servicenow

UNIVERSITÉ LAVAL

Mind the structure

An introduction to causal inference for machine learning

Alexandre Drouin

July 11, 2022

Outline

Part I: Introduction to causal inference

1. **What** is causal inference?
2. **Why** should you care about causality?
3. **How?** A graphical framework

Outline

Part I: Introduction to causal inference

1. **What** is causal inference?
2. **Why** should you care about causality?
3. **How?** A graphical framework

➔ This is a primer for the practical session that will happen this afternoon.

Outline

Part I: Introduction to causal inference

1. **What** is causal inference?
2. **Why** should you care about causality?
3. **How?** A graphical framework

➜ This is a primer for the practical session that will happen this afternoon.

Part II: Research

4. Causal discovery
5. Other tasks and work in progress

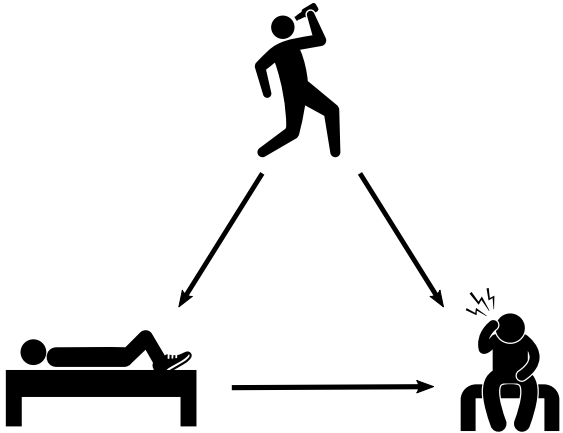
Part I

Introduction to causal inference

Correlation does not imply causation

Credit: Brady Neal

Correlation does not imply causation



Credit: Brady Neal

What is causal inference?

- **In short:** using data to estimate the effect of actions
  - Effect of a treatment on a disease
  - Effect of interest rates on inflation
  - Effect of acting according to some policy (e.g., public policy, RL)

More accurately

Causal inference is a field of study that explores how:

1. Data
2. Assumptions about this data

can be combined to answer certain types of questions (and which cannot be answered).

What is causal inference?

- **In short:** using data to estimate the effect of actions
  - Effect of a treatment on a disease
  - Effect of interest rates on inflation
  - Effect of acting according to some policy (e.g., public policy, RL)

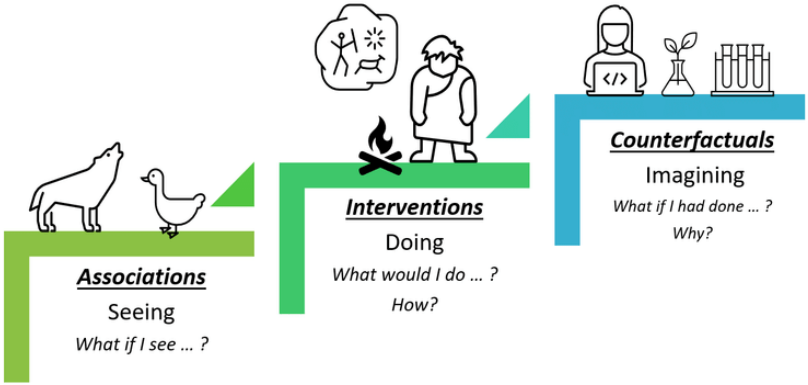
More accurately

Causal inference is a field of study that explores how:

1. Data
2. Assumptions about this data

can be combined to answer certain types of questions (and which cannot be answered).

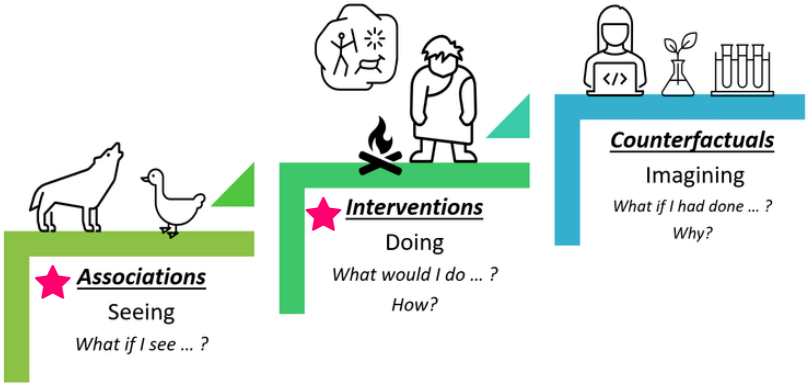
Different types of questions



Credit: Alycia N. Carey and Xintao Wu

Some questions are harder to answer than others

Different types of questions

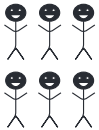


Credit: Alycia N. Carey and Xintao Wu

Today, we will focus on estimating the effect of interventions.

Interventional quantities for decision-making

- **Average treatment effect:** $\mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid \ do(A = a')]$
  - Effect of wearing masks on COVID-19 contaminations
  - Effect of a new user interface on customer satisfaction

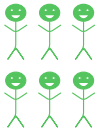


- **Conditional average treatment effect:** $\mathbb{E}[Y \mid X, do(A = a)] - \mathbb{E}[Y \mid \ X, do(A = a')]$
  - Effect on sales ($) of sending e-mail discounts to customers that are older than 30 yo.
  - Effect on survival of a treatment for individuals with a particular genotype.
  - Contextual bandits

- **Individualized treatment effect:** pushing the specificity of X to the limit.
  - Personalized medicine

Interventional quantities for decision-making

- **Average treatment effect:** $\mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid \ do(A = a')]$
  - Effect of wearing masks on COVID-19 contaminations
  - Effect of a new user interface on customer satisfaction

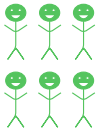


- **Conditional average treatment effect:** $\mathbb{E}[Y \mid X, do(A = a)] - \mathbb{E}[Y \mid \ X, do(A = a')]$
  - Effect on sales ($) of sending e-mail discounts to customers that are older than 30 yo.
  - Effect on survival of a treatment for individuals with a particular genotype.
  - Contextual bandits

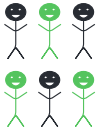
- **Individualized treatment effect:** pushing the specificity of X to the limit.
  - Personalized medicine

