CAUSAL GRAPHS WILL SAVE US ALL FROM BIG DATA

https://chield.excd.org

Seán Roberts
excd.lab, University of Bristol

Learning outcomes

What are causal graphs?

Why are they useful?

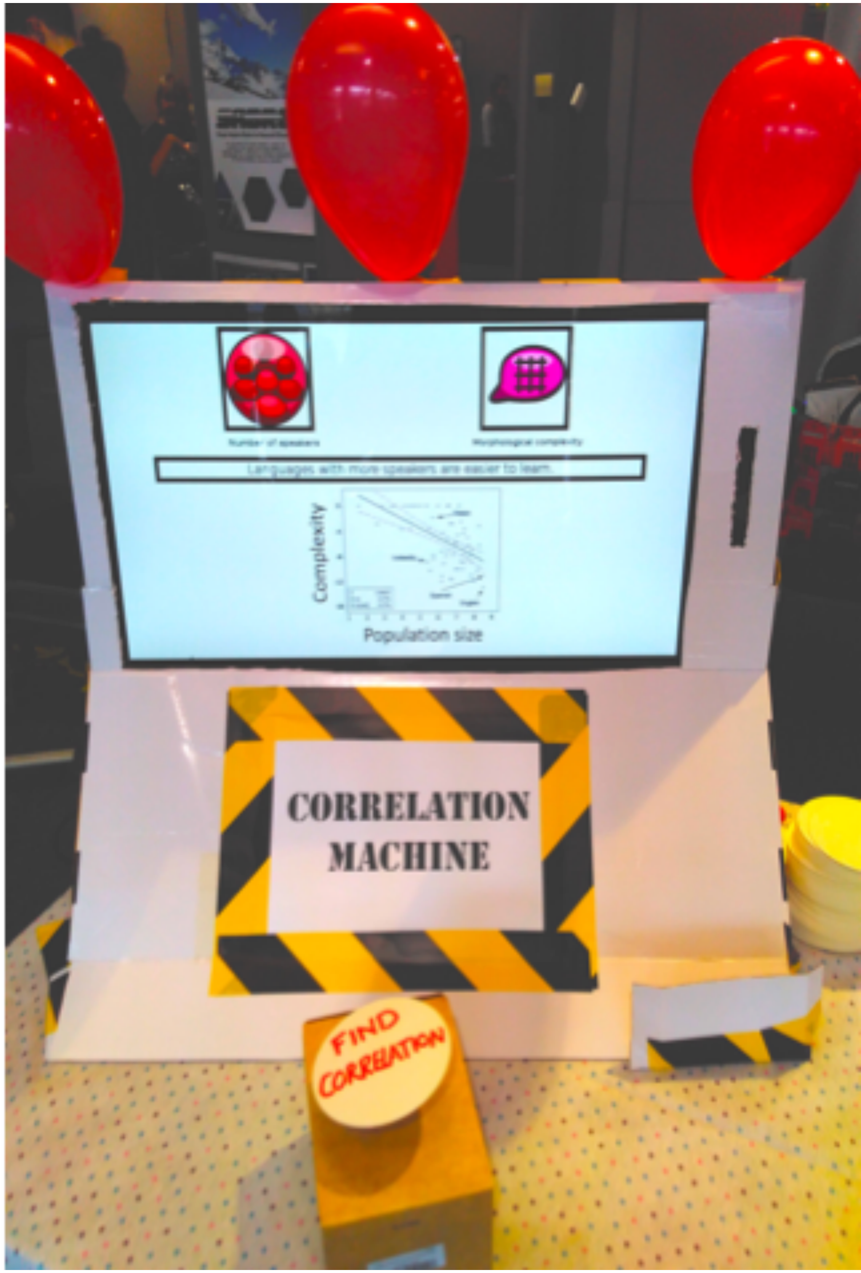
- Observation vs. Intervention
- Identify confounders
- Prevent spurious correlations
- Dealing with unobservable variables

How can causal graphs be applied in practice?

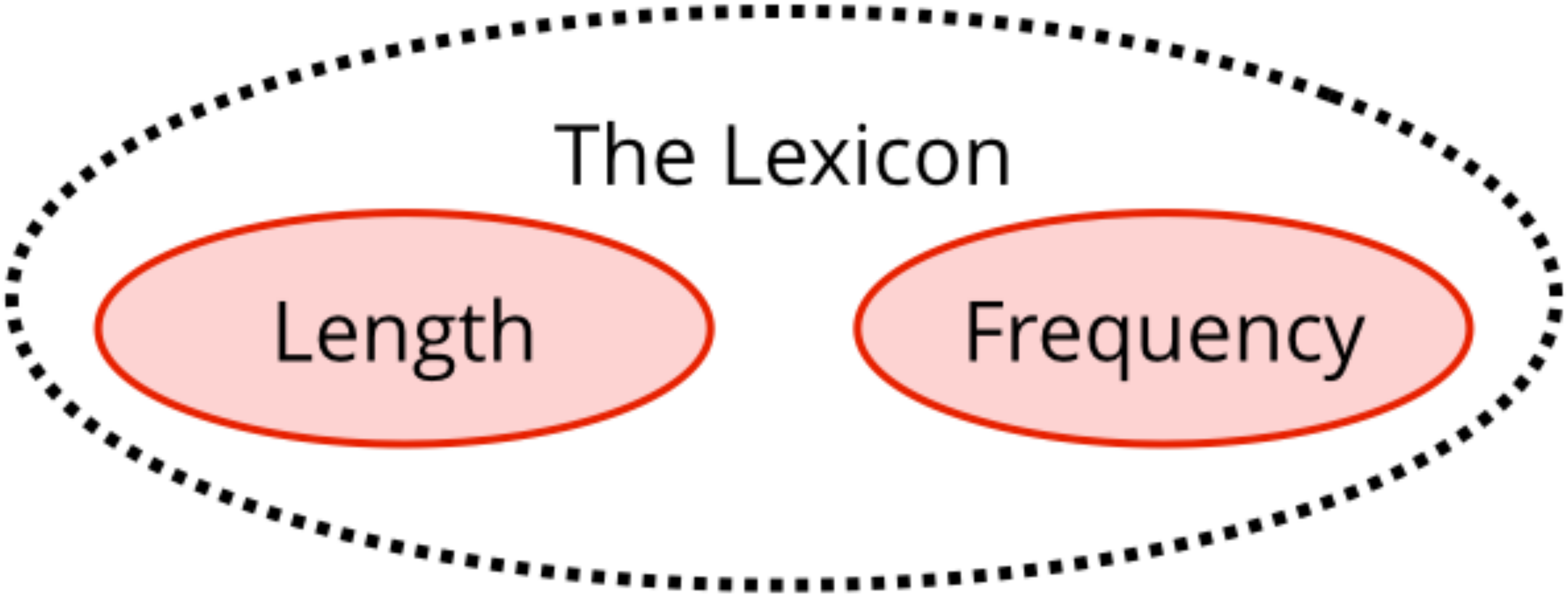
- Quick overview

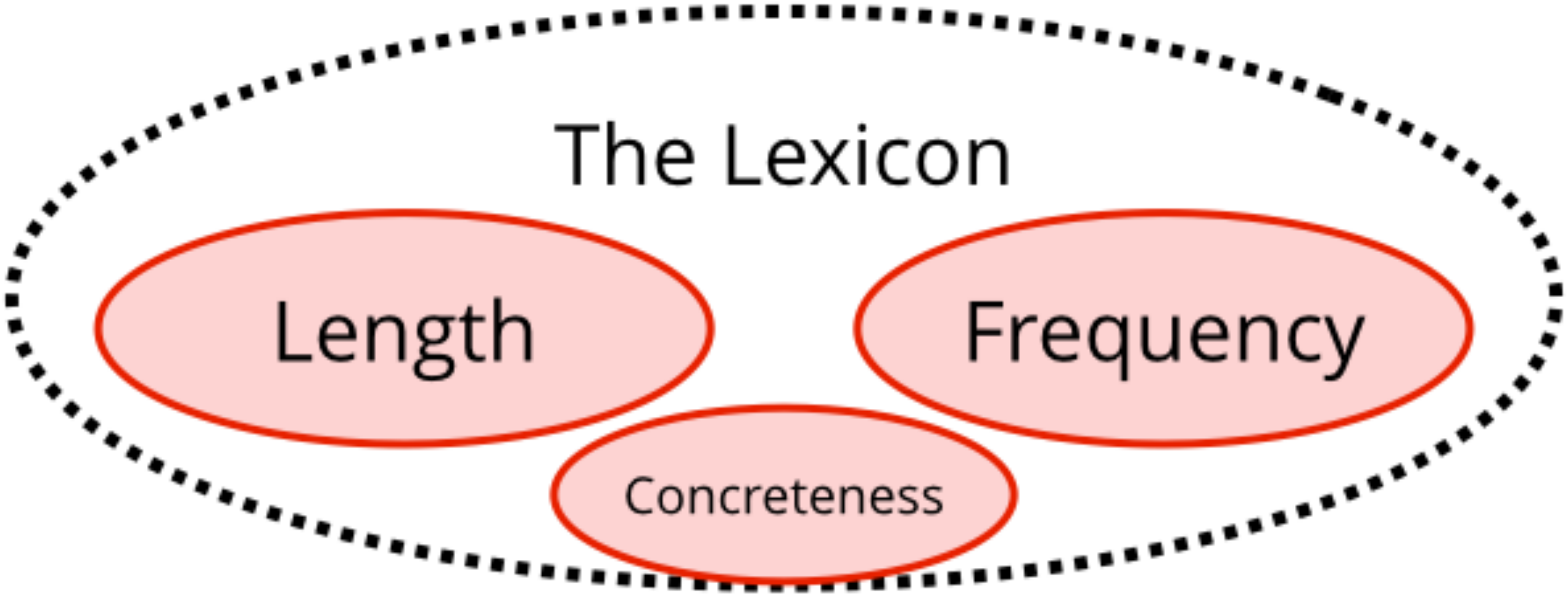
My Database of Causal Theories (CHIELD)

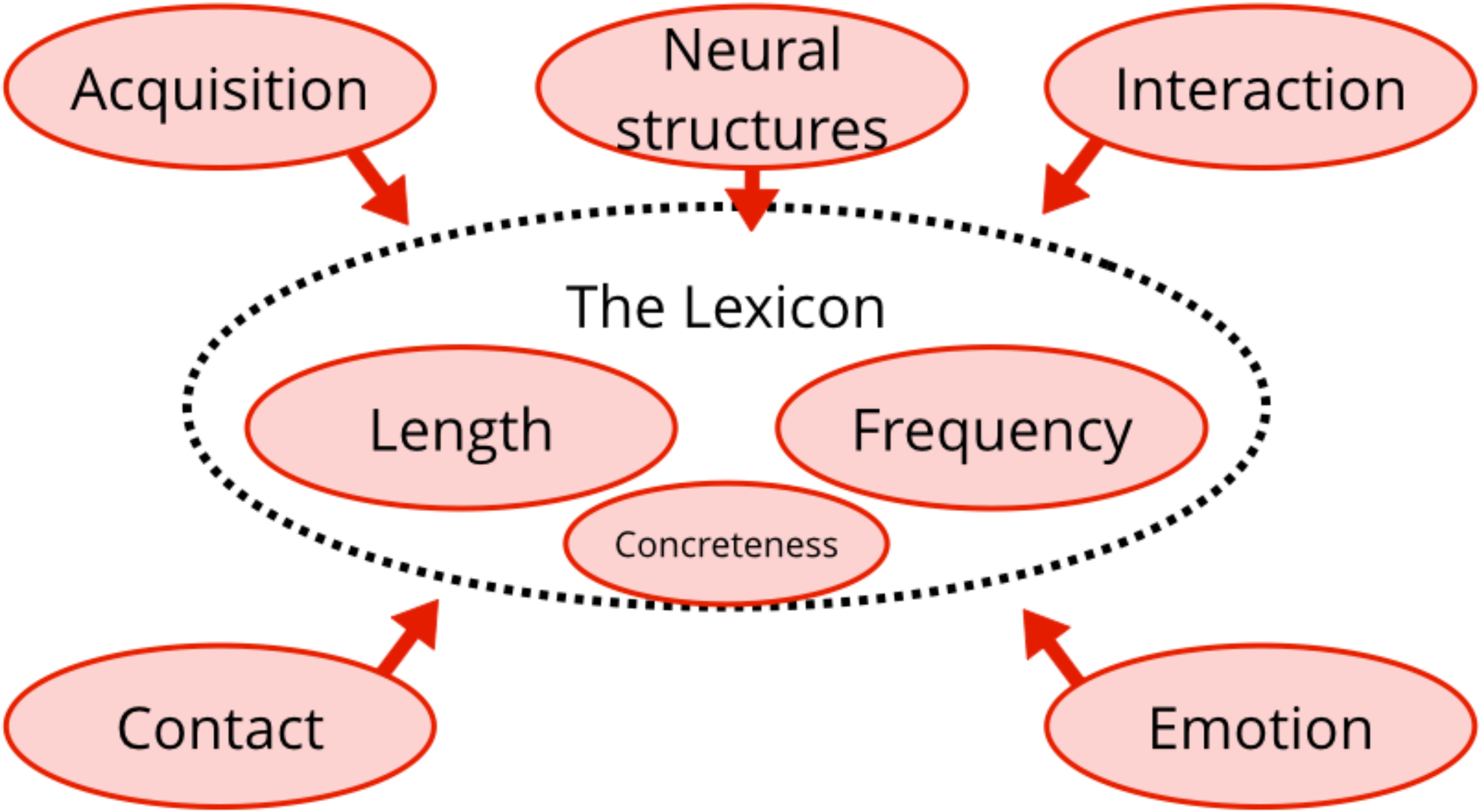
THE CORRELATION MACHINE

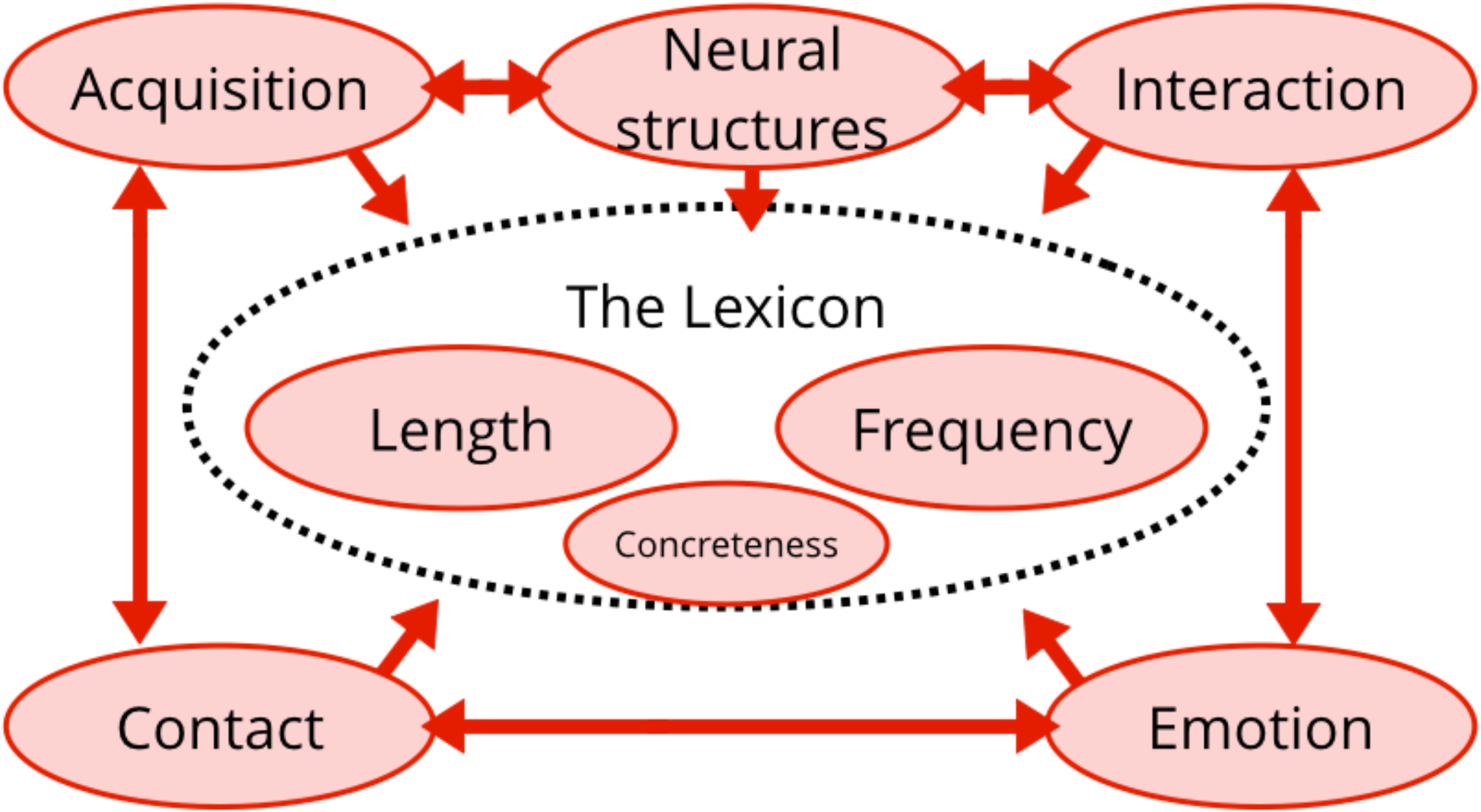



aeiou

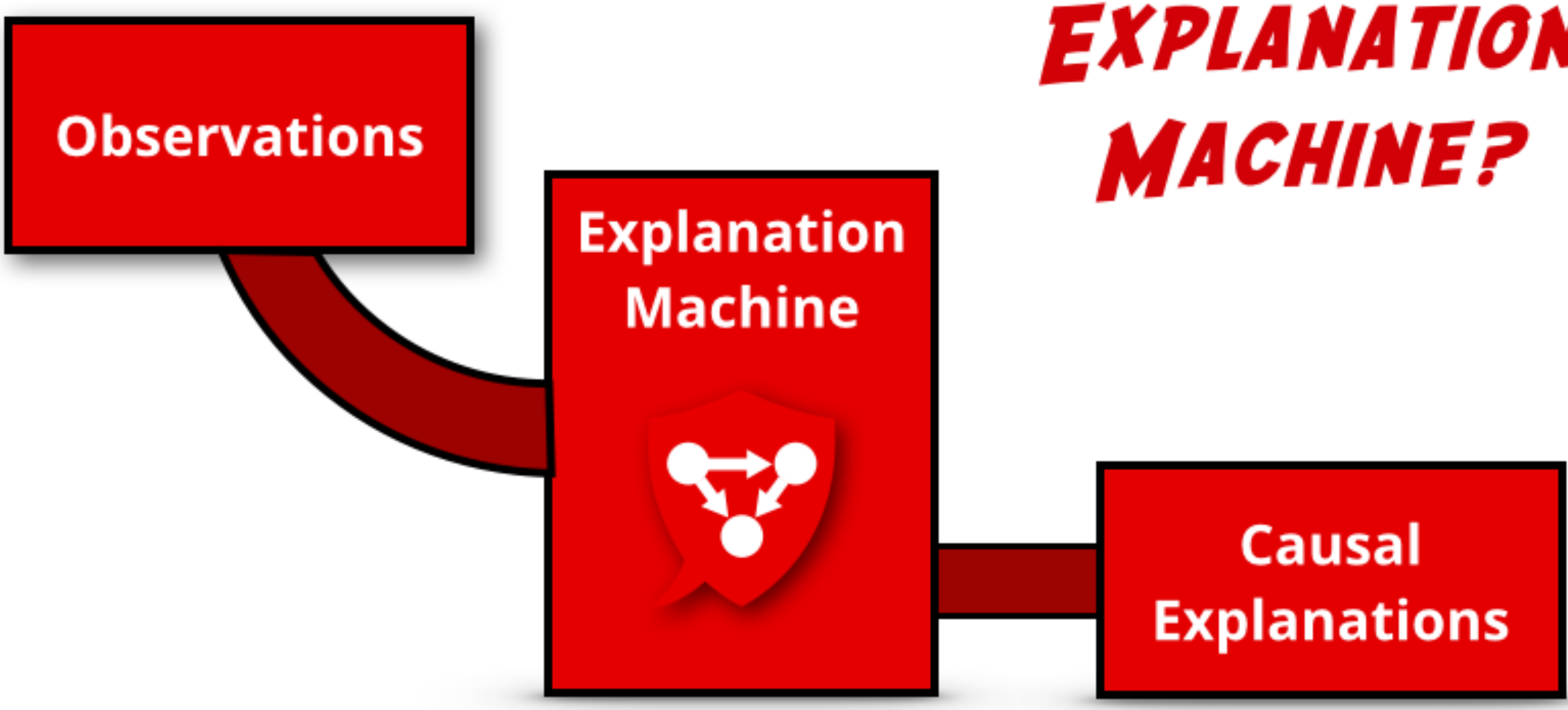

The Lexicon
Length
Frequency

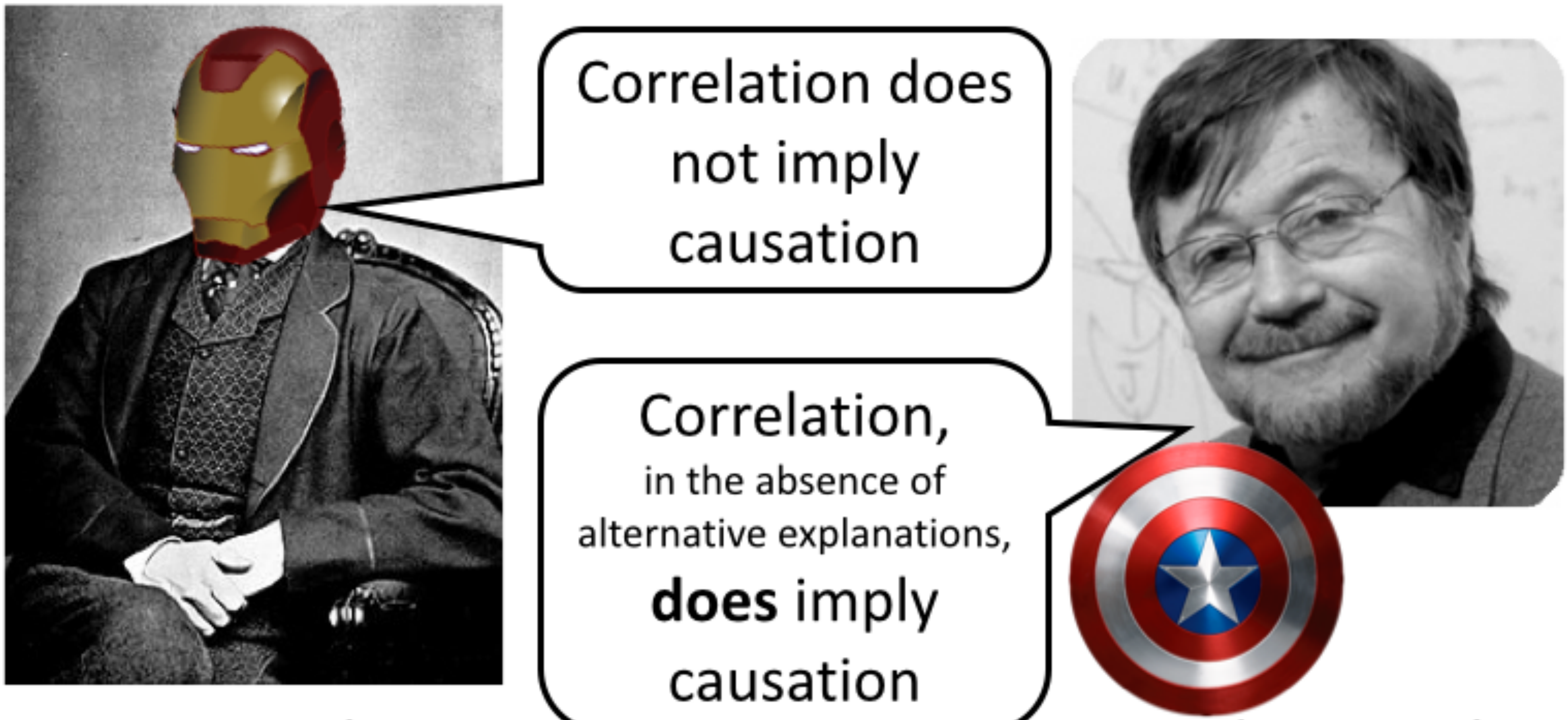

The Lexicon
Length
Frequency
Concreteness


Acquisition
Neural structures
Interaction
The Lexicon
Length
Frequency
Concreteness
Contact
Emotion


Acquisition
Neural structures
Interaction
The Lexicon
Length
Frequency
Concreteness
Contact
Emotion

THE EXPLANATION MACHINE?





Francis Galton

Judea Pearl

Observation vs. Intervention

Observation ("seeing")

Can look up in a probability table

P(stain | icecream)

= 125/(125+25) = 83%

Beach survey	*T-shirt stains*	*No T-shirt*
Ice cream	*125*	*25*
No Ice cream	*73*	*77*

Intervention ("doing")

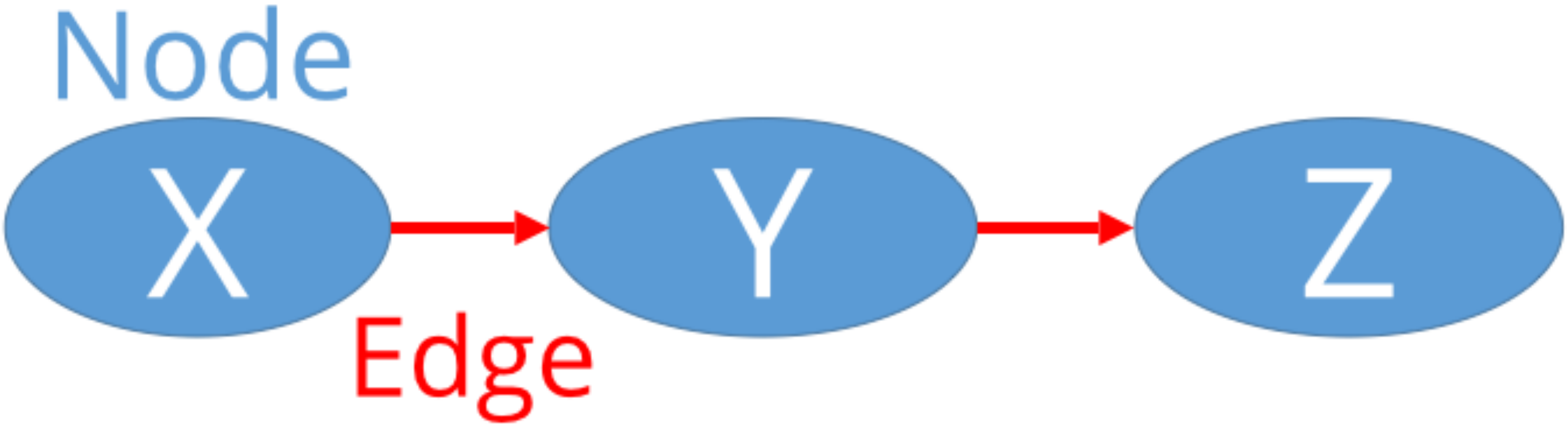
What is the probability of a stain if we force someone to have an icecream?

P(stain | *do*(icecream))

Randomised control experiment

What are causal graphs?

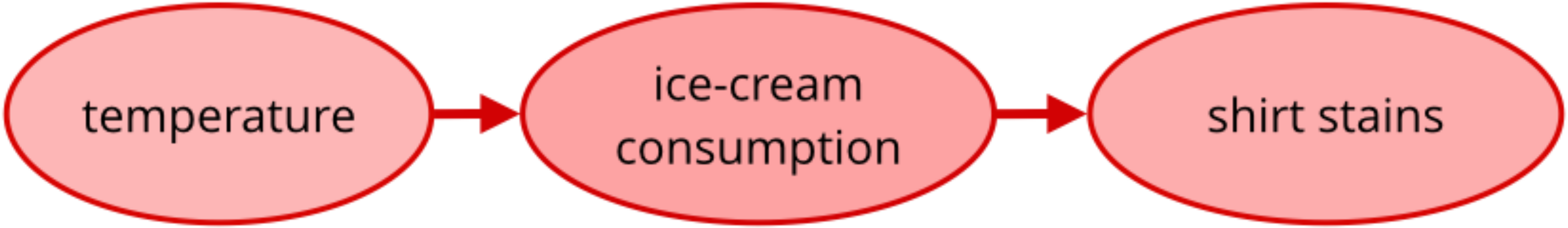
Causal graphs are a mathematical language which helps us think clearly



If we manipulated X in a suitable way, Y would change
Could be absolute, statistical, categorical ...

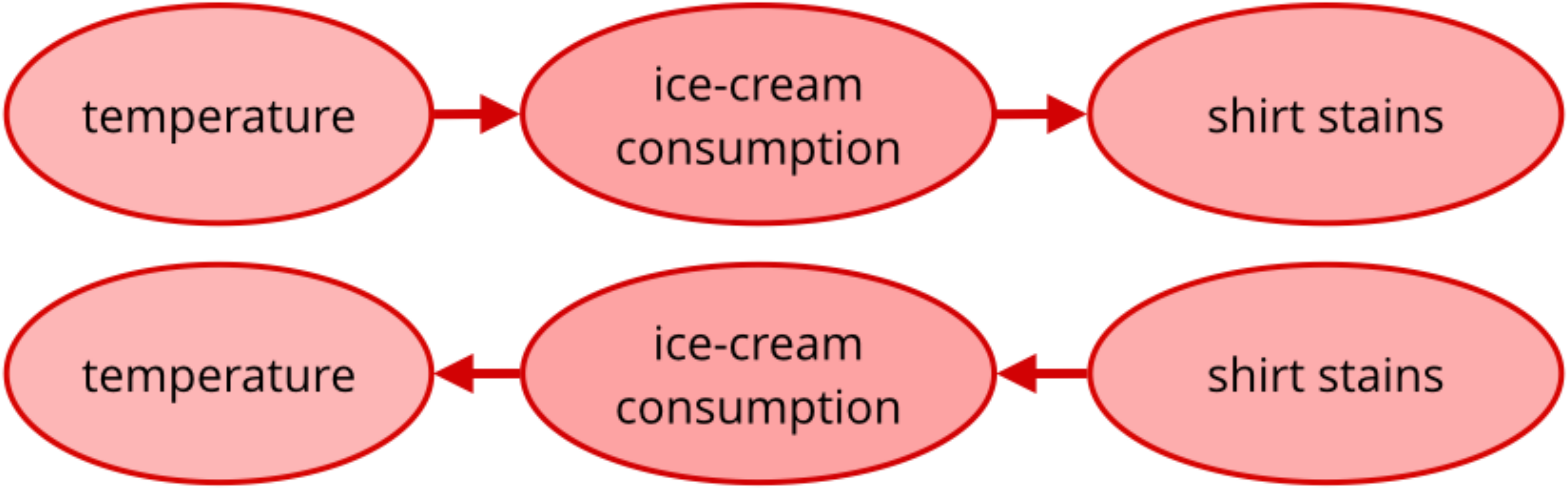
Causal graphs are mathematical expressions which help us think clearly about causality

"We hypothesise that high temperatures cause more ice-cream consumption, and more ice-cream consumption leads to more shirt stains."

