CS 295: Causal Reasoning

Rina Dechter

Identification of Causal Effect
The Back-Door Criterion

Primer chapter 3, Causality 1.3,3.1,3.2)

Outline (chapter 3)

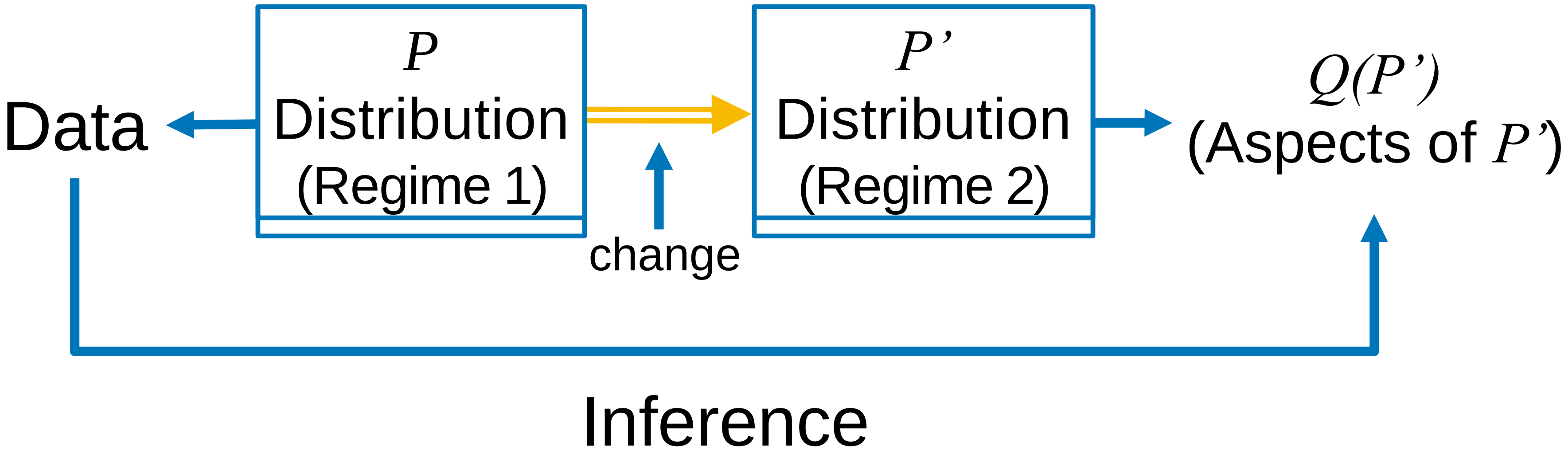
- The semantic of Intervention in SCM, the do operators
- How to determine P(Y|do(x)) given an SCM
- The back door criterion and the adjustment formula
- Identifiability

Target: to Determine the Effect of Interventions

- "Correlation is no causation", e.g., Increasing ice-cream sales is correlated with more crime, still selling more ice-cream will not cause more violence. Hot weather is a cause for both.
- **Randomized controlled experiments** are used to determine causation: all factors except a selected one of interest are kept static or random. So the outcome can only be influenced by the selected factor.
- Randomized experiments are often not feasible (we cannot randomize the weather), so how can we determine cause for wildfire?
- **Observational studies** must be used. But how we untangle correlation from causation?


CBSNEWS
Video
Science News
from universities, journals, and other research organizations
One Drink Of Red W
Drinks Are Stressf
To Circulation, But Two
Feb. 13, 2008 — One dr
alcohol slightly
but the positive
disappear with
Peter Munk Ca
Hospital.
The New York
INSIDE THE EM
OF BOYS
Home
News
Current Issue
Why I
One of the
the consis
By John Cl
Like
Correc
THE CON
Gra
Re
By RON
PRINT
In Texas it is illegal to take
41 Comments
Shares
The NEW ENGLAND JOURNAL of MEDICINE
HOME
ARTICLES & MULTIMEDIA
ISSUES
SPECIALTIES & TOPICS
FOR AUTHORS
CME
ORIGINAL ARTICLE
Association of Nut Consumption with Total and Cause-Specific Mortality
Bao, M.D., Sc.D., Jiali Han, Ph.D., Frank B. Hu, M.D., Ph.D., Edward L. Giovannucci, M.D., Sc.D., Meir J. Stampfer, M.D., Dr.P.H., Walter C. Willett, M.D., Dr.P.H., and Charles S. Fuchs, M.D., M.P.H.
N Engl J Med 2013; 369:2001-2011 | November 21, 2013 | DOI: 10.1056/NEJMoa1307352
Abstract
Article
References
Share:
BACKGROUND
Increased nut consumption has been associated with a reduced risk of major chronic diseases, including cardiovascular disease and type 2 diabetes mellitus. However, the association between nut consumption and mortality remains unclear.
Full Text of Background...
METHODS
We examined the association between nut consumption and
MEDIA IN THIS ARTICLE
Video
Nuts

Causal Inference — Connecting Different Worlds

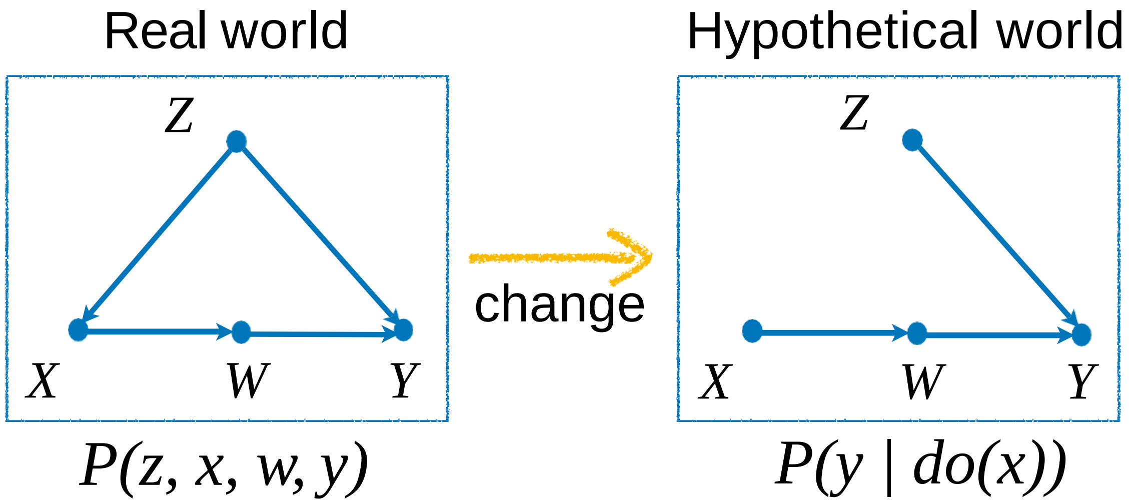


What happens when P changes?

e.g., Infer whether less people would get cancer if we ban smoking.

$Q = P(Cancer = true \mid do(Smoking = no))$ Not an aspect of P.

The Challenge of Causal Inference

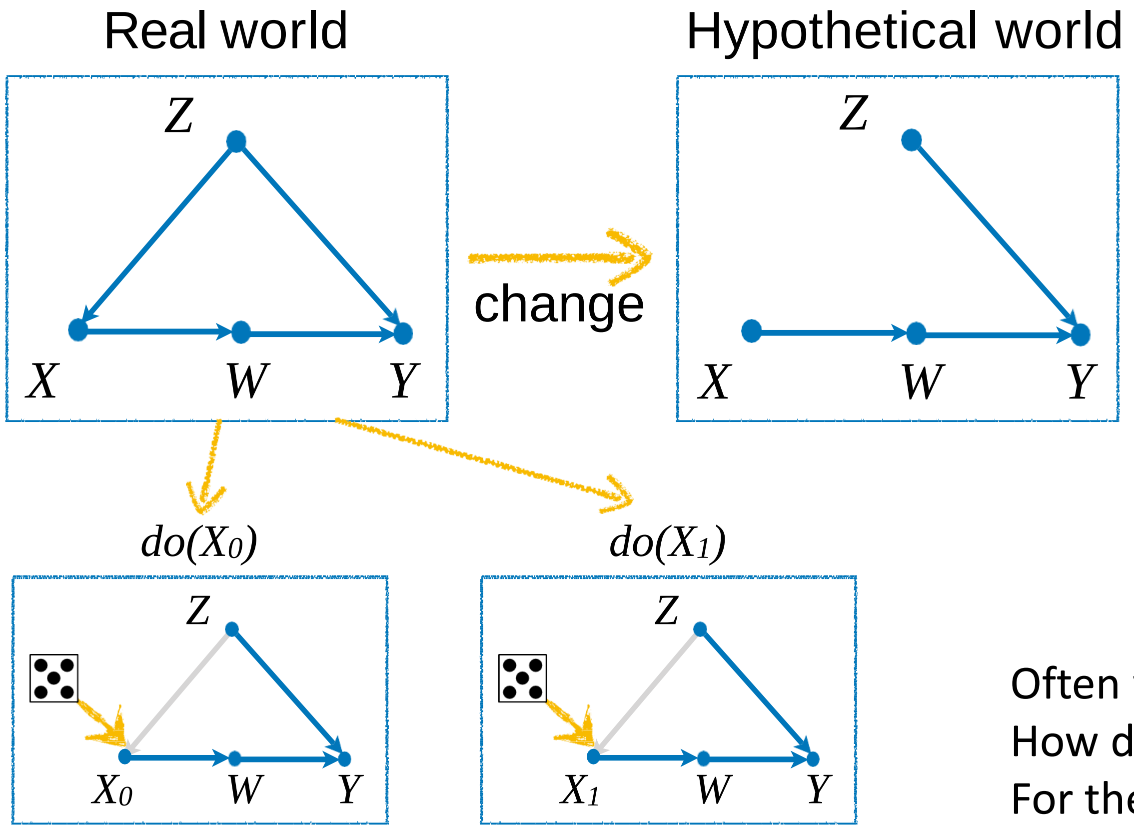


Z : age, sex
X : action
W : mediator
Y : outcome

- Goal: how much Y changes with X if we vary X between two different constants free of the influence of Z.
- These variations are called causal effects!

Bareinboim slides 2020

Method for Computing Causal Effects: Randomized Experiments



Z : age, sex

X : action

W : mediator

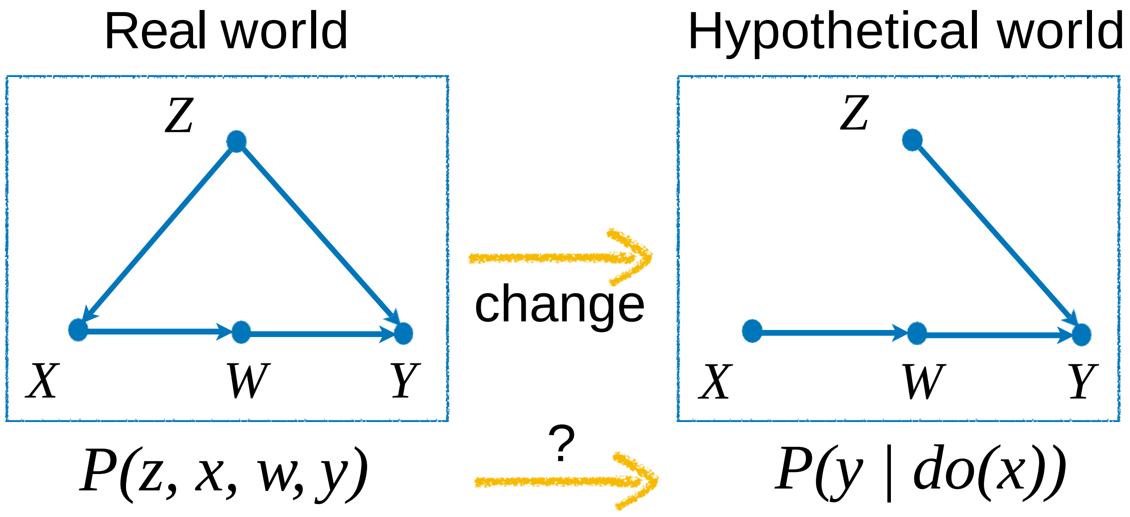
Y : outcome

Randomization:

$P(y \mid do(X_0))$ $\quad$ $P(y \mid do(X_1))$

Often we cannot do this:
How do we force people to smoke (and wait 20 years
For them to die or not
How can we change cholesterol levels…

Computing Causal Effects (l2) from Observational Data (l1)



Z : age, sex
X : action
W : mediator
Y : outcome

Questions:

* What is the relationship between $P(z, x, w, y)$ and $P(y \mid do(x))$?
* Is $P(y \mid do(x)) = P(y \mid x)$?

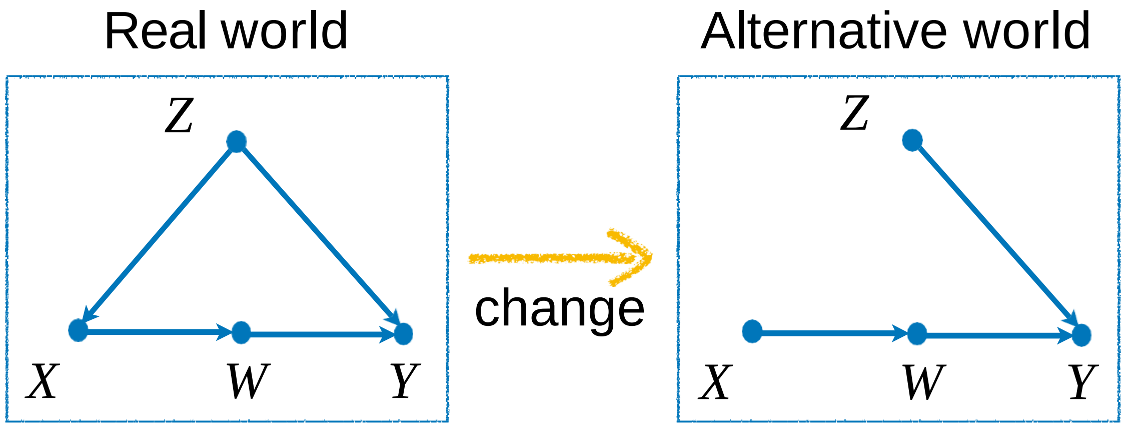
Causal Effects (formal)

Causal Effect (Def. 3.2.1 [C]):

Given two disjoint sets of variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$, the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$, denoted as $P(\mathbf{y} \mid do(\mathbf{x}))$, is a function from $\boldsymbol{X}$ to the space of probability distributions of $\boldsymbol{Y}$.

