Learning Causal Relations from Data

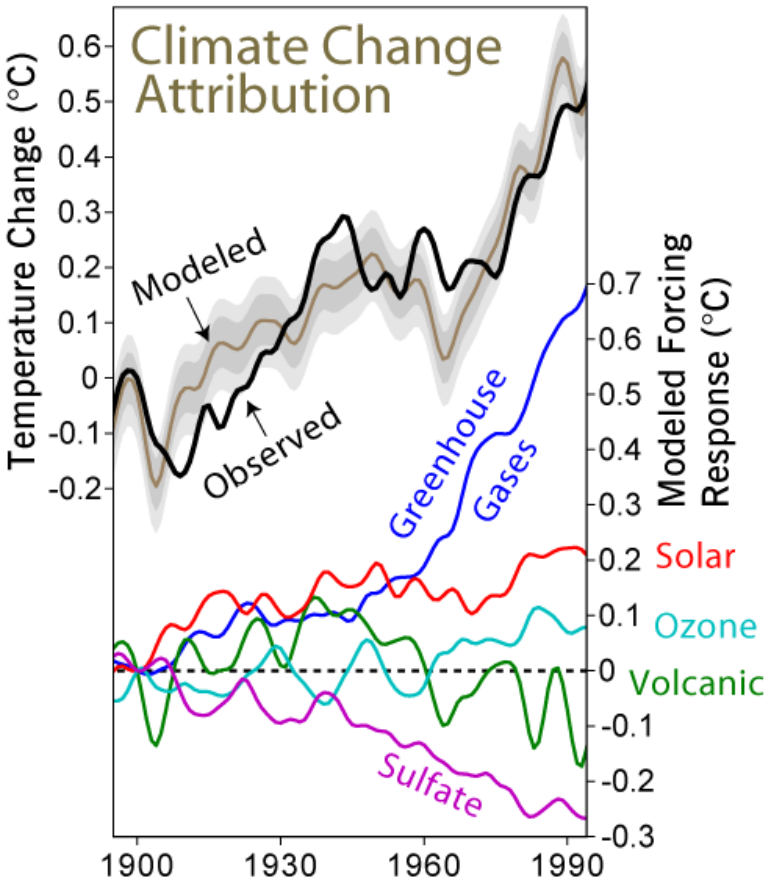
Joris Mooij

j.m.mooij@uva.nl

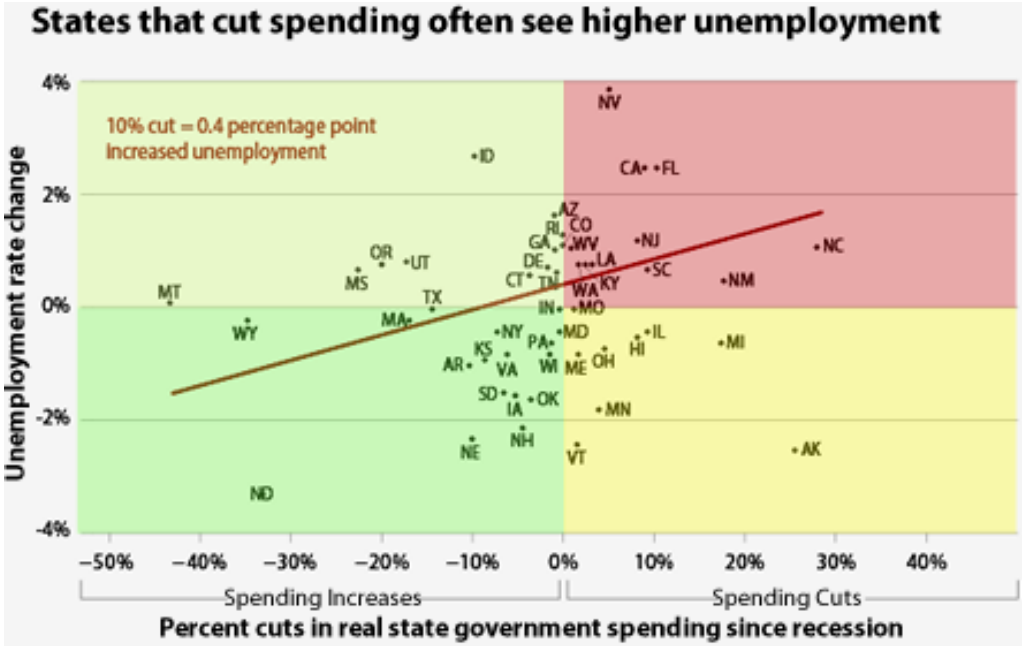
UNIVERSITY OF AMSTERDAM

Many questions in science are *causal*

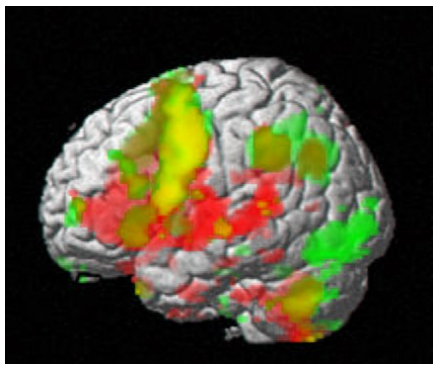
Climatology:

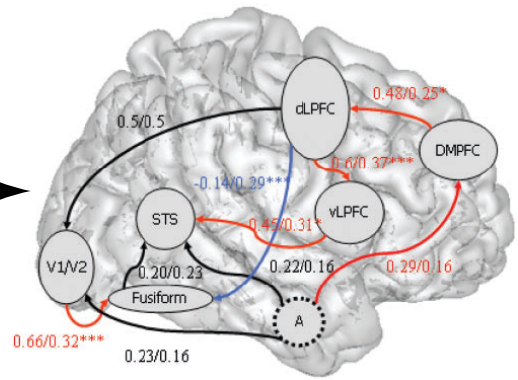


Economy:



Neuroscience:





Probabilistic inference vs. causal inference

Traditional statistics, machine learning

- Models the **distribution** of the data
- Focuses on prediction from **observations**
- Useful e.g. in medical diagnosis: given the symptoms of the patient, what is the most likely disease?

Causal modeling, reasoning, learning, inference

- Models the **mechanism** that generates the data
- Also allows to predict results of **interventions**
- Useful e.g. in medical treatment: if we treat the patient with a drug, will it cure the disease?

Causality is essential to answer questions of the type: *given the circumstances, what action should we take to achieve a certain goal?*

Causation or just correlation?

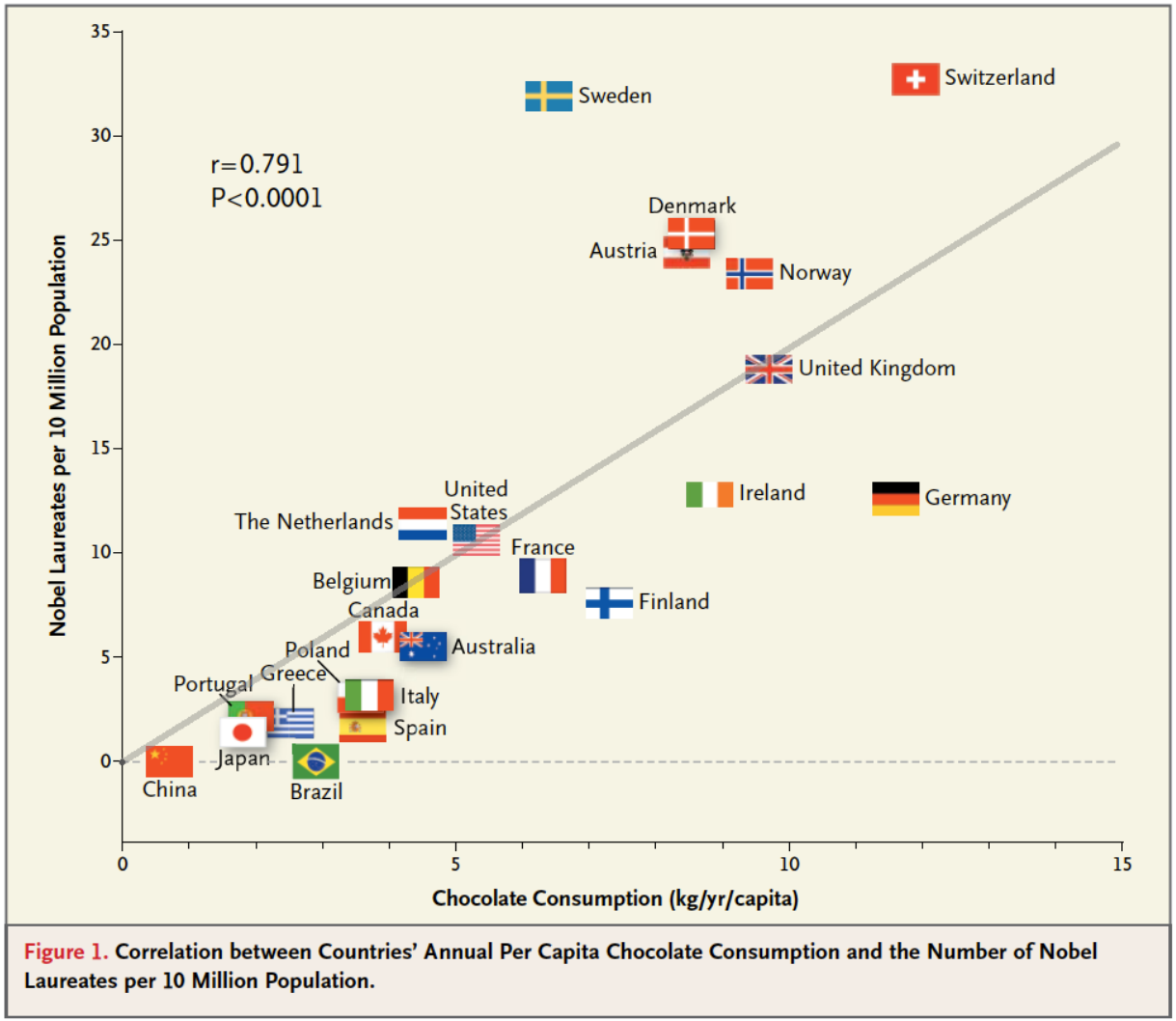


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: Messerli, New England Journal of Medicine (2012)

Causation or just correlation?

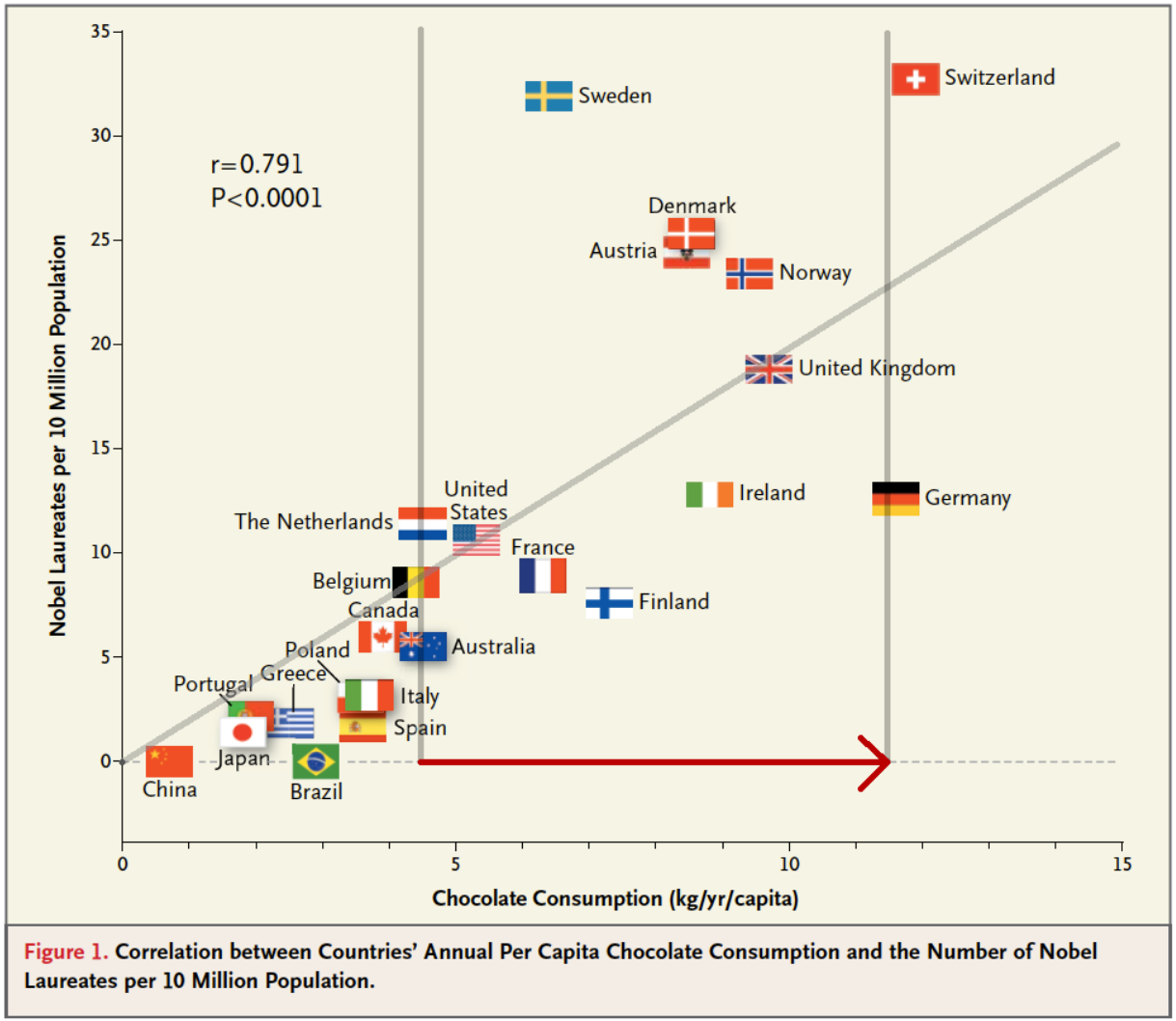


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: Messerli, New England Journal of Medicine (2012)

Should NWO fund chocolate for researchers?

