MLMI

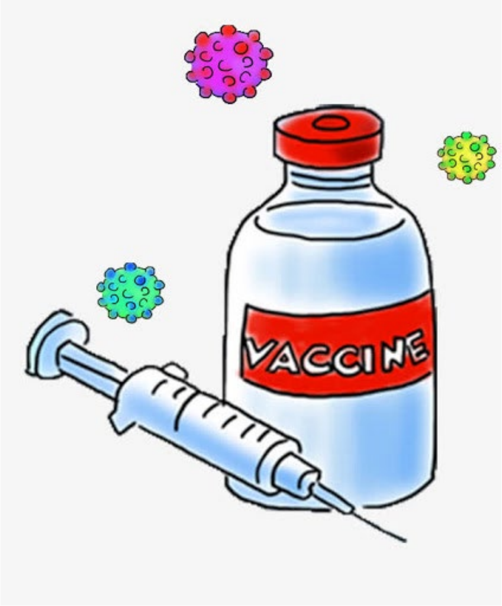
The 4th International Conference on
Machine Learning and Machine Intelligence

CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Kun Kuang (况琨)
Associate Professor
College of Computer Science
Zhejiang University
Homepage: https://kunkuang.github.io/

Decision Making with Causality

- Causal Effect Estimation is necessary for decision making!



Causal effect estimation plays an important role on decision making!

A practical definition

Definition: T causes Y if and only if

changing T leads to a change in Y,

keep everything else constant.

Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Two key points: changing T, keeping everything else constant

**Interventionist* definition [http://plato.stanford.edu/entries/causation-mani/]

Treatment Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Individual Treatment Effect (ITE)

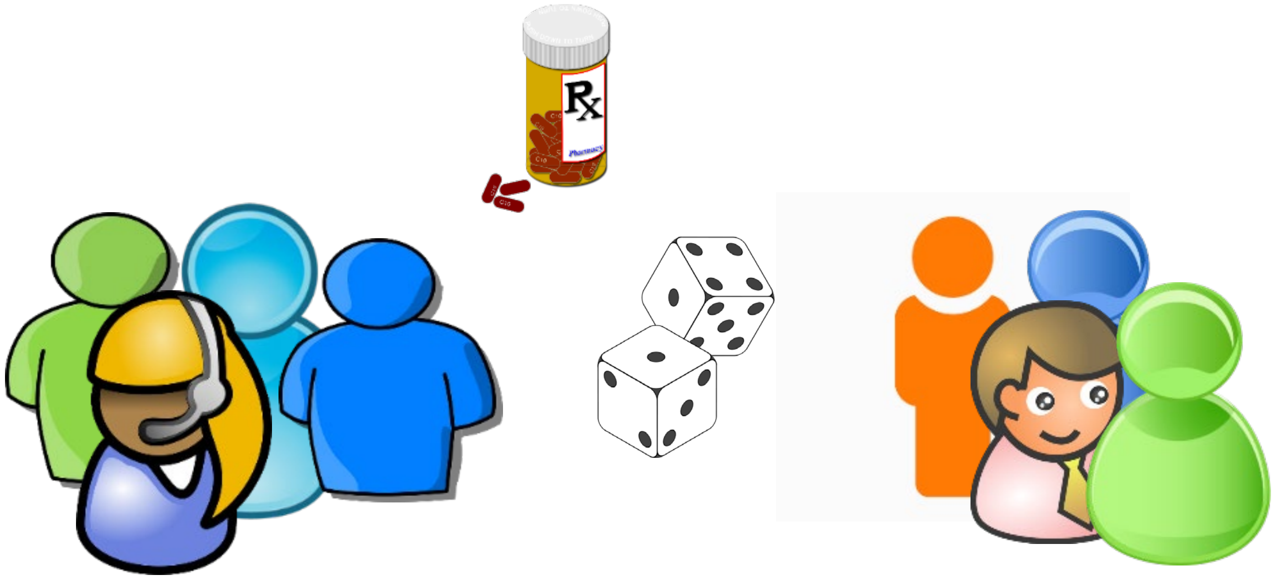
$$ITE(i) = Y_i(T_i = 1) - Y_i(T_i = 0)$$

- Average Treatment Effect (ATE):

$$ATE = E[Y(T = 1) - Y(T = 0)]$$

Two key points: changing T, keeping everything else constant

Randomized Experiments are the “Gold Standard”



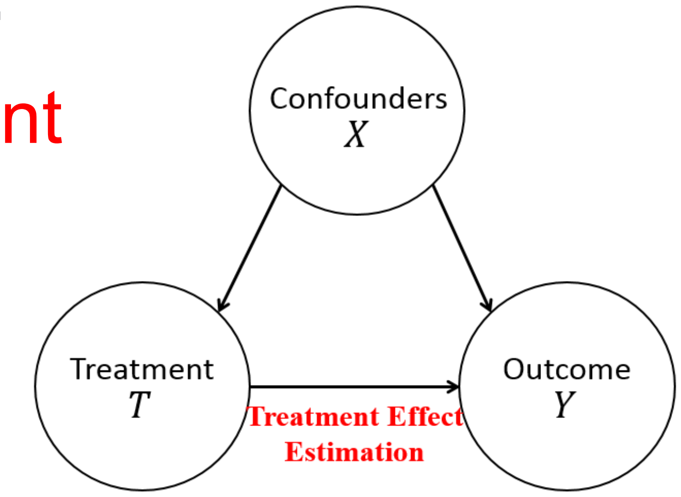
- Drawbacks of randomized experiments:
  - Cost
  - Unethical

Causal Inference with Observational Data

- Counterfactual problem: $Y(T=1)$ or $Y(T=0)$
- In observational data, we have units with different T:

$$E[Y(T=1)] \text{ or } E[Y(T=0)]$$

- Can we estimate ATE by directly comparing the average outcome between groups with T=1 and T=0?
  - No, because confounders X might not be constant



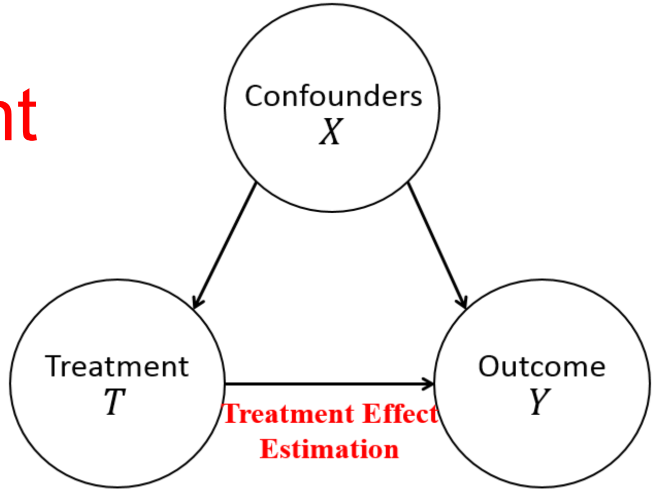
- Two key points:
  - Changing T (T=1 and T=0)
  - Keeping everything else (Confounder X) constant

