Targeted Learning for Data Adaptive Causal Inference in Observational and Randomized Studies

Maya Petersen and Mark van der Laan

Department of Biostatistics, University of California, Berkeley School of Public Health

Part I: From causal questions to the statistical estimation problem

Introduction using single time point interventions

Outline

- A general roadmap for tackling causal questions
- Introduction to structural causal models (SCM)/Causal Graphs
- Defining target causal quantities using counterfactuals
- Identifying causal effects as parameters of the observed data distribution

What's special about causal inference?

- Data + statistical assumptions= statistical inference
  - Conclusions about an underlying population
- Data + statistical assumptions + causal assumptions (non-testable) = causal inference
  - Conclusions about how the underlying population would change if conditions changed
    - Eg- if we changed the way treatment was assigned

A Roadmap for Causal Inference

1. Specify a Question, Causal Model, and its link to the Observed Data
2. Specify the Causal Quantity of Interest
3. Assess Identifiability
4. Commit to a Statistical Model and Target Parameter of the Observed Data Distribution
5. Estimate the Chosen Parameter of the Observed Data Distribution
6. Interpret Results

Defining the Statistical Estimation Problem

1. Specify a Question, Causal Model, and its link to the Observed Data
2. Specify the Causal Quantity of Interest
3. Assess Identifiability
4. Commit to a Statistical Model and Target Parameter of the Observed Data Distribution
5. Estimate the Chosen Parameter of the Observed Data Distribution
6. Interpret Results

Example: Abacavir and Cardiovascular Disease

- Analysis of observational data from several cohorts suggested abacavir use associated with increased risk of myocardial infarction among treated HIV-infected population
  - Other analyses found no evidence of such an association….
- Example of a causal question: Does use of abacavir (ABC) increase risk of myocardial infarction (MI)?

Specifying a Causal Model

- Causal Model is a way to represent background knowledge about the system you want to study
- Example:
  - What factors affect physicians' decisions to prescribe abacavir?
  - What are major determinants of myocardial infarction in this population?
- Structural Causal Models (SCM) are a formal way to represent this knowledge
  - Unify structural equation, causal graph, and counterfactual frameworks

Structural Causal Models: Motivation

- Provide a framework in which we can

1. Rigorously express causal assumptions
   - These are different from statistical assumptions
2. Define causal questions
3. Evaluate whether the data and assumptions together are sufficient to answer those questions

- Once we have succeeded in defining our question as a parameter of the observed data distribution (steps 1-4), we are back in the world of standard statistics (step 5)
  - Step 5 (estimation) is still a very hard problem

Definition: Structural Causal Model

1. Endogenous variables $X = \{X_1, ..., X_J\}$

– Variables that are meaningful for the scientific question, or about which you have some scientific knowledge

- E.g. We often (but not always) know the time ordering of these variables
- Includes all the variables you measure (or are considering measuring)
- Might also include some variables you do not/cannot observe

– Affected by other variables in the model

Definition: Structural Causal Model

2. Exogenous variables (Errors)

$$U = \{U_1, ..., U_J\}$$

– Not affected by other factors in the model

– All the unmeasured factors not included in X that go into determining the values that the X variables take

  - U collapses all these unknown factors into one variable

– We denote the distribution of these factors P_U

Definition: Structural Causal Model

3. Functions $F = \{f_{X_1}, \ldots, f_{X_J}\}$

– The functions *F* define a set of structural equations for each of the endogenous variables

– For each endogenous variable in X_j, we specify its parents $Pa(X_j)$: Endogenous variables that may affect the value of X_j

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, \ldots, J$$

$$Pa(X_j) \subseteq X \setminus X_j$$

– One option: include in $Pa(X_j)$ all variables that temporally/causally precede X_j

Structural Causal Model

- Given an input *u*, the functions *F* deterministically assign a value to each of the endogenous variables

- Our model says that the distribution of *(U,X)* is generated by

1. Drawing a multivariate U from a specific probability distribution P_U
2. Deterministically assigning *X* by plugging *U* into the set of functions *F*

– A given input *u* gives us a specific realization *x*

SCM Encode Causal Assumptions

- Assumptions about how the variables X were generated in the system we want to study
- What factors does "Nature" (or the "experiment" that generated the data in the system we want to study) consult when assigning a value to these variables?
  - What do we know about factors that determine whether an individual gets an MI?
  - What do we know about factors that affect whether a patient is prescribed abacavir?

Example: Abacavir and Cardiovascular Disease

- **Question**: Does use of abacavir (ABC) increase risk of myocardial infarction (MI)?
- To introduce concepts and notation, assume a simplified single time point data structure:
  - **A:** treatment with ABC at the start of follow up
  - **W**: patient covariates measured prior to decision whether to treat with ABC
    - Cardiovascular risk factors, renal disease, intravenous drug use….
  - **Y:** an indicator that a patient experiences an MI by the end of the study

Example: SCM for Point Treatment

- X={W, A, Y}
  - W=CHD Risk Factors,…
  - A=ABC use
  - Y= MI
- Errors: $U=(U_W,U_A,U_Y) \sim P_U$

- Structural equations:

$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$

- Distribution of (U,X) generated by:

1. Draw *U* from P_U
2. Generate *W* as a deterministic function of U_W
3. Generate *A* as a deterministic function of *W* and U_A
4. Generate *Y* as a deterministic function of *W*, *A*, U_Y

Non-Parametric Structural Equation Models

- The structural equations do not restrict the functional form of the causal relationships
  - Ex: $A=f_A(W,U_A)$ *vs.* $A=\beta_0+\beta_1 LDL+\beta_2 HTN+\ldots+U_A$
  - If you have real knowledge about the functional form of a structural equation, you can incorporate it
- Similarly, we do not impose unsupported assumptions on the error distribution
- **The use of non-parametric structural equation models allows us to respect the limits of our knowledge**

Assumptions on the SCM (1): Exclusion Restrictions

- We make assumptions by leaving X variables out of a given parent set
  - Excluding a variable from Pa(X_j) assumes it does not directly affect what value X_j takes
  - Leaving a variable in Pa(X_j) means it might (or might not) affect what value X_j takes

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), j = 1, ..., J$$

$$Pa(X_j) \subseteq X \setminus X_j$$

  - One option: include in Pa(X_j) all variables that temporally precede X_j

Assumptions on the SCM (2): Independence Assumptions

- Independence assumptions restrict the allowed distributions for P_U
- Ex. Assume U_A is independent of U_Y
  - Corresponds to saying that A and Y share no common causes outside other than those included in X
  - When might this be reasonable?