How to formalize, model, estimate and exploit causality?

Intuition

Let A and B be two variables of a system.

> *A causes B if external interventions that change A result in a change of B.*

How to formalize, model, estimate and exploit causality?

Intuition

Let A and B be two variables of a system.

> *A causes B if external interventions that change A result in a change of B.*

But...

- How to model causality mathematically?
- How to model causality and uncertainty (statistical causal modeling)?
- How to discover causal relations from data?
- How to estimate the strength of causal relations from data?
- How to use knowledge of the causal relations for making predictions?
- How to exploit knowledge of the causal relations for control and optimization purposes?

Part I

Statistical Causal Modeling

Statistical Causal Modeling

Given a standard measurable space $(\mathcal{X}, \Sigma)$.

Definition (sketch)

A statistical model is a family of probability distributions: $\theta \mapsto \mathbb{P}_\theta(\cdot)$ where $\mathbb{P}_\theta$ is a probability measure on $\mathcal{X}$ for each parameter value $\theta \in \Theta$.

Statistical Causal Modeling

Given a standard measurable space $(\mathcal{X}, \Sigma)$.

Definition (sketch)

A statistical model is a family of probability distributions: $\theta \mapsto \mathbb{P}_\theta(\cdot)$ where $\mathbb{P}_\theta$ is a probability measure on $\mathcal{X}$ for each parameter value $\theta \in \Theta$.

We need to add more structure to model causality.

Definition (sketch)

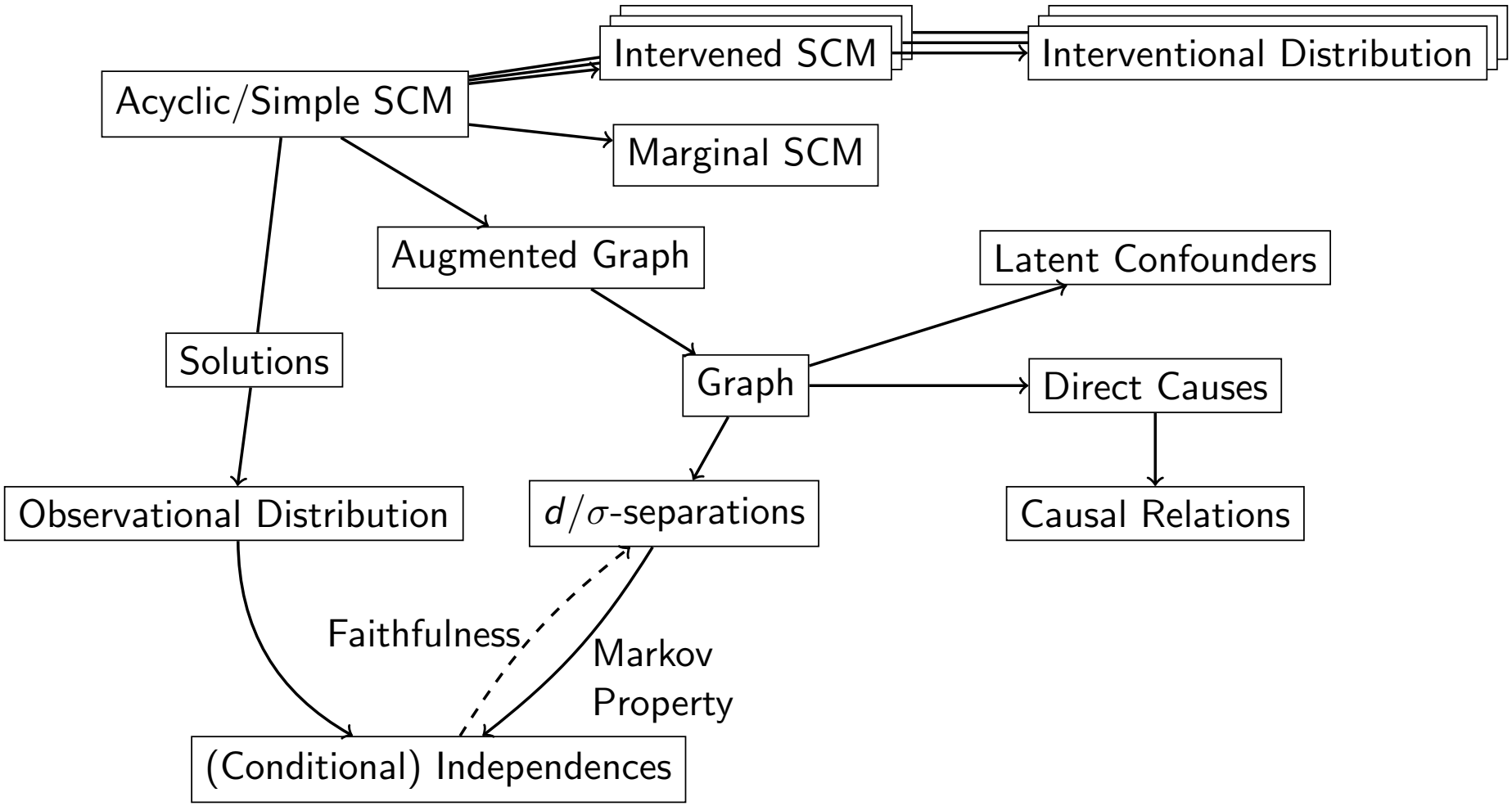
A statistical causal model is a family of statistical models, indexed by interventions:

$$(I, \theta) \mapsto \mathbb{P}_\theta\big(\,\cdot \mid \mathrm{do}(I)\big)$$

Here, $I \in \mathbb{I}$ represents an external intervention on a system. Modeling causality is done by imposing relationships on members of the family.

This enables us to model how probability distributions of system variables change under external interventions on the system.

Statistical Causal Modeling with Structural Causal Models



Structural Causal Models: Definition

Definition ([Wright, 1921, Strotz and Wold, 1960, Pearl, 2000])

A Structural Causal Model (SCM), also known as Structural Equation Model (SEM), is a tuple $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with:

1. a product of standard measurable spaces $\boldsymbol{\mathcal{X}} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$ (domains of the endogenous variables)
2. a product of standard measurable spaces $\boldsymbol{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ (domains of the latent exogenous variables)
3. a measurable mapping $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ (the causal mechanisms)
4. a product probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j \in \mathcal{J}} \mathbb{P}_{\mathcal{E}_j}$ on $\boldsymbol{\mathcal{E}}$ (the latent exogenous distribution)

Definition ([Bongers et al., 2016])

A pair of random variables $(\boldsymbol{X}, \boldsymbol{E}) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}}$ is a solution of SCM $\mathcal{M}$ if $\boldsymbol{E} \sim \mathbb{P}_{\boldsymbol{\mathcal{E}}}$ and the structural equations $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$ hold a.s..

Structural Causal Models: Example

Example

Structural Causal Model $\mathcal{M}$:

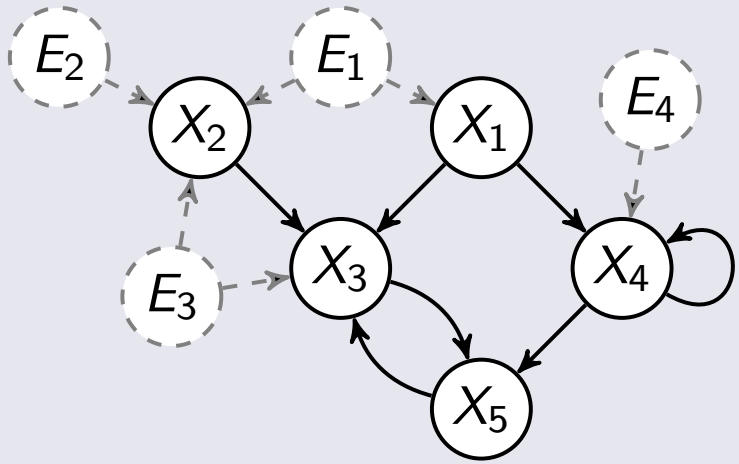
Formally:

$$\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle =$$
$$\langle \{1, \ldots, 4\}, \{1, \ldots, 5\},$$
$$\mathbb{R}^4, \mathbb{R}^5, (f_1, \ldots, f_5), \textstyle\prod_{j=1}^{4} \mathbb{P}_{\mathcal{E}_j} \rangle$$

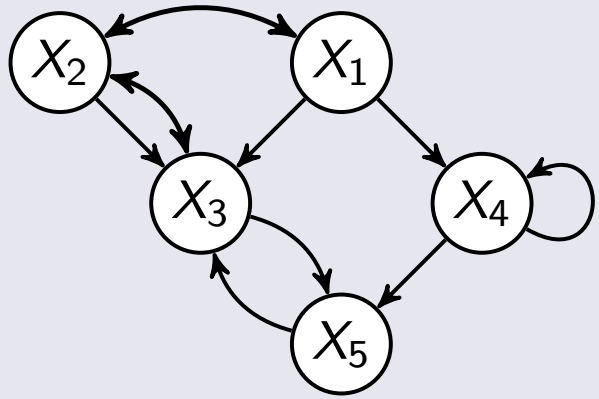
Informally:

$$\begin{aligned} E_1 &\sim \mathbb{P}_{\mathcal{E}_1} & X_1 &= f_1(E_1) \\ E_2 &\sim \mathbb{P}_{\mathcal{E}_2} & X_2 &= f_2(E_1, E_2, E_3) \\ E_3 &\sim \mathbb{P}_{\mathcal{E}_3} & X_3 &= f_3(X_1, X_2, X_5, E_3) \\ E_4 &\sim \mathbb{P}_{\mathcal{E}_4} & X_4 &= f_4(X_1, X_4, E_4) \\ & & X_5 &= f_5(X_3, X_4) \end{aligned}$$

Augmented graph $\mathcal{G}^a(\mathcal{M})$:



Graph $\mathcal{G}(\mathcal{M})$:



Interventions on a Structural Causal Model: Definition

To interpret an SCM as a statistical *causal* model, we also need to define its semantics under interventions.

Definition (Perfect Interventions [Pearl, 2000])

Let $\mathcal{M}$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables ("*intervention targets*") and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value ("*intervention values*"). The intervened SCM $\mathcal{M}_{\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)}$ is the same as $\mathcal{M}$, except with a modified causal mechanism $\tilde{\boldsymbol{f}}$ with components $\tilde{f}_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$:

$$\tilde{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \begin{cases} \xi_i & i \in I \\ f_i(\boldsymbol{x}, \boldsymbol{e}) & i \notin I. \end{cases}$$

Interpretation: The perfect intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$ enforces $\boldsymbol{X}_I$ to attain value $\boldsymbol{\xi}_I$ by completely overriding the default causal mechanisms that normally determine the values of the intervened variables, while leaving the other causal mechanisms invariant.

Interventions on a Structural Causal Model: Example

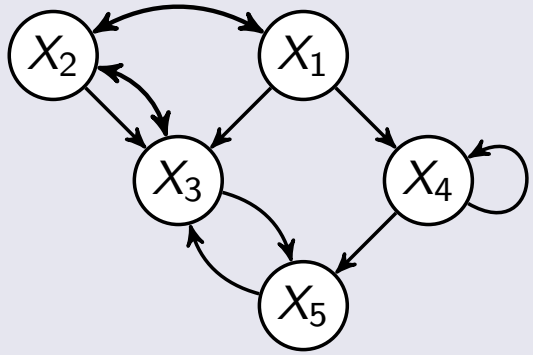
Example

Observational (no intervention):

SCM $\mathcal{M}$:

$$
\begin{aligned}
E_1 &\sim \mathbb{P}_{\mathcal{E}_1} & X_1 &= f_1(E_1) \\
E_2 &\sim \mathbb{P}_{\mathcal{E}_2} & X_2 &= f_2(E_1, E_2, E_3) \\
E_3 &\sim \mathbb{P}_{\mathcal{E}_3} & X_3 &= f_3(X_1, X_2, X_5, E_3) \\
E_4 &\sim \mathbb{P}_{\mathcal{E}_4} & X_4 &= f_4(X_1, X_4, E_4) \\
& & X_5 &= f_5(X_3, X_4)
\end{aligned}
$$

Graph $\mathcal{G}(\mathcal{M})$:

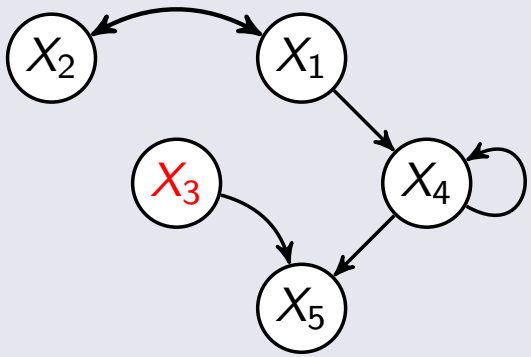


Intervention $\mathrm{do}(X_3 = \xi_3)$:

Intervened SCM $\mathcal{M}_{\mathrm{do}(X_3=\xi_3)}$:

$$
\begin{aligned}
E_1 &\sim \mathbb{P}_{\mathcal{E}_1} & X_1 &= f_1(E_1) \\
E_2 &\sim \mathbb{P}_{\mathcal{E}_2} & X_2 &= f_2(E_1, E_2, E_3) \\
E_3 &\sim \mathbb{P}_{\mathcal{E}_3} & X_3 &= \xi_3 \\
E_4 &\sim \mathbb{P}_{\mathcal{E}_4} & X_4 &= f_4(X_1, X_4, E_4) \\
& & X_5 &= f_5(X_3, X_4)
\end{aligned}
$$

Intervened graph $\mathcal{G}(\mathcal{M}_{\mathrm{do}(X_3=\xi_3)})$:

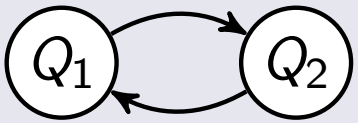


Causal cycles: Toy example

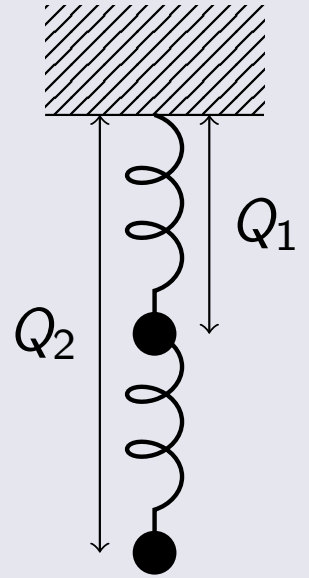
In many dynamical systems, feedback loops induce cyclic causality at equilibrium [Bongers and Mooij, 2018].