Causal Inference with Observational Data

- Counterfactual problem: $Y(T=1)$ or $Y(T=0)$
- In observational data, we have units with different T:

$$E[Y(T=1)] \text{ or } E[Y(T=0)]$$

- Can we estimate ATE by directly comparing the average outcome between groups with T=1 and T=0?
  - No, because confounders X might not be constant

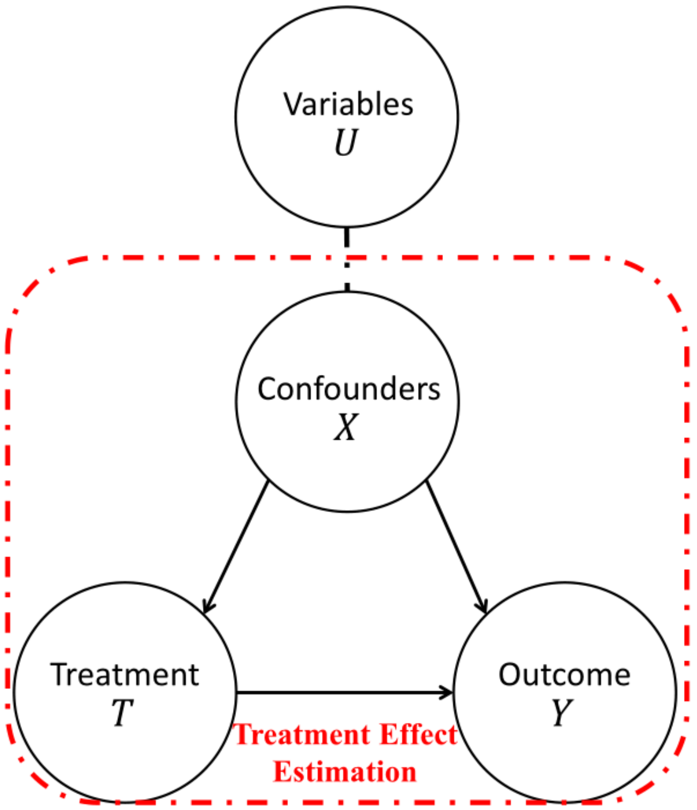


- Two key points:

Balancing Confounders' Distribution

Related Work

- Matching Methods
  - *Exactly Matching, Coarse Matching*
  - Poor performance in high dimensional settings
- Propensity Score based Methods
  - Propensity score $e(\mathbf{X}) = p(T = 1|\mathbf{X})$
  - *Matching, Weighting, Doubly Robust*
  - Treat all observed variables as confounders, and ignore the non-confounders
  - Mainly designed for binary treatment



(a) Previous Causal Framework.

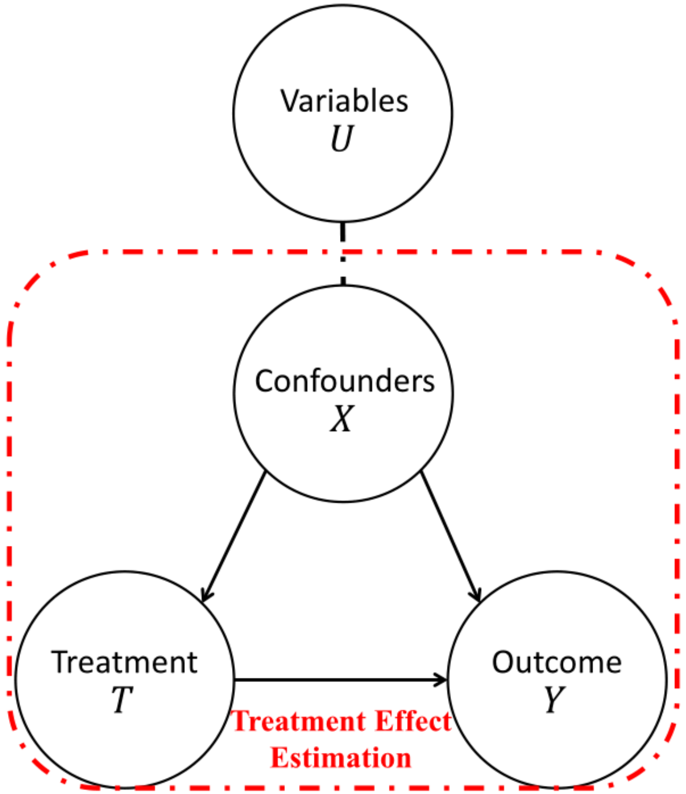
New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition (D^2VD)
- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV)
- Continuous treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD)

New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition (D^2VD)
- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV)
- Continuous treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD)

Previous Causal Framework



(a) Previous Causal Framework.

- Treat all observed variables **U** as confounders **X**
- Propensity Score Estimation:

$$e(\mathbf{U}) = p(T = 1|\mathbf{U}) = p(T = 1|\mathbf{X}) = e(\mathbf{X})$$

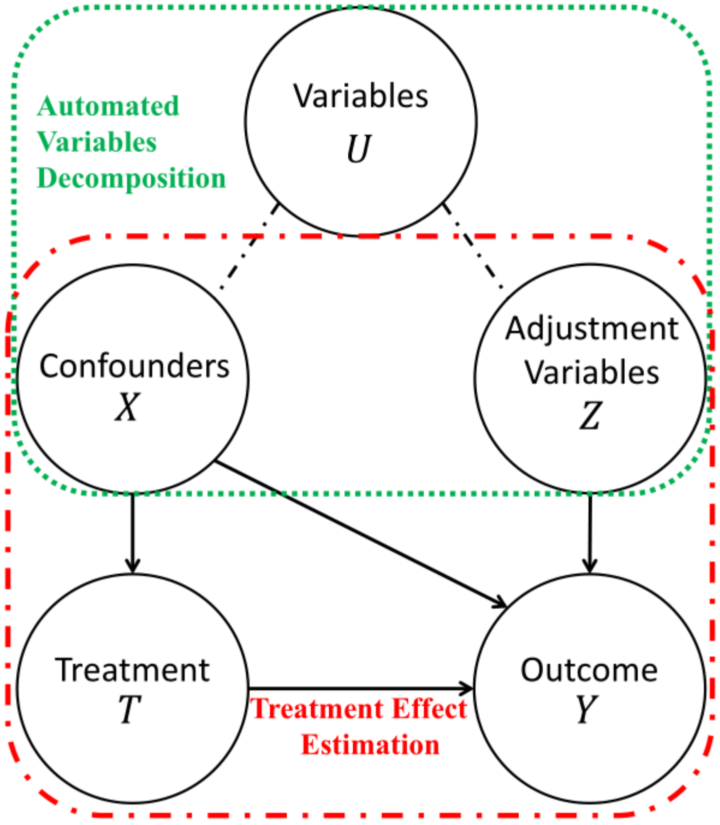
- Adjusted Outcome:

$$Y^{\star} = Y^{obs} \cdot \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} = Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

- IPW ATE Estimator:

$$\widehat{ATE}_{IPW} = \hat{E}(Y^{\star})$$

Our Causal Framework



(b) Our Causal Framework.

