Causal Inference under the rubric of Structural Causal Model

Yonghan Jung

Causal AI Lab

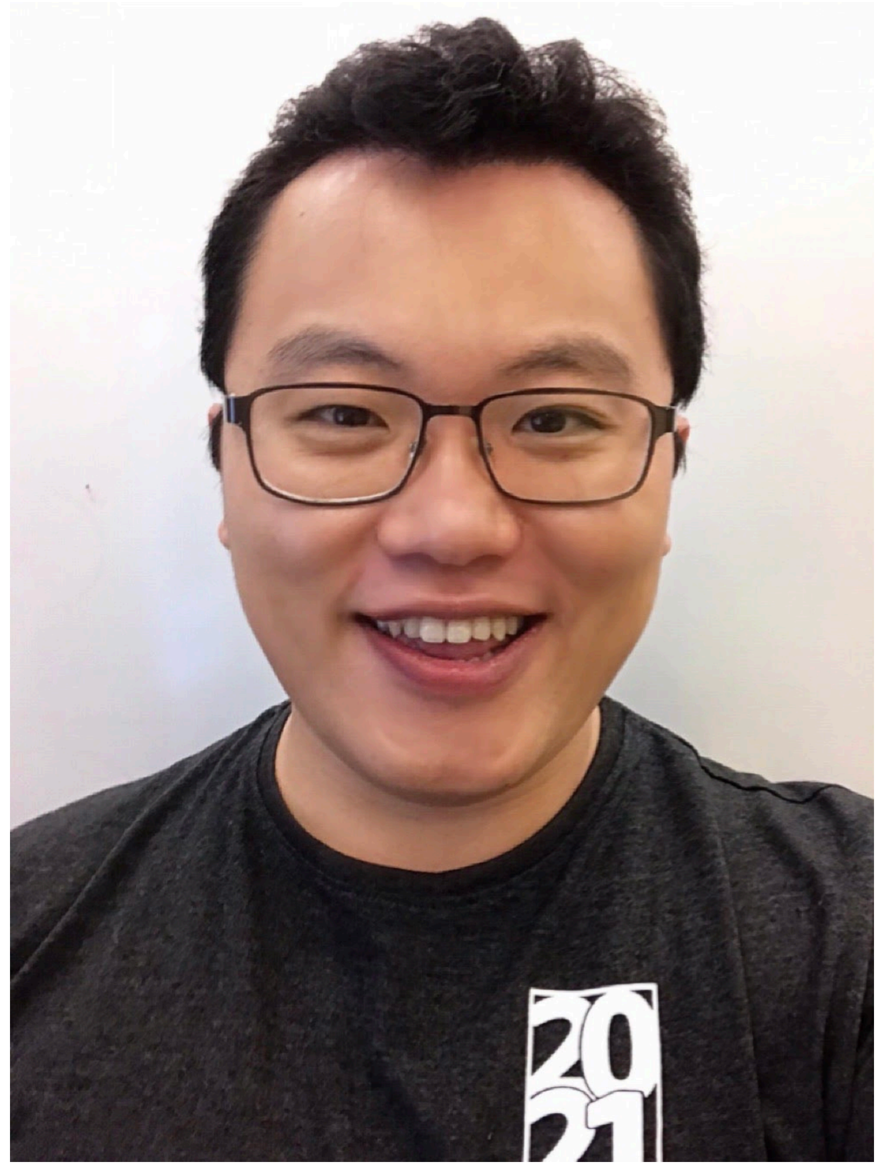
Purdue University

(http://yonghanjung.me/)

Korea Summer Session on Causal Inference 2021

My academic genealogy

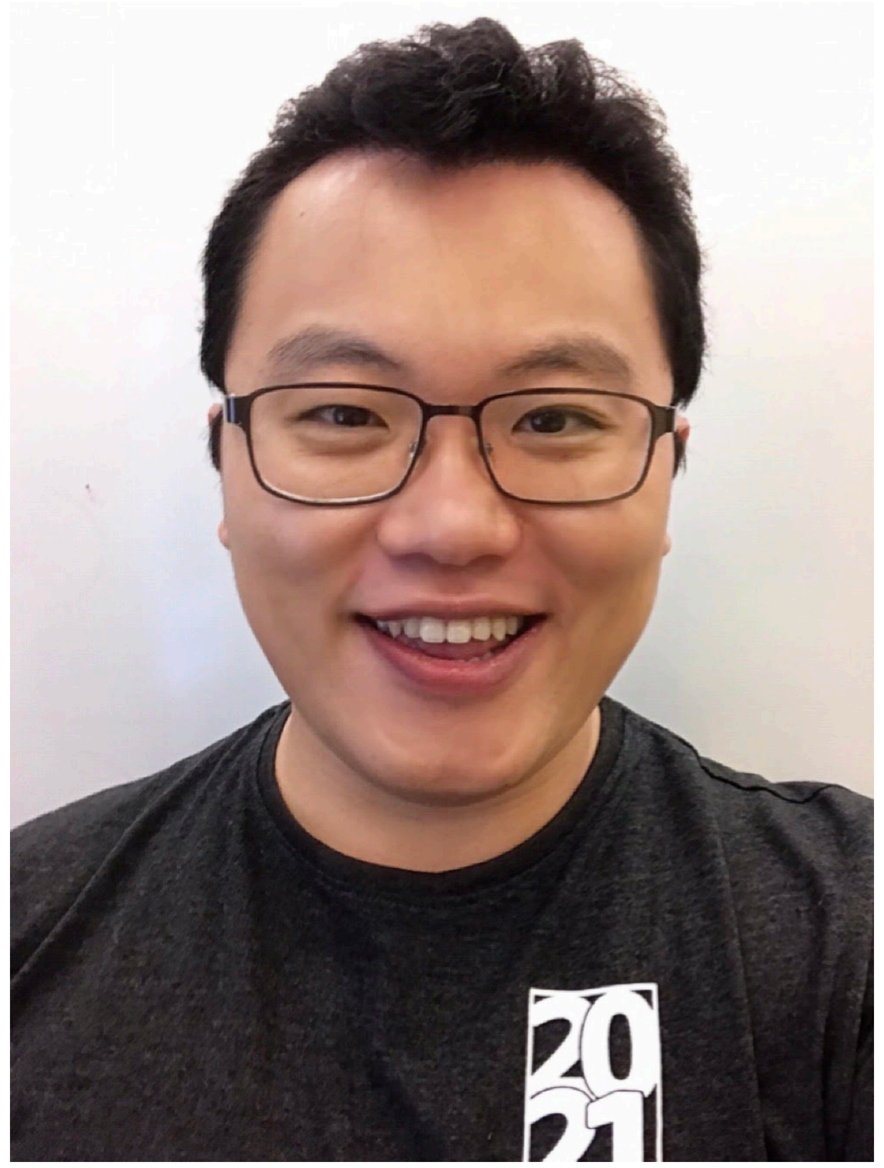
My academic genealogy



Yonghan Jung

http://yonghanjung.me

My academic genealogy



Advisor

Yonghan Jung

http://yonghanjung.me

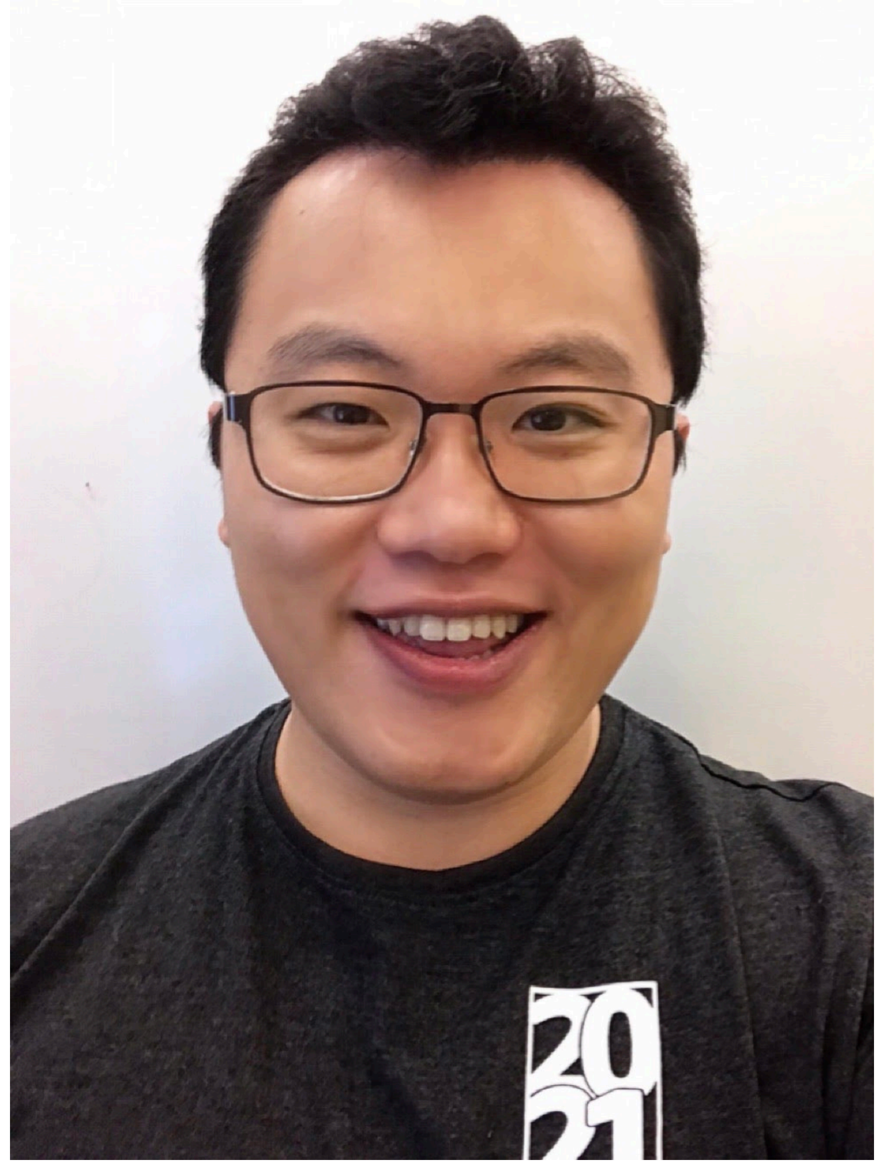
Elias Barenboim

Professor in
Columbia University

Director of CausalAI lab

http://causalai.net

My academic genealogy



Advisor

Advisor

Yonghan Jung

http://yonghanjung.me

Elias Barenboim

Professor in
Columbia University

Director of CausalAI lab

http://causalai.net

Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

Outline

Outline

1. **Structural Causal Models (SCMs)** — a new science of causality.

Outline

1. **Structural Causal Models (SCMs)** — a new science of causality.

Outline

1. **Structural Causal Models (SCMs)** — a new science of causality.

Outline

1. **Structural Causal Models (SCMs)** — a new science of causality.

2. **Causal effect identification** — what are conditions for estimate causal effects using data?

Outline

1. **Structural Causal Models (SCMs)** — a new science of causality.

2. **Causal effect identification** — what are conditions for estimate causal effects using data?

3. **Causal effect estimation** — how to estimate the causal effect sample efficiently?

Outline

1. **Structural Causal Models (SCMs)** — a new science of causality.

2. **Causal effect identification** — what are conditions for estimate causal effects using data?

3. **Causal effect estimation** — how to estimate the causal effect sample efficiently?

4. **My research themes**

1. Structural Causal Model (SCM)

Fundamental contribution to causal reasoning

Fundamental contribution to causal reasoning

Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

Fundamental contribution to causal reasoning

Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

"Creation of mathematical framework for causal inference"

— *Structural causal model* and its graphical representation using Bayesian networks (https://amturing.acm.org/award_winners/pearl_2658896.cfm)

Fundamental contribution to causal reasoning

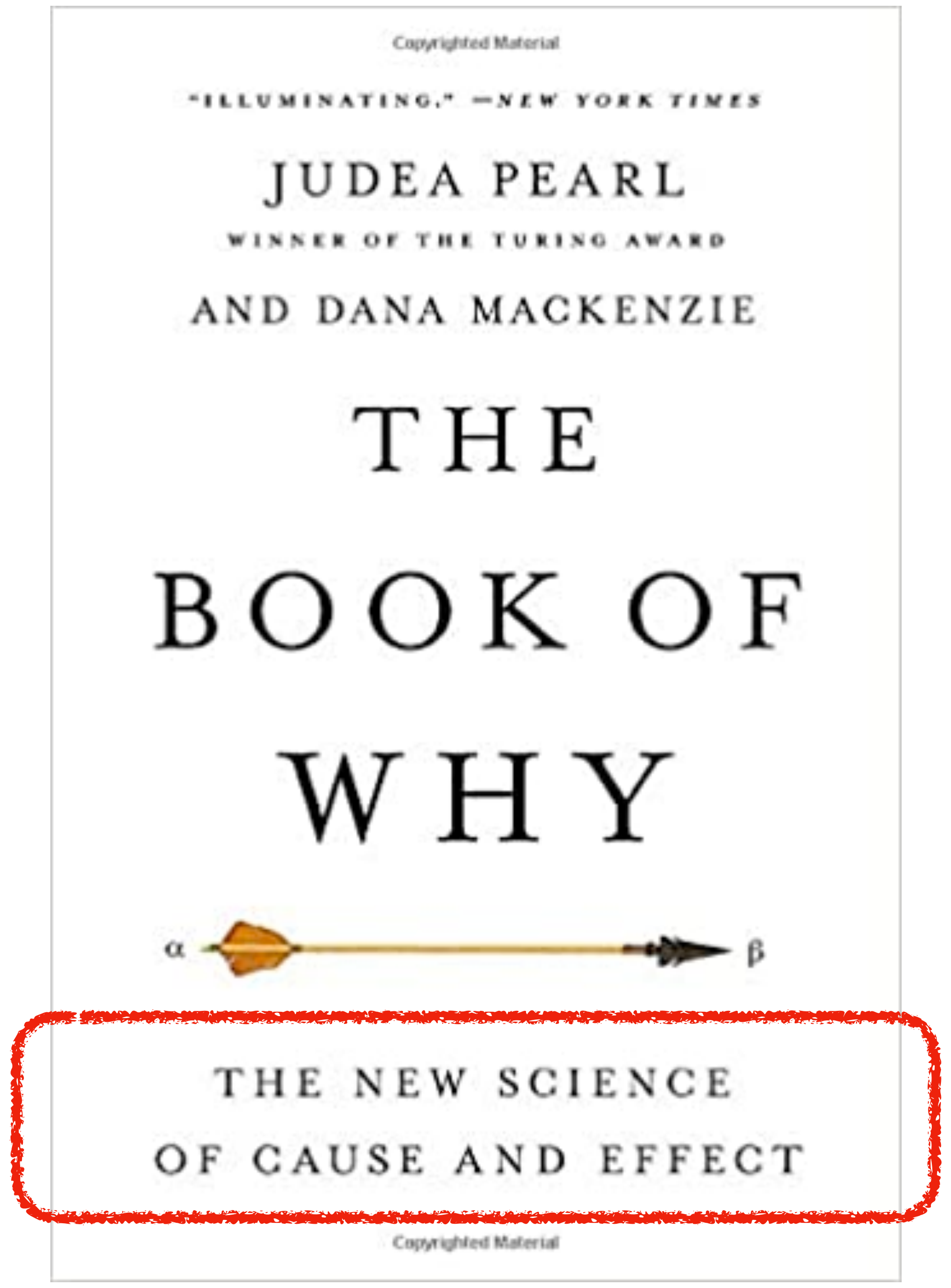
Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

"Creation of mathematical framework for causal inference"

— *Structural causal model* and its graphical representation using Bayesian networks (https://amturing.acm.org/award_winners/pearl_2658896.cfm)



Fundamental contribution to causal reasoning

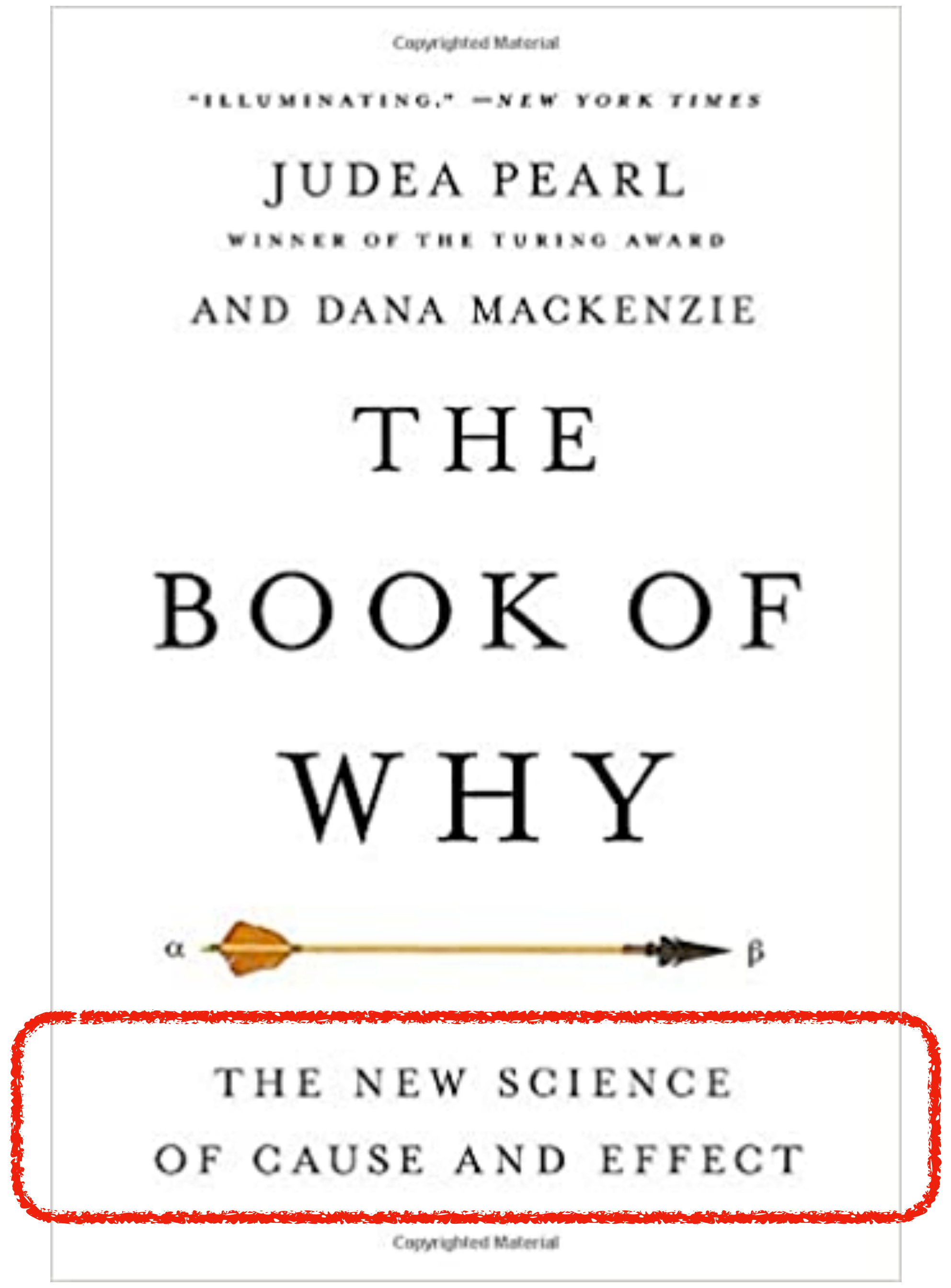
Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

"Creation of mathematical framework for causal inference"

— *Structural causal model* and its graphical representation using Bayesian networks (https://amturing.acm.org/award_winners/pearl_2658896.cfm)



"radical mathematical solution on causality"
— Nature

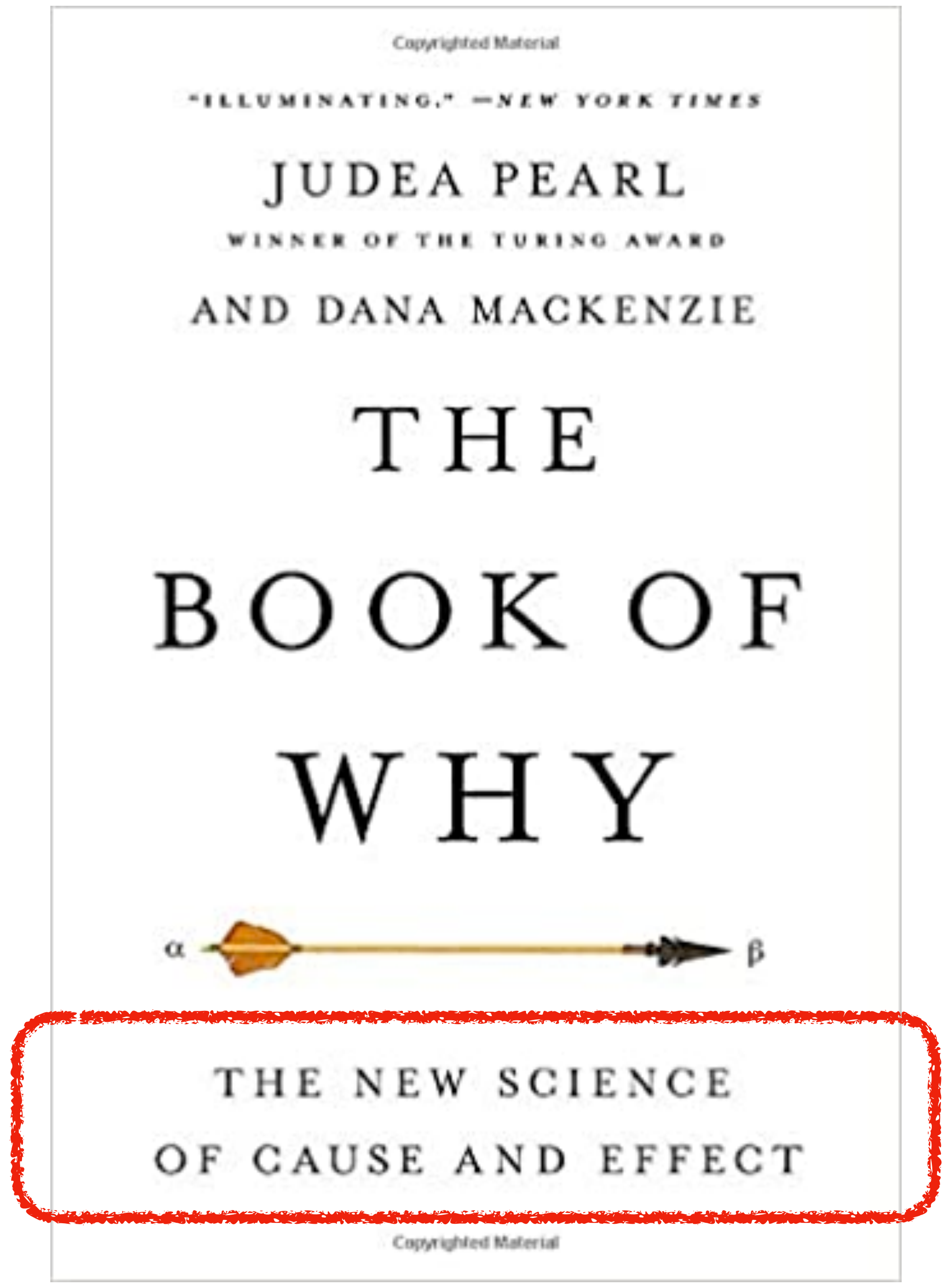
Fundamental contribution to causal reasoning

Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

“Creation of mathematical framework for causal inference"
— *Structural causal model* and its graphical representation using Bayesian networks (https://amturing.acm.org/award_winners/pearl_2658896.cfm)



“**radical mathematical solution on causality**”
— Nature

“***wonderful book has illuminating answers***”
— Daniel Kahneman, winner of the Nobel Memorial Prize in Economic Sciences

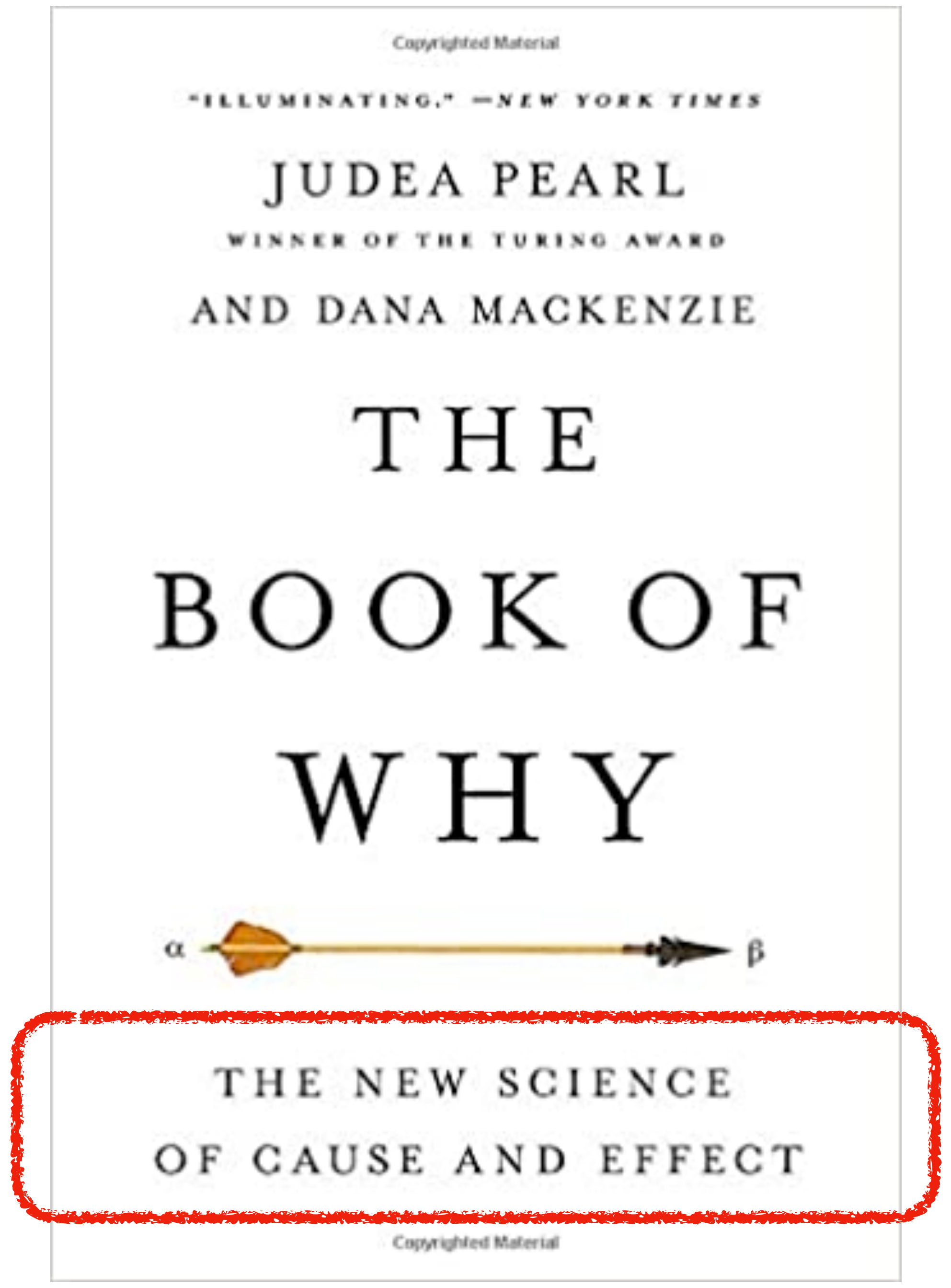
Fundamental contribution to causal reasoning

Judea Pearl

Professor in UCLA

Recipient of Turing Award 🏆

"Creation of mathematical framework for causal inference"
— *Structural causal model* and its graphical representation using Bayesian networks (https://amturing.acm.org/award_winners/pearl_2658896.cfm)



"**radical mathematical solution on causality**"
— Nature

"***wonderful book has illuminating answers***"
— Daniel Kahneman, winner of the Nobel Memorial Prize in Economic Sciences

"***elegant, powerful, controversial theory of causality***"
— American Mathematical Society

Many disagreements!

Many disagreements!

🔥 Kareem Carr 🔥
@kareem_carr

Controversial opinion: Causal inference is just another kind of statistical inference.

11:23 AM · Jul 30, 2021 · Twitter for iPhone

Many disagreements!

🔥 **Kareem Carr** 🔥
@kareem_carr

Controversial opinion: Causal inference is just another kind of statistical inference.

11:23 AM · Jul 30, 2021 · Twitter for iPhone

Mischaracterizations of statistics and statisticians

As noted above, Pearl and Mackenzie have a habit of putting down statisticians in a way that seems to reflect ignorance of our field.

Many disagreements!

Kareem Carr
@kareem_carr

Controversial opinion: Causal infere
kind of statistical inference.

11:23 AM · Jul 30, 2021 · Twitter for iPhone

Mischaracterizations of statistics and statisticians

As noted above, Pearl and Mackenzie have a habit of putting down
flect ignorance of our field.

Miguel Hernán
@_MiguelHernan

Replying to @_MiguelHernan

To define a causal effect, describe the hypothetical randomized experiment that you'd conduct to quantify the effect. If you can't describe the experiment (#TargetTrial), chances are you don't know what causal effect you are after.

My view, not Pearl's.
academic.oup.com/aje/article/18...

Many disagreements!

Kareem Carr
@kareem_carr

Controversial opinion: Causal infere…
kind of statistical inference.

11:23 AM · Jul 30, 2021 · Twitter for iPhone

Mischaracterizations of statistics and statisticians

As noted above, Pearl and Mackenzie have a habit of putting down
…flect ignorance of our field.

Miguel Hernán
@_MiguelHernan

Replying to @_MiguelHernan

To define a causal effect, describe the hypothetical randomized experiment that you'd conduct to quantify the effect. If you can't describe the experiment (#TargetTrial), chances are you don't know what causal effect you are after.

My view, not Pearl's.
academic.oup.com/aje/article/18…

"Is it a new science on Causality?"

Do we understand causality?

Do we understand causality?

Studied an association

Original Investigation | Nutrition, Obesity, and Exercise

February 15, 2019

Association Between Push-up Exercise Capacity and Future Cardiovascular Events Among Active Adult Men

Justin Yang, MD, MPH[1,2]; Costas A. Christophi, PhD[1,3]; Andrea Farioli, MD, PhD[4]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2019;2(2):e188341. doi:10.1001/jamanetworkopen.2018.8341

Do we understand causality?

Studied an association

Original Investigation | Nutrition, Obesity, and Exercise

February 15, 2019

Association Between Push-up Exercise Capacity and Future Cardiovascular Events Among Active Adult Men

Justin Yang, MD, MPH[1,2]; Costas A. Christophi, PhD[1,3]; Andrea Farioli, MD, PhD[4]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2019;2(2):e188341. doi:10.1001/jamanetworkopen.2018.8341

Interpreted as causation

STAYING HEALTHY

More push-ups may mean less risk of heart problems

May 01, 2019

How Pushups Can Help Men's Hearts

By Matt McMillen

Medically Reviewed by Michael W. Smith, MD on March 26, 2019

Do we understand causality?

Studied an association

Original Investigation | Nutrition, Obesity, and Exercise

February 15, 2019

Association Between Push-up Exercise Capacity and Future Cardiovascular Events Among Active Adult Men

Justin Yang, MD, MPH[1,2]; Costas A. Christophi, PhD[1,3]; Andrea Farioli, MD, PhD[4]; et al

Author Affiliations | Article Information

JAMA Netw Open. 2019;2(2):e188341. doi:10.1001/jamanetworkopen.2018.8341

Interpreted as causation

STAYING HEALTHY

More push-ups may mean less risk of heart problems

May 01, 2019

How Pushups Can Help Men's Hearts

By Matt McMillen

Medically Reviewed by Michael W. Smith, MD on March 26, 2019

What the public understood:

푸쉬업 하면 심장에 좋다는거죠??

So, what is causality? (1) — Correlation

So, what is causality? (1) — Correlation

David Hume

So, what is causality? (1) — Correlation

David Hume

"We may define a ***cause*** to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second" (1752)

So, what is causality? (1) — Correlation

David Hume

"We may define a ***cause*** to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second" (1752)

Roughly, if X happens and then Y happens, then X is a cause of Y.

So, what is causality? (1) — Correlation

David Hume

"We may define a ***cause*** to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second" (1752)

Roughly, if X happens and then Y happens, then X is a cause of Y.

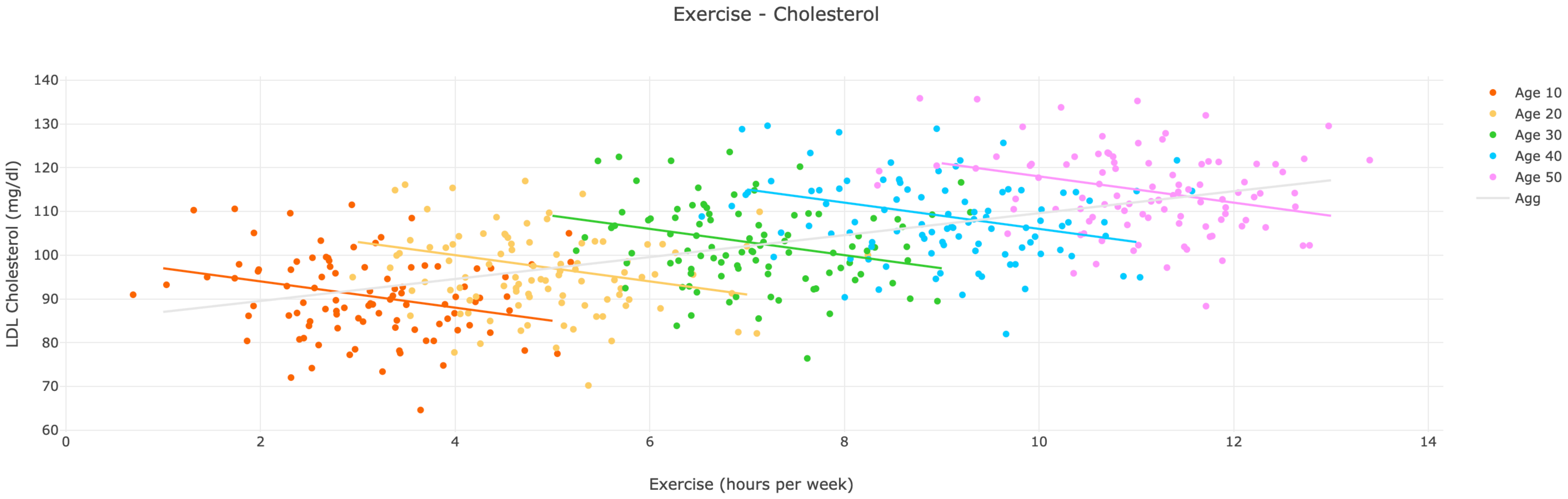
Correlation implies causation?

So, what is causality? (1) — Correlation

So, what is causality? (1) — Correlation

↑Exercise ⇒ ↑Cholesterol?

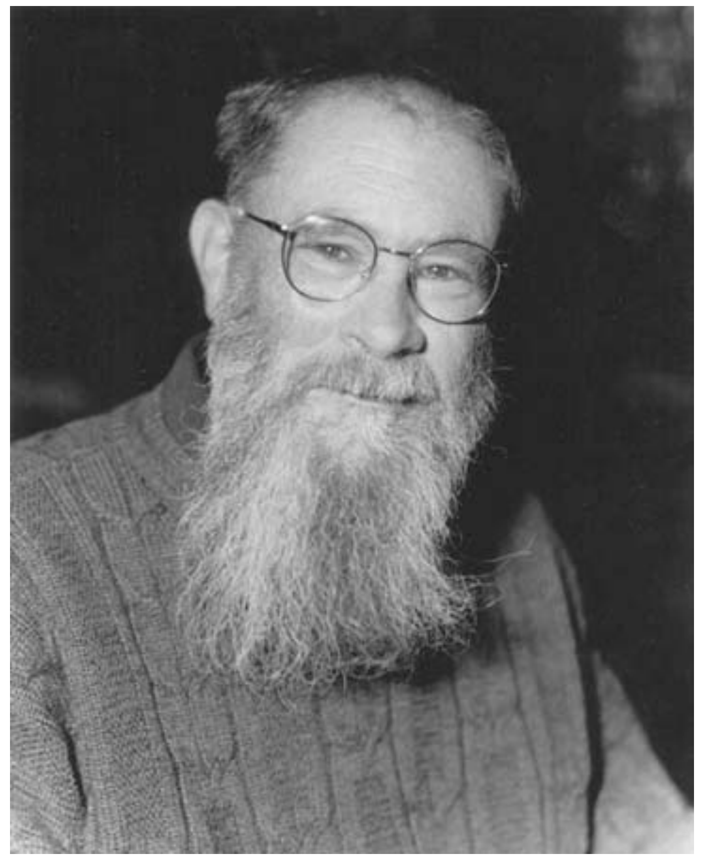
So, what is causality? (1) — Correlation



↑Exercise ⇒ ↓Cholesterol per age!

What is causality? (2) — Counterfactual

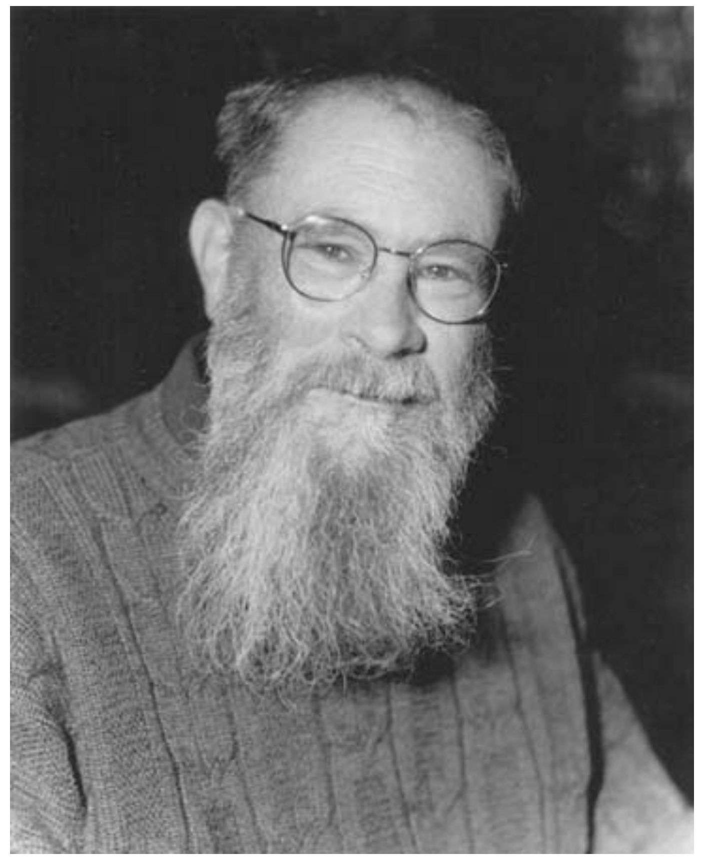
What is causality? (2) — Counterfactual



David Lewis

Donald Rubin

What is causality? (2) — Counterfactual

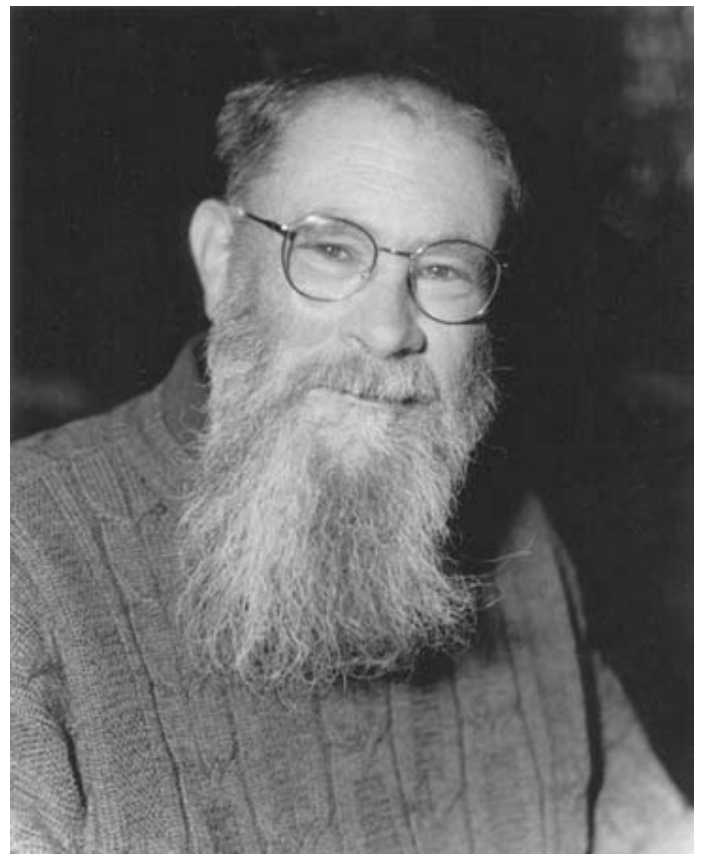


David Lewis

Donald Rubin

**Counterfactual (Lewis, 1973) or
PO-based causality (PO, Rubin, 1974)**

What is causality? (2) — Counterfactual



David Lewis

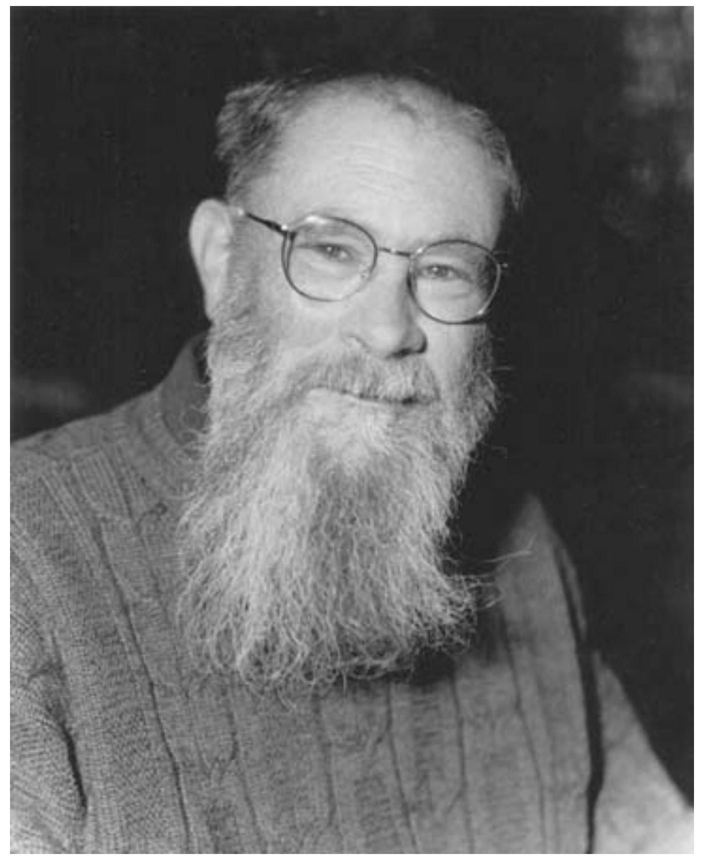
Donald Rubin

Counterfactual (Lewis, 1973) or
PO-based causality (PO, Rubin, 1974)

X is a cause of an outcome Y means

- If X had occurred, then Y would have occurred; and
- If X had not occurred, then Y would not have occurred;

What is causality? (2) — Counterfactual



David Lewis

Donald Rubin

Counterfactual (Lewis, 1973) or
PO-based causality (PO, Rubin, 1974)

X is a cause of an outcome Y means

- If X had occurred, then Y would have occurred; and
- If X had not occurred, then Y would not have occurred;

Potential outcome: Let Y_x denote Y if X had been set to x.

- X is a cause of Y if $Y_{X=1} = 1$ & $Y_{X=0} = 0$.

What is causality? (2) — Counterfactual (Example)

X is a cause of an outcome Y means

- If X had occurred, then Y would have occurred; and
- If X had not occurred, then Y would not have occurred;

What is causality? (2) — Counterfactual (Example)

X is a cause of an outcome Y means

- If X had occurred, then Y would have occurred; and
- If X had not occurred, then Y would not have occurred;

$X = 1$ $\qquad$ $Y_{X=1} = 1$

What is causality? (2) — Counterfactual (Example)

X is a cause of an outcome Y means

- If X had occurred, then Y would have occurred; and
- If X had not occurred, then Y would not have occurred;

$X = 1$ $\qquad$ $Y_{X=1} = 1$

$X = 0$ $\qquad$ $Y_{X=0} = 0$

What is causality? (2) — Counterfactual (Example)

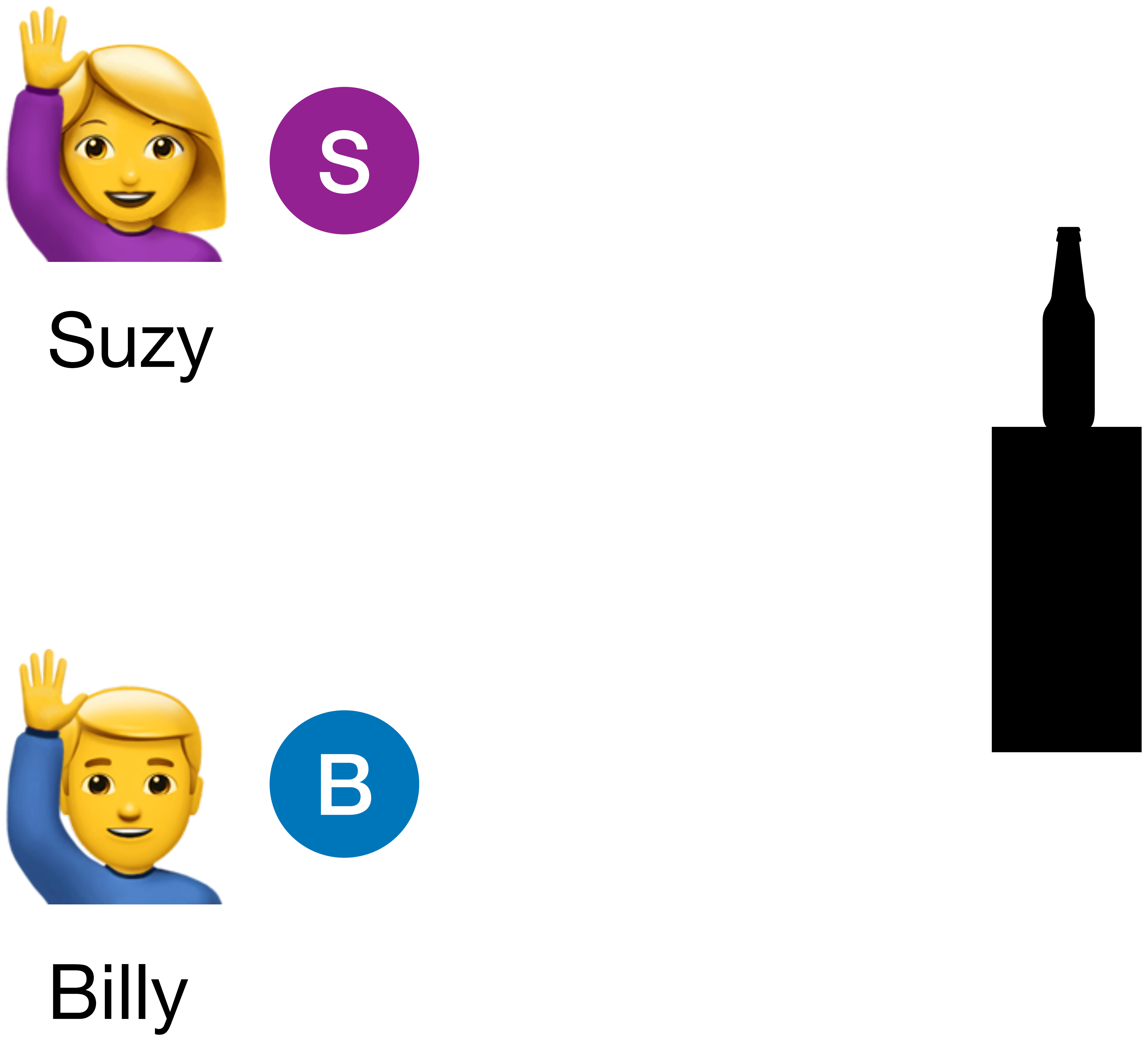
Sounds reasonable… Is indeed the PO-based definition capturing causation?

$X = 1$ $\qquad$ $Y_{X=1} = 1$

$X = 0$ $\qquad$ $Y_{X=0} = 0$

Peculiarity in Potential Outcome - 1

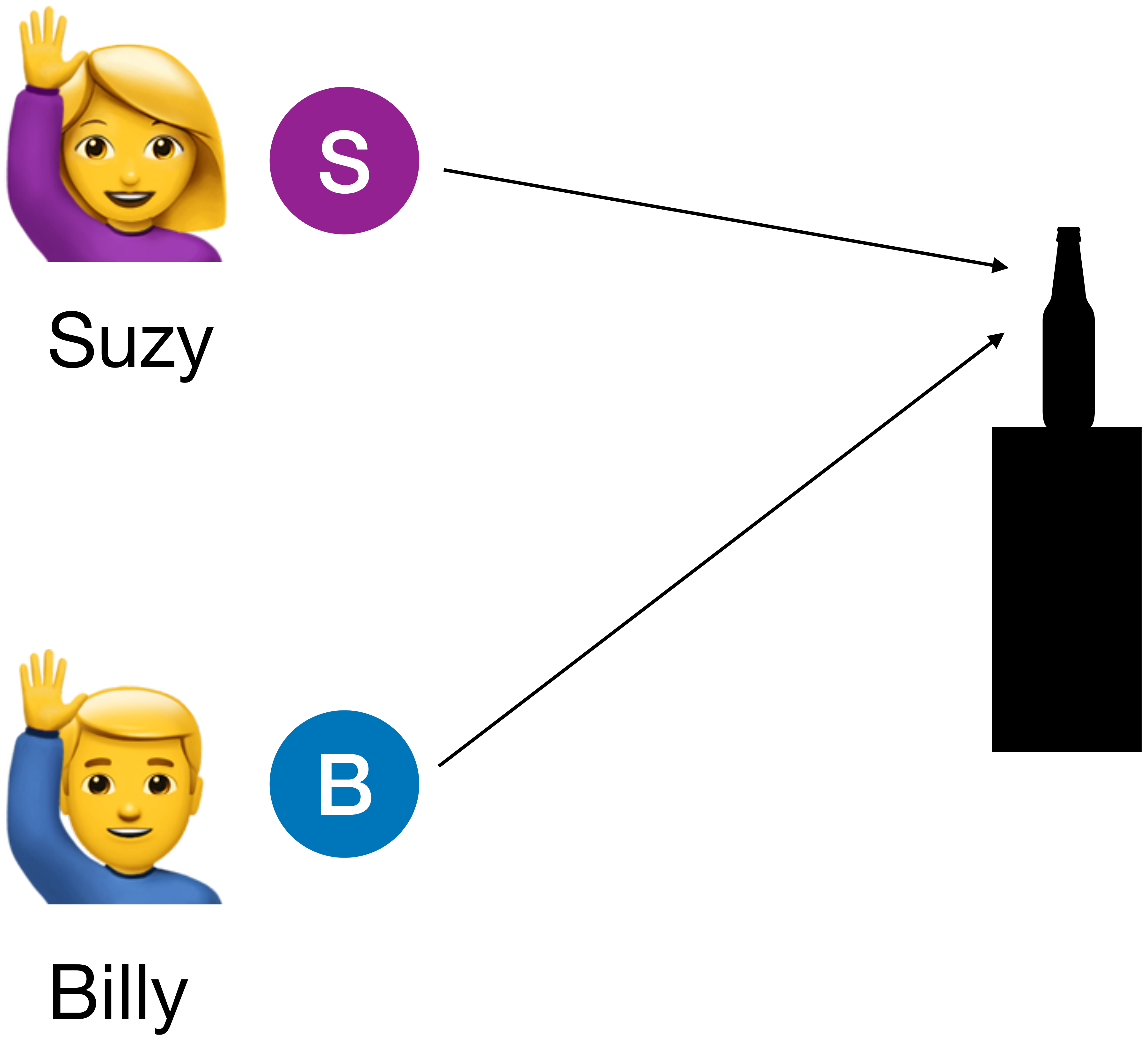
Peculiarity in Potential Outcome - 1



- Suzy and Billy throw the ball to the bottle on the tower

This example is from Lewis (2000)

Peculiarity in Potential Outcome - 1

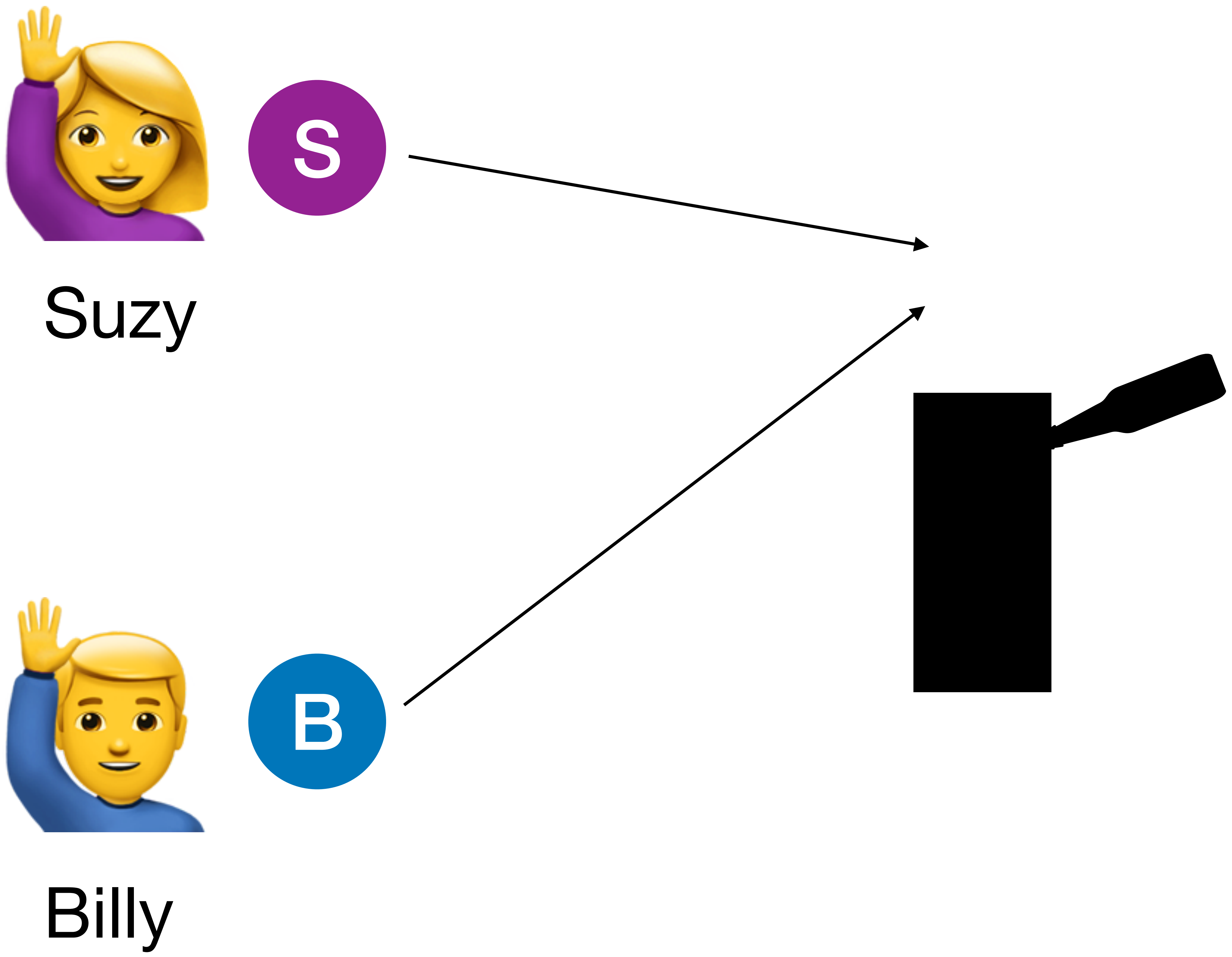


- Suzy and Billy throw the ball to the bottle on the tower

- They threw accurately to the bottle.

