Many questions in science are causal

**Climatology:**

![Graph showing climate change attribution](image)

**Economy:**

![Graph showing states that cut spending and unemployment](image)

**Medicine:**

![Image of medical equipment](image)

**Neuroscience:**

![Brain image](image)
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

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?
Contents

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. Extensions to $\sigma$-separation

7. Large-Scale Validation of Causal Discovery
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 \( X \) 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
- \( X_1 \) 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 \( X_3 \)
- \( X_2 \) causes \( X_1 \)
Let $\mathbf{V} = \{X_1, \ldots, X_N\}$ 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**

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

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

3. $X_1$ causes $X_2$;
   $X_1$ causes $X_2$ directly w.r.t. $\{X_1, X_2, X_3\}$
Each stone causes *all* subsequent stones to topple.
Each stone only *directly causes* the next neighboring stone to topple.

Causal graph:

\[ X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \cdots \rightarrow X_7 \rightarrow X_8 \rightarrow X_9 \]
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.
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.

The intervention changes the causal graph by removing all edges that point towards variables in $X$ (because none of the variables can now cause $X$).
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. $\{X, Y, H\}$
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**
Wealth might confound 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.
We denote latent confounders by **bidirected edges** in the causal graph:

**Example**

\[
\begin{align*}
X \rightarrow Y & \equiv X \rightarrow Y, \\
X \leftrightarrow Y & \equiv X \rightarrow Y, \\
X \rightarrow Y & \equiv X \rightarrow Y,
\end{align*}
\]
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**.
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:

```
Q1   Q2
```

- Cannot be modeled with acyclic causal model!
“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]
“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.” [McArthu, 2014]
A graph $\mathcal{G}$ that consists of directed and bidirected edges is called Directed Mixed Graph (DMG).

If $i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_k$ in $\mathcal{G}$ then $i_1$ is ancestor of $i_k$: $i_1 \in \text{an}_\mathcal{G}(i_k)$.

$\mathcal{G}$ 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).
Contents

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 Extensions to $\sigma$-separation

7 Large-Scale Validation of Causal Discovery
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
- $p(B = 1 \mid A = 1) > p(B = 1 \mid A = 0)$

then $A$ causes $B$.
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:

   \[ p(\text{recovery} | \text{drug}) > p(\text{recovery} | \text{no drug}) \]

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

   \[ p(\text{recovery} | \text{drug}, \text{male}) < p(\text{recovery} | \text{no drug}, \text{male}) \]

   \[ p(\text{recovery} | \text{drug}, \text{female}) < p(\text{recovery} | \text{no drug}, \text{female}) \]

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:

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

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

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!
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
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):

\[ X: \text{ the battery is charged} \]
\[ Y: \text{ the start engine is operational} \]
\[ S: \text{ the car starts} \]

Exogenous variables (latent, independent, binary):

\[ E_X \sim \text{Ber}(0.95) \]
\[ E_Y \sim \text{Ber}(0.99) \]
\[ E_S \sim \text{Ber}(0.999) \]

Structural equations (one per endogenous variable):

\[ X = f_X(E_X) = E_X \]
\[ Y = f_Y(E_Y) = E_Y \]
\[ S = f_S(X, Y, E_S) = X \land Y \land E_S \]
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:

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

Informally:

$$
X_1 = f_1(E_1) \\
X_2 = f_2(E_1, E_2) \\
X_3 = f_3(X_1, X_2, X_5, E_3) \\
X_4 = f_4(X_1, X_4, E_4) \\
X_5 = f_5(X_3, X_4, E_5)
$$

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

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

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

$$X_i = f_i(x_{\text{pa}_i^{\mathcal{I}}}, e_{\text{pa}_i^{\mathcal{J}}})$$

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 $G^a(M)$ of SCM $M$ 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 $G(M)$ of SCM $M$ 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 $\text{pa}_i^{\mathcal{J}} \cap \text{pa}_k^{\mathcal{J}} \neq \emptyset$. 
**Unique Solvability**

**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 $g_\mathcal{O} : X_{(\text{pa}_\mathcal{H}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}} \times \mathcal{E}_{\text{pa}_\mathcal{H}(\mathcal{O}) \cap \mathcal{J}} \rightarrow X_\mathcal{O}$ such that for $\mathbb{P}_\mathcal{E}$-almost every $e$ for all $x \in X$:

$$x_\mathcal{O} = g_\mathcal{O}(x_{(\text{pa}_\mathcal{H}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}}, e_{\text{pa}_\mathcal{H}(\mathcal{O}) \cap \mathcal{J}}) \iff x_\mathcal{O} = f_\mathcal{O}(x, e).$$

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

**Example**

An SCM with structural equations:

\[
\begin{align*}
X_1 &= X_1 \\
X_2 &= X_1 + X_3 \\
X_3 &= X_3 + 1
\end{align*}
\]

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

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

- 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 $i \in I$. 

\[\text{Perfect Interventions}\]
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):

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

Structural equations (one per endogenous variable):

\[ X = E_X \]
\[ Y = E_Y \]
\[ S = X \land Y \land E_S \]
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):

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

Structural equations (one per endogenous variable):

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

\( X = 1 \)
\( Y = E_Y \)
\( S = X \land Y \land E_S \)
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):

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

Structural equations (one per endogenous variable):

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

- $X = E_X$
- $Y = E_Y$
- $S = 0$
Observational (no intervention):

