TÜBINGEN
AI CENTER

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

A causal perspective on reliable and interpretable representation learning

Michel Besserve

MPI for Intelligent Systems, Tübingen, Germany

AISSAI Workshop Heterogeneous Data and Large Representation Models in Science

Oct. 2nd, 2024

Artificial Intelligence (AI) has challenges ahead

High hopes for AI to transform all economic sectors.

But many technical, strategic and societal issues ahead:

- reliability and ethical issues;
- data and compute requirements;
- pressure on resources and sustainability.


The ugly truth behind ChatGPT: AI is guzzling resources at planet-eating rates
Mariana Mazzucato

[The Guardian, May 30th 2024]

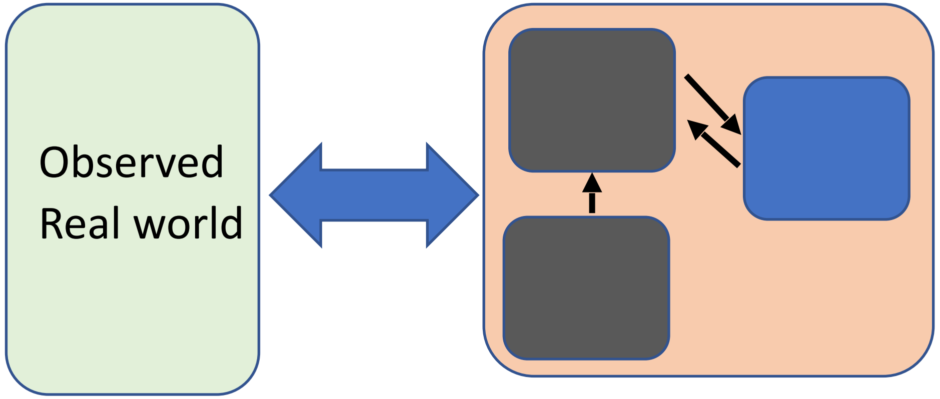
We need general frameworks to address these issues.

Causality allows to:

- phrase rigorously meaningful requirements for AI,
- derive theoretical guaranties for them,
- formulate problems and explanations in a language understandable by humans.

What is causality about?

- Representing knowledge about the world *and how it can be changed*.

