

# MLSS 2019: Causality

Joris Mooij

j.m.mooij@uva.nl

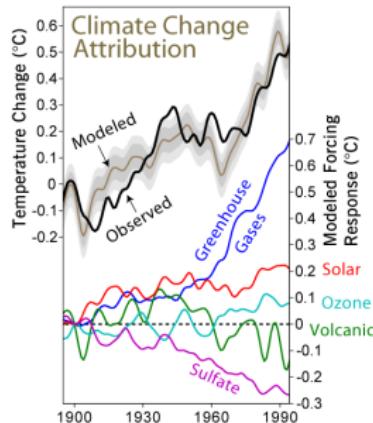


UNIVERSITY OF AMSTERDAM

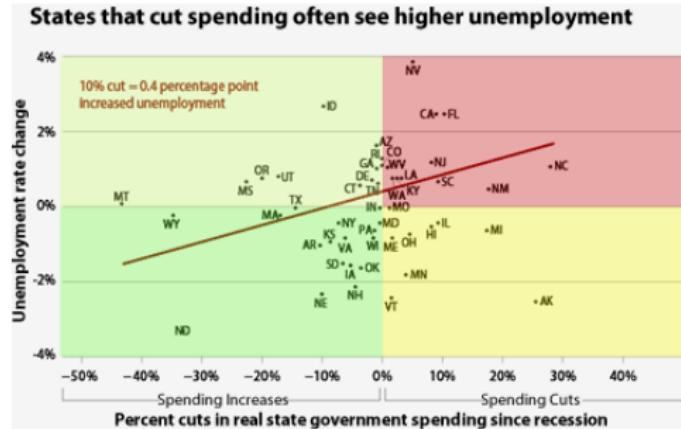
August 29-30, 2019

# Many questions in science are *causal*

## Climatology:



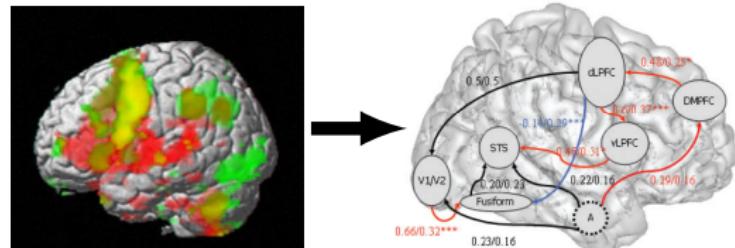
## Economy:



## Medicine:



## Neuroscience:



Causality is clearly an important notion in daily life and in science.

- But how should we formalize the notion of causality?
- How to reason about causality?
- How can we discover causal relations from data?
- How to obtain causal predictions?
- How do they differ from ordinary predictions in ML?

That is what you will learn in this tutorial!

## Probabilistic Inference (traditional statistics / machine learning)

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

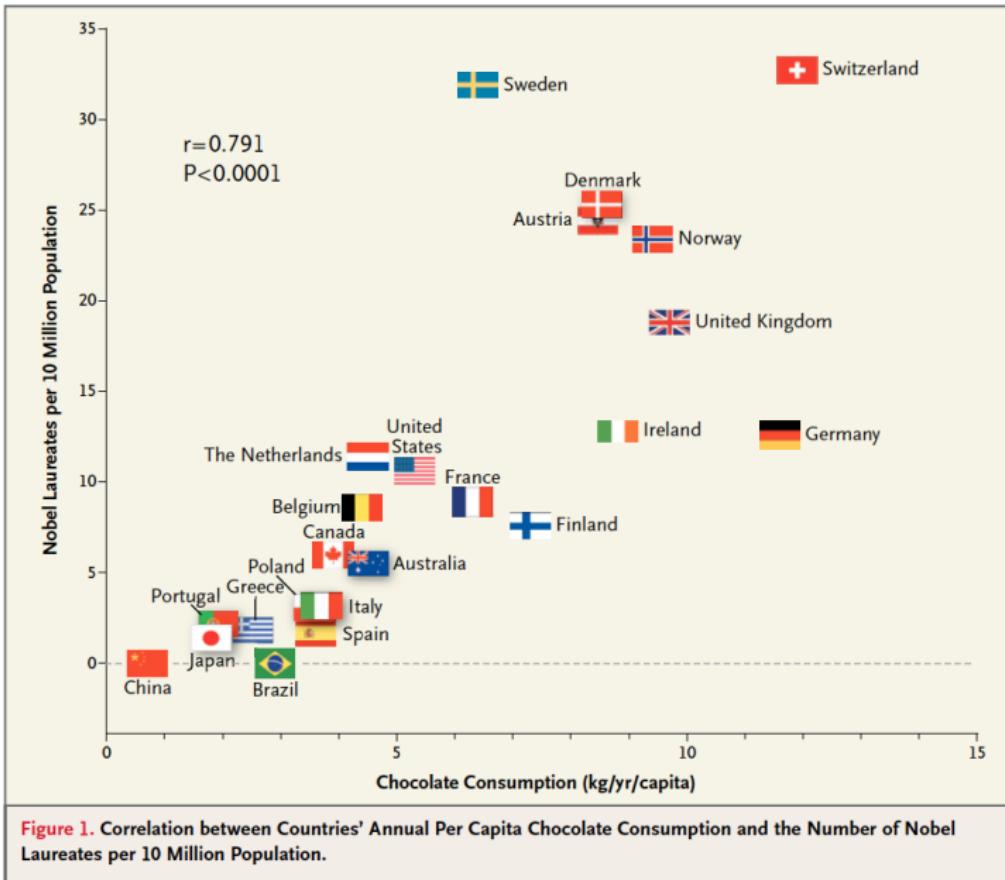
## Causal Inference

- 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?*

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

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Causation ≠ Correlation



# Causal relations

## Definition (Informal)

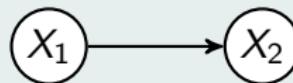
Let  $X$  and  $Y$  be two distinct variables of system.  $X$  causes  $Y$  if changing  $X$  (*intervening on  $X$* ) leads to a change of  $Y$ .

Causal graph represents causal relationships between variables graphically.

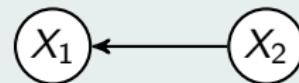
### Example



$X_1$  and  $X_2$  are causally unrelated



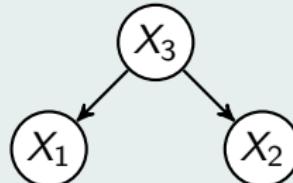
$X_1$  causes  $X_2$



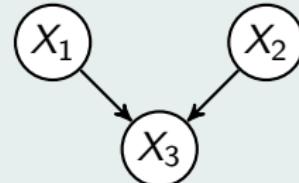
$X_2$  causes  $X_1$



$X_1$  and  $X_2$  cause each other



$X_1$  and  $X_2$  have a common cause  $X_3$



$X_1$  and  $X_2$  have a common effect  $X_3$

# Direct causation

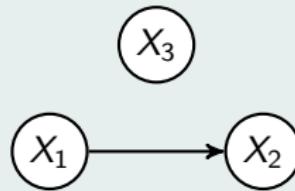
Let  $\mathbf{V} = \{X_1, \dots, X_N\}$  be a set of variables.

## Definition (Informal)

If  $X_i$  causes  $X_j$  even if all other variables  $\mathbf{V} \setminus \{X_i, X_j\}$  are held fixed at some values, then

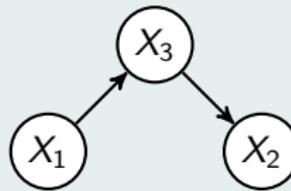
- we say that  $X_i$  causes  $X_j$  directly with respect to  $\mathbf{V}$
- we indicate this in the causal graph on  $\mathbf{V}$  by a directed edge  $X_i \rightarrow X_j$

## Example



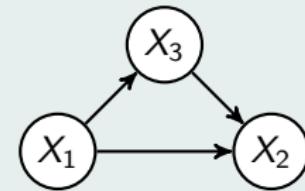
$X_1$  causes  $X_2$ ;

$X_1$  causes  $X_2$  directly  
w.r.t.  $\{X_1, X_2, X_3\}$



$X_1$  causes  $X_2$ ;

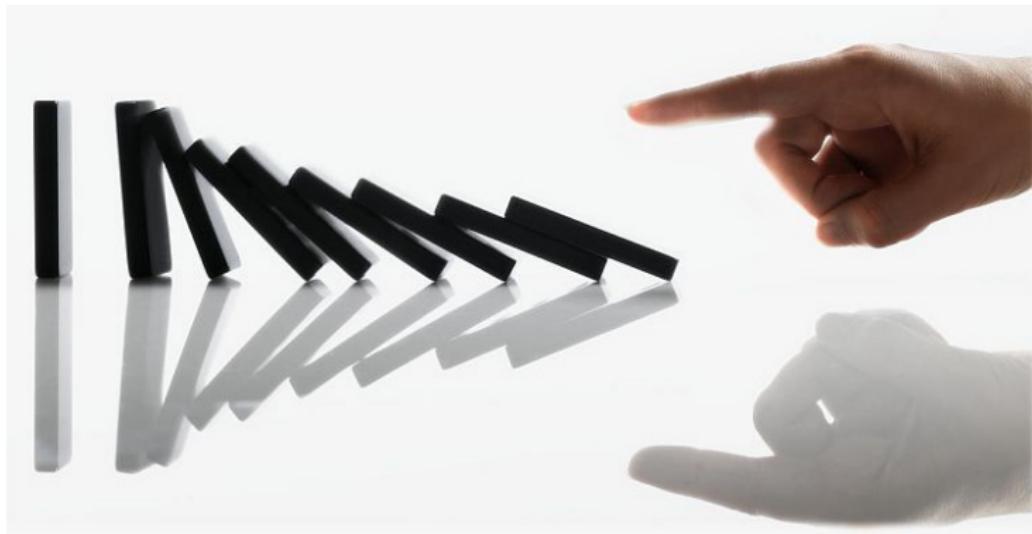
$X_1$  does not cause  $X_2$  directly  
w.r.t.  $\{X_1, X_2, X_3\}$



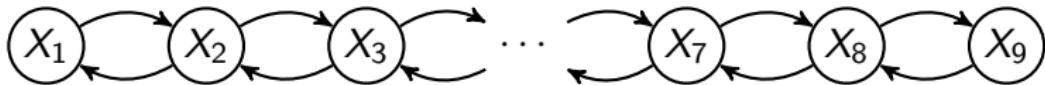
$X_1$  causes  $X_2$ ;

$X_1$  causes  $X_2$  directly  
w.r.t.  $\{X_1, X_2, X_3\}$

## Direct vs. indirect causation: Example



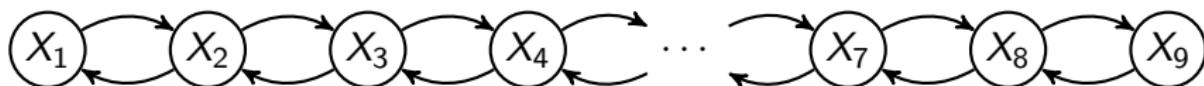
- Each stone causes *all* subsequent stones to topple.
- Each stone only **directly causes** the **next** neighboring stone to topple.
- Causal graph:



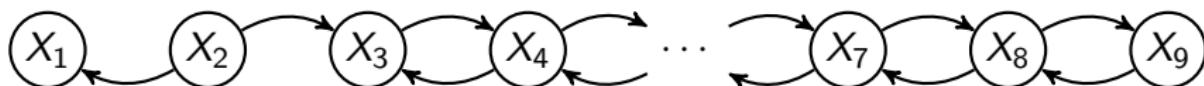
# Perfect interventions: Example

Suppose we **intervene** by keeping the second stone fixed in an “upright” position (e.g. by glueing it to the floor), an operation that we denote by  $\text{do}(X_2 = \text{upright})$ .

Before the intervention, the causal graph is:



After the intervention  $\text{do}(X_2 = \text{upright})$ , the causal graph is:



If we keep the second stone fixed, it is no longer affected by the other stones.

## Definition (Informal)

A perfect (“surgical”) intervention on a set of variables  $X \subseteq V$ , denoted  $\text{do}(X = \xi)$ , is an externally enforced change of the system that ensures that  $X$  takes on value  $\xi$  and leaves the rest of the system untouched.

The concept of perfect intervention assumes **modularity**: the causal system can be divided into two parts,  $X$  and  $V \setminus X$ , and we can make changes to one part while keeping the other part **invariant**.

## Note

The intervention changes the causal graph by removing all edges that point towards variables in  $X$  (because none of the variables can now cause  $X$ ).

# Confounders: Definition

Informally: a **confounder** is a latent common cause.

## Definition

Consider three variables  $X, Y, H$ .  $H$  confounds  $X$  and  $Y$  if:

- ①  $H$  causes  $X$  directly w.r.t.  $\{X, Y, H\}$
- ②  $H$  causes  $Y$  directly w.r.t.  $\{X, Y, H\}$

# Confounders: Definition

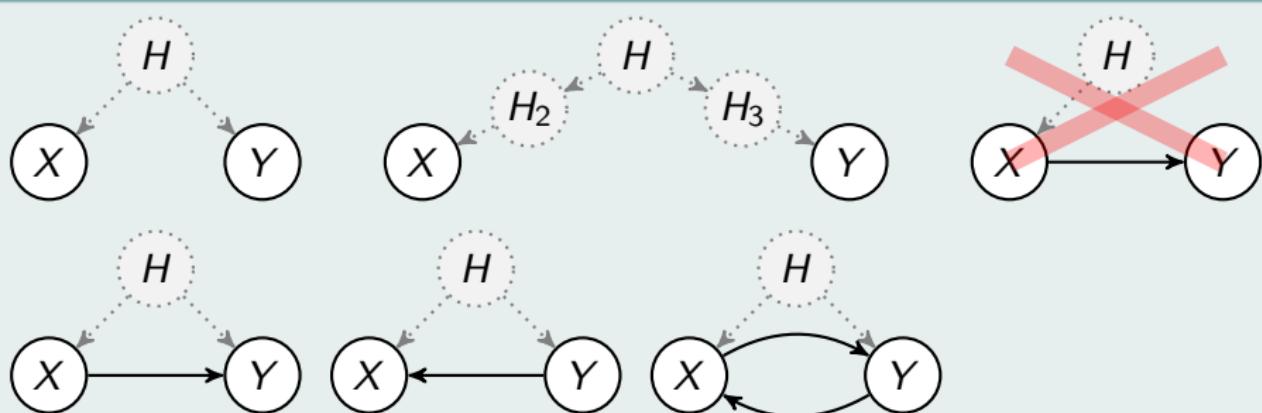
Informally: a **confounder** is a latent common cause.

## Definition

Consider three variables  $X, Y, H$ .  $H$  confounds  $X$  and  $Y$  if:

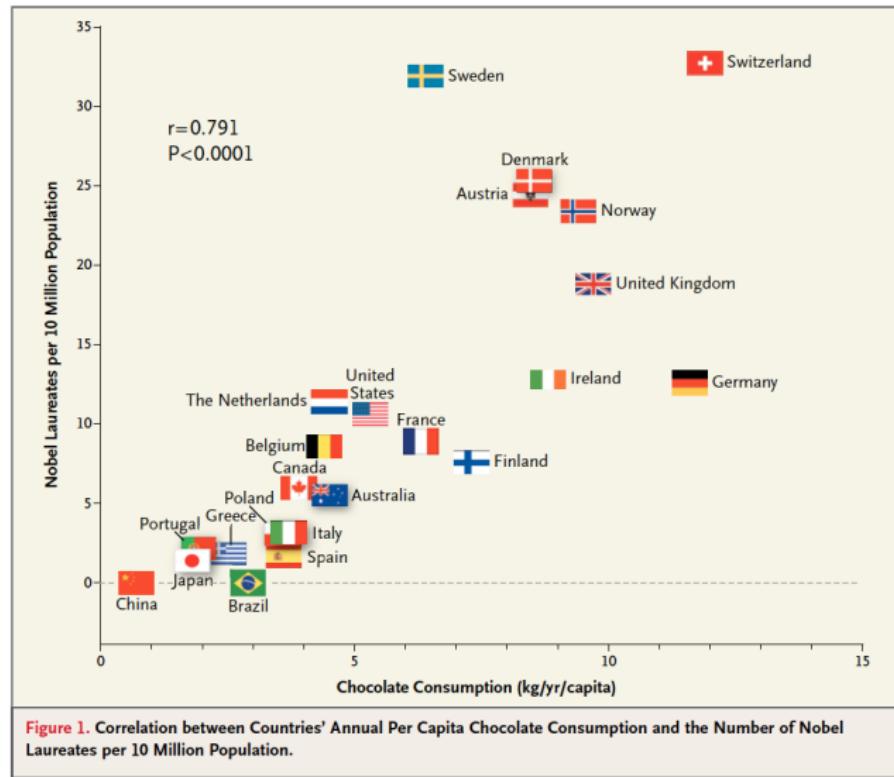
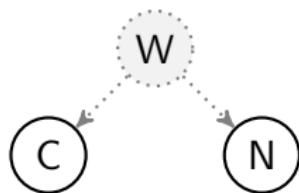
- ①  $H$  causes  $X$  directly w.r.t.  $\{X, Y, H\}$
- ②  $H$  causes  $Y$  directly w.r.t.  $\{X, Y, H\}$

## Example



# Confounders: Example

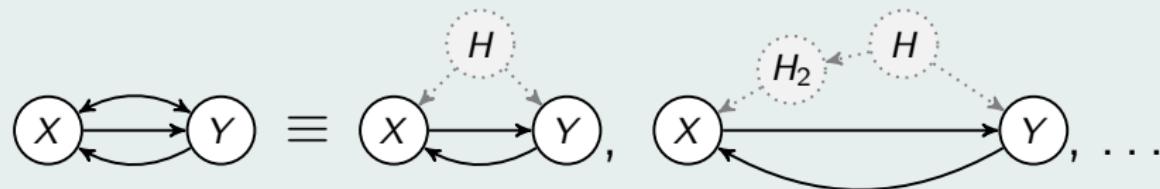
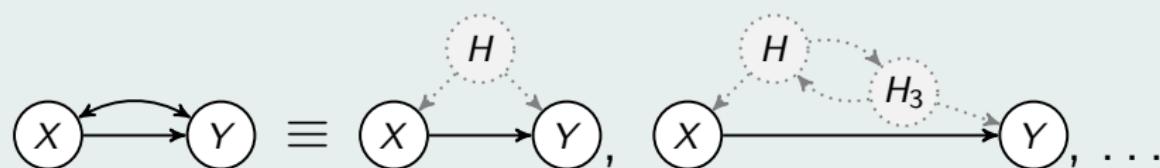
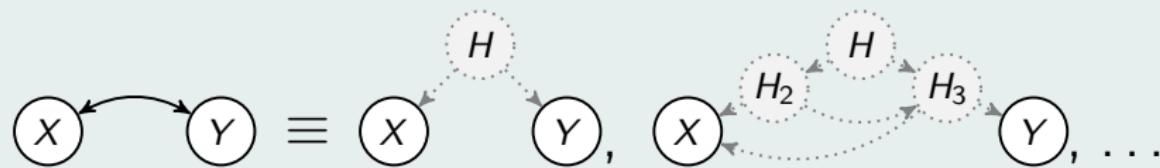
Wealth might confound chocolate consumption and Nobel prize winners.



# Confounders: Graphical notation

We denote latent confounders by **bidirected edges** in the causal graph:

Example



# Causal Cycles: Definition and Example

Let  $X, Y$  be two variables in a system.

## Definition

If  $X$  causes  $Y$  and  $X$  causes  $Y$ , then  $X$  and  $Y$  form a **causal cycle**.

# Causal Cycles: Definition and Example

Let  $X, Y$  be two variables in a system.

## Definition

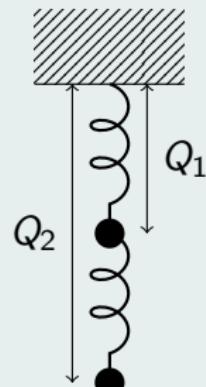
If  $X$  causes  $Y$  and  $X$  causes  $Y$ , then  $X$  and  $Y$  form a **causal cycle**.

## Example (Damped Coupled Harmonic Oscillators)

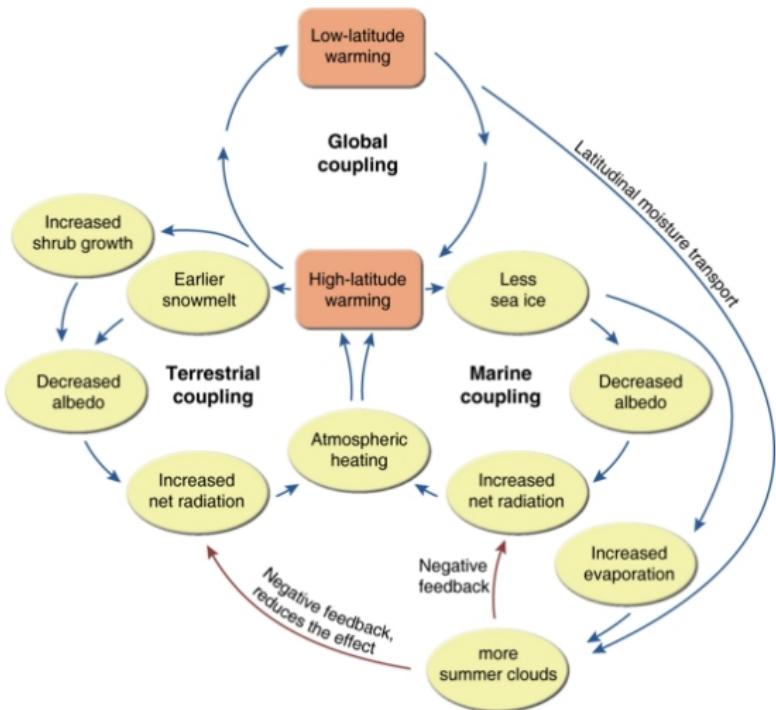
- Two masses, connected by a spring, suspended from the ceiling by another spring.
- Variables: vertical **equilibrium** positions  $Q_1$  and  $Q_2$ .
- $Q_1$  causes  $Q_2$ .
- $Q_2$  causes  $Q_1$ .
- Causal graph:



- Cannot be modeled with acyclic causal model!

