# Machine Learning meets Biostatistics II

## A crash course on Causal Inference

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

September, 2022 @ CCAIM Summer School

More information: `http://www.statslab.cam.ac.uk/~qz280/teaching`.

1. Randomization and potential outcomes
2. Undirected graphical models
3. Directed acyclic graphical (DAG) models
4. Causal DAGs
5. Why does causality matter?

# Outline

1. Randomization and potential outcomes

2. Undirected graphical models

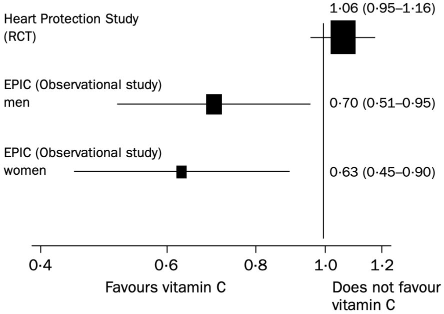
3. Directed acyclic graphical (DAG) models

4. Causal DAGs

5. Why does causality matter?

# Motivating example: Vitamin studies.

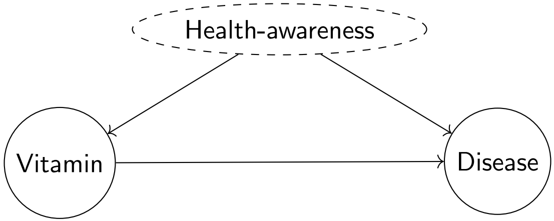
- In 1990s, several studies have found a strong inverse association of antioxidant vitamins with cardiovascular disease, cancer, and all-cause mortality.
- However, well conducted randomised controlled trials later have shown that supplementation with antioxidants does not protect against these diseases.



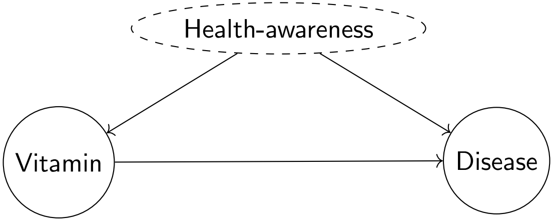
**Estimates of the effects of an increase of 15·7 μmol/L plasma vitamin C on coronary heart disease 5-year mortality**

What went wrong? (Figure from D. A. Lawlor *et al.*, *The Lancet* **363**, 1724–1727 (May 2004).)

# Confounder = Common cause of treatment and effect

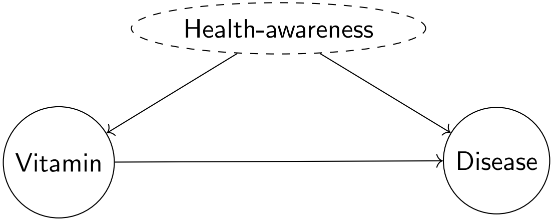


# Confounder = Common cause of treatment and effect



- How can we balance observed confounders? Better design (e.g. blocking).

# Confounder = Common cause of treatment and effect



- How can we balance observed confounders? Better design (e.g. blocking).
- How can we balance unobserved confounders (stochastically)? Randomization!

# Randomization as a basis of inference

Randomization is now widely regarded as the “gold standard” for causal inference. But in the early days, many people find it difficult to accept.

# Randomization as a basis of inference

Randomization is now widely regarded as the “gold standard” for causal inference. But in the early days, many people find it difficult to accept.

## Example

- Suppose a physician is allowed to administer a promising new drug to 5 out of 10 patients.

# Randomization as a basis of inference

Randomization is now widely regarded as the “gold standard” for causal inference. But in the early days, many people find it difficult to accept.

## Example

- Suppose a physician is allowed to administer a promising new drug to 5 out of 10 patients.
- The physician thinks the best way to prove the effectiveness of the drug is to give it to the 5 patients that they think are the most ill.

# Randomization as a basis of inference

Randomization is now widely regarded as the “gold standard” for causal inference. But in the early days, many people find it difficult to accept.

## Example

- Suppose a physician is allowed to administer a promising new drug to 5 out of 10 patients.
- The physician thinks the best way to prove the effectiveness of the drug is to give it to the 5 patients that they think are the most ill.
- What’s the flaw in this design?

