Causality and Machine Learning

David M. Blei
Departments of Computer Science and Statistics
Data Science Institute
Columbia University

Introduction

- Causality: Understanding the **effects** of a **hypothetical intervention**.

- Example: A father is considering signing his daughter up for “book of the month” club because he thinks she will read more and do better in school.

- Why is this a causal inference? The father is considering an intervention in the world in the hopes of changing an outcome.

- Will signing her up for the club cause more reading? Will more reading cause better grades? (Will better grades cause a happier life?)

- (Causality quickly turns philosophical, but we’ll avoid such rabbit holes.)

Here are some other causal questions.

- If I take this medicine, will my headache go away?
- If I deploy this new recommender, will my website's revenue increase?
- If I implement this education policy, will poverty decrease?
- If I make this change to the tax law, will unemployment decrease?
- If I use this algorithm to hire people, will it satisfy equal opportunities?
- If I show a user this article, will she click on it?
- If I enroll at Columbia, will I make more money?
- If students turn off their phones, will they learn more in class?
- If I stop eating french fries, will my cholesterol go down?
- If I encourage my friends to vote, will they vote?

These questions are about what happens when we intervene in the world.

We can also ask causal questions about the past.

- Would Hillary have won the election had she visited Michigan three days before it, given that she didnt and Trump won.
- Would double panes have prevented my windows from blowing through during the storm, given that I have single-pane windows and they did blow through.
- Would the supermarket have gone out of business had it raised the price of peanut butter instead of bananas, given that it did not go out of business and that it did raise the price of bananas.

Such questions are called counterfactuals. They too are questions about the effects of an intervention, even if it's one that requires a time machine.

- Causality is about understanding the **effects** of a **hypothetical intervention**.
  - “If I take this medicine, will my headache go away?”
  - “Would Hillary had won the election had she gone to Michigan three days before it, given that Trump won the election?”
- It can be a question about the future or the past.
- Note: This is an attitude, a stance, a philosophy.
- There might be others. (But today we will adopt it.)

For contrast, consider these predictive questions:

- Will I get a headache tomorrow?
- What is my website's expected revenue next quarter?
- How much poverty can we expect next year?
- What will the unemployment rate be next year?
- What will a subjects brain activity look like in the next hour?
- Will a particular user find and click on a particular article?
- How much money will a Columbia student make after graduation?
- How much will students learn in this class?

These are the kinds of questions answered by traditional ML/statistics.
They involve passive observation, but no intervention.

- We have intuitions about causality. Can we make them mathematical?
- Can we use data to answer causal questions?

- **Counterfactuals and Causal Inference (2nd edition)**
  S. Morgan and C. Winship (2015, Cambridge University Press)

- **Causality (2nd edition)**
  J. Pearl (2009, Cambridge University Press)

- **Causal Inference in Statistics: A Primer**
  J. Pearl, M. Glymour, and N. Jewell (2016, John Wiley & Sons)

- **The Book of Why**
  J. Pearl and D. MacKenzie (2018, Basic Books)

- **Causal Inference**
  M. Hernan and J. Robins (2019, Chapman & Hall/CRC)

- **Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction**
  G. Imbens and D. Rubin (2015, Cambridge University Press)

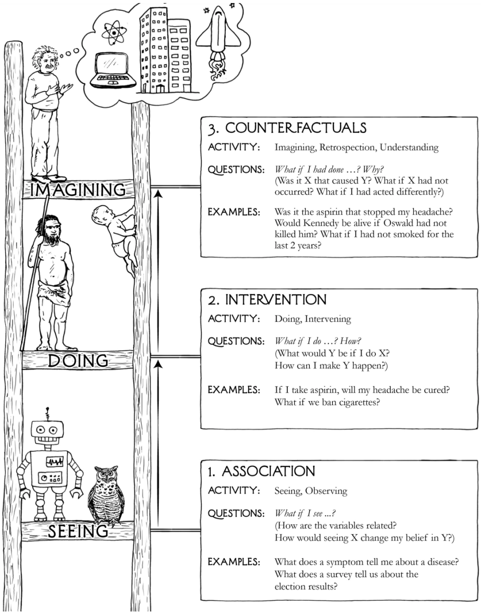
- **Elements of Causal Inference : Foundations and Learning Algorithms**
  J. Peters, D. Janzing, and B. Schoelkopf (2018, The MIT Press)

- **Advanced Data Analysis from an Elementary Point of View**
  C. Shalizi (2020, in preparation)

The science of causality discusses a **ladder of causation**.

- Is excercise correlated with good health? (observation/association)
- Will my health improve if I exercise? (intervention)
- Would I have still caught a cold on my trip had I exercised, given that I did catch a cold and I didn't exercise? (imagination)

Level 1 is classical ML/statistics. Levels 2 and 3 involve causality.



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

from Pearl and MacKenzie (2018)

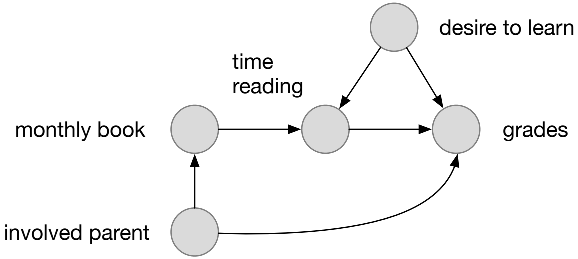
Topics in causality we will discuss

- Causal graphical models
- Identification with backdoor adjustment
- Estimation
  - General adjustment
  - Regression
  - Matching
  - Inverse propensity weights
- Structural causal models and counterfactuals
- Potential outcomes (a little)

Topics in causality we will not discuss

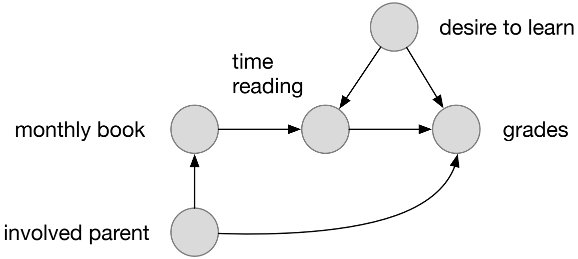
Structure learning, Instrumental variables, Negative controls, Front-door adjustment, Double machine learning, Sensitivity analysis, Multiple causality, Time-dependent treatments, Regression discontinuities, Causal forests, Invariant risk minimization, Experimental design, Single-world intervention graphs, Mediators and moderators, Overcoming selection bias, Reinforcement learning, Sufficiency of the propensity score, and more...

Causal Graphical Models



- A **causal graphical model (CGM)** is a directed acyclic graph (DAG) where
  - nodes represent variables
  - edges indicate a direct causal relationship
  - causal relationships are probabilistic

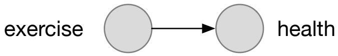
- CGMs help reason about the effects of interventions.



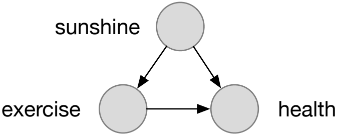
- Some rules:
  - Any common cause of two or more variables is in the graph.
  - If there is an edge then there is a direct cause.

- Causal graphs make assumptions about the world.
  - Notice: missing edges → nodes without common causes

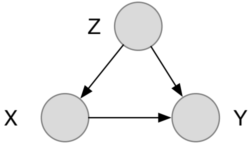
- For now, assume every variable is observed.



- Here is a simpler example.
- exercise is a direct cause of health

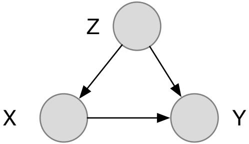


- But suppose sunshine is a common cause of exercise and health.
- It must be included in the graph.

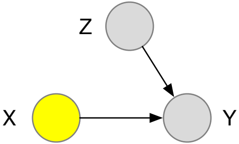


- In a CGM, the direct causal relationships are **probabilistic**.
- Each variable is drawn from a conditional given its parents, $\mathrm{p}(v_i \mid \mathrm{pa}(v_i))$.
- Here,

$$\mathrm{p}(x, y, z) = \mathrm{p}(z)\mathrm{p}(x \mid z)\mathrm{p}(y \mid x, z).$$

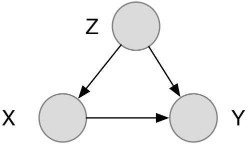


observation model

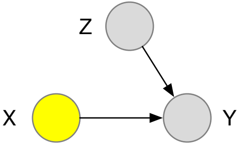


intervene on x

- What is an **intervention**?
- It is a “graph surgery.”
  - Remove the edges to the parents of the intervened node.
  - E.g., intervention of $X = \bar{x}$ removes $Z \rightarrow X$
  - The post-surgery graph is also a CGM; it is denoted “$\mathrm{do}(\bar{x})$.”
- A causal inference asks about $\mathrm{p}(y\,;\,\mathrm{do}(\bar{x}))$.



