Inferring causality from observations

Dominik Janzing[1] and Sebastian Weichwald[2]

1) Amazon Development Center, Tübingen, Germany
2) CoCaLa, University of Copenhagen, Denmark

September 2019

Online material

- Peters, Janzing, Schölkopf: *Elements of Causal Inference*, MIT Press 2017, free download as pdf at

  `https://mitpress.mit.edu/books/elements-causal-inference`

- 5-day course at a Summer School 2014 in Finland:

  `https://ei.is.tuebingen.mpg.de/publications/janzing14`

- 3 hours course (together with Bernhard Schölkopf) at the Machine Learning Summer School 2013

  `http://mlss.tuebingen.mpg.de/2013/speakers.html`

- 4 lectures on causality from Jonas Peters

  `https://stat.mit.edu/news/four-lectures-causality/`

Outline

1. **Motivation:** correlation versus causation
2. **Formalizing causality:** causal DAGs, functional causal models, Markov conditions, do-operator, potential outcomes
3. **Strong assumptions that enable causal discovery:** faithfulness, independence of mechanisms, additive noise, linear non-Gaussian models
4. **Macroscopic and microscopic causal models:** consistent coarse-graining of causal models
5. **Causal inference in time series:** Granger causality and its limitations
6. **Causal relations among individual objects:** algorithmic Markov conditions, analogy to probabilistic Markov conditions

(some applications in neuroscience are spread over the sections)

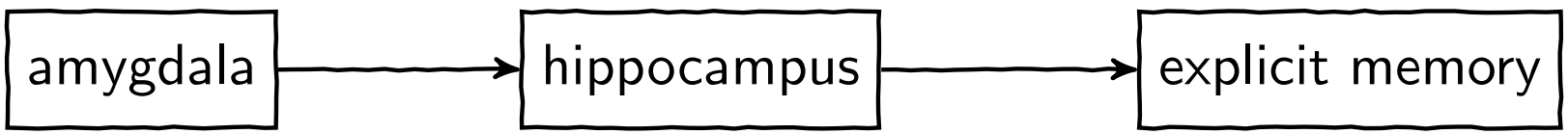
1. Motivation:

correlation versus causation

Why Causality?

Check out discussion sections for causal terminology sneaking in ;-)

Hippocampal activity in this study was correlated with amygdala activity, supporting the view that the amygdala **enhances** *explicit memory by* **modulating** *activity in the hippocampus.*



Drawing causal conclusions from statistical data

- challenging problem, ongoing research
- don't expect an algorithm to which you feed your data and the output is the causal structure
- applying existing algorithms in a sensible way requires deep understanding of the problems of causal inference
- this course will provide a basis for this

Can we infer causal relations from passive observations?

Study report less allergies for children who grew up without dishwasher

Hesselmar et al, Pediatrics March 2015, Vol135 / Issue 3

image source: Wikipedia 'Geschirrspülmaschine', author Christian Giersing

Possible explanations:

- stronger exposure to microbes helps development of immune system
- families without dishwasher tend to have different life style also in other regards

$\Rightarrow$ Relation between statistical and causal dependences is tricky

Statistical and causal statements...

...differ by **slight** rewording:

- **"children growing up without dishwasher are less likely to have allergies"**

- **"children growing up without dishwasher are less likely to have allergies because of missing dishwasher"**

Statistical and causal statements...

...differ by **slight** rewording:

- **"children growing up without dishwasher are less likely to have allergies"**

  statistical statement:
  can be tested by standard statistical tools

- **"children growing up without dishwasher are less likely to have allergies because of the missing dishwasher"**

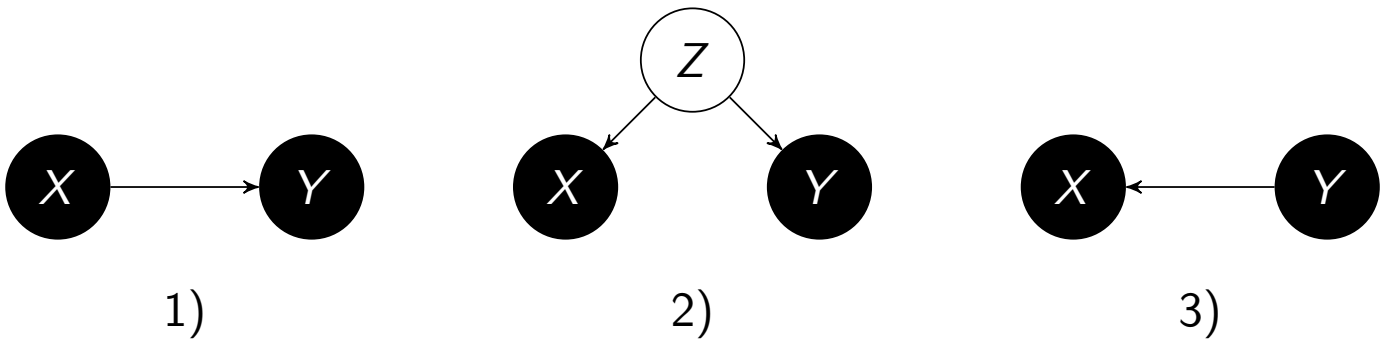
  causal statement:
  no standard methods available, the tutorial will give partial answers, don't expect simple ones!

...this raises the question...

does statistics tell us something about causality at all?

Reichenbach's principle of common cause (1956)

If two variables X and Y are statistically dependent then either



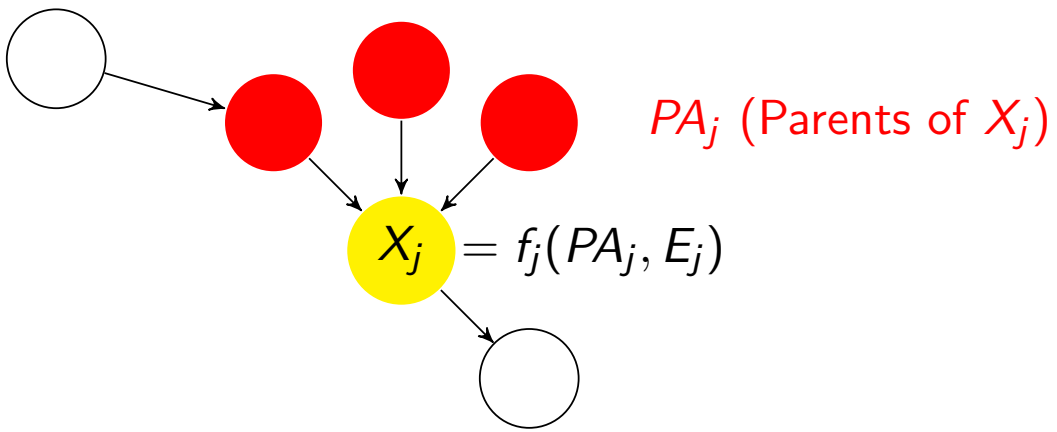
- every statistical dependence is due to a causal relation, we also call 2) "causal".
- distinction between 3 cases is a key problem in scientific reasoning.
- cases 1-3 can also occur simultaneously

2. Formalizing causality:

causal DAGs, functional causal models, Markov conditions, do-operator, potential outcomes

Functional model of causality Pearl et al

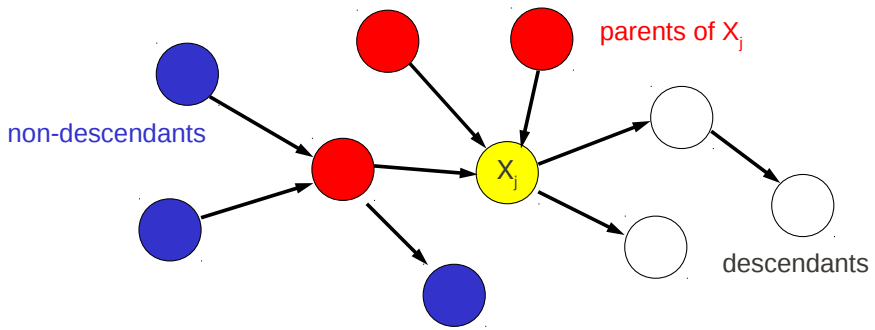
- every node X_j is a function of its parents PA_j and an unobserved noise term E_j
- f_j describes how X_j changes when parents are set to specific values



- all noise terms E_j are statistically independent (causal sufficiency)
- which properties of $P(X_1, \ldots, X_n)$ follow?

Causal Markov condition (4 equivalent versions) Lauritzen et al, Pearl

- **existence of a functional model**
- **local Markov condition:** every node is conditionally independent of its non-descendants, given its parents

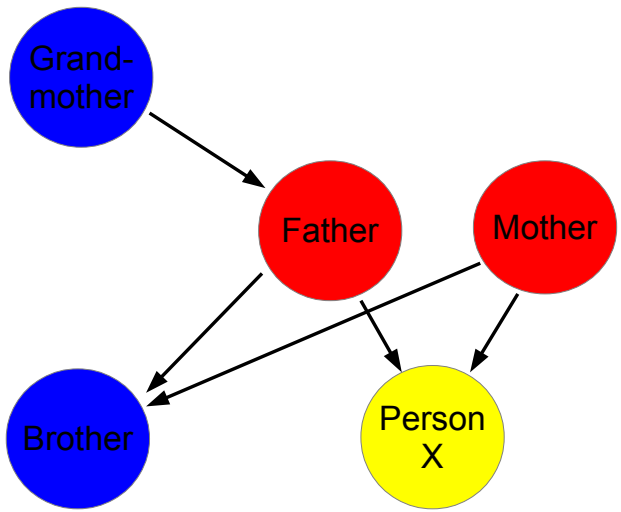


(information exchange with non-descendants involves parents)

- **global Markov condition:** describes all ind. via d-separation
- **Factorization:** $P(X_1, \ldots, X_n) = \prod_j P(X_j|PA_j)$

(every $P(X_j|PA_j)$ describes a causal mechanism)

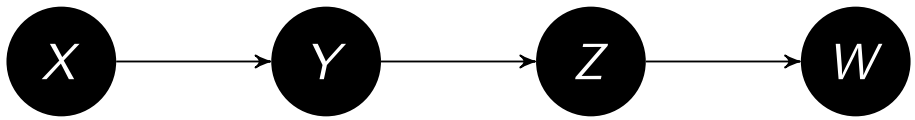
Metaphor for local Markov condition



If someone knows the genes of X's parents, neither the genes of the grandmother nor the genes of the brother contain additional information about X

Idea of the global Markov condition

conditional independences stated by the local Markov condition implies further conditional independences, e.g.



$$X \perp\!\!\!\perp W \,|\, Y$$

does not directly follow from the local Markov condition, although it's true

- intuitively reasonable: since the influence of X on W is intermediated by Y, the dependence disappears for fixed values of Y
- there are mathematical rules about which conditional independences imply further independences

Statistical independence vs. uncorrelatedness

- X, Y **independent:** probabilities factorize, i.e.

$$p(x, y) = p(x)p(y).$$

(difficult to test)

- X, Y **uncorrelated:** expectations factorize, i.e.

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

(easy to test: just compute empirical means)

independent implies uncorrelated but not vice versa
(note: physics literature is sometime sloppy about the difference)

Reformulation of statistical independence

- **factorizing probabilities:** $p(x, y) = p(x)p(y)$
- **knowing X does not change the distribution of Y:**

$$p(y|x) = p(y)$$

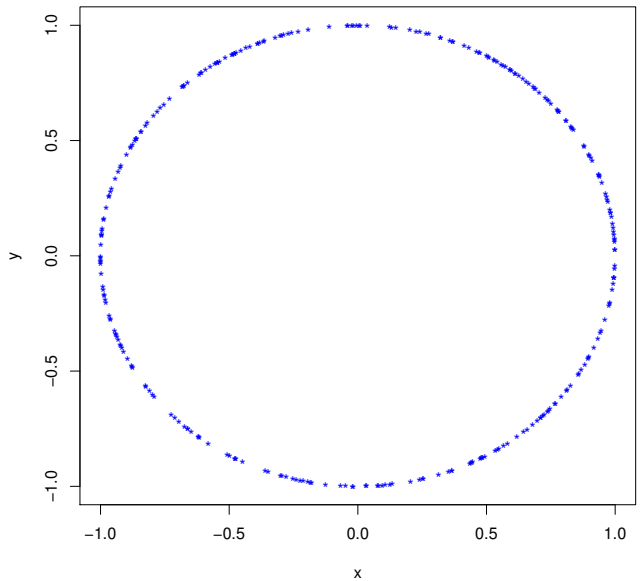
(X contains no information about Y and vice versa)

- **functions of X and Y are uncorrelated:**

$$\mathbb{E}[f(X) \cdot g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)] \quad \forall f, g$$

Dependence without correlation

Let $P_{X,Y}$ be uniform distribution on a circle:



- uncorrelated because $\mathbb{E}[XY] = 0$ and $\mathbb{E}[X] = 0, \mathbb{E}[Y] = 0$ for symmetry reasons
- X and Y are statistically dependent: knowing X reduces the possible values Y from $[-1, 1]$ to just two options

d-separation

(Pearl 1988)

Path = sequence of pairwise distinct nodes where consecutive ones are adjacent

A path q is said to be **blocked** by the set Z if

- q contains a *chain* $i \to m \to j$ or a *fork* $i \leftarrow m \to j$ such that the middle node is in Z, or
- q contains a *collider* $i \to m \leftarrow j$ such that the middle node is not in Z and such that no descendant of m is in Z.

Z is said to **d-separate** X and Y in the DAG G, formally

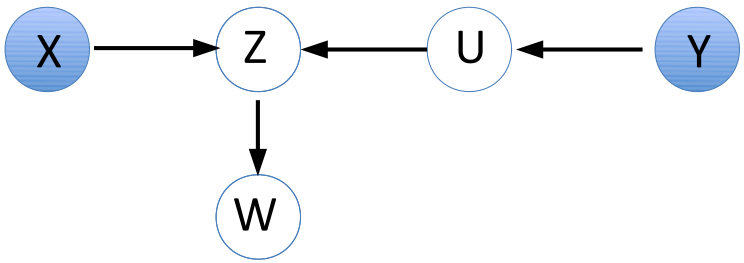
$$(X \perp\!\!\!\perp Y \,|Z)_G$$

if Z blocks every path from a node in X to a node in Y.

Example (blocking of paths)

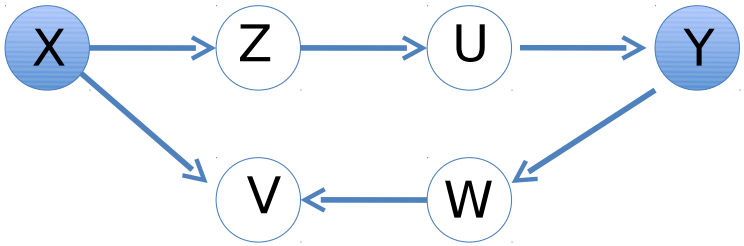
path from X to Y is blocked by conditioning on U or Z or both

Example (unblocking of paths)



- path from X to Y is blocked by $\emptyset$
- unblocked by conditioning on Z or W or both

Example (blocking and unblocking of paths)

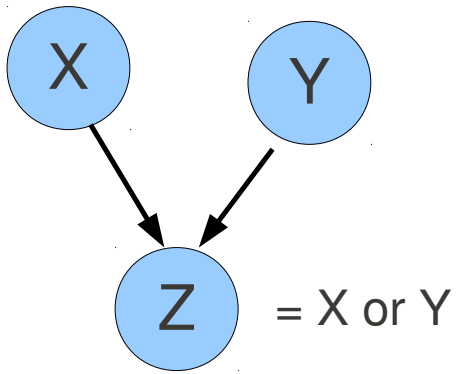


several options for blocking all paths between X and Y:

$$(X \perp\!\!\!\perp Y \,|\, ZW)_G$$
$$(X \perp\!\!\!\perp Y \,|\, ZUW)_G$$
$$(X \perp\!\!\!\perp Y \,|\, VZUW)_G$$

Unblocking by conditioning on common effects

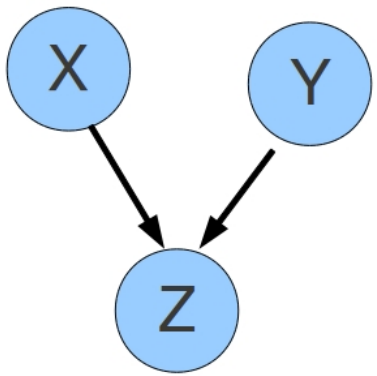
Berkson's paradox (1946), selection bias. Example: X, Y, Z binary

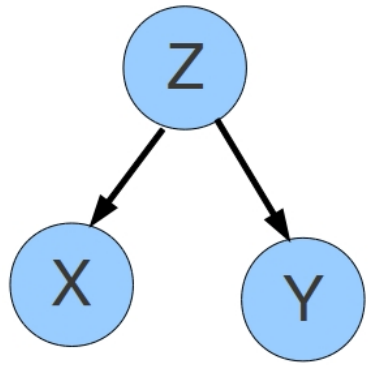


- assume language skills and science skills are independent a priori
- assume pupils go to high school if they have good skills in science or languages
- then there is a negative correlation between science skills and language skills in high school

Asymmetry with respect to inverting arrows

Reichenbach: The direction of time (1956)





$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Y \,|\, Z$

$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \,|\, Z$

Formalizing the difference between seeing and doing

- **observational probabilities:** $p(y|x)$ probability for $Y = y$, given that we observed $X = x$
- **interventional probabilities:** $p(y|do(x))$ probability for $Y = y$, given that we have set X to x.

confusing $p(y|x)$ with $p(y|do(x))$ is the reason for most of the common misconceptions about causality!

Pearl's do operator

how to compute $p(x_1, \ldots, x_n | do(x_i'))$:

- write $p(x_1, \ldots, x_n)$ as

$$\prod_{k=1}^{n} p(x_k | parents(x_k))$$

- replace $p(x_i | parents(x_i))$ with $\delta_{x_i, x_i'}$

$$p(x_1, \ldots, x_n | do(x_i')) = \prod_{k \neq i} p(x_k | parents(x_k)) \delta_{x_i, x_i'}$$

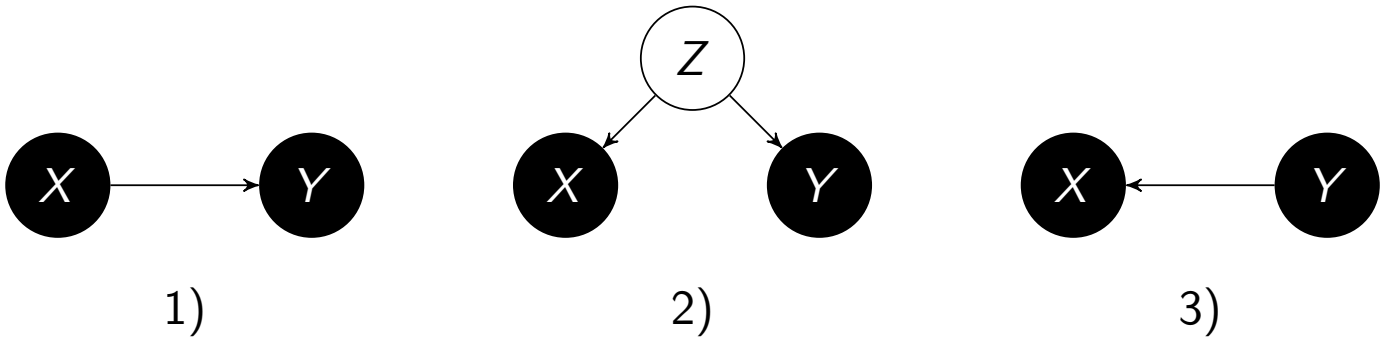
How to compute $p(x_j|do(x_i))$

marginalize over all $k \neq j$:

$$\begin{aligned} p(x_j|do(x_i')) &= \sum p(x_1, \ldots, x_n|do(x_i')) \\ &= \sum \prod_{k \neq i} p(x_k|parents(x_k))\delta_{x_i, x_i'} \end{aligned}$$

(sum runs over all $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$)

Simple examples



1. interventional and observational probabilities coincide (seeing is the same as doing)

$$p(y|do(x)) = p(y|x)$$

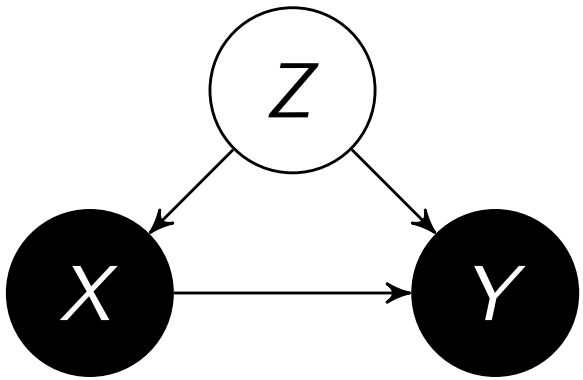
2. intervening on x does not change y

$$p(y|do(x)) = p(y) \neq p(y|x)$$

3. intervening on x does not change y

$$p(y|do(x)) = p(y) \neq p(y|x)$$

Most important case: confounder correction



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

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

- Population $\mathcal{U}$ of units $u \in \mathcal{U}$,

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

- Population $\mathcal{U}$ of units $u \in \mathcal{U}$,

  e. g. a patient group

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

- Population $\mathcal{U}$ of units $u \in \mathcal{U}$,

  e. g. a patient group

- Treatment variable $S : \mathcal{U} \to \{\mathrm{t}, \mathrm{c}\}$,

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

- Population $\mathcal{U}$ of units $u \in \mathcal{U}$,

  e. g. a patient group

- Treatment variable $S : \mathcal{U} \to \{\mathrm{t}, \mathrm{c}\}$,

  e. g. assignment to treatment/control

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

- Population $\mathcal{U}$ of units $u \in \mathcal{U}$,

  e. g. a patient group

- Treatment variable $S : \mathcal{U} \to \{\mathrm{t}, \mathrm{c}\}$,

  e. g. assignment to treatment/control

- Potential outcomes $Y : \mathcal{U} \times \{\mathrm{t}, \mathrm{c}\} \to \mathbb{R}$,

