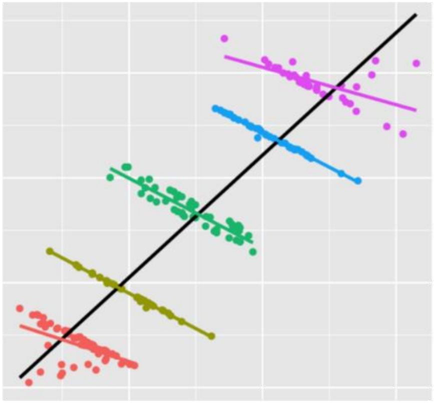
Artificial Intelligence

A Course About Foundations

Causal Inference

Marco Piastra

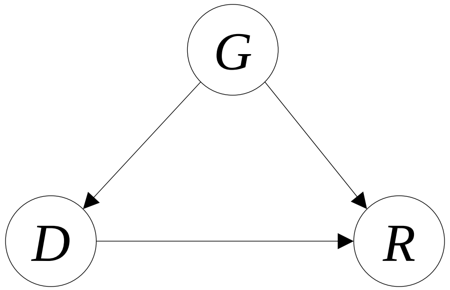
An Example:
Simpson's Paradox



[Image from https://skewthescript.org/data-projects/simpsons-paradox]

Causes and Effects: *say it with graphs*

- What is a cause?



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

Experimental data

- In both groups, recovery rates are *higher* if drug is administered...
- ... while in the entire population, recovery rates are *lower*

Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

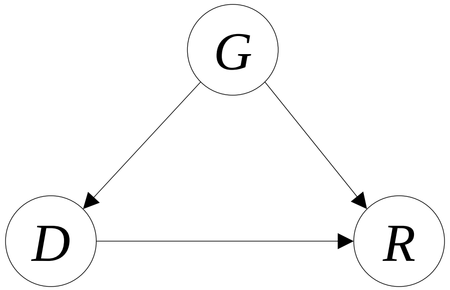
Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Causes and Effects: *say it with graphs*

- What is a cause?



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

Experimental data

- Note however that gender also influenced drug prescription...
- ... in fact, in this example, doctors were more likely to prescribe drug to males than to females

Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

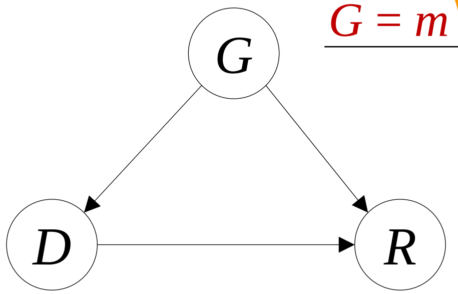
Causes and Effects: *say it with graphs*

- What is a cause?

Maximum Likelihood Estimation (CPTs)

$P(G)$

$G = f$	0.49
$G = m$	0.51



$P(D|G)$

	$G = f$	$G = m$
$D = 0$	0.23	0.76
$D = 1$	0.77	0.24

$P(R|G, D)$

	$G = f$ $D = 0$	$G = f$ $D = 1$	$G = m$ $D = 0$	$G = m$ $D = 1$
$R = 0$	0.31	0.27	0.13	0.07
$R = 1$	0.69	0.73	0.87	0.93

Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

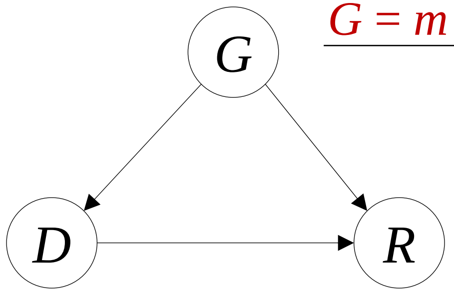
Causes and Effects: *say it with graphs*

- What is a cause?

Maximum Likelihood Estimation (CPTs)

$P(G)$

$G = f$	0.49
$G = m$	0.51



$P(D|G)$

	$G = f$	$G = m$
$D = 0$	0.23	0.76
$D = 1$	0.77	0.24

$P(R|G,D)$

	$G = f$, $D = 0$	$G = f$, $D = 1$	$G = m$, $D = 0$	$G = m$, $D = 1$
$R = 0$	0.31	0.27	0.13	0.07
$R = 1$	0.69	0.73	0.87	0.93

Using Graphical Model as a predictor

Case 1: Gender is observed

$$P(R=1|G=0,D=0) = 0.69$$
$$P(R=1|G=0,D=1) = 0.73$$
$$P(R=1|G=1,D=0) = 0.87$$
$$P(R=1|G=1,D=1) = 0.93$$

Prescribe drug, regardless

Case 2: Gender is <u>not</u> observed

$$P(R|D) = \frac{\sum_G P(R|G,D)P(D|G)P(G)}{\sum_{G,R} P(R|G,D)P(D|G)P(G)}$$

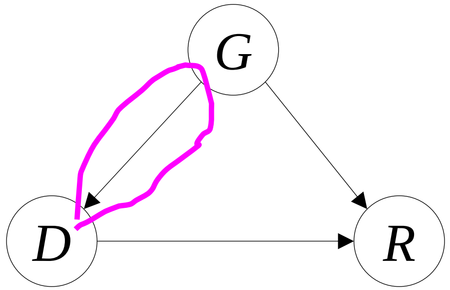
$$P(R=1|D=0) = 0.83$$
$$P(R=1|D=1) = 0.78$$

Do not prescribe drug, regardless

(ridiculous!)

Causes and Effects: *say it with graphs*

- What is a cause?



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

How can we solve the problem?

- The problem is due to the discrepancy in drug administration across genders
- An obvious solution would be *to repeat* the experiment with equal administration rates
- *In other words, we would sever this link*

Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Aside:
Dependence & Correlation

Expected Value of a Random Variable

Basic definition

$$\mathbb{E}_X[X] := \sum_{x \in \mathcal{X}} x \; P(X = x)$$

In a more concise notation:

$$\mathbb{E}[X] := \sum_x x \; P(x) = \mu_X$$

Continuous case

$$\mathbb{E}_X[X] := \int_{x \in \mathcal{X}} x \; p(x) dx$$

Probability density

Expectation is a linear operator

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[cX] = c\mathbb{E}[X]$$

Conditional expectation

$$\mathbb{E}_X[X|Y = y] = \mathbb{E}[X|Y = y] := \sum_{x \in \mathcal{X}} x \; P(X = x|Y = y)$$

Variance and Covariance

Variance

$$\mathrm{Var}(X) := \mathbb{E}_X[(X - \mathbb{E}_X[X])^2] = \mathbb{E}_X[(X - \mu_X)^2]$$

where: $\mu_X := \mathbb{E}_X[X]$

$$\mathrm{Var}(X) := \sum_{x \in \mathcal{X}} P(X = x)\ (x - \mu_X)^2$$

In a more concise notation:

$$\mathrm{Var}(X) := \sum_{x} P(x)\ (x - \mu_X)^2 = \sigma^2$$

Variance is *not* a linear operator

Conditional variance

$$\mathrm{Var}(X|Y = y) := \mathbb{E}_X[(X - \mathbb{E}_X[X|Y = y])^2\ |Y = y]$$

Covariance

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \ = \ \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

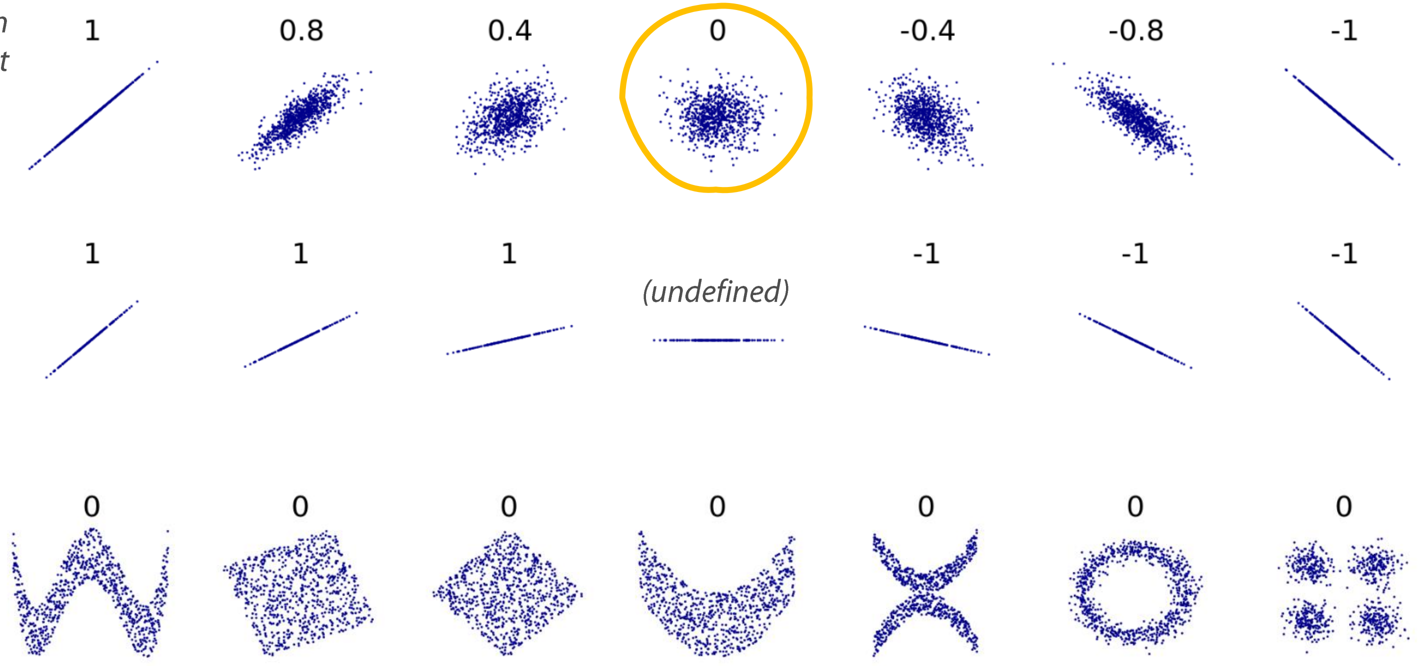
Correlation and Independence

Pearson's correlation:

$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Standard deviation : $\sigma_X = \sqrt{\mathrm{Var}(X)}$

Pearson's Correlation Coefficient



Independence between two random variables

$$< X \perp Y > \Rightarrow \; P(X,Y) = P(X)P(Y)$$

$$< X \perp Y > \Rightarrow \; \rho(X,Y) = 0$$

Zero correlation does NOT imply independence

$$\rho(X,Y) = 0 \;\not\Rightarrow < X \perp Y >$$

[Images from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient]

Correlation and Independence

Zero correlation does NOT imply independence

Does <u>independence</u> imply <u>zero correlation</u>?