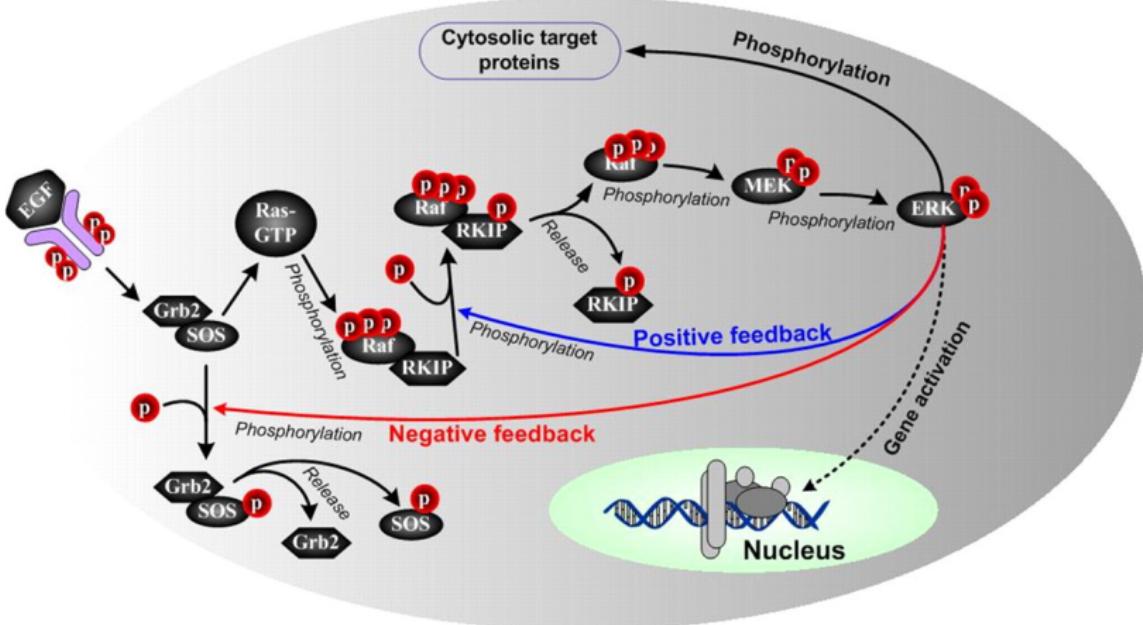


# Cycles: Relevance in Climatology



"Part of the uncertainty around future climates relates to important feedbacks between different parts of the climate system: air temperatures, ice and snow albedo (reflection of the sun's rays), and clouds." [Ahlenius, 2007]

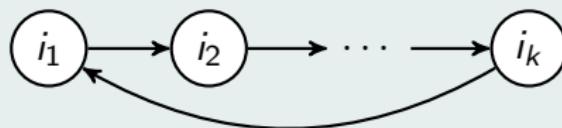
# Cycles: Relevance in Biology



"Feedback mechanisms may be critical to allow cells to achieve the fine balance between dysregulated signaling and uncontrolled cell proliferation (a hallmark of cancer) as well as the capacity to switch pathways on or off when needed for physiologic purposes." [McArthur, 2014]

## Definition

- A graph  $\mathcal{G}$  that consists of directed and bidirected edges is called **Directed Mixed Graph (DMG)**.
- If  $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k$  in  $\mathcal{G}$  then  $i_1$  is **ancestor** of  $i_k$ :  $i_1 \in \text{ang}_{\mathcal{G}}(i_k)$ .
- $\mathcal{G}$  is called **cyclic** if it contains a **directed cycle**:



- The **strongly-connected component** of a node  $i \in \mathcal{G}$  is the set of nodes  $j \in \mathcal{G}$  such that  $i$  and  $j$  are each other's ancestors.
- If  $\mathcal{G}$  does not contain such a directed cycle, it is called **acyclic**, and known as an **Acyclic Directed Mixed Graph (ADMG)**.
- If, in addition,  $\mathcal{G}$  does not contain any bidirected edges, it is called a **Directed Acyclic Graph (DAG)**.

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Defining Causality in terms of Probabilities?

When looking for a more quantitative treatment of causality, it is a natural idea to try to *define* causality in terms of probabilities.

A naïve example of such an attempt could be:

## Attempt at a definition

Given two binary random variables  $A, B$ . If

- $A$  precedes  $B$  in time, and
- $p(B = 1 | A = 1) > p(B = 1 | A = 0)$

then  $A$  causes  $B$ .

# Defining Causality in terms of Probabilities?

When looking for a more quantitative treatment of causality, it is a natural idea to try to *define* causality in terms of probabilities.

A naïve example of such an attempt could be:

## Attempt at a definition

Given two binary random variables  $A, B$ . If

- $A$  precedes  $B$  in time, and
- $p(B = 1 | A = 1) > p(B = 1 | A = 0)$

then  $A$  causes  $B$ .

This does not work, as exemplified by *Simpson's paradox*.

## Exercise

Please make Exercise 1.1.

## Example (Simpson's paradox)

We collect electronic patient records to investigate the effectiveness of a new drug against a certain disease. We find that:

- ① The probability of recovery is higher for patients that took the drug:

$$p(\text{recovery} \mid \text{drug}) > p(\text{recovery} \mid \text{no drug})$$

- ② For **both male and female** patients, the relation is **opposite**:

$$p(\text{recovery} \mid \text{drug, male}) < p(\text{recovery} \mid \text{no drug, male})$$

$$p(\text{recovery} \mid \text{drug, female}) < p(\text{recovery} \mid \text{no drug, female})$$

Does the drug cause recovery? I.e., would *you* use this drug if you are ill?

## Example (Simpson's paradox)

We collect electronic patient records to investigate the effectiveness of a new drug against a certain disease. We find that:

- ① The probability of recovery is higher for patients that took the drug:

$$p(\text{recovery} \mid \text{drug}) > p(\text{recovery} \mid \text{no drug})$$

- ② For **both male and female** patients, the relation is **opposite**:

$$p(\text{recovery} \mid \text{drug, male}) < p(\text{recovery} \mid \text{no drug, male})$$

$$p(\text{recovery} \mid \text{drug, female}) < p(\text{recovery} \mid \text{no drug, female})$$

Does the drug cause recovery? I.e., would *you* use this drug if you are ill?

*Note: Big data and deep learning do not help us here!*

# Quantitative Models of Causality

Problems like these have historically prevented statisticians from considering causality.

Nonetheless, different approaches have been proposed to model causality in a quantitative way:

- Potential outcome framework
- Causal Bayesian Networks
- **Structural Causal Models (SCMs)**

We will use SCMs, as they are arguably the most general of the three:

- SCMs can model **cycles** naturally (close connections to ODE models from physics, chemistry, biology, engineering, . . . )
- Acyclic SCMs are closed under **marginalization** (can efficiently handle latent variables)
- SCMs can model **counterfactuals** (provides alternative to potential outcome framework)
- SCMs generalize Causal Bayesian Networks

# Structural Causal Models: Concepts

SCMs turn things upside down: rather than defining causality in terms of probabilities, probability distributions are defined by a causal model, thereby avoiding traps like Simpson's paradox.

- The *system* we are modeling is described by **endogenous variables**; endogenous variables are:
  - observed,
  - modeled by **structural equations**.
- The *environment* of the system is described by **exogenous variables**; exogenous variables are:
  - latent (unobserved),
  - modeled by **probability distributions**,
  - *not caused* by endogenous variables,
  - provide the “source” of randomness.
- *Each endogenous variable has its own structural equation, which describes how this variable depends on its direct causes.*
- SCMs are equipped with a notion of **perfect intervention**, which gives them a *causal* semantics.

# Structural Causal Models: Example

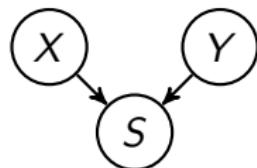
Endogenous variables (binary):

$X$ : the battery is charged

$Y$ : the start engine is operational

$S$ : the car starts

Causal graph:



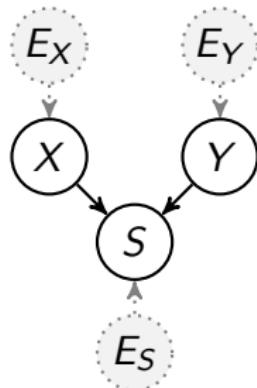
Exogenous variables (latent, independent, binary):

$$E_X \sim \text{Ber}(0.95)$$

$$E_Y \sim \text{Ber}(0.99)$$

$$E_S \sim \text{Ber}(0.999)$$

Augmented graph:



Structural equations (one per endogenous variable):

$$X = f_X(E_X) = E_X$$

$$Y = f_Y(E_Y) = E_Y$$

$$S = f_S(X, Y, E_S) = X \wedge Y \wedge E_S$$

# Structural Causal Models: Formal Definition

Definition ([Wright, 1921, Pearl, 2000, Bongers et al., 2018])

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

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

## Definition

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

# Structural Causal Models: Example

## Example

Structural Causal Model  $\mathcal{M}$ :

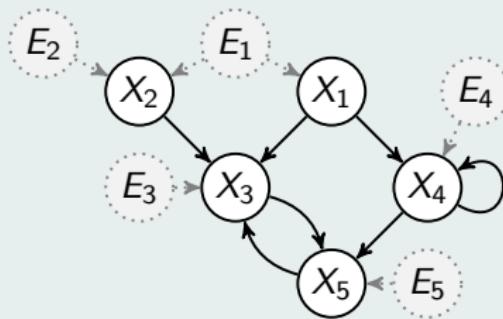
Formally:

$$(\mathcal{X}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}}) = (\prod_{i=1}^5 \mathbb{R}, \prod_{j=1}^5 \mathbb{R}, (f_1, \dots, f_5), \prod_{j=1}^5 \mathbb{P}_{\mathcal{E}_j})$$

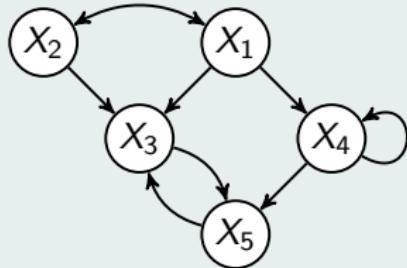
Informally:

$$\begin{array}{ll} X_1 = f_1(E_1) & E_1 \sim \mathbb{P}_{\mathcal{E}_1} \\ X_2 = f_2(E_1, E_2) & E_2 \sim \mathbb{P}_{\mathcal{E}_2} \\ X_3 = f_3(X_1, X_2, X_5, E_3) & E_3 \sim \mathbb{P}_{\mathcal{E}_3} \\ X_4 = f_4(X_1, X_4, E_4) & E_4 \sim \mathbb{P}_{\mathcal{E}_4} \\ X_5 = f_5(X_3, X_4, E_5) & E_5 \sim \mathbb{P}_{\mathcal{E}_5} \end{array}$$

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



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



# (Augmented) Graphs

## Definition

The components of the causal mechanism usually do not depend on *all* variables: for  $i \in \mathcal{I}$ ,

$$X_i = f_i(\mathbf{x}_{\text{pa}_i^{\mathcal{I}}}, \mathbf{e}_{\text{pa}_i^{\mathcal{J}}})$$

where  $f_i$  only depends on  $\text{pa}_i^{\mathcal{I}} \subseteq \mathcal{I}$  (the **endogenous parents of  $i$** ) and  $\text{pa}_i^{\mathcal{J}} \subseteq \mathcal{J}$  (the **exogenous parents of  $i$** ).

## Definition

The **augmented graph**  $\mathcal{G}^a(\mathcal{M})$  of SCM  $\mathcal{M}$  is a directed graph with nodes  $\mathcal{I} \cup \mathcal{J}$  and an edge  $k \rightarrow i$  iff  $k \in \text{pa}_i^{\mathcal{I}} \cup \text{pa}_i^{\mathcal{J}}$  is a parent of  $i \in \mathcal{I}$ .

## Definition

The **graph**  $\mathcal{G}(\mathcal{M})$  of SCM  $\mathcal{M}$  is a DMG with nodes  $\mathcal{I}$ , directed edges  $k \rightarrow i$  iff  $k \in \text{pa}_i^{\mathcal{I}}$ , and bidirected edges  $k \leftrightarrow i$  iff  $\text{pa}_i^{\mathcal{J}} \cap \text{pa}_k^{\mathcal{J}} \neq \emptyset$ .

## Definition

An SCM  $\mathcal{M}$  is said to be **uniquely solvable w.r.t.  $\mathcal{O} \subseteq \mathcal{I}$**  if there exists a measurable mapping  $\mathbf{g}_{\mathcal{O}} : \mathcal{X}_{(\text{pa}_{\mathcal{H}}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}} \times \mathcal{E}_{\text{pa}_{\mathcal{H}}(\mathcal{O}) \cap \mathcal{J}} \rightarrow \mathcal{X}_{\mathcal{O}}$  such that for  $\mathbb{P}_{\mathcal{E}}$ -almost every  $\mathbf{e}$  for all  $\mathbf{x} \in \mathcal{X}$ :

$$\mathbf{x}_{\mathcal{O}} = \mathbf{g}_{\mathcal{O}}(\mathbf{x}_{(\text{pa}_{\mathcal{H}}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}}, \mathbf{e}_{\text{pa}_{\mathcal{H}}(\mathcal{O}) \cap \mathcal{J}}) \iff \mathbf{x}_{\mathcal{O}} = \mathbf{f}_{\mathcal{O}}(\mathbf{x}, \mathbf{e}).$$

(Loosely speaking: if the structural equations for  $\mathcal{O}$  provide a unique solution for  $\mathbf{x}_{\mathcal{O}}$  in terms of the other variables).

## Example

An SCM with structural equations:

$$\begin{cases} X_1 = X_1 \\ X_2 = X_1 + X_3 \\ X_3 = X_3 + 1 \end{cases}$$

is uniquely solvable w.r.t.  $\{X_2\}$  but not w.r.t. any other subset.

For simplicity we will here assume only a special subclass of SCMs:

## Definition

We call an SCM  $\mathcal{M}$  **simple** if it is uniquely solvable with respect to any subset  $\mathcal{O} \subseteq \mathcal{I}$ .

## Lemma

*If  $\mathcal{G}(\mathcal{M})$  is acyclic,  $\mathcal{M}$  is simple.*

- The class of simple SCMs extends the class of acyclic SCMs by allowing for (weak) cyclic causal relations, while preserving most of the simplicity and convenience of acyclic SCMs.
- The theory for non-simple SCMs is considerably more involved [Bongers et al., 2018].
- Simple SCMs induce modular SCMs (mSCMs) [Forré and Mooij, 2017].

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

Definition (Perfect Interventions, [Pearl, 2000])

- The perfect intervention  $\text{do}(\mathbf{X}_I = \xi_I)$  enforces  $\mathbf{X}_I$  to attain value  $\xi_I$ .
- This changes the SCM  $\mathcal{M} = \langle \mathcal{X}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}} \rangle$  into the intervened SCM  $\mathcal{M}_{\text{do}(\mathbf{X}_I = \xi_I)} = \langle \mathcal{X}, \mathcal{E}, \tilde{\mathbf{f}}, \mathbb{P}_{\mathcal{E}} \rangle$  where

$$\tilde{f}_i(\mathbf{x}, \mathbf{e}) = \begin{cases} \xi_i & i \in I \\ f_i(\mathbf{x}_{\text{pa}_i^{\mathcal{I}}}, \mathbf{e}_{\text{pa}_i^{\mathcal{I}}}) & i \notin I. \end{cases}$$

- Interpretation: overrides default causal mechanisms that normally would determine the values of the intervened variables.
- In the (augmented) graph, the intervention removes all incoming edges with an arrowhead at any intervened variable  $i \in I$ .

# Interventions: Example

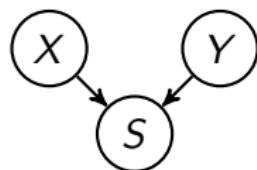
Endogenous variables (binary):

$X$ : the battery is charged

$Y$ : the start engine is operational

$S$ : the car starts

Causal graph:



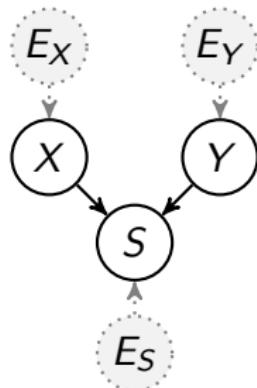
Exogenous variables (latent, independent, binary):

$$E_X \sim \text{Ber}(0.95)$$

$$E_Y \sim \text{Ber}(0.99)$$

$$E_Z \sim \text{Ber}(0.999)$$

Augmented graph:



Structural equations (one per endogenous variable):

$$X = E_X$$

$$Y = E_Y$$

$$S = X \wedge Y \wedge E_S$$

# Interventions: Example

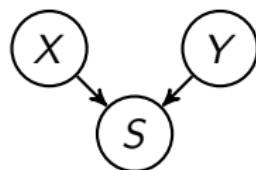
Endogenous variables (binary):

$X$ : the battery is charged

$Y$ : the start engine is operational

$S$ : the car starts

Causal graph:



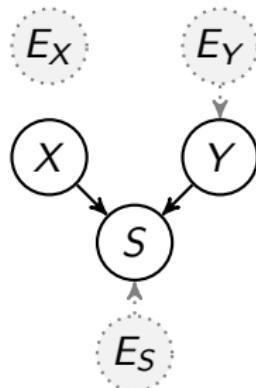
Exogenous variables (latent, independent, binary):

$$E_X \sim \text{Ber}(0.95)$$

$$E_Y \sim \text{Ber}(0.99)$$

$$E_Z \sim \text{Ber}(0.999)$$

Augmented graph:



Structural equations (one per endogenous variable):

after charging the battery  $\text{do}(X = 1)$ :

$$X = 1$$

$$Y = E_Y$$

$$S = X \wedge Y \wedge E_S$$

# Interventions: Example

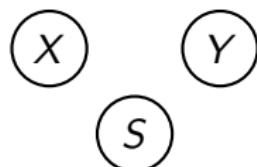
Endogenous variables (binary):

$X$ : the battery is charged

$Y$ : the start engine is operational

$S$ : the car starts

Causal graph:



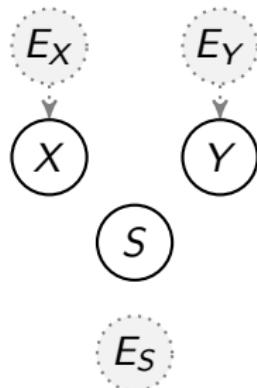
Exogenous variables (latent, independent, binary):

$$E_X \sim \text{Ber}(0.95)$$

$$E_Y \sim \text{Ber}(0.99)$$

$$E_Z \sim \text{Ber}(0.999)$$

Augmented graph:



Structural equations (one per endogenous variable):

after loosing the key  $\text{do}(S = 0)$ :

$$X = E_X$$

$$Y = E_Y$$

$$S = 0$$

# Interventions: Example

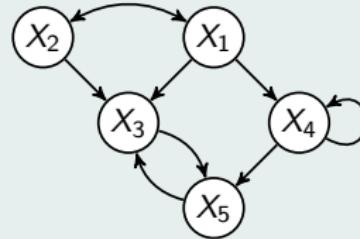
## Example

Observational (no intervention):

SCM  $\mathcal{M}$ :

$$\begin{array}{ll} X_1 = f_1(E_1) & E_1 \sim \mathbb{P}_{\mathcal{E}_1} \\ X_2 = f_2(E_1, E_2) & E_2 \sim \mathbb{P}_{\mathcal{E}_2} \\ X_3 = f_3(X_1, X_2, X_5, E_3) & E_3 \sim \mathbb{P}_{\mathcal{E}_3} \\ X_4 = f_4(X_1, X_4, E_4) & E_4 \sim \mathbb{P}_{\mathcal{E}_4} \\ X_5 = f_5(X_3, X_4, E_5) & E_5 \sim \mathbb{P}_{\mathcal{E}_5} \end{array}$$

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

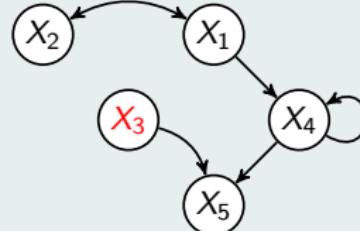


Intervention  $\text{do}(X_3 = \xi_3)$ :

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

$$\begin{array}{ll} X_1 = f_1(E_1) & E_1 \sim \mathbb{P}_{\mathcal{E}_1} \\ X_2 = f_2(E_1, E_2) & E_2 \sim \mathbb{P}_{\mathcal{E}_2} \\ \textcolor{red}{X_3 = \xi_3} & E_3 \sim \mathbb{P}_{\mathcal{E}_3} \\ X_4 = f_4(X_1, X_4, E_4) & E_4 \sim \mathbb{P}_{\mathcal{E}_4} \\ X_5 = f_5(X_3, X_4, E_5) & E_5 \sim \mathbb{P}_{\mathcal{E}_5} \end{array}$$

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



# Observational Distribution(s)

## Definition (Reminder)

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

## Definition

For  $(\mathbf{X}, \mathbf{E})$  a solution of SCM  $\mathcal{M}$ , we call  $\mathbb{P}^{\mathbf{X}}$  an **observational distribution of  $\mathcal{M}$** .

An important special case:

## Proposition

*If  $\mathcal{M}$  is simple, then its observational distribution exists and is unique.*

## Definition

Given a simple SCM  $\mathcal{M}$  and a fixed background measure on  $\mathbf{X}$ , we denote the density of the observational distribution as  $p_{\mathcal{M}}(\mathbf{x})$ .

# Interventional Distribution(s)

A perfect intervention on  $\mathcal{M}$  may change the distributions.

## Definition

We call the family of sets of observational distributions of  $\mathcal{M}_{\text{do}(X_I=\xi_I)}$  (for  $I \subseteq \mathcal{I}$ ,  $\xi_I \subseteq \mathcal{X}_I$ ) the **interventional distributions of  $\mathcal{M}$** .

## Proposition

If  $\mathcal{M}$  is simple, then all intervened SCMs  $\mathcal{M}_{\text{do}(X_I=\xi_I)}$  are simple, and hence all interventional distributions of a simple SCM exist and are unique.

## Definition ([Pearl, 2000])

Given a simple SCM  $\mathcal{M}$  and a fixed background measure on  $\mathbf{X}$ , we denote the density of the interventional distributions as  $p_{\mathcal{M}}(\mathbf{x} \mid \text{do}(\mathbf{X}_I = \xi_I))$

*Crucial difference with traditional probabilistic models: SCMs simultaneously model all distributions that are obtained under all perfect interventions on a system.*

# Self-cycles

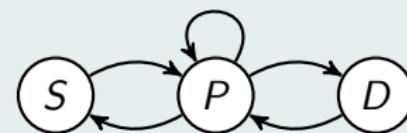
## Definition

We say  $\mathcal{M}$  has a **self-cycle** at  $i \in \mathcal{I}$  if  $i \in \text{pa}_{\mathcal{M}}^{\mathcal{I}}(i)$ .

## Example (Price-supply-demand)

Consider an SCM with three endogenous variables (Price, Supply and Demand) modeling a free market:

$$\begin{aligned}S &= \alpha P + E_S \\D &= \beta P + E_D \\P &= P + (S - D)\end{aligned}$$



The structural equation for  $P$  has a self-cycle that cannot be removed without changing the observational and interventional distributions.

Self-cycles complicate matters considerably [Bongers et al., 2018].

## Proposition

*Simple SCMs are equivalent to SCMs without self-cycles.*

## Definition

Let  $\mathcal{M}$  be a simple SCM. If  $i \rightarrow j \in \mathcal{G}(\mathcal{M})$  we call  $i$  a **direct cause of  $j$  according to  $\mathcal{M}$** .

We can now formalize our earlier informal definition of direct cause as a sufficient condition:

## Proposition

Let  $\mathcal{M}$  be a simple SCM. If there exist interventions  $\text{do}(\mathbf{X}_{\mathcal{I} \setminus \{j\}} = \xi)$  and  $\text{do}(\mathbf{X}_{\mathcal{I} \setminus \{j\}} = \xi')$  such that  $\xi_{\mathcal{I} \setminus \{i,j\}} = \xi'_{\mathcal{I} \setminus \{i,j\}}$  and  $\xi_i \neq \xi'_i$  such that

$$\mathbb{P}_{\mathcal{M}}(X_j | \text{do}(\mathbf{X}_{\mathcal{I} \setminus \{j\}} = \xi)) \neq \mathbb{P}_{\mathcal{M}}(X_j | \text{do}(\mathbf{X}_{\mathcal{I} \setminus \{j\}} = \xi'))$$

then  $i$  is a direct cause of  $j$  according to  $\mathcal{M}$ , i.e.,  $i \rightarrow j \in \mathcal{G}(\mathcal{M})$ .

(Interestingly, a necessary condition is not known)

# Causal Interpretation of Directed Paths

## Definition

Let  $\mathcal{M}$  be a simple SCM. If there exists a directed path  $i \rightarrow \dots \rightarrow j \in \mathcal{G}(\mathcal{M})$ , i.e., if  $i \in \text{an}_{\mathcal{G}(\mathcal{M})}(j)$ , then we call  $i$  a **cause of  $j$  according to  $\mathcal{M}$** .

We can now formalize our earlier informal definition of cause as a sufficient condition:

## Proposition

Let  $\mathcal{M}$  be a simple SCM. If there exist interventions  $\text{do}(X_i = \xi)$  and  $\text{do}(X_i = \xi')$  with  $\xi \neq \xi'$  such that

$$\mathbb{P}_{\mathcal{M}}(X_j | \text{do}(X_i = \xi)) \neq \mathbb{P}_{\mathcal{M}}(X_j | \text{do}(X_i = \xi'))$$

then  $i$  is a cause of  $j$  according to  $\mathcal{M}$ , i.e.,  $i \in \text{an}_{\mathcal{G}(\mathcal{M})}(j)$ .

