MSR AI Summer School: Causality I

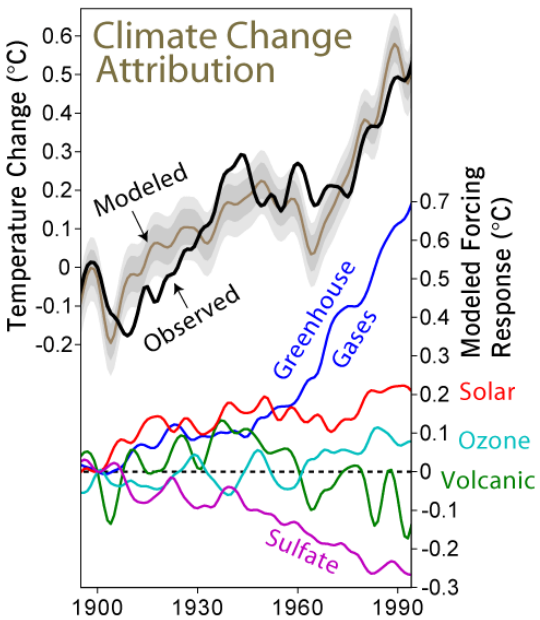
Joris Mooij
j.m.mooij@uva.nl

UNIVERSITY OF AMSTERDAM

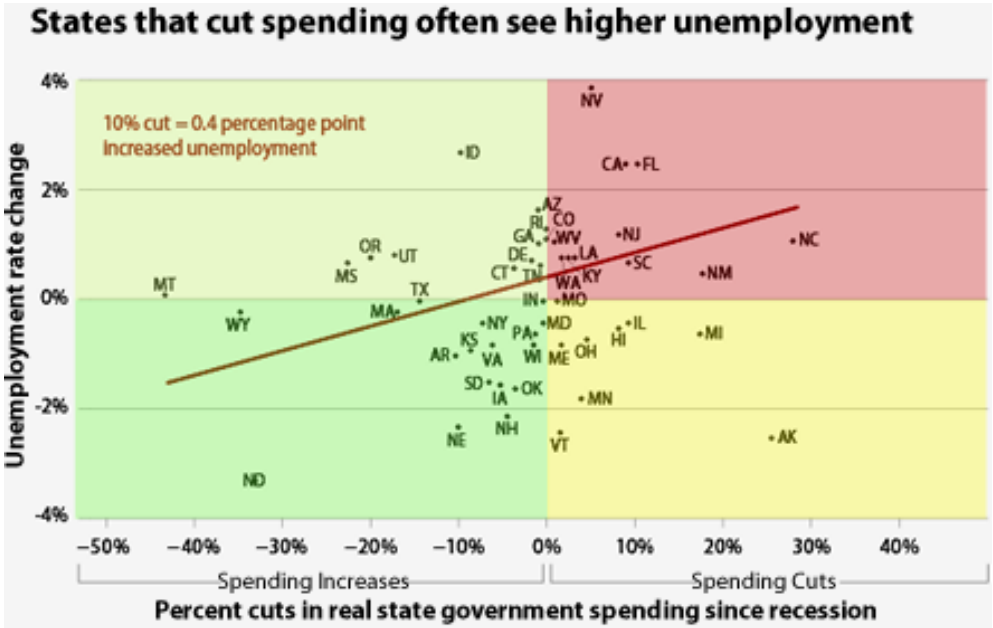
July 3rd, 2018

Many questions in science are *causal*

Climatology:

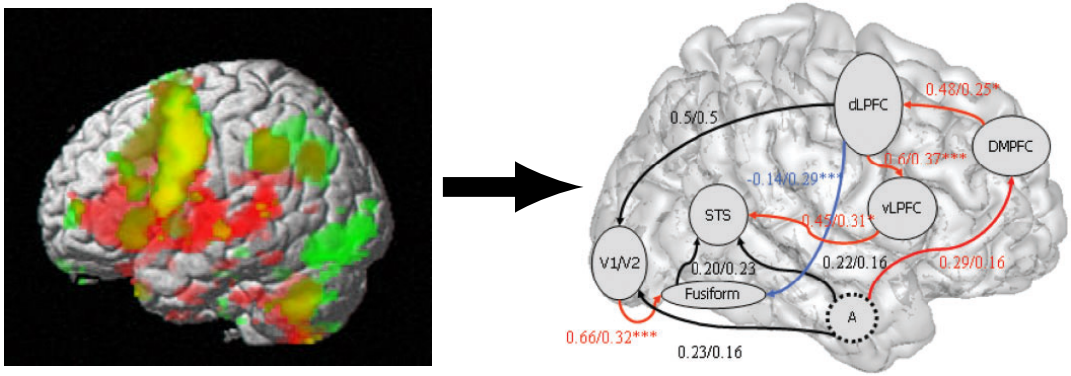


Economy:



Medicine:

Neuroscience:



Contents of this Tutorial

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 vs. Causal 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?*

Outline

1. **Qualitative Causality: Causal Graphs**
2. Quantifying Causality: Structural Causal Models
3. Markov Properties: From Graph to Conditional Independences
4. Causal Inference: Predicting Causal Effects

Causation $\neq$ Correlation

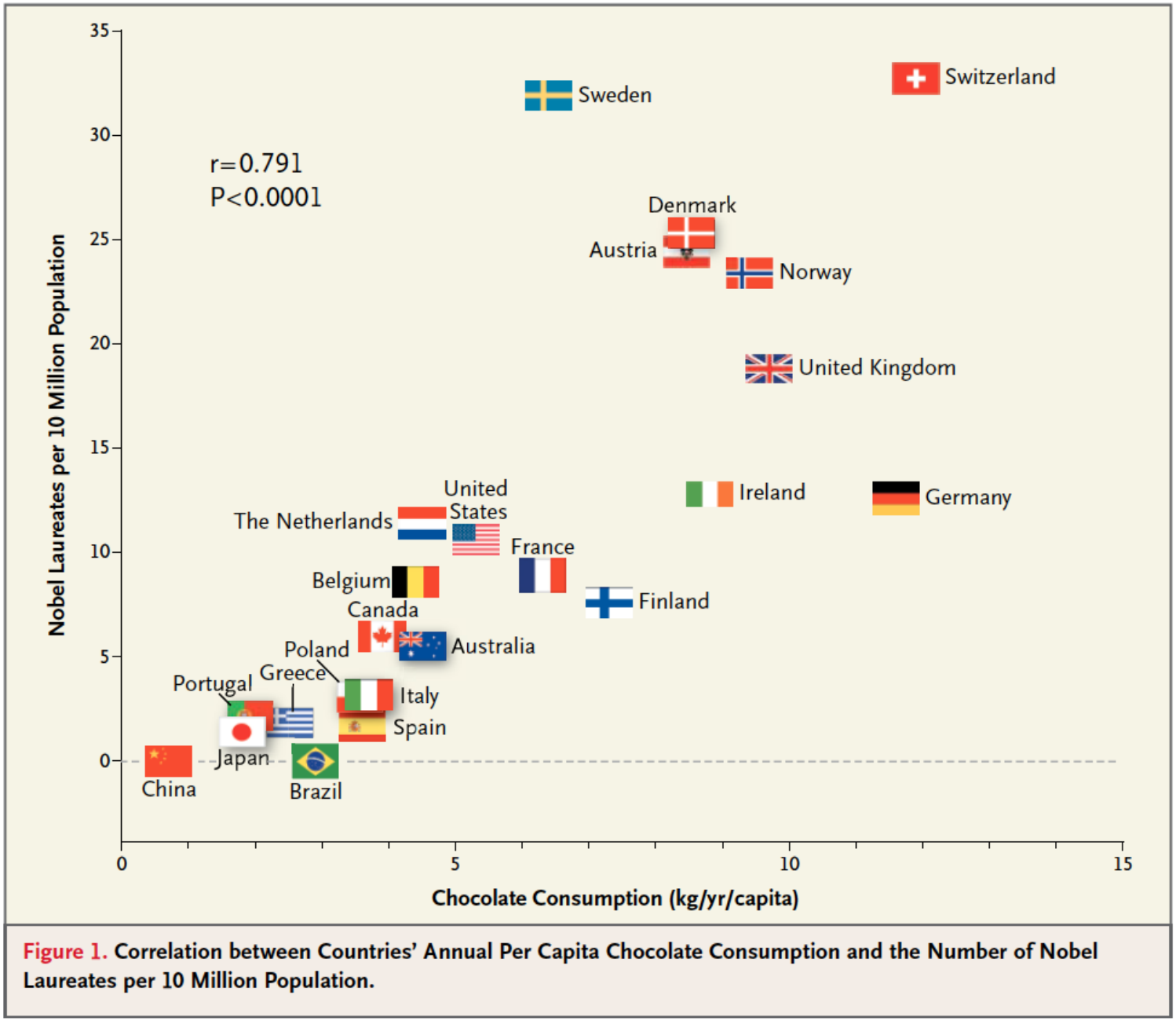


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

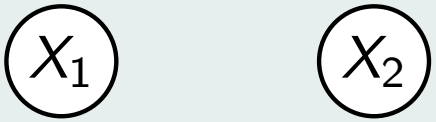
Causal Relations

Definition (Informal)

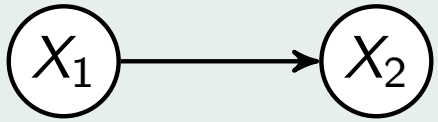
Let A and B be two distinct variables of system. A causes B ($A \dashrightarrow B$) if changing A (*intervening on* A) leads to a change of B.

Causal graph represents causal relationships between variables graphically.

