Imperial College London

Causality & Generative Models

Fabio De Sousa Ribeiro

fdesousa@ic.ac.uk

Miguel Monteiro

Nick Pawlowski

Daniel C. Castro

Tian Xia

Ben Glocker

BioMedIA

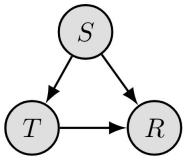
project mira ERC StG 2017 grant No. 757173

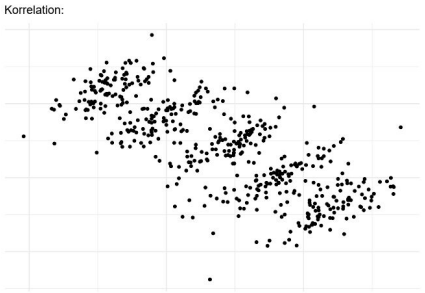
erc European Research Council

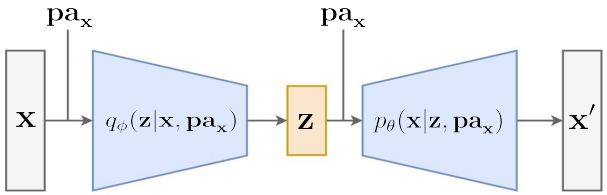
Microsoft

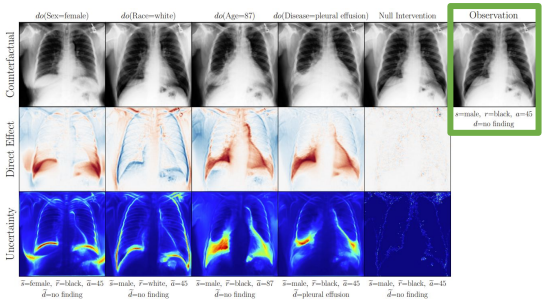
Outline

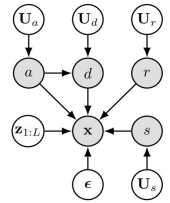
- ➔ Introduction & Motivation
  - Causality: The Ladder of Causation
  - Structural Causal Models
- ➔ Image Counterfactuals
  - Deep Causal Mechanisms for Structured Variables
  - Causal Mediation Analysis
- ➔ Case Studies: Medical Imaging
  - Brain Imaging
  - Chest X-ray Imaging
- ➔ Conclusion & Outlook











(a) Deep SCM for MIMIC-CXR. The variables in the causal graph are: age (a), sex (s), race (r), disease (d) and chest x-ray ($\mathbf{x}$). The disease d is pleural effusion.

What is Causality?

➔ Scientific inquiry is often motivated by **causal** questions:

I. How effective is a treatment in preventing a disease?

II. If I take this pill, will my headache be gone?

III. Would my grades have been better had I studied more?

Causality is the relationship between cause and effect

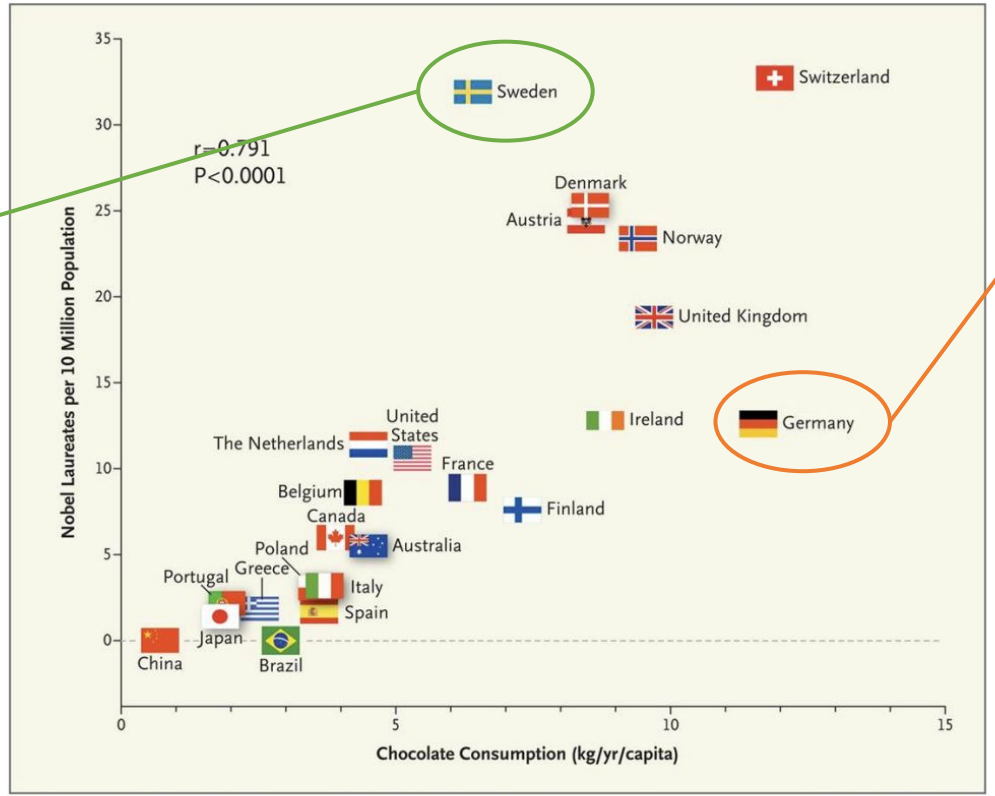
Example: chocolate

"Correlation does not imply causation."

Secretly the best chocolate?

Least effective chocolate?

Nobel Prize Correlates with Chocolate Consumption



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Inspired by Jonas Peters

Reichenbach's Common Cause Principle

Hans Reichenbach

I. Chocolate consumption causes Nobel prize win

$$C \rightarrow N$$

II. Nobel prize win causes chocolate consumption

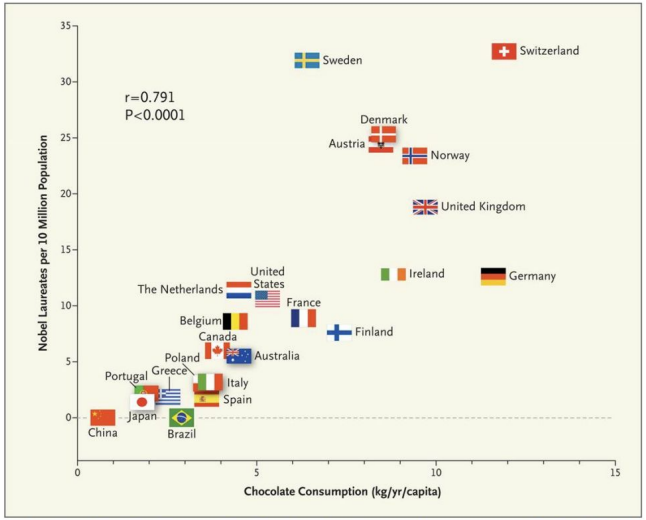
$$C \leftarrow N$$

III. Both are caused by an **unknown factor** U

$$C \leftarrow U \rightarrow N$$

Could U be GDP per capita or wealth per adult?

Nobel Prize Correlates with Chocolate Consumption



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 20

Simpson's Paradox

Edward Simpson

Kidney Stone Size	Treatment A	Treatment B
Small	**93%** (81/87)	87% (234/270)
Large	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

Charig CR, *et al*. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed). 1986 Mar 29.

Simpson's Paradox

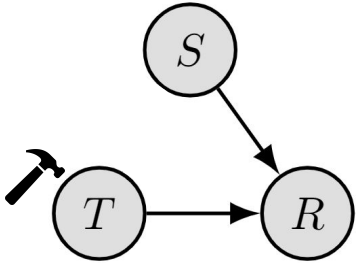
Edward Simpson

Kidney Stone Size	Treatment A	Treatment B
Small	**93%** (81/87)	87% (234/270)
Large	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

Charig CR, *et al*. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed). 1986 Mar 29.

Simpson's Paradox

Edward Simpson



- ➔ Stone size (S) is a **confounder**
- ➔ Make treatment (T) independent of size:
  - ◆ What's the recovery (R) rate if ***all*** subjects receive treatment **A** vs **B**?

Kidney Stone Size	Treatment A	Treatment B
Small	**93%** (81/87)	87% (234/270)
Large	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

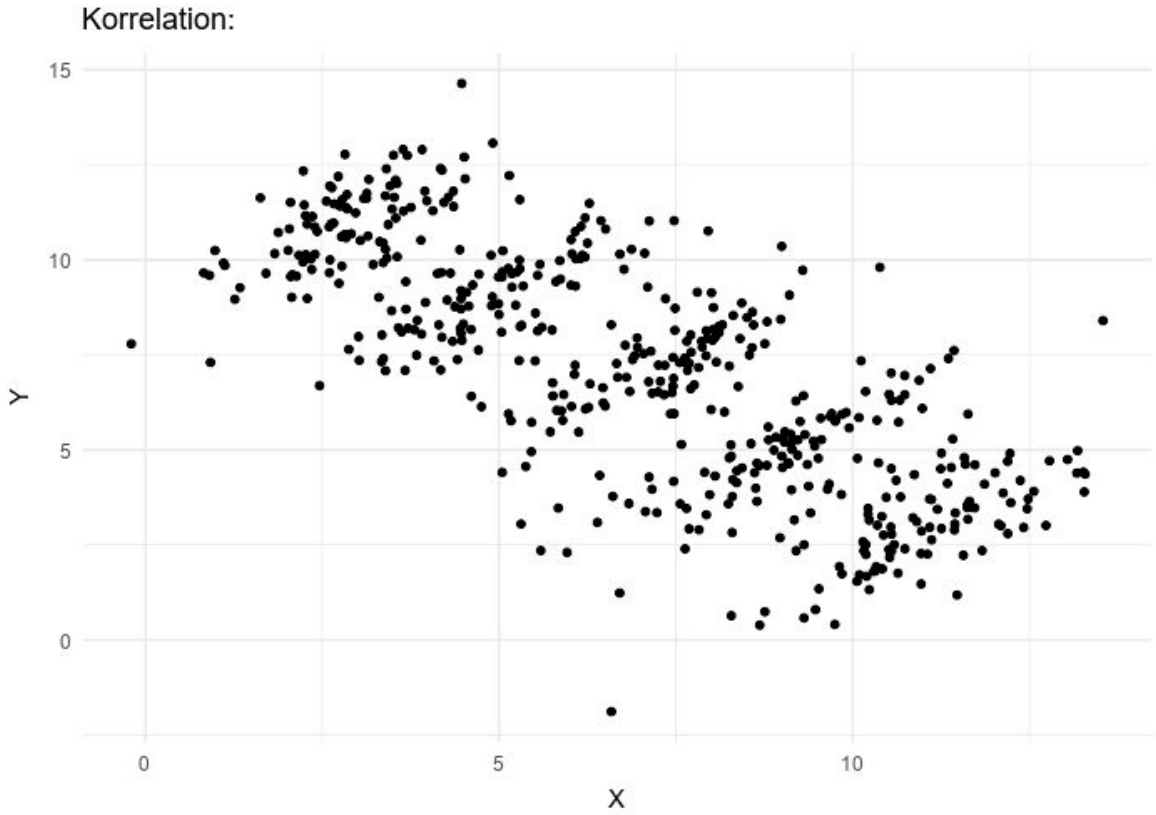
??

Patients with **large stones** received the **better treatment** (**A**), and those with **small stones** received the **inferior treatment** (**B**)

Charig CR, *et al*. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed). 1986 Mar 29.