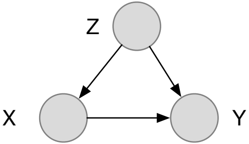
observation model



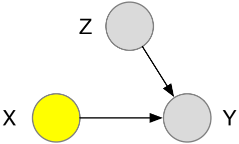
intervene on x

- Notice: We just assumed **modularity**.
- On intervention we only changed how X obtains its value.
- The other causal relationships do not change.
- We can also define “direct cause.” In a CGM,

$$\mathrm{p}(v_i \,|\, \mathrm{pa}(v_i)) = \mathrm{p}(v_i \,|\, \mathrm{do}(\mathrm{pa}(v_i)))$$

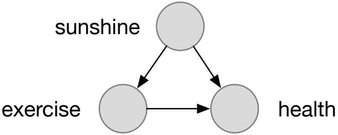


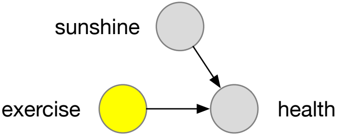
observation model

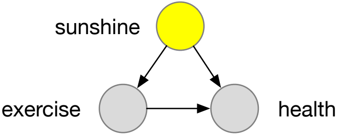


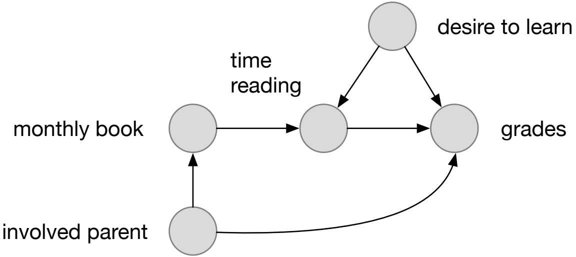
intervene on x

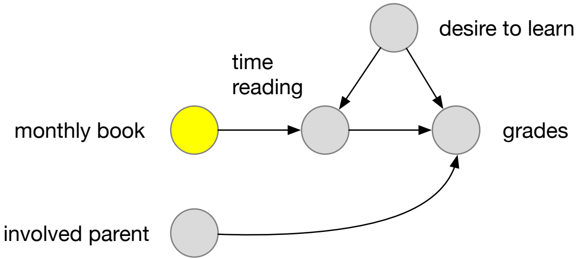
- Let’s consider some interventions in the example graphs.
- Think about how (and if) the distribution of other variables changes.
- Think about the assumption of modularity.

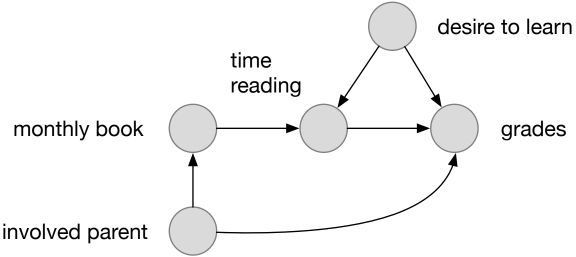

sunshine
exercise
health

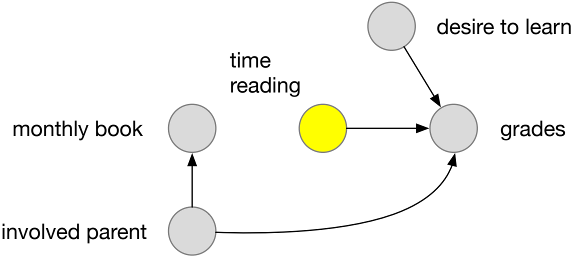

sunshine
exercise
health

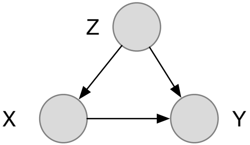

sunshine
exercise
health


desire to learn
time
reading
monthly book
grades
involved parent

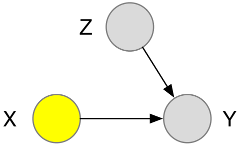

desire to learn
time
reading
monthly book
grades
involved parent


desire to learn
time
reading
monthly book
grades
involved parent


desire to learn
time
reading
monthly book
grades
involved parent

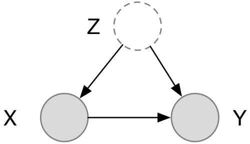


observation model

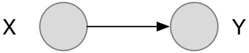


intervene on x

- How is a **causal GM** different from a **probabilistic GM**?
- Both provide a factorized family of joint distributions.
  - indexed by the conditional distributions at each node
- But a PGM can factorize the joint in a way that does not reflect causality.
- Pearl puts it well:
  - PGMs are carriers of independence.
  - CGMs are oracles for interventions.

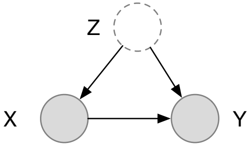


an unobservable common cause

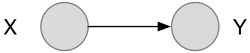


accurate PGM of observations

- Let’s dig into the difference between a CGM and PGM.
- Suppose, in the true model, Z is a common cause of X and Y.
- But suppose Z is unobservable. We only observe X and Y.
  - (Notice the notation for an unobservable variable.)

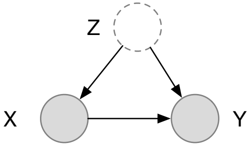


an unobservable common cause

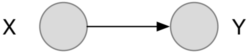


accurate PGM of observations

- The PGM that omits Z provides a perfectly good distribution of the joint.
  - All possible joints $\mathrm{p}(x, y)$ can be represented in this way.
- But it is not an accurate causal GM.
- Why? It does not contain the common cause of X and Y.



an unobservable common cause



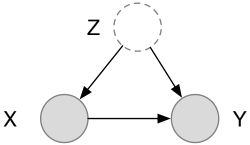
accurate PGM of observations

- If treated as a CGM, the 2-node model equates intervening and conditioning,

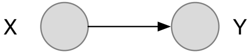
$$p_2(y\,;\,\mathrm{do}(\bar{x})) = p(y \,|\, x) = \int p(y \,|\, x, z)\, p(z \,|\, x)\, \mathrm{d}z.$$

- But this does not reflect reality. The real interventional distribution is

$$p(y\,;\,\mathrm{do}(\bar{x})) = \int p(y \,|\, x, z)\, p(z)\, \mathrm{d}z.$$

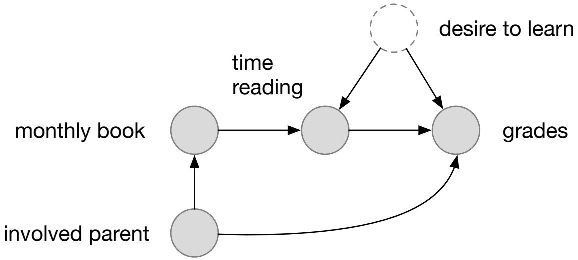


an unobservable common cause



accurate PGM of observations

- To reiterate: On the right is a perfectly good PGM.
- But it does not posit the (unobservable) common cause.
- If treated causally, it makes incorrect predictions about the intervention.

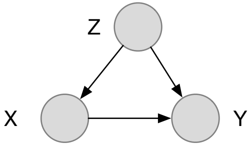


- Classical ML needs “only” to posit a valid statistical model of the data.

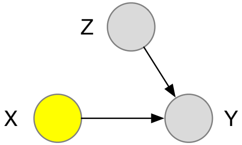
- Causality requires common causes between variables, even if unobservable.

- Unobserved common causes can indicate when CI is not possible.

- Omitting a common cause from the model can lead to confounding.



observation model



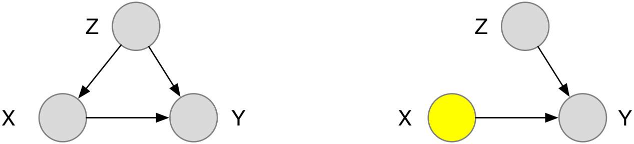
intervene on x

- Consider again the true causal model, and assume everything is observable.

- We saw the interventional distribution is

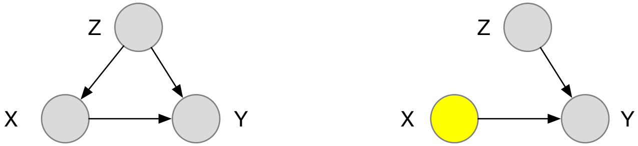
$$\mathrm{p}(y\,;\,\mathrm{do}(\bar{x})) = \int \mathrm{p}(y \mid x, z)\,\mathrm{p}(z)\,\mathrm{d}z$$

- **Identification**: We wrote the result of an intervention in terms of the distribution of observations.



$$\mathrm{p}(y\,;\,\mathrm{do}(x)) = \int \mathrm{p}(y \mid x, z)\,\mathrm{p}(z)\,\mathrm{d}z$$

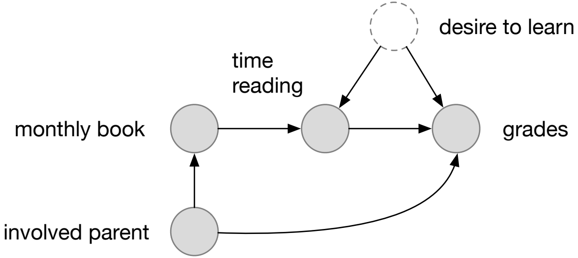
- Identification helps link causal inferences to observational data.
  - Consider a dataset $\{(x_i, y_i, z_i)\}_{i=1}^{n}$ from the causal model.
  - Use it to estimate $\mathrm{p}(z)$ and $\mathrm{p}(y \mid x, z)$.
  - Then use those estimates in the expression for $\mathrm{p}(y\,;\,\mathrm{do}(\bar{x}))$.
- **Estimation**: ML/statistics/data science meets causality



$$p(y\,;\,\mathrm{do}(x)) = \int p(y \mid x, z)\, p(z)\, \mathrm{d}z$$

- Example:
  - Collect data $\{(x_i, y_i, z_i)\}_{i=1}^{n}$.
  - fit a function $\hat{f}(x, z)$ to estimate $p(y \mid x, z)$.
  - use Monte Carlo to estimate $p(z)$.
- Then estimate the intervention distribution

$$p(y\,;\,\mathrm{do}(\bar{x})) \approx \frac{1}{n} \sum_{i=1}^{n} \hat{f}(\bar{x}, z_i)$$



