Causality for Machine Learning

INVICTA Spring School

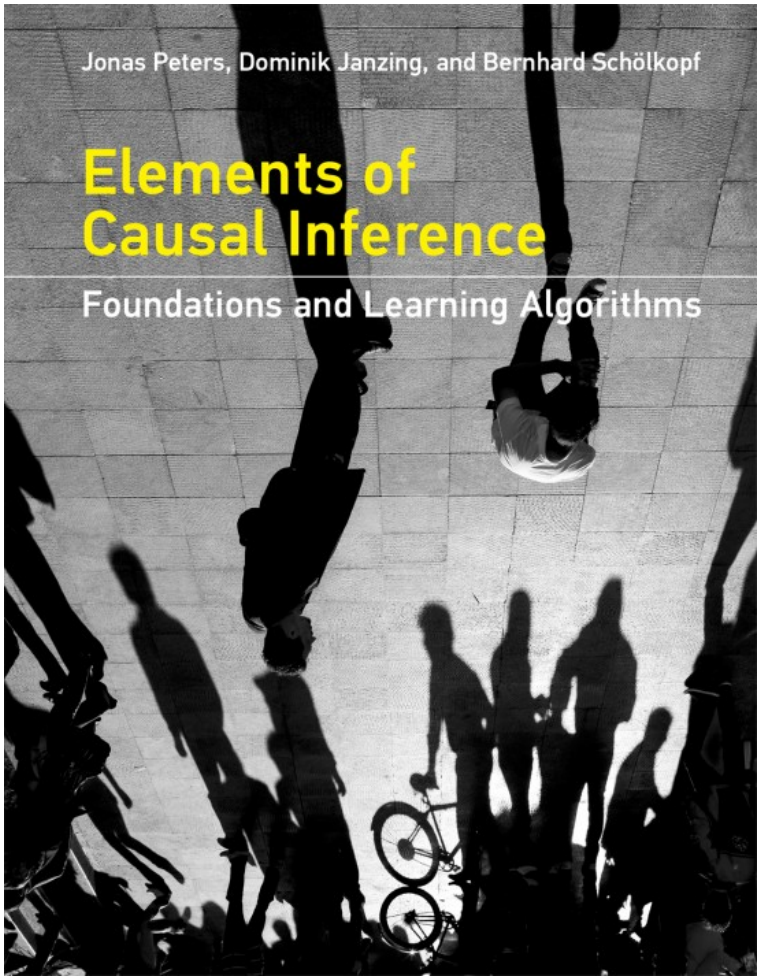
Julius von Kügelgen

ETH Zürich

20 March 2024

Disclaimer: Sources

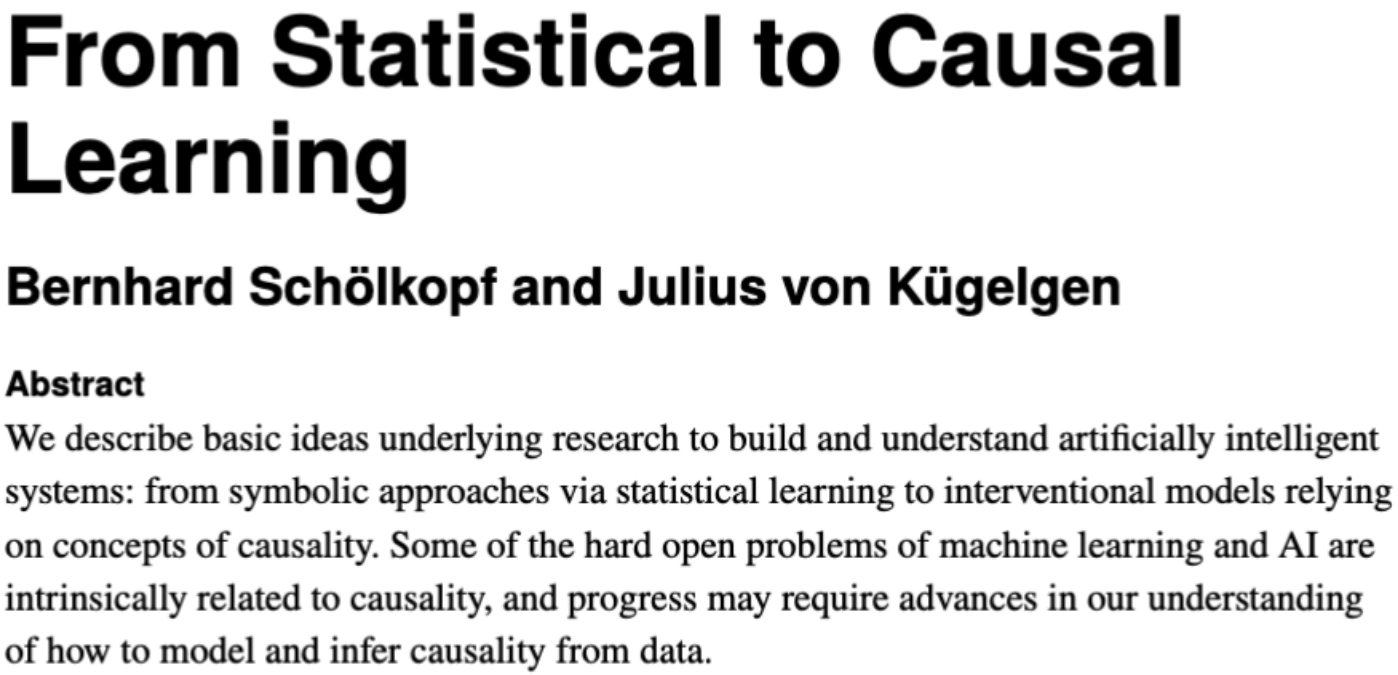
- Much of the material (incl. many figures) is from Peters et al. (2017)
- Biased treatment of causal inference with focus on cause-effect models and connections to machine learning.
- Meant to complement other works such as those of Pearl (2009); Spirtes et al. (2000); Imbens and Rubin (2015).



A free PDF version of the book is available online via:
`https://mitp-content-server.mit.edu/books/content/sectbyfn?collid=books_pres_0&id=11283&fn=11283.pdf`

Disclaimer: Sources

Other material is based on the following review article:



From Statistical to Causal Learning

Bernhard Schölkopf and Julius von Kügelgen

Abstract

We describe basic ideas underlying research to build and understand artificially intelligent systems: from symbolic approaches via statistical learning to interventional models relying on concepts of causality. Some of the hard open problems of machine learning and AI are intrinsically related to causality, and progress may require advances in our understanding of how to model and infer causality from data.

In: *Proceedings of the International Congress of Mathematicians*, 2022.
`https://arxiv.org/abs/2204.00607`

Outline

A First Definition of Causation

Philosophers have thought about causation for a very long time, leading (amongst others) to counterfactual definitions such as

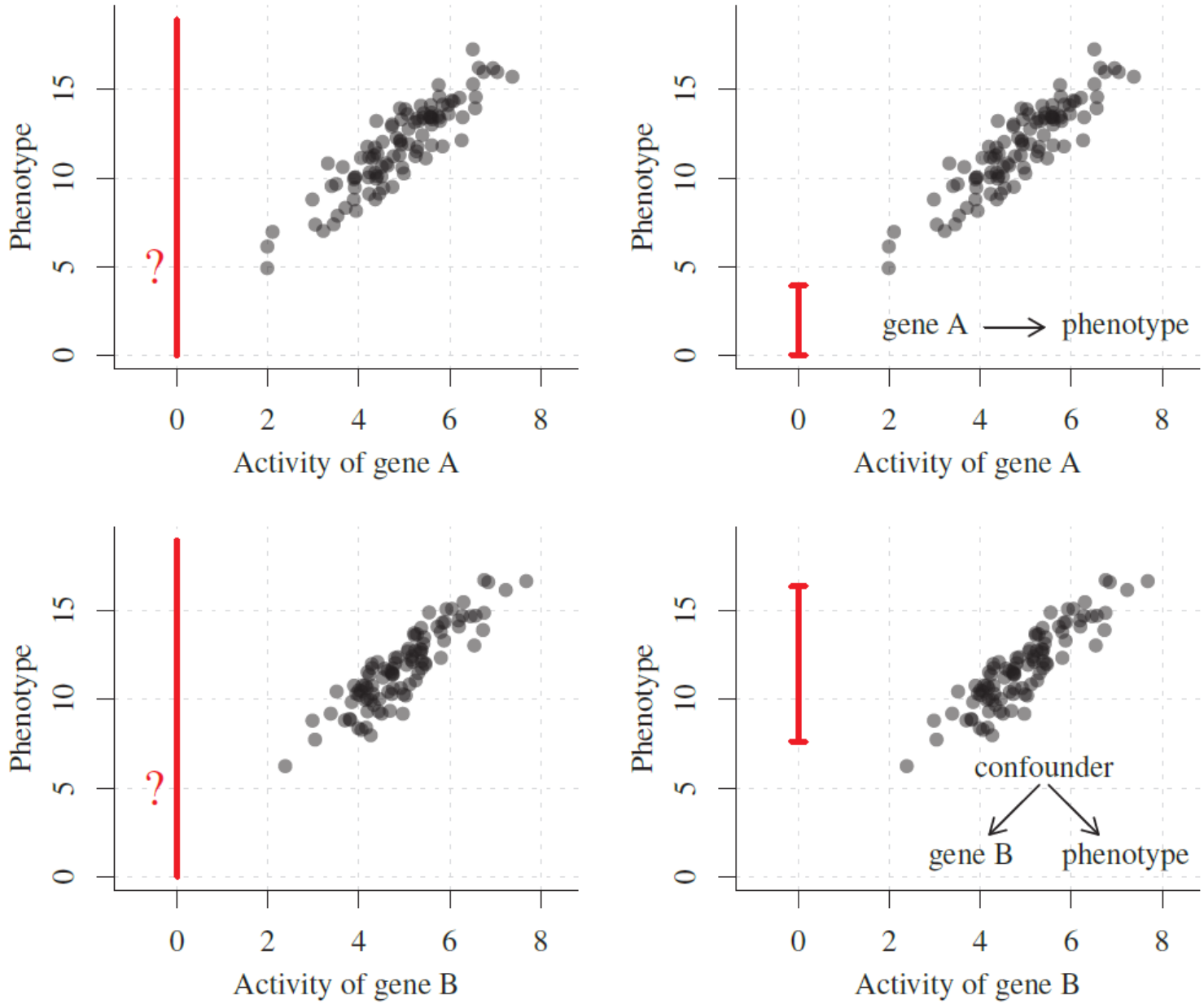
> *"We may define a cause to be an object followed by another [...] where, if the first object had not been, the second never had existed."—Hume (1748)*

> *"We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well"—Lewis (1973)*

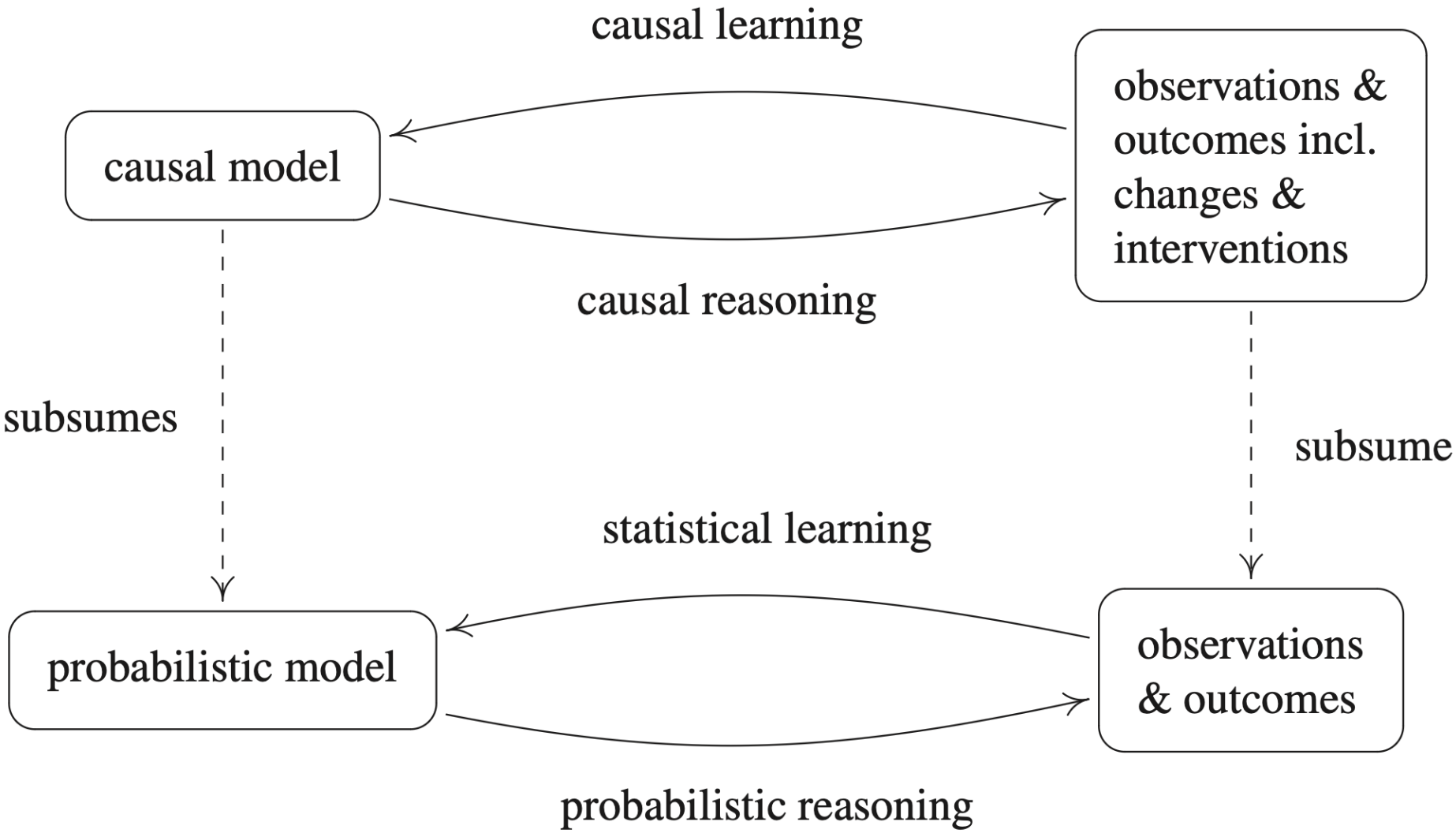
Definition (Total Causal Effect (informal))

X has a (total) causal effect on Y if there are distinct values $x \neq x'$ such that the distribution of Y differs when changing X from x to x'.

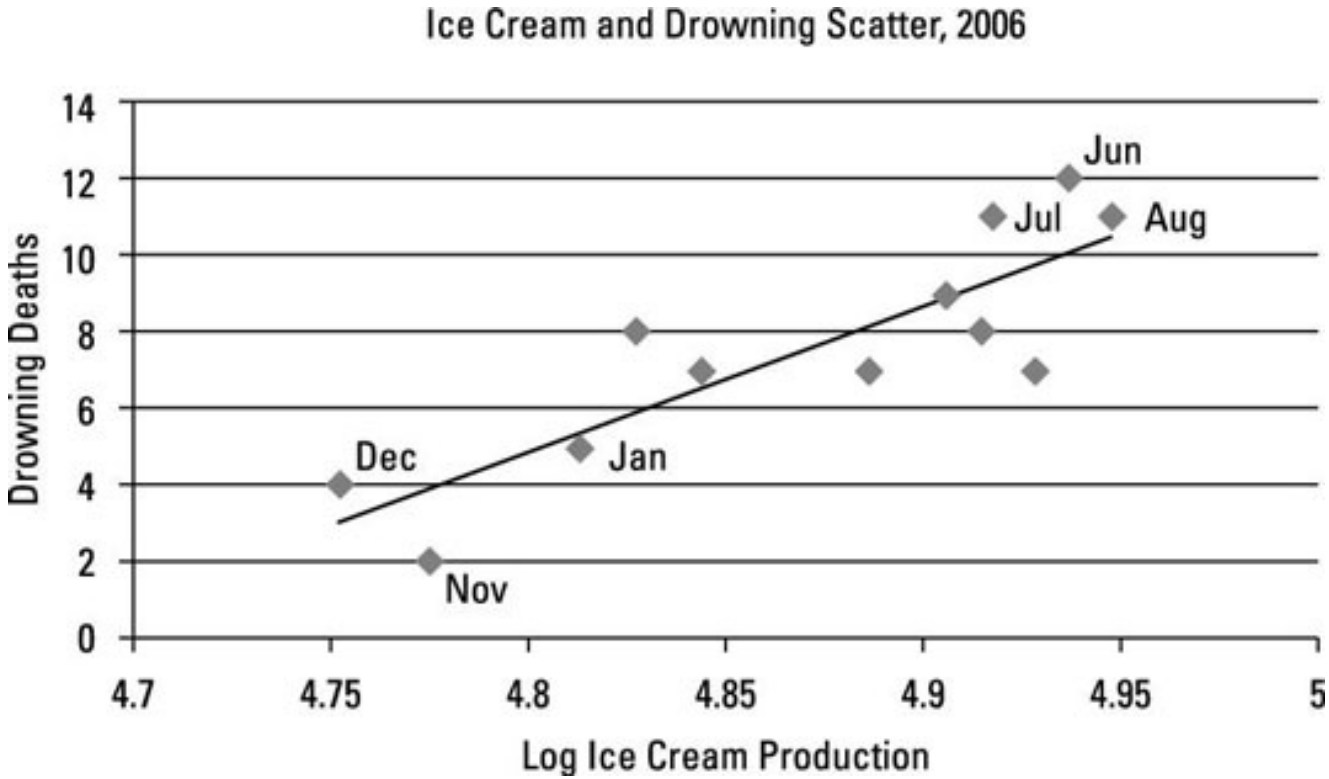
Why Causal Structure Matters: Interventions



Big Picture


causal learning
causal model
observations & outcomes incl. changes & interventions
causal reasoning
subsumes
subsume
statistical learning
probabilistic model
observations & outcomes
probabilistic reasoning

Correlation $\neq$ Causation

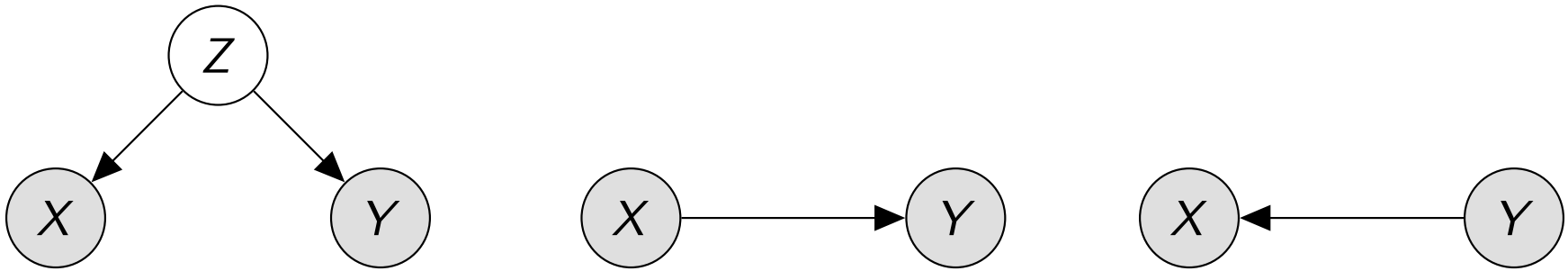


Should you avoid swimming after eating ice cream? Does hearing about drowning deaths increase people's appetite for ice cream?

From Correlation to Causation

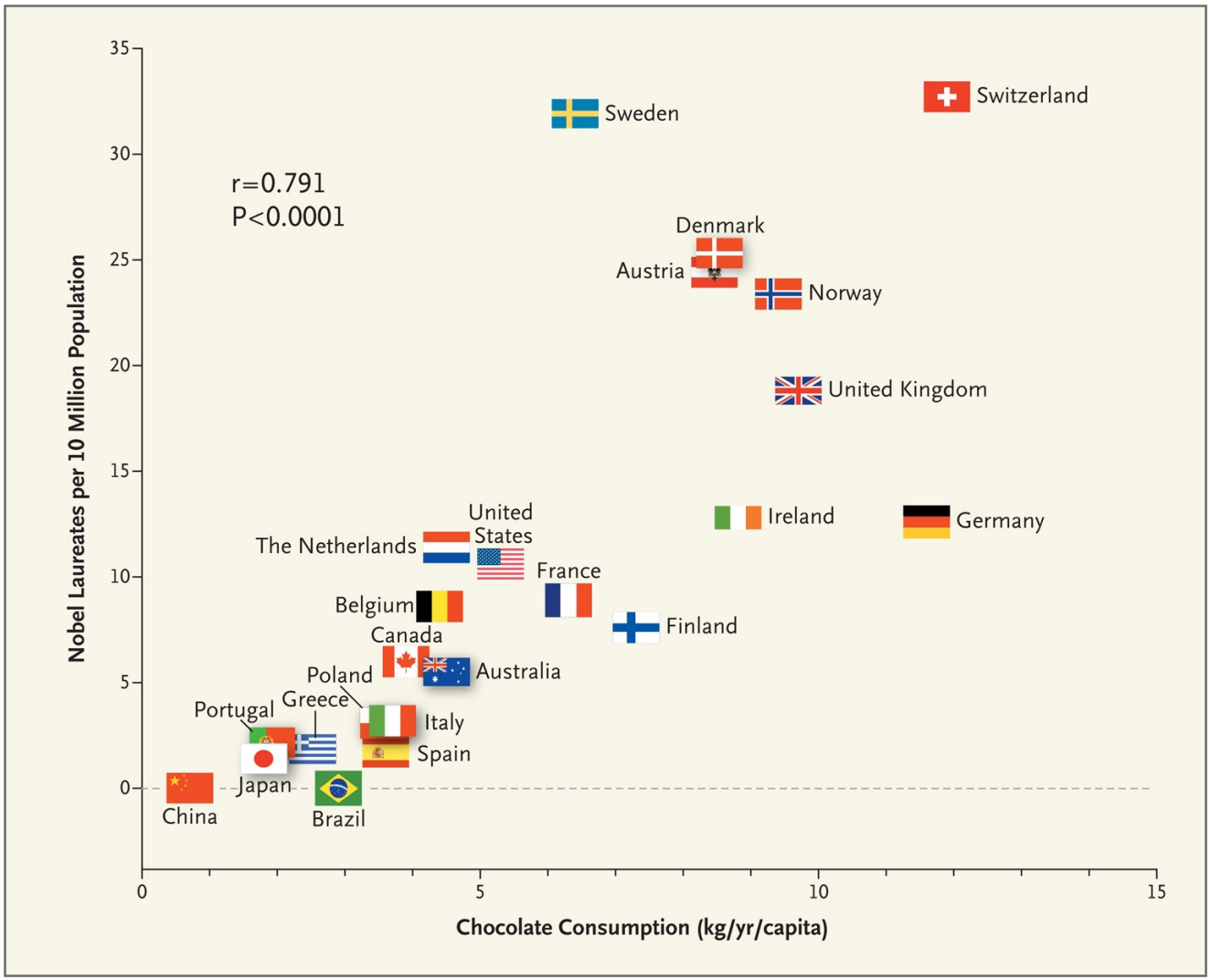
Principle (Reichenbach's common cause principle)

If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a third variable Z that causally influences both. (As a special case, Z may coincide with either X or Y.)
Furthermore, this variable Z screens X and Y from each other in the sense that given Z they become independent, $X \perp\!\!\!\perp Y | Z$.



The common cause principle (Reichenbach, 1956) establishes a link between statistical properties and possible causal structures.

