But, why?

an introduction to causal inference from observational data

Jilles Vreeken

13 September 2019

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Questions of the day

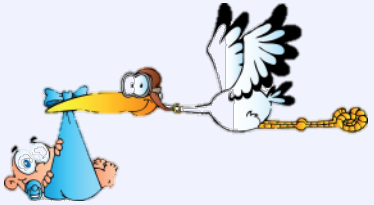
What is **causation**,
how can we **measure** it,
and how can **discover** it?

Causation

'the relationship between something that happens or exists and the thing that causes it'

(Merriam-Webster)

Correlation vs. Causation



Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

Correlation does not tell us anything about **causality**

Instead, we should talk about **dependence**.

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

Dependence vs. Causation

What is causal inference?

*'reasoning to the conclusion that something is, or is **likely** to be, the cause of something else'*

Godzillian different definitions of 'cause' and 'effect'

- equally many inference frameworks
- all require (strong) assumptions
- many highly specific

Causal Inference

Naïve approach

If

$$p(cause)p(\,effect \mid cause\,) > p(effect)p(\,cause \mid effect\,)$$

then $cause \rightarrow effect$

Naïve approach fails

Both are equal as they are simply factorizations of $p(cause, effect)$

~~If~~

$$p(cause)p(\,effect \mid cause\,) = p(effect)p(\,cause \mid effect\,)$$

~~then *cause → effect*~~

Naïve approach

If

$$p(cause)p(effect \mid cause) > p(effect)p(cause \mid effect)$$

then $cause \rightarrow effect$

Naïve approach fails

Depends on distribution and domain size of data, not on causal effect

~~If~~

$$p(cause)p(effect \mid cause) <> p(effect)p(cause \mid effect)$$

~~then *cause → effect*~~

Naïve approach

If

$$p(cause)p(effect \mid cause) > p(effect)p(cause \mid effect)$$

then $cause \rightarrow effect$

Naïve approach fails

~~If~~

$$\frac{p(\,effect \mid cause\,)}{p(effect)} > \frac{p(cause \mid effect)}{p(cause)}$$

~~then *cause* → *effect*~~

But do we know **for sure** that the lhs is higher when when *cause* → *effect* ?
What about differences in domain sizes, complexities of distributions, etc

The Ultimate Test

Randomized controlled trials are the de-facto standard for determining whether X causes Y

- treatment $X \in \{0,1, \dots\}$, potential effect Y and co-variates Z

Simply put, we

1. gather a **large population** of test subjects
2. **randomly split** the population into two equally sized groups A and B, making sure that Z is **equally distributed** between A and B
3. **apply treatment** $X = 0$ to group A, and treatment $X = 1$ to group B
4. **determine** whether Y and X are dependent

If $Y \not\perp\!\!\!\perp X$, we conclude that X causes Y

The Ultimate Test

Randomized controlled trials are the de-facto standard f

- treatmen

Simply pu

1. gather a
2. **random** ups A and B, making B
3. **apply tre** to group B
4. **determine** whether Y and X are dependent

If $Y \not\perp\!\!\!\perp X$, we conclude that X causes Y

Ultimate, but not ideal

- Often impossible or unethical
- Large populations needed
- Difficult to control for Z

Do, or do not

Observational $p(y \mid x)$

- distribution of Y given that we **observe** variable X takes value x
- what we usually estimate, e.g. in regression or classification
- **can be inferred from data** using Bayes' rule $p(y \mid x) = \frac{p(x,y)}{p(x)}$

Interventional $p(y \mid do(x))$

- distribution of Y given that we **set** the value of variable X to x
- describes the distribution of Y we would observe if we would **intervene** by artificially forcing X to take value x, but otherwise use the original **data generating process** ($\neq p(x, y, \dots)$!)
- the conditional distribution of Y we would get through a randomized control trial!

Same old, same old?

In general, $p(y \mid do(x))$ and $p(y \mid x)$ are **not the same**

Let's consider my espresso machine

- y actual pressure in the boiler
- x pressure measured by front gauge

Now,

- if the barometer works well, $p(y \mid x)$ will be unimodal around x
- intervening on the barometer, e.g. moving its needle up or down, however, has **no effect** on the actual pressure and hence, $p(y \mid do(x)) = p(y)$

What do you want?

Before we go into a lot more detail, what do we want?

If you just want to predict, $p(y \mid x)$ is great

- e.g. when 'interpolating' Y between its cause and its effects is fine
- also, boring, because lots of cool methods exist

If you want to act on x, you really want $p(y \mid do(x))$

- for example, for drug administration, or discovery
- also, exciting, not so many methods exist

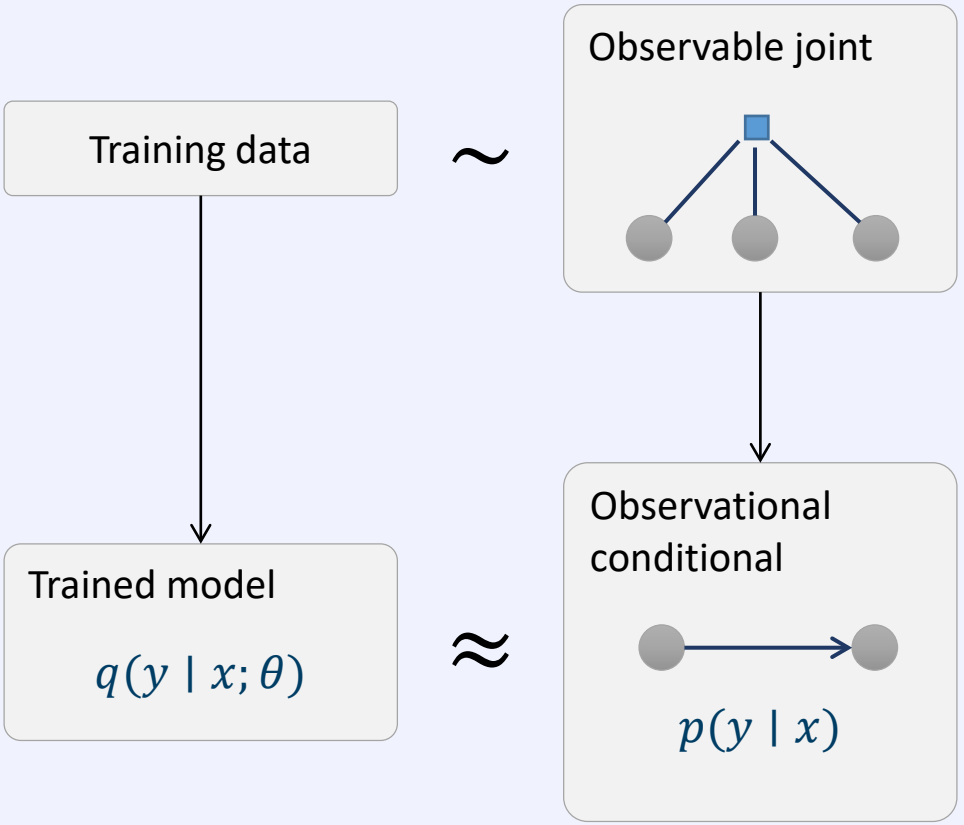
Observational Data

Even if we cannot directly access $p(y \mid do(x))$
e.g. through randomized trials, it does **exist**

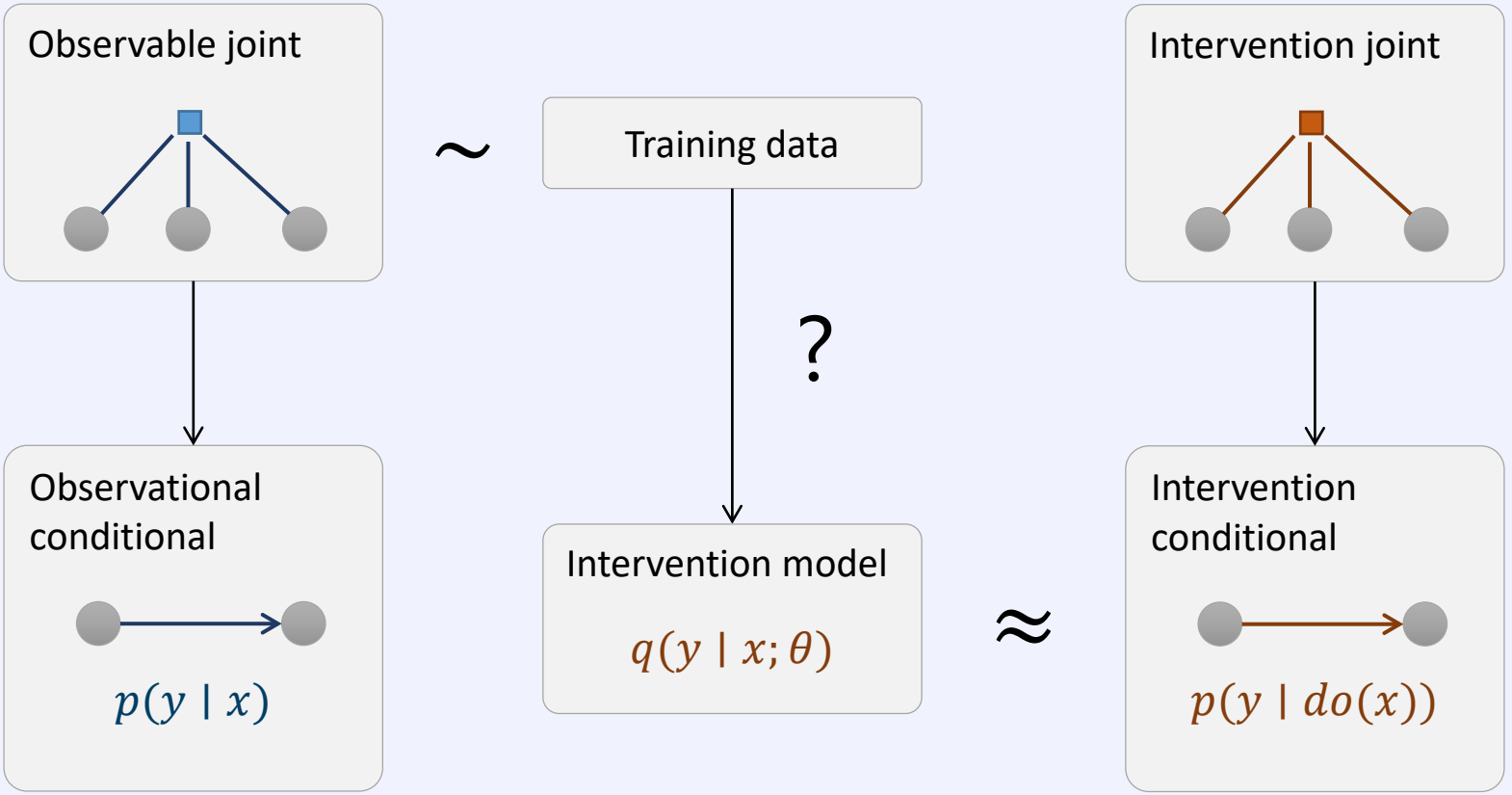
The main point of causal inference and do-calculus is:

If we cannot measure $p(y \mid do(x))$ *directly*
in a randomized trial, can we estimate it
based on data we observed outside
of a controlled experiment?