- **Domain knowledge**

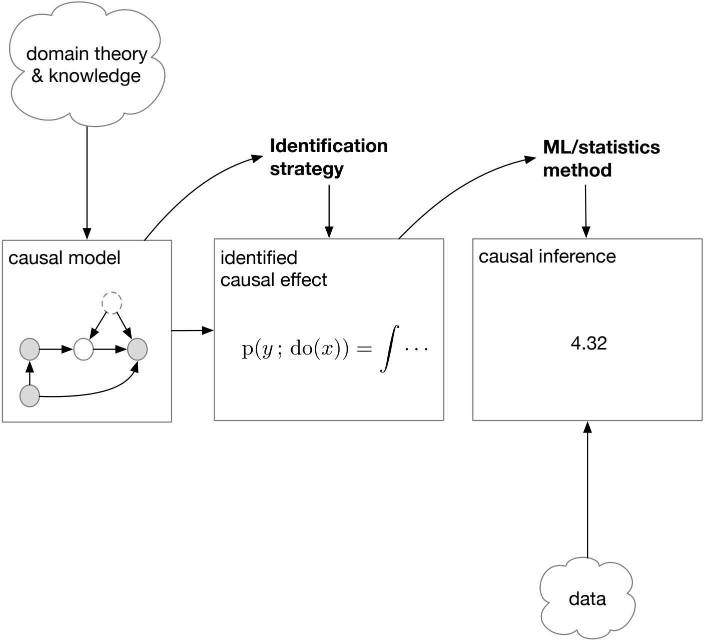
  provides a model, including what is observable and unobservable.

- **Identification strategies**

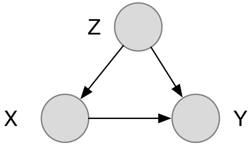
  express a causal quantity of interest in terms of the observational distribution.

- **Estimation methods**

  use an identification formula and data to approximate the causal quantity.


domain theory & knowledge
Identification strategy
ML/statistics method
causal model
identified causal effect
$\mathrm{p}(y\,;\,\mathrm{do}(x)) = \int \cdots$
causal inference
4.32
data

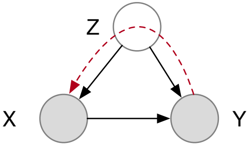
Identification with Backdoor Adjustment



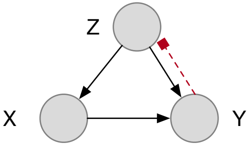
- One of the most useful identification formulas is **backdoor adjustment**.
- Consider a CGM. If a set of variables $\mathbf{z}$ satisfies the backdoor criteria, then

$$\mathrm{p}(y\,;\,\mathrm{do}(\bar{x})) = \int \mathrm{p}(y\,|\,\bar{x}, \mathbf{z})\,\mathrm{p}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

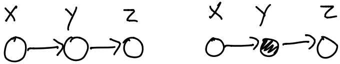
- We saw an example in the three-node model.



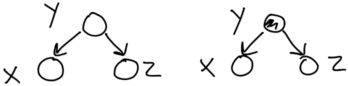
- A backdoor path between X and Y is a path that ends in an edge going into X, i.e., a parent of X.
- Intuition: backdoor paths are how statistical information can confuse causal inferences, i.e., how X and Y can be non-causally associated.



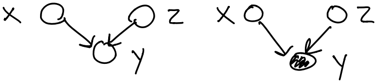
- A blocked door path is a backdoor path that is blocked in the classical graphical models sense (d-separation).
- It is the pattern of observation—which nodes are shaded—that determines whether information can flow from Y to X along the backdoor path.



$X \perp\!\!\!\perp Z \mid Y$

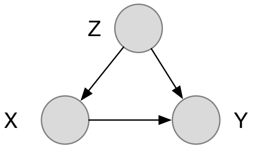


$X \perp\!\!\!\perp Z \mid Y$

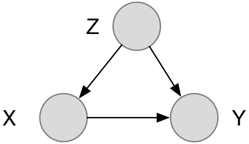


$X \perp\!\!\!\perp Z$

- Let’s review d**-separation**.
- Statistical dependence (which is a liquid) flows in a graphical model.
- How it flows depends on the pattern of conditioning
- Three motifs: mediation, mutual dependence, mutual causation.



- A set of variables $\mathbf{Z}$ satisfies the backdoor criterion (Pearl, 2009) if
  - All directed paths between X and Y are unperturbed.
  - $\mathbf{Z}$ blocks every backdoor path between X and Y.
  - $\mathbf{Z}$ does not introduce any new backdoor paths between X and Y.
- Intuition: The first condition preserves causal links between X and Y. The second and third ensure that there is no other association between X and Y.



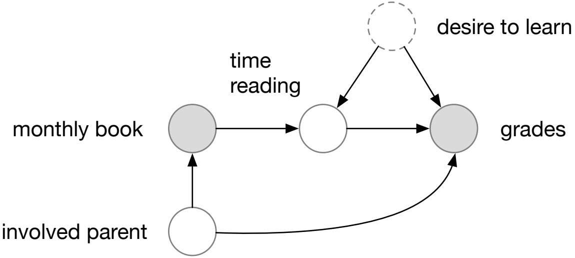
- If we can find a set of variables that satisfies the backdoor criterion then we can write the intervention distribution in terms of the observation distribution.

- Again,

$$\mathrm{p}(y\,;\,\mathrm{do}(\bar{x})) = \int \mathrm{p}(\mathbf{z})\,\mathrm{p}(y\,|\,\bar{x},\mathbf{z})\,\mathrm{d}\mathbf{z}$$

- This identification formula helps estimate properties of the intervention distribution using data from the observation distribution.

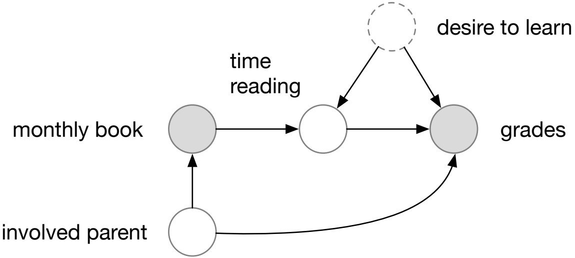
- Note: In an RCT, the empty set satisfies the backdoor criterion.



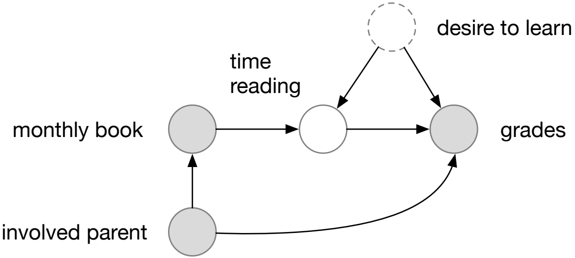
- Use domain knowledge to write a causal graphical model.
  - observable (solid stroke)
  - unobservable (dotted stroke)
  - un-unobservable (double stroke)

- Then solve the puzzle of which sets $\mathbf{Z}$ satisfy the backdoor criterion.

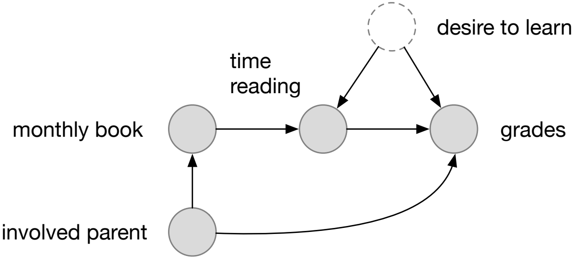
- Given the model and $\mathbf{Z}$, write the adjustment formula for $\mathrm{p}(y\,;\,\mathrm{do}(\bar{x}))$.



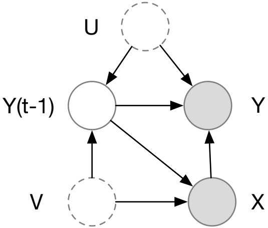
- Should I enroll R in book of the month?
- My involvement is confounding!



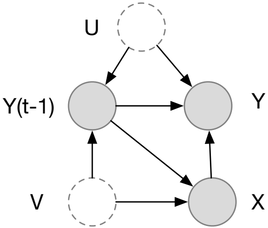
- Should I enroll R in book of the month?
- Conditioning on involvement identifies the effect.



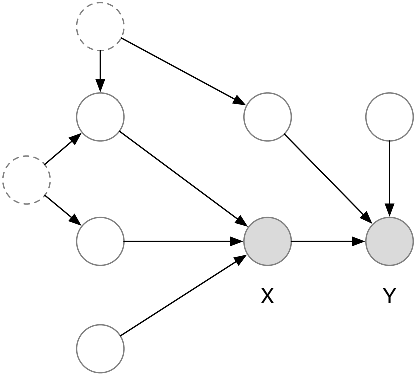
- Should I enroll R in book of the month?
- Don’t just condition on everything!