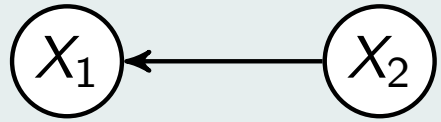
Example



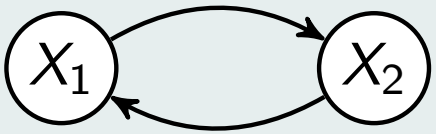
X_1 and X_2 are causally unrelated



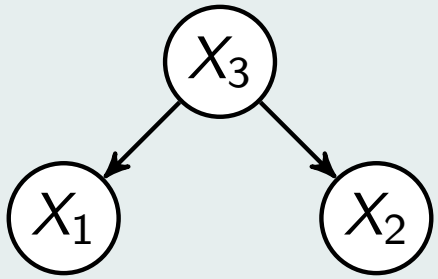
X_1 causes X_2



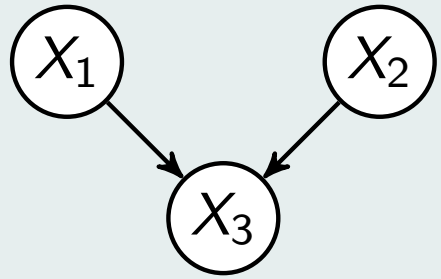
X_2 causes X_1



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

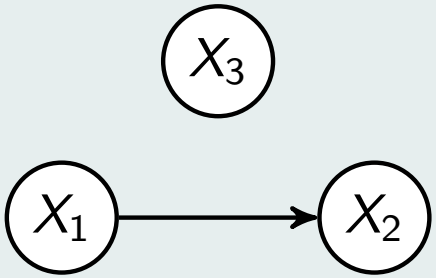
Let $\boldsymbol{V} = \{X_1, \ldots, X_N\}$ be a set of variables.

Definition

If X_i causes X_j even if all other variables $\boldsymbol{V} \setminus \{X_i, X_j\}$ are hold fixed at arbitrary values, then

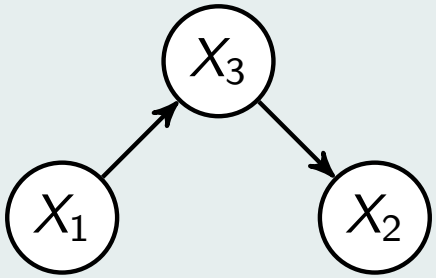
- we say that X_i causes X_j directly with respect to $\boldsymbol{V}$
- we indicate this in the causal graph on $\boldsymbol{V}$ by a directed edge $X_i \to X_j$

Example



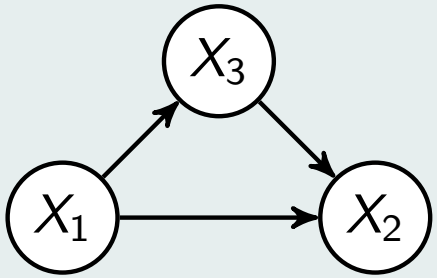
X_1 causes X_2;

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. $\{X_1, X_2, X_3\}$

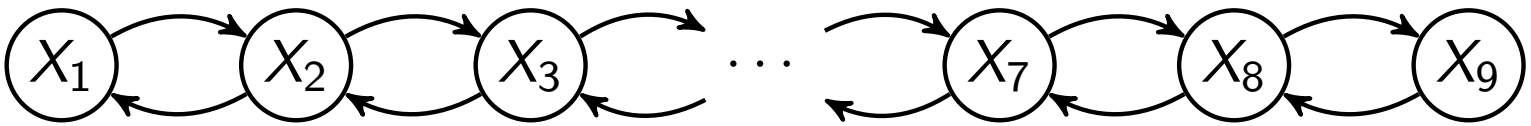


X_1 causes X_2;

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

Definition (Informal)

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

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

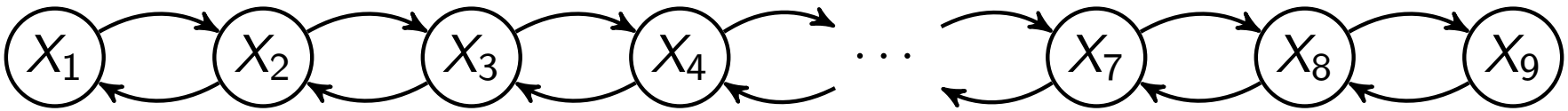
Note

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

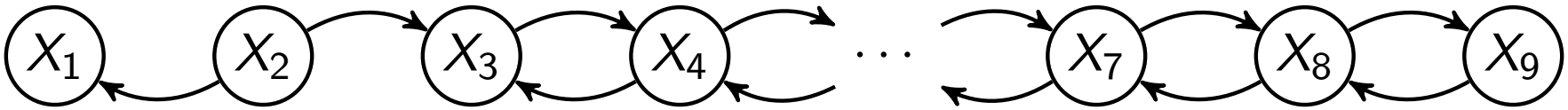
Perfect interventions: Example

Consider the 9 dominoes stones, and the perfect intervention that enforces X_2 to be in "upright" position.

Before the intervention, the causal graph is:



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



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

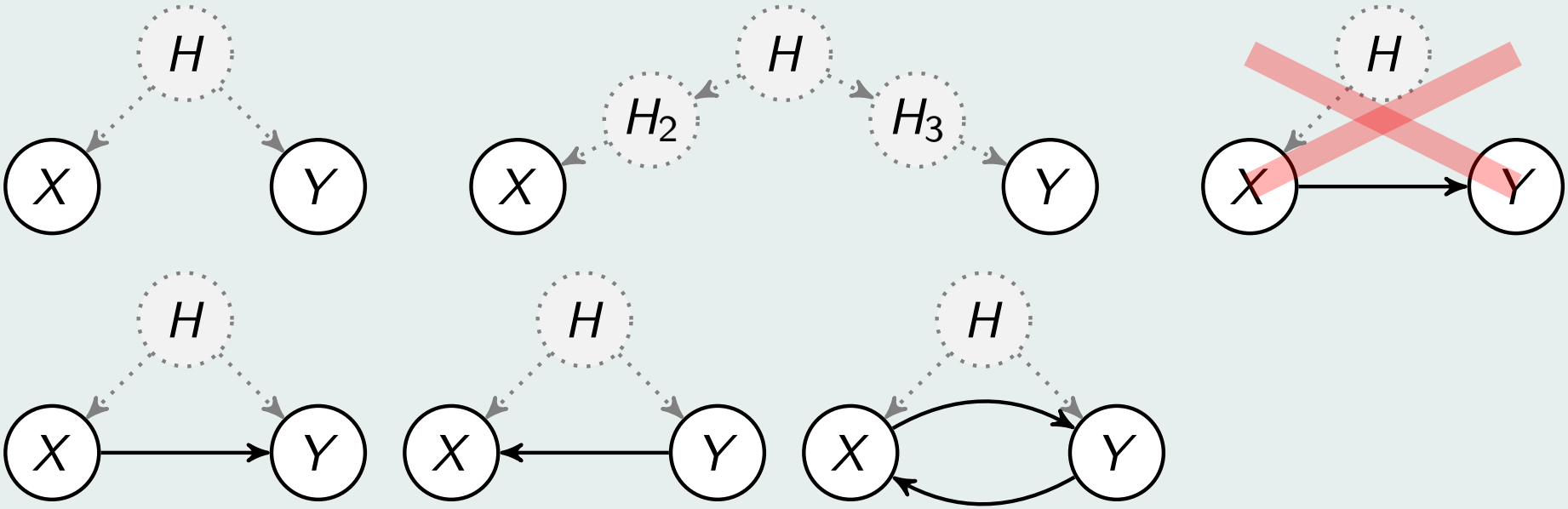
Informally: a confounder is a latent common cause.

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 confounds chocolate consumption and Nobel prize winners.