$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

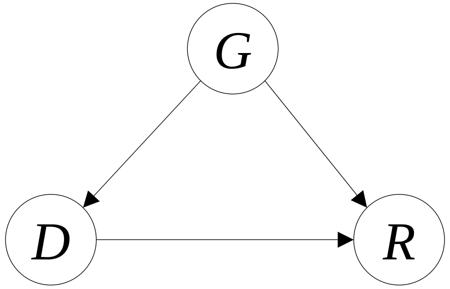
$$\begin{aligned}
\mathrm{Cov}(X,Y) &:= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\
&= \mathbb{E}[XY] - \mu_Y\mathbb{E}[X] - Y\mu_X\mathbb{E}[Y] + \mu_X\mu_Y \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \sum_{x,y} xy\, P(x,y) - \sum_x x\, P(x) \sum_y y\, P(y) \\
&= \sum_{x,y} xy\, P(x,y) - \sum_{x,y} xy\, P(x)P(y)
\end{aligned}$$

Covariance

So, the answer is <u>*yes*</u>: *the last term must be zero if the two variables are independent*

Structural Causal Models

Probabilistic Graphical Model



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

$$P(G, R, D) = P(G)P(D|G)P(R|G, D)$$

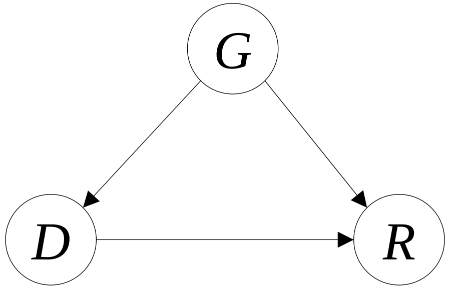
Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

From Graphical Model to Structural Equations



G is *any measure* (= discrete/continuous)
D is *any measure* (= discrete/continuous)
R is *any measure* (= discrete/continuous)

$$P(G, R, D) = P(G)P(D|G)P(R|G, D)$$

** first approximation*

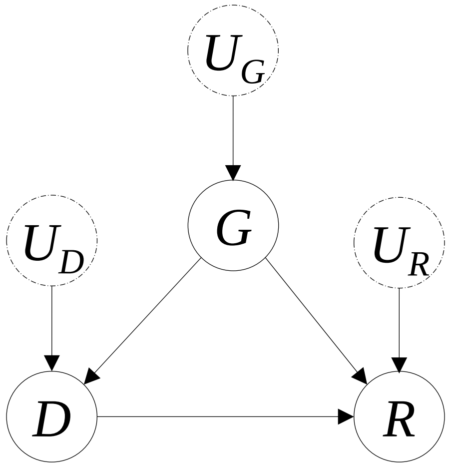
Structural Equations

$$D = f_D(G)$$

$$R = f_R(G, D)$$

How can these two things
be reconciled with probability?
Functions are <u>deterministic</u>

From Graphical Model to Structural Equations



G is *any measure* (= discrete/continuous)
D is *any measure* (= discrete/continuous)
R is *any measure* (= discrete/continuous)

$$P(G, R, D) = P(G)P(D|G)P(R|G, D)$$

U_G, U_D and U_R are <u>unobservable</u>, <u>random</u> variables
The probability distribution is the <u>observable</u> aspect of the structural equations

** second approximation*

Structural Equations

$$G = U_G$$

$$D = f_D(G, U_D)$$

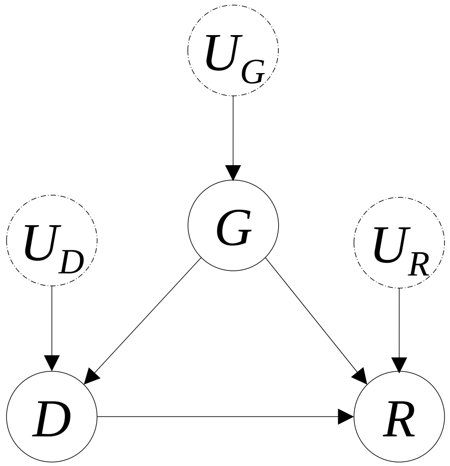
$$R = f_R(G, D, U_R)$$

Causal? Functions could be invertible
Example:

$$D = k + \beta_g G + U_D$$

$$G = \frac{1}{\beta_g}(D - k - U_D)$$

From Graphical Model to Structural Causal Model



G is *any measure* (= discrete/continuous)
D is *any measure* (= discrete/continuous)
R is *any measure* (= discrete/continuous)

$$P(G, R, D) = P(G)P(D|G)P(R|G, D)$$

U_G, U_D and U_R are <u>unobservable</u>, <u>random</u> variables
The probability distribution is the <u>observable</u> aspect of the structural causal model

Structural Causal Model (SCM)

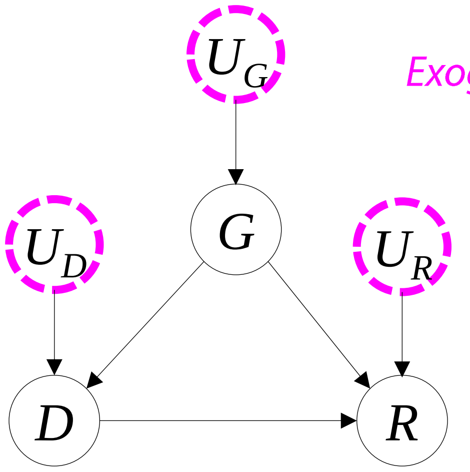
$$G := U_G$$

$$D := f_D(G, U_D)$$

$$R := f_R(G, D, U_R)$$

Force directions, in keeping with causation assumptions

From Graphical Model to Structural Causal Model



Exogenous variables

Structural Causal Model (SCM)

$$G := U_G$$

$$D := f_D(G, U_D)$$

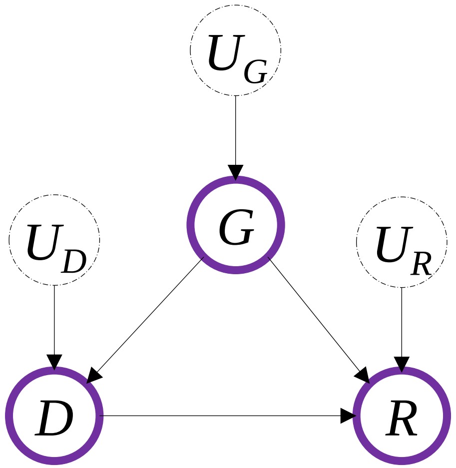
$$R := f_R(G, D, U_R)$$

G is *any measure* (= discrete/continuous)
D is *any measure* (= discrete/continuous)
R is *any measure* (= discrete/continuous)

$$P(G, R, D) = P(G)P(D|G)P(R|G, D)$$

U_G, U_D and U_R are <u>unobservable</u>, <u>random</u> variables
The probability distribution is the <u>observable</u> aspect of the structural causal model

From Graphical Model to Structural Causal Model



Endogenous variables

Structural Causal Model (SCM)

$$G := U_G$$

$$D := f_D(G, U_D)$$

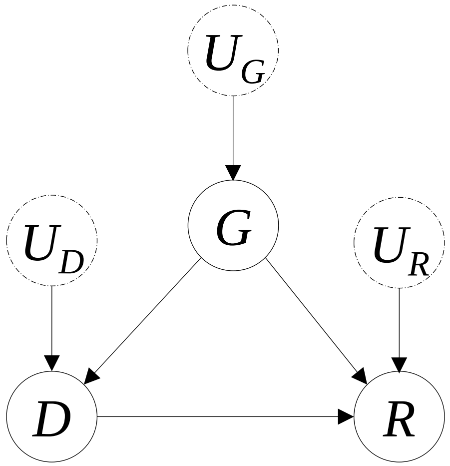
$$R := f_R(G, D, U_R)$$

G is *any measure* (= discrete/continuous)
D is *any measure* (= discrete/continuous)
R is *any measure* (= discrete/continuous)

$$P(G, R, D) = P(G)P(D|G)P(R|G, D)$$

U_G, U_D and U_R are <u>unobservable</u>, <u>random</u> variables
The probability distribution is the <u>observable</u> aspect of the structural causal model

Structural Causal Model



Structural Causal Model (SCM)

$$G := U_G$$

$$D := f_D(G, U_D)$$

$$R := f_R(G, D, U_R)$$

Structural Causal Model (SCM) definition

1) A set of *endogenous* variables
2) A set of *exogenous* variables
3) A set of *structural equations*

An SCM *induces* a graphical model with a probability distribution P over *endogenous* variables

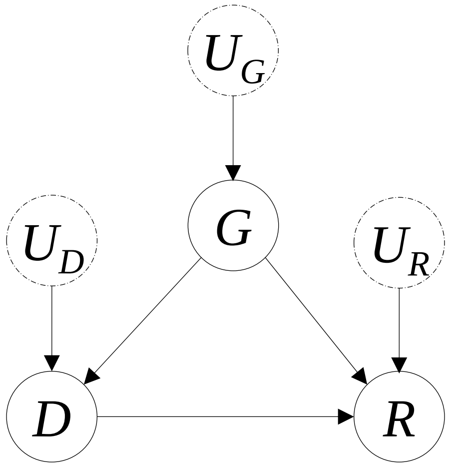
Structural Causal Model (formal definition)

Structural Causal Model (SCM), formally

1) A set of *endogenous* variables $\boldsymbol{X} := \{X_1, X_2, \cdots, X_n\}$
2) A set of *exogenous* variables $\boldsymbol{U} := \{U_1, U_2, \cdots, U_n\}$
3) A set of *structural equations* $\boldsymbol{f} := \{f_1, f_2, \cdots, f_n\}$

An SCM $\mathcal{M}$ *induces* a graphical model $\mathcal{G}$ with a probability distribution $P(\boldsymbol{X})$

Structural Causal Model



induces ←

Structural Causal Model (SCM)

$$G := U_G$$

$$D := f_D(G, U_D)$$

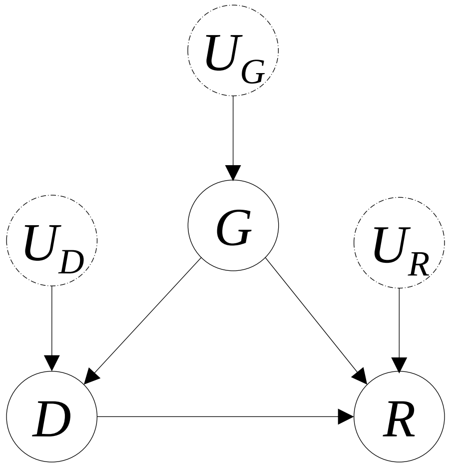
$$R := f_R(G, D, U_R)$$

The graphical model induced is uniquely defined

Further questions:

