CAUSAL LEARNING IN HEALTHCARE

Irene BALELLI
irene.balelli@inria.fr

RECAP FROM LESSON 1– KEY CONCEPTS

- ✓ DAG, paths, Markov blanket
- ✓ Local Markov Condition: factorisation of the joint distribution induced by the DAG
- ✓ Chain; fork; v-structure
- ✓ Blocking path: mediator or confounder; not a (descendant of) collider. Intuition: X--Y is blocked by Z if conditioning on Z makes X and Y independent.
- ✓ D-separation: extend this definition to sets of nodes
- ✓ Causal sufficiency and fairness
- ✓ MEC and PAG
- ✓ Constraint-based Causal Discovery: PC and FCI

TOWARD GOAL-ORIENTED CAUSALITY

- Treatment-Outcome
- Climbing the ladder
- The fundamental problem of causal inference

FROM EGO-LESS TO GOAL-ORIENTED GRAPH

Recall: in the first part of this course, we discussed about causal graphs and the possibility of recovering them from observations → ego-less set of causal relationships

Today:
- We will suppose the causal graph as already given
- We will give special roles to some nodes therein

T **- the Treatment:** an intervenable variable that we are interested in/can manipulate (e.g., a drug administration/dosage, a policy change, an habit,…)

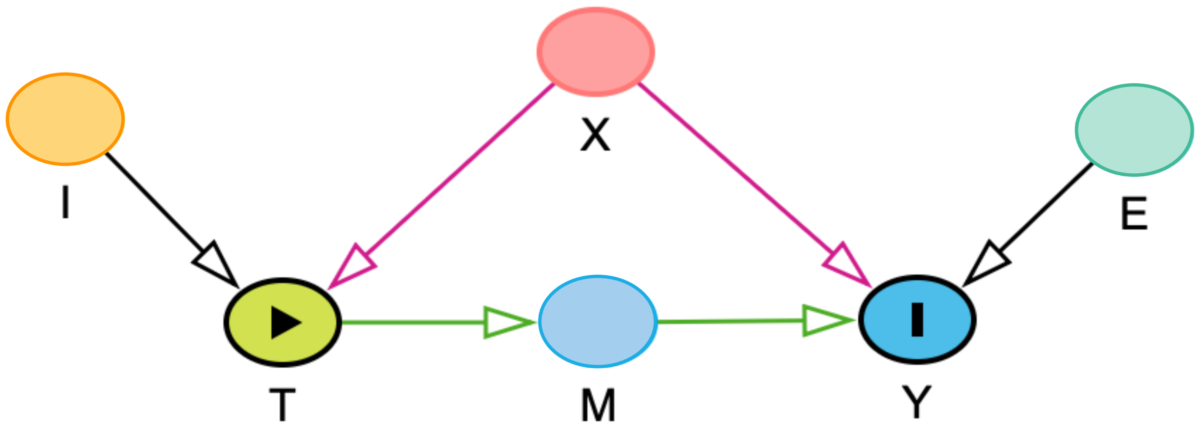
Y **- the Outcome:** the quantity of interest we want to analyse and measure (e.g., patient recovery rate, mortality, risk of stroke,…)

FROM EGO-LESS TO GOAL-ORIENTED GRAPH

T **- the Treatment:** an intervenable variable that we are interested in/can manipulate (e.g., a drug administration/dosage, a policy change, an habit,…)

Y **- the Outcome:** the quantity of interest we want to analyse and measure (e.g., patient recovery rate, mortality, risk of stroke,…)

- the remaining variables are named after their relative role with respect to treatment and outcome paths (**confounders**, **instrumental variables**, **effect modifiers**, **mediators**)



FROM EGO-LESS TO GOAL-ORIENTED GRAPH

T **- the Treatment:** an intervenable variable that we are interested in/can manipulate (e.g., a drug administration/dosage, a policy change, an habit,…)

Y **- the Outcome:** the quantity of interest we want to analyse and measure (e.g., patient recovery rate, mortality, risk of stroke,…)

Objective: can we identify and quantify the Treatment's effect on the Output? How does a change in the Treatment *propagate* through the system to produce a change in the Outcome? → the **Treatment Effect**

The answer to this question implies developing strategies to handle the **fundamental problem of causal inference**

THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

For a subject (or unit) i, and a binary treatment $T \in \{0,1\}$ (e.g. drug/no drug), let us define the potential outcome $Y_i(t), t = 0,1$ as the outcome of subject i when receiving treatment t.

