Credit: Tom Gauld

hessian.AI

CAUSE

TECHNISCHE UNIVERSITÄT DARMSTADT

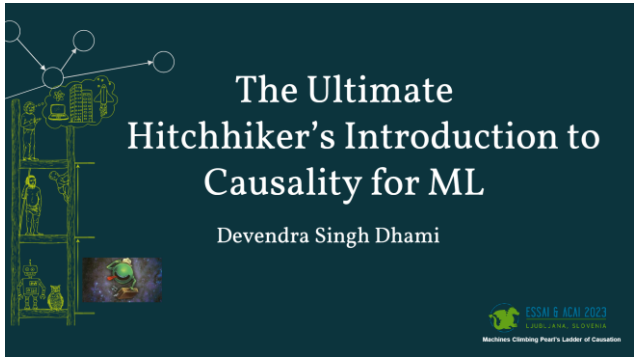
PHILIPPS-UNIVERSITÄT MARBURG

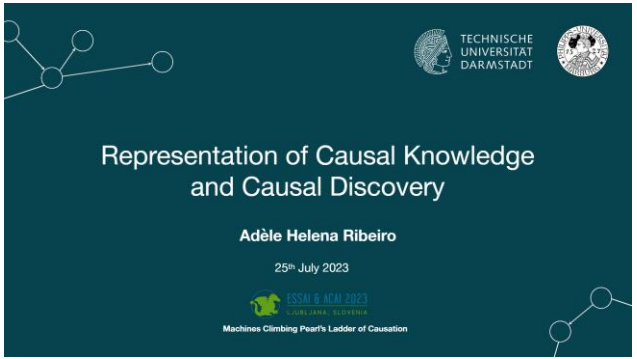
Machines Climbing Pearl's Ladder of Causation- Day 5

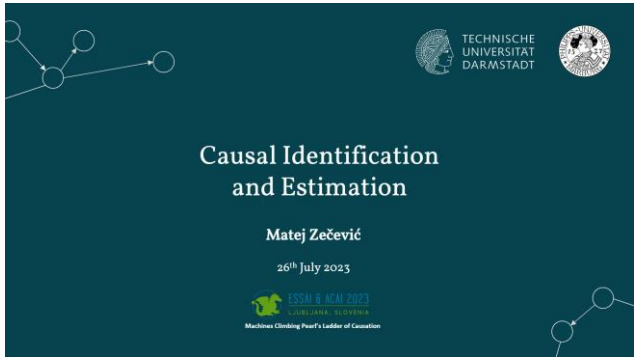
Devendra Singh Dhami
Matej Zečević
Adèle Ribeiro

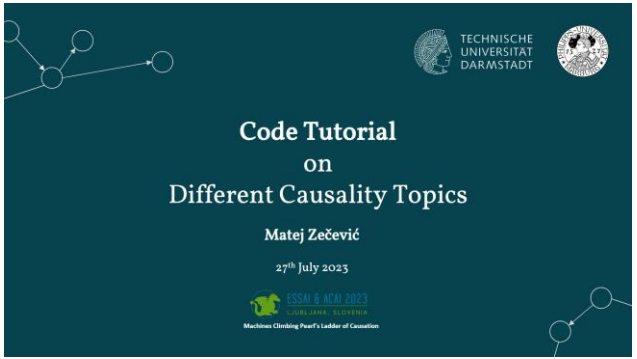
ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA

We have Seen…….


The Ultimate
Hitchhiker's Introduction to
Causality for ML
Devendra Singh Dhami
ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA
Machines Climbing Pearl's Ladder of Causation


TECHNISCHE
UNIVERSITÄT
DARMSTADT
Representation of Causal Knowledge
and Causal Discovery
Adèle Helena Ribeiro
25th July 2023
ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA
Machines Climbing Pearl's Ladder of Causation


TECHNISCHE
UNIVERSITÄT
DARMSTADT
Causal Identification
and Estimation
Matej Zečević
26th July 2023
ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA
Machines Climbing Pearl's Ladder of Causation


TECHNISCHE
UNIVERSITÄT
DARMSTADT
Code Tutorial
on
Different Causality Topics
Matej Zečević
27th July 2023
ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA
Machines Climbing Pearl's Ladder of Causation

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

Devendra Singh Dhami
Adèle Ribeiro

ESSAI & ACAI 2023
LJUBLJANA, SLOVENIA

Causality
And
More

Causal Discovery
And
Identification

Causal
Explanations

Causal
Generative
Modelling

Causal Machine Learning

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

Causal Discovery And Identification

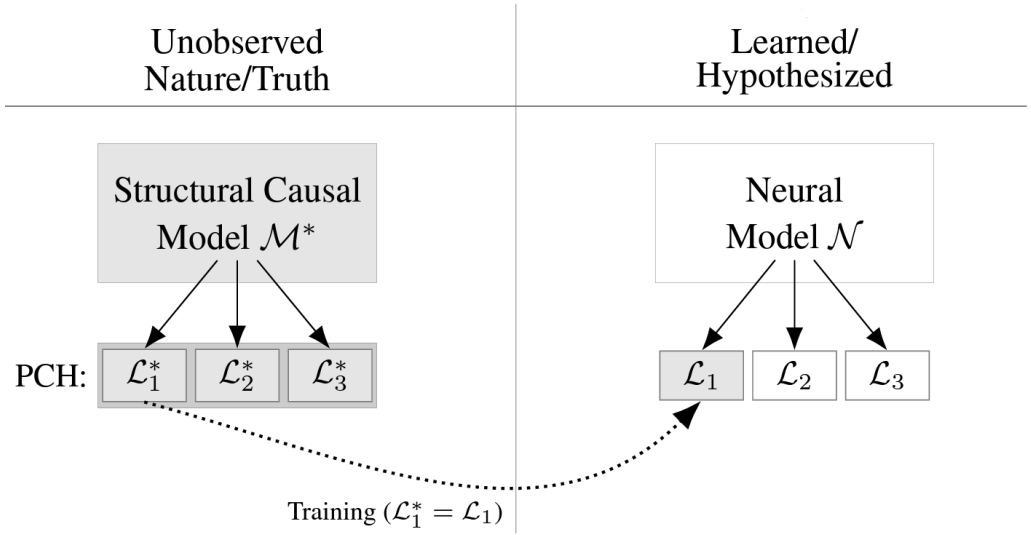
Causal Machine Learning

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

Causal Machine Learning

Causal
Generative
Modelling

The Neural-Causal Connection



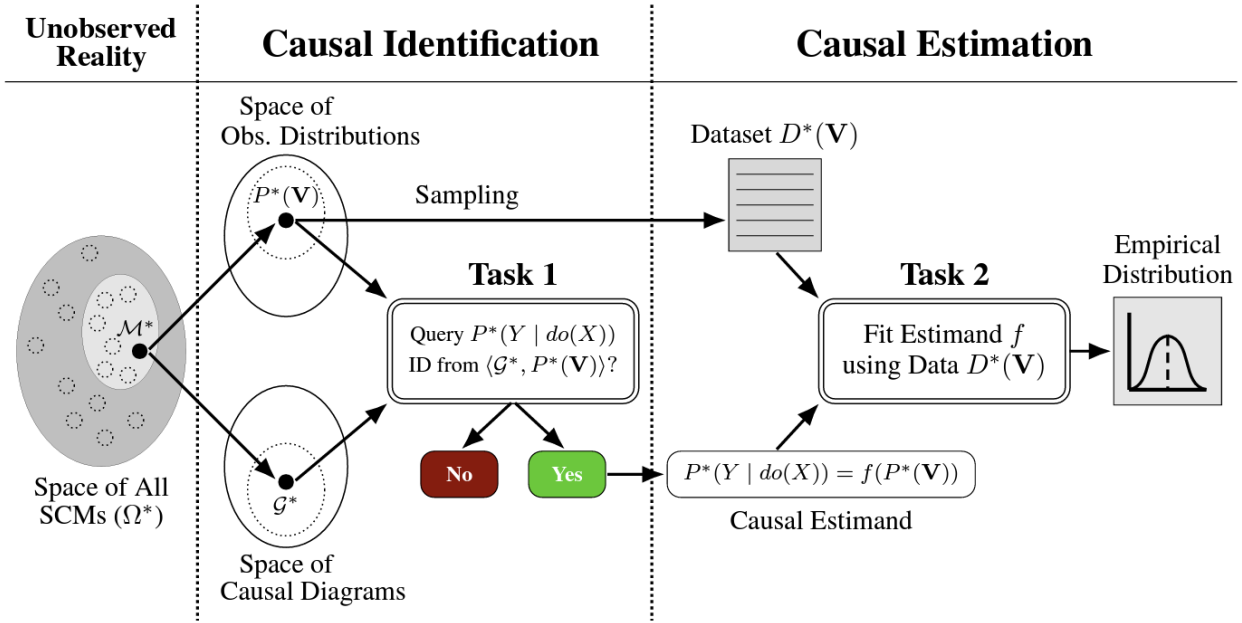
- Neural Network -> universal approximators
- Neural Networks should be able to learn any SCM

- Unsurprisingly, not the case

Xia et al., The Causal-Neural Connection: Expressiveness, Learnability, and Inference, NeurIPS 2021

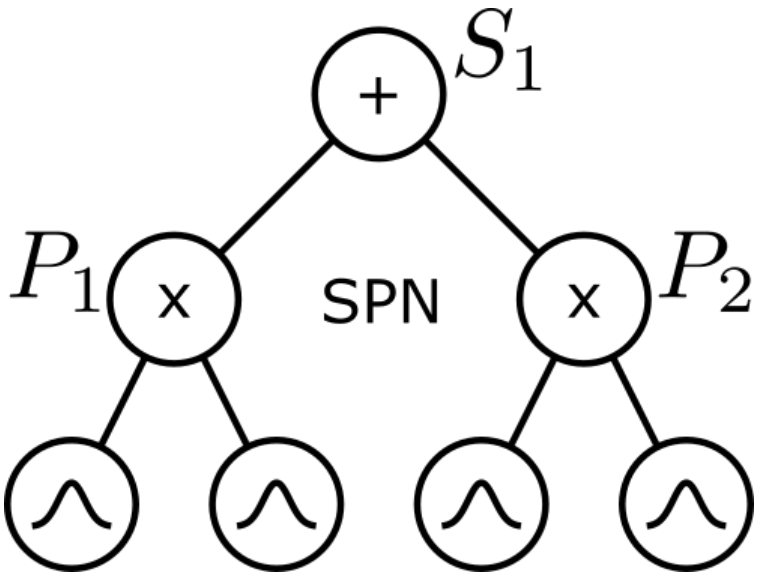
Not so simple after all….

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



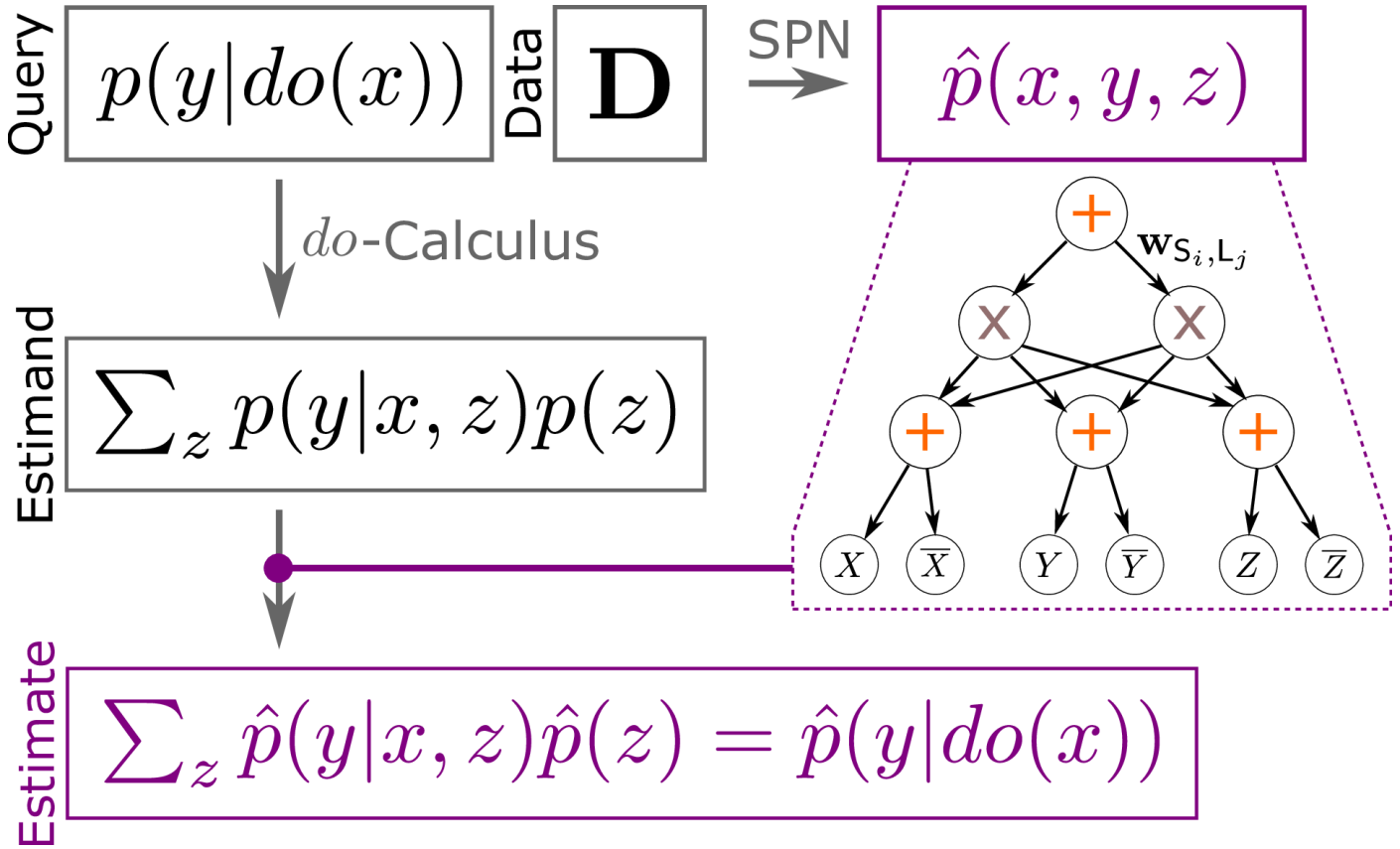
Sum-Product Networks (SPNs)

- 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

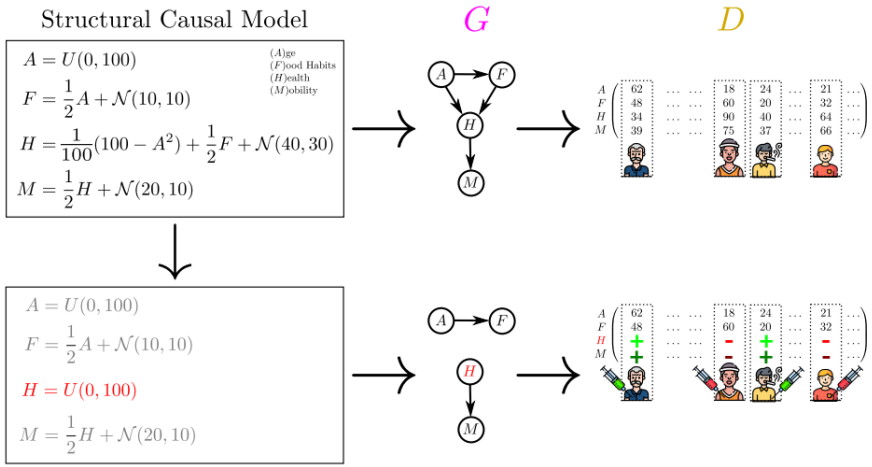


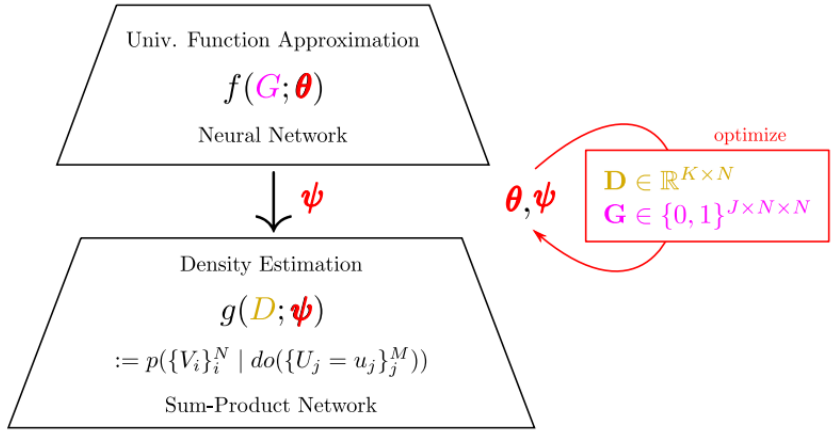
Poon and Domingos, Sum-Product Networks: A New Deep Architecture, UAI 2011

