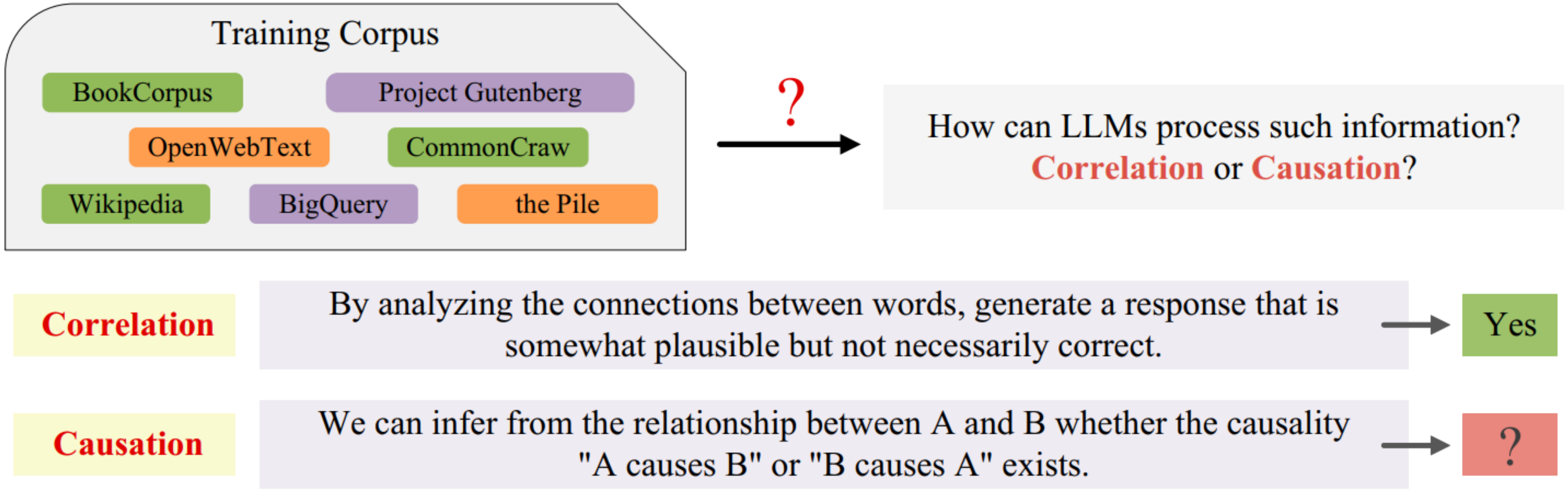
Outlier Robust Causal Discovery

서울대학교 DSML Lab 송의종

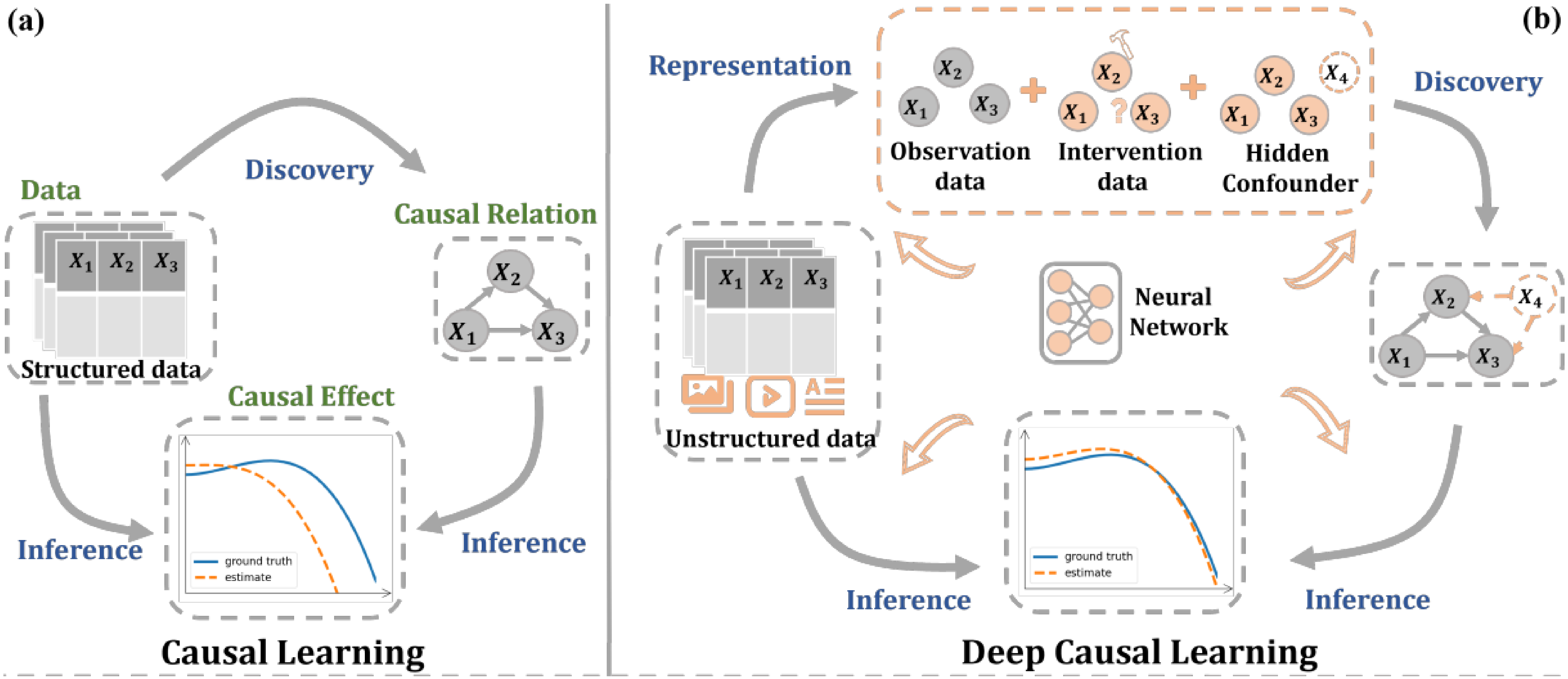
Recent Works of Causal Learning in Age of AI

LLM and Causation



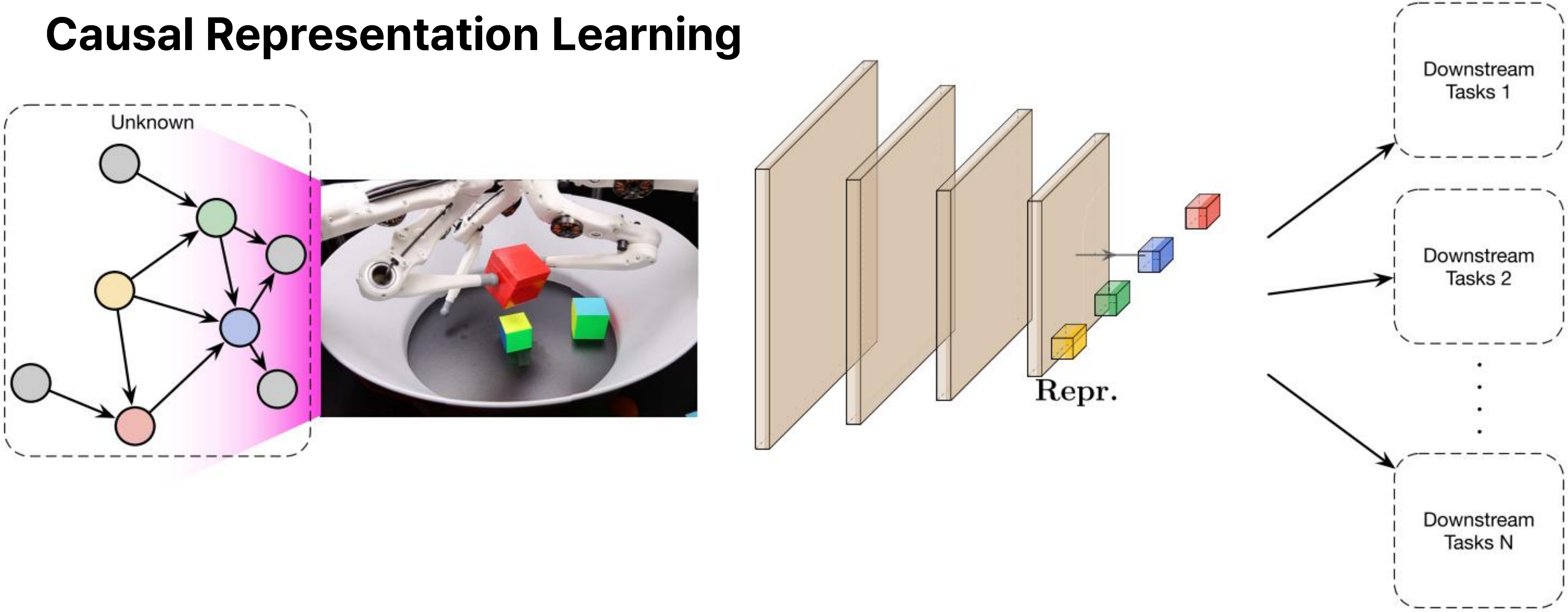
Jiao, Licheng et al. "Causal Inference Meets Deep Learning: A Comprehensive Survey." Research (Washington, D.C.) vol. 7 0467. 10 Sep. 2024, doi:10.34133/research.0467

