Department of Computing

Imperial College of Science, Technology and Medicine

Causality and Generative Models

This week's paper is [1]

Author: Anton Zhitomirsky

Contents

1 Causality **2**

1.1 What is Causlity? . . . 2
1.2 Reichenbach's Common Cuase Principle . . . 2
1.3 Simpson's Paradox . . . 3
1.4 Predictive Modelling . . . 4
1.4.1 Example . . . 5
1.5 Ladder of Causation . . . 6
1.6 Structural Causal Models . . . 6
1.7 Distribution Math . . . 8
1.7.1 Computing Counterfactuals Example . . . 10
1.8 Deep Structural Causal Models . . . 10
1.8.1 Normalizing Flows . . . 11
1.8.2 VAEs . . . 13
1.9 Evaluating Counterfactual . . . 20
1.9.1 Composition Axiom . . . 21
1.9.2 Effectiveness Axiom . . . 22
1.9.3 Reversibility Axiom . . . 22

Bibliography **24**

1 Causality

1.1 What is Causlity?

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

Figure 1: Correlation doesn't imply casuation

1.2 Reichenbach's Common Cuase Principle

Reichenbach's Common Cause Principle

Hans Reichenbach

I. Chocolate consumption causes Nobel prize win

$$C \rightarrow N$$

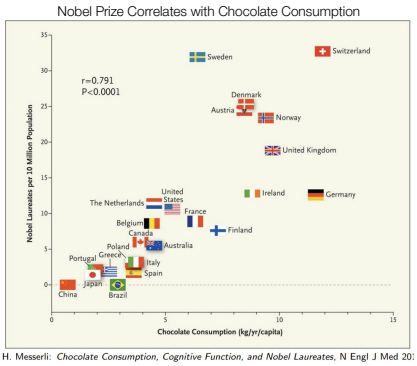
II. Nobel prize win causes chocolate consumption

$$C \leftarrow N$$

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

$$C \leftarrow U \rightarrow N$$

Could U be GDP per capita or wealth per adult?



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

Figure 2: there is a case that countries that eat the most chocolate have the most Nobel laureates. So there are three possible explinations for the correlations. This is the **Reichenbach's Common Cause Principle**

1.3 Simpson's Paradox

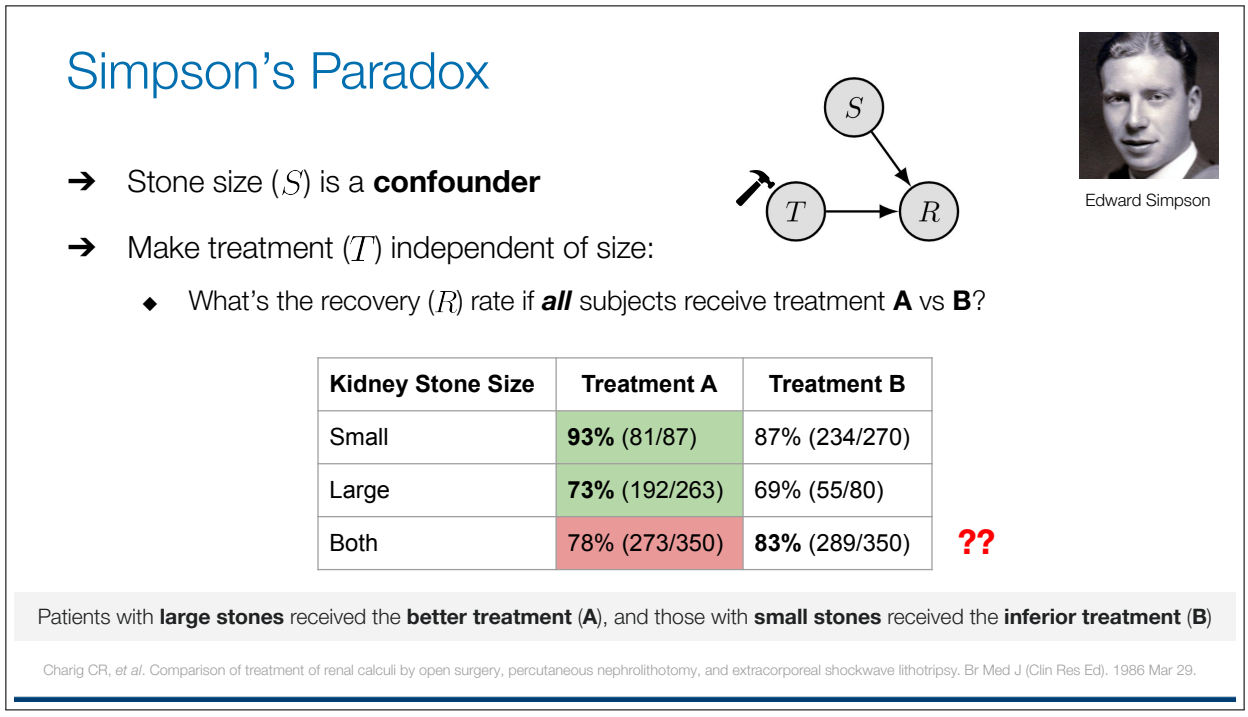


Figure 3: Another paradox includes two treatments for different sizes of kidneys. Treatment A has better recovery rates, but Treatment B has an average better recovery rate. This is the **Simpson's Paradox**. The reason is that the stone size is a confounding variable. What happened was that the doctors were prescribing the 'better' treatments for the pateitns with the worst conditions, which caused a bias.

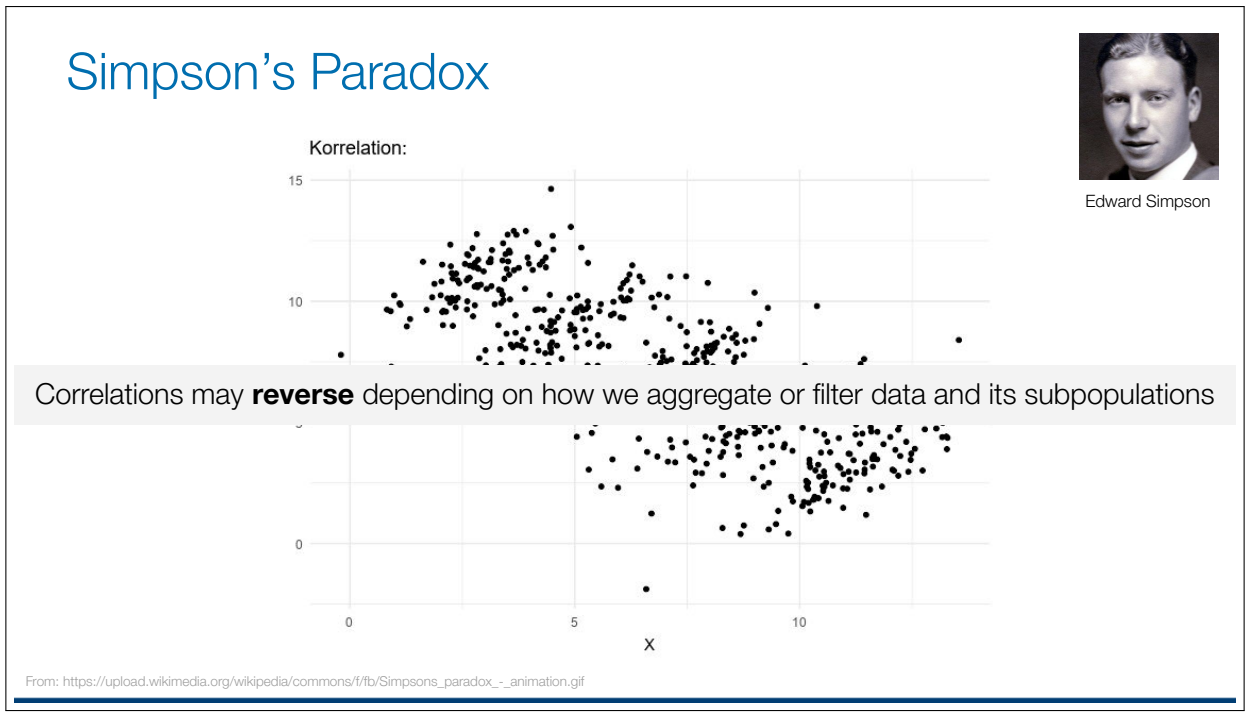


Figure 4: depending on the way you aggregate on your populations, the correlations may reverse (e.g. a downward slope vs many upwards sloping lines)

1.4 Predictive Modelling

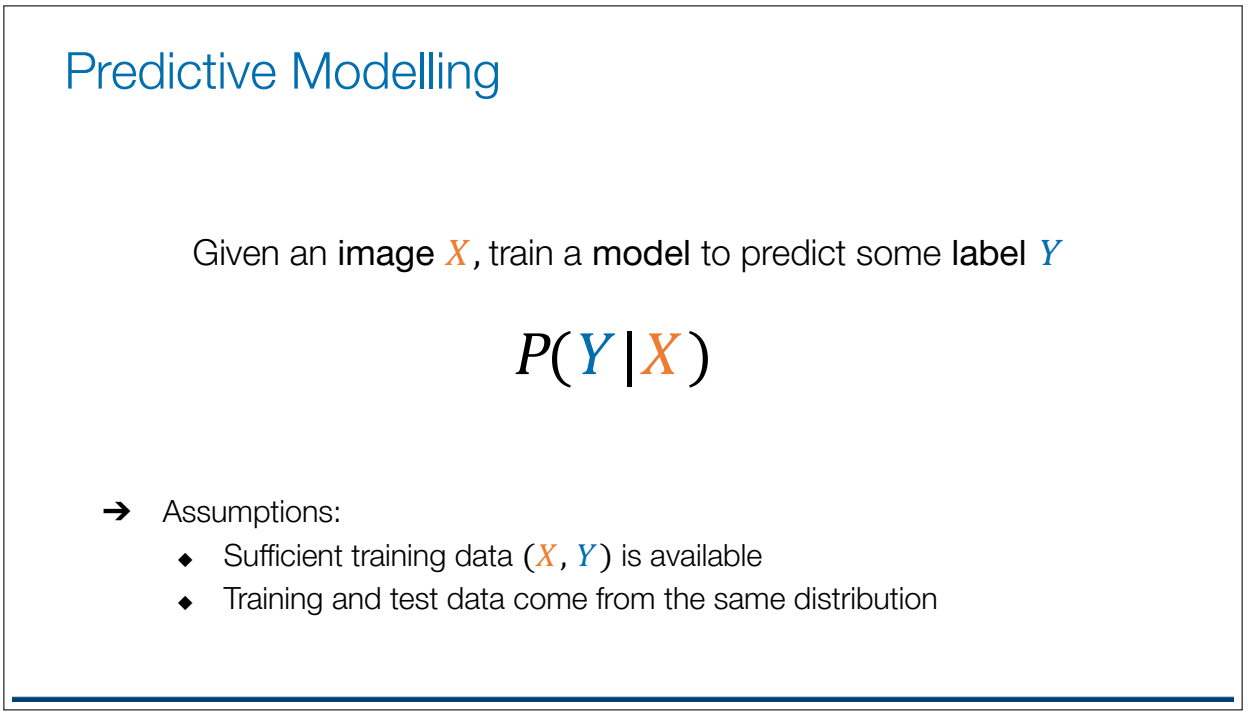


Figure 5: The goal is to estimate $P(Y|X)$ given some assumptions.