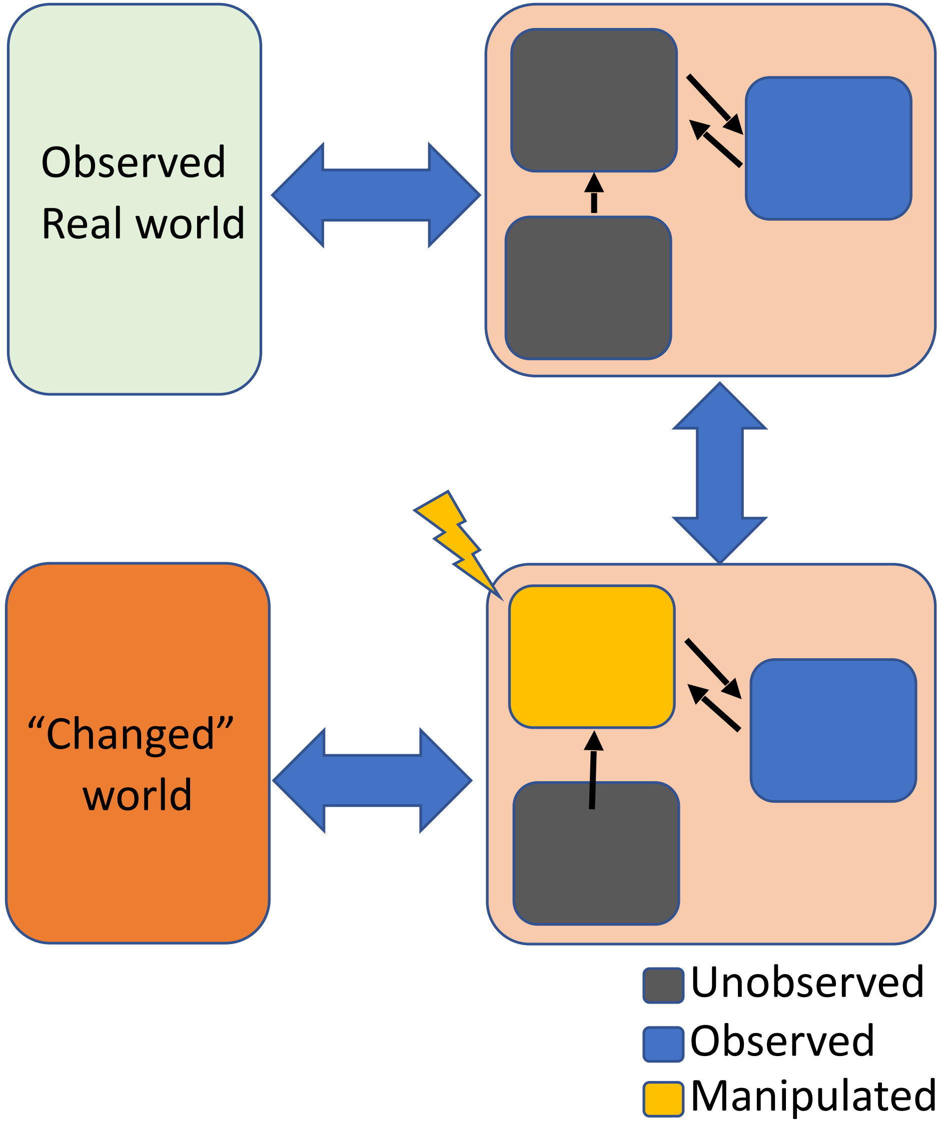


- Specificity : **Modularity** of the system
- Modules are called “mechanisms”

Unobserved
Observed
Manipulated

What is causality about?

- Representing knowledge about the world *and how it can be changed*.



- Specificity : **Modularity** of the system
- Modules are called "mechanisms"

➢Make plausible changes to a mechanism to:

  ➢Explain "why things happen",

  ➢Make predictions.

Structural causal models

Causal graph

Show the influence of variables on each other

Mechanisms = “Structural assignments”

Computes child node value as a function of parents

$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

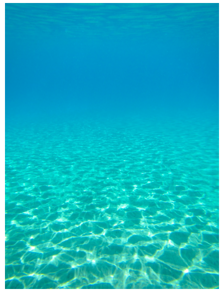
Independent
“Exogenous”
variables

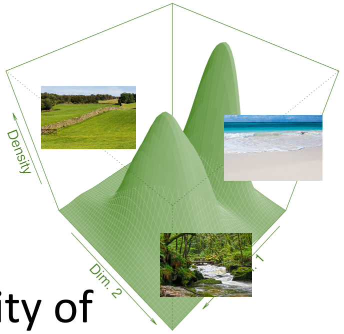
Structural causal models

Causal graph

Show the influence of variables on each other

X_1





Probability of background image

Mechanisms = "Structural assignments"

Computes child node value as a function of parents

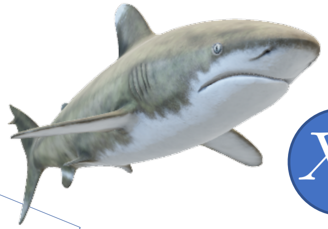
$$\begin{cases} \boxed{X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1)} \\ X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

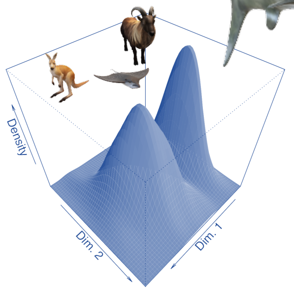
Independent "Exogenous" variables

Structural causal models

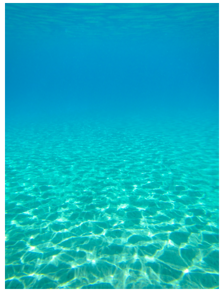
Causal graph

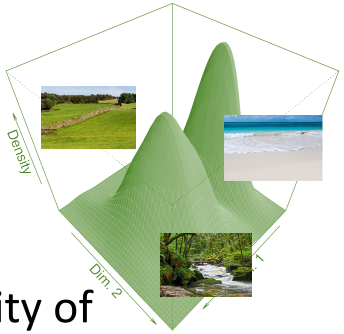
Show the influence of variables on each other





$$X_2 \longleftarrow X_1$$





Probability of background image

Probability of animal **given background**

Mechanisms = "Structural assignments"