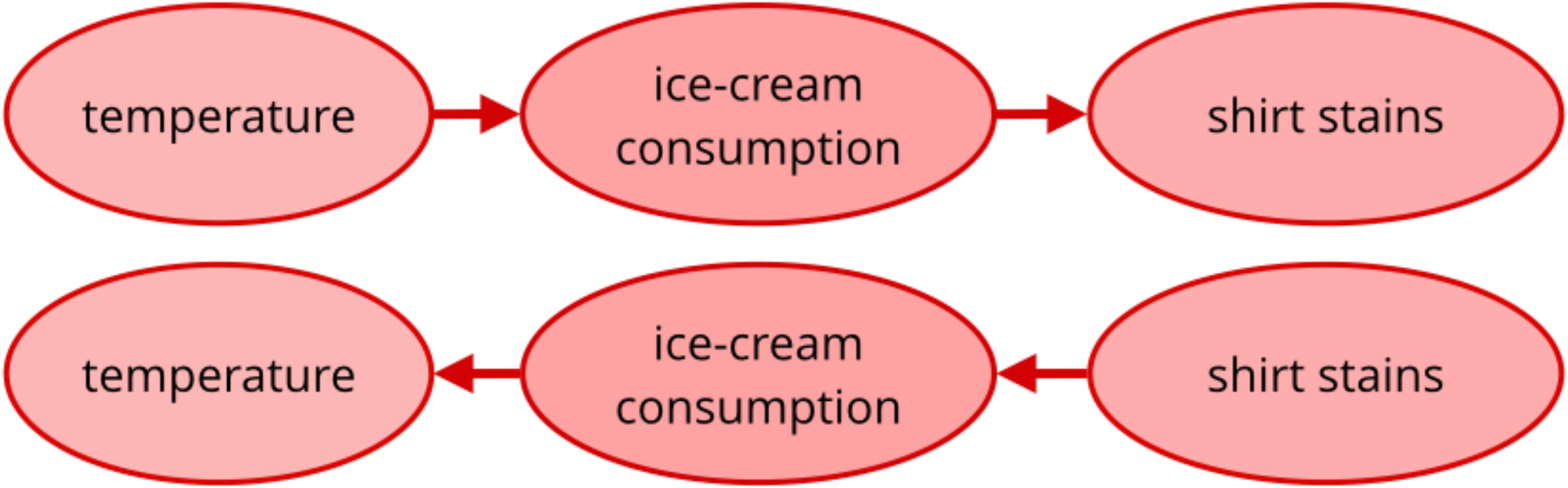


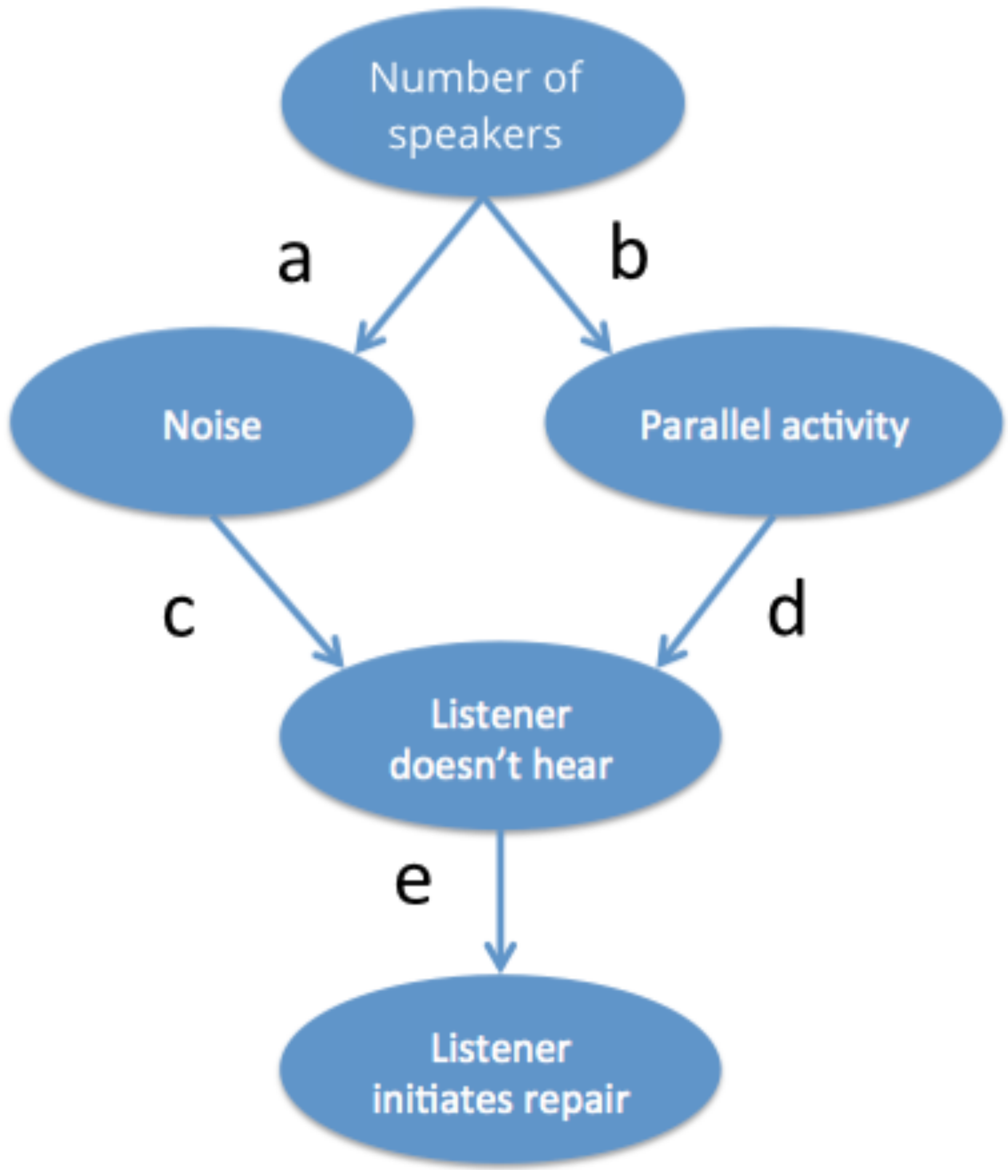
If we manipulated temperature in a suitable way, ice-cream consumption would change.
Could be absolute, statistical, categorical ...

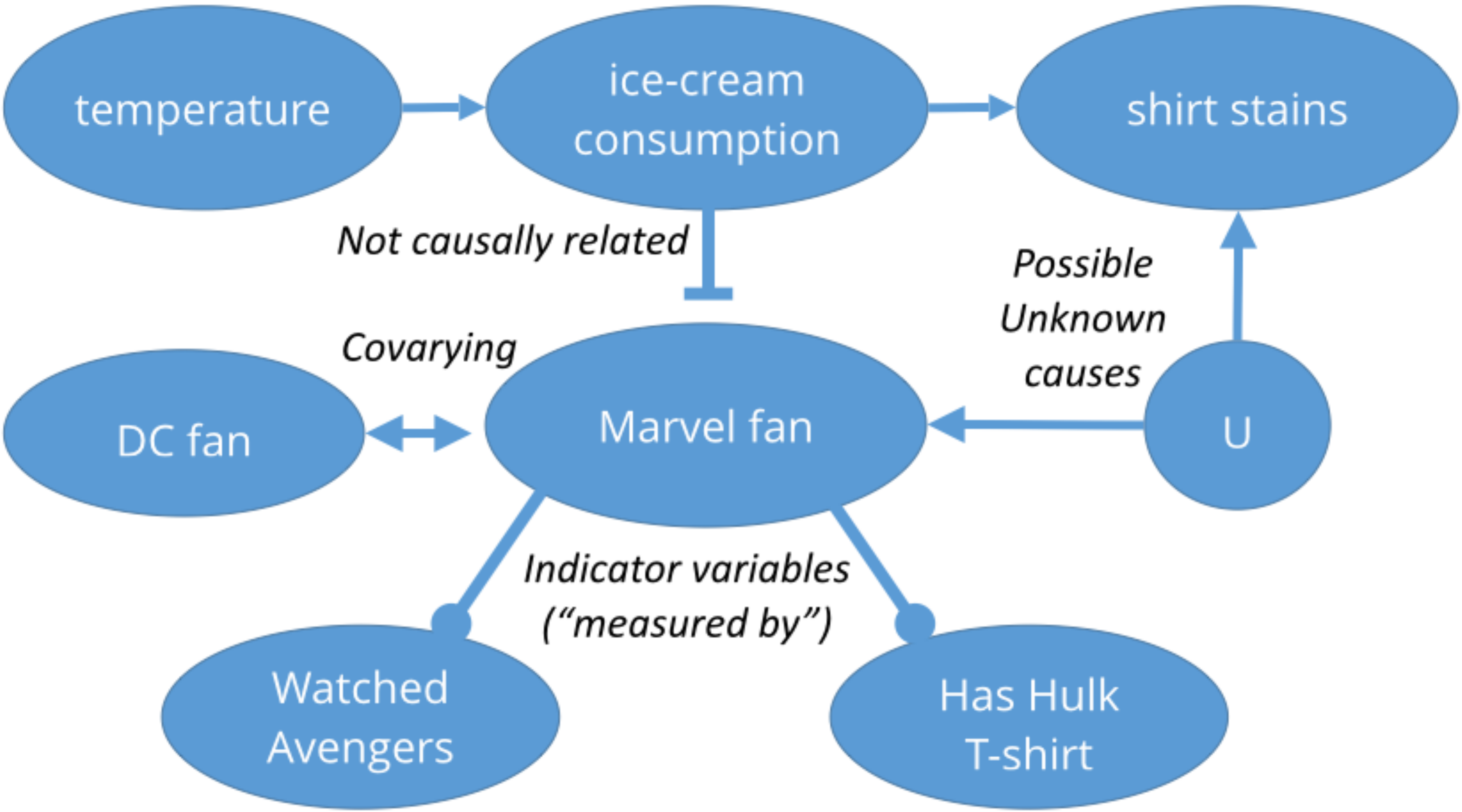
Causal graphs are mathematical expressions which help us think clearly about causality



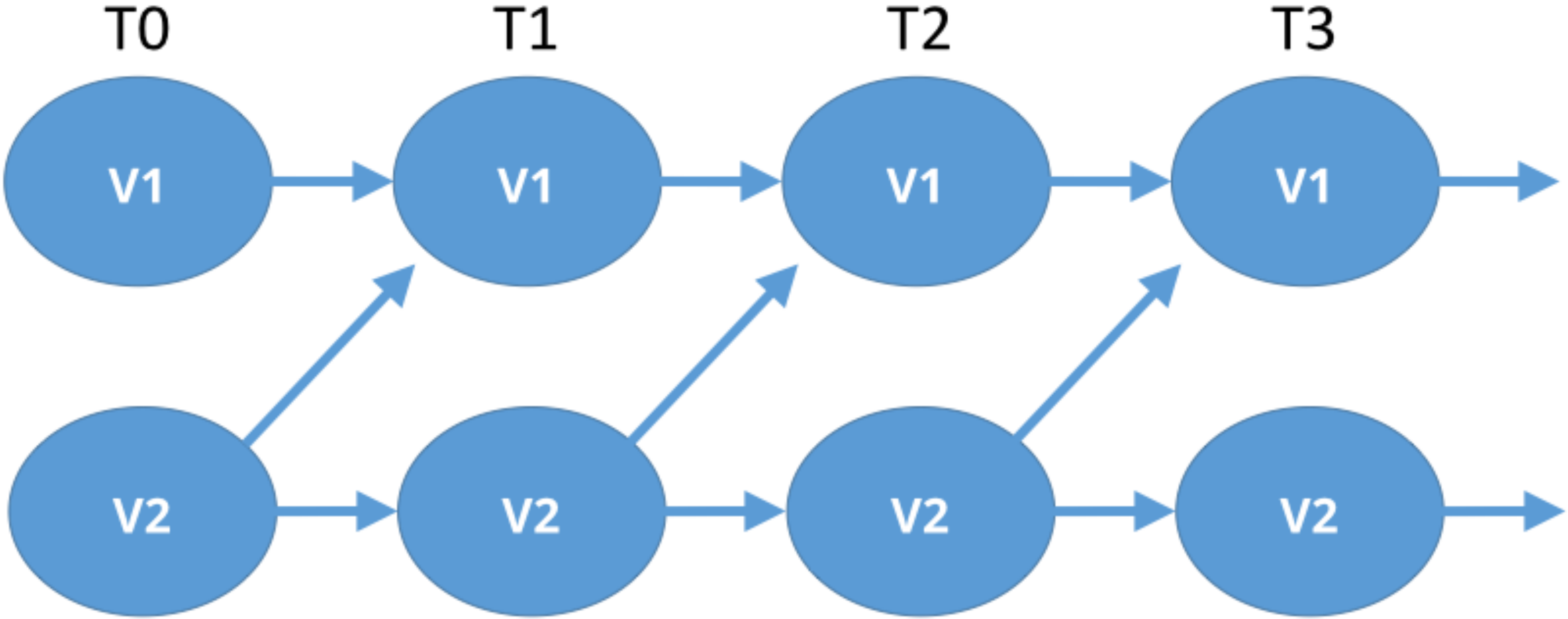
Causal graphs are mathematical expressions which help us think clearly about causality




Number of speakers
a
b
Noise
Parallel activity
c
d
Listener doesn't hear
e
Listener initiates repair


temperature
ice-cream consumption
shirt stains
Not causally related
Possible Unknown causes
Covarying
DC fan
Marvel fan
U
Indicator variables ("measured by")
Watched Avengers
Has Hulk T-shirt

Time



Why use causal graphs?

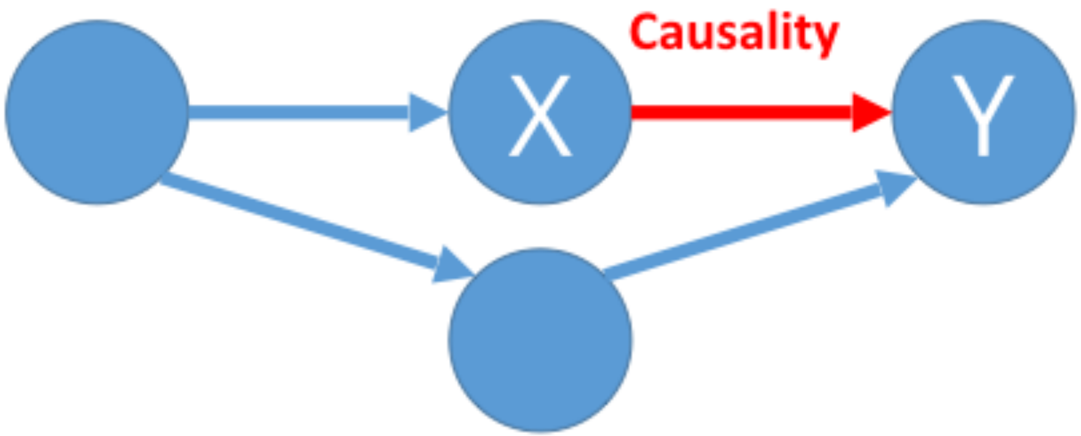
Look at intervention, not just observation

Confounding: Determine which variables to control for

Spurious correlations: Deal with colliders

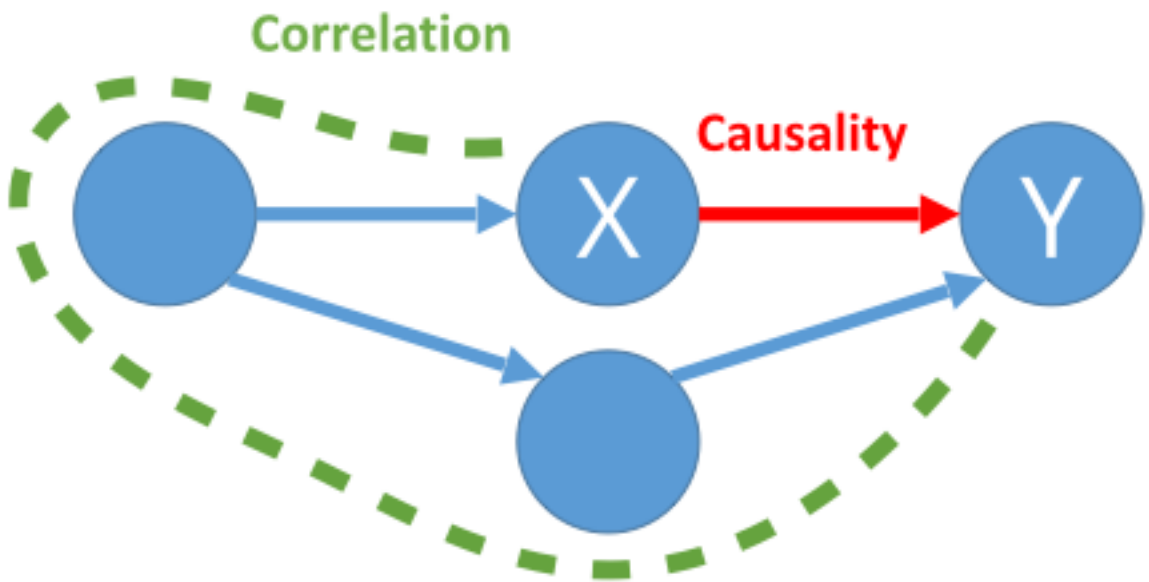
Confounding

Causality follows the arrows
But correlation (*noncausal* information) flows both ways



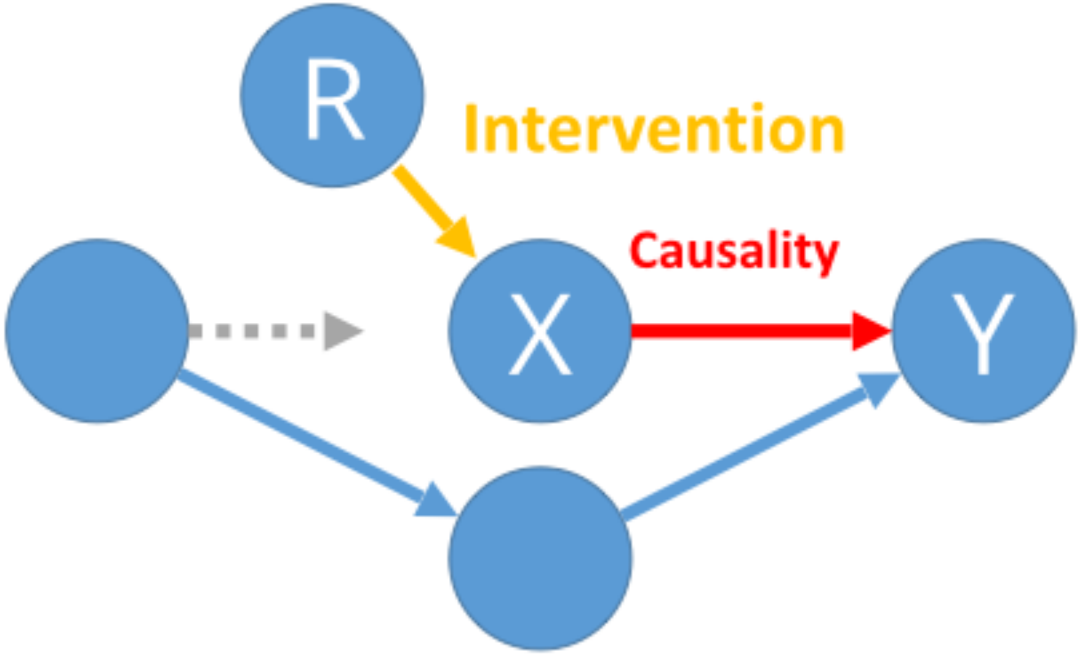
Confounding

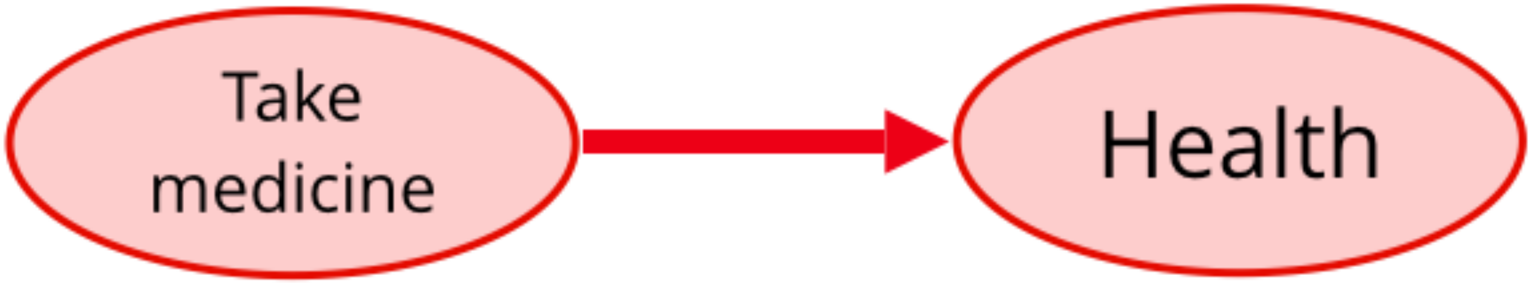
Causality follows the arrows
But correlation (*noncausal* information) flows both ways

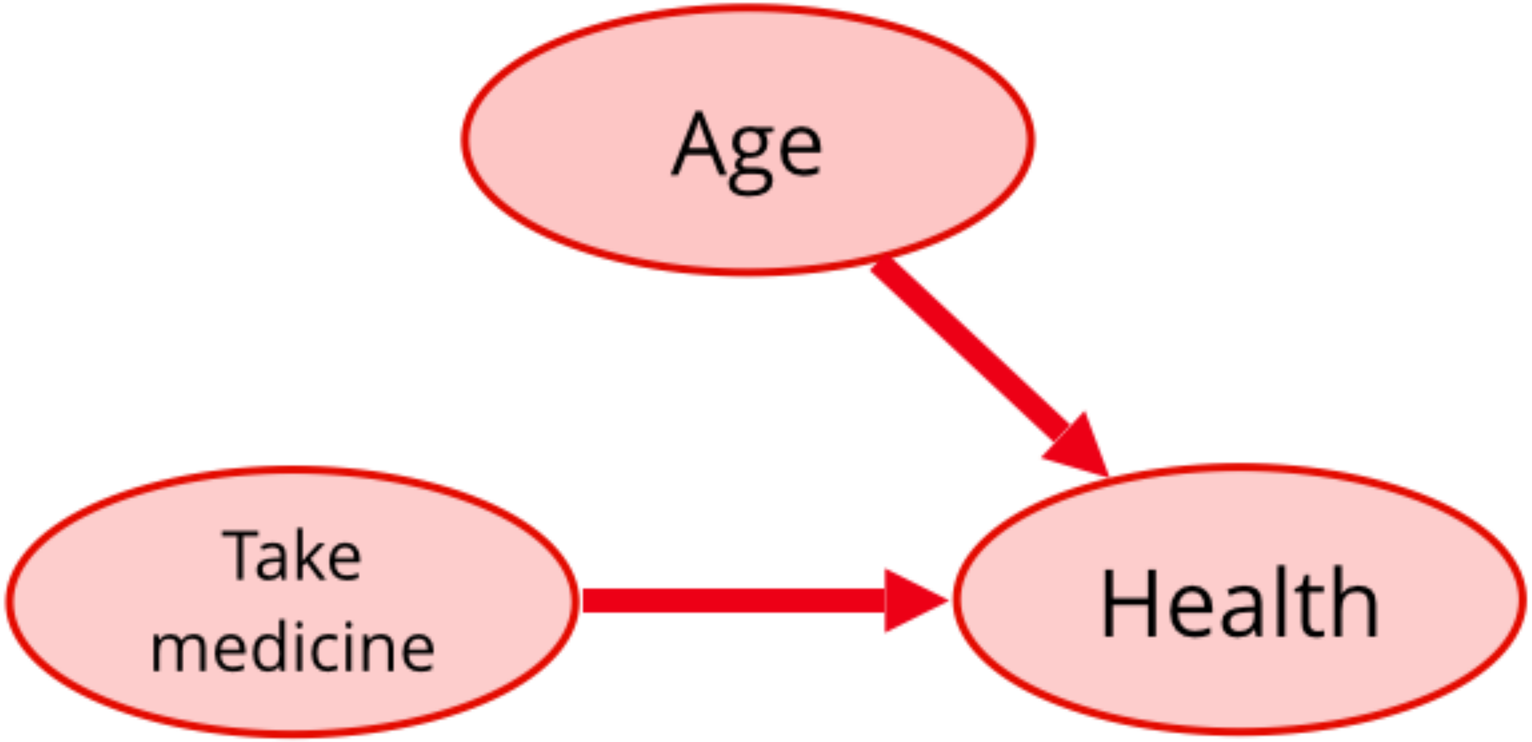


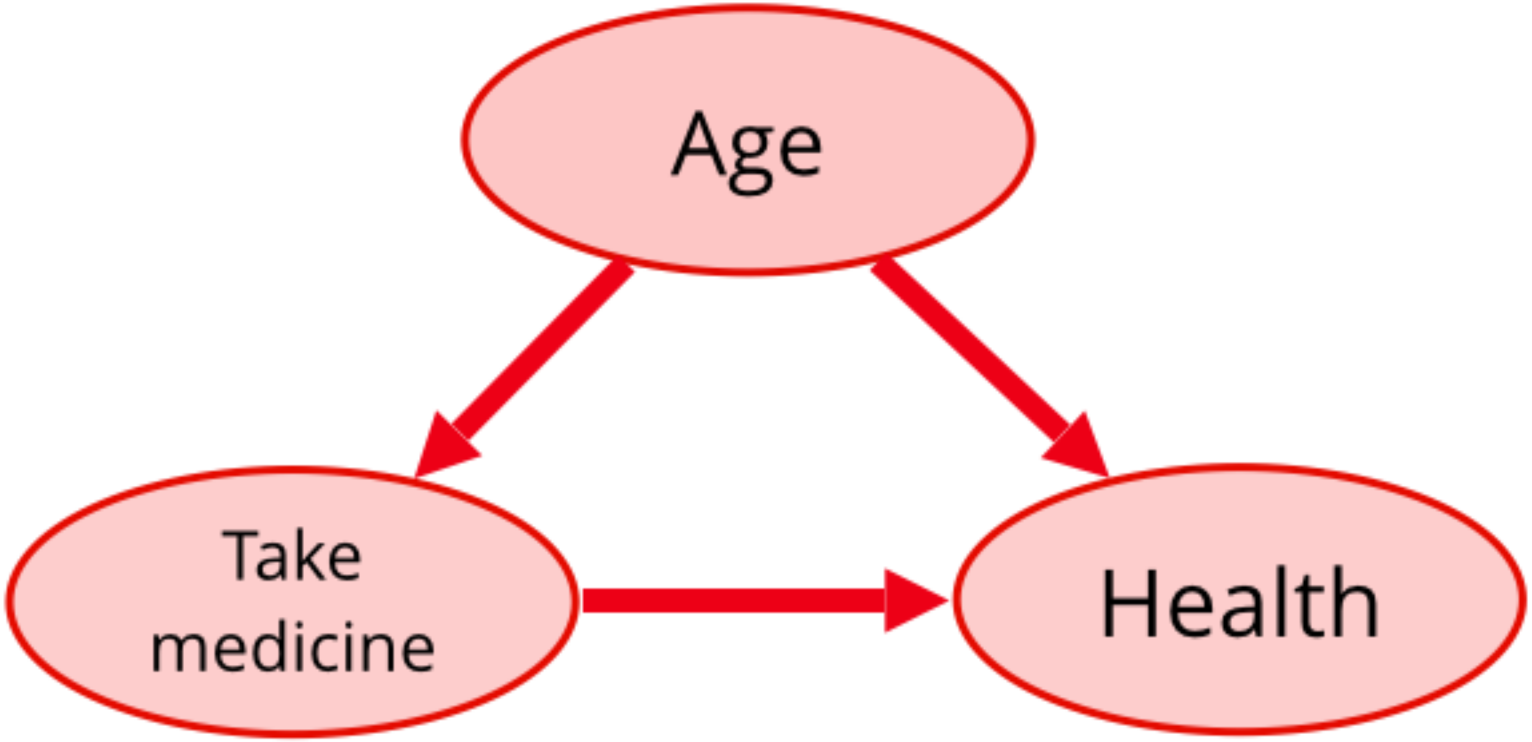
Confounding

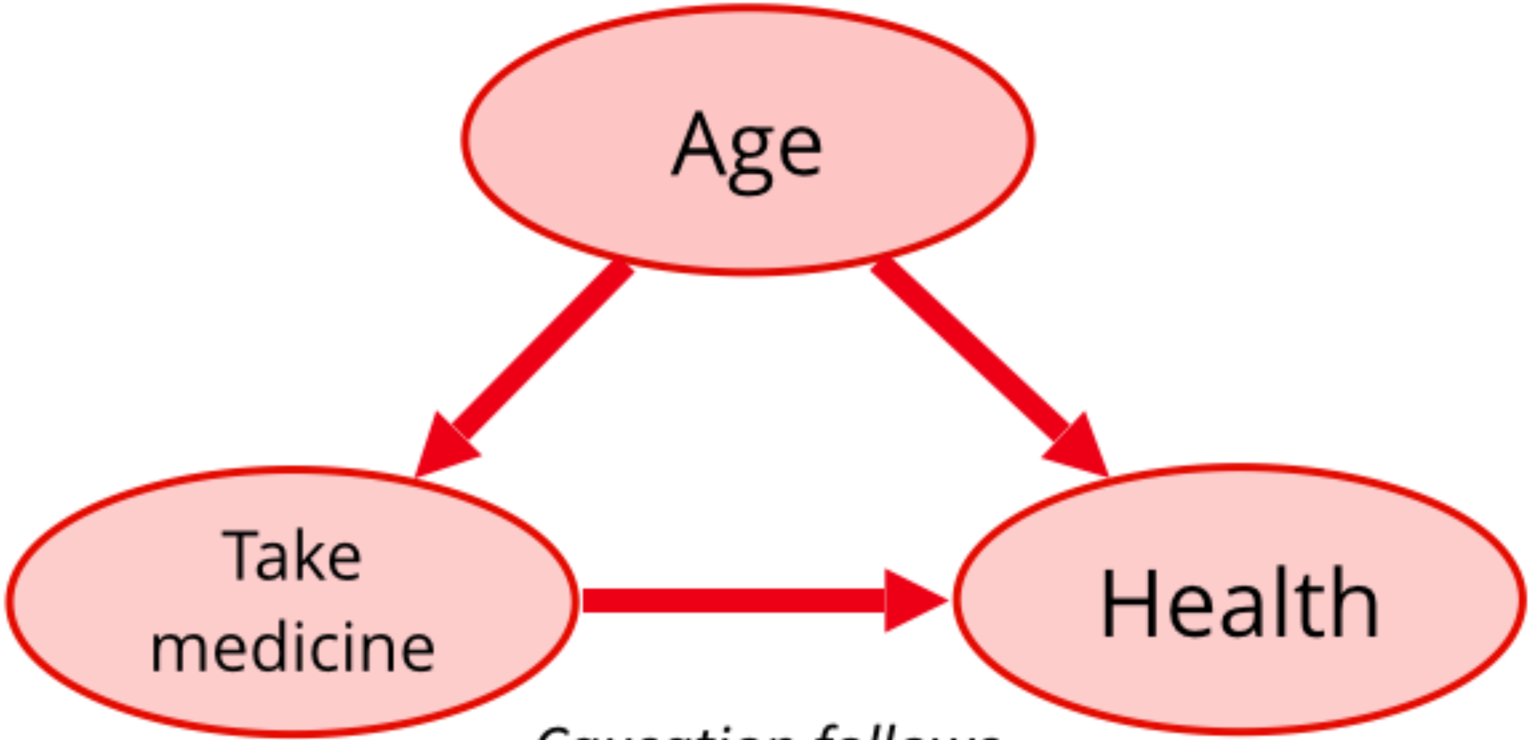
Causality follows the arrows
But correlation (*noncausal* information) flows both ways
Intervention breaks the flow of correlations