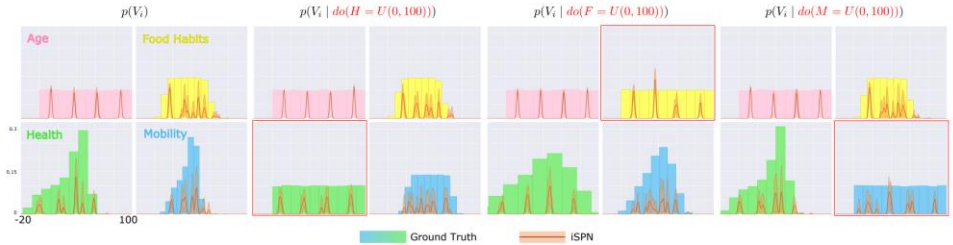
Vanilla SPN Causal Inference



Let's Stay at the Higher Rung: Interventional SPNs



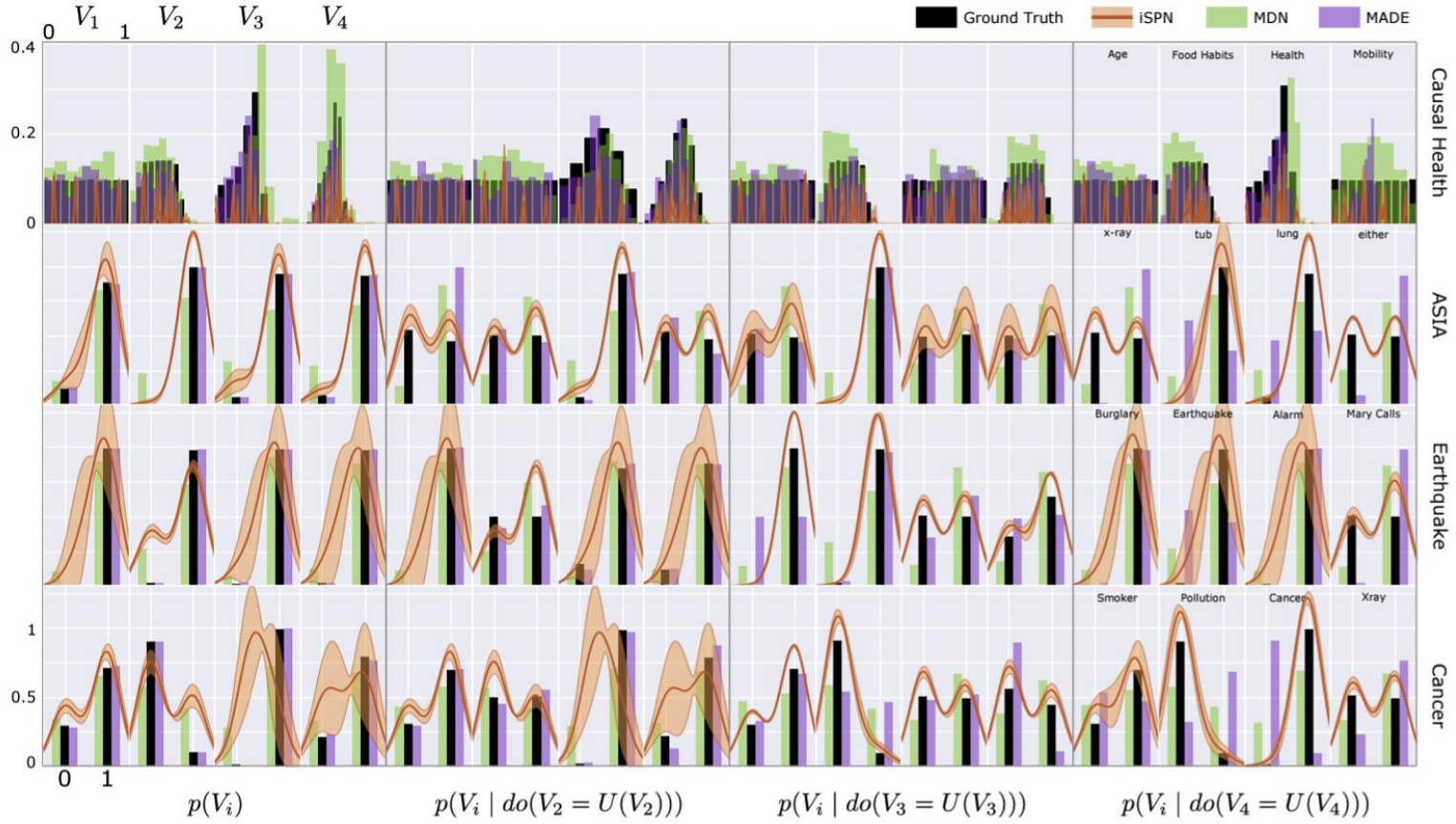




Confounding	Conditioning	Ground Truth[1]	CausalML	DoWhy	iSPN	
No	0.0374	0.0397 (0.04)	0.0397	0.0397	0.0347 ✓	ATE(asia, tub)
No	0.9271	0.9337 (0.93342)*	0.9337	0.9337	‡0.9139 ✓	ATE(Burglary, Alarm)
Yes	0.6766	0.6703 (0.667586)	0.6703	0.6697	0.6551 ✓	ATE(bronc, dysp)
Yes	−0.0457	0.0537 (0.05)	−0.0454	0.0538	0.0545 ✓	ATE(Surgery, Recovery)

Zečević et al., Interventional Sum-Product Networks: Causal Inference with Tractable Probabilistic Models, NeurIPS 2021

Show Me It Works



Show Me It Works

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

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.

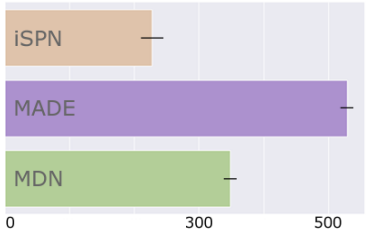
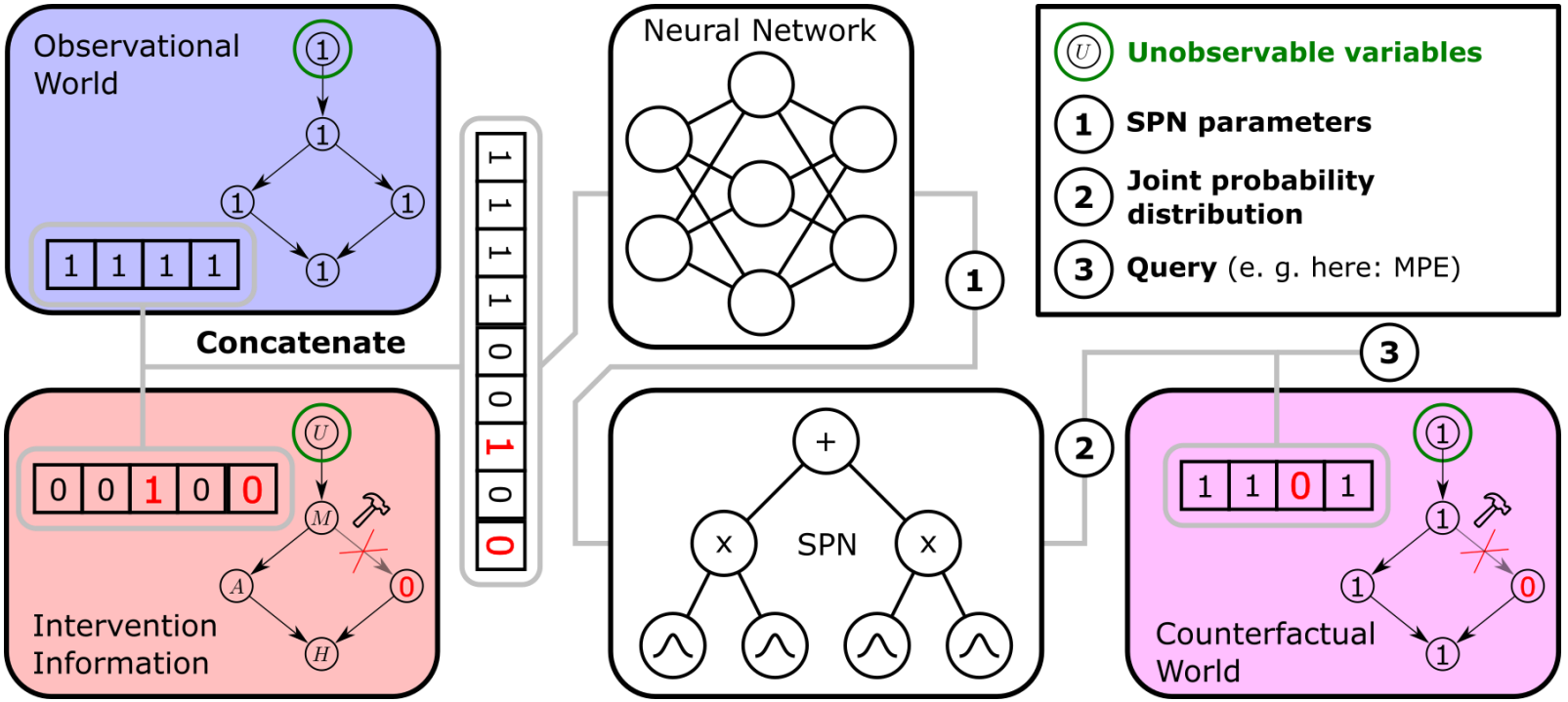


Figure 3: **Mean Running Times in sec. till convergence (Causal Health)** for 50 full passes. More data sets results in supplementary.

Counterfactual Sum-Product Networks



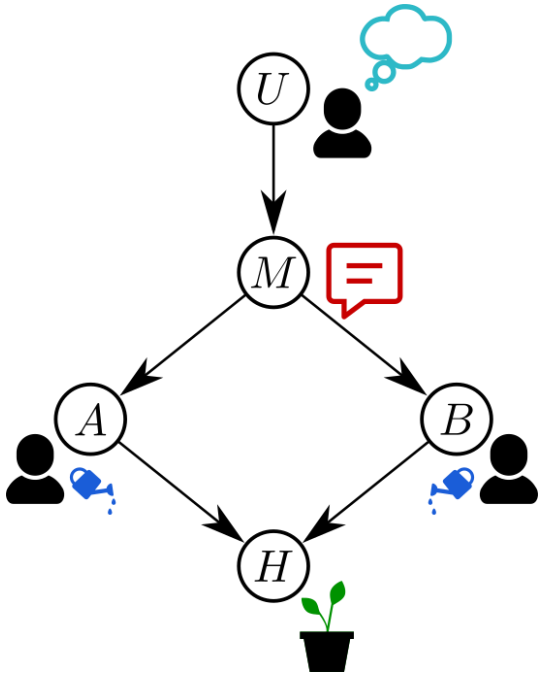
Busch et al., Computing Counterfactuals using Sum-Product Networks, Fundamental Challenges in Causality Colloquium 2023

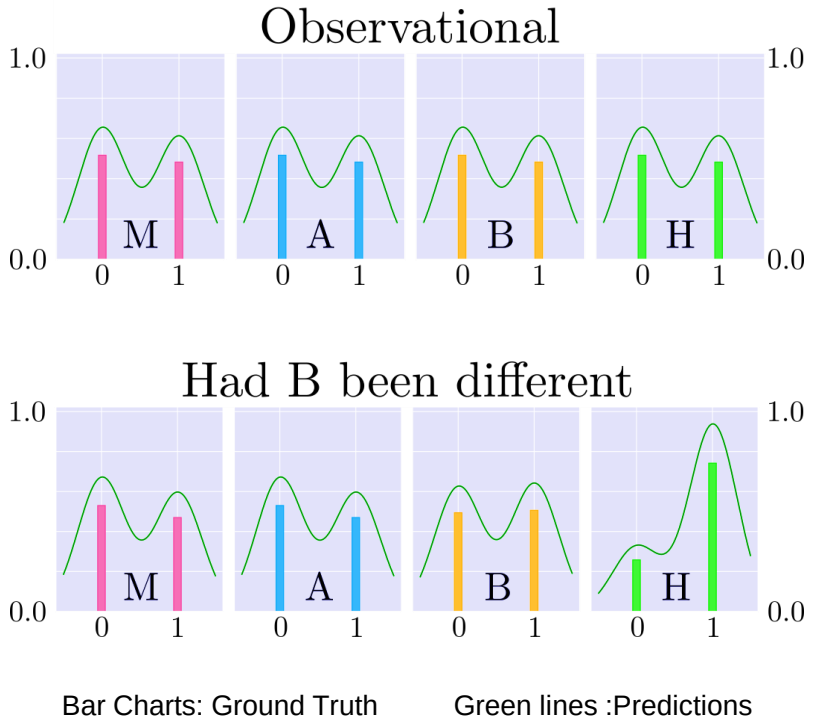
More formally

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

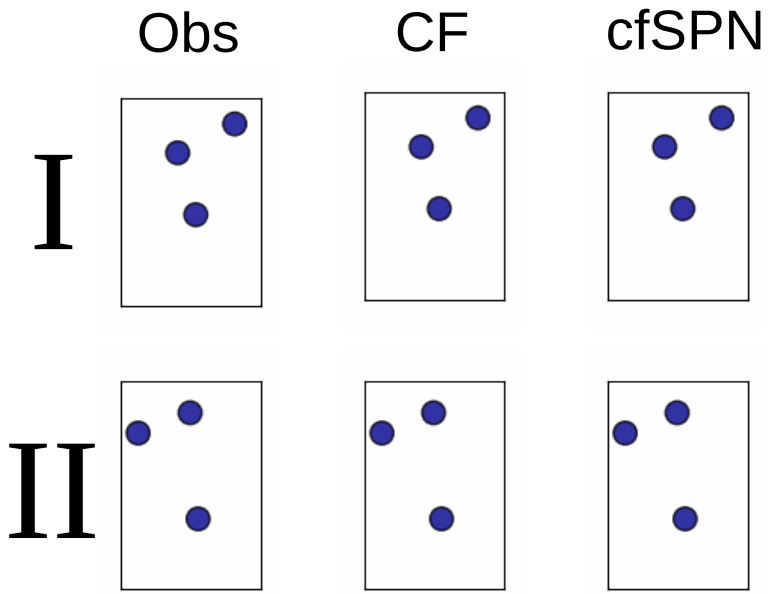




Bar Charts: Ground Truth Green lines :Predictions

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

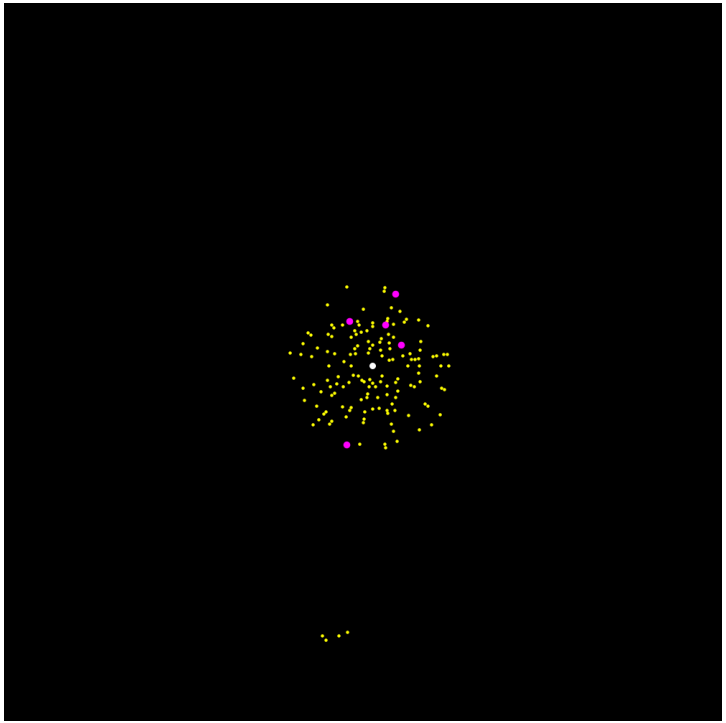


Based on the code repository for particle simulation:
https://github.com/ineporozhnii/particles_in_a_box