(Interestingly, a necessary condition is not known)

# Causal Interpretation of Bidirected Edges

## Definition

Let  $\mathcal{M}$  be a simple SCM. If there exists a bidirected edge  $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$ , then we call  $i$  and  $j$  **confounded** according to  $\mathcal{M}$ .

We can formulate a sufficient condition for confoundedness:

## Proposition

Let  $\mathcal{M}$  be a simple SCM. If  $j \rightarrow i \notin \mathcal{G}(\mathcal{M})$  and there exist an intervention  $\text{do}(\mathbf{X}_{\mathcal{I} \setminus \{i,j\}} = \boldsymbol{\xi})$  such that

$$\mathbb{P}_{\mathcal{M}}(X_j | \text{do}(i, x_i), \text{do}(\mathcal{I} \setminus \{i,j\}, \boldsymbol{\xi})) \neq \mathbb{P}_{\mathcal{M}}(X_j | X_i = x_i, \text{do}(\mathcal{I} \setminus \{i,j\}, \boldsymbol{\xi}))$$

then  $i$  and  $j$  are confounded according to  $\mathcal{M}$ .

(Again, a necessary condition is not known)

# Marginalization: “Integrating out” a subsystem (Example)

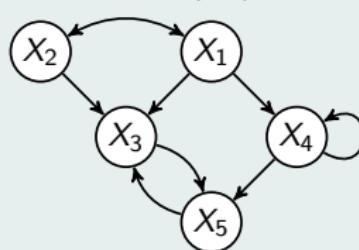
## Example

SCM for complete system:

Structural Causal Model  $\mathcal{M}$ :

$$\begin{array}{ll} X_1 = f_1(E_1) & E_1 \sim \mathbb{P}_{\mathcal{E}_1} \\ X_2 = f_2(E_1, E_2) & E_2 \sim \mathbb{P}_{\mathcal{E}_2} \\ X_3 = f_3(X_1, X_2, X_5, E_3) & E_3 \sim \mathbb{P}_{\mathcal{E}_3} \\ X_4 = f_4(X_1, X_4, E_4) & E_4 \sim \mathbb{P}_{\mathcal{E}_4} \\ X_5 = f_5(X_3, X_4, E_5) & E_5 \sim \mathbb{P}_{\mathcal{E}_5} \end{array}$$

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

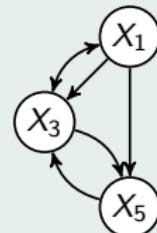
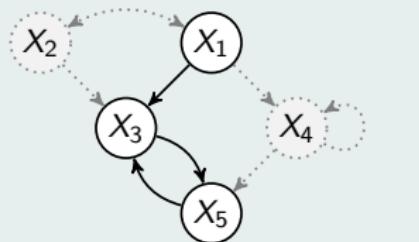


Marginalizing out  $\{X_2, X_4\}$ :

Marginalization  $\mathcal{M}^{\setminus\{2,4\}}$ :

$$\begin{array}{ll} X_1 = f_1(E_1) & E_1 \sim \mathbb{P}_{\mathcal{E}_1} \\ X_3 = f_3(X_1, g_2(E_1, E_2), X_5, E_3) & E_2 \sim \mathbb{P}_{\mathcal{E}_2} \\ & E_3 \sim \mathbb{P}_{\mathcal{E}_3} \\ X_5 = f_5(X_3, g_4(X_1, E_4), E_5) & E_4 \sim \mathbb{P}_{\mathcal{E}_4} \\ & E_5 \sim \mathbb{P}_{\mathcal{E}_5} \end{array}$$

Graph  $\mathcal{G}(\mathcal{M}^{\setminus\{2,4\}})$ :



# Marginalization: Substituting equations

Given a simple SCM  $\mathcal{M}$  and a subset of its endogenous variables  $\mathcal{L} \subseteq \mathcal{I}$ , with complement  $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$ , we can always “substitute out” the structural equations for  $\mathcal{L}$ :

$$\begin{aligned}\mathbf{X} &= \mathbf{f}(\mathbf{X}, \mathbf{E}) \\ \iff &\begin{cases} \mathbf{X}_{\mathcal{L}} = \mathbf{f}_{\mathcal{L}}(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{O}}, \mathbf{E}) \\ \mathbf{X}_{\mathcal{O}} = \mathbf{f}_{\mathcal{O}}(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{O}}, \mathbf{E}) \end{cases} \\ \iff &\begin{cases} \mathbf{X}_{\mathcal{L}} = \mathbf{g}_{\mathcal{L}}(\mathbf{X}_{\mathcal{O}}, \mathbf{E}) \\ \mathbf{X}_{\mathcal{O}} = \mathbf{f}_{\mathcal{O}}(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{O}}, \mathbf{E}) \end{cases} \\ \iff &\begin{cases} \mathbf{X}_{\mathcal{L}} = \mathbf{g}_{\mathcal{L}}(\mathbf{X}_{\mathcal{O}}, \mathbf{E}) \\ \mathbf{X}_{\mathcal{O}} = \mathbf{f}_{\mathcal{O}}(\mathbf{g}_{\mathcal{L}}(\mathbf{X}_{\mathcal{O}}, \mathbf{E}), \mathbf{X}_{\mathcal{O}}, \mathbf{E}) \end{cases}\end{aligned}$$

all hold a.s., where  $\mathbf{g}_{\mathcal{L}} : \mathcal{X}_{\mathcal{O}} \times \mathcal{E} \rightarrow \mathcal{X}_{\mathcal{L}}$  is the explicit solution of the structural equations for  $\mathbf{X}_{\mathcal{L}}$ , i.e.,

$$\mathbf{X}_{\mathcal{L}} = \mathbf{g}_{\mathcal{L}}(\mathbf{X}_{\mathcal{O}}, \mathbf{E}) \iff \mathbf{X}_{\mathcal{L}} = \mathbf{f}_{\mathcal{L}}(\mathbf{X}_{\mathcal{L}}, \mathbf{X}_{\mathcal{O}}, \mathbf{E}) \text{ a.s..}$$

# Marginalization of an SCM

Definition ([Bongers et al., 2018])

Let  $\mathcal{M} = \langle \mathcal{X}_{\mathcal{I}}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}} \rangle$  be a simple SCM,  $\mathcal{L} \subseteq \mathcal{I}$  a subset of endogenous variables and  $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$ . Then the **marginalization of  $\mathcal{M}$  on  $\mathcal{I} \setminus \mathcal{L}$**  is defined as the SCM  $\mathcal{M}^{\setminus \mathcal{L}} := \langle \mathcal{X}_{\mathcal{I} \setminus \mathcal{L}}, \mathcal{E}, \mathbf{f}^{\setminus \mathcal{L}}, \mathbb{P}_{\mathcal{E}} \rangle$ , where the marginal causal mechanism  $\mathbf{f}^{\setminus \mathcal{L}}$  is obtained by substitution:  
$$\mathbf{f}^{\setminus \mathcal{L}}(\mathbf{x}_{\mathcal{O}}, \mathbf{e}) := \mathbf{f}_{\mathcal{O}}(\mathbf{g}_{\mathcal{L}}(\mathbf{x}_{\mathcal{O}}, \mathbf{e}), \mathbf{x}_{\mathcal{O}}, \mathbf{e}).$$

## Definition

For a DMG  $\mathcal{G}$  and a subset  $\mathcal{L} \subseteq \mathcal{I}$  of nodes, the **latent projection**  $\mathcal{G}^{\setminus \mathcal{L}}$  is defined as the DMG with nodes  $\mathcal{I} \setminus \mathcal{L}$  and edges

- $i \rightarrow j$  iff there is a directed path  $i \rightarrow \ell_1 \rightarrow \cdots \rightarrow \ell_k \rightarrow j$  in  $\mathcal{G}$  with  $\ell_1, \dots, \ell_k \in \mathcal{L}$
- $i \leftrightarrow j$  iff there is a path  $i \leftarrow \ell_1 \leftarrow \cdots \leftarrow \ell_{k_1} \leftrightarrow \ell_{k_1+1} \rightarrow \cdots \rightarrow \ell_{k_2} \rightarrow j$  in  $\mathcal{G}$  with  $\ell_1, \dots, \ell_{k_1}, \dots, \ell_{k_2} \in \mathcal{L}$

# Marginalization of an SCM: Properties

The marginalization preserves the causal semantics (restricted to the remaining part of the system,  $\mathcal{I} \setminus \mathcal{L}$ ):

Theorem ([Bongers et al., 2018])

Let  $\mathcal{M} = \langle \mathcal{X}_{\mathcal{I}}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}} \rangle$  be a simple SCM and  $\mathcal{L} \subseteq \mathcal{I}$  a subset of endogenous variables.

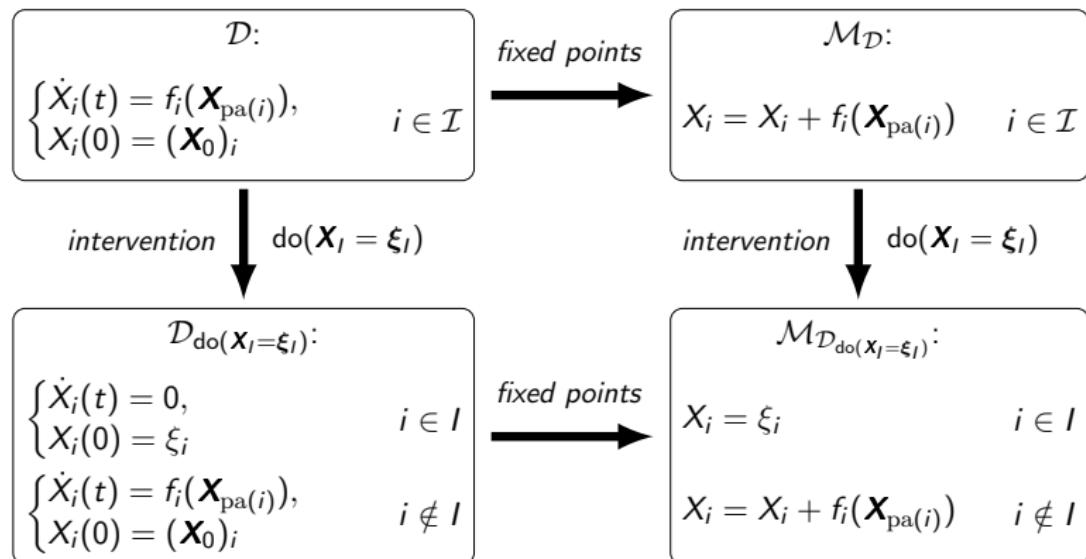
- The marginalization  $\mathcal{M}^{\setminus \mathcal{L}}$  is **interventionally equivalent** to  $\mathcal{M}$  w.r.t.  $\mathcal{I} \setminus \mathcal{L}$ . I.e., the observational distribution and all interventional distributions of  $\mathcal{M}$ , marginalized onto  $\mathcal{X}_{\mathcal{I} \setminus \mathcal{L}}$ , coincide with the corresponding ones of  $\mathcal{M}^{\setminus \mathcal{L}}$ .
- The graph  $\mathcal{G}(\mathcal{M}^{\setminus \mathcal{L}})$  of the marginalization of  $\mathcal{M}$  on  $\mathcal{I} \setminus \mathcal{L}$  is always a subgraph of the latent projection of  $\mathcal{G}(\mathcal{M})$  on  $\mathcal{I} \setminus \mathcal{L}$  (some edges may cancel out).
- The marginal SCM  $\mathcal{M}^{\setminus \mathcal{L}}$  is simple.

# Modeling ODE fixed points with an SCM

Strong motivation for (cyclic) SCMs:

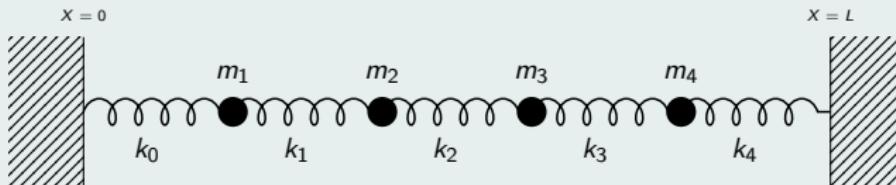
Theorem ([Mooij et al., 2013, Bongers and Mooij, 2018])

An ODE describing a dynamical system induces an SCM that models its equilibrium states, and how these change under perfect interventions.



# From ODE to SCM: Example 1

## Example (Damped coupled harmonic oscillators)



- ODE  $\mathcal{D}$ :

$$\ddot{X}_i = \frac{k_i}{m_i}(X_{i+1} - X_i - l_i) - \frac{k_{i-1}}{m_i}(X_i - X_{i-1} - l_{i-1}) - b_i \dot{X}_i$$

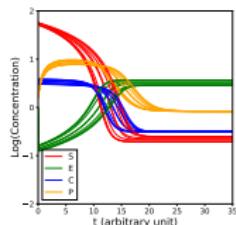
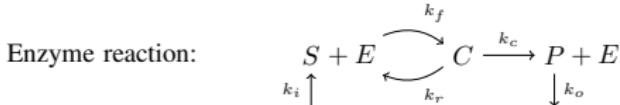
- Structural Equations of induced SCM  $\mathcal{M}_{\mathcal{D}}$ :

$$X_i = \frac{k_i(X_{i+1} - l_i) + k_{i-1}(X_{i-1} + l_{i-1})}{k_i + k_{i+1}}$$

- Graph of induced SCM  $\mathcal{G}(\mathcal{M}_{\mathcal{D}})$ :

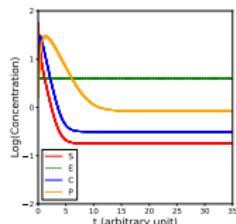


# From ODE to SCM: Example 2

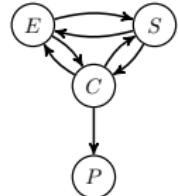


Random differential equations:

$$\begin{aligned}\frac{d}{dt}S &= k_i - k_f ES + k_r C \\ \frac{d}{dt}E &= -k_f ES + (k_r + k_c)C \\ \frac{d}{dt}C &= k_f ES - (k_r + k_c)C \\ \frac{d}{dt}P &= k_c C - k_o P\end{aligned}$$



More generally, any chemical reaction can be modeled as an SCM at equilibrium. (Note: the SCM is in general *underspecified*, i.e., it does not retain all information about the equilibrium states of the dynamical system [Blom & Mooij, 2018]).



$\downarrow \text{do}(E = \eta)$

Intervened RDE:

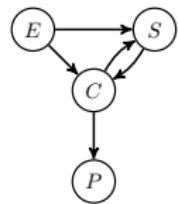
$$\begin{aligned}\frac{d}{dt}S &= k_i - k_f ES + k_r C \\ \frac{d}{dt}E &= \eta \\ \frac{d}{dt}C &= k_f ES - (k_r + k_c)C \\ \frac{d}{dt}P &= k_c C - k_o P\end{aligned}$$

$t \rightarrow \infty$

$\downarrow \text{do}(E = \eta)$

Intervened SCM:

$$\begin{aligned}S &= k_i k_f^{-1} E^{-1} - k_r k_f^{-1} E^{-1} C \\ E &= \eta \\ C &= k_f (k_r + k_c)^{-1} E S \\ P &= k_c k_o^{-1} C\end{aligned}$$



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

## Definition

Given a simple SCM  $\mathcal{M}$  and let  $\mathbf{E} \sim \mathbb{P}_{\mathcal{E}}$ . For any subset  $I \subseteq \mathcal{I}$  and value  $\xi_I$ , define the **potential outcome**  $X_{\xi_I} := g_{\mathcal{M}_{\text{do}(X_I=\xi_I)}}(\mathbf{E})$ .

Also, we can connect SCMs to causal Bayesian networks:

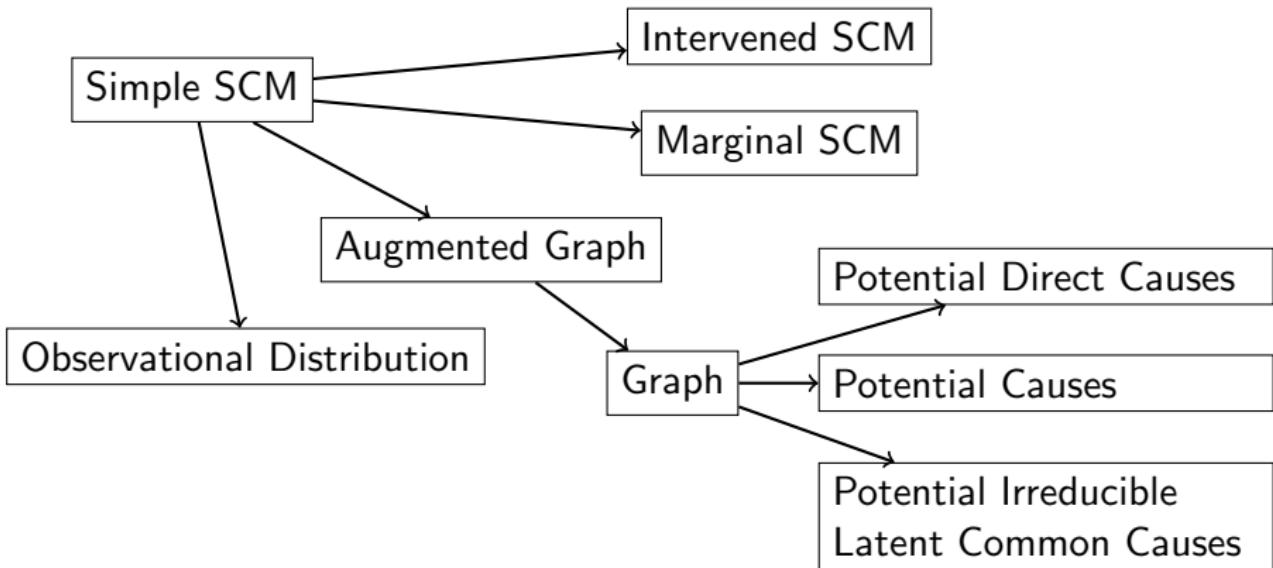
## Proposition

Given a simple SCM  $\mathcal{M}$  with a graph  $\mathcal{G}(\mathcal{M})$  that is

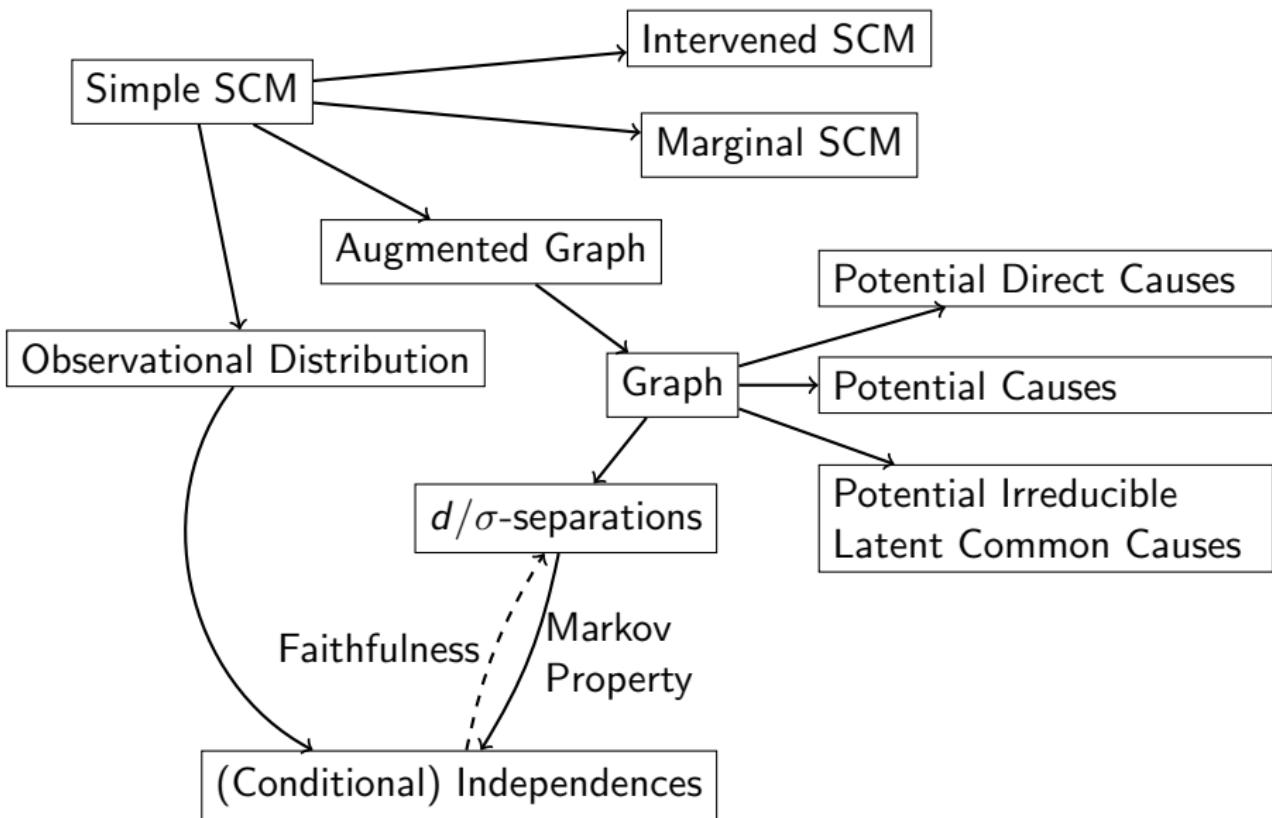
- acyclic (i.e., has no directed cycles), and
- causally sufficient (i.e., it has no bidirected edges).

Then  $\mathcal{M}$  induces a **Causal Bayesian Network**  $\langle \mathcal{G}(\mathcal{M}), p_{\mathcal{M}}(\mathbf{X}) \rangle$ . Vice versa, for every Causal Bayesian Network there exists an SCM that induces it.

# Simple SCMs: Overview



# Simple SCMs: Overview



- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

## Definition (Independence)

Given two random variables  $X, Y$ , we write  $X \perp\!\!\!\perp Y$  and say that  $X$  is independent of  $Y$  if

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

Intuitively,  $X$  is independent of  $Y$  if we do not learn anything about  $X$  when told the value of  $Y$  (or vice versa).

# (Conditional) independences

## Definition (Independence)

Given two random variables  $X, Y$ , we write  $X \perp\!\!\!\perp Y$  and say that  $X$  is independent of  $Y$  if

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

Intuitively,  $X$  is independent of  $Y$  if we do not learn anything about  $X$  when told the value of  $Y$  (or vice versa).

## Definition (Conditional Independence)

Given a third random variable  $Z$ , we write  $X \perp\!\!\!\perp Y | Z$  and say that  $X$  is (conditionally) independent from  $Y$ , given  $Z$ , if

$$p(x, y | Z = z) = p(x | Z = z)p(y | Z = z).$$

Intuitively,  $X$  is independent of  $Y$  if, given the value of  $Z$ , we do not learn anything new about  $X$  when told the value of  $Y$ .

## Definition (Paths, Ancestors)

Let  $\mathcal{G}$  be a directed mixed graph.

- A **path**  $q$  is a sequence of adjacent edges in which no node occurs more than once.
- A **directed path** is of the form  $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k$ .
- If there is a directed path from  $X$  to  $Y$ ,  $X$  is called an **ancestor** of  $Y$ .
- The ancestors of  $Y$  are denoted  $\text{an}_{\mathcal{G}}(Y)$ , and include  $Y$ .

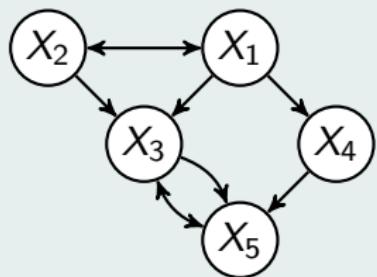
# (Directed) Paths

## Definition (Paths, Ancestors)

Let  $\mathcal{G}$  be a directed mixed graph.