SCM $\mathcal{M}$:

\[
\begin{align*}
X_1 &= f_1(E_1) \\
X_2 &= f_2(E_1, E_2) \\
X_3 &= f_3(X_1, X_2, X_5, E_3) \\
X_4 &= f_4(X_1, X_4, E_4) \\
X_5 &= f_5(X_3, X_4, E_5)
\end{align*}
\]

$E_1 \sim \mathbb{P}_{\xi_1}$

$E_2 \sim \mathbb{P}_{\xi_2}$

$E_3 \sim \mathbb{P}_{\xi_3}$

$E_4 \sim \mathbb{P}_{\xi_4}$

$E_5 \sim \mathbb{P}_{\xi_5}$

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

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

\[
\begin{align*}
X_1 &= f_1(E_1) \\
X_2 &= f_2(E_1, E_2) \\
X_3 &= \xi_3 \\
X_4 &= f_4(X_1, X_4, E_4) \\
X_5 &= f_5(X_3, X_4, E_5)
\end{align*}
\]

$E_1 \sim \mathbb{P}_{\xi_1}$

$E_2 \sim \mathbb{P}_{\xi_2}$

$E_3 \sim \mathbb{P}_{\xi_3}$

$E_4 \sim \mathbb{P}_{\xi_4}$

$E_5 \sim \mathbb{P}_{\xi_5}$

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

**Definition (Reminder)**

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

**Definition**

For \((X, E)\) a solution of SCM \(\mathcal{M}\), we call \(P^X\) 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 \(X\), we denote the density of the observational distribution as \(p_\mathcal{M}(x)\).
A perfect intervention on $\mathcal{M}$ may change the distributions.

**Definition**

We call the family of sets of observational distributions of $\mathcal{M}_{do(X_i=\xi_i)}$ (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 $\mathcal{M}_{do(X_i=\xi_i)}$ 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 $\mathcal{X}$, we denote the density of the interventional distributions as $p_{\mathcal{M}}(x | \text{do}(X_i = \xi_i))$.

**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 $i \in \text{pa}_{\mathcal{M}}(i)$.

**Example (Price-supply-demand)**

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

\[
\begin{align*}
S &= \alpha P + E_S \\
D &= \beta P + E_D \\
P &= P + (S - D)
\end{align*}
\]

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.*
Causal Interpretation of Direct Edges

**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 $\text{do}(X_{\mathcal{I}\setminus\{j\}} = \xi)$ and $\text{do}(X_{\mathcal{I}\setminus\{j\}} = \xi')$ such that $\xi_{\mathcal{I}\setminus\{i,j\}} = \xi'_{\mathcal{I}\setminus\{i,j\}}$ and $\xi_i \neq \xi'_i$ such that

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

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 $i \rightarrow \cdots \rightarrow j \in \mathcal{G}(\mathcal{M})$, 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

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

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 $\text{do}(X_{\mathcal{I}\setminus\{i,j\}} = \xi)$ such that

\[
P_{\mathcal{M}}(X_j \mid \text{do}(i, x_i), \text{do}(\mathcal{I} \setminus \{i,j\}, \xi)) \neq P_{\mathcal{M}}(X_j \mid X_i = x_i, \text{do}(\mathcal{I} \setminus \{i,j\}, \xi))
\]

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 $\mathcal{M}$:

\[
\begin{align*}
X_1 &= f_1(E_1) & E_1 &\sim P_{\varepsilon_1} \\
X_2 &= f_2(E_1, E_2) & E_2 &\sim P_{\varepsilon_2} \\
X_3 &= f_3(X_1, X_2, X_5, E_3) & E_3 &\sim P_{\varepsilon_3} \\
X_4 &= f_4(X_1, X_4, E_4) & E_4 &\sim P_{\varepsilon_4} \\
X_5 &= f_5(X_3, X_4, E_5) & E_5 &\sim P_{\varepsilon_5}
\end{align*}
\]

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

Marginalization $\mathcal{M}\backslash\{2,4\}$:

\[
\begin{align*}
X_1 &= f_1(E_1) & E_1 &\sim P_{\varepsilon_1} \\
X_3 &= f_3(X_1, g_2(E_1, E_2), X_5, E_3) & E_3 &\sim P_{\varepsilon_3} \\
X_5 &= f_5(X_3, g_4(X_1, E_4), E_5) & E_4 &\sim P_{\varepsilon_4} \\
\end{align*}
\]
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 $\mathcal{L}$:

\[
X = f(X, E)
\]

\[
\iff \begin{cases}
X_L = f_L(X_L, X_O, E) \\
X_O = f_O(X_L, X_O, E)
\end{cases}
\]

\[
\iff \begin{cases}
X_L = g_L(X_O, E) \\
X_O = f_O(X_L, X_O, E)
\end{cases}
\]

\[
\iff \begin{cases}
X_L = g_L(X_O, E) \\
X_O = f_O(g_L(X_O, E), X_O, E)
\end{cases}
\]

all hold a.s., where $g_L : \mathcal{X}_O \times \mathcal{E} \to \mathcal{X}_L$ is the explicit solution of the structural equations for $X_L$, i.e.,

\[
X_L = g_L(X_O, E) \iff X_L = f_L(X_L, X_O, E) \text{ a.s.}
\]
Marginalization of an SCM

Definition ([Bongers et al., 2018])

Let $\mathcal{M} = \langle X_\mathcal{I}, E, f, P_E \rangle$ be a simple SCM, $L \subseteq \mathcal{I}$ a subset of endogenous variables and $O = \mathcal{I} \setminus L$. Then the marginalization of $\mathcal{M}$ on $\mathcal{I} \setminus L$ is defined as the SCM $\mathcal{M}^{\setminus L} := \langle X_{\mathcal{I} \setminus L}, E, f^{\setminus L}, P_E \rangle$, where the marginal causal mechanism $f^{\setminus L}$ is obtained by substitution:

$f^{\setminus L}(x_O, e) := f_O(g_L(x_O, e), x_O, e)$.