# Randomization as a basis of inference

Randomization is now widely regarded as the “gold standard” for causal inference. But in the early days, many people find it difficult to accept.

## Example

- Suppose a physician is allowed to administer a promising new drug to 5 out of 10 patients.
- The physician thinks the best way to prove the effectiveness of the drug is to give it to the 5 patients that **they think** are the most ill.
- What’s the flaw in this design?

Randomization introduces an **objective basis of inference** which anyone else can use.

# A mathematical formalization of causal inference

- The treatment (e.g. vitamin) and outcome (e.g. disease status) of the $i$th individual are represented by two variables, $A_i$ and $Y_i$, respectively.

# A mathematical formalization of causal inference

- The treatment (e.g. vitamin) and outcome (e.g. disease status) of the $i$th individual are represented by two variables, $A_i$ and $Y_i$, respectively.

### Key concept: Potential/Counterfactual outcome

Let $Y_i(a)$ be the value of the outcome of individual $i$ **if the treatment is $A_i = a$.**

# A mathematical formalization of causal inference

- The treatment (e.g. vitamin) and outcome (e.g. disease status) of the $i$th individual are represented by two variables, $A_i$ and $Y_i$, respectively.

## Key concept: Potential/Counterfactual outcome

Let $Y_i(a)$ be the value of the outcome of individual $i$ **if the treatment is $A_i = a$.**
There are two ways to interpret this:

- **Prospectively**, $Y_i(a)$ is the (potential) value of $Y_i$ if we assign treatment value $a$ to this individual.
- **Retrospectively**, $Y_i(a)$ is the (counterfactual) value of $Y_i$ had this individual received treatment value $a$.

# A mathematical formalization of causal inference

- The treatment (e.g. vitamin) and outcome (e.g. disease status) of the $i$th individual are represented by two variables, $A_i$ and $Y_i$, respectively.

## Key concept: Potential/Counterfactual outcome

Let $Y_i(a)$ be the value of the outcome of individual $i$ **if the treatment is $A_i = a$.**
There are two ways to interpret this:

- **Prospectively**, $Y_i(a)$ is the (potential) value of $Y_i$ if we assign treatment value $a$ to this individual.
- **Retrospectively**, $Y_i(a)$ is the (counterfactual) value of $Y_i$ had this individual received treatment value $a$.

Some call this the **Neyman-Rubin causal model**.

# The inferential problem

Under the N-R model, we are interested in making inference about $Y_i(1) - Y_i(0)$.
(e.g. Will the disease status be different if we do or do not take vitamin supplements?)

# The inferential problem

Under the N-R model, we are interested in making inference about $Y_i(1) - Y_i(0)$.
(e.g. Will the disease status be different if we do or do not take vitamin supplements?)

- A common presumption for statistical inference is the stable unit treatment value assumption (SUTVA): $Y_i = Y_i(A_i)$ for all $i$.

# The inferential problem

Under the N-R model, we are interested in making inference about $Y_i(1) - Y_i(0)$.
(e.g. Will the disease status be different if we do or do not take vitamin supplements?)

- A common presumption for statistical inference is the stable unit treatment value assumption (SUTVA): $Y_i = Y_i(A_i)$ for all $i$.
- This links potential/counterfactual outcomes with realized/factual outcomes.
- This can be violated, for example, if there is **interference** (e.g. if we are studying the effect of a vaccine).

# The inferential problem

Under the N-R model, we are interested in making inference about $Y_i(1) - Y_i(0)$.
(e.g. Will the disease status be different if we do or do not take vitamin supplements?)

- A common presumption for statistical inference is the stable unit treatment value assumption (SUTVA): $Y_i = Y_i(A_i)$ for all $i$.
- This links potential/counterfactual outcomes with realized/factual outcomes.
- This can be violated, for example, if there is **interference** (e.g. if we are studying the effect of a vaccine).

## Fundamental problem of causal inference

**Only one potential outcome can ever be observed!**

# The inferential problem

Under the N-R model, we are interested in making inference about $Y_i(1) - Y_i(0)$.
(e.g. Will the disease status be different if we do or do not take vitamin supplements?)