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Ingredients:

- Population $\mathcal{U}$ of units $u \in \mathcal{U}$,

  e. g. a patient group

- Treatment variable $S : \mathcal{U} \to \{\mathrm{t}, \mathrm{c}\}$,

  e. g. assignment to treatment/control

- Potential outcomes $Y : \mathcal{U} \times \{\mathrm{t}, \mathrm{c}\} \to \mathbb{R}$,

  e. g. survival times $Y_{\mathrm{t}}(u)$ and $Y_{\mathrm{c}}(u)$ of patient u

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association,* 1986)

Fundamental problem of causal inference:

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Possible remedy assumptions:

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Possible remedy assumptions:

- Unit homogeneity: $Y_{\mathrm{t}}(u_1) = Y_{\mathrm{t}}(u_2)$ and $Y_{\mathrm{c}}(u_1) = Y_{\mathrm{c}}(u_2)$

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Possible remedy assumptions:

- Unit homogeneity: $Y_{\mathrm{t}}(u_1) = Y_{\mathrm{t}}(u_2)$ and $Y_{\mathrm{c}}(u_1) = Y_{\mathrm{c}}(u_2)$
- Causal transience: can measure $Y_{\mathrm{t}}(u)$ and $Y_{\mathrm{c}}(u)$ sequentially

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Possible remedy assumptions:

- Unit homogeneity: $Y_{\mathrm{t}}(u_1) = Y_{\mathrm{t}}(u_2)$ and $Y_{\mathrm{c}}(u_1) = Y_{\mathrm{c}}(u_2)$
- Causal transience: can measure $Y_{\mathrm{t}}(u)$ and $Y_{\mathrm{c}}(u)$ sequentially

"Statistical solution": Average Treatment Effect $\mathbb{E}[Y_{\mathrm{t}}] - \mathbb{E}[Y_{\mathrm{c}}]$

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Possible remedy assumptions:

- Unit homogeneity: $Y_{\mathrm{t}}(u_1) = Y_{\mathrm{t}}(u_2)$ and $Y_{\mathrm{c}}(u_1) = Y_{\mathrm{c}}(u_2)$
- Causal transience: can measure $Y_{\mathrm{t}}(u)$ and $Y_{\mathrm{c}}(u)$ sequentially

"Statistical solution": Average Treatment Effect $\mathbb{E}[Y_{\mathrm{t}}] - \mathbb{E}[Y_{\mathrm{c}}]$

- Can observe $\mathbb{E}[Y_{\mathrm{t}}|S = \mathrm{t}]$ and $\mathbb{E}[Y_{\mathrm{c}}|S = \mathrm{c}]$

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

Possible remedy assumptions:

- Unit homogeneity: $Y_{\mathrm{t}}(u_1) = Y_{\mathrm{t}}(u_2)$ and $Y_{\mathrm{c}}(u_1) = Y_{\mathrm{c}}(u_2)$
- Causal transience: can measure $Y_{\mathrm{t}}(u)$ and $Y_{\mathrm{c}}(u)$ sequentially

"Statistical solution": Average Treatment Effect $\mathbb{E}[Y_{\mathrm{t}}] - \mathbb{E}[Y_{\mathrm{c}}]$

- Can observe $\mathbb{E}[Y_{\mathrm{t}}|S = \mathrm{t}]$ and $\mathbb{E}[Y_{\mathrm{c}}|S = \mathrm{c}]$
- which, when randomly assigning treatments, i. e. $(Y_{\mathrm{t}}, Y_{\mathrm{c}}) \perp\!\!\!\perp S$,

Potential Outcomes Framework

(PW Holland, Statistics and Causal Inference. *Journal of the American Statistical Association*, 1986)

Fundamental problem of causal inference:
For each unit u we get to observe *either* $Y_{\mathrm{t}}(u)$ *or* $Y_{\mathrm{c}}(u)$ and hence the treatment effect $Y_{\mathrm{t}}(u) - Y_{\mathrm{c}}(u)$ cannot be computed.

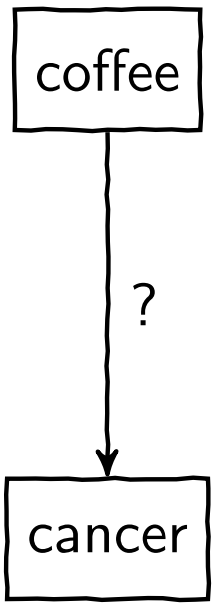
Possible remedy assumptions:

- Unit homogeneity: $Y_{\mathrm{t}}(u_1) = Y_{\mathrm{t}}(u_2)$ and $Y_{\mathrm{c}}(u_1) = Y_{\mathrm{c}}(u_2)$
- Causal transience: can measure $Y_{\mathrm{t}}(u)$ and $Y_{\mathrm{c}}(u)$ sequentially

"Statistical solution": Average Treatment Effect $\mathbb{E}[Y_{\mathrm{t}}] - \mathbb{E}[Y_{\mathrm{c}}]$

- Can observe $\mathbb{E}[Y_{\mathrm{t}}|S = \mathrm{t}]$ and $\mathbb{E}[Y_{\mathrm{c}}|S = \mathrm{c}]$
- which, when randomly assigning treatments, i. e. $(Y_{\mathrm{t}}, Y_{\mathrm{c}}) \perp\!\!\!\perp S$,
- is equal to $\mathbb{E}[Y_{\mathrm{t}}]$ and $\mathbb{E}[Y_{\mathrm{c}}]$.

Potential Outcomes Framework



Potential Outcomes Framework

- Split population $\mathcal{U}$ into

Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$

Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$
  - 'consumed lots': $S(u) = \blacksquare$

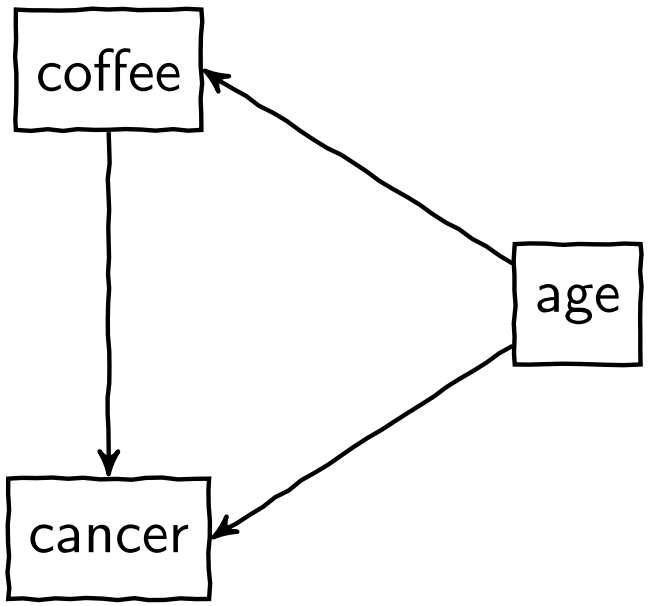
Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$
  - 'consumed lots': $S(u) = \blacksquare$
- Observe whether they suffer from cancer or not, $Y \in \{0, 1\}$

Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$
  - 'consumed lots': $S(u) = \blacksquare$
- Observe whether they suffer from cancer or not, $Y \in \{0, 1\}$
- Assume older units have higher cumulative coffee consumption as well as an increased risk of cancer

Potential Outcomes Framework



Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$
  - 'consumed lots': $S(u) = \blacksquare$
- Observe whether they suffer from cancer or not, $Y \in \{0, 1\}$
- Assume older units have higher cumulative coffee consumption as well as an increased risk of cancer
  - $(Y_{\square}, Y_{\blacksquare}) \not\perp\!\!\!\perp S$

Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$
  - 'consumed lots': $S(u) = \blacksquare$
- Observe whether they suffer from cancer or not, $Y \in \{0, 1\}$
- Assume older units have higher cumulative coffee consumption as well as an increased risk of cancer
  - $(Y_\square, Y_\blacksquare) \not\perp\!\!\!\perp S$
  - $\mathbb{E}[Y_\square | S = \square] < \mathbb{E}[Y_\square]$

Potential Outcomes Framework

- Split population $\mathcal{U}$ into
  - 'consumed little': $S(u) = \square$
  - 'consumed lots': $S(u) = \blacksquare$
- Observe whether they suffer from cancer or not, $Y \in \{0, 1\}$
- Assume older units have higher cumulative coffee consumption as well as an increased risk of cancer
  - $(Y_\square, Y_\blacksquare) \not\perp\!\!\!\perp S$
  - $\mathbb{E}[Y_\square | S = \square] < \mathbb{E}[Y_\square]$

$\implies$ $\mathbb{E}[Y_\blacksquare | S = \blacksquare] - \mathbb{E}[Y_\square | S = \square]$ systematically overestimates the effect of cumulative coffee consumption on cancer

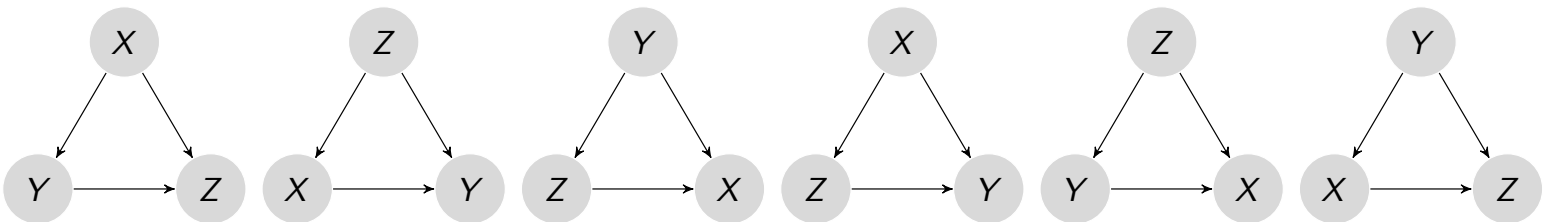
3. Strong assumptions that enable causal discovery:

faithfulness, independence of mechanisms, additive noise, linear non-Gaussian models

Causal discovery from observational data

Can we infer G from $P(X_1, \ldots, X_n)$?

- MC only describes which sets of DAGs are consistent with P
- $n!$ many DAGs are consistent with any distribution



- reasonable rules for preferring simple DAGs required

Independence of mechanisms (ICM)

The conditionals $P(X_j|PA_j)$ in the causal factorization $P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P(X_j|PA_j)$ represent independent mechanisms in nature

- **independent change:** they change independently across data sets
- **no information:** they contain no information about each other, formalization by algorithmic information theory: shortest description of $P(X_1, \ldots, X_n)$ is given by separate descriptions of $P(X_j|PA_j)$

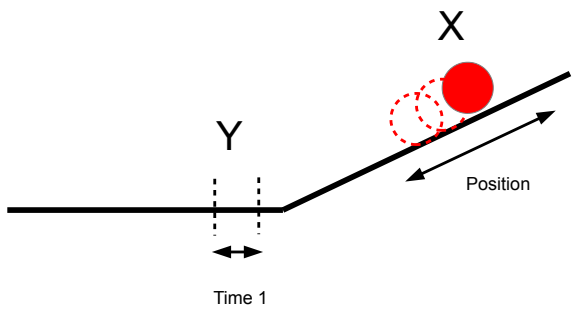
(see Peters, Janzing, Schölkopf: *Elements of Causal Inference* for historical overview)

ICM for the bivariate case

- both $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ may change across environments
- but they change independently
- knowing how $P(\text{cause})$ has changed does not provide information about if and how $P(\text{effect}|\text{cause})$ has changed
- knowing how $P(\text{effect}|\text{cause})$ has changed does not provide information about if and how $P(\text{cause})$ has changed

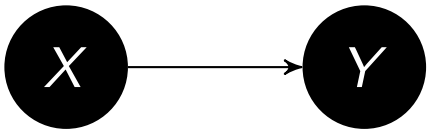
Independent changes in the real world: ball track

relation between initial position (cause) and speed (effect) measured between two light barriers

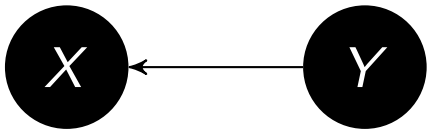


- $P(\text{cause})$ changes if another child plays
- $P(\text{effect}|\text{cause})$ changes if the light barriers are mounted at a different position
- hard to think of operations that change $P(\text{effect})$ without affecting $P(\text{cause}|\text{effect})$ or vice versa

Implications of ICM for causal and anti-causal learning



causal learning:
predict effect from cause



anticausal learning:
predict cause from effect

- **Causal learning:**
  predict properties of a molecule from its structure
- **Anticausal learning:** tumor classification, image segmentation

Hypothesis: SSL only works for anticausal learning. Confirmed by screening performance studies in the literature.

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij: On causal and anticausal learning, ICML 2012

Anti-causal prediction: why unlabelled points may help

- let Y be some class label e.g. $y \in \{male, female\}$
- Let X be a feature influenced by Y, e.g. height
- observe that P_X is bimodal

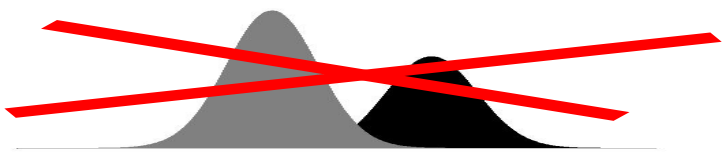
- probably the two modes correspond to the two classes (idea of cluster algorithms)

(can easily be confirmed by observing a *small* number of labeled points)

Causal prediction: why unlabelled points don't help

- let Y be some class label of an effect $y \in \{sick, healthy\}$
- Let X be a feature influencing Y, e.g. a risk factor like blood pressure
- observe that P_X is bimodal

- no reasons to believe that the modes correspond to the two classes

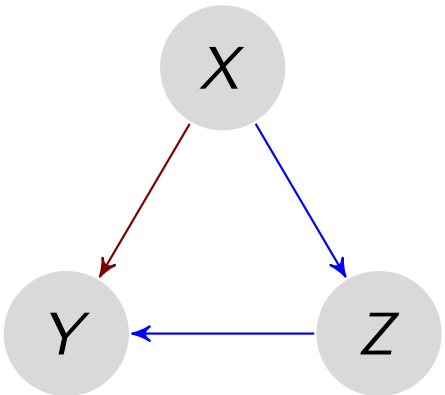


Causal faithfulness as implication of ICM

Spirtes, Glymour, Scheines, 1993

Prefer those DAGs for which all observed conditional independences are implied by the Markov condition

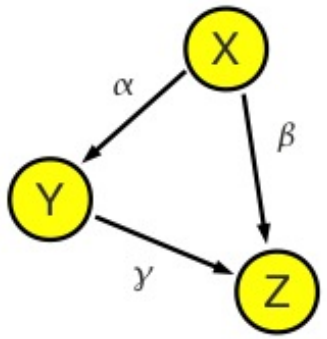
- **Idea:** generic choices of parameters yield faithful distributions
- **Example:** let $X \perp\!\!\!\perp Y$ for the DAG



- not faithful, direct and indirect influence compensate

Examples of unfaithful distributions

cancellation of direct and indirect influence in linear models



$$\begin{aligned} Y &= \alpha X + N_Y \\ Z &= \beta X + \gamma X + N_Z \end{aligned}$$

with independent X, N_Y, N_Z

X and Z are independent if $\beta + \alpha\gamma = 0$

Conditional-independence based causal inference

Spirtes, Glymour, Scheines and Pearl:

Causal Markov condition + Causal faithfulness:
accept only those DAGs as causal hypotheses for which:

- **all** independences are true that are required by the Markov condition
- **only** those independences are true

identifies causal DAG up to Markov equivalence class
(DAGs that imply the same conditional independences)

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X$

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y$

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc,* 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y \implies$ existence of path between S and Y w/o collider

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc*, 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y \implies$ existence of path between S and Y w/o collider
  - $S \perp\!\!\!\perp Y | X$

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc*, 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y \implies$ existence of path between S and Y w/o collider
  - $S \perp\!\!\!\perp Y | X \implies$ all paths between S and Y blocked by X

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc*, 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y \implies$ existence of path between S and Y w/o collider
  - $S \perp\!\!\!\perp Y | X \implies$ all paths between S and Y blocked by X
- Can rule out cases such as $S \to X \leftarrow h \to Y$

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc*, 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y \implies$ existence of path between S and Y w/o collider
  - $S \perp\!\!\!\perp Y | X \implies$ all paths between S and Y blocked by X
- Can rule out cases such as $S \to X \leftarrow h \to Y$
- Can formally prove that X indeed is a cause of Y

Hidden Confounding and CI-based CI in Neuroimaging

(S Weichwald et al., *NeuroImage*, 2015; M Grosse-Wentrup et al., *NeuroImage*, 2016; S Weichwald et al., *IEEE ST SigProc*, 2016)

- Randomised stimulus S
- Observe neural activity X and Y

$\rightsquigarrow$ Estimate $\mathbb{P}^{\varnothing}_{S,X,Y}$

- Assume we find
  - $S \not\perp\!\!\!\perp X \implies$ existence of path between S and X w/o collider
  - $S \not\perp\!\!\!\perp Y \implies$ existence of path between S and Y w/o collider
  - $S \perp\!\!\!\perp Y | X \implies$ all paths between S and Y blocked by X
- Can rule out cases such as $S \to X \leftarrow h \to Y$
- Can formally prove that X indeed is a cause of Y

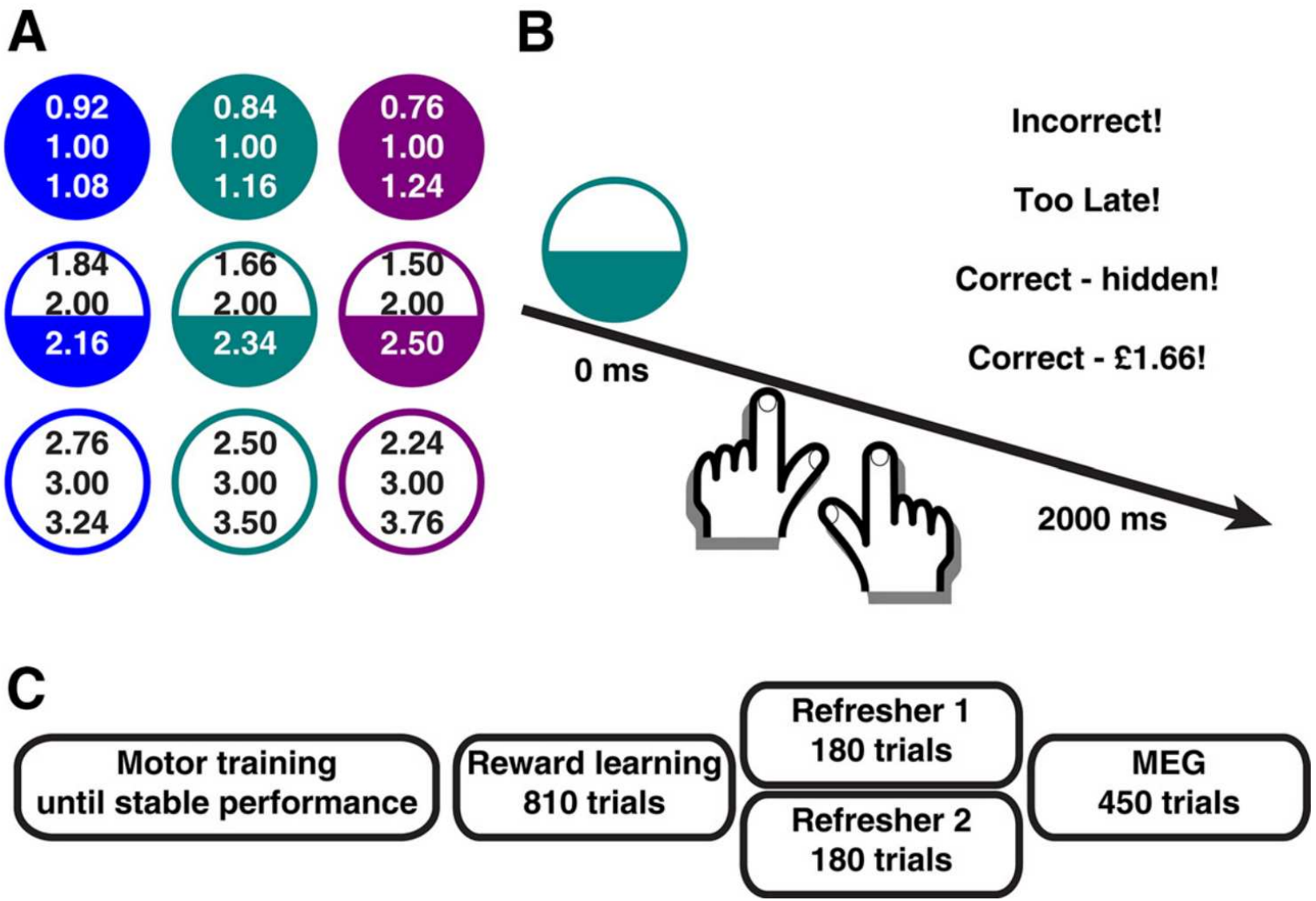
$\implies$ Robust against hidden confounding