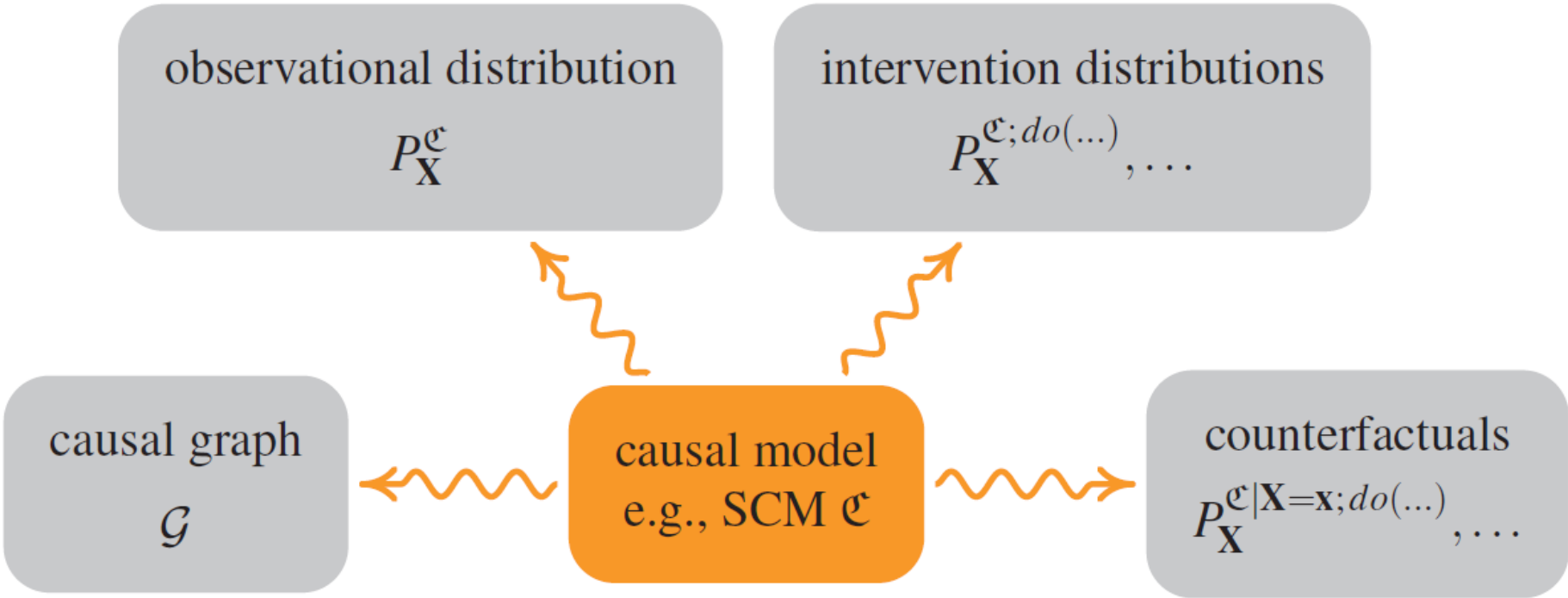
Another Famous Spurious Correlation[1]



[1] Franz H Messerli. "Chocolate consumption, cognitive function, and Nobel laureates". In: *N Engl J Med* 367.16 (2012), pp. 1562–1564.

Outline

What Does a Causal Model Entail?



A Structured Family of Distributions

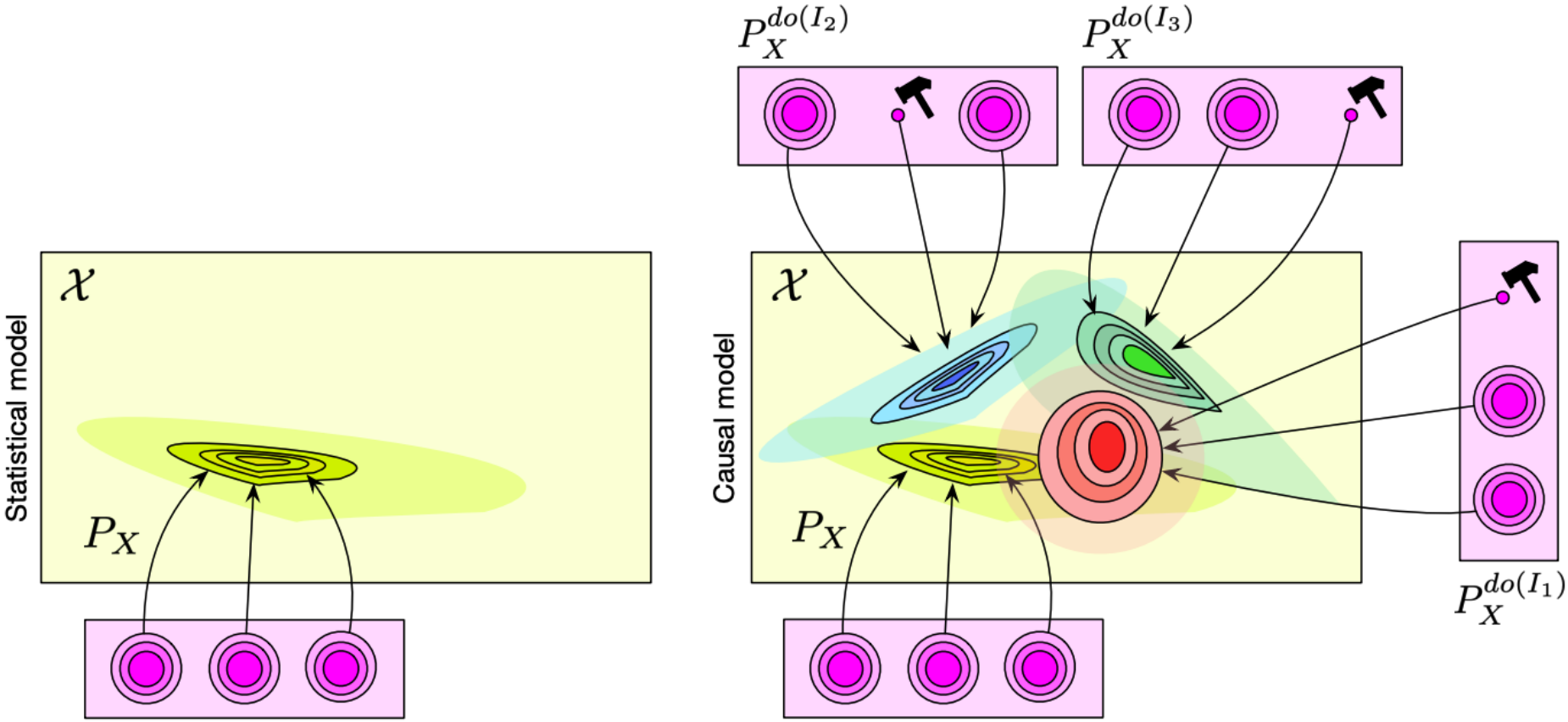


Figure from: Schölkopf et al., 2021

Structural Causal Model (SCM)

Definition (Markovian SCM)

A Markovian structural causal model (SCM) $\mathfrak{C} = (\boldsymbol{S}, P_{\boldsymbol{N}})$ over observables $\boldsymbol{X} = \{X_1, ..., X_n\}$ consists of (i) a collection $\boldsymbol{S}$ of structural equations,

$$X_j := f_j(\mathbf{PA}_j, N_j), \qquad \text{for} \qquad j = 1, ..., n$$

where $\mathbf{PA}_j \subseteq \{X_1, ..., X_n\} \setminus \{X_j\}$ are the causal parents (direct causes) of X_j; and (ii) a factorizing joint distribution $P_{\boldsymbol{N}} = P_{N_1} \times ... \times P_{N_n}$ over jointly independent ("exogenous") noise variables $\boldsymbol{N} = (N_1, ..., N_n)$.

The functions $f_j(\cdot)$ are deterministic. All stochasticity in the system comes from the noise $\boldsymbol{N} \sim P_{\boldsymbol{N}}$, which induces observational, interventional, and counterfactual distributions of $\boldsymbol{X}$ through (manipulations of) $\boldsymbol{S}$.

Structural Causal Model (SCM)

Definition (Markovian SCM)

A Markovian structural causal model (SCM) $\mathfrak{C} = (\boldsymbol{S}, P_{\boldsymbol{N}})$ over observables $\boldsymbol{X} = \{X_1, ..., X_n\}$ consists of (i) a collection $\boldsymbol{S}$ of structural equations,

$$X_j := f_j(\mathbf{PA}_j, N_j), \qquad \text{for} \qquad j = 1, ..., n$$

where $\mathbf{PA}_j \subseteq \{X_1, ..., X_n\} \setminus \{X_j\}$ are the causal parents (direct causes) of X_j; and (ii) a factorizing joint distribution $P_{\boldsymbol{N}} = P_{N_1} \times ... \times P_{N_n}$ over jointly independent ("exogenous") noise variables $\boldsymbol{N} = (N_1, ..., N_n)$.

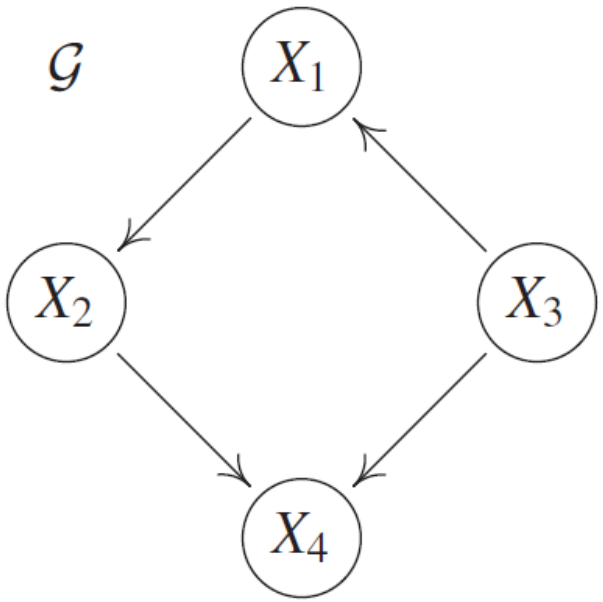
The functions $f_j(\cdot)$ are deterministic. All stochasticity in the system comes from the noise $\boldsymbol{N} \sim P_{\boldsymbol{N}}$, which induces observational, interventional, and counterfactual distributions of $\boldsymbol{X}$ through (manipulations of) $\boldsymbol{S}$.

Structural Causal Model (SCM)

Example of an SCM with 4 variables. The corresponding causal graph $\mathcal{G}$ is obtained by drawing a directed edge from each node in $\mathbf{PA}_j$ to X_j. We will assume $\mathcal{G}$ does not contain cycles, i.e., it is a directed acyclic graph (DAG).

$$
\begin{aligned}
X_1 &:= f_1(X_3, N_1) \\
X_2 &:= f_2(X_1, N_2) \\
X_3 &:= f_3(N_3) \\
X_4 &:= f_4(X_2, X_3, N_4)
\end{aligned}
$$

- $N_1, \ldots, N_4$ jointly independent
- $\mathcal{G}$ is acyclic



Ancestral sampling: to sample $\boldsymbol{X}$, first draw $\boldsymbol{N} \sim P_{\boldsymbol{N}}$, then iteratively compute X_i following the (partial) causal order induced by the graph.

Causal Markov condition

An SCM with associated DAG G implies the following causal factorisation of the observational distribution:

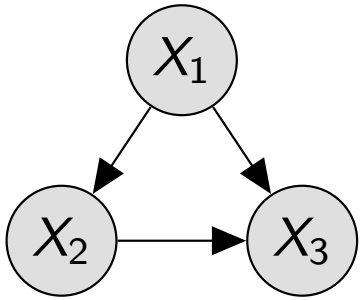
$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{PA}_i)$$

where $\mathbf{PA}_i = \{X_j : (X_j \to X_i) \in G\}$ denotes the set of parents, or direct causes, of X_i in G. This is equivalent to:

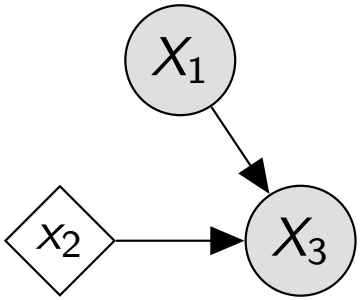
Definition (Causal Markov condition)

A distribution p satisfies the causal Markov condition w.r.t. a DAG G if every variable is conditionally independent of its non-descendants in G given its parents in G.

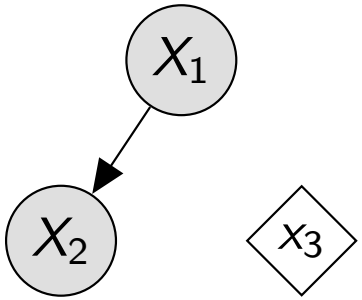
Hard Interventions and Graph Surgery



(a) G



(b) $G' : do(X_2 = x_2)$



(c) $G'' : do(X_3 = x_3)$

When a variable is intervened upon, $do(\cdot)$, and set to a constant (white diamonds), this removes any influence from other variables, captured graphically by removing all incoming edges.

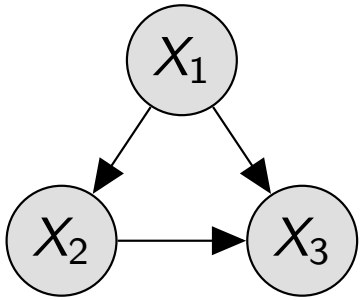
Interventional distributions are computed as conditional distributions, factorised over the corresponding post-intervention graph:

$$p_G(X_3|do(X_2 := x_2)) = p_{G'}(X_3|X_2 = x_2)$$

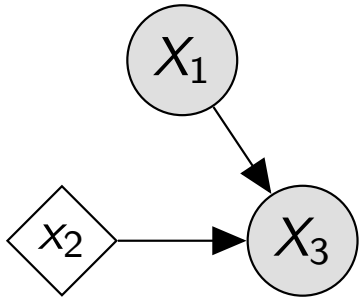
$\implies$ g-formula/truncated factorisation/manipulation theorem:

$$p(\mathbf{X} \mid do(\mathbf{x}_\mathcal{I})) = \mathbb{I}\{\mathbf{X}_\mathcal{I} = \mathbf{x}_\mathcal{I}\} \prod_{i \notin \mathcal{I}} p(X_i \mid \mathbf{PA}_i)$$

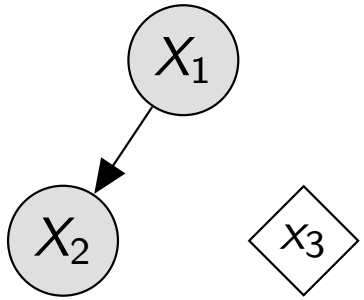
Hard Interventions and Graph Surgery



(a) G



(b) $G' : do(X_2 = x_2)$



(c) $G'' : do(X_3 = x_3)$

When a variable is intervened upon, $do(\cdot)$, and set to a constant (white diamonds), this removes any influence from other variables, captured graphically by removing all incoming edges.

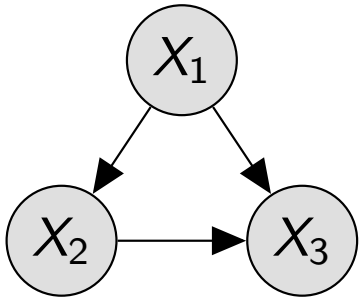
Interventional distributions are computed as conditional distributions, factorised over the corresponding post-intervention graph:

$$p_G(X_3|do(X_2 := x_2)) = p_{G'}(X_3|X_2 = x_2)$$

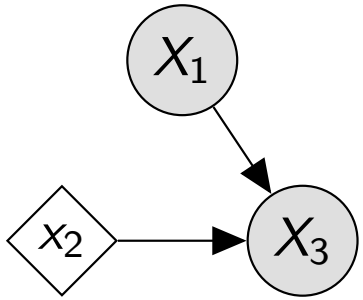
$\implies$ g-formula/truncated factorisation/manipulation theorem:

$$p(\mathbf{X} \mid do(\mathbf{x}_\mathcal{I})) = \mathbb{I}\{\mathbf{X}_\mathcal{I} = \mathbf{x}_\mathcal{I}\} \prod_{i \notin \mathcal{I}} p(X_i \mid \mathbf{PA}_i)$$

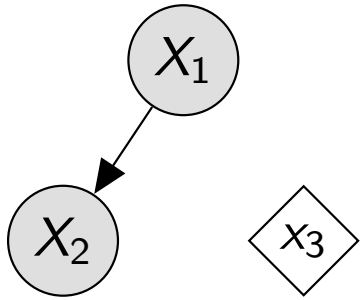
Hard Interventions and Graph Surgery



(a) G



(b) $G' : do(X_2 = x_2)$



(c) $G'' : do(X_3 = x_3)$

When a variable is intervened upon, $do(\cdot)$, and set to a constant (white diamonds), this removes any influence from other variables, captured graphically by removing all incoming edges.

Interventional distributions are computed as conditional distributions, factorised over the corresponding post-intervention graph:

$$p_G(X_3|do(X_2 := x_2)) = p_{G'}(X_3|X_2 = x_2)$$

$\implies$ g-formula/truncated factorisation/manipulation theorem:

$$p(\mathbf{X} \mid do(\mathbf{x}_\mathcal{I})) = \mathbb{I}\{\mathbf{X}_\mathcal{I} = \mathbf{x}_\mathcal{I}\} \prod_{i \notin \mathcal{I}} p(X_i \mid \mathbf{PA}_i)$$

Simpson's Paradox: Kidney Stones Example[2]

Table: Recovery rates (Y) for different treatments (T) and stone sizes (Z).

	Overall	Small Stones	Large Stone
Treatment A	78%	**93%**	**73%**
Treatment B	**83%**	87%	69%

[2] Charig et al., 1986.

Simpson's Paradox: Kidney Stones Example[2]

Table: Recovery rates (Y) for different treatments (T) and stone sizes (Z).

	Overall	Small Stones	Large Stone
Treatment A	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B	**83%** (289/350)	87% (234/270)	69% (55/80)

[2] Charig et al., 1986.

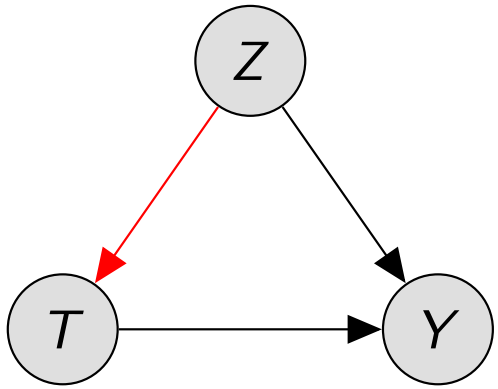
Simpson's Paradox: Kidney Stones Example[2]

Table: Recovery rates (Y) for different treatments (T) and stone sizes (Z).

	Overall	Small Stones	Large Stone
Treatment A	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B	**83%** (289/350)	87% (234/270)	69% (55/80)

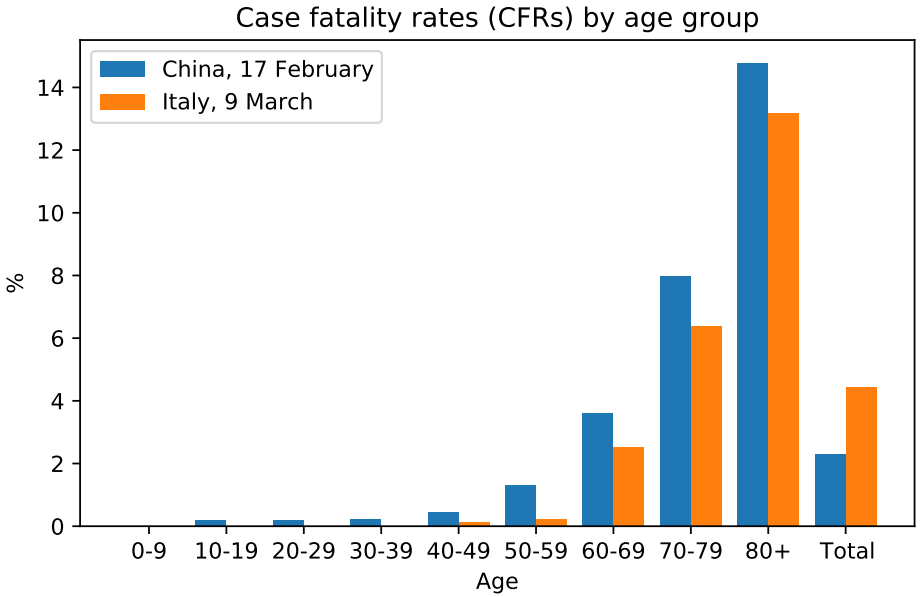
Need to adjust for confounder Z:

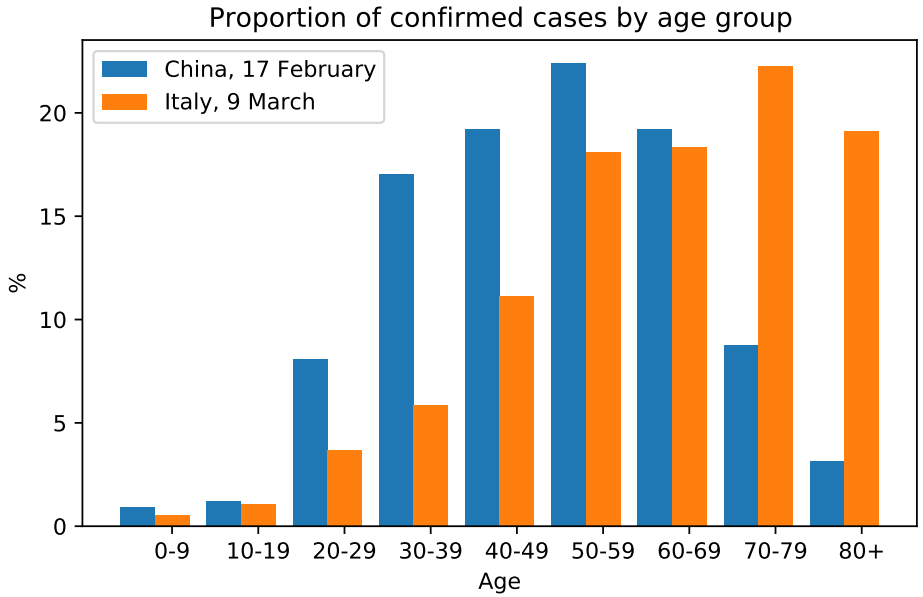
$$
\begin{aligned}
p\big(y|do(t)\big) &= \sum_{z\in\mathcal{Z}} p\big(y|do(t),z\big)p\big(z|do(t)\big) \\
&= \sum_{z\in\mathcal{Z}} p(y|t,z)p(z) \\
&\neq \sum_{z\in\mathcal{Z}} p(y|t,z)p(z|t) = p(y|t)
\end{aligned}
$$



[2] Charig et al., 1986.

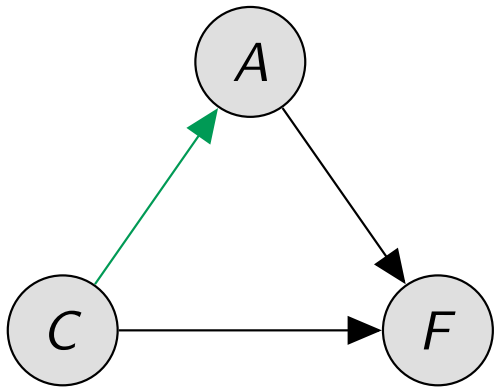
Simpson's Paradox: Covid-19 Example[3]





Assumed causal graph: age A acts as a *mediator* of the effect of country C on case fatality F

$\Longrightarrow$ do *not* adjust for age, if interested in total causal effect: $p(f|do(c)) = p(f|c)$.



[3] von Kügelgen et al., 2021.

Computing counterfactuals with SCMs

Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X, U_Y \sim \mathcal{N}(0, 1). \tag{1}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?".

Computing counterfactuals with SCMs

Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X, U_Y \sim \mathcal{N}(0, 1). \tag{1}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?".

1) Abduction: Updating the noise using the observed evidence via (1), we obtain the counterfactual SCM $\mathcal{M}^{X=2,Y=6.5}$,

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5). \tag{2}$$

Computing counterfactuals with SCMs

Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X, U_Y \sim \mathcal{N}(0, 1). \tag{1}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?".

1) Abduction: Updating the noise using the observed evidence via (1), we obtain the counterfactual SCM $\mathcal{M}^{X=2,Y=6.5}$,

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5). \tag{2}$$

2) Action: Performing the intervention $do(X := 1)$ in (2) then leads to

$$X := 1, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5).$$

Computing counterfactuals with SCMs

Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X, U_Y \sim \mathcal{N}(0,1). \tag{1}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?".