Deep Causal Learning



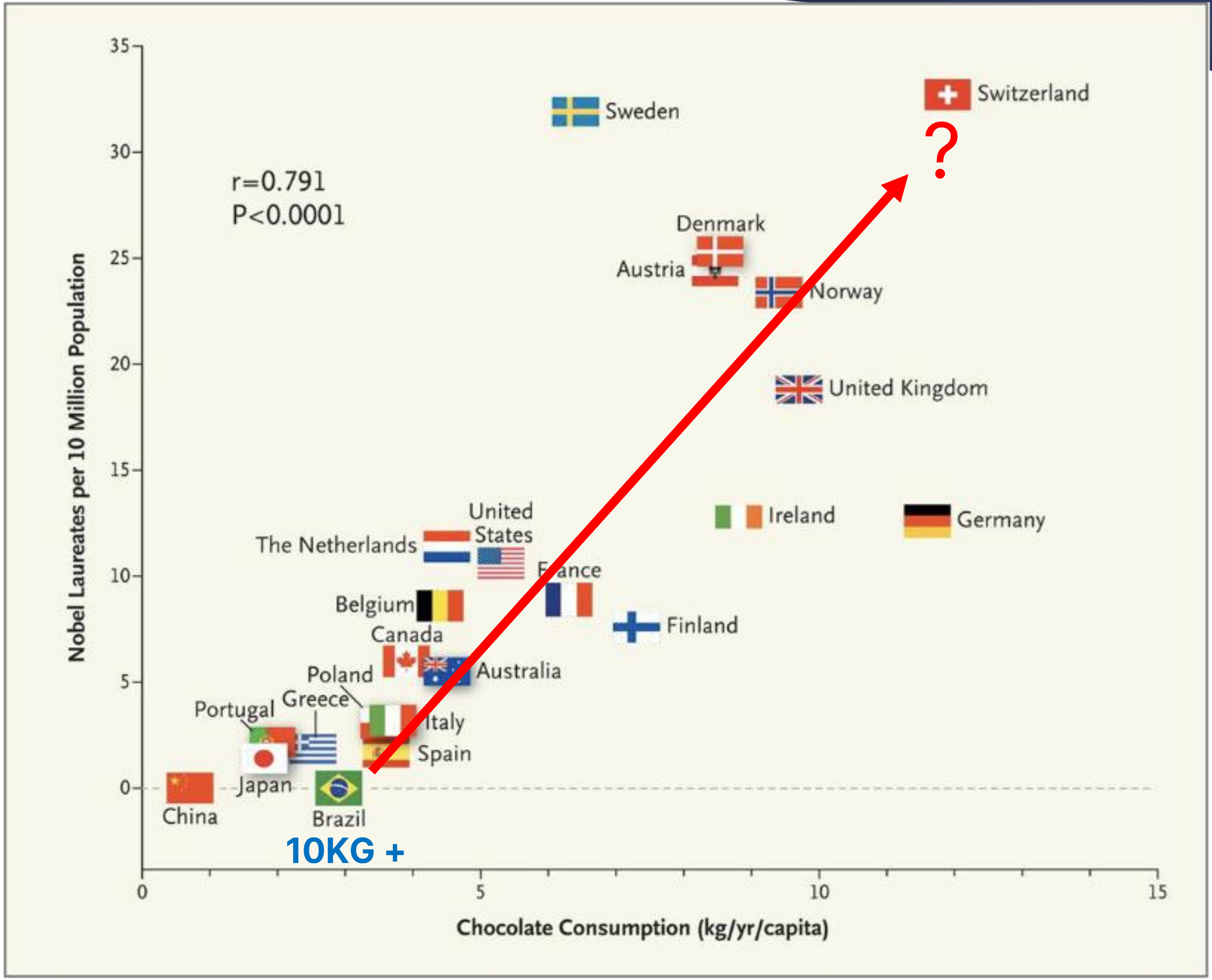
Zizhen Deng, Hu Tian, Xiaolong Zheng, and Daniel Dajun Zeng. 2025. Deep Causal Learning: Representation, Discovery and Inference. ACM Comput. Surv. 58, 2, Article 28 (January 2026),

Causal Representation Learning

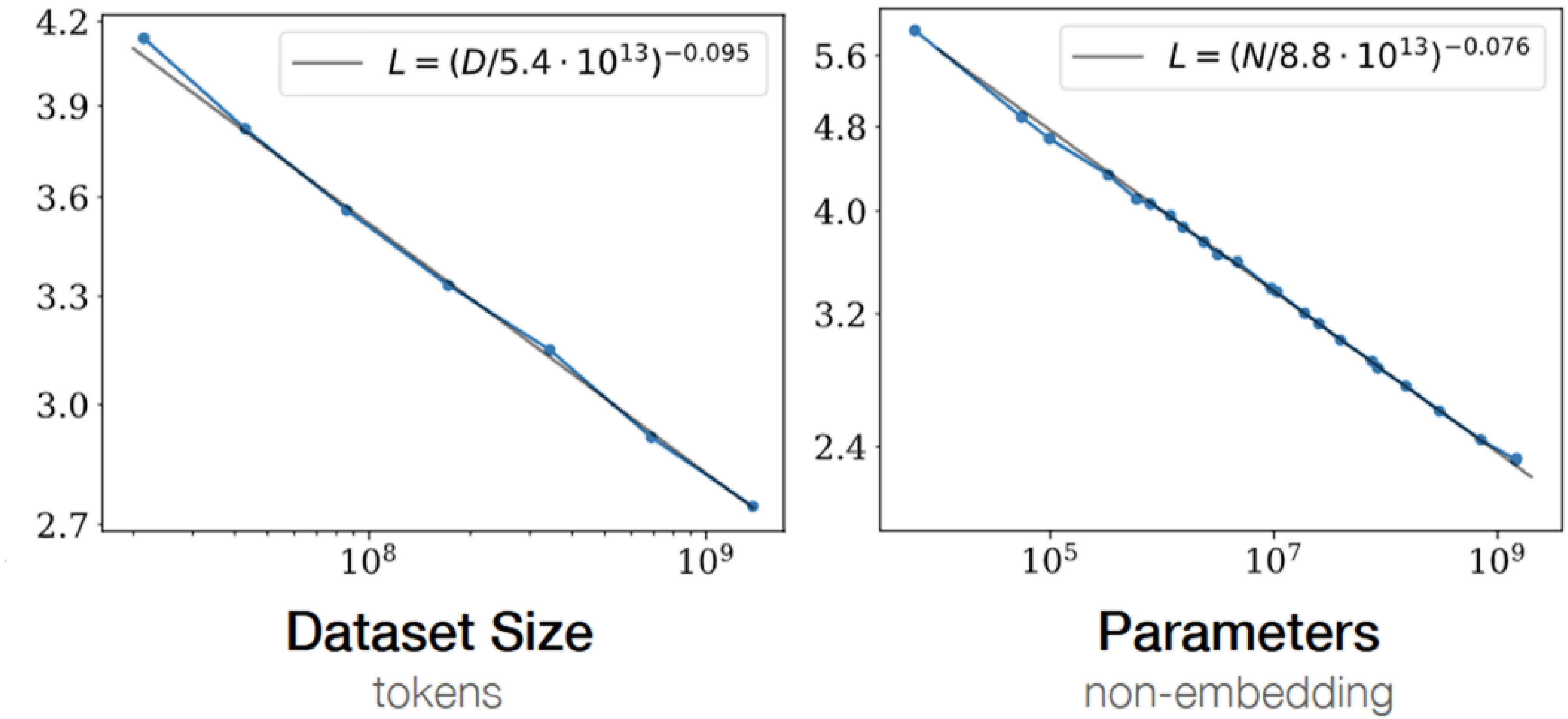


Schölkopf, Bernhard, et al. "Toward causal representation learning." Proceedings of the IEEE 109.5 (2021): 612-634.

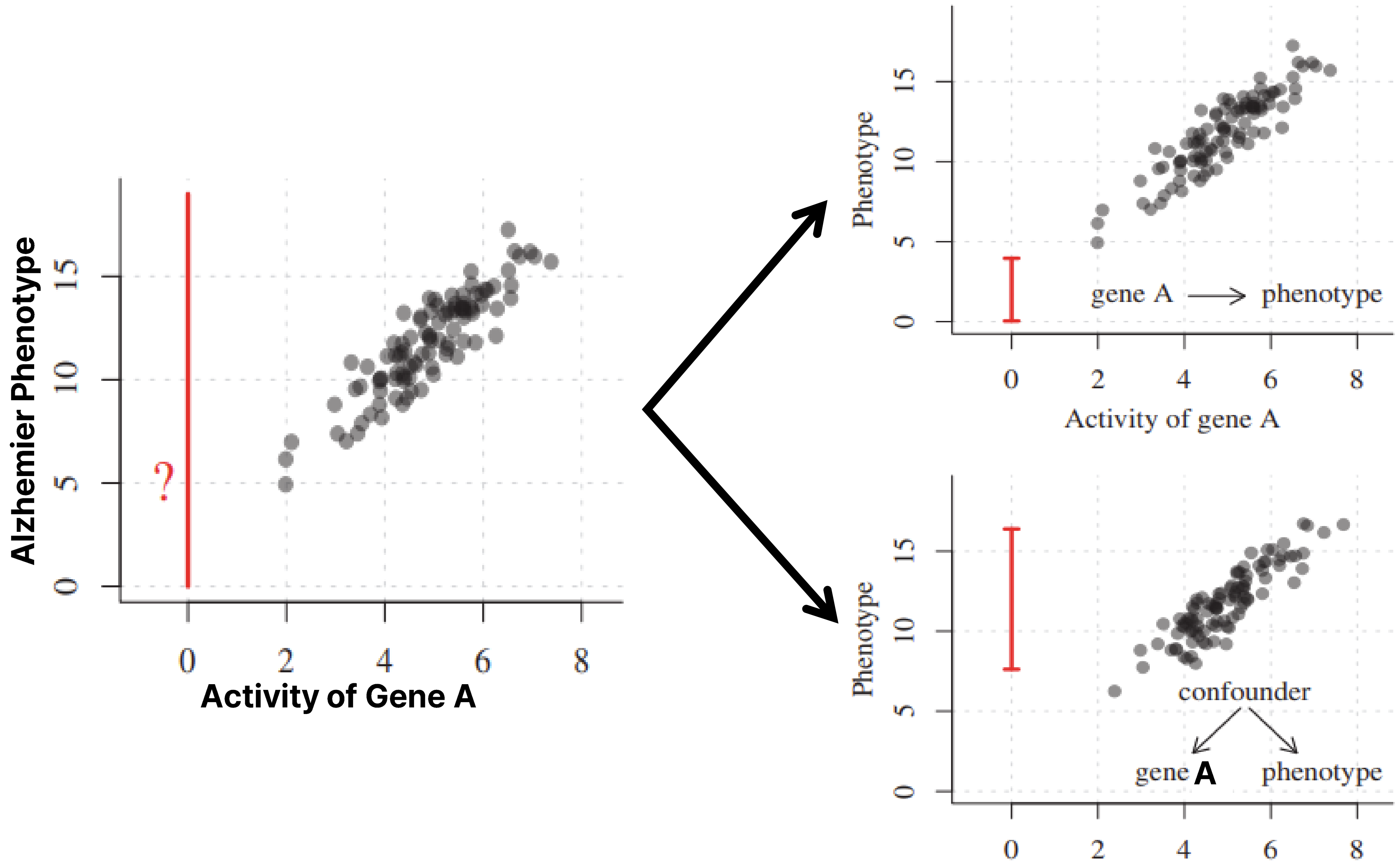
Motivation of Causal Learning



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012



Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).



