

I'D BE HAPPY TO EXPLAIN MY RESEARCH. DO YOU WANT THE SHORT, SIMPLIFIED VERSION THAT YOU CAN ACTUALLY UNDERSTAND, OR THE COMPLETE, BEWILDERINGLY COMPLEX VERSION THAT YOU CAN PRETEND TO UNDERSTAND?

TOM GAULD for NEW SCIENTIST

Credit: Tom Gauld



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT



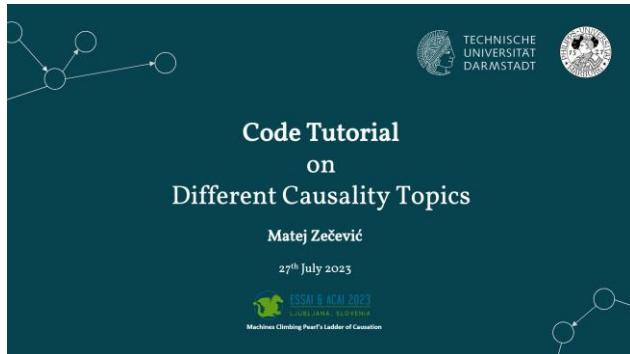
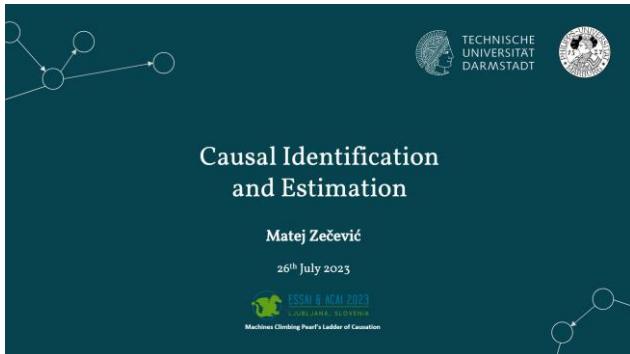
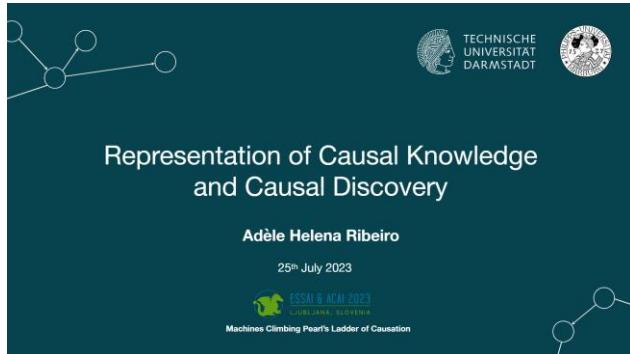
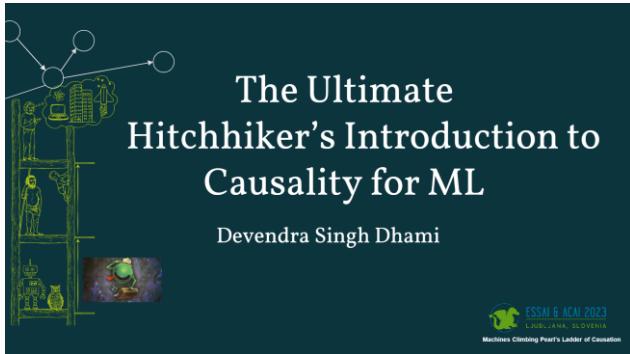
# Machines Climbing Pearl's Ladder of Causation- Day 5

Devendra Singh Dhami  
Matej Zečević  
Adèle Ribeiro



Machines Climbing Pearl's Ladder of Causation

# We have Seen.....





# Current State of Research in Causality (in AI/ML)

Devendra Singh Dhami  
Adèle Ribeiro



Machines Climbing Pearl's Ladder of Causation

**Causality  
And  
More**

**Causal  
Explanations**



**Causal Discovery  
And  
Identification**

**Causal  
Generative  
Modelling**

**Causal Machine Learning**



Causal Discovery  
And  
Identification

## Causal Machine Learning

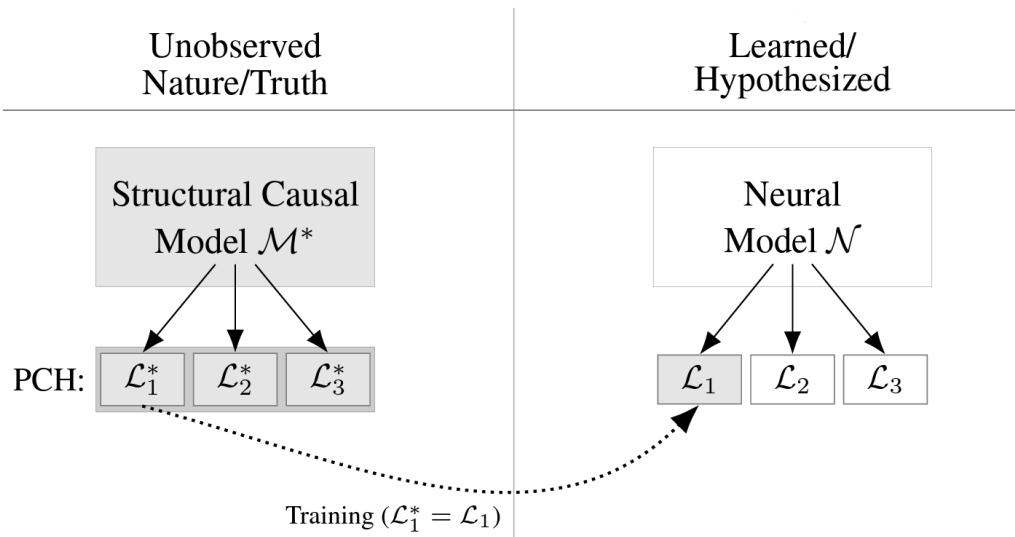
Causal Machine Learning: A Survey and Open Problems, Kaddour et al., 2022



**Causal Machine Learning**

**Causal  
Generative  
Modelling**

# The Neural-Causal Connection



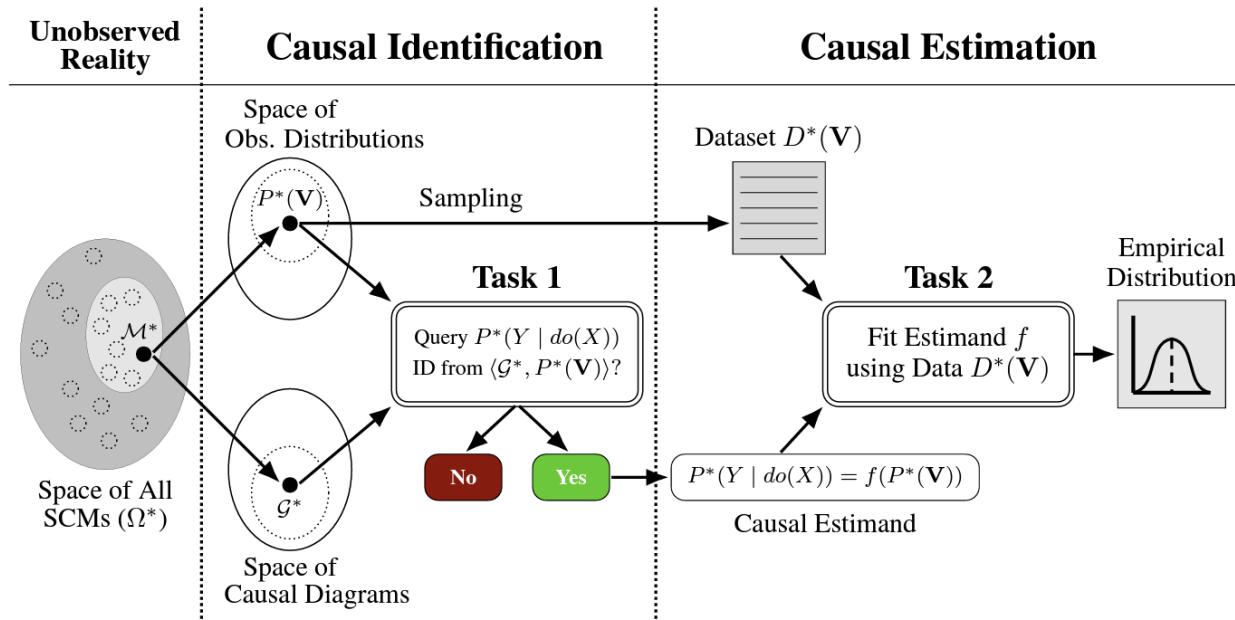
- Neural Network  $\rightarrow$  universal approximators
- Neural Networks should be able to learn any SCM



- Unsurprisingly, not the case

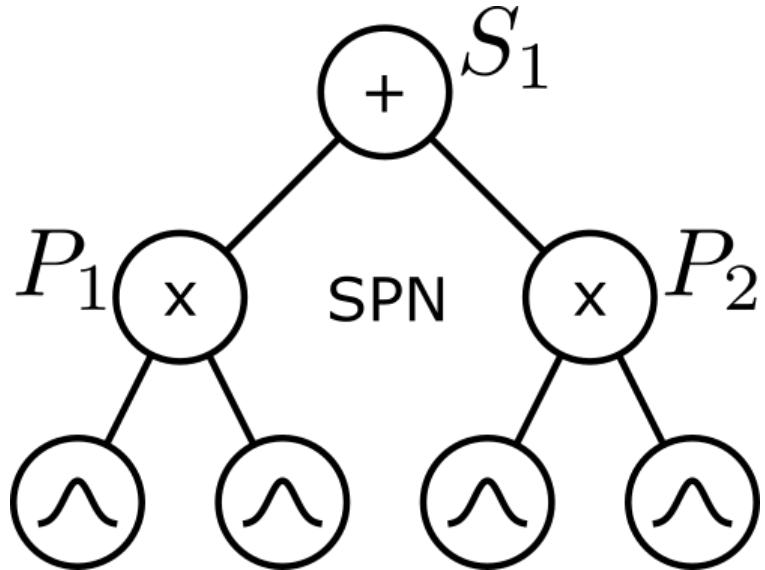
# Not so simple after all....

- Need a special class of SCMs
- Can act as a proxy for the true, unobserved SCM

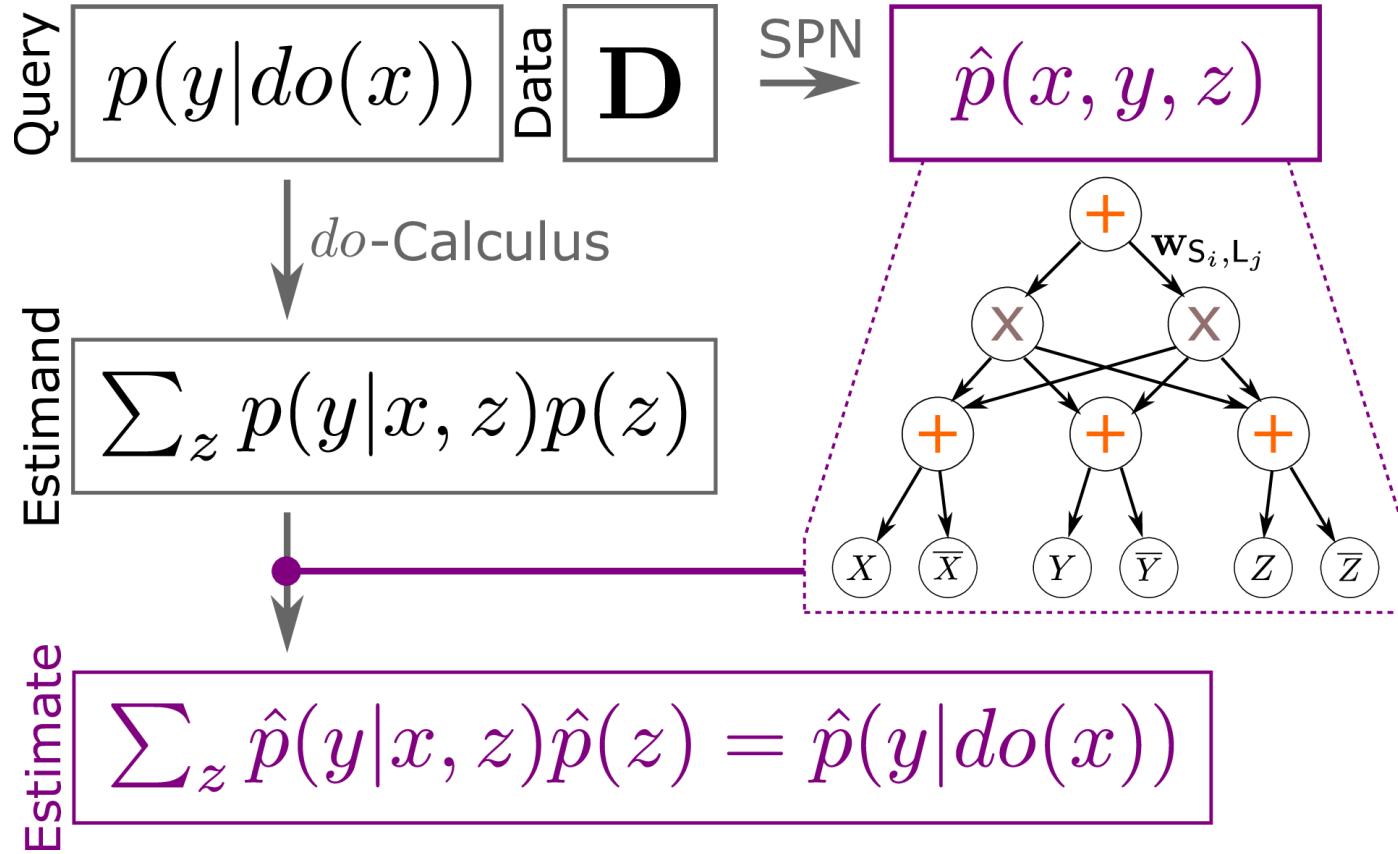


# Sum-Product Networks (SPNs)

- Graphical model
- Calculate joint probability distribution
- Leaf nodes, sum nodes, product nodes
  - Leaf node: Probability distribution for a single variable
  - Sum node: Mixture of child distributions
  - Product node: Product (independence) of child nodes
- Why SPN? Fast: Inference in linear time



# Vanilla SPN Causal Inference



# Let's Stay at the Higher Rung: Interventional SPNs

Structural Causal Model

$$A = U(0, 100)$$

$$F = \frac{1}{2}A + \mathcal{N}(10, 10)$$

$$H = \frac{1}{100}(100 - A^2) + \frac{1}{2}F + \mathcal{N}(40, 30)$$

$$M = \frac{1}{2}H + \mathcal{N}(20, 10)$$

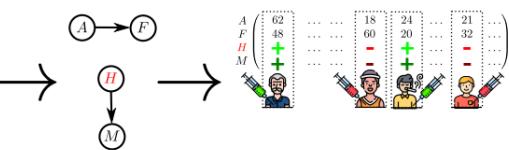
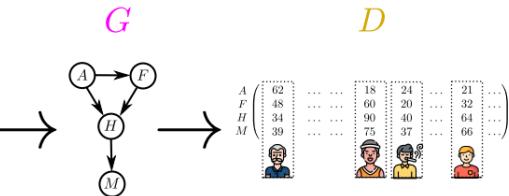


$$A = U(0, 100)$$

$$F = \frac{1}{2}A + \mathcal{N}(10, 10)$$

$$\textcolor{red}{H} = \textcolor{red}{U}(0, 100)$$

$$M = \frac{1}{2}H + \mathcal{N}(20, 10)$$



Univ. Function Approximation

$$f(\textcolor{magenta}{G}; \theta)$$

Neural Network

$$\psi$$

optimize

$$\textcolor{blue}{D} \in \mathbb{R}^{K \times N}$$

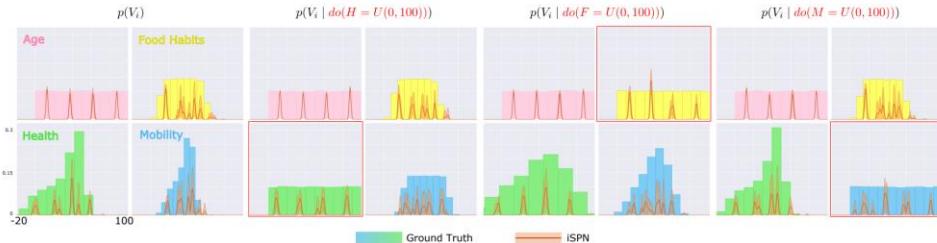
$$\textcolor{magenta}{G} \in \{0, 1\}^{J \times N \times N}$$

Density Estimation

$$g(\textcolor{blue}{D}; \psi)$$

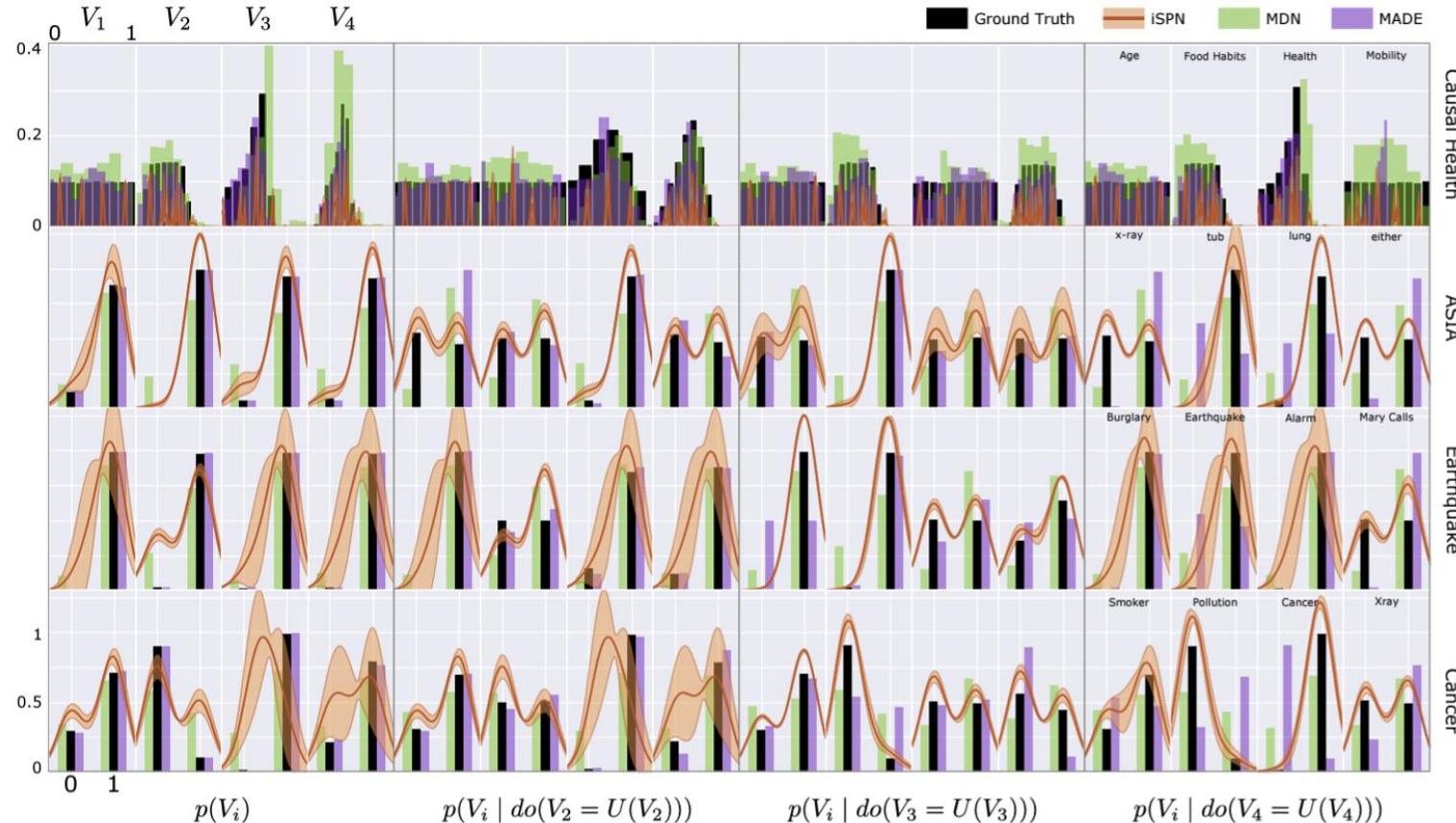
$$:= p(\{V_i\}_i^N \mid do(\{U_j = u_j\}_j^M))$$

Sum-Product Network



Confounding	Conditioning	Ground Truth <sup>†</sup>	CausalML	DoWhy	iSPN
No	0.0374	0.0397 (0.04)	0.0397	0.0397	0.0347 ✓ ATE(asia, tub)
No	0.9271	0.9337 (0.93342)*	0.9337	0.9337	0.9139 ✓ ATE(Burglary, Alarm)
Yes	0.6766	0.6703 (0.667586)	0.6703	0.6697	0.6551 ✓ ATE(brone, dysp)
Yes	-0.0457	0.0537 (0.05)	-0.0454	0.0538	0.0545 ✓ ATE(Surgery, Recovery)

# Show Me It Works



# Show Me It Works

Method \ Query	$V_1$	$V_2$	$V_3$	$V_4$
<b>iSPN</b>	.001 ± .00	.007 ± .01	.003 ± .00	.013 ± .01
<b>MADE</b>	.588 ± .59	.108 ± .16	.015 ± .02	.105 ± .12
<b>MDN</b>	.178 ± .14	.263 ± .14	.184 ± .12	.079 ± .01

Table 1: **Jensen-Shannon-Divergence Evaluation of Estimated Interventional Distributions.** Numerical pendant to Fig.4, mean and standard deviation per  $p(V_{j \setminus i} \mid do(V_i = U(V_i)))$  where  $U$  is the uniform distribution across all data sets. Lower=better.

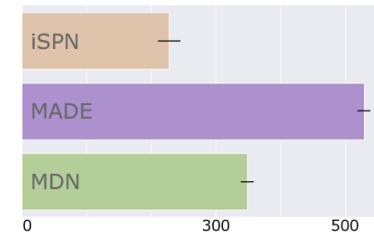
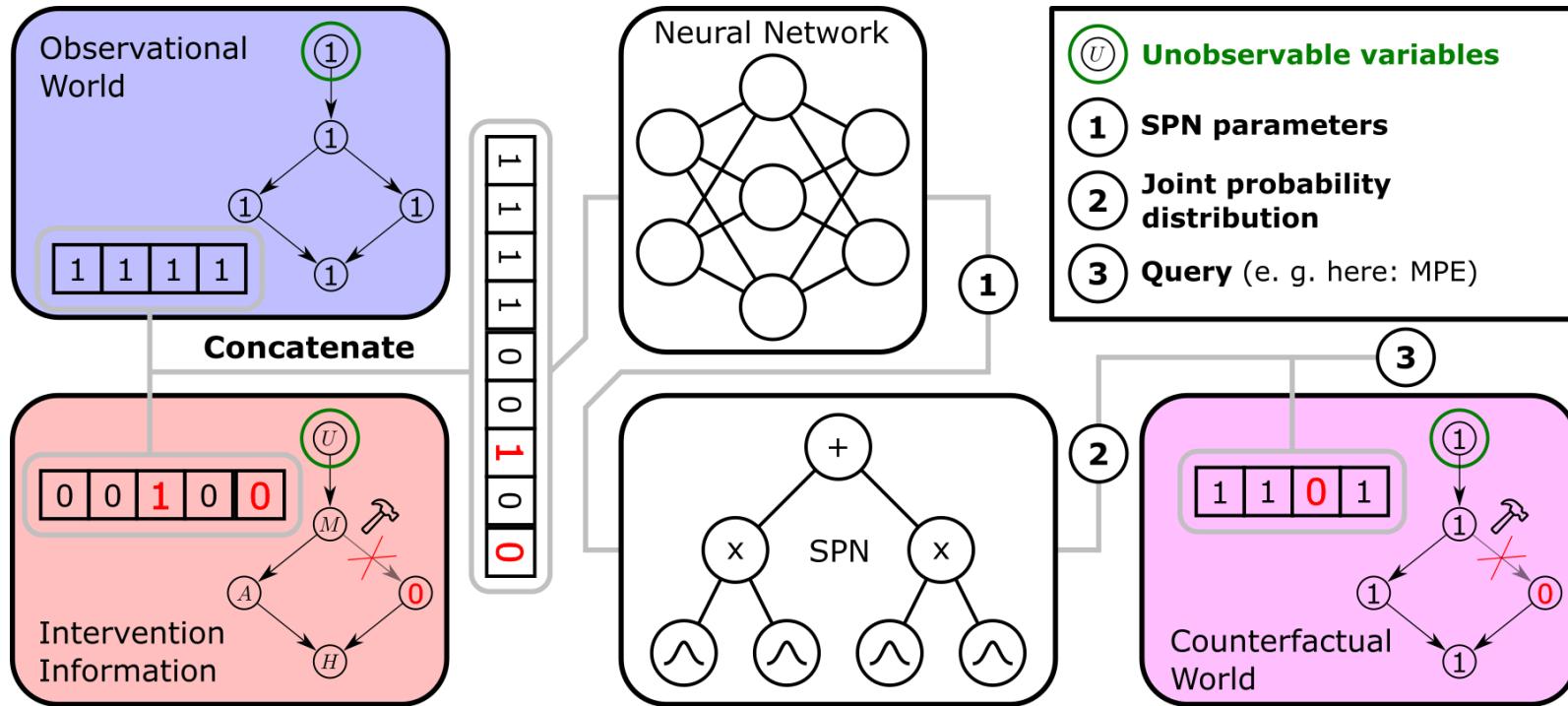


Figure 3: **Mean Running Times in sec. till convergence (Causal Health)** for 50 full passes. More data sets results in supplementary.

# Counterfactual Sum-Product Networks

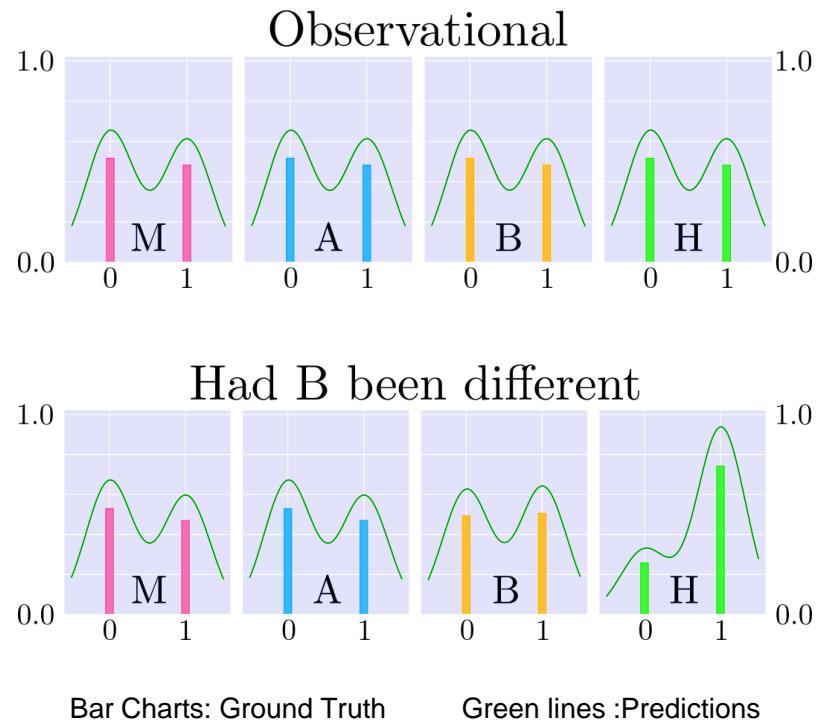
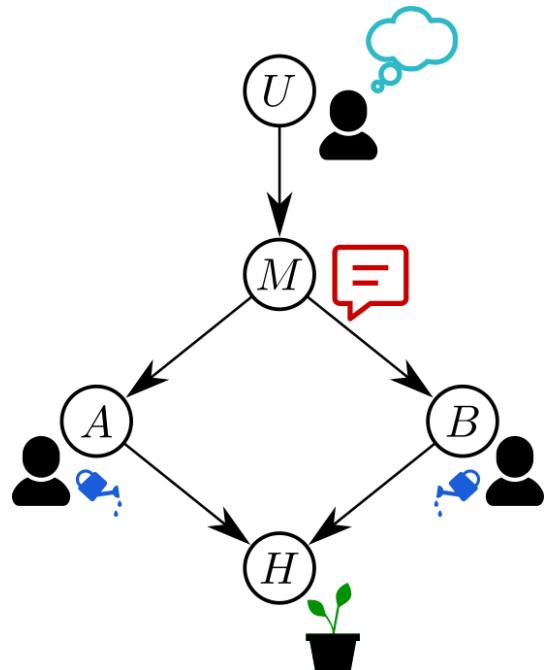


## More formally

**Definition 6.** A counterfactual sum-product network (cf-SPN) is the joint model  $m(\mathbf{D}) = g(\mathbf{D}^{cf}; \psi = f(\mathbf{D}^{orig}; \theta))$ , where  $g(\cdot)$  is a SPN,  $f(\cdot)$  a non-parametric function approximator and  $\psi = f(\mathbf{G})$  are shared parameters.

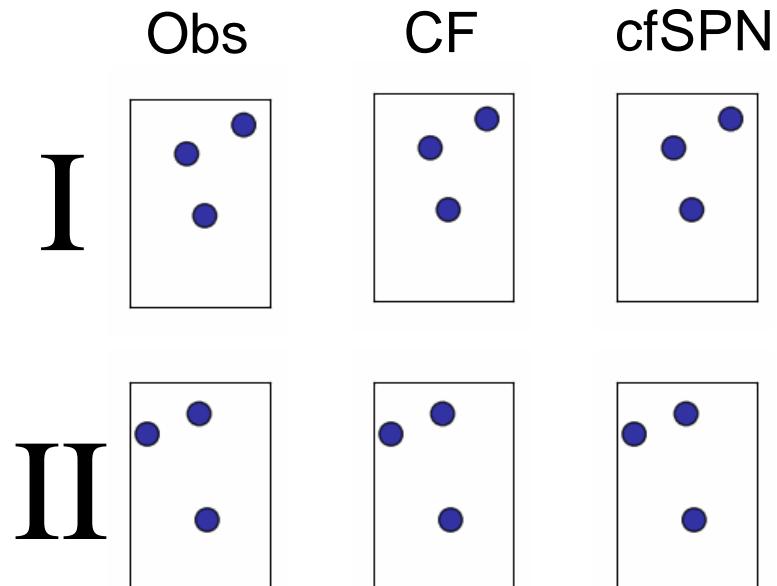
**Proposition 3.** Assuming autonomy and invariance, a cf-SPN  $m(\mathbf{D})$  where  $\mathbf{D} = (\mathbf{D}^{orig} | \mathbf{D}^{cf})$  is able to identify any counterfactual distribution  $p^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}(\mathbf{V}_i = \mathbf{v}_i \mid do(\mathbf{U}_j = \mathbf{u}_j))$ , permitted by a SCM  $\mathfrak{C}$  through counterfactuals, with knowledge of the original world variables  $\mathbf{x} \in \mathbf{D}^{cf}$  and corresponding counterfactual data  $\mathbf{D}^{orig}$  generated from the counterfactual SCMs by modelling the distribution  $p^{\mathfrak{C}|\mathbf{X}=\mathbf{x}, do(\mathbf{U}_j=\mathbf{u}_j)}(\mathbf{V}_i = \mathbf{v}_i)$ .