More on assumptions to come…

- Assumptions (at least on P_U) will be necessary if we want to make causal inferences with observational data
  - We will come back to this when we talk about identifiability
- Our goals

1. Whenever possible, restrict our assumptions to those supported by our knowledge
2. When we have to make more questionable ("convenience") assumptions
   - Make them explicitly so that we can evaluate them better and interpret results appropriately
   - Limit them to (causal) assumptions that do not change the statistical model

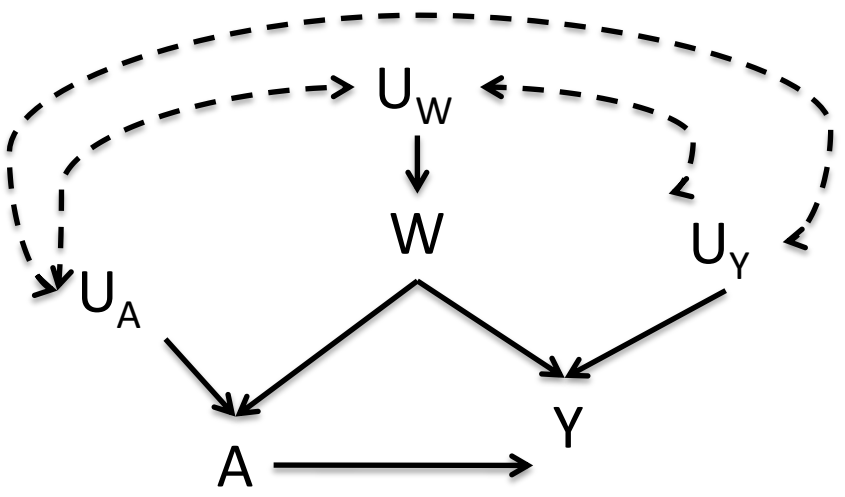
Structural Causal Model

- Defines set of allowed distributions for (U,X)
- Specifically, this is the set of possible distributions $P_{U,X}$ defined by
  - All the joint distributions P_U compatible with any independence assumptions
  - All the specifications of the functions $F=(f_{Xj}: j)$ compatible with any exclusion restrictions
- We will call this model $\mathcal{M}^{\mathcal{F}}$
  - Each distribution included in the model is indexed by a specific distribution P_U and specific functions F

Structural Model Defines a Graph

- Connect parents to children with an arrow
  - Makes the asymmetry of the equations explicit
- Each endogenous X variable has an error (U)
- Correlations between errors encoded in dashed lines/double headed errors.

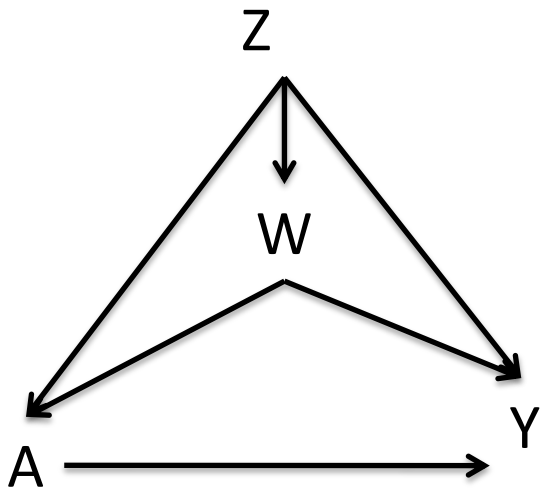
$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y)$$



Alternative Representation

- Include as a node any unmeasured common cause of at least 2 of the X variables
  - Doesn't have to represent a specific variable that you understand well
  - Just an alternative way to express there may be such a variable (or variables)
- The remaining errors will be independent
  - Customarily omitted from the graph

$$Z = f_Z(U_Z)$$
$$W = f_W(Z, U_W)$$
$$A = f_A(Z, W, U_A)$$
$$Y = f_Y(Z, A, W, U_Y)$$



Defining a Target Causal Parameter

- Recall our motivation:

  experimental conditions under which we observe a system ≠ experimental conditions we are most interested in
- The process of translating our background knowledge into a SCM required us to be specific about our knowledge of existing experimental conditions
- The process of translating our scientific question into a target causal parameter requires us to be specific about our ideal experimental conditions

Defining a Target Causal Parameter

- Step 1. Decide which variable or variables we want to intervene on
  - "Exposure" or "Treatment"
  - We are interested in a system that modifies the way these variables are generated
  - For now focus on one variable at a single time point
    - Lots of times you are interested in intervening on more than one variable/time point
    - We will get to that
  - We refer to this variable as the intervention variable, and typically use "A" to represent it

Defining a Target Causal Parameter

- Step 2. Decide what kind of intervention we are interested in
  - For now, we will focus on "static" interventions
    - Interventions that deterministically set A equal to some fixed value(s) of interest
  - There are other options
    - E.g. dynamic interventions: Set A in response to the values of other variables
- Step 3. Specify an outcome (or outcomes)
  - Again, we'll focus on a single outcome at a single time point for now

Example: Abacavir and Cardiovascular Disease

- **Question**: Does use of abacavir (ABC) increase risk of myocardial infarction (MI)?