Interventional quantities for decision-making

- **Average treatment effect:** $\mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid \ do(A = a')]$
  - Effect of wearing masks on COVID-19 contaminations
  - Effect of a new user interface on customer satisfaction



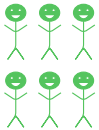
- **Conditional average treatment effect:** $\mathbb{E}[Y \mid X, do(A = a)] - \mathbb{E}[Y \mid \ X, do(A = a')]$
  - Effect on sales ($) of sending e-mail discounts to customers that are older than 30 yo.
  - Effect on survival of a treatment for individuals with a particular genotype.
  - Contextual bandits



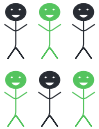
- **Individualized treatment effect:** pushing the specificity of X to the limit.
  - Personalized medicine

Interventional quantities for decision-making

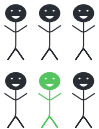
- **Average treatment effect:** $\mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid \ do(A = a')]$
  - Effect of wearing masks on COVID-19 contaminations
  - Effect of a new user interface on customer satisfaction



- **Conditional average treatment effect:** $\mathbb{E}[Y \mid X, do(A = a)] - \mathbb{E}[Y \mid \ X, do(A = a')]$
  - Effect on sales ($) of sending e-mail discounts to customers that are older than 30 yo.
  - Effect on survival of a treatment for individuals with a particular genotype.
  - Contextual bandits



- **Individualized treatment effect:** pushing the specificity of X to the limit.
  - Personalized medicine



Why should you care about causality?

A motivational example

Recovery rate of patients with kidney stones (after 5 months)

Treatment	E[$Y \mid A$]
Yes ($A = 1$)	78% (273/250)
No ($A = 0$)	**83%** (289/350)

Example adapted from Julious & Mullee [1994]

A motivational example

Recovery rate of patients with kidney stones (after 5 months)

Treatment	E[$Y \mid A$]
Yes ($A = 1$)	78% (273/250)
No ($A = 0$)	**83%** (289/350)

Example adapted from Julious & Mullee [1994]

Overall: The treatment seems detrimental.

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = -5\%$$

A motivational example

Recovery rate of patients with kidney stones (after 5 months)

Treatment	E[Y | A]	E[Y | A, Z = mild]	E[Y | A, Z = severe]
Yes (A = 1)	78% (273/250)	**93%** (81/87)	**73%** (192/263)
No (A = 0)	**83%** (289/350)	87% (234/270)	69% (55/80)

Example adapted from Julious & Mullee [1994]

Mild illness: the treatment is beneficial.

$$\mathbb{E}[Y \mid A = 1, Z = \text{mild}] - \mathbb{E}[Y \mid A = 0, Z = \text{mild}] = 6\%$$

A motivational example

Recovery rate of patients with kidney stones (after 5 months)

Treatment	E[Y | A]	E[Y | A, Z = mild]	E[Y | A, Z = severe]
Yes (A = 1)	78% (273/250)	**93%** (81/87)	**73%** (192/263)
No (A = 0)	**83%** (289/350)	87% (234/270)	69% (55/80)

Example adapted from Julious & Mullee [1994]

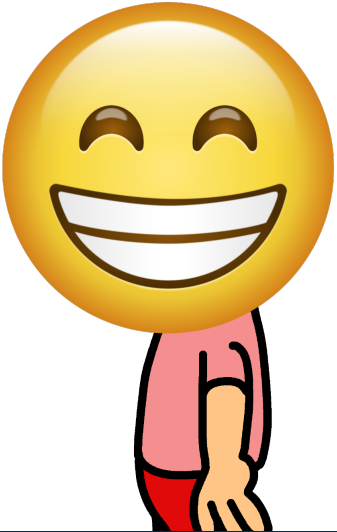
Severe illness: the treatment is beneficial.

$$\mathbb{E}[Y \mid A = 1, Z = \text{severe}] - \mathbb{E}[Y \mid A = 0, Z = \text{severe}] = 4\%$$

Doctors act according to a policy

Doctors act according to a policy

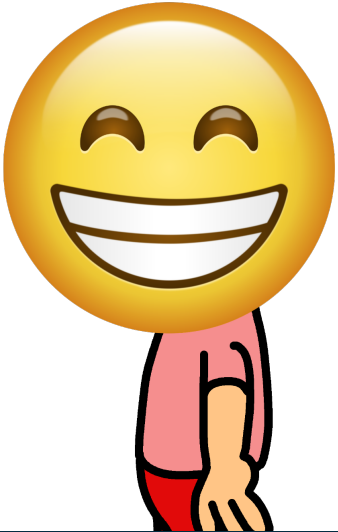
$Z = \text{mild}$



Doctors act according to a policy

You'll be fine, $A = 0$.

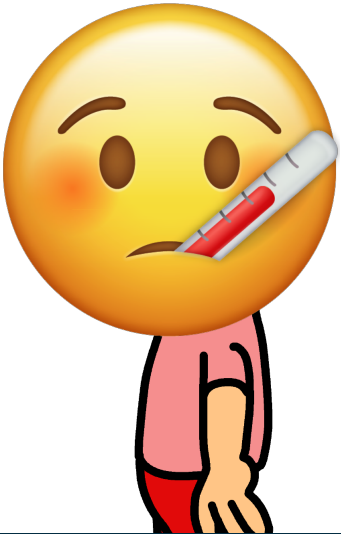
$Z = \text{mild}$



Doctors act according to a policy

Doctors act according to a policy

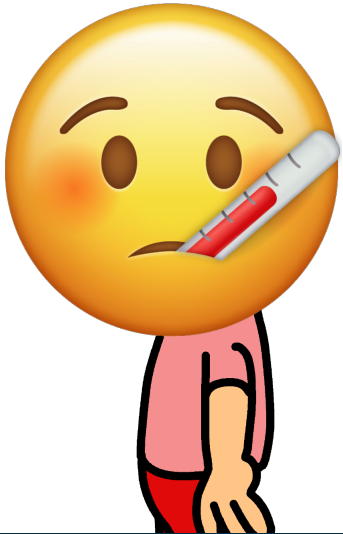
$Z = \text{severe}$



Doctors act according to a policy

You need treatment, $A = 1$!