Simpson’s Paradox

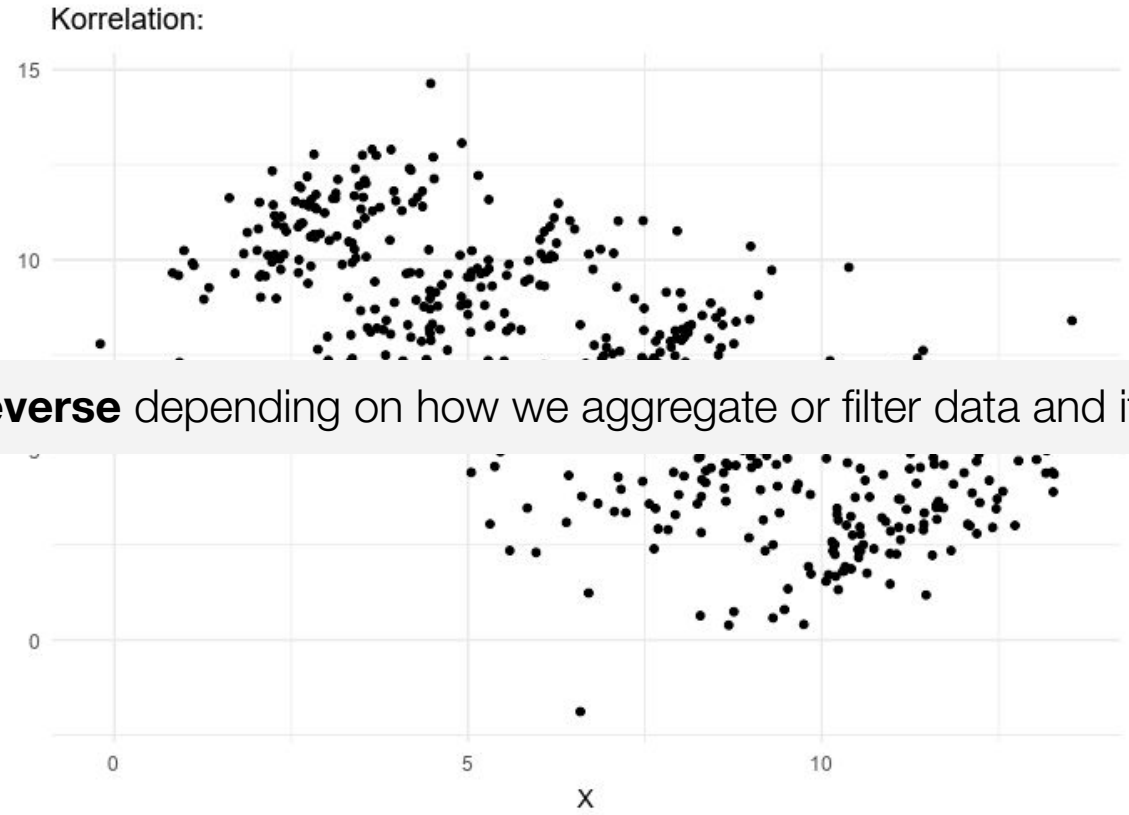
Edward Simpson



From: https://upload.wikimedia.org/wikipedia/commons/f/fb/Simpsons_paradox_-_animation.gif

Simpson's Paradox

Edward Simpson



Correlations may **reverse** depending on how we aggregate or filter data and its subpopulations

From: https://upload.wikimedia.org/wikipedia/commons/f/fb/Simpsons_paradox_-_animation.gif

Predictive Modelling

Given an **image** X, train a **model** to predict some **label** Y

$$P(Y \mid X)$$

➔ Assumptions:

- Sufficient training data (X, Y) is available
- Training and test data come from the same distribution

A Causal Perspective: Predictive Modelling

What is the **causal relationship** between image X and label Y?

$$P(Y \mid X)$$

$X \rightarrow Y$

causal
(predict effect from cause)

$Y \rightarrow X$

anti-causal
(predict cause from effect)

Example: Skin Lesion Classification

X – dermascopic image

Y – biopsy-derived diagnosis

$$X \overset{?}{\leftrightarrow} Y$$

causal or anti-causal?

$$P(Y \mid X)$$

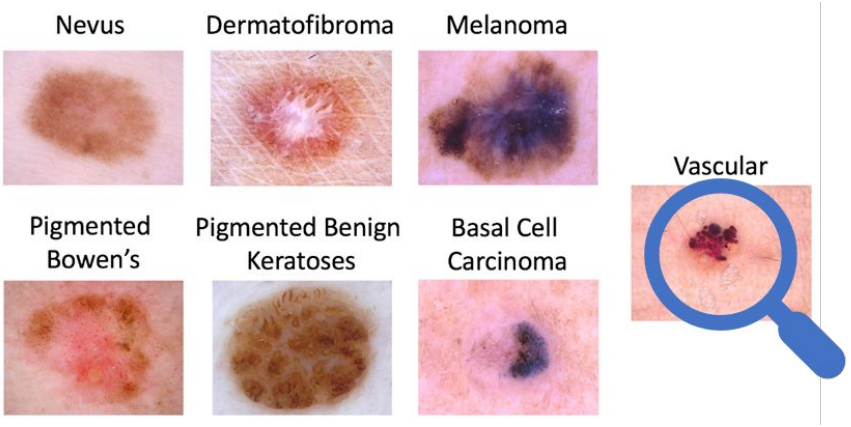


Image: ISIC 2018 challenge (https://challenge2018.isic-archive.com/task3/)

Example: Skin Lesion Classification

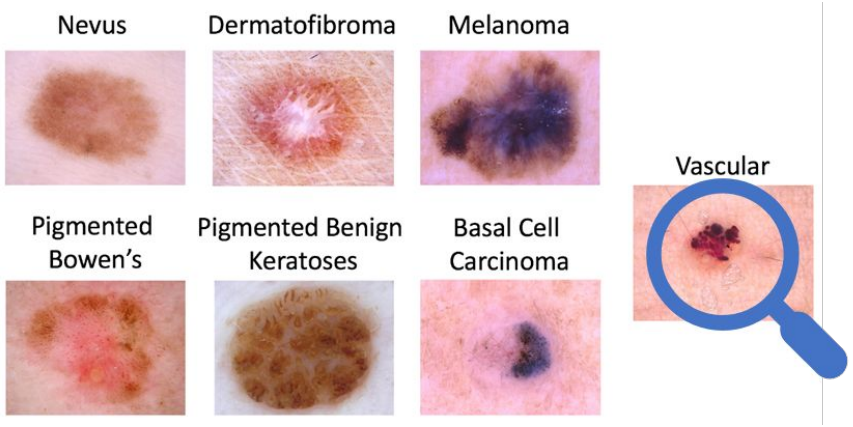
X – dermascopic image

Y – biopsy-derived diagnosis

$$X \leftarrow Y$$

anti-causal
(predict cause from effect)

$$P(Y \mid X)$$



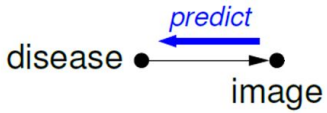
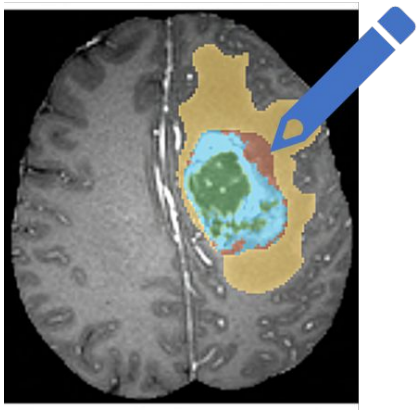


Image: ISIC 2018 challenge (https://challenge2018.isic-archive.com/task3/)

Example: Brain Tumour Segmentation

X – structural brain MRI

Y – manually drawn contour



$$X \overset{?}{\leftrightarrow} Y$$

causal or anti-causal?

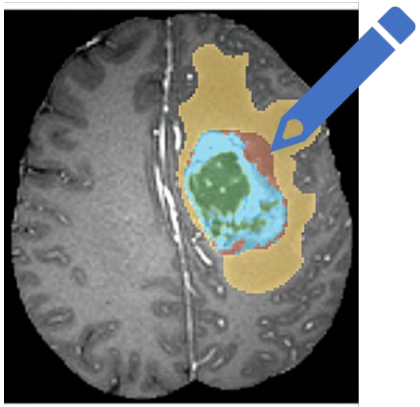
$$P(Y \mid X)$$

Image: ISIC 2018 challenge (https://challenge2018.isic-archive.com/task3/)

Example: Brain Tumour Segmentation

X – structural brain MRI

Y – manually drawn contour



$$X \rightarrow Y$$

causal
(predict effect from cause)

$$P(Y \mid X)$$

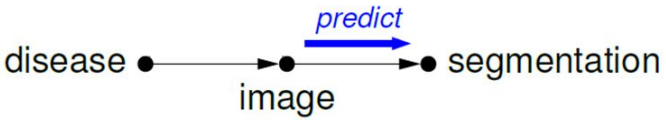
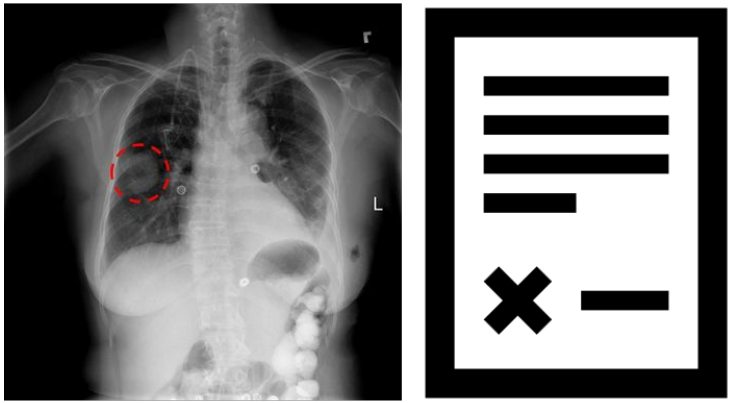


Image: ISIC 2018 challenge (https://challenge2018.isic-archive.com/task3/)

Example: Radiology Reports

X – chest X-ray

Y – diagnosis extracted from report



$$X \overset{?}{\leftrightarrow} Y$$

causal or anti-causal?

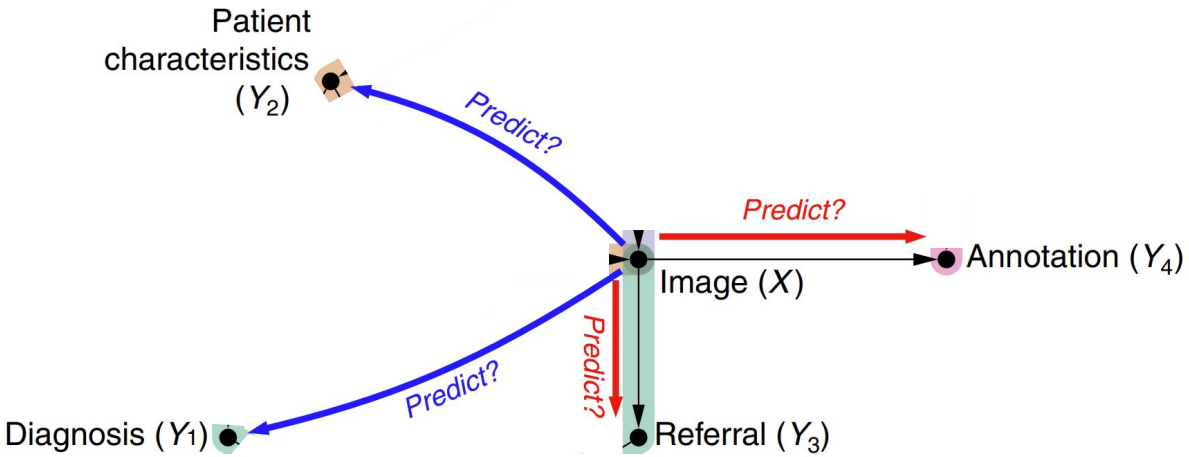
Not clear whether the diagnosis was derived from the image, or from another lab result...