- A common presumption for statistical inference is the stable unit treatment value assumption (SUTVA): $Y_i = Y_i(A_i)$ for all $i$.
- This links potential/counterfactual outcomes with realized/factual outcomes.
- This can be violated, for example, if there is **interference** (e.g. if we are studying the effect of a vaccine).

## Fundamental problem of causal inference

**Only one potential outcome can ever be observed!**

| $i$ | $Y_i(0)$ | $Y_i(1)$ | $A_i$ | $Y_i$ |
|---|---|---|---|---|
| 1 | ? | **1** | 1 | 1 |
| 2 | **0** | ? | 0 | 0 |
| 3 | ? | **0** | 1 | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

# Imputation of potential outcomes

We are interested in testing the null hypothesis $H_0 : Y_i(0) = Y_i(1)$ for all $i$.

# Imputation of potential outcomes

We are interested in testing the null hypothesis $H_0 : Y_i(0) = Y_i(1)$ for all $i$.
Under $H_0$, we may impute all the potential outcomes by $Y_i(0) = Y_i(1) = Y_i$.

# Imputation of potential outcomes

We are interested in testing the null hypothesis $H_0 : Y_i(0) = Y_i(1)$ for all $i$.
Under $H_0$, we may impute all the potential outcomes by $Y_i(0) = Y_i(1) = Y_i$.

## Example

| $i$ | $Y_i(0)$ | $Y_i(1)$ | $A_i$ | $Y_i$ |
|---|---|---|---|---|
| 1 | **1** | 1 | 1 | 1 |
| 2 | 0 | **0** | 0 | 0 |
| 3 | **0** | 0 | 1 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# Causal identification

Suppose $(A_i, Y_i(0), Y_i(1), X_i), i = 1, \dots, n$ are independent and identically distributed. We say the causal effect of $A$ on $Y$ have **no unmeasured confounders** if

$$A_i \perp\!\!\!\perp Y_i(a) \mid X_i, \text{ for } a = 0, 1.$$

# Causal identification

Suppose $(A_i, Y_i(0), Y_i(1), X_i), i = 1, \ldots, n$ are independent and identically distributed. We say the causal effect of $A$ on $Y$ have **no unmeasured confounders** if

$$A_i \perp\!\!\!\perp Y_i(a) \mid X_i, \text{ for } a = 0, 1.$$

## Theorem (Identification of average treatment effect)

*Assuming SUTVA, no unmeasured confounders, and positivity (i.e. $0 < \mathbb{P}(A_i = 1 \mid X_i) < 1$), we have*

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \mathbb{E}[Y_i \mid A_i = 1, X_i = x] - \mathbb{E}[Y_i \mid A_i = 0, X_i = x].$$

# Causal identification

Suppose $(A_i, Y_i(0), Y_i(1), X_i), i = 1, \ldots, n$ are independent and identically distributed. We say the causal effect of $A$ on $Y$ have **no unmeasured confounders** if

$$A_i \perp\!\!\!\perp Y_i(a) \mid X_i, \text{ for } a = 0, 1.$$

## Theorem (Identification of average treatment effect)

*Assuming SUTVA, no unmeasured confounders, and positivity (i.e. $0 < \mathbb{P}(A_i = 1 \mid X_i) < 1$), we have*

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \mathbb{E}[Y_i \mid A_i = 1, X_i = x] - \mathbb{E}[Y_i \mid A_i = 0, X_i = x].$$

*Proof: For any $a$ and $x$,*

$$\begin{aligned} Y_i(a) \mid \boldsymbol{X}_i = x &\overset{d}{=} Y_i(a) \mid \boldsymbol{X}_i = x, A_i = a && \textit{(by unconfoundedness and positivity)} \\ &\overset{d}{=} Y_i \mid \boldsymbol{X}_i = x, A_i = a. && \textit{(by SUTVA)} \end{aligned}$$

# Outline

1. Randomization and potential outcomes

2. Undirected graphical models

3. Directed acyclic graphical (DAG) models

4. Causal DAGs

5. Why does causality matter?

# Contingency tables and conditional independence

## A simple example

- Observed three discrete random variables (e.g., genotypes): $(A_i, B_i, C_i)$, $i = 1, \dots, n$.
- Data as a **contingency table**: $Y_{abc} = \sum_{i=1}^{n} 1_{\{A_i=a, B_i=b, C_i=c\}}$ ($a/b/c$ is a level of $A/B/C$).
- Let $\pi_{abc} = \mathbb{P}(A = a, B = b, C = c)$. It is common to model the counts by $Y_{abc} \overset{\text{ind}}{\sim} \text{Poisson}(\mu \cdot \pi_{abc})$.