1. What is the intervention variable?
2. What is the intervention?
3. What is the outcome?

Counterfactuals

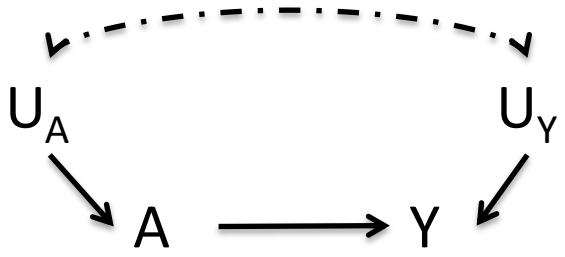
- Y_a for an individual is the value that variable Y would have taken for that individual if that individual had received treatment A=a
  - "Counterfactual" because the individual may not have actually received treatment A=a
  - Also referred to as "Potential Outcomes"

Counterfactuals can be derived from the SCM

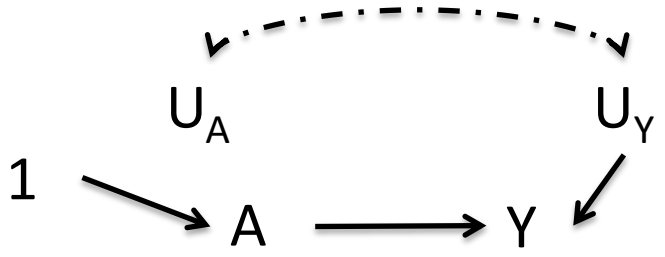
- Structural equations are autonomous
  - Changing one function does not change the other functions
- Can intervene on part of the system and see how changes are transmitted through the rest of the system
  - To make inferences about data generated by the same system under different conditions, we have to know which parts of the system will change and which parts will stay the same

Interventions on the SCM

- The autonomy of structural equations means that we can make a targeted modification to the set of equations in order to represent our intervention of interest
- Ex. Intervene on the system to set A=1
  - Replace f_A with constant function *A=1*



$$A = f_A(U_A)$$
$$Y = f_Y(A, U_Y)$$



$$A = 1$$
$$Y = f_Y(1, U_Y)$$

Counterfactuals derived from SCM

- $Y_a(u)$ is defined as the solution to the equation f_Y under an intervention on the system of equations to set $A=a$ (with input $U=u$)
  - We can think of u as the background factors of each subject
  - $Y_a(u)$ is a realization
    - It is implied by F and u
- P_U and F induce a probability distribution on Y_a just as they do on Y
  - $Y_a=Y_a(U)$ is the post-intervention (or counterfactual) random variable

Ex: Counterfactuals derived from SCM

- Endogenous variables:

  X={W, A, Y}
  - W=CHD Risk Factors,…
  - A=ABC use
  - Y= MI
- Errors: $U=(U_W,U_A,U_Y)\sim P_U$

- Post-intervention Structural equations

$$W = f_W(U_W)$$
$$A = a$$
$$Y_a = f_Y(W, a, U_Y)$$

- Interventions of interest: Set A=1 and A=0
- Counterfactuals of Interest:

$$Y_a = f_Y(W, a, U_Y),\ a \in \mathcal{A} = \{0, 1\}$$

where $\mathcal{A}$ refers to treatment levels of interest

Defining a target causal parameter

1. Decide which variable we want to "intervene on" and what the interventions of interest are
2. Decide outcome of interest

Steps 1 and 2 define our counterfactual outcomes of interest (and our SCM defines a model for the distribution of these counterfactuals)

3. Specify what parameter of the distribution of these counterfactual outcomes we are interested in… (our target causal quantity)

Example: Average Treatment Effect:

- How would expected outcome have differed if everyone in the population had been treated vs. if no one in the population had been treated?
  - This is a common target of inference.
  - This is what many RCTs are trying to estimate….

$$E_{U,X}Y_1 - E_{U,X}Y_0$$

Distribution of Y_a is given by P_U and F, or alternatively, by $P_{U,X}$

Examples: Other counterfactual parameters

- For binary Y:
  - Causal Relative Risk $E_{U,X}Y_1 / E_{U,X}Y_0$
  - Causal Odds Ratio $\dfrac{E_{U,X}Y_1/(1-E_{U,X}Y_1)}{E_{U,X}Y_0/(1-E_{U,X}Y_0)}$
- May be interested in a causal effect within certain strata of the population…

$$E_{U,X}(Y_1 - Y_0|V), V \subset W$$

Marginal Structural Models

- Specify a (working) model for $E(Y_a)$ or $E(Y_a|V)$
- Useful when interested in
  - Dose response curves for multi-level/continuous exposures
  - Effect modification by multi-level covariates
- Ex. A: Abacavir dose

  Y: Renal function

$$E_{U,X}(Y_a) = m(a|\beta)$$

$$m(a|\beta) = \beta_0 + \beta_1 a$$

$$\beta(P_{U,X}|m) \equiv \arg\min_{\beta} E_{U,X}\left[\sum_{a\in\mathcal{A}}(Y_a - m(a|\beta))^2\right]$$

Specify the Observed Data

- Simple Abacavir Example: Observed data for a given subject: O=(W,A,Y)
  - Baseline covariates W= CHD risk factors
  - Exposure A= ABC Use
  - Outcome Y= MI

- Later today, we will address missing data, longitudinal data, right censoring and time to event outcomes…

Linking the Observed Data to the SCM

- Defining the statistical estimation problem requires specifying the link between endogenous variables X and the observed data O
  - In other words, we specify how the observed data were generated by the data generating system encoded in our SCM
- For our simple example, O=X
  - Can specify other links as well