$Z =$ severe

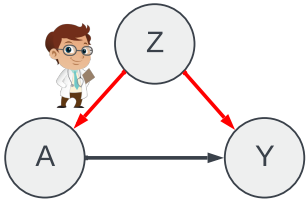


The $A \rightarrow Y$ relationship is confounded

Treatment: $A \in \{0, 1\}$

Recovery: $Y \in \{0, 1\}$

Severity: $Z \in \{\text{mild}, \text{severe}\}$



Treatment	E[Y | A]	E[Y | A, Z = mild]	E[Y | A, Z = severe]
Yes (A = 1)	78% (273/250)	**93%** (81/87)	**73%** (192/263)
No (A = 0)	**83%** (289/350)	87% (234/270)	69% (55/80)

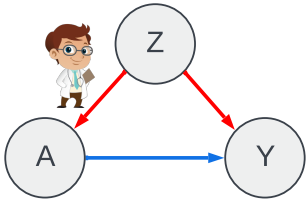
Confounding: patients with a severe illness are more likely to receive treatment and also more likely to have a bad outcome.

The $A \rightarrow Y$ relationship is confounded

Treatment: $A \in \{0, 1\}$

Recovery: $Y \in \{0, 1\}$

Severity: $Z \in \{\text{mild}, \text{severe}\}$



Treatment	E[Y | A]	E[Y | A, Z = mild]	E[Y | A, Z = severe]
Yes (A = 1)	78% (273/250)	**93%** (81/87)	**73%** (192/263)
No (A = 0)	**83%** (289/350)	87% (234/270)	69% (55/80)

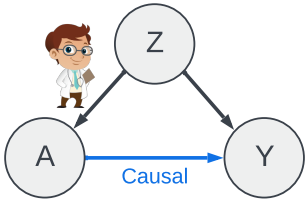
Confounding: patients with a severe illness are more likely to receive treatment and also more likely to have a bad outcome.

The $A \rightarrow Y$ relationship is confounded

Treatment: $A \in \{0, 1\}$

Recovery: $Y \in \{0, 1\}$

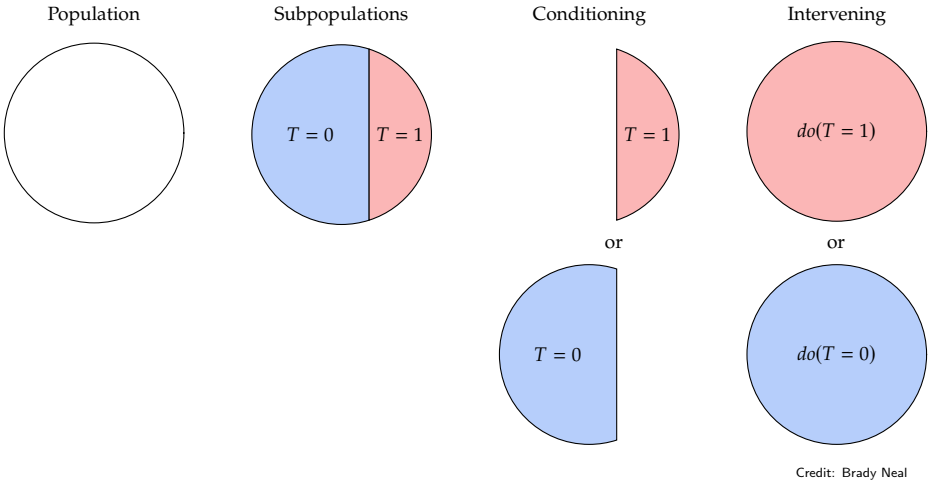
Severity: $Z \in \{\text{mild}, \text{severe}\}$



Treatment	E[Y | A]	E[Y | A, Z = mild]	E[Y | A, Z = severe]
Yes (A = 1)	78% (273/250)	**93%** (81/87)	**73%** (192/263)
No (A = 0)	**83%** (289/350)	87% (234/270)	69% (55/80)

Confounding: patients with a severe illness are more likely to receive treatment and also more likely to have a bad outcome.

Illustration of the difference between conditioning and intervening



Credit: Brady Neal

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working?
- How many issues can we expect to solve if I make 100 employees work today?
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big?

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working?
- How many issues can we expect to solve if I make 100 employees work today?
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big?

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working? associative
- How many issues can we expect to solve if I make 100 employees work today?
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big?

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working? associative
- How many issues can we expect to solve if I make 100 employees work today?
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big?

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working? associative
- How many issues can we expect to solve if I make 100 employees work today? interventional
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big?

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working? associative
- How many issues can we expect to solve if I make 100 employees work today? interventional
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big?

Quiz time: are these questions associative or interventional?

Suppose that you manage a customer support center and want to study its efficiency:

- What is the expected number of issues solved per hour if we have 100 employees working? associative
- How many issues can we expect to solve if I make 100 employees work today? interventional
- Given that we solved 1000 issues and had 100 employees working today, how many could we have solved with a team twice as big? counterfactual

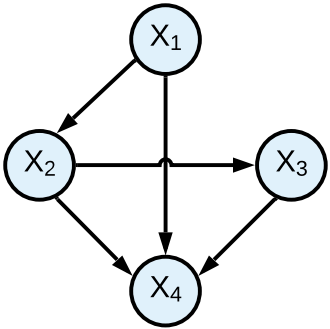
How? A graphical framework

Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $P(X) = \prod_{i=1}^{d} P(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$

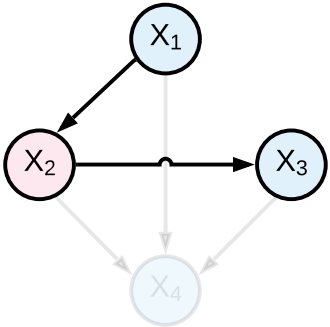
Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $P(X) = \prod_{i=1}^{d} P(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$



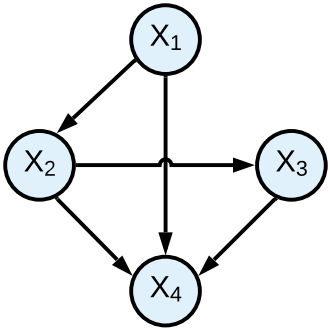
Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $P(X) = \prod_{i=1}^{d} P(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$