Alzhemier Phenotype
?
Activity of Gene A
Phenotype
gene A → phenotype
Activity of gene A
Phenotype
confounder
gene A
phenotype

Contents

1. Introduction to Causal Inference

1.1. Causal Reasoning & Learning

[Statistical Learning & Reasoning]

statistical learning

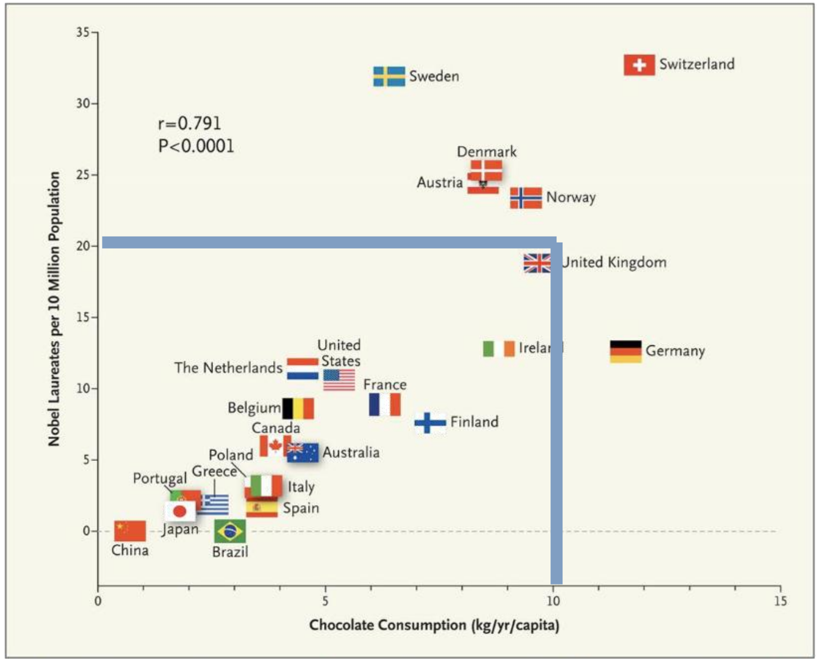
Statistical Model ⇄ **Observations**

probabilistic reasoning

노벨상 = 2초콜릿 + ϵ

초콜릿 섭취량=10

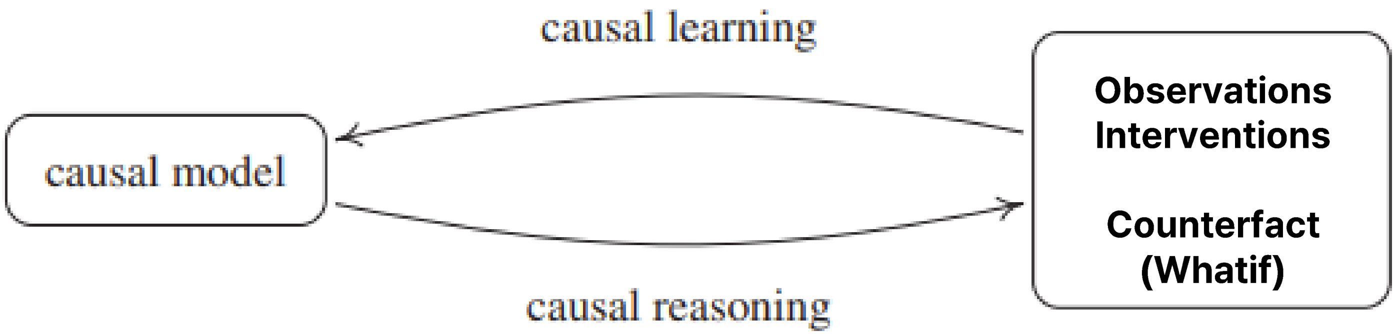
노벨상 = 20



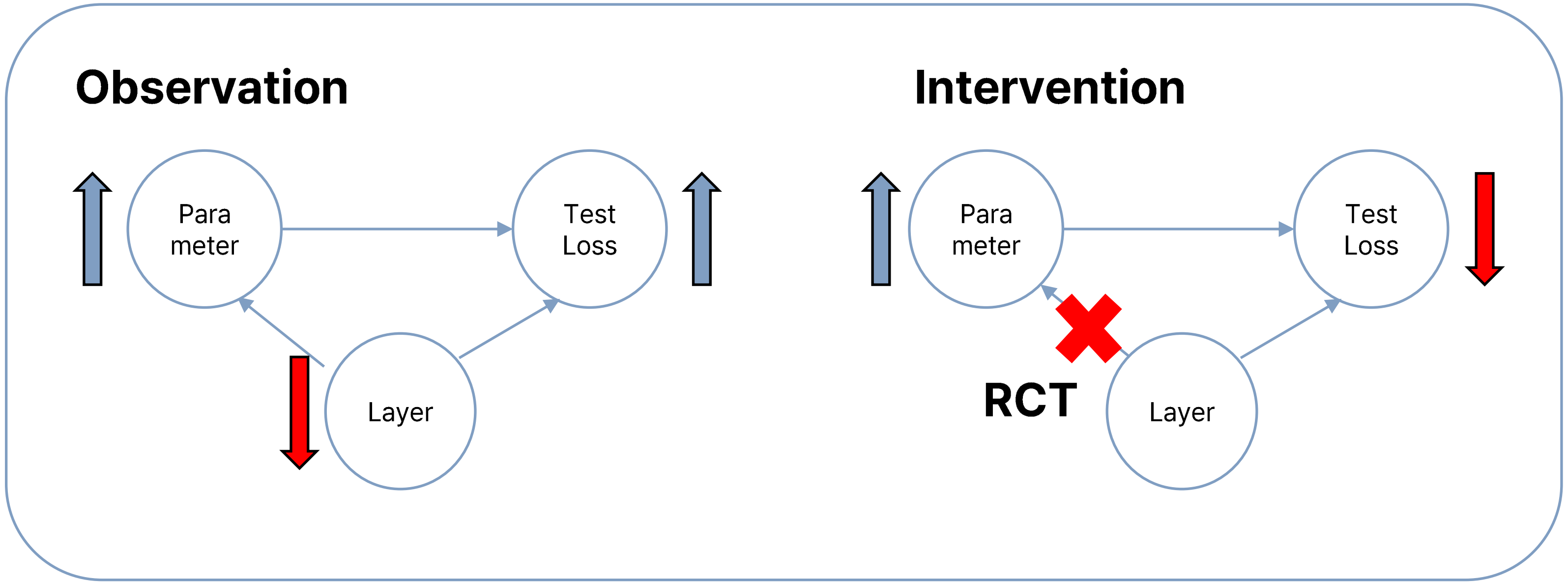
F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

1.1. Causal Reasoning & Learning

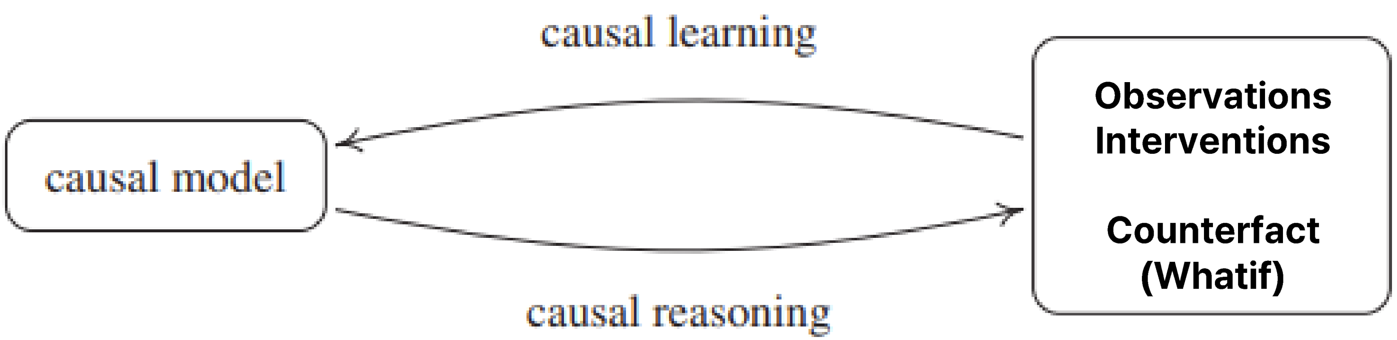
[Causal Learning & Reasoning]



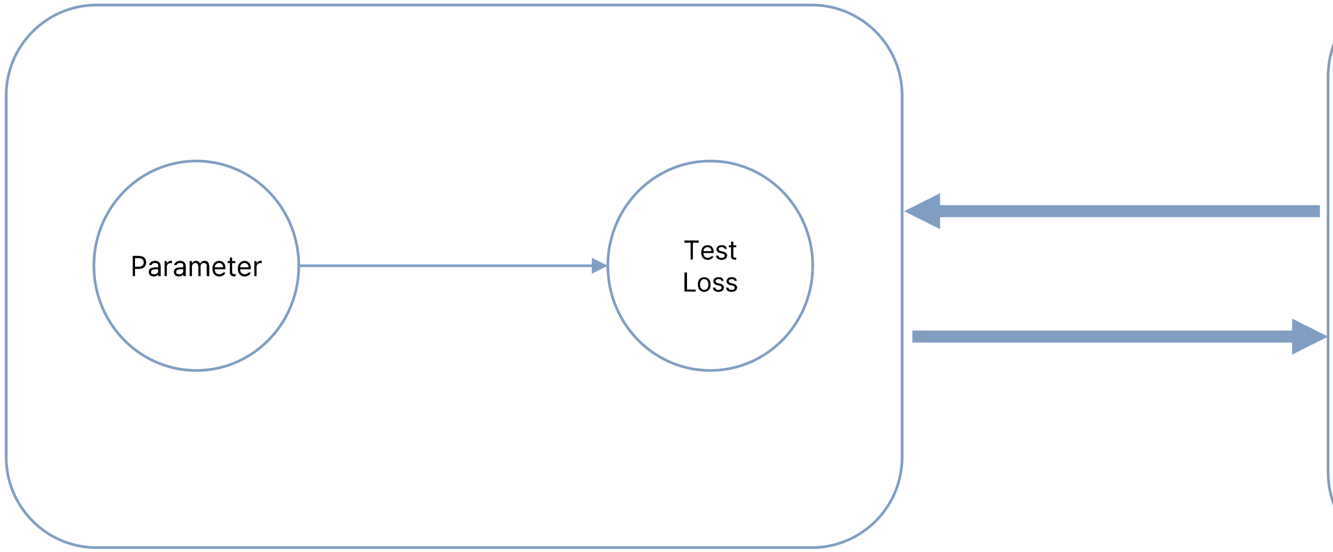
Intervention(개입) Data

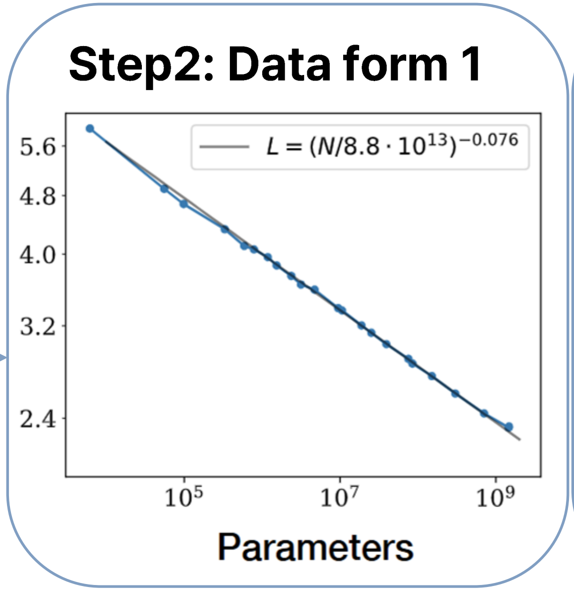


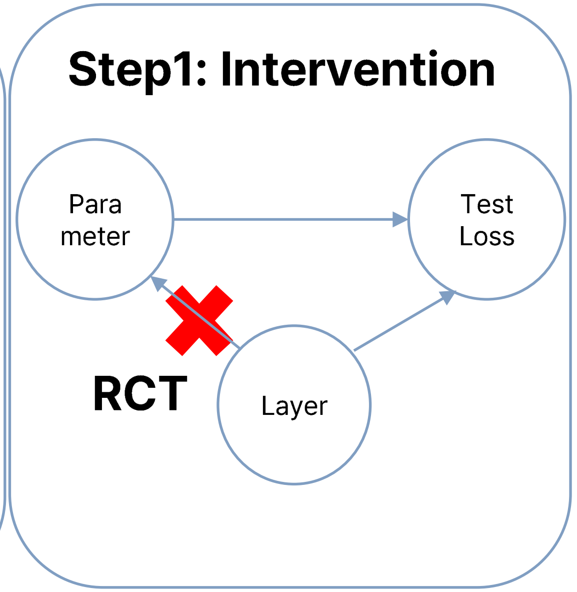
[Causal Learning & Reasoning]