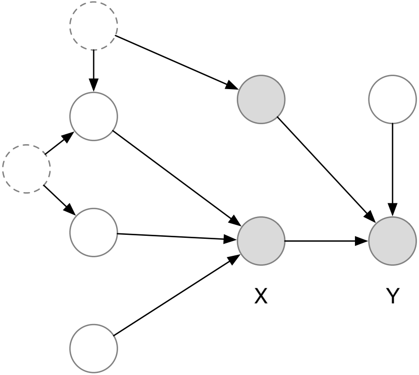
- “A patient’s time series”
- What can we condition on?



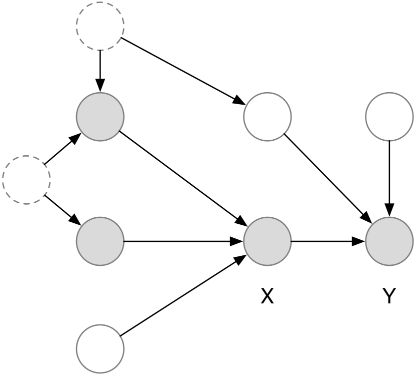
- “A patient’s time series”
- (No way to identify the causal effect)



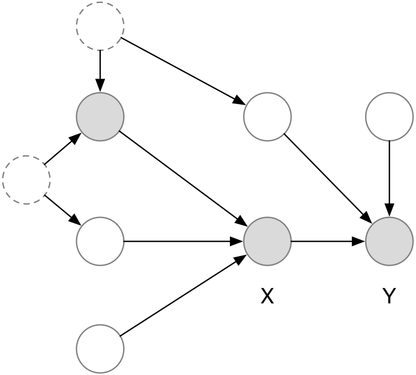
- A big model from Winship and Morgan (2015)
- Is there an admissible set?



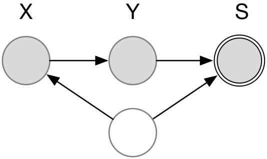
- A big model from Winship and Morgan (2015)
- One node does it.

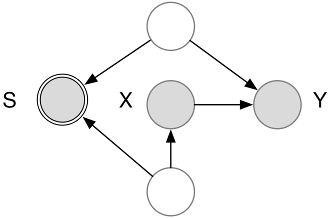


- A big model from Winship and Morgan (2015)
- Or these two nodes.

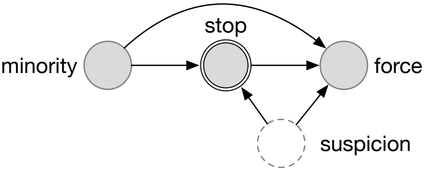


- A big model from Winship and Morgan (2015)
- Just this one does not block the backdoor.

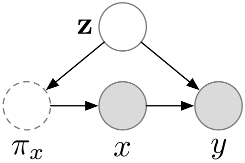




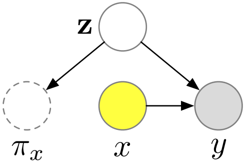
- Sometimes we must model the variable that selects into a study.
- It might open spurious dependence between X and Y (even conditionally).



- This is a graph of real-world importance (Knox et al., 2020)
- Assess the causal effect of being perceived as a minority on the use of force.
- The danger of selection bias: no way to identify the direct effect



observed



intervention

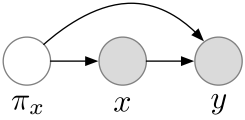
- Let's **sketch the proof** of the backdoor adjustment formula.

- Assume $\mathbf{Z}$ satisfies the backdoor criterion. Then

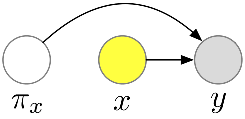
$$Y \perp \pi_x \mid X, \mathbf{Z}$$
$$X \perp \mathbf{Z} \mid \pi_x$$

- The first is because it is only through π_x that a backdoor path can be open.

- The second is because $\mathbf{Z}$ cannot be downstream of X.



observed



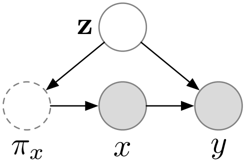
intervention

- Here's a little lemma: adjustment by the parents of X,

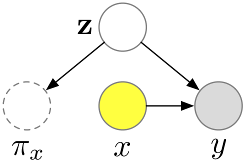
$$\begin{aligned} \mathrm{p}(y\,;\,\mathrm{do}(\bar{x})) &= \int \mathrm{p}(y \,|\, \pi_x\,;\,\mathrm{do}(\bar{x}))\,\mathrm{p}(\pi_x\,;\,\mathrm{do}(\bar{x}))\,\mathrm{d}\pi_x \\ &= \int \mathrm{p}(y \,|\, \bar{x}, \pi_x)\,\mathrm{p}(\pi_x)\,\mathrm{d}\pi_x. \end{aligned}$$

- The second line follows from the graphical model and the definition of "do."

- Key: Replaced intervention distributions with observation distributions.



observed



intervention

$$
\begin{aligned}
\mathrm{p}(y\,;\,\mathrm{do}(\bar{x})) &= \int \mathrm{p}(\pi_x)\,\mathrm{p}(y\,|\,\bar{x},\pi_x)\;\mathrm{d}\pi_x \\
&= \int \mathrm{p}(\pi_x)\int \mathrm{p}(\mathbf{z}\,|\,\bar{x},\pi_x)\,\mathrm{p}(y\,|\,\mathbf{z},\bar{x},\pi_x)\;\mathrm{d}\mathbf{z}\;\mathrm{d}\pi_x \\
&= \int \mathrm{p}(\pi_x)\int \mathrm{p}(\mathbf{z}\,|\,\pi_x)\,\mathrm{p}(y\,|\,\mathbf{z},\bar{x})\;\mathrm{d}\mathbf{z}\;\mathrm{d}\pi_x \\
&= \int \mathrm{p}(\mathbf{z})\,\mathrm{p}(y\,|\,\mathbf{z},\bar{x}).
\end{aligned}
$$

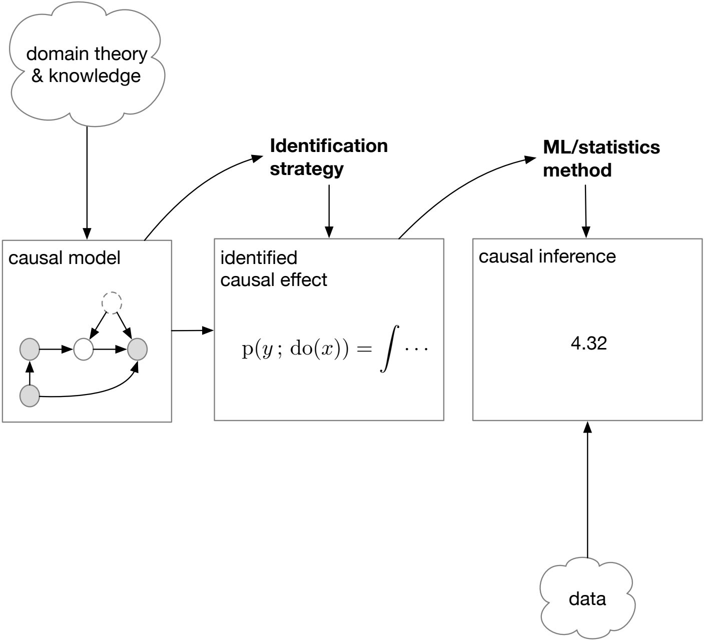
- Adjustment comes with an important assumption, which is called **overlap**.
- The adjustment formula (again),

$$p(y\,;\,\mathrm{do}(\bar{x})) = \int p(\mathbf{z})\, p(y \,|\, \bar{x}, \mathbf{z})\, d\mathbf{z}$$

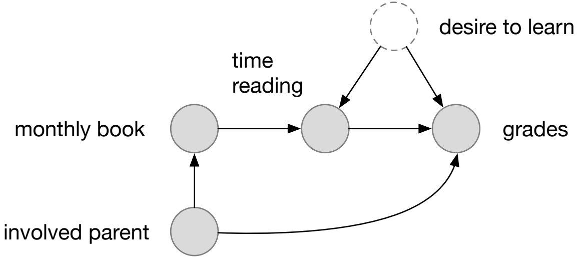
- The conditional $p(y \,|\, \bar{x}, \mathbf{z})$ cannot condition on a zero-probability set,

$$p(\bar{x}, \mathbf{z}) = p(\mathbf{z}) p(\bar{x} \,|\, \mathbf{z}) > 0.$$

- Important: For all values of $\mathbf{z}$, $p(\bar{x} \,|\, \mathbf{z}) > 0$.
  - Each patient has positive conditional probability of treatment (and not)
  - A parent has positive conditional probability of enrolling in BoM (and not)


domain theory & knowledge
Identification strategy
ML/statistics method
causal model
identified causal effect
$\mathrm{p}(y\,;\,\mathrm{do}(x)) = \int \cdots$
causal inference
4.32
data

Estimation and the Backdoor Criterion



- **Domain knowledge**

  provides a model, including what is observable and unobservable.

- **Identification strategies**

  express a causal quantity of interest in terms of the observational distribution.

- **Estimation methods**

  use an identification formula and data to approximate the causal quantity.

