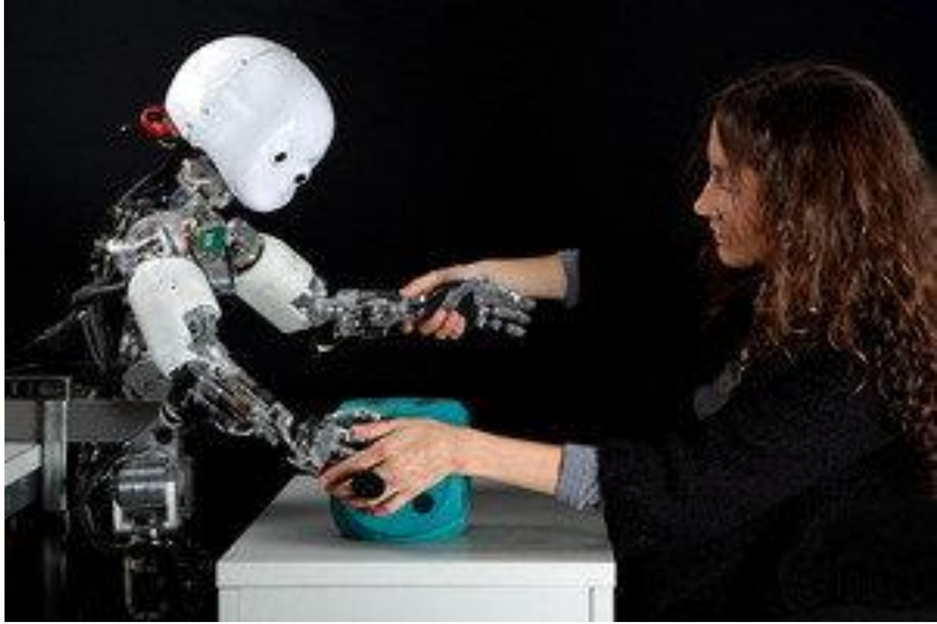
Causal inference and its connection to Machine Learning

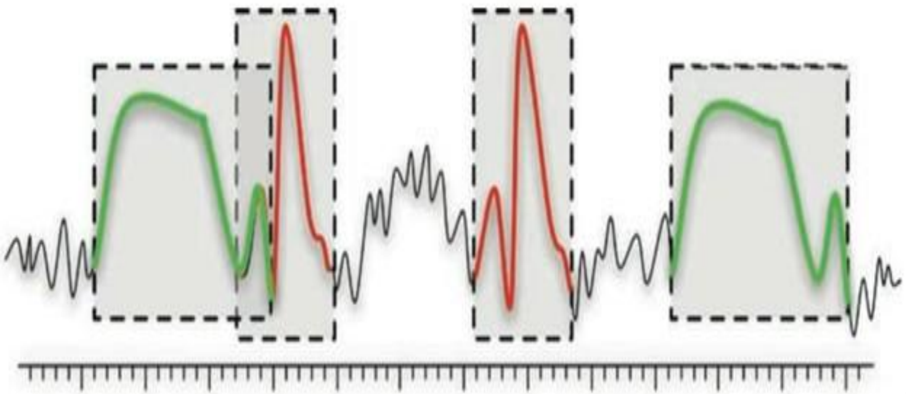
Sepideh Pashami

June 26, 2023

Rapid development of AI and autonomous systems toward human intelligence

Machine Learning is good at ….





Machine Learning is not enough!

- ML acts on observation.
  - No ML today can answer such questions about interventions **not encountered before**.
- ML (mainly) focuses on statistical relations (**correlation**) between variables.
  - It need to consider causal relations in data.
- ML can’t take human like decisions as systems get more autonomous
  - Combination of **ML and symbolic** collaboration of data and model is needed.

Human Intelligence

“Humans have the ability to

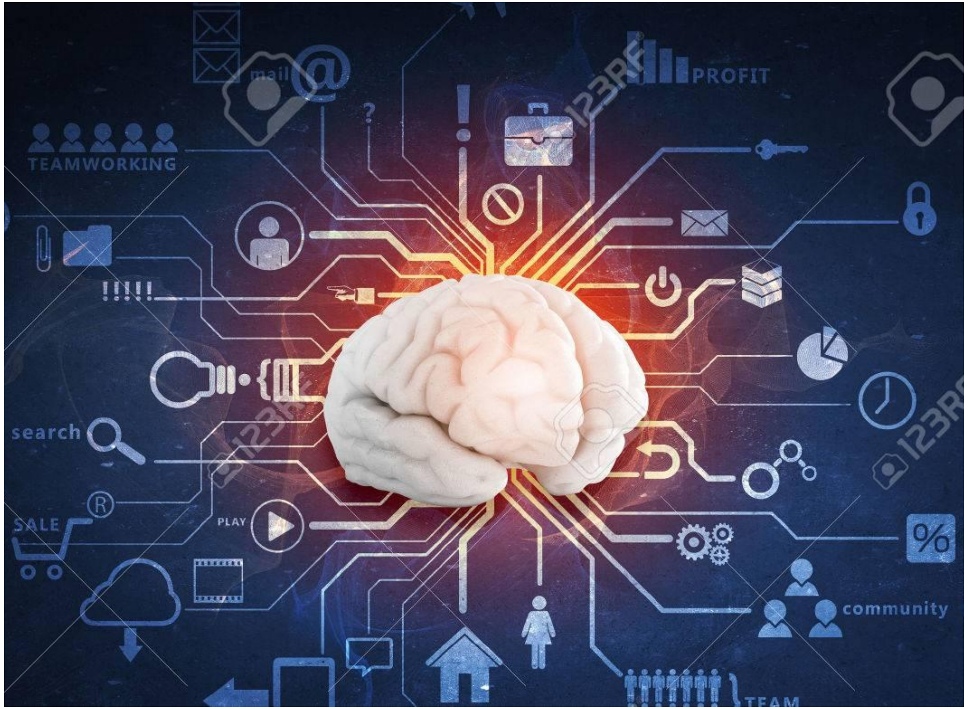
(1) choreograph a mental representation of their environment,

(2) interrogate that representation,

(3) distort it by mental acts of ***imagination*** and

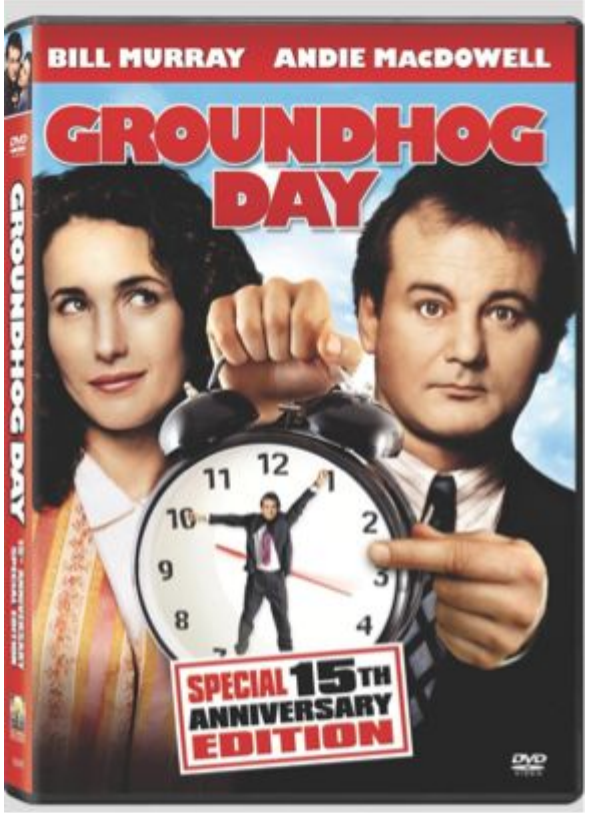
(4) finally answer *‘What if?’* kind of questions.”

Judea Pearl, 2018



Learning from imagination?

- **In fiction**
  - Groundhog day
    - Phil is trapped in a time loop
    - He experience different outcomes of his actions during a day.
- **In reality**
  - We observe
    - I took aspirin two hours ago, my headache has passed.
  - We can not observe
    - the case I didn’t take an aspirin. What would happen?



Causal inference

Inferring the effects of any treatment/policy/intervention/etc.

Why do we need causal inference?

How effective is a given treatment in **preventing** a disease?

Did the new tax law **cause** our sales to go up, or was it our advertising campaign?

What is the health-care cost **attributable to** obesity?

Can hiring records prove an employer is guilty of a **policy** of sex discrimination?

I am about to quit my job, **should I**?

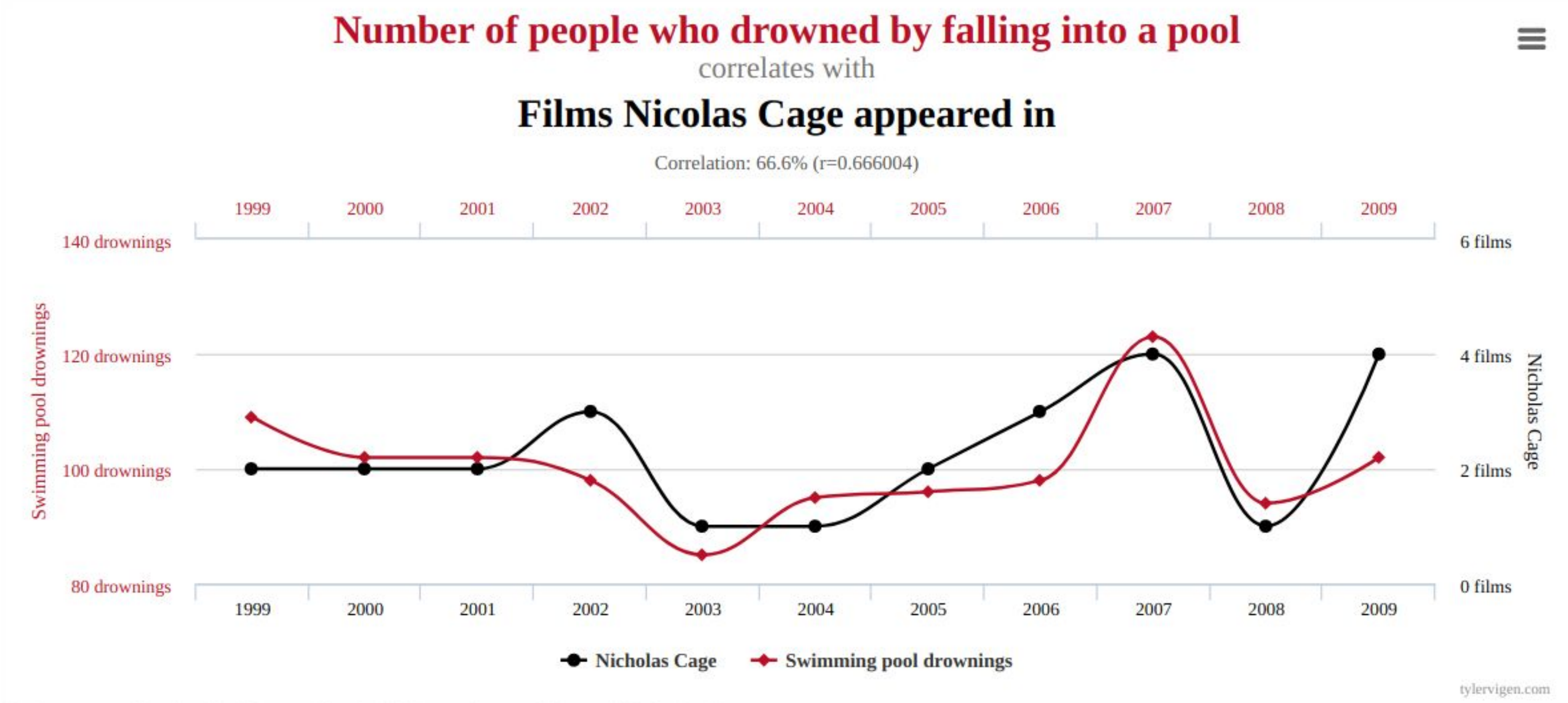
Pearl & Mackenzie. The book of why. 2019

Causal Hierarchy

Level	Typical Activity	Typical Questions	Examples
Association	Seeing	What is? How would seeing X changes my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
Intervention	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes? What happens if we double the price?
Counterfactuals	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Correlation does not imply causation!

Correlation is not causation!



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

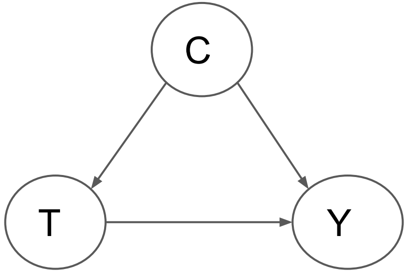
Simpson’s Paradox - Ex 1

New disease: COVID

Treatment T: A(0) and B(1)

Condition C: mild (0) or severe (1)

Outcome Y: alive(0) or dead(1)



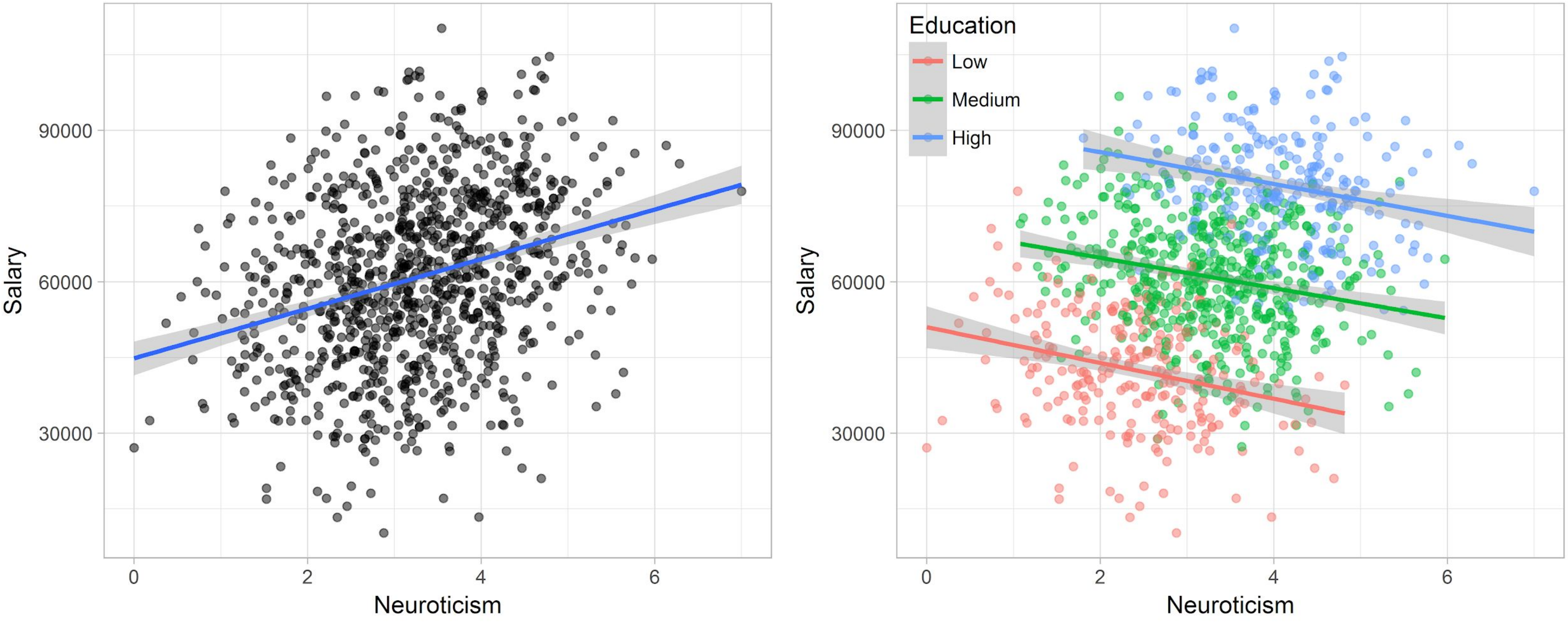
Condition

		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)
		$\mathbb{E}[Y\|T, C=0]$	$\mathbb{E}[Y\|T, C=1]$	$\mathbb{E}[Y\|T]$

1400/1500(0.15) +
100/1500(0.30) = 0.16

50/550(0.10) +
500/550(0.20) = 0.19

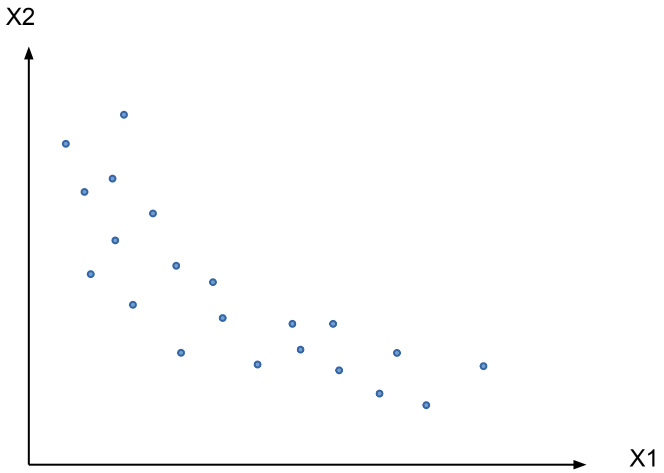
Simpson's paradox - Ex 2



Causality and ML

What do you think is the causal relation between X1 and X2?

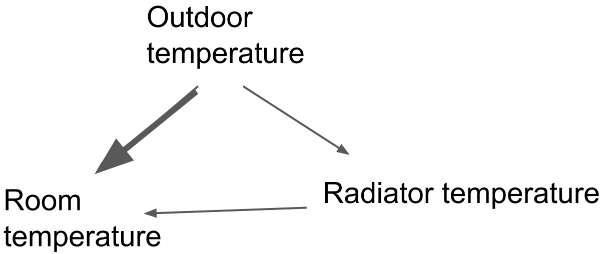
X1->X2
Or
X2->X1

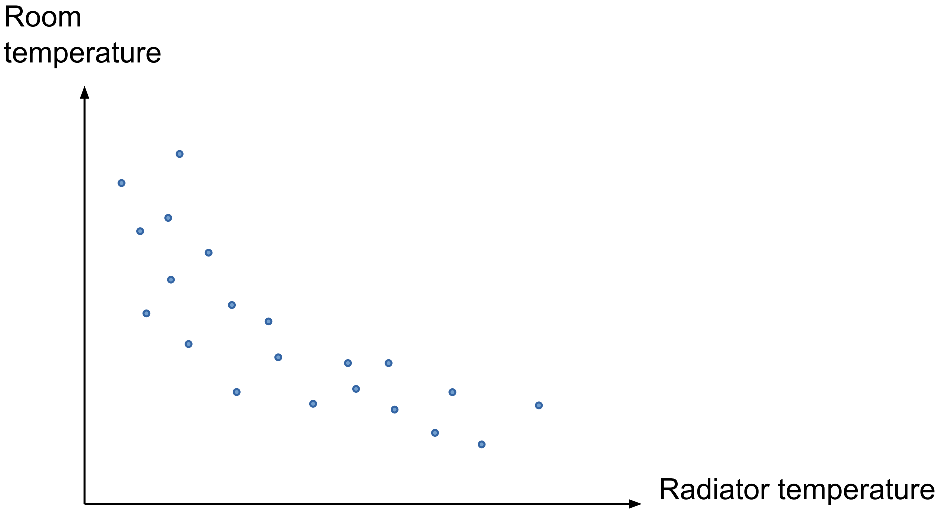


Causality and ML

Do you think our conclusion is correct if I tell you?

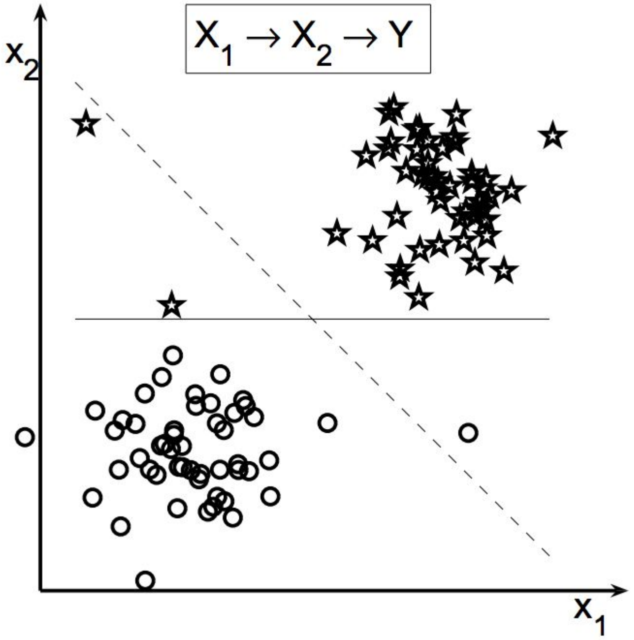
- X2: room temperature
- X1: radiator temperature





Causality and ML

Build a better predictor by using only variable X2



Causal feature selection, Guyon et. al.