# Contingency tables and conditional independence

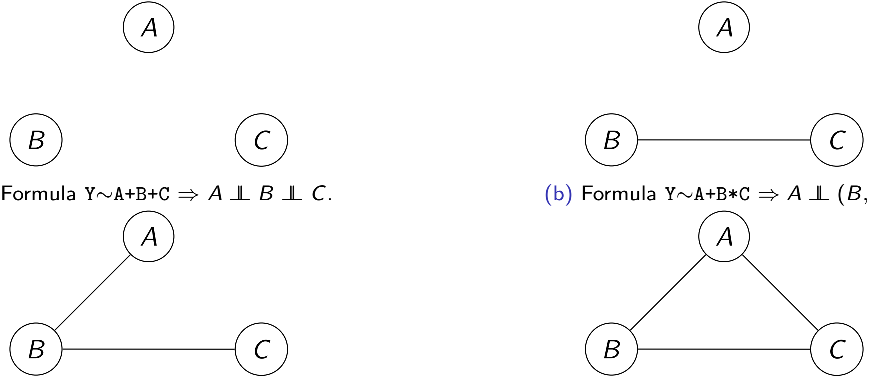
## A simple example

- Observed three discrete random variables (e.g., genotypes): $(A_i, B_i, C_i)$, $i = 1, \dots, n$.
- Data as a **contingency table**: $Y_{abc} = \sum_{i=1}^{n} 1_{\{A_i = a, B_i = b, C_i = c\}}$ ($a/b/c$ is a level of $A/B/C$).
- Let $\pi_{abc} = \mathbb{P}(A = a, B = b, C = c)$. It is common to model the counts by $Y_{abc} \overset{\text{ind}}{\sim} \text{Poisson}(\mu \cdot \pi_{abc})$.

| `glm` formula in R | Poisson log-linear model | Joint distribution | Independence |
|---|---|---|---|
| `Y∼A+B+C` | $\log \mu_{abc} = \log \mu + \log \pi_a + \log \pi_b + \log \pi_c$ | $\pi_{abc} = \pi_a \pi_b \pi_c$ | $A \perp\!\!\!\perp B \perp\!\!\!\perp C$ |
| `Y∼A+B*C` | $\log \mu_{abc} = \log \mu + \log \pi_a + \log \pi_{bc}$ | $\pi_{abc} = \pi_a \pi_{bc}$ | $A \perp\!\!\!\perp (B, C)$ |
| `Y∼A*B+B*C` | $\log \mu_{abc} = \log \mu + \log \pi_{ab} + \log \pi_{bc}$ | $\pi_{abc} = \pi_{ab} \pi_{bc}$ | $A \perp\!\!\!\perp C \mid B$ |
| `Y∼A*B+B*C+C*A` | $\log \mu_{abc} = \log \mu + \log \pi_{ab} + \log \pi_{bc} + \log \pi_{ac}$ | $\pi_{abc} = \pi_{ab} \pi_{bc} \pi_{ac}$ | No (but no three-way interaction) |
| `Y∼A*B*C` | $\log \mu_{abc} = \log \mu + \log \pi_{abc}$ | $\pi_{abc} = \pi_{abc}$ | No |

# Undirected graphical models

- Add an edge if there is an interaction in the joint distribution.
- Blocking all paths ⇒ conditional independence.



(a) Formula $\texttt{Y}\sim\texttt{A+B+C} \Rightarrow A \perp\!\!\!\perp B \perp\!\!\!\perp C$.

(b) Formula $\texttt{Y}\sim\texttt{A+B*C} \Rightarrow A \perp\!\!\!\perp (B, C)$.