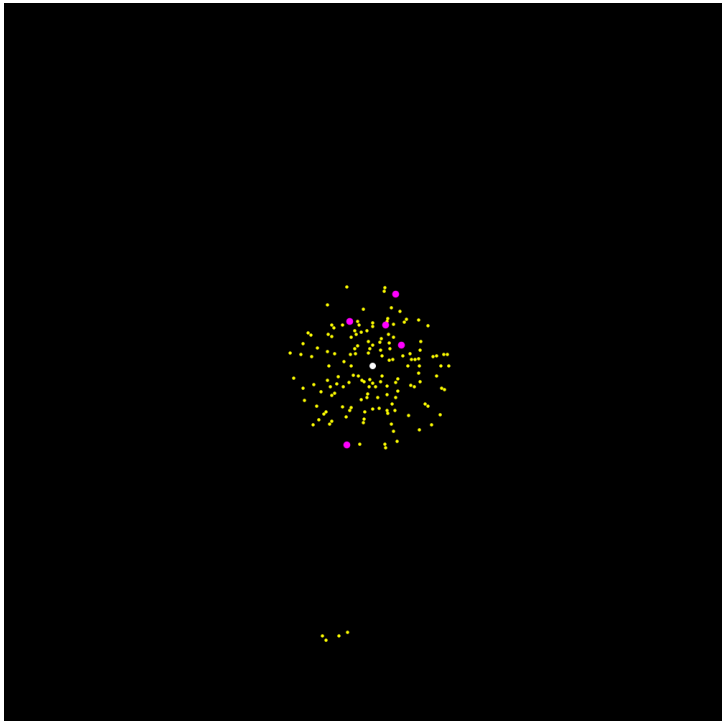
Experiment: Galaxies Collision

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

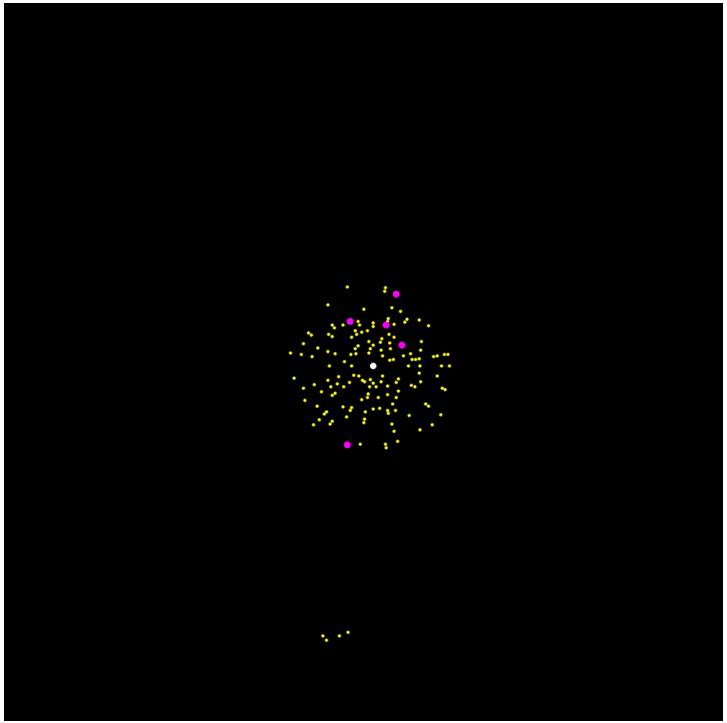
- Scaling is all you need?



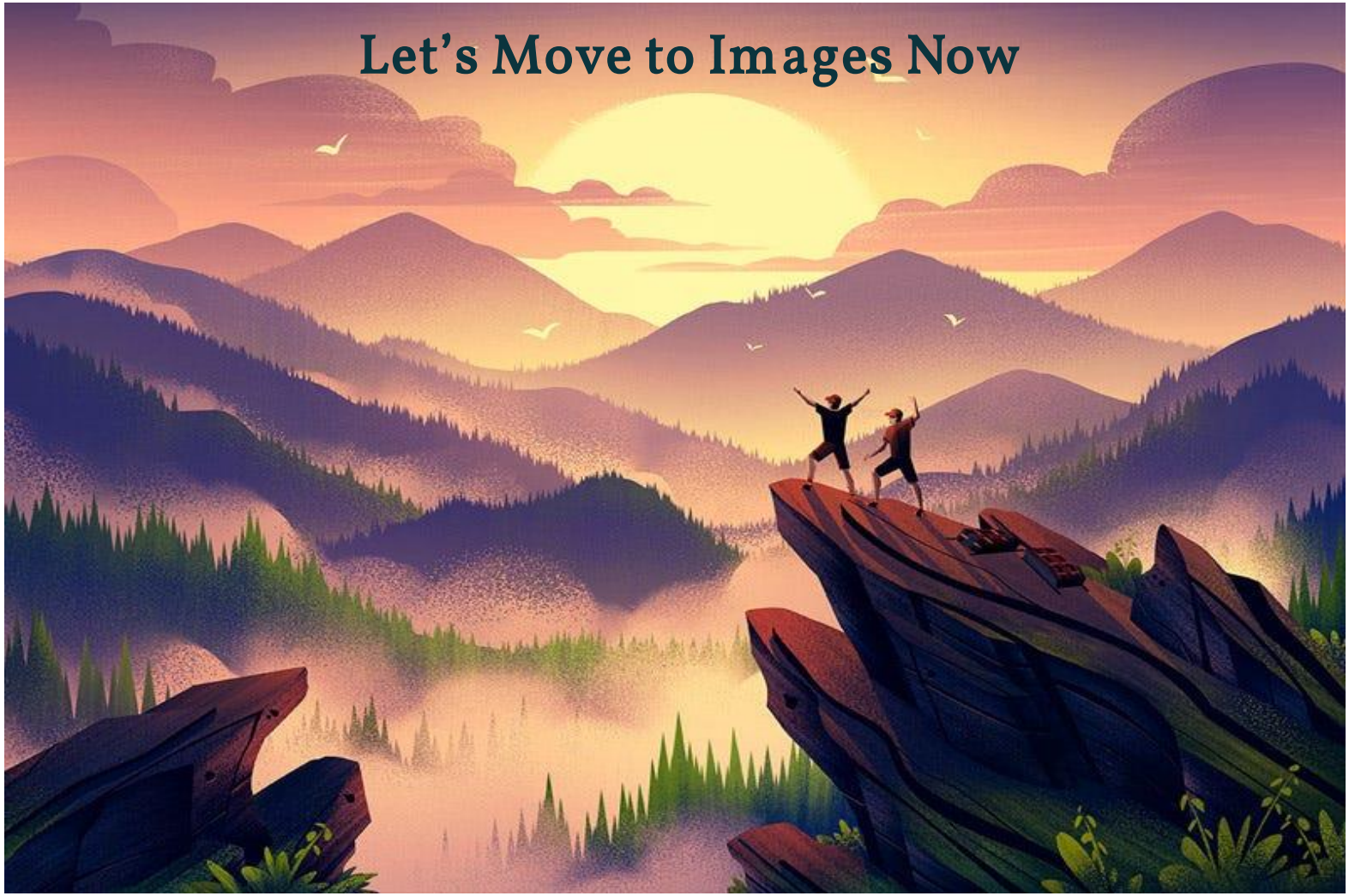
Counterfactual



counterfactualSPN

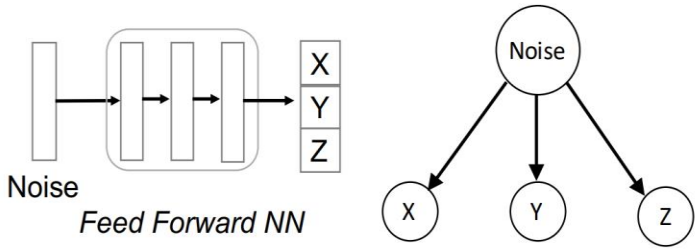


Let's Move to Images Now

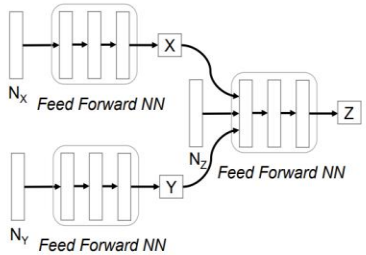


Causal Generative Adversarial Networks

- 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.



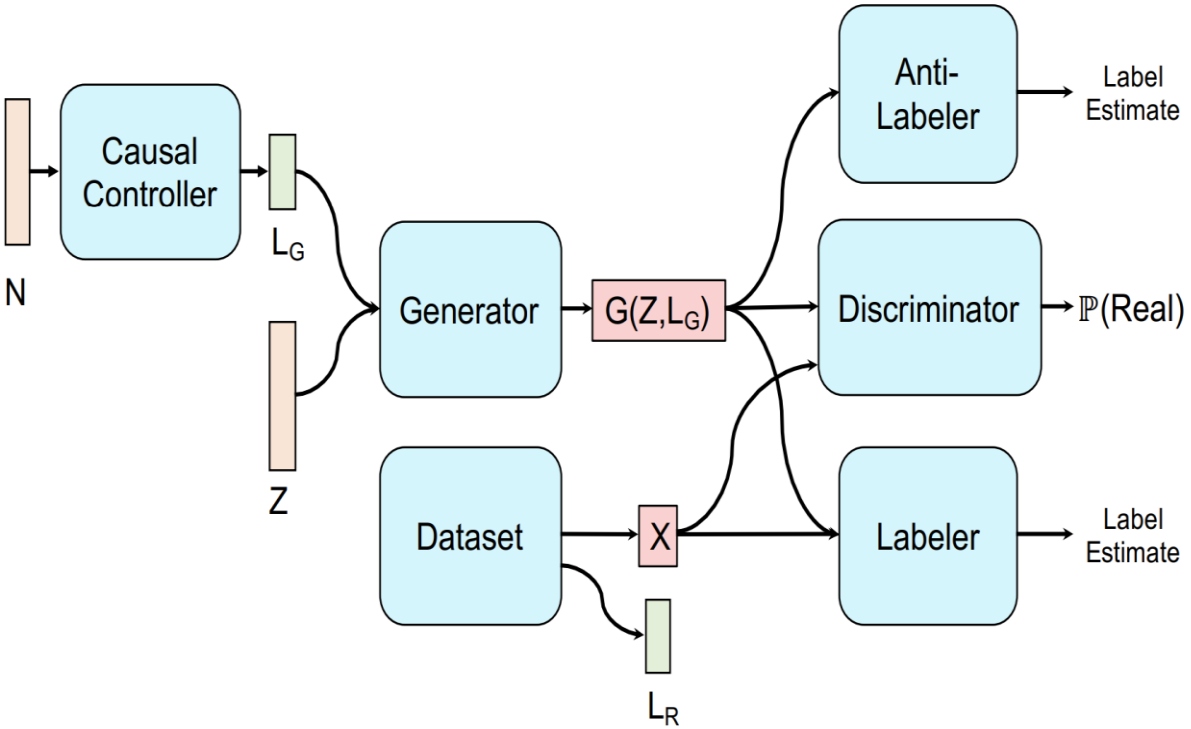
(a) Standard generator architecture and the causal graph it represents



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

Kocaoglu et al., CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training, ICLR 2018

CausalGAN: Architecture



CausalGAN: Loss Functions

- For a fixed generator

Anti- Labeler: $\max_{D_{LG}} \rho \mathbb{E}_{x \sim p_{\mathrm{g}}^{1}(x)} \left[\log(D_{LG}(x))\right] + (1-\rho)\mathbb{E}_{x \sim p_{\mathrm{g}}^{0}(x)} \left[\log(1 - D_{LG}(x)\right]$

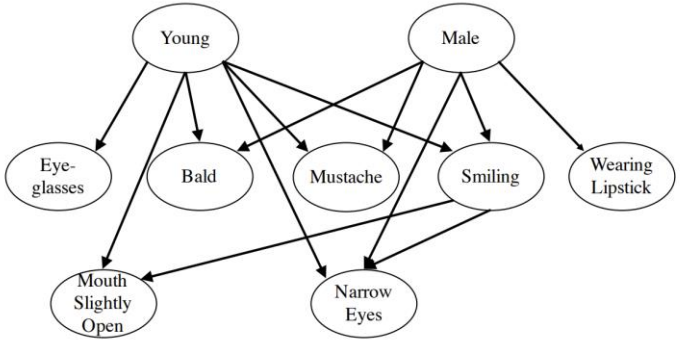
Labeler: $\max_{D_{LR}} \rho \mathbb{E}_{x \sim p_{\mathrm{data}}^{1}(x)} \left[\log(D_{LR}(x))\right] + (1-\rho)\mathbb{E}_{x \sim p_{\mathrm{data}}^{0}(x)} \left[\log(1 - D_{LR}(x)\right]$

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

- For a fixed labeler anti-labeler and discriminator

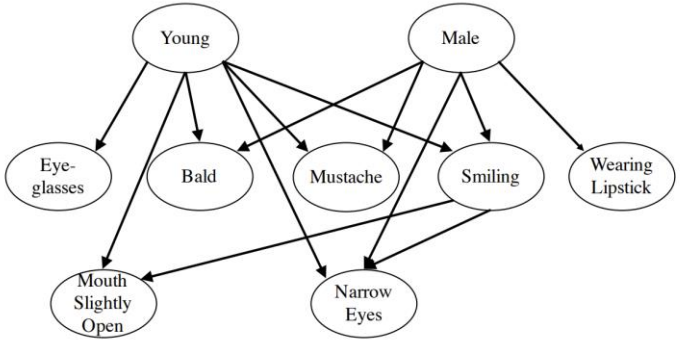
Generator $$\begin{aligned}\min_{G} \mathbb{E}_{x \sim p_{\mathrm{data}}(x)} \left[\log(D(x))\right] + \mathbb{E}_{x \sim p_{\mathrm{g}}(x)} \left[\log\left(\frac{1 - D(x)}{D(x)}\right)\right] \\ - \rho \mathbb{E}_{x \sim p_{g}^{1}(x)} \left[\log(D_{LR}(X))\right] - (1-\rho)\mathbb{E}_{x \sim p_{g}^{0}(x)} \left[\log(1 - D_{LR}(X))\right] \\ + \rho \mathbb{E}_{x \sim p_{g}^{1}(x)} \left[\log(D_{LG}(X))\right] + (1-\rho)\mathbb{E}_{x \sim p_{g}^{0}(x)} \left[\log(1 - D_{LG}(X))\right].\end{aligned}$$

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

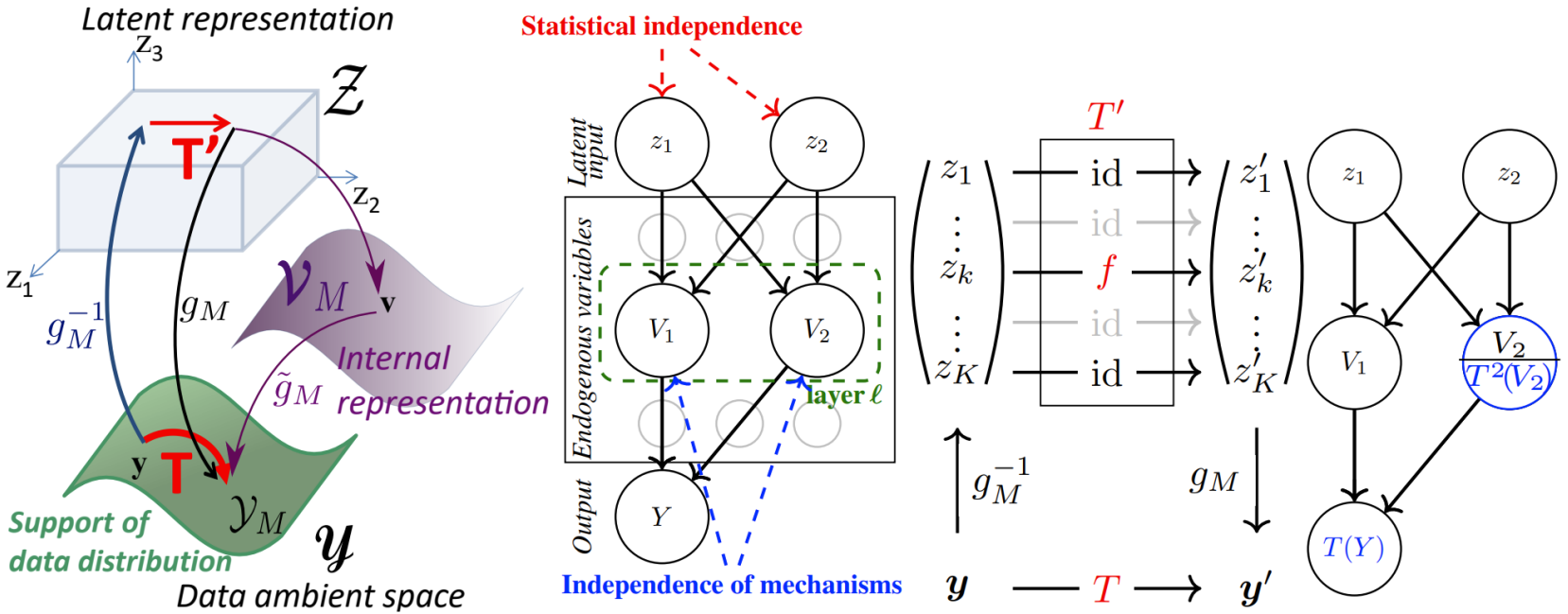
(a) Intervening vs Conditioning on Mouth Slightly Open, Top: Intervene Mouth Slightly Open=1, Bottom: Condition Mouth Slightly Open=1

