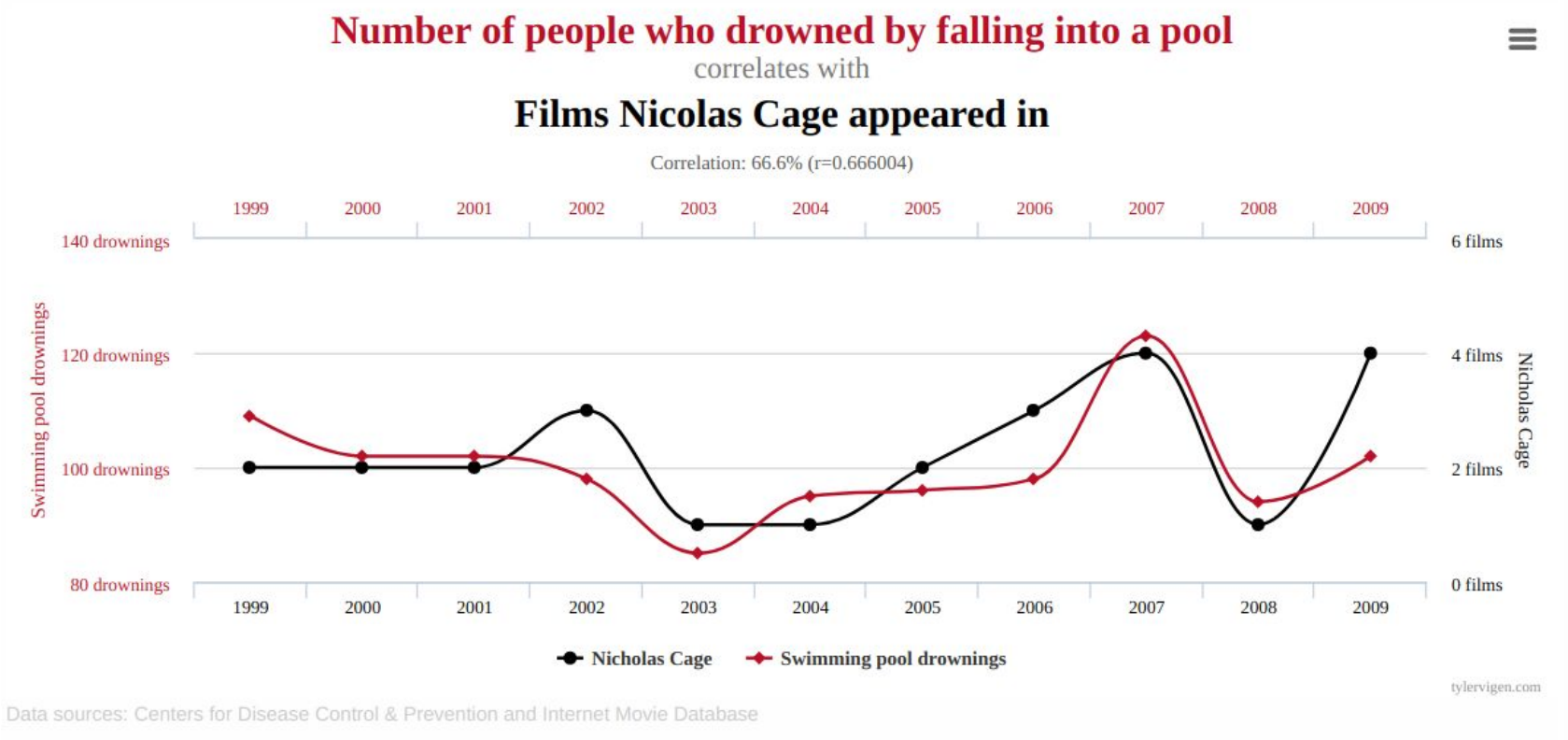
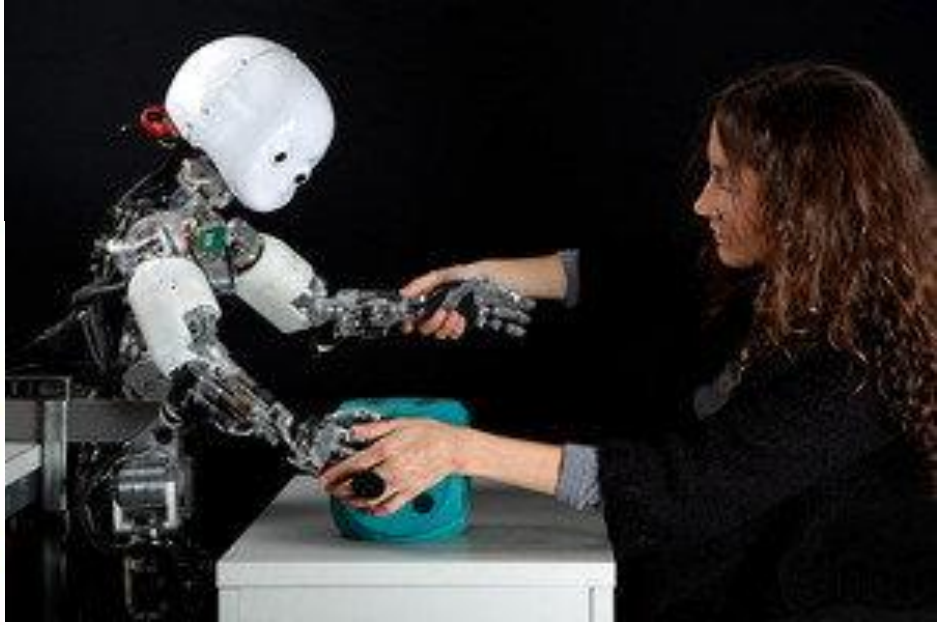
Causal inference in machine learning