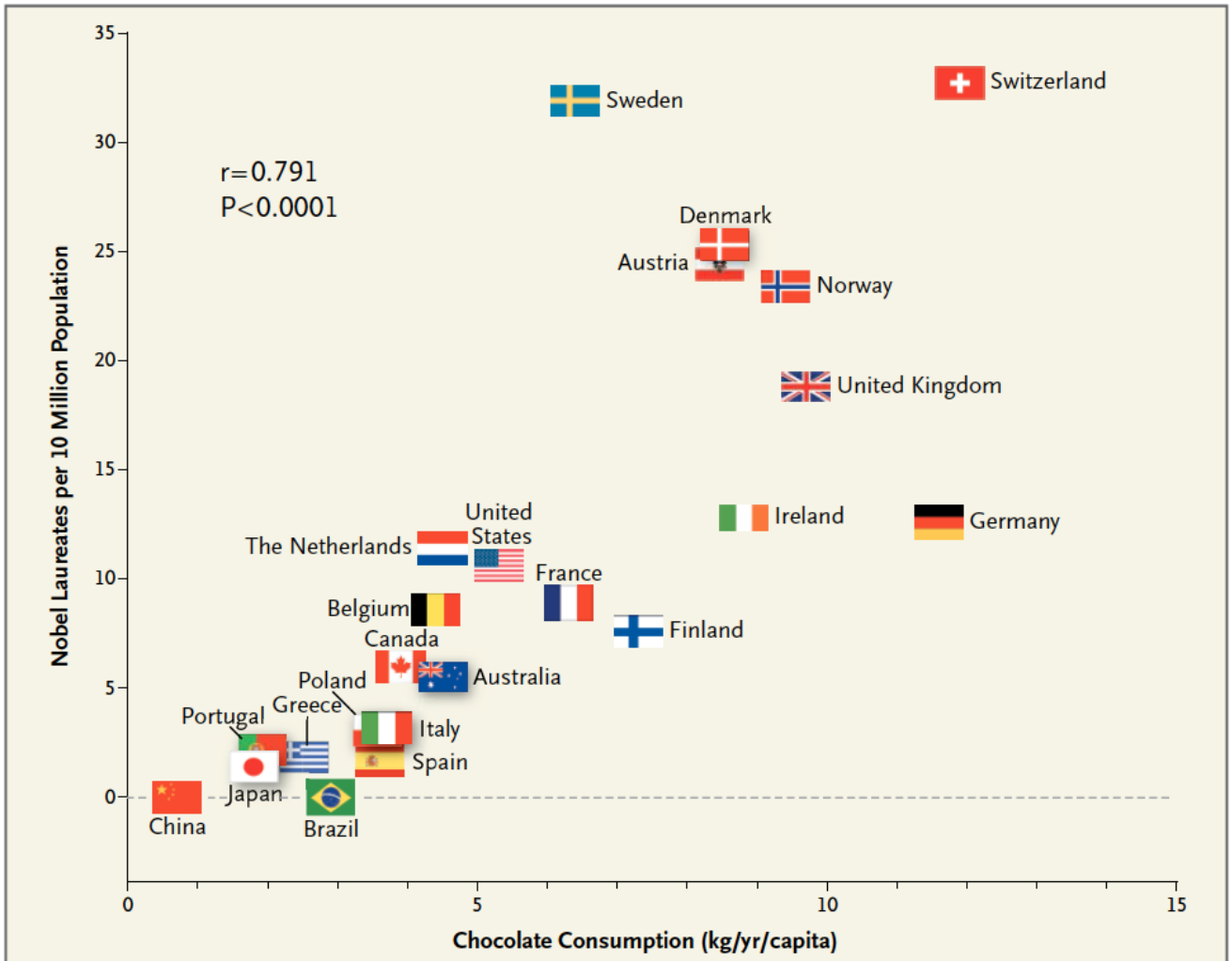
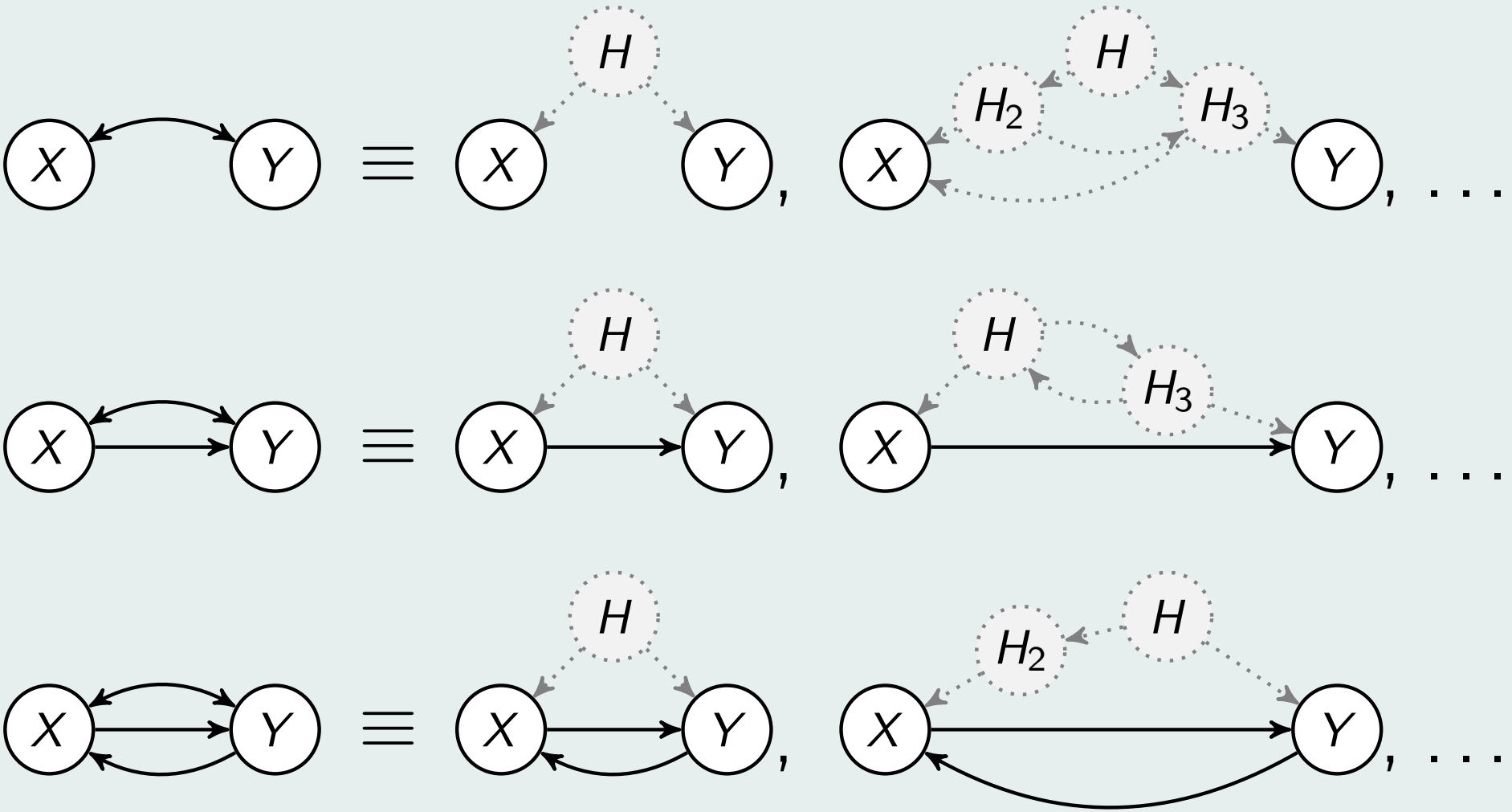


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Confounders: Graphical notation

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

Example



Cycles: Definitions

Let A, B be two variables in a system.

Definition

If A causes B and B causes A, then A and B are involved in a causal cycle.

Cycles: Definitions

Let A, B be two variables in a system.

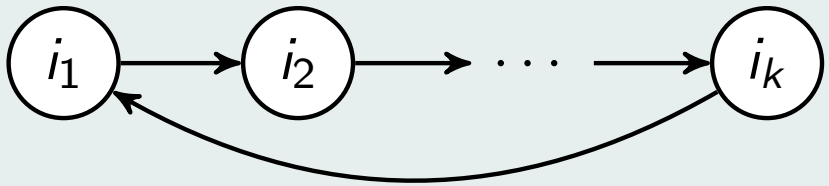
Definition

If A causes B and B causes A, then A and B are involved in a causal cycle.

Let $\mathcal{G}$ be a Directed Mixed Graph with directed and bidirected edges.

Definition

$\mathcal{G}$ is cyclic if it contains a directed cycle

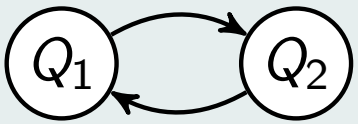


If $\mathcal{G}$ does not contain such a directed cycle, it is called acyclic, and known as an Acyclic Directed Mixed Graph (ADMG). If in addition, $\mathcal{G}$ does not contain any bidirected edges, it is called a Directed Acyclic Graph (DAG).

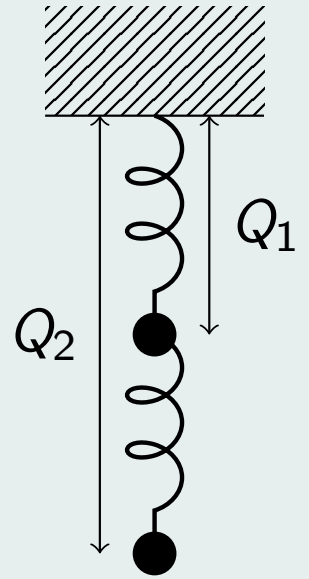
Cycles: Toy example

Example (Damped Coupled Harmonic Oscillators)

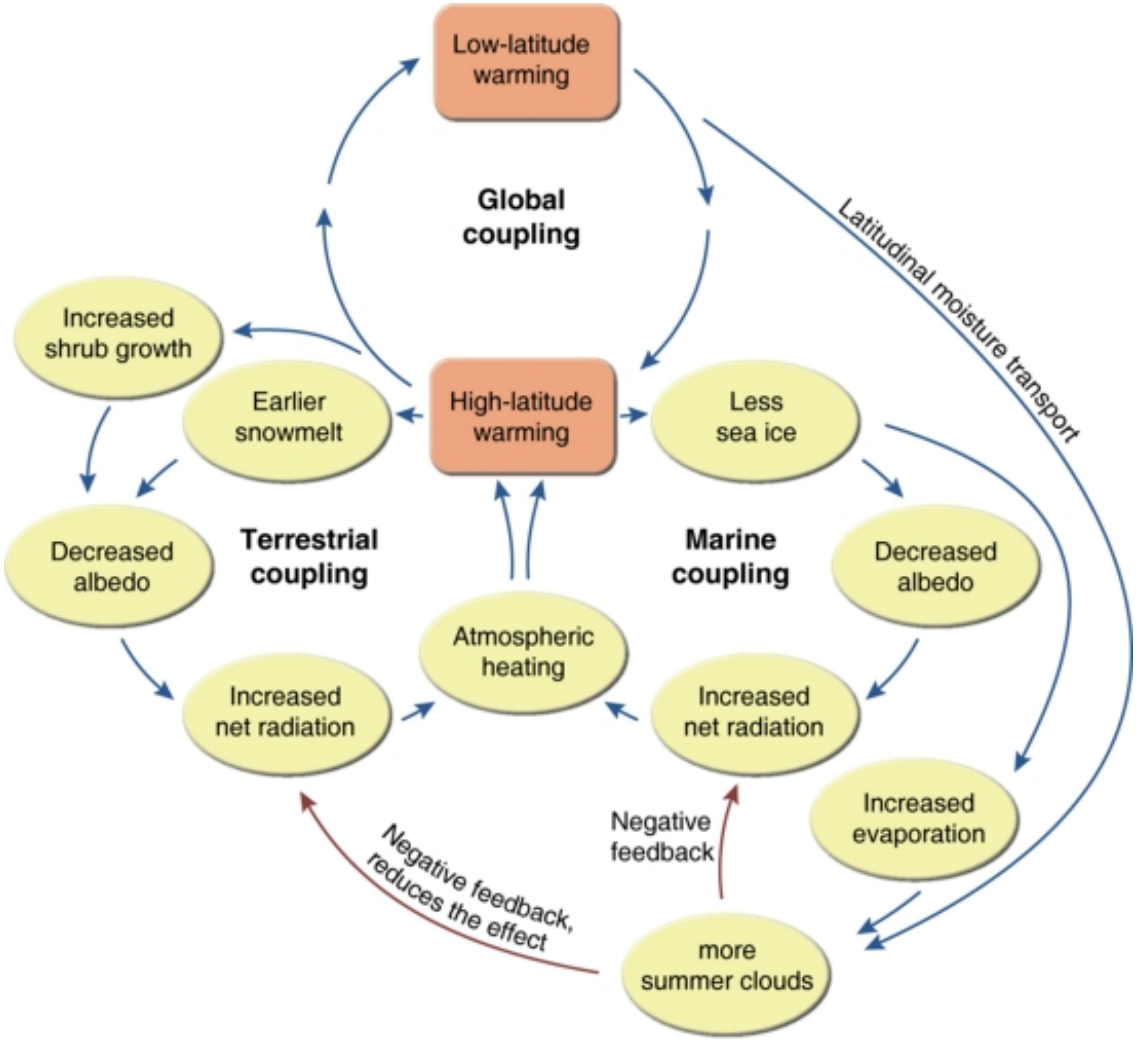
- Two masses, connected by a spring, suspended from the ceiling by another spring.
- Variables: vertical **equilibrium** positions Q_1 and Q_2.
- Q_1 causes Q_2.
- Q_2 causes Q_1.
- Causal graph:



- Cannot be modeled with acyclic causal model!