Counterfactuals uncover the modular structure of DGMs

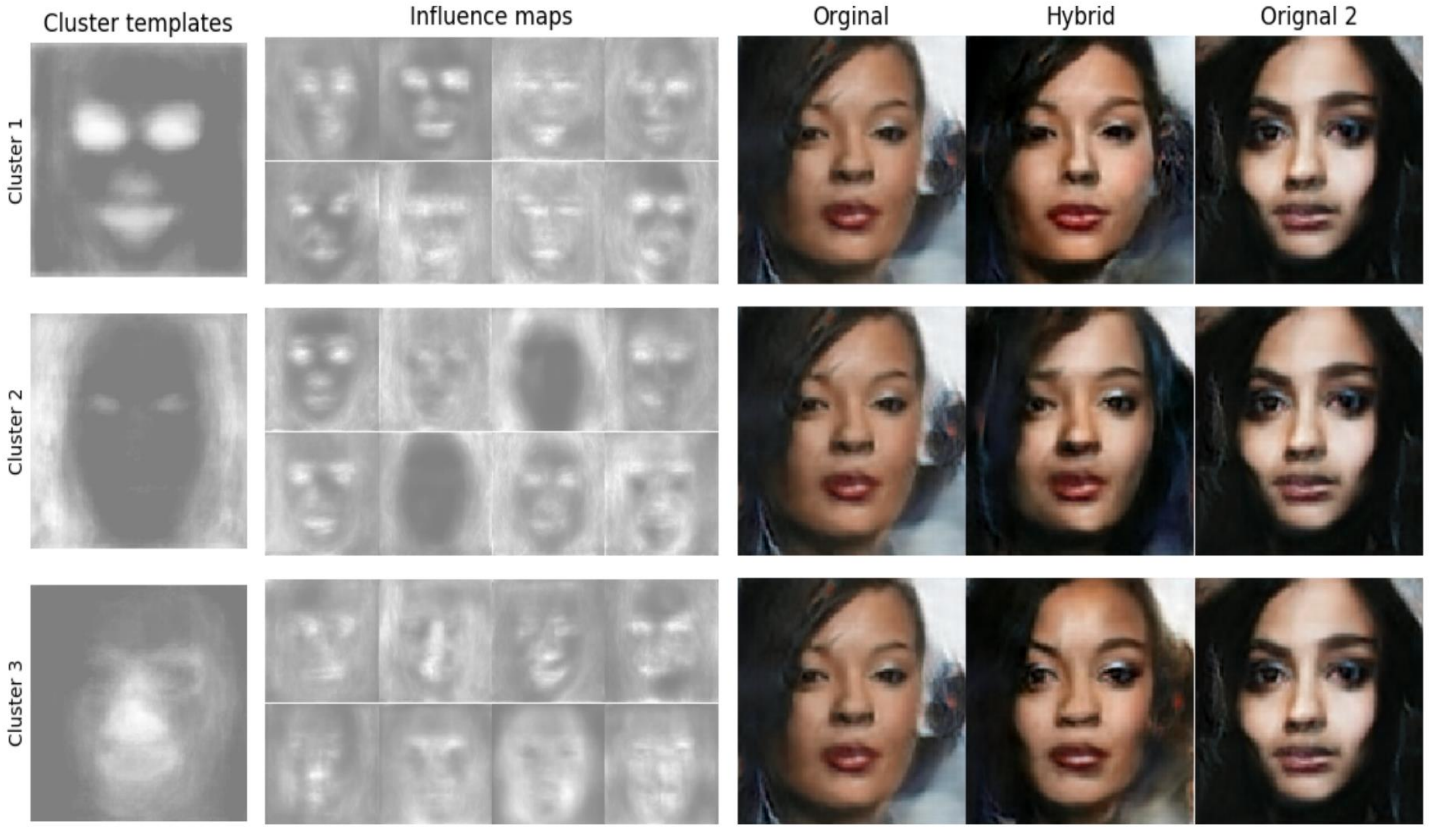
- Enables meaningful and controllable transformations in the data space without requiring explicit supervision
- **Challenge**: Deep generative models can provide latent representations of complex image datasets, but manipulating these representations for controlled transformations is challenging without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

Besserve et al., Counterfactuals uncover the modular structure of deep generative models, ICLR 2020

Causal Generative Models: Architecture



Causal Generative Models: Results



Still a Developing Field

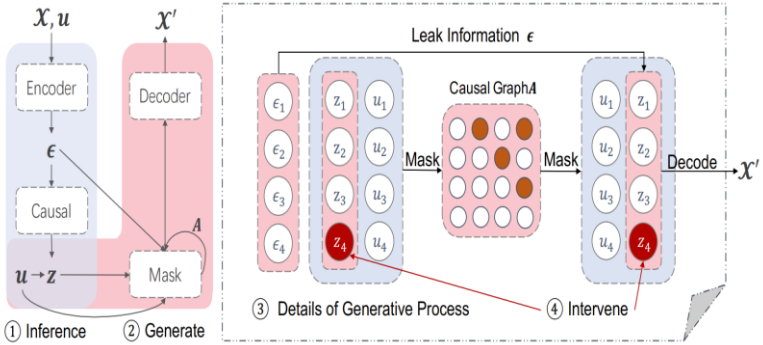


Figure 2. Model structure of CausalVAE. The encoder takes observation $\mathbf{x}$ as inputs to generate independent exogenous variable ϵ, whose prior distribution is assumed to be standard Multivariate Gaussian. Then it is transformed by the Causal Layer into causal representations $\mathbf{z}$ (Eq. 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

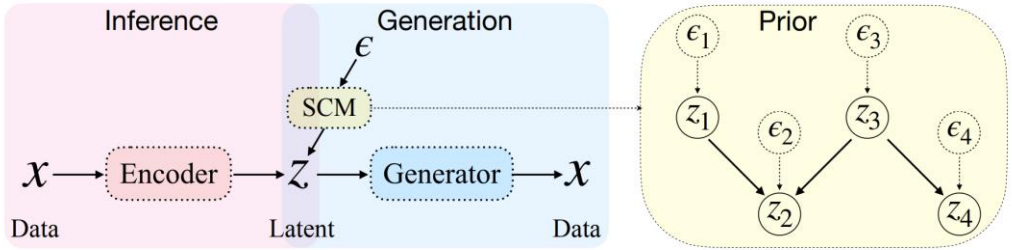


Figure 1: Model structure of a BGM (left) with an SCM prior (right).

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 ☺

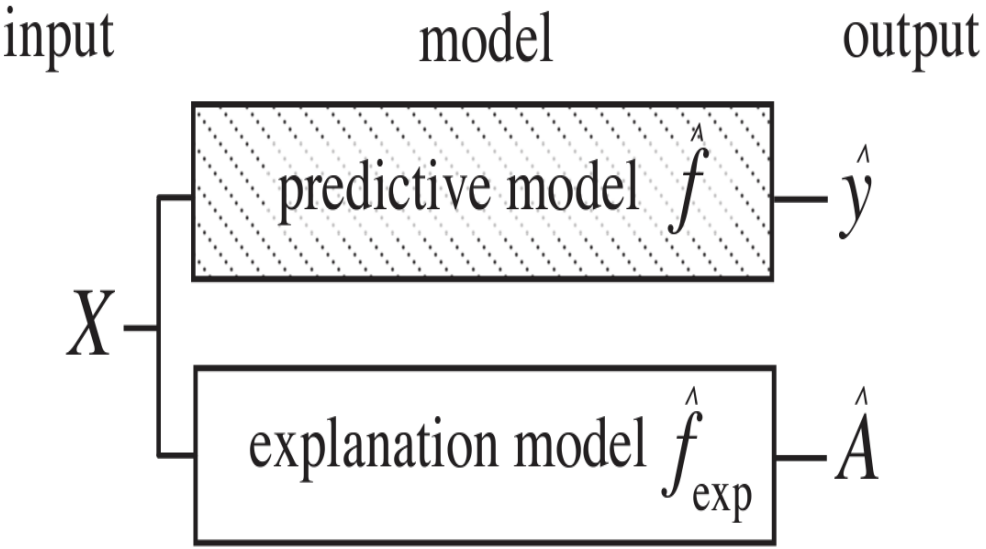
Causal Machine Learning

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 $A_F(\mathbf{x}) = (a_1, \ldots, a_n) \in \mathbb{R}^n$ is an attribution of predictive model F when a_i is the contribution of x_i for prediction $F(\mathbf{x})$ (with $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$). 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."

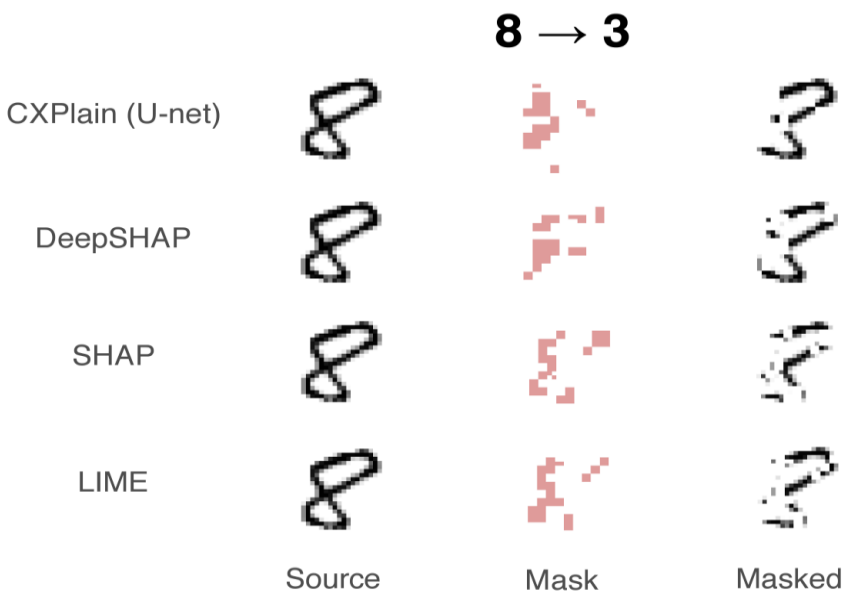
CXPlain

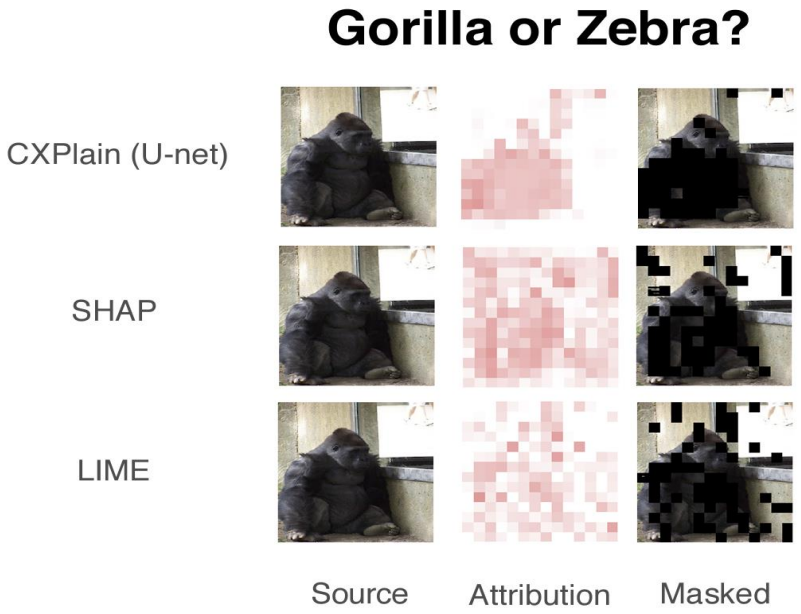


Main Contribution: A causal objective that optimizes explanation models to learn to explain another predictive model

Schwab and Karlen, CXPlain: Causal Explanations for Model Interpretation under Uncertainty, NeurIPS 2019

CXPlain: Results





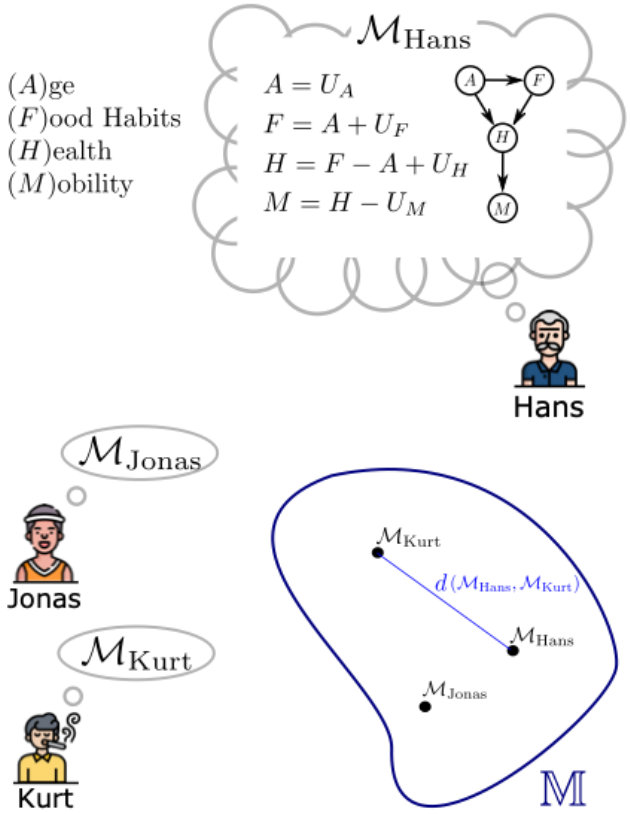
eXplainable Interactive Learning (XIL)