1) Abduction: Updating the noise using the observed evidence via (1), we obtain the counterfactual SCM $\mathcal{M}^{X=2,Y=6.5}$,

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5). \tag{2}$$

2) Action: Performing the intervention $do(X := 1)$ in (2) then leads to

$$X := 1, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5).$$

3) Prediction: computing the pushforward gives the result $p(Y_{X=1} \mid X = 2, Y = 6.5) = \delta(3.5)$, so "$Y$ would have been 3.5".

Computing counterfactuals with SCMs

Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X, U_Y \sim \mathcal{N}(0,1). \tag{1}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?".

1) Abduction: Updating the noise using the observed evidence via (1), we obtain the counterfactual SCM $\mathcal{M}^{X=2,Y=6.5}$,

$$X := U_X, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5). \tag{2}$$

2) Action: Performing the intervention $do(X := 1)$ in (2) then leads to

$$X := 1, \qquad Y := 3X + U_Y, \qquad U_X \sim \delta(2), \qquad U_Y \sim \delta(0.5).$$

3) Prediction: computing the pushforward gives the result $p(Y_{X=1} \mid X = 2, Y = 6.5) = \delta(3.5)$, so "$Y$ would have been 3.5".

Note: This differs from $p(Y \mid do(X = 1)) = \mathcal{N}(3, 1)$, since the factual observation helped determine the background state ($U_X = 2, U_Y = 0.5$).

Potential Outcome (PO) Framework

Proposed by Neyman for randomized studies and later popularized and extended to observational settings by Rubin and others; it is popular within statistics and epidemiology.

Main idea: view causal inference as a *missing data problem*.

For each individual (or unit) i and treatment t there is a PO $Y_i(t)$ capturing what would happen if individual i received treatment t:

i	T_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	7	?	?
2	0	?	8	?
3	1	3	?	?
4	1	6	?	?
5	0	?	4	?
6	0	?	1	?

Individual Treatment Effects

Definition (ITE)

The ITE for individual i under a binary treatment is defined as

$$\tau_i = Y_i(1) - Y_i(0).$$

"fundamental problem of causal inference": only one of the POs is ever observed for each i, the other becomes counterfactual

$$Y_i = TY_i(1) + (1 - T)Y_i(0).$$

$$Y_i^{\text{CF}} = (1 - T)Y_i(1) + TY_i(0).$$

SCM perspective:

$$Y_i(t) = Y \mid do(T = t) \qquad \text{in an SCM with} \qquad \mathbf{U} = \mathbf{u}_i,$$

Average Treatment Effects

Definition (Treatment effects)

The conditional average treatment effect (CATE), conditioned on (a subset of) features $\mathbf{x}$, is defined as

$$\tau(\mathbf{x}) := \mathbb{E}[Y \mid \mathbf{x}, do(T = 1)] - \mathbb{E}[Y \mid \mathbf{x}, do(T = 0)] = \mathbb{E}[Y(1) - Y(0) \mid \mathbf{x}].$$

The average treatment effect (ATE) is the population average,

$$\tau := \mathbb{E}[\tau(\mathbf{X})] = \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)] = \mathbb{E}[Y(1) - Y(0)].$$

Identification in the PO Framework

When can we estimate causal effects from purely observational data?

Assumption (Overlap/common treatment support)

For any treatment t and any configuration of features $\mathbf{x}$, it holds that: $0 < p(T = t \mid \mathbf{X} = \mathbf{x}) < 1$.

Assumption (Conditional ignorability/no hidden confounding)

Given a treatment $T \in \{0, 1\}$, potential outcomes $Y(0), Y(1)$, and observed covariates $\mathbf{X}$, we have:

$$Y(0) \perp\!\!\!\perp T \mid \mathbf{X} \quad \textit{and} \quad Y(1) \perp\!\!\!\perp T \mid \mathbf{X}. \tag{3}$$

Together, these assumptions guarantee identifiability.

Identification in the Graphical view

Graphical criterion:

Proposition (Valid adjustment sets (Shpitser et al., 2010))

Under causal sufficiency, a set $\mathbf{Z}$ *is a valid adjustment set for the causal effect of a singleton treatment* T *on an outcome* Y *in the sense of*

$$p(y \mid do(t)) = \sum_{\mathbf{z}} p(\mathbf{z}) p(y|t, \mathbf{z}).$$

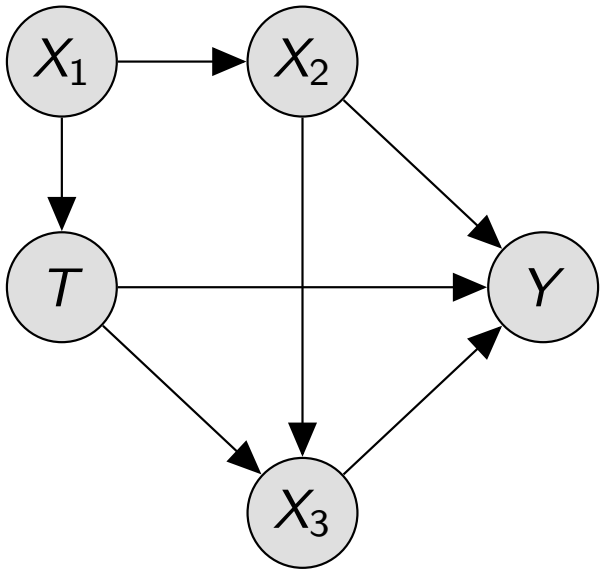
if and only if the following two conditions hold:

(i) $\mathbf{Z}$ *blocks all non-directed paths from* T *to* Y*;*

(ii) $\mathbf{Z}$ *contains no descendant of any node on a directed path from* T *to* Y *(except for descendants of* T *which are not on a directed path from* T *to* Y*).*

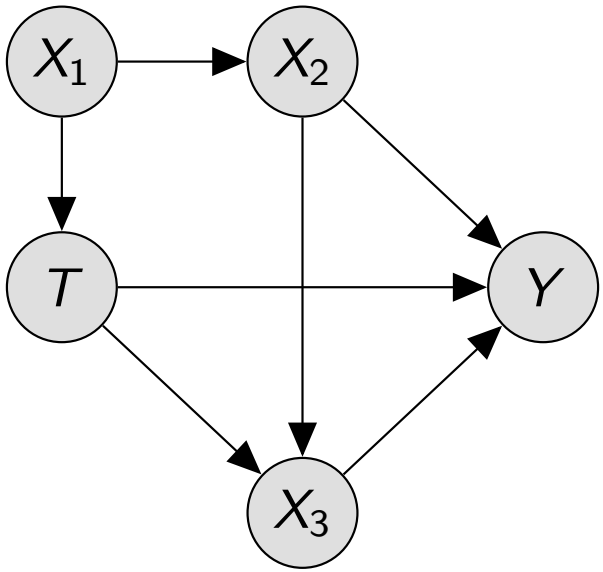
Valid Adjustment Sets: Example

Treatment effect estimation with three observed covariates X_1, X_2, X_3:



Valid Adjustment Sets: Example

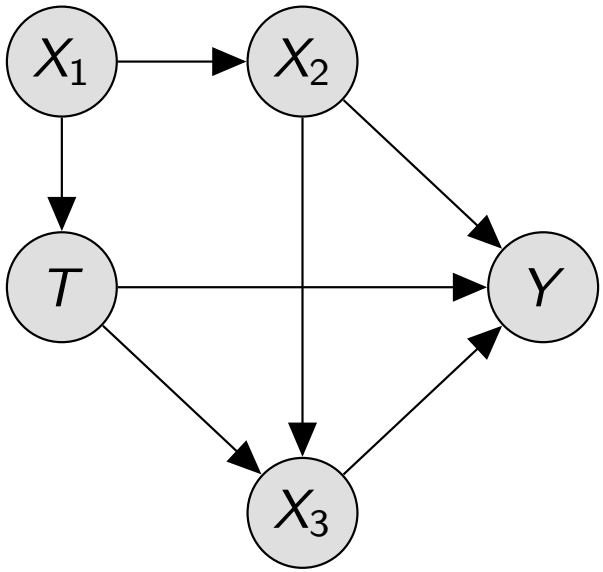
Treatment effect estimation with three observed covariates X_1, X_2, X_3:



Here, the valid adjustment sets for $T \to Y$ are $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$.

Valid Adjustment Sets: Example

Treatment effect estimation with three observed covariates X_1, X_2, X_3:



Here, the valid adjustment sets for $T \to Y$ are $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$.

Including X_3 opens the *non-directed path* $T \to X_3 \leftarrow X_2 \to Y$ and lies on the directed path $T \to X_3 \to Y$, both of which can introduce bias.

(Average) Treatment Effect Estimators

m_1 observations $(\boldsymbol{x}_i, y_i, t_i = 1,)$ from treatment group
m_0 observations $(\boldsymbol{x}_i, y_i, t_i = 0)$ from control group

RCT estimator (Gold standard; no adjustment necessary):

$$\hat{\tau}_{\text{RCT}} = \frac{1}{m_1} \sum_{i \,:\, t_i = 1} y_i - \frac{1}{m_0} \sum_{i \,:\, t_i = 0} y_i .$$

(Average) Treatment Effect Estimators

m_1 observations $(\boldsymbol{x}_i, y_i, t_i = 1,)$ from treatment group
m_0 observations $(\boldsymbol{x}_i, y_i, t_i = 0)$ from control group

RCT estimator (Gold standard; no adjustment necessary):

$$\hat{\tau}_{\text{RCT}} = \frac{1}{m_1} \sum_{i \,:\, t_i = 1} y_i - \frac{1}{m_0} \sum_{i \,:\, t_i = 0} y_i.$$

Regression adjustment: Fit $\hat{f}$ to $\mathbb{E}[Y \mid \mathbf{Z} = \mathbf{z}, T = t]$ where $\boldsymbol{Z}$ is a valid adjustment set; then use $\hat{f}(\mathbf{z}, t)$ to impute counterfactual outcomes as $\hat{y}_i^{\text{CF}} = \hat{f}(\mathbf{z}_i, 1 - t_i)$

$$\hat{\tau}_{\text{regression-adj.}} = \frac{1}{m_1} \sum_{i \,:\, t_i = 1} \left(y_i - \hat{f}(\mathbf{z}_i, 0)\right) + \frac{1}{m_0} \sum_{i \,:\, t_i = 0} \left(\hat{f}(\mathbf{z}_i, 1) - y_i\right),$$

(Average) Treatment Effect Estimators

m_1 observations $(\boldsymbol{x}_i, y_i, t_i = 1,)$ from treatment group
m_0 observations $(\boldsymbol{x}_i, y_i, t_i = 0)$ from control group

Inverse probability weighting (IPW) estimator (aka propensity score):

$$\hat{\tau}_{\text{IPW}} = \frac{1}{m_1} \sum_{i \,:\, t_i=1} \frac{y_i}{p(T = 1 \mid \mathbf{Z} = \mathbf{z}_i)} - \frac{1}{m_0} \sum_{i \,:\, t_i=0} \frac{y_i}{p(T = 0 \mid \mathbf{Z} = \mathbf{z}_i)}.$$

(Average) Treatment Effect Estimators

m_1 observations $(\boldsymbol{x}_i, y_i, t_i = 1,)$ from treatment group
m_0 observations $(\boldsymbol{x}_i, y_i, t_i = 0)$ from control group

Inverse probability weighting (IPW) estimator (aka propensity score):

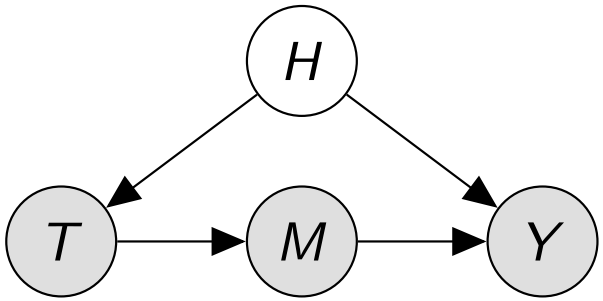
$$\hat{\tau}_{\text{IPW}} = \frac{1}{m_1} \sum_{i\,:\,t_i=1} \frac{y_i}{p(T = 1 \mid \mathbf{Z} = \mathbf{z}_i)} - \frac{1}{m_0} \sum_{i\,:\,t_i=0} \frac{y_i}{p(T = 0 \mid \mathbf{Z} = \mathbf{z}_i)}.$$

Nearest neighbour matching: match and contrast each individual i with the most similar one, $j(i)$, from the opposite treatment group based on $\boldsymbol{Z}$

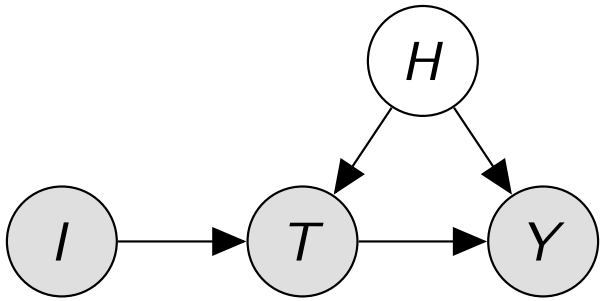
$$\hat{\tau}_{\text{NN-matching}} = \frac{1}{m_1} \sum_{i\,:\,t_i=1} (y_i - y_{j(i)}) + \frac{1}{m_0} \sum_{i\,:\,t_i=0} (y_{j(i)} - y_i).$$

Causal Reasoning with Hidden Confounders/No Overlap

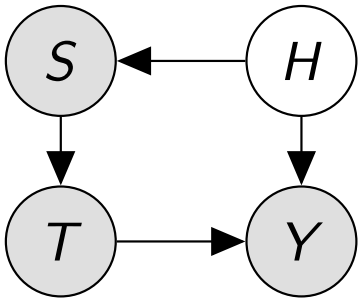
In general, interventional distributions not identifiable (from observational data) with hidden confounders or no overlap between treatment groups.



(a) Front-door



(b) IV



(c) RDD

$$p(y \mid do(t)) = \sum_m p(m \mid t) \sum_{t'} p(t')p(y \mid m, t').$$

Definition (IV)

A variable I is a valid instrument for estimating the effect of treatment T on outcome Y confounded by a hidden variable H if all of the following three conditions hold: (i) $I \perp\!\!\!\perp H$; (ii) $I \not\perp\!\!\!\perp T$; and (iii) $I \perp\!\!\!\perp Y \mid T$.

Instrumental Variables (IV)

Example (Linear IV with 2SLS)

Consider the linear SCM defined by

$$T := aI + bH + U_T, \qquad Y := cH + dT + U_Y.$$

with U_T, U_Y independent noise terms. Then, since $I \perp\!\!\!\perp H$, linear regression of T on I recovers the coefficient a via $\hat{T} = aI$. Substituting for T in the structural equation for Y gives

$$Y := daI + (c + bd)H + U_Y + dU_T.$$

A second linear regression of Y on $\hat{T} = aI$ recovers the causal effect d because $(I \perp\!\!\!\perp H) \implies (\hat{T} \perp\!\!\!\perp H)$, whereas a naive regression of Y on T would give a different result, as $T \not\perp\!\!\!\perp H$.

Regression Discontinuity Design

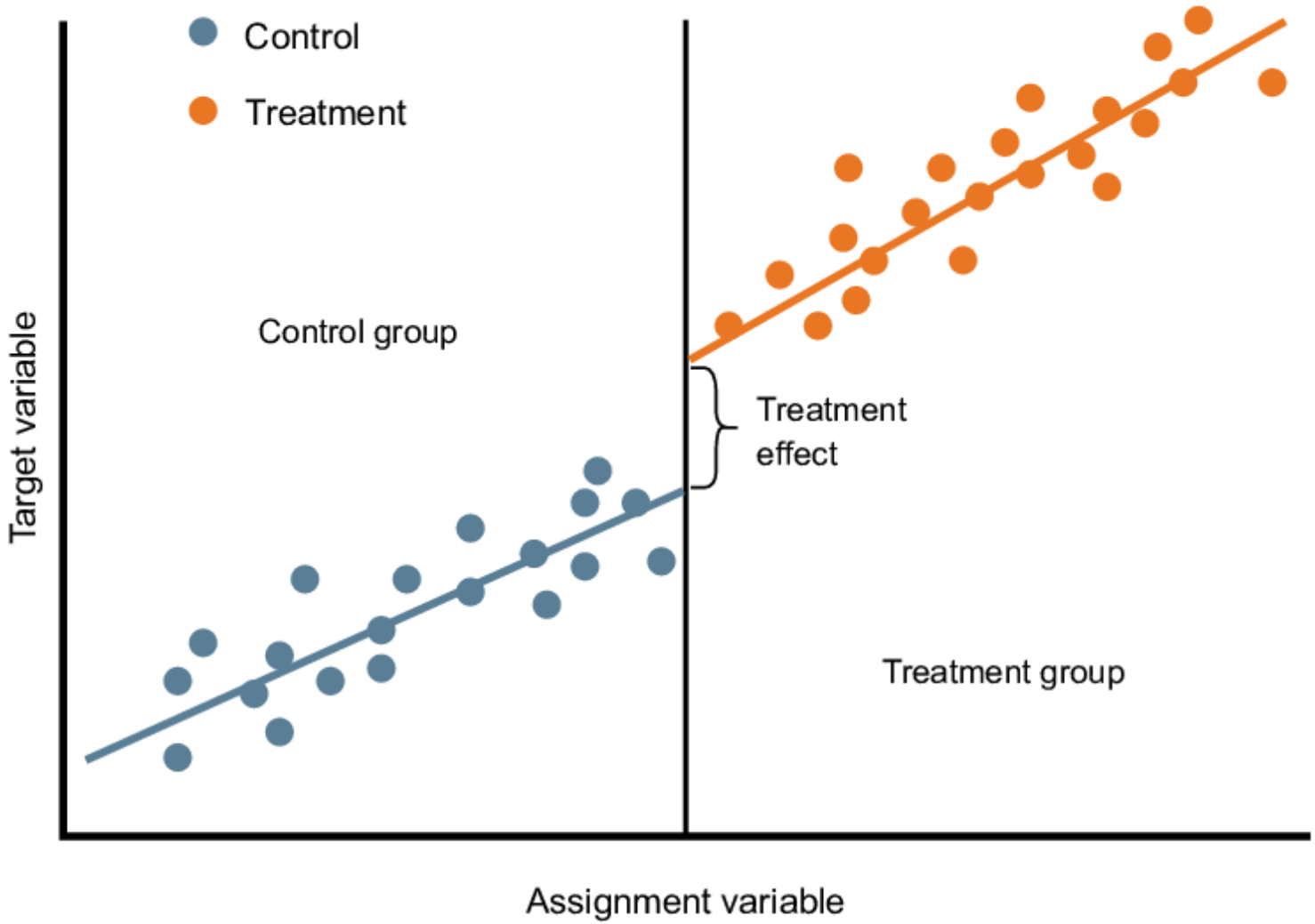
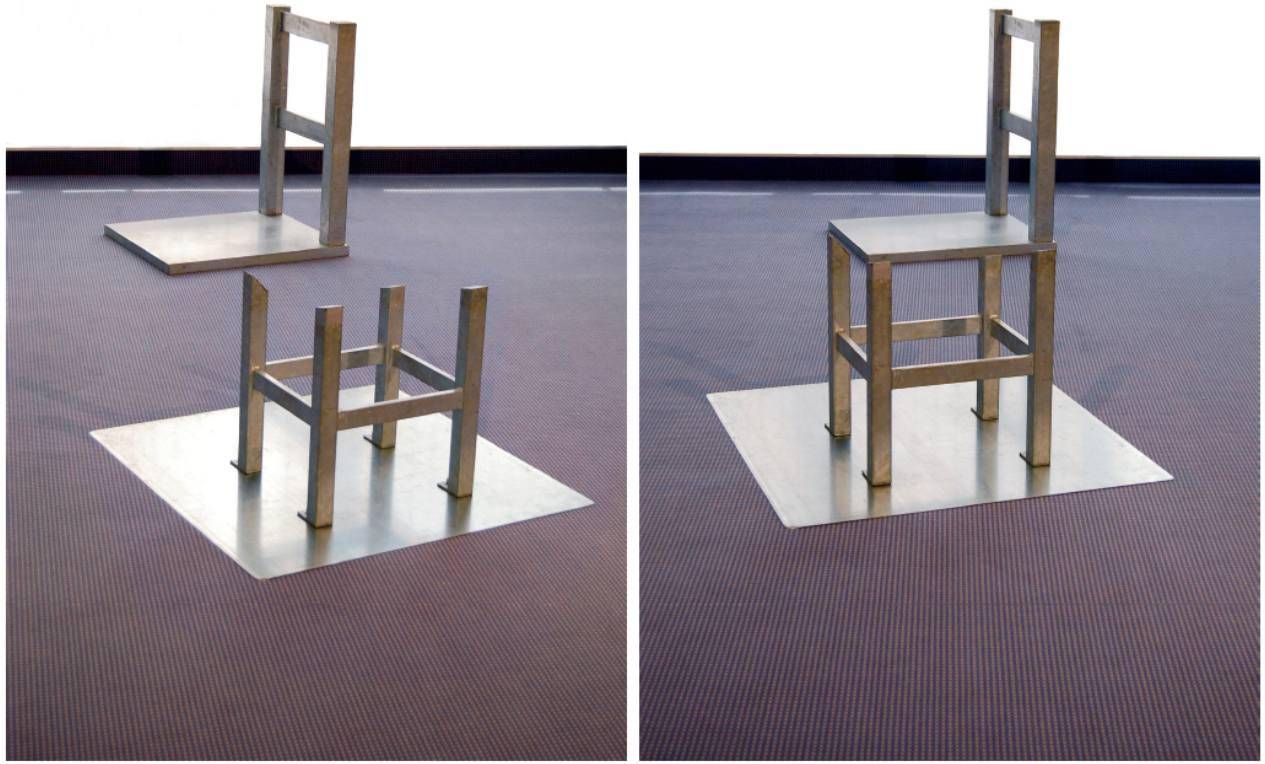


Figure credit: Liu, Yuchu, et al. "Bayesian causal inference in automotive software engineering and online evaluation." (2022).

Outline

Motivating Example 1: Beuchet chair

Generic viewpoint assumption[4]: the object and the mechanism by which it is perceived (e.g., viewpoint, illumination,...) are independent.



The generic viewpoint assumption is violated on the right, leading to an illusion known as the Beuchet chair (image courtesy of Markus Elsholz).

[4] Freeman, 1994

Motivating Example 2: Altitude and Temperature

Suppose we have estimated the joint density $p(a, t)$ of altitude A and temperature T of a sample of cities. Can express this in two ways.

$$\begin{aligned} p(a, t) &= p(a|t)p(t) && (T \to A) \\ &= p(t|a)p(a) && (A \to T) \end{aligned}$$

Which is the causal one?

- What would happen if we intervened on T or A? Which of the conditionals is invariant across different datasets?

Motivating Example 2: Altitude and Temperature

Suppose we have estimated the joint density $p(a, t)$ of altitude A and temperature T of a sample of cities. Can express this in two ways.

$$\begin{aligned} p(a, t) &= p(a|t)p(t) && (T \to A) \\ &= p(t|a)p(a) && (A \to T) \end{aligned}$$

Which is the causal one?

- What would happen if we intervened on T or A? Which of the conditionals is invariant across different datasets?
- Does $p(a)$ contain information about $p(t|a)$? Does $p(t)$ contain information about $p(a|t)$?

Motivating Example 2: Altitude and Temperature

Suppose we have estimated the joint density $p(a, t)$ of altitude A and temperature T of a sample of cities. Can express this in two ways.