What does imply causation?

Potential outcome

Inferring the effect of treatment/policy on some outcome

Example:

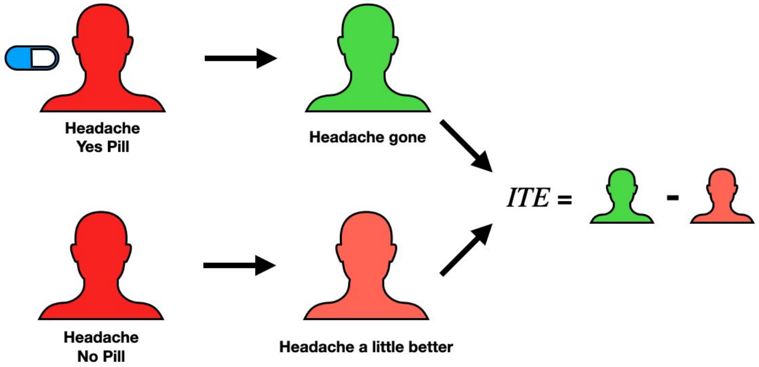
$$do(T=1) \longrightarrow Y_i|_{do(T=1)} = Y_i(1)$$

- T: observed treatment
  - T=1, T=0 → taking a pill, not taking a pill
- i: individual / sample
- Y: observed outcome
  - $Y_i(1)$ -> potential outcome of taking the pill

Causal effect

Individual treatment effect (ITE):

$$Y_i(1) - Y_i(0)$$



	$Y_i(1)$	$Y_i(0)$
Person 1	1	0
Person 2	0	1
Person 3	1	1
Person 4	0	0
Person 5	0	0
Person 6	0	1
Person 7	0	1
Person 8	1	1
Person 9	1	0
Person 10	1	0

Fundamental problem of causal inference

One outcome is **factual** and the other is **counterfactual**

Individual treatment effect (**ITE**):

$$Y_i(1) - Y_i(0) = ?$$

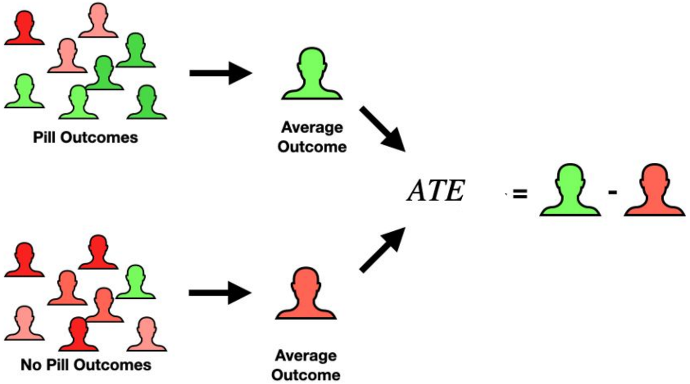
	$Y_i(1)$	$Y_i(0)$
Person 1	?	0
Person 2	0	?
Person 3	1	?
Person 4	?	0
Person 5	?	0
Person 6	?	1
Person 7	0	?
Person 8	1	?
Person 9	1	?
Person 10	?	0

Average treatment effect

Average treatment effect (**ATE**):

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

$$\cancel{=} \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$$

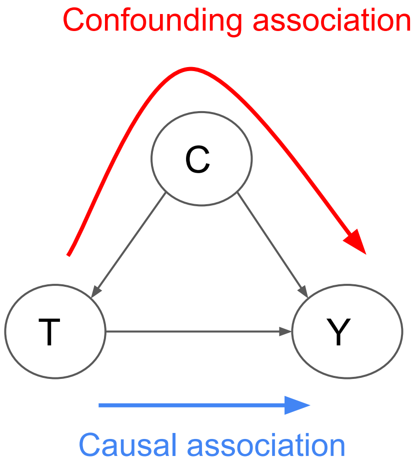


	$Y_i(1)$	$Y_i(0)$
Person 1	?	0
Person 2	0	?
Person 3	1	?
Person 4	?	0
Person 5	?	0
Person 6	?	1
Person 7	0	?
Person 8	1	?
Person 9	1	?
Person 10	?	0

Average treatment effect

Average treatment effect (**ATE**):

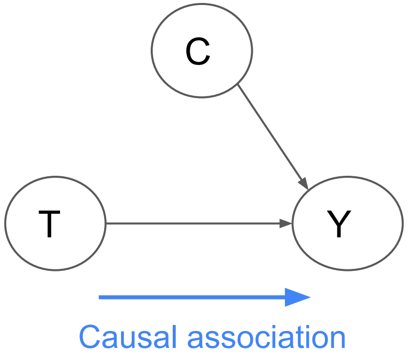
$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0)] &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &\neq \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]\end{aligned}$$

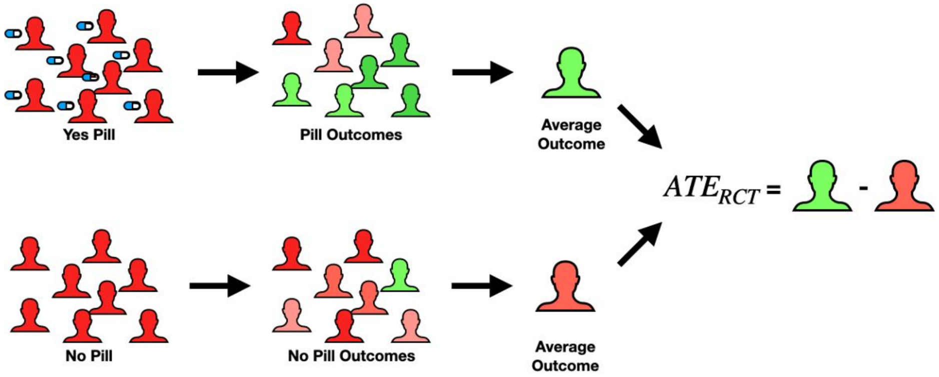


Randomised control trials (RCTs)

Treatment groups or control groups are selected randomly

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0)] &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]\end{aligned}$$





How can we discover causal relations?

- Correlation:
  - It is raining -> people probably carry open umbrellas
  - People carry open umbrellas -> It is probably raining

Not enough

- Intervention:
  - Will it rain if we ban umbrella?
  - Would it have rained if we had banned umbrellas?

Too difficult

- Randomized trials
  - Randomly split people in two groups
  - Force one group to carry the umbrella and force another group not to carry.
  - Measure the correlation of the rain

Sometimes impractical

- Observational data only
  - Selection on existing data

How to measure causal effect in observational studies?

If the Treatment is not assigned randomly, we need to adjust / control for confounders (W).

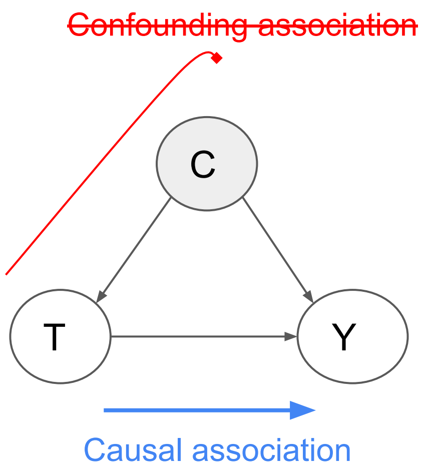
W is a sufficient adjustment set, if we have:

$$[Y(0), Y(1) \perp T | W]$$

then,

$$\mathbb{E}[Y(t)|W = w] = \mathbb{E}[Y|do(T = t), W = w] = \mathbb{E}[Y|t, w]$$

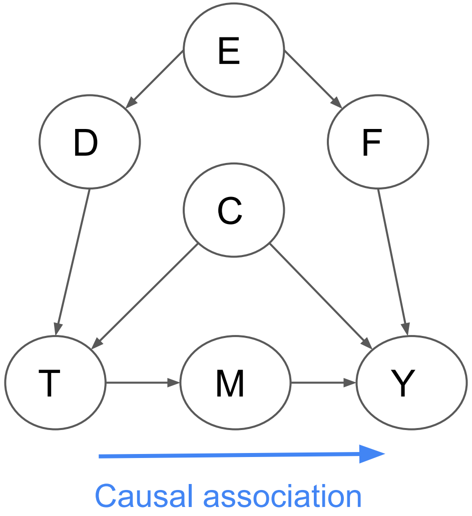
$$\mathbb{E}[Y(t)] = \mathbb{E}[Y|do(T = t)] = \mathbb{E}_W\mathbb{E}[Y|t, W]$$



Unconfoundedness Assumption

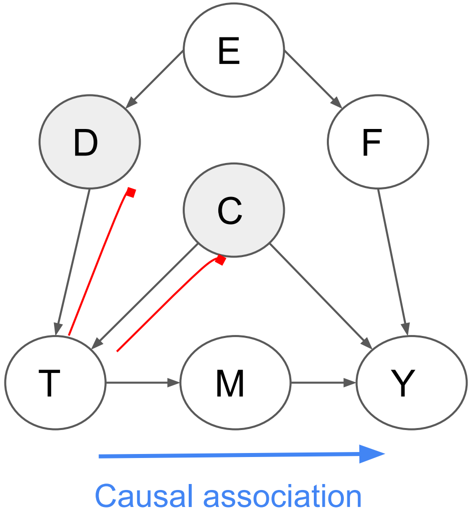
Backdoor adjustment

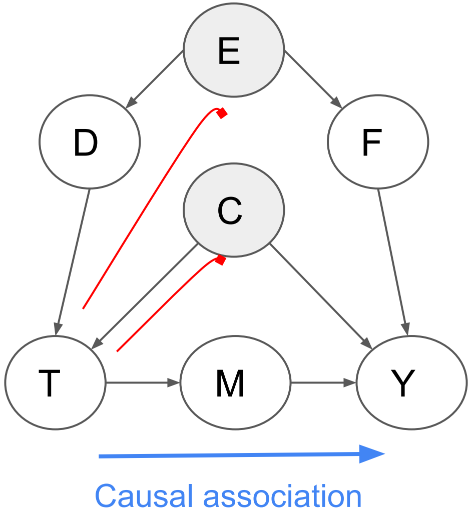
What are the minimum nodes that we need block (condition for) to remove all confounding associations between treatment and outcome.



Backdoor adjustment

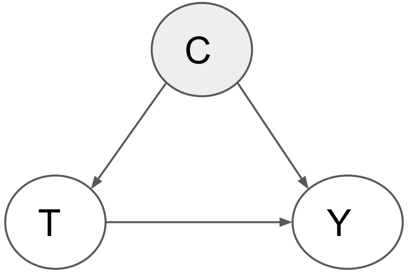
What are the minimum nodes that we need block (condition for) to remove all confounding associations between treatment and outcome.





Simpson’s Paradox - Ex 1

$$\mathbb{E}[Y|do(T=t)] = \mathbb{E}_C\mathbb{E}[Y|t,C] = \sum_c \mathbb{E}[Y|t,c]P(c)$$



		Condition				
		Mild	Severe	Total	Causal	
Treatment	A	15% (210/1400)	30% (30/100)	**16%** **(240/1500)**	19.4%	1450/2050(0.15) + 600/2050(0.30) ~= 0.19
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)	**12.9%**	1450/2050(0.10) + 600/2050(0.20) ~= 0.12
		$\mathbb{E}[Y \mid T, C=0]$	$\mathbb{E}[Y \mid T, C=1]$	$\mathbb{E}[Y \mid T]$	$\mathbb{E}[Y \mid do(t)]$	

Regression adjustment and unconfoundedness

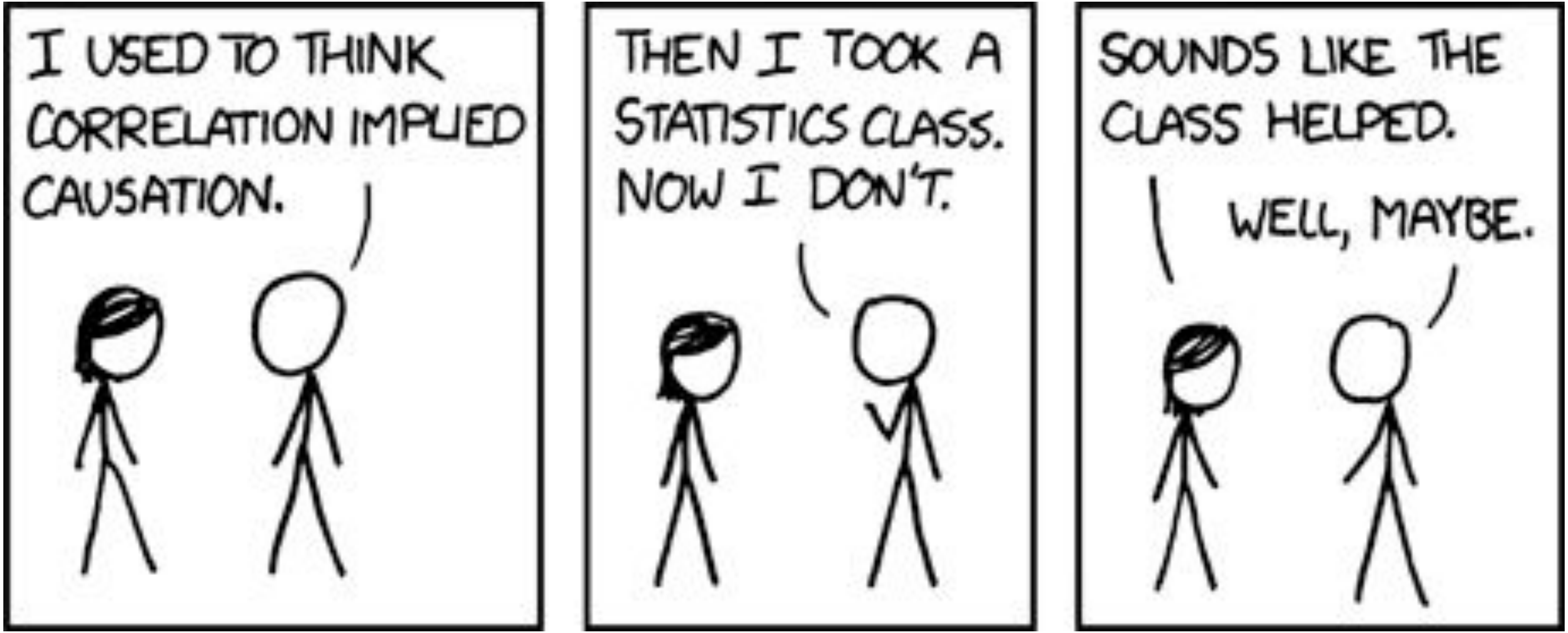
Identifying ATE: $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W[\mathbb{E}[Y|T=1,W] - \mathbb{E}[Y|T=0,W]]$

Estimation: $\frac{1}{n}\sum_w [\mathbb{E}[Y|T=1,w] - \mathbb{E}[Y|T=0,w]] = \frac{1}{n}\sum_w \underbrace{\mathbb{E}[Y(1)|w]}_{} - \underbrace{\mathbb{E}[Y(0)|w]}_{}]$

Model: Predicting subset of Y from W
on particular treatment group

Linear regression: $\hat{\mu}_{(t)}(w) = \beta_{(t)}w$

Estimation of ATE: $\frac{1}{n}\sum_i (\hat{\mu}_{(1)}(w_i) - \hat{\mu}_{(0)}(w_i)) = (\hat{\beta}_{(1)} - \hat{\beta}_{(0)})\frac{1}{n}\sum_i w_i = (\hat{\beta}_{(1)} - \hat{\beta}_{(0)})\bar{W}$


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Causal graphical models

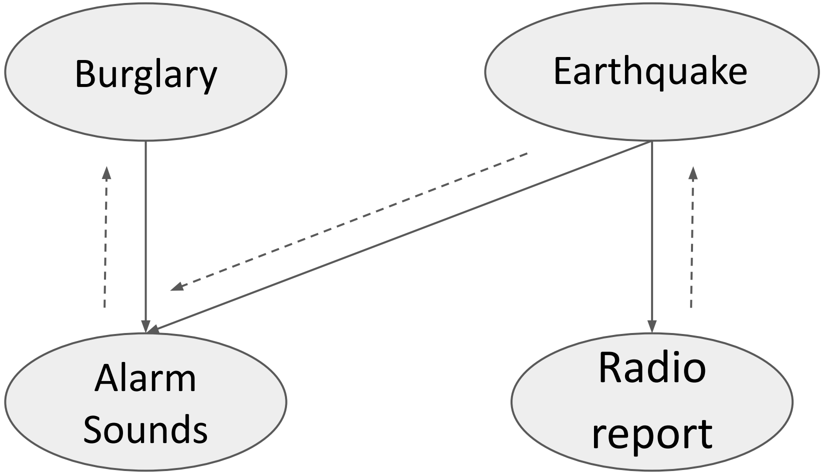
Holmes is told that the **burglary alarm** in his house is gone of.

He rushes into his car and heads for home. On his way, the radio reports a small earthquake.

He knows that earthquake has tendency to turn the burglar alarm on.

He returns to his work leaving his neighbors the pleasures of the noise.

Example by Pearl and Jensen


Burglary
Earthquake
Alarm Sounds
Radio report

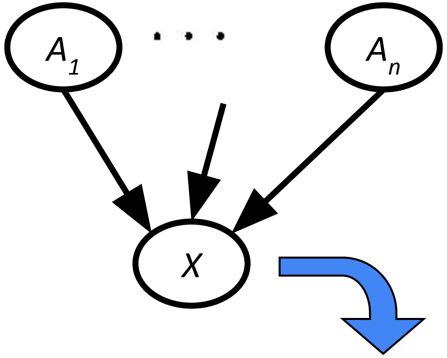
sherlockstuff | tumblr

BAYESIAN NETWORKS

Bayesian network representation

Directed acyclic graph (DAG)

- Nodes: variables
- Edges: interactions
  - Indicate “direct influence” between variables
- Formally: encode conditional independence
- For now: imagine that arrows mean direct noisy causation (in general, they are not causal!)



$P(X|A_1 \ldots A_n)$

Encodes joint distribution

- Set of conditional probability tables for each node in the graph

How to construct a Bayesian Network?

Expert determines nodes and links

How to construct a Bayesian Network?

- Expert determines nodes and links
- Estimating the network from the data:
  - **Score-based** learning
    - (1) defines a criterion to evaluate how well the Bayesian network fits the data, then (2) searches over the space of DAGs for a structure achieving the maximal score.
      - Bayesian scoring functions, e.g. BD (Bayesian Dirichlet) (1995), K2 (1992)
      - Information-theoretic scoring functions, e.g. LL (Log-likelihood) (1912-22), MDL/BIC (Minimum description length/Bayesian Information Criterion) (1978), MIT (Mutual Information Tests) (2006)

How to construct a Bayesian Network?

- Expert determines nodes and links
- Estimating the network from the data:
  - **Score-based** learning
  - **constraint-based** learning
    - employs the independence test to identify a set of edge constraints for the graph and then finds the best DAG that satisfies the constraints
    - PC algorithm (2000), FCI (2001), …

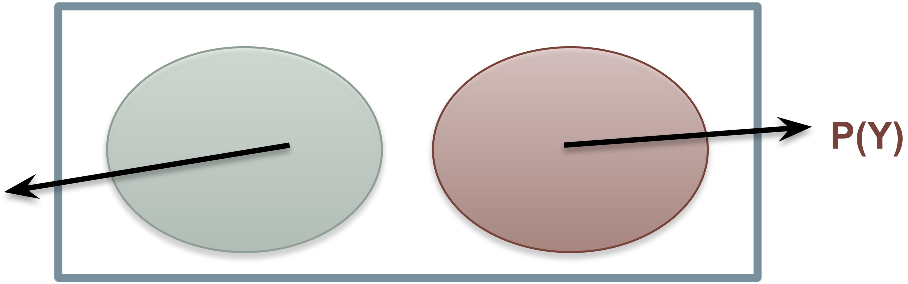
Independence

Two variables are *independent* if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

Another form: $\forall x, y : P(x|y) = P(x)$

We write: $X \perp\!\!\!\perp Y$



Conditional Independence

X is conditionally independent of Y given Z

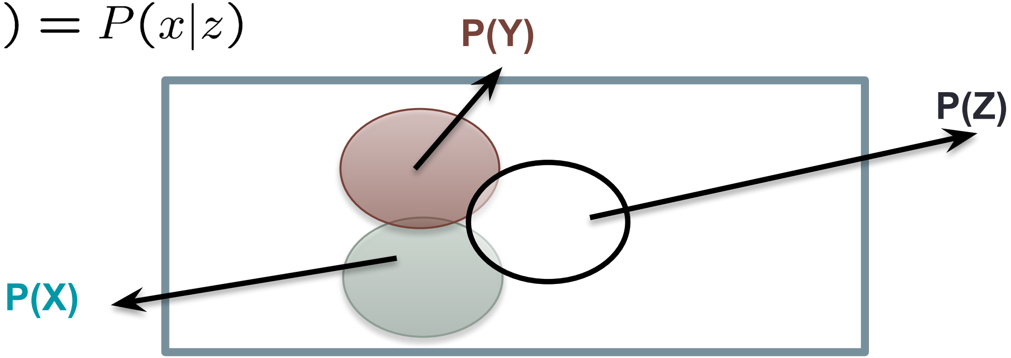
if and only if:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

We write $X \perp\!\!\!\perp Y | Z$



Causal networks

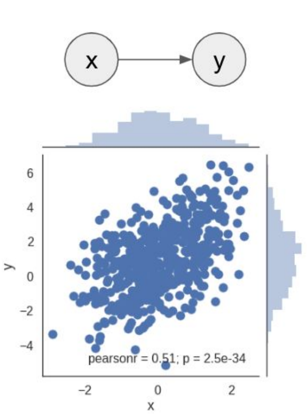
Bayesian networks are usually used to represent causal relationships. This is, however, not strictly necessary: a directed edge from node i to node j does not require that X_i is causally dependent on X_j.