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

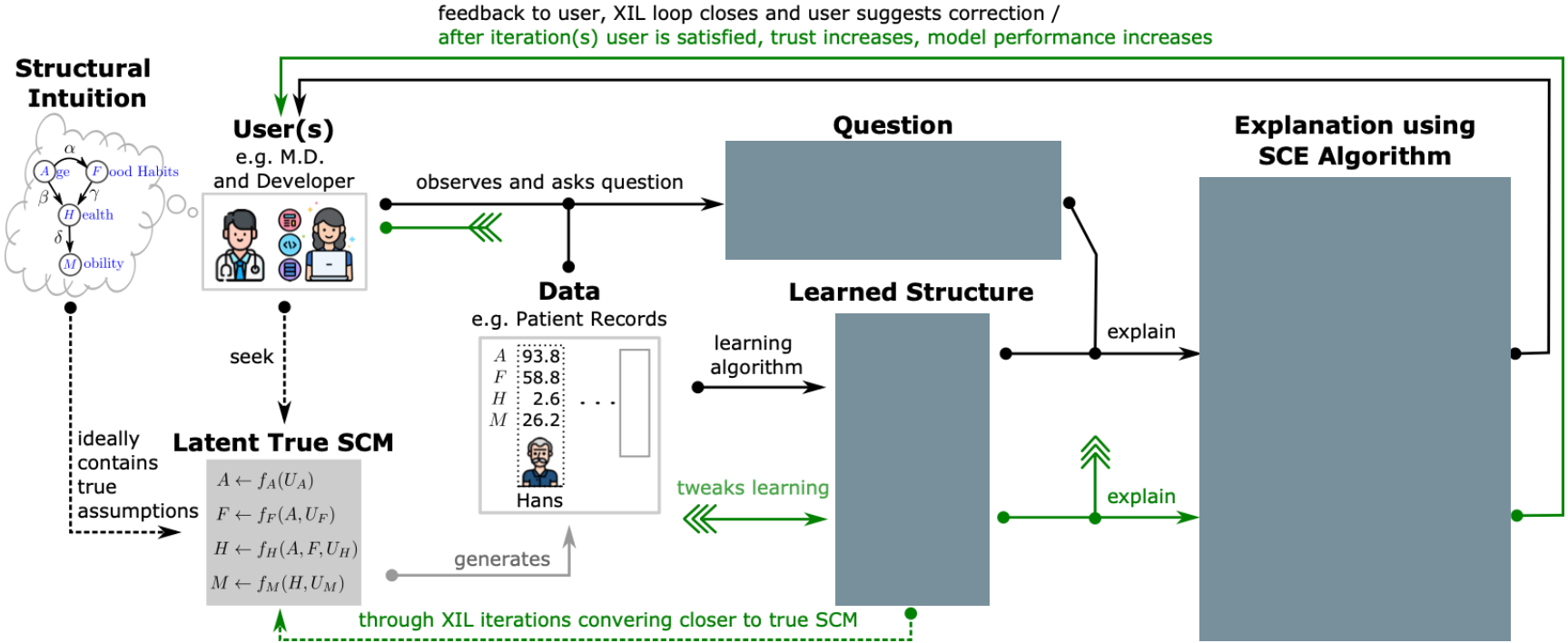
$f \leftarrow \textsc{Fit}(\mathcal{L})$
repeat
  $x \leftarrow \textsc{SelectQuery}(f, \mathcal{U})$
  $\hat{y} \leftarrow f(x)$
  $\hat{z} \leftarrow \textsc{Explain}(f, x, \hat{y})$
  Present x, $\hat{y}$, and $\hat{z}$ to the user
  Obtain y and explanation correction $\mathcal{C}$
  $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c} \leftarrow \textsc{ToCounterExamples}(\mathcal{C})$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x, y)\} \cup \{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c}$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus (\{x\} \cup \{\bar{x}_i\}_{i=1}^{c})$
  $f \leftarrow \textsc{Fit}(\mathcal{L})$
until budget T is exhausted or f is good enough
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 μ^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 allows 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 $\alpha_{X \to Y}$ 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 influental 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* $C_{XY} := (\alpha_{X \to Y}, x, y, \mu^X, \mu^Y)$ *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* $\mathcal{Z}_X = \{|\alpha_{Z \to X}| : Z \in \mathrm{Pa}_X\}$ *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:* $\quad R_1(s(\alpha_{X \to Y}), 0) \wedge (R_2(y, \mu^Y) \vee R_1(x, \mu^X))$,

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

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

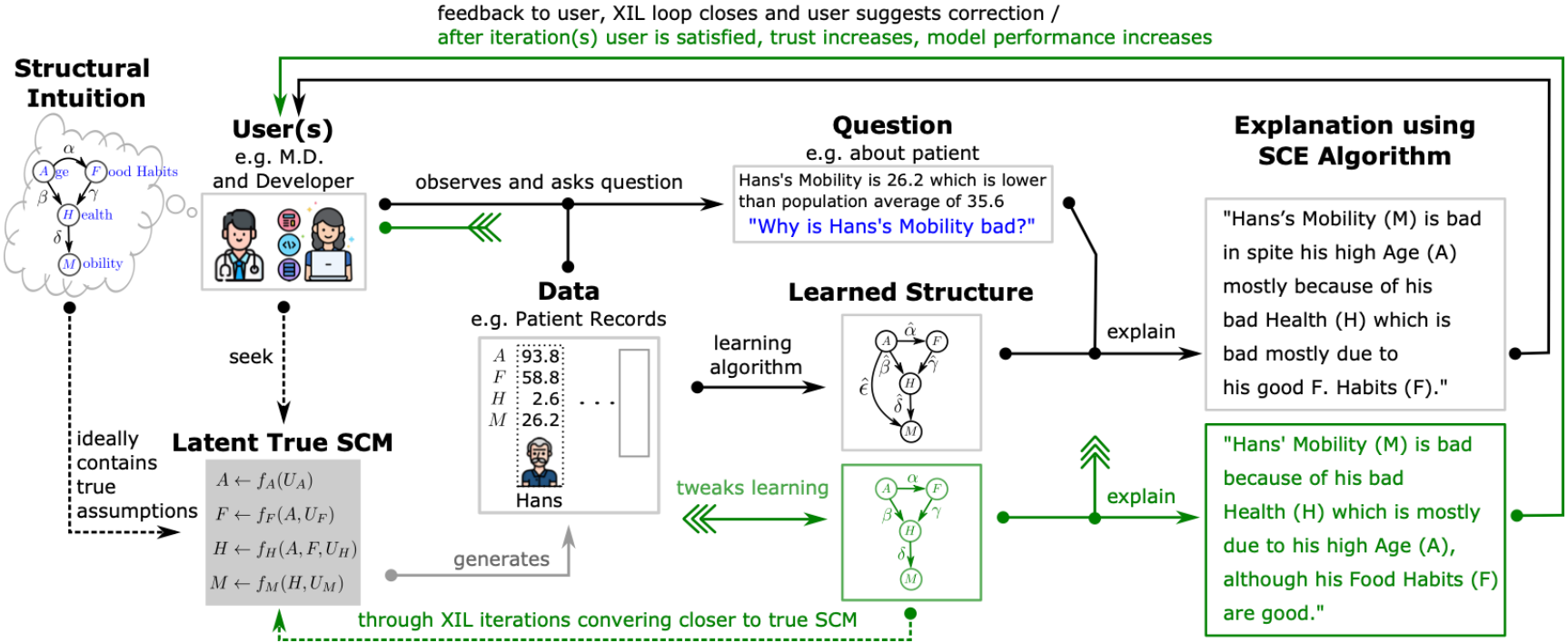
indicating for each rule ER$i(\cdot) \in \{-1, 0, 1\}$ *how the causal relation* $X \to Y$ *satisfies that rule.*

Pronouncing the Rules + Inspiration

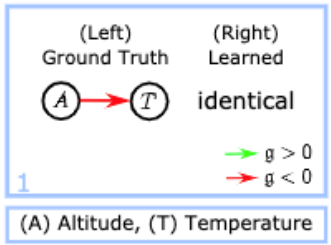
*ER*1	Excitation	"Y because of X [being low/high]"
*ER*2	Inhibition	"Y although X [is low/high]"
*ER*3	Preference	"mostly" + *ER*1 or *ER*2 pronunciation

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

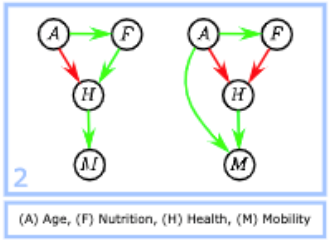
The Causal XIL Loop for SCE



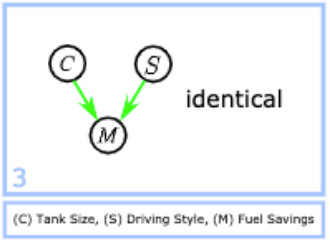
Quality of Learned Explanations



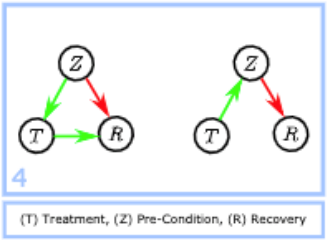
(A) Altitude, (T) Temperature



(A) Age, (F) Nutrition, (H) Health, (M) Mobility



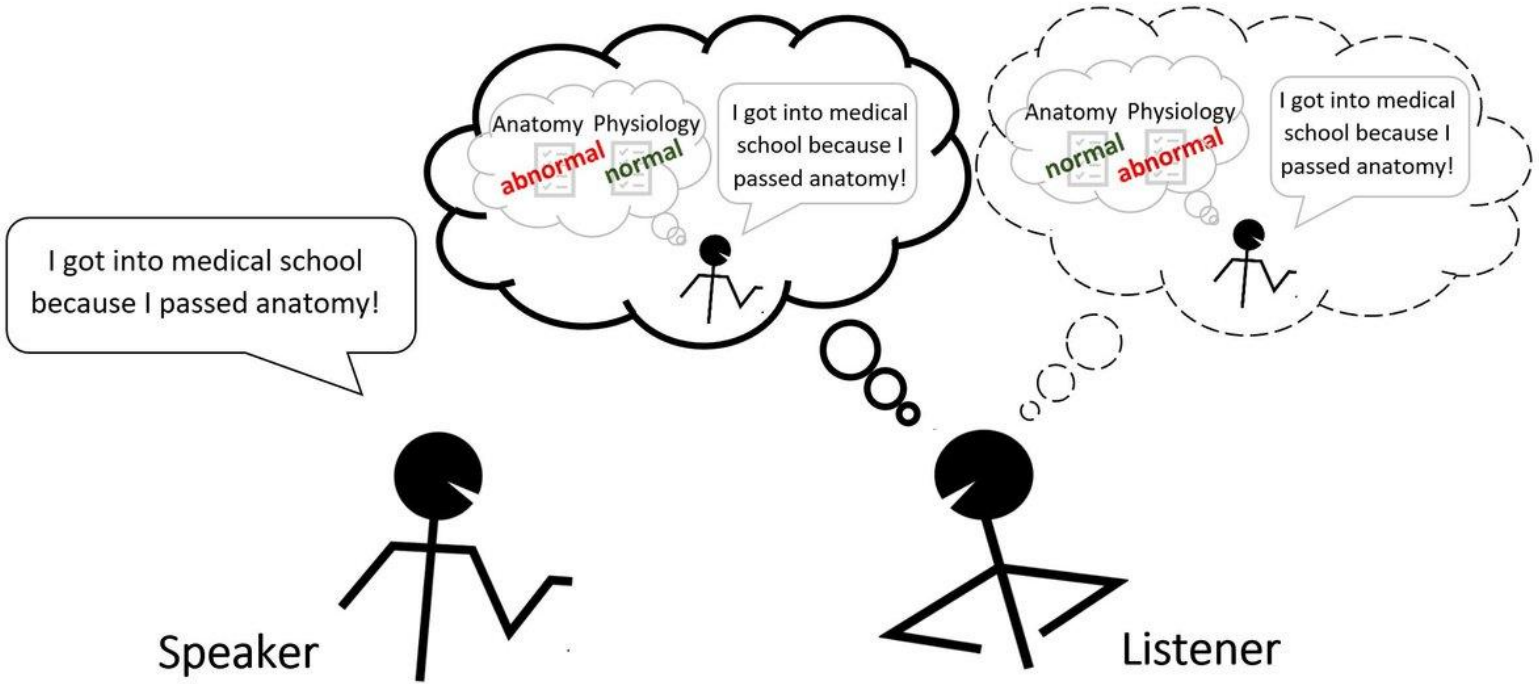
(C) Tank Size, (S) Driving Style, (M) Fuel Savings



(T) Treatment, (Z) Pre-Condition, (R) Recovery

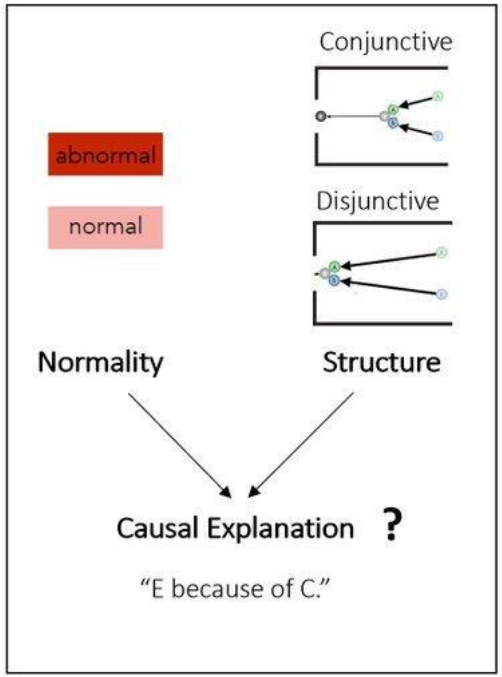
1	*"Why is the Temperature at the Matterhorn low?"*	(*Question*)
	"The Temperature at the Matterhorn is low because of the high Altitude."	(Ground Truth)
	"The Temperature at the Matterhorn is low because of the high Altitude."	**(Learned)**
2	*"Why is Hans's Mobility bad?"*	
	"Hans's Mobility is bad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

Normal vs Abnormal Causal Explanations

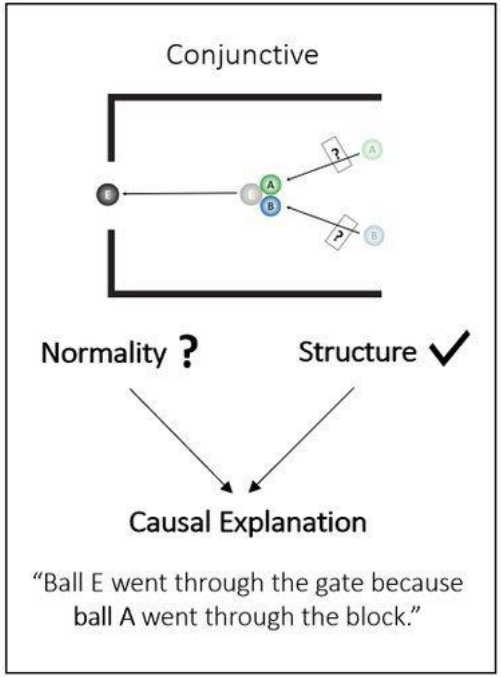


Kirfel et al., Inference from Explanation, Journal of Experimental Psychology: General 2022

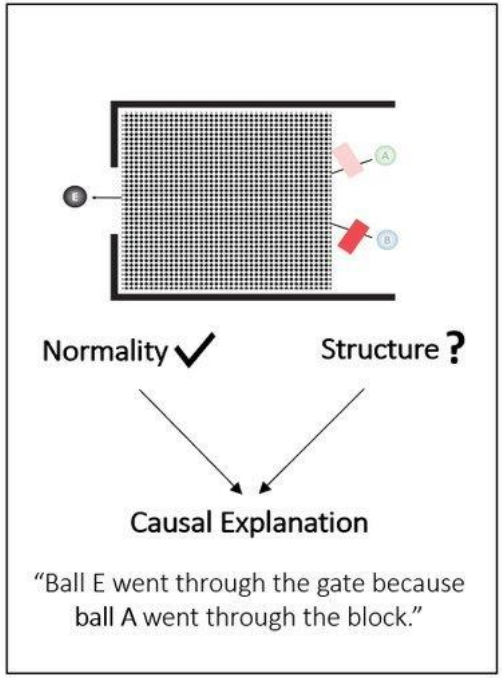
Normal vs Abnormal Causal Explanations



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

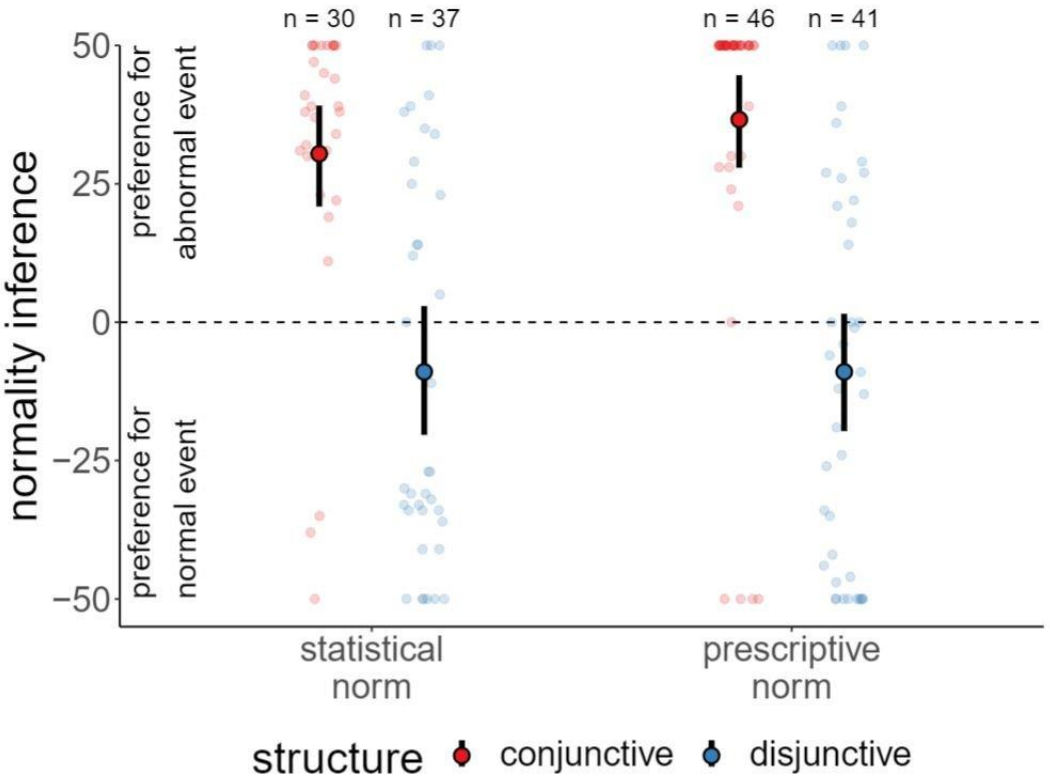


Hypothesis 2: Inference about normality from causal explanation and structure.