Definition

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

- $i \to j$ iff there is a directed path $i \to \ell_1 \to \cdots \to \ell_k \to j$ in $\mathcal{G}$ with $\ell_1, \ldots, \ell_k \in L$.
- $i \leftrightarrow j$ iff there is a path $i \leftarrow \ell_1 \leftarrow \cdots \leftarrow \ell_{k_1} \leftrightarrow \ell_{k_1+1} \to \cdots \to \ell_{k_2} \to j$ in $\mathcal{G}$ with $\ell_1, \ldots, \ell_{k_1}, \ldots, \ell_{k_2} \in L$. 
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 $\mathcal{M} = \langle \mathbf{X}_\mathcal{I}, \mathcal{E}, f, \mathbb{P}_\mathcal{E} \rangle$ 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 $\mathbf{X}_{\mathcal{I} \setminus \mathcal{L}}$, 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.

\[ D: \]
\[
\begin{align*}
\dot{X}_i(t) &= f_i(X_{pa(i)}), \\
X_i(0) &= (X_0)_i, \quad i \in I
\end{align*}
\]

intervention \rightarrow \text{do}(X_i = \xi_I)

\[ \mathcal{D}_{do(X_I = \xi_I)}: \]
\[
\begin{align*}
\dot{X}_i(t) &= 0, \\
X_i(0) &= \xi_i, \quad i \in I \\
\dot{X}_i(t) &= f_i(X_{pa(i)}), \\
X_i(0) &= (X_0)_i, \quad i \notin I
\end{align*}
\]

\[ \mathcal{M}_{D}: \]
\[
X_i = X_i + f_i(X_{pa(i)}) \quad i \in I
\]

intervention \rightarrow \text{do}(X_I = \xi_I)

\[ \mathcal{M}_{D_{do(X_I = \xi_I)}}: \]
\[
\begin{align*}
X_i &= \xi_i, \quad i \in I \\
X_i &= X_i + f_i(X_{pa(i)}) \quad i \notin I
\end{align*}
\]
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 \( \mathcal{G}(\mathcal{M}_\mathcal{D}) \):**
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]).
SCMs and other Causal Modeling Frameworks

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

**Definition**

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

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** $\langle \mathcal{G}(\mathcal{M}), p_\mathcal{M} \rangle$. Vice versa, for every Causal Bayesian Network there exists an SCM that induces it.
We can now express “correlation does not imply causation” (or, as Pearl says, “seeing is not doing”) more precisely:

\[ p(y \mid \text{do}(X = x)) \neq p(y \mid X = x) \quad \text{in general} \]

Do we really need to introduce this additional interventional semantics (“the do-operator”) on top of the notion of conditioning that we already are so familiar with in probability theory?

Not necessarily: we can introduce additional variables to get a purely probabilistic model that can mimic the SCM.
Extending an SCM with Intervention Variables

Definition

Given a simple SCM $\mathcal{M}$ with discrete endogenous domains $\mathcal{X}_i$. Define an extended SCM $\hat{\mathcal{M}}$ by (i) for each endogenous variable $X_i$ with $i \in I$, add an endogenous intervention variable $C_i$, taking values in the space $\mathcal{X}_i \cup \{\emptyset\}$; (ii) replace the causal mechanism $f$ by $\hat{f}$ with:

$$\hat{f}_{X_i}(x, c, e) = \begin{cases} 
  c_i & c_i \in \mathcal{X}_i \\
  f_i(x, e) & c_i = \emptyset
\end{cases} \quad \text{("set by perfect intervention")}
$$

and $\hat{f}_{C_i}(\epsilon_i) = \epsilon_i$ where $\epsilon_i \sim P_{C_i}$ with strictly positive density.

Proposition

For any intervention target $I \subseteq \mathcal{I}$ and intervention value $\xi_I \in \mathcal{X}_I$:

$$p_{\mathcal{M}}(x \mid \text{do}(X_I = \xi_I)) = p_{\hat{\mathcal{M}}}(x \mid C_I = \xi_I, C_{\mathcal{I}\setminus I} = \emptyset)$$

All interventional distributions of $\mathcal{M}$ can be obtained by conditioning $p_{\hat{\mathcal{M}}}$. 
Simple SCMs: Overview

SCM → Intervened SCM → Interventional Distribution
SCM → Marginal SCM
SCM → Augmented Graph
Observational Distribution → Graph

(Potential) Direct Causes
(Potential) Causes
(Potential) Confounders
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 (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 $X \perp \perp Y \mid Z$ and say that $X$ is (conditionally) independent from $Y$, given $Z$, if

$$p(x, y \mid Z = z) = p(x \mid Z = z)p(y \mid Z = z).$$

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 $i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_k$.
- If there is a directed path from $X$ to $Y$, $X$ is called an **ancestor** of $Y$.
- The ancestors of $Y$ are denoted $\text{ang}(Y)$, and include $Y$. 