For each realization $\mathbf{x}$ of $\boldsymbol{X}$, $P(\mathbf{y} \mid do(\mathbf{x}))$ gives the probability $\boldsymbol{Y} = \boldsymbol{y}$ induced by deleting from the model all equations corresponding to variables in $\boldsymbol{X}$ and substituting $\boldsymbol{X} = \mathbf{x}$ in the remaining equations.

Computing Causal Effects from Observational Data



Z : age, sex

X : action

W : mediator

Y : outcome

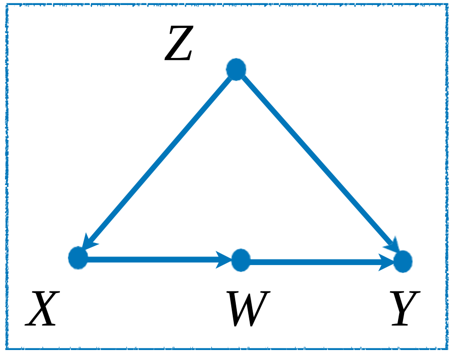
$$M = \begin{cases} Z = f_Z(u_z) \\ X = f_X(z, u_x) \\ W = f_W(x, u_w) \\ Y = f_Y(w, z, u_y) \end{cases}$$

$do(X=x)$

$$M_x = \begin{cases} Z = f_Z(u_z) \\ \cancel{X = f_X(z, u_x)} \quad X = x \\ W = f_W(x, u_w) \\ Y = f_Y(w, z, u_y) \end{cases}$$

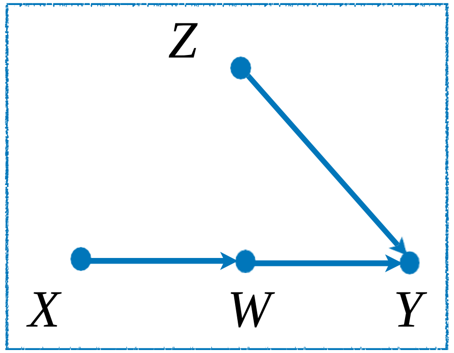
Computing Causal Effects from Observational Data

Real world



change

Alternative world



Z : age, sex

X : action

W : mediator

Y : outcome

$P(v) =$

$P(z) \times P(x \mid z) \times P(w \mid x) \times P(y \mid w, z)$

$do(X=x)$

$P_x(v) =$

$P(z) \times$

~~$P(x)$~~ $\times$ equal to 1 in M_x

$P(w \mid x) \times$

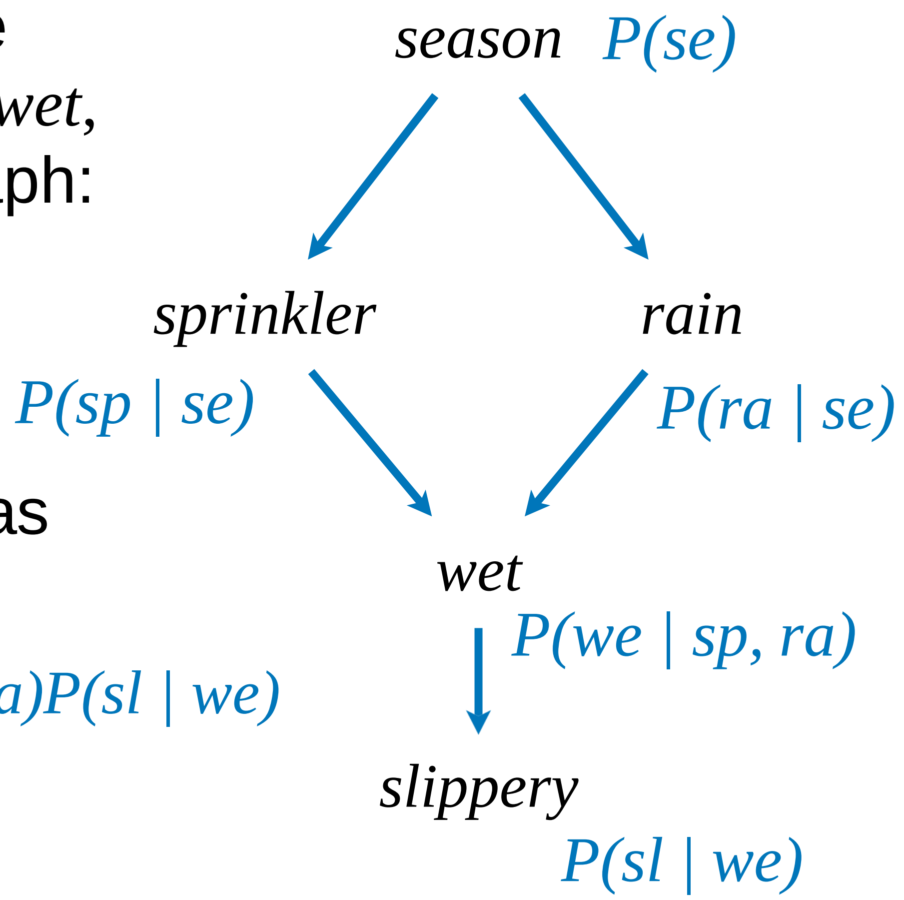
$P(y \mid w, z)$

Outline (chapter 3)

- The semantic of Intervention in SCM, the do operators
- How to determine P(Y|do(x)) given an SCM
- The back door criterion and the adjustment formula
- Identifiability

Computing Causal Effects from Observational Data

Consider a distribution over the variables: *season*, *sprinkler*, *rain*, *wet*, and *slippery*; and the causal graph:



This distribution decomposes as

$P(\mathbf{v}) =$

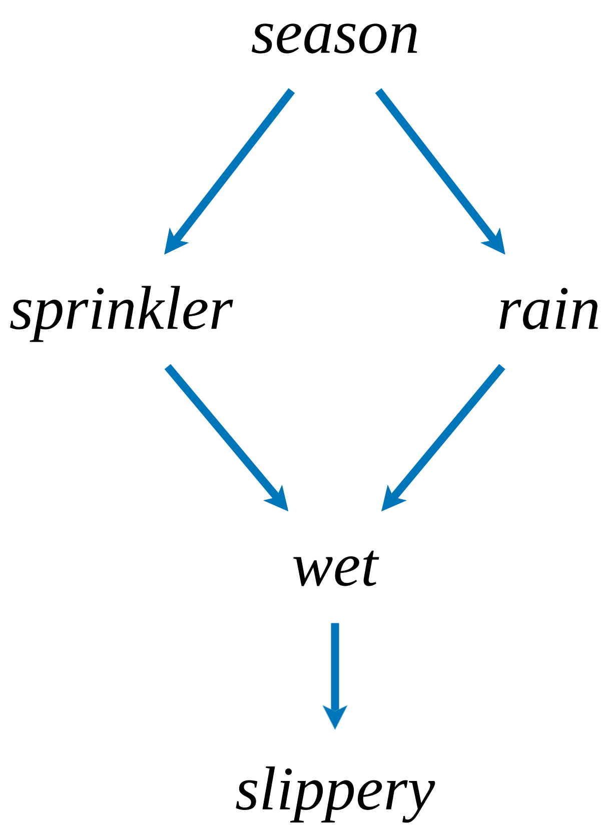
$P(se)P(sp \mid se)P(ra \mid se)P(we \mid sp, ra)P(sl \mid we)$

Computing Causal Effects from Observational Data

Queries:

$Q_1 = P(wet \mid Sprinkler = on)$

$Q_2 = P(wet \mid do(Sprinkler = on))$

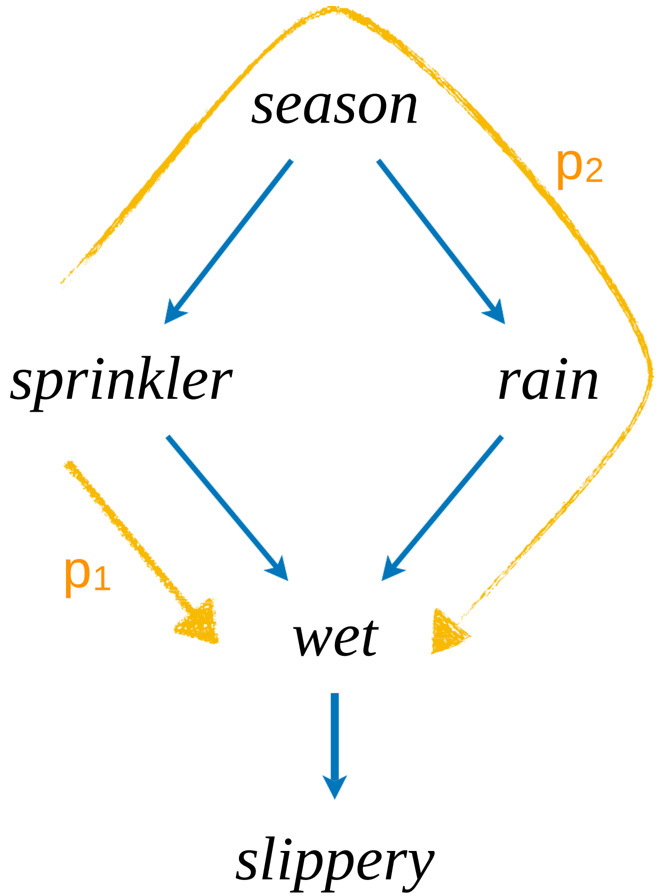


Computing Causal Effects from Observational Data

Queries:

$Q_1 = P(wet \mid Sprinkler = on)$
$= P(p_1) + P(p_2)$

$Q_2 = P(wet \mid do(Sprinkler = on))$



Computing Causal Effects from Observational Data

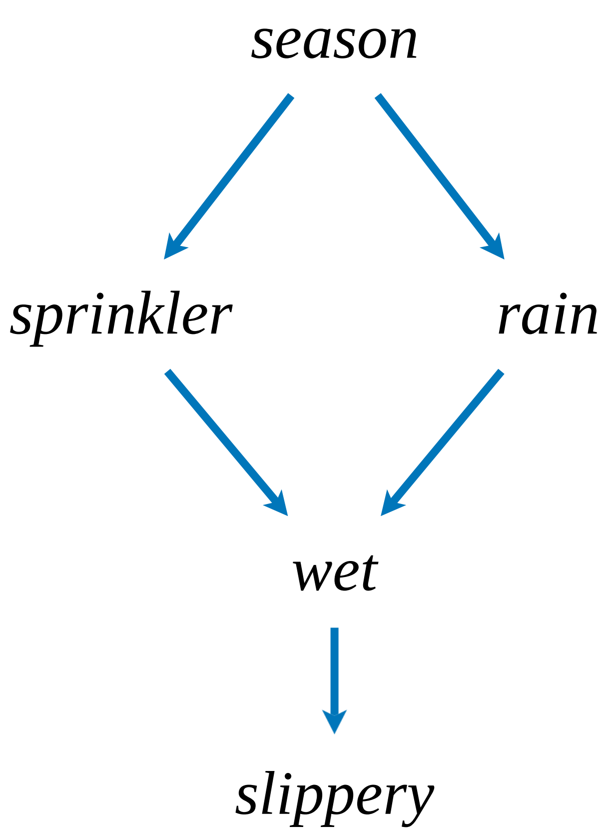
Queries:

$Q_1 = P(wet \mid Sprinkler = on)$

$$= \frac{\sum_{se,ra} P(we \mid Sp = on, ra) P(Sp = on \mid se) P(ra \mid se) P(se)}{\sum_{se} P(Sp = on \mid se) P(se)}$$

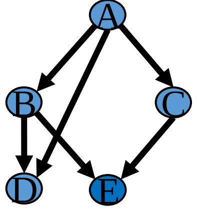
$Q_2 = P(wet \mid do(Sprinkler = on))$

You can do algorithm bucket elimination to infer Q1.



Bucket elimination

Algorithm *BE-bel* (Dechter 1996)



$$P(A \mid E=0) = \alpha \sum_{E=0,D,C,B} P(A)\cdot P(B \mid A)\cdot P(C \mid A)\cdot P(D \mid A,B)\cdot P(E \mid B,C)$$

$\sum_b \prod$ ← Elimination operator

bucket B: P(b|a) P(d|b,a) P(e|b,c)

bucket C: P(c|a) $\lambda^B(a,d,c,e)$

bucket D: $\lambda^C(a,d,e)$

bucket E: e=0 $\lambda^D(a,e)$

bucket A: P(a) $\lambda^E(a)$

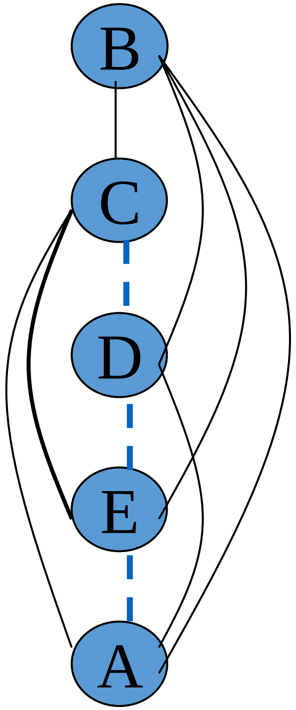
$P(e=0)$

$P(a,e=0)$

W*=4

"induced width"

(max clique size)



$$P(a \mid e=0) = \frac{P(a,e=0)}{P(e=0)}$$

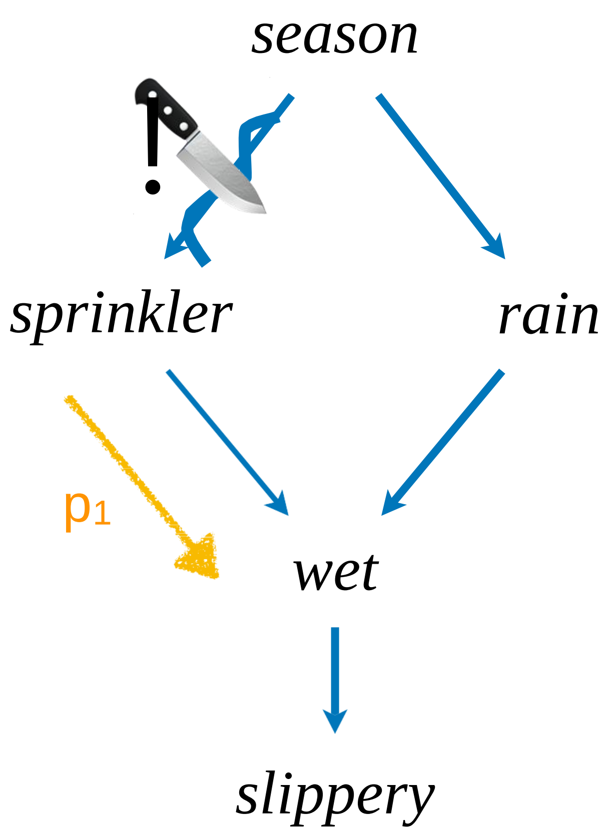
Computing Causal Effects from Observational Data

Queries:

$Q_1 = P(wet \mid Sprinkler = on)$

$$= \frac{\sum_{se,ra} P(we \mid Sp = on, ra) P(Sp = on \mid se) P(ra \mid se) P(se)}{\sum_{se} P(Sp = on \mid se) P(se)}$$

$Q_2 = P(wet \mid do(Sprinkler = on))$



You can do algorithm bucket elimination to infer Q2.

