Causal Discovery from Observation Data: Introduction and Some Recent Results

Mário A. T. Figueiredo

Instituto Superior Técnico,
Universidade de Lisboa, Portugal

Instituto de Telecomunicações
Lisboa, Portugal

MPML 2023, IST, Lisboa

Nothing exists of which it cannot be asked what is the cause (or reason) why it exists.

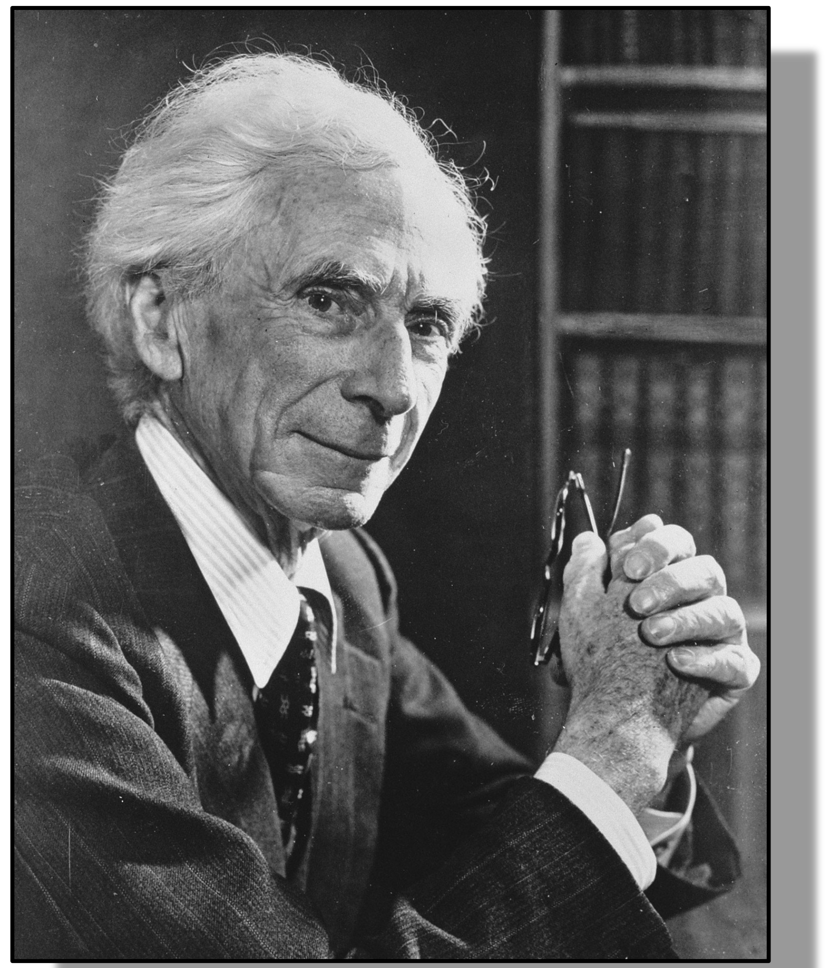
Gottfried Wilhelm Leibniz, 1720

Baruch Spinoza, 1663

We can find no true or existent fact, no true assertion, without there being a sufficient reason why it is thus and not otherwise, although most of the time these reasons cannot be known to us.

- Causality is a deep, controversial topic

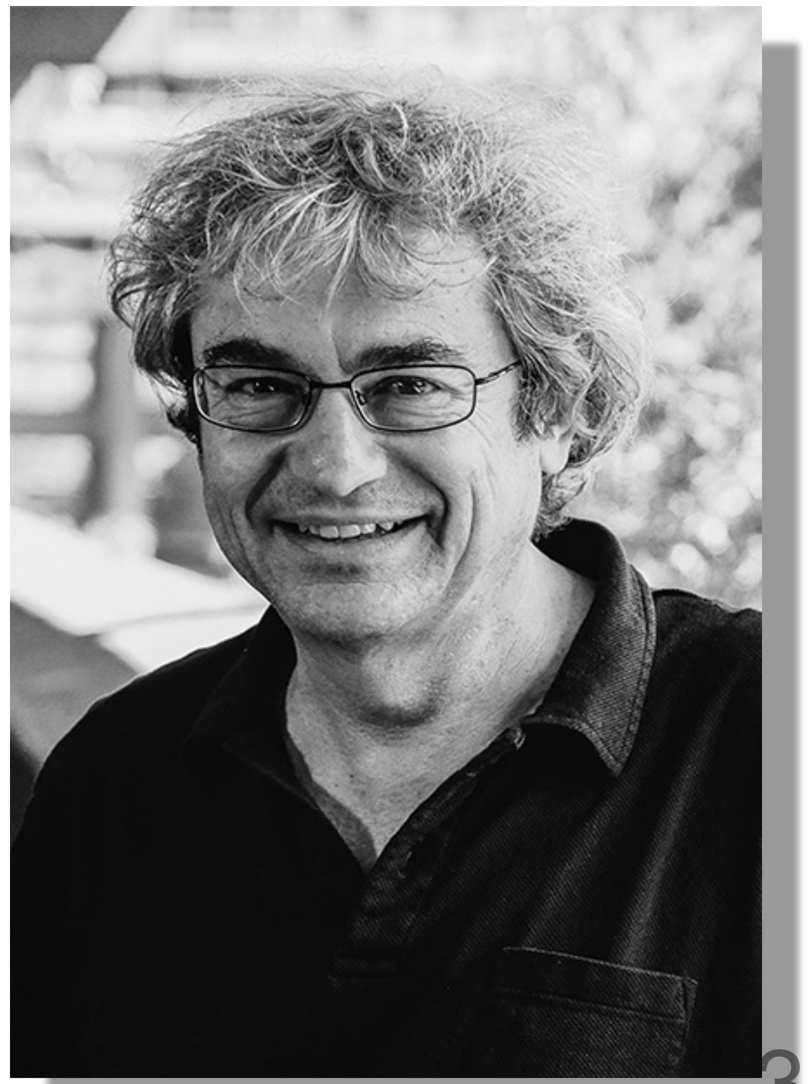
All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never appears. Dr James Ward … makes this a ground of complaint against physics … To me, it seems that … the reason why physics has ceased to look for causes is that, in fact, there are no such things. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm. (Russell 1913, p. 1).[2]



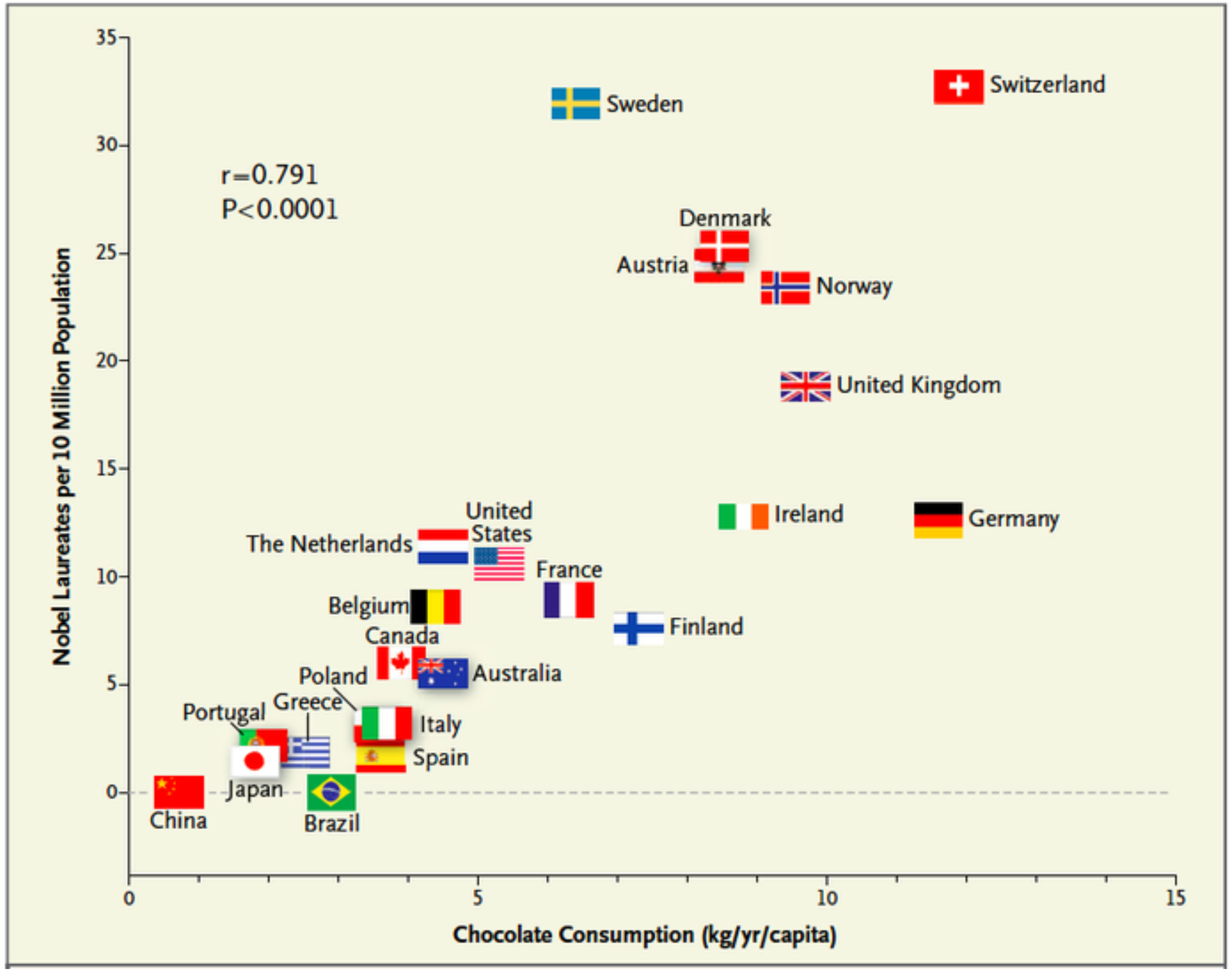
Back to Reichenbach (2022)

Carlo Rovelli
Aix-Marseille University, Université de Toulon, CPT-CNRS, Marseille, France,
Department of Philosophy and the Rotman Institute of Philosophy, Western University, London ON, Canada,
and Perimeter Institute, 31 Caroline Street N, Waterloo ON, Canada

In his 1956 book 'The direction of Time', Hans Reichenbach offered a comprensive analysis of the physical ground of the *direction of time*, the *notion of physical cause*, and the relation between the two. I review its conclusions and argue that at the light of recent advances Reichenbach analysis provides the best account of the physical underpinning of these notions. I integrate recent results in cosmology, and relative to the physical underpinning of records and agency into Reichenbach's account, and discuss which questions it leaves open.



- Correlation vs causation

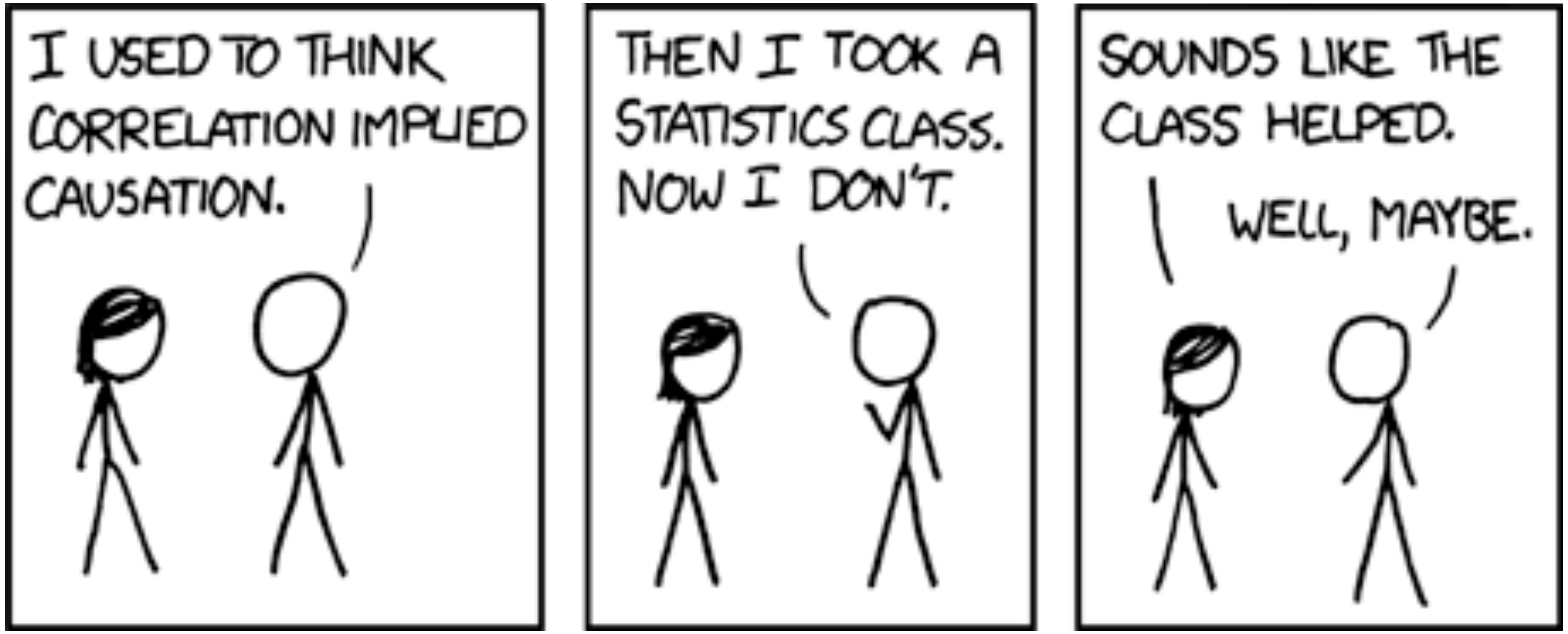


Why scatter plots suggest causality, and what we can do about it.

Carl T. Bergstrom, Jevin D. West

(2018)

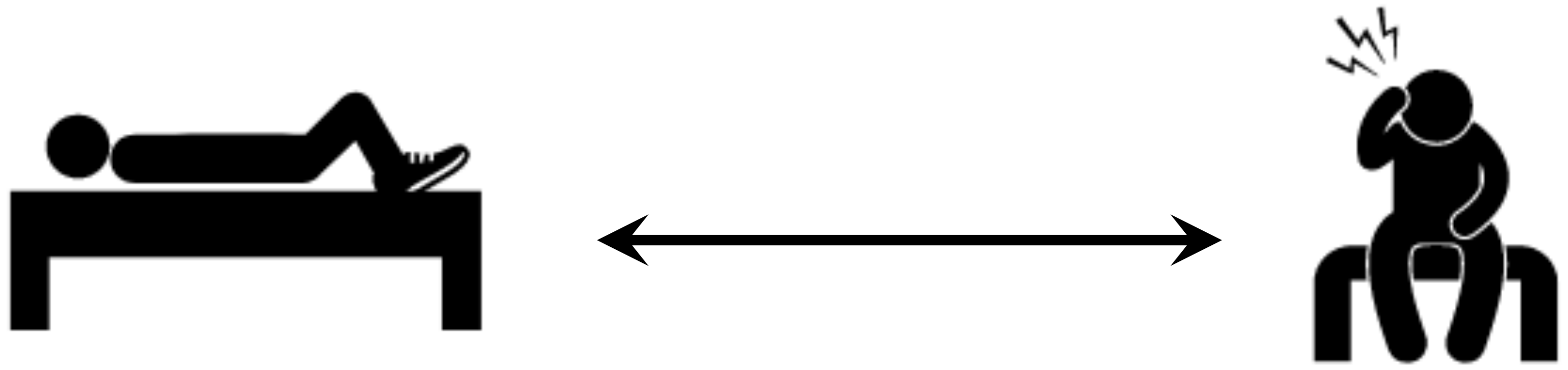
Abstract—Scatter plots carry an implicit if subtle message about causality.
This is a problem for the public understanding of scientific results and perhaps also for professional scientists


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

xkcd.com

- Causal Discovery

  - Suppose we have statistical evidence: people who wake up with their shoes on are more likely to have a headache.
  - Does sleeping with the shoes on cause headache?
  - Or vice-versa?
  - Or is it just a coincidence?



Picture credits: “Introduction to Causal Inference" (Brady Neal, 2022)

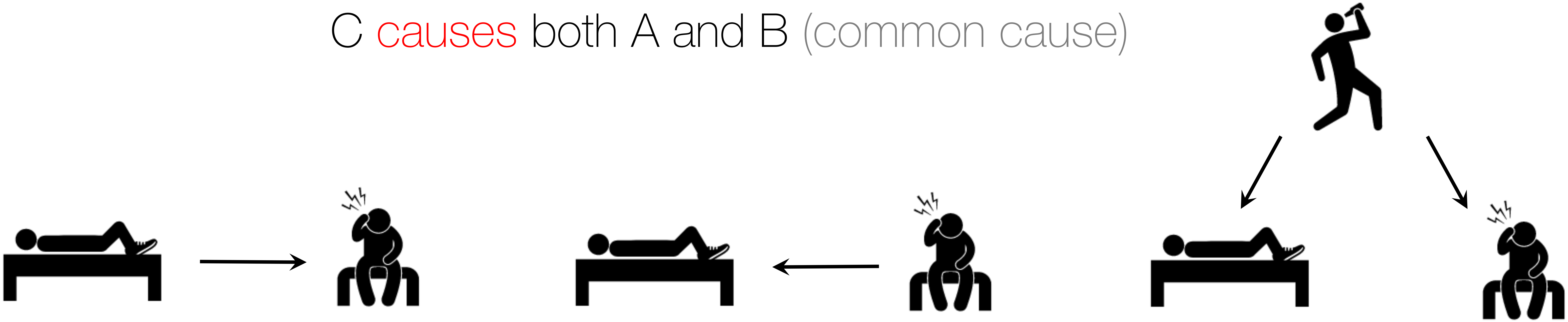
- Reichenbach's Common Cause Principle (1956)

  - If two events A and B are dependent, then

    A causes B, or

    B causes A, or

    C causes both A and B (common cause)



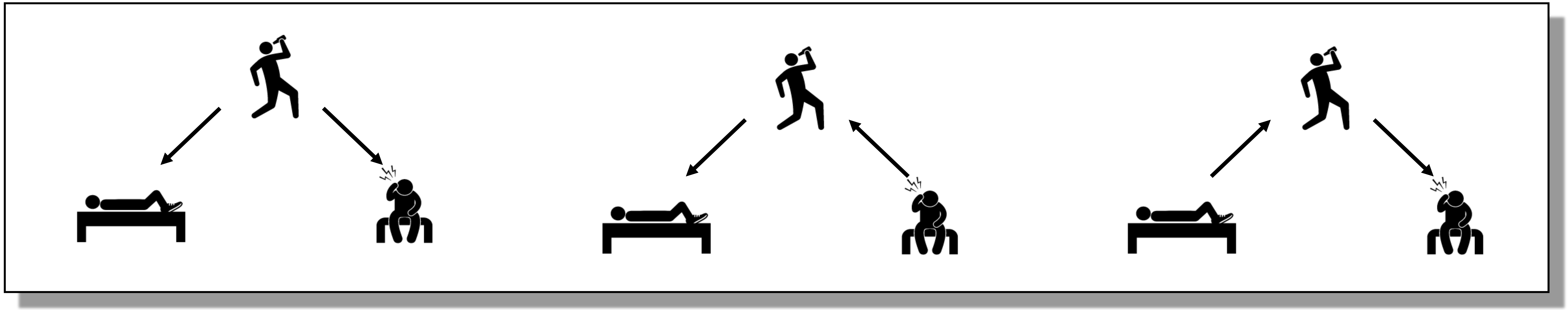
• Conditional Independence

- Suppose we find that  **independent from**  **, given** 

- P(shoes, headache | drunk) $\simeq$ P (shoes | drunk) P(headache | drunk)
  (conditional independence)

- Compatible with 3 causal mechanisms:

(Markov equivalence class)

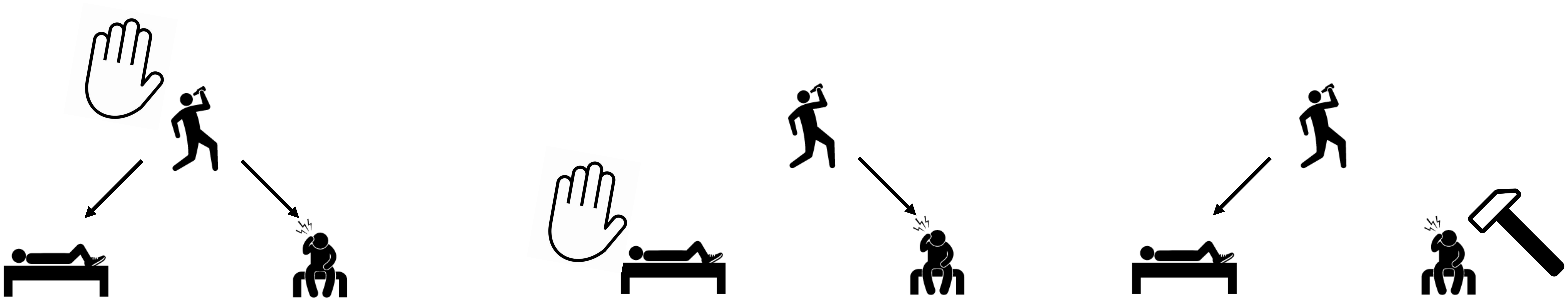


How to find the true one?

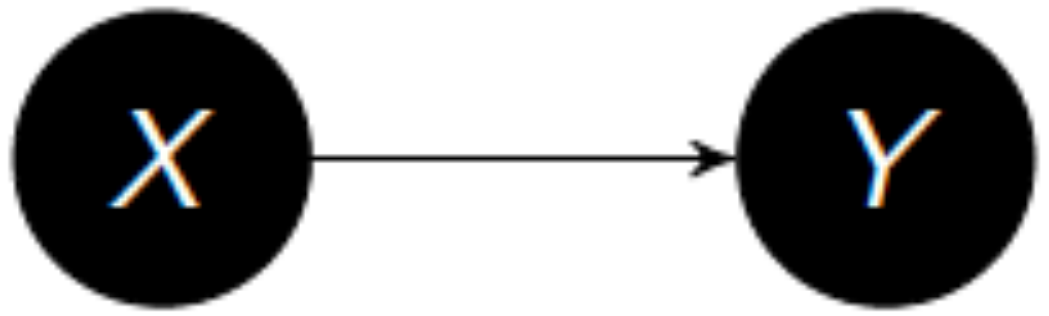
• Causal Discovery from Observations

• Causal Discovery from Observations

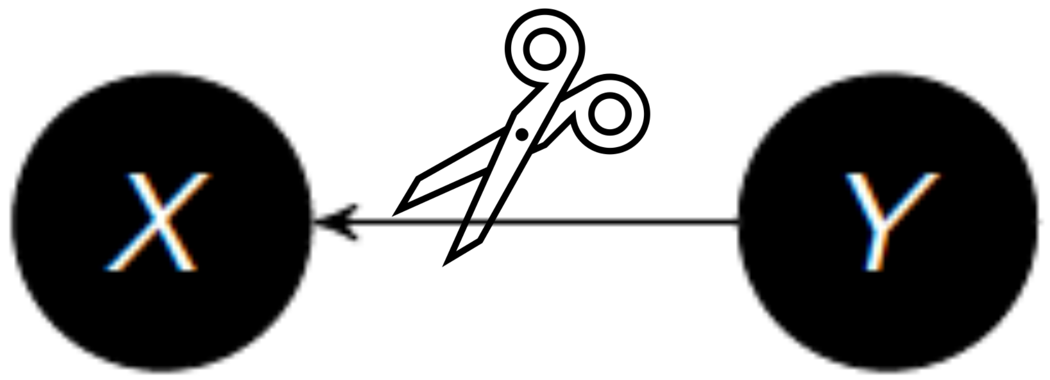
- Interventions
  - Science: hypotheses ➞ experiments (interventions)
  - Causal discovery with intervention data
  - Often impossible, impractical, unethical, …



- Formalizing interventions:
  Pearl's "do-calculus"
  - Key insight: observation ≠ intervention
  - Intervention, do(x), cut the input arrows; set the value.

