

CAUSAL INFERENCE

An Overview, And the Role of Probabilistic Modelling

Gaussian Process and Uncertainty Quantification
Summer School 2020

Speaker: Ricardo Silva, University College London



LET ME GIVE YOU THREE PROBLEMS

1. Mr X is a smoker. How long will he live?
2. Is it worthwhile for Mr X to stop smoking?
3. Mr X just died of lung cancer. Would he be alive if he hasn't been a smoker?

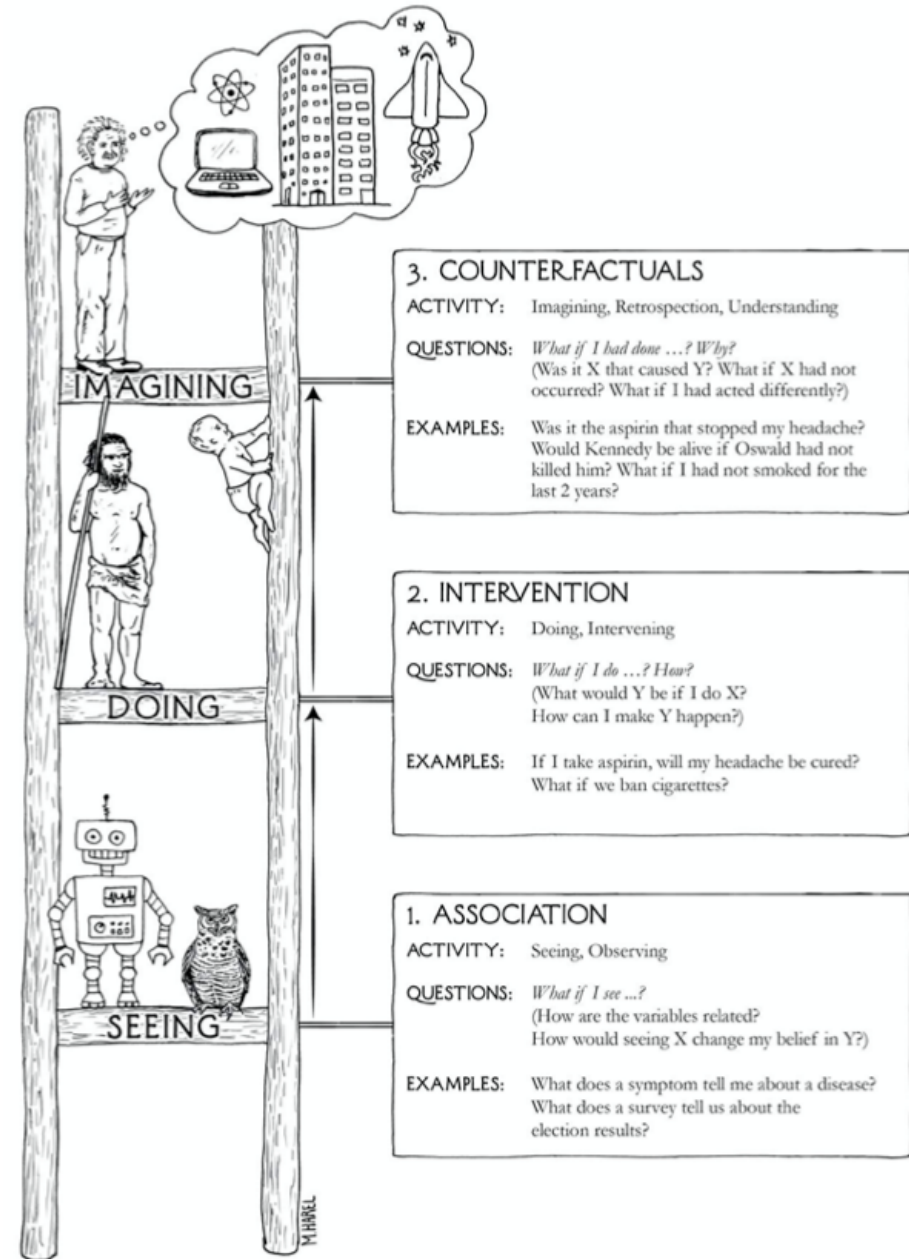
THREE CLASSES OF PROBLEM

1. **Prediction by observation:** see state of the system, predict outcome.
2. **Prediction with intervention:** see state of the system, **hypothesize an intervention** on it, predict outcome.
3. **Explanation by counterfactuals:** see final state of the system, **conjecture what would have happened** if an intervention that did not take place had actually taken place.

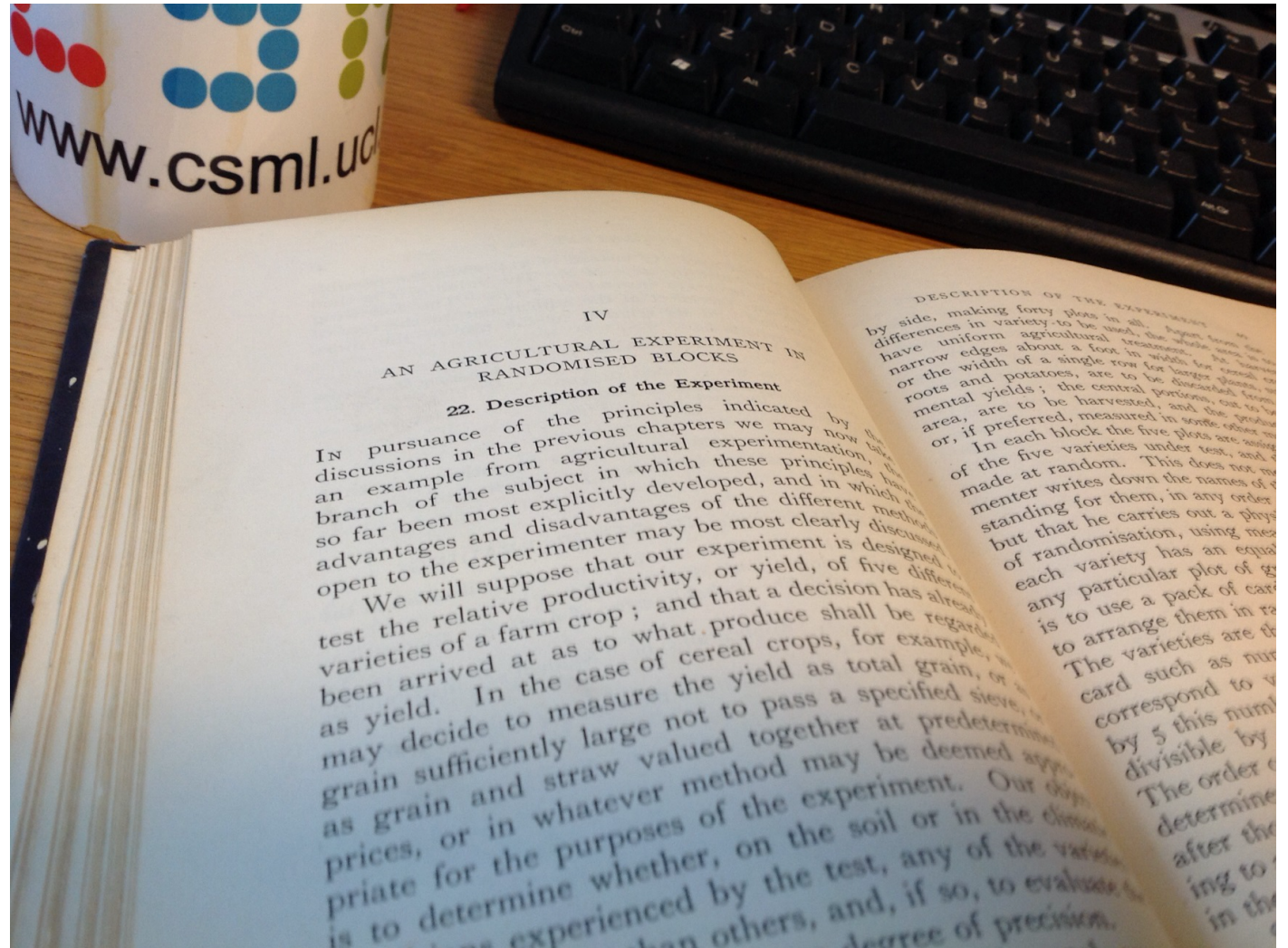
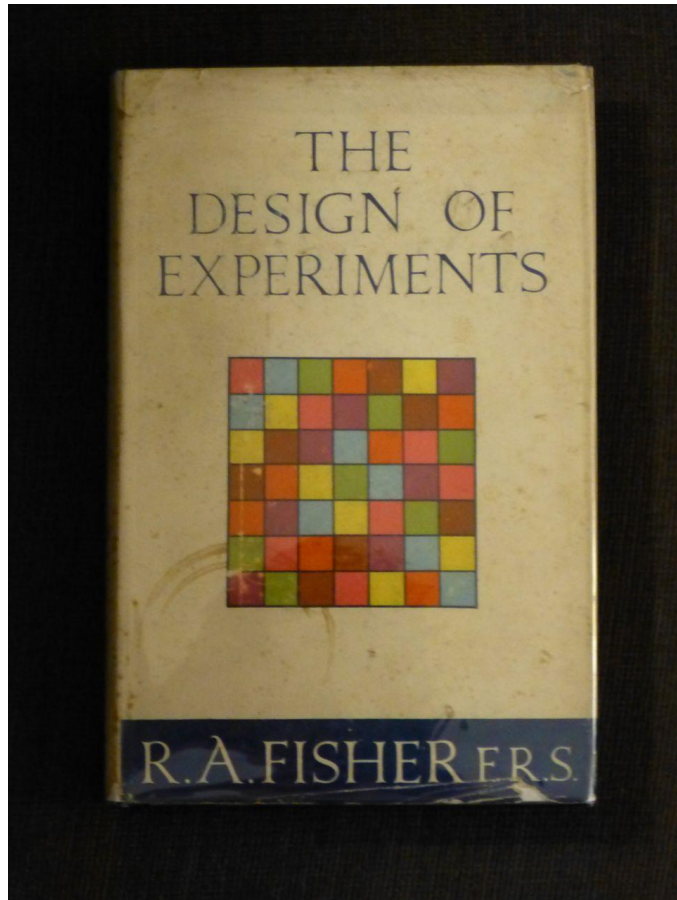
PEARL'S “LADDER OF CAUSATION”

Pearl and Mackenzie (2018). *The Book of Why, Basic Books.*

The most distinctive aspect of Ladders 2 & 3 is that we imagine (hypothetically or counterfactually) an action taken by **someone outside the system.**



LEARNING FROM ACTIONS



THE PROBLEM OF **CONFOUNDING**

Say that a medication seems to be killing off patients prematurely in a particular hospital: mortality rates are higher among those who take it.

Does it mean the medication is indeed bad?

THE PROBLEM OF CONFOUNDING

Not necessarily: perhaps the medication is being prescribed to those who were most at risk to begin with!

We say drug-taking and health-outcome are **confounded** in the sense that a common cause can explain their association (and as you know, “**association is not causation**” etc. etc.)

If we know which observed factors confound this association, we can account for them.

The problem gets particularly **hard if there are hidden variables** responsible for confounding.

Dealing with confounding is perhaps the central problem of causal inference (it's not the only one).

RANDOMIZED CONTROLLED TRIALS (RCTS)

Say we have **treatments** and **outcomes** variables.

For simplicity, in most cases we will assume treatment is a binary variable X and the outcome is a single variable Y .

Assume **we can control X in any way we want**. But it may be hard to believe the treatment selection is not **confounded** with Y .

Fisher introduced the idea of randomization. Pick values for X based on (say) the flip of a coin. Assess differences on Y . Breaking possible confounding is what a **randomized controlled trial (RCT)** is aiming at.

A FORMAL LANGUAGE

In Fisher's setup, we can imagine that there are two "parallel worlds" corresponding to the **potential outcomes** of Y under each possible course of action.

Each potential outcome is **denoted** by $Y(x)$ for possible choices x of X .

What the selection of X does is to select which "world" we are visiting and observing. **The rest is missing data.**

X	$Y(0)$	$Y(1)$
<u>0</u>	<u>1.6</u>	<u>?</u>
0	-0.3	?
1	?	2.1
0	0.8	?
1	?	1.7

INFERENCE, THE CLASSICAL WAY

The only source of randomness is the flip of the coin. We assume the potential outcomes are fixed. We just “visit” them. Think of stopping people in the street to ask whom they are going to vote for, for instance.

This led to famous setups in Statistics, such as “**Fisher’s sharp null hypothesis**”:

- assume the **null hypothesis** $Y(0) = Y(1)$ for all data points. Under the null, there is no missing data!
- the distribution of the average of each column of selected potential outcomes can be computed. Averaging is taken over all assignments to the X column
- how unlikely is the gap between the observed column averages under the null? This can be computed!

IS THIS REALLY RELEVANT?

Of course it is. It's the mainstream (and regulated) way of assessing effectiveness of a drug, for example. It makes minimal assumptions. By all means, I'd demand a RCT to assess a Covid-19 vaccine, with as few assumptions as possible!

But notice the issues of classical inference such as RCT + Fisher's test:

- This is only about a specific group of units, under a specific time! It makes zero claims of out-of-sample behavior. **This is anathema for machine learning/AI.**
- Who cares about the null? But why should we care about the **effect size** if we assume the potential outcomes were fixed to begin with?
- Theoretically, all a ("successful") classical RCT for drug approval does is to say **"there exists a group of people that at some point in time under specific circumstances showed the property that they could have better than zero benefits, on average, in taking this-treatment compared to that-treatment"**.
- That's a mouthful you don't see in drug advertisements!

IS THIS REALLY RELEVANT?

Why such a low bar?

Confounding is a major issue, so we want RCTs.

But they are expensive (**sometimes not possible!**). So we end up both with not that large of a sample size and – worse – often a non-representative sample of a population of interest

- How diverse is the design: white, middle-aged men only? Who volunteers: relatives of affected people, low-income individuals looking for compensation, high-income individuals with spare time, etc.?

This can of course be much less of a problem in some domains. For instance, we are experimented with constantly as part of the web commerce global environment.

- But even then, practical issues: we can have a very complex space of “things to try”, and we don’t want to pester people by experimenting with them constantly and chasing them for consent.

OBSERVATIONAL STUDIES

This is something Fisher and others were vehemently against: infer causal claims without RCTs. We shouldn't buy this oversimplification. The smoking and lung cancer link is perhaps the most famous case.

Much of causal inference concerns observational studies, or even RCTs where the intervention is not as precise as we can afford it to be. Donald Rubin extended **the basic idea of potential outcome to the more general setup** that includes observational studies.