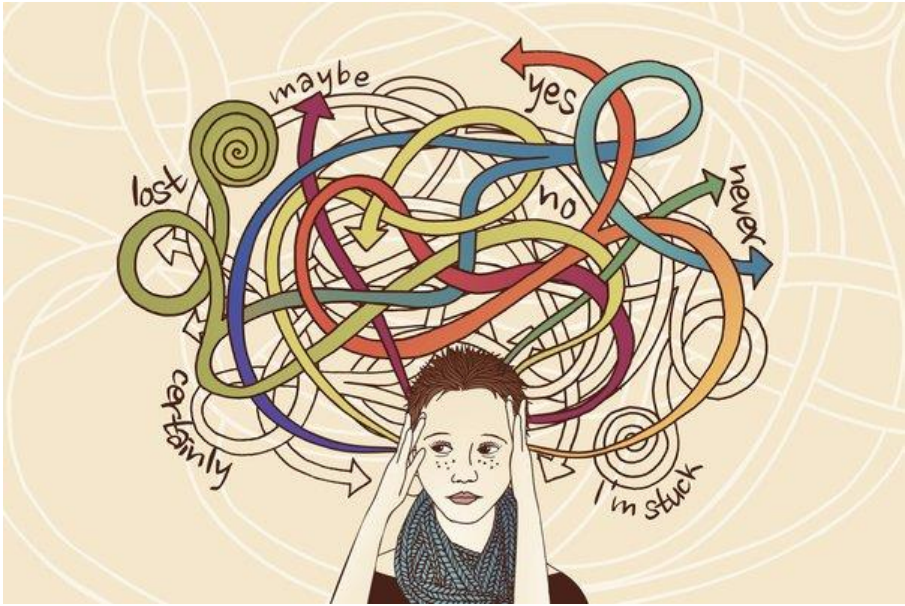
Hypothesis 3: Inference about structure from causal explanation and normality.

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

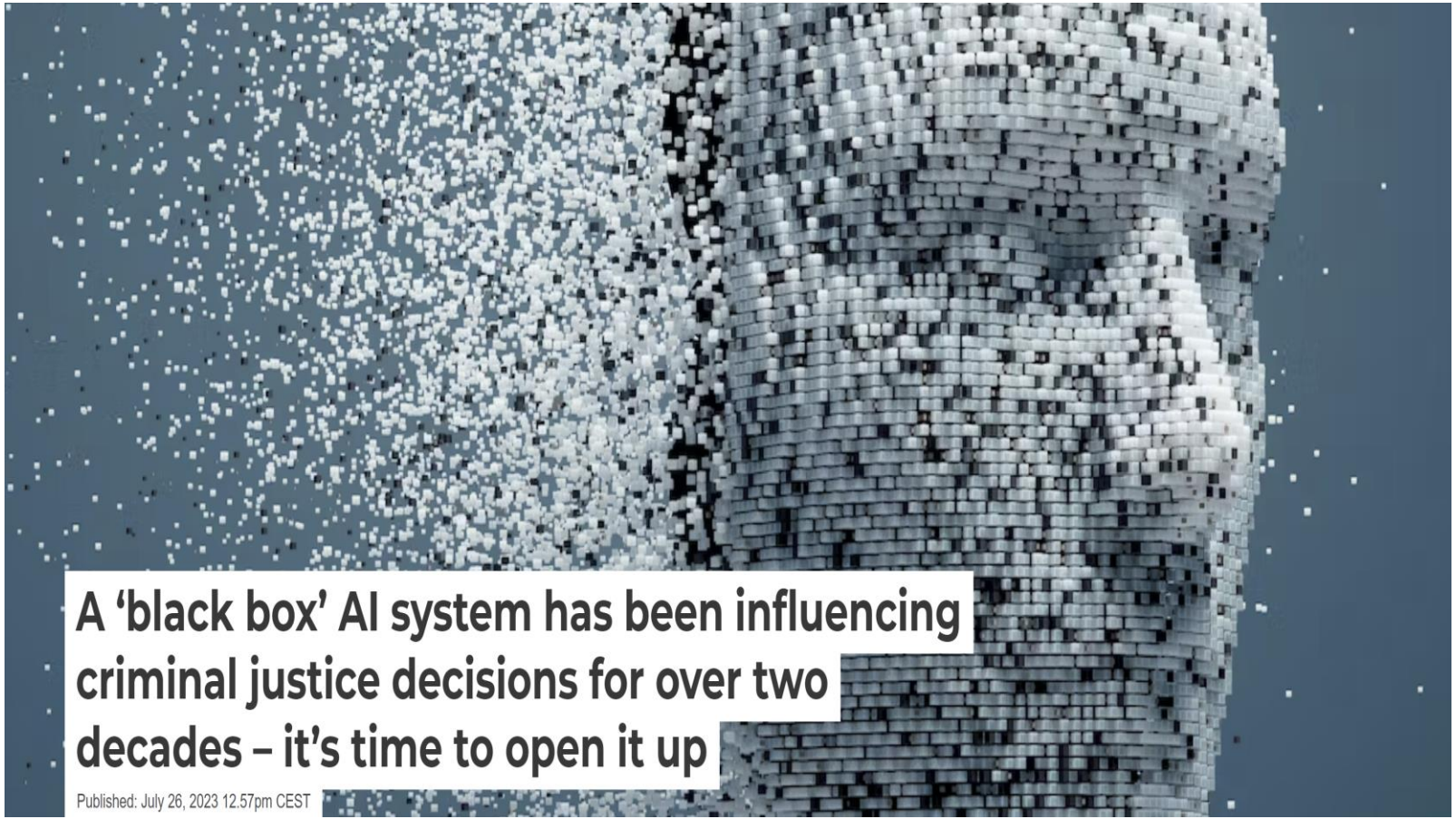
Causal Machine Learning

Explainability

Robustness

Causality

Fairness

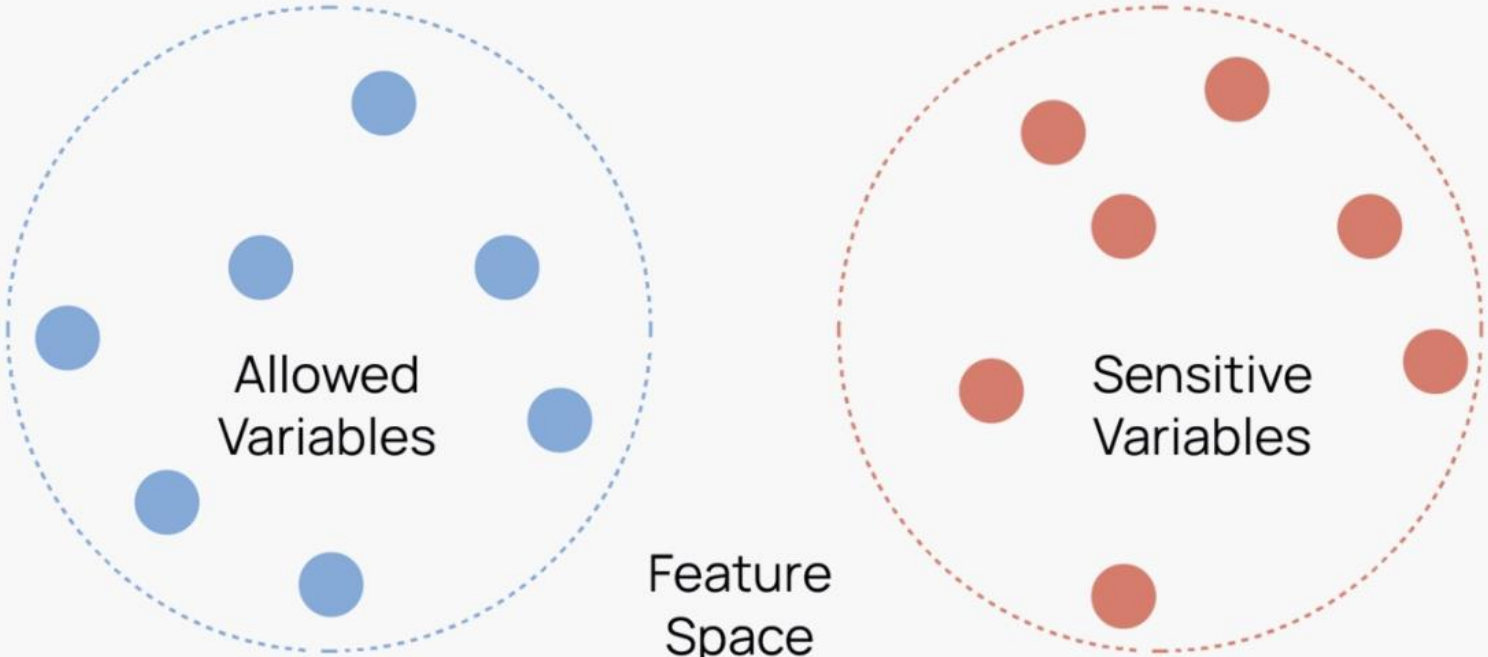


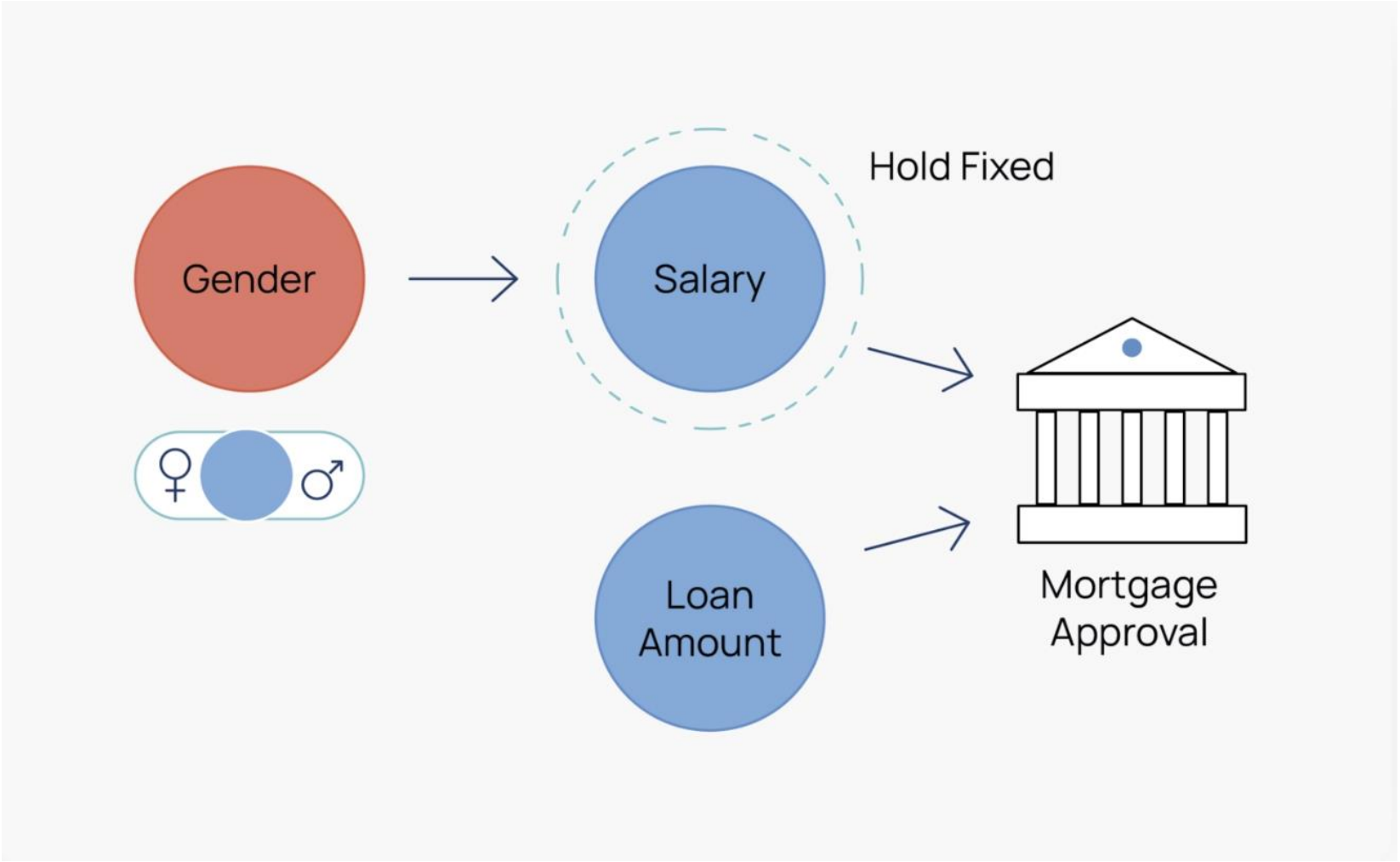
https://theconversation.com/a-black-box-ai-system-has-been-influencing-criminal-justice-decisions-for-over-two-decades-its-time-to-open-it-up-200594

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.

https://www.causalens.com/white-paper/certifying-ai-fairness/


Allowed
Variables
Sensitive
Variables
Feature
Space


Hold Fixed
Gender
Salary
Loan Amount
Mortgage Approval

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.

Carey and Wu., The Causal Fairness Field Guide: Perspectives From Social and Formal Sciences, Frontiers 2022

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.