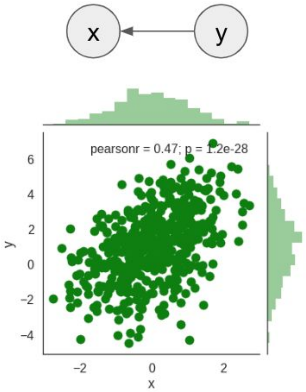
This is demonstrated by the fact that many causal mechanism construct the same distributions

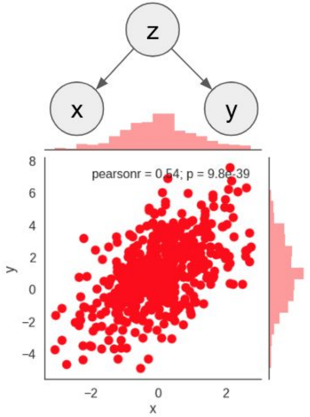
```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```







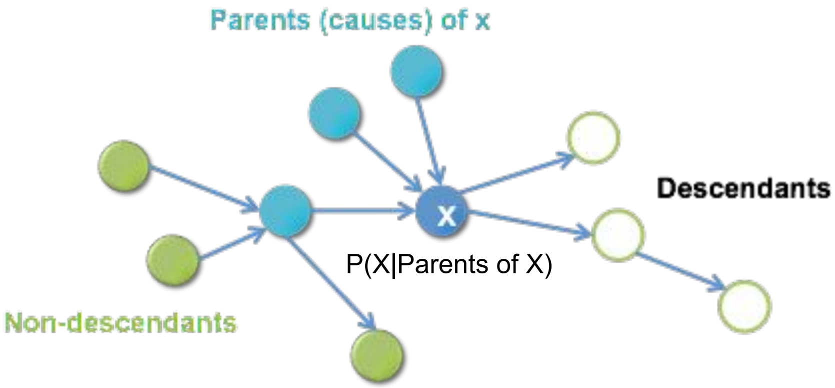
Cause and Effect

particular independent variable (the cause) has an effect on the dependent variable of interest (the effect)

> A variable X is said to be a cause of a variable Y if Y can change in response to changes in X.
>
> Y = f (X, noise)

Causal Model (Pearl et al.)

- Set of variables X1, . . . ,Xn on a directed acyclic graph G.
- Arrows = direct causal links (come from either the expert or the data)
- X = f(Parents of x, Noise)



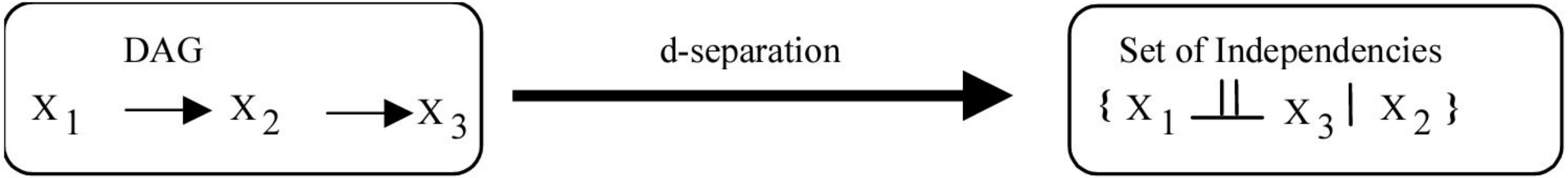
- Implies p(X1, . . . ,Xn) with particular conditional independence structure:
  - **Causal Markov condition**:

X independent of non-descendants, given parents

$$P(x|PA_x, non - decendents) = P(x|PA_x)$$

d-separation

If sets of variables X and Y are d-separated by a set Z in the DAG G, then X and Y are conditionally independent conditional on Z.



DAG

$X_1 \longrightarrow X_2 \longrightarrow X_3$

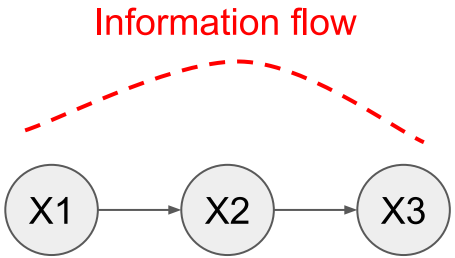
d-separation

Set of Independencies

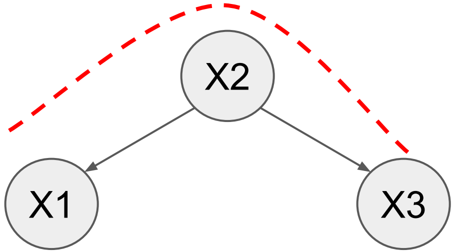
$\{ X_1 \perp\!\!\!\perp X_3 \mid X_2 \}$

d-separation

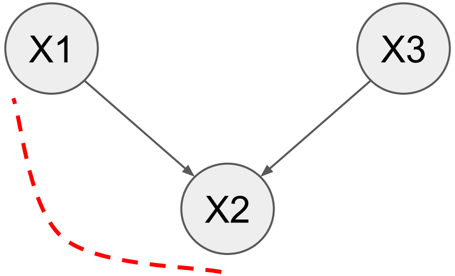
Two (sets of) nodes X and Y are d-separated by a set of nodes Z if all of the paths between (any node in) X and (any node in) Y are blocked by Z.



Chain



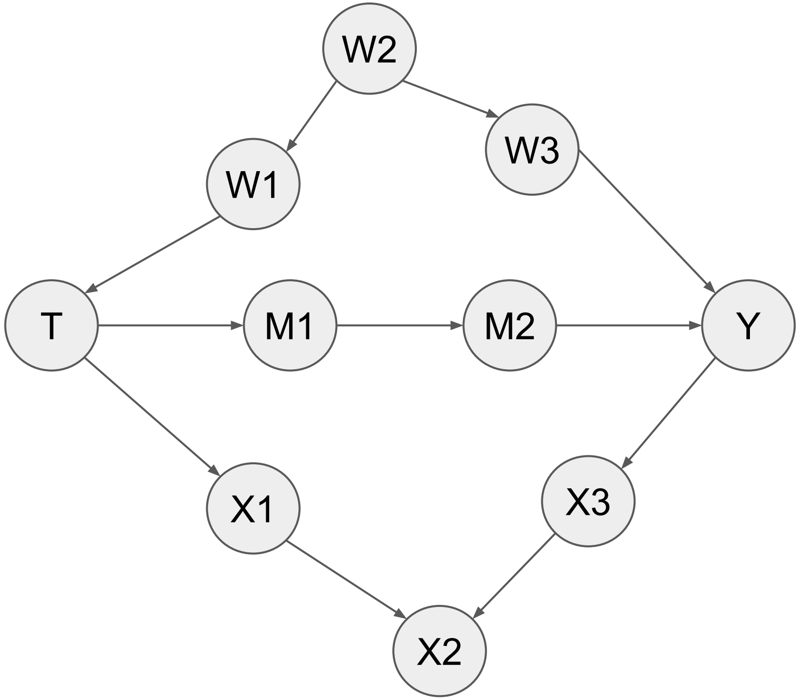
Fork



Collider

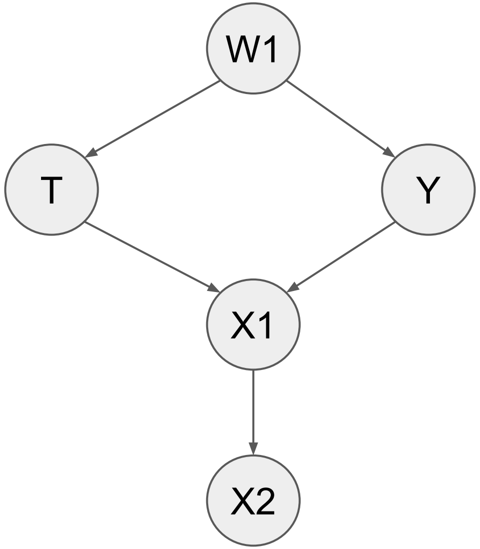
d-separation example I

- What is the d-separation set between T and Y?



d-separation example II

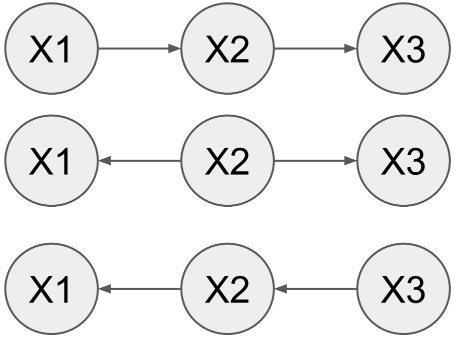
- What is the d-separation set between T and Y?



Markov equivalent classes

Graphs with same skeleton and same conditional independence.

E.g. $X_1 \perp\!\!\!\perp X_3 | X_2$ and $X_1 \not\!\perp\!\!\!\perp X_3$

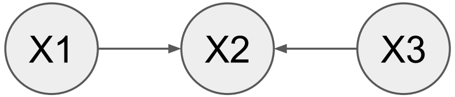


They can not being distinguished only by looking at conditional independence.

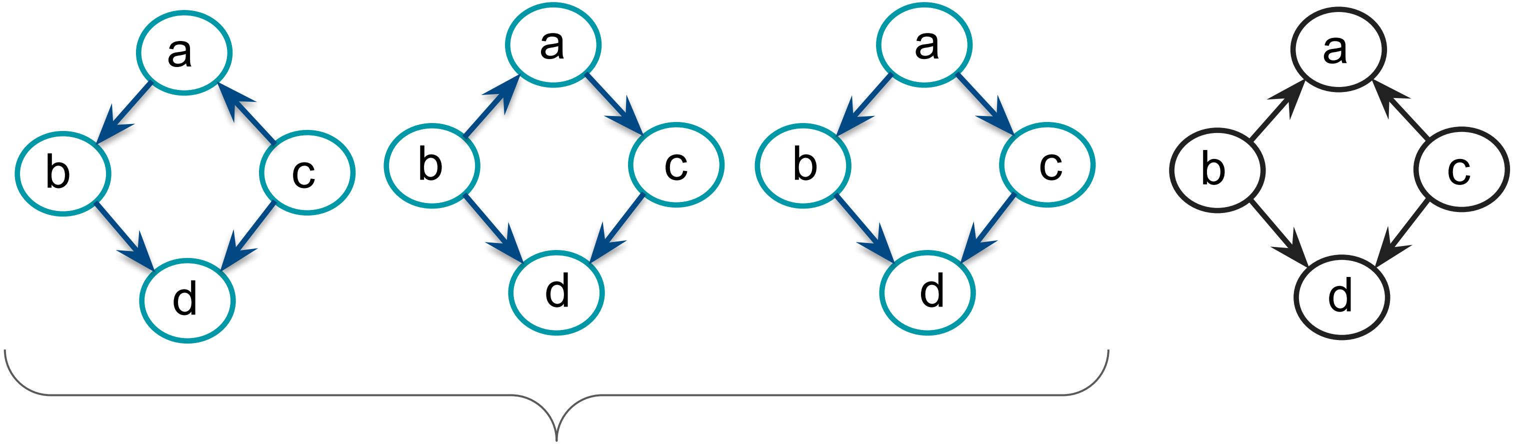
Markov equivalent classes

Graphs with same skeleton and same conditional independence.

Collider case $X_1 \not\perp\!\!\!\perp X_3 | X_2$ and $X_1 \perp\!\!\!\perp X_3$



Markov equivalent classes, e.g.

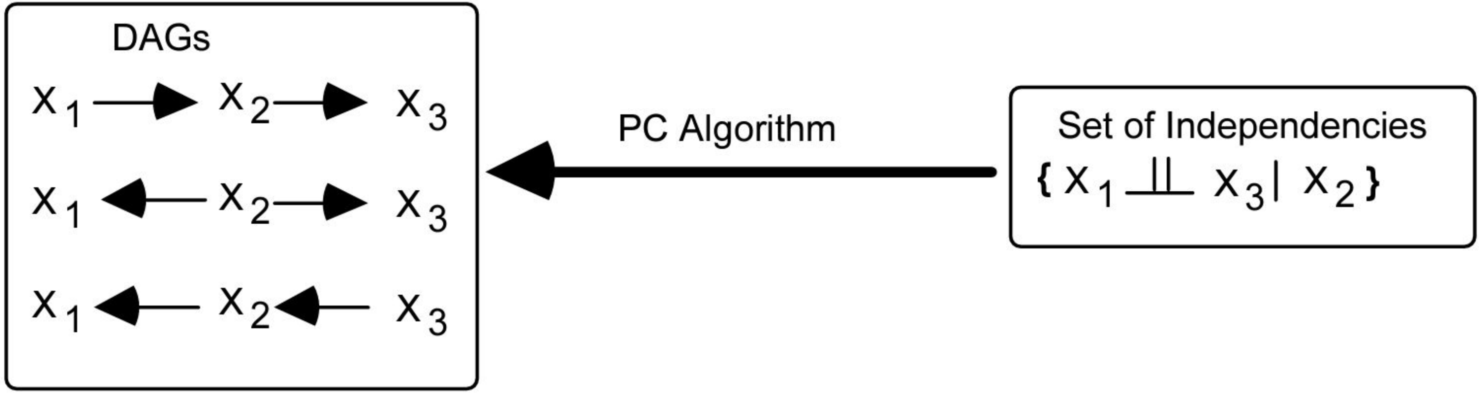


Markov equivalence class

How to create graphs in a same Markov equivalent class?

Construct Causal Network from data - PC algorithm

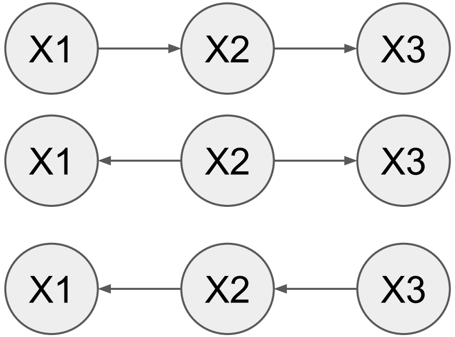
Performing **condition independence test** on different subset variables to calculate the causal directions:


DAGs
$X_1 \rightarrow X_2 \rightarrow X_3$
$X_1 \leftarrow X_2 \rightarrow X_3$
$X_1 \leftarrow X_2 \leftarrow X_3$
PC Algorithm
Set of Independencies
$\{X_1 \perp\!\!\!\perp X_3 | X_2\}$

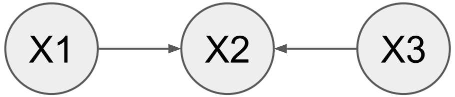
Construct Causal Network from data - PC algorithm

Performing **condition independence test** on different subset variables to calculate the causal directions:

X1 is independent of X3 given X2



X1 and X3 are independent, they become dependent given X2

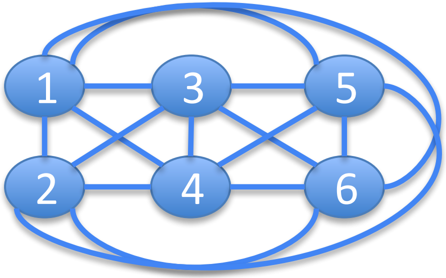


Construct Causal Network from data - PC assumptions

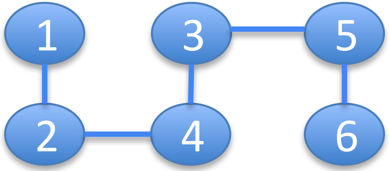
- Markov assumption
  - A variable is independent of non-descendants, given its parents
- Faithfulness
  - Two path between two variable should not cancel out each other effects
- Causal sufficiency
  - There are no unobserved confounders of any of the variables in the graph
- Acyclicity
  - There are no cycles in the graph

Causal graph from observational data

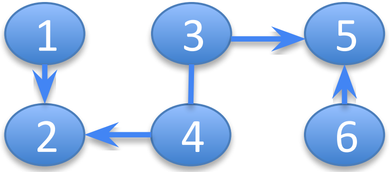
PC algorithm: conditional independence based algorithm



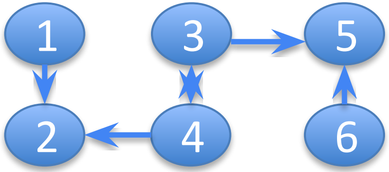
Initialize with a fully connected un-oriented graph



Step1: An edge a-b is deleted if $a \perp b|c$



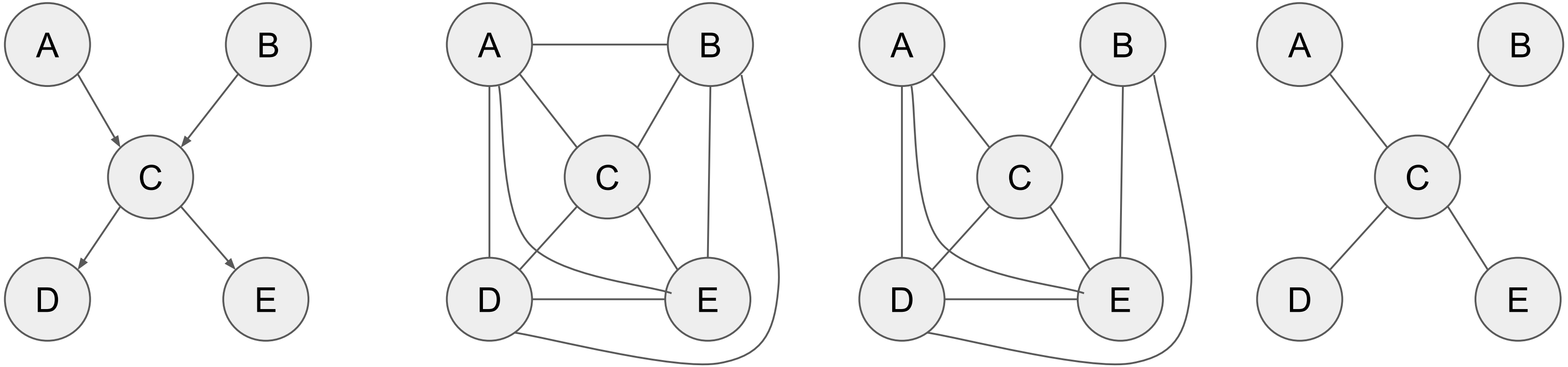
Step 2: Orient edges in “collider” triplets



Step 3: Further orient edges with a constraint-propagation

PC - Step 1: Identifying the skeleton

Start with complete undirected graph and remove edges X – Y where $X \perp\!\!\!\perp Y|Z$ for some (potentially empty) conditioning set Z, starting with the empty conditioning set and increasing the size.



Ground Truth

Step 1

$A \perp\!\!\!\perp B|\{\}$

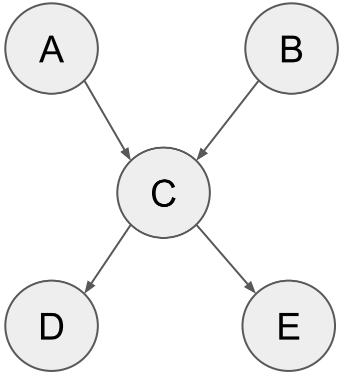
$X \perp\!\!\!\perp Y|\{C\}, where\ X, Y \in \{A, B, D, E\}$

PC - Step 2: Identifying the colliders

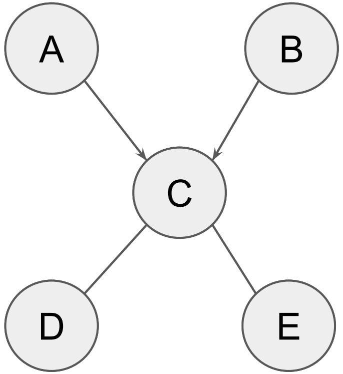
Now for any paths X – Z – Y in our working graph where the following are true:

1. We discovered that there is no edge between X and Y in our previous step.
2. Z was not in the conditioning set that makes X and Y conditionally independent

Then, we know X – Z – Y forms a collider



Ground Truth



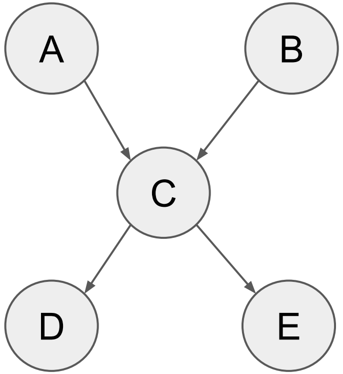
$A \perp\!\!\!\perp B | \{\}$

$A \not\perp\!\!\!\perp B | \{C\}$

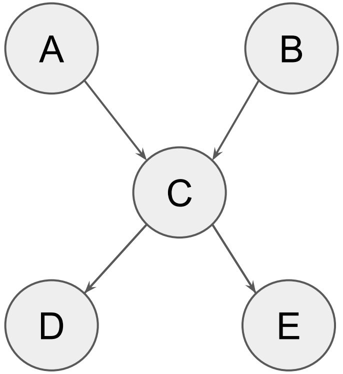
PC - Step 3: Orienting remaining edges

Idea: use fact that we discovered all colliders

Any edge Z-Y part of a partially directed path of the form X->Z-Y, where there is no edge connecting X and Y can be oriented as Z->Y



Ground Truth



Removing some assumptions

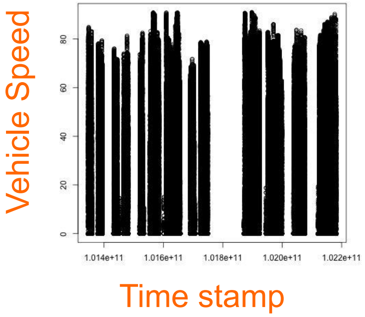
- No assumed causal sufficiency: FCI algorithm (Spirtes et al., 2001)