Computes child node value as a function of parents

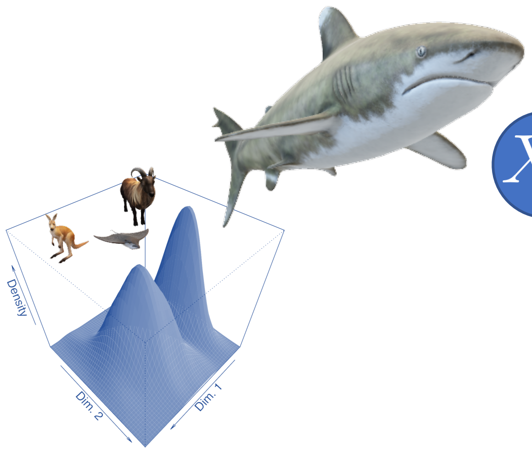
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ \boxed{X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

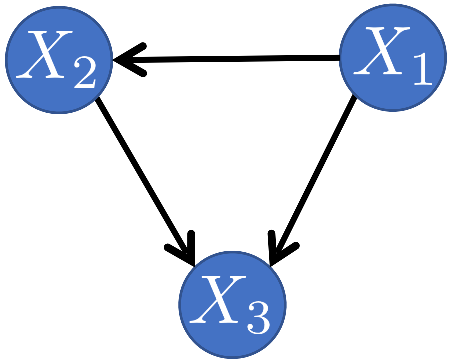
Independent "Exogenous" variables

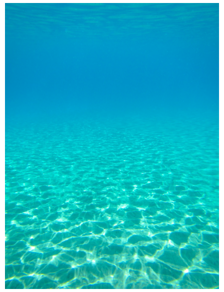
Structural causal models

Causal graph

Show the influence of variables on each other

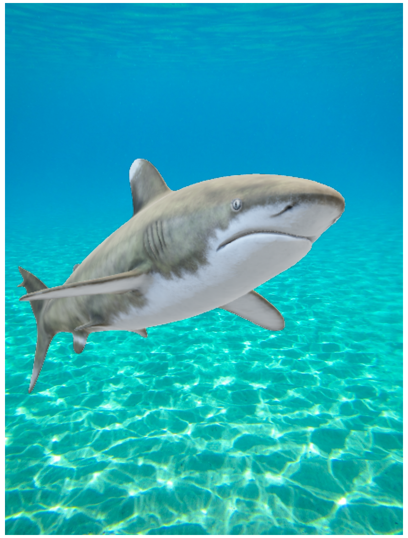






Probability of background image

Probability of animal **given background**



Mechanisms = "Structural assignments"

Computes child node value as a function of parents

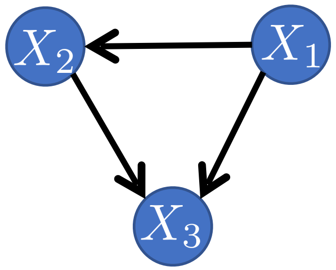
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ \boxed{X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3)} \end{cases}$$

Independent "Exogenous" variables

Samples of object-background combination images.

