Causal Models

Marco Piastra

Graphical Models: *dependence* and *independence*

Chain Factorization

- **Univariate factorization of a Joint Probability Distribution**

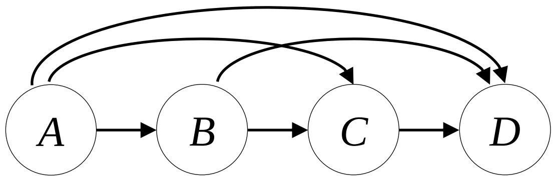
From the definition of conditional probability

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$

Any joint probability distribution can be factorized in a way such that each factor is *univariate* (i.e. one random variable as independent) conditional distribution.

- Each factorization depends on an arbitrary *sequence* of the *random variables*
- Hence factorizations are not *unique*: any sequence produces a legitimate factorization of the same kind

Graphical equivalent



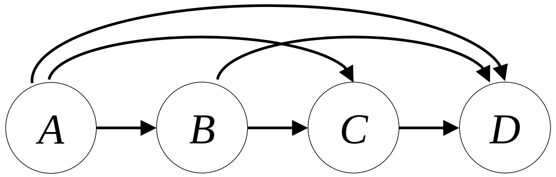
In this <u>oriented</u> graph:

- each node represents a random variable (and the corresponding *univariate* factor)
- each arc represents a conditioning of a random variable over another one (i.e. *dependence*)

Chain Factoriaztion

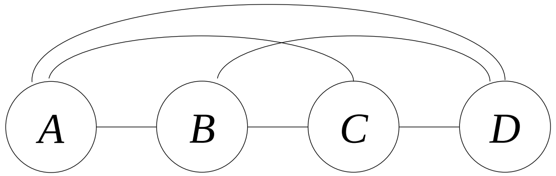
- **Graphical model**

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$



This graph:

- is *acyclic*: if you follow the arrows, you will never return to the same node
- is *completely connected*: if you ignore arc orientations, every node is connected to any other node



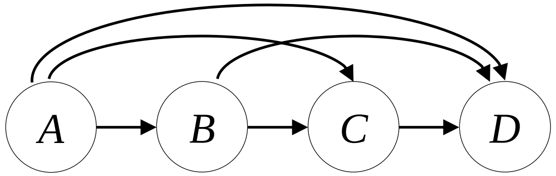
Any *univariate factorization* can be represented by a *graphical model*

Every *completely connected*, *acyclic* and *oriented graph* represents a *univariate factorization*

Chain Factorization and Independence Assumptions

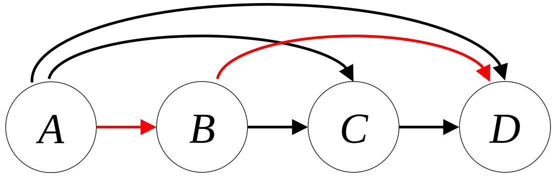
- **Graphical model**

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C)$$



- **Independence**

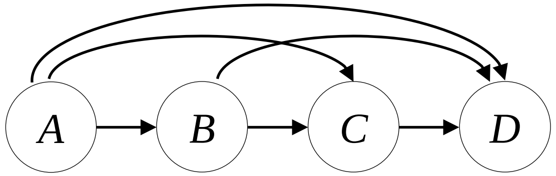
Let's remove a few arcs from the graph and rewrite the factorization accordingly



Chain Factorization and Independence Assumptions

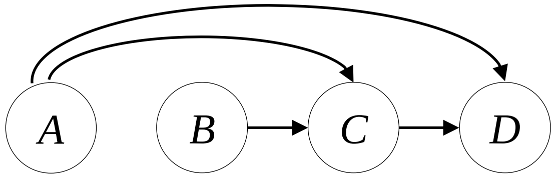
- **Graphical model**

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$



- **Independence**

Let's remove a few arcs from the graph and rewrite the factorization accordingly



$$P(A,B,C,D) = P(A)P(B)P(C|A,B)P(D|A,C)$$

The latter holds true only if

$$P(B|A) = P(A) \qquad \langle A \perp B \rangle$$

Independence

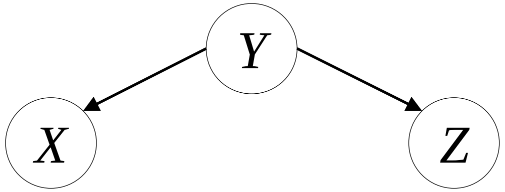
$$P(D|A,B,C) = P(D|A,C) \qquad \langle B \perp D \,|A,C \rangle$$

Conditional Independence

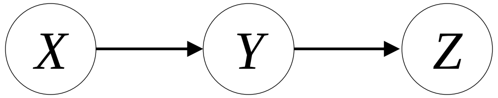
Graphical models and independence assumptions

- Structural equivalence

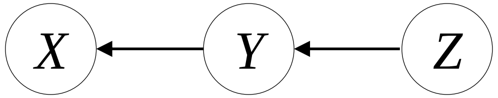
Different *structures*, different factorizations, same *independence* assumptions:



$P(Y)P(X|Y)P(Z|Y) \Rightarrow \langle X \perp Z | Y \rangle$

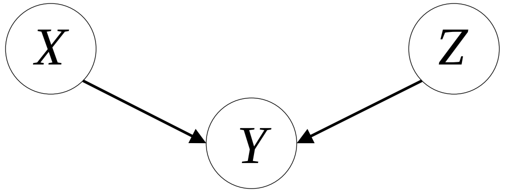


$P(X)P(Y|X)P(Z|Y) \Rightarrow \langle X \perp Z | Y \rangle$



$P(Z)P(Y|Z)P(X|Y) \Rightarrow \langle X \perp Z | Y \rangle$

Yet, this structure implies a different independence assumption:



$$P(X)P(Z)P(Y|X,Z) \Rightarrow \langle X \perp Z \rangle$$

Graphical models and independence assumptions

- Equivalence criterion

  Two graphical models share the same independence assumptions when:

  1) they share the same *undirected* structure (i.e., *skeleton*)
  2) they share the same *joins* (a.k.a. *colliders*)

  (*) This holds true when some independence is expressed (i.e., if some links are missing).
  Any DAG built out of a clique will be equivalent, regardless of joins
  (i.e., no independence assumptions represented anyway)

From *dependence* to *causation*

Causes and Effects: the Simpson's Paradox [1922]

- Does physical exercise prevent cholesterol?