This example is from Lewis (2000)

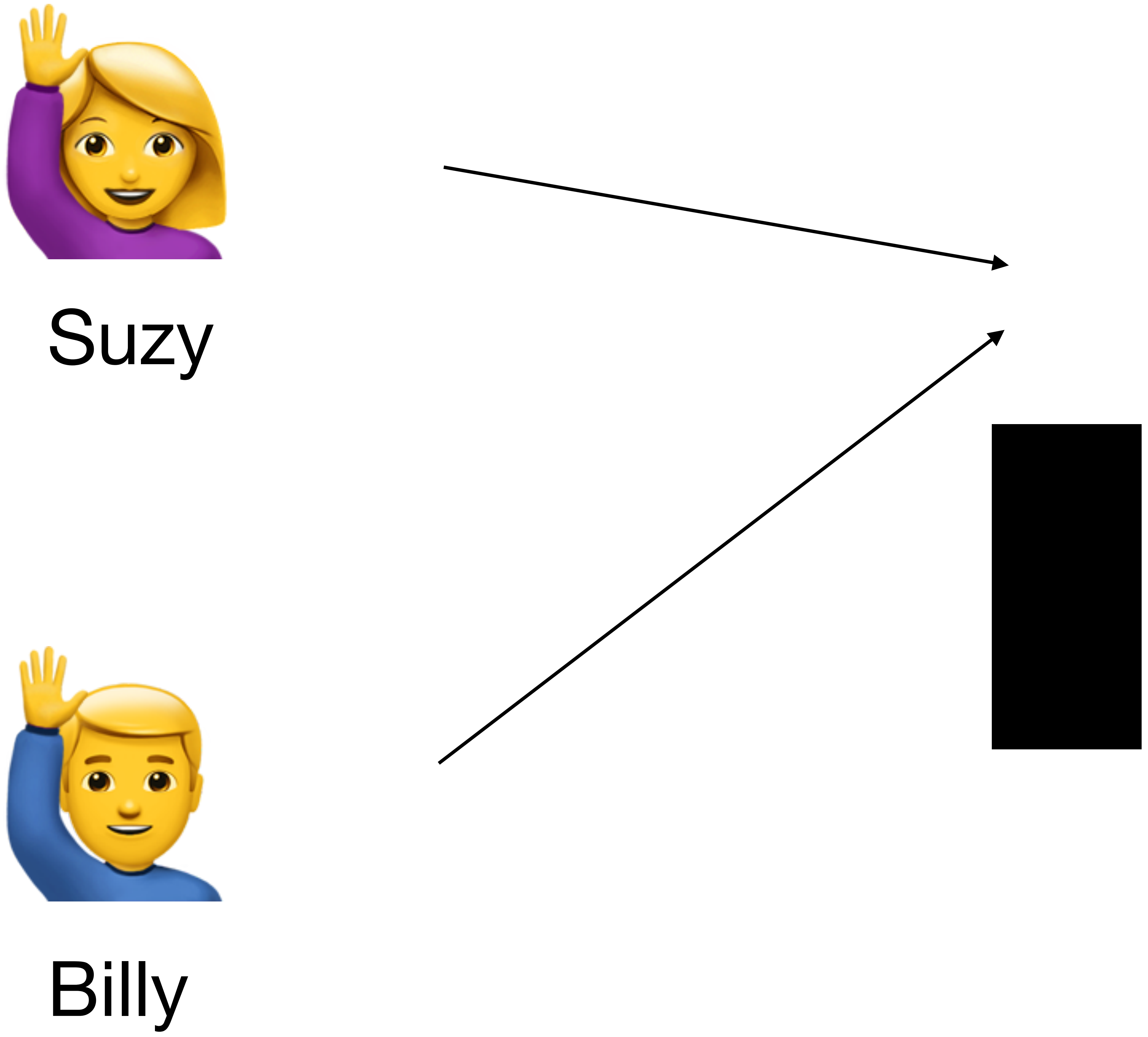
Peculiarity in Potential Outcome - 1



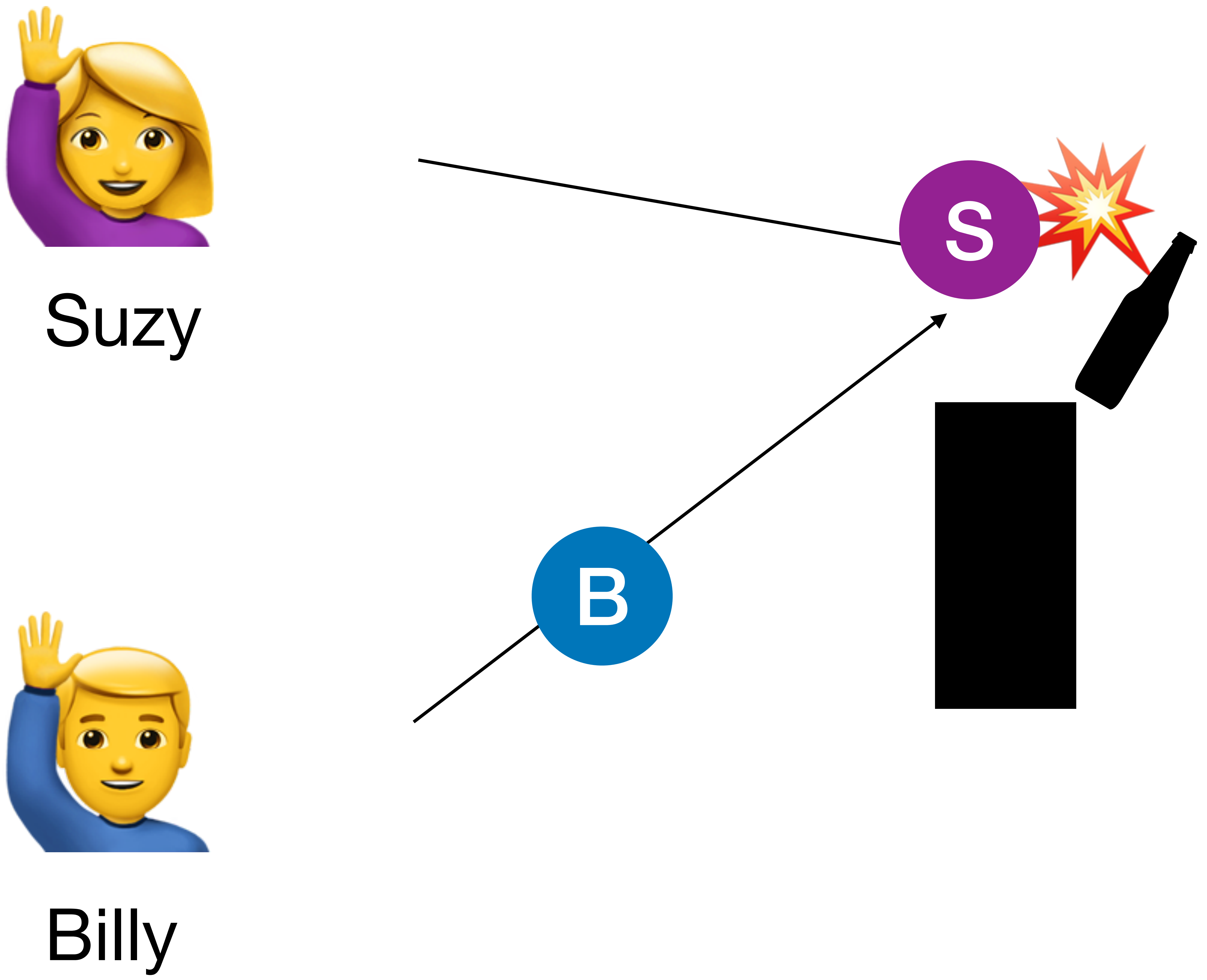
- Suzy and Billy throw the ball to the bottle on the tower
- They threw accurately to the bottle.
- The bottle will fall off once got hit.

This example is from Lewis (2000)

Peculiarity in Potential Outcome - 2

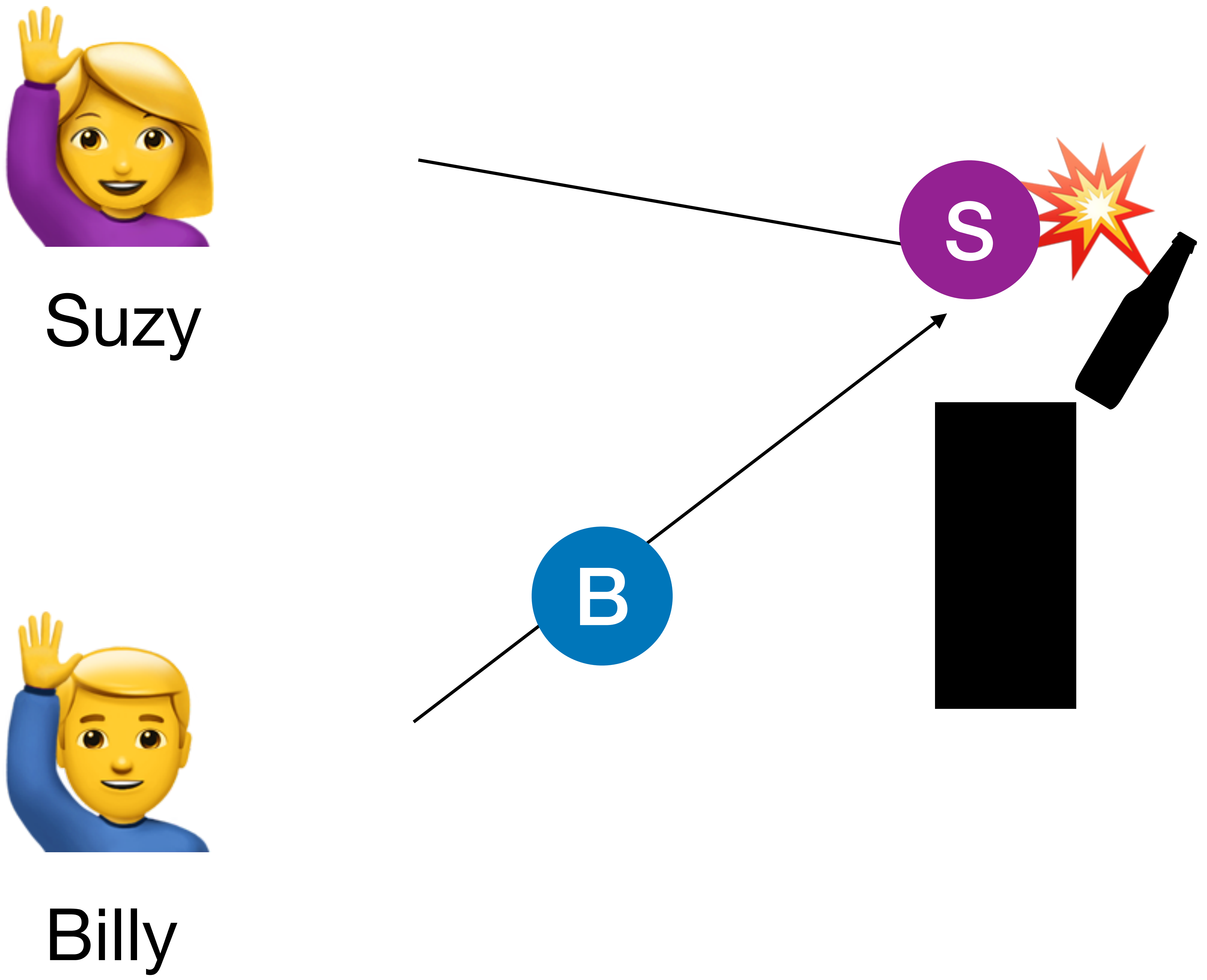


Peculiarity in Potential Outcome - 2



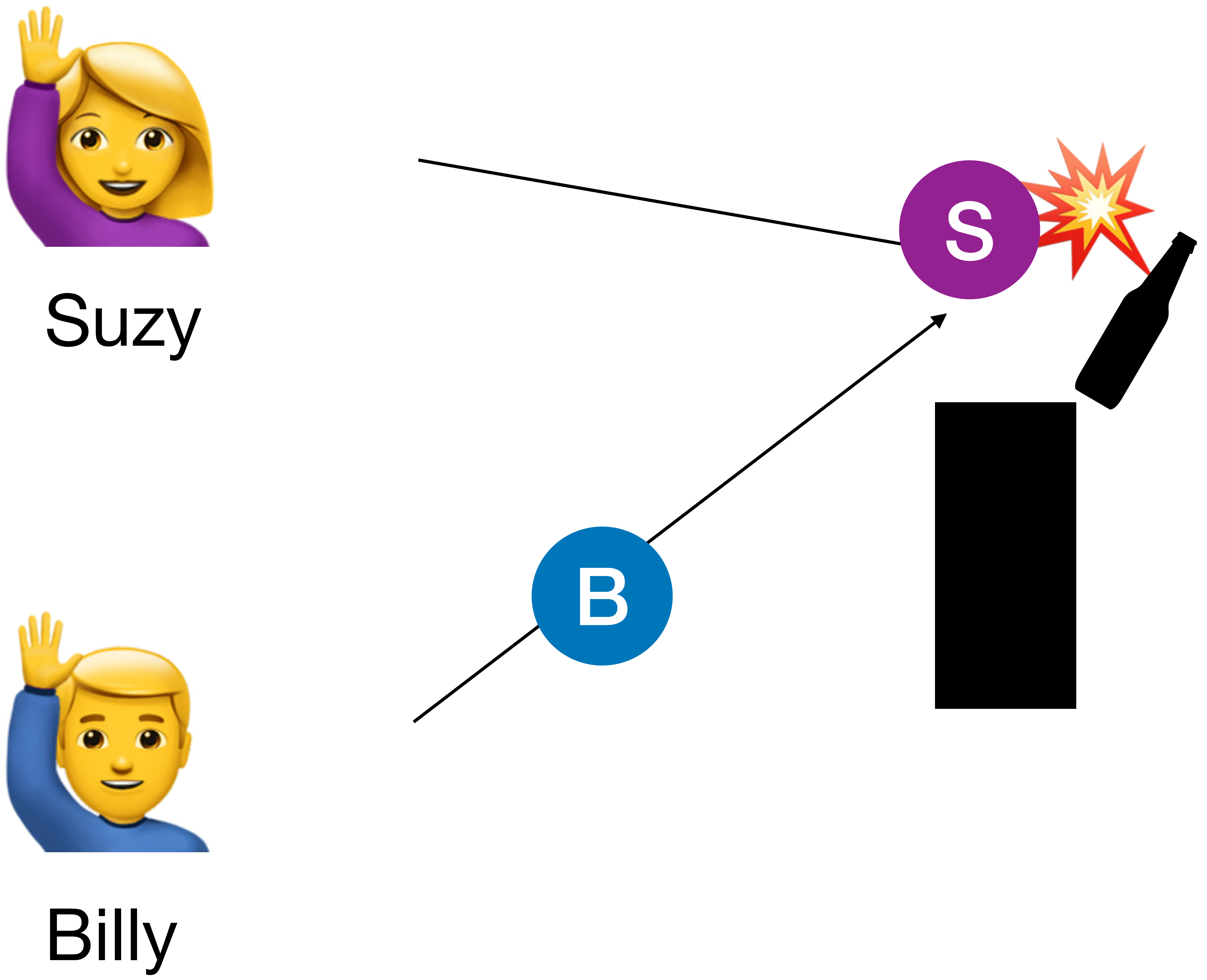
- Suppose Suzy's ball hits the bottle first.

Peculiarity in Potential Outcome - 2



- Suppose Suzy’s ball hits the bottle first.
- Then, Billy’s ball doesn’t hit the bottle.

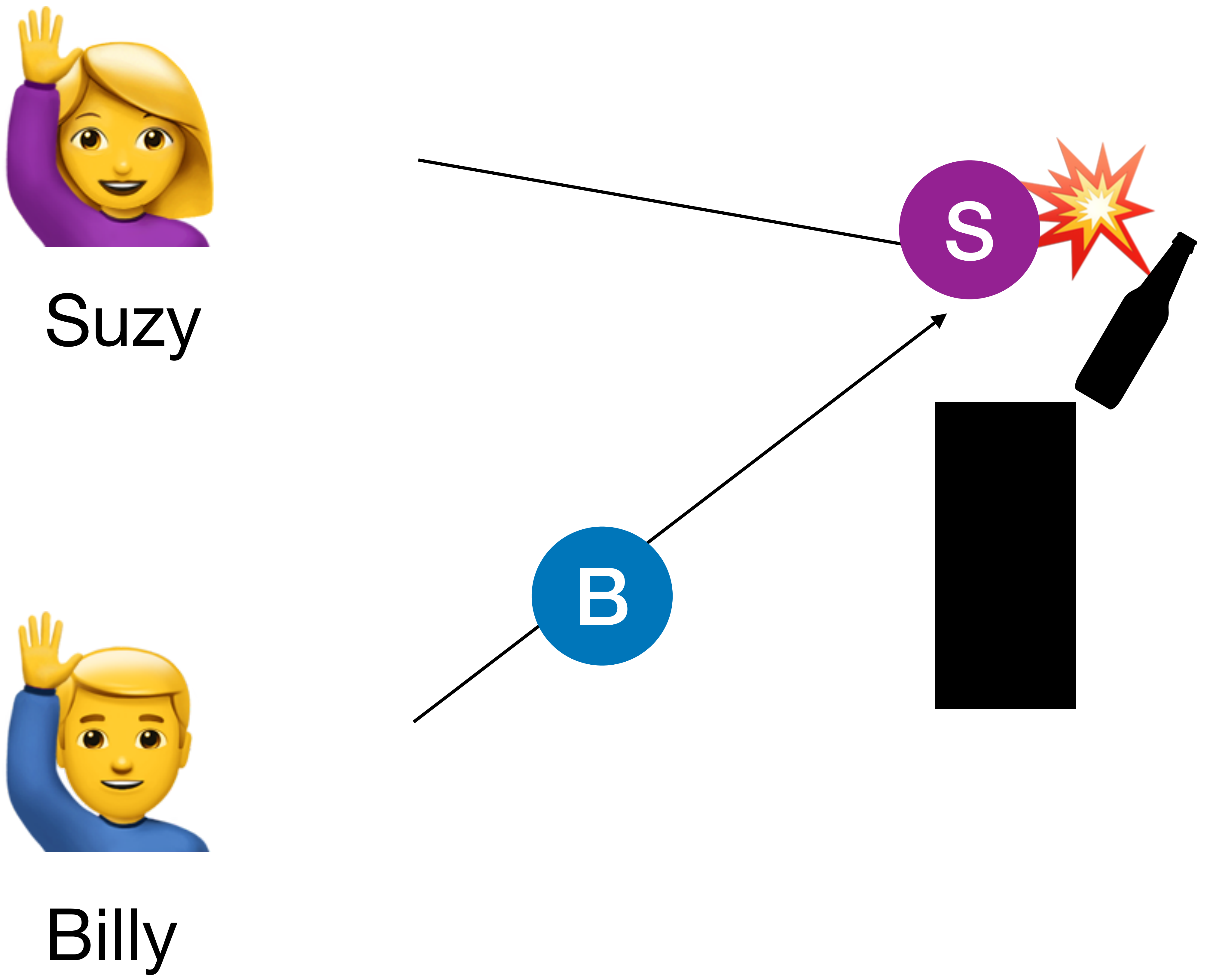
Peculiarity in Potential Outcome - 2



- Suppose Suzy's ball hits the bottle first.
- Then, Billy's ball doesn't hit the bottle.

Q. Is Suzy throwing a ball a cause of the bottle falling off?

Peculiarity in Potential Outcome - 2

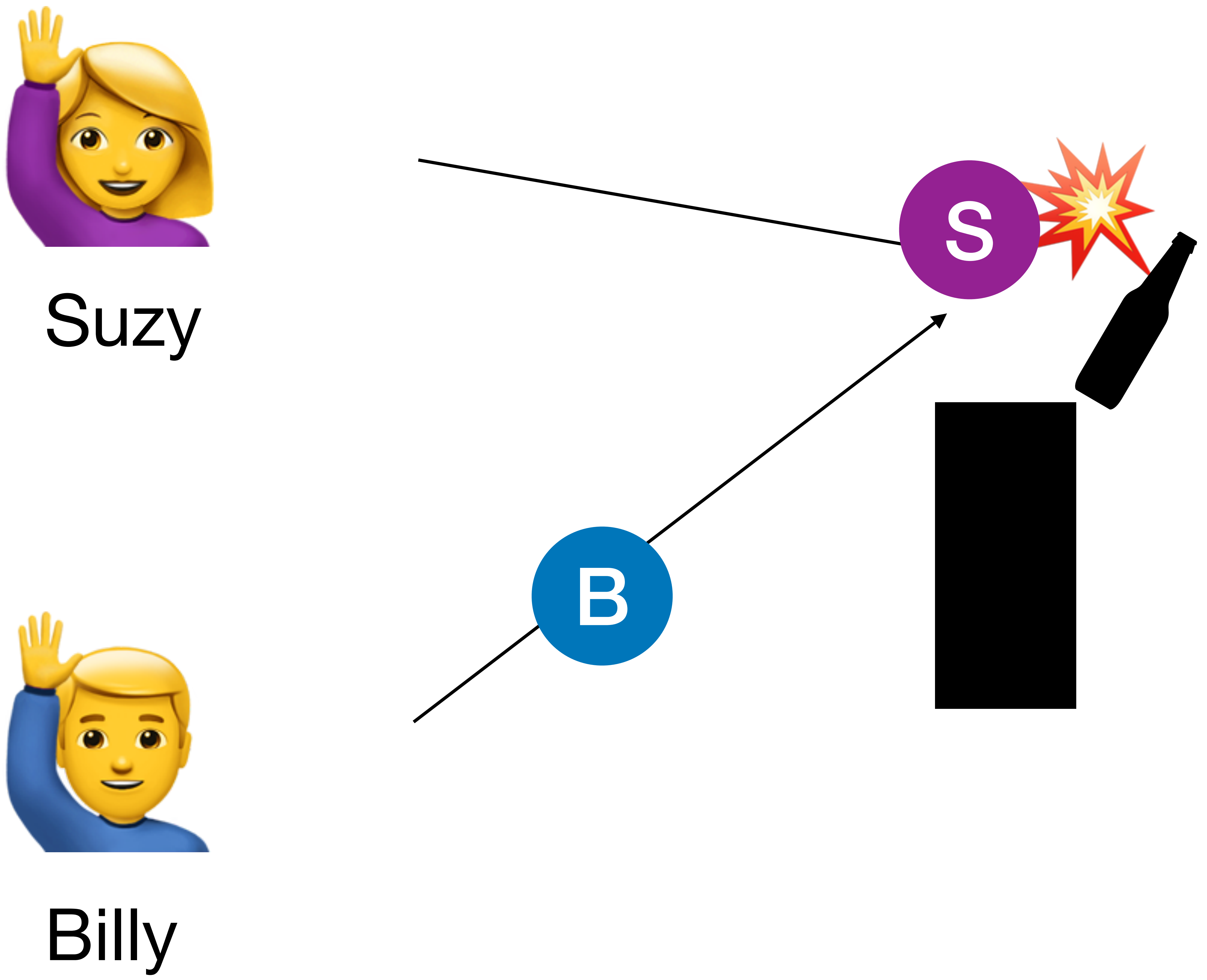


- Suppose Suzy's ball hits the bottle first.
- Then, Billy's ball doesn't hit the bottle.

Q. Is Suzy throwing a ball a cause of the bottle falling off?

A. Yes. Because Suzy threw a ball, it hits the bottle, and the bottle fell off.

Peculiarity in Potential Outcome - 2



- Suppose Suzy's ball hits the bottle first.
- Then, Billy's ball doesn't hit the bottle.

Q. Is Suzy throwing a ball a cause of the bottle falling off?

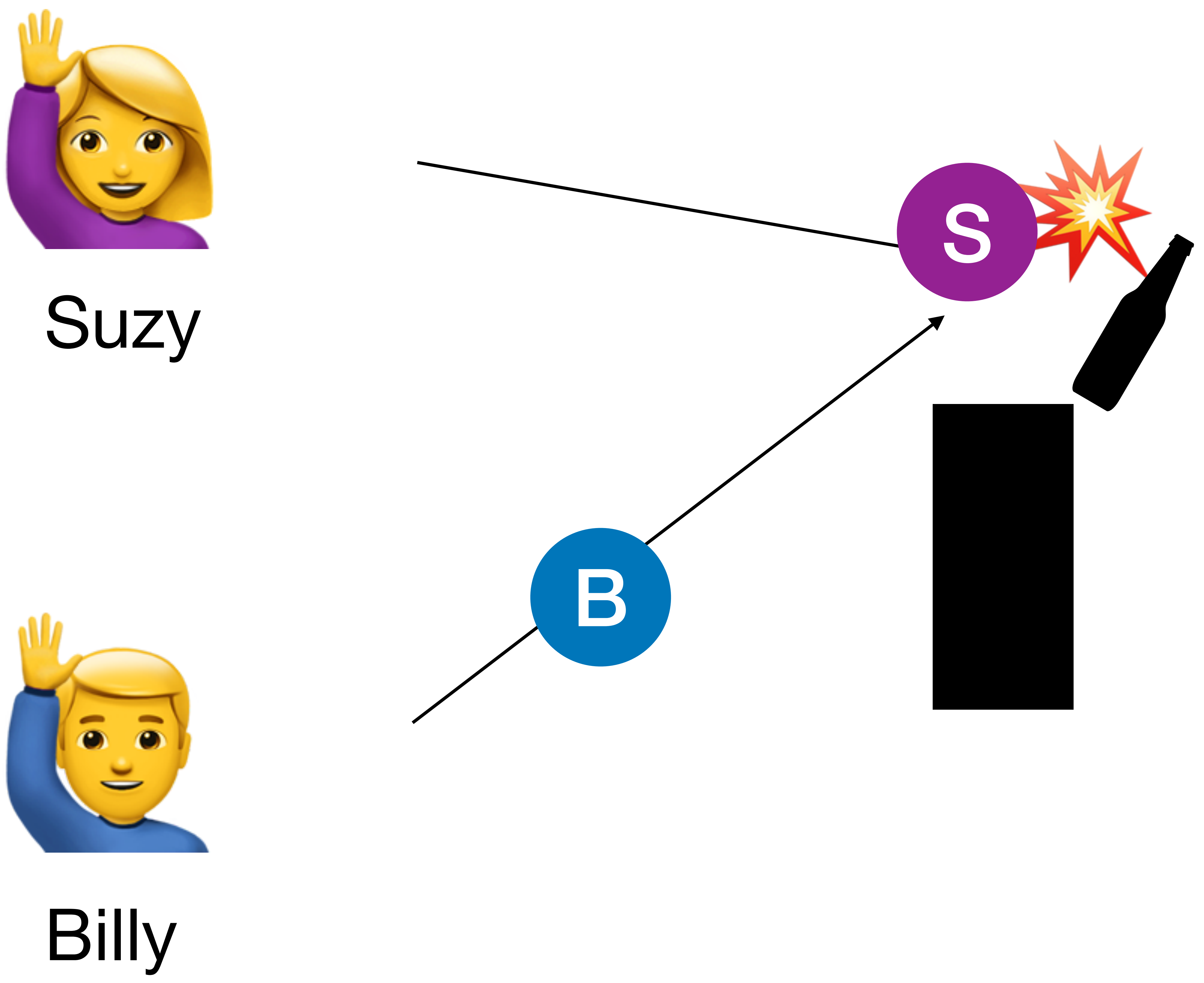
A. Yes. Because Suzy threw a ball, it hits the bottle, and the bottle fell off.

Too obvious!

Peculiarity in Potential Outcome - 3

What would the PO-based causality say?

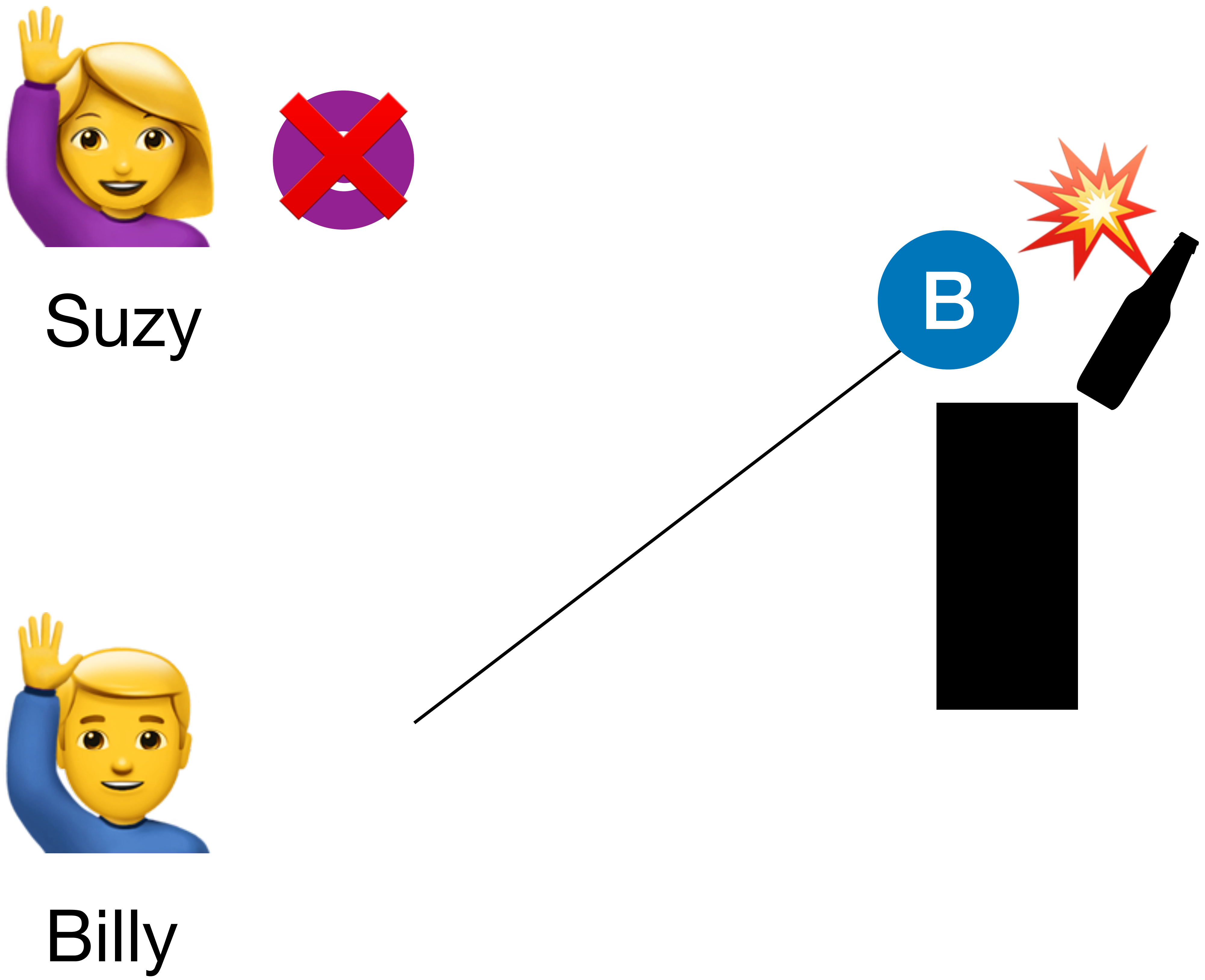
Peculiarity in Potential Outcome - 3



What would the PO-based causality say?

- If Suzy had thrown the ball ($ST = T$), the bottle would fall off ($BFO = T$).

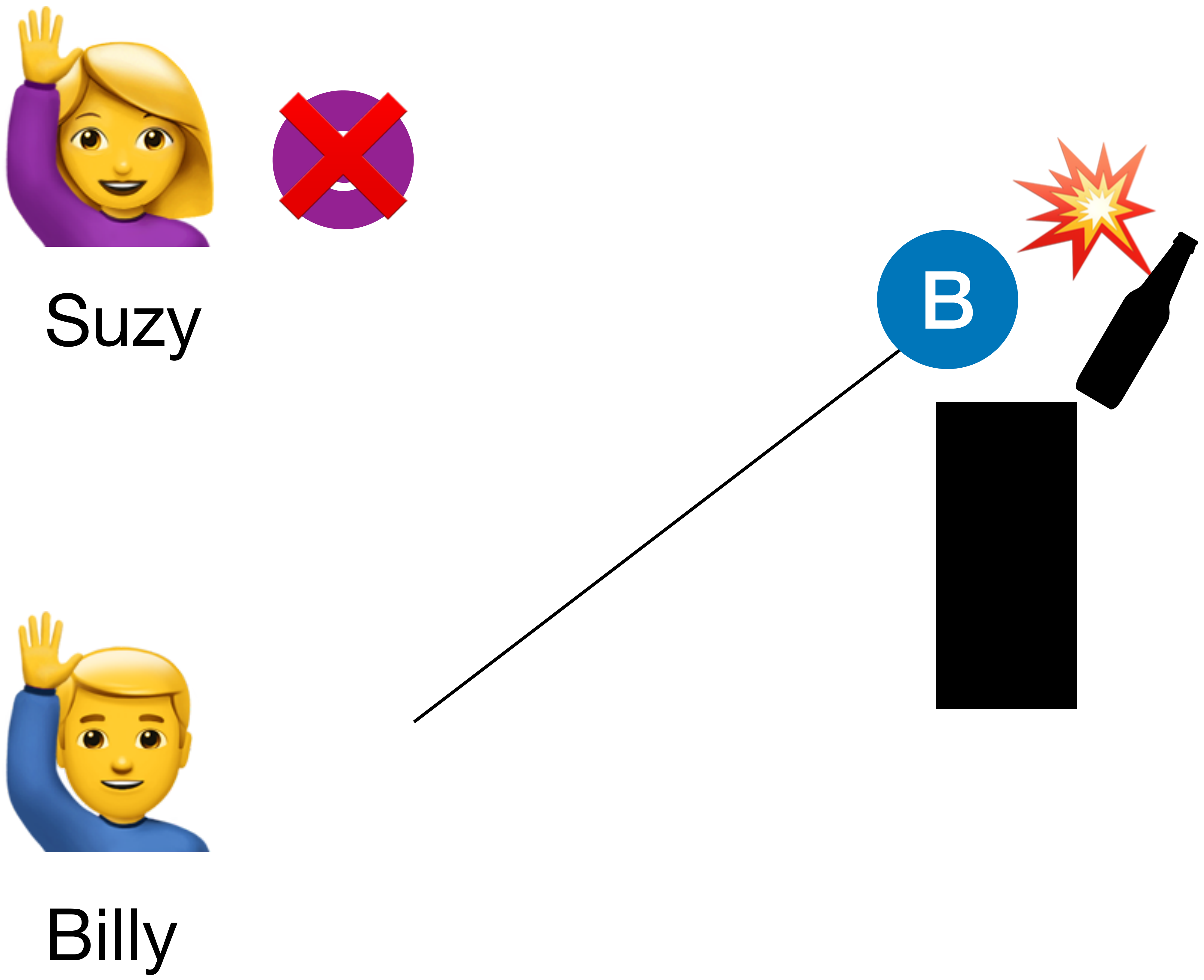
Peculiarity in Potential Outcome - 3



What would the PO-based causality say?

- If Suzy had thrown the ball ($ST = T$), the bottle would fall off ($BFO = T$).
- If Suzy hadn't thrown the ball ($ST = F$), the bottle would fall off ($BFO = T$).

Peculiarity in Potential Outcome - 3



What would the PO-based causality say?

- If Suzy had thrown the ball ($ST = T$), the bottle would fall off ($BFO = T$).
- If Suzy hadn't thrown the ball ($ST = F$), the bottle would fall off ($BFO = T$).
- $ST = T \rightarrow BFO = T$ and
  $ST = F \rightarrow BFO = T$

Peculiarity in Potential Outcome - 3

Suzy

Billy

What would the PO-based causality say?

T), the

$= F$), the

By the PO-based causality definition, Suzy's ball throwing is not a cause, because $ST = 1$ or $ST = 0$ doesn't make any change.

$$ST = F \rightarrow BFO = T$$

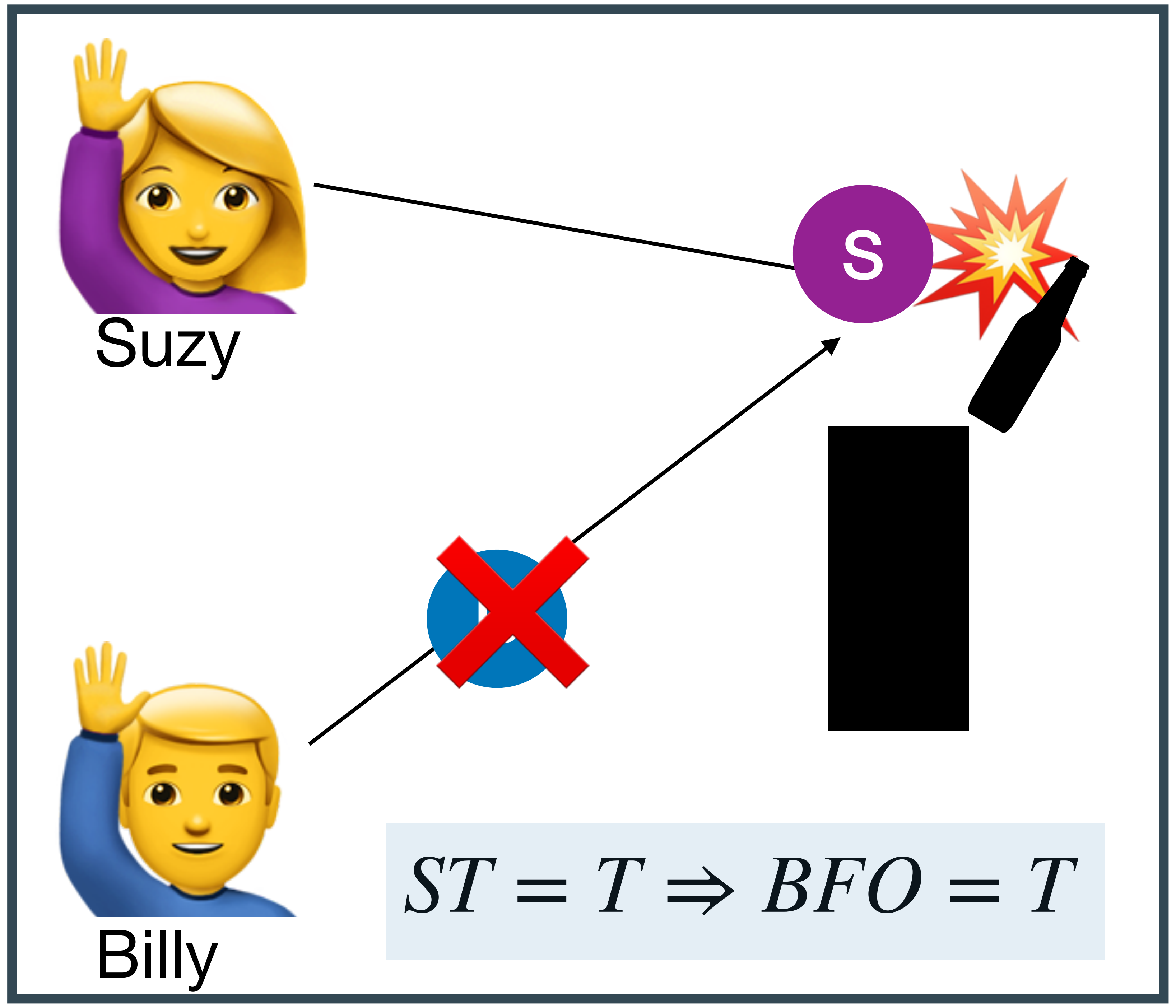
Toward a modern causality

Toward a modern causality

The counterfactual only holds under the situation where Billy’s ball didn’t hit the bottle.

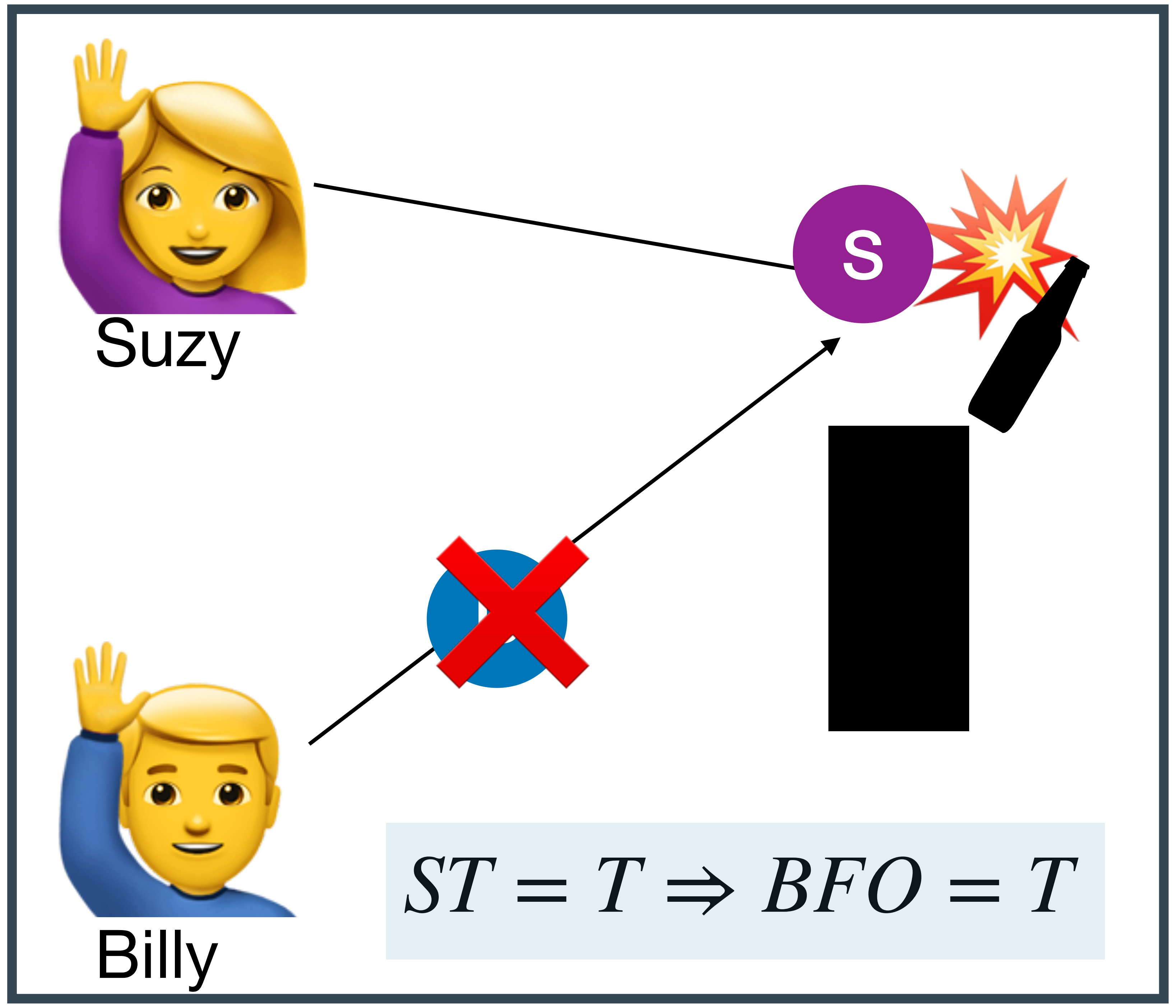
Toward a modern causality

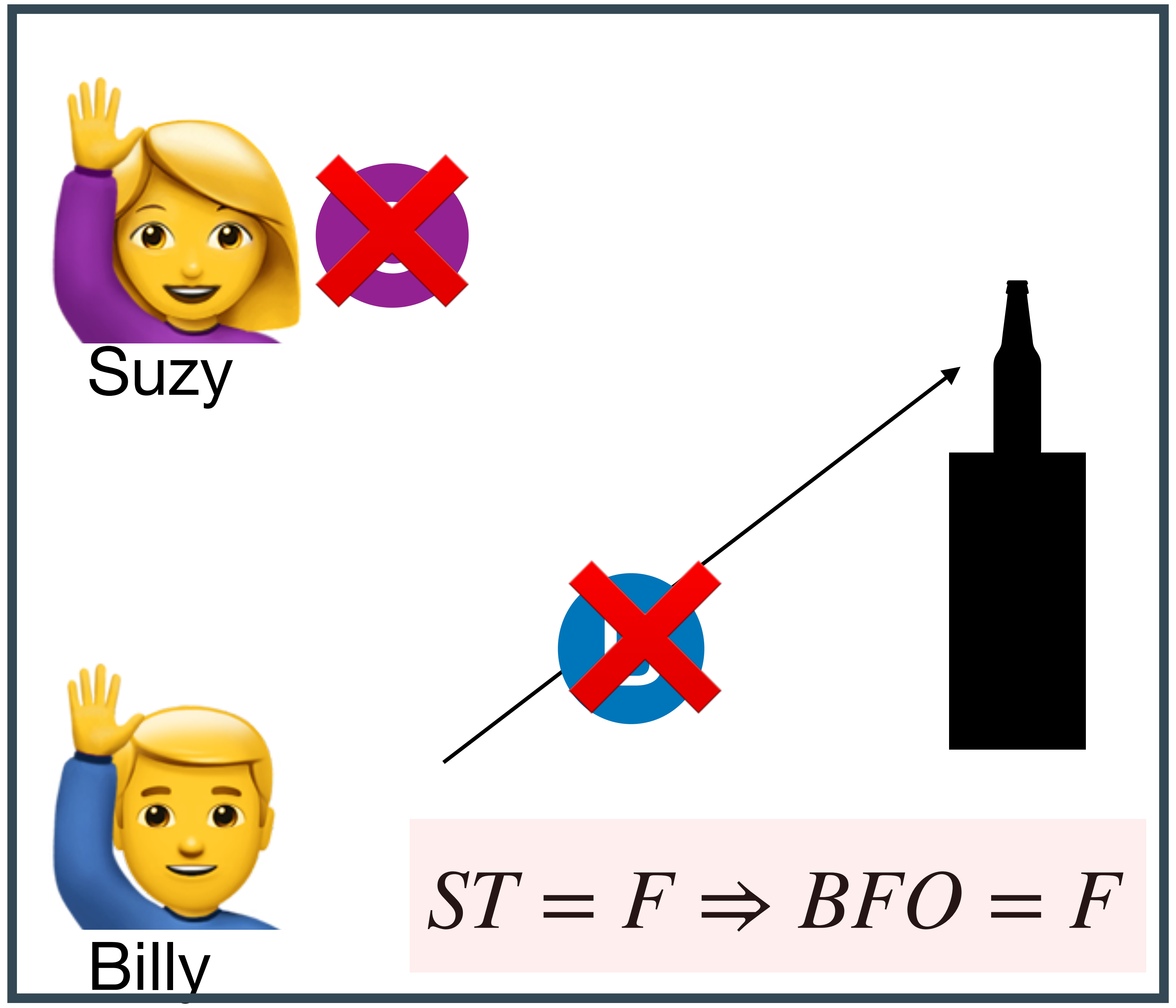
The counterfactual only holds under the situation where Billy's ball didn't hit the bottle.



Toward a modern causality

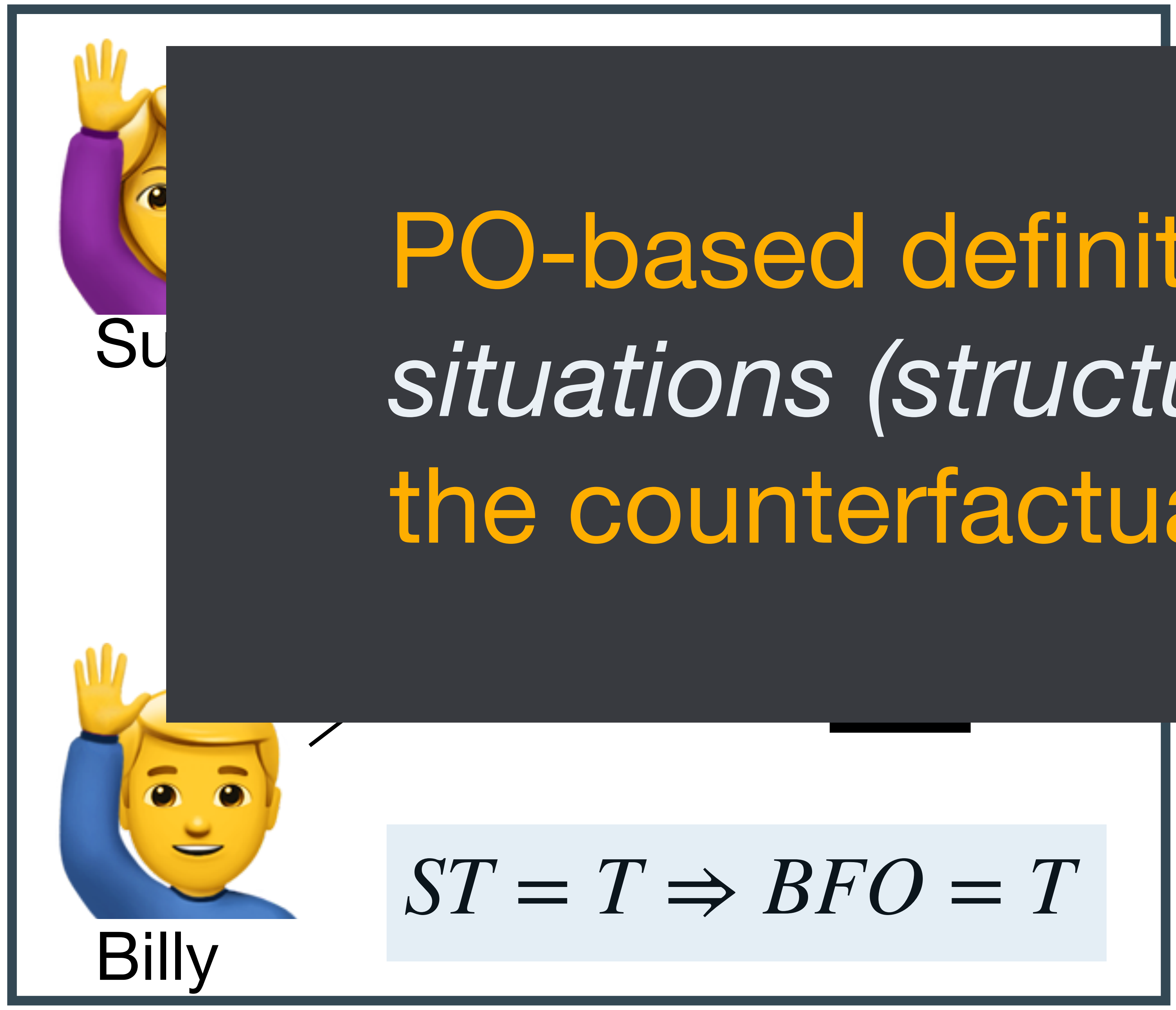
The counterfactual only holds under the situation where Billy's ball didn't hit the bottle.

