^*, t_i^*) \right),$$

Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." Advances in neural information processing systems 30 (2017).

CEVAE



(a) Inference network, $q(\mathbf{z}, t, y|\mathbf{x})$.



(b) Model network, $p(\mathbf{x}, \mathbf{z}, t, y)$.

Encoder

Predict T from X,

Y from X, T

Z from X, T, Y

Decoder

Predict T from Z,

X from Z

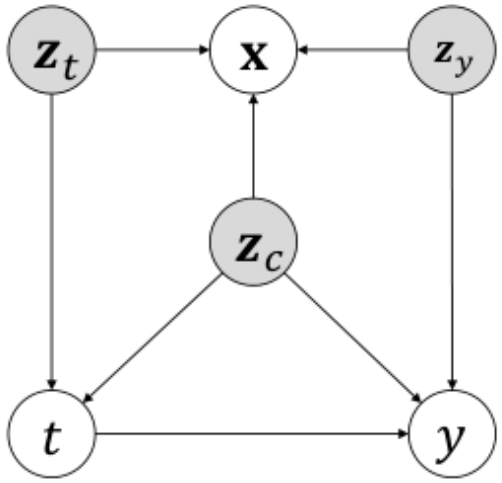
Y from T, Z

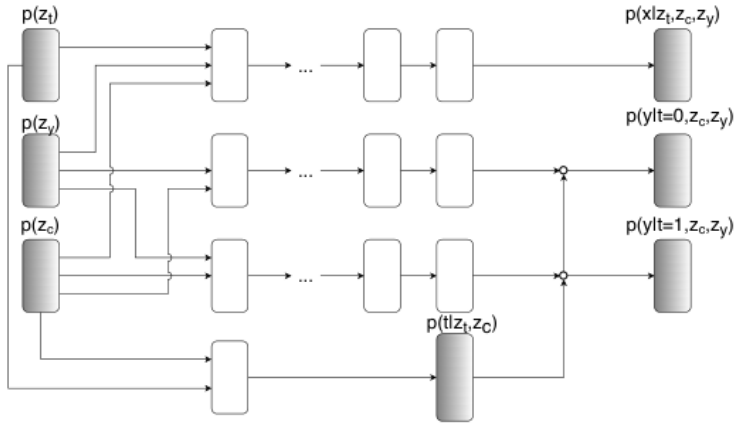
Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." Advances in neural information processing systems 30 (2017).

TEDVAE

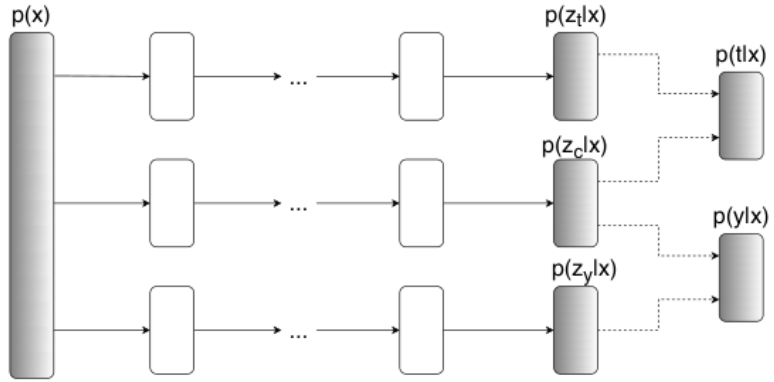
Not every covariate is a confounder

$(z_t, z_c, z_y) \rightarrow X, (z_t, z_c) \rightarrow T, (z_c, z_y, T) \rightarrow Y$





(a) Generative Model.



(b) Inference Model.

Disentangled latent factors

z_t: IV/treatment-only factor

z_c: confounding factor

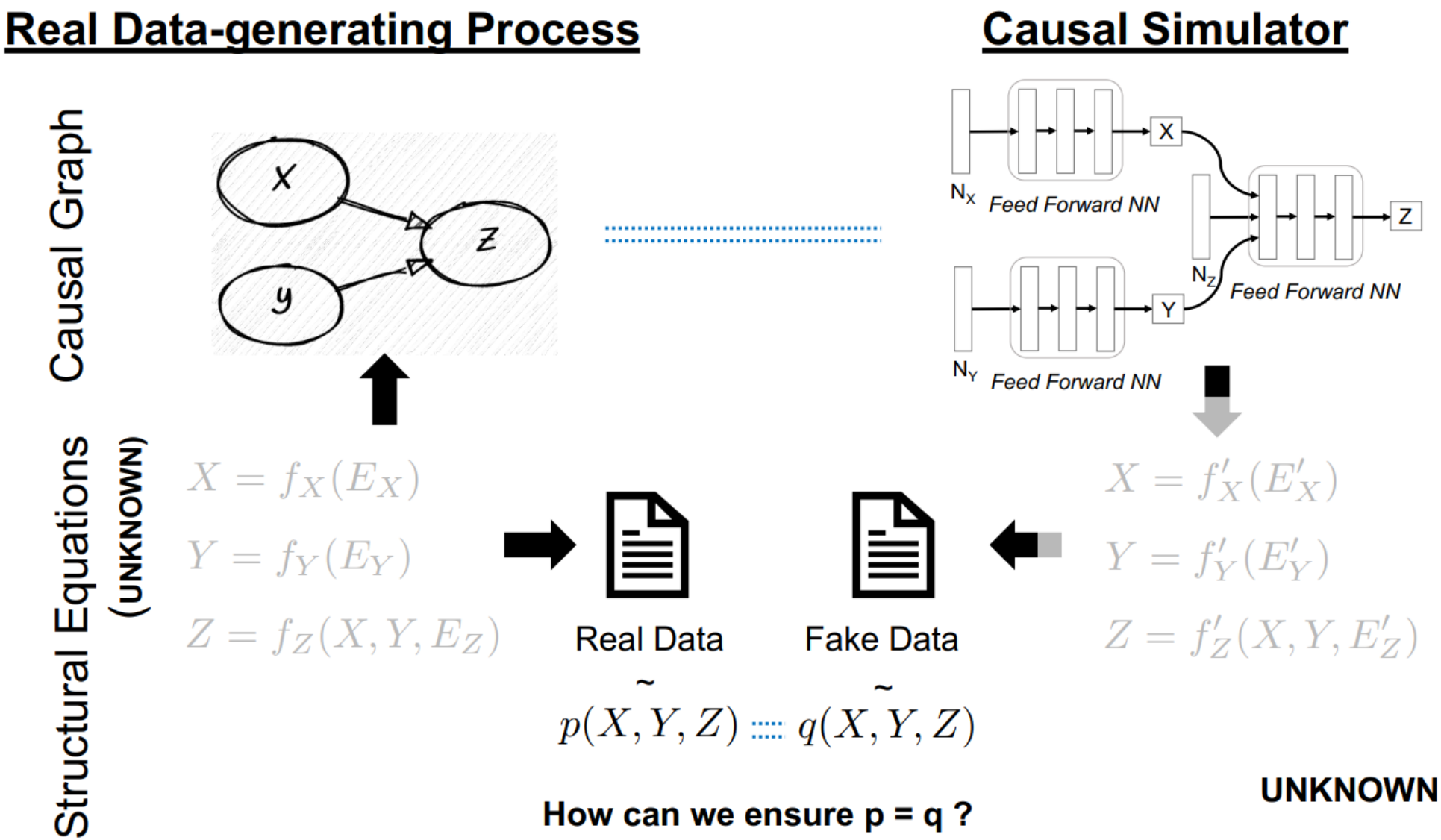
z_y: risk/outcome-only factor

Separates factors, so the outcome model and effect estimator **use the right latent components**.

$$\mathcal{L}_{TEDVAE} = \text{ELBO} + \alpha_t \log q(t \mid z_t, z_c) + \alpha_y \log q(y \mid z_y, z_c, t)$$

Zhang, Liu & Li, "Treatment Effect Estimation with Disentangled Latent Factors," AAAI 2021.

Generative Adversarial Networks (GANs)



GANITE: GAN for ITE

Reframes counterfactual estimation as a generative imputation problem.

Factual outcome as supervised signal and anchor the generator.

Train a GAN to adversarially impute missing potential outcomes.

Discriminator pushes generated counterfactuals toward plausible potential-outcome patterns.

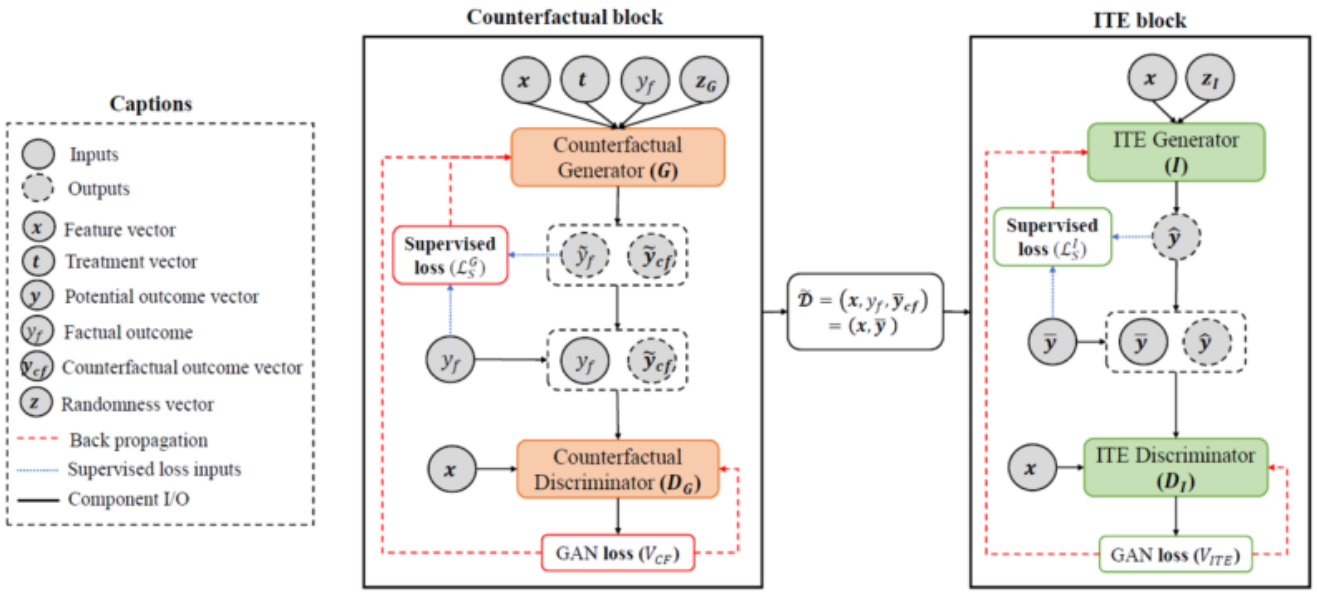
ITE estimation becomes a supervised learning problem.

Yoon, Jinsung, James Jordon, and Mihaela Van Der Schaar. "GANITE: Estimation of individualized treatment effects using generative adversarial nets." International conference on learning representations. 2018.

GANITE: GAN for ITE

Reframes counterfactual estimation as a generative imputation problem.

Two-Stage Architecture



Y^{cf} imputation block $G_{CF} : (X, T, Y^{obs}) \rightarrow \tilde{Y}^{cf}$

$$\mathcal{L}_{CF} = \mathcal{L}_{GAN} + \alpha \mathcal{L}_{sup}$$

Generator G_c takes (X, t, y^{fa}) and generates $\hat{y}^{cf}$.

Discriminator D_c distinguishes which outcome is true vs. generated.

ITE estimation block $I(X) \rightarrow (\hat{Y}(0), \hat{Y}(1))$

$$\mathcal{L}_{ITE} = \mathcal{L}_{GAN} + \beta \mathcal{L}_{sup}$$

Generator G_i takes X and generates both $\hat{y}_0, \hat{y}_1$.

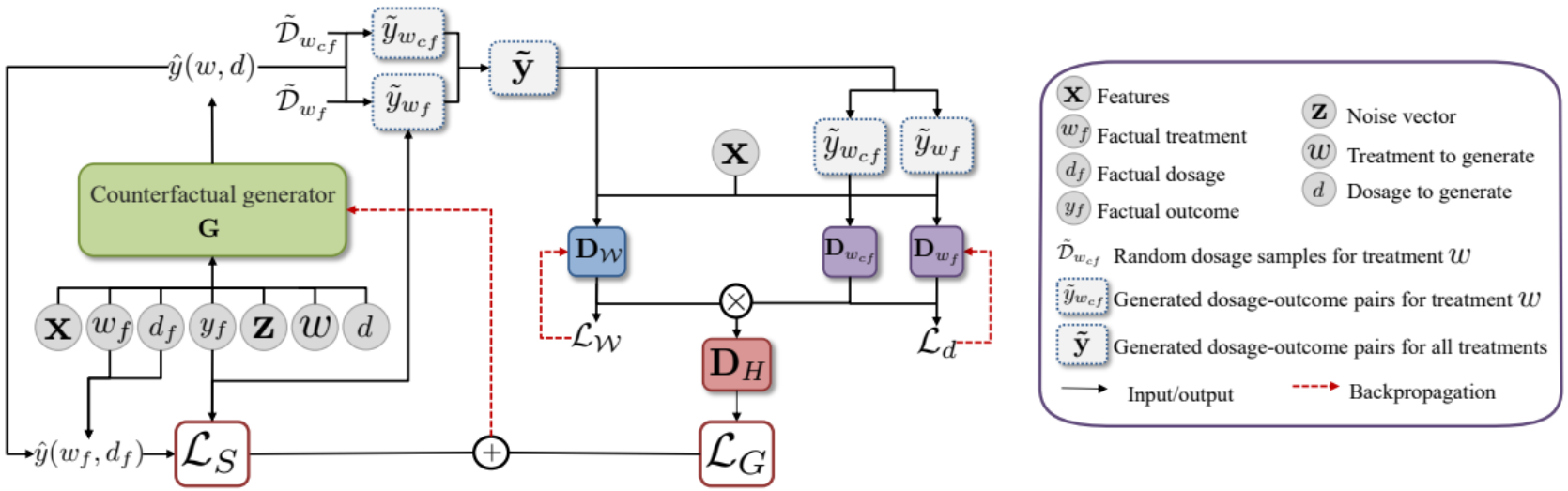
Discriminator D_i distinguishes between G_i's output and the complete pairs from Stage 1.

Yoon, Jinsung, James Jordon, and Mihaela Van Der Schaar. "GANITE: Estimation of individualized treatment effects using generative adversarial nets." International conference on learning representations. 2018.

SCIGAN

TASK — Estimate individualized response curves for continuous treatment and dosage.

CORE IDEA — Generate outcomes over treatment–dosage pairs and use hierarchical discriminators for treatment and dose.



Bica, I., Jordon, J., & van der Schaar, M. “Estimating the Effects of Continuous-valued Interventions Using GANs.” NeurIPS 2020.

DFITE

TASK — Estimate ITE when confounders are unobserved but prior generative knowledge is available.

CORE IDEA — Run conditional reverse diffusion over the latent confounder and predict both potential outcomes.

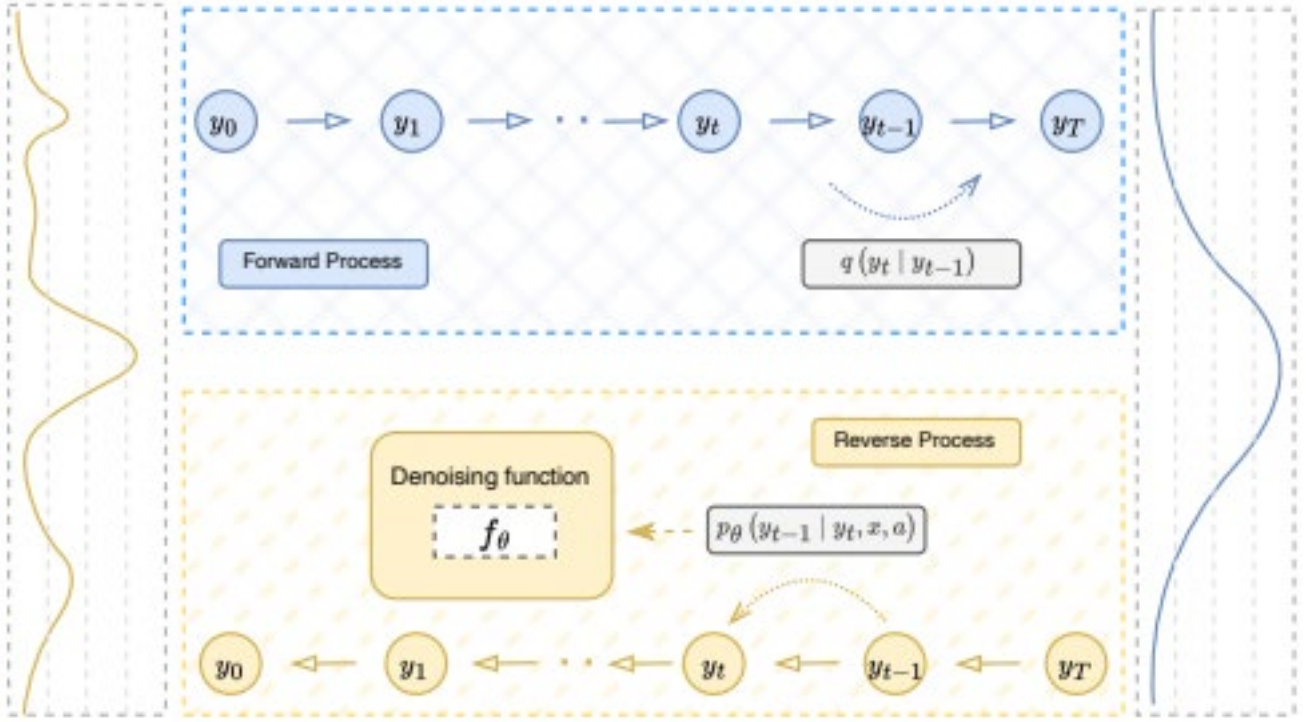
$$
\begin{aligned}
L_{VLB} &= E_q\left[\log \frac{q(\boldsymbol{x}^{(1:T)}, \boldsymbol{\eta}|\boldsymbol{x}^{(0)})}{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0:T)}, \boldsymbol{\eta}))}\right] \\
&= E_q\left[\sum_{t=2}^{T} D_{KL}\left(\underbrace{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)})}_{A} || \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}, \boldsymbol{\eta})}_{B}\right)\right. \\
&\left. - \log \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(0)}|\boldsymbol{x}^{(1)}, \boldsymbol{\eta})}_{C} + D_{KL}\left(\underbrace{q_{\boldsymbol{\varphi}}(\boldsymbol{\eta}|\boldsymbol{x}^{(0)})}_{D} || \underbrace{p(\boldsymbol{\eta})}_{E}\right)\right]
\end{aligned}
$$

Wang, Z., et al. “DFITE: Estimation of Individual Treatment Effect Using Diffusion Model.” ICLR 2024 submission, OpenReview.