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 $i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_k$.
- If there is a directed path from $X$ to $Y$, $X$ is called an ancestor of $Y$.
- The ancestors of $Y$ are denoted $\text{ang}(Y)$, and include $Y$.

Example

\[
\begin{array}{c}
\text{X}_2 \rightarrow \text{X}_1 \\
\text{X}_3 \rightarrow \text{X}_4 \\
\text{X}_5
\end{array}
\quad
\begin{align*}
\text{X}_1 & \rightarrow \text{X}_3 \leftarrow \text{X}_1 \text{ is not a path.} \\
\text{X}_1 & \leftrightarrow \text{X}_2 \rightarrow \text{X}_3 \text{ is a path.} \\
\text{X}_1 & \rightarrow \text{X}_4 \rightarrow \text{X}_5 \text{ is a directed path.} \\
\text{X}_4 & \rightarrow \text{X}_5 \leftarrow \text{X}_3 \text{ is not a directed path.} \\
\text{The ancestors of } \text{X}_3 \text{ are } \{\text{X}_1, \text{X}_2, \text{X}_3\}.
\end{align*}
\]
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 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.
Colliders and non-colliders

Definition (Colliders)

Let $G$ be a directed mixed graph, and $q$ a path on $G$.

- 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 $G$ be a directed mixed graph. Given a path $q$ on $G$, 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

$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$ is $d$-blocked by $\emptyset$.

$X_3 \rightarrow X_5 \leftarrow X_4$ 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**

Let $G$ be a directed mixed graph. Given a path $q$ on $G$, 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 $G$ be a directed mixed graph. For three sets $X$, $Y$, $Z$ of nodes in $G$, 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 $X \perp_G Y \mid Z$. 

Example

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

$X_3$ 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 $\{X_1, X_5\}$. 
**Definition (d-separation)**

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

**Example**

- $X_3$ and $X_4$ are \textit{d-separated} by $\{X_1\}$.
- $X_3$ and $X_4$ are \textit{d-separated} by $\{X_1, X_2\}$.
- $X_3$ and $X_4$ are not \textit{d-separated} by $\emptyset$.
- $X_3$ and $X_4$ are not \textit{d-separated} by $\{X_1, X_5\}$. 
Please make Exercise 1.2
For an acyclic SCM, the Global Markov Property holds:

\[
X \perp_{\mathcal{G}(\mathcal{M})} Y \mid Z \quad \implies \quad X \perp_{\mathcal{P}_\mathcal{M}} Y \mid Z
\]

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

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

![Diagram showing causality with storks and babies](image)
Assuming that $p(X, Y)$ is generated by an acyclic SCM, we can easily prove Reichenbach’s Principle by applying the Global Markov property:

(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**

\[ X: \text{the battery is charged} \]
\[ Y: \text{the start engine is operational} \]
\[ S: \text{the car starts} \]

- 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.
Faithfulness Assumption

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

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

$$X, Y \perp_{G(\mathcal{M})} Z \implies X \perp_{P_M} Y \mid Z$$

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

**Definition (Faithfulness Assumption)**

For all subsets $X, Y, Z$ of nodes,

$$X, Y \perp_{G(\mathcal{M})} Z \iff X \perp_{P_M} Y \mid Z$$

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

Example (Parameter cancellation)
Consider an SCM $\mathcal{M}$:

\[
\begin{align*}
X &= E_X \\
Y &= X + E_Y \\
Z &= X - Y + E_Z
\end{align*}
\]

Then:

\[
Z \perp\!\!\!\!\perp_{p_{\mathcal{M}}} X \text{ but } Z \not\in_{\mathcal{G}(\mathcal{M})} X.
\]

Example (Deterministic relation)
Consider an SCM $\mathcal{M}$:

\[
\begin{align*}
X &= E_X \\
Y &= X \\
Z &= Y + E_Z
\end{align*}
\]

Then:

\[
Z \perp\!\!\!\!\perp_{p_{\mathcal{M}}} Y \mid X \text{ but } Z \not\in_{\mathcal{G}(\mathcal{M})} Y \mid X.
\]
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 Extensions to $\sigma$-separation

7 Large-Scale Validation of Causal Discovery
One important task ("causal inference") is the prediction of causal effects.

**Definition**

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

**Special cases:**

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

**Note:** In general, since $p(y \mid \text{do}(X = x)) \neq p(y \mid X = x)$, 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)
Experimentation (e.g., Randomized Controlled Trials, A/B-testing, ...) provides the gold standard for causal effect estimation.
If we cannot do experiments... Can we express $p(y \mid \text{do}(X = x))$ in terms of the observational distribution?