$$P(Y \mid X)$$

Meta-information is needed to establish causal relationships

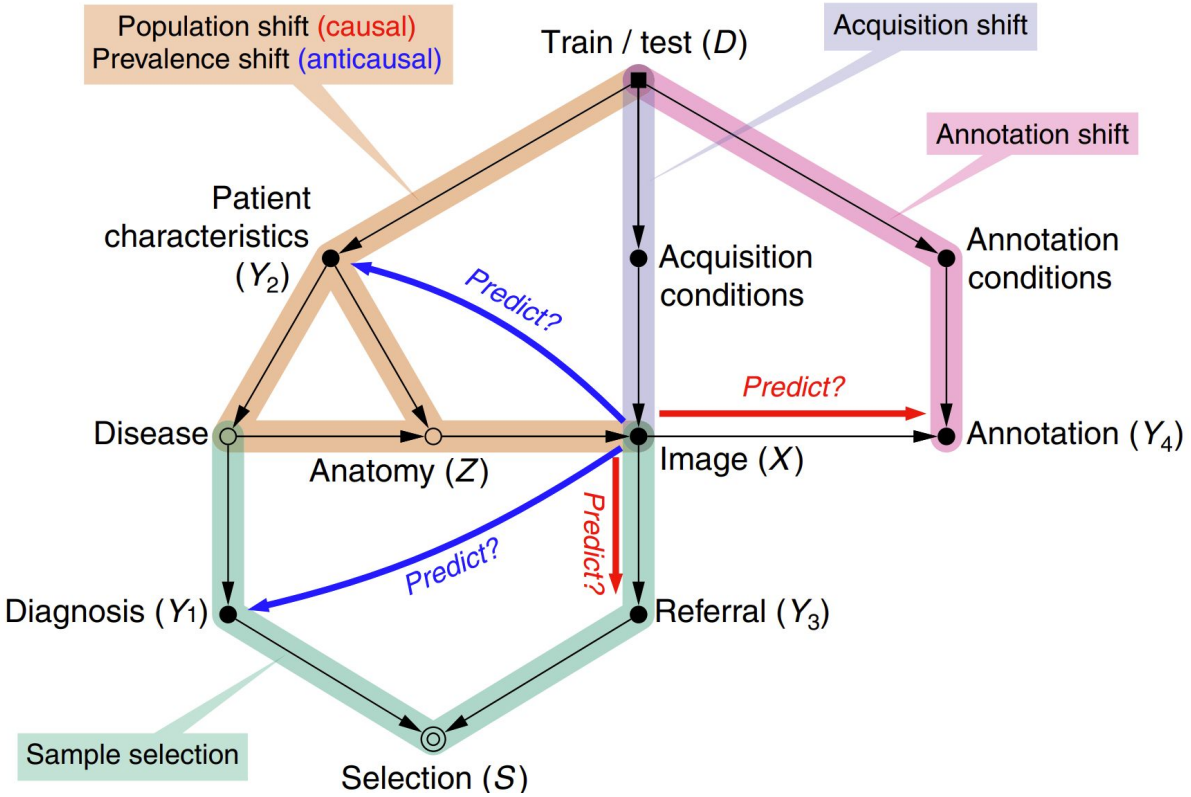
Image: ISIC 2018 challenge (https://challenge2018.isic-archive.com/task3/)

Data Generating Process



Castro, D. C., Walker, I., & Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, *11*(1), 3673.

Data Generating Process



!!

Castro, D. C., Walker, I., & Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, *11*(1), 3673.

Ladder of Causation

Counterfactuals

What if X had not occurred?

I. Counterfactual reasoning

II. Deduce causes for observed events

3.

Intervention

What will Y be if i **do** X?

I. Use experiments to identify causal effects

II. Crucial for planning and policy making

2.

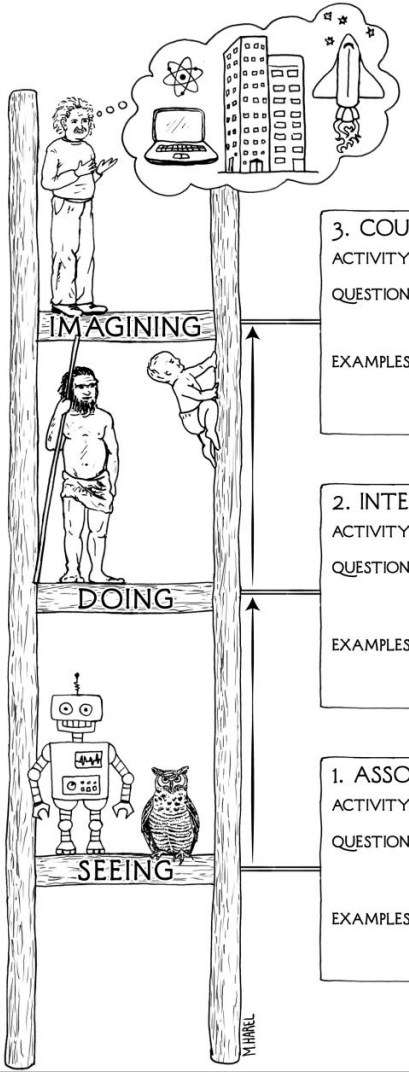
Association

How would **seeing** X change my belief in Y?

I. Modelling correlations $P(Y|X)$

II. Modern ML excels at this task

1.



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?* (How are the variables related? How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

Judea Pearl

Pearl & Mackenzie. *The Book of Why*

Structural Causal Models

➔ A Structural Causal Model (SCM) is a triple: $\mathcal{M} := \langle X, U, F \rangle$

- Two sets of variables:

$$X = \{x_1, \ldots, x_N\} \qquad U = \{u_1, \ldots, u_N\}$$

- A set of functions known as **causal mechanisms**:

$$F = \{f_1, \ldots, f_N\}$$

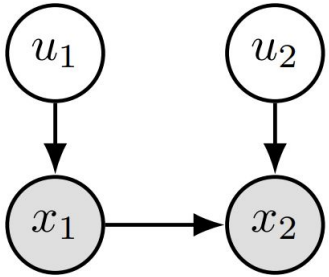
- The value of each variable is a function of its **parents** (direct causes):

$$x_k := f_k(\mathbf{pa}_k, u_k), \qquad k = 1, \ldots, N$$

Structural Causal Models

➔ Example: A simple SCM:

- x_1, x_2 are **endogenous** whereas u_1, u_2 are **exogenous**

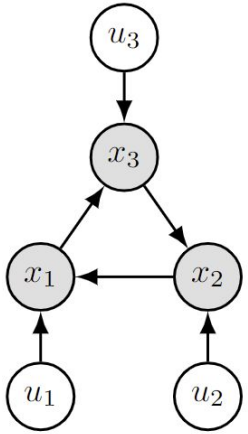


$$x_1 := f_1(u_1), \qquad u_1 \sim \mathcal{N}(0,1)$$

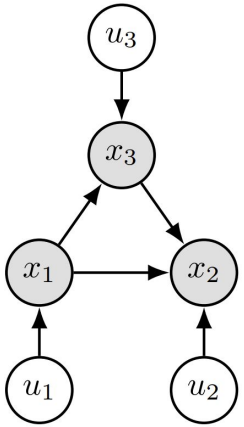
$$x_2 := f_2(x_1, u_2), \qquad u_2 \sim \mathcal{N}(0,1)$$

Structural Causal Models

➔ Acyclic SCMs can be represented by a **Directed Acyclic Graphs** (DAGs), with edges pointing from causes to effects



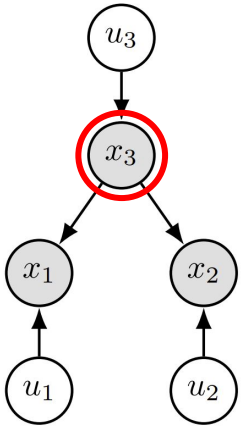
(a) Cyclic



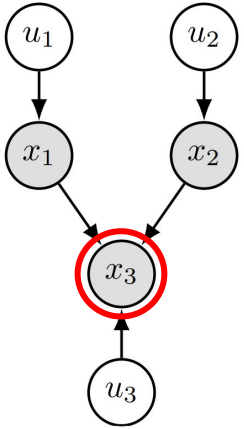
(b) Acyclic

Structural Causal Models

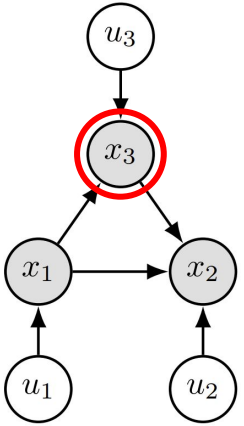
➔ Acyclic SCMs can be represented by a **Directed Acyclic Graphs** (DAGs), with edges pointing from causes to effects



(a) Confounder



(b) Collider



(c) Mediator

Observational Distribution (rung 1.)

➔ SCMs with jointly **independent** exogenous noises are called **Markovian**:

$$P(u_1, \ldots, u_N) = \prod_{k=1}^{N} P(u_k)$$

➔ Markovian SCMs induce a unique joint **observational distribution** over the endogenous variables:

$$P_{\mathcal{M}}(x_1, \ldots, x_N) = \prod_{k=1}^{N} P_{\mathcal{M}}(x_k \mid \mathbf{pa}_k)$$

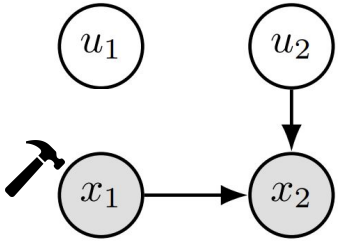
➔ Each variable is **independent** of its non-descendants given its direct causes (*causal Markov condition*)

Interventional Distribution (rung 2.)

➔ SCMs **predict** the **causal effects** of actions via **interventions**

➔ **Interventions** answer causal questions like:

- E.g. what would x_2 be if we set $x_1 := c$?



➔ **Interventions** replace one or more of the structural assignments and are denoted with the *do* operator: $do(x_k := c)$

➔ This induces a submodel $\mathcal{M}_c$ and its entailed distribution is known as the **interventional distribution**: $P_{\mathcal{M}_c}(X \mid do(c))$

Counterfactuals (rung 3.)

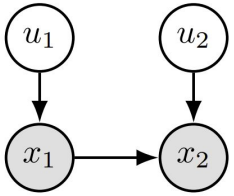
➔ SCMs can consider **hypothetical** scenarios:

- Given that we observed (x_1, x_2), what would x_2 have been had x_1 been c?
- All else being equal, would I have been late had I not missed the train?

➔ **Counterfactual inference** involves three steps:

1. **Abduction:** Update $P(U)$ given observed evidence, i.e. infer posterior $P(U \mid X)$
2. **Action:** Perform an intervention e.g. $do(\widetilde{x}_k := c)$ and obtain the submodel $\mathcal{M}_c$
3. **Prediction:** Use the modified model $\langle \mathcal{M}_c, P(U \mid X) \rangle$ to compute counterfactuals

Example: Computing Counterfactuals



$$x_1 := f_1(u_1) = 1 + u_1$$

$$x_2 := f_2(x_1, u_2) = 3x_1 + u_2$$

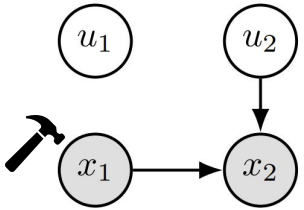
Q: Given we observed $\{x_1{=}2, x_2{=}4\}$, what would x_2 have been had x_1 been 5?