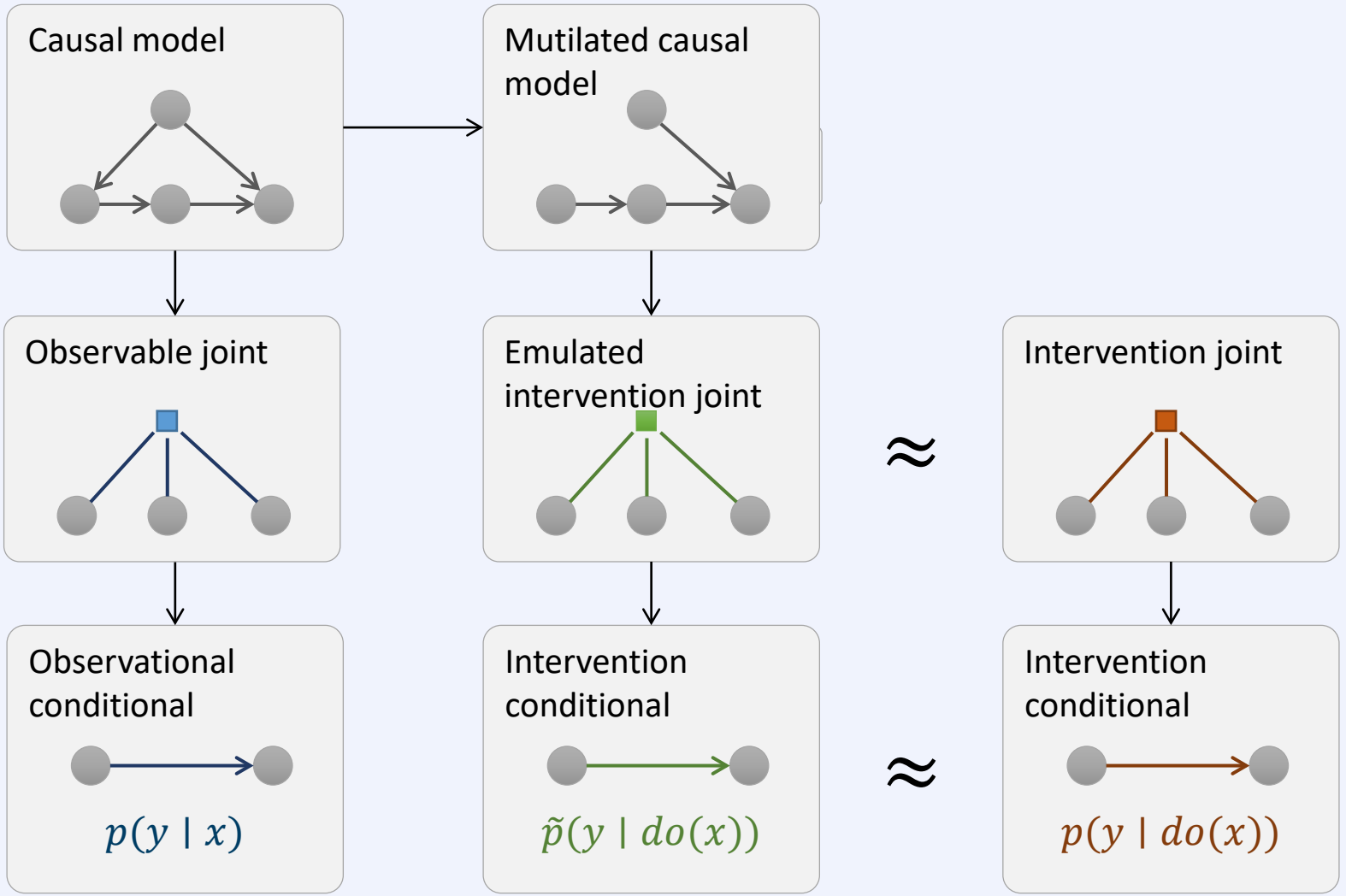
Standard learning setup



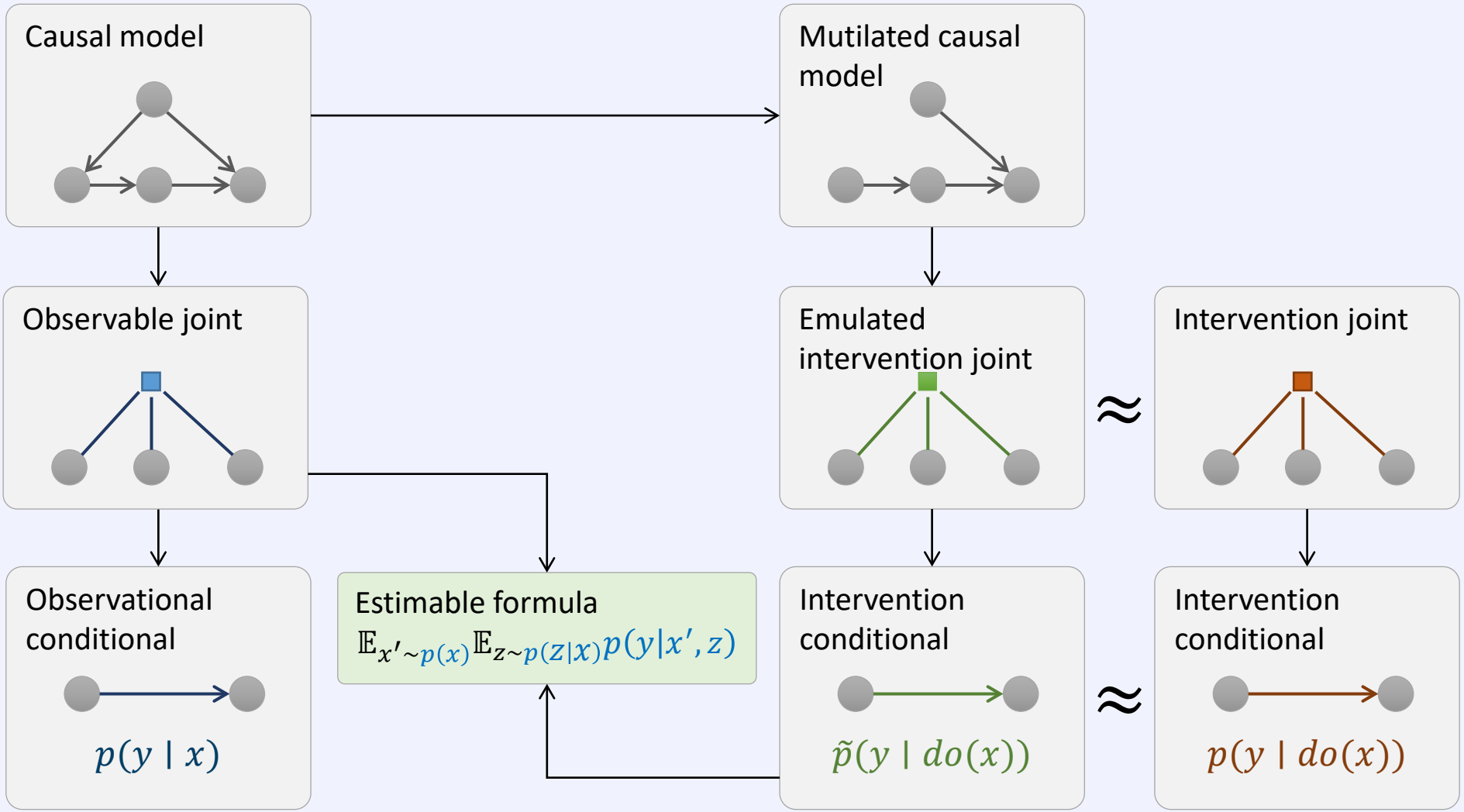
Causal learning setup



Causal learning goal

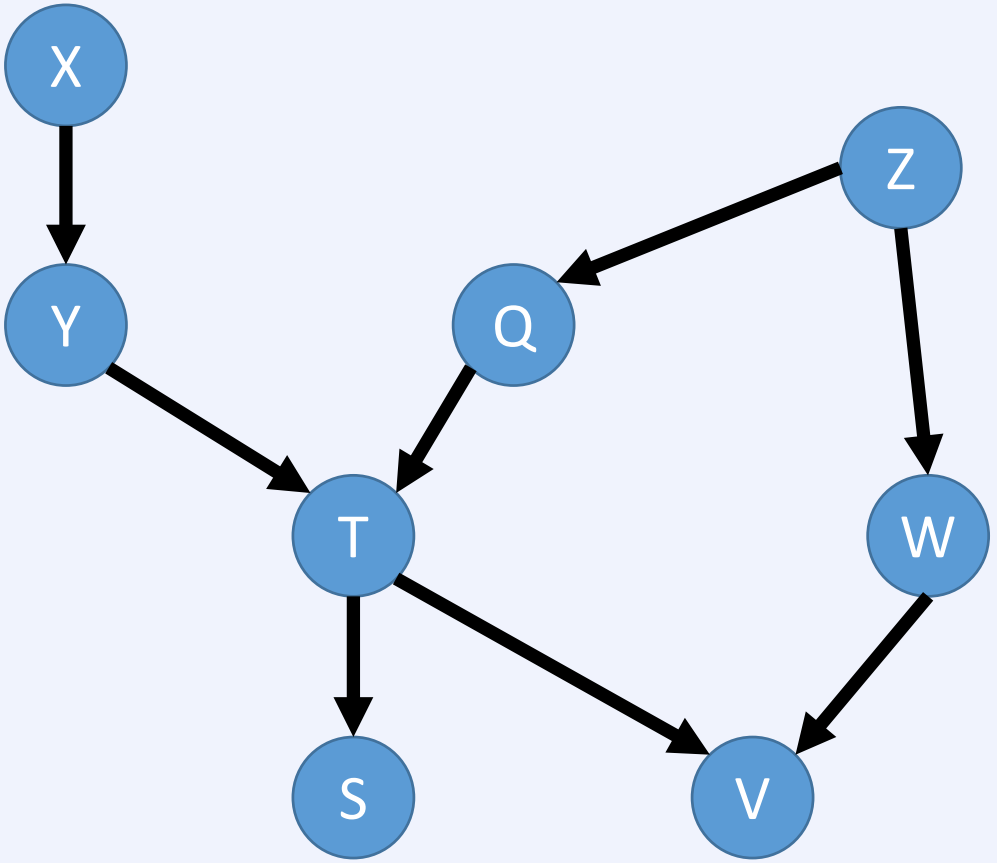


If through **do-calculus** we can derive an equivalent of $\tilde{p}(y \mid do(x))$ without any do's, we can **estimate it from observational data alone** and call $\tilde{p}(y \mid do(x))$ **identifiable.**



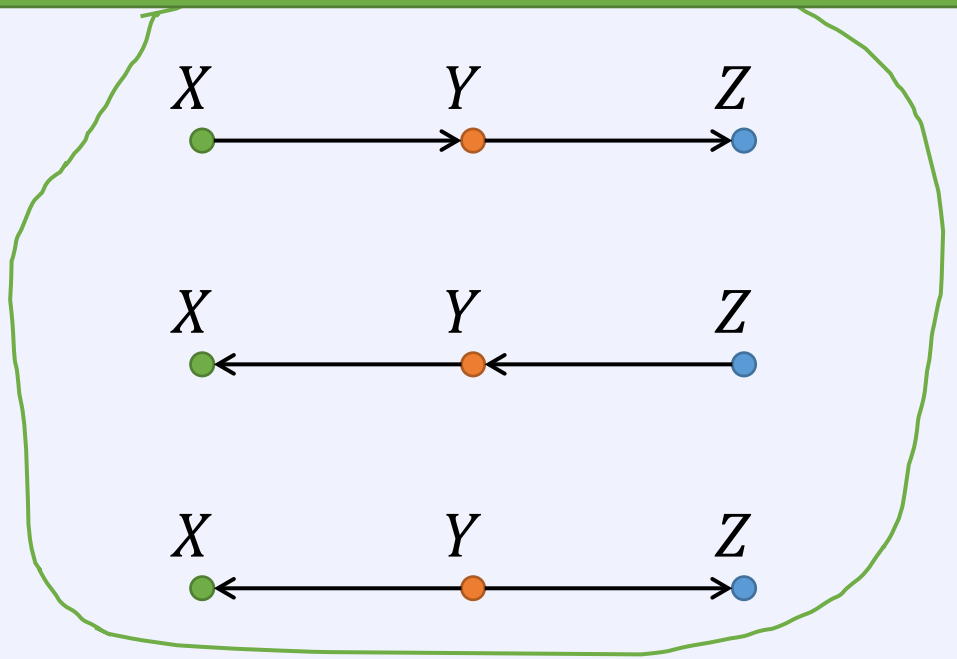
Causal Discovery

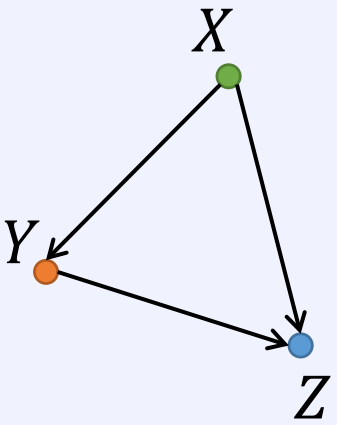
Causal Discovery

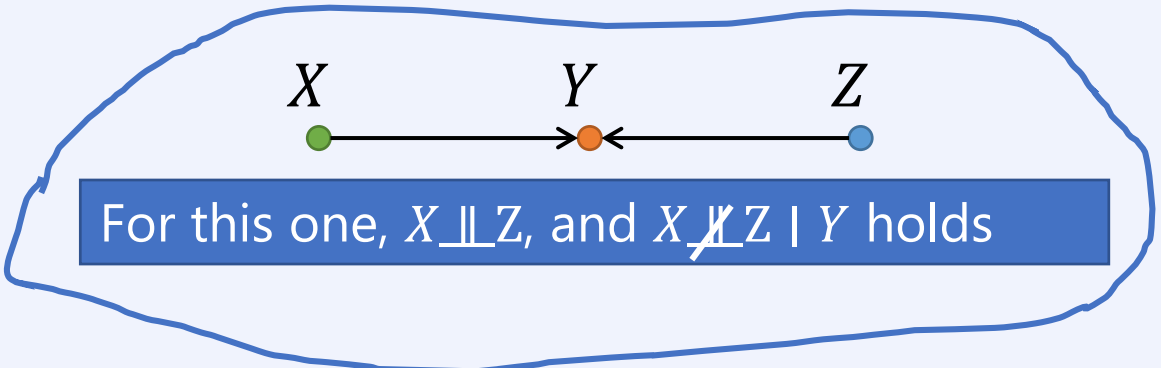


Choices…

For these three, $X \not\perp\!\!\!\perp Z$, and $X \perp\!\!\!\perp Z \mid Y$ holds





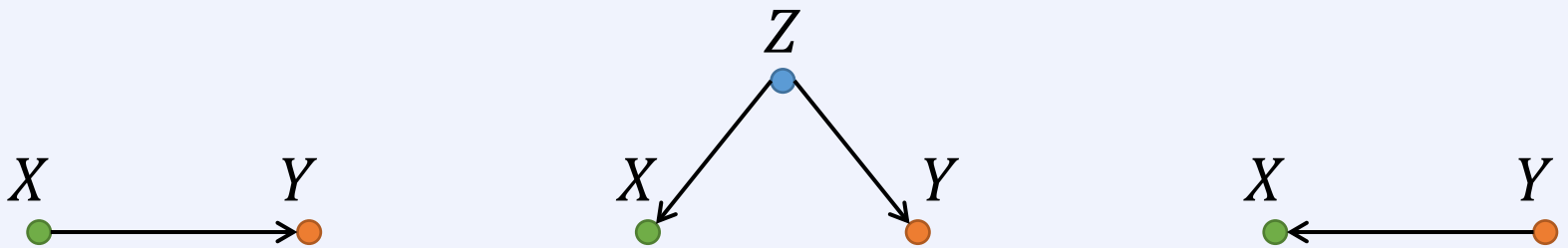


For this one, $X \perp\!\!\!\perp Z$, and $X \not\perp\!\!\!\perp Z \mid Y$ holds

Statistical Causality

Reichenbach's
common cause principle
links causality and probability

if X and Y are statistically dependent then either



When Z **screens** X and Y from each other,
given Z, X and Y become **independent**.

Causal Markov Condition

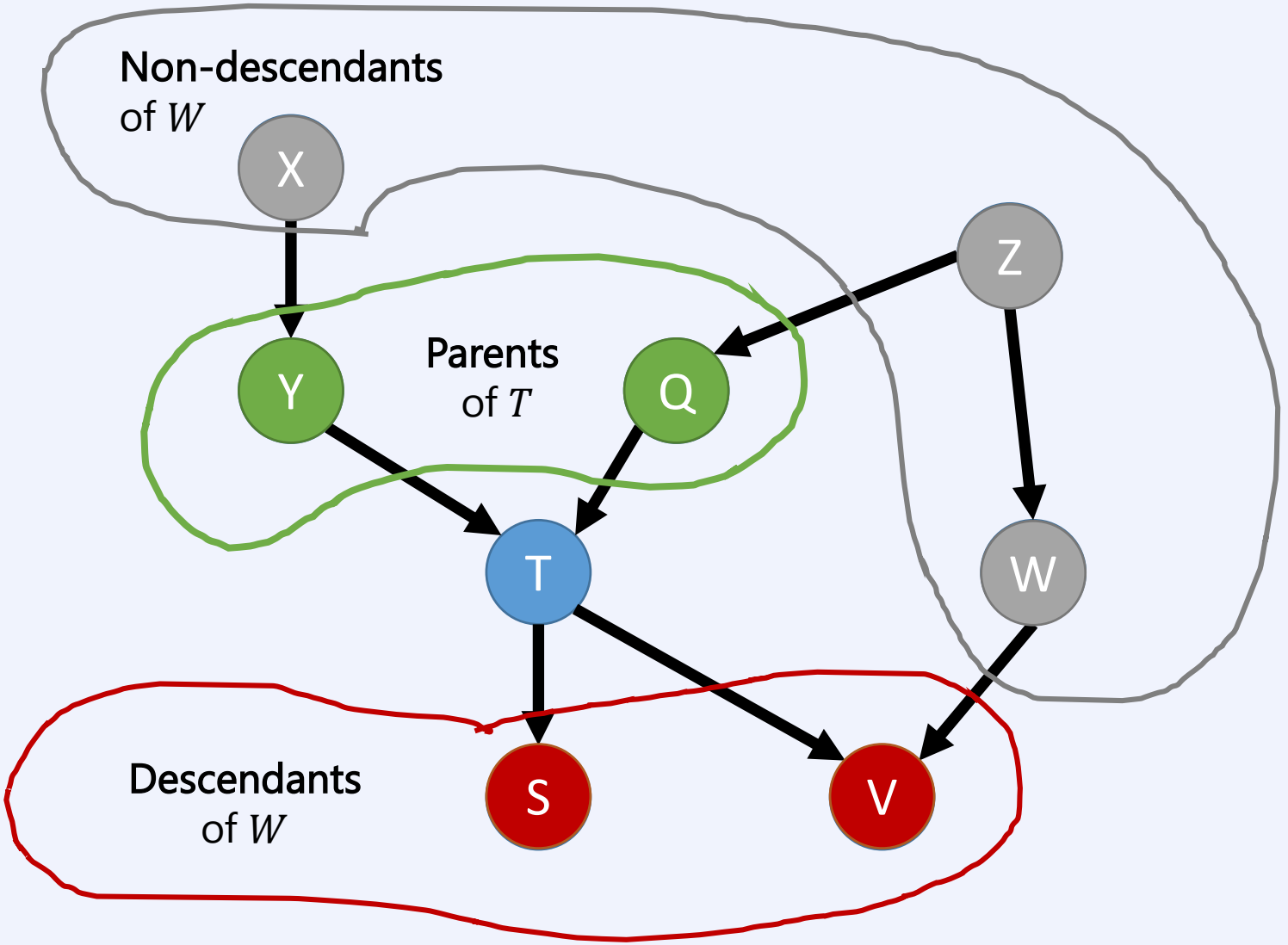
Any distribution generated by a Markovian model M can be factorized as