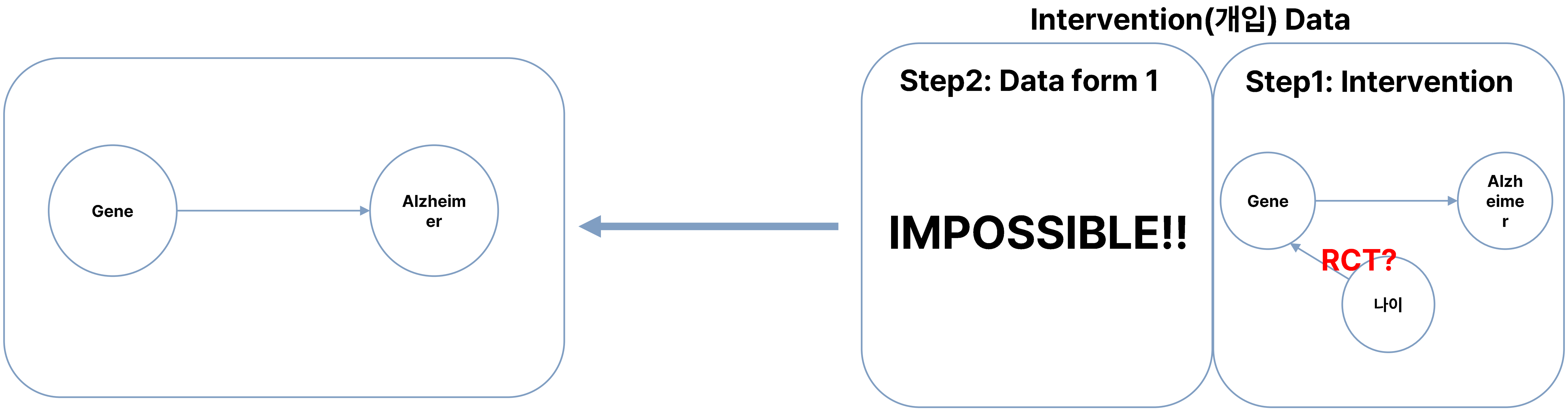
Intervention(개입) Data



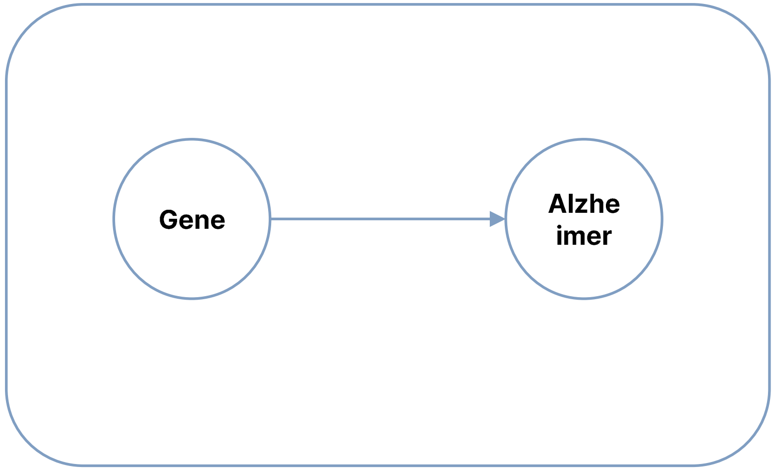




[Causal Learning & Reasoning]

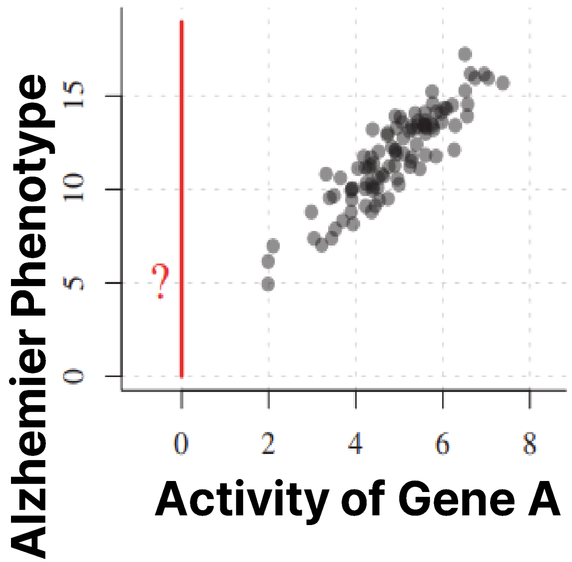


[Causal Learning & Reasoning]



Impossible?

Observational(관측) Data



[Causal Learning from Observation]

- $X = \varepsilon_X,\ Y = aX + \varepsilon_Y,\ \varepsilon_X \perp \varepsilon_Y,$
- $\mathrm{Var}(\varepsilon_X) = \mathrm{Var}(\varepsilon_Y) = \sigma^2$

$\Downarrow$

- $\mathrm{Var}(X) = \sigma^2\ \leq\ \mathrm{Var}(Y) = \sigma^2(1 + a^2)$

- $X = bY + \varepsilon_X,\ Y = \varepsilon_Y,\ \varepsilon_X \perp \varepsilon_Y,$
- $\mathrm{Var}(\varepsilon_X) = \mathrm{Var}(\varepsilon_Y) = \sigma^2$

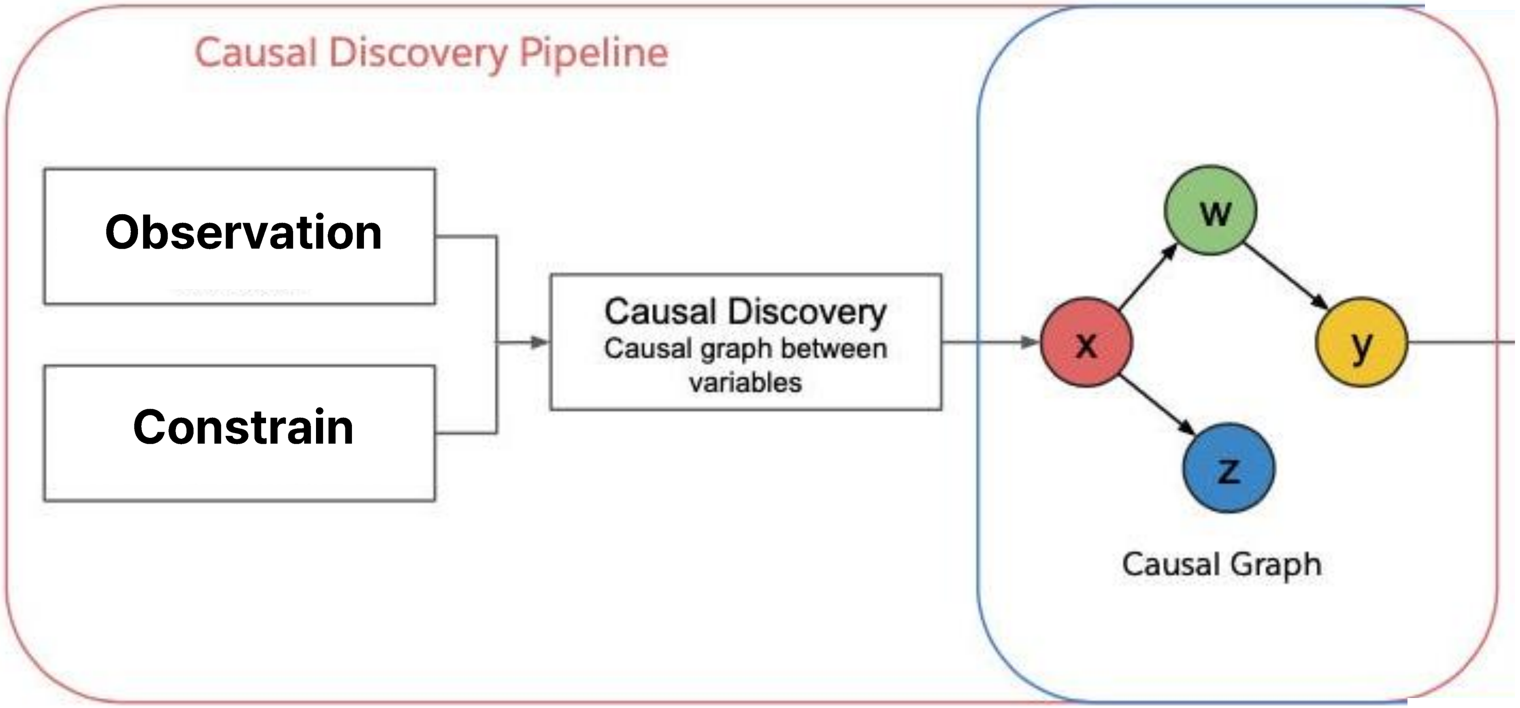
$\Downarrow$

- $\mathrm{Var}(Y) = \sigma^2\ \leq\ \mathrm{Var}(X) = \sigma^2(1 + b^2)$

Now We can Identify Cause and Effect!!