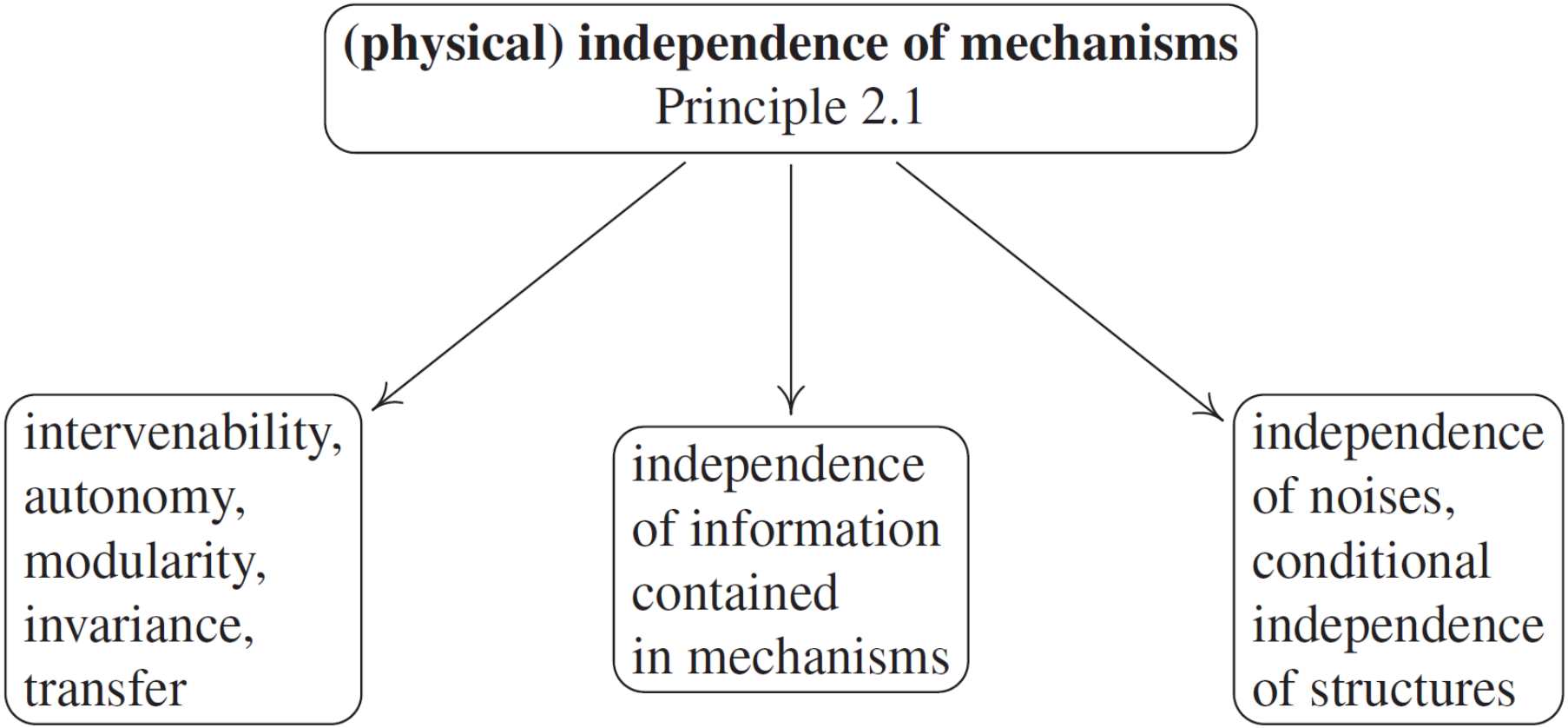
$$\begin{aligned} p(a,t) &= p(a|t)p(t) && (T \to A) \\ &= p(t|a)p(a) && (A \to T) \end{aligned}$$

Which is the causal one?

- What would happen if we intervened on T or A? Which of the conditionals is invariant across different datasets?
- Does $p(a)$ contain information about $p(t|a)$? Does $p(t)$ contain information about $p(a|t)$?
- Which of the two SCMs leads to independent noise terms, $N_A \perp\!\!\!\perp N_T$?

$$\begin{aligned} A &:= N_A \\ T &:= f_T(A, N_T) \end{aligned} \qquad \text{vs.} \qquad \begin{aligned} T &:= N_T \\ A &:= f_A(T, N_A) \end{aligned}$$

Principle of Independent Mechanisms



Principle of Independent Mechanisms

The previous considerations can be seen as part of a fundamental principle.

Principle (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 of cause and mechanism[a].

[a] Daniusis et al., 2010.

Outline

The Two-Variable Case: Learning Cause-Effect Models

Focus on the two variable case: distinguish between cause C and effect E.

$$C := N_C$$
$$E := f_E(C, N_E)$$
$$N_C \perp\!\!\!\perp N_E$$

- This is arguably the most fundamental setting: insights can be transferred to higher-dimensional scenarios.
- No non-trivial conditional independences, so statistical information not helpful (recall the Common Cause Principle)

Why Additional Assumptions Are Required

Proposition (Non-uniqueness of graph structures)

For every joint distribution $P_{X,Y}$ of two real-valued variables, $\exists$ SCM

$$X := N_X, \qquad Y := f_Y(X, N_Y), \qquad N_Y \perp\!\!\!\perp N_X,$$

where f_Y is a measurable function and N_Y is a real-valued variable.

Why Additional Assumptions Are Required

Proposition (Non-uniqueness of graph structures)

For every joint distribution $P_{X,Y}$ of two real-valued variables, $\exists$ SCM

$$X := N_X, \qquad Y := f_Y(X, N_Y), \qquad N_Y \perp\!\!\!\perp N_X,$$

where f_Y is a measurable function and N_Y is a real-valued variable.

Proof.

Use the (inverse) conditional cumulative distribution function (CDF)

$$F_{Y|x}(y) := P(Y \leq y | X = x),$$

to define

$$f_Y(x, n_Y) := F^{-1}_{Y|x}(n_Y).$$

Then, let N_Y be uniform on $[0, 1]$ and independent of $N_X = X$. □

Why Additional Assumptions Are Required

Proposition (Non-uniqueness of graph structures)

For every joint distribution $P_{X,Y}$ of two real-valued variables, $\exists$ SCM

$$X := N_X, \qquad Y := f_Y(X, N_Y), \qquad N_Y \perp\!\!\!\perp N_X,$$

where f_Y is a measurable function and N_Y is a real-valued variable.

Proof.

Use the (inverse) conditional cumulative distribution function (CDF)

$$F_{Y|x}(y) := P(Y \leq y | X = x),$$

to define

$$f_Y(x, n_Y) := F^{-1}_{Y|x}(n_Y).$$

Then, let N_Y be uniform on $[0, 1]$ and independent of $N_X = X$. □

Without restricting the function class, both causal directions are possible!

Intuition on Non-Identifiability in the General Case

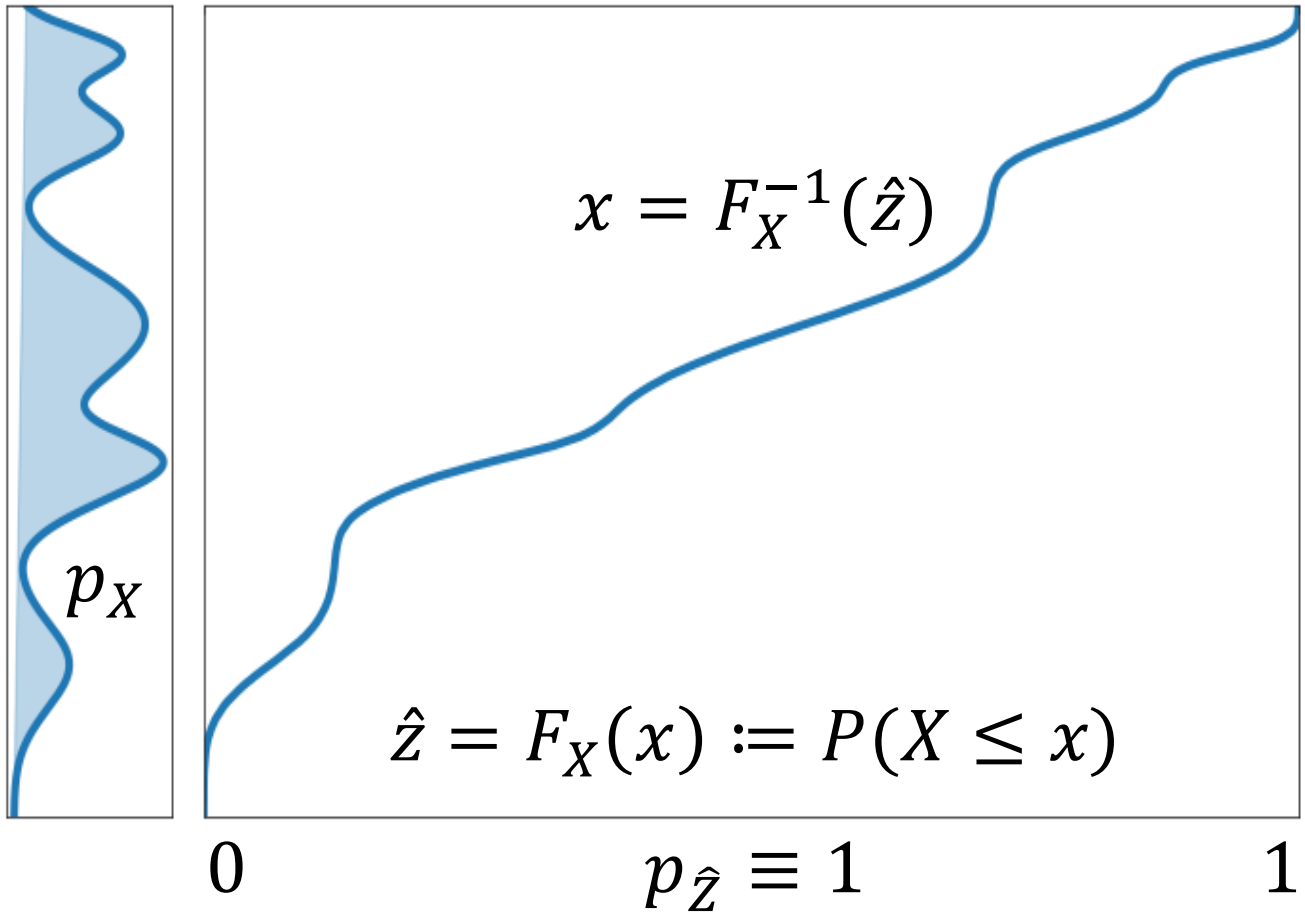


Figure from: Gresele et al. (2021)

Linear Non-Gaussian Acyclic Model (LiNGAM)

Theorem (Identifiability of linear non-Gaussian models)

Assume that $P_{X,Y}$ admits the linear model

$$Y = \alpha X + N_Y, \qquad N_Y \perp\!\!\!\perp X,$$

with continuous X, N_Y, and Y. Then $\exists \beta \in \mathbb{R}$ and N_X s.t.

$$X = \beta Y + N_X, \qquad N_X \perp\!\!\!\perp Y,$$

if and only if N_Y and X are Gaussian.

The multi-variate generalization also holds (Shimizu et al., 2006).

Darmois-Skitovich Theorem

The identifiability result for LiNGAMs follows from the following powerful characterisation of the Gaussian distribution.

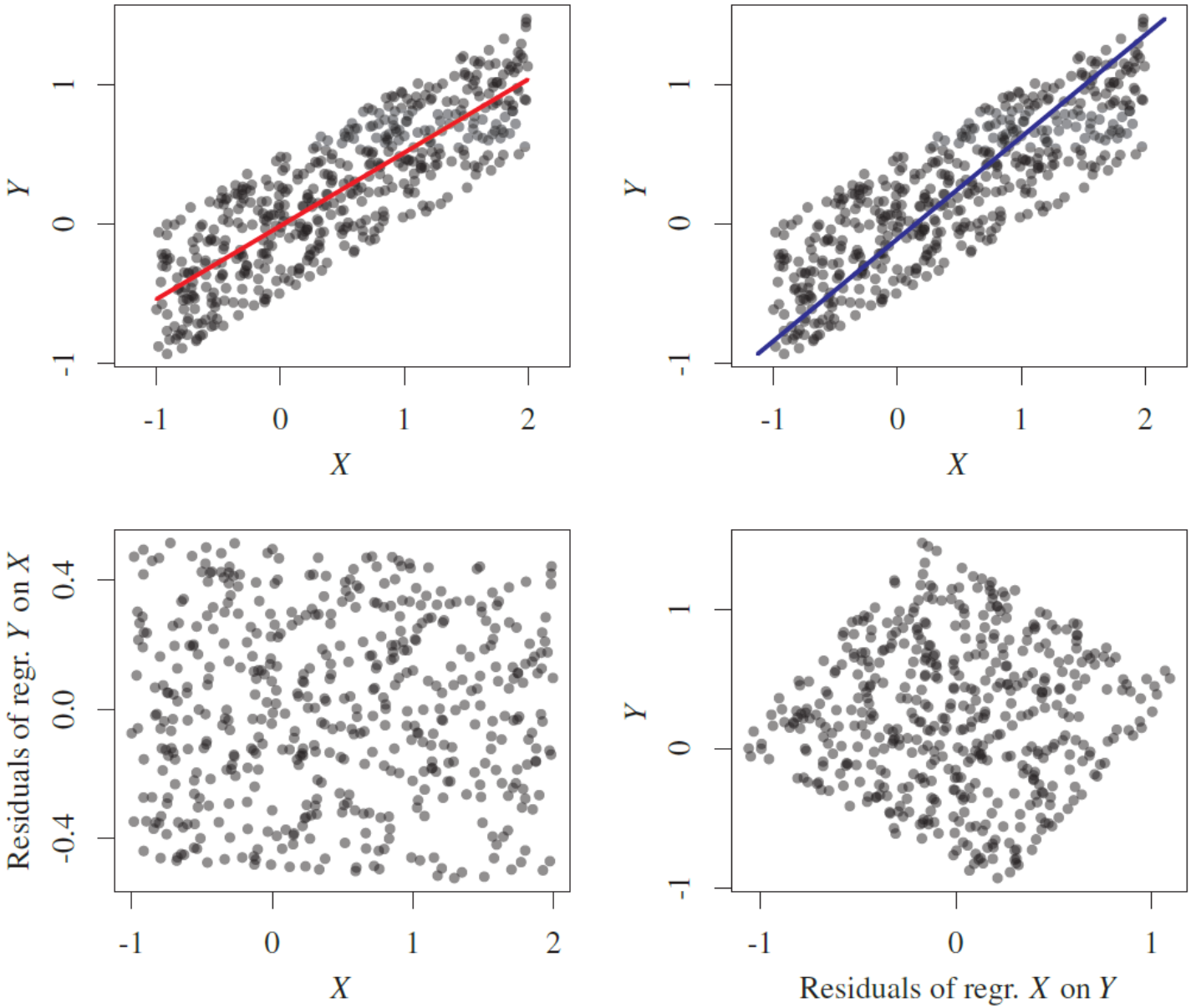
Theorem (Darmois (1953) and Skitovič (1954))

For $k \geq 2$, let $Z_1, ..., Z_k$ be mutually independent, non-degenerate random variables, and let $a_1, ..., a_k$ and $b_1, ..., b_k$ be non-vanishing coefficients ($a_j \neq 0 \neq b_j$ for all j). If the two linear combinations

$$\hat{Z}_1 = \sum_{j=1}^{k} a_j Z_j \,, \qquad \hat{Z}_2 = \sum_{j=1}^{k} b_j Z_j$$

are independent, then all Z_j are Gaussian.

Causal Discovery with LiNGAMs: Intuition



More Complex Additive Noise Models (ANMs)

Similar identifiability results (with considerably more complicated conditions) have been shown for other model classes, such as:

- nonlinear additive noise models (Hoyer et al., 2009),

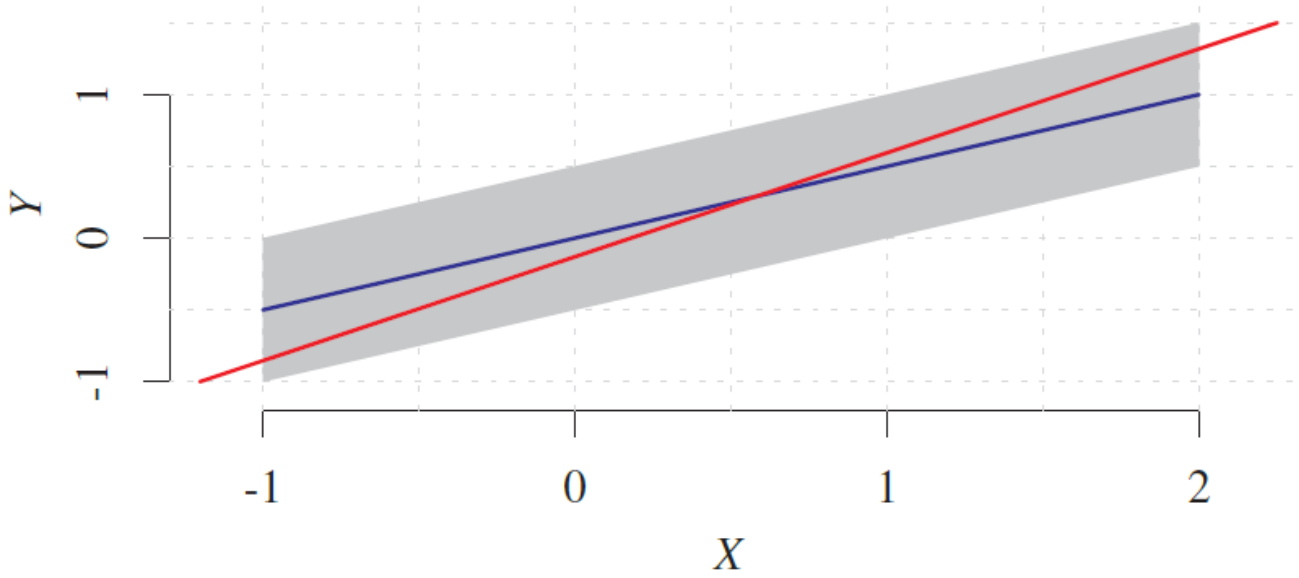
$$Y := f_Y(X) + N_Y, \qquad N_Y \perp\!\!\!\perp X,$$

- post-nonlinear models (Kun Zhang and Hyvärinen, 2009),

$$Y := g_Y(f_Y(X) + N_Y), \qquad N_Y \perp\!\!\!\perp X,$$

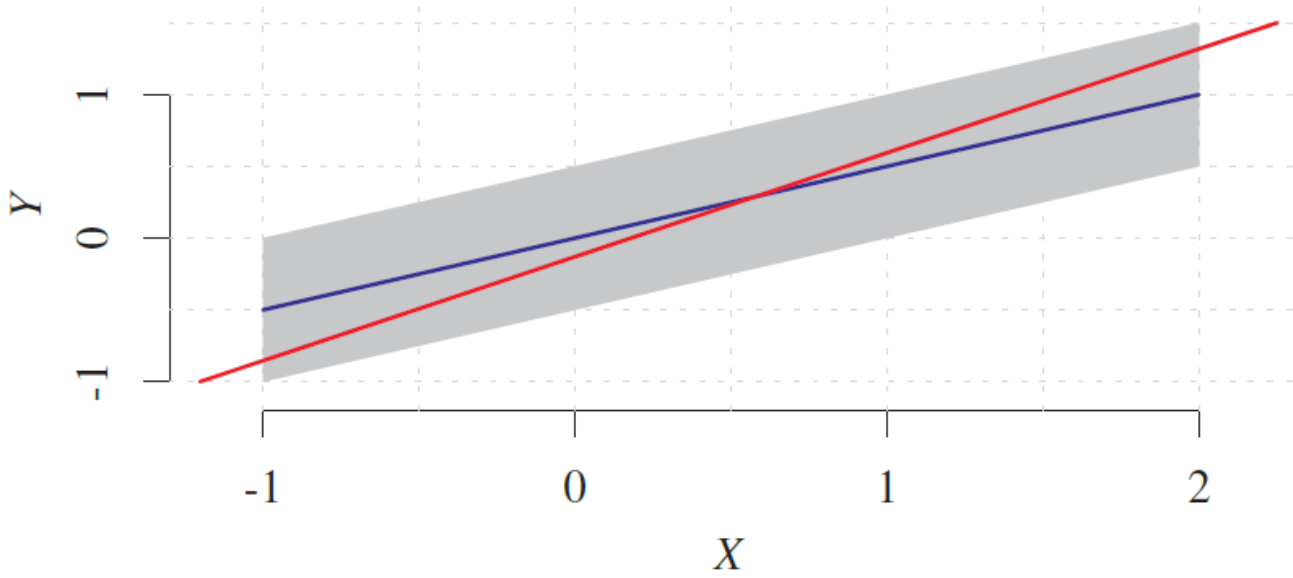
- discrete non-linear ANMs with either X or $Y \in \mathbb{Z}$ (or $\mathbb{Z}/m\mathbb{Z}$).

Causal Discovery with ANMs: Algorithm



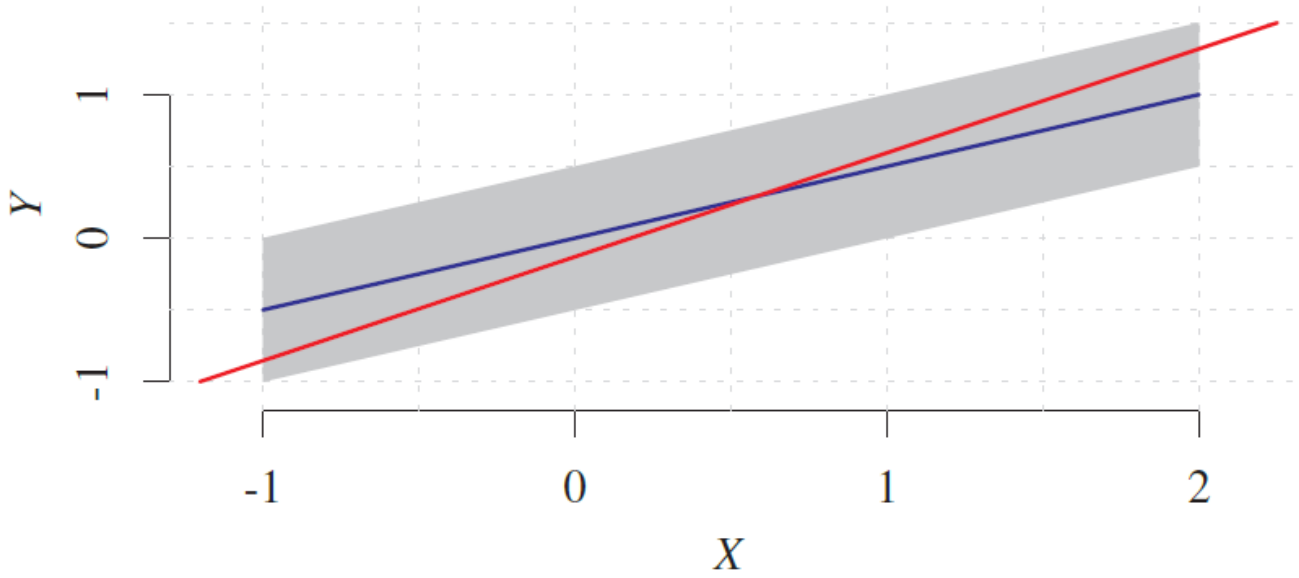
1. Regress Y on X writing $Y = \hat{f}_Y(X) + N_Y$.