- No assumed acyclicity: CCD algorithm (Richardson, 1996)
- Neither causal sufficiency nor acyclicity: SAT-based causal discovery (Hyttinen et al., 2013; 2014)

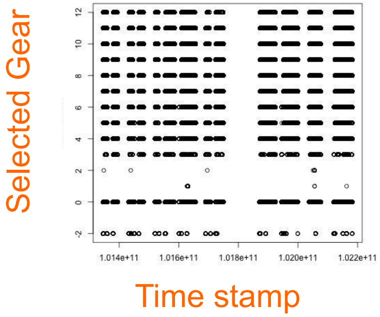
Challenges of conditional-based causal discovery

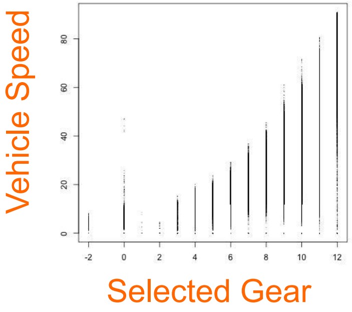
- Independence-based causal discovery algorithms rely on accurate conditional independence testing.
    - Conditional independence testing is simple if we have infinite data.
    - However, it is a quite hard problem with finite data, and it can sometimes require a lot of data to get accurate test results (Shah & Peters, 2020).
- Markov equivalence class is hard to interpret
- Causal graphs are not robust to input data, noise, and outliers
- In many real world application, the true causal graph is not known

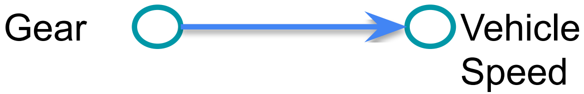
Causal discovery in real world setting

Causal relation between vehicle speed and selected gear



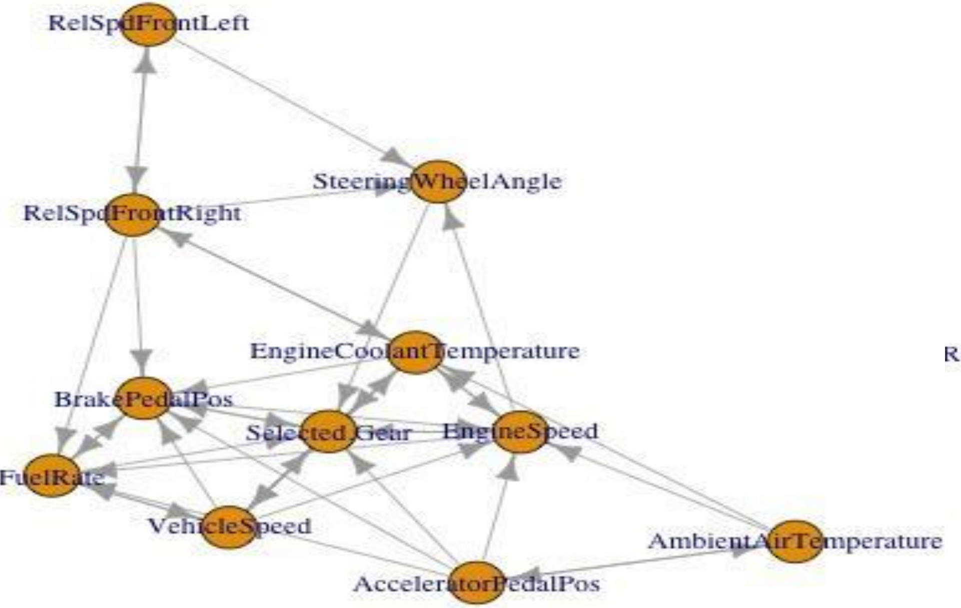




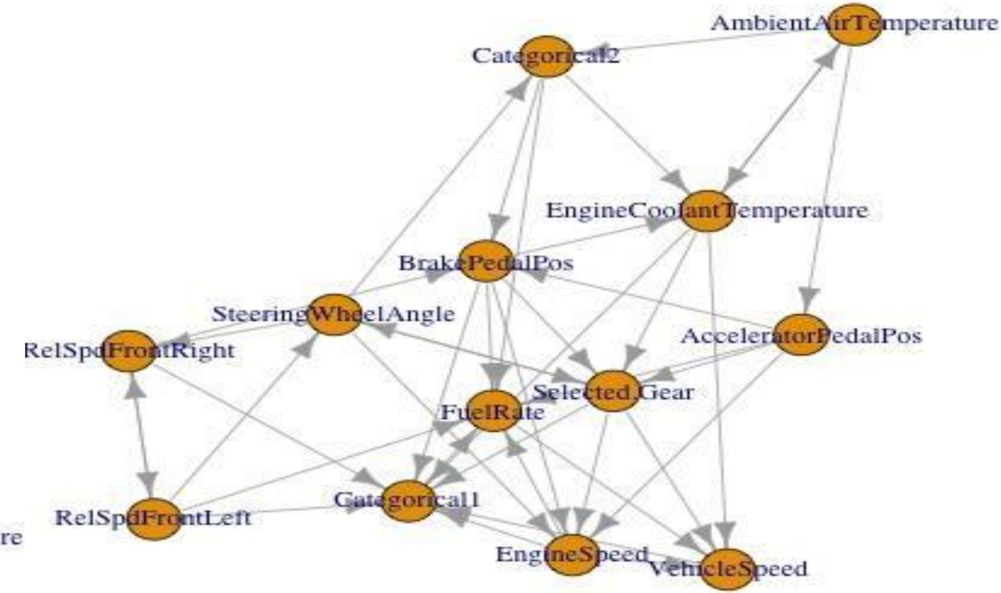


Causal discovery in real world setting

- 6 busses , 11 signals
- latent variables: “idle run” and “model year”



Without categorical variables



With categorical variables

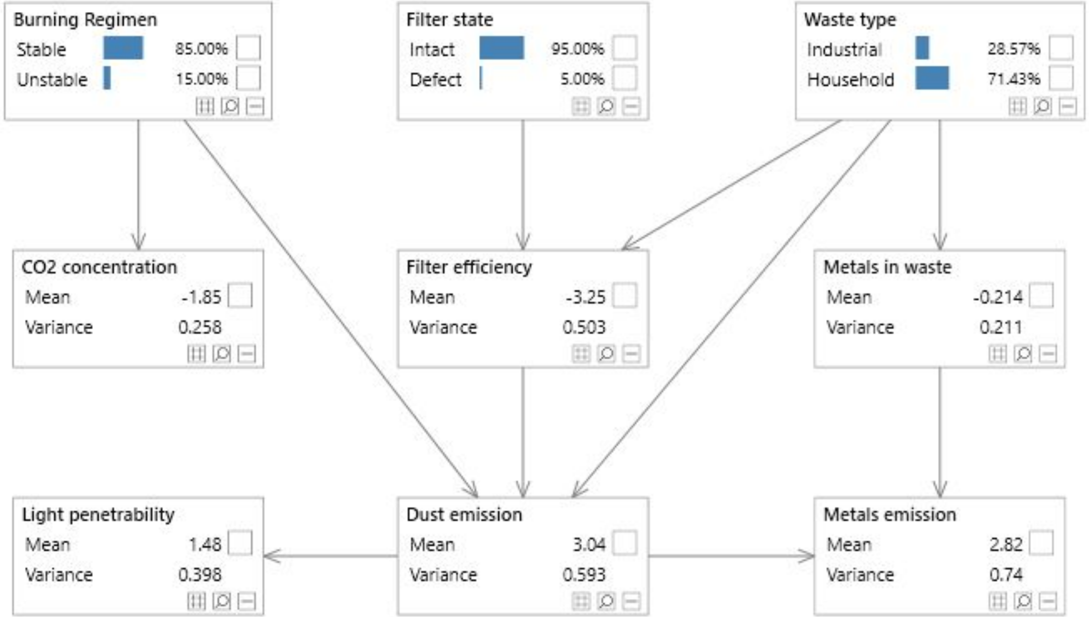
Causal discovery is still worthwhile

Get insight about the cause

- If we do inference
- If we ask the proper counterfactual question
- Evaluate the results based on “usefulness”
- Calculate causal effect

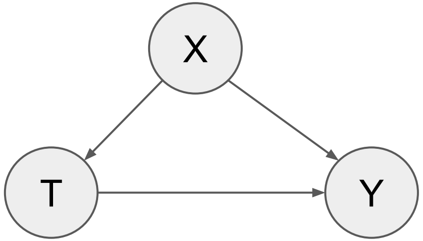
Inference

Given a network describing $P(X_1, X_2, , X_n)$, what is $P(x_i|x_j, x_k)$?



Inference

We can do intervention on a node, e.g. calculating P(y|do(t))

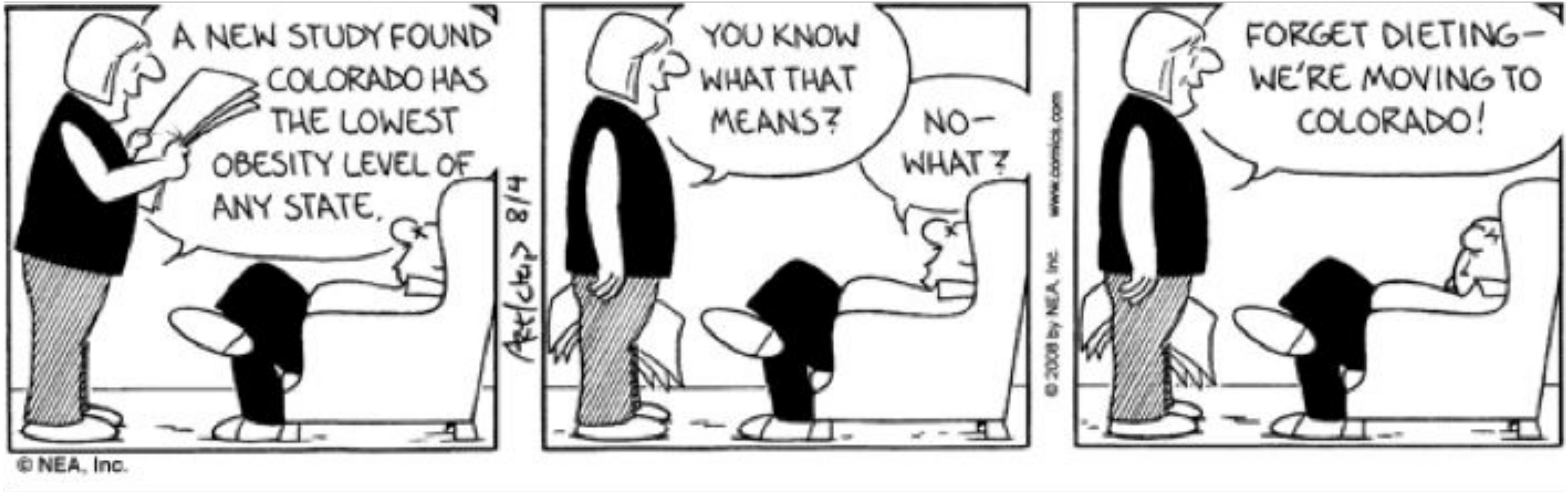


- Bayesian Factorization: P(x,t,y) = P(x)P(t|x)P(y|x,t)
- Because of intervention on t, P(t|x)=1

  P(x,y|do(t)) = P(x)P(y|x,t)

- Marginalization: $P(Y|do(t)) = \sum_x P(x)P(y|x,t)$

Correlation vs. Causation


A NEW STUDY FOUND COLORADO HAS THE LOWEST OBESITY LEVEL OF ANY STATE.
Art/cleirz 8/4
YOU KNOW WHAT THAT MEANS?
NO- WHAT?
© 2008 by NEA, Inc. www.comics.com
FORGET DIETING- WE'RE MOVING TO COLORADO!
© NEA, Inc.

Complementary materials

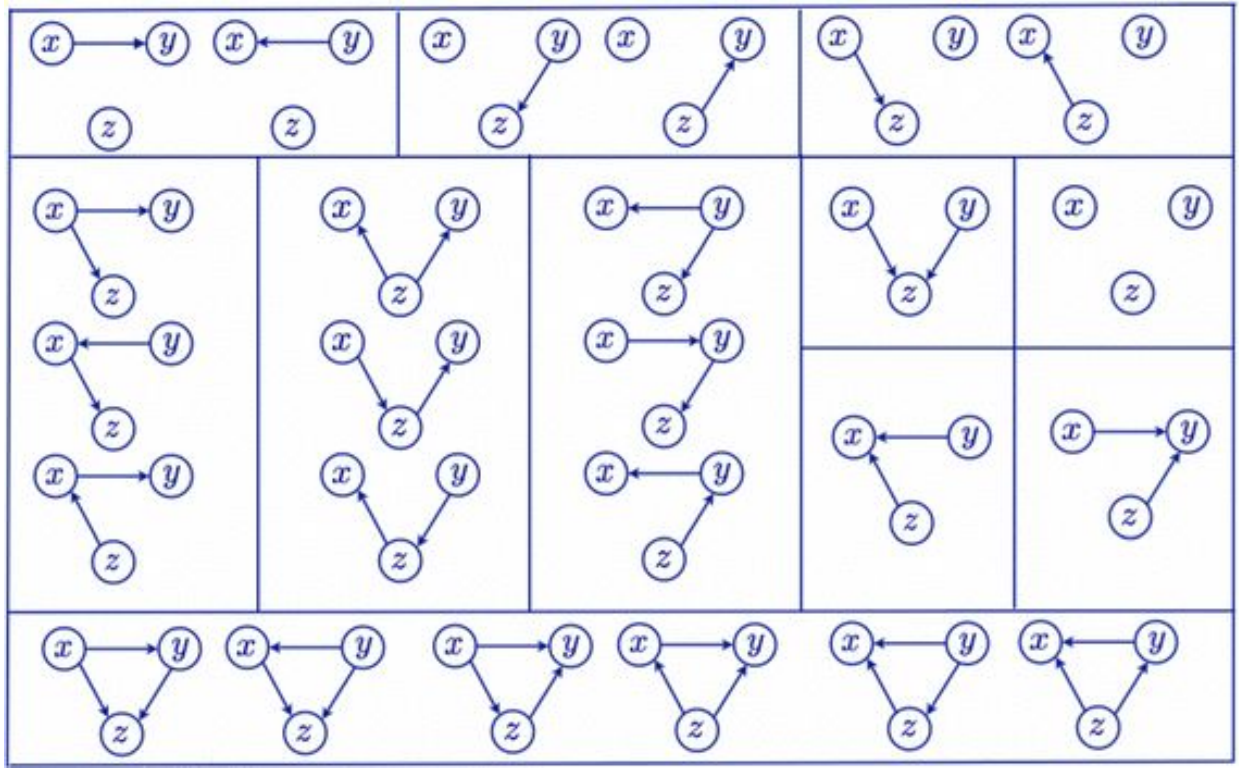
An Introduction to Causal Inference, Richard Scheines
https://www.cmu.edu/dietrich/philosophy/docs/scheines/introtocausalinference.pdf

Structure learning for Bayesian networks
https://ermongroup.github.io/cs228-notes/learning/structure/

Other Causal Discovery Approaches

Issues with Independence-Based Causal Discovery

- Requires **faithfulness assumption**
- **Large samples** can be necessary for conditional independence tests
- Only identifies the **Markov equivalence class**



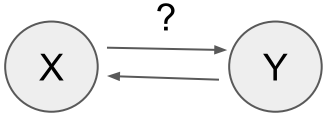
Can we do better than Markov equivalence class?

If we have **multinomial distributions** (Meek, 1995) or **linear Gaussian structural equations** $X_j = f_j(PA_j, N_j)$ (Geiger & Pearl, 1988), we can only identify a graph up to its Markov equivalence class.

- What about non-Gaussian structural equations?
- Or nonlinear structural equations?

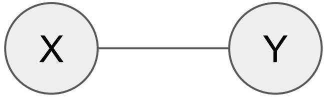
Two Variable Case

Is the causal direction from X to Y or the reverse?



- There exist functions, f_X and f_Y such that:
  - $Y = f_Y(X, U_Y), X \perp\!\!\!\perp U_Y$
  - $X = f_X(Y, U_X), Y \perp\!\!\!\perp U_X$
  - X, Y, U_X , U_Y are real-valued random variables

Note: Without **extra assumptions about the parametric form**, we can not recognise the direction.



Linear Non-Gaussian Setting

Linear Non-Gaussian Assumption

Recall: We cannot hope to identify the graph more precisely than the Markov equivalence class in the linear Gaussian noise setting (Geiger & Pearl, 1988).

What if the **noise is non-Gaussian**?

Linear Non-Gaussian Assumption:

All structural equations (causal mechanisms that generate the data) are of the following form:

$$Y = f(X) + U$$

where f is a linear function, X is independent of U, and U is distributed as some non-Gaussian

Linear Non-Gaussian Assumption

Theorem (Shimizu et al., 2006):

In the linear non-Gaussian setting,

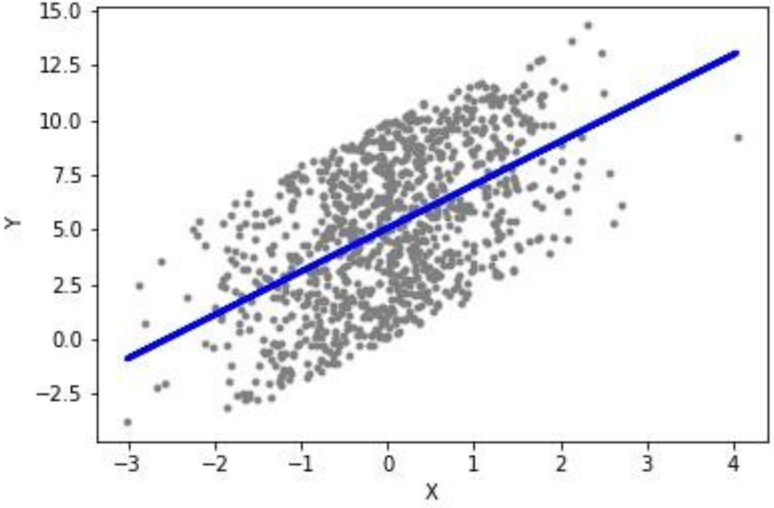
- if it exists the following function

$$Y = f(X) + U, X \perp\!\!\!\perp U$$

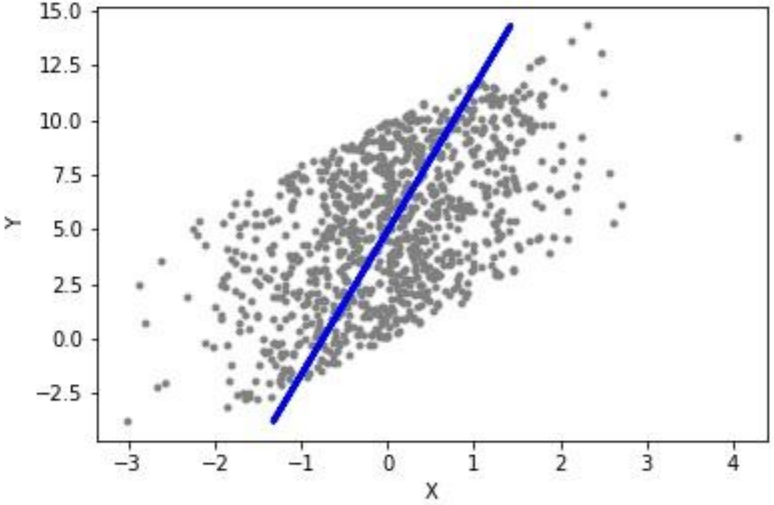
- then, it does not exist the following relation on the reverse direction

$$X = g(Y) + \tilde{U}, Y \perp\!\!\!\perp \tilde{U}$$

Example of Linear Non-Gaussian Setting

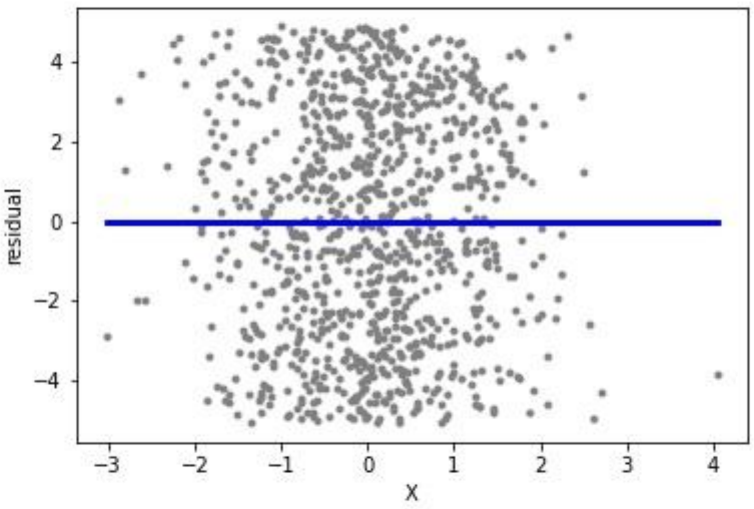


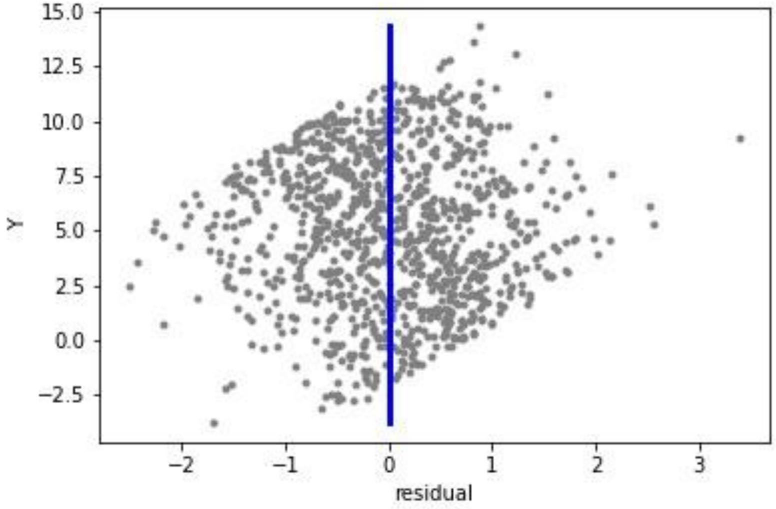
$$Y = f(X) + U$$



$$X = g(Y) + \tilde{U}$$

Example of Linear Non-Gaussian Setting





$X \perp\!\!\!\perp U$

$Y \not\!\perp\!\!\!\perp \tilde{U}$

LiNGAM algorithm

Linear, Non-Gaussian, Acyclic causal Models based on purely observational, continuous-valued data. Assumptions are

(a) there are no hidden confounders

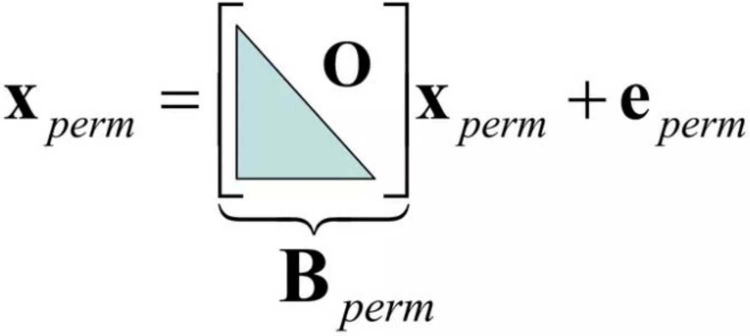
(b) the error terms are non-gaussian.

Under these conditions it can be shown that the **full generating model can be identified** in the limit of an infinite sample.

https://www.cs.helsinki.fi/group/neuroinf/lingam/JMLR06.pdf

LinGAM algorithm

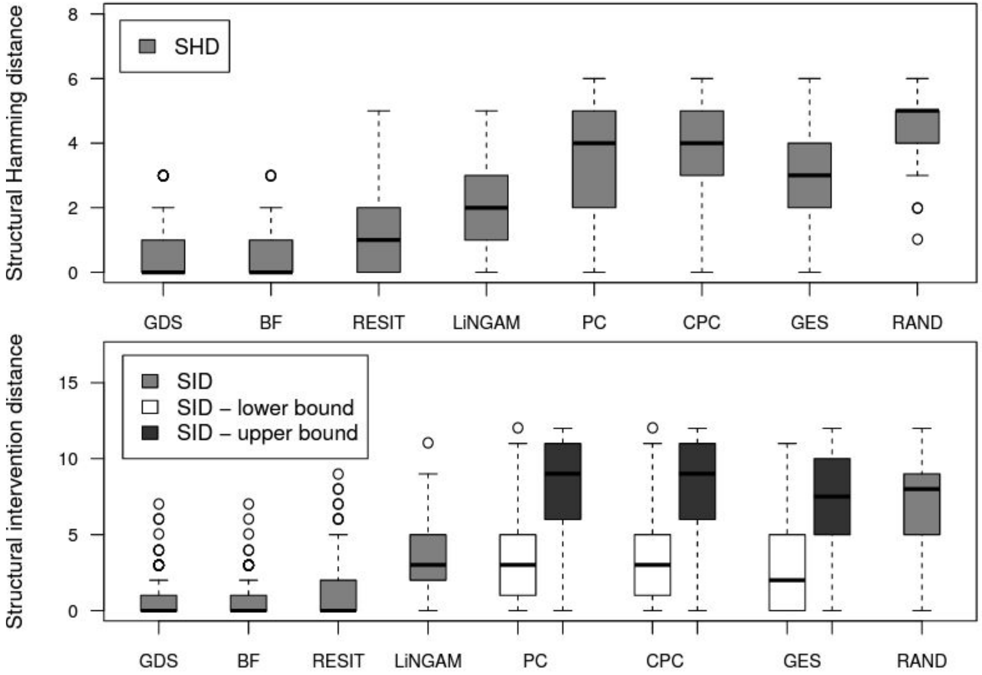
1. The observed variables xi, i ∈{1,...,m} can be arranged in a causal order, such that no later variable causes any earlier variable.