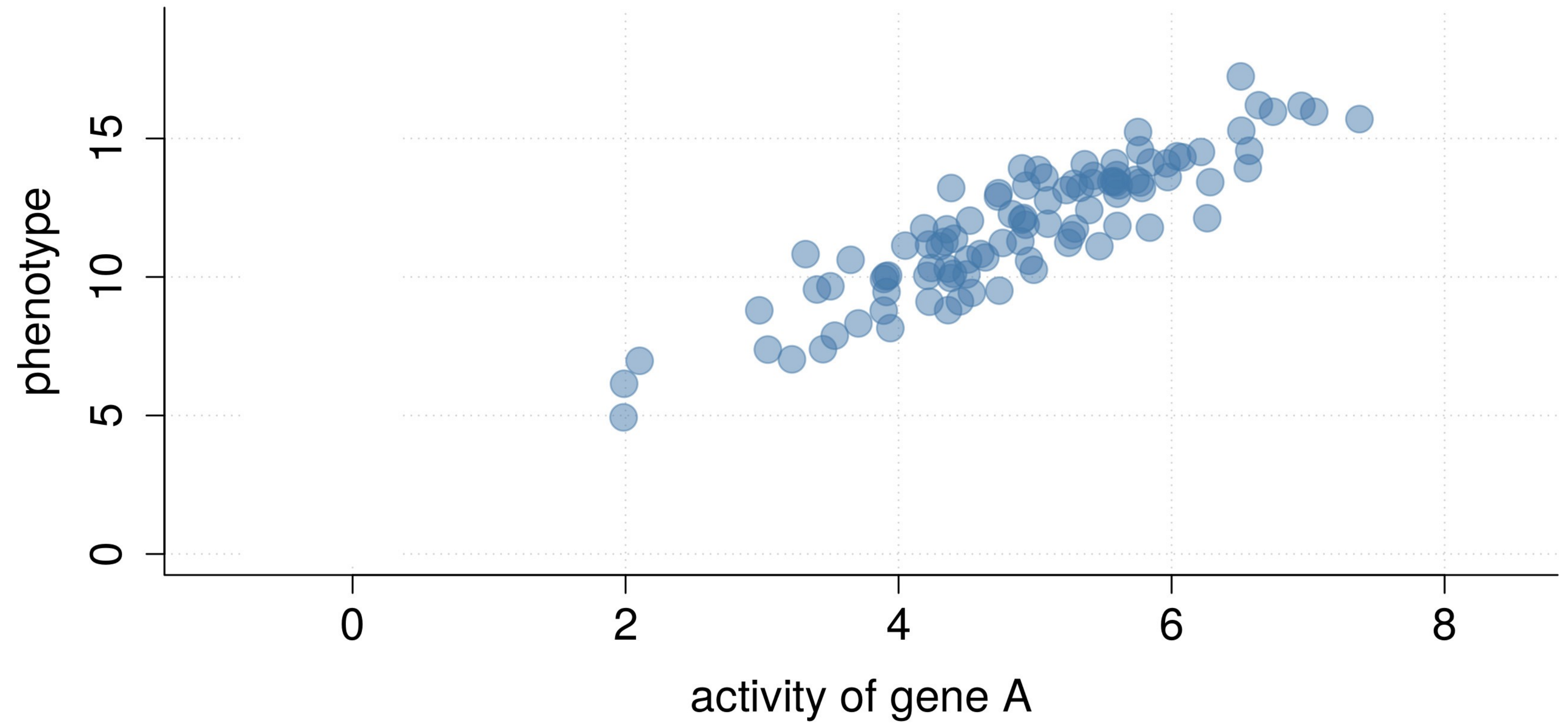


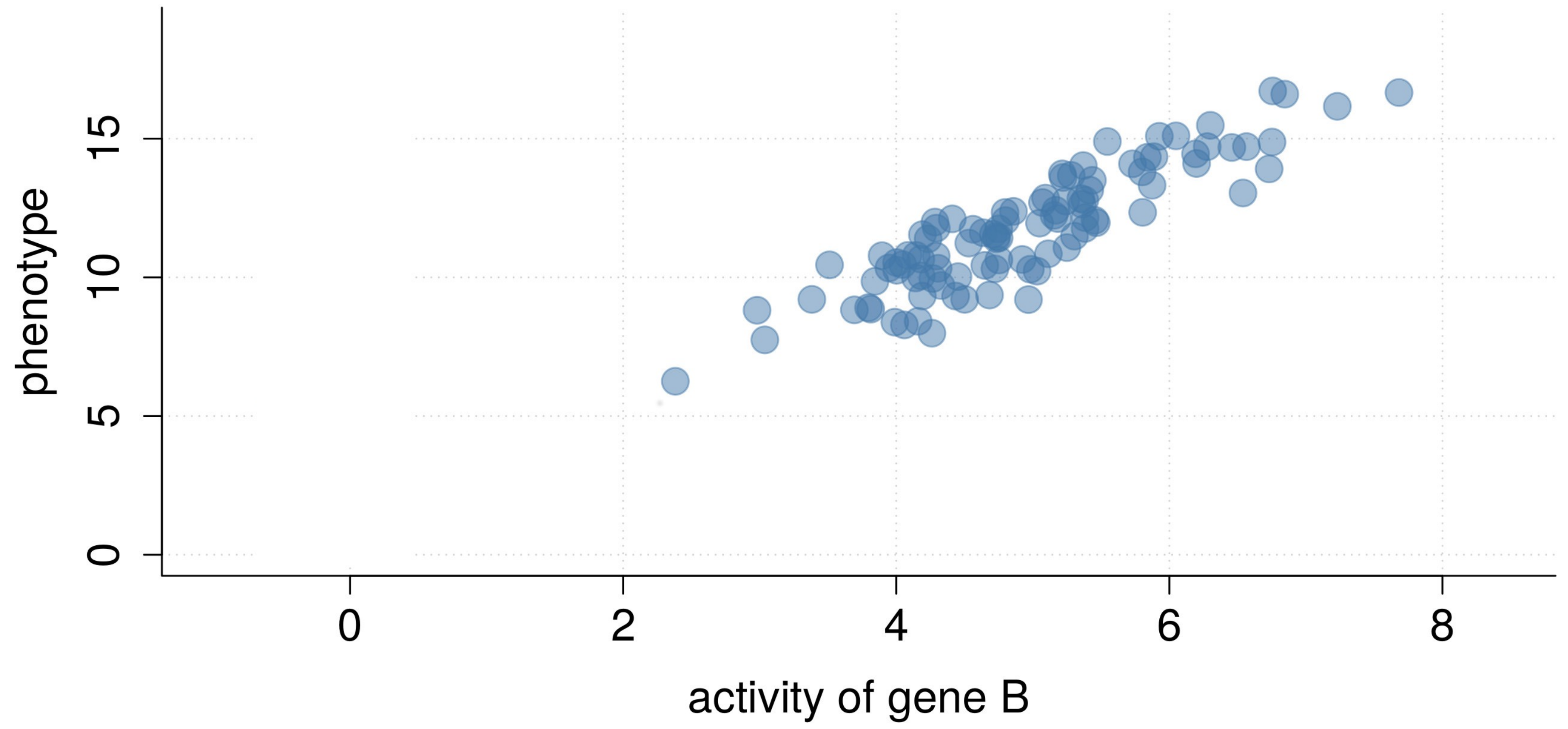
$$\mathbb{P}\big(Y = y | \mathrm{do}(X = x)\big) = \mathbb{P}\big(Y = y | X = x\big)$$



$$\mathbb{P}\big(Y = y | \mathrm{do}(X = x)\big) = \mathbb{P}\big(Y = y\big) \neq \mathbb{P}\big(Y = y | X = x\big)$$

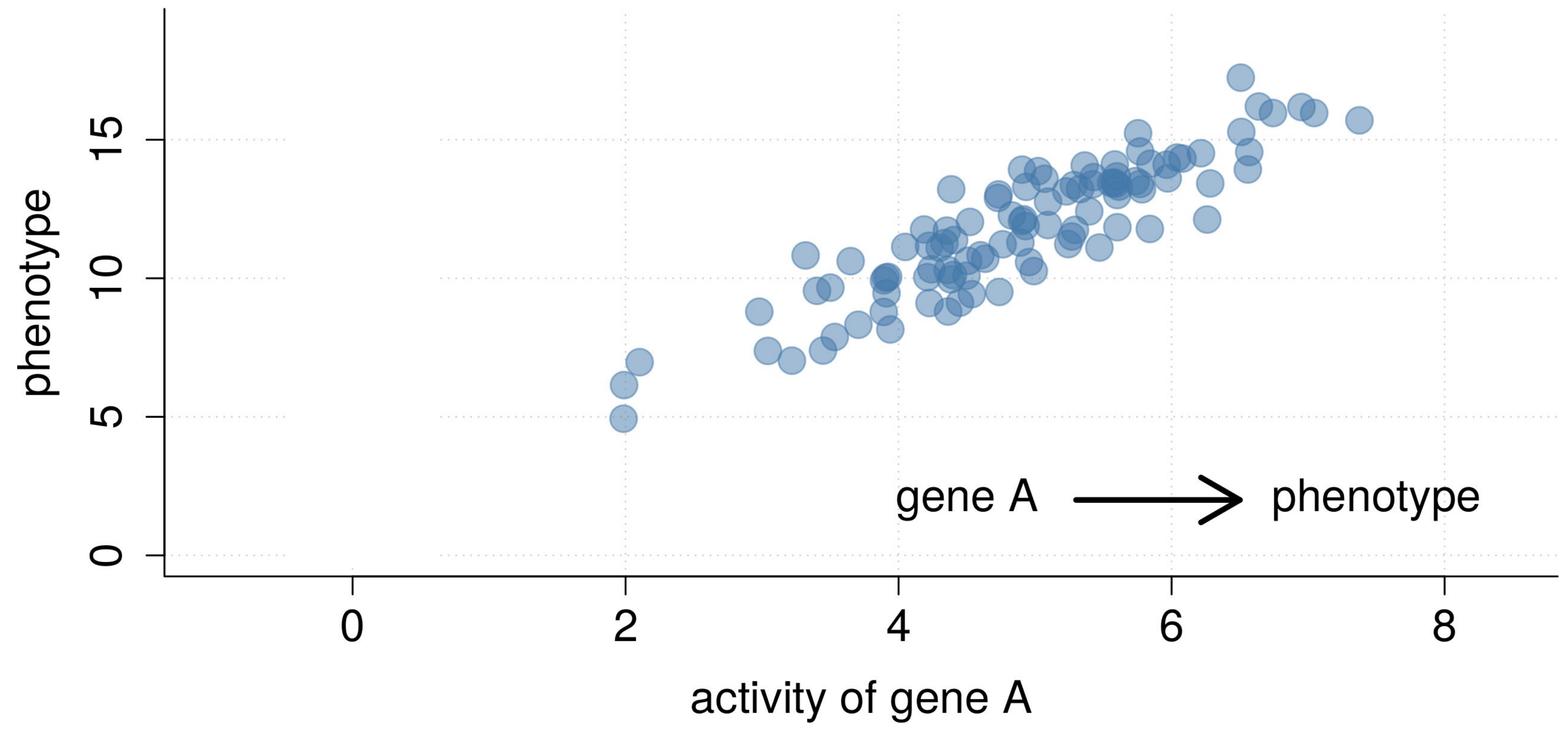
• Gene Knockout

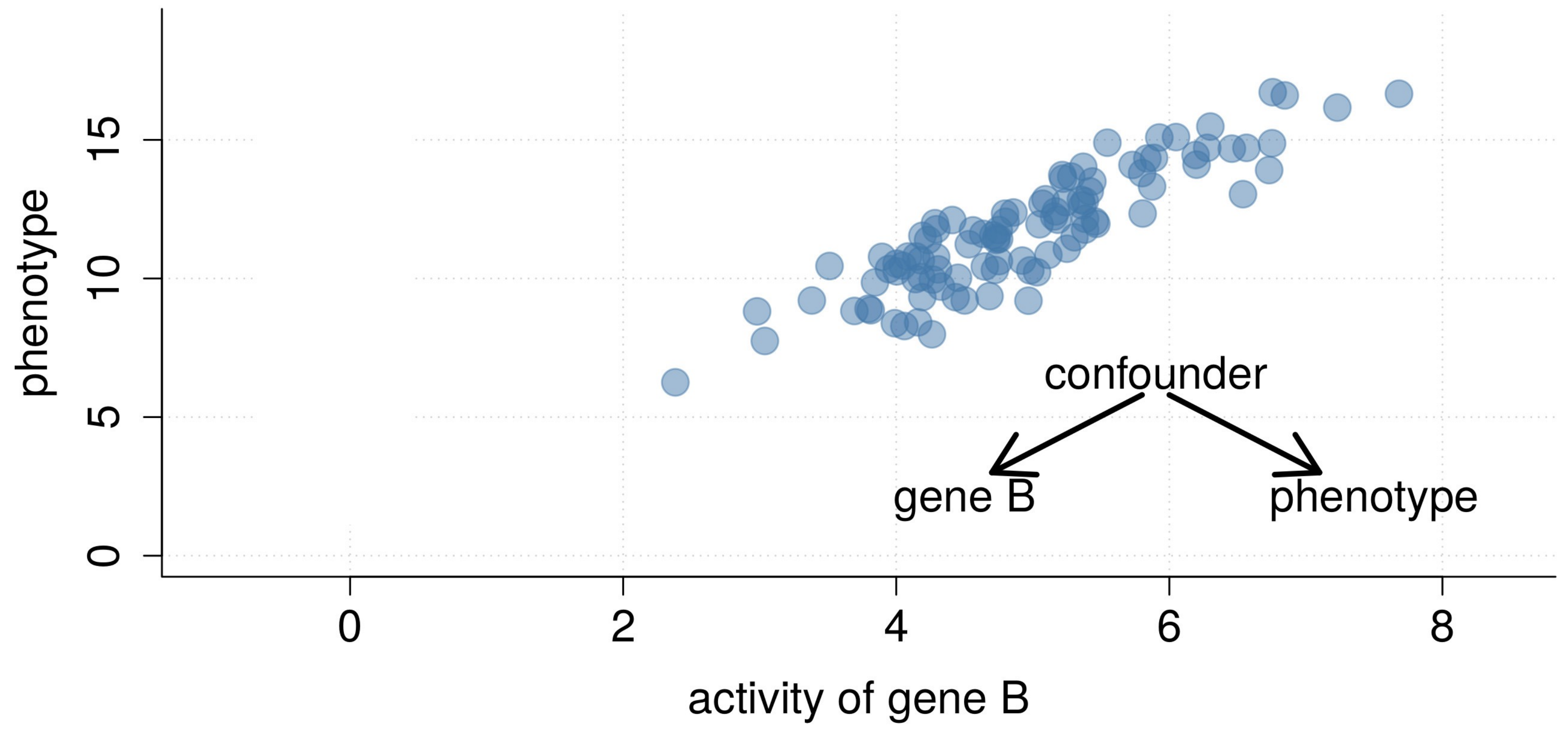




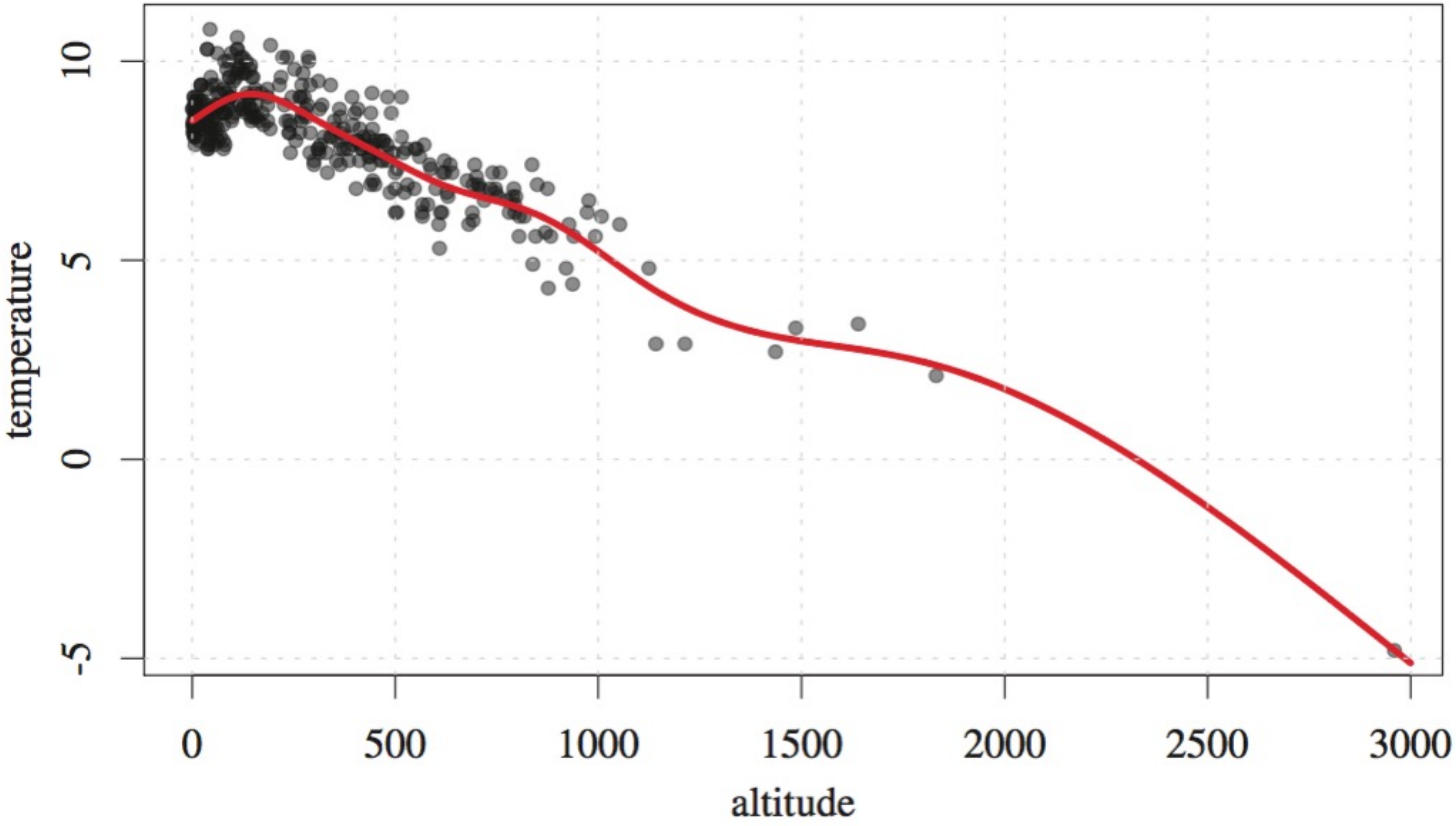
delete gene A (force to 0); predict the phenotype.

delete gene B (force to 0); predict the phenotype.