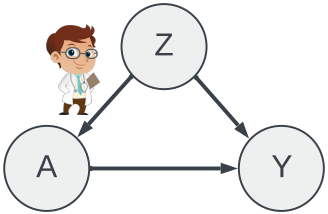
Causal Bayesian networks

- Random vector $X = (X_1, ..., X_d)$
- Let $\mathcal{G}$ be a directed acyclic graph (DAG)
  - d vertices (one per X_i)
  - edges indicate causal relationships
- Encodes (conditional) independence constraints (via *d-separation*, see Koller & Friedman [2009])
- Distribution P_X: $P(X) = \prod_{i=1}^{d} P(X_i | X_{\pi_i^{\mathcal{G}}})$, where $\pi_i^{\mathcal{G}}$ = parents of i in $\mathcal{G}$



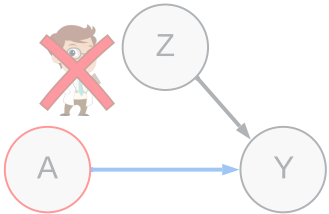
Interventions: graph surgery

Observations



$$P(a, y, z) = P(z)\,P(a \mid z)\,P(y \mid a, z)$$

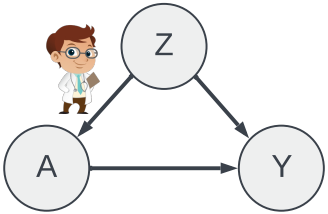
Intervention



$$P'(a, y, z) = P(z)\,P'(a)\,P(y \mid a, z)$$

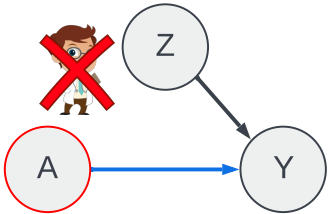
Interventions: graph surgery

Observations



$$P(a, y, z) = P(z)\, P(a \mid z)\, P(y \mid a, z)$$

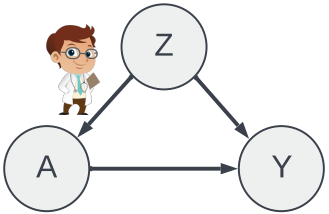
Intervention



$$P'(a, y, z) = P(z)\, P'(a)\, P(y \mid a, z)$$

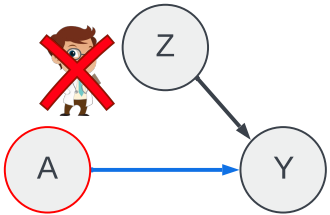
Interventions: graph surgery

Observations



$$P(a, y, z) = P(z)\,P(a \mid z)\,P(y \mid a, z)$$

Intervention

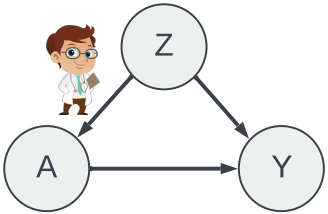


$$P'(a, y, z) = P(z)\,P'(a)\,P(y \mid a, z)$$

⚠ **Important:** notice how conditionals that are not under intervention are invariant across distributions (a.k.a, modularity/autonomy).

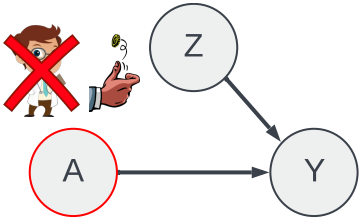
Interventions: graph surgery

Observations



$$P(a, y, z) = P(z)\, P(a \mid z)\, P(y \mid a, z)$$

Intervention



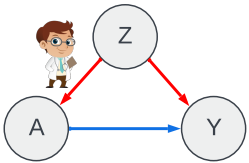
$$P'(a, y, z) = P(z)\, P'(a)\, P(y \mid a, z)$$

Randomization is one way that can be used to obtain such a graph

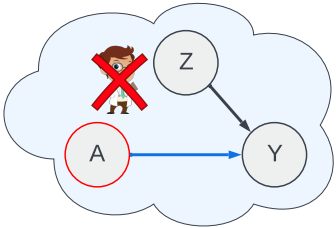
Causal inference from observational data

- **Objective:** estimate the effect of an intervention: $\mathbb{E}[Y \mid do(A = a)]$
  - ↳ randomization is not always possible
    e.g., life threatening, detrimental to the economy, etc.

Observational distribution



Interventional distribution

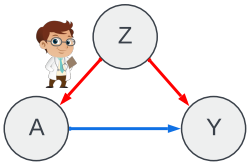


- **How?** Identification: transform a causal estimand (with $do(.)$) into a purely statistical one (no $do(.)$).

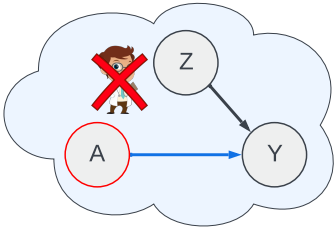
Causal inference from observational data

- **Objective:** estimate the effect of an intervention: $\mathbb{E}[Y \mid do(A = a)]$
  ↳ randomization is not always possible
  e.g., life threatening, detrimental to the economy, etc.

Observational distribution



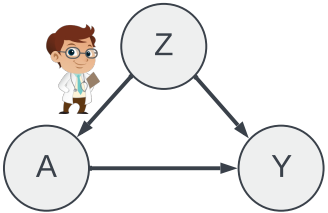
Interventional distribution



- **How?** Identification: transform a causal estimand (with $do(.)$) into a purely statistical one (no $do(.)$).

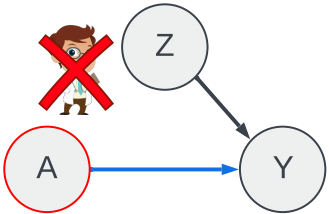
Interventions: truncated factorization

Observations



$$P(a, y, z) = P(z)\, P(a \mid z)\, P(y \mid a, z)$$

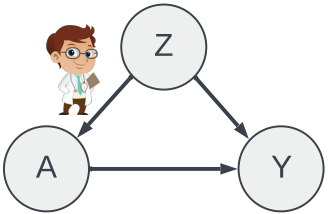
Intervention



$$P(a, y, z|do(A = a')) = P(z)\, \delta_{a=a'}\, P(y \mid a, z)$$

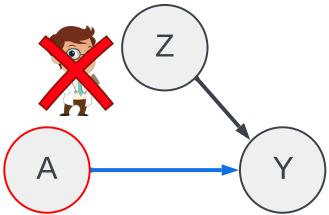
Interventions: truncated factorization

Observations



$$P(a, y, z) = P(z)\, P(a \mid z)\, P(y \mid a, z)$$

Intervention



$$P(a, y, z | do(A = a')) = P(z)\, \delta_{a=a'}\, P(y \mid a, z)$$

Truncated factorization: the general expression for such interventional distributions in CBNs

$$P(x_1, \ldots, x_d \mid do(X_i = x_i')) = \prod_{j=1; j \neq i}^{d} P(x_j \mid x_{\pi_j^{\mathcal{G}}}) \cdot \delta_{x_i = x_i'}$$