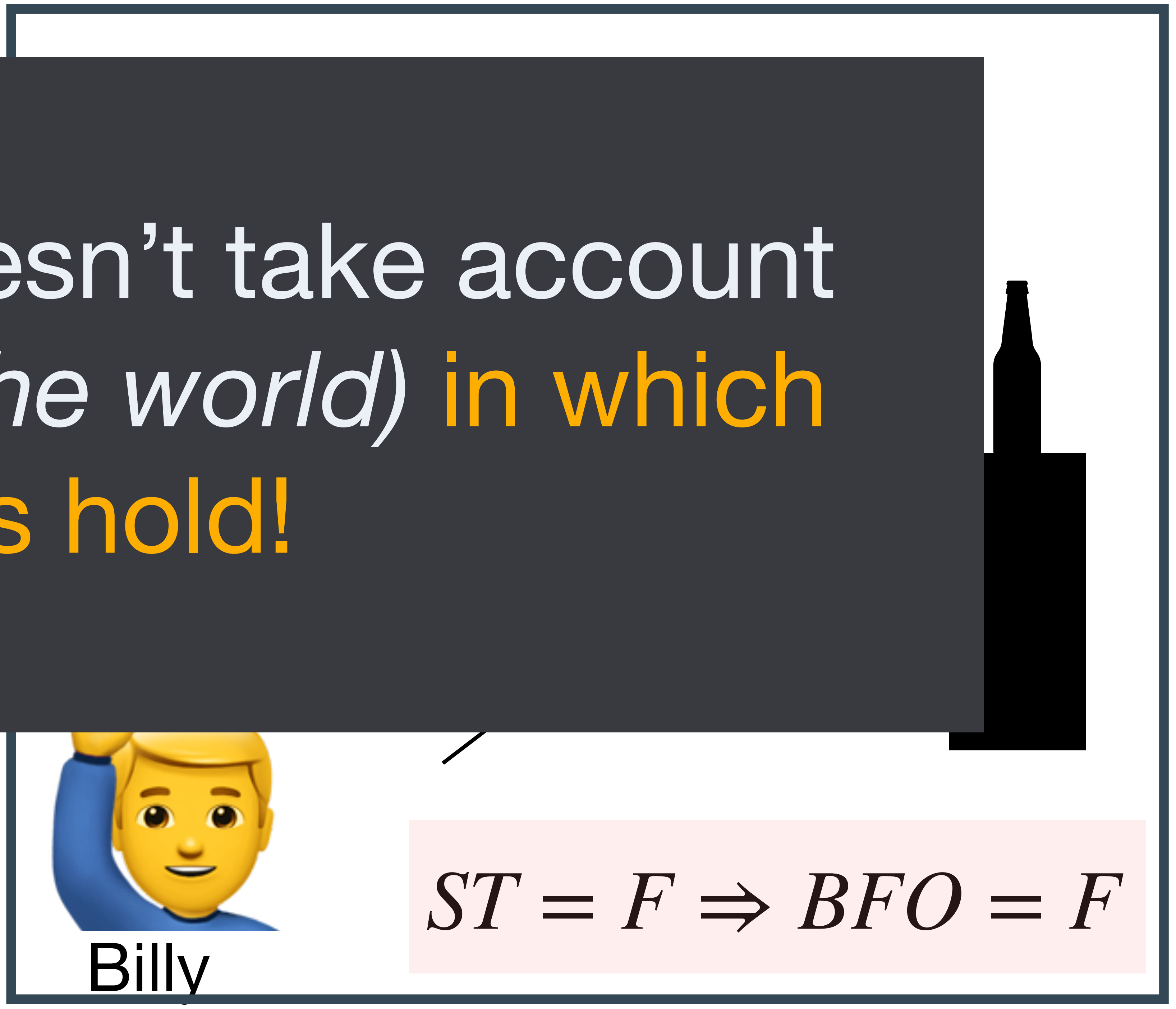




Toward a modern causality

The counterfactual only holds under the situation where Billy's ball didn't hit the bottle.





PO-based definition doesn't take account *situations (structure of the world)* in which the counterfactual logics hold!

Causal Model

Causal Model

A modern causality *taking account of situations* is developed by Pearl and his colleagues [Pearl, 2000].

Causal Model

A modern causality *taking account of situations* is developed by Pearl and his colleagues [Pearl, 2000].

Causal Model $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F} \rangle$

Causal Model

A modern causality *taking account of situations* is developed by Pearl and his colleagues [Pearl, 2000].

Causal Model $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F} \rangle$

- $\mathbf{V}$: A set of endogenous (observable) variables.

Causal Model

A modern causality *taking account of situations* is developed by Pearl and his colleagues [Pearl, 2000].

Causal Model $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F} \rangle$

- $\mathbf{V}$: A set of endogenous (observable) variables.
- $\mathbf{U}$: A set of exogenous (latent) variables.

Causal Model

A modern causality *taking account of situations* is developed by Pearl and his colleagues [Pearl, 2000].

Causal Model $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F} \rangle$

- $\mathbf{V}$: A set of endogenous (observable) variables.
- $\mathbf{U}$: A set of exogenous (latent) variables.
- $\mathbf{F}$: A set of structural equations $\{f_{V_i}\}_{V_i \in \mathbf{V}}$ determining the value of $V_i \in \mathbf{V}$, where $V_i \leftarrow f_{v_i}(PA_{V_i}, U_{V_i})$ for some $PA_{V_i} \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$.

Causal Model

A modern causality *taking account of situations* is developed by Pearl and his colleagues [Pearl, 2000].

Causal Model $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F} \rangle$

- $\mathbf{V}$: A set of endogenous (observable) variables.
- $\mathbf{U}$: A set of exogenous (latent) variables.
- $\mathbf{F}$: A set of structural equations $\{f_{V_i}\}_{V_i \in \mathbf{V}}$ determining the value of $V_i \in \mathbf{V}$, where $V_i \leftarrow f_{v_i}(PA_{V_i}, U_{V_i})$ for some $PA_{V_i} \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$.

💡 $\mathbf{U}$ describes '*context*' — By fixing $\mathbf{U} = \mathbf{u}$, values $\mathbf{V} = \mathbf{v}$ are completely determined.

Causal Model — Example - 1

ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off

Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off

Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow f_{BH}(BT) = BT \wedge (\neg SH)$$

ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off

Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow f_{BH}(BT) = BT \wedge (\neg SH)$$

$$BFO \leftarrow f_{BFO}(SH, BH) = SH \vee BH$$

ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off

Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow f_{BH}(BT) = BT \wedge (\neg SH)$$

$$BFO \leftarrow f_{BFO}(SH, BH) = SH \vee BH$$

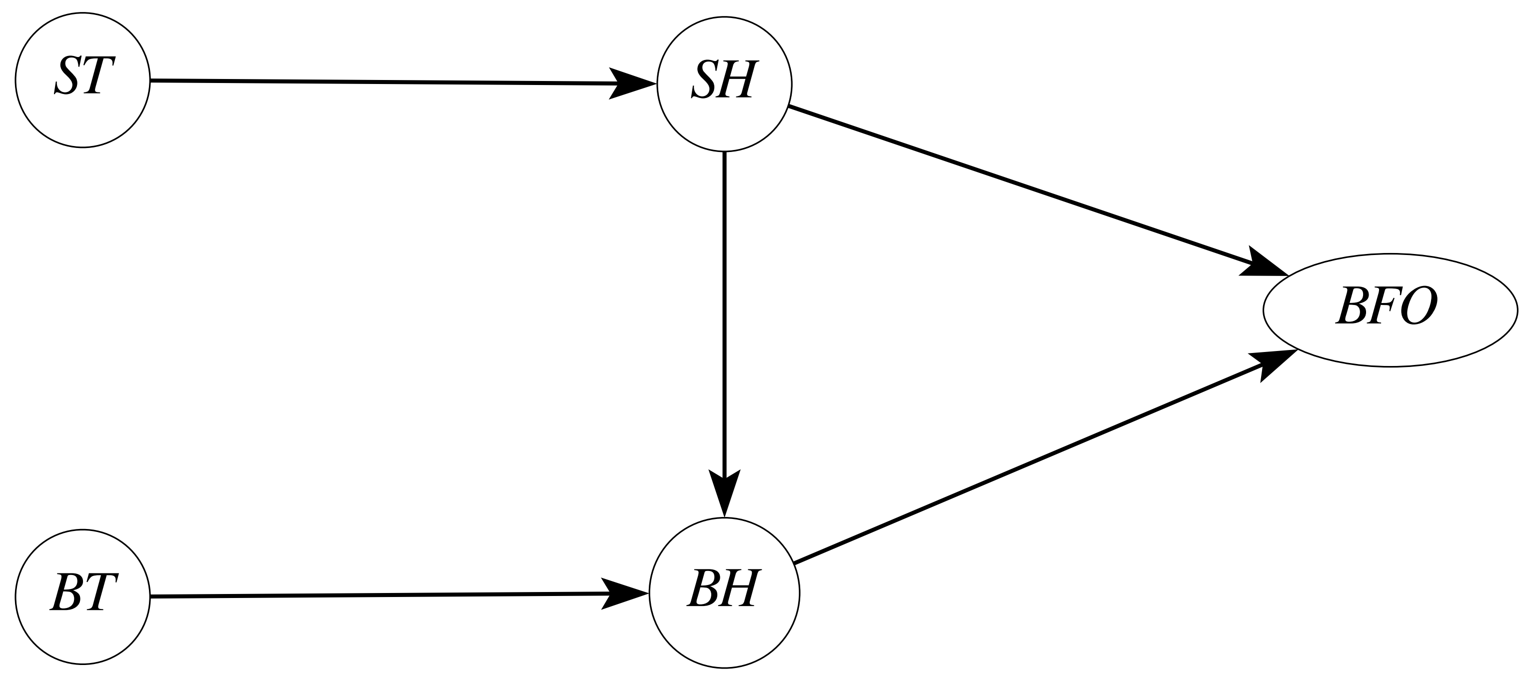
ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off



Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow f_{BH}(BT) = BT \wedge (\neg SH)$$

$$BFO \leftarrow f_{BFO}(SH, BH) = SH \vee BH$$

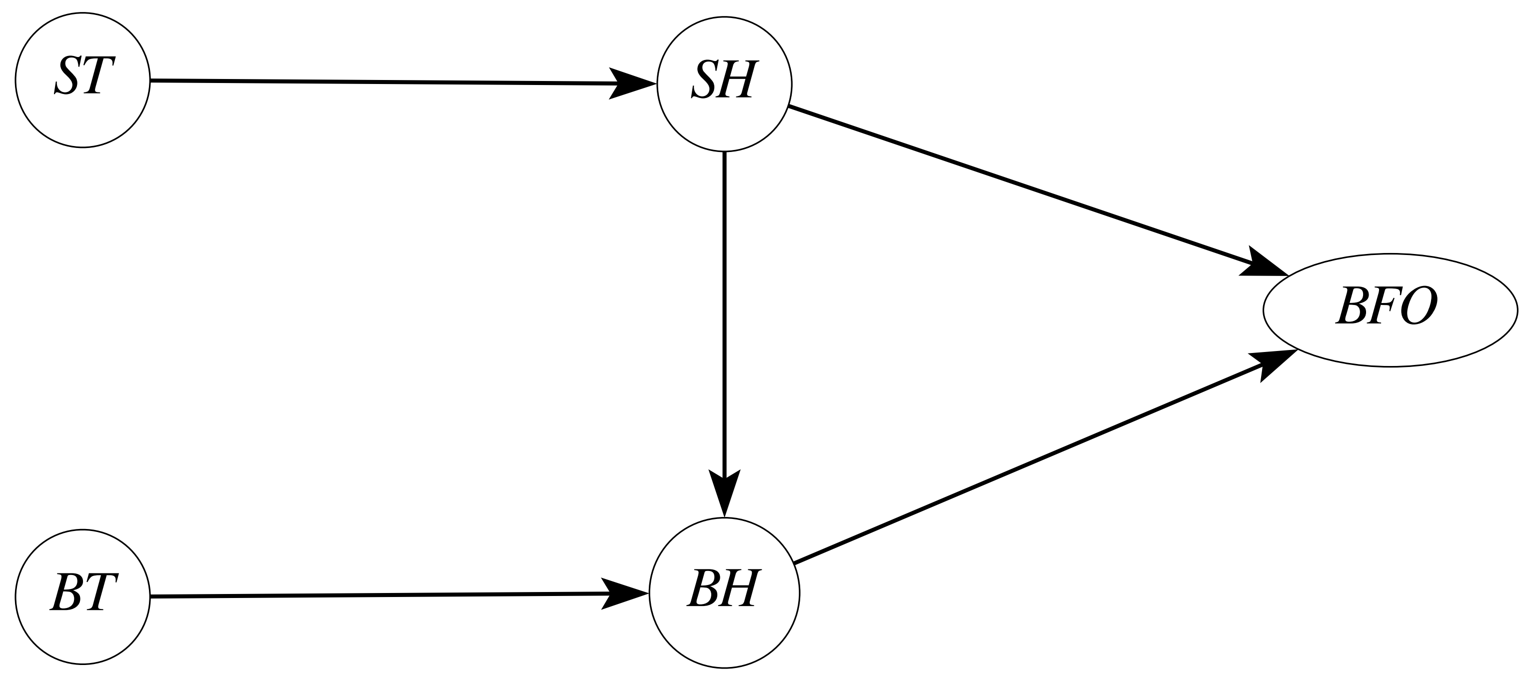
ST: Suzy Throws $\in \{T, F\}$

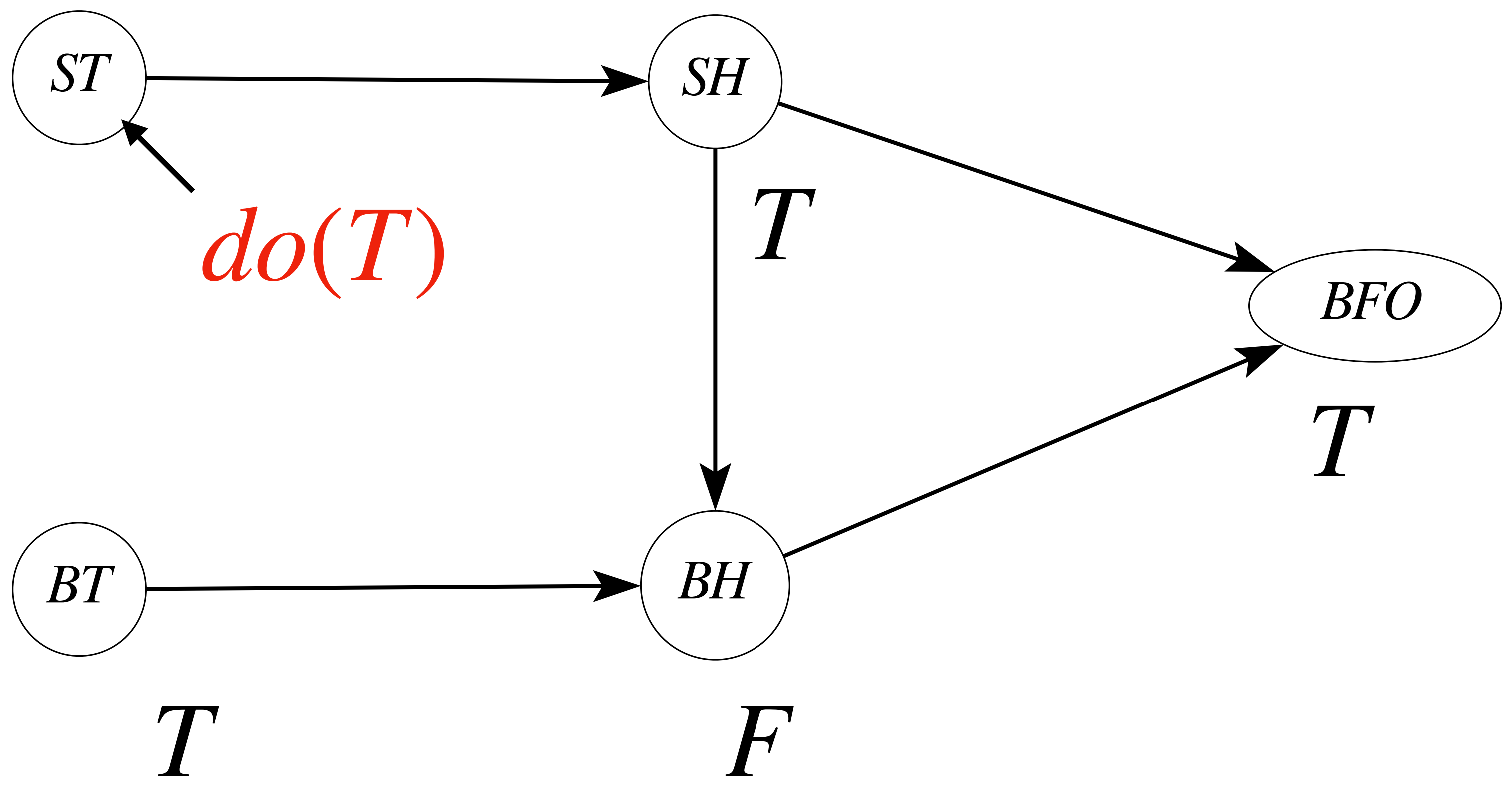
BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off





Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow f_{BH}(BT) = BT \wedge (\neg SH)$$

$$BFO \leftarrow f_{BFO}(SH, BH) = SH \vee BH$$

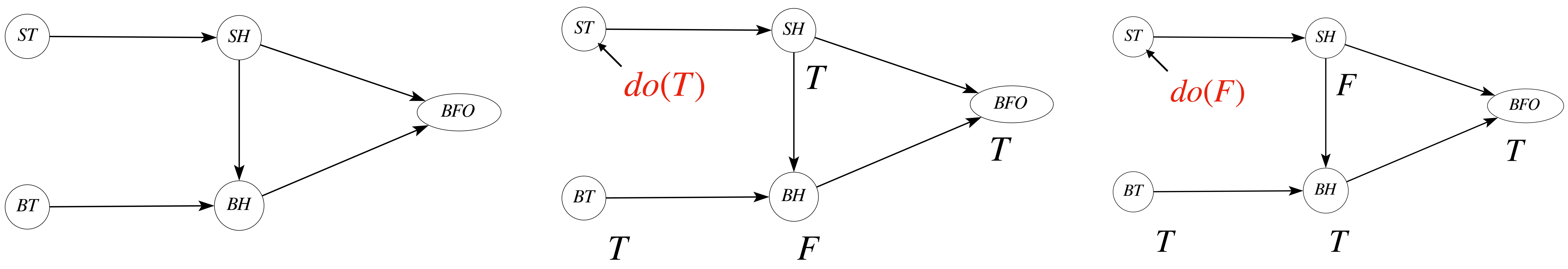
ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off



Causal Model — Example - 1

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow f_{BH}(BT) = BT \wedge (\neg SH)$$

$$BFO \leftarrow f_{BFO}(SH, BH) = SH \vee BH$$

ST: Suzy Throws $\in \{T, F\}$

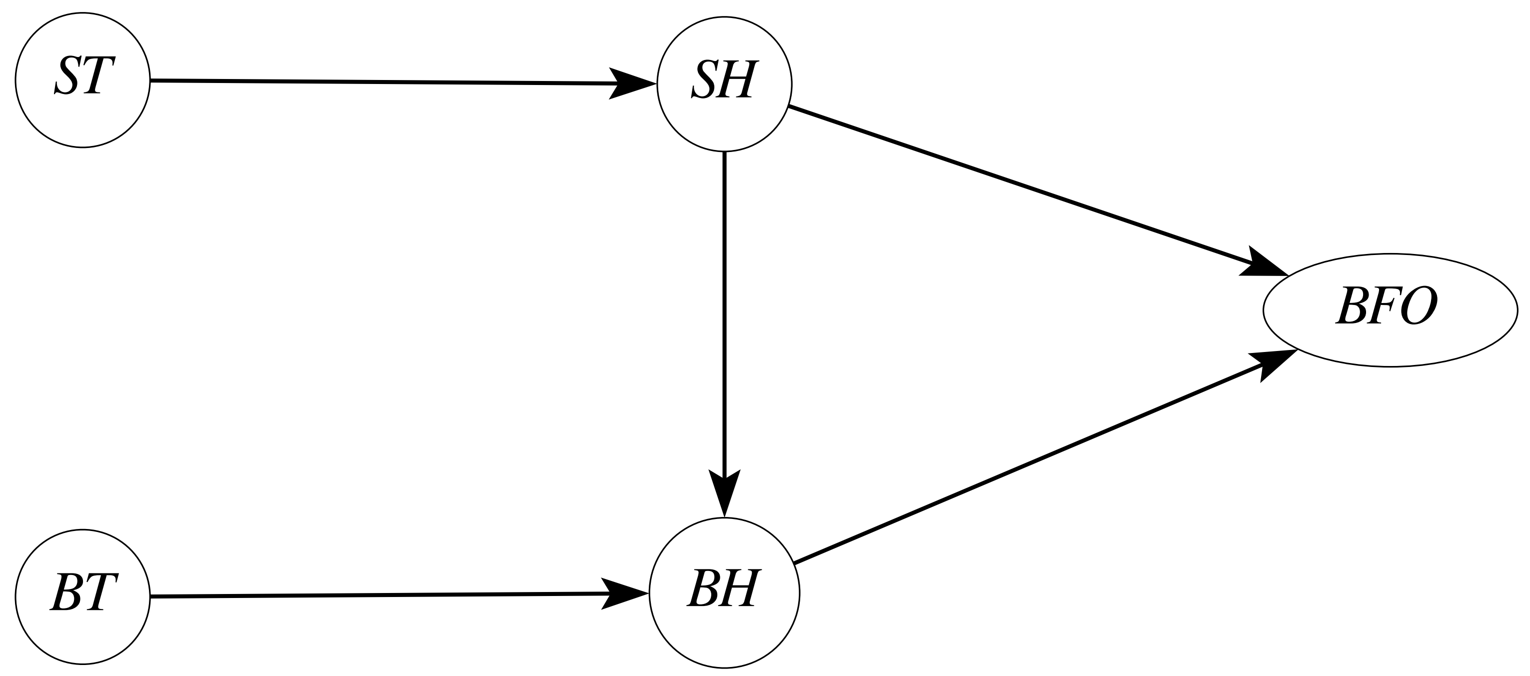
BT: Billy Throws $\in \{T, F\}$

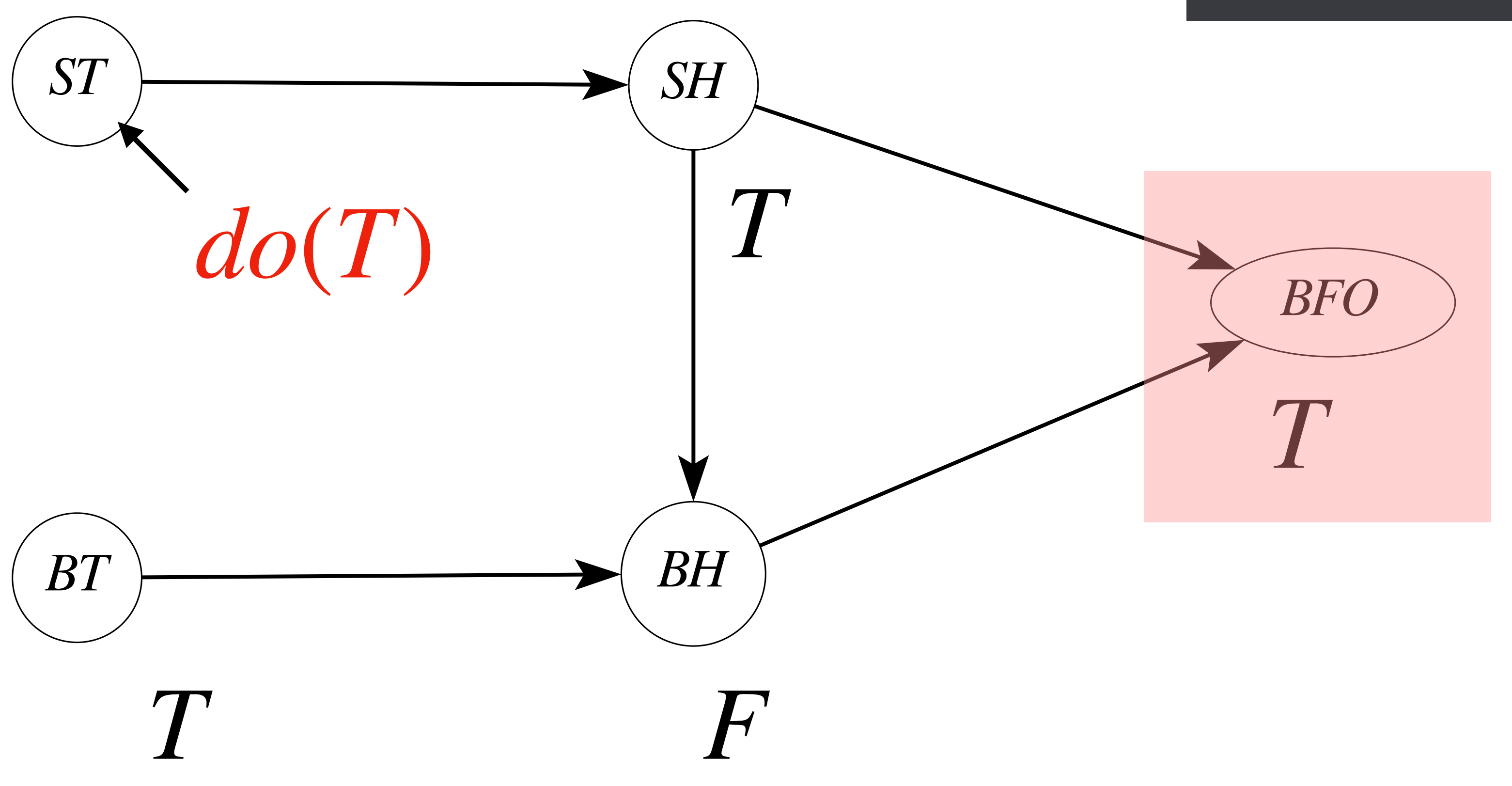
SH: Suzy's ball hit the bottle

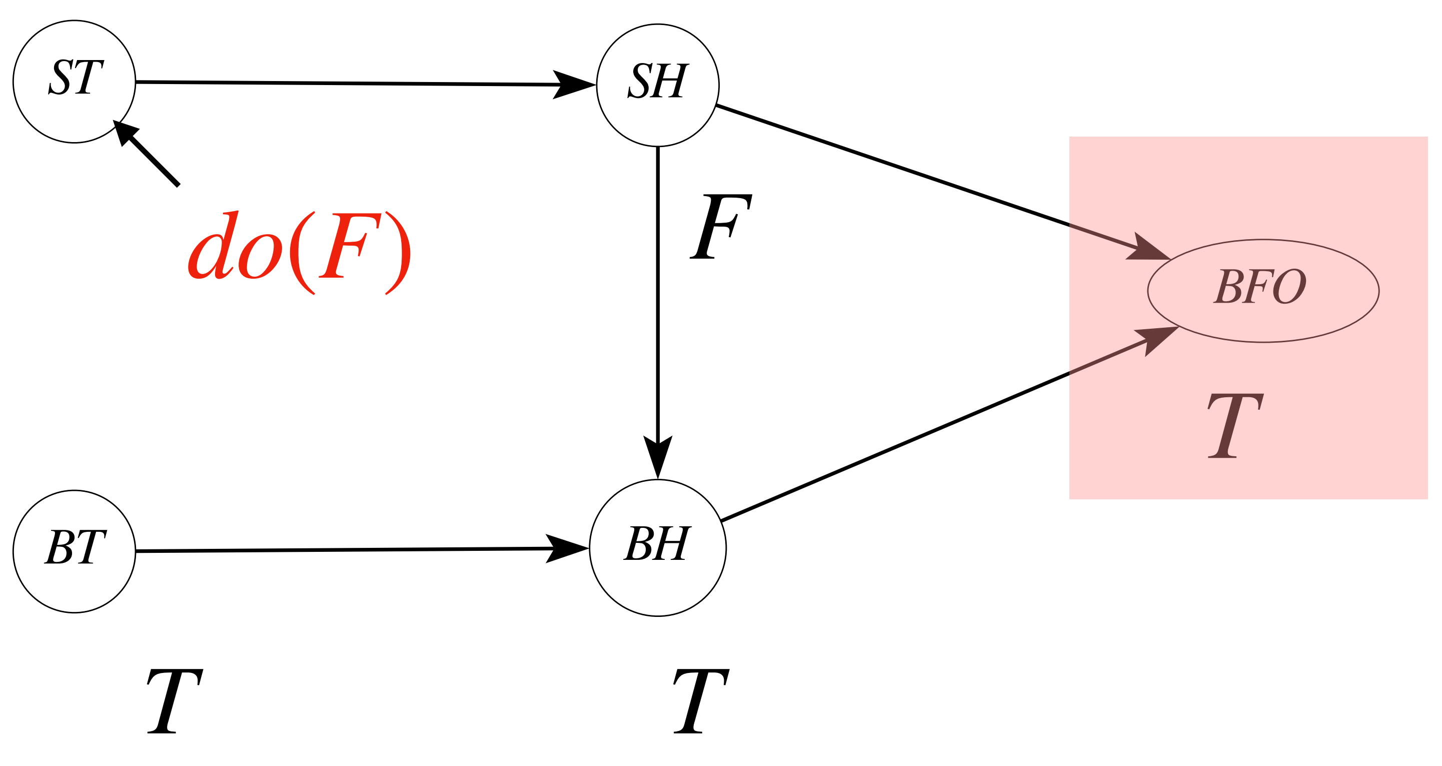
BH: Billy'

BFO: Bo

Suzy's throwing is not a cause, according to PO-based definition.







Causal Model — Example - 2

$SH \leftarrow f_{SH}(ST) = ST$

$BFO \leftarrow f_{BFO}(SH, F) = SH \vee BH$

ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off

Causal Model — Example - 2

$SH \leftarrow f_{SH}(ST) = ST$

$BH \leftarrow F$

$BFO \leftarrow f_{BFO}(SH, F) = SH \vee BH$

ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off

Causal Model — Example - 2

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow F$$

$$BFO \leftarrow f_{BFO}(SH, F) = SH \vee BH$$

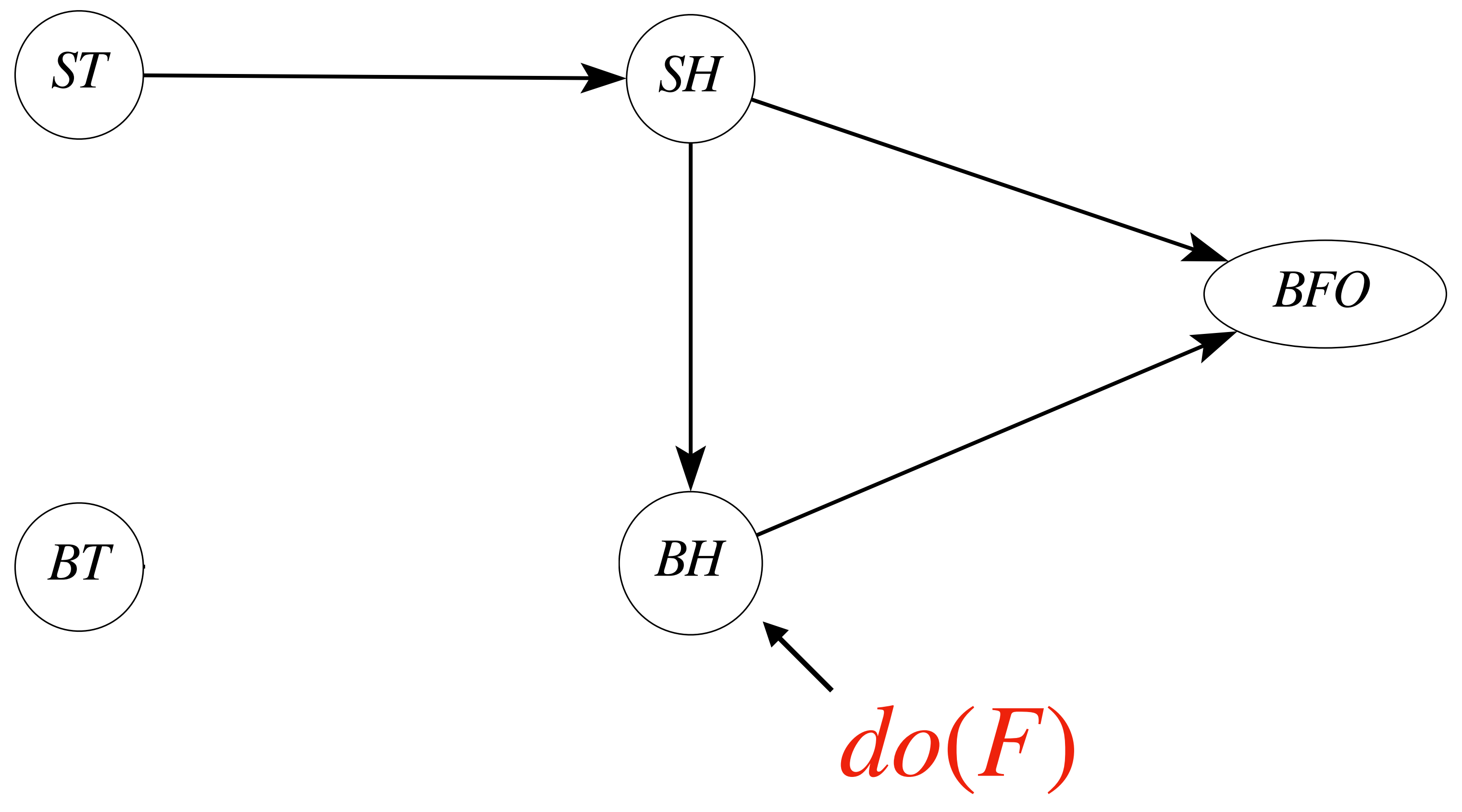
ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off



Causal Model — Example - 2

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow F$$

$$BFO \leftarrow f_{BFO}(SH, F) = SH \vee BH$$

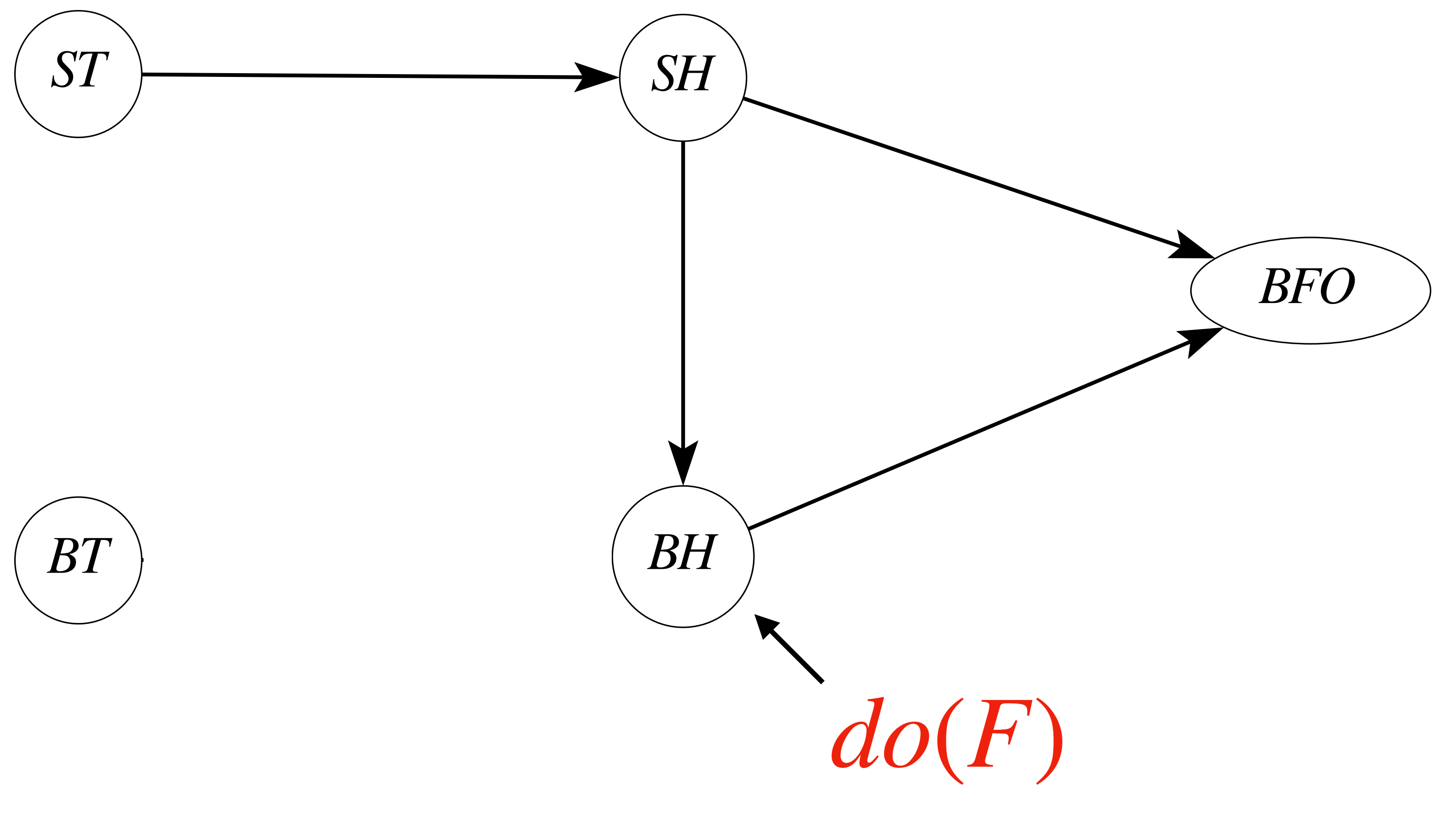
ST: Suzy Throws $\in \{T, F\}$

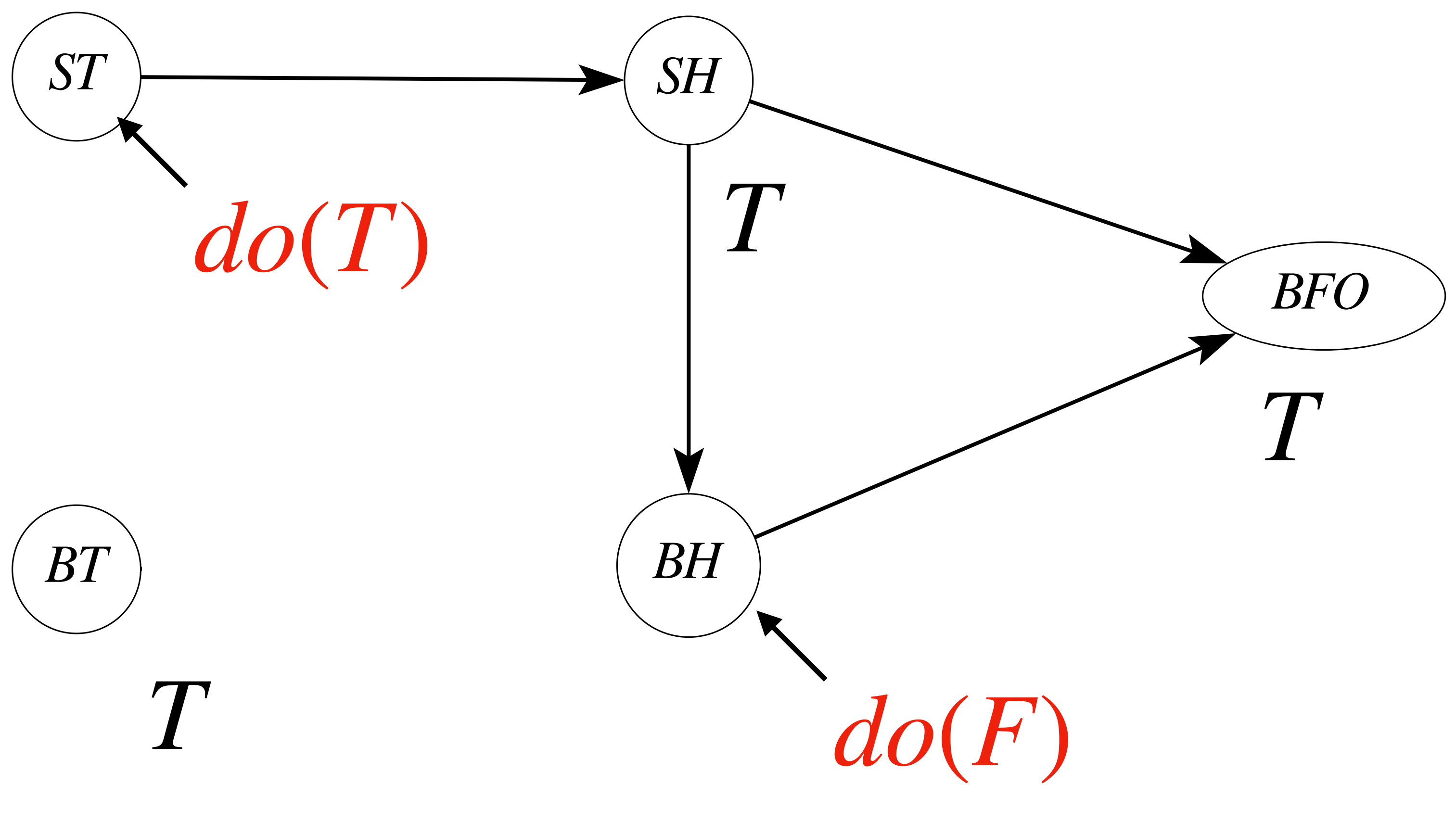
BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off





Causal Model — Example - 2

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow F$$

$$BFO \leftarrow f_{BFO}(SH, F) = SH \vee BH$$

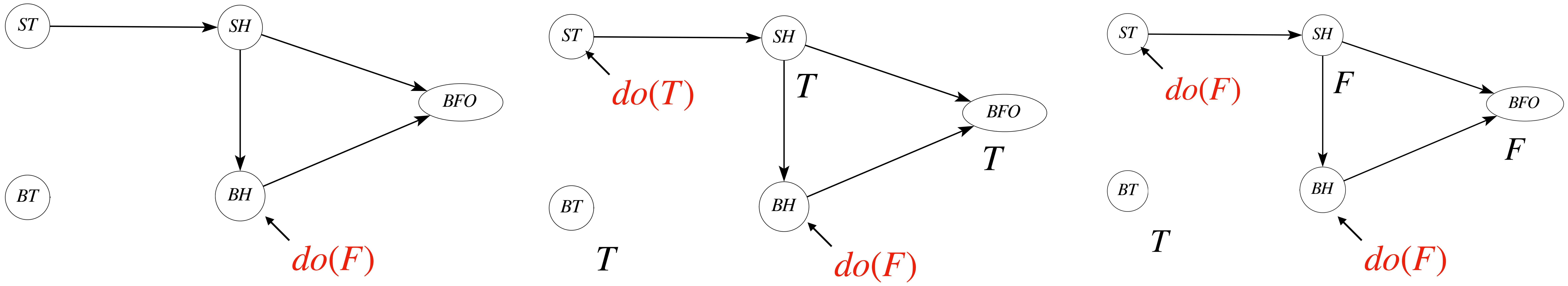
ST: Suzy Throws $\in \{T, F\}$

BT: Billy Throws $\in \{T, F\}$

SH: Suzy's ball hit the bottle

BH: Billy's ball hit the bottle

BFO: Bottle Fall Off



Causal Model — Example - 2

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow F$$

$$BFO \leftarrow f_{BFO}(SH, F) = SH \vee BH$$

ST: Suzy Throws $\in \{T, F\}$

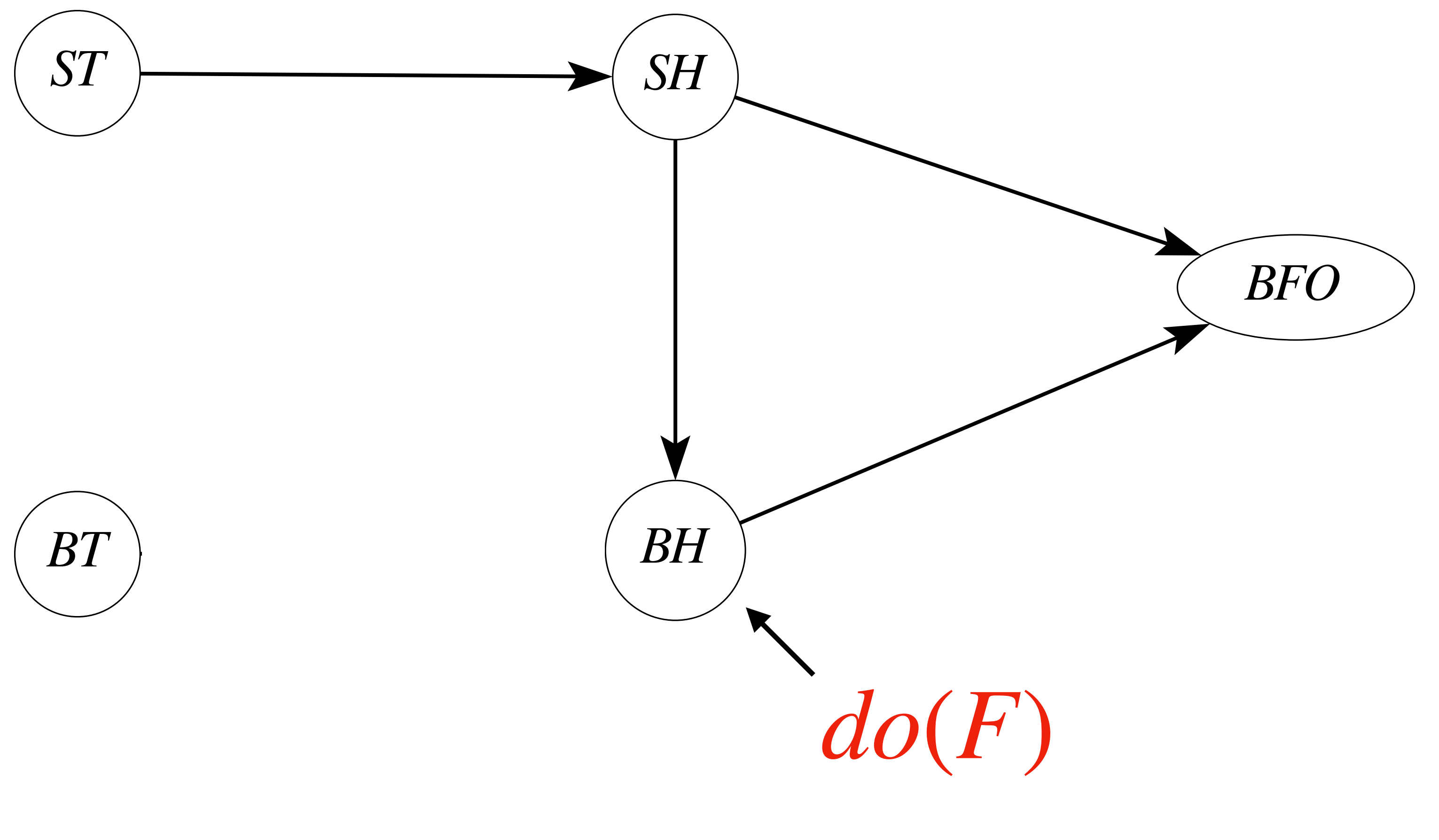
BT: Billy Throws $\in \{T, F\}$

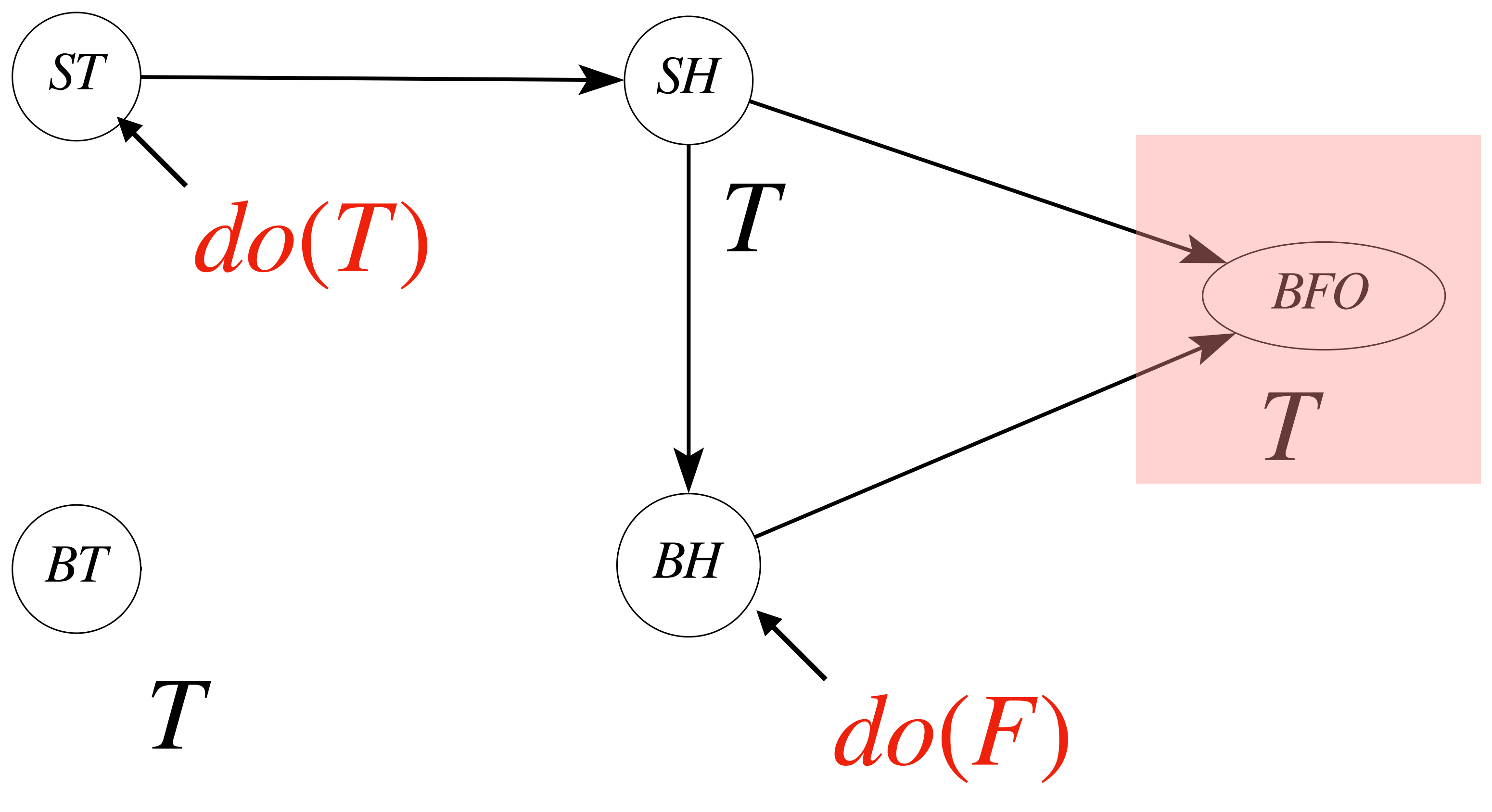
SH: Suzy's ball hit the bottle

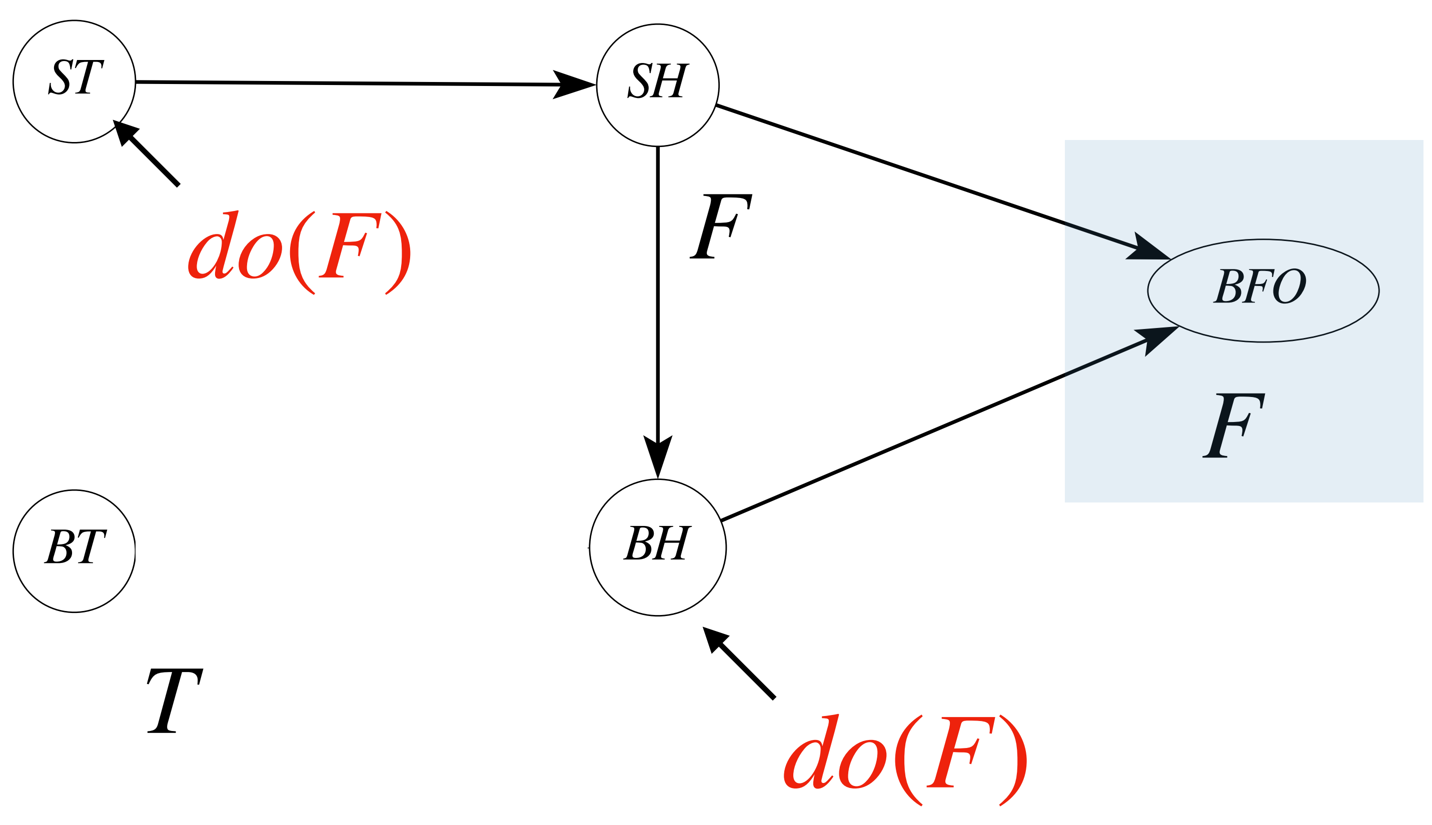
BH:

BFO

Under the situation where Billy's ball didn't hit the bottle, Suzy's throwing is a cause.