- This is known as the **Rubin Causal Model**, or **Neyman-Rubin model**, as Neyman outlined the PO idea originally.

Observational data is particularly important if:

- We can't do a RCT for ethical (or technological/economical) reasons, as in the smoking case!
- We strive to provide predictive claims in a less coarse way ("this treatment is better than that treatment **for someone of your age**")
- We want to gather evidence of how a treatment *actually* works outside of a lab!
- We want to assess impact on variables nobody thought to measure at trial time, such as particular side-effects
- We want to collect evidence of what trials to design, as the **action space** someone can choose from may be too large to think through systematically

A VIEW FROM MACHINE LEARNING/AI

The Machine Learning/AI community got into causal inference through a totally different route. It didn't have the baggage of thinking about in-sample estimates as in Fisher's exact test, nor thinking of RCTs as fundamental.

This means they started by completely by-passing potential outcomes/counterfactuals. **In predicting the effects of interventions, counterfactuals are literally “entities multiplied with necessity”**. This will be the focus in this talk.

There must still be something non-standard about any causal language, as causality is not just probability. But instead of independence constraints between factials and counterfactuals, **it suffices to postulate independencies between random variables and intervention variables**.

Framing them in a graph becomes much simpler, giving rise to causal graphs. But remember my First Law of Graphical Models: **the drawing is “just” syntactic sugar!**

CAUSAL GRAPHS

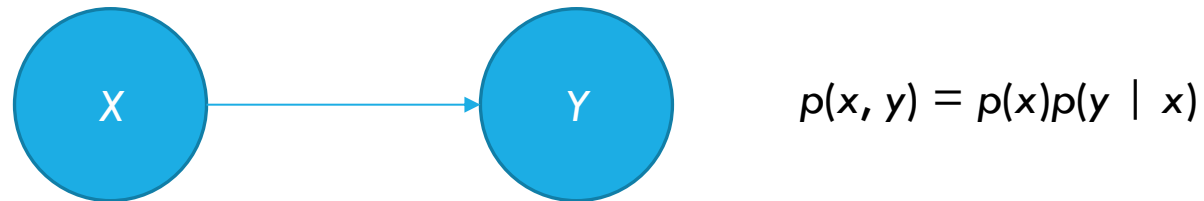
As before, vertices represent random variables. But they can also represent **intervention variables**: “indices” that are fixed by some **external** agent. **It has no causes.**

In the most classical setting, **we define an operation that replaces a random variable with an intervention variable** set at a particular level. The **“do” operator** of Pearl is the most popular formalism, which we will follow.

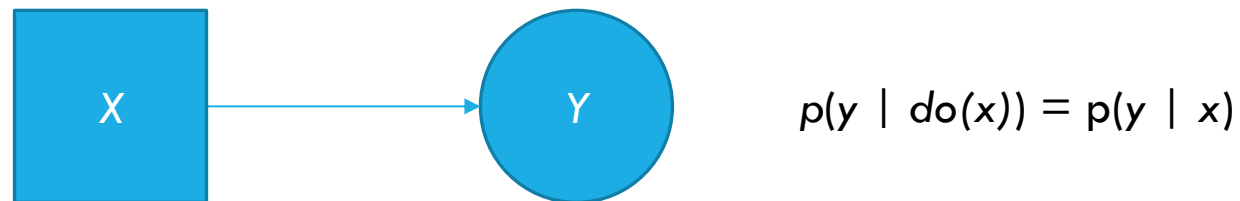
The asymmetry of cause and effect follows naturally from the notion of intervention variables (and do-operators).

ENCODING X CAUSES Y

Use DAGs (directed acyclic graphs) models. This literally boils down to making X a **parent** of Y , and the model is given by the usual DAG factorization.



How is this connected to an intervention? Consider the operation $do(x)$, which replaces X with a fixed value x (we will use squares to represent intervention variables).

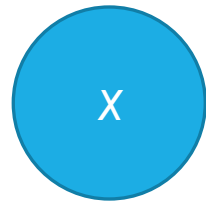


Short for $p(y \mid do(X = x))$, based on context

ENCODING X CAUSES Y

What changes from the “natural state” to the “do(x)” regime? The corresponding factor of X gets erased, all others remain the same.

Implication for $do(y)$:



$p(x \mid do(y)) = p(x)$ because $X \perp\!\!\!\perp Y$ in the intervened graph.

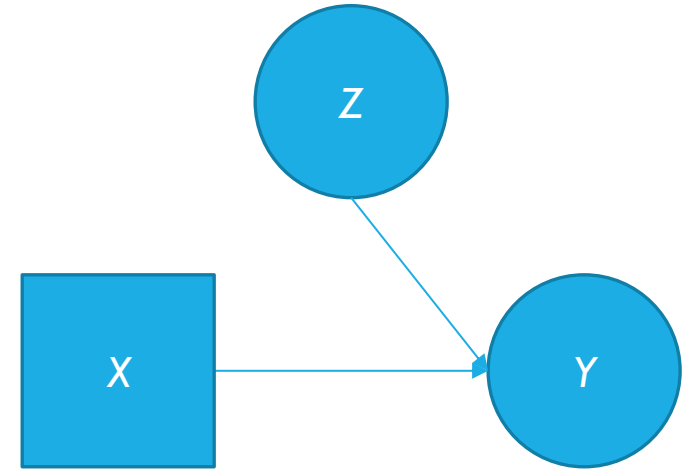
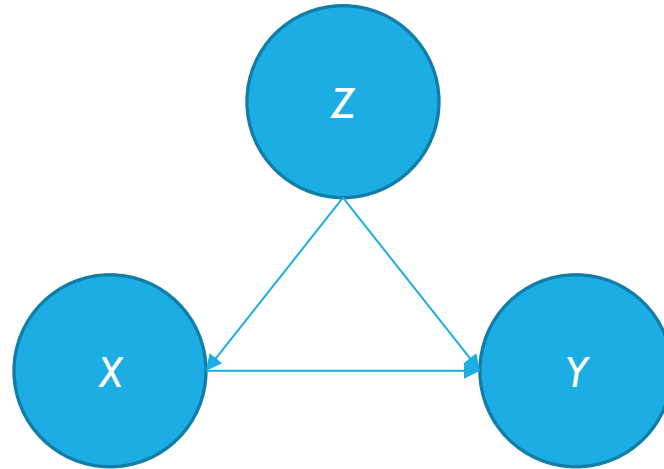
The rules for reading-off independencies are defined to be the same regardless whether we are talking about random variables or intervention variables, but be aware that in the literature it is commonly the case that intervention variables are *not* explicitly represented in a different way.

“ Y does not cause X ” is operationalized as $p(x \mid do(y), context) = p(x \mid context)$ for all y and $context$. The causal graph follows from the independencies raised. The *do* operator/intervention is a **primitive**.

DERIVING IGNORABILITY

The philosophy of causal graphical modeling is that

“We write the model based on what we postulate as How Things Happen. Only then we derive which e.g. ignorability conditions hold, if any.”



$$p(x, y, z) = p(z)p(x | z)p(y | x, z) \quad (\text{always true})$$

$$p(y | x) = \sum_z p(y | x, z)p(z | x) \quad (\text{always true})$$

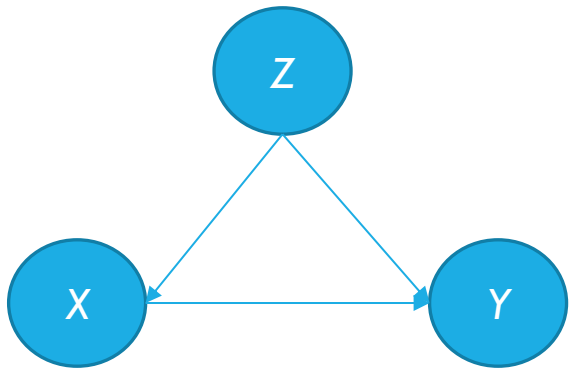
conditional ignorability

$$p(y | do(x)) = \sum_z p(y | x, z)p(z) \quad (\text{warranted by causal graph})$$

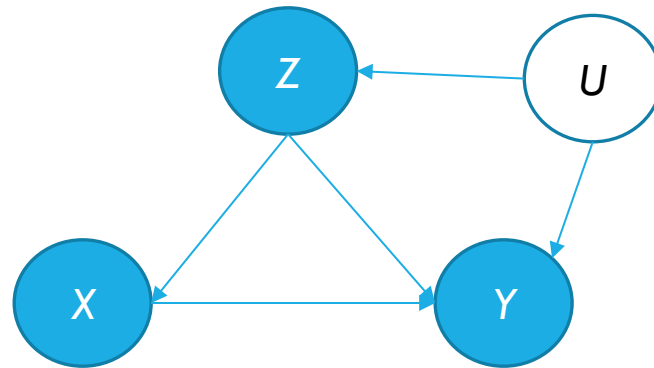
$$p(y | do(x), z) = p(y | x, z) \quad (\text{warranted by causal graph})$$

HOW MUCH DETAIL SHOULD I SPECIFY?

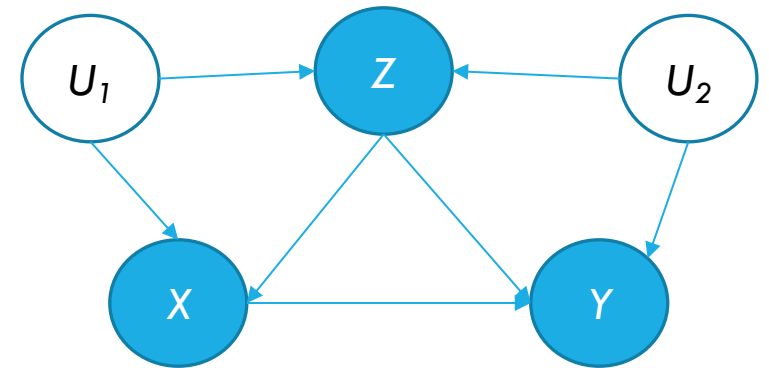
There is nothing stopping you from postulating a **set** of graphs and claiming ignorability holds only if it is implied by all plausible graphs.



Conditional ignorability holds
 $p(y | do(x), z) = p(y | x, z)$



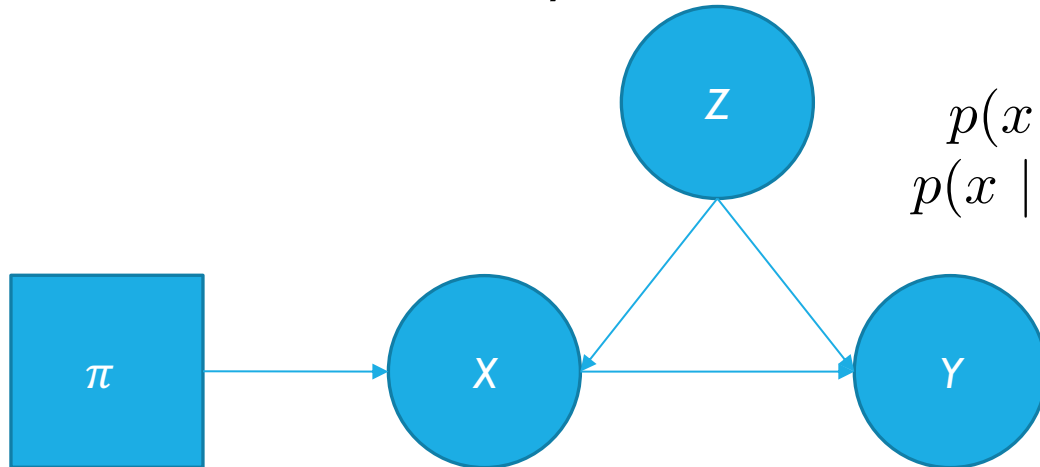
Conditional ignorability holds
 $p(y | do(x), z) = p(y | x, z)$



Conditional ignorability does *not* hold (in general)

CONTEXTUAL INTERVENTIONS AND EXTERNAL INTERVENTION VERTICES

It is very helpful to think of interventions as separate vertices from the random variables modified by them.



$$\begin{aligned} p(x \mid \pi = x^*, z) &= I(x = x^*) \\ p(x \mid \pi = \text{idle}, z) &= p(x \mid z) \quad (\text{“natural regime”}) \end{aligned}$$

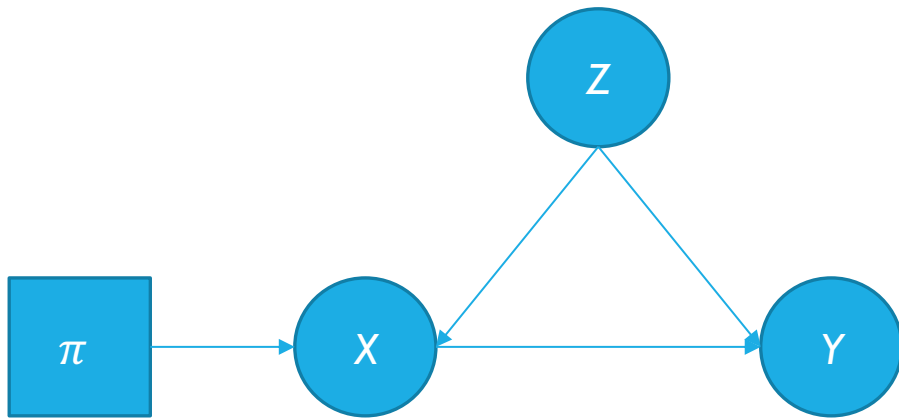
Please notice that I use the “ $do(x)$ ” notation to modify a symbol X that also takes the name of a random variable. I abandon the “ do ” when the variable is always an intervention variable.

See Dawid (2020) for a recent survey, <https://arxiv.org/abs/2004.12493>

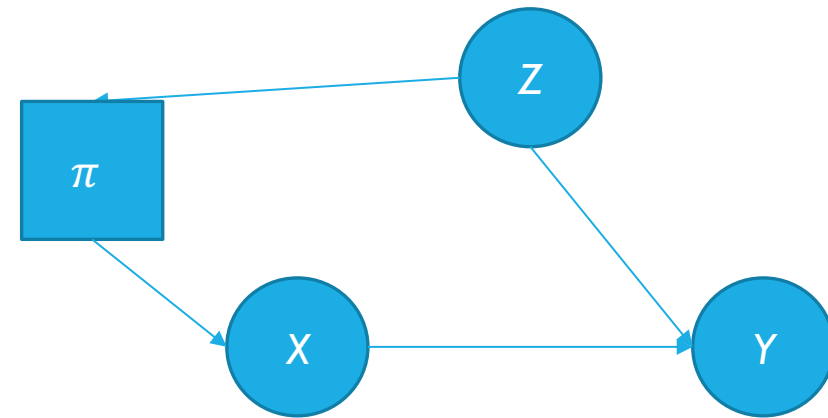
CONTEXTUAL INTERVENTIONS AND EXTERNAL INTERVENTION VERTICES

One reason is that it easily allows for the expression of context-driven policies (i.e., give drug to patient only if no history of heart disease).