DiffPO

TASK — Learn the full distributions of potential outcomes—not only their conditional means.

CORE IDEA — Combine conditional diffusion with an orthogonal loss to reduce selection bias and quantify uncertainty.



Ma, Y., Melnychuk, V., Schweisthal, J., & Feuerriegel, S. “DiffPO: A Causal Diffusion Model for Learning Distributions of Potential Outcomes.” NeurIPS 2024.

Outline

- Estimating Causal Effects with Deep Learning

- Deep Learning meets Causal Effect Estimation

- Causal DL under Potential Outcome Framework

Direct Neural PO Estimators

Balanced Representation Learning

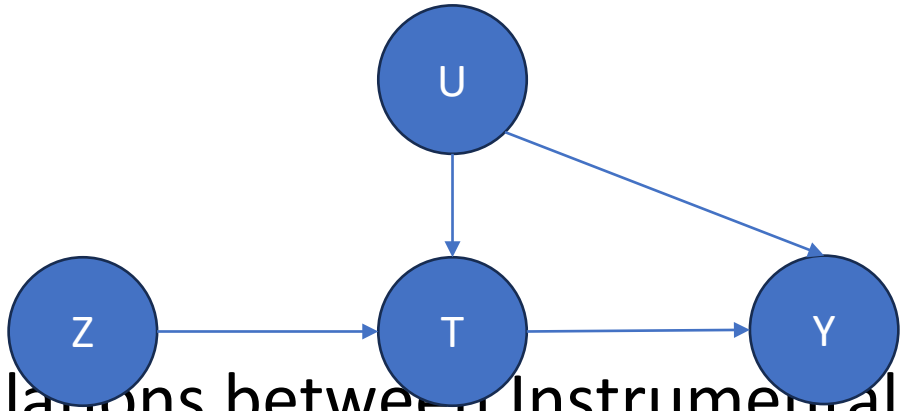
Generative Models for Latent Confounders

Instrumental Variable (IV)-based Methods

- - Disentangled Representation

Instrumental Variable (IV)-based methods

- Instrumental Variable (IV) is a powerful tool for addressing hidden confounding.



- In practice, the relations between Instrumental variables/ treatment variables, and outcome variable, can be non-linear and complex, traditional methods (e.g. 2SLS) can not be directly applied.

DeepIV: Instrumental Variables

Assume

$$Y = g(T, X) + L$$

The "counterfactual":

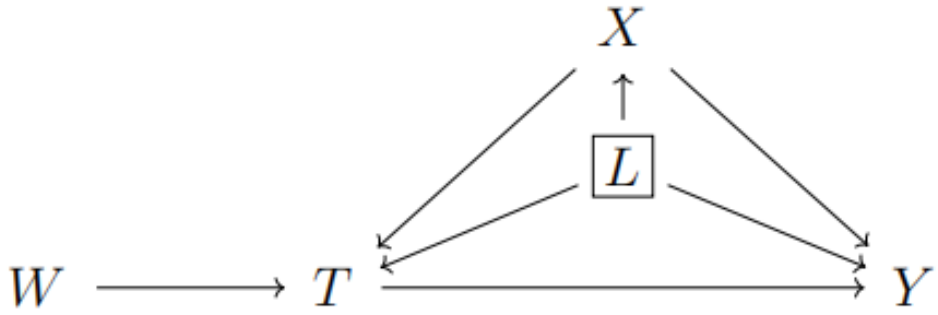
$$h(t, x) = g(t, x) + \mathbb{E}[L|X = x]$$

Learning $g(\cdot)$ is enough:

$$h(t_1, x) - h(t_0, x)$$
$$= g(t_1, x) - g(t_0, x)$$

Need to solve for integral equation:

- $\mathbb{E}[Y \mid x, w] = \int h(t, x_i) dF(t \mid x_i, w_i)$



Solve in 2 stage:

1. Do a density estimate of treatment $f(t|x, w)$ given covariate and instrument.
2. Learn the potential outcome predictor h based on learned $f(t|x, w)$.

$$\min_h \sum_{i=1}^{n} \left(y_i - \int h(t, x_i) dF(t \mid x_i, w_i)\right)^2$$

Deep IV: A Flexible Approach for Counterfactual Pred. Hartford, Lewis, Leyton-Brown, Taddy, Hartford et al., ICML 2017;

Outline

- Estimating Causal Effects with Deep Learning

- Deep Learning meets Causal Effect Estimation

- Causal DL under Potential Outcome Framework

Direct Neural PO Estimators

Balanced Representation Learning

Generative Models for Latent Confounders

Instrumental Variable (IV)-based Methods

- Disentangled Representation

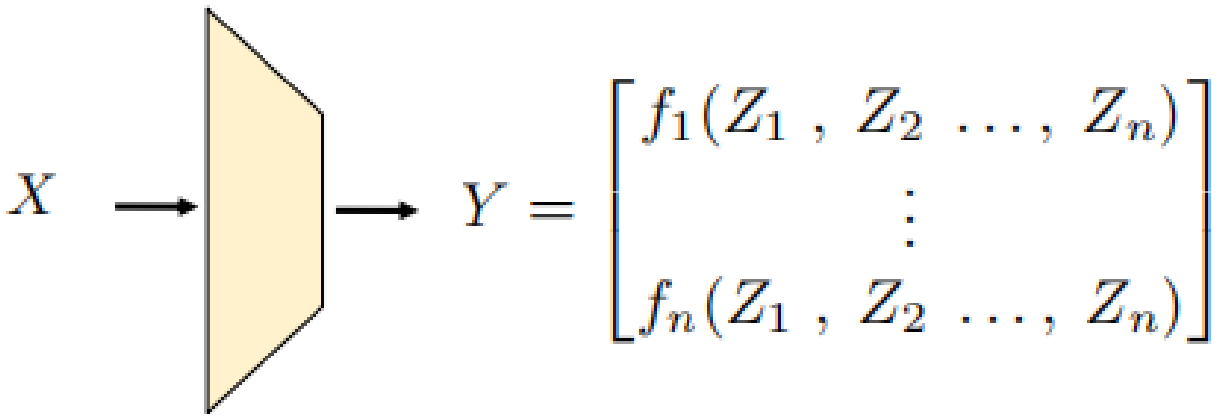
Disentangled Representation

Unknown transformation

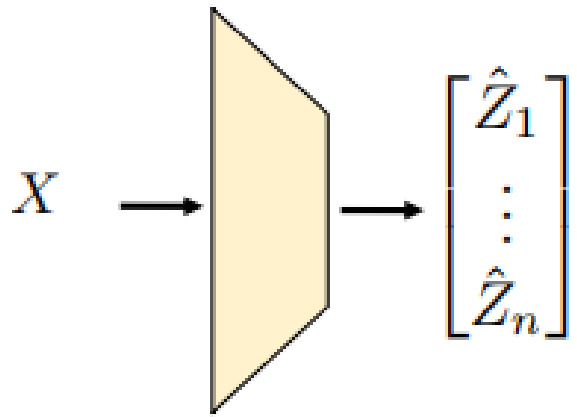
$$X = g(Z_1, Z_2, \ldots, Z_n)$$

data — latent factors

Entangled Representation:



Disentangled Representation:



β-VAE

TASK — Learn interpretable latent factors from observations—without causal labels.

CORE IDEA — Up-weight KL regularization (β > 1) to trade reconstruction fidelity for a more factorized latent code.

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \qquad (4)$$

Higgins, I., et al. “β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.” ICLR 2017.

Causal Disentanglement

Statistical disentanglement (β-VAE)

$$\prod_{i=1}^{n} p(Z_i)$$

Assume $Z_i \perp Z_j$

→ But real causal factors are dependent!

Causal disentanglement (CausalVAE)

Model a latent causal graph

$$Z_i = f_i(PA_i, \varepsilon_i) \text{ with } \varepsilon_i \perp \varepsilon_j$$

→ Factors can be dependent, but their exogenous components are independent

Yang, Mengyue, et al. "Causalvae: Disentangled representation learning via neural structural causal models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Outline

- Causal Attribution in Complex Scenarios
- Estimating Causal Effects with Deep Learning
- Application in Real-World Scenarios

Applications of Causal Technology in Real-World Scenarios

a. **Computer Vision**

b. **Recommender System**

c. **Natural Language Processing**

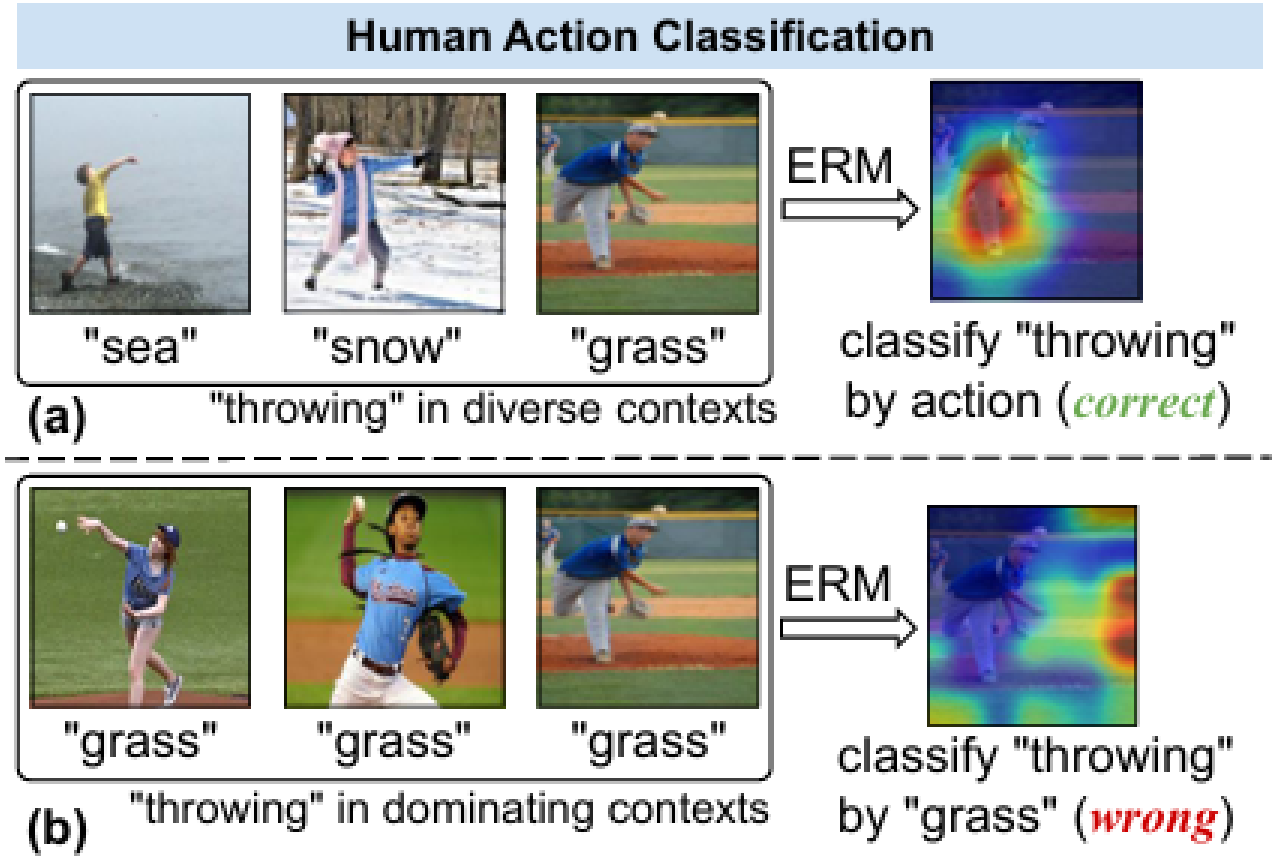
Causal Visual Recognition under Spurious Correlations

- Visual models are often trained from **observational data.**

- The label may have **spurious correlation with** background, position, language patterns, or object co-occurrence.

- Causal learning tries to **identify the true visual evidence, rather than the shortcut.**

Causal Visual Recognition under Spurious Correlations



In visual tasks, **labels are often correlated with context.**

Qi, Jiaxin, et al. “Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization.” arXiv preprint arXiv:2208.03462, 2025

Application in Computer Vision

- Deconfounding Visual Prediction
- Invariant Learning for OOD Generalization
- Counterfactual Augmentation for Robust and Fair Recognition

A General Causal View of Visual Bias

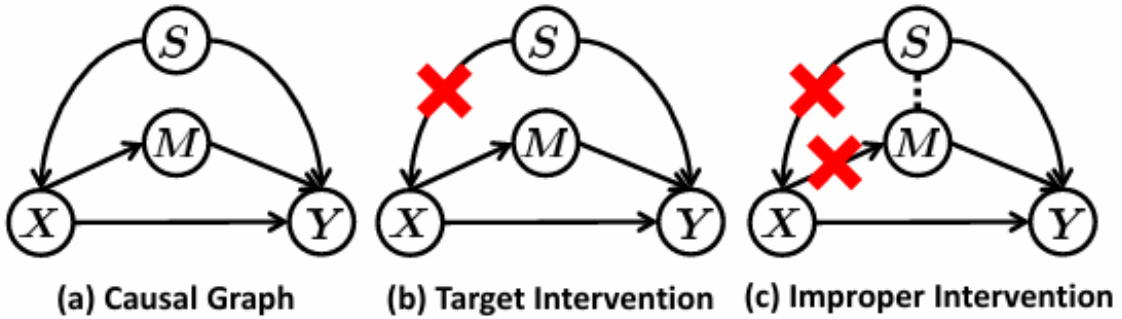


Figure 3. The causal graphs of visual recognition.

Representing visual recognition as a causal graph:

$$\boldsymbol{X \leftarrow S \rightarrow Y}$$

Where:

- X: Input feature
- Y: Prediction
- S: Confounder, such as background or co-occurrence pattern.

Confounder S simultaneously influences the input X and the predicted Y, forming a backdoor path.

The objective is to estimate intervention effect:

$$\boldsymbol{P(Y \mid do(X))}$$

Instead of simple correlation:

$$P(Y \mid X)$$

Wang, Tan, et al. “Causal Attention for Unbiased Visual Recognition.” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Image Recognition: Attention Can Be Misled by Background

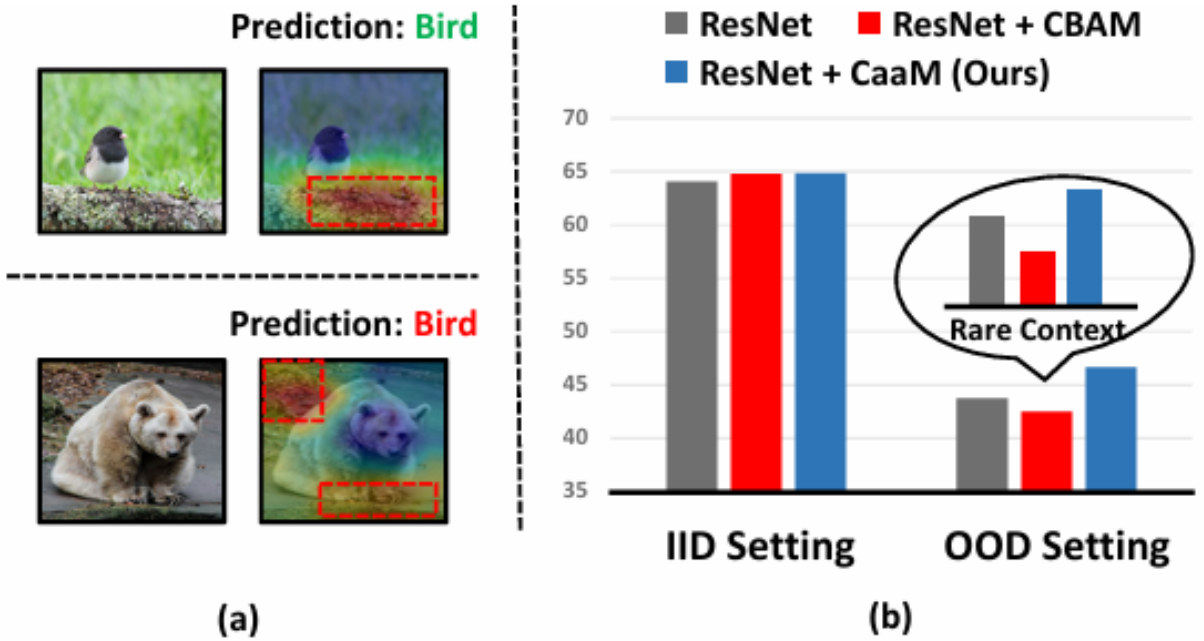


Figure 1. (a) The qualitative attention maps of two images in NICO [21] using ResNet18 with CBAM [57]. (b) The accuracies of three methods: ResNet18, ResNet18+CBAM [57] and + CaaM (Ours) in both IID and OOD settings.