2. The value assigned to each variable xi is a linear function of the values already assigned to the earlier variables, plus a ‘disturbance’ (noise) term ei, and plus an optional constant term ci.
3. The disturbances ei are all continuous-valued random variables with non-Gaussian distributions of non-zero variances, and the ei are independent of each other.



Idea: Find a permutation of variables that sort them based on causal order

https://www.cs.helsinki.fi/group/neuroinf/lingam/JMLR06.pdf

Which one is better?



Box plots of the SHD between the estimated structure (either DAG or CPDAG) and the correct DAG for $p = 4$ and $n = 100$ for linear non-Gaussian SEMs (top). The SID is computed between the correct DAG and the estimated DAG (bottom). Some methods estimate only the Markov equivalence class. We then compute the SID to the "best" and to the "worst" DAG within the equivalence class; therefore a lower and an upper bound are shown.

https://jmlr.org/papers/volume15/peters14a/peters14a.pdf

Nonlinear Additive Noise Setting

Nonlinear Additive Noise Setting

Recall: We cannot hope to identify the graph more precisely than the Markov equivalence class in the linear Gaussian noise setting (Geiger & Pearl, 1988).

What if the structural equations are nonlinear?

Nonlinear additive noise assumption:

$Y = f(X) + U, X \perp\!\!\!\perp U$ where f is a **nonlinear function**.

Theorem (Hoyer et al. 2008): Under the Markov assumption, causal sufficiency, acyclicity, the nonlinear additive noise assumption, and a technical condition from Hoyer et al. (2008), we can identify the causal graph.

Nonlinear Additive Noise Setting

Algorithm:

1. test whether x and y are statistically independent.

2. If not, calculate a model Y = f(X)+U , e.g. using a nonlinear regression of y on x.

3. calculating the corresponding residuals U = Y − f(X), and testing whether U is independent of X. If yes, there is a causal link from X to Y, otherwise there is no link in this direction.

4. then, similarly test whether the reverse model X = g(Y)+U’ fits the data.

https://proceedings.neurips.cc/paper_files/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf

Abalone dataset

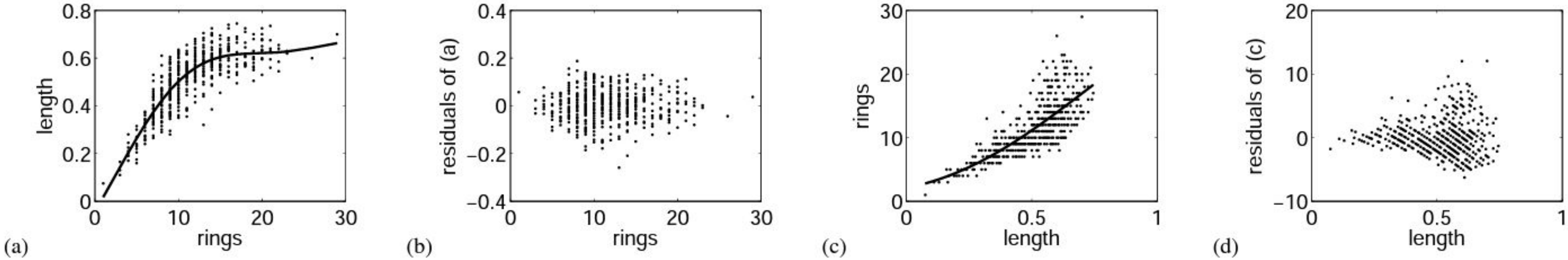
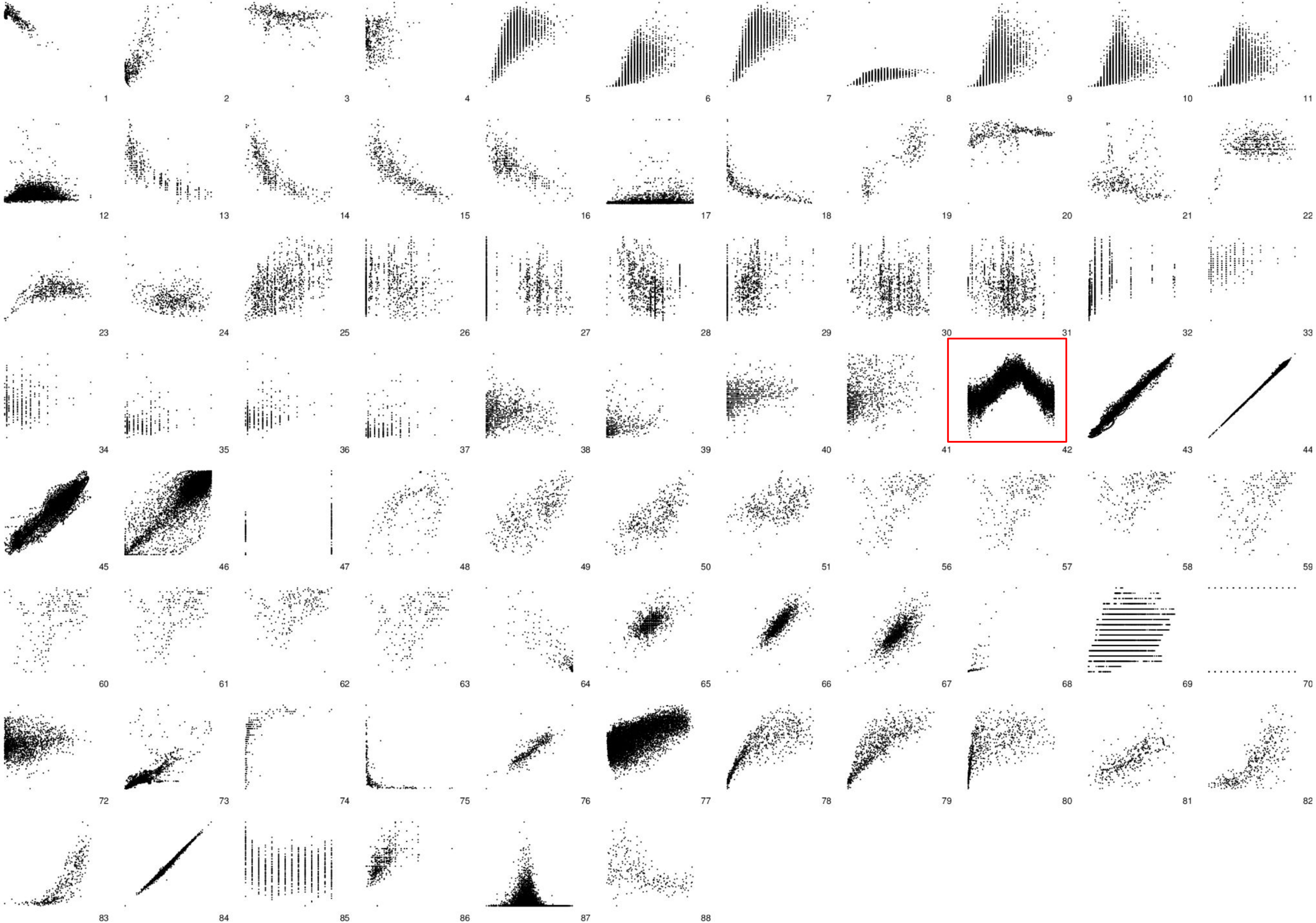
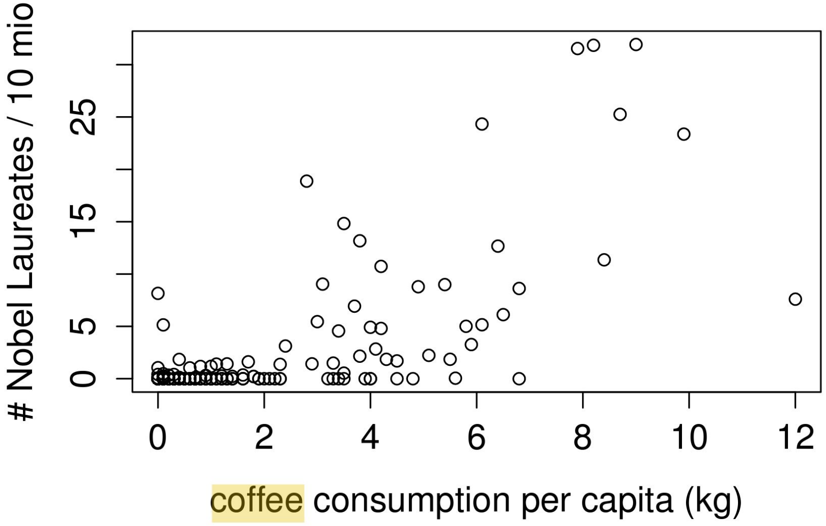


Figure 4: Abalone data: (a) forward fit corresponding to "age (rings) causes length"; (b) residuals for forward fit; (c) backward fit corresponding to "length causes age (rings)"; (d) residuals for backward fit.

https://proceedings.neurips.cc/paper_files/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88



Correlation: 0.698

Coffee -> Nobel Prize: Dependent residuals (p-value of 5.1*10^-78)
Nobel Prize -> Coffee: Dependent residuals (p-value of 3.1*10^-12)

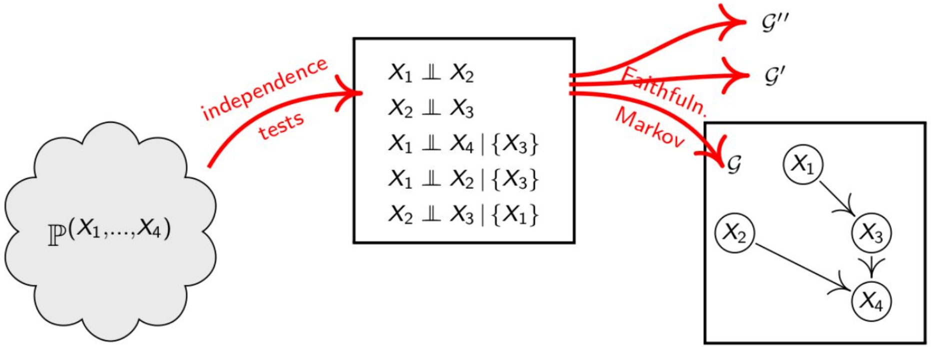
Cause-Effect Pairs

- For each pair of variables (Xi,Yi), the two possible additive noise models are tested that correspond with the two different possible causal directions, Xi → Yi and Yi → Xi.
- For both directions, the functional relationship are estimated by performing regression. The goodness-of-fit is then evaluated by testing independence of the residuals and the inputs.
- then, the pairs are ranked according to the highest of the two p-values of the pair.
- In this way, the trade off is created between accuracy, i.e., percentage of correct decisions, versus the amount of decisions taken.

https://jmlr.org/papers/volume15/peters14a/peters14a.pdf

Summary of ideas for causal discovery from observational data

Idea 1: independence-based methods



$X_1 \perp\!\!\!\perp X_2$

$X_2 \perp\!\!\!\perp X_3$

$X_1 \perp\!\!\!\perp X_4 \mid \{X_3\}$

$X_1 \perp\!\!\!\perp X_2 \mid \{X_3\}$

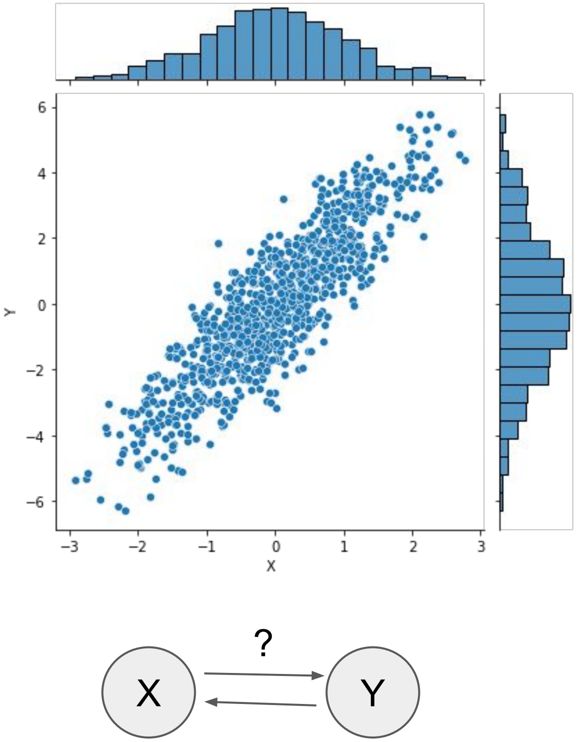
$X_2 \perp\!\!\!\perp X_3 \mid \{X_1\}$

Idea 2: additive noise

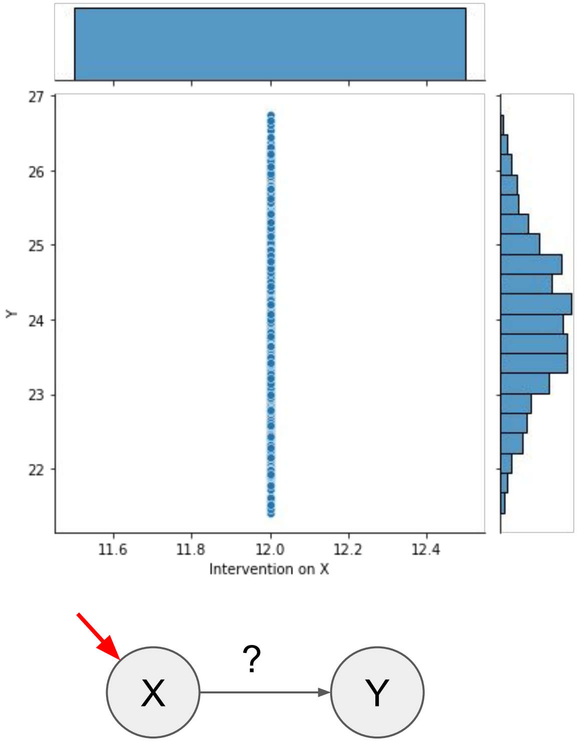
$$
\begin{aligned}
X_1 &= f_1(X_3) + N_1 \\
X_2 &= N_2 \\
X_3 &= f_3(X_2) + N_3 \\
X_4 &= f_4(X_2, X_3) + N_4
\end{aligned}
$$

Causal Discovery from Interventions

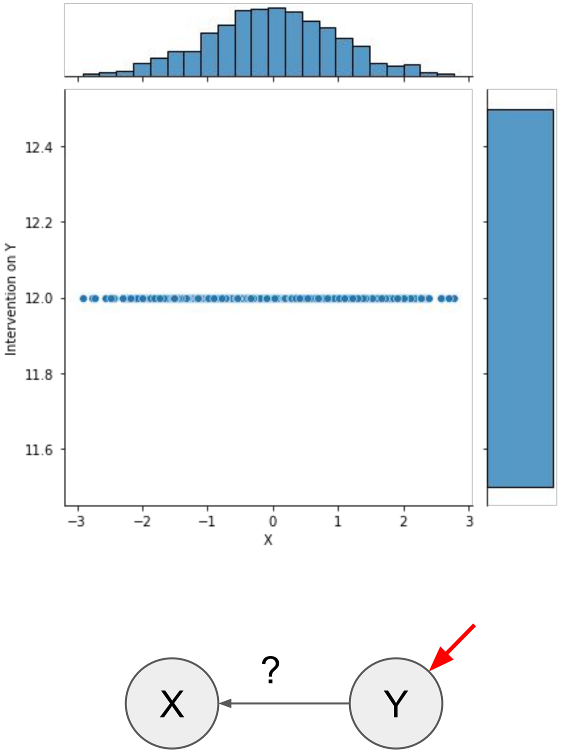
Causal discovery from intervention: two variable case



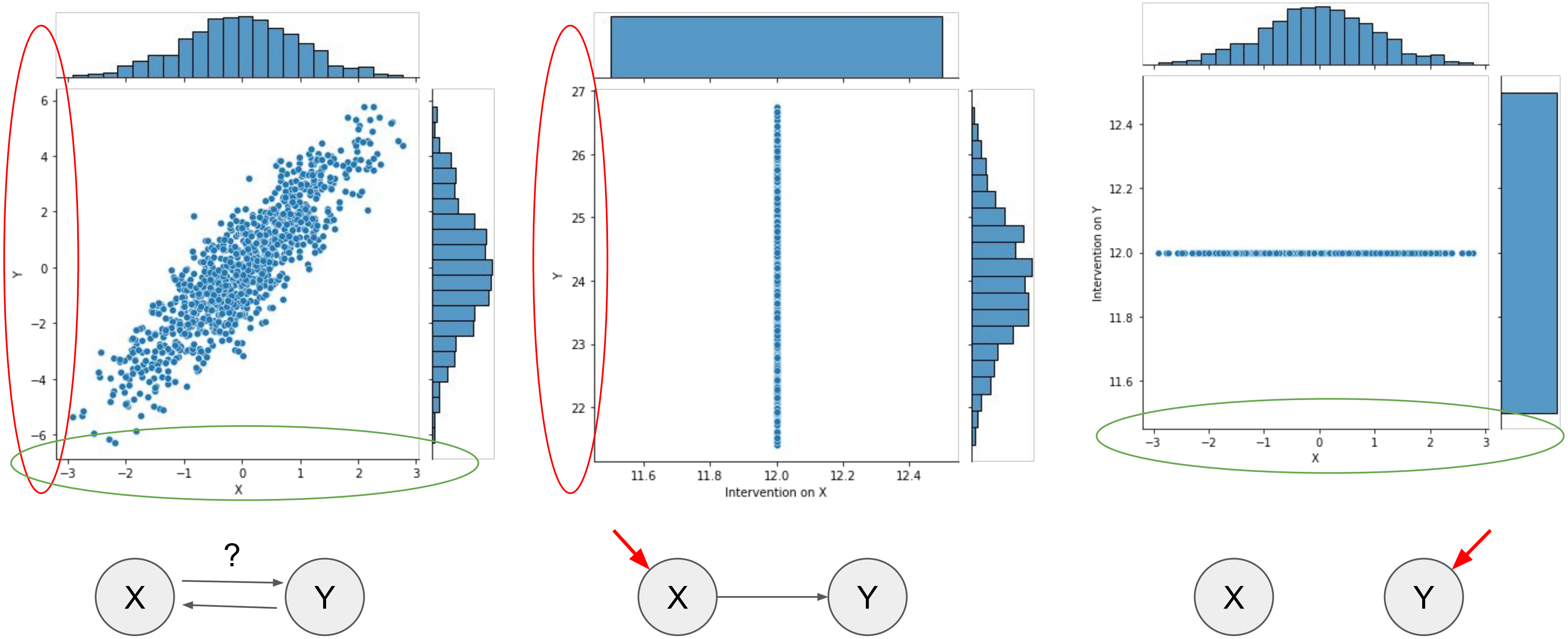
Intervention on X: X = 12



Intervention on Y: Y = 12



Causal discovery from intervention: two variable case



Intervention on X: X = 12

Intervention on Y: Y = 12

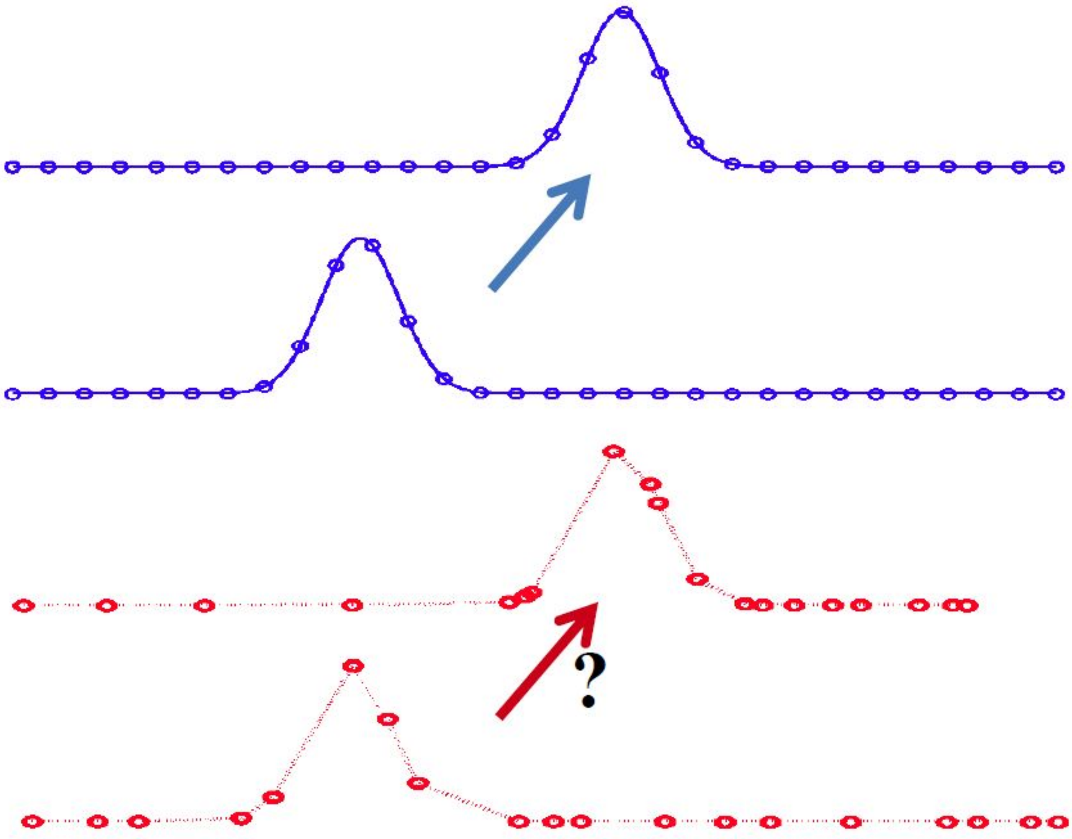
Causal discovery from intervention

- We can alway find underlying causal graph using series of interventions on skeleton:
    - Eberhardt et al. (2005) found that $\lfloor log_2(n) \rfloor + 1$ multi-node interventions are necessary for finding causal graph in worst case scenario (complete graph).
- Intervention can be structural or parametric:
    - E.g. Y=noise or Y=f(PA, N) -> Y=g(PA, N)

Causality in Time series

Estimating the causal generating processes for time series **is not close to solved**

Any of the methods described previously can be used on time series. But their accuracies are sensitive to all of the factors just mentioned.



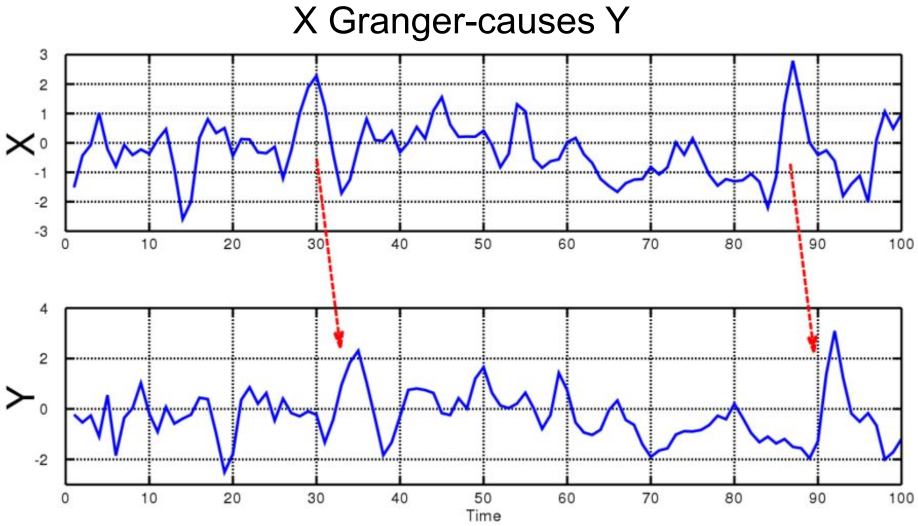
Causality in time series is challenging

Finding the causal dynamics is challenging because:

- the generating process may be non-linear
- the data acquisition rate may be much slower than the underlying rate of changes
- there may be measurement error
- the system may be non-stationary
- there may be unmeasured confounding causes

Granger causality

Does X causes Y or Y causes X?