Representing intervention variables as having “causes” in the graph is weird. It contradicts the very notion of an external agent with free will!



$$\begin{aligned} p(x \mid \pi = v, z) &= f(v, z) & v \neq \text{idle} \\ p(x \mid \pi = \text{idle}, z) &= p(x \mid z) & \text{("natural regime")} \end{aligned}$$



Please, no!

GRAPHICAL IDENTIFIABILITY

The other main reason: graphical ways of reading-off ignorability (and other criteria) for **causal identification**.

Causal identification just means we can **reduce** a causal expression to a probabilistic-only expression (i.e., a function of random variables only, free of intervention variables).

We did some algebraic manipulations before. But can we just apply a graph algorithm to e.g. test whether conditional ignorability holds. That is, when does

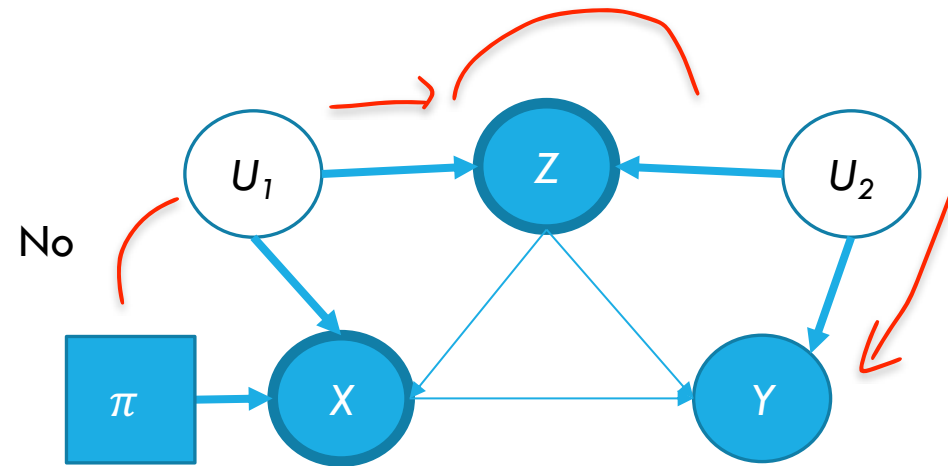
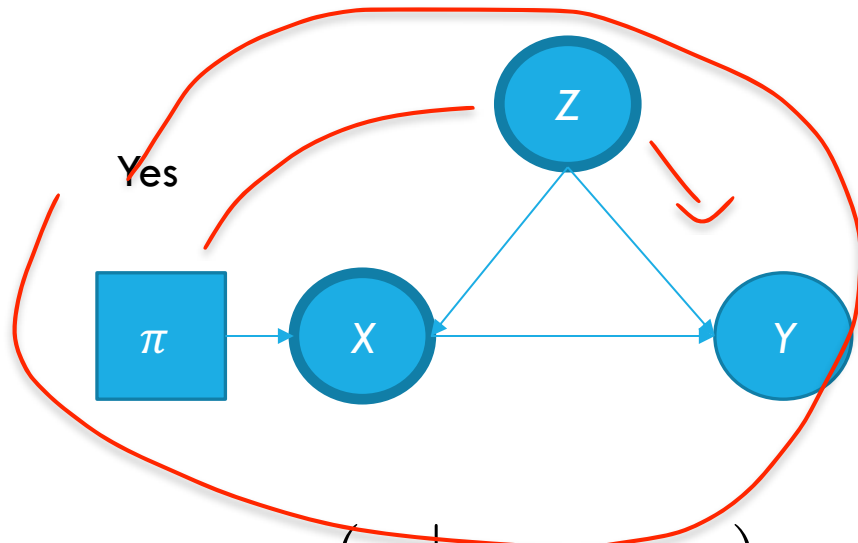
$$p(y \mid do(x), z) = p(y \mid x, z) ?$$

It turns out this is easier by explicitly adding an intervention vertex in the graph.

A GRAPHICAL CRITERION FOR CONDITIONAL IGNORABILITY

Do Z and X **d-separate** π from Y ? That is, can I read-off $\pi \perp\!\!\!\perp Y \mid \{X, Z\}$ from the graph?

Yes? Good. That's it!



$$p(y \mid \underline{\pi = x}, z) = p(y \mid \pi = x, \underline{x}, z) = \underline{p(y \mid x, z)}$$

THE BACK-DOOR CRITERION

Related to that, consider the most basic query: $p(y \mid do(x))$. Is this identifiable from my assumptions?

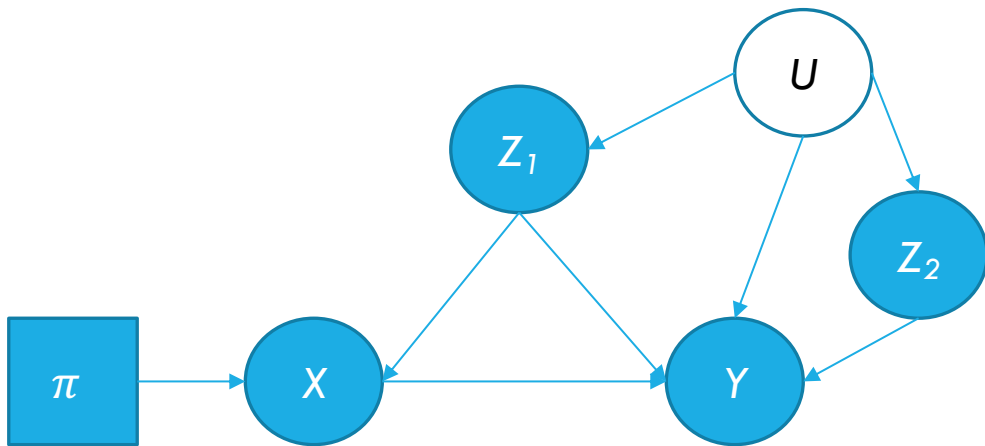
Let us define a valid adjustment set Z^* as one that satisfies the following:

1. Is any element of Z^* a descendant of X ? No? Good.
2. Do Z^* and X d-separate π from Y ? Yes? Good.

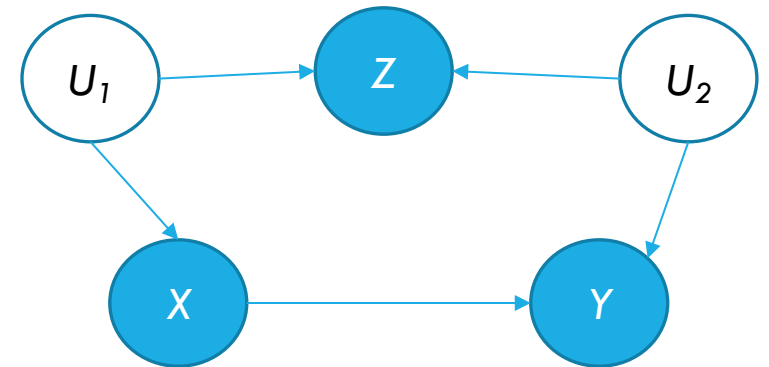
$$\begin{aligned} \underline{p(y \mid do(x))} &= \underline{p(y \mid \underline{\pi} = x)} = \sum_{z^*} p(y \mid \pi = x, \underline{z^*}) \underline{p(z^* \mid \pi = x)} \\ &= \sum_{z^*} p(y \mid \pi = x, x, z^*) \underline{p(z^*)} \quad \leftarrow \begin{array}{l} Z^* \text{ not a} \\ \text{descendant of } \pi \end{array} \\ &= \underline{\sum_{z^*} p(y \mid x, z^*) p(z^*)} \quad \leftarrow \begin{array}{l} \text{The "back-door", or} \\ \text{"standardization" adjustment} \end{array} \end{aligned}$$

EXAMPLE

$\{Z_1\}$ is a valid adjustment set. $\{Z_2\}$ is not.
What about $\{Z_1, Z_2\}$?



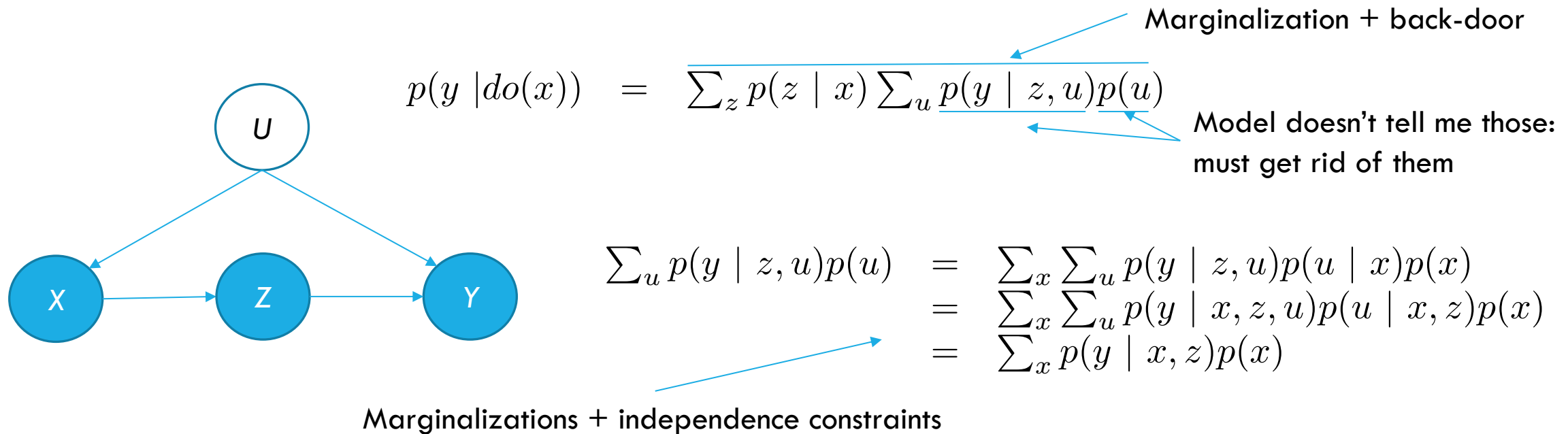
The empty set is a valid adjustment set. $\{Z\}$ is not!



DAG-IDENTIFIABILITY

I myself call “**DAG-identifiability**” the task of solving identifiability only with the graph structure. It is *not* the only way of doing it. See Niki’s talk later.

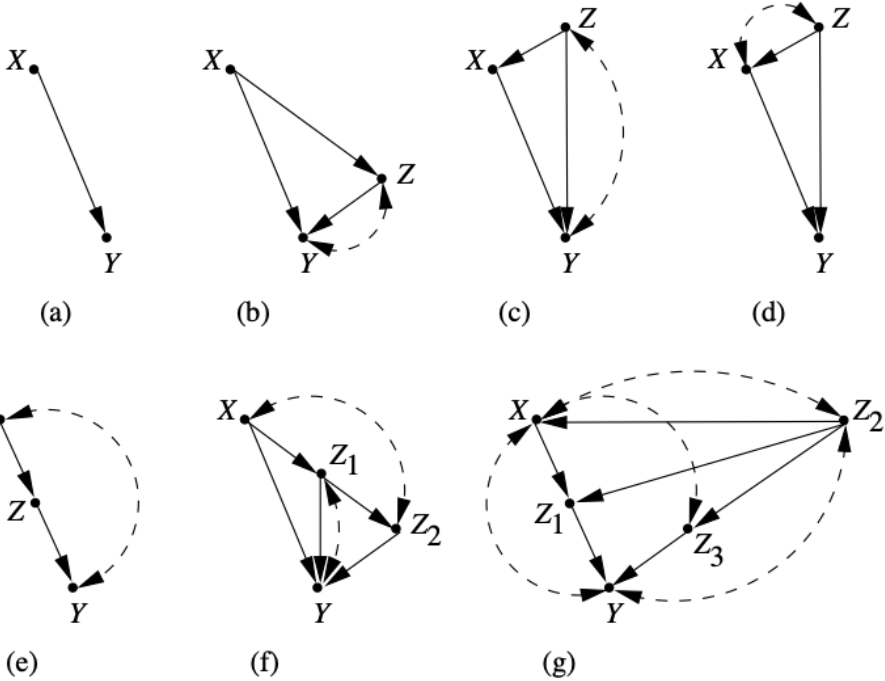
Another example, which allows for hidden confounding: **the front-door criterion**.



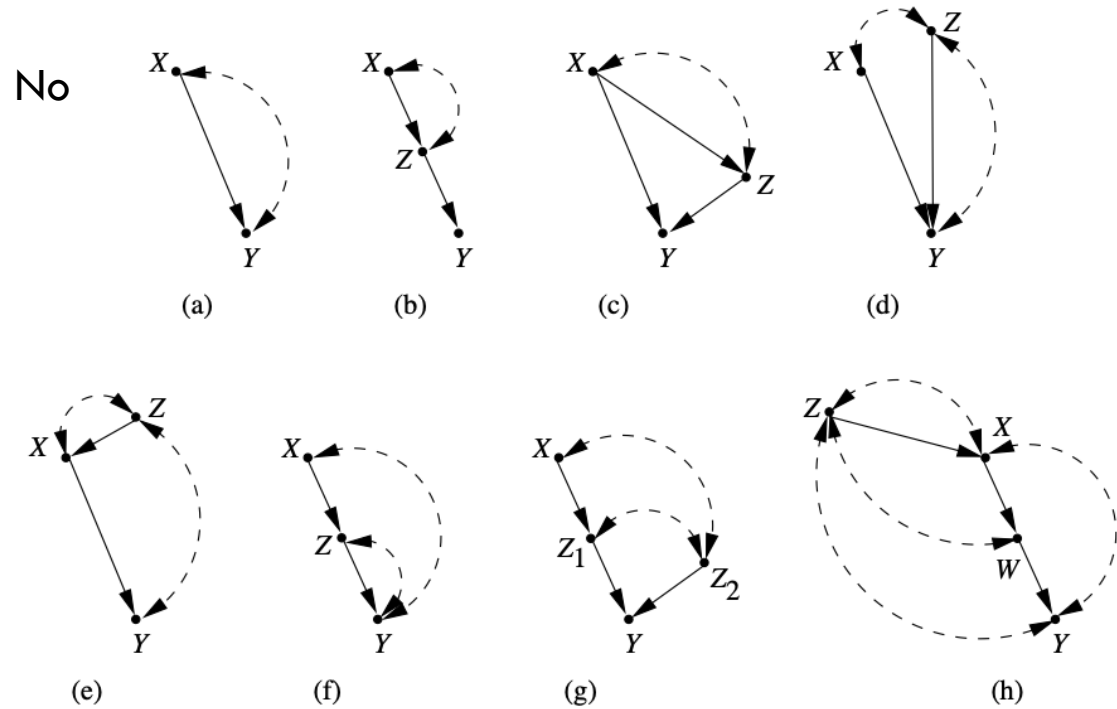
DAG-IDENTIFIABILITY

We apply combinations of conditioning, marginalization, and independence constraints to get to any such a reduction, if one exists.