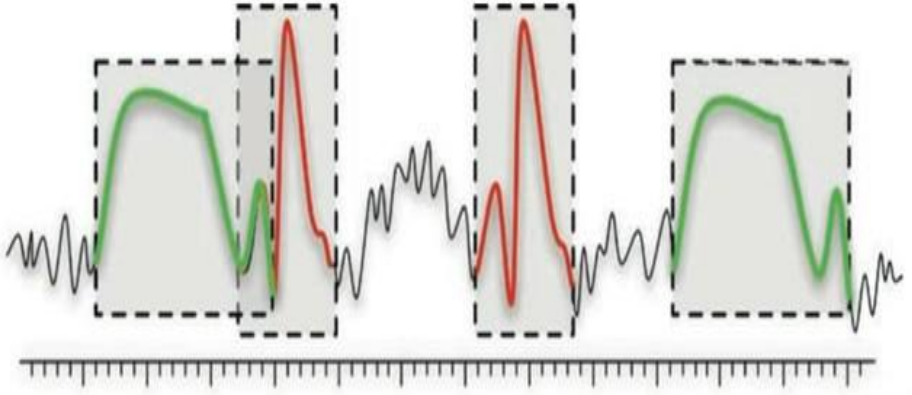
Sepideh Pashami

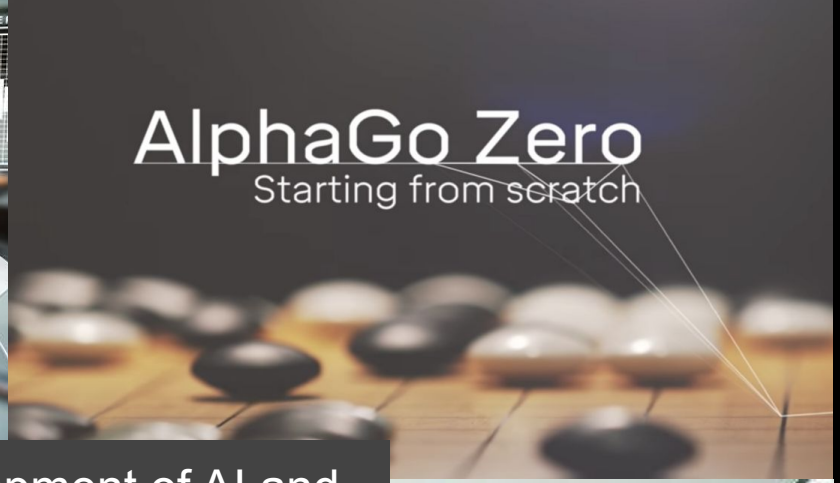
Correlation is not causation!



Machine Learning is good at ….







Rapid development of AI and autonomous systems toward human intelligence

Human Intelligence

"Humans have the ability to

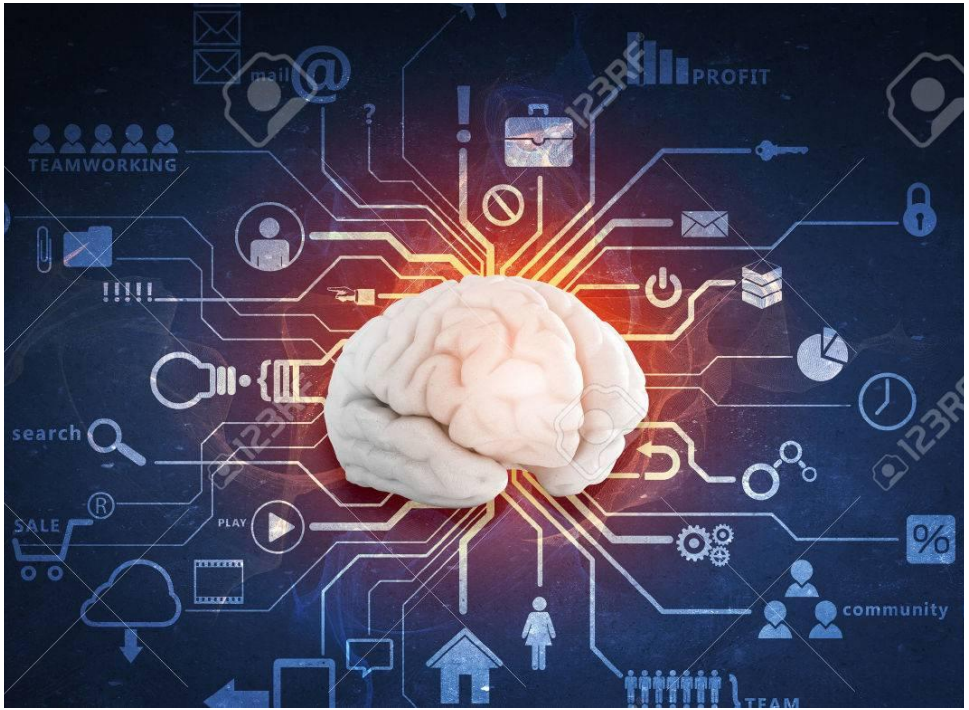
(1) choreograph a mental representation of their environment,