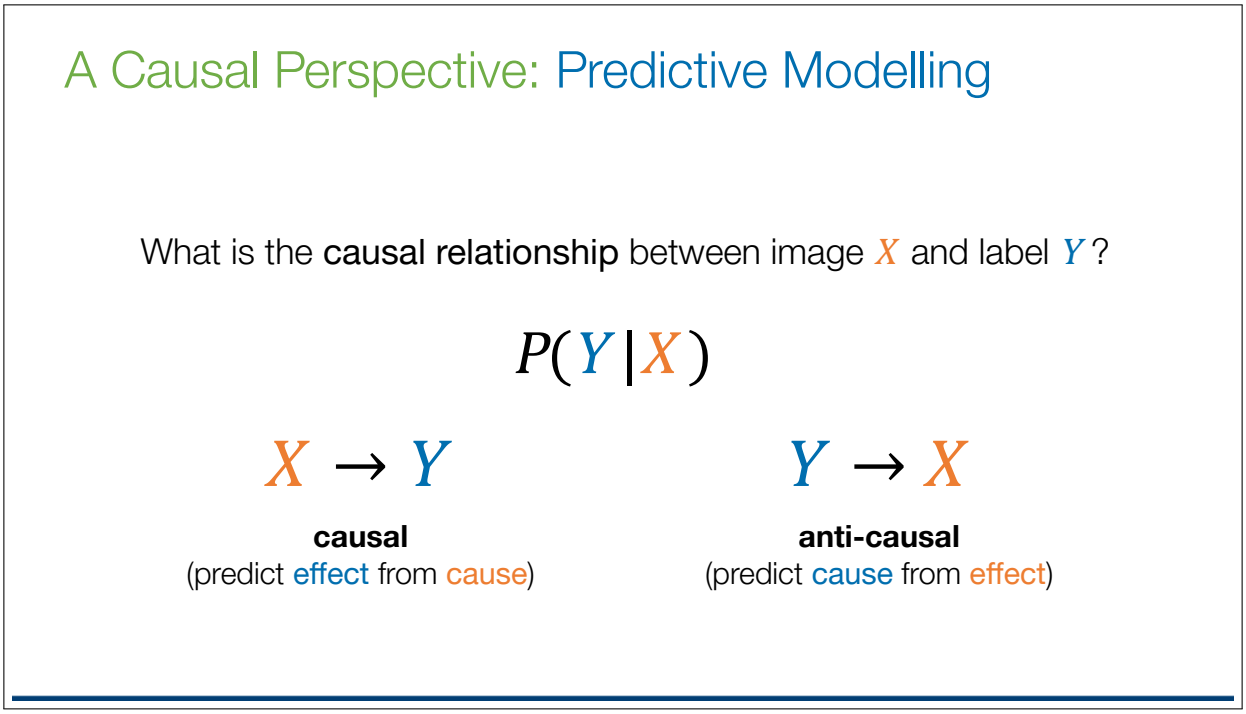
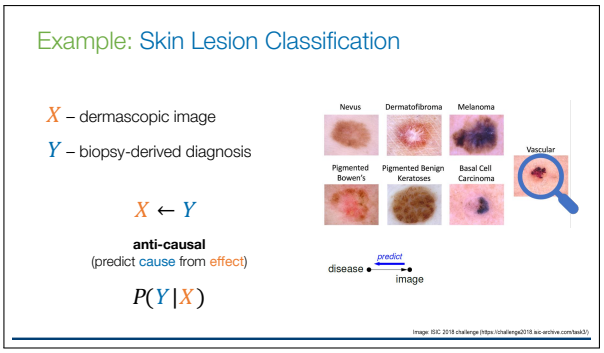
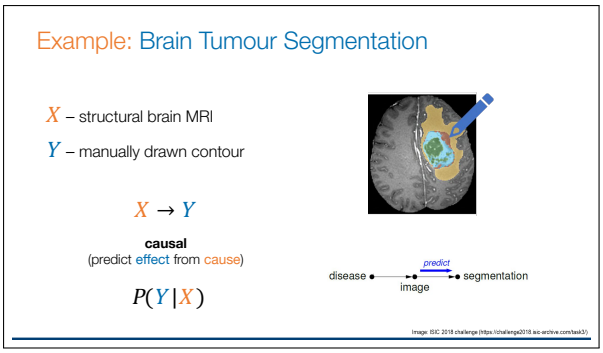


Figure 6: Firstly we are predicting the effeect from the cause, and secondly we are predicting the cause from the effect.

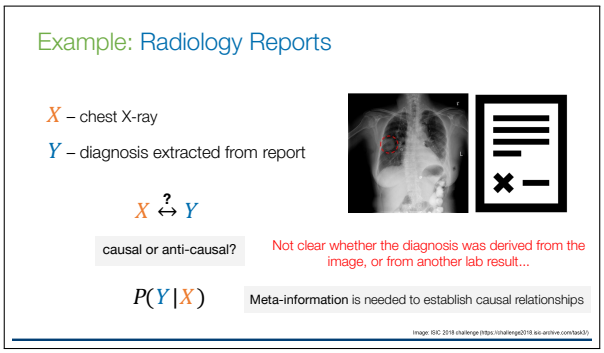
1.4.1 Example



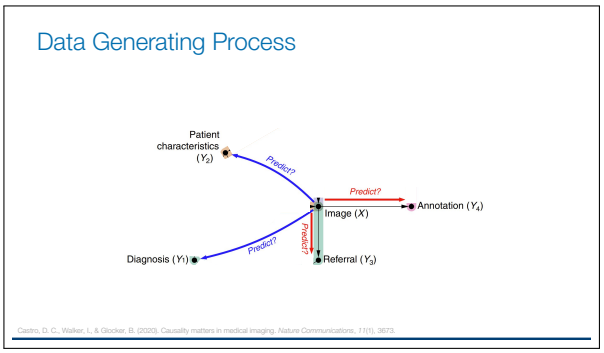
(a) the disease makes the image look a certain way, however, the thing we are predicting comes from a biopsy, not the image.



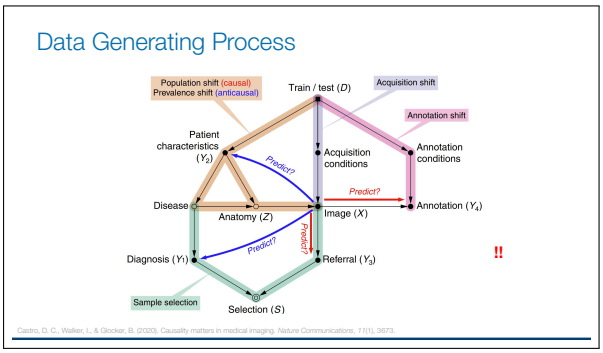
(b) the segmentation is fully determined by the iamge. Therefore, the image causes the segmentation map



(c) sometimes you need more information to determine



(d) we need to be careful about our data generating process



(e) complications may arise when you annotate the data, becuase we get annotation shifts. When we acquire the data there can be acquisition shifts and so on. E.g. an annotator might annotate inconsistently to the annotations.

1.5 Ladder of Causation

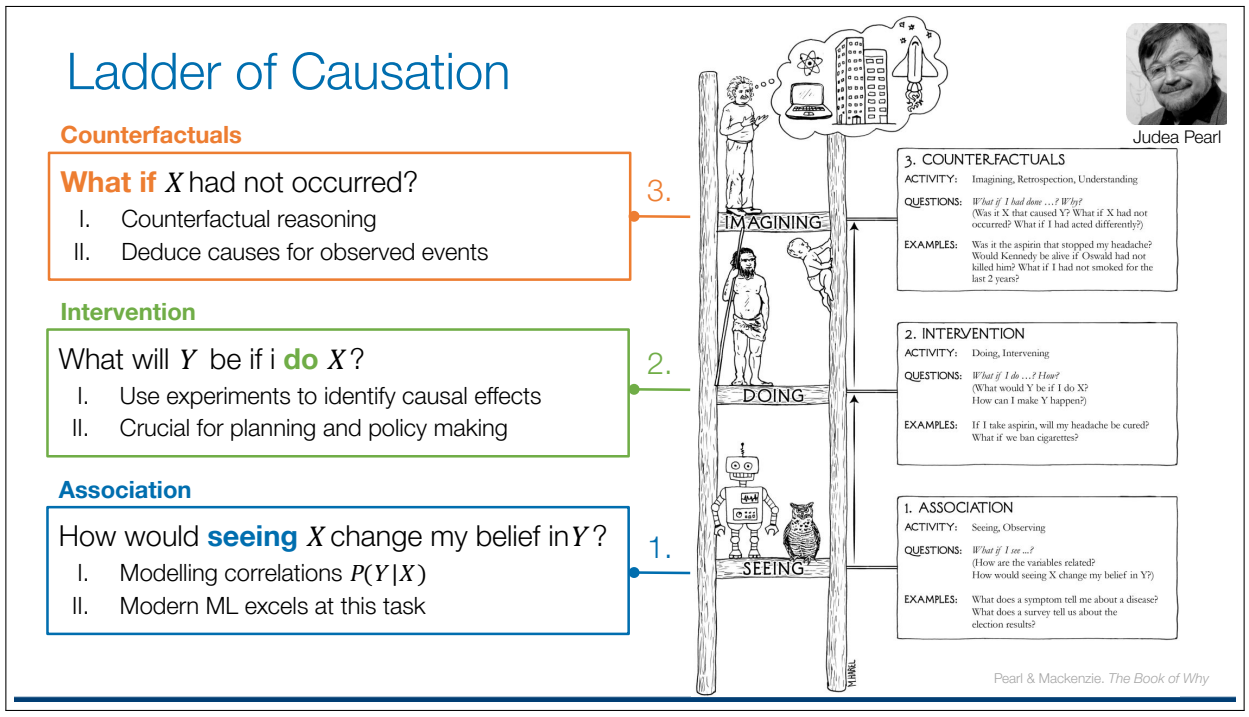


Figure 7: "the first rung deals with association, so how would seeing X change my belief in Y? E.g. in deep learning models excel at this task if we have enough data and a big enough neural network. The second rung deals with intervention. So this is about asking questions like what will Y be if i do X? The final does counterfactuals, these are hypothetical questions like what would happen if X had not occurred.