Our ultimate goal is to estimate the individual causal effect: $ITE_i = Y_i(1) - Y_i(0)$

To do so our dream dataset would be something like this:

$\boldsymbol{i}$
1
2
3
4
5
…
N

$\boldsymbol{Y_i(1)}$	$\boldsymbol{Y_i(0)}$	$\boldsymbol{ITE_i}$
1	0	1
1	1	0
1	0	1
0	0	0
0	0	0
1	1	0
1	0	1

THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

For a subject i, and a binary treatment $T \in \{0,1\}$ (e.g. drug/no drug), let us define the potential outcome $Y_i(t), t = 0,1$ as the outcome of subject i when receiving treatment t.

Our goal is to estimate the individual causal effect: $ITE_i = Y_i(1) - Y_i(0)$

But in reality we get this:

Treatment assignment for subject i

ITE_i is never observed!

$\boldsymbol{i}$	$\boldsymbol{T_i}$	$\boldsymbol{Y_i(1)}$	$\boldsymbol{Y_i(0)}$	$\boldsymbol{ITE_i}$
1	1	1	?	?
2	0	?	1	?
3	0	?	0	?
4	1	0	?	?
5	1	0	?	?
…	0	?	1	?
N	1	1	?	?

THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

For a subject i, and a binary treatment $T \in \{0,1\}$ (e.g. drug/no drug), let us define the potential outcome $Y_i(t), t = 0,1$ as the outcome of subject i when receiving treatment t.

Our goal is to estimate the individual causal effect: $ITE_i = Y_i(1) - Y_i(0)$

$\Longrightarrow$ Causal inference is fundamentally a missing data problem

- We do observe: $Y_i^{\mathrm{obs}} \coloneqq T_i Y_i(1) + (1 - T_i) Y_i(0)$
- We switch our goal from estimating the individual effect to estimating an Average Treatment Effect (ATE) at a population level: $ATE = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Population-level potential outcomes

THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

For a subject i, and a binary treatment $T \in \{0,1\}$ (e.g. drug/no drug), let us define the potential outcome $Y_i(t), t = 0,1$ as the outcome of subject i when receiving treatment t.

Our goal is to estimate the individual causal effect: $ITE_i = Y_i(1) - Y_i(0)$

$\Longrightarrow$ Causal inference is fundamentally a missing data problem

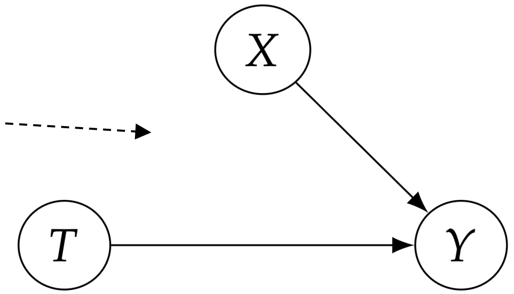
- We do observe: $Y_i^{\text{obs}} \coloneqq T_i Y_i(1) + (1 - T_i) Y_i(0)$
- We switch our goal from estimating and individual effect to estimating an Average Treatment Effect (ATE) at a population level: $ATE = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