Interventions and counterfactuals

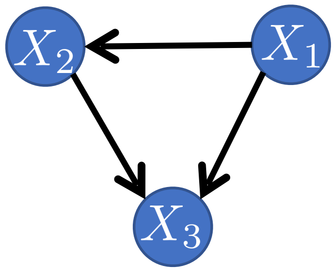
Unintervened model



$$
\begin{cases}
X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\
X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\
X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3)
\end{cases}
$$

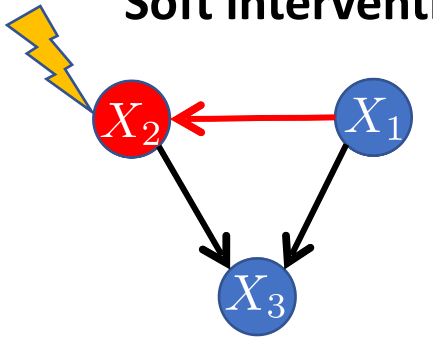
Interventions and counterfactuals

Unintervened model



$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ \boxed{X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

Soft intervention: « *only sharks when in the sea* »



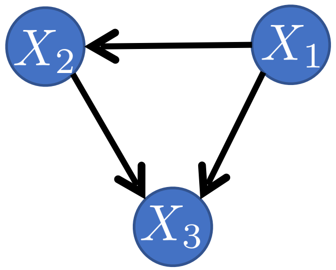
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ \boxed{X_2 := h_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$





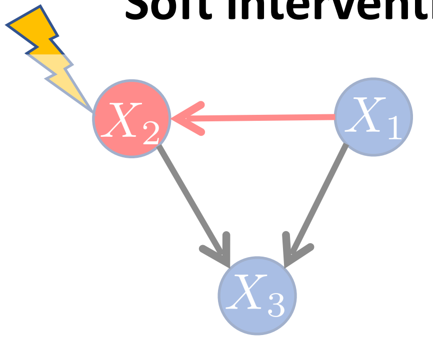
Interventions and counterfactuals

Unintervened model



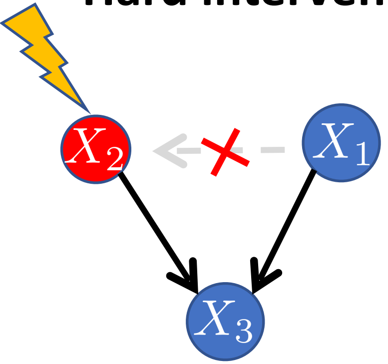
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ \boxed{X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

Soft intervention: « *only sharks when in the sea* »



$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := h_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

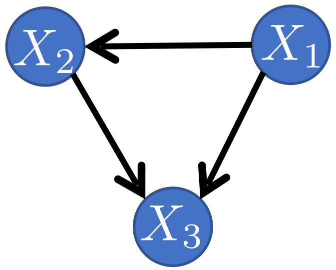
Hard intervention: « *only Sharks, everywhere* »



$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ \boxed{X_2 := g_{2,\boldsymbol{\theta_2}}(\epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

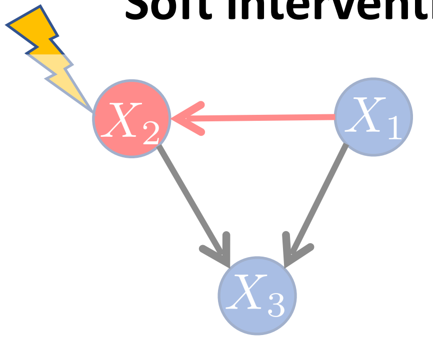
Interventions and counterfactuals

Unintervened model

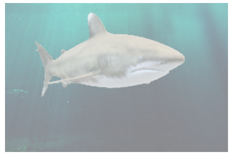


$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\boxed{\epsilon_1}) \\ \boxed{X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \boxed{\epsilon_3}) \end{cases}$$

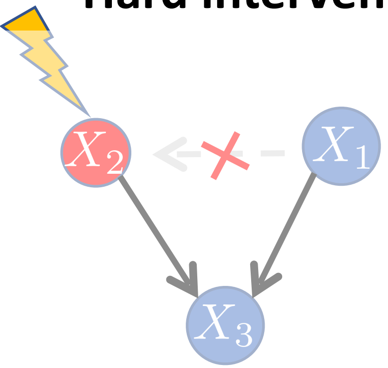
Soft intervention: « *only sharks when in the sea* »