Apparently not: correlation is *positive*

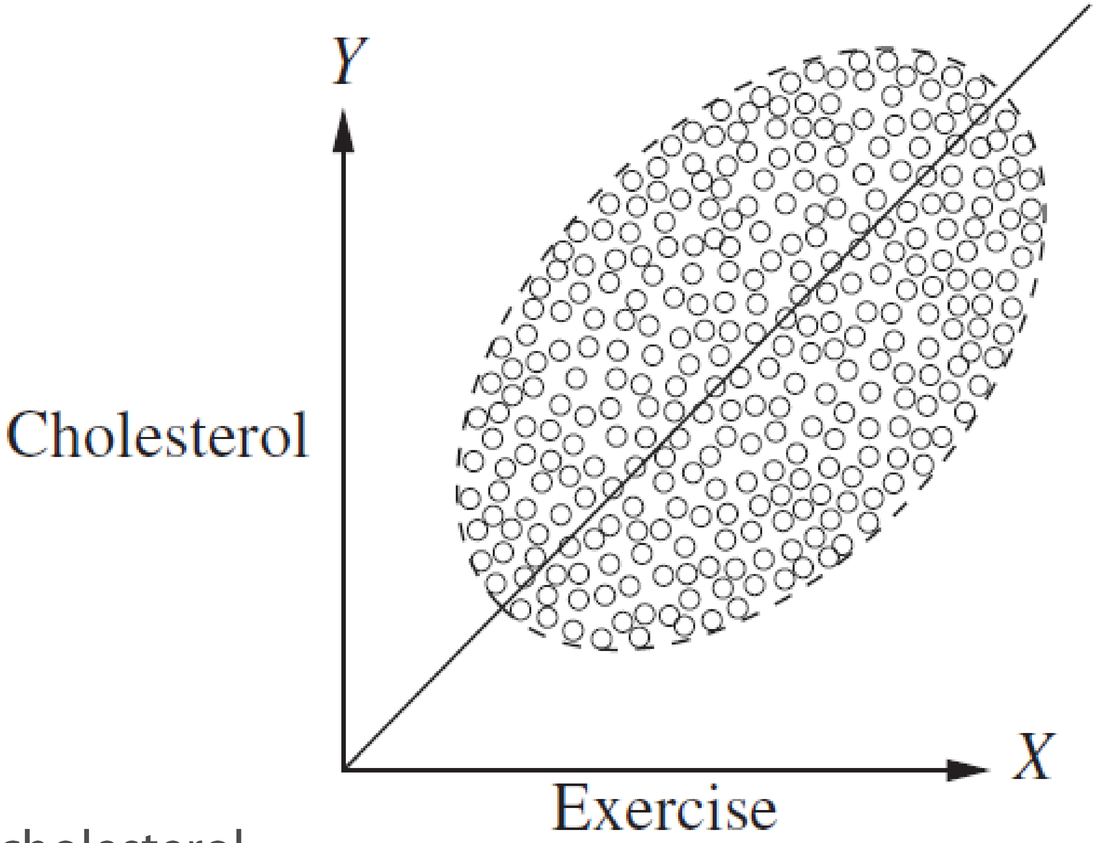
$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where:

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

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

standard deviation



In words:
more physical exercise corresponds to *(causes?)* more cholesterol ...

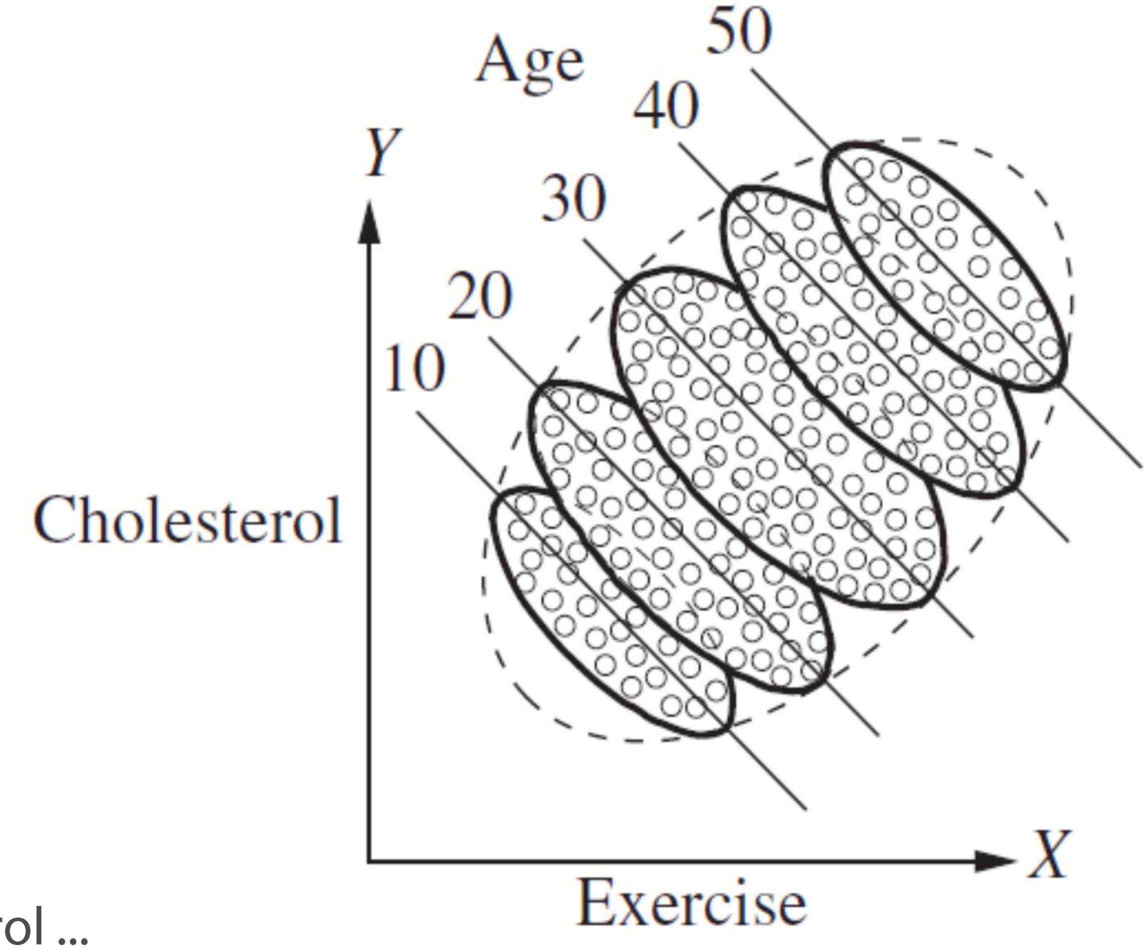
[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Causes and Effects: the Simpson's Paradox [1922]

- Does physical exercise prevent cholesterol?

  Maybe yes if we consider another variable...

  Correlation in Age subgroups is *negative*

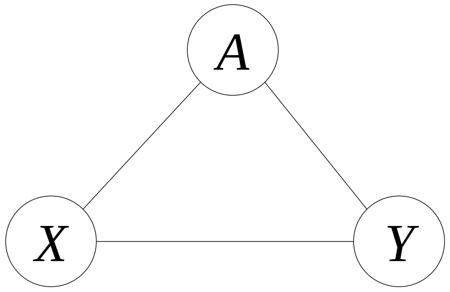


In words:
more exercise corresponds to *(causes?)* <u>less</u> cholesterol ...

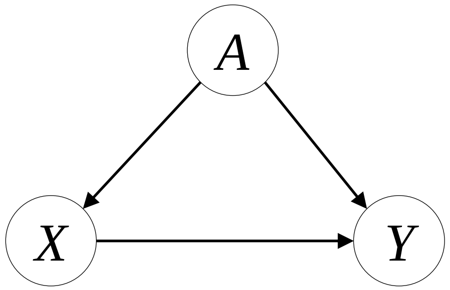
[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Causes and Effects: *say it with graphs*

- Does physical exercise prevent cholesterol?

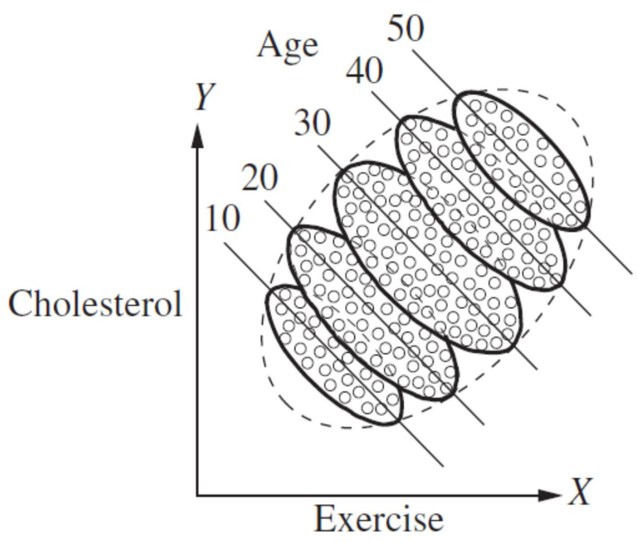


Undirected structure (a clique): no independence assumptions.
All DAGs built form it will be equivalent (just different factorizations)



Does this DAG make more sense from a <u>causal</u> viewpoint?