Example (Damped Coupled Harmonic Oscillators)

- Two masses, connected by a spring, suspended from the ceiling by another spring.
- Variables: vertical **equilibrium** positions Q_1 and Q_2.
- Q_1 causes Q_2.
- Q_2 causes Q_1.
- Causal graph:

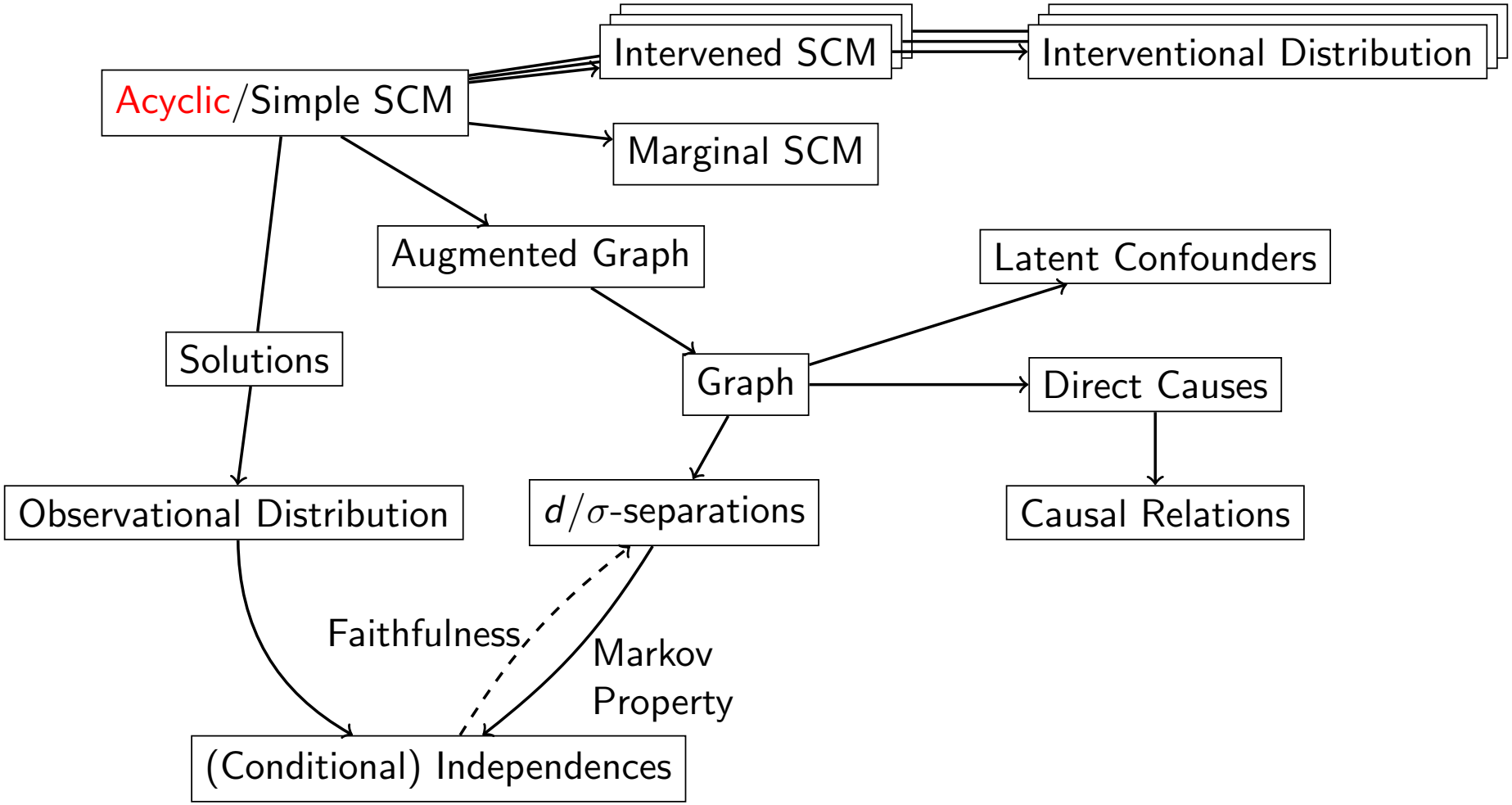


- Cannot be modeled with acyclic causal model!



In time-series modeling, fast dynamical interactions can also lead to "instantaneous" causal cycles.

Acyclic Structural Causal Models



An SCM $\mathcal{M}$ is called acyclic if its graph $\mathcal{G}(\mathcal{M})$ is acyclic.

Note: everything can be generalized to the class of *simple* SCMs (which can be cyclic).

Structural Causal Model as Statistical Causal Model

Remark

Let $\mathcal{M}$ be an acyclic SCM. Then $\mathcal{M}$ induces a statistical causal model:

- For each perfect intervention, we obtain a unique distribution

$$\mathbb{P}_{\mathcal{M}_{\text{do}(\boldsymbol{X}_I=\boldsymbol{\xi}_I)}}(\boldsymbol{X}) =: \mathbb{P}_{\mathcal{M}}(\boldsymbol{X} \mid \text{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I))$$

  for solutions $(\boldsymbol{X}_{\text{do}(\boldsymbol{X}_I=\boldsymbol{\xi}_I)}, \boldsymbol{E})$ of $\mathcal{M}_{\text{do}(\boldsymbol{X}_I=\boldsymbol{\xi}_I)}$.
- The "parameters" are the causal mechanism $\boldsymbol{f}$ and the latent exogenous distribution $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$.

Note that the SCM "ties together" all interventional distributions:

$$\begin{array}{ccc} \mathcal{M} & \xrightarrow{\text{intervention}} & \mathcal{M}_{\text{do}(\boldsymbol{X}_I=\boldsymbol{\xi}_I)} \\ \Big\downarrow{\scriptstyle \text{induced distribution}} & & \Big\downarrow{\scriptstyle \text{induced distribution}} \\ \mathbb{P}_{\mathcal{M}} & & \mathbb{P}_{\mathcal{M}_{\text{do}(\boldsymbol{X}_I=\boldsymbol{\xi}_I)}} \end{array}$$

Causal Interpretation of the Graph of an SCM

The graph $\mathcal{G}(\mathcal{M})$ of an acyclic SCM $\mathcal{M}$ can be interpreted causally:

Pattern in Graph	Causal interpretation
$i \to j$	i is a direct cause of j
$i_1 \to i_2 \to \cdots \to i_n$	i_1 is a cause of i_n
$i \leftrightarrow j$	i and j are confounded

We then refer to it as the causal graph.

Definition (d-separation [Pearl, 2000])

In an ADMG $\mathcal{G}$, a walk (i.e., a finite sequence of adjacent edges)

$$i_1 \; \begin{matrix} \leftarrow \\ \rightarrow \\ \leftrightarrow \end{matrix} \; \cdots \; \begin{matrix} \leftarrow \\ \rightarrow \\ \leftrightarrow \end{matrix} \; i_n$$

is called blocked by a set of nodes Z iff

- one or both end nodes i_1, i_n are in Z, or
- it contains a collider $i_{k-1} \begin{matrix} \rightarrow \\ \leftrightarrow \end{matrix} i_k \begin{matrix} \leftarrow \\ \leftrightarrow \end{matrix} i_{k+1}$ with $i_k \notin Z$, or
- it contains a non-collider with $i_k \in Z$,

For three sets of nodes A, B, Z, we say that A is d-separated from B by Z in $\mathcal{G}$, denoted $A \perp_{\mathcal{G}} B \mid Z$, if every walk that starts in A and ends in B is blocked by Z.

Cornerstone of causal reasoning: Markov Property

The Markov property allows one to read off conditional independences between endogenous variables directly from the causal graph.

Theorem (Markov Property (d-separation criterion))

Let $\mathcal{M}$ be an acyclic SCM. Then, for any subsets $A, B, Z \subseteq \mathcal{I}$ of endogenous variables:

$$A \underset{\mathcal{G}(\mathcal{M})}{\perp} B \,|\, Z \implies \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{M}}}{\perp\!\!\!\perp} \boldsymbol{X}_B \,|\, \boldsymbol{X}_Z$$

i.e., if A is d-separated from B by Z in $\mathcal{G}(\mathcal{M})$ then A is conditionally independent of B given Z in $\mathbb{P}_{\mathcal{M}}$.

Important consequences of the Markov Property are:

- Pearl's do-calculus [Pearl, 2000]
- Adjustment criteria for estimating causal effects [Pearl, 2000]
- Identification algorithm for identifying causal effects [Tian, 2002]

Faithfulness and Occam's razor

Definition

An acyclic SCM $\mathcal{M}$ is called faithful if the converse of the Markov Property also holds, i.e., if

$$A \underset{\mathcal{G}(\mathcal{M})}{\perp} B \,|\, Z \iff A \underset{\mathbb{P}_{\mathcal{M}}}{\perp\!\!\!\perp} B \,|\, Z$$

for all subsets $A, B, Z \subseteq \mathcal{I}$.

For certain classes of SCMs, faithfulness has been shown to hold *generically* [Meek, 1995].

The faithfulness assumption, often used in causal discovery, provides a particular form of Occam's razor.

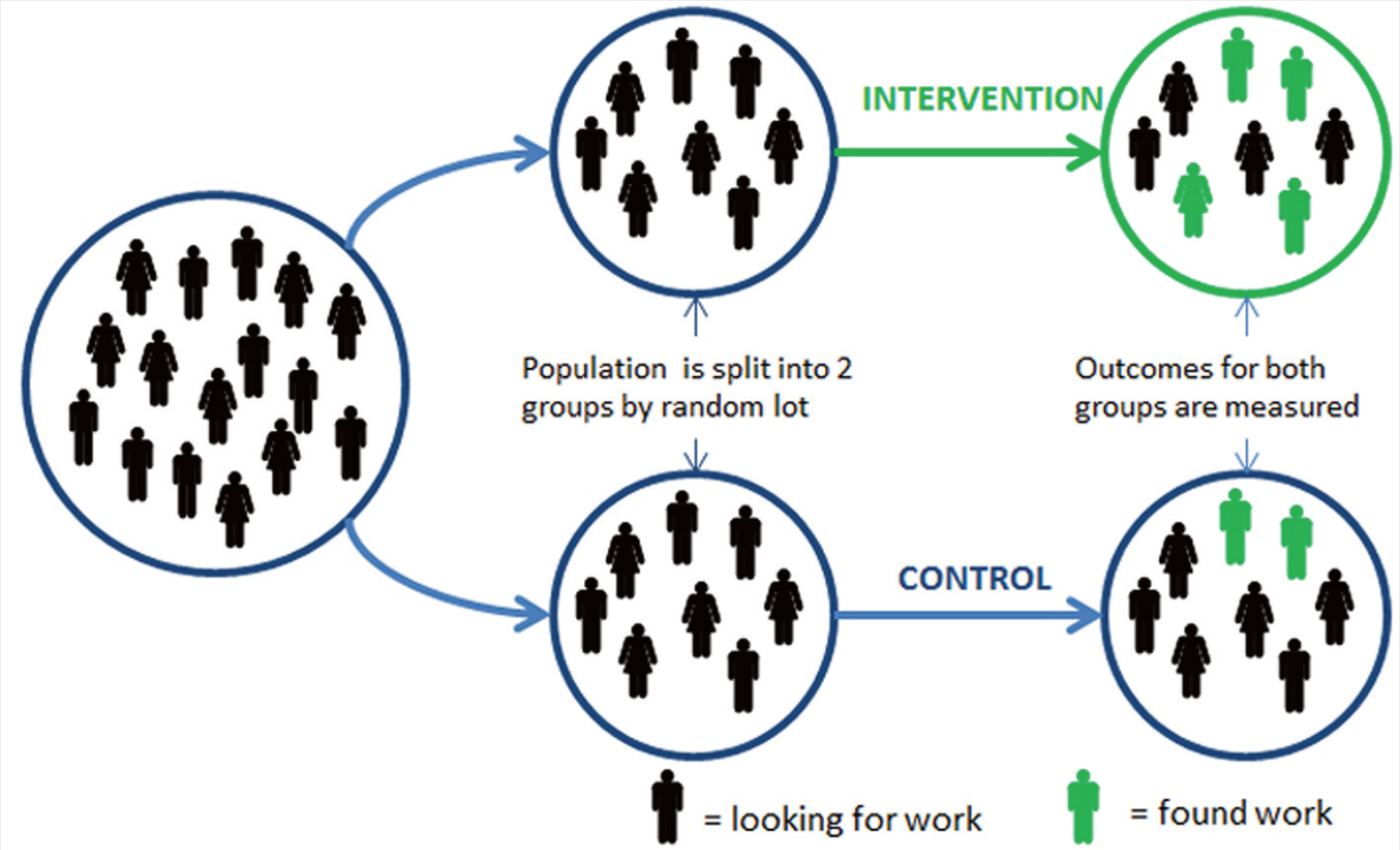
Part II

Causal Discovery: Estimating the Causal Graph from Data

1. **Randomized Controlled Trials**
2. Purely Observational Data
3. Exploiting Background Knowledge

Causal discovery by experimentation

Experimentation (e.g., Randomized Controlled Trials, A/B-testing, . . .) provides the gold standard for causal discovery (Fisher, 1935).