Application: Neural Dynamics of Reward Prediction

(Bach, Symmonds, Barnes, and Dolan. *Journal of Neuroscience,* 2017)

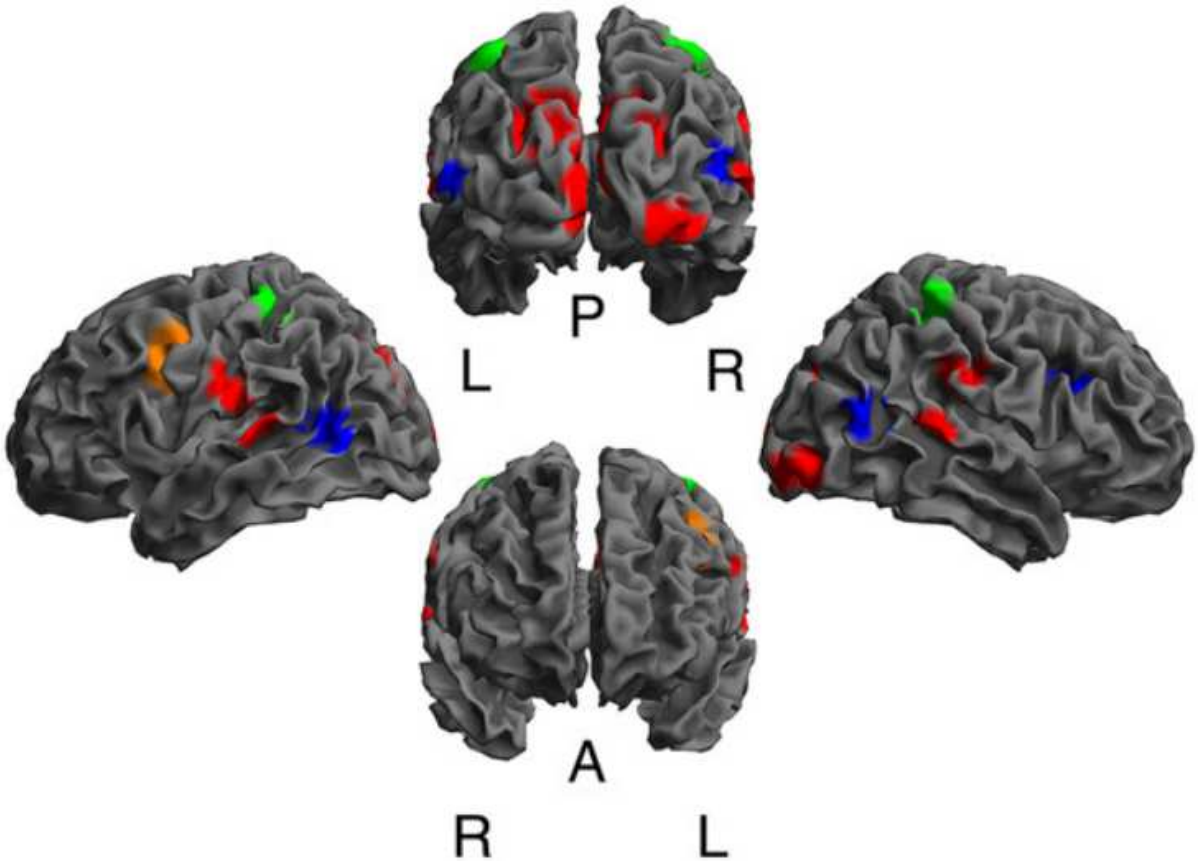
Application: Neural Dynamics of Reward Prediction

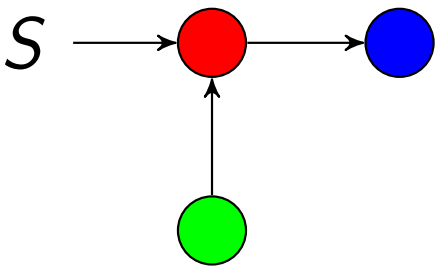
(Bach, Symmonds, Barnes, and Dolan. *Journal of Neuroscience,* 2017)



Application: Neural Dynamics of Reward Prediction

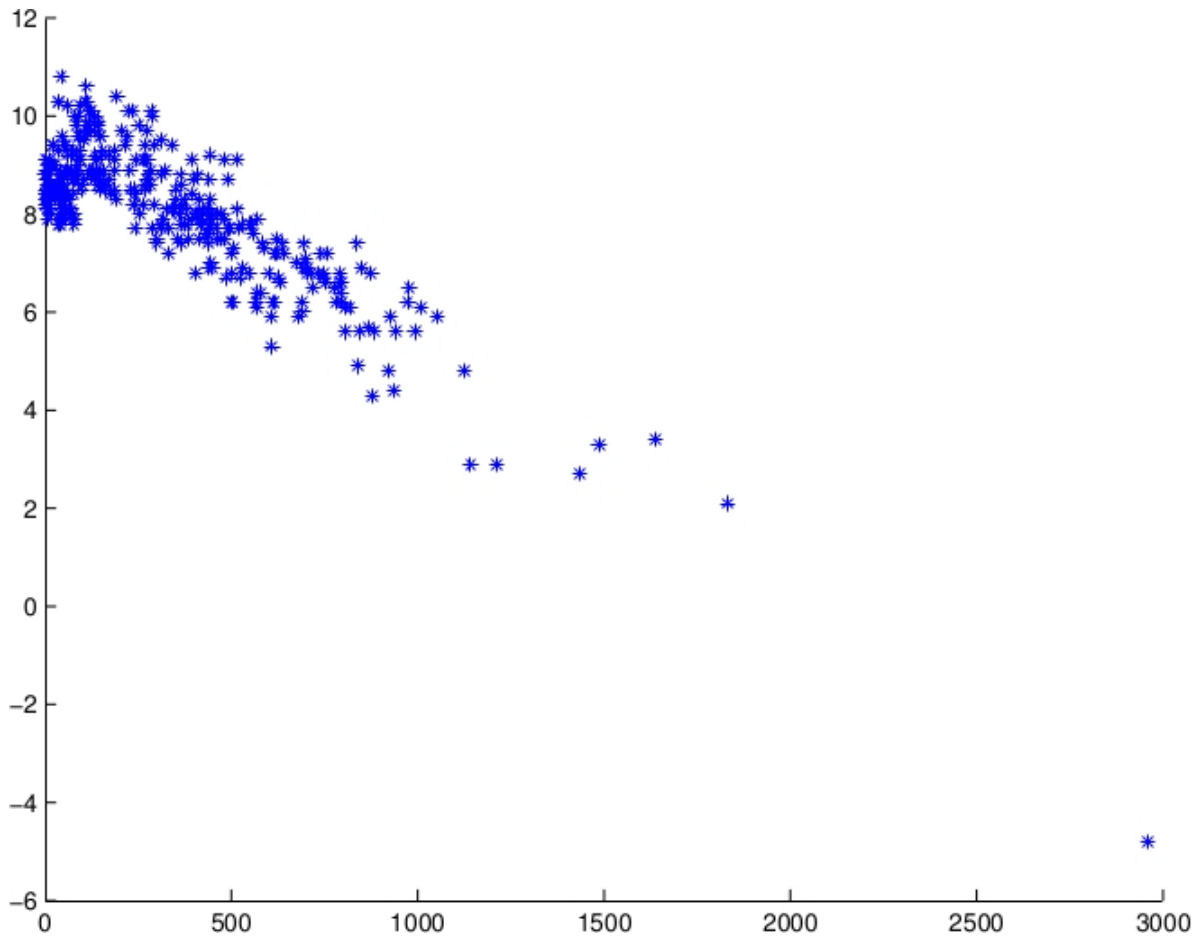
(Bach, Symmonds, Barnes, and Dolan. *Journal of Neuroscience,* 2017)



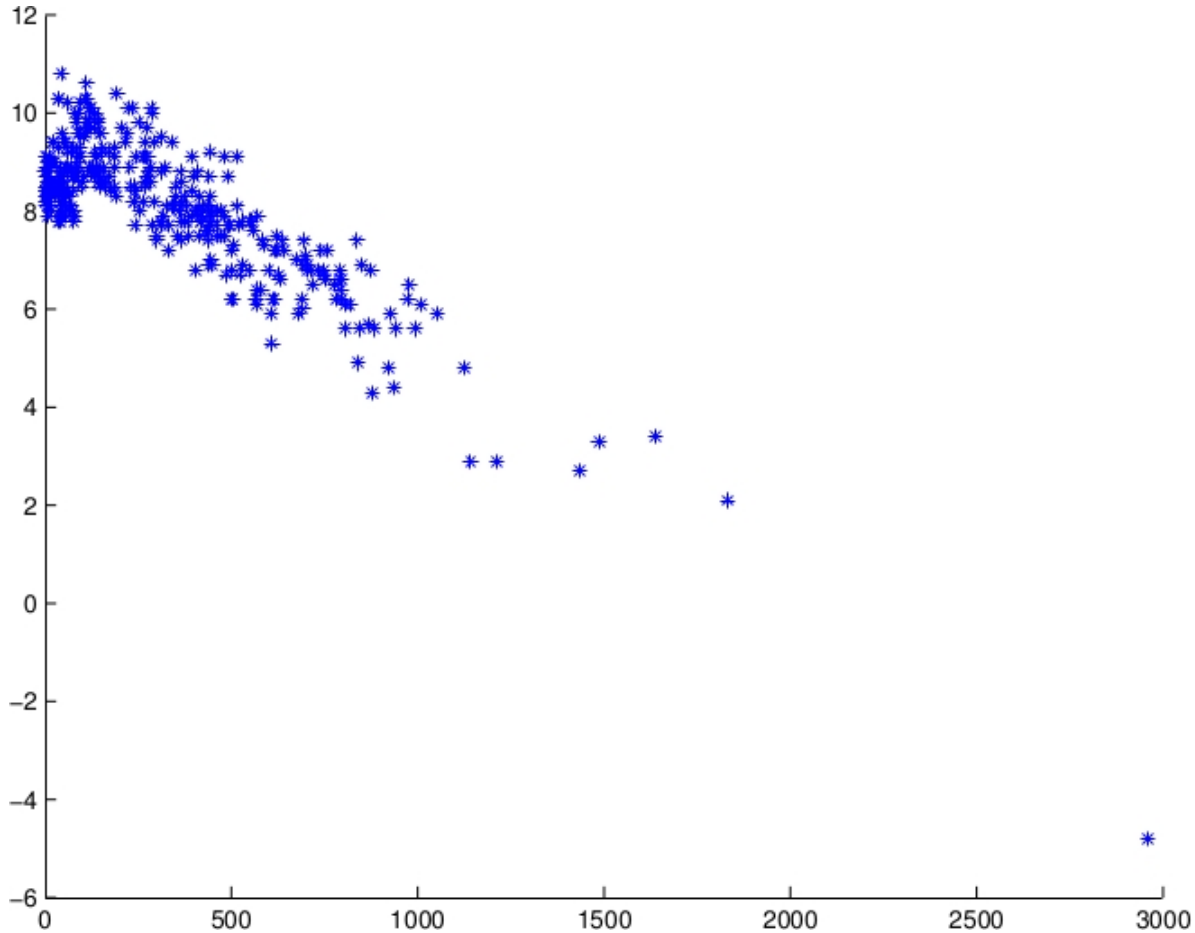


What can be said beyond Markov condition and faithfulness?

What's the cause and what's the effect?

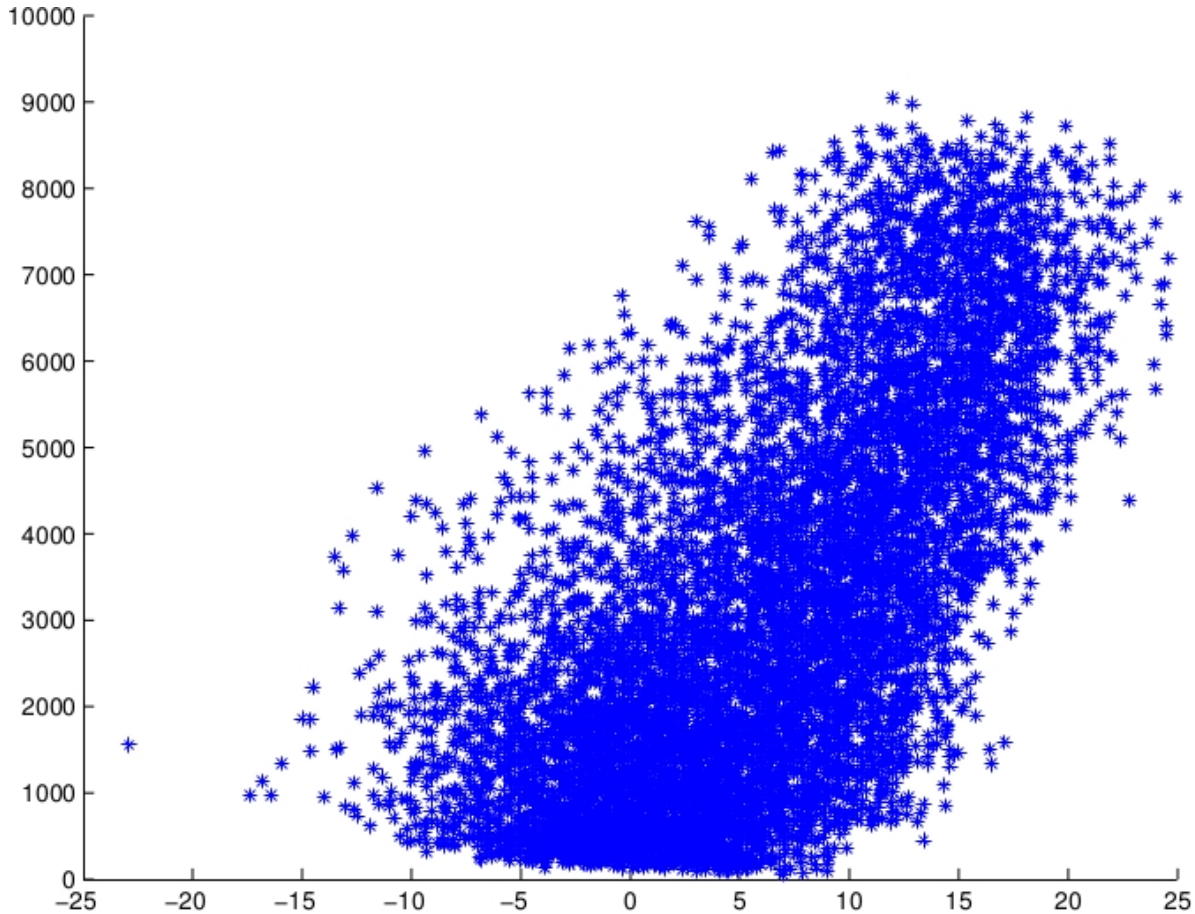


What's the cause and what's the effect?

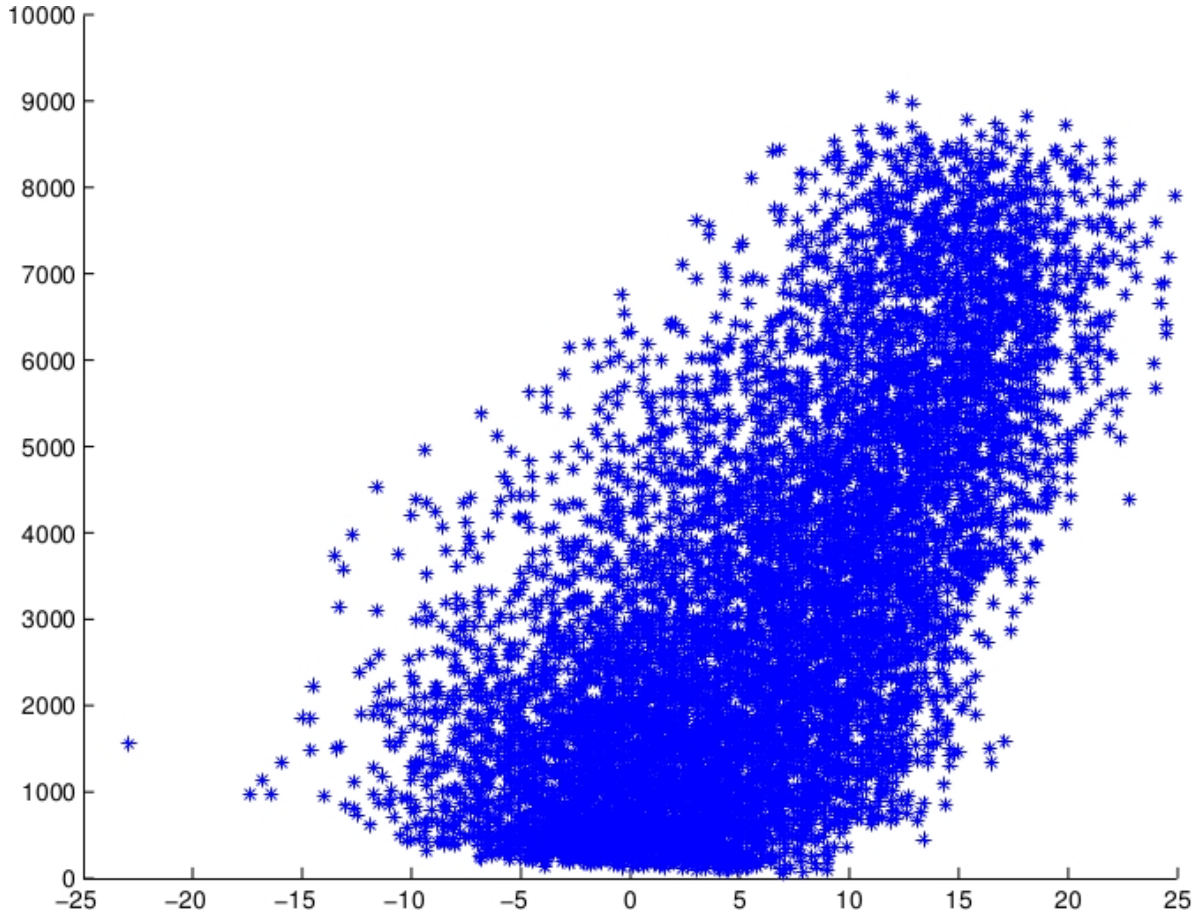


X (Altitude) $\rightarrow$ Y (Temperature)

What's the cause and what's the effect?

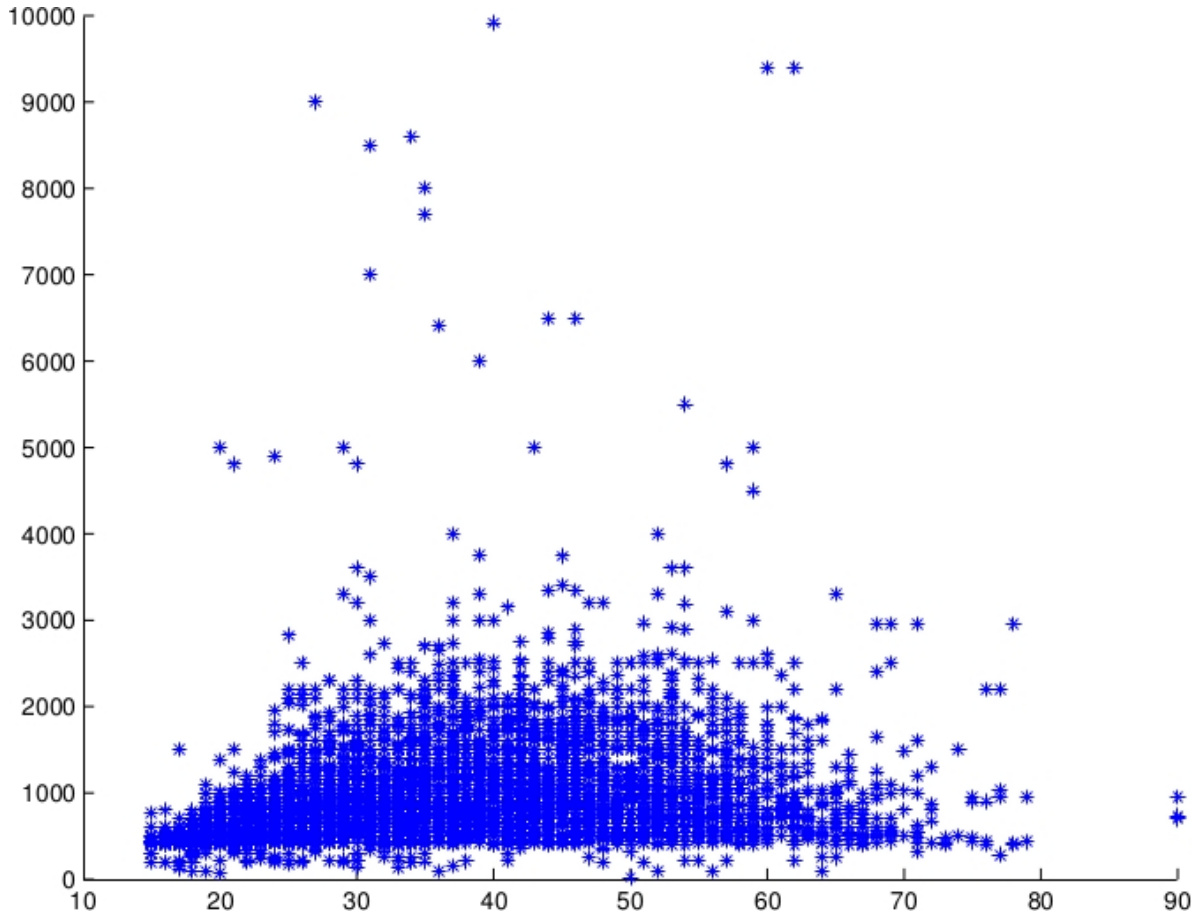


What's the cause and what's the effect?