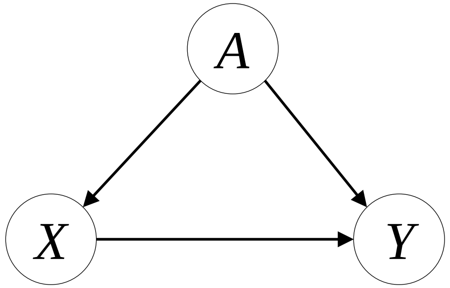
And what does this mean, after all?

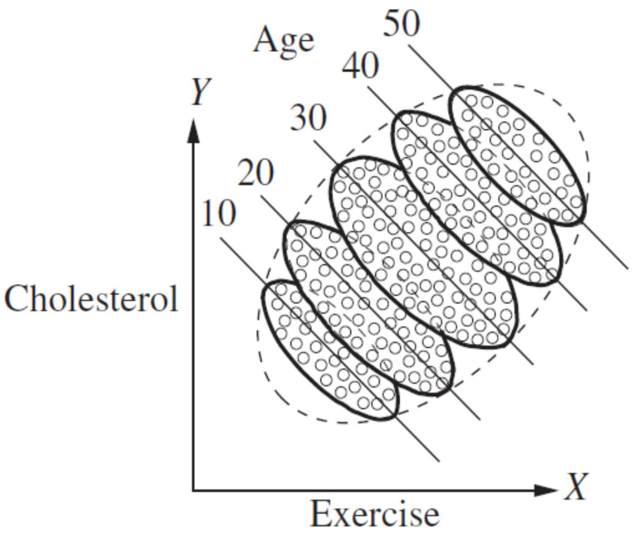


[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Causes and Effects: *say it with graphs*

- What is a cause?





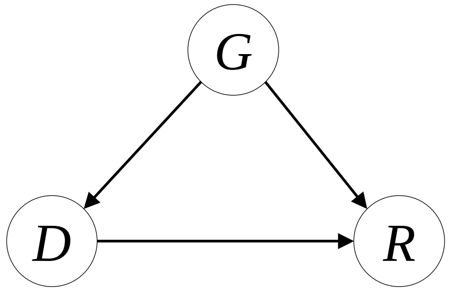
A variable X is said to be a <u>*cause*</u> of a variable Y
if Y can change in response to changes in X

In a *Causal Graphical Model* (CGM), each parent is a <u>*direct cause*</u> of all of its children

(*) *Independence assumptions are hard to elicit from data, whereas causal assumptions are <u>impossible</u> to elicit.*
No observation will tell us what could happen if we changed the state of things (counterfactuals)

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

- What is a cause? (*Another example*)



Variable G is biological gender (= male / female)
Variable D is drug administration (= yes / no)
Variable R is recovery from illness (= yes / no)

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	

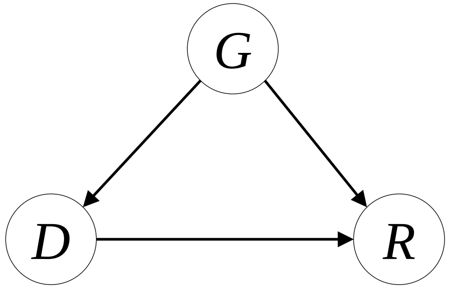
Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		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? (*Another example*)



Variable G is biological gender (= male / female)
Variable D is drug administration (= yes / no)
Variable R is recovery from illness (= yes / no)

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]

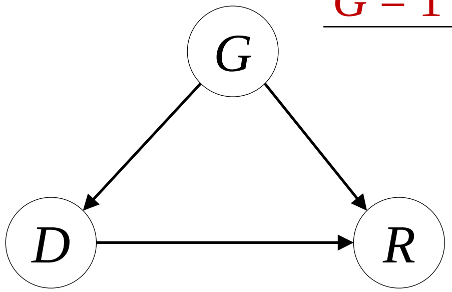
Causes and Effects: *say it with graphs*

- What is a cause? (*Another example*)

  Maximum Likelihood Estimation (CPTs)

$P(G)$

$G = 0$	0.49
$G = 1$	0.51



$P(D|G)$

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

$P(R|G, D)$

	$G = 0$, $D = 0$	$G = 0$, $D = 1$	$G = 1$, $D = 0$	$G = 1$, $D = 1$
$R = 0$	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	

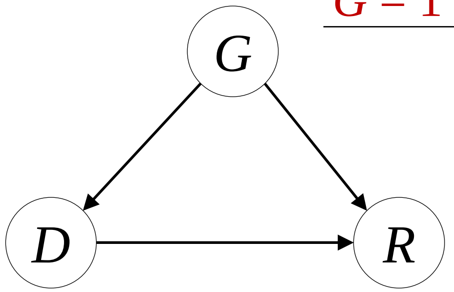
	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

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

- What is a cause? (*Another example*)

Maximum Likelihood Estimation (CPTs)

$P(G)$	
$G = 0$	0.49
$G = 1$	0.51



$P(D\|G)$	$G = 0$	$G = 1$
$D = 0$	0.23	0.76
$D = 1$	0.77	0.24

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

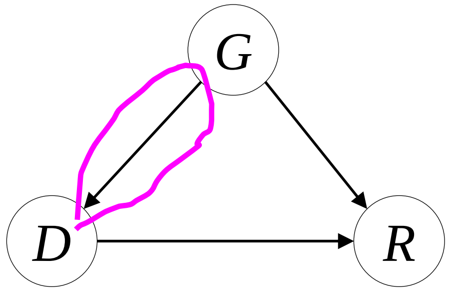
$$P(R|D)=\frac{\sum_G P(R|G,D)P(D|G)P(G)}{\sum_{G,R} P(R|G,D)P(D|G)P(G)}$$

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

Do not prescribe drug, regardless

(ridiculous!)

- What is a cause? (*Another example*)



Variable G is biological gender (= male / female)
Variable D is drug administration (= yes / no)
Variable R is recovery from illness (= yes / no)

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	

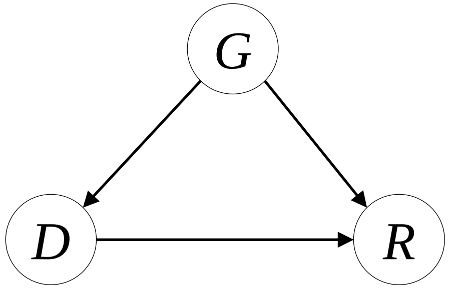
Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		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]

Causation and observations

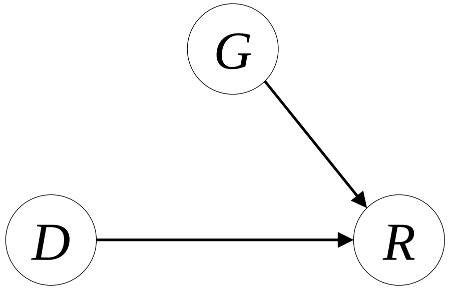
- **Confounders**



In this example, the problem is that G represents a 'common cause' of both D and R
It is a *confounder*, if we are interested in the causal link from D to R

In a controlled experiment, we could administer drug *at random*, regardless of G

In this case we would have:

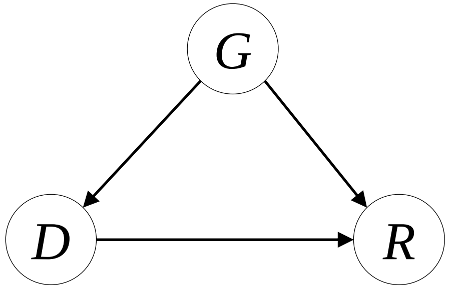


$$< D \perp G > \implies P(D|G) = P(G)$$

Can we always neutralize confounders in this way?