(2) interrogate that representation,

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

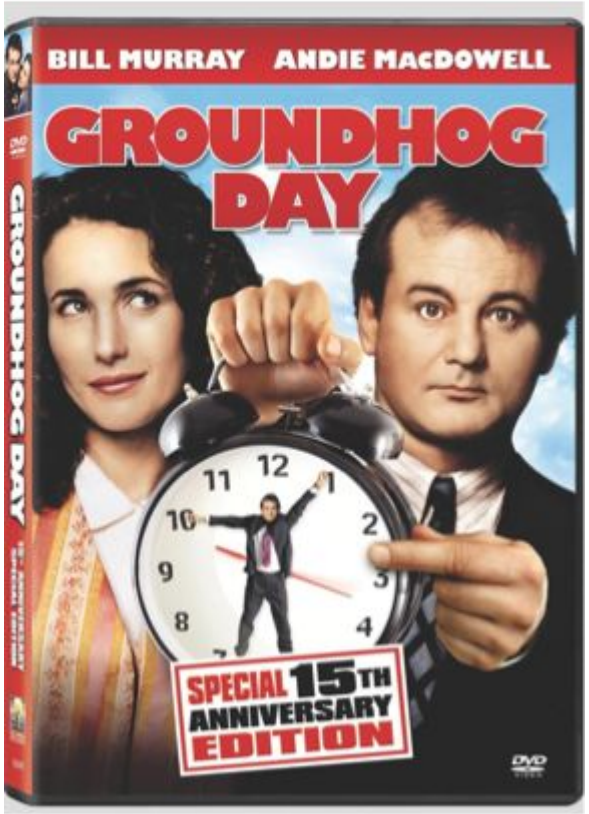
(4) finally answer *'What if?'* kind of questions."

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?



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?

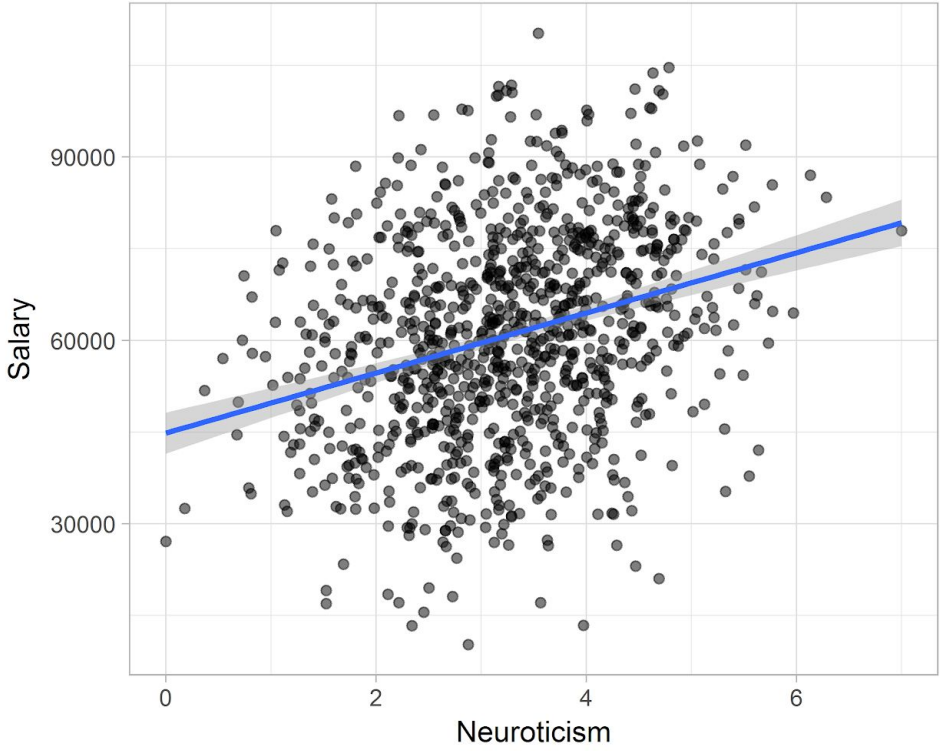
I am about to quit my job, **should I**?

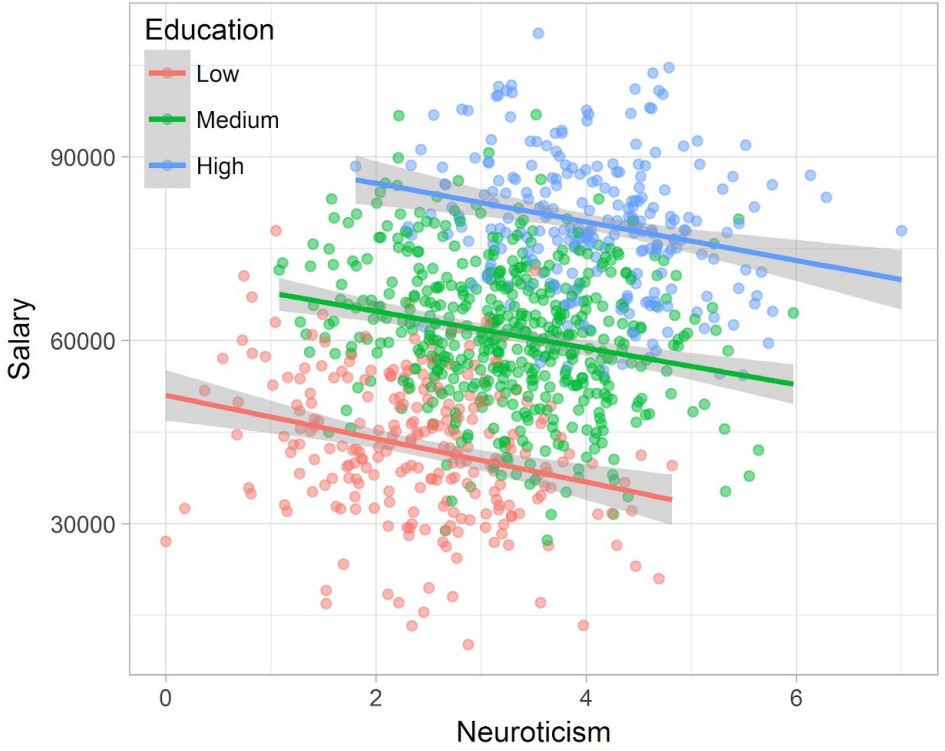
Pearl & Mackenzie. The book of why. 2019

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?