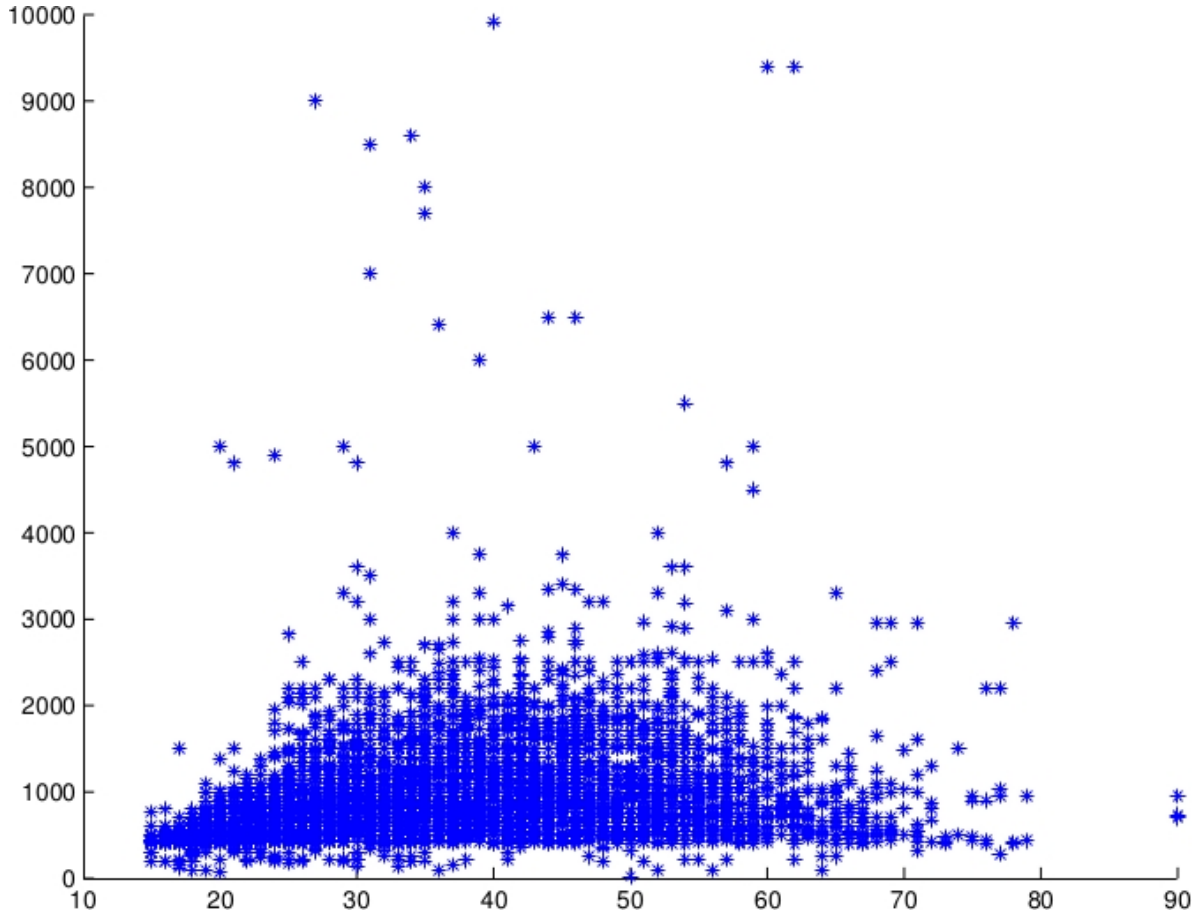


Y (Solar Radiation) $\rightarrow$ X (Temperature)

What's the cause and what's the effect?



What's the cause and what's the effect?



X (Age) $\rightarrow$ Y (Income)

Hence...

- there are asymmetries between cause and effect apart from those formalized by the causal Markov condition

- new methods that employ these asymmetries need to be developed

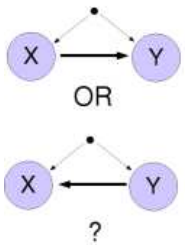
Database with cause effect pairs

Database with cause-effect pa ×

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

Diese Seite ist auf Englisch Soll sie übersetzt werden? Nein Übersetzen

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 g 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 de 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 w 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.

Note: pair0001 - pair0041 were taken from the UCI Machine Learning Repository, so if you use these data sets please refer to their webpage. Here you will find their citation policy.

If you have any comments, questions or suggestions for additional data sets, please contact Dominik Janzing.

Data	Description	Scatter plot (PDF)
pair0001.txt	pair0001_des.txt	pair0001.pdf
pair0002.txt	pair0002_des.txt	pair0002.pdf
pair0003.txt	pair0003_des.txt	pair0003.pdf
pair0004.txt	pair0004_des.txt	pair0004.pdf
pair0005.txt	pair0005_des.txt	pair0005.pdf
pair0006.txt	pair0006_des.txt	pair0006.pdf
pair0007.txt	pair0007_des.txt	pair0007.pdf

Idea of the website

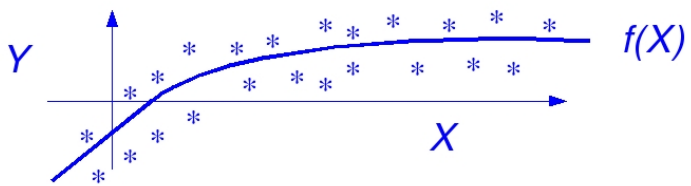
- to evaluate novel causal inference methods
- inspire development of novel methods
- provide data where ground truth is obvious to non-experts (as opposed to many data sets on economy, biology)
- should grow further (contains 105 pairs currently)
- ground truth discussed in: J. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf: *Distinguishing cause from effect using observational data: methods and benchmarks,* Journal of Machine Learning Research, 2016.

Non-linear additive noise based inference

Hoyer, Janzing, Mooij, Peters,Schölkopf, 2008

- Assume that the effect is a function of the cause up to an additive noise term that is statistically independent of the cause:

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



- there will, in the generic case, be no model

$$X = g(Y) + N_X \quad \text{with} \quad N_X \perp\!\!\!\perp Y\,,$$

even if f is invertible! (proof is non-trivial)

Note...

$$Y = f(X, N_Y) \quad \text{with} \quad N_Y \perp\!\!\!\perp X$$

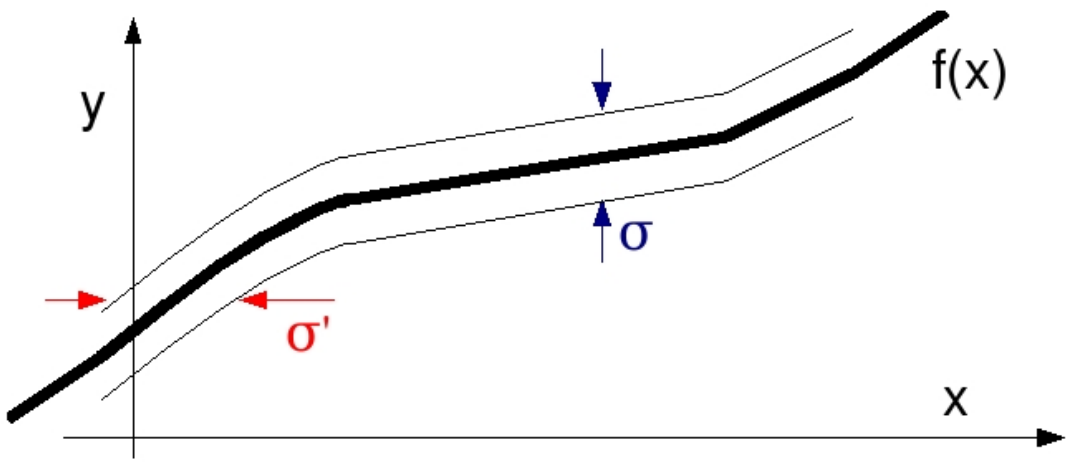
can model **any** conditional $P_{Y|X}$

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

restricts the class of $P_{Y|X}$

Intuition

- additive noise model from X to Y imposes that the width of noise is constant in x.
- for non-linear f, the width of noise won't be constant in y at the same time.



Causal inference method:

Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

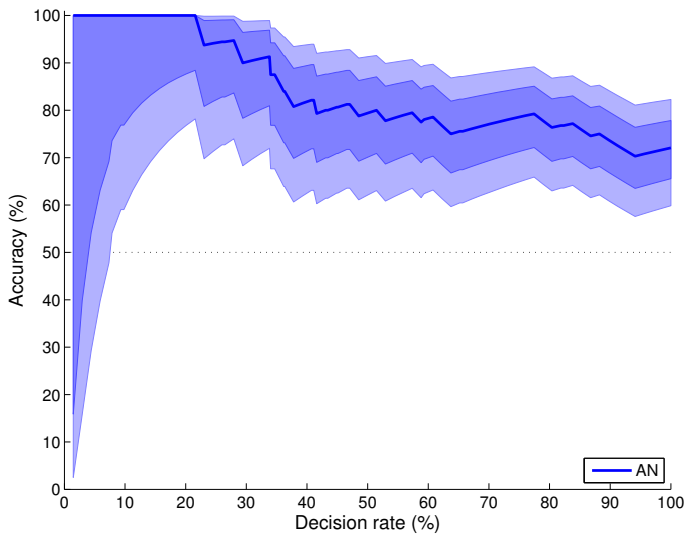
- Compute a function f as non-linear regression of Y on X, i.e., $f(x) := \mathbb{E}[Y|x]$.
- Compute the noise

$$N_Y := Y - f(X)$$

- check whether N_Y and X are statistically independent (uncorrelated is not sufficient, method requires tests that are able to detect higher order dependences)
- performed better than chance on real data with known ground truth

Extensive evaluation

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



- if the algorithm decides in all cases, about 75% decisions are right
- if it only decides in 'the most obvious' 20% of the cases, the fraction gets close ot 100%

Justification of the method

- we don't claim that every causal influence can be described by an additive noise model
- we only claim 'if there is an additive noise model from one direction but not the other the former is likely to be the causal direction'
- if nature chooses P_{cause} and $P_{\mathrm{effect|cause}}$ independently it is unlikely that the result is a joint distribution $P_{\mathrm{effect,cause}}$ that admits an additive noise model from effect to cause

Some theoretical support

Assume $Y = f(X) + N_Y$ with $N_Y \perp\!\!\!\perp X$

- Then P_Y and $P_{X|Y}$ are related:

$$\frac{\partial^2}{\partial y^2} \log p(y) = -\frac{\partial^2}{\partial y^2} \log p(x|y) - \frac{1}{f'(x)} \frac{\partial^2}{\partial x \partial y} \log p(x|y)\,.$$

  $\Rightarrow \frac{\partial^2}{\partial y^2} \log p(y)$ can be computed from $p(x|y)$ knowing $f'(x_0)$ for one specific x_0

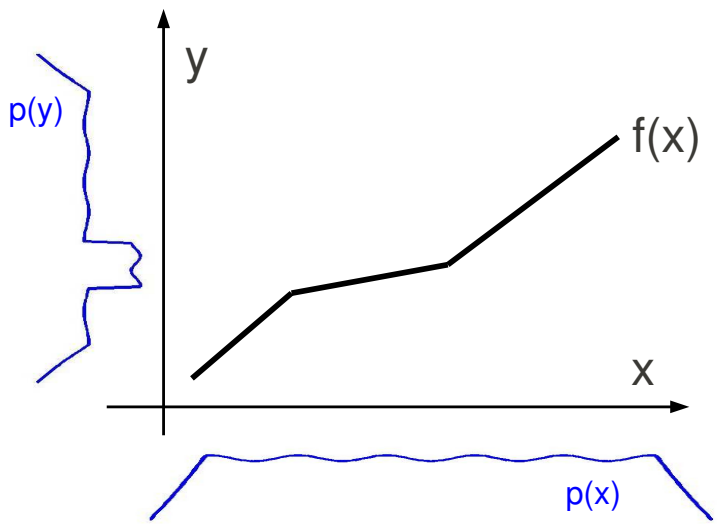
- $P_{X|Y}$ almost determines P_Y
- We reject $Y \to X$ (provided that P_Y is complex) because we assume that nature chooses P_{cause} and $P_{\text{effect}|\text{cause}}$ independently

Janzing, Steudel: Justifying additive noise-based causal inference via algorithmic information theory, OSID (2010)

Inferring deterministic causality

Daniusis, Janzing,... *UAI 2010*, Janzing et al. *AI 2012*

- Problem: infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the right causal model
- Idea: if $X \to Y$ then f and the density p_X are chosen independently "by nature"
- Hence, peaks of p_X do not correlate with the slope of f
- Then, peaks of p_Y correlate with the slope of f^{-1}



Inferring causal structure via ICA

(Shimizu et al. (2006))

A linear acyclic SCM

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} 0 & b_{12} & \dots & b_{1d} \\ 0 & 0 & \dots & b_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} + \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{pmatrix}$$

with mutually independent components $S_1, \dots, S_d$

Inferring causal structure via ICA

(Shimizu et al. (2006))

A linear acyclic SCM

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} 0 & b_{12} & \dots & b_{1d} \\ 0 & 0 & \dots & b_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} + \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{pmatrix}$$

with mutually independent components $S_1, \dots, S_d$

is closely linked to ICA (Independent Component Analysis) as per

Inferring causal structure via ICA

(Shimizu et al. (2006))

A linear acyclic SCM

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} 0 & b_{12} & \dots & b_{1d} \\ 0 & 0 & \dots & b_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} + \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{pmatrix}$$

with mutually independent components $S_1, \dots, S_d$

is closely linked to ICA (Independent Component Analysis) as per

$$X = \mathbf{B} \cdot X + S$$

Inferring causal structure via ICA

(Shimizu et al. (2006))

A linear acyclic SCM

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} 0 & b_{12} & \dots & b_{1d} \\ 0 & 0 & \dots & b_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} + \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{pmatrix}$$

with mutually independent components $S_1, \dots, S_d$

is closely linked to ICA (Independent Component Analysis) as per

$$X = \mathbf{B} \cdot X + S$$

$$\iff \qquad (\mathrm{Id} - \mathbf{B}) \cdot X = S$$

Inferring causal structure via ICA

(Shimizu et al. (2006))

A linear acyclic SCM

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} 0 & b_{12} & \dots & b_{1d} \\ 0 & 0 & \dots & b_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix} + \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{pmatrix}$$

with mutually independent components $S_1, \dots, S_d$

is closely linked to ICA (Independent Component Analysis) as per

$$\begin{aligned} & X = \mathbf{B} \cdot X + S \\ \iff \quad & (\mathrm{Id} - \mathbf{B}) \cdot X = S \\ \iff \quad & X = (\mathrm{Id} - \mathbf{B})^{-1} \cdot S \end{aligned}$$

Inferring causal structure via ICA

(Shimizu et al. (2006))

LiNGAM: Linear Non-Gaussian Acyclic Model $X = \mathbf{B}X + S$

Identify $\mathbf{B}$ via two steps:

Inferring causal structure via ICA

(Shimizu et al. (2006))

LiNGAM: Linear Non-Gaussian Acyclic Model $X = \mathbf{B}X + S$

Identify $\mathbf{B}$ via two steps:

1. infer $(\mathrm{Id} - \mathbf{B})$ up to scaling and permutation via ICA

Inferring causal structure via ICA

(Shimizu et al. (2006))

LiNGAM: Linear Non-Gaussian Acyclic Model $X = \mathbf{B}X + S$

Identify $\mathbf{B}$ via two steps:

1. infer $(\mathrm{Id} - \mathbf{B})$ up to scaling and permutation via ICA

   Non-Gaussianity!

Inferring causal structure via ICA

(Shimizu et al. (2006))

LiNGAM: Linear Non-Gaussian Acyclic Model $X = \mathbf{B}X + S$

Identify $\mathbf{B}$ via two steps:

1. infer $(\mathrm{Id} - \mathbf{B})$ up to scaling and permutation via ICA

   Non-Gaussianity!

2. resolve scaling and permutation to obtain $\mathbf{B}$

Inferring causal structure via ICA

(Shimizu et al. (2006))

LiNGAM: Linear Non-Gaussian Acyclic Model $X = \mathbf{B}X + S$

Identify $\mathbf{B}$ via two steps:

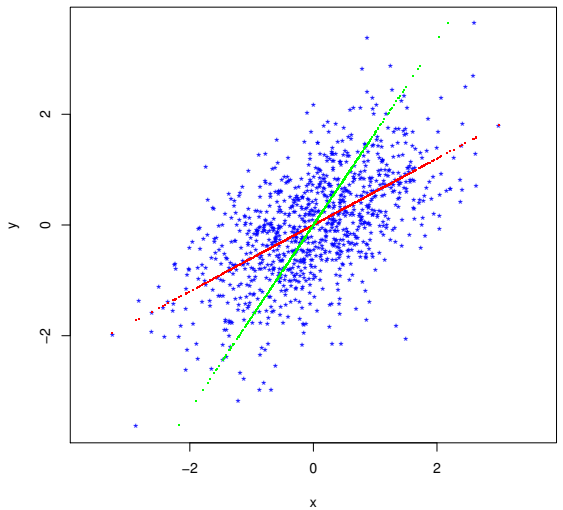
1. infer $(\mathrm{Id} - \mathbf{B})$ up to scaling and permutation via ICA

   Non-Gaussianity!

2. resolve scaling and permutation to obtain $\mathbf{B}$

   Acyclicity!