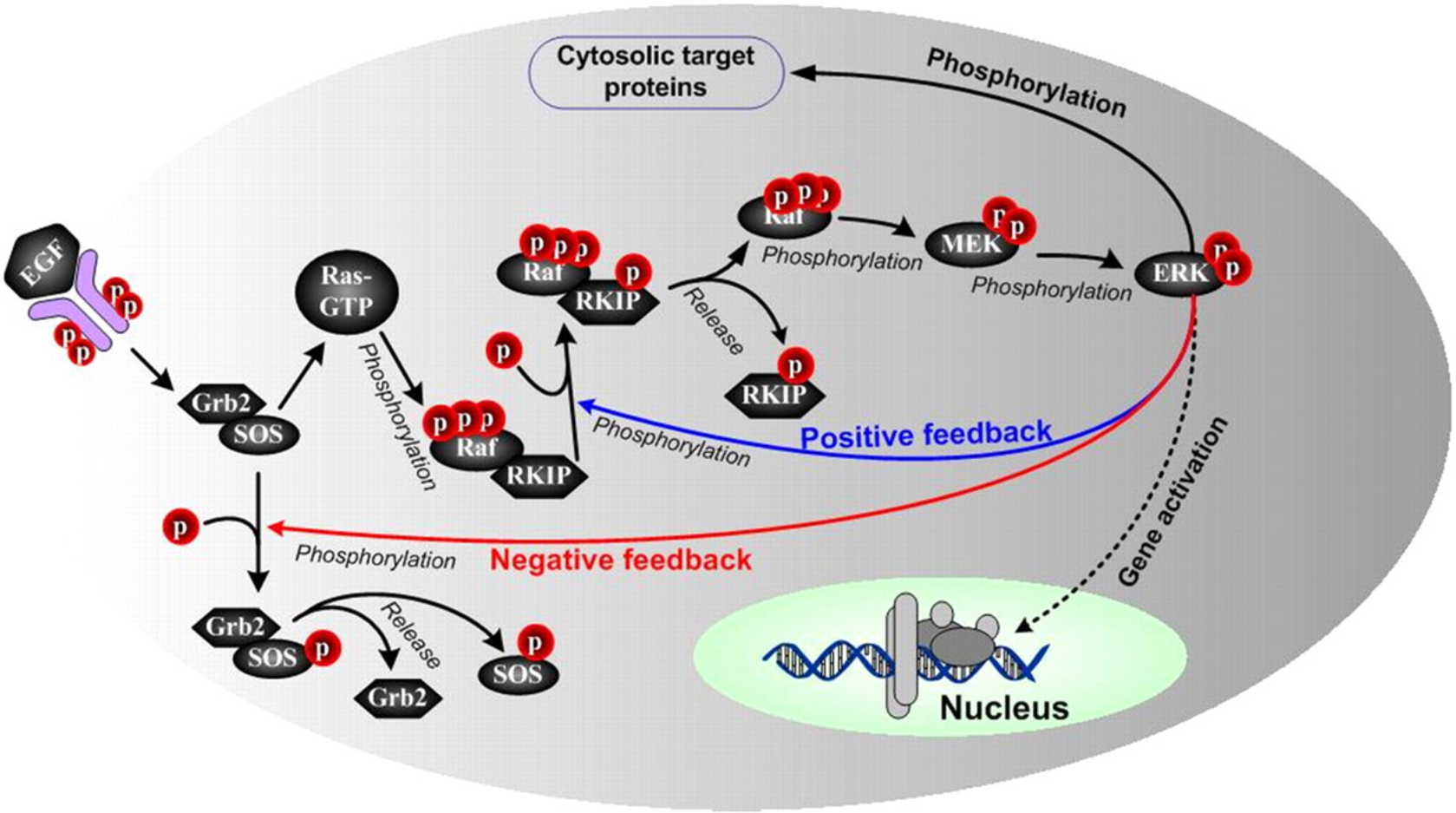


Cycles: Relevance in Climatology



"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]

Outline

1. Qualitative Causality: Causal Graphs
2. **Quantifying Causality: Structural Causal Models**
3. Markov Properties: From Graph to Conditional Independences
4. Causal Inference: Predicting Causal Effects

Defining Causality in terms of Probabilities?

It is a natural idea to try to *define* causality in terms of probabilities.

A naïve example of such an attempt could be:
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*.

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**:

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

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

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

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**:

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

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

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 discuss the latter modeling framework, because it is arguably the most general of the three.

Structural Causal Models: Concepts

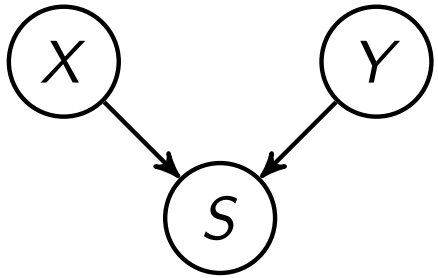
SCMs turn things around: rather than defining causality in terms of probabilities, probability distributions are defined by a causal model, thereby avoiding traps like Simpson's paradox.

- 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,
  - modeled by probability distributions,
  - *not caused* by endogenous variables.
- Each endogenous variable has its own structural equation, which describes how this variable depends causally on other variables.
- SCMs are equipped with a notion of perfect intervention, which gives them a *causal* semantics.

Structural Causal Models: Example

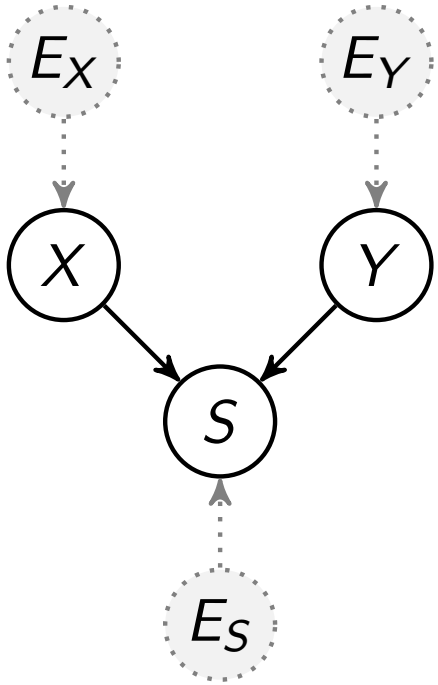
Endogenous variables (binary):

X: the battery is charged
Y: the start engine is operational
S: the car starts



Exogenous variables (latent, independent, binary):

$$E_X \sim \mathrm{Ber}(0.95)$$
$$E_Y \sim \mathrm{Ber}(0.99)$$
$$E_S \sim \mathrm{Ber}(0.999)$$



Structural equations (one per endogenous variable):

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

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 $\mathcal{M} = \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with:

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

Definition

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

Structural Causal Models: Example

Example

Structural Causal Model $\mathcal{M}$:

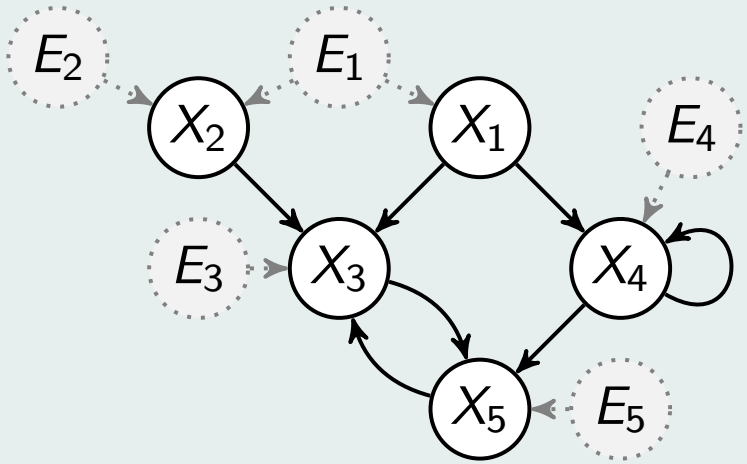
Formally:

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

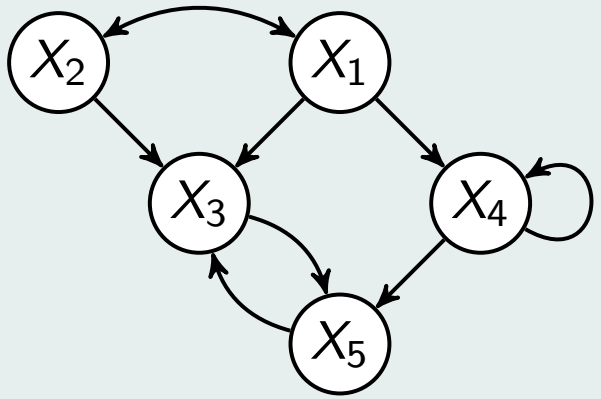
Informally:

$$\begin{array}{ll} X_1 = f_1(E_1) & \mathbb{P}^{E_1} = \ldots \\ X_2 = f_2(E_1, E_2) & \mathbb{P}^{E_2} = \ldots \\ X_3 = f_3(X_1, X_2, X_5, E_3) & \mathbb{P}^{E_3} = \ldots \\ X_4 = f_4(X_1, X_4, E_4) & \mathbb{P}^{E_4} = \ldots \\ X_5 = f_5(X_3, X_4, E_5) & \mathbb{P}^{E_5} = \ldots \end{array}$$

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



Functional graph $\mathcal{G}(\mathcal{M})$:



(Augmented) Functional Graphs

Definition

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

$$X_i = f_i(\boldsymbol{X}_{\mathrm{pa}_i^{\mathcal{I}}}, \boldsymbol{E}_{\mathrm{pa}_i^{\mathcal{J}}})$$

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

Definition

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

Definition

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

Causal Graph

Proposition

If $\mathcal{M}$ has no self-loops, the causal graph of $\mathcal{M}$ is a subgraph of the functional graph $\mathcal{G}(\mathcal{M})$.