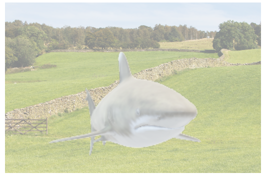
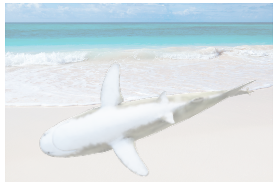
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := h_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

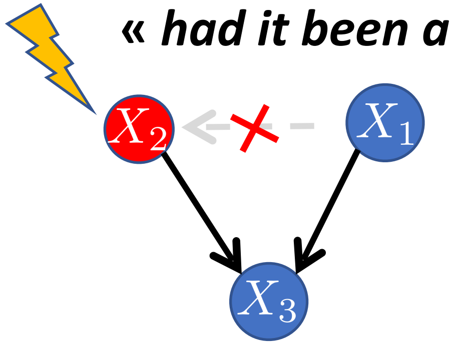

Hard intervention: « *only Sharks, everywhere* »



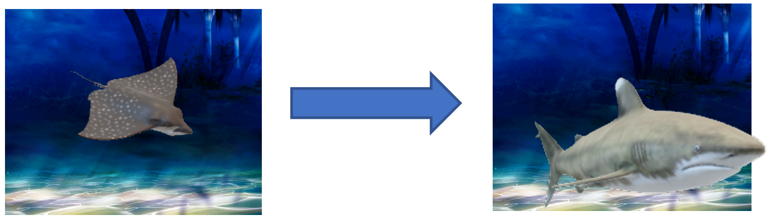
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := g_{2,\boldsymbol{\theta_2}}(\epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$


Counterfactuals:
« *had it been a shark instead of a stingray* »



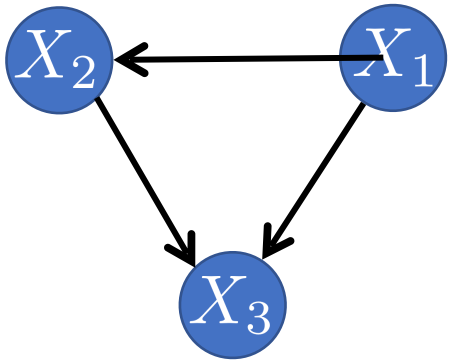
$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\boxed{\epsilon_1}) \\ \boxed{X_2 := g_{2,\boldsymbol{\theta_2}}(\epsilon_2)} \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \boxed{\epsilon_3}) \end{cases}$$



Causal inference

Causal graph

Show the influence of variables on each other



Mechanisms = "Structural equations"

Compute child node value as a function of parents

$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

Independent "Exogenous" variables

Three types of inference tasks based on data:

Causal discovery: "finding the arrows"

Causal effect estimation: "how strong is causation?"

Causal representation learning: "finding the variables"

[= building a **mapping** from the observations to causal variables]

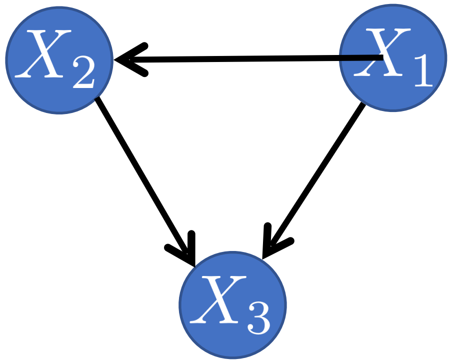
Three types of dataset:

Observational, interventional, counterfactual.

Causal inference

Causal graph

Show the influence of variables on each other



Mechanisms = "Structural equations"

Compute child node value as a function of parents

$$\begin{cases} X_1 := f_{1,\boldsymbol{\theta_1}}(\epsilon_1) \\ X_2 := f_{2,\boldsymbol{\theta_2}}(X_1, \epsilon_2) \\ X_3 := f_{3,\boldsymbol{\theta_3}}(X_1, X_2, \epsilon_3) \end{cases}$$

Independent "Exogenous" variables

Three types of inference tasks based on data:

Causal discovery: "finding the arrows"

Causal effect estimation: "how strong is causation?"