1.6 Structural Causal Models

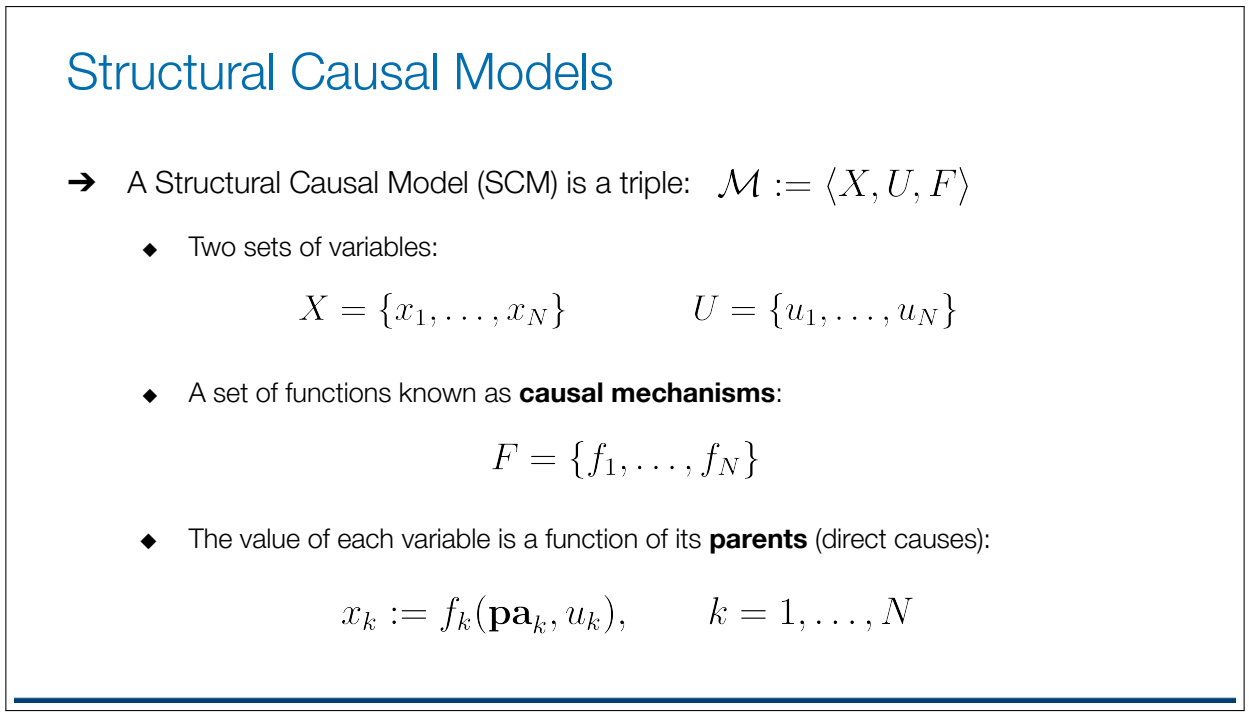


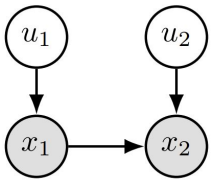
Figure 8: X is the set of observations, and U are unobvserved conditions (e.g. background, unaware or things you have no control over.) Lastly, the idea is that each X variable is a function of its direct causes via these causal mechanisms.

For example:

Structural Causal Models

➔ Example: A simple SCM:

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

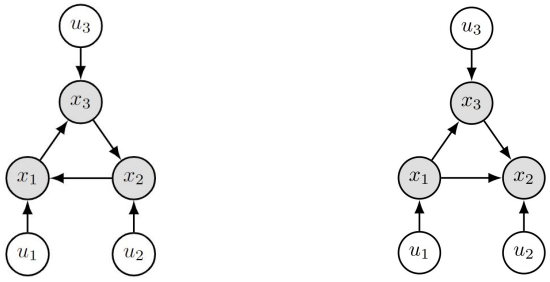


$$x_1 := f_1(u_1), \qquad u_1 \sim \mathcal{N}(0,1)$$
$$x_2 := f_2(x_1, u_2), \qquad u_2 \sim \mathcal{N}(0,1)$$

Figure 9: endogenous: caused by any variable in the model, **exogenous**: caused by factors external to the model or independent of the model.

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

Figure 10: We can represent Acyclic SCMs as a DAGs

Structural Causal Models

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

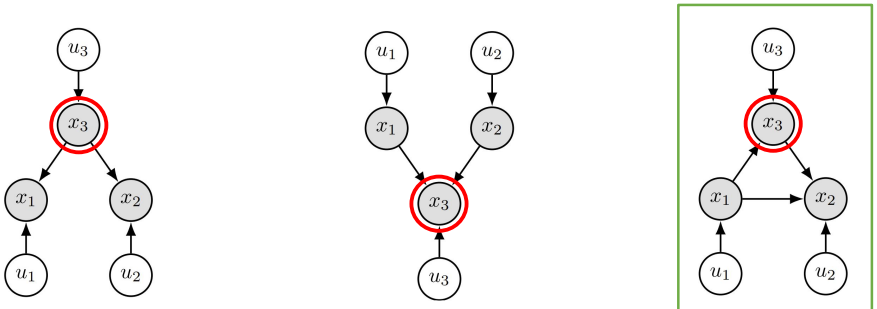


Figure 11: a) both the kidney and chocolate example, b) when you consider on x_3 you would also experience correlation between x_1 and x_2. c) x_3 mediates the causal effect of x_1 on x_2.

1.7 Distribution Math

Observational Distribution (rung 1.)

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

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

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

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

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

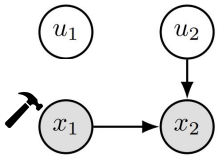
Figure 12: 1) distribution of the noise terms factorize accross every node in the graph. 2) these also factorize over every node in the graph.

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**: $P_{\mathcal{M}_c}(X \mid do(c))$

Figure 13: we intervene with the do operator. This produces a submodel. This is generall ydifferent to the observational distribution we start out with before interveneing on anything.

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?

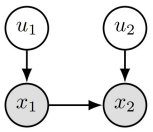
➔ **Counterfactual inference** involves three steps:

1. **Abduction:** Update $P(U)$ given observed evidence, i.e. infer posterior $P(U \mid X)$
2. **Action:** Perform an intervention e.g. $do(\tilde{x}_k := c)$ 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

Figure 14: a way to solve this is counterfactual inference.

1.7.1 Computing Counterfactuals Example

Example: Computing Counterfactuals

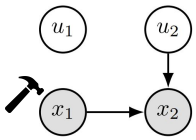


$$x_1 := f_1(u_1) = 1 + u_1$$
$$x_2 := f_2(x_1, u_2) = 3x_1 + u_2$$

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

1.8 Deep Structural Causal Models

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?

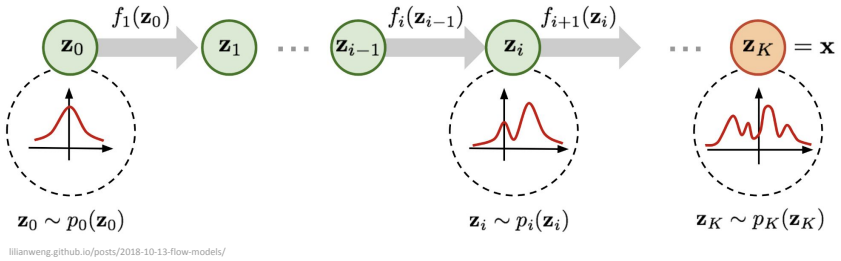
[1] Pawlowski, Castro, Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. *NeurIPS* 2020

Figure 15: Can we use DGMs to learn structural clausal mechanisms? Abducition is challenging in medicine.

1.8.1 Normalizing Flows

Deep Mechanisms: Normalizing Flows

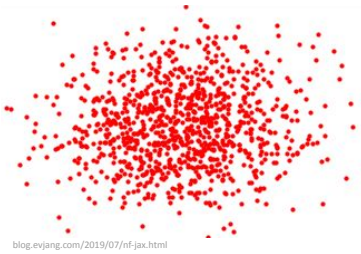
➔ Normalizing Flows (NFs) build complex probability distributions via successive (invertible) transformations of simple distributions



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



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

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

Figure 16: Here the dataset is transforming into a spiral after multiple transformations. We attempt to maximilse the likelihood of observing the data. The benefit is that they allow us to learn invertible mechanisms and deterministic abduction then we can invert it (since we constrained it to be invertible) and find the exogenous noise term u_x given the x

Intuition: Normalizing Flows

➔ Change-of-variables formula:

$$p(x) = p(u)\left|\frac{du}{dx}\right|$$

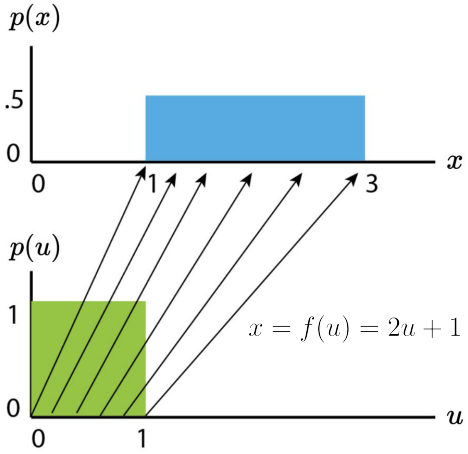
volume correction

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

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

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

$X = f(U)$



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

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

Figure 17: "Lets say that we have two random variables which are related by a deterministic mapping. Lets say this function is $f(u) = 2u+1 = x$. The change in variables says that we can evaluate the density of X as the density of u times the volume correction term which tries to preseve density when we apply this function. Therefore, we invert the function and differentiate it. After this we can see that for any value of u the density of x is just one half because we have $1 \times \frac{1}{2}$."

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!

Figure 18: This concept is what we use to train the noramlizing cluase. We take the log of it and use that as the training objective.