$$p(X_1, X_2, \ldots, X_n) = \prod_i p(X_i \mid pa_i)$$

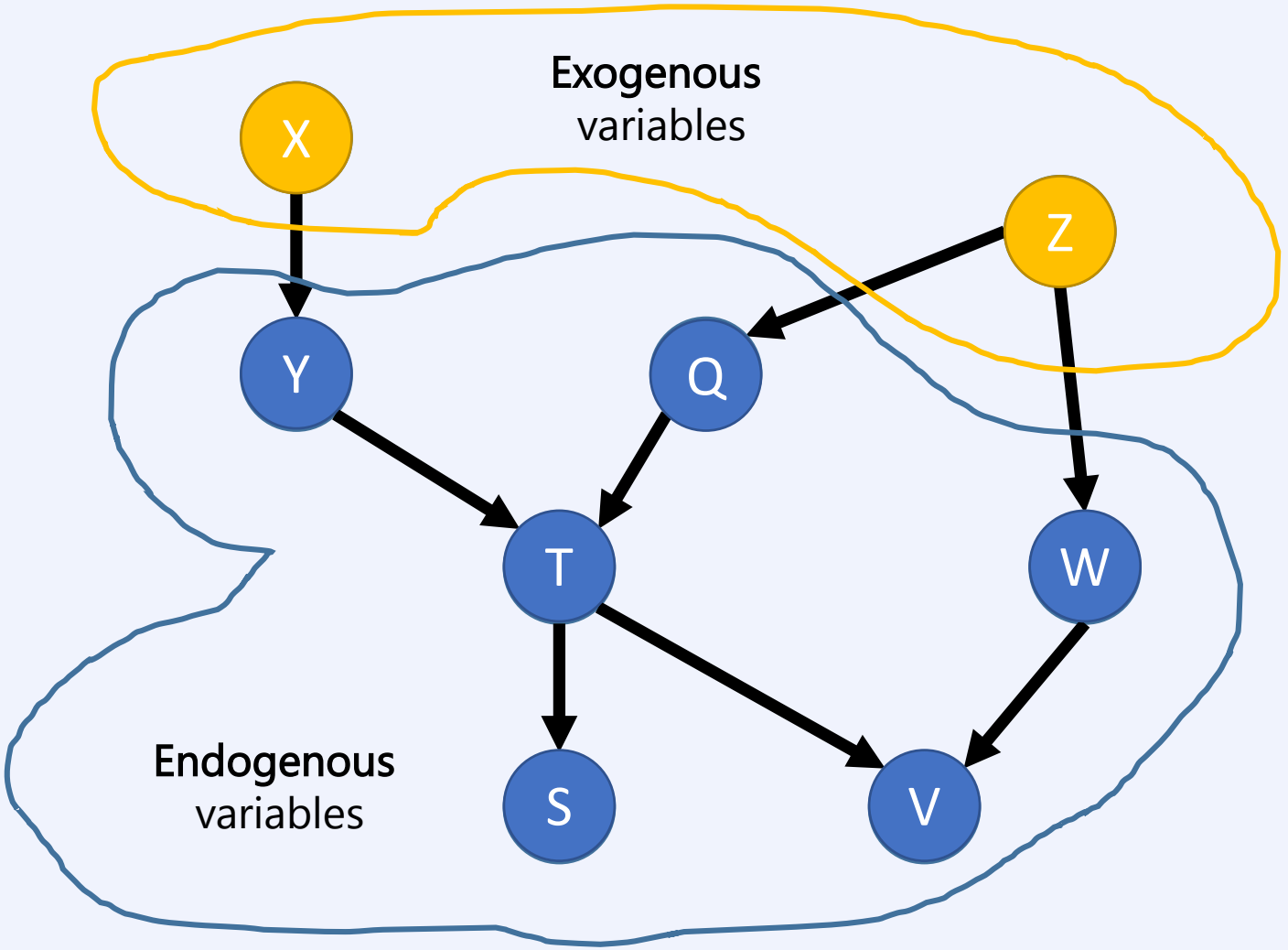
where $X_1, X_2, \ldots, X_n$ are the **endogenous** variables in M, and pa_i are (values of) the **endogenous** "parents" of X_i in the causal diagram associated with M

(Spirtes, Glymour, Scheines 1982; Pearl 2009)

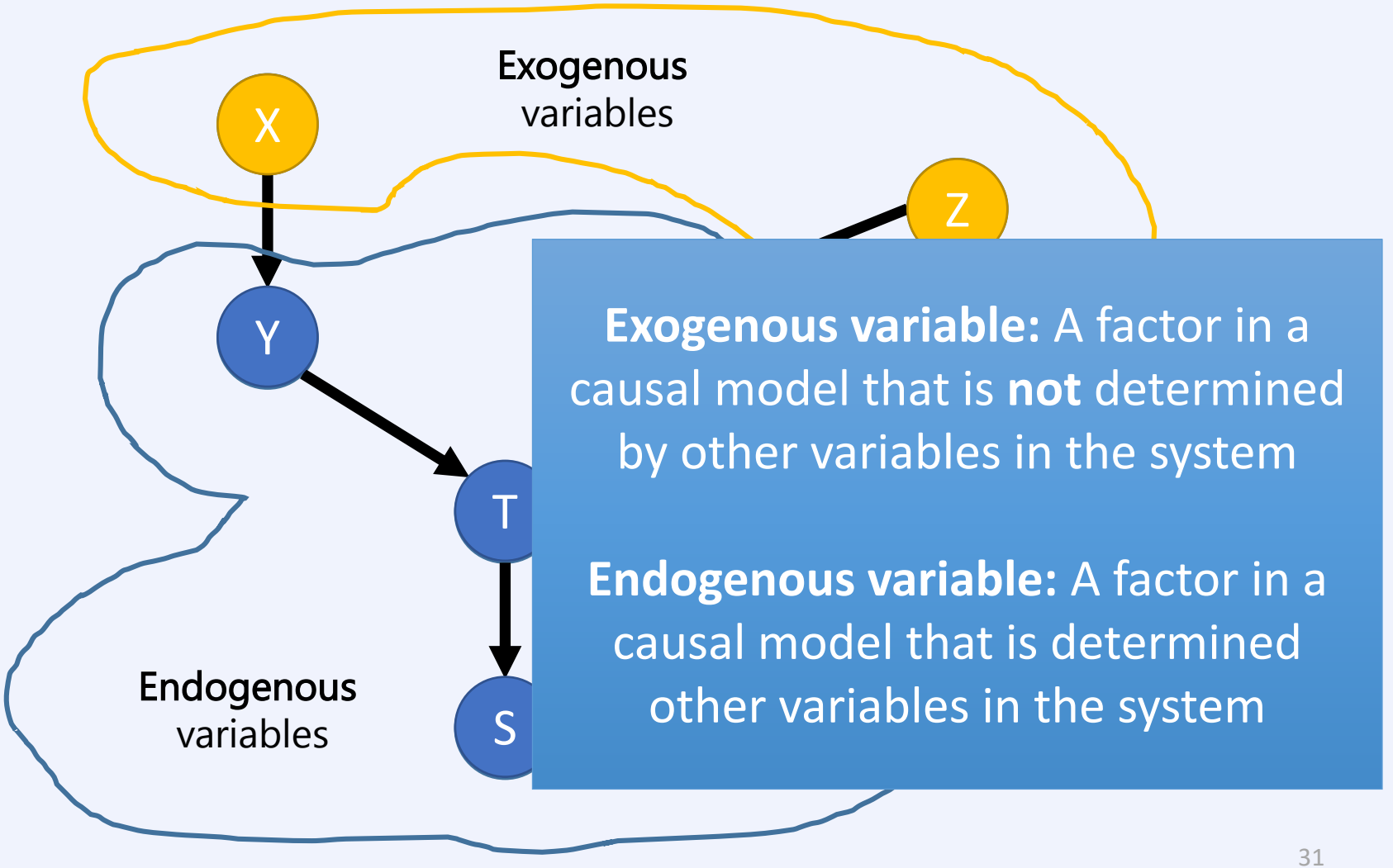
Types of Nodes



Causal Discovery



Causal Discovery



Exogenous variable: A factor in a causal model that is **not** determined by other variables in the system

Endogenous variable: A factor in a causal model that is determined other variables in the system

Causal Markov Condition

Any distribution generated by a Markovian model M can be factorized as

$$p(X_1, X_2, \ldots, X_n) = \prod_i p(X_i \mid pa_i)$$

where $X_1, X_2, \ldots, X_n$ are the **endogenous** variables in M, and pa_i are (values of) the **endogenous** "parents" of X_i in the causal diagram associated with M

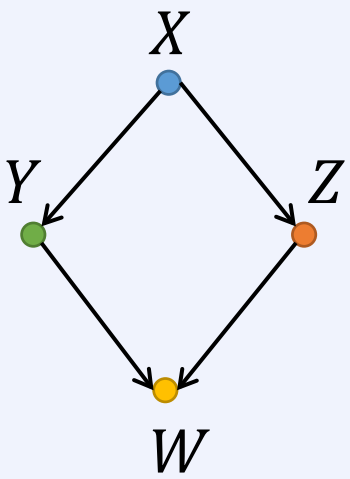
(Spirtes, Glymour, Scheines 1982; Pearl 2009)

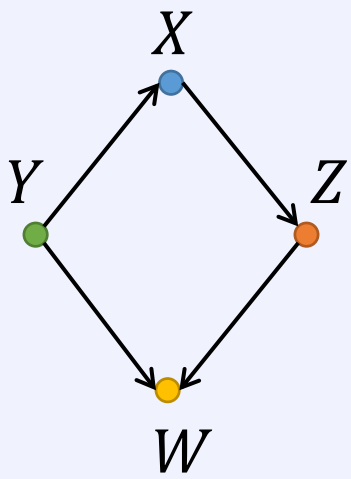
In other words…

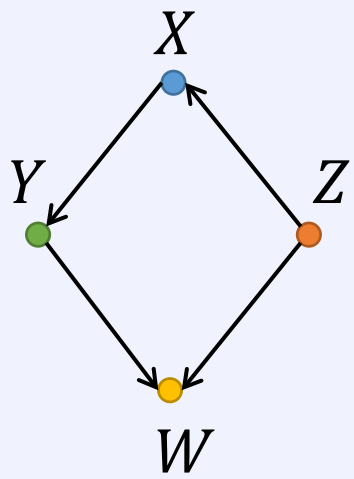
For all distinct variables X and Y in the variable set V, if X does not cause Y, then $P(X \mid Y, pa_X) = P(X \mid pa_X)$

That is, we can **weed out** edges from a causal graph – we can identify DAGs **up to** Markov equivalence classes.

Which is great, although we are **unable** to choose among these







Constraint-Based Causal Discovery

The PC algorithm is one of the most well-known, and most relied upon causal discovery algorithms

- proposed by **P**eter Spirtes and **C**lark Glymour

Assumes the following

1) data-generating distribution has the causal Markov property on graph G
2) data generating distribution is faithful to G
3) every member of the population has the same distribution
4) all relevant variables are in G
5) there is only one graph G to which the distribution is faithful

Constraint-Based Causal Discovery

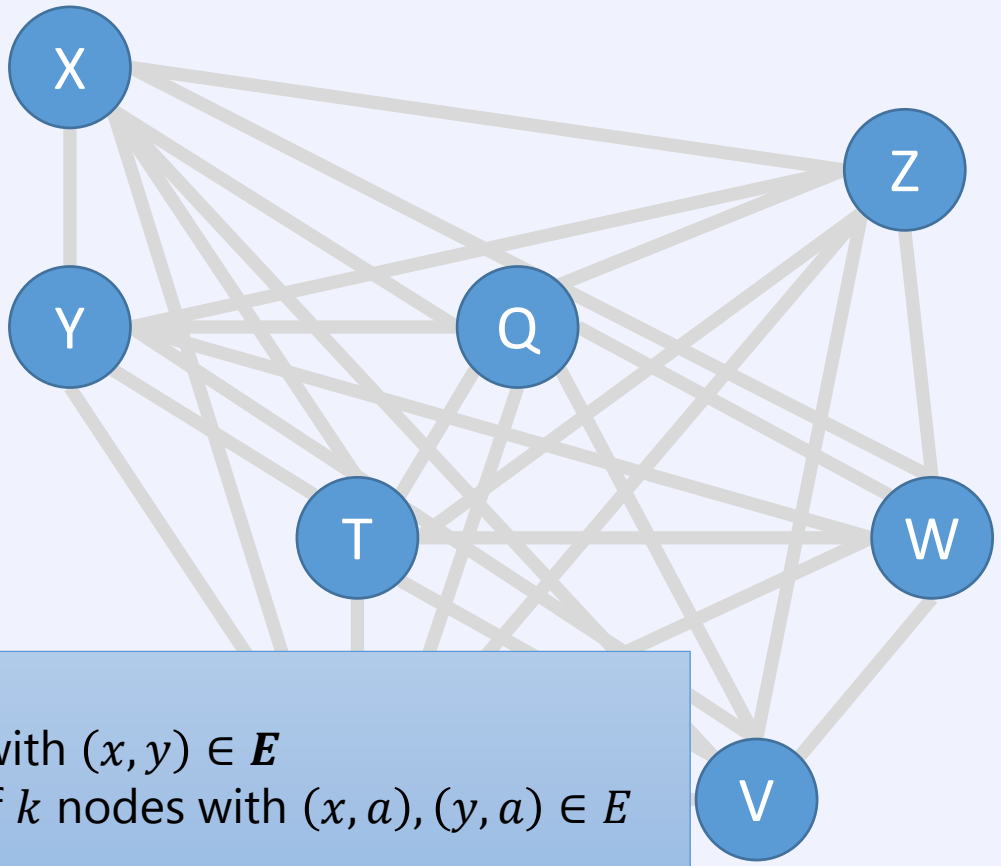
The PC algorithm is one of the most well-known, and most relied upon causal discovery algorithms

- proposed by **P**eter Spirtes and **C**lark Glymour

Two main steps

1) use conditional independence tests to determine the undirected causal graph (aka the skeleton)
2) apply constraint-based rules to direct (some) edges

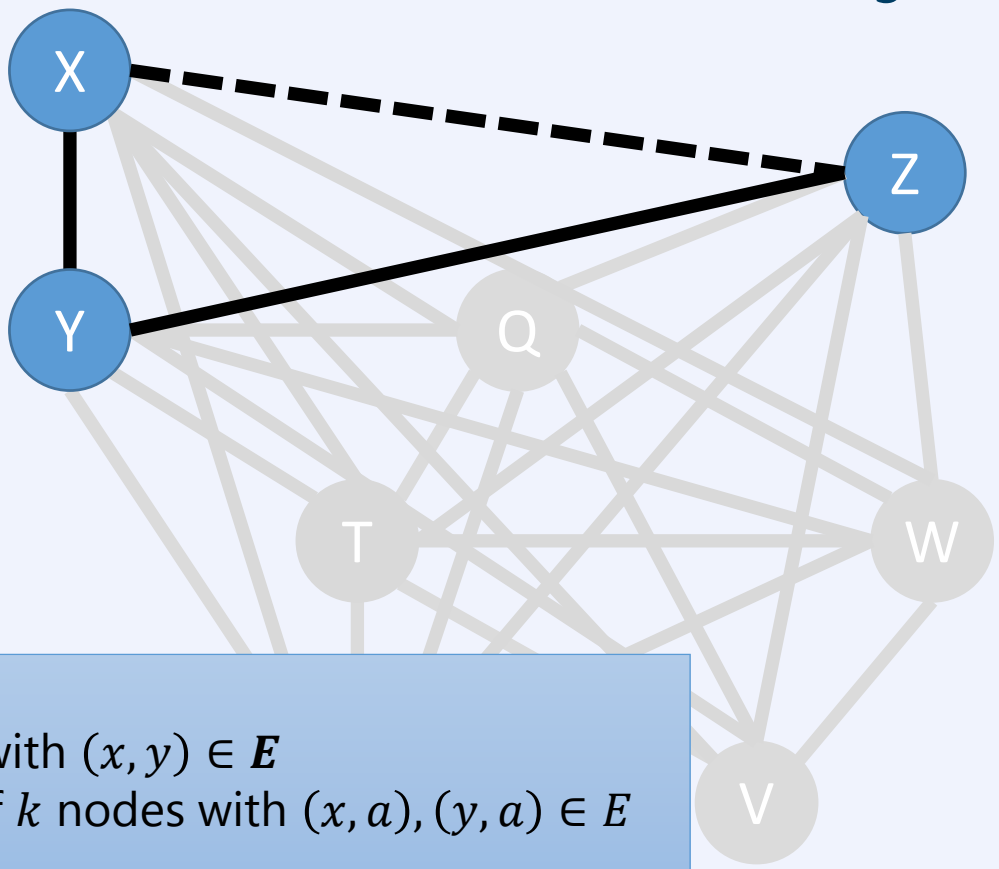
Step 1: Discover the Skeleton



for $k = 0$ *to* n
  for all $X, Y \in \boldsymbol{V}$ with $(x, y) \in \boldsymbol{E}$
    for all $\boldsymbol{A} \subseteq \boldsymbol{V}$ of k nodes with $(x, a), (y, a) \in E$
      if $X \perp\!\!\!\perp Y \mid A$
        remove (x, y) from $\boldsymbol{E}$