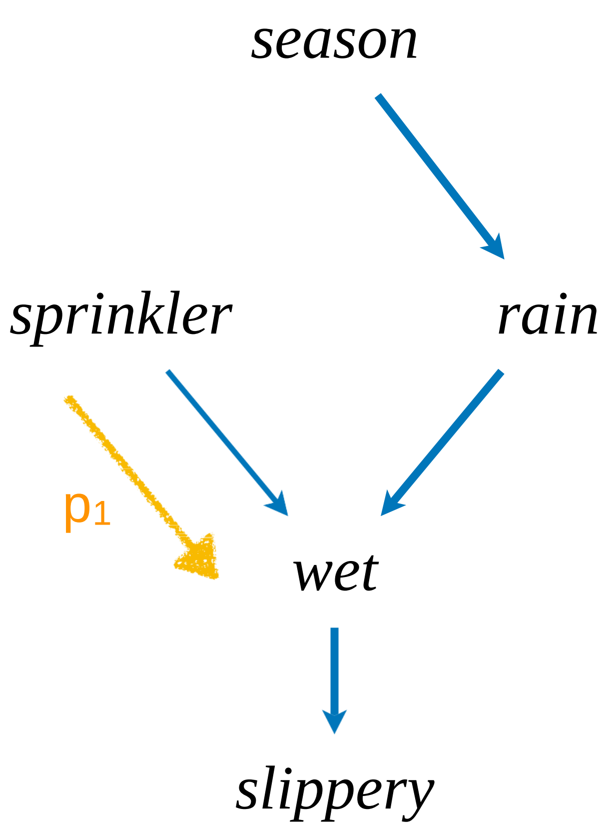
Computing Causal Effects from Observational Data

Queries:

$$Q_1 = P(wet \mid Sprinkler = on)$$

$$= \frac{\sum_{se,ra} P(we \mid Sp = on, ra) P(Sp = on \mid se) P(ra \mid se) P(se)}{\sum_{se} P(Sp = on \mid se) P(se)}.$$

$$Q_2 = P(wet \mid do(Sprinkler = on))$$
$$= P(p_1)$$



You can do algorithm bucket elimination to infer Q1.

Computing Causal Effects from Observational Data

Queries:

$$Q_1 = P(wet \mid Sprinkler = on)$$

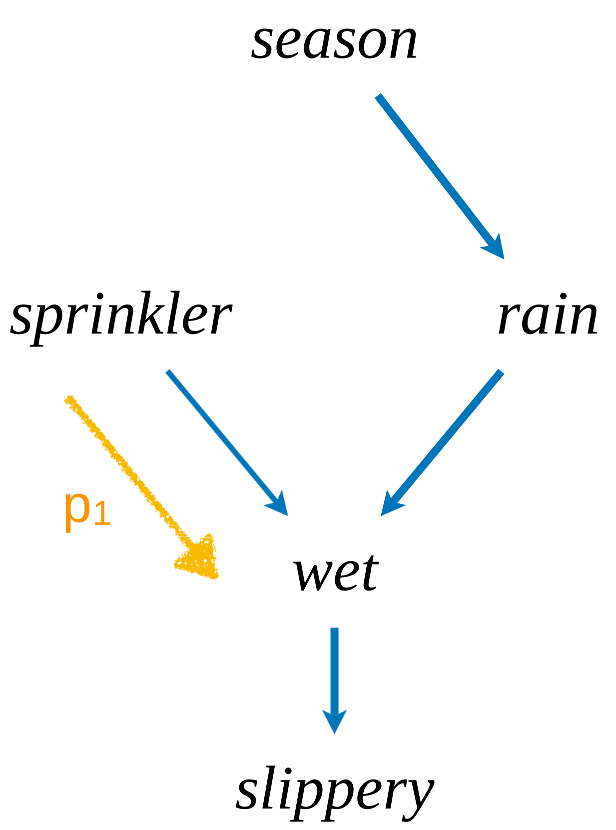
$$= \frac{\sum_{se,ra} P(we \mid Sp = on, ra) P(Sp = on \mid se) P(ra \mid se) P(se)}{\sum_{se} P(Sp = on \mid se) P(se)}$$

$$Q_2 = P(wet \mid do(Sprinkler = on))$$

$$= \frac{\sum_{se,ra} P(we \mid Sp = on, ra) P(Sp = on) P(ra \mid se) P(se)}{P(Sp = on)}$$

equal to 1

$$= \sum_{se,ra} P(we \mid Sp = on, ra) P(ra \mid se) P(se)$$



You can do algorithm bucket elimination to infer Q2.

Truncated Factorization Product (Operationalizing Interventions)

Corollary (Truncated Factorization, Manipulation Thm., G-comp.):

The distribution generated by an intervention $do(\mathbf{X}=\mathbf{x})$ (in a Markovian model M) is given by the truncated factorization:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{V_i \in \mathbf{V} \setminus \mathbf{X}\}} P(v_i \mid pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

Truncated Factorization Formula

The truncated product,

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{V_i \in \mathbf{V} \setminus \mathbf{X}\}} P(v_i \mid pa_i) \Bigg|_{\mathbf{X}=\mathbf{x}}$$

can be rewritten as:

$$P(\mathbf{v} \mid do(\mathbf{x})) = \frac{P(\mathbf{v})}{P\left(\mathbf{x} \mid pa_{\mathbf{x}}\right)} \Bigg|_{\mathbf{X}=\mathbf{x}}$$

Also equivalent to:

$$P(\mathbf{v} \mid do(\mathbf{x})) = P\left(\mathbf{v} \mid \mathbf{x}, pa_{\mathbf{x}}\right) P\left(pa_{\mathbf{x}}\right) \Big|_{\mathbf{X}=\mathbf{x}}$$

The transformation between the observation and interventional distributions can be seen as a re-weighing process.

Intervention vs. Conditioning, The Ice-Cream Story

Conditioning P(X=x|Y=y)
Intervening P(X=x| do(Y=y))

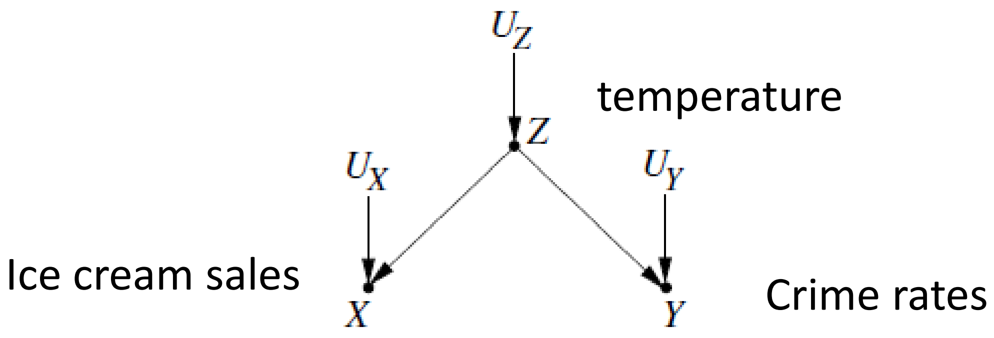


Figure 3.1: A graphical model representing the relationship between temperature (Z), ice cream sales (X), and crime rates (Y)

When we intervene to fix a value of a variable,
We curtail the natural tendencies of the variable to vary
In response to other variables in nature.

- **This corresponds to a surgery of the model**
- i.e. varying Z will not affect X
- intervention is different than conditioning.
- **Intervention depends on the structure of the graph.**

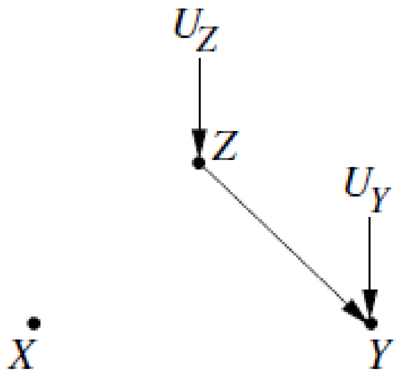
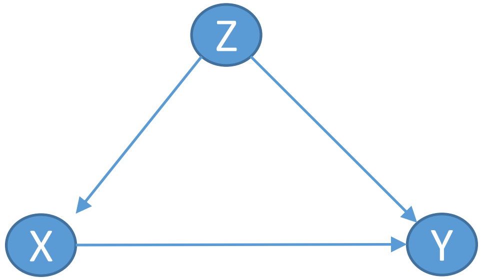


Figure 3.2: A graphical model representing an intervention on the model in Figure 3.1 that lowers ice cream sales

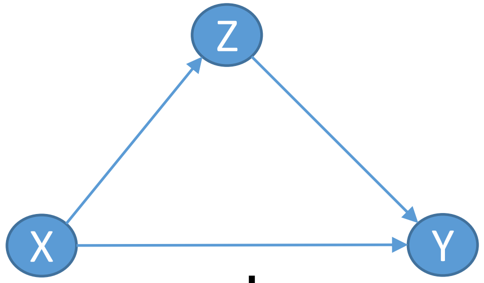
Intervention vs Conditioning, The Surgery Operation

Conditioning P(Y=y|X=x)

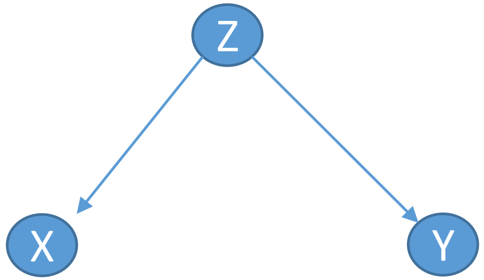
The Simpson story



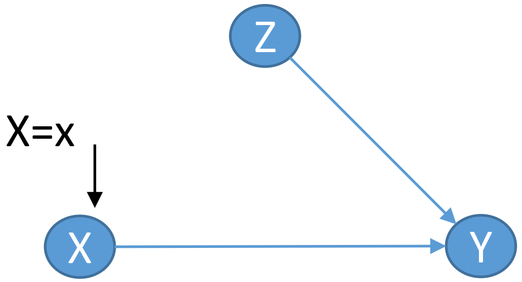
The blood pressure story

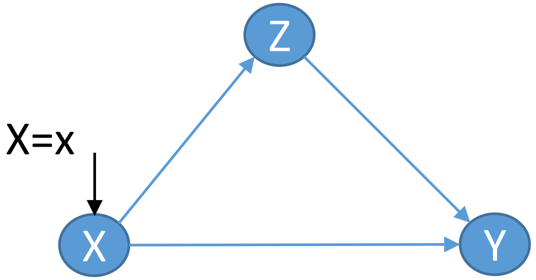


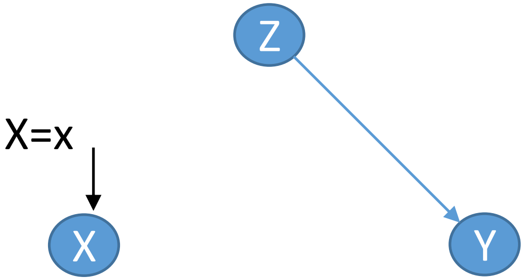
The ice-cream story



Intervening P(Y=y| do(X=x))







Intervention vs. Conditioning…

In notation, we distinguish between cases where a variable X takes a value x naturally and cases where we fix $X = x$ by denoting the latter $do(X = x)$. So $P(Y = y|X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, while $P(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to make $X = x$. In the distributional terminology, $P(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x. On the other hand, $P(Y = y|do(X = x))$ represents the population distribution of Y if *everyone in the population* had their X value fixed at x. We similarly write $P(Y = y|do(X = x), Z = z)$ to denote the conditional probability of $Y = y$, given $Z = z$, in the distribution created by the intervention $do(X = x)$.

Do operation and graph surgery can help determine causal effect

We make an assumption that intervention has no side-effect. Namely, assigning a variable by intervention does not affect other variables in a direct way.

The Adjustment Formula

To find out how effective the drug is in the population, we imagine a hypothetical intervention by which we administer the drug uniformly to the entire population and compare the recovery rate to what would obtain under the complementary intervention, where we prevent everyone from using the drug.