Causation and observations

- **Counterfactuals, potential outcomes**



In many circumstances, data are acquired in an *uncontrolled* ways: they are mere *observations*

We might still circumvent the problem if we knew would have happened if actions were *different (i.e., counterfactuals* or *potential outcomes)*

It may be seen as a problem of missing data in the dataset:

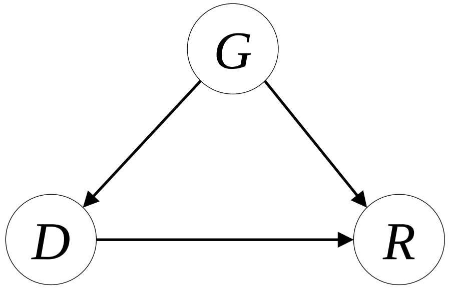
Subject	G	D	$R(D=0)$	$R(D=1)$
1	0	1	?	1
2	1	1	?	0
3	1	0	1	?
4	0	1	?	1
5	0	0	0	?
…	…	…	…	…
N	1	0	1	?

factual outcomes

counterfactual outcomes

Causation and observations

- **Counterfactuals, potential outcomes**



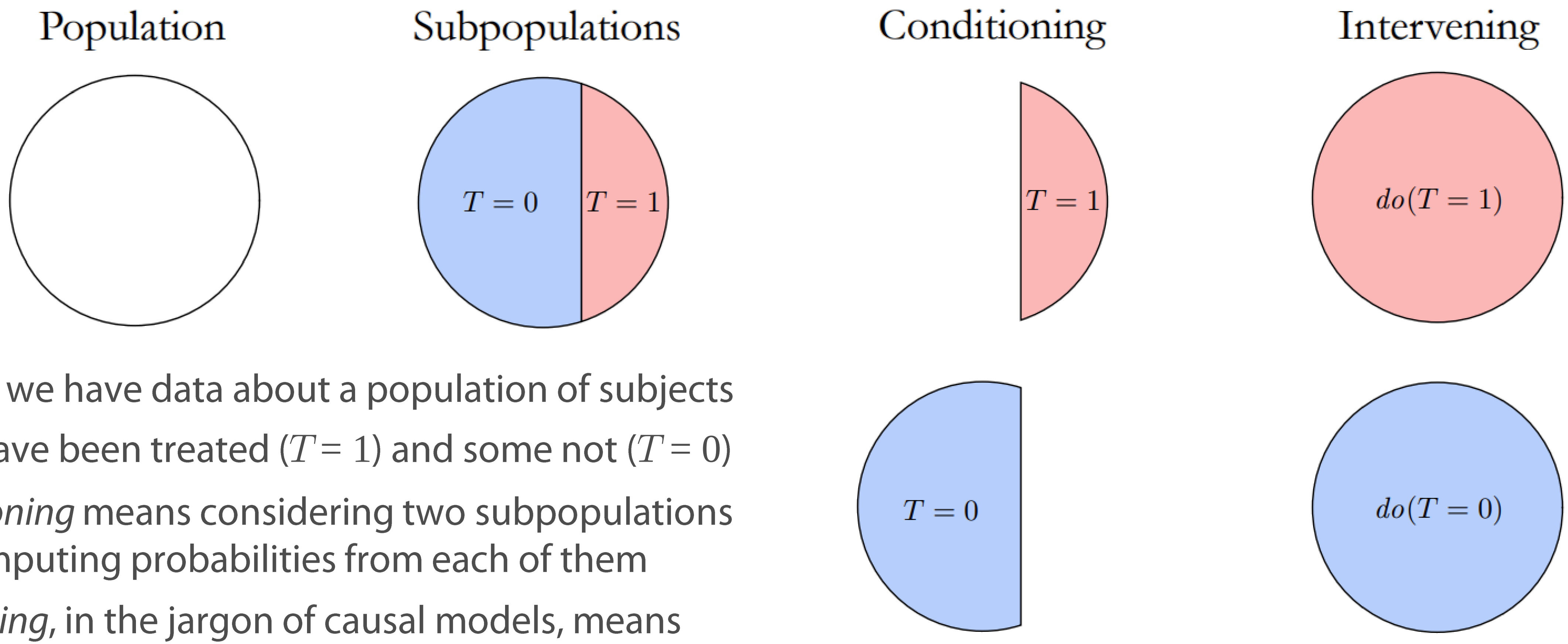
In many circumstances, data are acquired in an uncontrolled ways: they are mere *observations*

Can we work around all of this,
even with data from uncontrolled (i.e., observational) experiments?

Causal Models
(do-calculus)

Causation and Conditionals

- **Conditioning and Intervening**



Assume we have data about a population of subjects

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

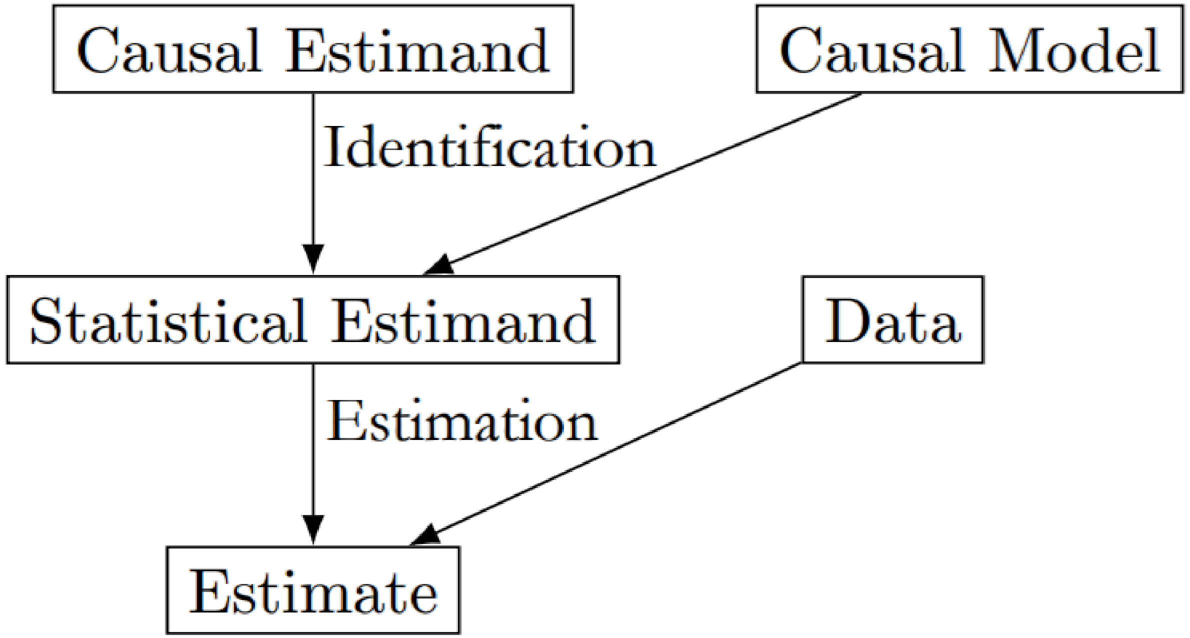
Conditioning means considering two subpopulations and computing probabilities from each of them

Intervening, in the jargon of causal models, means assuming that every subject in the population has been treated or not (*potential outcomes*)