- Suppose X is binary and call it the treatment.
- One causal quantity that is often of interest is the **average treatment effect**.
- Average difference in Y between treating ($X = 1$) and not treating ($X = 0$),

$$\mathrm{ATE} = \mathbb{E}[Y\,;\,\mathrm{do}(x=1)] - \mathbb{E}[Y\,;\,\mathrm{do}(x=0)]$$

- Aside: Generalize backdoor adjustment to expectations,

$$\mathbb{E}[f(Y); \mathrm{do}(\bar{x})] = \int \mathbb{E}[f(Y) \,|\, X = \bar{x}, \mathbf{Z} = \mathbf{z}] \; \mathrm{p}(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- Find an admissible set $\mathbf{Z}$ and define

$$\mu(x, \mathbf{z}) \triangleq \mathbb{E}[Y \,|\, X = x, \mathbf{Z} = \mathbf{z}].$$

- Backdoor adjustment provides an identification formula,

$$\mathrm{ATE} = \int (\mu(1, \mathbf{z}) - \mu(0, \mathbf{z})) \; \mathrm{p}(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

- We have identified the ATE. How do we estimate it?
- First use data to estimate $\hat{\mu}(x, \mathbf{z}) \approx \mathbb{E}[Y \mid X = x, \mathbf{Z} = z]$
- Then approximate $\mathrm{p}(\mathbf{z})$ with its empirical distribution (i.e., use Monte Carlo).
- With these two decisions, an estimate of the ATE is

$$\text{ATE} \approx \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}(1, \mathbf{z}_i) - \hat{\mu}(0, \mathbf{z}_i)).$$

1. Use data to estimate

$$\mathbb{E}[Y \mid X = x, \mathbf{Z} = \mathbf{z}] \approx \hat{\mu}(x, \mathbf{z})$$

2. Approximate

$$\text{ATE} \approx \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}(1, \mathbf{z}_i) - \hat{\mu}(0, \mathbf{z}_i)).$$

- Other than to fit $\hat{\mu}$, we do not use the assigned treatment x_i. Rather, we consider $\hat{\mu}(x, z_i)$ under both treatment (x=1) and control (x=0).

- Notice that $\hat{\mu}(x, \mathbf{z}_i)$ is a generic estimator of the conditional expectation. It is not tied to the causal model. (We can use regression, neural networks, etc.)

1. Use data to estimate

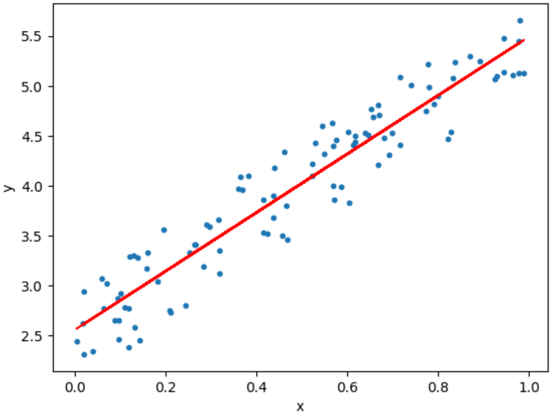
$$\mathbb{E}[Y \mid X = x, \mathbf{Z} = \mathbf{z}] \approx \hat{\mu}(x, \mathbf{z})$$

2. Approximate

$$\text{ATE} \approx \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}(1, \mathbf{z}_i) - \hat{\mu}(0, \mathbf{z}_i)).$$

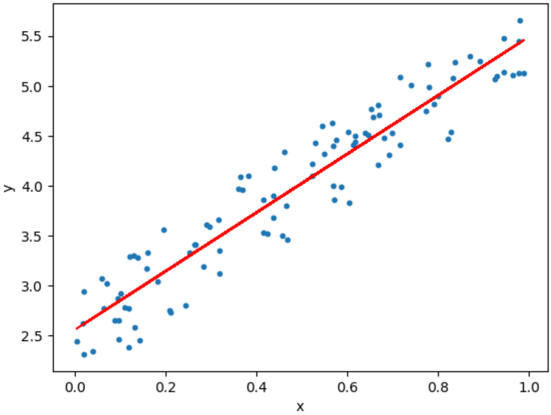
Consider this “causal inference workflow”:

- Use domain knowledge to determine the model.
- Use theory to find observed $\mathbf{Z}$ that satisfies the backdoor criterion.
- Use ML/statistics and data to approximate $\mathbb{E}[Y \mid X = x, \mathbf{Z} = \mathbf{z}]$.



towardsdatascience.com

- What is the relationship between **causal inference** and **regression**?

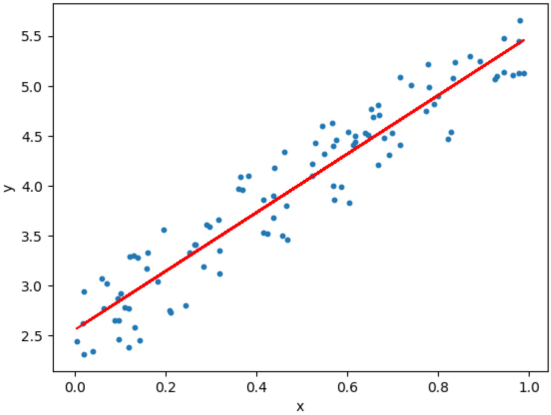


towardsdatascience.com

- Consider the effect of a one-unit change in the treatment x,

$$1\text{TE} = \mathbb{E}[Y\,;\,\text{do}(\bar{x}+1)] - \mathbb{E}[Y\,;\,\text{do}(\bar{x})].$$

(When x is binary, the ATE is an instance of such an effect.)



towardsdatascience.com

- Given an admissible set, we can identify the 1TE by adjustment

$$1\mathrm{TE} = \int (\mu(\bar{x}+1,\mathbf{z}) - \mu(\bar{x},\mathbf{z}))\, \mathrm{p}(\mathbf{z})\, \mathrm{d}\mathbf{z}.$$

- Fit a linear regression model of the conditional expectation,

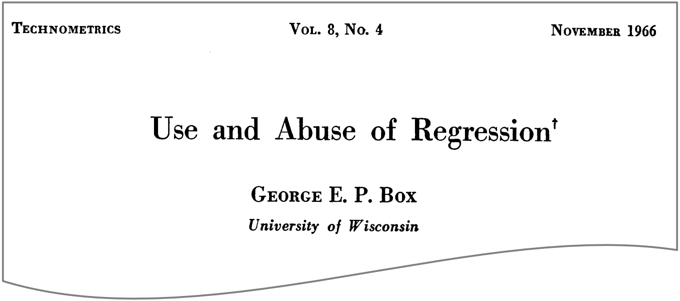
  $$\mathbb{E}[Y \mid X = x, \mathbf{Z} = \mathbf{z}] = \mu(x, \mathbf{z}) \approx \hat{\beta} x + \hat{\eta} \cdot \mathbf{z}. \tag{1}$$

  The covariates are $(x, \mathbf{z})$.

- Substitute the fitted regression into the identification formula,

  $$\begin{aligned} \text{1TE} &\approx \int (\hat{\mu}(\bar{x}+1, \mathbf{z}) - \hat{\mu}(\bar{x}, \mathbf{z}))\, \mathrm{p}(\mathbf{z})\, \mathrm{d}\mathbf{z} \\ &= \hat{\beta}. \end{aligned}$$

- This is what is meant by reporting a coefficient and “adjusting for $\mathbf{z}$.”



TECHNOMETRICS VOL. 8, NO. 4 NOVEMBER 1966

Use and Abuse of Regression†

GEORGE E. P. BOX

University of Wisconsin

- The coefficient is only causal if $\mathbf{z}$ blocks all backdoor paths.
- Beware: Don't just include everything you observe.
  - It might not be enough to block all backdoor paths.
  - It might open a backdoor path (e.g., via a collider).
  - It might block a causal path (e.g., via a mediator).

1. Use regression to estimate

$$\mathbb{E}[Y \,|X = x, \mathbf{Z} = \mathbf{z}] \approx \hat{\beta}x + \hat{\eta} \cdot \mathbf{z}$$

2. Approximate

$$1\mathrm{TE} \approx \hat{\beta}$$

- Thanks to linearity, we did not need to take a Monte Carlo estimate over $\mathrm{p}(\mathbf{z})$.
- Notice (yet again) the clear lines between identification and estimation.
  - No assumption about the true causal model (e.g., linearity)
  - But linear regression must capture the conditional expectation.

- Another commonly used estimation method is **matching**.
- Recall that $\mu(\bar{x}, \mathbf{z}) = \mathbb{E}[Y \,|\, X = \bar{x}, \mathbf{Z} = \mathbf{z}]$.
- Given data, consider a Monte Carlo approximation of the ATE,

$$\begin{aligned} \text{ATE} &= \int \mu(1, \mathbf{z}) - \mu(0, \mathbf{z})\, \mathrm{p}(\mathbf{z})\; \mathrm{d}\mathbf{z} \qquad\qquad \text{(from adjustment)} \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \mu(1, \mathbf{z}_i) - \mu(0, \mathbf{z}_i). \end{aligned}$$

- Here is the Monte Carlo approximation of the ATE

$$\text{ATE} \approx \frac{1}{n}\sum_{i=1}^{n} \mu(1, \mathbf{z}_i) - \mu(0, \mathbf{z}_i).$$

- Consider a datapoint i where $x_i = 1$, the individual received treatment.

- A <u>match</u> for i is another individual i^* where:
  - the adjusting variables are the same, $\mathbf{z}_i = \mathbf{z}_{i^*}$
  - but they did *not* receive the treatment, $x_{i^*} = 0$.