- A **path**  $q$  is a sequence of adjacent edges in which no node occurs more than once.
- A **directed path** is of the form  $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k$ .
- If there is a directed path from  $X$  to  $Y$ ,  $X$  is called an **ancestor** of  $Y$ .
- The ancestors of  $Y$  are denoted  $\text{an}_{\mathcal{G}}(Y)$ , and include  $Y$ .

## Example



$X_1 \rightarrow X_3 \leftarrow X_1$  is not a path.

$X_1 \leftrightarrow X_2 \rightarrow X_3$  is a path.

$X_1 \rightarrow X_4 \rightarrow X_5$  is a directed path.

$X_4 \rightarrow X_5 \leftarrow X_3$  is not a directed path.

The ancestors of  $X_3$  are  $\{X_1, X_2, X_3\}$ .

## Definition (Colliders)

Let  $\mathcal{G}$  be a directed mixed graph, and  $q$  a path on  $\mathcal{G}$ .

- A **collider** on  $q$  is a (non-endpoint) node  $X$  on  $q$  with precisely two arrowheads pointing towards  $X$  on the adjacent edges:

$$\rightarrow X \leftarrow, \quad \rightarrow X \leftrightarrow, \quad \leftrightarrow X \leftarrow, \quad \leftrightarrow X \leftrightarrow$$

- A **non-collider** on  $q$  is any node on the path which is not a collider.

## Definition (Colliders)

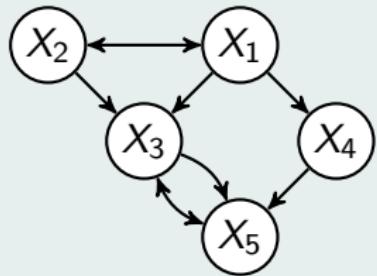
Let  $\mathcal{G}$  be a directed mixed graph, and  $q$  a path on  $\mathcal{G}$ .

- A **collider** on  $q$  is a (non-endpoint) node  $X$  on  $q$  with precisely two arrowheads pointing towards  $X$  on the adjacent edges:

$$\rightarrow X \leftarrow, \quad \rightarrow X \leftrightarrow, \quad \leftrightarrow X \leftarrow, \quad \leftrightarrow X \leftrightarrow$$

- A **non-collider** on  $q$  is any node on the path which is not a collider.

## Example



The path  $X_3 \rightarrow X_5 \leftarrow X_4$  contains a collider  $X_5$ .  
The path  $X_1 \leftrightarrow X_2 \rightarrow X_3$  contains no collider.  
 $X_5$  is a non-collider on  $X_5 \leftrightarrow X_3 \leftarrow X_1$ .

## Definition

Let  $\mathcal{G}$  be a directed mixed graph. Given a path  $q$  on  $\mathcal{G}$ , and a set of nodes  $S$ , we say that  $S$   **$d$ -blocks**  $q$  if  $q$  contains

- a non-collider which is in  $S$ , or
- a collider which is *not* an ancestor of  $S$ .

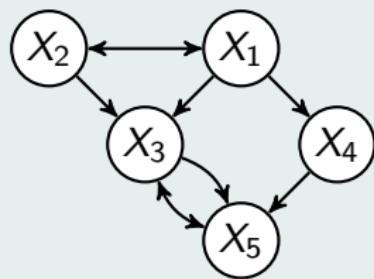
# $d$ -Blocked paths

## Definition

Let  $\mathcal{G}$  be a directed mixed graph. Given a path  $q$  on  $\mathcal{G}$ , and a set of nodes  $S$ , we say that  $S$   **$d$ -blocks**  $q$  if  $q$  contains

- a non-collider which is in  $S$ , or
- a collider which is *not* an ancestor of  $S$ .

## Example



$X_3 \rightarrow X_5 \leftarrow X_4$  is  $d$ -blocked by  $\emptyset$ .

$X_3 \rightarrow X_5 \leftarrow X_4$  is  $d$ -blocked by  $\{X_1\}$ .

$X_3 \rightarrow X_5 \leftarrow X_4$  is not  $d$ -blocked by  $\{X_5\}$ .

$X_3 \leftarrow X_2 \leftrightarrow X_1 \rightarrow X_4$  is  $d$ -blocked by  $\{X_1\}$ .

$X_3 \leftarrow X_2 \leftrightarrow X_1 \rightarrow X_4$  is not  $d$ -blocked by  $\{X_5\}$ .

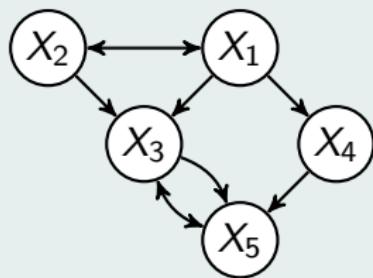
## Definition ( $d$ -separation)

Let  $\mathcal{G}$  be a directed mixed graph. For three sets  $X, Y, Z$  of nodes in  $\mathcal{G}$ , we say that  $X$  and  $Y$  are  $d$ -separated by  $Z$  iff all paths between a node in  $X$  and a node in  $Y$  are  $d$ -blocked by  $Z$ , and write  $X \perp_{\mathcal{G}} Y | Z$ .

## Definition ( $d$ -separation)

Let  $\mathcal{G}$  be a directed mixed graph. For three sets  $X, Y, Z$  of nodes in  $\mathcal{G}$ , we say that  $X$  and  $Y$  are  $d$ -separated by  $Z$  iff all paths between a node in  $X$  and a node in  $Y$  are  $d$ -blocked by  $Z$ , and write  $X \perp_{\mathcal{G}} Y | Z$ .

## Example



$X_3$  and  $X_4$  are  $d$ -separated by  $\{X_1\}$ .

$X_3$  and  $X_4$  are  $d$ -separated by  $\{X_1, X_2\}$ .

$X_3$  and  $X_4$  are not  $d$ -separated by  $\emptyset$ .

$X_3$  and  $X_4$  are not  $d$ -separated by  $\{X_1, X_5\}$ .

Please make Exercise 1.2

## Theorem

For an *acyclic* SCM, the following Global Markov Property holds:

$$X, Y \perp\!\!\!\perp Z \underset{\mathcal{G}(\mathcal{M})}{\implies} X \perp\!\!\!\perp Y | Z$$

for all subsets  $X, Y, Z$  of nodes.

In words: every d-separation in the graph  $\mathcal{G}(\mathcal{M})$  of  $\mathcal{M}$  implies a (conditional) independence in the (unique) observational distribution associated to  $\mathcal{M}$ .

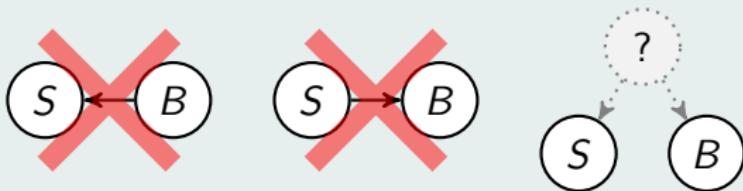
For *cyclic* SCMs, the notion of d-separation is too strong in general. A weaker notion called  *$\sigma$ -separation* has to be used instead [Forré and Mooij, 2017]. For simple SCMs, a global Markov condition using  $\sigma$ -separation can then be shown to hold.

## Reichenbach's Principle of Common Cause

The dependence  $X \not\perp\!\!\!\perp Y$  implies that  $X \rightarrow Y$ ,  $Y \rightarrow X$ , or  $X \leftrightarrow Y$  (or any combination of these three).

### Example

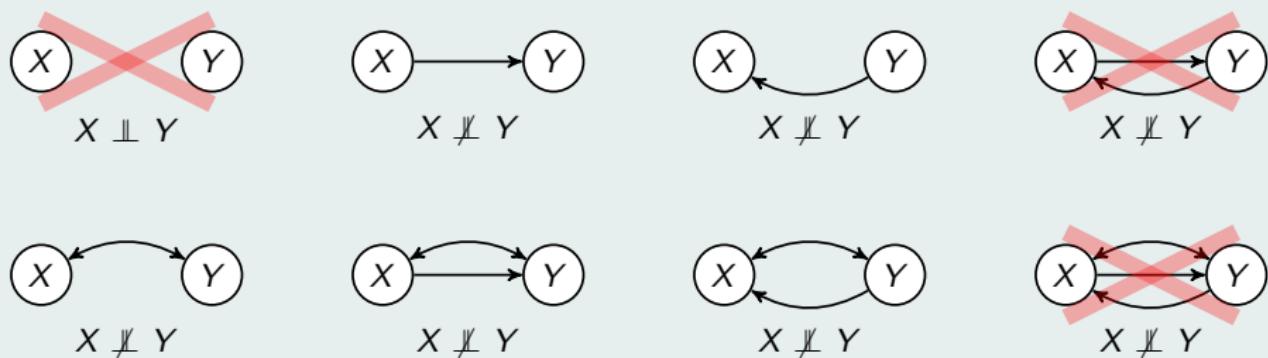
- Significant correlation ( $p = 0.008$ ) between human birth rate and number of stork populations in European countries [Matthews, 2000]
- Most people nowadays do not believe that storks deliver babies (nor that babies deliver storks)
- There must be some confounder explaining the correlation



# Proof of Reichenbach's Principle

Assuming that  $p(X, Y)$  is generated by an acyclic SCM, we can easily prove Reichenbach's Principle by applying the Global Markov property:

## Proof



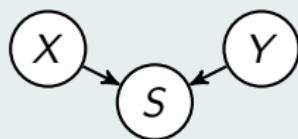
(The proof can be extended to include the cyclic case)

Reichenbach's Principle may fail in case of *selection bias*.

## Definition

If a data set is obtained by only including samples conditional on some event, **selection bias** may be introduced.

## Example



$X$ : the battery is charged

$Y$ : the start engine is operational

$S$ : the car starts

- A car mechanic (who only observes cars for which  $S = 0$ ) will observe a dependence between  $X$  and  $Y$ :  $X \not\perp\!\!\!\perp Y | S$ .
- When the car mechanic invokes Reichenbach's Principle without realizing that he is selecting on the value of  $S$  (maybe  $S$  is a latent variable), a wrong conclusion will be drawn.

# Faithfulness Assumption

Let  $\mathcal{M}$  be an **acyclic** SCM.

We have seen that the *Global Markov Property* holds:

$$\mathbf{X}, \mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \quad \underset{\mathcal{G}(\mathcal{M})}{\Rightarrow} \quad \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$$

for all subsets  $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$  of nodes.

## Definition (Faithfulness Assumption)

For all subsets  $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$  of nodes,

$$\mathbf{X}, \mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \quad \underset{\mathcal{G}(\mathcal{M})}{\Leftarrow} \quad \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$$

**Note:** Faithfulness holds **generically**, i.e., up to measure-zero sets of parameters [Meek, 1995]. In other words, SCM parameters need to be *carefully tuned* in order to violate the faithfulness assumption.

# Faithfulness Violations

Faithfulness violations may occur e.g. in case of parameter cancellations or deterministic relations.

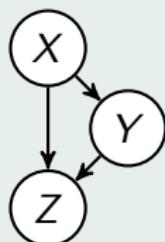
## Example (Parameter cancellation)

Consider an SCM  $\mathcal{M}$ :

$$X = E_X$$

$$Y = X + E_Y$$

$$Z = X - Y + E_Z$$



Then:

$Z \perp\!\!\!\perp_{p_{\mathcal{M}}} X$  but  $Z \not\perp\!\!\!\perp_{\mathcal{G}(\mathcal{M})} X$ .

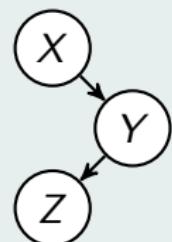
## Example (Deterministic relation)

Consider an SCM  $\mathcal{M}$ :

$$X = E_X$$

$$Y = X$$

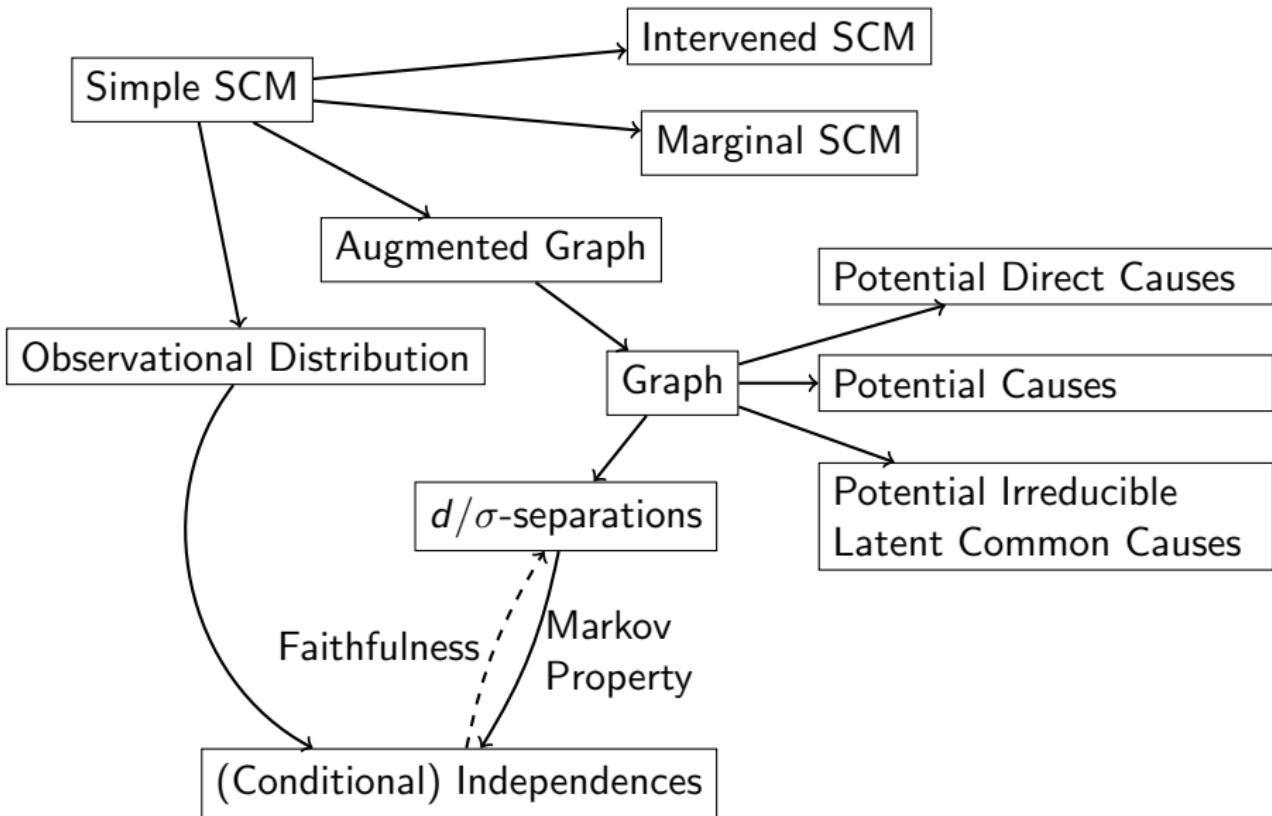
$$Z = Y + E_Z$$



Then:

$Z \perp\!\!\!\perp_{p_{\mathcal{M}}} Y | X$  but  $Z \not\perp\!\!\!\perp_{\mathcal{G}(\mathcal{M})} Y | X$ .

# Simple SCMs: Overview



- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Causal Inference: Predicting Causal Effects

One important task (“*causal inference*”) is the prediction of causal effects.

## Definition

The **causal effect** of  $X$  on  $Y$  is defined as  $p(y \mid \text{do}(X = x))$ .

Special cases:

- $X$  binary:  $\mathbb{E}(Y \mid \text{do}(X = 1)) - \mathbb{E}(Y \mid \text{do}(X = 0))$
- $X, Y$  linearly related:  $\frac{\partial}{\partial x} \mathbb{E}(Y \mid \text{do}(X = x))$

# Causal Inference: Predicting Causal Effects

One important task (“*causal inference*”) is the prediction of causal effects.

## Definition

The **causal effect** of  $X$  on  $Y$  is defined as  $p(y | \text{do}(X = x))$ .

Special cases:

- $X$  binary:  $\mathbb{E}(Y | \text{do}(X = 1)) - \mathbb{E}(Y | \text{do}(X = 0))$
- $X, Y$  linearly related:  $\frac{\partial}{\partial x} \mathbb{E}(Y | \text{do}(X = x))$

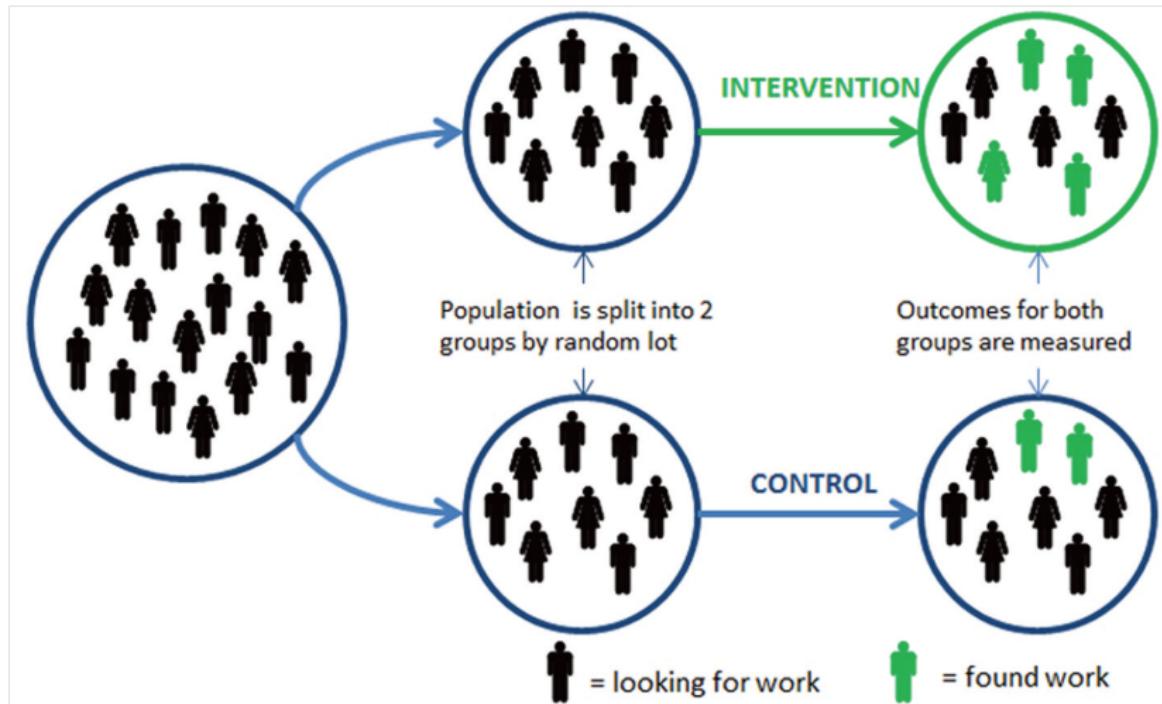
**Note:** In general, since  $p(y | \text{do}(X = x)) \neq p(y | X = x)$ , we cannot use standard supervised learning (regression, classification) for this task.

Two approaches can be used:

- Experimentation (Randomized Controlled Trials, A/B-testing)
- Apply the Back-door Criterion (if causal graph is known)

# Causal discovery by experimentation

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



# Identifiability: Example

If we cannot do experiments... Can we express  $p(y | \text{do}(X = x))$  in terms of the observational distribution?

## Example



$$p(y | \text{do}(X = x))$$

=

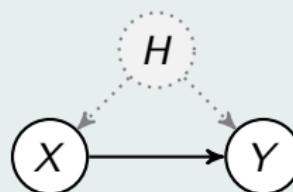
$$p(y | X = x)$$

Yes!

# Identifiability: Example

If we cannot do experiments... Can we express  $p(y | \text{do}(X = x))$  in terms of the observational distribution?

## Example



$$p(y | \text{do}(X = x))$$

=

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

$$p(y | \text{do}(X = x)) = \int p(h)p(y | x, h) dh$$

≠

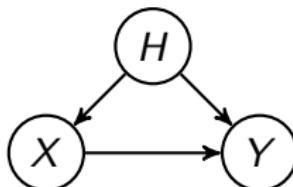
$$p(y | X = x) = \int p(h | x)p(y | x, h) dh$$

Yes!

No!

# Adjustment for covariates

We have seen that for the following causal graph,



adjusting for the confounder  $H$ , yields the causal effect of  $X$  on  $Y$ :

$$\int p(h)p(y | x, h) dh = p(y | \text{do}(X = x))$$

More generally, given a causal graph: which covariates  $\mathbf{H}$  could we adjust for in order to express the causal effect of  $X$  on  $Y$  in terms of the observational distribution?

A sufficient condition is given by the **Back-door Criterion**.

# The Back-Door Criterion

Theorem (Back-Door Criterion [Pearl, 2000])

For an *acyclic* SCM, nodes  $X$ ,  $Y$  and set of nodes  $\mathbf{H}$ : if

- ①  $X, Y \notin \mathbf{H}$ ;
- ②  $X$  is not an ancestor of any node in  $\mathbf{H}$  in  $\mathcal{G}(\mathcal{M})$ ;
- ③  $\mathbf{H}$  *d-blocks* all *back-door paths*  $X \leftarrow \dots Y$  and  $X \leftrightarrow \dots Y$  in  $\mathcal{G}(\mathcal{M})$   
(i.e., all paths between  $X$  and  $Y$  that start with an arrowhead at  $X$ ).

then the causal effect of  $X$  on  $Y$  can be obtained by adjusting for  $H$ :

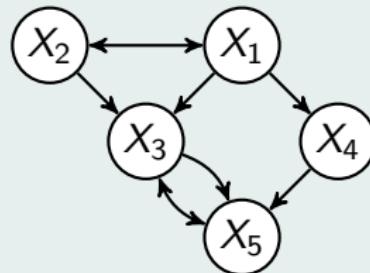
$$p(y \mid \text{do}(X = x)) = \int p(y \mid x, \mathbf{h}) p(\mathbf{h}) d\mathbf{h} \left( = \sum_{\mathbf{h}} p(y \mid x, \mathbf{h}) p(\mathbf{h}) \right).$$

For the special case  $\mathbf{H} = \emptyset$ , this should be read as:

$$p(y \mid \text{do}(X = x)) = p(y \mid x).$$

# The Back-door Criterion: Example

## Example



The sets of variables that are admissible for adjustment to get the causal effect of  $X_2$  on  $X_5$  are:  $\{X_1\}$ ,  $\{X_1, X_4\}$ . Therefore:

$$\begin{aligned} p(x_5 \mid \text{do}(X_2 = x_2)) &= \int p(x_5 \mid x_1, x_2) p(x_1) dx_1 \\ &= \int p(x_5 \mid x_1, x_2, x_4) p(x_1, x_4) dx_1 dx_4 \end{aligned}$$

Some sets of variables that are *not* admissible for adjustment to get the causal effect of  $X_2$  on  $X_5$  are:  $\{X_3\}$ ,  $\{X_1, X_3\}$ .

Please make Exercise 3

# Simpson's Paradox

Remember Simpson's paradox:

## Example (Simpson's paradox)

We collect electronic patient records to investigate the effectiveness of a new drug against a certain disease. We find that:

- ① The probability of recovery is higher for patients that took the drug:

$$p(\text{recovery} \mid \text{drug}) > p(\text{recovery} \mid \text{no drug})$$

- ② For **both male and female** patients, the relation is **opposite**:

$$p(\text{recovery} \mid \text{drug, male}) < p(\text{recovery} \mid \text{no drug, male})$$

$$p(\text{recovery} \mid \text{drug, female}) < p(\text{recovery} \mid \text{no drug, female})$$

Does the drug cause recovery? I.e., would *you* use this drug if you are ill?

The answer depends on the causal relationships between the variables!

# Resolving Simpson's paradox

The crux to resolving Simpson's paradox is to realize:

Seeing  $\neq$  doing

- $p(R = 1 | D = 1)$ : the probability that somebody recovers, given the observation that the person took the drug.
- $p(R = 1 | \text{do}(D = 1))$ : the probability that somebody recovers, if we force the person to take the drug.