Kusner et al., Counterfactual Fairness, NeurIPS 2017

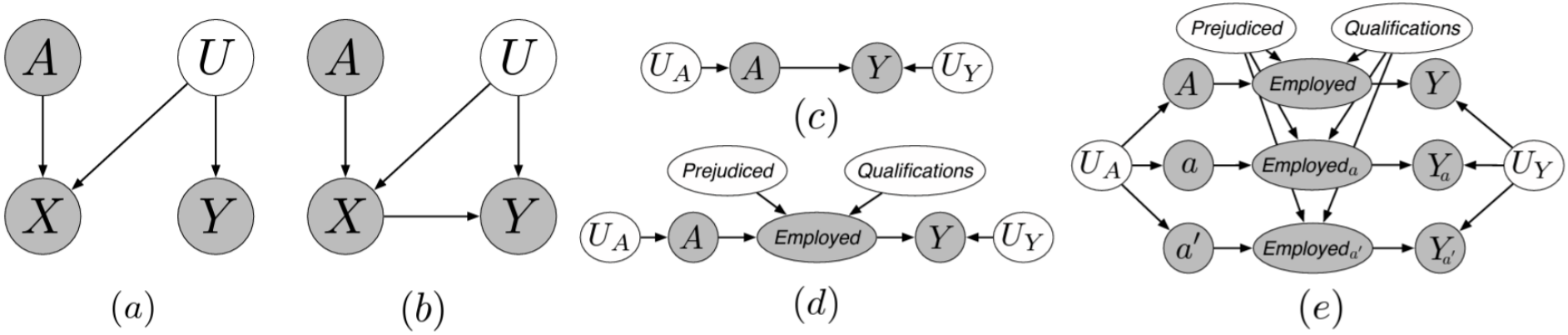
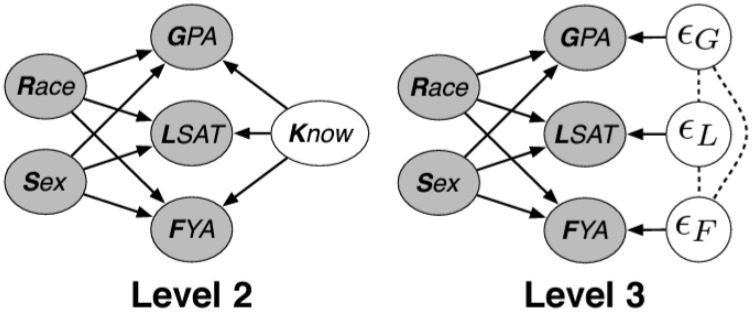
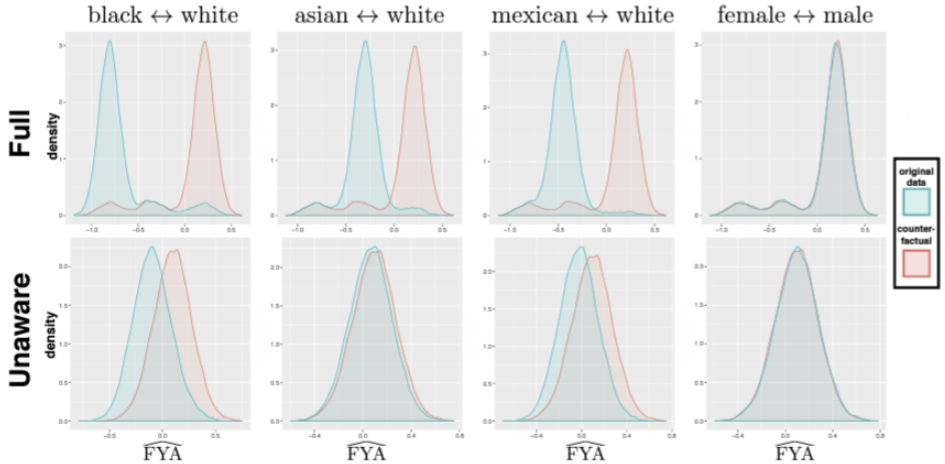


Figure 1: (a), (b) Two causal models for different real-world fair prediction scenarios. See Section 3.1 for discussion. (c) The graph corresponding to a causal model with A being the protected attribute and Y some outcome of interest, with background variables assumed to be independent. (d) Expanding 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

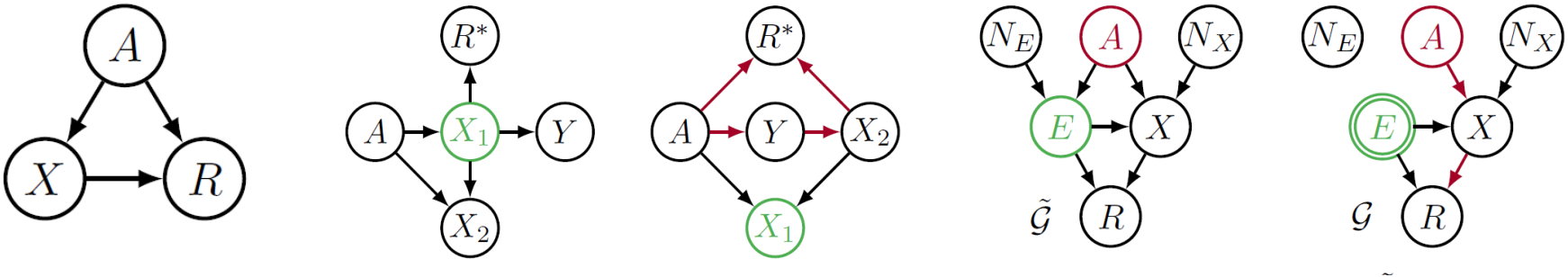




- Full: all attributes; Unaware: no protected attribute
- Blue distribution is density of predicted attribute for the original data and the red distribution is this density for the counterfactual data
- Models are unfair

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?”



Kilbertus et al., Avoiding Discrimination through Causal Reasoning, NeurIPS 2017

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 JAKE.FAWKES@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Robin J. Evans EVANS@STATS.OX.AC.UK
Department of Statistics, University of Oxford

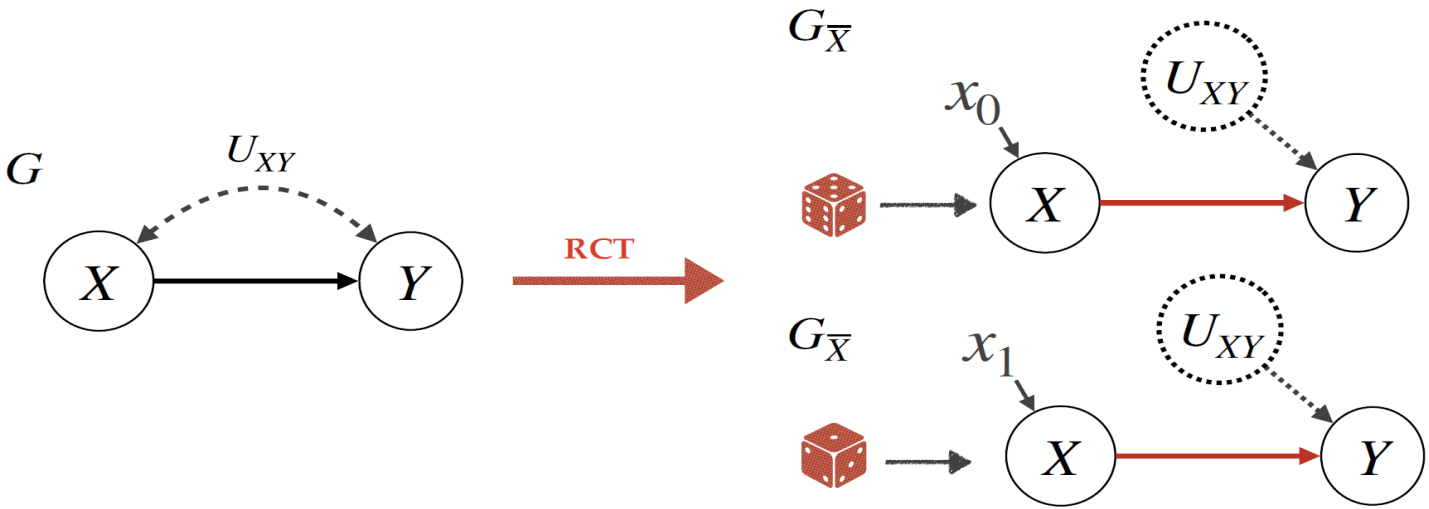
Dino Sejdinovic DINO.SEJDINOVIC@STATS.OX.AC.UK
Department of Statistics, University of Oxford

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

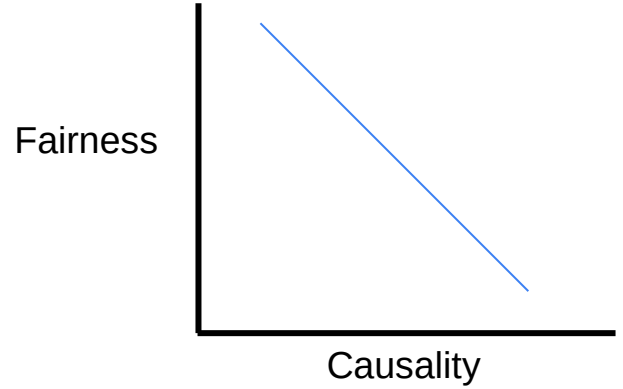


Causality + Fairness and else

Causal Conceptions of Fairness and their Consequences

Hamed Nilforoshan [*1] Johann Gaebler [*1] Ravi Shroff [2] Sharad Goel [3]

ICML Best Paper Award 2022



(**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

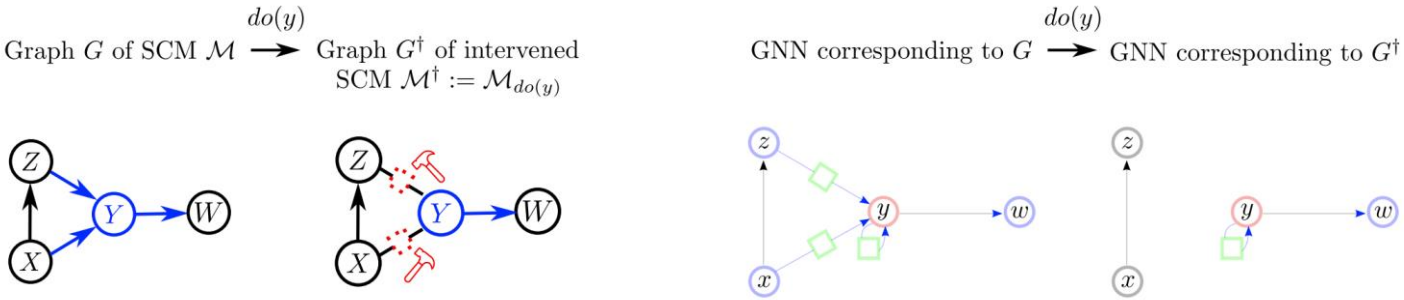
Judea Pearl
@yudapearl

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
> 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



Zečević et al., Interventions in Graph Neural Networks Lead to New Neural Causal Models (Under Review)

Intervention akin to SCM within the GNN computation rule

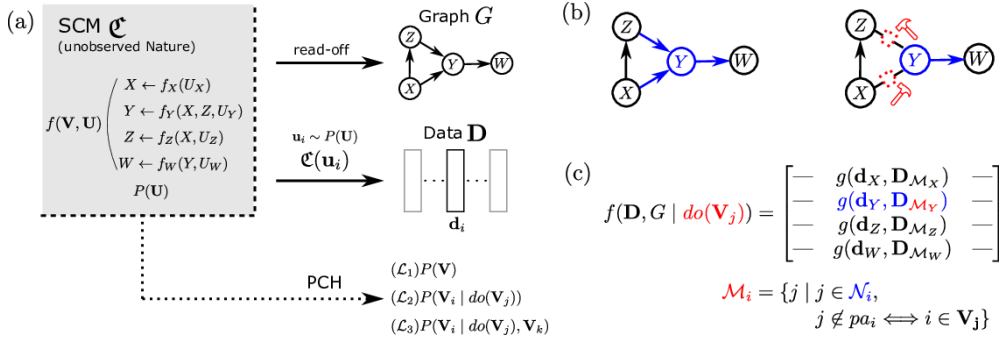
Definition 1. (Interventions within GNN.) An intervention $\mathbf{x}$ on the corresponding set of variables $\mathbf{X} \subseteq \mathbf{V}$ within a GNN layer $f(\mathbf{D}, \mathbf{A}_G)$, denoted by $f(\mathbf{D}, \mathbf{A}_G | do(\mathbf{X} = \mathbf{x}))$, is defined as a modified layer computation,

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

where the intervened local neighborhood is given by

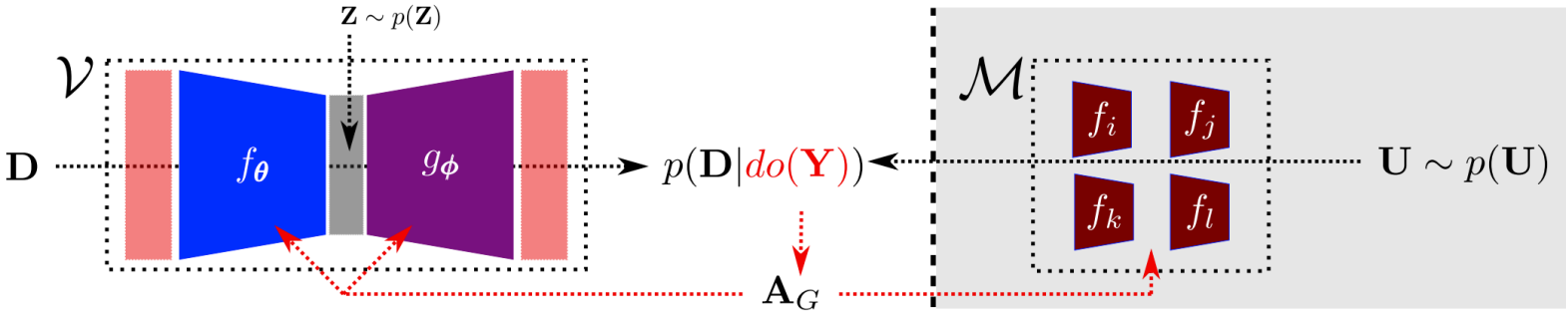
$$\mathcal{M}_i^G = \{j \mid j \in \mathcal{N}_i^G, j \notin pa_i \iff i \in \mathbf{X}\}$$

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



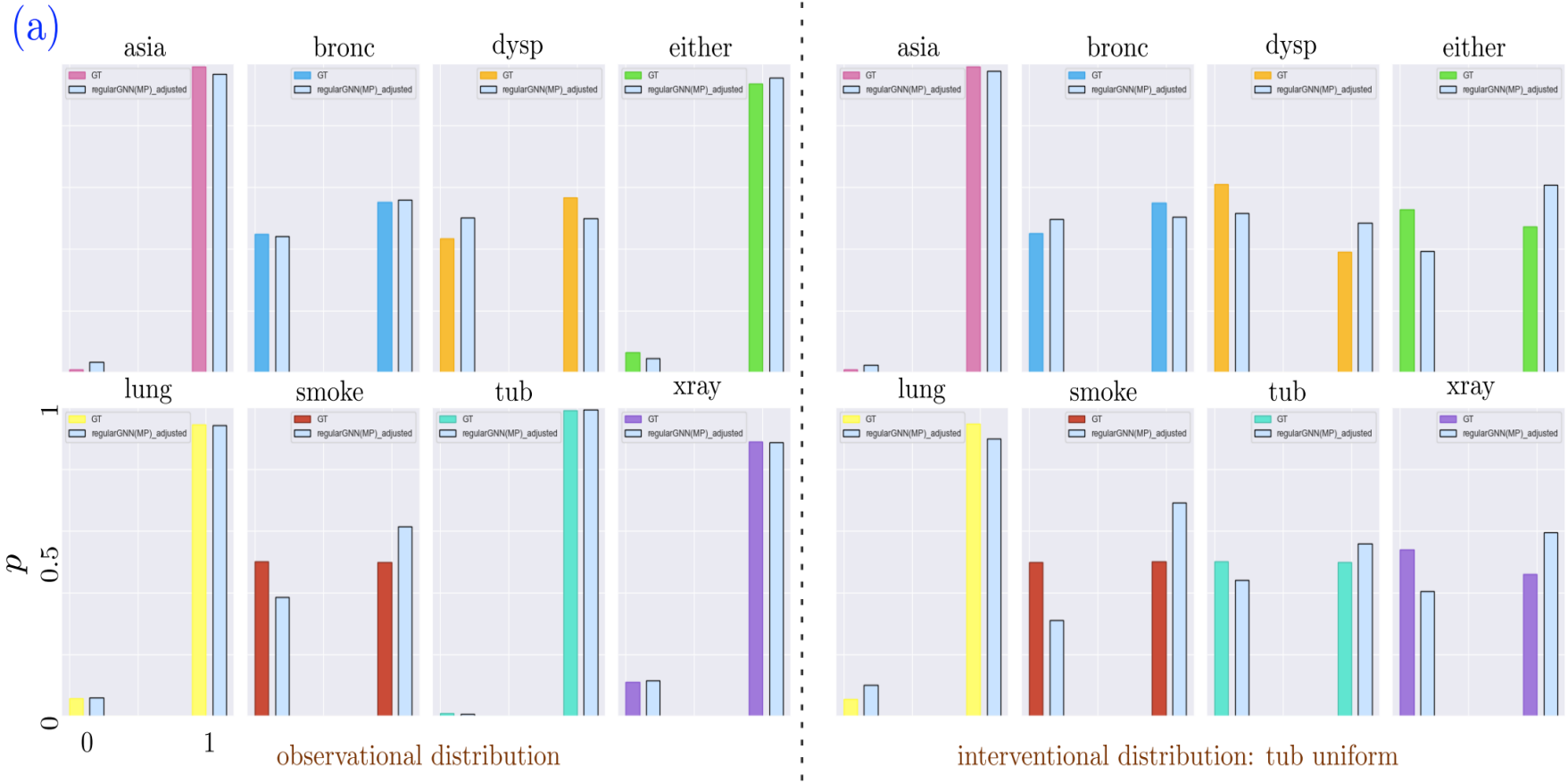
- No discertion between different interventions on same variable
- Considers only structural properties
- This interventional computation layer can be embedded into any system
- We show it via Variational Graph Autoencoder and call it iVGAE

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?