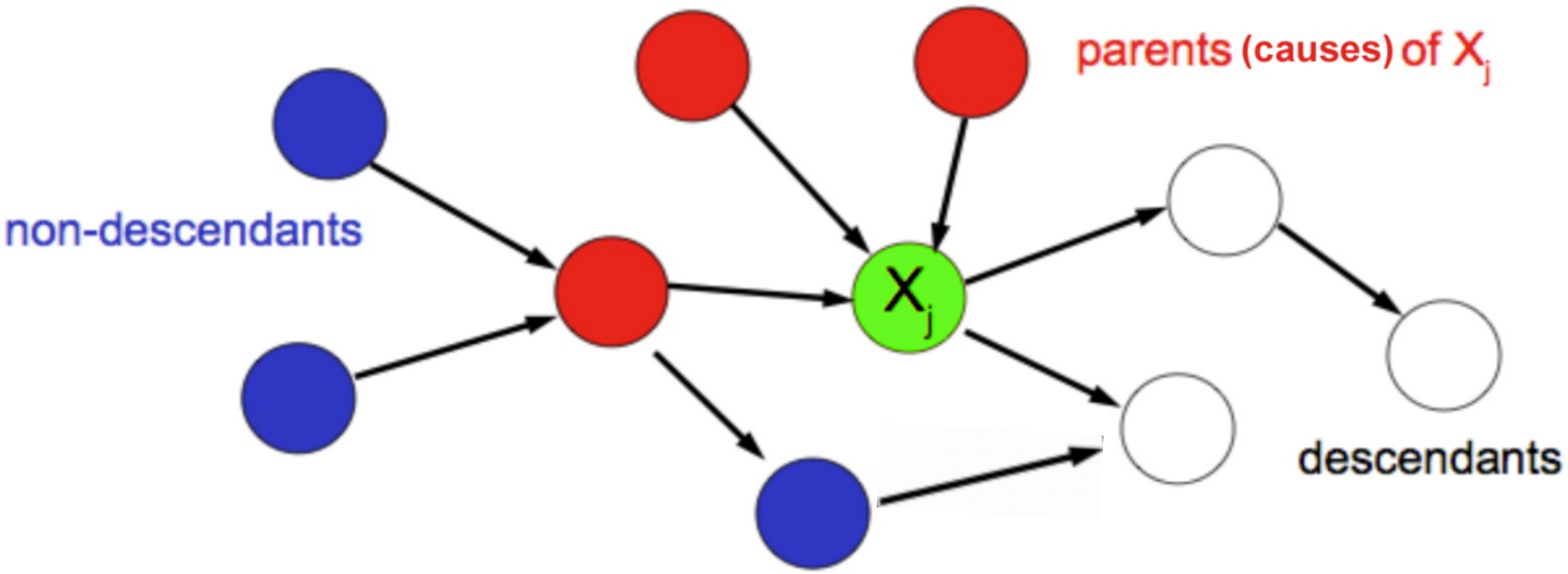




Example by Jonas Peters


10
5
0
-5
temperature
0
500
1000
1500
2000
2500
3000
altitude

- Structural Causal Model (SCM)
  - Directed acyclic graph (DAG): $\mathcal{G} = (\mathcal{V}, \mathcal{E}), \;\; \mathcal{V} = (X_1, ..., X_n)$



$$X_j \leftarrow f_j\left(X_{pa_j}, U_j\right)$$

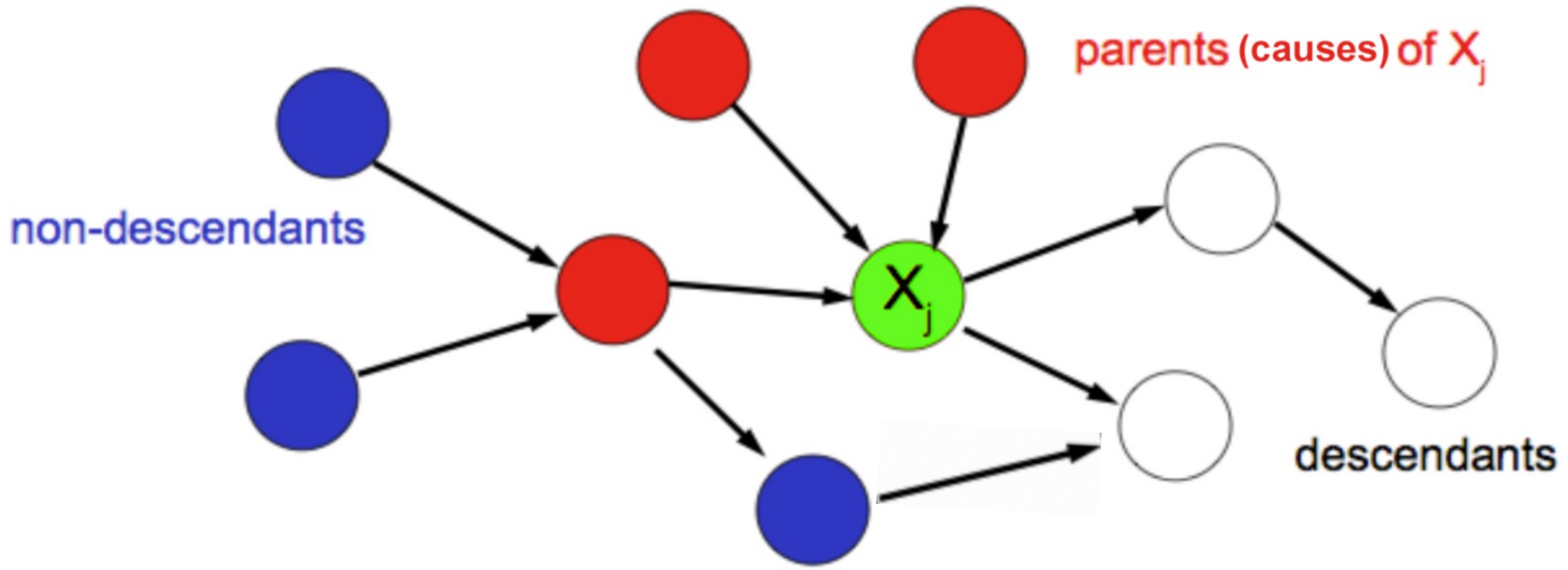
U_i are mutually independent

exogenous/unexplained variables; noise

  - Every f_j is a causal mechanism

• Structural Causal Model (SCM)

- Each mechanism entails a local conditional $P(X_j | X_{pa_j})$



$$X_j \leftarrow f_j\left(X_{pa_j}, U_j\right)$$

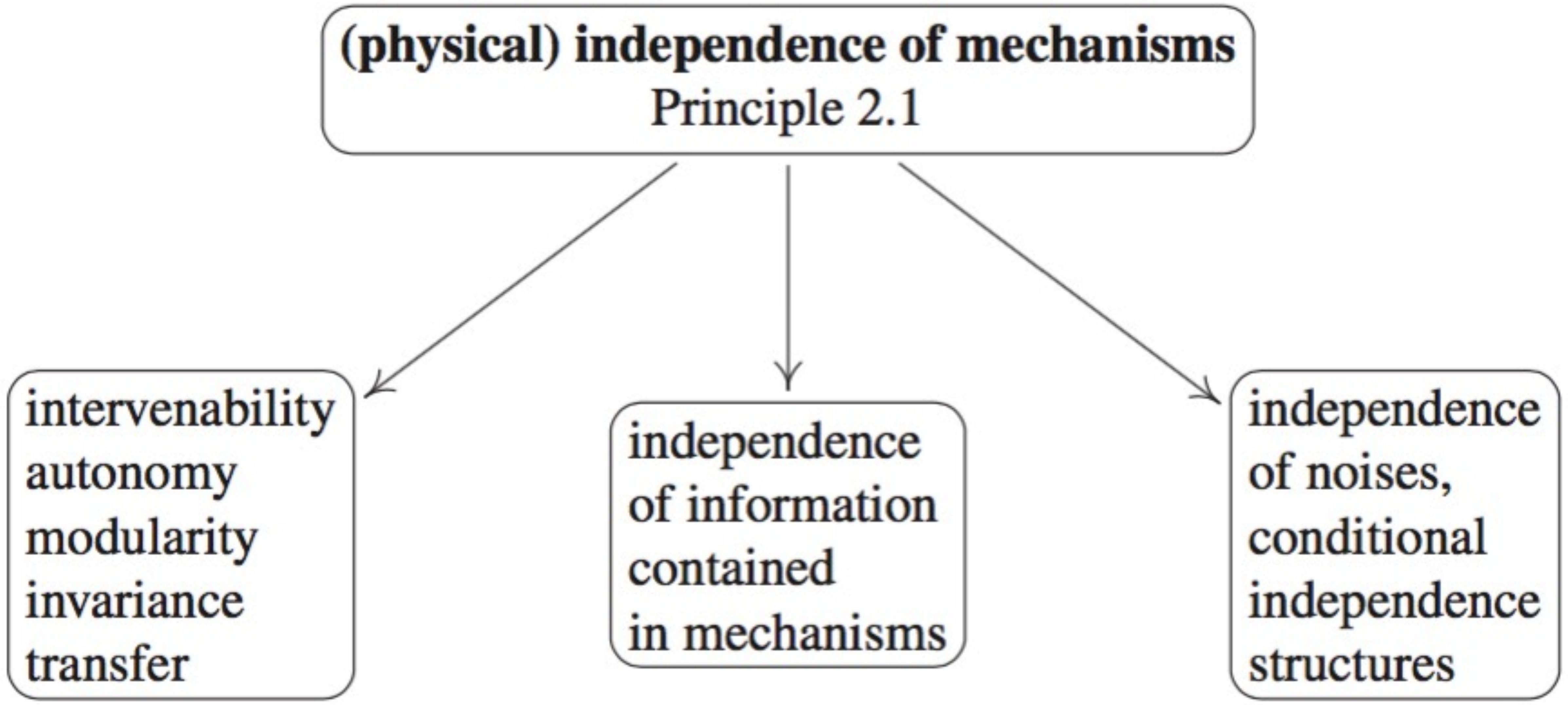
U_i are mutually indendent

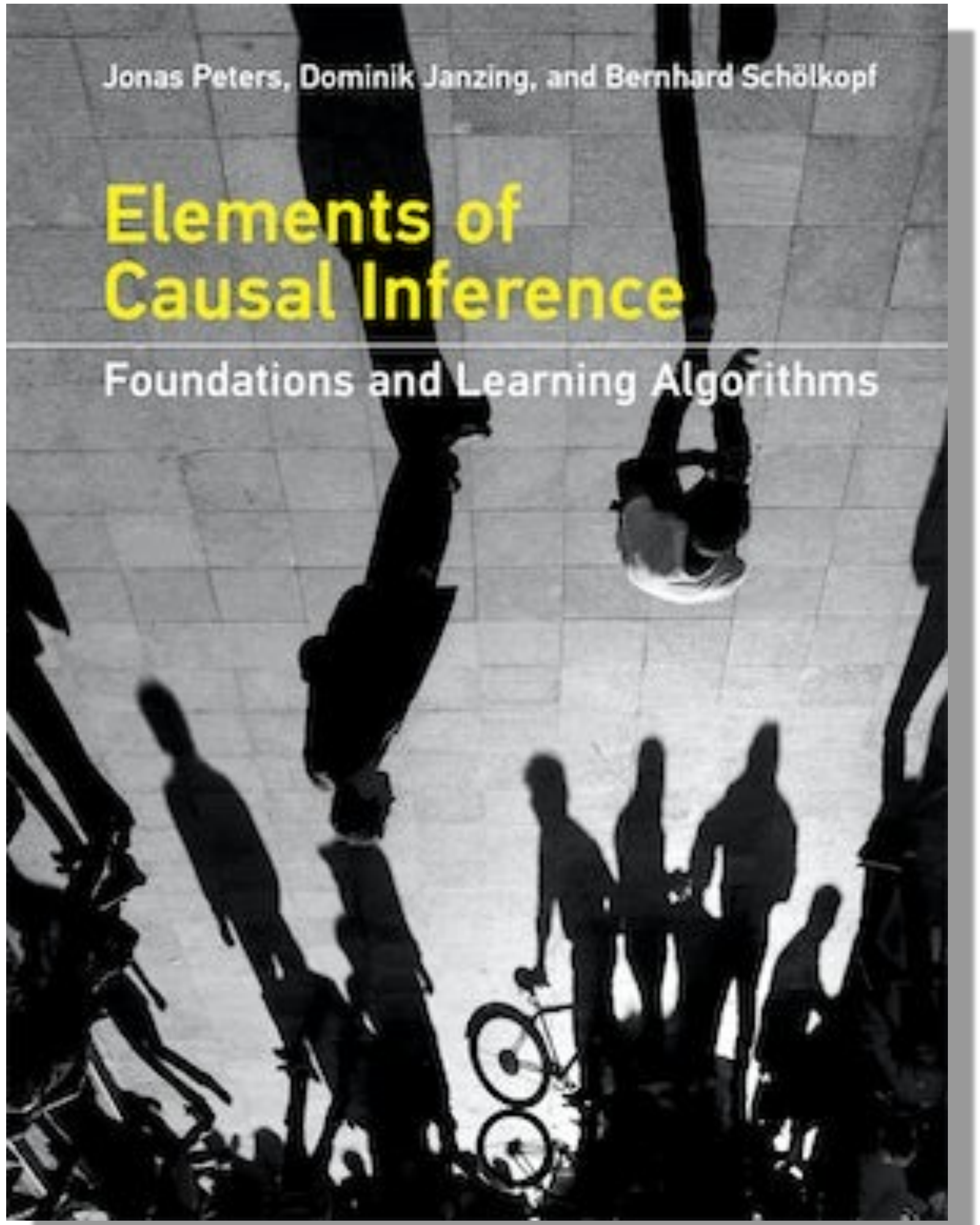
exogenous/unexplained variables; noise

- Joint distribution: $P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P(X_j | X_{pa_j})$

Principle 2.1 (Independent Mechanisms) *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.





- The Beuchet chair illusions (and others): the brain assumes the viewed object (the cause) and the viewing angle (the observation mechanism) are independent.

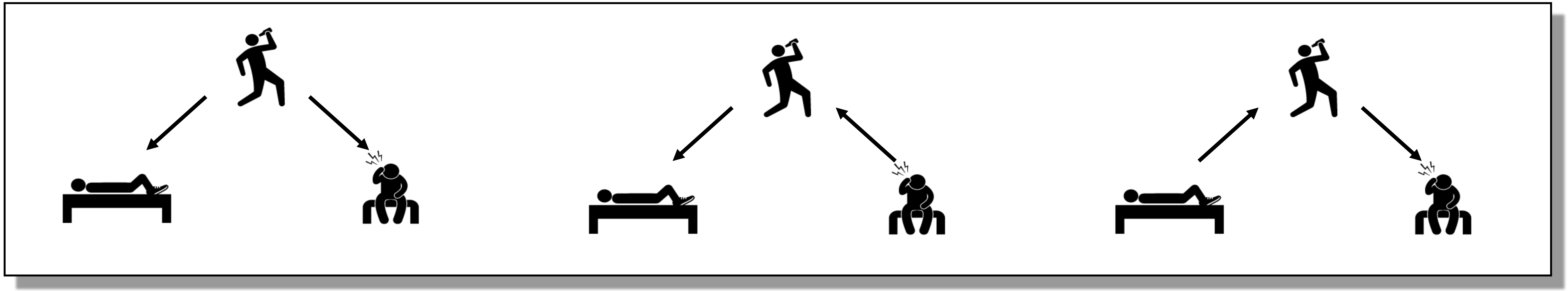
Graphical Causal Discovery

- Can we recover $\mathcal{G}, f_1, \ldots, f_n, P(U_1), \ldots, P(U_n)$

  from $P(X_1, \ldots, X_n)$ or from samples/data?

- In general, only up to the Markov equivalence class

Markov Equivalence Class



- Interventions in SCMs
  - Hard interventions: $do(X_j = x_j)$

    i.e. replace $X_j \leftarrow f_j(X_{pa_j}, U_j)$ with $X_j \leftarrow x_j$

  - Soft interventions: many other possibilities, e.g.

    replace $X_j \leftarrow f_j(X_{pa_j}, U_j)$ with $X_j \leftarrow \tilde{U}_j$

Most important case: confounder correction

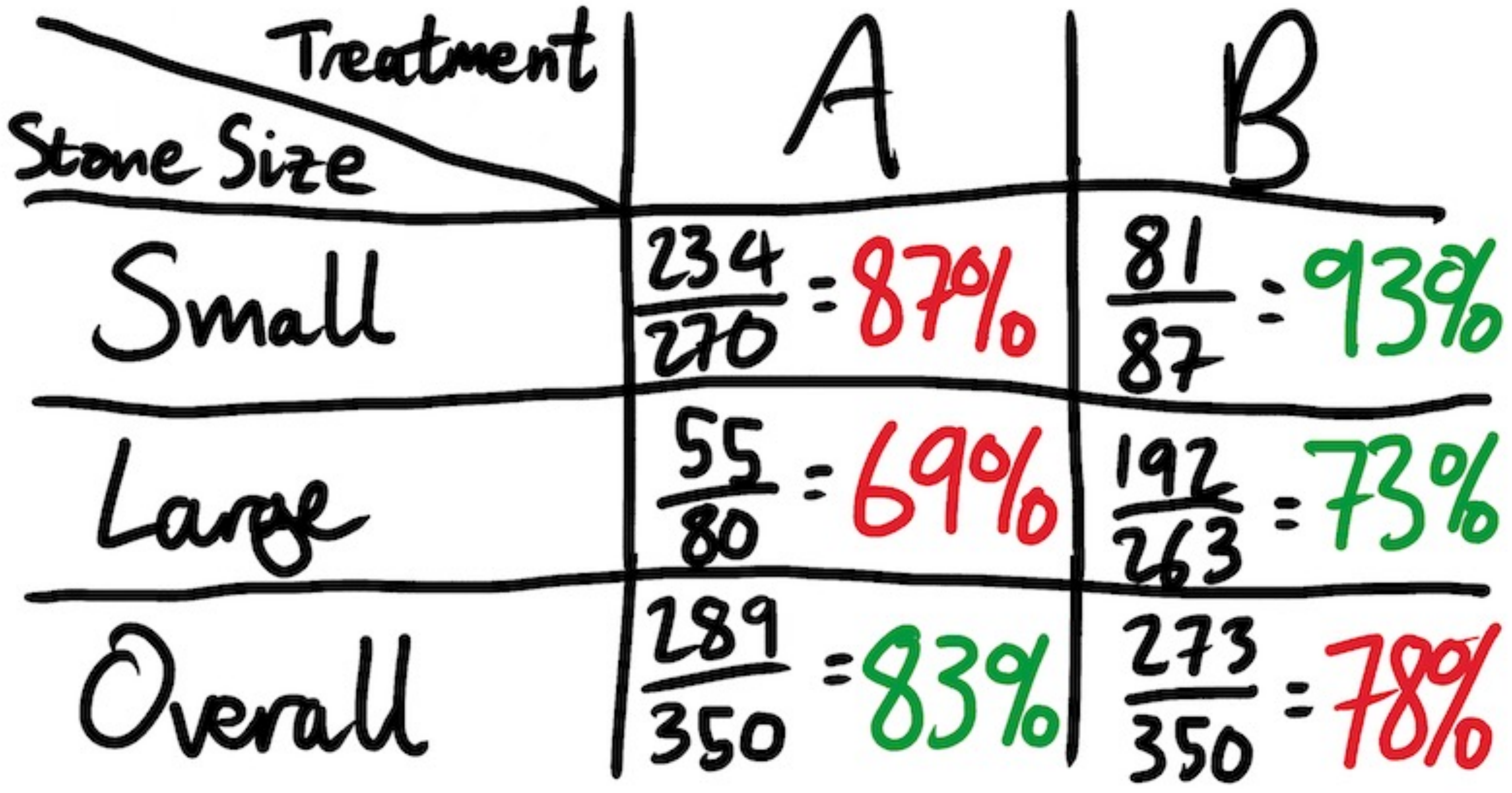
$$p(y|do(x)) = \sum_z p(y|x,z)p(z) \neq \sum_z p(y|x,z)p(z|x) = p(y|x)$$

Adapted from Janzing and Weichald, 2019.

- Core idea behind randomized trials (medicine, economics, A/B testing, …)
- Key aspect: assignment of X is independent of Z. Possibility: random

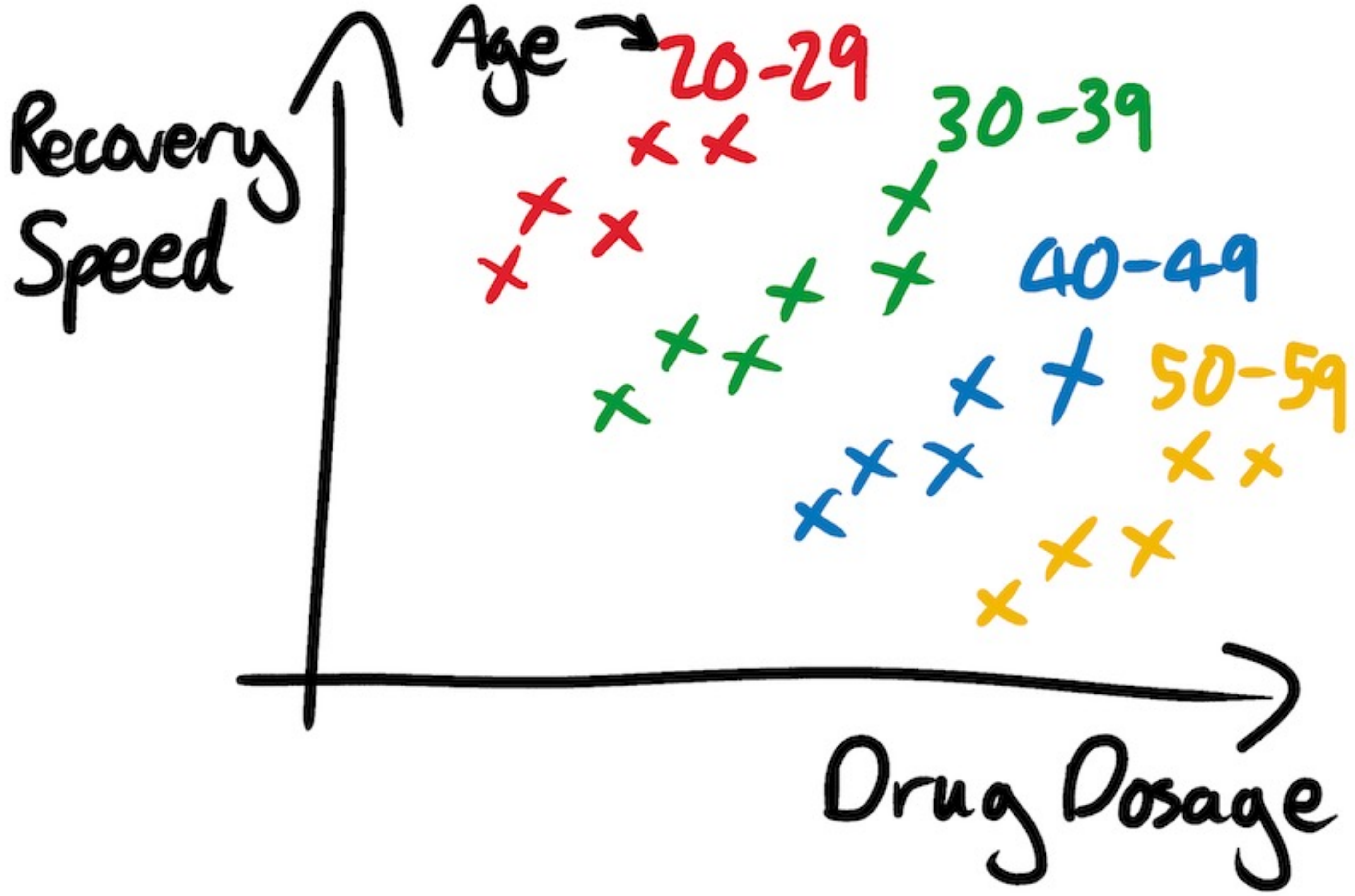
• Why are RTs needed? Simpson’s “paradox”

- Two treatments for kidney stones: A and B
- Two types of stones: small and large
- Success rates:



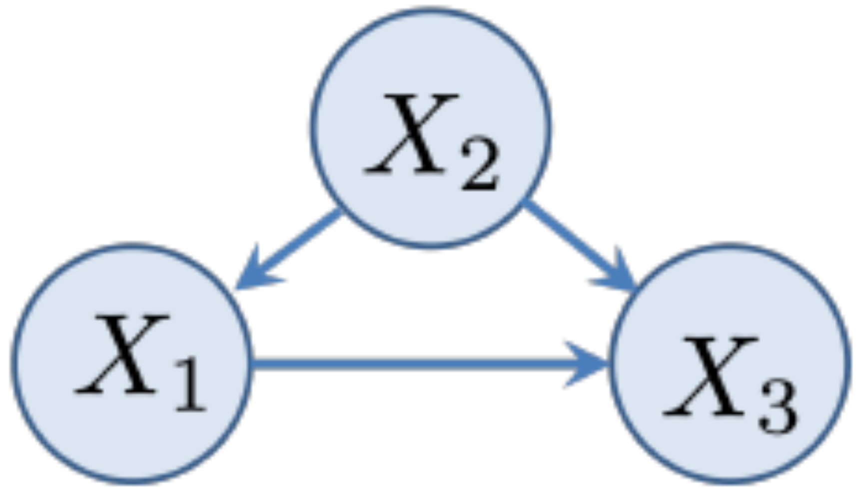
Picture from robertheaton.com

- Simpson's "paradox" with hidden confounder

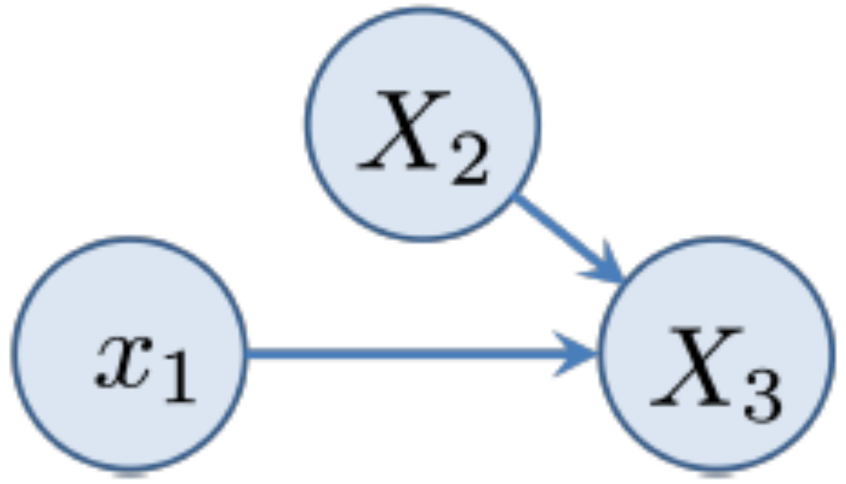


Picture from roberttheaton.com

- Marginalizing vs adjusting/controlling



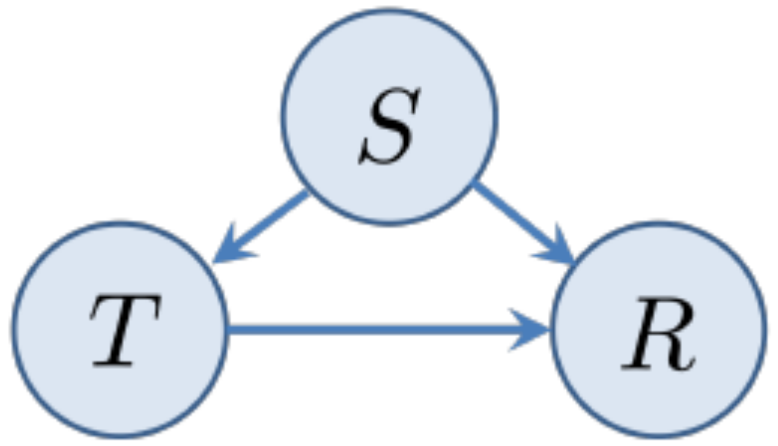
$$\mathbb{P}\big(X_3 = x_3 | X_1 = x_1\big) = \sum_{x_2} \mathbb{P}(X_3 = x_3 | X_1 = x_1, X_2 = x_2)\, \mathbb{P}(X_2 = x_2 | X_1 = x_1)$$

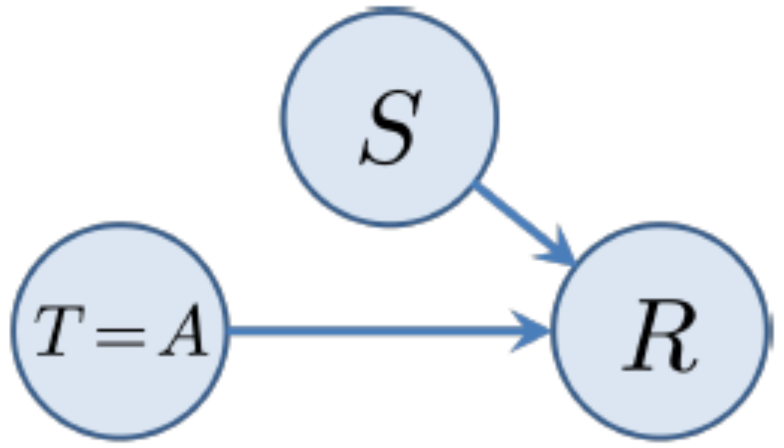


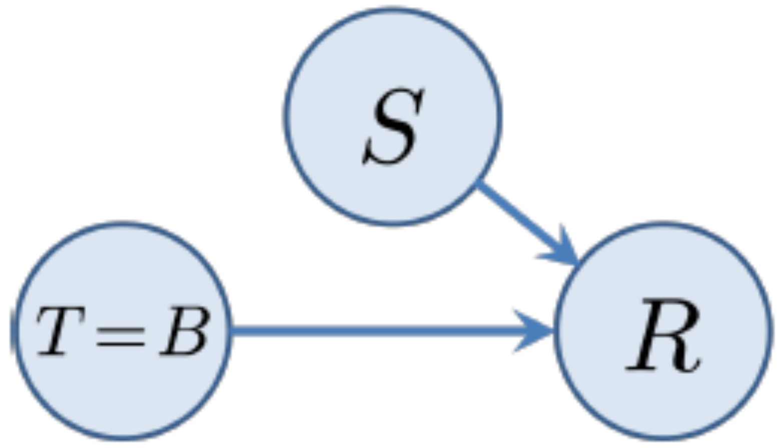
$$\mathbb{P}\big(X_3 = x_3 | \mathrm{do}(X_1 = x_1)\big) = \sum_{x_2} \mathbb{P}(X_3 = x_3 | X_1 = x_1, X_2 = x_2)\, \mathbb{P}(X_2 = x_2)$$

Revisiting Simpson's "Paradox"

Stone Size \ Treatment	A	B
Small	$\frac{234}{270}$ = 87%	$\frac{81}{87}$ = 93%
Large	$\frac{55}{80}$ = 69%	$\frac{192}{263}$ = 73%
Overall	$\frac{289}{350}$ = 83%	$\frac{273}{350}$ = 78%



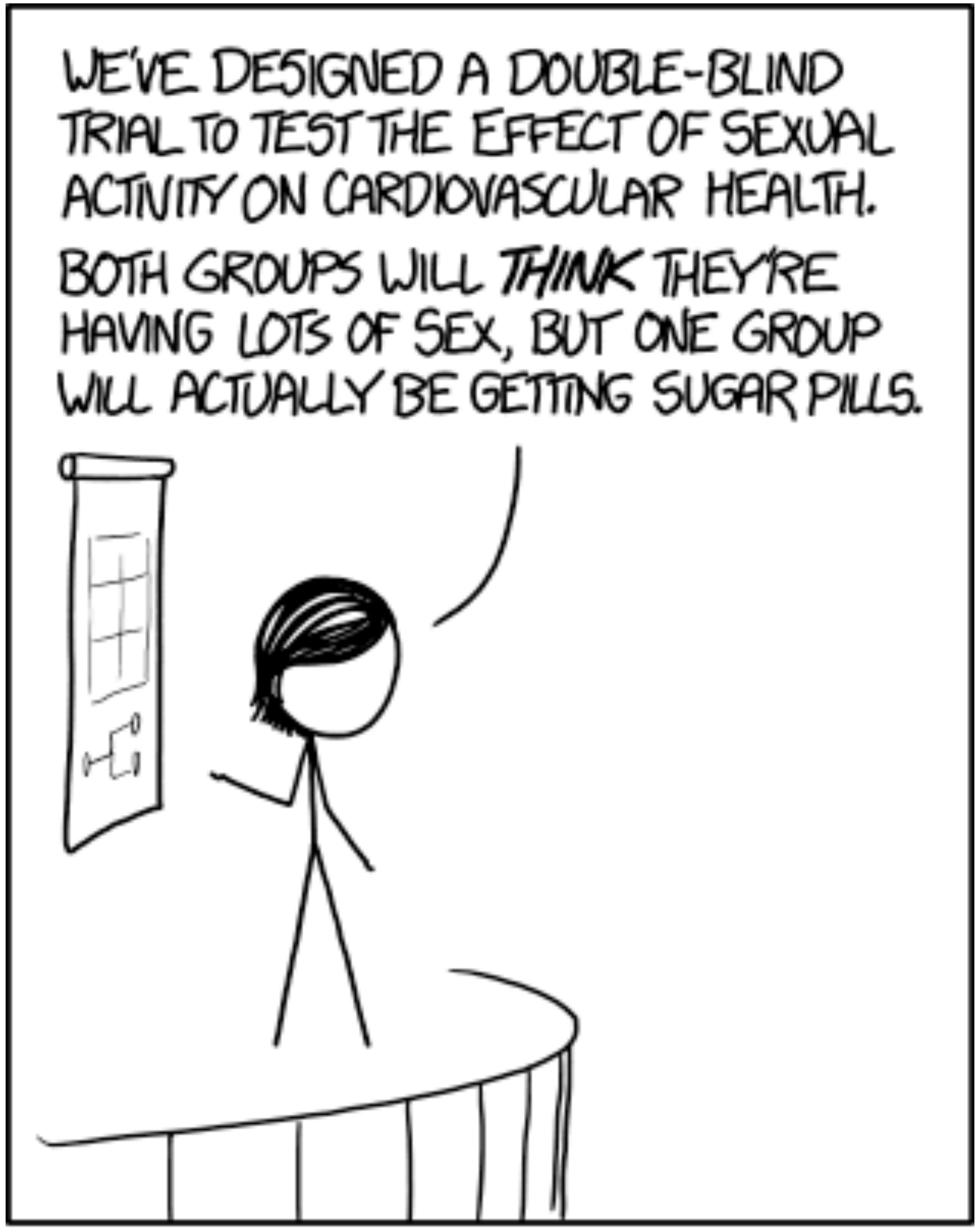




- Variables: $T \in \{A, B\}$, $S = \{s, l\}$, result $R \in \{1, 0\}$.
- Adjustment formulas:

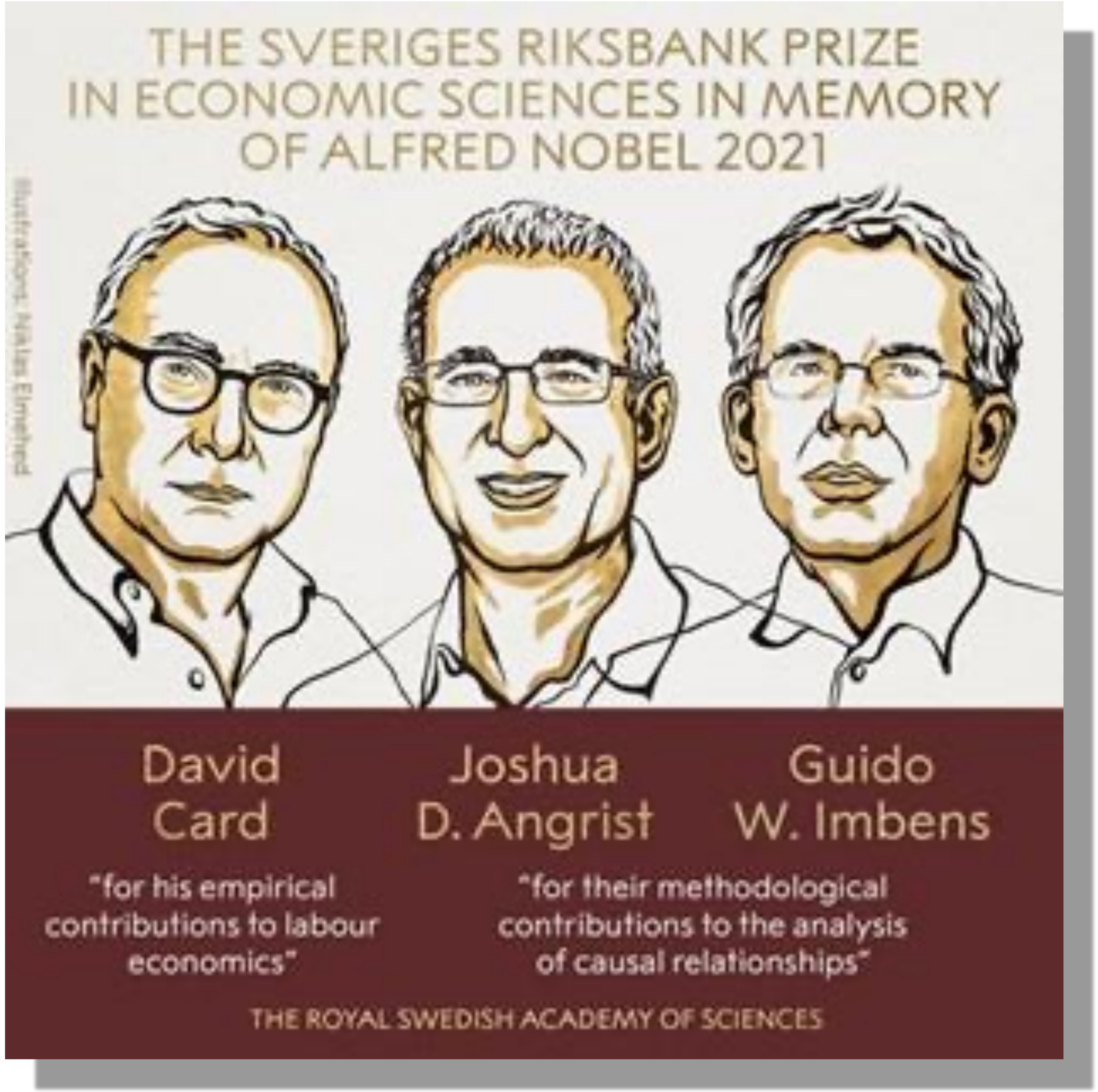
$$\begin{aligned}
&\mathbb{P}\big(R = 1|\text{do}(T = A)\big) \\
&\quad = \mathbb{P}(R{=}1|T{=}A, S{=}s)\,\mathbb{P}(S = s) + \mathbb{P}(R{=}1|T{=}A, S = l)\,\mathbb{P}(S = l) \\
&\quad = 0.87 \times \frac{357}{700} + 0.69 \times \frac{343}{700} \simeq 0.782 \\
&\mathbb{P}\big(R = 1|\text{do}(T = B)\big) \\
&\quad = \mathbb{P}(R{=}1|T{=}B, S{=}s)\,\mathbb{P}(S = s) + \mathbb{P}(R{=}1|T{=}B, S = l)\,\mathbb{P}(S = l) \\
&\quad = 0.93 \times \frac{357}{700} + 0.73 \times \frac{343}{700} \simeq 0.832
\end{aligned}$$

- Controlling/adjusting for size: treatment B is better than A.



THE LIMITATIONS OF BLIND TRIALS

xkcd.com


THE SVERIGES RIKSBANK PRIZE
IN ECONOMIC SCIENCES IN MEMORY
OF ALFRED NOBEL 2021
David
Card
Joshua
D. Angrist
Guido
W. Imbens
"for his empirical
contributions to labour
economics"
"for their methodological
contributions to the analysis
of causal relationships"
THE ROYAL SWEDISH ACADEMY OF SCIENCES

• The “Simplest” Problem: Cause-Effect

- How to distinguish between

  $X \rightarrow Y$ and $X \leftarrow Y$

  only from observations?

- Usual assumptions: no hidden confounders,
  i.e, one of the two hypotheses is “true”.

• The Cause-Effect Problem

- Without interventions, we need model assumptions
- A foundational principle (mentioned above):

  independence of cause and mechanism (ICM)

- What is meant by independence? not statistical, but functional:

  P(cause) and P(effect | cause) ignore each other!

- Several instantiations: information geometry, algorithmic (Kolmogorov) complexity, stochastic complexity (MDL), …

- Additive Noise Models
  - Instantiation of the ICM: additive noise model (ANM)
  - Very simple SCM (real variables):

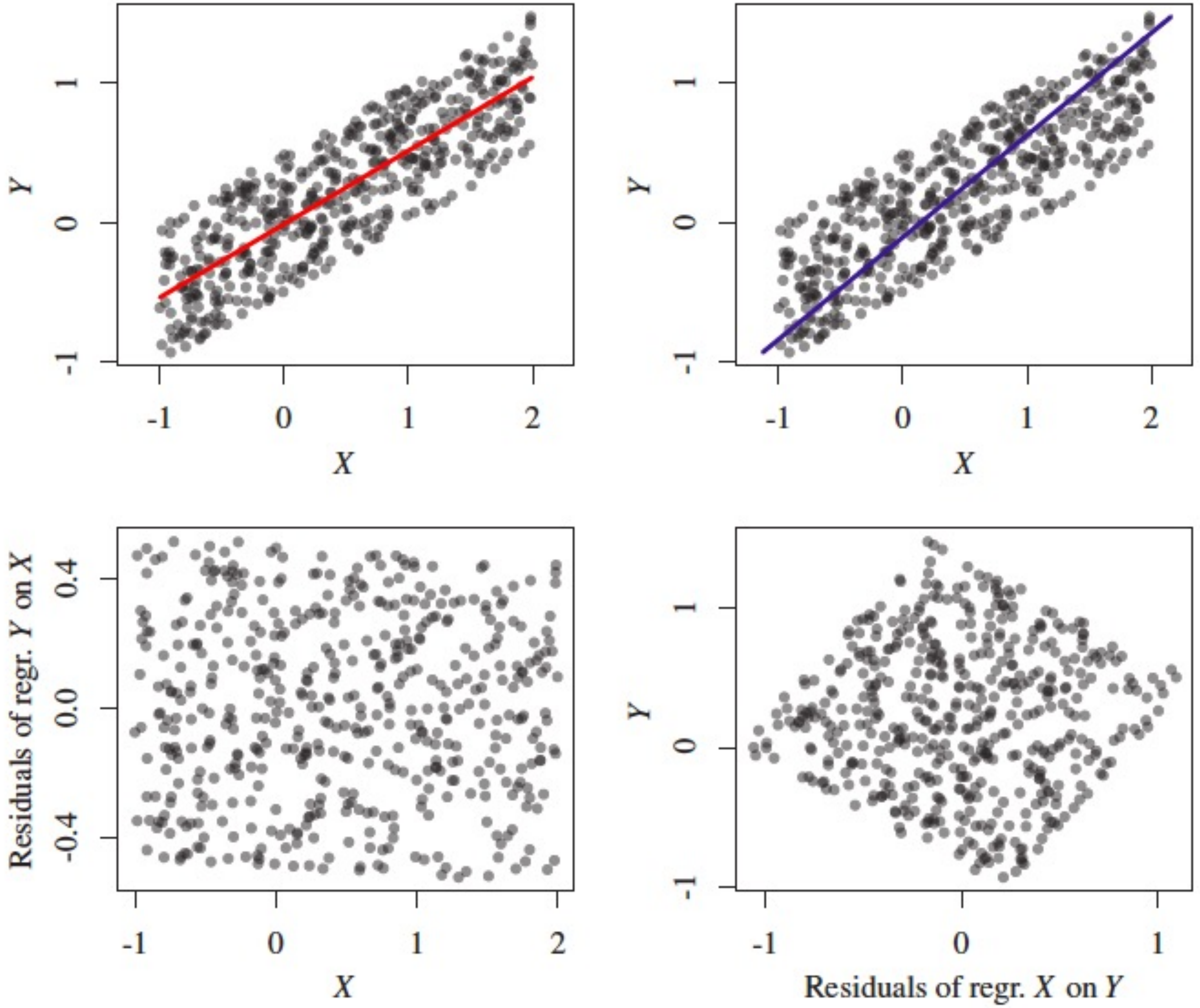
$$Y \leftarrow f(X) + N_Y \quad \text{with} \quad N_Y \perp\!\!\!\perp X$$

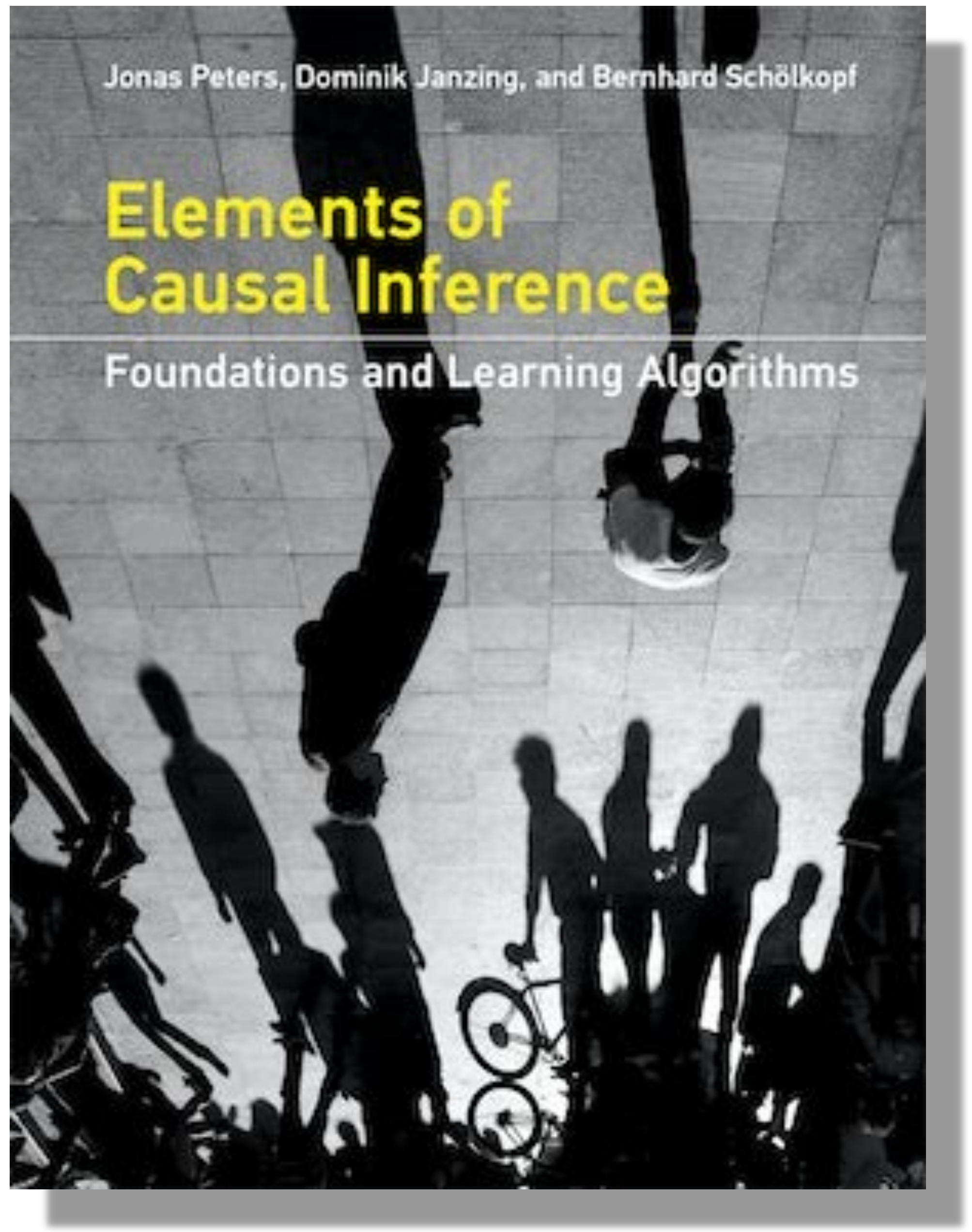
  - Identifiability: if there is ANM from X to Y,

    in general, there is no ANM from Y to X

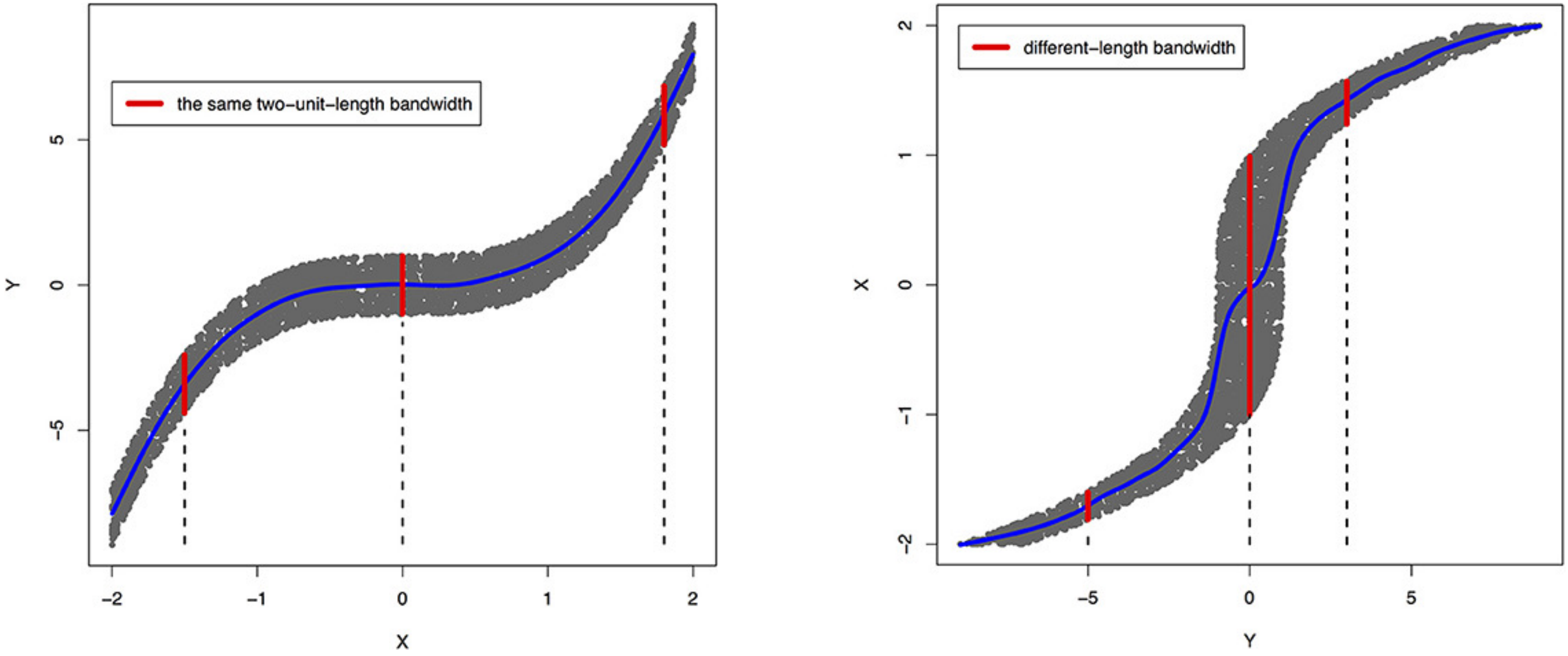
Peters et al: Causal Discovery with Continuous Additive Noise Models, JMLR 2014