Simpson's paradox only manifests itself if we misinterpret correlation as causation by identifying  $p(r | D = d)$  with  $p(r | \text{do}(D = d))$ .

We should prescribe the drug if

$$p(R = 1 | \text{do}(D = 1)) > p(R = 1 | \text{do}(D=0)).$$

How to find the causal effect of the drug on recovery?

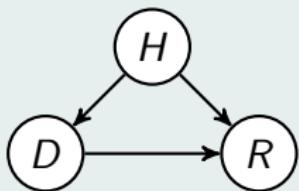
- ① Randomized Controlled Trials
- ② Back-Door Criterion (requires knowledge of causal graph)

# Exercise 4

Please make Exercise 4

# Back-Door Criterion for Simpson's paradox

## Example (Scenario 1)



$R$ : Recovery  
 $D$ : Took drug  
 $H$ : Gender

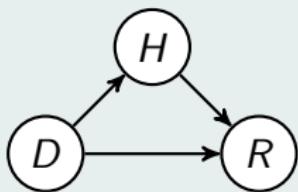
- There is one back-door path:  $D \leftarrow H \rightarrow R$ , which is  $d$ -blocked by  $\{H\}$ .
- $D$  is not an ancestor of  $H$ .
- Therefore, adjust for  $\{H\}$  to obtain causal effect of drug on recovery:

$$p(r \mid \text{do}(D = d)) = \sum_h p(r \mid D = d, H = h)p(h)$$

- So in scenario I, you should **not** take the drug: for both males and females, taking the drug lowers the probability of recovery.

# Back-Door Criterion for Simpson's paradox

## Example (Scenario 2)



$R$ : Recovery  
 $D$ : Took drug  
 $H$ : Gender

- There are no back-door paths.
- $D$  is an ancestor of  $H$ .
- Do **not** adjust for  $\{H\}$  to obtain causal effect of drug on recovery:

$$p(r \mid \text{do}(D = d)) = p(r \mid D = d)$$

- So in scenario II, you **should** take the drug: in the general population, taking the drug increases the probability of recovery.

(If you think gender-changing drugs are unlikely, replace “gender” by “high/low blood pressure”, for example).

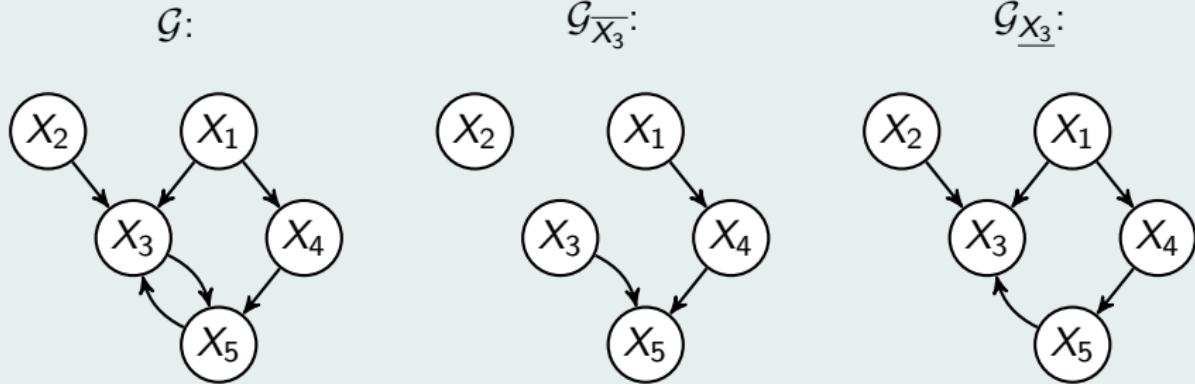
# Mutilated graphs

## Definition

Given a DMG  $\mathcal{G}$  and a subset  $X$  of nodes in  $\mathcal{G}$ , we define

- $\mathcal{G}_{\overline{X}}$  to be  $\mathcal{G}$  without the incoming edges on nodes in  $X$ ;
- $\mathcal{G}_{\underline{X}}$  to be  $\mathcal{G}$  without the outgoing edges from nodes in  $X$ .

## Example



# Do-calculus [Pearl, 2000]

Pearl formulated three rules (the “**do-calculus**”) that can be used in addition to the usual rules for probabilistic reasoning:

## ① Ignoring observations:

$$p(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w}, \mathbf{z}) = p(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w}) \quad \text{if } \mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}$$

## ② Action/observation exchange:

$$p(\mathbf{y} \mid \text{do}(\mathbf{x}), \text{do}(\mathbf{z}), \mathbf{w}) = p(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}) \quad \text{if } \mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}$$

## ③ Ignoring actions:

$$p(\mathbf{y} \mid \text{do}(\mathbf{x}), \text{do}(\mathbf{z}), \mathbf{w}) = p(\mathbf{y} \mid \text{do}(\mathbf{x}), \mathbf{w}) \quad \text{if } \mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W}$$

where  $\mathbf{Z}(\mathbf{W}) = \mathbf{Z} \setminus \text{An}_{\mathcal{G}_{\overline{\mathbf{X}}}}(\mathbf{W})$ .

The do-calculus allows us to reason with (probabilistic) causal statements, given (partial) knowledge of the causal structure. These rules are more powerful than the Back-door Criterion for causal prediction purposes.

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Causal Discovery

We have seen how to perform causal reasoning, given the causal model.  
But how do we get the causal model in the first place?

Establishing causal relations from data (“causal discovery”) is one of the fundamental tasks in science.

# Causal Discovery

We have seen how to perform causal reasoning, given the causal model.  
But how do we get the causal model in the first place?

Establishing causal relations from data (“causal discovery”) is one of the fundamental tasks in science.

Since the pioneering work by Peirce and Fisher, the gold standard for causal discovery is a **randomized, controlled experiment**.



# Causal Discovery

We have seen how to perform causal reasoning, given the causal model.  
But how do we get the causal model in the first place?

Establishing causal relations from data (“causal discovery”) is one of the fundamental tasks in science.

Since the pioneering work by Peirce and Fisher, the gold standard for causal discovery is a **randomized, controlled experiment**.



More recently, causal discovery methods from **purely observational** data have been developed, starting with the work of Spirtes, Glymour, Scheines, Pearl and others.



# Causal Discovery

We have seen how to perform causal reasoning, given the causal model.  
But how do we get the causal model in the first place?

Establishing causal relations from data (“causal discovery”) is one of the fundamental tasks in science.

Since the pioneering work by Peirce and Fisher, the gold standard for causal discovery is a **randomized, controlled experiment**.



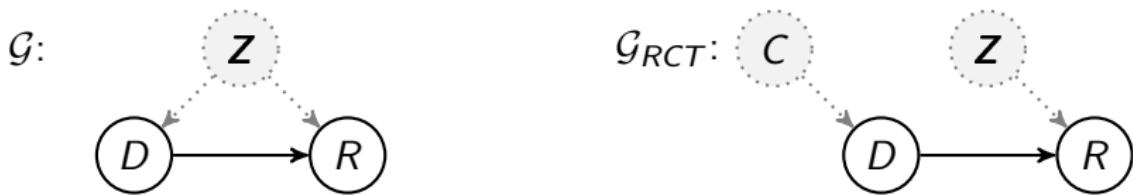
More recently, causal discovery methods from **purely observational** data have been developed, starting with the work of Spirtes, Glymour, Scheines, Pearl and others.



These ideas have inspired causal discovery methods that combine observational and interventional data in various ways.

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Randomized Controlled Trials [Fisher, 1935]

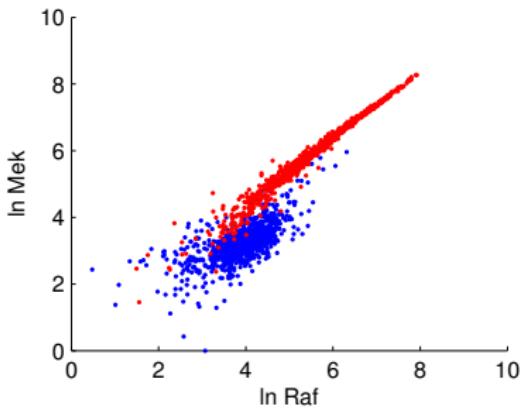


$R$ : Recovery,  $D$ : Drug,  $Z$ : latent confounders (e.g., genetics),  $C$ : coin flip.

- Divide patients into two groups: **treatment** and **control** randomly (e.g., by a coin flip).
- Patients in the treatment group are forced to take a drug, and patients in the control group are forced to not take the drug (but to take a placebo instead):  $D = C$ .
- Estimating the causal effect of the drug now becomes a standard statistical exercise, as  $p(R | D = C) = p(R | \text{do}(D = C))$ .
- The RCT intervention breaks any back-door paths, if existent.

All *evidence-based* medicine is based on this idea.

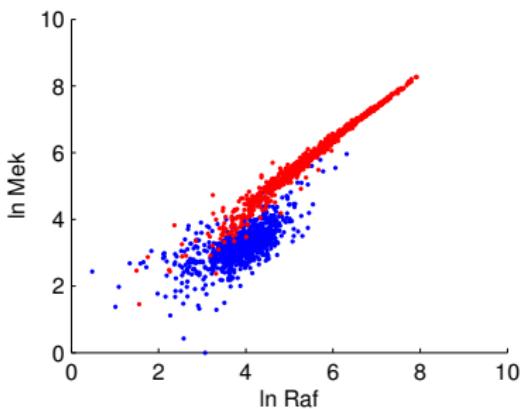
# Causal Discovery by Experimentation: Example



- Each dot is a measurement in a single human immune system cell
- Raf: abundance of phosphorylated Raf
- Mek: abundance of phosphorylated Mek
- blue = baseline,  
red = reagent U0126 added

**Question: What is the causal relation between Raf and Mek?**

# Causal Discovery by Experimentation: Example

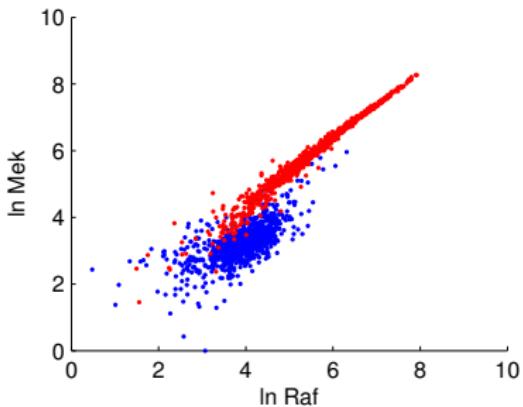


- Each dot is a measurement in a single human immune system cell
- Raf: abundance of phosphorylated Raf
- Mek: abundance of phosphorylated Mek
- blue = baseline,  
red = reagent U0126 added

**Question: What is the causal relation between Raf and Mek?**

*Hint: U0126 inhibits Raf.*

# Causal Discovery by Experimentation: Example



- Each dot is a measurement in a single human immune system cell
- Raf: abundance of phosphorylated Raf
- Mek: abundance of phosphorylated Mek
- blue = baseline,  
red = reagent U0126 added

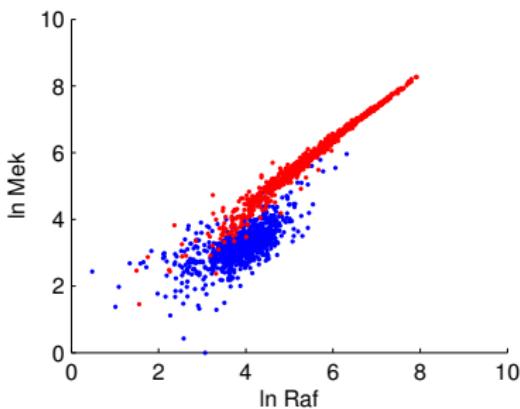
**Question: What is the causal relation between Raf and Mek?**

*Hint: U0126 inhibits Mek.*

**Answer: Mek causes Raf**

*(Changing activity of Mek changes abundance of Raf.)*

# Causal Discovery by Experimentation: Example



- Each dot is a measurement in a single human immune system cell
- Raf: abundance of phosphorylated Raf
- Mek: abundance of phosphorylated Mek
- blue = baseline,  
red = reagent U0126 added

**Question: What is the causal relation between Raf and Mek?**

*Hint: U0126 inhibits Mek.*

**Answer: Mek causes Raf**

*(Changing activity of Mek changes abundance of Raf.)*

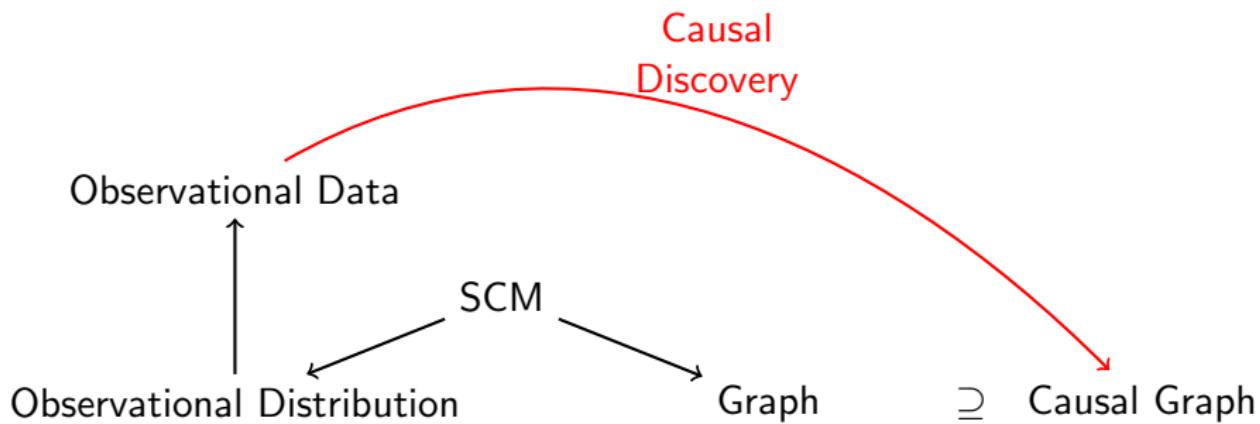
**Note:** How did we know that “U0126 inhibits Mek” in the first place?

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Causal Discovery from Observational Data

Experiments can be expensive, time-consuming, unethical, impractical or even infeasible.

Intriguing alternative: causal discovery from **purely observational data** [Spirtes et al., 2000, Pearl, 2000]!



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

## Conditional-independence constraint-based

*Independence patterns in the data constrain the possible causal graphs.*

- LCD (Cooper, 1997)
- Y-Structures (Mani & Cooper, 2004)
- PC (Spirtes & Gleimour & Scheines, 2000), IC (Pearl, 2000)
- FCI (Spirtes & Meek & Richardson, 1995; Zhang, 2008)
- ...

## General constraint-based

*Similar, but exploiting more general types of constraints in the data.*

- Verma constraints (Robins (1986), Verma & Pearl (1990), Tian & Pearl (2002))
- Nested Markov Models (Richardson, Evans, Robins, Shpitser (2017))
- Algebraic Constraints (Van Ommen & Mooij (2017))
- ...

## Likelihood-based approaches

*Score penalized likelihoods of possible causal graphs and select the best one(s).*

- Bayesian Network Learning (Heckerman, Geiger, Chickering, 1995)
- Greedy Equivalence Search (Chickering, 2002)
- ...

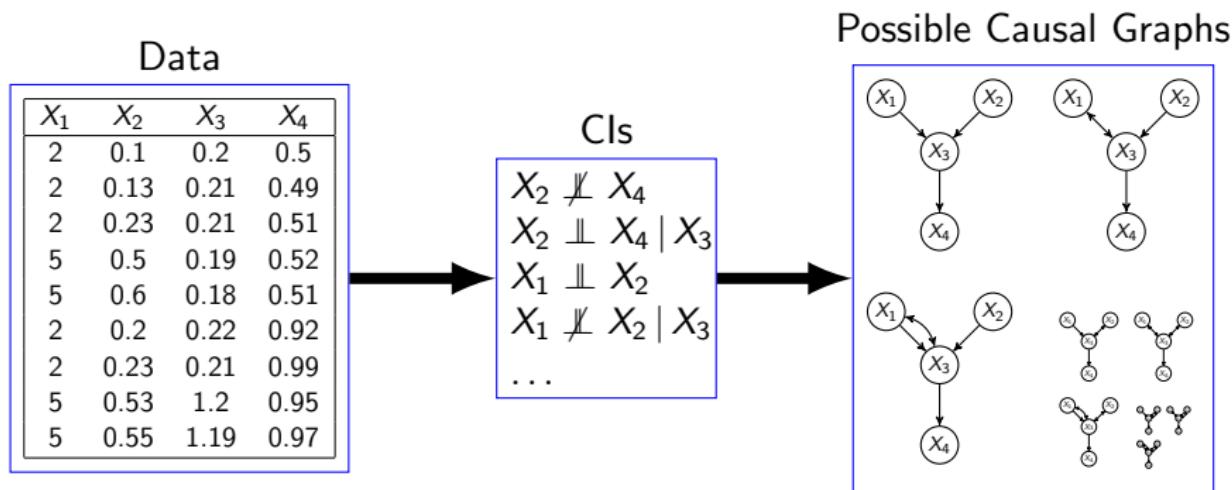
## Restrictions on functional causal relations and noise distributions

*Minimize the “complexity” of causal models.*

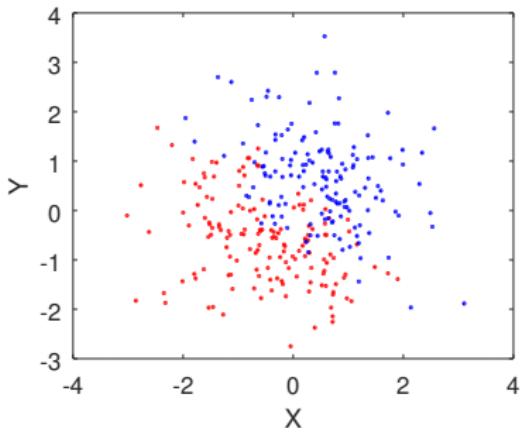
- LINGAM (Kano, Shimizu, 2003; Shimizu *et al.*, 2006)
- Additive Noise Models (Hoyer *et al.*, 2006)
- Post-Nonlinear Model (Zhang & Hyvärinen, 2009)
- ...

# 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 et al., 2000].



# Causal Discovery from Observational Data: V-Structure

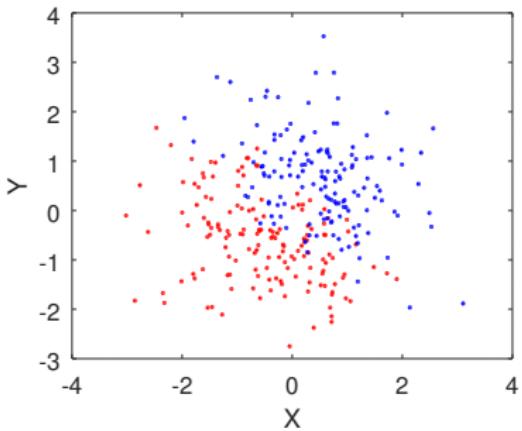


blue:  $Z = 0$ , red:  $Z = 1$

$$\begin{aligned} X \perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Y | Z, \\ X \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Z | Y, \\ Y \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z | X. \end{aligned}$$

**Question:** What is the causal relation between  $X$ ,  $Y$  and  $Z$ ?

# Causal Discovery from Observational Data: V-Structure

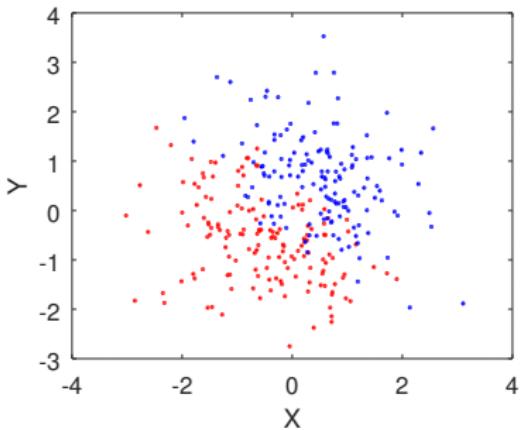


blue:  $Z = 0$ , red:  $Z = 1$

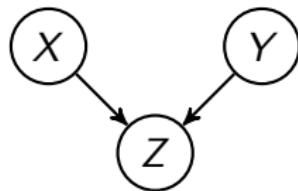
$$\begin{aligned} X \perp\!\!\!\perp Y, X \not\perp\!\!\!\perp Y | Z, \\ X \not\perp\!\!\!\perp Z, X \not\perp\!\!\!\perp Z | Y, \\ Y \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z | X. \end{aligned}$$

**Question:** What is the causal relation between  $X$ ,  $Y$  and  $Z$ ?

*Hint: Assume an acyclic, faithful SCM without latent confounders generated the data, and assume no selection bias or measurement error*



blue:  $Z = 0$ , red:  $Z = 1$



$$\begin{aligned} X \perp\!\!\!\perp Y, \quad X \not\perp\!\!\!\perp Y \mid Z, \\ X \not\perp\!\!\!\perp Z, \quad X \not\perp\!\!\!\perp Z \mid Y, \\ Y \not\perp\!\!\!\perp Z, \quad Y \not\perp\!\!\!\perp Z \mid X. \end{aligned}$$

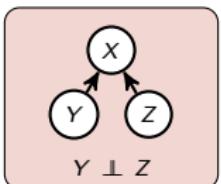
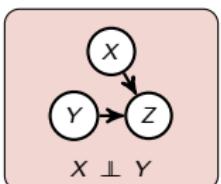
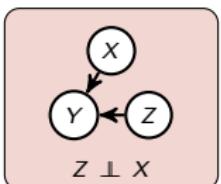
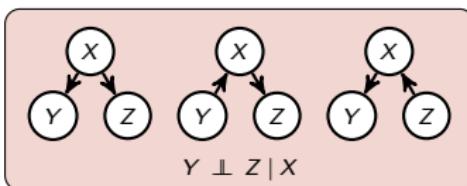
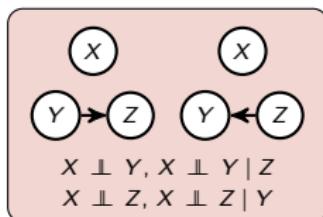
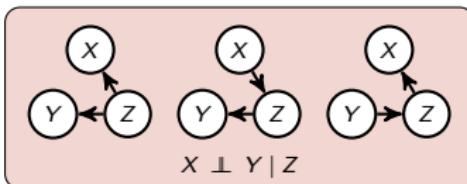
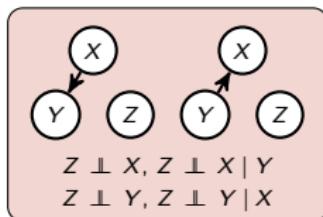
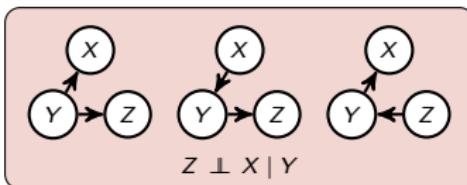
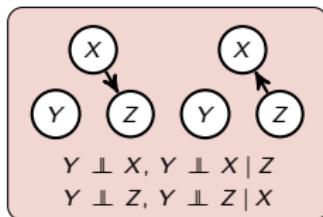
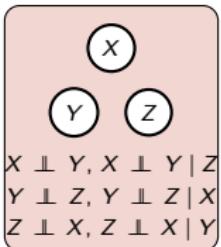
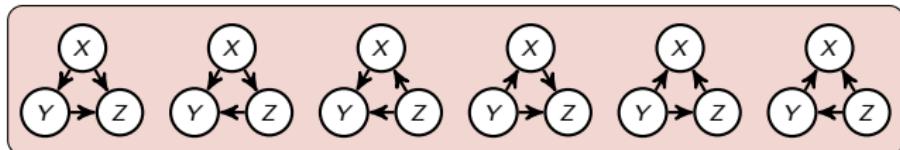
**Question:** What is the causal relation between  $X$ ,  $Y$  and  $Z$ ?