Causal Discovery with ANMs: Algorithm



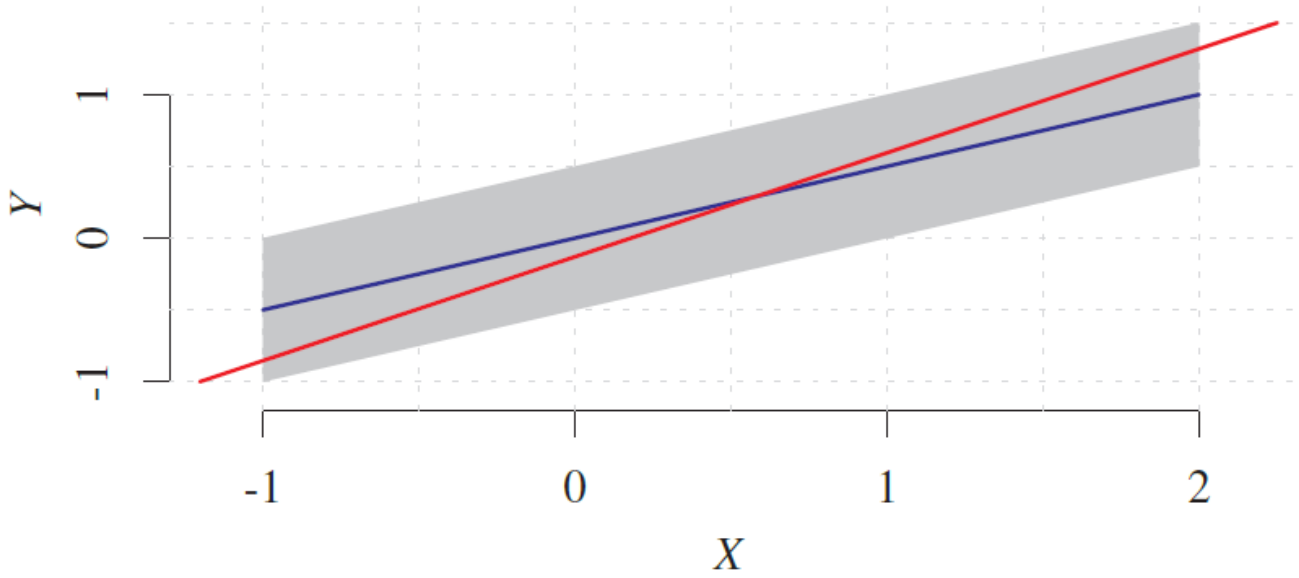
1. Regress Y on X writing $Y = \hat{f}_Y(X) + N_Y$.
2. Test the residuals $N_Y = Y - \hat{f}_Y(X)$ for independence of X.

Causal Discovery with ANMs: Algorithm



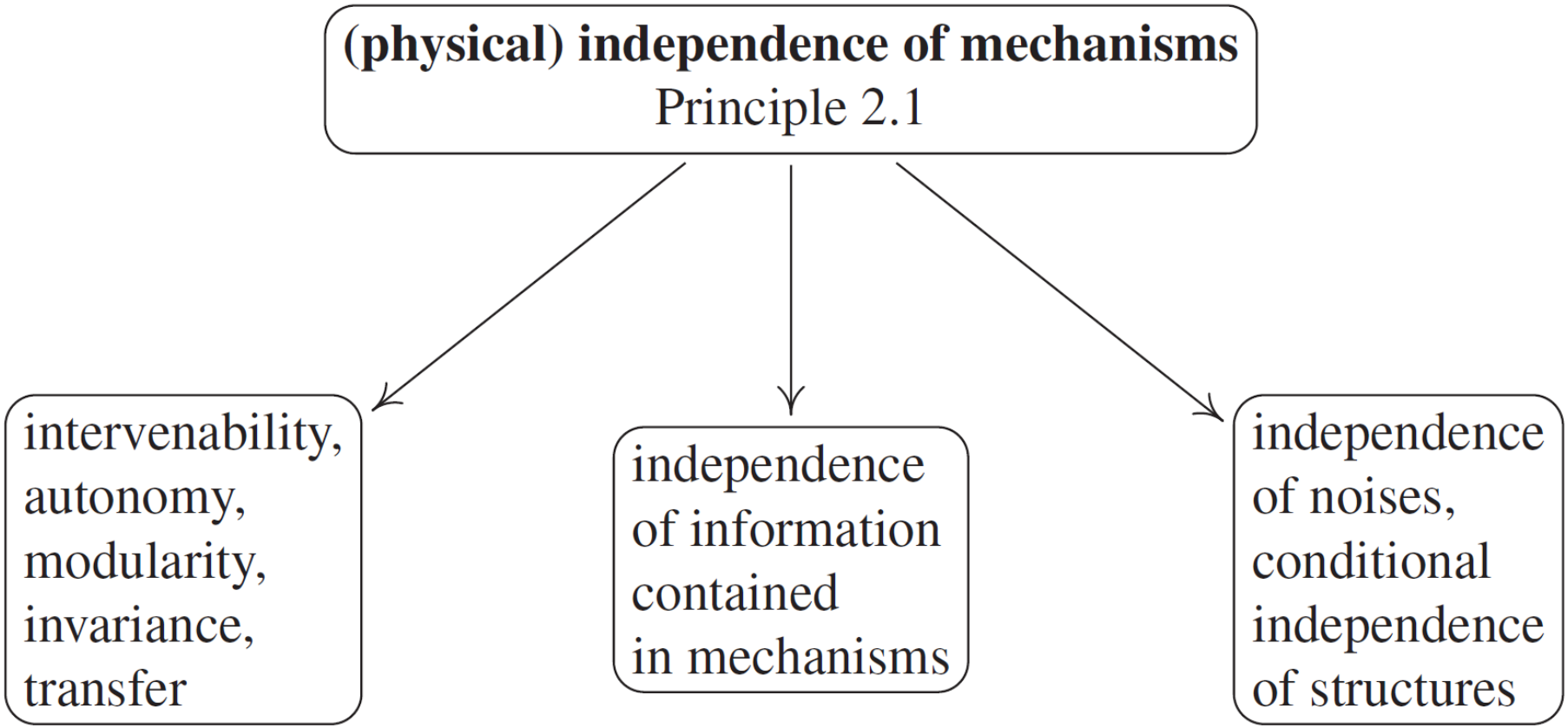
1. Regress Y on X writing $Y = \hat{f}_Y(X) + N_Y$.
2. Test the residuals $N_Y = Y - \hat{f}_Y(X)$ for independence of X.
3. Repeat steps 1 and 2 with the roles of X and Y interchanged.

Causal Discovery with ANMs: Algorithm



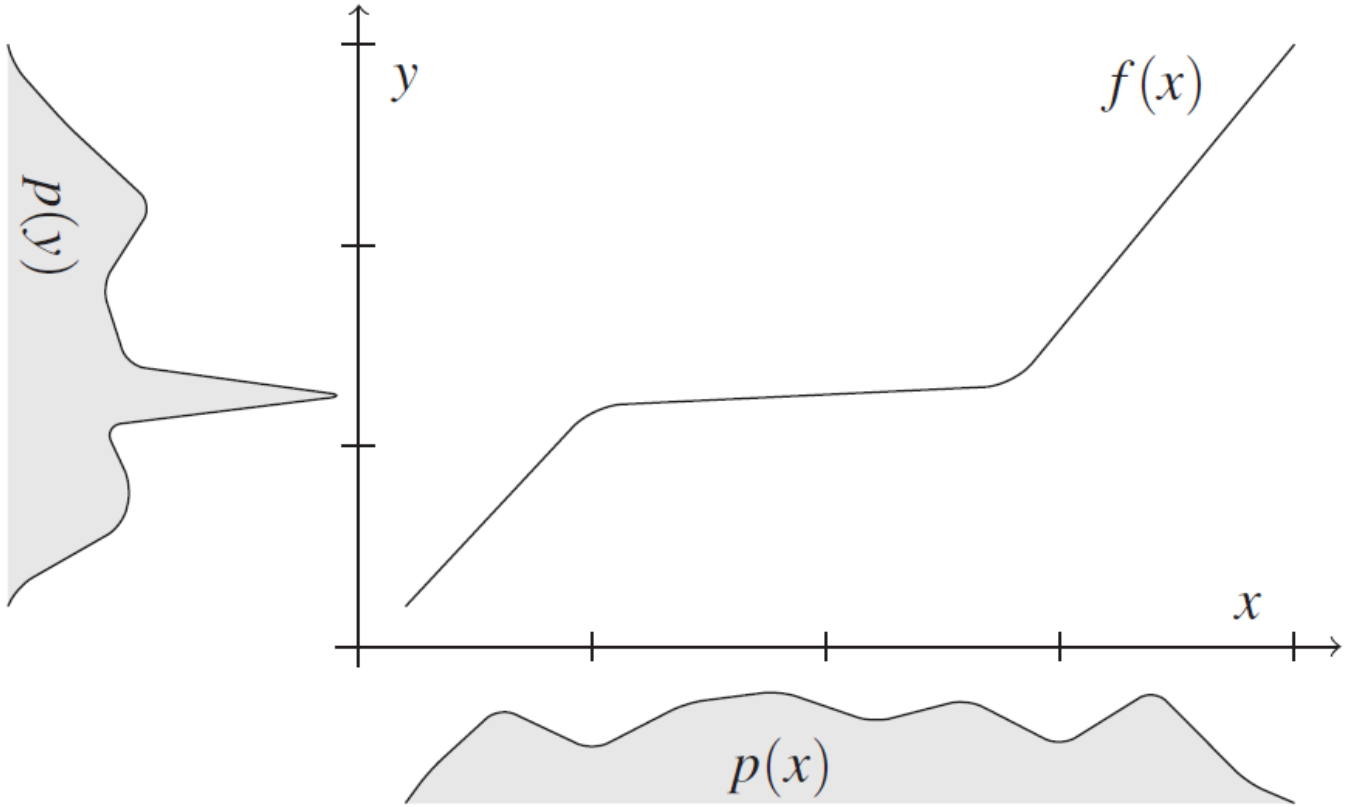
1. Regress Y on X writing $Y = \hat{f}_Y(X) + N_Y$.
2. Test the residuals $N_Y = Y - \hat{f}_Y(X)$ for independence of X.
3. Repeat steps 1 and 2 with the roles of X and Y interchanged.
4. If independence is accepted in one direction and rejected in the other, accept the former as the correct causal one.

Principle of Independent Mechanisms



Information-Geometric Causal Inference (IGCI)

Idea: make use of independence of cause and mechanisms by measuring how much information is shared between them.



Sketch of IGCI: if $p(x)$ and f are chosen independently, then $p(y)$ contains information about f and therefore f^{-1}.

IGCI Definition

Definition (IGCI)

$P_{X,Y}$ is said to satisfy an IGCI model from X to Y if the following hold:

- $Y = f(X)$ for some diffeomorphism[a] f of $[0,1]$ that is strictly monotonic and satisfies $f(0) = 0$ and $f(1) = 1$;
- P_X has the strictly positive continuous density p_X, such that

$$\mathrm{cov}[\log f', p_X] = 0,$$

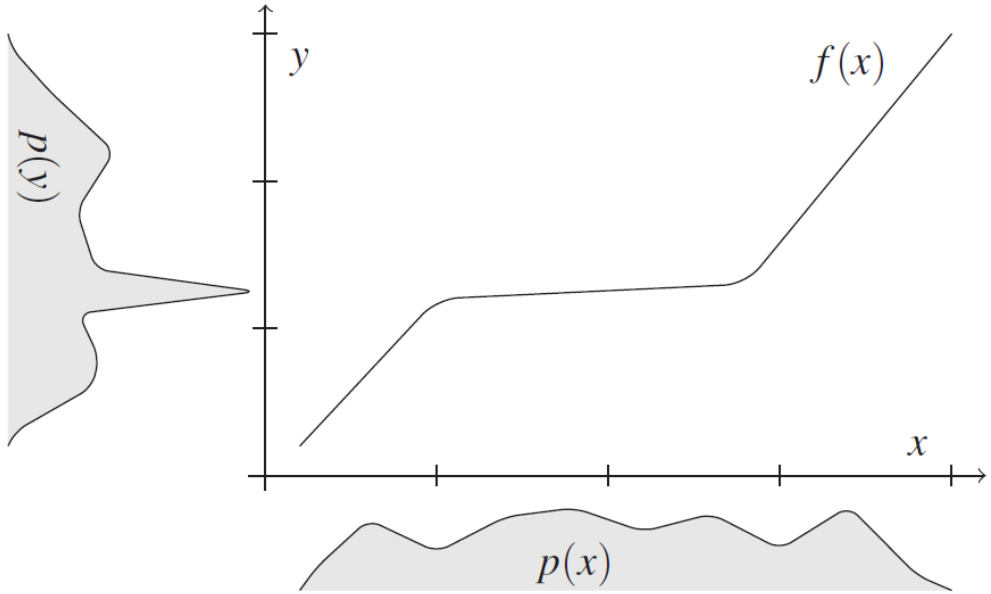
where $\log f'$ and p_X are considered as random variables on the probability space $[0,1]$ endowed with the uniform distribution.

[a] A function is called a diffeomorphism if it is differentiable and bijective and it has a differentiable inverse.

Note that:

$$\mathrm{cov}[\log f', p_X] = \int_0^1 f'(x) p_X(x) \mathrm{d}x - \int_0^1 f'(x) \mathrm{d}x.$$

IGCI: Identifiability Result



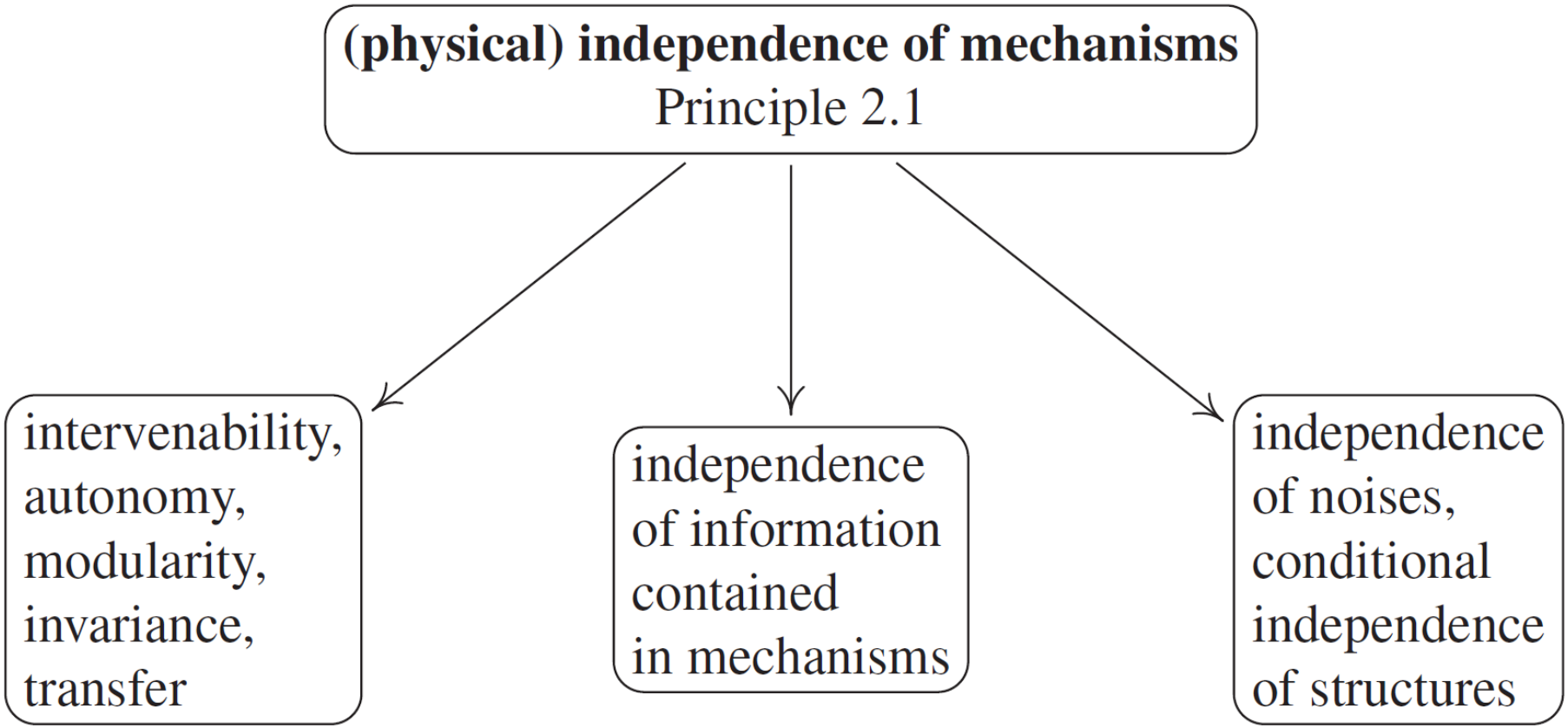
Theorem (Identifiability of IGCI models)

Assume the distribution $P_{X,Y}$ admits an IGCI model from X to Y. Then the inverse function f^{-1} satisfies

$$cov[\log f^{-1'}, p_Y] \geq 0$$

with equality if and only if f is the identity.

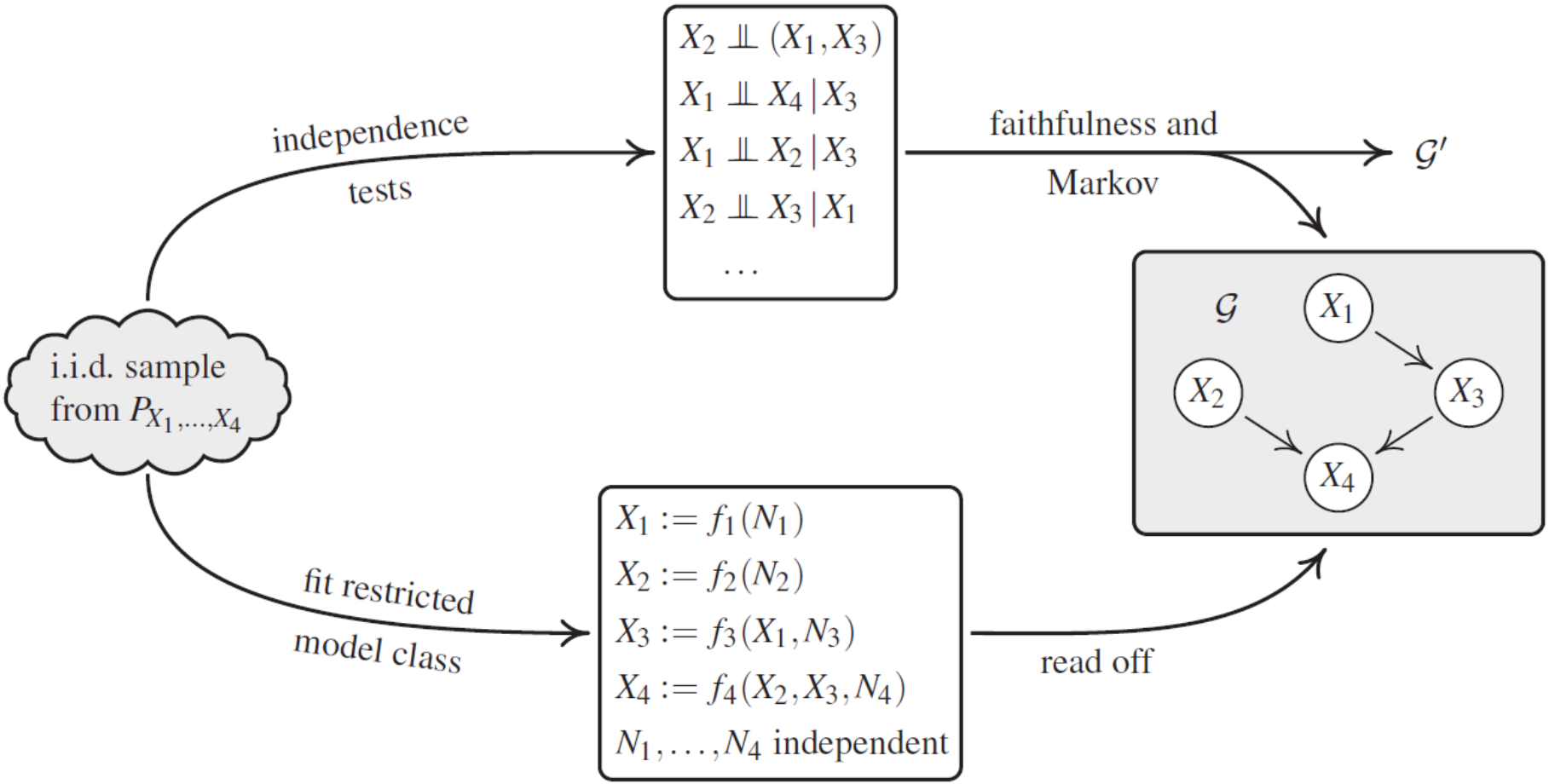
Principle of Independent Mechanisms



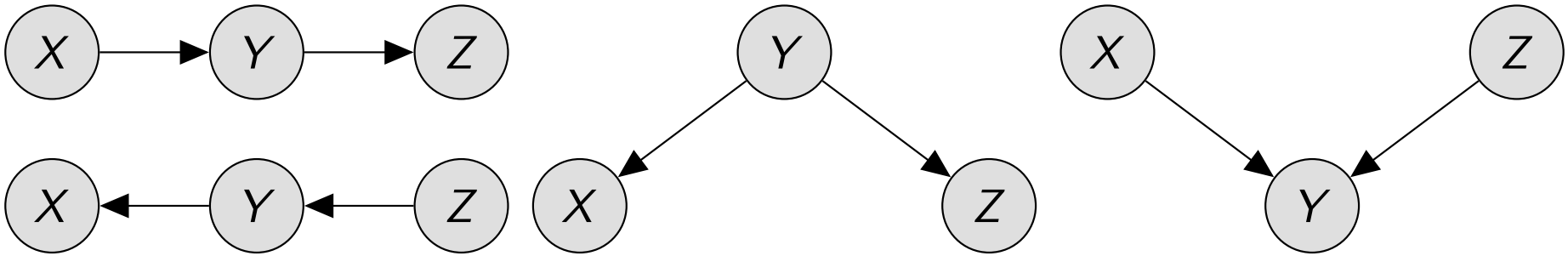
Outline

Learning Multivariate Causal Models: Overview

Two main approaches for learning multivariate causal models *from observational data alone*:



Markov Equivalence of Graphs



(a) Chains (b) Fork (c) Collider / v-structure

(a) and (b) all imply $X \perp\!\!\!\perp Z \mid Y$ (and no others). They thus form a *Markov equivalence class*, meaning they cannot be distinguished using conditional independence testing alone.

(c) implies $X \perp\!\!\!\perp Z$ (but $X \not\perp\!\!\!\perp Z \mid Y$) and forms its own Markov equivalence class, so it can be uniquely identified from observational data.

$\rightarrow$ v-structures are helpful for causal discovery: two graphs are Markov equivalent iff. they share the same skeleton and v-structures.

Faithfulness

Assumption (Faithfulness)

We say that p is faithful to G if the only (conditional) independences satisfied by p are those implied by G (via d-separation).

Faithfulness can be seen as the converse of the causal Markov condition. Together, they constitute a one-to-one correspondence between graphical separation in G and conditional independence in p.

Example (Violation of faithfulness)

$$X_1 := N_1, \qquad X_2 := \alpha X_1 + U_2, \qquad X_3 := \beta X_1 + \gamma X_2 + U_3$$

with $N_1, N_2, N_3 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. By substitution, we obtain

$$X_3 = (\beta + \alpha\gamma) X_1 + \gamma N_2 + N_3.$$

Hence, $X_3 \perp\!\!\!\perp X_1$ whenever $\beta + \alpha\gamma = 0$.

Multivariate Models via Restricting the Function Class

Overview of identifiability results and their conditions:

Type of structural assignment		Condition on funct.	DAG identif.
(General) SCM:	$X_j := f_j(X_{\mathbf{PA}_j}, N_j)$	—	✗
ANM:	$X_j := f_j(X_{\mathbf{PA}_j}) + N_j$	nonlinear	✓
CAM:	$X_j := \sum_{k \in \mathbf{PA}_j} f_{jk}(X_k) + N_j$	nonlinear	✓
Linear Gaussian:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✗
Lin. G., eq. error var.:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✓

Score-Based Causal Discovery

Assign a score to each graph G from a set of candidate graphs (usually the set of all DAGs).

The score S is supposed to reflect how well G explains the observed data $\mathbf{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$.

$$\hat{G} = \arg\max_G S(G \mid \mathbf{D}).$$

Examples:

$$S_{\text{BIC}}(G \mid \mathbf{D}) = \log p(\mathbf{D} \mid G, \hat{\theta}^{\text{MLE}}) - \frac{k}{2} \log m$$

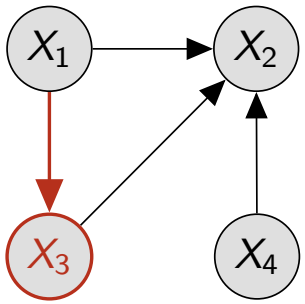
$$S_{\text{BAYES}}(G \mid \mathbf{D}) = p(\mathbf{D} \mid G) = \int_\Theta p(\mathbf{D} \mid G, \theta) p(\theta \mid G) \mathrm{d}\theta.$$

Drawback: exponential search space; e.g., the number of DAGs for $n = 5$ and $n = 10$ nodes is 29281 and 4175098976430598143, respectively.