# Experiments (1000 samples, averaged)



# Experiment: Particle Collision

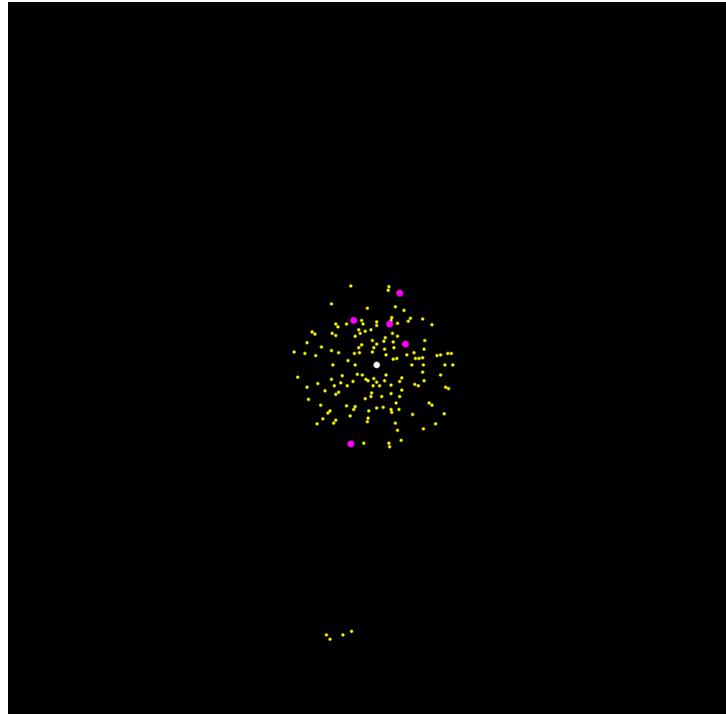
- More difficult problem: Particle simulation with collisions
- Goal: cfSPN prediction should match true counterfactual simulation (CF)
- I: Move the bottom particle to the right after some timesteps
- II: Change the velocity of the top particle to slightly upwards at the start



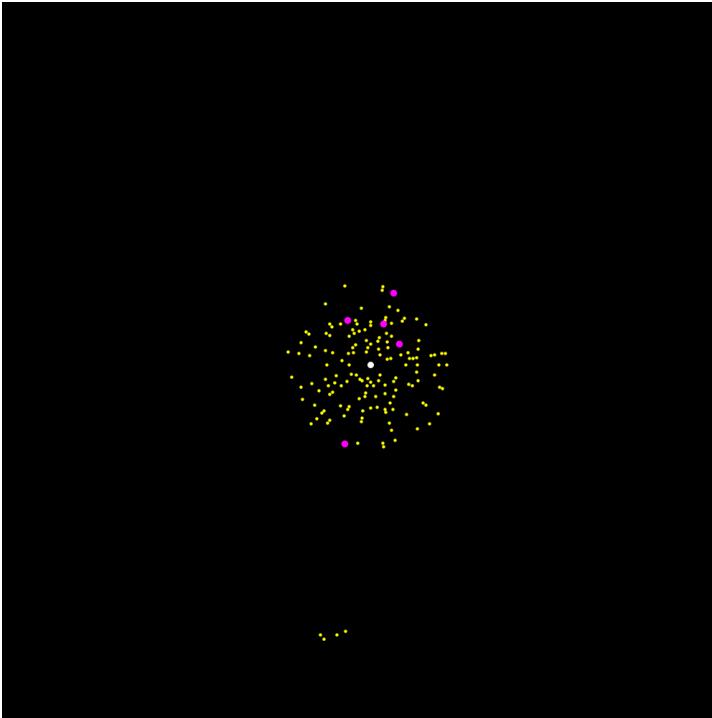
Based on the code repository for particle simulation:  
[https://github.com/ineporozhnii/particles\\_in\\_a\\_box](https://github.com/ineporozhnii/particles_in_a_box)

# Experiment: Galaxies Collision

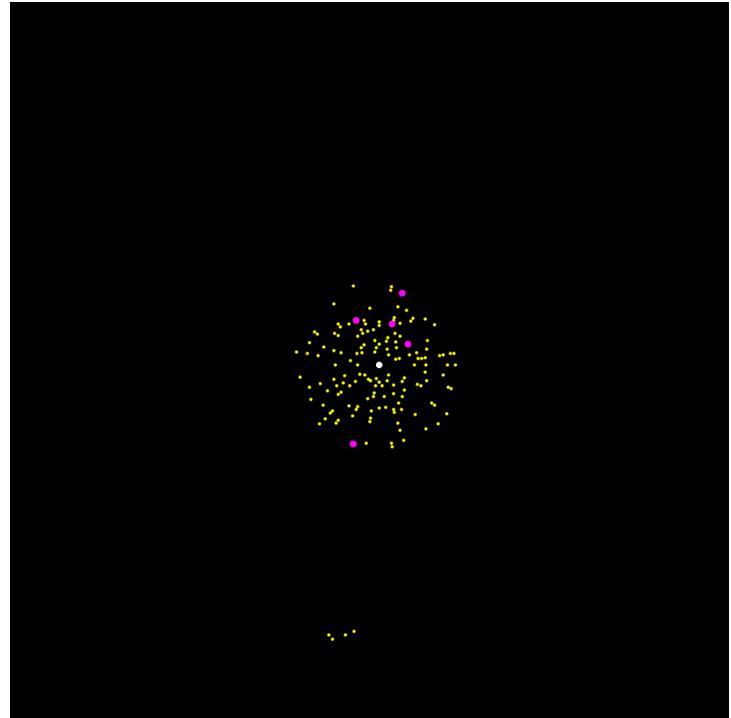
- A much more difficult problem: Particle simulation with gravity collision
- Goal: cfSPN prediction should match true counterfactual simulation (CF)
- Intervene on a few stars
- Scaling is all you need?



Counterfactual



counterfactualSPN

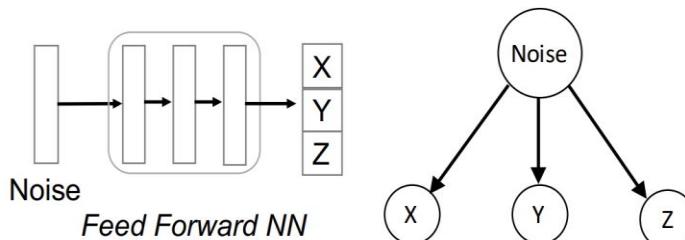


# Let's Move to Images Now

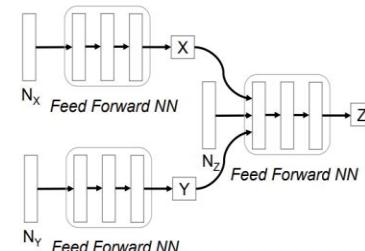
A vibrant sunset over a mountain range. In the foreground, two people stand on a large, craggy rock formation, their arms raised in excitement. The sky is filled with large, soft clouds bathed in orange and yellow light from the setting sun. Several birds are scattered across the sky. The mountains in the background are silhouetted against the bright sky, with dense green forests visible on their slopes. The overall atmosphere is one of adventure and natural beauty.

# Causal Generative Adversarial Networks

- An adversarial training approach: Causality on label space
- Ensuring alignment of generator architecture with the causal graph, adversarial training can achieve generative models that accurately represent observational and interventional distributions
- Proposed CausalGAN and CausalBEGAN architectures improve conditional GAN performance in generating images conditioned on labels.

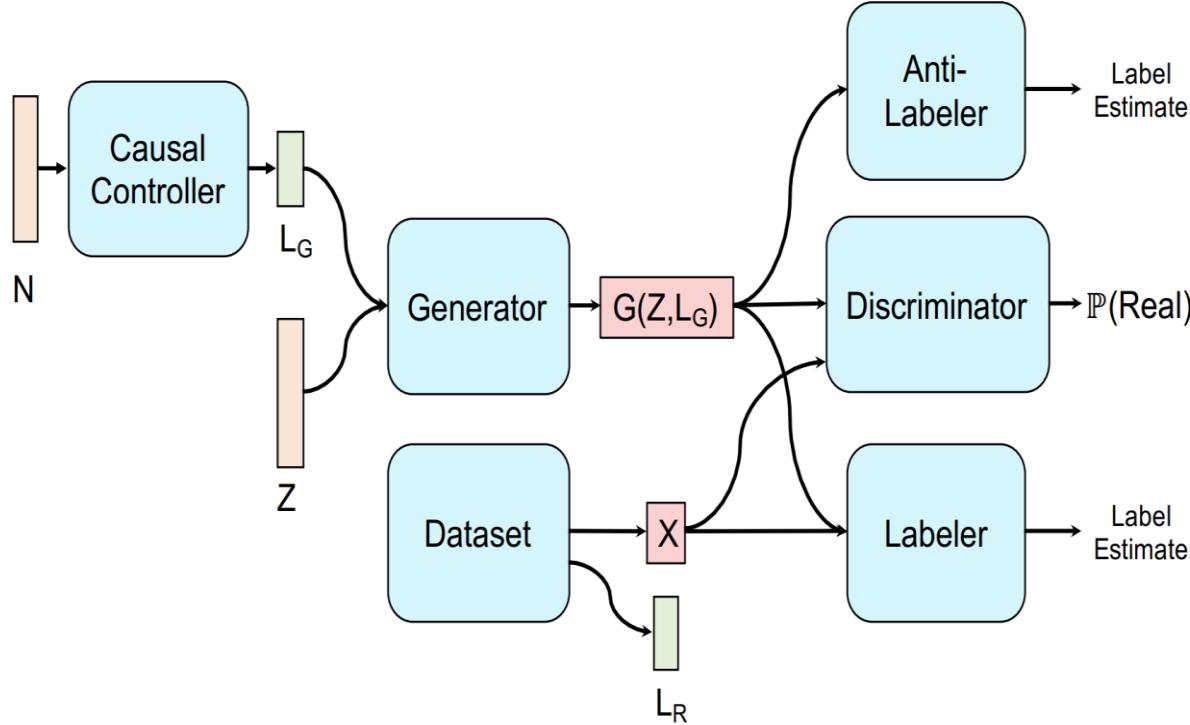


(a) Standard generator architecture and the causal graph it represents



(b) Generator neural network architecture that represents the causal graph  $X \rightarrow Z \leftarrow Y$

# CausalGAN: Architecture



# CausalGAN: Loss Functions

- For a fixed generator

$$\text{Anti- Labeler: } \max_{D_{LG}} \rho \mathbb{E}_{x \sim p_g^1(x)} [\log(D_{LG}(x))] + (1 - \rho) \mathbb{E}_{x \sim p_g^0(x)} [\log(1 - D_{LG}(x))]$$

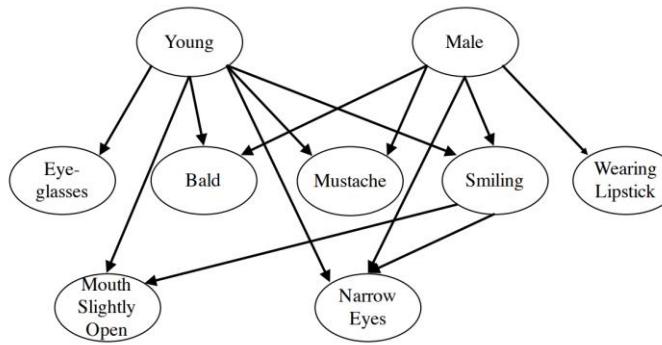
$$\text{Labeler: } \max_{D_{LR}} \rho \mathbb{E}_{x \sim p_{\text{data}}^1(x)} [\log(D_{LR}(x))] + (1 - \rho) \mathbb{E}_{x \sim p_{\text{data}}^0(x)} [\log(1 - D_{LR}(x))]$$

$$\text{Discriminator: } \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log(D(x))] + \mathbb{E}_{x \sim p_g(x)} \left[ \log \left( \frac{1 - D(x)}{D(x)} \right) \right]$$

- For a fixed labeler anti-labeler and discriminator

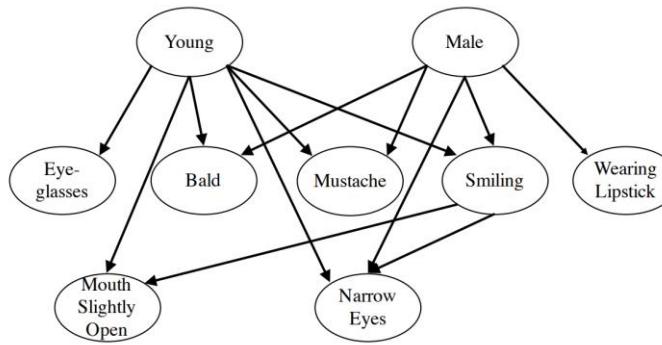
$$\begin{aligned} \text{Generator} \quad & \min_G \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log(D(x))] + \mathbb{E}_{x \sim p_g(x)} \left[ \log \left( \frac{1 - D(x)}{D(x)} \right) \right] \\ & - \rho \mathbb{E}_{x \sim p_g^1(x)} [\log(D_{LR}(X))] - (1 - \rho) \mathbb{E}_{x \sim p_g^0(x)} [\log(1 - D_{LR}(X))] \\ & + \rho \mathbb{E}_{x \sim p_g^1(x)} [\log(D_{LG}(X))] + (1 - \rho) \mathbb{E}_{x \sim p_g^0(x)} [\log(1 - D_{LG}(X))]. \end{aligned}$$

# CausalGAN: Results



(a) Intervening vs Conditioning on Mustache, Top: Intervene Mustache=1, Bottom: Condition Mustache=1

# CausalGAN: Results



(a) Intervening vs Conditioning on Mustache, Top: Intervene Mustache=1, Bottom: Condition Mustache=1



(a) Intervening vs Conditioning on Bald, Top: Intervene Bald=1, Bottom: Condition Bald=1



(a) Intervening vs Conditioning on Wearing Lipstick, Top: Intervene Wearing Lipstick=1, Bottom: Condition Wearing Lipstick=1



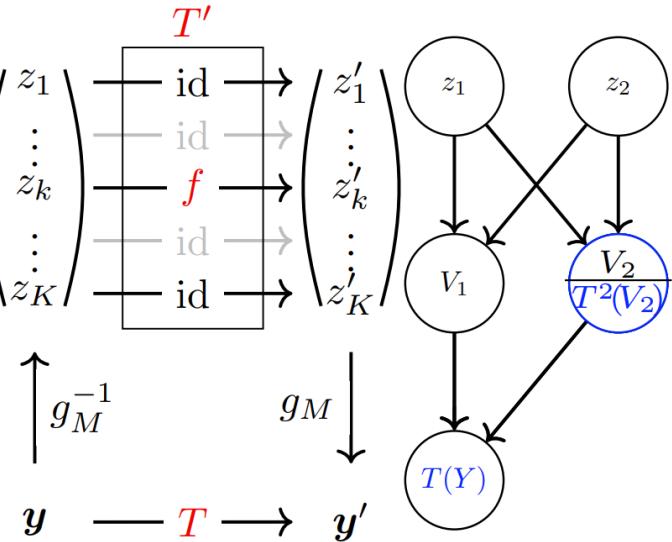
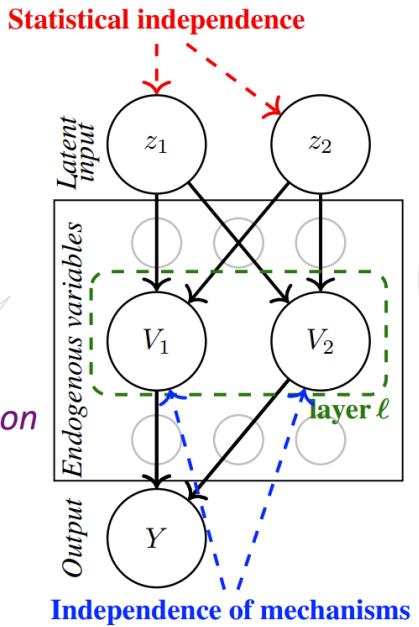
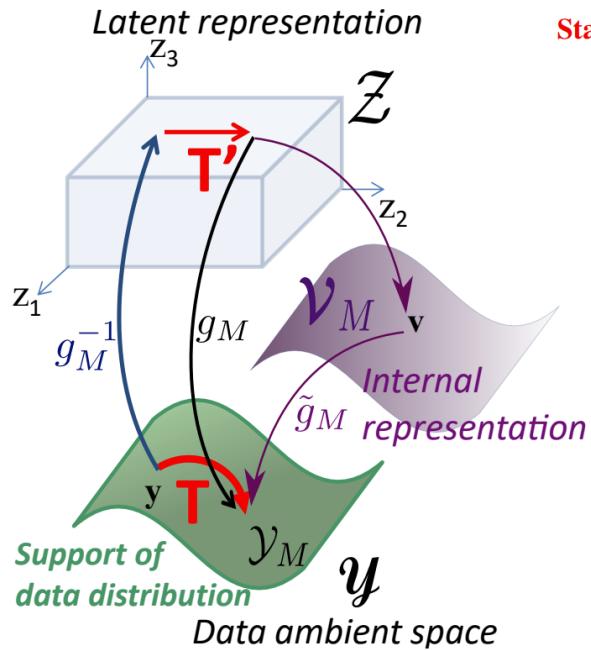
(a) Intervening vs Conditioning on Mouth Slightly Open, Top: Intervene Mouth Slightly Open=1, Bottom: Condition Mouth Slightly Open=1

# Counterfactuals uncover the modular structure of DGMs

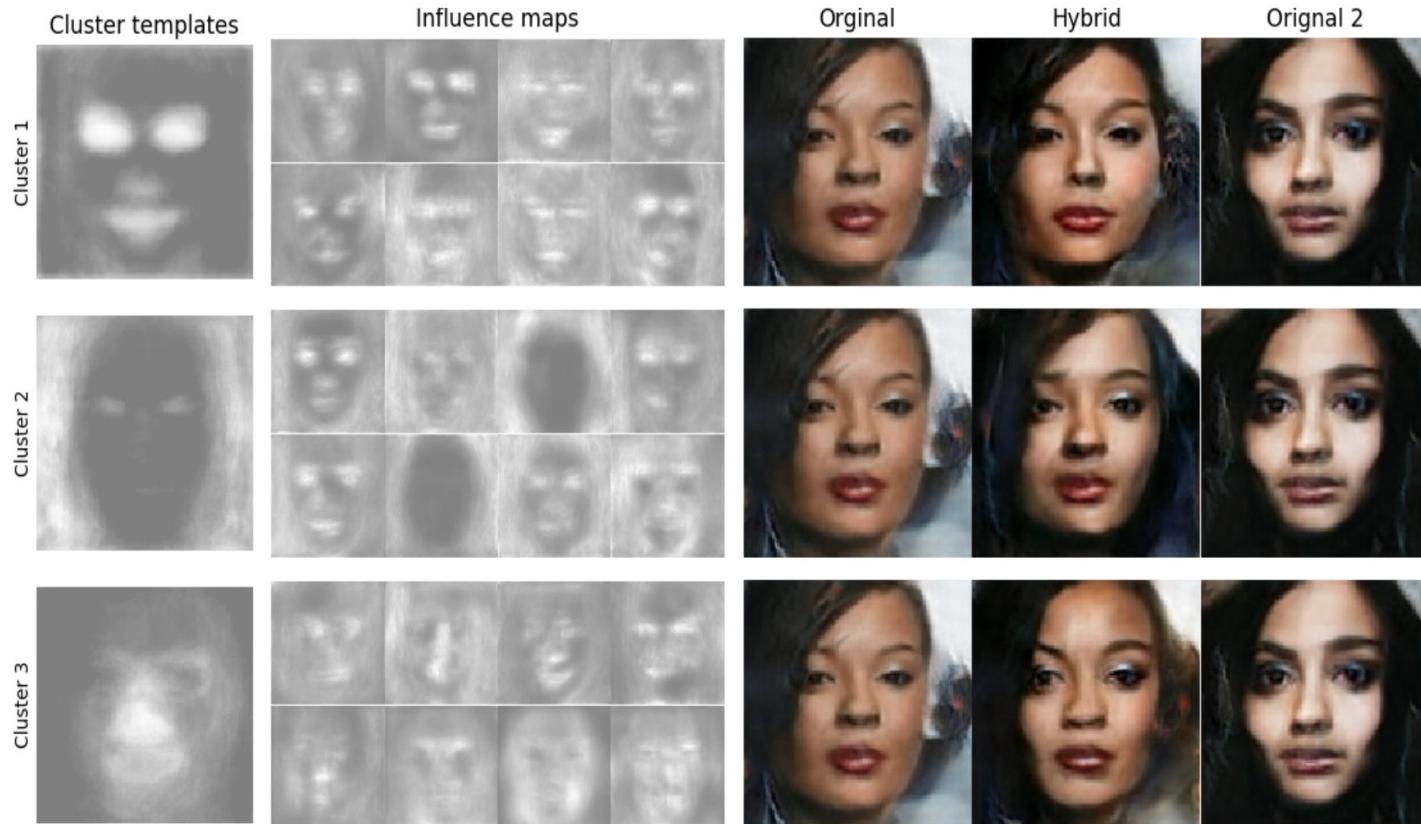
- Enables meaningful and controllable transformations in the data space without requiring explicit supervision
- **Challenge:** Deep generative models can provide latent representations of complex image datasets, but manipulating these representations for controlled transformations is challenging without supervision
- By applying counterfactual manipulations, the modular structure of the network is uncovered, composed of disentangled groups of internal variables



# Causal Generative Models: Architecture



# Causal Generative Models: Results



# Still a Developing Field

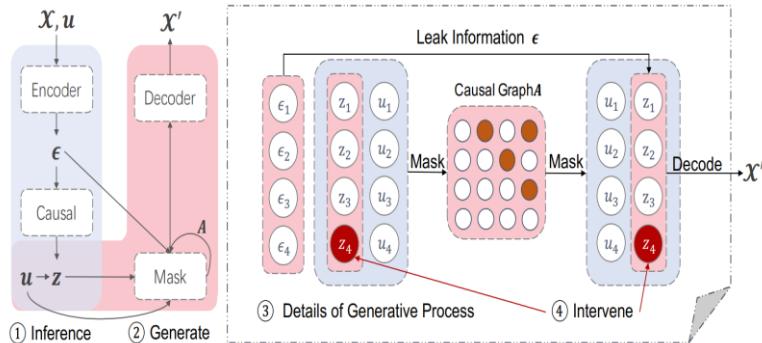


Figure 2. Model structure of CausalVAE. The encoder takes observation  $\mathbf{x}$  as inputs to generate independent exogenous variable  $\epsilon$ , whose prior distribution is assumed to be standard Multivariate Gaussian. Then it is transformed by the Causal Layer into causal representations  $\mathbf{z} = \epsilon + \mathbf{u}$  (Eq. 1) with a conditional prior distribution  $p(\mathbf{z}|\mathbf{u})$ . A Mask Layer is then applied to  $\mathbf{z}$  to resemble the SCM in Eq. 2. After that,  $\mathbf{z}$  is taken as the input of the decoder to reconstruct the observation  $\mathbf{x}$ .

# CausalVAE

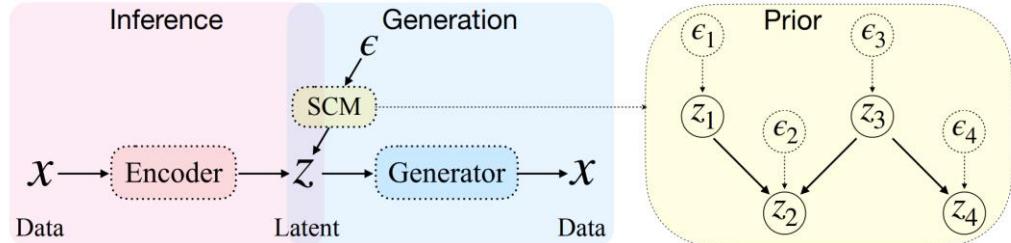


Figure 1: Model structure of a BGM (left) with an SCM prior (right).

# Disentangled gEnerative cAusal Representation (DEAR)

# Lots of Open Questions

- Are the causal generative models really causal?
- How to capture the inherent causality in language?
- Causal image generators
- Causal Circuits
- Continual Causality



Prompt: Make it a dark **star-lit** night.

## (Some) References

- Zhou et al., On the Opportunity of Causal Deep Generative Models: A Survey and Future Directions, <https://arxiv.org/abs/2301.12351>
- Anciukevičius et al., Unsupervised Causal Generative Understanding of Images, NeurIPS 2022
- Davis and Marcus, Causal generative models are just a start, Behavioral and Brain Sciences, 2017
- Goudet et al., Causal Generative Neural Networks, <https://arxiv.org/abs/1711.08936>
- Google ☺



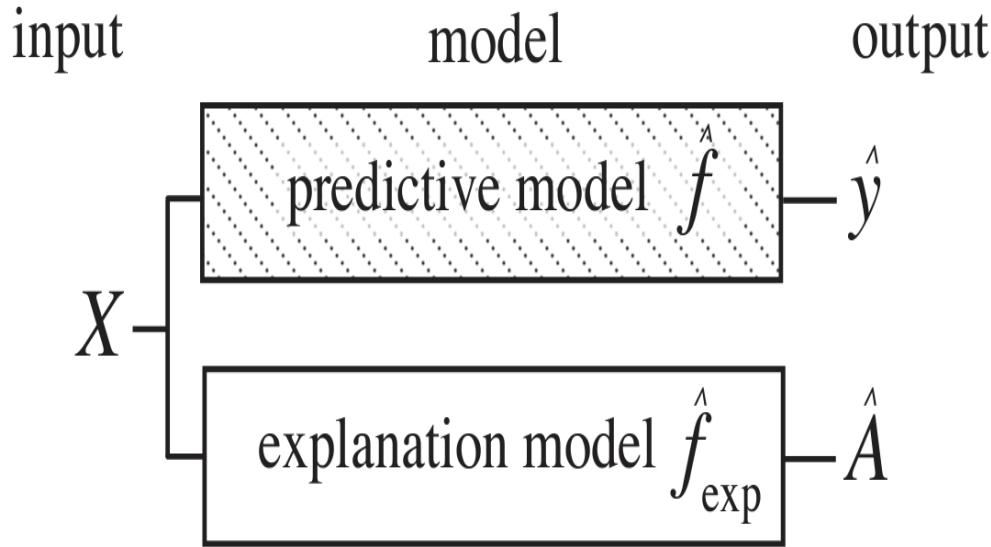
**Causal Machine Learning**

**Causal  
Explanations**

# eXplainable Artificial Intelligence (XAI)

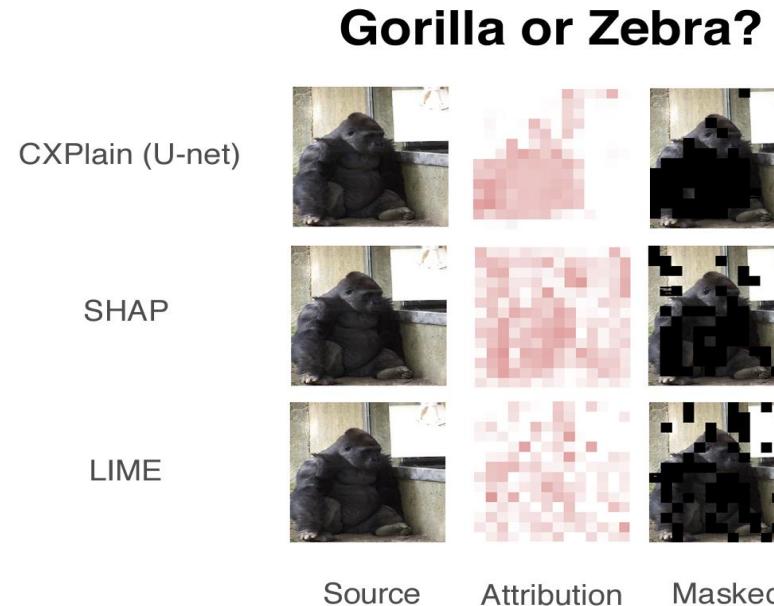
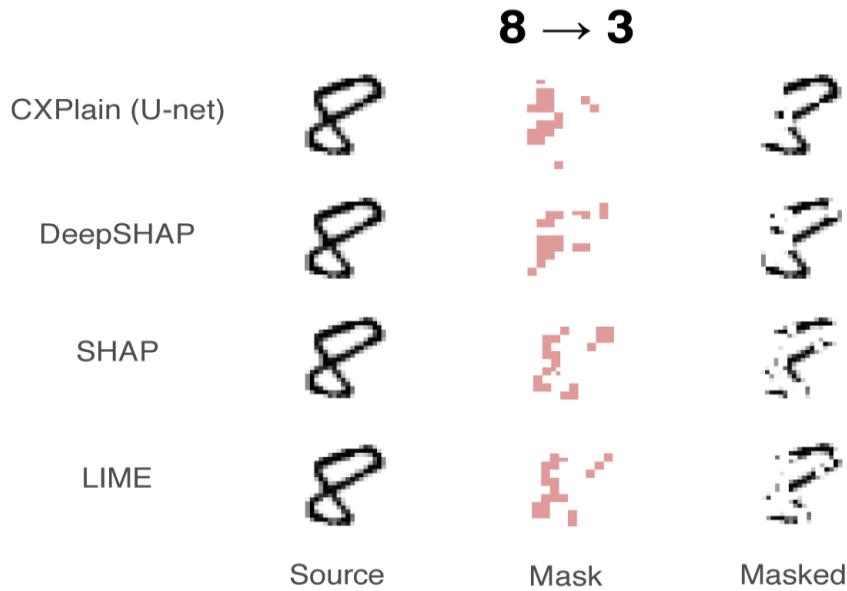
**Explainable and Interpretable AI/ML.** A great body of work within deep learning has provided visual means for explanations of how a neural model came up with its decision i.e., importance estimates for a model’s prediction are being mapped back to the original input space e.g. raw pixels in the arguably standard use-case of computer vision (Selvaraju et al., 2017; Schulz et al., 2020). Formally defined in (Sundararajan et al.), we simply have that  $A_F(\mathbf{x}) = (a_1, \dots, a_n) \in \mathbb{R}^n$  is an attribution of predictive model  $F$  when  $a_i$  is the contribution of  $x_i$  for prediction  $F(\mathbf{x})$  (with  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ ). Recently, Stammer et al. (2021) argued that such explanations are insufficient for any task that requires symbolic-level knowledge while comparing the existing state of explanations to “children that are only able to point fingers but lack articulation”. (Stammer et al., 2021) therefore proposed a neuro-symbolic explanation scheme to revise and ultimately circumvent “Clever Hans” like behavior from learned models in a XIL user-model loop (Teso & Kersting, 2019). On the causal end, (Schwab & Karlen, 2019) proposed a model-agnostic approach that can generate explanations following the idea of Granger causality (which is very different from Pearlian causality as it captures “temporal relatedness” which holds in their setting as input precedes output). Specifically, they train a surrogate model to capture to what degree certain inputs cause outputs in the model to be explained. They achieve this by simply comparing the prediction loss of the model for the original input  $\mathcal{L}(y, \hat{y}_X)$  (where  $X$  is the input) with the alternate prediction loss when a certain feature  $i$  is being removed  $\mathcal{L}(y, \hat{y}_{X \setminus \{i\}})$ . On the Pearlian side of explanations, the arguably closest works on explainable AI/ML can be found in research around fairness (Kusner et al., 2017; Plecko & Bareinboim, 2022). For instance, Karimi et al. (2020) investigated how to best find a counterfactual that flips a decision of interest e.g. an applicant for a credit is rejected and the question is now which counterfactual setting (changes to the applicant) would have resulted in a credit approval. From a purely causal viewpoint, our work might be compared to the definitions of Halpern (2016) for “actual causation.”