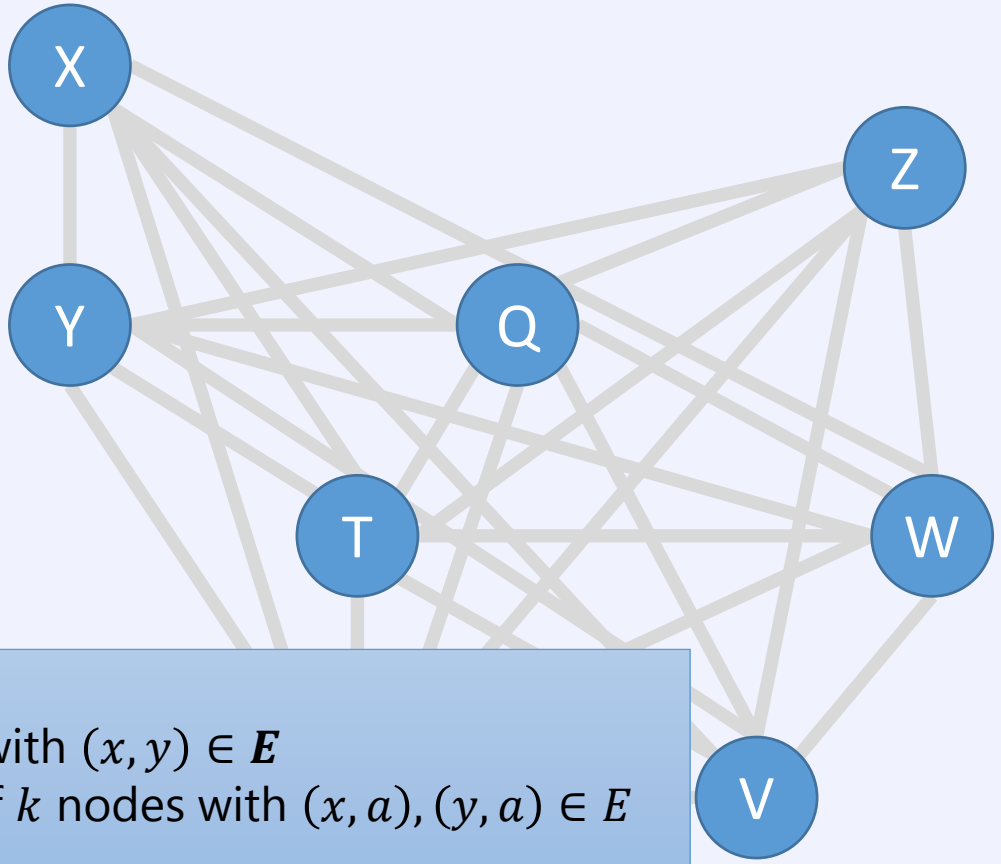
Step 1: Discover the Skeleton

$X \perp\!\!\!\perp Z \mid Y$?→ no causal edge



for $k = 0$ *to* n
  for all $X, Y \in \boldsymbol{V}$ with $(x, y) \in \boldsymbol{E}$
    for all $\boldsymbol{A} \subseteq \boldsymbol{V}$ of k nodes with $(x, a), (y, a) \in E$
      if $X \perp\!\!\!\perp Y \mid A$
        remove (x, y) from $\boldsymbol{E}$

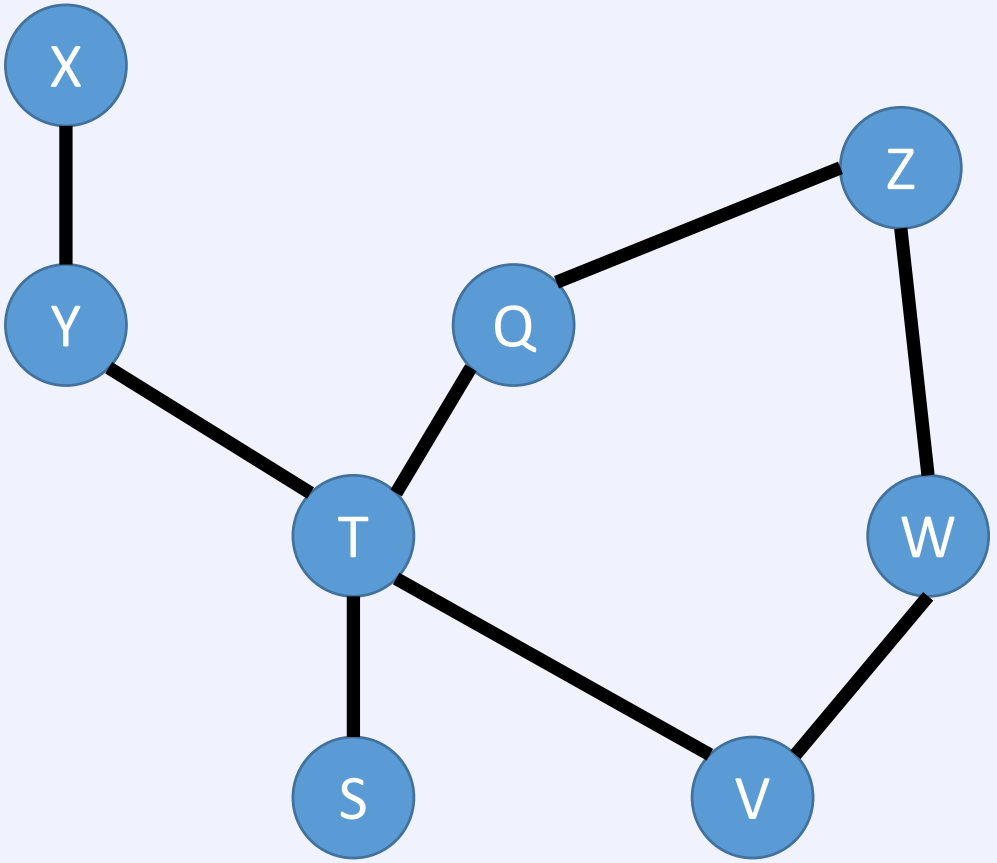
Step 1: Discover the Skeleton



for $k = 0$ *to* n
  for all $X, Y \in \boldsymbol{V}$ with $(x, y) \in \boldsymbol{E}$
    for all $\boldsymbol{A} \subseteq \boldsymbol{V}$ of k nodes with $(x, a), (y, a) \in E$
      if $X \perp\!\!\!\perp Y \mid A$
        remove (x, y) from $\boldsymbol{E}$

Step 1: Discover the Skeleton

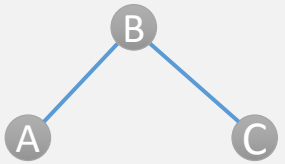
We now have the **causal skeleton**



Step 2: Orientation

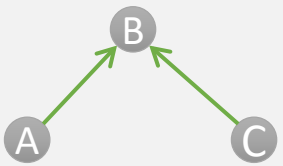
We now identify all **colliders** $X \to Y \leftarrow Z$
considering all relevant pairs **once**

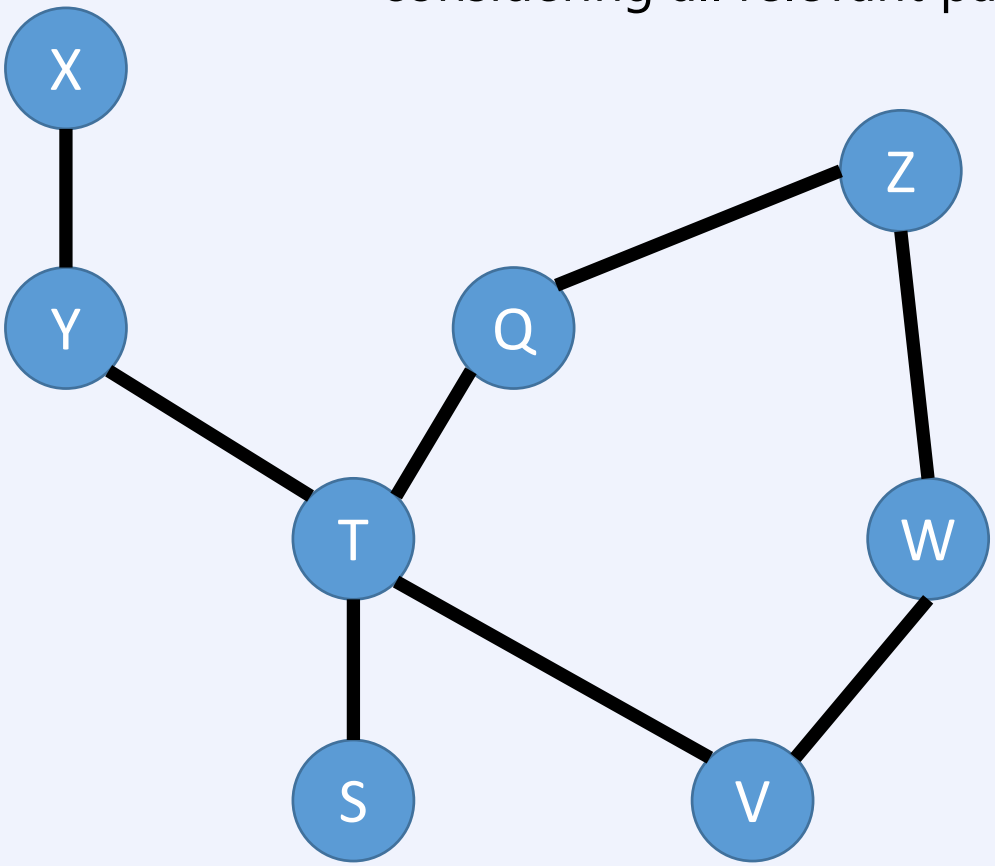
Rule 1



$A \not\perp\!\!\!\perp C$

$A \not\perp\!\!\!\perp C \mid B$

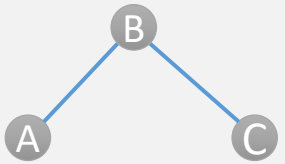




Step 2: Orientation

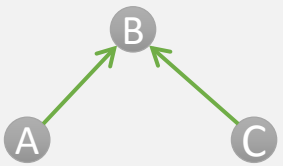
We now identify all **colliders** $X \rightarrow Y \leftarrow Z$
considering all relevant pairs **once**