Identification: parent adjustment (or standardization)

$$P(y \mid do(A = a')) = \sum_a \sum_z P(a, y, z \mid do(A = a'))$$

$$< \text{Truncated factorization} >$$

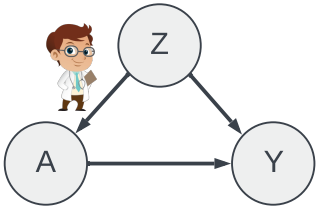
$$= \sum_a \sum_z P(y \mid a', z) \cdot P(z) \cdot \delta_{a=a'}$$

$$< \text{Density is zero for all } a \neq a' >$$

$$= 0 + \sum_z P(y \mid a', z) \cdot P(z) \cdot 1$$

$$< \text{Cleaning up a bit} >$$

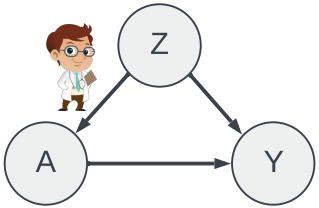
$$= \sum_z P(y \mid a', z) \cdot P(z)$$



$$ATE(a, a') = \mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid do(A = a')]$$

$$= \sum_z \Big[\mathbb{E}[Y \mid a, z] - \mathbb{E}[Y \mid a', z] \Big] P(z)$$

Identification: parent adjustment (or standardization)



$$P(y \mid do(A = a')) = \sum_a \sum_z P(a, y, z \mid do(A = a'))$$

$$< \text{Truncated factorization} >$$

$$= \sum_a \sum_z P(y \mid a', z) \cdot P(z) \cdot \delta_{a=a'}$$

$$< \text{Density is zero for all } a \neq a' >$$

$$= 0 + \sum_z P(y \mid a', z) \cdot P(z) \cdot 1$$

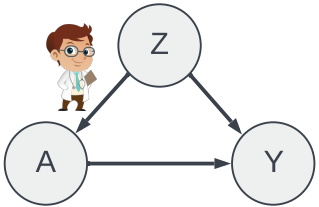
$$< \text{Cleaning up a bit} >$$

$$= \sum_z P(y \mid a', z) \cdot P(z)$$

$$ATE(a, a') = \mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid do(A = a')]$$

$$= \sum_z \Big[\mathbb{E}[Y \mid a, z] - \mathbb{E}[Y \mid a', z] \Big] P(z)$$

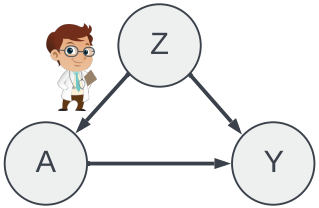
Identification: parent adjustment (or standardization)



$$
\begin{aligned}
P(y \mid do(A = a')) &= \sum_a \sum_z P(a, y, z \mid do(A = a')) \\
&< \text{Truncated factorization} > \\
&= \sum_a \sum_z P(y \mid a', z) \cdot P(z) \cdot \delta_{a=a'} \\
&< \text{Density is zero for all } a \neq a' > \\
&= 0 + \sum_z P(y \mid a', z) \cdot P(z) \cdot 1 \\
&< \text{Cleaning up a bit} > \\
&= \sum_z P(y \mid a', z) \cdot P(z)
\end{aligned}
$$

$$
\begin{aligned}
ATE(a, a') &= \mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid do(A = a')] \\
&= \sum_z \Big[\mathbb{E}[Y \mid a, z] - \mathbb{E}[Y \mid a', z] \Big] P(z)
\end{aligned}
$$

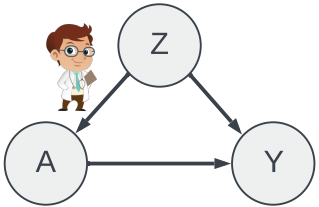
Identification: parent adjustment (or standardization)



$$
\begin{aligned}
P(y \mid do(A = a')) &= \sum_a \sum_z P(a, y, z \mid do(A = a')) \\
&< \text{Truncated factorization} > \\
&= \sum_a \sum_z P(y \mid a', z) \cdot P(z) \cdot \delta_{a=a'} \\
&< \text{Density is zero for all } a \neq a' > \\
&= 0 + \sum_z P(y \mid a', z) \cdot P(z) \cdot 1 \\
&< \text{Cleaning up a bit} > \\
&= \sum_z P(y \mid a', z) \cdot P(z)
\end{aligned}
$$

$$
\begin{aligned}
ATE(a, a') &= \mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid do(A = a')] \\
&= \sum_z \Big[\mathbb{E}[Y \mid a, z] - \mathbb{E}[Y \mid a', z] \Big] P(z)
\end{aligned}
$$

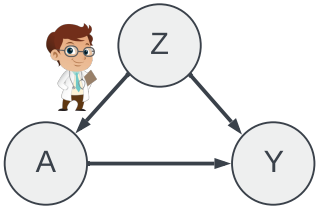
Identification: parent adjustment (or standardization)



$$
\begin{aligned}
P(y \mid do(A = a')) &= \sum_a \sum_z P(a, y, z \mid do(A = a')) \\
&< \text{Truncated factorization} > \\
&= \sum_a \sum_z P(y \mid a', z) \cdot P(z) \cdot \delta_{a=a'} \\
&< \text{Density is zero for all } a \neq a' > \\
&= 0 + \sum_z P(y \mid a', z) \cdot P(z) \cdot 1 \\
&< \text{Cleaning up a bit} > \\
&= \sum_z P(y \mid a', z) \cdot P(z)
\end{aligned}
$$

$$
\begin{aligned}
ATE(a, a') &= \mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid do(A = a')] \\
&= \sum_z \Big[\mathbb{E}[Y \mid a, z] - \mathbb{E}[Y \mid a', z] \Big] P(z)
\end{aligned}
$$