(c) Formula $\texttt{Y}\sim\texttt{A*B+B*C} \Rightarrow A \perp\!\!\!\perp C \mid B$.

(d) Formula $\texttt{Y}\sim\texttt{A*B+B*C+C*A}$ or $\texttt{Y}\sim\texttt{A*B*C}$.

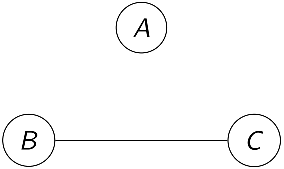
# Undirected graphical models: Rigorous definitions

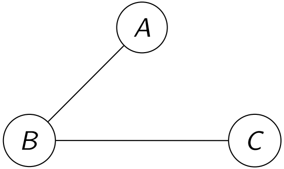
## Basic theorem: Hammersley-Clifford

Suppose $\boldsymbol{X}$ has a positive mass/density function $f_{\boldsymbol{X}}(\cdot)$, then

$$\underbrace{f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{\text{clique } C \subseteq V} \psi_C(\boldsymbol{x}_C) \text{ for some } \psi_C(\cdot), C \subseteq V}_{f \text{ factories according to } \mathcal{G}} \Longleftrightarrow \underbrace{J \perp\!\!\!\perp K \mid L \; [\mathcal{G}] \Rightarrow \boldsymbol{X}_J \perp\!\!\!\perp \boldsymbol{X}_K \mid \boldsymbol{X}_L, \forall \text{distinct } J, K, L \subset V}_{\text{``Global Markov property''}}.$$

# Undirected graphical models: Examples





| glm formula | Poisson log-linear model | Joint distribution | Independence |
|---|---|---|---|
| `Y∼A+B*C` | $\log \mu_{abc} = \log \mu + \log \pi_a + \log \pi_{bc}$ | $\pi_{abc} = \pi_a \pi_{bc}$ | $A \perp\!\!\!\perp (B, C)$ |
| `Y∼A*B+B*C` | $\log \mu_{abc} = \log \mu + \log \pi_{ab} + \log \pi_{bc}$ | $\pi_{abc} = \pi_{ab} \pi_{bc}$ | $A \perp\!\!\!\perp C \mid B$ |

- Verify that the joint distribution factories according to the corresponding graph.
- Verify conditional independence by graph separation.

# Outline

1 Randomization and potential outcomes

2 Undirected graphical models

3 Directed acyclic graphical (DAG) models

4 Causal DAGs

5 Why does causality matter?

# DAG models

## Graph terminology

- **Directed graph** = all edges are directed.
- **Path** is a sequence of distinct, adjacent nodes. **Directed path** = all arrows are going "forward".
- **Cycle** is a directed path with the modification that the first and last nodes are the same.
- **Directed acyclic graph (DAG)** = directed graph with no cycles.
- If $A \rightarrow B$, $A \in pa(B)$ **parent set** of $B$; $B \in ch(A)$ **child set** of $A$.
- **Ancestors** = parents, parents of parents, ...; **Descendants** = children, children of children, ....

# DAG models

## Graph terminology

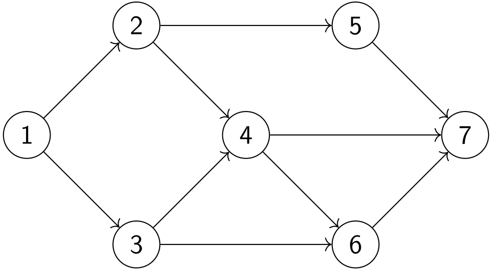
- **Directed graph** = all edges are directed.
- **Path** is a sequence of distinct, adjacent nodes. **Directed path** = all arrows are going "forward".
- **Cycle** is a directed path with the modification that the first and last nodes are the same.
- **Directed acyclic graph (DAG)** = directed graph with no cycles.
- If $A \rightarrow B$, $A \in pa(B)$ **parent set** of $B$; $B \in ch(A)$ **child set** of $A$.
- **Ancestors** = parents, parents of parents, ...; **Descendants** = children, children of children, ....