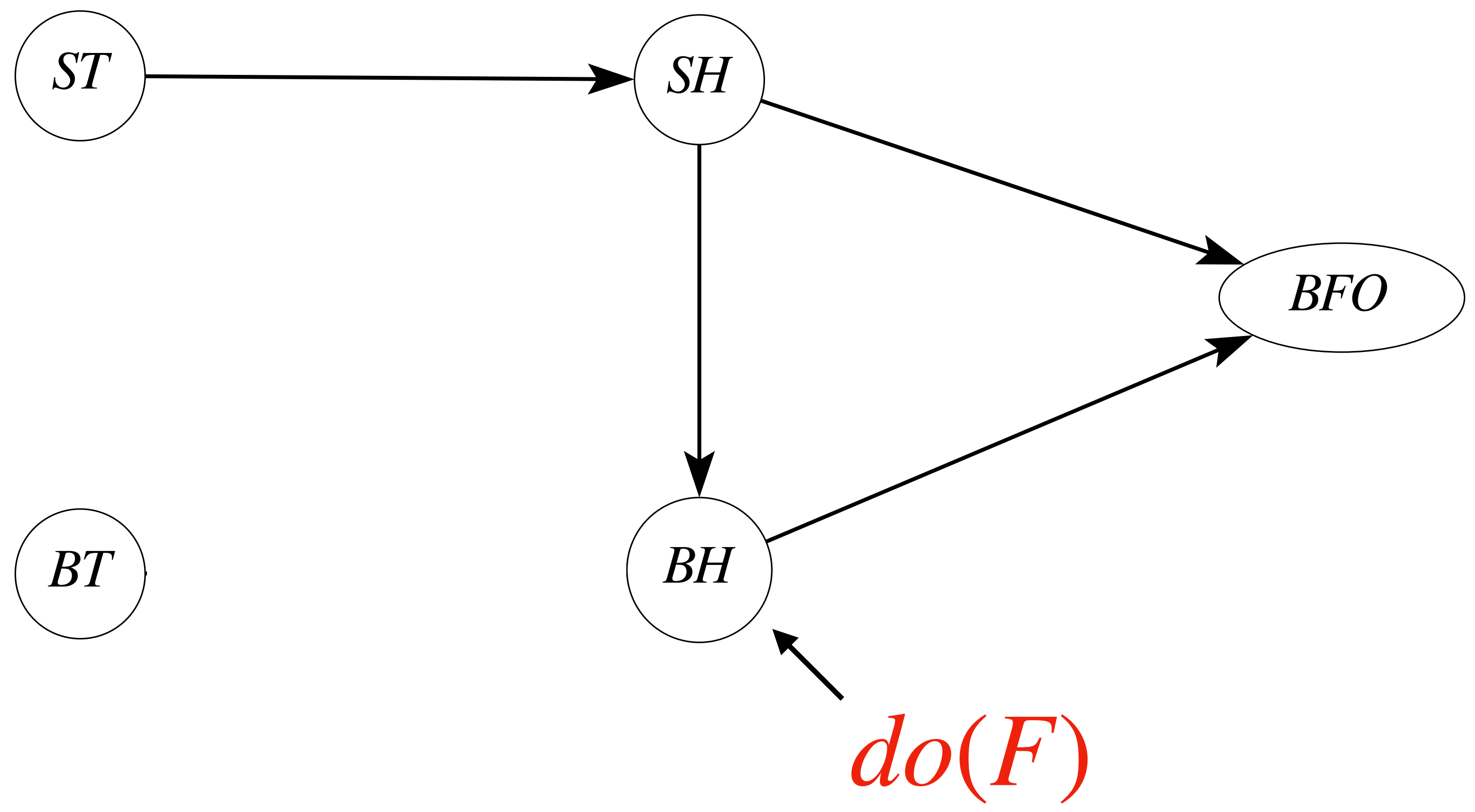


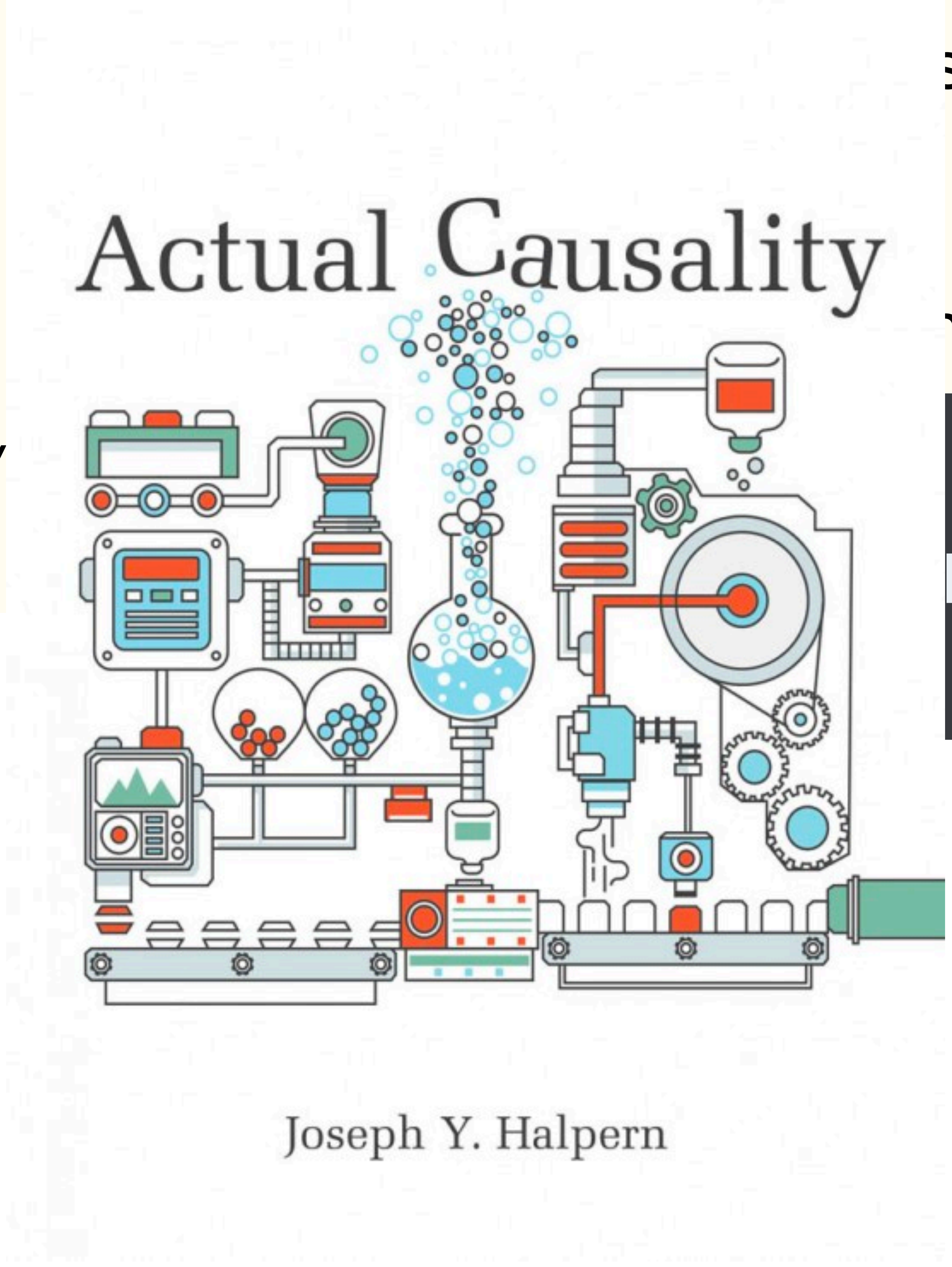
Causal Model — Example - 2

$$SH \leftarrow f_{SH}(ST) = ST$$

$$BH \leftarrow F$$

$$BFO \leftarrow f_{BFO}(SH, F) = SH \vee$$



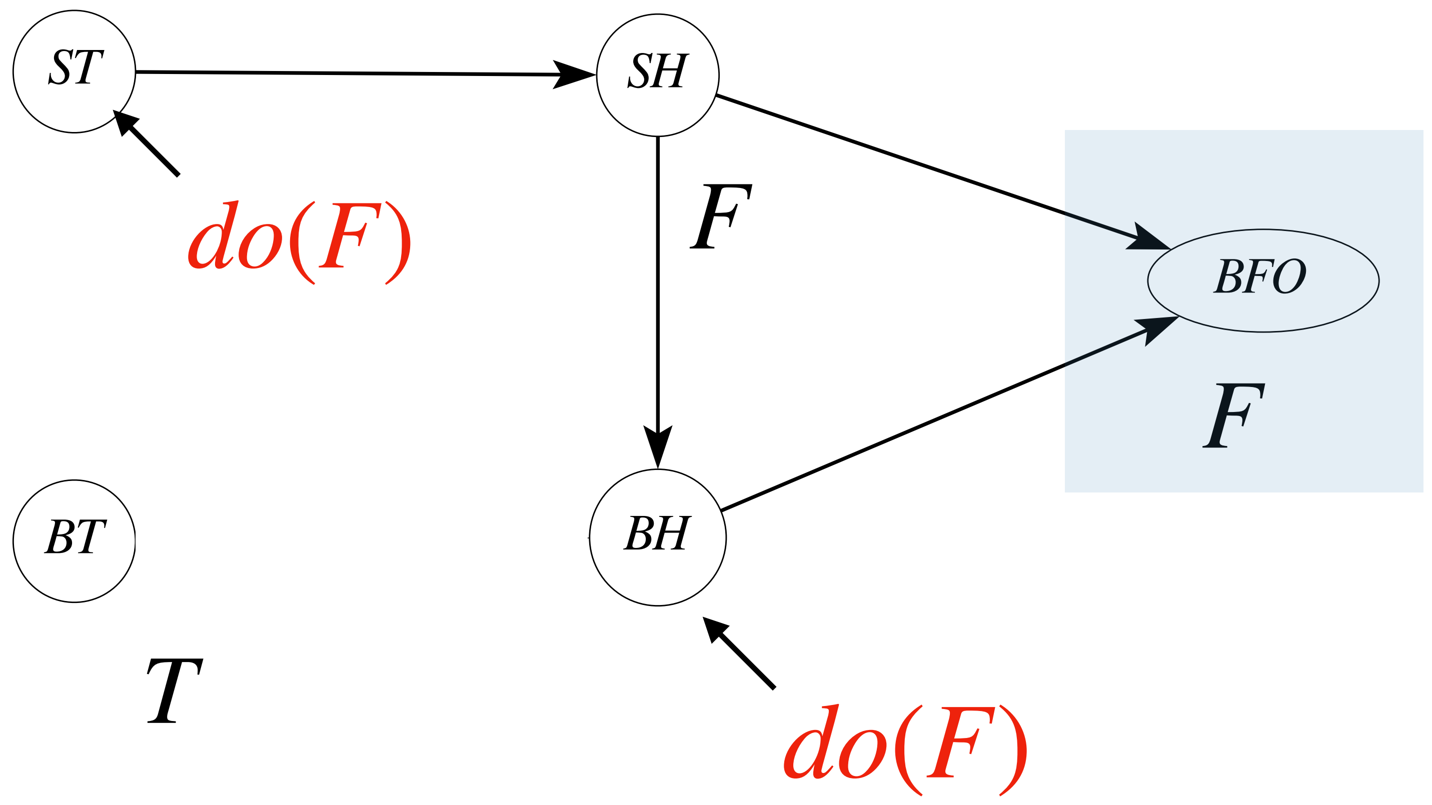


s $\in \{T, F\}$

$\in \{T, F\}$

it the bottle

situation where Billy's ball
he bottle, Suzy's throwing is



Structural Causal Model (SCM)

Structural Causal Model (SCM) permits *probabilistic uncertainties* in the context $\mathbf{U} = \mathbf{u}$.

Structural Causal Model $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{u}) \rangle$

- $\mathbf{V}$: A set of endogenous (observable) variables.
- $\mathbf{U}$: A set of exogenous (latent) variables.
- $\mathbf{F}$: A set of structural equations $\{f_{V_i}\}_{V_i \in \mathbf{V}}$ determining the value of $V_i \in \mathbf{V}$, where $V_i \leftarrow f_{v_i}(PA_{V_i}, U_{V_i})$ for some $PA_{V_i} \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$.
- $P(\mathbf{u})$: A probability measure for $\mathbf{U}$.

SCM as a unified language

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

Indeed, SCM subsumes the PO-based causality, because the potential outcome can be equivalently defined using the SCM.

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

Indeed, SCM subsumes the PO-based causality, because the potential outcome can be equivalently defined using the SCM.

SCM subsumes Potential outcome Y_x ("Y if X had been x")

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

Indeed, SCM subsumes the PO-based causality, because the potential outcome can be equivalently defined using the SCM.

SCM subsumes Potential outcome Y_x ("Y if X had been x")

- Given the SCM $\mathscr{M}$, let $\mathscr{M}_{do(x)}$ denote the SCM inducing by fixing $X = x$ in $\mathscr{M}$.

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

Indeed, SCM subsumes the PO-based causality, because the potential outcome can be equivalently defined using the SCM.

SCM subsumes Potential outcome Y_x ("Y if X had been x")

- Given the SCM $\mathscr{M}$, let $\mathscr{M}_{do(x)}$ denote the SCM inducing by fixing $X = x$ in $\mathscr{M}$.
- Let Y be induced from the SCM $\mathscr{M}$.

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

Indeed, SCM subsumes the PO-based causality, because the potential outcome can be equivalently defined using the SCM.

SCM subsumes Potential outcome Y_x ("Y if X had been x")

- Given the SCM $\mathscr{M}$, let $\mathscr{M}_{do(x)}$ denote the SCM inducing by fixing $X = x$ in $\mathscr{M}$.
- Let Y be induced from the SCM $\mathscr{M}$.
- Then Y_x is induced from $\mathscr{M}_{do(x)}$.

SCM as a unified language

So far, we see that SCM can fill the lacuna missed by PO-based causality.

Indeed, SCM subsumes the PO-based causality, because the potential outcome can be equivalently defined using the SCM.

SCM subsumes Potential outcome Y_x ("Y if X had been x")

- Given the SCM $\mathscr{M}$, let $\mathscr{M}_{do(x)}$ denote the SCM inducing by fixing $X = x$ in $\mathscr{M}$.
- Let Y be induced from the SCM $\mathscr{M}$.
- Then Y_x is induced from $\mathscr{M}_{do(x)}$. (Roughly, $Y_x = Y \mid do(x)$)

SCM as an axiomatic characterization

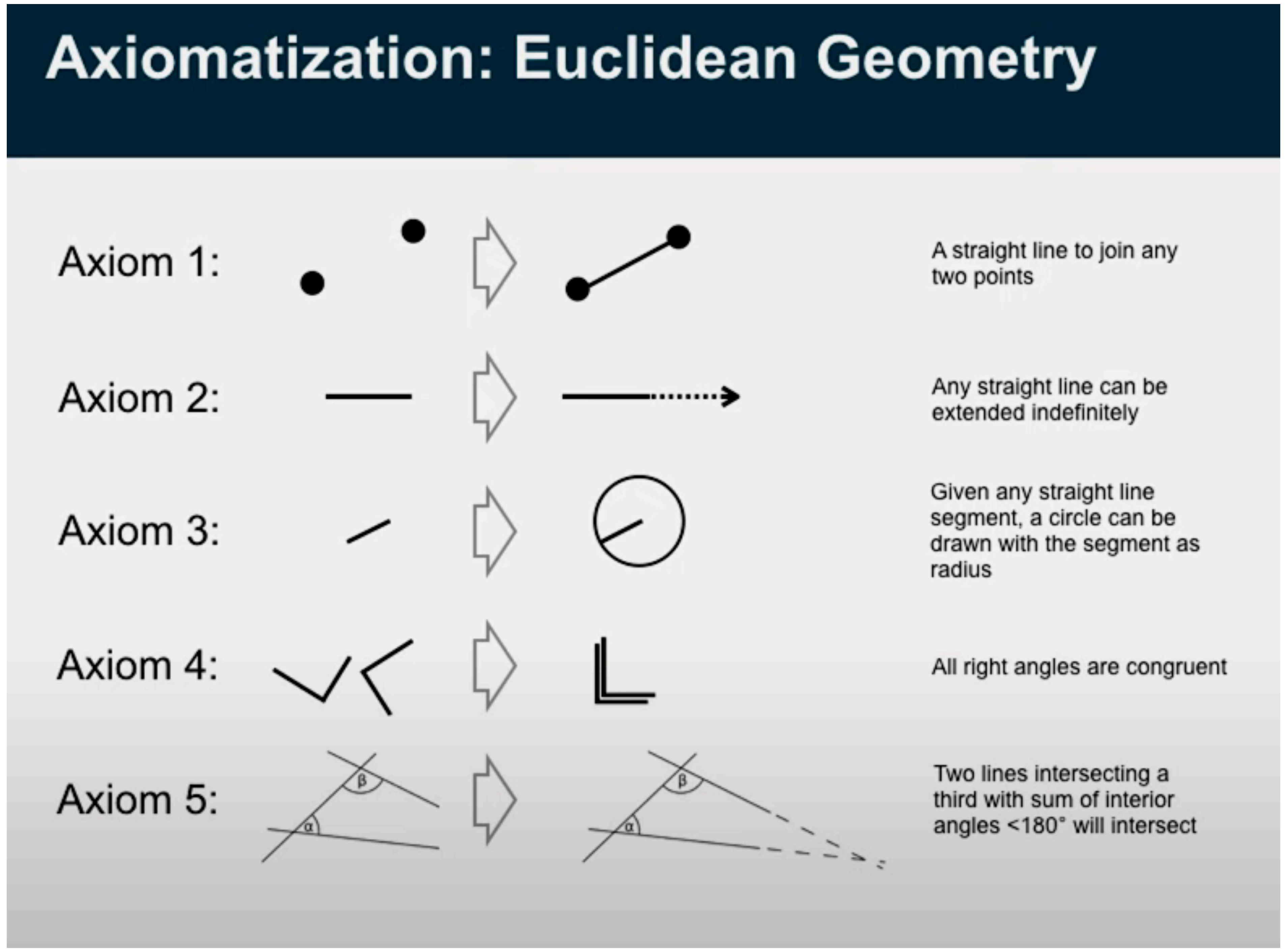
SCM as an axiomatic characterization

Axiomatic logic — Logic systems (or theories) starting from very simple (even trivial) properties.

SCM as an axiomatic characterization

Axiomatic logic — Logic systems (or theories) starting from very simple (even trivial) properties.

Why axiomatization? Consider Euclidean Geometry. Any theories (or logic system) of geometry agreeing with these axioms are equivalent to Euclidian Geometry!



(Frederick Eberhardt)

SCM as an axiomatic characterization

SCM as an axiomatic characterization

Axioms for counterfactual [Galles, Pearl & Halpern]

SCM as an axiomatic characterization

Axioms for counterfactual [Galles, Pearl & Halpern]

For a given context $\mathrm{U} = \mathbf{u}$, suppose we take *acyclicity* (no cycle loop) as truth.

$$(\text{cause} \rightarrow \text{effect}) \ \& \ (\text{effect} \nrightarrow \text{cause})$$

SCM as an axiomatic characterization

Axioms for counterfactual [Galles, Pearl & Halpern]

For a given context $\mathbf{U} = \mathbf{u}$, suppose we take *acyclicity* (no cycle loop) as truth.

$$\text{(cause} \rightarrow \text{effect)} \;\&\; \text{(effect} \nrightarrow \text{cause)}$$

Composition: In the hypothetical population where X is fixed to x for all units, any W equals to W_x.

$$Y_x(\mathbf{u}) = Y_{x, W_x(\mathbf{u})}(\mathbf{u})$$

If we had treated all patients a drug ($X = 1$), then patients' blood pressure (BP, W) would be W_x.

SCM as an axiomatic characterization

Axioms for counterfactual [Galles, Pearl & Halpern]

For a given context $\mathbf{U} = \mathbf{u}$, suppose we take *acyclicity* (no cycle loop) as truth.

(cause $\rightarrow$ effect) & (effect $\nrightarrow$ cause)

Composition: In the hypothetical population where X is fixed to x for all units, any W equals to W_x.

$$Y_x(\mathbf{u}) = Y_{x, W_x(\mathbf{u})}(\mathbf{u})$$

If we had treated all patients a drug ($X = 1$), then patients' blood pressure (BP, W) would be W_x.

Effectiveness: In the hypothetical population where X is fixed to x, for any context, $X = x$.

$$X_x(\mathbf{u}) = x$$

SCM as an axiomatic characterization

SCM is a **sound and complete** framework satisfying these axioms!

$$X_x(\mathbf{u}) = x$$

SCM as an axiomatic characterization

SCM is a **sound and complete** framework satisfying these axioms!

👏 SCM can subsume any causal theories agreeing with these axioms.

$$X_x(\mathbf{u}) = x$$

SCM as an axiomatic characterization

SCM is a **sound and complete** framework satisfying these axioms!

👏 SCM can subsume any causal theories agreeing with these axioms.

💁‍♀️ This is why Pearl's causality is acknowledged as a 'revolution' or 'new science' on causality.

$$X_x(\mathbf{u}) = x$$

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question
L1 (Association)	See	$P(y \mid x)$	What does the symptom tells about my headache?

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question
L1 (Association)	Classification Regression Observational study	$P(y \mid x)$	What does the symptom tells about my headache?

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question
L1 (Association)	Classification Regression Observational study	$P(y \mid x)$	What does the symptom tells about my headache?
L2 (Intervention)	 Do	$P(y \mid do(x))$	What if I took the aspirin, will my headache be cured?

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question
L1 (Association)	Classification Regression Observational study	$P(y \mid x)$	What does the symptom tells about my headache?
L2 (Intervention)	Reinforcement Learning Randomized trial	$P(y \mid do(x))$	What if I took the aspirin, will my headache be cured?

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question
L1 (Association)	Classification Regression Observational study	$P(y \mid x)$	What does the symptom tells about my headache?
L2 (Intervention)	Reinforcement Learning Randomized trial	$P(y \mid do(x))$	What if I took the aspirin, will my headache be cured?
L3 (Counterfactual)	 Retrospect	$P(y_x \mid x', y')$	Given that I didn't take the aspirin and didn't get cured, what if I did?

Three hierarchy in human cognition

Human cognition	Task	Quantity	Question
L1 (Association)	Classification Regression Observational study	$P(y \mid x)$	What does the symptom tells about my headache?
L2 (Intervention)	Reinforcement Learning Randomized trial	$P(y \mid do(x))$	What if I took the aspirin, will my headache be cured?
L3 (Counterfactual)	Structural Causal Model	$P(y_x \mid x', y')$	Given that I didn't take the aspirin and didn't get cured, what if I did?

Three hierarchy in human cognition

SCM as a suitable language to teach human cognition to AI

Human cognition	Task	Quantity	Question
L3 (Counterfactual)	Structural Causal Model	$P(y_x \mid x', y')$	Given that I didn't take the aspirin and didn't get cured, what if I did?

Pearl’s Causal Hierarchy (PCH)

Pearl's Causal Hierarchy (PCH)

Pearl's Causal Hierarchy [Bareinboim et al., 2020]

Pearl's Causal Hierarchy (PCH)

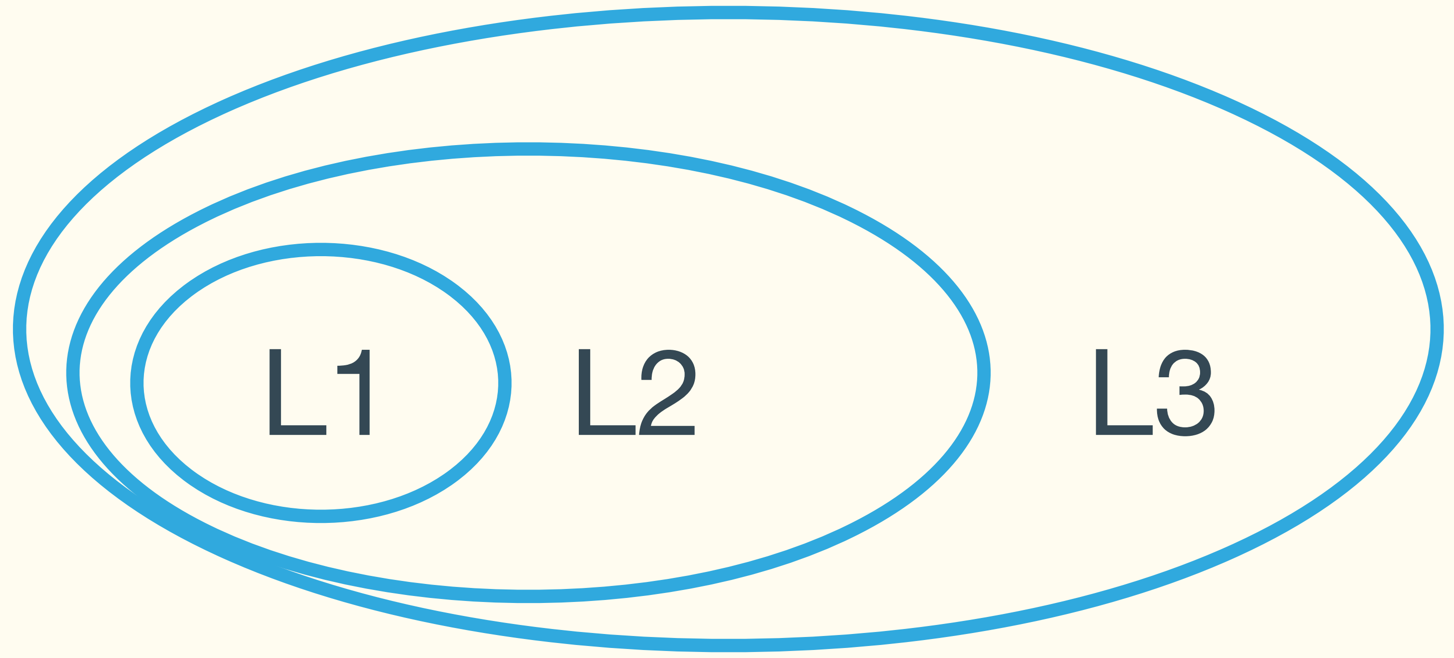
Pearl's Causal Hierarchy [Bareinboim et al., 2020]

- Structural Causal Model can represent all three layers (e.g., $\mathscr{M}$ is for L1, $\mathscr{M}_{do(x)}$ is for L2)

Pearl's Causal Hierarchy (PCH)

Pearl's Causal Hierarchy [Bareinboim et al., 2020]

- Structural Causal Model can represent all three layers (e.g., $\mathscr{M}$ is for L1, $\mathscr{M}_{do(x)}$ is for L2)

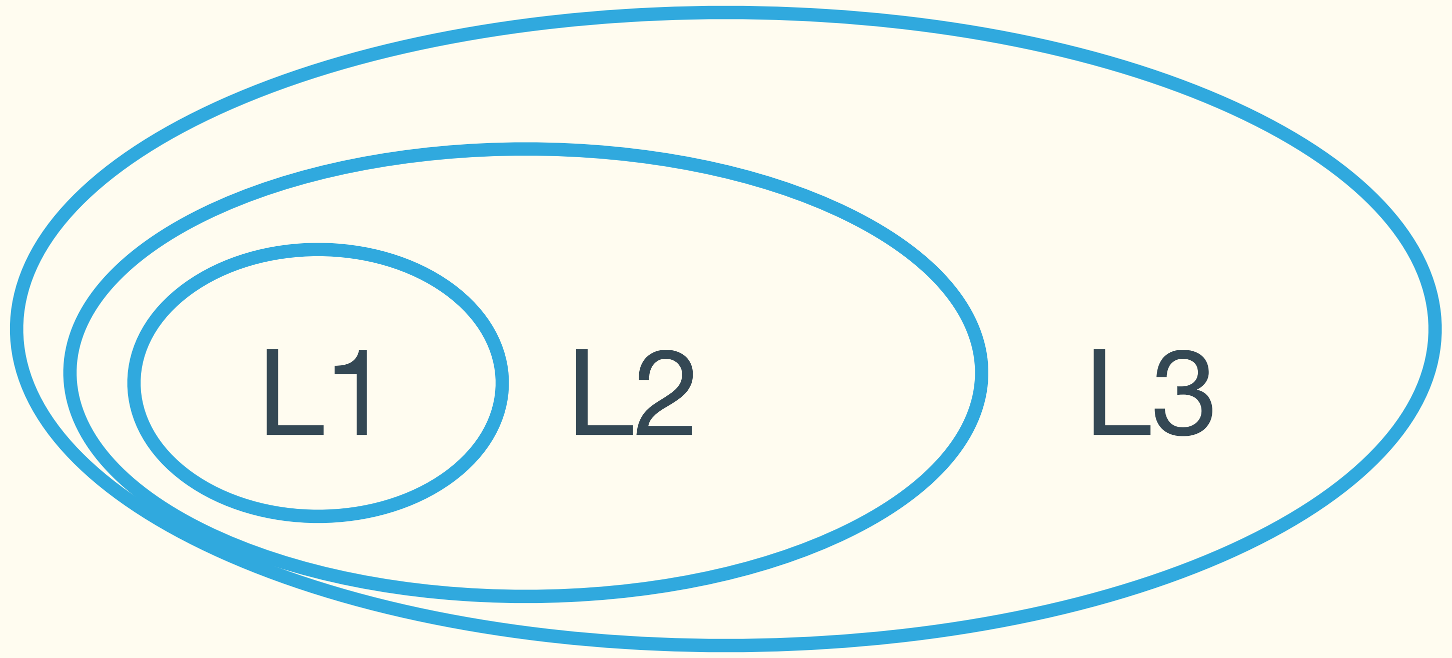


- With knowledge from lower layers, we cannot say anything about the higher layers.

Pearl’s Causal Hierarchy (PCH)

Pearl’s Causal Hierarchy [Bareinboim et al., 2020]

- Structural Causal Model can represent all three layers (e.g., $\mathscr{M}$ is for L1, $\mathscr{M}_{do(x)}$ is for L2)

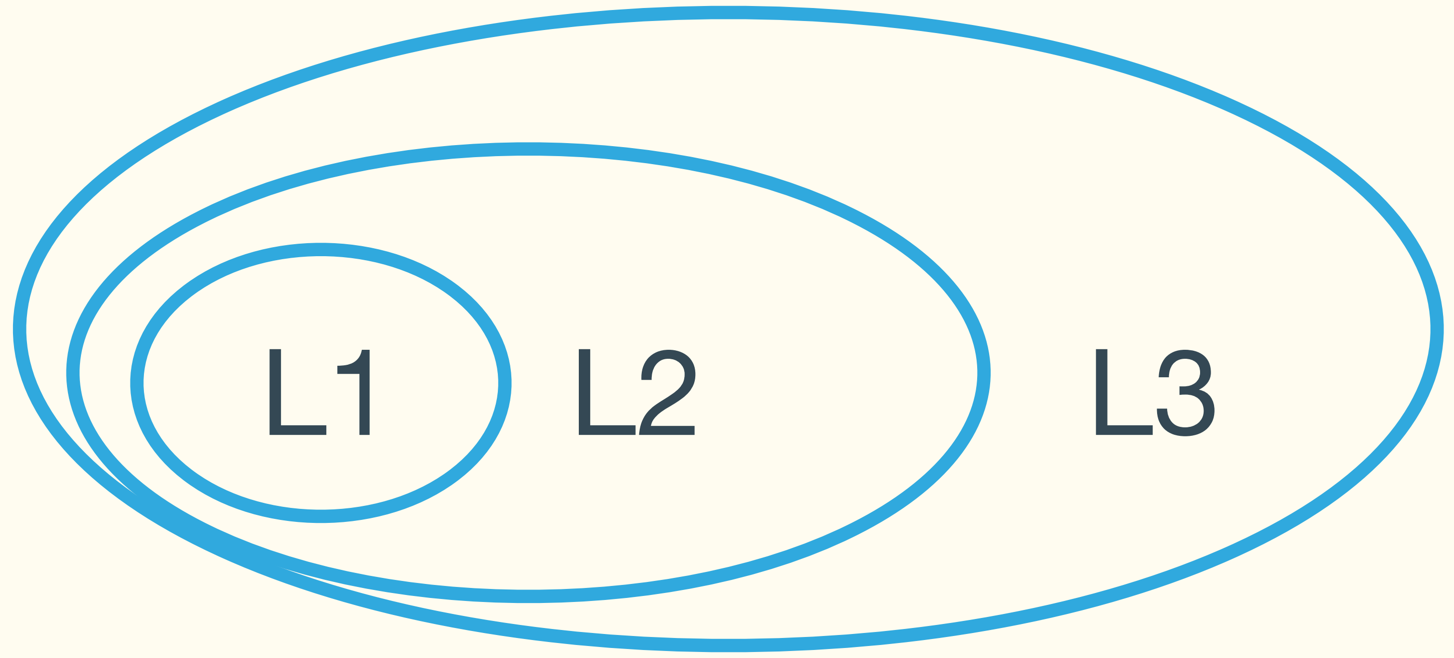


- With knowledge from lower layers, we cannot say anything about the higher layers.
  - Solely with the observational data (L1), we cannot answer ‘what-if’ question in L2.

Pearl's Causal Hierarchy (PCH)

Pearl's Causal Hierarchy [Bareinboim et al., 2020]

- Structural Causal Model can represent all three layers (e.g., $\mathscr{M}$ is for L1, $\mathscr{M}_{do(x)}$ is for L2)



- With knowledge from lower layers, we cannot say anything about the higher layers.
  - Solely with the observational data (L1), we cannot answer 'what-if' question in L2.
  - Solely with the interventional data (L2), we cannot answer retrospective/ counterfactual question in L3.

Causal inference through SCM

Causal inference through SCM

SCM

$$X \leftarrow f_X(U_X)$$

$$Y \leftarrow f_Y(X, U_Y)$$

$$P(U_X, U_Y)$$

Causal inference through SCM

SCM

$X \leftarrow f_X(U_X)$

$Y \leftarrow f_Y(X, U_Y)$

$P(U_X, U_Y)$

Graph

$X \rightarrow Y$

Qualitative description of SCM

Data

$\mathscr{D} = \{X_i, Y_i\}_{i=1}^{N} \sim P(x, y)$

Causal inference through SCM

SCM

$X \leftarrow f_X(U_X)$

$Y \leftarrow f_Y(X, U_Y)$

$P(U_X, U_Y)$

Graph

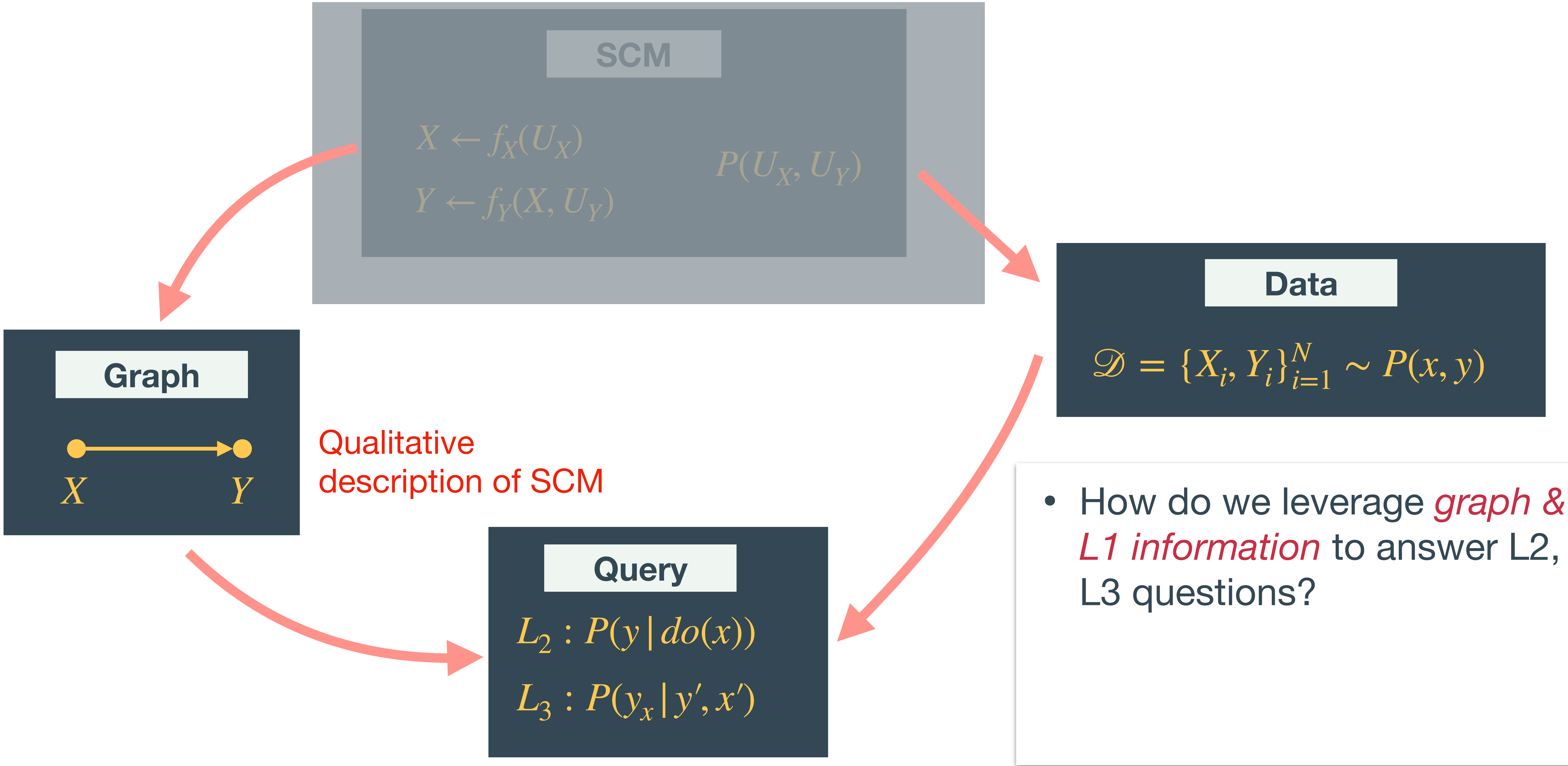
$X \rightarrow Y$

Qualitative description of SCM

Data

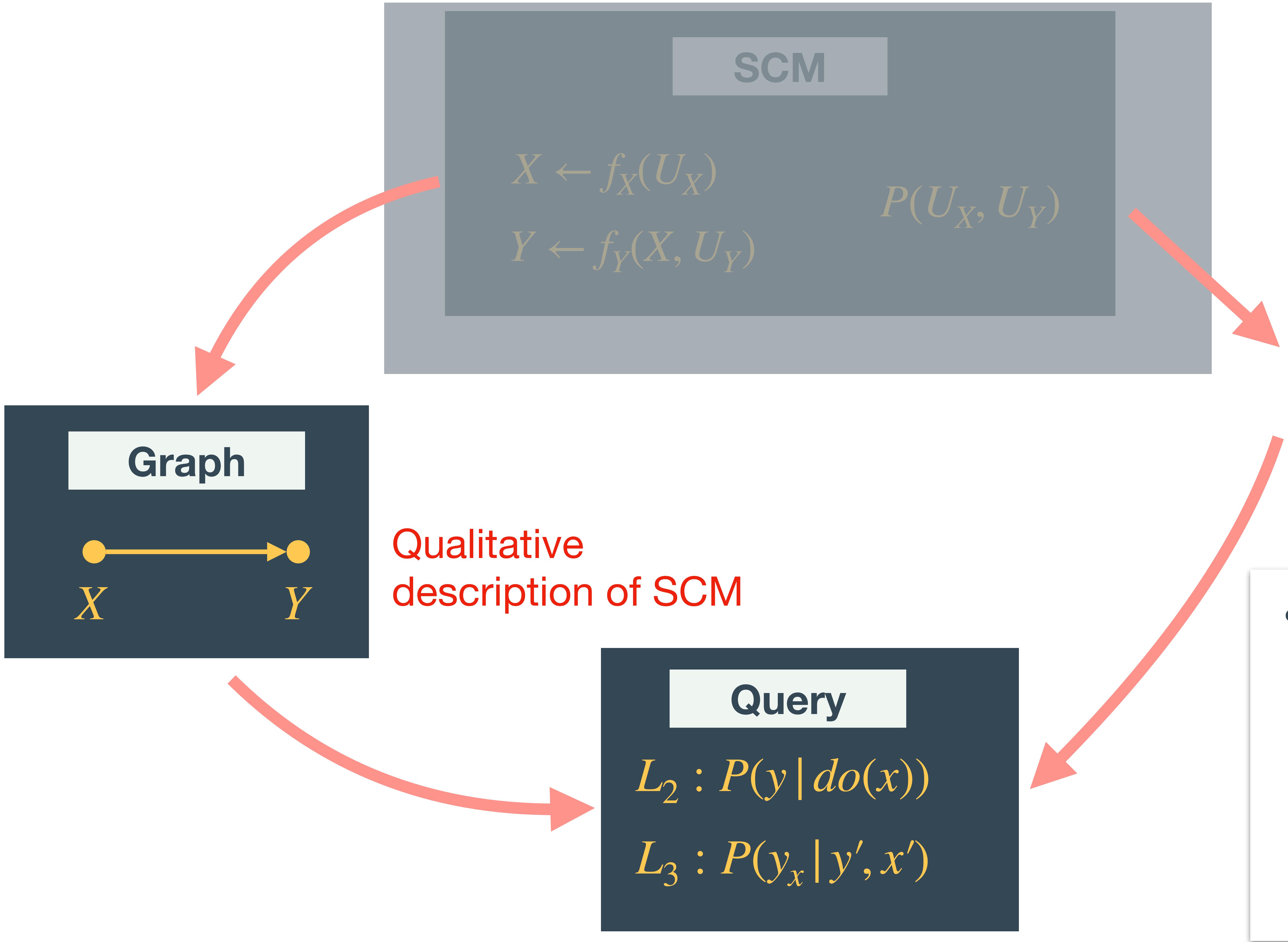
$\mathscr{D} = \{X_i, Y_i\}_{i=1}^{N} \sim P(x, y)$

Causal inference through SCM



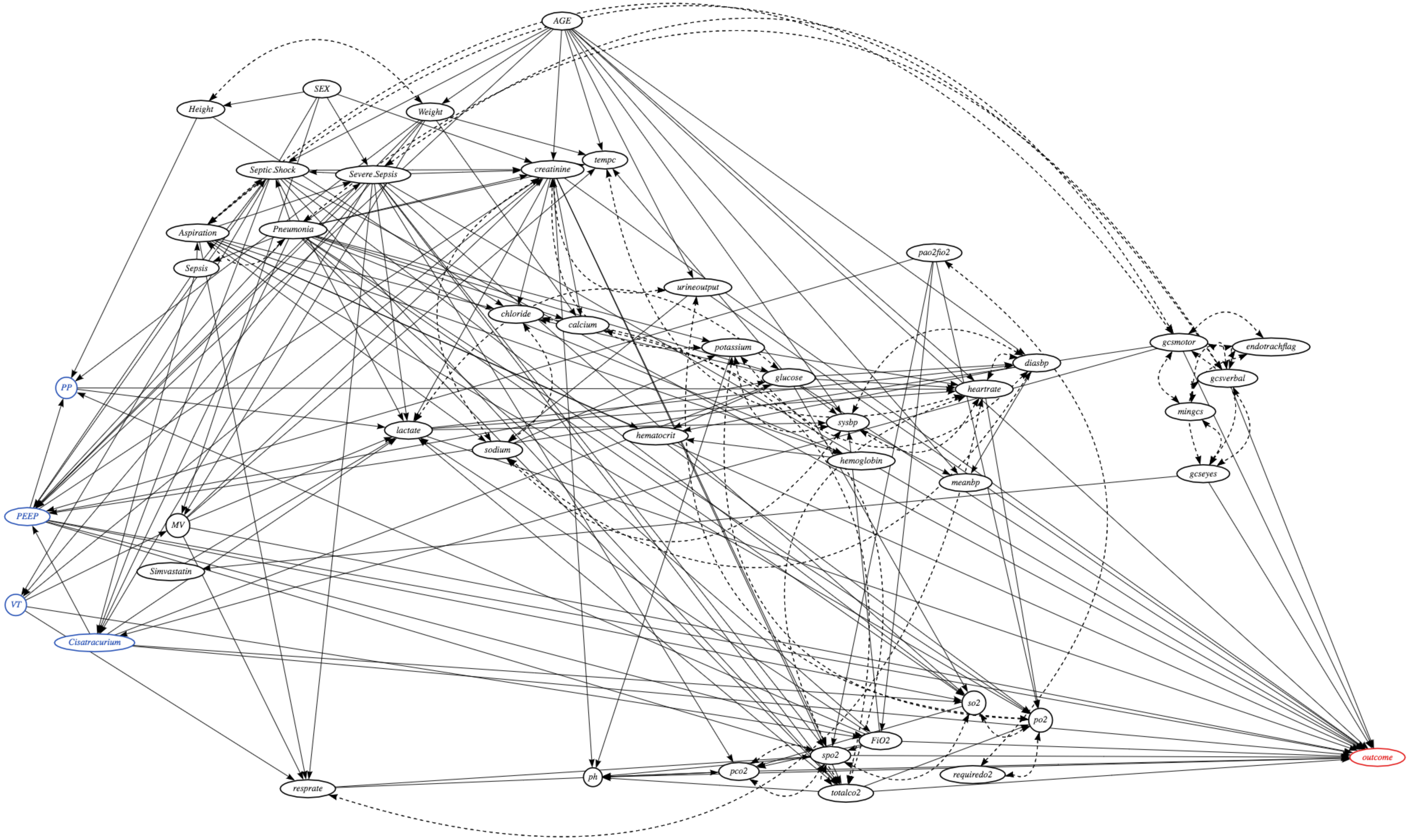
- How do we leverage *graph & L1 information* to answer L2, L3 questions?

Causal inference through SCM

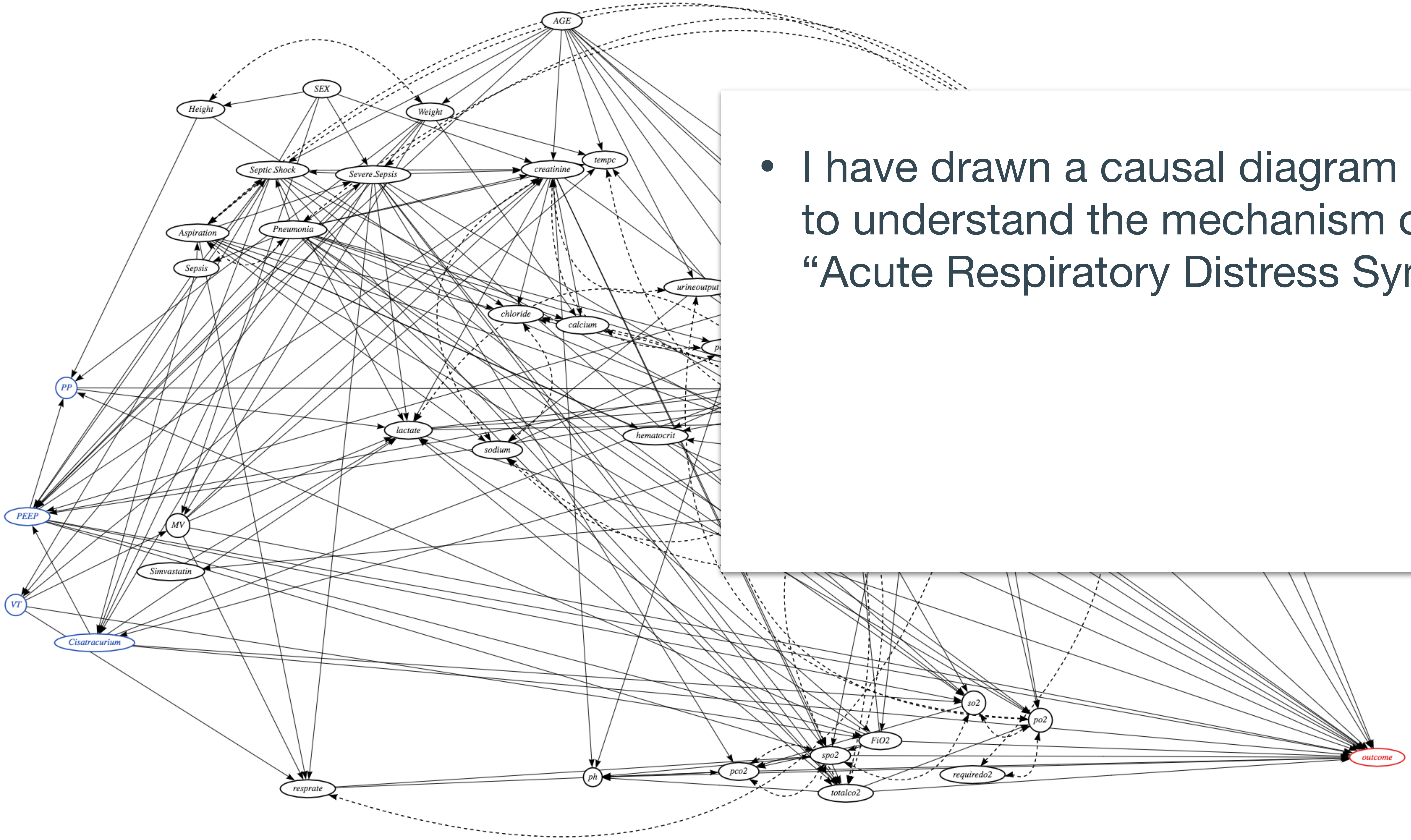


- How do we leverage *graph & L1 information* to answer L2, L3 questions?
- Without any info (i.e., no graphs), we cannot answer due to PCH)

Practical example

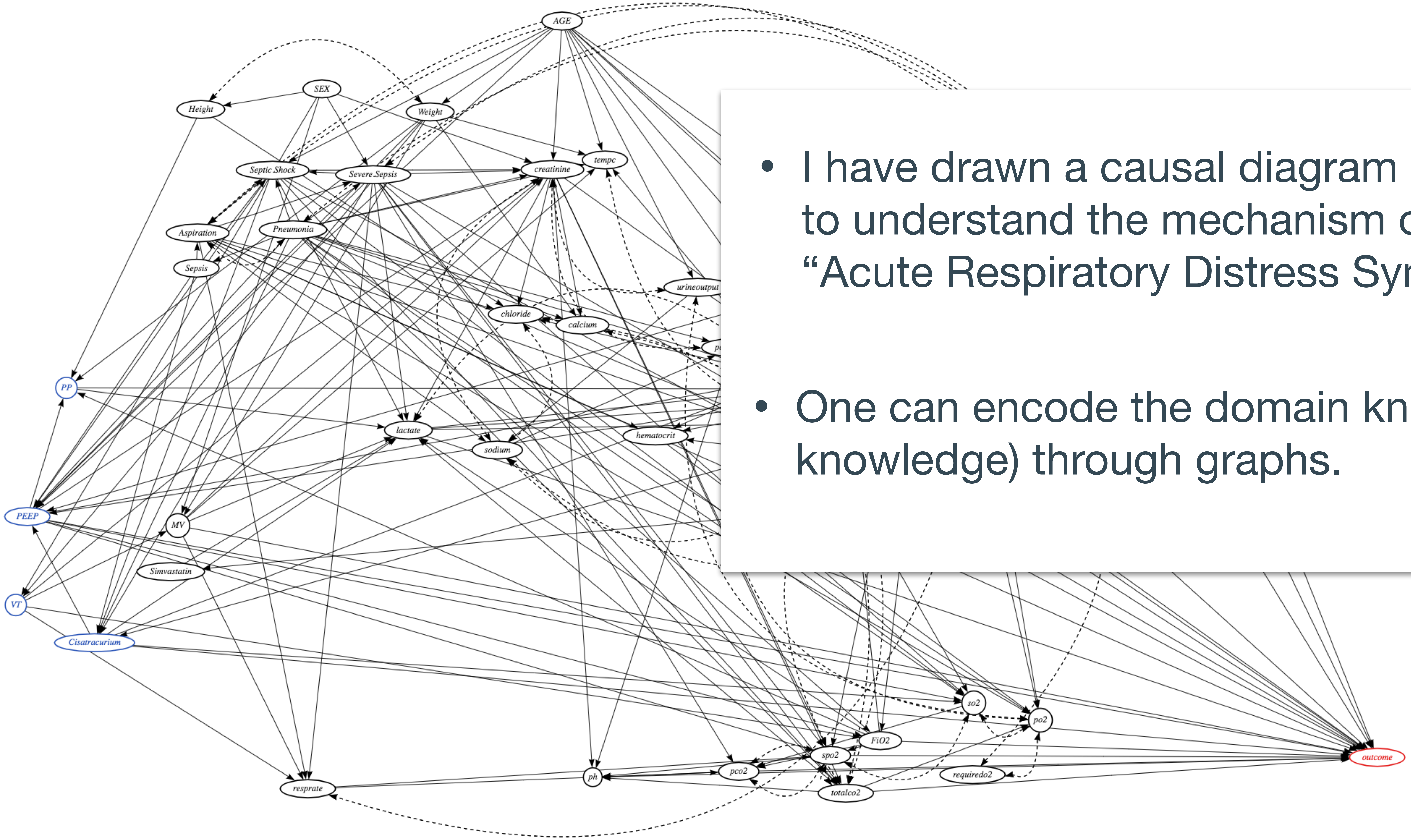


Practical example



- I have drawn a causal diagram with helps of clinicians, to understand the mechanism of the treatment effect in "Acute Respiratory Distress Syndrome (ARDS)"

Practical example



- I have drawn a causal diagram with helps of clinicians, to understand the mechanism of the treatment effect in “Acute Respiratory Distress Syndrome (ARDS)”

- One can encode the domain knowledge (clinician’s knowledge) through graphs.