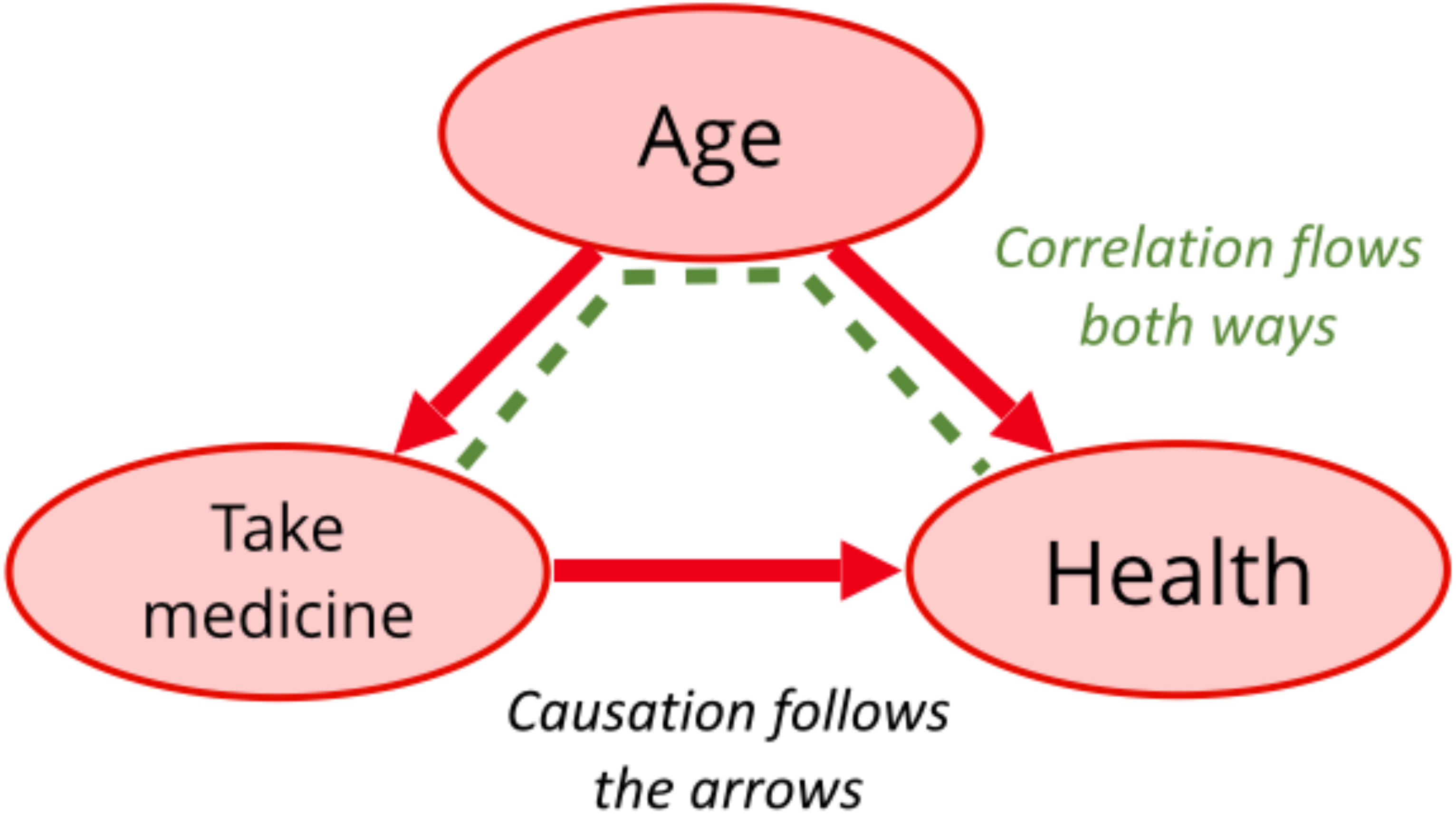

Take
medicine
Health


Age
Take
medicine
Health

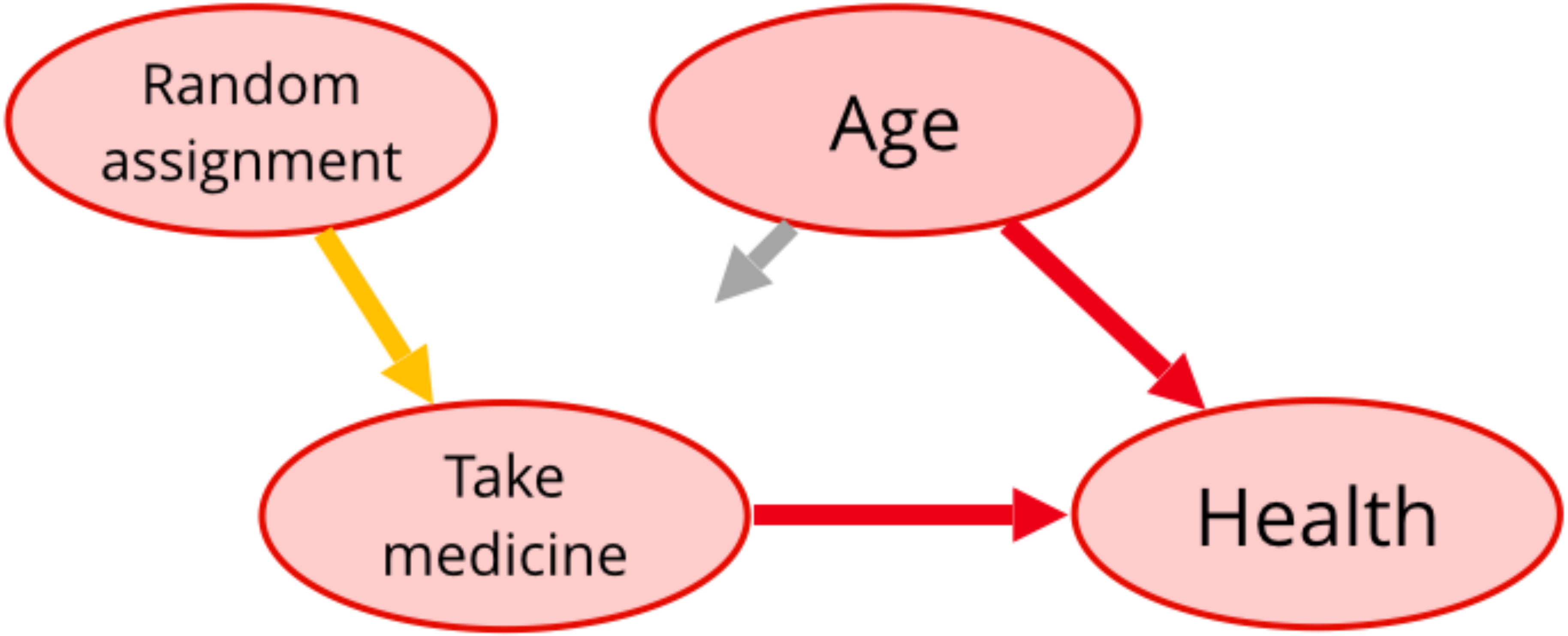

Age
Take
medicine
Health



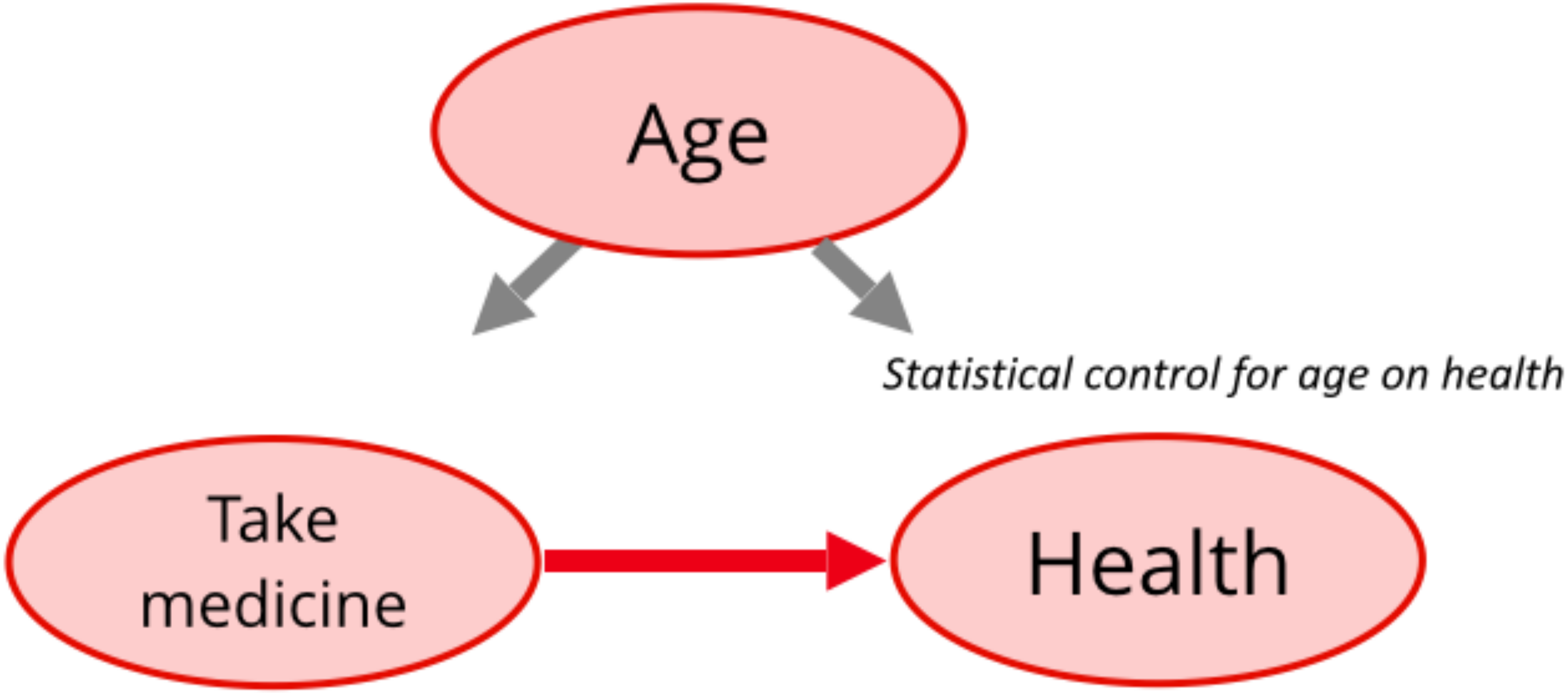
Causation follows the arrows


Age
Correlation flows
both ways
Take
medicine
Health
Causation follows
the arrows

Randomised control experiment



Statistical control

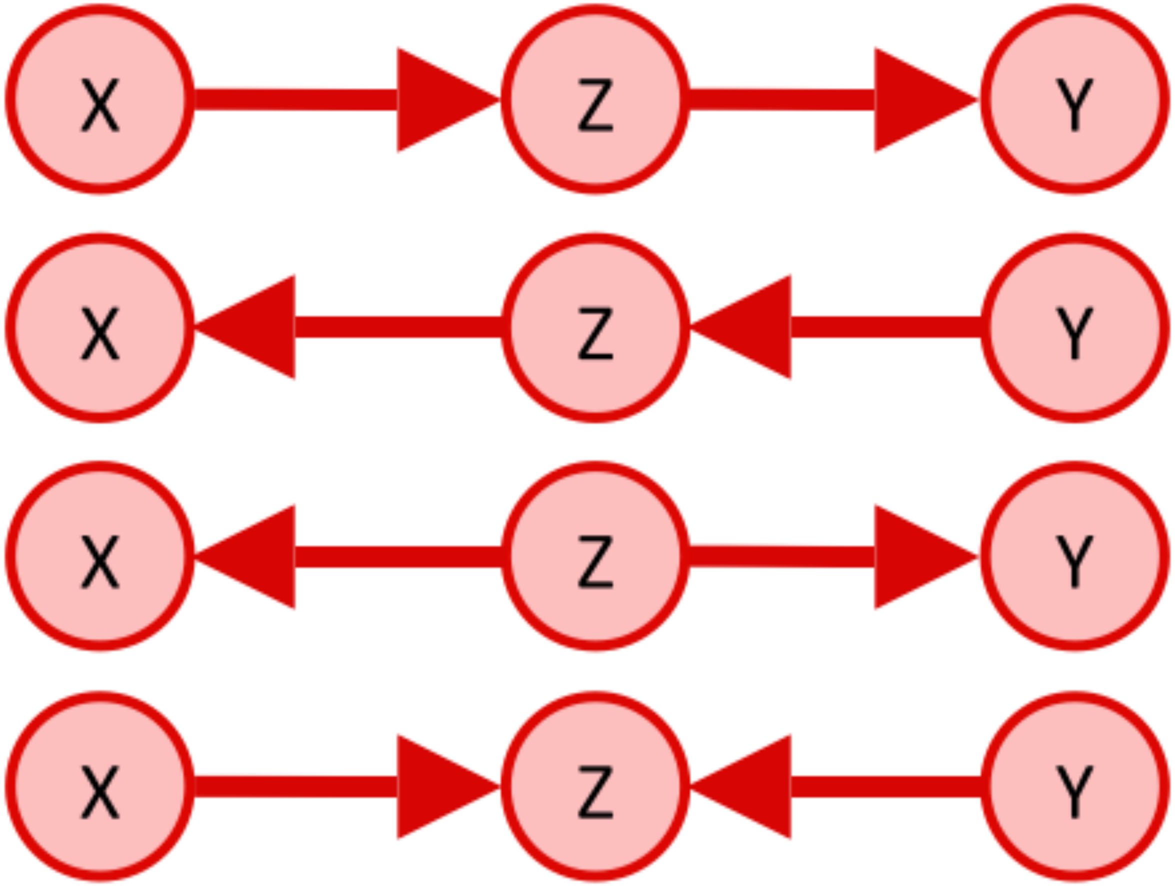


X is correlated with Y?

Yes

Yes

X is correlated with Y?

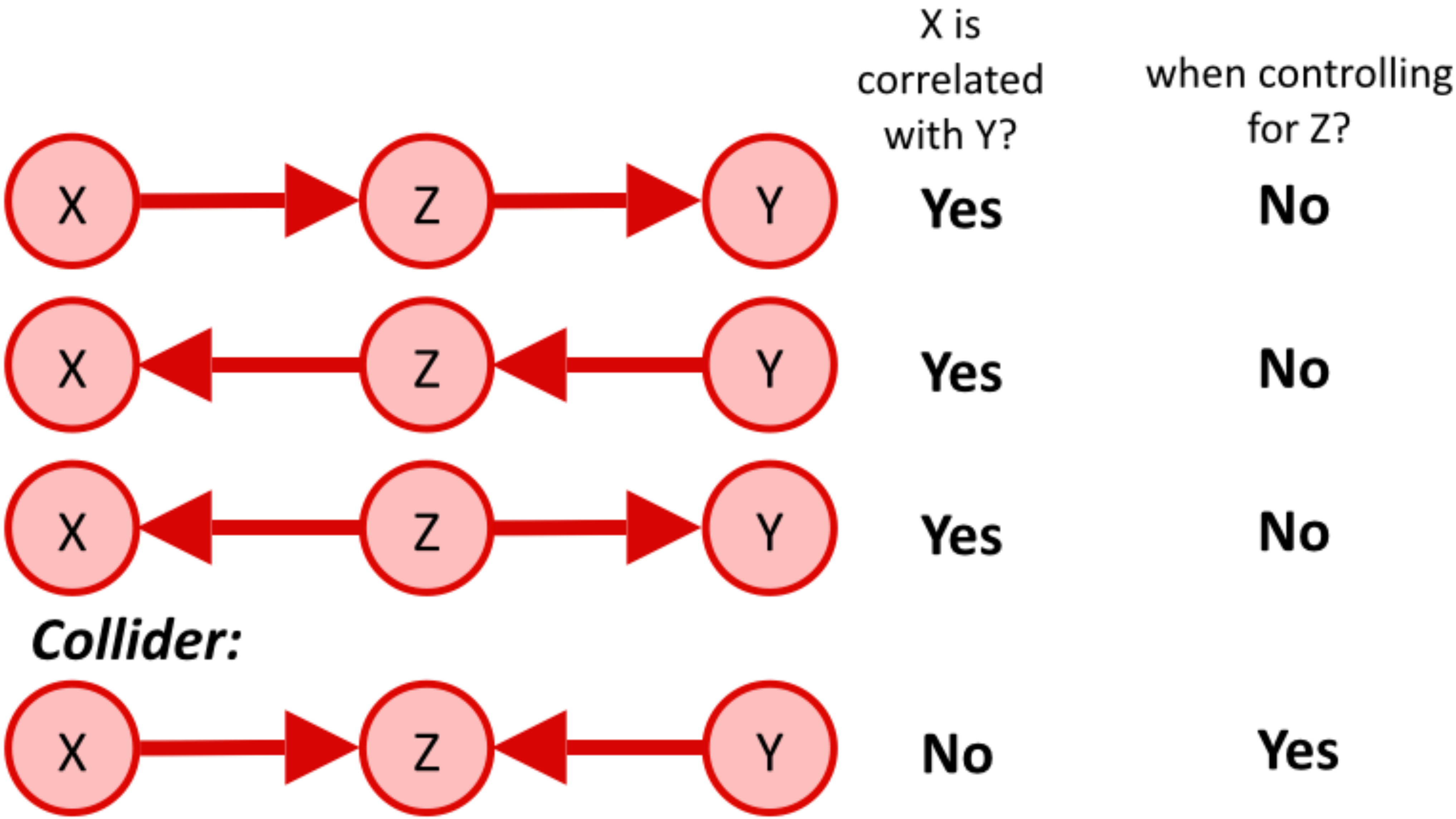


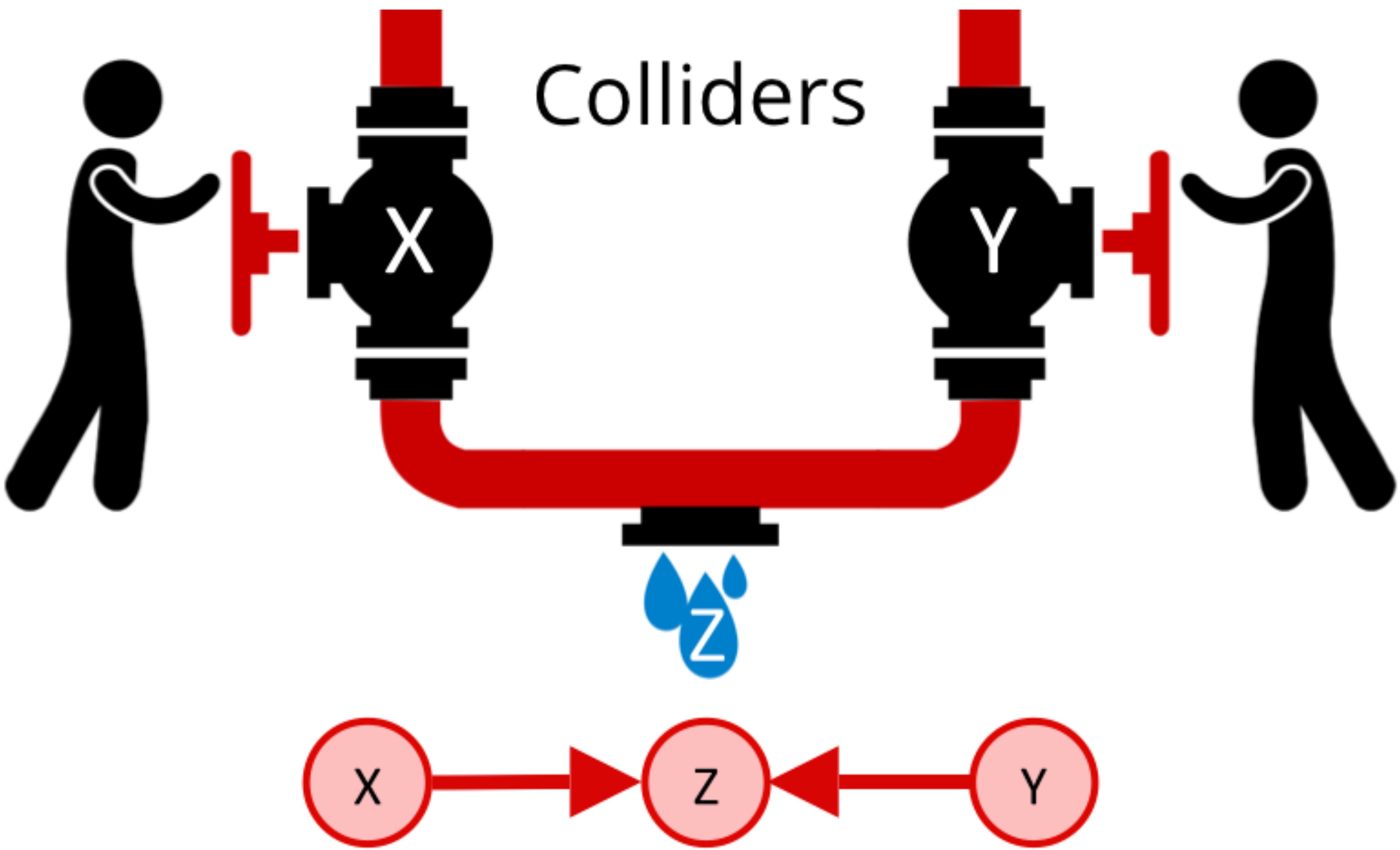
Yes

Yes

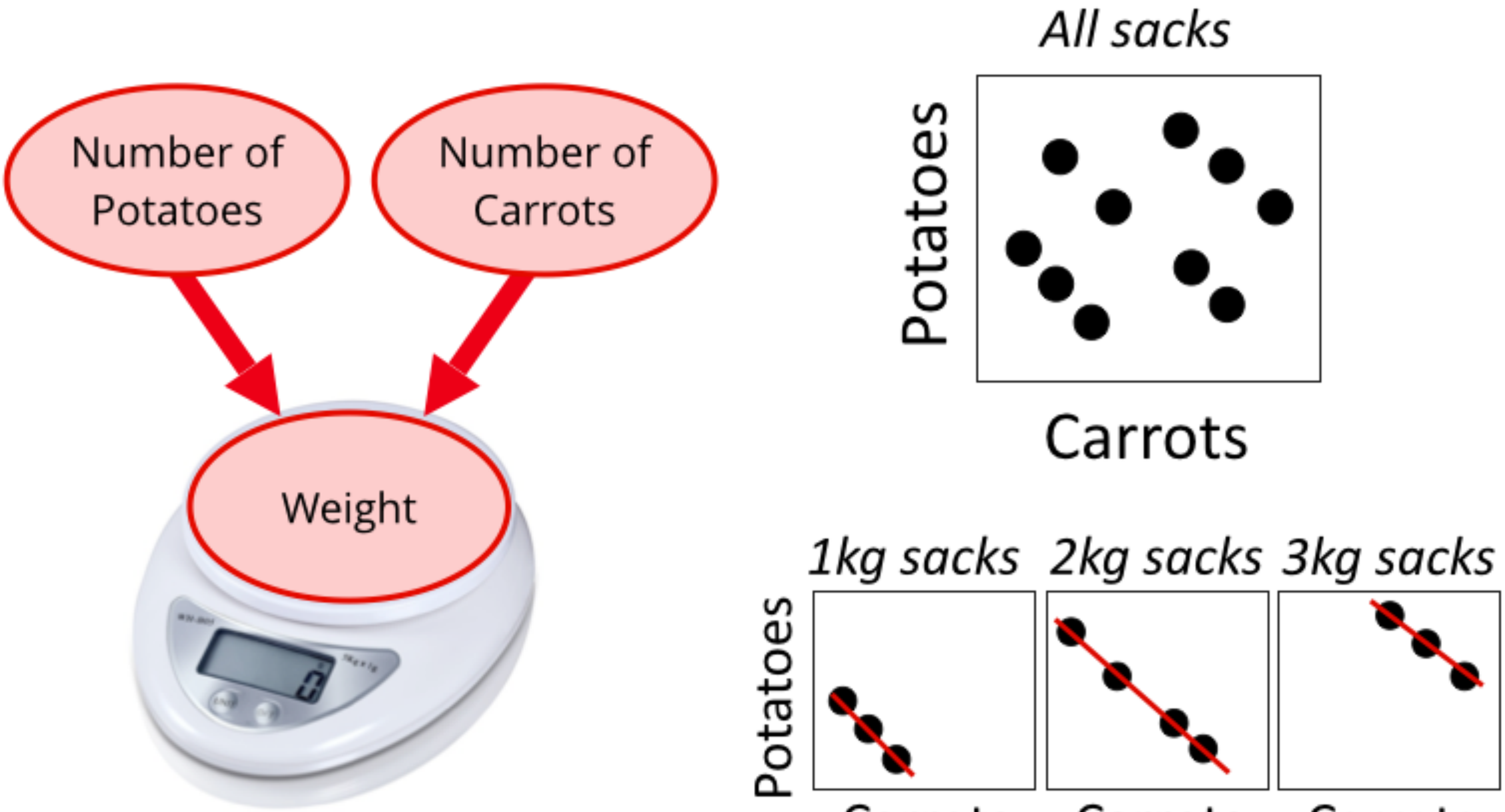
Yes

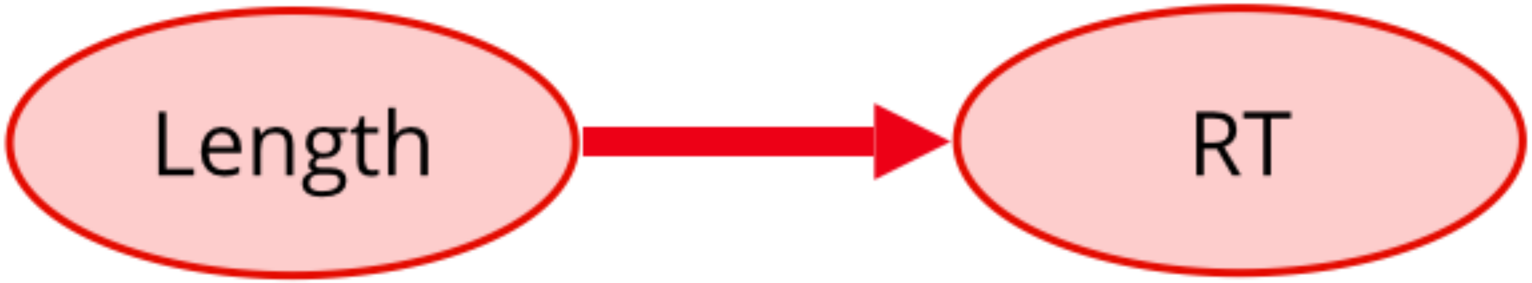
No

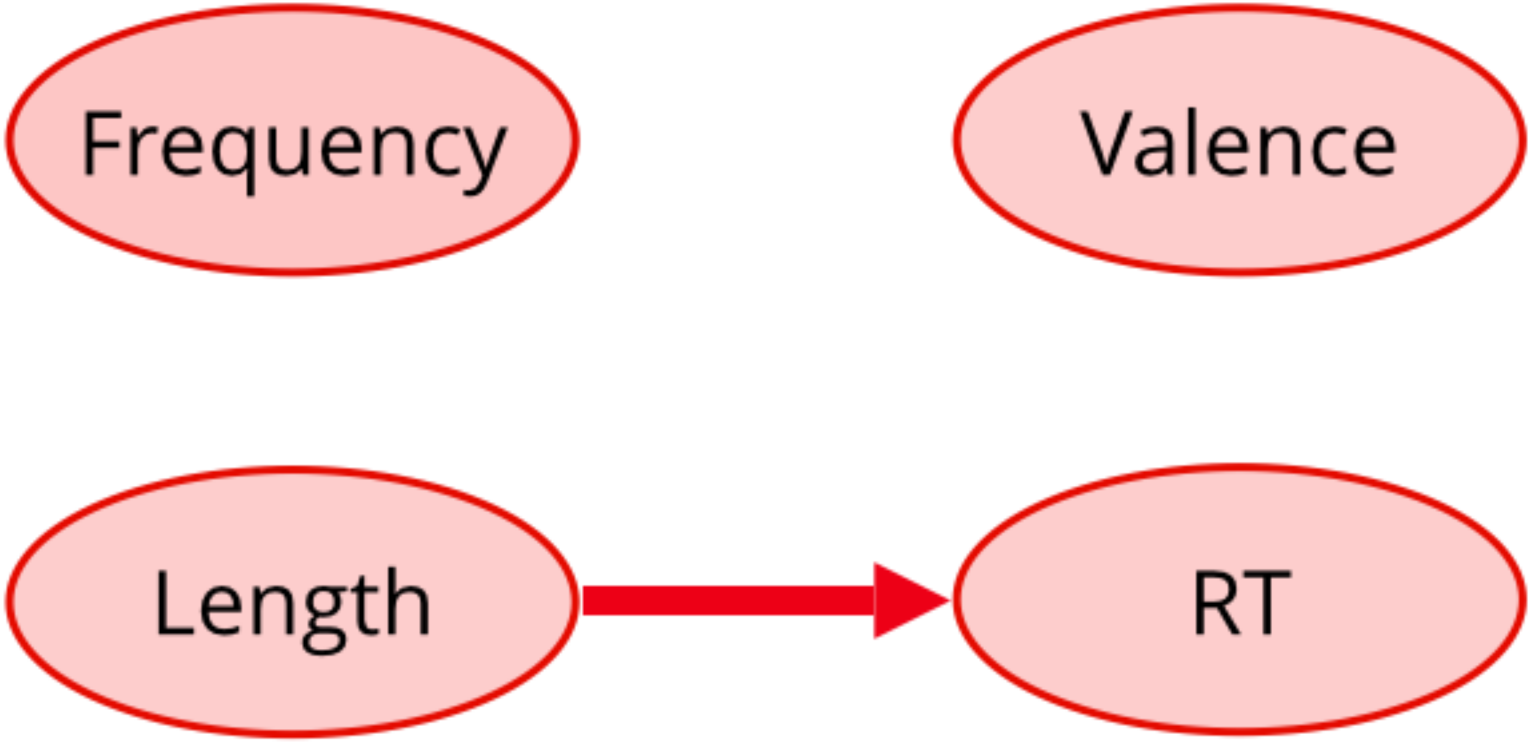

X is correlated with Y?
when controlling for Z?
X
Z
Y
Yes
No
X
Z
Y
Yes
No
X
Z
Y
Yes
No
Collider:
X
Z
Y
No
Yes