# CXPlain



**Main Contribution:** A causal objective that optimizes explanation models to learn to explain another predictive model

# CXPlain: Results



# eXplainable Interactive Learning (XIL)

---

**Algorithm 1** CAIPI takes as input a set of labelled examples  $\mathcal{L}$ , a set of unlabelled instances  $\mathcal{U}$ , and iteration budget  $T$ .

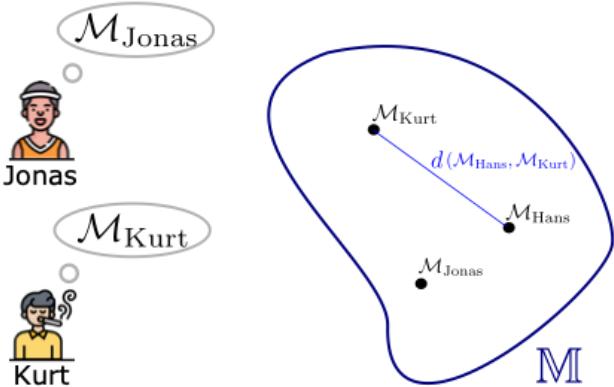
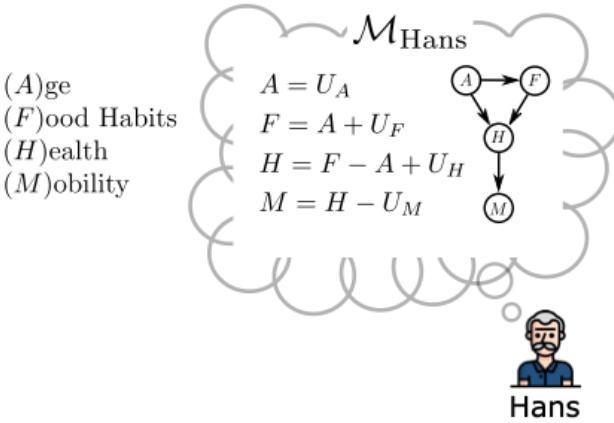
---

```
1:  $f \leftarrow \text{FIT}(\mathcal{L})$ 
2: repeat
3:    $x \leftarrow \text{SELECTQUERY}(f, \mathcal{U})$ 
4:    $\hat{y} \leftarrow f(x)$ 
5:    $\hat{z} \leftarrow \text{EXPLAIN}(f, x, \hat{y})$ 
6:   Present  $x$ ,  $\hat{y}$ , and  $\hat{z}$  to the user
7:   Obtain  $y$  and explanation correction  $\mathcal{C}$ 
8:    $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^c \leftarrow \text{TOCOUNTEREXAMPLES}(\mathcal{C})$ 
9:    $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x, y)\} \cup \{(\bar{x}_i, \bar{y}_i)\}_{i=1}^c$ 
10:   $\mathcal{U} \leftarrow \mathcal{U} \setminus (\{x\} \cup \{\bar{x}_i\}_{i=1}^c)$ 
11:   $f \leftarrow \text{FIT}(\mathcal{L})$ 
12: until budget  $T$  is exhausted or  $f$  is good enough
13: return  $f$ 
```

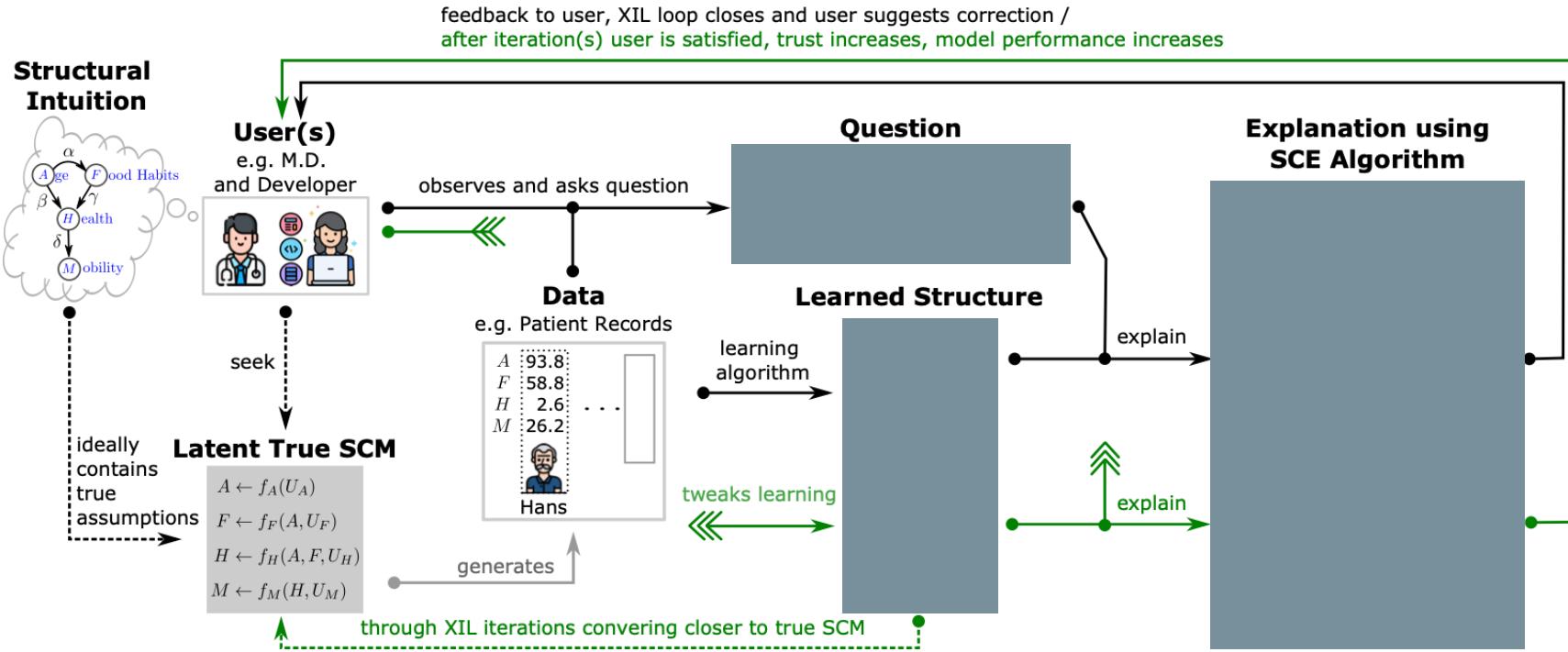
---

During interactions between the system and the user, three cases can occur: (1) **Right for the right reasons**: The prediction and the explanation are both correct. No feedback is requested. (2) **Wrong for the wrong reasons**: The prediction is wrong. As in active learning, we ask the user to provide the correct label. The explanation is also necessarily wrong, but we currently do not require the user to act on it. (3) **Right for the wrong reasons**: The prediction is correct but the explanation is wrong. We ask the user to provide an explanation *correction*  $\mathcal{C}$ .

# Philosophical Point: Human Thoughts in Terms of SCM



# The Causal XIL Loop (Before Our Derivation)



# Formalizing the Question Type

**Definition 1** (Why Question). *Let  $x \in Val(X)$  be an instance of  $X \in \mathbf{V}$  of SCM  $\mathcal{M}$ . Further, let  $\mu^X$  be the empirical mean for a set of samples ( $\mu^X := \frac{1}{n} \sum_i^n x_i$ ) and let  $R \in \{<, >\}$  be a binary ordering relation. We call  $Q_X := R(x, \mu^X)$  a (single) why question if  $Q_X$  is true.*

Checking back with the definition, we see that **Q1** defines a valid question for the Causal Hans example since  $Q_M := m_H < \mu^M = 26.2 < 35.6$  evaluates to true.

# Formulating an Explanation

**Explanation 1** (for Q1). *“Hans’s Mobility is bad because of his bad Health which is mostly due to his high Age although his Food Habits are good.”*

Explanation 1 is a truly causal answer to the observation about Hans’s mobility deficiency based on SCM  $\mathcal{M}$ . It captures both the existence and the “strength” of a causal relation. In the following we will capture and formalize our intuition that allowed us to derive Exp.1. This will allow us to move towards computing such causal explanations automatically.

# Generalizing the Key Ideas in Logic

We mainly used four ideas or pieces of knowledge in our argument above: (I) that there is a relative notion in the why question  $Q_M$  like “why ... bad?” that implicitly compares an individual (here, Hans) to the remaining population, (II) note that by definition there can only exist a causal effect from some variable to another *if and only if* one is the argument of the other in a structural equation of  $\mathcal{M}$ , (III) the causal effect  $\alpha_{X \rightarrow Y}$  allows us to assert whether the observed values for  $(x, y)$  are “surprising” or not (e.g. it was not surprising that  $m_H < \mu^M$  after observing  $h_H > 0$  and knowing that  $\gamma > 0$  since decreasing health means decreasing mobility in general and Hans is old), and (IV) that some causal effects are more important or influential than others (e.g. age versus food habits w.r.t. health). We can neatly collect all information from (I-III) in a single tuple which we call causal scenario.

**Definition 2.** *The tuple  $C_{XY} := (\alpha_{X \rightarrow Y}, x, y, \mu^X, \mu^Y)$  is called causal scenario.*

The (IV) point we can capture separately as will be shown below. Now, we finally express our build up intuition and understanding into rules expressed in first-order logic that will then allow us to compute causal explanations like Exp. 1 automatically.

**Definition 3** (Explanation Rules). *Let  $C_{XY}$  denote a causal scenario, let  $s(x) \in \{-1, 1\}$  be the sign of a scalar, let  $R_i \in \{<, >\}$  be a binary ordering relation and let  $\mathcal{Z}_X = \{|\alpha_{Z \rightarrow X}| : Z \in \text{Pa}_X\}$  be the set of absolute parental causal effects onto  $X$ . We define FOL-based rule functions as*

(ER1) *If  $R_1 \neq R_2$ , then:  $R_1(s(\alpha_{X \rightarrow Y}), 0) \wedge (R_2(y, \mu^Y) \vee R_1(x, \mu^X)),$*

(ER2) *If  $R_1 \neq R_2$ , then:  $R_1(s(\alpha_{X \rightarrow Y}), 0) \wedge R_1(y, \mu^Y) \wedge R_2(x, \mu^X)$ , and*

(ER3) *If  $|\mathcal{Z}_X| > 1$ , then  $Y \iff \arg \max_{Z \in \mathcal{Z}_X} Z$*

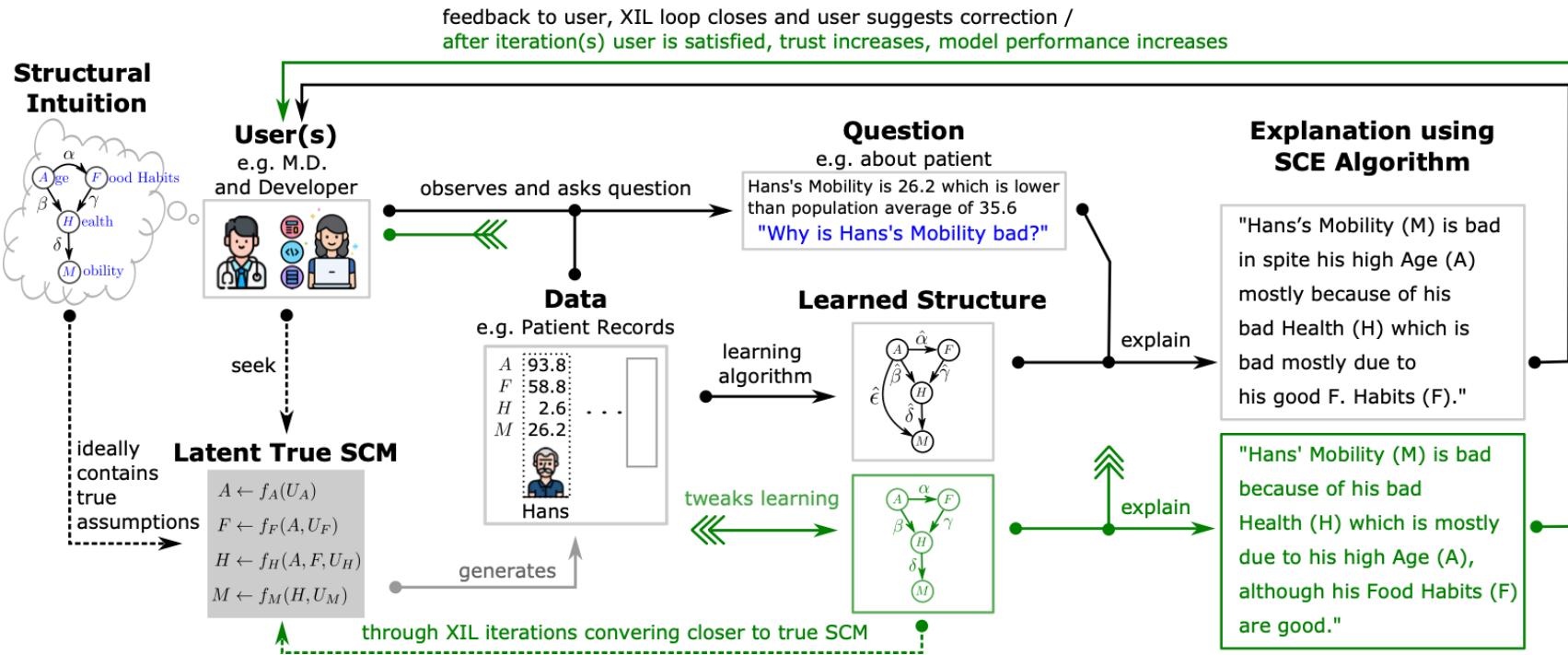
*indicating for each rule  $ERi(\cdot) \in \{-1, 0, 1\}$  how the causal relation  $X \rightarrow Y$  satisfies that rule.*

# Pronouncing the Rules + Inspiration

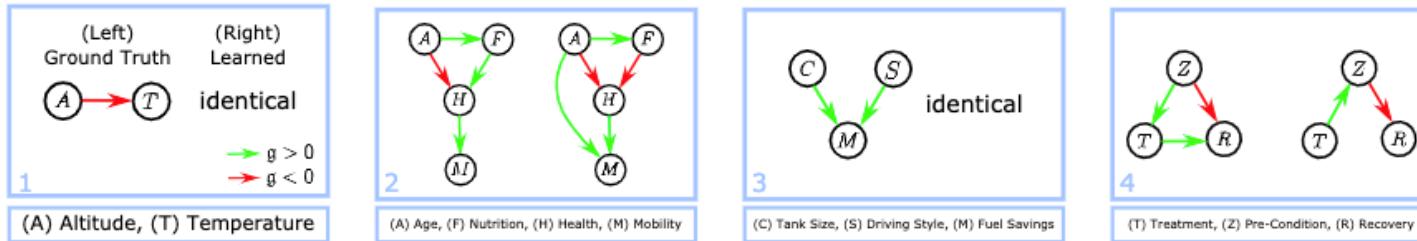
<i>ER1</i>	Excitation	“Y because of X [being low/high]”
<i>ER2</i>	Inhibition	“Y although X [is low/high]”
<i>ER3</i>	Preference	“mostly” + <i>ER1</i> or <i>ER2</i> pronunciation

Table 1: **Pronunciation Scheme.** Right shows the natural language reading of a rule’s activation.

# The Causal XIL Loop for SCE

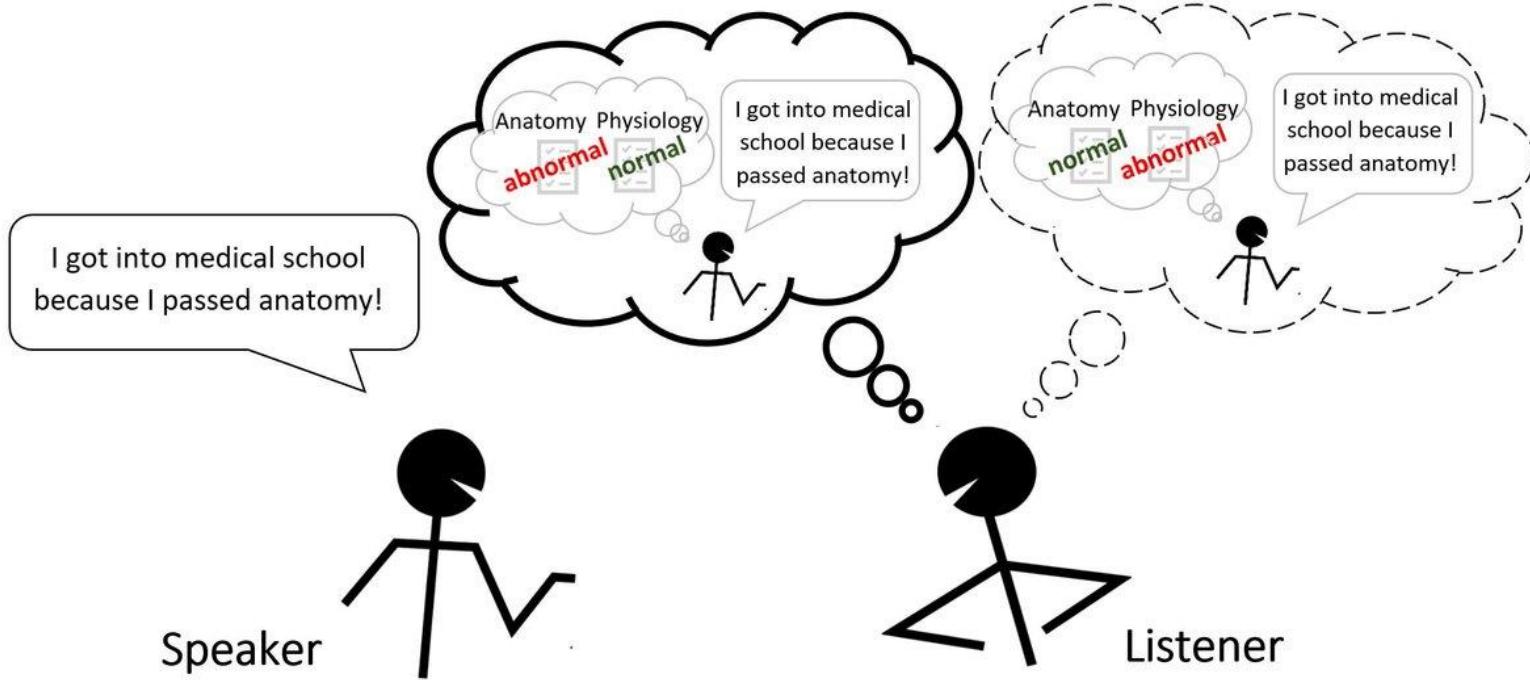


# Quality of Learned Explanations

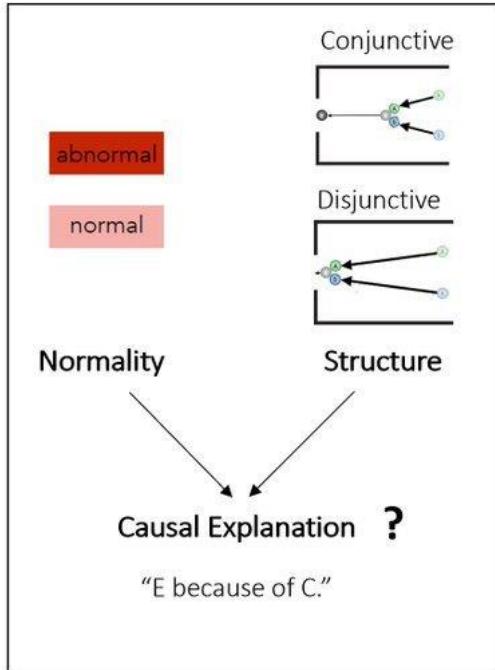


- |   | (Question)  | (Ground Truth)               | (Learned)   |
|---|---|------------------------------|---|
| 1 | “Why is the Temperature at the Matterhorn low?”<br>“The Temperature at the Matterhorn is low because of the high Altitude.”                                 | (Question)<br>(Ground Truth) | <b>“The Temperature at the Matterhorn is low because of the high Altitude.”</b>   |
| 2 | “Why is Hans’s Mobility bad?”<br>“Hans’s Mobility is bad because of his bad Health which is mostly due to his high Age, although his Food Habits are good.” |                              | <b>“Hans’s Mobility, in spite his high Age, is bad mostly because of his bad Health which is bad mostly due to his good Food Habits.”</b> |
| 3 | “Why is your personal car’s left Mileage low?”<br>“Your left Mileage is low because of your small Car and your bad Driving Style.”                          |                              | <b>“Your left Mileage is low because of your small Car and your bad Driving Style.”</b>   |
| 4 | “Why did Kurt not Recover?”<br>“Kurt did not Recover because of his bad Pre-condition, although he got Treatment.”  |                              | <b>“Kurt did not Recover because of his bad Pre-Condition, which were bad although he got Treatment.”</b>                                 |

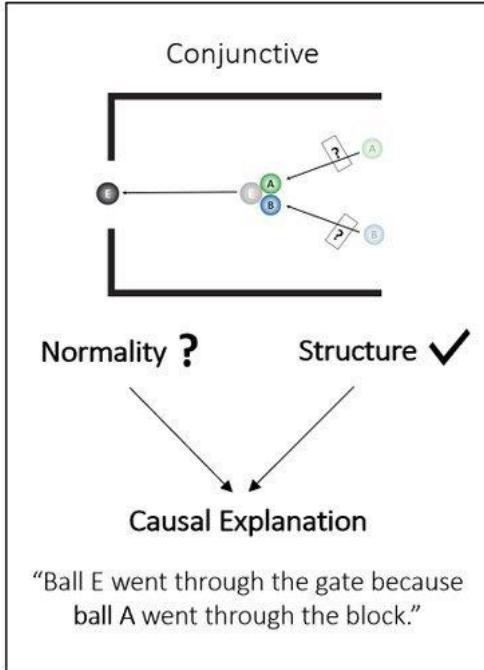
# Normal vs Abnormal Causal Explanations



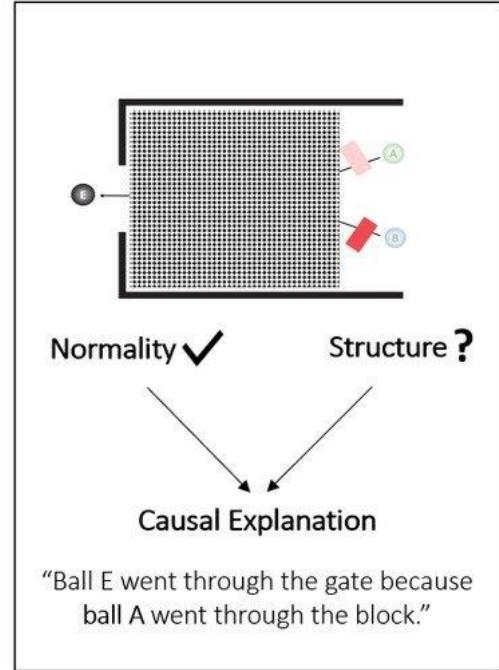
# Normal vs Abnormal Causal Explanations



**Hypothesis 1:** Influence of normality and structure on causal explanations.

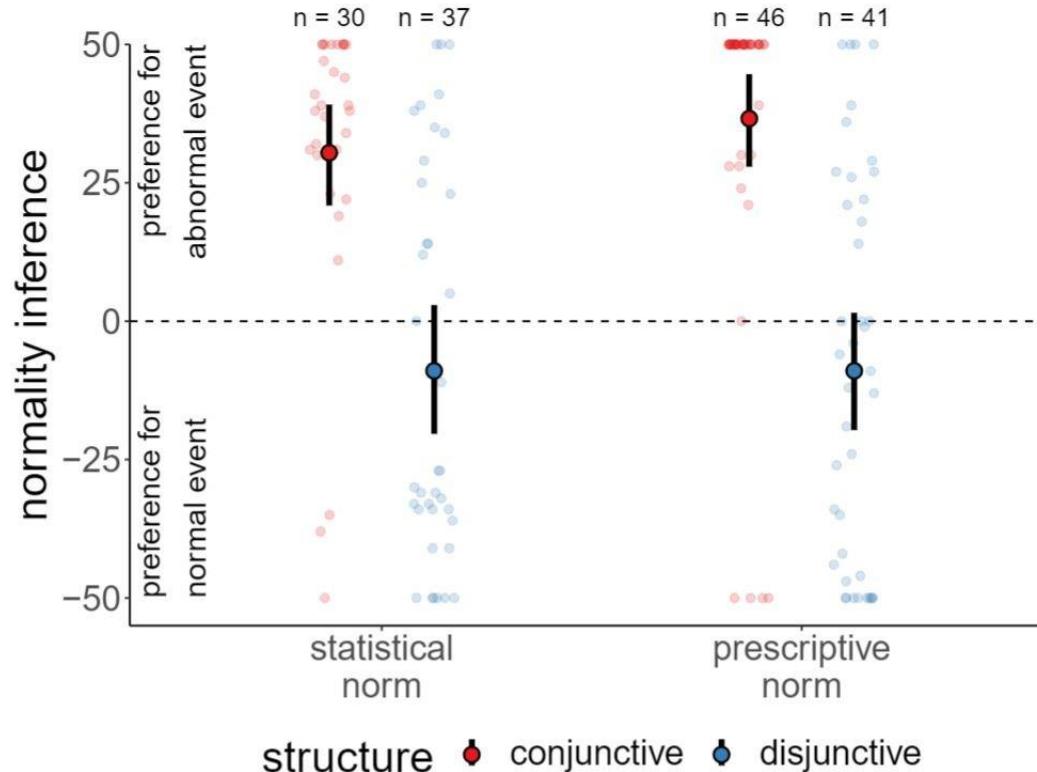


**Hypothesis 2:** Inference about normality from causal explanation and structure.



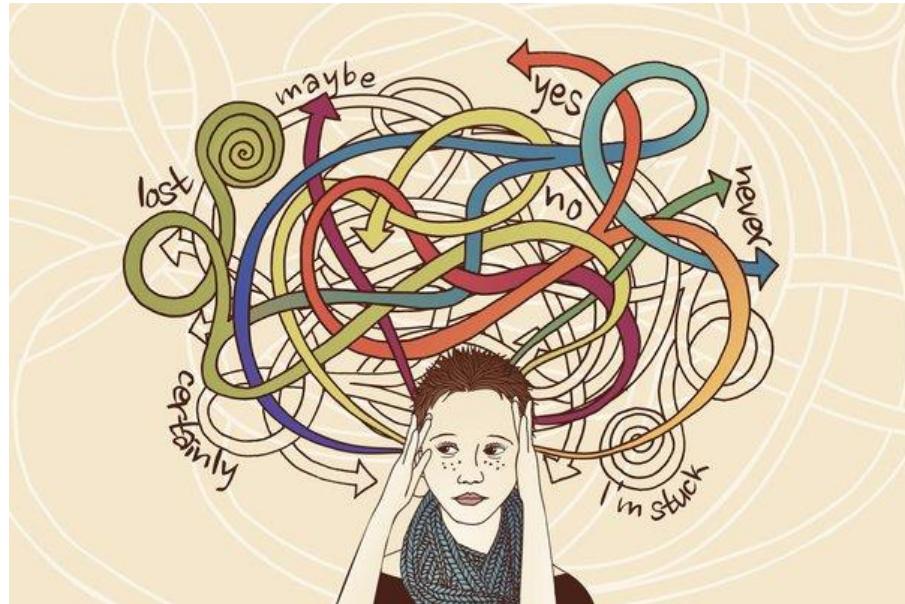
**Hypothesis 3:** Inference about structure from causal explanation and normality.