1.8.2 VAEs

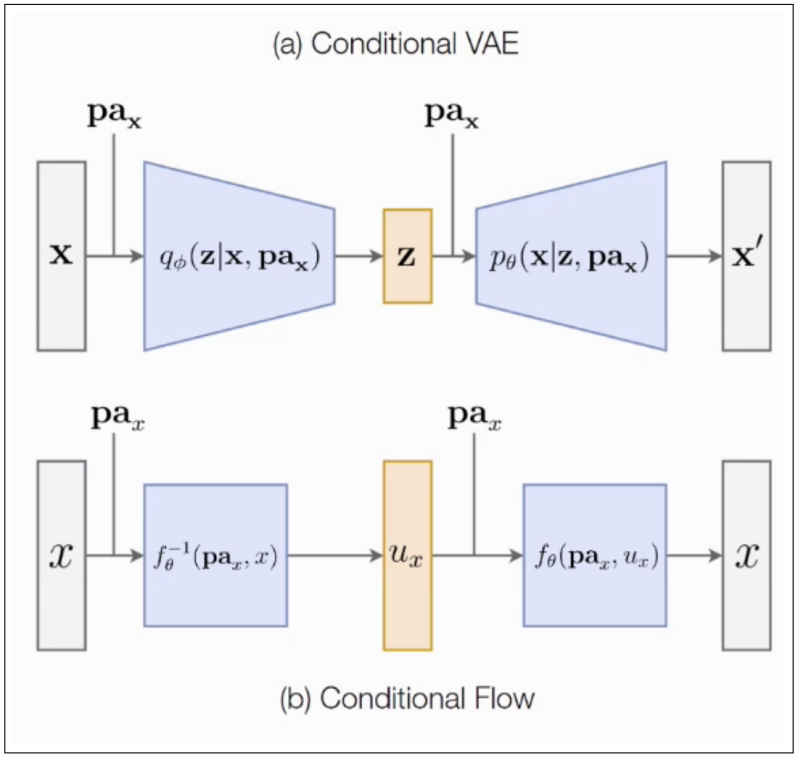


Figure 19: "conditional vae and the conditional flow look the same. The catch here, is that the attention step is now non-deterministic because the ecnoder produces a distribution of all these encodings and the decoder also produces a distirbution over pixel values."

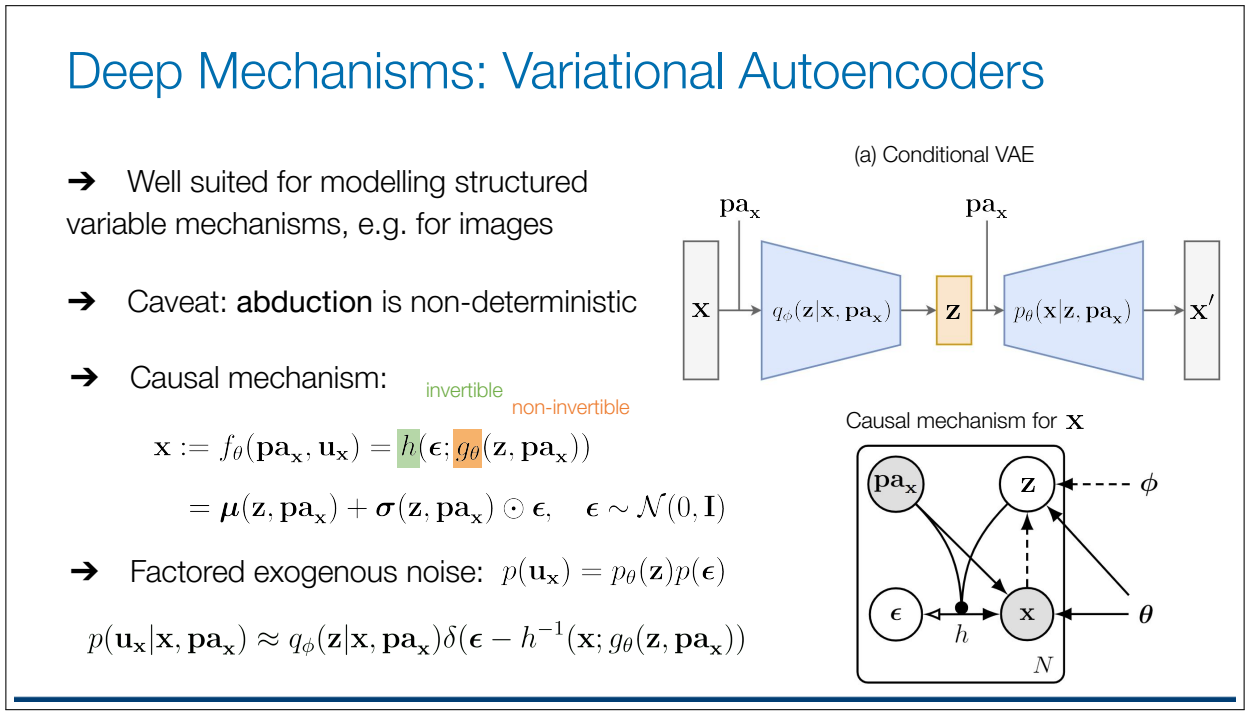


Figure 20: "We can define the causal mechanism for an image X. The parents of X here are meta infromation like age etc." We have g that is decoder model that predicts the per pixel mean and variances. and the H mechanism is how we smaple from this distribution. The encoder is approximately invertible. Therefore, we have a factorizeed exogenous noise decomposition. To calucalte the posterior it is apprxoimated by the encoder mapping and the inversion of the dterministic sampling procedure

Deep Mechanisms: Variational Autoencoders

➔ Example: computing counterfactuals:

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

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

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

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

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

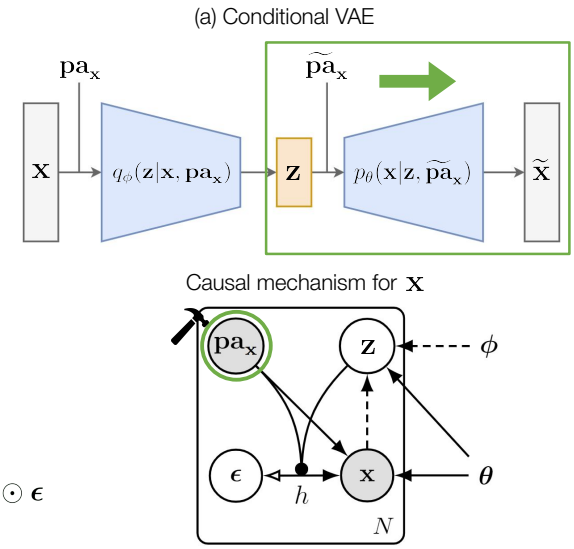


Figure 21: 1) encoding: pass it through encoder and receive z. Then decode. after you invert the sampling in the observation space so given its an image X we subtract the predicted mean and divide by the predicted variabce per pixel. 2) next step is to perform some intervention upsteam. e.g. if we have a causal graph with various dependencies between the parents, if we intervene on one of them and compute the value of the new parents, (the $\tilde{\cdot}$ values) and set these to 0 values. 3) in the final step we pass the new values through the decoder and plug values of epsilon and z that we abducted in the abduction step as well as the new values of the parents to condition with the decoder. *This applies to, if we want to see what the image would look like if we had a different age.*

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:

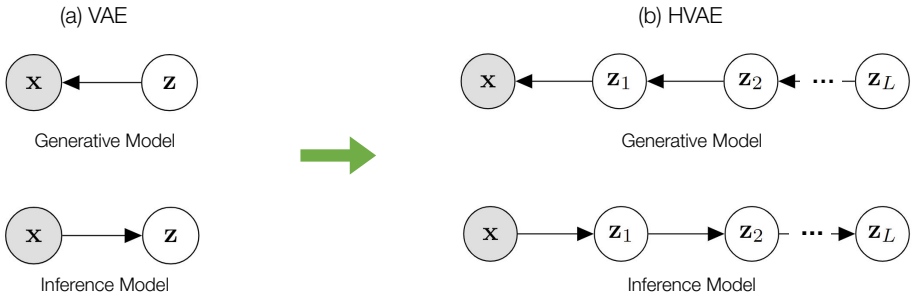


Figure 22: This is not scalable to high resolutions, so we need a more powerful generative model that is still amenable to principal deduction ut can also scale up to higher resolutions. A natural solution is the heirarchical vae; here we have L latent variables instead of just one.

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}

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

Figure 23: The goal is simliar; optimize model to be close to a given data distribution. The only difference from the VAE objective is that we now have a multitude of latent variables and factorize across the L layers.

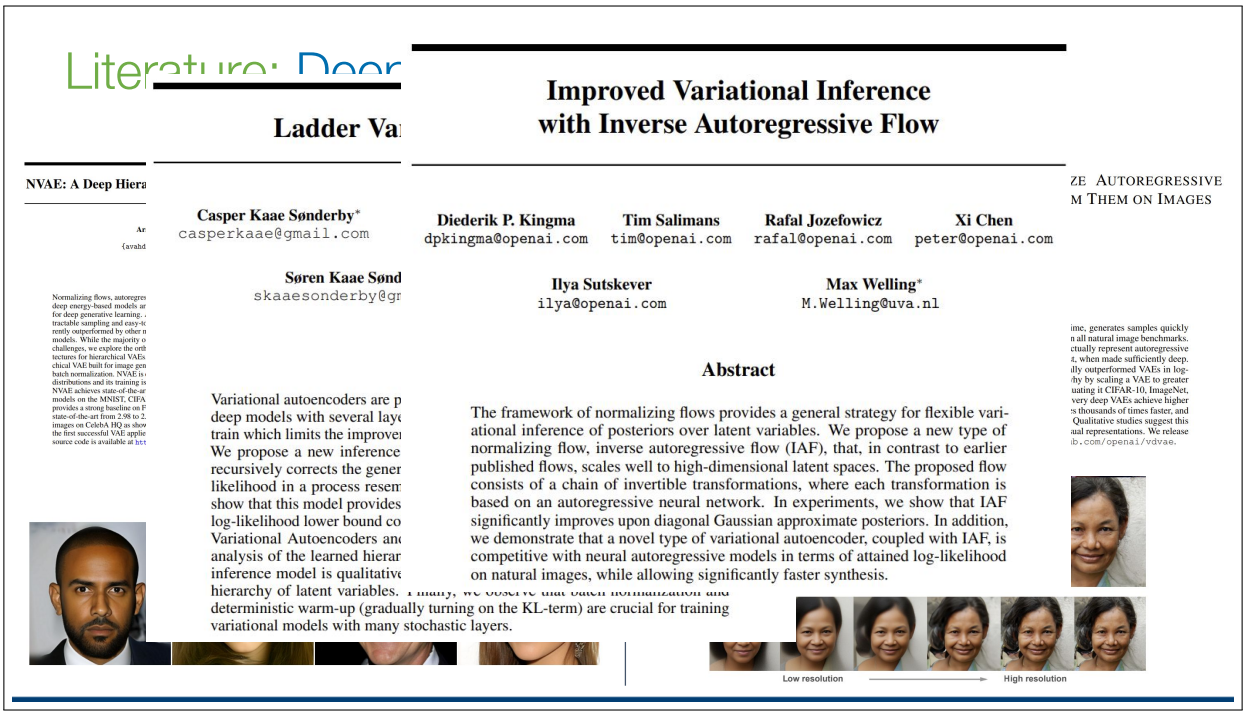


Figure 24: caption

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

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

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

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

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*.