Linking the Observed Data to the SCM

- We observe a sample of size n of the random variable O
  - For now we will work with independent samples
  - The framework is not restricted to this
- We assume our observed data were generated by sampling *n* times from the data generating system specified in our causal model
- This gives us n i.i.d. copies $O_1, O_2, \ldots, O_n$ drawn from true probability distribution P_0

The Statistical Model

- The model $\mathcal{M}^{\mathcal{F}}$ (set of possible distributions for U,X) implies a model (set of possible distributions) for O
- We refer to this set of possible distributions as the statistical model $\mathcal{M}$
- The true distribution P_0 of O is an element of $\mathcal{M}$

The Statistical Model

- Often, a model that respects the limits of our knowledge puts no restrictions on the set of allowed distributions for O
- **In this case our statistical model is non-parametric**
- **We need to respect this fact when we frame the statistical estimation problem**

Overview of Identifiability

- What we want (target of inference): $\Psi^F(P_{U,X})$
  - Ex. $$\Psi^F(P_{U,X}) = E_{U,X}(Y_1 - Y_0)$$
- What we have: a sample from the observed data distribution
  - Ex. n i.i.d. observations of $O \sim P_0$
  - Can use this to make inferences about parameters of the observed data distribution: $\Psi(P_0)$

Overview: Identifiability

- Identifiability in a nutshell:

  Can we write $\Psi^F(P_{U,X})$ as $\Psi(P_0)$?

- Slightly more formally, we need that:

  For each $P_{U,X}$ in $\mathcal{M}^{\mathcal{F}}$ (each $P_{U,X}$ compatible with the SCM) we have that $\Psi^F(P_{U,X}) = \Psi(P_0)$ for some Ψ

Identifiability for Point Treatment

- Focus here on identifiability for the effect of a single intervention (point treatment) when baseline covariates have been measured

- We will focus on one identifiability result:
  - "G-computation formula"
- Holds under
  - Randomization assumption
  - Backdoor criterion

Example: Identifiability Problem

- SCM $\mathcal{M}^{\mathcal{F}}$:
  - X=(W,A,Y); U=(U_W,U_A,U_Y)~P_U
  - F: No exclusion restrictions or independence assumptions
- Observe: O=(W,A,Y)~P_0
- Statistical model $\mathcal{M}$ is non-parametric
- Target: $\Psi^F(P_{U,X})=E_{U,X}(Y_1-Y_0)$
- Can we write $\Psi^F(P_{U,X,0})$ as a parameter of P_0?

Identifiability of Point Treatment Effects under the Randomization Assumption

- Randomization Assumption (RA):

$$Y_a \perp A|W$$

- Identifiability Result

$$P_0(Y = y|A = a, W = w) = P_{U,X}(Y_a = y|A = a, W = w)$$

By definition of counterfactuals

$$= P_{U,X}(Y_a = y|W = w)$$

Under the randomization assumption

Identifiability of Point Treatment Effects under the Randomization Assumption

- If the Randomization Assumption $Y_a \perp A|W$ holds then:

$$E_{U,X}(Y_a|W=w) = E_0(Y|A=a,W=w)$$

- This gives us the G-computation formula

$$E_{U,X}(Y_a) = \sum_w E_0(Y|A=a,W=w)P_0(W=w)$$

$\Psi^F(P_{U,X})$

$\Psi(P_0)$: "estimand"

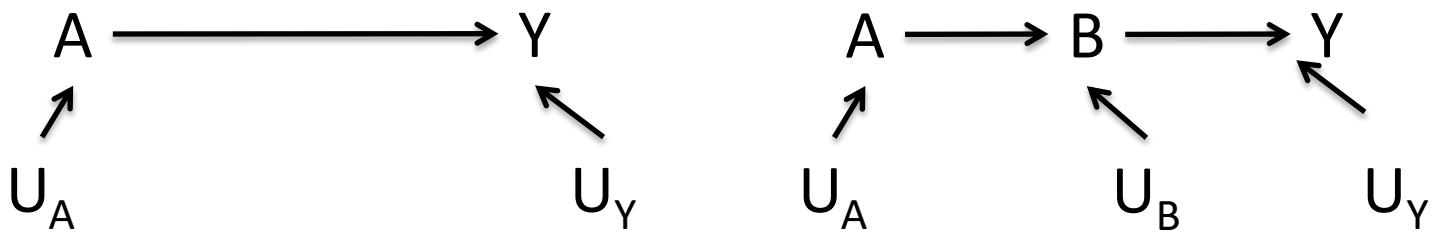
A graphical approach to identifiability: The Back-door Criterion

- Conditional on W, we want to be sure that any observed association between A and Y is due to the effect of A on Y we are interested in
- This means we need to

1. Block all spurious sources of association
2. Not create any new spurious sources of association
3. Leave the causal paths we are interested in unperturbed

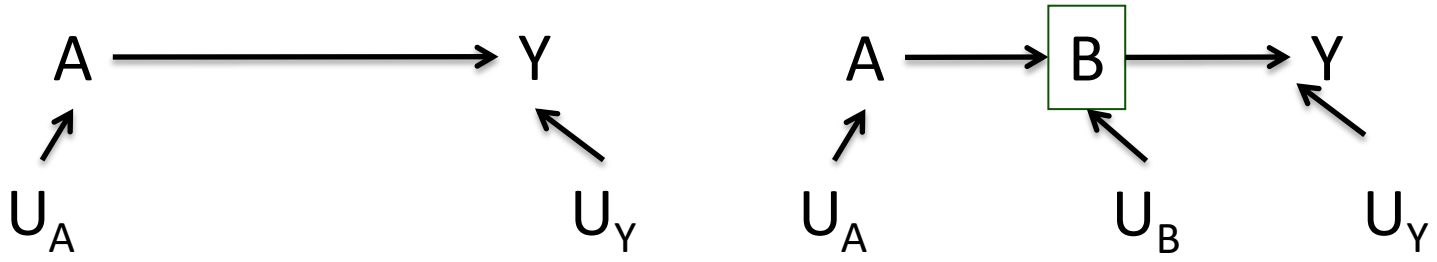
What causal structures can lead to dependence between two observed variables?