# Quality of Learned Explanations



# Lots of Open Questions

- Causal explanations over time
- Trustworthy causal explanations (a more formal framework)
- (Real) Causal explanations in deep learning
- Transferring causal explanations
- Benchmarks, benchmarks, benchmarks
- Take ideas from how humans explain



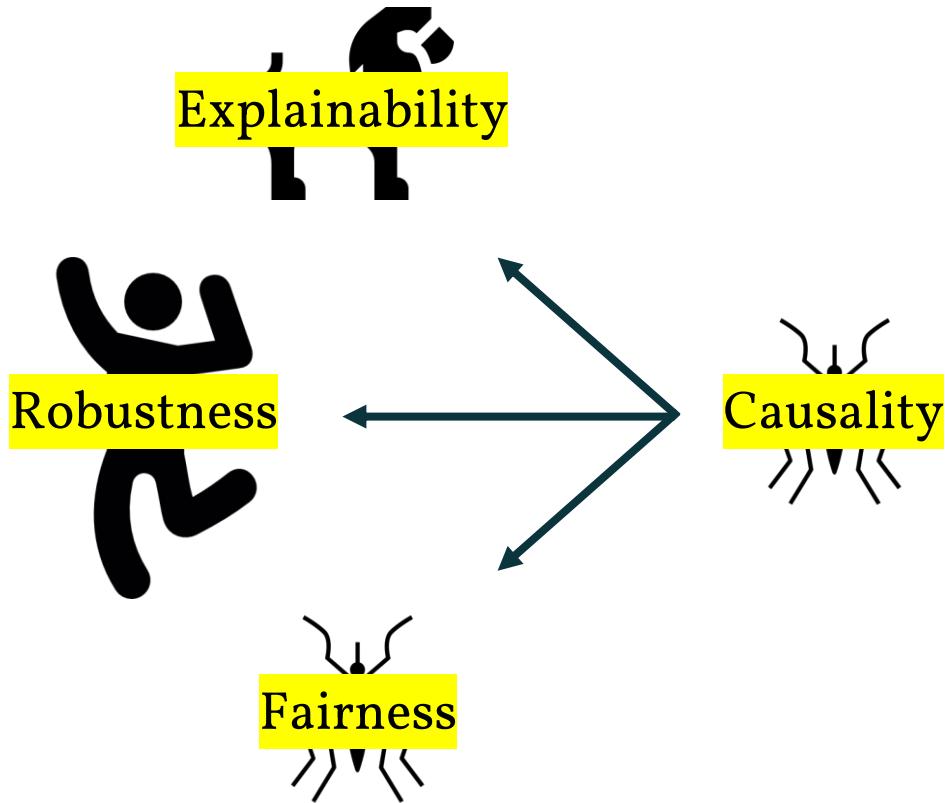
# (Some) References

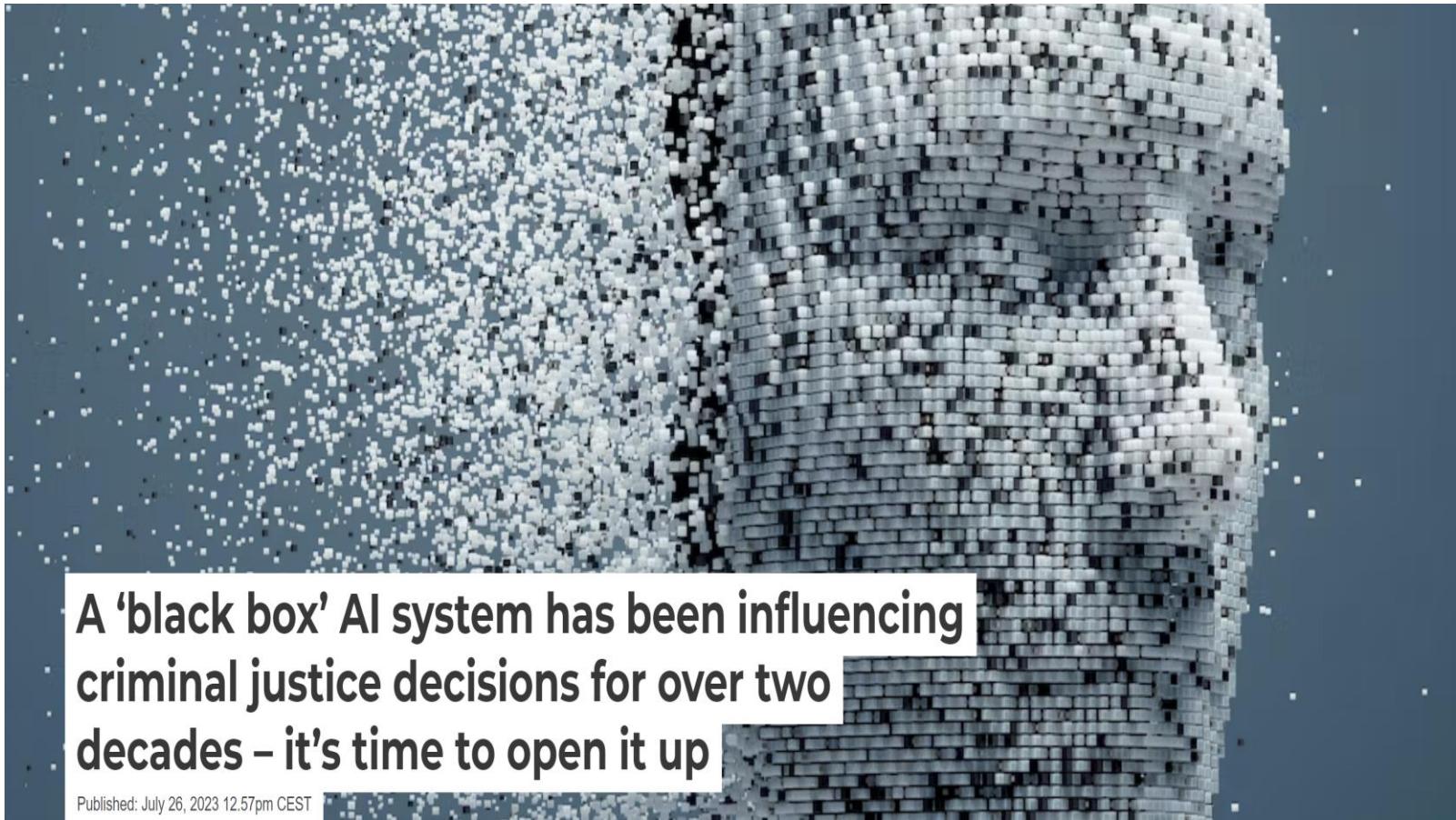
- Beckers, Causal Explanations and XAI, CLeaR 2022
- Moraffah et al., Causal Interpretability for Machine Learning Problems, Methods and Evaluation, KDD 2020
- Baron, Explainable AI and Causal Understanding: Counterfactual Approaches Considered, Minds and Machines 2023
- Grimsley, Causal and Non-Causal Explanations of Artificial Intelligence, 2020
- Faithful, Interpretable Model Explanations via Causal Abstraction  
(<http://ai.stanford.edu/blog/causal-abstraction/>)
- Google ☺

Causal  
Fairness



Causal Machine Learning





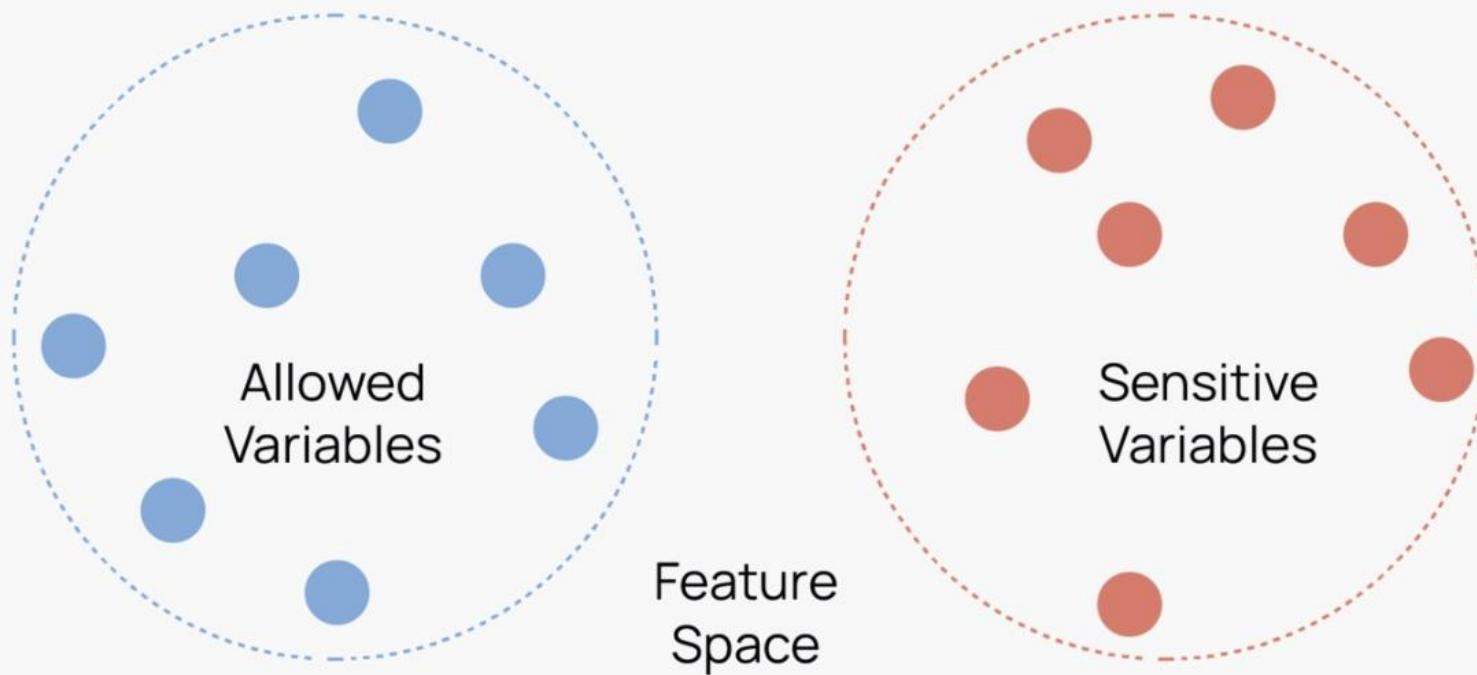
**A 'black box' AI system has been influencing criminal justice decisions for over two decades – it's time to open it up**

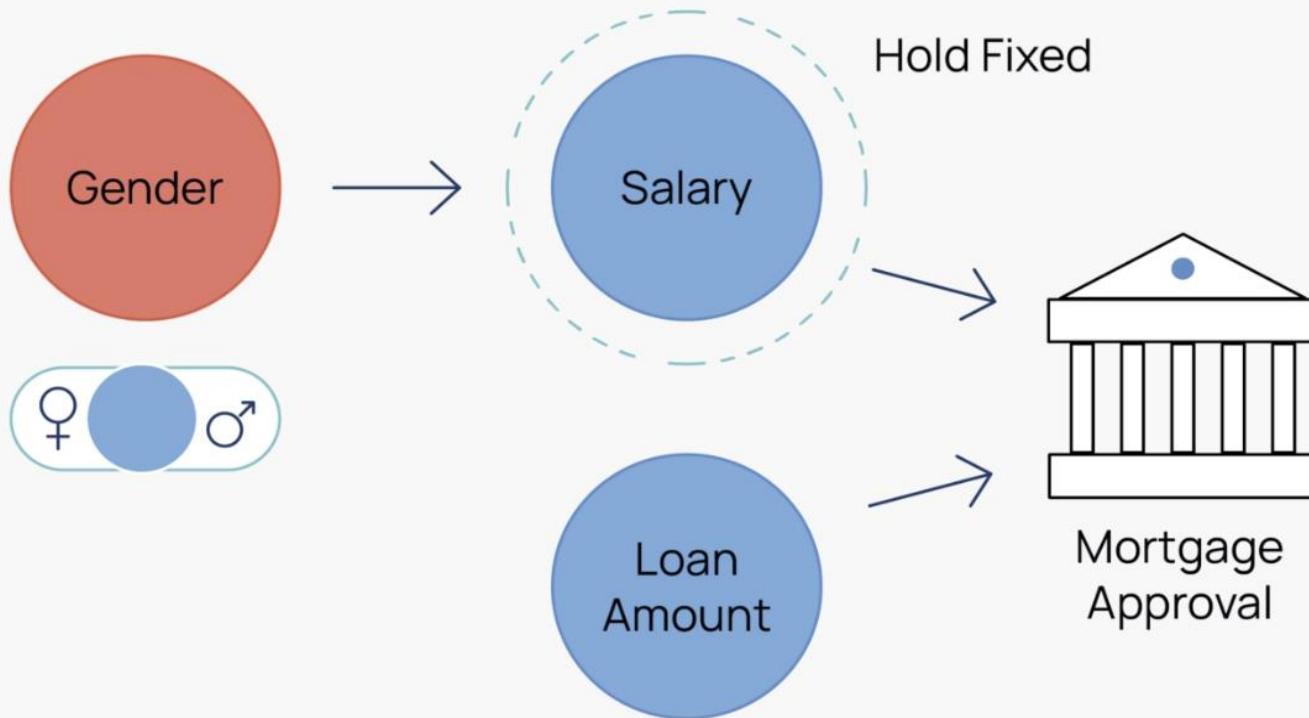
Published: July 26, 2023 12.57pm CEST

<https://theconversation.com/a-black-box-ai-system-has-been-influencing-criminal-justice-decisions-for-over-two-decades-its-time-to-open-it-up-200594>

One of the challenges of tackling AI bias and promoting fairness is formally defining these concepts. Researchers have proposed a vast number of definitions of fairness ([21 by one count](#)). Many of these definitions attempt to define fairness in terms of "parity" (i.e. equal treatment or outcomes) for people from different demographic groups. Troublingly, the different definitions are difficult to verify in the real world and are known to be incompatible with each other.

Causal AI approaches are simpler – roughly speaking, the idea is to check whether sensitive or protected characteristics (like race, gender, religion, and so on) are influencing the algorithm's decisions. If not, then the algorithm is fair.





Notion	Association	SCM	PO	Intervention	Counterfactual	$Y$ and $\hat{Y}$
Total variation	✓					
Total causal fairness		✓		✓		
Natural direct effect		✓		✓		
Natural indirect effect		✓		✓		
Path-specific causal fairness		✓		✓		
Direct causal fairness		✓		✓		
Indirect causal fairness		✓		✓		
Counterfactual fairness		✓			✓	
Counterfactual direct effect		✓			✓	
Counterfactual indirect effect		✓			✓	
Path-specific counterfactual fairness		✓			✓	
Proxy fairness		✓		✓		
Justifiable fairness		✓		✓		
Counterfactual direct error rate		✓			✓	✓
Counterfactual indirect error rate		✓			✓	✓
Individual equalized counterfactual odds		✓			✓	✓
Fair on average causal effect		✓		✓		
Fair on average causal effect on the treated		✓			✓	
Equal effort fairness		✓			✓	

SCM, structure causal model; PO, potential outcome. The last column describes whether the fairness notion involves both  $Y$  and  $\hat{Y}$  in their counterfactual quantity. A checkmark means that the causality-based fairness notion falls within the given category. For example, total causal fairness belongs to both the SCM framework and the intervention rung of Pearl's ladder of causation.

# Counterfactual Fairness

- Fairness Through Unawareness (FTA): An algorithm is fair so long as any protected attributes are not explicitly used in the decision-making process
- Individual Fairness (IF): An algorithm is fair if it gives similar predictions to similar individuals

**Definition 5** (Counterfactual fairness). *Predictor  $\hat{Y}$  is counterfactually fair if under any context  $X = x$  and  $A = a$ ,*

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \quad (1)$$

*for all  $y$  and for any value  $a'$  attainable by  $A$ .*

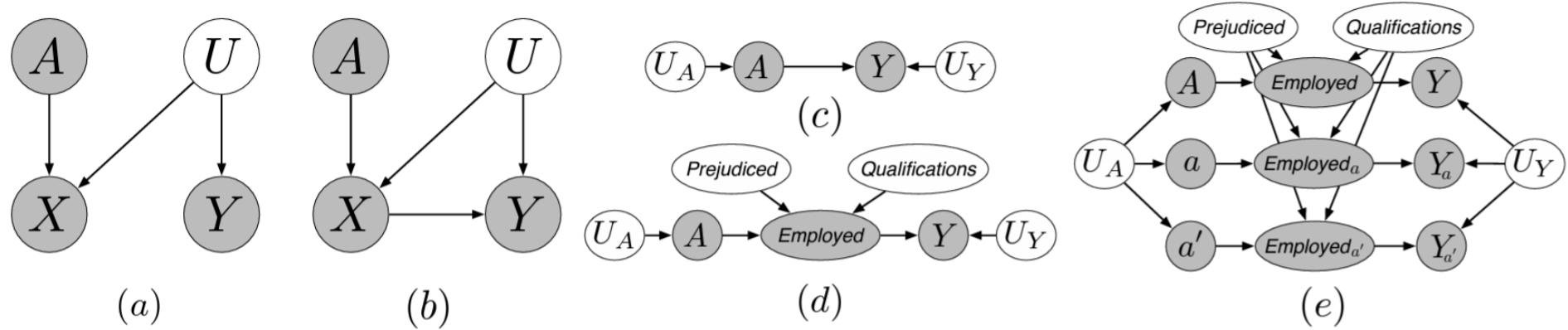
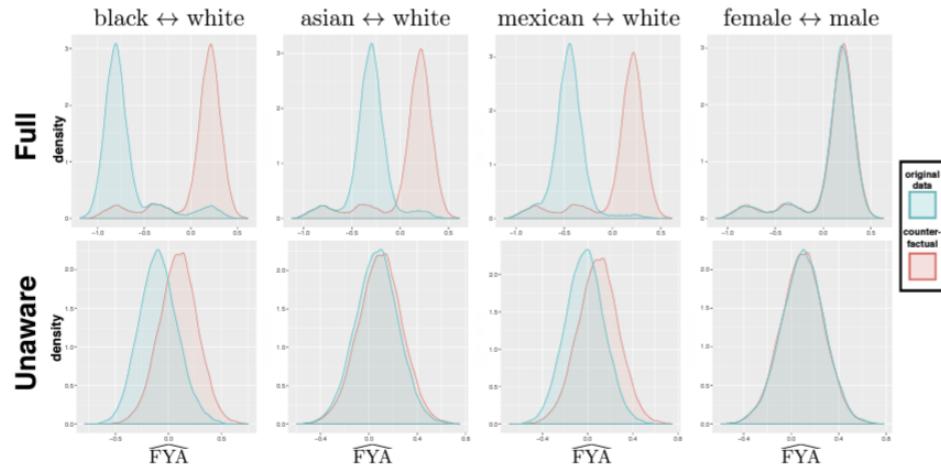
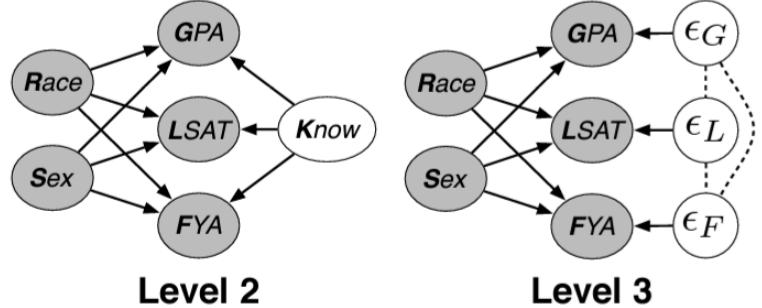


Figure 1: (a), (b) Two causal models for different real-world fair prediction scenarios. See Section 3.1 for discussion. (c) The graph corresponding to a causal model with  $A$  being the protected attribute and  $Y$  some outcome of interest, with background variables assumed to be independent. (d) Expanding the model to include an intermediate variable indicating whether the individual is employed with two (latent) background variables **Prejudiced** (if the person offering the job is prejudiced) and **Qualifications** (a measure of the individual's qualifications). (e) A twin network representation of this system [28] under two different counterfactual levels for  $A$ . This is created by copying nodes descending from  $A$ , which inherit unaffected parents from the factual world.

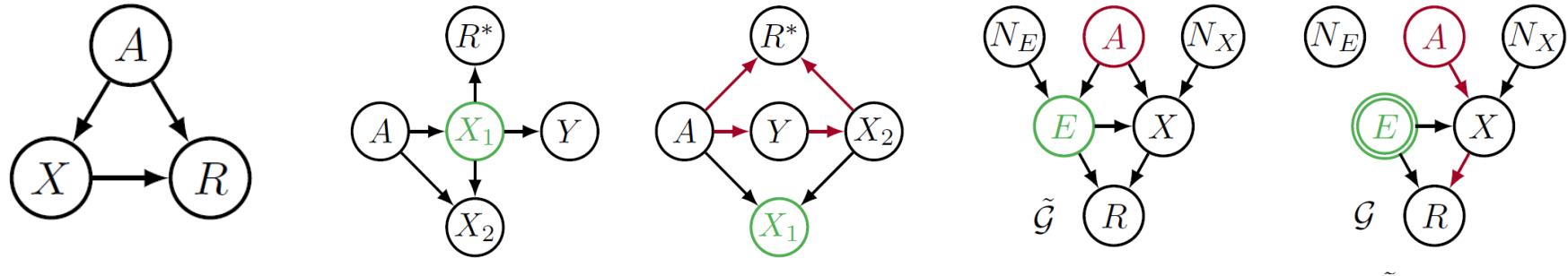
# Measuring Counterfactual Fairness



- Full: all attributes; Unaware: no protected attribute
- Blue distribution is density of predicted attribute for the original data and the red distribution is this density for the counterfactual data
- Models are unfair

# Avoiding Discrimination through Causal Reasoning

- Fairness from data generation perspective i.e. going beyond observational data
- “What do we need to assume about the causal data generating process?”



~

# Causality + Fairness and else

Proceedings of Machine Learning Research vol 140:1–15, 2022

1st Conference on Causal Learning and Reasoning

## Selection, Ignorability and Challenges with Causal Fairness

**Jake Fawkes**

*Department of Statistics, University of Oxford*

JAKE.FAWKES@STATS.OX.AC.UK

**Robin J. Evans**

*Department of Statistics, University of Oxford*

EVANS@STATS.OX.AC.UK

**Dino Sejdinovic**

*Department of Statistics, University of Oxford*

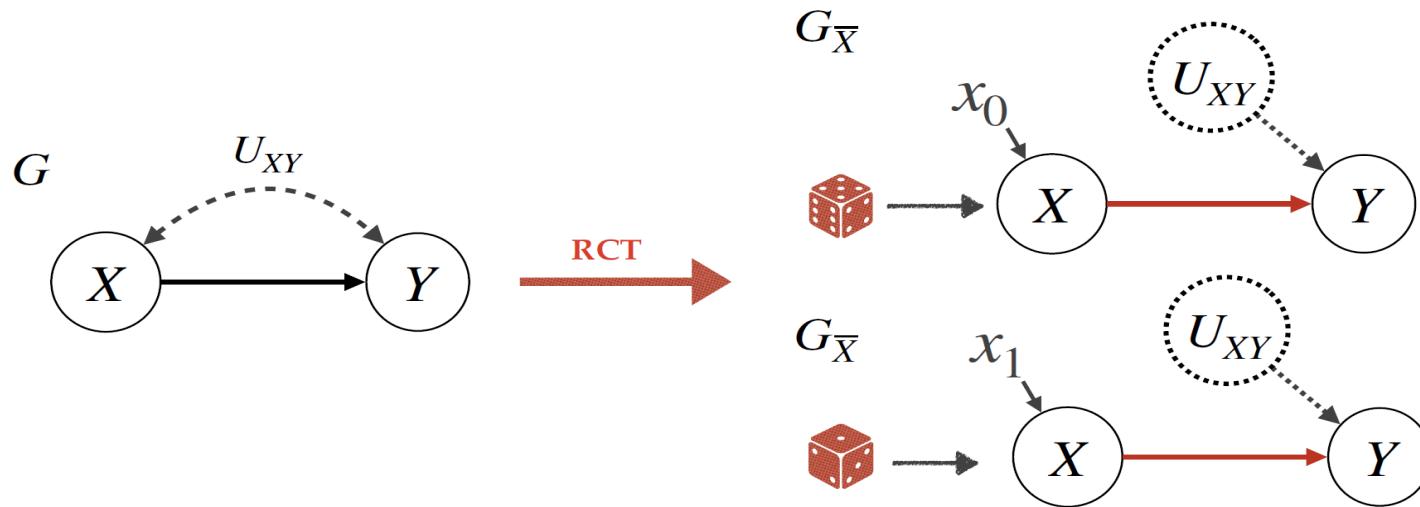
DINO.SEJDINOVIC@STATS.OX.AC.UK

Ignorability is a feature of an experiment design whereby the method of data collection does not depend on the missing data

# Remember? Randomized Experiments

## Recap: Randomized Experiments

Randomized Experiments / Control Trials (e.g. RCT) allow the identification of causal effects by leveraging randomization of the treatment assignment



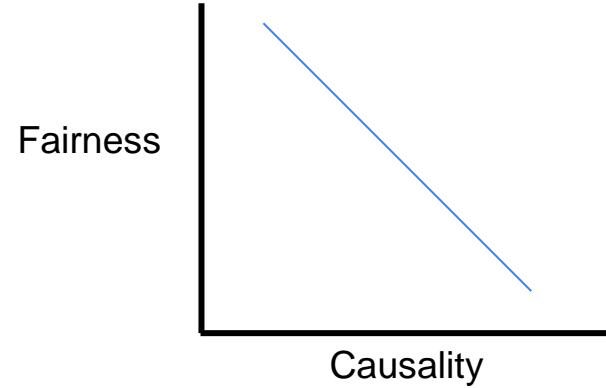
# Causality + Fairness and else

---

## Causal Conceptions of Fairness and their Consequences

---

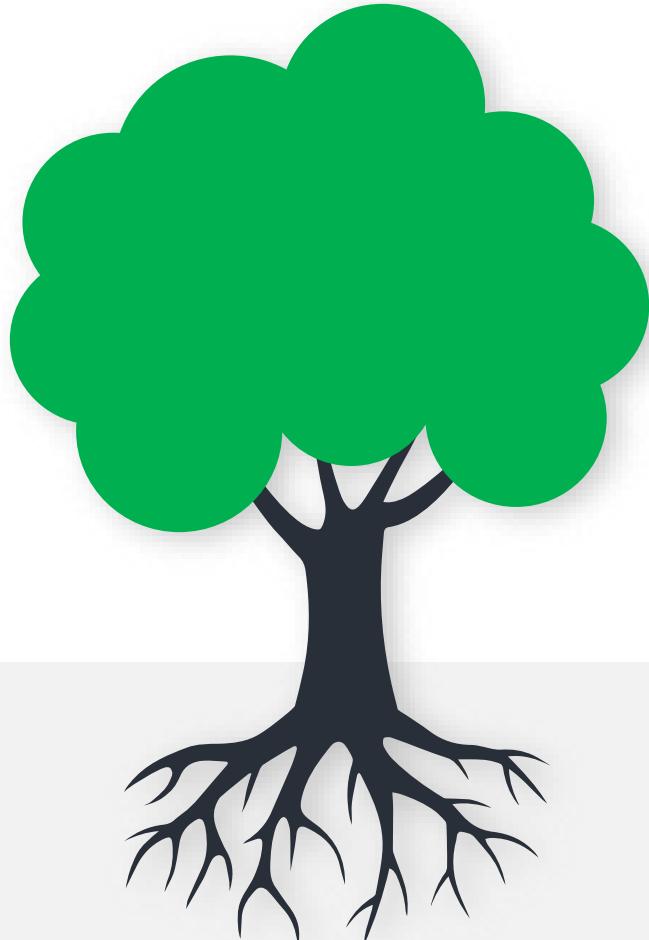
Hamed Nilforoshan<sup>\*1</sup> Johann Gaebler<sup>\*1</sup> Ravi Shroff<sup>2</sup> Sharad Goel<sup>3</sup>



# (Some) References

- Beckers et al., A Causal Analysis of Harm, NeurIPS 2022
- Carey and Wu, The Causal Fairness Field Guide: Perspectives From Social and Formal Sciences, Frontiers in Big Data 2022
- Su et al., A review of causality-based fairness machine learning, Intelligence and Robotics 2022
- Xu et al., Achieving Causal Fairness through Generative Adversarial Networks, IJCAI 2019
- <https://fairness.causalai.net/>
- Google ☺

**Causality  
Overdrive**



**Causal Machine Learning**



Judea Pearl ✅

@yudapearl

...

1/ This paper deserves attention for deriving new theoretical connections between GNNs and SCMs. Moreover, it allows causal information to be provided the natural way, i.e., via DAGs, then translated to GNN. What is not clear to this reader, though it may be implicit in the text,



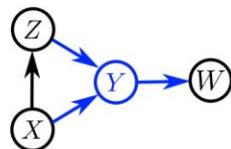
Mehmet Süzen @memosisland · Sep 11, 2021

#graph #deeplearning meets #causality via SCM  
[arxiv.org/abs/2109.04173](https://arxiv.org/abs/2109.04173)  
fyi @yudapearl @mmbronstein @eliasbareinboim

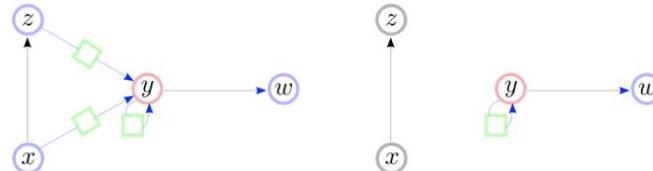
# Interventions in Graph Neural Networks Lead to New Neural Causal Models

- Inspiration
  - both SCMs and GNN share the following property: a graph
- GNN places an inductive bias on the structural relation of the input (a graph)
- Allows for specialized computation
- We establish a new model class, partial causal models (PCM) for GNN-based causal inference via interventions

Graph  $G$  of SCM  $\mathcal{M}$   $\xrightarrow{do(y)}$  Graph  $G^\dagger$  of intervened SCM  $\mathcal{M}^{\dagger} := \mathcal{M}_{do(y)}$



GNN corresponding to  $G$   $\xrightarrow{do(y)}$  GNN corresponding to  $G^\dagger$



# *Intervention* akin to SCM within the GNN computation rule

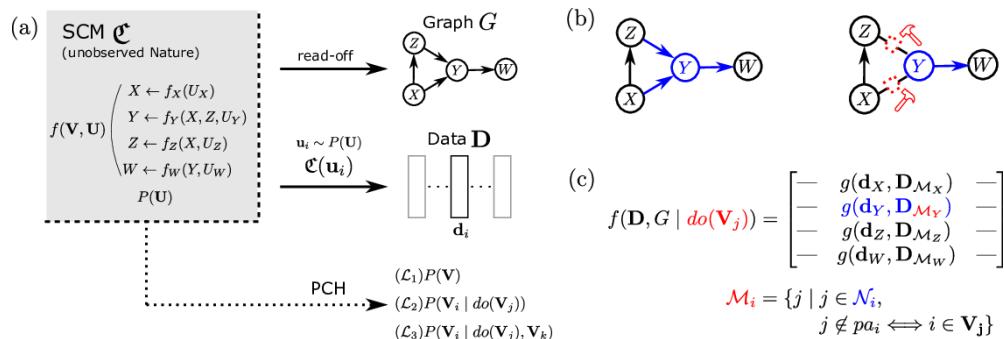
**Definition 1.** (Interventions within GNN.) An intervention  $\mathbf{x}$  on the corresponding set of variables  $\mathbf{X} \subseteq \mathbf{V}$  within a GNN layer  $f(\mathbf{D}, \mathbf{A}_G)$ , denoted by  $f(\mathbf{D}, \mathbf{A}_G | do(\mathbf{X} = \mathbf{x}))$ , is defined as a modified layer computation,

$$\mathbf{h}_i = \phi\left(\mathbf{d}_i, \bigoplus_{j \in \mathcal{M}_i^G} \psi(\mathbf{d}_i, \mathbf{d}_j)\right),$$

where the intervened local neighborhood is given by

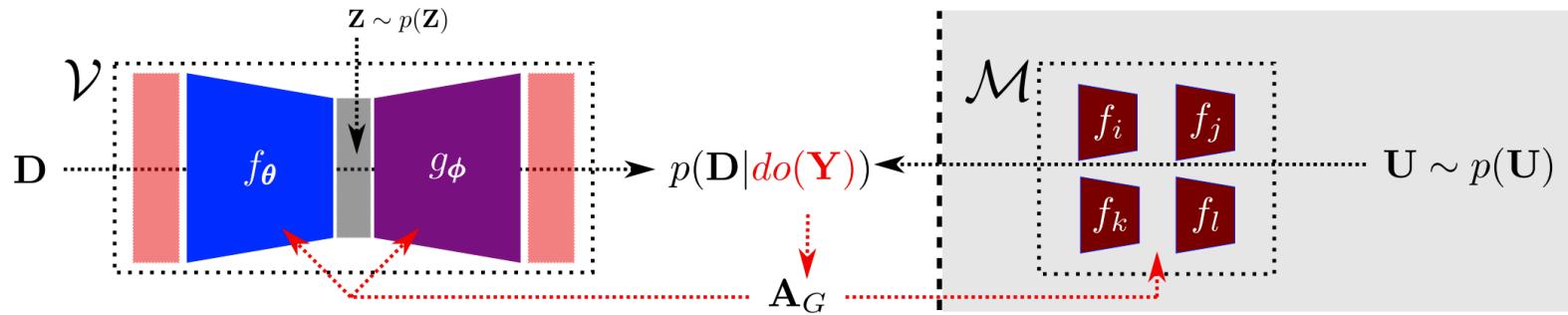
$$\mathcal{M}_i^G = \{j \mid j \in \mathcal{N}_i^G, j \notin pa_i \iff i \in \mathbf{X}\}$$

where  $\mathcal{N}^G$  denotes the regular graph neighborhood. A GNN layer that computes Eq.1 is said to be *interventional*.