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

Causation and Conditionals

- **Causal Model and Estimation**



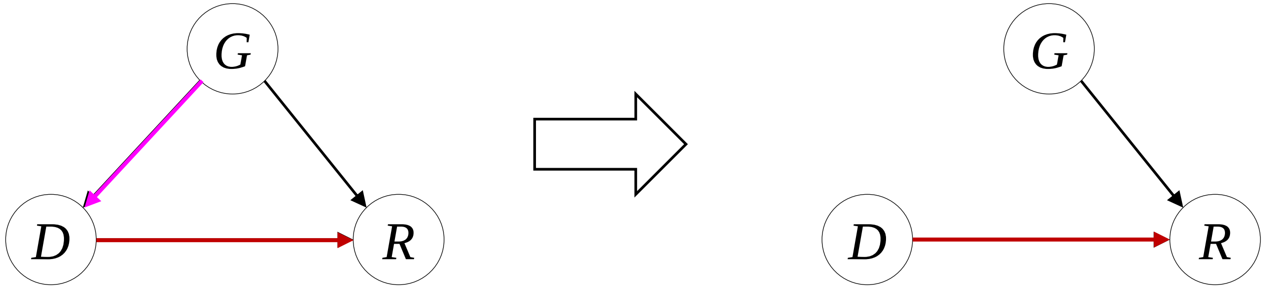
Basic principles:

1. Having selected what kind of causal effect we want to estimate
2. We start from a *Causal Graphical Model* (CGM)
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]

The Magic of Controlled Experiments

- **When association is causation**



In this *Causal Graphical Model*:

1. The causal effect we are interested 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

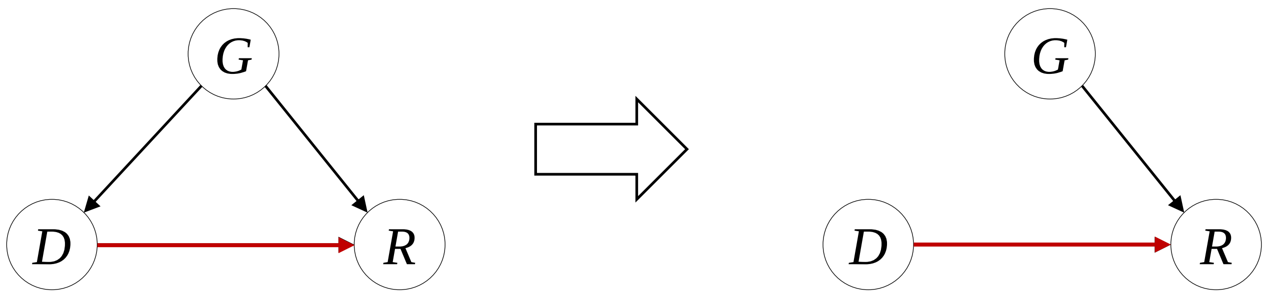
$$< D \perp G > \implies P(D|G=0) = P(D|G=1) = P(D)$$

4. *In other words, the CGM 'loses' the problematic link and the estimate becomes*

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

The Magic of Controlled Experiments

- **When association is causation**

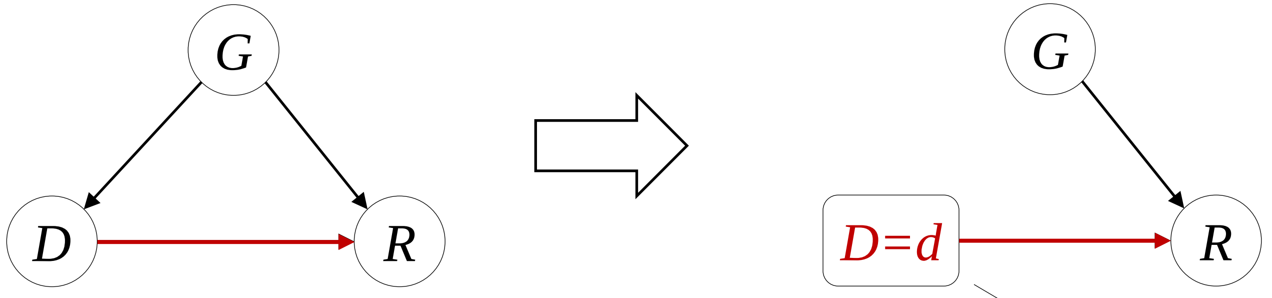


In *controlled experiments*, the principle is more general:

- by *randomizing* the administration of treatment
- we make the effects independent of any *confounders*
- be them observed or not

do-calculus

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



In this *Causal Graphical Model* (for an uncontrolled experiment):

1. Conditional probability:

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

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 hold true in general

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

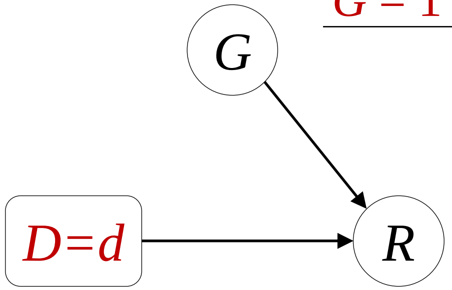
do-calculus

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

  (same observational probabilities, from MLE)

$P(G)$

$G = 0$	0.49
$G = 1$	0.51



$P(R|G,D)$

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

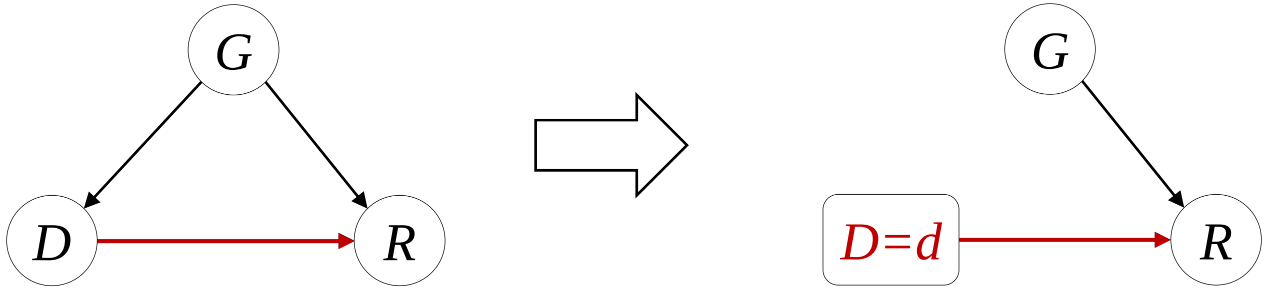
Using do-calculus

$$P(R=1|do(D=0)) = \sum_G P(G)P(R=1|G,D=0)$$
$$= 0.49 \cdot 0.69 + 0.51 \cdot 0.87 = 0.78$$

$$P(R=1|do(D=1)) = \sum_G P(G)P(R=1|G,D=1)$$
$$= 0.49 \cdot 0.73 + 0.51 \cdot 0.93 = 0.83$$

Prescribe drug, regardless

- **Compare two expressions**



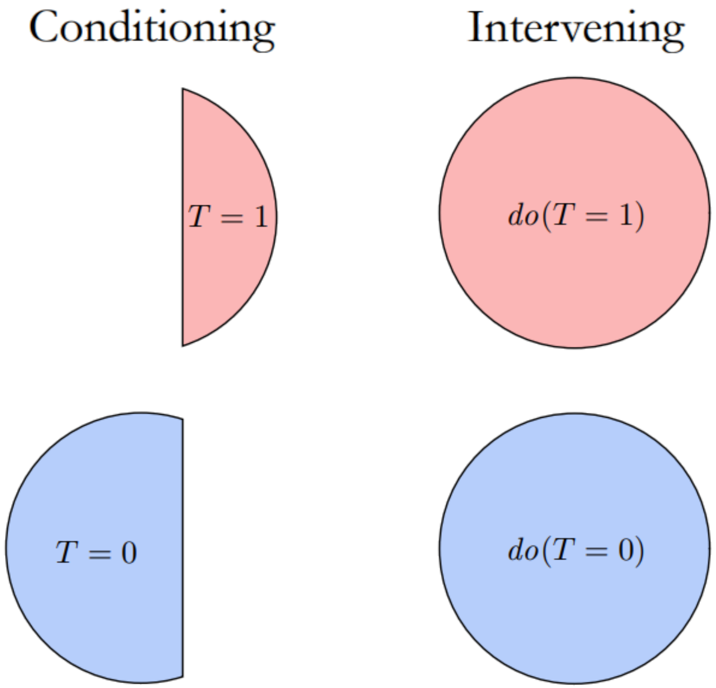
1. Conditional probability:

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

2. Intervention (do-calculus):

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

no normalization =
no conditional subspace



do-calculus:

Is it that simple?