1. Direct and Indirect Effects
   - An effect of A on Y can result in an association



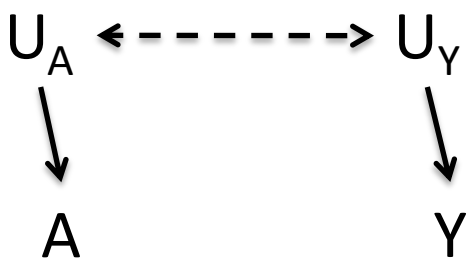
What causal structures can lead to dependence between two observed variables?

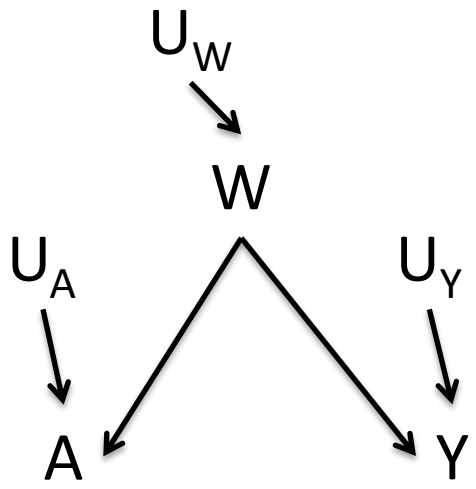
1. Direct and Indirect Effects
   - An effect of A on Y can result in an association
   - Conditioning on an intermediate "blocks" this source of dependence



What causal structures can lead to dependence between two observed variables?

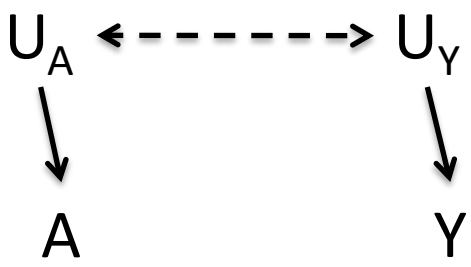
2. Shared common cause
   - A common cause (measured or unmeasured) of A and Y can result in an association
   - When the common cause is not included in X, it is represented through the correlation it induces between errors U

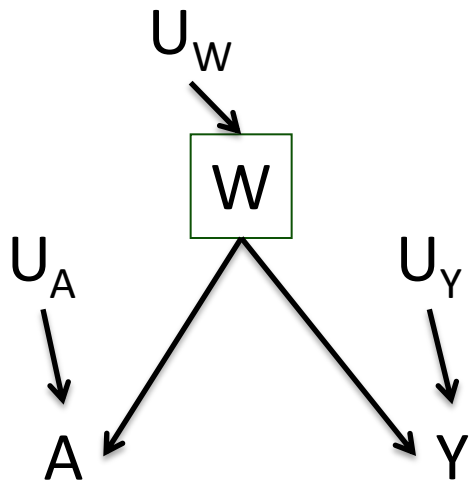




What causal structures can lead to dependence between two observed variables?

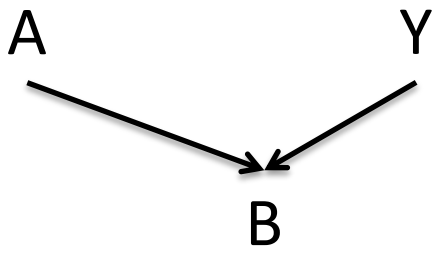
2. Shared common cause
   - A common cause (measured or unmeasured) of A and Y can result in an association
   - Conditioning on a common cause "blocks" this source of dependence

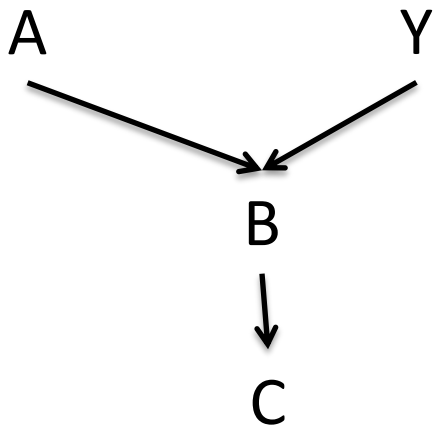




What causal structures can lead to dependence between two observed variables?

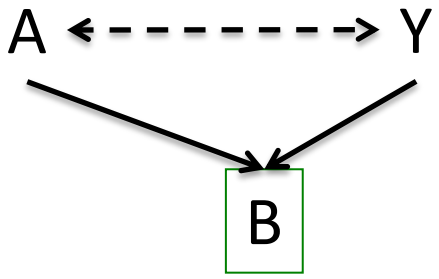
3. Conditioning on a Collider
   - Collider= “inverted fork” a->b<-c
   - A and Y are independent

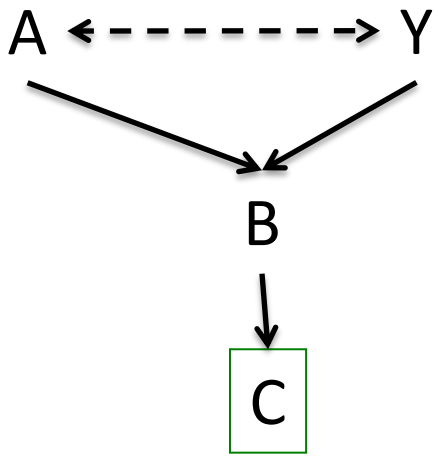




What causal structures can lead to dependence between two observed variables?