- No discretion between different interventions on same variable
- Considers only structural properties
- This interventional computation layer can be embedded into any system
- We show it via Variational Graph Autoencoder and call it iVGAE

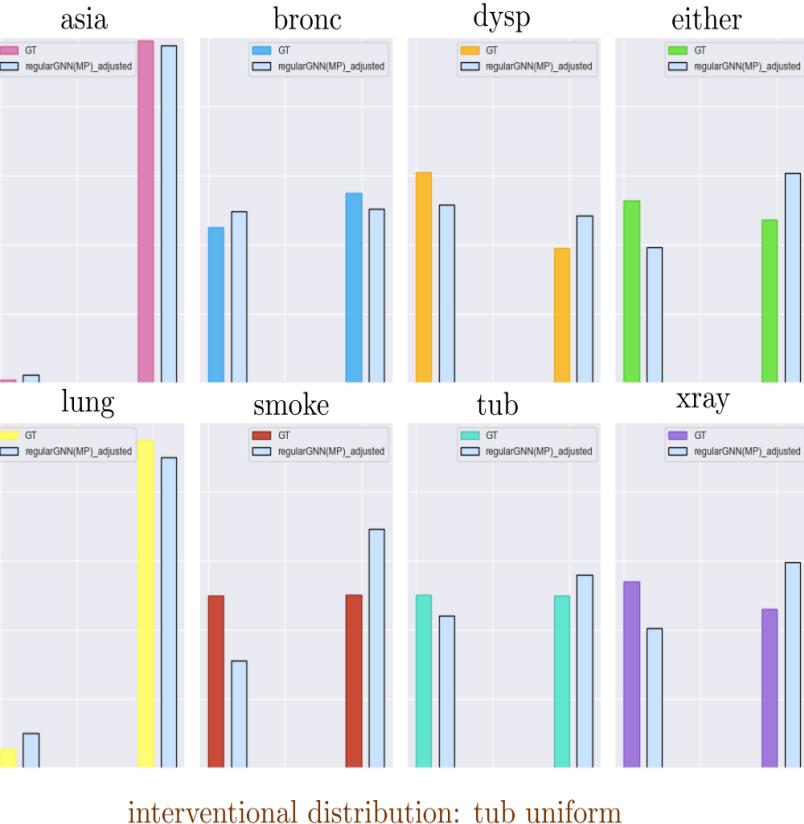
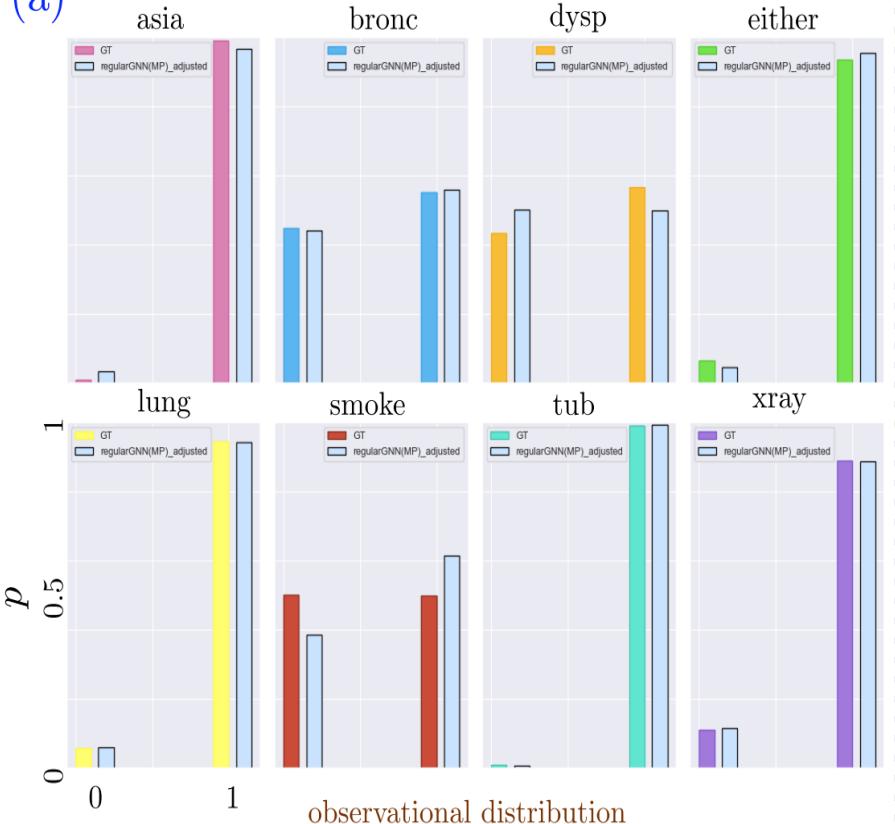
# Interventional Variational Graph Autoencoders



**Definition 4 (iVGAE).** Let  $\mathcal{V}$  be a VGAE with encoder  $q$  and decoder  $p$  being GNNs. If  $(q,p)$  are set to be interventional GNN layers (Def.1) modelling the latent variables and endogenous variables (data) respectively, then  $\mathcal{V}$  is also called interventional VGAE.

# Does iVGAe capture Interventions?

(a)





Judea Pearl

@yudapearl

...

My list of recent articles on causal inference:

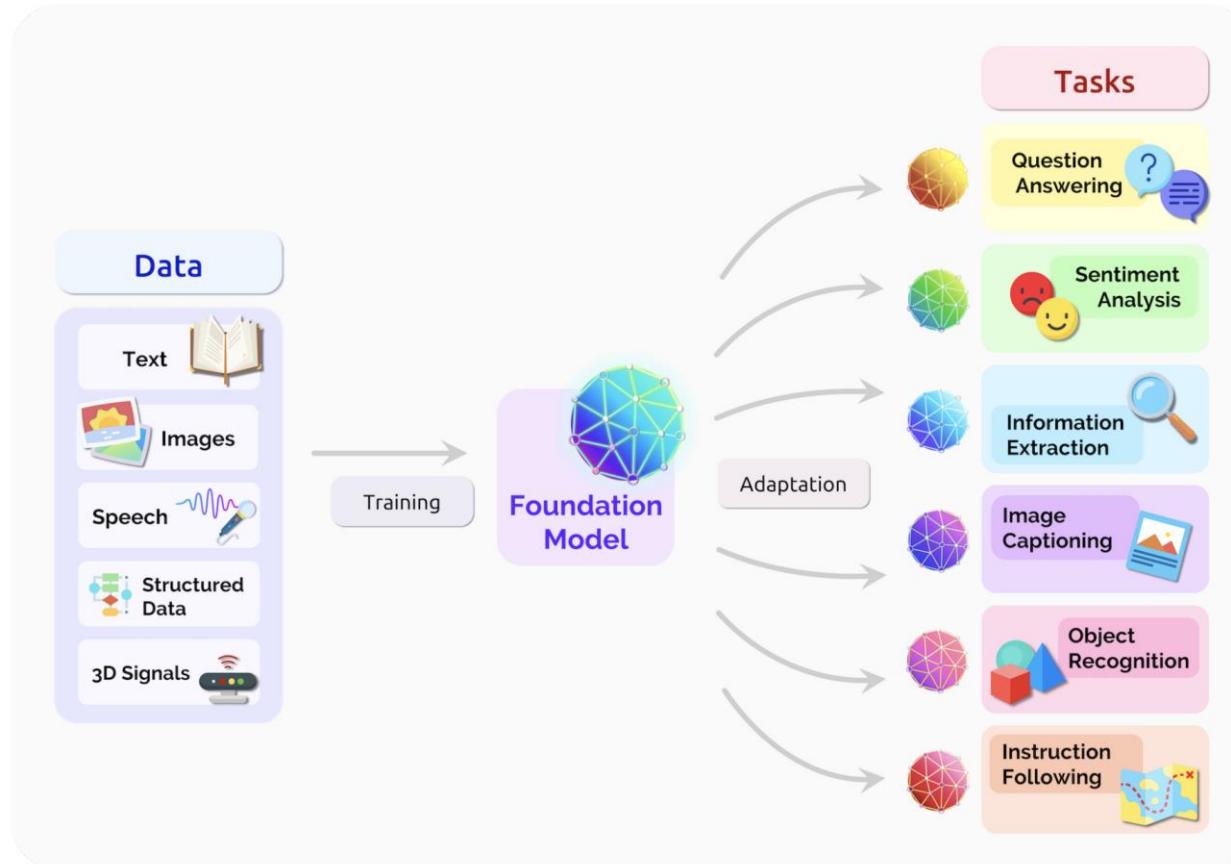
[ucla.in/39WglUm](https://ucla.in/39WglUm)

One paper catching my attention is "Can Foundation Models Talk Causality" [arxiv.org/pdf/2206.10591...](https://arxiv.org/pdf/2206.10591.pdf)

a topic discussed on our platform which made me wonder: How can one ask "Can X do Y?" when X is undefined?

8:39 AM · Jun 30, 2022

# Why Large Language Models?



# Related Work in Causality

---

## Can Large Language Models Infer Causation from Correlation?

---

Zhijing Jin<sup>1,2,\*</sup> Jiarui Liu<sup>3</sup> Zhiheng Lyu<sup>4</sup> Spencer Poff<sup>5</sup>  
Mrinmaya Sachan<sup>2</sup> Rada Mihalcea<sup>3</sup> Mona Diab<sup>5,†</sup> Bernhard Schölkopf<sup>1,†</sup>  
<sup>1</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany, <sup>2</sup>ETH Zürich,  
<sup>3</sup>University of Michigan, <sup>4</sup>University of Hong Kong, <sup>5</sup>Meta AI

released on arXiv *last month*

## Related Work in Causality

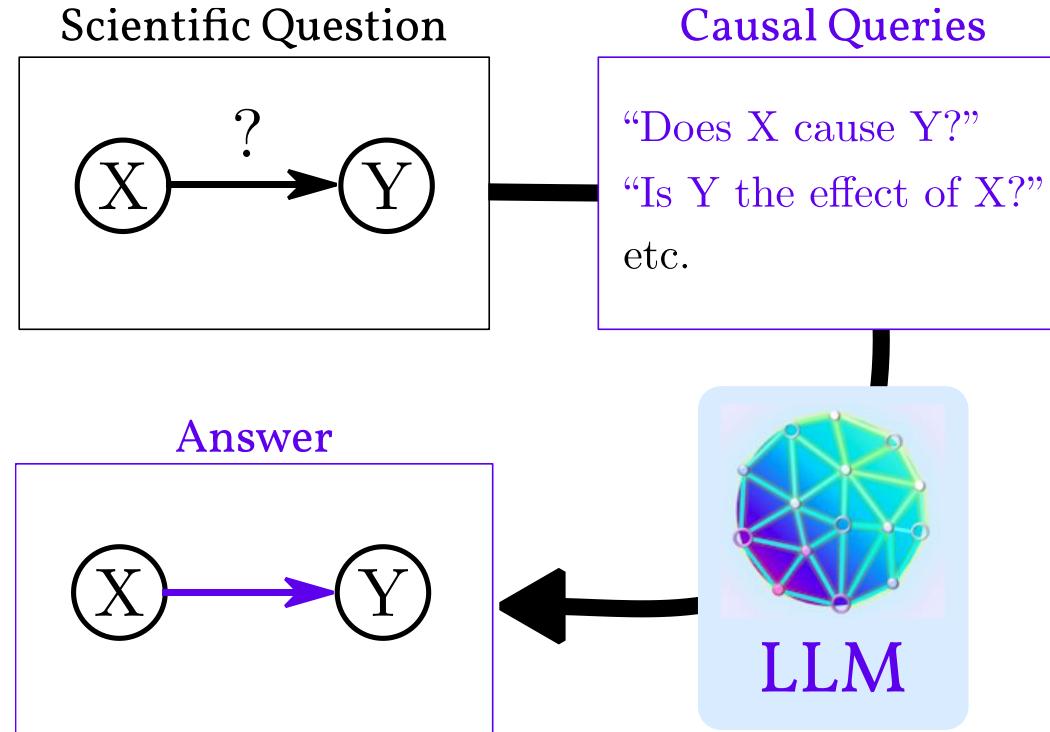
“[...] We evaluated an extensive list of LLMs on this new task, and showed that off-the-shelf LLMs perform poorly on this task.”

Tentative conclusion:

The literature seems to agree that LLMs  
are “castles in the air.” \*

\*except if the folks from the literature work for a company  
that builds/buys LLMs themselves

# An Initial, Naïve Inference Approach



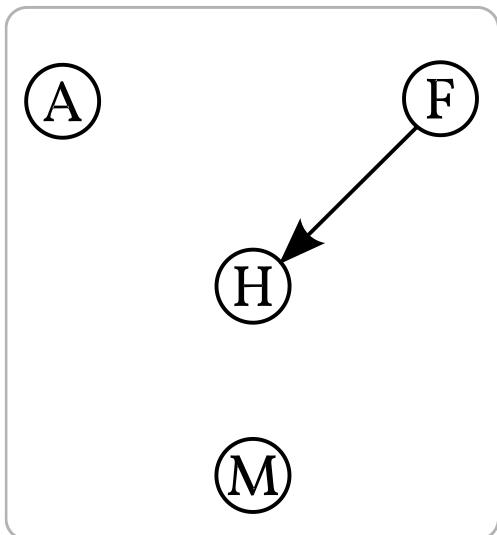
# Overall Poor. But if good, then surprisingly!

	Metric	Altitude	Health	Driving	Recovery	Cancer	Earthquake	LLM
Causal Graph	SID ↓	<b>0.80±0.40</b>	<b>7.20±0.75</b>	<b>3.00±0.89</b>	<b>4.00±1.79</b>	<b>11.80±4.66</b>	<b>11.40±1.50</b>	GPT-3
		1.20±0.98	10.60±1.85	6.00±0.00	5.40±1.20	<b>11.40±3.07</b>	16.00±3.63	Luminous
		1.60±0.80	10.80±2.40	5.00±1.26	5.80±0.40	16.80±1.94	15.60±5.95	OPT
ML	SHD ↓	0.80±0.40	<b>4.00±0.63</b>	<b>2.60±0.49</b>	<b>2.20±0.40</b>	<b>7.00±1.41</b>	<b>4.60±0.80</b>	GPT-3
		<b>0.60±0.49</b>	7.00±1.10	4.20±0.40	3.40±0.80	10.00±3.52	5.60±1.62	Luminous
		0.80±0.40	<b>7.40±1.20</b>	<b>3.40±1.20</b>	<b>4.00±0.00</b>	<b>13.20±1.60</b>	<b>8.60±3.01</b>	OPT
Edges	$F_1$ Score ↑	0.20±0.40	<b>0.47±0.14</b>	0.11±0.23	0.27±0.33	0.35±0.11	0.12±0.15	GPT-3
		<b>0.80±0.16</b>	<b>0.41±0.21</b>	<b>0.46±0.09</b>	<b>0.55±0.07</b>	<b>0.40±0.13</b>	<b>0.40±0.04</b>	Luminous
		0.73±0.13	<b>0.52±0.05</b>	<b>0.53±0.15</b>	<b>0.47±0.07</b>	<b>0.35±0.03</b>	<b>0.47±0.07</b>	OPT
Edges	Sparsity	0.90±0.20	<b>0.63±0.28</b>	0.77±0.31	0.70±0.31	0.65±0.16	0.93±0.07	GPT-3
		0.20±0.24	<b>0.22±0.35</b>	0.03±0.07	0.10±0.13	0.40±0.16	0.74±0.12	Luminous
		0.10±0.20	<b>0.05±0.10</b>	0.17±0.21	<b>0.07±0.13</b>	0.18±0.12	0.41±0.18	OPT
Edges	ADS ↑	0.50	<b>0.62</b>	<b>0.33</b>	<b>0.50</b>	<b>0.69</b>	0.00	GPT-3
		<b>1.00</b>	0.53	0.17	0.17	0.38	0.26	Luminous
		0.50	0.25	0.25	0.33	0.28	<b>0.47</b>	OPT

Table 2. Comparing LLMs prediction to existing ground truth causal structures. The metrics concerned with the causal graph structure (SID, SHD) reveal a closer match of GPT-3 predictions to the ground truth causal structures than for the other LLMs. High  $F_1$  Scores and low sparsity indicate densely connected graph prediction by Luminous and OPT. This can be desired for ML applications. The ADS reveals that all LLMs increase their decisiveness on edge directions when querying with asymmetric sentence templates.

# Remarkable Observation I: Query Wording Sensitivity

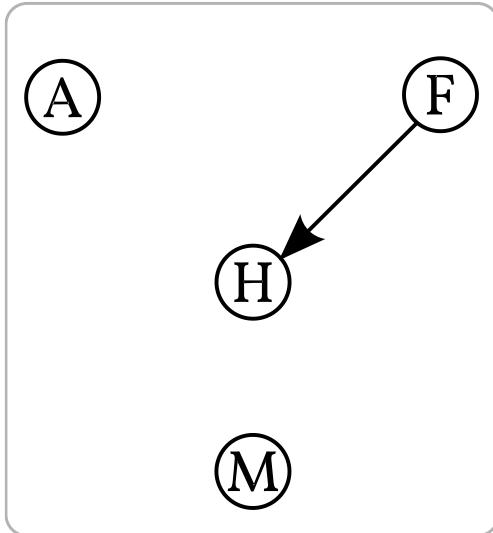
“Does  $X$  cause  $Y$ ? ”



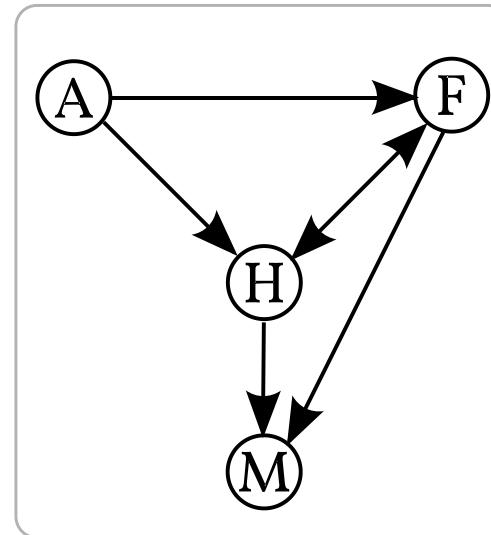
Legend: [A]ge    [N]utrition    [H]ealth    [M]obility

# Remarkable Observation I: Query Wording Sensitivity

“Does  $X$  cause  $Y$ ? ”

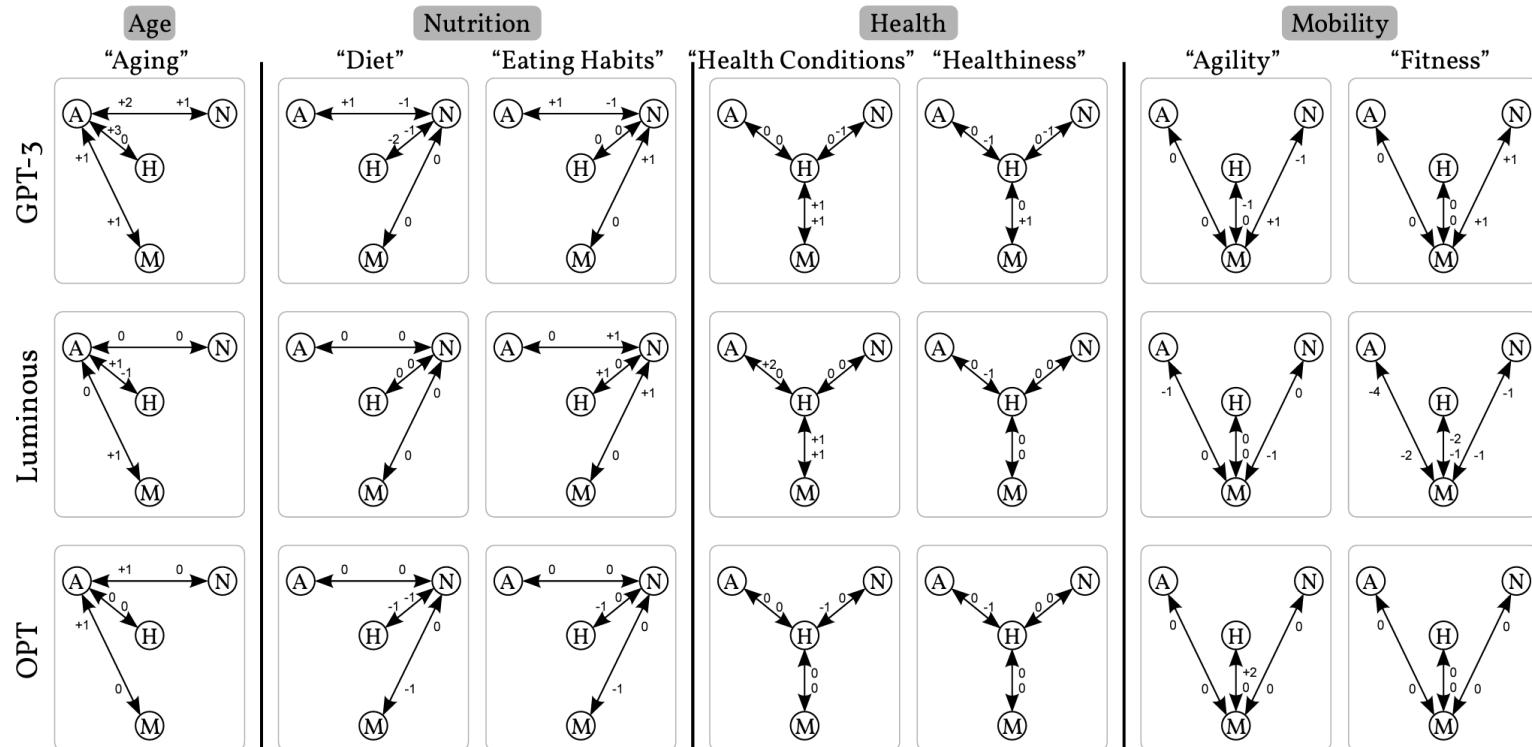


“Is there a causality  
between  $X$  and  $Y$ ? ”



Legend: [A]ge    [N]utrition    [H]ealth    [M]obility

# Remarkable Observation 2: Variable Naming Sensitivity



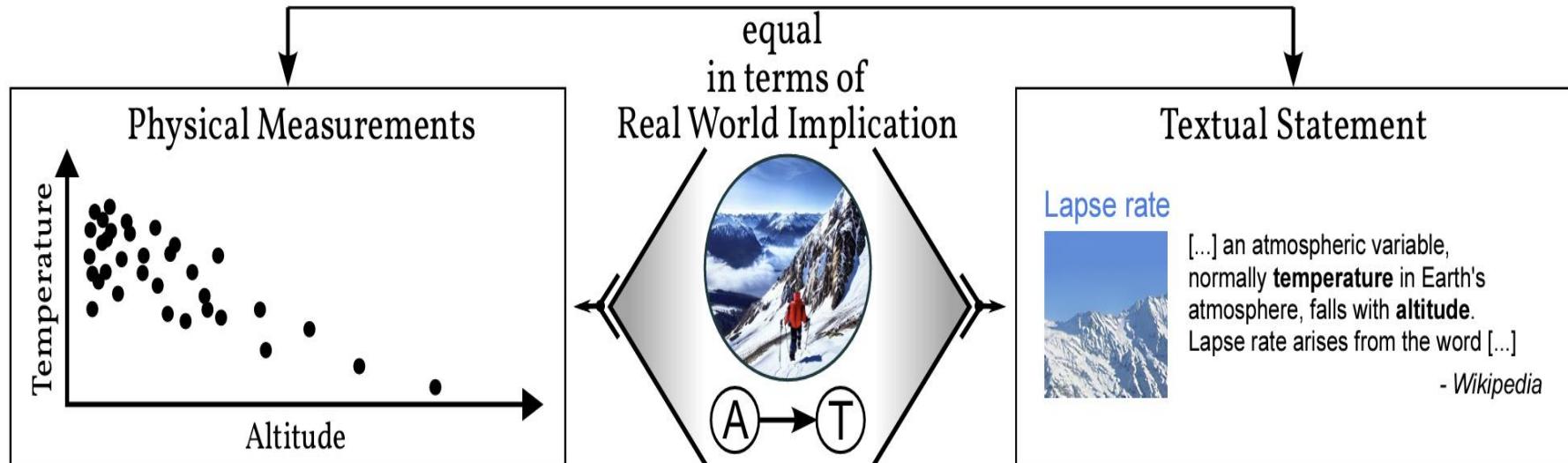
“If A causes B and B causes C. Does A cause C?”

GPT-4	Yes, in a sense, A does cause C indirectly through B. This is known as a causal chain. However, the relationship may not be as strong or direct as the relationship between A and B or B and C. Other factors could also influence <i>[end of output]</i>
GPT-3	It is possible that A causes C, but it is also possible that A and C are unrelated.
Luminous	A: The answer is no. <i>[continued]</i>
OPT	If A causes B and B causes C, then A causes C. If A causes B and B causes C, then A does not cause C. <i>[continued]</i>

“If A causes B and B causes C. Does A cause B?”

GPT-4	Yes, as per the initial statement, A causes B.
GPT-3	There is not enough information to answer this question.
Luminous	A: The answer is no. <i>[continued]</i>
OPT	If A causes B and B causes C, then A causes C. If A causes B and B does not cause C, then A does not cause C. <i>[repeating]</i>

# Ultimately, a Philosophical Question



‘understanding’

‘knowing’

# Alternative Analogy: Plato's Allegory of the Cave



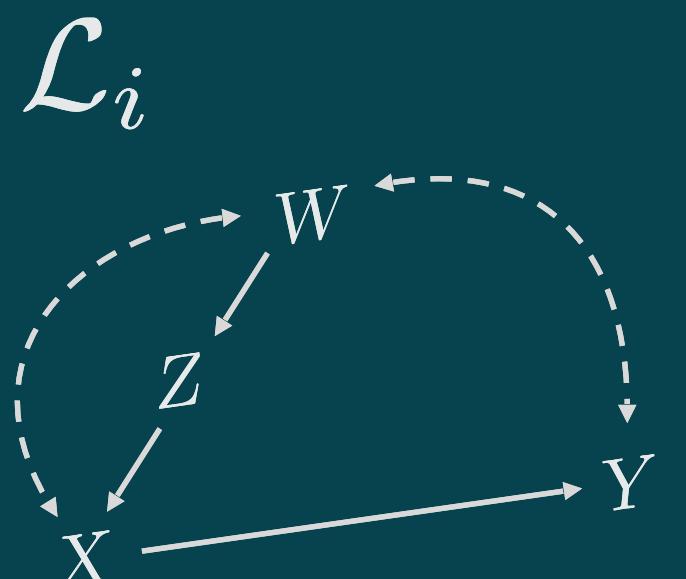
correct answers by LLM

“correct causal reality”



# Causal Discovery

Current State of the Art



# Causal Discovery - Current State of Art

- Learning from a combination of observational and experimental data
- Learning dynamic causal models from time-series data
- Extension to cyclic causal models

# Learning from Observational and Experimental Data

## **Under causal sufficiency:**

- Gonçalo Rui Alves Faria, Andre Martins, Mario A. T. Figueiredo. Differentiable Causal Discovery Under Latent Interventions. Proceedings of the First Conference on Causal Learning and Reasoning, PMLR 177:253-274, 2022.

## **Allowing unmeasured confounders:**

- Jaber, A., Kocaoglu, M., Shanmugam, K., Bareinboim, E. Causal Discovery from Soft Interventions with Unknown Targets: Characterization & Learning. In Advances in Neural Information Processing Systems 2020 ([Link](#)).
- Kocaoglu, M., Jaber, A., Shanmugam, K., Bareinboim, E. Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. 2019 ([Link](#)).

# Learning Dynamic Causal Models from Time-Series Data

## **SVAR-FCI – Malinsky, D. and Spirtes, P. (2018).**

- Assume the data-generating process is a structural vector autoregression (SVAR) with latent components.
- Allow for both “contemporaneous” causal relations and arbitrary unmeasured (“latent”) processes influencing observed variables
- Algorithm is a modification of the FCI algorithm to search for an equivalence class of Dynamic Ancestral Graphs.
  1. respects the time order of the variables by restricting possible conditioning sets to variables in the “present” or “past” time slices and prohibiting orientations backwards in time.
  2. enforces the repeating structure of the underlying dynamic DAG in both determining adjacencies and orientations.