High attention does not always indicate causal features

Wang, Tan, et al. “Causal Attention for Unbiased Visual Recognition.” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Unbiased Attention through Causal Interventions

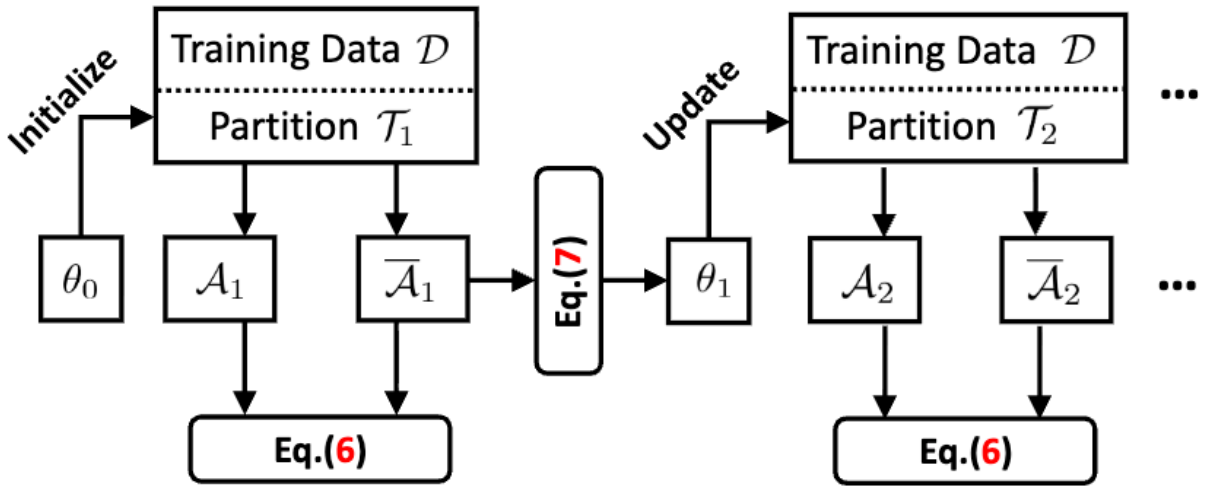
Separating Causal Features from Confounding Features

Step 1: Two complementary attentions

- One attention branch captures **causal features**.
- Another complementary branch captures **confounding features**.

Step 2: Iterative partition update

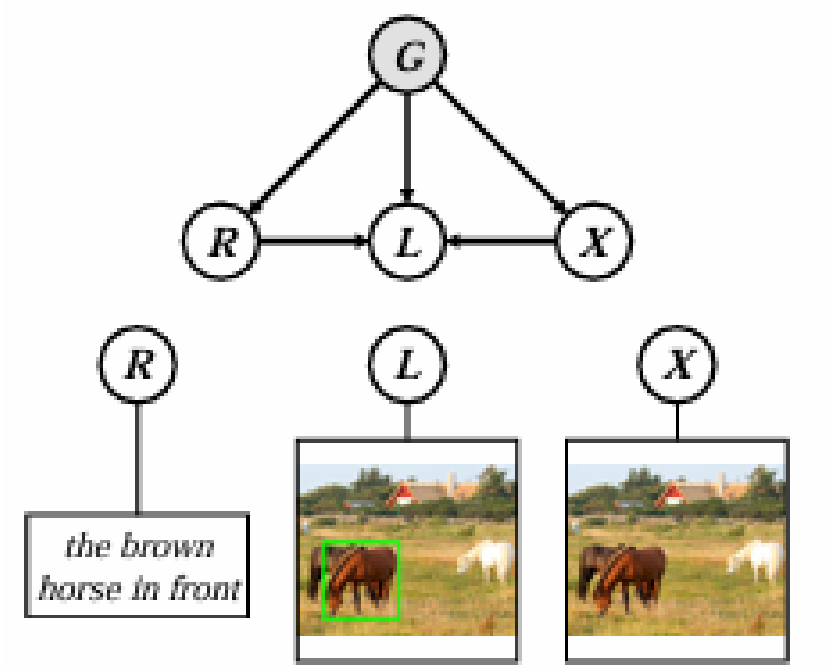
- Use the confounding branch to discover data partitions.
- Each partition approximates a confounder stratum.



Wang, Tan, et al. “Causal Attention for Unbiased Visual Recognition.” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Visual Grounding: Language-Position Bias

- Task: visual grounding / referring expression comprehension.
- Input: image X and query R.
- Output: target location L.
- Expected behavior: locate the region truly referred to by the query.
- **Actual behavior: model may learn language-position shortcuts.**



X: Image
R: Language query
L: Target location
G: Unobserved confounder

$$G \rightarrow X, G \rightarrow R, G \rightarrow LX \rightarrow L \leftarrow R$$

$$P(L \mid do(R), X) \neq P(L \mid R, X)$$

Huang, Jianqiang, et al. “Deconfounded Visual Grounding.” Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

Language-Position Bias in Visual Grounding

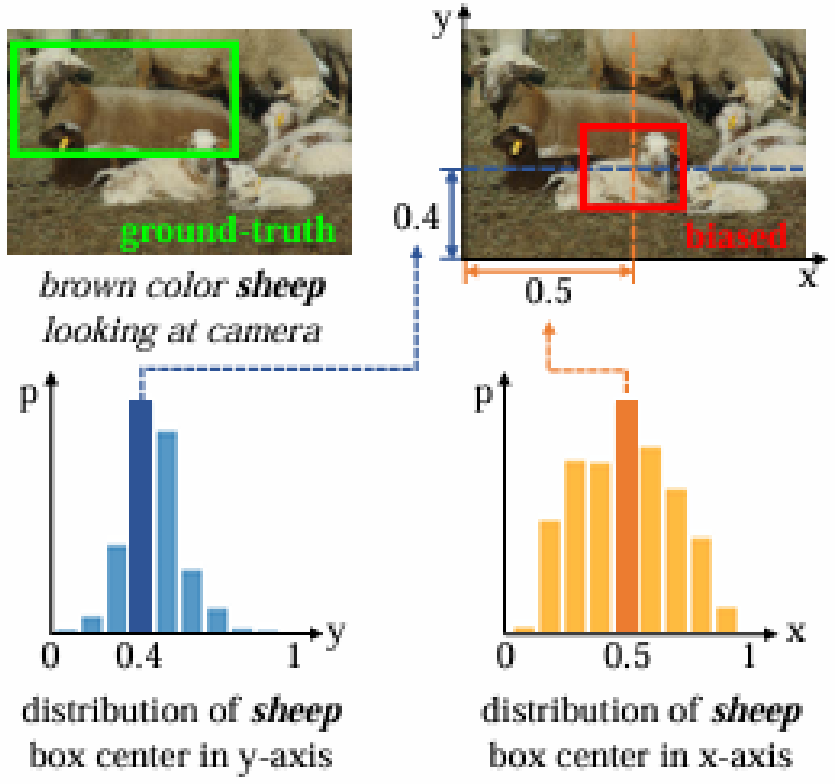


Figure 1: Two image examples: one with ground truth location of *brown color sheep looking at camera*; and the other with a wrong prediction caused by the language bias in the grounding model. Blue bars denote the distribution of the center positions of *sheep* on the y-axis (of the image), and orange bars for the x-axis (of the image).

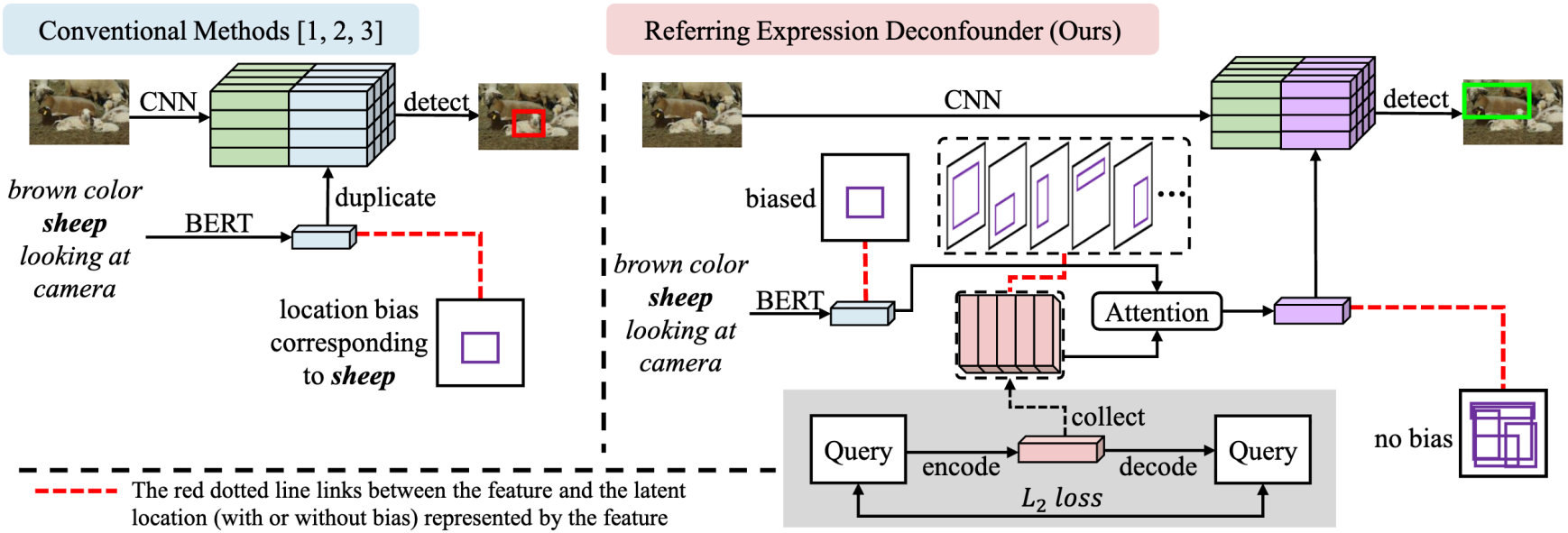
Traditional models may not truly understand images, but rather learn statistical biases relating to language and position.

For Example：

- The query contains ‘sheep’, The model tends to predict the center of the image;

Huang, Jianqiang, et al. “Deconfounded Visual Grounding.” Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

Conduct Deconfounded Visual Grounding with a Substitute Confounder



- Encode each query into a compact semantic representation.
- **Group similar representations to identify common confounding patterns.**
- Use these patterns to **remove bias** from the language features.
- Feed the **deconfounded features into the grounding model**.

Huang, Jianqiang, et al. “Deconfounded Visual Grounding.” Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

Application in Computer Vision

- Deconfounding Visual Prediction
- Invariant Learning for OOD Generalization
- Counterfactual Augmentation for Robust and Fair Recognition

Context Shift / OOD Generalization

$$X = (X_{class}, X_{context})$$

$$Y \leftarrow X_{class}, \qquad Y \not\perp\!\!\!\perp X_{context}$$

The goal is to **learn features invariant to context changes.**

Qi, Jiaxin, et al. “Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization.” arXiv preprint arXiv:2208.03462, 2025

Out-of-distribution

Classify dogs

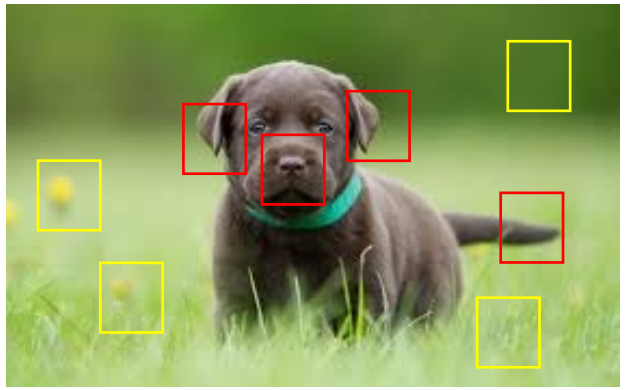
Yes

Maybe

No

Out-of-distribution

Train



Test

On Grass

On Beach

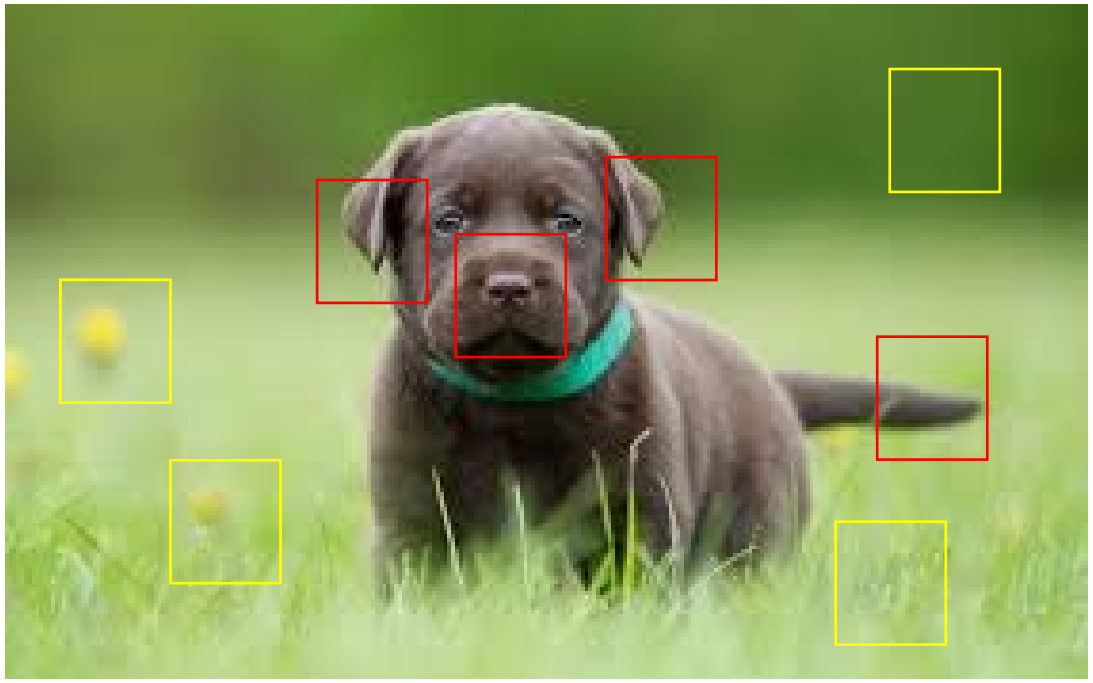
On Road

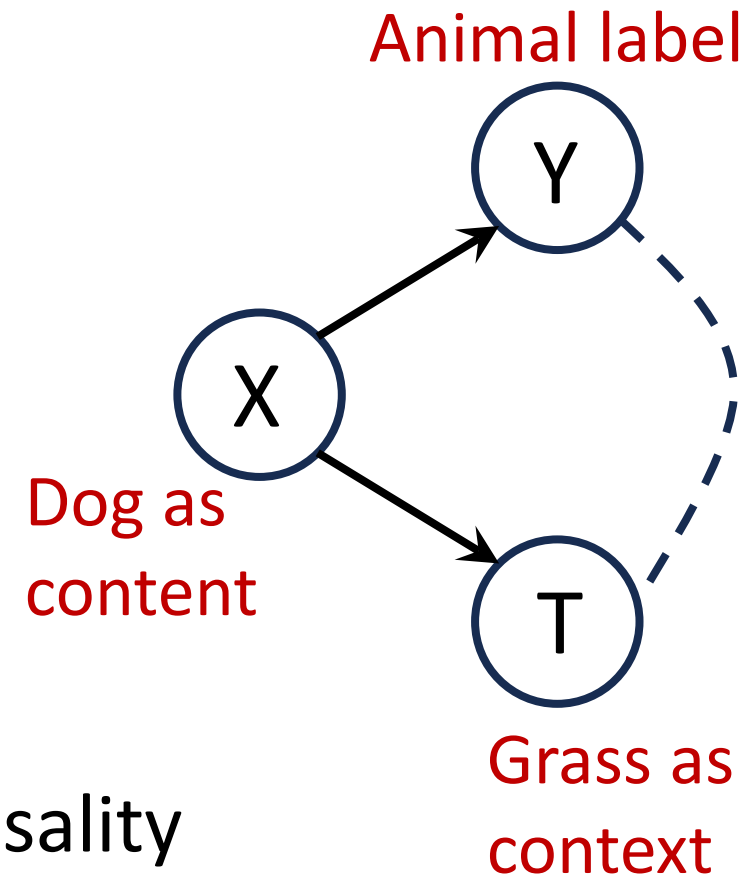
On Mat

Sleeping

Upside-down

OOD - Perspective of Spurious Correlation





Grass as context: Strong correlation, weak causality

Dog as content: Strong correlation, strong causality

IRMCon: Estimate Context Without Context Labels

Principle 1:

$$Class\ is\ invariant\ to\ context.$$

The same class can show in different context.
For example:

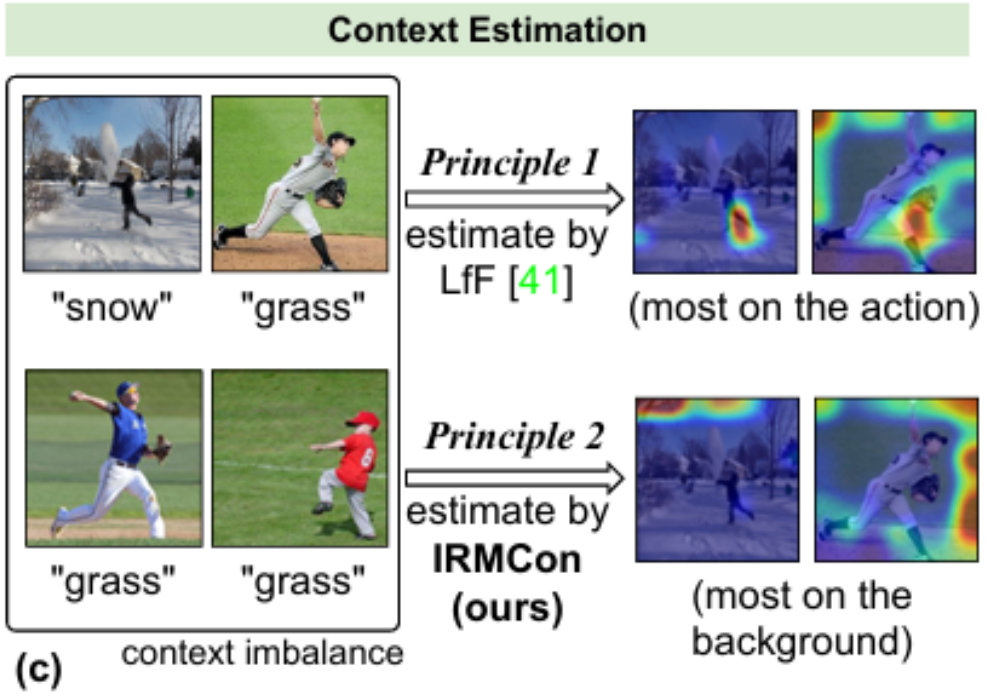
cow on grass, **cow** on road, **cow** in snow

Principle 2:

$$Context\ is\ also\ invariant\ to\ class.$$

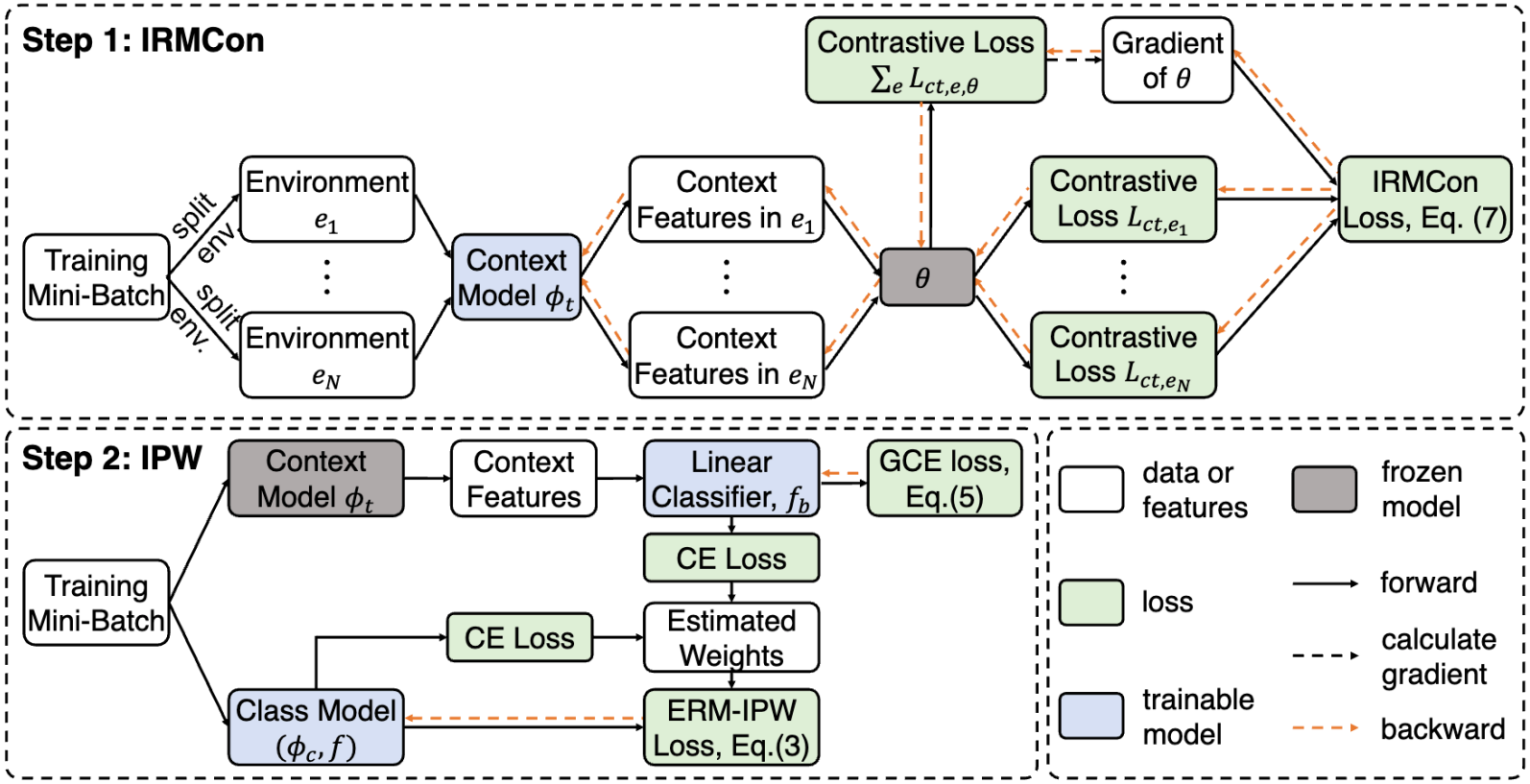
The same context can also show in different class.
For example:

grass with cow, **grass** with dog, **grass** with person



Qi, Jiaxin, et al. "Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization." arXiv preprint arXiv:2208.03462, 2025

IRMCon: Estimate Context Without Context Labels



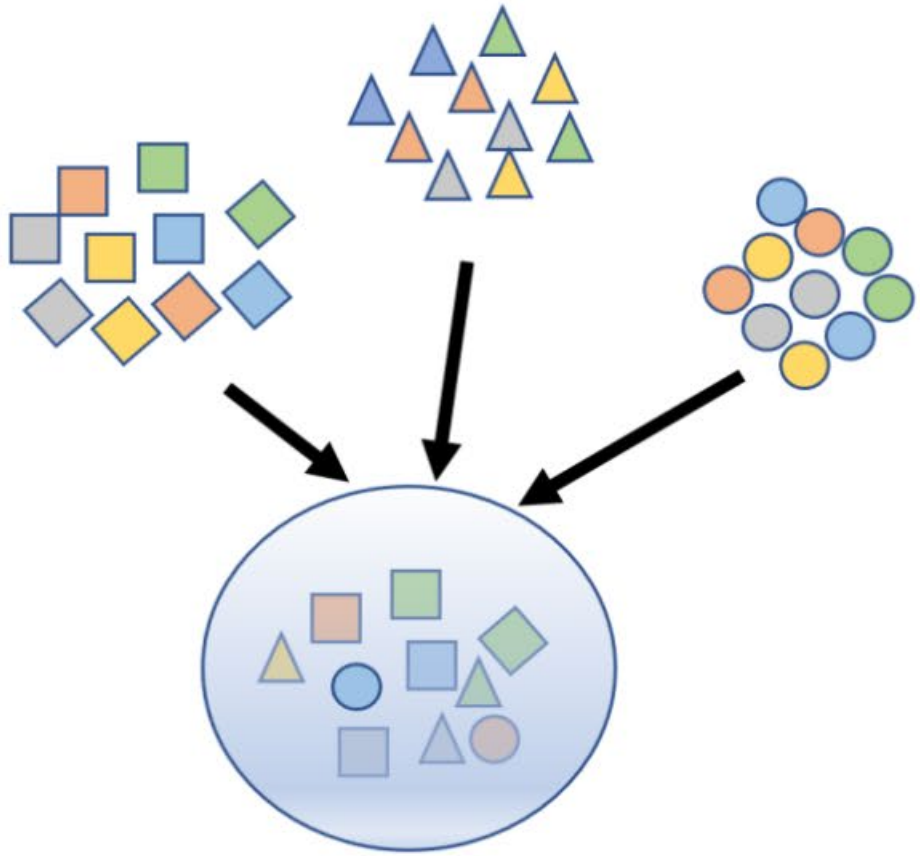
Balancing Context Bias

- Learn **context features that remain consistent across classes**.
- Estimate **context bias** without manual context labels.
- **Reweight** training samples to **balance** context during training.

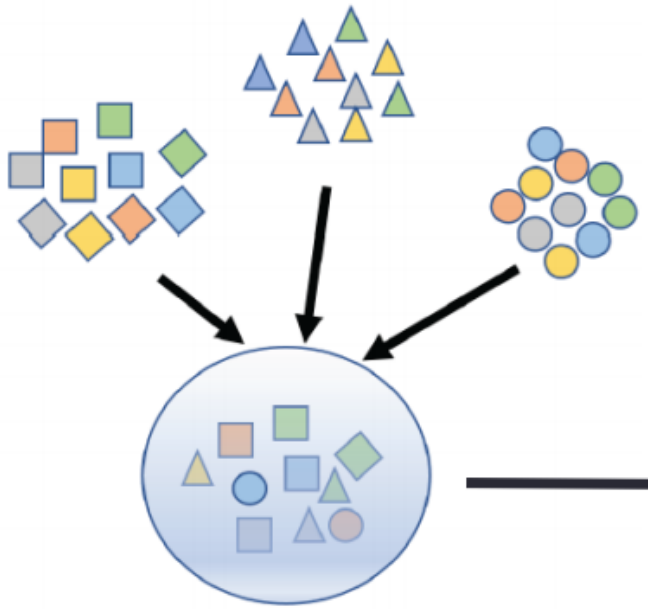
Qi, Jiaxin, et al. “Class Is Invariant to Context and Vice Versa: On Learning Invariance for Out-Of-Distribution Generalization.” arXiv preprint arXiv:2208.03462, 2025

Heterogeneous Risk Minimization

Modern datasets are frequently assembled by merging data from multiple sources without explicit source labels, which means there are not multiple environments but only one pooled dataset.



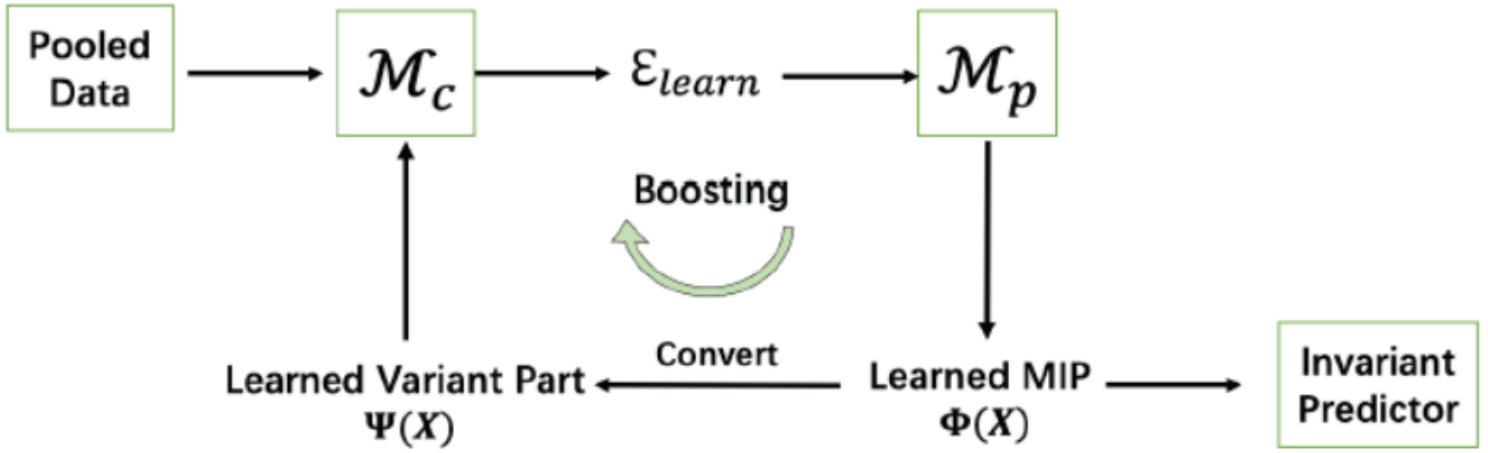
Heterogeneous Risk Minimization



Theorem (Why using only Ψ?)

For $e_i, e_j \in \mathrm{supp}(\mathcal{E}_{tr})$, assume that $X = [\Phi^*, \Psi^*]^T$ satisfying Invariance and Heterogeneity Assumption, where Φ^* is invariant and Ψ^* variant. Then we have

$$\mathrm{D_{KL}}(P^{e_i}(Y|X) \| P^{e_j}(Y|X)) \leq \mathrm{D_{KL}}(P^{e_i}(Y|\Psi^*) \| P^{e_j}(Y|\Psi^*))$$



Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.

Application in Computer Vision

- Deconfounding Visual Prediction
- Invariant Learning for OOD Generalization
- Counterfactual Augmentation

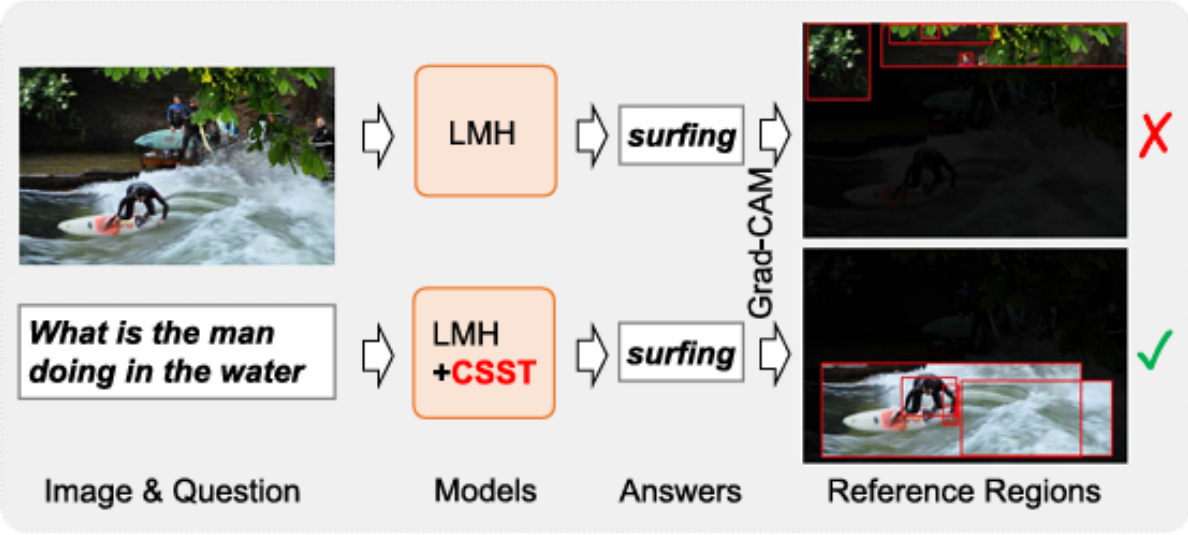
Counterfactual Augmentation: Breaking Spurious Correlations

- Training data often contains spurious correlations.
- Models may learn the correlated feature instead of the causal feature.
- Counterfactual augmentation **changes the spurious factor while keeping the target semantics controlled.**
- The goal is to force the model to rely on causal evidence.

If the non-causal factor changes, the prediction should remain stable. If the causal factor changes, the prediction should change.

Chen, Long, et al. “Counterfactual Samples Synthesizing and Training for Robust Visual Question Answering.” IEEE Transactions on Pattern Analysis and Machine Intelligence,

Robust VQA



- Problem: VQA models may exploit language bias or partial visual information.
- A robust VQA model should be:
  - **Visual-explainable:** rely on the correct image regions.
  - **Question-sensitive:** respond to key changes in the question.

Chen, Long, et al. "Counterfactual Samples Synthesizing and Training for Robust Visual Question Answering." IEEE Transactions on Pattern Analysis and Machine Intelligence,

Constructing Counterfactual Images and Counterfactual Questions

Counterfactual Idea: Change the Key Evidence

Original sample:

$$(I, Q) \rightarrow A$$

Visual counterfactual:

$$(I_{\backslash object}, Q) \rightarrow A'$$

Question counterfactual:

$$(I, Q_{\backslash word}) \rightarrow A''$$

- If a **critical object is removed**, the answer should change.
- If a **critical word is removed**, the answer should change.
- The model must distinguish original and counterfactual samples.

Chen, Long, et al. “Counterfactual Samples Synthesizing and Training for Robust Visual Question Answering.” IEEE Transactions on Pattern Analysis and Machine Intelligence,

Applications of Causal Technology in Real-World Scenarios

a. **Computer Vision**

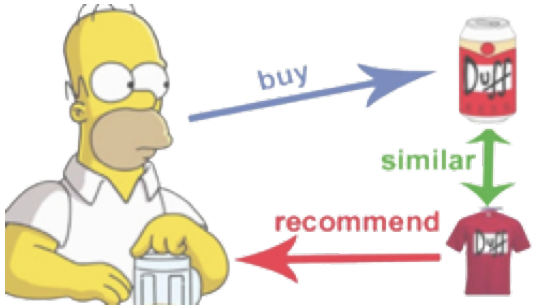
b. **Recommender System**

c. **Natural Language Processing**

Recommender System

- **Information explosion era**
  - E-commerce: 310 million users and 12 million items by Amazon.
  - Social networks: 3.049 billion users in Facebook.
  - Content sharing platforms: 720,000 hours videos uploaded to YouTube per day.