Key points (So far)

Key points (So far)

- SCM is a comprehensive framework for studying causality.

Key points (So far)

- SCM is a comprehensive framework for studying causality.
    - **Unified framework**: SCM subsumes PO-based causality.

Key points (So far)

- SCM is a comprehensive framework for studying causality.
  - **Unified framework**: SCM subsumes PO-based causality.
  - **Axiomatization**: SCM is the sound and complete language obeying axioms.

Key points (So far)

- SCM is a comprehensive framework for studying causality.
  - **Unified framework**: SCM subsumes PO-based causality.
  - **Axiomatization**: SCM is the sound and complete language obeying axioms.
- SCM is a suitable tool to represent human cognition and teach them to AI.

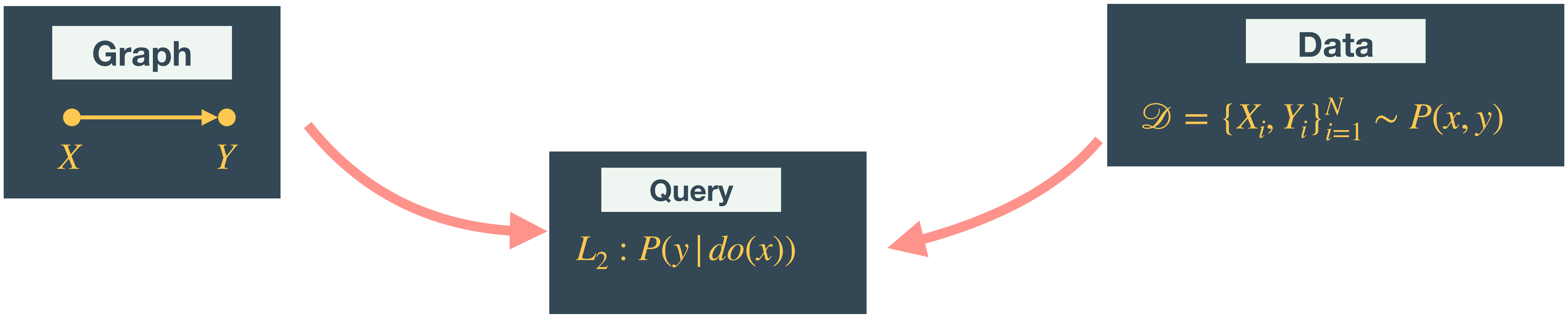
Key points (So far)

- SCM is a comprehensive framework for studying causality.
  - **Unified framework**: SCM subsumes PO-based causality.
  - **Axiomatization**: SCM is the sound and complete language obeying axioms.
- SCM is a suitable tool to represent human cognition and teach them to AI.

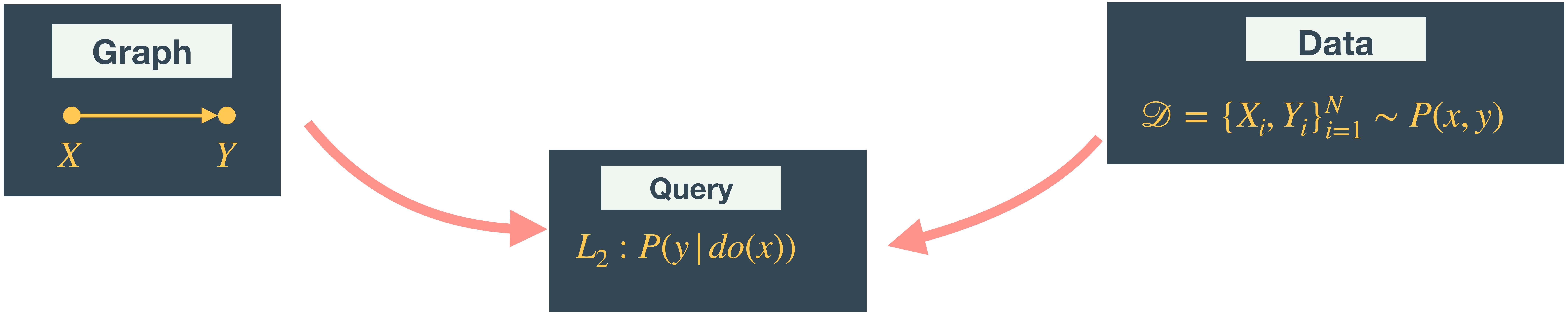
We answered Why Pearl’s Causality is revolutionary.

2. Causal effect identification

Causal effect identification

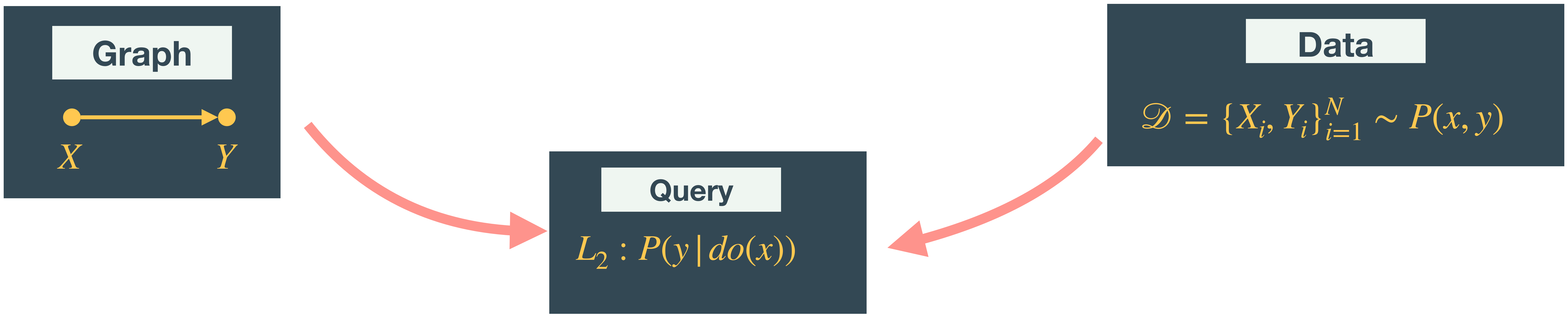


Causal effect identification



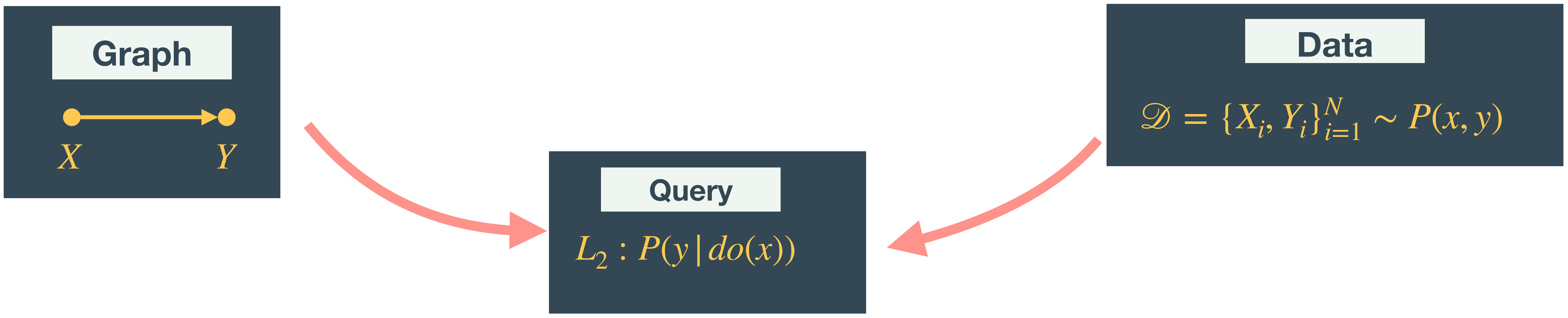
- In general, we cannot answer L2 query using the graph and data from L1 (PCH).

Causal effect identification



- In general, we cannot answer L2 query using the graph and data from L1 (PCH).
- By leveraging the graphical information, we may be able to answer!

Causal effect identification



- In general, we cannot answer L2 query using the graph and data from L1 (PCH).
- By leveraging the graphical information, we may be able to answer!
- **Causal effect identification (ID)** — Representing L2 distribution as something computable from L1 information (data drawn from the joint distribution) and graphical information.

$$\mathbb{E}[Y \mid do(X)] = \sum_z \mathbb{E}[Y \mid x, z] P(z)$$

Ignorability — Identification in PO

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been set to x in the hypothetical population.

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been set to x in the hypothetical population.

In PO, the only thing we know about Y_x is this:

- Y_x is observed when $X = x$.
- Y_x is missing when $X = x'$

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been set to x in the hypothetical population.

In PO, the only thing we know about Y_x is this:

- Y_x is observed when $X = x$.
- Y_x is missing when $X = x'$

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been ation.

b/c PO doesn't formalize the data generating process

In PO, the only thing we know about Y_x is this:

- Y_x is observed when $X = x$.
- Y_x is missing when $X = x'$

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been set to x in the hypothetical population.

In PO, the only thing we know about Y_x is this:

- Y_x is observed when $X = x$.
- Y_x is missing when $X = x'$

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been set to x in the hypothetical population.

In PO, the only thing we know about Y_x is this:

- Y_x is observed when $X = x$.
- Y_x is missing when $X = x'$

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

How can we estimate $\mathbb{E}[Y_{X=1}]$ — An expectation of Y if all population takes $X = 1$?

Ignorability — Identification in PO

Potential outcome Y_x: Y if X had been set to x in the hypothetical population.

In PO, the only thing we know about Y_x is this:

- Y_x is observed when $X = x$.
- Y_x is missing when $X = x'$

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

How can we estimate $\mathbb{E}[Y_{X=1}]$ — An expectation of Y if all population takes $X = 1$?

😱 Nontrivial, because of missing data (NA)!

Ignorability — How PO treats causality

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

Ignorability — How PO treats causality

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

We can see X as a missingness indicator ($X = 0$ means $Y_{X=1}$ missing).

Ignorability — How PO treats causality

X	Y	Y_{X=1}	Y_{X=0}	Z (age)
1	1	1	NA	1
0	0	NA	0	1
1	1	1	NA	0
0	1	NA	1	0

We can see X as a missingness indicator ($X = 0$ means $Y_{X=1}$ missing).

Missingness at random (MAR) assumption [Rudin, 1974]: Missingness (X) is independent of missing variables $(Y_{X=1})$ given some variables Z. (i.e., missingness can be explained by Z). This is a widely used assumption for imputing the missing data.

$$Y_x \perp\!\!\!\perp X \mid Z$$

Ignorability — How PO treats causality

X	Y	Y_{X=1}	Y_{X=0}	Z (age)

Ignorability assumption.

An ignobility assumption states that Y_x and X are conditionally independent given Z.

$$Y_x \perp\!\!\!\perp X | Z$$

Missingness at random (MAR) assumption [Rudin, 1974]: Missingness (X) is independent of missing variables ($Y_{X=1}$) given some variables Z. (i.e., missingness can be explained by Z). This is a widely used assumption for imputing the missing data.

$$Y_x \perp\!\!\!\perp X | Z$$

Ignorability — How PO treats causality

X	Y	Y_{X=1}	Y_{X=0}	Z (age)

Ignorability assumption.

An ignobility assumption states that Y_x and X are conditionally independent given Z.

$$Y_x \perp\!\!\!\perp X \mid Z$$

Covariate adjustment - Identification w/ Ignoriability assumption

- If the ignorability assumption holds, then

$$\mathbb{E}[Y_x] = \sum_z \mathbb{E}[Y_x \mid z]P(z) = \sum_z \mathbb{E}[Y_x \mid x, z]P(z) = \sum_z \mathbb{E}[Y \mid x, z]P(z)$$

Ignorability — How PO treats causality

X	Y	Y_{X=1}	Y_{X=0}	Z (age)

Ignorability assumption.

An ignobility assumption states that Y_x and X are conditionally independent given Z.

$$Y_x \perp\!\!\!\perp X | Z$$

Covariate adjustment - Identification w/ Ignoriability assumption

- If the ignorability assumption holds, then

$$\mathbb{E}[Y_x] = \sum_z \mathbb{E}[Y_x | z]P(z) = \sum_z \mathbb{E}[Y_x | x, z]P(z) = \sum_z \mathbb{E}[Y | x, z]P(z)$$

L2 quantity

L1 quantity

Practical implication of ignorability?

Practical implication of ignorability?

What $Y_x \perp\!\!\!\perp X \mid Z$ means ("missingness of Y_x can be explained by Z") is unclear in practice.

Practical implication of ignorability?

What $Y_x \perp\!\!\!\perp X \,|\, Z$ means ("missingness of Y_x can be explained by Z") is unclear in practice.

What about $Z = \{$ variables correlated with {X,Y}$\}$? Can missingness of Y_x be explained by such Z?

Practical implication of ignorability?

What $Y_x \perp\!\!\!\perp X | Z$ means ("missingness of Y_x can be explained by Z") is unclear in practice.

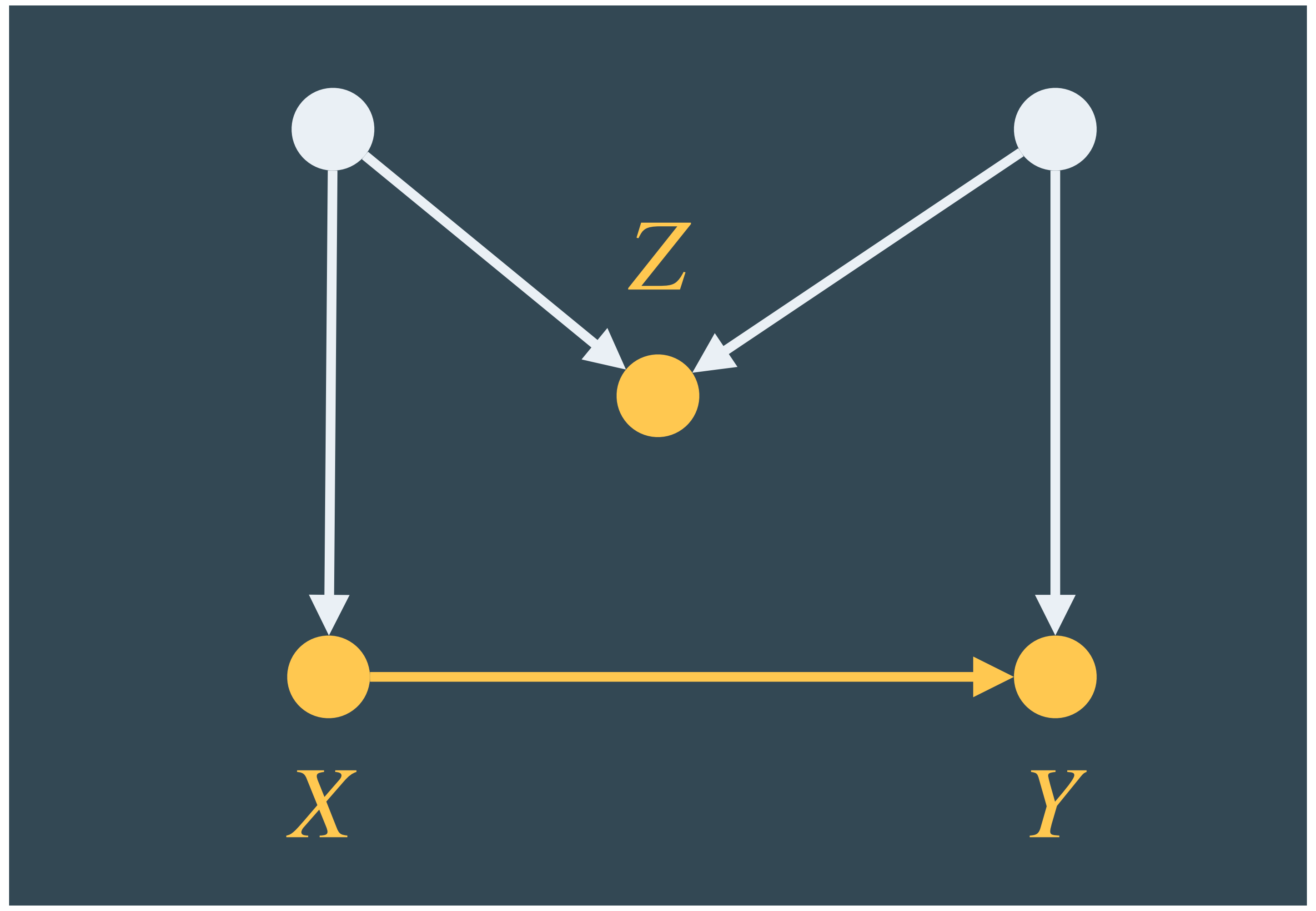
What about $Z = \{$ variables correlated with {X,Y}$\}$? Can missingness of Y_x be explained by such Z?

No!

Practical implication of ignorability?

What $Y_x \perp\!\!\!\perp X \mid Z$ means ("missingness of Y_x can be explained by Z") is unclear in practice.

What about $Z = \{$ variables correlated with {X,Y}$\}$? Can missingness of Y_x be explained by such Z?



"M-bias" [Pearl]: Counterexample that for such Z, still $Y_x \not\!\perp\!\!\!\perp X \mid Z$.

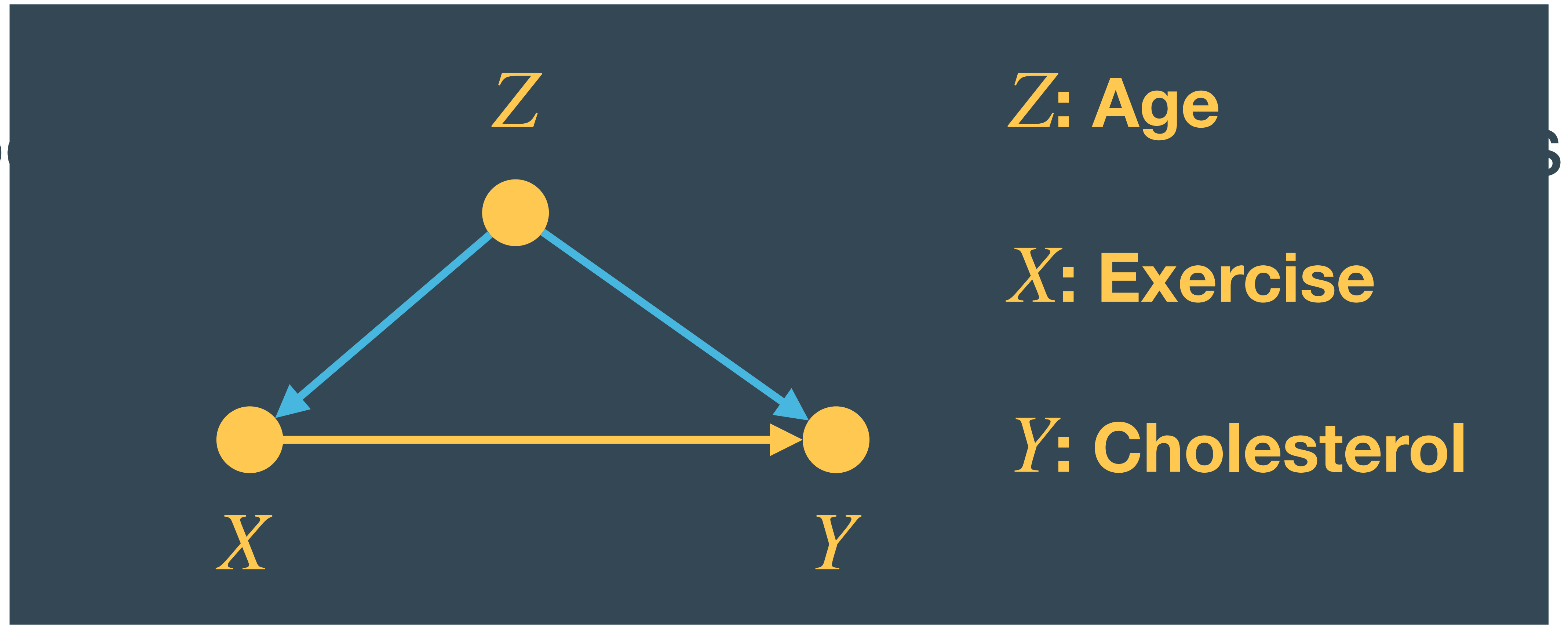
Back-door criterion

Back-door criterion

Pearl provides "Back-door criterion", a graphical criterion corresponding to the ignorability criterion.

Back-door criterion

Pearl provides “Back-do … sponding to the ignorability criterion.



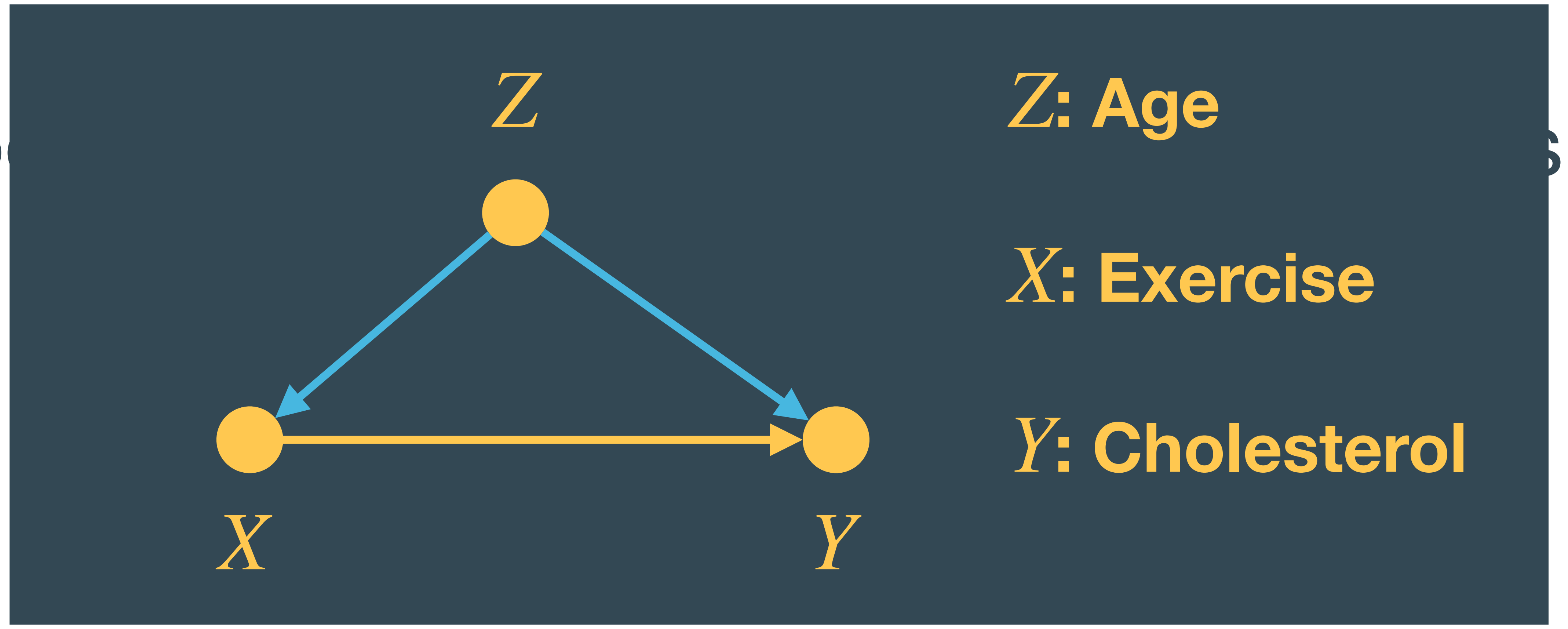
Back-door criterion

Given G, if all the **non-causal** path (or spurious path, indirect path) from X and Y is blocked by Z, then $\mathbb{E}[Y_x]$ (= $\mathbb{E}[Y \mid do(x)]$ in terms of SCM) is

$$\mathbb{E}[Y \mid do(x)] = \sum_z \mathbb{E}[Y \mid x, z] P(z) \, .$$

Back-door criterion

Pearl provides “Back-do … sponding to the ignorability criterion.



Back-door criterion

Given G, if all the **non-causal** path (or spurious path, indirect path) from X and Y is blocked by Z, then $\mathbb{E}[Y_x]$ (= $\mathbb{E}[Y \mid do(x)]$ in terms of SCM) is

$$\underbrace{\mathbb{E}[Y \mid do(x)]}_{\text{L2 quantity}} = \sum_z \underbrace{\mathbb{E}[Y \mid x, z]P(z)}_{\text{L1 quantity}} \,.$$

Beyond the Back-door

Beyond the Back-door

Recall that *No formal data generating process on* Y_x in the PO-based causality.

Beyond the Back-door

Recall that *No formal data generating process on* Y_x in the PO-based causality.

- Only information: Y_x can be viewed as missing data $\Rightarrow$ ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

Beyond the Back-door

Recall that *No formal data generating process on Y_x in the PO-based causality.*

- Only information: Y_x can be viewed as missing data $\Rightarrow$ ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

If $Y_x \not\!\perp\!\!\!\perp X \mid Z$, in the PO-framework, we can do nothing b/c no further information can be used.

Beyond the Back-door

Recall that *No formal data generating process on* Y_x in the PO-based causality.

- Only information: Y_x can be viewed as missing data $\Rightarrow$ ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

If $Y_x \not\perp\!\!\!\perp X \mid Z$, in the PO-framework, we can do nothing b/c no further information can be used.

Does this mean that the causal effect is NOT identifiable if $Y_x \not\perp\!\!\!\perp X \mid Z$?

Beyond the Back-door

Recall that *No formal data generating process on Y_x in the PO-based causality.*

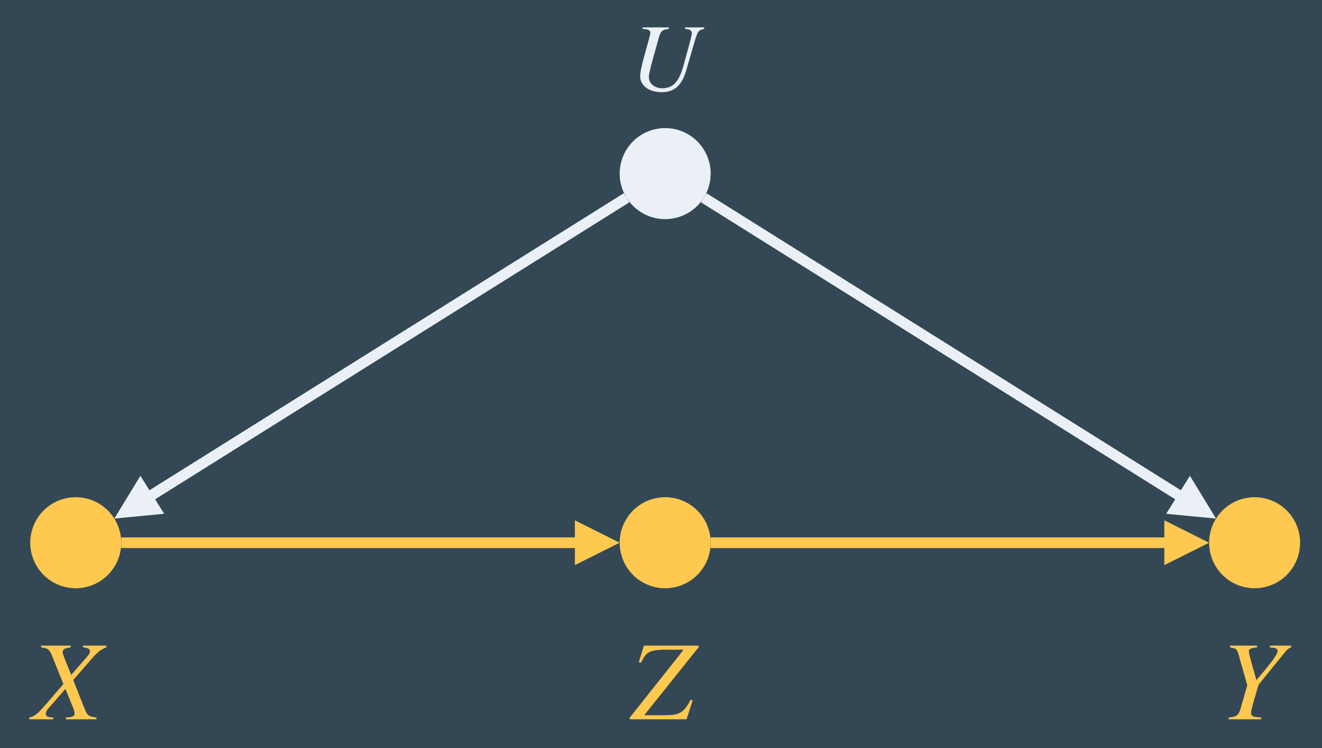
- Only information: Y_x can be viewed as missing data $\Rightarrow$ ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

If $Y_x \not\!\perp\!\!\!\perp X \mid Z$, in the PO-framework, we can do nothing b/c no further information can be used.

Does this mean that the causal effect is NOT identifiable if $Y_x \not\!\perp\!\!\!\perp X \mid Z$?

🙅‍♂️ No!

Front-door



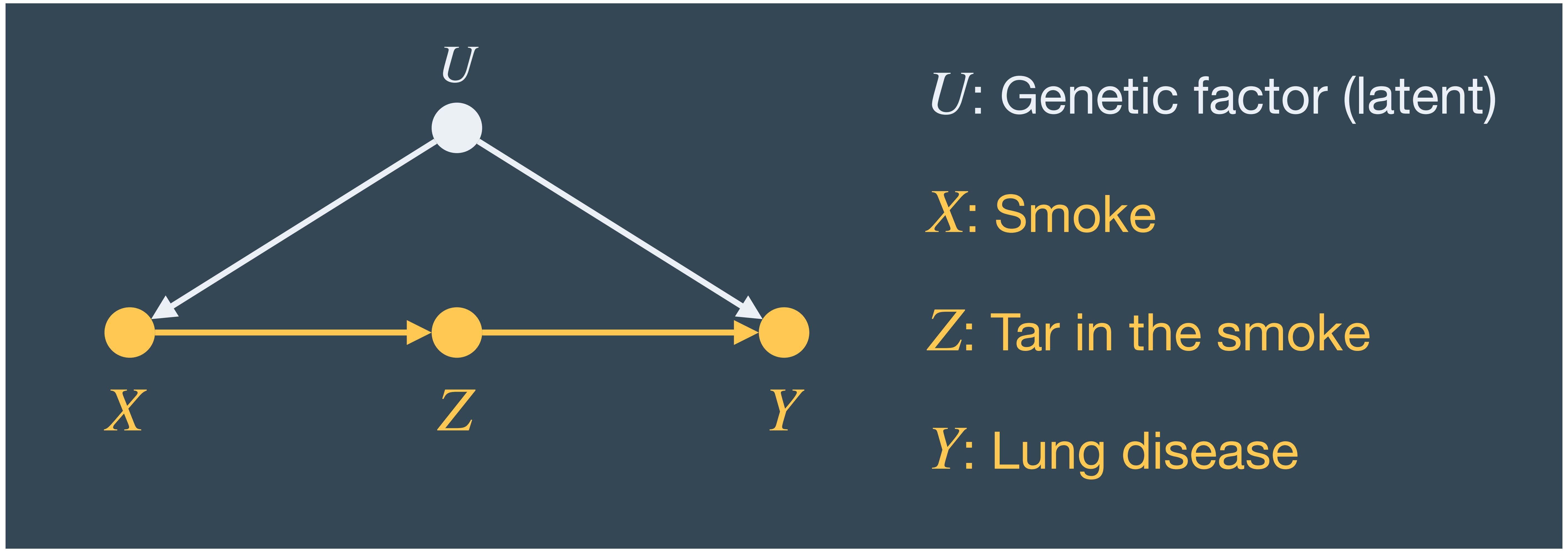
U: Genetic factor (latent)

X: Smoke

Z: Tar in the smoke

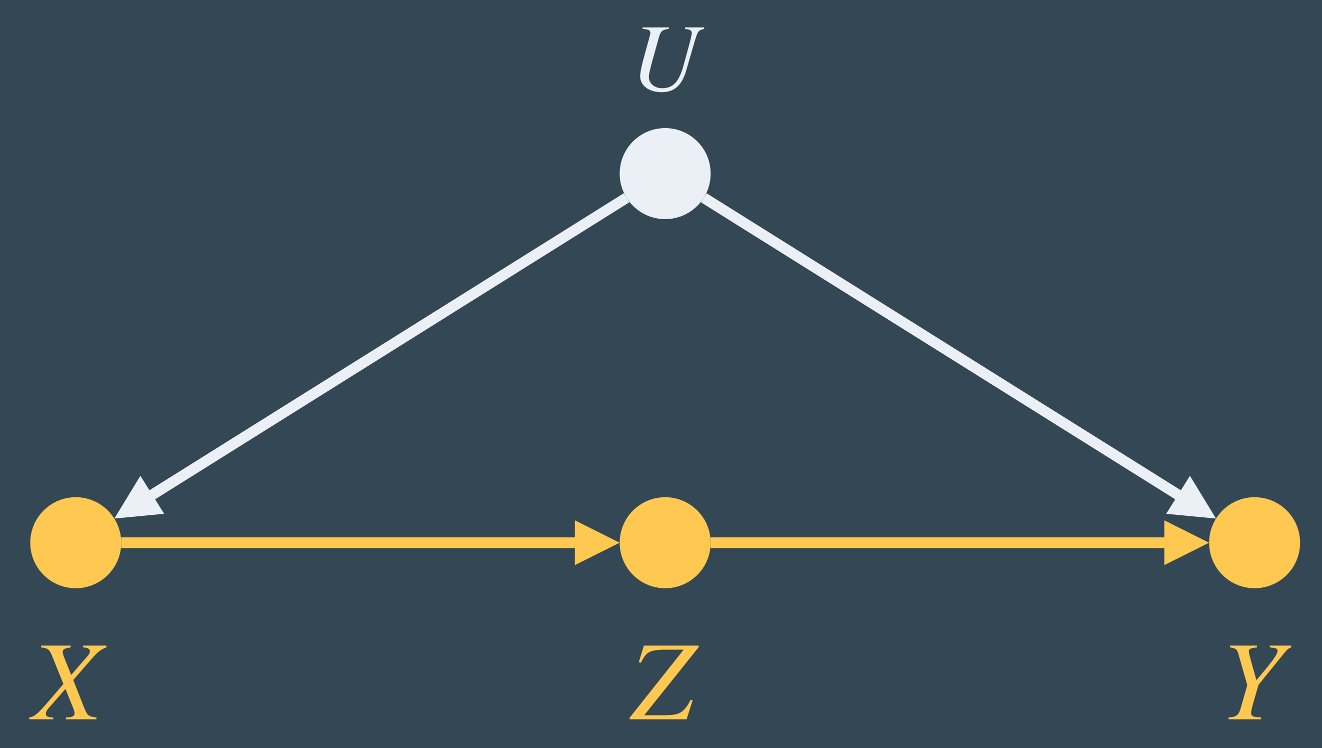
Y: Lung disease

Front-door



In Front-door graph [Pearl, 1995], the ignorability doesn't hold: $Y_x \not\perp\!\!\!\perp X \mid Z$

Front-door



U: Genetic factor (latent)

X: Smoke

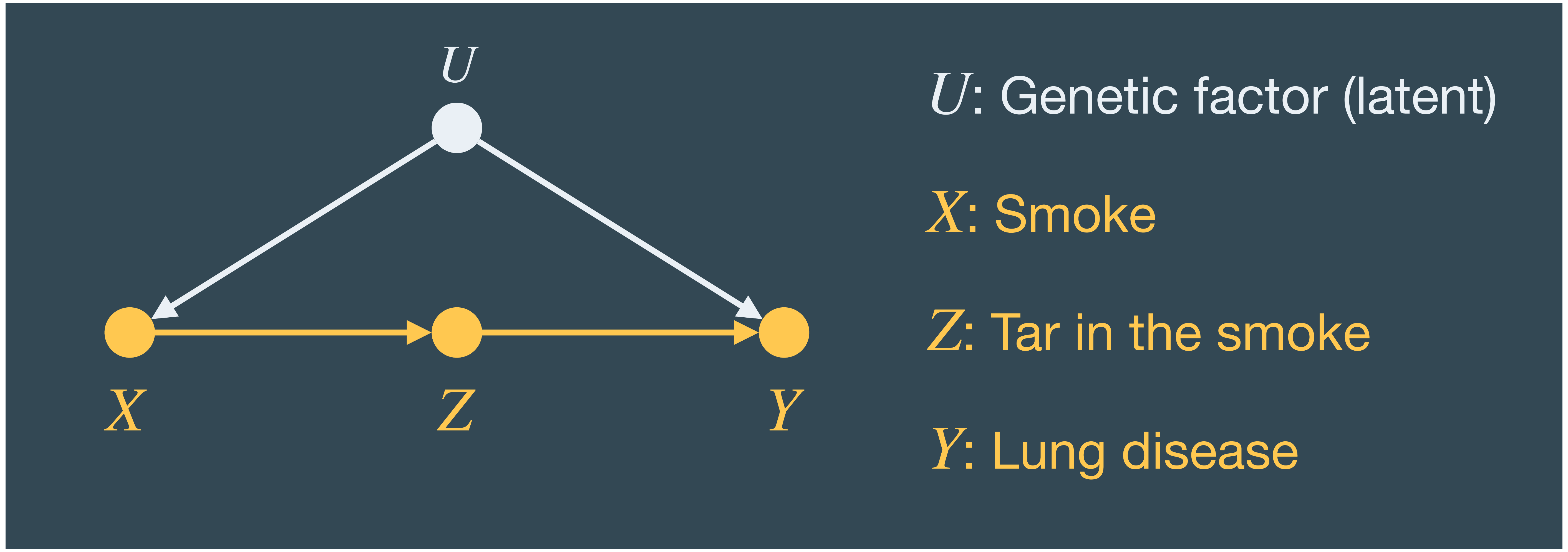
Z: Tar in the smoke

Y: Lung disease

In Front-door graph [Pearl, 1995], the ignorability doesn't hold: $Y_x \not\perp\!\!\!\perp X \mid Z$

However, $\mathbb{E}[Y_x] \equiv \mathbb{E}[Y \mid do(x)]$ is identifiable and given as

Front-door



In Front-door graph [Pearl, 1995], the ignorability doesn't hold: $Y_x \not\perp\!\!\!\perp X \mid Z$

However, $\mathbb{E}[Y_x] \equiv \mathbb{E}[Y \mid do(x)]$ is identifiable and given as

$$\mathbb{E}[Y \mid do(x)] = \sum_z P(z \mid x) \sum_{x'} \mathbb{E}[Y \mid x', z] P(x').$$

Front-door

U: Genetic factor (latent)

Front-door is the 1st example showing the insufficiency of the ignorability

However, $\mathbb{E}[Y_x] \equiv \mathbb{E}[Y \mid do(x)]$ is identifiable and given as

$$\mathbb{E}[Y \mid do(x)] = \sum_z P(z \mid x) \sum_{x'} \mathbb{E}[Y \mid x', z] P(x').$$

Pearl's do-calculus

Pearl's do-calculus

Motivated by Front-door example, Pearl [1995] developed three rules that can be used for identifying causal effect from a graph w/ $Y_x \not\perp\!\!\!\perp X \mid Z$.

Pearl's do-calculus

Motivated by Front-door example, Pearl [1995] developed three rules that can be used for identifying causal effect from a graph w/ $Y_x \not\perp\!\!\!\perp X \mid Z$.

Rule 1 (conditional independence):

$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}} \Rightarrow P(y \mid do(x), z, w) = P(y \mid do(x), w)$$

Pearl’s do-calculus

Motivated by Front-door example, Pearl [1995] developed three rules that can be used for identifying causal effect from a graph w/ $Y \not\perp\!\!\!\perp X \mid Z$

$G_{\overline{A}\underline{B}}$: A graph cutting incoming edges to A, and outgoing edges from B.

Rule 1 (conditional independenc

$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}} \Rightarrow P(y \mid do(x), z, w) = P(y \mid do(x), w)$$

Pearl's do-calculus

Motivated by Front-door example, Pearl [1995] developed three rules that can be used for identifying causal effect from a graph w/ $Y_x \not\perp\!\!\!\perp X \mid Z$.

Rule 1 (conditional independence):

$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}} \Rightarrow P(y \mid do(x), z, w) = P(y \mid do(x), w)$$

Rule 2 (Doing/seeing interchange):

$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}} \Rightarrow P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$$

Pearl's do-calculus

Motivated by Front-door example, Pearl [1995] developed three rules that can be used for identifying causal effect from a graph w/ $Y_x \not\perp\!\!\!\perp X \mid Z$.

Rule 1 (conditional independence):

$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}} \Rightarrow P(y \mid do(x), z, w) = P(y \mid do(x), w)$$

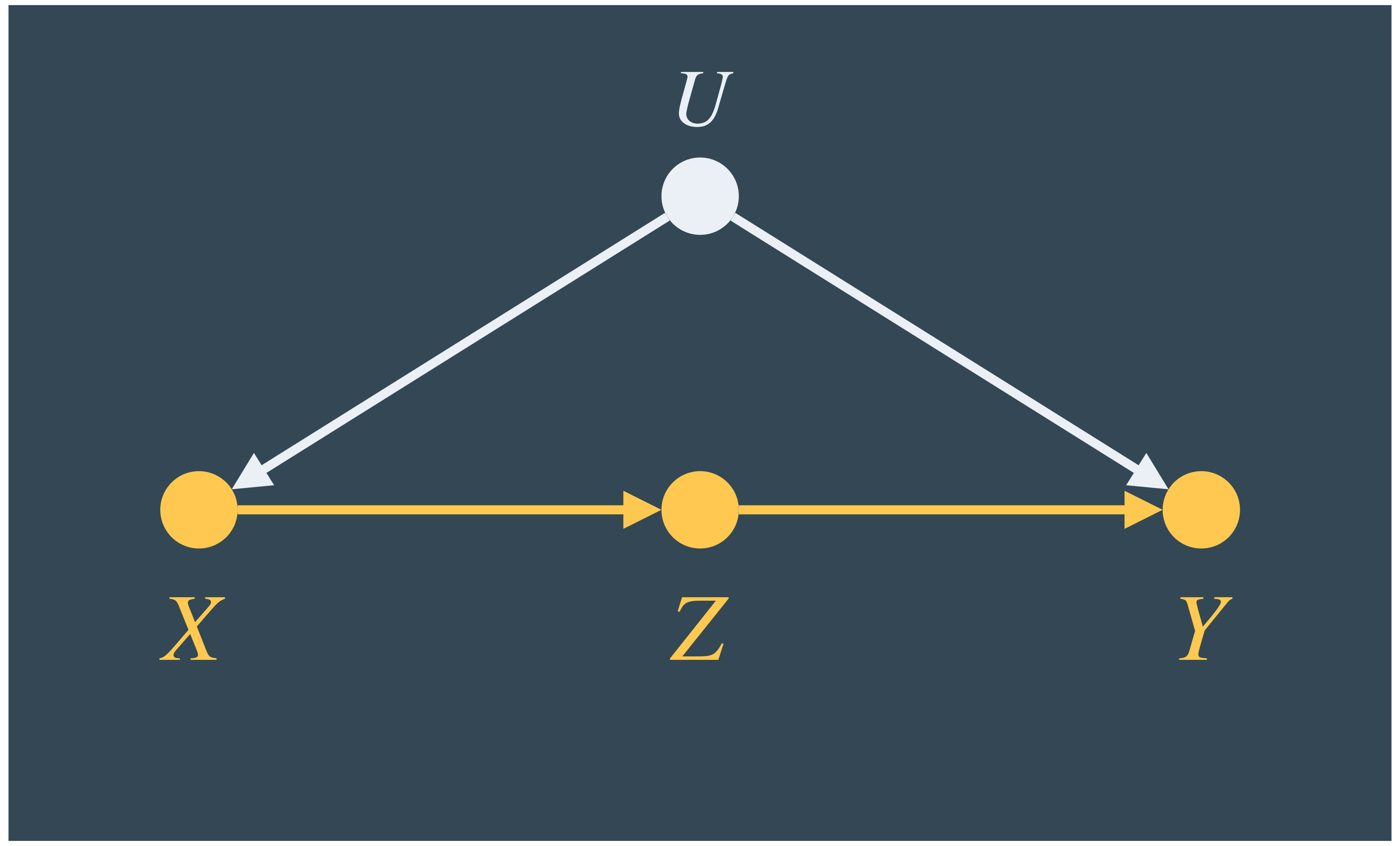
Rule 2 (Doing/seeing interchange):

$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}} \Rightarrow P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$$

Rule 3 (conditional independence for interventions)

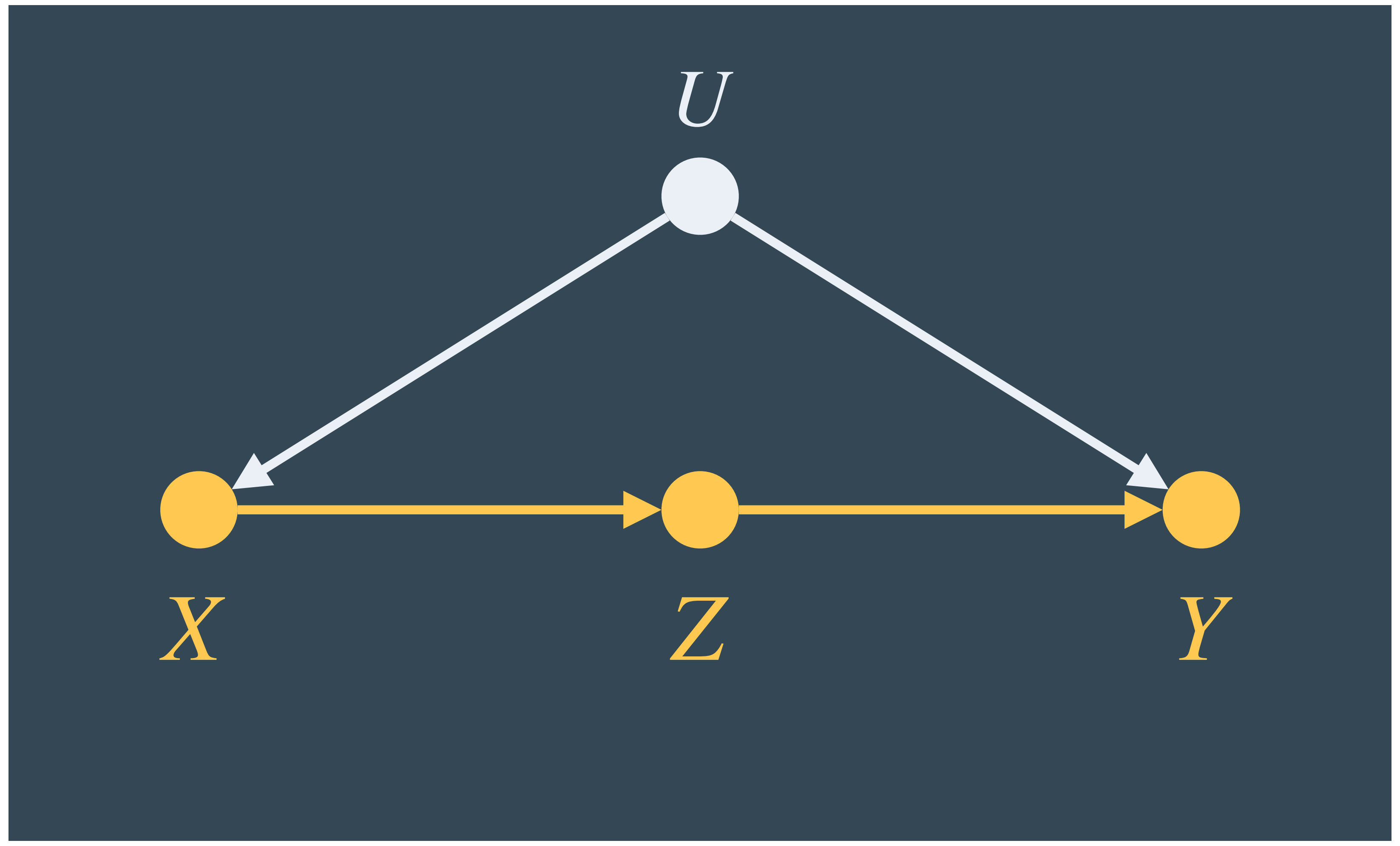
$$(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}, \overline{Z \setminus An(W)_{G_{\overline{X}}}}}} \Rightarrow P(y \mid do(x), do(z), w) = P(y \mid do(z), w)$$

Front-door —Identification through do-calculus



Front-door —Identification through do-calculus

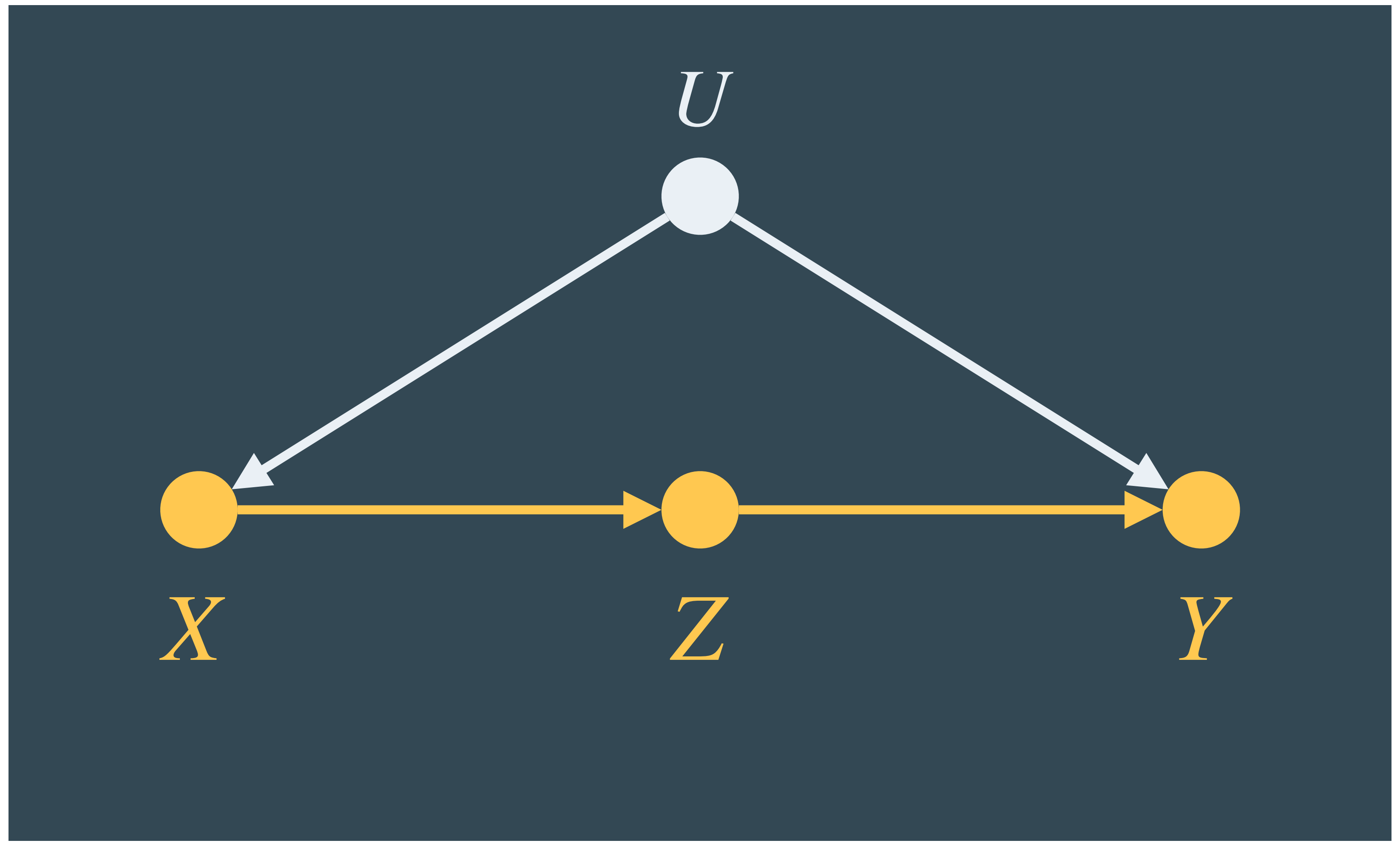
$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$ Marginalization