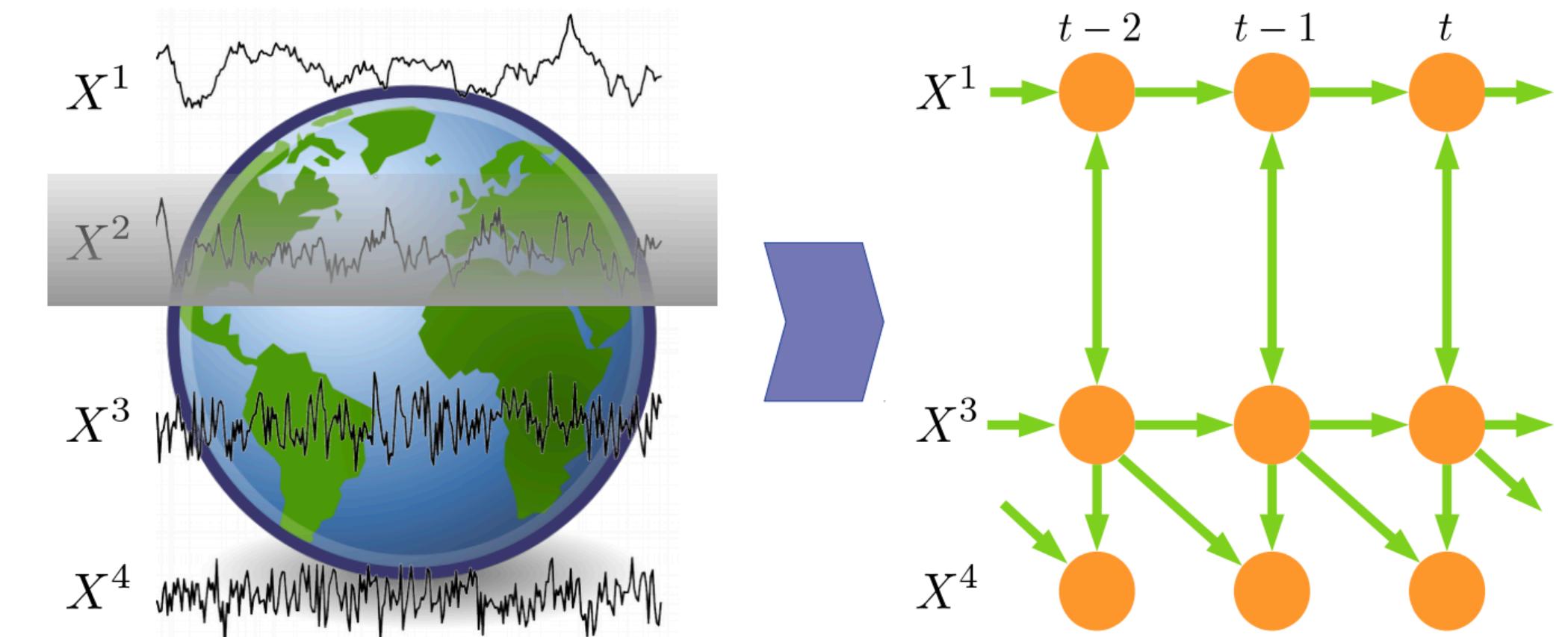
Malinsky, D. and Spirtes, P. (2018). *Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding*. Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery, volume 92 of Proceedings of Machine Learning Research, pages 23–47, London, UK. PMLR.

# Learning Dynamic Causal Models from Time-Series Data

**Nice Overview:** Runge, J., Gerhardus, A., Varando, G. et al. Causal inference for time series. Nat Rev Earth Environ (2023). ([Link](#))

## Under causal sufficiency:

- **PCMCI:** J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, D. Sejdinovic, Detecting and quantifying causal associations in large nonlinear time series datasets. Sci. Adv. 5, eaau4996 (2019). ([Link](#))
- **PCMCI+:** J. Runge (2020): Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020, Toronto, Canada, 2019, AUAI Press, 2020. ([Link](#))

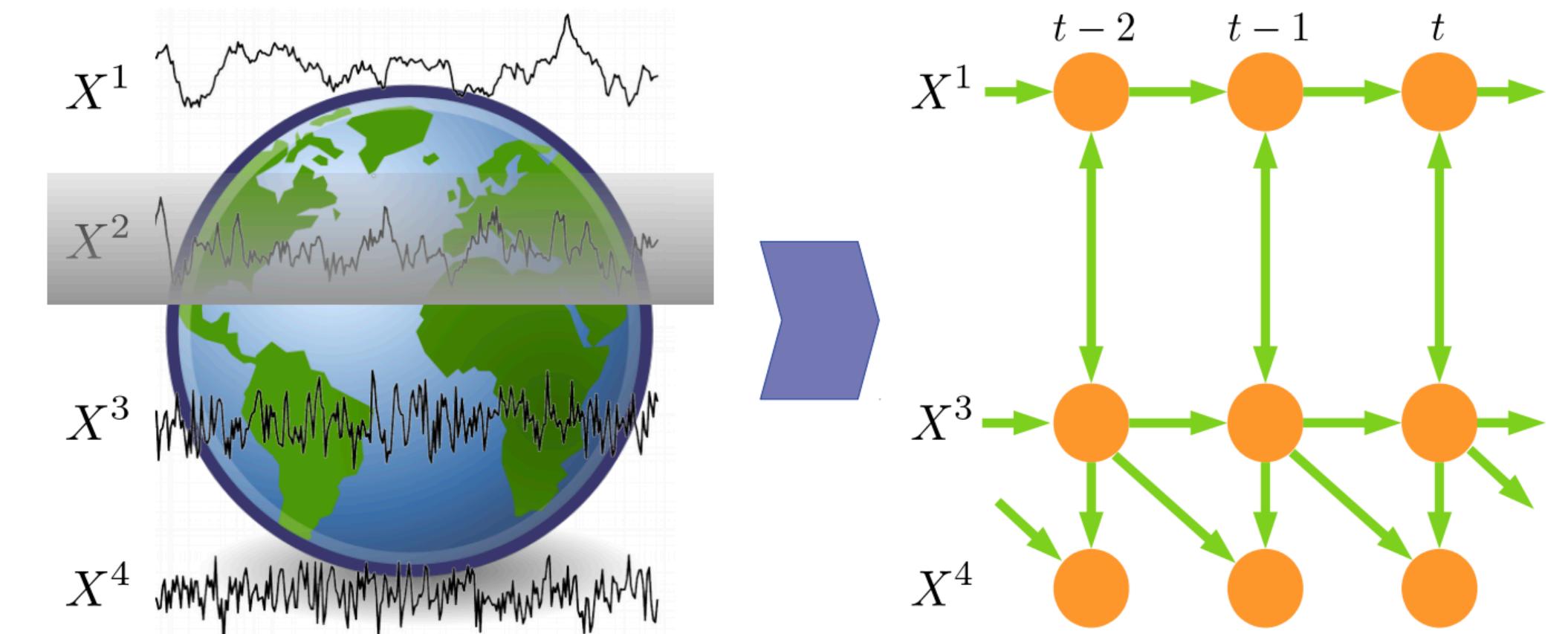


# Learning Dynamic Causal Models from Time-Series Data

**Nice Overview:** Runge, J., Gerhardus, A., Varando, G. et al. Causal inference for time series. Nat Rev Earth Environ (2023). ([Link](#))

## Allowing unmeasured confounders:

- **Latent PCMCI (L-PCMCI):** Gerhardus, A. & Runge, J. High-recall causal discovery for autocorrelated time series with latent confounders Advances in Neural Information Processing Systems, 2020, 33. ([Link](#))



# Learning Dynamic Causal Models from Time-Series Data

Method	Assumptions	Output
	(in addition to Causal Markov Condition and Faithfulness)	
PCMCI	Causal stationarity, no contemporaneous causal links, no hidden variables	Directed lagged links, undirected contemporaneous links (for tau_min=0)
PCMCIplus	Causal stationarity, no hidden variables	Directed lagged links, directed and undirected contemp. links (Time series CPDAG)
LPCMCI	Causal stationarity	Time series PAG

<https://github.com/jakobrunge/tigramite>

# Extension to Cyclic Causal Models

Surprisingly, the output of the Fast Causal Inference (FCI) algorithm is correct if it is applied to observational data generated by a system that involves feedback.

Joris M. Mooij, Tom Claassen. *Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles*.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR 124:1159-1168, 2020.

# Extension to Cyclic Causal Models

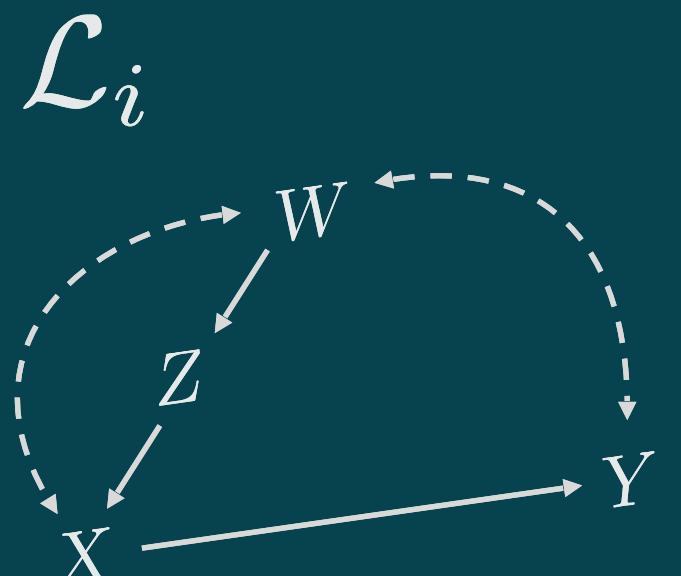
- Surprisingly, the output of the **Fast Causal Inference (FCI) algorithm** is correct if it is applied to observational data generated by a system that involves feedback.
- In other words, it learns a more general class of directed mixed graphs (DMGs), where feedback loops may be present.
- The work focuses on **Simple SCMs**, which admit (sufficiently weak) cyclic interactions but retain many of the convenient properties of acyclic SCMs. In particular, they satisfy the  $\sigma$ -separation Markov property and there is no selection bias.

Joris M. Mooij, Tom Claassen. *Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles*. Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR 124:1159-1168, 2020.



# Causal Identification

## Current State of the Art

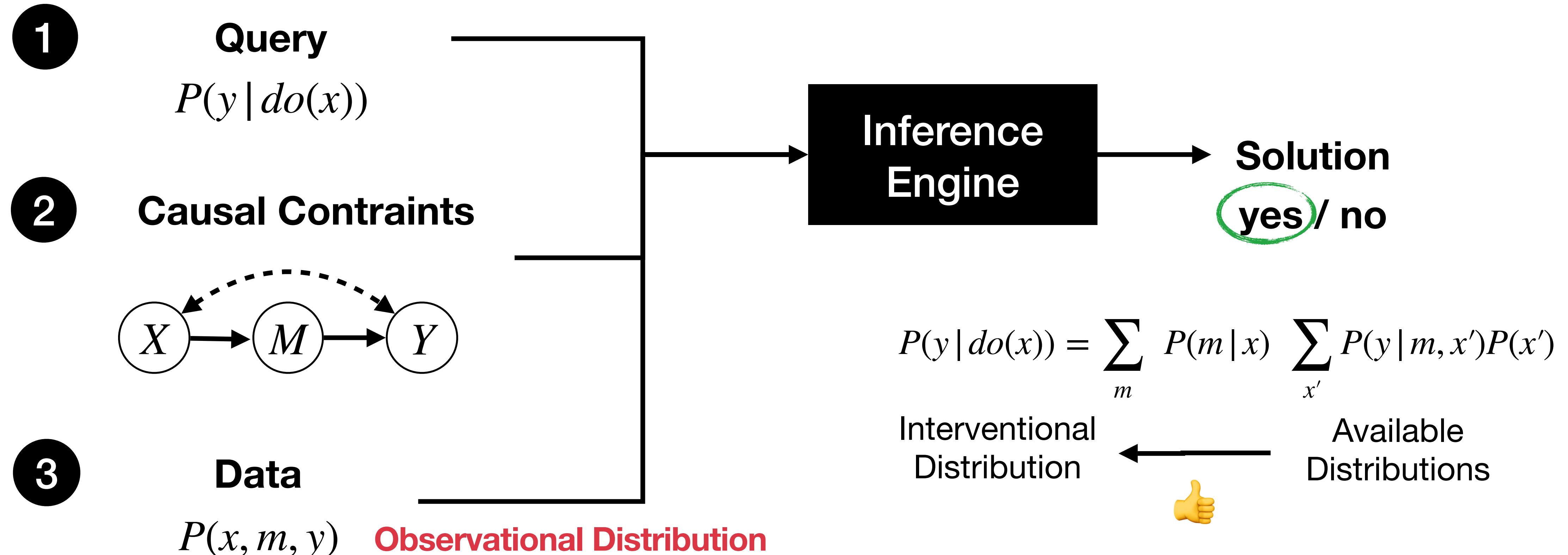


# Causal Identification - Current State of Art

- Identification from observational and experimental data
- Identification in partially understood domains (C-DAGs)
- Partial identification of (bounds on) causal effects
- Identification in Markov Equivalence Classes (PAGs)

# Classical Causal Effect Identification

Identification from observational data



- Tian, J. and Pearl, J. A General Identification Condition for Causal Effects. In Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002), pp. 567–573, Menlo Park, CA, 2002. AAAI Press/MIT Press.

# Causal Inference and Data-Fusion



## Causal inference and the data-fusion problem

Elias Bareinboim<sup>a,b,1</sup> and Judea Pearl<sup>a</sup>

<sup>a</sup>Department of Computer Science, University of California, Los Angeles, CA 90095; and <sup>b</sup>Department of Computer Science, Purdue University, West Lafayette, IN 47907

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved March 15, 2016 (received for review June 29, 2015)

<http://causalfusion.net>

**Fusion<sup>(β)</sup>**

**Summary**

Treatment :  $X$   
 Outcome :  $Y$   
 Adjusted :  
 Query :  $P_X(Y)$   
[Show More Details](#)

**Editor**

Graphical Structural  
[Refresh](#)

```

1 <NODES>
2 X 215,-205
3 Y 380,-205
4 Z_1 255,-270
5 Z_2 300,-340
6
7 <EDGES>
8 X -- Z_2 0.7
9 Y-- Z_2 -0.7
Populations
  
```

**Confounding Analysis**

Admissible Sets  
 Admissibility Test  
 Instrumental Variables  
 IV Admissibility Test

**Path Analysis**

D-Separation  
 Causal Paths  
 Confounding Paths  
 Biasing Paths

**Do-Calculus Analysis**

Do-Inspector  
 Do-Separation

**σ-Calculus Analysis**

σ-Inspector  
 σ-Separation

**Compute** The causal effect of  $X$  on  $Y$  conditional on  with do :  (Query:  $P_X(Y)$  from  $P(v)$ ) Non-Parametric  Clear

1 
$$P_X(Y) = \frac{\sum_{Z_2} P(X,Y|Z_1,Z_2)P(Z_2)}{\sum_{Z_2} P(X|Z_1,Z_2)P(Z_2)}$$

**Load**  
 Estimation  
 Derivation  
 Remove

**Fusion<sup>(β)</sup>**

**Summary**

- Treatment :  $X$
- Outcome :  $Y$
- Adjusted :
- Query :  $P_X(Y)$

Show More Details

**Editor**

Graphical Structural Refresh

```

1 <NODES>
2 X 170,-210
3 Y 375,-210
4 Z_1 225,-275
5 Z_2 225,-350
6 Z_3 275,-210
7
8 <EDGES>
9 X --> Z_1
  X --> Z_3
  Z_1 --> Z_2
  Z_1 --> Z_3
  Z_2 --> Z_3
  Z_3 --> Y

```

Populations

**Confounding Analysis**

- Admissible Sets
- Admissibility Test
- Instrumental Variables
- IV Admissibility Test

**Path Analysis**

- D-Separation
- Causal Paths
- Confounding Paths
- Biasing Paths

**Do-Calculus Analysis**

- Do-Inspector
- Do-Separation

**σ-Calculus Analysis**

- σ-Inspector
- σ-Separation

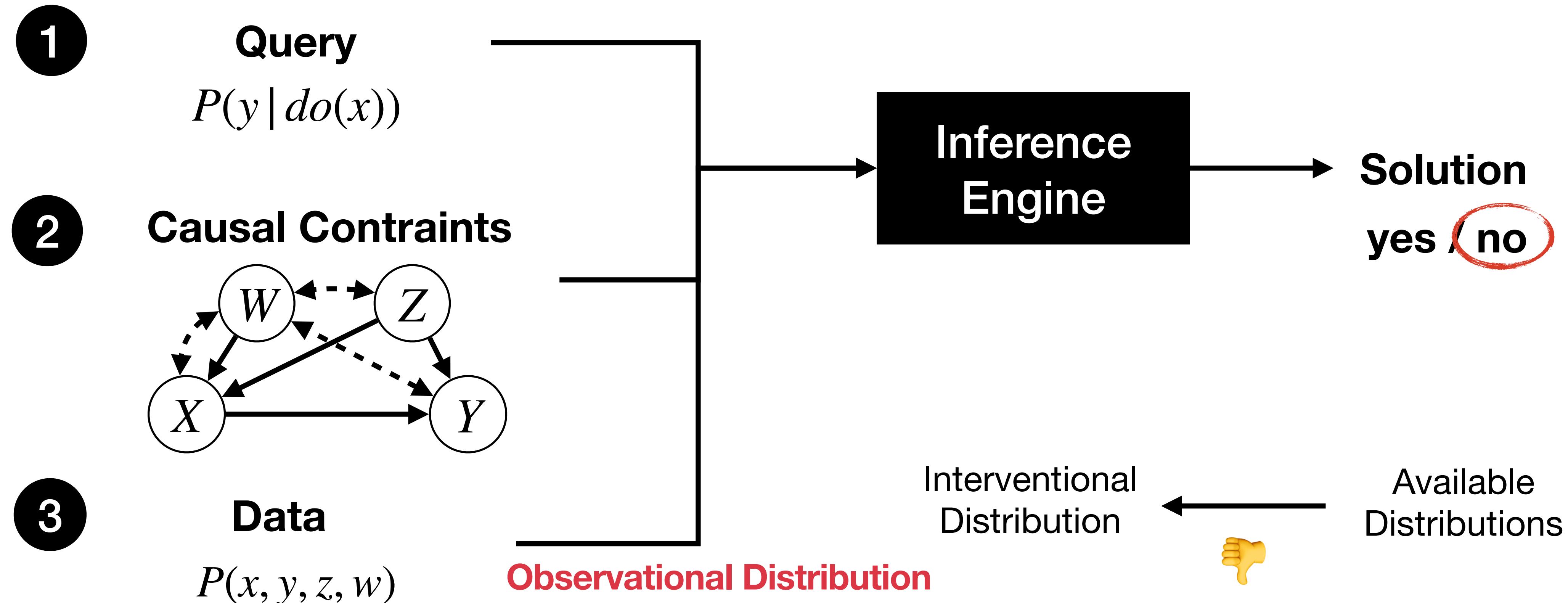
**Compute** The causal effect of  $X$  on  $Y$  conditional on  with do :  (Query:  $P_X(Y)$  from  $P(v)$ ) Non-Parametric  Clear

1 
$$P_X(Y) = \sum_{Z_2, Z_3} P(Y|X, Z_1, Z_2, Z_3) P(Z_2) \sum_{Z_1} P(Z_3|X, Z_1) P(Z_1)$$

**Load**  
Derivation  
Remove

# Classical Causal Effect Identification

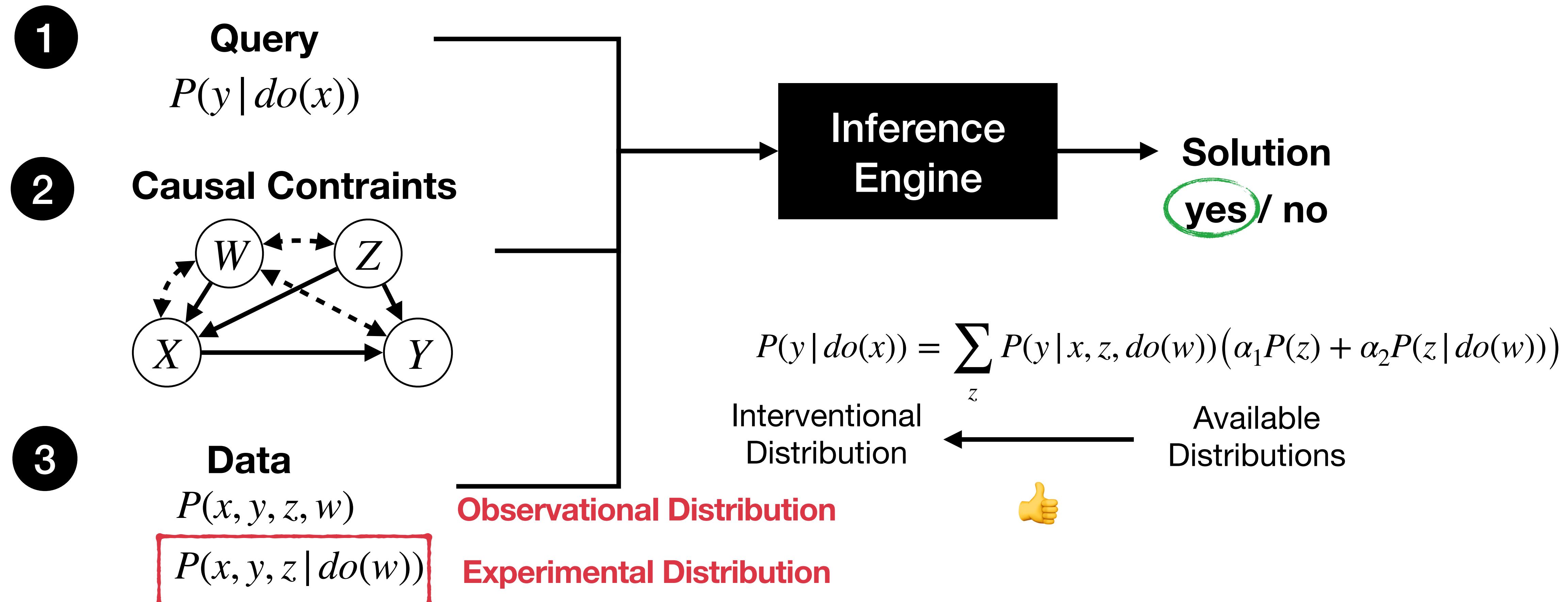
Identification from observational data



- Tian, J. and Pearl, J. (2002) A General Identification Condition for Causal Effects. In Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI), pp. 567–573, Menlo Park, CA. AAAI Press/MIT Press.

# General Causal Effect Identification

Identification from observational and experimental data



- Lee, S., Correa, J., and Bareinboim, E. (2019). General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 35, Tel Aviv, Israel. AUAI Press. [Link](#)

**Fusion<sup>(β)</sup>**

**Summary**

Treatment :  $X$   
 Outcome :  $Y$   
 Adjusted :  
 Query :  $P_X(Y)$

Show More Details

**Editor**

Graphical Structural Refresh

```

1 <NODES>
2 X 2796,-126
3 Y 2964,-126
4 W 2838,-186
5 Z 2916,-186
6
7 <EDGES>
8 X --> W 0.49033418064818
9
  
```

Populations

**Confounding Analysis**

Admissible Sets  
 Admissibility Test  
 Instrumental Variables  
 IV Admissibility Test

**Path Analysis**

D-Separation  
 Causal Paths  
 Confounding Paths  
 Biasing Paths

**Do-Calculus Analysis**

Do-Inspector  
 Do-Separation

**σ-Calculus Analysis**

σ-Inspector  
 σ-Separation

**Compute** The causal effect of  $X$  on  $Y$  conditional on  with do :  (Query:  $P_X(Y)$  from  $P(v), P_W(v)$ ) Non-Parametric  Clear

1 
$$P_X(Y) = \sum_Z P_W(Y|X, Z) \left( w_1^{(*)} P(Z) + w_2^{(*)} P_W(Z) \right)$$

**Load**  
**Derivation**  
**Remove**

# Identification from Observational and Experimental Data

## **Identification from observational and experimental data:**

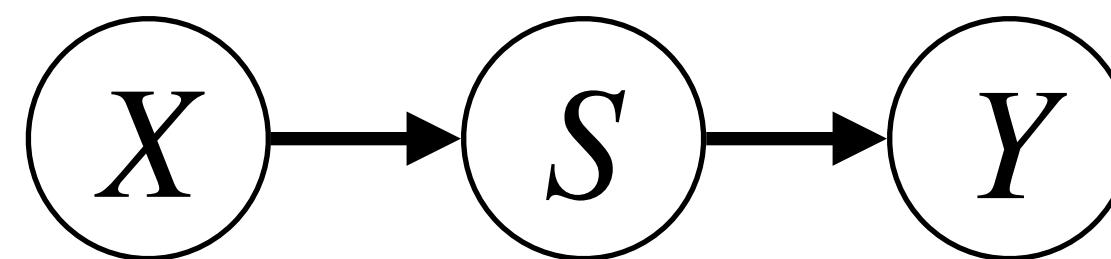
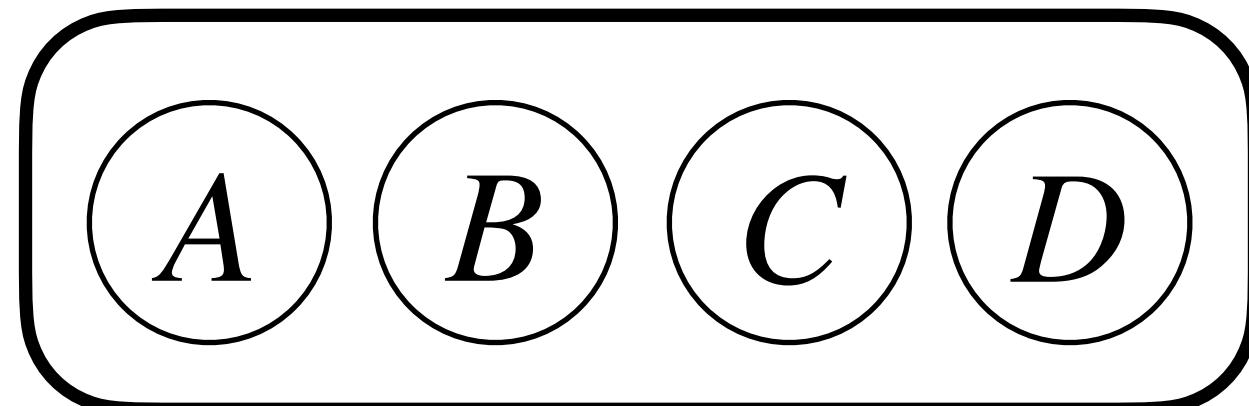
Lee, S., Correa, J., and Bareinboim, E. (2019). General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 35, Tel Aviv, Israel. AUAI Press. [Link](#)

## **Identification of stochastic/soft (and possibly imperfect) interventions:**

Correa, J. and Bareinboim, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY. AAAI Press. [Link](#)

# Identification in Partially Understood Systems

- (A) Age
- (B) Blood pressure
- (C) Comorbidities
- (D) Medication history
- (X) Lisinopril
- (S) Sleep Quality
- (Y) Stroke

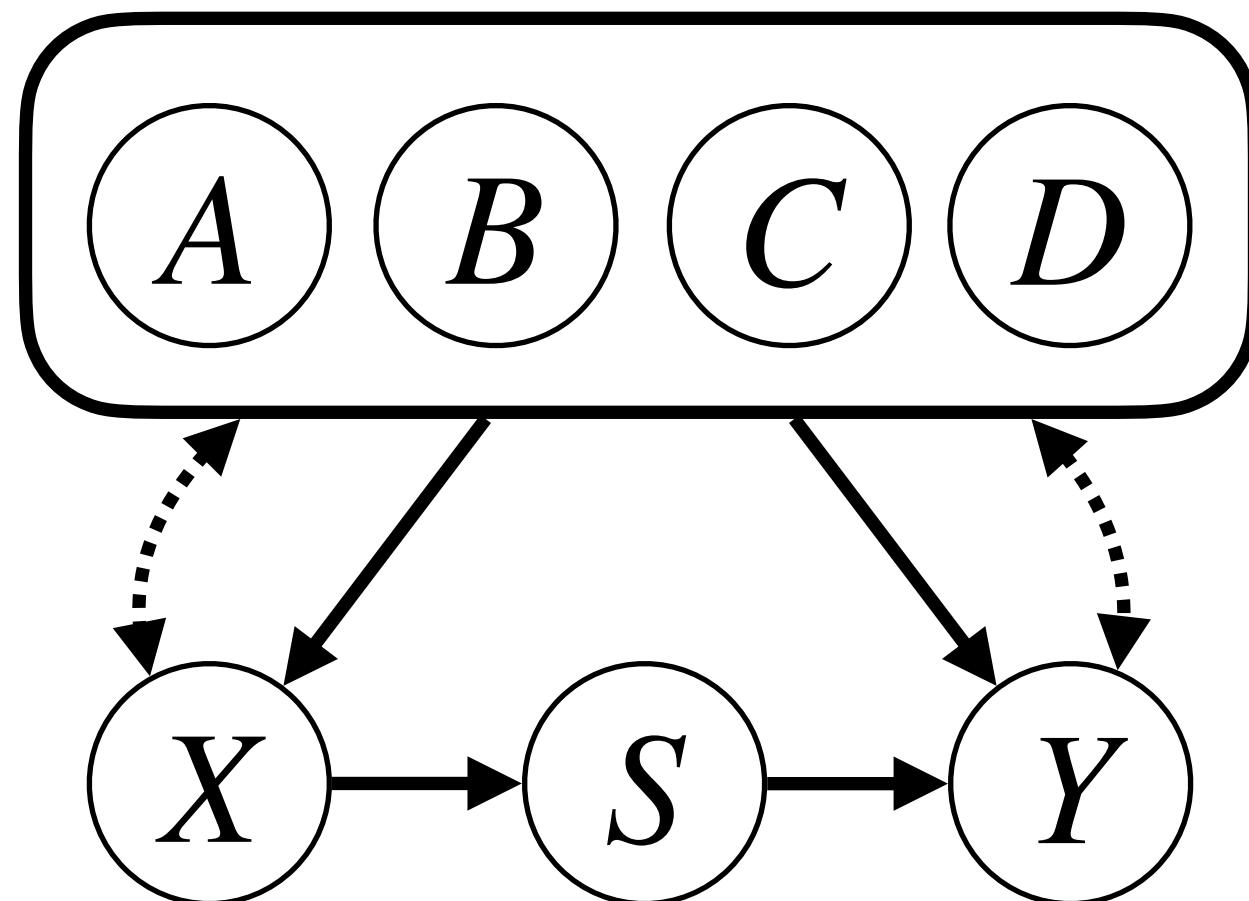


A causal diagram cannot be specified given the existing knowledge!