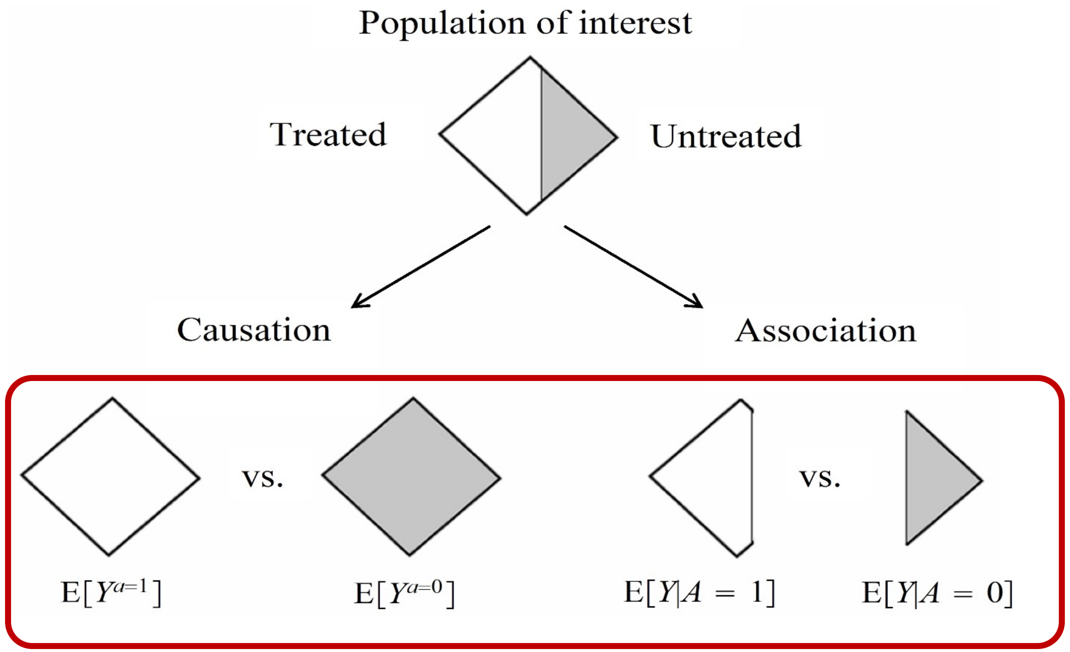
Population-level potential outcomes

Is $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$?

RANDOMIZED CONTROLLED TRIAL (RDT)

- Subjects are randomly assigned to the treatment vs control group $\Longrightarrow T$ is a root node (no parents)
- The path between T and Y is open!
- In this case: $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$



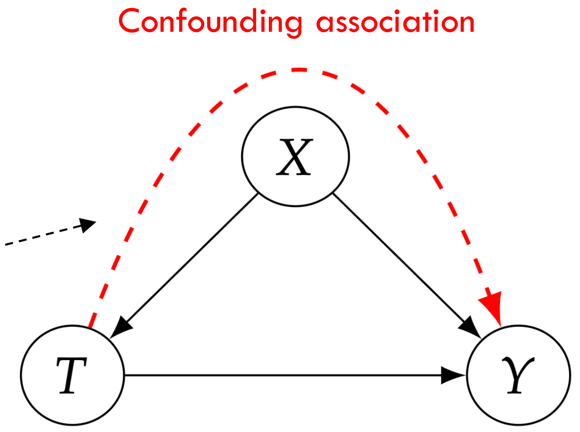


Statistical association is enough to establish causation **in RCTs**!

OBSERVATIONAL STUDIES

- Randomization is not always possible:
  - Ethical reasons
  - Cost and Time
  - Infeasibility

- The path between T and Y is typically **blocked** by confounding



- In this case, statistical association is no longer equivalent to causation (Simpson's paradox), i.e.:

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \neq \underbrace{\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]}$$

The treatment effect is confounded by the (multivariate) variable X (e.g., age, clinical history, disease stage, comorbidities,…), which both affect treatment assignment and outcome.

➢ We need to be more cautious and perform a causal analysis to isolate the causal treatment effect.

TWO APPROACHES TO CAUSAL INFERENCE

1. **The Potential Outcomes Framework (Rubin-Neyman)** – a statistical approach
   - **SUTVA:** No interference between units (your treatment doesn't affect my outcome), and only one version of the treament – $Y = TY(1) + (1 - T)Y(0)$
   - **Ignorability** (or **unconfoundedness**): Treatment assignment is "as good as random" given observed covariates X, i.e., there is no unmeasured confounder – $(Y(0), Y(1)) \perp\!\!\!\perp T | X$
   - **Positivity:** Every unit has a non-zero probability of receiving any treatment – $\forall\ t, x, i,\ 0 < \mathbb{P}(T_i = t | X_i = x) < 1$