**Example**

\[
p(y \mid \text{do}(X = x)) = p(y \mid X = x)
\]

Yes!
Identifiability: Example

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

Example

\[ p(y \mid \text{do}(X = x)) = p(y \mid X = x) \]

Yes!

\[ p(y \mid \text{do}(X = x)) = \int p(h)p(y \mid x, h) \, dh \]

\[ p(y \mid X = x) = \int p(h \mid x)p(y \mid x, h) \, dh \]

No!
We have seen that for the following causal graph,

![Causal Graph](image)

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

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

More generally, given a causal graph: which variables $H$ 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

Let $\mathcal{M}$ be an acyclic SCM $\mathcal{M}$ with disjoint subsets of endogenous variables $\{X\}, \{Y\}, H$. Let $\hat{\mathcal{G}}$ be $\mathcal{G}(\mathcal{M})$ extended with an intervention node $I_X \rightarrow X$. If

1. $H \perp_{\hat{\mathcal{G}}} I_X$;
2. $Y \perp_{\hat{\mathcal{G}}} I_X | \{X\} \cup H$,

then $H$ is called admissible for adjustment to find the causal effect of $X$ on $Y$, i.e., this causal effect is given by:

$$p_\mathcal{M}(y | \text{do}(X = x)) = \int p_\mathcal{M}(y | x, h)p_\mathcal{M}(h) \, dh.$$ 

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

$$p_\mathcal{M}(y | \text{do}(X = x)) = p_\mathcal{M}(y | x).$$
The Back-door Criterion: Example

Example

\[ \mathcal{G}(\mathcal{M}): \]

\[ \hat{\mathcal{G}}: \]

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:

\[
p(x_5 \mid \text{do}(X_2 = x_2)) = \int p(x_5 \mid x_1, x_2)p(x_1) \, dx_1
\]

\[
= \int p(x_5 \mid x_1, x_2, x_4)p(x_1, x_4) \, dx_1 \, dx_4
\]

Some sets of variables that are \textit{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 2.2
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

- $p(R = 1 \mid D = 1)$: the probability that somebody recovers, given the observation that the person took the drug.
- $p(R = 1 \mid \text{do}(D = 1))$: 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 $p(r \mid D = d)$ with $p(r \mid \text{do}(D = d))$.

We should prescribe the drug if

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

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

1. Randomized Controlled Trials
2. Back-door Criterion (requires knowledge of causal graph)
Please make Exercise 2.3
Back-door Criterion for Simpson’s paradox

Example (Scenario 1)

- $I_D \perp H$;
- $I_D \perp R | D, H$;
- Therefore, adjust for $\{H\}$ to obtain causal effect of drug on recovery:

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

- So in scenario 1, 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)

\[ I_D \not\perp H \] (\( H \) is not admissible for adjustment);
\[ I_D \perp R \mid D; \]
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 2, 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).
Pearl formulated three rules (the “do-calculus”) that can be used in addition to the usual rules for probabilistic reasoning. For acyclic SCMs:

1. **Inserting/deleting observations:**

   \[ p(y \mid x, z, \text{do}(w)) = p(y \mid z, \text{do}(w)) \quad \text{if} \quad Y \perp \hat{G}_{\text{do}(W)} X \mid Z \]

2. **Inserting/deleting actions:**

   \[ p(y \mid \text{do}(x), z, \text{do}(w)) = p(y \mid z, \text{do}(w)) \quad \text{if} \quad Y \perp I_X \mid Z. \]

3. **Action/observation exchange:**

   \[ p(y \mid \text{do}(x), z, \text{do}(w)) = p(y \mid x, z, \text{do}(w)) \quad \text{if} \quad Y \perp I_X \mid X, Z \]

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 Extensions to $\sigma$-separation
7 Large-Scale Validation of 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.
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.
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, Gleimour, 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, Gleimour, 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. Extensions to \(\sigma\)-separation
7. Large-Scale Validation of Causal Discovery
Randomized Controlled Trials [Fisher, 1935]

$G$: $Z$ $\rightarrow$ $R$

$G_{RCT}$: $C$ $\rightarrow$ $D$ $\rightarrow$ $R$

$R$: 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 $p(R \mid D = C) = p(R \mid do(D = C))$.
- Gold standard for causal discovery.

All evidence-based medicine is based on this idea.
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 Extensions to $\sigma$-separation

7 Large-Scale Validation of Causal Discovery
Controlled 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))
- ...

Approaches to Causal Discovery from Observational Data
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)
- ...

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

$X \perp Y$, $X \not\perp Y \mid Z$,

$X \not\perp Z$, $X \not\perp Z \mid Y$,

$Y \not\perp Z$, $Y \not\perp Z \mid X$.

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

\[ \text{X} \perp \perp \text{Y}, \text{X} \perp \perp \text{Z}, \text{Y} \perp \perp \text{Z} \]

\[ \text{X} \perp \perp \text{Y} | \text{Z}, \text{Y} \perp \perp \text{Z} | \text{X} \]

\[ \text{Z} \perp \perp \text{X} | \text{Y} \]

\[ \text{Z} \perp \perp \text{X} \]

\[ \text{X} \perp \perp \text{Y} | \text{Z} \]

\[ \text{Y} \perp \perp \text{Z} | \text{X} \]

\[ \text{Y} \perp \perp \text{Z} \]

\[ \text{X} \perp \perp \text{Y} | \text{Z} \]

\[ \text{X} \perp \perp \text{Y} | \text{Z} \]

\[ \text{Y} \perp \perp \text{Z} | \text{X} \]

\[ \text{Y} \perp \perp \text{Z} \]
Please make Exercise 2.4
Causal Discovery from Observational Data: Y-Structure

\[ X_1 \perp X_2 \]
\[ X_1 \not\perp X_2 \mid X_3 \]
\[ X_1 \not\perp X_4 \]
\[ X_1 \perp X_4 \mid X_3 \]
\[ X_2 \not\perp X_4 \]
\[ X_2 \perp X_4 \mid X_3 \]

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.
Causal Discovery from Observational Data: Y-Structure