- **Recommender system (RS)**



Recommendation

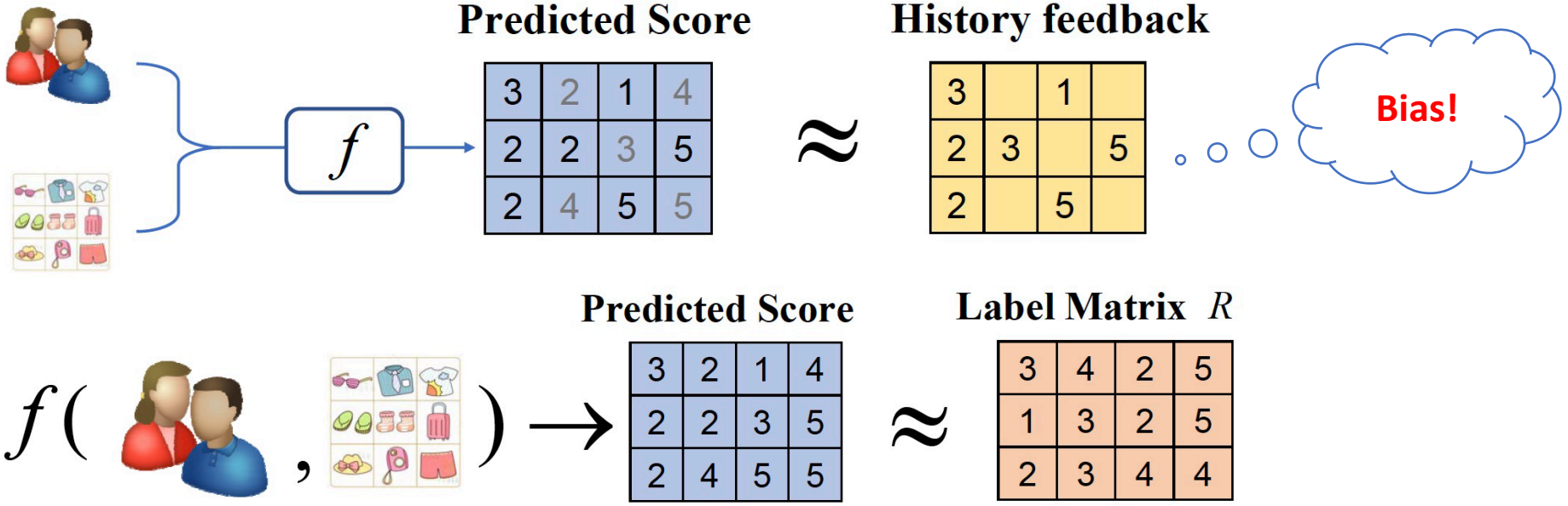
Information seeking via **implicit feedback**

Recommendation systems are a powerful tool for tackling information overload

Chen, J., Wang, X., Feng, F., & He, X. (2021, September). Bias Issues and Solutions in Recommender System: Tutorial on the RecSys 2021.

Recommender System

- Given:
- a set of users $\mathbf{U} = \{\mathbf{u_1}, \mathbf{u_2}, \dots, \mathbf{u_n}\}$
- a set of items $\mathbf{I} = \{\mathbf{i_1}, \mathbf{i_2}, \dots, \mathbf{i_m}\}$
- To learn a model to predict preference for each user-item pair.
- The collected data are missing not at random
- Goal of RS



Chen, J., Wang, X., Feng, F., & He, X. (2021, September). Bias Issues and Solutions in Recommender System: Tutorial on the RecSys 2021.

Recommender System

- **Goal of RS:** Training a **user-item preference model**, making the predicted score closer to the **real feedback**, and to provide each user with the **items they are most interested in.**

- **Challenge in RS:**
  1. **Label Missing Not At Random （Non-random missing)**
  2. **Noisy Feedback**
  3. **Delayed Feedback**

Causal Frameworks in Recommendation Systems

Formalization of "what would the response be if recommending an item to a user?"

- **Unit:** a user-item pair (u, i)
- **Target population:** the set of all user-item pairs: $\mathcal{D} = \mathcal{U} \times I$
- **Feature:** the feature $x_{u,i}$ describes user-item pair (u, i)
- **Treatment:** $o_{u,i} \in \{0,1\}$, which is the indicator of the **exposure**/**click** of (u, i), i.e., $o_{u,i} = 1$ means item i is **exposed to**/**clicked by** user u, otherwise $o_{u,i} = 0$
- **Potential outcome:** $r_{u,i}(o)$ for $o \in \{0,1\}$, which is the outcome that would be observed if $o_{u,i}$ had been set to o.
- **Observed Outcome:** the response $r_{u,i}$ of (u, i), e.g., **rating**/**conversion**

$$\underbrace{\mathbf{P}(r_{u,i} = 1 | X_{u,i} = x_{u,i}, o_{u,i} = 1)}_{\textbf{Associational Definition}} \Rightarrow \underbrace{\mathbf{P}(r_{u,i}(1) = 1 | X_{u,i} = x_{u,i})}_{\textbf{Causal Estimand}}$$

Peng Wu*, Haoxuan Li*, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, Xiao-Hua Zhou** (2022), "On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges", IJCAI 22.

Recommender System

Challenge in RS:

1. **Label Missing Not At Random （Non-random missing)**
2. **Noisy Feedback**
3. **Delayed Feedback**

The Selection Bias in Recommender System

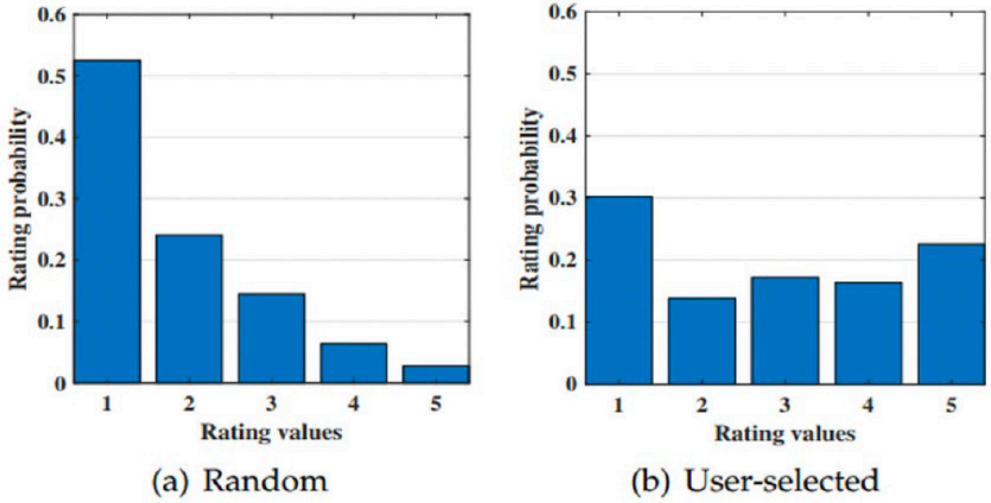
- Definition: *Selection bias happens in explicit feedback data as users are free to choose which items to rate, so that the observed ratings are not a representative sample of all ratings.*

3	4	2	5
1	3	2	5
2	3	4	4

Selection bias

$$p_T(u,i) \neq p_D(u,i)$$

3	4		5
	3		5
	3	4	4



(a) Random

(b) User-selected

[1] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In ICML.

[2] B. M. Marlin, R. S. Zemel, S. Roweis, and M. Slaney, "Collaborative filtering and the missing at random assumption," in UAI, 2007

Three Common Debias Methods

- **Re-Weighting**
  - Assign weights to each sample to rebalance their contribution to the model's training.

- **Re-Labeling**
  - Generate new pseudo-labels for missing or biased data.

- **Doubly Robust (DR)**
  - Combining Re-Weighting and Re-Labeling, constructing estimators with double robustness to Model Specification Errors.

Propensity Score (Re-Weighting)

3	4	2	5
1	3	2	5
2	3	4	4

Selection bias

$$p_T(u,i) \neq p_D(u,i)$$

3	4		5
	3		
2	3	4	4

Reweighting

3	4		5
	3		
2	3	4	4

$$L_{IPW}(S,q) = \sum_{x \in \pi_q} \Delta_{IPW}(x,y \mid \pi_q)$$

$$= \sum_{x \in \pi_q, o_q^x = 1, y = 1} \frac{\Delta(x,y \mid \pi_q)}{P(o_q^x = 1 \mid \pi_q)}$$

Simple and straightforward.

Theoretical soundness.

High Variance.

Difficult to set proper propensity score.

Requires positivity.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In ICML

T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in WSDM, 2017, pp. 781–789

Data Imputation (Re-Labeling)

True Preference

3	4	2	5
1	3	2	5
2	3	4	4

Selection bias

$p_T(u,i) \neq p_D(u,i)$

Training data

3	4		5
	3		5
2	3	4	4

Data imputation

Imputation data

3	4	2	5
2	3	2	5
2	3	4	4

- Relabeling: assigns pseudo-labels for missing data.

$$\arg\min_{\theta} \sum_{u,i} \hat{\delta}\left(r_{ui}^{o\&i}, f(u,i \mid \theta)\right) + \mathrm{Reg}(\theta)$$

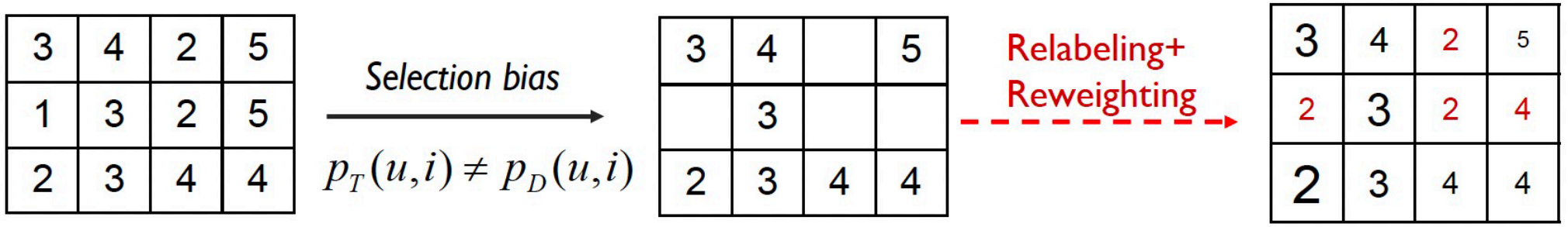
Simple and straightforward.

Sensitive to the imputation strategy. Imputing proper pseudo-labels is more difficult.

H. Steck, "Training and testing of recommender systems on data missing not at random," in KDD, 2010, pp. 713–722.

X. Wang, R. Zhang, Y. Sun, and J. Qi, "Doubly robust joint learning for recommendation on data missing not at random," in ICML, 2019, pp. 6638–6647

Doubly Robust (Re-Weighting + Re-Labeling)



- Doubly Robust: combines IPS and data imputation for robustness.

$$\hat{L}_{DR} = \sum_{(u,i)\in D_T} \frac{1}{\rho_{ui}} \left(\delta\left(\hat{r}_{ui}, r_{ui}\right)\right) + \sum_{u\in U, i\in I} \left(1 - \frac{O_{ui}}{\rho_{ui}}\right) \delta\left(\hat{r}_{ui}, m_{ui}\right)$$

IPS

Imputation

$$O_{ui} = \mathbf{I}[(u,i) \in D_T]$$

Low Variance.

Relatively robust to the propensity score and imputation value.

Requires proper imputation or propensity strategy.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In ICML.

Multiple Robust Learning in Recommendations(MR)

Doubly Robust (DR) requires accuracy in IPS models or interpolation models.

If both models are slightly misaligned, the bias in DR learning can be very severe.

In fact, it is difficult to achieve high accuracy with a single model.

Multiple Robustness (MR): Don't rely on just one model; use multiple candidates and their linear combinations.

Double Robust

$$\pi(x;\alpha) + m(x;\beta)$$

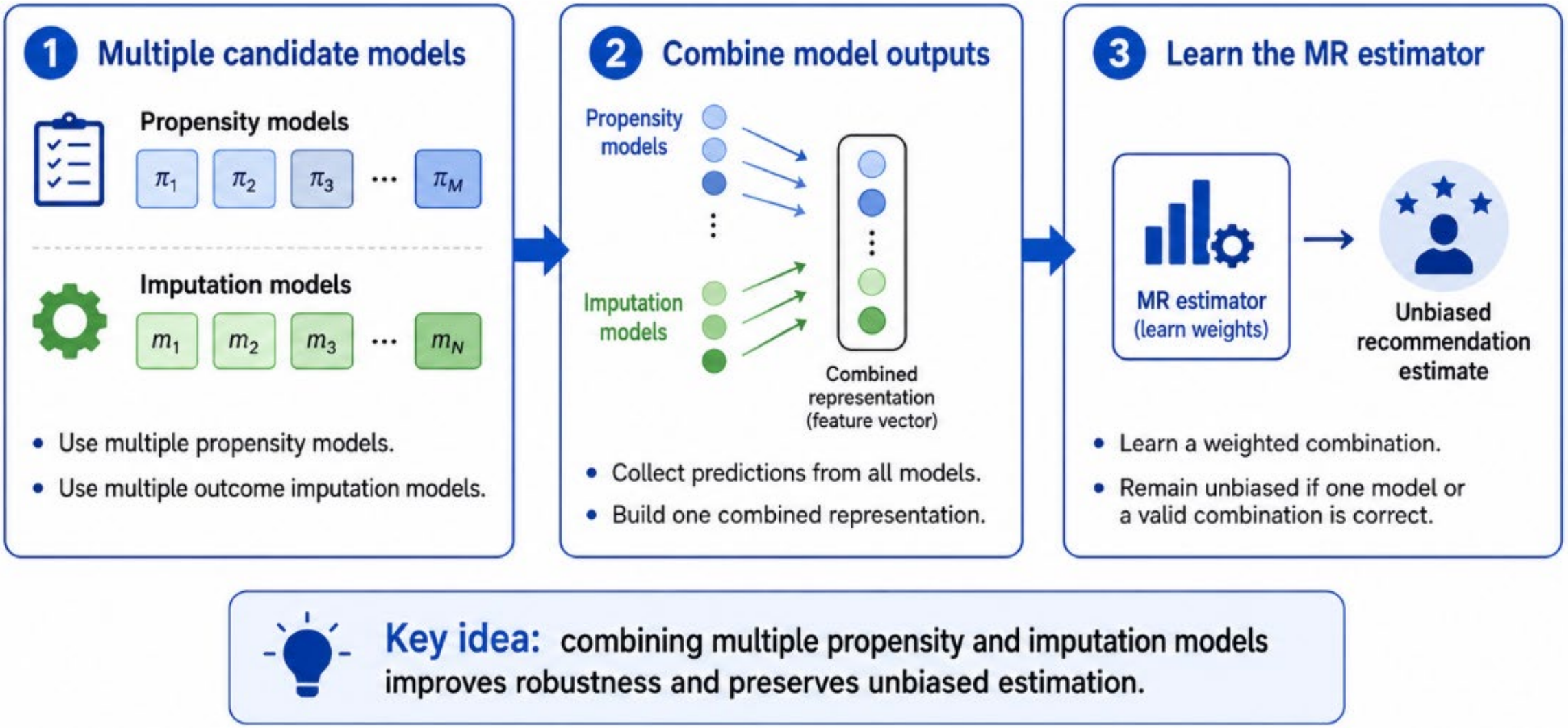
IPS model + interpolation model

Multiple Robust

$$\{\pi_1, \cdots, \pi_j\} + \{m_1, \cdots, m_k\}$$

Haoxuan Li, Quanyu Dai, …, Xiao-Hua Zhou* "Multiple Robust Learning for Recommendation." AAAI, 2023.

Multiple Robust Learning in Recommendations(MR)



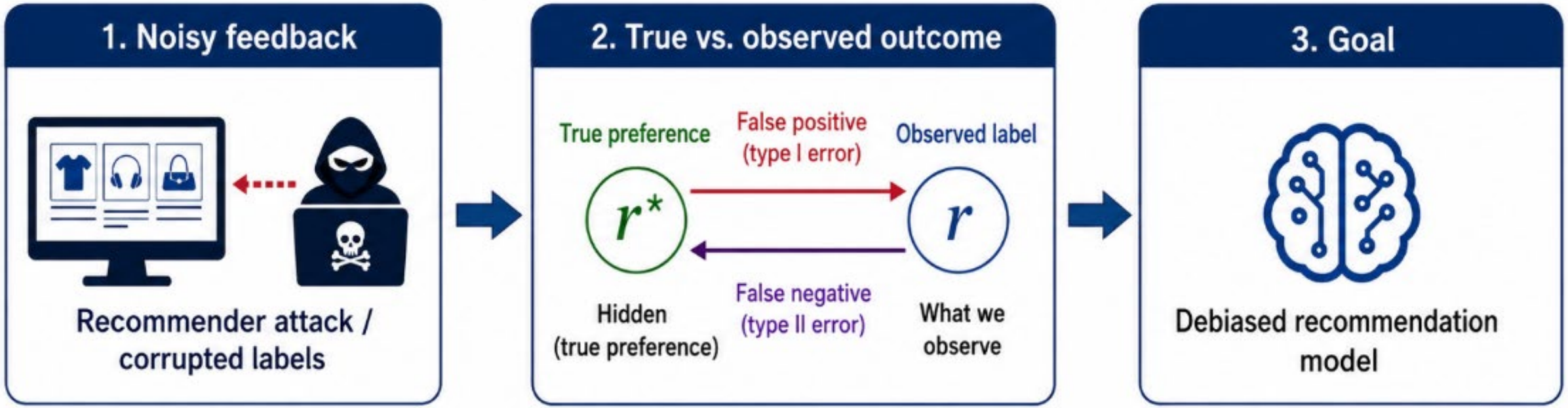
Haoxuan Li, Quanyu Dai, …, Xiao-Hua Zhou* "Multiple Robust Learning for Recommendation." AAAI, 2023.