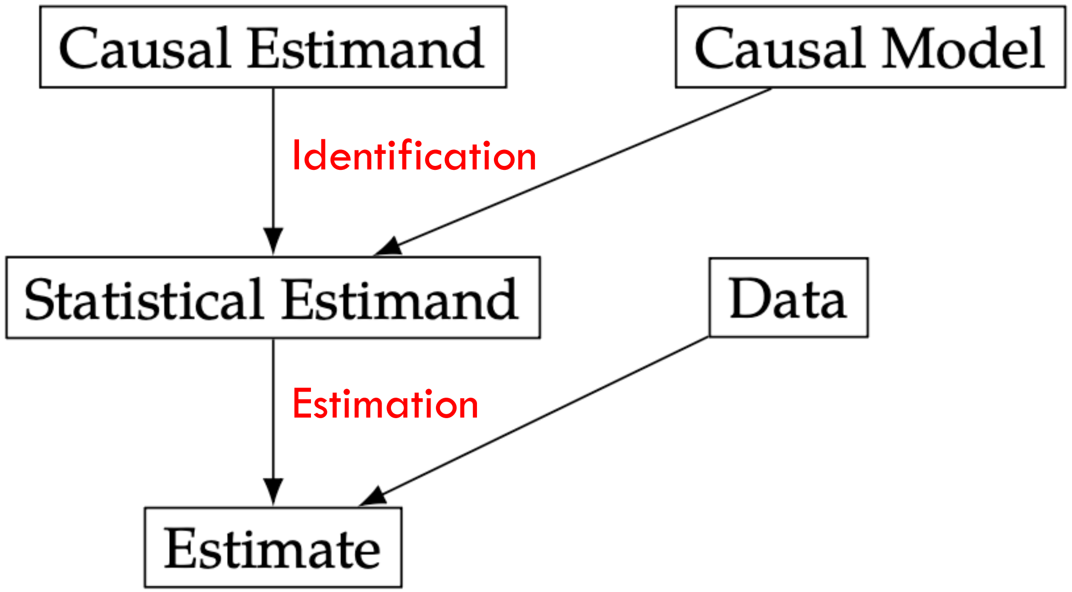
2. **The Do-Calculus & SCMs (Pearl)** – a probabilistic approach
   - **The** *do***-operator:** the probability of observing an outcome Y given that we have observed a treatment assignement $\mathrm{T} = t$ is formally distinguished from the probability of observing an outcome Y given that we have actively switched the treatment assignement to $\mathrm{T} = t$ with the *do*-operator: $\mathbb{P}(Y = y | T = t) \neq \mathbb{P}(Y = y | do(T = t))$.
   - **Modularity** (**independent mechanisms/invariance**)**:** If we intervene on a node, say X_k, only the mechanism $\mathbb{P}(X_k | PA_k)$ changes. All other mechanisms remain unchanged. This is modelled through the graphical representation of the *do*-operator which sirurgically remove all incoming edges of X_k.
   - **The Mechanism:** Represent the world as $Y = f(T, X, U_Y)$ – Structural Causal Model (SCM).

TWO APPROACHES TO CAUSAL INFERENCE

Why both matters?

PO framework primarily focus on **estimation**, while the *do*-calculus primarily focus on **identification** of the causal effect.

- **Rubin's Strength:** Rigorous statistical estimation. Once we decide *what* to control for, Rubin gives us the tools (Propensity Scores, Matching, Weighting) to get the most accurate number.
- **Pearl's Strength:** Scientific transparency. The DAG tells us *which* variables to adjust for (and which NOT to adjust for, like colliders) before we touch the data.