Front-door —Identification through do-calculus

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

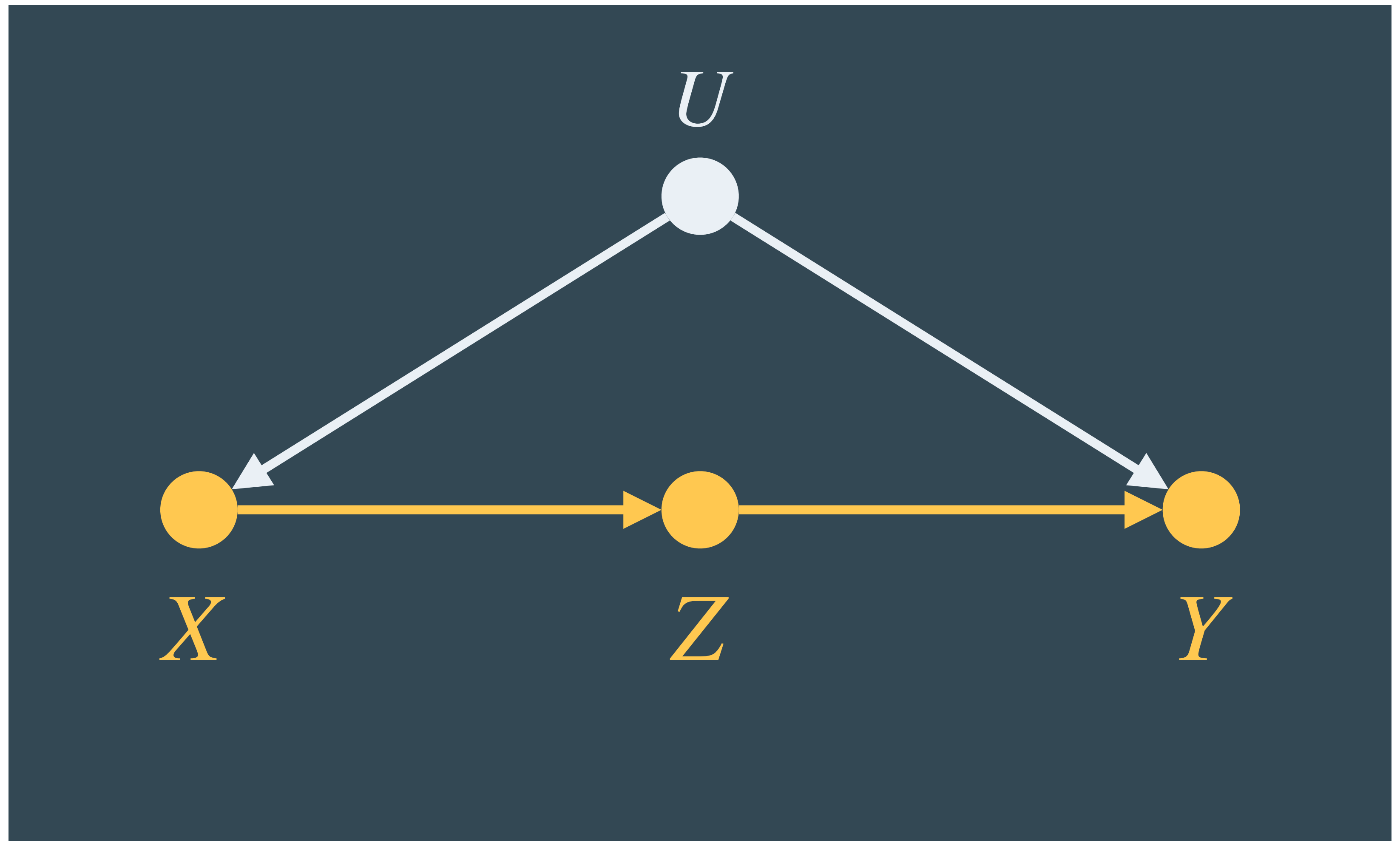


Front-door —Identification through do-calculus

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(x), do(z)) P(z \mid x)$$ R2



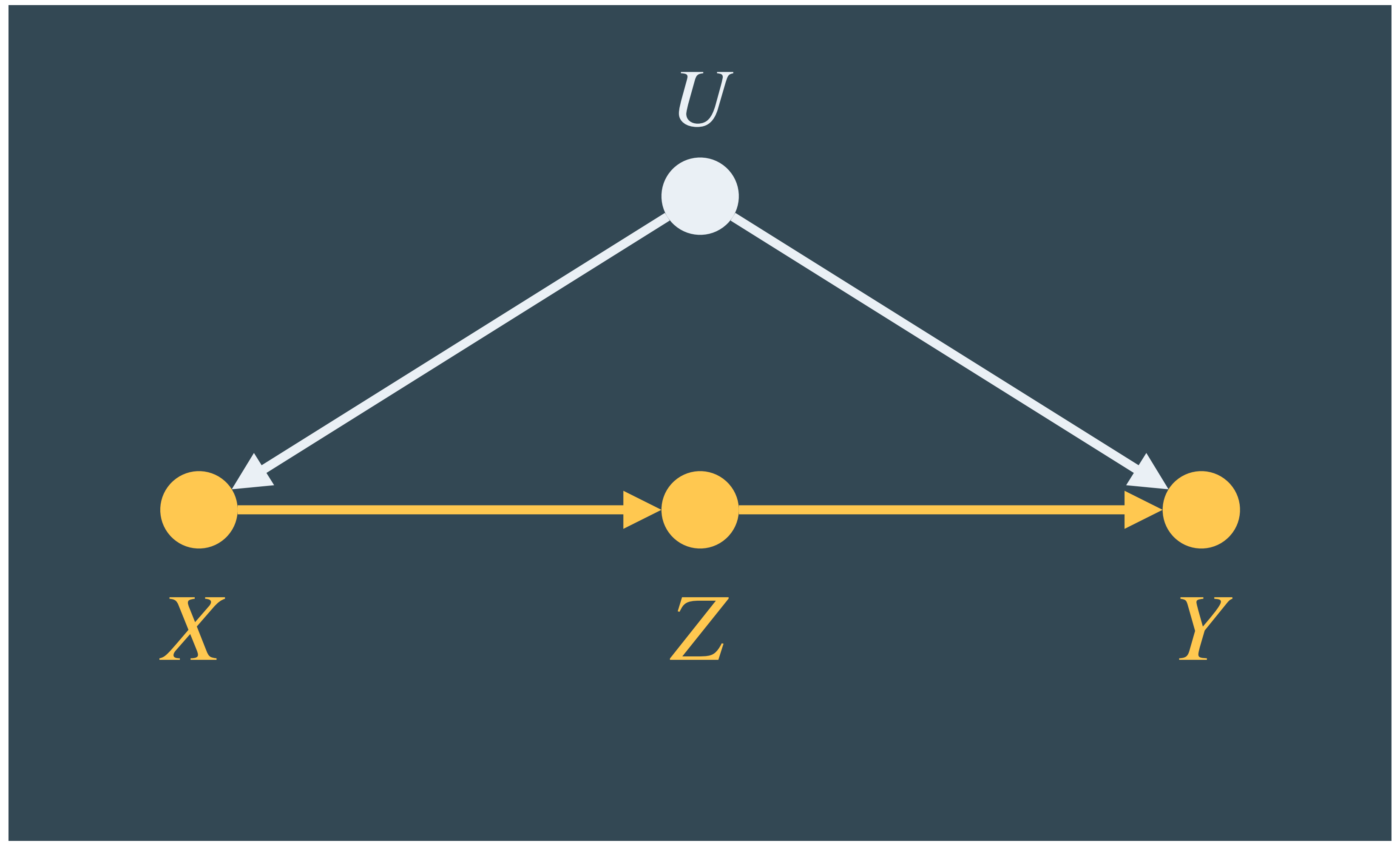
Front-door —Identification through do-calculus

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(x), do(z)) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(z)) P(z \mid x)$$ R3



Front-door —Identification through do-calculus

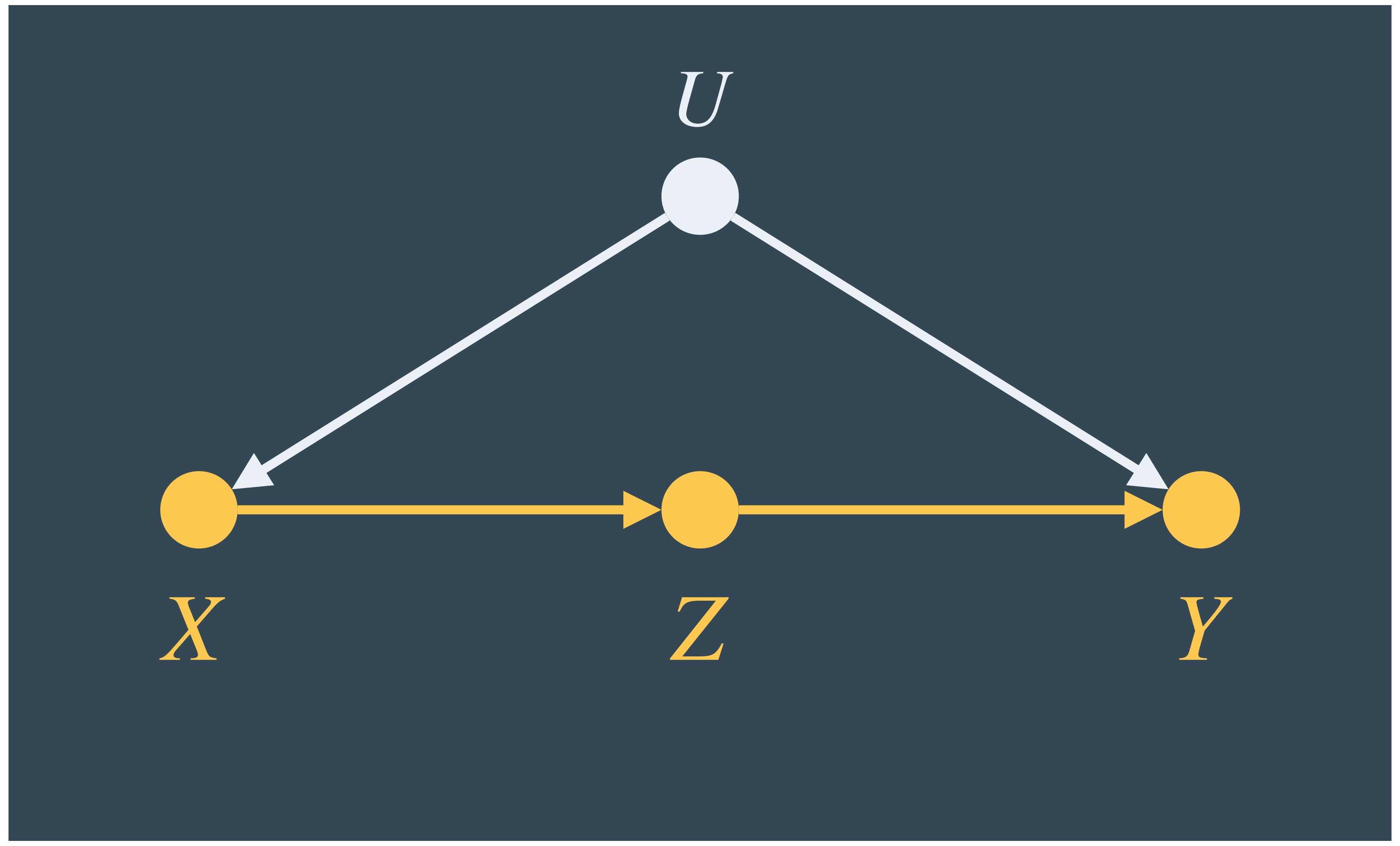
$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(x), do(z)) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(z)) P(z \mid x)$$ R3

$$= \sum_z \left(\sum_{x'} P(y \mid do(z), x') P(x' \mid do(z)) \right) P(z \mid x).$$ Marginalization



Front-door —Identification through do-calculus

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

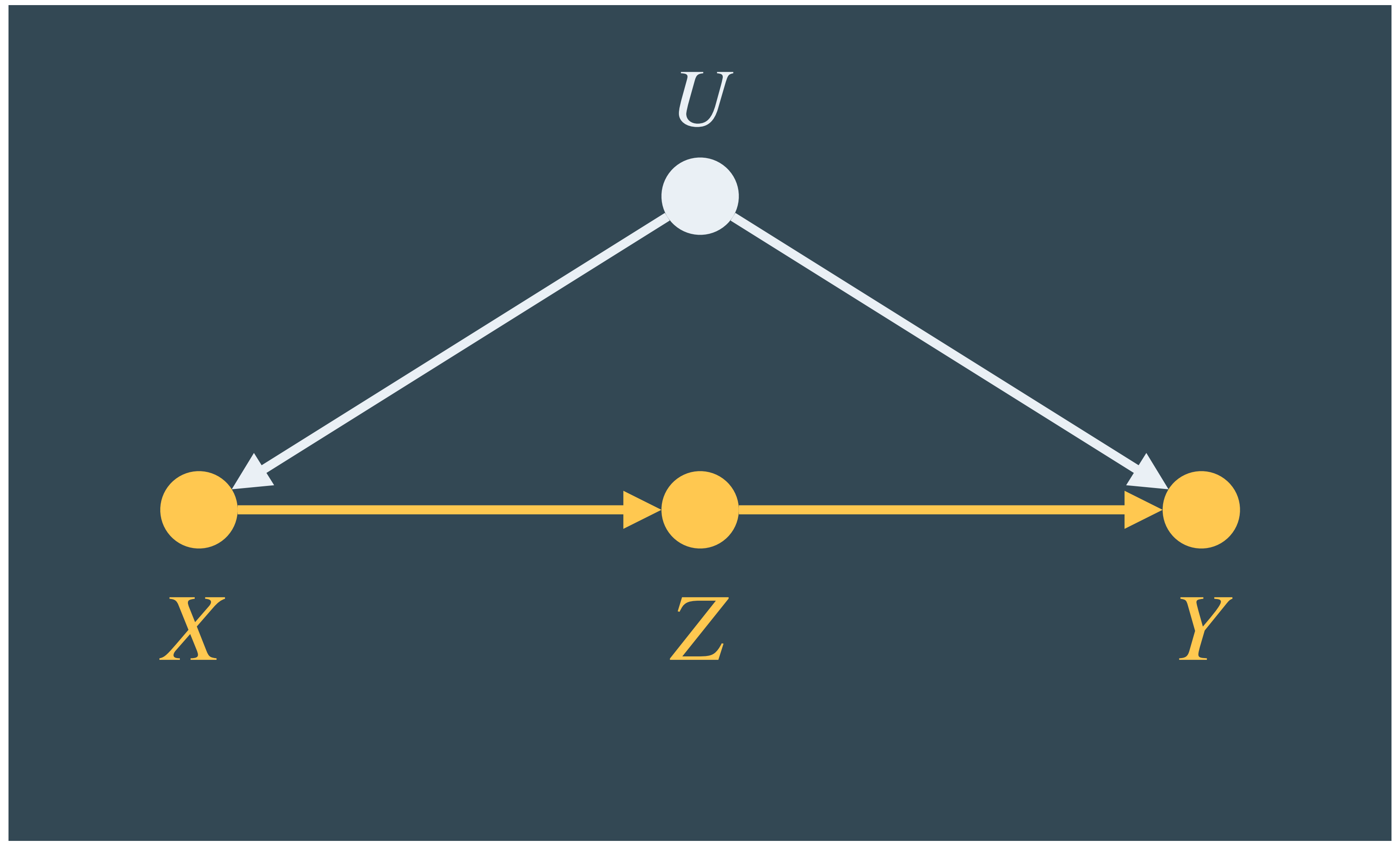
$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(x), do(z)) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(z)) P(z \mid x)$$ R3

$$= \sum_z \left(\sum_{x'} P(y \mid do(z), x') P(x' \mid do(z)) \right) P(z \mid x).$$ Marginalization

$$= \sum_z \left(\sum_{x'} P(y \mid z, x') P(x' \mid do(z)) \right) P(z \mid x).$$ R2



Front-door —Identification through do-calculus

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

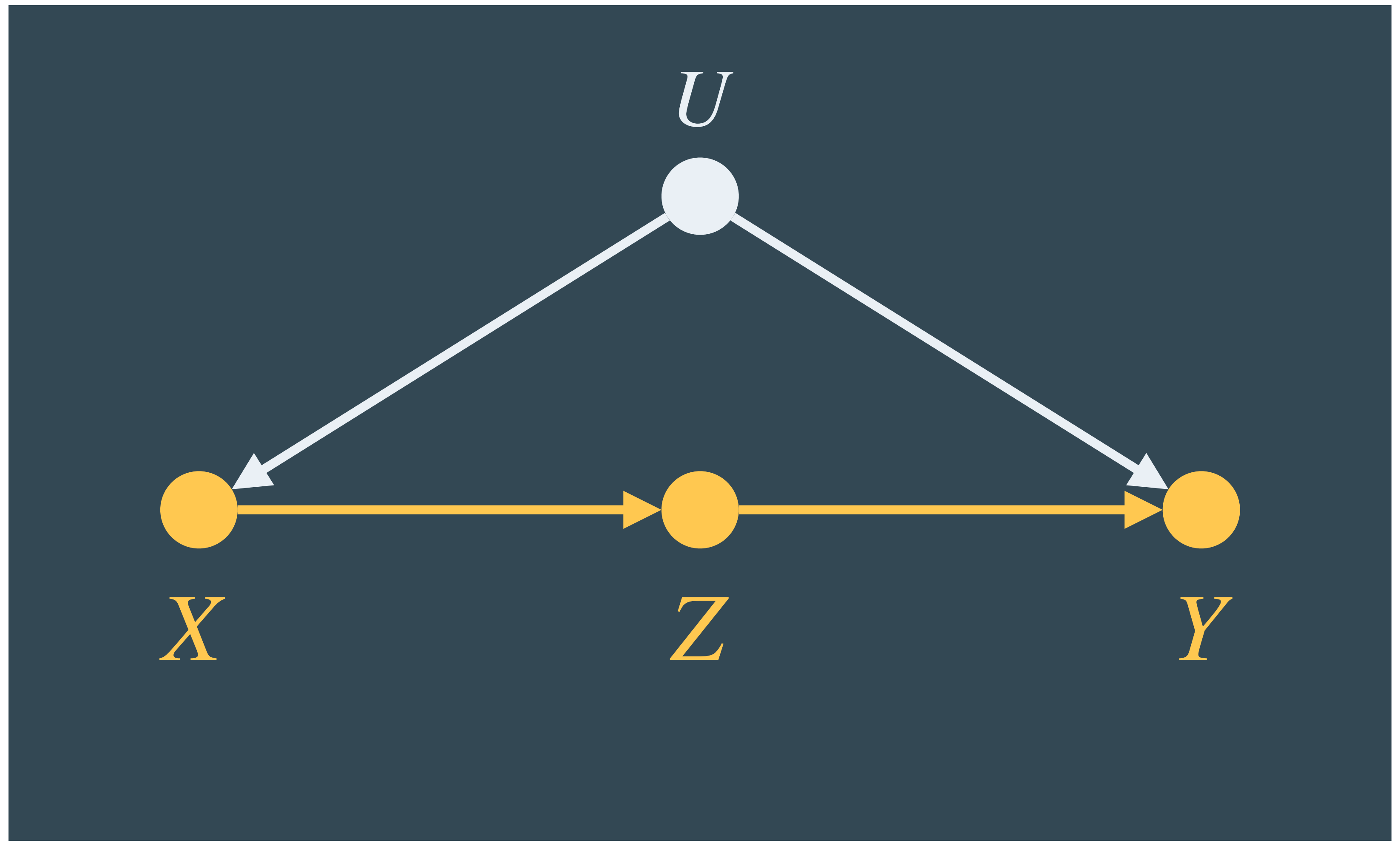
$$= \sum_z P(y \mid do(x), do(z)) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(z)) P(z \mid x)$$ R3

$$= \sum_z \left(\sum_{x'} P(y \mid do(z), x') P(x' \mid do(z)) \right) P(z \mid x).$$ Marginalization

$$= \sum_z \left(\sum_{x'} P(y \mid z, x') P(x' \mid do(z)) \right) P(z \mid x).$$ R2

$$= \sum_z \left(\sum_{x'} P(y \mid z, x') P(x') \right) P(z \mid x).$$ R3



Front-door —Identification through do-calculus

$$P(y \mid do(x)) = \sum_z P(y \mid do(x), z) P(z \mid do(x))$$ Marginalization

$$= \sum_z P(y \mid do(x), z) P(z \mid x)$$ R2

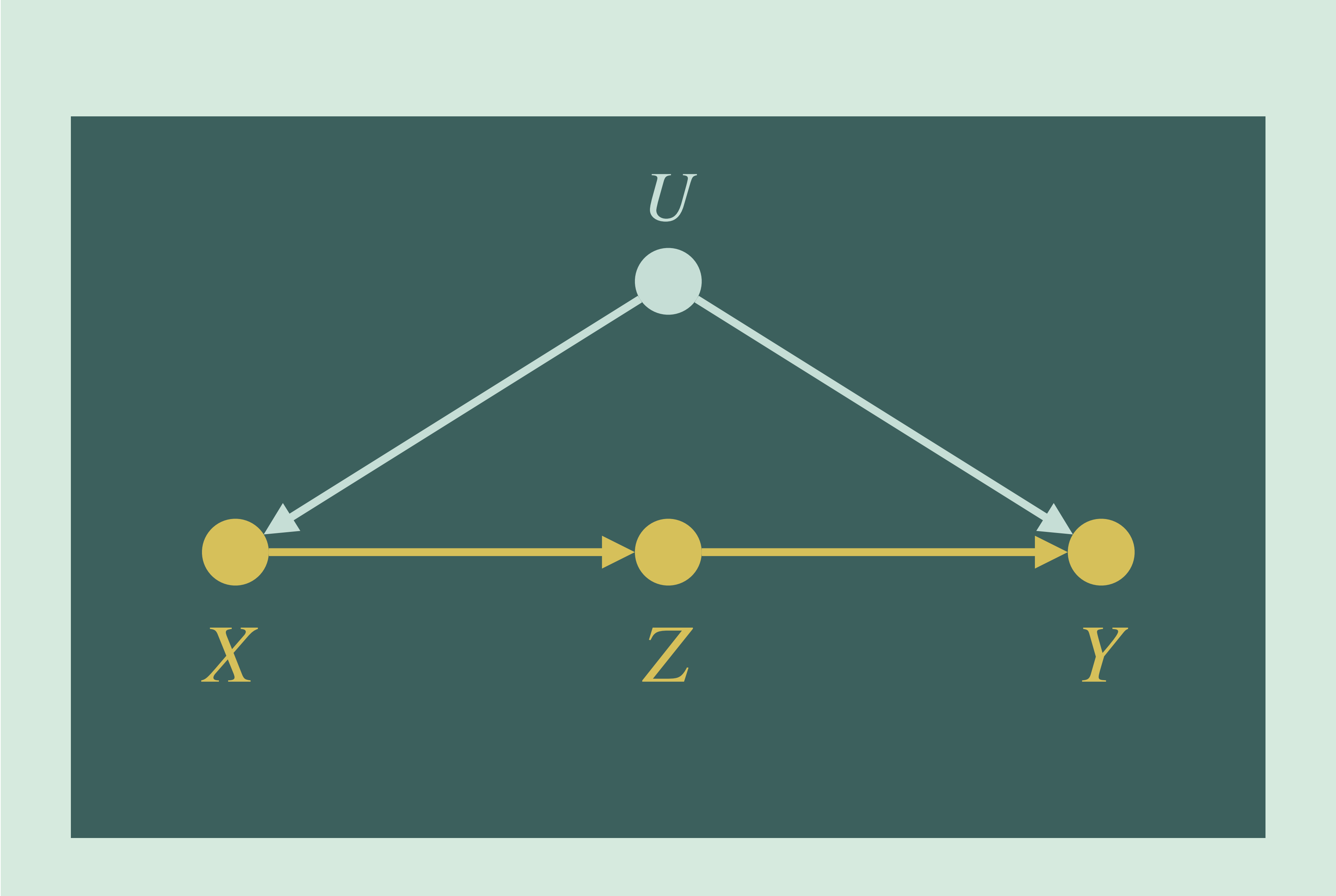
$$= \sum_z P(y \mid do(x), do(z)) P(z \mid x)$$ R2

$$= \sum_z P(y \mid do(z)) P(z \mid x)$$ R3

$$= \sum_z \left(\sum_{x'} P(y \mid do(z), x') P(x' \mid do(z)) \right) P(z \mid x).$$ Marginalization

$$= \sum_z \left(\sum_{x'} P(y \mid z, x') P(x' \mid do(z)) \right) P(z \mid x).$$ R2

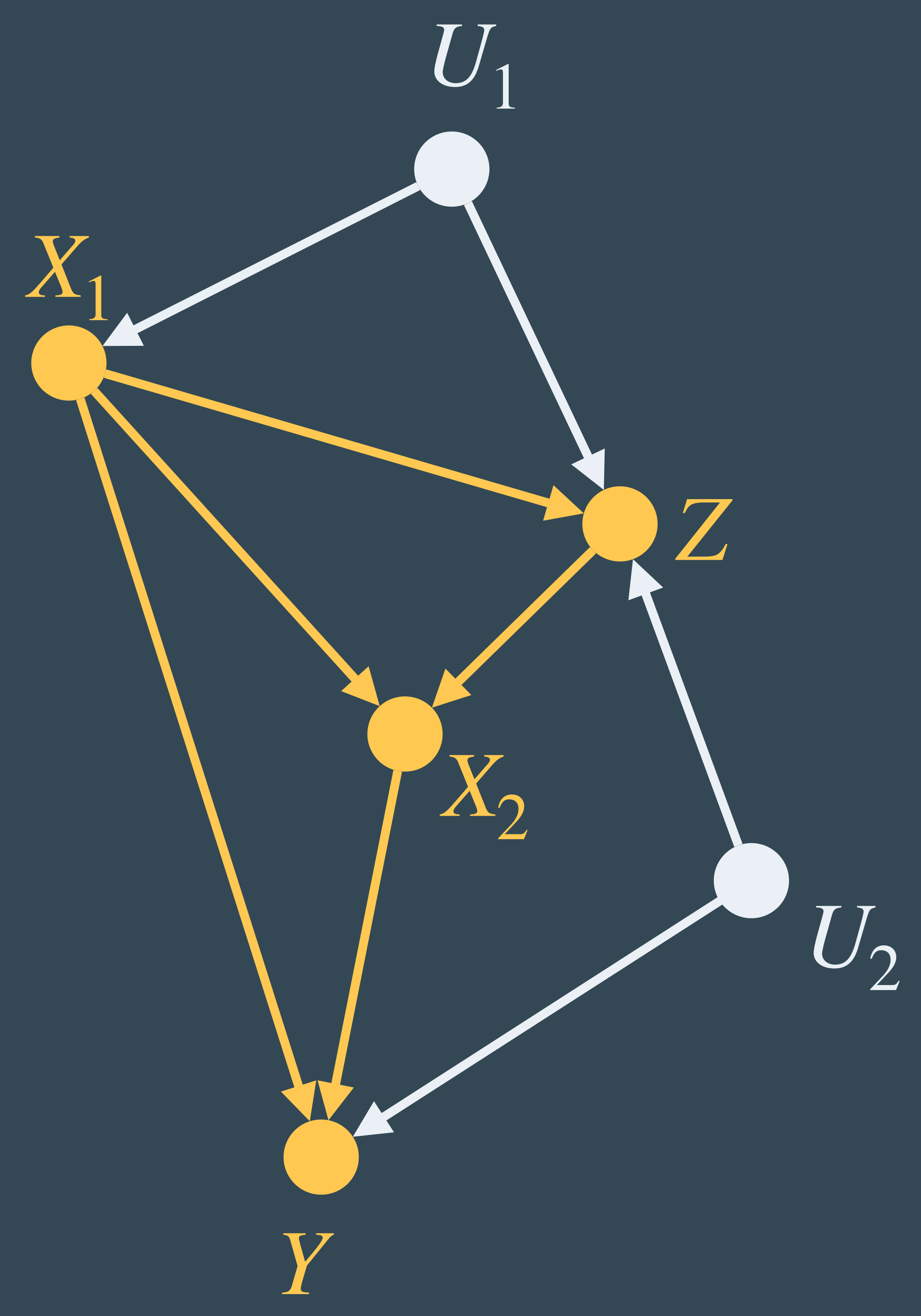
$$= \sum_z \left(\sum_{x'} P(y \mid z, x') P(x') \right) P(z \mid x).$$ R3



L2 quantity is represented as an L1 quantity given the graph through do-calculus rules.

What about other graphs?

What about other graphs?



X_1, X_2: Treatment at time $1{,}2$

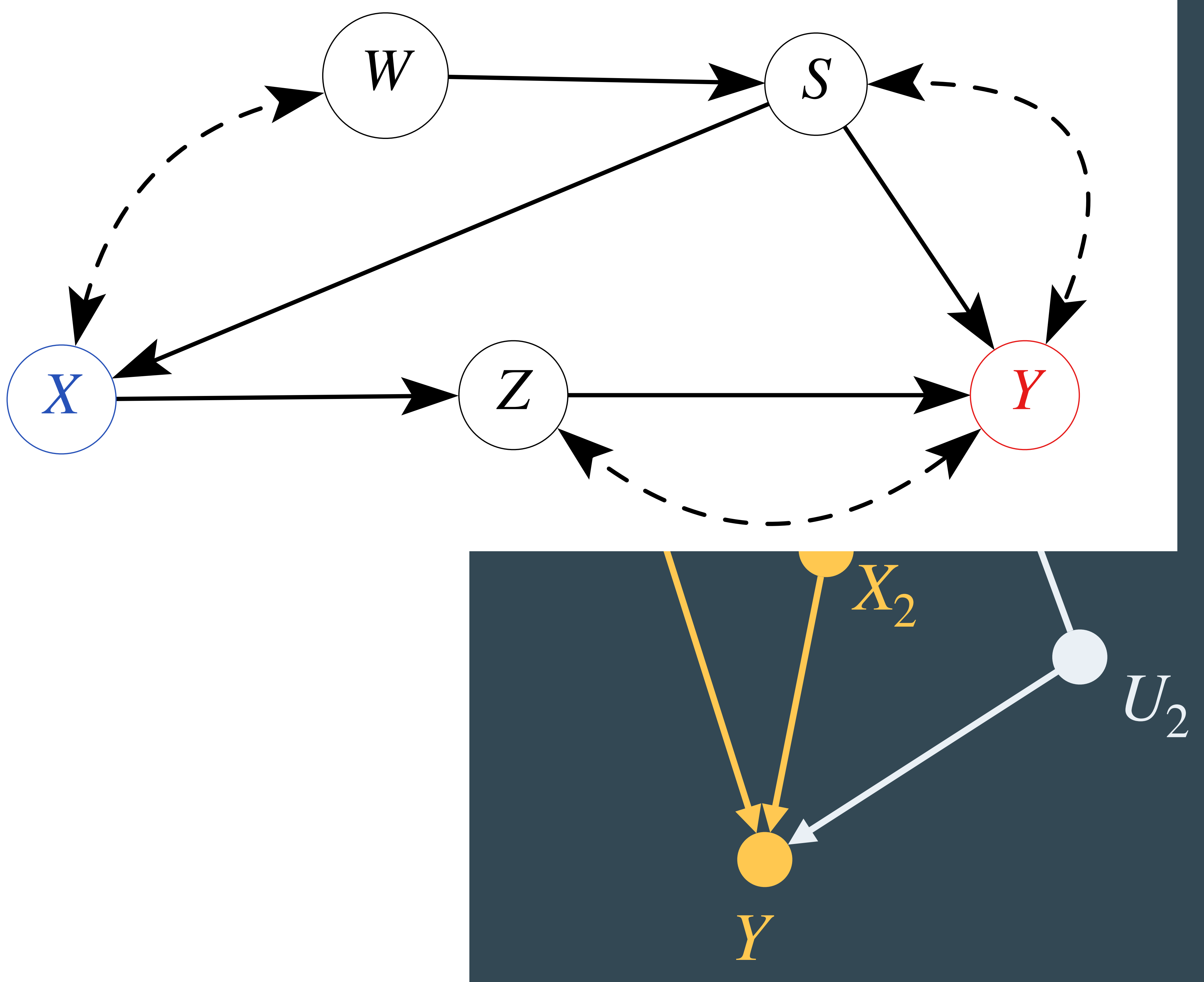
Z: Physiologic response.

Y: Survival

U_1: Patients' history

U_2: Genetic factor

What about other graphs?



X_1, X_2: Treatment at time $1{,}2$

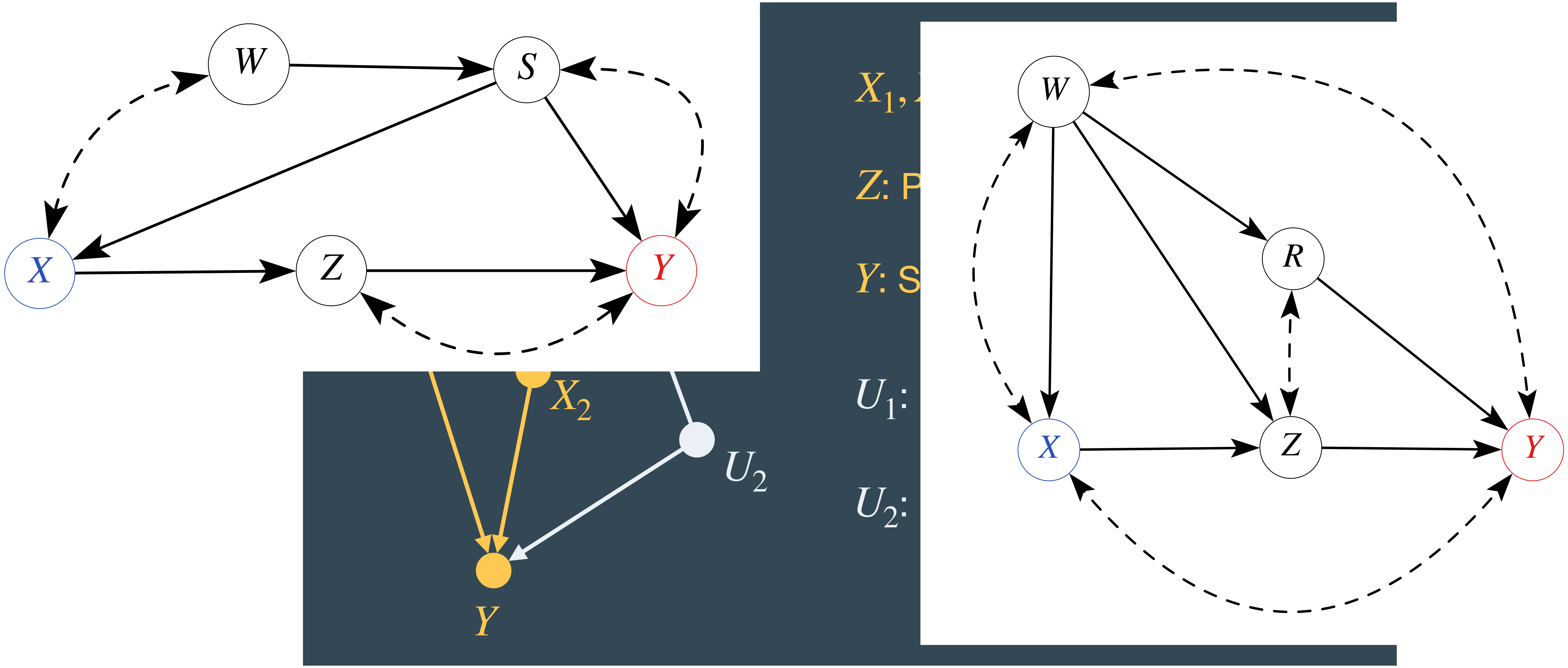
Z: Physiologic response.

Y: Survival

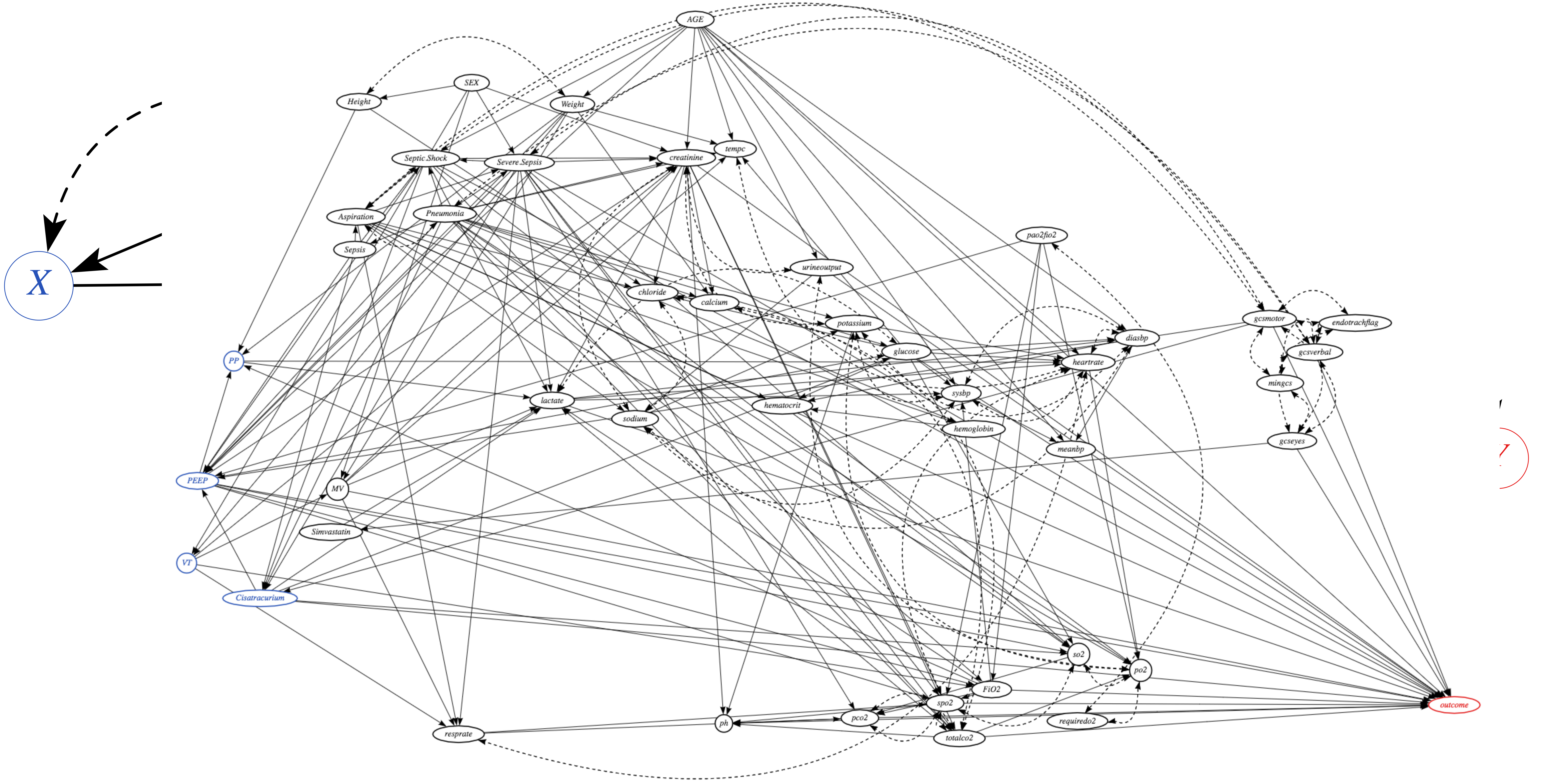
U_1: Patients' history

U_2: Genetic factor

What about other graphs?



What about other graphs?



What about other graphs?

What about other graphs?

Nature's data generating process can be arbitrary.

What about other graphs?

Nature's data generating process can be arbitrary.

$\Rightarrow$ Causal graphs can be arbitrary.

What about other graphs?

Nature's data generating process can be arbitrary.

$\Rightarrow$ Causal graphs can be arbitrary.

Q1. Can we determine identifiability using do-calculus for arbitrary graphs?

What about other graphs?

Nature's data generating process can be arbitrary.

$\Rightarrow$ Causal graphs can be arbitrary.

Q1. Can we determine identifiability using do-calculus for arbitrary graphs?

Q2. If so, how do we find a correct procedure for applying do-calculus?

Complete identification solution

Complete identification solution

Q1. Can we determine identifiability using do-calculus for arbitrary graphs?

Complete identification solution

Q1. Can we determine identifiability using do-calculus for arbitrary graphs?

A1. Yes. Do-calculus is **complete** (i.e., the causal effect is identifiable if and only if it can be derived through do-calculus) [Tian, 2002], [Valtorta and Huang, 2006] ,[Shpitser and Pearl, 2006]

Complete identification solution

Q1. Can we determine identifiability using do-calculus for arbitrary graphs?

A1. Yes. Do-calculus is **complete** (i.e., the causal effect is identifiable if and only if it can be derived through do-calculus) [Tian, 2002], [Valtorta and Huang, 2006] ,[Shpitser and Pearl, 2006]

Q2. How do we find a correct procedure for applying do-calculus?

Complete identification solution

Q1. Can we determine identifiability using do-calculus for arbitrary graphs?

A1. Yes. Do-calculus is **complete** (i.e., the causal effect is identifiable if and only if it can be derived through do-calculus) [Tian, 2002], [Valtorta and Huang, 2006] ,[Shpitser and Pearl, 2006]

Q2. How do we find a correct procedure for applying do-calculus?

A2. There is an algorithm! (https://www.causalfusion.net/login)

Key points (So far)

Key points (So far)

- Under SCM frameworks, a sound and complete algorithm (i.e., identifiable if and only the algorithm works) for determining identifiability (CausalFusion) exists.

Key points (So far)

- Under SCM frameworks, a sound and complete algorithm (i.e., identifiable if and only the algorithm works) for determining identifiability (CausalFusion) exists.
    - In the PO-based causality, no formal data generating process (DGP) on Y_x

ID via ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

Key points (So far)

- Under SCM frameworks, a sound and complete algorithm (i.e., identifiable if and only the algorithm works) for determining identifiability (CausalFusion) exists.

  - In the PO-based causality, no formal data generating process (DGP) on Y_x

    ID via ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

  - SCM frameworks allow to encode knowledge on DGPs.

Key points (So far)

- Under SCM frameworks, a sound and complete algorithm (i.e., identifiable if and only the algorithm works) for determining identifiability (CausalFusion) exists.
  - In the PO-based causality, no formal data generating process (DGP) on Y_x

    ID via ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

  - SCM frameworks allow to encode knowledge on DGPs.
  - Since Nature's DGP is arbitrary, causal graphs can be arbitrary.

Key points (So far)

- Under SCM frameworks, a sound and complete algorithm (i.e., identifiable if and only the algorithm works) for determining identifiability (CausalFusion) exists.
    - In the PO-based causality, no formal data generating process (DGP) on Y_x

      ID via ignorability assumption ($Y_x \perp\!\!\!\perp X \mid Z$)

    - SCM frameworks allow to encode knowledge on DGPs.
    - Since Nature's DGP is arbitrary, causal graphs can be arbitrary.
    - That's why SCM frameworks engenders causal effect identification problems.

Task of Identification

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

Task of Identification

1 query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$$

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

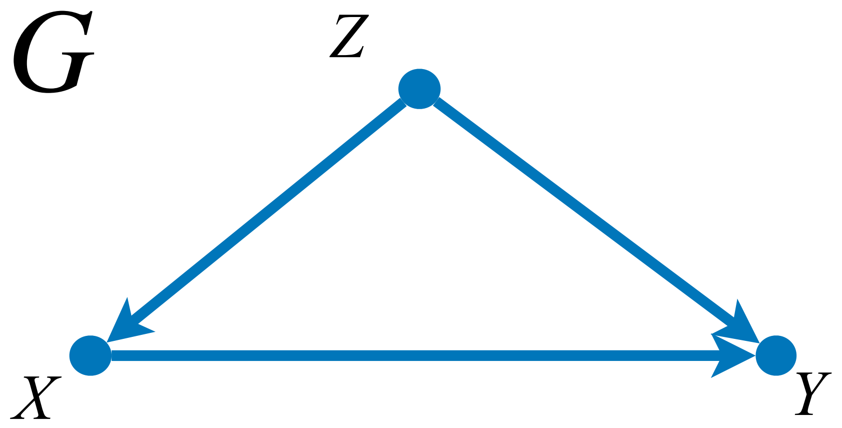
Task of Identification

1 query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$$

2 graph

G



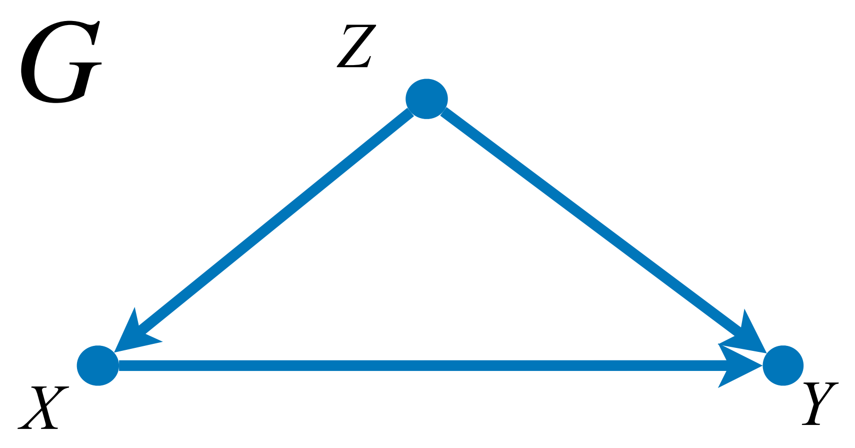
With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

Task of Identification

1 query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$$

2 graph



3 probability

$$P(\mathbf{V})$$

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

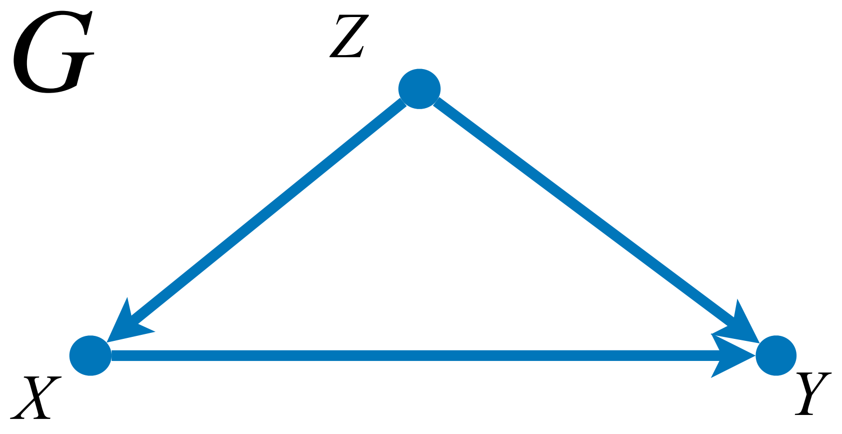
Task of Identification

1 query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$$

2 graph

G



3 probability

$$P(\mathbf{V})$$

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

ID (G, P, Q)

Task of Identification

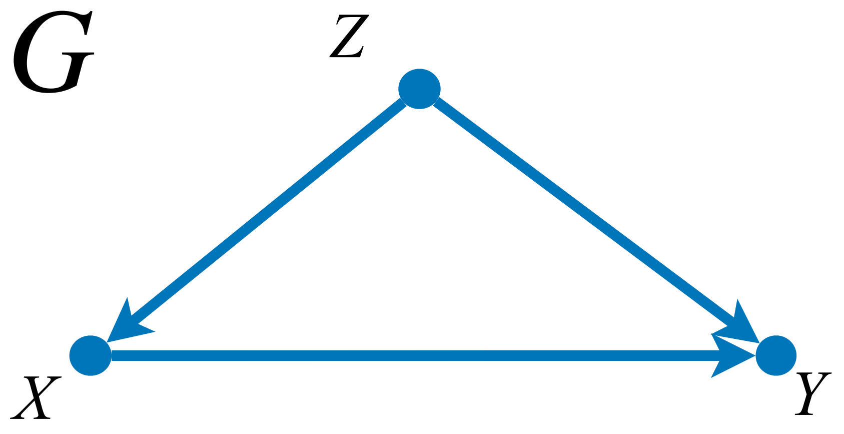
1 query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$$

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

2 graph

G



3 probability

$$P(\mathbf{V})$$

ID $\left(G, P, Q\right)$ → solution

yes / no

Task of Identification

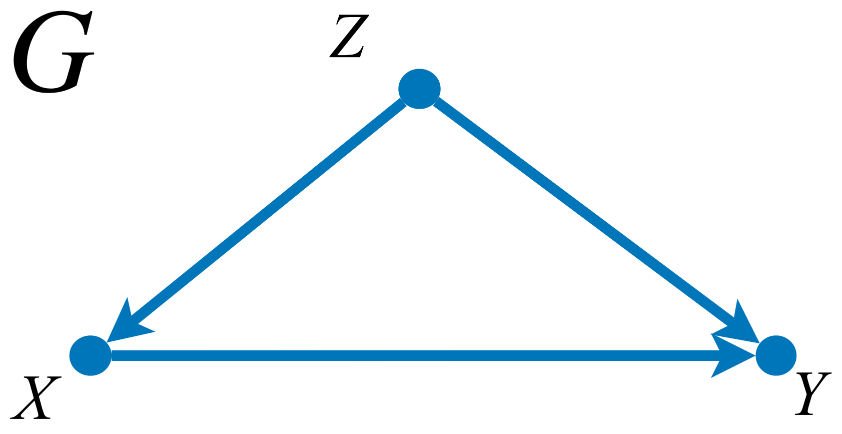
1. query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \mid do(\mathbf{x}))$$

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

2. graph

G



3. probability

$$P(\mathbf{V})$$

$$\boxed{\text{ID}\,(G, P, Q)}$$

solution

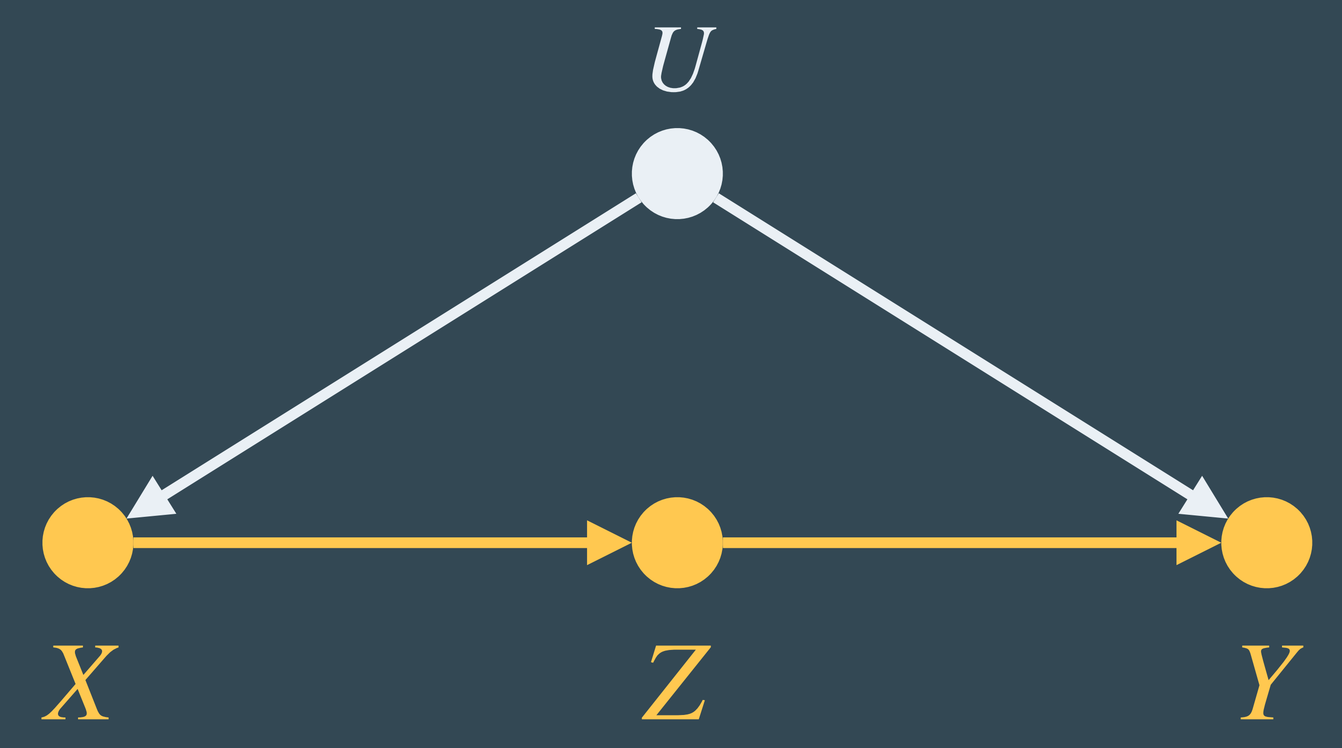
yes / no

Causal Functional

$$P_{\mathbf{x}}(\mathbf{y}) = f(P)$$

3. Causal effect estimation

Front-door — How to estimate?