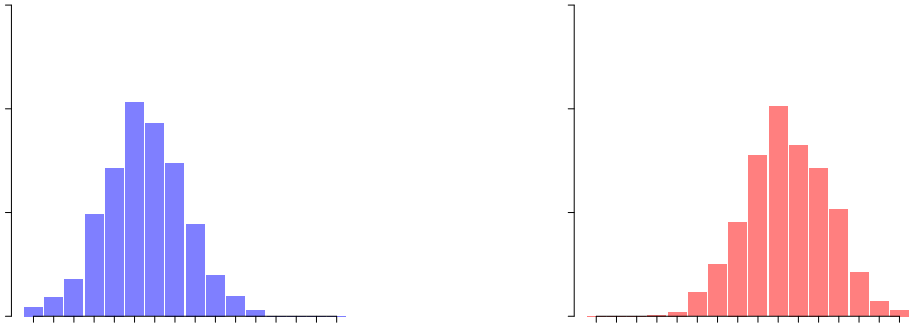


Two equivalent points of view

(a) Separate data sets

Control ($C = 0$):
X
-0.2
0.6
-1.7
...

Intervention ($C = 1$):
X
-0.3
1.8
-0.1
...

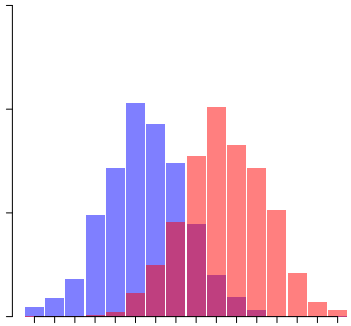


Two-sample test:

Is $\mathbb{P}(X \mid C = 0) = \mathbb{P}(X \mid C = 1)$?

(b) Pooled data

C	X
0	-0.2
0	0.6
0	-1.7
0	...
1	-0.3
1	1.8
1	-0.1
1	...



Independence test:

Is $X \perp\!\!\!\perp C$?

Proposition ("In RCTs, correlation implies causation")

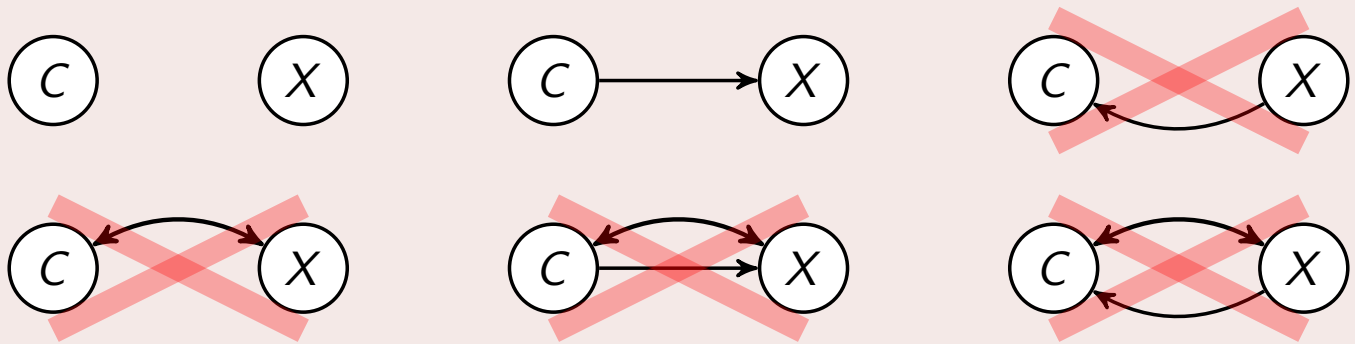
For an acyclic SCM $\mathcal{M}$ with two variables $\{C, X\}$, the RCT assumptions

- $C \leftarrow X \notin \mathcal{G}(\mathcal{M})$ ("*outcome does not cause treatment*")
- $C \leftrightarrow X \notin \mathcal{G}(\mathcal{M})$ ("*outcome and treatment are unconfounded*")

imply that if $C \not\perp\!\!\!\perp_{\mathbb{P}(\mathcal{M})} X$, then $C \to X \in \mathcal{G}(\mathcal{M})$. The causal effect of C on X is then:

$$\mathbb{P}_{\mathcal{M}}\big(X \mid \mathrm{do}(C = c)\big) = \mathbb{P}_{\mathcal{M}}(X \mid C = c).$$

Proof



Part III

Causal Discovery: Estimating the Causal Graph from Data

1. Randomized Controlled Trials
2. **Purely Observational Data**
3. Exploiting Background Knowledge

Causal discovery from purely observational data

Intriguing alternative: causal discovery from purely observational data (Spirtes & Gleimour & Scheines (2000), Pearl (2000), . . .)

No more experiments necessary. . . ! Two approaches:

Constraint-based

Look for certain patterns of (conditional) independences in data, which constrain the possible causal graphs.

Likelihood-based

Score likelihoods of different causal graphs and select the best one(s).

Disclaimer: Works only under strong assumptions and with (possibly very) large sample sizes.

Constraint-based Causal Discovery

From the pattern of conditional independences in the data we can reconstruct a set of possible underlying causal graphs, even when allowing for latent confounders (Spirtes, Gleimour, Scheines; 2000).

Data

X_1	X_2	X_3	X_4
2	0.1	0.2	0.5
2	0.13	0.21	0.49
2	0.23	0.21	0.51
5	0.5	0.19	0.52
5	0.6	0.18	0.51
2	0.2	0.22	0.92
2	0.23	0.21	0.99
5	0.53	1.2	0.95
5	0.55	1.19	0.97

CIs

$X_2 \not\perp\!\!\!\perp X_4$

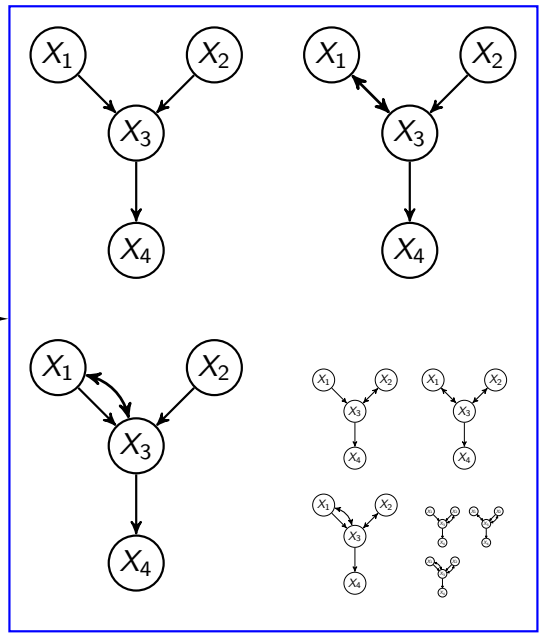
$X_2 \perp\!\!\!\perp X_4 \mid X_3$

$X_1 \perp\!\!\!\perp X_2$

$X_1 \not\perp\!\!\!\perp X_2 \mid X_3$

$\ldots$

Possible Causal Graphs



Hardness of Causal Discovery

ASD [Hyttinen et al., 2014] solves the inverse problem (from conditional independences to causal graph) by a general-purpose optimizer, taking into account the strength of the dependences.

d	Number of DAGs with d nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	33277190122710759173617757331126112588358307625842190258354677350 5

Table B.1: The number of DAGs depending on the number d of nodes, taken from `http://oeis.org/A003024` [OEIS Foundation Inc., 2017]. The length of the numbers grows faster than any linear term.

Source: Peters, Janzing & Schölkopf (2017)

The combinatorial explosion is even worse when allowing for confounders and cycles!

Constraint-based Causal Discovery Algorithm

The (Augmented) FCI algorithm [Spirtes *et al.*, 2000, Spirtes *et al.*, 1999, Ali *et al.*, 2005, Zhang, 2008] is one of the "classical" algorithms:

$\mathcal{R}0a$ If $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$, then $X \not\perp\!\!\!\perp Y$, $Sep(X,Y) \leftarrow \mathbf{Z}$.

$\mathcal{R}0b$ If $X *\!\!-\!\!* Z \circ\!\!-\!\!* Y$ and $X \not\perp\!\!\!\perp Y$, then if $Z \notin Sep(X,Y)$, then $X *\!\!\rightarrow Z \leftarrow\!\!* Y$.

$\mathcal{R}1$ If $X *\!\!\rightarrow Z \circ\!\!-\!\!* Y$, and $X \not\perp\!\!\!\perp Y$, then $Z \longrightarrow Y$.

$\mathcal{R}2a$ If $Z \longrightarrow X *\!\!\rightarrow Y$ and $Z *\!\!-\!\!\circ Y$, then $Z *\!\!\rightarrow Y$.

$\mathcal{R}2b$ If $Z *\!\!\rightarrow X \longrightarrow Y$ and $Z *\!\!-\!\!\circ Y$, then $Z *\!\!\rightarrow Y$.

$\mathcal{R}3$ If $X *\!\!\rightarrow Z \leftarrow\!\!* Y$, $X *\!\!-\!\!\circ W \circ\!\!-\!\!* Y$, $X \not\perp\!\!\!\perp Y$, and $W *\!\!-\!\!\circ Z$, then $W *\!\!\rightarrow Z$.

$\mathcal{R}4a$ If $u = \langle X, .., Z_k, Z, Y \rangle$ is a discriminating path between X and Y for Z, and $Z \circ\!\!-\!\!* Y$, then if $Z \in Sep(X,Y)$, then $Z \longrightarrow Y$.

$\mathcal{R}4b$ Idem, if $Z \notin Sep(X,Y)$ then $Z_k \longleftrightarrow Z \longleftrightarrow Y$.

$\mathcal{R}5$ If $u = \langle Z, X, .., W, Y, Z, X \rangle$ is an uncov. circle path, then $Z - Y$ (idem for all edges on u).

$\mathcal{R}6$ If $X - Z \circ\!\!-\!\!* Y$, then orient as $Z -\!\!* Y$.

$\mathcal{R}7$ If $X -\!\!\circ Z \circ\!\!-\!\!* Y$, and $X \not\perp\!\!\!\perp Y$, then $Z -\!\!* Y$.

$\mathcal{R}8a$ If $Z \longrightarrow X \longrightarrow Y$ and $Z \circ\!\!\rightarrow Y$, then $Z \longrightarrow Y$.

$\mathcal{R}8b$ If $Z -\!\!\circ X \longrightarrow Y$ and $Z \circ\!\!\rightarrow Y$, then $Z \longrightarrow Y$.

$\mathcal{R}9$ If $Z \circ\!\!\rightarrow Y$, $u = \langle Z, X, W, .., Y \rangle$ is an uncov. p.d. path, and $X \not\perp\!\!\!\perp Y$, then $Z \longrightarrow Y$.

$\mathcal{R}10$ If $Z \circ\!\!\rightarrow Y$, $X \longrightarrow Y \longleftarrow W$, $u_1 = \langle Z, S, .., X \rangle$ and $u_2 = \langle Z, V, .., W \rangle$ are uncov. p.d. paths, (possibly with $S = X$ and/or $V = W$), then if $S \not\perp\!\!\!\perp V$, then $Z \longrightarrow Y$.

Input : independence oracle for $\mathbf{V}$
Output : complete PAG $\mathcal{P}$ over $\mathbf{V}$
1: $\mathcal{P} \leftarrow$ fully $\circ\!\!-\!\!\circ$ connected graph over $\mathbf{V}$
2: **for all** $\{X, Y\} \in \mathbf{V}$ **do**
3: search *in some clever way* for a $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$
4: $\mathcal{P} \leftarrow \mathcal{R}0a$ (eliminate $X \not\perp\!\!\!\perp Y$)
5: record $Sep(X,Y) \leftarrow \mathbf{Z}$
6: **end for**
7: $\mathcal{P} \leftarrow \mathcal{R}0b$ (unshielded colliders)
8: **repeat** $\mathcal{P} \leftarrow \mathcal{R}1 - \mathcal{R}4b$ **until** finished
9: $\mathcal{P} \leftarrow \mathcal{R}5$ (uncovered circle paths)
10: **repeat** $\mathcal{P} \leftarrow \mathcal{R}6 - \mathcal{R}7$ **until** finished
11: **repeat** $\mathcal{P} \leftarrow \mathcal{R}8a - \mathcal{R}10$ **until** finished

Algorithm 1: Augmented FCI algorithm

Source: [Claassen & Heskes, 2011]

Part IV

Causal Discovery: Estimating the Causal Graph from Data

1. Randomized Controlled Trials
2. Purely Observational Data
3. **Exploiting Background Knowledge**

Local Causal Discovery (LCD)

Theorem ([Cooper, 1997])

Let $\mathcal{M}$ be an acyclic, faithful SCM. If for three endogenous variables $i, j, k \in \mathcal{I}$:

- *X_j and X_k do not cause X_i,*
- *$X_i \not\perp\!\!\!\perp X_j$ and $X_j \not\perp\!\!\!\perp X_k$,*
- *$X_i \perp\!\!\!\perp X_k \,|\, X_j$,*

then X_j causes X_k and $\mathbb{P}\big(X_k \,|\, \mathrm{do}(X_j = x_j)\big) = \mathbb{P}(X_k \,|\, X_j = x_j)$.