Bivariate Gaussian and Indeterminacies of ICA



The same distribution can be described as

$$X = N_X \qquad \text{or} \qquad X = \beta \cdot Y + N_X$$
$$Y = \alpha \cdot X + N_Y \qquad\qquad Y = N_Y$$

where N_X and N_Y are suitable independent Gaussian distributions

Linear non-Gaussian models

Kano & Shimizu 2003

Theorem

Let $X \not\perp\!\!\!\perp Y$. Then $P_{X,Y}$ admits linear models in both directions, i.e.,

$$\begin{aligned} Y &= \alpha X + N_Y \quad \text{with } N_Y \perp\!\!\!\perp X \\ X &= \beta Y + N_X \quad \text{with } N_X \perp\!\!\!\perp Y\,, \end{aligned}$$

if and only if $P_{X,Y}$ is bivariate Gaussian

- if $P_{X,Y}$ is non-Gaussian, there can be a linear model in at most one direction.
- LINGAM: causal direction is the one that admits a linear model

LiNGAM and confounding-robust ICA

(Pfister*, Weichwald*, et al. (2018) arXiv:1806.01094)

LiNGAM $\qquad X = \mathbf{B}X + S$

where S has mutually independent components

LiNGAM and confounding-robust ICA

(Pfister*, Weichwald*, et al. (2018) arXiv:1806.01094)

LiNGAM $\qquad X = \mathbf{B}X + S$

Confounded LiNGAM $\qquad X = \mathbf{B}X + S + H$

where S has mutually independent components

LiNGAM and confounding-robust ICA

(Pfister*, Weichwald*, et al. (2018) arXiv:1806.01094)

LiNGAM $\qquad X = \mathbf{B}X + S$

Confounded LiNGAM $\qquad X = \mathbf{B}X + S + H$

where S has mutually independent components
and H is *group-wise stationary* confounding

LiNGAM and confounding-robust ICA

(Pfister*, Weichwald*, et al. (2018) arXiv:1806.01094)

$$\text{LiNGAM} \qquad X = \mathbf{B}X + S$$
$$\iff X = (\mathrm{Id} - \mathbf{B})^{-1} S$$

$$\text{Confounded LiNGAM} \qquad X = \mathbf{B}X + S + H$$

where S has mutually independent components
and H is *group-wise stationary* confounding

LiNGAM and confounding-robust ICA

(Pfister*, Weichwald*, et al. (2018) arXiv:1806.01094)

$$\text{LiNGAM} \qquad X = \mathbf{B}X + S$$
$$\iff X = (\mathrm{Id} - \mathbf{B})^{-1} S$$

$$\text{Confounded LiNGAM} \qquad X = \mathbf{B}X + S + H$$
$$\iff X = (\mathrm{Id} - \mathbf{B})^{-1}(S + H)$$

where S has mutually independent components
and H is *group-wise stationary* confounding

LiNGAM and confounding-robust ICA

(Pfister*, Weichwald*, et al. (2018) arXiv:1806.01094)

$$\text{LiNGAM} \qquad X = \mathbf{B}X + S$$
$$\iff X = (\mathrm{Id} - \mathbf{B})^{-1} S$$

$$\text{Confounded LiNGAM} \qquad X = \mathbf{B}X + S + H$$
$$\iff X = (\mathrm{Id} - \mathbf{B})^{-1}(S + H)$$

where S has mutually independent components
and H is *group-wise stationary* confounding

$\rightsquigarrow$ coroICA allows to identify the confounded LiNGAM model
and accounts for dependencies due to H
if H is group-wise stationary

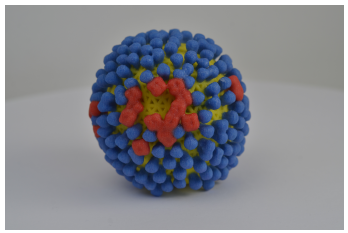
4. Macroscopic and microscopic causal models:

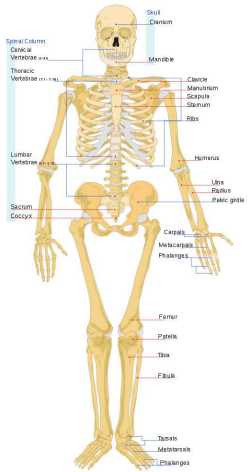
consistent coarse-graining of causal models

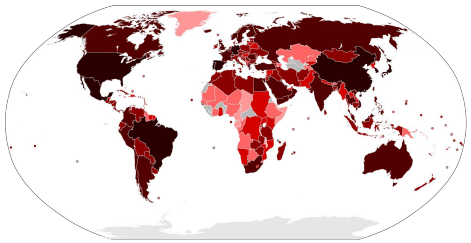
Models at different levels

Fine-grained ⟷ Coarse-grained

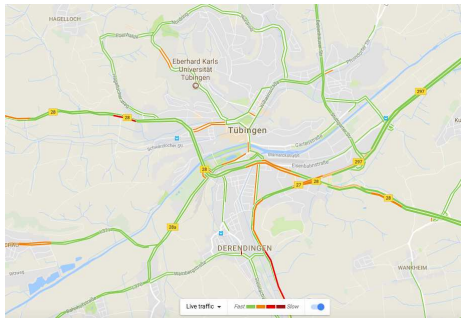
Disease

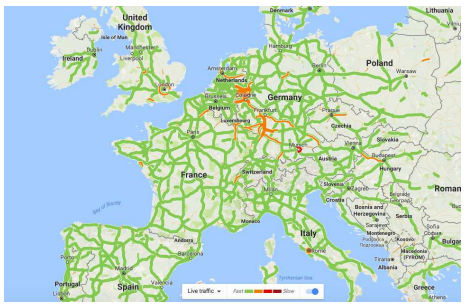




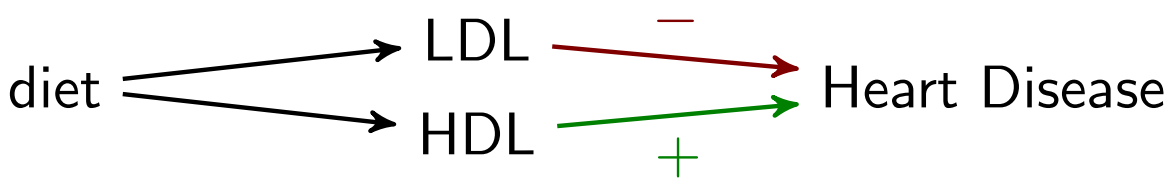


Traffic

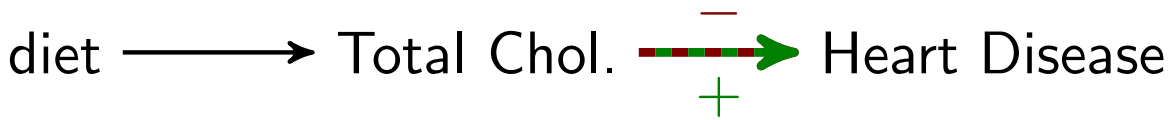


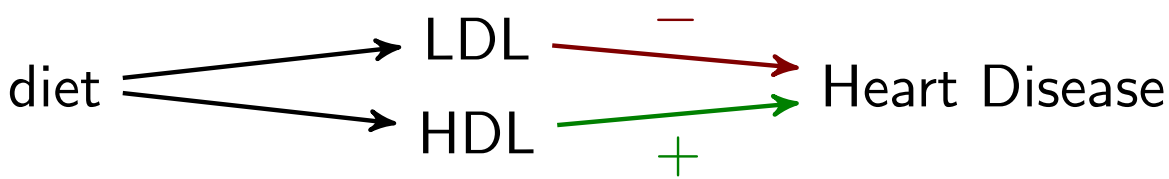


What can go wrong? Cholesterol and Heart Disease

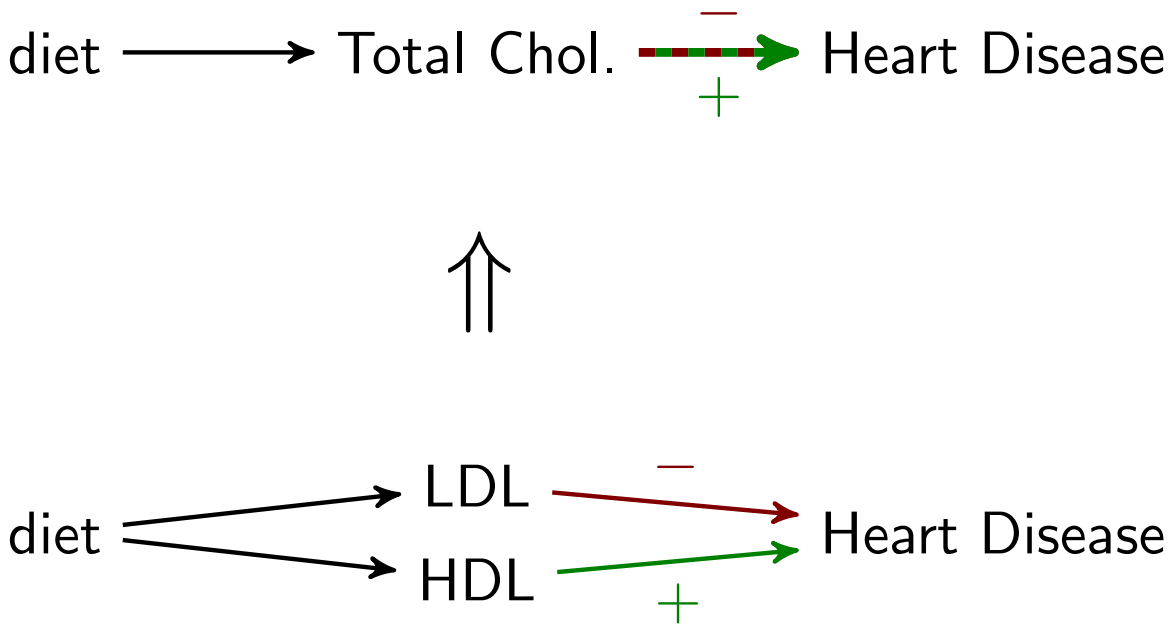


What can go wrong? Cholesterol and Heart Disease



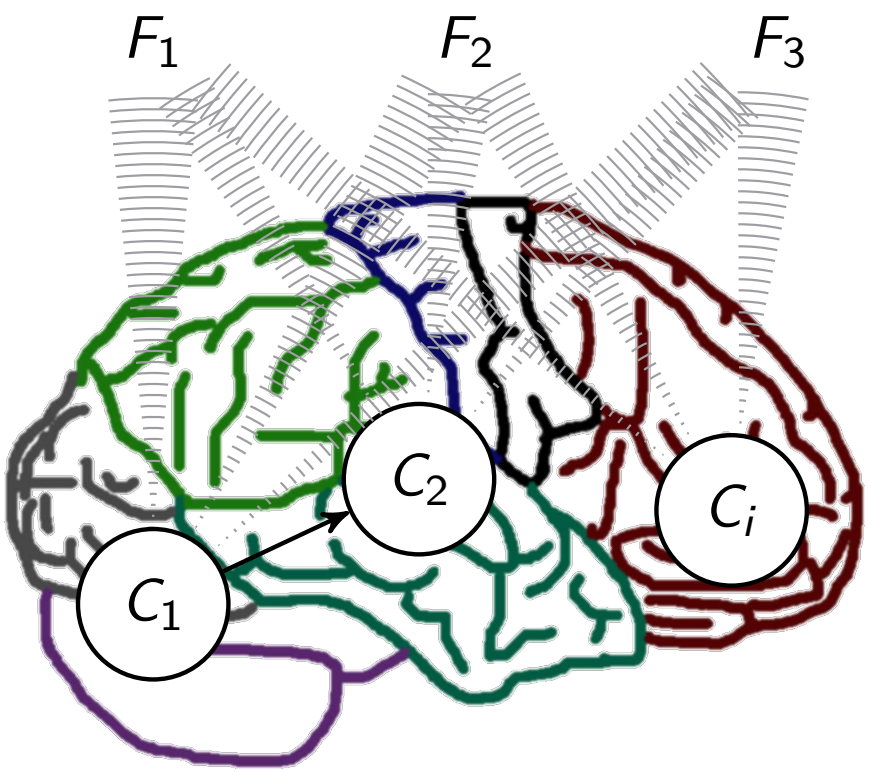


What can go wrong? Cholesterol and Heart Disease



Incorrectly 'transforming' the model can lead to problems.

Limited ability to observe breaks causal reasoning

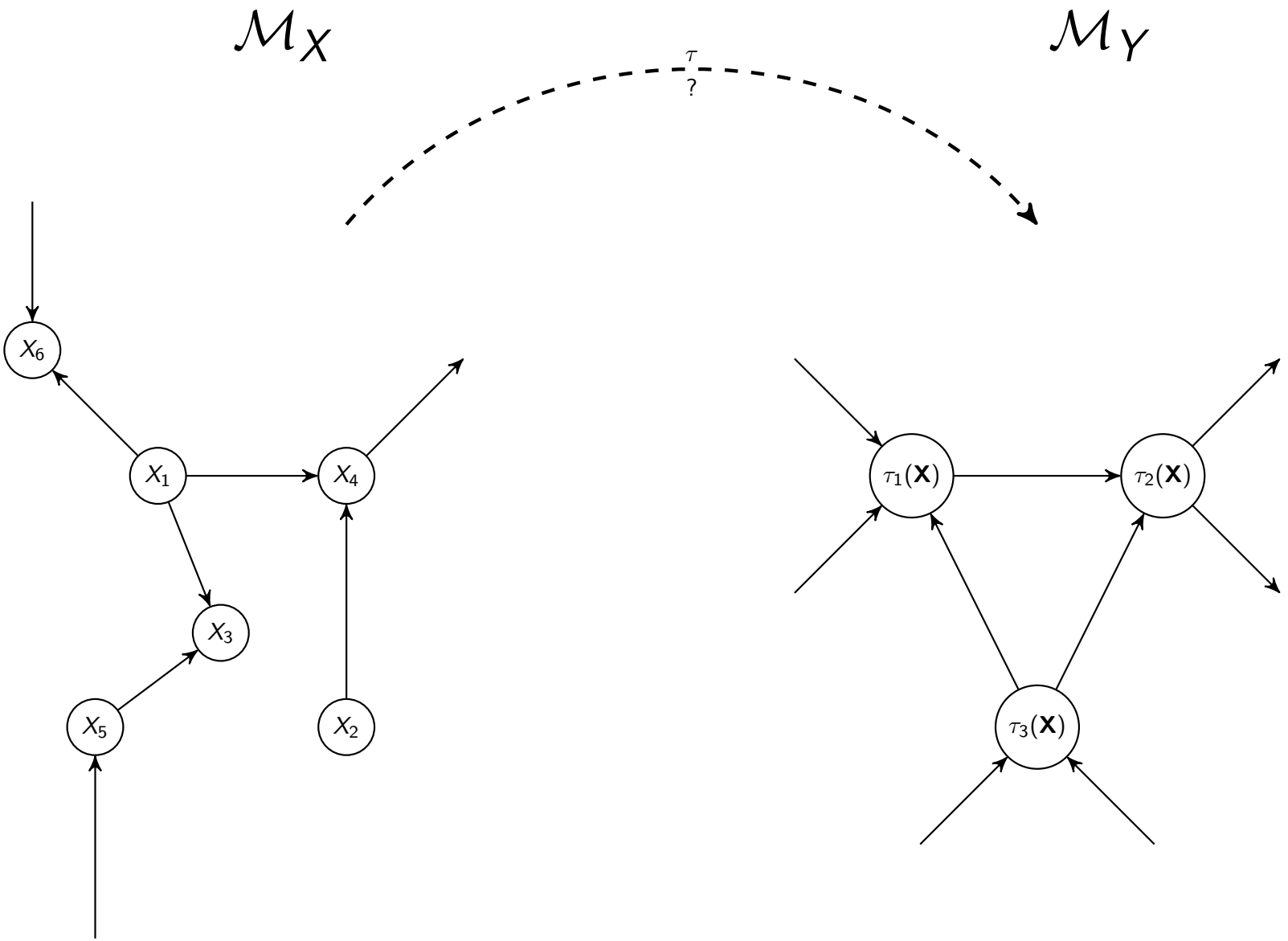


observed linear mixture

linear mixing

causal variables

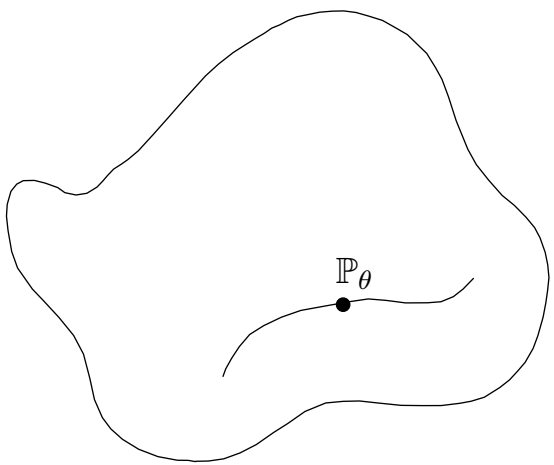
Transformations of causal models



Causal Models as Posets of Distributions

"Normal" Probabilistic Model:

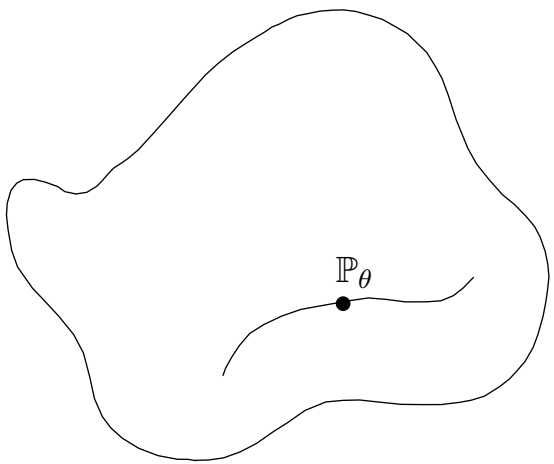
$\mathcal{M}_X : \theta \mapsto \mathbb{P}_\theta$



Causal Models as Posets of Distributions

"Normal" Probabilistic Model:

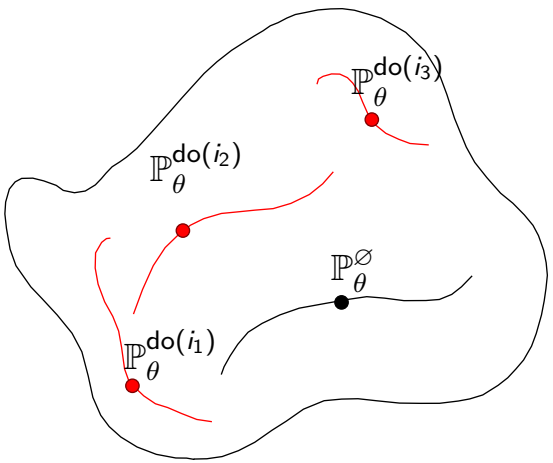
$\mathcal{M}_X : \theta \mapsto \mathbb{P}_\theta$



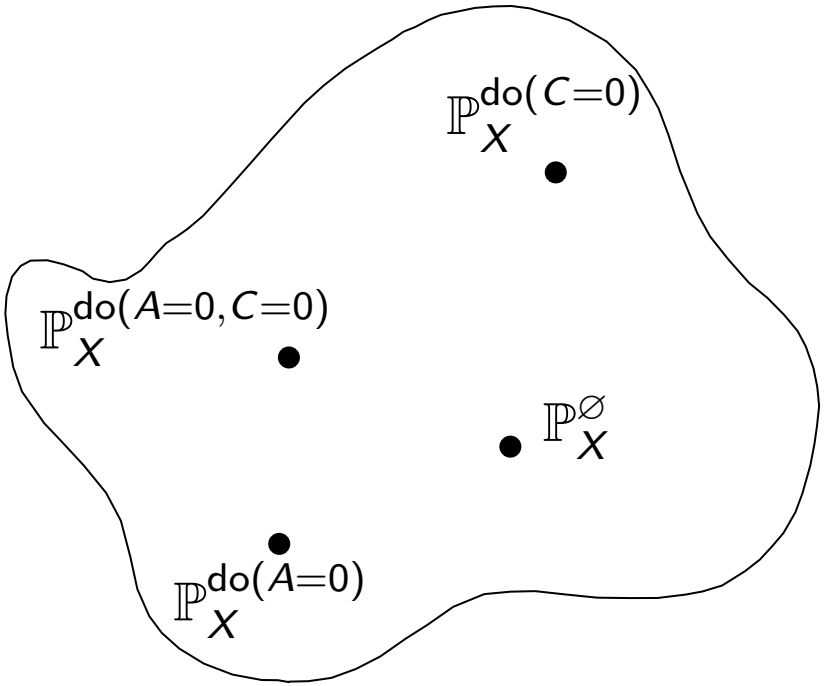
Causal Model:

$\mathcal{M}_X : \theta \mapsto \{\mathbb{P}_\theta^{\mathsf{do}(i)} \ : \ i \in \mathcal{I}_X\}$

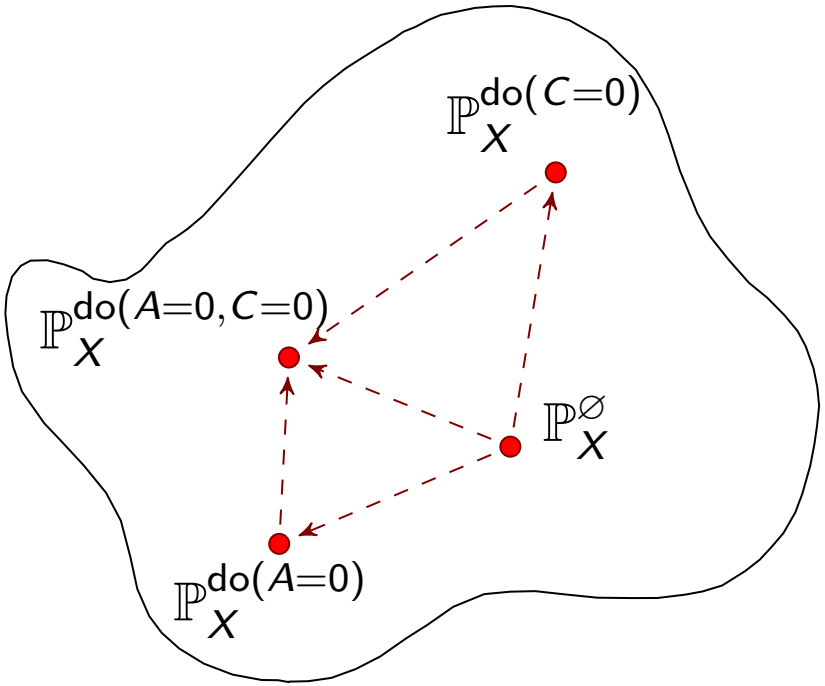
$\mathcal{I}_X$ is set of interventions.