TWO APPROACHES TO CAUSAL INFERENCE

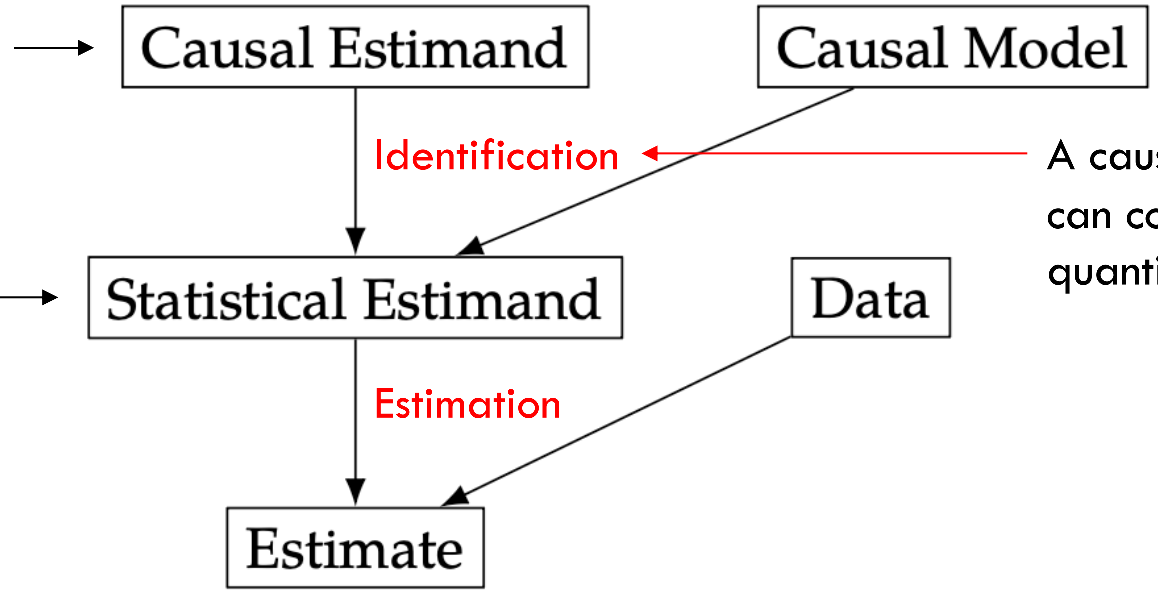
Why both matters?

PO framework primarily focus on **estimation**, while the *do*-calculus primarily focus on **identification** of the causal effect.

- **Rubin's Strength:** Rigorous statistical estimation. Once we decide *what* to control for, Rubin gives us the tools (Propensity Scores, Matching, Weighting) to get the most accurate number.
- **Pearl's Strength:** Scientific transparency. The DAG tells us *which* variables to adjust for (and which NOT to adjust for, like colliders) before we touch the data.

e.g., $\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]$

e.g., $\mathbb{E}[Y|T=0], \mathbb{E}[Y|T=1]$



A causal quantity is identifiable if we can compute it from purely statistical quantities!

TWO APPROACHES TO CAUSAL INFERENCE

Some parallels between the two approaches

Concept	Rubin (Potential Outcomes)	Pearl (Graphical Models)
Causal Effect	$\mathbb{E}[Y(1) - Y(0)]$	$\mathbb{E}[Y \mid do(T=1)] - \mathbb{E}[Y \mid do(T=0)]$
Assumption	SUTVA/Ignorability/Positivity	Modularity/d-separation
Conditioning	Adjusting for "Confounders"	Blocking non-causal paths in the DAG
Inference	Missing Data Problem	Graph Surgery/*do*-calculus

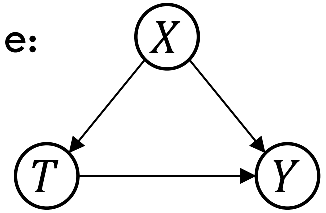
PEARL’S FRAMEWORK

- Identification and the *do*-operator
- Structural causal models (SCMs)

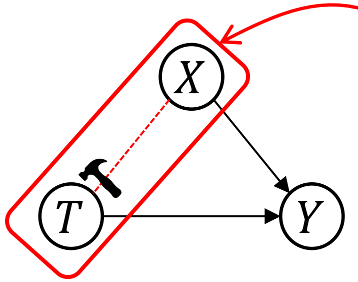
IDENTIFICATION AND THE do-OPERATOR

The *do*-operator (e.g., $do(T = t)$) formally differentiates the seeing (association) from the doing (causation).
Graphically, this is represented through a *surgery*: deleting all incoming edges in T.

Example:



Let us now do an intervention by setting $T = t$



Graph surgery: we are physically removing all edges pointing to T.