We want to estimate the "causal effect difference," or "average causal effect" (ACE).

$$P(Y = 1 | do(X = 1)) - P(Y = 1 | do(X = 0)) \quad (3.1)$$

We need a causal story articulated by a graph (for the Simpson story):

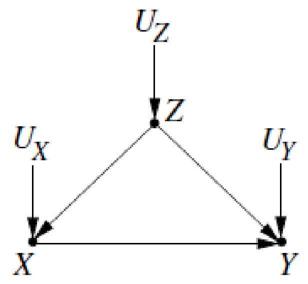
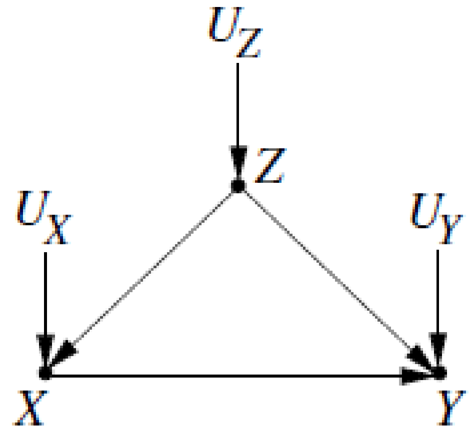


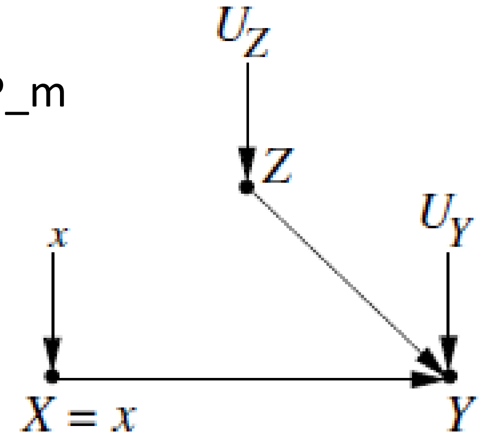
Figure 3.3: A graphical model representing the effects of a new drug, with Z representing gender, X standing for drug usage, and Y standing for recovery

Definition of Intervention and Graph Surgery: The Adjustment Formula



- We simulate the intervention in the form of a graph surgery.
- The causal effect $P(Y = y|do(X = x))$ equals to the conditional probability $P_m(Y = y|X = x)$ that prevails in the manipulated model of the figure below

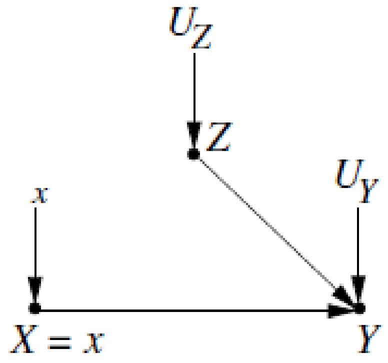
P_m



Important: the random functions for Z and Y remain invariant

$$P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x) \quad \text{and} \quad P_m(Z = z) = P(Z = z)$$

The Adjustment Formula



$$P(Y = y|do(X = x))$$

$$= P_m(Y = y|X = x) \qquad \text{(by definition)} \tag{3.2}$$

$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z|X = x) \tag{3.3}$$

$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z) \tag{3.4}$$

Equation (3.3) is obtained from Bayes' rule by conditioning on and summing over all values of $Z = z$ (as in Eq. (1.19)), while (Eq. 3.4) makes use of the independence of Z and X in the modified model.

Finally, using the invariance relations, we obtain a formula for the causal effect, in terms of preintervention probabilities:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{3.5}$$

Equation (3.5) is called the *adjustment formula* and as you can see, it computes the association between X and Y for each value z of Z, then averages over those values. This procedure is referred to as "adjusting for Z" or "controlling for Z."

The Adjustment Formula (in the Simpson story)

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \qquad (3.5)$$

The right hand-side can be estimated from the data since it has only conditional probabilities.

If we had a randomized controlled experiments on X (taking the drug) we would not need adjustment Because the data is already generated from the manipulated distribution. Namely it will yield P(Y=y|do(x)) From the data of the randomized experiment.

In practice adjustment is sometime used in randomized experiments to reduce sampling variations (Cox 1958). (This means: If the input is samples from the joint distribution over X,Y and Z we can estimate the P(y|x) directly. Or, we can first estimate P(y|x,s) and also P(z) and perform the summation.)

In the Simpson example:

Table 1.1 Results of a study into a new drug, with gender taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

$$P(Y=1|do(X=1)) = P(Y=1|X=1,Z=1)P(Z=1) + P(Y=1|X=1,Z=0)P(Z=0)$$

Substituting the figures given in Table 1.1 we obtain

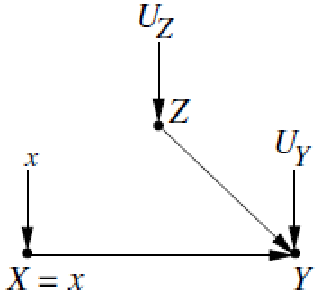
$$P(Y=1|do(X=1)) = \frac{0.93(87+270)}{700} + \frac{0.73(263+80)}{700}$$

=0.832

while, similarly,

$$P(Y=1|do(X=0)) = \frac{0.87(87+270)}{700} + \frac{0.69(263+80)}{700}$$

=0.7818



We get that the Average Causal Effect (ACE):

$$ACE = P(Y=1|do(X=1)) - P(Y=1|do(X=0)) = 0.832 - 0.7818 = 0.0502$$

A more informal interpretation of ACE is that it is the difference in the fraction of the population that would recover if everyone took the drug compared to when no one takes the drug.

The Blood Pressure Example

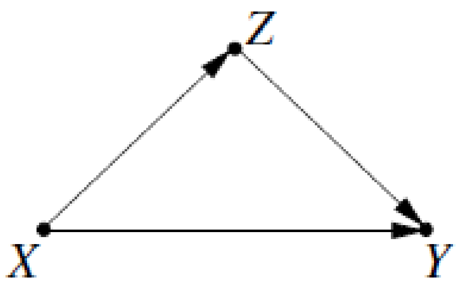


Figure 3.5: A graphical model representing the effects of a new drug, with X representing drug usage, Y representing recovery, and Z representing blood pressure (measured at the end of the study). Exogenous variables are not shown in the graph, implying that they are mutually independent

P(Y=y | do(X=x) = ?

Here the "surgery on X changes nothing. So,

This means that no surgery is required; the conditions under which data were obtained were such that treatment was assigned "as if randomized." If there was a factor that would make subjects prefer or reject treatment, such a factor should show up in the model; the absence of such a factor gives us the license to treat X as a randomized treatment.

$$P(Y = y|do(X = x)) = P(Y = y|X = x),$$

To Adjust or not to Adjust?

Rule 1 (The Causal Effect Rule) *Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by*

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, PA = z)P(PA = z) \qquad (3.6)$$

Where z ranges over all the combinations of values that the variables n PA take.

So, the causal graph helps determine the parents PA!

But, in many cases some of the parents are unobserved so we cannot perform the calculation.

Luckily we can often adjust for other variables substituting for the unmeasured variables in PA(X), and this Can be decided via the graph.

Multiple Interventions, the Truncated Product Rule

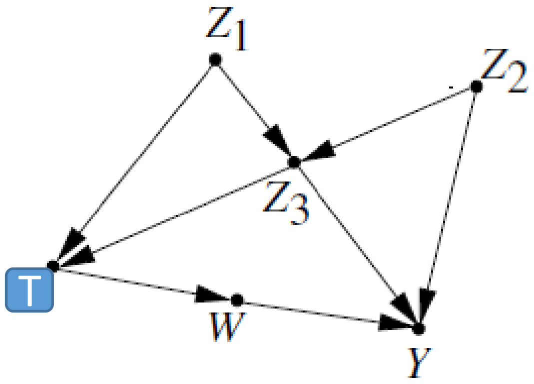
Often we have multiple interventions that may not correspond to disconnected variables.
We will use the product decomposition. We write **the product truncated formula**

$$P(x_1, x_2, \ldots, x_n | do(x)) = \prod_i P(x_i | pa_i) \qquad \text{for all } i \text{ with } X_i \text{ not in } X.$$

Example:

$$P(z_1, z_2, w, y | do(T = t, Z_3 = z_3)) = P(z_1)P(z_2)P(w|t)P(y|w, z_3, z_2)$$

where we have deleted the factors $P(t|z_1, z_3)$ and $P(z_3|z_1, z_2)$ from the product.



Multiple Interventions and the Truncated Product Rule

preintervention distribution in the model of Figure 3.3 is given by

$$P(x, y, z) = P(z)P(x|z)P(y|x, z)$$

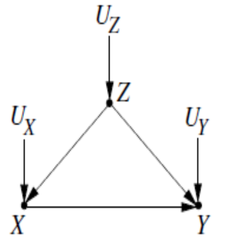


Figure 3.3: A graphical model representing the effects of a new drug, with Z representing gender, X standing for drug usage, and Y standing for recovery

whereas the postintervention distribution, governed by the model of Figure 3.4 is given by the product

$$P(z, y|do(x)) = P_m(z)P_m(y|x, z) = P(z)P(y|x, z) \quad (3.9)$$

with the factor $P(x|z)$ purged from the product, since X becomes parentless as it is fixed at $X = x$. This coincides with the adjustment formula, because to evaluate $P(y|do(x))$ we need to marginalize (or sum) over z, which gives

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

Outline (chapter 3)

- The semantic of Intervention in SCM, the do operators
- How to determine P(Y|do(x)) given an SCM
- The identification problem
- The back door criterion and the adjustment formula

The Identification Problem

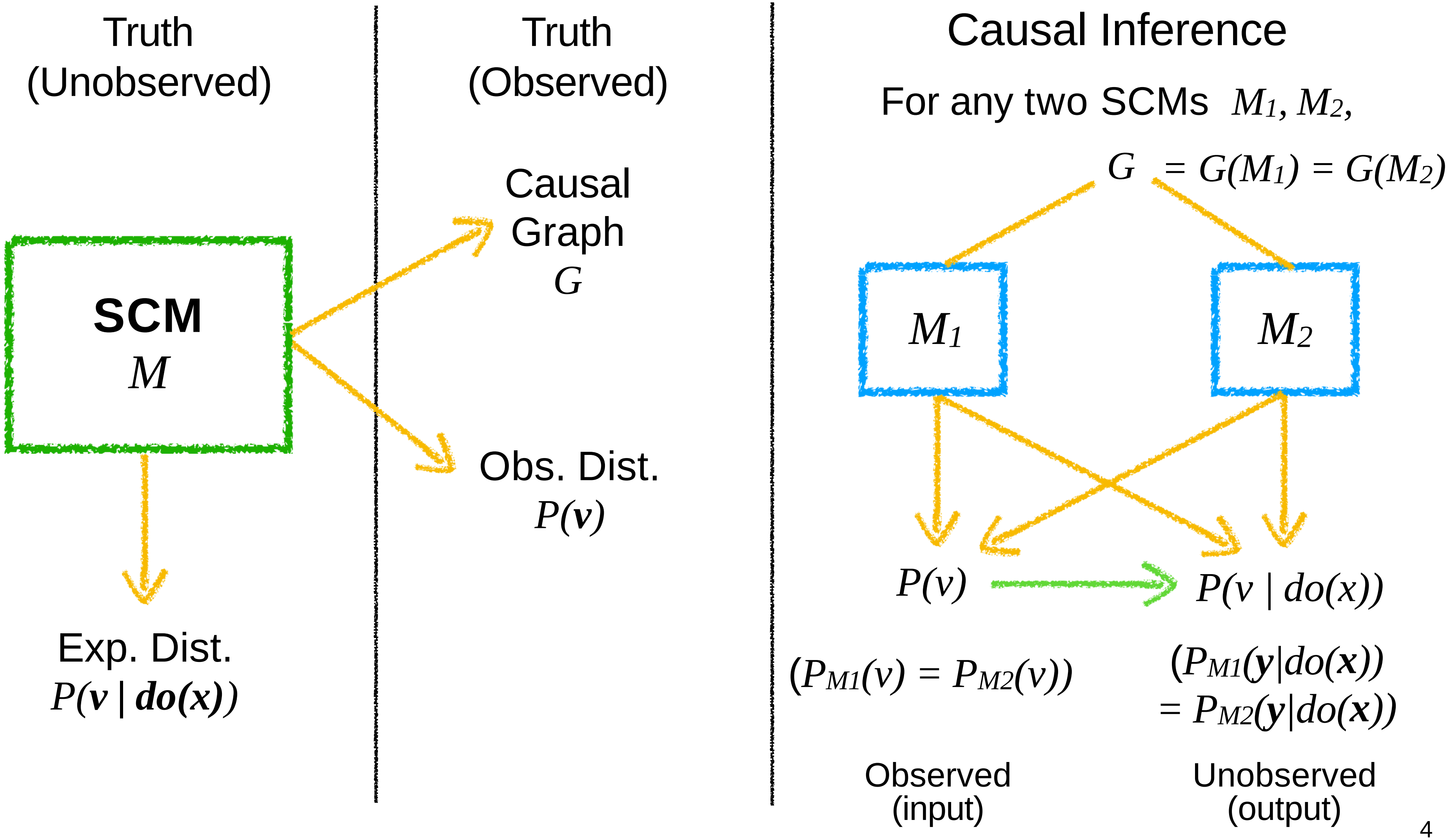
Causal Effect Identifiability (Def. 3.2.2)

The causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ is said to be identifiable from a causal diagram G if the quantity $P(\mathbf{y} \mid do(\mathbf{x}))$ can be computed uniquely from a positive probability of the observed variables.

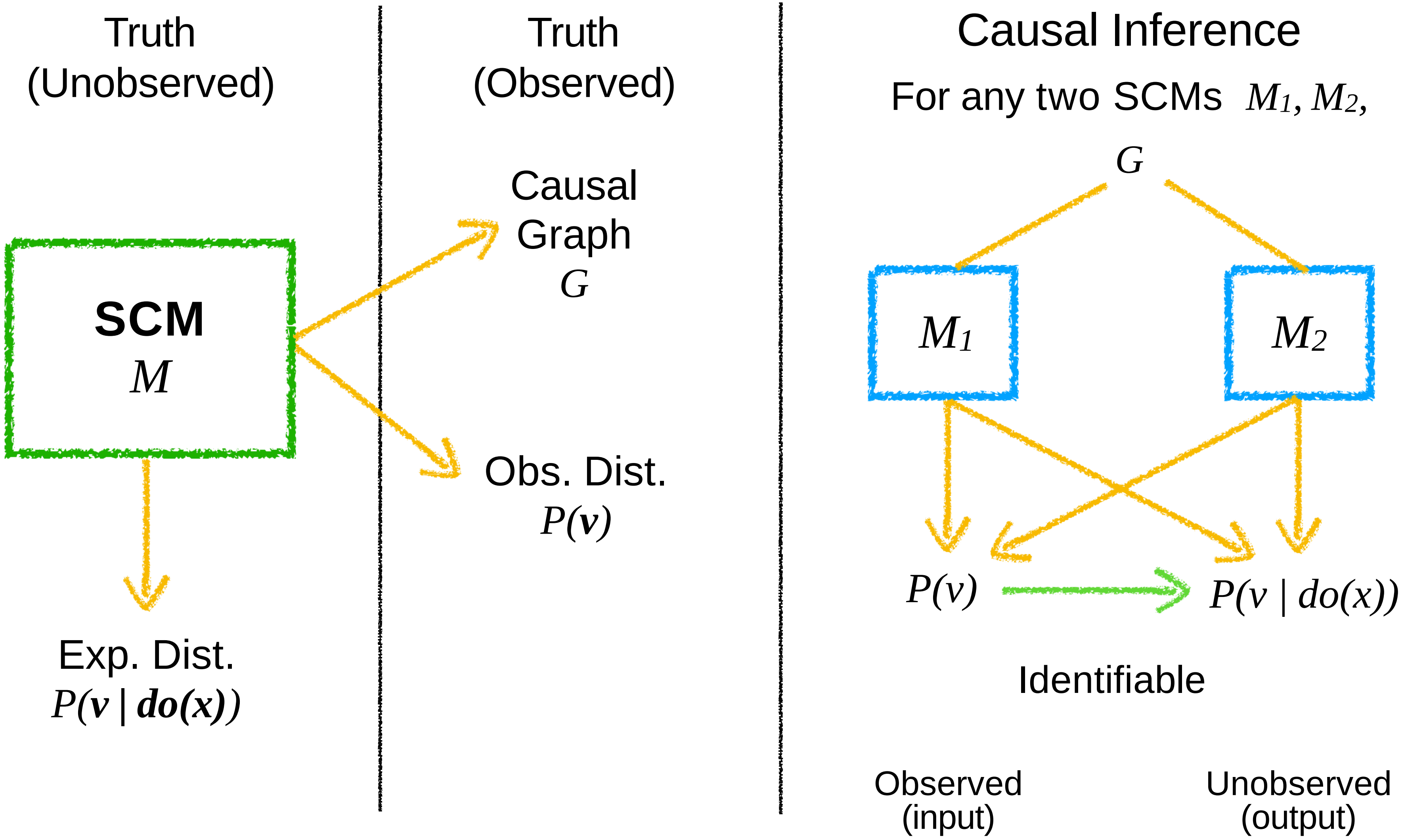
That is, if for every pair of models M_1 and M_2 inducing G, $P_{M_1}(\mathbf{y} \mid do(\mathbf{x})) = P_{M_2}(\mathbf{y} \mid do(\mathbf{x}))$, whenever $P_{M_1}(\mathbf{v}) = P_{M_2}(\mathbf{v}) > 0$.