*Hint: Assume an acyclic, faithful SCM without latent confounders generated the data, and assume no selection bias or measurement error*

**Answer:**  $X$  causes  $Z$ ;  $Y$  causes  $Z$ ;  $X$  and  $Y$  causally unrelated

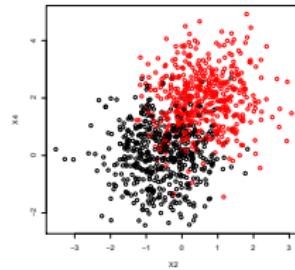
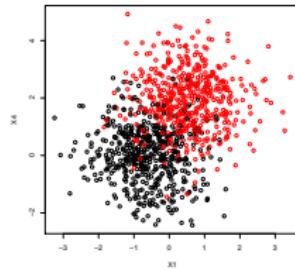
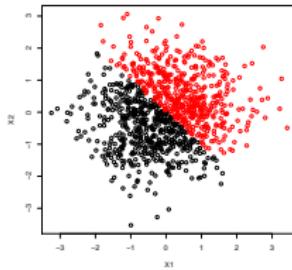
**Note:** Strong assumptions, but no experiments needed!

# Markov equivalence classes for three variables



Please make Exercise 5

# Causal Discovery from Observational Data: Y-Structure



$$\begin{aligned} X_1 &\perp\!\!\!\perp X_2 \\ X_1 &\not\perp\!\!\!\not X_2 \mid X_3 \end{aligned}$$

$$\begin{aligned} X_1 &\not\perp\!\!\!\not X_4 \\ X_1 &\perp\!\!\!\perp X_4 \mid X_3 \end{aligned}$$

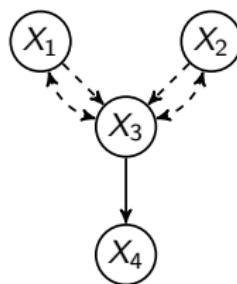
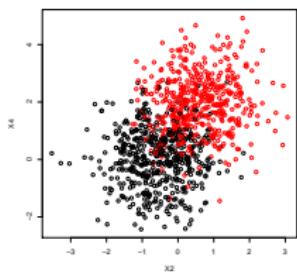
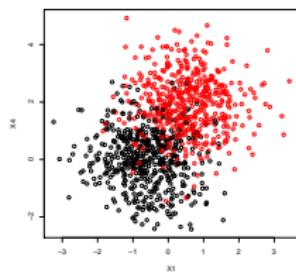
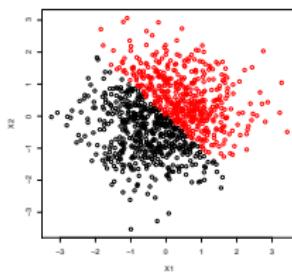
$$\begin{aligned} X_2 &\not\perp\!\!\!\not X_4 \\ X_2 &\perp\!\!\!\perp X_4 \mid X_3 \end{aligned}$$

black:  $X_3 = 0$ , red:  $X_3 = 1$

**Question: What is the causal relation between  $X_3$  and  $X_4$ ?**

*Hint: Assume an acyclic, faithful SCM generated the data, and assume no selection bias or measurement error.*

# Causal Discovery from Observational Data: Y-Structure



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

$$X_1 \not\perp\!\!\!\perp X_4 \\ X_1 \perp\!\!\!\perp X_4 | X_3$$

$$X_2 \not\perp\!\!\!\perp X_4 \\ X_2 \perp\!\!\!\perp X_4 | X_3$$

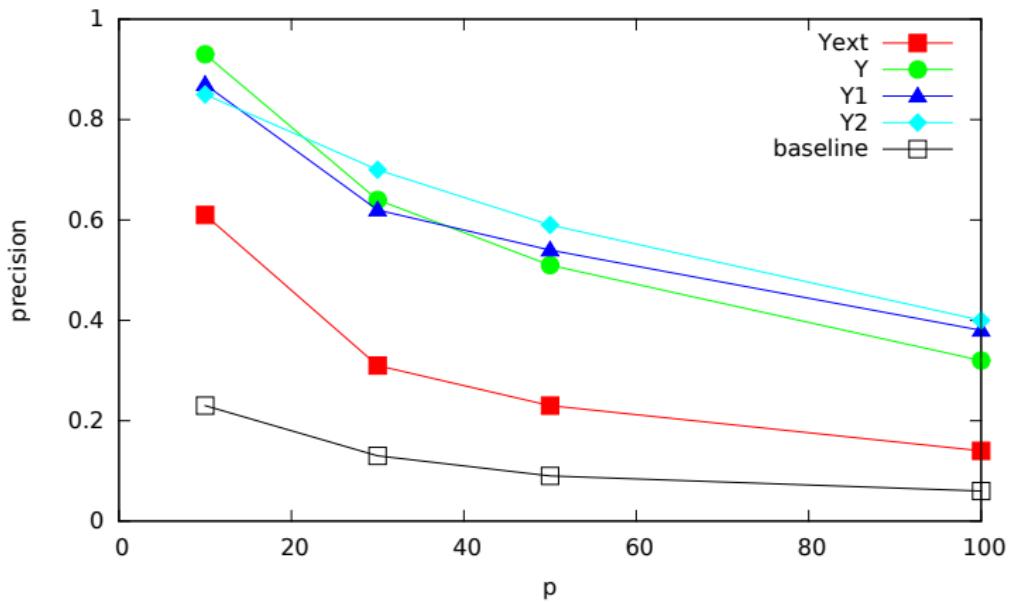
black:  $X_3 = 0$ , red:  $X_3 = 1$

**Question: What is the causal relation between  $X_3$  and  $X_4$ ?**

*Hint: Assume an acyclic, faithful SCM generated the data, and assume no selection bias or measurement error.*

**Answer:  $X_3$  causes  $X_4$  and they are not confounded. Hence, the causal effect of  $X_3$  on  $X_4$  satisfies  $p(x_4 | \text{do}(X_3 = x_3)) = p(x_4 | x_3)$ .**

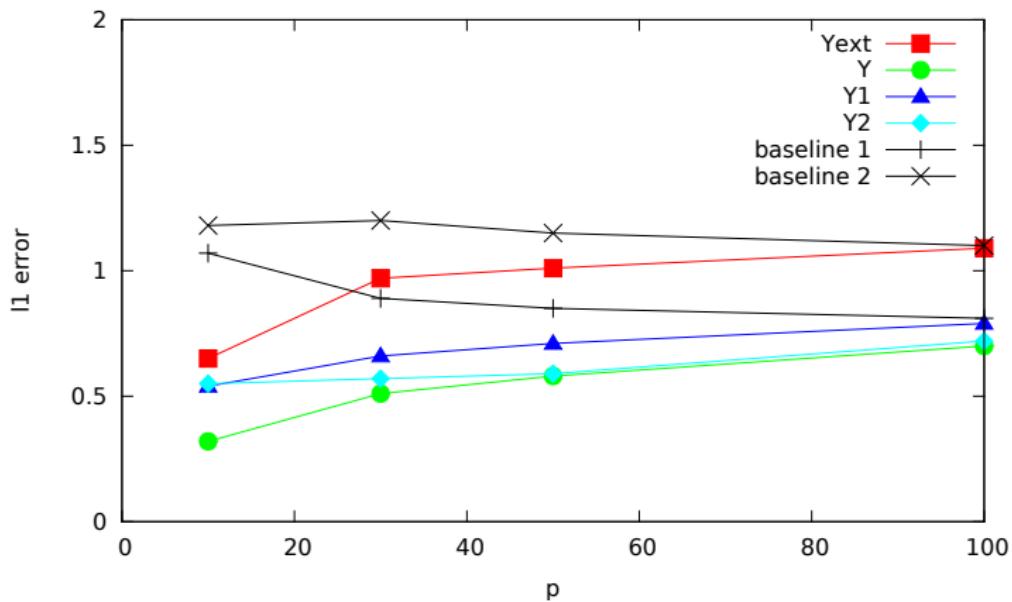
Precision of prediction  $X$  causes  $Y$ :



Baseline: random guessing

# Y-structures: Empirical Performance II

Causal prediction error for  $\mathbb{E}(Y | \text{do}(X = x))$ :



Baseline 1:  $p(y | \text{do}(X = x)) = p(y)$ , Baseline 2:  $p(y | \text{do}(X = x)) = p(y|x)$

# Hardness of Causal Discovery

$d$	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	332771901227107591736177573311261125883583076258421902583546773505

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.

# (Augmented) FCI

[Spirtes *et al.*, 2000, Spirtes *et al.*, 1999, Ali *et al.*, 2005, Zhang, 2008]

- $\mathcal{R}0a$  If  $X \perp\!\!\!\perp Y | \mathbf{Z}$ , then  $X \not\asymp Y$ ,  $Sep(X, Y) \leftarrow \mathbf{Z}$ .
- $\mathcal{R}0b$  If  $X *-* Z \circ-* Y$  and  $X \not\asymp Y$ , then if  $Z \notin Sep(X, Y)$ , then  $X *-\rightarrow Z \leftarrow-* Y$ .
- $\mathcal{R}1$  If  $X *-\rightarrow Z \circ-* Y$ , and  $X \not\asymp Y$ , then  $Z \rightarrow Y$ .
- $\mathcal{R}2a$  If  $Z \rightarrow X *-\rightarrow Y$  and  $Z *-\circ Y$ , then  $Z *-\rightarrow Y$ .
- $\mathcal{R}2b$  If  $Z *-\rightarrow X \rightarrow Y$  and  $Z *-\circ Y$ , then  $Z *-\rightarrow Y$ .
- $\mathcal{R}3$  If  $X *-\rightarrow Z \leftarrow-* Y$ ,  $X *-\circ W \circ-* Y$ ,  $X \not\asymp Y$ , and  $W *-\circ Z$ , then  $W *-\rightarrow Z$ .
- $\mathcal{R}4a$  If  $u = \langle X, \dots, Z_k, Z, Y \rangle$  is a discriminating path between  $X$  and  $Y$  for  $Z$ , and  $Z \circ-* Y$ , then if  $Z \in Sep(X, Y)$ , then  $Z \rightarrow Y$ .
- $\mathcal{R}4b$  Idem, if  $Z \notin Sep(X, Y)$  then  $Z_k \leftrightarrow Z \leftrightarrow Y$ .
- $\mathcal{R}5$  If  $u = \langle Z, X, \dots, W, Y, Z, X \rangle$  is an uncov. circle path, then  $Z \rightarrow Y$  (idem for all edges on  $u$ ).
- $\mathcal{R}6$  If  $X \rightarrow Z \circ-* Y$ , then orient as  $Z *-\rightarrow Y$ .
- $\mathcal{R}7$  If  $X \circ-\rightarrow Z \circ-* Y$ , and  $X \not\asymp Y$ , then  $Z \rightarrow Y$ .
- $\mathcal{R}8a$  If  $Z \rightarrow X \rightarrow Y$  and  $Z \circ-\rightarrow Y$ , then  $Z \rightarrow Y$ .
- $\mathcal{R}8b$  If  $Z \circ-\rightarrow X \rightarrow Y$  and  $Z \circ-\rightarrow Y$ , then  $Z \rightarrow Y$ .
- $\mathcal{R}9$  If  $Z \circ-\rightarrow Y$ ,  $u = \langle Z, X, W, \dots, Y \rangle$  is an uncov. p.d. path, and  $X \not\asymp Y$ , then  $Z \rightarrow Y$ .
- $\mathcal{R}10$  If  $Z \circ-\rightarrow Y$ ,  $X \rightarrow Y \leftarrow W$ ,  $u_1 = \langle Z, S, \dots, X \rangle$  and  $u_2 = \langle Z, V, \dots, W \rangle$  are uncov. p.d. paths, (possibly with  $S = X$  and/or  $V = W$ ), then if  $S \not\asymp V$ , then  $Z \rightarrow Y$ .

**Input** : independence oracle for  $\mathbf{V}$

**Output** : complete PAG  $\mathcal{P}$  over  $\mathbf{V}$

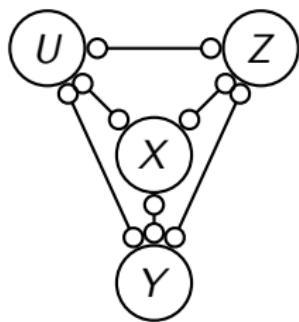
- 1:  $\mathcal{P} \leftarrow$  fully  $\circ-\circ$  connected graph over  $\mathbf{V}$
- 2: **for all**  $\{X, Y\} \in \mathbf{V}$  **do**
- 3:     search in some clever way for a  $X \perp\!\!\!\perp Y | \mathbf{Z}$
- 4:          $\mathcal{P} \leftarrow \mathcal{R}0a$  (eliminate  $X \not\asymp Y$ )
- 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]

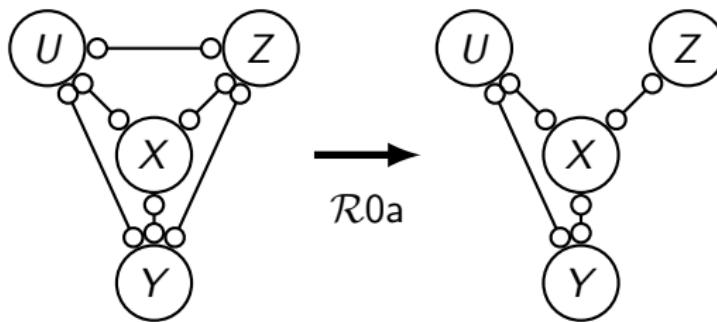
# FCI: Example (“Extended Y-structure”)

Independences:  $Z \perp\!\!\!\perp U$ ,  $Z \perp\!\!\!\perp Y | X$



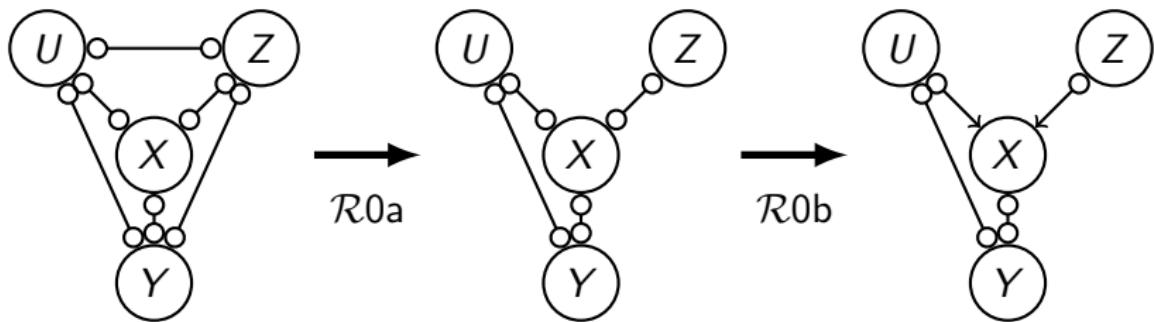
# FCI: Example (“Extended Y-structure”)

Independences:  $Z \perp\!\!\!\perp U$ ,  $Z \perp\!\!\!\perp Y | X$



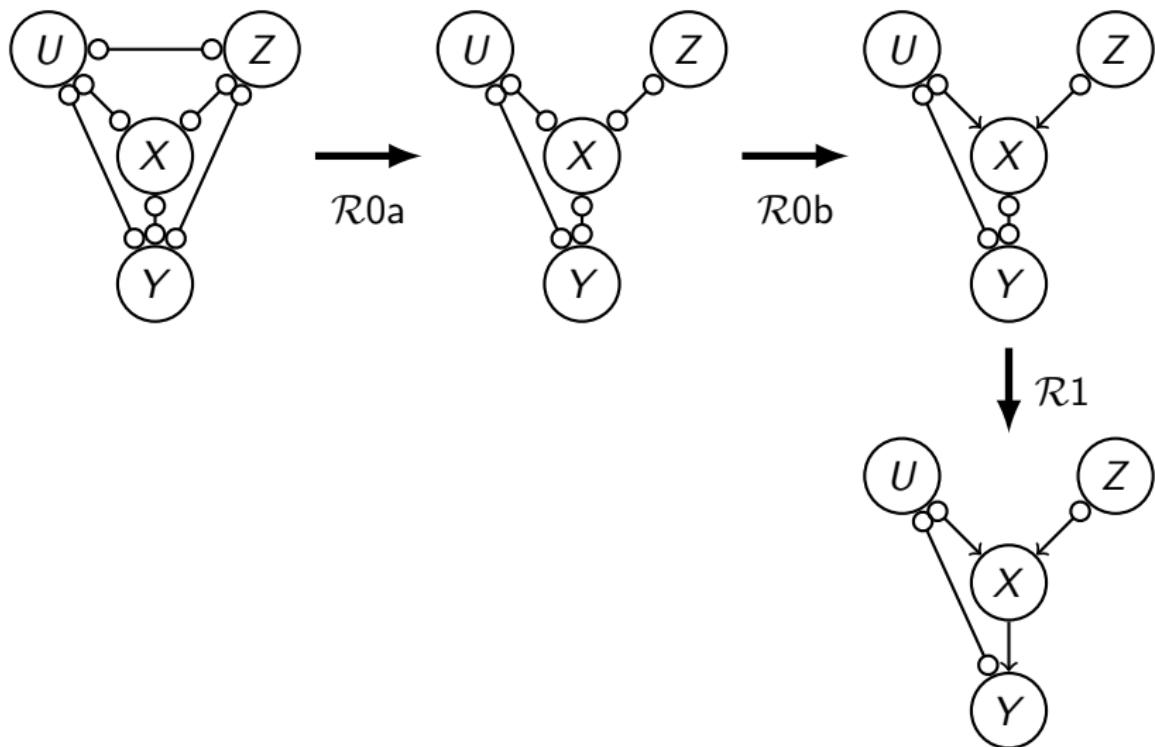
# FCI: Example (“Extended Y-structure”)

Independences:  $Z \perp\!\!\!\perp U$ ,  $Z \perp\!\!\!\perp Y | X$



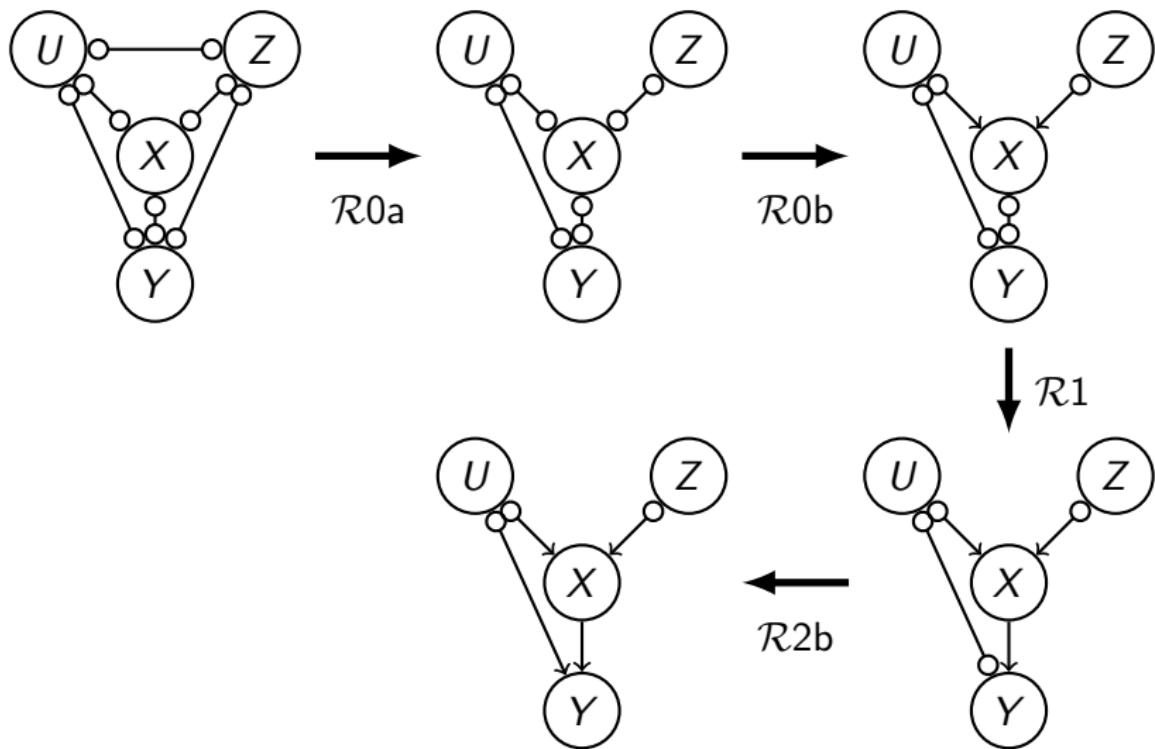
# FCI: Example (“Extended Y-structure”)

Independences:  $Z \perp\!\!\!\perp U$ ,  $Z \perp\!\!\!\perp Y | X$

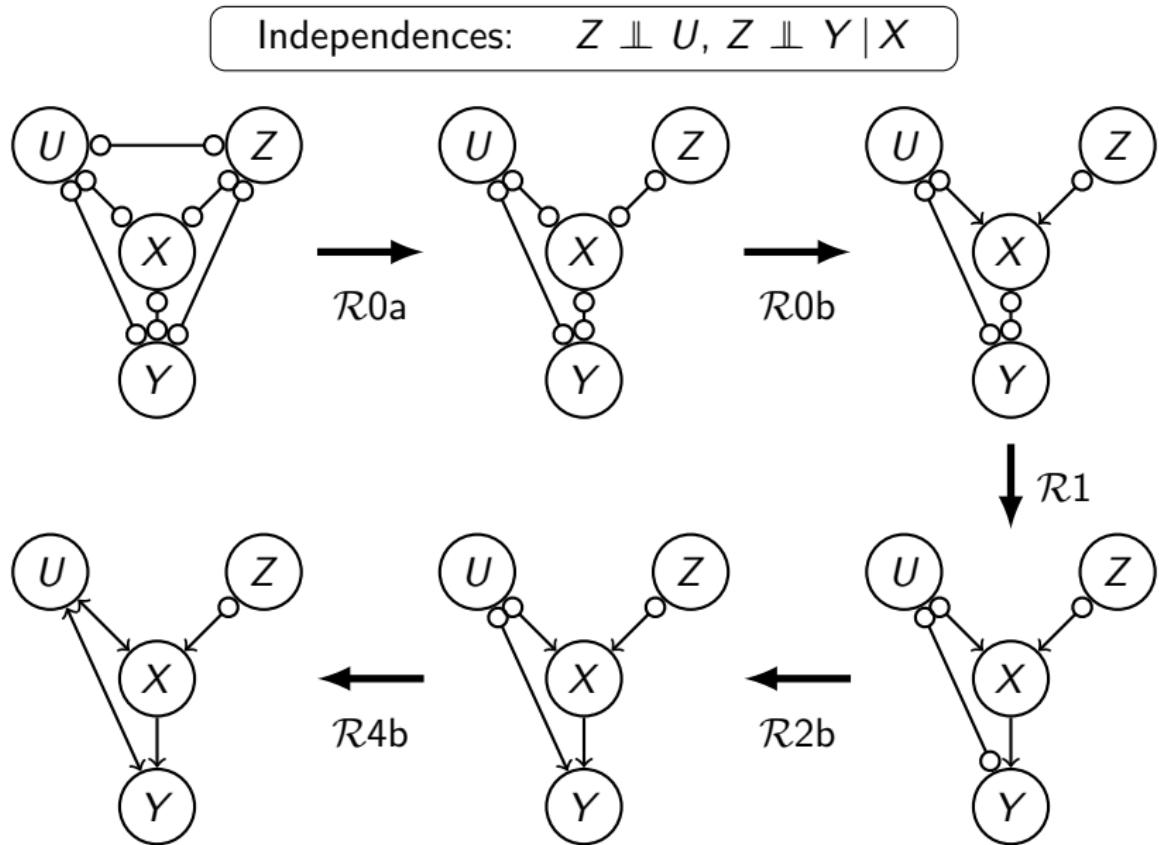


# FCI: Example (“Extended Y-structure”)

Independences:  $Z \perp\!\!\!\perp U$ ,  $Z \perp\!\!\!\perp Y | X$



# FCI: Example (“Extended Y-structure”)



# Local Causal Discovery (LCD)

Local Causal Discovery: simple causal discovery algorithm (Cooper, 1997).