Local Causal Discovery (LCD)

Theorem ([Cooper, 1997])

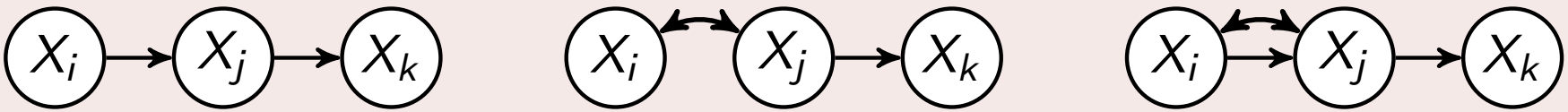
Let $\mathcal{M}$ be an acyclic, faithful SCM. If for three endogenous variables $i, j, k \in \mathcal{I}$:

- *X_j and X_k do not cause X_i,*
- *$X_i \not\perp\!\!\!\perp X_j$ and $X_j \not\perp\!\!\!\perp X_k$,*
- *$X_i \perp\!\!\!\perp X_k \,|\, X_j$,*

then X_j causes X_k and $\mathbb{P}\big(X_k \,|\, \mathrm{do}(X_j = x_j)\big) = \mathbb{P}(X_k \,|\, X_j = x_j)$.

Proof.

The only possible causal graphs of the marginalized $\mathcal{M}_{\{i,j,k\}}$ are:



Now apply the do-calculus. □

Part V

A Unifying Framework: Joint Causal Inference

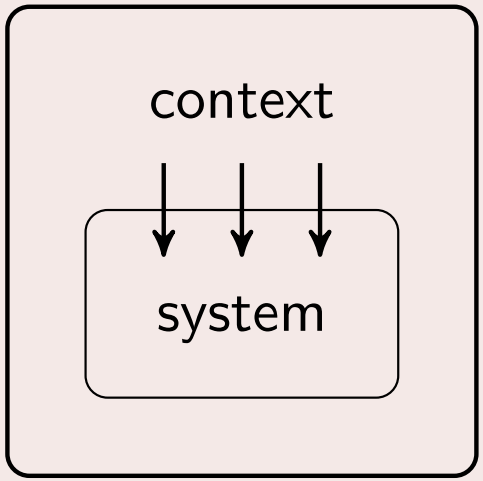
Joint Causal Inference: the main idea

JCI generalizes the idea of RCTs to multiple context and system variables. Distinguish:

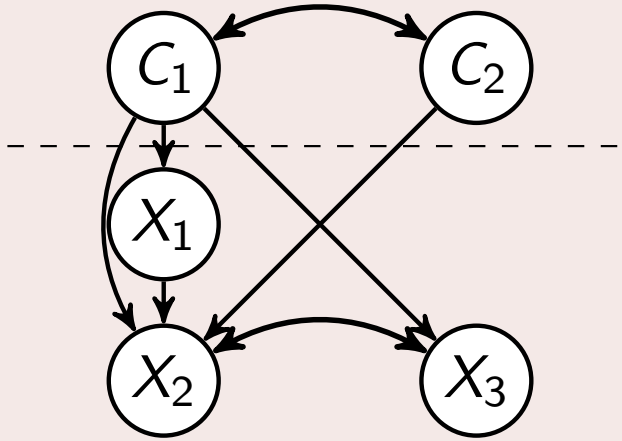
- **System** variables $\{X_i\}_{i \in \mathcal{I}}$ that model the system of interest.
- **Context** variables $\{C_k\}_{k \in \mathcal{K}}$ that model the context of the system,

Main idea

JCI reduces modeling a system *in* its environment to modeling the meta-system consisting of the system *and* its environment:



meta-system



JCI Assumptions

The boundary between system and context is chosen by the modeler. We assume that the context variables are observed and exogenous while the system variables are endogenous.

JCI Assumptions

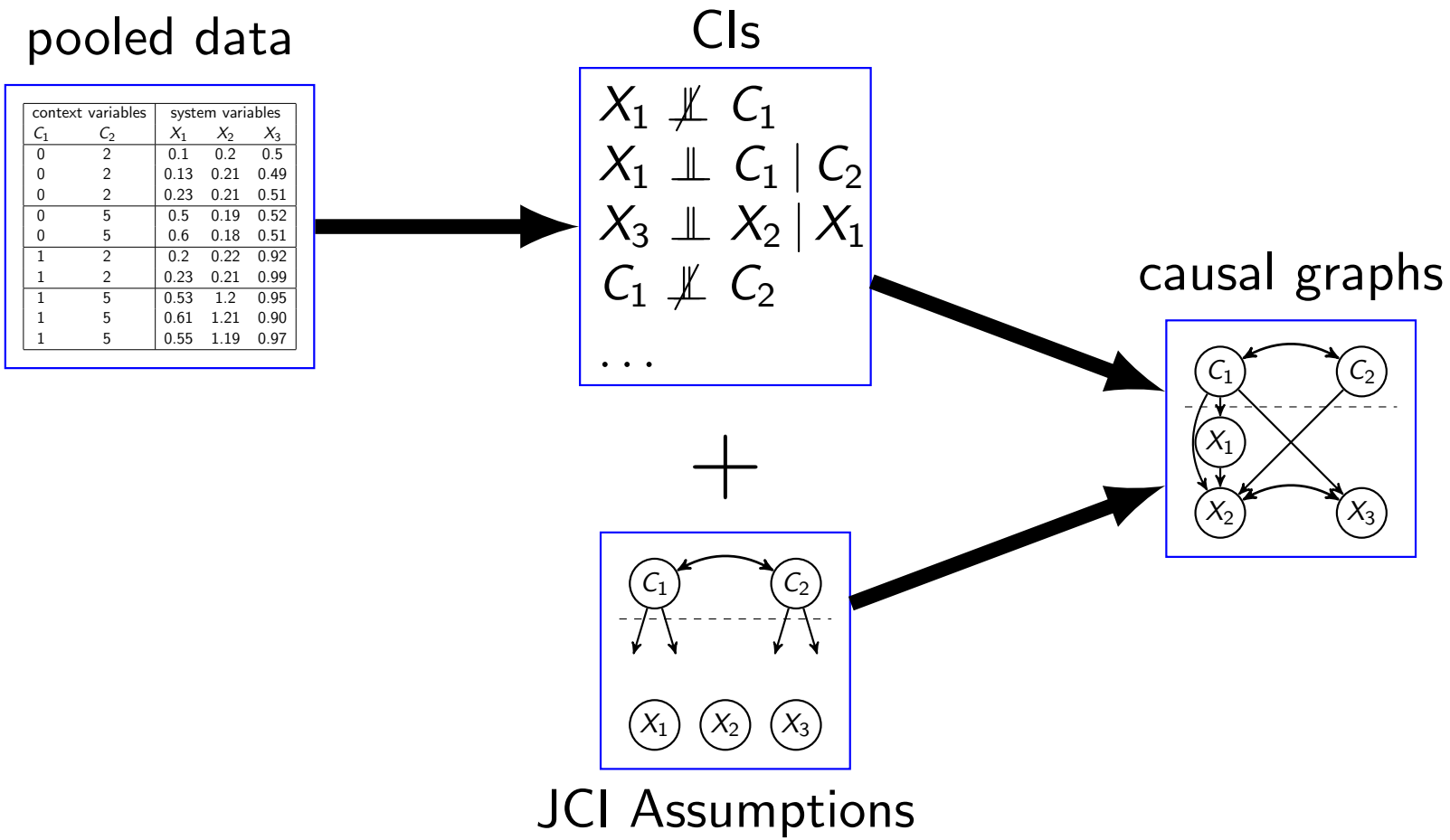
The causal graph $\mathcal{G}(\mathcal{M})$ that includes both system variables $\{X_i\}_{i \in \mathcal{I}}$ and context variables $\{C_k\}_{k \in \mathcal{K}}$, which jointly models the system and its environment, satisfies:

1. $C_k \leftarrow X_i \notin \mathcal{G}(\mathcal{M})$ for all $k \in \mathcal{K}, i \in \mathcal{I}$ ("*the system does not affect its context*"), and
2. $C_k \leftrightarrow X_i \notin \mathcal{G}(\mathcal{M})$ for all $k \in \mathcal{K}, i \in \mathcal{I}$ ("*context and system are unconfounded*").

The second assumption is optional because it can be violated if the context is not complete and no randomization has been performed.

Joint Causal Inference

Question: How to estimate the causal graph from the data?



context variables		system variables		
C_1	C_2	X_1	X_2	X_3
0	2	0.1	0.2	0.5
0	2	0.13	0.21	0.49
0	2	0.23	0.21	0.51
0	5	0.5	0.19	0.52
0	5	0.6	0.18	0.51
1	2	0.2	0.22	0.92
1	2	0.23	0.21	0.99
1	5	0.53	1.2	0.95
1	5	0.61	1.21	0.90
1	5	0.55	1.19	0.97

Answer: Simply apply a standard constraint-based causal discovery method (designed for purely observational data) on the *pooled* data, and incorporate the JCI assumptions.

Implementing JCI

The following algorithms can be seen as causal discovery algorithms implementing special cases of the JCI framework:

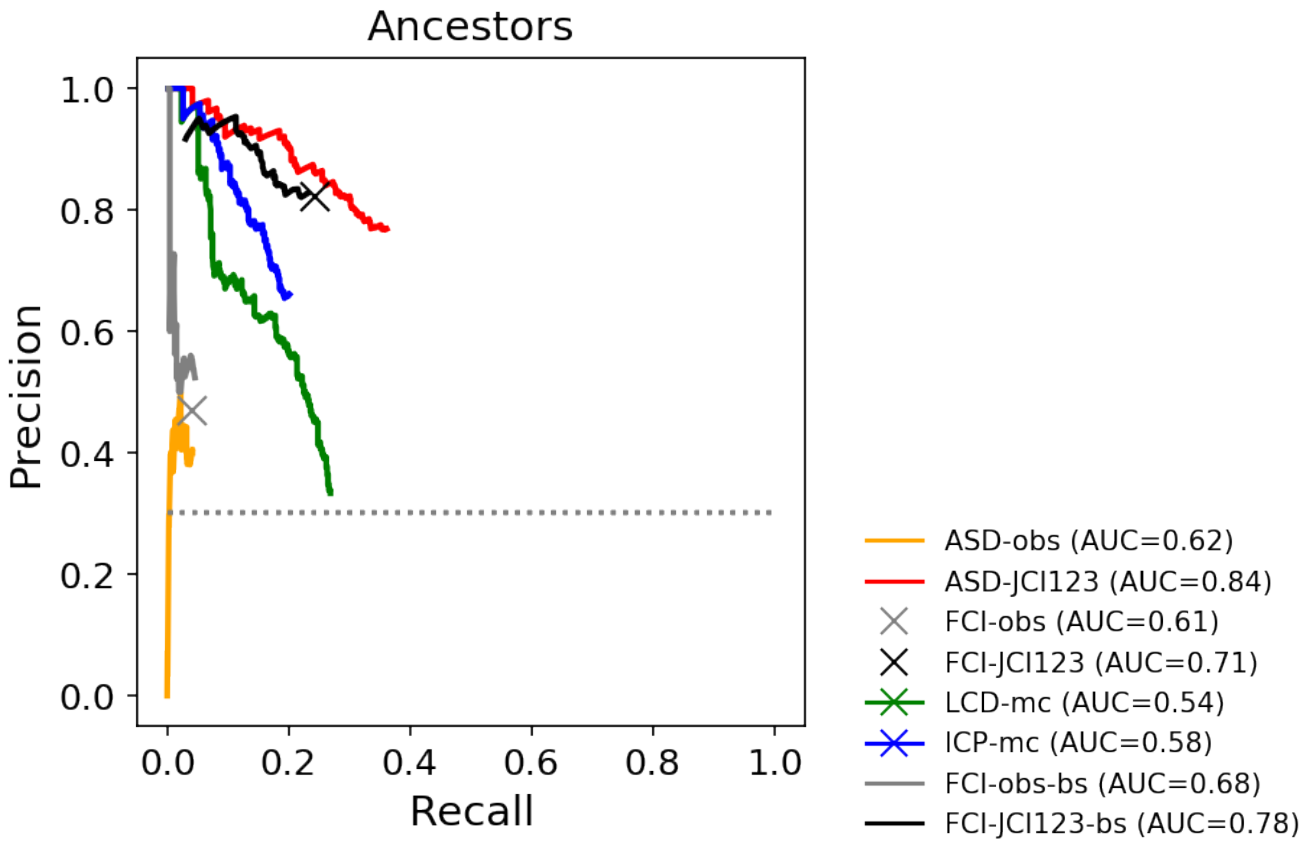
Method	$\#\{\mathcal{K}\}$	$\#\{\mathcal{I}\}$
FCI	0	≥ 2
ASD	0	≥ 2
RCT	1	1
LCD	1	2
ICP	1	≥ 2

Novel methods [Mooij et al., 2020]:

Method	$\#\{\mathcal{K}\}$	$\#\{\mathcal{I}\}$
ASD-JCI	≥ 0	≥ 2
FCI-JCI	≥ 0	≥ 2

Evaluation on simulated data (I)