• Additive Noise Models: Illustration





- # Additive Noise Models: Illustration



- # Additive Noise Models: Identifiability

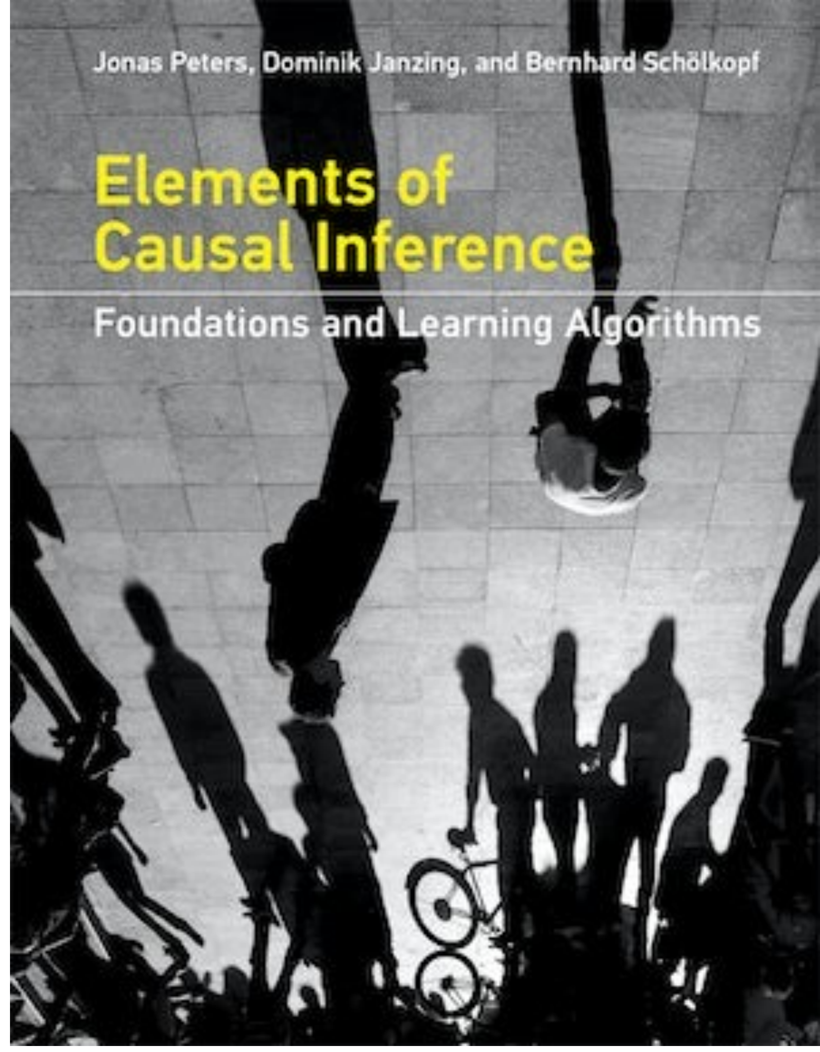
Theorem 1 (Identifiability of ANMs) *For the purpose of this theorem, let us call the ANM smooth if N_Y and X have strictly positive densities p_{N_Y} and p_X and f_Y, p_{N_Y}, and p_X are three times differentiable.*
Assume that $P_{Y|X}$ admits a smooth ANM from X to Y, and there exists a $y \in \mathbb{R}$ such that

$$(\log p_{N_Y})''(y - f_Y(x)) f_Y'(x) \neq 0$$

for all but countably many values x. Then, the set of log densities $\log p_X$ for which the obtained joint distribution $P_{X,Y}$ admits a smooth ANM from Y to X is contained in a 3-dimensional affine space.

(Hoyer, Janzing, Mooij, Peters, Schölkopf, 2008)

• Additive Noise Models: Identifiability



- For linear non-Gaussian ANM (LiNGAM) it is simpler.

Theorem 4.2 (Identifiability of linear non-Gaussian models) *Assume that $P_{X,Y}$ admits the linear model*

$$Y = \alpha X + N_Y, \quad N_Y \perp\!\!\!\perp X, \tag{4.1}$$

with continuous random variables X, N_Y, and Y. Then there exist $\beta \in \mathbb{R}$ and a random variable N_X such that

$$X = \beta Y + N_X, \quad N_X \perp\!\!\!\perp Y, \tag{4.2}$$

if and only if N_Y and X are Gaussian.

- Proof hinges on a classical results for Gaussians:
  Kac-Bernstein (1939) and Darmois-Skitovic theorems (1953, 1954)

- Closely related to independent component analysis (ICA)

(Shimizu et al, 2006)

Additive Noise Models: How to Apply

- Perform regression of Y on X: estimate $\hat{f}(x) \simeq \mathbb{E}[Y|X=x]$
- Compute resulting residual/noise: $\hat{N}_Y = Y - \hat{f}(X)$
- Perform regression of X on Y: estimate $\hat{g}(y) = \mathbb{E}[X|Y=y]$
- Compute resulting residual/noise: $\hat{N}_X = X - \hat{g}(Y)$
- Select $X \to Y$ if N_Y is more independent of X, than N_X is of Y; e.g., using mutual information:

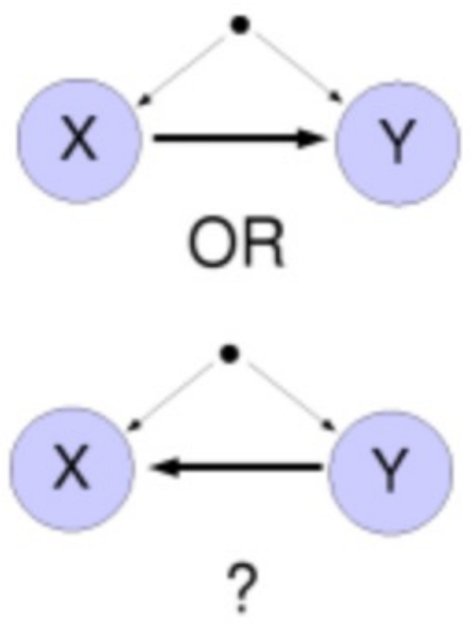
$$I(N_Y;X) \overset{?}{\lesseqgtr} I(N_X;Y)$$

Database with cause-effect pa ×

webdav.tuebingen.mpg.de/cause-effect/

Apps Bookmarks Dicionário Pribera... Boyd on writing Electronic library.... Colibri V3 - Video... ScholarOne TCI Other Bookmarks

Database with cause-effect pairs



This is a growing database with different data for testing causal detection algorithms. The goal here is to distinguish between cause and effect. We searched for data sets with known ground truth. However, we do not guarantee that all provided ground truths are correct. The datafiles are .txt-files and contain two variables, one is the cause and the other the effect. For every example there exists a description file where you can find the ground truth and how the data was derived.

Note that not always the first column is the cause and the second the effect. This is indicated in a meta-data file. Please look at README for further explanations. We also suggest a weighting factor for some pairs which are very similar if you want to calculate the overall performance.

To get all data files at once download all data as a zip file.

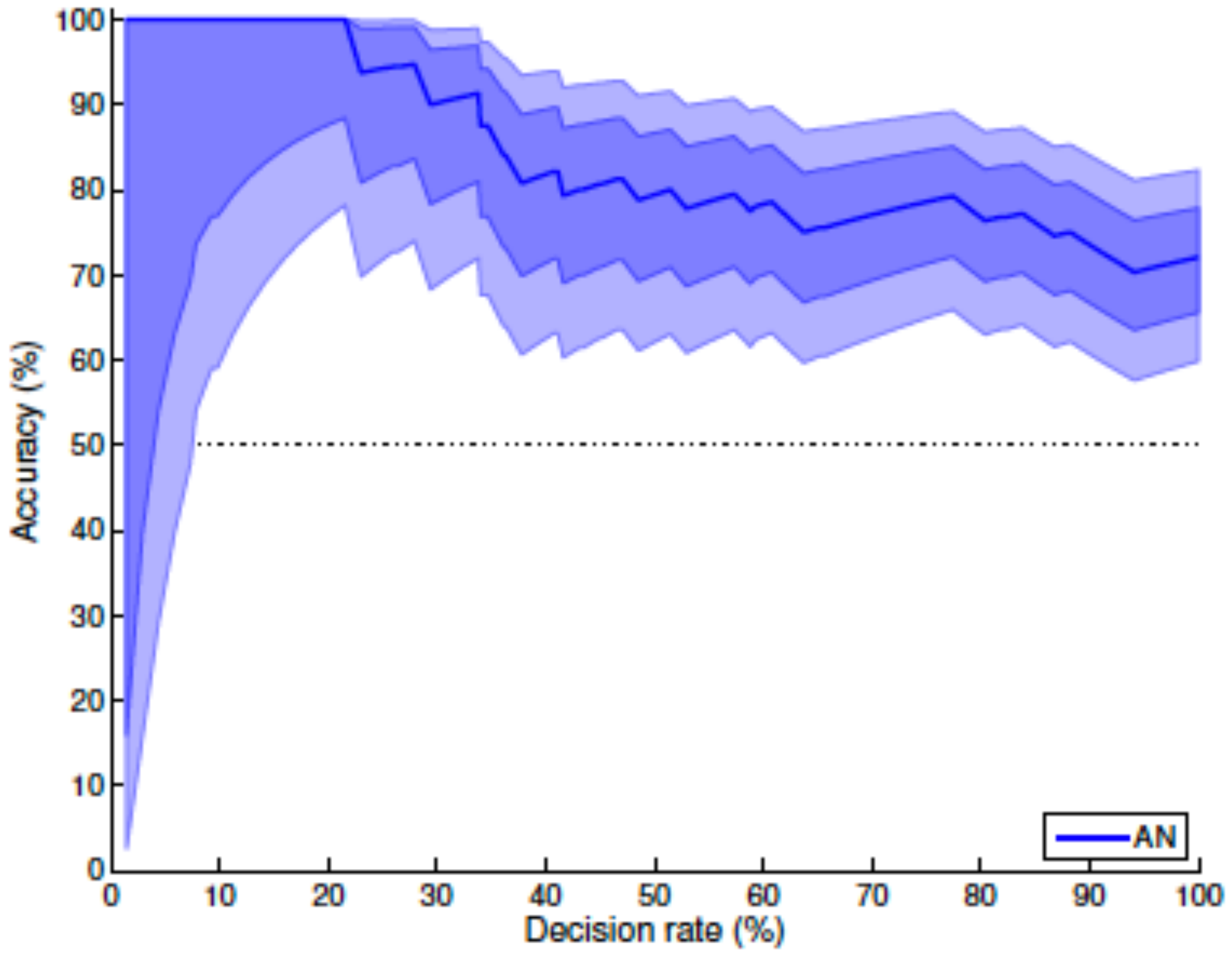
When you use this data set in a publication, please cite the following paper (which also contains much more detailed information regarding this data set in the supplement):

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schoelkopf:
"Distinguishing cause from effect using observational data: methods and benchmarks",
Journal of Machine Learning Research 17(32):1-102, 2016

• Additive Noise Models: How to Apply

- Practical aspects: how to do the regressions?
- How to measure independence?

Peters, Mooij, Janzing, Schölkopf: Causal Discovery with Continuous Additive Noise Models, JMLR 20014



- Extended to discrete data (in $\mathbb{Z}$) by Peters, Janzing, Scholkopf (2011)

• Extending to Categorical Variables

- For categorical variables, no addition is defined; no ANM
- ANM for continuous variables satisfy:

  Proposition 4 *If real-valued variables X and Y admit an ANM from X to Y, then the conditional differential entropy $h(Y|X) = h(N_Y)$, independently of the distribution of X.*

- Agrees with independence of cause and mechanism.
- Can we do the same for categorical variables?

Proceedings of Machine Learning Research vol TBD:1–20, 2023 2nd Conference on Causal Learning and Reasoning

Distinguishing Cause from Effect on Categorical Data: The Uniform Channel Model

Mário A. T. Figueiredo, MARIO.FIGUEIREDO@TECNICO.ULISBOA.PT
Catarina Oliveira CATARINA.A.OLIVEIRA@TECNICO.ULISBOA.PT
Instituto de Telecomunicações and LUMLIS (Lisbon ELLIS Unit),
Instituto Superior Técnico, Universidade de Lisboa, Portugal

Uniform Channel Model (UCM)

- Based on the analogy with a communication channel

- Channel (stochastic) matrix: $\theta^{X\to Y}_{x,y} = \mathbb{P}(Y = y | X = x)$
- UCM: rows of $\theta^{X\to Y}$ are permutations of each other
- Agreement with ICM:

Proposition 5 *If $\theta^{X\to Y}$ corresponds to a UC (each row of $\theta^{X\to Y}$ is a permutation of a vector $\gamma \in \Delta_{|\mathcal{Y}|-1}$), then the conditional entropy $H(Y|X) = H(\gamma)$, independently of p_X.*

• UCM as a Structural Causal Model

- Arbitrary marginal $\mathbb{P}(X = x)$
- UCM conditional: $\theta_{x,y}^{X \to Y} = \mathbb{P}(Y = y | X = x)$
- The corresponding joint $\mathbb{P}(X = x, Y = y)$ is entailed by SCM

$$Y \leftarrow f(X, U_Y) \qquad U_Y \perp\!\!\!\perp X$$

with U_Y taking values in the same set as Y

- If the conditional is not a UCM, such an SCM is not possible.

UCM: Identifiability (Binary Case)

- Binary UCM (binary symmetric channel) $\boldsymbol{\theta}^{X \to Y} = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}.$
- Marginal: $\mathbb{P}(X = 1) = \beta$
- Reverse channel (from Bayes law):

$$\boldsymbol{\theta}^{Y \to X} = \begin{bmatrix} \frac{(1-\alpha)\beta}{(1-\alpha)\beta + \alpha(1-\beta)} & \frac{\alpha(1-\beta)}{(1-\alpha)\beta + \alpha(1-\beta)} \\ \frac{\alpha\beta}{\alpha\beta + (1-\alpha)(1-\beta)} & \frac{(1-\alpha)(1-\beta)}{\alpha\beta + (1-\alpha)(1-\beta)} \end{bmatrix}$$

- Conditions for $\boldsymbol{\theta}^{Y \to X}$ being UCM have zero measure:

$$\{(\alpha, \beta) \in [0,1]^2 \, : \, \alpha = 0 \vee \alpha = 1/2 \vee \alpha = 1 \vee \beta = 0 \vee \beta = 1/2 \vee \beta = 1\}$$

• UCM: Identifiability (General Case)

Theorem 9 *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two categorical random variables with a joint pmf such that the conditional $\boldsymbol{\theta}^{X \to Y}$ corresponds to a UC. Assume also that the marginals have full support*[5]*: $p_Y(y) \neq 0$, for any $y \in \mathcal{Y}$, and $p_X(x) \neq 0$, for any $x \in \mathcal{X}$. Further assume that the rows of the channel matrix $\boldsymbol{\theta}^{X \to Y}$ are not all equal to each other (i.e., X and Y are not independent*[6]*). Then, the set of parameters such that the reverse channel $\boldsymbol{\theta}^{Y \to X}$ is also a UCM has zero Lebesgue measure.*

UCM causal inference principle for categorical variables: given two categorical variables X and Y, if the conditional pmf $\boldsymbol{\theta}^{X \to Y}$ corresponds to a UCM, but the conditional pmf $\boldsymbol{\theta}^{Y \to X}$ does not, then we infer the causal direction to be $X \to Y$.

• Applying UCM to Data: Channel Estimates

- Independent samples: $(x_1, y_1), \ldots, (x_N, y_N)$
- Count matrix: $N_{x,y} =$ number of samples s.t. $x_i = x \wedge y_i = y$
- Matrix estimate, without constraint: $\hat{\theta}_{x,y} = \frac{N_{x,y}}{N_x} = \frac{N_{x,y}}{\sum_y N_{x,y}}$
- Matrix estimate, with UCM constraint:

Sort each row of $N_{x,y}$: $\hat{\tau}_x$ *is such that* $N_{x,\hat{\tau}_x(1)} \geq \cdots \geq N_{x,\hat{\tau}_x(|\mathcal{Y}|)}$

Compute: $\hat{\gamma}_y = \frac{1}{N} \sum_{x \in \mathcal{X}} N_{x,\hat{\tau}_x(y)}$ and $\hat{\theta}^{X \to Y}_{x,y} = \hat{\gamma}_{\hat{\sigma}_x(y)}$

$\hat{\sigma}_x = \hat{\tau}_x^{-1}$

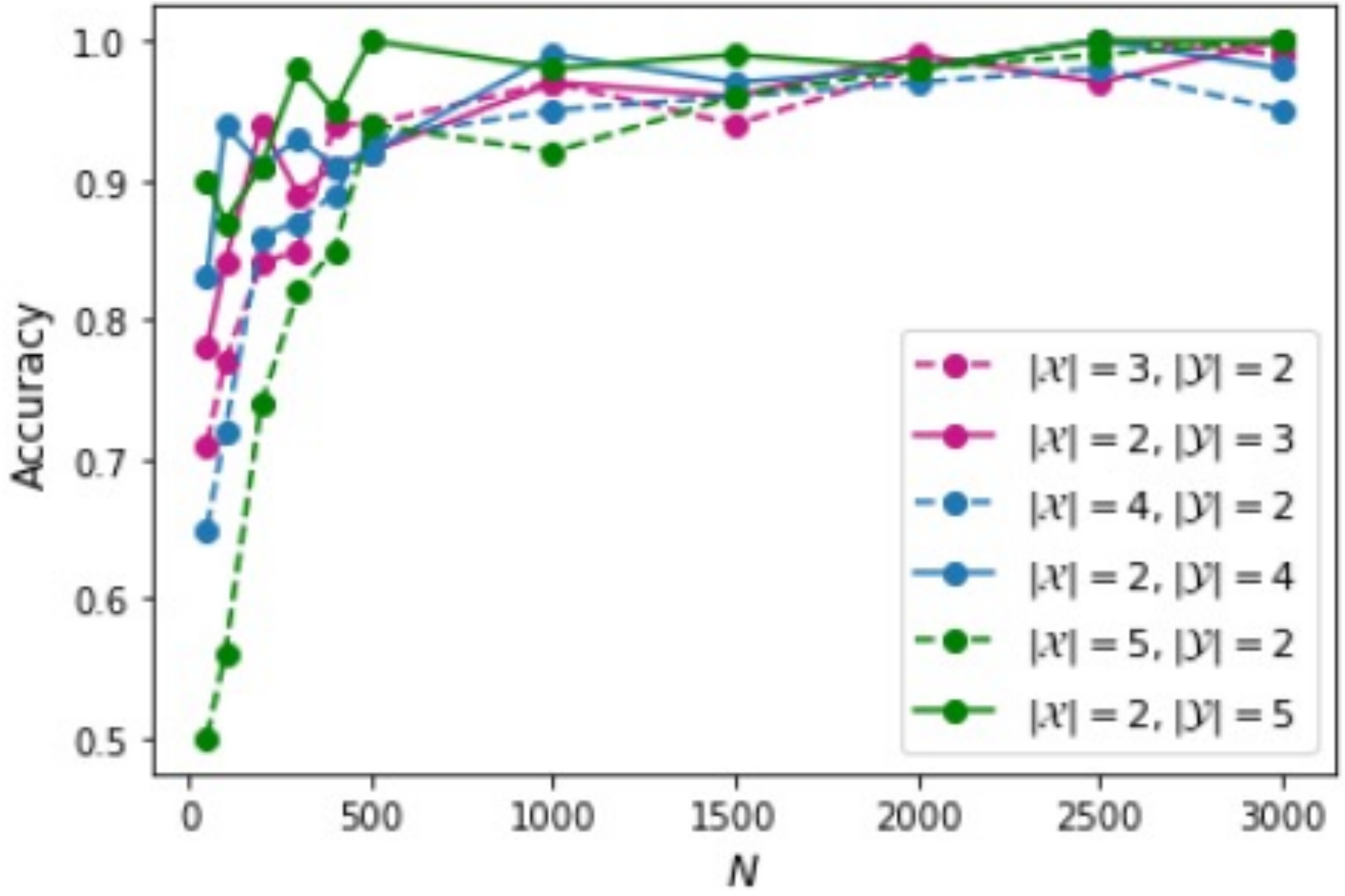
• Applying UCM to Data: Criterion

- Statistical tests: likelihood ratio tests (LRT) for UCM vs non-UCM
- Let $\mathrm{p}^{X \to Y}$ and $\mathrm{p}^{Y \to X}$ be p-values for RLT in both directions.
- Choose some significance threshold α (e.g., 0.05)

  - If $\mathrm{p}^{X \to Y} \geq \alpha$ and $\mathrm{p}^{Y \to X} < \alpha$, declare $X \to Y$.
  - If $\mathrm{p}^{X \to Y} < \alpha$ and $\mathrm{p}^{Y \to X} \geq \alpha$, declare $Y \to X$.
  - If $\mathrm{p}^{X \to Y} < \alpha$ and $\mathrm{p}^{Y \to X} < \alpha$, declare "undecided: wrong model".
  - If $\mathrm{p}^{X \to Y} \geq \alpha$ and $\mathrm{p}^{Y \to X} \geq \alpha$, declare "undecided: both directions possible".