3. Conditioning on a Collider
   - Collider= “inverted fork” a->b<-c
   - Conditioning on a common effect (descendent) of A and Y can result in association between A and Y
     - Berkson’s bias/ selection bias





The Back-door Criterion

- Conditional on W, we want to be sure that any observed association between A and Y is due to the effect of A on Y we are interested in
- This tells us what characteristics W should have

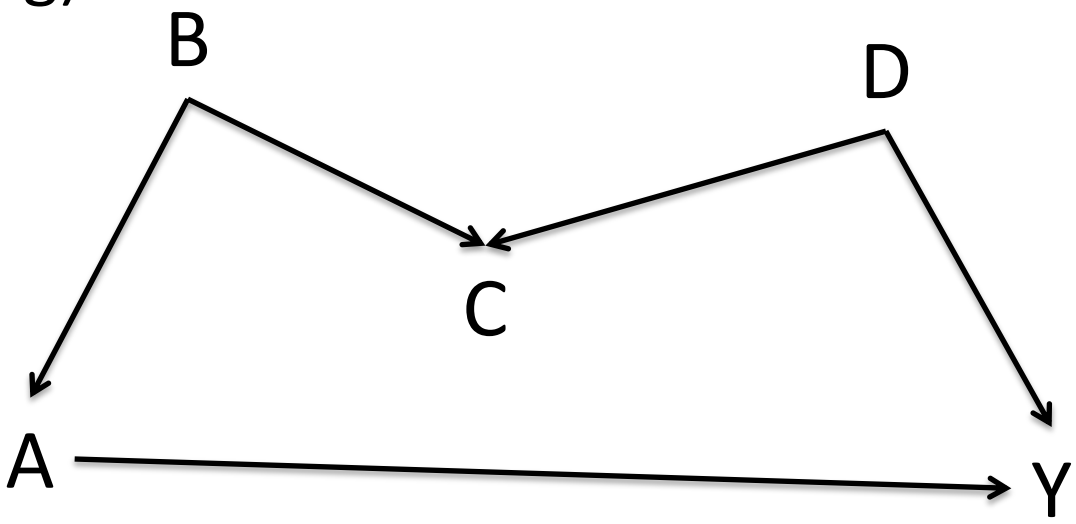
1. W should block any association between A and Y that arises from unmeasured common causes
2. W should not create any new non-causal associations between A and Y
3. W should not block any of the effect of A on Y

Back-door criterion

- A set of variables W satisfies the back door criterion with respect to (A,Y) if
  1. **No node in W is a descendent of A**
  - Motivation:
    1. Avoid blocking the path of interest
    2. Avoid introducing spurious sources of dependence
  2. **W blocks all "backdoor" paths from A to Y**
  - Backdoor path= path with arrow into A
  - Motivation: Block all sources of spurious association between A and Y (due to common causes)

Example

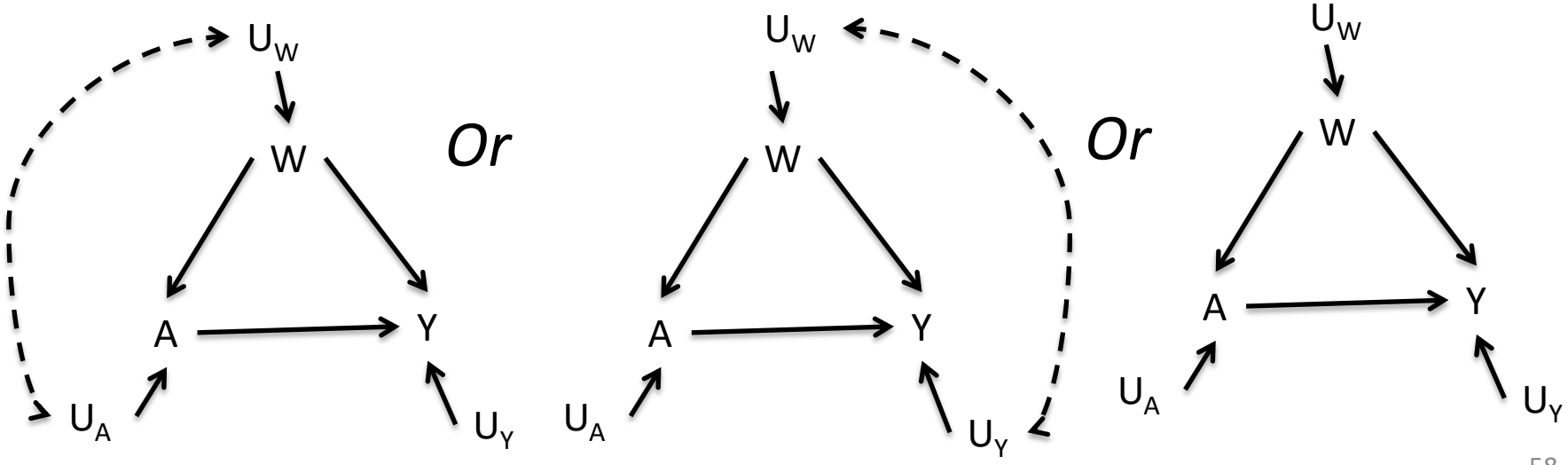
- Back door criterion satisfied for the effect of A on Y by:
  - {} (nothing)?
  - {C}?
  - {B,C}?



Summary: Identifiability for Point Treatment Effects

- Under what sets of independence assumptions will the G-computation identifiability result hold?

$$E_{F_X}(Y_a) = \sum_w E_0(Y|A=a, W=w)P_0(W=w)$$



Positivity Assumption

- Need $E_0(Y|A=a,W=w)$ to be well-defined for all possible values (a,w)
- In non-parametric model, each treatment of interest occur must with some positive probability for each possible covariate history
- Let $g_0(a|W)$ denote $P_0(A=a|W)$
- Positivity assumption:

$$\inf_{a \in \mathcal{A}} g_0(a|W) > 0 \text{ P-a.e.}$$

Our initial model assumptions are not sufficient. Now what?