1) *Which <u>functions</u>?*
2) *How are the random variables U_G, U_D and U_R <u>distributed</u>?*
3) *Are they <u>dependent</u> (or correlated)?*
4) *Is the SCM <u>identifiable</u> from observed data?*

Structural Equation Model (linear functions)



a special case, linear

Structural Equation Model (SEM)

$$G := U_G$$

$$D := k_1 + \beta_1 G + U_D$$

$$R := k_2 + \beta_2 G + \beta_3 D + U_R$$

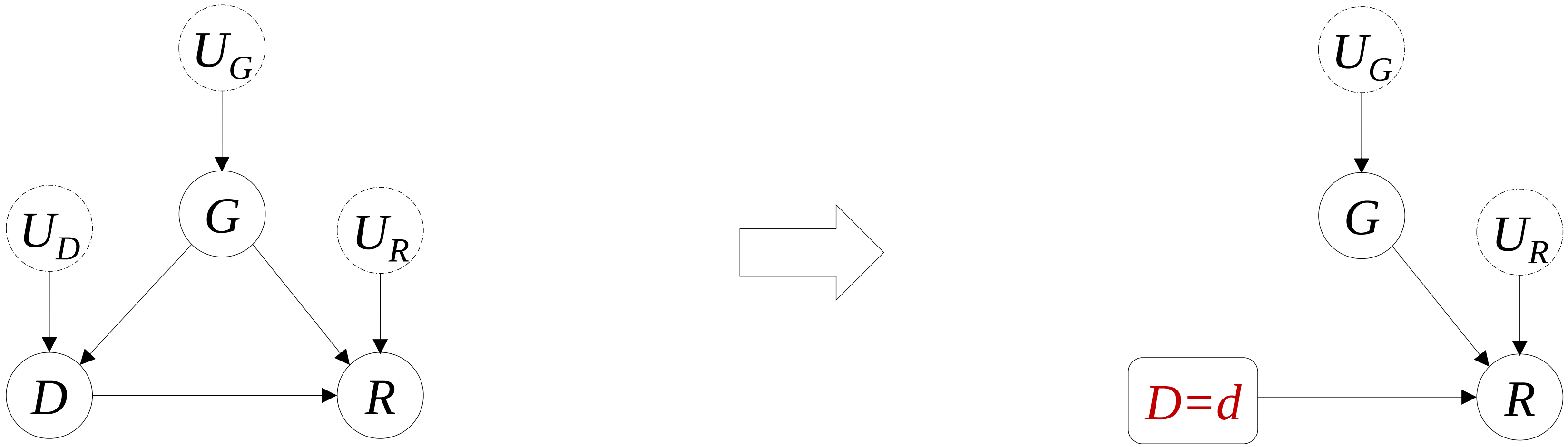
Assumptions:

1) *All functions are <u>linear</u>*
2) *All random variables U_G, U_D and U_R are <u>normally distributed</u>*
3) *All random variables are <u>uncorrelated</u>*

Under further, specific conditions a SEM <u>is</u> identifiable from observed data

more in general, however, this is not true of any SCM

Intervention in a Structural Causal Model



Structural Equation Model (SEM)

$$
\begin{aligned}
G &:= U_G \\
D &:= f_D(G, U_D) \\
R &:= f_R(G, D, U_R)
\end{aligned}
\quad\Longrightarrow\quad
\begin{aligned}
G &:= U_G \\
&\boxed{D := d} \\
R &:= f_R(G, D, U_R)
\end{aligned}
$$

An intervention on an SCM creates a new sub-model by changing one or more structural equations
It induces a new graph

Causal Model

A ***causal model***
is a conceptual tool that we can align with actual ***observations***,
it allows us to perform virtual ***interventions***, to estimate their effects,
and to evaluate possible ***counterfactual*** worlds
("*What if one or more aspects were different from what observed?*")

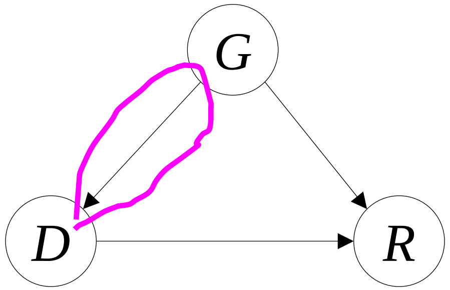
All of this in a precise and formal framework,
in which each inference step *can be performed,*
under specific prerequisites

Using probability theory *as the basic formalism*

Inference through Intervention

The Origin of the Problem

- Probabilistic Graphical Model



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

How can we solve the problem?

- The problem is due to the discrepancy in drug administration across genders
- An obvious solution would be *to repeat* the experiment with equal administration rates
- *In other words, we would sever this link*

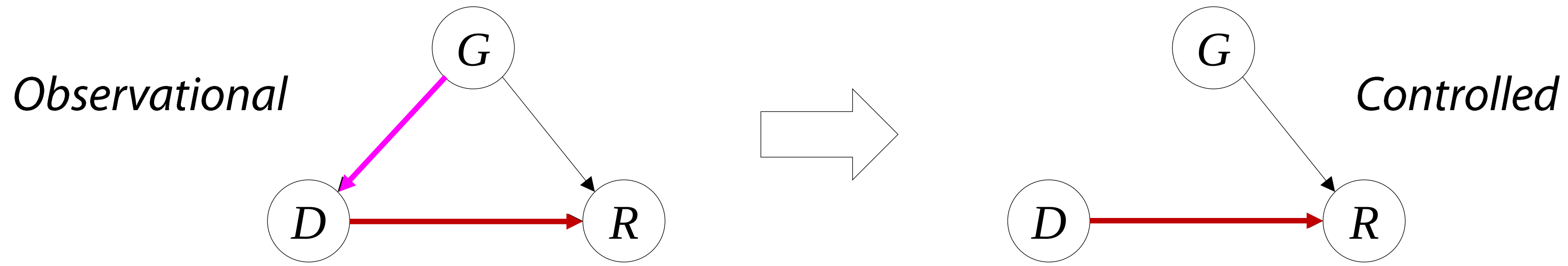
Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

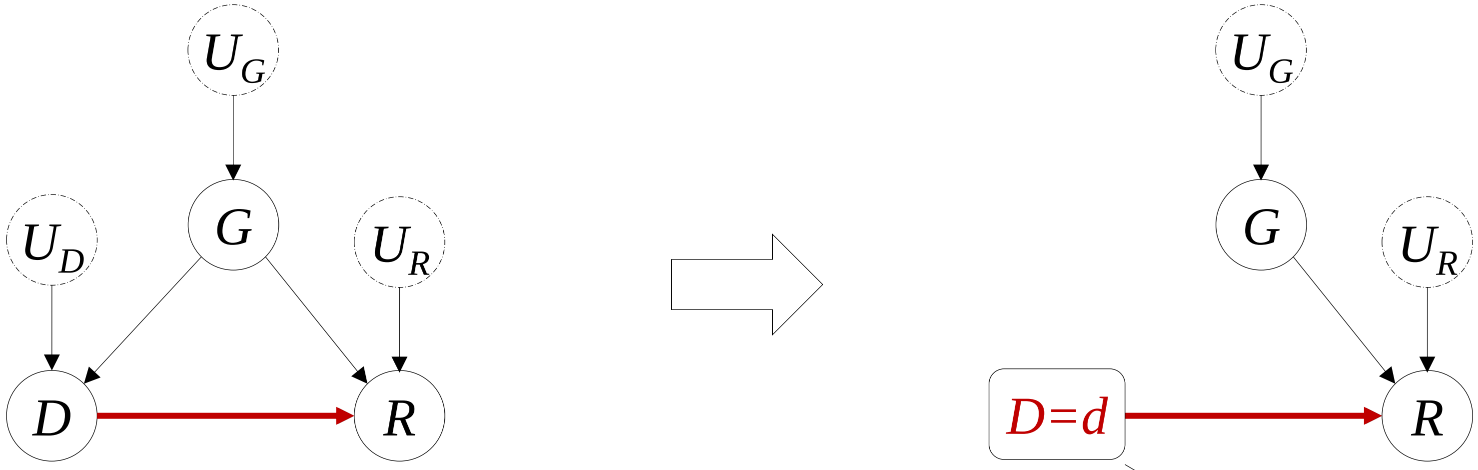
The Magic of Controlled Experiments



In this *Graphical Model*:

1. The causal effect we are interested in is that of D over R
2. The link between G and D is *problematic*: we know that $P(D|G=0) \neq P(D|G=1)$
3. In a ***controlled experiment***, D is administered *at random* , therefore
$$< D \perp G > \implies P(D|G=0) = P(D|G=1) = P(D)$$
4. In other words, the graphical model 'loses' the problematic link
$$P(G,R,D) = P(G)P(D)P(R|G,D)$$
5. The conditional estimate then becomes
$$P(R|D) = \frac{P(R,D)}{P(D)} = \frac{\sum_G P(G)P(D)P(R|G,D)}{P(D)} = \sum_G P(G)P(R|G,D)$$

The Magic of a Structural Model

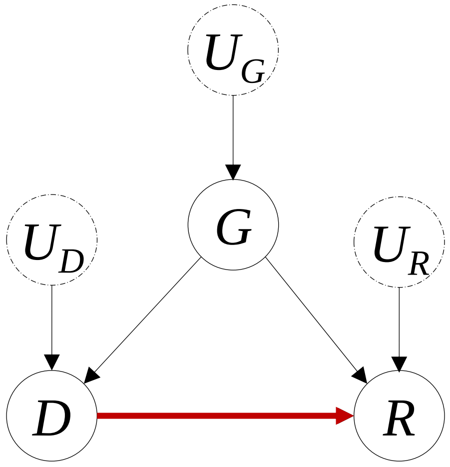


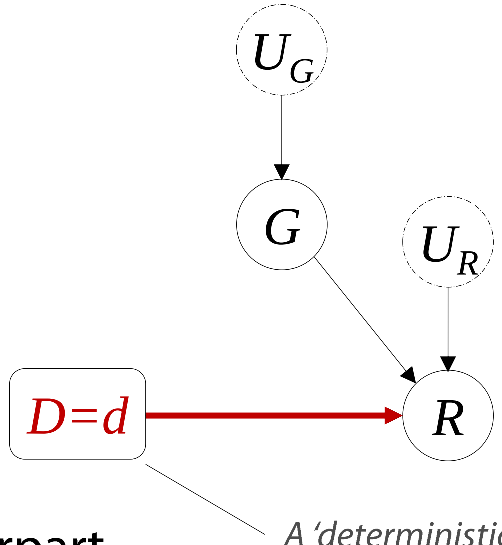
Assume that this *Graphical Model* has a *Structural* counterpart

$$
\begin{aligned}
G &:= U_G \\
D &:= f_D(G, U_D) \\
R &:= f_R(G, D, U_R)
\end{aligned}
\qquad \Rightarrow \qquad
\begin{aligned}
G &:= U_G \\
D &= d \\
R &:= f_R(G, D, U_R)
\end{aligned}
$$