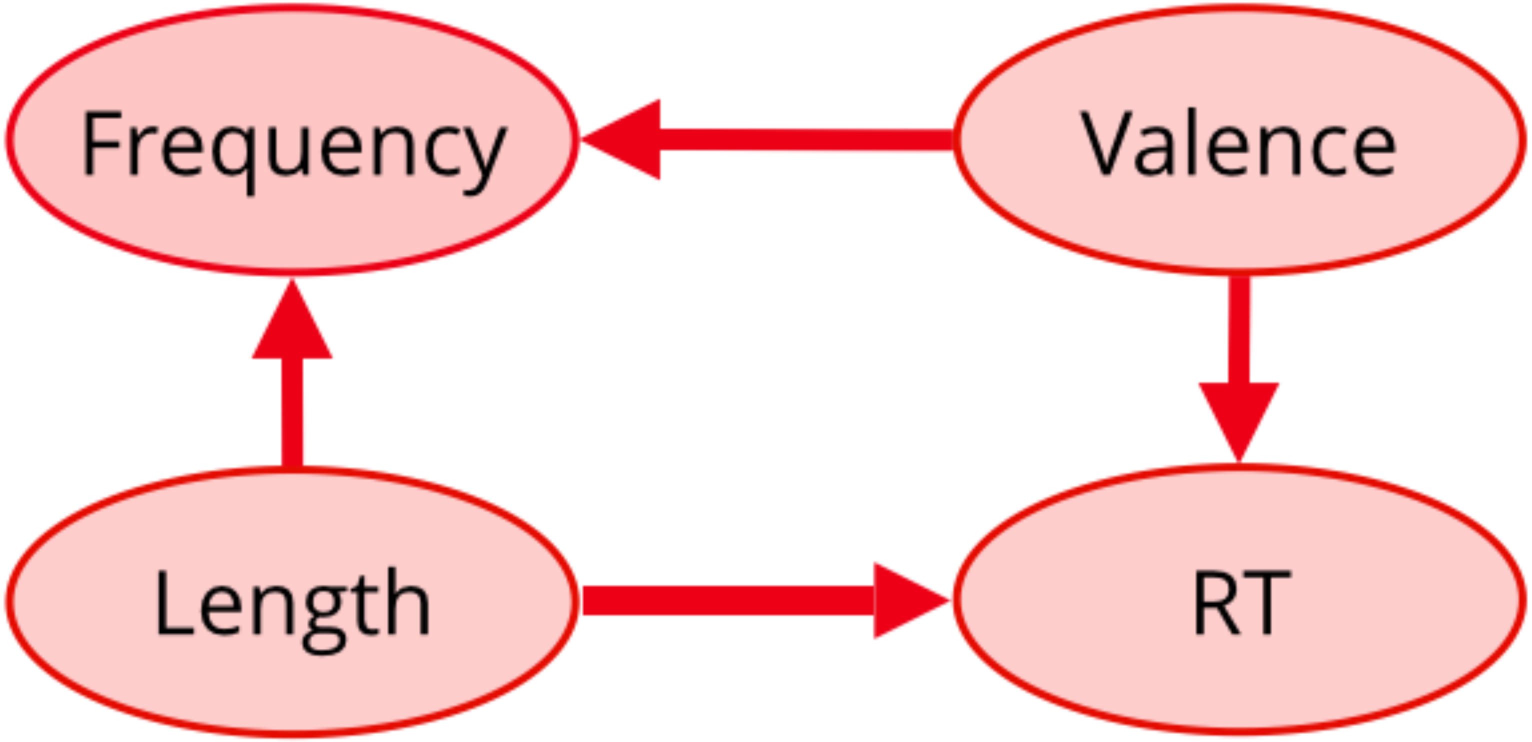


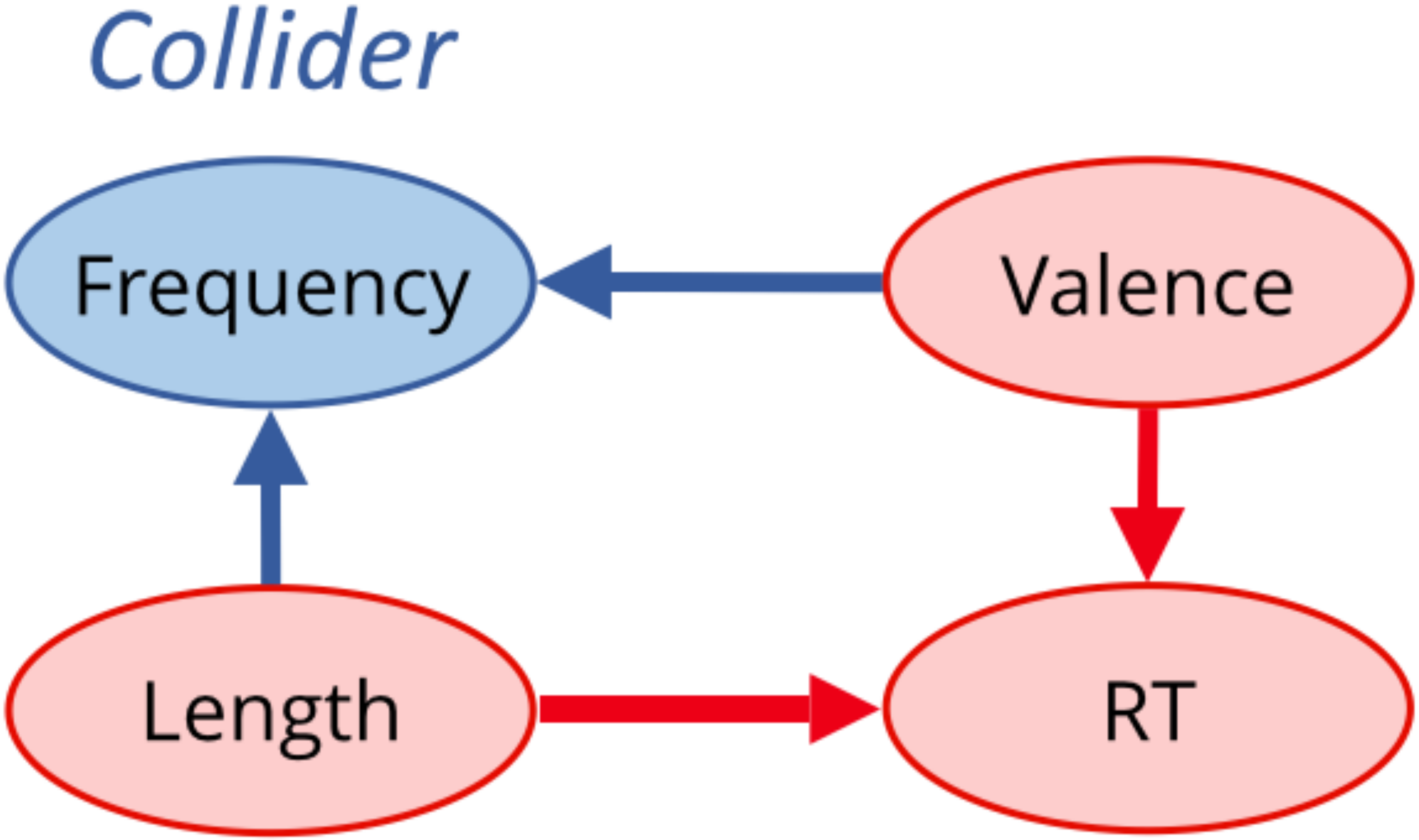
Colliders
X
Y
Z
X
Z
Y

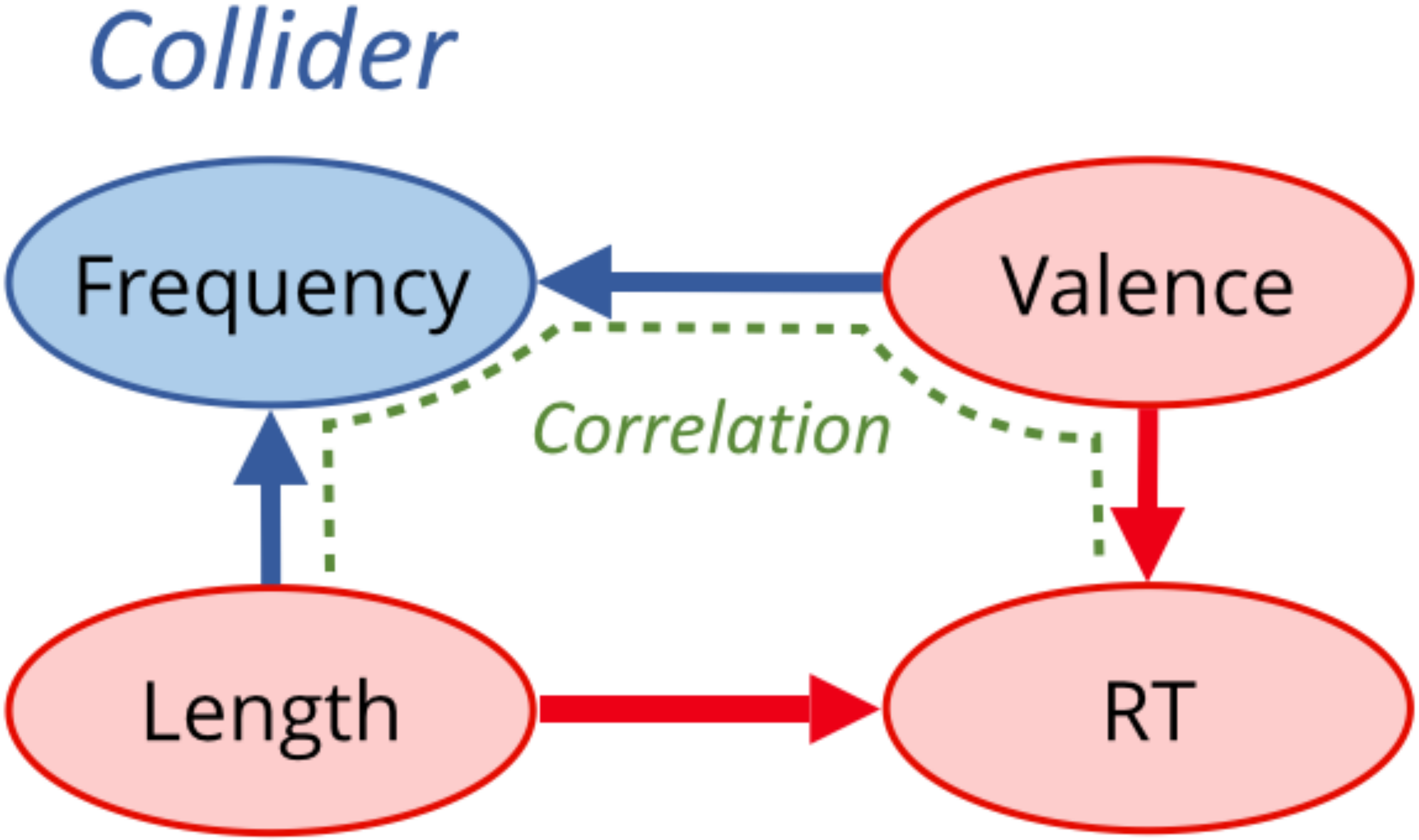

Number of Potatoes
Number of Carrots
Weight
All sacks
Potatoes
Carrots
1kg sacks
2kg sacks
3kg sacks
Potatoes
Carrots
Carrots
Carrots

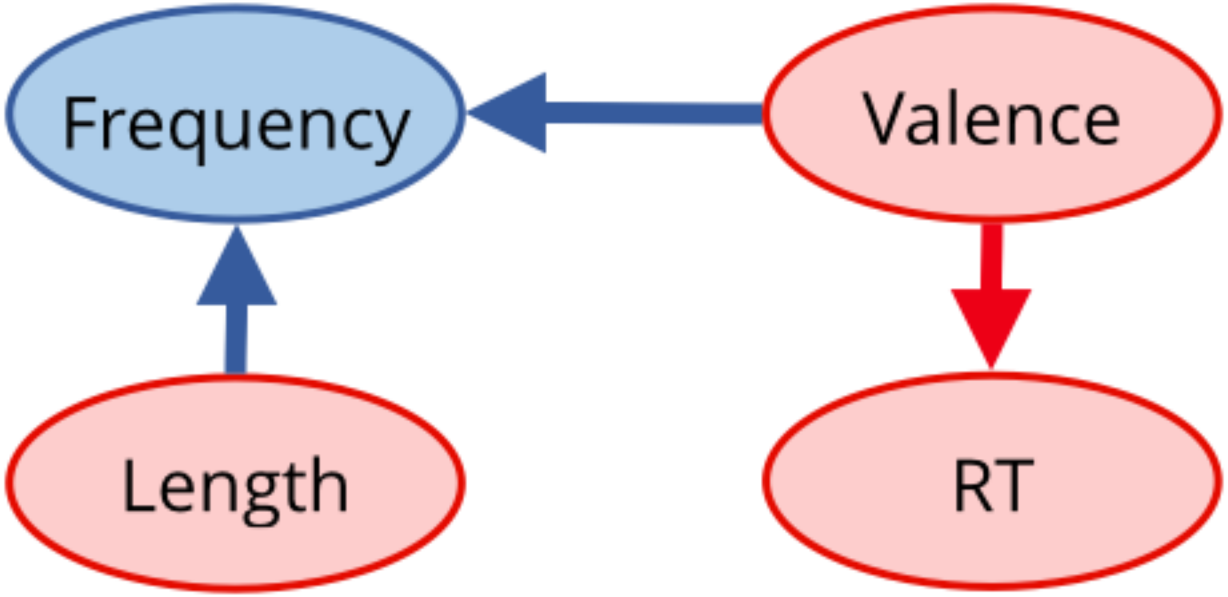

Length
RT


Frequency
Valence
Length
RT


Frequency
Valence
Length
RT


Collider
Frequency
Valence
Length
RT


Collider
Frequency
Valence
Correlation
Length
RT



Create some hypothetical variables

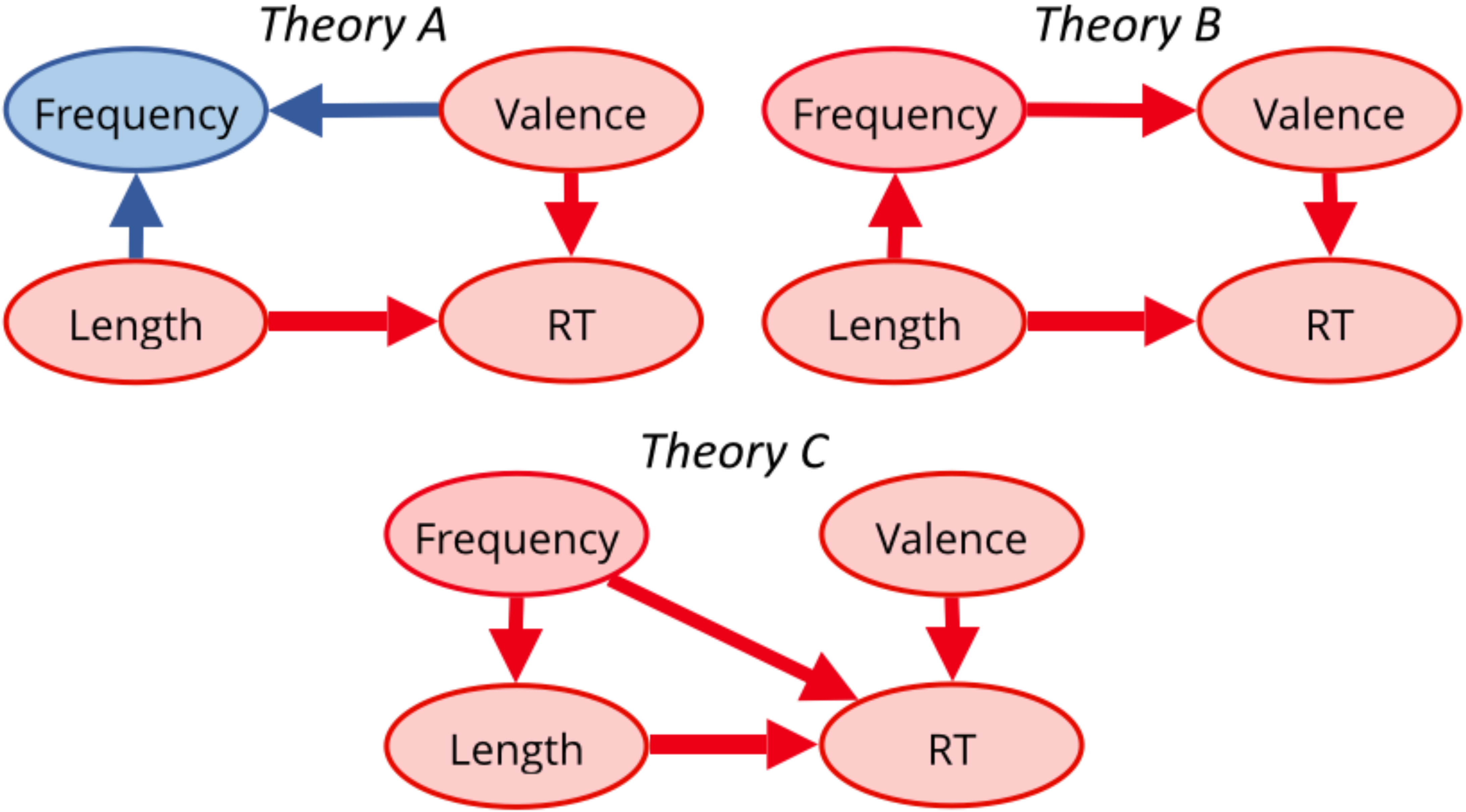
```
n = 200
length =  sample(1:7, n, replace = T)
valence = sample(1:7, n, replace = T)
freq = length + valence + rnorm(n)
RT = valence + rnorm(n)
```

Model without controls:

```
summary(lm(RT ~ length))
            Estimate Std. Error t value Pr(>|t|)
length      -0.03436    0.07971  -0.431    0.667
```

Model controlling for frequency:

```
summary(lm(RT ~ length + freq))
            Estimate Std. Error t value Pr(>|t|)
length      -0.83004    0.06520 -12.730   <0.001 ***
freq         0.85081    0.04647  18.310   <0.001 ***
```



Theory A
Frequency
Valence
Length
RT
Theory B
Frequency
Valence
Length
RT
Theory C
Frequency
Valence
Length
RT

Identifying confounding variables

Causal (but vague)

Cartwright: Control for anything that is “causally relevant”

Declarative

Any variable that is correlated with both X and Y.

Procedural

Noncollapsibility: Try controlling for Z. If it makes a difference, then Z is a confounder.

Identifying confounding variables

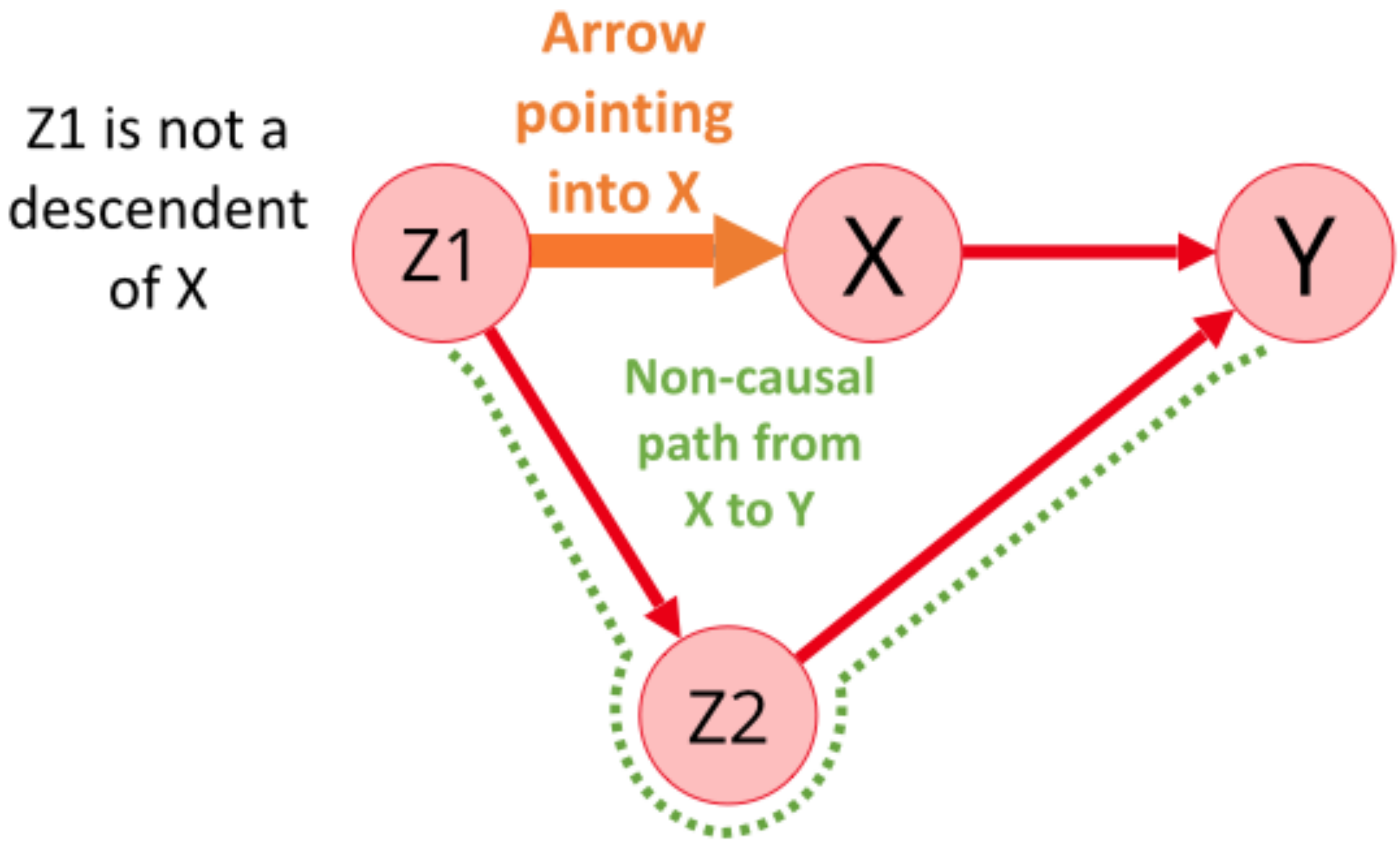
We can observe P(Y | X)
We want to find P(Y | do(X))
Confounds are anything that leads to a difference between these.

Block every noncausal path between X and Y
Without blocking any causal paths
Block all back-door paths:
Control for all variables on a path from X to Y that starts with an arrow pointing to X.
And where the variable is not a descendant of X
See Shrier & Platt (2008)

Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

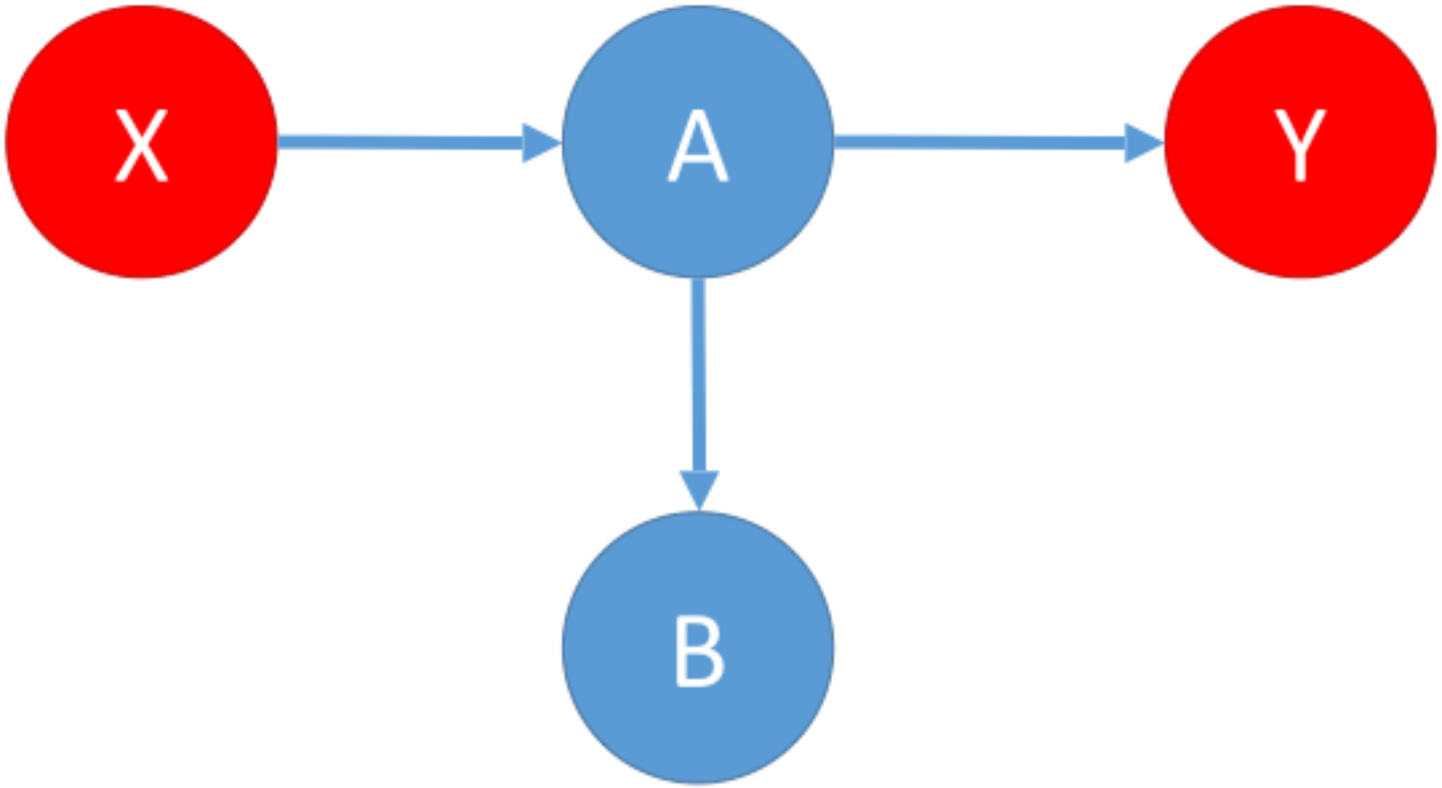
And where the variable is not a descendant of X



Block Back Door Paths:

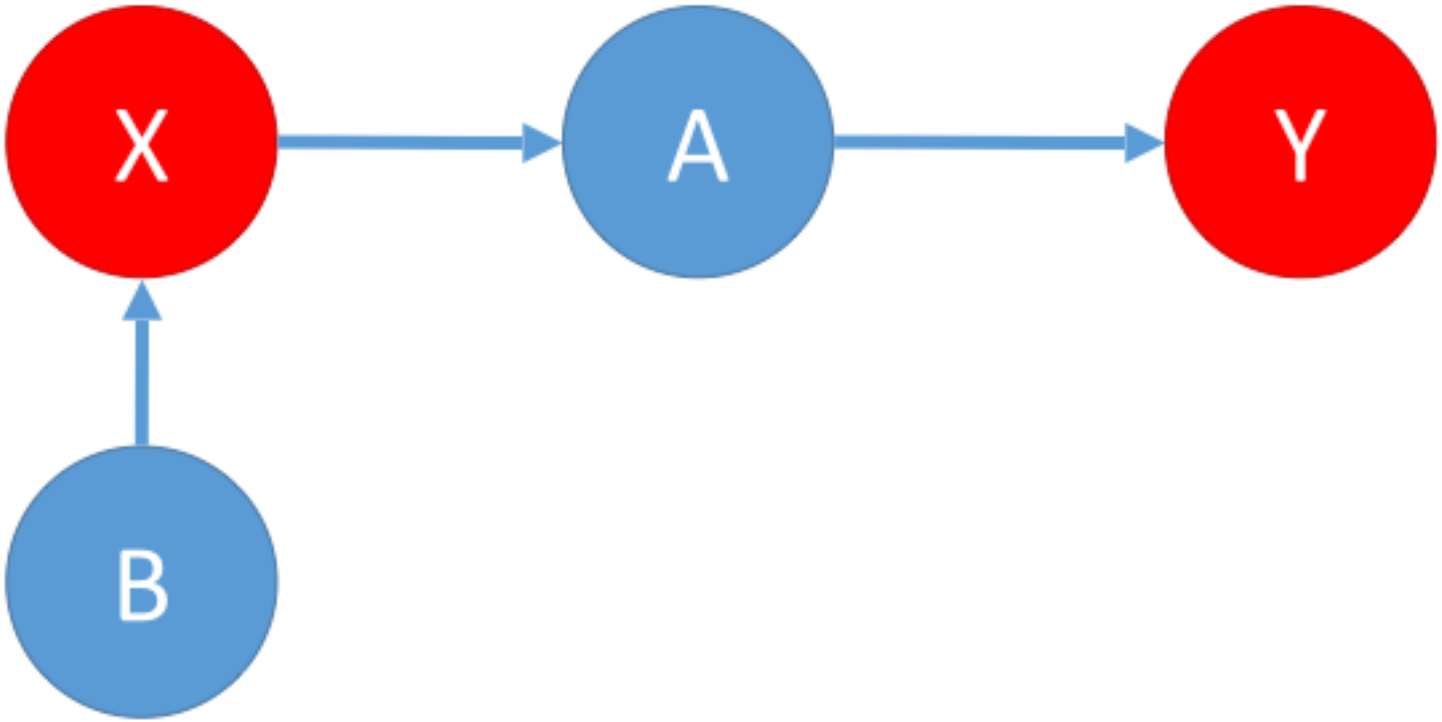
Variables on a path from X to Y that starts with an arrow pointing to X.

And where the variable is not a descendant of X



Block Back Door Paths:

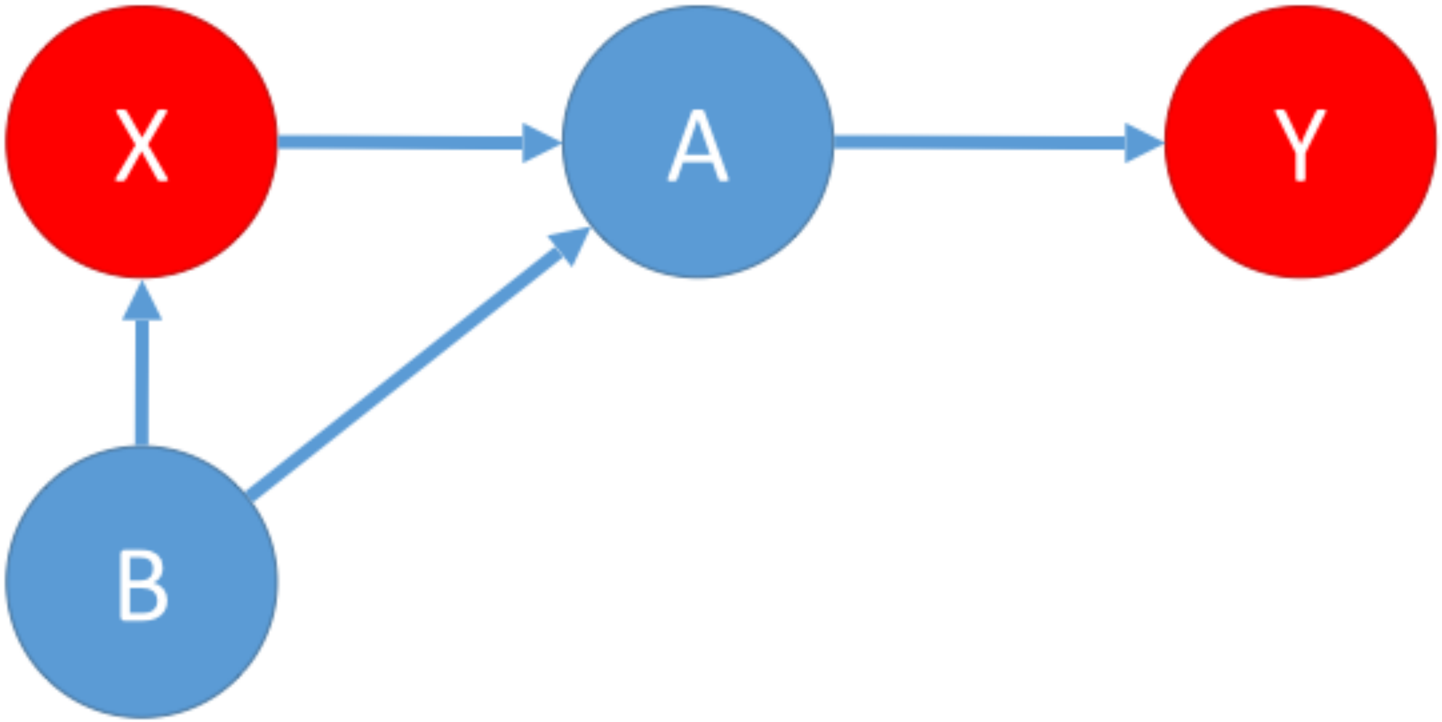
Variables on a path from X to Y that starts with an arrow pointing to X.
And where the variable is not a descendant of X



Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

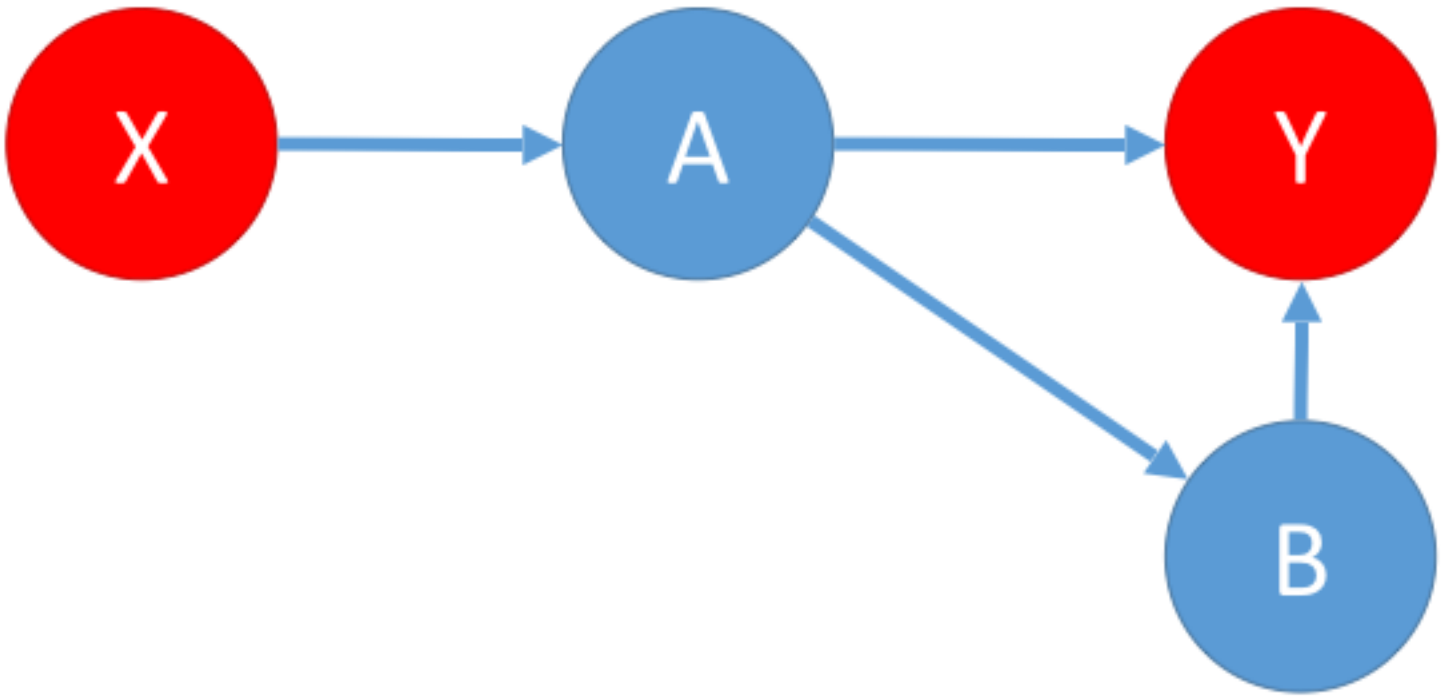
And where the variable is not a descendant of X



Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

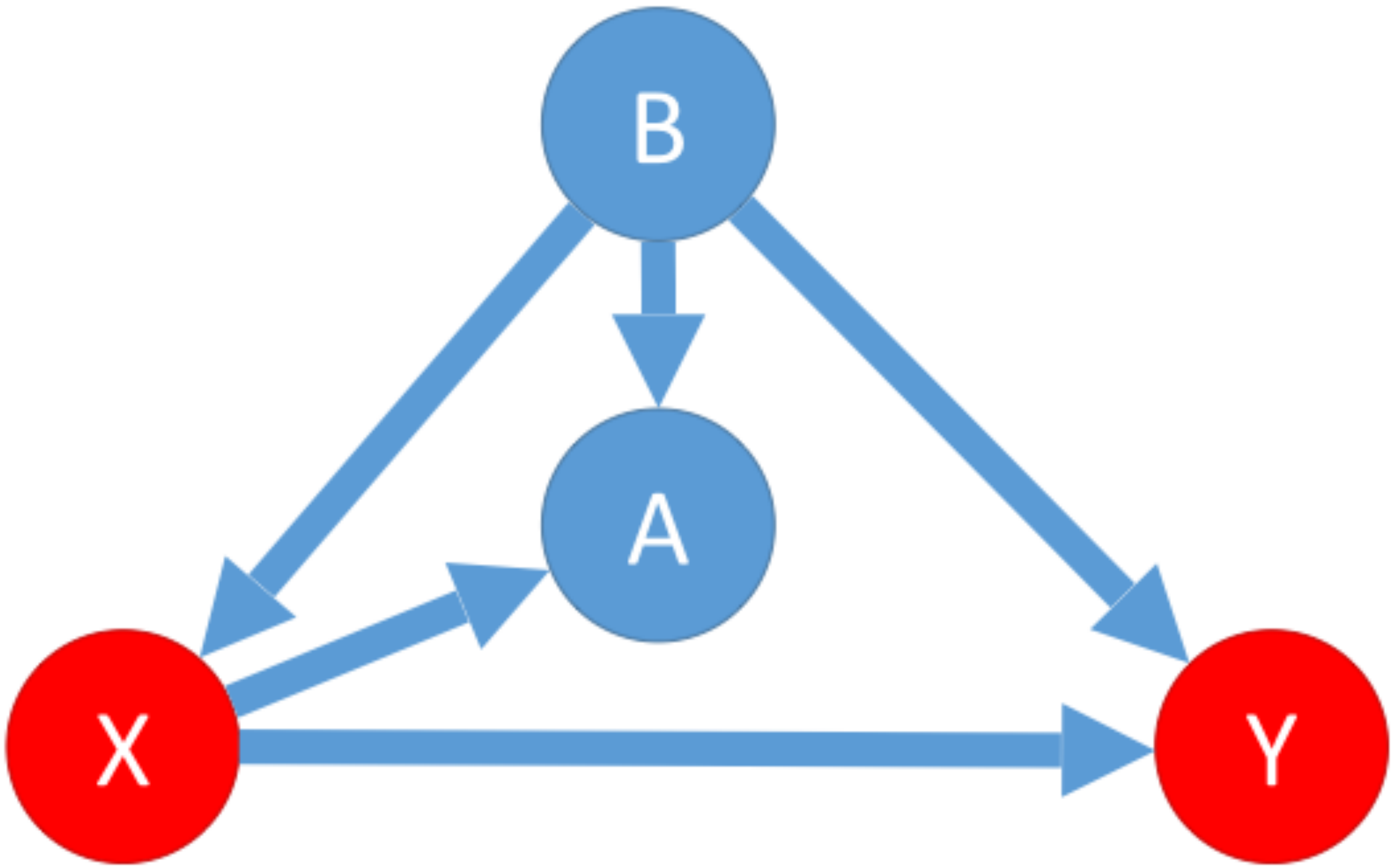
And where the variable is not a descendant of X



Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

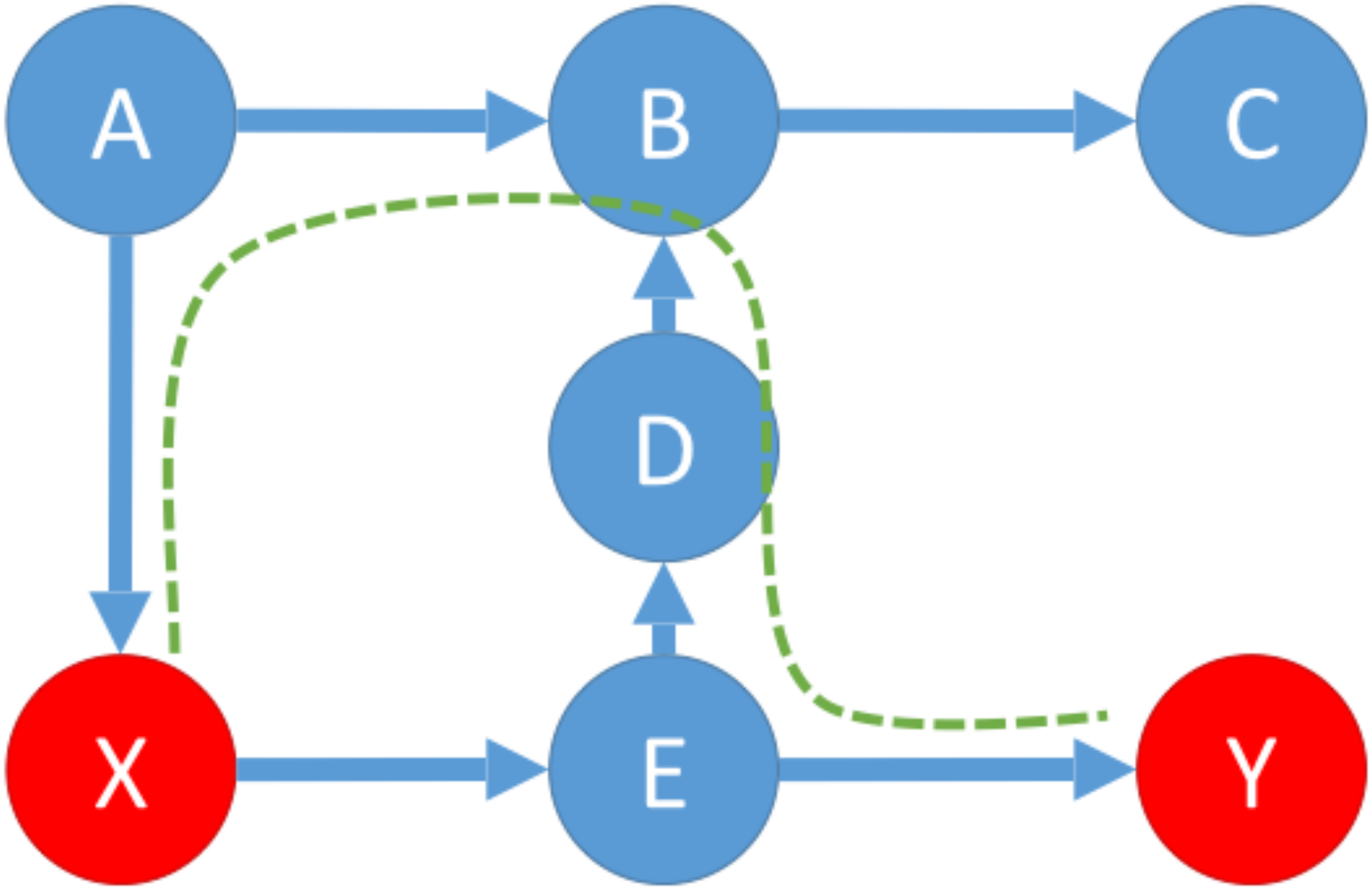
And where the variable is not a descendant of X



Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

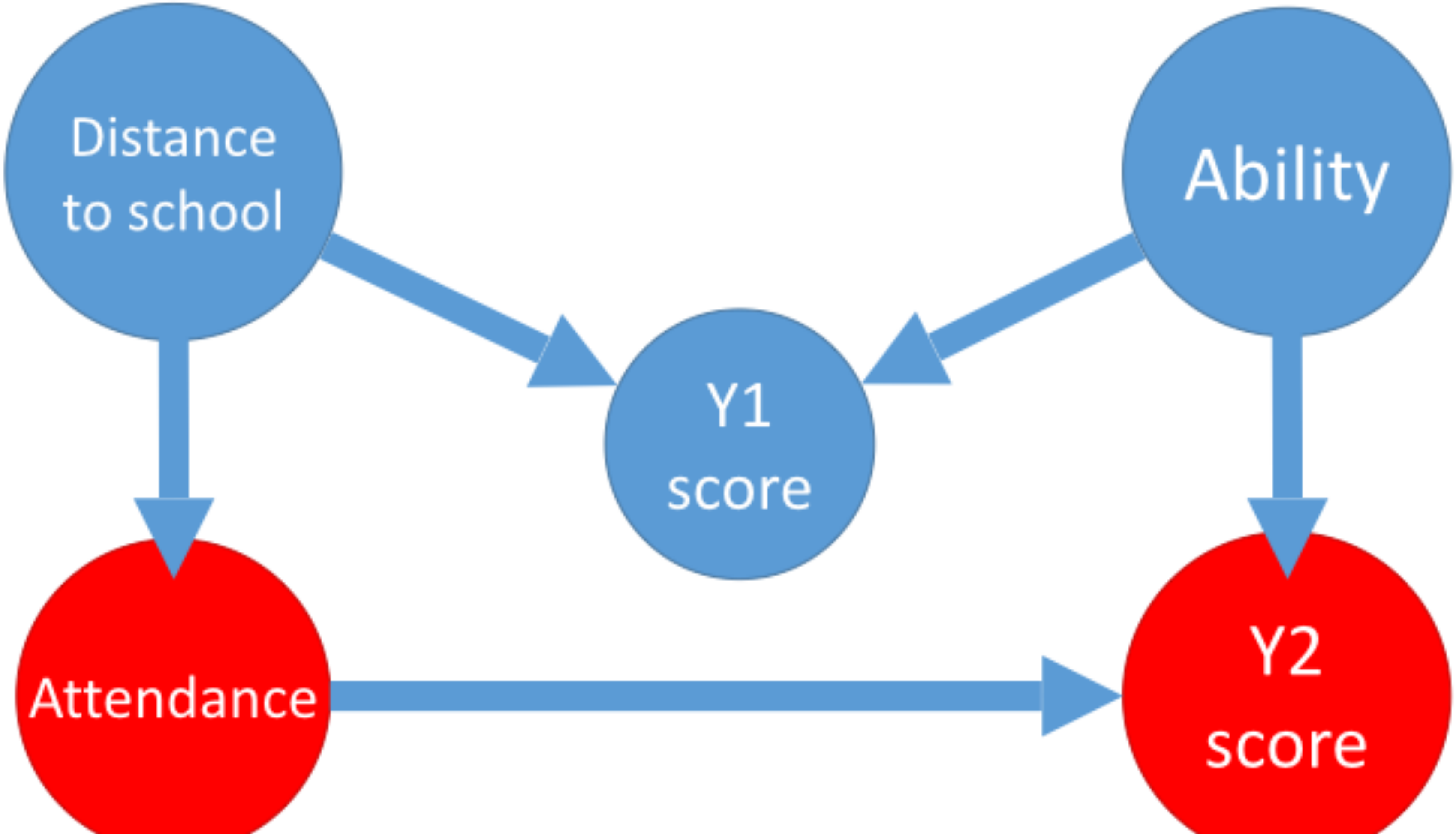
And where the variable is not a descendant of X



Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

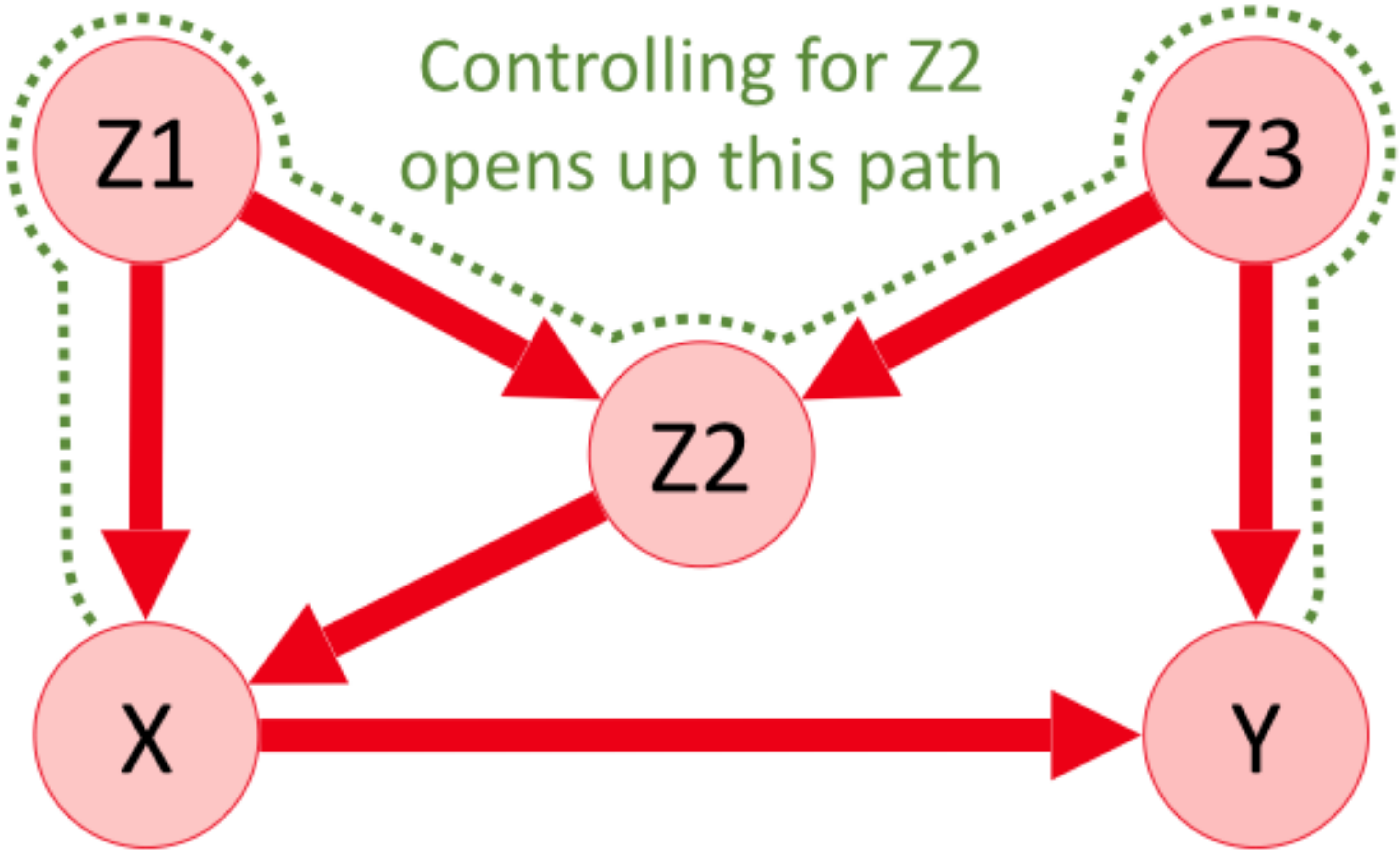
And where the variable is not a descendant of X



Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

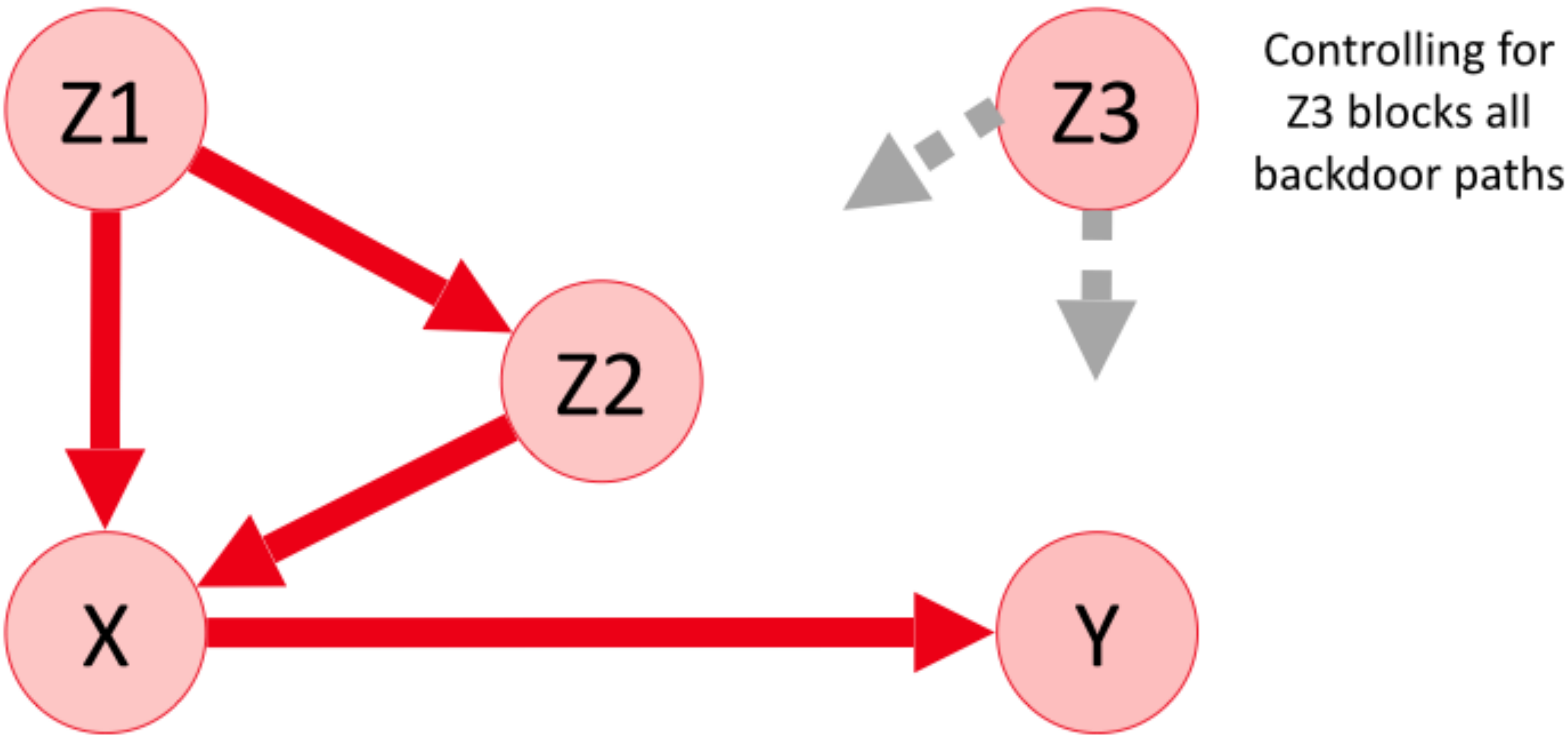
And where the variable is not a descendant of X



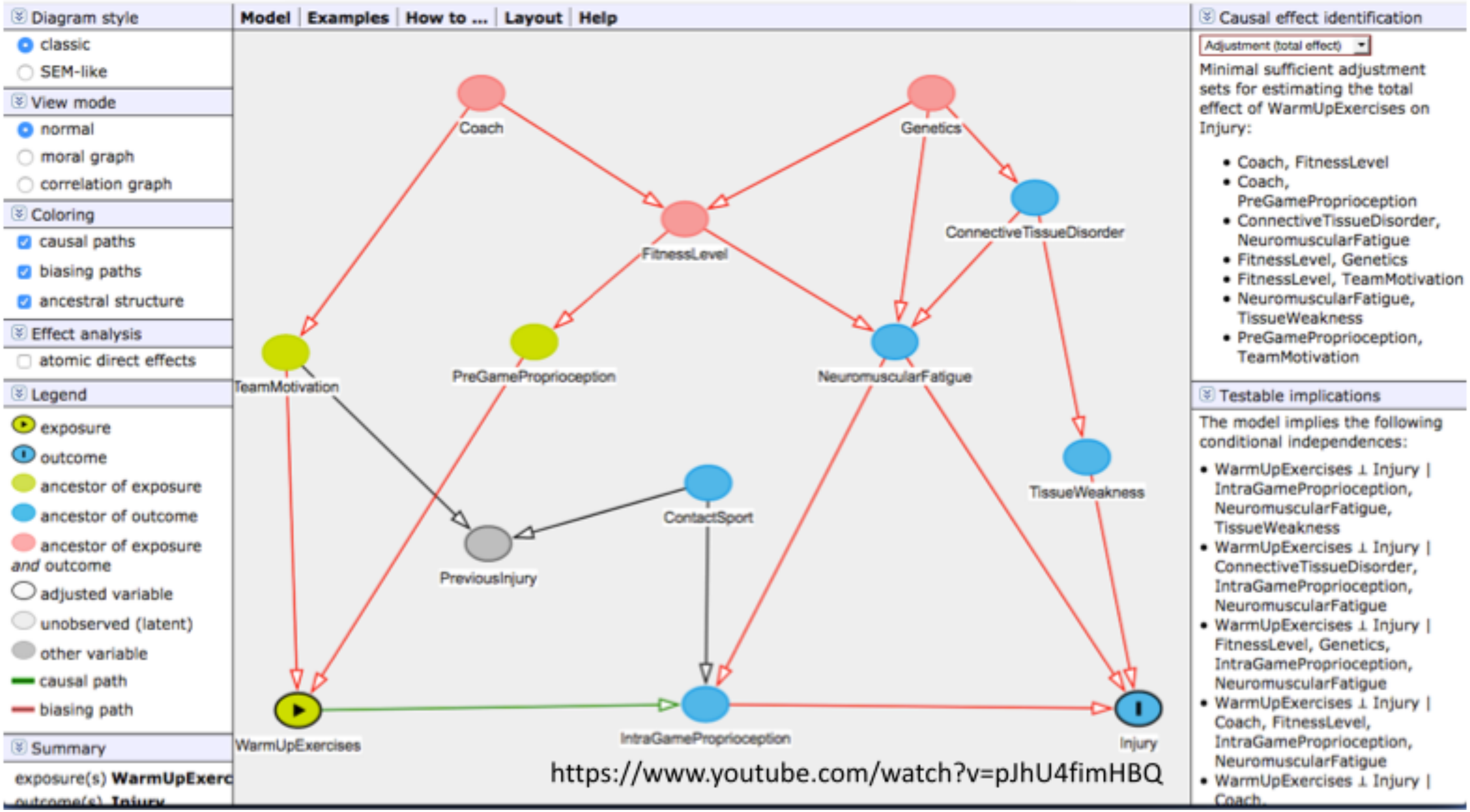
Block Back Door Paths:

Variables on a path from X to Y that starts with an arrow pointing to X.

And where the variable is not a descendant of X



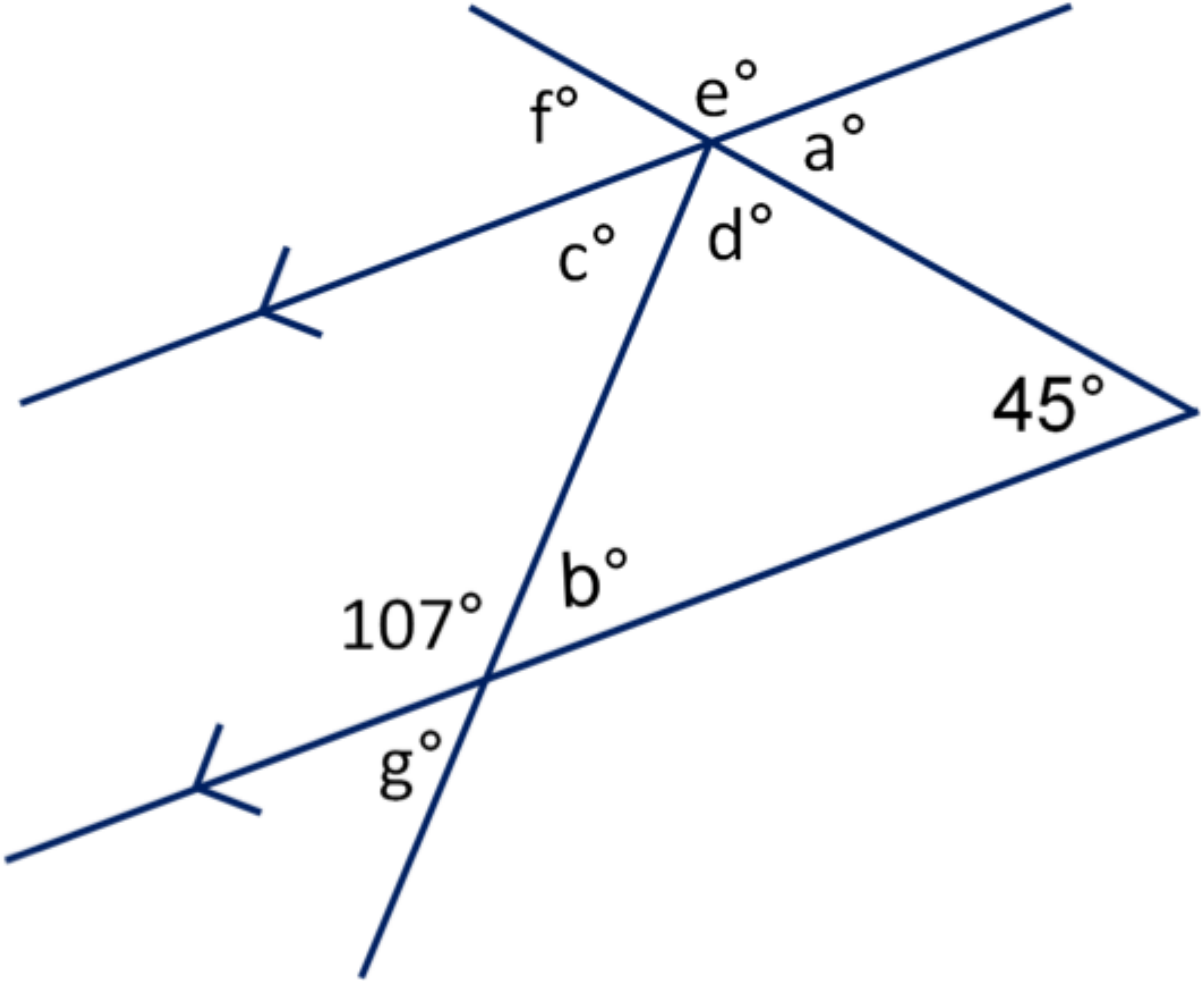
Dagitty http://www.dagitty.net/



https://www.youtube.com/watch?v=pJhU4fimHBQ

Causal Axioms

Geometric axioms +
geometric calculus
= ability to solve
problems that looked
impossible before.



Causal Axioms

Causal calculus aims to do the same:

Exchange probabilities based on intervention (do(X))

for probabilities based on observation (see(X))

Rule1: $P(Y \mid do(X), Z, W) = P(Y \mid do(X), Z)$

if W is irrelevant to Y

Rule 2: $P(Y \mid do(X), Z) = P(Y|X,Z)$

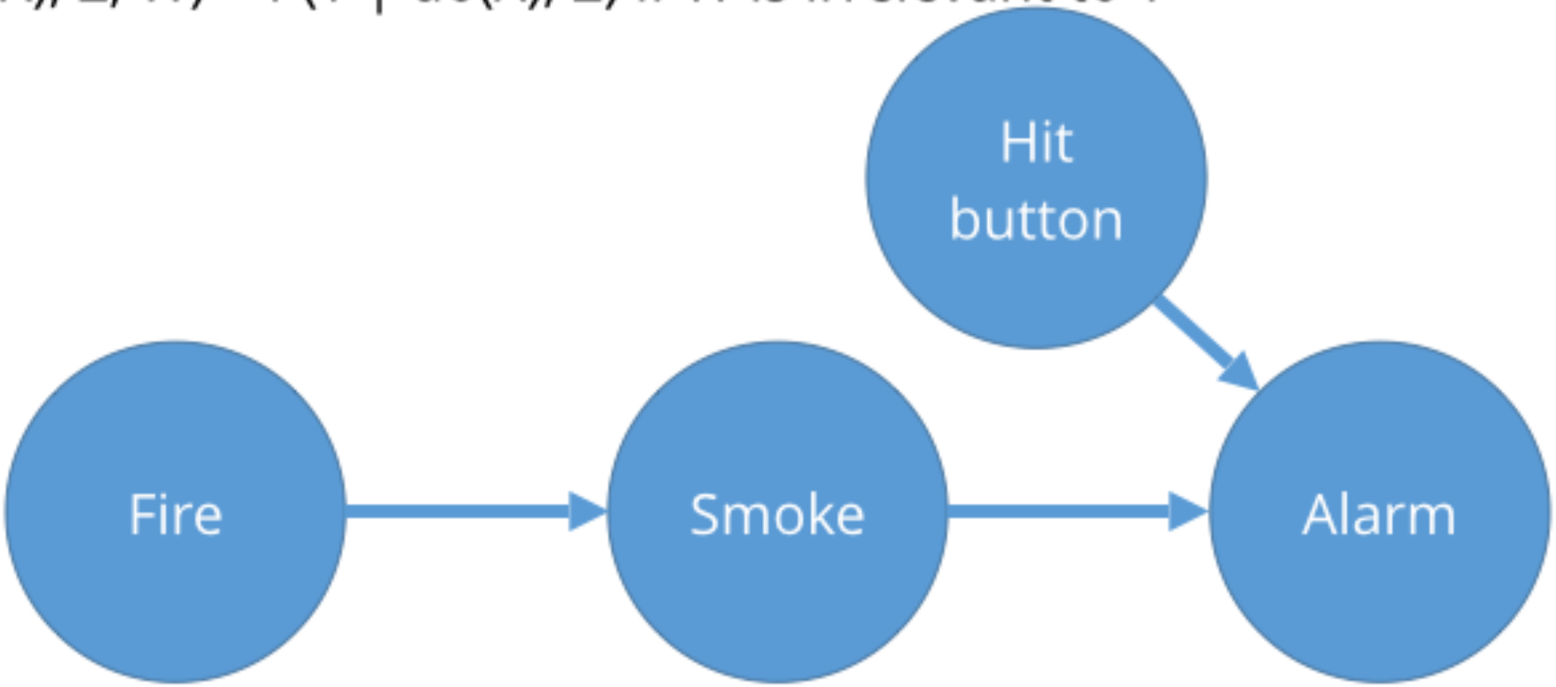
if Y is independent of Z given X and W if all connections are severed from X