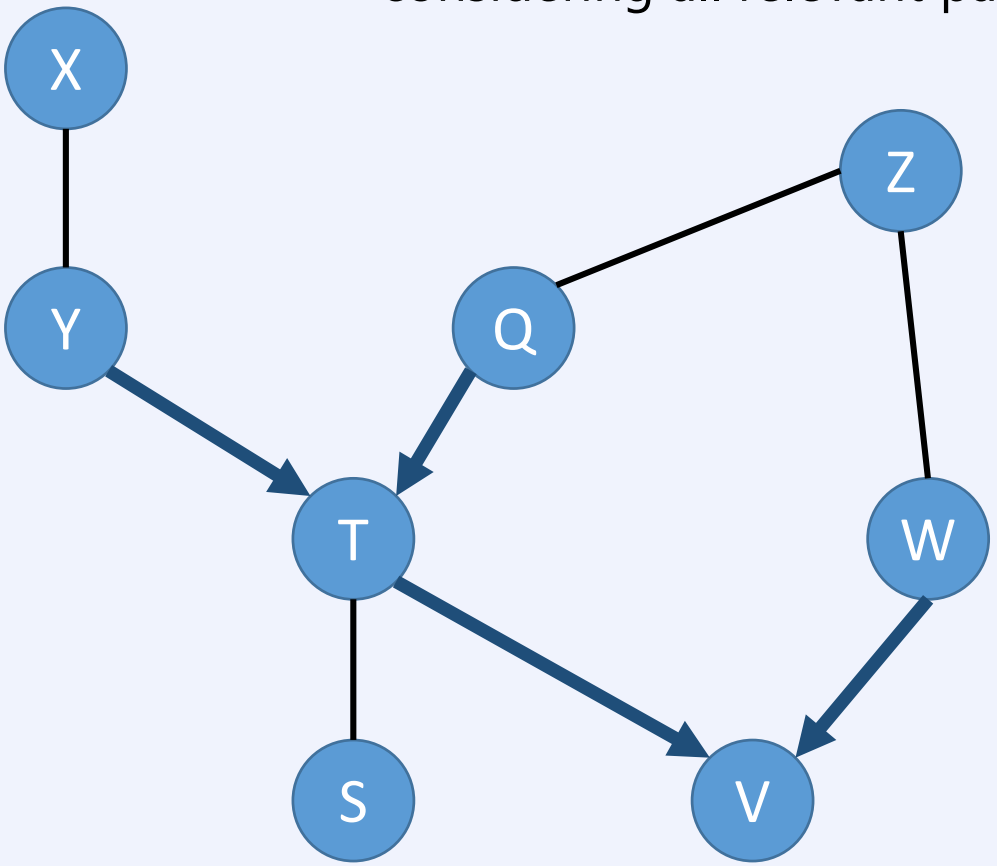
Rule 1



$A \perp\!\!\!\perp C$

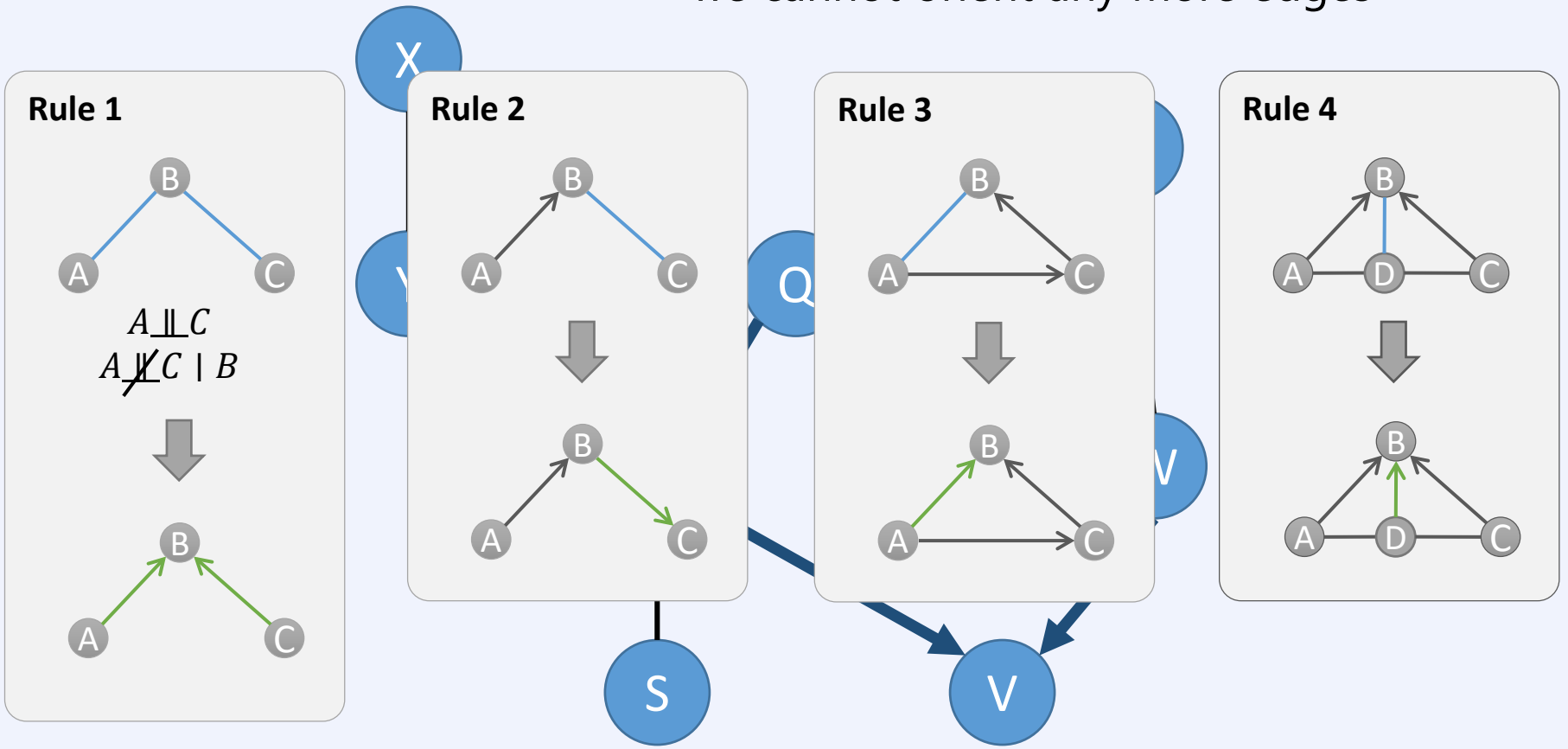
$A \not\perp\!\!\!\perp C \mid B$





Step 2: Orientation

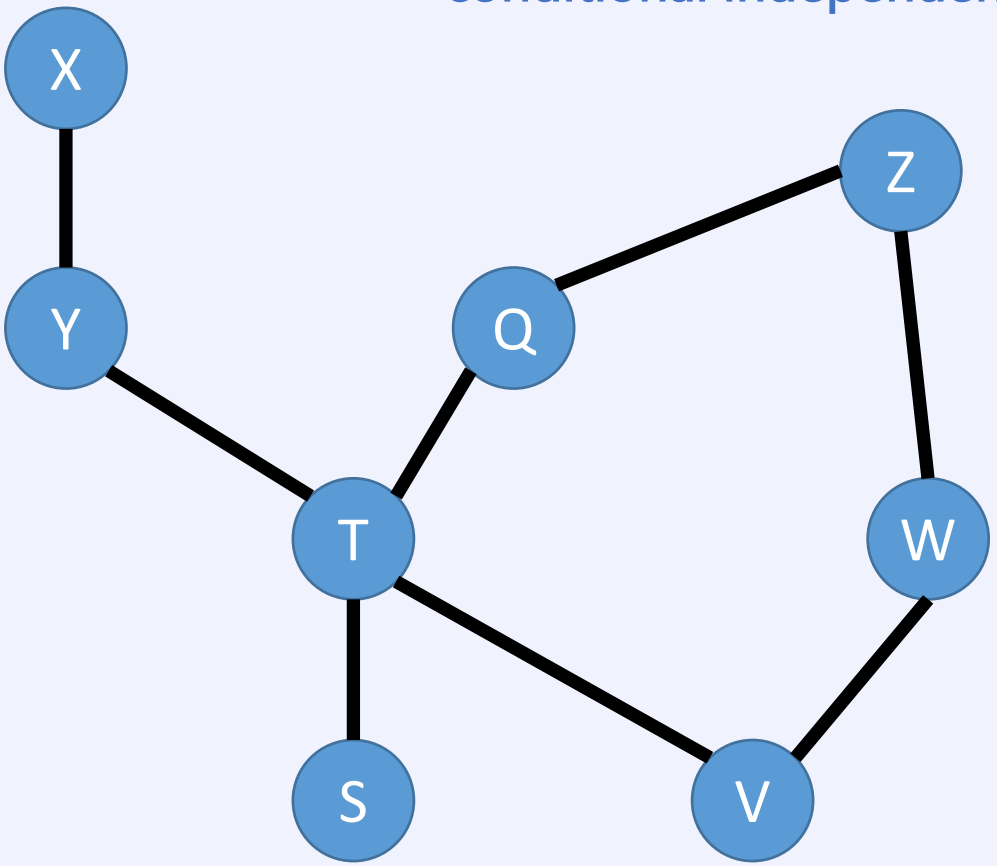
We then iteratively apply Rules 2—3 until we cannot orient any more edges



Causal Inference

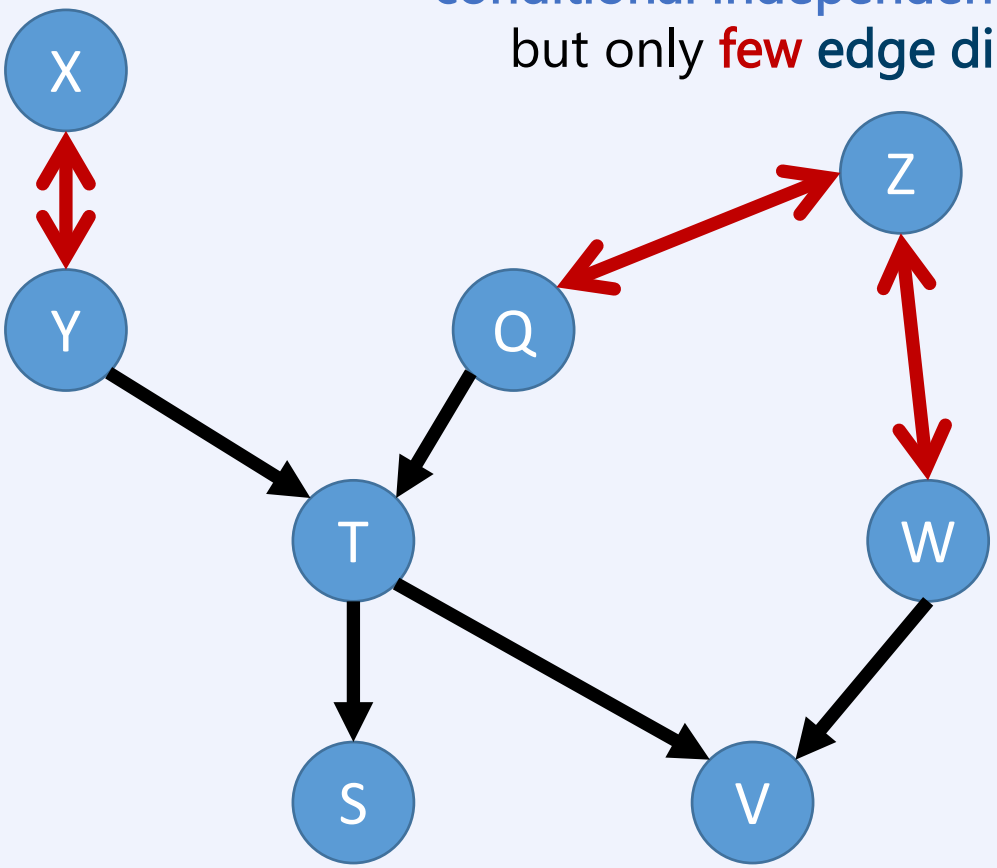
Causal Inference

We can find the **causal skeleton** using **conditional independence** tests



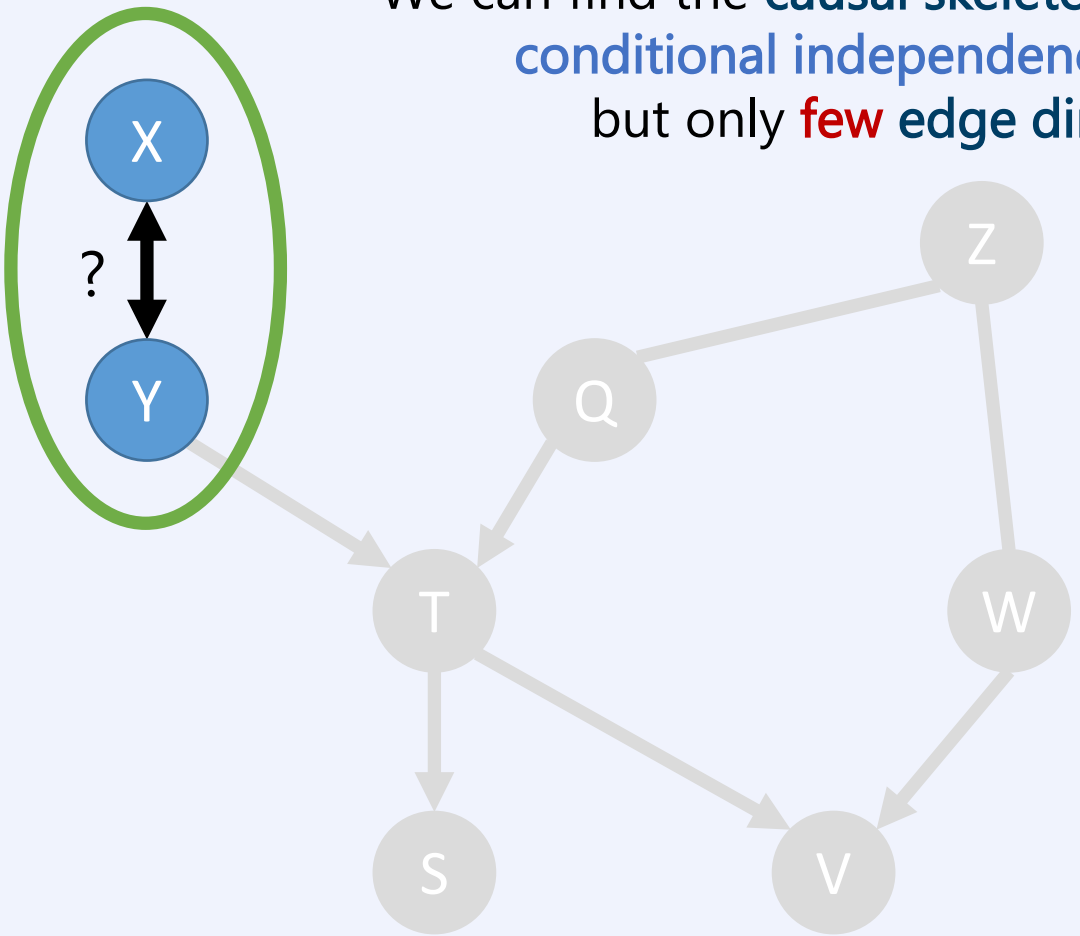
Causal Inference

We can find the **causal skeleton** using **conditional independence** tests, but only **few edge directions**



Causal Inference

We can find the **causal skeleton** using **conditional independence** tests, but only **few** edge directions



Three is a crowd

Traditional causal inference methods
rely on **conditional independence tests**
and hence require *at least* **three** observed variables

That is, they **cannot** distinguish between
$X \to Y$ and $Y \to X$
as $p(x)p(y \mid x) = p(y)p(x \mid y)$
are just factorisations of $p(x, y)$

Can we infer the causal direction between pairs?

Wiggle Wiggle

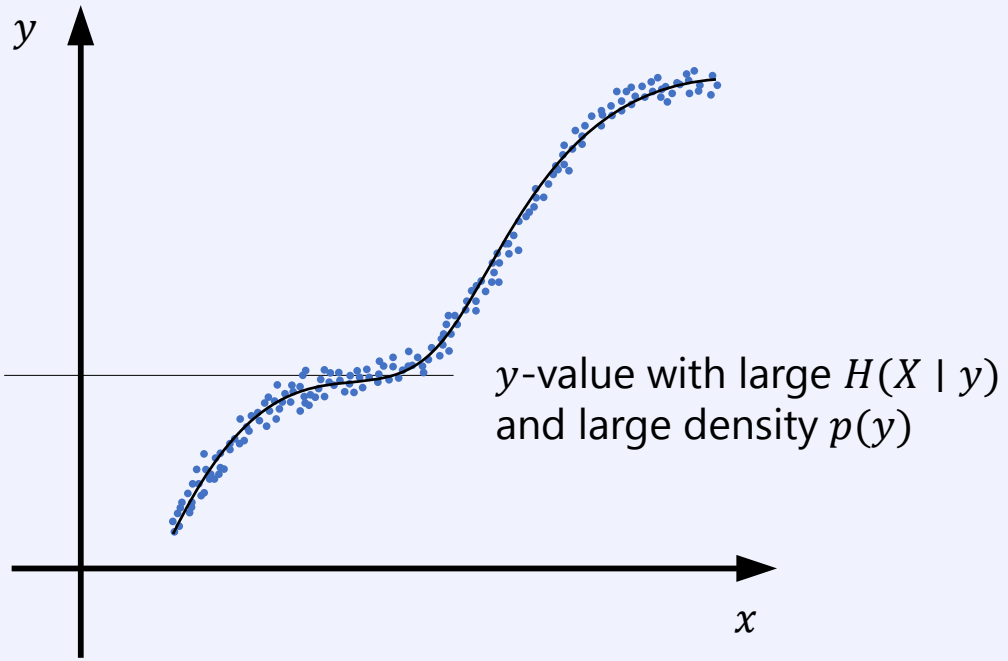
Let's take another look at the definition of causality.

'the relationship between something
that happens or exists and the thing that causes it'

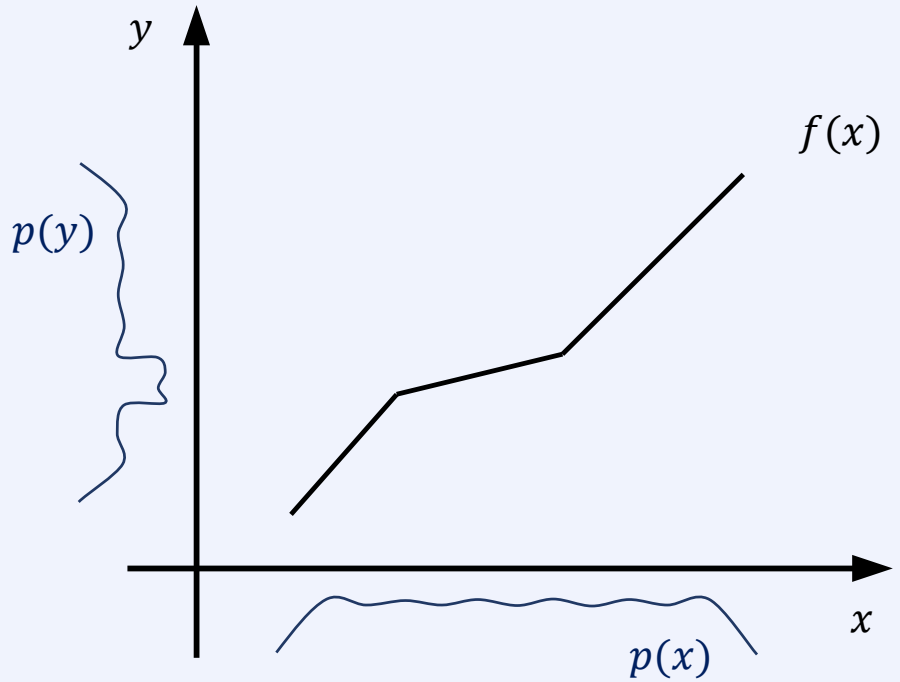
From the do-calculus it follows that if X cause Y,
we can wiggle Y by wiggling X,
while when we cannot wiggle X by wiggling Y.