- Separateness Assumption:
  - All observed variables U can be decomposed into two sets: Confounders **X**, and Adjustment Variables **Z**
- Propensity Score Estimation:

$$e(\mathbf{X}) = p(T = 1|\mathbf{X})$$

- Adjusted Outcome:

$$Y^{+} = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

- Our D²VD ATE Estimator:

$$\widehat{ATE}_{D^2VD} = \hat{E}(Y^{+})$$

Kuang K, Cui P, Li B, et al. Treatment effect estimation with data-driven variable decomposition [C]//AAAI, 2017 (and extended to TKDE 2020)

Data-Driven Variable Decomposition (D²VD)

$$minimize \quad \|Y^+ - h(\mathbf{U})\|^2 \quad \text{where} \quad Y^+ = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

$$\boxed{e(\mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\beta)} \qquad \phi(\mathbf{Z}) = \mathbf{Z}\alpha,}$$

Replace X, Z with U $\quad h(\mathbf{U}) = \mathbf{U}\gamma,$

$$minimize \quad \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2, \quad \text{where} \quad W(\beta) := \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))}$$

$$s.t. \quad \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau,$$

$$\|\alpha\|_1 \leq \lambda, \ \|\beta\|_1 \leq \delta, \ \|\gamma\|_1 \leq \eta, \ \boxed{\|\alpha \odot \beta\|_2^2 = 0.}$$

α, β, γ

- Adjustment variables: $\mathbf{Z} = \{\mathbf{U}_i : \hat{\alpha}_i \neq 0\}$
- Confounders: $\mathbf{X} = \{\mathbf{U}_i : \hat{\beta}_i \neq 0\}$
- Treatment Effect: $\widehat{ATE}_{D^2VD} = E(\mathbf{U}\hat{\gamma})$

Data-Driven Variable Decomposition (D^2VD)

Bias Analysis:

Our D^2VD algorithm is unbiased to estimate causal effect

THEOREM 1. *Under assumptions 1-4, we have*

$$E(Y^+|X, Z) = E(Y(1) - Y(0)|X, Z).$$

Variance Analysis:

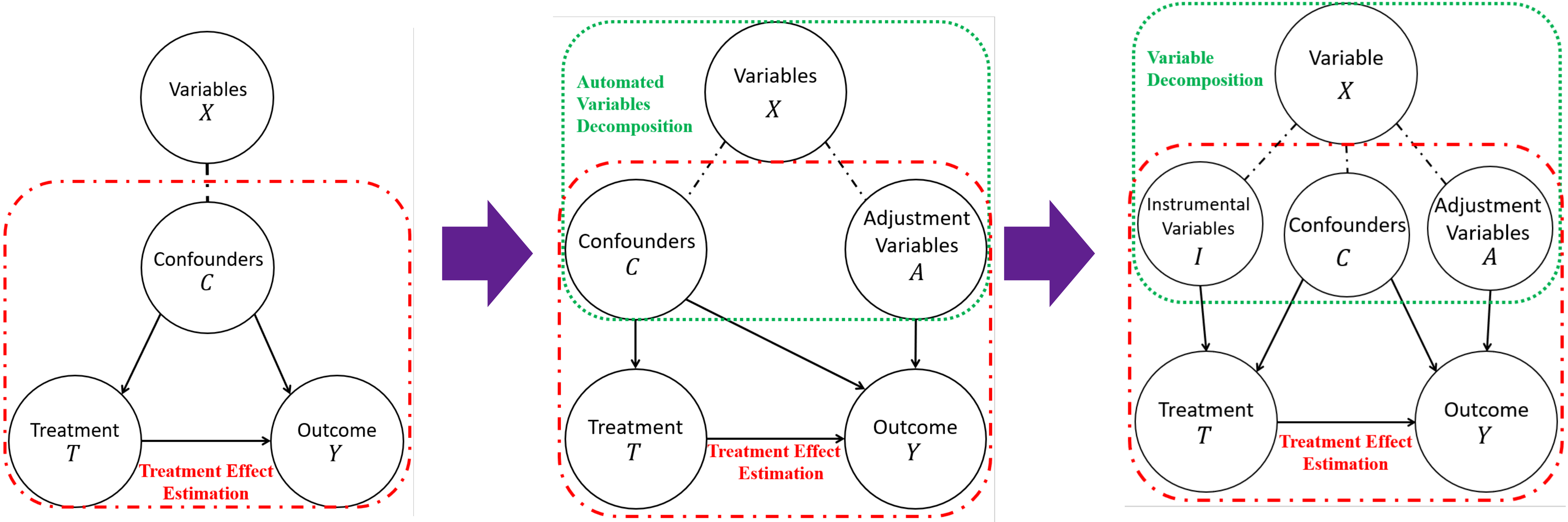
The asymptotic variance of Our D^2VD algorithm is smaller

THEOREM 2. *The asymptotic variance of our adjusted estimator $\widehat{ATE}_{adj}$ is no greater than IPW estimator $\widehat{ATE}_{IPW}$:*

$$\sigma^2_{adj} \leq \sigma^2_{IPW}.$$

Kun Kuang, Peng Cui, Hao Zou, Bo Li, Jianrong Tao, Fei Wu, and Shiqiang Yang. Data-Driven Variable Decomposition for Treatment Effect Estimation, TKDE, 2020

Learning Decomposed Representation for Counterfactual Inference

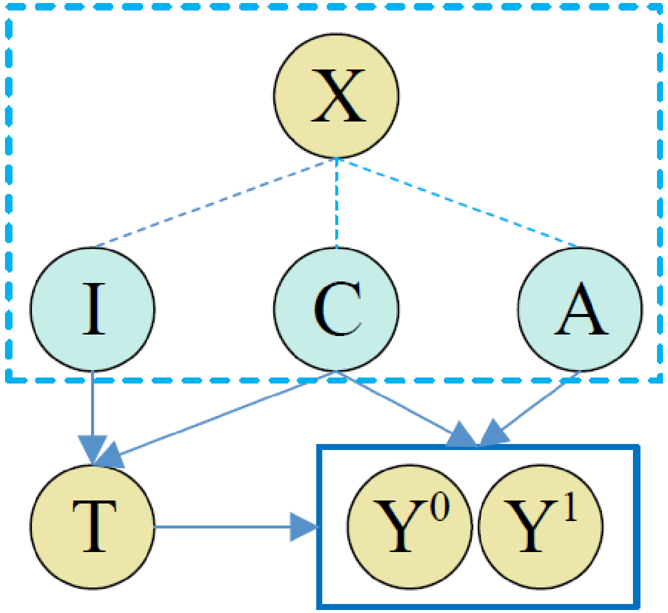


Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Learning Decomposed Representation for Counterfactual Inference