In that case, generically:

- The directed edges in $\mathcal{G}(\mathcal{M})$ represent direct causal effects w.r.t. $\mathcal{I}$;
- The bidirected edges in $\mathcal{G}(\mathcal{M})$ represent the existence of confounders w.r.t. $\mathcal{I}$;

A particular case of interest is:

Definition

We call the SCM $\mathcal{M}$ acyclic if $\mathcal{G}(\mathcal{M})$ is acyclic.

If in addition, $\mathcal{G}(\mathcal{M})$ doesn't have bidirected edges, this leads to a causal Bayesian network.

Interventions

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 $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$ enforces $\boldsymbol{X}_I$ to attain value $\boldsymbol{\xi}_I$.
- This changes the SCM $\mathcal{M} = \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ into the intervened SCM $\mathcal{M}_{\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)} = \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ where

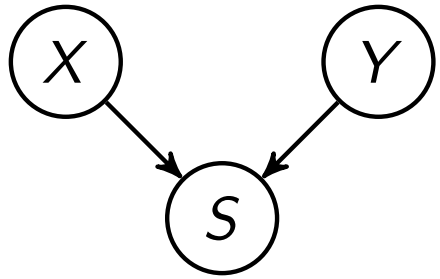
$$\tilde{f}_i = \begin{cases} \xi_i & i \in I \\ f_i(\boldsymbol{X}_{\mathrm{pa}_i^{\mathcal{I}}}, \boldsymbol{E}_{\mathrm{pa}_i^{\mathcal{J}}}) & i \notin I. \end{cases}$$

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

Interventions: Example

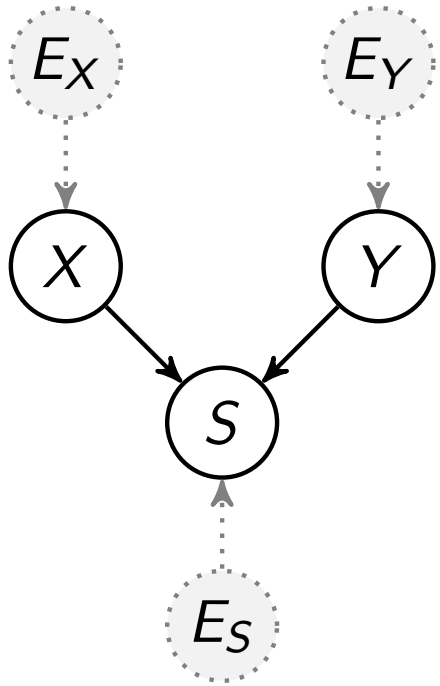
Endogenous variables (binary):

X: the battery is charged
Y: the start engine is operational
S: the car starts



Exogenous variables (latent, independent, binary):

$$E_X \sim \mathrm{Ber}(0.95)$$
$$E_Y \sim \mathrm{Ber}(0.99)$$
$$E_Z \sim \mathrm{Ber}(0.999)$$



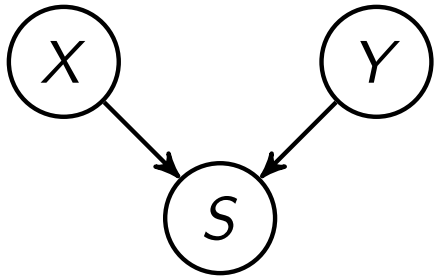
Structural equations (one per endogenous variable):

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

Interventions: Example

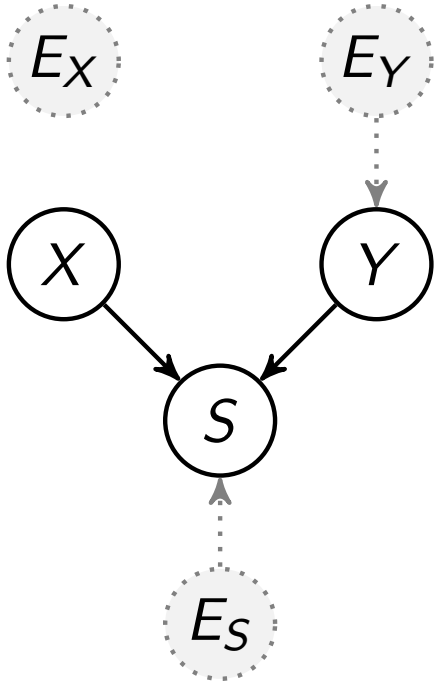
Endogenous variables (binary):

X: the battery is charged
Y: the start engine is operational
S: the car starts



Exogenous variables (latent, independent, binary):

$$E_X \sim \mathrm{Ber}(0.95)$$
$$E_Y \sim \mathrm{Ber}(0.99)$$
$$E_Z \sim \mathrm{Ber}(0.999)$$



Structural equations (one per endogenous variable): after charging the battery $\mathrm{do}(X_1 = 1)$:

$$X = 1$$
$$Y = E_Y$$
$$S = X \wedge Y \wedge E_S$$

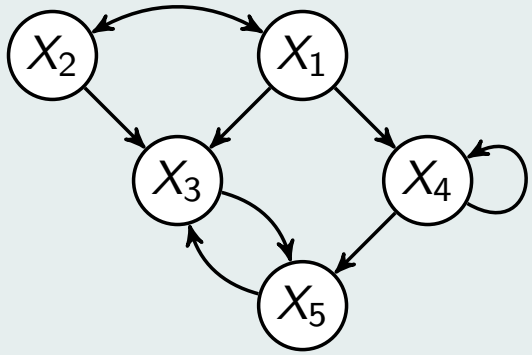
Example

Observational (no intervention):

Structural Causal Model $\mathcal{M}$:

$$
\begin{aligned}
X_1 &= f_1(E_1) & \mathbb{P}^{E_1} &= \dots \\
X_2 &= f_2(E_1, E_2) & \mathbb{P}^{E_2} &= \dots \\
X_3 &= f_3(X_1, X_2, X_5, E_3) & \mathbb{P}^{E_3} &= \dots \\
X_4 &= f_4(X_1, X_4, E_4) & \mathbb{P}^{E_4} &= \dots \\
X_5 &= f_5(X_3, X_4, E_5) & \mathbb{P}^{E_5} &= \dots
\end{aligned}
$$

Functional graph $\mathcal{G}(\mathcal{M})$:

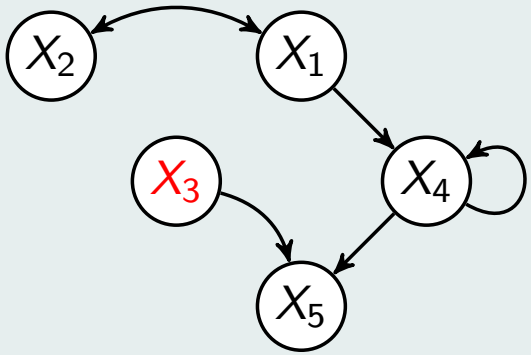


Intervention $\mathrm{do}(X_3 = \xi_3)$:

Structural Causal Model $\mathcal{M}_{\mathrm{do}(X_3=\xi_3)}$:

$$
\begin{aligned}
X_1 &= f_1(E_1) & \mathbb{P}^{E_1} &= \dots \\
X_2 &= f_2(E_1, E_2) & \mathbb{P}^{E_2} &= \dots \\
X_3 &= \xi_3 & \mathbb{P}^{E_3} &= \dots \\
X_4 &= f_4(X_1, X_4, E_4) & \mathbb{P}^{E_4} &= \dots \\
X_5 &= f_5(X_3, X_4, E_5) & \mathbb{P}^{E_5} &= \dots
\end{aligned}
$$

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



Observational Distributions

Remember:

Definition

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

Definition

We call the set of probability distributions of the solutions $\boldsymbol{X}$ of an SCM $\mathcal{M}$ the observational distributions of $\mathcal{M}$.

An important special case:

Proposition

If $\mathcal{M}$ is acyclic, then it its observational distribution exists and is unique. We denote its marginal density on $\boldsymbol{X}$ simply by $p(\boldsymbol{x})$.

Interventional Distributions

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

Definition

We call the family of sets of probability distributions of the solutions of $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ (for $I \subseteq \mathcal{I}$, $\boldsymbol{\xi}_I \subseteq \boldsymbol{\mathcal{X}}_I$) the interventional distributions of $\mathcal{M}$.

Crucial difference with common statistical models: SCMs *simultaneously* model the distributions under all perfect interventions on a system.

Interventional Distributions

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

Definition

We call the family of sets of probability distributions of the solutions of $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ (for $I \subseteq \mathcal{I}$, $\boldsymbol{\xi}_I \subseteq \boldsymbol{\mathcal{X}}_I$) the interventional distributions of $\mathcal{M}$.