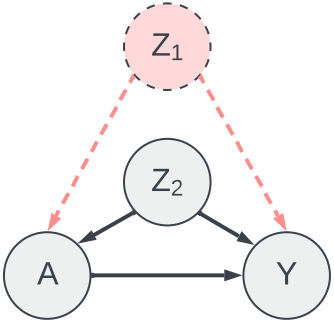
Identification: parent adjustment (or standardization)



$$
\begin{aligned}
P(y \mid do(A = a')) &= \sum_a \sum_z P(a, y, z \mid do(A = a')) \\
&< \text{Truncated factorization} > \\
&= \sum_a \sum_z P(y \mid a', z) \cdot P(z) \cdot \delta_{a=a'} \\
&< \text{Density is zero for all } a \neq a' > \\
&= 0 + \sum_z P(y \mid a', z) \cdot P(z) \cdot 1 \\
&< \text{Cleaning up a bit} > \\
&= \sum_z P(y \mid a', z) \cdot P(z)
\end{aligned}
$$

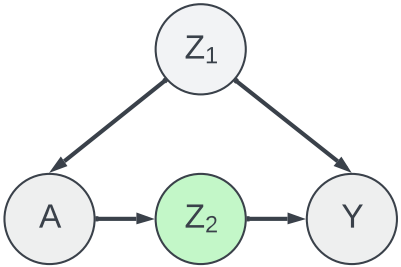
$$
\begin{aligned}
ATE(a, a') &= \mathbb{E}[Y \mid do(A = a)] - \mathbb{E}[Y \mid do(A = a')] \\
&= \sum_z \Big[\mathbb{E}[Y \mid a, z] - \mathbb{E}[Y \mid a', z] \Big] P(z)
\end{aligned}
$$

Identification: parent adjustment (unmeasured confounder)



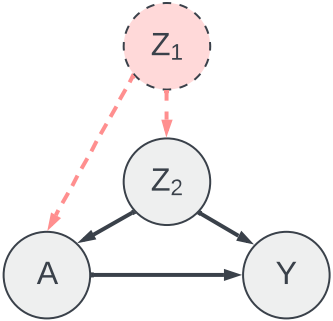
Not identifiable! (and unverifiable)

Identification: don't just adjust for any variable



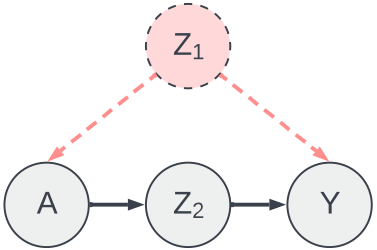
Conditioning on a mediator blocks the $A \to Y$ path!

Identification: back-door adjustment



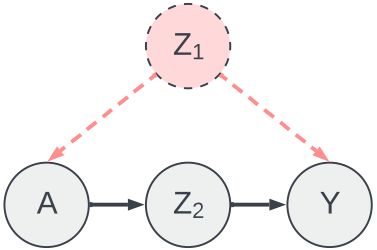
$$P(y \mid do(A = a')) = \sum_{z_2} P(y \mid a', z_2) \cdot P(z_2)$$

Identification: front-door adjustment



$$P(y \mid do(A = a')) = \sum_{z_2} P(z_2 \mid a') \cdot \sum_{a} P(y \mid a, z_2) \cdot P(a)$$

Identification: front-door adjustment



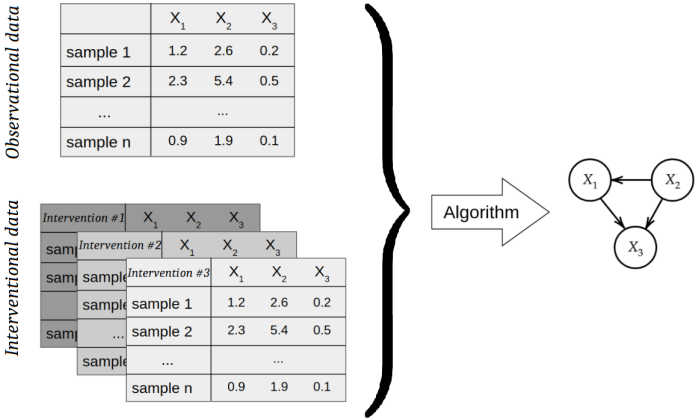
$$P(y \mid do(A = a')) = \sum_{z_2} P(z_2 \mid a') \cdot \sum_{a} P(y \mid a, z_2) \cdot P(a)$$

General case: do-calculus [Pearl, 2009]

Part II

Research

Causal discovery: finding causal relationships in data



Sacrifices to make: assumptions

To make this possible, we need to make assumptions. Some common ones include:

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Sacrifices to make: assumptions

To make this possible, we need to make assumptions. Some common ones include:

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Sacrifices to make: assumptions

To make this possible, we need to make assumptions. Some common ones include:

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Sacrifices to make: assumptions

To make this possible, we need to make assumptions. Some common ones include:

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

Sacrifices to make: assumptions

To make this possible, we need to make assumptions. Some common ones include:

- **Causal sufficiency:** no hidden confounding variables
- **Markov property:** *d-separation* in the graph implies *conditional independence*

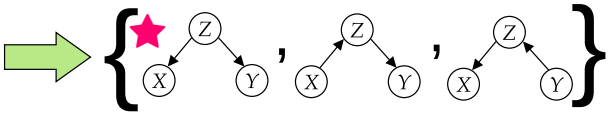
  $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp X_2 \,|\, Z$
- **Faithfulness:** *conditional independence* implies *d-separation* in the graph

  $X_1 \perp\!\!\!\perp X_2 \,|\, Z \Longrightarrow X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 \,|\, Z$

These last two assumptions guarantee an equivalence between properties of the data and properties of the graph

Challenge: uncertainty in the graph structure

X	Y	Z
1.21	1.58	0.33
1.50	1.84	0.51
.	.	.
.	.	.
.	.	.
0.96	1.07	0.11

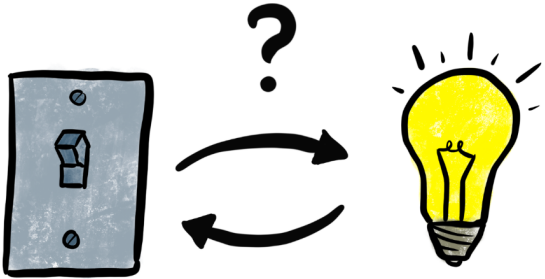


Without making more assumptions, observational data only allows identification up to a Markov equivalence class (MEC) [Verma & Pearl, 1991]

Assumptions on variable types can improve identification

Brouillard, P., Taslakian, P., Lacoste, A., Lachapelle, S., & Drouin, A. (2022).
Typing assumptions improve identification in causal discovery. CLeaR 2022.