In a line of principle, we could fix (by *intervention*) the value of D in the structural part
The other equation components would remain *unaltered*

The Magic of a Structural Model





Assume that this *Graphical Model* has a *Structural* counterpart

$$\begin{aligned} G &:= U_G \\ D &:= f_D(G, U_D) \\ R &:= f_R(G, D, U_R) \end{aligned} \quad \Rightarrow \quad \begin{aligned} G &:= U_G \\ D &= d \\ R &:= f_R(G, D, U_R) \end{aligned}$$

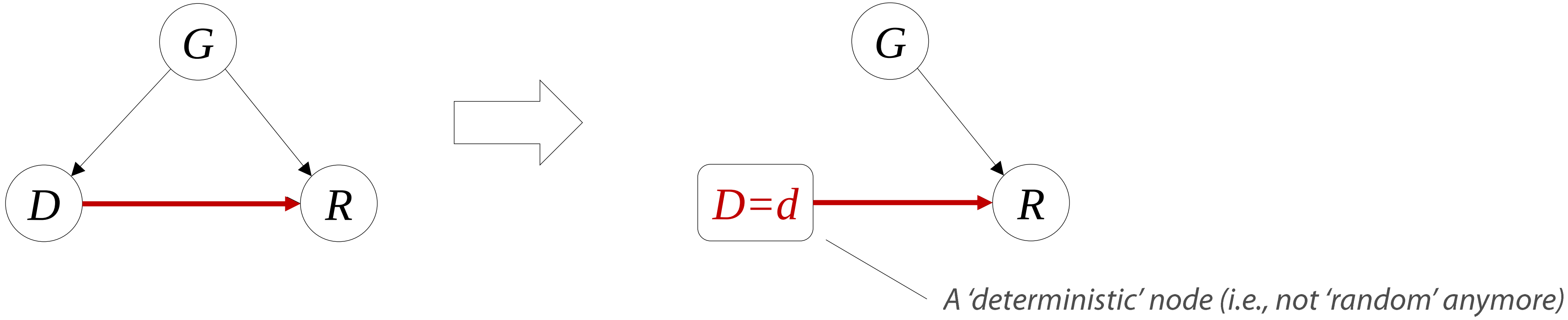
The corresponding joint probability distribution becomes:

$$P(G, R, D) = P(G)P(D|G)P(R|G, D) \quad \Rightarrow \quad P(G, R, D = d) = P(G)P(R|G, D = d)$$

do-calculus

do-calculus (*the intuition*)

- **From Conditional (pre-intervention) to Intervention Probability**



In this *Graphical Model* (for an <u>uncontrolled</u> experiment):

1. Conditional probability:

$$P(R|D=d) = \frac{\sum_G P(G)P(R|G,D=d)P(D=d|G)}{\sum_G P(G)P(D=d|G)}$$

2. Intervention (do-calculus, *this is new*)

$$P(R|do(D=d)) := \sum_G P(G)P(R|G,D=d)$$

These two expression would be identical if

$$P(D=d|G)=1$$

which cannot be true in general

3. *This is equivalent to* $P(R|D=d)$ *in a <u>modified</u> graphical in which we 'enforce intervention'*

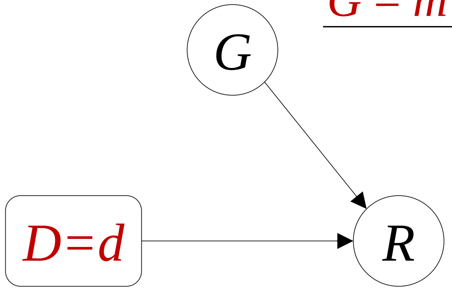
do-calculus

- **From Conditional (pre-intervention) to Intervention Probability**

(same observational probabilities, from MLE)

$P(G)$

$G = f$	0.49
$G = m$	0.51



$P(R|G, D)$

	$G = f$, $D = 0$	$G = f$, $D = 1$	$G = m$, $D = 0$	$G = m$, $D = 1$
$R = 0$	0.31	0.27	0.13	0.07
$R = 1$	0.69	0.73	0.87	0.93

Using do-calculus

$$P(R=1|do(D=0)) = \sum_G P(G)P(R=1|G, D=0)$$

$$= 0.49 \cdot 0.69 + 0.51 \cdot 0.87 = 0.78$$

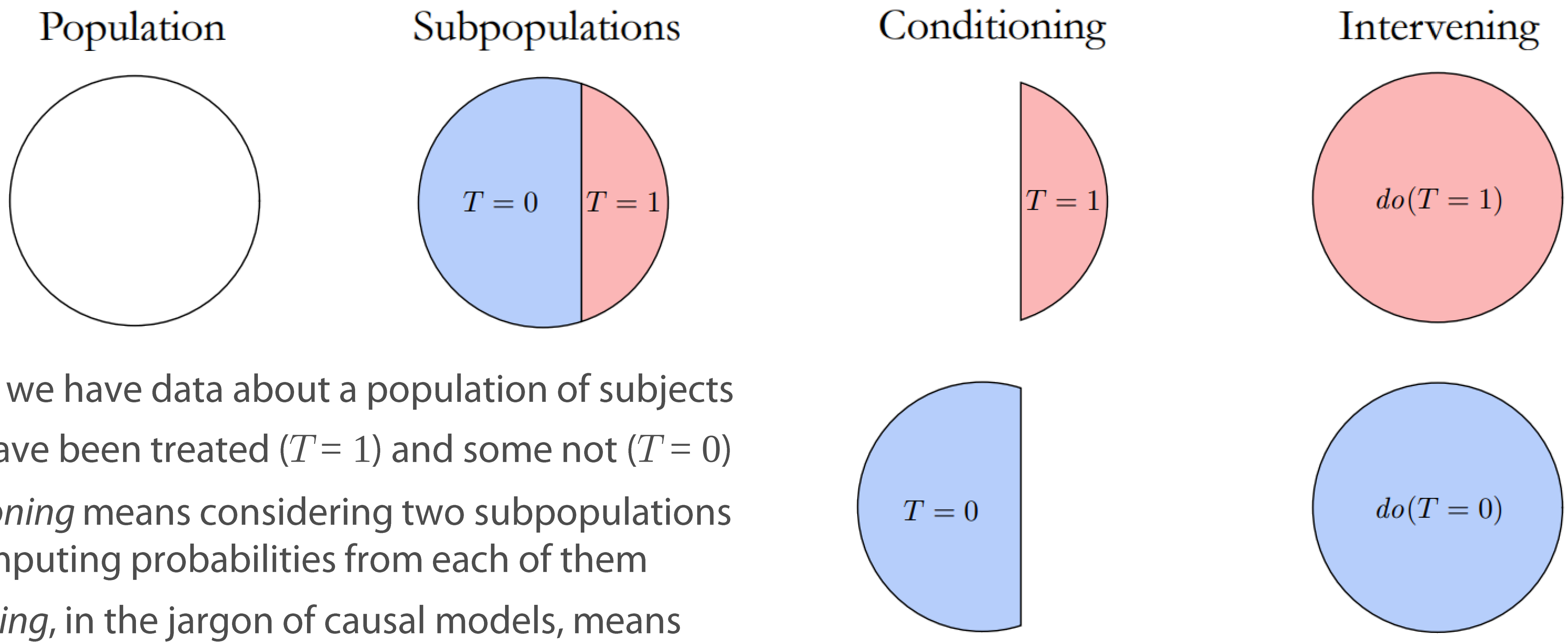
$$P(R=1|do(D=1)) = \sum_G P(G)P(R=1|G, D=1)$$

$$= 0.49 \cdot 0.73 + 0.51 \cdot 0.93 = 0.83$$

Prescribe drug, regardless

Causation and Conditionals

- **Conditioning and Intervening**



Assume we have data about a population of subjects

Some have been treated ($T = 1$) and some not ($T = 0$)

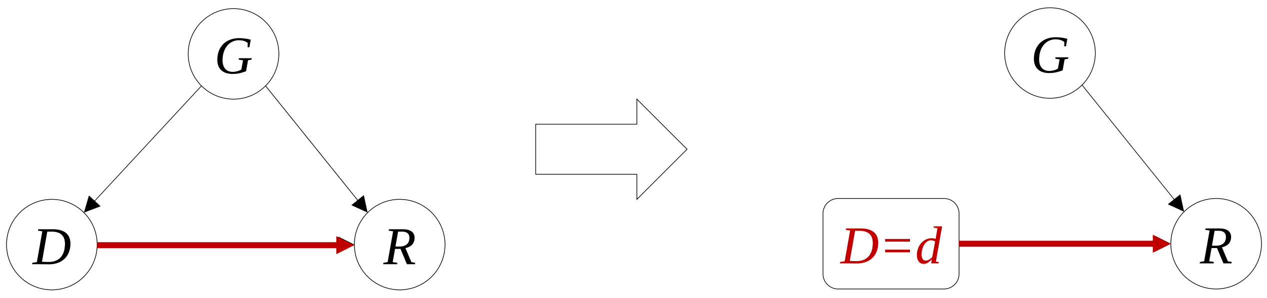
Conditioning means considering two subpopulations and computing probabilities from each of them

Intervening, in the jargon of causal models, means assuming that every subject in the population has either been treated or not treated