Yes



No



THE DO-CALCULUS AND THE ID ALGORITHM

There are ways of deriving such results using a handful of rules, and an algorithm that uses them to always provide the correct answer (including “I don’t know”).

The rules are called the **do calculus** (due to Judea Pearl), the algorithm is known as the **ID algorithm** (due to Jian Tian with Pearl). The proof of **completeness** was done independently by Yimin Huang with Marco Valtorta, and by Ilya Shpitser with Pearl.

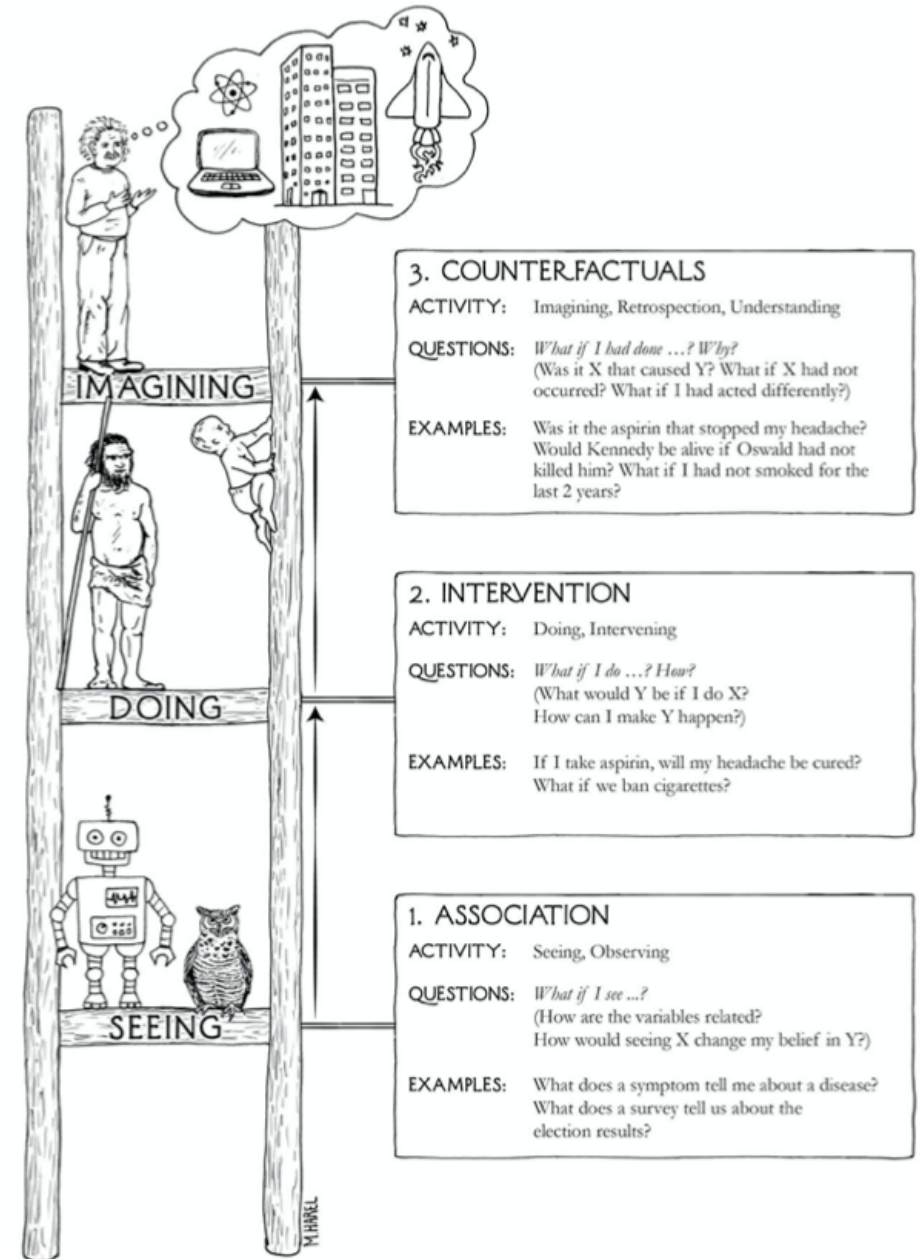
These are some tools that implement the ID algorithm and extensions:

- Causal Fusion (graphical interface): <https://causalfusion.net/>
- DAGitty (graphical interface and R package): <http://www.dagitty.net>
- dosearch (R package): <https://www.rdocumentation.org/packages/dosearch/versions/1.0.4>

THE THIRD LEVEL: COUNTERFACTUALS AND THE STRUCTURAL CAUSAL MODEL

Counterfactuals are baked into the primitives of Neyman-Rubin potential outcome modeling, including the consistency axiom.

Pearl's **Structural Causal Model** (SCM) framework is a constructive way of deriving potential outcomes, independence assumptions, and consistency, from a more fundamental starting point: **structural equations**.



STRUCTURAL CAUSAL MODELS

For every vertex V in a causal graph, postulate a structural equation f_v

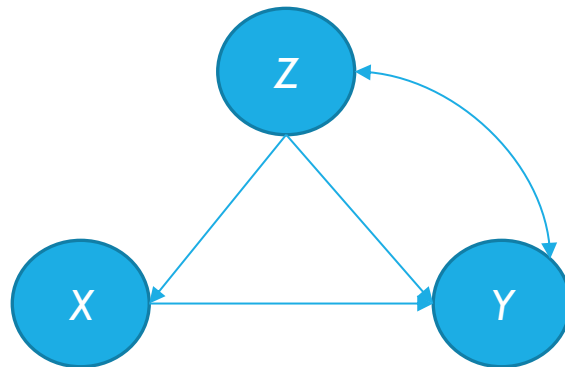
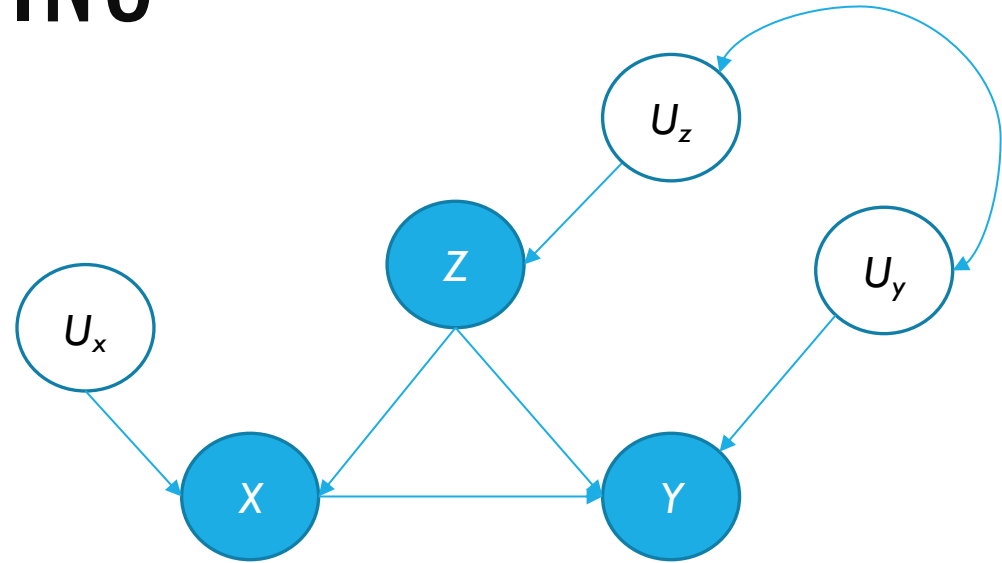
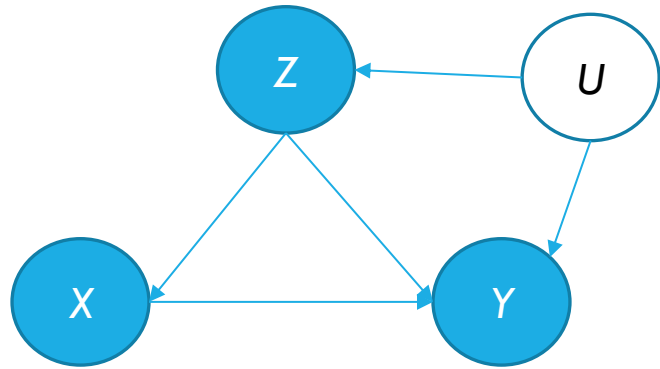
$$V = f_v(\text{parents}(V), U_v)$$

where “parents” are the set of vertices pointing to V in the original graph.

U_v needs not to be a single variable. If anything, it can be infinite-dimensional! Such variables have no causes from “inside” the system, and sometimes are called **exogenous**. Confusingly, as they don’t need to be independent of each other. I’ll call them **background** variables.

We can represent the background factors in the graph too, with bi-directed edges among them or between the respective observables.

THREE WAYS OF REPRESENTING UNMEASURED CONFOUNDING



WHAT DO WE GAIN, AND WHAT DO WE PAY?

As I've mentioned, potential outcomes are optional if all we want is an interventional distribution.

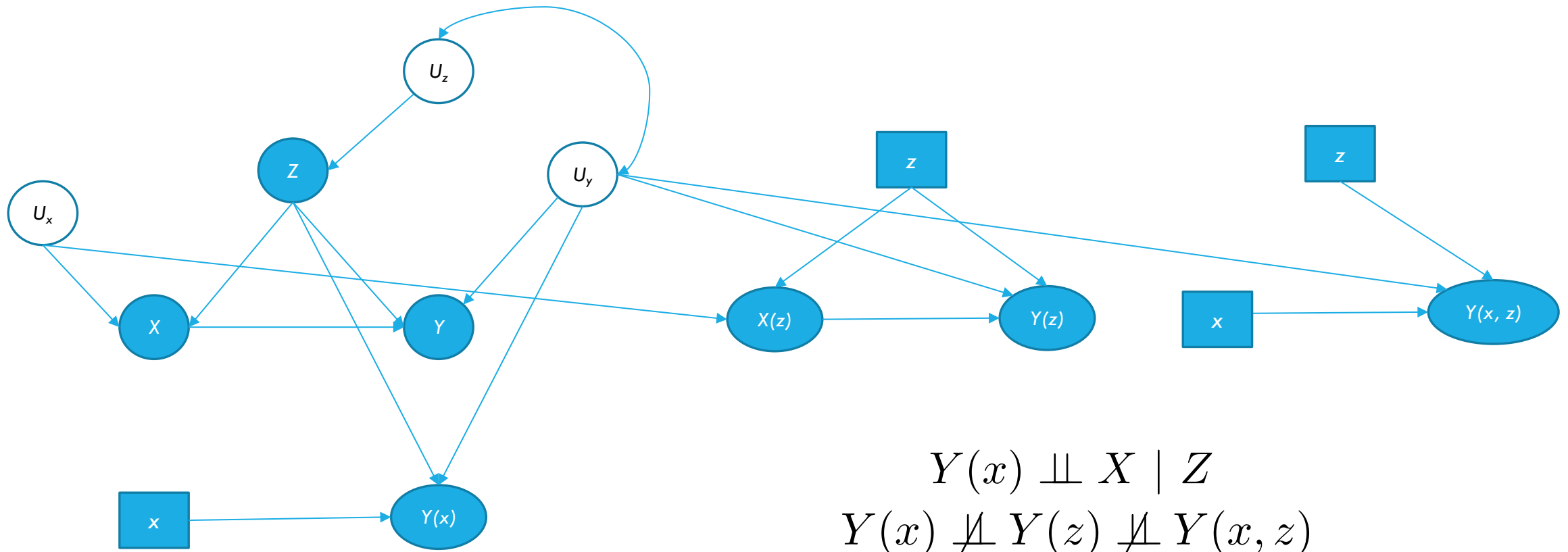
Two main reasons for a SCM as opposed to a plain causal graphical model:

- When we are genuinely interested in counterfactual estimands
- To make easier to add stronger assumptions that aid identifiability (not fundamental, but handy)

As this is based in determinism, it requires **assumptions that cannot be fully testable**, unless we were to measure every single possible cause under the sun!

TWIN NETWORKS

A graphical construction of joint distributions of factu-als and counterfactuals.