Figure 25: boxed term is MSE loss. point being here is that it is a hierarchcal VAE with a different inference model.

Deep Mechanisms: Conditional HVAEs

➔ Generative model structures:

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

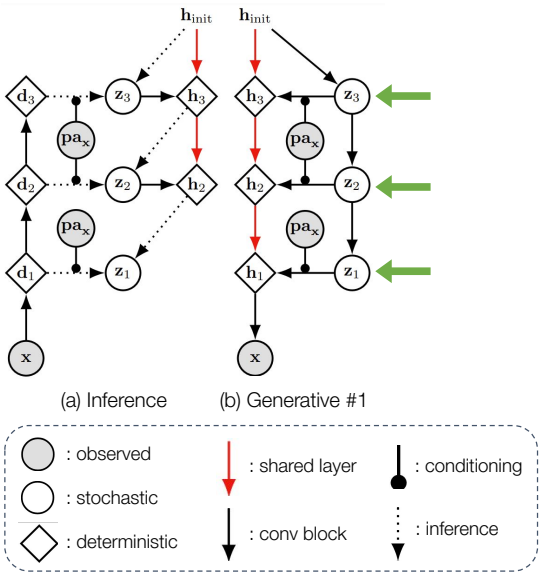


Figure 26: "We had to figure out a couple of ways to condition these VAEs. The first, is to retrain the role of the latent variable as being part of the exogenous noise for X. in this case, the latent varibales are independent o fhte values of the parents. So when inferring these lantent vairbales they on't get acecso the vlaues of the parents.'

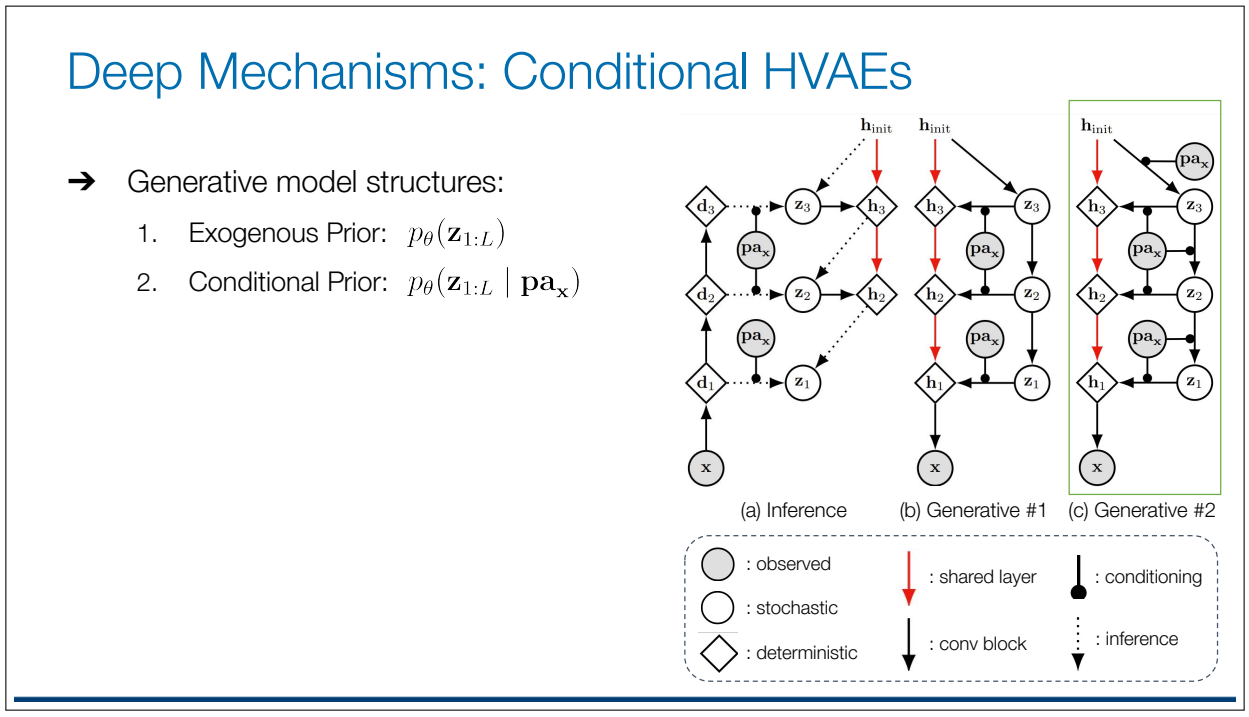


Figure 27: The second option is to include a conditional prior, so the latent varibles do see the values of the parents.

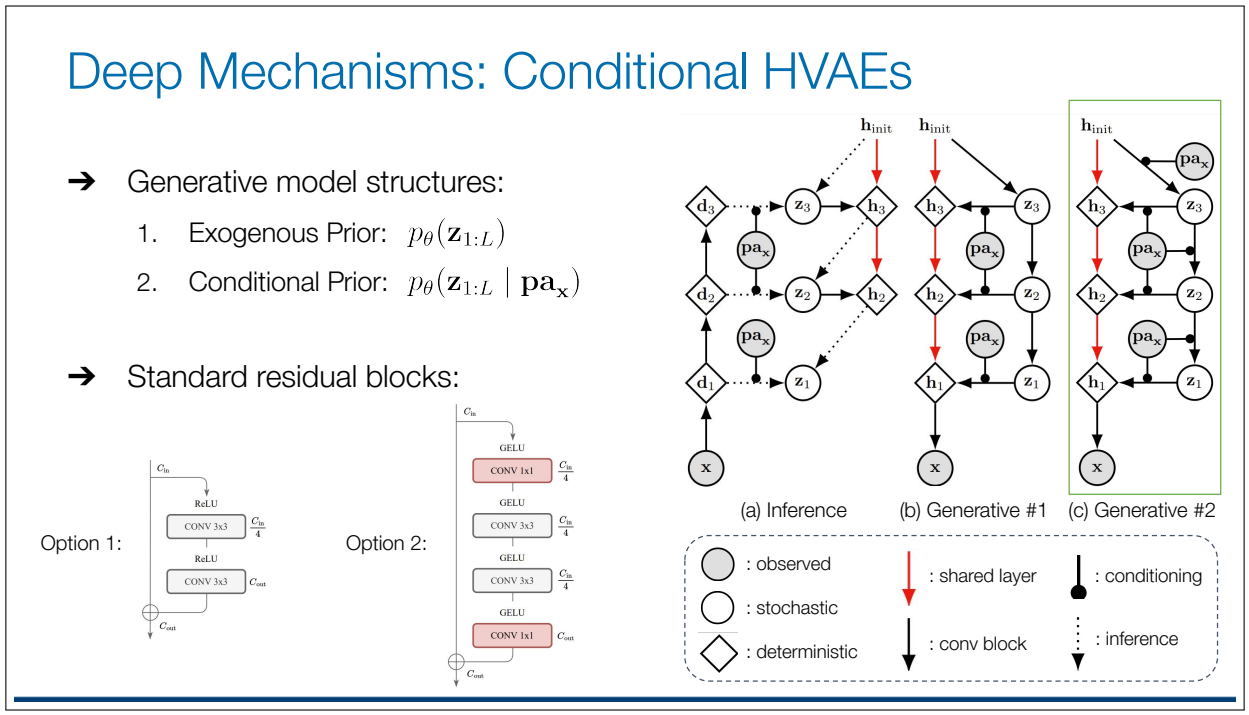


Figure 28: The way you construct this is by stacking residula blueing blocks into a u-net style architecture. These should then be trained to predict the mean and variabce of the latent variables.

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:

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

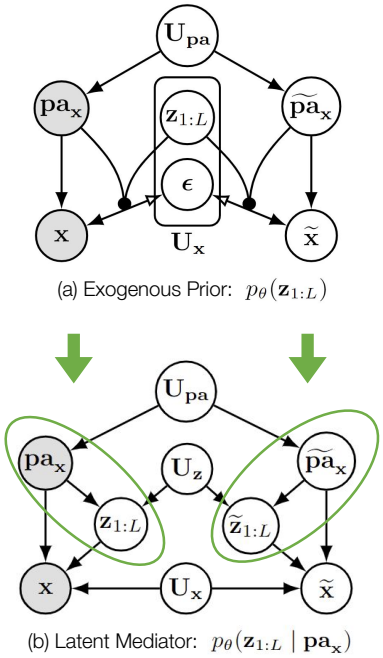


Figure 29: using this conditional prior as an implication induces a latent mediateor. These latent varibales are no longer part of the exogenous noise, this is becuase they are caused by things within the model (there is a directed arrow from the parents to the latent variables now). This is still Markovian because the way you give rise to these latent vairblaes still depends on some external independent noise when you sample them.

In these diagrams, the grey nodes are observed variables, the middle are exogenous noise u terms and on the right we have unobserved counterfactual values. The key thing is that the exogenous noise is shared between the factual observed and counterfactual branches.

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.

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

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

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

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

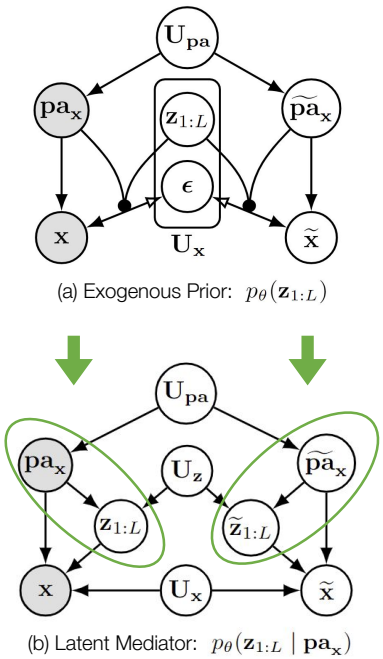


Figure 30: this allows you to apply causal mediation analysis defintion above. the main takeaway, is how the parents affect the image vs how the latent variables affect the image. We can visualise this and understand what our interventions are doing when we're chanign things and confusing counterfactuals.