[Image from https://www.bradyneal.com/causal-inference-course]

do-Calculus

- **Compare two expressions**



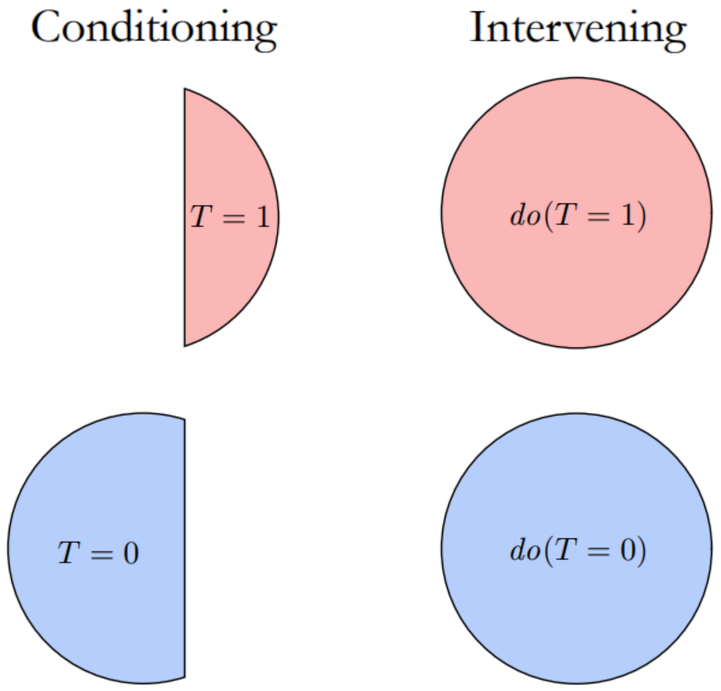
1. Conditioning:

$$P(R|D=d) = \frac{\sum_G P(G)P(R|G,D=d)P(D=d|G)}{\sum_G P(G)P(D=d|G)}$$

2. Intervening:

$$P(R|do(D=d)) := \sum_G P(G)P(R|G,D=d)$$

no normalization =
no conditional subspace



Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable*
  iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y

  This criterion is necessary and sufficient for <u>effect identifiability</u>

  Then:

$$P(Y|do(T=t)) = \sum_{\boldsymbol{W}} P(Y|T=t,\boldsymbol{W})P(\boldsymbol{W})$$

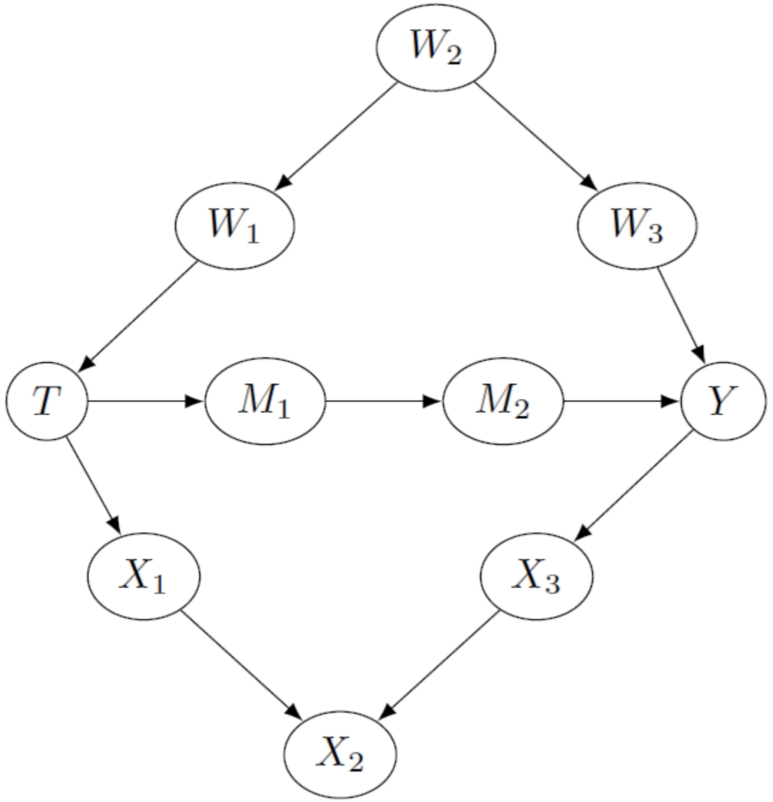
  In words, under the above conditions, the <u>causal effect</u> can be estimated statistically, from data

Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable*
  iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y



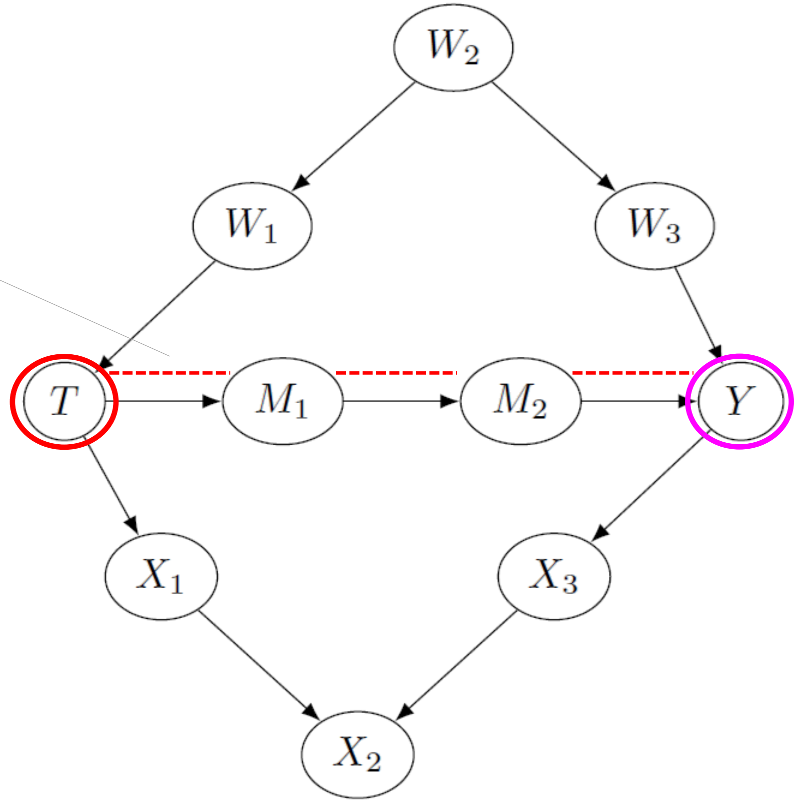
Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y

This is the causal effect that we are interested in



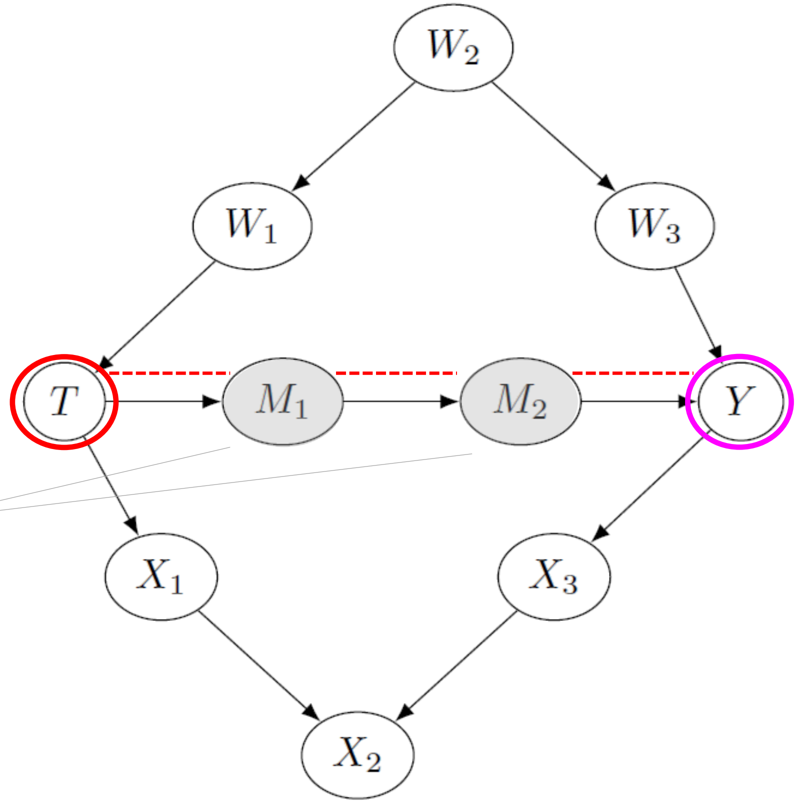
Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y

For identifiability *<u>none</u> of the variables along the causal path must be in the adjustment set*



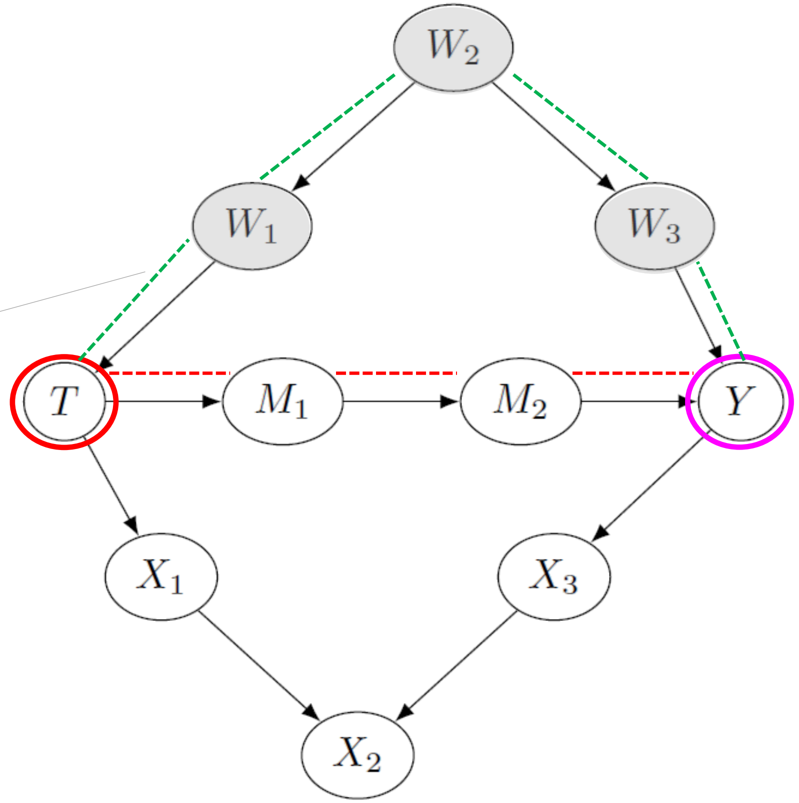
Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y