Simpson's paradox





Applications of causal inference

- **Law**: Counterfactual reasoning for increasing transparency of automated solutions
- Data-driven **policy making**: measuring the effects of interventions, rather than looking for mere correlation
- **Medical** decision making: Distinguishing causal effects of treatment from results.
- **Epidemiological** studies: an exercise in measurement of an effect rather than as a criterion-guided process for deciding whether an effect is present or not.

How can we discover causal relations?

- Correlation:
  - It is raining -> people probably carry open umbrellas
  - People carry open umbrellas -> It is probably raining
- Intervention:
  - Will it rain if we ban umbrella?
  - Would it have rained if we had banned umbrellas?
- 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

Not enough

Too difficult

Sometimes impractical

Causal Inference Based on Observations

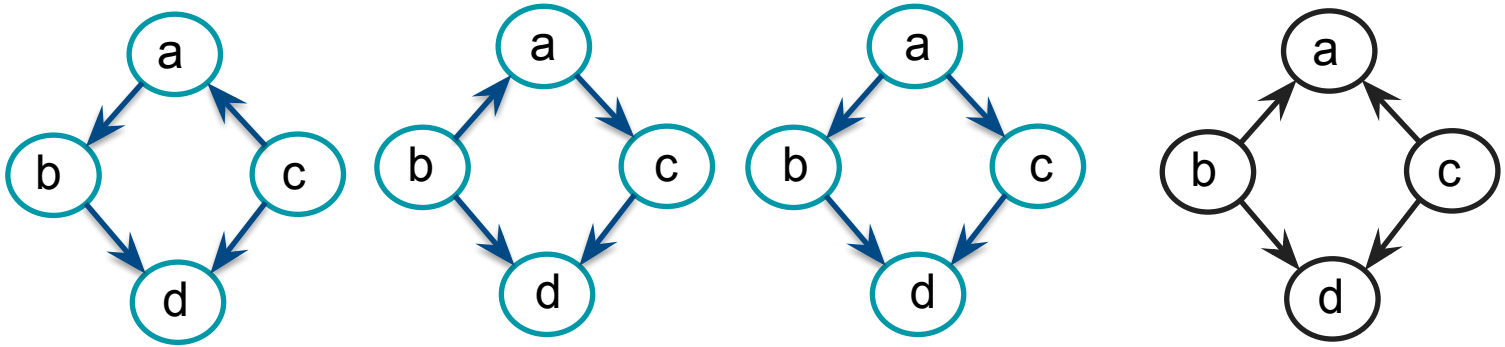
Counterfactual reasoning using graphical representation

Is it possible?

Can we infer causal links from purely observational data?

- NO!
- Assuming faithfulness (and conditional independence tests), can **estimate a Markov equivalence class** containing the true causal graph. [Pearl, 2000]

Markov equivalent classes

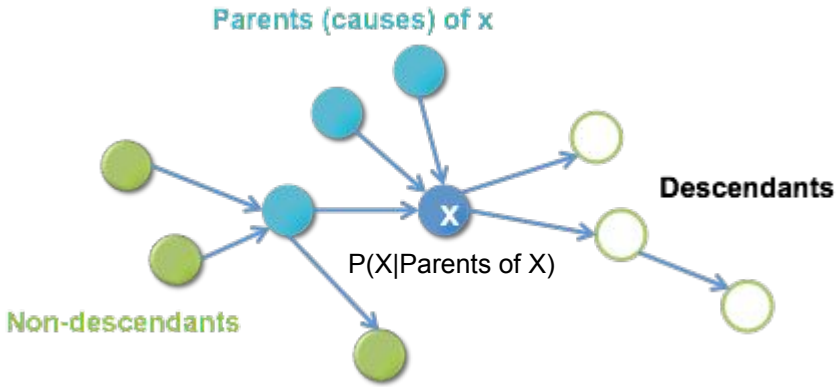


Causal Model (Pearl et al.)

- 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)

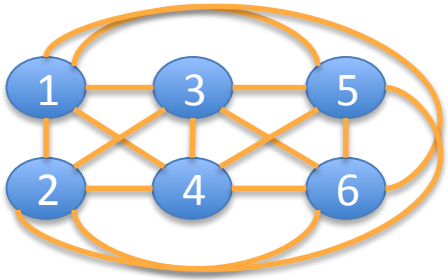
- Implies p(X1, . . . ,Xn) with particular conditional independence structure:
  - Causal Markov condition:

    X independent of non-descendants, given parents

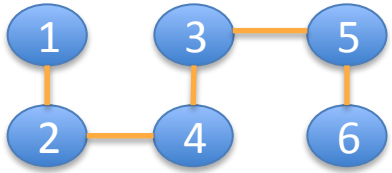


Causal graph from observational data

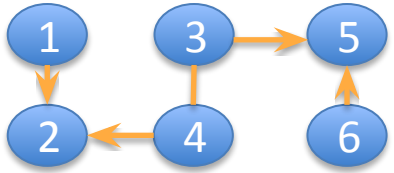
PC algorithm: conditional independence based algorithm



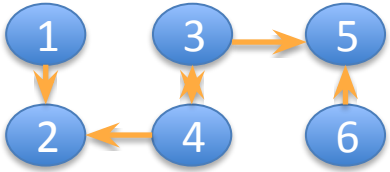
Initialize with a fully connected un-oriented graph



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

Complications

- Needs large amounts of data
- Needs all relevant variables to be known
- No feedback loops
- No symmetries that make correlations cancel out

Complications

- Needs large amounts of data
- Needs all relevant variables to be known
- No feedback loops
- No symmetries that make correlations cancel out

Approach

- Use an invariant measure of correlation – needs less data
- Bayesian analysis of correlation – incremental and measure of uncertainty
- Time series analysis with Markov chains – to unroll loops
- Higher order correlations – to break symmetries

An Invariant Conditional Independence Test

- Most methods for causal discovery requires relatively **large amounts of data**
- A constraint based method **condition on a quite large set** of other variables
  - This splits up the data set in many small parts
- This means that each conditional independence test is performed on a fraction of the available data, leading to low significance
- the estimates from many smaller sets are pooled, risking to miss individual results that are significant.