But… we only have observational data jointly over X, Y,
and cannot do any wiggling ourselves…

May The Noise Be With You



May The Noise Be With You



"If the **structure** of density of $p(x)$ is not correlated with the slope of f, then the flat regions of f induce peaks in $p(y)$.

The causal hypothesis $Y \rightarrow X$ is thus implausible because the causal mechanism f^{-1} appears to be adjusted to the "input" distribution $p(y)$."

(Janzing et al. 2012)

Independence of Input and Mechanism

If X causes Y,
the **marginal distribution** of the **cause**, $p(X)$
and the **conditional distribution** of
the **effect given the cause**, $p(Y|X)$
are **independent**

That is, if $X \rightarrow Y$
$p(X)$ **contains no information** about $p(Y|X)$

Additive Noise Models

Whenever the joint distribution $p(X, Y)$ **admits** a model in one direction, i.e. there exists an f and N such that

$$Y = f(X) + N \text{ with } N \perp\!\!\!\perp X,$$

but does **not admit** the reversed model, i.e. for all g and $\widetilde{N}$ we have

$$X = g(Y) + \widetilde{N} \text{ with } \widetilde{N} \not\perp\!\!\!\perp Y$$

We can infer $X \to Y$

(Shimizu et al. 2006, Hoyer et al. 2009, Zhang & Hyvärinen 2009)

ANMs and Identifiability

When are ANMs identifiable?

- what do we need to assume about the data generating process for ANM-based inference to make sense?
- for which functions f and what noise distributions $\mathcal{N}$ are ANMs identifiable from observational data?

✗ Linear functions and Gaussian noise

✓ Linear functions and non-Gaussian noise

✓ For most cases of non-linear functions and any noise

(Shimizu et al. 2006, Hoyer et al. 2009, Zhang & Hyvärinen 2009)

Additive Noise Models

Whenever the joint distribution $p(X, Y)$ **admits** a model in one direction, i.e. there exists an f and N such that

$$Y = f(X) + N \text{ with } N \perp\!\!\!\perp X,$$

but does **not admit** the reversed model, i.e. for all g and $\widetilde{N}$ we have

$$X = g(Y) + \widetilde{N} \text{ with } \widetilde{N} \not\perp\!\!\!\perp Y$$

How do we determine or use this in practice?

(Shimizu et al. 2006, Hoyer et al. 2009, Zhang & Hyvärinen 2009)

Independence of Input and Mechanism

If X causes Y,
the **marginal distribution** of the **cause**, $p(X)$
and the **conditional distribution** of
the **effect given the cause**, $p(Y|X)$
are **independent**

That is, if $X \rightarrow Y$
$p(X)$ **contains no information** about $p(Y|X)$

(Sgouritsa et al 2015)

Plausible Markov Kernels

In other words, if we observe that

$$p(cause)p(effect \mid cause)$$

is simpler than

$$p(effect)p(cause \mid effect)$$

then it is likely that $cause \rightarrow effect$

How to robustly measure ‘simpler’?

(Sun et al. 2006, Janzing et al. 2012)

Kolmogorov Complexity

$$K(s)$$

The Kolmogorov complexity of a binary string s
is the length of the shortest program p^*
for a universal Turing Machine U
that generates s and **halts**.

(Kolmogorov, 1963)

Algorithmic Markov Condition

If $X \to Y$, we have,

up to an additive constant,

$$K\big(p(X)\big) + K\big(p(Y|X)\big) \leq K\big(p(Y)\big) + K\big(p(X|Y)\big)$$

That is, we can do **causal inference** by identifying the factorization of the joint with the **lowest Kolmogorov complexity**

(Janzing & Schölkopf, IEEE TIT 2012)

Univariate and Numeric

Two-Part MDL

The Minimum Description Length (MDL) principle

given a model class $\mathcal{M}$, the best model $M \in \mathcal{M}$
is the M that minimises

$$L(M) + L(D \mid M)$$

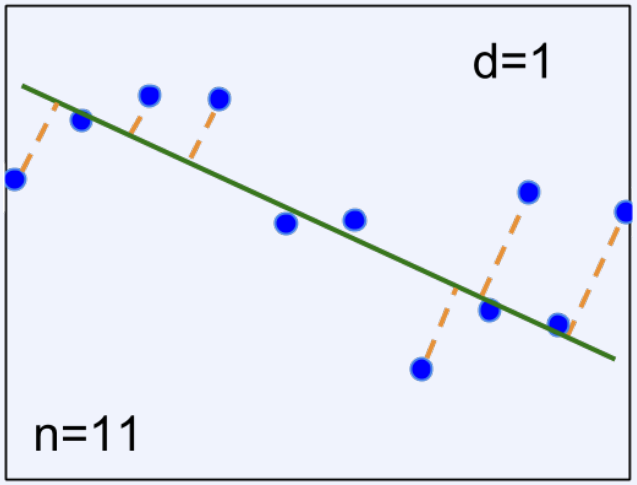
in which

$L(M)$ is the length, in bits, of the description of M

$L(D \mid M)$ is the length, in bits, of the description of the data when encoded using M

(see, e.g., Rissanen 1978, 1983, Grünwald, 2007)

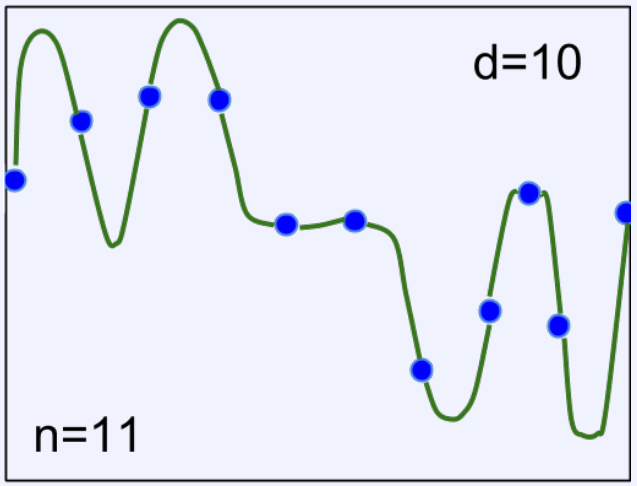
MDL and Regression



$$L(M) + L(D|M)$$

$a_1 x + a_0$ errors

VS.



$a_{10} x^{10} + a_9 x^9 + \ldots + a_0$ { }

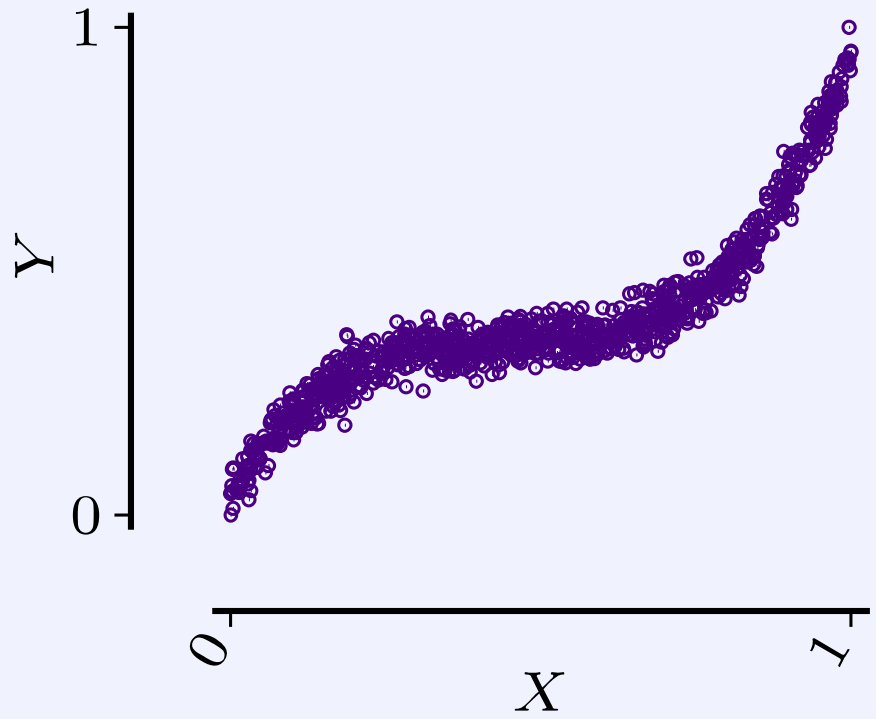
Modelling the Data

We model Y as

$$Y = f(X) + \mathcal{N}$$

As f we consider **linear**, **quadratic**, **cubic**, **exponential**, and **reciprocal** functions, and model the noise using a 0-mean Gaussian. We choose the f that minimizes

$$L(Y \mid X) = L(f) + L(\mathcal{N})$$



SLOPE – computing $L(Y \mid X)$

Linear in #data points

```
1  F = ∅;
2  f_g ← fit global function and add f_g to F;
3  for each function type t do
4  |   F_t ← F;
5  |   for x ∈ X, count(x) > δ do
6  |   |   f_l ← fit local function on x̃ of x;
7  |   |   if adding f_l to F_t reduces overall costs then
8  |   |   |   F_t = F_t ∪ f_l;
9  |   |   end
10 |   end
11 |   F ← min(F, F_t);
12 end
13 return costs of Y given F and X;
```

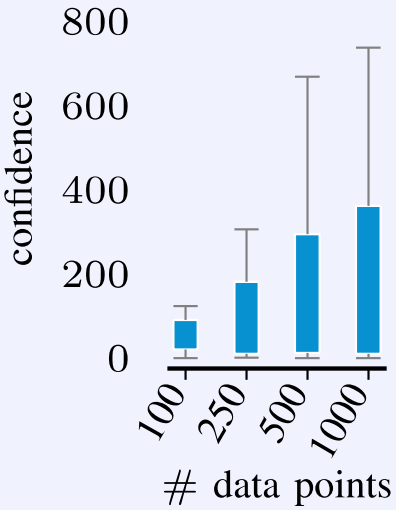
Confidence and Significance

How certain are we?

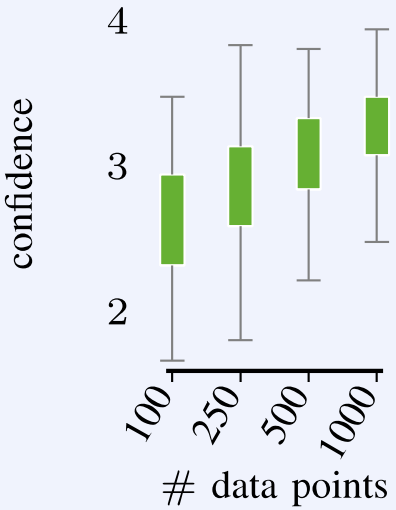
$$\mathbb{C} = |\underbrace{L(X) + L(Y \mid X)}_{L(X \to Y)} - \underbrace{L(Y) + L(X \mid Y)}_{L(Y \to X)}|$$

- the higher the more certain

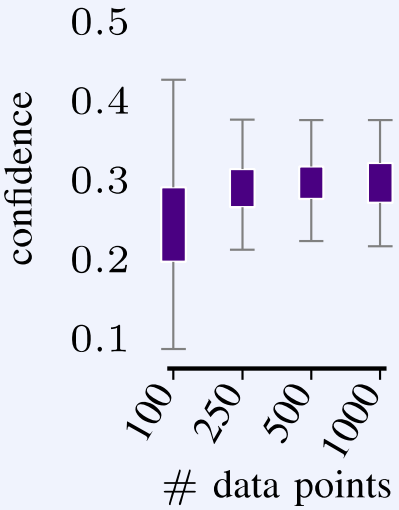
Confidence Robustness



RESIT
(HSIC idep.)



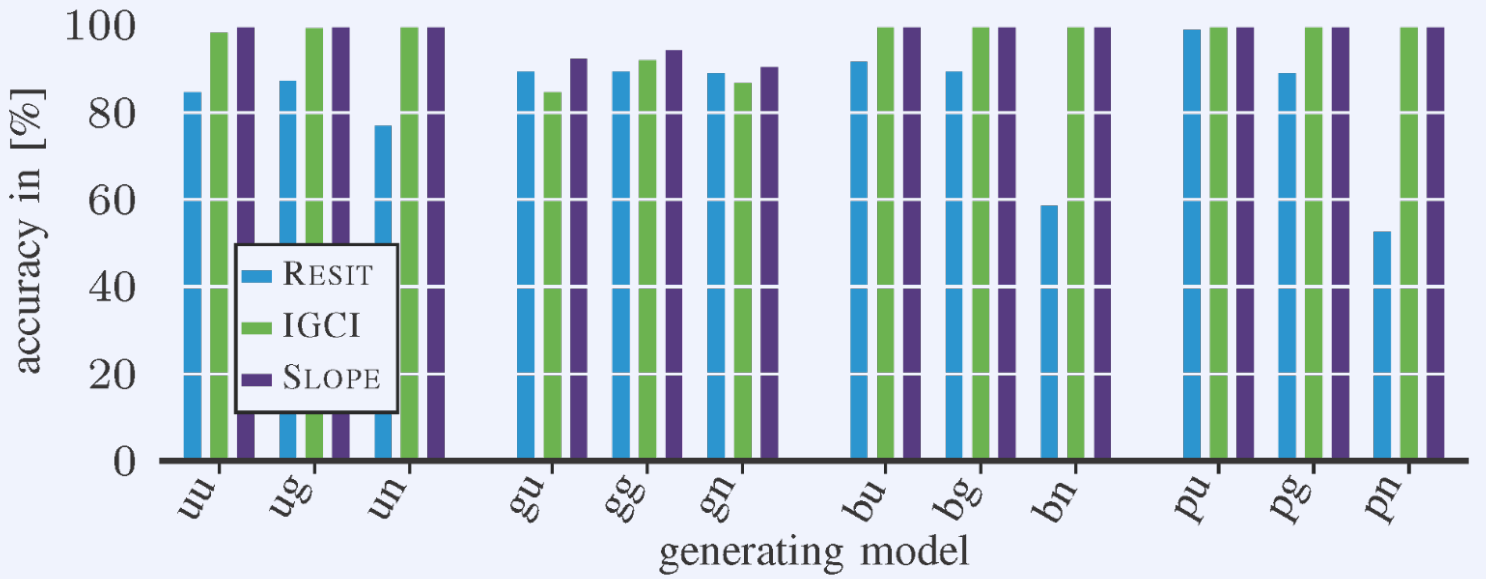
IGCI
(Entropy)



SLOPE
(Compression)

Putting SLOPE to the test

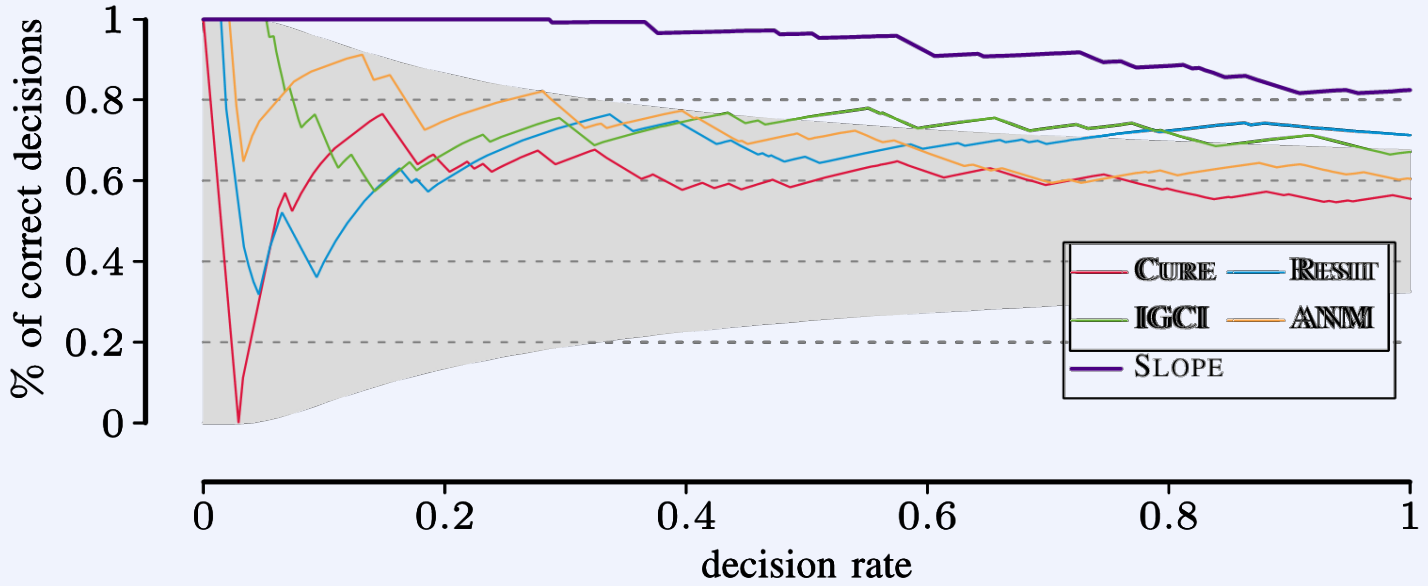
We first evaluate using an ANM, with linear, cubic, or reciprocal functions, sampling X and noise as indicated.