- Notice that $y_i - y_{i^*}$ is an unbiased estimate of $\mu(1, \mathbf{z}_i) - \mu(0, \mathbf{z}_i)$.

1. For each datapoint i:

    Find a match i^* where $\mathbf{z}_i = \mathbf{z}_{i^*}$ and x_{i^*} is opposite x_i.

2. Estimate the ATE as

$$\text{ATE} \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - y_{i^*}).$$

1. For each datapoint i:

   Find a match i^* where $\mathbf{z}_i = \mathbf{z}_{i^*}$ and x_{i^*} is opposite x_i.

2. Estimate the ATE as

$$\text{ATE} \approx \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{i^*}).$$

- The adjusting variables $\mathbf{z}$ still must satisfy the backdoor criterion.
- Multiple datapoints can be matched to the same datapoint.
- In practice, we can use approximate matches.

- The final estimation strategy we discuss is **inverse propensity weighting**
- There are many ways to derive it. One is through importance sampling.

- The goal of **importance sampling** is to approximate $\mathbb{E}_{\mathrm{p}}[f(\mathbf{Z}, X, Y)]$.
- Consider $\mathrm{p}(\mathbf{z}, x, y)$ and $\mathrm{q}(\mathbf{z}, x, y)$. (No causality for a moment.)
- Importance sampling writes the p-expectation with a q-expectation,

$$\begin{aligned}
\mathbb{E}_{\mathrm{p}}[f(\mathbf{Z}, X, Y)] &= \int \mathrm{p}(\mathbf{z}, x, y) f(\mathbf{z}, x, y) \,\mathrm{d}\mathbf{z}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int \mathrm{p}(\mathbf{z}, x, y) f(\mathbf{z}, x, y) \frac{\mathrm{q}(\mathbf{z}, x, y)}{\mathrm{q}(\mathbf{z}, x, y)} \,\mathrm{d}\mathbf{z}\,\mathrm{d}x\,\mathrm{d}y \\
&= \mathbb{E}_{\mathrm{q}}\left[\frac{\mathrm{p}(\mathbf{Z}, X, Y)}{\mathrm{q}(\mathbf{Z}, X, Y)} f(\mathbf{Z}, X, Y)\right]
\end{aligned}$$

- Then use Monte Carlo,

$$\mathbb{E}_{\mathrm{p}}\left[f(\mathbf{Z}, X, Y)\right] \approx \frac{1}{n} \sum_{i=1}^{n} w_i f(\mathbf{z}_i, x_i, y_i), \qquad (\mathbf{z}_i, x_i, y_i) \sim \mathrm{q}(\cdot)$$

- The weights are called importance weights,

$$w_i = \frac{\mathrm{p}(\mathbf{z}_i, x_i, y_i)}{\mathrm{q}(\mathbf{z}_i, x_i, y_i)}$$

- Punchline: IS uses samples from $\mathrm{q}(\cdot)$ to approximate a p-expectation.

- Back to causality, our goal is to calculate an interventional expectation

$$\mathbb{E}[f(\mathbf{Z}, X, Y)\,;\, \mathrm{do}(\bar{x}))] \qquad \text{e.g., } f(\cdot) = y$$

- Consider the target as intervention and the proposal as observation,

$$\begin{aligned} \mathrm{p}(\cdot) &= \mathrm{p}(\mathbf{z}, x, y\,;\, \mathrm{do}(\bar{x})) \\ \mathrm{q}(\cdot) &= \mathrm{p}(\mathbf{z}, x, y) \end{aligned}$$

- The importance weights simplify,

$$\frac{\mathrm{p}(\cdot)}{\mathrm{q}(\cdot)} = \frac{\mathrm{p}(\mathbf{z}\,;\, \mathrm{do}(\bar{x}))}{\mathrm{p}(\mathbf{z})} \frac{\mathrm{p}(x \mid \mathbf{z}\,;\, \mathrm{do}(\bar{x}))}{\mathrm{p}(x \mid \mathbf{z})} \frac{\mathrm{p}(y \mid \mathbf{z}, x\,;\, \mathrm{do}(\bar{x}))}{\mathrm{p}(y \mid \mathbf{z}, x)} \quad = \frac{1(x = \bar{x})}{\mathrm{p}(x \mid \mathbf{z})}$$

- Notice where we needed that $\mathbf{z}$ satisfies the backdoor criterion.

1. Use data to estimate

$$\mathrm{p}(x \mid \mathbf{z}) \approx \hat{\sigma}(\mathbf{z})$$

2. Approximate

$$\mathbb{E}[Y\,;\,\mathrm{do}(\bar{x})] \approx \frac{1}{n}\sum_{i=1}^{n} w_i y_i \qquad\qquad w_i = \frac{1(x_i = \bar{x})}{\hat{\sigma}(\bar{x}, \mathbf{z}_i)}$$

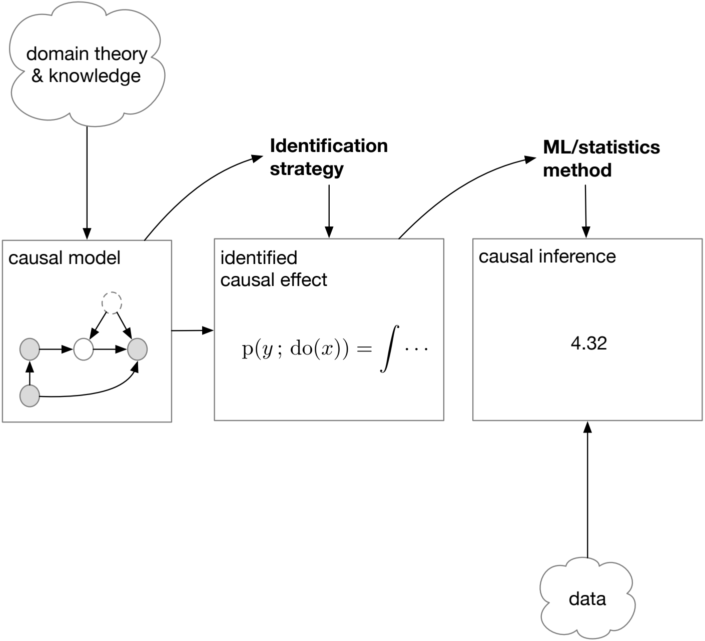
1. Use data to estimate

$$\mathrm{p}(x \mid \mathbf{z}) \approx \hat{\sigma}(\mathbf{z})$$

2. Approximate

$$\mathbb{E}[Y ; \mathrm{do}(\bar{x})] \approx \frac{1}{n} \sum_{i=1}^{n} w_i y_i \qquad w_i = \frac{1(x_i = \bar{x})}{\hat{\sigma}(\bar{x}, \mathbf{z}_i)}$$

- Only requires estimating the propensity score, $\mathrm{p}(x \mid \mathbf{z})$
- Can consider other interventions, e.g., a randomized experiment
- Variance can be high, especially if $\mathrm{p}(x \mid \mathbf{z})$ is extreme. (Try clipping.)
- In practice $\sum_i (1/w_i)$ replaced $1/n$; the derivation is similar

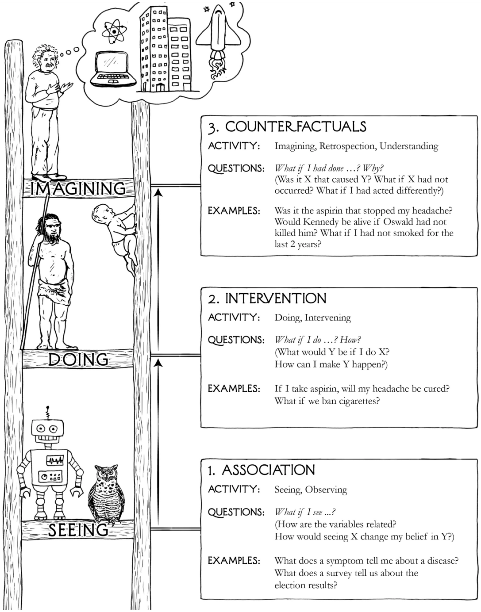

domain theory & knowledge
Identification strategy
ML/statistics method
causal model
identified causal effect
$\mathrm{p}(y\,;\,\mathrm{do}(x)) = \int \cdots$
causal inference
4.32
data

We discussed four options for estimation. Each has different requirements.

- **General adjustment**: Estimate $\mathrm{p}(\mathbf{z})$ and $\mathbb{E}[Y \mid \mathbf{Z} = \mathbf{z}, X = x]$.
- **Regression adjustment**: Estimate $\mathbb{E}[Y \mid \mathbf{Z} = \mathbf{z}, X = x]$ with regression.
- **Matching**: Find $\mathbf{z}$-matches for each i.
- **IPW**: Estimate $\mathrm{p}(x \mid \mathbf{z})$.

All of them require $\mathbf{z}$ that satisfies the backdoor criterion.