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. The causal effect of $X_3$ on $X_4$ satisfies $p(x_4 \mid do(X_3 = x_3)) = p(x_4 \mid x_3)$. 

\[ X_1 \perp X_2 \]
\[ X_1 \nparallel X_2 \mid X_3 \]
\[ X_1 \nparallel X_4 \]
\[ X_1 \perp X_4 \mid X_3 \]
\[ X_2 \nparallel X_4 \]
\[ X_2 \perp X_4 \mid X_3 \]

black: $X_3 = 0$, red: $X_3 = 1$
<table>
<thead>
<tr>
<th>$d$</th>
<th>Number of DAGs with $d$ nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>543</td>
</tr>
<tr>
<td>5</td>
<td>29281</td>
</tr>
<tr>
<td>6</td>
<td>3781503</td>
</tr>
<tr>
<td>7</td>
<td>1138779265</td>
</tr>
<tr>
<td>8</td>
<td>783702329343</td>
</tr>
<tr>
<td>9</td>
<td>1213442454842881</td>
</tr>
<tr>
<td>10</td>
<td>4175098976430598143</td>
</tr>
<tr>
<td>11</td>
<td>3160345936418917607425</td>
</tr>
<tr>
<td>12</td>
<td>521939651343829405020504063</td>
</tr>
<tr>
<td>13</td>
<td>18676600744432035186664816926721</td>
</tr>
<tr>
<td>14</td>
<td>1439428141044398334941790719839535103</td>
</tr>
<tr>
<td>15</td>
<td>237725265553410354992180218286376719253505</td>
</tr>
<tr>
<td>16</td>
<td>83756670773733320287699303047996412235223138303</td>
</tr>
<tr>
<td>17</td>
<td>6270792119692388989446452602494921906963551482675201</td>
</tr>
<tr>
<td>18</td>
<td>99421195322159515895228914592354524516555026878588305014783</td>
</tr>
<tr>
<td>19</td>
<td>332771901227107591736177573311261125883583076258421902583546773505</td>
</tr>
</tbody>
</table>

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

Source: [Peters et al., 2017]
State-of-the-art Causal Discovery: (Augmented) FCI


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

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

\( \mathcal{R}1 \) If \( X \leftarrow Z \rightarrow Y \), and \( X \not\perp\!
\!\perp Y \), then \( Z \rightarrow Y \).

\( \mathcal{R}2a \) If \( Z \rightarrow X \rightarrow Y \) and \( Z \leftarrow Y \), then \( Z \rightarrow Y \).

\( \mathcal{R}2b \) If \( Z \rightarrow X \rightarrow Y \) and \( Z \leftarrow Y \), then \( Z \rightarrow Y \).

\( \mathcal{R}3 \) If \( X \leftarrow Z \rightarrow Y \), \( X \leftarrow W \rightarrow Y \), \( X \not\perp\!
\!\perp Y \), and \( W \leftarrow Y \), then \( W \rightarrow Z \).

\( \mathcal{R}4a \) If \( u = (X,\ldots,Z_k,Z,Y) \) is a discriminating path between \( X \) and \( Y \) for \( Z \), and \( Z \leftarrow Y \), then if \( Z \in Sep(X,Y) \), then \( Z \rightarrow Y \).

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

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

\( \mathcal{R}6 \) If \( X \leftarrow Z \rightarrow Y \), then orient as \( Z \rightarrow Y \).

\( \mathcal{R}7 \) If \( X \leftarrow Z \rightarrow Y \), and \( X \not\perp\!
\!\perp Y \), then \( Z \rightarrow Y \).

\( \mathcal{R}8a \) If \( Z \rightarrow X \rightarrow Y \) and \( Z \leftarrow Y \), then \( Z \rightarrow Y \).

\( \mathcal{R}8b \) If \( Z \leftarrow X \rightarrow Y \) and \( Z \leftarrow Y \), then \( Z \rightarrow Y \).

\( \mathcal{R}9 \) If \( Z \leftarrow Y \), \( u = (Z,X,W,\ldots,Y) \) is an uncov. p.d. path, and \( X \not\perp\!
\!\perp Y \), then \( Z \rightarrow Y \).

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

Source: [Claassen & Heskes, 2011]
FCI: Example ("Extended Y-structure")

Independences: $Z \perp U, Z \perp Y \mid X$
FCI: Example ("Extended Y-structure")

Independences: $Z \perp U$, $Z \perp Y \mid X$
FCI: Example ("Extended Y-structure")

Independences: \( Z \perp U, Z \perp Y \mid X \)
FCI: Example ("Extended Y-structure")

Independences: $Z \perp U, Z \perp Y \mid X$

$\mathcal{R}0a$

$\mathcal{R}0b$

$\mathcal{R}1$
FCI: Example ("Extended Y-structure")

Independences: $Z \perp U$, $Z \perp Y \mid X$

\[ R_{0a} \]
\[ R_{0b} \]
\[ R_{1} \]
\[ R_{2b} \]
FCI: Example ("Extended Y-structure")

Indepedences: \( Z \perp \perp U, Z \perp \perp Y \mid X \)
Local Causal Discovery: simple causal discovery algorithm (Cooper, 1997).