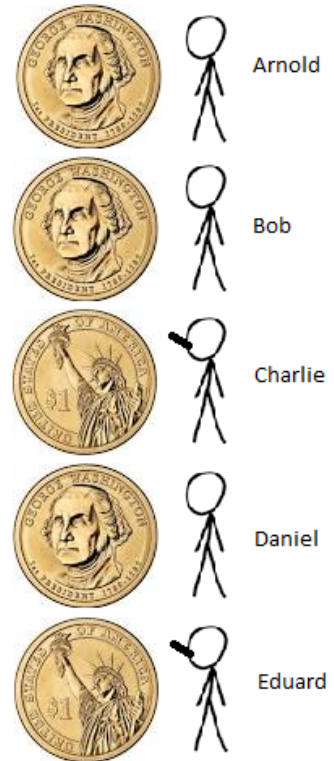
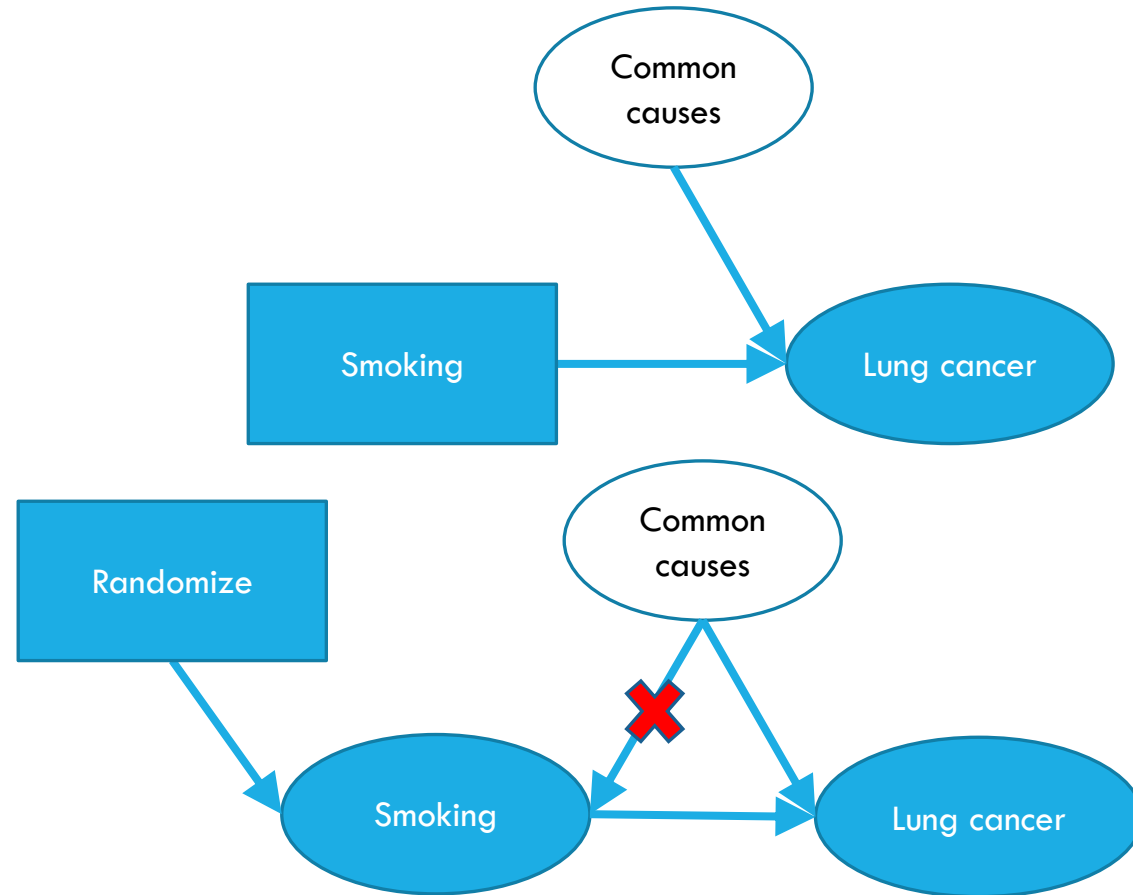
ESTIMANDS AND THE CAUSAL PIPELINE



ESTIMANDS

As in any statistical problem, the **estimand** is the unknown we want to estimate. It can be as simple as the expected outcome under intervention.

RCTs are not fundamental. They're "just" a way of sampling data. **It is an important and very useful device, but it doesn't define the question** in the same way that a sampling strategy doesn't define the target of a survey.



THE AVERAGE TREATMENT EFFECT AND OTHER ESTIMANDS

We saw the **expected outcome under intervention**:

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

This could as well be another summary of the distribution $p(y \mid do(x))$.

We also saw the **average treatment effect (ATE)**:

$$\mathbb{E}[Y \mid do(X = x)] - \mathbb{E}[Y \mid do(X = x')]$$

We can also ask the same questions conditioning on some **pre-treatment** outcomes, for instance the **conditional average treatment effect (CATE)**:

$$\mathbb{E}[Y \mid do(X = x), Z = z] - \mathbb{E}[Y \mid do(X = x'), Z = z]$$

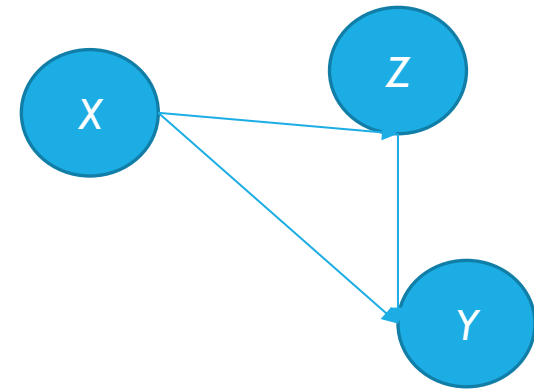
CONDITIONAL OUTCOMES

What about when Z happens *after* treatment? Is

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

sensible? **It depends what you want to do with it.**

Do you want a prediction of Y taking place in the not-so-near future, after intervening on X and waiting for the event corresponding to Z ? That's fine.



CAUSAL EFFECTS

What about we want to **compare outcomes under different levels of intervention**? This is not only a prediction under an intervention, but a **causal effect**.

Will the CATE still make sense when Z is post-treatment? We have two ways of seeing why this doesn't make sense.

From a **decision-theoretical perspective**, it makes no sense to contrast decisions using **information unavailable at the moment of the decision**.

From a counterfactual perspective, it also doesn't make sense. For any single individual, $Z(x) = f_z(x, U_z)$. We can't have $Z(x) = Z(x')$ for any non-trivial structural equation for a single individual. Hence, this **CATE is just comparing two disjoint groups of people who happen to coincide on Z** . This again is of dubious interest, to say the least. We say this type of estimand is “not a causal effect”.

CASE STUDY: SIMPSON'S PARADOX

Let's give an example on the role a causal model clarifying the choice of particular estimands.

Consider **Simpson's paradox**.

$$\begin{aligned} Pr(E | F, C) &< Pr(E | F, \neg C) \\ Pr(E | \neg F, C) &< Pr(E | \neg F, \neg C) \\ Pr(E | C) &> Pr(E | \neg C) \end{aligned}$$

Would you recommend the drug?
Which quantity is the relevant one?

	Combined	<i>E</i>	$\neg E$		Recovery Rate
(a)	Drug (<i>C</i>)	20	20	40	50%
	No drug ($\neg C$)	16	24	40	40%
		36	44	80	
	Males	<i>E</i>	$\neg E$		Recovery Rate
(b)	Drug (<i>C</i>)	18	12	30	60%
	No drug ($\neg C$)	7	3	10	70%
		25	15	40	
	Females	<i>E</i>	$\neg E$		Recovery Rate
(c)	Drug (<i>C</i>)	2	8	10	20%
	No drug ($\neg C$)	9	21	30	30%
		11	29	40	

Figure 6.1, Pearl (2009). *Causality*, CUP

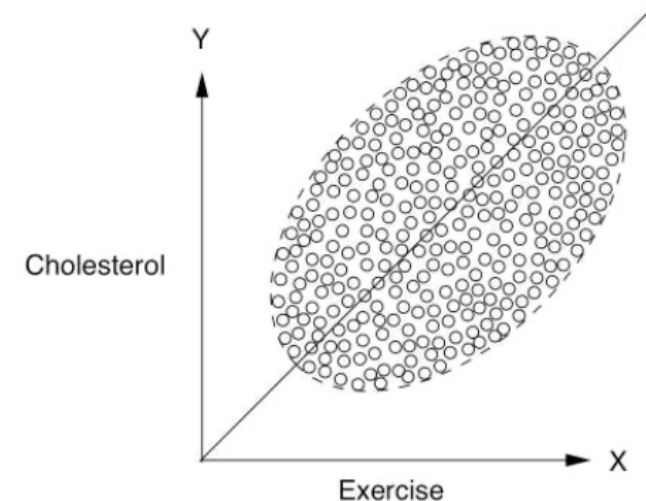
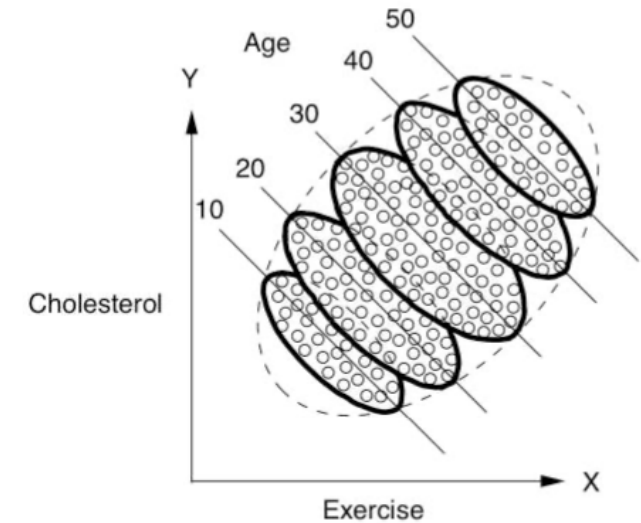
CASE STUDY: SIMPSON'S PARADOX

There is nothing paradoxical about association reversal.

What then, should guide us to select (actually, define!) the appropriate functional?

Notice: this has nothing to do with estimation. We assume we have access to the population distribution.

Pearl and Mackenzie (2018). *The Book of Why*, Chapter 6.



MODELING THE SYSTEM

Your assumptions and your question should be independent. Your question boils down to what is higher, $Pr(E \mid do(C))$ or $Pr(E \mid do(\neg C))$?

The answer = assumptions + question. What's yours?

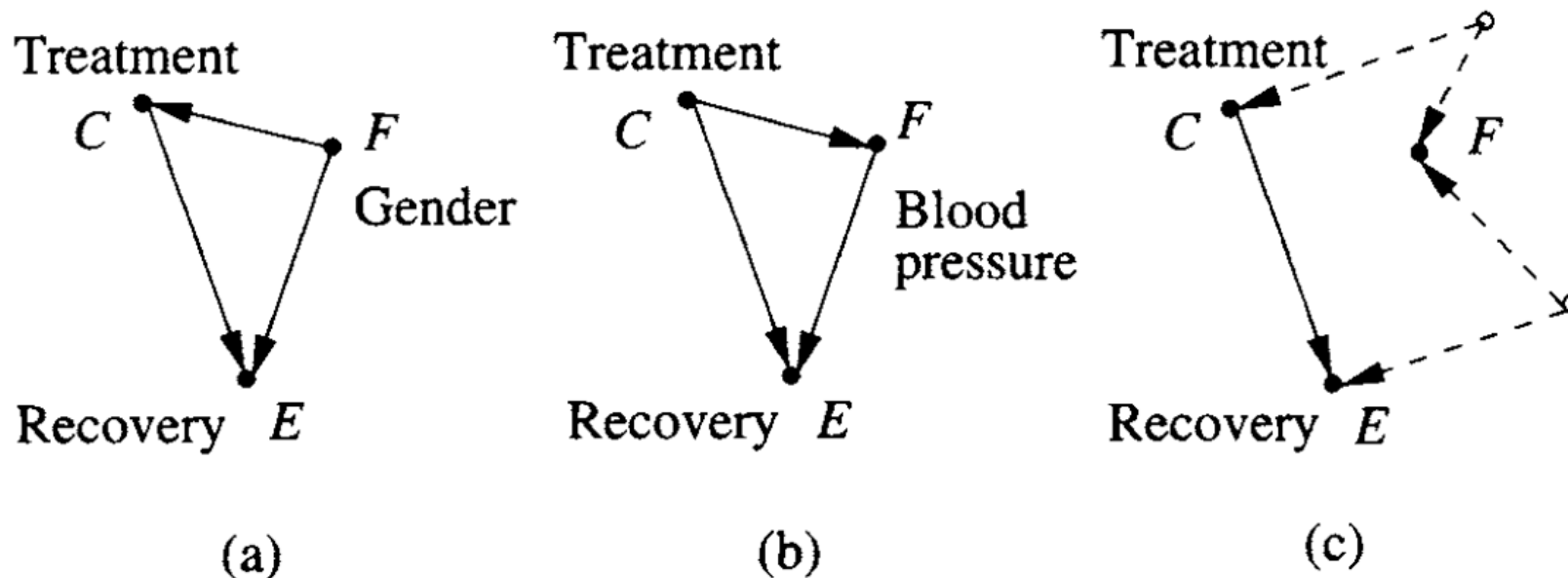


Figure 6.2, Pearl (2009). *Causality*, CUP

DISARMING THE (PSYCHOLOGICAL) PARADOX WITH CAUSAL REASONING

Why does it *feel* like a paradox?

Let our population have subpopulations F and $\neg F$. And let our treatment C not cause changes in the distribution of the subpopulations.

$$Pr(F \mid do(C)) = Pr(F \mid do(\neg C)) = Pr(F)$$

Then for outcome E , it is impossible to have, simultaneously,

$$\begin{aligned} Pr(E \mid do(C), F) &< Pr(E \mid do(\neg C), F) \\ Pr(E \mid do(C), \neg F) &< Pr(E \mid do(\neg C), \neg F) \\ Pr(E \mid do(C)) &> Pr(E \mid do(\neg C)) \end{aligned}$$

LESSONS: THE CAUSAL PIPELINE

Keep this cheat sheet to help you formulate and solve your problems of interest. Consider separating the following components within your solution:

1. **The Estimand:** the quantity you want to learn. For example, the ATE.
2. **The Model:** assumptions linking observable signal to the Estimand by solving a (partial) identification problem. For instance, a SCM.
3. **The Estimator:** the way data is used to infer an estimate of the Estimand given the assumptions of the Model. For instance, regression is often used, as we will see.
4. **The Algorithms:** the computational procedures to solve the identification problem (mapping Model to Estimand), and to compute the output of the Estimator. For instance, the **ID algorithm** and optimization methods.

You will be a happier researcher if you separate these ingredients before pipelining them...

**ESTIMATING CAUSAL EFFECTS:
THE SINGLE-SHOT,
NO UNMEASURED CONFOUNDERS CASE**

THE SINGLE-SHOT, MEASURED CONFOUNDERS CASE

This is by far the simplest and most common case: we have treatment X , outcome Y , and a bunch of covariates Z that satisfy the back-door criterion for $p(y \mid \text{do}(x), z)$.