- $\big(\mathrm{Var}(X) < \mathrm{Var}(Y) \Leftrightarrow X \rightarrow Y\big)\ \wedge\ \big(\mathrm{Var}(Y) < \mathrm{Var}(X) \Leftrightarrow Y \rightarrow X\big)$

[Causal Discovery and Inference]

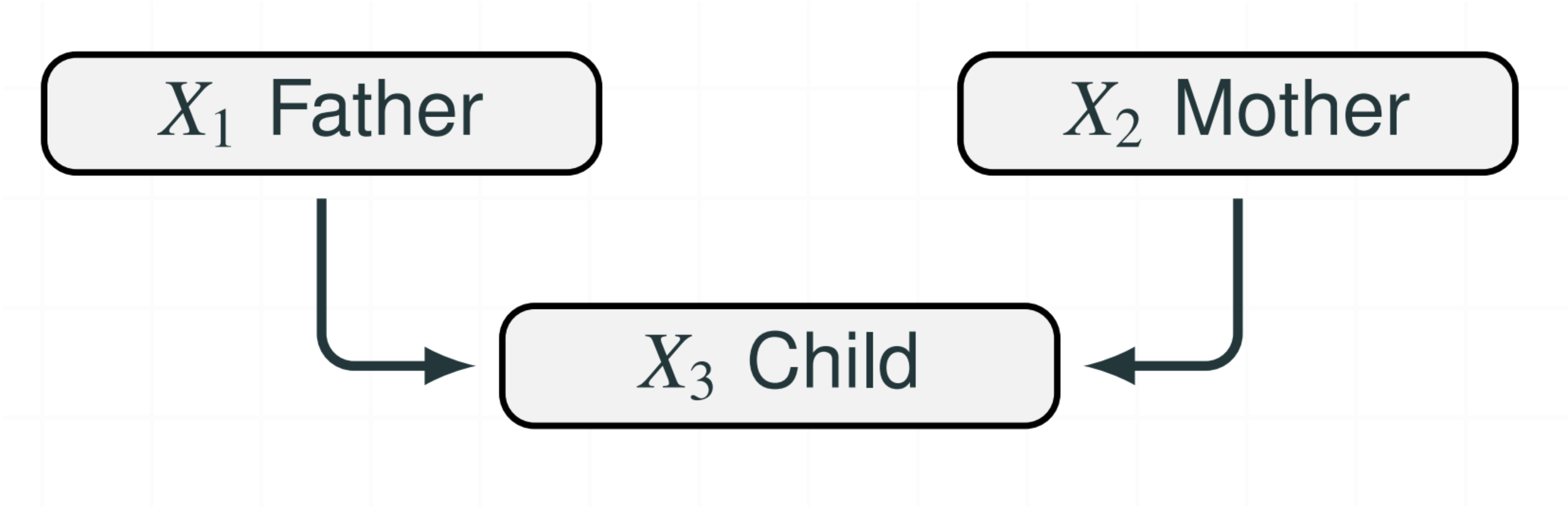


2. Outlier Robust Causal Discovery

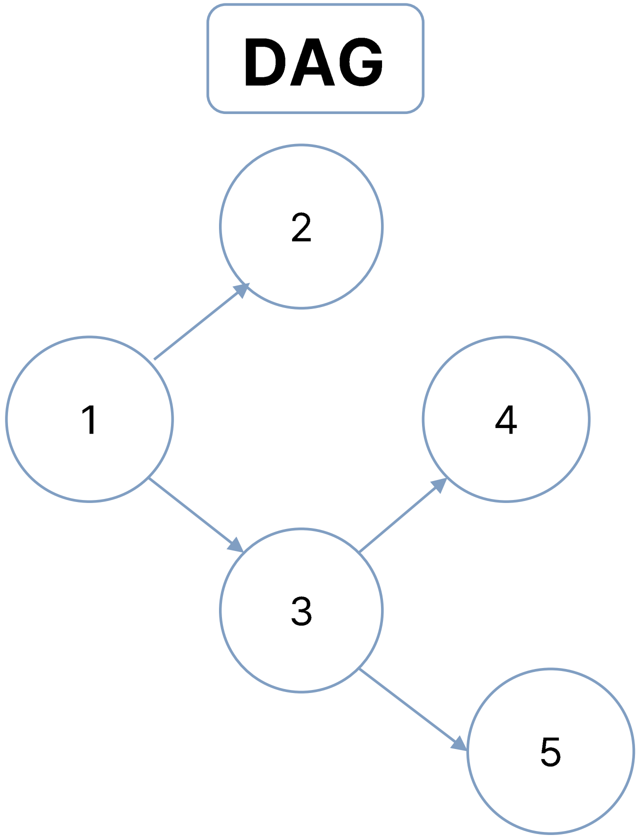
[Table of Contents]

Outlier Robust Algorithm → Learning → Causal Discovery in Outlier Model

[Definitions]



[Definitions]



Graph Structure

- Parent (Pa): $1 \rightarrow \{2, 3\},\ Pa(3) = \{1\}$
- Child (Ch): $Ch(3) = \{4, 5\}$
- Ancestor (An): $An(4) = \{1, 3\}$
- Descendant (De): $De(1) = \{2, 3, 4, 5\}$

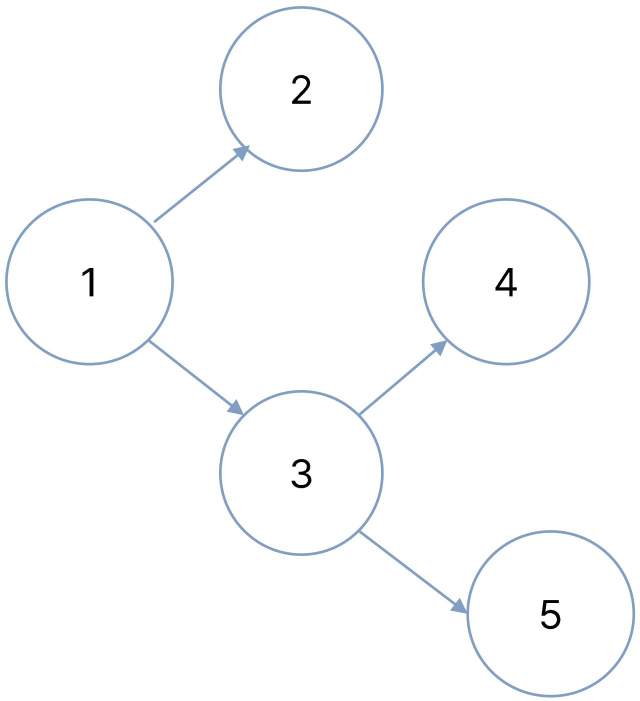
Orderings

- $\pi = (1, 2, 3, 4, 5) \quad \text{or} \quad (1, 3, 2, 5, 4)$

Maximum Indegree

- $d_{\text{in}} = 1$

[Definitions]



Graph Structure

- Parent (Pa): $1 \rightarrow \{2, 3\}$, $Pa(3) = \{1\}$
- Child (Ch): $Ch(3) = \{4, 5\}$
- Ancestor (An): $An(4) = \{1, 3\}$
- Descendant (De): $De(1) = \{2, 3, 4, 5\}$

Orderings

- $\pi = (1, 2, 3, 4, 5)$ or $(1, 3, 2, 5, 4)$

Maximum Indegree

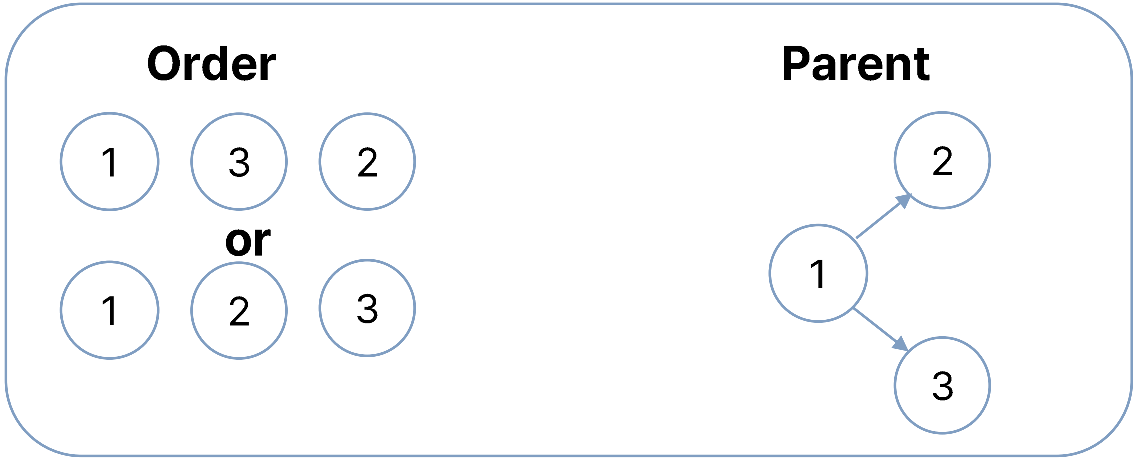
- $d_{\text{in}} = 1$

[Causal Discovery]

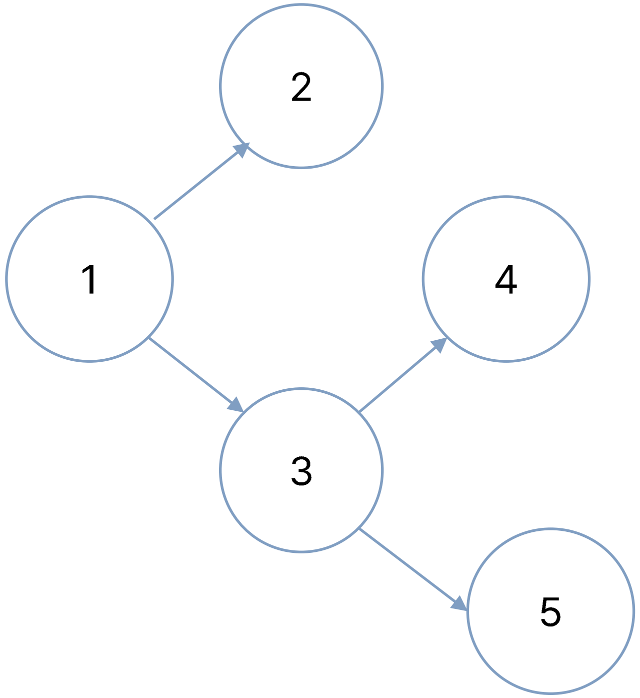
Order-based & Two-Stage (Order → Parent)

Definition. First infer a ordering, then select parent set.
Key characteristics. If the order is correct, search is fast.
Representative methods (mixed classic+recent).

- *K2* (G. F. Cooper & E. Herskovits, 1992)
- *Ordering-Based Search* (M. Teyssier & D. Koller, 2005)
- *BOSS — Best Order Score Search* (B. Andrews, J. Ramsey, R. Sanchez-Romero, J. Camchong, & E. Kummerfeld, 2023)



[Backward Selection]



Backward Selection: Order then Parents

- **Order fixed:** $(1, 2, 3, 4, 5)$

Parent selection (check & pick)

- 5: $\{1, 2, 3, 4\}$ Check $\Rightarrow$ $5 \leftarrow 3$
- 4: $\{1, 2, 3\}$ Check $\Rightarrow$ $4 \leftarrow 3$
- 3: $\{1, 2\}$ Check $\Rightarrow$ $3 \leftarrow 1$
- 2: $\{1\}$ Check $\Rightarrow$ $2 \leftarrow 1$

[Causal Learning]