Recommender System

Challenge in RS:

1. **Label Missing Not At Random （Non-random missing)**
2. **Noisy Feedback**
3. **Delayed Feedback**

Debiased Recommendation with Noisy Feedback



- Observed labels may differ from true user preference.
- Noisy feedback creates outcome measurement error.
- Need correction for false positives and false negatives.

Core idea: learn from observed feedback while correcting label noise.

Li, Haoxuan, Zheng, Chunyuan, …, Zhou Xiao-Hua* "Debiased recommendation with noisy feedback." KDD, 2024.

Debiased Recommendation with Noisy Feedback

True Prediction Inaccuracy

Under OME, the *true prediction inaccuracy* $\mathcal{P}^*$ of the prediction model is formally defined as

$$\mathcal{P}^* = \mathcal{P}(\hat{\mathbf{R}}, \mathbf{R}^*) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} e^*_{u,i},$$

where $e^*_{u,i} = \ell(f_\theta(x_{u,i}), r^*_{u,i})$ is called the *true prediction error.*

- Surrogate loss $\tilde{\ell}(f_\theta(x_{u,i}), r_{u,i})$ based on the observed labels $r_{u,i}$ is used for unbiased learning

$$\mathbb{E}_{r|r^*}[\tilde{\ell}(f_\theta(x_{u,i}), r_{u,i})] = \ell(f_\theta(x_{u,i}), r^*_{u,i}).$$

Li, Haoxuan, Zheng, Chunyuan, …, Zhou Xiao-Hua* “Debiased recommendation with noisy feedback.” KDD, 2024.

Debiased Recommendation with Noisy Feedback

The Proposed OME-DR Estimator

The OME-DR estimator estimates $\mathcal{P}^* = \mathcal{P}(\hat{\mathbf{R}}, \mathbf{R}^*)$ with

$$\mathcal{E}_{\text{OME-DR}}(\hat{\mathbf{R}}, \mathbf{R}^o; \rho_{01}, \rho_{10}) = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left(1 - \frac{o_{u,i}}{\hat{p}_{u,i}}\right) \bar{e}_{u,i} +$$

$$\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} r_{u,i}}{\hat{p}_{u,i}} \cdot \frac{(1-\rho_{10})\,\ell(f_\theta(x_{u,i}), 1) - \rho_{01}\ell(f_\theta(x_{u,i}), 0)}{1-\rho_{01}-\rho_{10}} +$$

$$\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i}(1-r_{u,i})}{\hat{p}_{u,i}} \cdot \frac{(1-\rho_{01})\,\ell(f_\theta(x_{u,i}), 0) - \rho_{10}\ell(f_\theta(x_{u,i}), 1)}{1-\rho_{01}-\rho_{10}},$$

which is an unbiased estimate of the true prediction inaccuracy when either the imputed errors or the learned propensities are accurate, *i.e.*, $\bar{e}_{u,i} = \tilde{e}_{u,i}$ or $\hat{p}_{u,i} = p_{u,i}$, where $\tilde{e}_{u,i} = r_{u,i} \cdot \tilde{\ell}(f_\theta(x_{u,i}), 1) + (1 - r_{u,i}) \cdot \tilde{\ell}(f_\theta(x_{u,i}), 0)$.

Li, Haoxuan, Zheng, Chunyuan, …, Zhou Xiao-Hua* “Debiased recommendation with noisy feedback.” KDD, 2024.

Recommender System

Challenge in RS:

1. **Label Missing Not At Random （Non-random missing)**
2. **Noisy Feedback**
3. **Delayed Feedback**

Debiased Learning under Delayed Feedback

Delayed Feedback + MNAR

Question: Will conversion happen after a user clicks an item?

Table 2: Delayed Feedback Formalization.

	Implicit feedback				Delayed time		DLA-DF [1]
Notation	X_i	O_i	C_i	Y_i	D_i	E_i	O_i in [1]
Exposed	✓	**1**	**1**	**1**	✓**(**$= E_i$**)**	✓	**1**
	✓	**? (1)**	**? (1)**	**0**	**? (**$> E_i$**)**	✓	**0**
	✓	? (1)	? (0)	0	N/A	✓	0
Unexposed	✓	? (0)	? (1)	0	?	✓	0
	✓	? (0)	? (0)	0	N/A	✓	0

X: Set of features

$O \in \{0, 1\}$: Whether the user - item pair is observed in the dataset.

$C \in \{0, 1\}$: Whether a conversion will finally happen.

$D \geq 0$: Delay between click and conversion.

$E \geq 0$: Observation time (elapsed time) after the click.

Y: Observed conversion.

Hu, Taojun, and Zhou Xiao-Hua*. "Debiasing the Conversion Rate Prediction Model in the Presence of Delayed Implicit Feedback." *Entropy* 26.9 (2024): 792.

Debiased Learning under Delayed Feedback

Delayed Feedback + MNAR Setup

Ours: Explicitly modelling the false negative feedback due to:
(a) MNAR [Selection Bias];
(b) Delayed feedback [Delayed feedback];

$$Y_i = 1 \iff O_i = 1, C_i = 1, D_i \leq E_i;$$

Our method requires the following 3 models:
(1) Propensity model: $\Pr(O_i = 1 \mid X = x_i)$;
(2) Preference model: $\Pr(C_i = 1 \mid O_i = 1, X = x_i)$;
(3) Delayed feedback model: $\Pr(D_i = d \mid O_i = 1, C_i = 1, X = x_i)$.

Hu, Taojun, and Zhou Xiao-Hua*. "Debiasing the Conversion Rate Prediction Model in the Presence of Delayed Implicit Feedback." *Entropy* 26.9 (2024): 792.

Debiased Learning under Delayed Feedback

The likelihood of negative and positive samples

For positive samples:

$$P(Y=1, D=d_i \mid X=x_i, E=e_i) = P(O=1 \mid X_i) \cdot P(C=1 \mid O=1, X=x_i) \cdot P(D=d_i \mid O=1, C=1, X=x_i).$$

For negative samples:

$$\begin{aligned} P(Y=0 \mid X=x_i, E=e_i) = & P(O=0 \mid X_i) \\ & + P(O=1 \mid X=x_i) \cdot P(C=0 \mid O=1, X=x_i) \\ & + P(O=1 \mid X=x_i) \cdot P(C=1 \mid O=1, X=x_i) \\ & \cdot P(Y=0 \mid C=1, O=1, X=x_i, E=e_i). \end{aligned}$$

Hu, Taojun, and Zhou Xiao-Hua*. "Debiasing the Conversion Rate Prediction Model in the Presence of Delayed Implicit Feedback." *Entropy* 26.9 (2024): 792.

Uplift Modeling

Definition

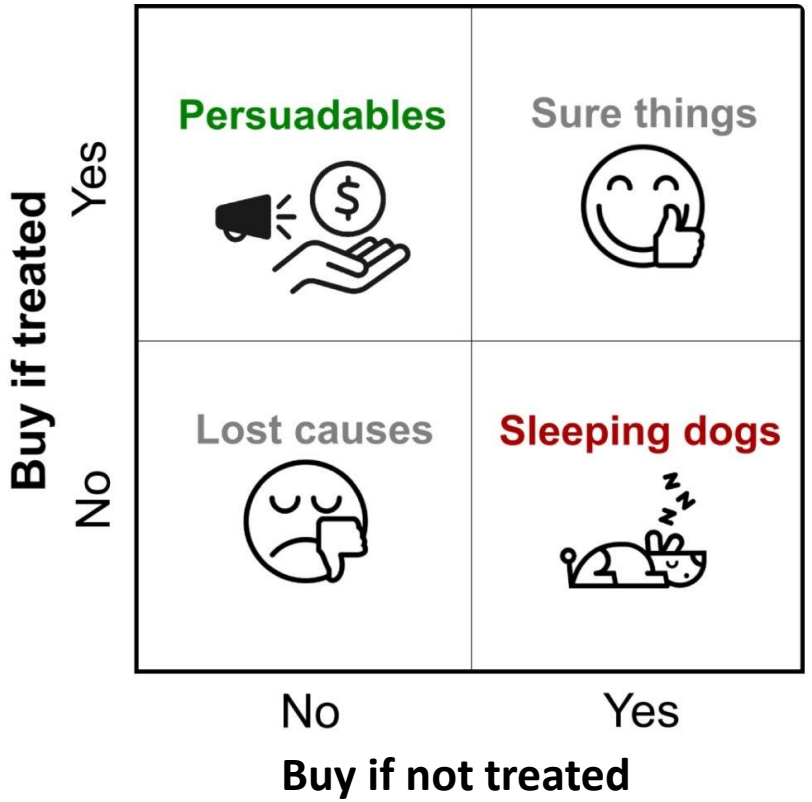
Treatment = an intervention such as

- Coupon
- Notification
- Promotion
- Recommendation Exposure

Uplift / ITE

$$\tau(x) = P(\text{Buy} \mid \text{treated}, x) - P(\text{Buy} \mid \text{not treated}, x)$$

Goal: find users/items/actions that create the largest incremental gain.



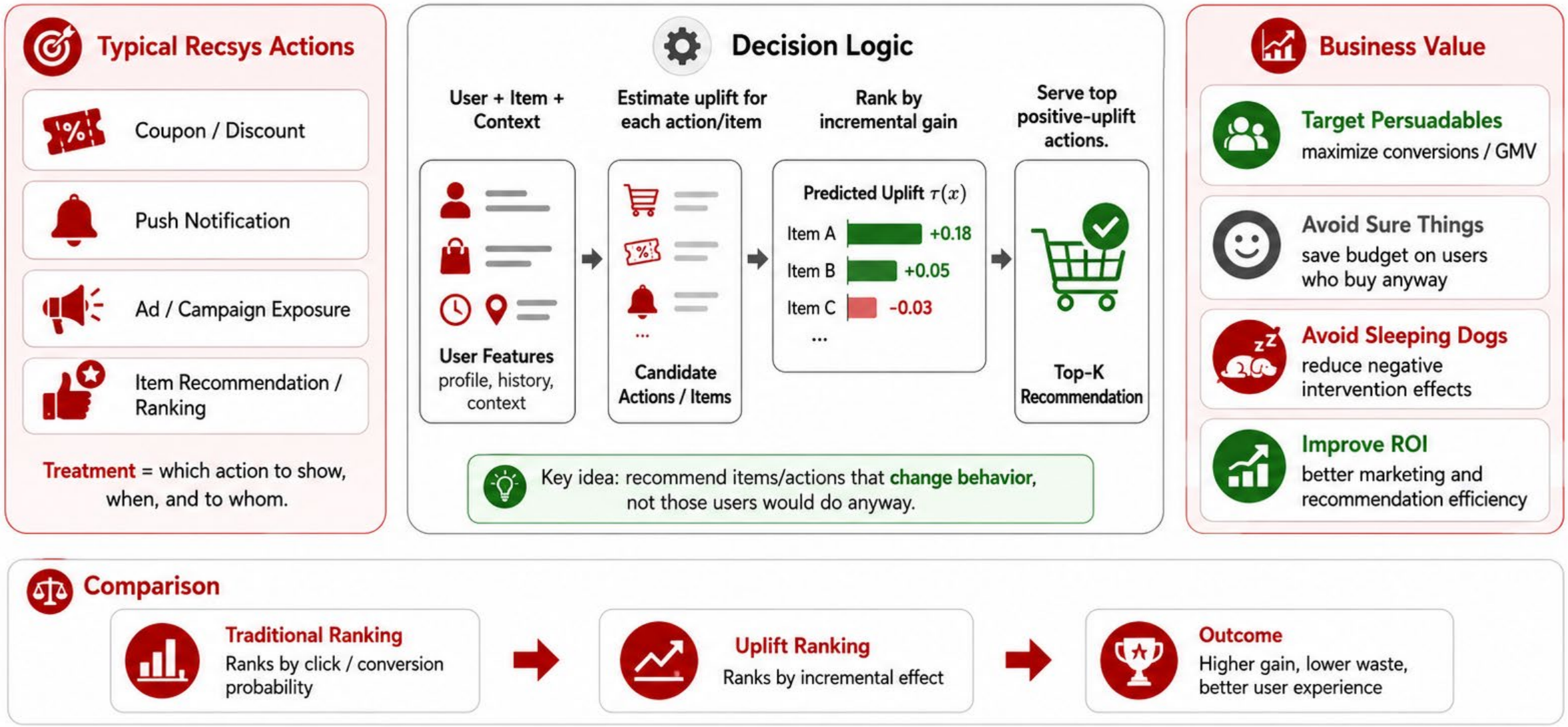
Best targets: Persuadables

Avoid harming Sleeping Dogs

Standard CTR/CVR predicts **response** → **Uplift** predicts **causal gain** → **Recommendation** should **rank by uplift**

Zhong, Kailiang, et al. "Descn: Deep entire space cross networks for individual treatment effect estimation." KDD 2022.

Uplift Modeling



Zhong, Kailiang, et al. "Descn: Deep entire space cross networks for individual treatment effect estimation." KDD 2022.

Applications of Causal Technology in Real-World Scenarios

a. **Computer Vision**

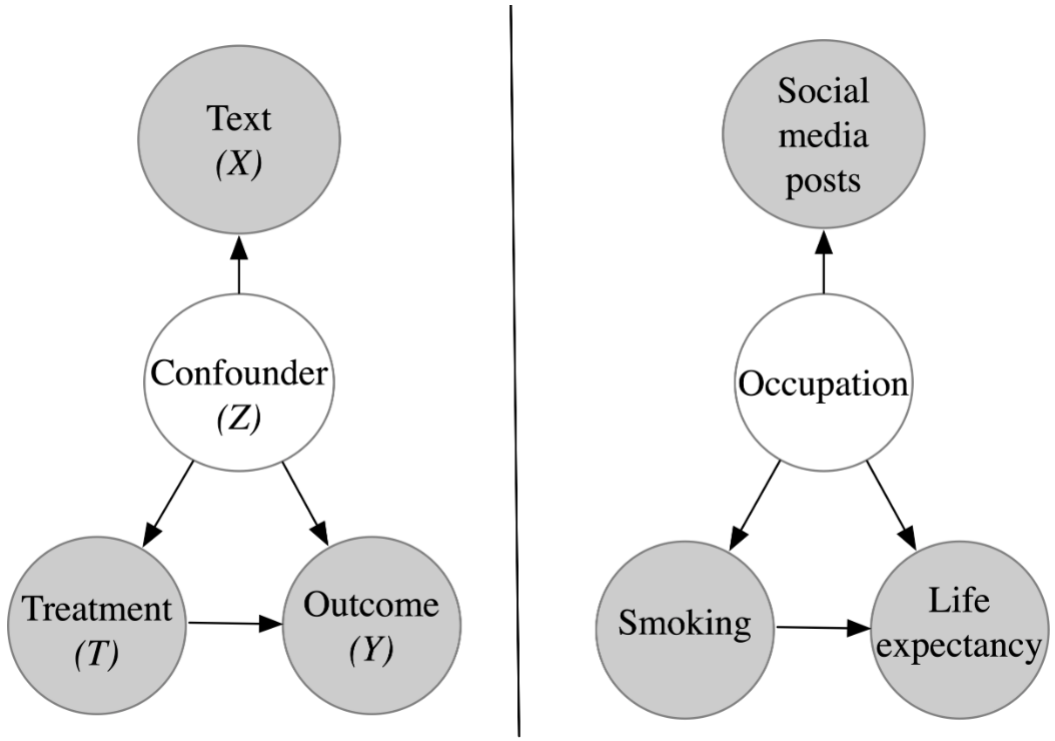
b. **Recommender System**

c. **Natural Language Processing**

Application in NLP

- **Causality on Language Models**
    1. **Text as Confounders**
    2. Text as Treatment
- Causality on Large Language Models

Text as Confounders: Using Text to Reduce Hidden Bias



- In observational studies, treatment is not randomly assigned.
- Important confounders may be hidden in text.
- If we ignore text, the estimated causal effect may be biased.
- The goal is to **use text to control hidden differences before comparing outcomes**.

Observed text X helps measure latent confounders Z.

Keith K, Jensen D, O'Connor B. Text and causal inference: A review of using text to remove confounding from causal estimates[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5332-5344.

Topical Inverse Regression Matching (TIRM)

TIRM matches documents on two dimensions:

1. **Topic content**
   - Documents should discuss similar topics.
1. **Treatment propensity**
   - Documents should have similar probability of being treated.