Bareiboim slides

The Identification Problem (II)



The Identification Problem (II)



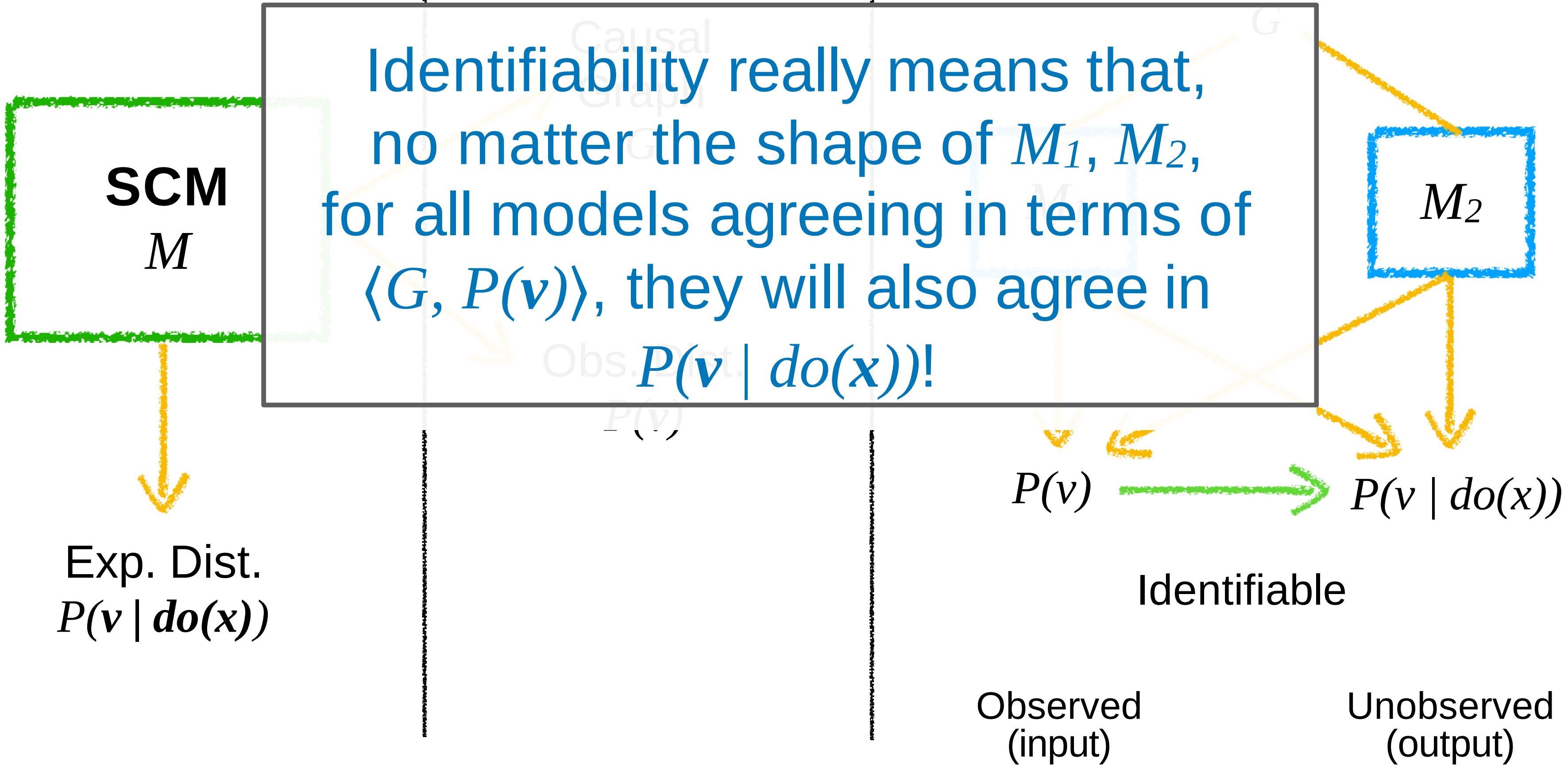
The Identification Problem (II)

Truth (Unobserved)

Truth (Observed)

Causal Inference

For any two SCMs M_1, M_2,

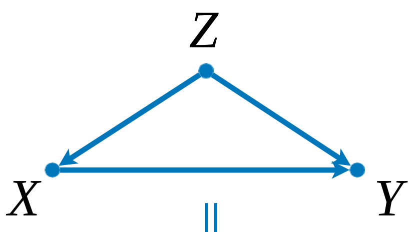


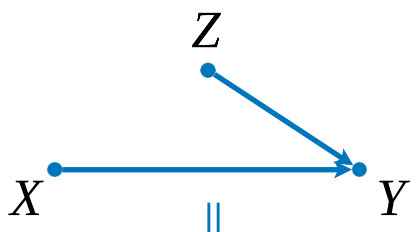
Identifiability really means that, no matter the shape of M_1, M_2, for all models agreeing in terms of $\langle G, P(\mathbf{v})\rangle$, they will also agree in $P(\mathbf{v} \mid do(\mathbf{x}))$!

4
5

Example. Identifiable Effect

- Consider any two pair of models compatible with the following graph and the same observational distribution $P(\mathbf{v})$:





$\Downarrow$ $\qquad$ $\Downarrow$

$P(\mathbf{v}) = P(z)P(x|z)P(y|x,z)$ $\qquad$ $P(\mathbf{v}|do(x)) = P(z)P(y|x,z) \implies P(y \mid do(x)) = \sum_z P(z)P(y \mid x,z)$

$$M^{(1)} = \begin{cases} Z \leftarrow f_z^{(1)}(u_z) \\ X \leftarrow f_x^{(1)}(z, u_x) \\ Y \leftarrow f_y^{(1)}(x, z, u_y) \end{cases} \quad \overset{do(x)}{\longmapsto} \quad M^{(1)} = \begin{cases} Z \leftarrow f_z^{(1)}(u_z) \\ X \leftarrow x \\ Y \leftarrow f_y^{(1)}(x, z, u_y) \end{cases}$$

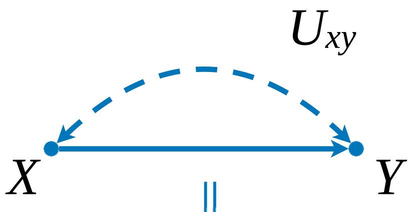
$$M^{(2)} = \begin{cases} Z \leftarrow f_z^{(2)}(u_z) \\ X \leftarrow f_x^{(2)}(z, u_x) \\ Y \leftarrow f_y^{(2)}(x, z, u_y) \end{cases} \qquad M^{(2)} = \begin{cases} Z \leftarrow f_z^{(2)}(u_z) \\ X \leftarrow x \\ Y \leftarrow f_y^{(2)}(x, z, u_y) \end{cases}$$

No matter what the specific functions or P(u) are, as long as M_1, M_2 agree in $\langle G, P(\mathbf{v})\rangle$, they will also agree in $P(z)$ and $P(y|x,z)$, hence in $P(\mathbf{v} \mid do(\mathbf{x}))$!

4
7

Example. Non-identifiable Effect

- Consider the pair of models compatible with the following graph *G* and observational distribution *P(**v**)*:



$\Downarrow$

$$P(\mathbf{v}) = \sum_{u_{xy}} P(y \mid x, u_{xy}) P(x \mid u_{xy}) P(u_{xy})$$

$$M^{(1)} = \begin{cases} X \leftarrow U_{xy} \\ Y \leftarrow (X \oplus U_{xy}) \vee U_y \end{cases}$$

$$M^{(2)} = \begin{cases} X \leftarrow U_{xy} \\ Y \leftarrow U_y \end{cases}$$

$$P^{(j)}(U_i = 1) = 1/2, \quad i = \{x, y, xy\}, j = \{1,2\}$$

U_y	U_{xy}	X	$Y^{(1)}$	$Y^{(2)}$	$P(\mathbf{v},\mathbf{u})$
0	0	0	0	0	1/8
0	1	1	0	0	1/8
1	0	0	1	1	1/8
1	1	1	1	1	1/8
0	0	0	0	0	1/8
0	1	1	0	0	1/8
1	0	0	1	1	1/8
1	1	1	1	1	1/8

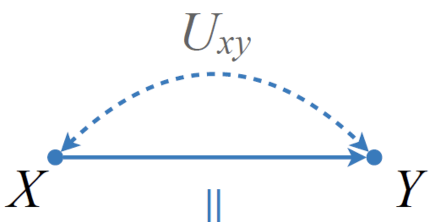
$\Downarrow$

X	$Y^{(1)}$	$Y^{(2)}$	$P^{(1)}(\mathbf{v})$	$P^{(2)}(\mathbf{v})$
0	0	0	1/4	1/4
0	1	1	1/4	1/4
1	0	0	1/4	1/4
1	1	1	1/4	1/4

They match in $P(\mathbf{v})$, that is, $P^{(1)}(\mathbf{v})=P^{(2)}(\mathbf{v})$!

Example. Non-identifiable Effect

- Consider the pair of models compatible with the following graph *G* and observational distribution $P(\mathbf{v})$:



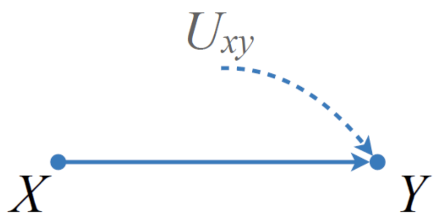
$$P(\mathbf{v}) = \sum_{u_{xy}} P(y \mid x, u_{xy}) P(x \mid u_{xy}) P(u_{xy})$$

$$M^{(1)} = \begin{cases} X \leftarrow U_{xy} \\ Y \leftarrow (X \oplus U_{xy}) \vee U_y \end{cases}$$

$$M^{(2)} = \begin{cases} X \leftarrow U_{xy} \\ Y \leftarrow U_y \end{cases}$$

$$P^{(j)}(U_i = 1) = 1/2, \quad i = \{x, y, xy\}, j = \{1,2\}$$

$$\overset{do(x)}{\longmapsto}$$



$$P(\mathbf{v} \mid do(x)) = \sum_{u_{xy}} P(y \mid x, u_{xy}) P(u_{xy})$$

$$M^{(1)} = \begin{cases} X \leftarrow x \\ Y \leftarrow (X \oplus U_{xy}) \vee U_y \end{cases}$$

$$M^{(2)} = \begin{cases} X \leftarrow x \\ Y \leftarrow U_y \end{cases}$$

U_y	U_{xy}	X	$Y^{(1)}$	$Y^{(2)}$	$P(\mathbf{v},\mathbf{u} \mid do(x))$
0	0	x	$0 \oplus x$	0	1/8
0	1	x	$1 \oplus x$	0	1/8
1	0	x	1	1	1/8
1	1	x	1	1	1/8
0	0	x	$0 \oplus x$	0	1/8
0	1	x	$1 \oplus x$	0	1/8
1	0	x	1	1	1/8
1	1	x	1	1	1/8

$$\Downarrow$$

$Y^{(1)}$	$Y^{(2)}$	$P^{(1)}(\mathbf{y} \mid do(x))$	$P^{(2)}(\mathbf{y} \mid do(x))$
0	0	1/4	1/2
1	1	3/4	1/2

Even though both models induce G and have the same $P(\mathbf{v})$, the effect $P^{(1)}(y \mid do(x)) \neq P^{(2)}(y \mid do(x))$!

Let’s study how to ***decide*** whether a causal effect is ***identifiable***…

Identification in Markovian Models

Theorem. Given the causal diagram G of any Markovian model that all variables are measured, the causal effect $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ is identifiable for every subsets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ and is obtained from the truncated factorization, i.e.,

$$P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{\{V_i \in \mathbf{V} \setminus \mathbf{X}\}} P\left(v_i \mid pa_i\right)$$

Sum over all variables not in $\boldsymbol{X} \cup \boldsymbol{Y}$

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{v} \setminus (\mathbf{x} \cup \mathbf{y})} \prod_{\{V_i \in \mathbf{V} \setminus \mathbf{X}\}} P\left(v_i \mid pa_i\right)$$

Adjustment by Direct Parents

Thm. Given a causal diagram G of any Markovian system, the causal quantity $Q = P(\mathbf{y} \mid do(\mathbf{x}))$ is identifiable whenever $\{\boldsymbol{X}, \boldsymbol{Y}, Pa_x\} \subseteq \boldsymbol{V}$, that is, whenever $\boldsymbol{X}$, $\boldsymbol{Y}$, and all the parents of variables $\boldsymbol{X}$ are measured. The expression of Q is then obtained by adjustment for PA_x, or

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{pa_{\mathbf{x}}} P\left(\mathbf{y} \mid \mathbf{x}, pa_{\mathbf{x}}\right) P\left(pa_{\mathbf{x}}\right)$$

Quiz: 1) derive from previous slide
2) derive for non-Markovian models

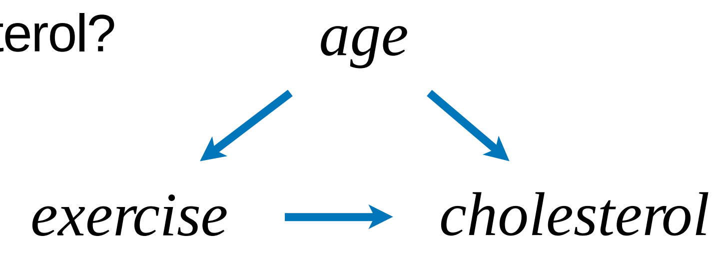
How could adjustment help in real data analysis?