Crucial difference with common statistical models: SCMs *simultaneously* model the distributions under all perfect interventions on a system.

Definition

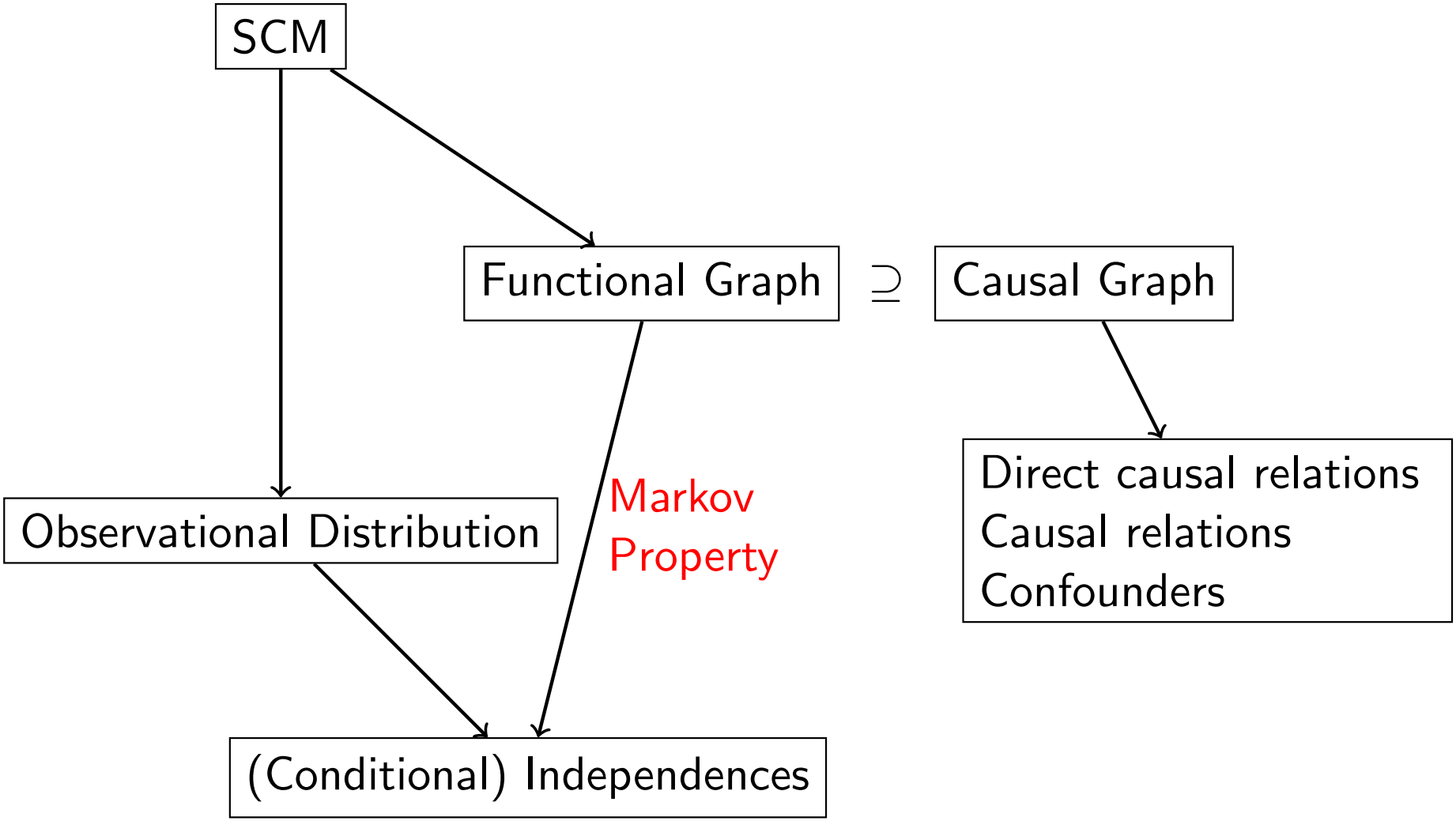
If $\mathcal{M}$ is acyclic, all its interventional distributions exist and are unique. Following [Pearl, 2000], we denote their densities by $p\big(\boldsymbol{x} \mid \mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)\big)$.

We can now express "correlation does not imply causation" (or, as Pearl says, "seeing is not doing") more precisely:

$$p\big(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{X} = \boldsymbol{x})\big) \neq p(\boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) \qquad \text{in general}$$

Representations of acyclic SCMs

For acyclic SCMs, we get the following relationships:



Outline

1. Qualitative Causality: Causal Graphs
2. Quantifying Causality: Structural Causal Models
3. **Markov Properties: From Graph to Conditional Independences**
4. Causal Inference: Predicting Causal Effects

(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).

(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.

(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 path in which each edge is of the form $\cdots \to \cdots$ is called directed.
- If there is a directed path from X to Y, X is called a ancestor of Y.
- The ancestors of Y are denoted $\mathrm{an}_{\mathcal{G}}(Y)$, and include Y.

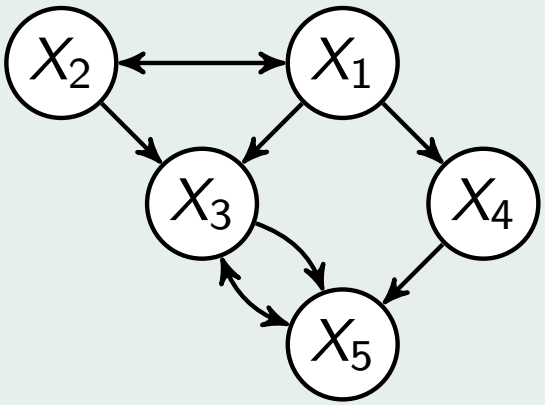
(Directed) Paths

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 path in which each edge is of the form $\cdots \to \cdots$ is called directed.
- If there is a directed path from X to Y, X is called a ancestor of Y.
- The ancestors of Y are denoted $\mathrm{an}_{\mathcal{G}}(Y)$, and include Y.

Example



$X_1 \to X_3 \leftarrow X_1$ is not a path.
$X_1 \leftrightarrow X_2 \to X_3$ is a path.

$X_1 \to X_4 \to X_5$ is a directed path.
$X_4 \to X_5 \leftarrow X_3$ is not a directed path.

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

Colliders and non-colliders

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 arrow heads 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.

Colliders and non-colliders

Definition (Colliders)

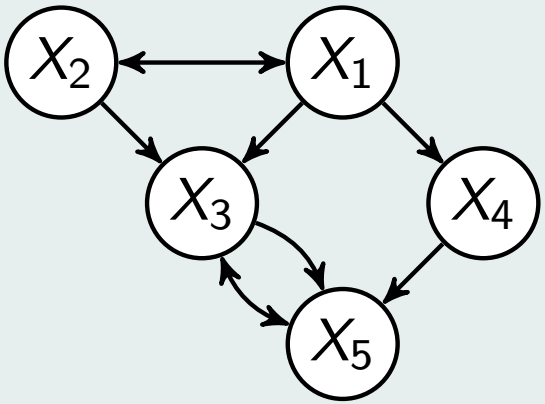
Let $\mathcal{G}$ be a directed mixed graph, and q a path on $\mathcal{G}$.

- A collider on q is a (non-endpoint) node X on q with precisely two arrow heads 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 $X_5 \leftrightarrow X_3 \leftarrow X_1$.

Blocked paths

Definition

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

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

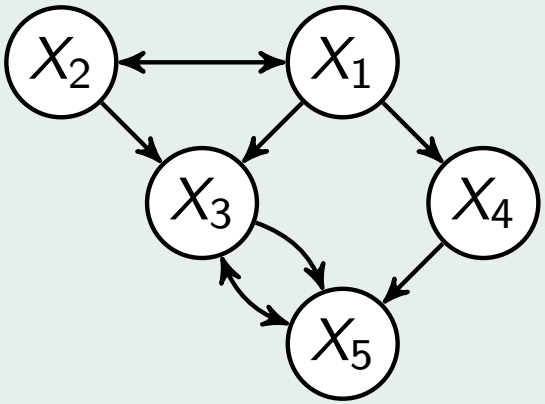
Blocked paths

Definition

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

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

Example



$X_3 \to X_5 \leftarrow X_4$ is blocked by $\emptyset$.
$X_3 \to X_5 \leftarrow X_4$ is blocked by $\{X_1\}$.
$X_3 \to X_5 \leftarrow X_4$ is not blocked by $\{X_5\}$.

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

d-separation

Definition (d-separation)

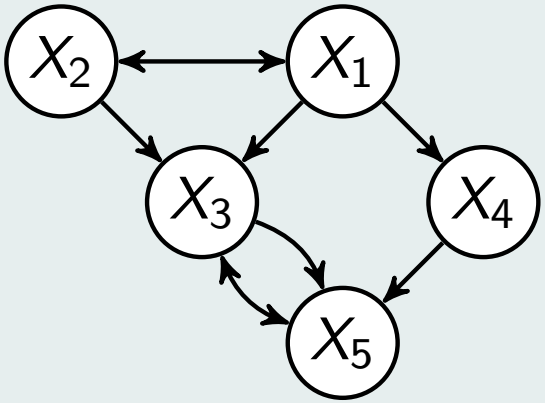
Let $\mathcal{G}$ be a directed mixed graph. For three sets of nodes $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ of nodes, we say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ iff all paths between a node in $\boldsymbol{X}$ and a node in $\boldsymbol{Y}$ are blocked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

d-separation

Definition (d-separation)

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

Example



X_2 and X_4 are not d-separated by $\emptyset$.
X_2 and X_4 are d-separated by X_1.
X_2 and X_4 are not d-separated by X_3.
X_2 and X_4 are d-separated by $\{X_1, X_3\}$. X_2 and X_4 are not d-separated by $\{X_1, X_3, X_5\}$.