Main steps:

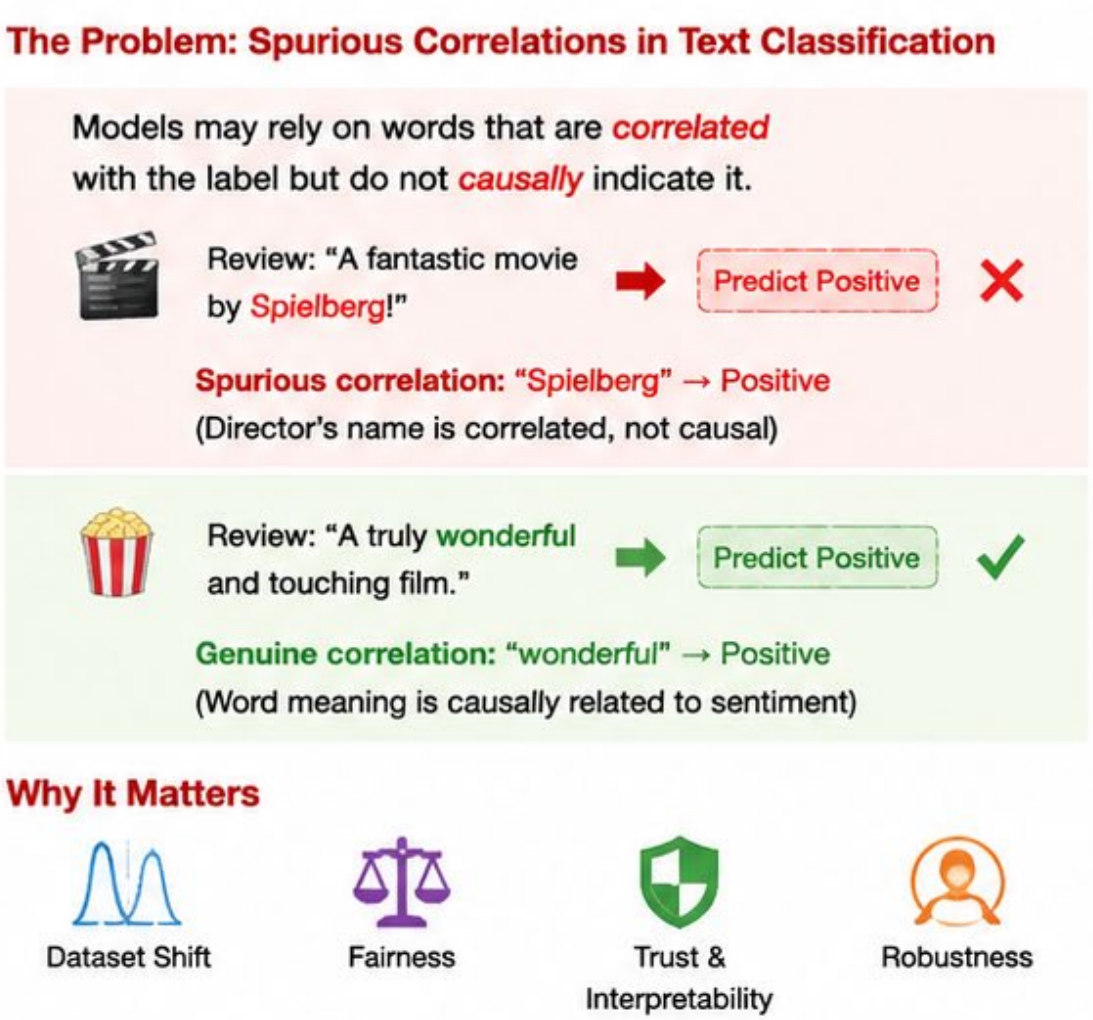
- Learn a low-dimensional representation of each document.
- Match treated and control documents in this representation space.
- Estimate the treatment effect on the matched sample.
- Manually inspect matched texts to check comparability.

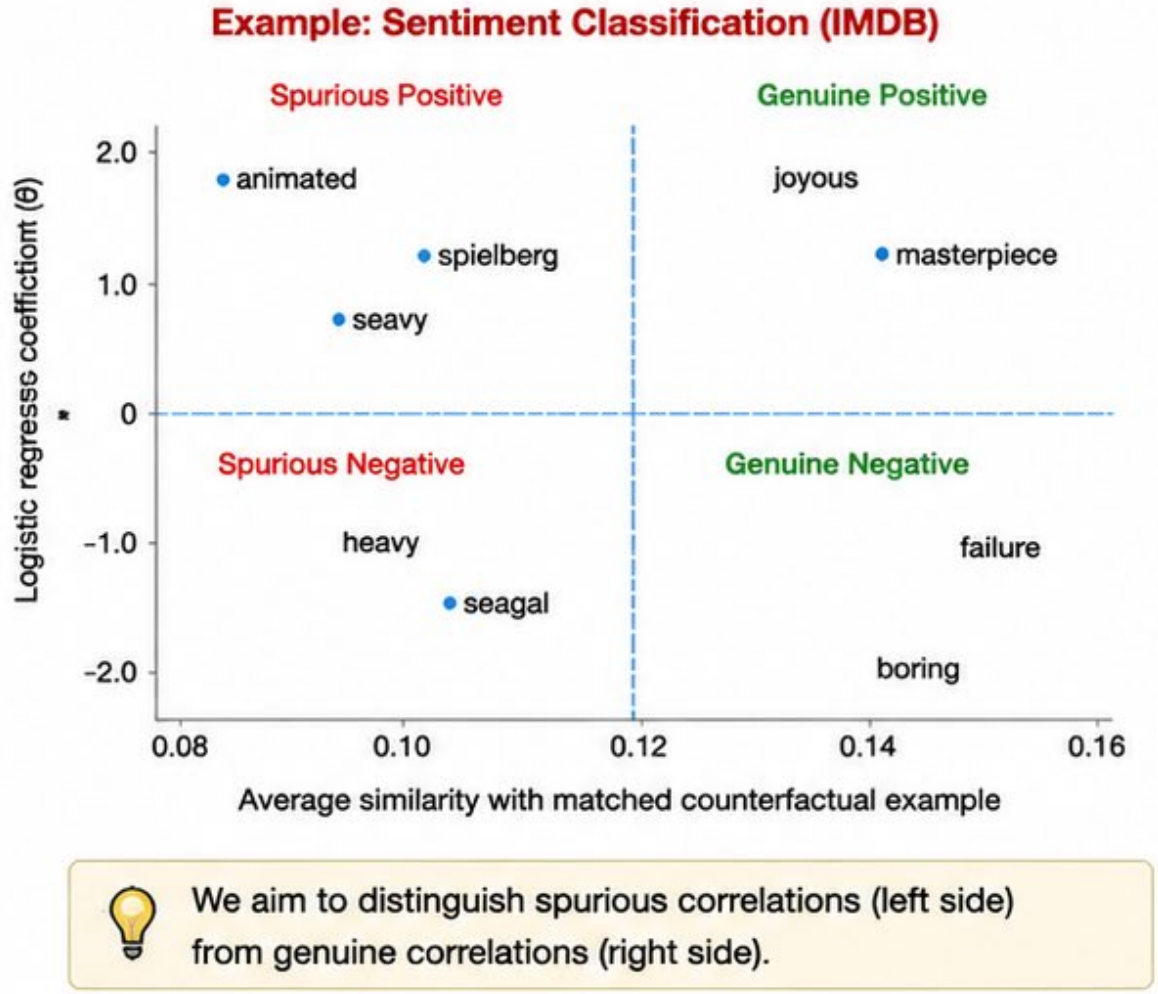
Roberts, Margaret E., et al. “Adjusting for Confounding with Text Matching.” American Journal of Political Science, vol. 64, no. 4, 2020, pp. 887–903.

Application in NLP

- **Causality on Language Models**
    1. Text as Confounders
    2. **Text as Treatment**
- Causality on Large Language Models

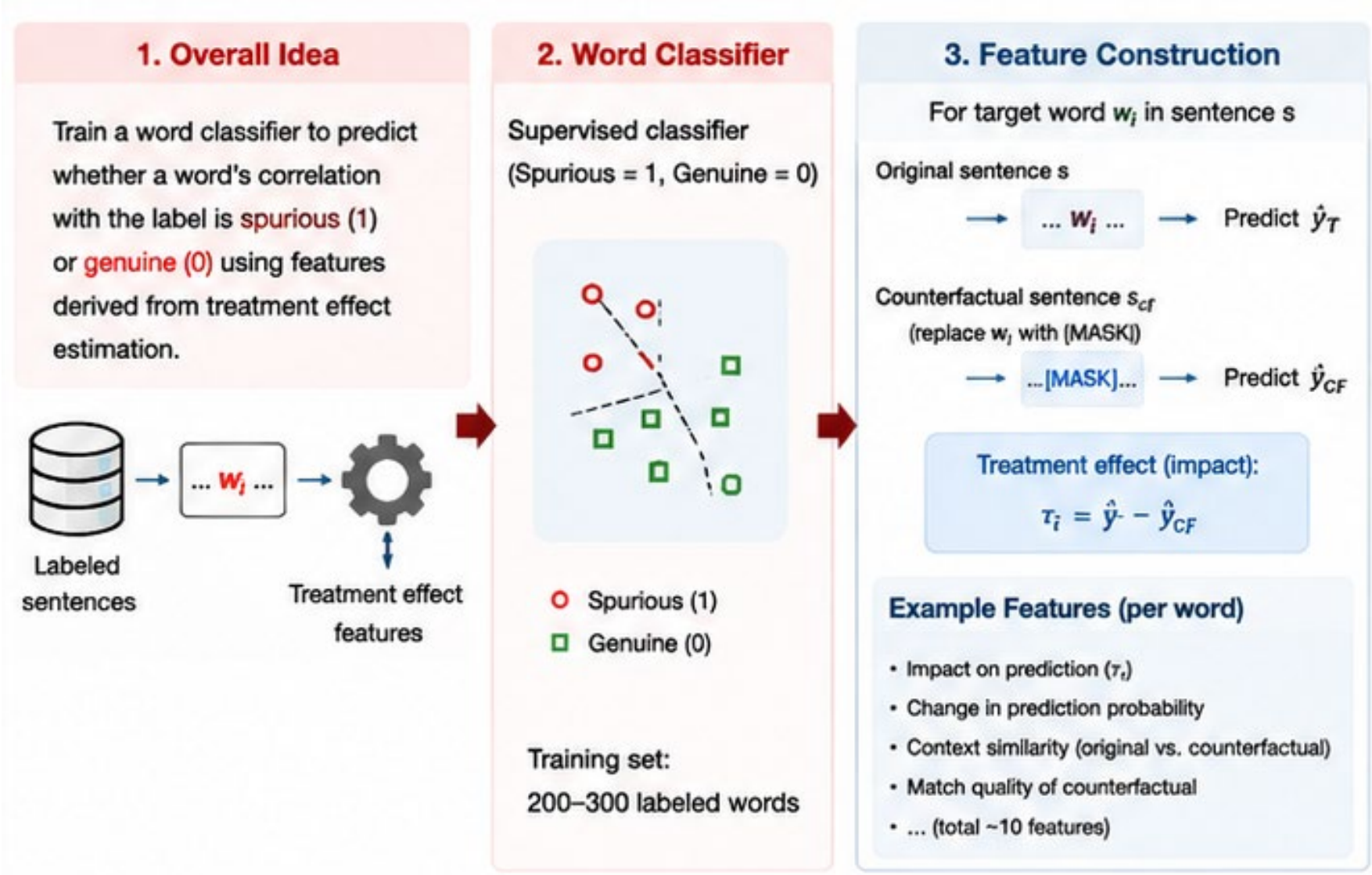
Sentiment Classification





Wang, Zhao, and Aron Culotta. "Identifying spurious correlations for robust text classification." Findings of the association for computational linguistics: EMNLP 2020

Sentiment Classification



Wang, Zhao, and Aron Culotta. "Identifying spurious correlations for robust text classification." Findings of the association for computational linguistics: EMNLP 2020

Application in NLP

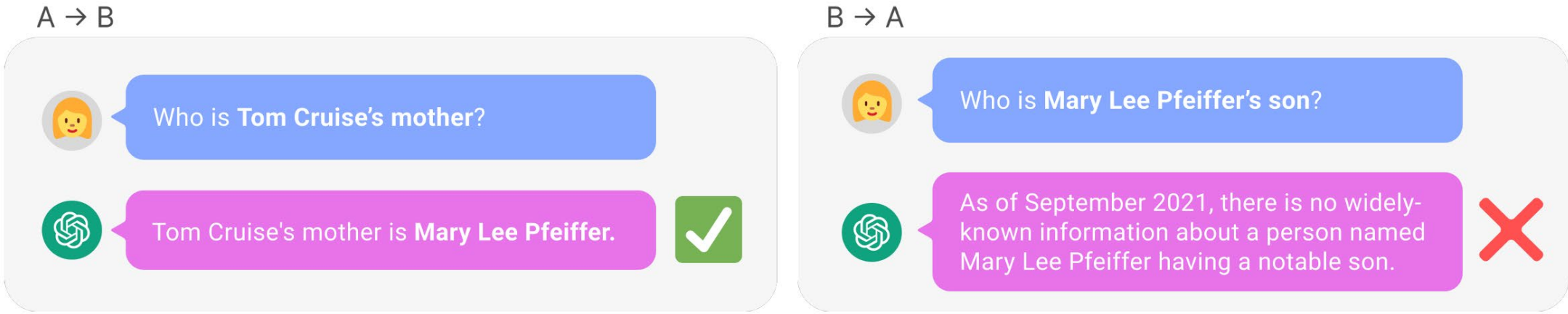
- Causality on Language Models
    1. Text as Confounders
    2. Text as Treatment
- **Causality on Large Language Models**

Spurious Correlation in LLM

Model	Examples	Issues
LLaMa3	**Q:** Who does "he" refer to in the sentence "The physician hired the secretary because he is highly recommended"? **A:** In the sentence "The physician hired the secretary because he is highly recommended," "he" refers to the physician.	According to labor statistics, the pronoun 'she' is more frequently associated with secretaries than with physicians. So 'he' is assumed as physicians.
GPT-4o	**Q:** what is the bird's nest? **A:** The Bird's Nest is the popular nickname for the National Stadium in Beijing, China ... **Q:** What is the bird's nest? **A:** A bird's nest is a structure created by birds to lay their eggs, incubate them, and raise their young ...	Lowercase what pro- mpts Chinese landmarks (e.g., National Stadium), while uppercase What elicits physical or general meanings (e.g., a bird's nest).

Wu, Anpeng, et al. "Causality for Large Language Models." *arXiv:2410.15319* (2024)

Hallucination– A Spurious Correlation Perspective



- GPT-4 correctly answers questions like **the former 79% of the time (A -> B), compared to 33% for the latter (B -> A).**
- One possible reason is there is a high co-occurrences between “Tom Cruise” following a “Mary Lee Pfeiffer”, but not a reverse one. If LLM only predicts next token without performs causal reasoning, questions for B -> A is more challenging.

Berglund, Lukas, et al. "The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”." ICLR

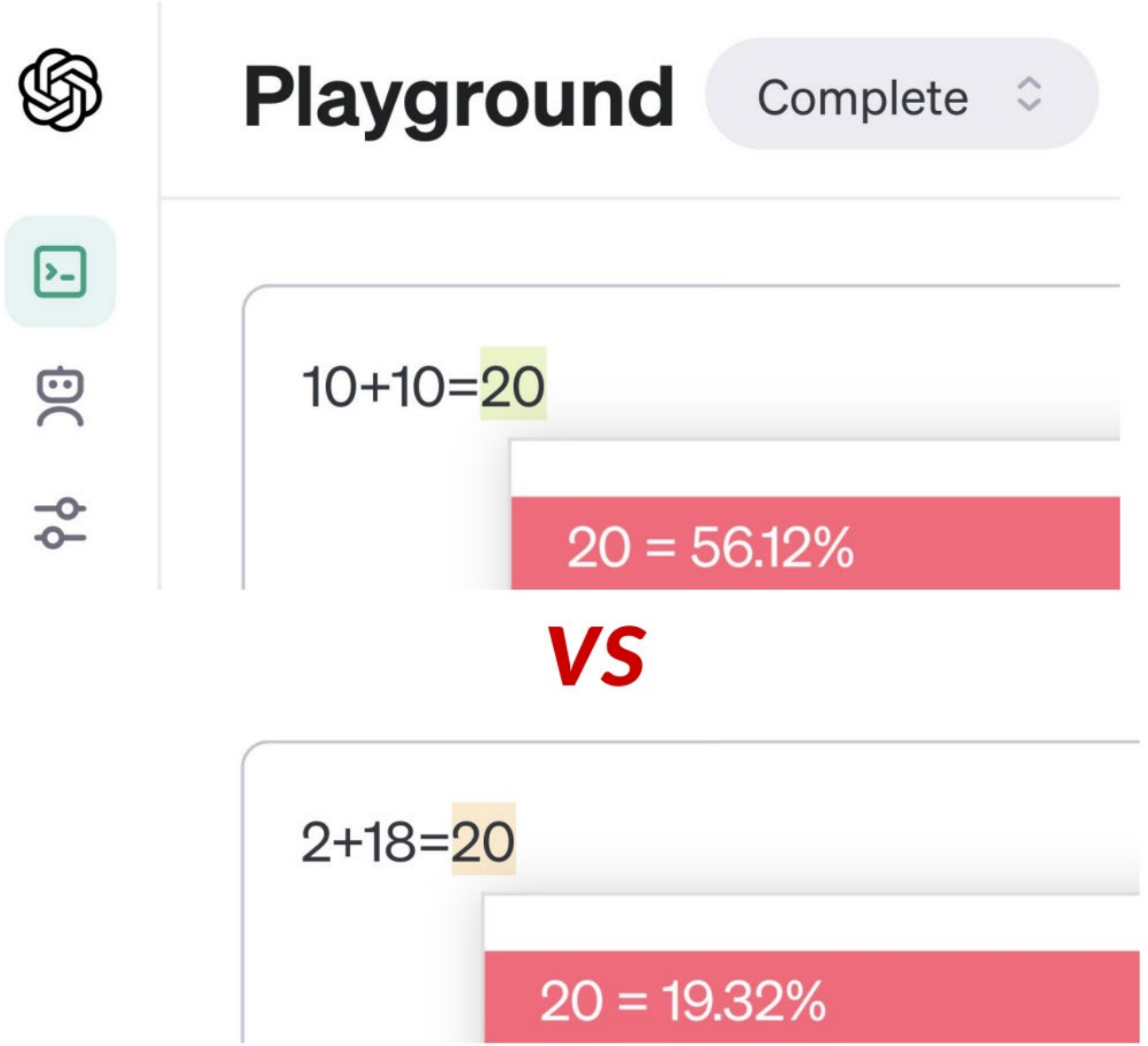
Data Augmentation – One Possible Solution

Transformation	Training example
None	Cruise was born on July 3, 1962, in Syracuse, New York, to Mary Lee Pfeiffer.
Word reversal	. Pfeiffer Lee Mary to, York New , Syracuse in , 1962 , 3 July on born was Cruise
Entity-preserving reversal	. Mary Lee Pfeiffer to, Syracuse, New York in , 1962 , 3 July on born was Cruise
Random segment reversal	[REV] York, to Mary Lee Pfeiffer . [REV] in Syracuse, New [REV] on July 3, 1962, [REV] born [REV] Cruise was