(Resit by Peters et al, 2014; IGCI by Janzing et al 2012)

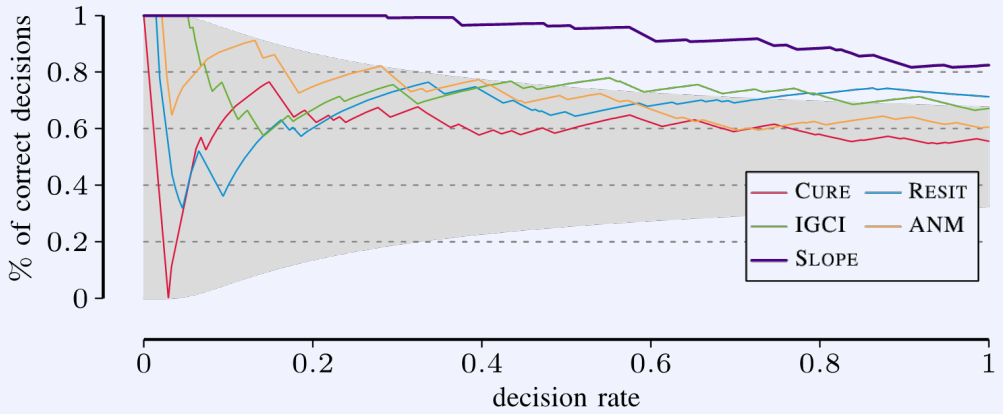
Performance on Benchmark Data

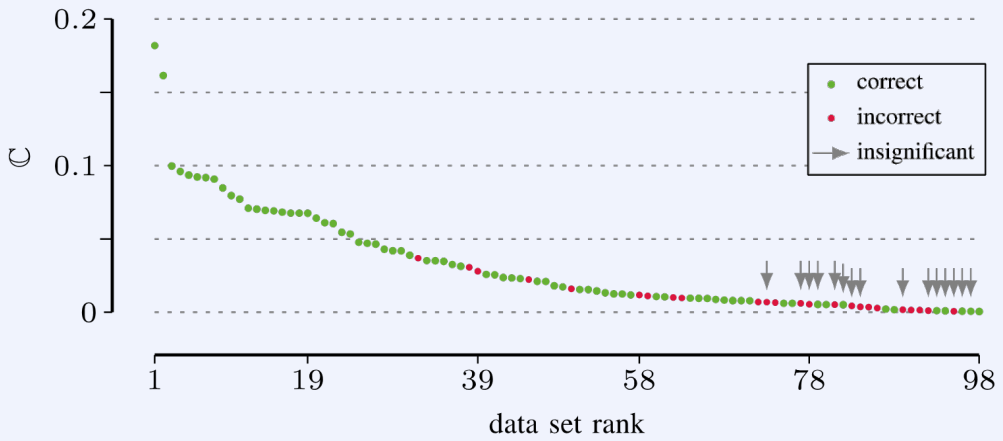
(Tübingen 97 univariate numeric cause-effect pairs, weighted)



Performance on Benchmark Data

(Tübingen 97 univariate numeric cause-effect pairs, weighted)



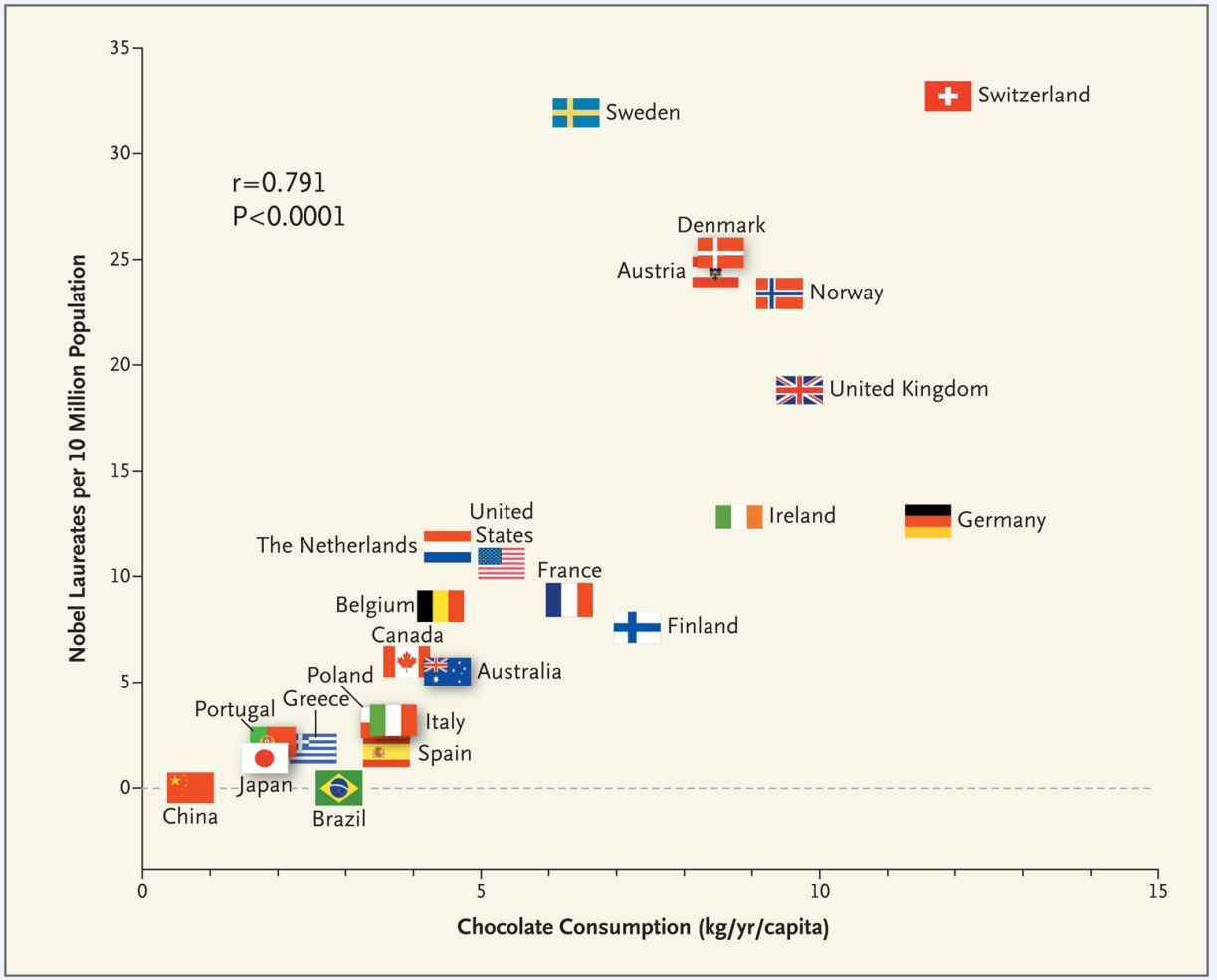


Inferences of state of the art algorithms **ordered** by **confidence** values.

SLOPE is 85% accurate with $\alpha = 0.001$

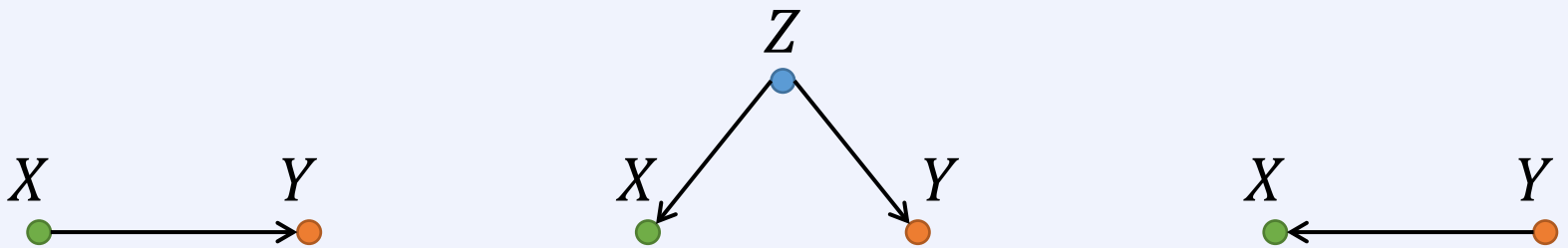
Detecting Confounding

Does Chocolate Consumption cause Nobel Prizes?



Reichenbach

If X and Y are statistically dependent then either

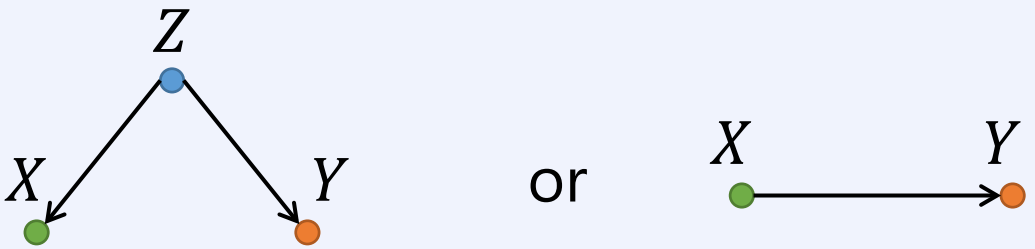


How can we distinguish these cases?

(Reichenbach, 1956)

Conditional Independence Tests

If we have measured everything relevant
then testing $X \perp\!\!\!\perp Y | Z$ for all possible Z
lets us decide whether



Problem: It's impossible to measure everything relevant

Why not just find a confounder?

We would like to be able to infer a $\hat{Z}$ such that

$$X \perp\!\!\!\perp Y \mid \hat{Z}$$

if and only if X and Y are actually confounded

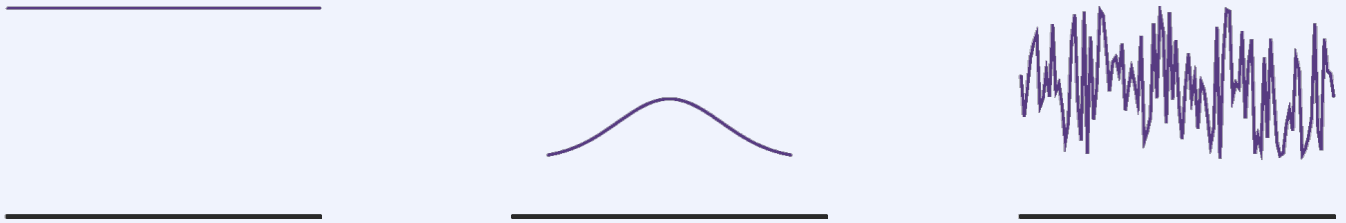
Problem: Finding such a $\hat{Z}$ is **too easy**.
$\hat{Z} = X$ always works

Kolmogorov Complexity

$K(P)$ is the length of the shortest program computing P

$$K(P) = \min_{p} \left\{ |p| : p \in \{0,1\}^*, |\mathcal{U}(p, x, q) - P(x)| < \frac{1}{q} \right\}$$

This shortest program p^* is the best compression of P



From the Markov Condition…

An admissible causal network for $X_1, \ldots, X_m$ is G satisfying

$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(\, X_i \mid pa_i \,)$$

Problem: How do we find a simple factorization?

…to the Algorithmic Markov Condition

The simplest causal network for $X_1, \ldots, X_m$ is G^* satisfying

$$K(P(X_1, \ldots, X_m)) = \sum_{i=1}^{m} K(P(X_i \mid pa_i^*))$$

Postulate: G^* corresponds to the true generating process

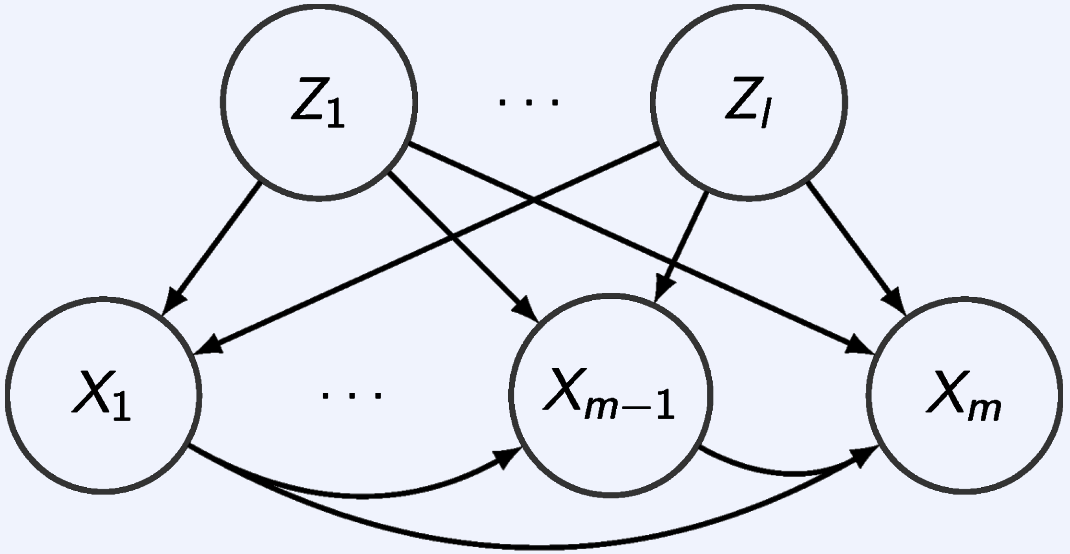
(Janzing & Schölkopf, 2010)

AMC with Confounding

We can also include latent variables

$$K\big(P(\boldsymbol{X},\boldsymbol{Z})\big) = \sum_{i=1}^{m} K\big(P(X_i \mid pa_i')\big) + \sum_{j=1}^{l} K\Big(P\big(Z_j\big)\Big)$$

We don't know $P(\cdot)$



$$P(\boldsymbol{X}, \boldsymbol{Z}) = P(\boldsymbol{Z}) \prod_{i=1}^{m} P(X_i \mid \boldsymbol{Z})$$