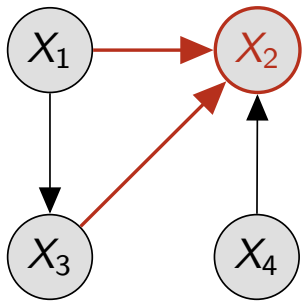
Learning from Sparse Domain Shifts[5]

Given datasets $\mathcal{D}^e \sim P^e_{\boldsymbol{X}}$ over observables $\boldsymbol{X} = \{X_1, ..., X_d\}$, resulting from *soft interventions* on an *unknown* subset $\mathcal{I}^e$ of mechanisms:

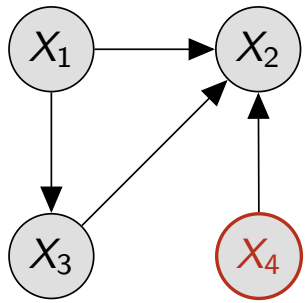
$$P^e_{\boldsymbol{X}}(X_1, ..., X_d) = \prod_{j \in \mathcal{I}^e} \underbrace{P^e_{\boldsymbol{X}}(X_j \mid \mathbf{Pa}_j)}_{\text{Changed mechanism}} \prod_{j \in [d] \setminus \mathcal{I}^e} \underbrace{P_{\boldsymbol{X}}(X_j \mid \mathbf{Pa}_j)}_{\text{Base mechanism}}.$$



$\mathcal{D}^1$



$\mathcal{D}^2$



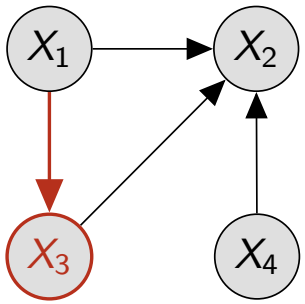
$\mathcal{D}^3$

[5] Perry et al., 2022.

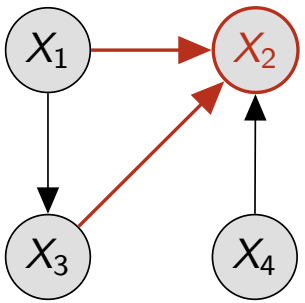
Learning from Sparse Domain Shifts[5]

Given datasets $\mathcal{D}^e \sim P^e_{\boldsymbol{X}}$ over observables $\boldsymbol{X} = \{X_1, ..., X_d\}$, resulting from *soft interventions* on an *unknown* subset $\mathcal{I}^e$ of mechanisms:

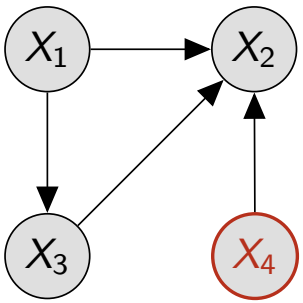
$$P^e_{\boldsymbol{X}}(X_1, ..., X_d) = \prod_{j \in \mathcal{I}^e} \underbrace{P^e_{\boldsymbol{X}}(X_j \mid \mathbf{Pa}_j)}_{\text{Changed mechanism}} \prod_{j \in [d] \setminus \mathcal{I}^e} \underbrace{P_{\boldsymbol{X}}(X_j \mid \mathbf{Pa}_j)}_{\text{Base mechanism}}.$$



$\mathcal{D}^1$



$\mathcal{D}^2$



$\mathcal{D}^3$

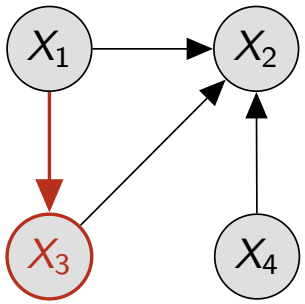
\+ *Sparse Mechanism Shift Hypothesis* (Schölkopf et al., 2021)

[5] Perry et al., 2022.

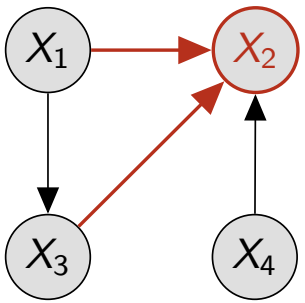
Learning from Sparse Domain Shifts[5]

Given datasets $\mathcal{D}^e \sim P^e_{\boldsymbol{X}}$ over observables $\boldsymbol{X} = \{X_1, ..., X_d\}$, resulting from *soft interventions* on an *unknown* subset $\mathcal{I}^e$ of mechanisms:

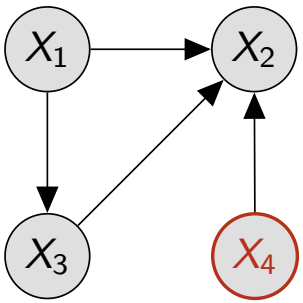
$$P^e_{\boldsymbol{X}}(X_1, ..., X_d) = \prod_{j \in \mathcal{I}^e} \underbrace{P^e_{\boldsymbol{X}}(X_j \mid \mathbf{Pa}_j)}_{\text{Changed mechanism}} \prod_{j \in [d] \setminus \mathcal{I}^e} \underbrace{P_{\boldsymbol{X}}(X_j \mid \mathbf{Pa}_j)}_{\text{Base mechanism}}.$$



$\mathcal{D}^1$



$\mathcal{D}^2$



$\mathcal{D}^3$

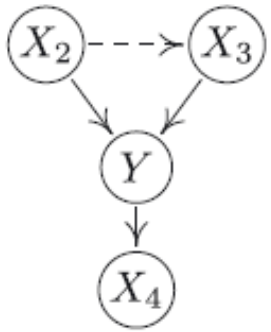
\+ *Sparse Mechanism Shift Hypothesis* (Schölkopf et al., 2021)

Thm: observed $\boldsymbol{X}$ + sufficiently many environments $\implies$ true DAG G

[5] Perry et al., 2022.

Causal Inference by Using Invariant Prediction: Idea

$$X_2^1 = 0.3\varepsilon_2^1$$
$$X_3^1 = X_2^1 + 0.2\varepsilon_3^1$$
$$Y^1 = -0.7X_2^1 + 0.6X_3^1 + 0.1\varepsilon_Y^1$$
$$X_4^1 = -0.5Y^1 + 0.5X_3^1 + 0.1\varepsilon_4^1$$
$$\varepsilon_Y^1, \varepsilon_2^1, \varepsilon_3^1, \varepsilon_4^1 \overset{iid}{\sim} \mathcal{N}(0,1)$$

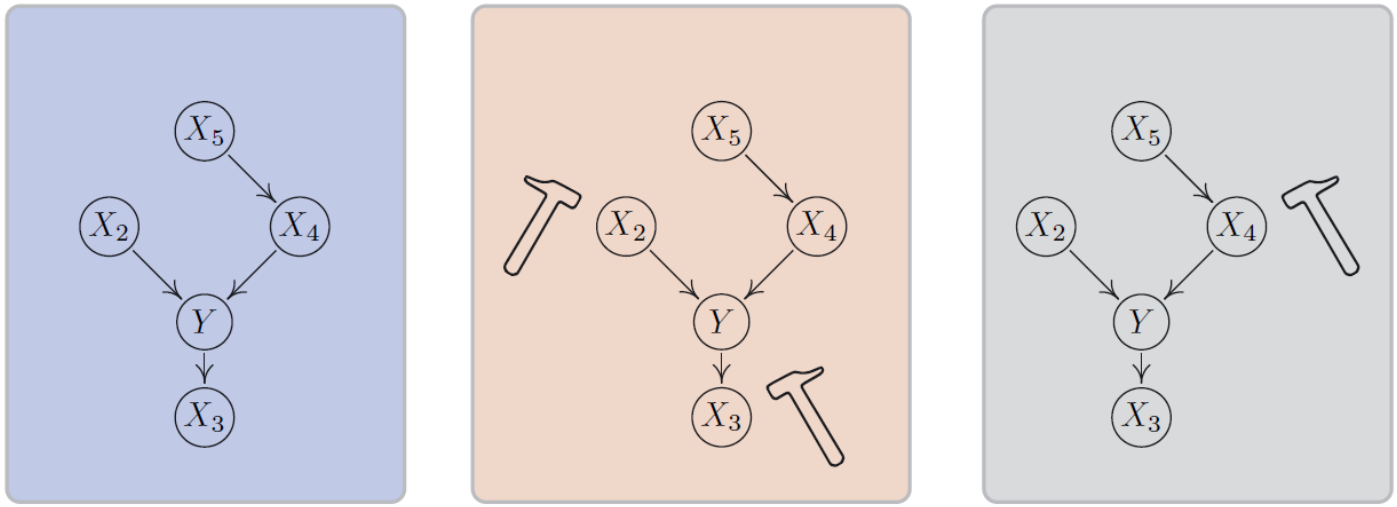


$$X_2^2 = 0.3\varepsilon_2^2$$
$$X_3^2 = 0.4\varepsilon_3^2$$
$$Y^2 = -0.7X_2^2 + 0.6X_3^2 + 0.1\varepsilon_Y^2$$
$$X_4^2 = -0.5Y^2 + 0.5X_3^2 + 0.1\varepsilon_4^2$$
$$\varepsilon_Y^2, \varepsilon_2^2, \varepsilon_3^2, \varepsilon_4^2 \overset{iid}{\sim} \mathcal{N}(0,1)$$

"What is the difference between a prediction that is made with a causal model and that with a non-causal model? Suppose that we intervene on the predictor variables or change the whole environment. The predictions from a causal model will in general work as well under interventions as for observational data. [...] We propose to exploit this invariance of a prediction under a causal model for causal inference."[6]

[6] Peters et al., 2016.

Causal Inference by Using Invariant Prediction: Details



Q: Which of d predictors $\boldsymbol{X}$ are the causal parents of a target variable Y?

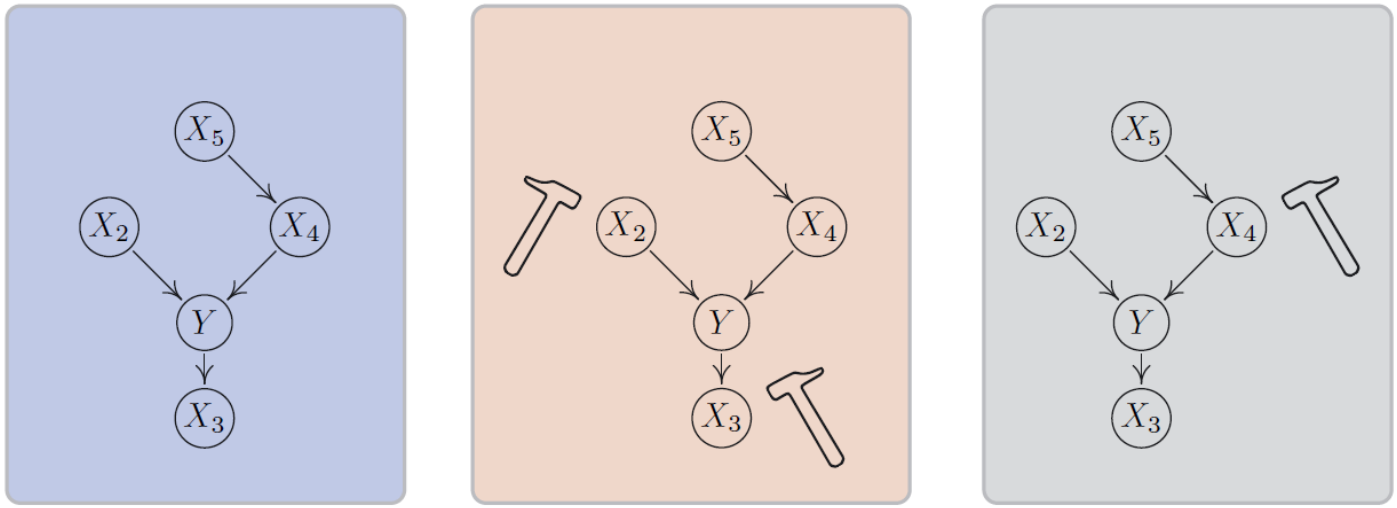
Assume that we are given data from different environments/interventions,

$$(\boldsymbol{X}^e, Y^e) \sim P_{\boldsymbol{X}^e, Y^e} \quad \text{for} \quad e \in \mathcal{E}.$$

Provided Y was not intervened on, its parents are invariant predictors,

$$P_{Y^e|\boldsymbol{PA}_Y^e} = P_{Y^f|\boldsymbol{PA}_Y^f} \quad \forall e, f \in \mathcal{E}.$$

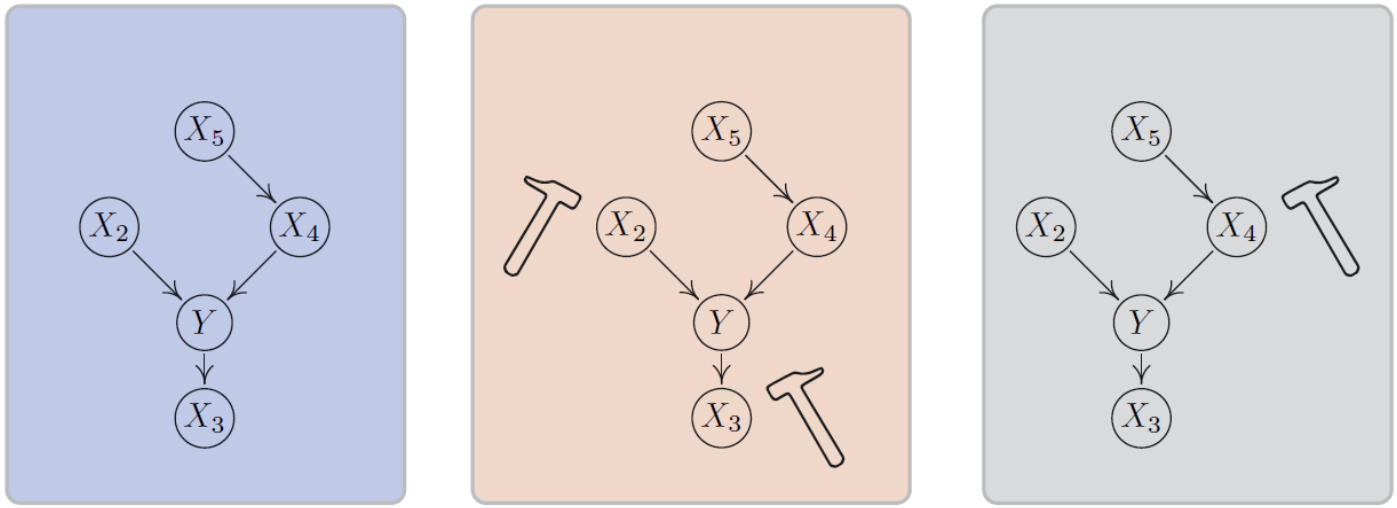
Causal Inference by Using Invariant Prediction: Details



Consider the collection $\mathcal{S}$ of all sets $S \subseteq \{1, ..., d\}$ of variables leading to invariant prediction (at significance level α),

$$P_{Y^e|\boldsymbol{X}_S^e} = P_{Y^f|\boldsymbol{X}_S^f} \quad \forall e, f \in \mathcal{E} \quad \text{and} \quad \forall S \in \mathcal{S}.$$

Causal Inference by Using Invariant Prediction: Details



Consider the collection $\mathcal{S}$ of all sets $S \subseteq \{1, ..., d\}$ of variables leading to invariant prediction (at significance level α),

$$P_{Y^e|\boldsymbol{X}_S^e} = P_{Y^f|\boldsymbol{X}_S^f} \quad \forall e, f \in \mathcal{E} \quad \text{and} \quad \forall S \in \mathcal{S}.$$

Then the variables appearing in all $S \in \mathcal{S}$ are direct causes of Y with high probability $(1 - \alpha)$,

$$\bigcap_{S \in \mathcal{S}} S \subseteq \boldsymbol{PA}_Y.$$

Outline

Causal Time Series

Multivariate time series $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ as stationary stochastic process.

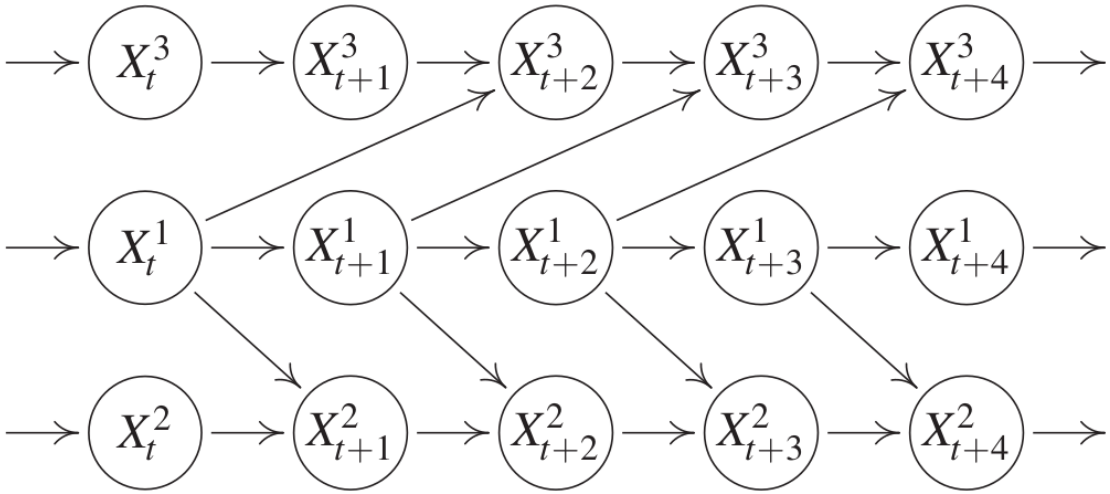


Figure 10.1: Example of a time series with no instantaneous effects.

Often interested in summary graph:

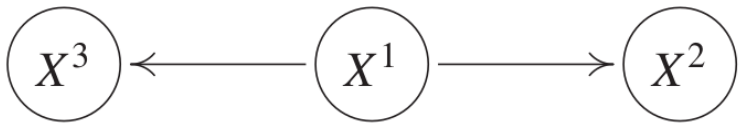


Figure 10.3: Summary graph of the full time graphs shown in Figures 10.1 and 10.2.

Granger Causality

Granger Causality $\leftarrow$ Conditional (In)dependence + Direction of Time (no mention of intervention!)

"X has causal influence on Y whenever past values of X help in predicting Y from its own past." Formally:

$$X \quad \text{Granger-causes} \quad Y \iff Y_t \not\perp\!\!\!\perp X_{\text{past}(t)} \,|\, Y_{\text{past}(t)}$$

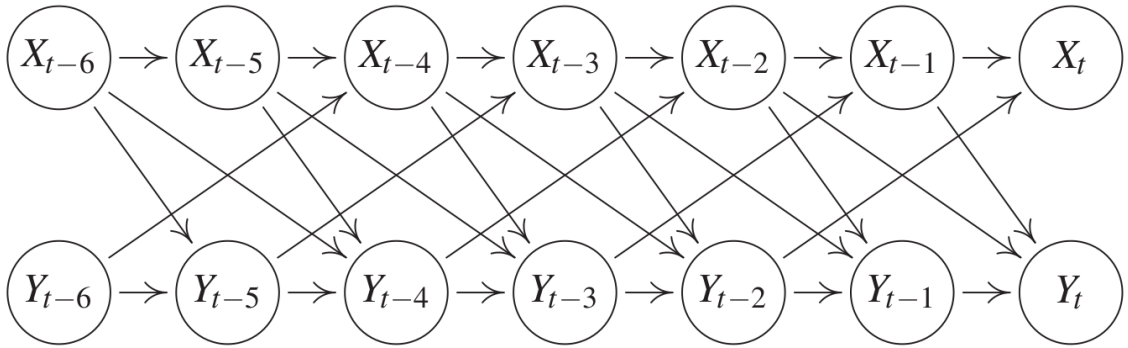


Figure 10.6: Typical scenario, in which Granger causality works: if all arrows from X to Y were missing, Y_t would be conditionally independent of the past values of X, given its own past. Here, Y_t does depend on the past values of X, given its own past. Thus, condition (10.4) proves the existence of an influence from X to Y.

Granger Causality

Granger Causality $\leftarrow$ Conditional (In)dependence + Direction of Time (no mention of intervention!)

"X has causal influence on Y whenever past values of X help in predicting Y from its own past." Formally:

$$X \quad \text{Granger-causes} \quad Y \iff Y_t \not\perp\!\!\!\perp X_{\text{past}(t)} \,|\, Y_{\text{past}(t)}$$

In practice: fit two regression models, e.g., in the linear case,

$$Y_t = \sum_{i=1}^{q} a_i Y_{t-i} + N_t$$

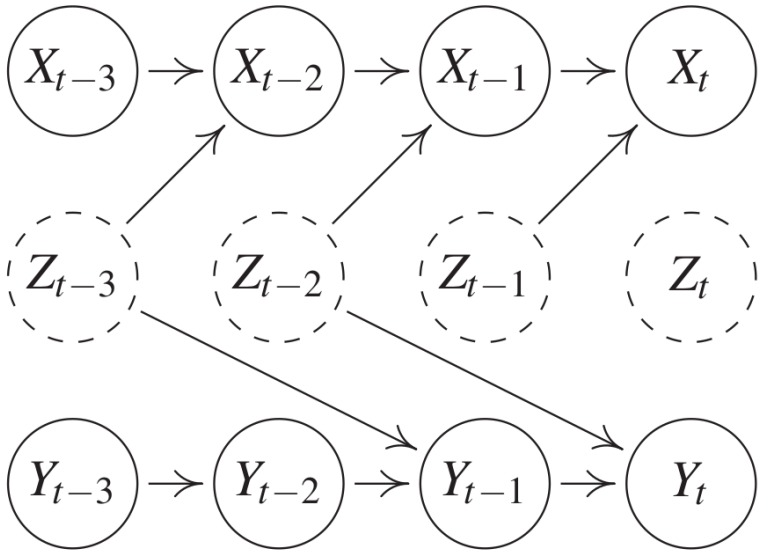
$$Y_t = \sum_{i=1}^{q} a_i Y_{t-i} + \sum_{i=1}^{q} b_i X_{t-i} + M_t$$

and compare the variances of N_t and M_t.

Limitations of Granger Causality: Hidden Confounders

Important to condition on all relevant variables! Multivariate version:

$$X^j \quad \text{Granger-causes} \quad X^k \iff X_t^k \not\perp\!\!\!\perp X^j_{\text{past}(t)} | \mathbf{X}^{-j}_{\text{past}(t)}$$

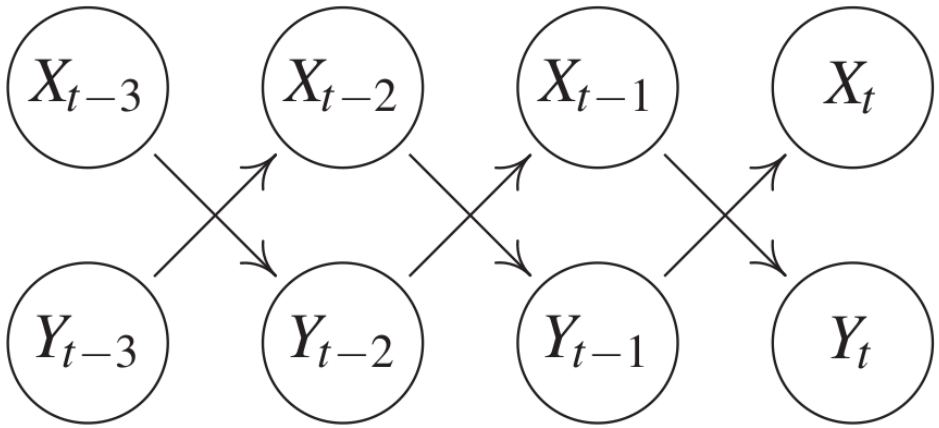


(a) Due to the hidden common cause Z, Granger causality erroneously infers causal influence from X to Y.