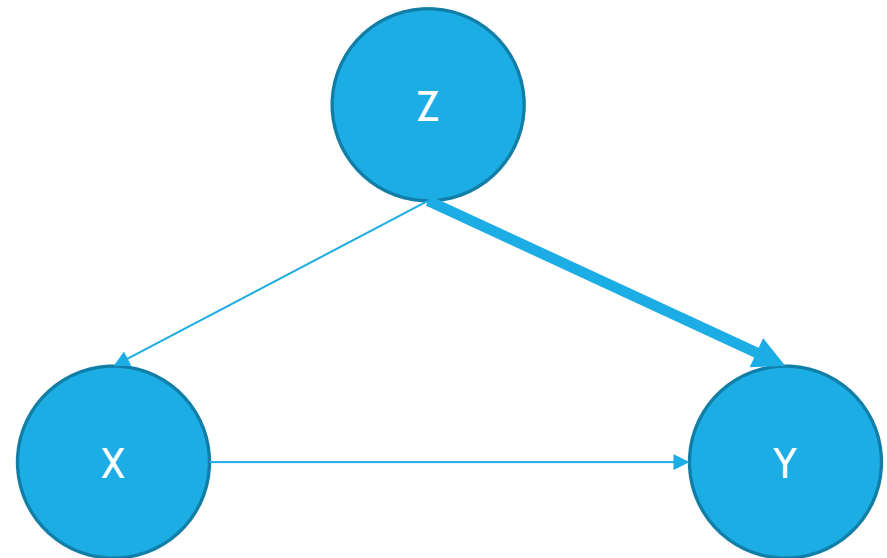
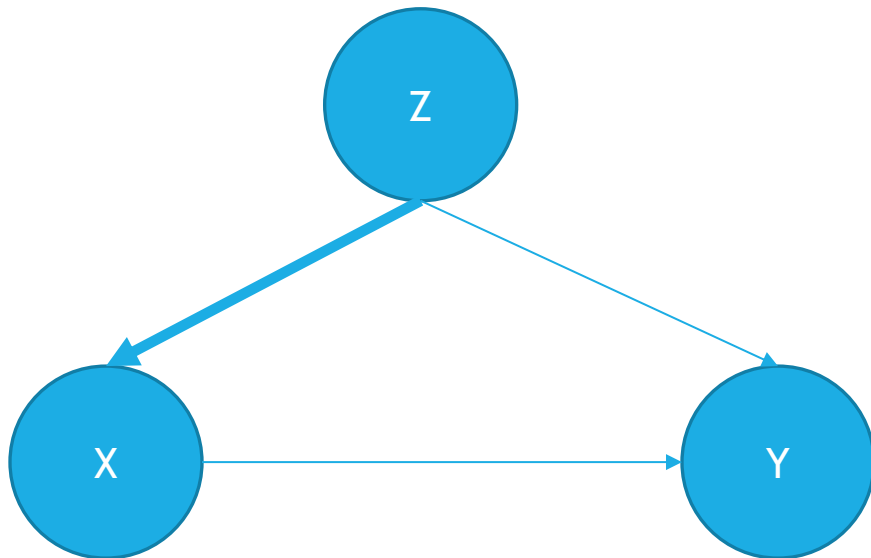
The most common causal effect of interest will cover is the ATE/CATE when X is binary, or more generally, when we have a “control” value x vs a “treatment” value x' .

Estimation in causal inference is related to, but not quite the same as, prediction. You maybe be interested in the expectation of the outcome of a treatment by itself, or just the contrast with respect to the baseline. This can motivate different methods.

THE TWO MAIN BUILDING BLOCKS

Z can be high dimensional. What now?

We can start with one that focus on the **treatment assignment model**, $p(x | z)$, or on the **outcome model**, $p(y | x, z)$.



BUILDING ON THE TREATMENT ASSIGNMENT MODEL

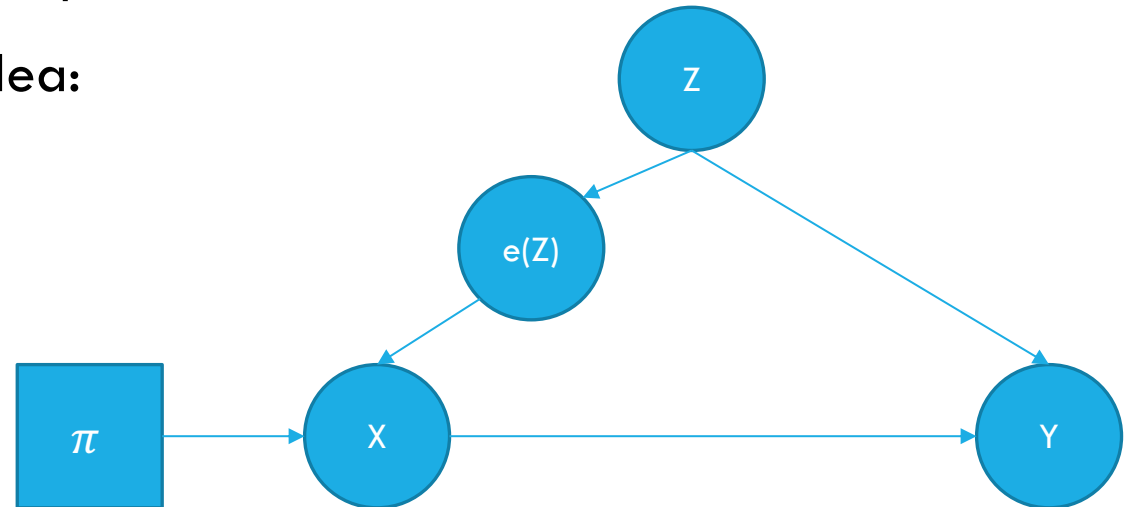
Let's do a sleight of hand and define a quantity that is also a function of the model, the **propensity score**:

$$e(z) = Pr(X = 1 \mid Z = z)$$

(yes, we can generalize it to non-binary X).

It is not hard to show that $\{e(z)\}$ is a valid adjustment set for the back-door criterion!

Here is a very informal depiction of the idea:



THE INVERSE PROBABILITY WEIGHTING (IPW) ADJUSTMENT

Recall the difference between $p(y \mid x)$ and $p(y \mid do(x))$ as a function of the model:

$$p(y \mid x) = \sum_z p(y \mid x, z)p(z \mid x) \quad p(y \mid do(x)) = \sum_z p(y \mid x, z)p(z)$$

The estimation problem is the fact that we don't know any of these factors. This suggests the following idea:

1. Recognize that the data follows the distribution providing the factors on the left, above
2. That being the case, "reweight" the data by replacing the factor $p(z \mid x)$ with $p(z)$.

IPW IDENTITIES IN THE POPULATION

back-door formula

$$\begin{aligned}\mathbb{E}[Y \mid do(X = 1)] &= \sum_{y,z} y \times p(y \mid 1, z)p(z) \\ &= \sum_{x,y,z} I(x = 1) \times y \times p(y \mid x, z)p(z) \\ &= \sum_{x,y,z} I(x = 1) \times y \times p(y \mid x, z) \frac{p(x \mid z)}{p(x \mid z)} p(z) \\ &= \sum_{x,y,z} I(x = 1) \times y \times \frac{1}{p(x \mid z)} \times p(x, y, z) \\ &= \mathbb{E} \left[\frac{I(X = 1)Y}{p(X \mid Z)} \right]\end{aligned}$$

a useful trick that will allow us to get rid of the outcome model!

AN ESTIMATE VIA IPW

The beauty of that derivation is that it completely disregards $p(y \mid x, z)$. So we can get all we need from

1. The **empirical distribution** of (X, Y, Z)

That is, each data point in the training data get probability $1 / n$, and anything out of the data gets probability zero

2. An estimate of $p(x \mid z)$, that we can get by plug-in whatever method we want

$$\begin{aligned}\widehat{ATE} &= \widehat{\mathbb{E}}[Y \mid do(X = 1)] - \widehat{\mathbb{E}}[Y \mid do(X = 0)] \\ &= \frac{1}{n} \sum_{i=1}^n \frac{I(x^{(i)} = 1)y^{(i)}}{\hat{p}(x^{(i)} \mid z^{(i)})} - \frac{1}{n} \sum_{i=1}^n \frac{I(x^{(i)} = 0)y^{(i)}}{\hat{p}(x^{(i)} \mid z^{(i)})}\end{aligned}$$

ATTACKING THE PROBLEM WITH THE OUTCOME MODEL

This is more straightforward. Say that

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

Therefore

$$\begin{aligned}\mathbb{E}[Y \mid do(x)] &= \sum_{y, z} y \times p(y \mid x, z)p(z) \\ &= \sum_z \mathbb{E}[Y \mid x, z]p(z) \\ &= \sum_z f(x, z)p(z)\end{aligned}$$

ESTIMATION

Again, we can get an estimate by just using the empirical distribution (it suffices to be the one of Z alone) and plugging-in an estimate of the regression function.

$$\begin{aligned}\widehat{ATE} &= \widehat{\mathbb{E}}[Y \mid do(X = 1)] - \widehat{\mathbb{E}}[Y \mid do(X = 0)] \\ &= \frac{1}{n} \sum_{i=1}^n \hat{f}(1, z^{(i)}) - \frac{1}{n} \sum_{i=1}^n \hat{f}(0, z^{(i)})\end{aligned}$$

SO, WHICH ONE IS BEST?

Each is theoretically sound, and each can fail badly.

Misspecification can affect both, but empirically it tends to affect outcome models more strongly. In many cases, it is also the case that in some problems we know the treatment assignment model (when it is actually logged. We will see examples in **bandit models** later one).

IPW can suffer of major variance because of the denominator. Outcome models are less prone to high variance, but they cannot escape the problem of **lack of overlap**.

THE LACK OF OVERLAP PROBLEM

We saw that causal learning from observational data is a type of extrapolation problem. But that's only half of the story.

Another major source of extrapolation typically arises by **the mere fact that X and Z are associated in observational data.**

That's because a functional like the **ATE requires evaluating combinations of (X, Z) that do not need to have much support in the observational data!**

When reading causal inference books/papers/package documentations you will find one typical assumption, the **overlap assumption:**

$$Pr(X = x \mid z) > 0 \text{ for all points } (x, z) \text{ of interest}$$

THE LACK OF OVERLAP PROBLEM

This looks inoffensive but hides something potentially much more harmful. **When in the data this probability is close to zero (or 1)**, in the corresponding ATE estimate

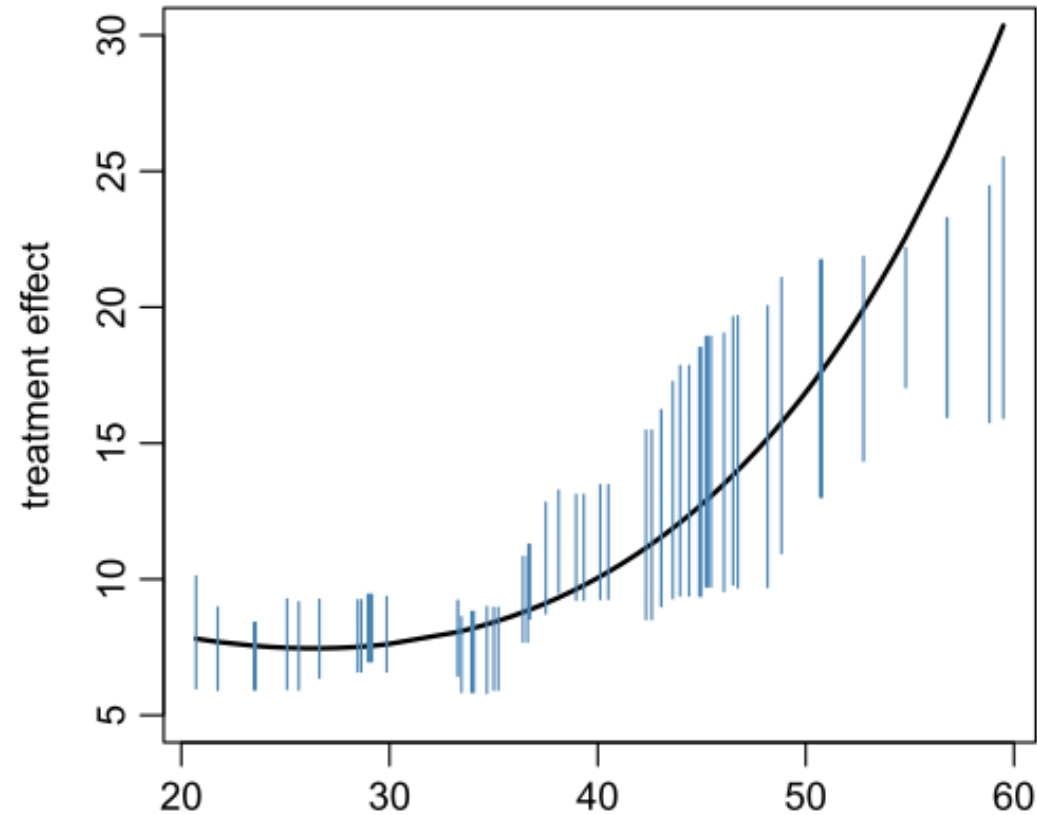
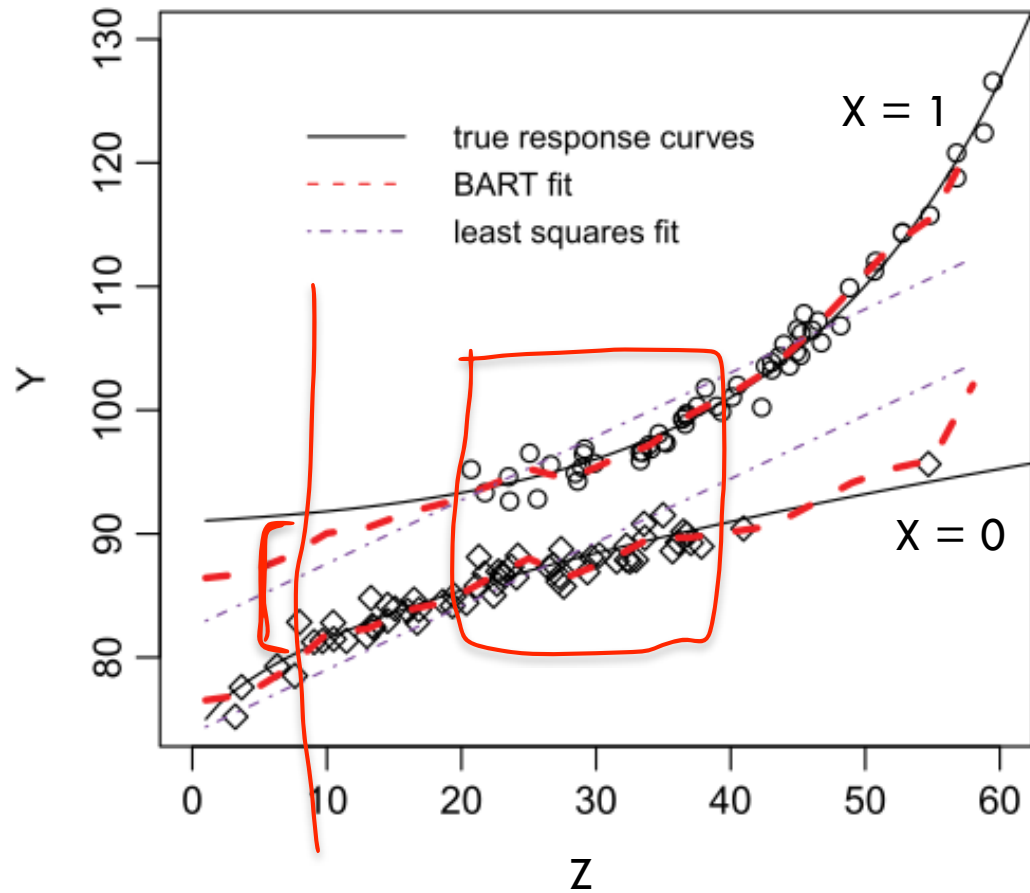
$$\widehat{ATE} = \widehat{\mathbb{E}}[Y \mid do(X = 1)] - \widehat{\mathbb{E}}[Y \mid do(X = 0)]$$

at least one of the two terms above might be meaningless, in the sense that the point estimate might be pretty much a wild extrapolation.