We say the distribution of $\boldsymbol{X}$ **factories according to a DAG** $\mathcal{G}$ (also called a **Bayesian network**) if its density satisfies

$$f(\boldsymbol{x}) = \prod_{i \in V} f_{i|pa(i)}(x_i \mid \boldsymbol{x}_{pa(i)}),$$

where $f_{i|pa(i)}(x_i \mid \boldsymbol{x}_{pa(i)})$ is the conditional density of $X_i$ given $\boldsymbol{X}_{pa(i)}$.

# DAG factorisation: Examples



$$f(\boldsymbol{x}) = f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_4 \mid x_2, x_3)f(x_5 \mid x_2)f(x_6 \mid x_3, x_4)f(x_7 \mid x_4, x_5, x_6).$$

(To simplify notation, we omit the subscripts indexing density functions.)

# DAG models: Conditional independence

In undirected graphical models, factorisation is equivalent to the global Markov property (conditional independence by graph separation). How do we test $\boldsymbol{X}_J \perp\!\!\!\perp \boldsymbol{X}_K \mid \boldsymbol{X}_L$ in DAG models?

# DAG models: Conditional independence

In undirected graphical models, factorisation is equivalent to the global Markov property (conditional independence by graph separation). How do we test $\boldsymbol{X}_J \perp\!\!\!\perp \boldsymbol{X}_K \mid \boldsymbol{X}_L$ in DAG models?

## Conditional independence: Intuitions

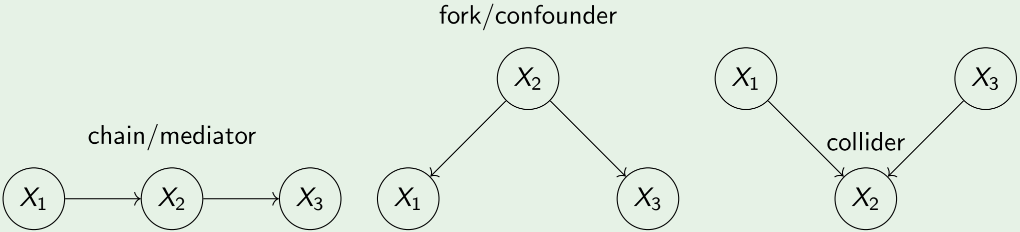


Figure: Possible DAGs with 3 vertices and 2 edges.

- $X_1 \perp\!\!\!\perp X_3$ is true in graph 3 but not in 1 & 2.
- $X_1 \perp\!\!\!\perp X_3 \mid X_2$ is true in graphs 1 & 2 but not in 3.

Exercise: verify these by using the DAG factorisation.

# Graphical criteria

Suppose we are interested in testing $\boldsymbol{X}_J \perp\!\!\!\perp \boldsymbol{X}_K \mid \boldsymbol{X}_L$.

## Converting to undirected graph

1. Obtain the subgraph containing $J$, $K$, $L$, and their ancestors;
2. Moralisation: join parents with a common child; then ignores all direction edges.
3. Examine whether $L$ blocks $J$ from $K$.

# Graphical criteria

Suppose we are interested in testing $\boldsymbol{X}_J \perp\!\!\!\perp \boldsymbol{X}_K \mid \boldsymbol{X}_L$.

### Converting to undirected graph

1. Obtain the subgraph containing $J$, $K$, $L$, and their ancestors;
2. Moralisation: join parents with a common child; then ignores all direction edges.
3. Examine whether $L$ blocks $J$ from $K$.

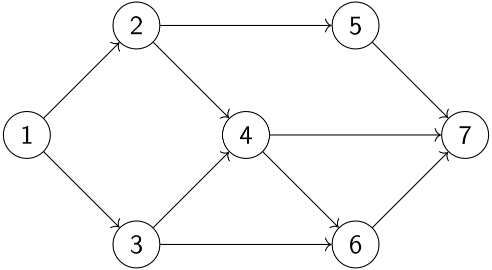
### d-separation

- In $B \rightarrow A \leftarrow C$, $A$ is called a **collider**.
- A path is **blocked** by $L \subseteq V$ if there exists $A$ on the path such that **either**
  - $A$ is not a collider and $A \in L$; **or**
  - $A$ is a collider and $A$ and all its descendants are not in $L$;
- $J$ and $K$ are **d-separated** by $L$ (written as $J \perp\!\!\!\perp K \mid L\ [\mathcal{G}]$) if every path from $J$ to $K$ is blocked by $L$.