Extension of Odds ratio

$$Q_{XY} = \frac{p_{11}p_{00}}{p_{01}p_{10}}$$

Odds ratio

$$Q_{XYZ} = \frac{p_{111}p_{001}p_{010}p_{100}}{p_{011}p_{101}p_{110}p_{000}}$$

Third order odds ratio

$$Q_{XY\ldots Z} = \frac{\prod_{x,y,\ldots,z:\text{even \#0}} p_{xy\ldots z}}{\prod_{x,y,\ldots,z:\text{odd \#0}} p_{xy\ldots z}}$$

Interaction between multivariate variables are defined using multivariate extension of odds-ratio

The invariant interaction test

A measure which invariant to the values of conditioning variables

Using a Bayesian approach, we estimate the distribution of S from the observational data and then look at the mass in the tail beyond the H0

$$S_{XY\ldots Z} = 2^k \delta(Q_{XY\ldots Z})$$

$$S = \frac{\sqrt{Q_{XY}} - 1}{\sqrt{Q_{XY}} + 1}$$

The k-order correlation measure is obtained by

Bayesian distribution of correlation

$$P(S \mid \mathbf{D}_1, \ldots, \mathbf{D}_K) = \int_{p_x^{(1)}, p_y^{(1)}, \ldots, p_x^{(K)}, p_y^{(K)}} P(S, p_x^{(1)}, p_y^{(1)}, \ldots, p_x^{(K)}, p_y^{(K)} \mid \mathbf{D}_1, \ldots, \mathbf{D}_K)$$

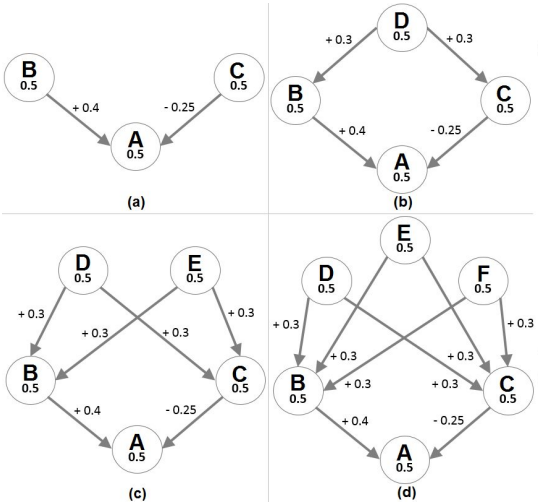
Experiment results

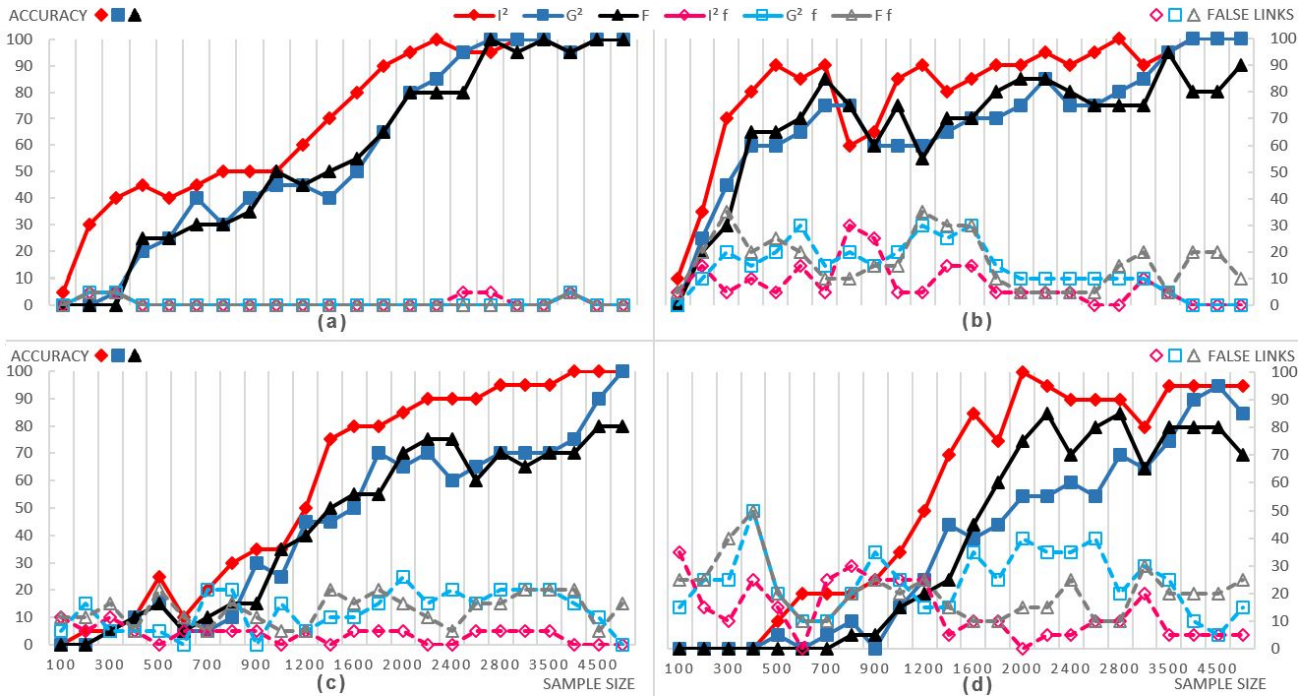
The proposed CI test (I^2) is invariant to the amount of data