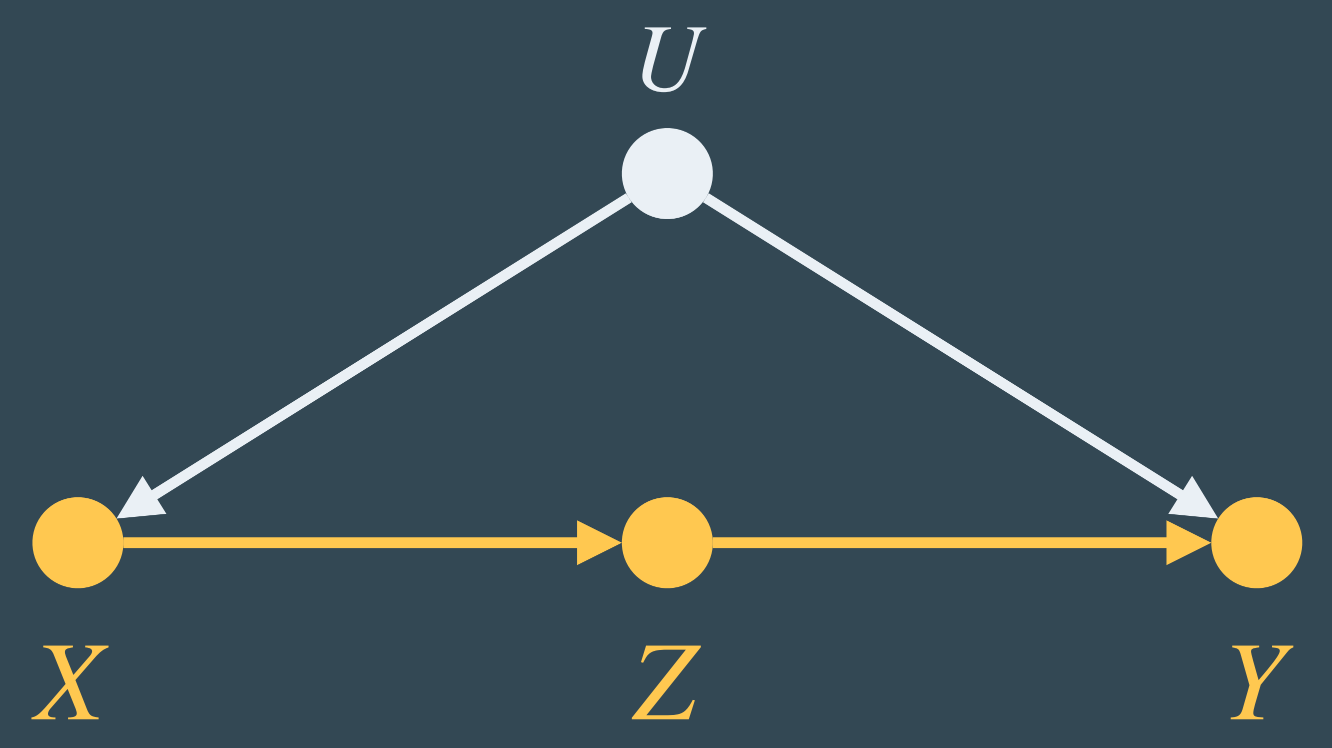
U: Genetic factor (latent)

X: Smoke

Z: Tar in the smoke

Y: Lung disease

Front-door — How to estimate?



U: Genetic factor (latent)

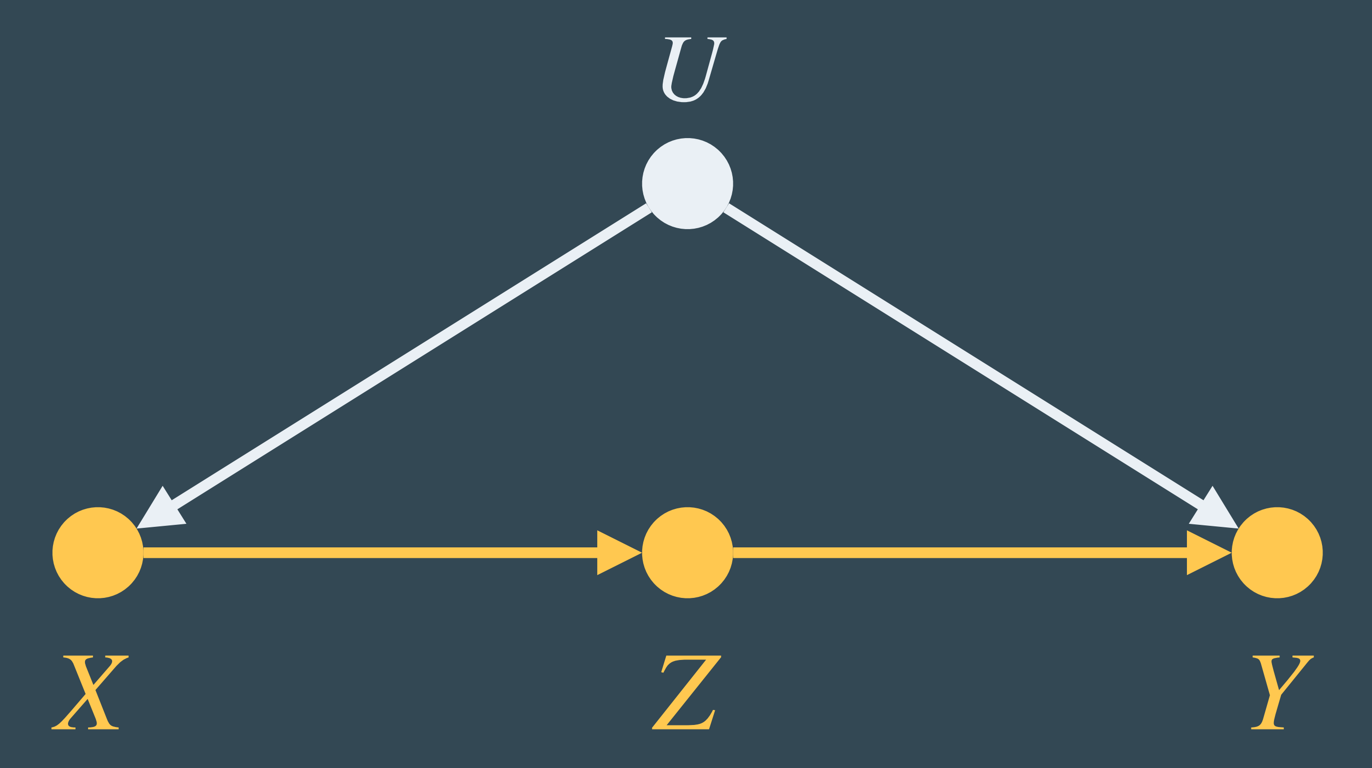
X: Smoke

Z: Tar in the smoke

Y: Lung disease

$$\mathbb{E}[Y \mid do(x)] = \sum_z P(z \mid x) \sum_{x'} \mathbb{E}[Y \mid x', z] P(x').$$

Front-door — How to estimate?



U: Genetic factor (latent)

X: Smoke

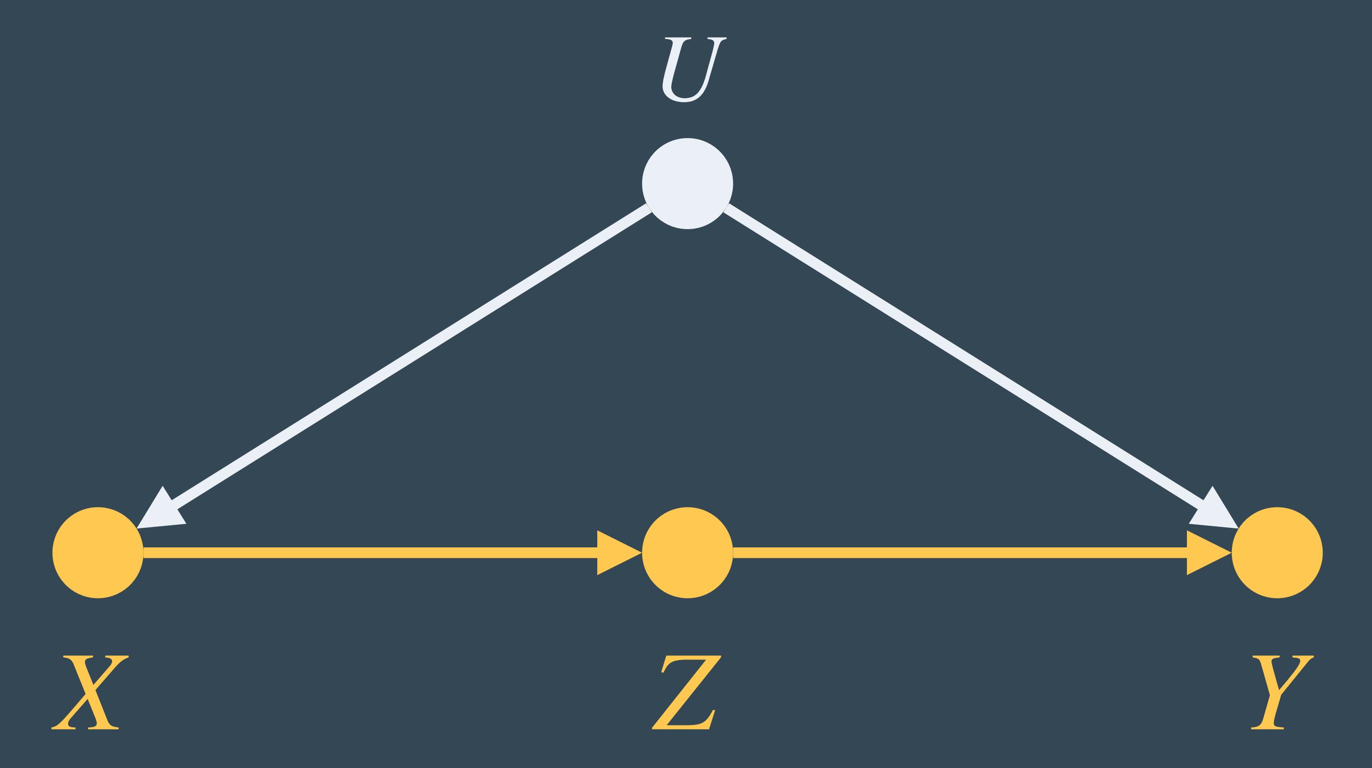
Z: Tar in the smoke

Y: Lung disease

$$\mathbb{E}[Y \mid do(x)] = \sum_z P(z \mid x) \sum_{x'} \mathbb{E}[Y \mid x', z] P(x').$$

Instead of the L1 distribution $P(x, y, z)$, we are only given finite samples $D = \{X_i, Y_i, Z_i\}_{i=1}^N$ from P.

Front-door — How to estimate?



U: Genetic factor (latent)

X: Smoke

Z: Tar in the smoke

Y: Lung disease

$$\mathbb{E}[Y \mid do(x)] = \sum_z P(z \mid x) \sum_{x'} \mathbb{E}[Y \mid x', z] P(x').$$

Instead of the L1 distribution $P(x, y, z)$, we are only given finite samples $D = \{X_i, Y_i, Z_i\}_{i=1}^N$ from P.

We must estimate the ID quantity ("Causal functional") from the dataset D.

Task of Estimation

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

Task of Estimation

$\text{ID}\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$$P(y \mid do(x)) = f(P)$$

Task of Estimation

$\boxed{\text{ID}\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)}$

$\downarrow$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$$P(y \mid do(x)) = f(P)$$

2 Data

$$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$$

Task of Estimation

ID $\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$$P(y \mid do(x)) = f(P)$$

Estimation Engine

EST$(f(P), D)$

2 Data

$$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$$

Task of Estimation

ID $\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$$P(y \mid do(x)) = f(P)$$

Estimation Engine

EST$(f(P), D)$

Estimand

$$g(P) = f(P)$$

2 Data

$$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$$

Task of Estimation

ID $\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a
reliable estimate T_N for

Alternative representation for amenable estimation

1 Causal Functional

$P(y \mid do(x)) = f(P)$

Estimation Engine

EST$(f(P), D)$

Estimand

$g(P) = f(P)$

2 Data

$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$

Task of Estimation

ID $\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$$P(y \mid do(x)) = f(P)$$

2 Data

$$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$$

Estimation Engine

EST$(f(P), D)$

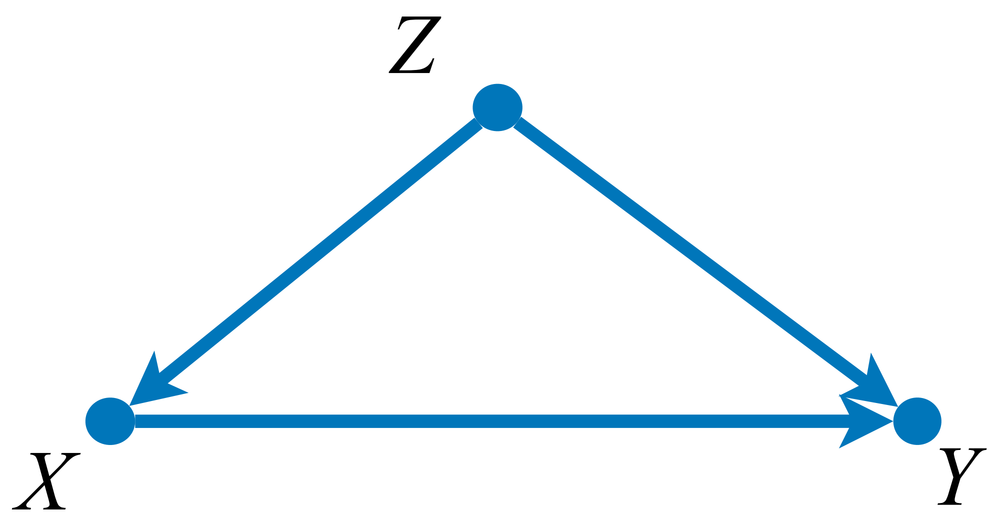
Estimand

$$g(P) = f(P)$$

Estimator

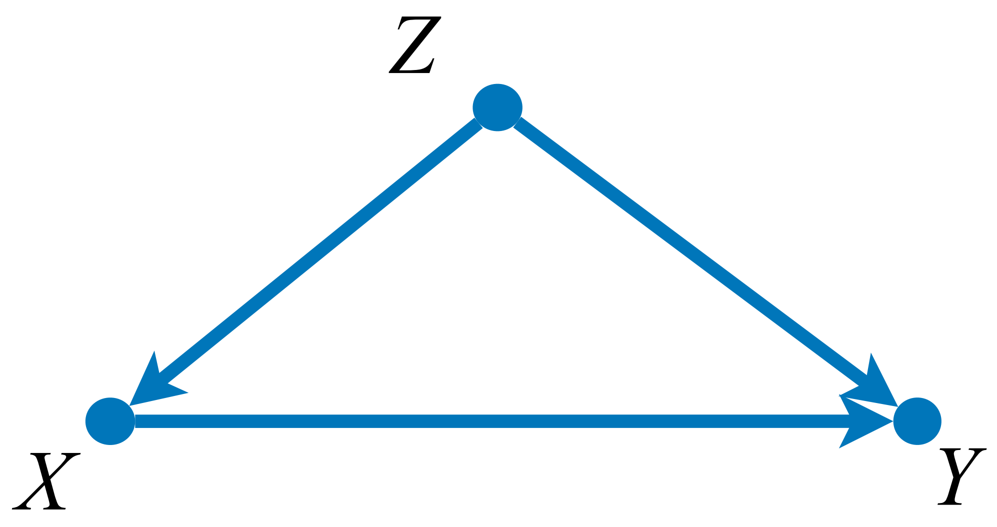
$$T_N = \widehat{g(P)}$$

Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

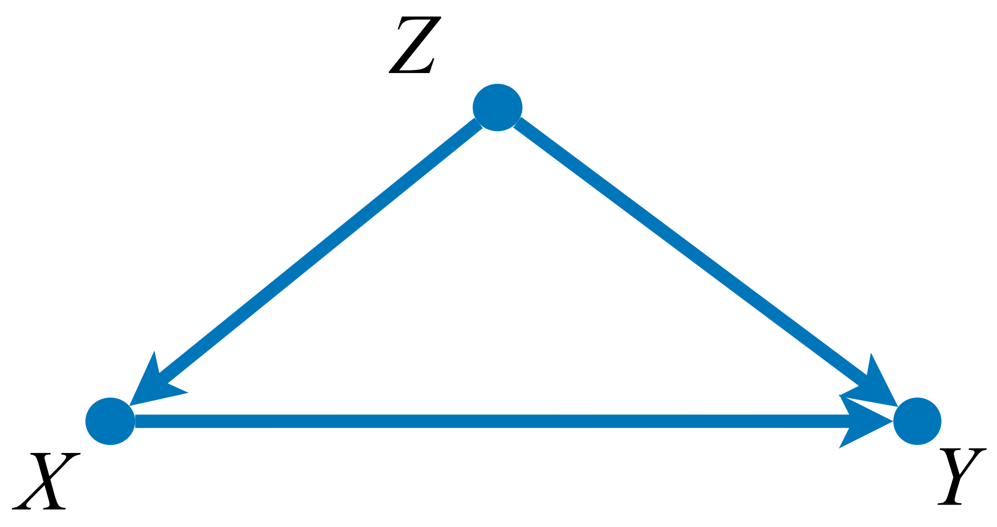
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

$$\sum_z P(y \mid x, z)P(z) = \sum_z P(y \mid x, z)P(x \mid z)P(z)\frac{1}{P(x \mid z)}$$

Classic BD estimator:
1. Inverse probability weighting (IPW)

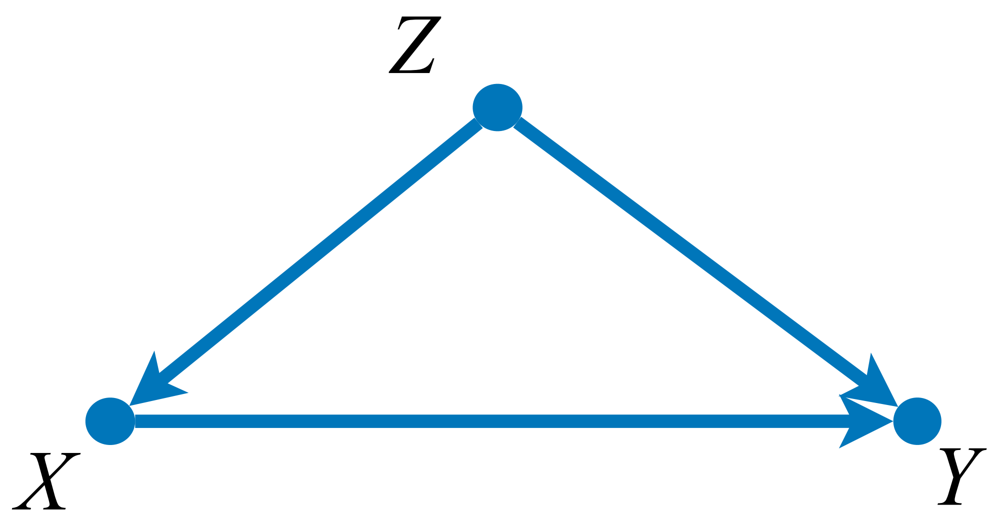


$$P_x(y) = \sum_z P(y|x,z)P(z)$$

$$\sum_z P(y|x,z)P(z) = \sum_z \underline{P(y|x,z)P(x|z)P(z)}\frac{1}{P(x|z)}$$

$$= \sum_z P(z,x,y)\frac{1}{P(x|z)}$$

Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

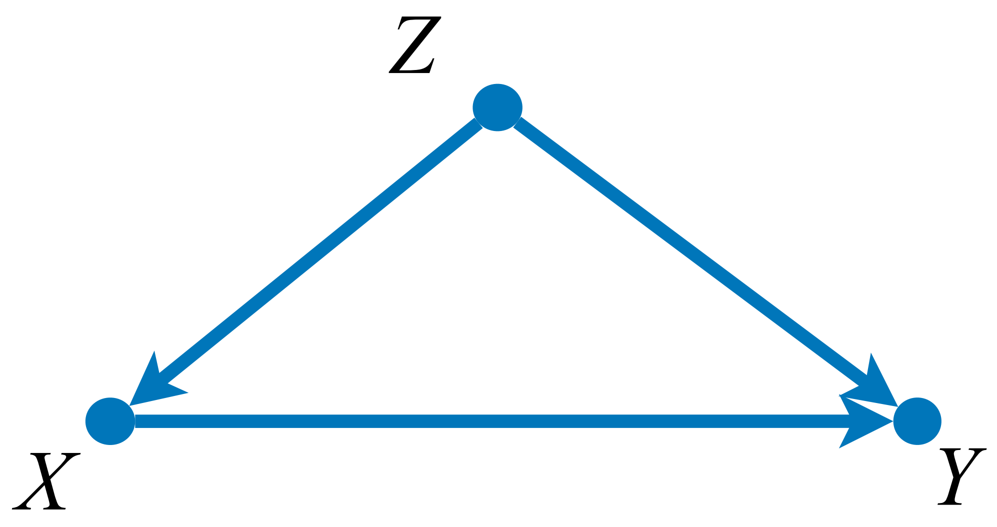
$$\sum_z P(y \mid x, z) P(z) = \sum_z \underline{P(y \mid x, z) P(x \mid z) P(z)} \frac{1}{P(x \mid z)}$$

$$= \sum_z P(z, x, y) \frac{1}{P(x \mid z)}$$

$I_x(x') = 1$ if $x = x'$; otherwise 0.

$$= \sum_{z, x', y'} P(z, x', y') \frac{I_x(x')}{P(x \mid z)} I_y(y')$$

Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

$$\sum_z P(y \mid x, z) P(z) = \sum_z P(y \mid x, z) {\color{blue}P(x \mid z)} P(z) \frac{1}{{\color{blue}P(x \mid z)}}$$

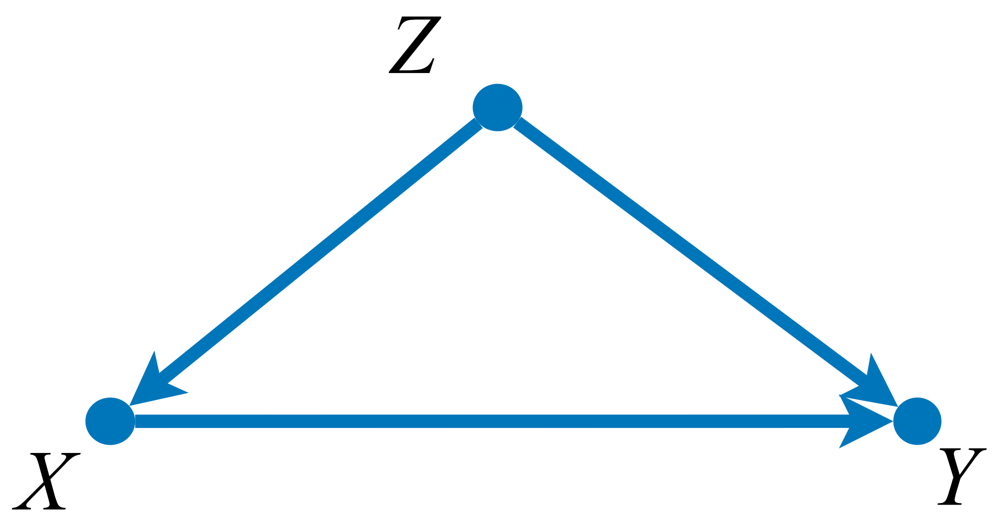
$$= \sum_z P(z, x, y) \frac{1}{P(x \mid z)}$$

$I_x(x') = 1$ if $x = x'$;
otherwise 0.

$$= \sum_{z, {\color{blue}x', y'}} P(z, {\color{blue}x', y'}) \frac{{\color{blue}I_x(x')}}{P(x \mid z)} I_y(y')$$

$$= \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y) \right]$$

Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

$$\sum_z P(y \mid x, z)P(z) = \sum_z \underline{P(y \mid x, z)P(x \mid z)P(z)}\frac{1}{P(x \mid z)}$$

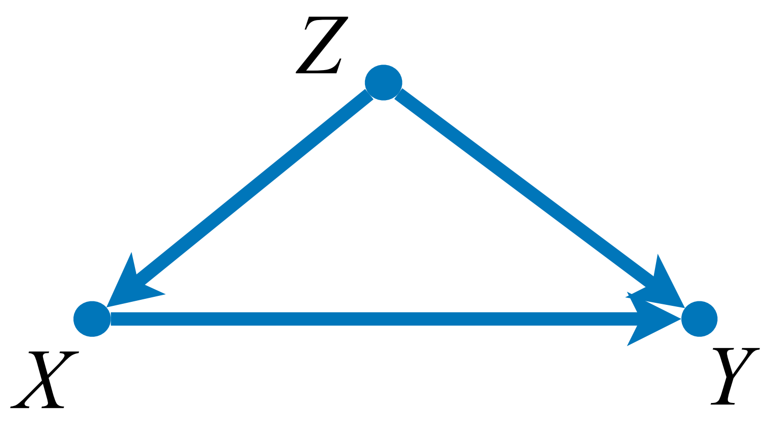
$$= \sum_z P(z, x, y)\frac{1}{P(x \mid z)}$$

$I_x(x') = 1$ if $x = x'$;
otherwise 0.

$$= \sum_{z,x',y'} P(z, x', y')\frac{I_x(x')}{P(x \mid z)}I_y(y')$$

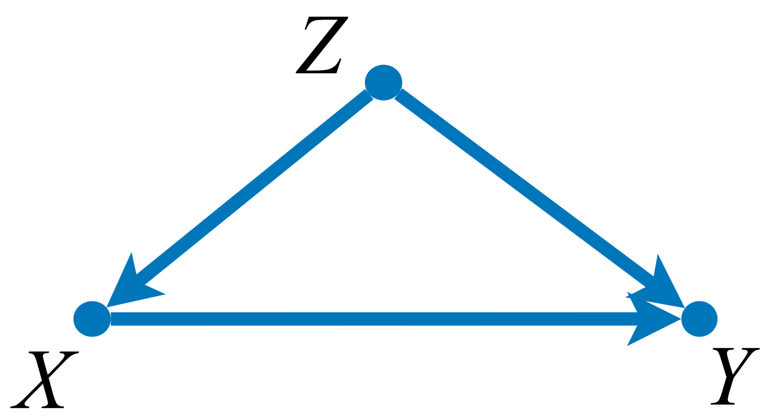
$$= \mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right] = g(P)$$

Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\mathsf{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

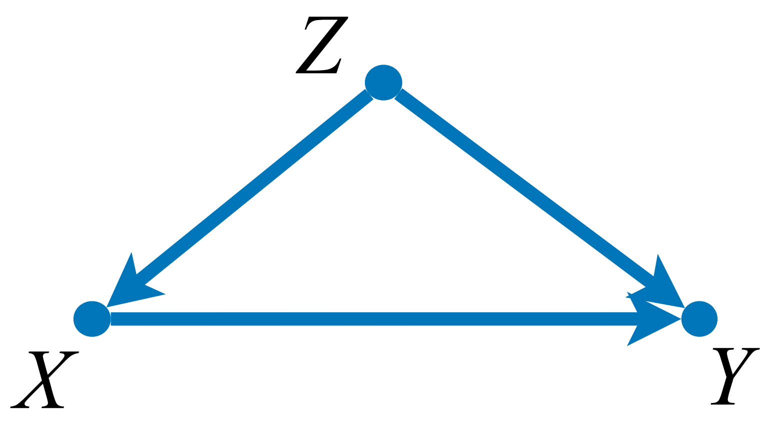
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\textsf{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$

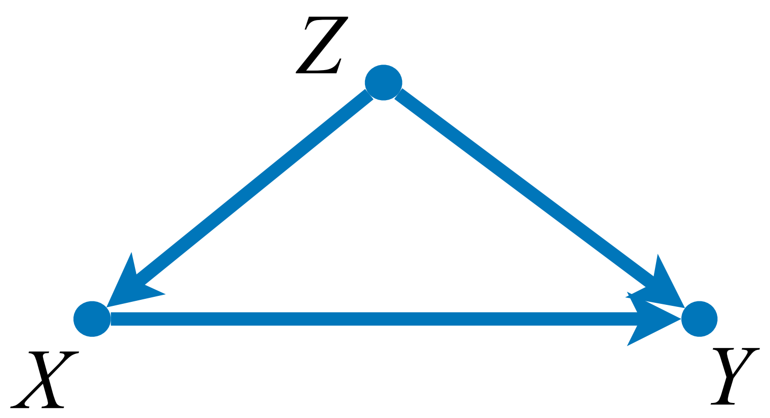
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y)\right]$

Classic BD estimator:
1. Inverse probability weighting (IPW)

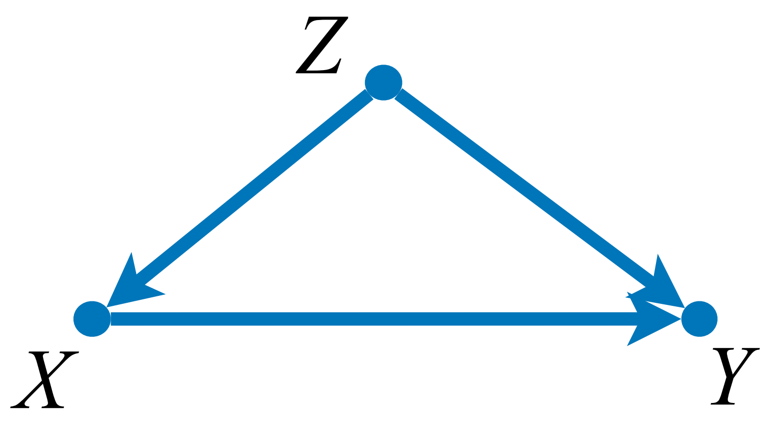


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \Big[\frac{I_x(}{P(X}$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y) \right]$

$\mathbb{E}_D[f(\mathbf{W})] \equiv \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{W}_{(i)})$, an empirical expectation of $f(\mathbf{W})$ using samples.

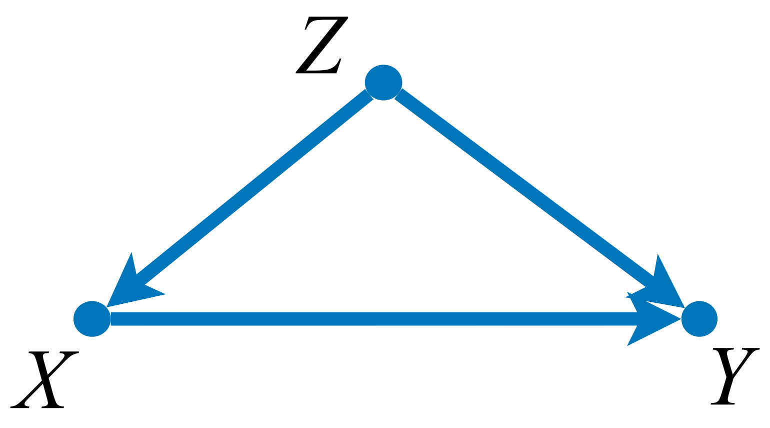
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\mathsf{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y)\right]$

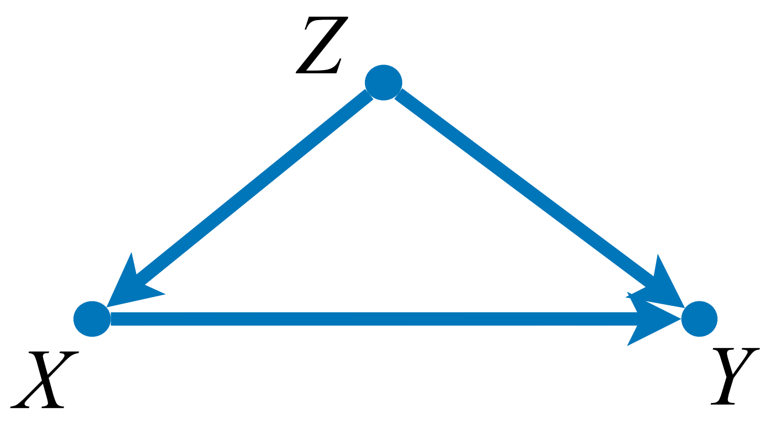
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\mathrm{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y) \right]$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y) \right]$
For consistent estimation	$\hat{P}(x \mid z)$ converges to $P(x \mid z)$.

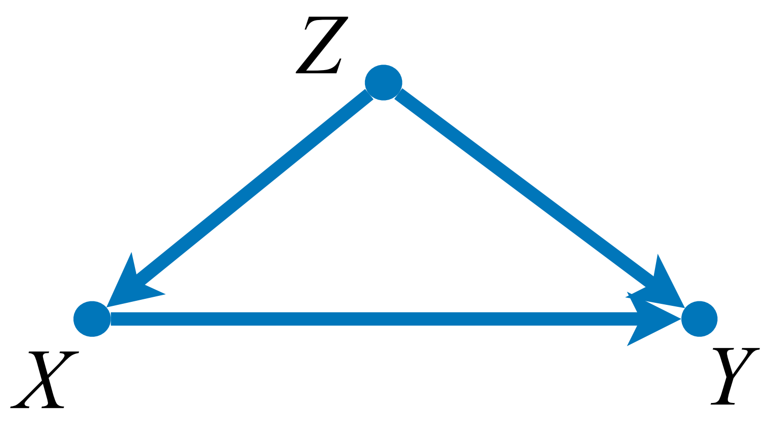
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y)\right]$
For consistent estimation	$\hat{P}(x \mid z)$ converges to $P(x \mid z)$.
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

Classic BD estimator:
1. Inverse probability weighting (IPW)

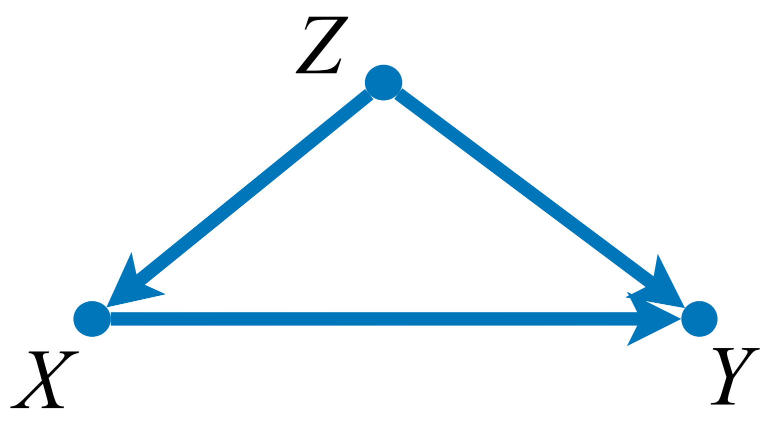


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	
For consistent estimation	
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

- A function class s.t. complexities of functions are restricted. (e.g., linear/logistic regression, smooth parametric functions).

Classic BD estimator:
1. Inverse probability weighting (IPW)

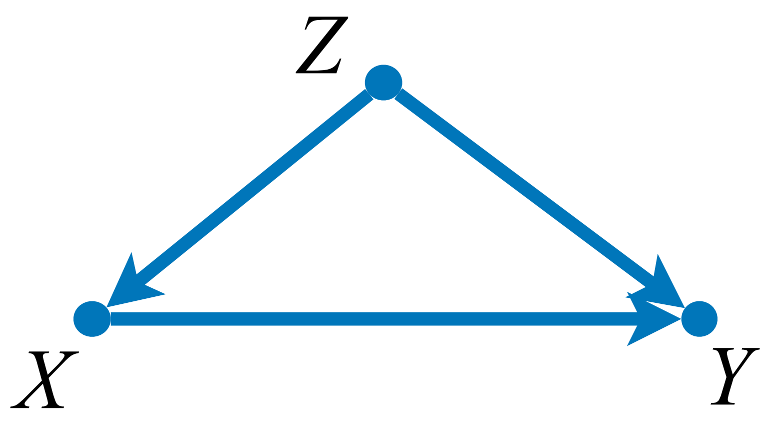


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y) \right]$
Estimator (T_N)	
For consistent estimation	
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

- A function class s.t. complexities of functions are restricted. (e.g., linear/logistic regression, smooth parametric functions).
- It's unclear modern flexible/complicated ML methods (e.g., neural networks) are in this class.

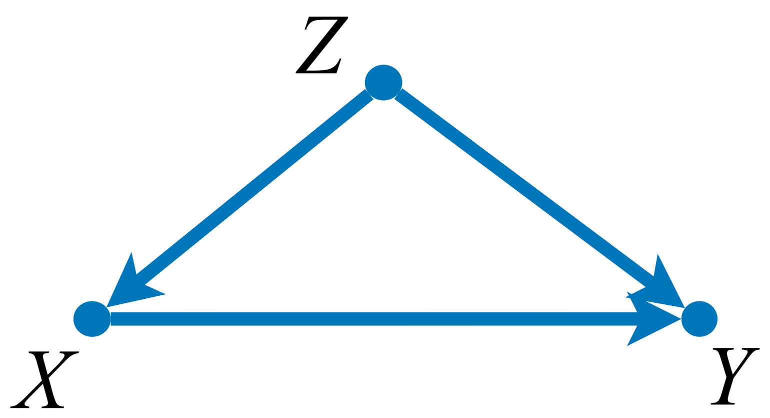
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

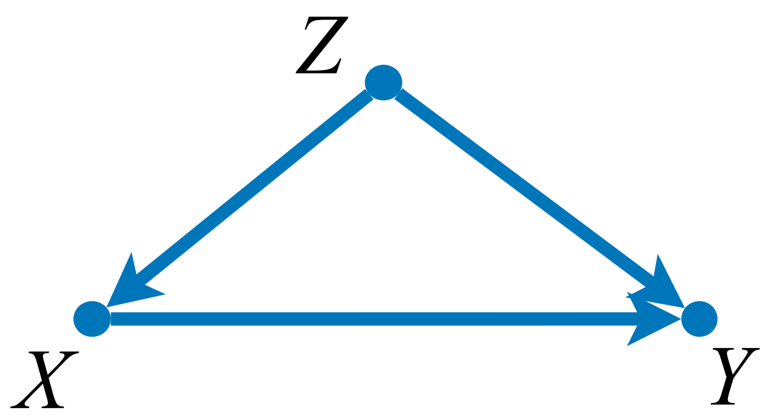
Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y)\right]$
For consistent estimation	$\hat{P}(x \mid z)$ converges to $P(x \mid z)$.
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

Classic BD estimator: 2. Outcome-regression (OR)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Classic BD estimator:
2. Outcome-regression (OR)

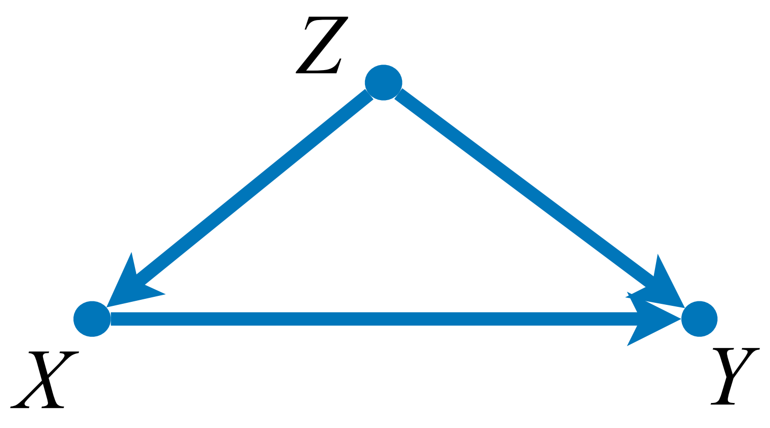


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P[P(y \mid x, Z)]$

Expectation over Z.
$= \sum_z P(y \mid x, z)P(z)$

Classic BD estimator:
2. Outcome-regression (OR)

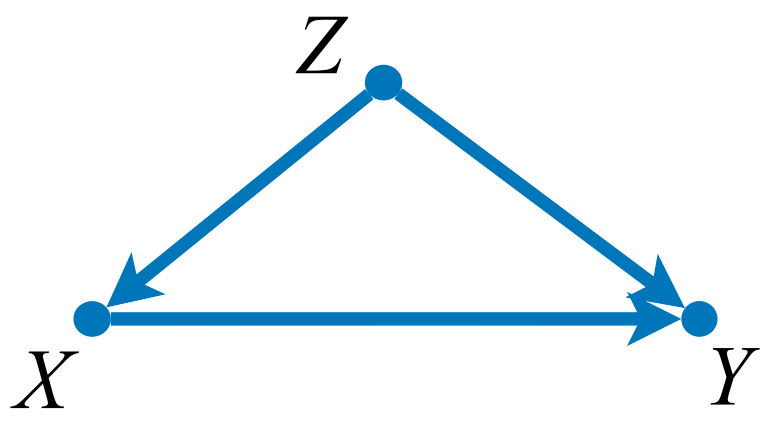


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P[P(y \mid x, Z)]$
Estimator (T_N)	$\mathbb{E}_D\left[\hat{P}(y \mid x, Z)\right]$

Expectation over Z.
$= \sum_z P(y \mid x, z) P(z)$

Classic BD estimator: 2. Outcome-regression (OR)

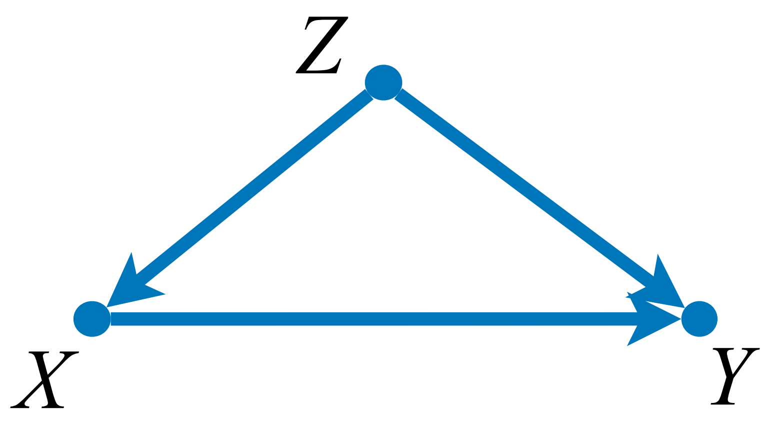


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Expectation over Z.
$= \sum_z P(y \mid x, z)P(z)$

Estimand ($g(P)$)	$\mathbb{E}_P[P(y \mid x, Z)]$
Estimator (T_N)	$\mathbb{E}_D\left[\hat{P}(y \mid x, Z)\right]$
For correct estimation	$\hat{P}(y \mid x, z)$ converges to $P(y \mid x, z)$.

Classic BD estimator: 2. Outcome-regression (OR)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Expectation over Z.
$= \sum_z P(y \mid x, z)P(z)$

Estimand ($g(P)$)	$\mathbb{E}_P[P(y \mid x, Z)]$
Estimator (T_N)	$\mathbb{E}_D\left[\hat{P}(y \mid x, Z)\right]$
For correct estimation	$\hat{P}(y \mid x, z)$ converges to $P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(y \mid x, z) \to P(y \mid x, z)$ fast; and (2) $\{\hat{P}(y \mid x, z), P(y \mid x, z)\}$ in Donsker Function class.

Comparison- Classic BD estimators

	IPW	OR
Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X\|Z)} \cdot I_y(Y)\right]$	$\mathbb{E}_P\left[P(y\|x,Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X\|Z)} \cdot I_y(Y)\right]$	$\mathbb{E}_D\left[\hat{P}(y\|x,Z)\right]$
For correct estimation	$\hat{P}(x\|z) \rightarrow P(x\|z)$	$\hat{P}(y\|x,z) \rightarrow P(y\|x,z)$
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x\|z) \rightarrow P(x\|z)$ fast; (2) $\{\hat{P}(x\|z), P(x\|z)\}$ in Donsker class.	(1) $\hat{P}(y\|x,z) \rightarrow P(y\|x,z)$ fast; (2) $\{\hat{P}(y\|x,z), P(y\|x,z)\}$ in Donsker class.

Comparison- Classic BD estimators

	IPW	**OR**
Estimand ($g(P)$)	$\mathbb{E}_P\left[\dfrac{I_x(X)}{P(X\mid Z)}\cdot I_y(Y)\right]$	$\mathbb{E}_P\left[P(y\mid x,Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\dfrac{I_x(X)}{\hat{P}(X\mid Z)}\cdot I_y(Y)\right]$	$\mathbb{E}_D\left[\hat{P}(y\mid x,Z)\right]$
For correct estimation	**(Single chance of being correct)** If $\hat{P}(x\mid z)$ ($\hat{P}(y\mid x,z)$) is misspecified, then T_N fails to converge	
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x\mid z)\rightarrow P(x\mid z)$ fast; (2) $\{\hat{P}(x\mid z), P(x\mid z)\}$ in Donsker class.	(1) $\hat{P}(y\mid x,z)\rightarrow P(y\mid x,z)$ fast; (2) $\{\hat{P}(y\mid x,z), P(y\mid x,z)\}$ in Donsker class.

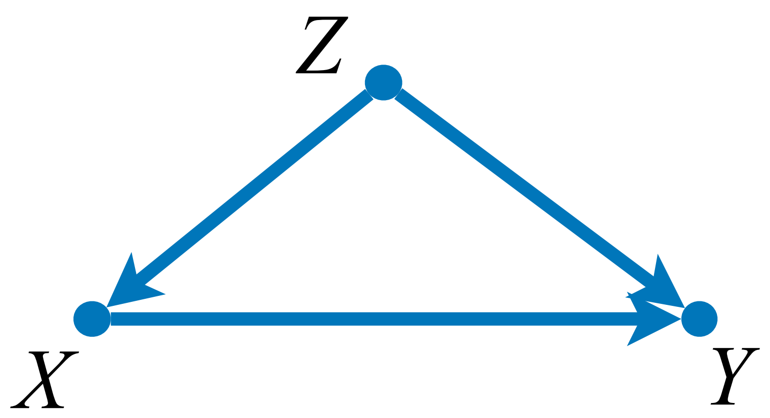
Comparison- Classic BD estimators

IPW

OR

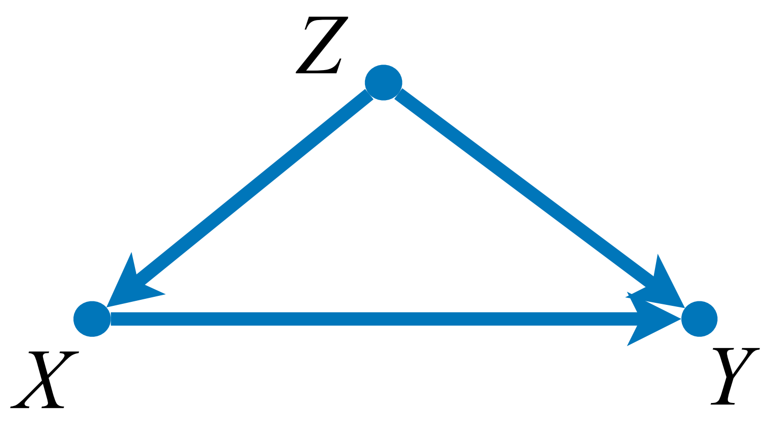
	IPW	OR
convergence	(2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker class.	(2) $\{\hat{P}(y \mid x, z), P(y \mid x, z)\}$ in Donsker class.

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

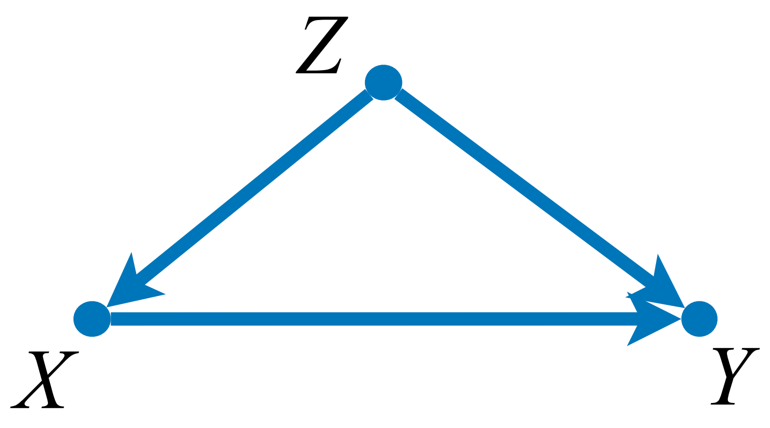
Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

- Consider the following estimand: $$g(P) \equiv \mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$$

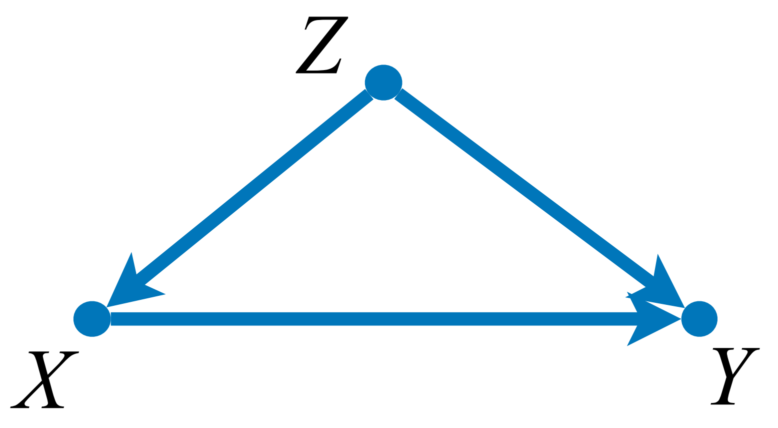
Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $$g(P) \equiv \mathbb{E}_P \left[\overset{\textbf{IPW}}{\frac{I_x(X)}{P(X \mid Z)}(I_y(Y)} - P(y \mid X, Z)) + \overset{\textbf{OR}}{P(y \mid x, Z)} \right]$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y|x,z)P(z)$$

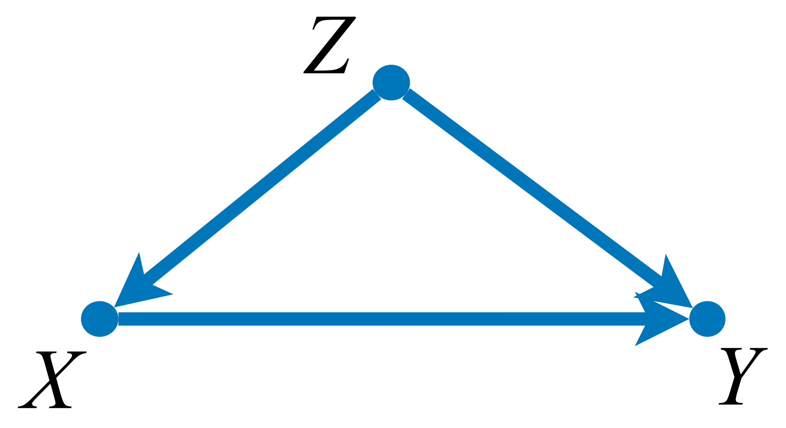
- Consider the following estimand:

IPW　　　　OR

$$g(P) \equiv \mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z)\right]$$

- This $g(P)$ is a valid estimand (i.e., $g(P) = \sum_z P(y|x,z)P(z)$) even when $P(x|z)$ or $P(y|x,z)$ are misspecified.