• UCM: Synthetic Experiments

- Random uniform channels $X \to Y$
- 100 datasets for each configuration of cardinalities



• UCM on Benchmark Data

- 112 cause-effect pairs with categorical variables
- Comparison with:
  DC (distance correlation) by Liu and Chan, 2016.
  HCR (hidden compact representation) by Cai et al. 2018.
- Average accuracy (notice that random choice yields 1/3 accuracy)

UCM	DC	HCR
0.61	0.41	0. 47

• UCM on Real Data

Table 2: Results on real data. Wrong decisions are shown in red; UWM stands for "undecided: wrong model". Month is a cyclic variable, thus a CUC was used in the $Y \to X$ direction.

Dataset	X	Y	UCM	DC	HCR
Adult	Occupation	Income	UWM	$X \to Y$	$X \to Y$
Adult	Work Class	Income	UWM	$X \to Y$	$X \to Y$
Acute Inflammation	Inflam. of urinary bladder	Lumbar pain	$Y \to X$	Inconcl.	Inconcl.
Acute Inflammation	Inflam. of urinary bladder	Nausea	$Y \to X$	Inconcl.	Inconcl.
Acute Inflammation	Inflam. of urinary bladder	Burning urethra	$Y \to X$	Inconcl.	Inconcl.
Pittsburgh Bridges	Material	Lanes	$X \to Y$	$Y \to X$	$X \to Y$
Pittsburgh Bridges	Purpose	Type	UWM	$Y \to X$	$X \to Y$
Temperature	Month	Temperature	$X \to Y$	$X \to Y$	$Y \to X$
Horse Colic	Abdomen Status	Surgical Lesion	UWM	$X \to Y$	$X \to Y$

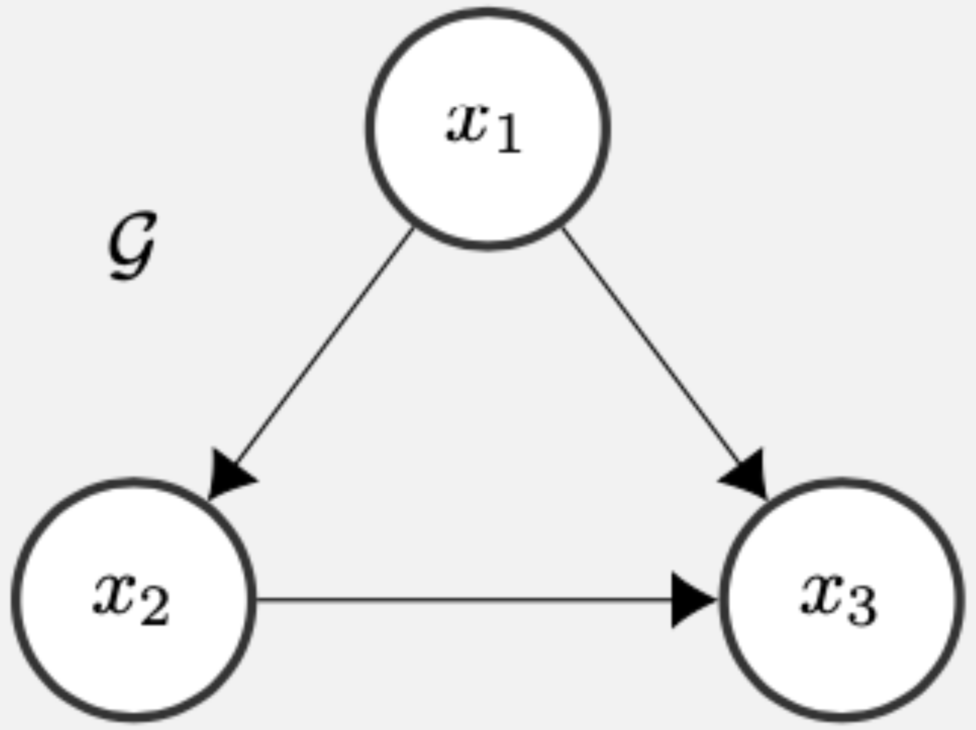
Proceedings of Machine Learning Research vol 140:1–22, 2022 1st Conference on Causal Learning and Reasoning

Differentiable Causal Discovery Under Latent Interventions

Gonçalo R. A. Faria[*†] GONCALORAFARIA@TECNICO.ULISBOA.PT
André F. T. Martins [*†‡] ANDRE.T.MARTINS@TECNICO.ULISBOA.PT
Mário A. T. Figueiredo [*†] MARIO.FIGUEIREDO@TECNICO.ULISBOA.PT
[*] *Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit), Universidade de Lisboa, Portugal.*
[†] *Instituto de Telecomunicações, Lisboa, Portugal.*
[‡] *Unbabel, Lisboa, Portugal.*

Editors: Bernhard Schölkopf, Caroline Uhler and Kun Zhang

• Score-based methods



$$\begin{cases} x_1 & :=\mathrm{NN}(\mathcal{G}_{11}x_1, \mathcal{G}_{12}x_2, \mathcal{G}_{13}x_3, \epsilon_1; \theta_1) \\ x_2 & :=\mathrm{NN}(\mathcal{G}_{21}x_1, \mathcal{G}_{22}x_2, \mathcal{G}_{23}x_3, \epsilon_2; \theta_2) \\ x_3 & :=\mathrm{NN}(\mathcal{G}_{31}x_1, \mathcal{G}_{32}x_2, \mathcal{G}_{33}x_3, \epsilon_3; \theta_3) \end{cases}$$

Train the NNs and estimate $\mathcal{G}_{ij}$ such that graph is DAG

Similar to LASSO (sparsity) for all of the variables at once w/ acyclicity constraints.

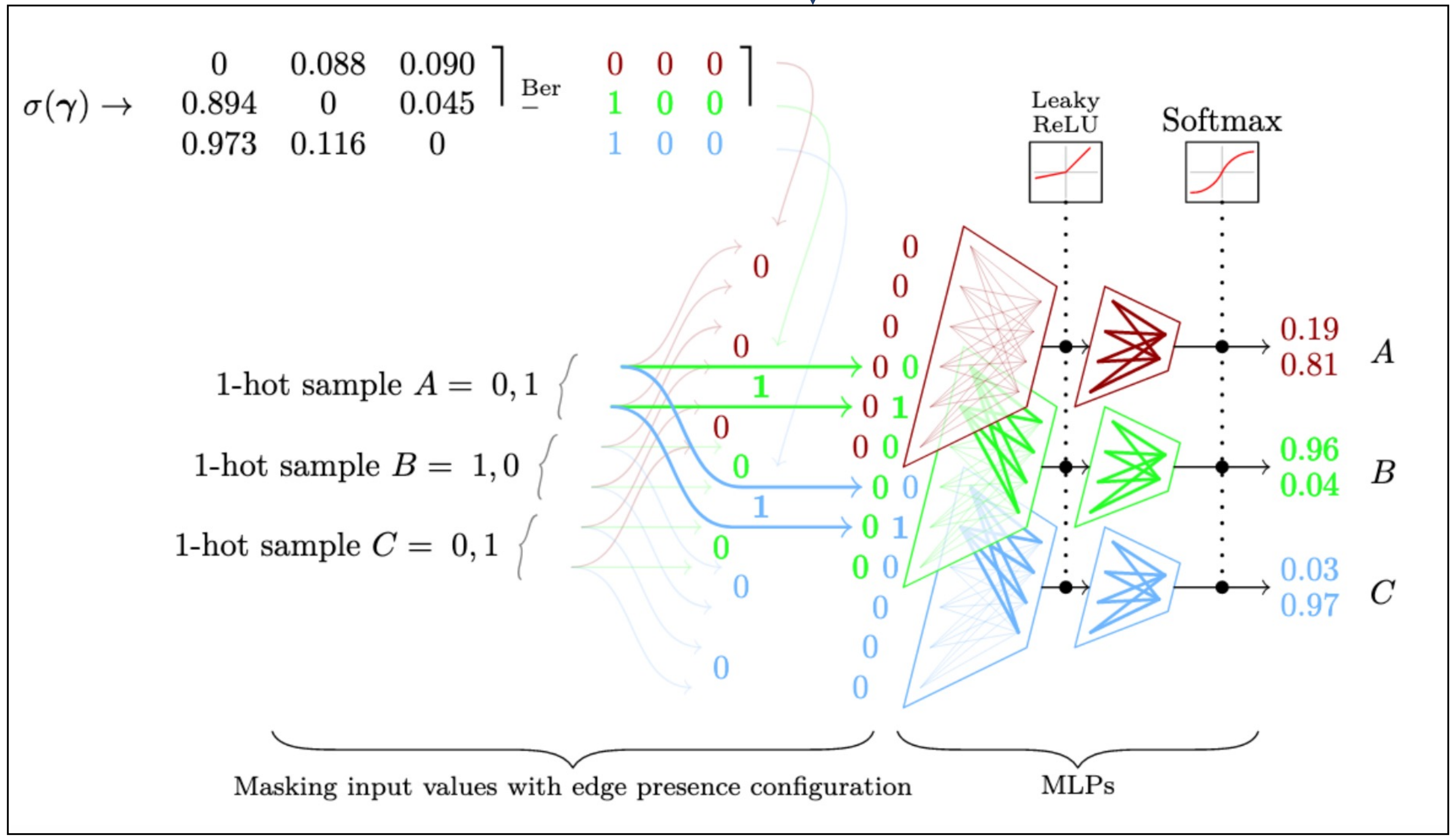
Ke et. al. 2020

• Score-Based Methods with Deep Learning

$$\Lambda^* = \arg\max_{\Lambda} S(\Lambda)$$

$$S(\Lambda) = \max_{\theta} \mathbb{E}_{\mathcal{G}\sim \mathrm{DAG}(\mathcal{G};\Lambda)} \left[\log p(\mathcal{D}|\mathcal{G},\theta) + \log p(\mathcal{G}) \right]$$

(most of the time)
approximated w/ fully
factorized Bernoulli



Penalize dense
graphs/enforce
acyclicity

Ke et. al. 2020

• Causal Discovery with Interventional Data

$$\Lambda^*, \tilde{\mathcal{M}}^* = \arg\max_{\Lambda, \tilde{\mathcal{M}}} S(\Lambda, \tilde{\mathcal{M}})$$

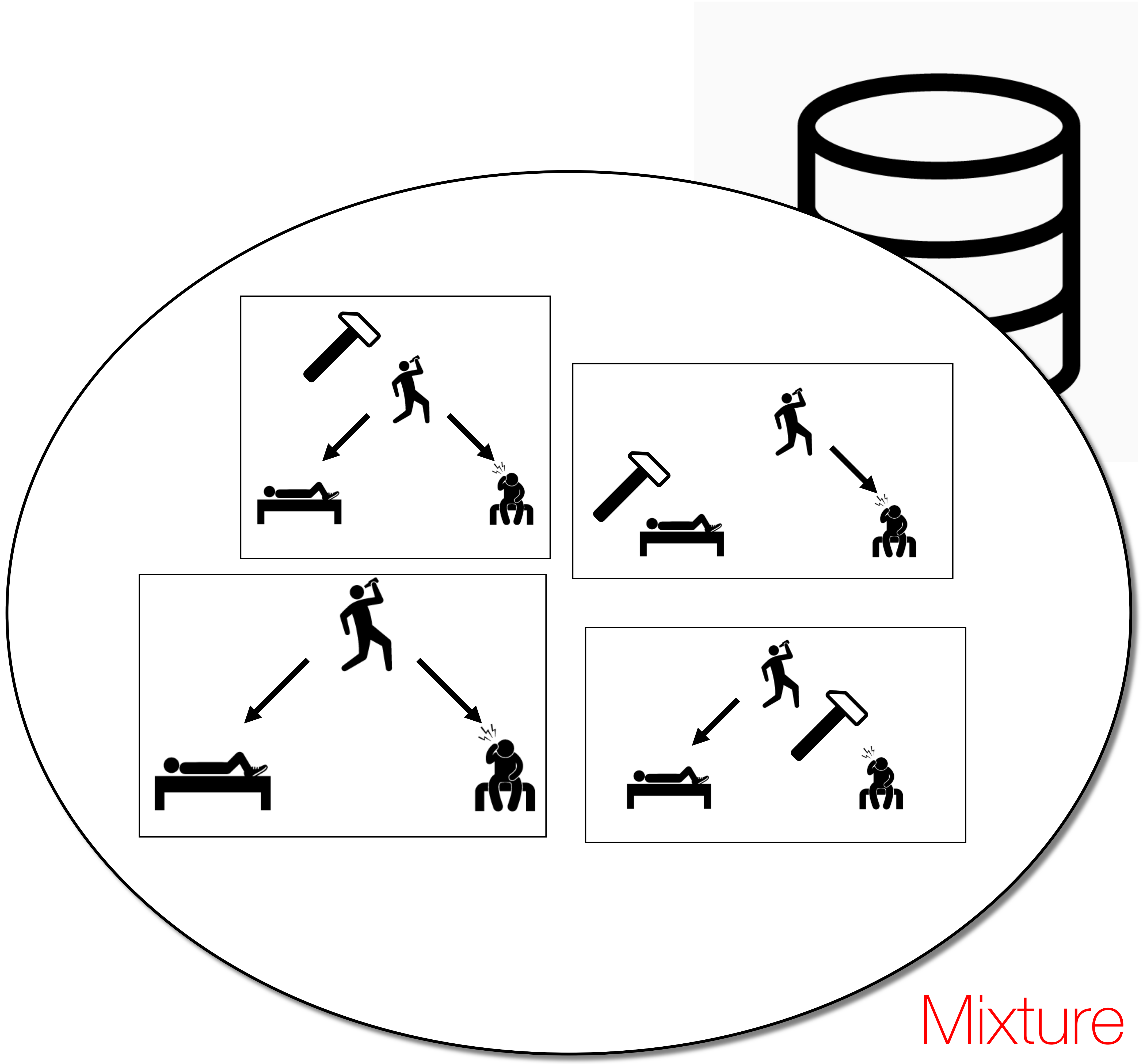
$$S(\Lambda, \tilde{\mathcal{M}}) = \max_{\theta,\phi} \mathbb{E}_{\mathcal{G}\sim \mathrm{DAG}(\mathcal{G};\Lambda)}\Big[\mathbb{E}_{k\sim p(k)}\big[\log p(\mathcal{D}^k|\mathcal{G}, \theta^k, \tilde{\mathcal{M}}^{(k)}) + \log p(\mathcal{G})\big]\Big]$$

Intervention specific parameters

Intervention variables

Brouillard et. al. 2020

• Latent interventions



For each sample:

- do not know correspondence to intervention regime;

For each intervention:

- do not know experimental conditions

Mixture of experimental regimes.

• Intervention Recovery

- Given the causal graph, ...
- recover interventions and correspondences,
- propose joint distribution.

Infinite mixture of intervention SCMs + **Prior distribution for the interventions** (Dirichlet process)

- # Approximate Posterior Inference

Dataset

Graph

$$\boxed{p(z^{(i)}, \tilde{\mathcal{M}} | \mathcal{D}, \mathcal{G}; \theta) = p(z^{(i)} | \tilde{\mathcal{M}}, x^{(i)}, \mathcal{G}; \theta) p(\tilde{\mathcal{M}} | \mathcal{D}, \mathcal{G}; \theta)}$$

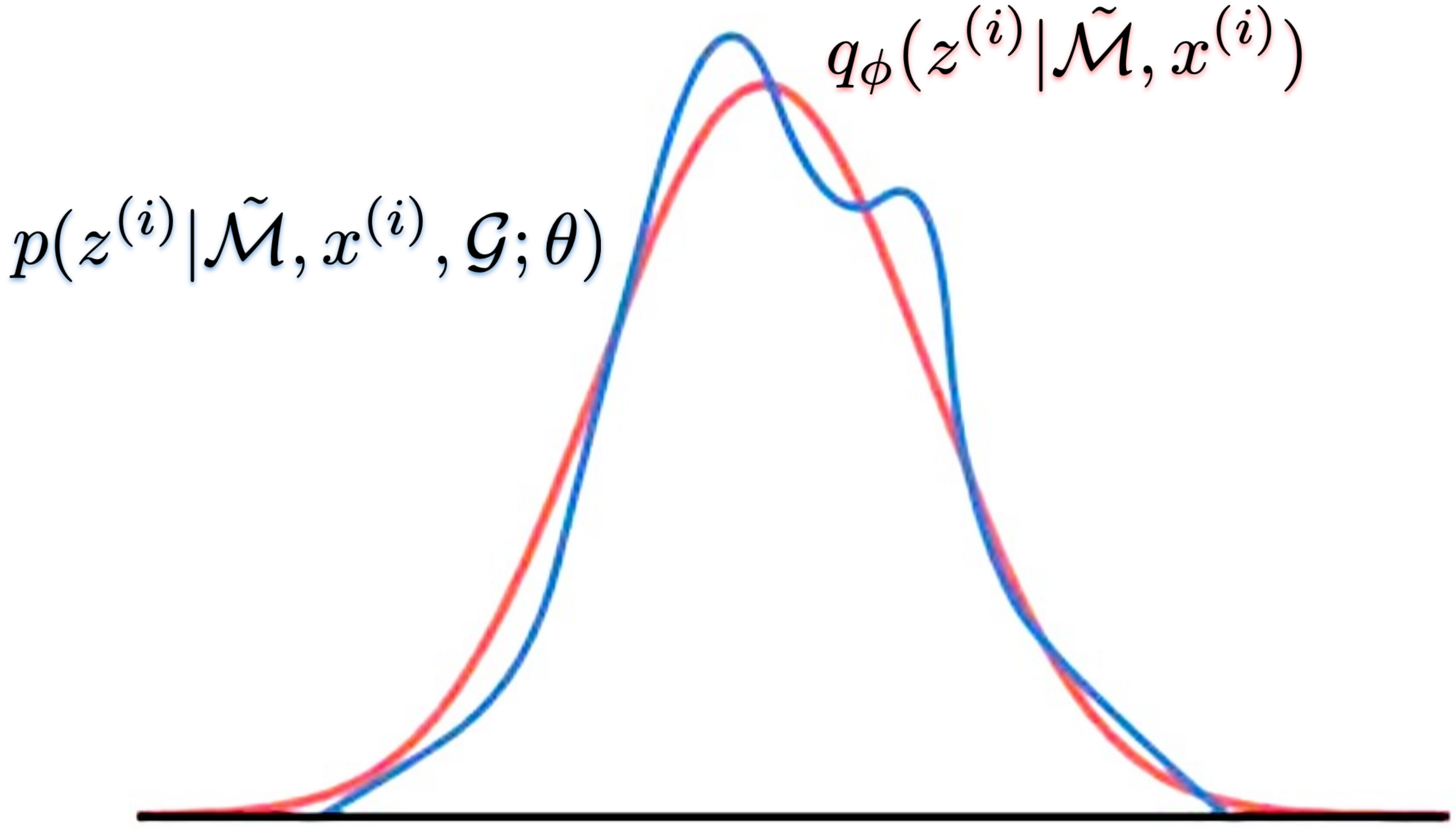
Correspondence

Latent interventions

Sample

• Approximate Posterior Inference

- Variational inference

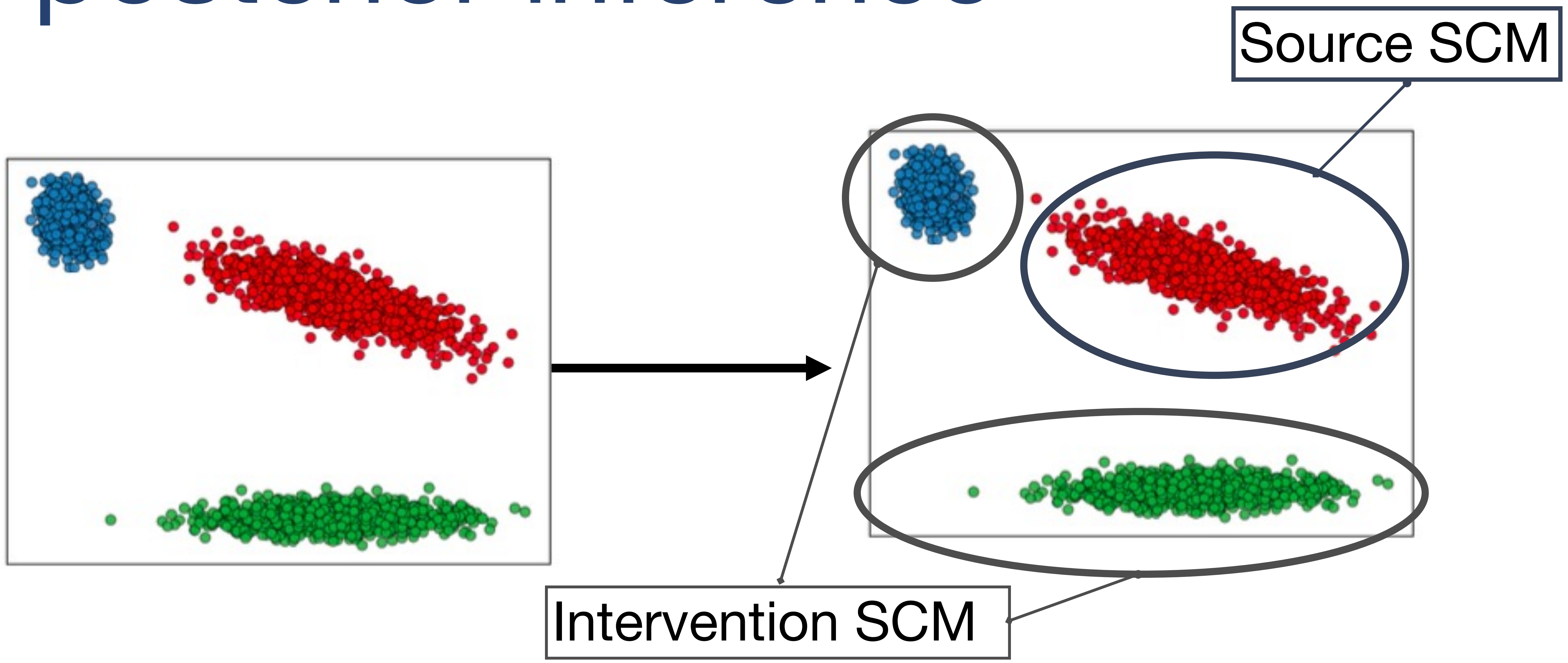


• Approximate posterior inference

- Variational inference

  Similar to a clustering problem

(each cluster has few degrees of freedom)



highly overlapping clusters

- Evidence lower bound (ELBO):

$$\max_{\theta,\phi} \sum_{i=1}^{N} \mathrm{ELBO}_{q(z^{(i)},\tilde{\mathcal{M}};\phi)}\left(x^{(i)},\mathcal{G};\theta\right) \text{ -}$$

- Score-Based Method

$$\Lambda^* = \arg\max_{\Lambda} S(\Lambda)$$

$$S(\Lambda) = \max_{\theta,\phi} \mathbb{E}_{\mathcal{G}\sim\mathrm{DAG}(\mathcal{G};\Lambda)}\Big[\sum_{i=1}^{N} \mathrm{ELBO}_{q(z^{(i)},\tilde{\mathcal{M}};\phi)}\big(x^{(i)},\mathcal{G};\theta\big) + \log p(\mathcal{G})\Big]$$

• Model Variants

- **“latent”** (new)

For each sample:

- do not know correspondence to intervention regime;

For each intervention:

- do not know experimental conditions

- Model variants

- “latent” (new)
- **“unknown”**

For each sample:

- ~~do not~~ know correspondence to intervention regime;

For each intervention:

- do not know experimental conditions

• Model variants

- “latent” (new)
- “unknown”
- **“known”**

For each sample:

- ~~do not~~ know correspondence to intervention regime;

For each intervention:

- ~~do not~~ know experimental conditions

• Model variants

- “latent” (new)
- “unknown”
- “known”
- **“observational”**

assume there is no intervention data.