This path is non-causal and needs to be blocked by the adjustment set

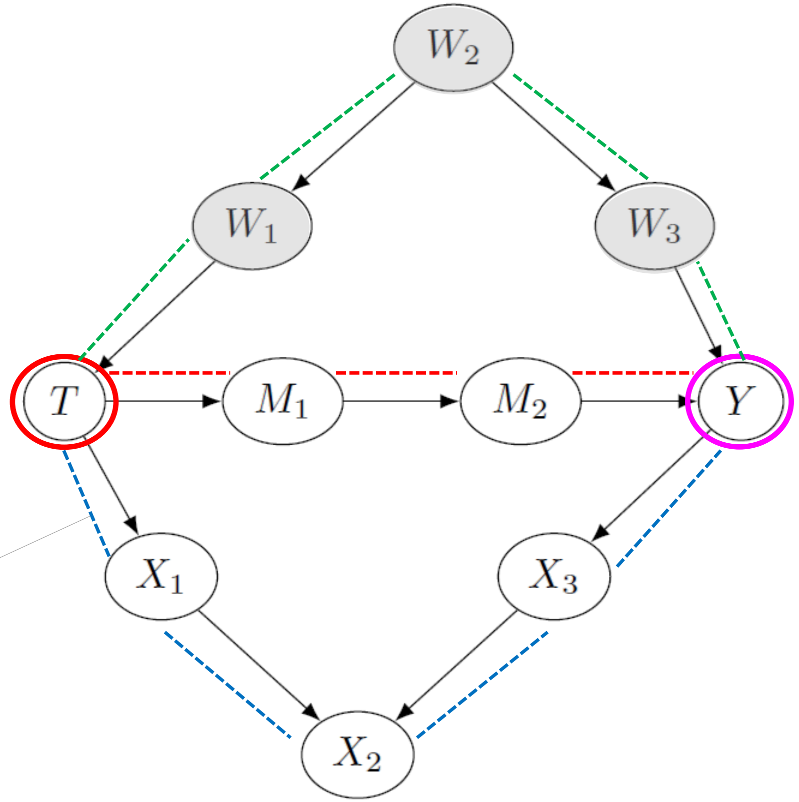


Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y



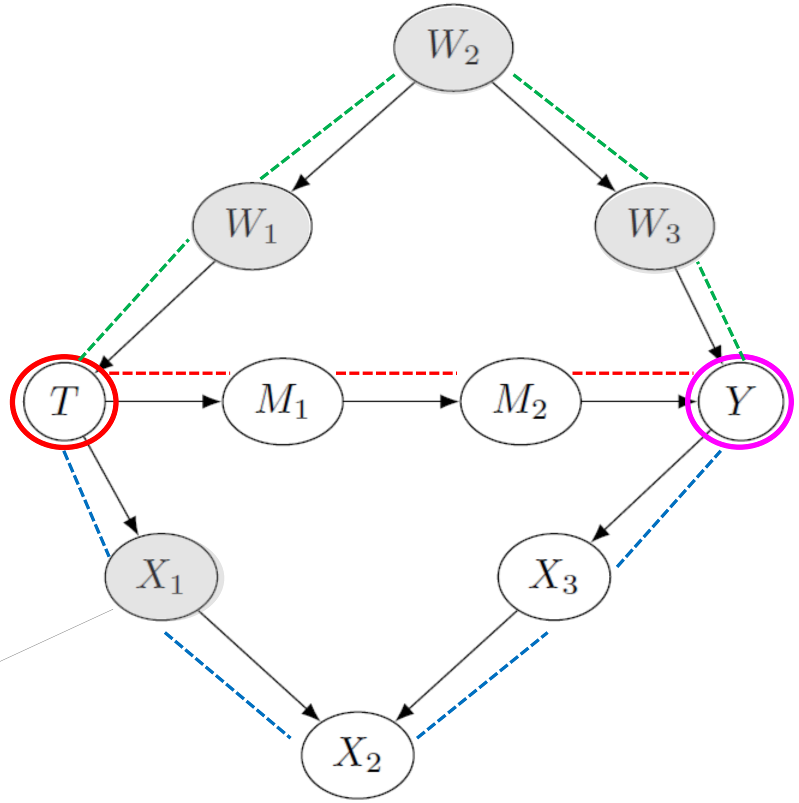
This path is non causal
yet it <u>blocked</u> AS IS:
the collider blocks it

Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y



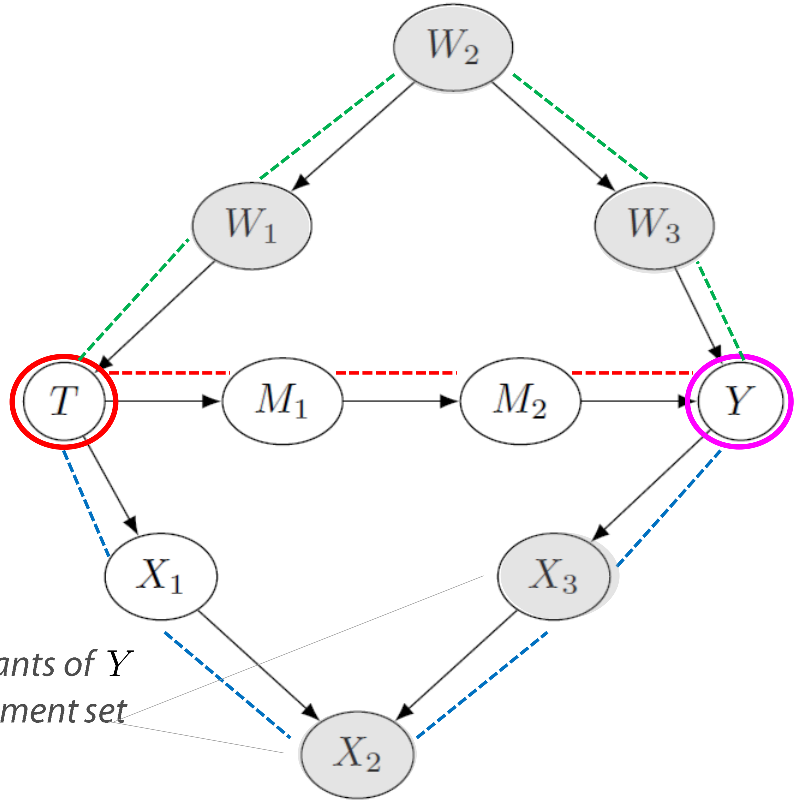
The observation of this variable is not problematic

Identifiability

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y



None of the descendants of Y must be in the adjustment set

Identification

- **Adjustment Set Criterion with *observed* and *unobserved* variables**

More in general, in practical cases,
there can be both observed and unobserved (possibly hidden) variables

An *adjustment set* can be composed of both:

$$\boldsymbol{W} = \boldsymbol{W}_{obs} \cup \boldsymbol{W}_{hid}$$

Then, if $\boldsymbol{W}$ satisfies the Adjustment Set Criterion:

$$P(Y|do(T=t), \boldsymbol{W}_{obs}) = \sum_{\boldsymbol{W}_{hid}} P(Y|T=t, \boldsymbol{W}_{hid}, \boldsymbol{W}_{obs}) P(\boldsymbol{W}_{hid})$$

When there are no *observed* variables in the adjustment set:

$$P(Y|do(T=t)) = \sum_{\boldsymbol{W}} P(Y|T=t, \boldsymbol{W}) P(\boldsymbol{W})$$

Likewise, when there are no *unobserved* variables in the adjustment set:

$$P(Y|do(T=t), \boldsymbol{W}) = P(Y|T=t, \boldsymbol{W})$$

Estimating Effects

Expected effects of different interventions can be estimated via do-calculus

In general, the *expected effect* on Y of treatment T will be

$$\mathbb{E}[Y|do(T=t), \boldsymbol{W}_{obs}] := \sum_{y \in \mathcal{Y}} y \; P(Y|do(T=t), \boldsymbol{W}_{obs})$$

where $\boldsymbol{W} = \boldsymbol{W}_{obs} \cup \boldsymbol{W}_{hid}$ is a valid *adjustment set*

Differences in effects can be measured by comparing expected effects.

As a special case, when $T \in \{0, 1\}$

- The *Conditional Average Treatment Effect* (CATE) is defined as:

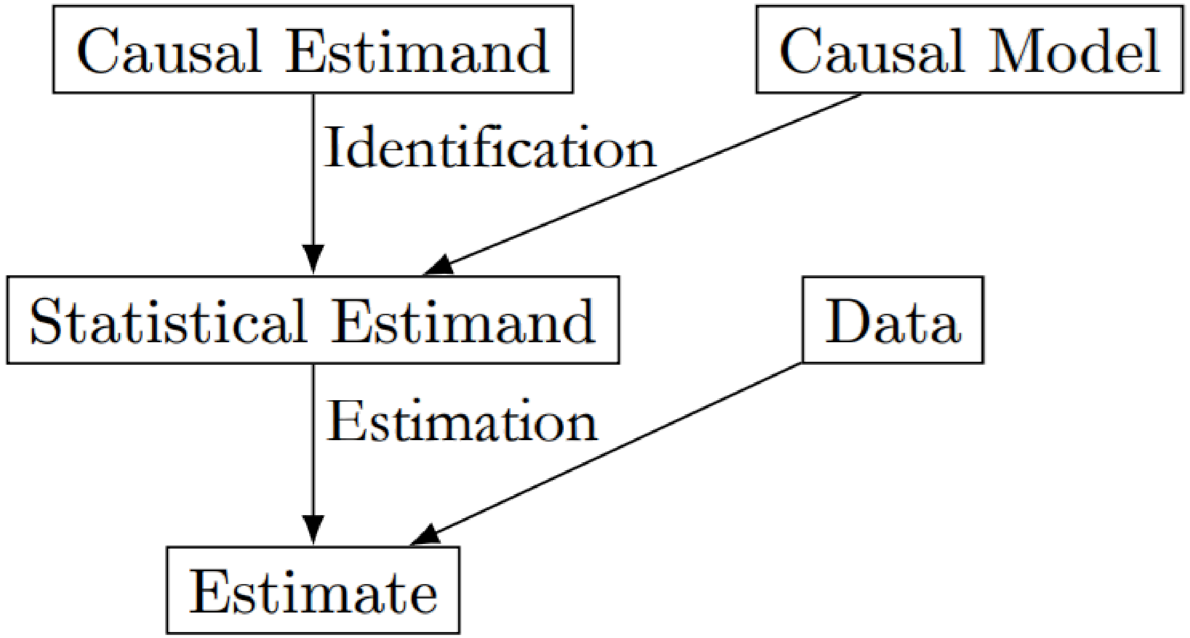
$$\tau(\boldsymbol{W}_{obs}) := \mathbb{E}[Y|do(T=1), \boldsymbol{W}_{obs}] - \mathbb{E}[Y|do(T=0), \boldsymbol{W}_{obs}]$$

- The *Average Treatment Effect* (ATE) is defined as:

$$\mathbb{E}[\tau(\boldsymbol{W})] := \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)]$$