Example: X = price of butter; Z = price of milk; Y = price of cheese

Limitations of Granger Causality: Deterministic Relations

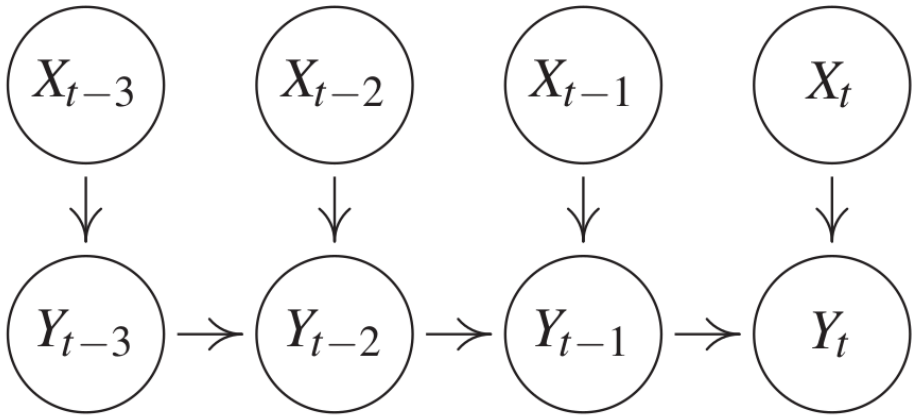
Determinism introduces additional independences $\rightarrow$ problematic for conditional independence-based causal inference.



(b) Granger causality erroneously infers neither causal influence from X to Y nor from Y to X if the influence from X_t on Y_{t+1} and the one from Y_t to X_{t+1} are deterministic.

Limitations of Granger Causality: Instantaneous Effects

If time resolution is not fine enough, "instantaneous" (up to our resolution) effects of the form $X_t \to Y_t$ may not be picked up by Granger causality.



(a) Granger causality cannot detect the influence of X on Y because the past of X influences Y_t only via the past of Y.

Takeaways

- Time ordering makes causal reasoning easier by imposing ordering constraints.
- Sometimes skeleton and temporal ordering are enough.
- Care should be taken with regard to hidden confounders, deterministic relations and instantaneous effects.
- If possible, best to combine temporal and interventional information.

Outline

Machine Learning for Causal Inference: Overview

In practice, causal inference from finite data requires methods for:

- regression, i.e., estimating a function such as a conditional mean

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, do(T = t)] \approx \hat{f}(\boldsymbol{x}, t)$$

  e.g., for regression adjustment or propensity score weighting

- (conditional) density estimation,

$$\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(m)} \sim P^*(\boldsymbol{Z}) \approx P_\theta(\boldsymbol{Z})$$

  e.g,., for covariate adjustment or assessing overlap.

Machine Learning for Causal Inference: Overview

In practice, causal inference from finite data requires methods for:

- regression, i.e., estimating a function such as a conditional mean

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, do(T = t)] \approx \hat{f}(\boldsymbol{x}, t)$$

  e.g., for regression adjustment or propensity score weighting

- (conditional) density estimation,

$$\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(m)} \sim P^*(\boldsymbol{Z}) \approx P_\theta(\boldsymbol{Z})$$

  e.g,., for covariate adjustment or assessing overlap.

ML methods are particularly useful for nonlinear, high-dimensional settings.

Machine Learning for Causal Inference: Overview

In practice, causal inference from finite data requires methods for:

- regression, i.e., estimating a function such as a conditional mean

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, do(T = t)] \approx \hat{f}(\boldsymbol{x}, t)$$

  e.g., for regression adjustment or propensity score weighting

- (conditional) density estimation,

$$\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(m)} \sim P^*(\boldsymbol{Z}) \approx P_\theta(\boldsymbol{Z})$$

  e.g,., for covariate adjustment or assessing overlap.

ML methods are particularly useful for nonlinear, high-dimensional settings.

Other applications include, e.g., kernel-based conditional independence tests for constraint-based causal discovery (Zhang et al., 2011).

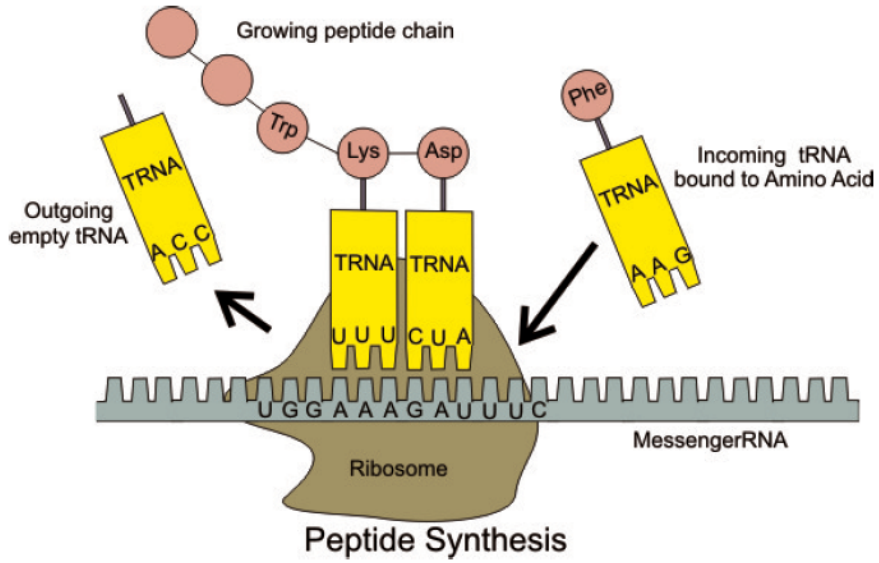
Causality for Machine Learning: Overview

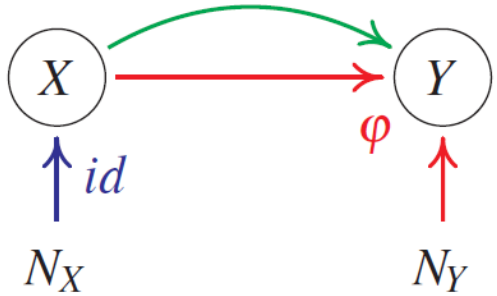
Causal models provide a factorisation of a complex system into independent and autonomous modules. This can be helpful for:

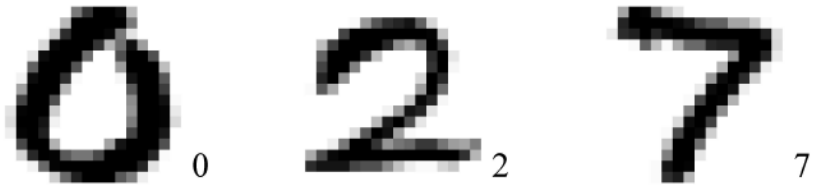
- extracting shared information from unlabelled data, e.g., in semi-supervised learning
- adapting to distribution shifts, e.g., in domain adaptation
- transferring knowledge, e.g., in continual or multi-task learning
- planning and reasoning, e.g., in algorithmic fairness, recourse, interpretability, or explainability

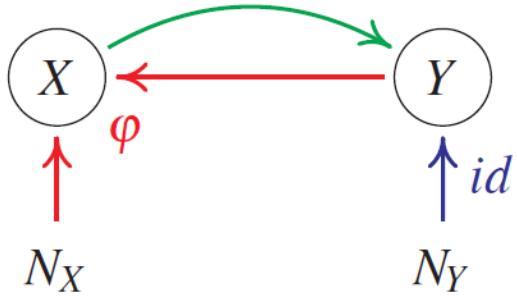
Causal and Anticausal Learning (Schölkopf et al., 2012)

In a causal learning setting (top) the direction of prediction (green arrow) is aligned with the causal direction (red arrows).









In an anticausal learning setting (bottom) this is not the case.

Semi-Supervised Learning (SSL)

In semi-supervised learning (SSL) we are given additional unlabelled samples $\{x_1, ..., x_m\}$ from P_X.

Q: Can we improve our estimate of $P_{Y|X}$ via a better estimate of P_X (obtainable from the additional unlabelled data)?

[7] even if P_X does not tell us anything about $P_{Y|X}$, knowing P_X can still help for better estimating Y in the sense that we obtain lower risk in a learning scenario

Semi-Supervised Learning (SSL)

In semi-supervised learning (SSL) we are given additional unlabelled samples $\{x_1, ..., x_m\}$ from P_X.

Q: Can we improve our estimate of $P_{Y|X}$ via a better estimate of P_X (obtainable from the additional unlabelled data)?

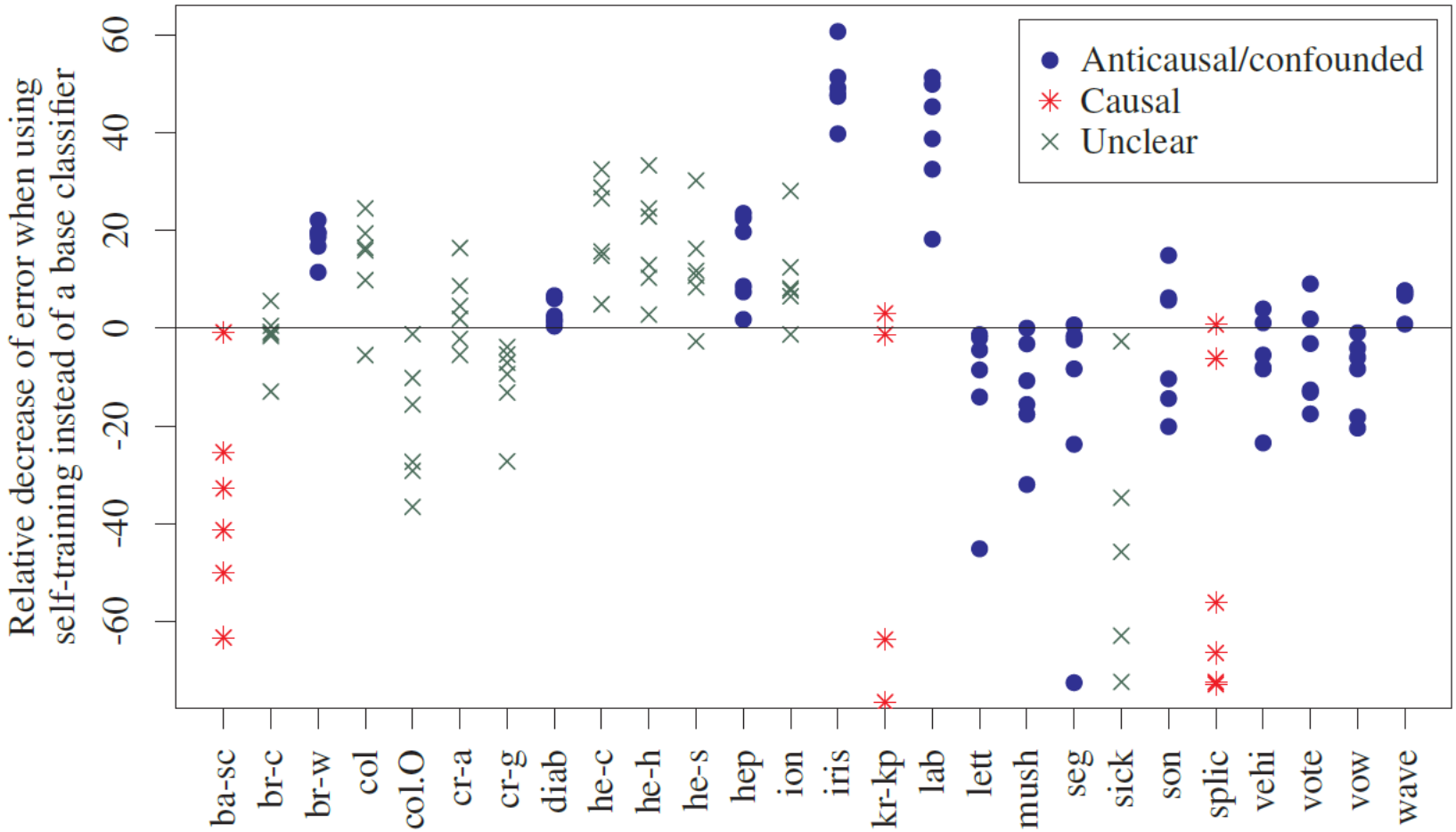
A: According to the *independence of cause and mechanism*:

- if the learning task is causal ($X \to Y$) then P_X and $P_{Y|X}$ share no information, and SSL should not be possible[7];
- if the learning task is anticausal ($Y \to X$) then P_Y and $P_{X|Y}$ are independent, but P_X and $P_{Y|X}$ may share information, so SSL is in principle possible.

[7] even if P_X does not tell us anything about $P_{Y|X}$, knowing P_X can still help for better estimating Y in the sense that we obtain lower risk in a learning scenario

SSL: Empirical Results

Results of SSL on benchmark data sets, grouped by causal structure. Data points correspond to different base classifiers.



Domain Adaptation & Covariate Shift

In the domain adaptation setting we are given labelled data from a source distribution,

$$(\boldsymbol{X}^S, Y^S) \sim P_{\boldsymbol{X}^S, Y^S} =: P^S,$$

and wish to make predictions for a target distribution P^T for which only unlabelled data is available,

$$\boldsymbol{X}^T \sim P_{\boldsymbol{X}^T}.$$

Domain Adaptation & Covariate Shift

In the domain adaptation setting we are given labelled data from a source distribution,

$$(\boldsymbol{X}^S, Y^S) \sim P_{\boldsymbol{X}^S, Y^S} =: P^S,$$

and wish to make predictions for a target distribution P^T for which only unlabelled data is available,

$$\boldsymbol{X}^T \sim P_{\boldsymbol{X}^T}.$$

A commonly used and well-studied assumption is *covariate shift*,

$$P_{\boldsymbol{X}^S} \neq P_{\boldsymbol{X}^T}, \quad \text{but} \quad P_{Y^S|\boldsymbol{X}^S} = P_{Y^T|\boldsymbol{X}^T} =: P_{Y|\boldsymbol{X}}.$$

Domain Adaptation & Covariate Shift

In the domain adaptation setting we are given labelled data from a source distribution,

$$(\boldsymbol{X}^S, Y^S) \sim P_{\boldsymbol{X}^S, Y^S} =: P^S,$$

and wish to make predictions for a target distribution P^T for which only unlabelled data is available,

$$\boldsymbol{X}^T \sim P_{\boldsymbol{X}^T}.$$

A commonly used and well-studied assumption is *covariate shift*,

$$P_{\boldsymbol{X}^S} \neq P_{\boldsymbol{X}^T}, \quad \text{but} \quad P_{Y^S|\boldsymbol{X}^S} = P_{Y^T|\boldsymbol{X}^T} =: P_{Y|\boldsymbol{X}}.$$

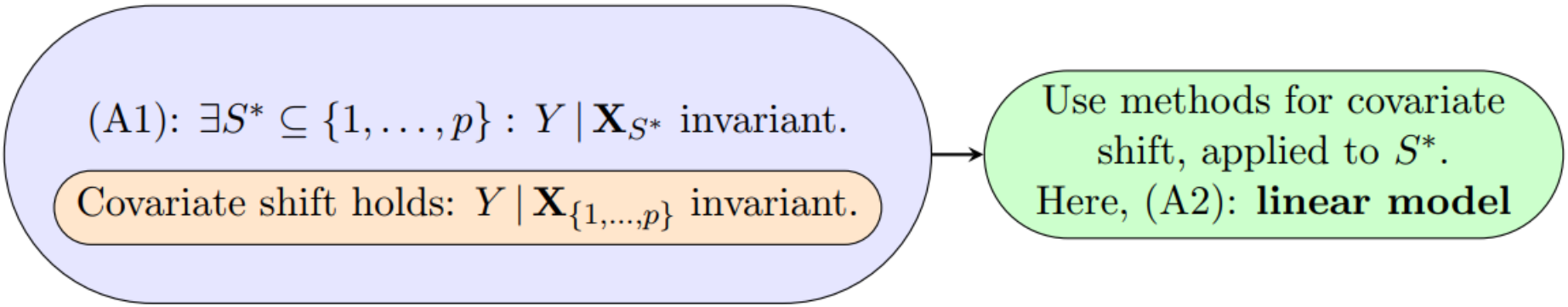
According to the *independence of cause and mechanism*, this is only justified in a causal learning setting, $\boldsymbol{X} \to Y$, where a change in $P_{\boldsymbol{X}}$ does not influence or inform $P_{Y|\boldsymbol{X}}$.

Invariant Models for Causal Transfer Learning

Rojas-Carulla et al. (2018) consider different transfer learning tasks,

Method	Training data from	Test domain
Domain generalization	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D+1$
Multi-task learning	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T \in \{1, \ldots, D\}$
Asymmetric multi-task learning	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D$

and relax the covariate shift assumption to hold only for a subset of variables S^* which are invariant for prediction.



They prove that in this set-up, predicting Y using only $\boldsymbol{X}_{S^*}$ is optimal in an adversarial setting.

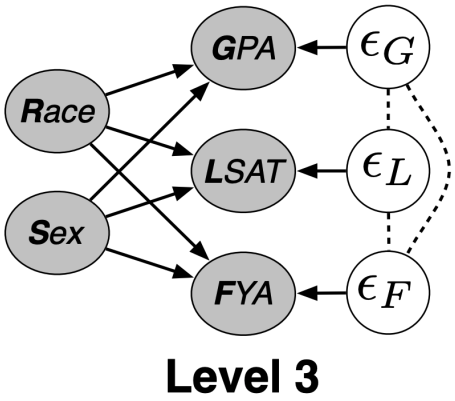
Algorithmic Fairness[8]

Let A be a protected/sensitive attribute (e.g., race, sex, age, religion).

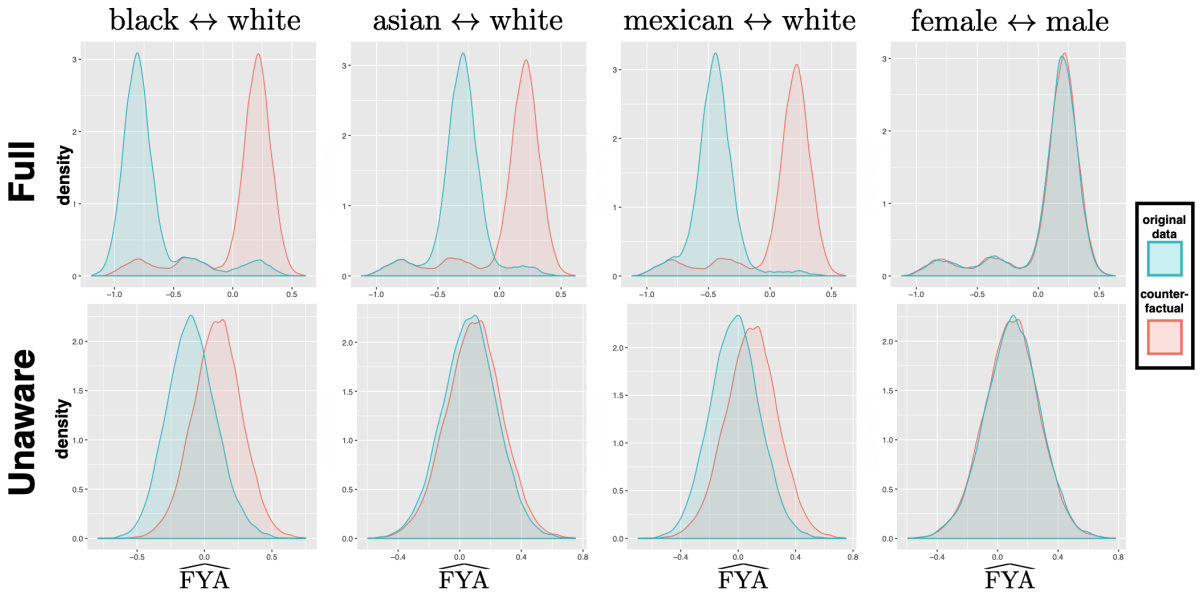
Q: Does predictor $\hat{Y} = h(\boldsymbol{X}, A)$ discriminate based on A?

A: $\hat{Y}$ is counterfactually fair if the prediction had not changed had the protected attribute A taken on a different value a' instead of a, i.e, if

$$\forall \boldsymbol{x}, a, a' : \quad P\left(\hat{Y}_{A \leftarrow a'} | \boldsymbol{X} = \boldsymbol{x}, A = a\right) = P\left(\hat{Y}_{A \leftarrow a} | \boldsymbol{X} = \boldsymbol{x}, A = a\right)$$



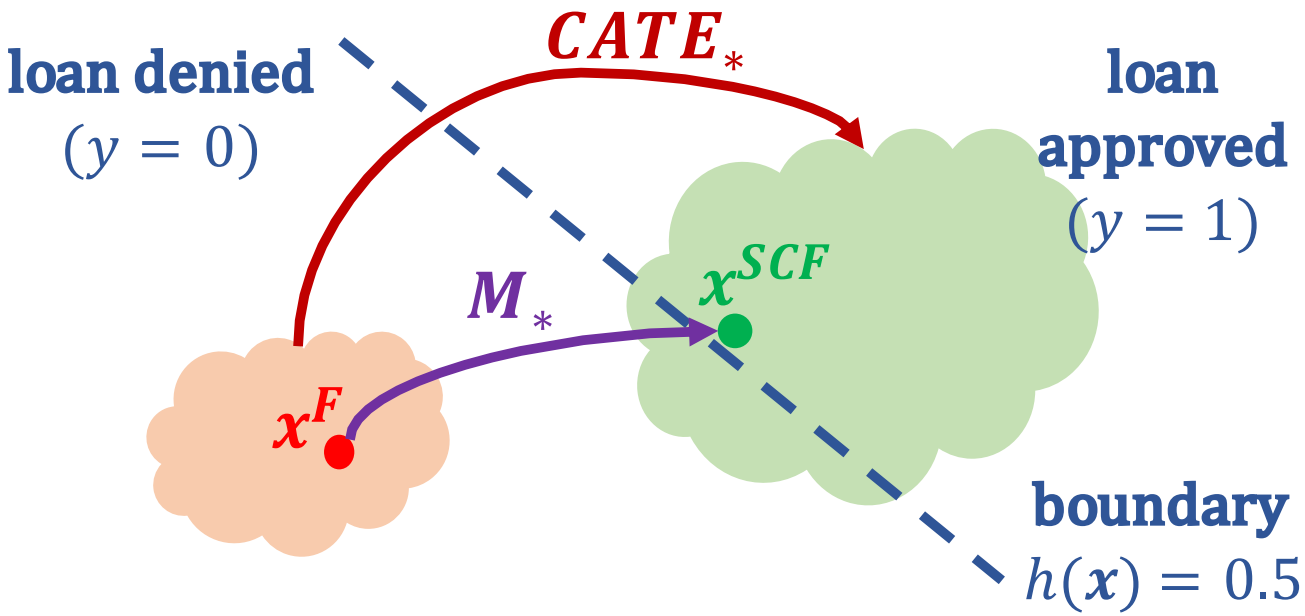
Level 3



[8] Kusner et al., 2017.

Algorithmic Recourse[9]

Why was the loan application of individual $\boldsymbol{x}^F$ rejected by ML model h?
What could they have done to obtain a more favourable outcome?