(not so fast…)

do-Calculus

- **In general, in a Causal Graphical Model**
  1. Joint Probability Distribution

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

where $\{X_1, X_2, \ldots, X_n\}$ is the set of random variables in the model

  2. Intervention (do-calculus):

$$P(\{X_i\}_{i \neq k} | do(X_k = x_k)) = \prod_{i \neq k} P(X_i \mid parents(X_i))|_{X_k = x_k}$$

In general, do-calculus allows translating a ***causal estimand*** into a ***statistical estimand***, hence a *probability*

Under which conditions such translation is effective and justified?

do-Calculus

- **In general, in a Causal Graphical Model**

1. Joint Probability Distribution

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

   where $\{X_1, X_2, \ldots, X_n\}$ is the set of random variables in the model

2. Intervention (do-calculus):

$$P(\{X_i\}_{i \neq k} | do(X_k = x_k)) = \prod_{i \neq k} P(X_i \mid parents(X_i))|_{X_k = x_k}$$

In general, do-calculus allows translating a ***causal estimand*** into a ***statistical estimand***, hence a *probability*

Under which conditions such translation is effective and justified?

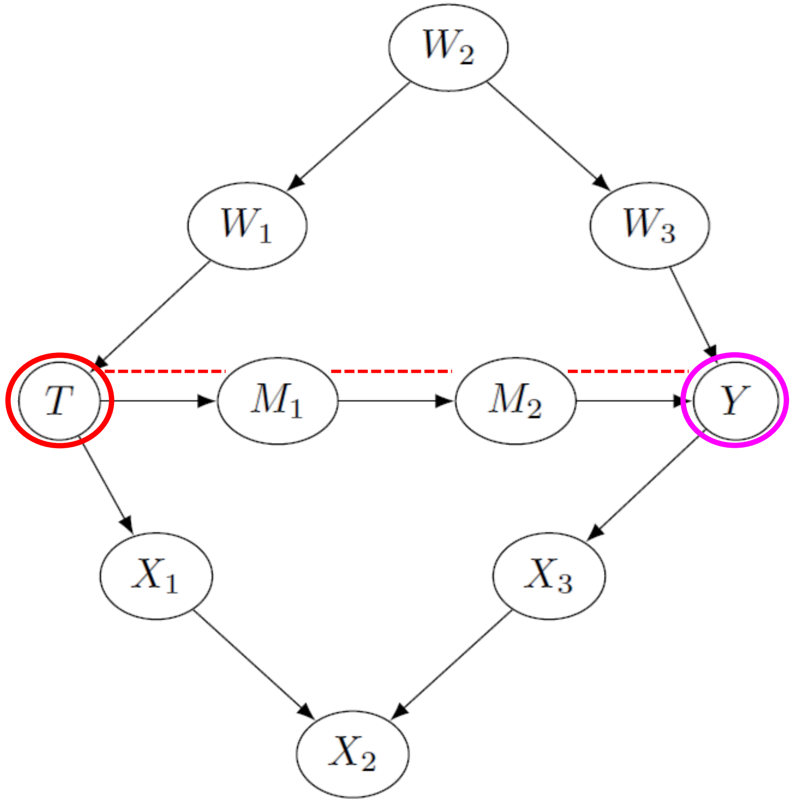
Identification

- **Causal Effect**

  In a more general Causal Graphical Model:

  1. Assume T over Y is the *causal effect* of interest
  2. Variables M_1 and M_2 are *mediators* of such effect
  3. All other variables in the model are *confounders*
  4. *Identify* the causal effect of T over Y we need to block any other paths, except the one of interest

     In the sense of *graphical models*...



Identification

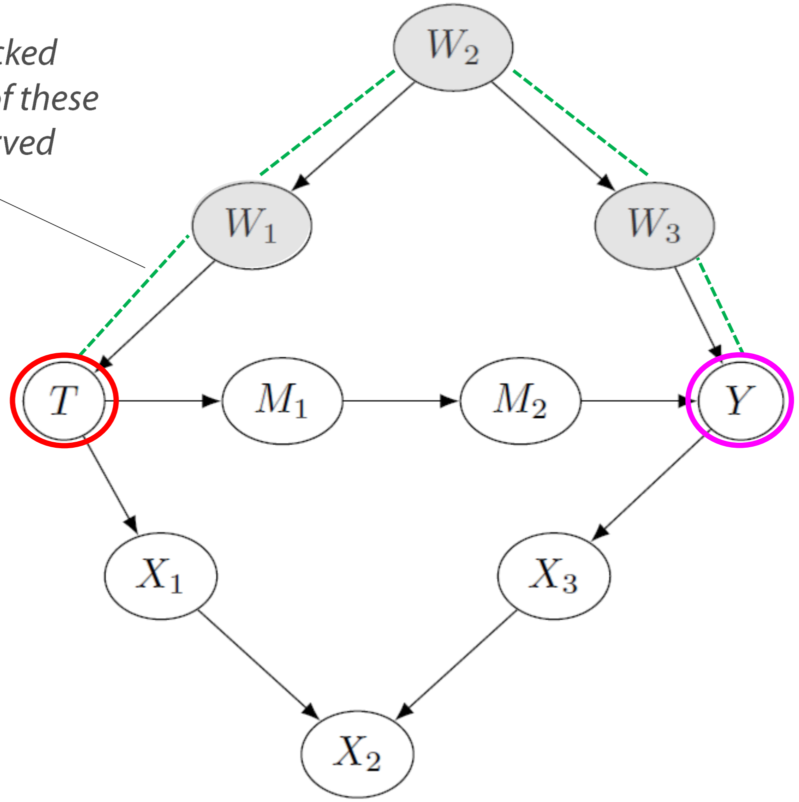
- **Causal Effect**

In a more general Causal Graphical Model:

1. Assume T over Y is the *causal effect* of interest
2. Variables M_1 and M_2 are *mediators* of such effect
3. All other variables in the model are *confounders*
4. *Identify* the causal effect of T over Y we need to block any other paths, except the one of interest

In the sense of *graphical models*...