Structural Causal Models and Counterfactuals



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

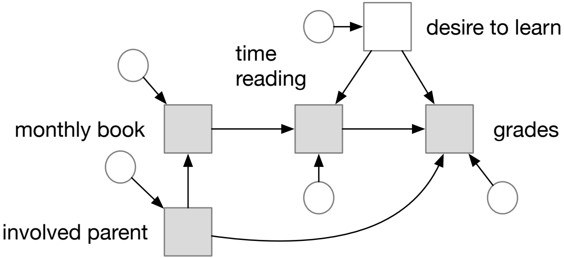
ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

from Pearl and MacKenzie (2018)

- Would Hillary have won the election had she visited Michigan three days before it, given that it was Trump who won the election.
- Would double panes on my windows have prevented them from blowing through during the hurricane, given that I have single-pane windows and they did blow through.
- Would the supermarket have gone out of business had it raised the price of peanut butter instead of bananas, given that it did not go out of business and that it did raise the price of bananas.
- Would this applicant have been admitted had they been white, given their test score was high, they were not admitted, and they are not white.



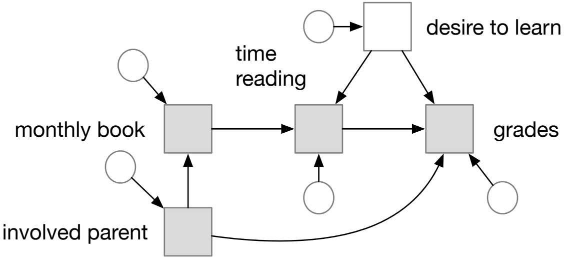
- A **structural causal model** M contains three components.

- Endogenous variables $V_{1:k}$ are organized in a DAG.

- Background variables $U_{1:k}$ are determined by “factors outside the model.”

- Functions $F_{1:k}$ maps U to V.
  The ith function takes the parents of v_i and the ith background variable u_i,

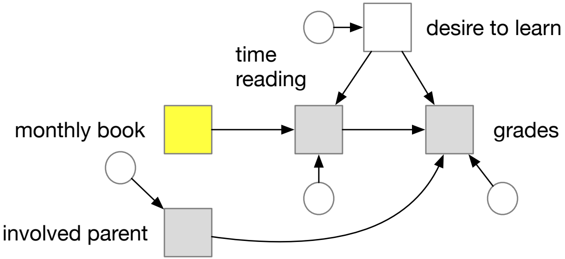
$$v_i = f_i(v_{\pi_i}, u_i), \qquad\qquad i = 1 \dots k.$$



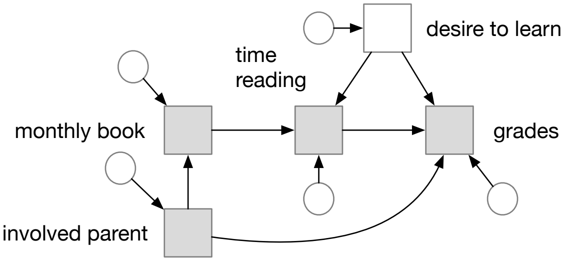
- In a **probabilistic SCM**, the background variables are drawn independently,

$$\mathrm{p}(u) = \prod_{i=1}^{k} \mathrm{p}_i(u)$$

- They can come from different distributions.
- They are the only probabilistic component of the model.



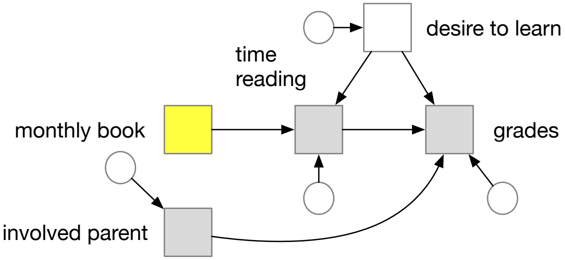
- An **intervention** on the ith variable changes the function that determines it. (Let's not allow arguments to be added.)
- Let Y_{M} denote the variable Y in model M.
- Consider an intervention model M'. Let $Y_{\mathrm{M}'}$ denote variable Y in that model.
- When we do causal inference, we ask questions about $Y_{\mathrm{M}'}$.



- In the definition, Y_{M} is a function of X, Z, and its background noise u_y.

- But we can unroll that function recursively. It is a function of the all of the background variables (u_x, u_y, u_z),

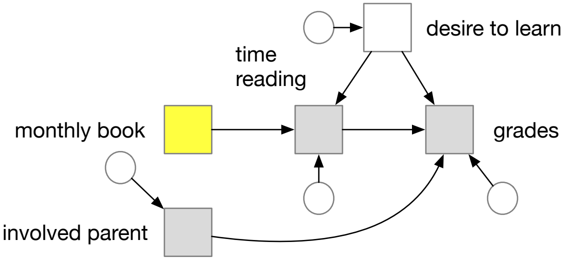
$$\begin{aligned} Y_{\mathrm{M}} &= f_y(x, z, u_y) \\ &= f_y(f_x(u_x, f_z(u_z)), f_z(u_z), u_y) \\ &\triangleq Y_{\mathrm{M}}(u_x, u_y, u_z). \end{aligned}$$



- We can do the same for $Y_{\mathrm{M}'}$ in the intervened model where we $\mathrm{do}(x = \bar{x})$.

$$\begin{aligned} Y_{\mathrm{M}'} &= f_y(\bar{x}, z, u_y) \\ &= f_y(\bar{x}, f_z(u_z), u_y) \\ &\triangleq Y_{\mathrm{M}'}(u_y, u_z). \end{aligned}$$

- Note: $\mathbb{E}[Y_{\mathrm{M}'}] = \mathbb{E}[Y \,;\, \mathrm{do}(\bar{x})]$

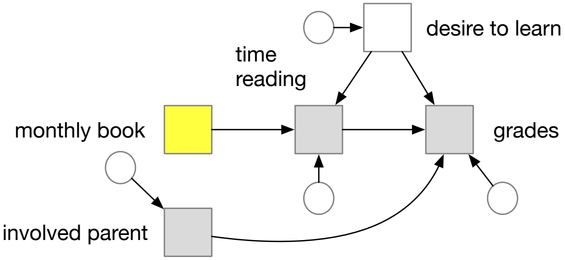


- Each variable is a different function of the same background variables,

$$Y_{\mathrm{M}}(\mathbf{u}) = f_y(f_x(u_x, f_z(u_z)), f_z(u_z), u_y) \qquad \text{(observation)}$$
$$Y_{\mathrm{M}'}(\mathbf{u}) = f_y(\bar{x}, f_z(u_z), u_y) \qquad \text{(intervention)}$$

- **Key idea**: The noise variables $\mathbf{u}$ simultaneously define the variables in all possible interventional worlds.

- If I hand you $\mathbf{u}$, you can tell me what would have happened under each intervention (including the observed non-intervention).



$$\mathbf{u}_i \sim \mathrm{p}(\mathbf{u}) \quad ; \quad (z_i, x_i, y_i) = f(\mathbf{u}).$$

- The **fundamental law of counterfactuals** says that y_i is the value it would have been had we intervened in the world and set $x = x_i$. E.g., consider a fixed $\mathbf{u}_i$ where $x_i = \bar{x}$. Then Y_{M} and $Y_{\mathrm{M}'}$ have the same value.
- An SCM is a CGM. But SCMs assert more than “direct causes.”

- **SEMs can express counterfactuals about the past.**
- Consider a variable in an intervened model, $Y_{\mathrm{M}'}(\mathbf{u})$.

  "My grades if I was enrolled in the book-of-the-month club."

- Its distribution is defined through the noise variables,

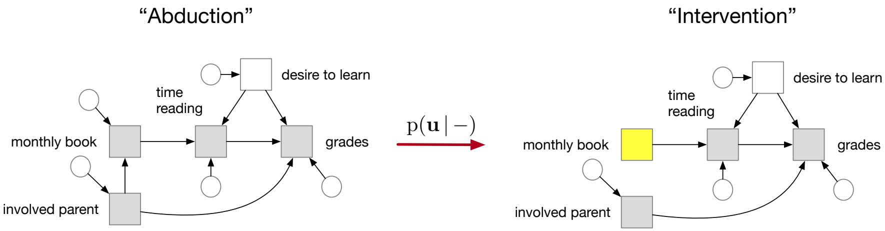
$$\mathrm{P}(Y_{\mathrm{M}'}(\mathbf{u}) = \bar{y}) = \int \mathrm{p}(\mathbf{u})\, 1(Y_{M'}(\mathbf{u}) = \bar{y})\, \mathrm{d}\mathbf{u}.$$

- We can condition on other variables, $\mathrm{P}(Y_{\mathrm{M}'} = \bar{y} \,|\, X_{\mathrm{M}} = \bar{x})$.

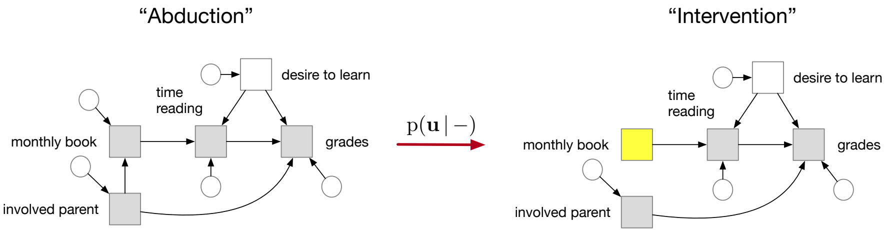
  "My grades if I had been enrolled in the book-of-the-month club, given that in reality I was not enrolled in the book-of-the-month club"

- Let's write it out,

$$\mathrm{p}(Y_{\mathrm{M}'} = \bar{y} \,|\, X_{\mathrm{M}} = \bar{x}) = \int \mathrm{p}(\mathbf{u} \,|\, X_{\mathrm{M}}(\mathbf{u}) = \bar{x})\, 1(Y_{\mathrm{M}'}(\mathbf{u}) = \bar{y})\, \mathrm{d}\mathbf{u}$$



- Intuitively, computing this counterfactual involves two steps
  - "Abduction": Calculate $p(\mathbf{u} \,|\, X_{\mathrm{M}}(\mathbf{u}) = \bar{x})$
  - "Intervention": Use the conditional to calculate $Y_{\mathrm{M}'}(\mathbf{u})$
- Note: Abduction comes from the observed world M.
- Put differently: An assumed SEM and the law of counterfactuals lets us reason about the past under intervention.