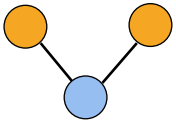


Credit: Philippe Brouillard

Assumptions on variable types can improve identification

Brouillard, P., Taslakian, P., Lacoste, A., Lachapelle, S., & Drouin, A. (2022).
Typing assumptions improve identification in causal discovery. CLeaR 2022.



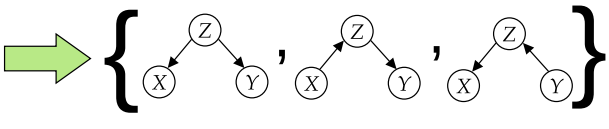
Idea: attribute a type to each variable and constrain how members of different types can interact (*type consistency*)

Assumptions on variable types can improve identification

Brouillard, P., Taslakian, P., Lacoste, A., Lachapelle, S., & Drouin, A. (2022).
Typing assumptions improve identification in causal discovery. CLeaR 2022.




X	Y	Z
1.21	1.58	0.33
1.50	1.84	0.51
⋮	⋮	⋮
0.96	1.07	0.11



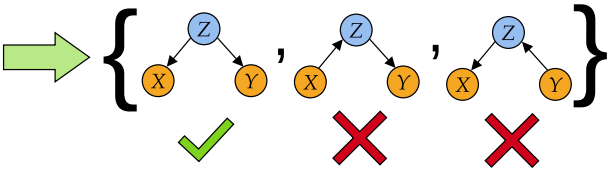
Contribution: we adapt DAGs, MECs, algorithms to include types

Assumptions on variable types can improve identification

Brouillard, P., Taslakian, P., Lacoste, A., Lachapelle, S., & Drouin, A. (2022).
Typing assumptions improve identification in causal discovery. CLeaR 2022.



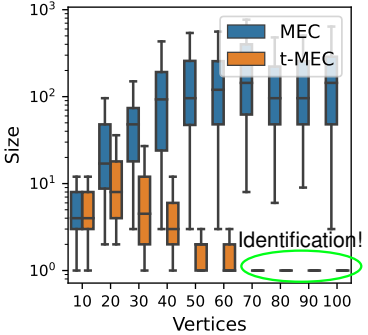

X	Y	Z
1.21	1.58	0.33
1.50	1.84	0.51
.	.	.
.	.	.
.	.	.
0.96	1.07	0.11



tMEC: type-based relationships invalidate some Markov-equivalent graphs

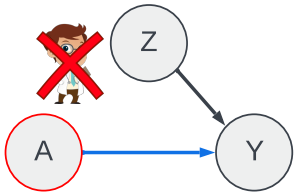
Assumptions on variable types can improve identification

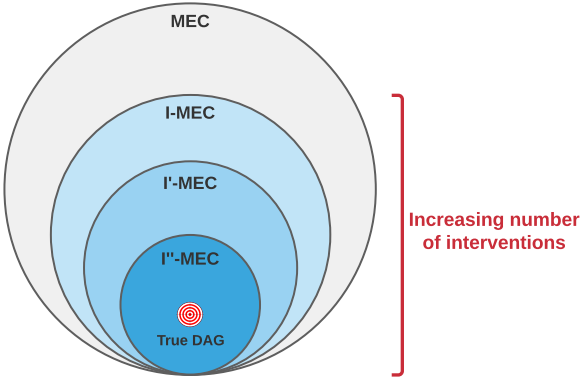
Brouillard, P., Taslakian, P., Lacoste, A., Lachapelle, S., & Drouin, A. (2022).
Typing assumptions improve identification in causal discovery. CLeaR 2022.



Theorem: under some conditions (including a fixed number of types), the size of the t-MEC goes to 1 exponentially fast as the number of vertices increases.

Interventions can also reduce uncertainty





- Interventions reveal invariances: I-MEC $\subseteq$ MEC (see Eberhardt et al. [2005])
- **Example:** gene knockout/knockdown experiments in biology [Dixit et al., 2016]

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020). *Differentiable causal discovery from interventional data*. NeurIPS 2020.




- **Score-based causal discovery:** find the DAG that maximizes a score function ($\mathcal{S}$)
  - E.g., data likelihood + sparsity prior
  - Consistency: need to demonstrate that the score leads to the true solution

$$\hat{\mathcal{G}} \in \arg\max_{\mathcal{G} \in DAG} \mathcal{S}(\mathcal{G})$$

- **Problem:** search space grows superexponentially with the number of variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020). *Differentiable causal discovery from interventional data*. NeurIPS 2020.




- **Score-based causal discovery:** find the DAG that maximizes a score function ($\mathcal{S}$)
  - E.g., data likelihood + sparsity prior
  - Consistency: need to demonstrate that the score leads to the true solution

$$\hat{\mathcal{G}} \in \operatorname*{arg\,max}_{\mathcal{G} \in DAG} \mathcal{S}(\mathcal{G})$$

- **Problem:** search space grows superexponentially with the number of variables

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020).
Differentiable causal discovery from interventional data. NeurIPS 2020.




$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

$\downarrow$ Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$, with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right]$$

$\downarrow$ Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020).
Differentiable causal discovery from interventional data. NeurIPS 2020.




$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

$\downarrow$ Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$, with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right]$$

$\downarrow$ Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020).
Differentiable causal discovery from interventional data. NeurIPS 2020.




$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

$\downarrow$ Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$, with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0 \right]$$

$\downarrow$ Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020).
Differentiable causal discovery from interventional data. NeurIPS 2020.