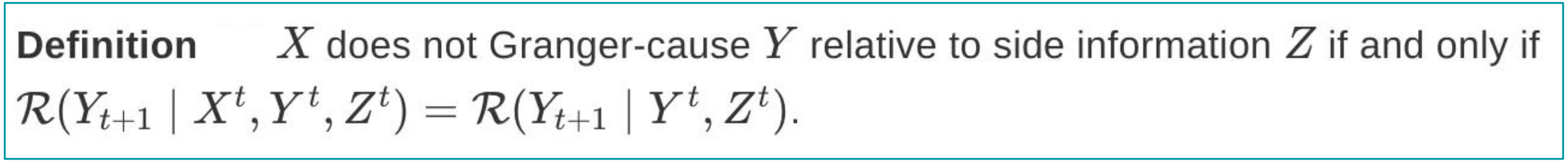

X Granger-causes Y

Granger defined the causality relationship based on two principles:

- The cause happens prior to its effect.
- The cause has unique information about the future values of its effect.

Granger causality

A signal X is said to **Granger-cause** Y if the future realizations of Y can be better explained using the past information from X and Y rather than Y alone.

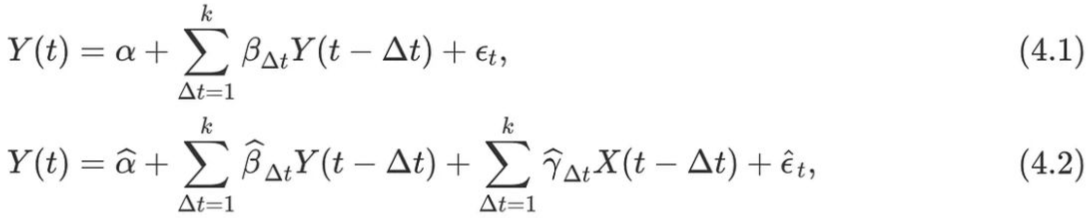


Definition X does not Granger-cause Y relative to side information Z if and only if $\mathcal{R}(Y_{t+1} \mid X^t, Y^t, Z^t) = \mathcal{R}(Y_{t+1} \mid Y^t, Z^t)$.

Standard Granger-causality tests assume a functional form in the relationship among the causes and effects and are implemented by fitting **autoregressive** models.

Granger causality

Consider **the linear vector-autoregressive (VAR) equations** (Wiener 1956; Granger 1969):



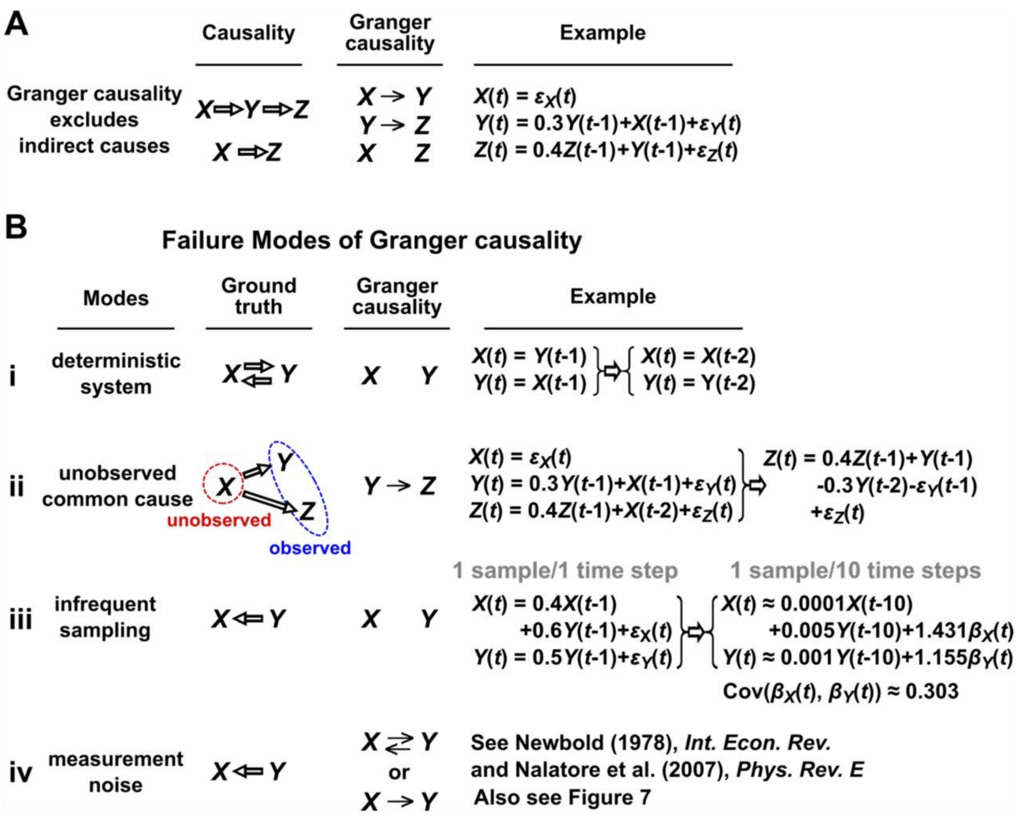
$$Y(t) = \alpha + \sum_{\Delta t=1}^{k} \beta_{\Delta t} Y(t - \Delta t) + \epsilon_t, \tag{4.1}$$

$$Y(t) = \widehat{\alpha} + \sum_{\Delta t=1}^{k} \widehat{\beta}_{\Delta t} Y(t - \Delta t) + \sum_{\Delta t=1}^{k} \widehat{\gamma}_{\Delta t} X(t - \Delta t) + \hat{\epsilon}_t, \tag{4.2}$$

k is the number of lags considered.

X does not G-cause Y if and only if the prediction errors of X in the restricted Eq. (4.1)and unrestricted regression models Eq. (4.2) are equal (i.e., they are statistically indistinguishable).

Challenges of Granger causality



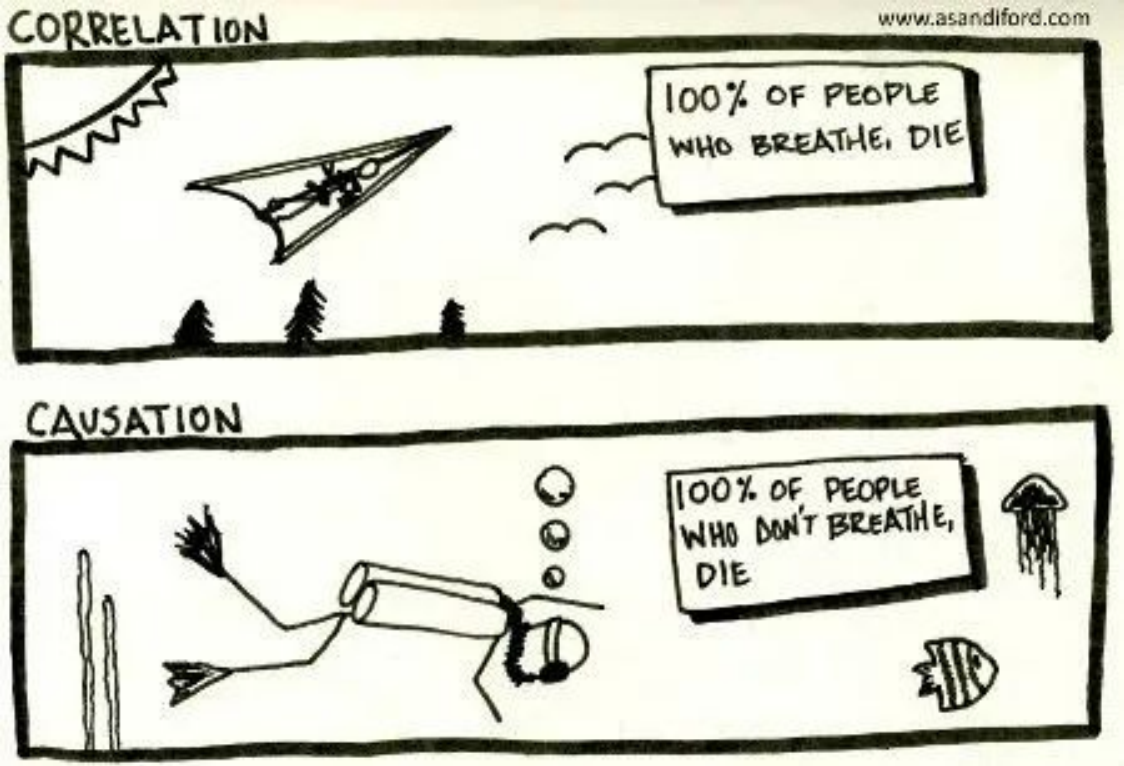
A

	Causality	Granger causality	Example
Granger causality excludes indirect causes	$X \Rightarrow Y \Rightarrow Z$ $X \Rightarrow Z$	$X \to Y$ $Y \to Z$ $X \quad Z$	$X(t) = \varepsilon_X(t)$ $Y(t) = 0.3Y(t\text{-}1)+X(t\text{-}1)+\varepsilon_Y(t)$ $Z(t) = 0.4Z(t\text{-}1)+Y(t\text{-}1)+\varepsilon_Z(t)$

B

Failure Modes of Granger causality

	Modes	Ground truth	Granger causality	Example
i	deterministic system	$X \rightleftarrows Y$	$X \quad Y$	$\left.\begin{array}{l} X(t) = Y(t\text{-}1) \\ Y(t) = X(t\text{-}1) \end{array}\right\} \Rightarrow \left\{\begin{array}{l} X(t) = X(t\text{-}2) \\ Y(t) = Y(t\text{-}2) \end{array}\right.$
ii	unobserved common cause	$X \Rightarrow Y$ $X \Rightarrow Z$ unobserved observed	$Y \to Z$	$\left.\begin{array}{l} X(t) = \varepsilon_X(t) \\ Y(t) = 0.3Y(t\text{-}1)+X(t\text{-}1)+\varepsilon_Y(t) \\ Z(t) = 0.4Z(t\text{-}1)+X(t\text{-}2)+\varepsilon_Z(t) \end{array}\right\} \Rightarrow \begin{array}{l} Z(t) = 0.4Z(t\text{-}1)+Y(t\text{-}1) \\ \quad -0.3Y(t\text{-}2)-\varepsilon_Y(t\text{-}1) \\ \quad +\varepsilon_Z(t) \end{array}$
				1 sample/1 time step → 1 sample/10 time steps
iii	infrequent sampling	$X \Leftarrow Y$	$X \quad Y$	$\left.\begin{array}{l} X(t) = 0.4X(t\text{-}1) \\ \quad +0.6Y(t\text{-}1)+\varepsilon_X(t) \\ Y(t) = 0.5Y(t\text{-}1)+\varepsilon_Y(t) \end{array}\right\} \Rightarrow \left\{\begin{array}{l} X(t) \approx 0.0001X(t\text{-}10) \\ \quad +0.005Y(t\text{-}10)+1.431\beta_X(t) \\ Y(t) \approx 0.001Y(t\text{-}10)+1.155\beta_Y(t) \end{array}\right.$ $\mathrm{Cov}(\beta_X(t), \beta_Y(t)) \approx 0.303$
iv	measurement noise	$X \Leftarrow Y$	$X \rightleftarrows Y$ or $X \to Y$	See Newbold (1978), *Int. Econ. Rev.* and Nalatore et al. (2007), *Phys. Rev. E* Also see Figure 7


CORRELATION
www.asandiford.com
100% OF PEOPLE WHO BREATHE, DIE
CAUSATION
100% OF PEOPLE WHO DON'T BREATHE, DIE

What is the connection between causality and ML?

Sepideh Pashami - 20230508

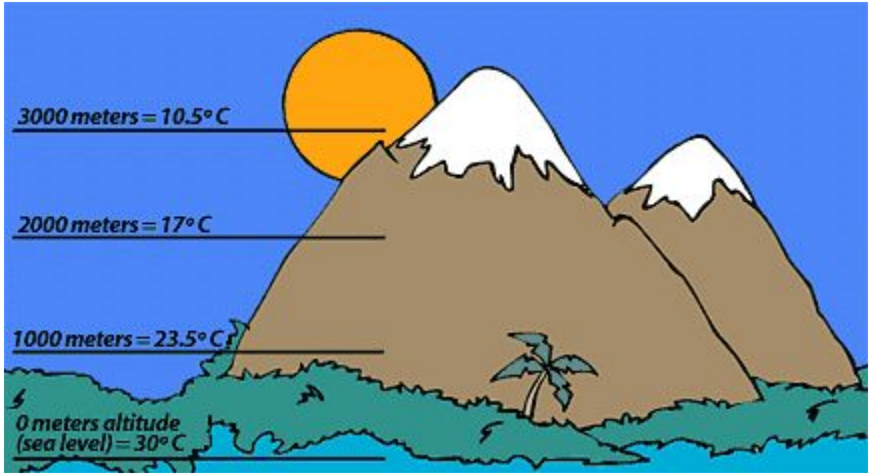
Imagine we have a dataset with the **altitude** and **the average annual temperature** from different cities in a country. So we have the joint probability p(a,t).

T→A? $p(a,t) = p(a|t)p(t)$? Or,

A → T? $p(a,t) = p(t|a)p(a)$?

Intervention 1: elevate all the cities in our dataset. What is the effect on the average annual temperature?

Intervention 2: change the city's temperature using a giant air conditioner. What is the effect on altitude?

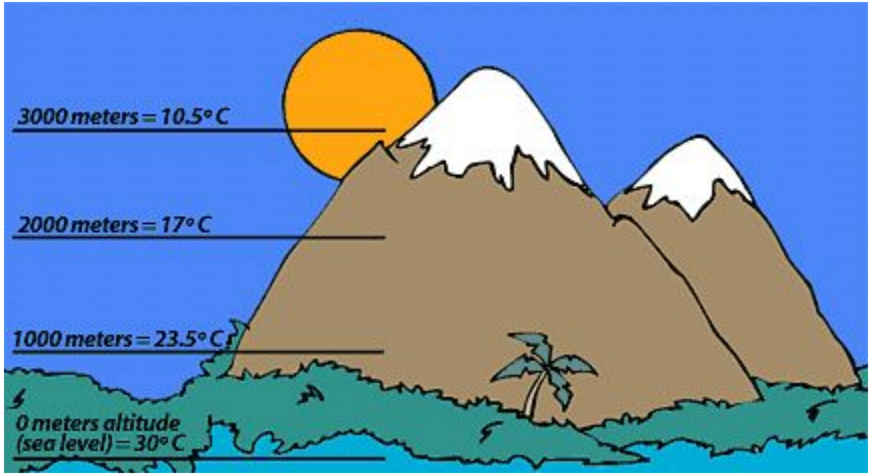


Peters, et al. Elements of causal inference: foundations and learning algorithms (2017)

- Instead of thinking about this data coming from a single country, let’s now imagine we have different datasets coming from two countries. E.g. Brazil and Germany
- Shouldn’t the relationship between altitude and temperature be equal no matter where we measure it?

This is an **invariance** under different countries

In another word, if we get causal inference correctly, we can reuse the same **relationship** learned for Brazil to Germany.



Independent Causal Mechanisms (ICM) Principle

The causal generative process of a system’s variables is composed of autonomous modules that do not inform or influence each other.

- In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

E.g. In our altitude and temperature example: $p(t|a)$ and $p(a)$ are “independent”.

Causality for Machine Learning, https://arxiv.org/pdf/1911.10500.pdf

Independent Causal Mechanisms (ICM) Principle

Applied to the causal factorization, the principle tells us that the factors should be "**independent**" in the sense that:

- Changing (or intervening upon) one mechanism $p(X_i|PA_i)$ does not change the other mechanisms $p(X_j|PA_j)(i \neq j)$
- Knowing some other mechanisms $p(X_i|PA_i)$ $(i \neq j)$ does not give us information about a mechanism $p(X_j|PA_j)$