Rule 3: $P(Y \mid do(X)) = P(Y)$

if there is no path from X to Y with only forward-directed arrows

Rule 1: Delete observations

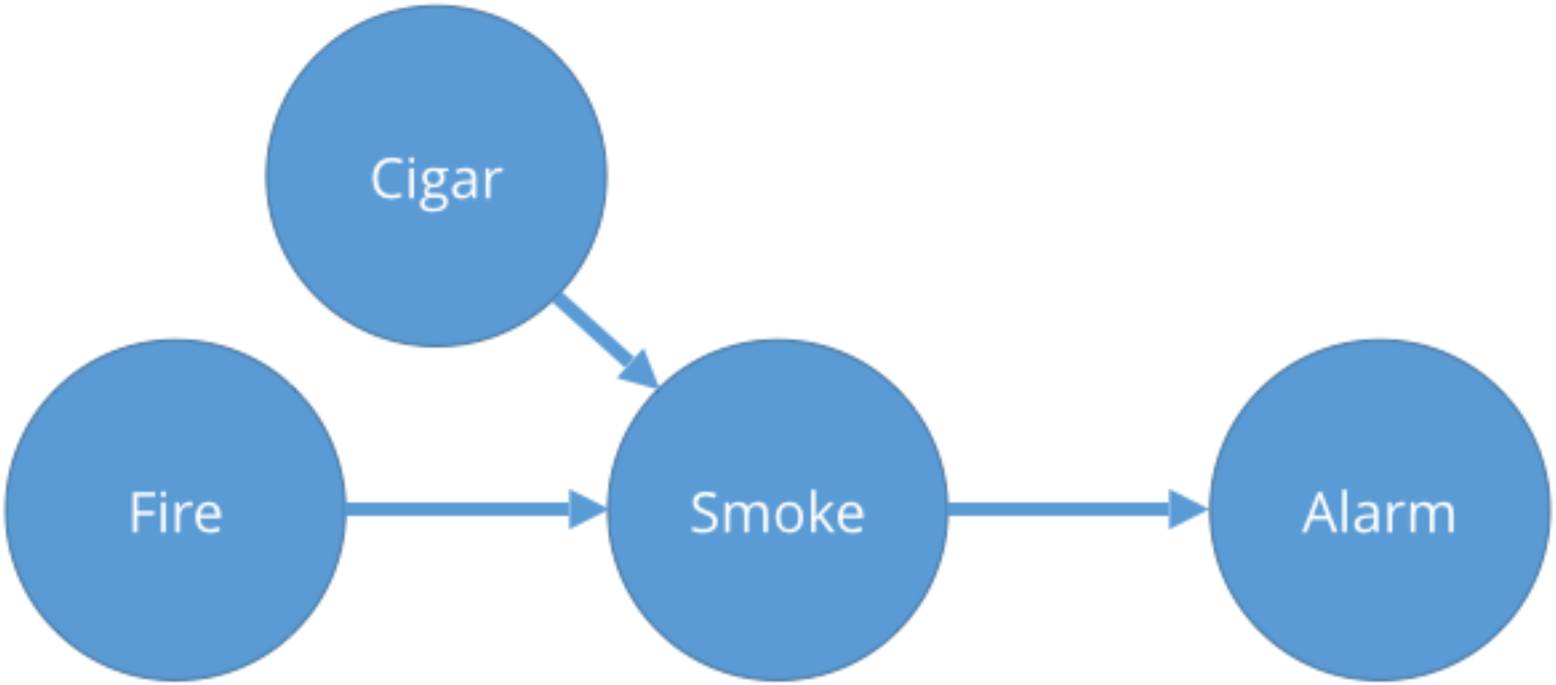
P(Y | do(X), Z, W) = P(Y | do(X), Z) if W is irrelevant to Y



P(Alarm | do(hit button), Smoke, Fire) =
P(Alarm | do(hit button), Smoke)

Rule 2: Exchange interventions with observations

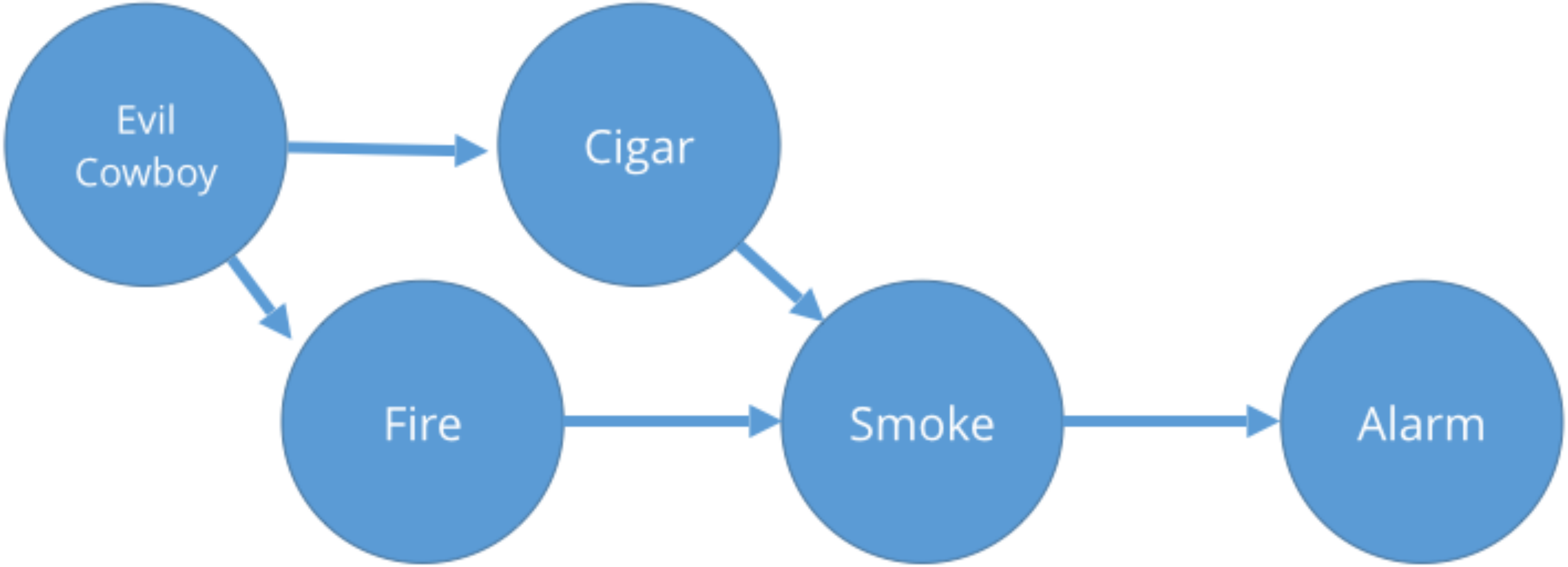
P(Y | do(X), Z) = P(Y|X,Z) if Y is independent of Z given X and W if all connections are severed from X



P(Alarm | *do*(Cigar), Smoke) =
P(Alarm | Cigar, Smoke)

Rule 2: Exchange interventions with observations

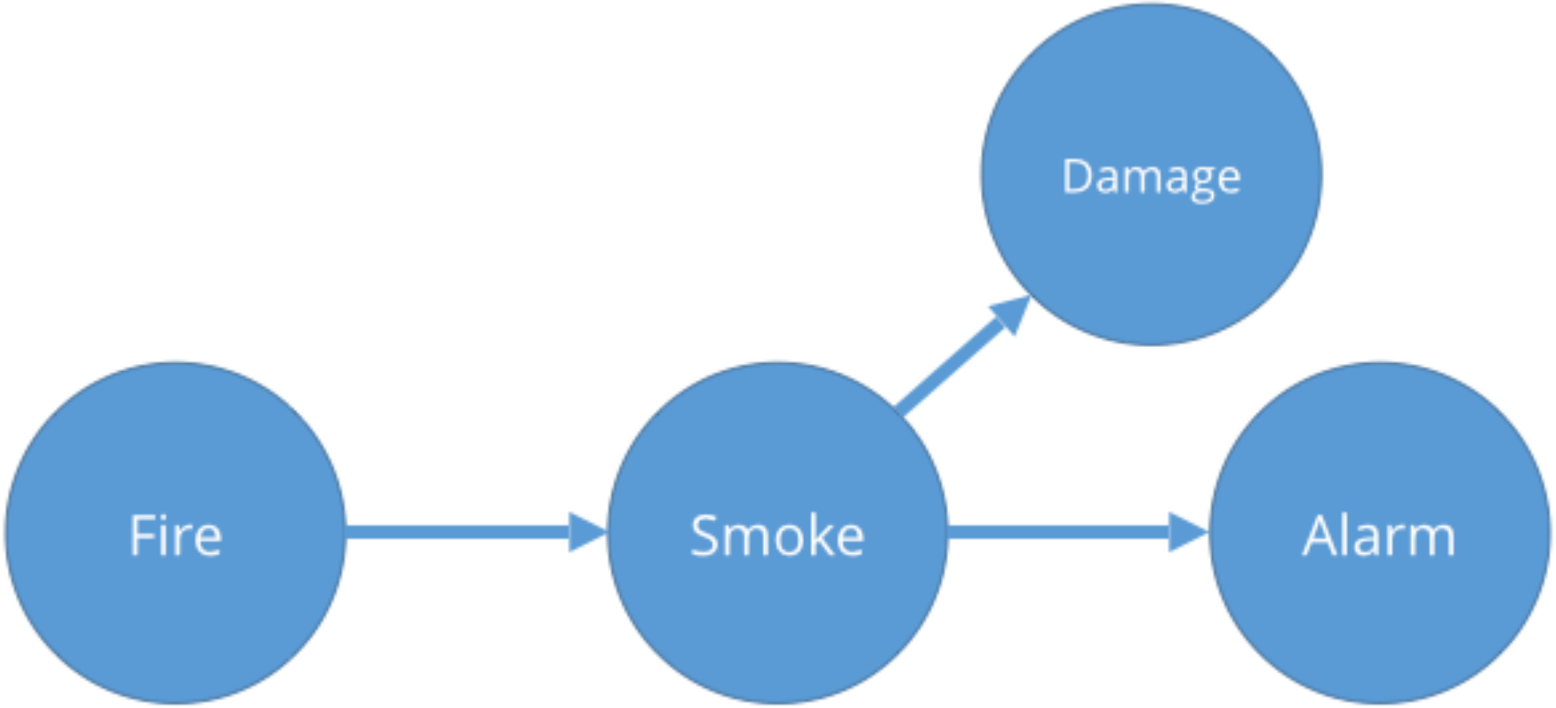
P(Y | do(X), Z) = P(Y|X,Z) if Y is independent of Z given X and W if all connections are severed from X



P(Alarm | *do*(Cigar), Smoke) ≠
P(Alarm | Cigar,   Smoke)

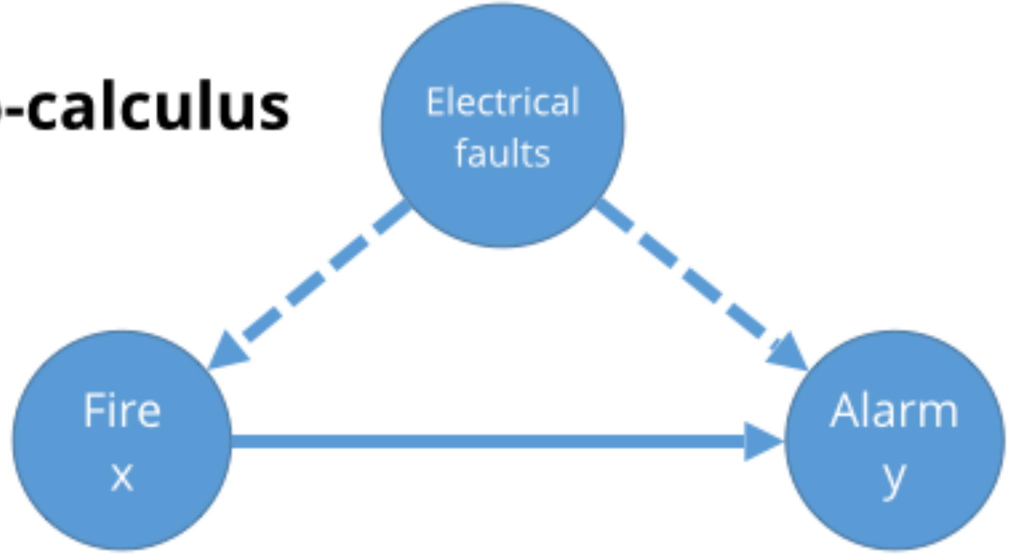
Rule 3: Delete interventions

P(Y | do(X)) = P(Y) if there is no path from X to Y with only forward-directed arrows



P(Alarm | *do*(Damage)) =
P(Alarm)

Applying the do-calculus



Calculation with interventions

$$
\begin{aligned}
P(y|do\{x\}) &= \sum_{z} P(y|do\{x\}, z)P(z|do\{x\}) && \text{Probability axioms}\\
&= \sum_{z} P(y|do\{x\}, do\{z\})P(z|do\{x\}) && \text{Rule 2}\\
&= \sum_{z} P(y|do\{x\}, do\{z\})P(z|x) && \text{Rule 2}\\
&= \sum_{z} P(y|do\{z\})P(z|x) && \text{Rule 3}\\
&= \sum_{x'}\sum_{z} P(y|do\{z\}, x')P(x'|do\{z\})P(z|x) && \text{Probability axioms}\\
&= \sum_{x'}\sum_{z} P(y|z, x')P(x'|do\{z\})P(z|x) && \text{Rule 2}\\
&= \sum_{x'}\sum_{z} P(y|z, x')P(x')P(z|x) && \text{Rule 3}
\end{aligned}
$$

Calculation with no interventions

Eliminating interventions

These axioms are complete: they can discover a solution if one exists (Shpitser, 2008).

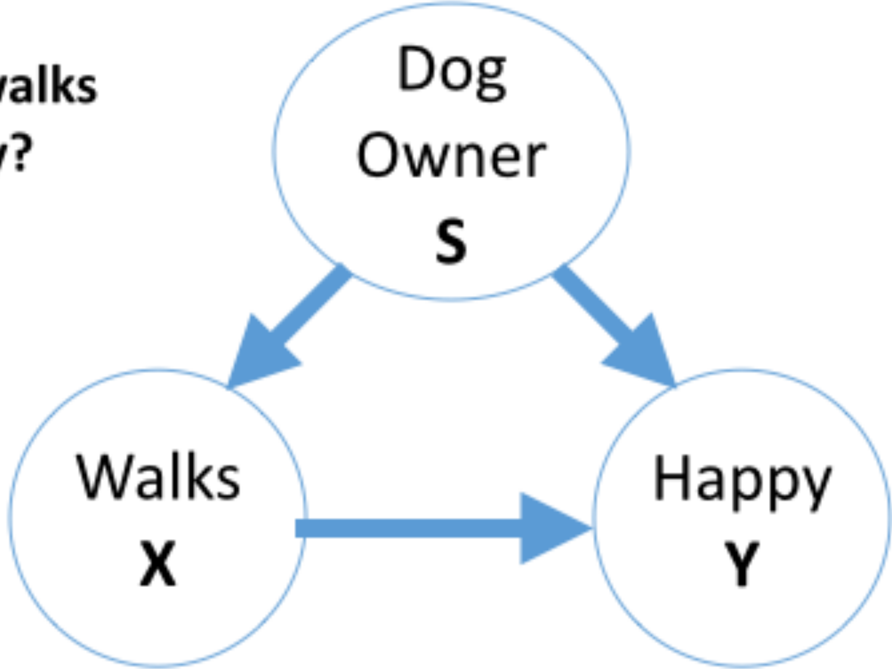
Algorithms exist to discover the solution in polynomial time (Shpitser, 2008).

If there is no solution, then we must do an experiment.

Algorithms exist to tell us what variables to experiment on, or discover other variables to manipulate if we can't manipulate the target (Bareinbolm, 2012)

Back door adjustment

Does going on walks make you happy?



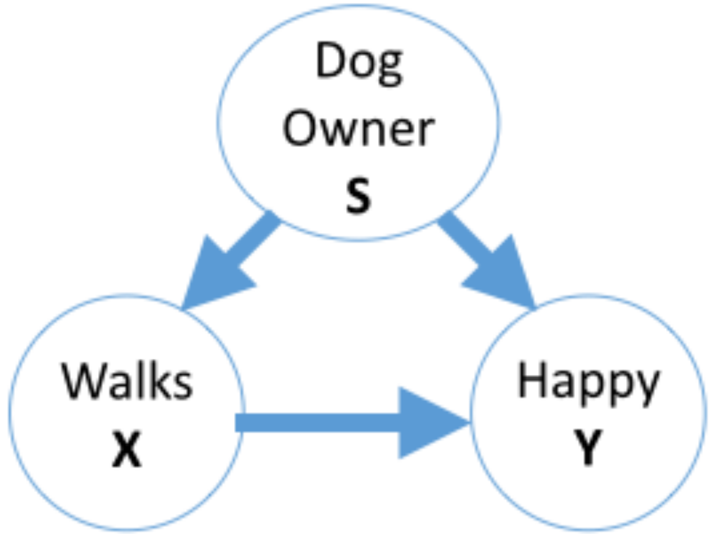
S is a set of variables that satisfies the back door criterion

$$P(Y \mid do(\mathbf{X} = x)) = \Sigma_s P(Y \mid \mathbf{X} = x, S = s)\, P(S = s)$$

Probability of Y if we intervene so that **X** is x

Probability of observing Y when **X** is x, for each possible value of S

Back door adjustment



$$P(\text{Happy} \mid do(\text{Walks} = \text{Yes})) =$$

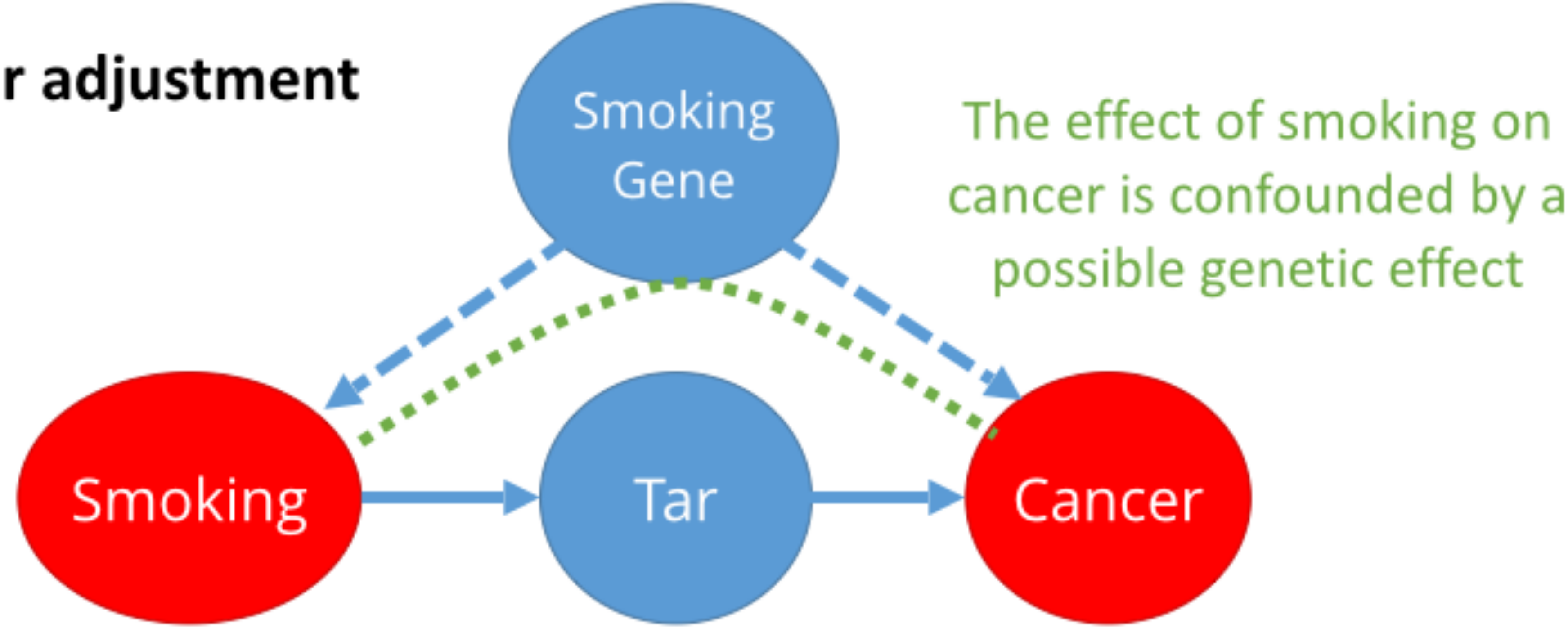
$$\Sigma_s\, P(\text{Happy} \mid \text{Walks} = \text{Yes}, \text{Dog} = s)\, P(\text{Dog} = s)$$

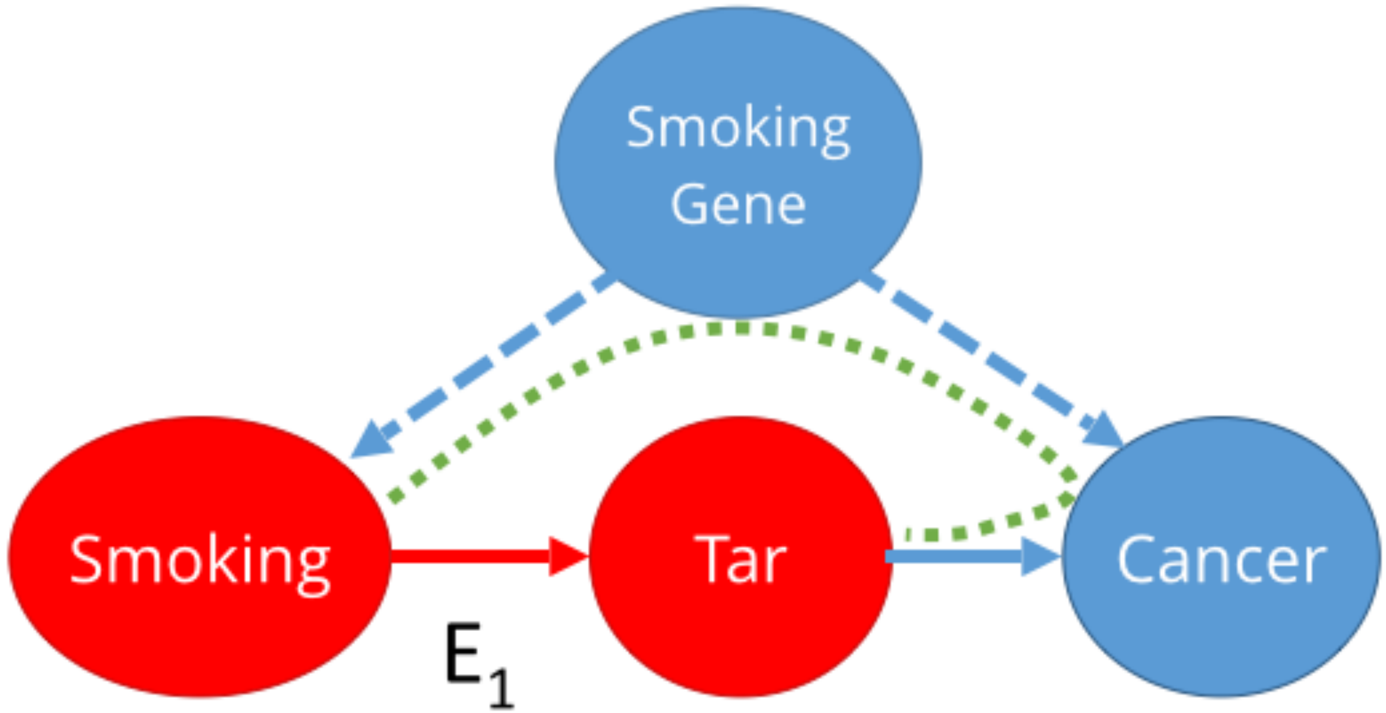
$$= \ P(\text{Happy} \mid \text{Walks} = \text{Yes}, \text{Dog} = \text{Yes})\, P(\text{Dog} = \text{Yes})$$

$$+\, P(\text{Happy} \mid \text{Walks} = \text{Yes}, \text{Dog} = \text{No})\, P(\text{Dog} = \text{No})$$

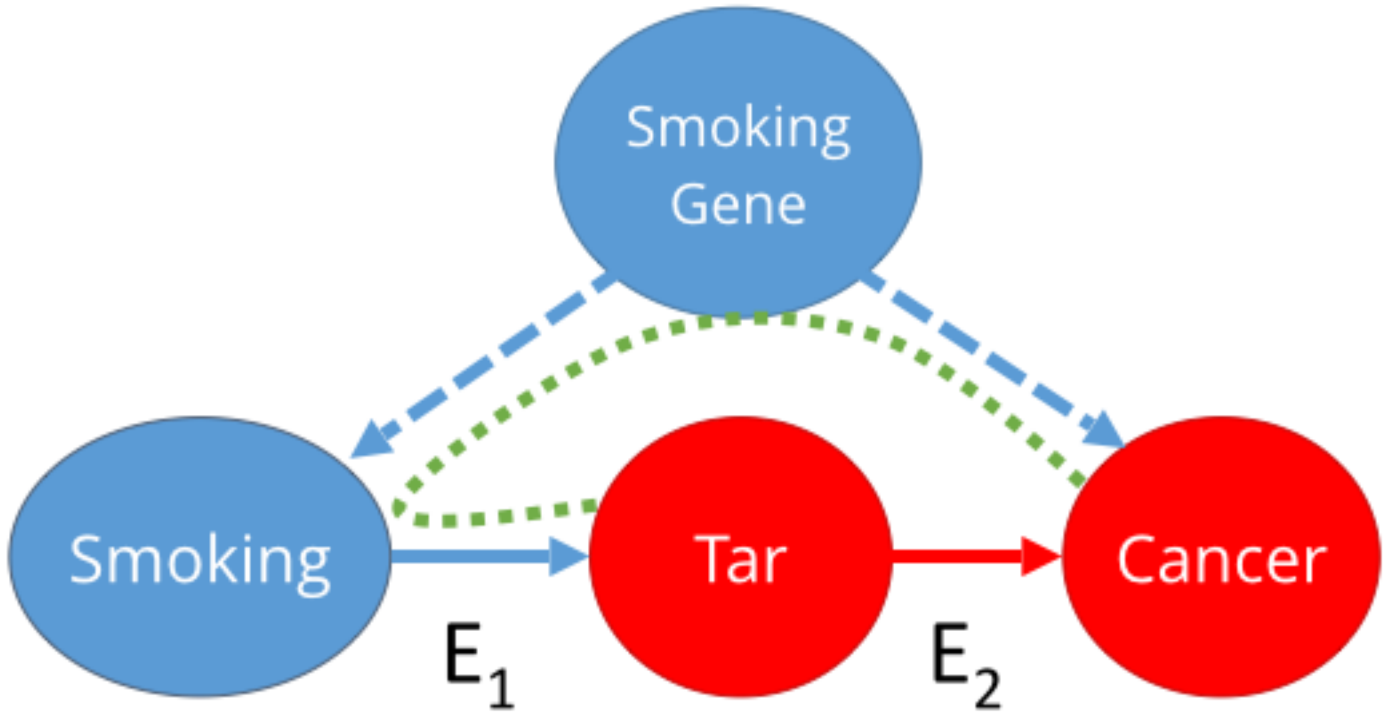
= Probability of being happy if you walk and have a dog, weighted by the probability of having a dog
+ Probaility of being happy if you walk and don't have a dog, weighted by the probability of not having a dog

Front door adjustment

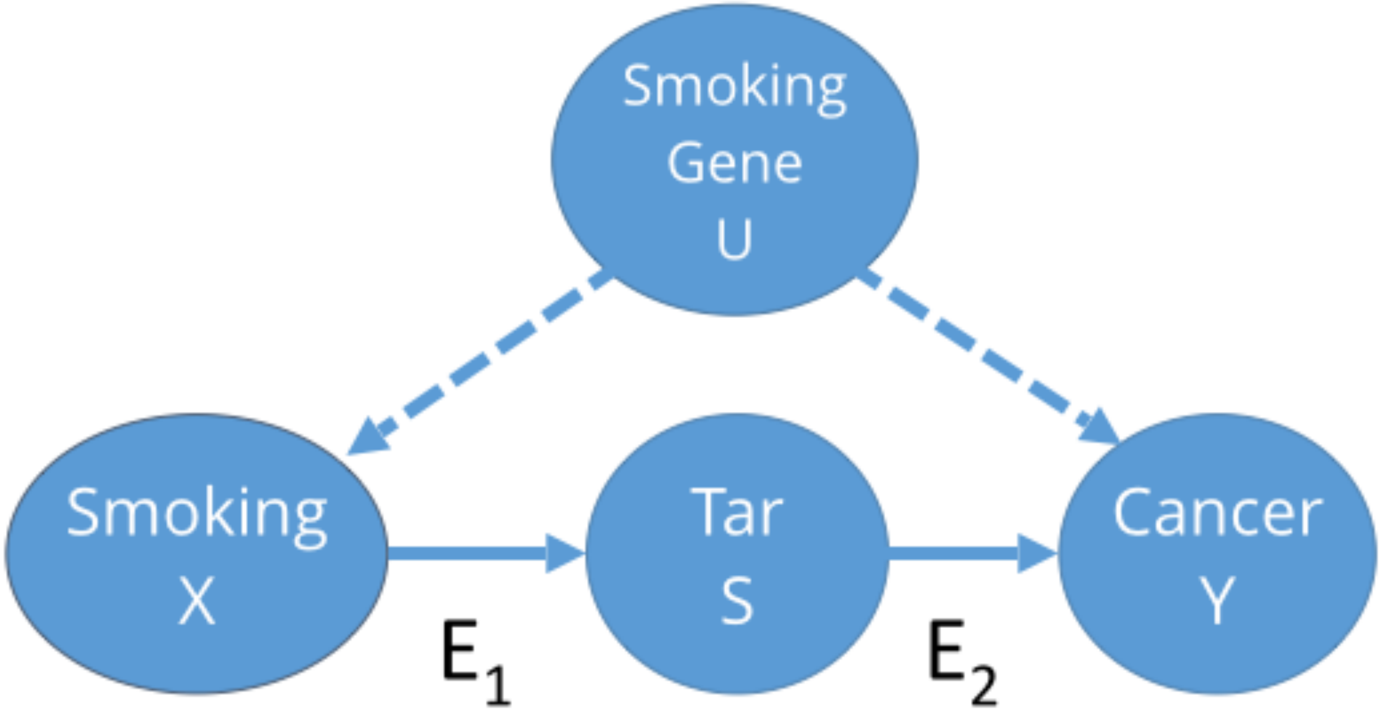




There is a back-door path from Smoking to Tar
BUT it's blocked by the collider in cancer
So we can just use the observed probabilities
Average causal effect E_1 = P(Tar | Smoking) – P(Tar | No smoking)



There is a back-door path from Tar to Cancer
But we can block it by controlling for smoking.
Average causal effect E_2 = P(Cancer | do(Tar)) – P(Cancer | do(No tar))