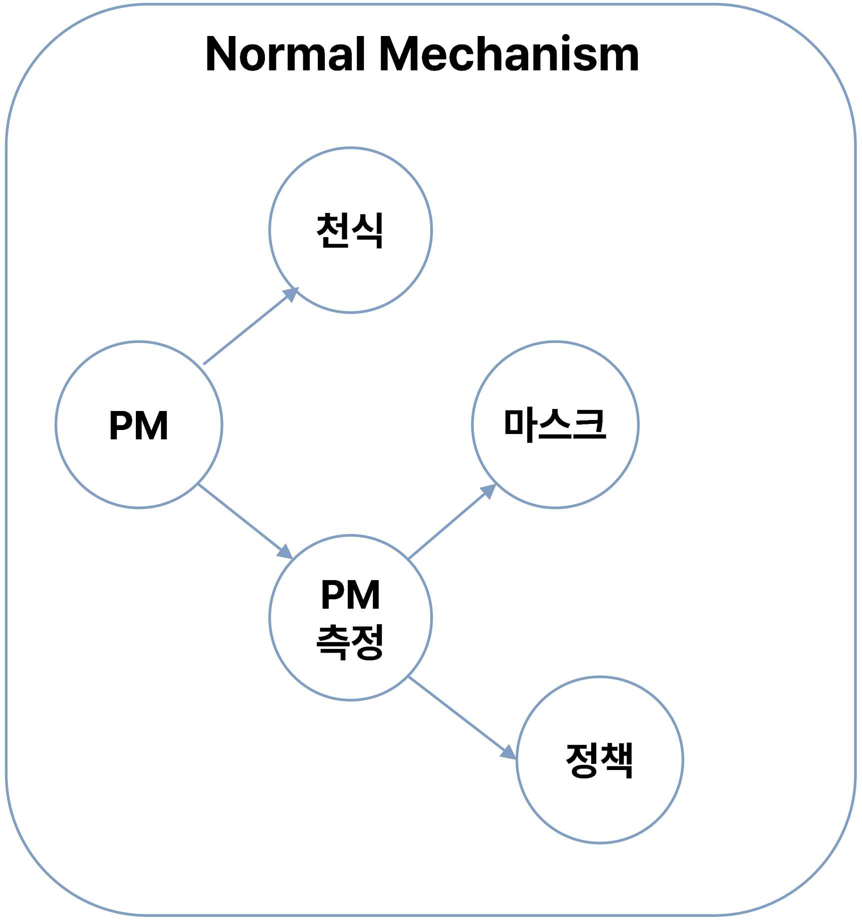
Differentiable DAG Learning

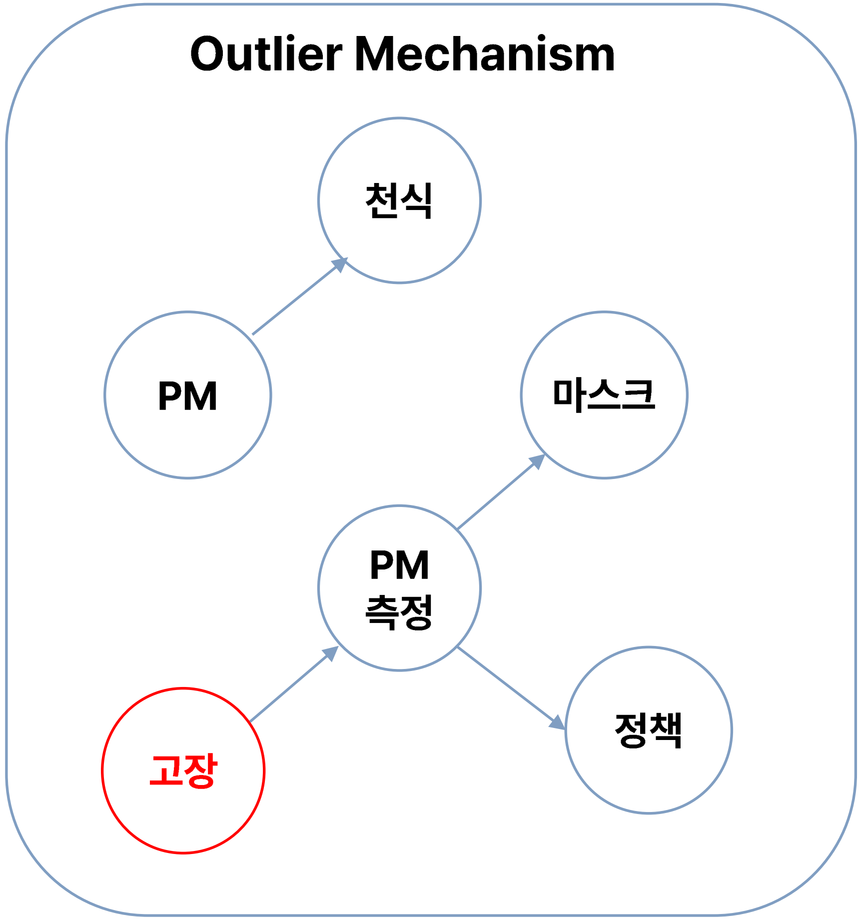
Definition. learning as continuous optimization with a smooth acyclicity constraint.
Key characteristics. Works well with deep learnin models.
Representative methods (mixed classic+recent).

- *NOTEARS* (X. Zheng, B. Aragam, P. K. Ravikumar, & E. P. Xing, 2018)
- *GOLEM* (I. Ng, A. E. Ghassami, & K. Zhang, 2020)
- *SDCD* — Stable Differentiable Causal Discovery (Achille Nazaret, Justin Hong, Elham Azizi, David M. Blei, 2024)

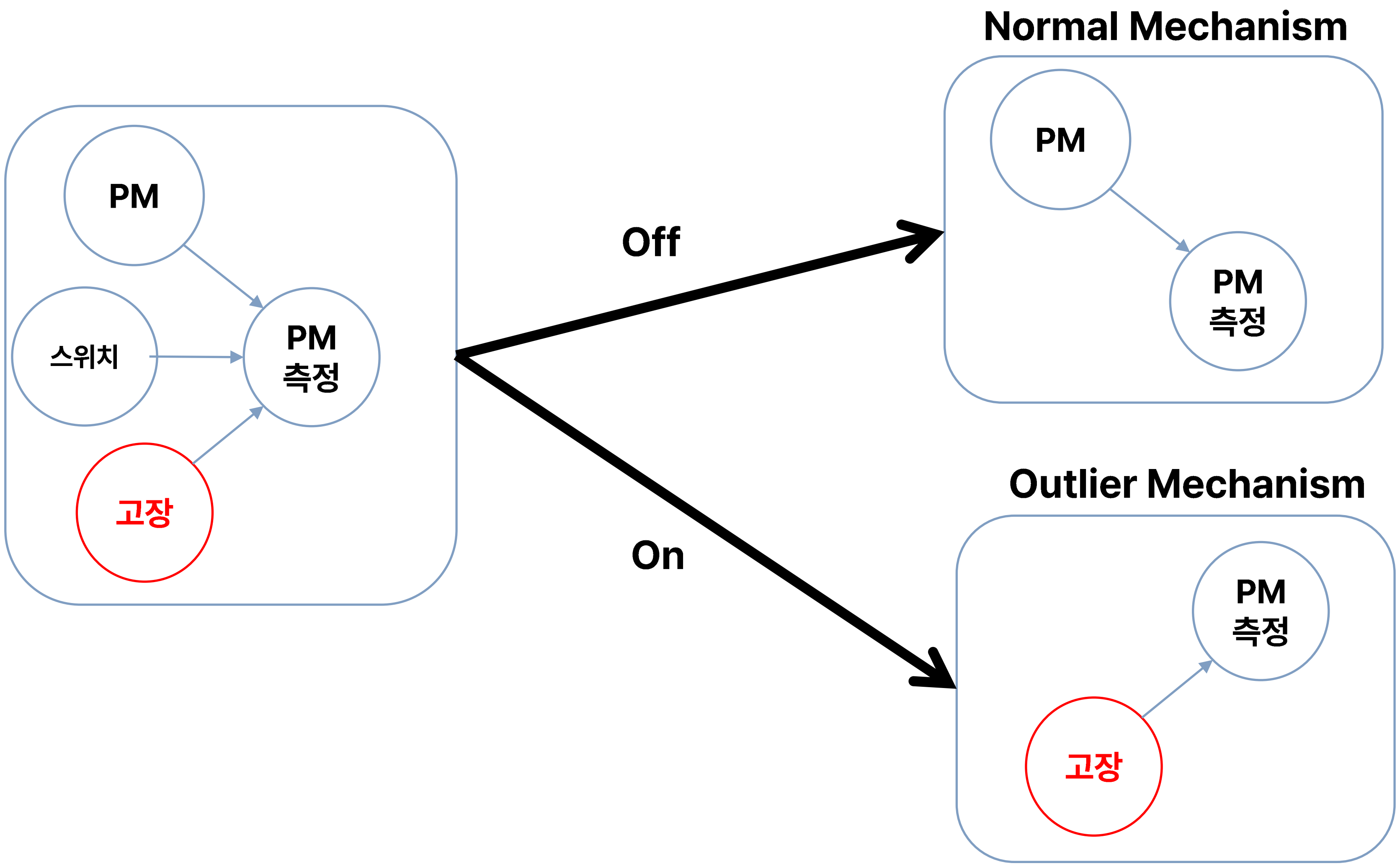
Discrete		**Continuous!!**
$\min_{W \in \mathbb{R}^{d \times d}} \; F(W)$ subject to $\mathsf{G}(W) \in \mathsf{DAGs}$	$\Longleftrightarrow$	$\min_{W \in \mathbb{R}^{d \times d}} \; F(W)$ subject to $h(W) = 0,$

[Outlier]


Normal Mechanism
천식
PM
마스크
PM
측정
정책


Outlier Mechanism
천식
PM
마스크
PM
측정
고장
정책

[Outlier]


PM
스위치
PM
측정
고장
Off
On
Normal Mechanism
PM
PM
측정
Outlier Mechanism
PM
측정
고장

2.1 . Outlier Model

[Cellwise Contaminated SEM]

CCSEM

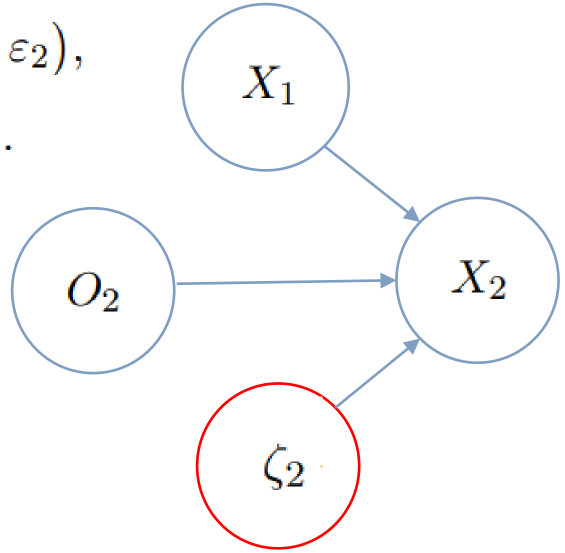
Structural assignments (for $j = 1, \dots, p$**):**

$$X_j = O_j\,\zeta_j + (1 - O_j)\Big(\textstyle\sum_{k \in \mathrm{Pa}_G^X(j)} \beta_{k,j}\, X_k + \varepsilon_j \Big).$$
$$O_j = \mathbf{1}\{ U_j \le q_j\big(O_{\mathrm{Pa}_G^O(j)}\big) \}.$$
$$\zeta_j = g_j\big(\mathrm{Pa}_G(\zeta_j)\big) + \xi_j.$$

Errors: $\varepsilon_j \sim N(0, \sigma_j^2)$ and are independent across j.

$$X_2 = O_2\,\zeta_2 + (1 - O_2)(\beta X_1 + \varepsilon_2),$$

$$O_2 \sim \mathrm{Bernoulli}(0.2), \quad X_1 = \varepsilon_1.$$



2.2 . Outlier Robust Algorithm

[Structure of Algorithm]

Step1: Order Recovery

$$\widehat{\pi}_r \;:=\; \arg \max_{j \in V \setminus \{\widehat{\pi}_{p+1-r}, \dots, \widehat{\pi}_p\}} \left[\, \widehat{\Sigma}^{\widehat{G}_j(r)}_{S'_j(r)} \,\right]^{-1}_{j,j}$$

Step2: Parent Recovery

$$\widehat{\mathrm{Pa}}(j) \;:=\; \mathrm{Supp}\big(\,\widehat{\theta}_j(r)\,\big),$$

- where $\widehat{\Sigma}^{G_j(r)}_{S'_j(r)}$ is the sample covariance of $X^{G_j(r)}_{S'_j(r)}$.
- $X^{\widehat{G}_j(r)}_{S'_j(r)} \in \mathrm{R}^{|\widehat{G}_j(r)| \times |S'_j(r)|}$ (genuine observations).
- $S'_j(r) \;:=\; \{j\} \,\cup\, \mathrm{Supp}(\theta^*_j(r))$.
- $\widehat{G}_j(r) \;:=\;$ estimated good indices via residual thresholding.