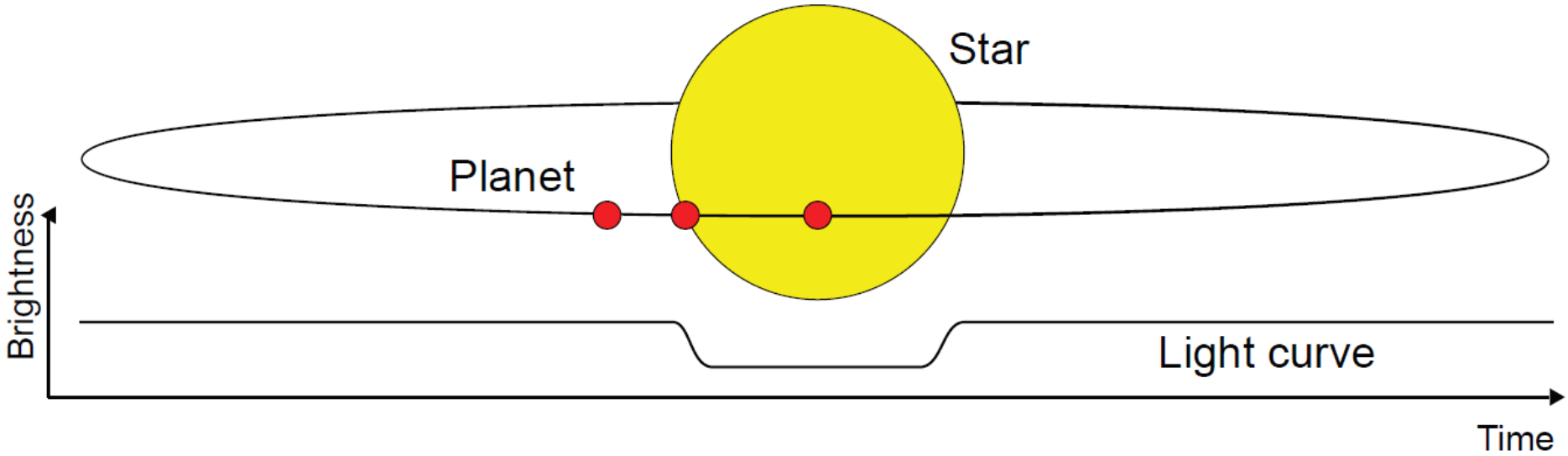
Counterfactual/interventional optimization problem:

$$\min_a \ \text{cost}(a) \qquad \text{subject to} \qquad h\left(\boldsymbol{x}^F_{do(a)}\right) > 0.5$$

[9] Karimi et al., 2020.

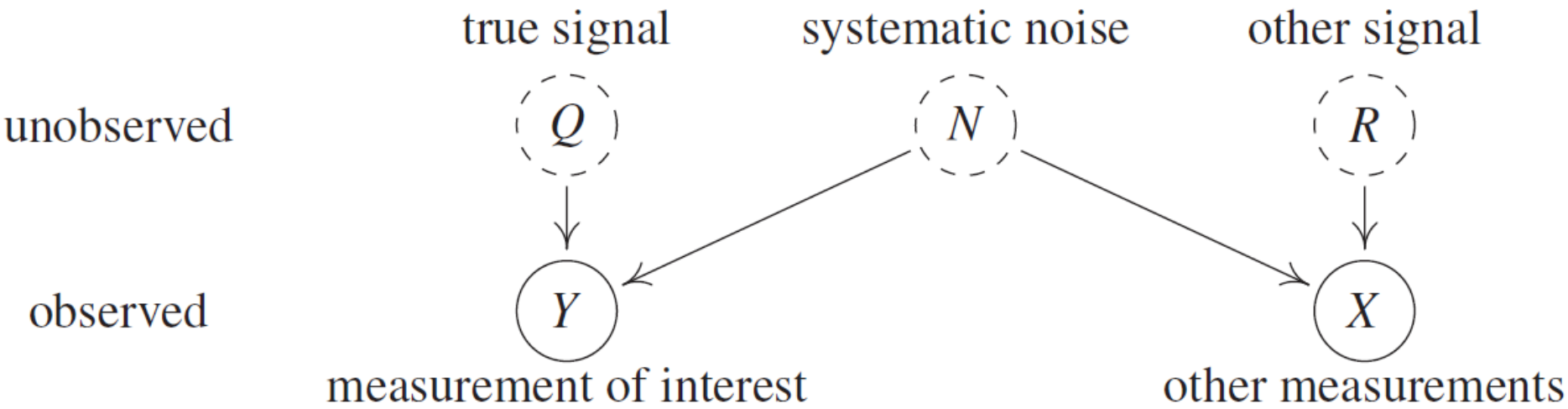
Example: Exoplanet Search

- Kepler space observatory launched in 2009 monitoring brightness of ≈150,000 stars in search for exoplanets.
- Planets orbiting a star can be detected from periodic decreases in light intensity due to occlusion.
- But measurements are corrupted by systematic noise from the telescope making signals hard to detect.



Denoising by Half-Sibling Regression

Idea: use the information contained in other measurements corrupted by the same systematic noise ('half-siblings') to denoise the signal of interest.



All information shared between X and Y must be due to N, so estimate:

$$\hat{Q} := Y - \mathbb{E}[Y|X].$$

Outline

Combining Causality with Representation Learning[10]

Classic AI:
symbols provided a priori; rules provided a priori.

Machine Learning:
representations (symbols) learned from data; only include *statistical* information.

Causal Modeling:
structural causal models assume the causal variables (symbols) are given.

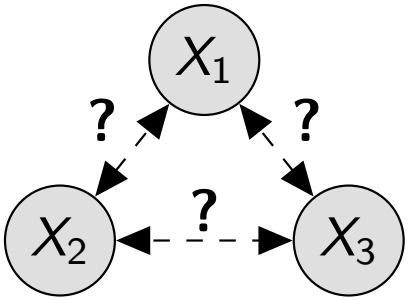
Causal Representation Learning:
capture interventions, reasoning, planning— *"Thinking is acting in an imagined space"* (Konrad Lorenz)

Symbolic AI $\rightarrow$ Learning-based systems

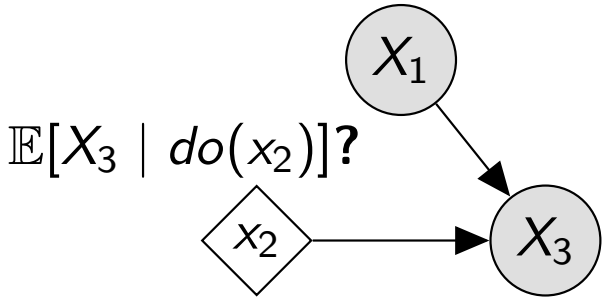
Classical causal inference $\rightarrow$ causal representation learning

[10] Schölkopf and von Kügelgen, 2022.

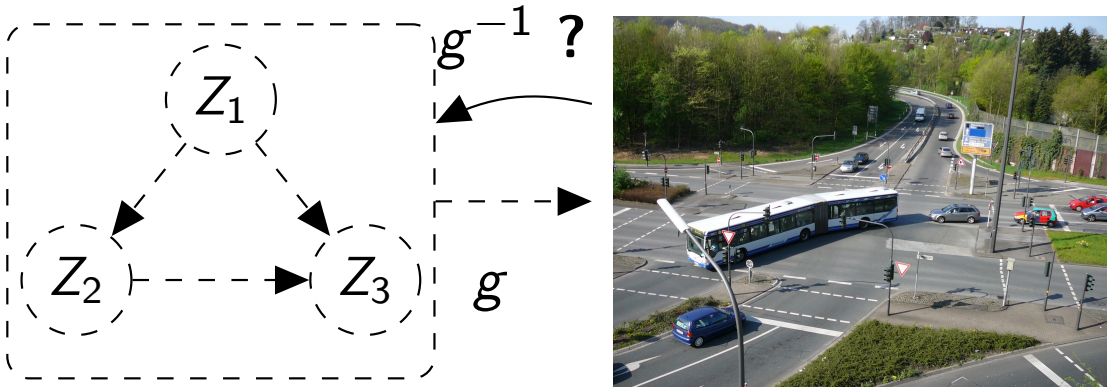
Categorisation of Different Causal Learning Tasks[11]



(a) Causal Discovery



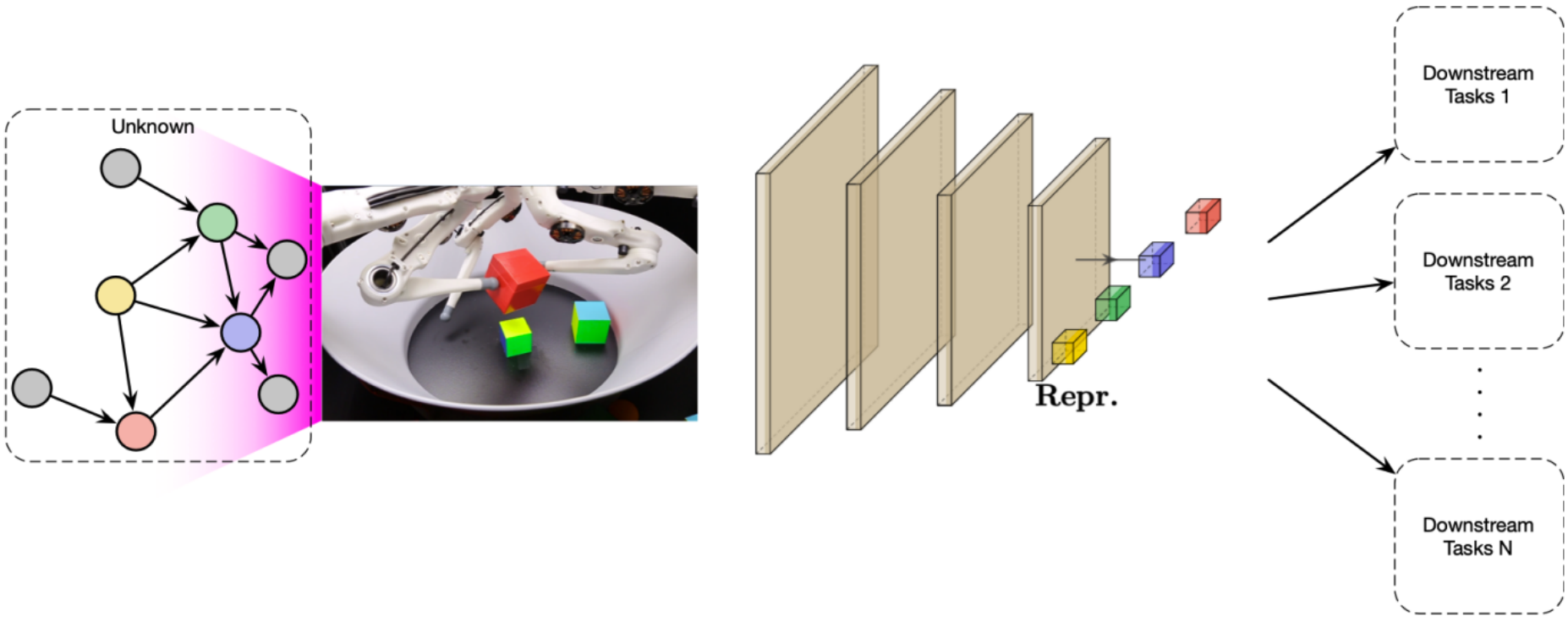
(b) Causal Reasoning



(c) Causal Representation Learning

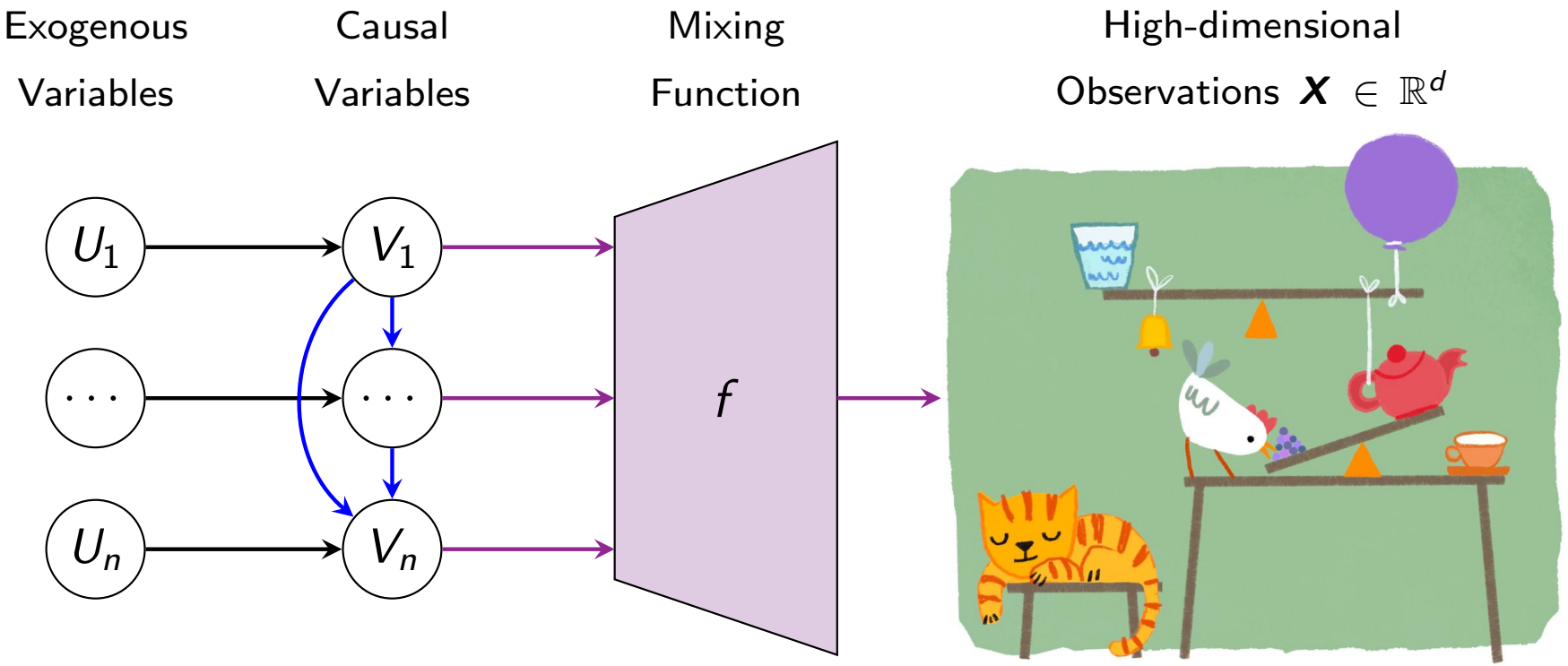
[11] Schölkopf and von Kügelgen, 2022.

Causal Representation Learning Overview [12]



[12] Schölkopf et al., 2021.

Causal Representation Learning: Formal Setup



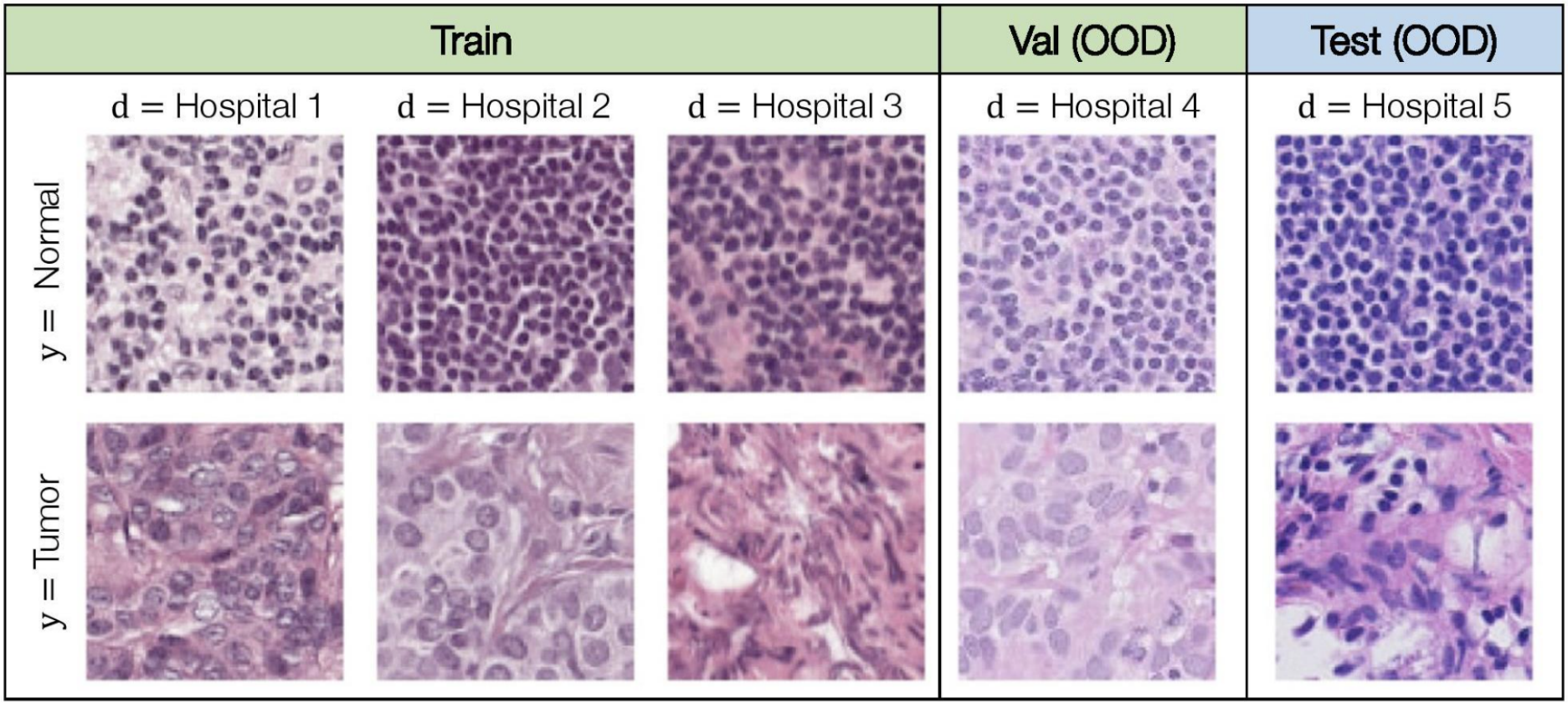
Goal: Given $\boldsymbol{X} = f(\boldsymbol{V})$, recover $(V_1, \ldots, V_n)$ and their causal graph G.

Special Case: If G is known to be empty/trivial, reduces to ICA.

Problem: Solution highly non-unique in general $\rightarrow$ model not identifiable $\rightarrow$ **need additional assumptions and/or rich non-i.i.d. data.**

Multi-Domain Data as an Interesting Learning Signal

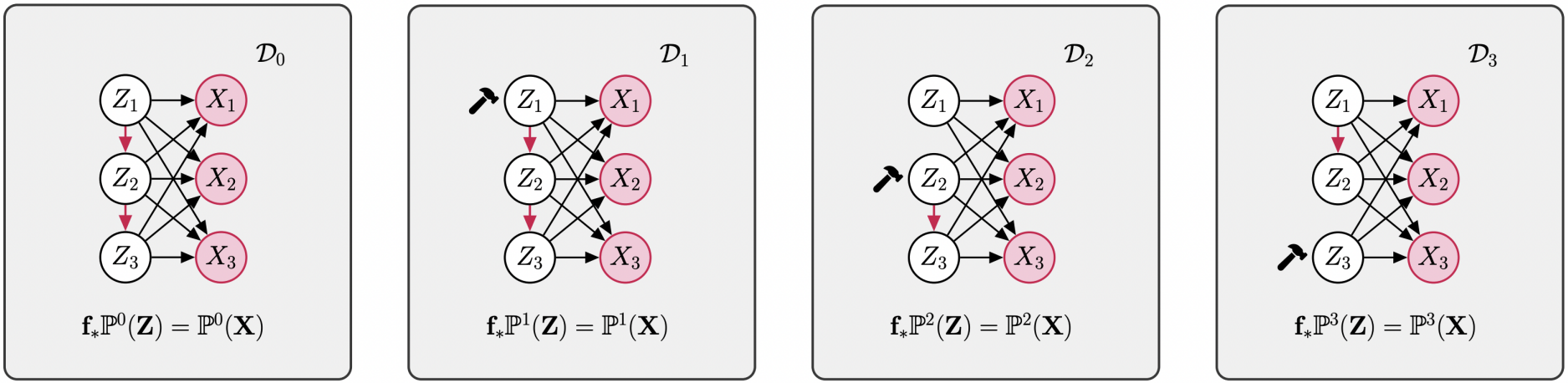
Idea: Have multiple datasets from different domains or environments.[13]



[13] Camelyon17 dataset (Bandi et al., 2018) from the WILDS benchmark suite (Koh et al., 2021) for domain generalisation (Muandet et al., 2013)

Multi-Environment Setup: Formalisation

Idea: Environments arise from interventions in a *shared causal model*.



$$\forall e \in \mathcal{E}: \quad \boldsymbol{X} := f(\boldsymbol{V})$$

$$\boldsymbol{V} \sim P^e(\boldsymbol{V}) = \underbrace{\prod_{i \in \mathcal{I}^e} P_i^e\left(V_i \mid \boldsymbol{V}_{\mathrm{pa}(i)}\right)}_{\text{stochastic interventions}} \underbrace{\prod_{j \notin \mathcal{I}^e} P_j\left(V_j \mid \boldsymbol{V}_{\mathrm{pa}(j)}\right)}_{\text{base causal mechanisms}}$$

Method: look for representations with sparse mechanism changes.

Outline

Summary

- Causal models are a rich model type, capturing not only an observational but also interventional (and counterfactual) distributions arising from manipulations of the system.
- Learning causal models from (observational) data is hard and requires additional assumptions (e.g., no hidden confounding, additive noise)
- Once we have learned a causal model, we can use it for planning and reasoning about interventions and counterfactuals.
- Important limitation & exciting future direction: combine the principled framework of causal modelling with ML strengths in learning representations of complex, high-dimensional data.

References I

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. "From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge". In: *IEEE Transactions on Medical Imaging* (2018).

Clive R Charig, David R Webb, Stephen Richard Payne, and John E Wickham. "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy.". In: *Br Med J (Clin Res Ed)* 292.6524 (1986), pp. 879–882.

P Daniusis, D Janzing, J Mooij, J Zscheischler, B Steudel, K Zhang, and B Schölkopf. "Inferring Deterministic Causal Relations". In: *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. AUAI Press. 2010, pp. 143–150.

George Darmois. "Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire". In: *Revue de l'Institut international de statistique* (1953), pp. 2–8.

William T Freeman. "The Generic Viewpoint Assumption in a Framework for Visual Perception". In: *Nature* 368.6471 (1994), p. 542.

References II


Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. "Independent mechanism analysis, a new concept?". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 28233–28248.


Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. "Nonlinear causal discovery with additive noise models". In: *Advances in neural information processing systems*. 2009, pp. 689–696.


David Hume. *An enquiry concerning human understanding*. 1748.


Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.


Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. "Algorithmic recourse under imperfect causal knowledge: a probabilistic approach". In: *Advances in Neural Information Processing Systems 33*. 2020.


Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. "Wilds: A benchmark of in-the-wild distribution shifts". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5637–5664.


Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4066–4076.

References III


David Lewis. "Causation". In: *The Journal of Philosophy* 70.17 (1973), pp. 556–567.


Franz H Messerli. "Chocolate consumption, cognitive function, and Nobel laureates". In: *N Engl J Med* 367.16 (2012), pp. 1562–1564.


Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. "Domain Generalization via Invariant Feature Representation". In: *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. 2013, pp. 10–18.


Judea Pearl. *Causality: Models, Reasoning, and Inference*. 2nd edition. Cambridge university press, New York, NY, 2009.


Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. "Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis". In: *Advances in Neural Information Processing Systems 36*. 2022.


Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5 (2016), pp. 947–1012.


Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

References IV

Hans Reichenbach. *The Direction of Time*. University of California Press, Berkeley, CA, 1956.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. "Invariant models for causal transfer learning". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 1309–1342.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. "On causal and anticausal learning". In: *Proceedings of the 29th International Conference on Machine Learning*. 2012, pp. 459–466.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. "Toward causal representation learning". In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

Bernhard Schölkopf and Julius von Kügelgen. "From Statistical to Causal Learning". In: *arXiv preprint arXiv:2204.00607* (2022).

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. "A linear non-Gaussian acyclic model for causal discovery". In: *Journal of Machine Learning Research* 7.Oct (2006), pp. 2003–2030.

References V

Ilya Shpitser, Tyler VanderWeele, and James M Robins. "On the validity of covariate adjustment for estimating causal effects". In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. 2010, pp. 527–536.

Viktor Pavlovich Skitovič. "Linear forms of independent random variables and the normal distribution law". In: *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 18.2 (1954), pp. 185–200.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. 2nd edition. MIT press, Cambridge, MA, 2000.

Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. "Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects". In: *IEEE Transactions on Artificial Intelligence* 2.1 (2021), pp. 18–27.

K Zhang, J Peters, D Janzing, and B Schölkopf. "Kernel-based Conditional Independence Test and Application in Causal Discovery". In: *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. AUAI Press. 2011, pp. 804–813.

Kun Zhang and Aapo Hyvärinen. "On the identifiability of the post-nonlinear causal model". In: *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press. 2009, pp. 647–655.