	% of data	I^2	G^2	F
1.	100%	0.008064	0.01140	0.01247
2.	50% × 2	0.008071	0.04073	0.08111
3.	25% × 4	0.008086	0.17109	0.22948

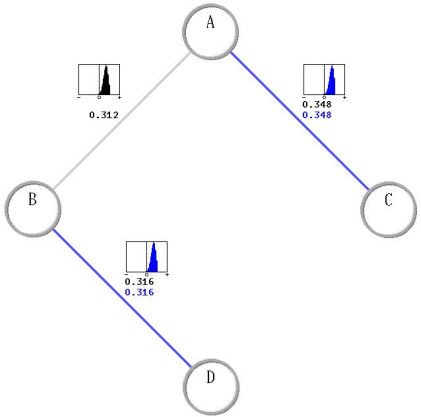
- Invariant Interaction (I^2)
- G2 test (G^2)
- Fisher's exact test (F)

Experiment results

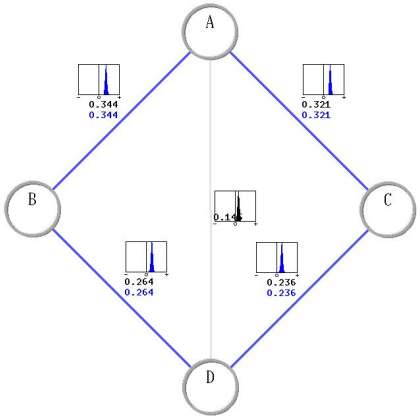




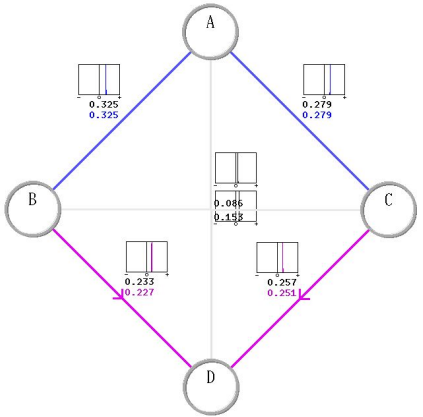
Incremental visualization of uncertainty