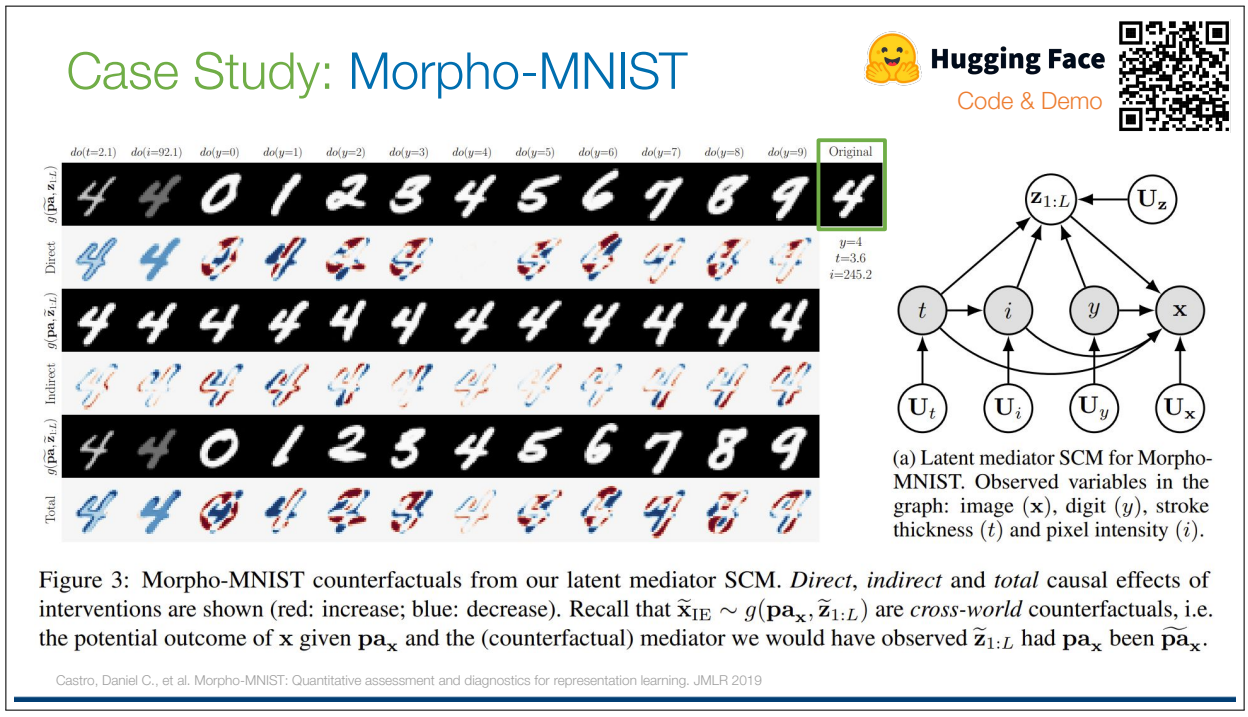


Figure 3: Morpho-MNIST counterfactuals from our latent mediator SCM. *Direct*, *indirect* and *total* causal effects of interventions are shown (red: increase; blue: decrease). Recall that $\widetilde{\mathbf{x}}_{\text{IE}} \sim g(\mathbf{pa}_\mathbf{x}, \widetilde{\mathbf{z}}_{1:L})$ are *cross-world*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

Figure 31: Here we have 4 observed variables and a predefined causal graph. Each column is a different intervention; so we can change the digits, the thickness and the intensity. We can see that when we change the thickness, it obeys the causal graph (thickness causes intensity but the same is not true the other way around).

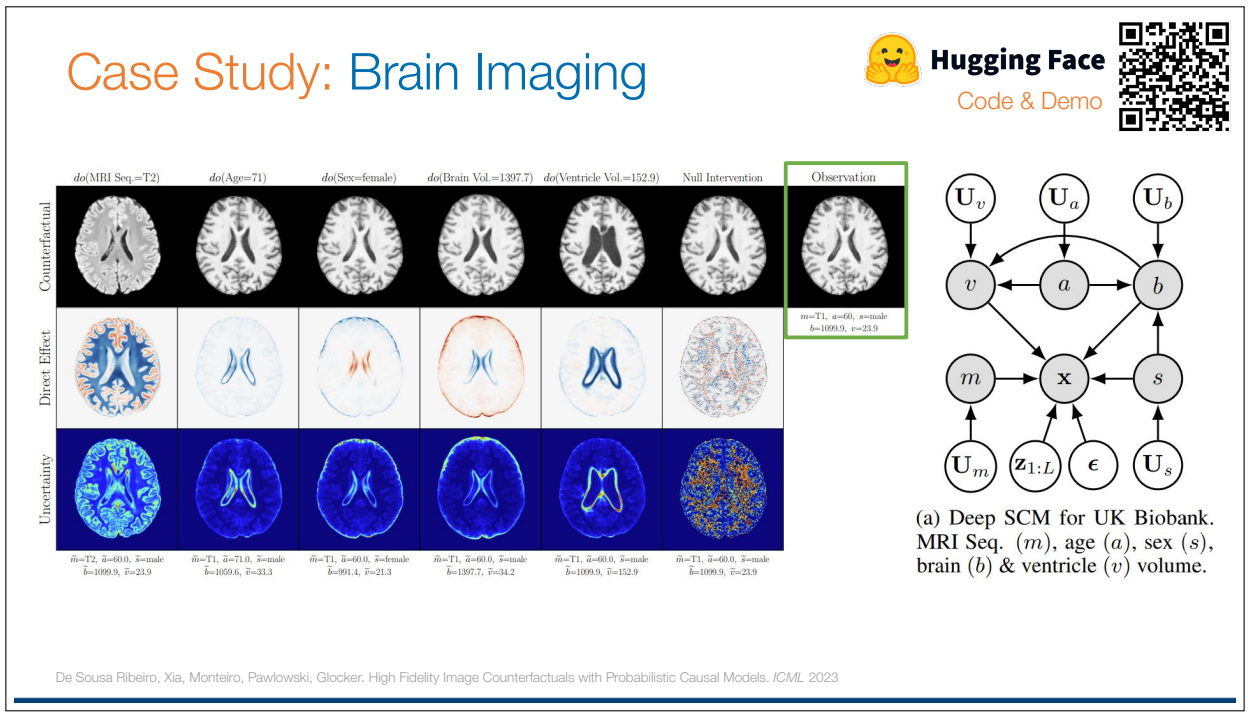


Figure 32: Similar example, we start with a medically assumed causal graph with 6 observed variables. X is one of the variables being the image, we also have ventricle volume, age, brain volume, MRI sqeuencing M, sex of patient S. Each column here is an intervention on a different attribute. If we change nothing we get a reconstruction back out. Idea here is we're not generating different brains, we generating this particular brain with different attributes.

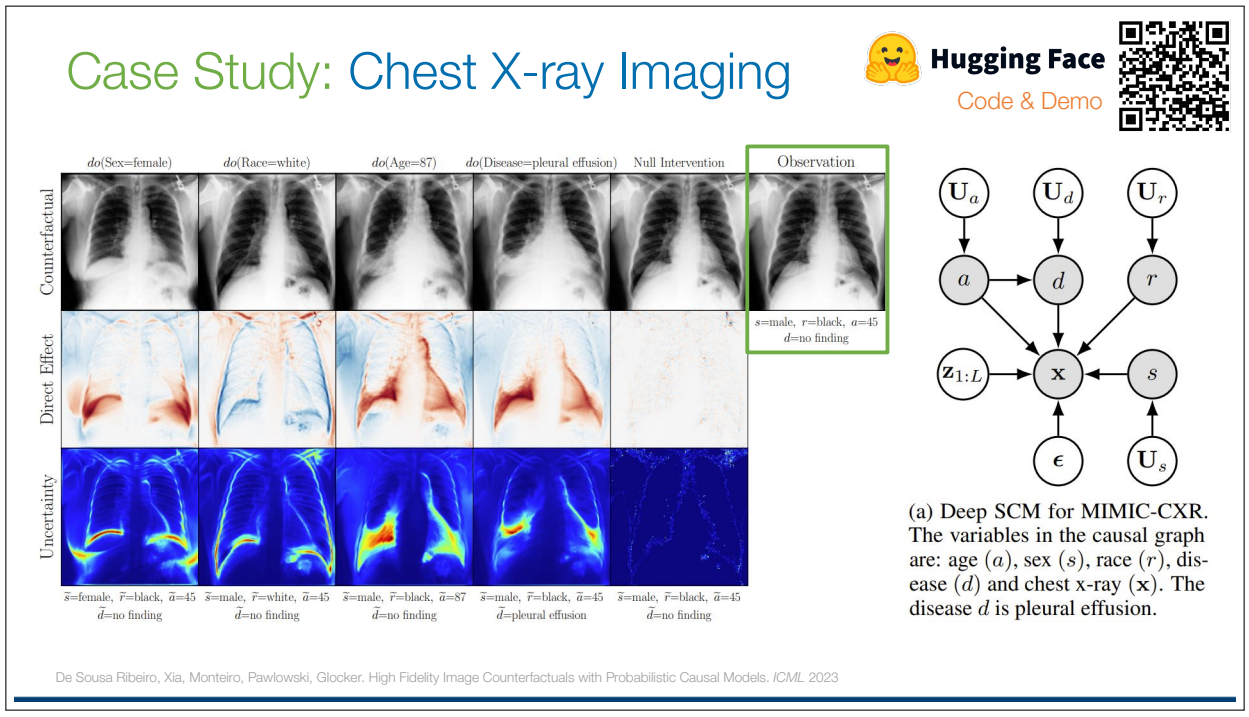


Figure 33: Here we have a different assumed causal graph, but once again it was medically informed.