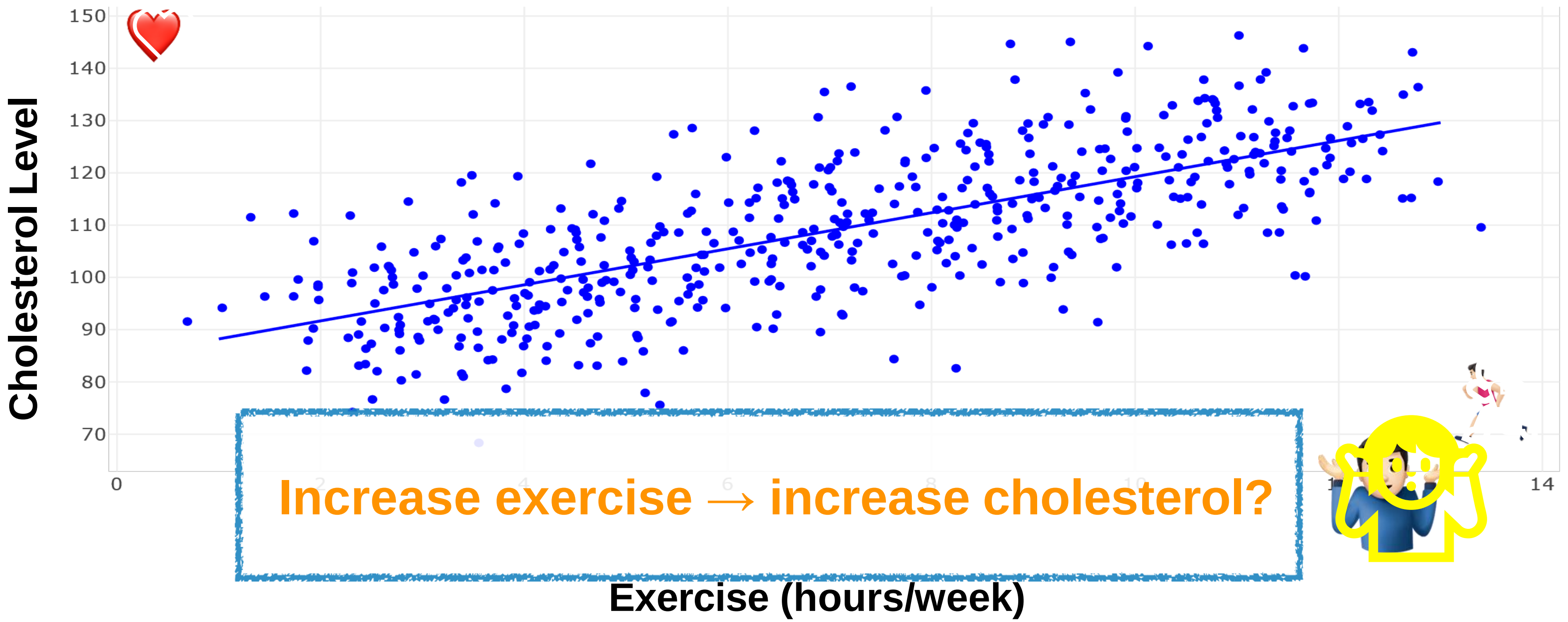
(The Problem of Confounding)

Confounding Bias

What's the causal effect of Exercise on Cholesterol?

What about *P(cholesterol | exercise)* ?

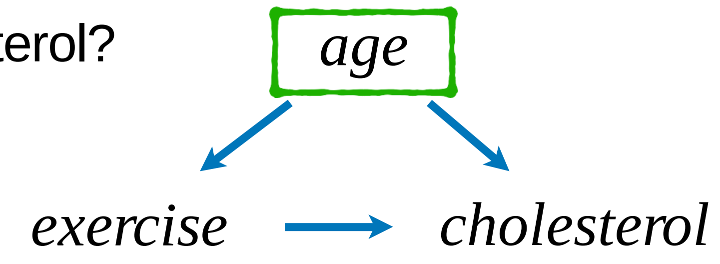


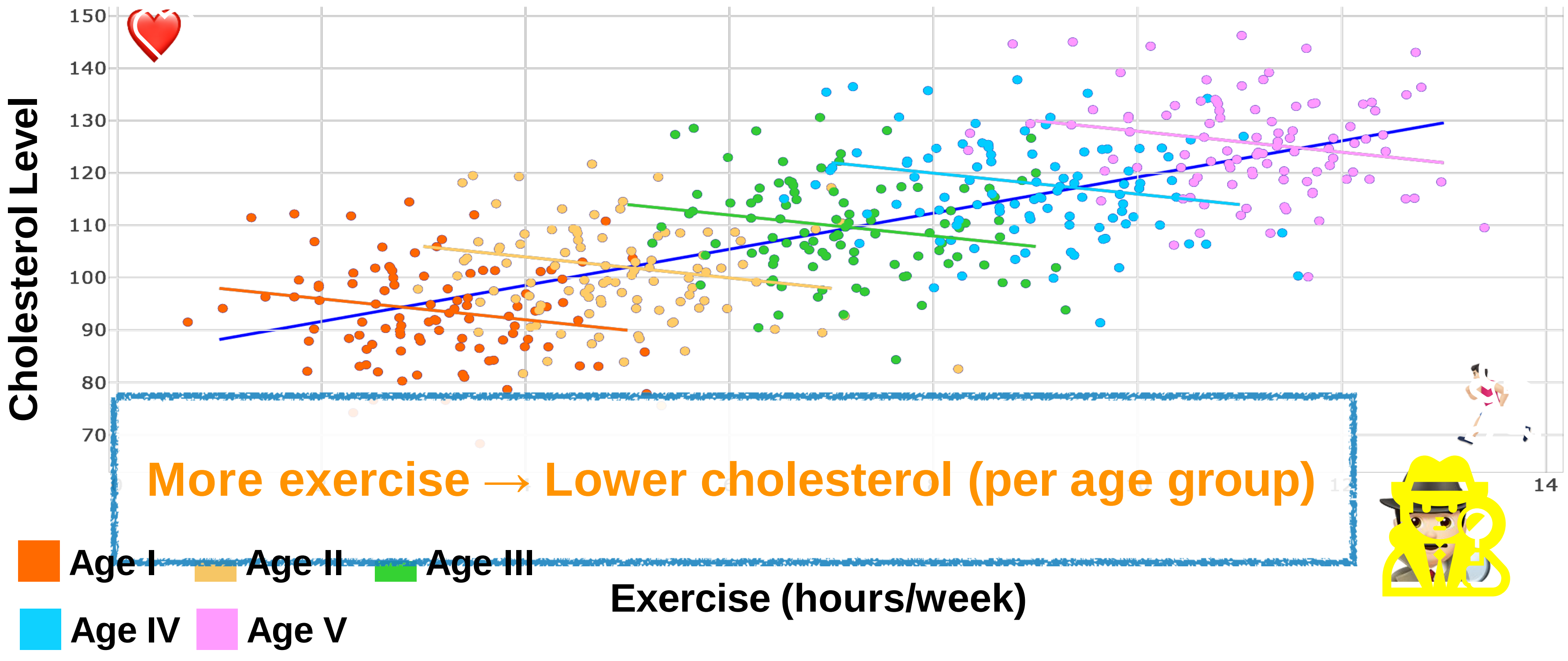


Confounding Bias

What's the causal effect of Exercise on Cholesterol?

What about *P(cholesterol | exercise)* ?

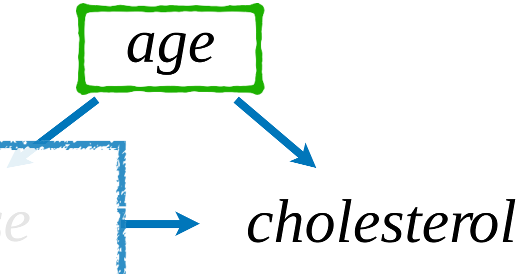




Confounding Bias

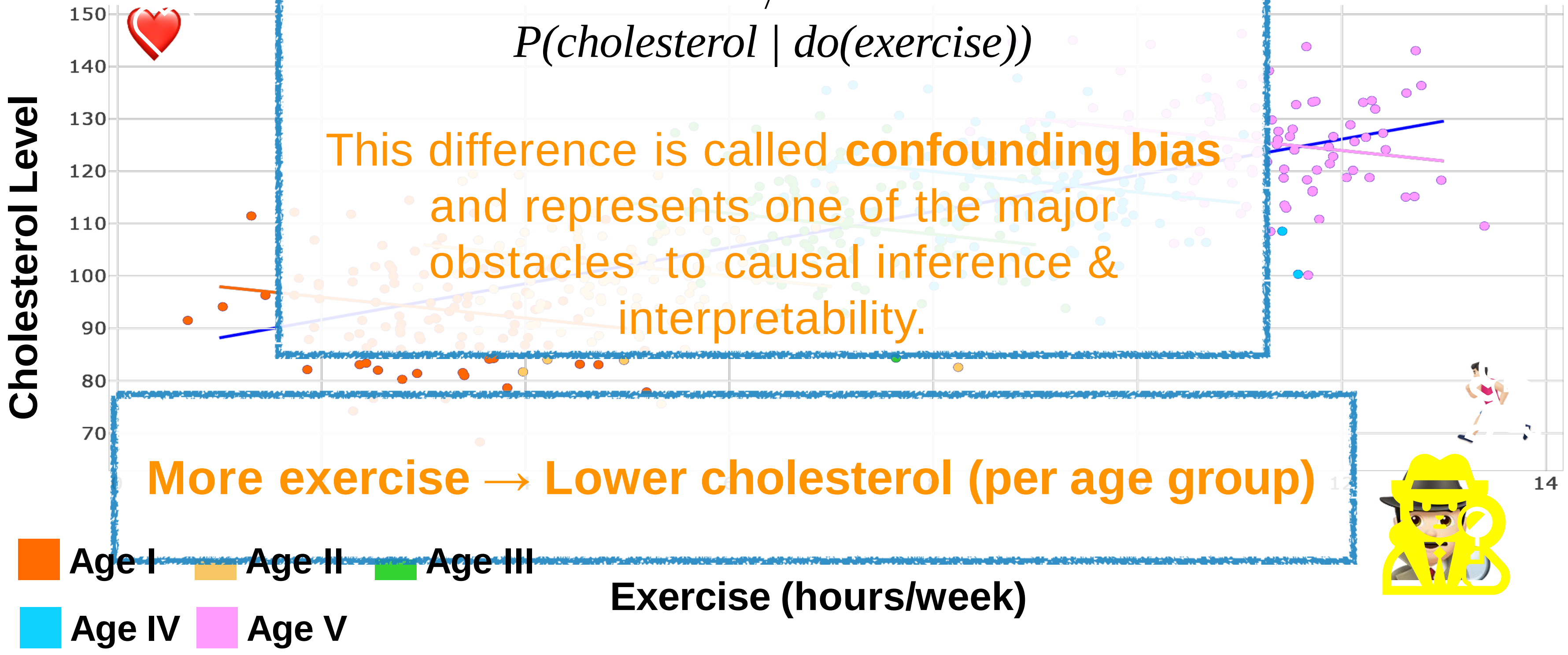
What's the causal effect of Exercise on Cholesterol?

What about *P(cholesterol | exercise)* ?



$$P(cholesterol \mid exercise) \neq P(cholesterol \mid do(exercise))$$

This difference is called **confounding bias** and represents one of the major obstacles to causal inference & interpretability.



More exercise → Lower cholesterol (per age group)

If Season is latent,
is the effect still computable?

Queries:

$Q_2 = P(wet \mid do(Sprinkler = on))$

$$= \frac{\sum_{se,ra} P(we \mid Sp = on, ra) P(Sp = on) P(ra \mid se) P(se)}{P(Sp = on)}$$

equal to 1

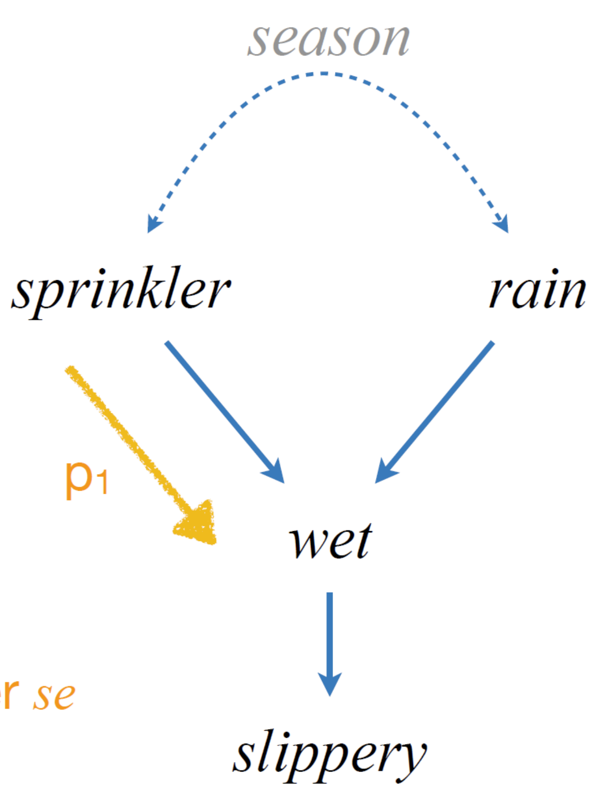
$$= \sum_{se,ra} P(we \mid Sp = on, ra) P(ra \mid se) P(se)$$

$$= \sum_{se,ra} P(we \mid Sp = on, ra) P(ra, se) \quad \text{chain rule}$$

$$= \sum_{ra} P(we \mid Sp = on, ra) \sum_{se} P(ra, se) \quad \text{summing over } se$$

$$= \sum_{ra} P(we \mid Sp = on, ra) P(ra)$$

Adjustment by Rain!



If Season is latent, is the effect still computable?