This will definitely be the case for IPW. For the outcome model approach, it *may* be OK if for example we are using a linear model that we believe holds for all plausible combinations of X and Y . But would you trust this assumption?

EXAMPLE: REGRESSION TREES

Linearity within data can be falsified, but linearity in extrapolation can't.



Hill (2011) "Bayesian nonparametric modeling for causal inference", *Journal of Computational and Graphical Statistics*.

NUISANCE PARAMETERS IN CAUSAL INFERENCE

Nuisance parameter is a term in Statistics that can be roughly described as “parameters we don’t care about, but which are there because they define/identify the ones we do care about.”

The most famous example might be the variance of the error term in a problem where all we care is the regression function. Likelihoods by definition will require the nuisance variance. Empirical Risk Minimization can get away without it.

It should separate inference of nuisances from inference of the parameters of interest as much as possible. This is particularly the case in many causal inference problems where the effects might be weak and the dimensionality of Z high.

EXAMPLE: LIKELIHOOD-BASED CATE IN *LINEAR* MODELS

It doesn't get any simpler than that.

Consider the following model for the outcome regression with a possible continuous treatment X :

$$Y = \alpha x + \beta^T z + \epsilon$$

It is not difficult to show that for any Δ ,

$$\mathbb{E}[Y \mid do(X = x), z] - \mathbb{E}[Y \mid do(X = x - \Delta), z] = \alpha \Delta$$

So, not only the variance of the error term would be a nuisance, but the entire vector β is! What are the implications? Pretty minimal if Z is small dimensional, but for high-dimensional Z we will want to do some sort of regularization/Bayesian modelling.

REGULARIZATION... AT A PRICE

Hahn et al. (2018). "Regularization and confounding in linear regression for treatment effect estimation". Bayesian Analysis.

When Z is high dimensional, we don't really have much of a choice but to regularize β . For instance, by ridge regression (or the Bayesian related idea of putting Gaussian priors on parameters).

This will **bias** the parameters, in the technical sense. More precisely,

$n \times p$ matrix of covariates

$$\text{bias}(\hat{\alpha}_{rr}) = - \left(\underbrace{(x^T x)^{-1} x^T \mathbf{Z}}_{\text{coefficients of one-at-a-time regression of each } Z_j \text{ on } x} \right) \left(\mathbf{I}_p + \mathbf{Z}^T \underbrace{(\mathbf{Z} - \hat{\mathbf{Z}}_x)}_{\text{\(n \times p\) matrix of residuals of the one-at-a-time regression}} \right)^{-1} \beta$$

Bias of ridge regression (rr):
difference between true value
and estimate, in expectation

coefficients of one-at-a-time
regression of each Z_j on x

$n \times p$ matrix of residuals of the
one-at-a-time regression

UNPACKING IT

The most striking feature of this bias is that the more strongly associated X and Z are, and the higher the dimensionality of Z is, the worse the bias is.

This wouldn't matter that much if the goal was purely predictive. We don't care about the bias of a particular parameter, but the total bias + variance across all parameters forming our prediction function.

But observational studies are the perfect storm!

- X and Z are associated, because otherwise Z wouldn't matter as a confounder
- We should expect many confounders, so Z is naturally high-dimensional
- We care mostly about the causal effect here, so what happens to the other parameters is irrelevant as long as I get a good causal effect
- But we can't just completely ignore the other parameters, because otherwise our causal effect parameter is not even defined

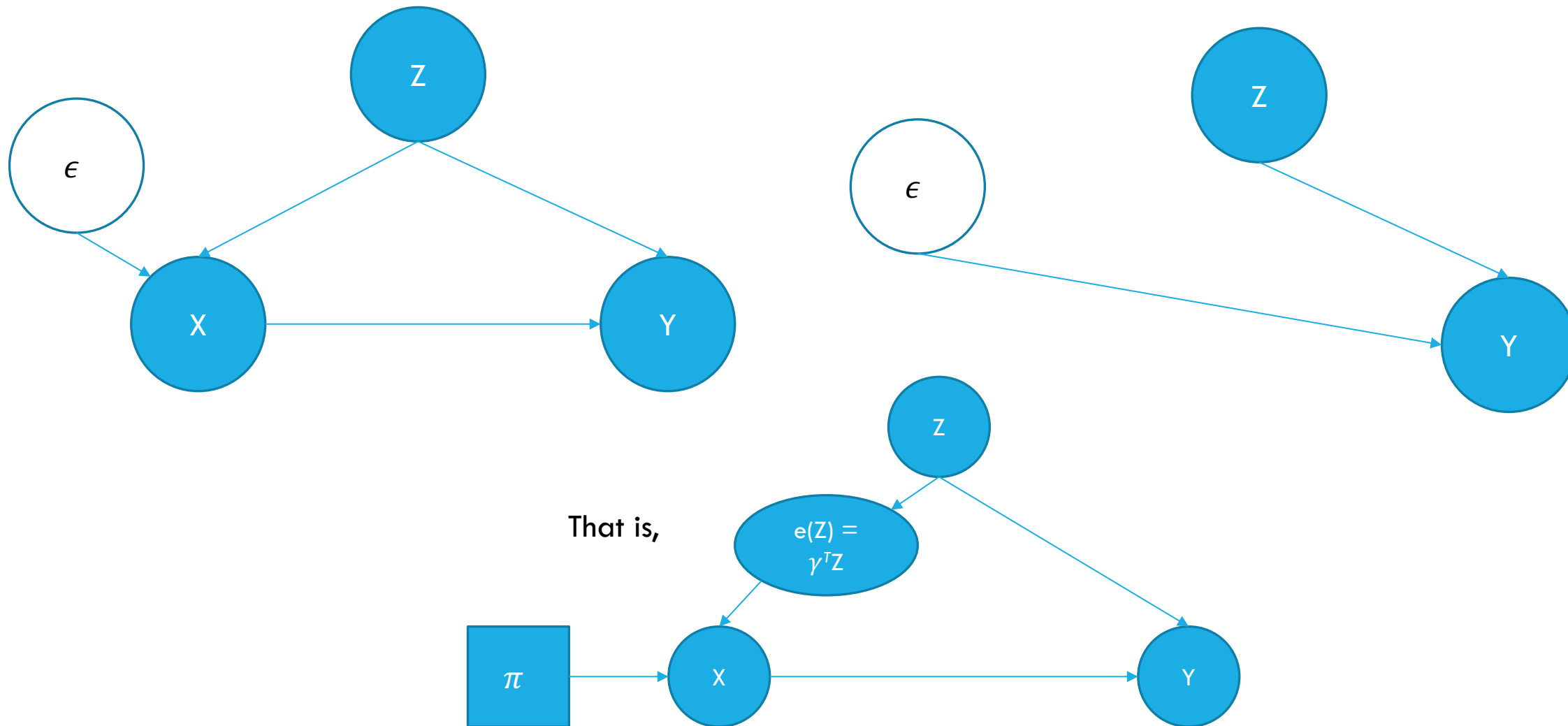
MITIGATING THE BIAS

The main idea behind what I'll describing is better intuitively understood in linear models, but it forms the basis of many methods:

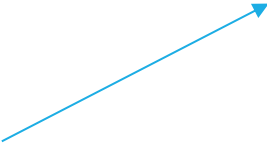
Use the treatment assignment model to modify the problem, so that the dependency between X and Z doesn't play a major role in a regression formulation of the back-door formula.

What does it mean? In the linear case, we can reparameterise the model so that we simulate an alternative regression function that still encodes the causal parameter.... but doesn't regress on X (directly).

REPARAMETERIZING THE LINEAR MODEL



REPARAMETERIZING THE LINEAR MODEL

$$Y = \alpha(X - \gamma^T Z) + (\beta + \alpha\gamma)^T Z + \epsilon_y = \alpha R_x + \beta_\alpha Z + \epsilon_y$$


solve Bayesian/regularized regression for the (R_x, Z) data, get alpha

NON-LINEAR MODELS

What would the equivalent be for non-linear models, particularly when the treatment is discrete?

The simplest idea is to just **include the propensity score as yet another covariate for your regression, then do regression**. It's not the "optimal" in a theoretical sense, but it's simple and may be just good enough.

In general, we will want to separate the "stuff we care" from "the stuff that we don't care about". Straightforward Likelihood + prior/regularization + Bayesian/cross-validation crank is not that off-the-shelf here.

GOING BEYOND NAÏVE LEARNING: NON-LINEARITIES WITH BINARY TREATMENT, FOR SIMPLICITY

Without loss of generality, we can rewrite the outcome model as

$$\mathbb{E}[Y \mid x, z] = f_0(z) + x\tau(z)$$

for x in $\{0, 1\}$. So CATE can be given by

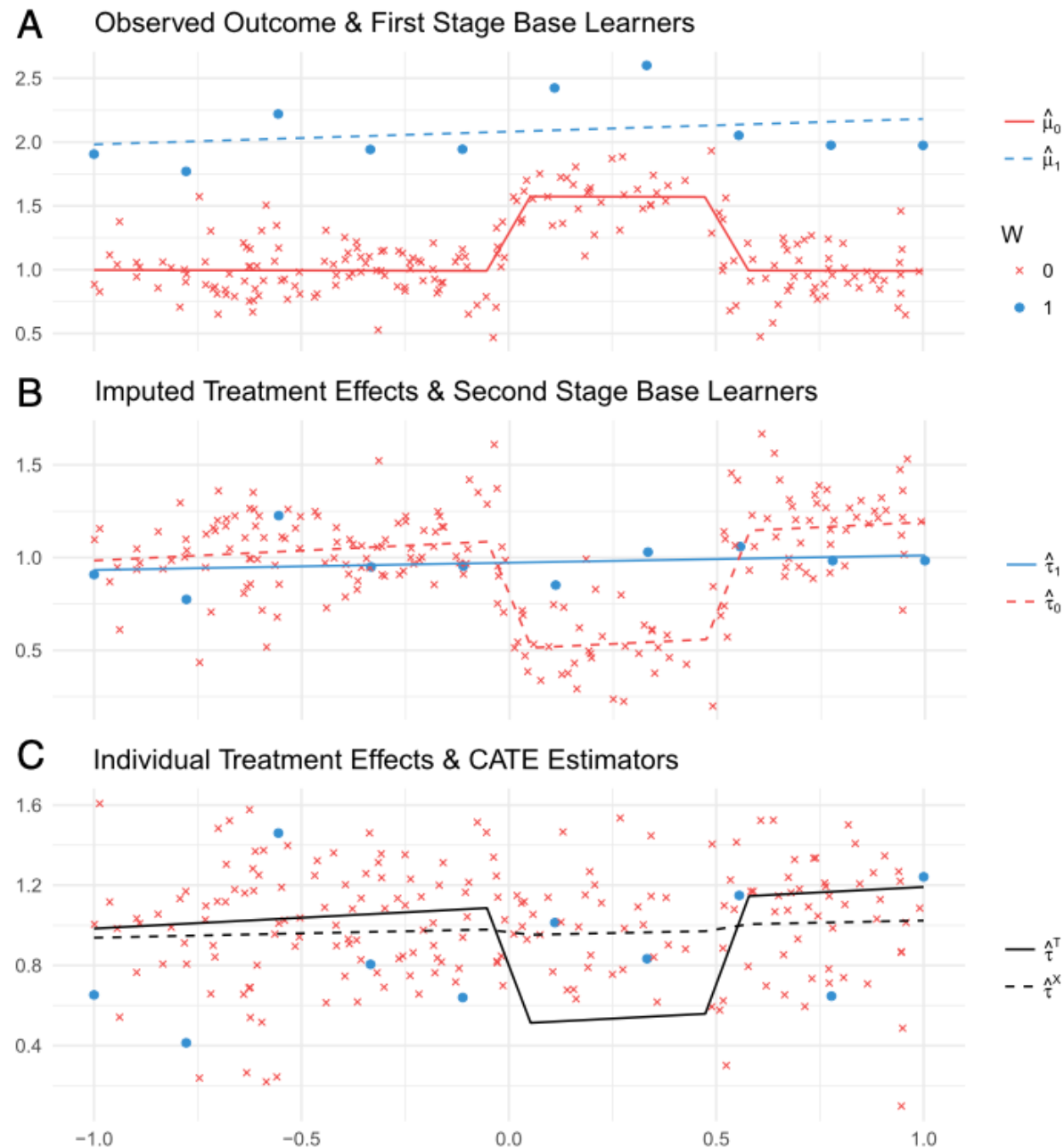
$$\mathbb{E}[Y \mid do(X = 1), z] - \mathbb{E}[Y \mid do(X = 0), z] = \tau(z)$$

Two things come to mind:

- The response function $\mathbb{E}[Y \mid do(x), z]$ can be complex and non-linear while the causal effect $\tau(z)$ can be smooth and very close to linear in z .
- In many applications, we don't even care about the **baseline response** $f_0(z)$.

EXAMPLE

Kunzel et al. (2019) *PNAS*,
<https://www.pnas.org/content/116/10/4156>



CUTTING TO THE CHASE: R-LEARNER

The idea of the R-learner is that we don't care about the inner structure of the contribution from Z , just its black-box behavior. So fit it separately!

Remember **Leo Breiman's Rashomon effect**: too many widely different model structures add up to (nearly) the same black-box behavior. It is a problem when we do care about “opening up” part of the black-box.

The next key idea: how to separate the nuisances “as much as we can” out of the causal effect.

THE NUISANCES

The high-dimensional contribution from Z can be accounted for in the following two regression functions:

$$\begin{aligned}e(z) &= Pr(X = 1 \mid z) \\m(z) &= \mathbb{E}[Y \mid z]\end{aligned}$$

The are linked to the model for Y given X and Z by the fact that if

$$Y = f_0(Z) + X\tau(Z) + error$$

where *error* is independent of X and Z and zero mean, then we can condition on X and Z **in stages**.

GLUING THEM TOGETHER

$$\begin{aligned}\mathbb{E}[Y \mid z] &= \mathbb{E}[f_0(Z) \mid z] + \mathbb{E}[X\tau(Z) \mid z] + \mathbb{E}[error \mid z] \\ m(z) &= f_0(z) + e(z)\tau(z)\end{aligned}$$

So, we have two ways of writing $m(z)$. Why do we care? Because

$$\begin{aligned}Y - m(Z) &= f_0(Z) + X\tau(Z) + error - m(Z) \\ &= (X - e(Z))\tau(Z) + error\end{aligned}$$

Can you guess where we are going?