Causal Models as Posets of Distributions

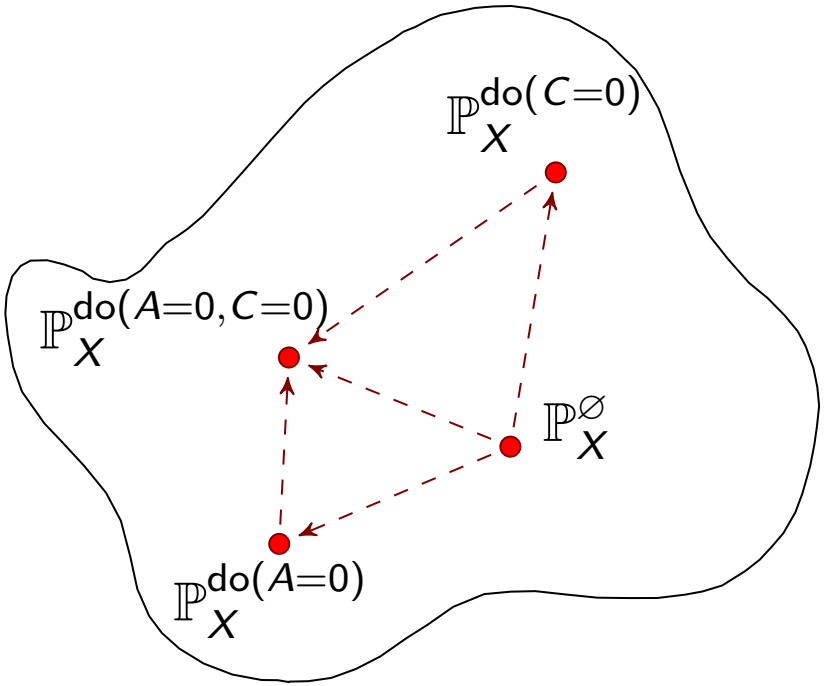


Causal Models as Posets of Distributions



$\mathcal{I}_X$ has partial ordering structure

Causal Models as Posets of Distributions



$\mathcal{I}_X$ has partial ordering structure

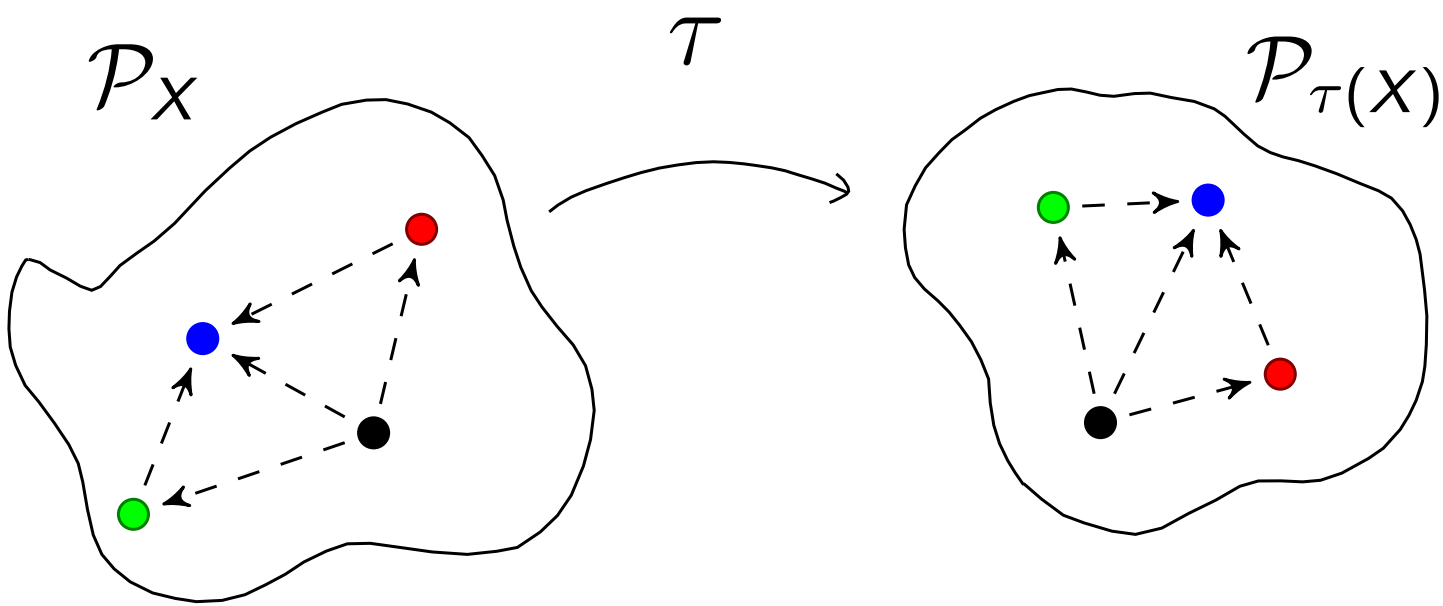
$\mathcal{M}_X$ implies the poset of distributions $\mathcal{P}_X := \left(\left\{ \mathbb{P}_X^{\mathrm{do}(i)} \quad : \quad i \in \mathcal{I}_X \right\}, \leq_X \right)$

Transformations of Structural Equation Models

Suppose we are given $\mathcal{M}_X$ and a 'measuring device' $\tau : \mathcal{X} \to \mathcal{Y}$

$X \sim \mathbb{P}_X$ an r.v. in $\mathcal{X} \implies \tau(X) \sim \mathbb{P}_{\tau(X)}$ is an r.v. in $\mathcal{Y}$

$$\tau : \mathcal{P}_X \to \mathcal{P}_{\tau(X)} = \left(\left\{ \mathbb{P}^i_{\tau(X)} \ : \ i \in \mathcal{I}_X \right\}, \leq_X \right)$$

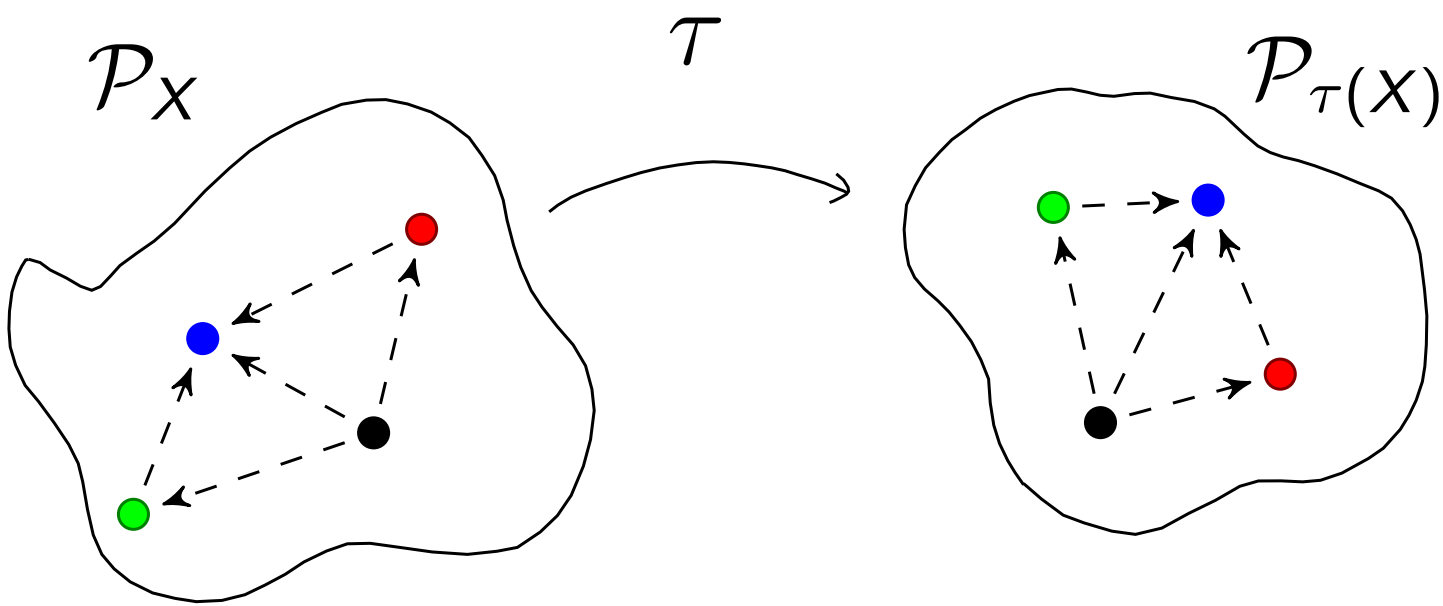


Transformations of Structural Equation Models

Suppose we are given $\mathcal{M}_X$ and a 'measuring device' $\tau : \mathcal{X} \to \mathcal{Y}$

$X \sim \mathbb{P}_X$ an r.v. in $\mathcal{X} \implies \tau(X) \sim \mathbb{P}_{\tau(X)}$ is an r.v. in $\mathcal{Y}$

$$\tau : \mathcal{P}_X \to \mathcal{P}_{\tau(X)} = \left(\left\{ \mathbb{P}^i_{\tau(X)} \ : \ i \in \mathcal{I}_X \right\}, \leq_X \right)$$



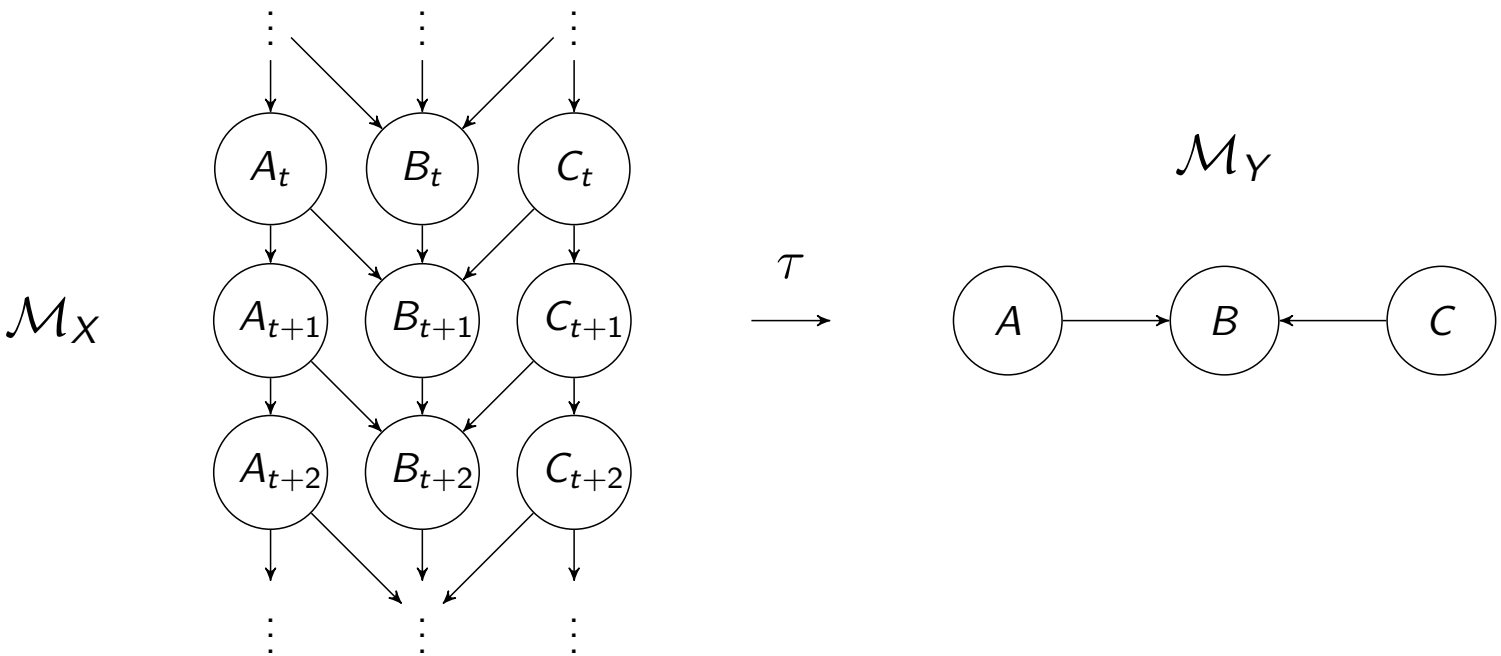
Does there exist an SEM $\mathcal{M}_Y$ with $\mathcal{P}_Y = \mathcal{P}_{\tau(X)}$?

If so, then $\mathcal{M}_Y$ will agree with our observations of $\mathcal{M}_X$ via τ.

In What Sense is Causal Reasoning Preserved?

Does there exist an SEM $\mathcal{M}_Y$ with $\mathcal{P}_Y = \mathcal{P}_{\tau(X)}$?

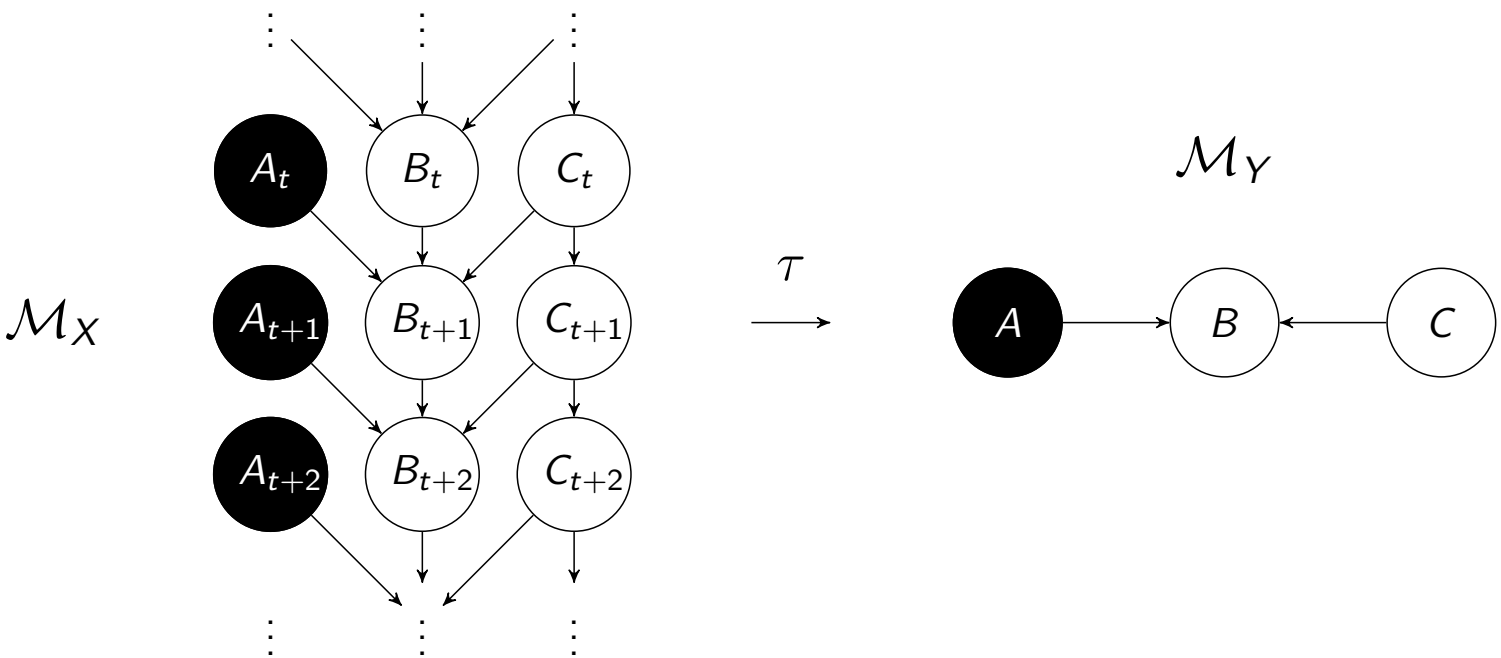
If so, then $\mathcal{M}_Y$ will agree with our observations of $\mathcal{M}_X$ via τ



In What Sense is Causal Reasoning Preserved?

Does there exist an SEM $\mathcal{M}_Y$ with $\mathcal{P}_Y = \mathcal{P}_{\tau(X)}$?

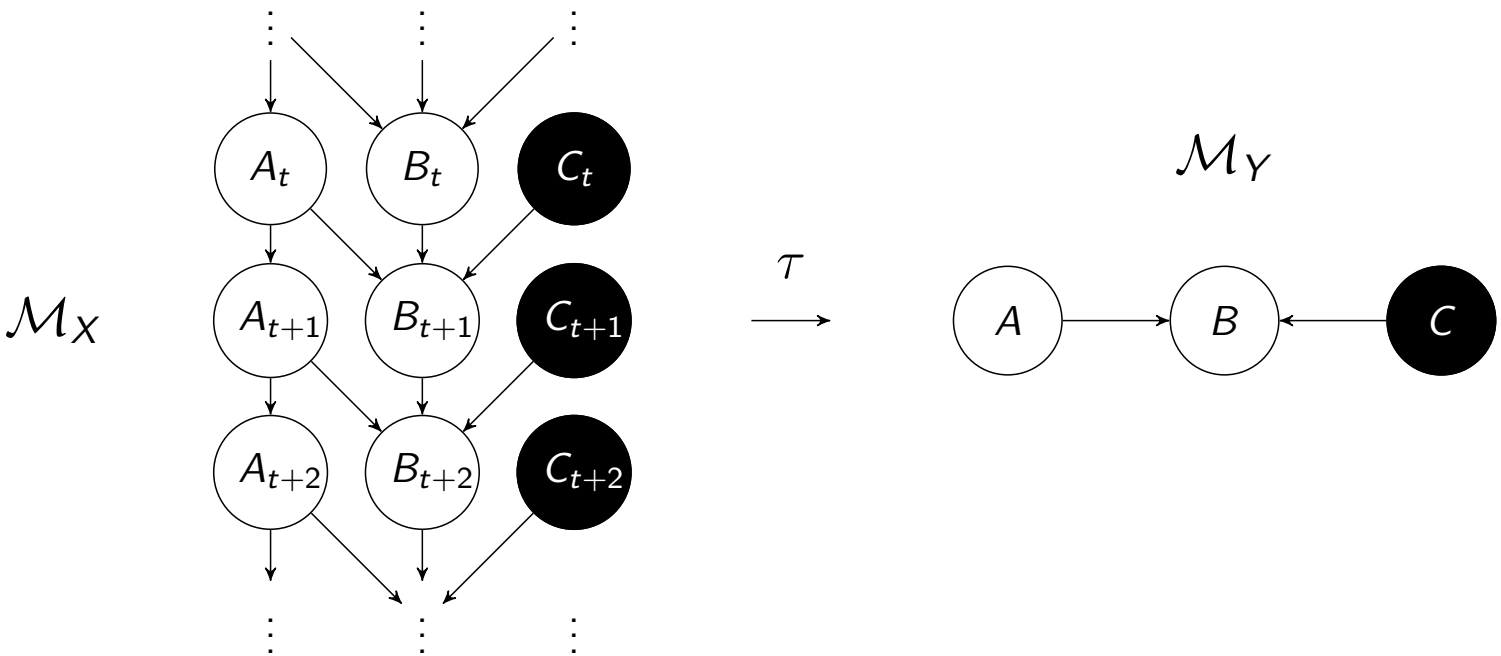
If so, then $\mathcal{M}_Y$ will agree with our observations of $\mathcal{M}_X$ via τ



In What Sense is Causal Reasoning Preserved?

Does there exist an SEM $\mathcal{M}_Y$ with $\mathcal{P}_Y = \mathcal{P}_{\tau(X)}$?

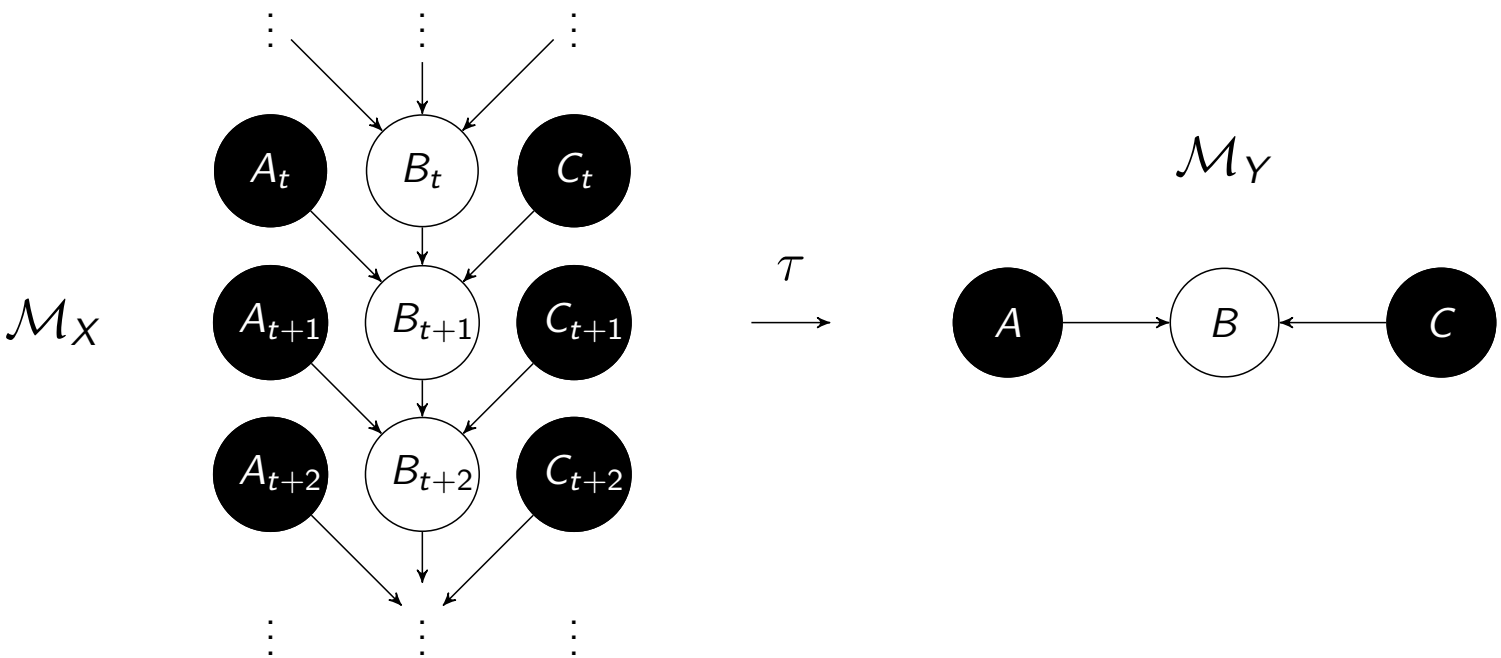
If so, then $\mathcal{M}_Y$ will agree with our observations of $\mathcal{M}_X$ via τ



In What Sense is Causal Reasoning Preserved?