- $\widehat{\theta}_j(r)$ is the solution of the ℓ_1-regularized LTS.
- $(\widehat{\theta}_j(r), \widehat{H}_j(r)) := \displaystyle\arg\min_{\theta_j \in \mathbb{R}^{|S_j(r)|},\; H \subset \{1,\dots,n\},\; \boxed{|H| = h}} \frac{1}{2h} \sum_{i \in H} \left(X^{(i)}_j - \langle X^{(i)}_{S_j(r)}, \theta_j \rangle\right)^2 + \lambda_j(r) \|\theta_j\|_1$

2.2 . Outlier Robust Algorithm

[Structure of Algorithm]

Step1: Order Recovery

$$\widehat{\pi}_r \ := \ \arg \max_{j \in V \setminus \{\widehat{\pi}_{p+1-r}, \ldots, \widehat{\pi}_p\}} \widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)})$$

Step2: Parent Recovery

$$\widehat{\mathrm{Pa}}(j) \ := \ \mathrm{Supp}(\,\widehat{\theta}_j(r)\,),$$

- where $\widehat{\Sigma}^{G_j(r)}_{S'_j(r)}$ is the sample covariance of $X^{G_j(r)}_{S'_j(r)}$.
- $X^{\widehat{G}_j(r)}_{S'_j(r)} \in \mathrm{R}^{|\widehat{G}_j(r)| \times |S'_j(r)|}$ (genuine observations).
- $S'_j(r) \ := \ \{j\} \cup \mathrm{Supp}(\theta^*_j(r))$.
- $\widehat{G}_j(r) \ :=$ estimated good indices via residual thresholding.

- $\widehat{\theta}_j(r)$ is the solution of the ℓ_1-regularized LTS.
- $(\widehat{\theta}_j(r), \widehat{H}_j(r)) := \underset{\theta_j \in \mathbb{R}^{|S_j(r)|},\ H \subset \{1,\ldots,n\},\ \boxed{|H|=h}}{\arg\min} \dfrac{1}{2h} \sum_{i \in H} \left(X_j^{(i)} - \langle X_{S_j(r)}^{(i)}, \theta_j \rangle\right)^2 + \lambda_j(r)\|\theta_j\|_1$

2.2 . Outlier Robust Algorithm

[Pseudo Code]

Algorithm 1: Robust Gaussian Linear SEM Learning Algorithm

Input : n independent samples, $X^{1:n}$
Output: Estimated graph structure, $\widehat{G} = (V, \widehat{E})$

Set $\widehat{\pi}_{p+1} = \emptyset$;
for $r = \{1, 2, ..., p-1\}$ **do**
  for $j \in V \setminus \{\widehat{\pi}_{p+1}, ..., \widehat{\pi}_{p+2-r}\}$ **do**
    $S_j(r) = V \setminus (\{j\} \cup \{\widehat{\pi}_{p+1}, ..., \widehat{\pi}_{p+2-r}\})$;
    Estimate $\widehat{\theta}_j(r)$ for ℓ_1-regularized LTS in Equation (2);
    Estimate truncated conditional variances $\widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)})$ using Equation (3);
  end
  Determine $\widehat{\pi}_{p+1-r} = \arg\max_j \widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)})$;
  Determine $\widehat{\mathrm{Pa}}(\widehat{\pi}_{p+1-r}) = \{k \in S_{\widehat{\pi}_{p+1-r}}(r) : [\widehat{\theta}_{\widehat{\pi}_{p+1-r}}(r)]_k \neq 0\}$;
end
Return: Estimate an edge set, $\widehat{E} = \cup_{r \in \{1,2,...,p-1\}} \{(k, \widehat{\pi}_{p+1-r}) : k \in \widehat{\mathrm{Pa}}(\widehat{\pi}_{p+1-r})\}$

2.3 . Consistency of Algorithm

[Order Recovery]

Assumption (Outlier–distance separation)

- For $r \in \{1,\dots,p-1\},\ j \in \{\pi_1,\dots,\pi_{p+1-r}\},\ T_j(r) := \{\pi_1,\dots,\pi_{p+1-r}\} \setminus \{j\}$.
- There exists $\eta_{\min} > 0$ such that

$$\min_{i \in B_j(r,o)} \left| X_j^{(i)} - \mathrm{E}\big[X_j^{(i)} \mid X_{T_j(r)}^{(i)}\big] \right| > \eta_{\min}.$$

- Bad–row set: $B_j(r,o) := \big\{ i \in \{1,\dots,n\} :\ \exists\, k \in \{j\} \cup T_j(r)\ \text{ with }\ o_k^{(i)} = 1 \big\}$.

Assumption (Truncated conditional–variance gap)

- For $r \in \{1,\dots,p-1\},\ j \in \{\pi_1,\dots,\pi_{p+1-r}\},\ T_j(r) := \{\pi_1,\dots,\pi_{p+1-r}\} \setminus \{j\}$.
- For any $\eta \in (0,\eta_{\min}),\ \exists\ \tau_{\min} > 0$ such that for every $\ell \in \mathrm{An}(j)$,

$$\mathrm{Var}\Big(X_j \,\Big|\, X_{T_j(r)},\ \big|X_j - \mathrm{E}(X_j \mid X_{T_j(r)})\big| < \eta\Big) - \mathrm{Var}\Big(X_\ell \,\Big|\, X_{T_j(r)},\ \big|X_\ell - \mathrm{E}(X_\ell \mid X_{T_j(r)})\big| < \eta\Big) > \tau_{\min}.$$

2.3 . Consistency of Algorithm

[Order Recovery]

Theorem (Recovery of the Ordering)

- Gaussian linear CCSEM with cellwise contamination. Assume (Outlier–distance separation) and (Truncated conditional–variance gap) hold.
- Suppose, for each $r \in \{1, \ldots, p-1\}$, $j \in \{\pi_1, \ldots, \pi_{p+1-r}\}$, the supports of $\theta_j^*(r)$ are provided, and choose $\eta \in (0, \eta_{\min})$.
- If the truncated–variance sample size $h' := \min_{j,r} |\widehat{G}_j(r)| \;\geq\; C'_\epsilon \, d^2 \log p$,

$\Rightarrow \;\; \Pr\big(\, \widehat{\pi} \in \Pi^* \,\big) \;\geq\; 1 - \epsilon,\;$ for any $\epsilon > 0$, with some constant $C'_\epsilon > 0$.

[Parent Recovery]

Assumption (Clean–row fraction under CCSEM)

- Fix iteration $r \in \{1, \ldots, p-1\}$ and $j \in \{\pi_1, \ldots, \pi_{p+1-r}\}$.
- Let $S_j(r)$ be the design set and define the bad–row set under config. o by $B_j(r, o) \;:=\; \{\, i \in \{1, \ldots, n\} : \; \exists\, k \in \{j\} \cup S_j(r) \;\text{ with }\; o_k^{(i)} = 1 \,\}$.
- There exists $\alpha_{\min} \in (0, 1]$ such that $|B_j(r, o)| \;\leq\; \min\{(1 - \alpha_{\min})\, h,\; n - h\}$.

2.3 . Consistency of Algorithm

[Parent Recovery]

Assumption (Spectral bound)

- $X^{B_j(r,o)}_{S_j(r)} \in \mathbb{R}^{|B_j(r,o)|\times(p-1)}$ collects rows $B_j(r,o)$ and columns $S_j(r)$.
- There exists $c_{\max} > 0$ such that

$$\max_{j,r} \max_{\substack{K_j \subset V\setminus\{j\} \\ |K_j| = |B_j(r,o)|}} \| X^{B_j(r,o)}_{K_j} \| \le c_{\max} \sqrt{|B_j(r,o)| \log p},$$

where $\|A\| := \sup_{\|v\|_2 \le 1} \|Av\|_2$ (spectral norm).

Assumption (Parameter bound)

- $\forall\, r \in \{1, \ldots, p-1\},\ j \in \{\pi_1, \ldots, \pi_{p+1-r}\},\ \exists\, c_1 > 0$ such that

$$\theta_{\min} < \min_{j,r} \|\theta^*_j(r)\|_{\min} \le \max_{j,r} \|\theta^*_j(r)\|_1 \le \rho \le \frac{c_1}{2}\sqrt{\frac{h}{\log p}},$$

- ρ is the tuning parameter used in the ℓ_1-regularized LTS step.

2.3 . Consistency of Algorithm

[Parent Recovery]

Theorem (Sign Recovery)

- Consider the ℓ_1-regularized LTS. with solution $\widehat{\theta}_j(r)$.
- Assume (Clean–row fraction), (Spectral bound), and (Parameter bound) hold.
- Then for any $\epsilon > 0$, $\exists\, c_\epsilon, \kappa_1 > 0$ such that

$$\lambda_j(r) \;=\; c_\epsilon \sqrt{\frac{\log p}{h}},$$

$$h \;\geq\; \left(\frac{c_\epsilon}{\kappa_1\, \alpha_{\min}\, \theta^2_{\min}}\right)^2 \left(\tfrac{9}{4}\, d \;+\; 4\,|B_j(r,o)|\right) \log p,$$

$$\Rightarrow\; \Pr\Big(\mathrm{sign}\big(\widehat{\theta}_j(r)\big) = \mathrm{sign}\big(\theta^*_j(r)\big)\Big) \;\geq\; 1 - \tfrac{\epsilon}{p^2}.$$

2.3 . Consistency of Algorithm

[Consistency]

Corollary (Consistency of the Algorithm)

- Setting: Gaussian linear CCSEM with cellwise contamination.
- Assume (Clean–row fraction), (Spectral bound), (Parameter bound), and (Truncated conditional–variance gap) hold, with appropriate tuning (λ, η).
- If $h = \Omega\big((d + |B|) \log p\big)$ and $h' = \Omega\big(d^2 \log p\big)$, where $|B| := \max_{j,r} |B_j(r, o)|$, $h' := \min_{j,r} |\widehat{G}_j(r)|$,

$\Rightarrow \Pr\big(\widehat{G} = G\big) \to 1 \quad (n \to \infty)$.