1. Abduction:

$$u_1 = f_1^{-1}(x_1) = x_1 - 1 \implies u_1 = 1$$

$$u_2 = f_2^{-1}(x_1, x_2) = x_2 - 3x_1 \implies u_2 = -2$$

2. Action: $\widetilde{x_1} := 5$



3. Prediction: $\widetilde{x}_2 = 3\widetilde{x_1} + u_2 = 3 \cdot 5 - 2 = 13$

Deep Structural Causal Models[1]

➔ Leverage **deep generative models** to learn SCM mechanisms:

$$x_k := f_k(\mathbf{pa}_k, u_k)$$

➔ Tractably estimate causal effects of interventions and perform counterfactual inference, i.e. answer “what if…?” type questions

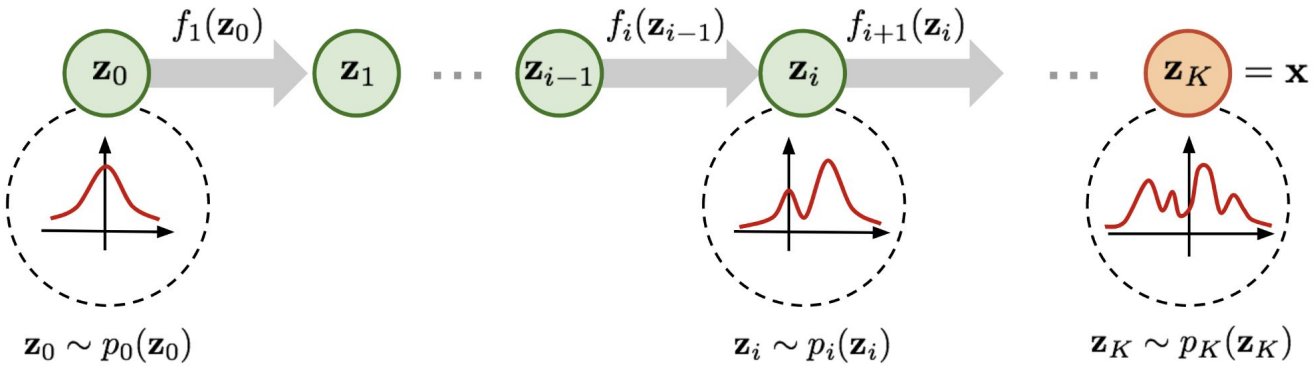
➔ **Abduction** is challenging in complex problems, e.g. medical imaging

Research Question: Can we generate plausible high-fidelity image counterfactuals of real-world data, and if so, how do we evaluate them?

[1] Pawlowski, Castro, Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. *NeurIPS* 2020

Deep Mechanisms: Normalizing Flows

➔ Normalizing Flows (NFs) build complex probability distributions via successive (invertible) transformations of simple distributions



lilianweng.github.io/posts/2018-10-13-flow-models/

Kobyzev, Prince, Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *PAMI* 2020

Papamakarios, Nalisnick, Rezende, Mohamed, Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *JMLR* 2021

Deep Mechanisms: Normalizing Flows

➔ Normalizing Flows (NFs) build complex probability distributions via successive (invertible) transformations of simple distributions

blog.evjang.com/2019/07/nf-jax.html

➔ **TLDR:** NFs enable **invertible** mechanisms and deterministic **abduction:**

$$x = f_\theta(\mathbf{pa}_x, u_x) \qquad u_x = f_\theta^{-1}(\mathbf{pa}_x, x)$$

Intuition: Normalizing Flows

➔ Change-of-variables formula:

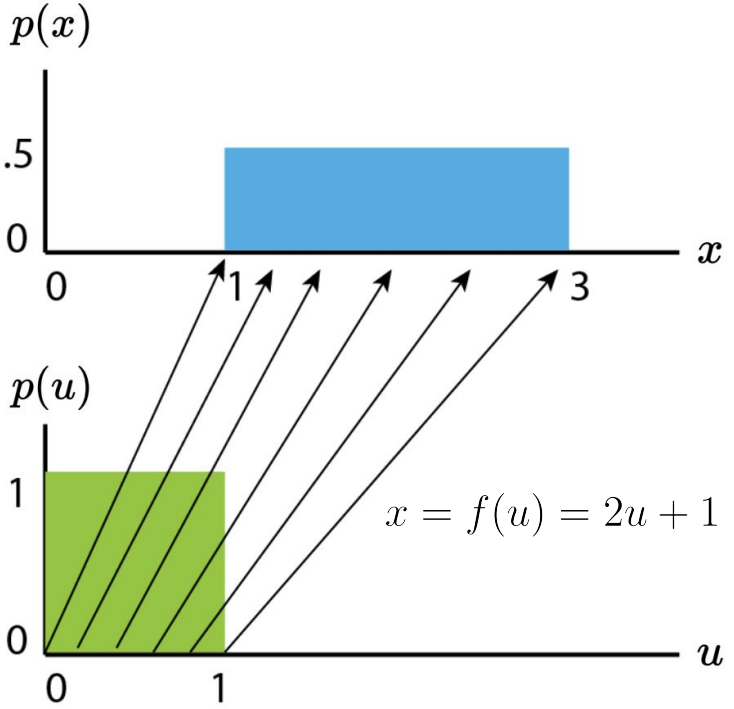
$$p(x) = p(u)\underbrace{\left|\frac{du}{dx}\right|}_{\text{volume correction}}$$

$$u = f^{-1}(x) = \frac{x-1}{2}$$

$$\frac{du}{dx} = \frac{df^{-1}(x)}{dx} = \frac{d}{dx}\left(\frac{x-1}{2}\right) = \frac{1}{2}$$

$$p(u) = p_U(f^{-1}(x)) = \frac{1}{1-0} = 1$$

$X = f(U)$



PDF of $\mathcal{U}[a, b]$ is $\frac{1}{b-a}$

Adapted figure from: blog.evjang.com/2018/01/nf1.html

Training Objective: Normalizing Flows

➔ Maximum (log) likelihood training objective:

$$\log p(x) = \log p(f_\theta^{-1}(x)) + \log \left| \frac{df_\theta^{-1}(x)}{dx} \right|$$

➔ Can condition on parents via a learned parameterised function:

$$\log p(x \mid \mathbf{pa}_x) = \log p(f_\theta^{-1}(\mathbf{pa}_x, x)) + \log \left| \det \left(\frac{\partial f_\theta^{-1}(\mathbf{pa}_x, x)}{\partial x} \right) \right|$$

➔ In multivariate settings, we need to compute the determinant of a **Jacobian** matrix, which can become pretty expensive!

Deep Mechanisms: Variational Autoencoders

➔ Well suited for modelling structured variable mechanisms, e.g. for images

➔ Caveat: **abduction** is non-deterministic

➔ Causal mechanism:

invertible non-invertible

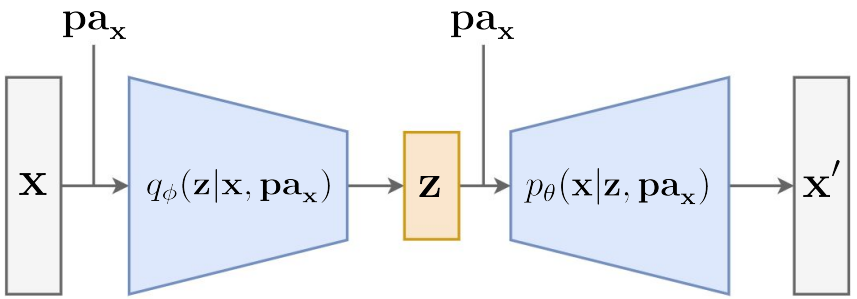
$$\mathbf{x} := f_\theta(\mathbf{pa_x}, \mathbf{u_x}) = h(\boldsymbol{\epsilon}; g_\theta(\mathbf{z}, \mathbf{pa_x}))$$

$$= \boldsymbol{\mu}(\mathbf{z}, \mathbf{pa_x}) + \boldsymbol{\sigma}(\mathbf{z}, \mathbf{pa_x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$$

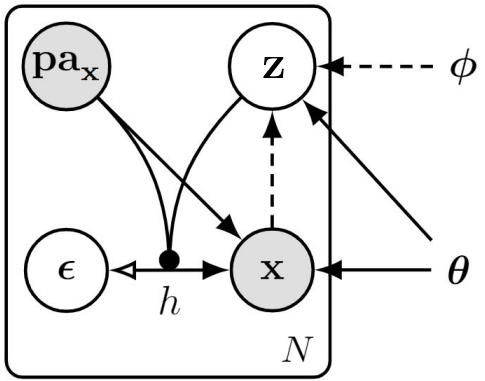
➔ Factored exogenous noise: $p(\mathbf{u_x}) = p_\theta(\mathbf{z})p(\boldsymbol{\epsilon})$

$$p(\mathbf{u_x}|\mathbf{x}, \mathbf{pa_x}) \approx q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{pa_x})\delta(\boldsymbol{\epsilon} - h^{-1}(\mathbf{x}; g_\theta(\mathbf{z}, \mathbf{pa_x}))$$

(a) Conditional VAE



Causal mechanism for $\mathbf{x}$



Deep Mechanisms: Variational Autoencoders

➔ Example: computing counterfactuals:

1. **Abduction:** $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{pa_x})$

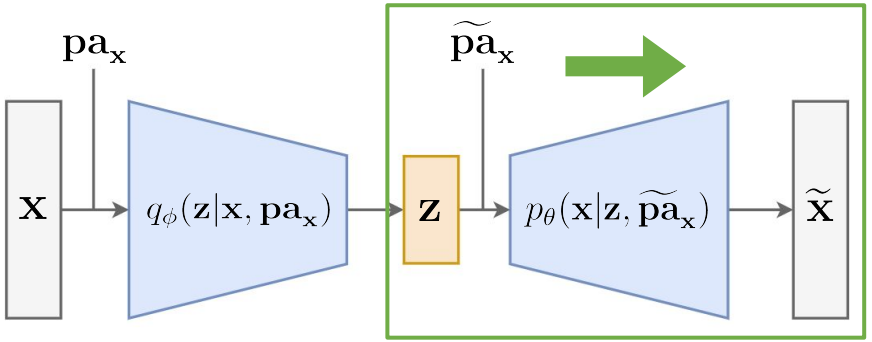
$$\boldsymbol{\epsilon} = h^{-1}(\mathbf{x}; g_\theta(\mathbf{z}, \mathbf{pa_x})) = \frac{\mathbf{x} - \boldsymbol{\mu}(\mathbf{z}, \mathbf{pa_x})}{\boldsymbol{\sigma}(\mathbf{z}, \mathbf{pa_x})}$$

2. **Action:** $do(\mathbf{pa_x} := \widetilde{\mathbf{pa}}_\mathbf{x})$

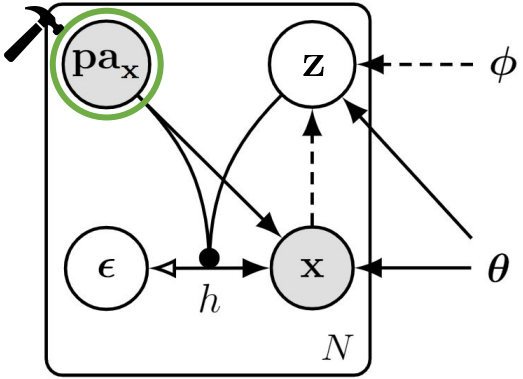
3. **Prediction:** $\widetilde{\mathbf{x}} \sim p_\theta(\widetilde{\mathbf{x}} \mid \mathbf{z}, \widetilde{\mathbf{pa}}_\mathbf{x})$

$$\widetilde{\mathbf{x}} := h(\boldsymbol{\epsilon}; g_\theta(\mathbf{z}, \widetilde{\mathbf{pa}}_\mathbf{x})) = \boldsymbol{\mu}(\mathbf{z}, \widetilde{\mathbf{pa}}_\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{z}, \widetilde{\mathbf{pa}}_\mathbf{x}) \odot \boldsymbol{\epsilon}$$