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y|x,z)P(z)$$

- Consider the following estimand:

$$g(P) \equiv \mathbb{E}_P\left[\underbrace{\frac{I_x(X)}{P(X|Z)}(I_y(Y)}_{\text{IPW}} - P(y|X,Z)) + \underbrace{P(y|x,Z)}_{\text{OR}}\right]$$

If $P(X|Z)$ is misspecified to $\tilde{P}(X|Z)$:

Augmented IPW (IPW + OR)

$$
\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & \\
X \arrow[rr] & & Y
\end{tikzcd}
$$

$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

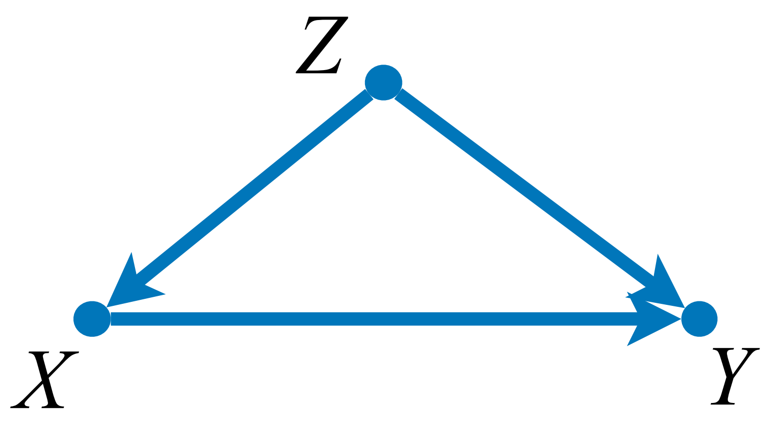
- Consider the following estimand:

$$g(P) \equiv \mathbb{E}_P \left[\underbrace{\frac{I_x(X)}{P(X \mid Z)}}_{\text{IPW}} (I_y(Y) - P(y \mid X, Z)) + \underbrace{P(y \mid x, Z)}_{\text{OR}} \right]$$

If $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand:

$$g(P) \equiv \mathbb{E}_P \left[\underbrace{\frac{I_x(X)}{P(X \mid Z)}(I_y(Y)}_{\text{IPW}} - P(y \mid X, Z)) + \underbrace{P(y \mid x, Z)}_{\text{OR}} \right]$$

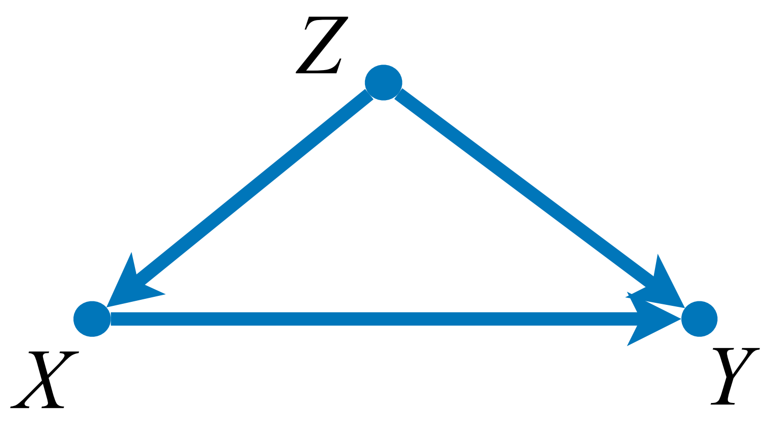
If $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$$

(Law of total expectation): Taking expectation to Y and (X, Z) in sequence.

$$= \mathbb{E}_{P(X,Z)} \left\{ \mathbb{E}_{P(Y \mid X,Z)} \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \middle| X, Z \right] \right\}$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand:

$$g(P) \equiv \mathbb{E}_P \left[\underbrace{\frac{I_x(X)}{P(X \mid Z)} (I_y(Y)}_{\text{IPW}} - P(y \mid X, Z)) + \underbrace{P(y \mid x, Z)}_{\text{OR}} \right]$$

If $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$$

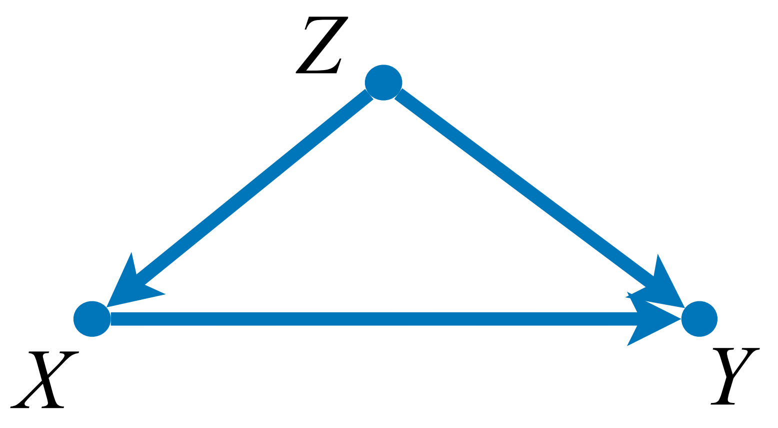
$$= \mathbb{E}_{P(X,Z)} \left\{ \mathbb{E}_{P(Y \mid X,Z)} \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \middle| X, Z \right] \right\}$$

$$= \mathbb{E}_{P(X,Z)} \left\{ \frac{I_x(X)}{\tilde{P}(X \mid Z)} (\cancel{P(y \mid X, Z) - P(y \mid X, Z)}) + P(y \mid x, Z) \right\}$$

(Law of total expectation):
Taking expectation to Y and (X, Z) in sequence.

Since
$\mathbb{E}_{P(Y \mid X,Z)}[I_y(Y) \mid X, Z] = P(y \mid X, Z)$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand:

IPW OR

$$g(P) \equiv \mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$$

If $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$:

$$\mathbb{E}_P\left[\frac{I_x(X)}{\tilde{P}(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$$

$$= \mathbb{E}_{P(X,Z)}\left\{\mathbb{E}_{P(Y \mid X,Z)}\left[\frac{I_x(X)}{\tilde{P}(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \,\middle|\, X, Z\right]\right\}$$

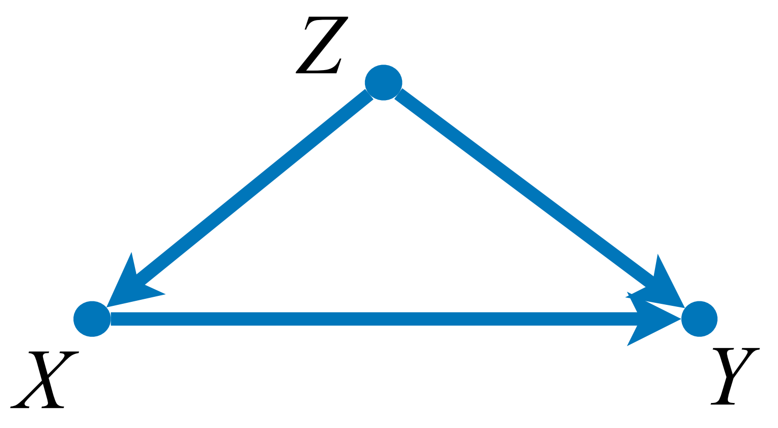
$$= \mathbb{E}_{P(X,Z)}\left\{\frac{I_x(X)}{\tilde{P}(X \mid Z)}(\cancel{P(y \mid X, Z) - P(y \mid X, Z)}) + P(y \mid x, Z)\right\}$$

$$= \mathbb{E}_P[P(y \mid x, Z)] = \sum_z P(y \mid x, z) P(z) = P_x(y)$$

(Law of total expectation):
Taking expectation to Y and (X, Z) in sequence.

Since
$\mathbb{E}_{P(Y \mid X,Z)}[I_y(Y) \mid X, Z] = P(y \mid X, Z)$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

- Consider the following estimand:

$$g(P) \equiv \mathbb{E}_P\left[\underbrace{\frac{I_x(X)}{P(X \mid Z)}(I_y(Y)}_{\text{IPW}} - P(y \mid X, Z)) + \underbrace{P(y \mid x, Z)}_{\text{OR}}\right]$$

If $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$:

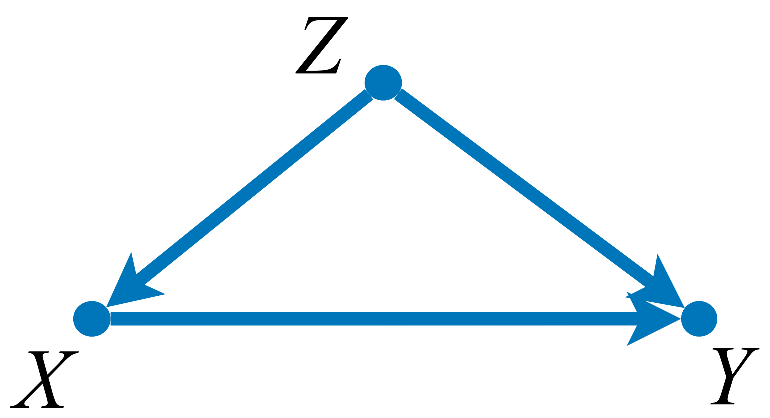
$$\mathbb{E}_P\left[\frac{I_x(X)}{\tilde{P}(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$$

(Law of total expectation):
Taking expectation to Y and
(X, Z) in sequence.

Takeaway: $g(P) = P_x(y)$ even if $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$

$$= \mathbb{E}_P[P(y \mid x, Z)] = \sum_z P(y \mid x, z)P(z) = P_x(y)$$

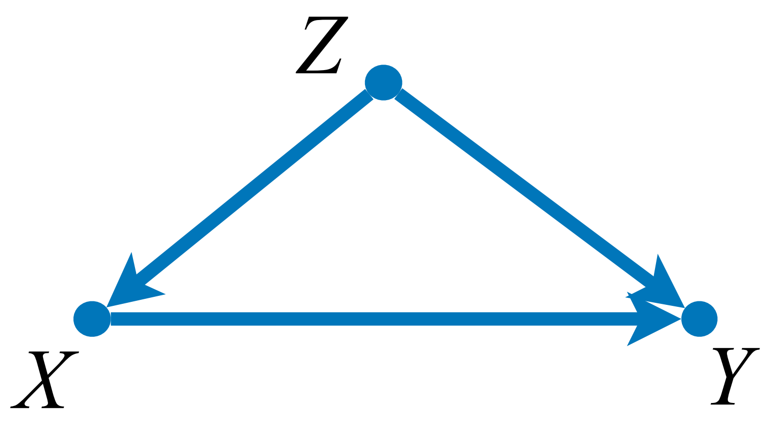
Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

Augmented IPW (IPW + OR)

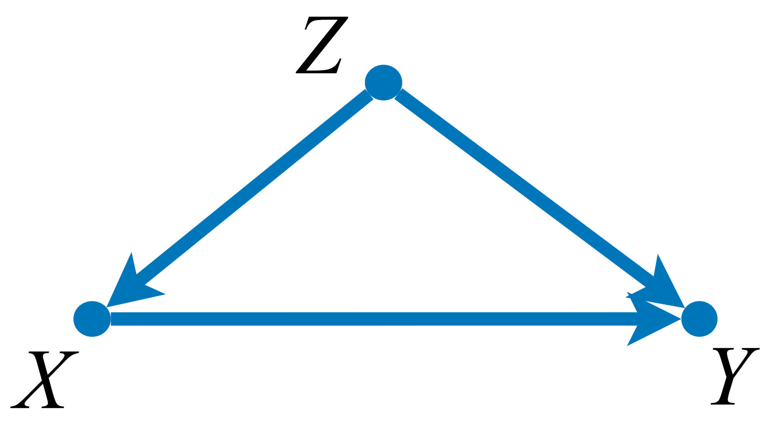


$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

Augmented IPW (IPW + OR)



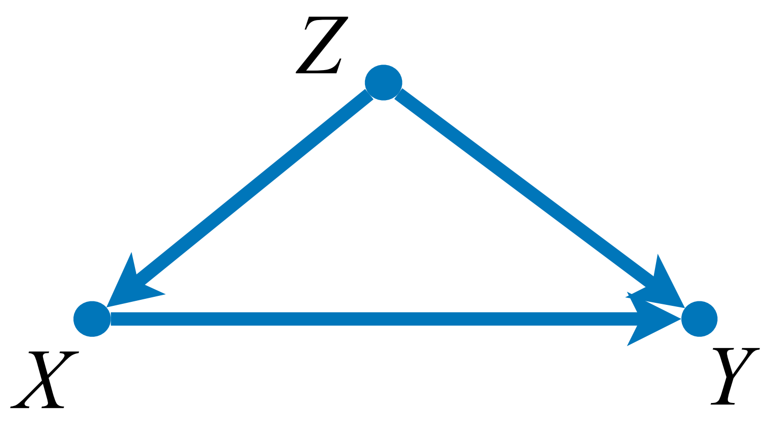
$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right]$$

Augmented IPW (IPW + OR)



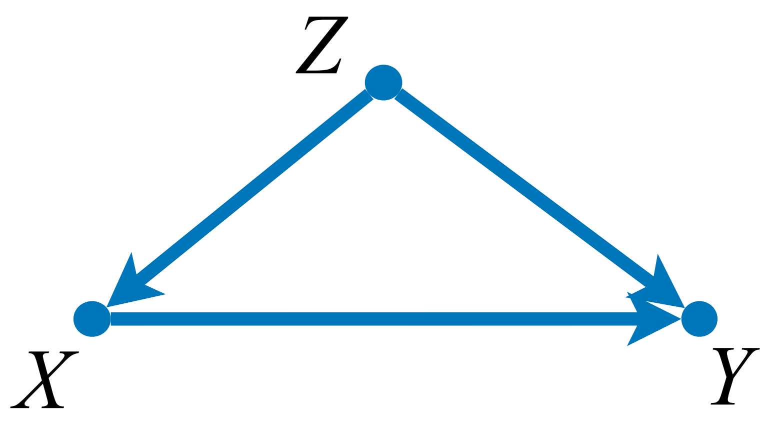
$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_{P(X,Z)} \left[\mathbb{E}_{P(Y \mid X,Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \,\middle|\, X, Z \right\} \right]$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

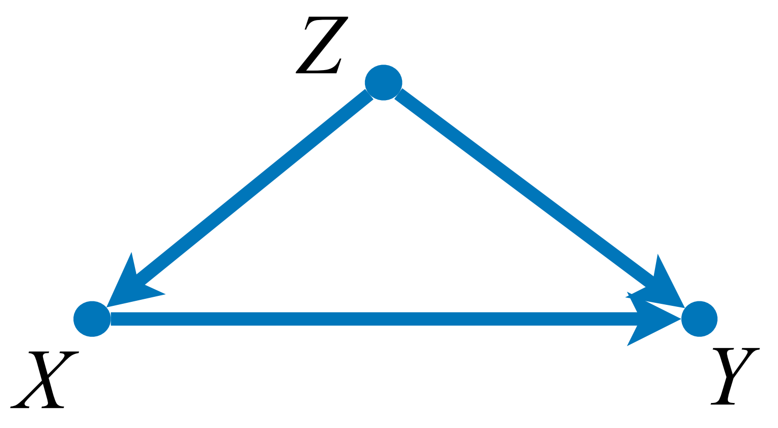
- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_{P(X,Z)} \left[\mathbb{E}_{P(Y \mid X, Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \middle| X, Z \right\} \right]$$

$$= \mathbb{E}_{P(X,Z)} \left[\frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right]$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

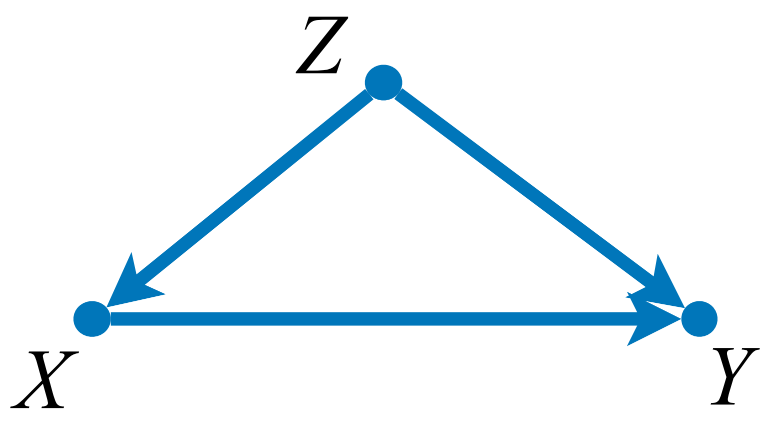
$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right]$$

$$= \mathbb{E}_{P(X,Z)} \left[\mathbb{E}_{P(Y \mid X,Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \middle| X, Z \right\} \right]$$

$$= \mathbb{E}_{P(X,Z)} \left[\frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right]$$

$$= \mathbb{E}_{P(Z)} \left[\mathbb{E}_{P(X \mid Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \middle| Z \right\} \right]$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \,|\, x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \,|\, Z)} (I_y(Y) - P(y \,|\, X, Z)) + P(y \,|\, x, Z) \right]$

If $P(y \,|\, X, Z)$ is misspecified to $\tilde{P}(y \,|\, X, Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \,|\, Z)} (I_y(Y) - \tilde{P}(y \,|\, X, Z)) + \tilde{P}(y \,|\, x, Z) \right]$$

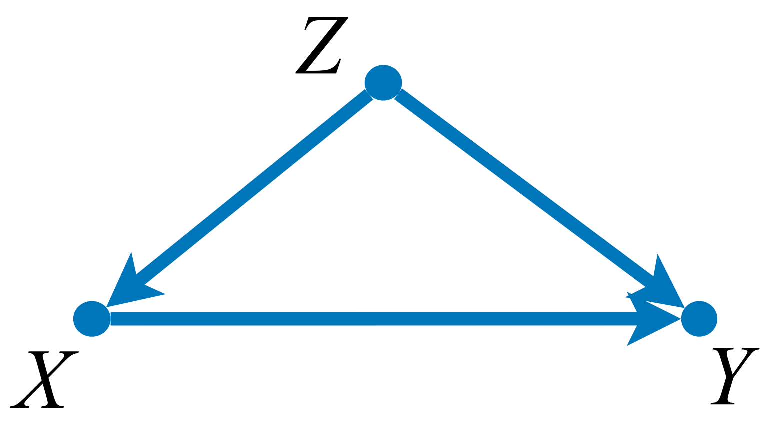
$$= \mathbb{E}_{P(X,Z)} \left[\mathbb{E}_{P(Y|X,Z)} \left\{ \frac{I_x(X)}{P(X \,|\, Z)} (I_y(Y) - \tilde{P}(y \,|\, X, Z)) + \tilde{P}(y \,|\, x, Z) \middle| X, Z \right\} \right]$$

$$= \mathbb{E}_{P(X,Z)} \left[\frac{I_x(X)}{P(X \,|\, Z)} (P(y \,|\, X, Z) - \tilde{P}(y \,|\, X, Z)) + \tilde{P}(y \,|\, x, Z) \right]$$

$$= \mathbb{E}_{P(Z)} \left[\mathbb{E}_{P(X|Z)} \left\{ \frac{I_x(X)}{P(X \,|\, Z)} (P(y \,|\, X, Z) - \tilde{P}(y \,|\, X, Z)) + \tilde{P}(y \,|\, x, Z) \middle| Z \right\} \right]$$

$$= \mathbb{E}_{P(Z)} \left[\cancel{\frac{P(x \,|\, Z)}{P(x \,|\, Z)}} (P(y \,|\, x, Z) - \tilde{P}(y \,|\, x, Z)) + \tilde{P}(y \,|\, x, Z) \right]$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

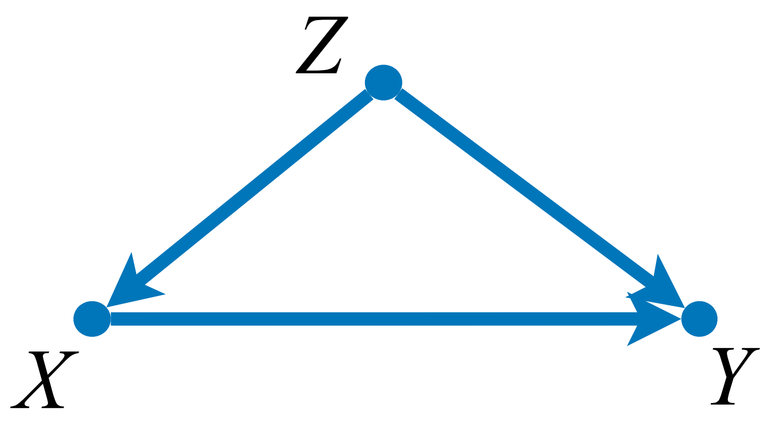
If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_{P(X,Z)} \left[\mathbb{E}_{P(Y \mid X,Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \middle| X, Z \right\} \right]$$

$$= \mathbb{E}_{P(X,Z)} \left[\frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_{P(Z)} \left[\mathbb{E}_{P(X \mid Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \middle| Z \right\} \right]$$

$$= \mathbb{E}_{P(Z)} \left[\frac{\cancel{P(x \mid Z)}}{\cancel{P(x \mid Z)}} (P(y \mid x, Z) - \tilde{P}(y \mid x, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_P[P(y \mid x, Z)] = \sum_z P(y \mid x, z) P(z) = P_x(y)$$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y|x,z)P(z)$$

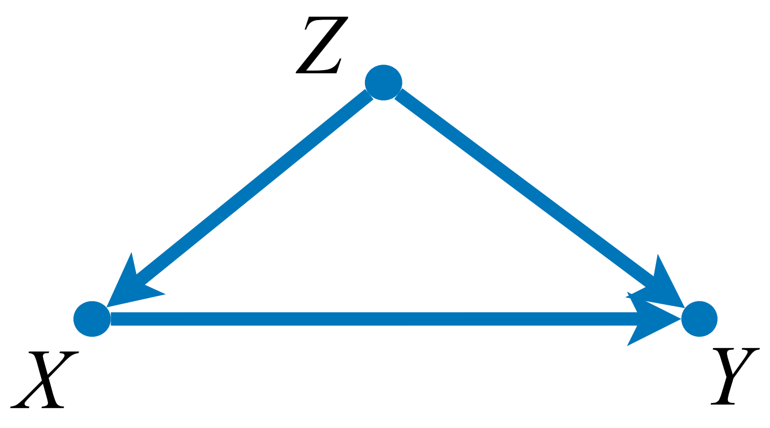
- Consider the following estimand: $g(P) \equiv \mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z)\right]$

If $P(y|X,Z)$ is misspecified to $\tilde{P}(y|X,Z)$:

Takeaways:

- $g(P) = P_x(y)$ even if $P(y|X,Z)$ is misspecified to $\tilde{P}(y,|X,Z)$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

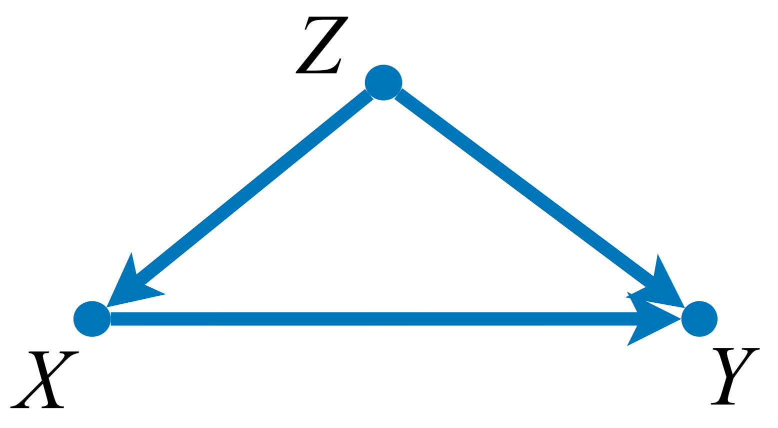
- Consider the following estimand: $g(P) \equiv \mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$

If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

Takeaways:

- $g(P) = P_x(y)$ even if $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y, \mid X, Z)$
- $g(P) = P_x(y)$ even if $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$

Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

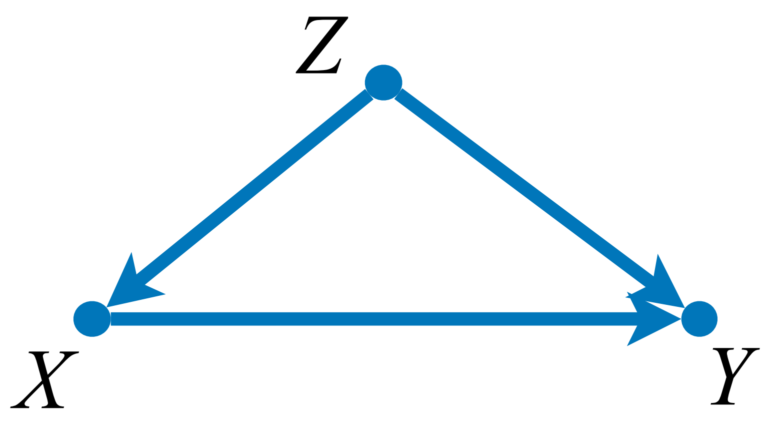
If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

Takeaways:

- $g(P) = P_x(y)$ even if $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y, \mid X, Z)$
- $g(P) = P_x(y)$ even if $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$

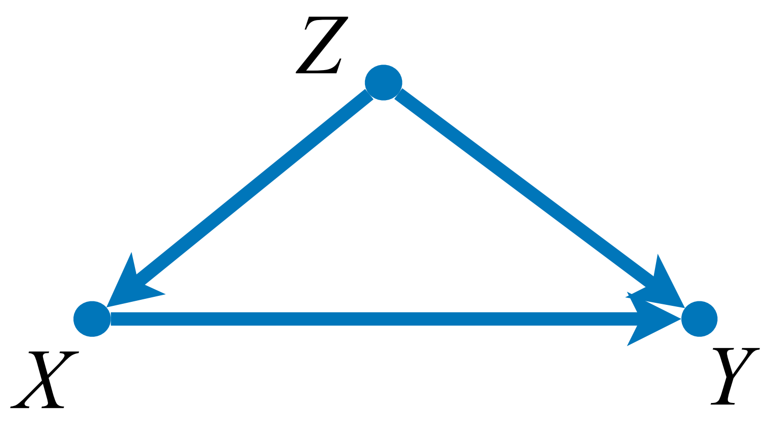
$\Longrightarrow$ **Doubly robustness:** An estimand $g(P)$ gives a double chance of being correct!

Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

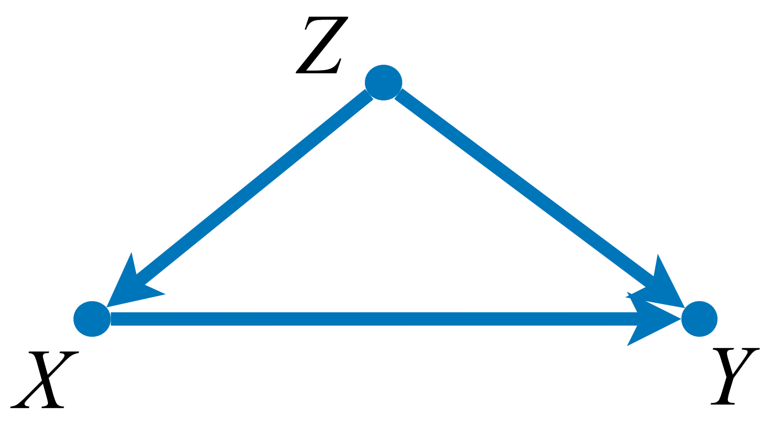
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$

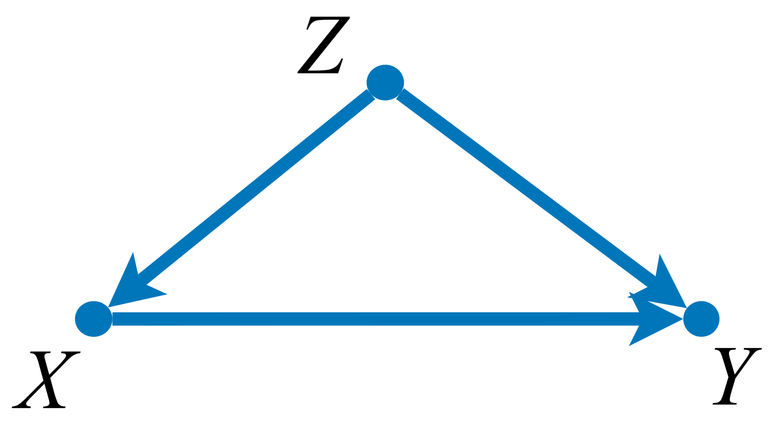
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)}(I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z)\right]$

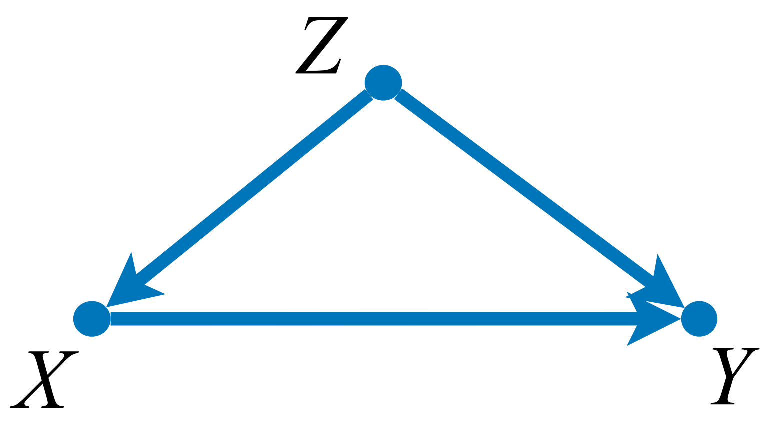
Classic BD estimator:
AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} (I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z) \right]$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; Or $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.

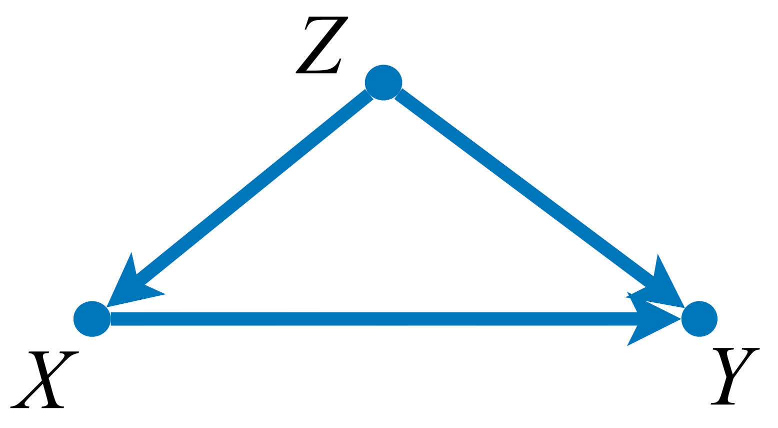
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)}(I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z)\right]$
For correct estimation	😁 "*Doubly robustness*" — Double chance of being correct!

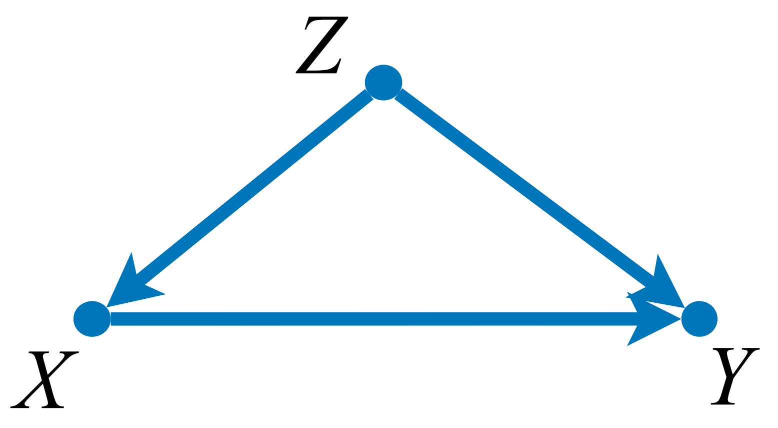
Classic BD estimator:
AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)}(I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z)\right]$
For correct estimation	😁 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😁 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \rightarrow \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate).

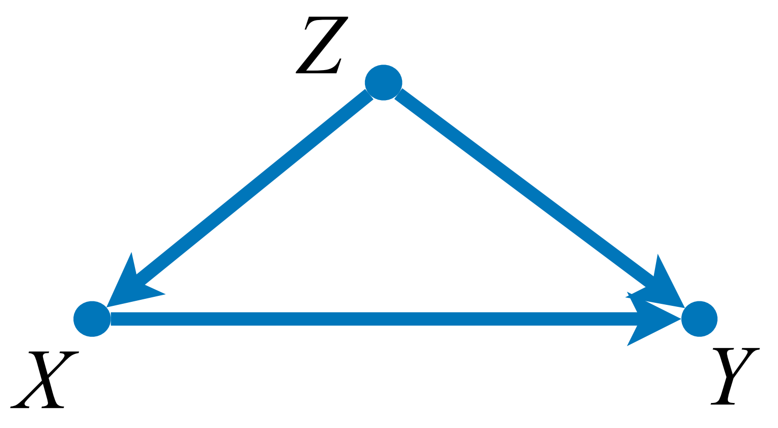
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)}(I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z)\right]$
For correct estimation	😁 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😁 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \rightarrow \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate). $\{\hat{P}(x \mid z), P(x \mid z), \hat{P}(y \mid x, z), P(y \mid x, z)\}$ in Donsker class.

Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)}(I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z)\right]$
For correct estimation	😁 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😁 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \to \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate). 😢 Unclear that modern ML methods are in Donsker.

Classic BD estimator: AIPW

Z

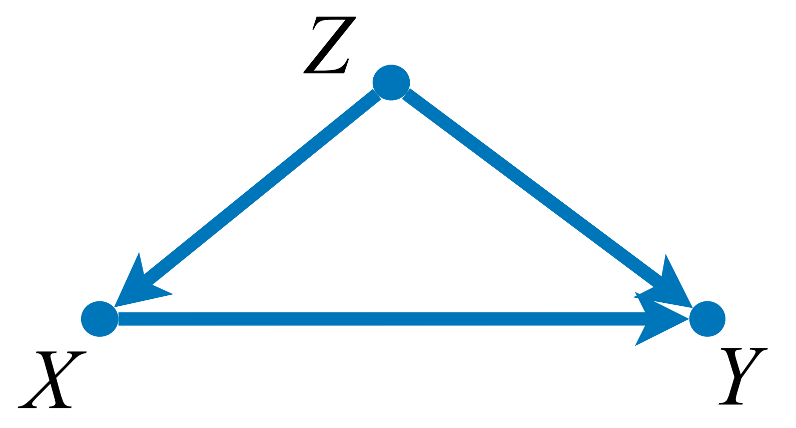
$$P_x(y) = \sum P(y \mid x, z)P(z)$$

For fast ($N^{-1/2}$ rate) convergence

relatively slowly ($N^{-1/4}$ rate).

😢 Unclear that modern ML methods are in Donsker.

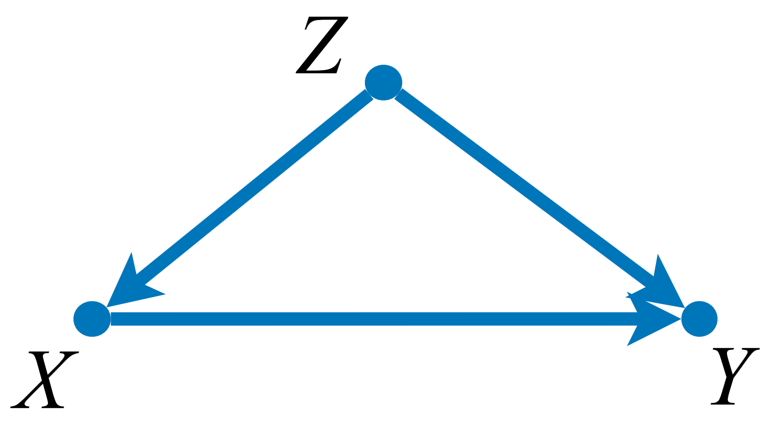
Double/Debiased Machine Learning



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; Or $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

Double/Debiased Machine Learning



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$
Estimator (T_N)	$\frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i} \left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)} (I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z) \right]$
For correct estimation	$\hat{P}(x \mid z) \rightarrow P(x \mid z)$; Or $\hat{P}(y \mid x, z) \rightarrow P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

Double/Debiased Machine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.

od$(P) = \sum_{z} P(y \mid x, z) P(z)$

$)) + P(y \mid x, Z)\Big]$

Estimator (T_N)	$\frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i} \left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)} (I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z) \right]$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; Or $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

Double/Debiased Machine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x|z), \hat{P}_i(y|x,z)\}$.

od$(P) = \sum_z P(y|x,z)P(z)$

$)) + P(y|x,Z)\Big]$

Estimator (T_N)	$\frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X\|Z)}(I_y(Y) - \hat{P}_{1-i}(y\|X,Z)) + \hat{P}_{1-i}(y\|x,Z)\right]$
For correct estimation	$\hat{P}(x\|z) \to P(x\|z)$; Or $\hat{P}(y\|x,z) \to P(y\|x,z)$.
For fast ($N^{-1/2}$ rate) convergence	

Double/Debiased Machine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x \mid z), \hat{P}_i(y \mid x, z)\}$.
3. Evaluate h using samples in D_i with models $\{\hat{P}_{1-i}(x \mid z), \hat{P}_{1-i}(y \mid x, z)\}$ trained through D_{1-i} (i.e., samples for evaluation / training are distinct)

od$(P) = \sum_z P(y \mid x, z) P(z)$

$)) + P(y \mid x, Z)\Big]$

Estimator (T_N)	$\frac{1}{2}\sum_{i \in \{0,1\}} \underbrace{\mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right]}_{\equiv h}$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; Or $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

Double/Debiased Machine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

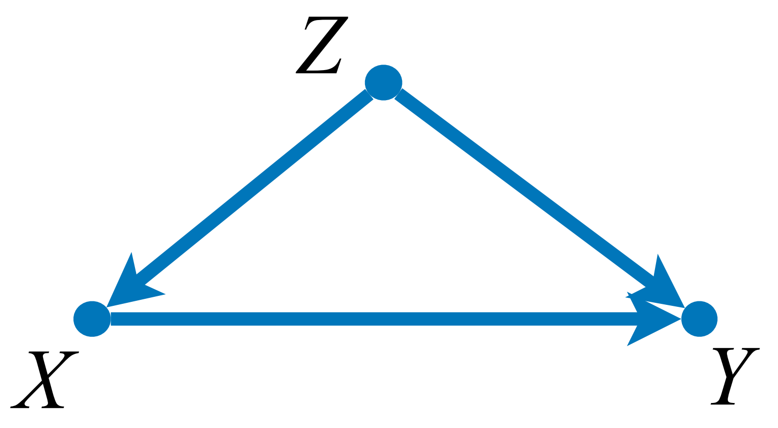
1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x \mid z), \hat{P}_i(y \mid x, z)\}$.
3. Evaluate h using samples in D_i with models $\{\hat{P}_{1-i}(x \mid z), \hat{P}_{1-i}(y \mid x, z)\}$ trained through D_{1-i} (i.e., samples for evaluation / training are distinct)
4. Take an empirical expectation of each h over D_i (i.e., $\mathbb{E}_{P_{D_i}}$) and divide it half.

od$(P) = \sum_{z} P(y \mid x, z) P(z)$

$)) + P(y \mid x, Z)\Big]$

Estimator (T_N)	$\frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i} \left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)} (I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z) \right]$ $\equiv h$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; Or $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

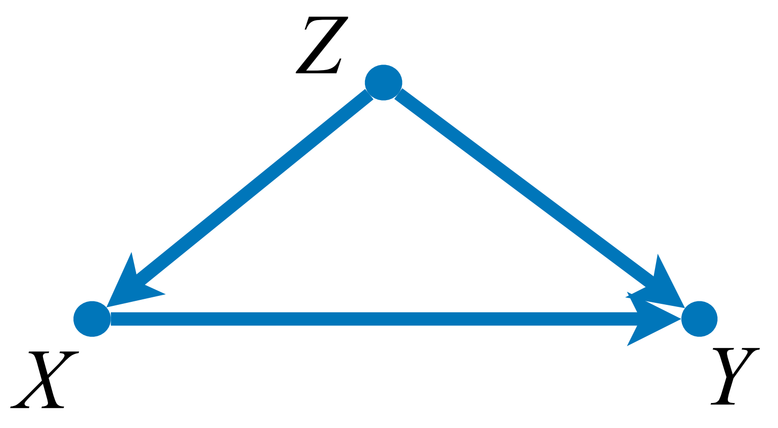
Double/Debiased Machine Learning



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right]$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; Or $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

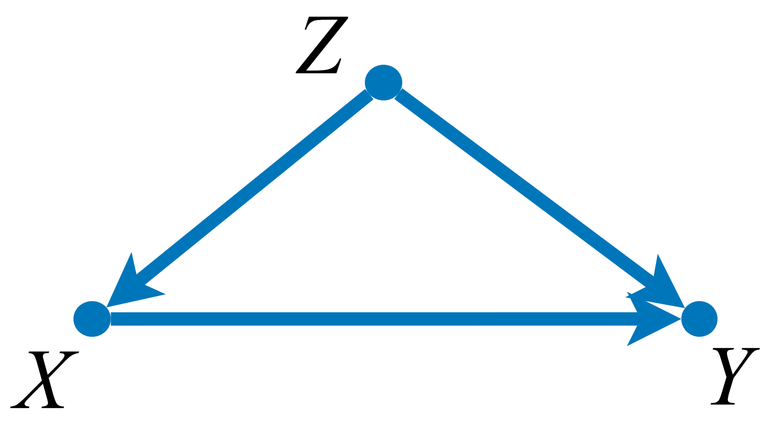
Double/Debiased Machine Learning



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z)P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\frac{1}{2}\sum_{i \in \{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right]$
For correct estimation	😁 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😁 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \rightarrow \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate).

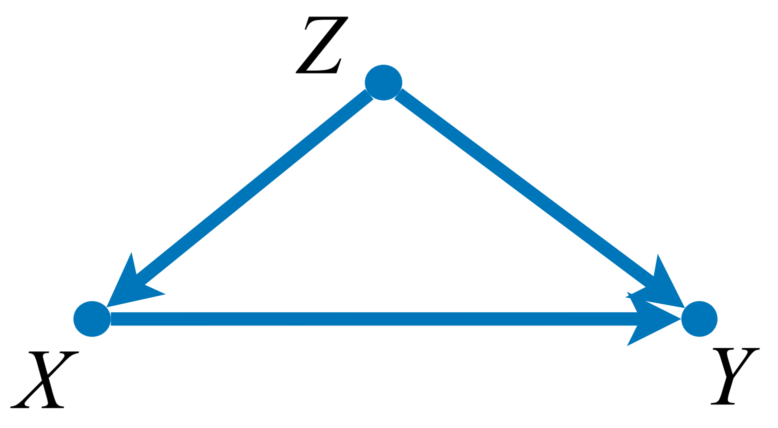
Double/Debiased Machine Learning



$$P_x(y) = f_{\mathrm{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\frac{1}{2}\sum_{i \in \{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right]$
For correct estimation	😁 … t! Functions not in Donsker class are prone to overfitting bias. Overfitting bias are mitigated by sample-splitting.
For fast ($N^{-1/2}$ rate) convergence	😁 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \rightarrow \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate).

Double/Debiased Machine Learning



$$P_x(y) = f_{\mathsf{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)\right]$
Estimator (T_N)	$\frac{1}{2}\sum_{i \in \{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right]$

Key result:

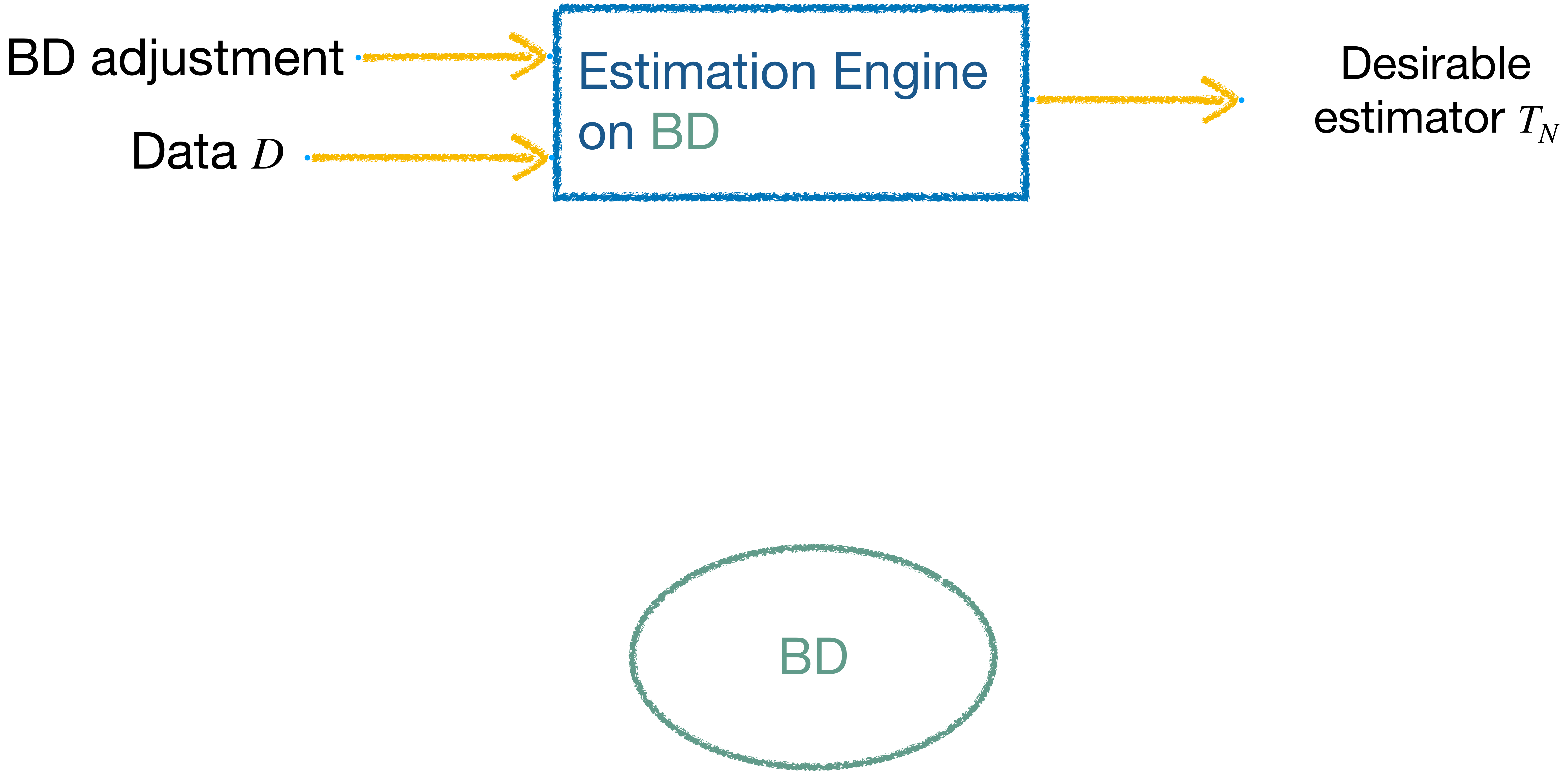
Double/debiased machine learning (DML) estimator for BD enjoys doubly robustness and debiasedness without restrictions on the function class!

4. My research theme

My research theme

My research theme

Under the BD setting,

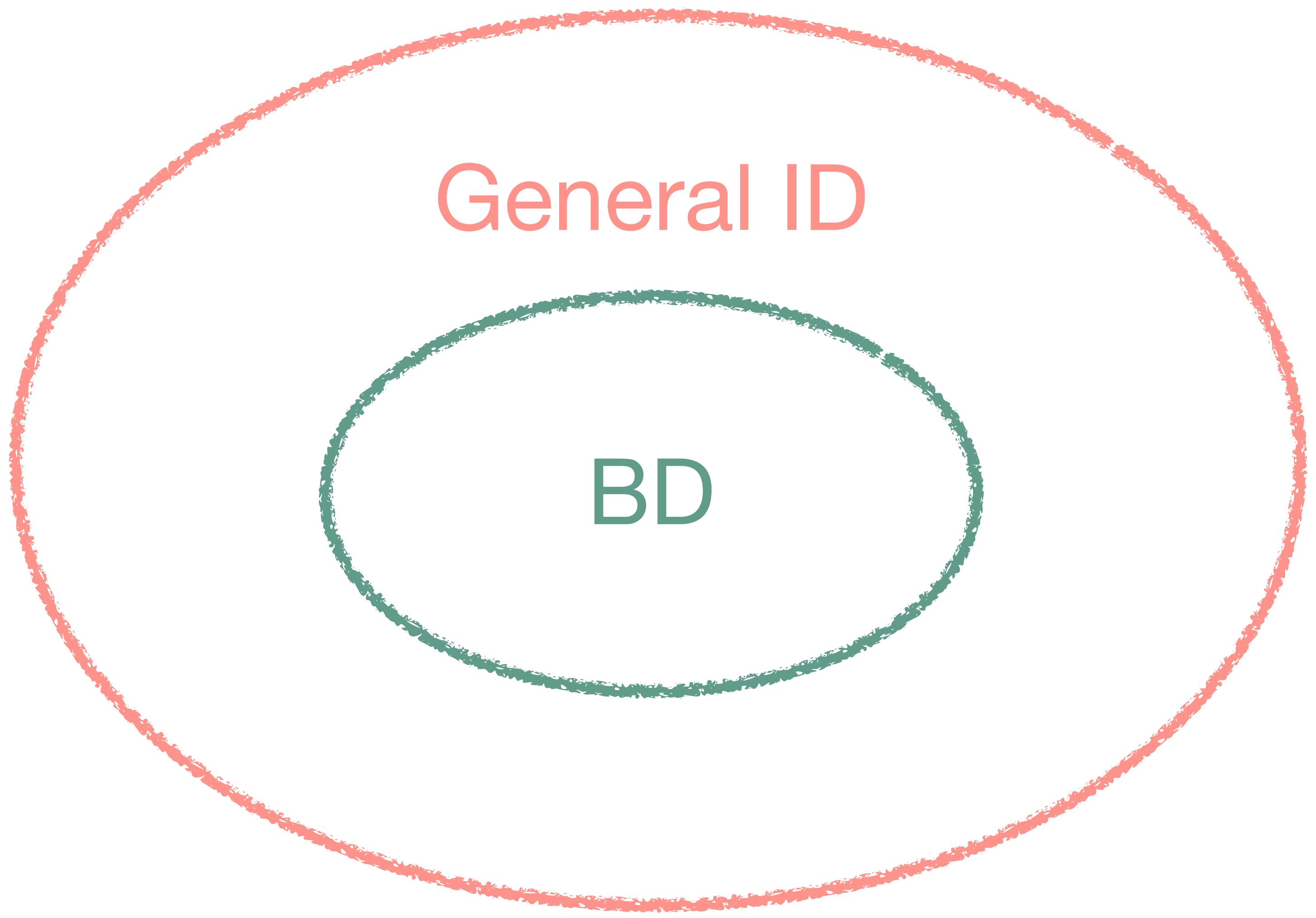


My research theme

Under the general identifiable setting (i.e., general $f(P) = P_x(y)$),

Functional $f(P)$ → **Estimation Engine on general ID** → ? Desirable estimator T_N

Data D → **Estimation Engine on general ID**



Estimators beyond the BD case

	Statistical properties		Causal properties		Graphical properties
	Doubly Robustness	Debiasedness	Beyond BD	General ID	Not assuming fully-specified graph

Estimators beyond the BD case

	Statistical properties		Causal properties		Graphical properties
	Doubly Robustness	Debiasedness	Beyond BD	General ID	Not assuming fully-specified graph
Jung, Tian, Bareinboim (2020a)	✘	✘	✔	✘	✘

Estimators beyond the BD case

	Statistical properties		Causal properties		Graphical properties
	Doubly Robustness	Debiasedness	Beyond BD	General ID	Not assuming fully-specified graph
Jung, Tian, Bareinboim (2020a)	✗	✗	✔	✗	✗
Fulcher et al., (2019) Bhattacharya et al (2020)	✔	✗	✔	✗	✗

Estimators beyond the BD case

	Statistical properties		Causal properties		Graphical properties
	Doubly Robustness	Debiasedness	Beyond BD	General ID	Not assuming fully-specified graph
Jung, Tian, Bareinboim (2020a)	✗	✗	✓	✗	✗
Fulcher et al., (2019) Bhattacharya et al (2020)	✓	✗	✓	✗	✗
Jung, Tian, Bareinboim (2020b)	✗	✗	✓	✓	✗

Estimators beyond the BD case

	Statistical properties		Causal properties		Graphical properties
	Doubly Robustness	Debiasedness	Beyond BD	General ID	Not assuming fully-specified graph
Jung, Tian, Bareinboim (2020a)	✗	✗	✓	✗	✗
Fulcher et al., (2019) Bhattacharya et al (2020)	✓	✗	✓	✗	✗
Jung, Tian, Bareinboim (2020b)	✗	✗	✓	✓	✗
Jung, Tian, Bareinboim (2021a))	✓	✓	✓	✓	✗

Estimators beyond the BD case

	Statistical properties		Causal properties		Graphical properties
	Doubly Robustness	Debiasedness	Beyond BD	General ID	Not assuming fully-specified graph
Jung, Tian, Bareinboim (2020a)	✘	✘	✔	✘	✘
Fulcher et al., (2019) Bhattacharya et al (2020)	✔	✘	✔	✘	✘
Jung, Tian, Bareinboim (2020b)	✘	✘	✔	✔	✘
Jung, Tian, Bareinboim (2021a))	✔	✔	✔	✔	✘
Jung, Tian, Bareinboim (2021b))	✔	✔	✔	✔	✔

Key points (So far)

Key points (So far)

- There have been many estimators for Back-door settings (ignorability).
  - IPW, OR, AIPW, Double machine learning, etc.

Key points (So far)

- There have been many estimators for Back-door settings (ignorability).
    - IPW, OR, AIPW, Double machine learning, etc.

- No estimators have been developed for the general ID setting.

Key points (So far)

- There have been many estimators for Back-door settings (ignorability).
  - IPW, OR, AIPW, Double machine learning, etc.

- No estimators have been developed for the general ID setting.

- My research theme is filling this lacuna — *developing an estimator for general ID functional.*

Conclusion

Conclusion

- Structural Causal Model (SCM) is a complete framework for causal inference.

Conclusion

- Structural Causal Model (SCM) is a complete framework for causal inference.
- In SCM, causal effect identification is important. There is a complete solution for ID problems.

Conclusion

- Structural Causal Model (SCM) is a complete framework for causal inference.
- In SCM, causal effect identification is important. There is a complete solution for ID problems.
- Since causal effect estimation problems have remained open, I have solved the estimation problems for general ID settings.