Causal representation learning: "finding the variables"

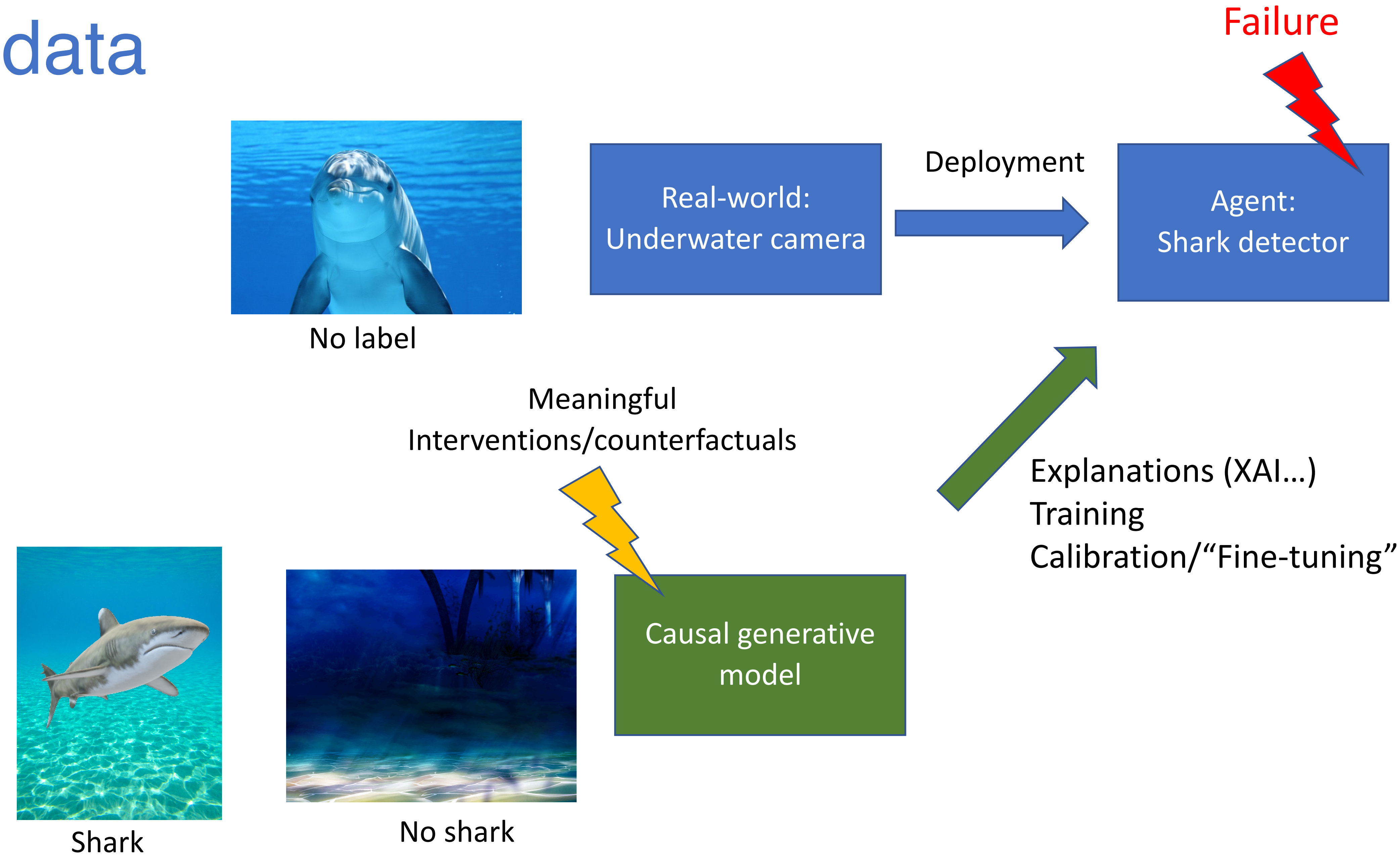
[= building a **mapping** from the observations to causal variables]

←This presentation

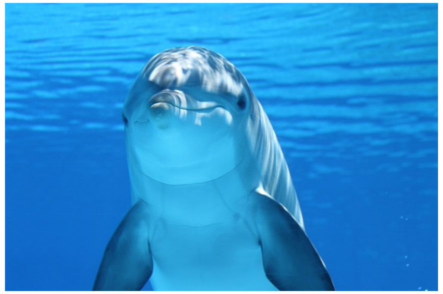
Three types of dataset:

Observational, interventional, counterfactual.