80 samples

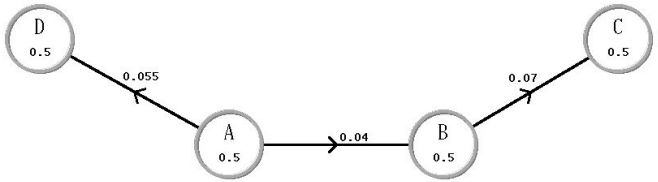


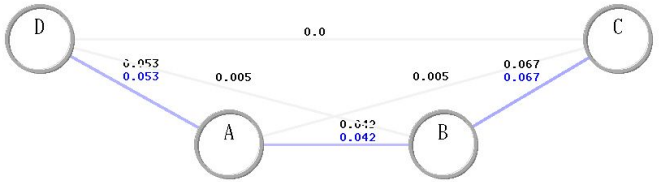
400 samples



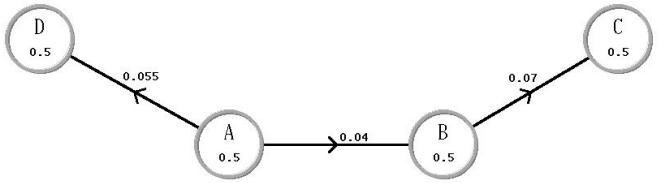
3000 samples

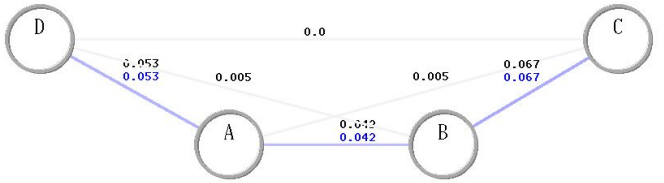
Time series

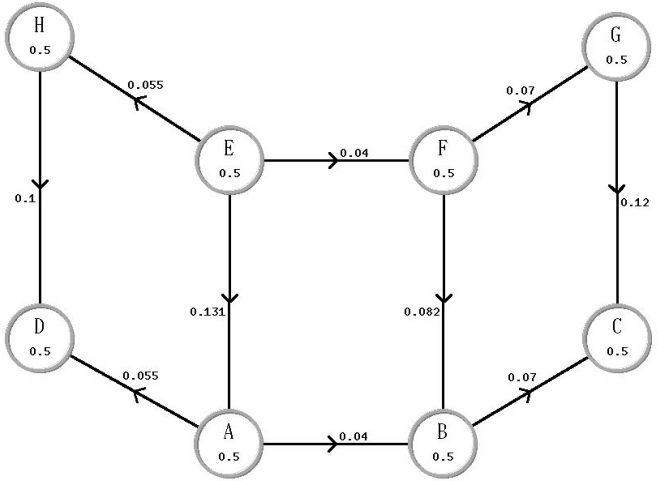

D
0.5
C
0.5
0.055
0.07
A
0.5
0.04
B
0.5

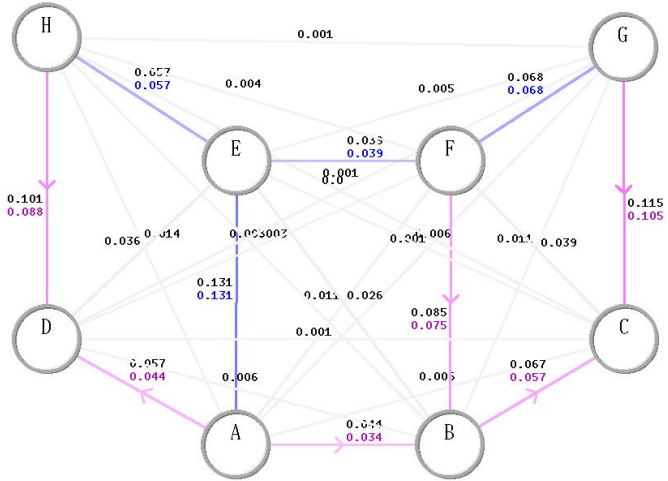

D
C
0.0
0.053
0.053
0.005
0.005
0.067
0.067
A
0.042
0.042
B

Time series


D
0.5
C
0.5
0.055
0.07
A
0.5
0.04
B
0.5


D
C
0.0
0.053
0.053
0.005
0.005
0.067
0.067
A
0.042
0.042
B


H
0.5
G
0.5
0.055
0.07
E
0.5
0.04
F
0.5
0.1
0.12
0.131
0.082
D
0.5
C
0.5
0.055
0.07
A
0.5
0.04
B
0.5


H
0.001
G
0.057
0.057
0.004
0.005
0.068
0.068
E
0.039
0.039
F
0.001
0.0
0.101
0.088
0.115
0.105
0.036
0.014
0.131
0.131
0.026
0.085
0.075
D
0.001
C
0.057
0.044
0.006
0.005
0.067
0.057
A
0.034
0.034
B

Can Causality solve open problems of ML?

1

Can we answer counterfactual questions based on observations only?

Answering counterfactual questions

- Deep generative models have proven successful at designing realistic images
- Providing a disentangle latent representation of the data using Generative models
- Statistical independent is too restrictive, they rely on counterfactual manipulation

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

2

Can we develop automatic data-driven machine learning algorithms?

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]

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

3

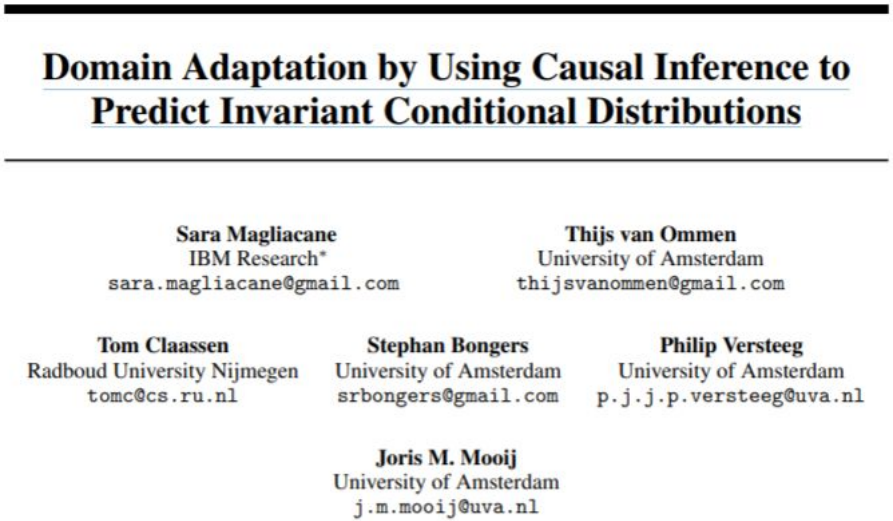
Can we perform domain adaptation using causal relation?

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

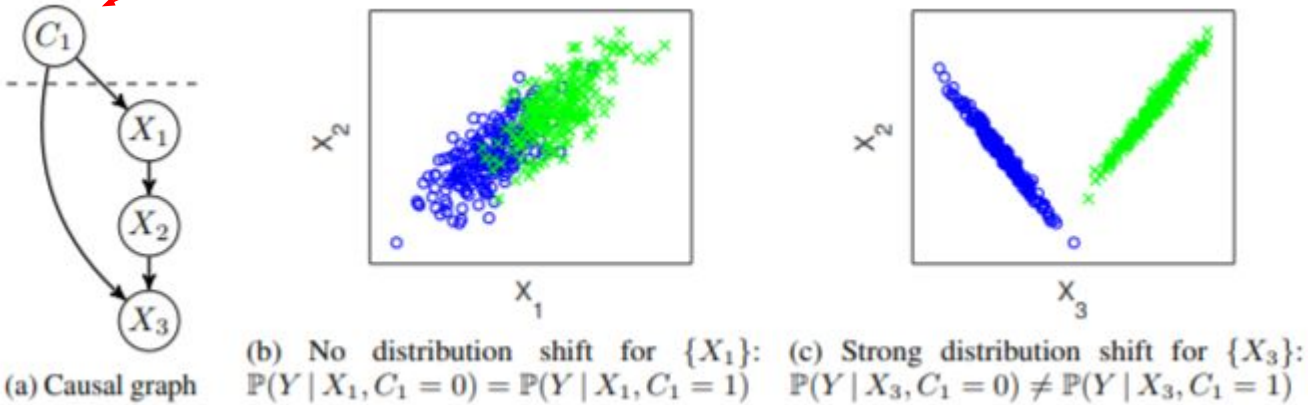
Philip Versteeg
University of Amsterdam
p.j.j.p.versteeg@uva.nl