(a) Conditional VAE



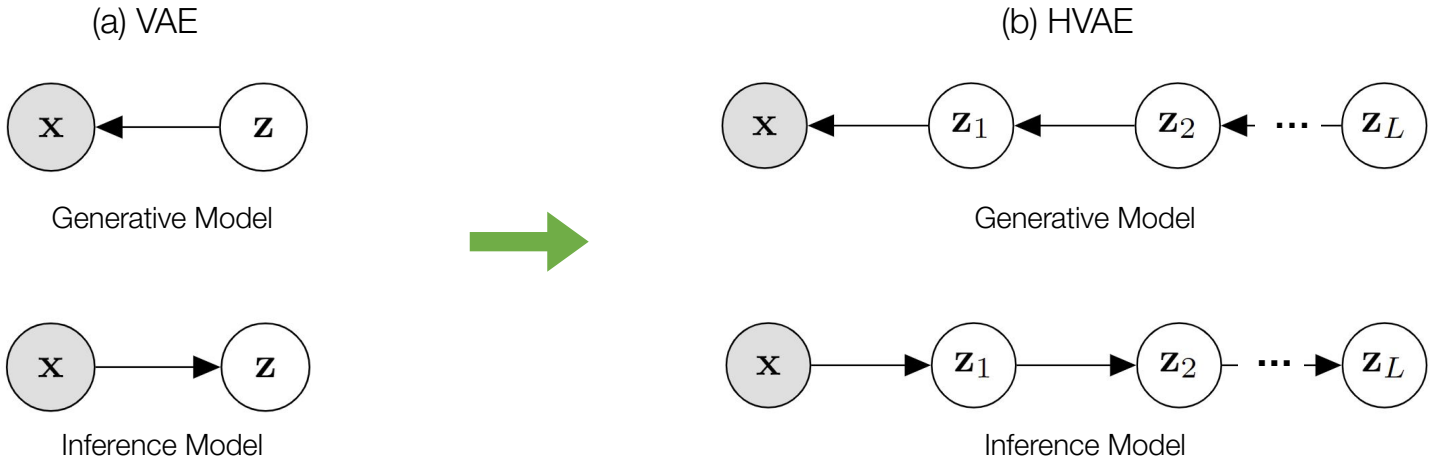
Causal mechanism for $\mathbf{x}$



Deep Mechanisms: Hierarchical VAEs

➔ To produce high-fidelity **image counterfactuals**, we require a powerful generative model that is amenable to principled **abduction**

➔ **Hierarchical VAEs** extend VAEs to multiple layers of latent variables:



Deep Mechanisms: Hierarchical VAEs

➔ To produce high-fidelity **image counterfactuals**, we require a powerful generative model that is amenable to principled **abduction**

➔ **Hierarchical VAEs** extend VAEs to multiple layers of latent variables:

$$\underbrace{p(\mathbf{x}, \mathbf{z}_{1:L})}_{\text{Generative Model}} = p(\mathbf{x} \mid \mathbf{z}_{1:L}) p(\mathbf{z}_L) \prod_{i=1}^{L-1} p(\mathbf{z}_i \mid \mathbf{z}_{>i})$$

➔ **Goal:** Optimize our model p to be close to a given data distribution p_{data}

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}_{1:L} \mid \mathbf{x})} \left[\log p(\mathbf{x} \mid \mathbf{z}_{1:L}) \right] - D_{\text{KL}}(q(\mathbf{z}_{1:L} \mid \mathbf{x}) \parallel p(\mathbf{z}_{1:L})) =: \text{ELBO}(\mathbf{x})$$

Literature: Deep

Ladder Va

NVAE: A Deep Hiera

Ar
{avahd

Normalizing flows, autoregres
deep energy-based models ar
for deep generative learning.
tractable sampling and easy-tc
rently outperformed by other n
models. While the majority o
challenges, we explore the orth
tectures for hierarchical VAEs.
chical VAE built for image gen
batch normalization. NVAE is
distributions and its training is
NVAE achieves state-of-the-ar
models on the MNIST, CIFA
provides a strong baseline on F
state-of-the-art from 2.98 to 2.
images on CelebA HQ as show
the first successful VAE applie
source code is available at htt

Casper Kaae Sønderby*
casperkaae@gmail.com

Søren Kaae Sønd
skaaesonderby@gm

Variational autoencoders are p
deep models with several laye
train which limits the improve
We propose a new inference
recursively corrects the gener
likelihood in a process resem
show that this model provides
log-likelihood lower bound co
Variational Autoencoders and
analysis of the learned hierar
inference model is qualitative
hierarchy of latent variables. Finally, we observe that batch normalization and
deterministic warm-up (gradually turning on the KL-term) are crucial for training
variational models with many stochastic layers.

Improved Variational Inference with Inverse Autoregressive Flow

Diederik P. Kingma
dpkingma@openai.com

Tim Salimans
tim@openai.com

Rafal Jozefowicz
rafal@openai.com

Xi Chen
peter@openai.com

Ilya Sutskever
ilya@openai.com

Max Welling*
M.Welling@uva.nl

Abstract

The framework of normalizing flows provides a general strategy for flexible variational inference of posteriors over latent variables. We propose a new type of normalizing flow, inverse autoregressive flow (IAF), that, in contrast to earlier published flows, scales well to high-dimensional latent spaces. The proposed flow consists of a chain of invertible transformations, where each transformation is based on an autoregressive neural network. In experiments, we show that IAF significantly improves upon diagonal Gaussian approximate posteriors. In addition, we demonstrate that a novel type of variational autoencoder, coupled with IAF, is competitive with neural autoregressive models in terms of attained log-likelihood on natural images, while allowing significantly faster synthesis.

ZE AUTOREGRESSIVE
M THEM ON IMAGES

ime, generates samples quickly
n all natural image benchmarks.
ctually represent autoregressive
t, when made sufficiently deep.
lly outperformed VAEs in log-
hy by scaling a VAE to greater
luating it CIFAR-10, ImageNet,
very deep VAEs achieve higher
s thousands of times faster, and
Qualitative studies suggest this
ual representations. We release
b.com/openai/vdvae.

Quick Detour: Connection to Diffusion Models

➔ A **Diffusion Model** is a type of Hierarchical VAE where:

I. The encoder $q(\mathbf{z}_{1:T} \mid \mathbf{x})$ is fixed rather than learned

II. Each latent variable $\mathbf{z}_t$ has the same dimensionality as $\mathbf{x}$

III. A single (denoising) model is shared amongst all layers in the hierarchy

$$-\text{ELBO}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{x})}\left[-\log p(\mathbf{x} \mid \mathbf{z}_{1:T})\right] + D_{\text{KL}}(q(\mathbf{z}_{1:T} \mid \mathbf{x}) \parallel p(\mathbf{z}_{1:T}))$$

$$= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})}\left[-\log p(\mathbf{x} \mid \mathbf{z}_1)\right] + D_{\text{KL}}(q(\mathbf{z}_T \mid \mathbf{x}) \parallel p(\mathbf{z}_T)) + \mathcal{L}_T(\mathbf{x})$$ diffusion loss

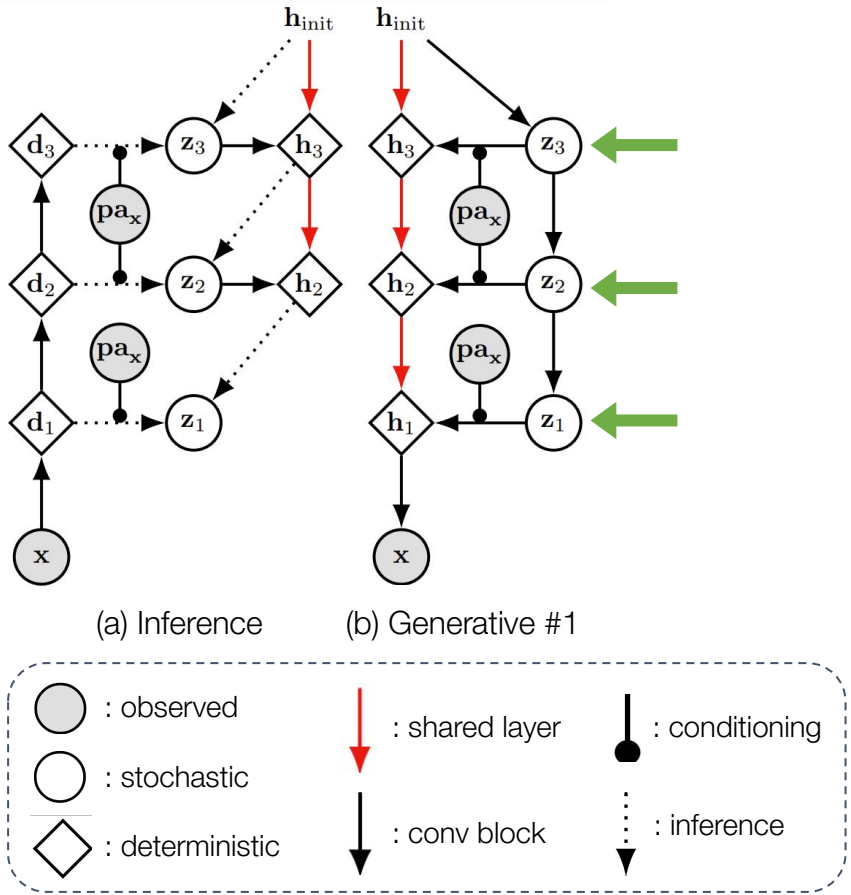
$$\mathcal{L}_T(\mathbf{x}) = \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}\left[D_{\text{KL}}(q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{x}) \parallel p(\mathbf{z}_{t-1} \mid \mathbf{z}_t))\right] = \frac{T}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t\sim U\{1,T\}}\left[w(t)\,\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t; t)\|_2^2\right]$$

Kingma, Diederik, et al. Variational diffusion models. NeurIPS 2021; De Sousa Ribeiro, F. & Glocker, B. (2024). Demystifying Variational Diffusion Models. *arXiv preprint arXiv:2401.06281*.

Deep Mechanisms: Conditional HVAEs

➔ Generative model structures:

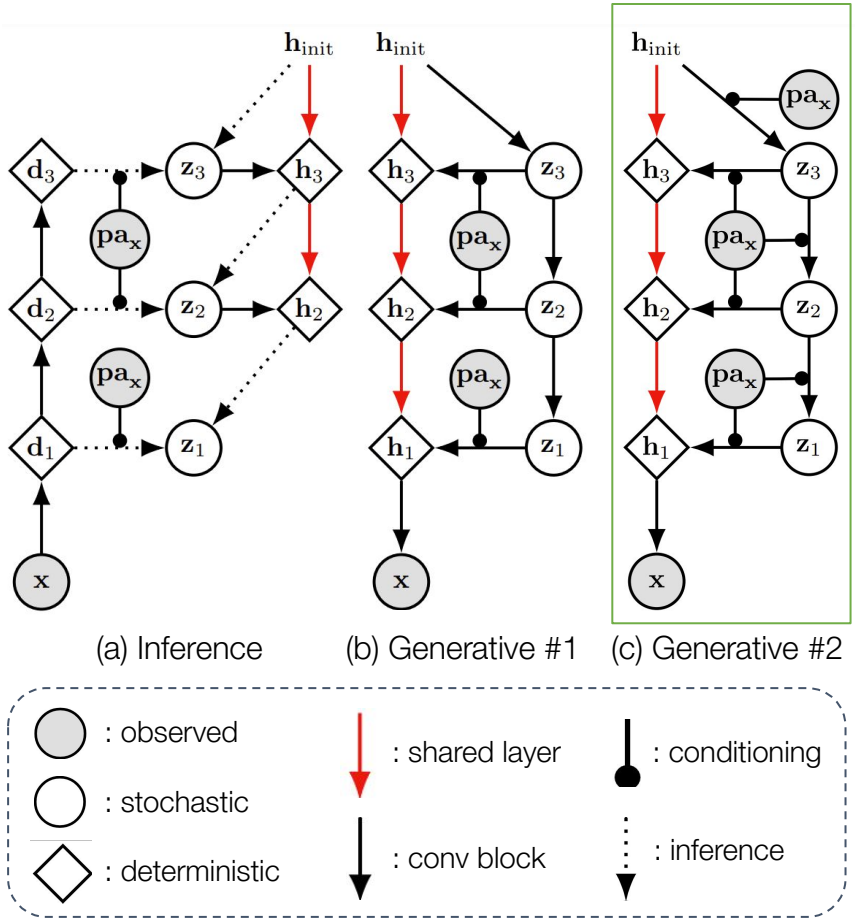
1. Exogenous Prior: $p_\theta(\mathbf{z}_{1:L})$