Independent Causal Mechanisms (ICM) Principle

Consider a Markov factorization with respect to causal DAG:

$$p(x_1, \ldots, X_d) = \prod_{i=1}^{d} p(x_i | x_{pa(i)})$$

Modularity suggests:

$p(x_1 | x_{pa(1)}), \ldots, p(x_d | x_{pa(d)})$ are independent.

Independent Causal Mechanisms (ICM) Principle

This principle subsumes several notions important to causality, including **separate intervenability of causal variables**, **modularity** and **autonomy of subsystems**, **entanglement**, and **invariance**.

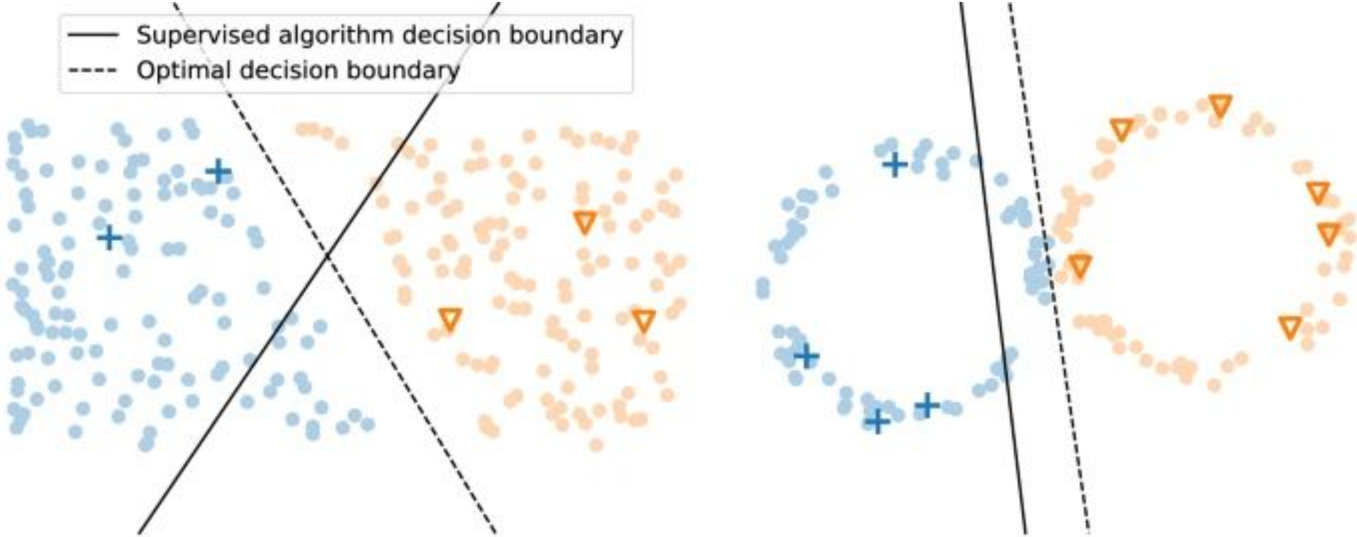
Note:

The dependence of two mechanisms $p(X_i|PA_i)$and $p(X_j|PA_j)$ does not coincide with the statistical dependence of the random variables X_i and X_j. Indeed, in a causal graph, many of the random variables will be dependent even if all the mechanisms are independent.

Semi-supervised learning

Small portion of data is labeled + lots of unlabeled data.

- We need some information in p(x) that improves p(y|x).


Supervised algorithm decision boundary
Optimal decision boundary

(a) Smoothness and low-density assumptions. **(b)** Manifold assumption.

Semi-supervised learning

According to Modularity assumption:

$p(x_1|x_{pa(1)}), \ldots, p(x_d|x_{pa(d)})$ are independent.

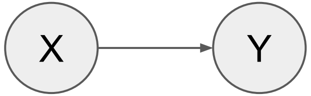
Special case for two variables:

- p(cause), p(effect|cause) are independent.
- p(effect), p(cause|effect) are not independent.

Semi-supervised learning

Semi-supervised learning from cause to effect does not work!

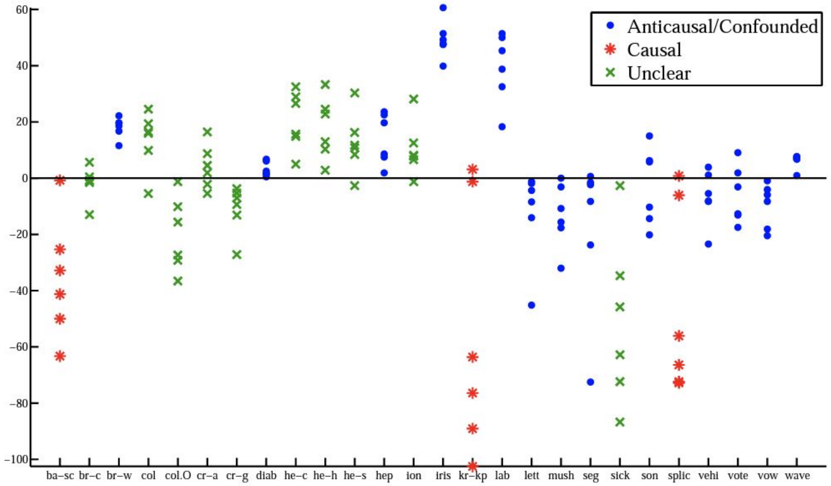
p(x,y) = p(x)p(y|x) or p(cause), p(effect|cause)



The ICM Principle posits that the modules in a joint distribution causal decomposition do not inform or influence each other. This means that in particular, p(x) (unlabelled data) should contain no additional information about p(y|x).

Semi-supervised learning

What about the opposite direction? Does semi-supervised learning work when we are predicting cause from the effect (anti-causal direction)?



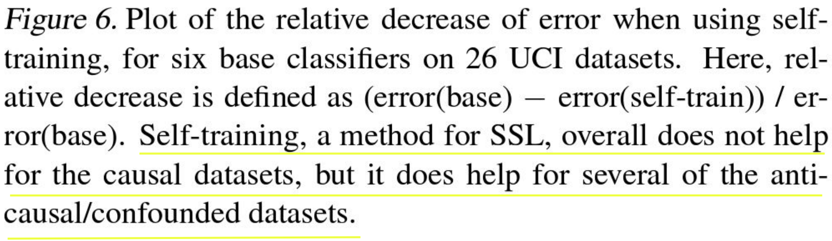
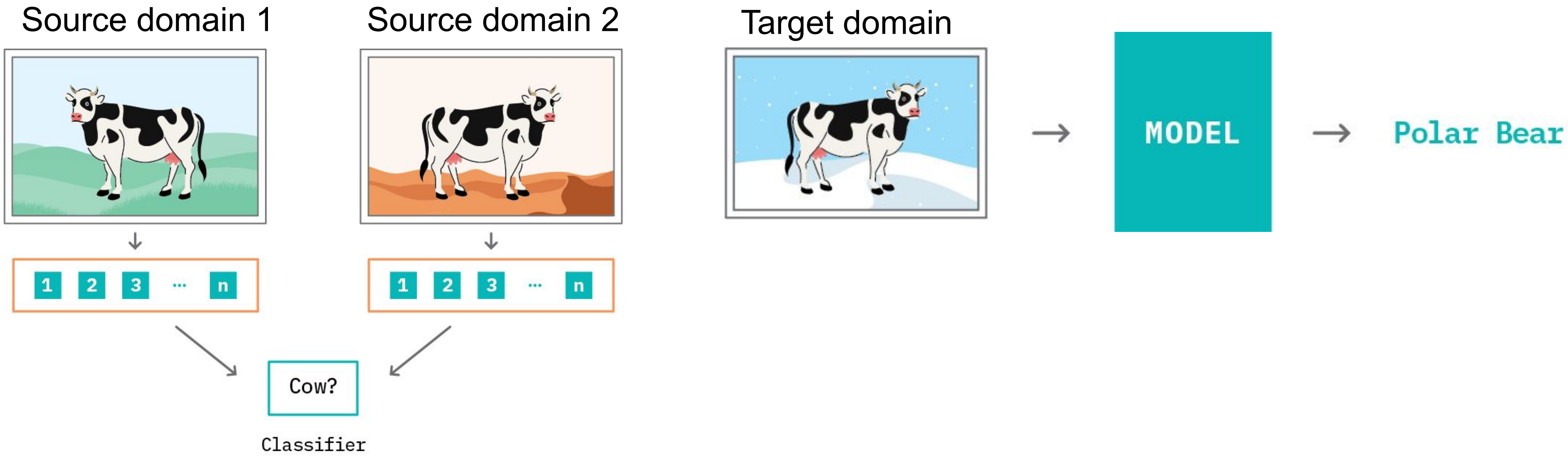


Figure 6. Plot of the relative decrease of error when using self-training, for six base classifiers on 26 UCI datasets. Here, relative decrease is defined as (error(base) − error(self-train)) / error(base). Self-training, a method for SSL, overall does not help for the causal datasets, but it does help for several of the anti-causal/confounded datasets.

On Causal and Anti-causal Learning, https://icml.cc/2012/papers/625.pdf

Domain adaptation

We are looking for invariant predictor trained on source domains for classifying animals in the target domain:

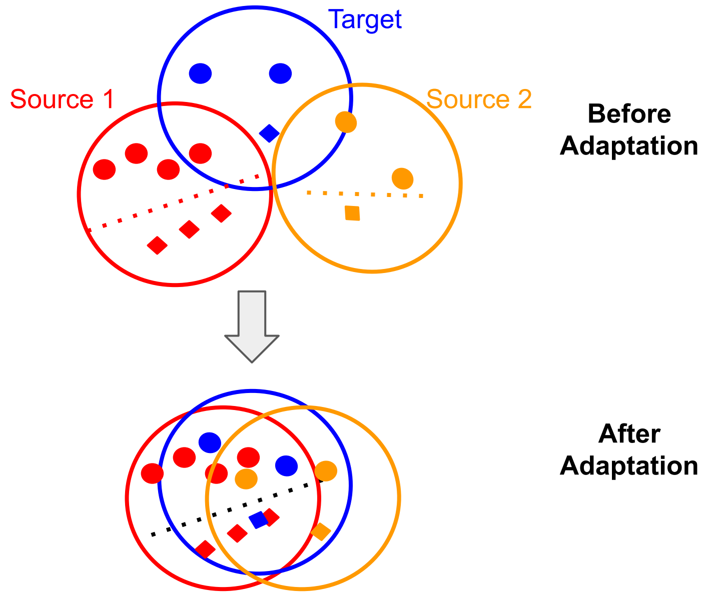


Domain adaptation

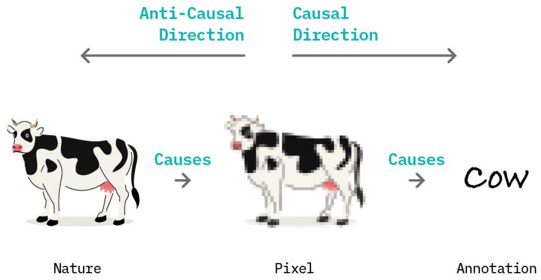
Domain adaptation is the ability to apply an algorithm trained in one or more "source domains" to a different (but related) "target domain".

- Often there is not enough data in target domain to train from scratch.

A **domain shift** is a change in the data distribution between source and target datasets.



Domain adaptation



Causal direction:

So when things like domain shift happen, it becomes just a matter of a different input distribution to our invariant mechanism. Independent causal mechanism can stay invariant under different conditions.

Anti-causal direction:

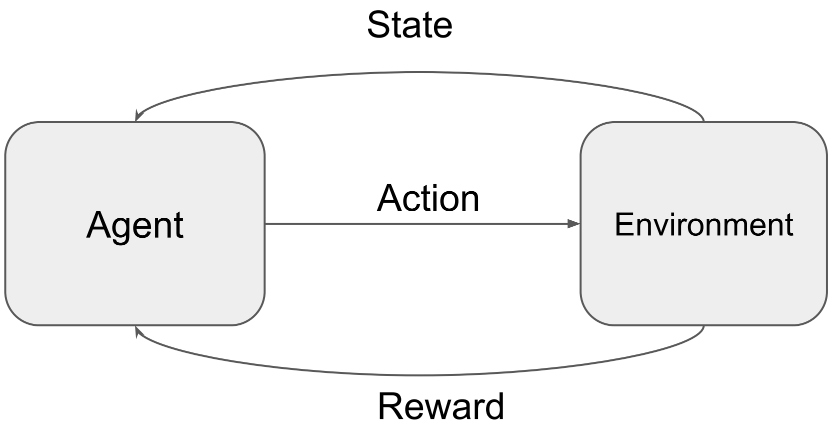
let’s say P(Cause|Effect), then a domain shift (P(Effect) change) is going to also change the learnt mechanism.

Reinforcement Learning

The machine is given feedback concerning the decision it makes, but no information about possible alternatives

Reinforcement learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states -> actions that tells you what to do in a given state



- Receive feedback in the form of rewards.
- Agent's utility is defined by the reward function.
- Must (learn to) act so as to maximize expected rewards.

Reinforcement learning

Kidney stone example:

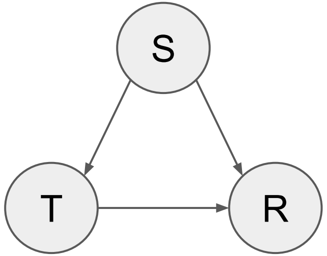
Given the condition of kidney stone (S), a treatment (T) will be selected in a way that it maximises the probability of recovery (R).

State

Action

Maximising reward

What would happen if ….? We want to intervene in the treatment.



$p(t,r,s) = p(s)\, p(t|s)\, p(r|s,t)$

Intervene on treatment $p^*(t|s)$

Take home messages!

- Knowledge can be decomposed in informationally **independent** pieces (**mechanisms**, modules)
- **Semi-supervised learning** from cause to effect does not work!
- Invariant models for **domain adaptation, multi-task learning, transfer learning** can be readily achieved when the causal conditional probability has been learned.
- **Reinforcement learning** is closer to causality research than the machine learning mainstream in that it sometimes effectively directly estimates do-probabilities.

Can Causality solve open problems of ML?

1

Can we answer counterfactual questions based on observations only?

?

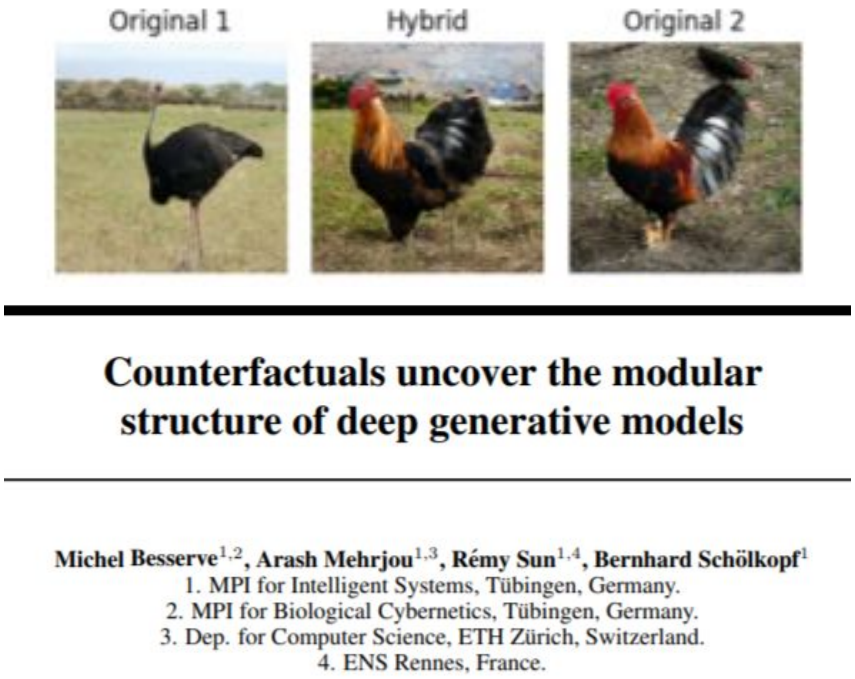
Many meaning full image properties are correlated.

- Elephant often has a green background.
- what if the background of an elephant was a city?

Answering counterfactual questions

- Deep generative models have proven successful at designing realistic images

- Providing a disentangle latent representation of the data using Generative models


Original 1
Hybrid
Original 2