$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

$\downarrow$

Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$,
with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda ||G||_0 \right]$$

$\downarrow$

Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020).
Differentiable causal discovery from interventional data. NeurIPS 2020.




$$\mathcal{S}_{I^*}(G) := \sup_{\phi} \sum_{k=1}^{K} \mathbb{E}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0$$

$\downarrow$

Relaxation where $G_{ij} \sim \text{Bernoulli}(\sigma(\Lambda_{ij}))$,
with $\sigma(\cdot) :=$ sigmoid function

$$\hat{\mathcal{S}}_{I^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{G \sim \sigma(\Lambda)} \left[\sum_{k=1}^{K} \mathop{\mathbb{E}}_{X \sim p^{(k)}} \log f^{(k)}(X; G, I^*, \phi) - \lambda \|G\|_0 \right]$$

$\downarrow$

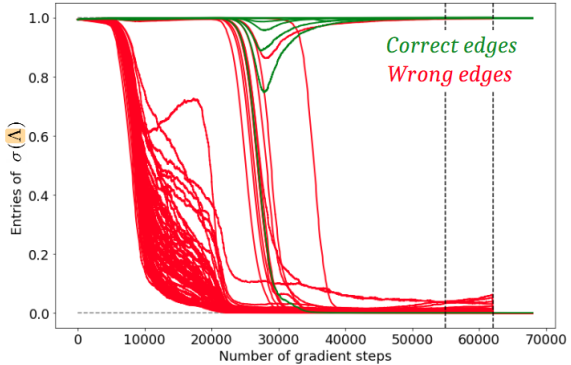
Optimize for Λ under acyclicity constraint

$$\sup_{\Lambda} \hat{\mathcal{S}}_{I^*}(\Lambda) \quad s.t. \quad \underbrace{\text{Tr}\left(e^{\sigma(\Lambda)}\right) - d = 0}_{\text{Acyclicity constraint}}$$

[Zheng et al., 2018]

Differentiable causal discovery with interventions

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020).
Differentiable causal discovery from interventional data. NeurIPS 2020.



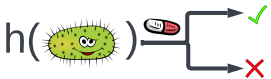
Probability of correct and incorrect edges w.r.t. learning step

Other tasks and work in progress

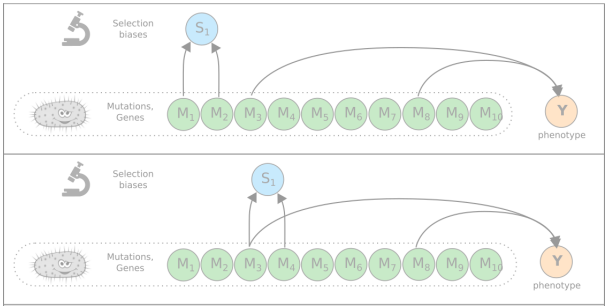
Learning robust models



- **Goal:** learn robust predictors of drug resistance in bacteria (multiple hospitals)



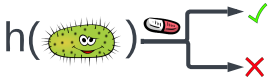
- **How:** use data from multiple environments and try to find invariances



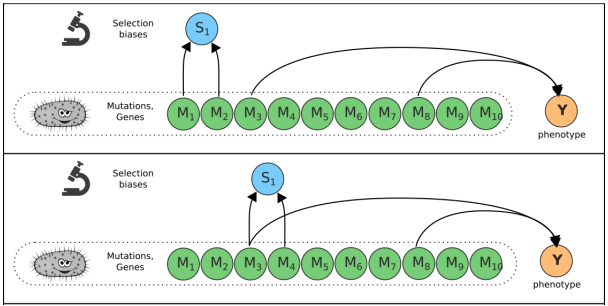
Learning robust models



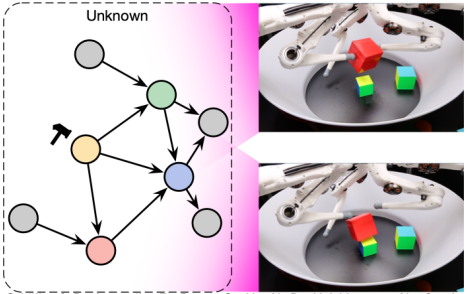
- **Goal:** learn robust predictors of drug resistance in bacteria (multiple hospitals)



- **How:** use data from multiple environments and try to find invariances



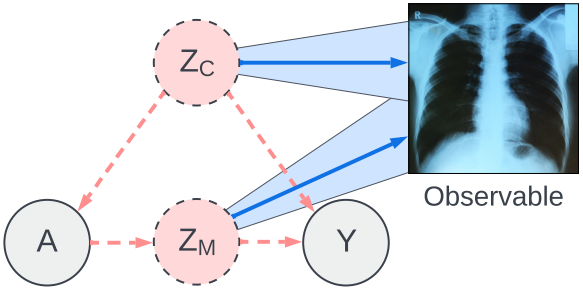
Causal representation learning



Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). *Toward causal representation learning*. Proceedings of the IEEE, 109(5), 612-634.

Given perceptual data, recover the state of an unknown causal system.

Causal representation learning (work in progress)



Recover and correctly segment latent confounders and mediators.

Conclusion

- Causal inference: estimating the effect of actions from data
- Need to be careful when using ML for decision making
- There exists a number of tasks where:
  - ML can help CI
  - CI can help ML
- Some interesting tasks that I didn't mention:
  - Causal reinforcement learning (use observational data)
  - Causal fairness (counterfactuals)

Thank you!

Conclusion

- Causal inference: estimating the effect of actions from data
- Need to be careful when using ML for decision making
- There exists a number of tasks where:
  - ML can help CI
  - CI can help ML
- Some interesting tasks that I didn't mention:
  - Causal reinforcement learning (use observational data)
  - Causal fairness (counterfactuals)

Thank you!

Conclusion

- Causal inference: estimating the effect of actions from data
- Need to be careful when using ML for decision making
- There exists a number of tasks where:
  - ML can help CI
  - CI can help ML
- Some interesting tasks that I didn't mention:
  - Causal reinforcement learning (use observational data)
  - Causal fairness (counterfactuals)

Thank you!

Conclusion

- Causal inference: estimating the effect of actions from data
- Need to be careful when using ML for decision making
- There exists a number of tasks where:
  - ML can help CI
  - CI can help ML
- Some interesting tasks that I didn't mention:
  - Causal reinforcement learning (use observational data)
  - Causal fairness (counterfactuals)

Thank you!

References

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7), 1853–1866.e17.

Eberhardt, F., Glymour, C., & Scheines, R. (2005). On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05 (pp. 178–184). Arlington, Virginia, USA: AUAI Press.

Julious, S. A. & Mullee, M. A. (1994). Confounding and simpson's paradox. *Bmj*, 309(6967), 1480–1481.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. MIT Press.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Verma, T. & Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department.

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems 31*.