- $\Psi^F(P_{U,X})$ is not identified under $\mathcal{M}^{\mathcal{F}}$
  - If we are honest with ourselves about the limits of what we know, this happens a lot!
- Options
  - Go get some more data/background research
  - Give up
- But…. Lots of questions require a timely "best guess" to inform ongoing decisions !?!
  - Goal: Get the best answer you can and be honest and transparent when interpreting results

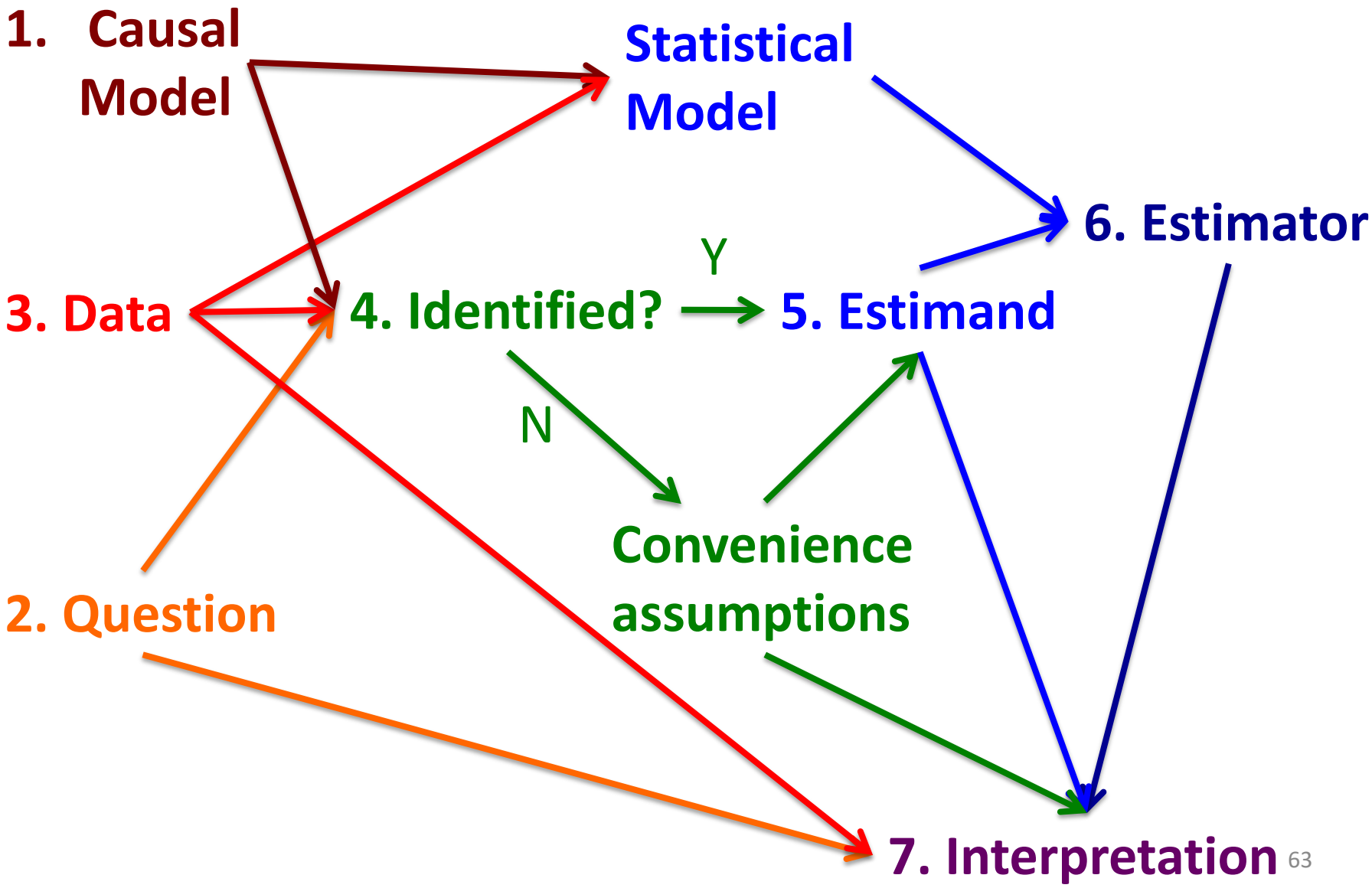
Our initial model assumptions are not sufficient. Now what?

- $\Psi^F(P_{U,X})$ is not identified under $\mathcal{M}^{\mathcal{F}}$
  - We know which additional assumptions would serve to identify $\Psi^F(P_{U,X})$
- We will use $\mathcal{M}^{\mathcal{F}^*}$ to refer to the original SCM + these additional assumptions
- This gives us a way to proceed, while keeping separate our real knowledge and our wished for identifiability assumptions
  - Useful in the interpretation stage!