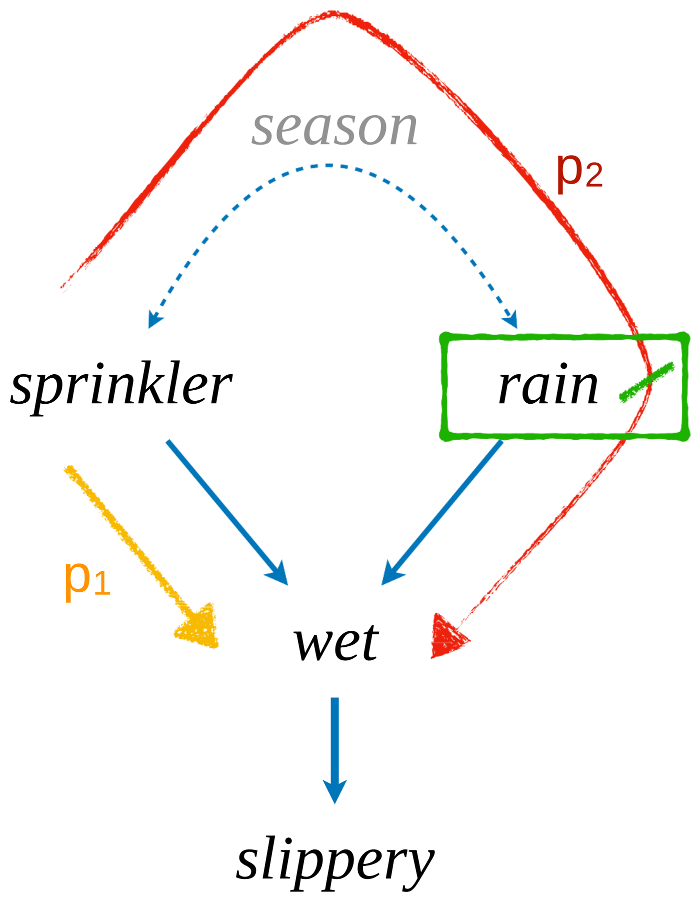
Queries:

$\mathbf{Q_2} = P(wet \mid do(Sprinkler = on))$

$$= \sum_{ra} P(we \mid Sp = on, ra)P(ra)$$

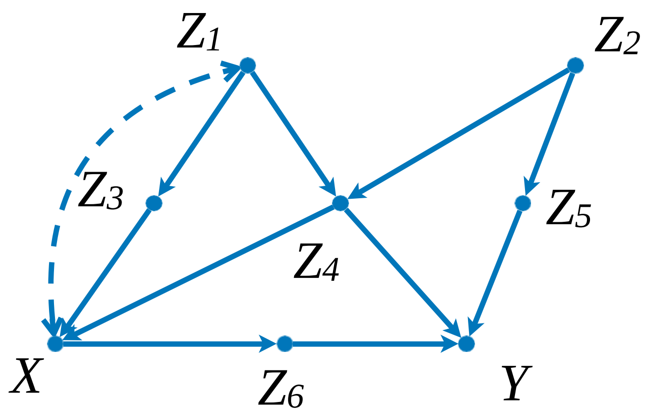
By conditioning on rain,

- p_2 (the non-causal path) is blocked, and
- p_1 (the causal path) remains unaffected!



Is Confounding Bias removable?

Goal: Find the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$, $Q = P(\mathbf{y}|do(\mathbf{x}))$, given measurements on variables $Z_1,\ldots, Z_k$,



where some of $\boldsymbol{X}$ parents are unobserved.

How can the target quantity Q be identified if only a subset of the parents is measured?

Outline (chapter 3)

- The semantic of Intervention in SCM, the do operators
- How to determine P(Y|do(x)) given an SCM
- The identification problem
- The back door criterion and the adjustment formula

Answer:
The Back-door Criterion

Definition 3.3.1 (Back-door Criterion)

A set $\boldsymbol{Z}$ satisfies the back-door criterion (bdc) w.r.t. to a pair of variables $\boldsymbol{X},\boldsymbol{Y}$ in a causal diagram G if:

(i) no node in $\boldsymbol{Z}$ is a descendent of $\boldsymbol{X}$; and

(ii) $\boldsymbol{Z}$ blocks every path between $\boldsymbol{X}$ and $\boldsymbol{Y}$ that contains an arrow into $\boldsymbol{X}$.

The Back-door Adjustment

Theorem 3.3.2 (Back-door Adjustment)

If a set $\boldsymbol{Z}$ satisfies the bdc w.r.t the pair $\boldsymbol{X}, \boldsymbol{Y}$, the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ is identifiable and given by:

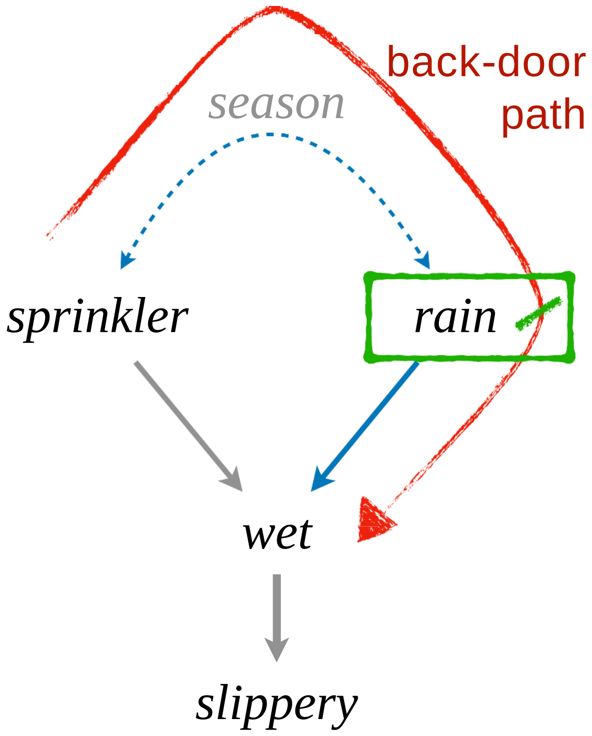
$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z})$$

Back-Door Sets as Substitutes of the Direct Parents of $\boldsymbol{X}$

Rain satisfies the back-door criterion relative to *Sprinkler* and *Wet*:

(i) *Rain* is not a descendant of *Sprinkler*, and

(ii) *Rain* blocks the only back-door path from *Sprinkler* to *Wet*.



Adjusting for the direct parents of *Sprinkler*, we have:

Direct derivation, showing it works

$$P(we \mid do(sp)) = \sum_{se} P(we \mid sp, se) P(se)$$

$$= \sum_{se,ra} P(we \mid sp, se, ra) P(ra \mid sp, se) P(se)$$

$(Sp \perp\!\!\!\perp Ra \mid Se)$

$(We \perp\!\!\!\perp Se \mid Ra, Sp)$

$$= \sum_{se,ra} P(we \mid sp, ra) P(ra \mid se) P(se)$$

$$= \sum_{ra} P(we \mid sp, ra) \sum_{se} P(ra, se) = \boxed{\sum_{ra} P(we \mid sp, ra) P(ra)}$$

Adjustment by *Rain*

Adjustment by Direct Parents → Back-door Adjustment

More Generally:

(i) no node in $\boldsymbol{Z}$ is a descendent of $\boldsymbol{X}$; and ⟹ $(X \perp\!\!\!\perp Z \mid Pa_x)$

(ii) $\boldsymbol{Z}$ blocks every path between $\boldsymbol{X}$ and $\boldsymbol{Y}$ that contains an arrow into $\boldsymbol{X}$. ⟹ $(Y \perp\!\!\!\perp Pa_x \mid Z, X)$

Then:

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{pa_{\mathbf{x}}} P(\mathbf{y} \mid \mathbf{x}, pa_{\mathbf{x}}) P(pa_{\mathbf{x}})$$

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

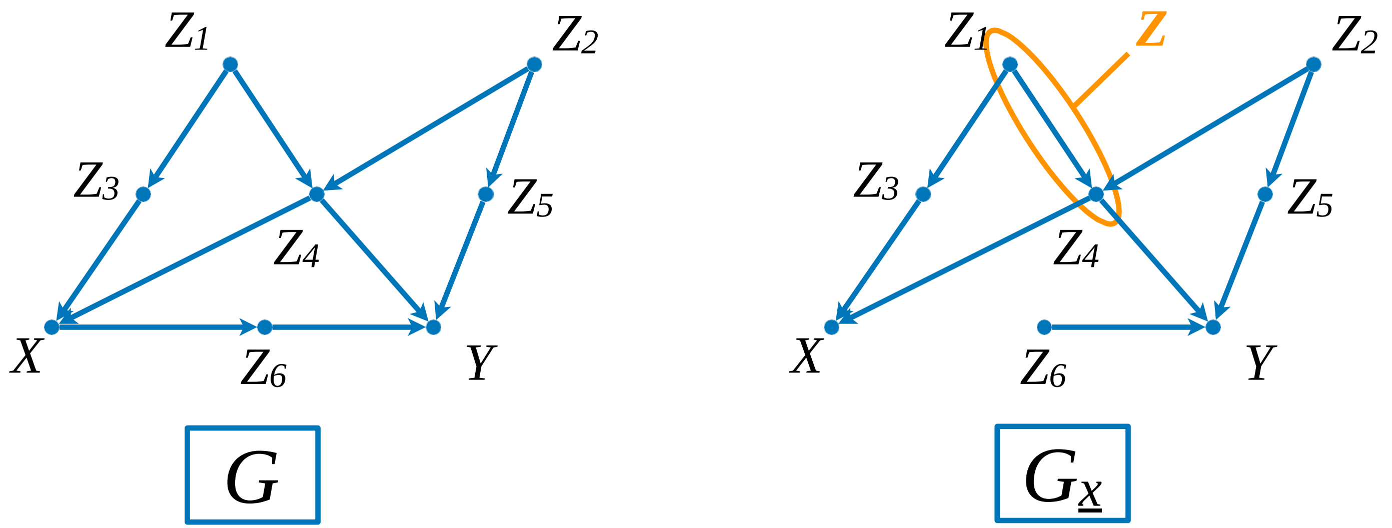
$$= \sum_{\mathbf{z}, pa_{\mathbf{x}}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z} \mid pa_{\mathbf{x}}) P(pa_{\mathbf{x}})$$

$$= \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \sum_{pa_{\mathbf{x}}} P(\mathbf{z}, pa_{\mathbf{x}}) = \boxed{\sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z})}$$

Adjustment by $\boldsymbol{Z}$ is equivalent to adjustment by direct parents whenever $\boldsymbol{Z}$ is bd-admissible!

How do we find these bd-sets? Graphical Condition

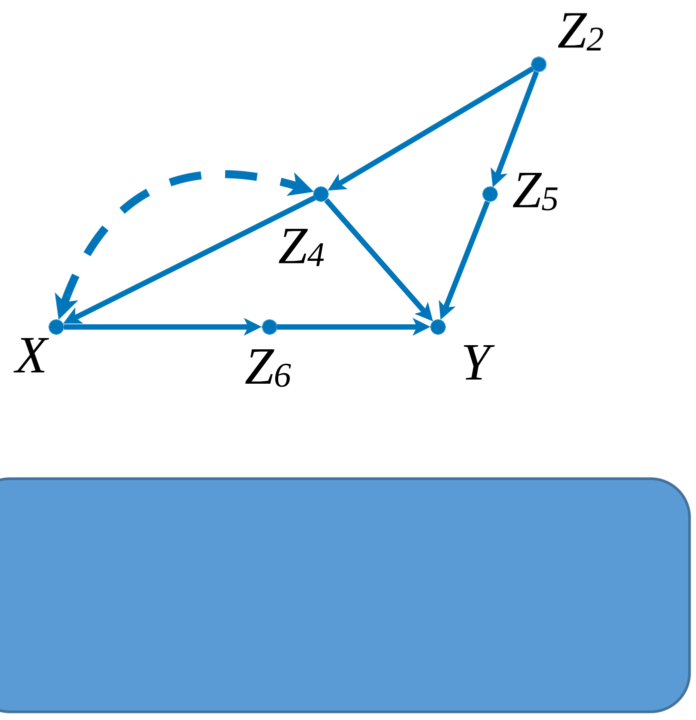
$P(y \mid do(x))$ is identifiable if there is a set $\boldsymbol{Z}$ that d-separates $\boldsymbol{X}$ from $\boldsymbol{Y}$ in $G_{\underline{X}}$ (the graph G where all arrows emanating from X are removed.)

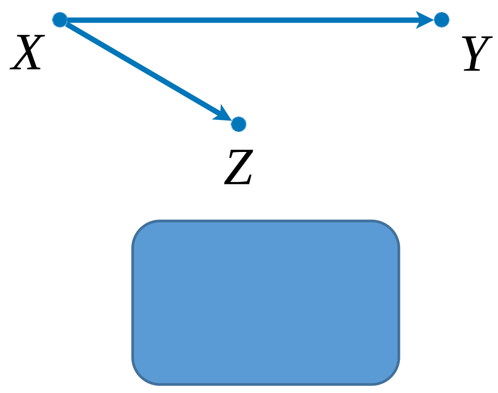


$$P(y \mid do(x)) = \sum_{z_1, z_4} P(y \mid x, z_1, z_4) P(z_1, z_4)$$

Back-door Examples

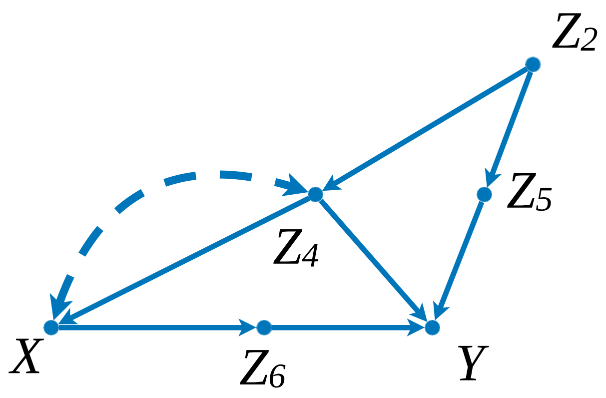
Are there admissible back-door sets (relative to $\boldsymbol{X}, \boldsymbol{Y}$) for the following graphs?

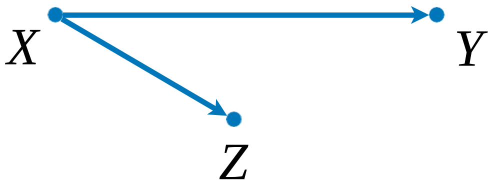




Back-door Examples

Are there admissible back-door sets (relative to $\boldsymbol{X},\boldsymbol{Y}$) for the following graphs?