Deep Mechanisms: Conditional HVAEs

➔ Generative model structures:

1. Exogenous Prior: $p_\theta(\mathbf{z}_{1:L})$
2. Conditional Prior: $p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa_x})$



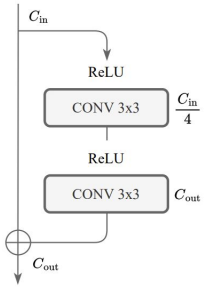
Deep Mechanisms: Conditional HVAEs

➔ Generative model structures:

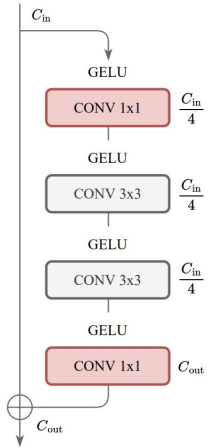
1. Exogenous Prior: $p_\theta(\mathbf{z}_{1:L})$
2. Conditional Prior: $p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa_x})$

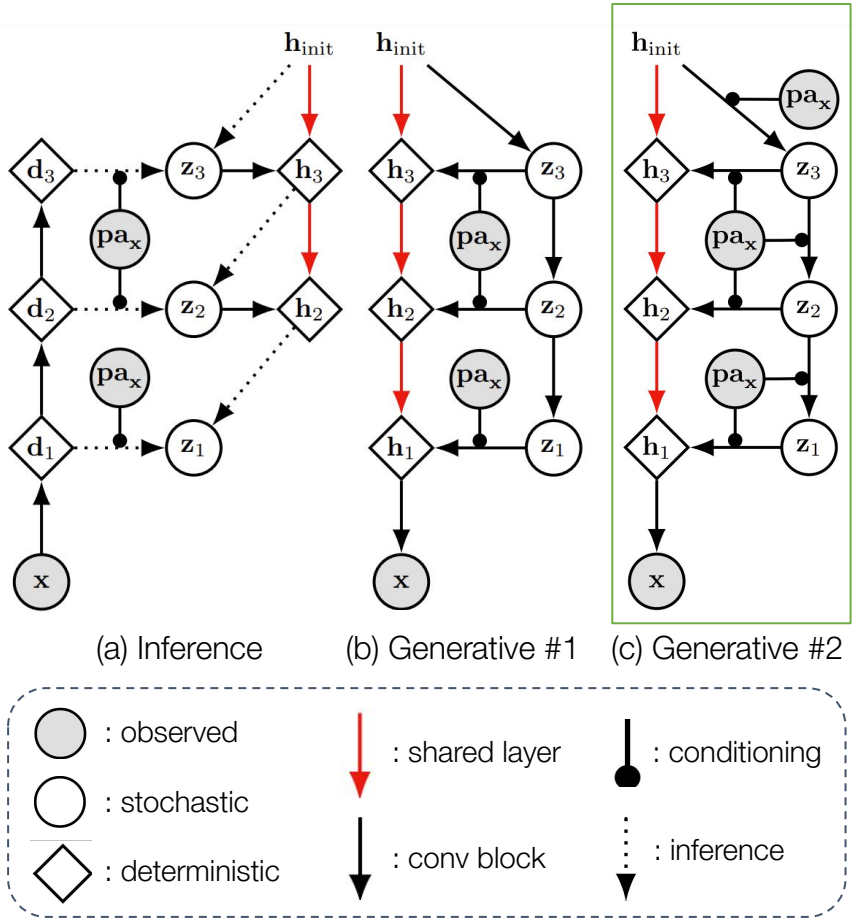
➔ Standard residual blocks:

Option 1:



Option 2:



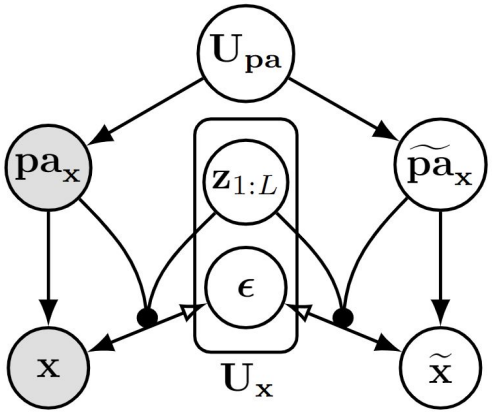


Latent Mediator Model

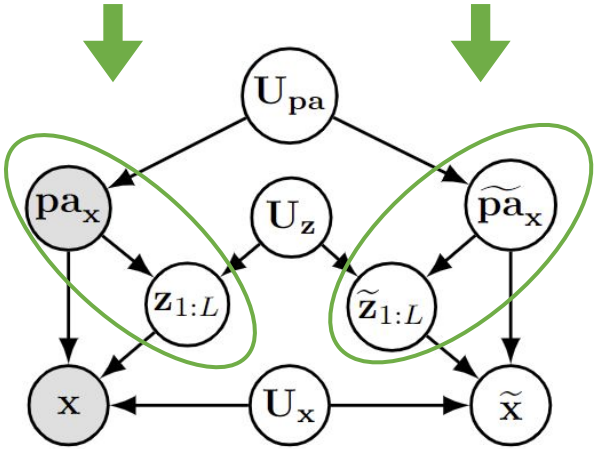
➔ The conditional prior $p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa}_\mathbf{x})$ induces a **latent mediator,** as $\mathbf{z}_{1:L}$ is no longer exogenous

➔ Nonetheless, the underlying SCM has a **Markovian** interpretation:

$$p(\mathbf{U}) = p(\mathbf{U}_\mathbf{x}) \left(\prod_{k=1}^{K} p(\mathbf{U}_{\mathbf{pa}_k}) \right) \left(\prod_{i=1}^{L} p(\mathbf{U}_{\mathbf{z}_i}) \right)$$



(a) Exogenous Prior: $p_\theta(\mathbf{z}_{1:L})$

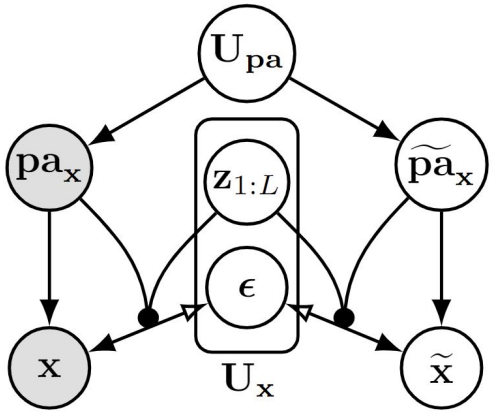


(b) Latent Mediator: $p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa}_\mathbf{x})$

Latent Mediator Model

Causal Mediation Analysis

The study of how a treatment effect is **mediated** by another variable, to help explain **why** or **how** an individual may respond to certain stimulus.



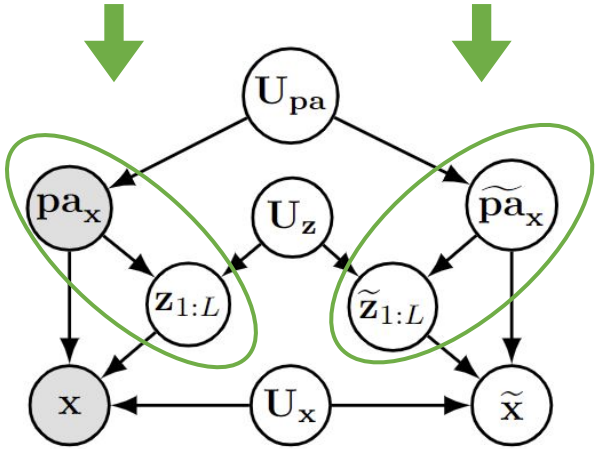
(a) Exogenous Prior: $p_\theta(\mathbf{z}_{1:L})$

➔ Enables estimation of **Direct (DE)**, **Indirect (IE)** and **Total (TE)** causal effects:

$$\mathrm{DE}_\mathbf{x}(\widetilde{\mathbf{pa}}_\mathbf{x}) = \mathbb{E}\left[g_\theta(\widetilde{\mathbf{pa}}_\mathbf{x}, \mathbf{z}_{1:L}) - g_\theta(\mathbf{pa}_\mathbf{x}, \mathbf{z}_{1:L})\right]$$

$$\mathrm{IE}_\mathbf{x}(\widetilde{\mathbf{z}}_{1:L}) = \mathbb{E}\left[g_\theta(\mathbf{pa}_\mathbf{x}, \widetilde{\mathbf{z}}_{1:L}) - g_\theta(\mathbf{pa}_\mathbf{x}, \mathbf{z}_{1:L})\right]$$

$$\mathrm{TE}_\mathbf{x}(\widetilde{\mathbf{pa}}_\mathbf{x}, \widetilde{\mathbf{z}}_{1:L}) = \mathbb{E}\left[g_\theta(\widetilde{\mathbf{pa}}_\mathbf{x}, \widetilde{\mathbf{z}}_{1:L}) - g_\theta(\mathbf{pa}_\mathbf{x}, \mathbf{z}_{1:L})\right]$$

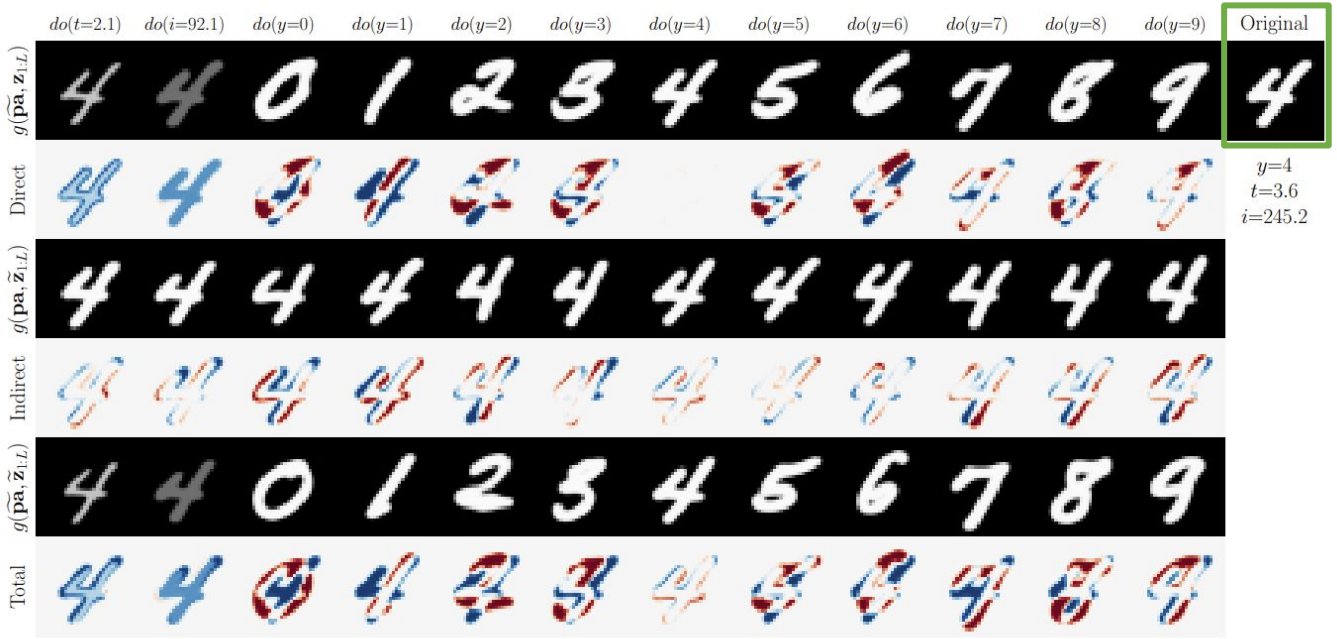


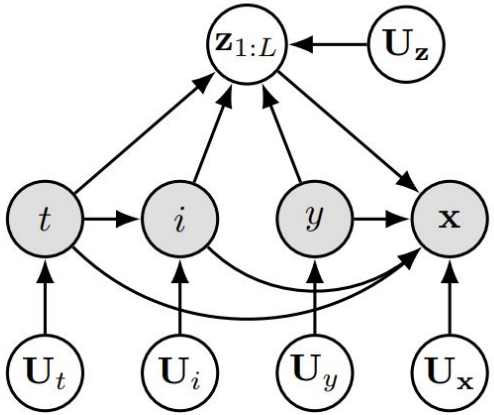
(b) Latent Mediator: $p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa}_\mathbf{x})$

Case Study: Morpho-MNIST

Hugging Face

Code & Demo





(a) Latent mediator SCM for Morpho-MNIST. Observed variables in the graph: image ($\mathbf{x}$), digit (y), stroke thickness (t) and pixel intensity (i).