### Theorem

1. These two criteria are equivalent.
2. Factorisation according to DAG $\mathcal{G} \iff \underbrace{J \perp\!\!\!\perp K \mid L\ [\mathcal{G}] \Rightarrow \boldsymbol{X}_J \perp\!\!\!\perp \boldsymbol{X}_K \mid \boldsymbol{X}_L, \forall \text{distinct } J, K, L \subset V}_{\text{Global Markov property}}$.

# Graph separation: Examples



1. $X_2 \not\perp\!\!\!\perp X_6 \mid X_4$ ($2 \leftarrow 1 \rightarrow 3 \rightarrow 6$ is unblocked);
2. $X_5 \not\perp\!\!\!\perp X_6 \mid X_4$ ($5 \leftarrow 2 \rightarrow 4 \leftarrow 3 \rightarrow 6$ and $5 \leftarrow 2 \leftarrow 1 \rightarrow 3 \rightarrow 6$ are unblocked);
3. $X_5 \perp\!\!\!\perp X_6 \mid \{X_3, X_4\}$;

Exercise: verify $X_2 \not\perp\!\!\!\perp X_6 \mid X_3$ and $X_2 \not\perp\!\!\!\perp X_7 \mid \{X_4, X_5\}$.

# Outline

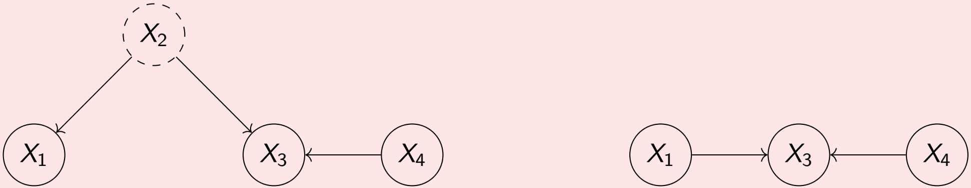
1. Randomization and potential outcomes
2. Undirected graphical models
3. Directed acyclic graphical (DAG) models
4. Causal DAGs
5. Why does causality matter?

# Causal inference: Correlation is not causation

- Up till now, graphs are used to model the distribution of observed data.
- However, the model may not generalise to other settings.

## Example

Imagine we have only observed $X_1, X_3, X_4$ (three proteins) but not $X_2$ (another protein).



(a) True causal DAG $\Rightarrow X_1 \perp\!\!\!\perp X_4, X_1 \not\perp\!\!\!\perp X_4 \mid X_3$.

(b) Encodes the same conditional independence relations.

Figure: Arrow in probabilistic DAG models $\neq$ causality.

# Causal DAGs

- A **causal graphical model** means that the (almost same) graph also holds under interventions.
- Example in last slide: $X_1 \to X_3 \leftarrow X_4$ is a probabilistic DAG but not a causal DAG.

# Causal DAGs

- A **causal graphical model** means that the (almost same) graph also holds under interventions.
- Example in last slide: $X_1 \to X_3 \leftarrow X_4$ is a probabilistic DAG but not a causal DAG.

## Formalising causality: Two cultures

### Structural equation models (SEMs)

$$X_j = g_j(\boldsymbol{X}_{pa(j)}, \epsilon_j),\; j \in V.$$

- $g_j(\cdot)$ describes how $X_j$ depends on its parents mechanically.
- $\epsilon_j$ is noise variable.
- Structural/causal: if we make an intervention and change some of $\boldsymbol{X}_{pa(j)}$, the equations still hold.

# Causal DAGs

- A **causal graphical model** means that the (almost same) graph also holds under interventions.
- Example in last slide: $X_1 \to X_3 \leftarrow X_4$ is a probabilistic DAG but not a causal DAG.

## Formalising causality: Two cultures

### Structural equation models (SEMs)

$$X_j = g_j(\boldsymbol{X}_{pa(j)}, \epsilon_j),\ j \in V.$$

- $g_j(\cdot)$ describes how $X_j$ depends on its parents mechanically.
- $\epsilon_j$ is noise variable.
- Structural/causal: if we make an intervention and change some of $\boldsymbol{X}_{pa(j)}$, the equations still hold.