- Model variants

  - “latent” (new)
  - “unknown”
  - “known”
  - “observational”
  - **“semi-supervised”** (new)

• Model variants

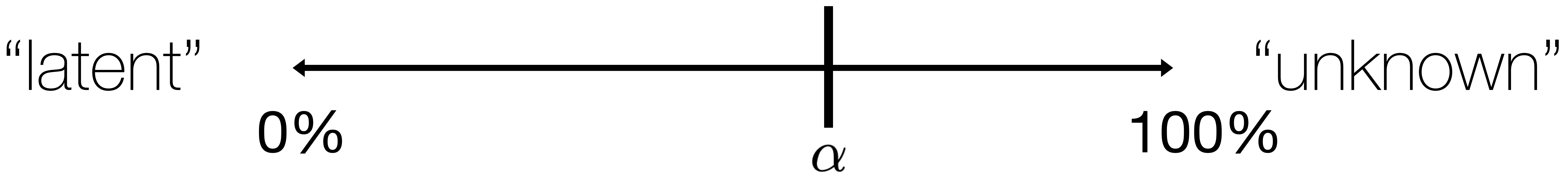
- “latent” (new)
- “unknown”
- “known”
- “observational”
- **“semi-supervised”** (new)

For each sample:

- (fraction) know correspondence to intervention regime;

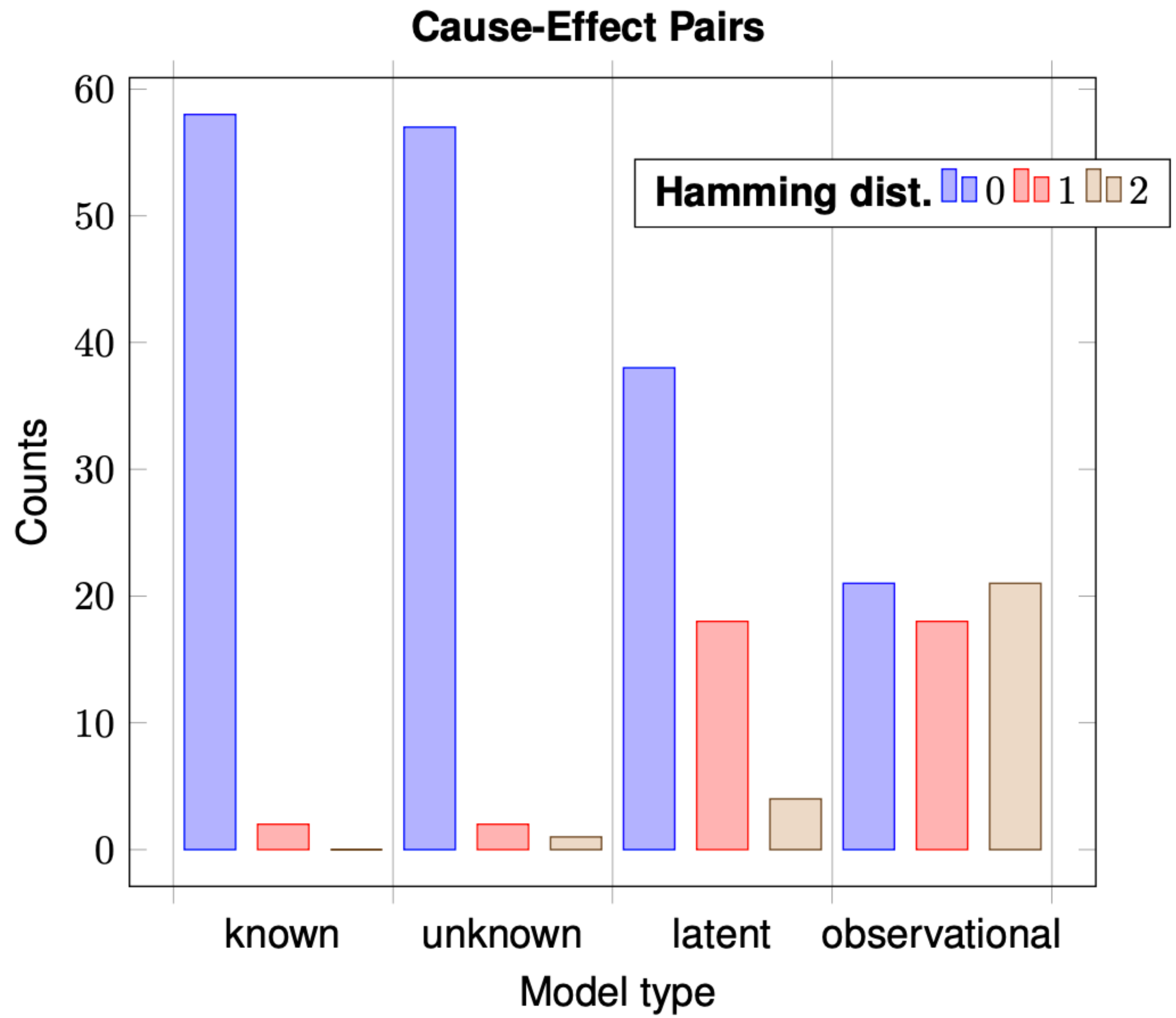
For each intervention:

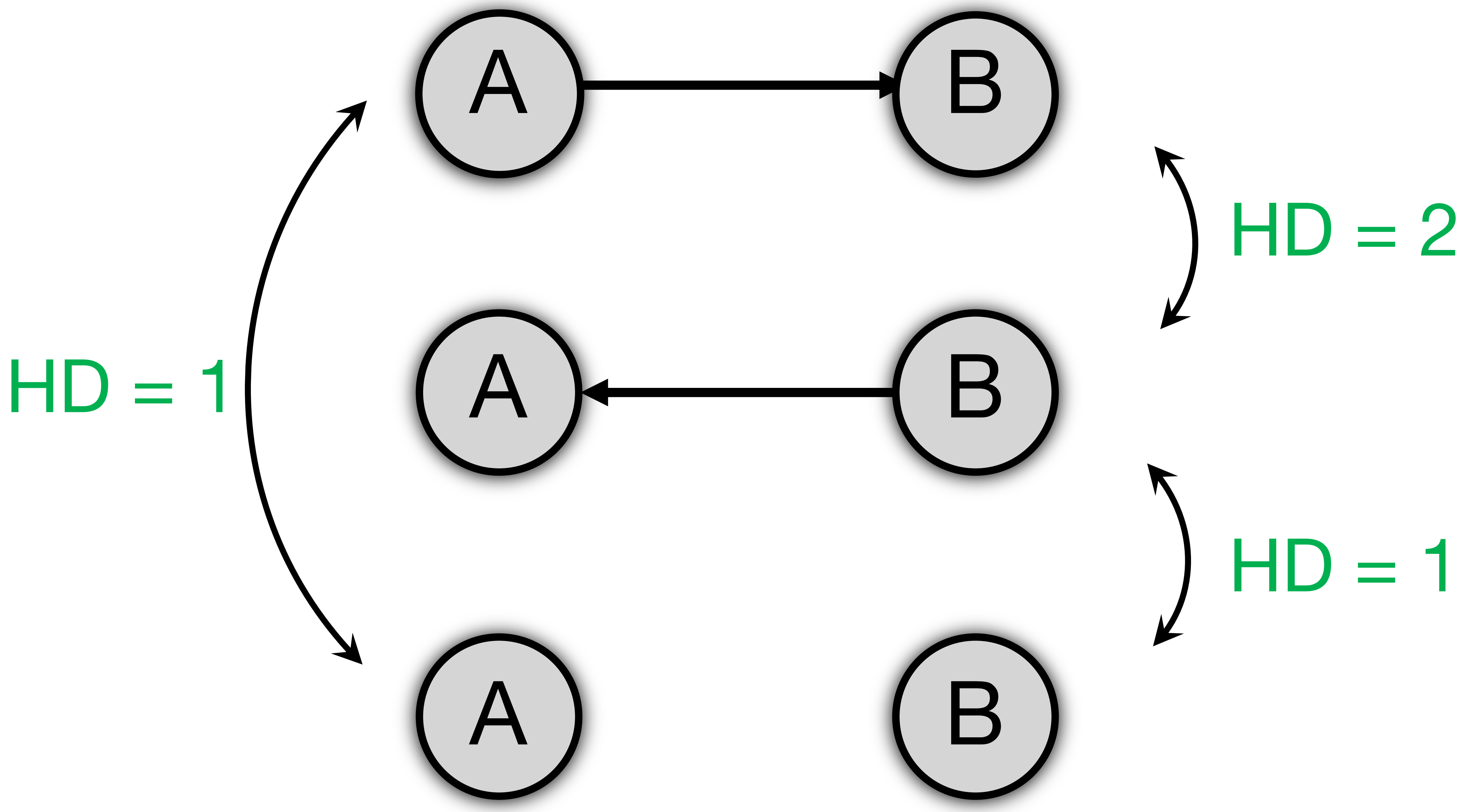
- do not know experimental conditions



Fraction of samples with correspondences.

- Experiments: Cause-Effect Pairs

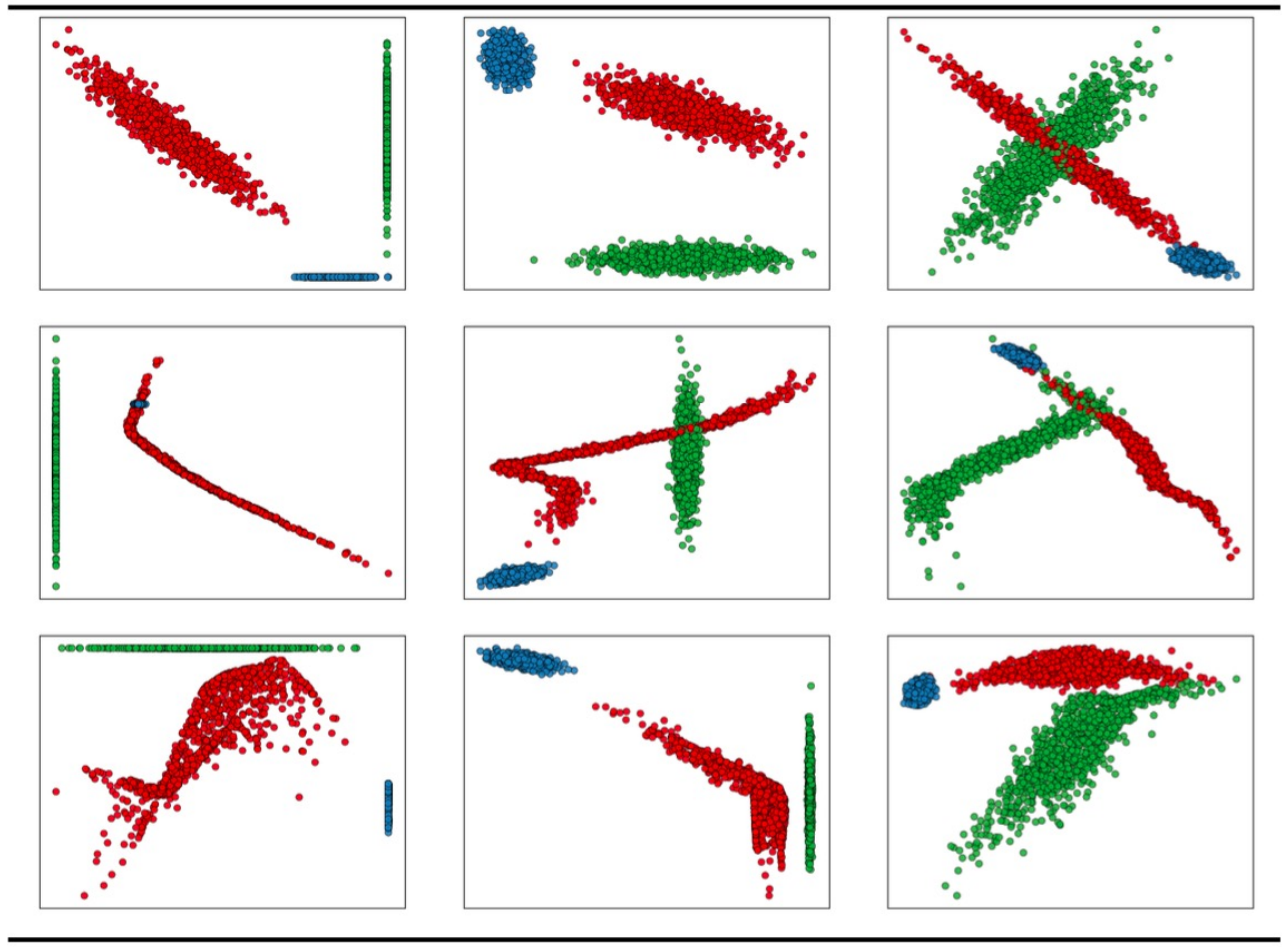




• Experiments on synthetic data

Model Type	e	**latent**	**unknown**	**known**	**observational**
		Stochastic Interventions:			
Linear Gaussian		5.9 ± 6.2	3.4 ± 3.2	0.5 ± 1.3	10.3 ± 7.8
Non-Linear Gaussian	1	12.2 ± 3.9	10.3 ± 2.5	7.0 ± 3.6	13.7 ± 3.8
Non-Linear Non-Gaussian		8.7 ± 6.6	8.0 ± 2.7	6.6 ± 2.2	11.3 ± 5.0
Linear Gaussian		27.2 ± 6.2	24.1 ± 5.8	15.6 ± 6.0	39.6 ± 5.0
Non-Linear Gaussian	4	35.8 ± 3.8	30.3 ± 5.3	27.7 ± 4.3	37.5 ± 5.2
Non-Linear Non-Gaussian		36.1 ± 4.4	35.5 ± 8.1	31.5 ± 5.6	40.2 ± 6.9
		Imperfect Interventions:			
Linear Gaussian		5.8 ± 4.2	6.2 ± 3.06	4.7 ± 3.6	10.4 ± 2.9
Non-Linear Gaussian	1	9.3 ± 2.4	8.9 ± 2.5	7.8 ± 3.9	10.5 ± 2.8
Non-Linear Non-Gaussian		8.8 ± 3.0	9.1 ± 3.5	7.9 ± 1.4	11.5 ± 5.4
Linear Gaussian		35.9 ± 8.3	29.7 ± 5.6	17.7 ± 7.9	39.1 ± 9.1
Non-Linear Gaussian	4	32.1 ± 6.0	32.6 ± 5.8	32.8 ± 5.4	39.8 ± 9.3
Non-Linear Non-Gaussian		30.4 ± 12.2	30.2 ± 11.2	25.8 ± 3.9	36.7 ± 9.8

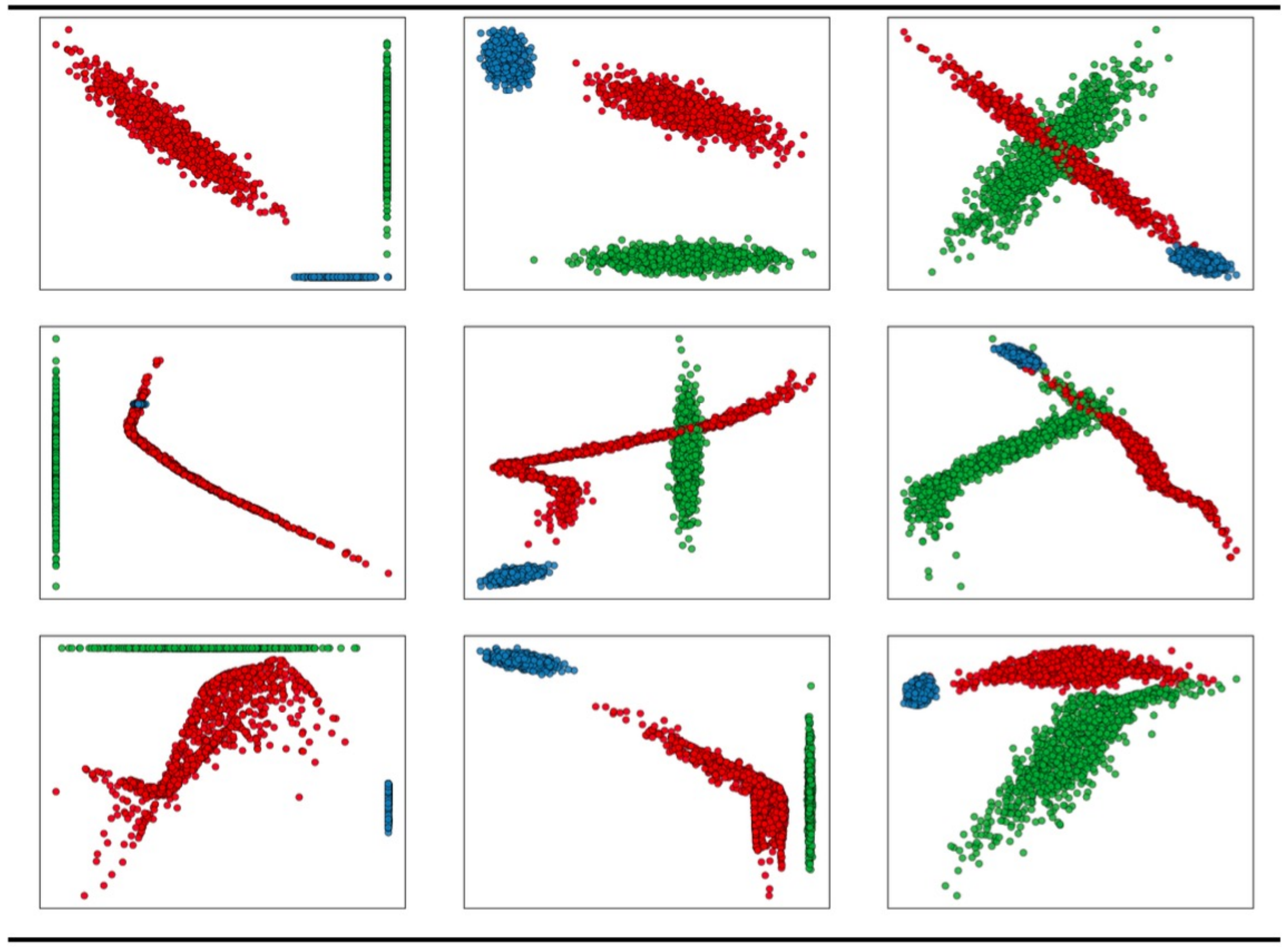
Table 4.1: Hamming distances on synthetic 10 variable SCMs.