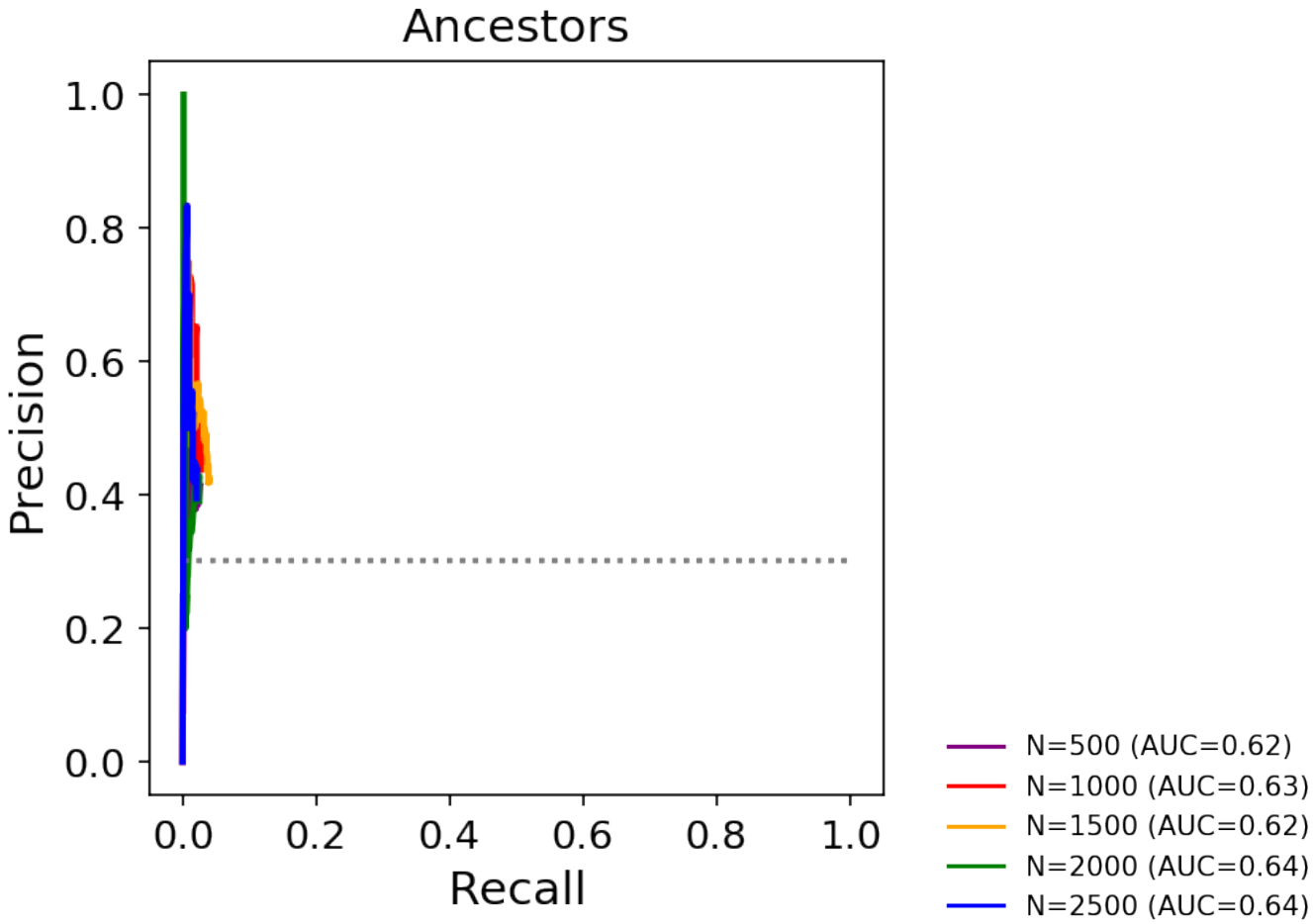
Random SCMs with 4 system variables, 2 context variables, random parameters, linear-Gaussian distribution, imperfect unknown interventions. Task: discover causal relations between system variables.



JCI outperforms purely observational discovery substantially.

Evaluation on simulated data (II)

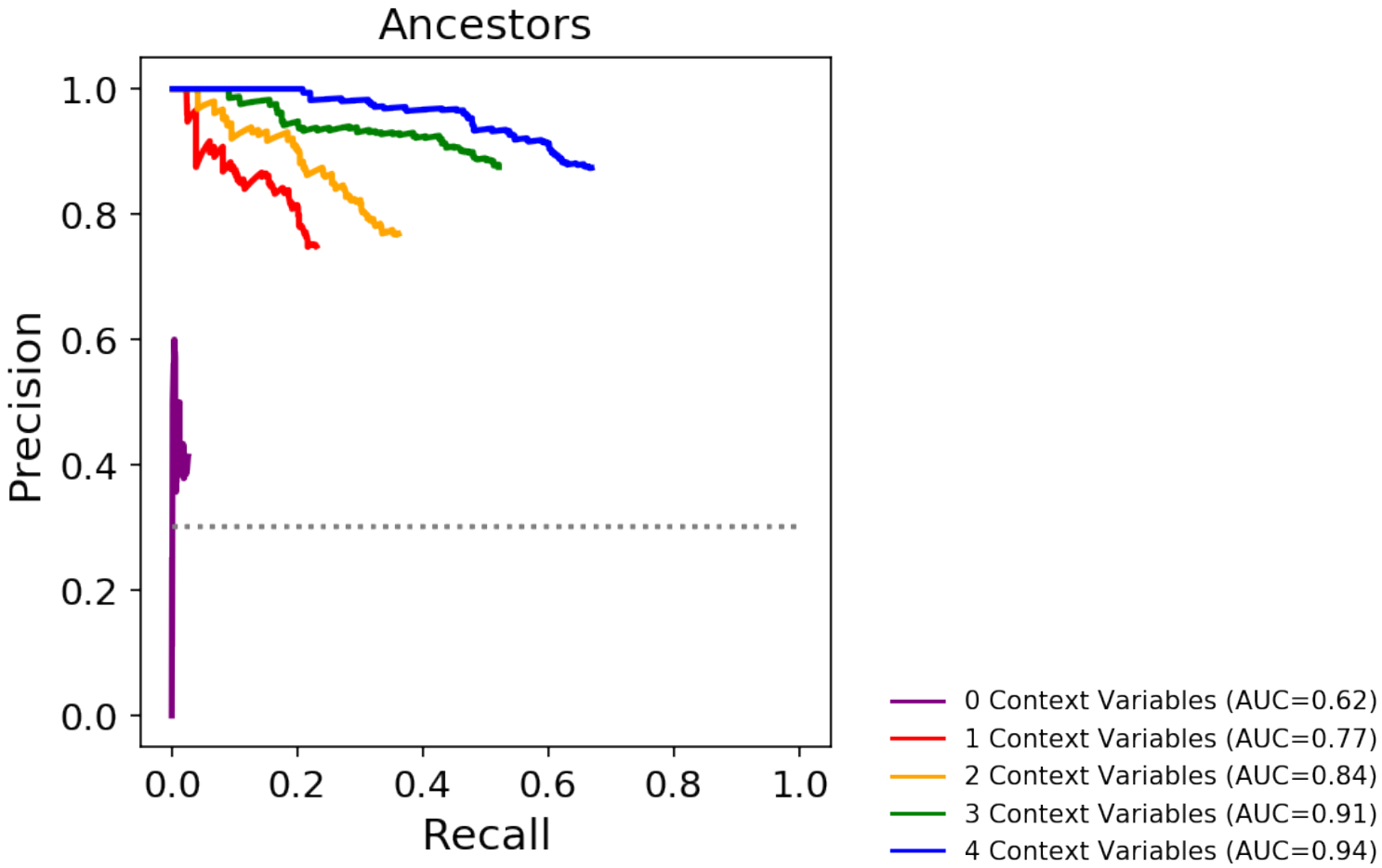
Increasing the number of observational samples does not help much (shown here: performance of ASD):



(Also, note that RCTs are not applicable.)

Evaluation on simulated data (III)

ASD-JCI with more context variables ($N_c = 500$ samples for each context) helps considerably:



Perturbing the system is extremely helpful for understanding it.

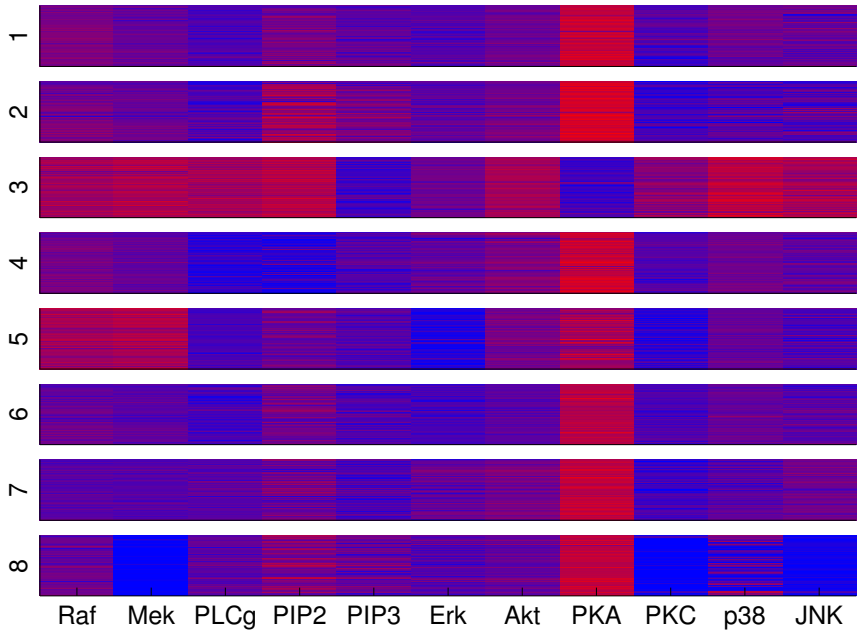
Part VI

Validation on real-world data

Understanding Protein Signaling

Protein Abundance Data:

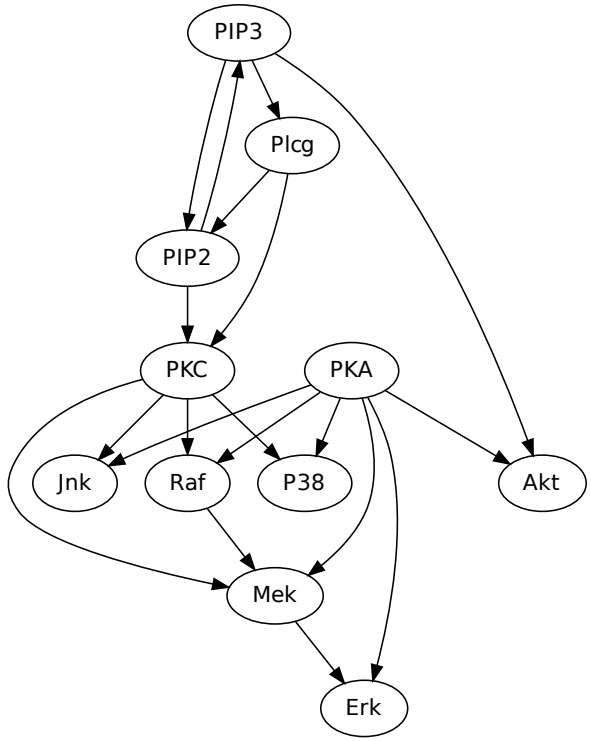
(Sachs *et al*, 2005)



Condition	Reagent	Intervention
1	-	observational
2	Akt-inhibitor	inhibits AKT activity
3	G0076	inhibits PKC activity
4	Psitectorigenin	inhibits PIP2 abundance
5	U0126	inhibits MEK activity
6	LY294002	inhibits PIP2/PIP3 activity
7	PMA	activates PKC + global
8	β2CAMP	activates PKA + global

Causal Graph:

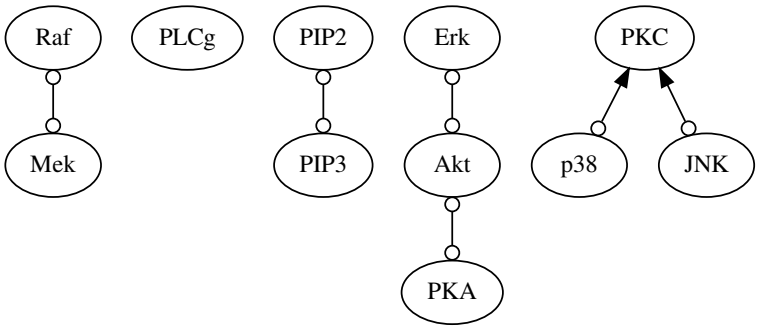
("Signalling network")



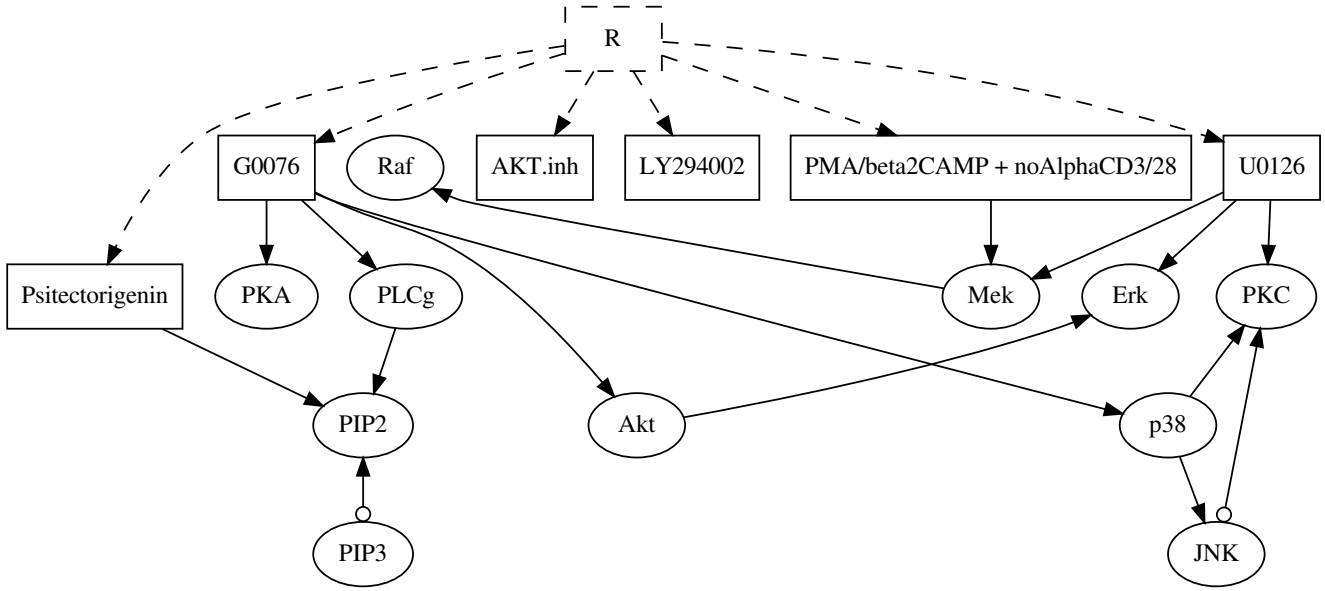
(depicted here: "consensus" network)