In particular, we will use probabilistic PCA

Kolmogorov is not computable

For data X, the Minimum Description Length principle identifies the best model $M \in \mathcal{M}$ by minimizing

$$L(X, M) = L(M) + L(X \mid M)$$

gives a statistically sound approximation to K

Decisions, decisions

If

$$L(\boldsymbol{X}, Y, \mid \mathcal{M}_{co}) < L(\boldsymbol{X}, Y \mid \mathcal{M}_{ca})$$

then we consider $\boldsymbol{X}, Y$ to be confounded

Decisions, decisions

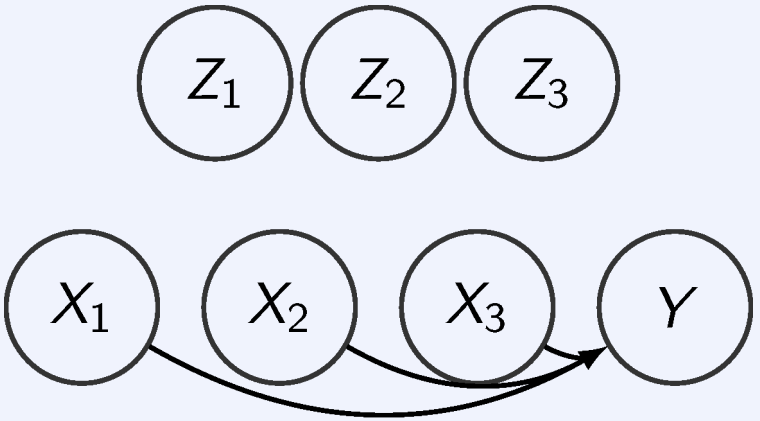
If

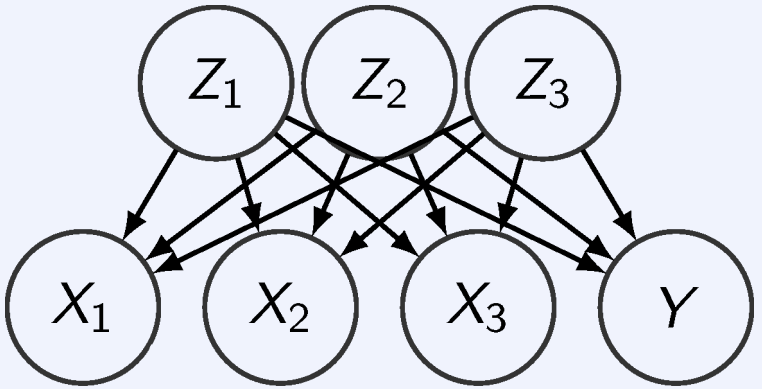
$$L(\boldsymbol{X}, Y, \mid \mathcal{M}_{co}) > L(\boldsymbol{X}, Y \mid \mathcal{M}_{ca})$$

then we consider $\boldsymbol{X}, Y$ to be causal

The difference can be interpreted as confidence

Confounding in Synthetic Data



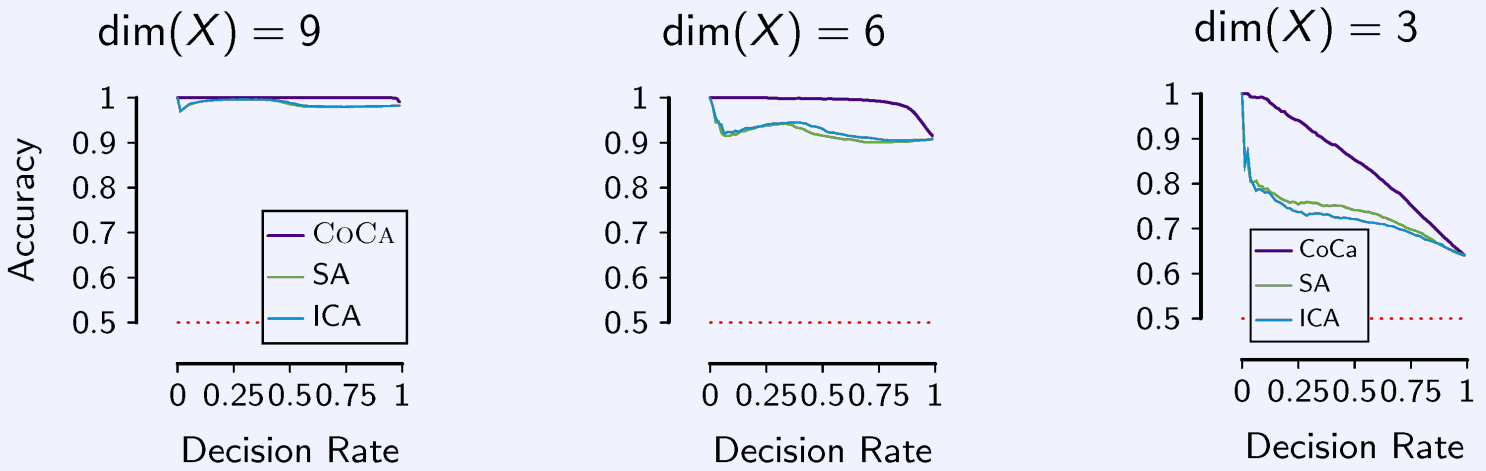


Synthetic Data: Results

There are only two other works directly related to ours

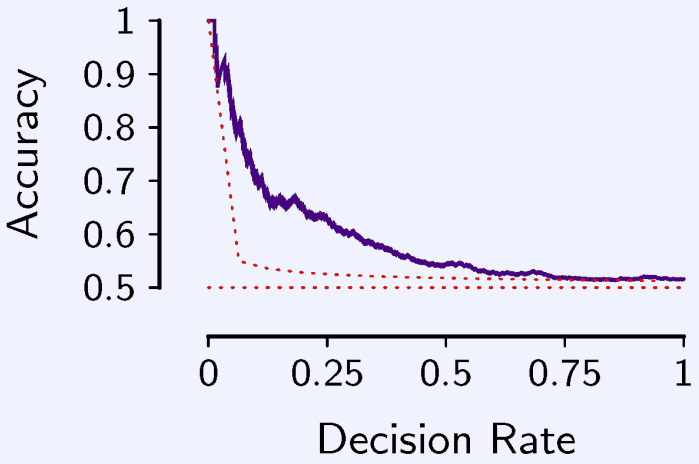
SA: Confounding strength in linear models using spectral analysis

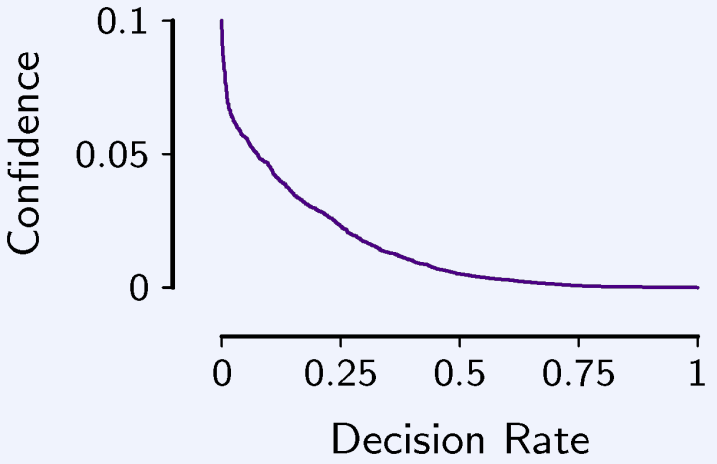
ICA: Confounding strength using independent component analysis



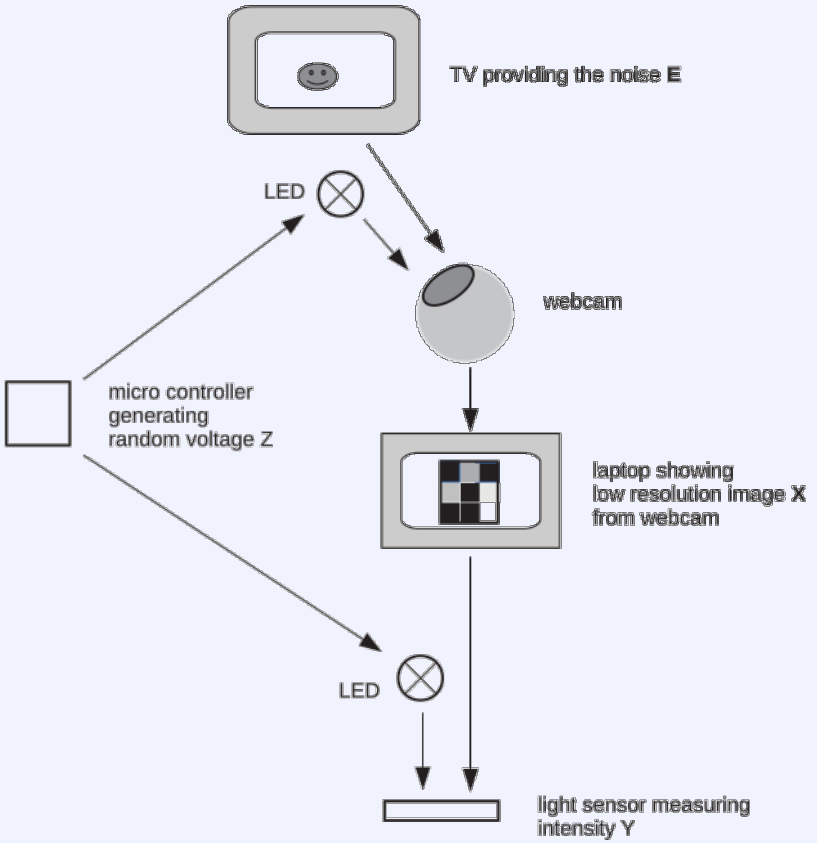
Confounding in Genetic Networks

More realistically, we consider gene regulation data



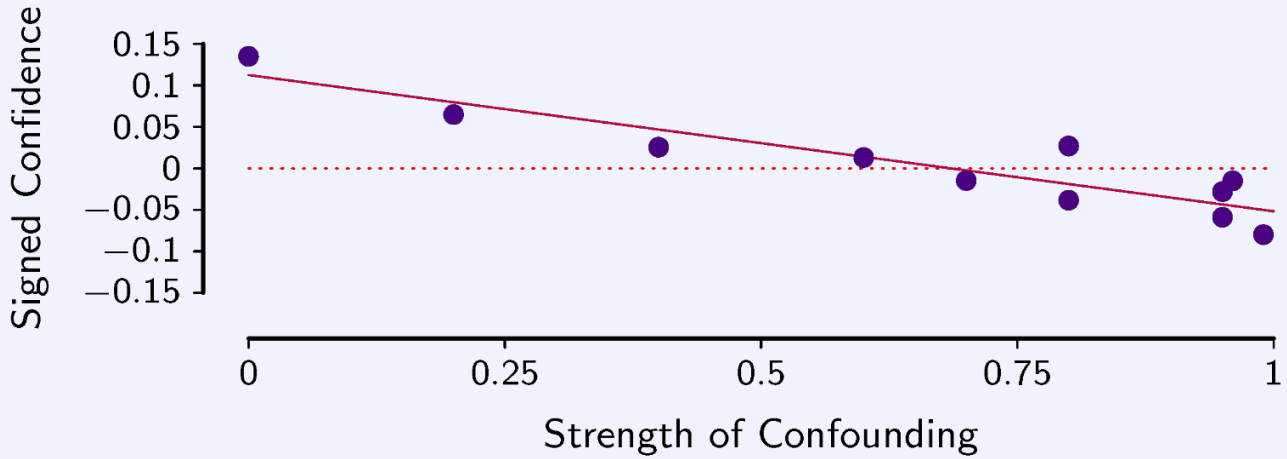


Optical Data

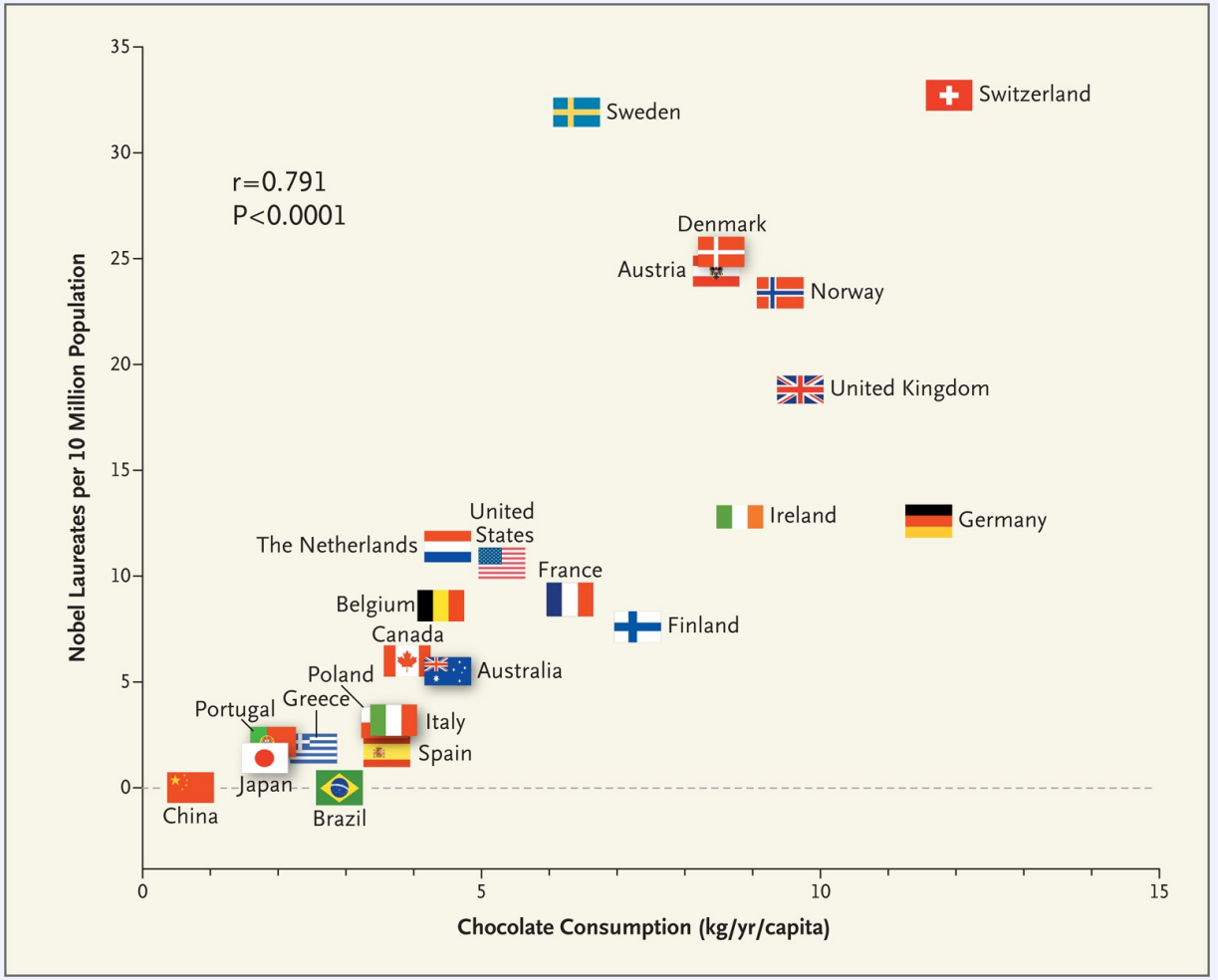


(Janzing & Schölkopf, 2017)

Optical Data



Wait! What about…



(Messerli 2012) 89

Conclusions

Causal inference from **observational data**

- necessary when making **decisions**, and to evaluate **what-if** scenarios
- **impossible without assumptions** about the causal model

Constraint-based **causal discovery**

- traditional approach based on **conditional independence testing**
- PC-algorithm discovers **causal skeleton** and **orients** (some) **edges**

Algorithmic Markov condition works very well in practice

- prefer simple explanations over complex ones
- consider complexity of both the model **and** the data

There is no causality without **assumptions**

- early work on relaxing e.g. causal sufficiency, determining confounding

Thank you!

"No causal claim can be established by a purely statistical method, be it propensity scores, regression, stratification, or any other distribution-based design"