- # Experiments on synthetic data

Model Type	e	latent	unknown	known	observational
		Stochastic Interventions:			
Linear Gaussian		5.9 ± 6.2	3.4 ± 3.2	0.5 ± 1.3	10.3 ± 7.8
Non-Linear Gaussian	1	12.2 ± 3.9	10.3 ± 2.5	7.0 ± 3.6	13.7 ± 3.8
Non-Linear Non-Gaussian		8.7 ± 6.6	8.0 ± 2.7	6.6 ± 2.2	11.3 ± 5.0
Linear Gaussian		27.2 ± 6.2	24.1 ± 5.8	15.6 ± 6.0	39.6 ± 5.0
Non-Linear Gaussian	4	35.8 ± 3.8	30.3 ± 5.3	27.7 ± 4.3	37.5 ± 5.2
Non-Linear Non-Gaussian		36.1 ± 4.4	35.5 ± 8.1	31.5 ± 5.6	40.2 ± 6.9
		Imperfect Interventions:			
Linear Gaussian		5.8 ± 4.2	6.2 ± 3.06	4.7 ± 3.6	10.4 ± 2.9
Non-Linear Gaussian	1	9.3 ± 2.4	8.9 ± 2.5	7.8 ± 3.9	10.5 ± 2.8
Non-Linear Non-Gaussian		8.8 ± 3.0	9.1 ± 3.5	7.9 ± 1.4	11.5 ± 5.4
Linear Gaussian		35.9 ± 8.3	29.7 ± 5.6	17.7 ± 7.9	39.1 ± 9.1
Non-Linear Gaussian	4	32.1 ± 6.0	32.6 ± 5.8	32.8 ± 5.4	39.8 ± 9.3
Non-Linear Non-Gaussian		30.4 ± 12.2	30.2 ± 11.2	25.8 ± 3.9	36.7 ± 9.8

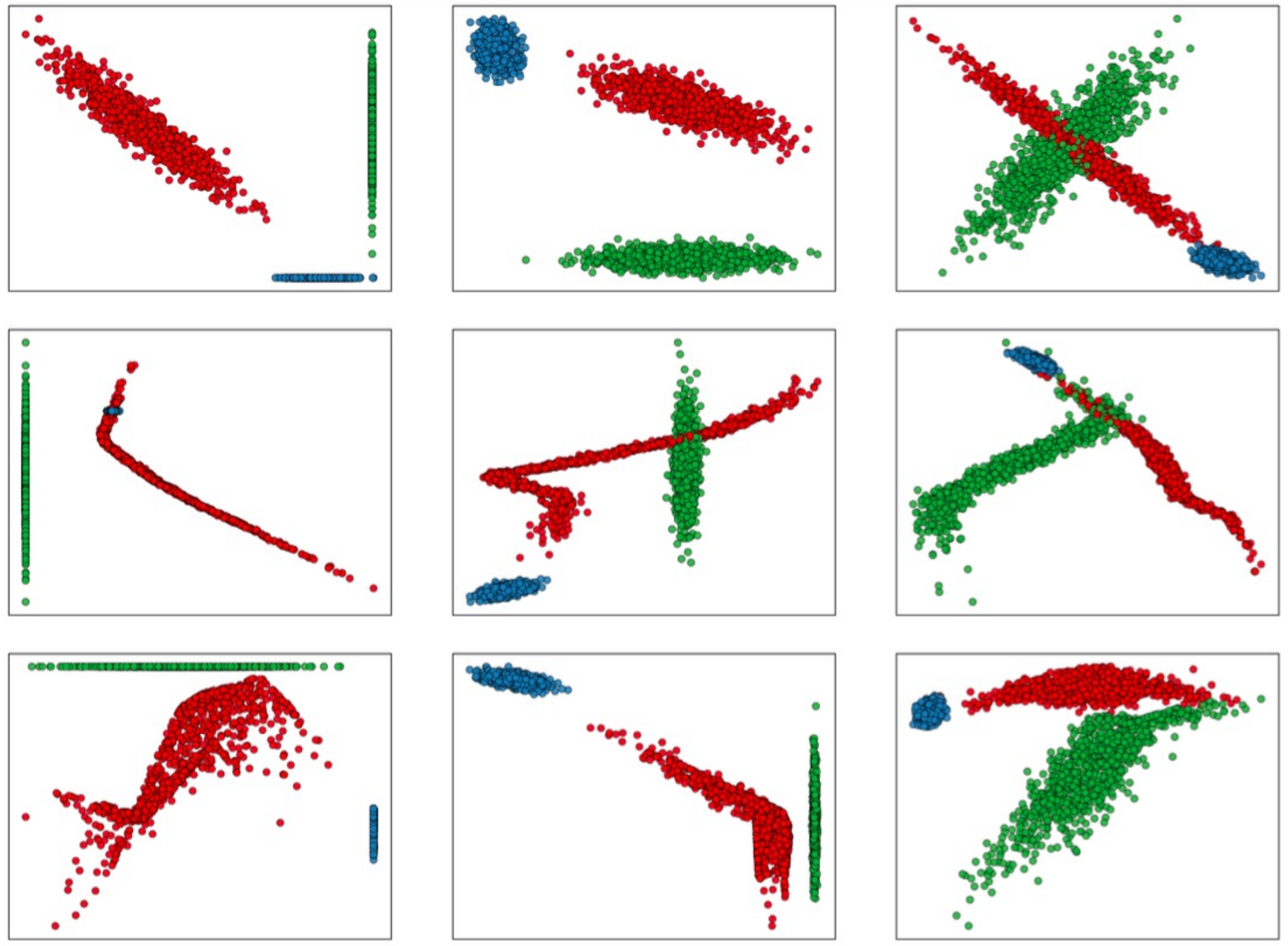
Table 4.1: Hamming distances on synthetic 10 variable SCMs.



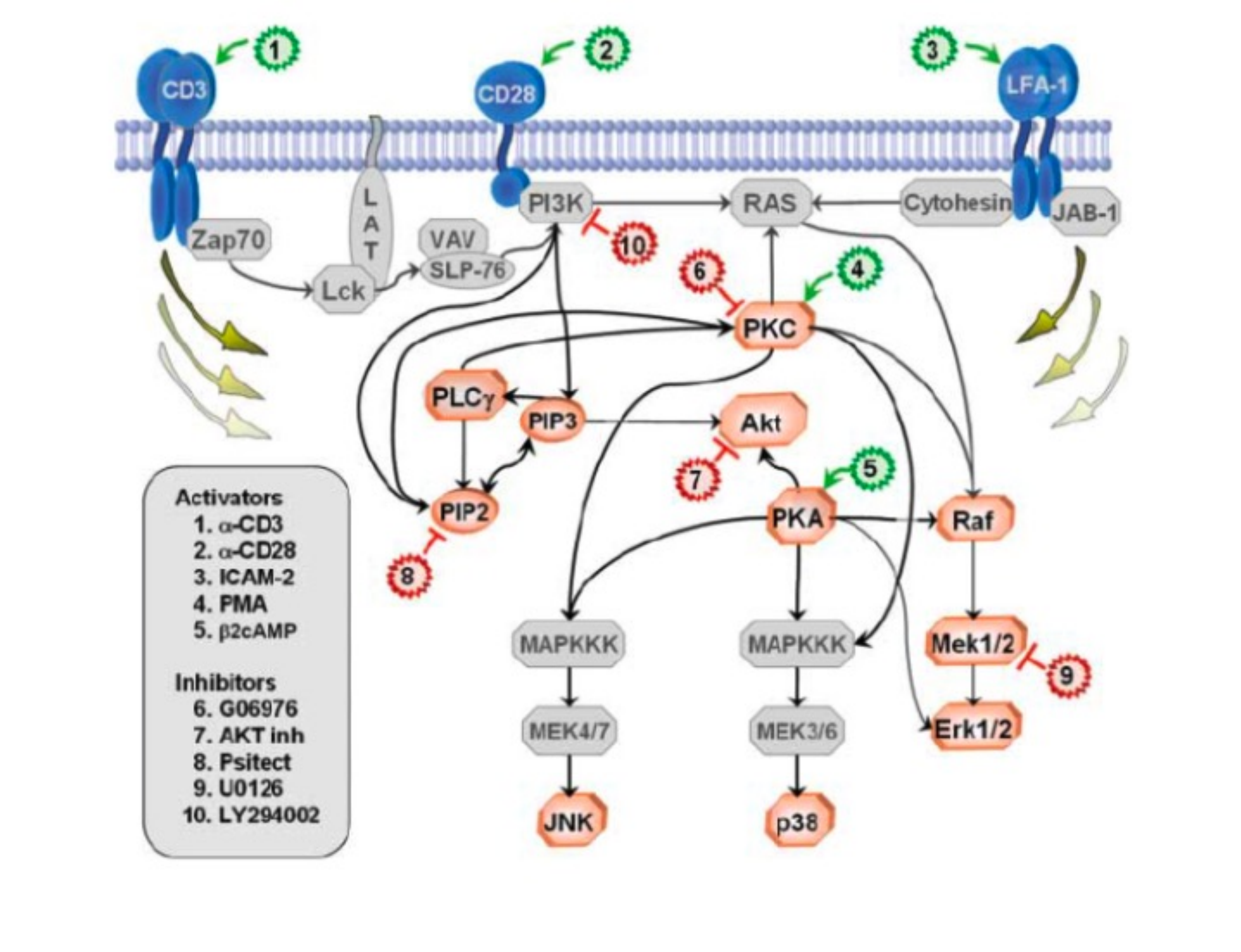
• Experiments on synthetic data

Model Type	e	**latent**	**unknown**	**known**	**observational**
		Stochastic Interventions:			
Linear Gaussian		5.9 ± 6.2	3.4 ± 3.2	0.5 ± 1.3	10.3 ± 7.8
Non-Linear Gaussian	1	12.2 ± 3.9	10.3 ± 2.5	7.0 ± 3.6	13.7 ± 3.8
Non-Linear Non-Gaussian		8.7 ± 6.6	8.0 ± 2.7	6.6 ± 2.2	11.3 ± 5.0
Linear Gaussian		27.2 ± 6.2	24.1 ± 5.8	15.6 ± 6.0	39.6 ± 5.0
Non-Linear Gaussian	4	35.8 ± 3.8	30.3 ± 5.3	27.7 ± 4.3	37.5 ± 5.2
Non-Linear Non-Gaussian		36.1 ± 4.4	35.5 ± 8.1	31.5 ± 5.6	40.2 ± 6.9
		Imperfect Interventions:			
Linear Gaussian		5.8 ± 4.2	6.2 ± 3.06	4.7 ± 3.6	10.4 ± 2.9
Non-Linear Gaussian	1	9.3 ± 2.4	8.9 ± 2.5	7.8 ± 3.9	10.5 ± 2.8
Non-Linear Non-Gaussian		8.8 ± 3.0	9.1 ± 3.5	7.9 ± 1.4	11.5 ± 5.4
Linear Gaussian		35.9 ± 8.3	29.7 ± 5.6	17.7 ± 7.9	39.1 ± 9.1
Non-Linear Gaussian	4	32.1 ± 6.0	32.6 ± 5.8	32.8 ± 5.4	39.8 ± 9.3
Non-Linear Non-Gaussian		30.4 ± 12.2	30.2 ± 11.2	25.8 ± 3.9	36.7 ± 9.8

Table 4.1: Hamming distances on synthetic 10 variable SCMs.



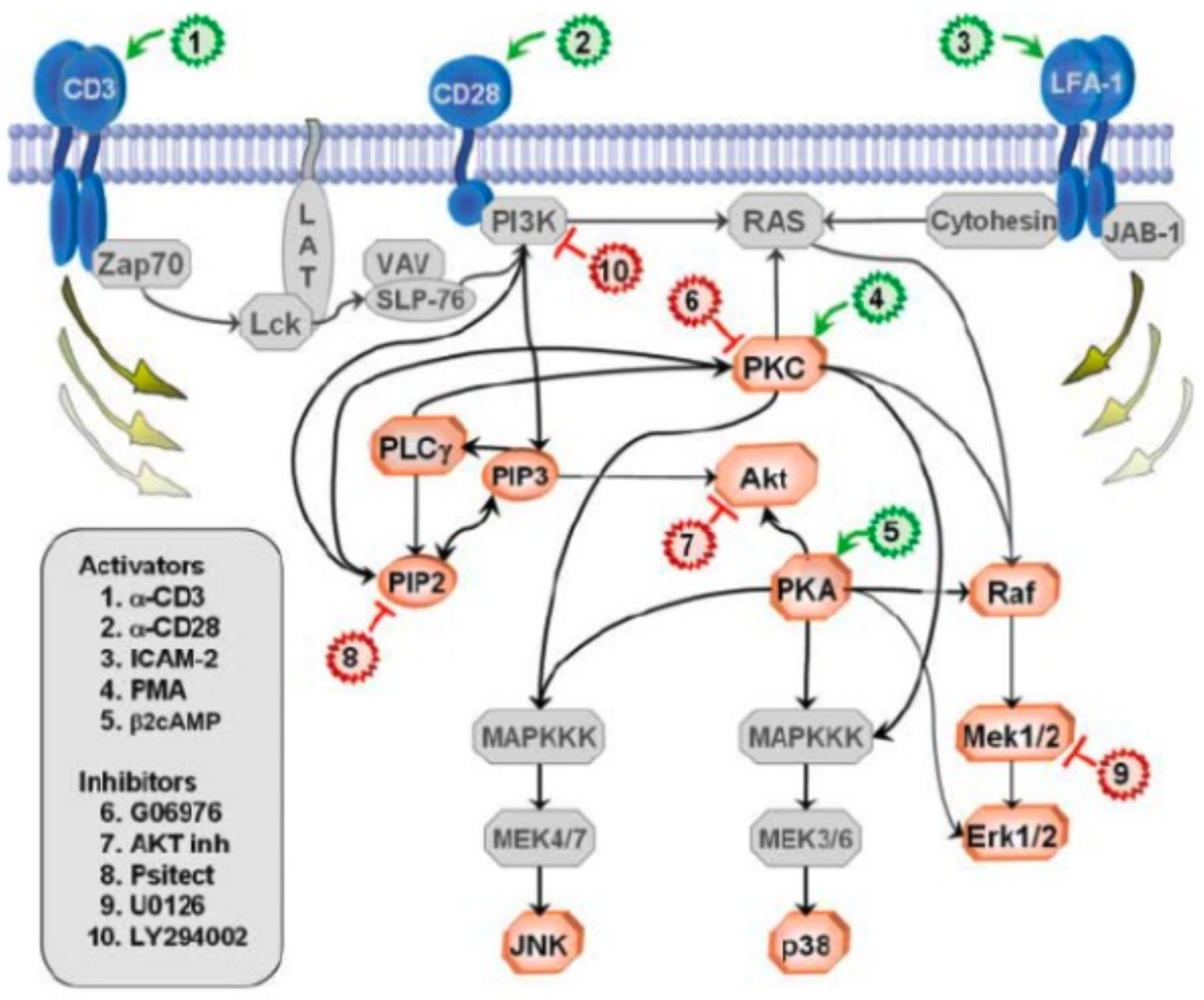
• Experiments on Real Data



ground truth causal model.

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian (**ours**)	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian (**ours**)	19	7	11	8	0	0.42
Imperfect Normalizing Flow (**ours**)	30	9	9	21	1	0.38
Perfect Linear Gaussian (**ours**)	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian (**ours**)	24	11	7	17	1	0.48
Perfect Normalizing Flow (**ours**)	23	7	11	12	2	0.38

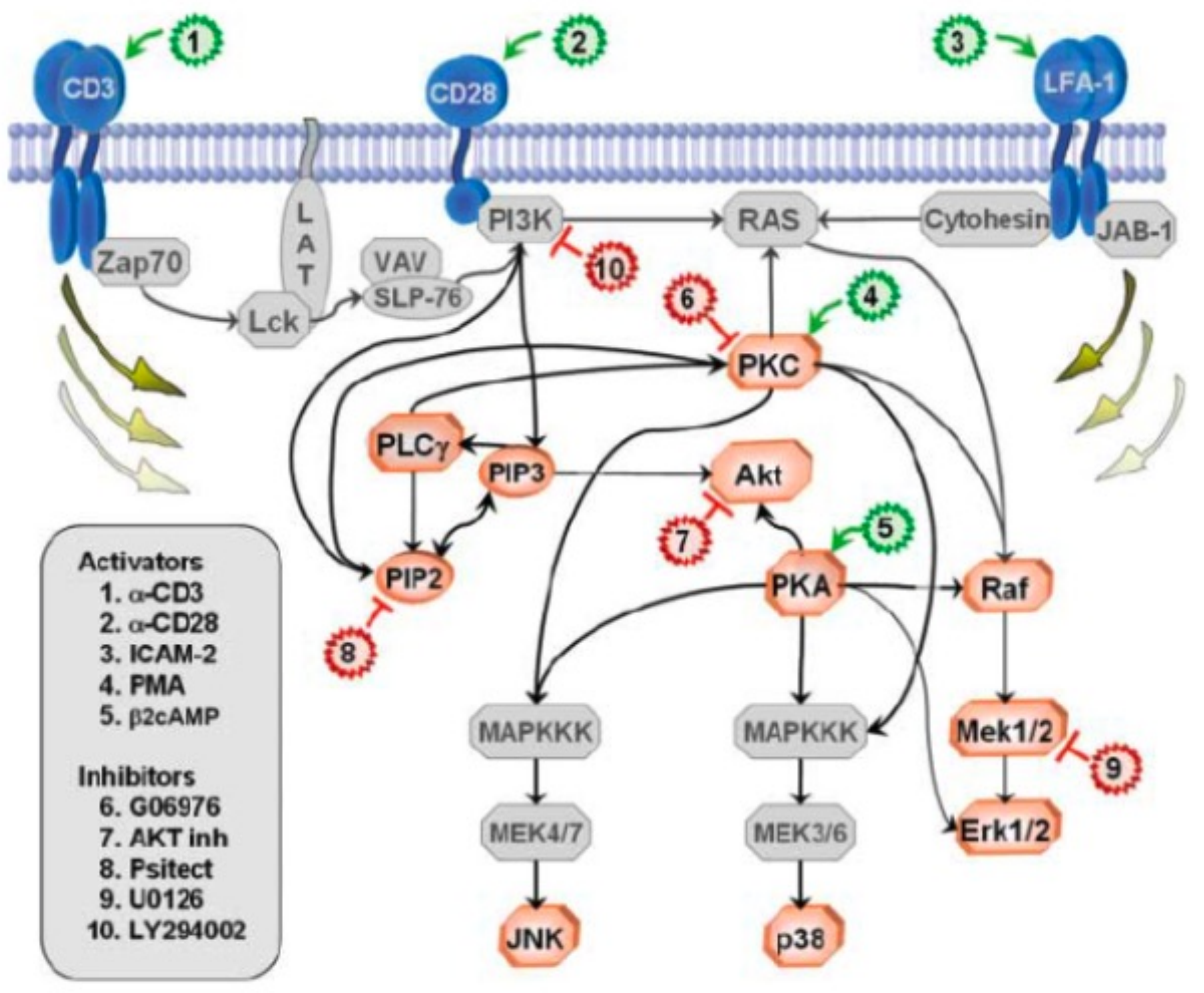
• Experiments on real data



ground truth causal model.

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian **(ours)**	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian **(ours)**	19	7	11	8	0	0.42
Imperfect Normalizing Flow **(ours)**	30	9	9	21	1	0.38
Perfect Linear Gaussian **(ours)**	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian **(ours)**	24	11	7	17	1	0.48
Perfect Normalizing Flow **(ours)**	23	7	11	12	2	0.38

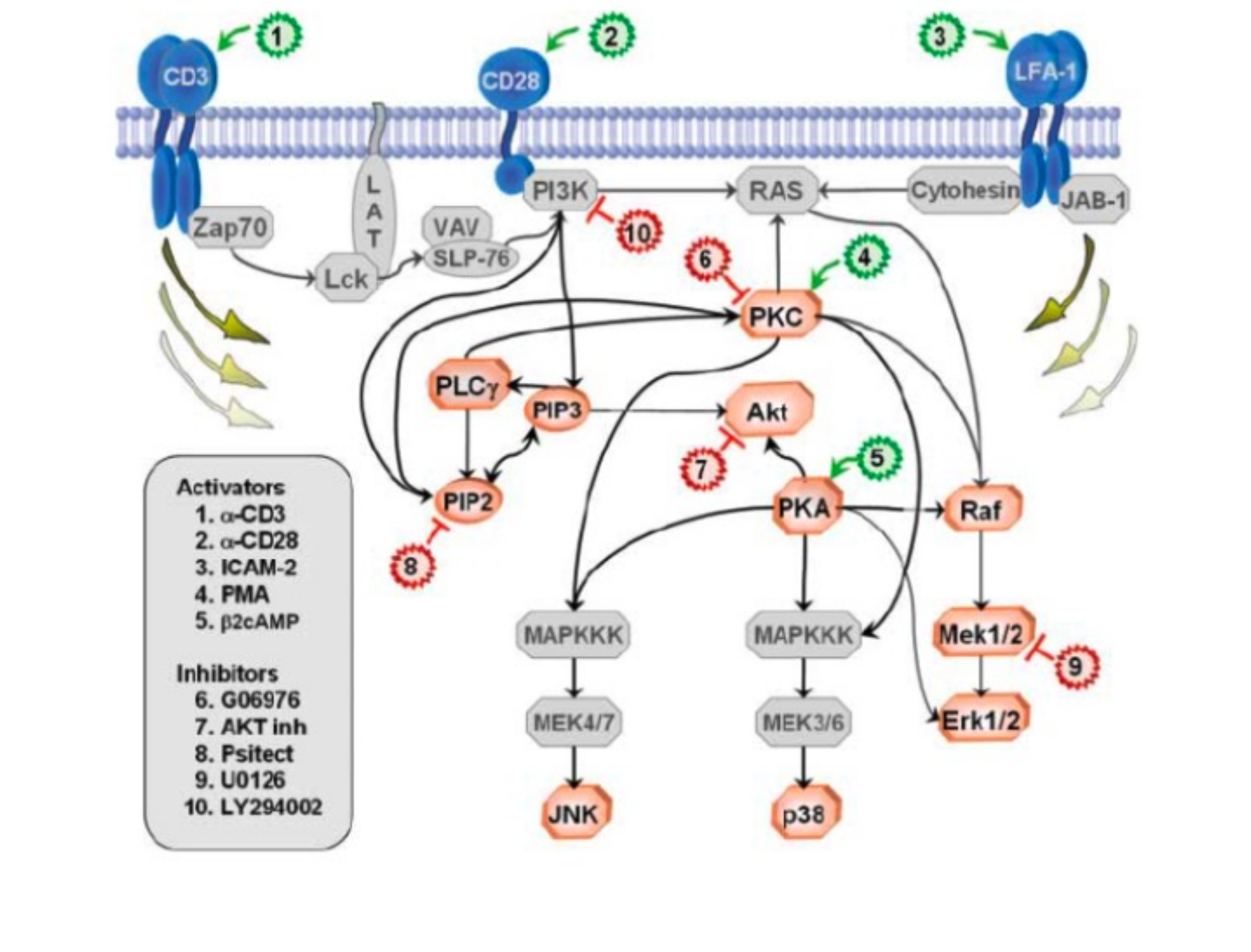
- # Experiments on real data



ground truth causal model.

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	$\boxed{0.51}$
IGSP (Wang et al., 2017)	18	4	6	5	7	$\boxed{0.42}$
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian (**ours**)	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian (**ours**)	19	7	11	8	0	$\boxed{0.42}$
Imperfect Normalizing Flow (**ours**)	30	9	9	21	1	0.38
Perfect Linear Gaussian (**ours**)	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian (**ours**)	24	11	7	17	1	$\boxed{0.48}$
Perfect Normalizing Flow (**ours**)	23	7	11	12	2	0.38

• Experiments on real data



ground truth causal model.

Edges in the wrong direction

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian **(ours)**	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian **(ours)**	19	7	11	8	0	0.42
Imperfect Normalizing Flow **(ours)**	30	9	9	21	1	0.38
Perfect Linear Gaussian **(ours)**	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian **(ours)**	24	11	7	17	1	0.48
Perfect Normalizing Flow **(ours)**	23	7	11	12	2	0.38

Can Large Language Models Infer Causation from Correlation?

Zhijing Jin[1,2,*] **Jiarui Liu**[3] **Zhiheng Lyu**[4] **Spencer Poff**[5]
Mrinmaya Sachan[2] **Rada Mihalcea**[3] **Mona Diab**[5,†] **Bernhard Schölkopf**[1,†]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany, [2]ETH Zürich,
[3]University of Michigan, [4]University of Hong Kong, [5]Meta AI

6.05836v1 [cs.CL] 9 Jun 2023

Abstract

Causal inference is one of the hallmarks of human intelligence. While the field of CausalNLP has attracted much interest in the recent years, existing causal inference datasets in NLP primarily rely on discovering causality from empirical knowledge (e.g. commonsense knowledge). In this work, we propose the first benchmark dataset to test the pure causal inference skills of large language models (LLMs). Specifically, we formulate a novel task CORR2CAUSE, which takes a (set of) correlational statements and determines the causal relationship between the variables. We curate a large-scale dataset of more than 400K samples, on which we evaluate seventeen existing LLMs. Through our experiments, we identify a key shortcoming of LLMs in terms of their causal inference skills, and show that these models achieve almost close to random performance on the task. This shortcoming is somewhat mitigated when we try to re-purpose LLMs for this skill via finetuning, but we find that these models still fail to generalize – they can only perform causal inference in in-distribution settings when variable names and textual expressions used in the queries are similar to those in the training set, but fail in out-of-distribution settings generated by perturbing these queries. CORR2CAUSE is a challenging task for LLMs, and would be helpful in guiding future research on improving LLMs' pure reasoning skills and generalizability.[1]

THANKS!