1.9 Evaluating Counterfactual

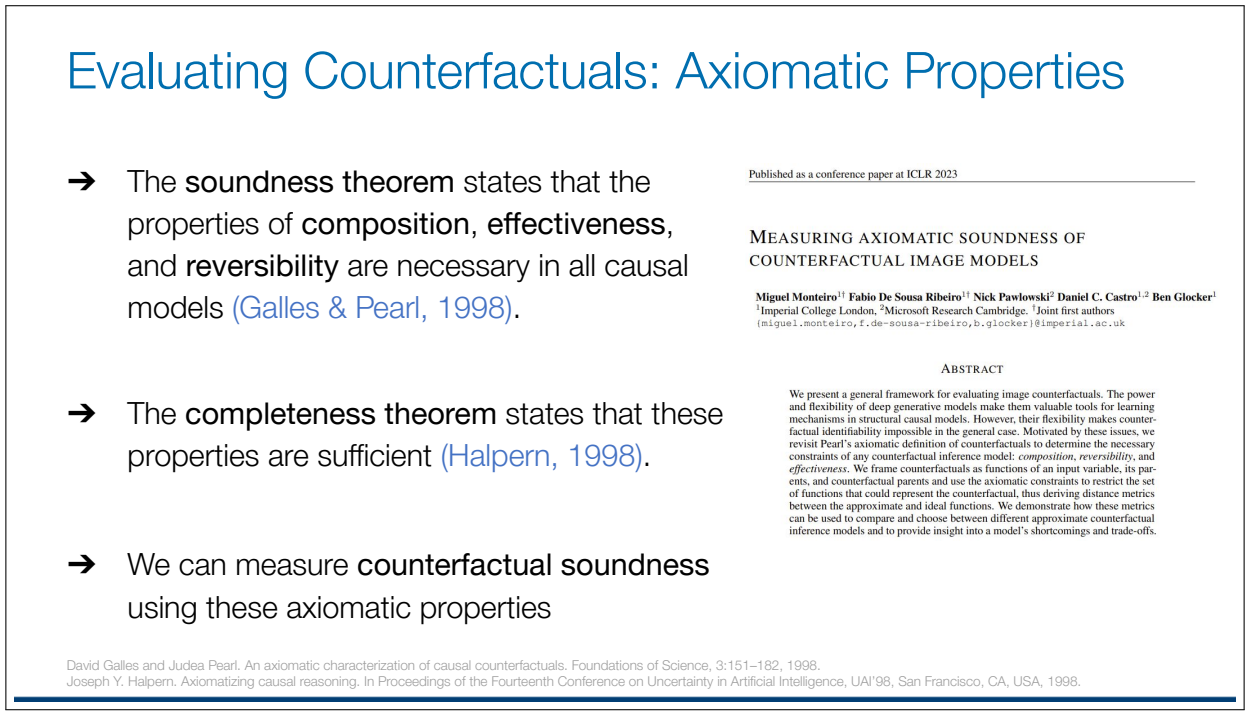


Figure 34: How do we evaluate these counterfactuals? There are a couple theorems above.

1.9.1 Composition Axiom

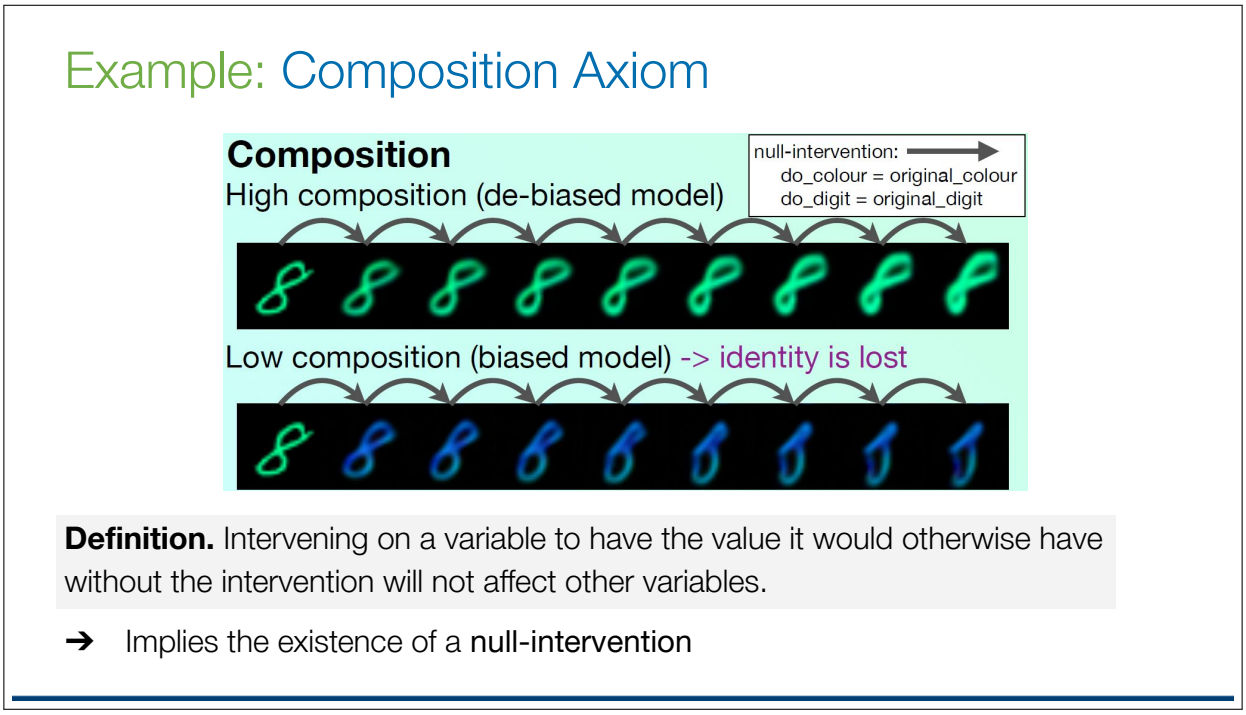


Figure 35: Definition above, we can see an example above with high composition; a de-biased model. If we continuously apply the null intervention iteratively and try to reconstruct the image, if the identities are well preserved across tehse iterated applications mechanisms, then we known that the model has high composition.
Conversely, if it has low composition the identity is lost.

1.9.2 Effectiveness Axiom

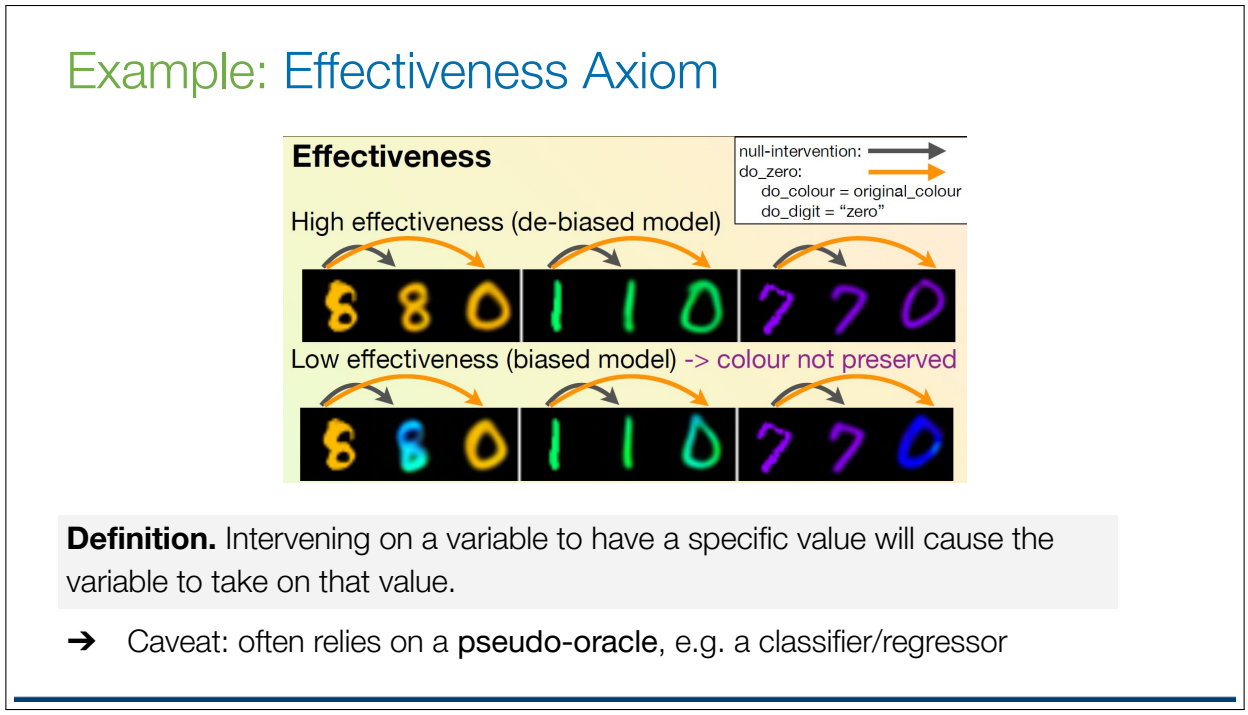


Figure 36: Definition above. An example with high effectiveness, if we want to change digit identitiy it will preserve colour, if we want to change colour then it should preserve digit identity. So a model with low effectiveness will be biased.

1.9.3 Reversibility Axiom

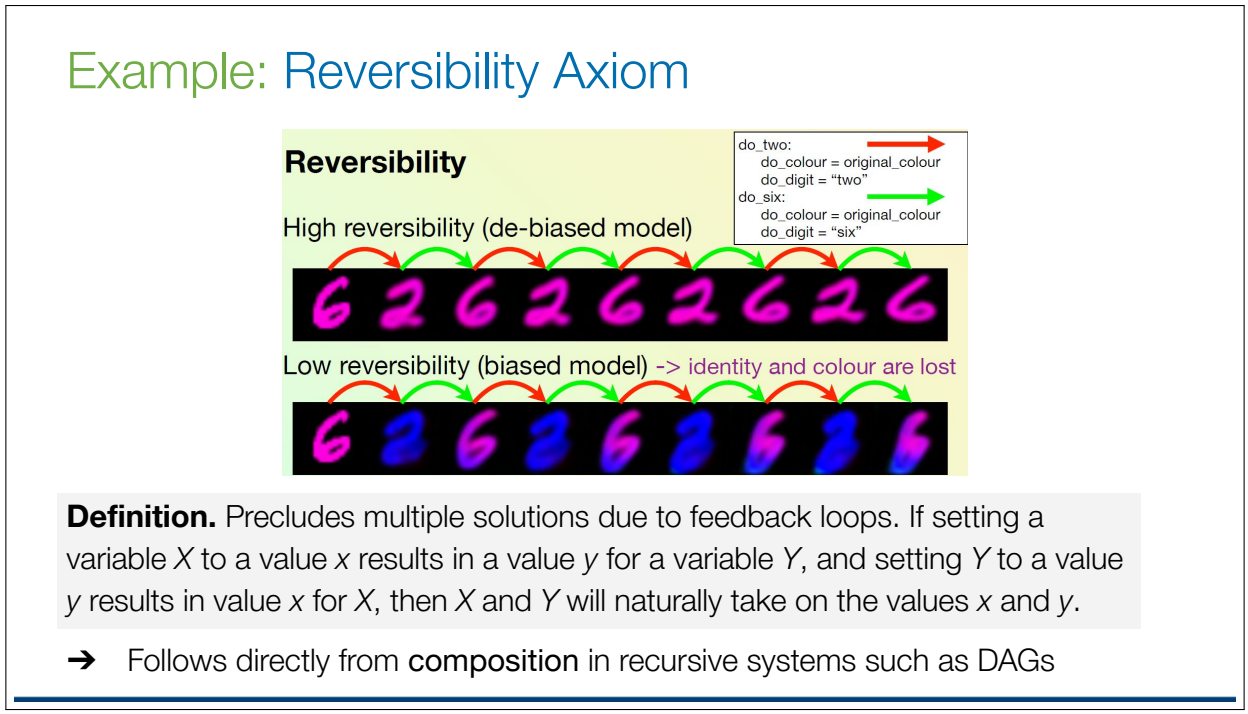


Figure 37: It has to do with cyclic consistency. A model with good reversibility will be able to do this, otherwise it color information will be lost. The reversibility follows from the composition axiom if you use a directed acyclic graph.