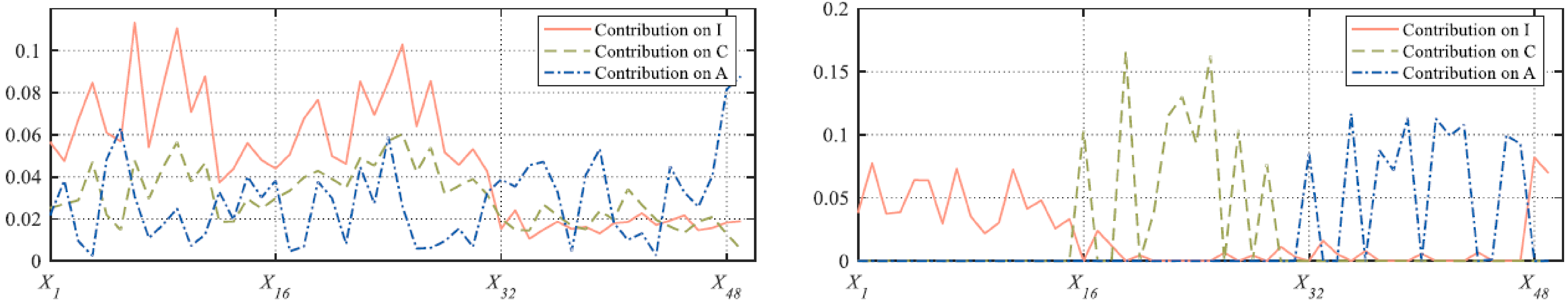
- Three decomposed representation networks
  - $I(X),\ C(X),\ A(X)$
- Three decomposition and balancing regularizers
  - Confounder identification: $A(X) \perp T, I(X) \perp Y \mid T$
  - Confounder balancing: $w \cdot C(X) \perp T$
- Two regression networks
  - $Y(T=1),\ Y(T=0)$
- Orthogonal Regularizer for Decomposition

$$\mathcal{L}_O = \bar{I}_W^T \cdot \bar{C}_W + \bar{C}_W^T \cdot \bar{A}_W + \bar{A}_W^T \cdot \bar{I}_W$$



Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Learning Decomposed Representation for Counterfactual Inference



(a) DR-CFR in Syn_16_16_16_3000

(b) DeR-CFR in Syn_16_16_16_3000

Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Learning Decomposed Representation for Counterfactual Inference

Table 1: The results on IHDP.

IHDP				
Mean +/- Std	**Within-sample**		**Out-of-sample**	
Methods	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
CFR-MMD	0.702 +/- 0.037	0.284 +/- 0.036	0.795 +/- 0.078	0.309 +/- 0.039
CFR-WASS	0.702 +/- 0.034	0.306 +/- 0.040	0.798 +/- 0.088	0.325 +/- 0.045
CFR-ISW	0.598 +/- 0.028	0.210 +/- 0.028	0.715 +/- 0.102	0.218 +/- 0.031
SITE	0.609 +/- 0.061	0.259 +/- 0.091	1.335 +/- 0.698	0.341 +/- 0.116
DR-CFR	0.657 +/- 0.028	0.240 +/- 0.032	0.789 +/- 0.091	0.261 +/- 0.036
DeR-CFR	**0.444 +/- 0.020**	**0.130 +/- 0.020**	**0.529 +/- 0.068**	**0.147 +/- 0.022**

Table 2: Ablation studies of DeR-CFR.

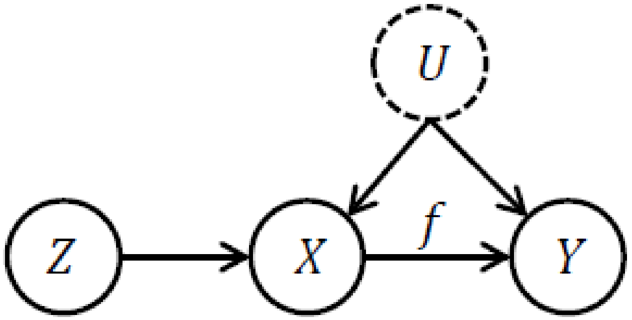
$\mathcal{L}_A$	$\mathcal{L}_I$	$\mathcal{L}_{C\_B}$	$\mathcal{L}_O$	PEHE	
				Within-sample	Out-of-sample
✓	✓	✓	✓	**0.444 +/- 0.020**	**0.529 +/- 0.068**
✓	✓	✓		0.478 +/- 0.033	0.542 +/- 0.053
✓	✓		✓	0.482 +/- 0.039	0.565 +/- 0.075
✓		✓	✓	0.479 +/- 0.030	0.560 +/- 0.071
	✓	✓	✓	0.635 +/- 0.035	0.858 +/- 0.133

Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition (D^2VD)
- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV)
- Continuous treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD)

AutoIV: Counterfactual Learning with Unobserved Confounders via Automatically generating IVs



Conditions of IV (instrumental variable)

- Relevance: $P(X|Z) \neq P(X)$
- Exclusion: $P(Y|Z,X,U) \neq P(Y|X,U)$
- Unconfounded: $Z \perp U$

2SLS:

First Stage: regressing X on Z $\quad \hat{X} = \hat{g}(Z)$

Second Stage: regressing Y on $\hat{X}$ $\quad \hat{Y} = \hat{f}(\hat{X})$

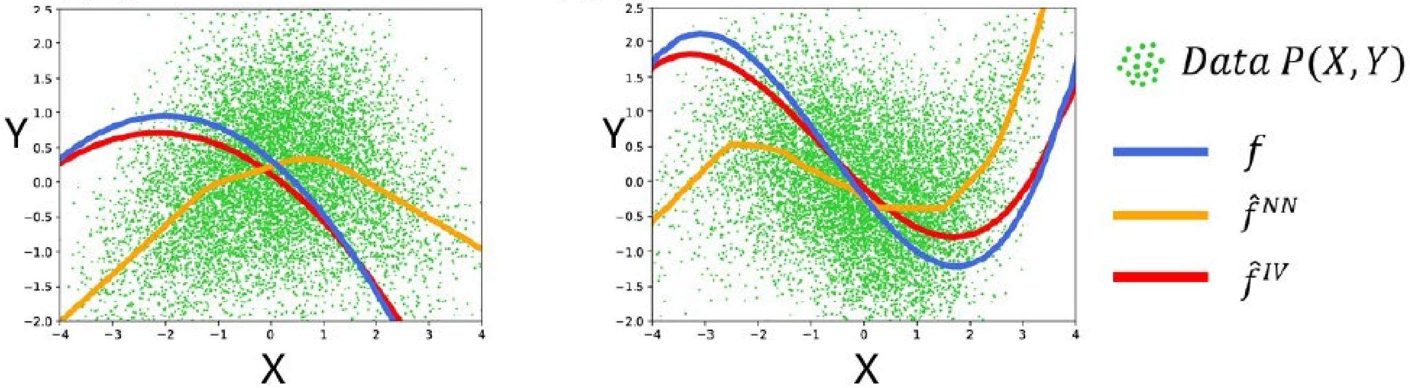
$Z \sim \mathcal{N}(0,1)$

$U \sim \mathcal{N}(0,1)$

$X = Z + U$

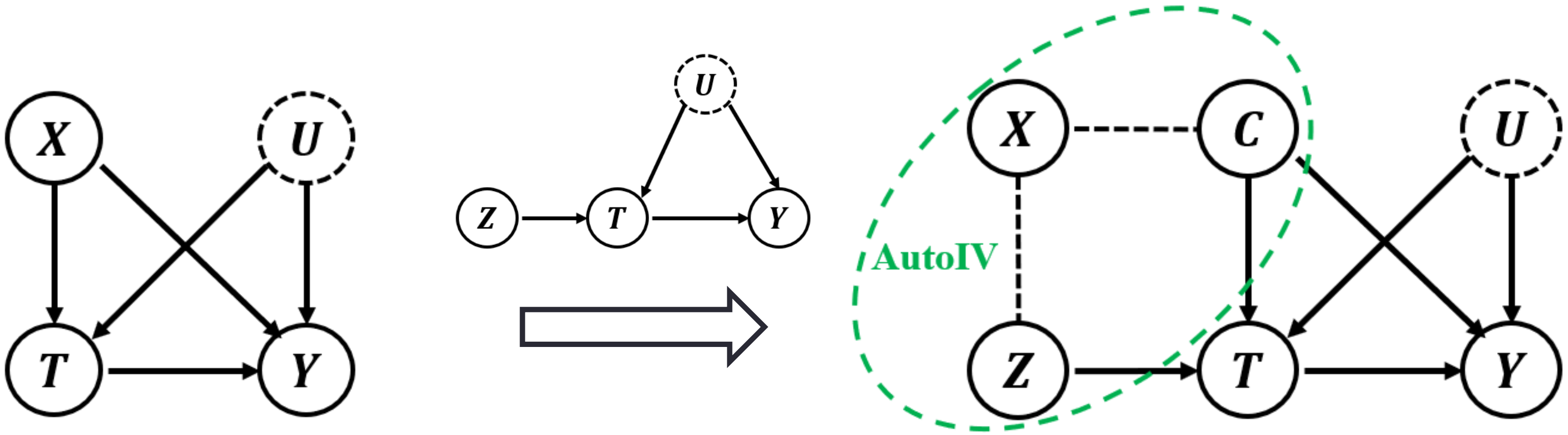
$Y = f(X) + U$

$f(X) = -0.1 \cdot X^2 - 0.4 \cdot X$ $\qquad$ $f(X) = 0.05 \cdot X^3 + 0.1 \cdot X^2 - 0.8 \cdot X$



But these methods require a pre-defined IV and find a valid IV is very hard.

AutoIV: Counterfactual Learning with Unobserved Confounders via Automatically generating IVs



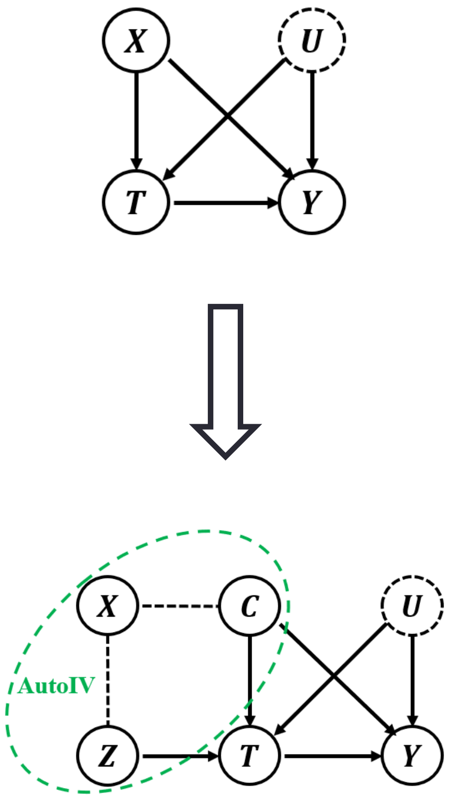
Conditions of IV

- Relevance: $P(T|Z) \neq P(T)$
- Exclusion: $P(Y|Z,T,C) \neq P(Y|T,C)$
- Unconfounded: $Z \perp C$

Mutual Information Representation Learning

Yuan J, Wu A, Kuang K, et al. Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition[J]. arXiv preprint arXiv:2107.05884, 2021.

AutoIV: Counterfactual Learning with Unobserved Confounders via Automatically generating IVs