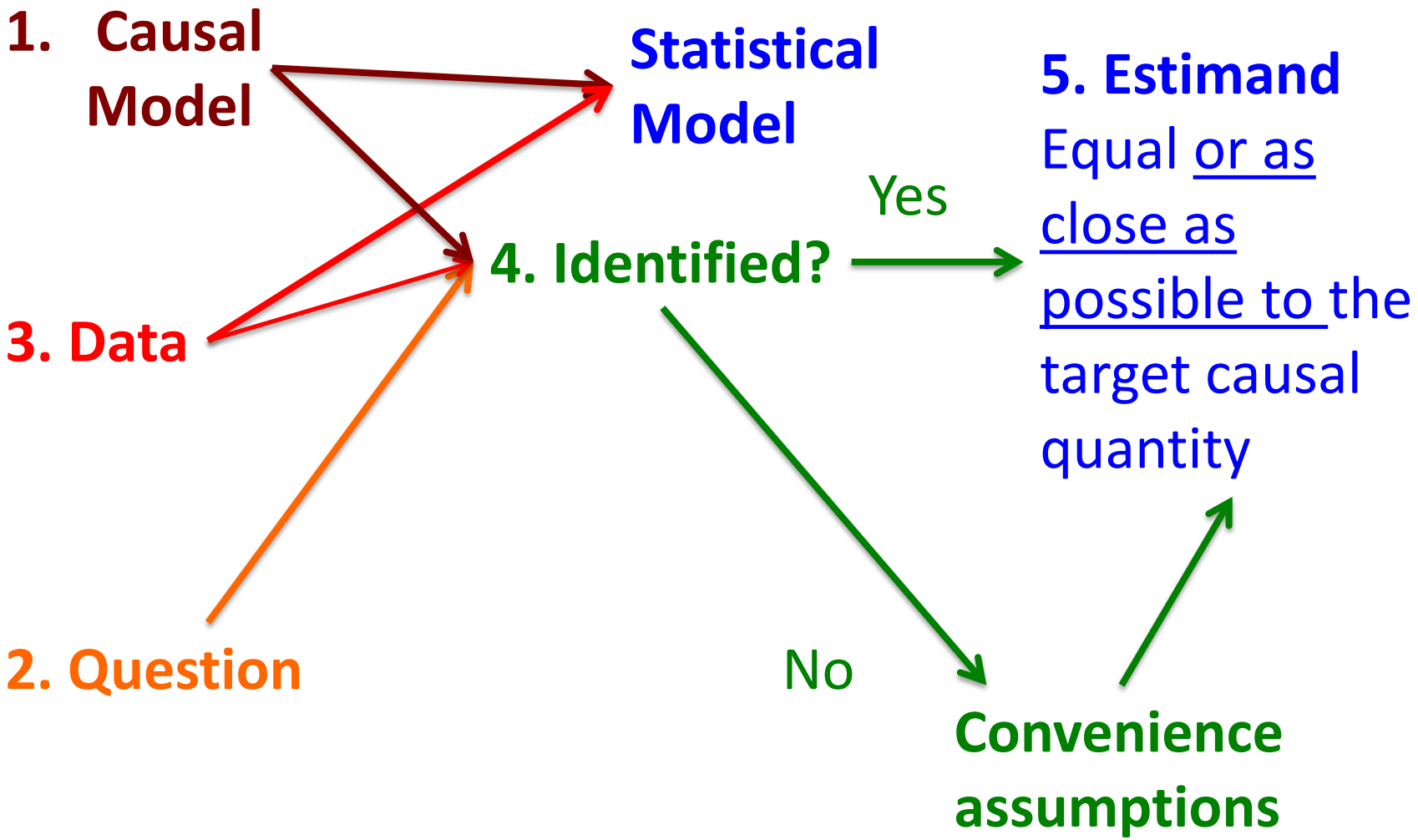
Commit to statistical model and target parameter of the observed data

- The Causal model $\mathcal{M}^{\mathcal{F}}$ implies a statistical model $\mathcal{M}$ for the distribution of the observed data $O \sim P_0$
  - Preference for statistical model implied by $\mathcal{M}^{\mathcal{F}}$ vs. $\mathcal{M}^{\mathcal{F}^*}$ (ensures that at least get a statistical estimation problem that respects the limits of our knowledge)
- Our identifiability result provides us with a target parameter of the observed data distribution (or estimand) $\Psi(P_0)$
- The statistical estimation problem is now defined

A Roadmap….


1. Causal Model
Statistical Model
6. Estimator
Y
3. Data
4. Identified?
5. Estimand
N
Convenience assumptions
2. Question
7. Interpretation

A Roadmap….


1. Causal Model
Statistical Model
5. Estimand
Equal or as close as possible to the target causal quantity
Yes
4. Identified?
3. Data
2. Question
No
Convenience assumptions

So when is a path blocked?

- Path= set of connected edges (any directionality)
- A path is blocked if
  - It has a non-collider that has been conditioned on

Or

  - It has a collider *and* neither the collider nor a descendent has been conditioned on

What does our model assume?

- Example 1:

$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, U_Y)
\end{aligned}
$$

W=Flu virus
A= Headache
Y=Cough

- Example 2:

$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$

W= Parental education
A= Random selection to receive school voucher
Y=Test scores

Assume U_A independent of U_Y?

- Example 1:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, U_Y) \end{aligned}$$

W=Flu virus
A= Headache
Y=Cough

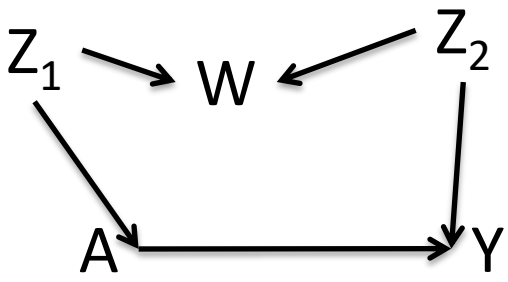
- Example 2:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(U_A) \\ Y &= f_Y(W, A, U_Y) \end{aligned}$$

W= Parental education
A= Random selection to receive school voucher
Y=Test scores

Conditioning on the whole past and only the past is not always a good idea…

- Ex 1. O=(W,A,Y); W occurs before A
  - RA fails conditional on W
  - RA holds conditional on {}



- Ex 2. O=(W,A,L,Y); L occurs after A
  - RA fails conditional on W
  - RA holds conditional on (W,L)