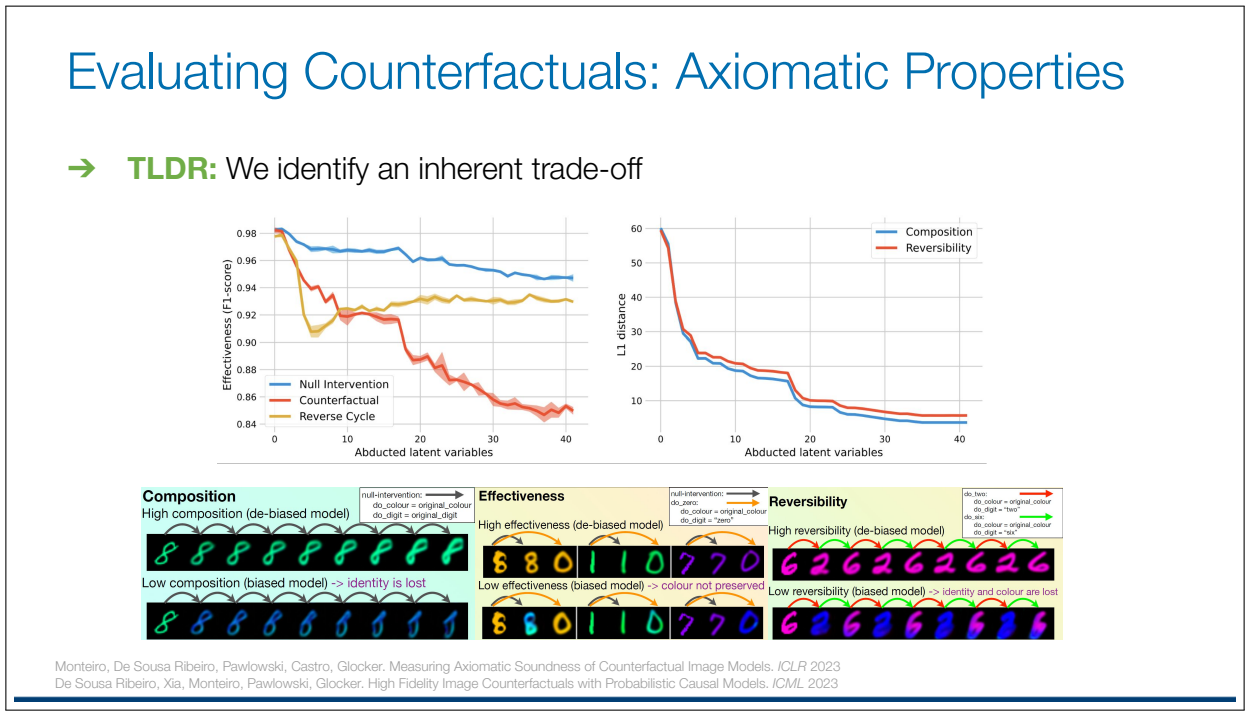


Figure 38: We identified an inherent trade-off between composition and effectiveness – when you train counterfactual inference models you want to be good at all of these things, but there seems to be a trade off between composition and effectiveness specifically.

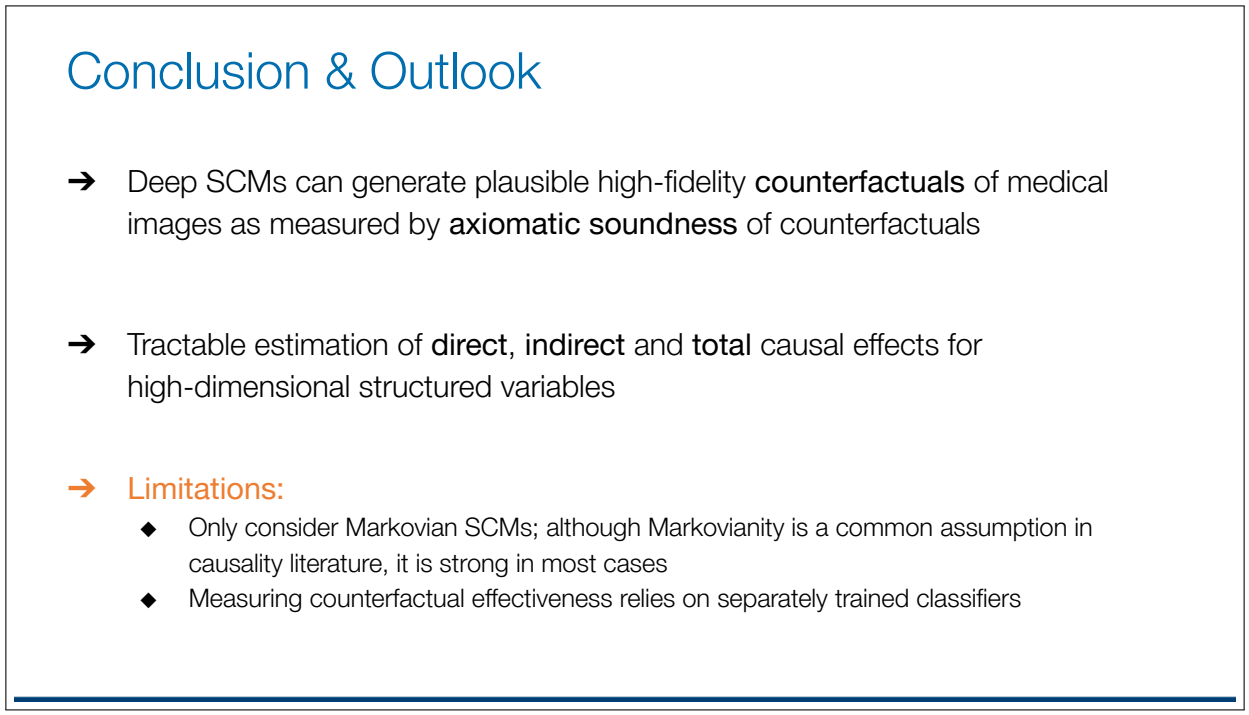


Figure 39: caption

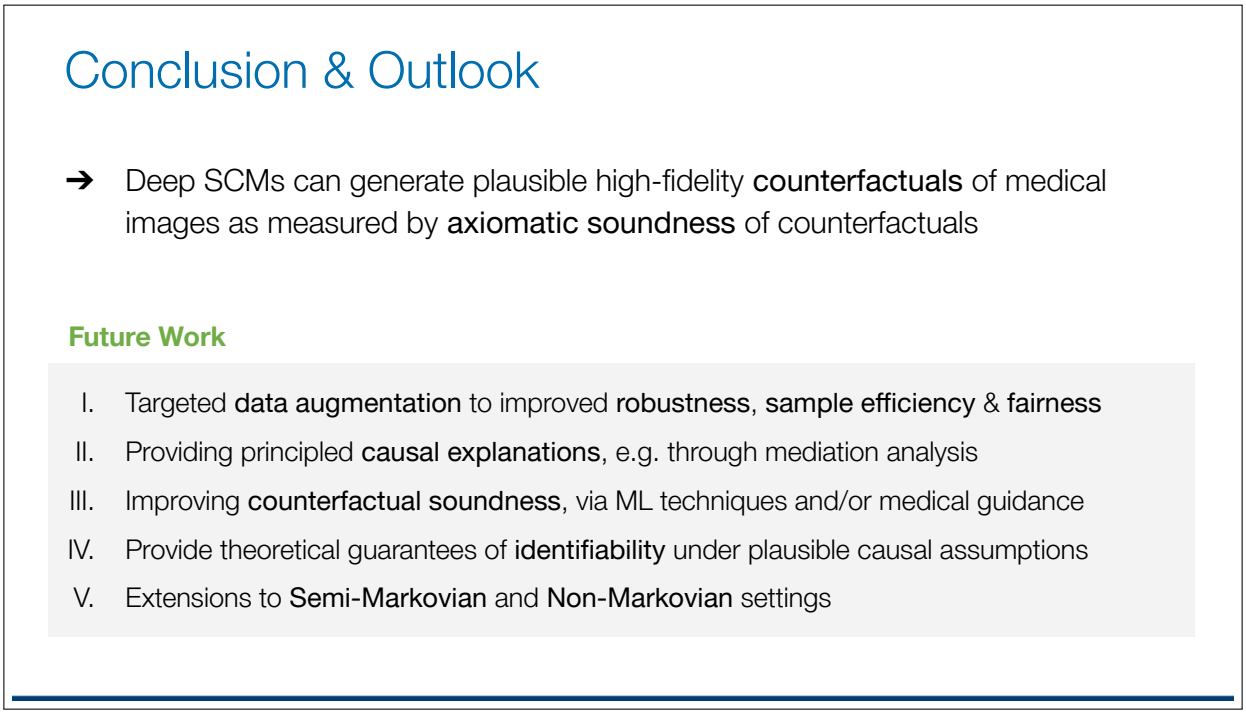


Figure 40: caption

References

[1] Daniel C Castro, Ian Walker, and Ben Glocker. “Causality matters in medical imaging”. en. In: *Nat Commun* 11.1 (July 2020), p. 3673.