Does there exist an SEM $\mathcal{M}_Y$ with $\mathcal{P}_Y = \mathcal{P}_{\tau(X)}$?

If so, then $\mathcal{M}_Y$ will agree with our observations of $\mathcal{M}_X$ via τ



Compositions of interventions are preserved!

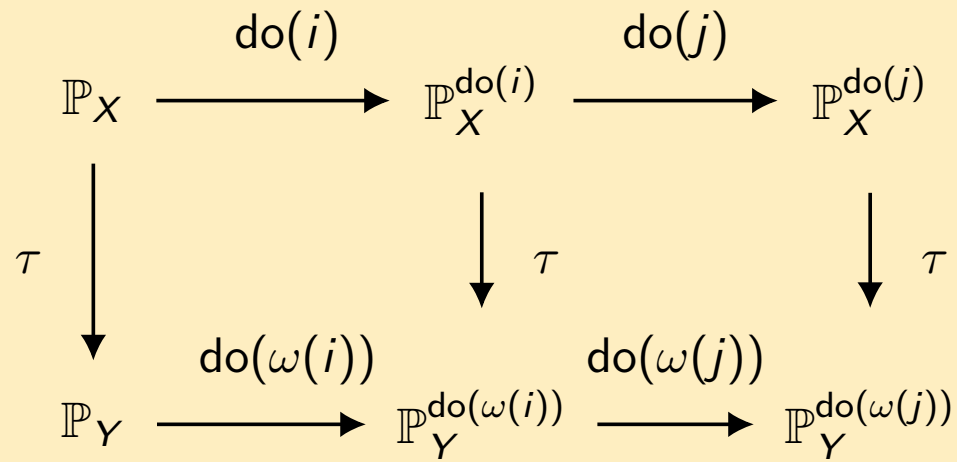
Definition (Exact Transformations between SEMs)

Let $\mathcal{M}_X$ and $\mathcal{M}_Y$ be SEMs and $\tau : \mathcal{X} \to \mathcal{Y}$ be a function. We say $\mathcal{M}_Y$ is an *exact* τ-transformation of $\mathcal{M}_X$ if there exists a *surjective order-preserving* map $\omega : \mathcal{I}_X \to \mathcal{I}_Y$ such that

$$\mathbb{P}^{i}_{\tau(X)} = \mathbb{P}^{\mathrm{do}(\omega(i))}_{Y} \quad \forall i \in \mathcal{I}_X$$

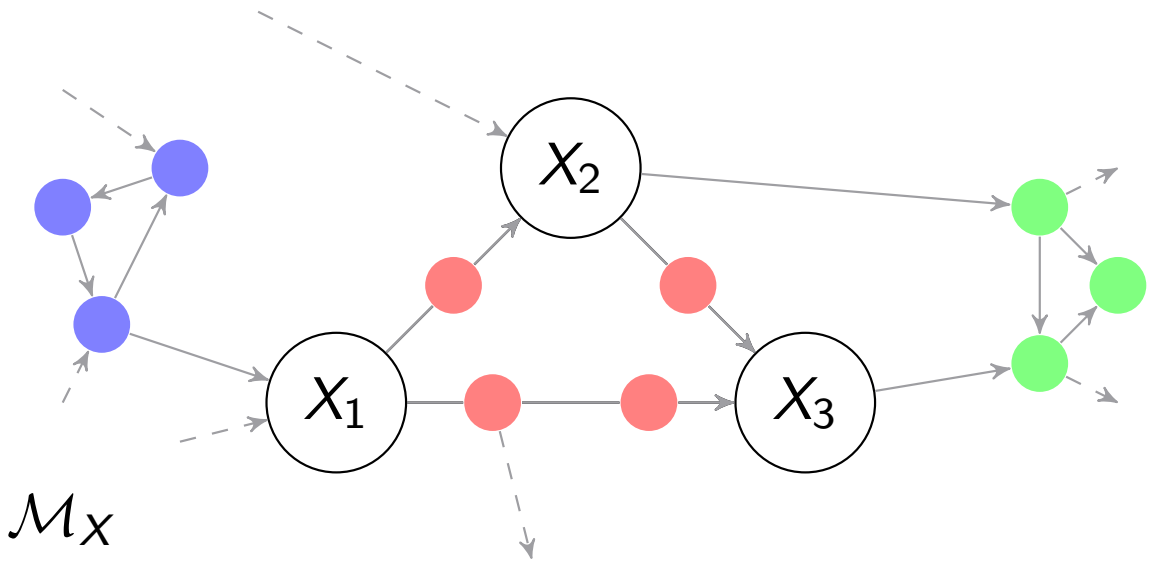
Theorem

The following diagram commutes:



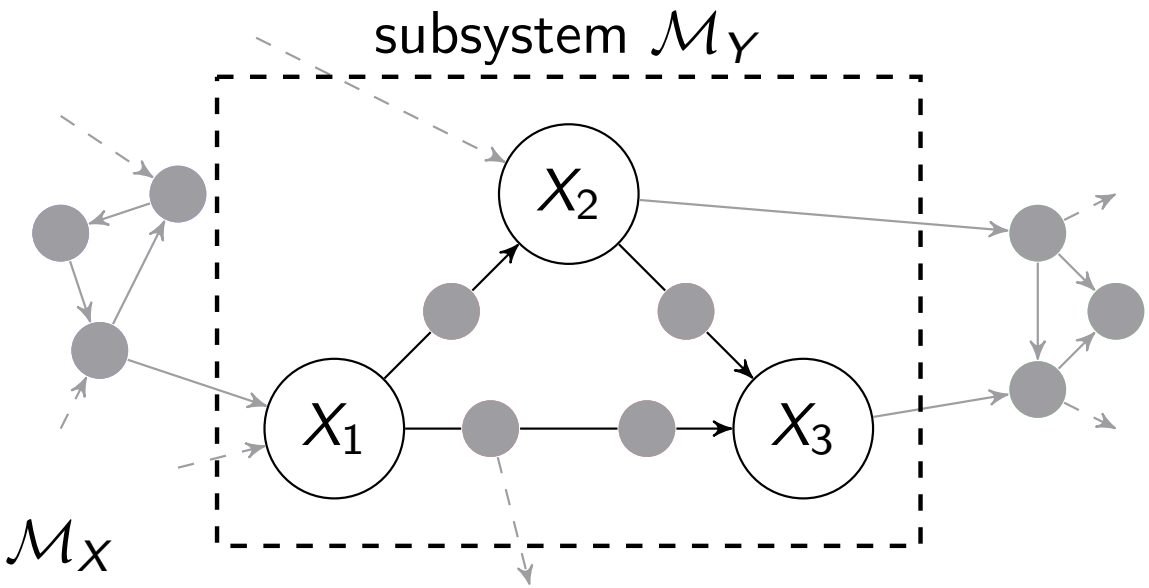
Transformations for Pragmatic Causal Models

- Marginalisation of variables



Transformations for Pragmatic Causal Models

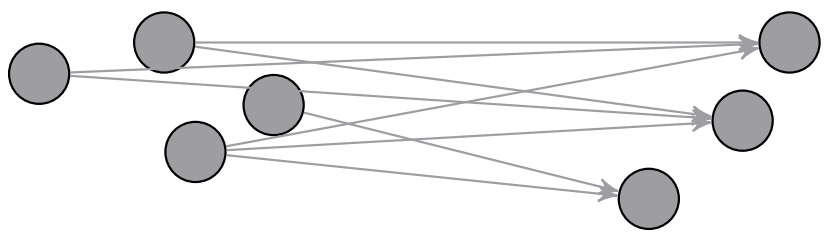
- Marginalisation of variables



Transformations for Pragmatic Causal Models

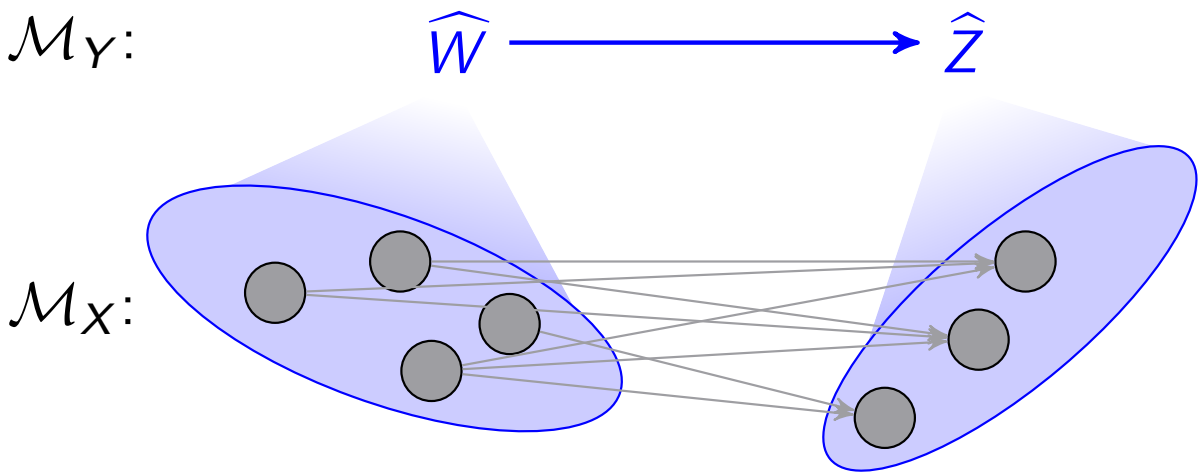
- Marginalisation of variables
- Micro- to macro-level and aggregate features

$\mathcal{M}_X$:



Transformations for Pragmatic Causal Models

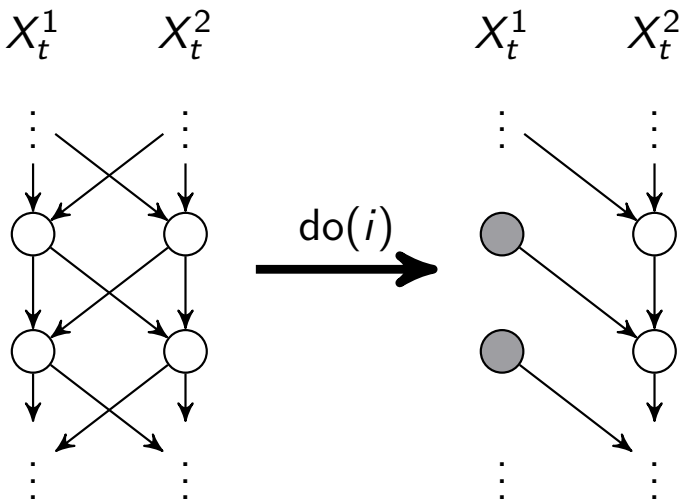
- Marginalisation of variables
- Micro- to macro-level and aggregate features



Transformations for Pragmatic Causal Models

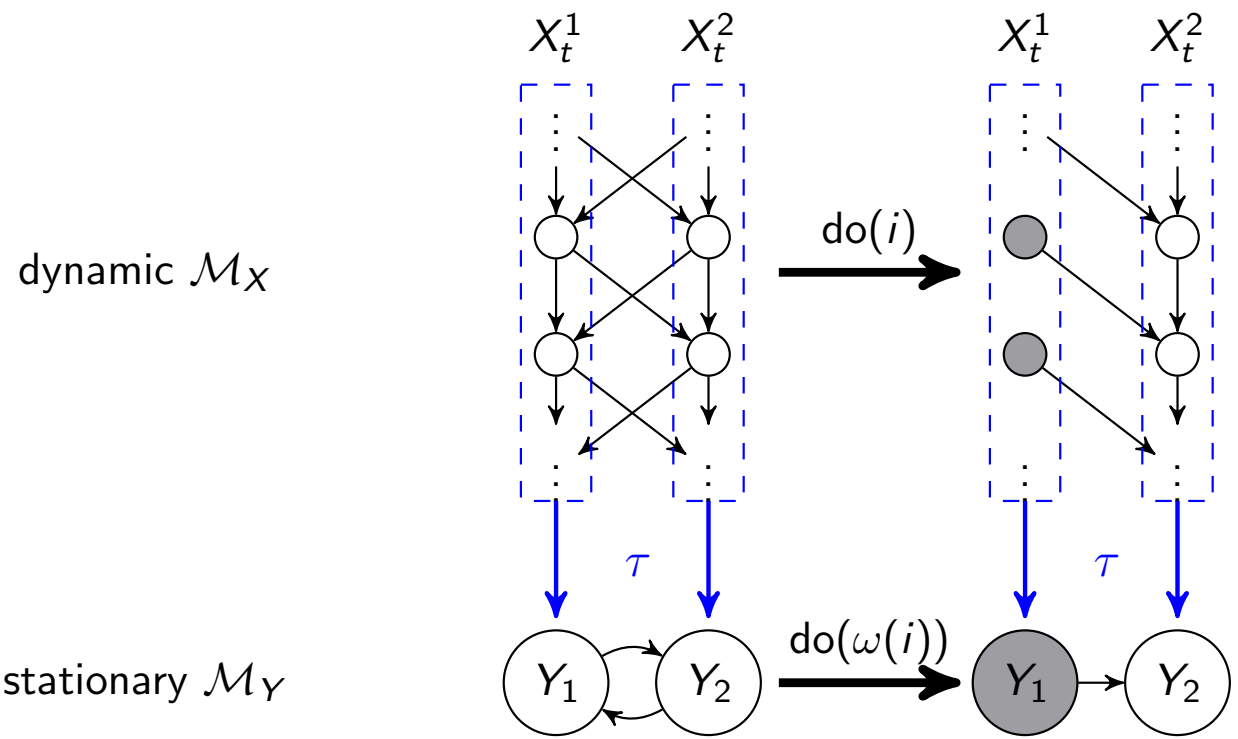
- Marginalisation of variables
- Micro- to macro-level and aggregate features
- Stationary behaviour of dynamical processes

dynamic $\mathcal{M}_X$



Transformations for Pragmatic Causal Models

- Marginalisation of variables
- Micro- to macro-level and aggregate features
- Stationary behaviour of dynamical processes



5. Causal inference in time series:

Granger causality and its limitations

Granger Causality

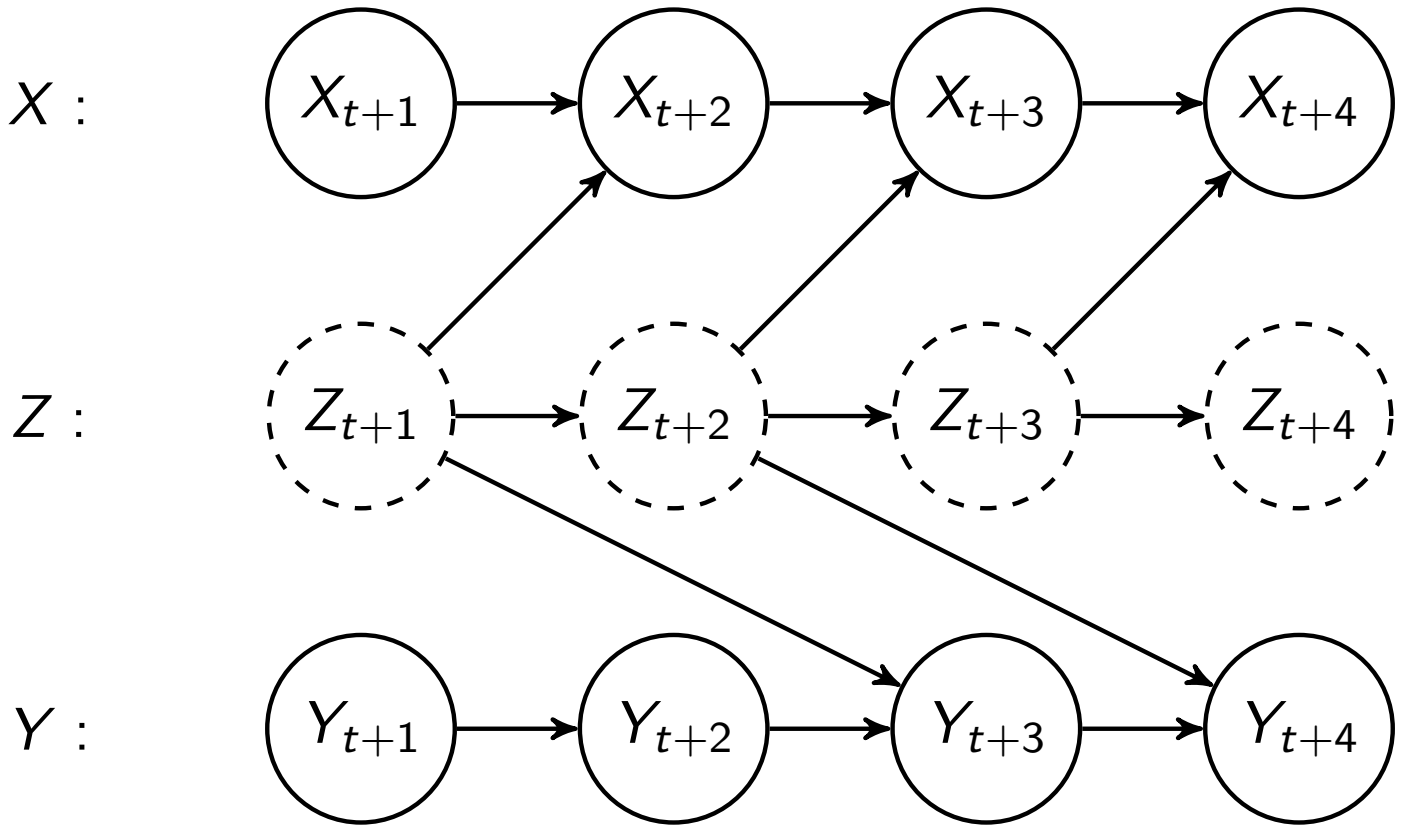
Simplified Definition: One stochastic process X is causal to a second Y if the autoregressive predictability of the second process at a given time point is improved by *including* measurements from the past of the first, i. e. if

$$\text{PredAcc}[Y_t|Y_{<t}] \;<\; \text{PredAcc}[Y_t|Y_{<t}, X_{<t}]$$

(*not* by C Granger)

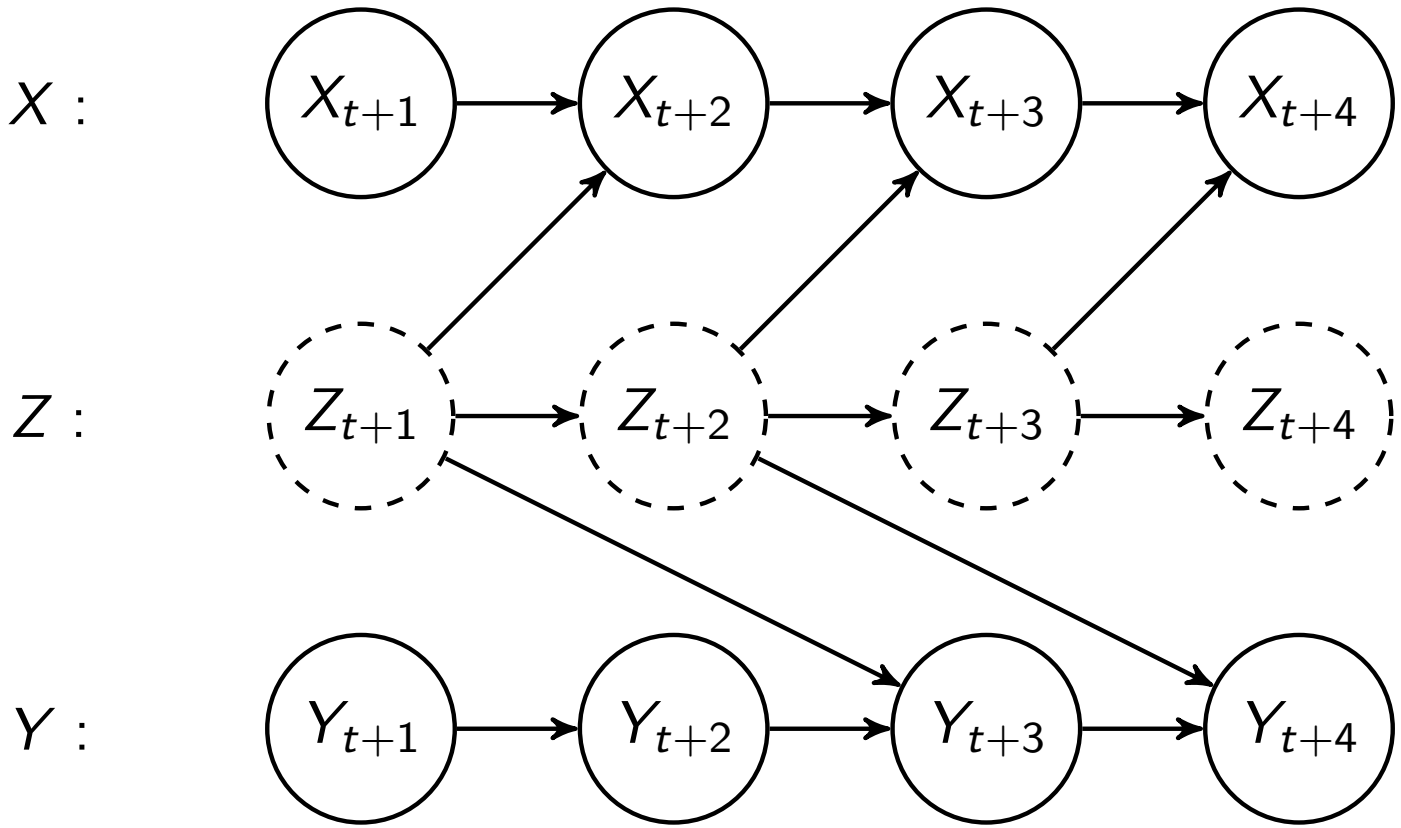
Granger Causality

(J Peters et al. Causal discovery on time series using restricted structural equation models. *NIPS*, 2013)



Granger Causality

(J Peters et al. Causal discovery on time series using restricted structural equation models. *NIPS*, 2013)



$$\text{PredAcc}[Y_t|Y_{<t}] \quad < \quad \text{PredAcc}[Y_t|Y_{<t}, X_{<t}]$$

Granger causality erroneously infers causal influence from X to Y!