## Definition

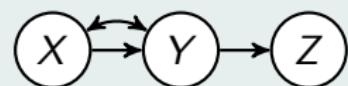
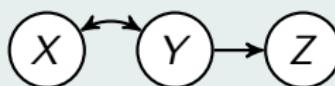
If for three variables  $X, Y, Z$ :

$$Y \notin \text{an}(X) \wedge Z \notin \text{an}(X) \wedge X \not\perp\!\!\!\perp Y \wedge Y \not\perp\!\!\!\perp Z \wedge X \perp\!\!\!\perp Z | Y,$$

then  $(X, Y, Z)$  is an LCD triplet.

## Theorem

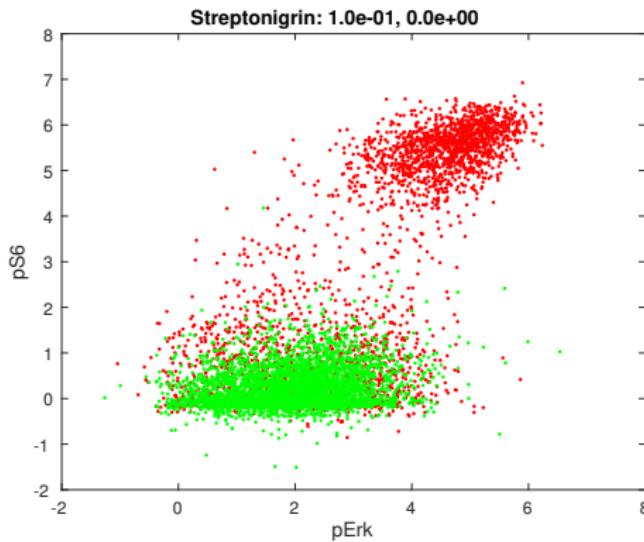
If an acyclic, faithful SCM generated the data without selection bias or measurement error, the only causal graphs that yield an LCD triplet are:



Therefore,  $Y \in \text{an}(Z)$  and  $p(Z | \text{do}(Y = y)) = p(Z | Y = y)$ .

# LCD: Example

- $p\text{Erk}$ : abundance of phosphorylated Erk in each cell
- $p\text{S6}$ : abundance of phosphorylated S6 in cell
- $I$ : green = baseline, red = PMA-IONO activator added



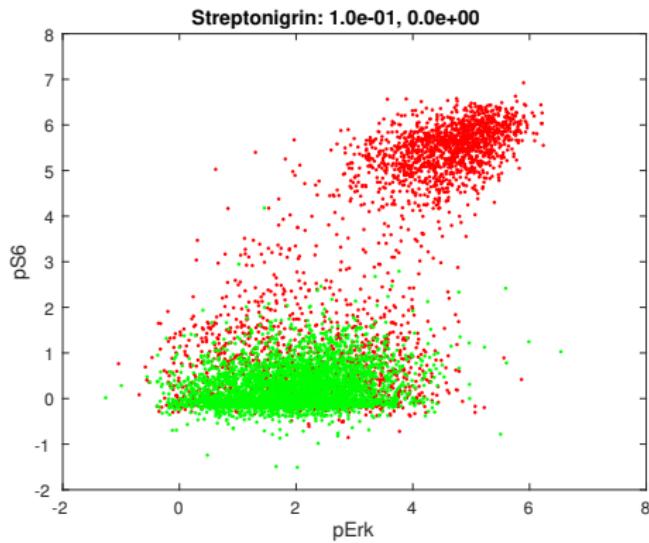
What is the causal relation?

$(X, Y, Z)$  is  
LCD triplet iff:

$$\begin{aligned}Y &\notin \text{an}(X) \\Z &\notin \text{an}(X) \\X &\not\perp\!\!\!\perp Y \\Y &\not\perp\!\!\!\perp Z \\X &\perp\!\!\!\perp Z \mid Y\end{aligned}$$

# LCD: Example

- $p\text{Erk}$ : abundance of phosphorylated Erk in each cell
- $pS6$ : abundance of phosphorylated S6 in cell
- $I$ : green = baseline, red = PMA-IONO activator added



$(X, Y, Z)$  is  
LCD triplet iff:

$$\begin{aligned}Y &\notin \text{an}(X) \\Z &\notin \text{an}(X) \\X &\not\perp\!\!\!\perp Y \\Y &\not\perp\!\!\!\perp Z \\X &\perp\!\!\!\perp Z \mid Y\end{aligned}$$

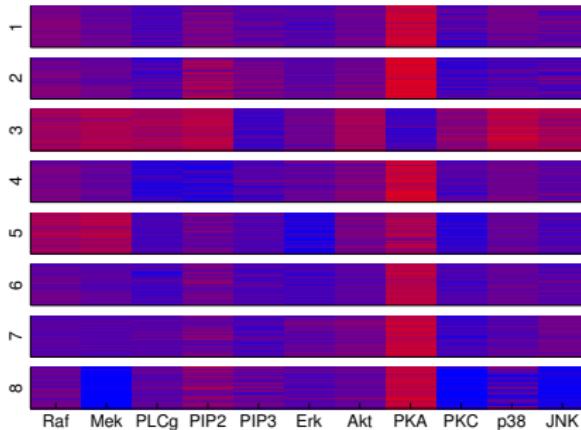
What is the causal relation? LCD triplet  $(I, pS6, p\text{Erk})$ , so  $pS6 \rightarrow p\text{Erk}$ .

**Note:** no prior knowledge on the effects of PMA-IONO needed!

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

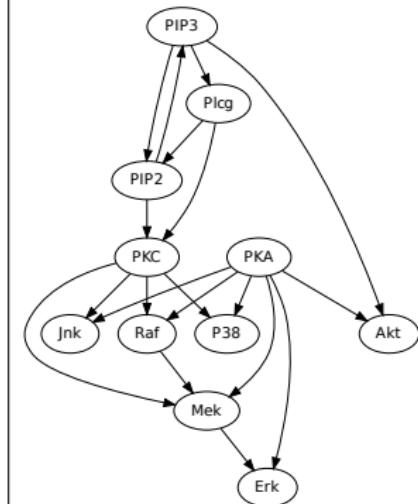
# Causal Discovery: Example Application

## Protein Abundance Data: (Sachs et al, 2005)



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

## Causal Graph: ("Signalling network")



(depicted here: "consensus" network)

# Causal Discovery from Multiple Contexts

	Latent confounders	Nonlinear mechanisms	Cycles	Perfect interventions	Mechanism changes	Activity interventions	Side effects	Other context changes	Unknown intervention/context targets	Learns intervention/context targets	Multiple system variables	Different variables per context	Combination strategy
(Fisher, 1935)	+	+	+	+	+	+	+	+	+	+	-	-	b
(Cooper and Yoo, 1999)	-	+	-	+	-	-	-	-	-	-	+	-	b
(Tian and Pearl, 2001)	-	+	-	-	+	-	-	+	-	-	+	-	b
(Sachs et al., 2005)	-	+	-	+	-	-	-	-	-	-	+	-	b
(Eaton and Murphy, 2007)	-	+	-	+	+	+	+	+	+	+	+	-	b
(Chen et al., 2007)	+	+	+	+	+	+	+	+	+	+	+	-	b
(Claassen and Heskes, 2010)	+	+	-	-	+	+	+	+	+	-	+	+	a
(Tillman and Spirtes, 2011)	+	+	-	-	+	+	+	+	+	-	+	+	a
(Hauser and Bühlmann, 2012)	-	+	-	+	-	-	-	-	-	-	+	-	b
(Hyttinen et al., 2012)	+	-	-	+	-	-	-	-	-	-	+	-	a
(Mooij and Heskes, 2013)	-	±	+	+	-	-	-	-	-	-	+	-	b
(Hyttinen et al., 2014)	+	+	+	+	-	-	-	-	-	-	+	+	a
(Triantafillou and Tsamardinos, 2015)	+	+	-	+	-	-	-	-	-	-	+	+	a
(Rothenhäusler et al., 2015)	+	-	±	±	-	-	-	+	+	+	+	-	a
(Peters et al., 2016)	±	±	±	+	+	+	+	+	+	-	+	-	b
(Oates et al., 2016a)	-	-	-	-	-	-	-	+	-	-	+	-	b
(Zhang et al., 2017)	-	+	-	+	+	+	+	+	+	+	+	-	b
JCI	+	+	+	+	+	+	+	+	+	+	+	±	b
JCI-LCD (Cooper, 1997)	+	+	+	+	+	+	+	+	+	+	+	-	b
JCI-HEJ	+	+	±	+	+	+	+	+	+	+	+	-	b
JCI-FCI	+	+	-	+	+	+	+	+	+	+	+	-	b

## Question

Can we combine the ideas of the “classical” approach to causal discovery based on experimentation with the “modern” approach based on conditional independences?

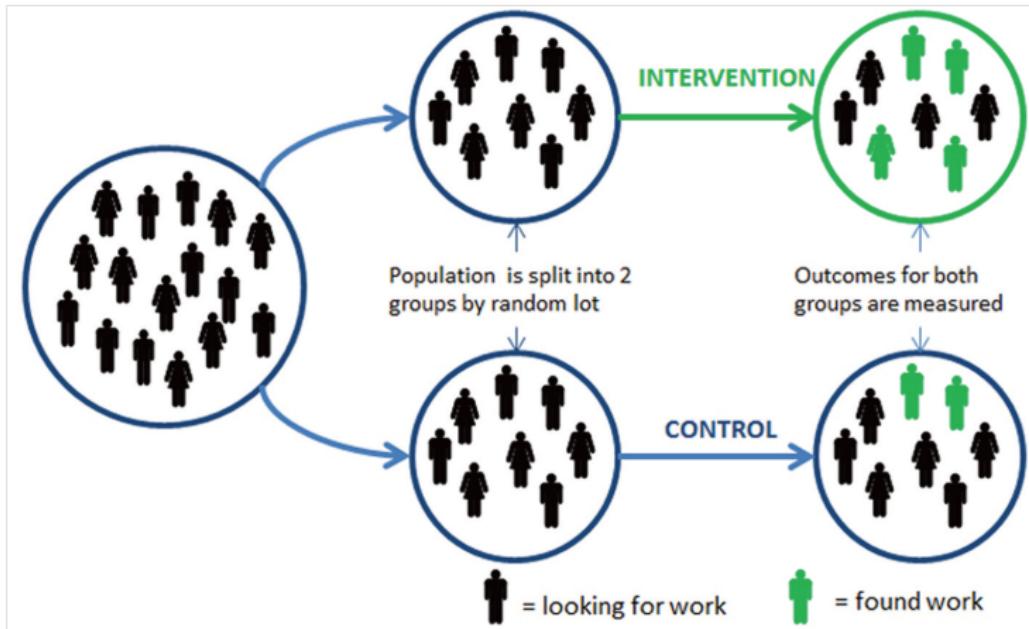
We hope to:

- obtain reliability of “classical” approach
- exploit conditional independences in the data to reduce the number of experiments necessary

## Answer

We propose **Joint Causal Inference**, a framework for causal discovery, that achieves this.

# Randomized Controlled Trials, or A/B-testing



$C_1$	$X_1$
0	1
0	0
0	1
0	0
0	0
0	0
0	0
0	0
1	0
1	0
1	1
1	1
1	0
1	1
1	0
1	1

Two variables: **context** variable  $C_1$ , **system** variable  $X_1$

$C_1$ : 0=control, 1=intervention

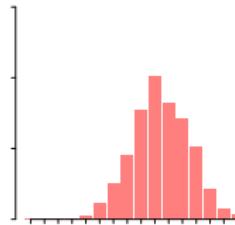
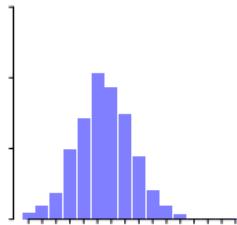
$X_1$ : 0=looking for work, 1=found work

# Two equivalent points of view

(a) Separate data sets

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

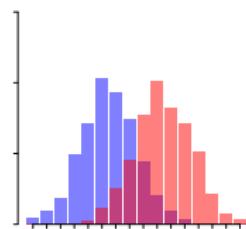
Drug ( $C = 1$ ):
$X$
-0.3
1.8
-0.1
...



Two-sample test:  
Is  $p(x \mid \text{do}(C = 0)) = p(x \mid \text{do}(C = 1))$ ?

(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  $X \perp C$ ?

## Proposition

Suppose  $C$  (treatment) and  $X$  (outcome) can be modeled with a Structural Causal Model. The Randomized Controlled Trial assumptions

- $X$  does not cause  $C$  (*because  $X$  happens after  $C$* )
- $X$  and  $C$  are unconfounded (*because of the randomization*)
- no selection bias (*measure and analyze all samples*)

imply that if  $C \not\perp\!\!\!\perp X$ , then  $C$  causes  $X$  (*correlation implies causation*).

# Causal Inference for Randomized Controlled Trial

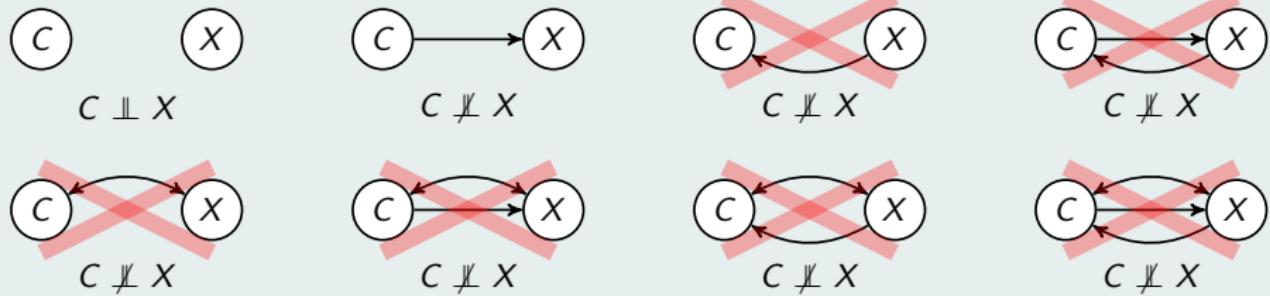
## Proposition

Suppose  $C$  (treatment) and  $X$  (outcome) can be modeled with a Structural Causal Model. The Randomized Controlled Trial assumptions

- $X$  does not cause  $C$  (*because  $X$  happens after  $C$* )
- $X$  and  $C$  are unconfounded (*because of the randomization*)
- no selection bias (*measure and analyze all samples*)

imply that if  $C \not\perp\!\!\!\perp X$ , then  $C$  causes  $X$  (*correlation implies causation*).

## Proof



# JCI: Two types of variables

## Definition

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

Distinguish:

- **Context variables**  $\{C_i\}_{i \in \mathcal{I}}$  that model the context of the system,
- **System** variables  $\{X_j\}_{j \in \mathcal{J}}$  that model the system of interest.

## Example

Data for 3 observed system variables in 4 experimental conditions:

### System variables:

$X_1$ : salary

$X_2$ : drug abuse

$X_3$ : depression

### *no interventions:*

$X_1$	$X_2$	$X_3$
0.1	0.2	0.5
0.13	0.21	0.49
0.23	0.21	0.51

### *only back-to-work program:*

$X_1$	$X_2$	$X_3$
0.2	0.22	0.92
0.23	0.21	0.99

### Context variables:

$C_1$ : back-to-work program

$C_2$ : psychotherapy

### *only psychotherapy:*

$X_1$	$X_2$	$X_3$
0.5	0.19	0.52
0.6	0.18	0.51

### *both interventions:*

$X_1$	$X_2$	$X_3$
0.53	1.2	0.95
0.61	1.21	0.90
0.55	1.19	0.97

# JCI: Pooling the data

After explicitly adding the context variables, we pool the data:

## Example

*no interventions:*

$X_1$	$X_2$	$X_3$
0.1	0.2	0.5
0.13	0.21	0.49
0.23	0.21	0.51

*only psychotherapy:*

$X_1$	$X_2$	$X_3$
0.5	0.19	0.52
0.6	0.18	0.51

*only back-to-work program:*

$X_1$	$X_2$	$X_3$
0.2	0.22	0.92
0.23	0.21	0.99

*both interventions:*

$X_1$	$X_2$	$X_3$
0.53	1.2	0.95
0.61	1.21	0.90
0.55	1.19	0.97

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

**System variables:** **Context variables:**

$X_1$ : salary

$C_1$ : back-to-work program

$X_2$ : drug abuse

$C_2$ : psychotherapy

$X_3$ : depression

## JCI Assumptions (Intuitive formulation)

We are modelling a generic setting in which the experimenter decides on the performed interventions *before* the measurements are performed, and this decision does not depend on anything else that might affect the system of interest.

## Formal JCI Assumptions

The causal graph  $\mathcal{G}$  that includes both system variables  $\{X_1, \dots, X_p\}$  and context variables  $\{C_1, \dots, C_d\}$ , which jointly models the experimental design and the system in *all* experimental conditions, satisfies:

- no variable directly causes any context variable  $C_i$ , and
- none of the pairs  $\{X_k, C_i\}$  of system and context variables is confounded, and
- each pair of context variables  $\{C_i, C_j\}$  is confounded.

Furthermore, we assume the absence of selection bias.

# Joint Causal Inference

**Question:** How can we discover

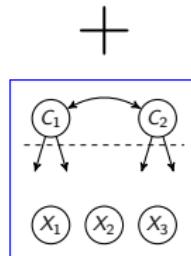
pooled data

interv. 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

Cl's

$$\begin{aligned} X_1 &\not\perp\!\!\!\perp C_1 \\ X_1 &\perp\!\!\!\perp C_1 \mid C_2 \\ X_3 &\perp\!\!\!\perp X_2 \mid X_1 \\ C_1 &\not\perp\!\!\!\perp C_2 \\ &\dots \end{aligned}$$

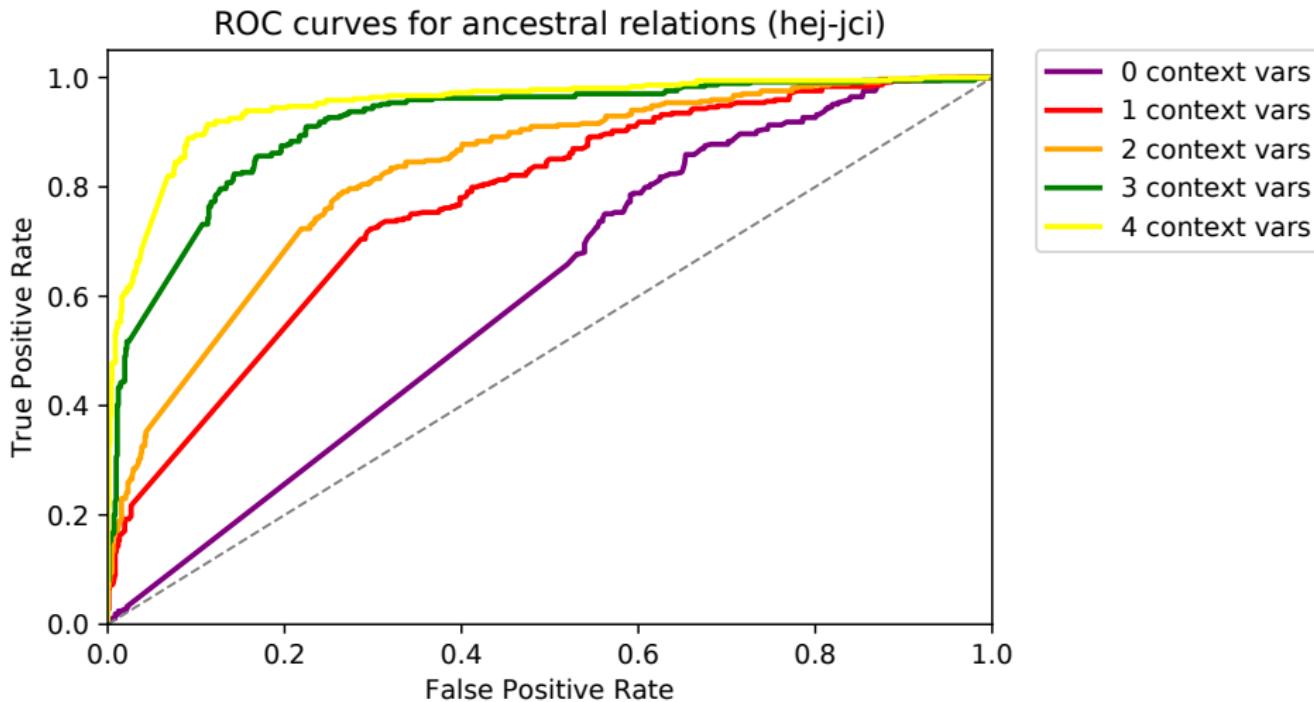
causal graphs



JCI Assumptions

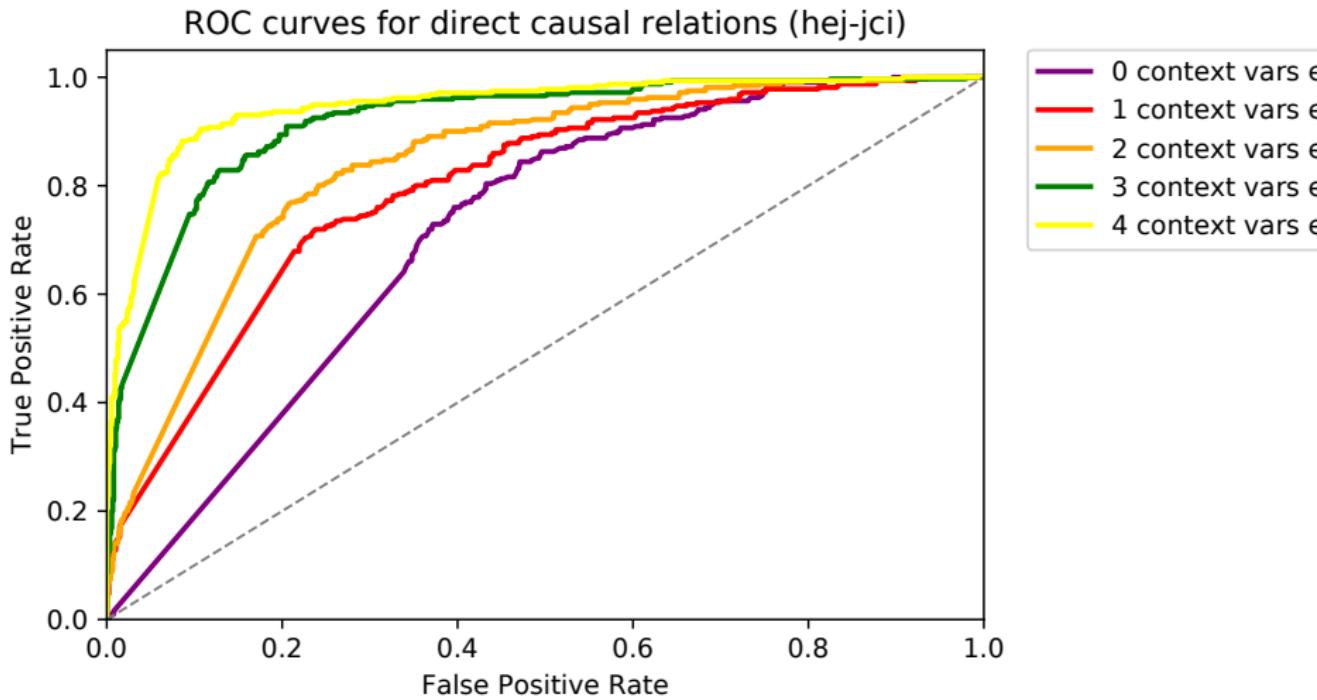
**Answer:** Simply apply a standard constraint-based causal discovery method (designed for purely observational data) on the *pooled* data, and incorporate the JCI assumptions as background knowledge.