Causation and Conditionals

- **Causal Model and Estimation**

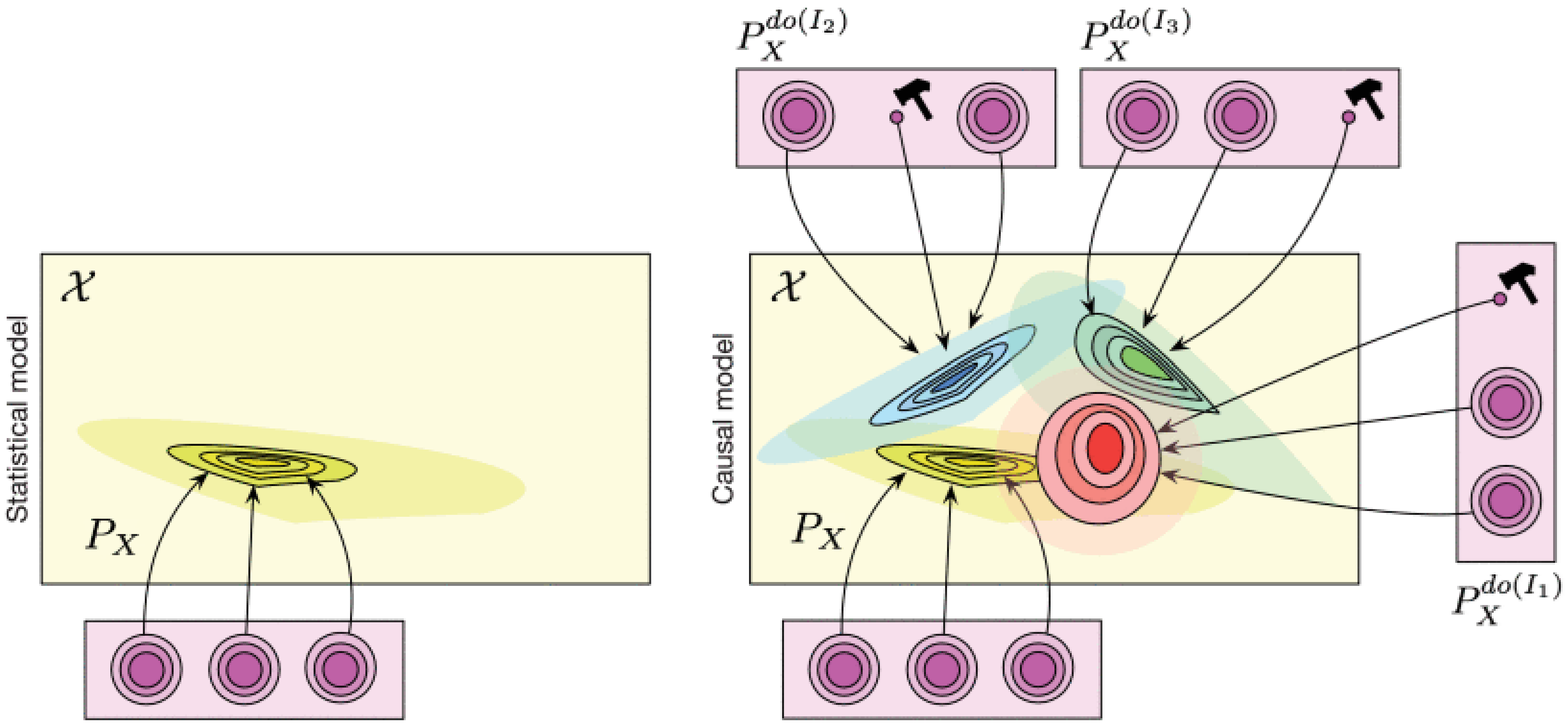


Basic principles:

1. Having selected what kind of causal effect we want to estimate
2. We start from a *Causal Model*
3. To translate the estimate into a statistical estimand, (*Identification*)
4. We use then *observational* data to compute the estimate: a *probability* or an *expected value*

[Image from https://www.bradyneal.com/causal-inference-course]

Statistical and Structural (Causal) Models



A graphical model without a structural component describes one probability distribution

A *structural* (*causal*) model represents *a family* of probability distributions,
one per each possible *intervention*

[Image from https://ieeexplore.ieee.org/abstract/document/9363924]

Inference through Counterfactuals (Causal)

Counterfactuals?

The ladder of causal inference [J. Pearl, *Causation*, Cambridge University Press, 2009]

- **Prediction**

  Given the probability distribution $P(\boldsymbol{X})$ *and some observations* $\boldsymbol{X}_o = \boldsymbol{x}_o$
  determine the probability $P(\boldsymbol{X}_u = \boldsymbol{x}_u \mid \boldsymbol{X}_o = \boldsymbol{x}_o)$ *for some unobserved variables* $\boldsymbol{X}_u$

- **Intervention**

  Intervene (i.e., force a change in value) on some variables $\boldsymbol{X}_i$ *and determine*
  the probability of effects $P(\boldsymbol{X}_e = \boldsymbol{x}_e \mid do(\boldsymbol{X}_i = \boldsymbol{x}_i))$

- **Counterfactual**

  Having observed $\boldsymbol{X}_o = \boldsymbol{x}_o$ *and its effects* $\boldsymbol{X}_e = \boldsymbol{x}_e$ *, what could be the probability*
  of different effects $\boldsymbol{x}'_e \neq \boldsymbol{x}_e$ *if some conditions* $\boldsymbol{X}_c \subseteq \boldsymbol{X}_o$ *were <u>different</u>?*

Counterfactual Inference

- **Counterfactual**

  Having observed $\boldsymbol{X}_o = \boldsymbol{x}_o$ *and its effects* $\boldsymbol{X}_e = \boldsymbol{x}_e$ *, what could be the probability of different effects* $\boldsymbol{x}'_e \neq \boldsymbol{x}_e$ *if some conditions* $\boldsymbol{X}_c \subseteq \boldsymbol{X}_o$ *were <u>different</u>?*

  A few relevant aspects:

  - *Prediction and Intervention occur <u>in the same world</u>, whereas counterfactuals require <u>alternative worlds</u>*
  - *Conceptually, counterfactuals relate to <u>potential outcomes</u> ("what could it be the outcome, were the condition different?")*
  - *Counterfactual inference can be performed at either <u>individual</u> or <u>population</u> level (more to follow)*

Counterfactual Inference

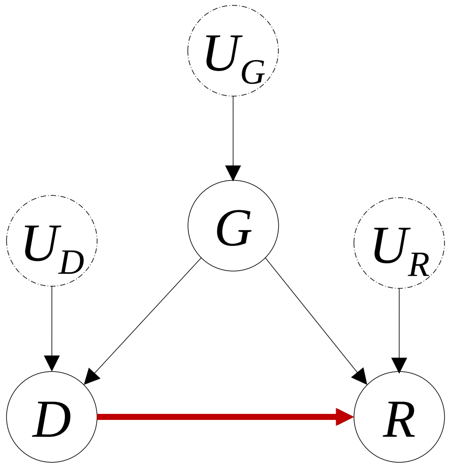
From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations for* $\boldsymbol{X}_c$ *with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$

Counterfactual Inference *(linear case)*

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations for* $\boldsymbol{X}_c$ *with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$



Example, in the linear case:

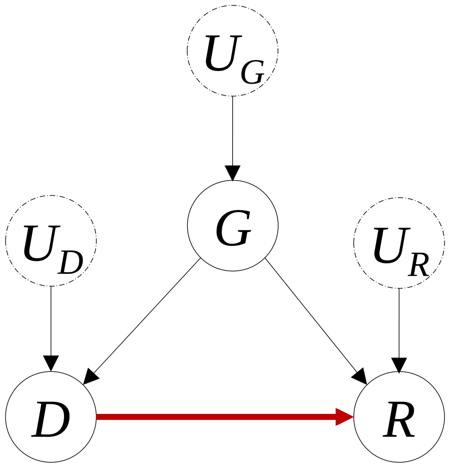
$$G := U_G$$
$$D := k_1 + \beta_1 G + U_D$$
$$R := k_2 + \beta_2 G + \beta_3 D + U_R$$

Assume all parameters are known (complete identification of $\mathcal{M}$ *)*

Counterfactual Inference *(linear case)*

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations for* $\boldsymbol{X}_c$ *with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$



Example, in the linear case:

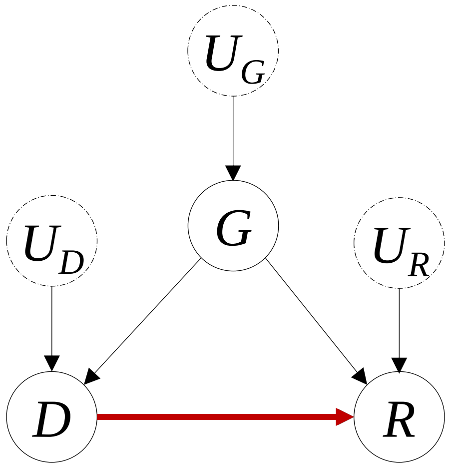
$$u_G = g_o$$
$$u_D := d_o - k_1 - \beta_1 g_o$$
$$u_R := r_o - k_2 - \beta_2 g_o + \beta_3 d_o$$

Replace with observed values and solve for $\boldsymbol{U}$

Counterfactual Inference *(linear case)*

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations for* $\boldsymbol{X}_c$ *with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$



Example, in the linear case:

$$g = u_G$$

$$d = \boxed{d_c} \quad \text{counterfactual value}$$

$$\text{effect} \quad \boxed{r_e} = k_2 + \beta_2 g_o + \beta_3 d_c + u_R$$

Plug back values $\boldsymbol{u}$, *impose counterfactual value* d_c *and compute the resulting effect* r_e

Counterfactual Inference, probabilistic case

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$

More in general, even keeping the assumption of complete identification of $\mathcal{M}$ *,*
what happens if some functions are not one-to-one for the values of $\boldsymbol{U}$ *?*

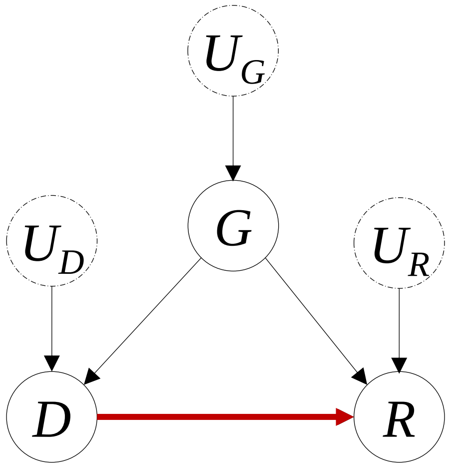
Note: this is the main point in which the causal hypothesis becomes fundamental

CAUSALITY: functions are considered as non-invertible, regardless

Counterfactual Inference, probabilistic case

From J. Pearl, Primer, 2016 (see suggested readings)