How can we identify  $P(y | do(x))$  in this case?

# Cluster Causal Diagrams (C-DAGs)

- (A) Age
- (B) Blood pressure
- (C) Comorbidities
- (D) Medication history
- (X) Lisinopril
- (S) Sleep Quality
- (Y) Stroke



$\{\{X\}, \{S\}, \{Y\}, \{A, B, C, D\}\}$

A *cluster causal diagram*  $G_C$  over a given partition  $\mathbf{C} = \{\mathbf{C}_1, \dots, \mathbf{C}_k\}$  of  $\mathbf{V}$  is compatible with a causal diagram  $G$  over  $\mathbf{V}$  if **for every**  $\mathbf{C}_i, \mathbf{C}_j \in \mathbf{C}$ :

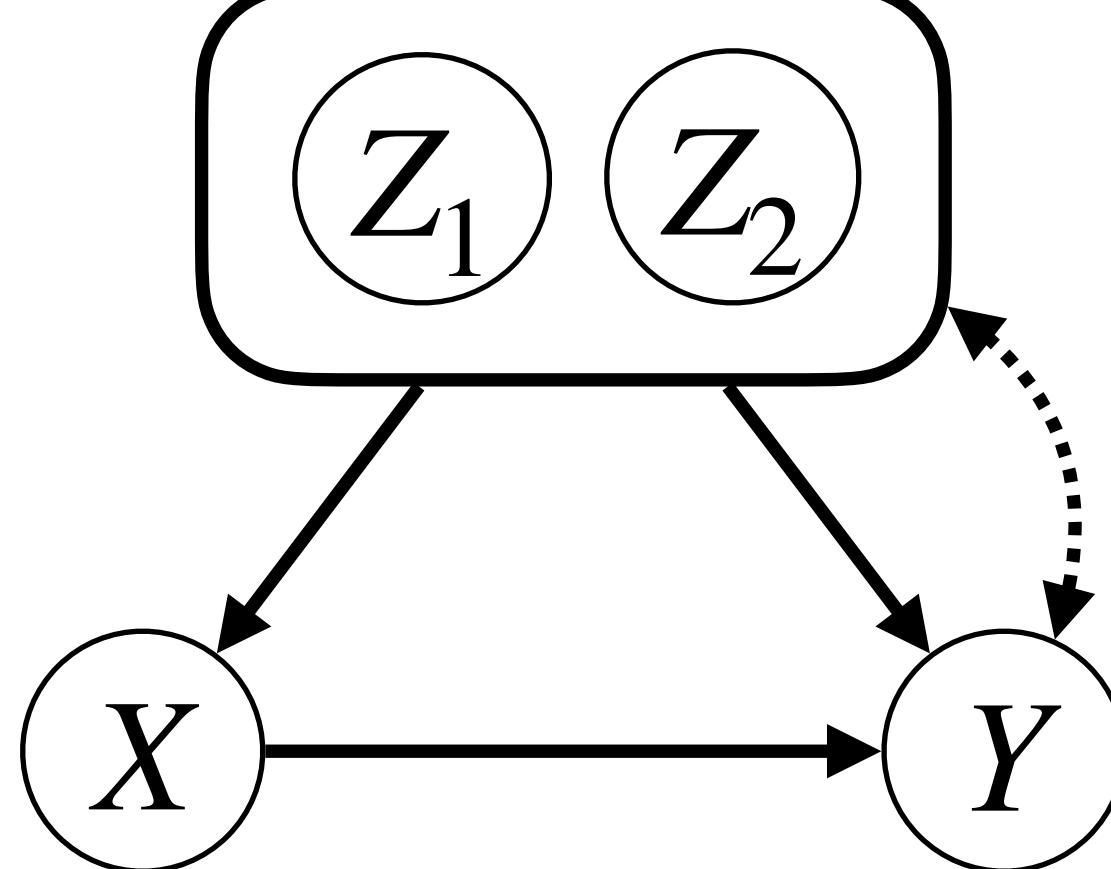
- $\mathbf{C}_i \rightarrow \mathbf{C}_j$  if  $\exists V_i \in \mathbf{C}_i$  and  $V_j \in \mathbf{C}_j$  such that  $V_i \rightarrow V_j$
- $\mathbf{C}_i \leftrightarrow \mathbf{C}_j$  if  $\exists V_i \in \mathbf{C}_i$  and  $V_j \in \mathbf{C}_j$  such that  $V_i \leftrightarrow V_j$

and  $G_C$  contains no cycles.

# Effect Identifiability given a C-DAG

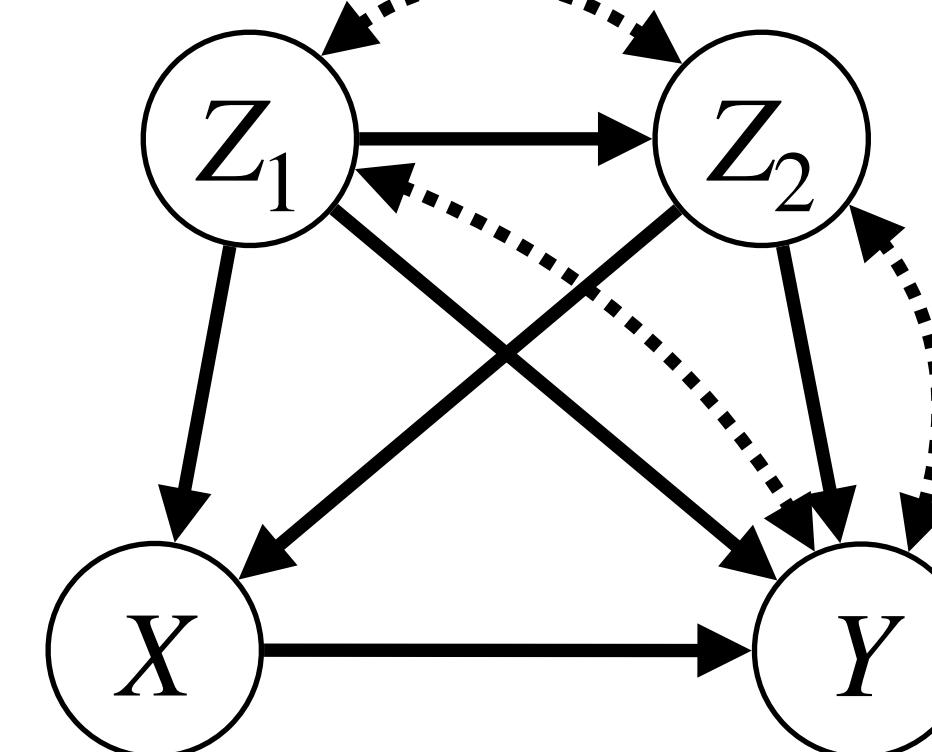
Simple evaluation of the **validity**  
of the Backdoor Criterion /  
Conditional Exchangeability

$G_C$



$$P(y | do(x)) = \sum_{\mathbf{z}} P(y | x, \mathbf{z}) P(\mathbf{z})$$

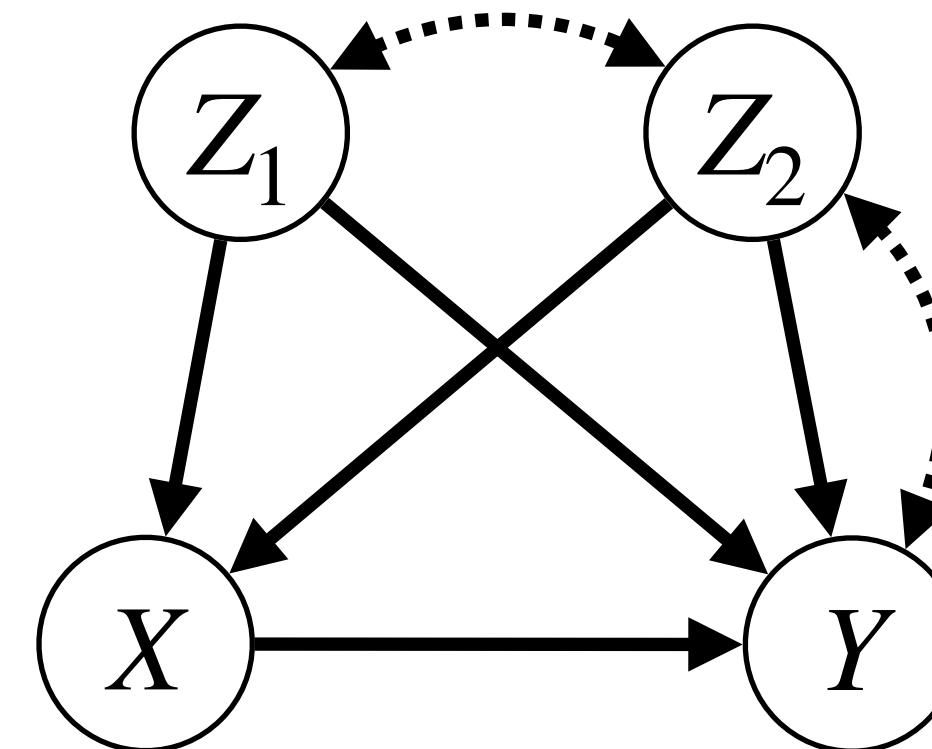
$G_1$



$$P(y | do(x)) =$$

$$\sum_{z_1, z_2} P(y | x, z_1, z_2) P(z_1, z_2)$$

$G_2$

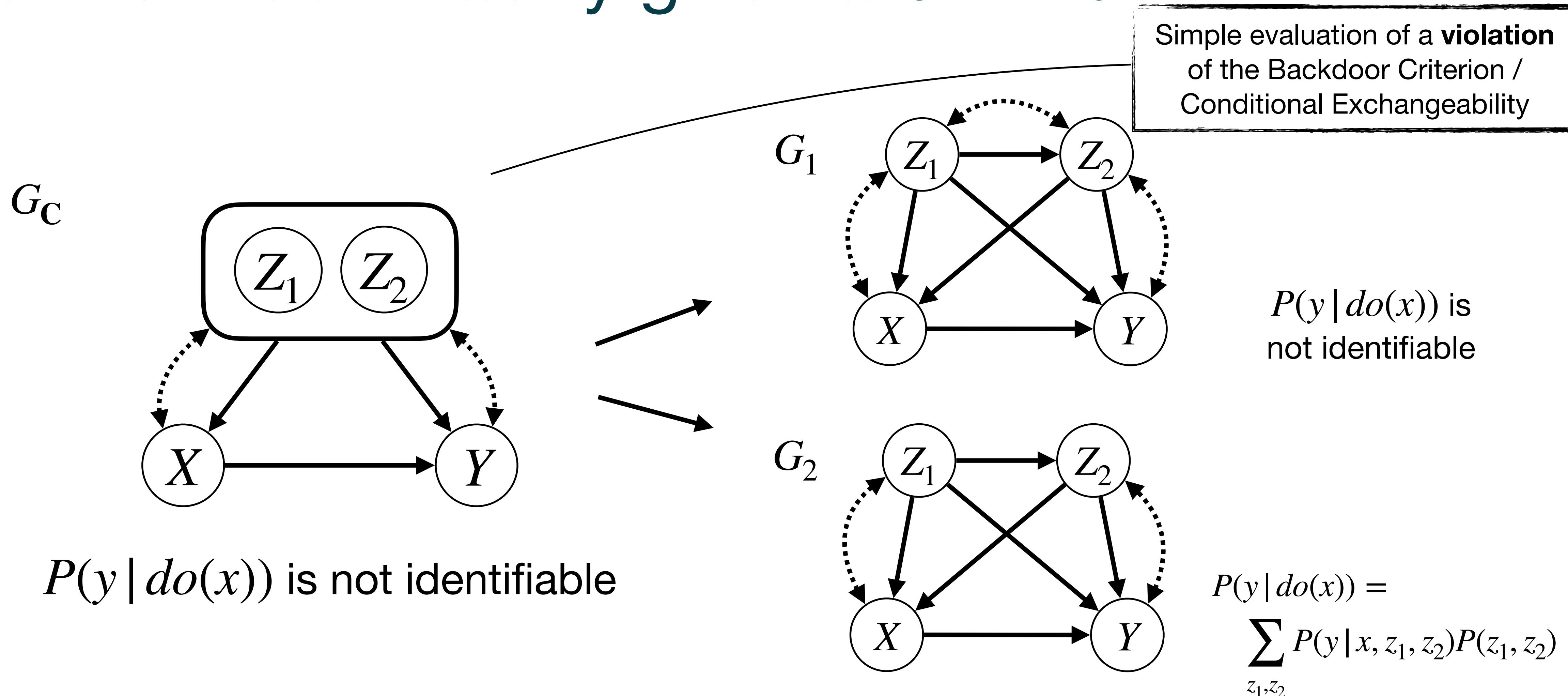


$$P(y | do(x)) =$$

$$\sum_{z_1, z_2} P(y | x, z_1, z_2) P(z_1, z_2)$$

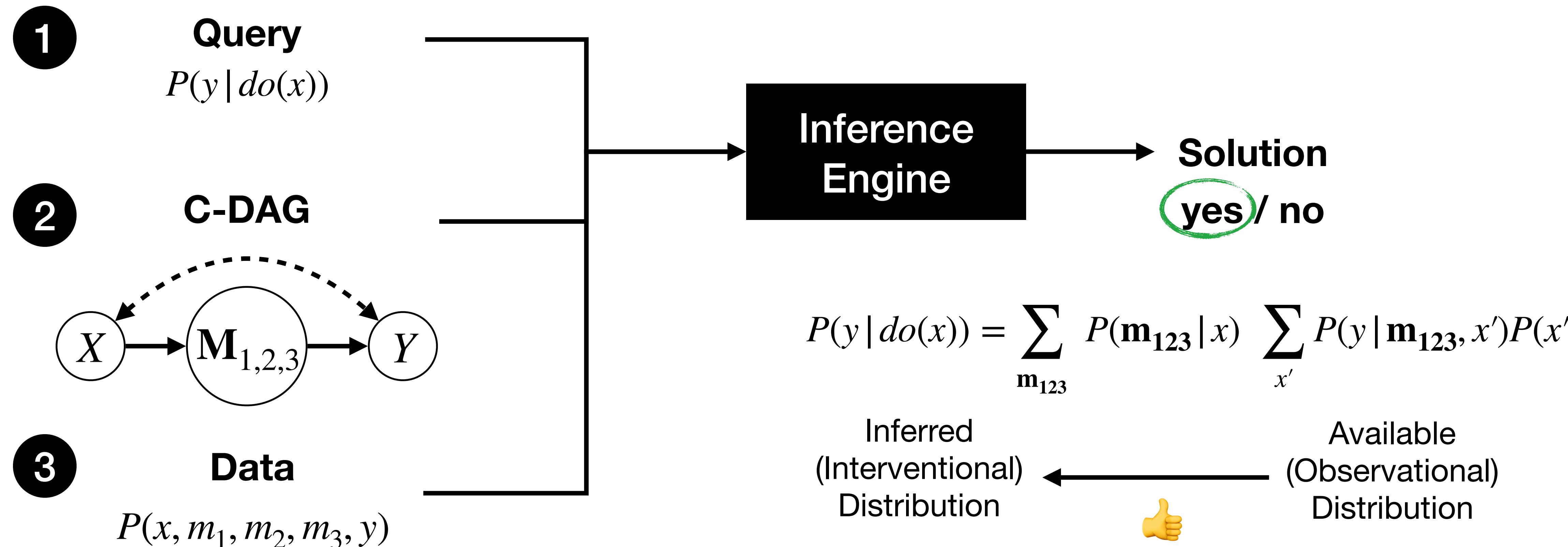
An identifiable effect in a C-DAG  $G_C$  is identifiable in all compatible causal diagrams  $G$  using the same identification formula!

# Effect Non-Identifiability given a C-DAG



A non-identifiable effect in a C-DAG  $G_C$  implies that there exists at least one compatible causal diagrams  $G$  in which the effect is non-identifiable.

# Identification of Causal Effects from C-DAGs



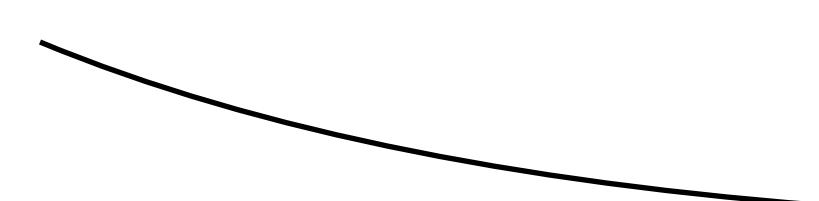
# Partial Identification of Causal Effects

## For Discrete Treatments:

- Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. Proceedings of the 39th International Conference on Machine Learning, PMLR 162, pages 26548–26558, 2022.

## For Continuous Treatments:

- Padh, Kirtan; Zeitler, Jakob; Watson, David; Kusner, Matt; Silva, Ricardo; Kilbertus, Niki; (2023) Stochastic causal programming for bounding treatment effects. In: (Proceedings) CLeaR 2023 : 2nd Conference on Causal Learning and Reasoning. ([Link](#))



**Partial Identification  
in C-DAGs!**

# What if *no* knowledge is available?



Can we learn a causal diagram  $\mathcal{G}$  from observational data?

## Causal Discovery (Lecture 2):

In non-parametric settings, we can't learn the true causal diagram, but algorithms such as the Fast Causal Inference (FCI) can learn a graphical representation of its *Markov equivalence class*!

Can we identify causal effects from the equivalence class?

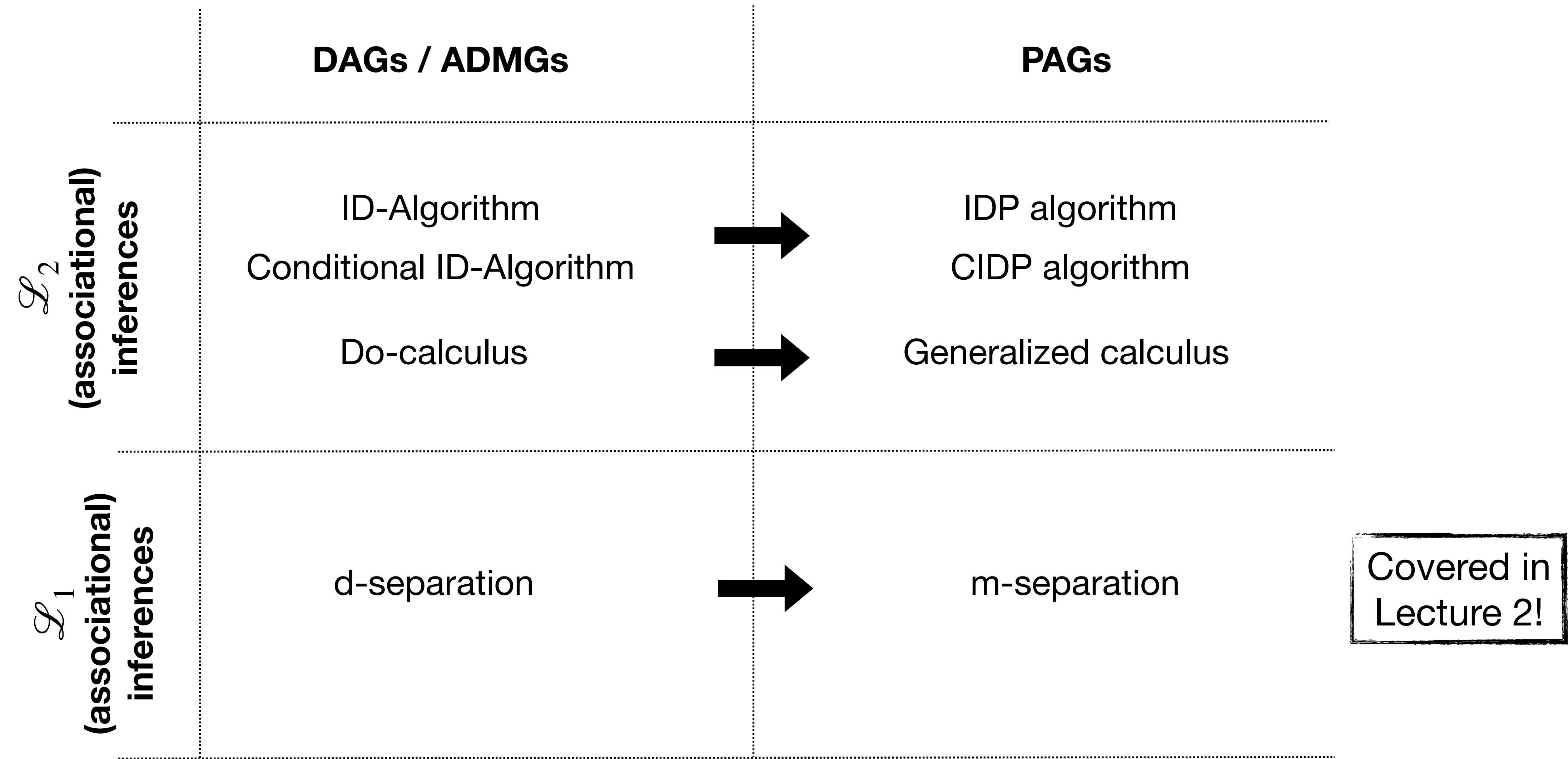
## Effect Identification:

Yes! Recently, we proposed **generalized do-calculus and complete algorithms** for the identification of marginal and conditional causal effect in PAGs!

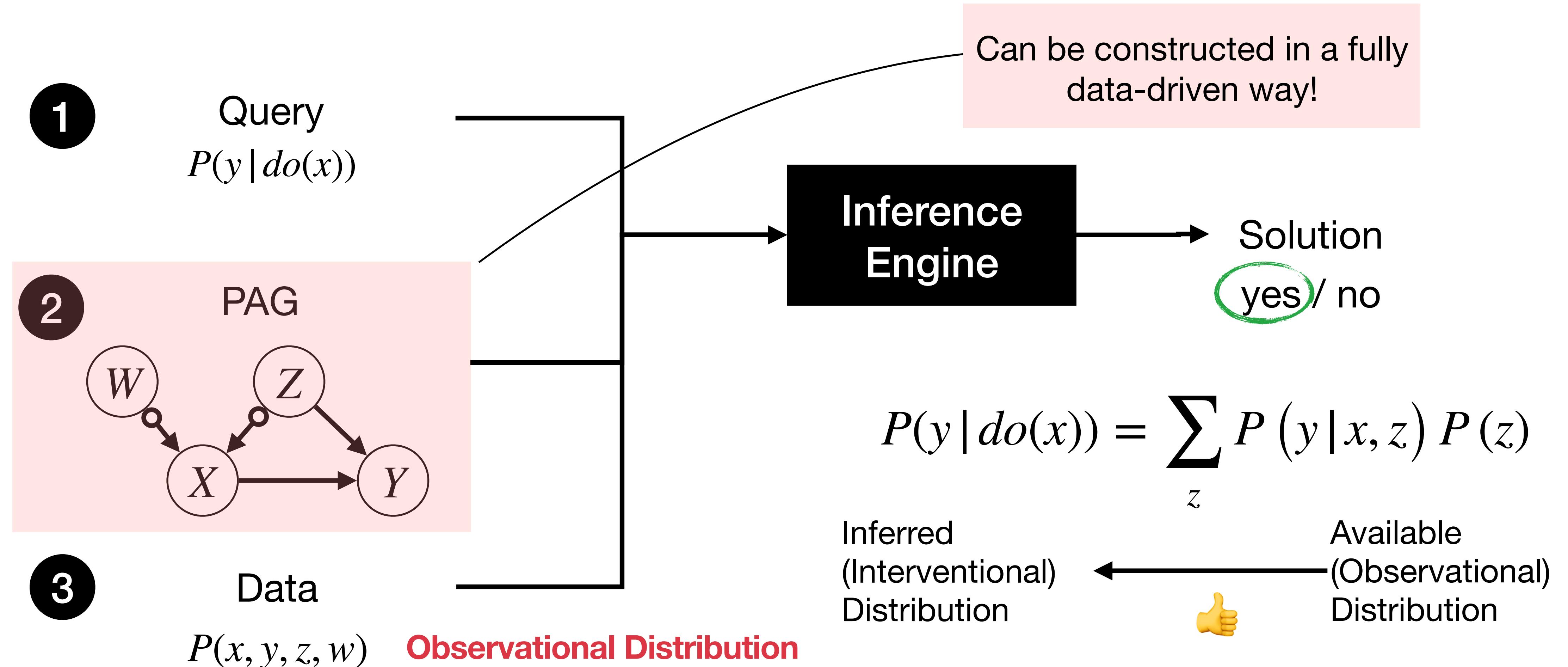
Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. [Link](#)

Jaber A., Ribeiro A. H., Zhang, J., Bareinboim, E. (2022) Causal Identification under Markov Equivalence - Calculus, Algorithm, and Completeness. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, NeurIPS 2022. [\(Link\)](#)

# Generalized Inference for PAGs

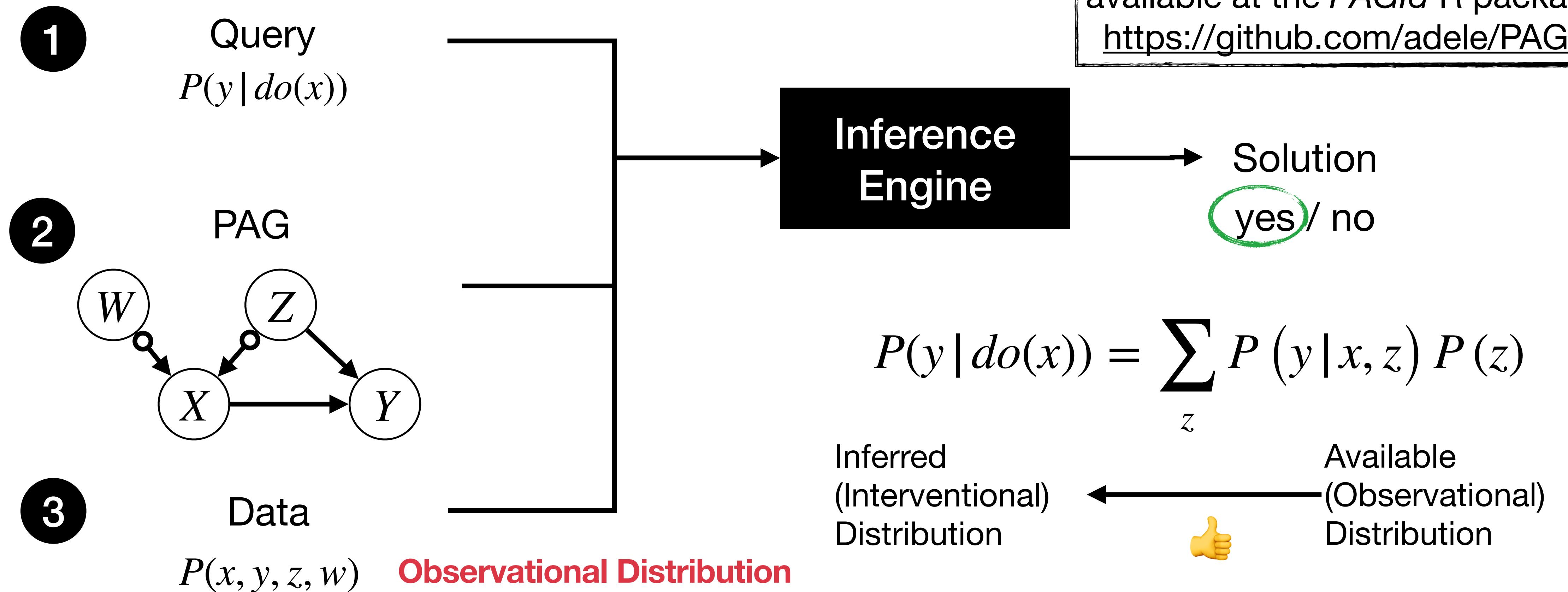


# General Identification in Markov Equivalence Classes



Jaber A., Ribeiro A. H., Zhang, J., Bareinboim, E. Causal Identification under Markov Equivalence - Calculus, Algorithm, and Completeness. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, NeurIPS 2022. ([Link](#))

# General Identification in Markov Equivalence Classes

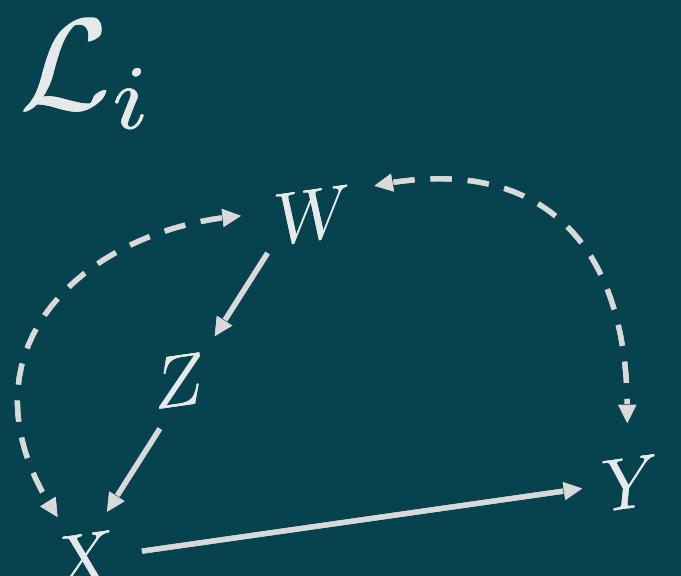


Jaber A., Ribeiro A. H., Zhang, J., Bareinboim, E. Causal Identification under Markov Equivalence - Calculus, Algorithm, and Completeness. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, NeurIPS 2022. ([Link](#))