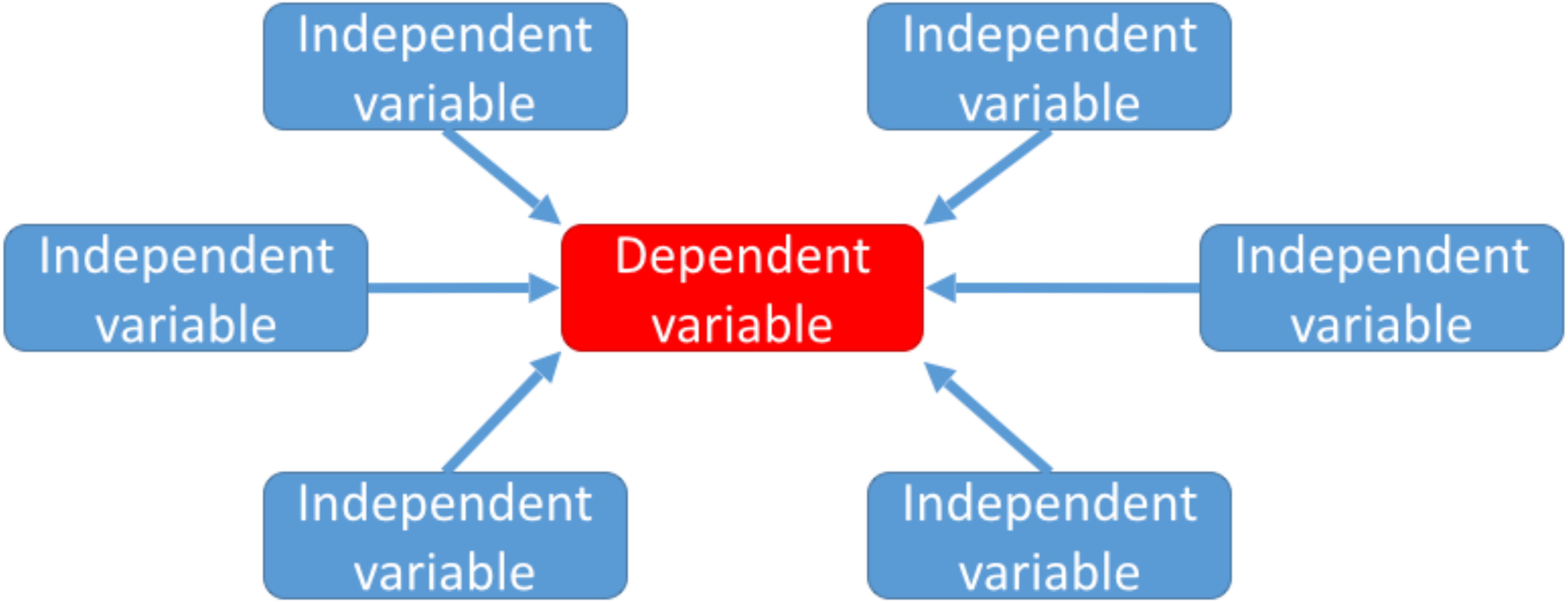


$$\Pr(Y|do(X=x)) = \sum_s \Pr(S=s|X=x) \sum_{x'} \Pr(Y|X=x', S=s) \Pr(X=x')$$

Fitting real data

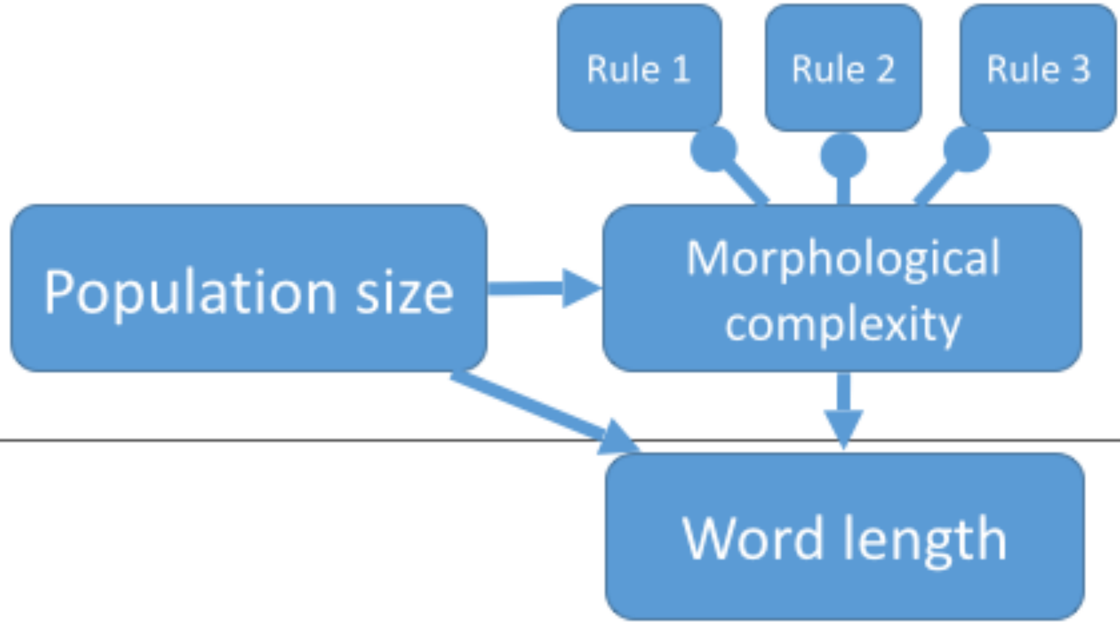
Model fitting in R

Standard assumption of linear model:



Model fitting in R

Convert a causal graph to a structural equation model




```
library(lavaan)
library(semPlot)
```

Define

```
model <- "
MorphologicalComplexity =~ Rule1 + Rule2 + Rule3
MorphologicalComplexity ~ PopulationSize
WordLength ~ MorphologicalComplexity + PopulationSize"
```

Fit

```
fit <- sem(model, data=d)
```

Statistics

```
summary(fit, standardized=TRUE)
```

Plot

```
semPaths(fit, 'std')
```

Inferring causal graphs

PC algorithm (Sprites et al., 2000; Kalisch et al., 2012)

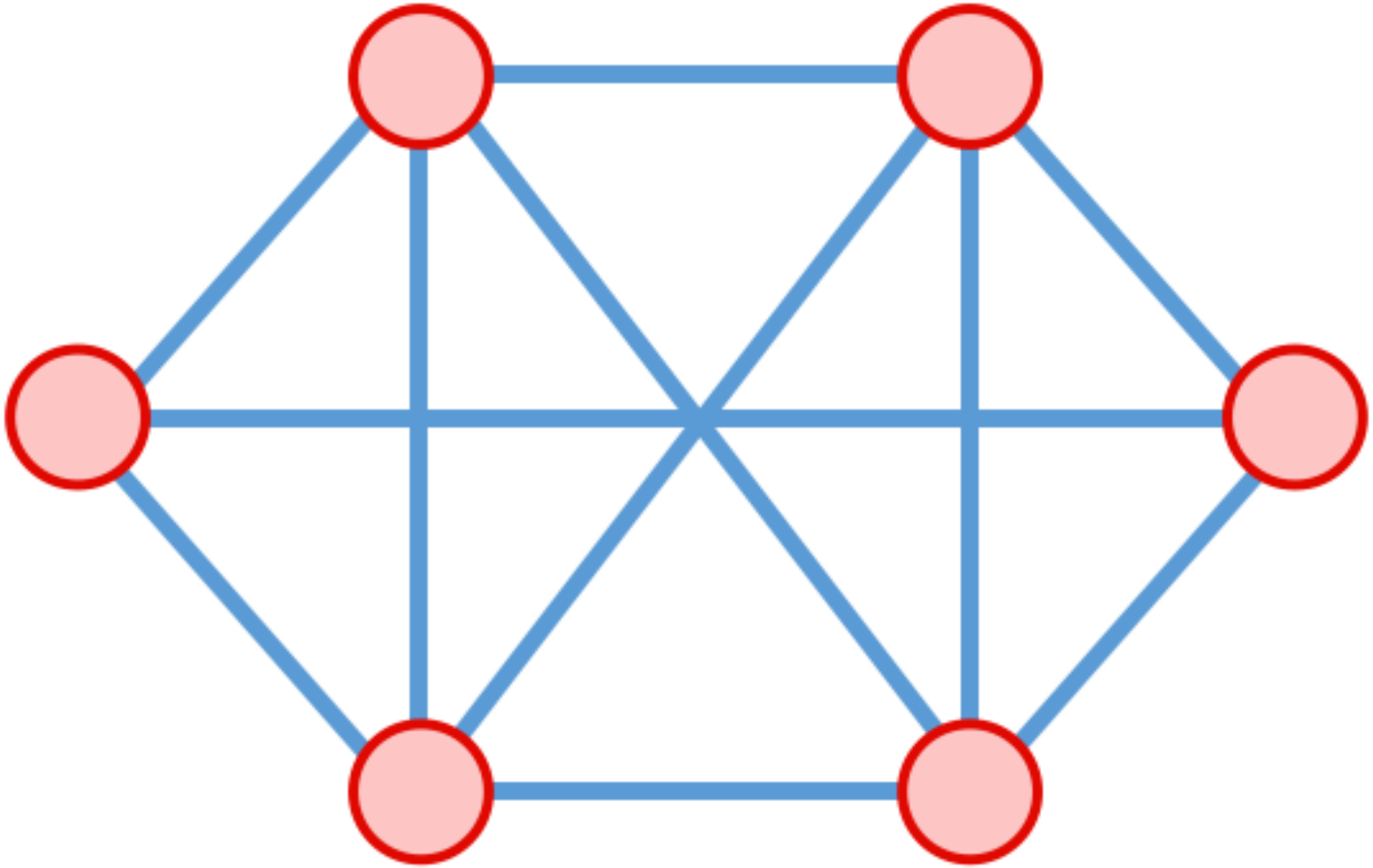
Start with fully connected graph

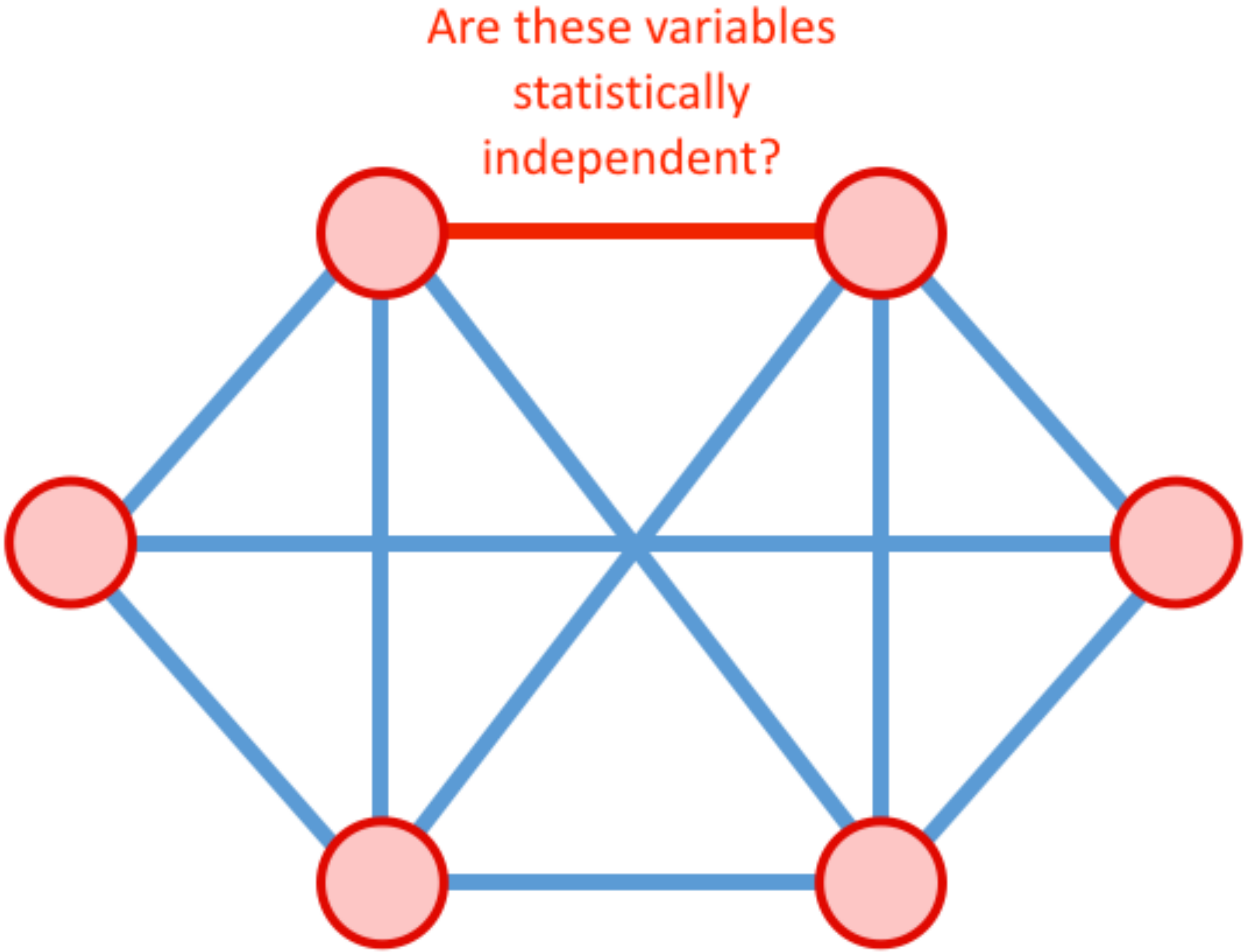
For each pair of variables:

- Try to find evidence that the variables are independent:
  - no correlation,
  - or correlation is explained by a set of other variables
- Any statistical test can be used (e.g. conditional independence)
- If variables are independent, remove the edge.

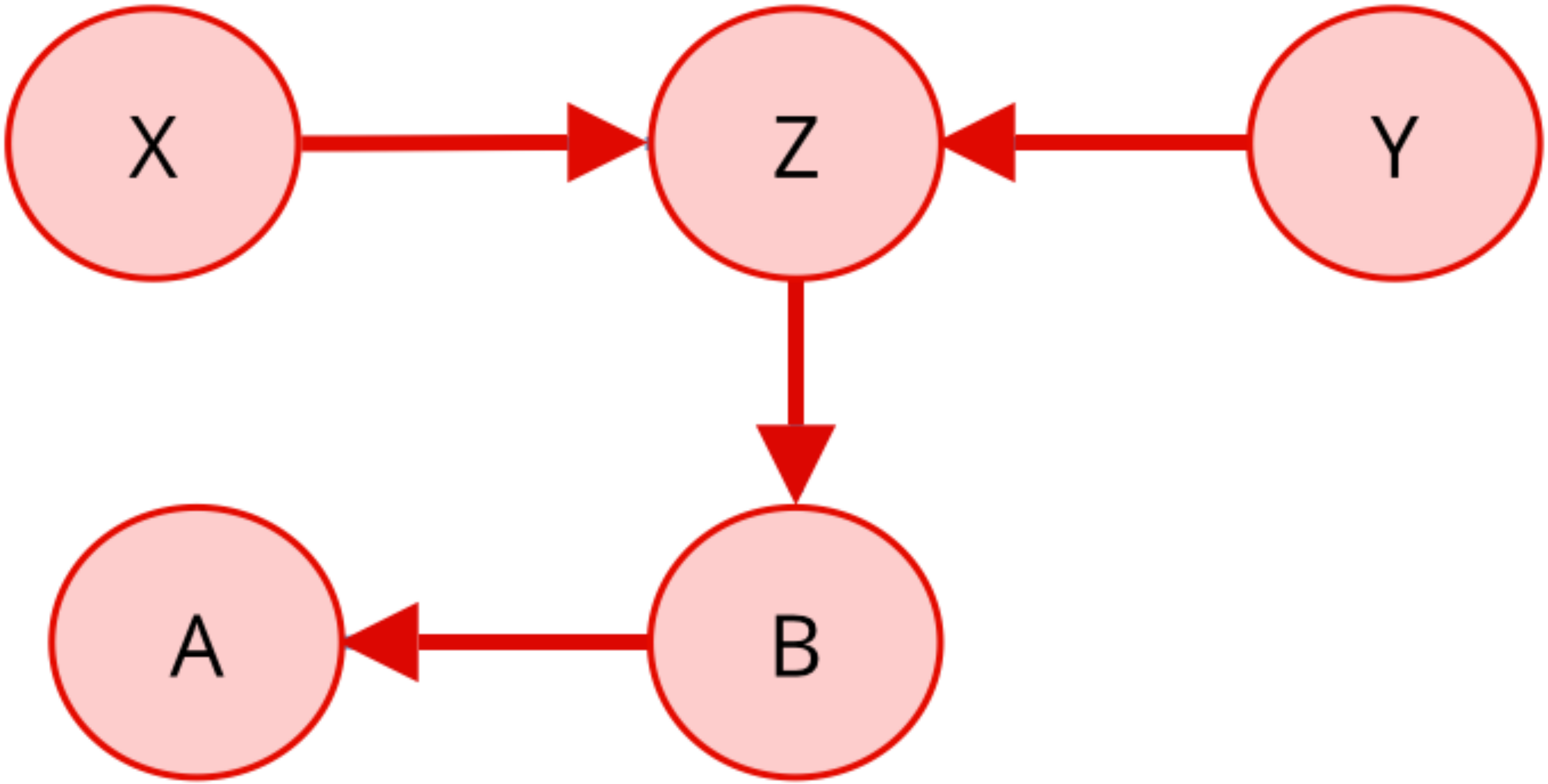
The PC algorithm is an efficient way of performing only the tests which need to be done.

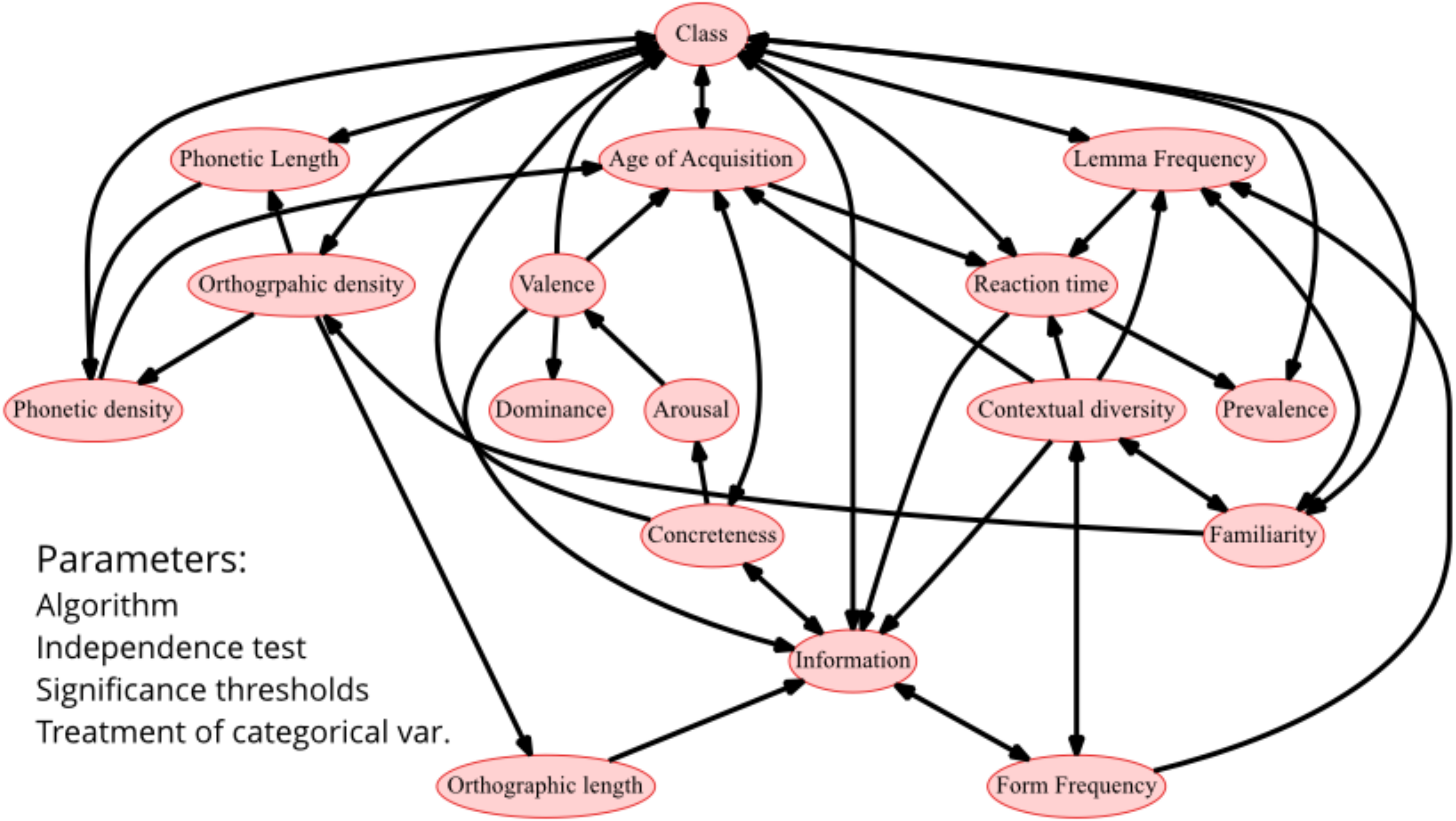
Results in a 'skeleton' graph




Are these variables
statistically
independent?

Orienting the edges





Parameters:
Algorithm
Independence test
Significance thresholds
Treatment of categorical var.

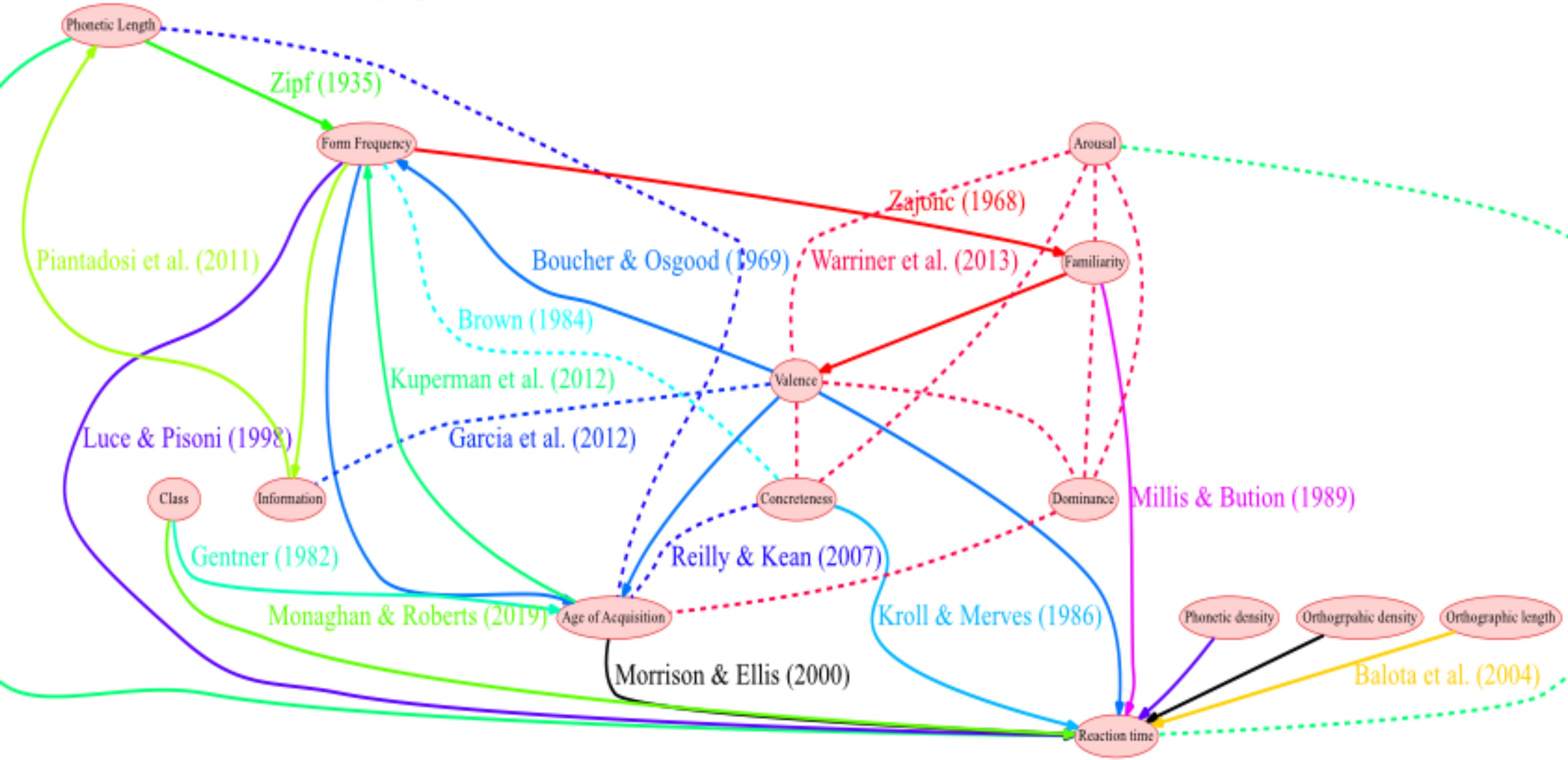
DANGER

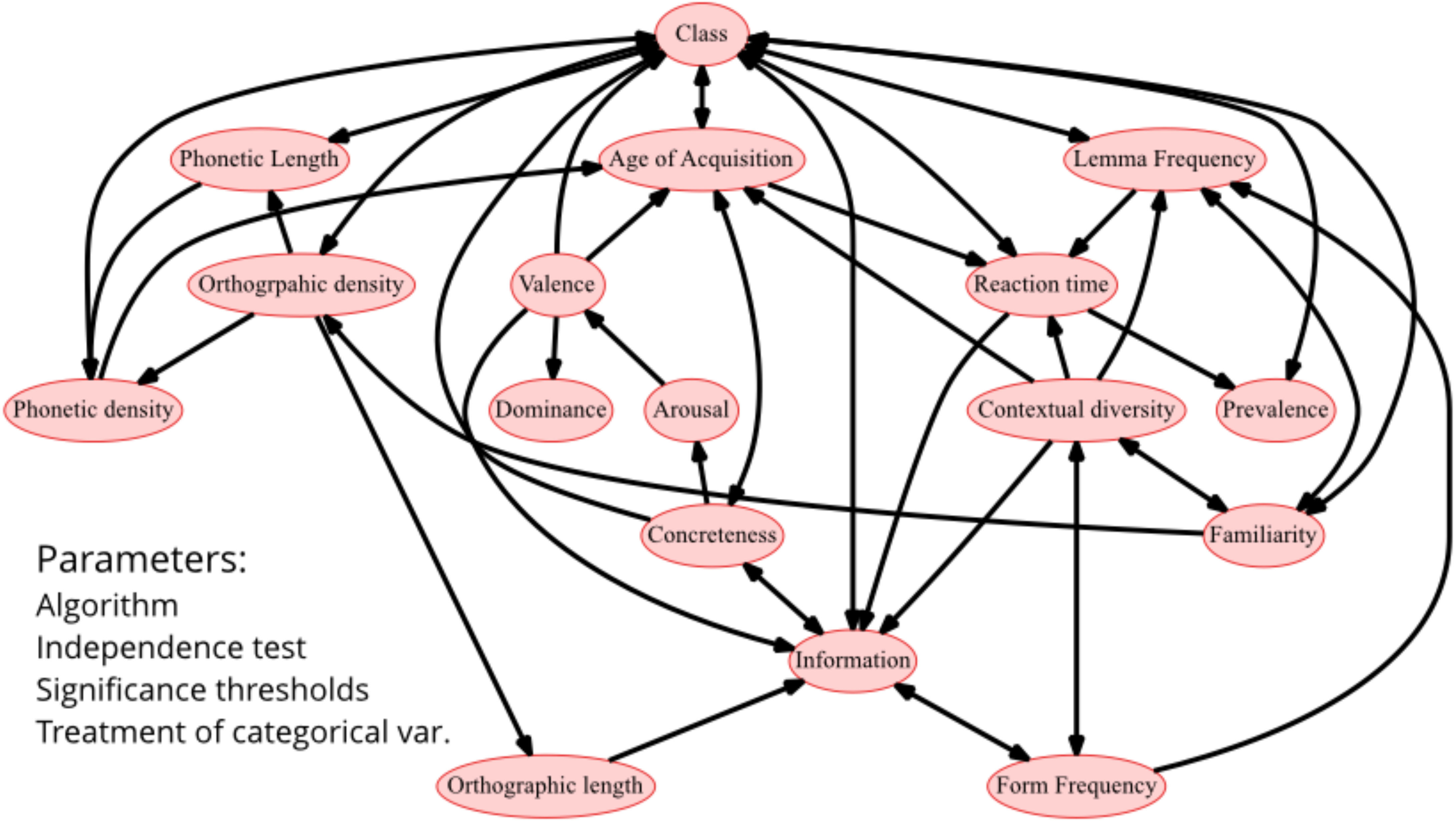
Causal discovery is tricky

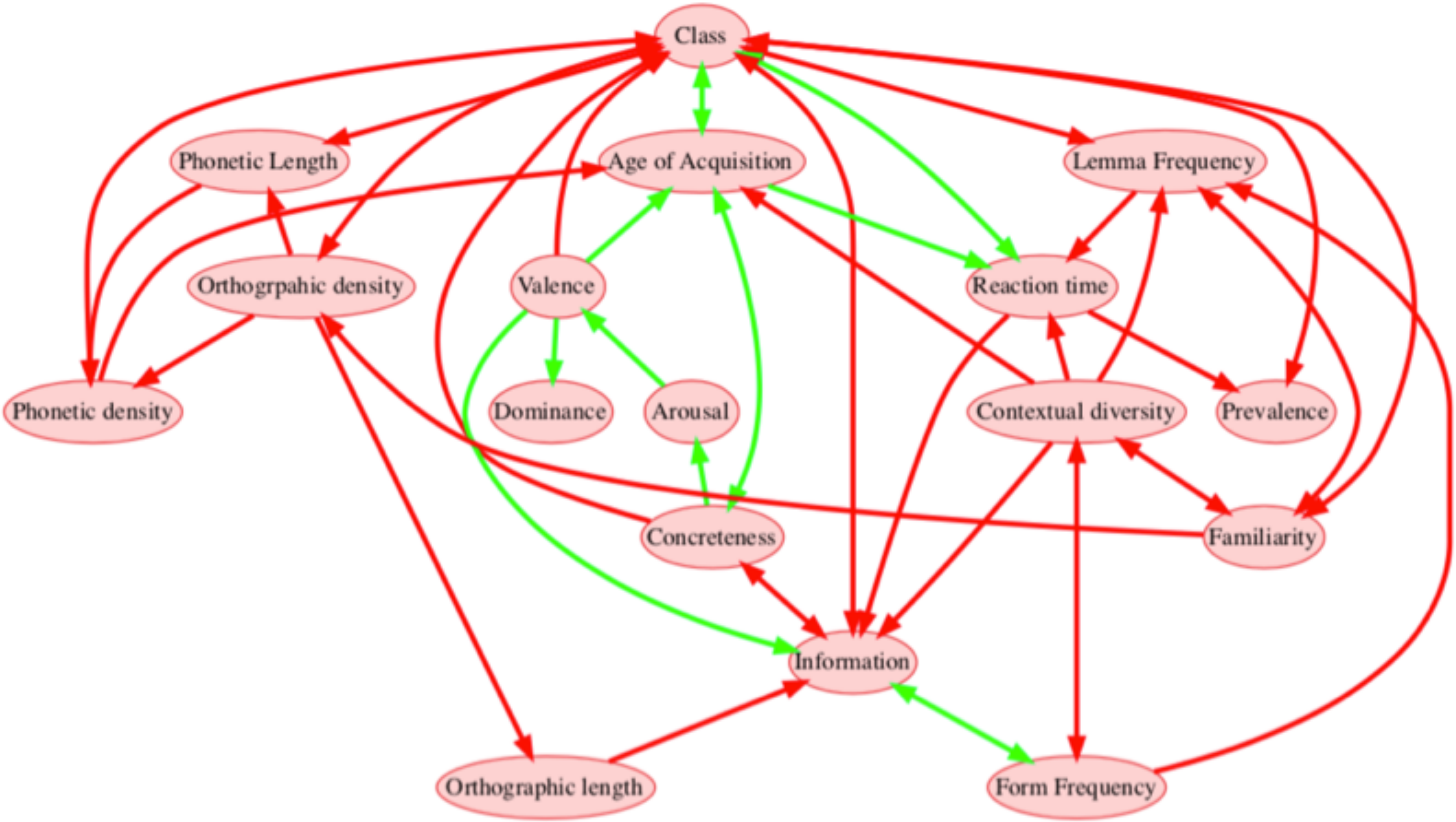
- Requires a lot of data
- Orientation of edges is not robust
- It’s easy to justify patterns

How can you protect yourself from ad-hoc storytelling?