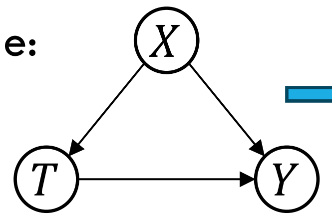
T is no longer influenced by its *natural causes*: it is set to a constant by the experimenter.

This induces a change in the distribution!

IDENTIFICATION AND THE do-OPERATOR

The *do*-operator (e.g., $do(T = t)$) formally differentiates the seeing (association) from the doing (causation). Graphically, this is represented through a *surgery*: deleting all incoming edges in T.

Example:



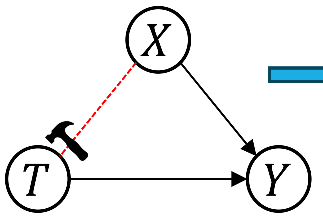
Bayesian factorisation according to the DAG:

$$\mathbb{P}(Y, T, X) = \mathbb{P}(Y|T, X)\mathbb{P}(T|X)\mathbb{P}(X)$$

The probability of having Y when seeing $T = t$ is given by:

$$\underline{\mathbb{P}(Y|T = t)} = \sum_x \mathbb{P}(Y|T = t, X = x)\mathbb{P}(T = t|X = x)\mathbb{P}(X = x)$$

Let us now do an intervention by setting $T = t$



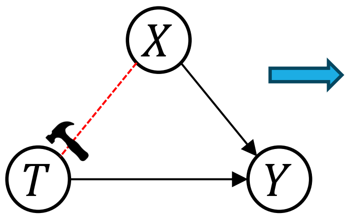
$$\underline{\mathbb{P}(Y|do(T = t))} = \sum_x \mathbb{P}(Y|T = t, X = x)\mathbb{P}(X = x)$$

IDENTIFICATION AND THE do-OPERATOR

The *do*-operator (e.g., $do(T = t)$) formally differentiates the seeing (association) from the doing (causation). Graphically, this is represented through a *surgery*: deleting all incoming edges in T.

- The modularity assumption states that an intervention on $T, do(T = t)$ only changes the mechanism of T ($\mathbb{P}(T = t|PA_T) = 1;\ \mathbb{P}(T = k|Pa_T) = 0$, if $k \neq t$), leaving other mechanisms (like $X \to Y$) unchanged.

- **Association ($\mathbb{P}(Y|T = t)$):** we observe $T = t$, this provide also information about X, which in turns provide information about $Y \Longrightarrow \mathbb{P}(Y|T = t)$ contains **spurious associations**!
- **Causation ($\mathbb{P}(Y|do(T = t))$):** we force $T = t$, the link $X \to T$ is broken and the information no longer flows to the common cause $\Longrightarrow \mathbb{P}(Y|do(T = t))$ describes the true **causal association**!



$$\mathbb{P}(Y|do(T = t)) = \sum_x \mathbb{P}(Y|T = t, X = x)\mathbb{P}(X = x)$$

THE BACKDOOR CRITERION

How can we compute $\mathbb{P}(Y|do(T=t))$ from observational data? How to make the causal effect identifiable?

Backdoor path: a noncausal path between treatment T and outcome Y, with an arrow pointing to T (e.g., $T \leftarrow X \rightarrow Y$)

Backdoor criterion: A set $\boldsymbol{Z} \not\supset \{T, Y\}$ satisfies the backdoor criterion if:

- $\boldsymbol{Z}$ contains no descendant of T
- $\boldsymbol{Z}$ blocks all paths from T to Y entering T through the backdoor ($T \leftarrow \cdots$), i.e., $\boldsymbol{Z}$ blocks all backdoor paths from T to Y

$\Longrightarrow$ $\boldsymbol{Z}$ is a valid **adjustment set** for (T, Y), i.e., **conditioning on $\boldsymbol{Z}$ suffices to control for confounding**, i.e. If $\boldsymbol{Z}$ satisfies the backdoor criterion relative to (T, Y), then the distribution $\mathbb{P}(Y|do(T=t))$ is identifiable from observational data.