THE R LEARNER

1. Estimate $\hat{m}(z), \hat{e}(z)$
2. Create pseudo-data $\tilde{Y}^{(i)} = Y^{(i)} - \hat{m}(Z^{(i)}), \tilde{X}^{(i)} = X^{(i)} - \hat{e}(Z^{(i)})$
3. Apply any optimization method you want to solve

$$\hat{\tau} \leftarrow \operatorname{argmin}_{\tau} \sum_{i=1}^d (\tilde{Y}^{(i)} - \tilde{X}^{(i)} \tau(Z^{(i)}))^2$$

using whatever training and representation of τ we want, including a deep neural net.

This can be implemented by wrapping up standard algorithms.

CROSS-FITTING

Is this it? Almost there.

We don't want to reuse the same data that was used to build the pseudo-data in Steps 1-2 in Step 3. But we can just do **cross-fitting**:

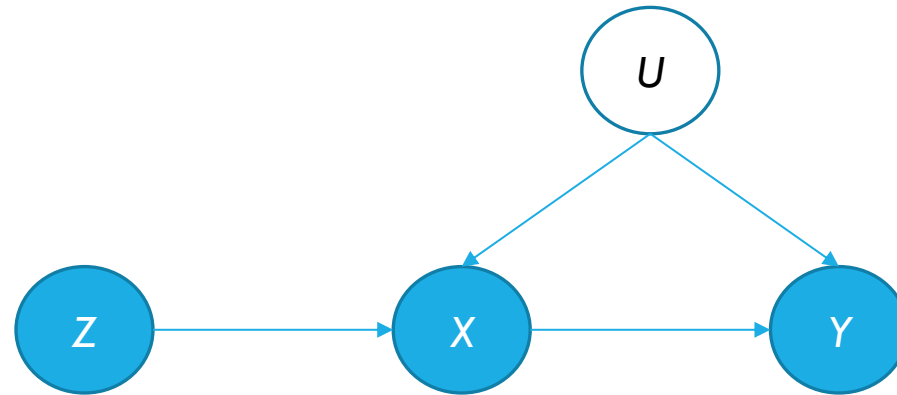
- Split the data in k folds just like we do with cross-validation.
- Use the k th fold for Step 3, while the remaining is used for Steps 1-2.
- Average over the k estimates to produce the final estimate.

The usual Bayesian way of doing it would be dubious from an orthodox perspective, but see the **cut operator** in a software package like BUGs.

WHEN IDENTIFIABILITY FAILS

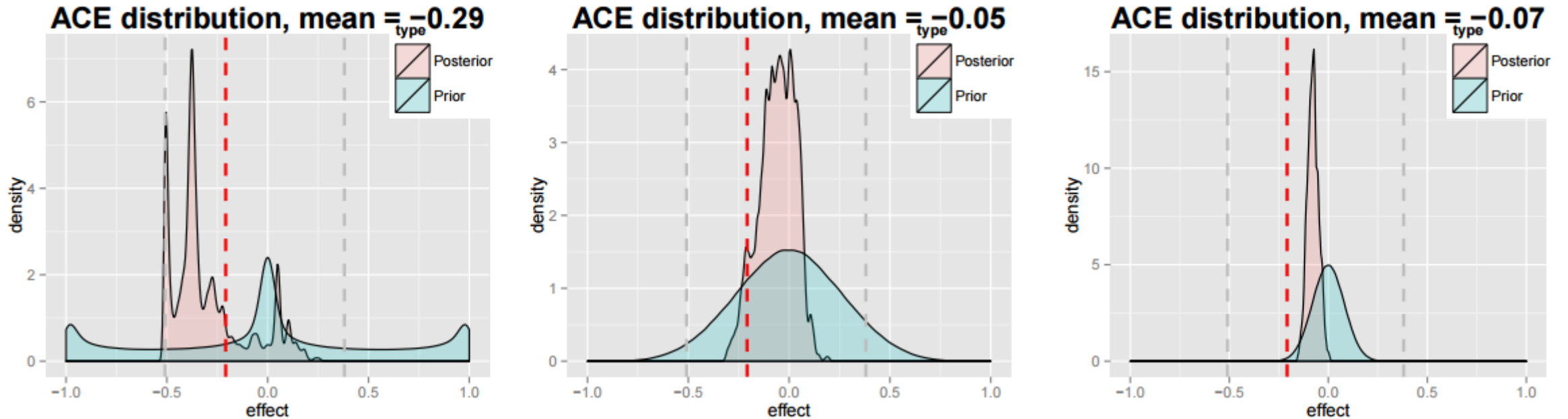
In some cases we can get **bounds** on the causal effect (again, Niki's talk FTW).

The most well-known case is the **instrumental variable structure**.



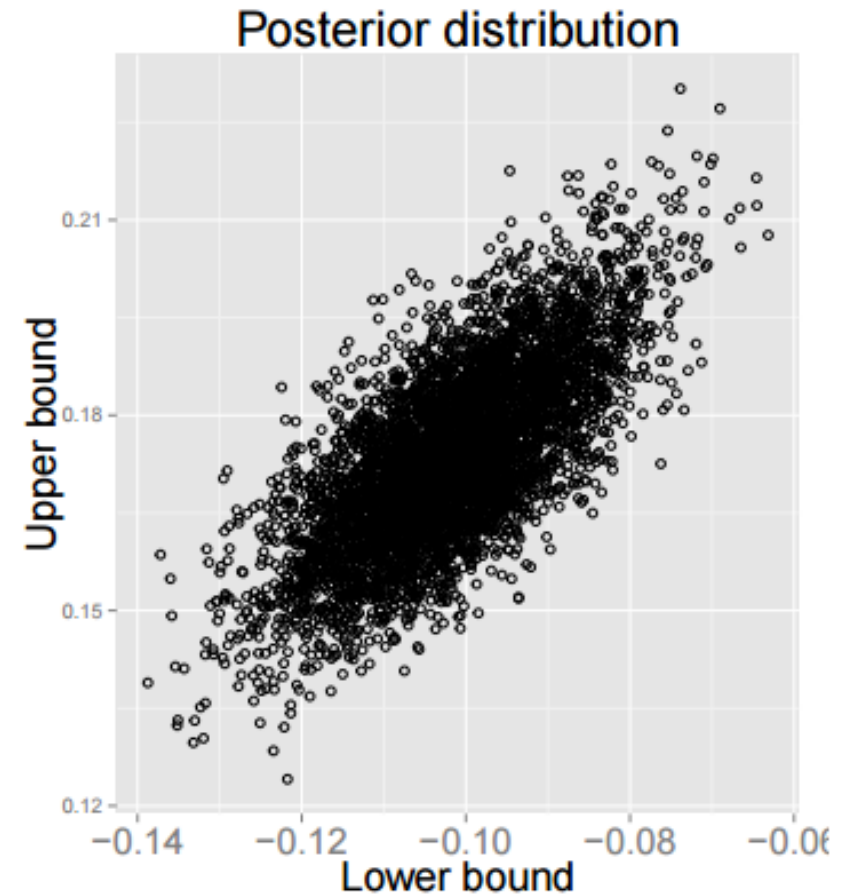
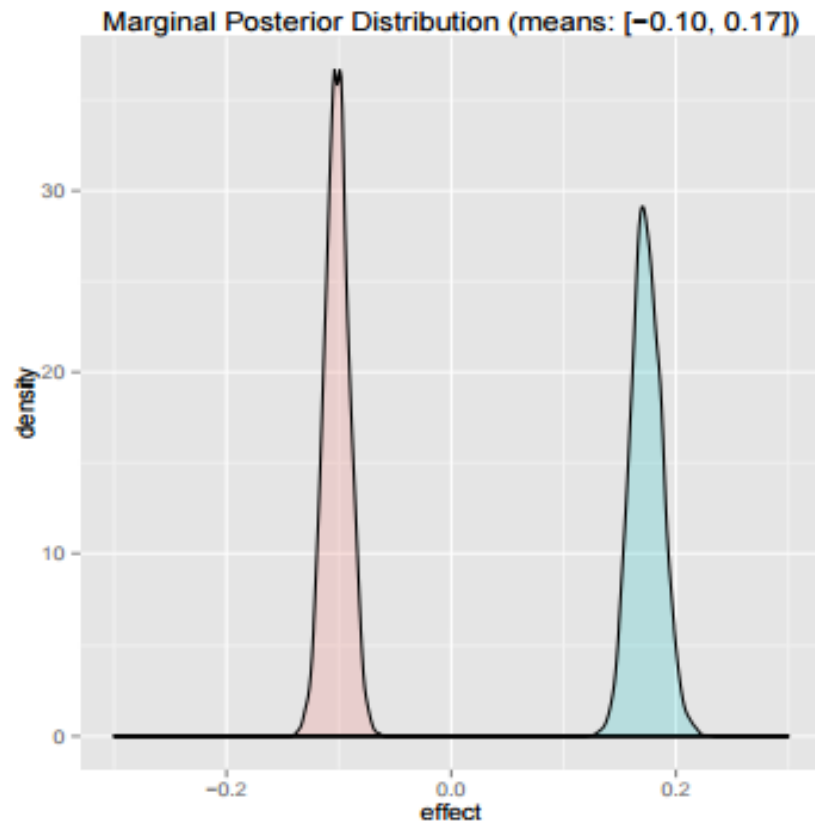
$$f_{lower}(p_{xyz}(\cdot, \cdot, \cdot)) \leq ATE \leq f_{upper}(p_{xyz}(\cdot, \cdot, \cdot))$$

IDENTIFIABILITY BY PRIORS: IF YOU MUST DO IT, TAKE IT SERIOUSLY



Silva and Evans (2006). "Causal inference through a witness protection program." *JMLR*

DO INFERENCE ON THE BOUNDS, INSTEAD



Silva and Evans (2006). "Causal inference through a witness protection program." *JMLR*

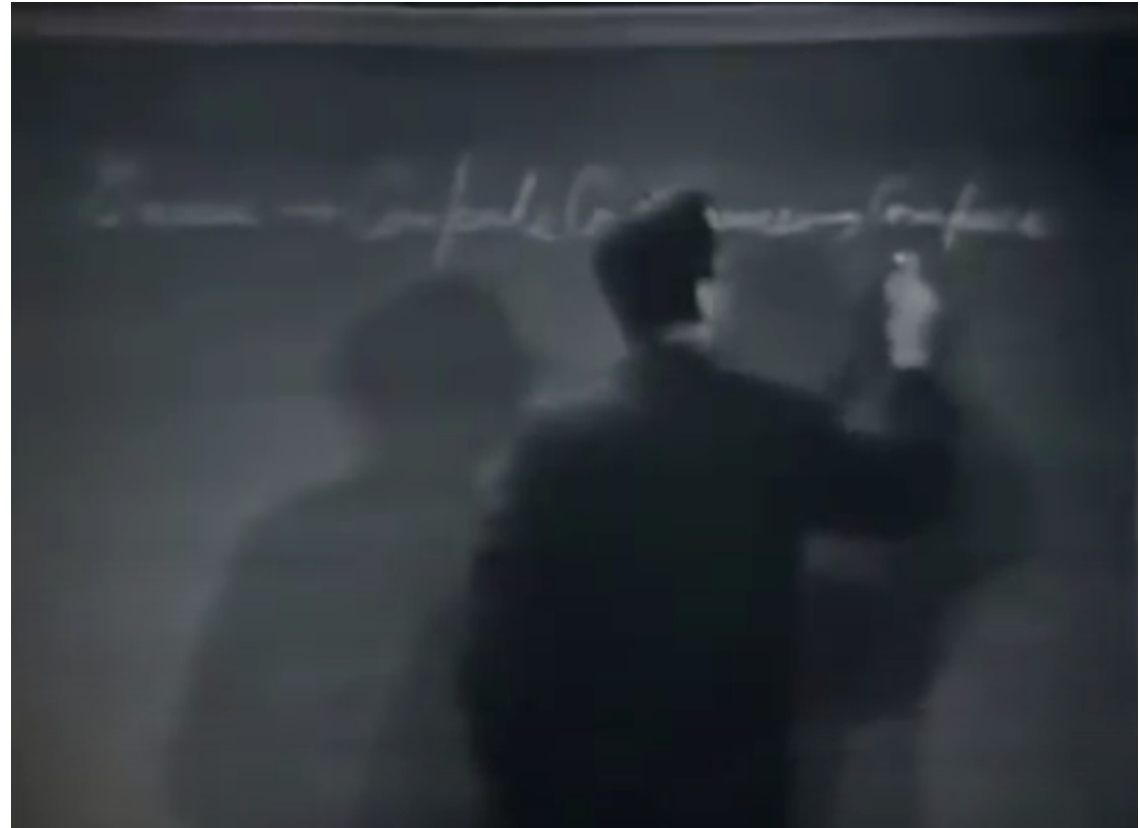
CONCLUSION



CAUSAL INFERENCE AND THE SCIENTIFIC METHOD

“First, we guess it. Then we compute the consequences of the guess... to see what it would imply. Then we compare the computation results to Nature or we say compare to experiment or experience, compare it directly with observations to see if it works.”

Richard Feynman



<https://www.youtube.com/watch?v=OKmimDq4cSU&feature=youtu.be>

IF YOU ARE A METHODOLOGIST, REMEMBER:

The domain expert has the last word. Not you.

As machine learners/AI folks, our primary goal should be to provide **languages** to express assumptions, **algorithms** to compute the consequences of such assumptions, and **inferential procedures** to report the resulting uncertainty and test what can be tested.

If a client/VC/reviewer complains your method gives too much freedom for a domain expert to put (consequential) assumptions in, tell them politely to find a new day job.

IF YOU ARE (ALSO) THE “TRUE” SCIENTIST IN CHARGE, REMEMBER:

Be aware that the result of an observational study is only as good as its assumptions.

“One-size-fits-all” tools, that is, those that work only under a specific type of assumptions, are fine if properly understood. But are you adopting them for convenience or because they are appropriate?

Can you take more than one starting point and see how your conclusions vary? If not, why not?

WHAT I DID NOT COVER

Sequential problems a.k.a. reinforcement learning. Off-policy evaluation in the causal inference literature started back in the mid-80s. See for instance,

- Hernan and Robins (2020) *Causal Inference: What If*. Chapman & Hall. Draft at <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>
- Chakraborty and Moodie (2012). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer.

Many (many) other ways of achieving identifiability and their respective learning methods (more on instrumental variables, proxies, differences-in-differences, synthetic controls, discontinuity designs, matching etc.) and combinations of regimes

Causal discovery, in its many guises and families of assumptions

THANK YOU