Application of FCI and FCI-JCI

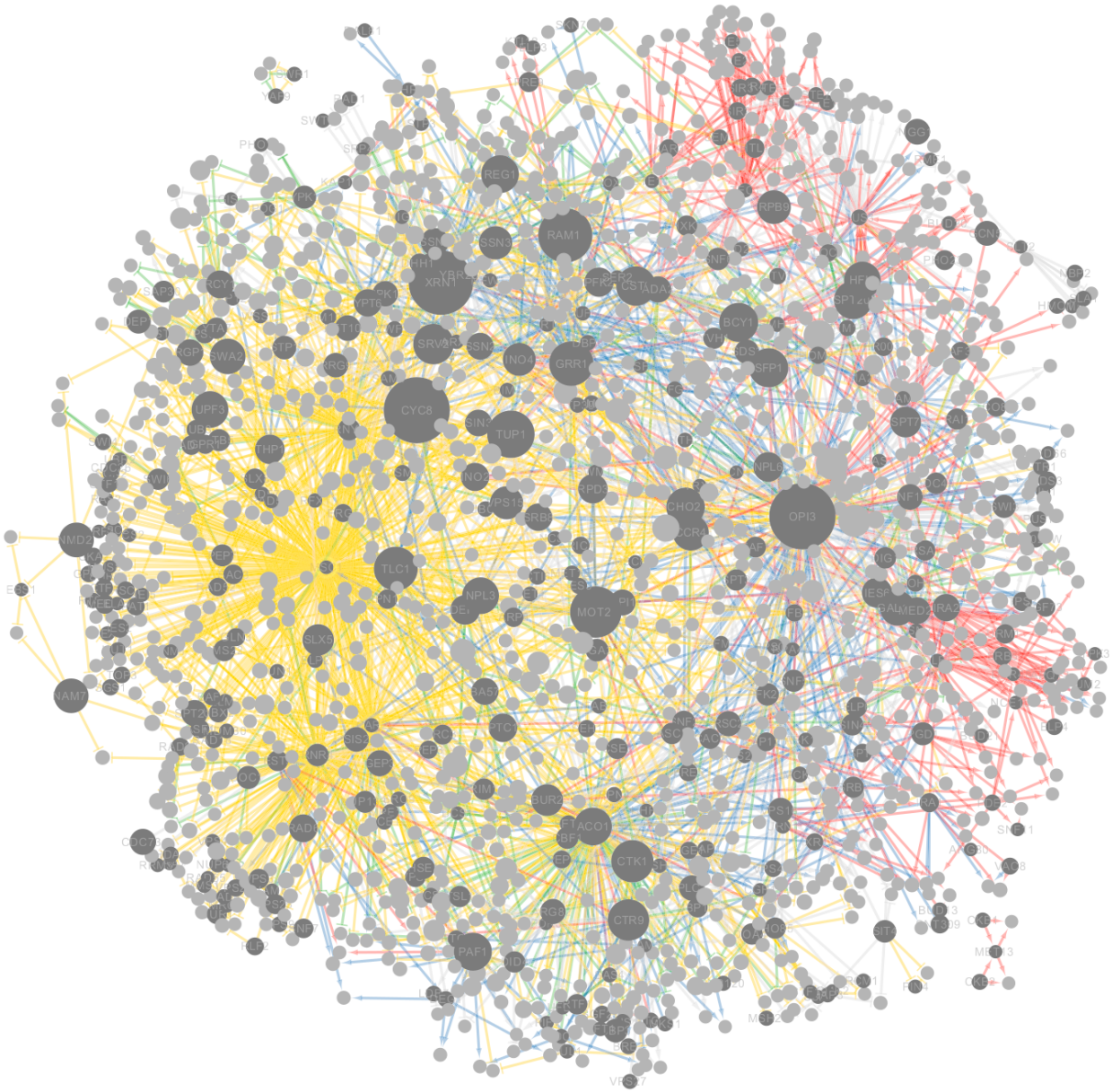
Only observational data (FCI):



All data (FCI-JCI):



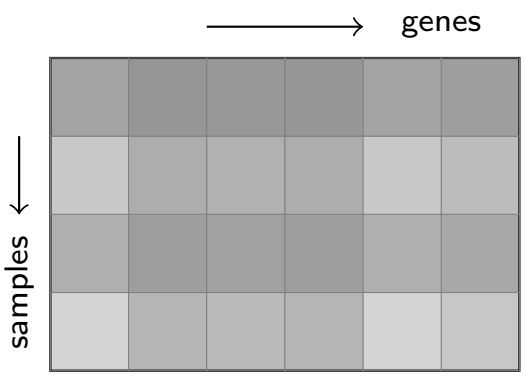
Gene Regulatory Network = Causal Graph



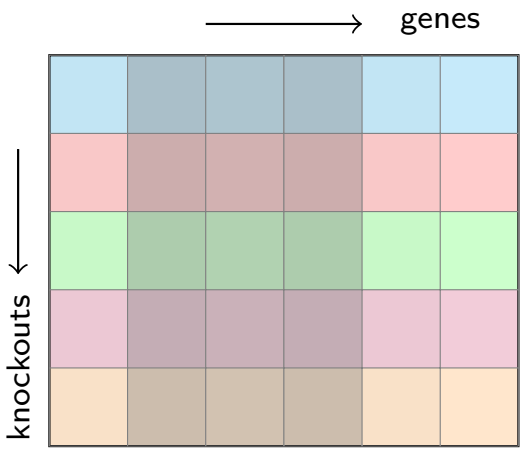
Source: [Kemmeren et al., 2014]

Causal Discovery of Gene Regulatory Networks

observational:
(wild-type vs. wild-type):



interventional:
(mutant vs. wild-type):

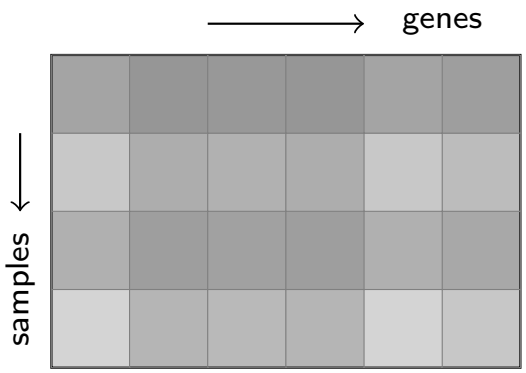


Large-scale Single Gene Knockout Micro-Array Data [Kemmeren et al., 2014]:

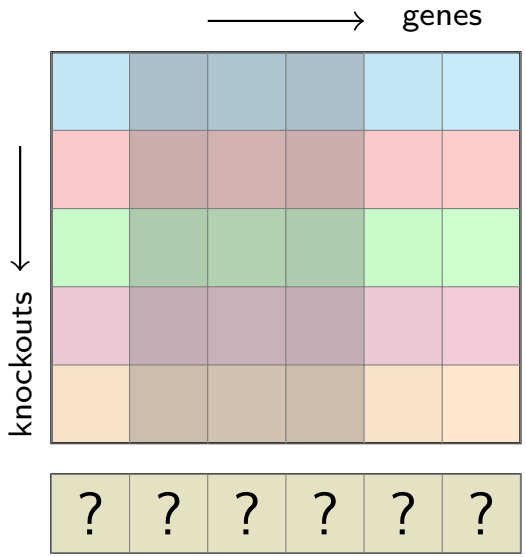
- ∼6,500 variables (gene expression)
- ∼260 observational samples (wild-type vs. wild-type)
- ∼1,500 interventional samples (single-gene knockouts/knockdowns)

Causal Discovery of Gene Regulatory Networks

observational:
(wild-type vs. wild-type):



interventional:
(mutant vs. wild-type):



Large-scale Single Gene Knockout Micro-Array Data [Kemmeren et al., 2014]:

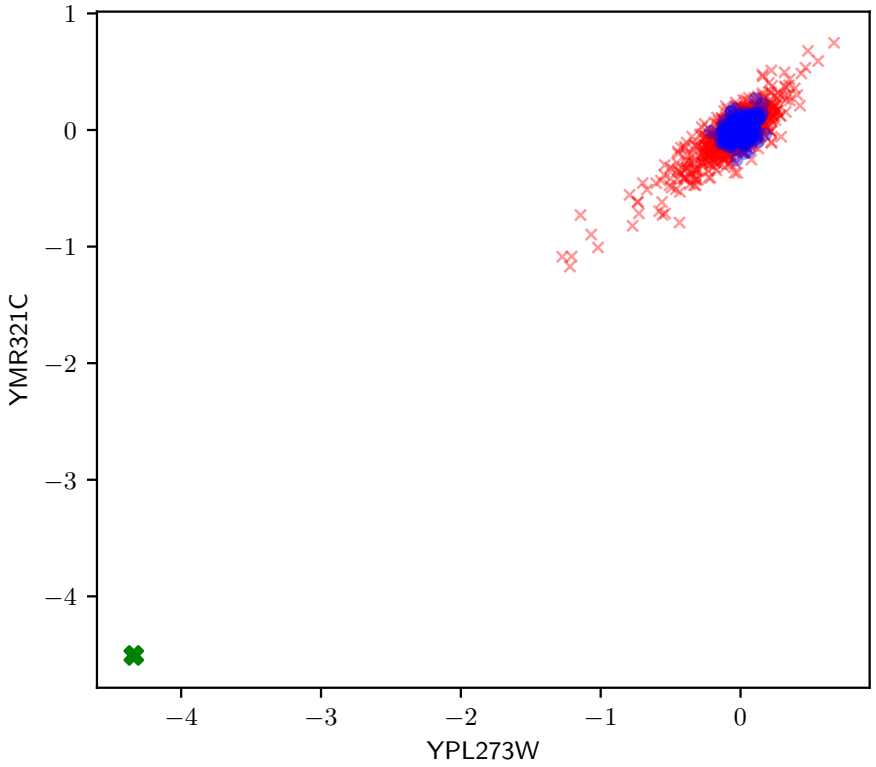
- ∼6,500 variables (gene expression)
- ∼260 observational samples (wild-type vs. wild-type)
- ∼1,500 interventional samples (single-gene knockouts/knockdowns)

Challenge

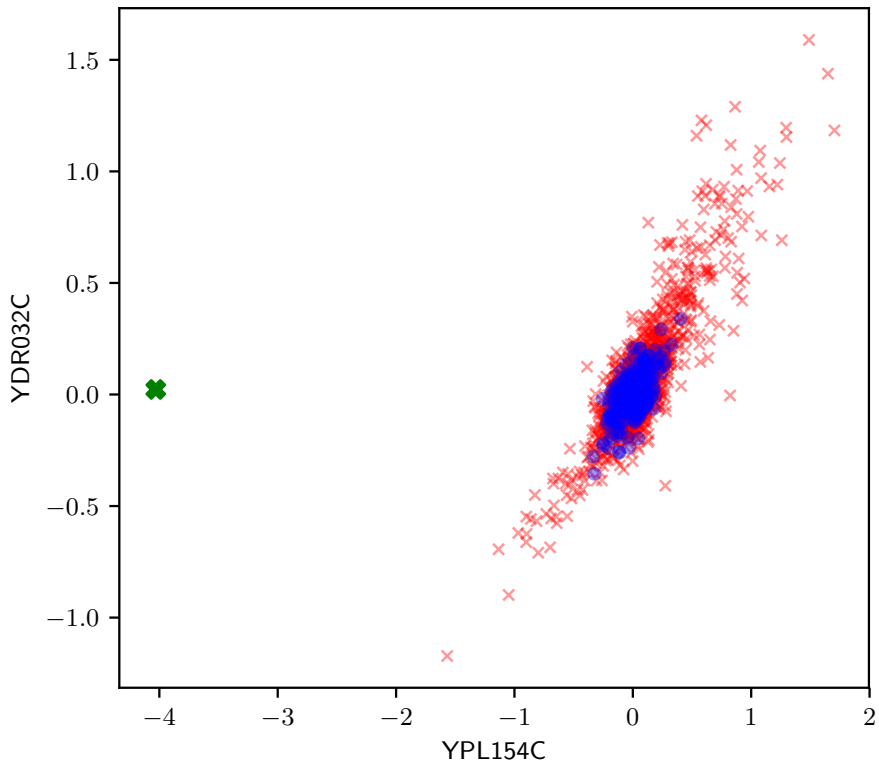
Can we, in a purely data-driven way (without using biological knowledge), predict which genes strongly change their expression when we knock-out a given gene (without using any data corresponding to that particular knock-out experiment)?

Causation or correlation?