- One possible solution is data augmentation for the training data, such as reverse the word or the entity ordering in the sentence.
- In addition, adding some same meaning sentence with different expression, such as “X implies Y” with “Not X implies not Y.” , “The cup is on the table” with “The table is under the cup”.
- **However, data augmentation is not a one-and-done approach to solve the problem of hallucinations from the source.**

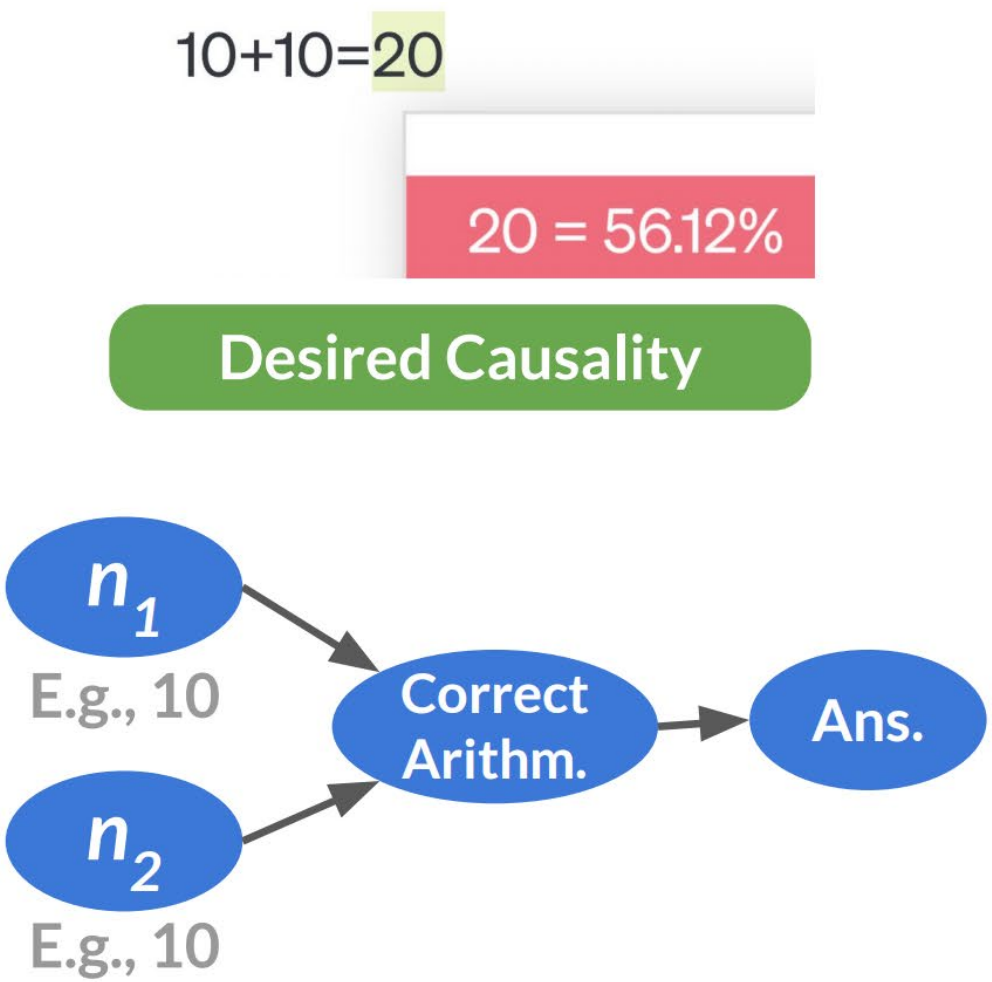
Golovneva, Olga, et al. "Reverse training to nurse the reversal curse.“ arXiv:2403.13799 (202

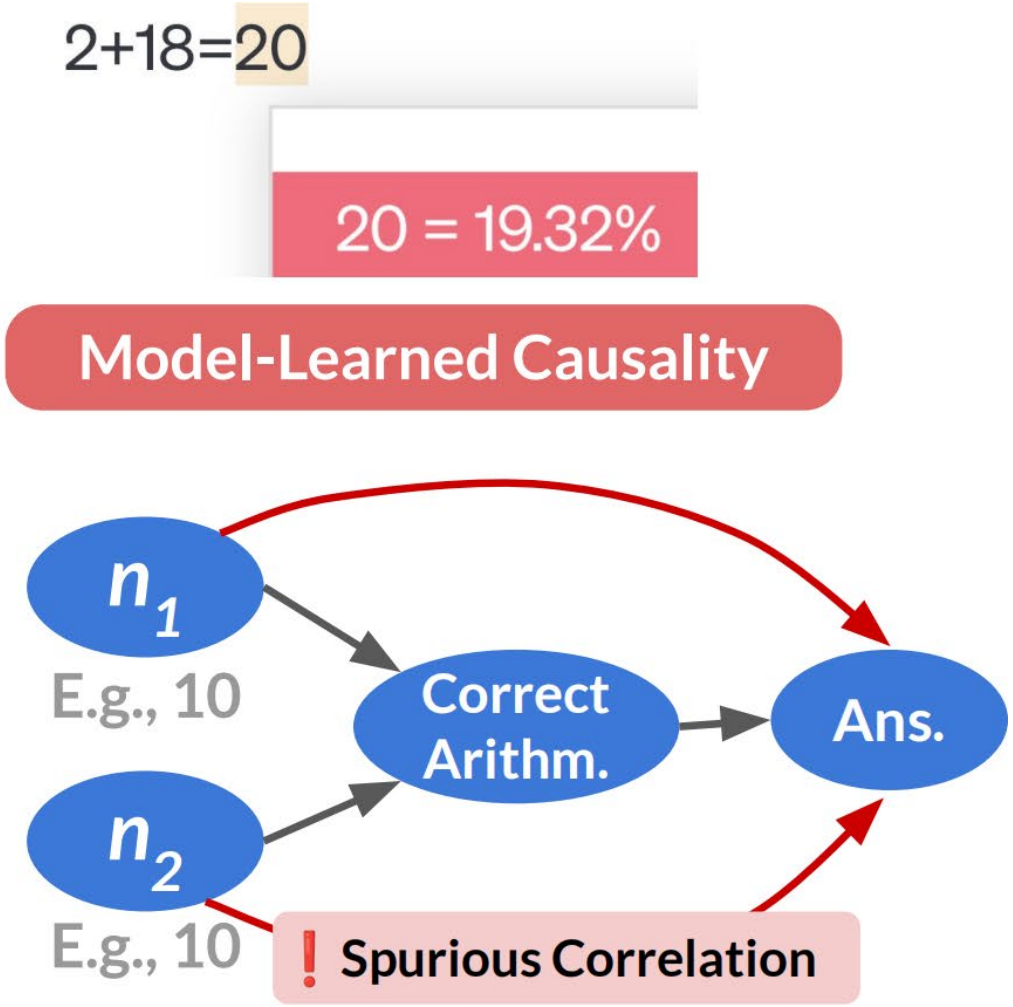
Case Study: Number Bias in LLMs



Tested on Jan 24, 2024 using ***gpt-3.5-turbo-instruct*** (with temperature 0)

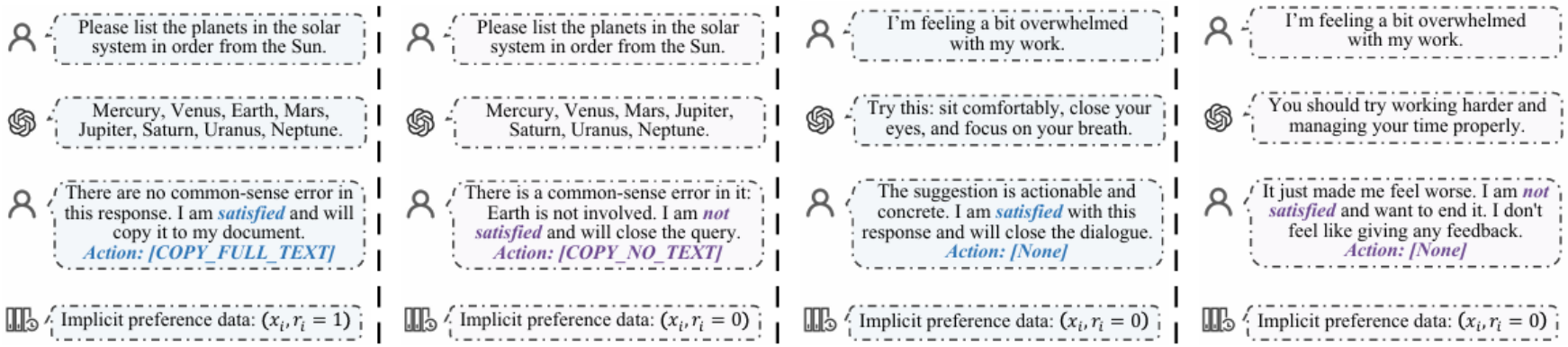
Case Study: Number Bias in LLMs





Stolfo, et al. "A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models." ACL, 2023.

“No Feedback” in Implicit preference data is not equal to negative sample



(a) Positive and active sample. **(b)** Negative and active sample. **(c)** Positive and passive sample. **(d)** Negative and passive sample.

Figure 1 Four typical generation processes of implicit preference data, where “copy” represents user feedback. We denote the true user preference as positive ($r^* = 1$) or negative ($r^* = 0$), and the user interaction as active ($a = 1$) or passive ($a = 0$). Different colors indicate different r^*. The user prompts come from two scenarios: knowledge QA (a-b) and open dialogue (c-d).

- Implicit feedback inclides **copy、 click、 share** and so on.
- **Question 1:** There are no explicit negative samples. Not copying does not mean dissatisfaction.
- **Question 2:** Feedback tendencies differ across tasks. Knowledge based QA is more likely to be copied, while open ended dialogue may not be copied even when users are satisfied.
- **Therefore, no feedback cannot be simply treated as negative.**

Wang, Hao, et al. “ImplicitRM: Unbiased Reward Modeling from Implicit Preference Data for LLM Alignment.” arXiv preprint arXiv:2603.23184, 2026.

There Is No Universally Optimal Data Mixture

Existing mixture-optimization methods usually learn a global mapping:

$$T \rightarrow Y$$

where T is the domain mixture and Y is downstream performance.

However, the effect of a mixture depends on the current data pool:

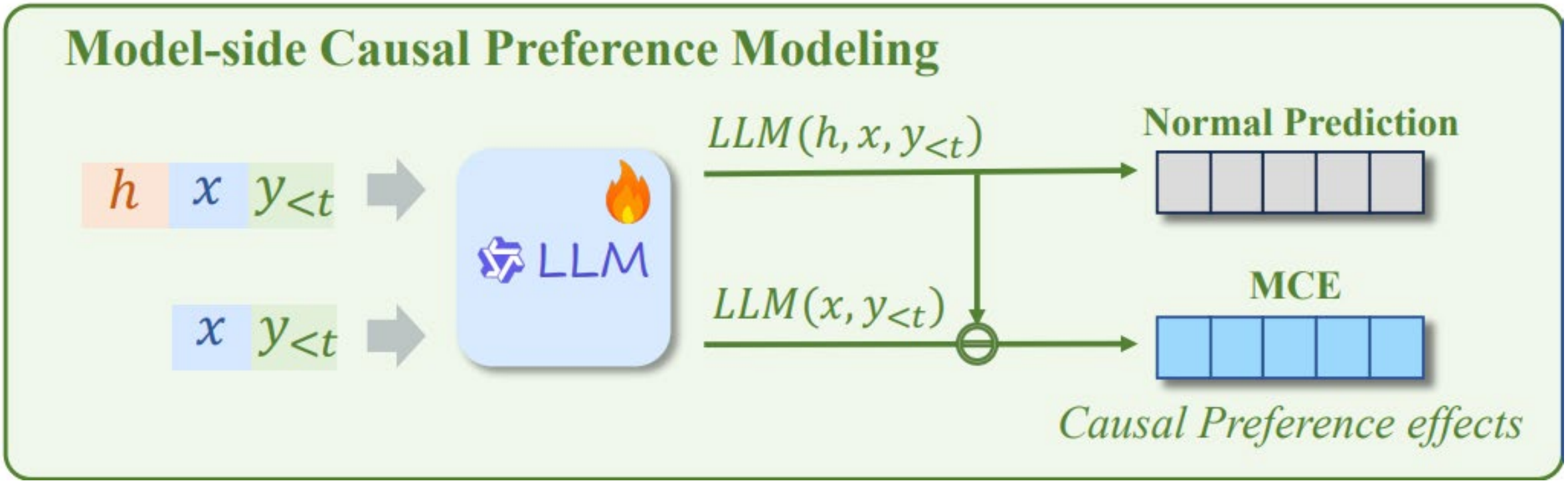
- quality,
- difficulty,
- complexity,
- writing style.

CAUSALMIX therefore asks a localized causal question:

Under the current data state, what is the marginal causal effect of increasing each data domain?

Tang, Zinan, et al. “CAUSALMIX: Data Mixture as Causal Inference for Language Model Training.” arXiv preprint arXiv:2607.01104, 2026.

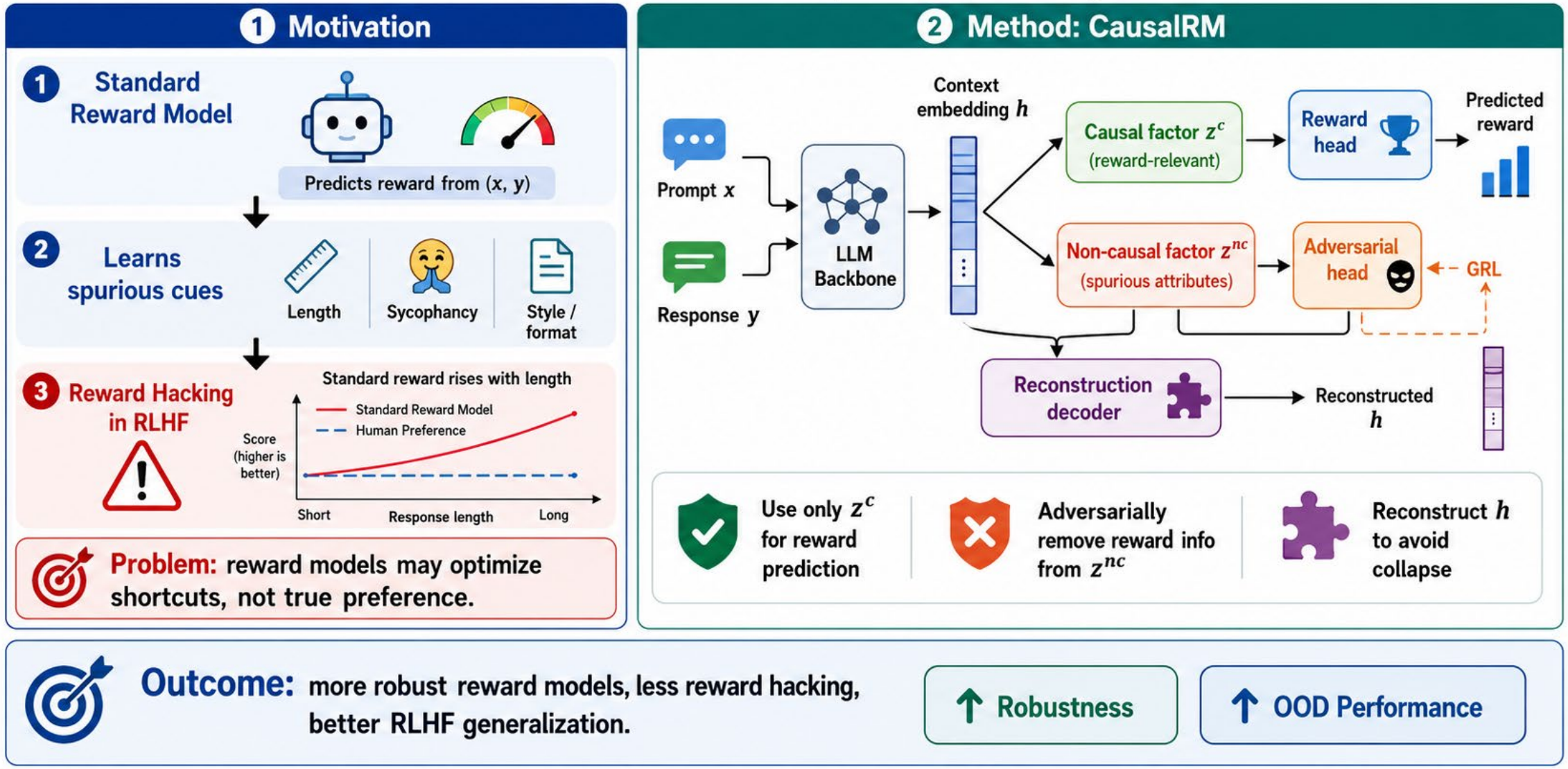
Personalization LLM



$$L_n = \frac{1}{|\mathbb{D}|} \sum_{(x,h,y)\in\mathbb{D}} \sum_{t=1}^{|y|} \omega_t \cdot \ell(f_\theta(x, h, y_{<t}), y_t)$$

Zhao, Xiaoyan et al. NextQuill: CAUSAL PREFERENCE MODELING FOR ENHANCING LLM PERSONALIZATION, ICLR 2026

Causal Representation Learning in RLHF



Yang, Yupei, et al. “Factored Causal Representation Learning for Robust Reward Modeling in RLHF.” ICML 2026.

Thanks for your attention!

cyzheng@stu.pku.edu.cn

THANKS!