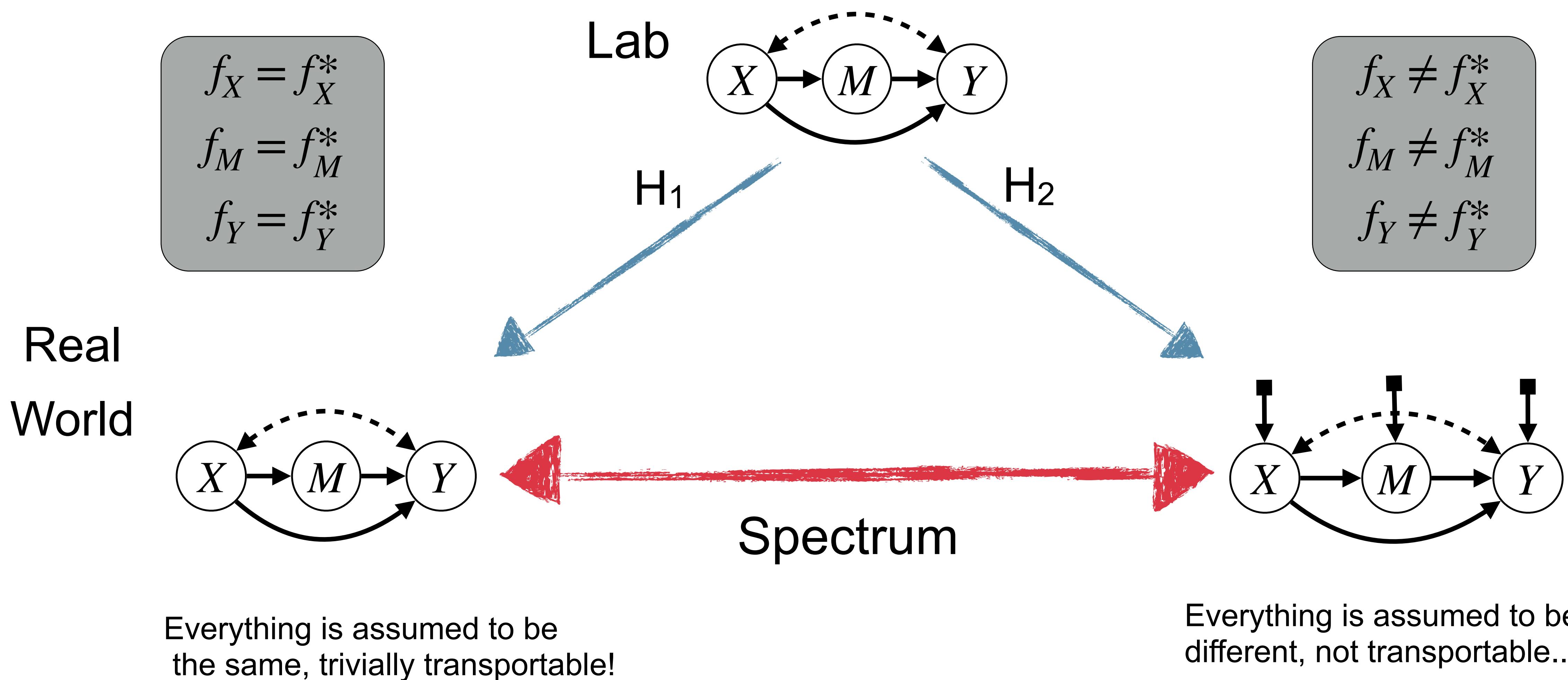


# Causal Transportability

Current State of the Art

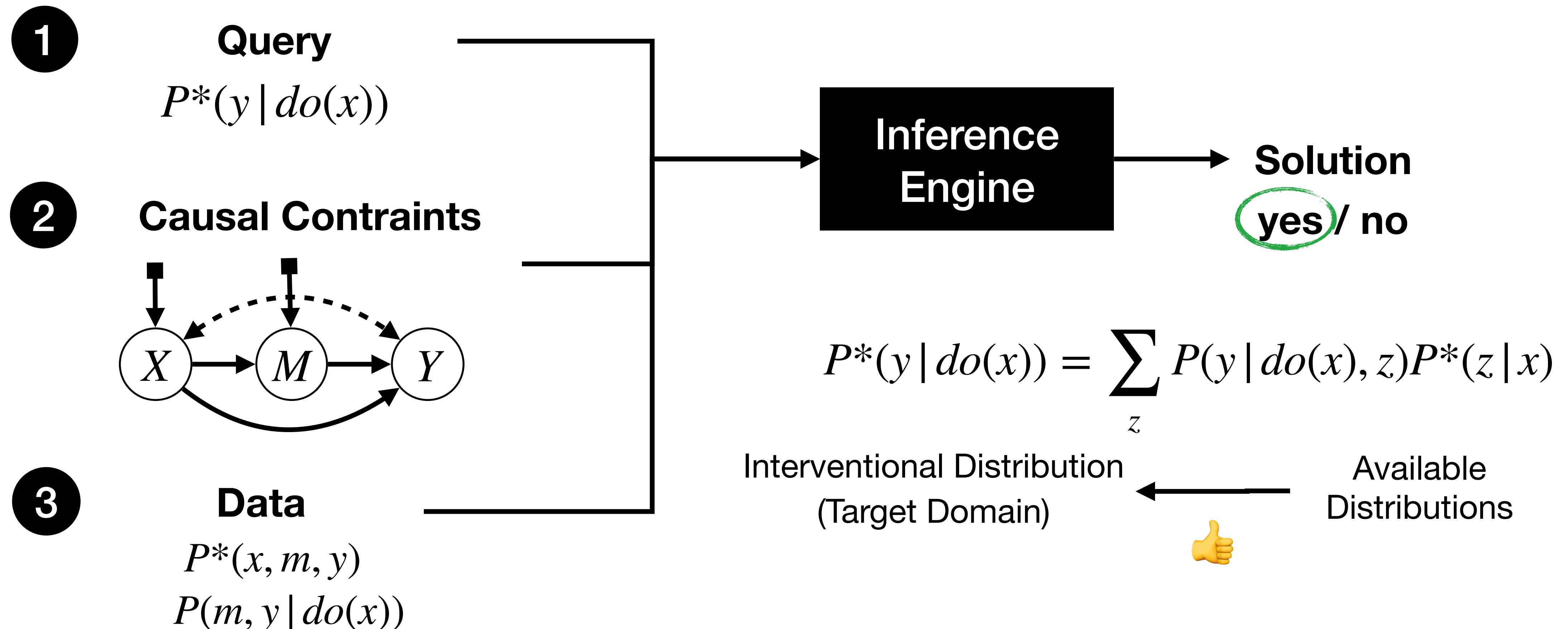


# Generalizability: from the “Lab” to the “Real World”



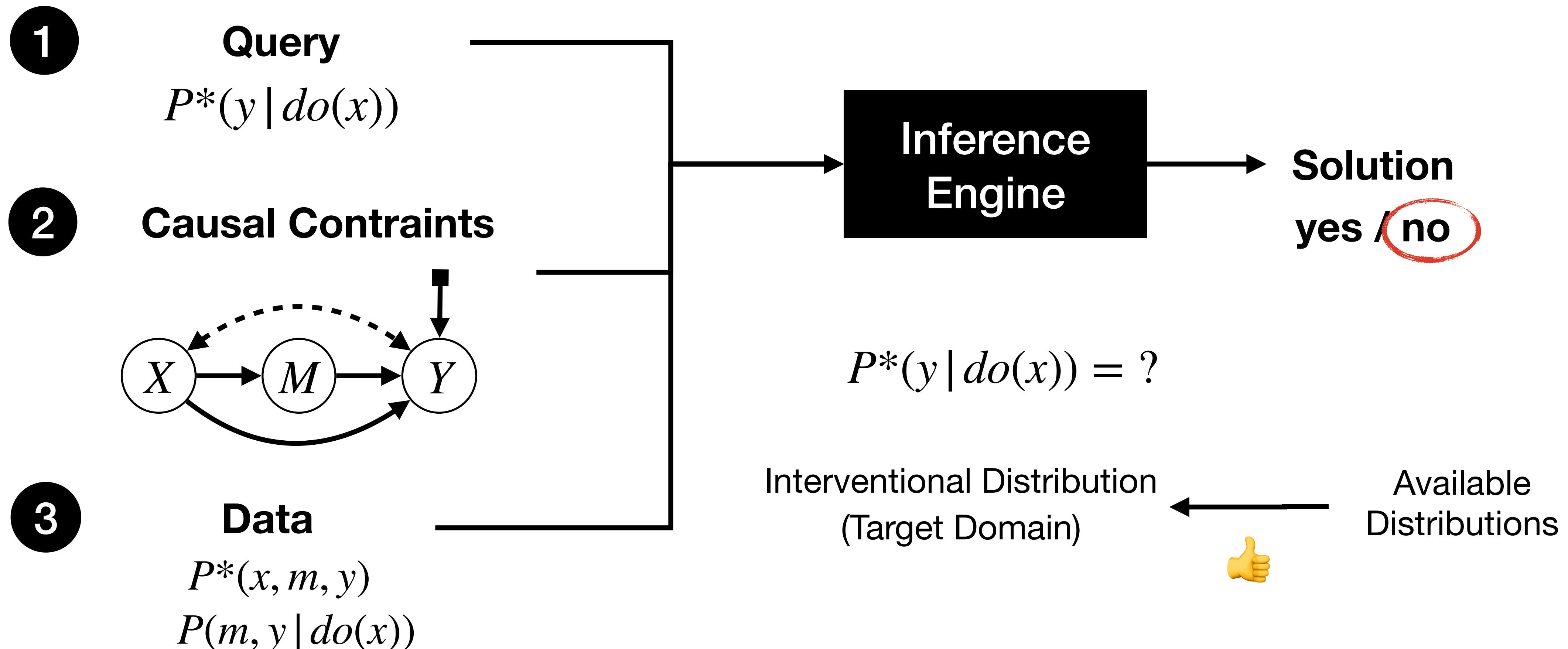
\* The lab stands for any environment, population, domain, setting.

# Effect Transportability - Workflow



- Lee, S., Correa, J., and Bareinboim, E. (2020). General Transportability – Synthesizing Observations and Experiments from Heterogeneous Domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY. AAAI Press. [Link](#)

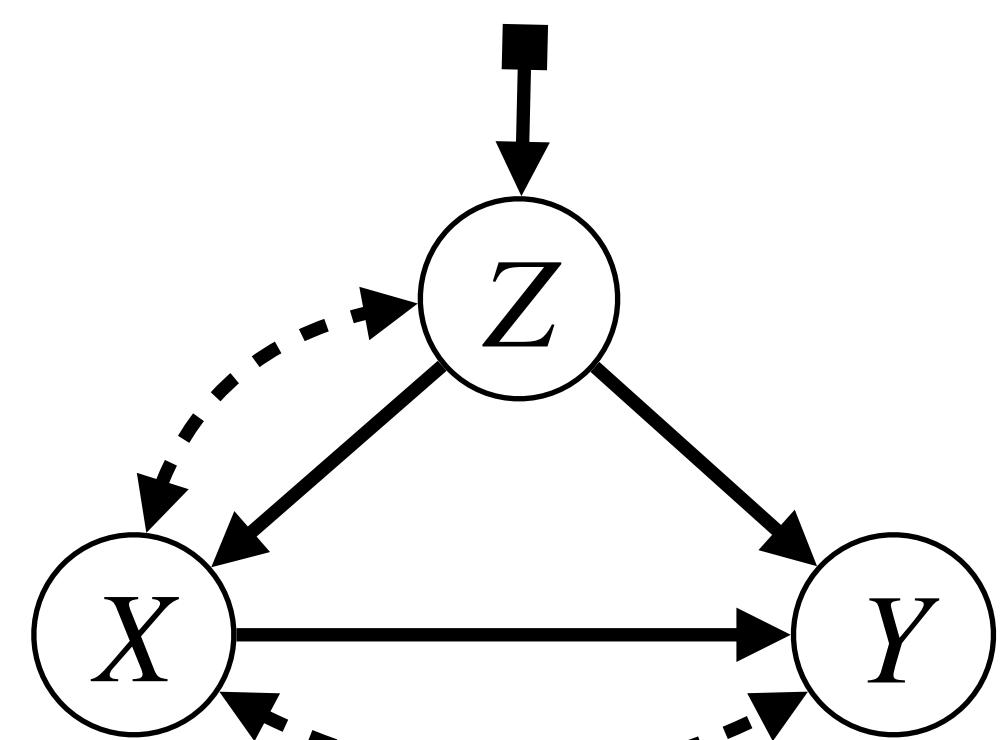
# Effect Transportability - Workflow



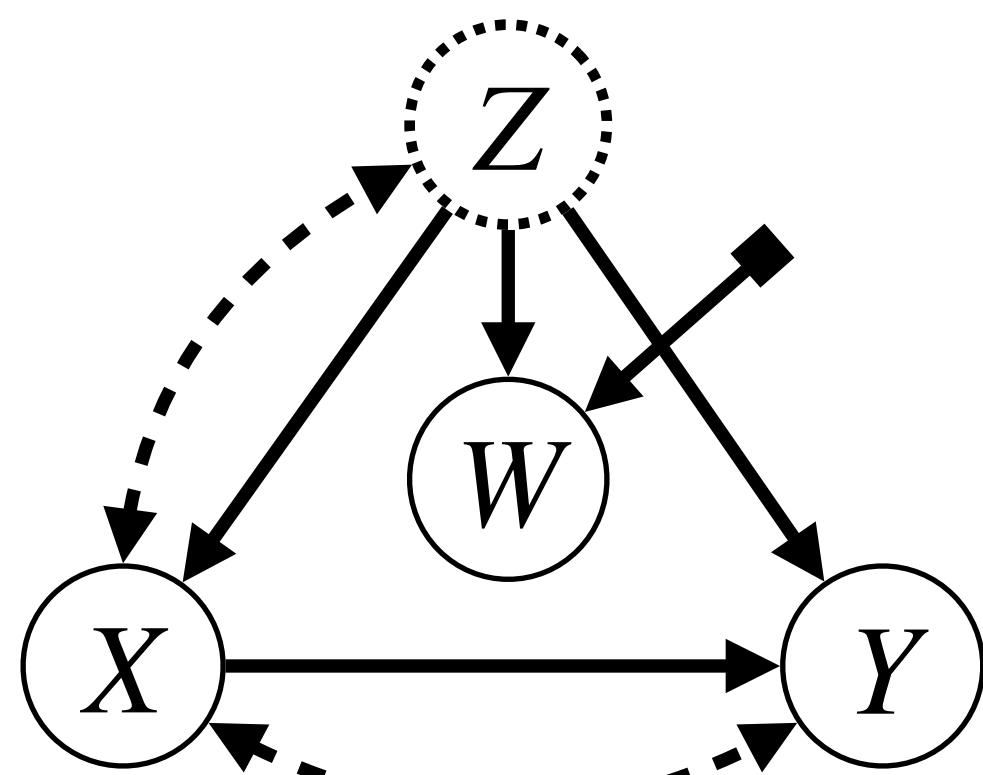
- Lee, S., Correa, J., and Bareinboim, E. (2020). General Transportability – Synthesizing Observations and Experiments from Heterogeneous Domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY. AAAI Press. [Link](#)

# Sensitivity to the Causal Assumptions

$X$ : Treatment

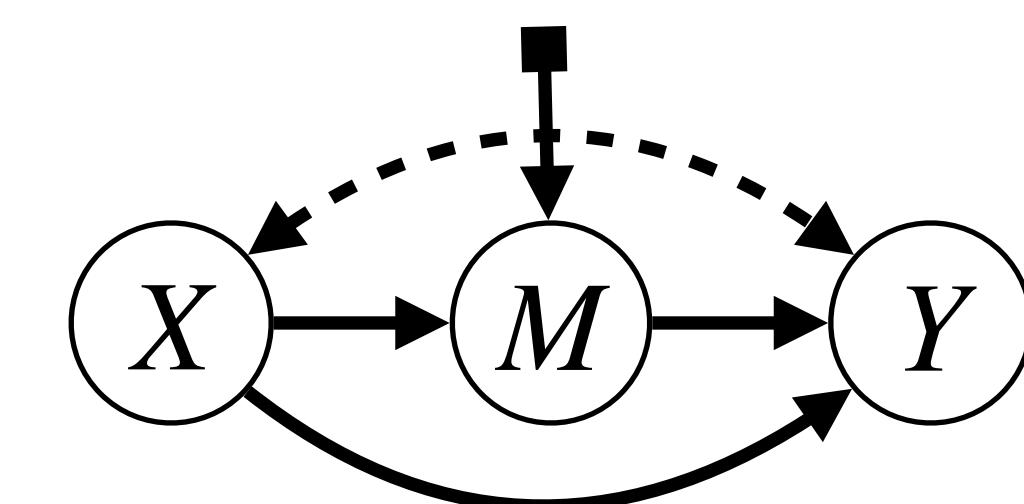


$Y$ : Outcome



$\Pi \rightarrow \Pi^*$

USA      Germany



$Z$ : Age

$$P^*(y | do(x)) = \sum_z P(y | do(x), z) P^*(z)$$

$W$ : Education

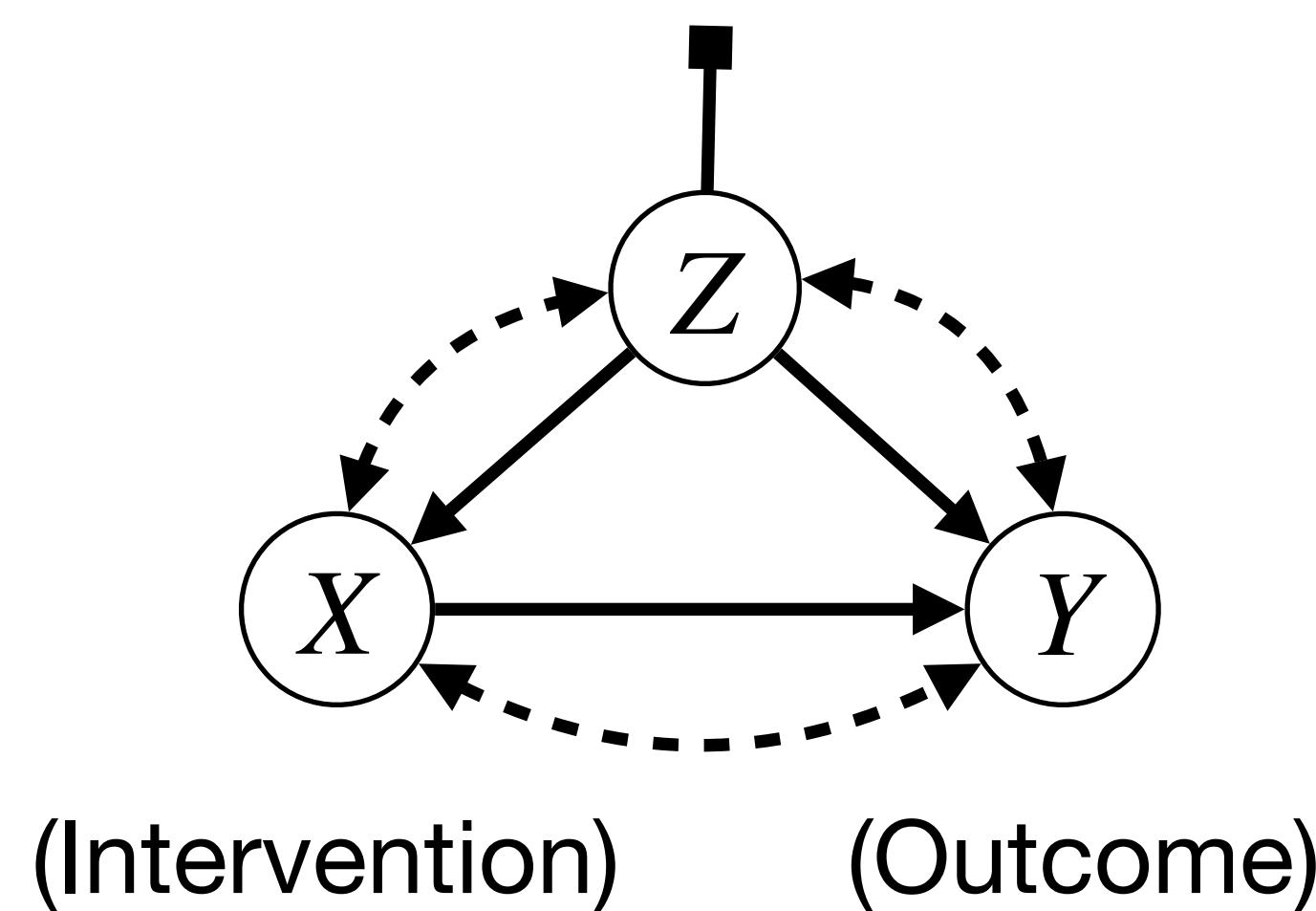
$$P^*(y | do(x)) = P(y | do(x))$$

$M$ : Biological Mediator

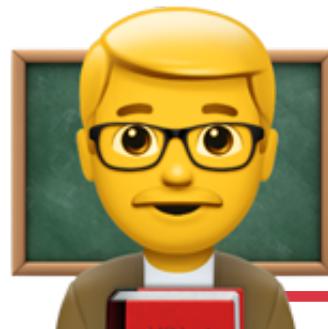
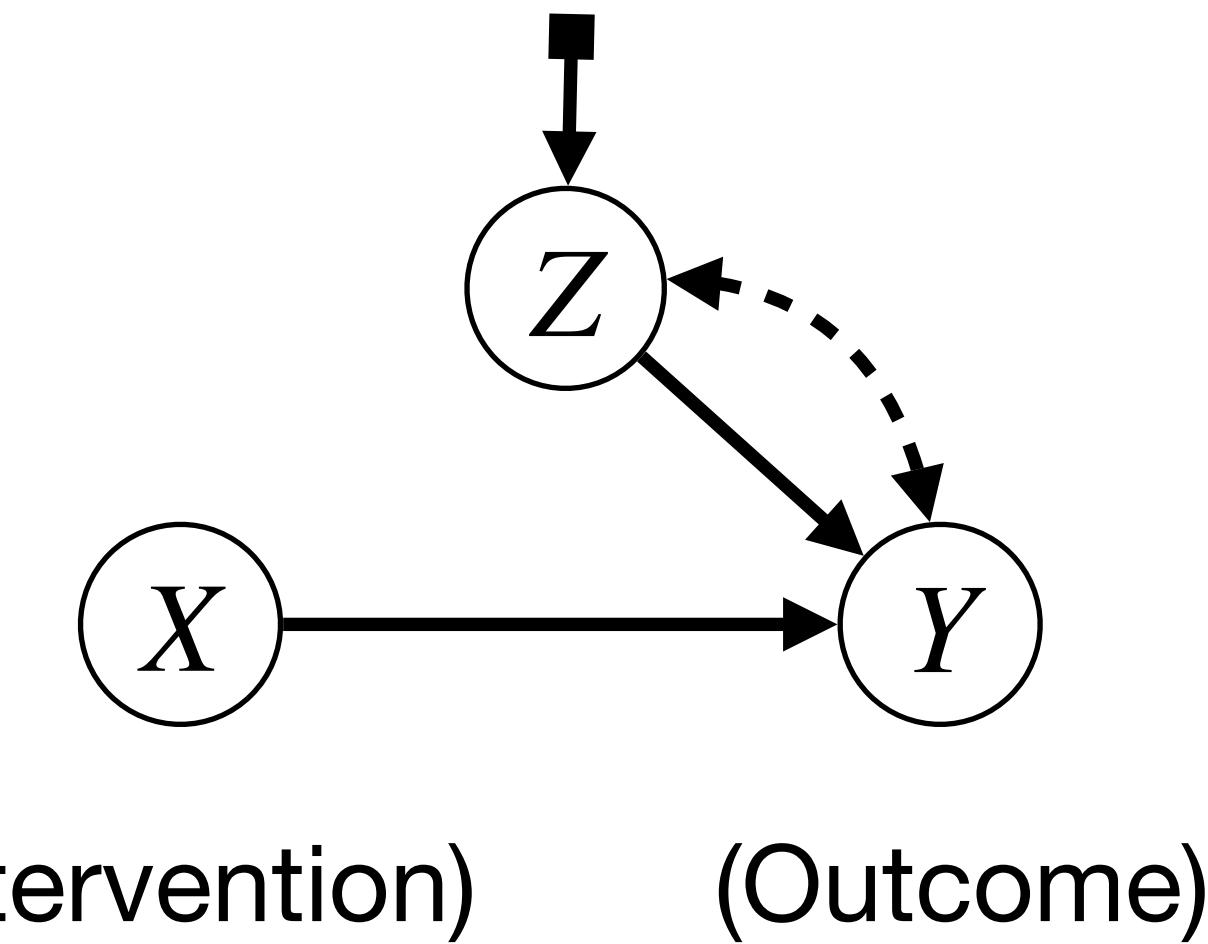
$$P^*(y | do(x)) = \sum_z P(y | do(x), m) P^*(m | x)$$

# Is the Gold Standard Golden?

Before randomization



After randomization



**Lesson.** Even if we have a perfect RCT, one still needs to go through a transportability exercise. TR theory is unavoidable.

Fusion (β)

Summary  
Treatment :  $X$   
Outcome :  $Y$   
Adjusted :  
Query :  $P_X^*(Y)$   
[Show More Details](#)

Editor  
Populations  
Target  $\pi^*$   
S  $\pi$  [Edit](#) [Delete](#)  
 Hide label \* [Add Population](#)

Datasets

The causal effect of  $X$  on  $Y$  conditional on  with do  $Z^* : \equiv Z : \equiv$  (Query:  $P_X^*(Y)$  from  $P^*(v), P(v), P_X(v)$ ) Non-Parametric  Clear

1 
$$P_X^*(Y) = \sum_Z P_X(Y|Z) P^*(Z)$$

Confounding Analysis  
Admissible Sets  
Admissibility Test  
Instrumental Variables  
IV Admissibility Test

Path Analysis  
D-Separation  
Causal Paths  
Confounding Paths

Do-Calculus Analysis  
Do-Inspector  
Do-Separation

σ-Calculus Analysis  
σ-Inspector  
σ-Separation

Testable Implications

Load  
Derivation  
Remove

Load

Fusion (β)

Treatment :  $X$   
Outcome :  $Y$   
Adjusted :  
Query :  $P$

Show More

Editor  
Populations  
Target  
 $S \quad \pi$   
 Hide label  $*$   
+ Add Population  
Datasets

Compute

Summary  
Treatment :  $X$   
Outcome :  $Y$   
Adjusted :  
Query :  $P_X^*(Y)$   
Show More Details

Editor  
Populations  
Target  $\pi^*$   
 $S \quad \pi$   
 Hide label  $*$   
+ Add Population  
Datasets

Compute  
The causal effect of  $X$  on  $Y$  conditional on  $Z$  with do  $Z^* : \equiv Z : \equiv$  (Query:  $P_X^*(Y)$  from  $P^*(v), P(v), P_X(v)$ ) Non-Parametric  Clear

1

$P_X^*(Y) = P_X(Y)$

Load  
Derivation  
Remove

Load

```

graph TD
    S((S)) --> Z((Z))
    S((S)) --> X((X))
    X((X)) --> Y((Y))
    Z((Z)) --> Y((Y))
    S((S)) -.-> W((W))
    W((W)) -.-> Z((Z))
    W((W)) -.-> Y((Y))
  
```

The screenshot shows the CausalFusion web application interface, which includes multiple tabs and panels for causal inference analysis.

**Summary Panel:**

- Treatment:  $X$
- Outcome:  $Y$
- Adjusted:
- Query:  $P$
- Show More

**Editor Panel:**

- Treatment:  $X$
- Outcome:  $Y$
- Adjusted:
- Query:  $P_X^*(Y)$
- Show More Details

**Populations Panel:**

- Target
- $S \quad \pi$
- Hide label  $*$
- + Add Population
- Datasets

**Compute Panel:**

- Compute
- 1

**Summary Panel (Second Tab):**

- Treatment:  $X$
- Outcome:  $Y$
- Adjusted:
- Query:  $P_X^*(Y)$
- Show More Details

**Editor Panel (Second Tab):**

- Treatment:  $X$
- Outcome:  $Y$
- Adjusted:
- Query:  $P_X^*(Y)$
- Show More Details

**Populations Panel (Second Tab):**

- Target
- $\pi^*$
- Hide label  $*$
- + Add Population
- Datasets

**Diagram Area:**

```

graph LR
    X((X)) --> Z((Z))
    S[Instrumental Variable] --> Z
    Z --> Y((Y))
    S -.-> Y
  
```

**Confounding Analysis Panel:**

- Admissible Sets
- Admissibility Test
- Instrumental Variables
- IV Admissibility Test

**Path Analysis Panel:**

- D-Separation
- Causal Paths
- Confounding Paths

**Do-Calculus Analysis Panel:**

- Do-Inspector
- Do-Separation

**σ-Calculus Analysis Panel:**

- σ-Inspector
- σ-Separation

**Testable Implications Panel:**

- Non-Parametric
- Clear

**Equation Panel:**

$$P_X^*(Y) = \sum_Z P_X(Y|Z) P^*(Z|X)$$

**Tool Buttons:**

- Load
- Derivation
- Remove
- Load

**Bottom Navigation:**

- 1

# More on Causal Transportability

## Transportability of causal effects:

Lee, S., Correa, J., and Bareinboim, E. (2020). General Transportability – Synthesizing Observations and Experiments from Heterogeneous Domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY. AAAI Press. [Link](#)

## Transportability of stochastic/soft interventions:

Correa, J. and Bareinboim, E. (2020). General transportability of soft interventions: Completeness results. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10902–10912, Vancouver, Canada. Curran Associates, Inc. [Link](#)

## Counterfactual transportability:

Correa, J. Lee, S., Bareinboim. E. (2022). Counterfactual Transportability: A Formal Approach. In *Proceedings of the 39th International Conference on Machine Learning*.

# Other Lines of Work and Research Questions

- Robustness against violations of the underlying assumptions (e.g., faithfulness)
- Bayesian approaches for modeling of uncertainty
- Causal Data-Fusion:
  - Learning and inference from a combination of data + prior knowledge
  - Learning and inference from multiple heterogeneous datasets and populations
- Causality and Continual Learning <https://www.continualcausality.org/>
- And much more...
  - Counterfactual identification <https://causalai.net>
  - Causal Fairness <https://fairness.causalai.net>
  - Causal reinforcement learning <https://crl.causalai.net>



Thank You Very Much for Staying With Us !!  
Hope we succeeded in “**causing**” your thinking to go  
towards the importance of causality in AI/ML

Please feel free to reach out:

Devendra Singh Dhami ([devendra.dhami@tu-darmstadt.de](mailto:devendra.dhami@tu-darmstadt.de)),

@devendratweetin

Matej Zečević ([matej.zecevic@tu-darmstadt.de](mailto:matej.zecevic@tu-darmstadt.de))

@matej\_zecevic

Adèle Ribeiro ([adele.ribeiro@uni-marburg.de](mailto:adele.ribeiro@uni-marburg.de))

@adelehr