[Simulations]

Simulation Settings (CCSEM)

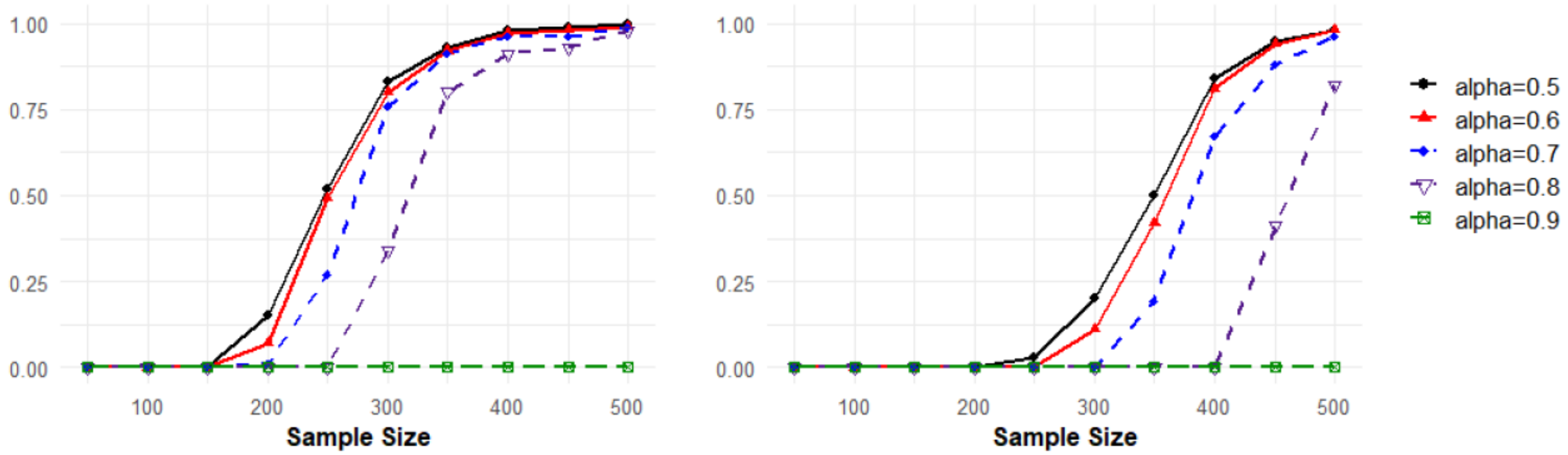
- **Graph size:** $p \in \{5, 10, \ldots, 25\}$; 100 realizations per p.
- **Sparsity:** max degree $d \leq 4$, min indegree $= 1$; no isolated nodes.

- **Edge weights:** $\beta_{k,j}$ i.i.d. uniform in $(-1, -0.75) \cup (0.75, 1)$.
- **Noise:** $\sigma_j^2 = 0.75$ for all j.

- **Outliers:** $B \in \{1, 30, 60, 90\}$,
  values generated from $N(100,\ \sigma_j^2)$.

2.4 . Simulations and Real-Data

[Number of Trimmed Sample]



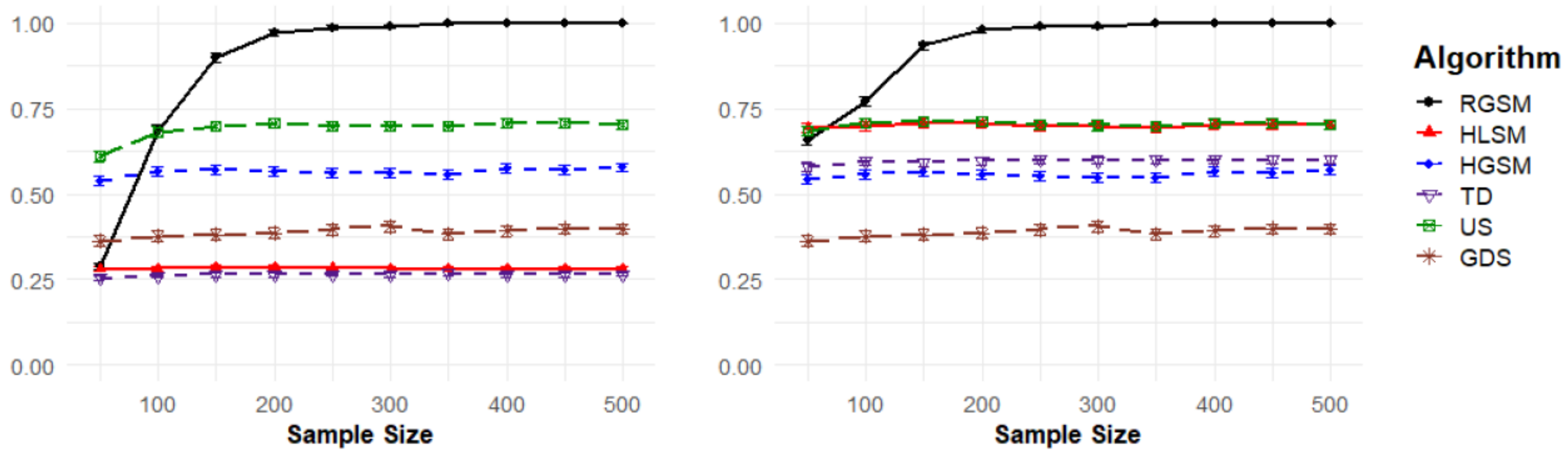
(c) $|B| = 60$

(d) $|B| = 90$

Figure 2: Performance of the proposed algorithm with various fractions of the sample size, $\alpha \in \{0.5, 0.6, ..., 0.9\}$, for learning 10-node corrupted Gaussian linear SEMs with the different maximum number of bad samples ($|B| \in \{1, 30, 60, 90\}$) on the first element of the ordering. The empirical probability of successful graph recovery is shown versus sample size.

2.4 . Simulations and Real-Data

[Comparison with Other Algorithms]



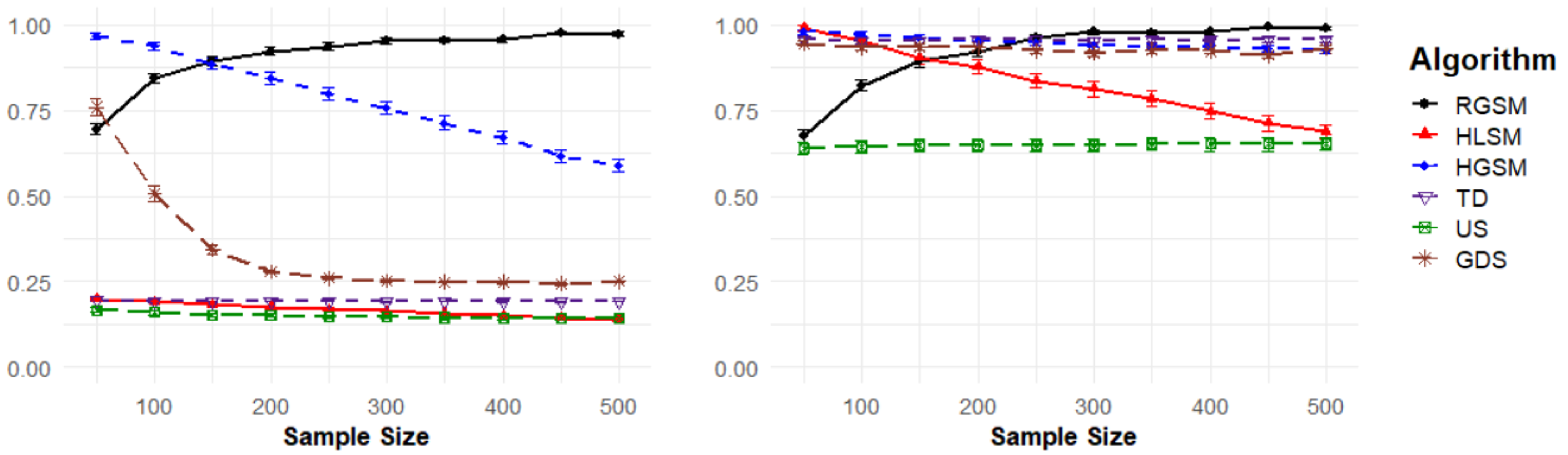
(c) Precision: $|B| = 30$

(d) Recall: $|B| = 30$

Figure 6: Comparison of the proposed algorithm (RGSM) against the HLSM, HGSM, TD, and US algorithms in terms of the average precision and recall for learning 10-node corrupted Gaussian linear SEMs with $|B| \in \{1, 30\}$.

2.4 . Simulations and Real-Data

[All Outliers]



(a) Precision (b) Recall

Figure 10: Comparison of the proposed algorithm (RGSM) against the HLSM, HGSM, TD, and US algorithms in terms of average Hamming distance for learning 10-node corrupted Gaussian linear SEMs when all observation are outliers.

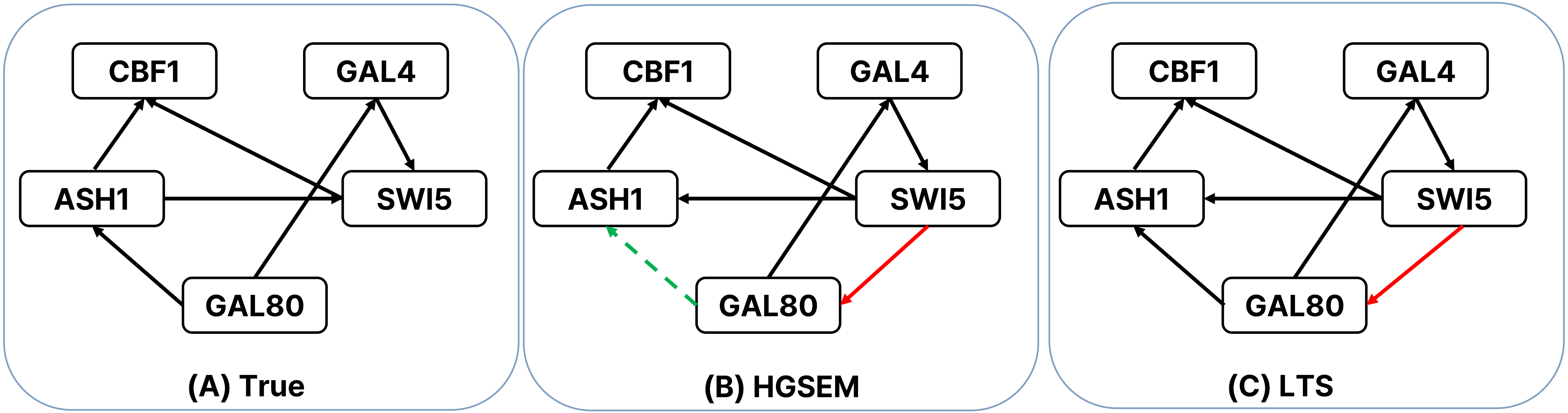
[Real-Data]

IRMA Yeast Real Data: Dataset

- **IRMA Yeast Network:**
  - 5 genes (CBF1, GAL4, SWI5, GAL80, ASH1)
  - Ground-truth DAG with 8 directed edges
- **Experimental Conditions:**
  - "Off" (glucose medium, network repressed)
  - "On" (galactose medium, network induced)
- **Measurements:**
  - 21 time-points per condition ($n = 42$ samples)
  - Steady-state mRNA expression (DNA microarrays)

2.4 . Simulations and Real-Data

[Real-Data]



감사합니다.