True positive:



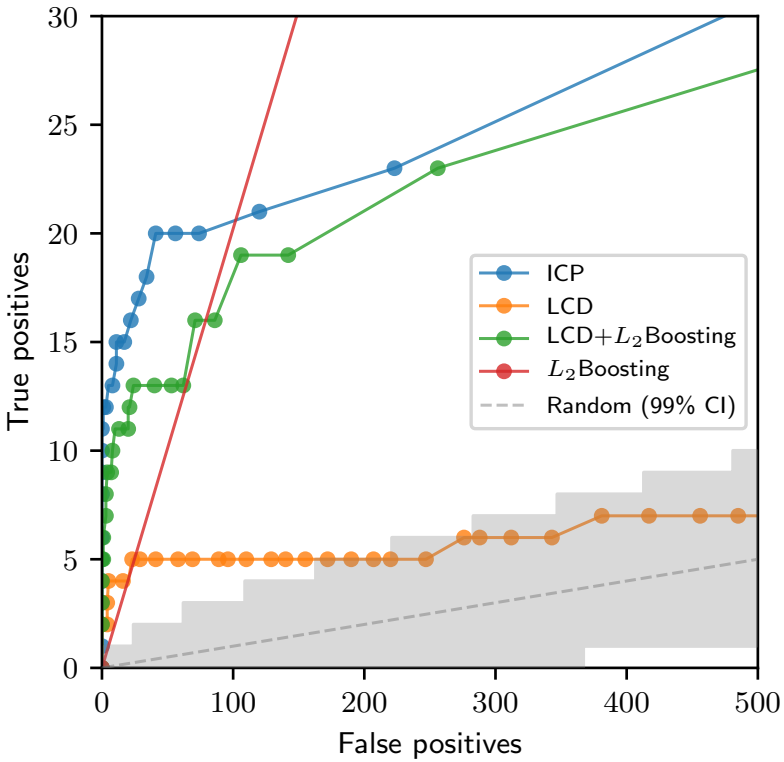
False positive:



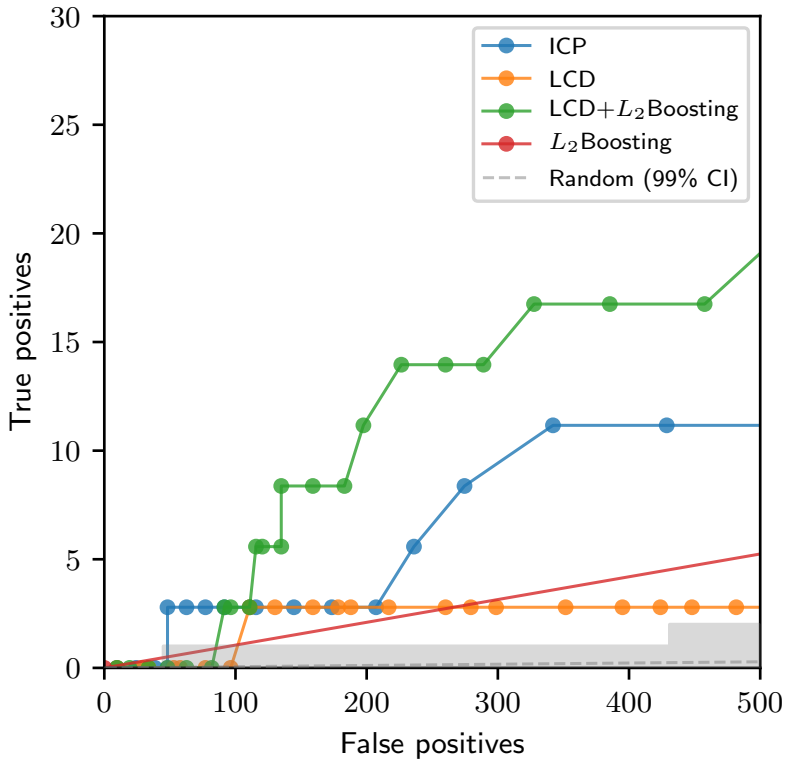
(Training data: Observational and Interventional. Test data: single intervention.)

Idea: introduce binary context variable ($C = 0$: observational; $C = 1$: interventional). JCI Assumption 1 seems justified, so apply LCD or ICP.

Beyond RCTs: First successful large-scale validation



Prediction Error (Internal Validation),



SGD DB (External Validation)

ICP: [Meinshausen et al., 2016];
LCD: HD-LCD [Versteeg and Mooij, 2019]

Part VII

Extensions for Cycles

Extensions to the cyclic case

The whole theory can be extended to allow for cycles. Major complication:

Complication

For cyclic SCMs, induced distributions may not exist or may not be unique, and this may change under interventions [Bongers et al., 2016].

We introduced the class of simple SCMs and showed that for this class, we

- get a generalized Markov property (replacing d-separation with σ-separation) [Forré and Mooij, 2017]
- retain a causal interpretation of the graph [Bongers et al., 2016];
- can define marginalized SCMs [Bongers et al., 2016];
- get a generalized do-calculus, adjustment criteria and identifiability algorithm [Forré and Mooij, 2019];
- can easily generalize causal discovery algorithms (RCT, ASD, LCD, ICP, FCI) [Forré and Mooij, 2018, Mooij et al., 2020];
- the JCI framework still applies [Mooij et al., 2020].

Conclusion

Causality is an important notion in daily life and in science, but underexplored in statistics and machine learning.

We discussed three approaches to causal discovery:

- Randomized Controlled Trials (A/B-testing), the gold standard.
- Constraint-based causal discovery from purely observational data.
- Approaches that also make use of causal background knowledge.

We introduced Joint Causal Inference, which:

- generalizes the idea of RCT to multiple context and system variables;
- does not require knowledge of the intervention targets and types;
- allows to exploit the strong signal in (partially) experimental data;
- proposed novel implementations (ASD-JCI, FCI-JCI);
- also works in case of cycles (assuming simple SCMs).

For details, see `https://arxiv.org/abs/1611.10351`

Thanks!

PhD students:

Philip Versteeg

Stephan Bongers

Tineke Blom

Noud de Kroon

PI:

Joris Mooij

Postdocs:

Thijs van Ommen

Patrick Forré

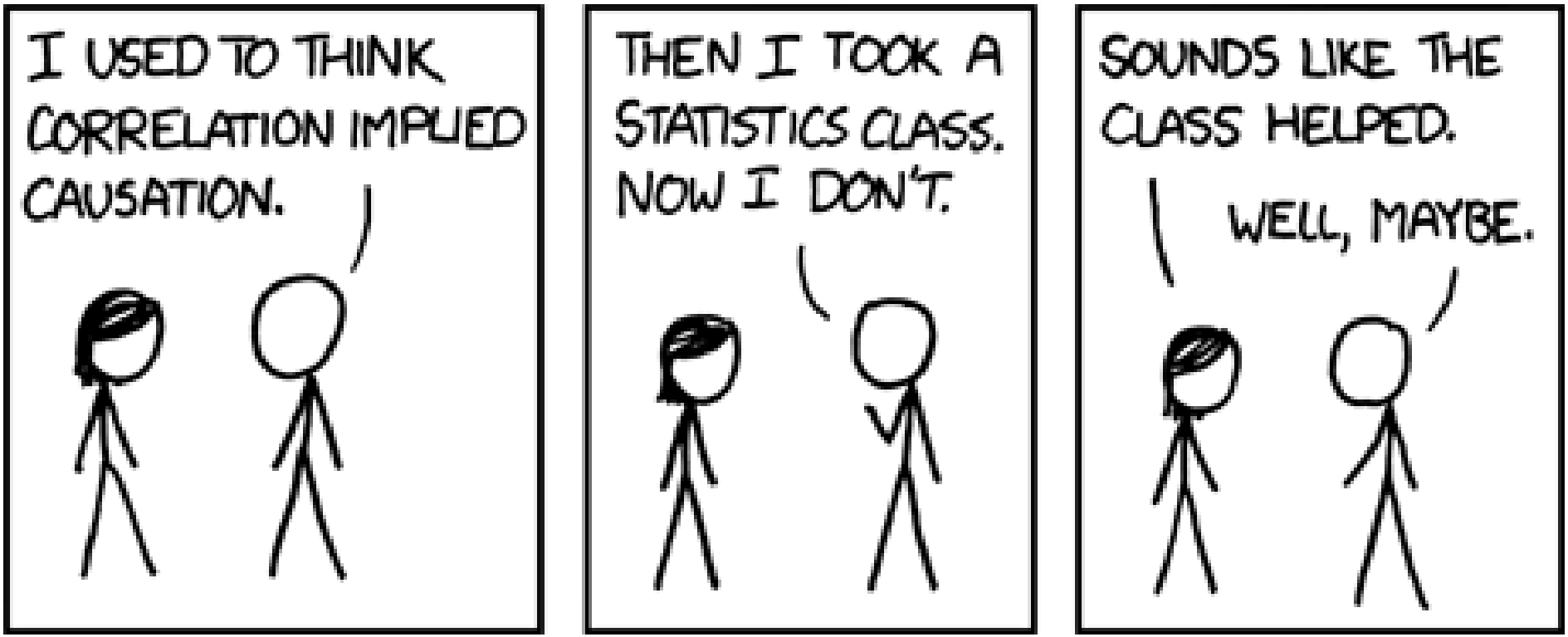
Sara Magliacane

Guest:

Tom Claassen

Funded by NWO VIDI grant 639.072.410, ERC Starting Grant n^{o} 639466

Take home message?



Source: xkcd.com

Thank you for your attention!

References I

Bongers, S. and Mooij, J. M. (2018).

From random differential equations to structural causal models: the stochastic case.
arXiv.org preprint, arXiv:1803.08784v2 [cs.AI].

Bongers, S., Peters, J., Schölkopf, B., and Mooij, J. M. (2016).

Structural causal models: Cycles, marginalizations, exogenous reparametrizations and reductions.
arXiv.org preprint, arXiv:1611.06221 [stat.ME].

Cooper, G. F. (1997).

A simple constraint-based algorithm for efficiently mining observational databases for causal relationships.
Data Mining and Knowledge Discovery, 1(2):203–224.

Forré, P. and Mooij, J. M. (2017).

Markov properties for graphical models with cycles and latent variables.
arXiv.org preprint, arXiv:1710.08775 [math.ST].

Forré, P. and Mooij, J. M. (2018).

Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders.
In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

Forré, P. and Mooij, J. M. (2019).

Causal calculus in the presence of cycles, latent confounders and selection bias.
arXiv.org preprint, arXiv:1901.00433 [stat.ML].

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014).

Constraint-based causal discovery: Conflict resolution with answer set programming.
In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 340–349, Quebec City, Quebec, Canada.

References II

Kemmeren, P., Sameith, K., van de Pasch, L., Benschop, J., Lenstra, T., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C., van Heesch, S., Kashani, M., Ampatziadis-Michailidis, G., Brok, M., Brabers, N., Miles, A., Bouwmeester, D., van Hooff, S., van Bakel, H., Sluiters, E., Bakker, L., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M., and Holstege, F. (2014).
Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors.
Cell, 157:740–752.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018).
Domain adaptation by using causal inference to predict invariant conditional distributions.
In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31 (NeurIPS2018)*, pages 10869–10879. Curran Associates, Inc.

Meek, C. (1995).
Strong completeness and faithfulness in Bayesian networks.
In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*.

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. (2016).
Methods for causal inference from gene perturbation experiments and validation.
Proceedings of the National Academy of Sciences of the United States of America, 113(27):7361–7368.

Mooij, J. M., Magliacane, S., and Claassen, T. (2020).
Joint causal inference from multiple contexts.
arXiv.org preprint, https://arxiv.org/abs/1611.10351v5 [cs.LG].
Forthcoming in Journal of Machine Learning Research.

Pearl, J. (2000).
Causality: Models, Reasoning, and Inference.
Cambridge University Press.

References III

Strotz, R. and Wold, H. (1960).

Recursive vs. nonrecursive systems: An attempt at synthesis.
Econometrica, 28(2):417–427.

Tian, J. (2002).

Studies in Causal Reasoning and Learning.
PhD thesis, University of California, Los Angeles.

Versteeg, P. J. and Mooij, J. M. (2019).

Boosting local causal discovery in high-dimensional expression data.
arXiv.org preprint, arXiv:1910.02505v2 [stat.ML].
Accepted for publication in BIBM 2019.

Wright, S. (1921).

Correlation and causation.
Journal of Agricultural Research, 20:557–585.

From SCMs to Potential Outcomes

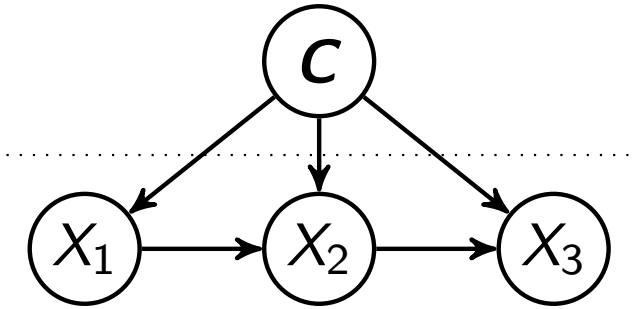
We can connect SCMs to the potential outcome framework (popular in the statistical literature):

Definition

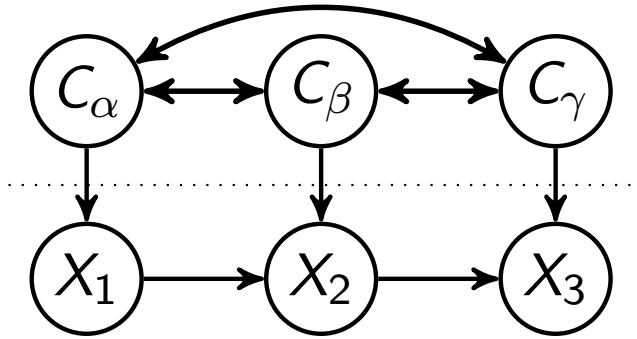
Given a simple SCM $\mathcal{M}$ and let $\boldsymbol{E} \sim \mathbb{P}_{\mathcal{E}}$. For any subset $I \subseteq \mathcal{I}$ and value $\boldsymbol{\xi}_I$, define the potential outcome $\boldsymbol{X}_{\boldsymbol{\xi}_I} := \boldsymbol{g}_{\mathcal{M}_{\text{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)}}(\boldsymbol{E})$.

Why multiple context variables?

In general, one can learn more about a system when using multiple context variables than when using a single one.



Not identifiable



Identifiable under JCI Assumptions 1,2