This path is blocked whenever any of these nodes are observed

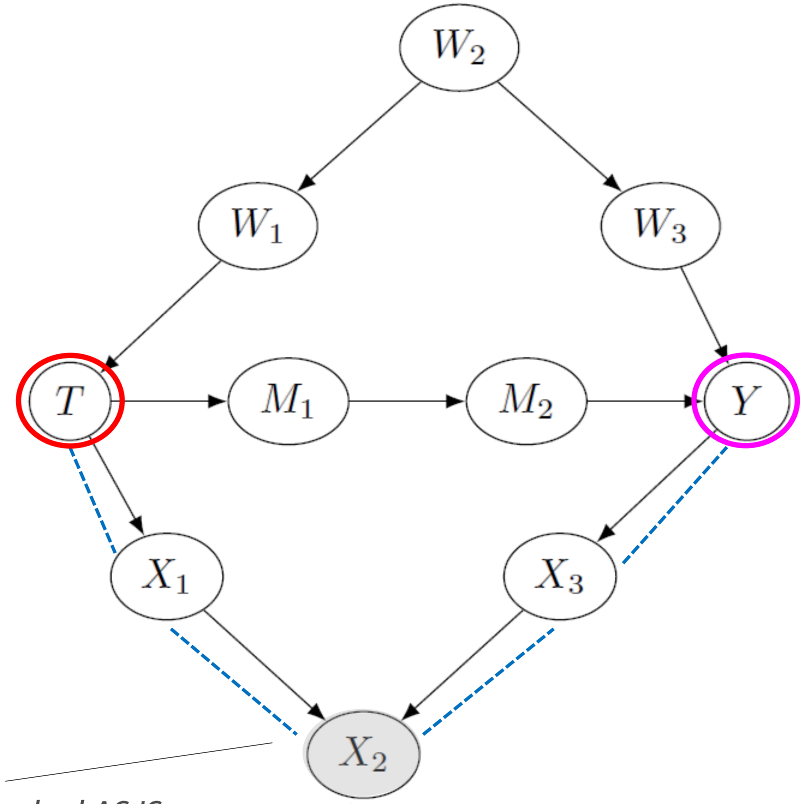


- **Causal Effect**

In a more general Causal Graphical Model:

1. Assume T over Y is the *causal effect* of interest
2. Variables M_1 and M_2 are *mediators* of such effect
3. All other variables in the model are *confounders*
4. *Identify* the causal effect of T over Y we need to block any other paths, except the one of interest

In the sense of *graphical models*...



This path is blocked AS IS: the collider blocks it

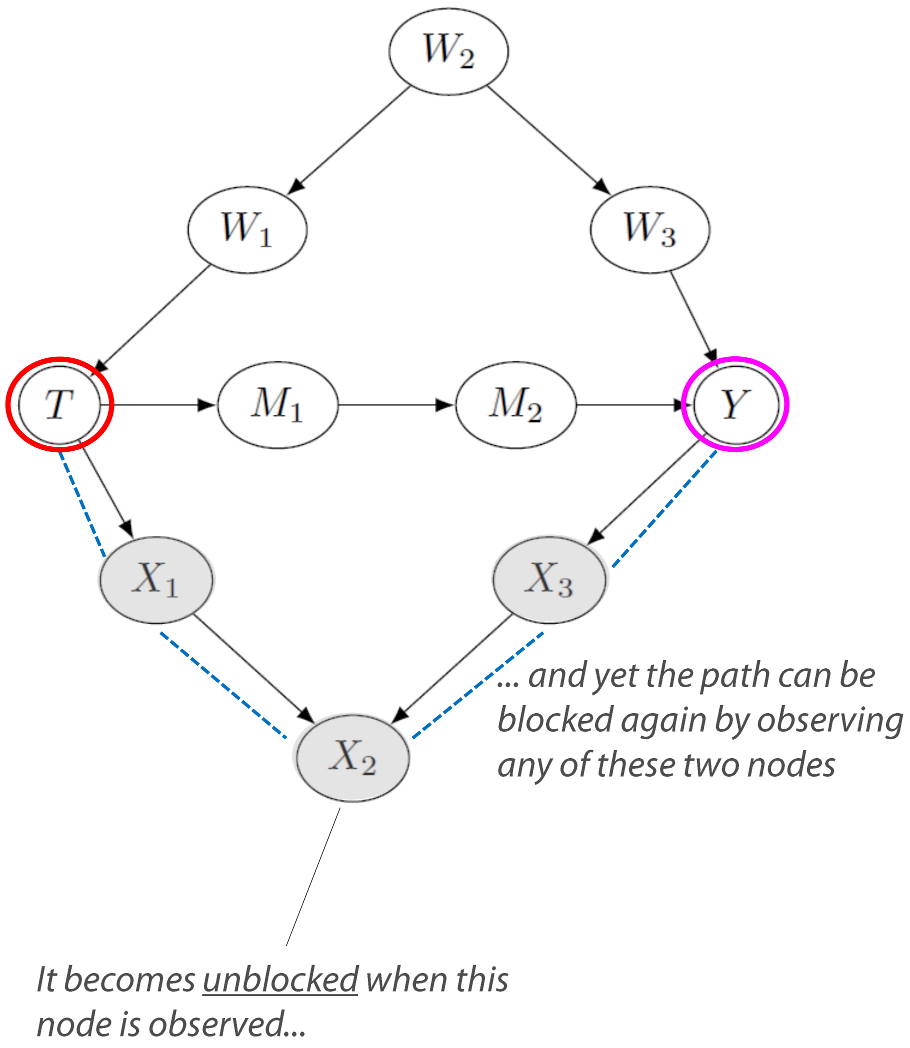
It becomes unblocked when this node is observed...

- **Causal Effect**

In a more general Causal Graphical Model:

1. Assume T over Y is the *causal effect* of interest
2. Variables M_1 and M_2 are *mediators* of such effect
3. All other variables in the model are *confounders*
4. *Identify* the causal effect of T over Y we need to block any other paths, except the one of interest

   In the sense of *graphical models*...



Identification

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

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

  - no *mediating* variable M in the *causal path*, nor any of its descendants, are in W
  - the variables in W block *(in the sense of graphical models)*
    all the non-causal paths between T and Y

  This criterion is necessary and sufficient for identifiability

  Then:

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

  In words, the causal effect can be estimated statistically, from data

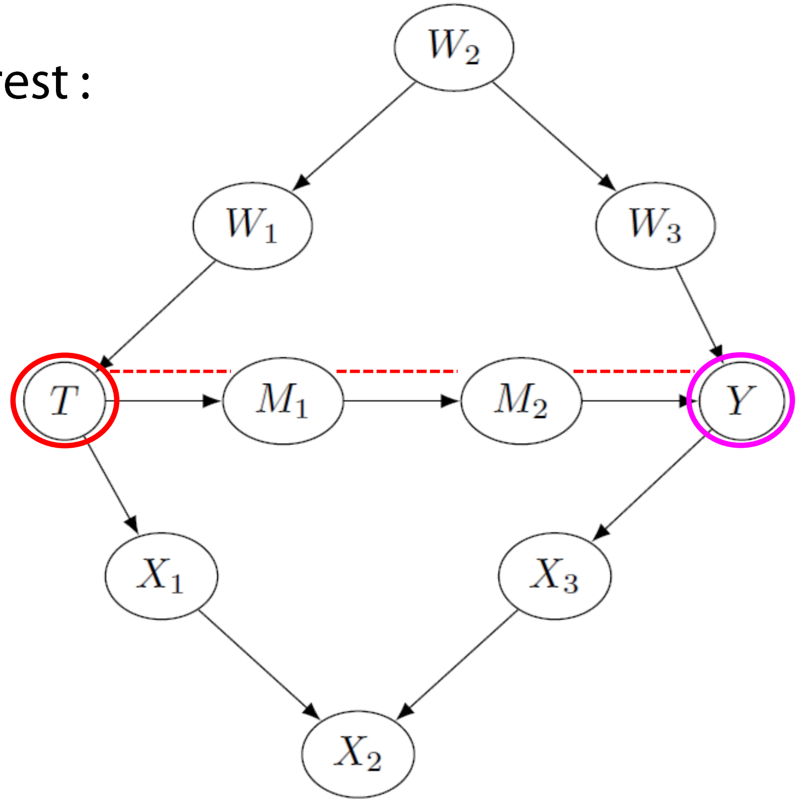
(*) *An earlier (and weaker) version of this is called 'back-door criterion'* [Pearl, 1993]