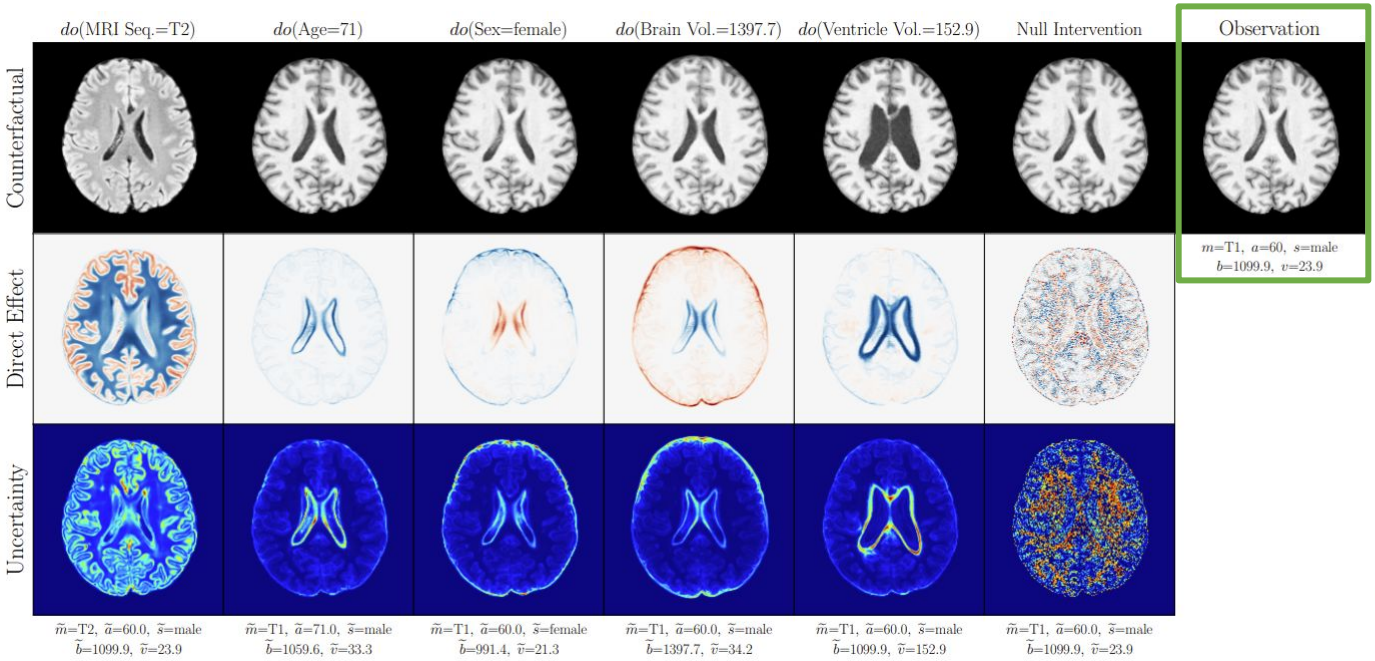
Figure 3: Morpho-MNIST counterfactuals from our latent mediator SCM. *Direct*, *indirect* and *total* causal effects of interventions are shown (red: increase; blue: decrease). Recall that $\widetilde{\mathbf{x}}_{\text{IE}} \sim g(\mathbf{pa}_{\mathbf{x}}, \widetilde{\mathbf{z}}_{1:L})$ are *cross-world* counterfactuals, i.e. the potential outcome of $\mathbf{x}$ given $\mathbf{pa}_{\mathbf{x}}$ and the (counterfactual) mediator we would have observed $\widetilde{\mathbf{z}}_{1:L}$ had $\mathbf{pa}_{\mathbf{x}}$ been $\widetilde{\mathbf{pa}}_{\mathbf{x}}$.

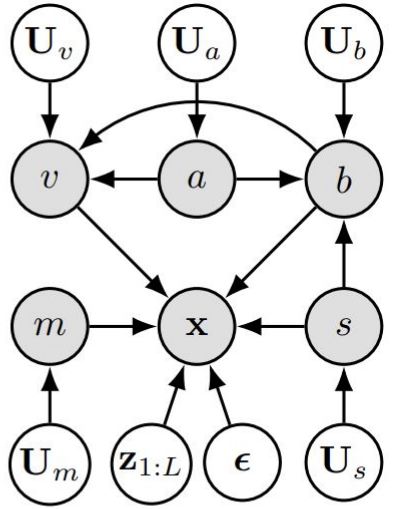
Castro, Daniel C., et al. Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. JMLR 2019

Case Study: Brain Imaging

Hugging Face

Code & Demo



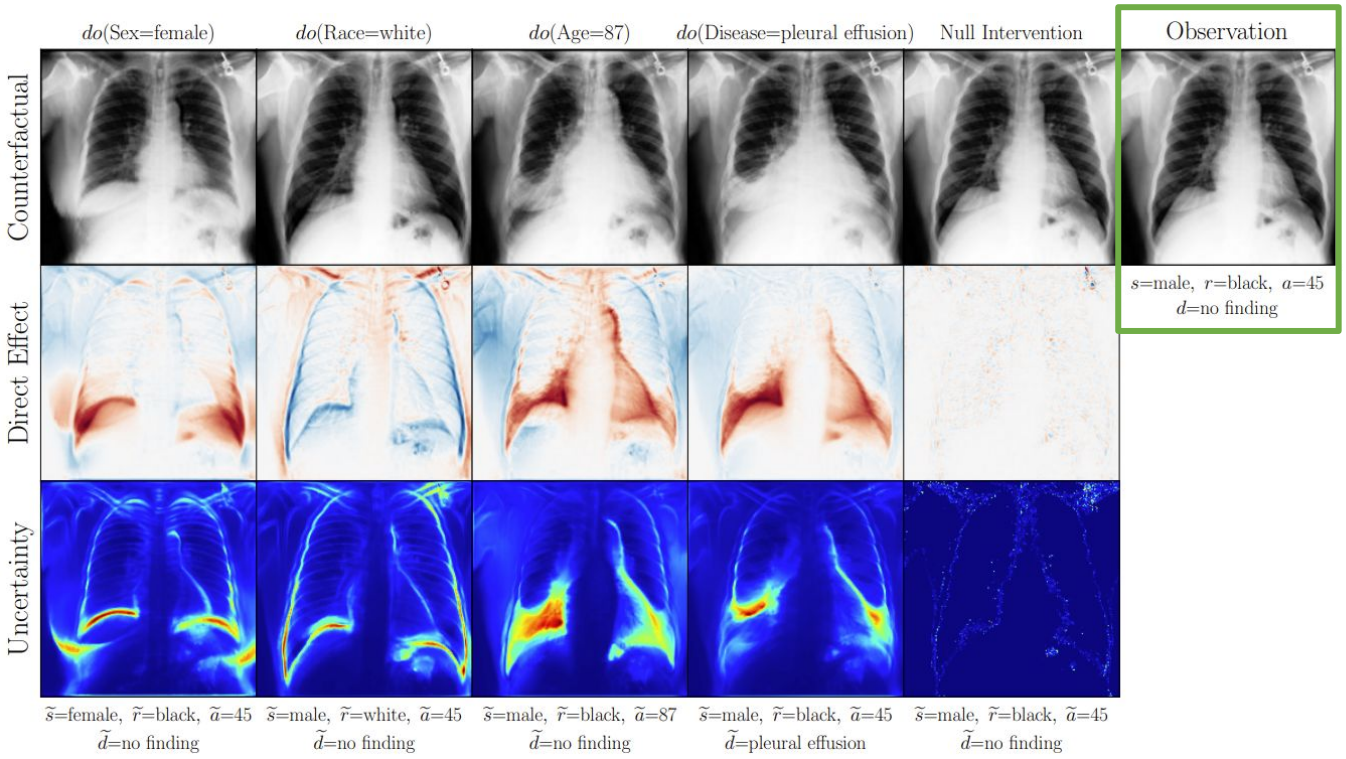


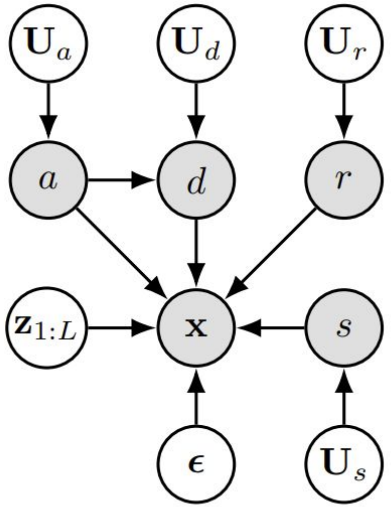
(a) Deep SCM for UK Biobank. MRI Seq. (m), age (a), sex (s), brain (b) & ventricle (v) volume.

Case Study: Chest X-ray Imaging

Hugging Face

Code & Demo





(a) Deep SCM for MIMIC-CXR. The variables in the causal graph are: age (a), sex (s), race (r), disease (d) and chest x-ray $(\mathbf{x})$. The disease d is pleural effusion.