**Definition**

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

$$Y \notin \text{an}(X) \land Z \notin \text{an}(X) \land X \not\perp Y \land Y \not\perp Z \land X \perp Z \mid Y,$$

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:

$$\begin{align*}
X &\rightarrow Y \rightarrow Z \\
X &\leftrightarrow Y \rightarrow Z \\
X &\rightarrow Y \leftrightarrow Z \\
X &\rightarrow Y \rightarrow Z
\end{align*}$$

Therefore, $Y \in \text{an}(Z)$ and $p(Z \mid \text{do}(Y = y)) = p(Z \mid Y = y)$. 
Please make Exercise 2.5
**Protein Abundance Data:**
(Sachs *et al.*, 2005)

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

**Causal Graph:**
("Signalling network")

(depicted here: "consensus" network)
Question: What is the causal relation between Raf and Mek?

Each dot is a measurement in a single human immune system cell.

- Raf: abundance of phosphorylized Raf
- Mek: abundance of phosphorylized Mek
- blue = baseline,
  red = reagent U0126 added

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 phosphorylized Raf
- Mek: abundance of phosphorylized Mek

- blue = baseline,
  red = reagent U0126 added

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

*Hint: U0126 inhibits Mek.*
Causal Discovery by Experimentation: Example

Each dot is a measurement in a single human immune system cell
- Raf: abundance of phosphorylized Raf
- Mek: abundance of phosphorylized 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.)
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?
**LCD: Example**

- **pErk**: abundance of phosphorylized Erk in each cell
- **pS6**: abundance of phosphorylized S6 in cell
- **I**: green = baseline, red = PMA-IONO activator added

\[ (X, Y, Z) \] is LCD triplet iff:

- \( Y \not\in \text{an}(X) \)
- \( Z \not\in \text{an}(X) \)
- \( X \perp Y \)
- \( Y \perp Z \)
- \( X \perp Z \mid Y \)

**What is the causal relation?**
LCD: Example

- $pErk$: abundance of phosphorylized Erk in each cell
- $pS6$: abundance of phosphorylized S6 in cell
- $I$: green = baseline, red = PMA-IONO activator added

What is the causal relation? LCD triplet $(I, pS6, pErk)$, so $pS6 \rightarrow pErk$.

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

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

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

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 [Mooij et al., 2019], a framework for causal discovery, that achieves this.
Randomized Controlled Trials, or A/B-testing

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 ($C = 0$):

<table>
<thead>
<tr>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2</td>
</tr>
<tr>
<td>0.6</td>
</tr>
<tr>
<td>-1.7</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

Drug ($C = 1$):

<table>
<thead>
<tr>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.3</td>
</tr>
<tr>
<td>1.8</td>
</tr>
<tr>
<td>-0.1</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

Two-sample test:

Is $p(x \mid \text{do}(C = 0)) = p(x \mid \text{do}(C = 1))$?

(b) Pooled data

<table>
<thead>
<tr>
<th>$C$</th>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.2</td>
</tr>
<tr>
<td>0</td>
<td>0.6</td>
</tr>
<tr>
<td>0</td>
<td>-1.7</td>
</tr>
<tr>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>-0.3</td>
</tr>
<tr>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>1</td>
<td>-0.1</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

Independence test:

Is $X \perp \perp C$?
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).
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 $C \perp\!\!\!\!\perp X$, then $C$ causes $X$ (correlation implies causation).

Proof

\[ C \perp\!\!\!\!\perp X \]
\[ C \not\perp\!\!\!\!\perp X \]
\[ C \not\perp\!\!\!\!\perp X \]
JCI: Two types of variables

**Definition**

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

- **Context variables** \( \{ C_i \}_{i \in I} \) that model the context of the system,
- **System** variables \( \{ X_j \}_{j \in J} \) 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

**Context variables:**

- \( C_1 \): back-to-work program
- \( C_2 \): psychotherapy

<table>
<thead>
<tr>
<th></th>
<th>no interventions:</th>
<th>only back-to-work program:</th>
<th>only psychotherapy:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>0.1</td>
<td>0.2</td>
<td>0.53</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0.13</td>
<td>0.21</td>
<td>0.52</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0.23</td>
<td>0.21</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_1 )</td>
<td>0.2</td>
<td>0.22</td>
<td>0.53</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0.23</td>
<td>0.21</td>
<td>0.52</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0.92</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_1 )</td>
<td>0.5</td>
<td>1.2</td>
<td>0.95</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0.61</td>
<td>1.21</td>
<td>0.90</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0.55</td>
<td>1.19</td>
<td>0.97</td>
</tr>
</tbody>
</table>
After explicitly adding the context variables, we pool the data:

**Example**

<table>
<thead>
<tr>
<th>Variables</th>
<th>no interventions:</th>
<th>only psychotherapy:</th>
<th>only back-to-work program:</th>
<th>both interventions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.19</td>
<td>0.22</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.22</td>
<td>0.21</td>
<td>0.55</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.2</td>
<td>0.19</td>
<td>0.22</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.18</td>
<td>0.21</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>1.19</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.5</td>
<td>0.52</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>0.49</td>
<td>0.51</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>0.51</td>
<td>0.99</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

**System variables:**
- $X_1$: salary
- $X_2$: drug abuse
- $X_3$: depression

**Context variables:**
- $C_1$: back-to-work program
- $C_2$: psychotherapy
JCI Assumptions (Intuitive formulation)

We are modelling a generic setting in which the experimenter decides on the performed interventions \emph{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 $G$ that includes both system variables $\{X_1, \ldots, X_p\}$ and context variables $\{C_1, \ldots, C_d\}$, which jointly models the experimental design and the system in \emph{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.
**Question:** How can we now reconstruct the causal graph from the data?