$\mathbf{Z} = \{Z_4, Z_2\}, \{Z_4, Z_5\}, \{Z_4, Z_2, Z_5\}$

$\mathbf{Z} = \varnothing$

Recaping The Backdoor Criterion

[redacted] Under what conditions does a causal story permit us to compute the causal effect of one variable on another, from data obtained by passive observations, with no interventions? Since we have decided to represent causal stories with graphs, the question becomes a graph-theoretical problem: Under what conditions is the structure of the causal graph sufficient for computing a causal effect from a given data set?

3.3 The Backdoor Criterion

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

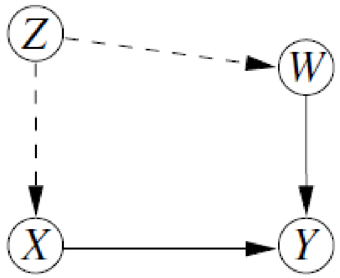
If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Rationale:

1. We block all spurious paths between X and Y.
2. We leave all directed paths from X to Y unperturbed.
3. We create no new spurious paths.

More Examples for Backdoors



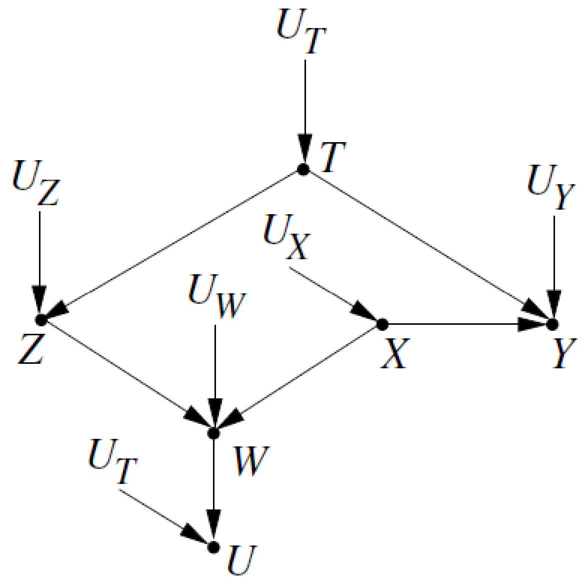
P(Y|do(X))?

Figure 3.6: A graphical model representing the relationship between a new drug (X), recovery (Y), weight (W), and an unmeasured variable Z (socioeconomic status)

W is a backdoor. Therefore we can compute:

$$P(Y = y|do(X = x)) = \sum_{w} P(Y = y|X = x, W = w)P(W = w)$$

Examples



P(Y|do(X))?

No backdoors between X and Y and therefore: P(Y|do(X))= P(Y|X)

What if we adjust for W? … wrong!!!

But what if we want to determine P(Y|do(X),w)? What do we do with the spurious path $X \rightarrow W \leftarrow Z$ <- $T \rightarrow Y$?

if we condition on T, we would block the spurious path $X \rightarrow W \leftarrow Z$ <- $T \rightarrow Y$. We can compute:

$$P(Y=y|do(X=x), W=w) = \sum_t P(Y=y|X=x, W=w, T=t)P(T=t|W=w)$$

Example: W can be post-treatment pain

Adjusting for Colliders?

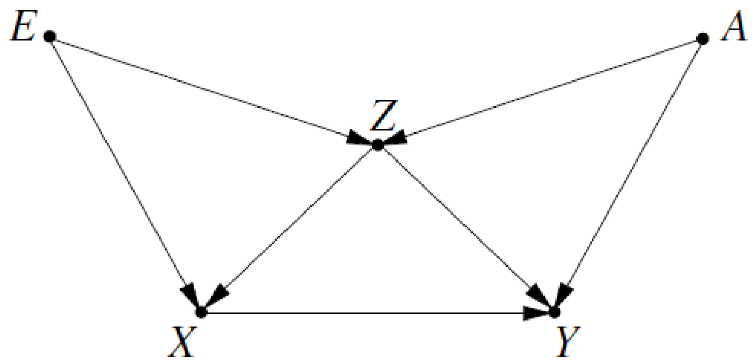


Figure 3.7: A graphical model in which the backdoor criterion requires that we condition on a collider (Z) in order to ascertain the effect of X on Y

There are 4 backdoor paths. We must adjust for Z, and one of E or A or both

Outline (chapter 3)

- The semantic of Intervention in SCM, the do operators
- How to determine P(Y|do(x)) given an SCM
- The identification problem
- The back door criterion and the adjustment formula
- Computing bd: Inverse probability weighting

Evaluating BD adjustment

- The backdoor provides a criterion for deciding *when* a set of covariates $\mathbf{Z}$ is admissible for adjustment, i.e.,

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

- In practice, how should backdoor expressions be evaluated?

- There are sample & computational challenges entailed by the eval. of such expressions since one needs to
  - estimate the different distributions, and
  - evaluate them, summing over a possibly high-dimensional $\mathbf{Z}$ (i.e., time $O(exp(|\mathbf{Z}|))$).

Inverse Probability Weighting (IPW)

- Let's rewrite the bd-expression,

$$P(\mathbf{y} \mid do(\mathbf{X} = \mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z})$$

$$= \sum_{\mathbf{z}} \frac{P(\mathbf{y}, \mathbf{x}, \mathbf{z})}{P(\mathbf{x}, \mathbf{z})} P(\mathbf{z})$$

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

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

Entries of the joint distribution

Fit a function *g(z)* that approximates this probability

Inverse Probability Weighting (IPW)

- Assume we have N samples, then

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} \frac{P(\mathbf{y}, \mathbf{x}, \mathbf{z})}{P(\mathbf{x} \mid \mathbf{z})}$$

$$= \sum_{\mathbf{z}} \frac{\frac{1}{N} \sum_{i=1}^{N} 1_{\mathbf{Y}_i=\mathbf{y}, \mathbf{X}_i=\mathbf{x}, \mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$$

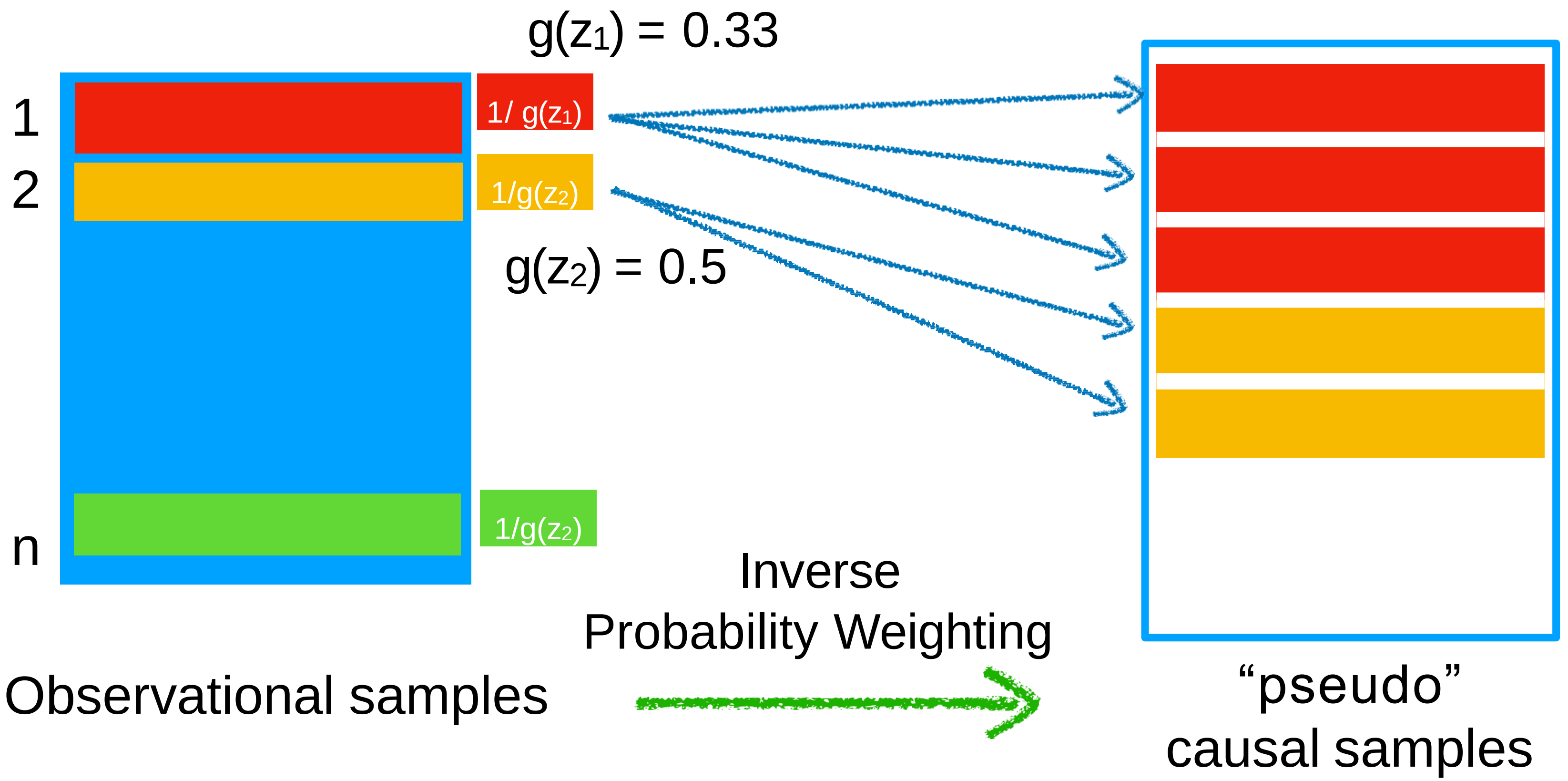
$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathbf{z}} \frac{1_{\mathbf{Y}_i=\mathbf{y}, \mathbf{X}_i=\mathbf{x}, \mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{1_{\mathbf{Y}_i=\mathbf{y}, \mathbf{X}_i=\mathbf{x}, \mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$$

→ Requires time proportional to the number of samples N

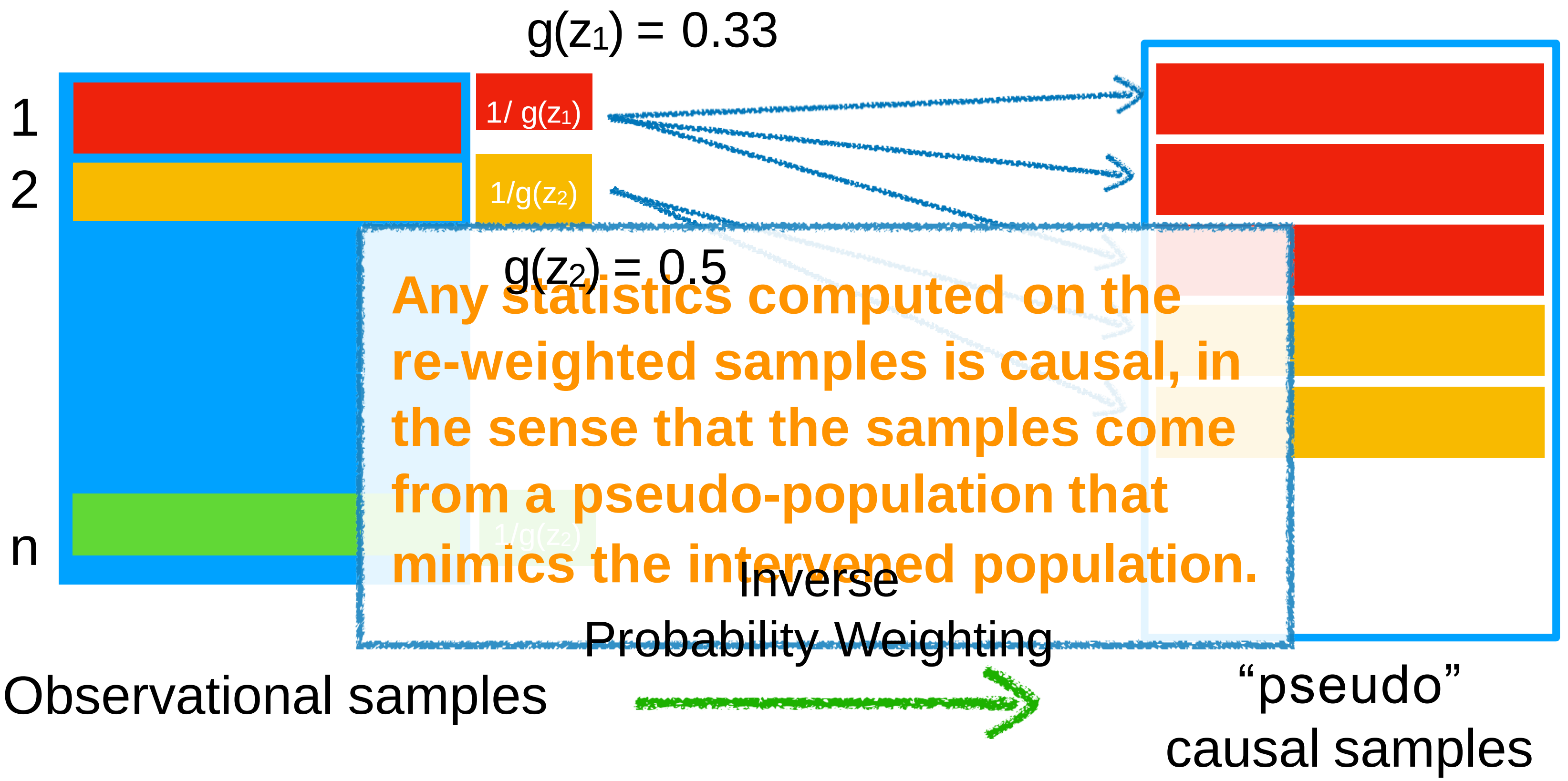
Inverse Probability Weighting (IPW)

- In practice, evaluating the expr. can be seen as: $\frac{1}{N}\sum_{i=1}^{N}\frac{1_{\mathbf{Y}_i=\mathbf{y},\mathbf{X}_i=\mathbf{x},\mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$



Inverse Probability Weighting (IPW)

- In practice, evaluating the expr. can be seen as: $\frac{1}{N}\sum_{i=1}^{N}\frac{1_{\mathbf{Y}_i=\mathbf{y},\mathbf{X}_i=\mathbf{x},\mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$



Any statistics computed on the re-weighted samples is causal, in the sense that the samples come from a pseudo-population that mimics the intervened population.