Note:

$$\mathbb{P}(Y|do(T=t)) = \sum_x \mathbb{P}(Y|T=t, X=x)\mathbb{P}(X=x)$$

This formula "simulates" a randomized trial by averaging over the strata of the confounder.

We blocked the backdoor path $T \leftarrow X \rightarrow Y$ by conditioning on X and marginalizing it out

THE BACKDOOR CRITERION

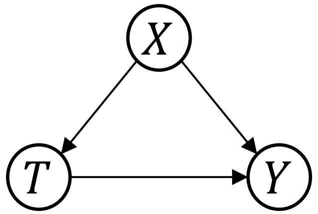
How can we compute $\mathbb{P}(Y|do(T = t))$ from observational data? How to make the causal effect identifiable?

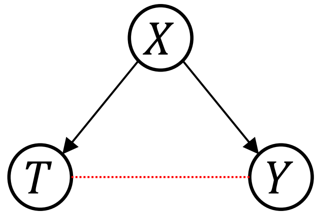
Backdoor path: a noncausal path between treatment T and outcome Y, with an arrow pointing to T (e.g., $T \leftarrow X \rightarrow Y$)

Backdoor criterion: A set $\boldsymbol{Z} \not\supset \{T, Y\}$ satisfies the backdoor criterion if:

- $\boldsymbol{Z}$ contains no descendant of T
- $\boldsymbol{Z}$ blocks all paths from T to Y entering T through the backdoor ($T \leftarrow \cdots$), i.e., $\boldsymbol{Z}$ blocks all backdoor paths from T to Y

Relatioship with d-separation: Backdoor criterion and d-separation are tightly related. Indeed, if a set $\boldsymbol{Z}$ satisfies the backdoor criterion relative to (T, Y), then T and Y are d-separated by $\boldsymbol{Z}$ in the graph $G_{\overline{T}}$ obtained from the original graph by removing all arrows emanating from T.



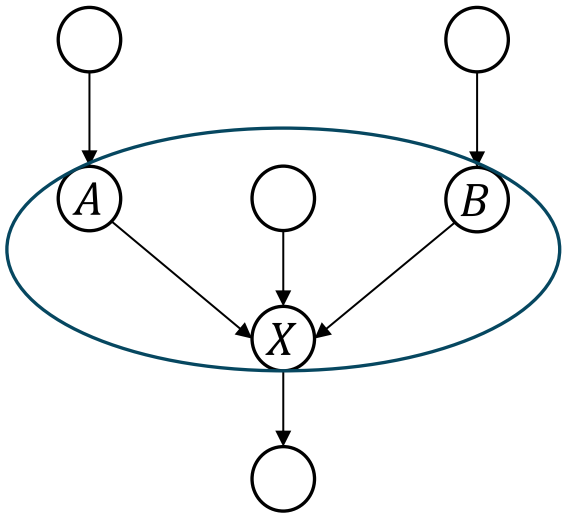


$\{X\}$ d-separate T and Y in $G_{\overline{T}}$:

- $\{X\}$ is a valid adjustment set
- $\mathbb{P}(Y|do(T = t))$ is identifiable if we observe samples from X, T, Y

STRUCTURAL CAUSAL MODELS (SCM)

Causal mechanisms and directed causes: a formal description of the mechanism of action leading to an effect.



$$X = f(\underbrace{A, B, \ldots}_{Pa_X})$$

$$\Updownarrow$$

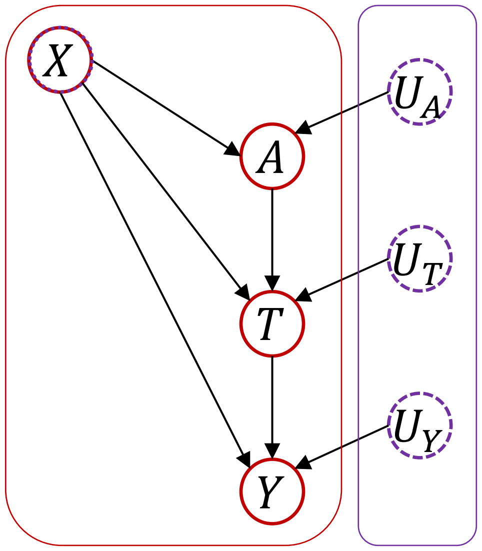
$$\mathbb{P}(X | Pa_X)$$

This has not to be seen as the symmetrical = sign (i.e. $x = ay \Leftrightarrow y = \frac{1}{a}x$), but rather as a **variable assignment which can not be reversed**!