Global Markov Property

Theorem

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

$$\boldsymbol{X}, \boldsymbol{Y} \text{ d-separated by } \boldsymbol{Z} \quad \Longrightarrow \quad \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{Z}$$

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

For *cyclic* SCMs, the notion of d-separation is too strong in general. A weaker notion called σ-separation has to be used instead [Forré and Mooij, 2017]. Under additional solvability conditions, a global Markov condition using σ-separation can be shown to hold.

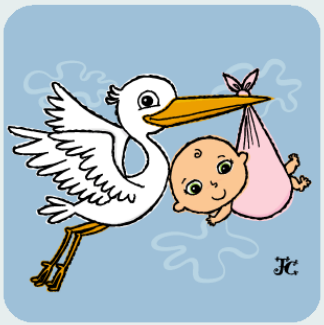
Reichenbach's Principle

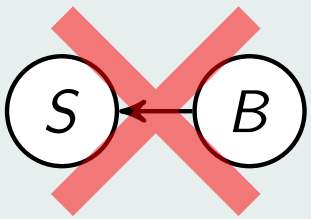
Reichenbach's Principle of Common Cause

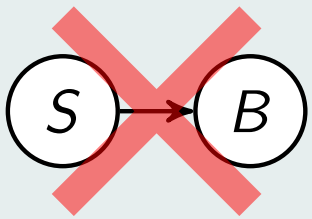
The dependence $X \not\perp\!\!\!\perp Y$ implies that $X \to Y$, $Y \to X$, or $X \leftrightarrow Y$ (or any combination of these three).

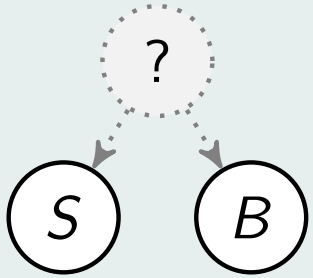
Example

- 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





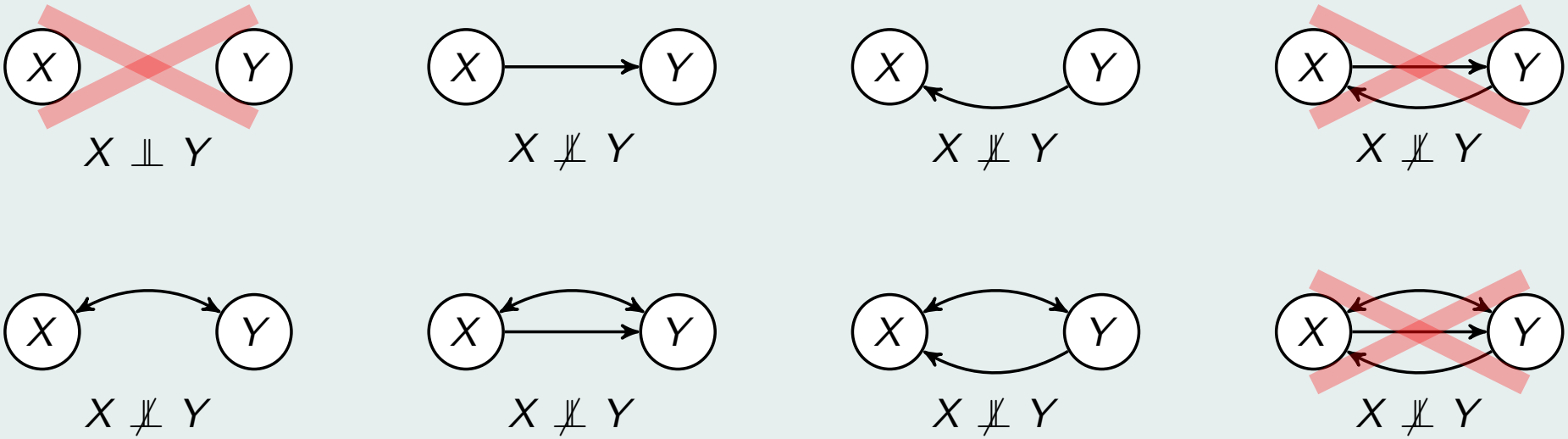




Proof of Reichenbach's Principle

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)

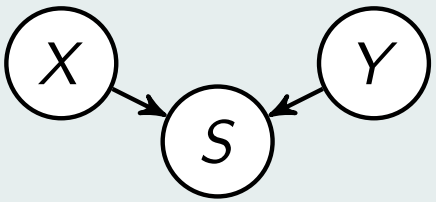
Selection Bias

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: $X \not\perp\!\!\!\perp Y \mid S$.
- 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.

Outline

1. Qualitative Causality: Causal Graphs
2. Quantifying Causality: Structural Causal Models
3. Markov Properties: From Graph to Conditional Independences
4. **Causal Inference: Predicting Causal Effects**

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

Special cases:

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

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

Special cases:

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

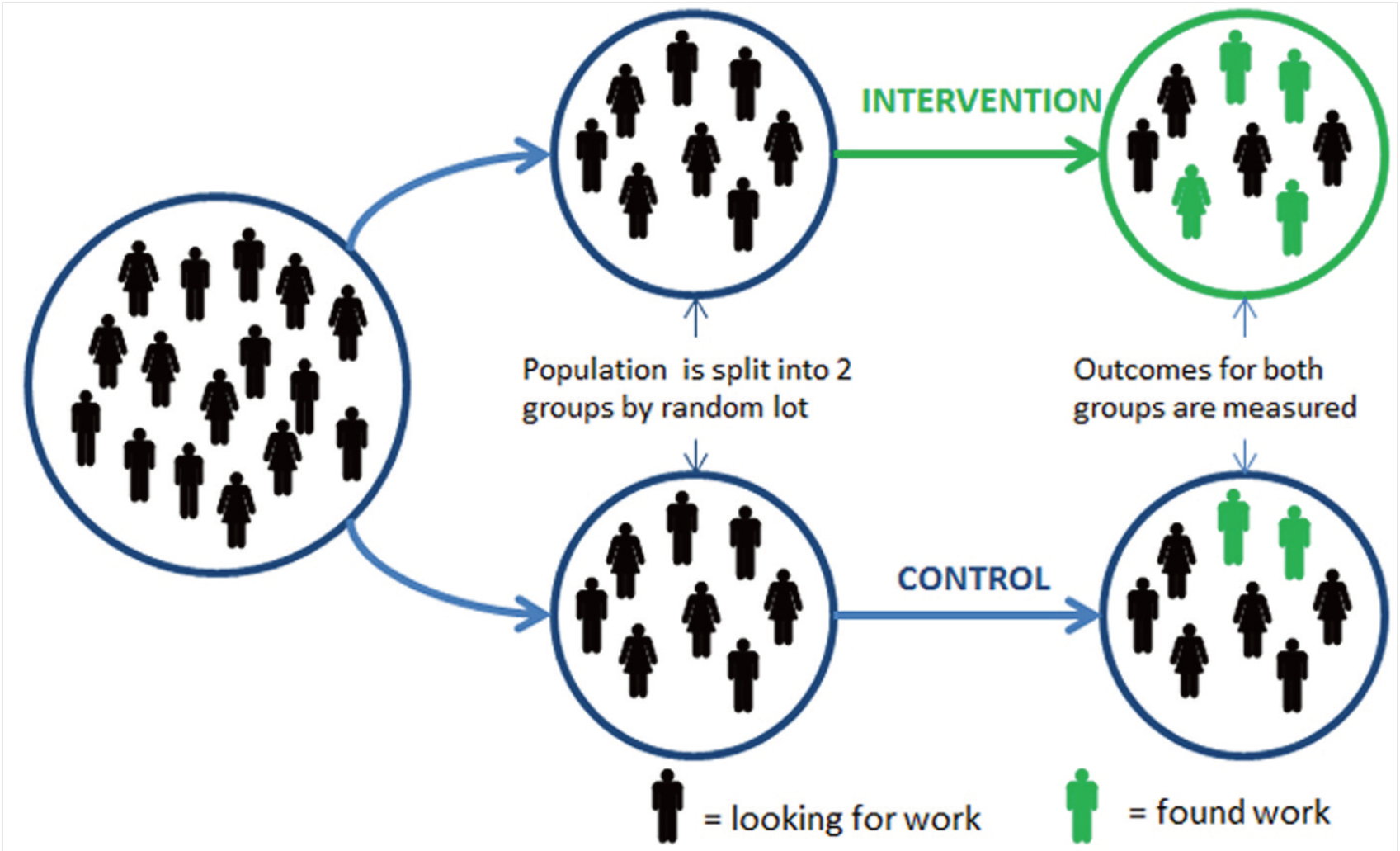
Note: In general, since $p\big(y \mid \mathrm{do}(X = x)\big) \neq p(y \mid X = x)$, we cannot use standard supervised learning (regression, classification) for this task.

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.