# Evaluation on simulated data I



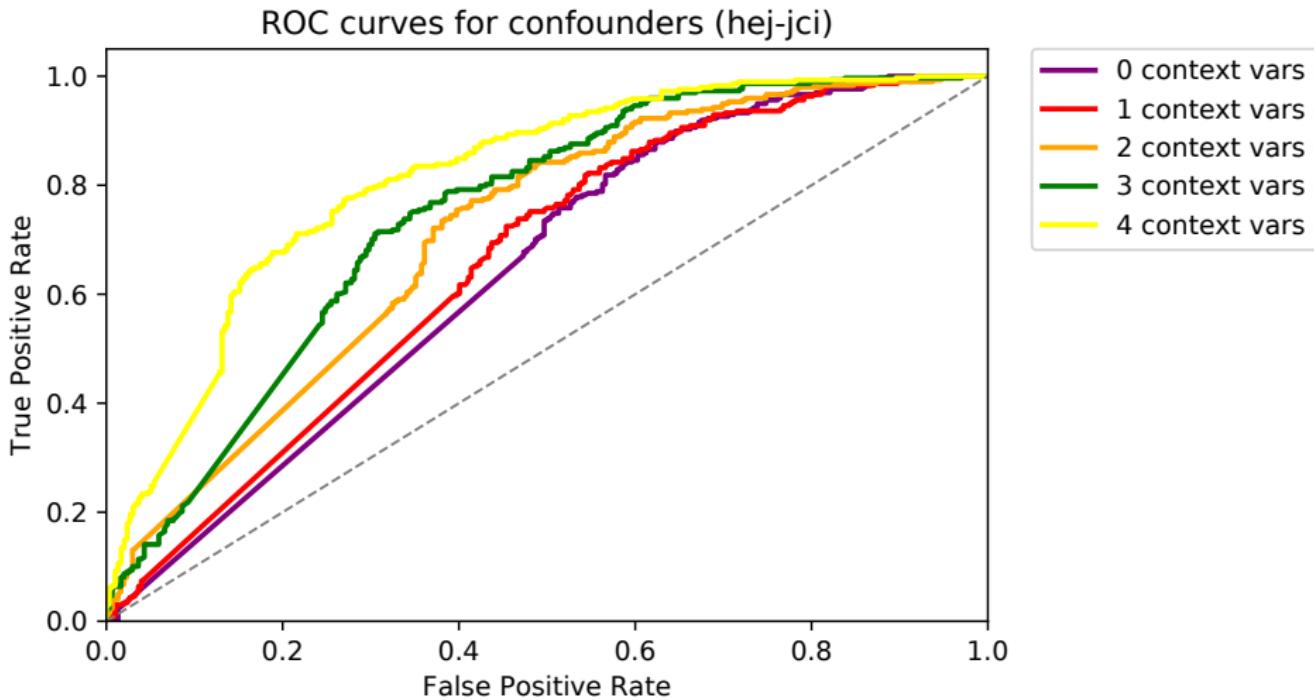
(4 system variables, 500 samples in each data set)

## Evaluation on simulated data II



(4 system variables, 500 samples in each data set)

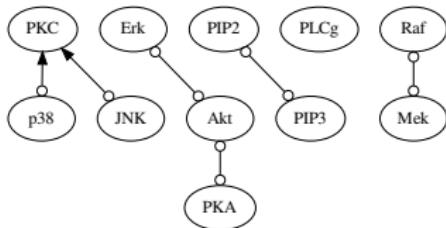
# Evaluation on simulated data III



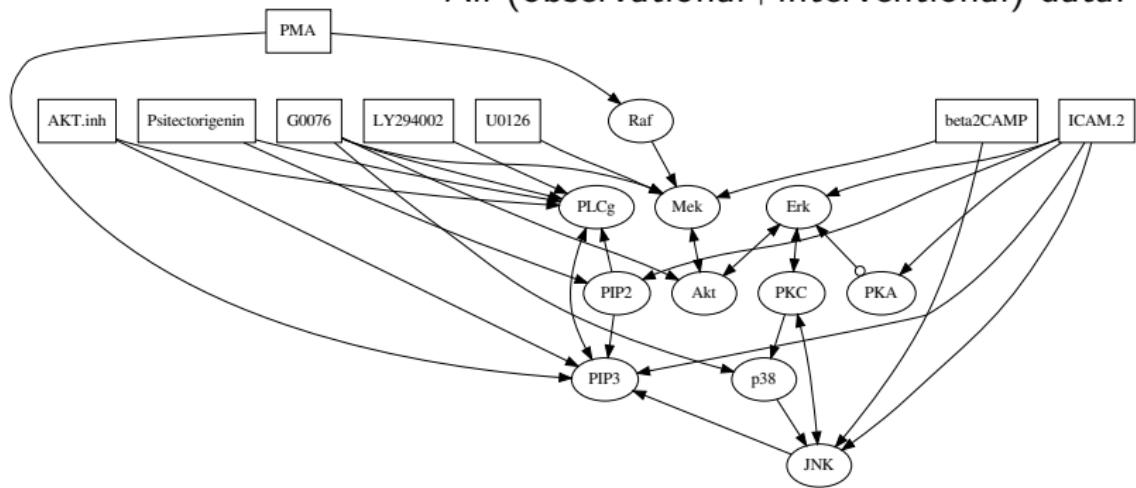
(4 system variables, 500 samples in each data set)

# Evaluation on real-world flow cytometry data

Only observational data:



All (observational+interventional) data:



- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# The generalized directed global Markov property

Given the importance of the Markov property, the first thing we need is a Markov property for cyclic SCMs.

We introduce a notion  **$\sigma$ -separation** that generalizes d-separation:

- $\sigma$ -separation implies d-separation.
- For acyclic graph,  $\sigma$ -separation is equivalent to d-separation.

Inspired by ideas by [Spirtes, 1996], we show:

**Theorem ([Forré and Mooij, 2017])**

If an SCM  $\mathcal{M}$  is uniquely solvable w.r.t. every strongly connected component in  $\mathcal{G}(\mathcal{M})$ , then the **generalized directed global Markov property** holds for any solution  $\mathbf{X}$  of  $\mathcal{M}$  with respect to the graph  $\mathcal{G}(\mathcal{M})$ :

$$A \perp\!\!\!\perp_{\mathcal{G}(\mathcal{M})}^{\sigma} B \mid Z \implies \mathbf{X}_A \perp\!\!\!\perp_{\mathbb{P}^{\mathbf{X}}} \mathbf{X}_B \mid \mathbf{X}_Z \quad A, B, Z \subseteq \mathcal{I}.$$

# Markov properties: $\sigma$ -separation

Definition ( $\sigma$ -separation, [Forré and Mooij, 2017])

In a DMG  $\mathcal{G}$ , a path

$$i_1 \xrightarrow{\quad} \cdots \xrightarrow{\quad} i_n$$
$$\Leftrightarrow \qquad \Leftrightarrow$$

is called  **$\sigma$ -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  $i_{k-1} \xleftrightarrow{\quad} i_k \xleftrightarrow{\quad} i_{k+1}$  with  $i_k \notin \text{ang}(Z)$ , or
- it contains a non-collider with  $i_k \in Z$ :

$$i_{k-1} \xrightarrow{\quad} i_k \rightarrow i_{k+1}, \quad i_{k-1} \leftarrow i_k \xrightarrow{\quad} i_{k+1},$$
$$\Leftrightarrow \qquad \Leftrightarrow$$

where the child  $i_{k+1}$  (resp.  $i_{k-1}$ ) is not in  $\text{scg}(i_k)$ .

We say that  $A$  is  $\sigma$ -separated from  $B$  by  $Z$ , denoted  $A \perp^\sigma B | Z$ , if every path with one end node in  $A$  and one end node in  $B$  is  $\sigma$ -blocked by  $Z$ .

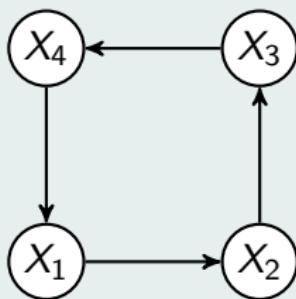
# Markov properties: Example

## Example

SCM  $\mathcal{M}$ :

$$\begin{aligned}X_1 &= f_1(X_4, E_1) = X_4 + E_1 \\X_2 &= f_2(X_1, E_2) = X_1 \cdot E_2 \\X_3 &= f_3(X_2, E_3) = X_2 + E_3 \\X_4 &= f_4(X_3, E_4) = X_3 \cdot E_4\end{aligned}$$

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



$$X_1 \perp^d X_3 | X_2, X_4$$

but

$$X_1 \not\perp^\sigma X_3 | X_2, X_4$$

So for any solution  $\mathbf{X}$  of the SCM  $\mathcal{M}$ , in general we do not have that  $X_1 \perp \!\!\! \perp X_3 | X_2, X_4$ .

In general: No  $\sigma$ -separations between nodes within the same strongly connected component.

Stronger statements can be derived for special cases:

Theorem ([Forré and Mooij, 2017])

If an SCM  $\mathcal{M}$  satisfies at least one of the following three conditions:

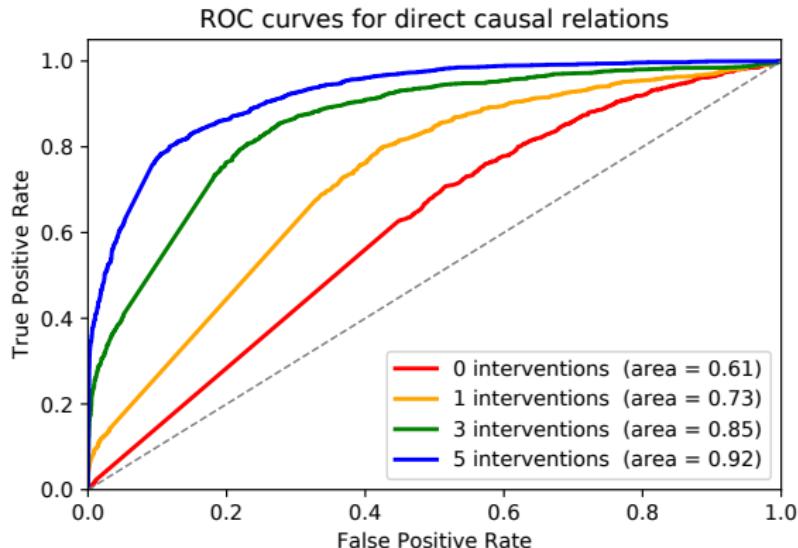
- ①  $\mathcal{M}$  is linear, its exogenous variables have a density with respect to Lebesgue measure, and  $\mathcal{M}$  is solvable w.r.t.  $\mathcal{I}$ ;
- ② all endogenous variables are discrete-valued,  $\mathcal{M}$  is uniquely solvable w.r.t. each ancestral subgraph of  $\mathcal{G}(\mathcal{M})$ ;
- ③  $\mathcal{M}$  is acyclic;

then the directed global Markov property holds for any solution  $\mathbf{X}$  of  $\mathcal{M}$  with respect to the graph  $\mathcal{G}(\mathcal{M})$ :

$$A \underset{\mathcal{G}(\mathcal{M})}{\perp\!\!\!\perp}^d B | Z \implies \mathbf{X}_A \underset{\mathbb{P}^{\mathbf{X}}}{\perp\!\!\!\perp} \mathbf{X}_B | \mathbf{X}_Z \quad A, B, Z \subseteq \mathcal{I}.$$

# Results on Synthetic Data [Forré and Mooij, 2018]

[Forré and Mooij, 2018]: the first causal discovery algorithm that can handle cycles, nonlinear relationships, latent confounding variables and data from different (interventional) contexts.

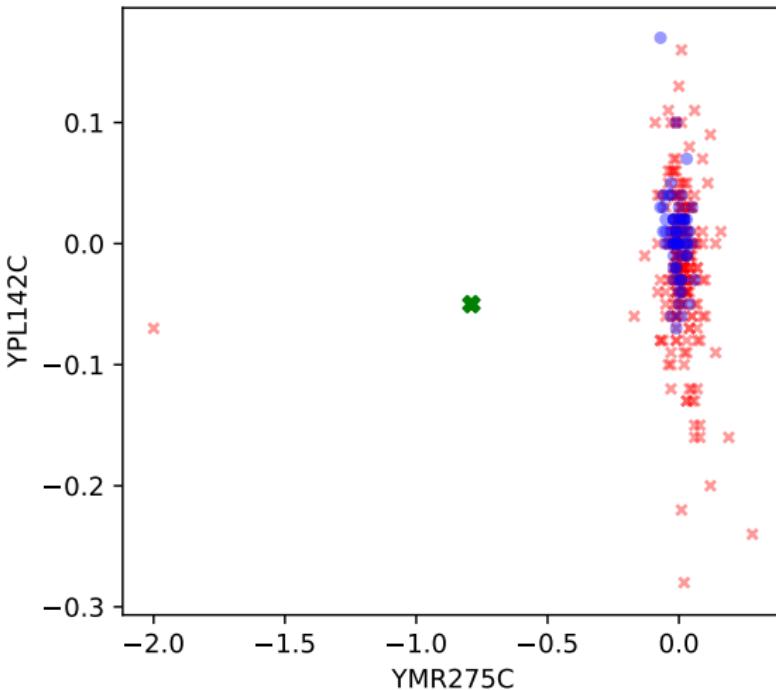


ROC curves for detecting direct causal relations from observational and interventional data, for varying numbers of interventional data sets.

- 1 Informal Causal Modeling: Causal Graphs
- 2 Causal Modeling: Structural Causal Models
- 3 Markov Properties: From Graph to Conditional Independences
- 4 Causal Inference: Predicting Causal Effects
- 5 Causal Discovery: From Data to Causal Graph
  - Causal Discovery by Experimentation
  - Causal Discovery from Observational Data
  - Causal Discovery from Multiple Contexts
- 6 Dealing with Cycles
- 7 Large-Scale Validation of Causal Discovery

# Example Gene Expression Data

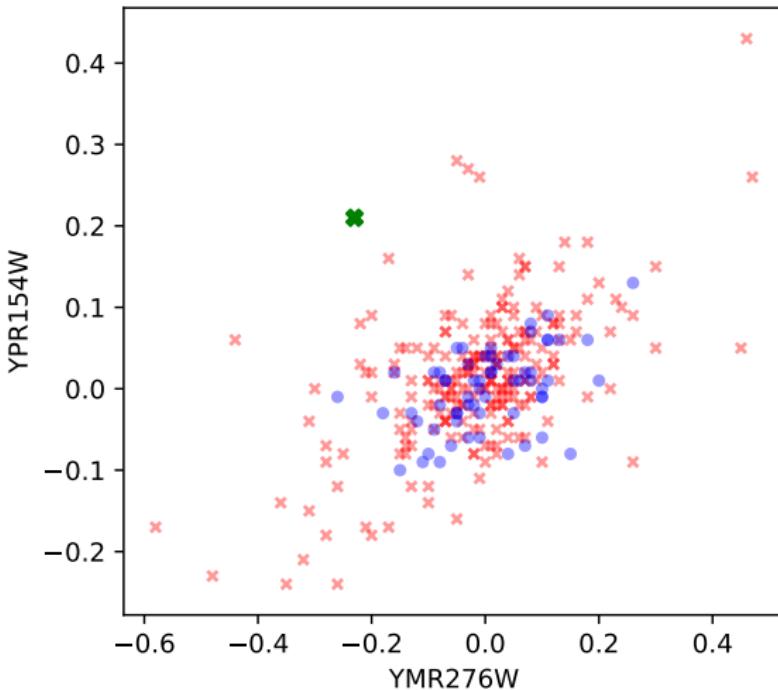
YMR275C does not cause YPL142C:



Blue: observational; Red: interventional; Green: knockout of gene X.

# Example Gene Expression Data

YMR276W causes YPR154W:

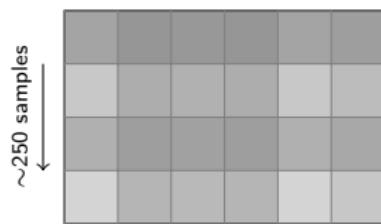


Blue: observational; Red: interventional; Green: knockout of gene X.

# Causal Discovery from Large-Scale Micro-Array Data

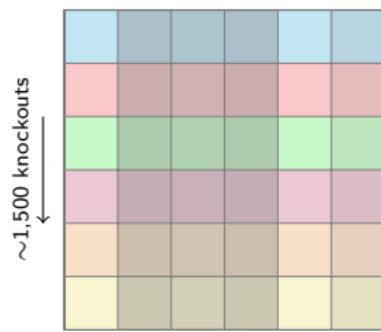
Observational:

~6,000 genes  
→



Interventional:

~6,000 genes  
→



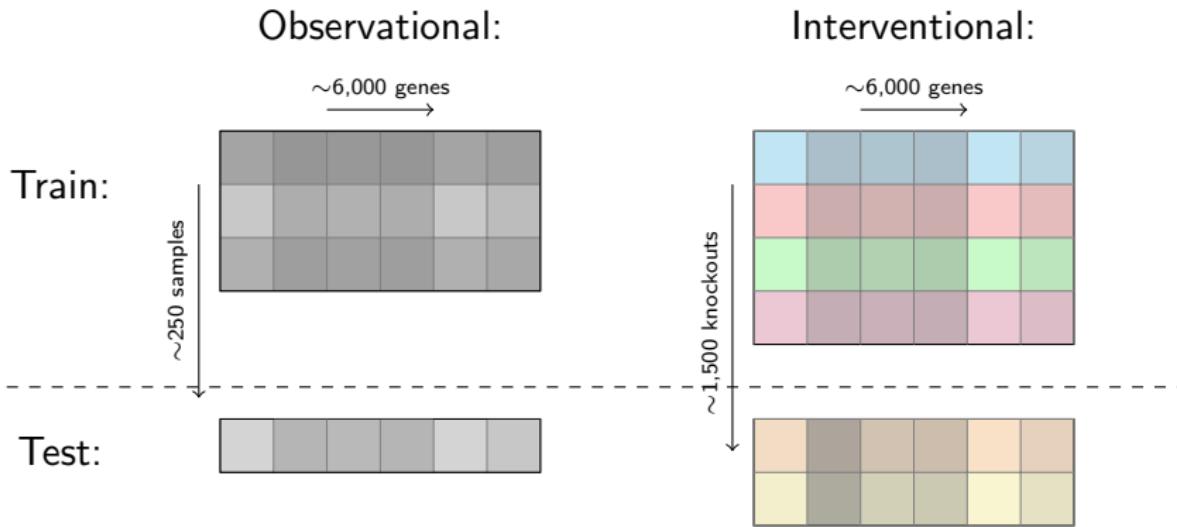
Large-scale Micro-Array Gene Expression Data  
(Kemmeren *et al.*, 2014):

- Variables  $i \in \{1, \dots, p\}$   
(*population gene expression levels*)  
 $p = 6170$
- Observational samples  $X_{in}, n = 1 \dots N_{\text{obs}}$   
(*wild-type vs. wild-type*)  
 $N_{\text{obs}} = 262$
- Interventional samples  $X_i^k, k = 1 \dots N_{\text{int}}$   
(*single-gene knockouts/knockdowns*)  
 $N_{\text{int}} = 1462$   
one sample for every knocked out gene

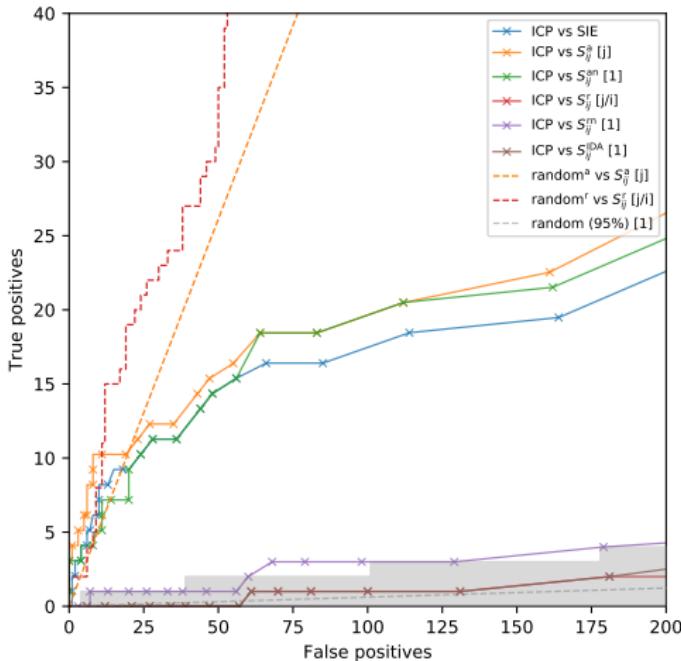
**Task:** Predict from the data which gene expression levels change when a certain gene is knocked out.

# $k$ -fold Cross-validation

Using 5-fold cross-validation, we split the data into a training set used to make predictions, and a test set used to define a ground truth for validating the predictions.



# ICP Performance



ICP outperforms baselines (for the 0.61% strongest effects) for certain ground truth scores (absolute normalized, SIE)

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

In this tutorial, you have learned how to:

- formalize the notion of causality;
- reason about causality;
- discover causal relations from data;
- make causal predictions
- that *seeing* is not the same as *doing*

This was just a sample of topics in an exciting research field. There is still much more to learn and to discover!



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. (2018).

Theoretical aspects of cyclic structural causal models.

*arXiv.org preprint*, arXiv:1611.06221v2 [stat.ME].



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].

## Further reading II

-  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)*.
-  Mooij, J. M., Janzing, D., and Schölkopf, B. (2013).  
From ordinary differential equations to structural causal models: the deterministic case.  
In Nicholson, A. and Smyth, P., editors, *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 440–448. AUAI Press.
-  Pearl, J. (2000).  
*Causality: Models, Reasoning, and Inference*.  
Cambridge University Press.



Peters, J., Janzing, D., and Schölkopf, B. (2017).

*Elements of Causal Inference: Foundations and Learning Algorithms.*  
The MIT Press.



Spirites, P., Glymour, C., and Scheines, R. (2000).

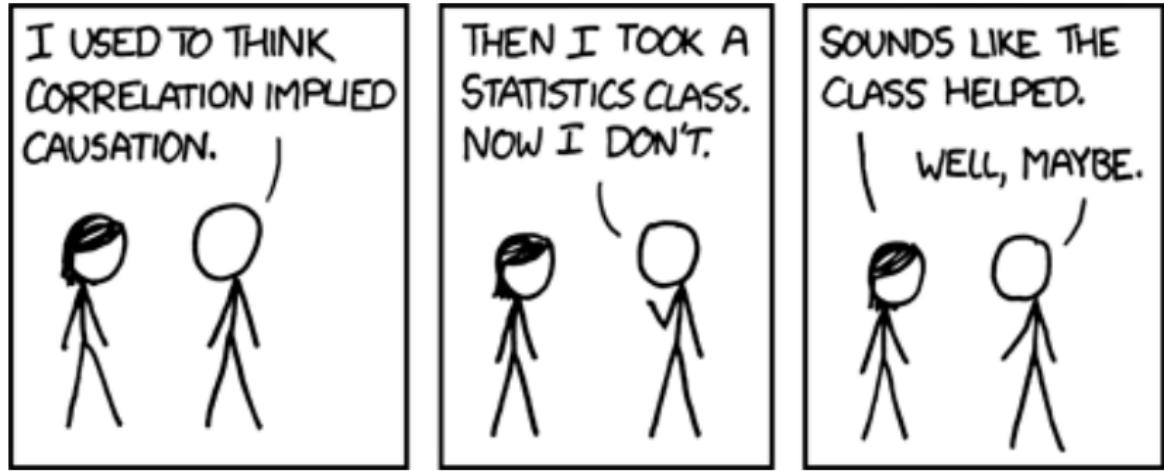
*Causation, Prediction, and Search.*  
The MIT Press.



Wright, S. (1921).

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

# Thank you for your attention!



Randall Munroe, [www.xkcd.org](http://www.xkcd.org)