### Counterfactuals/Potential outcomes

For $k \in pa(j)$, recursively define

$$X_j(x_k) = g_j(x_k, \boldsymbol{X}_{pa(j)\setminus\{k\}}(x_k), \epsilon_j).$$

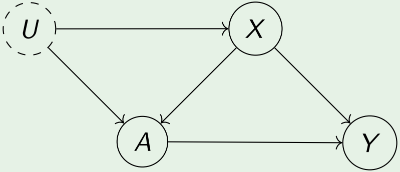
- For example, in the graph $X_1 \to X_2 \to X_3$ $X_1 \to X_3$, we have

$$X_2(x_1) = g_2(x_1, \epsilon_2),$$
$$X_3(x_1) = g_3(x_1, X_2(x_1), \epsilon_3).$$

- May define causal effect of $X_1$ on $X_3$ as $X_3(x_1) - X_3(x_1')$.

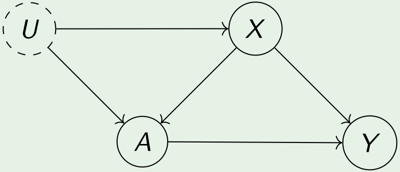
# Graphical criterion for causal identification

Theorem (Backdoor adjustment/Confounder adjustment, Pearl)



# Graphical criterion for causal identification

## Theorem (Backdoor adjustment/Confounder adjustment, Pearl)



We have $\mathbb{E}[Y(A=1) - Y(A=0) \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid A = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A = 0, \boldsymbol{X} = \boldsymbol{x}]$ if

- $\boldsymbol{X}$ **blocks all "backdoor" paths** from $A$ to $Y$ (paths with an arrow into $A$).
- $\boldsymbol{X}$ contains no descendants of $A$.

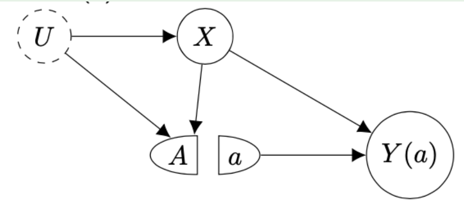
Proof: Under these graphical conditions, $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}$. That is, there are **no unmeasured confounders**!

# Single-world intervention graphs (SWIGs)

It turns out that there is a nice unification of the potential outcome and graphical approaches to causal inference: Given a causal DAG, the "single-world" counterfactuals (potential outcomes under the same intervention) will factorize according to a modified graph:

- Split the intervention node into two halves: a random half that inherits all incoming arrows and a fixed half that inherits all outgoing arrows.
- Change (the downstream) variables to the corresponding counterfactuals.

## Example



In this example, $A$ and $Y(a)$ are d-separated by $X$.

# Outline

1. Randomization and potential outcomes

2. Undirected graphical models

3. Directed acyclic graphical (DAG) models

4. Causal DAGs

5. Why does causality matter?

# Connections to medicine

It is fair to say that causal inference (especially the potential outcomes approach) is ubiquitous in clinical research and practice.

- **Randomized clinical trials** were developed after theoretical advancements in the **design and analysis of experiments**.
- In **epidemiology**, it is essential to distinguishing causality from correlation by identifying the correct **confounders**.
- Much of **precision medicine** is about inferring different aspects of the **conditional average treatment effect** $\mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X}]$.
- Another related problem in **precision medicine** is **dynamic treatment regimes**, where we are interested in designing the optimal sequence of treatment based on information we collected about the patients.
- When there are concerns about **unmeasured confounders**, **instrumental variables** provide a useful strategy to (partially) identify the causal effect.

# Connections to machine learning

To develop **artificial intelligence**, **graphical models** were brought in to computer science in 1980s. They are now ubiquitous in machine learning.

- Graphical rules such as **d-separation** were developed in hope that we can **make reasoning automatic**.
- Graphical algorithms such as **message passing** were developed to make probabilistic inference on graphs. They are now widely used in **Bayesian inference**.
- In reinforcement learning, **policy evaluation** is closely related to **causal effect estimation**.
- **Transfer learning** is closely related to **generalizability** of causal inference.