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

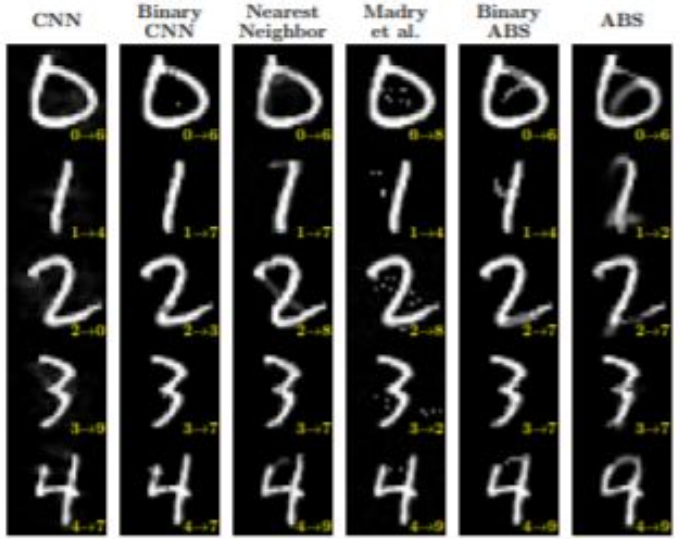
Deep neural networks (DNNs) are susceptible to minimal adversarial perturbations

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

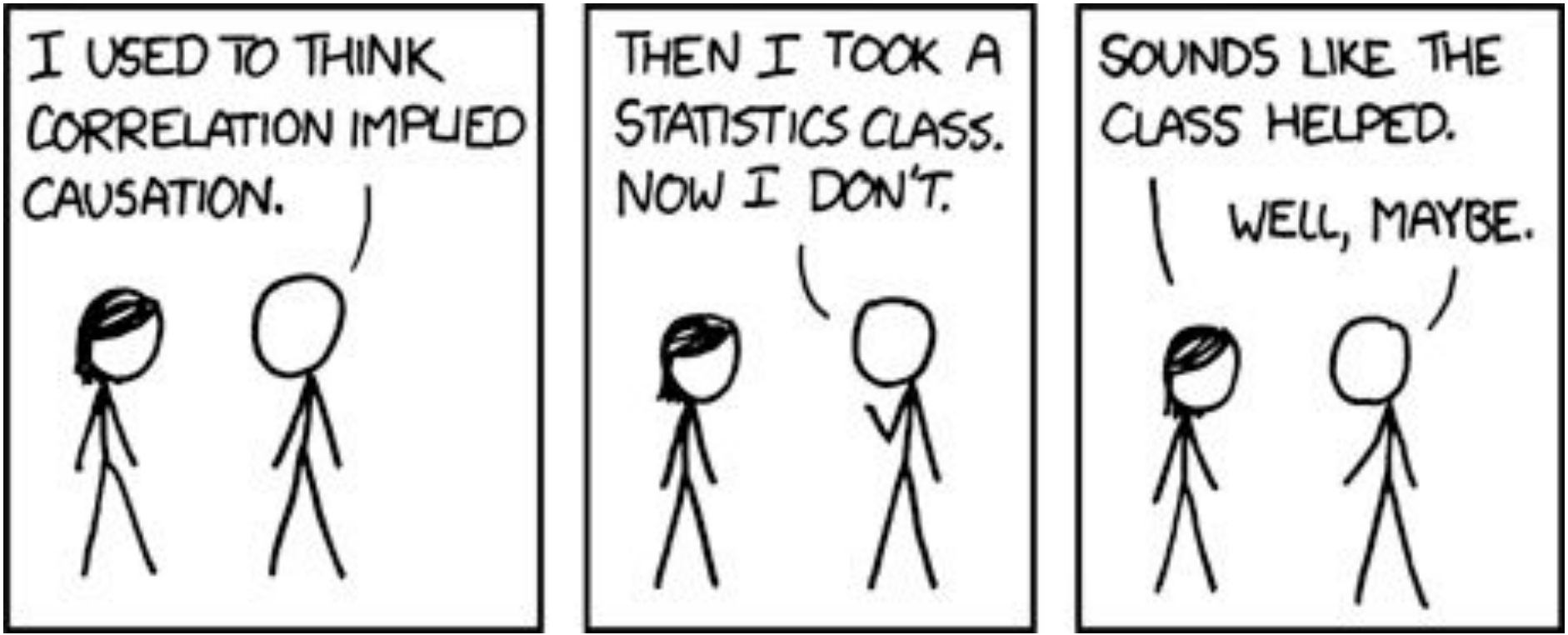


Discovery of causal relations from observational data in real world setting

1. Answering counterfactual questions
    a. Besserve et al 2018. Counterfactuals uncover the modular structure of deep generative models
2. Automatic data-driven algorithms
    a. Parascandolo et al 2018. Learning Independent Causal Mechanisms
3. Improving domain adaptation
    a. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions
4. Increasing robustness and security
    a. Schott et al 2018, Towards the first adversarially robust neural network model on MNIST

Discovery of causal relations from observational data in real world setting

1. Answering counterfactual questions
    a. Besserve et al 2018. Counterfactuals uncover the modular structure of deep generative models
2. Automatic data-driven algorithms
    a. Parascandolo et al 2018. Learning Independent Causal Mechanisms
3. Improving domain adaptation
    a. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions
4. Increasing robustness and security
    a. Schott et al 2018, Towards the first adversarially robust neural network model on MNIST
5. Increasing explainability
    a. Harradon et al 2018, Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations
6. Decreasing a need for huge amount of data
    a. Holst et al 2018. An Invariant Bayesian Conditional Independent Test for more Sensitive Causal Discovery.


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