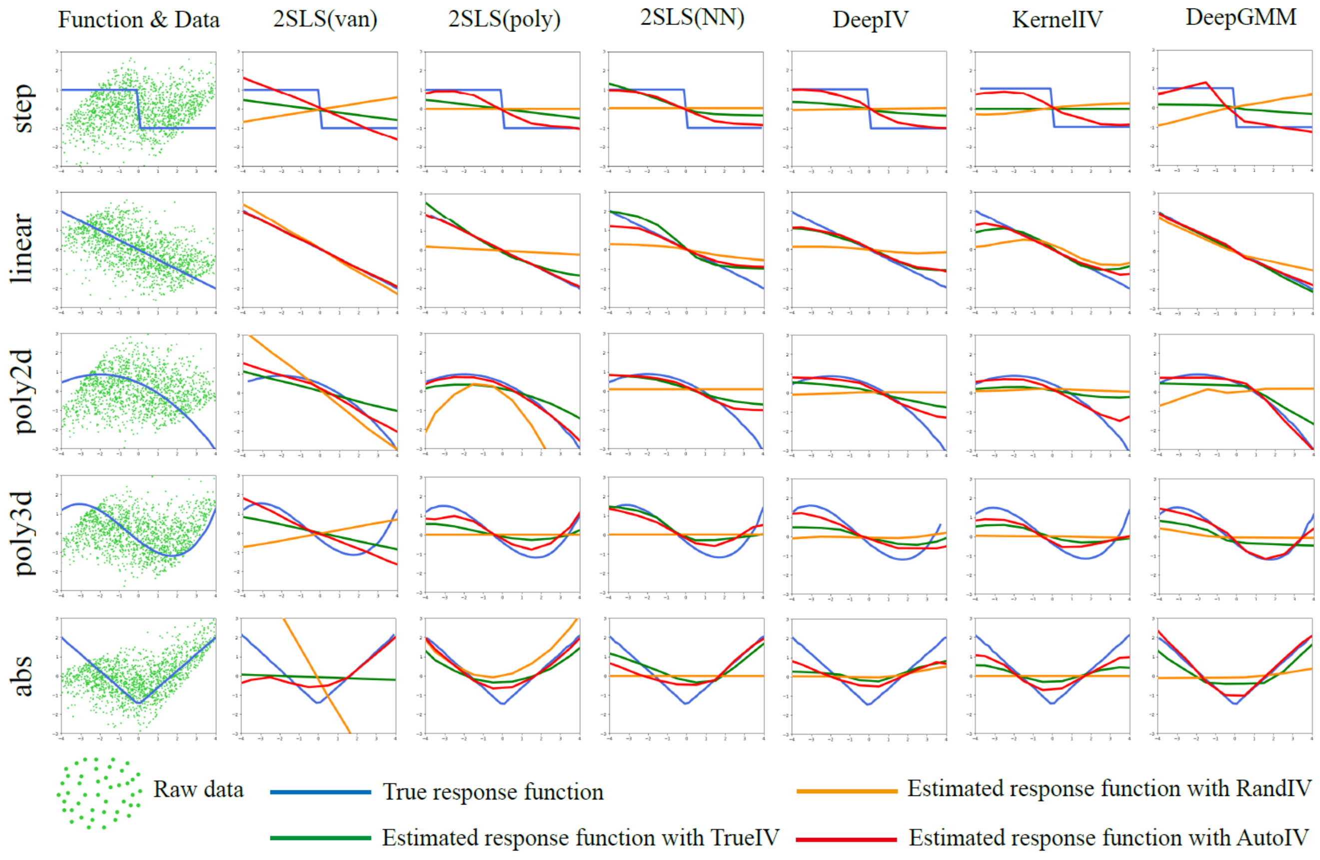


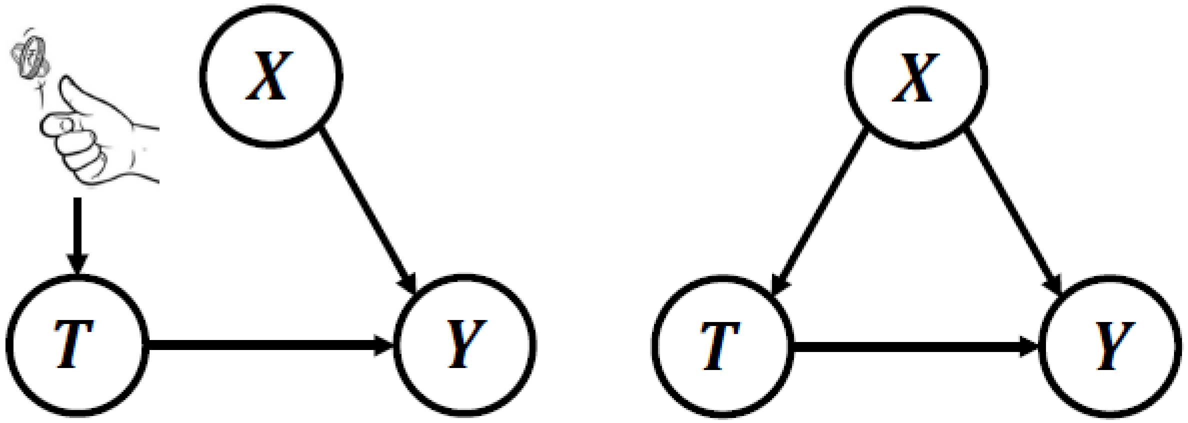
Figure 2: Response function prediction in low-dimensional scenarios.

Yuan J, Wu A, Kuang K, et al. Auto IV: Counterfactual Prediction via Automatic Instrumental Variable Decomposition[J]. arXiv preprint arXiv:2107.05884, 2021.

New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition (D^2VD)
- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV)
- Continuous treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD)

Continuous Treatment Effect Estimation



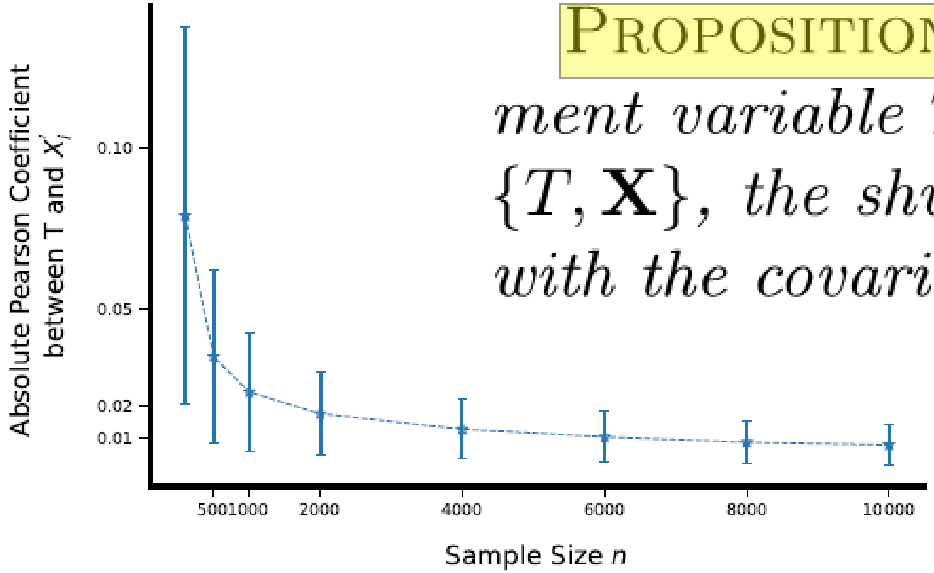
(a) Randomized Controlled Trial (RCT) (b) Observational Studies

- Binary Treatment
  - T=0 or T=1
  - $T \perp X$: confounder balancing
- Multi-valued Treatment
  - T=0,1,2,…
  - $T \perp X$: confounder balancing
- Continuous Treatment
  - How to make $T \perp X$?