Causal Inference for RCT

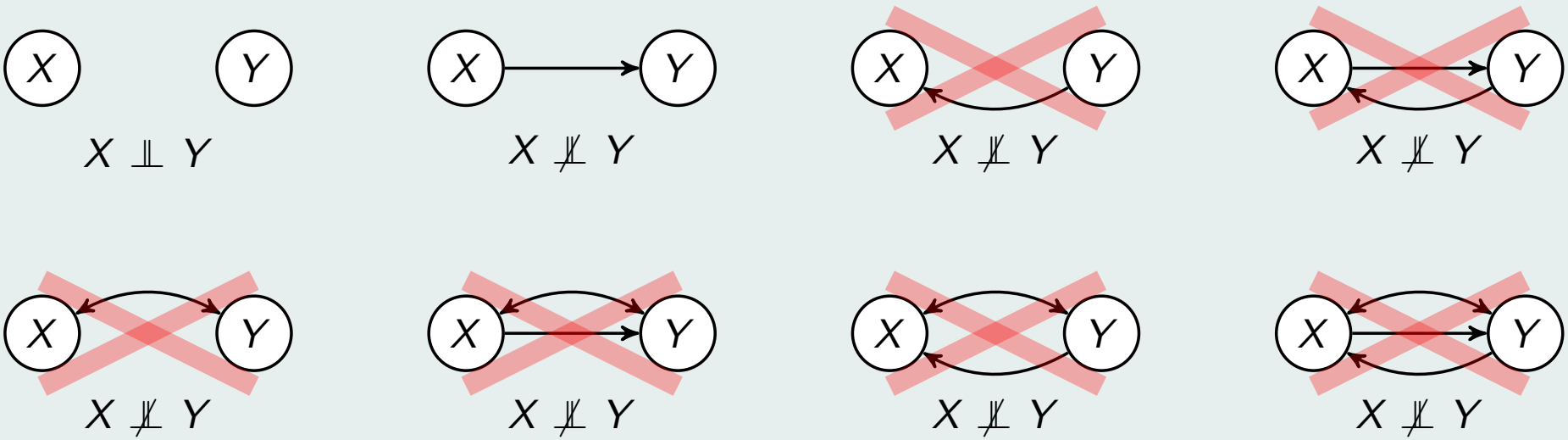
Proposition

The RCT assumptions

- Y does not cause X ($\Longleftarrow$ *X precedes Y in time*)
- Y and X are unconfounded ($\Longleftarrow$ *randomization*)
- no selection bias ($\Longleftarrow$ *study design*)

imply that $X \not\perp\!\!\!\perp Y$ iff X causes Y, and $p\big(Y \mid \mathrm{do}(X = x)\big) = p(y \mid X = x)$.

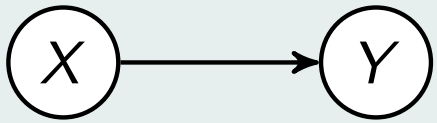
Proof



Identifiability: Example

Can we express $p\big(y \mid \mathrm{do}(X = x)\big)$ in terms of the observational density?

Example



$$p\big(y \mid \mathrm{do}(X = x)\big)$$

$$=$$

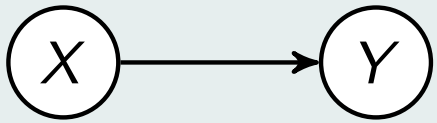
$$p(y \mid X = x)$$

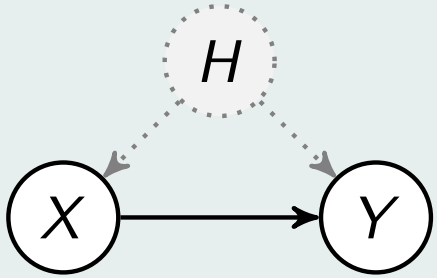
Yes!

Identifiability: Example

Can we express $p\big(y \mid \mathrm{do}(X = x)\big)$ in terms of the observational density?

Example





$$p\big(y \mid \mathrm{do}(X = x)\big) \qquad p\big(y \mid \mathrm{do}(X = x)\big) = \int p(h)p(y \mid x, h)\, dh$$

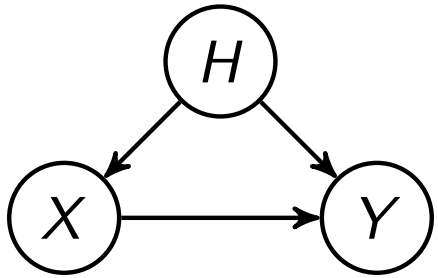
$$= \qquad \neq$$

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

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:

$$\int p(h)p(y \mid x, h)\, dh = p\big(y \mid \mathrm{do}(X = x)\big)$$

More generally, given a causal graph: which covariates $\boldsymbol{H}$ 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 $\boldsymbol{H}$: if

1. $X, Y \notin \boldsymbol{H}$;
2. X *is not an ancestor of any node in* $\boldsymbol{H}$;
3. $\boldsymbol{H}$ *blocks all back-door paths* $X \leftarrow \ldots Y$ *(i.e., all paths between* X *and* Y *that start with an incoming edge on* X*).*

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

$$p\big(y \,|\, \mathrm{do}(X = x)\big) = \int p(y \,|\, x, \boldsymbol{h}) p(\boldsymbol{h})\, d\boldsymbol{h}.$$

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

$$p\big(y \,|\, \mathrm{do}(X = x)\big) = p(y \,|\, x).$$

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**:

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

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

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

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

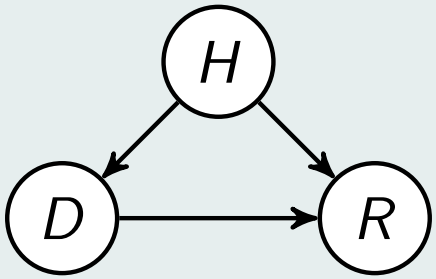
Simpson's paradox only manifests itself if we misinterpret correlation as causation by identifying $p(r \mid D = d)$ with $p\big(r \mid \text{do}(D = d)\big)$.

We should prescribe the drug if
$p\big(R = 1 \mid \text{do}(D = 1)\big) > p\big(R = 1 \mid \text{do(D=0)}\big)$.
How to find the causal effect of the drug on recovery?

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

Back-Door Criterion for Simpson's paradox

Example (Scenario 1)



R: Recovery
D: Took drug
H: Gender

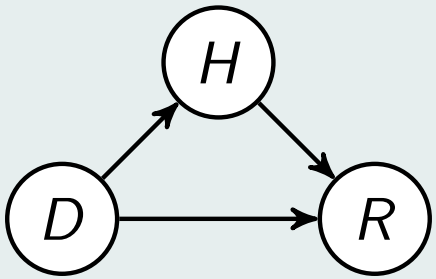
- There is one back-door path: $D \leftarrow H \rightarrow R$, which is blocked by $\{H\}$.
- D is not an ancestor of H.
- Therefore, adjust for $\{H\}$ to obtain causal effect of drug on recovery:

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

- 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:

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

- 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).

Conclusion

In Part I of this tutorial, we have discussed:

- Causal Modeling by means of Structural Causal Models
- Causal Reasoning by means of the Markov Property
- Causal Prediction by means of RCTs and the Back-Door Criterion

Part II of this tutorial will focus on:

- Causal Discovery: how to infer the causal graph from data?

Further reading I

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].

Pearl, J. (1999).

Simpson's paradox: An anatomy.

Technical Report R-264, UCLA Cognitive Systems Laboratory.

Pearl, J. (2000).

Causality: Models, Reasoning, and Inference.

Cambridge University Press.

Further reading II

Pearl, J. (2009).

Causal inference in statistics: An overview.

Statistics Surveys, 3:96–146.

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

Causation, Prediction, and Search.

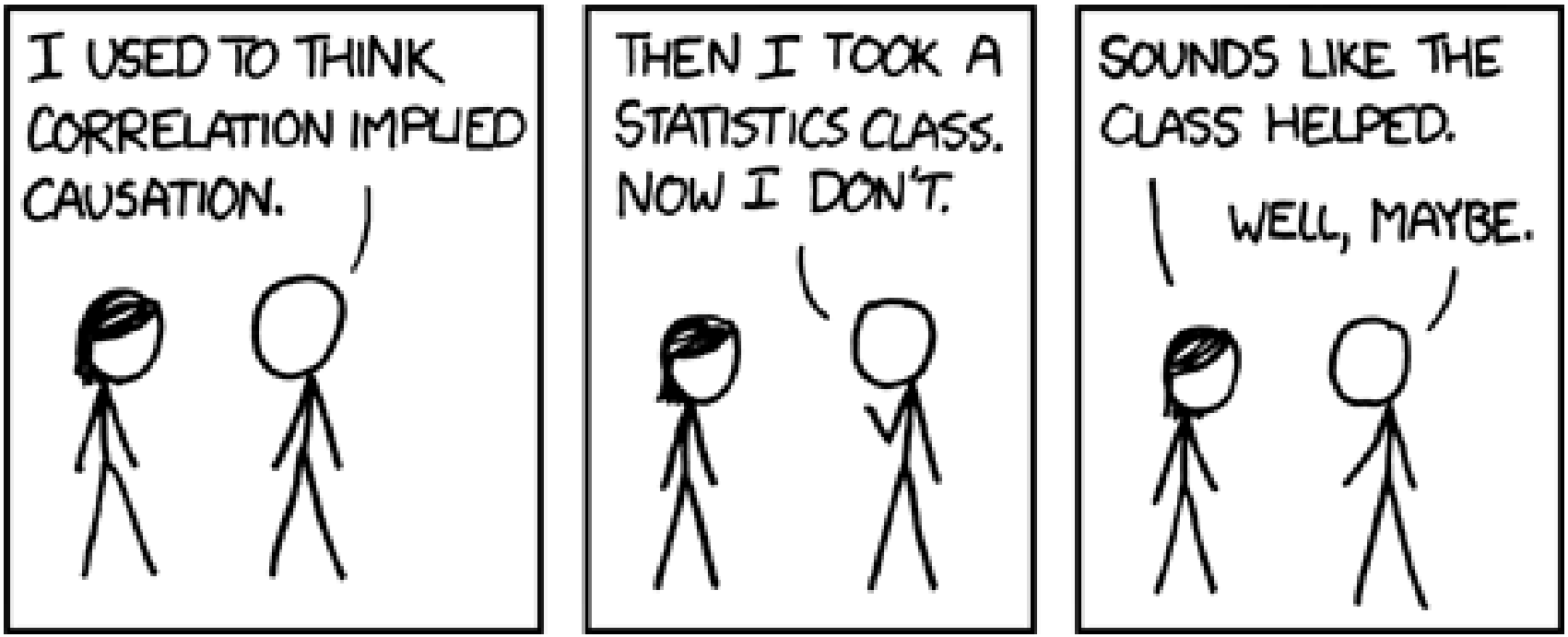
The MIT Press.

Wright, S. (1921).

Correlation and causation.

Journal of Agricultural Research, 20:557–585.

Thank you for your attention!



Randall Munroe, www.xkcd.org