Hypotheses about the lexicon



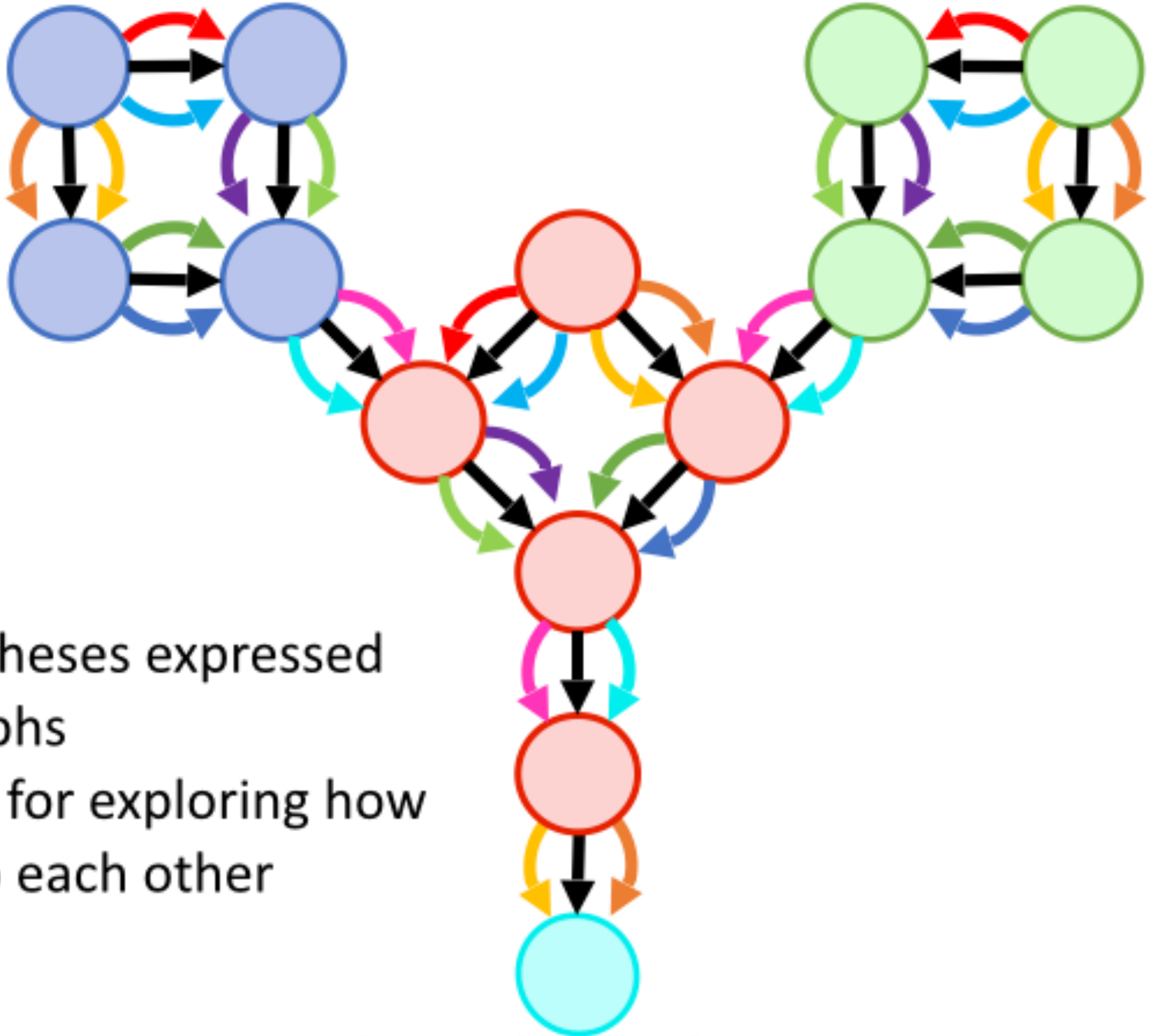

Class
Phonetic Length
Age of Acquisition
Lemma Frequency
Orthogrpahic density
Valence
Reaction time
Phonetic density
Dominance
Arousal
Contextual diversity
Prevalence
Concreteness
Familiarity
Parameters:
Algorithm
Independence test
Significance thresholds
Treatment of categorical var.
Information
Orthographic length
Form Frequency


Class
Phonetic Length
Age of Acquisition
Lemma Frequency
Orthogrpahic density
Valence
Reaction time
Phonetic density
Dominance
Arousal
Contextual diversity
Prevalence
Concreteness
Familiarity
Information
Orthographic length
Form Frequency

CHIELD

Causal Hypotheses in Evolutionary Linguistics Database

https://chield.excd.org

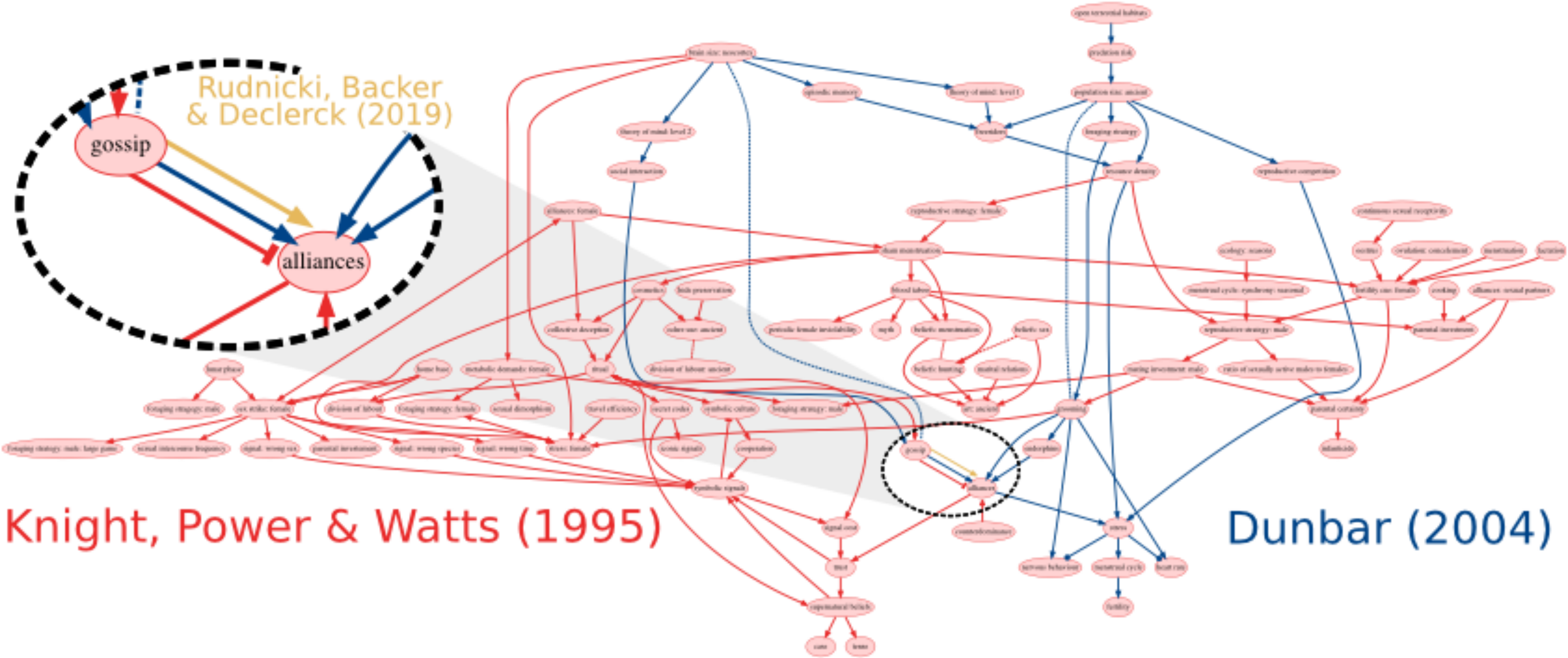


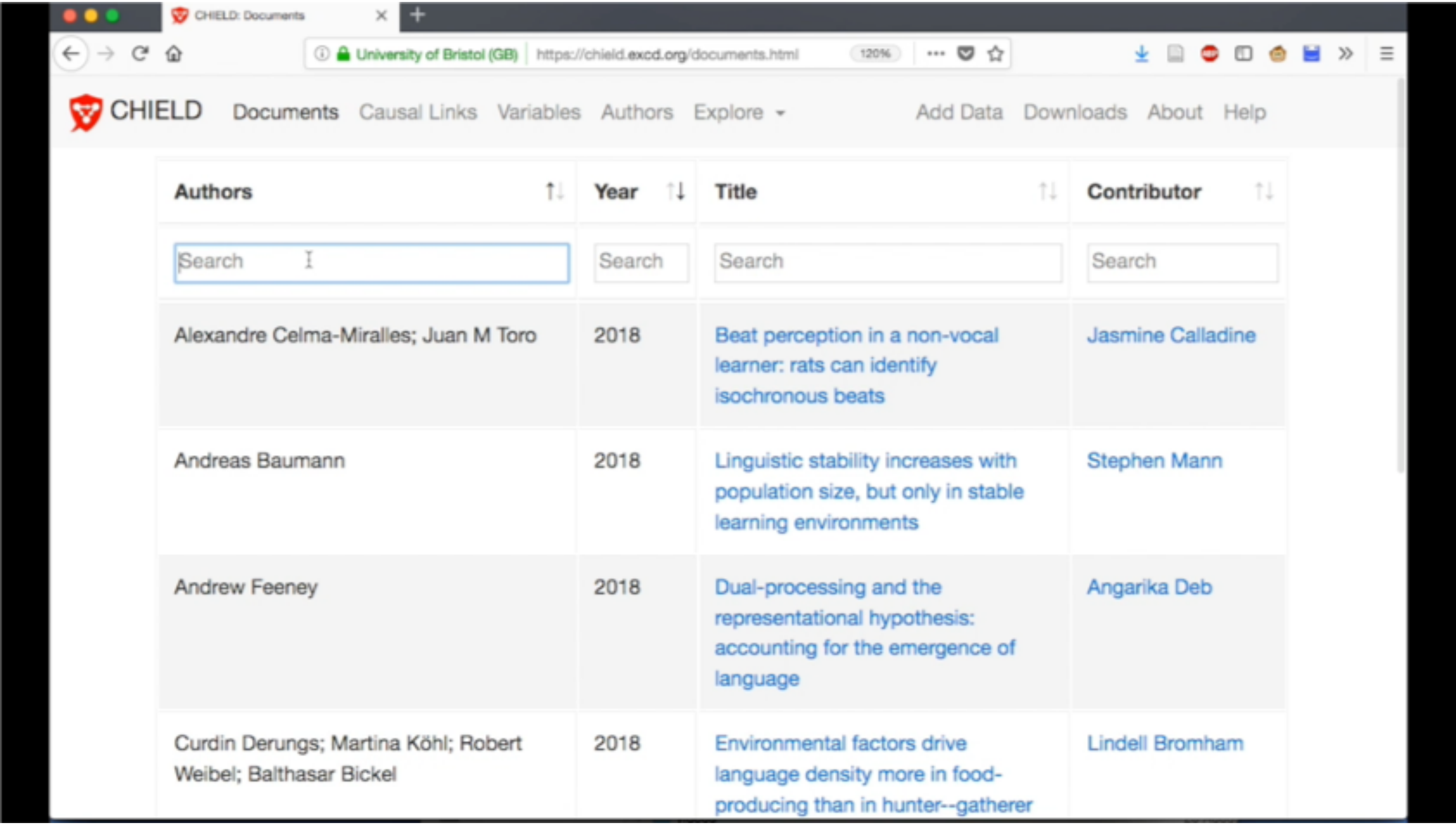
Aims:

- Collect hypotheses expressed as causal graphs
- Provide tools for exploring how they relate to each other

https://chield.excd.org

Compare theories, find critical differences and tests of those critical differences

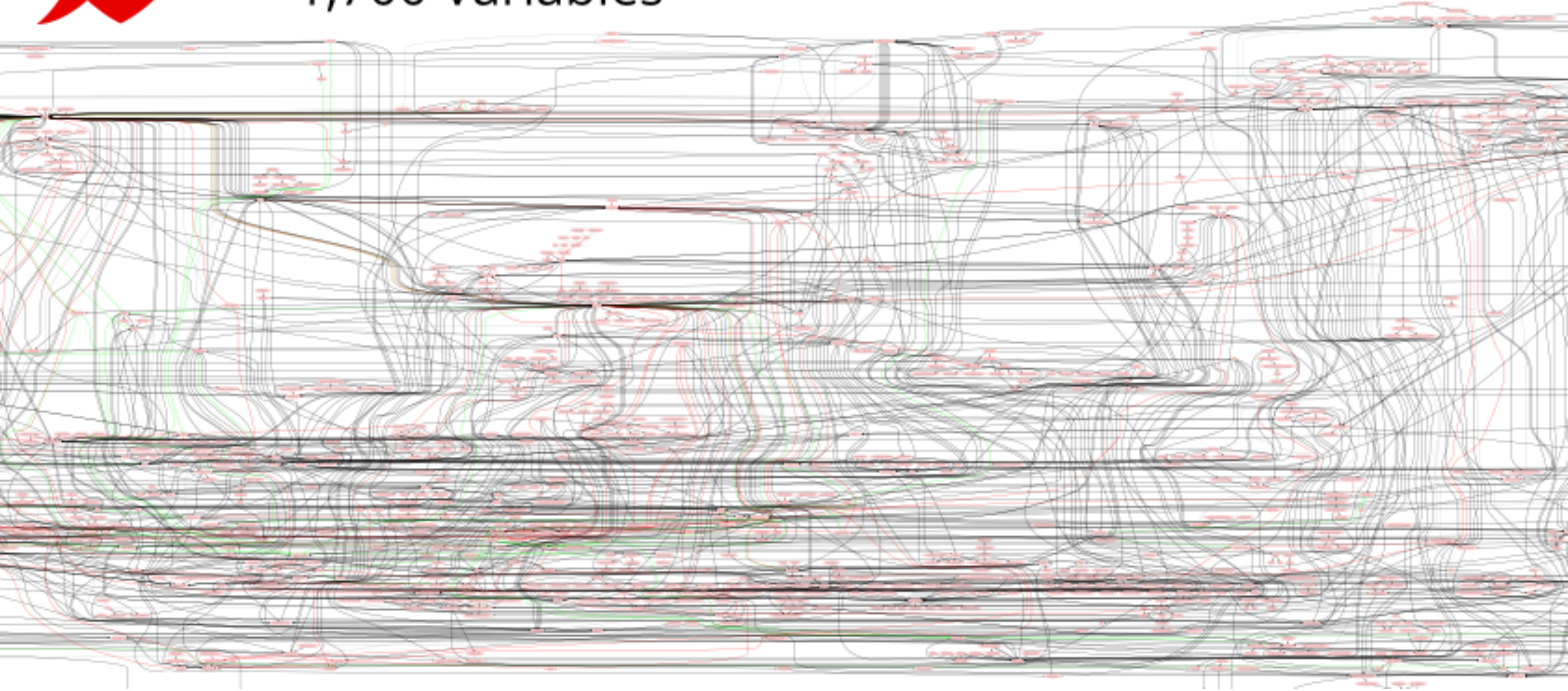



CHIELD: Documents
University of Bristol (GB)
https://chield.excd.org/documents.html
120%
CHIELD
Documents
Causal Links
Variables
Authors
Explore
Add Data
Downloads
About
Help
Authors
Year
Title
Contributor
Search
Search
Search
Search
Alexandre Celma-Miralles; Juan M Toro
2018
Beat perception in a non-vocal learner: rats can identify isochronous beats
Jasmine Calladine
Andreas Baumann
2018
Linguistic stability increases with population size, but only in stable learning environments
Stephen Mann
Andrew Feeney
2018
Dual-processing and the representational hypothesis: accounting for the emergence of language
Angarika Deb
Curdin Derungs; Martina Köhl; Robert Weibel; Balthasar Bickel
2018
Environmental factors drive language density more in food-producing than in hunter--gatherer
Lindell Bromham

400 Documents
3,400 Links
1,700 Variables

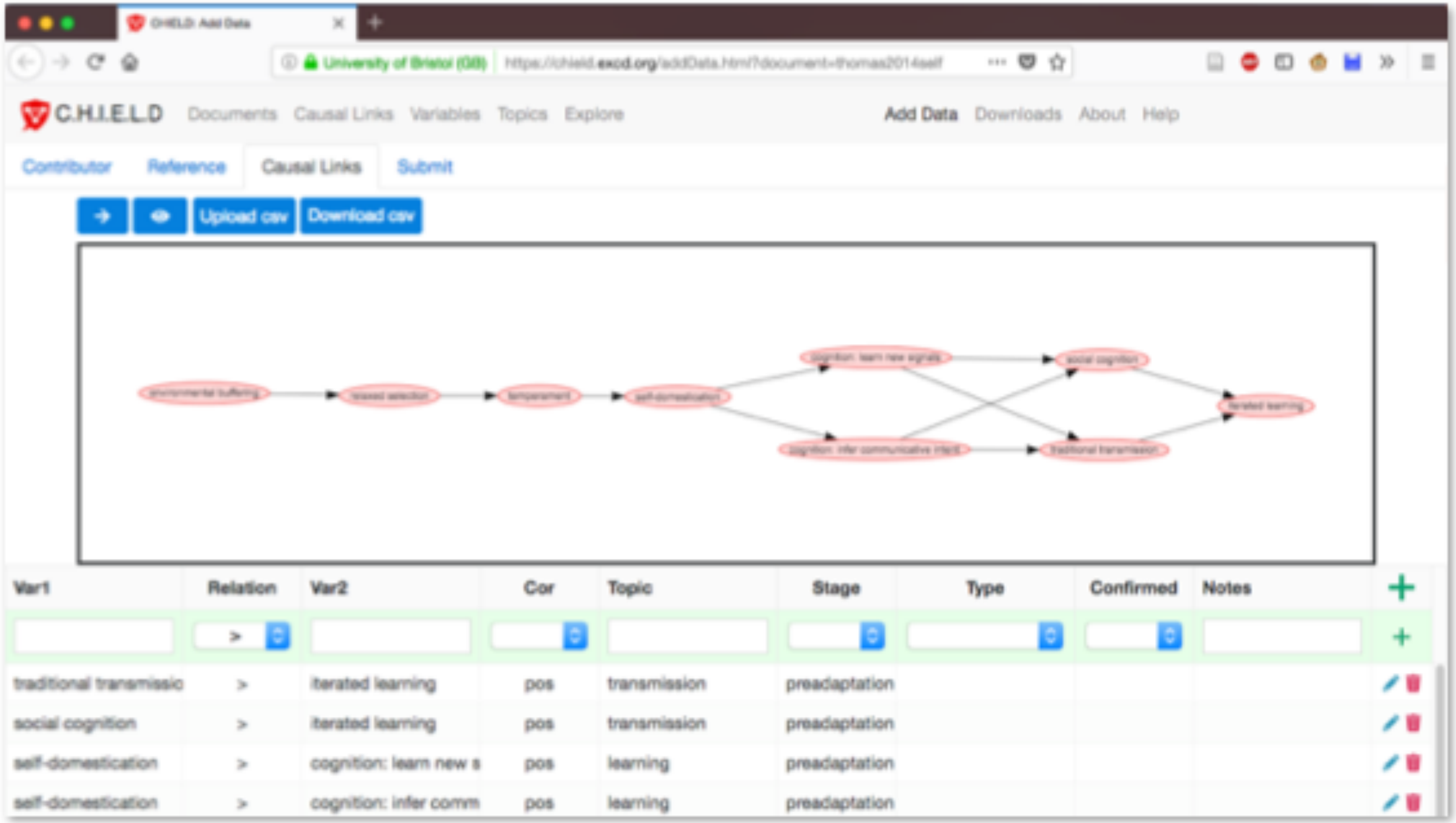
Database is live

https://chield.excd.org



Exporting

Contribute, Discuss, Edit



https://chield.excd.org

Problems

Controlling for some things removes spurious correlations

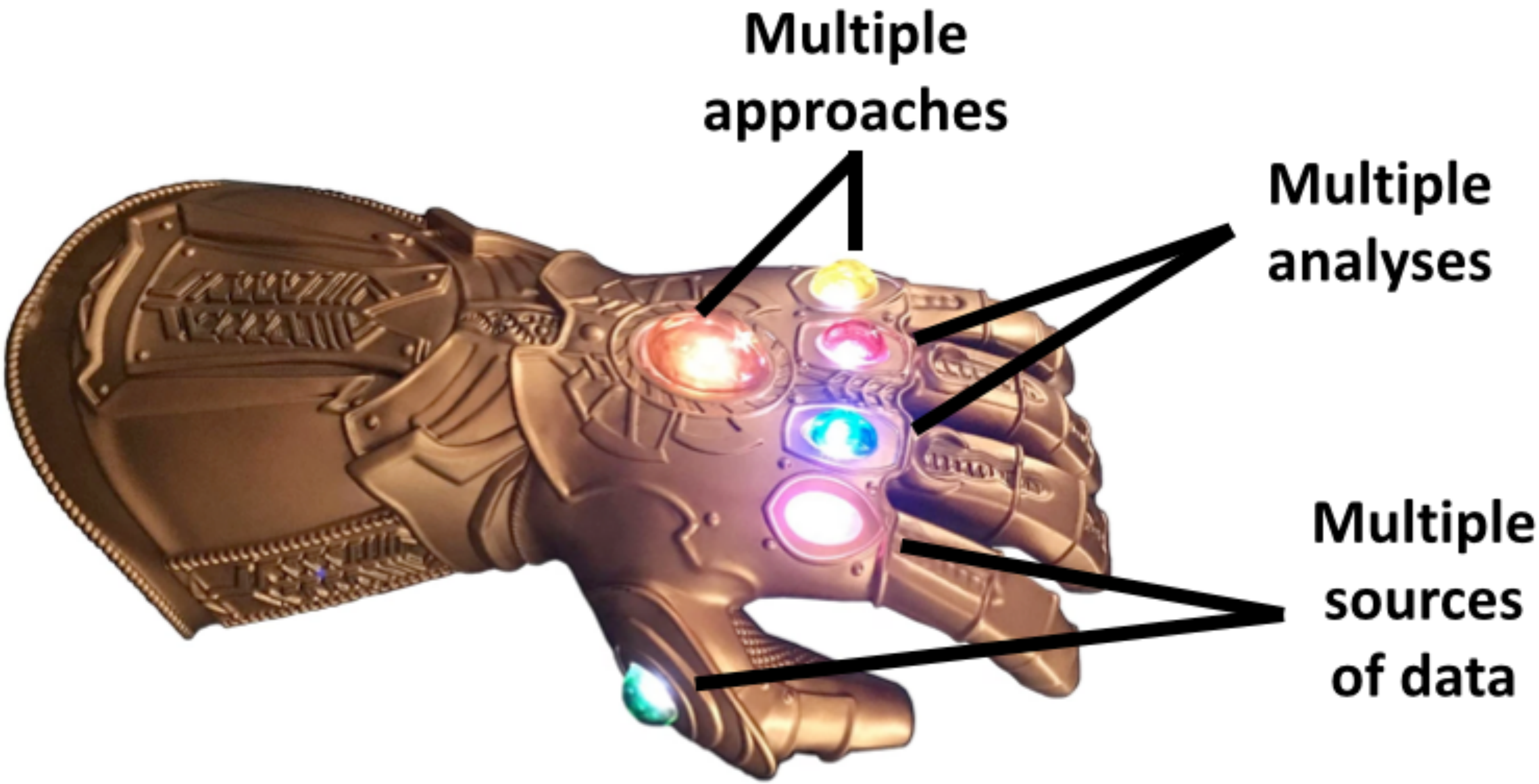
Controlling for other things **creates** spurious correlations

Solution: Build good causal models

Solution 1: Collaboration

- Best data
- Best methods
- Best theories
- More fun

Solution 2: Robust approach to explanation



Solution 3: Incremental approach

- Modest goals for papers
- Modest interpretations

Solution 4: Causal graphs

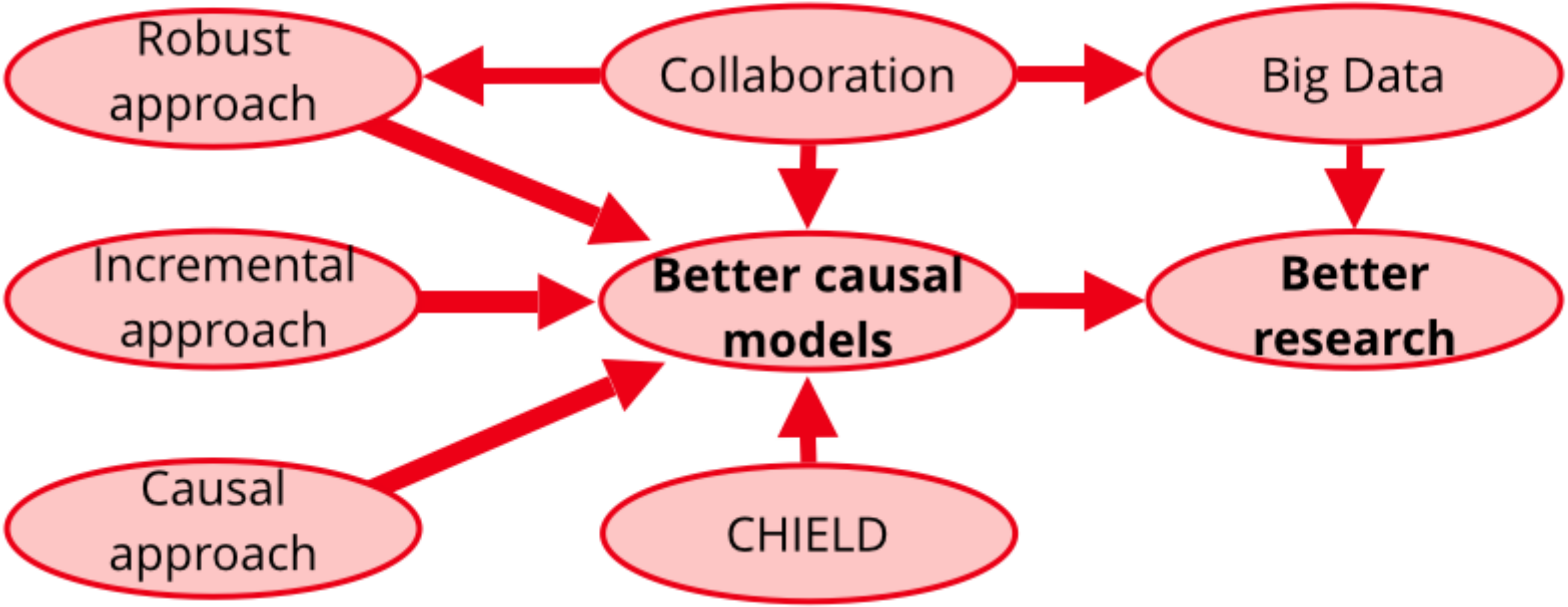
Causal graphs help us to:

- Be more explicit about our hypotheses
- Identify confounds
- Focus on critical differences between hypotheses
- Identify connections between theories

Conclusion

Big data brings many opportunities and challenges

Meeting them will require building better causal models



Recommended reading

- Rohrer, J. M. (2017). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data.
- Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic Books.
- Roberts, S. (2018). Robust, causal and incremental approaches to investigating linguistic adaptation. Frontiers in Psychology, 9, 166.

References

Bareinboim, E., & Pearl, J. (2012). Transportability of causal effects: Completeness results. In Twenty-Sixth AAAI Conference on Artificial Intelligence.

Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic Books.

Shrier, I. and Platt, R. W. (2008) . Reducing bias through directed acyclic graphs. BMC Medical Research Methodology, 8(70), 2008.

Shpitser, I. (2008). Complete identification methods for causal inference (Doctoral dissertation, UCLA).

Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. Journal of Machine Learning Research, 9(Sep), 1941-1979.

Textor, J., Benito van der Zander, Mark K. Gilthorpe, Maciej Liskiewicz, George T.H. Ellison. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. International Journal of Epidemiology 45(6):1887-1894, 2016.

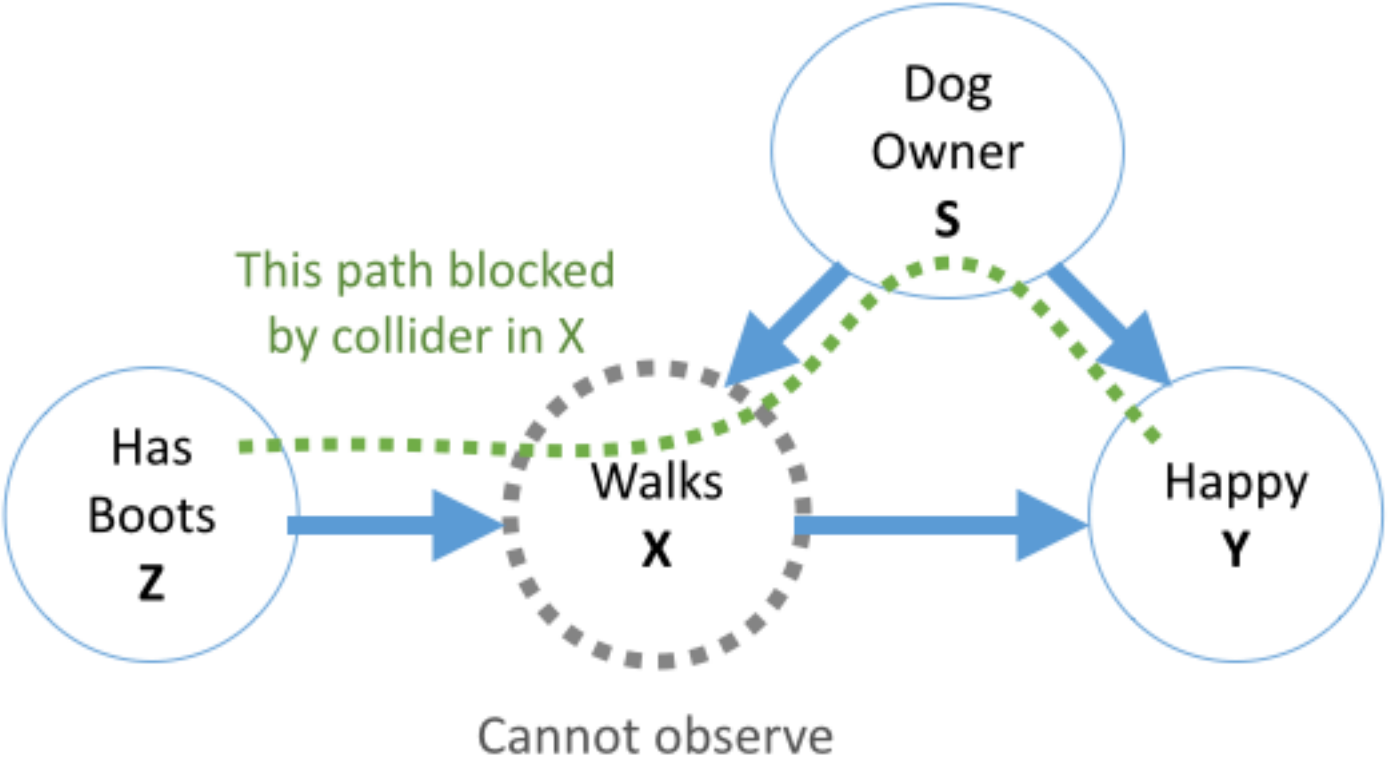
Instrumental variables

Variable Z is an instrument of X for a dependent variable Y if

Z causally influences X,

There are no unblocked alternative paths from Z to other determiners of Y (S)

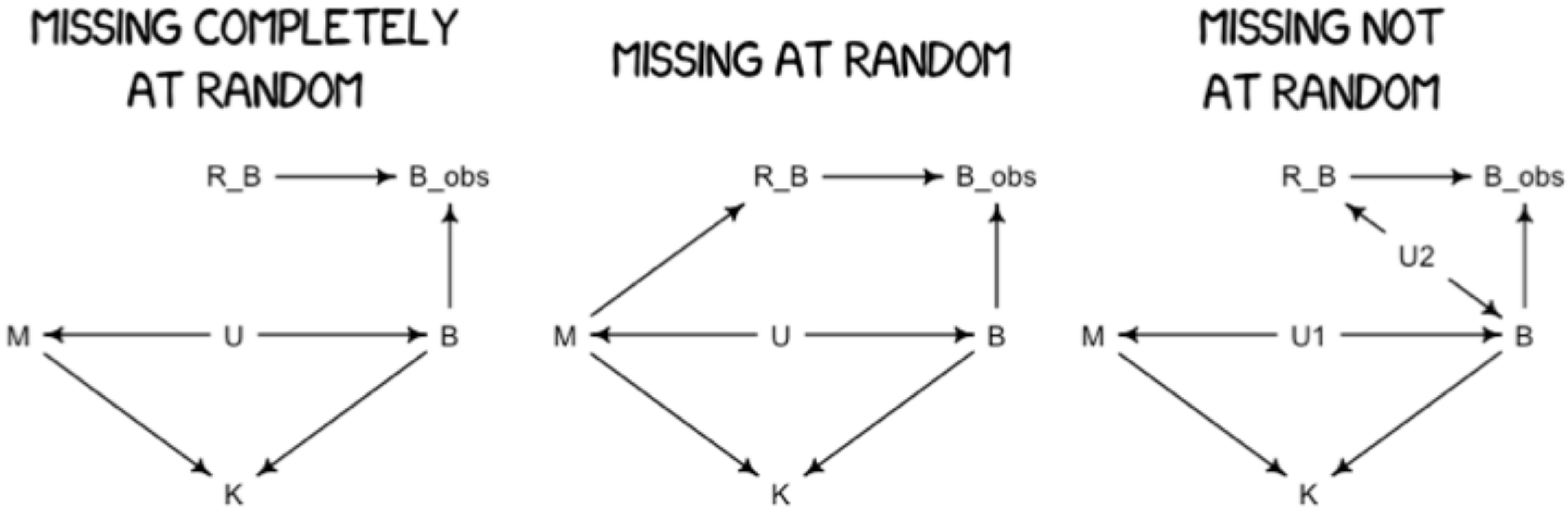
(Z causally influences Y, but only through X)



Z is correlated with Y,
but the only source of the
correlation is through X

Z simulates an intervention

Data imputation



McElreath (in prep)