STRUCTURAL CAUSAL MODELS (SCM)

A structural Causal Model completes the data-generating process supported by the DAG, by prescribing functional variable assignments to every variable included in the DAG, plus their corresponding (independent and unobserved) noise.

A DAG+SCM exactly describes HOW variables are generated through their causal interactions.



SCM: endogenous variables + exogenous variables + functions

$$
\begin{cases}
A = f_A(X, U_A) \\
T = f_T(X, A, U_T) \\
Y = f_Y(X, T, U_Y)
\end{cases}
$$

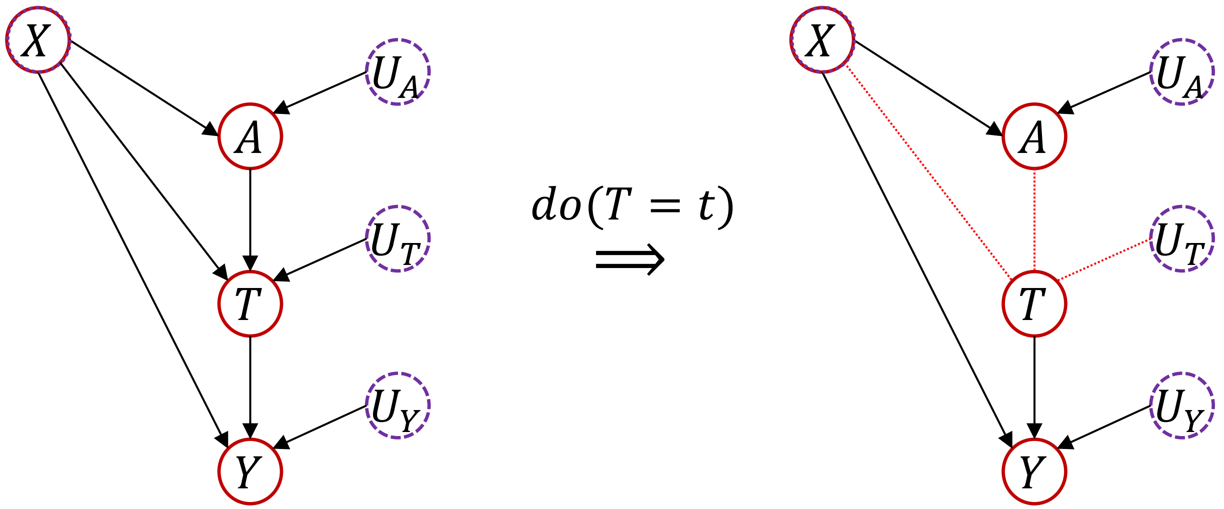
Exogenous variables: they correspond to mutually independent variables that describe the variability of the associated endogenous variable, which can not be explained with the causal relationships within the DAG.

Endogenous variables: they are included by design in the system. Their causal relationship is structurally represented by the DAG

STRUCTURAL CAUSAL MODELS (SCM)

A structural Causal Model completes the data-generating process supported by the DAG, by prescribing functional variable assignments to every variable included in the DAG, plus their corresponding (independent and unobserved) noise.

The *do*-operator, $do(T = t)$, means that we replace the assignment $T = f_T$ by $T = t$ in the intervened SCM.



$$\Longrightarrow \quad \begin{cases} A = f_A(X, U_A) \\ T = t \\ Y = f_Y(X, T, U_Y) \end{cases}$$

WHY SWITCH TO PO FRAMEWORK?

- Now that we have the tools to identify a causal effect (Is it identifiable? What should we control for?), we are interested in its quantitative estimation.

- Most causal inference libraries rely on the Potential Outcome (PO) framework.

- From Pearl's perspective, the PO $Y(t)$ is the solution for Y in the intervened DAG+SCM where we set $\mathrm{T} = t$.