Kuang K, Li Y, Li B, et al. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding[J]//DMKD 2021.

Continuous Treatment Effect Estimation

- Our goal: $T \perp X$
- Variable randomly shuffle to achieve independence



PROPOSITION 1. *By randomly shuffle the value of the treatment variable T over all samples in observed data $\mathbf{D}_{obs} = \{T, \mathbf{X}\}$, the shuffled treatment T' would become independent with the covariates $\mathbf{X}$ if sample size $n \rightarrow \infty$.*

Kuang K, Li Y, Li B, et al. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding[J]//DMKD 2021.

Continuous Treatment Effect Estimation

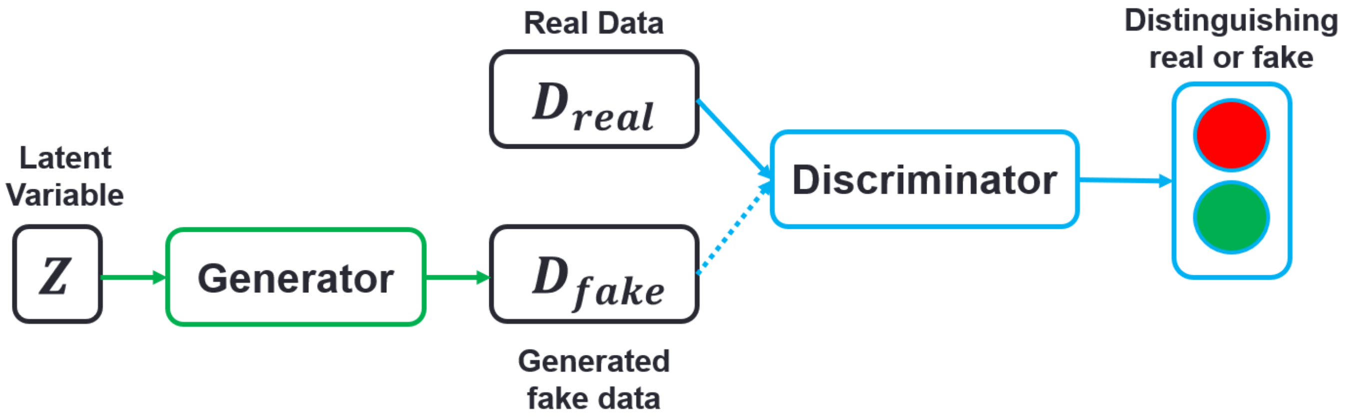
- Our goal: $T \perp X$
- "calibration" distribution generation
  - $\mathbf{D}_{cal} = \{T', \mathbf{X}\}$ on "calibration", we have $T' \perp X$
- "calibration" distribution approximation
  - Observed distribution: $\mathbf{D}_{obs} = \{T, \mathbf{X}\}$
  - Learning sample weights for distribution approximation

$$\mathbf{D}_{obs} = \{T, \mathbf{X}\} \xrightarrow{\text{sample weights } W} \mathbf{D}_{cal} = \{T', \mathbf{X}\}$$

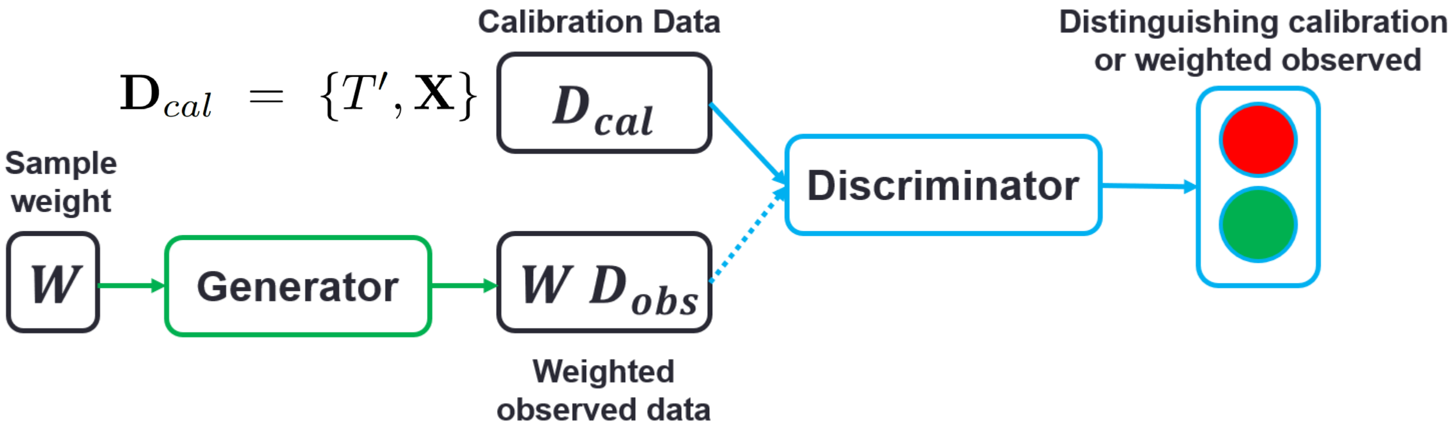
  - Such that: $W\ T \perp W\ X$

Idea from GAN mechanism

- Generative Adversarial Networks (GAN)



- Generative Adversarial De-confounding (GAD)



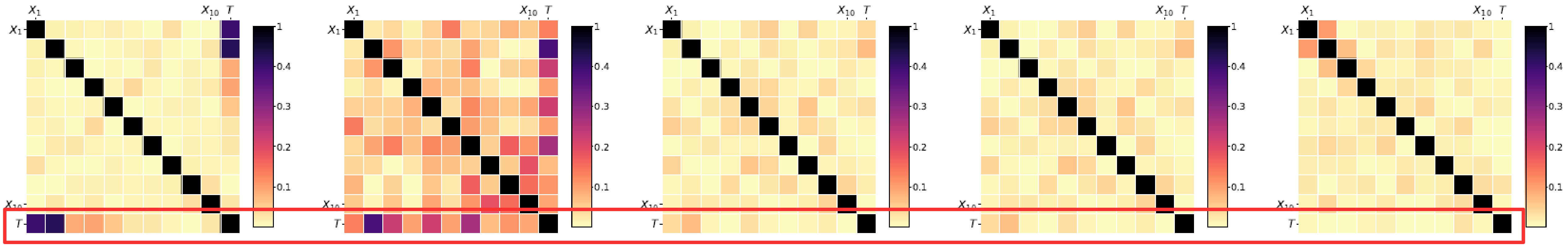
Generative Adversarial De-confounding (**GAD**)

- "Calibration" distribution: $\mathbf{D}_{cal} = \{T', \mathbf{X}\}$
- Observed distribution: $\mathbf{D}_{obs} = \{T, \mathbf{X}\}$
- Sample weights learning with GAD

$$
\begin{aligned}
L(\mathbf{w}, d) = & \mathbb{E}_{(t,x)\sim\mathbf{D}_{cal}}[l(d(t,x),\boxed{1})] \\
& + \mathbb{E}_{(t,x)\sim\mathbf{D}_{obs}}[\boxed{w_{(t,x)}} \cdot l(d(t,x),\boxed{0})], \\
s.t. \quad & \mathbb{E}_{(t,x)\sim\mathbf{D}_{obs}}[w_{(t,x)}] = 1, \mathbf{w} \succeq 0,
\end{aligned}
$$

Kuang K, Li Y, Li B, et al. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding[J]//DMKD 2021.