Granger Causality

Simplified Definition: One stochastic process X is causal to a second Y if the autoregressive predictability of the second process at a given time point is improved by *including* measurements from the past of the first, i. e. if

$$\text{PredAcc}[Y_t|Y_{<t}] \quad < \quad \text{PredAcc}[Y_t|Y_{<t}, X_{<t}]$$

(*not* by C Granger)

Granger Causality

(CWJ Granger, Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 1969)

Simplified Definition: One stochastic process X is causal to a second Y if the autoregressive predictability of the second process at a given time point is improved by *including* measurements from the past of the first, i. e. if

$$\text{PredAcc}[Y_t|Y_{<t}] \;<\; \text{PredAcc}[Y_t|Y_{<t}, X_{<t}]$$

(*not* by C Granger)

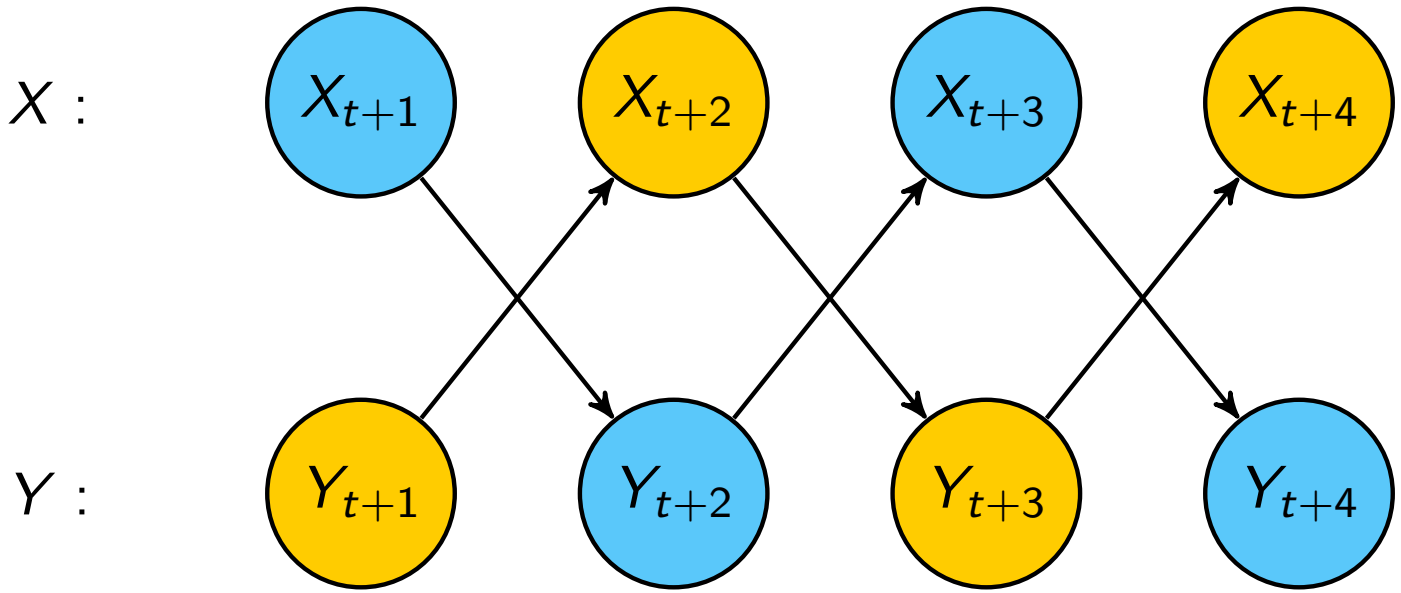
Granger's Definition: One stochastic process X is causal to a second Y if the predictability of the second process at a given time point is worsened by *removing* past measurements of the first from the universe's past, i. e. if

$$\text{PredAcc}[Y_t|\text{🌍}_{<t}] \;>\; \text{PredAcc}[Y_t|\text{🌍}_{<t} \setminus X_{<t}]$$

(by C Granger)

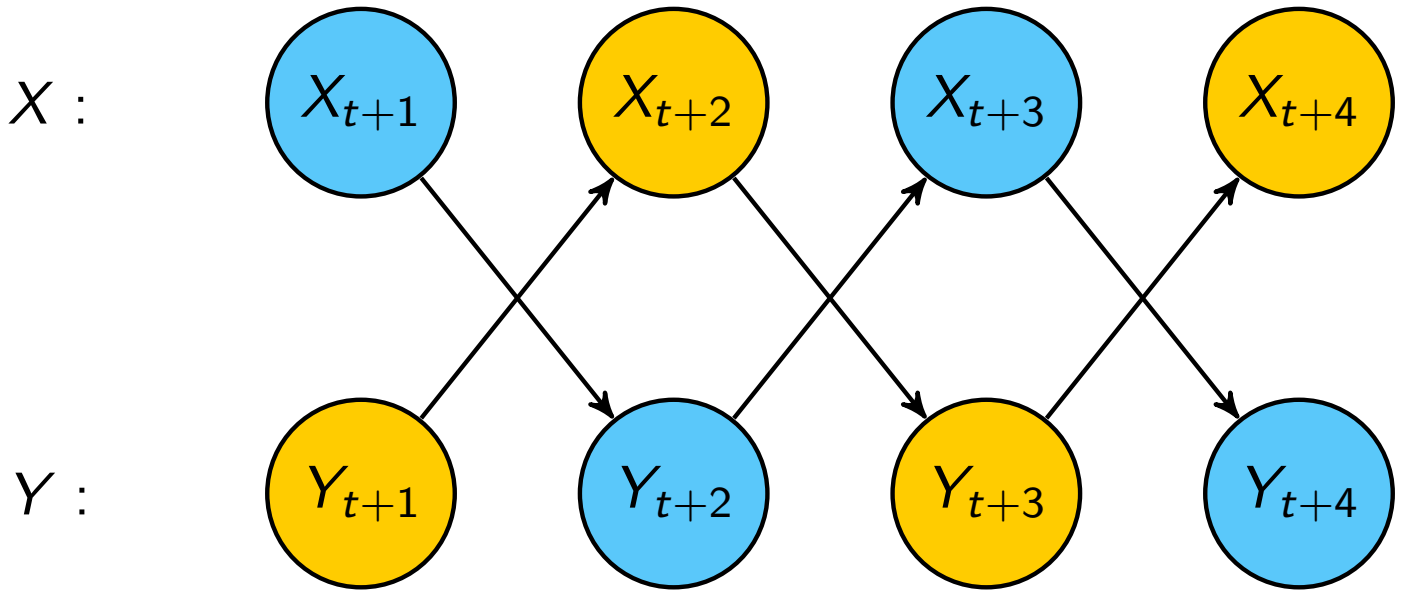
Granger Causality

(N Ay and D Polani, Information flows in causal networks. *Advances in Complex Systems*, 2008)



Granger Causality

(N Ay and D Polani, Information flows in causal networks. *Advances in Complex Systems*, 2008)



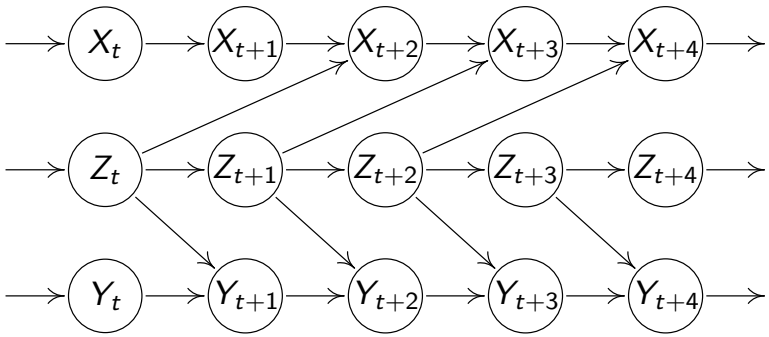
$$\text{PredAcc}[Y_t|\text{🌍}_{<t}] \not> \text{PredAcc}[Y_t|\text{🌍}_{<t} \setminus X_{<t}]$$

Granger causality fails to predict the effects of interventions!

Granger works under Markov and faithfulness

Assumptions:

- no hidden common causes
- no instantaneous effects



e.g. Theorem 10:3 in Peters, Janzing, Schölkopf: *Elements of Causal Inference*

If the distribution is Markov and faithful relative to the causal DAG,
then there exists arrows from $Y_{<t}$ to X_t if and only if Y Granger-causes X, i.e. $X_t \not\perp\!\!\!\perp Y_{<t} | \text{🌍}_{<t} \setminus Y_{<t}$

6. Causal relations among individual objects

algorithmic Markov conditions, analogy to probabilistic Markov conditions

causal conclusions in real life are not always based on statistics!

these 2 objects are similar...

– *why* are they so similar?

Conclusion: common history

similarities require an *explanation*

what kind of similarities require an explanation?

here we would *not* assume that anyone has copied the design...

..the pattern is too simple

- similarities require an explanation only if the pattern is sufficiently complex

consider a binary sequence

Experiment:
2 persons are instructed to write down a string with 1000 digits

Result:
Both write 1100100100001111110110101010001...
(all 1000 digits coincide)

the **naive** statistician concludes

"There must be an agreement between the subjects"

correlation coefficient 1 (between digits) is highly significant for sample size 1000 !

- reject statistical independence
- infer the existence of a causal relation

another mathematician recognizes...

$$11.001001000011111101101010100 1... = \pi$$

- subjects may have come up with this number independently because it follows from a simple law
- superficially strong similarities are not necessarily significant if the pattern is too simple

How do we measure simplicity versus complexity of patterns / objects?

Kolmogorov complexity

(Kolmogorov 1965, Chaitin 1966, Solomonoff 1964)
of a binary string x

- K(x) = length of the shortest program with output x (on a Turing machine)
- interpretation: number of bits required to describe the rule that generates x
  neglect string-independent additive constants; use $\stackrel{+}{=}$ instead of $=$
- strings x, y with low $K(x)$, $K(y)$ cannot have much in common
- $K(x)$ is uncomputable
- probability-free definition of information content

Conditional Kolmogorov complexity

- $K(y|x)$: length of the shortest program that generates y from the input x.
- number of bits required for describing y if x is given
- $K(y|x^*)$ length of the shortest program that generates y from x^*, i.e., the shortest compression x.
- subtle difference: x can be generated from x^* but not vice versa because there is no algorithmic way to find the shortest compression

Algorithmic mutual information

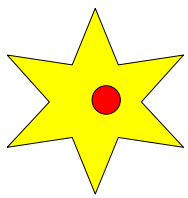
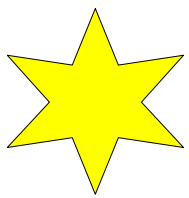
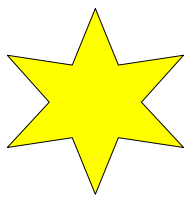
Chaitin, Gacs

Information of x about y (and vice versa)

- $I(x:y) := K(x) + K(y) - K(x,y)$
  $\stackrel{+}{=} K(x) - K(x|y^*) \stackrel{+}{=} K(y) - K(y|x^*)$

- Interpretation: number of bits saved when compressing x, y jointly rather than compressing them independently

Algorithmic mutual information: example

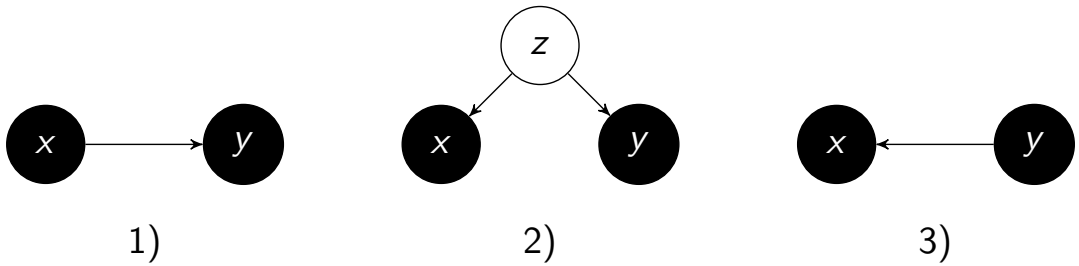
I( : ) = K()

Analogy to statistics:

- replace strings x, y (=objects) with random variables X, Y
- replace Kolmogorov complexity with Shannon entropy
- replace algorithmic mutual information $I(x : y)$ with statistical mutual information $I(X; Y)$

Causal Principle

If two strings x and y are algorithmically dependent then either



- every algorithmic dependence is due to a causal relation
- algorithmic analog to Reichenbach's principle of common cause
- distinction between 3 cases: use conditional independences on more than 2 objects

DJ, Schölkopf IEEE TIT 2010

conditional algorithmic mutual information

- $I(x : y|z) = K(x|z) + K(y|z) - K(x, y|z)$
- Information that x and y have in common when z is already given
- Formal analogy to statistical mutual information:

$$I(X : Y|Z) = S(X|Z) + S(Y|Z) - S(X, Y|Z)$$

- Define conditional independence:

$$I(x : y|z) \approx 0 :\Leftrightarrow x \perp\!\!\!\perp y|z$$

Algorithmic Markov condition

Postulate [DJ & Schölkopf IEEE TIT 2010]

Let $x_1, ..., x_n$ be some observations (formalized as strings) and G describe their causal relations.
Then, every x_j is conditionally algorithmically independent of its non-descendants, given its parents, i.e.,

$$x_j \perp\!\!\!\perp nd_j \,|pa_j^*$$

Equivalence of algorithmic Markov conditions

Theorem

For n strings $x_1, ..., x_n$ the following conditions are equivalent

- **Local Markov condition:**

$$I(x_j : nd_j | pa_j^*) \stackrel{+}{=} 0$$

- **Global Markov condition:**
  R d-separates S and T implies $I(S : T|R^*) \stackrel{+}{=} 0$
- **Recursion formula for joint complexity**

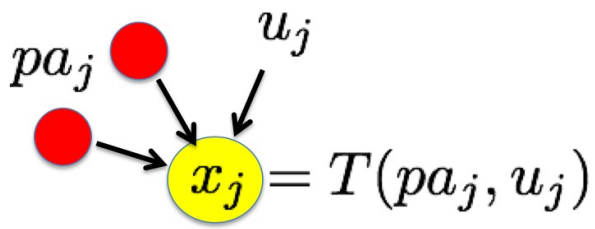
$$K(x_1, ..., x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j | pa_j^*)$$

$\rightarrow$ another analogy to statistical causal inference

Algorithmic model of causality

Given n causality related strings $x_1, \ldots, x_n$

- each x_j is computed from its parents pa_j and an unobserved string u_j by a Turing machine T



- all u_j are algorithmically independent
- each u_j describes the causal mechanism (the program) generating x_j from its parents
- u_j is the analog of the noise term in the statistical functional model

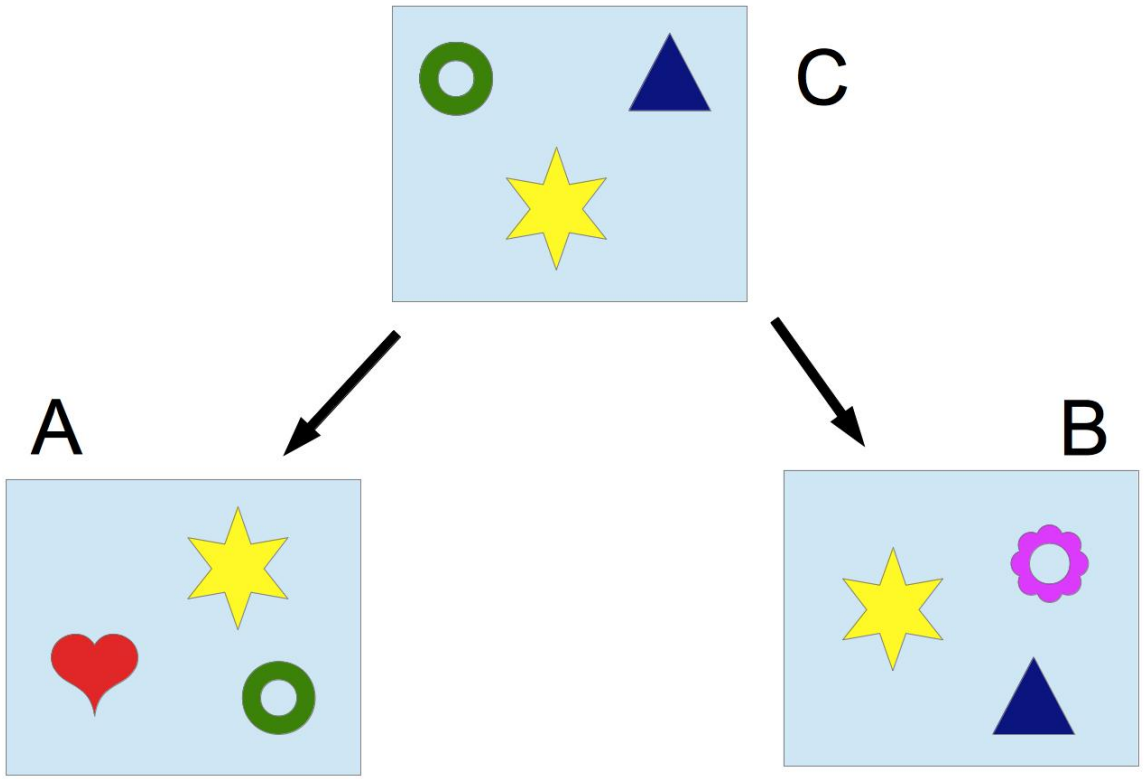
Algorithmic model of causality implies Markov condition

Theorem

If $x_1, \ldots, x_n$ are generated by an algorithmic model of causality according to the DAG G then they satisfy the 3 equivalent algorithmic Markov conditions.

Causal inference for single objects

3 carpets



conditional independence $A \perp\!\!\!\perp B \,|\, C$

Take home messages

- Graphical causal models do not *solve* the hard causal problems, but they provide a clear framework to address them
- Subject to *strong* assumptions, causal structure can also be inferred from passive observation
- However, machine learning is used to rely on strong assumptions

References


J. Pearl.
Causality.
Cambridge University Press, 2000.


P. Spirtes, C. Glymour, and R. Scheines.
Causation, Prediction, and Search.
Springer-Verlag, New York, NY, 1993.


Guido W. Imbens and Donald B. Rubin.
Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.
Cambridge University Press, 2015.


J. Peters, D. Janzing, and B. Schölkopf.
Elements of Causal Inference – Foundations and Learning Algorithms.
MIT Press, 2017.

Thank you for your attention!

note also the following competition:
`https://causeme.uv.es/neurips2019/`