**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

Ancestral causal relations

(4 system variables, 500 samples in each data set)
(4 system variables, 500 samples in each data set)
(4 system variables, 500 samples in each data set)
Evaluation on real-world flow cytometry data

Only observational data:

All (observational+interventional) data:
Given the importance of the Markov property, the first thing we need is a Markov property for cyclic SCMs. We introduced a notion \( \sigma \)-separation that generalizes d-separation [Forr´e and Mooij, 2017]:

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

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

**Theorem ([Forr´e and Mooij, 2017])**

For a simple SCM \( \mathcal{M} \), the generalized directed global Markov property holds for its observational distribution \( \mathbb{P}_\mathcal{M}(\mathbf{X}) \):

\[
\mathbf{X}_{A} \indep \mathbf{X}_{B} \mid \mathbf{X}_{Z} \quad \iff \quad \mathbf{X}_{A} \indep \mathbf{X}_{B} \mid \mathcal{I},
\]

\( A, B, Z \subseteq \mathcal{I} \).
Markov properties: $\sigma$-separation

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

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

$$i_1 \xleftarrow{} \cdots \xrightarrow{} i_n$$

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 $i_{k-1} \xrightarrow{} i_k \xleftarrow{} i_{k+1}$ with $i_k \notin \text{an}_\mathcal{G}(Z)$, or
- it contains a non-collider with $i_k \in Z$:

$$i_{k-1} \xleftarrow{} i_k \rightarrow i_{k+1}, \quad i_{k-1} \rightarrow i_k \xleftarrow{} i_{k+1},$$

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

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

**SCM $\mathcal{M}$:**

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

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

\[
\begin{align*}
X_1 &\perp^d X_3 \mid X_2, X_4 \\
\text{but} & \\
X_1 &\not\perp^\sigma X_3 \mid X_2, X_4
\end{align*}
\]

Indeed, as one can check explicitly, $X_1 \not\perp^p_{\mathcal{M}} X_3 \mid X_2, X_4$.

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

Stronger statements can be derived for special cases:

Theorem ([Forr´e and Mooij, 2017])

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

1. $\mathcal{M}$ is linear and its exogenous variables have a density with respect to Lebesgue measure, or
2. all endogenous variables are discrete-valued, or
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 $\mathcal{G}(\mathcal{M})$:

$$
\begin{align*}
A \not\perp_{\mathcal{G}(\mathcal{M})} B \mid Z & \quad \Longrightarrow \quad X_A \not\perp_{\mathcal{P}_{\mathcal{M}}} X_B \mid X_Z \\
A, B, Z & \subseteq \mathcal{I}.
\end{align*}
$$
By simply replacing d-separation with $\sigma$-separation, it turns out that one can directly extend the applicability from acyclic SCMs to (possibly cyclic) simple SCMs of:

- The Back-door Criterion [Forréd and Mooij, 2019];
- The do-Calculus [Forréd and Mooij, 2019];

Causal Discovery algorithms can be adapted, or turn out to need no modification:

- [Forréd and Mooij, 2018]: the first causal discovery algorithm that can handle cycles, nonlinear relationships, latent confounding variables and data from different (interventional) contexts.
- LCD, Y-structures, FCI and JCI all work out-of-the-box on simple SCMs [Mooij et al., 2019]
Contents

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. Extensions to $\sigma$-separation
7. Large-Scale Validation of Causal Discovery
Gene Regulatory Network = Causal Graph

Source: [Kemmeren et al., 2014]
observational: (wild-type vs. wild-type):

large-scale single gene knockout micro-array data [Kemmeren et al., 2014]:
- \(\sim 6,500\) variables (gene expression)
- \(\sim 260\) observational samples (wild-type vs. wild-type)
- \(\sim 1,500\) interventional samples (single-gene knockouts/knockdowns)

interventional: (mutant vs. wild-type):

challenge: can we, in a purely data-driven way (without using biological knowledge), predict which genes strongly change their expression when we knock-out a given gene (without using any data corresponding to that particular knock-out experiment)?
Causal Discovery of Gene Regulatory Networks

Large-scale Single Gene Knockout Micro-Array Data [Kemmeren et al., 2014]:
- \(\sim 6,500\) variables (gene expression)
- \(\sim 260\) observational samples (wild-type vs. wild-type)
- \(\sim 1,500\) interventional samples (single-gene knockouts/knockdowns)

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

### Observational:

- **Train:**
  - ~6,000 genes
  - ~250 samples

- **Test:**
  - 

### Interventional:

- **Train:**
  - ~6,000 genes
  - ~1,500 knockouts

- **Test:**
  - 

First successful validation of causal discovery

**ICP:** [Meinshausen et al., 2016]; **LCD:** high-dimensional version of LCD
Correlation: Causation or Confounding?

True positive:

False positive:

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

From random differential equations to structural causal models: the stochastic case.

Theoretical aspects of cyclic structural causal models.

Markov properties for graphical models with cycles and latent variables.
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)*.

Causal calculus in the presence of cycles, latent confounders and selection bias.
In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence (UAI-19)*.

Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors.

*Cell, 157*:740–752.


Methods for causal inference from gene perturbation experiments and validation.

From ordinary differential equations to structural causal models: the deterministic case.

Joint causal inference from multiple contexts.

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


Thank you for your attention!

Randall Munroe, www.xkcd.org