Continuous Treatment Effect Estimation



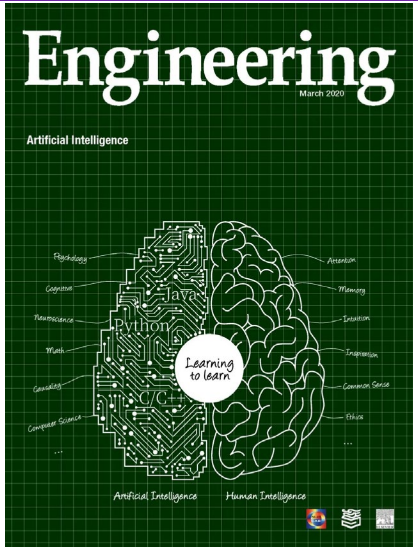
(a) Raw Data (b) $\text{IPW}_{unstable}$ (c) IPW_{stable} (d) CBGPS (e) GAD

Method	*TWINS*		
	BIAS_{MTEF}	RMSE_{MTEF}	RMSE_{ADRF}
OLS	0.208(0.079)	0.236(0.089)	0.686(0.350)
$\text{IPW}_{unstable}$	1.385(0.757)	1.532(0.890)	5.506(2.061)
IPW_{stable}	1.693(1.599)	1.878(1.849)	6.982(4.453)
ISMW	0.165(0.062)	0.181(0.069)	0.962(0.214)
CBGPS	0.187(0.137)	0.216(0.158)	0.683(0.380)
GAD	**0.127(0.039)**	**0.144(0.046)**	**0.383(0.091)**

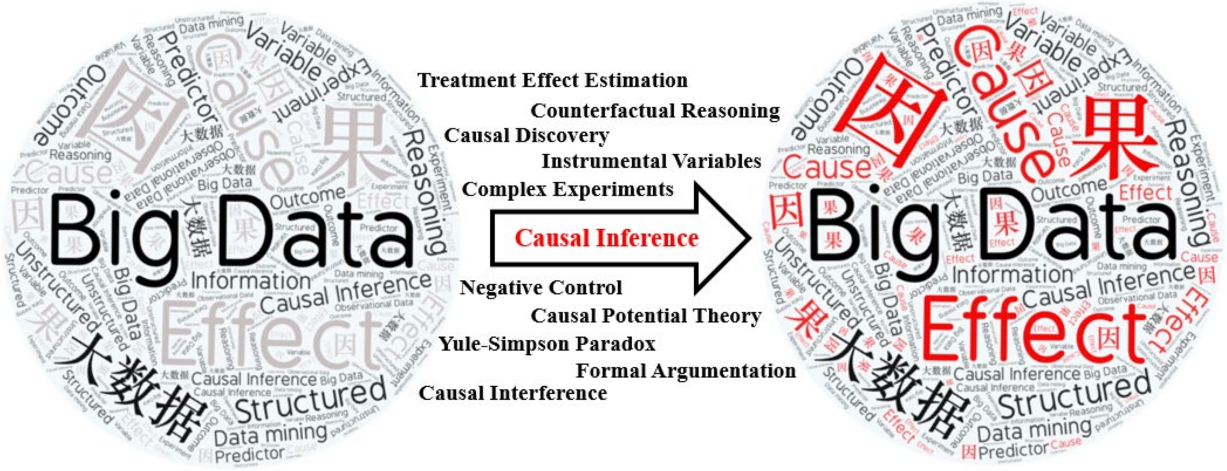
Summary: New challenges in Big Data era

- Automatically separate confounders
  - Not all observed variables are confounders
  - Data-Driven Variables Decomposition (D^2VD)
- Remove unobserved confounding bias
  - Not all confounders are observed
  - Automatic Instrumental Variable Decomposition (AutoIV)
- Continuous treatment effect estimation
  - Treatment variables are not always binary
  - Generative Adversarial De-confounding (GAD)

The official journal of the Chinese Academy of Engineering



Survey Paper：
Causal Inference（因果推理）



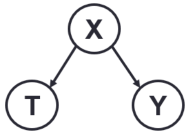
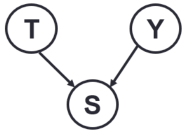
Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z. (2020). Causal Inference. *Engineering*. http://www.engineering.org.cn/ch/10.1016/j.eng.2019.08.016

Content

- Kun Kuang: Estimating average treatment effect: A brief review and beyond
- Lian Li: Attribution problems in counterfactual inference
- Zhi Geng: The Yule–Simpson paradox and the surrogate paradox
- Lei Xu: Causal potential theory
- Kun Zhang: Discovering causal information from observational data
- Beishui Liao and Huaxin Huang: Formal argumentation in causal reasoning and explanation
- Peng Ding: Causal inference with complex experiments
- Wang Miao: Instrumental variables and negative controls for observational studies
- Zhichao Jiang: Causal inference with interference

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z. (2020). Causal Inference. *Engineering*. http://www.engineering.org.cn/ch/10.1016/j.eng.2019.08.016

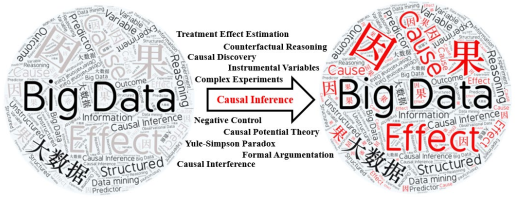
Sources of Correlation

Causation	Confounding	Sample Selection
		
Stable Actionable Explainable	Spurious Correlation: T is correlated with Y ignoring X	Spurious Correlation: T is correlated with Y given S

Causal Inference

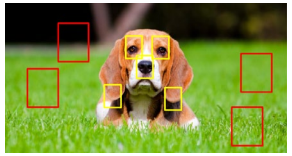
Causality Regularized Machine Learning

Draw Causation from Big Data



Causal Representation/Learning

Stable & Explainable



Fair & Actionable

Thank You!

Kun Kuang
kunkuang@zju.edu.cn
Homepage: https://kunkuang.github.io/