- Aside: This all falls out of basic probability theory.
- Think of $\mathbf{u}$ as the underlying atoms and their distribution as their measure.
- Variables like $Y_{\mathrm{M}'}(\mathbf{u})$ are just random variables, in the traditional sense.
- The probability space defines their joint distribution and their conditionals.

- Would Hillary have won the election had she visited Michigan three days before it, given that it was Trump who won the election.
- Would double panes on my windows have prevented them from blowing through during the hurricane, given that I have single-pane windows and they did blow through.
- Would the supermarket have gone out of business had it raised the price of peanut butter instead of bananas, given that it did not go out of business and that it did raise the price of bananas.
- Would this applicant have been admitted had they been white, given their test score was high, they were not admitted, and they are not white.

Unit	X	$Y_{\mathrm{do}(x=1)}$	$Y_{\mathrm{do}(x=0)}$
Student 1	yes	1	?
Student 2	no	?	0
Student 3	yes	?	1
Student 3	no	1	?
⋮			
Student n	yes	0	?

- The **Rubin-Neyman** causal model directly operates on the joint distribution of counterfactuals. This is called the **potential outcomes** framework.

- The joint distribution is $\mathrm{P}(Y_{\mathrm{do}(x=0)}, Y_{\mathrm{do}(x=1)}, X, \mathbf{Z})$.

- The idea: We observe incomplete data from this distribution.
  Causal inferences involve “filling in” the missing counterfactuals.

- The pattern of missingness can confound the estimates.

Unit	X	$Y_{\text{do}(x=1)}$	$Y_{\text{do}(x=0)}$
Student 1	yes	1	?
Student 2	no	?	0
Student 3	yes	?	1
Student 3	no	1	?
⋮			
Student n	yes	0	?

- **Ignorability** asserts that $X \perp (Y_0, Y_1) \mid \mathbf{Z}$.
  - Has the same effect as blocking backdoor paths
  - Note: In an RCT, ignorability holds unconditionally

- **Stable Unit Treatment Value Assumption (SUTVA)**
  - The units do not interfere with each other
  - The treatment is the same across units

Unit	X	$Y_{\mathrm{do}(x=1)}$	$Y_{\mathrm{do}(x=0)}$
Student 1	yes	1	?
Student 2	no	?	0
Student 3	yes	?	1
Student 3	no	1	?
⋮			
Student n	yes	0	?

- How do graphs and PO relate? Is one more expressive?

- Can all joints on interventions be defined via an SCM?
  - They can all be defined by the set of unrolled functions of noise.
  - But do all unrolled functions lead to an SCM?

- See Winship and Morgan (2015), Richardson and Robins (2013)

Discussion

- Causality is about understanding the **effects** of a **hypothetical intervention**.

- Example: A father is considering signing his daughter up for “book of the month” club because he thinks she will read more and do better in school.

- Why is this a causal inference? The father is considering an <u>intervention</u> in the world in the hopes of changing an <u>outcome</u>.

- Will signing her up for the club cause more reading? Will more reading cause better grades? (Will better grades cause a happier life?)

- (Note: Causality quickly turns philosophical, but we’ll avoid such rabbit holes.)

Here are some other causal questions.

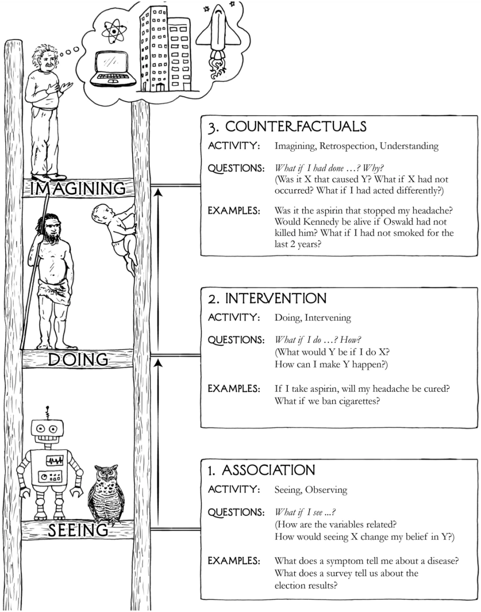
- If I take this medicine, will my headache go away?
- If I deploy this new recommender, will my website's revenue increase?
- If I implement this education policy, will poverty decrease?
- If I make this change to the tax law, will unemployment decrease?
- If I use this algorithm to hire people, will it satisfy equal opportunities?
- If I play music for this monkey, will he get smarter?
- If I show a user this article, will she click on it?
- If I enroll at Columbia, will I make more money?
- If students turn off their phones, will they learn more in class?
- If I stop eating french fries, will my cholesterol go down?
- If I encourage my friends to vote, will they vote?

These questions are about what happens when we intervene in the world.

We can also ask causal questions about the past.

- Would Hillary have won the election had she visited Michigan three days before it, given that it was Trump who won the election.
- Would double panes on my windows have prevented them from blowing through during the hurricane, given that I have single-pane windows and they did blow through.
- Would the supermarket have gone out of business had it raised the price of peanut butter instead of bananas, given that it did not go out of business and that it did raise the price of bananas.

Such questions are called counterfactuals. They too are questions about the effects of an intervention, even if it’s one that requires a time machine.



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

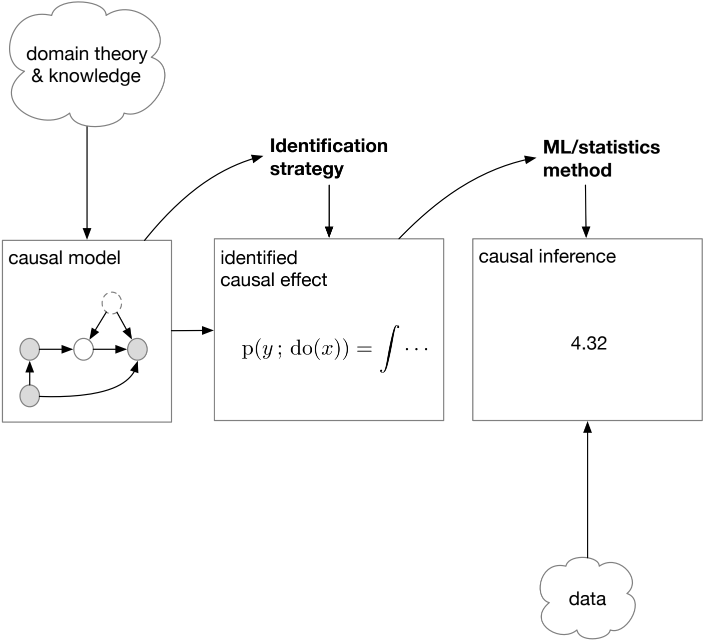
1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?* (How are the variables related? How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

from Pearl and MacKenzie (2018)


domain theory & knowledge
Identification strategy
ML/statistics method
causal model
identified causal effect
$\mathrm{p}(y\,;\,\mathrm{do}(x)) = \int \cdots$
causal inference
4.32
data

Topics in causality we discussed today

- Causal graphical models
- Identification with backdoor adjustment
- Estimation
  - General adjustment
  - Regression
  - Matching
  - Inverse propensity weights
- Structural causal models and counterfactuals
- Potential outcomes (a little)

Topics in causality we did not discuss today

- Structure learning
- Instrumental variables
- Negative controls
- Front-door adjustment
- Double machine learning
- Sensitivity analysis
- Multiple causality
- Time-dependent treatments
- Regression discontinuities
- Causal forests
- Invariant risk minimization
- Experimental design
- Single-world intervention graphs
- Mediators and moderators
- Overcoming selection bias
- Reinforcement learning
- Sufficiency of the propensity score
- and more...

- **Counterfactuals and Causal Inference (2nd edition)**
  S. Morgan and C. Winship (2015, Cambridge University Press)

- **Causality (2nd edition)**
  J. Pearl (2009, Cambridge University Press)

- **Causal Inference in Statistics: A Primer**
  J. Pearl, M. Glymour, and N. Jewell (2016, John Wiley & Sons)

- **The Book of Why**
  J. Pearl and D. MacKenzie (2018, Basic Books)

- **Causal Inference**
  M. Hernan and J. Robins (2019, Chapman & Hall/CRC)

- **Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction**
  G. Imbens and D. Rubin (2015, Cambridge University Press)

- **Elements of Causal Inference : Foundations and Learning Algorithms**
  J. Peters, D. Janzing, and B. Schoelkopf (2018, The MIT Press)

- **Advanced Data Analysis from an Elementary Point of View**
  C. Shalizi (2020, in preparation)