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$



Example, non-invertible case:

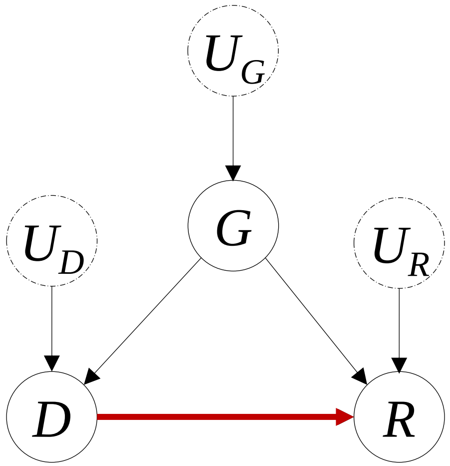
$$G := U_G$$
$$D := f_D(G, U_D)$$
$$R := f_R(G, D, U_R)$$

$\mathcal{M}$ *is still completely identified*

Counterfactual Inference, probabilistic case

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$



Example, non-invertible case:

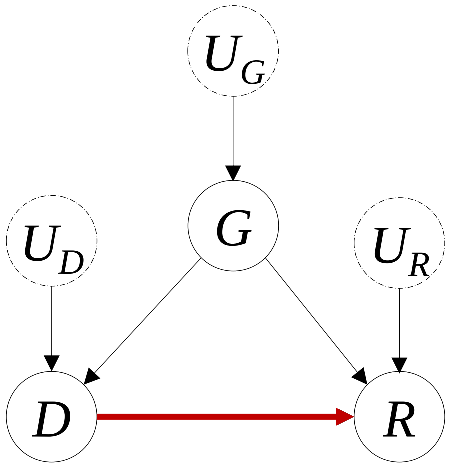
$$G = g_o$$
$$D = d_o$$
$$R = f_R(g_o, d_o, U_R) = r_o$$

Consider observed values

Counterfactual Inference, probabilistic case

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to determine the values $\boldsymbol{u}$ of the unobservable variables $\boldsymbol{U}$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model to compute* $\boldsymbol{x}_e$ *(the effects) by using* $\boldsymbol{u}$, $\boldsymbol{x}_c$ *and* $\boldsymbol{x}_o \setminus \boldsymbol{x}_c$



Example, non-invertible case:

$$G = g_o$$
$$D = d_o$$
$$R = f_R(g_o, d_o, \boxed{U_R}) = r_o$$

In general, it is not possible to solve for U_R
there may be a whole set of values of u_R
compatible with r_o

Counterfactual Inference, probabilistic case

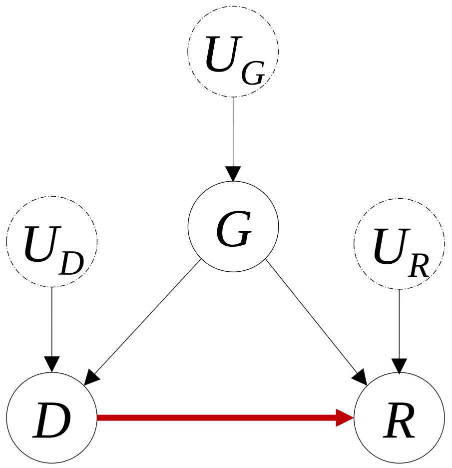
From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to update the probability distribution $P(\boldsymbol{U} \mid \boxed{\boldsymbol{X}_o = \boldsymbol{x}_o})$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boxed{\boldsymbol{X}_c = \boldsymbol{x}_c}$ *(counterfactual values)*
3. Prediction: *use the sub-model and the updated distribution to compute the probability distribution of possible effects* $\boldsymbol{x}_e$

Counterfactual Inference, probabilistic case

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to update the probability distribution $P(\boldsymbol{U} \mid \boldsymbol{X}_o = \boldsymbol{x}_o)$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model and the updated distribution to compute the probability distribution of possible effects* $\boldsymbol{x}_e$



Example, non-invertible case:

$$G = g_o$$

$$D = \boxed{d_c}$$

$$P(R_e = r_e) = \sum_{u_R \in \mathcal{U}_{R|r_e}} P(R = r_e, U_R = u_R | \boldsymbol{X}_o = \boldsymbol{x}_o)$$

where:

$$\mathcal{U}_{R|r_e} := \{u_R \ : \ f_R(g_o, d_c, u_R) = r_e\}$$

Counterfactual Inference, probabilistic case

From J. Pearl, *Primer*, 2016 *(see suggested readings)*

1. Abduction *(i.e., going in reverse)*: use a complete observation $\boldsymbol{x}_o$ to update the probability distribution $P(\boldsymbol{U} \mid \boldsymbol{X}_o = \boldsymbol{x}_o)$
2. Action: *create a sub-model of* $\mathcal{M}$ *by replacing the structural equations with* $\boldsymbol{X}_c = \boldsymbol{x}_c$ *(counterfactual values)*
3. Prediction: *use the sub-model and the updated distribution to compute the probability distribution of possible effects*

This is called <u>unit</u> or <u>individual-level</u> counterfactual inference
since it starts from the observation (possibly complete) of a specific case

It requires the complete identification of $\mathcal{M}$ *(including the distribution* $P(\boldsymbol{U}, \boldsymbol{X})$ *)*

Otherwise, there <u>are too many degrees of freedom</u>, and the inference problem is ill-posed

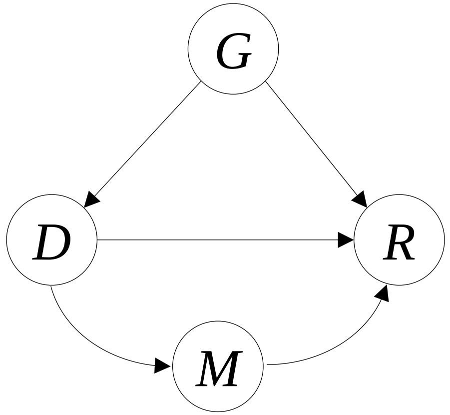
Counterfactual Inference, population level

What kind of counterfactual inference can be performed when the model $\mathcal{M}$ is NOT completely identified?

In other words, when what we have is the distribution $P(\boldsymbol{X})$ over endogenous variables as derived from actual observations?

Perhaps we should change the question somewhat:
what other kind of counterfactual inference could be useful in such case?

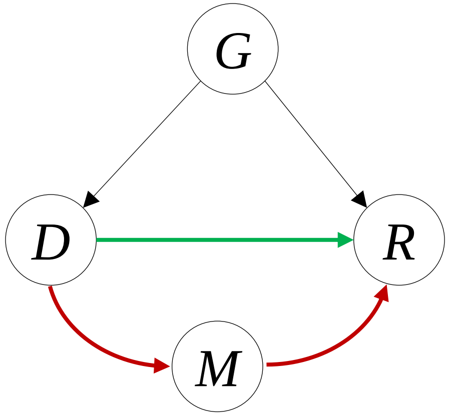
Path-Specific Counterfactual Evidence



Suppose that, as an extension to the previous model, we now assume that drug D has an observable side-effect M which also affects patient's recovery R

It is independent from gender G

Path-Specific Counterfactual Evidence



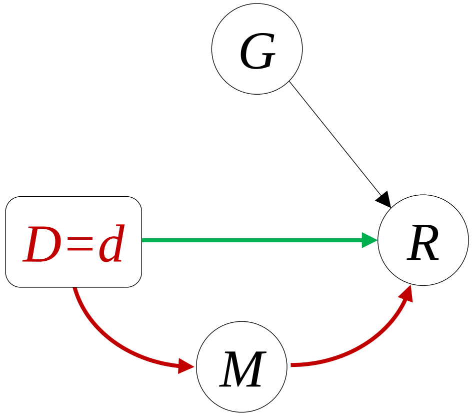
Suppose that, as an extension to the previous model, we now assume that drug D has an observable side-effect M which also affects patient's recovery R

It is independent from gender G

Now the model has two causal paths: one direct and another indirect, mediated by M

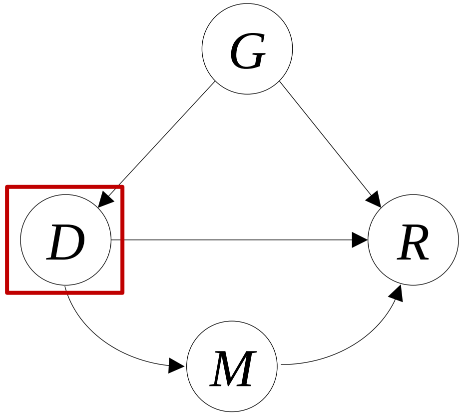
We might want to know what are the causal effects of each path in general, i.e., at the population level

Path-Specific Counterfactual Evidence



<u>Intervention</u> on D alone will not give the answer, as both paths need to be considered at once

Path-Specific Counterfactual Evidence



General idea (intuitive): splitting node D in two and letting different paths 'see' different values

SCM model $\mathcal{M}$

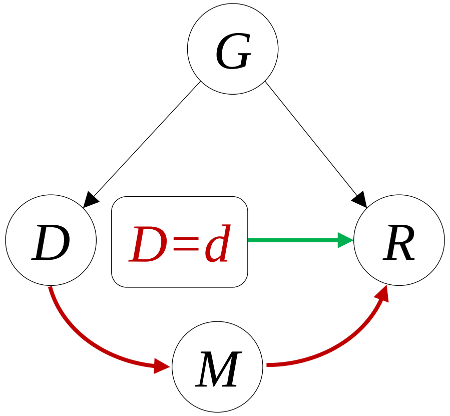
$$G := U_G$$

$$D := f_D(G, U_D)$$

$$M := f_M(D, U_M)$$

$$R := f_R(G, D, M, U_R)$$

Path-Specific Counterfactual Evidence



General idea (intuitive): splitting node D in two and letting different paths 'see' different values

Modified SCM model $\mathcal{M}'$

$$G := U_G$$

$$D := f_D(G, U_D)$$

$$M := f_M(D, U_M)$$

$$R := f_R(G, \boxed{D = d}, M, U_R)$$

When this is feasible, differences in path-specific effects can be evaluated from the observational distribution $P(\boldsymbol{X})$ alone

Counterfactual Inference
(see GeNIe 'berkeley_path_specific' attachment)

Download the GeNIe tool for free at: https://www.bayesfusion.com/genie/

Identifiability of Path-Specific Effects

- **Path-Specific Criterion (simplified)**

  In a SCM model $\mathcal{M}$, *path-specific effects* of T over Y with a *mediator* M are *identifiable* iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and M
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between M and Y

Extra requirements