Identification

- **Identifiable Causal Effect**

  In this example, assuming that T over Y is the *causal effect* of interest :

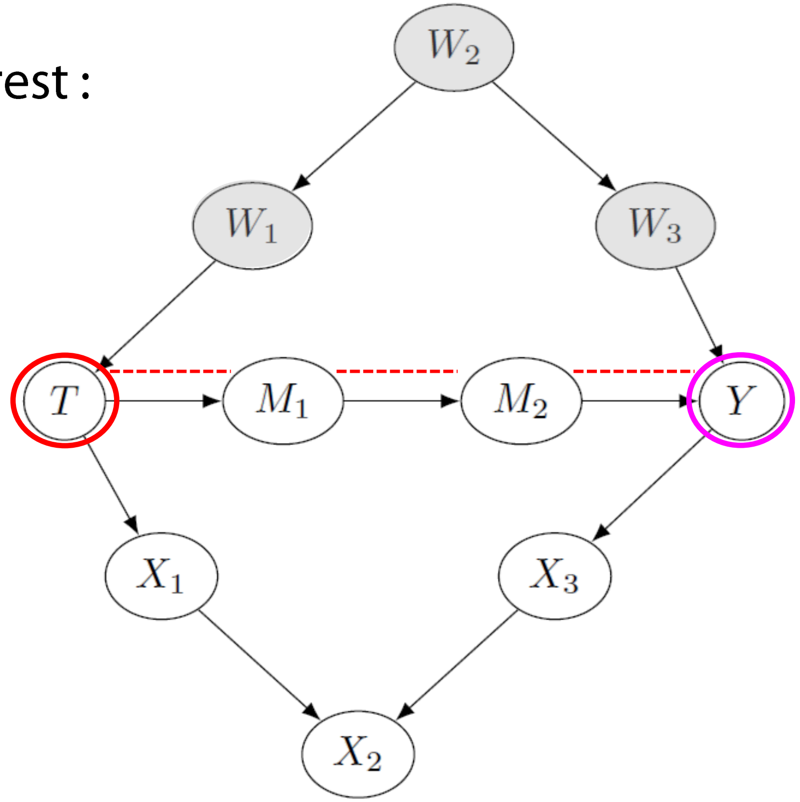
  1. The one in red is the *causal path*
     (there could be more than one)
  2. None of M_1 or M_2 should be in the adjustment set W



- **Identifiable Causal Effect**

  In this example, assuming that T over Y is the *causal effect* of interest :

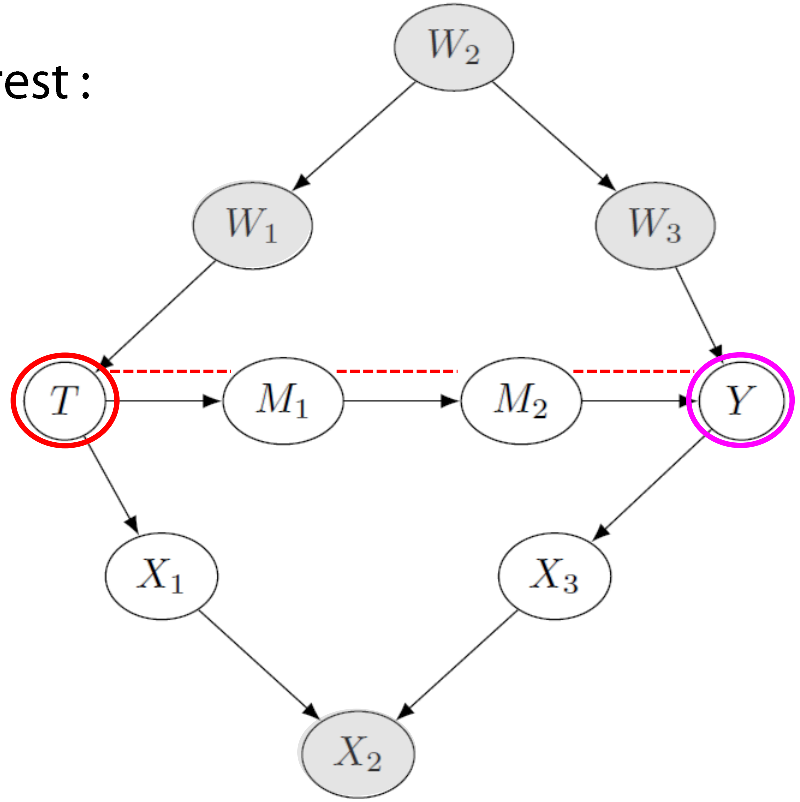
  1. The one in red is the *causal path* *(there could be more than one)*
  2. None of M_1 or M_2 should be in the adjustment set W
  3. Any non-empty subset of these three nodes is a valid *adjustment set* W



- **Identifiable Causal Effect**

In this example, assuming that T over Y is the *causal effect* of interest :

1. The one in red is the *causal path* *(there could be more than one)*
2. None of M_1 or M_2 should be in the adjustment set W
3. Any non-empty subset of these three nodes is a valid *adjustment set* W
4. Adding node X_2 makes it invalid

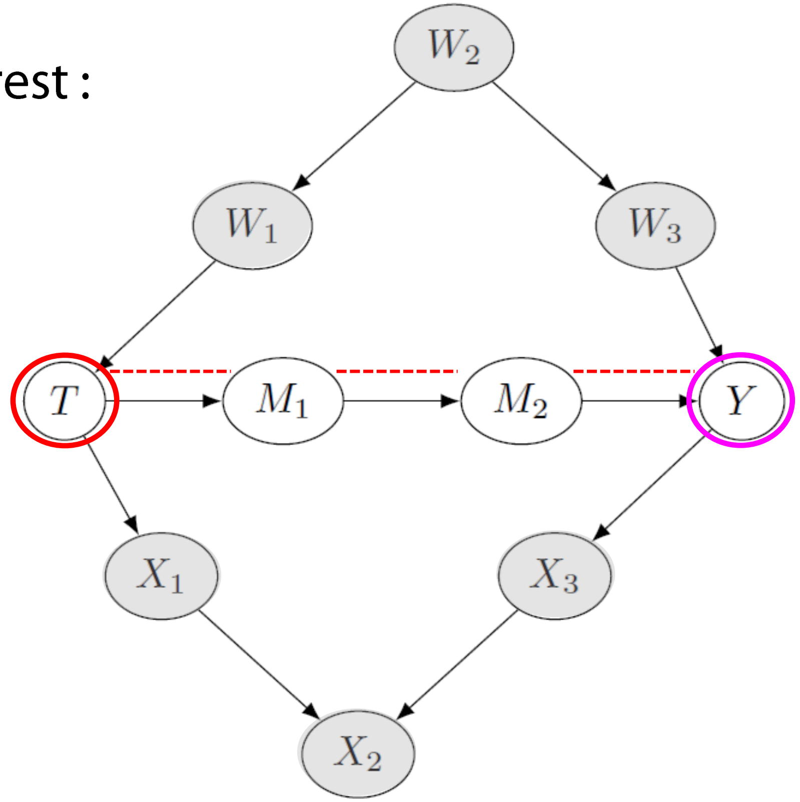


Identification

- **Identifiable Causal Effect**

In this example, assuming that T over Y is the *causal effect* of interest :

1. The one in red is the *causal path*
   (there could be more than one)
2. None of M_1 or M_2 should be in the adjustment set W
3. Any non-empty subset of these three nodes
   is a valid *adjustment set* W
4. Adding node X_2 makes it invalid
5. Adding any further blocking nodes makes W valid again



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

More in general, in practical cases,
there can be 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 W satisfies altogether 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:

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