De Sousa Ribeiro, Xia, Monteiro, Pawlowski, Glocker. High Fidelity Image Counterfactuals with Probabilistic Causal Models. *ICML* 2023

Evaluating Counterfactuals: Axiomatic Properties

- ➔ The **soundness theorem** states that the properties of **composition**, **effectiveness**, and **reversibility** are necessary in all causal models (Galles & Pearl, 1998).
- ➔ The **completeness theorem** states that these properties are sufficient (Halpern, 1998).
- ➔ We can measure **counterfactual soundness** using these axiomatic properties

Published as a conference paper at ICLR 2023

MEASURING AXIOMATIC SOUNDNESS OF COUNTERFACTUAL IMAGE MODELS

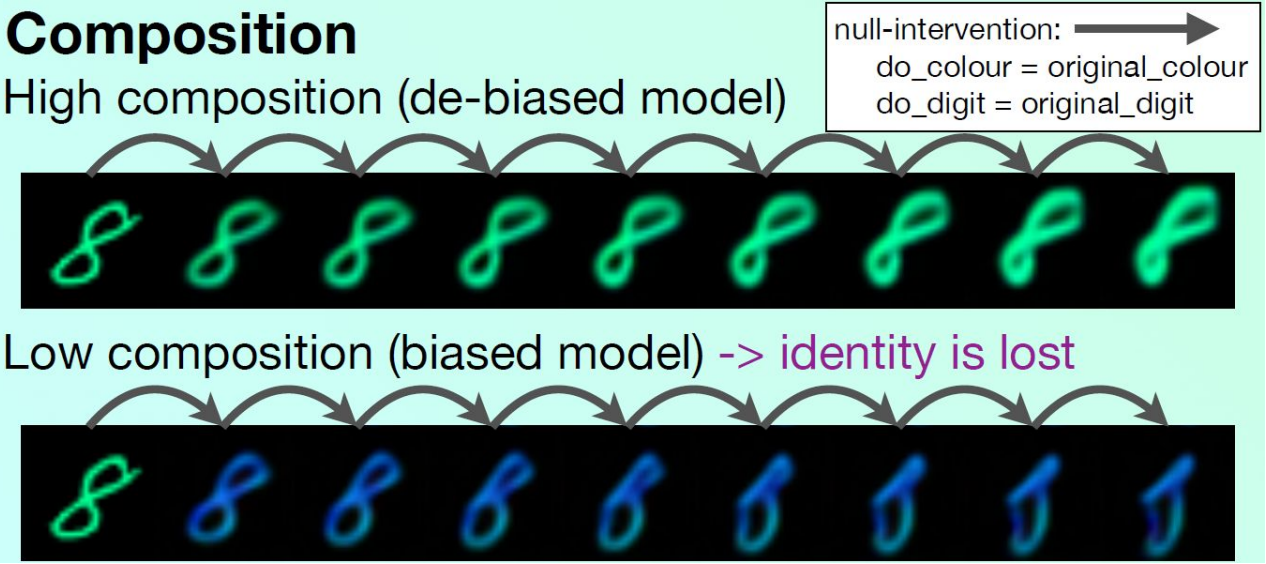
Miguel Monteiro[1†] **Fabio De Sousa Ribeiro**[1†] **Nick Pawlowski**[2] **Daniel C. Castro**[1,2] **Ben Glocker**[1]
[1]Imperial College London, [2]Microsoft Research Cambridge. [†]Joint first authors
{miguel.monteiro,f.de-sousa-ribeiro,b.glocker}@imperial.ac.uk

ABSTRACT

We present a general framework for evaluating image counterfactuals. The power and flexibility of deep generative models make them valuable tools for learning mechanisms in structural causal models. However, their flexibility makes counterfactual identifiability impossible in the general case. Motivated by these issues, we revisit Pearl's axiomatic definition of counterfactuals to determine the necessary constraints of any counterfactual inference model: *composition*, *reversibility*, and *effectiveness*. We frame counterfactuals as functions of an input variable, its parents, and counterfactual parents and use the axiomatic constraints to restrict the set of functions that could represent the counterfactual, thus deriving distance metrics between the approximate and ideal functions. We demonstrate how these metrics can be used to compare and choose between different approximate counterfactual inference models and to provide insight into a model's shortcomings and trade-offs.

David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. Foundations of Science, 3:151–182, 1998.
Joseph Y. Halpern. Axiomatizing causal reasoning. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98, San Francisco, CA, USA, 1998.

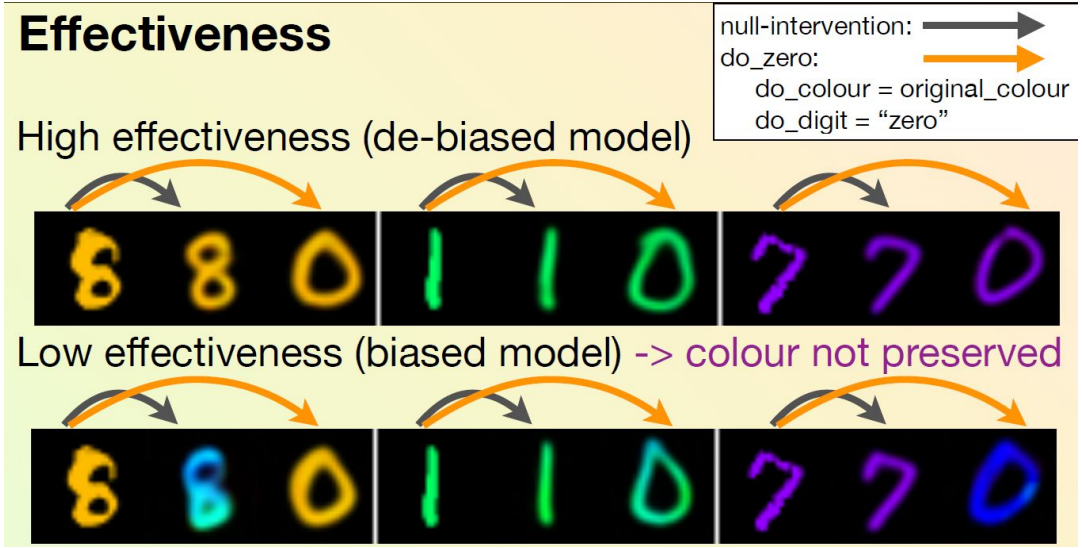
Example: Composition Axiom



Definition. Intervening on a variable to have the value it would otherwise have without the intervention will not affect other variables.

➔ Implies the existence of a **null-intervention**

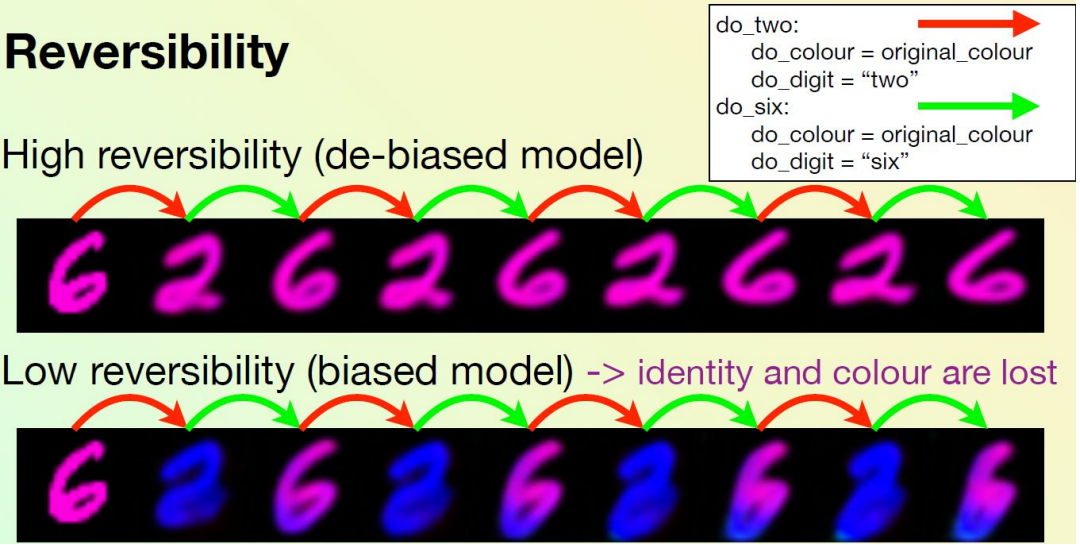
Example: Effectiveness Axiom



Definition. Intervening on a variable to have a specific value will cause the variable to take on that value.

➔ Caveat: often relies on a **pseudo-oracle**, e.g. a classifier/regressor

Example: Reversibility Axiom

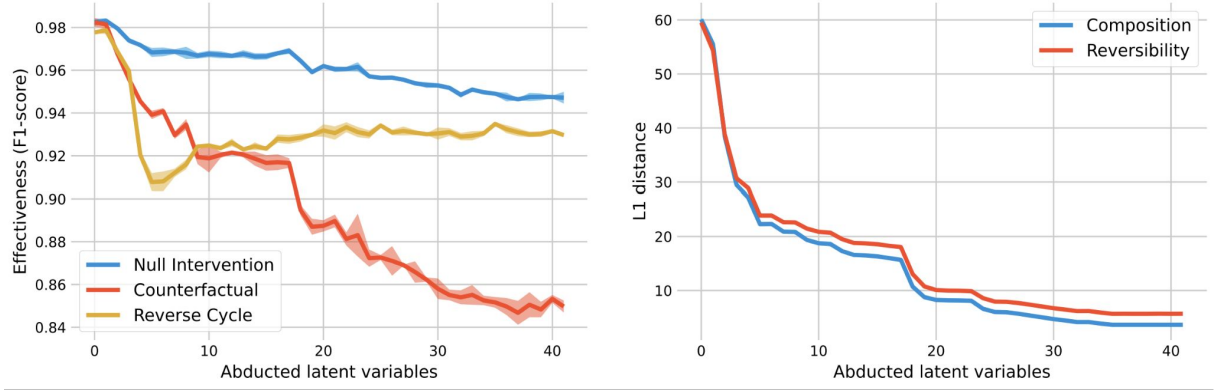


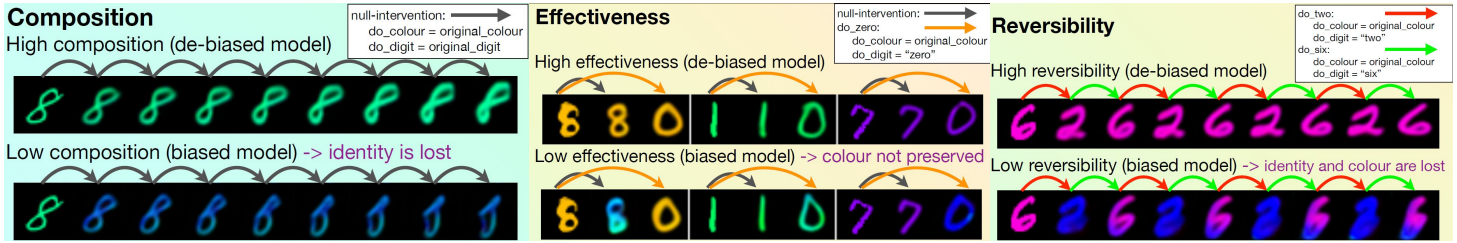
Definition. Precludes multiple solutions due to feedback loops. If setting a variable *X* to a value *x* results in a value *y* for a variable *Y*, and setting *Y* to a value *y* results in value *x* for *X*, then *X* and *Y* will naturally take on the values *x* and *y*.

➔ Follows directly from **composition** in recursive systems such as DAGs

Evaluating Counterfactuals: Axiomatic Properties

➔ **TLDR:** We identify an inherent trade-off





Monteiro, De Sousa Ribeiro, Pawlowski, Castro, Glocker. Measuring Axiomatic Soundness of Counterfactual Image Models. *ICLR* 2023
De Sousa Ribeiro, Xia, Monteiro, Pawlowski, Glocker. High Fidelity Image Counterfactuals with Probabilistic Causal Models. *ICML* 2023

Conclusion & Outlook

➔ Deep SCMs can generate plausible high-fidelity **counterfactuals** of medical images as measured by **axiomatic soundness** of counterfactuals

➔ Tractable estimation of **direct**, **indirect** and **total** causal effects for high-dimensional structured variables

➔ Limitations:

- Only consider Markovian SCMs; although Markovianity is a common assumption in causality literature, it is strong in most cases
- Measuring counterfactual effectiveness relies on separately trained classifiers

Conclusion & Outlook

➔ Deep SCMs can generate plausible high-fidelity **counterfactuals** of medical images as measured by **axiomatic soundness** of counterfactuals

Future Work

I. Targeted **data augmentation** to improved **robustness**, **sample efficiency** & **fairness**
II. Providing principled **causal explanations**, e.g. through mediation analysis
III. Improving **counterfactual soundness**, via ML techniques and/or medical guidance
IV. Provide theoretical guarantees of **identifiability** under plausible causal assumptions
V. Extensions to **Semi-Markovian** and **Non-Markovian** settings

Imperial College London

Hugging Face

Code & Demo

Thank you for listening

Happy to take questions!

fdesousa@ic.ac.uk


Fabio De Sousa Ribeiro


Miguel Monteiro


Nick Pawlowski


Daniel C. Castro


Tian Xia


Ben Glocker

BioMedIA

project mira ERC StG 2017 grant No. 757173

erc

Microsoft