Counterfactuals uncover the modular structure of deep generative models

Michel Besserve[1,2], **Arash Mehrjou**[1,3], **Rémy Sun**[1,4], **Bernhard Schölkopf**[1]
1. MPI for Intelligent Systems, Tübingen, Germany.
2. MPI for Biological Cybernetics, Tübingen, Germany.
3. Dep. for Computer Science, ETH Zürich, Switzerland.
4. ENS Rennes, France.

https://arxiv.org/pdf/1812.03253.pdf

Answering counterfactual questions

- Independent causal mechanism
  - Uncover a modular structure by manipulating latent representation in a way that changing one module doesn't change other modules
  - E.g. change smile of a person

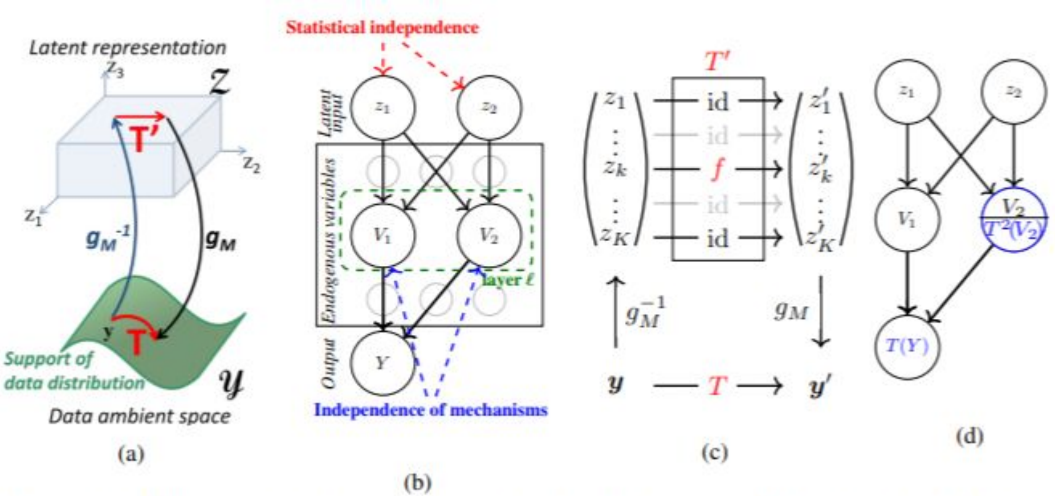


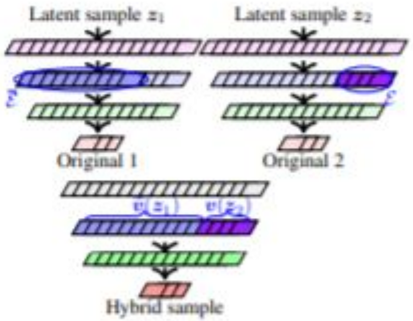
Figure 2: (a) Illustration of the generative mapping and a disentangled transformation. (b) Causal graph of an example CGM showing different types of independence between nodes. (c) Commutative diagram showing sparse transformation T' in latent space associated to a disentangled transformation T. (d) Illustration of intrinsic disentanglement with $\mathcal{E} = \{2\}$.

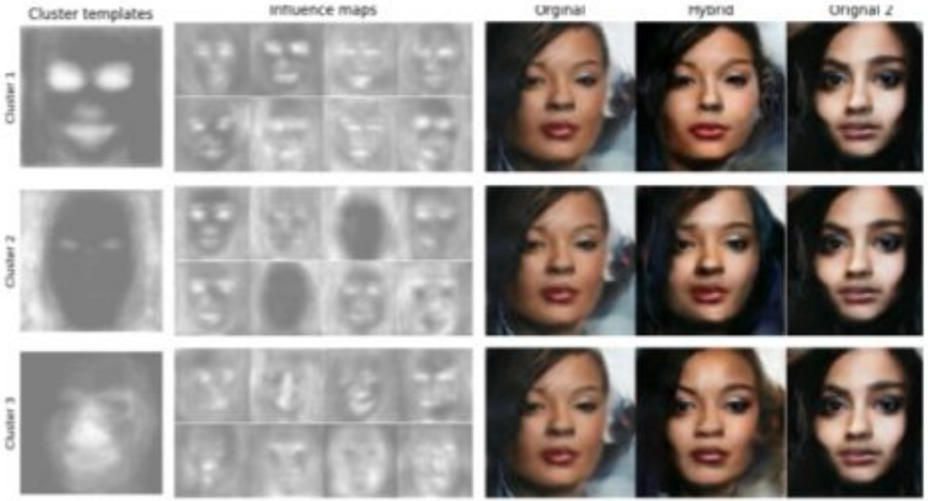
https://arxiv.org/pdf/1812.03253.pdf

Answering counterfactual questions

Generated image using causal disentanglement.

- Completely Unsupervised





https://arxiv.org/pdf/1812.03253.pdf

2

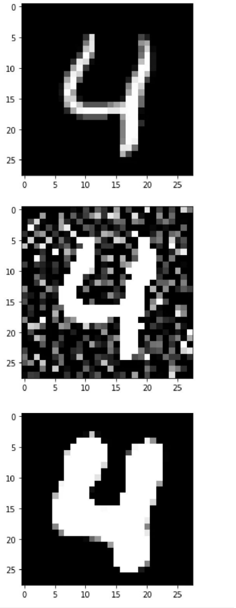
Can we learn independent causal mechanism automatically?

Human intelligence

Humans are able to recognize objects such as handwritten digits based on distorted inputs.

They can correctly label translated, corrupted, or inverted digits, without having to relearn them from scratch.

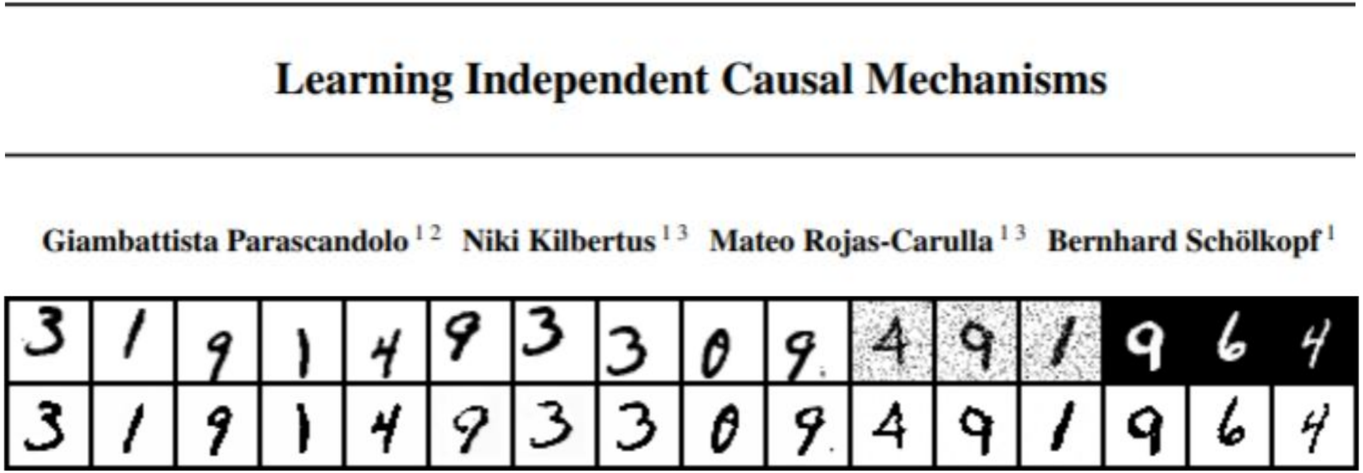
The same applies for new objects, essentially after having seen them once.



Automatic data-driven algorithms

Unsupervised transformation of digits by learning independent causal mechanism

The approach is based on a set of experts that compete for data generated by the mechanisms.



Learning Independent Causal Mechanisms

Giambattista Parascandolo[1 2] **Niki Kilbertus**[1 3] **Mateo Rojas-Carulla**[1 3] **Bernhard Schölkopf**[1]

3 1 9 1 4 9 3 3 0 9 4 9 1 9 6 4
3 1 9 1 4 9 3 3 0 9 4 9 1 9 6 4

https://arxiv.org/pdf/1712.00961.pdf

Automatic data-driven algorithms

The architecture using competing experts that automatically specialize on different image transformations

- Each example is fed to all experts independently and in parallel.
- Comparing the outputs of all experts and selecting the winning expert
- Weights of winning expert is updated and other experts stay unchanged. (The motivation behind competitively updating only the winning expert is to enforce specialization)

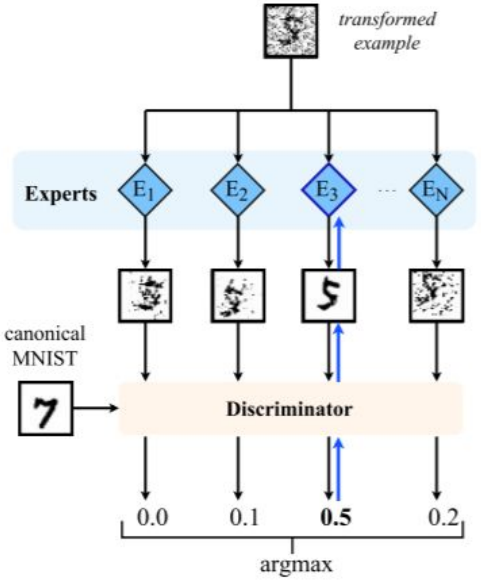


Figure 2. We show how a transformed example, here a noisy digit, is processed by a competition of experts. Only Expert 3 is specializing on denoising, it wins the example and gets trained on it, whereas the others perform translations and are not updated.

https://arxiv.org/pdf/1712.00961.pdf

3

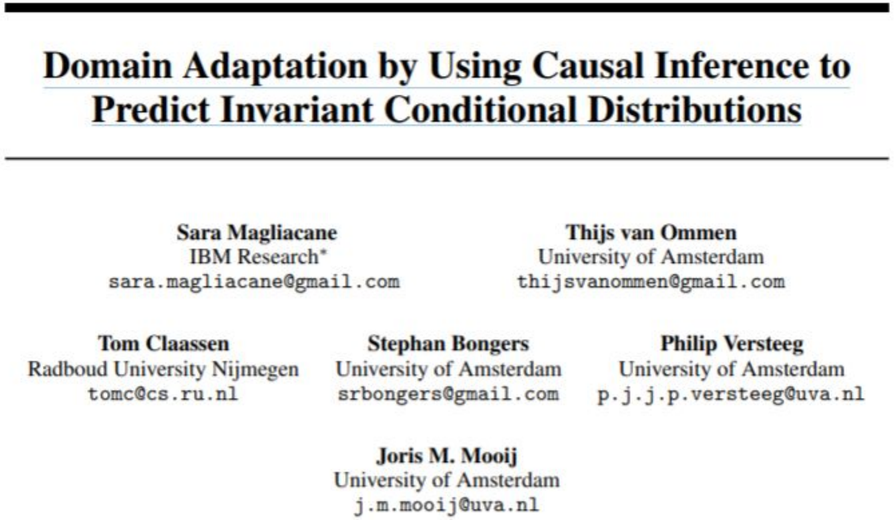
Can we perform domain adaptation using causal relation?

Improving domain adaptation

Standard feature selection methods rely only on predictive power

Selecting invariant features for source and target domains

Domain Invariant features found leveraging causal information



Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

Sara Magliacane
IBM Research*
sara.magliacane@gmail.com

Thijs van Ommen
University of Amsterdam
thijsvanommen@gmail.com

Tom Claassen
Radboud University Nijmegen
tomc@cs.ru.nl

Stephan Bongers
University of Amsterdam
srbongers@gmail.com

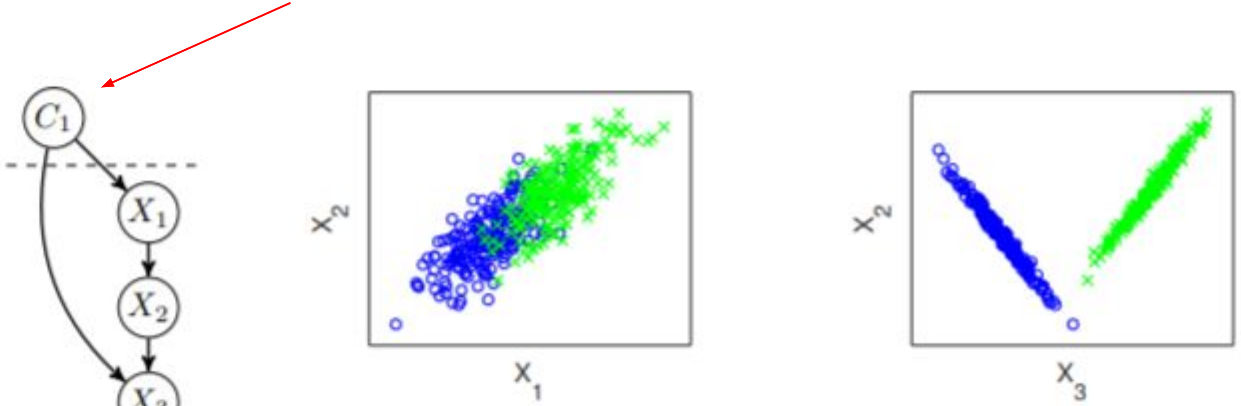
Philip Versteeg
University of Amsterdam
p.j.j.p.versteeg@uva.nl

Joris M. Mooij
University of Amsterdam
j.m.mooij@uva.nl

https://arxiv.org/pdf/1707.06422.pdf

Improving domain adaptation

Intervention causing distribution shift



(a) Causal graph

(b) No distribution shift for $\{X_1\}$: $\mathbb{P}(Y \mid X_1, C_1 = 0) = \mathbb{P}(Y \mid X_1, C_1 = 1)$

(c) Strong distribution shift for $\{X_3\}$: $\mathbb{P}(Y \mid X_3, C_1 = 0) \neq \mathbb{P}(Y \mid X_3, C_1 = 1)$

Predict Y from only features that make
Y and C1 independent

$$C_1 \perp Y \mid \boldsymbol{A}\ [\mathcal{G}]$$

https://arxiv.org/pdf/1707.06422.pdf

4

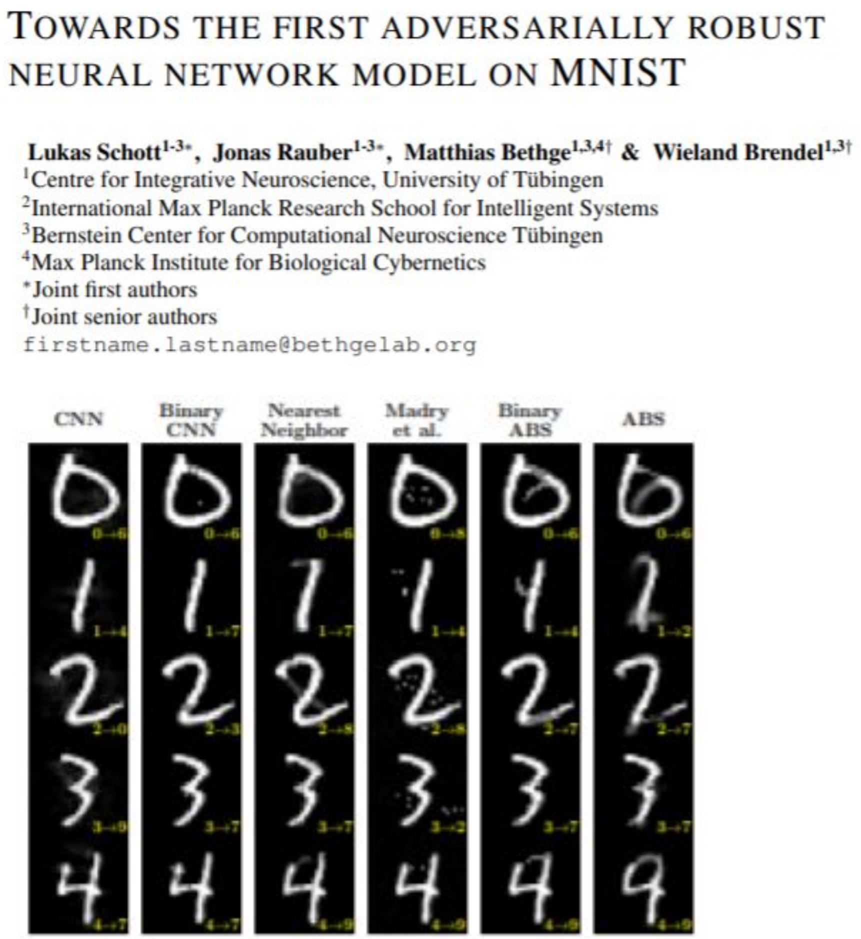
Can we increase robustness and security of Machine Learning algorithms?

Increasing robustness & security

Deep neural networks (DNNs) are susceptible to minimal adversarial perturbations

Using causality for creating adversarially robust NNs

https://arxiv.org/pdf/1805.09190.pdf



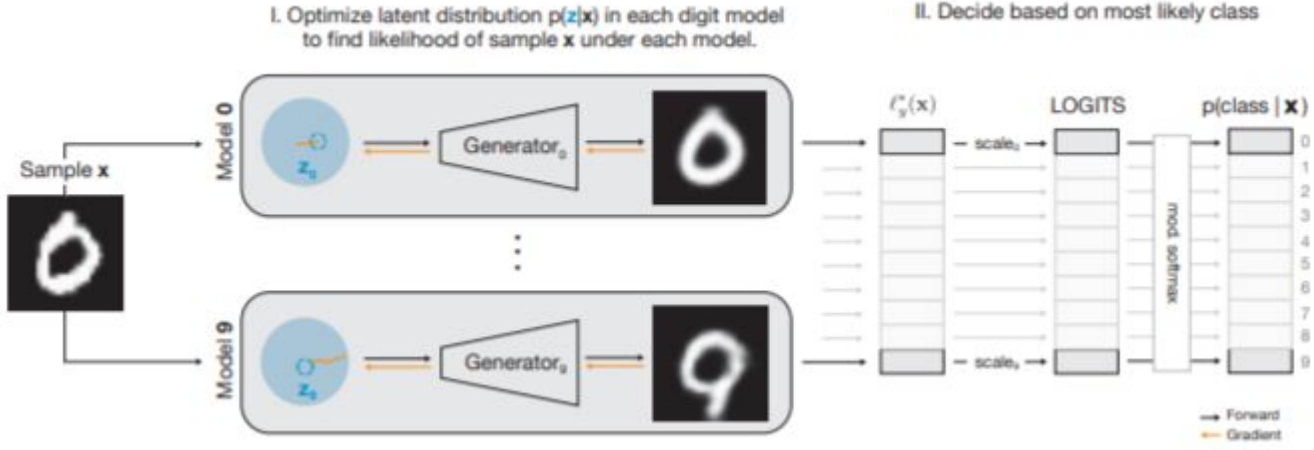
TOWARDS THE FIRST ADVERSARIALLY ROBUST NEURAL NETWORK MODEL ON MNIST

Lukas Schott[1-3*], **Jonas Rauber**[1-3*], **Matthias Bethge**[1,3,4†] & **Wieland Brendel**[1,3†]
[1]Centre for Integrative Neuroscience, University of Tübingen
[2]International Max Planck Research School for Intelligent Systems
[3]Bernstein Center for Computational Neuroscience Tübingen
[4]Max Planck Institute for Biological Cybernetics
[*]Joint first authors
[†]Joint senior authors
`firstname.lastname@bethgelab.org`

Increasing robustness & security

Machine Learning can benefit from causal and anticausal knowing structure / prediction tasks.

- Using Bayes’ rule to solve causal problem rather than anticausal.



https://arxiv.org/pdf/1805.09190.pdf