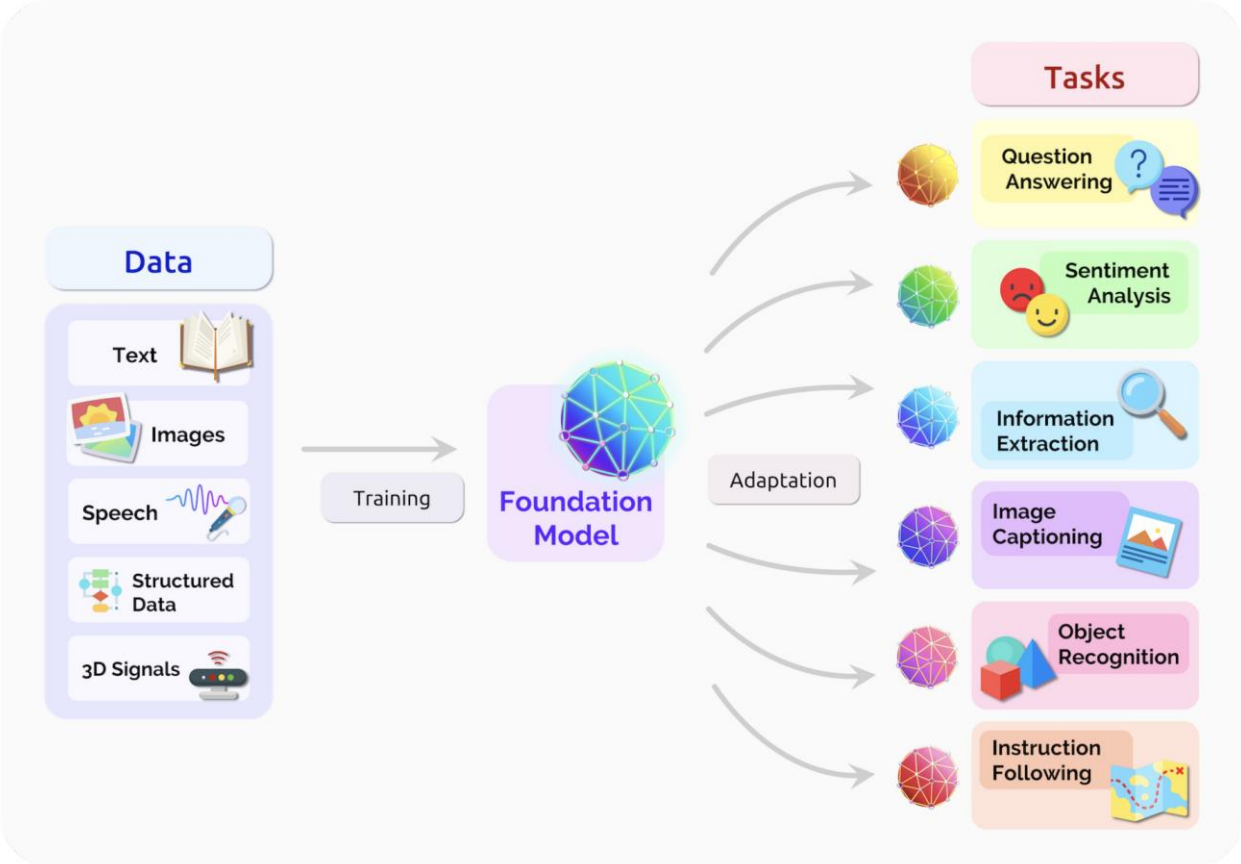


Judea Pearl
@yudapearl

My list of recent articles on causal inference:
ucla.in/39WglUm
One paper catching my attention is "Can Foundation Models Talk Causality" arxiv.org/pdf/2206.10591...
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[1,2,*] **Jiarui Liu**[3] **Zhiheng Lyu**[4] **Spencer Poff**[5]
Mrinmaya Sachan[2] **Rada Mihalcea**[3] **Mona Diab**[5,†] **Bernhard Schölkopf**[1,†]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany, [2]ETH Zürich,
[3]University of Michigan, [4]University of Hong Kong, [5]Meta AI

released on arXiv *last month*

Related Work in **Causality**

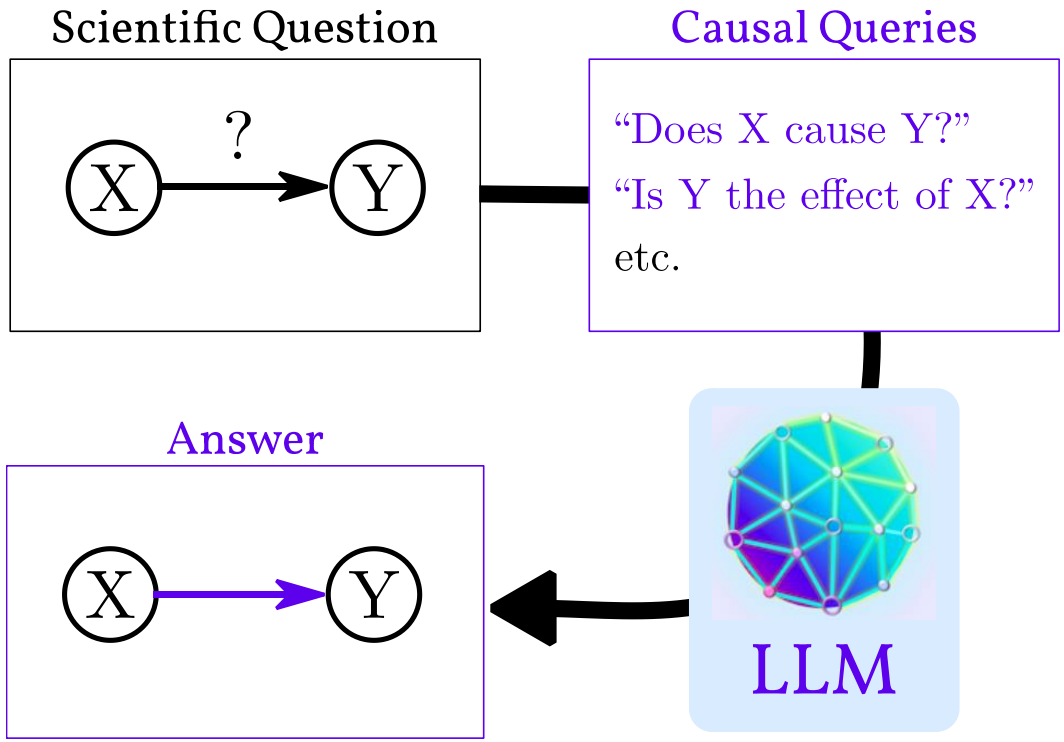
"[…] We evaluated an extensive list of LLMs on this new task, and showed that off-the-shelf LLMs perform poorly on this task."

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



Zečević et al., Causal Parrots: Large Language Models May Talk Causality But Are Not Causal (Under Review)

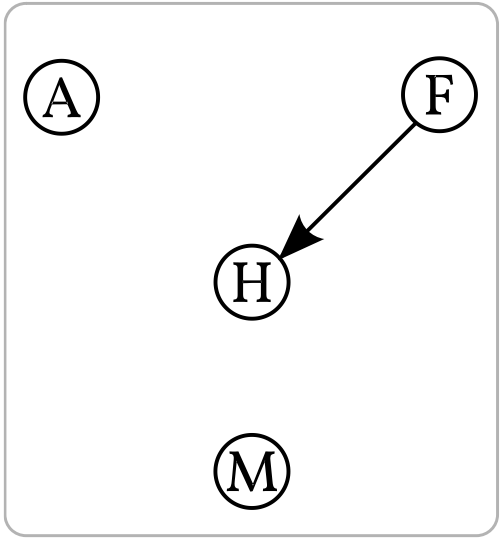
Overall Poor. But if good, then surprisingly!

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

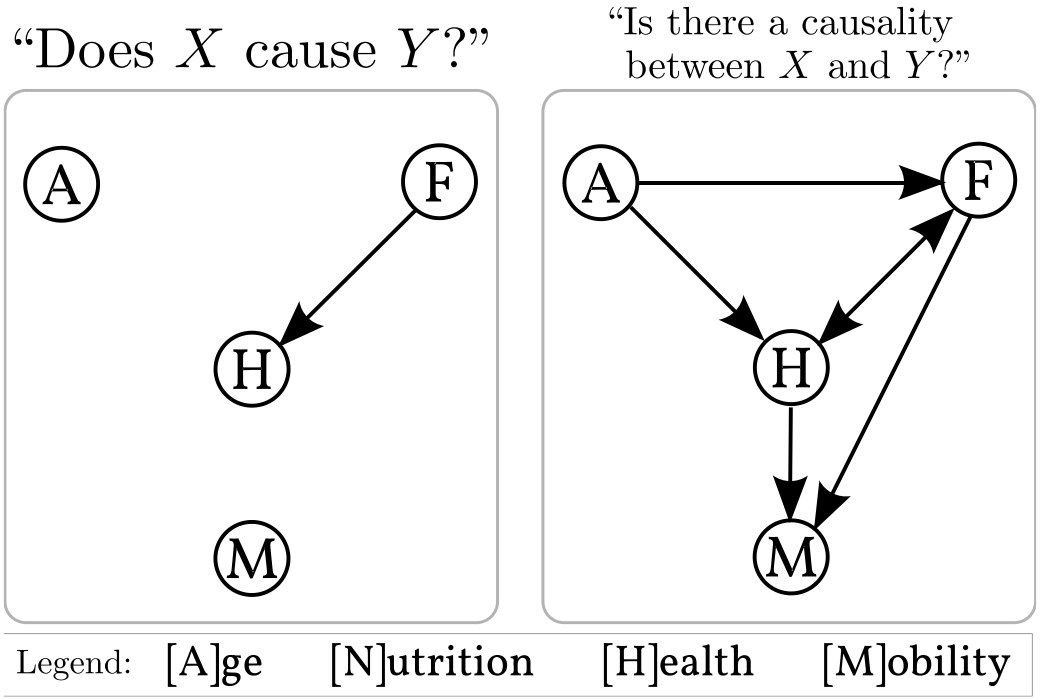
Remarkable Observation 1: Query Wording Sensitivity

"Does X cause Y?"

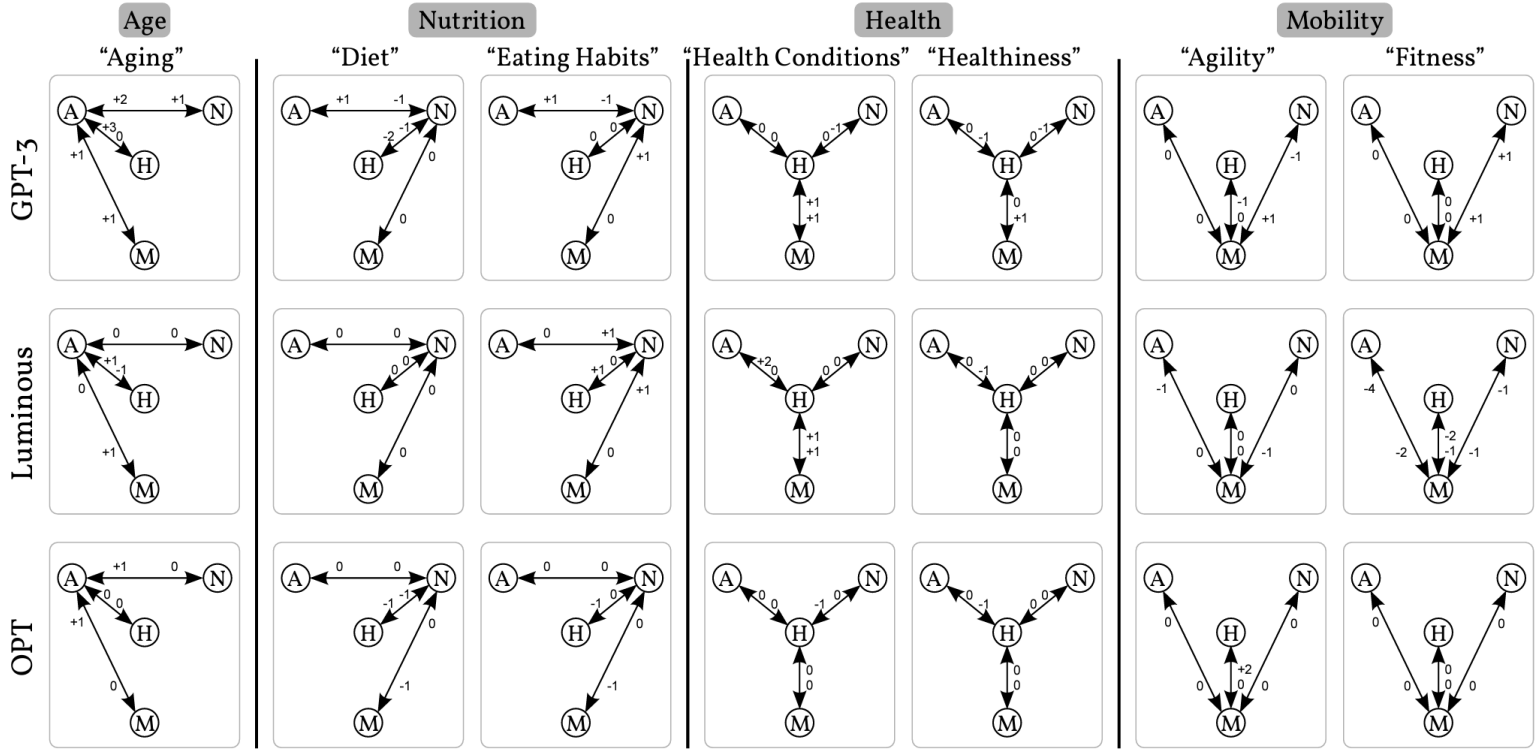


Legend: [A]ge [N]utrition [H]ealth [M]obility

Remarkable Observation 1: Query Wording Sensitivity



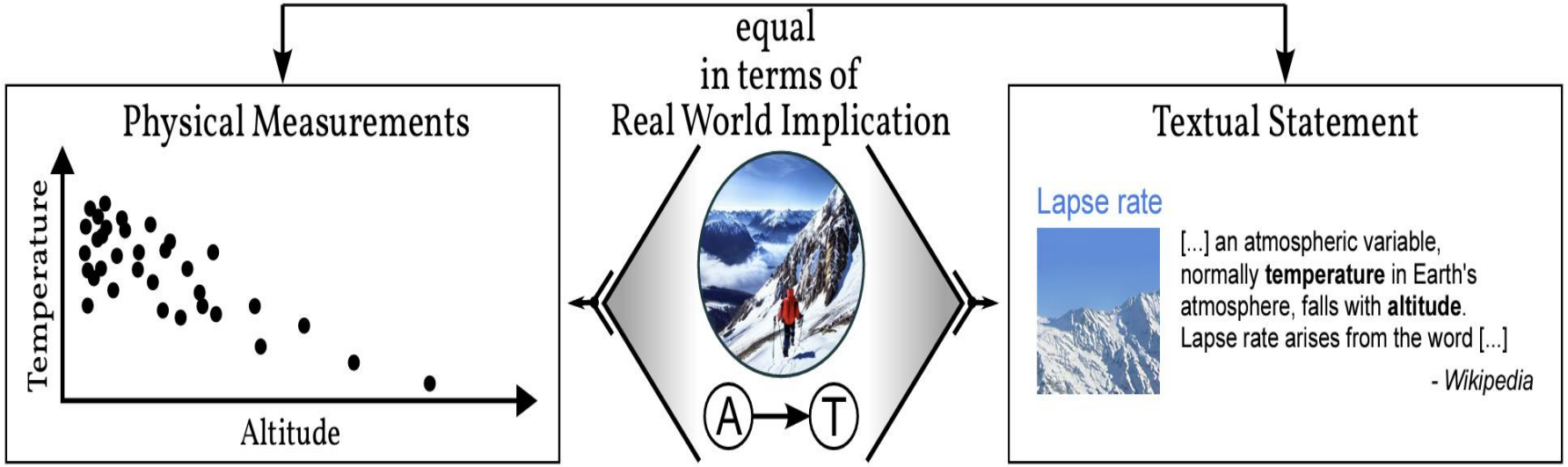
Remarkable Observation 2: Variable Naming Sensitivity



"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 *[end of output]*
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. *[continued]*
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. *[continued]*

"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. *[continued]*
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. *[repeating]*

Ultimately, a Philosophical Question



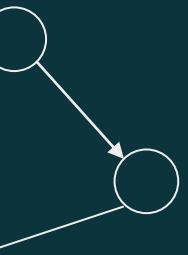
'understanding'

'knowing'

Alternative Analogy: Plato's Allegory of the Cave

correct answers by LLM

"correct causal reality"



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),
@devendratweetin
Matej Zečević (matej.zecevic@tu-darmstadt.de)
@matej_zecevic
Adèle Ribeiro (adele.ribeiro@uni-marburg.de)
@adelehr