Benefits of a causal model of the data



Benefits of a causal model of the data



No label

Real-world: Underwater camera

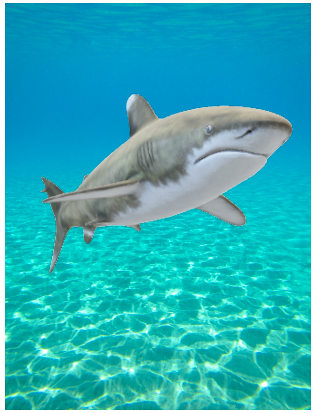
Deployment

Agent: Shark detector

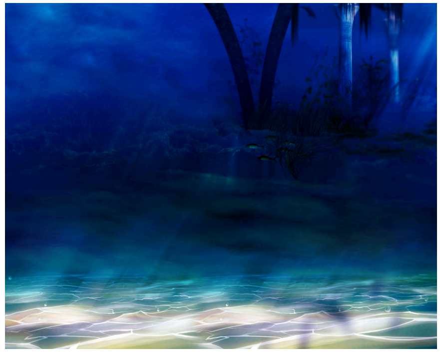
Failure

Meaningful Interventions/counterfactuals

Explanations (XAI…)
Training
Calibration/“Fine-tuning”



Shark



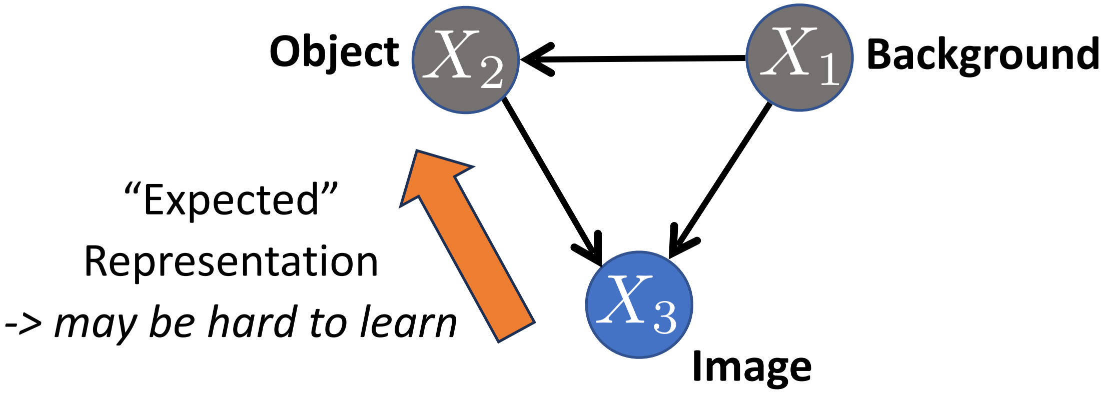
No shark

Causal generative model

An interactive tool to foster **reliable** and **interpretable** technologies and decisions.

Illustration: shortcut features

Features we are not « interested in » might be used to improve classification...



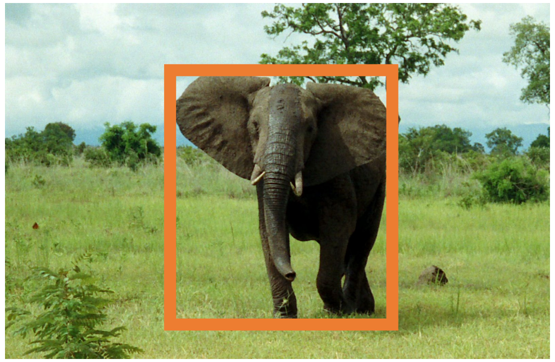
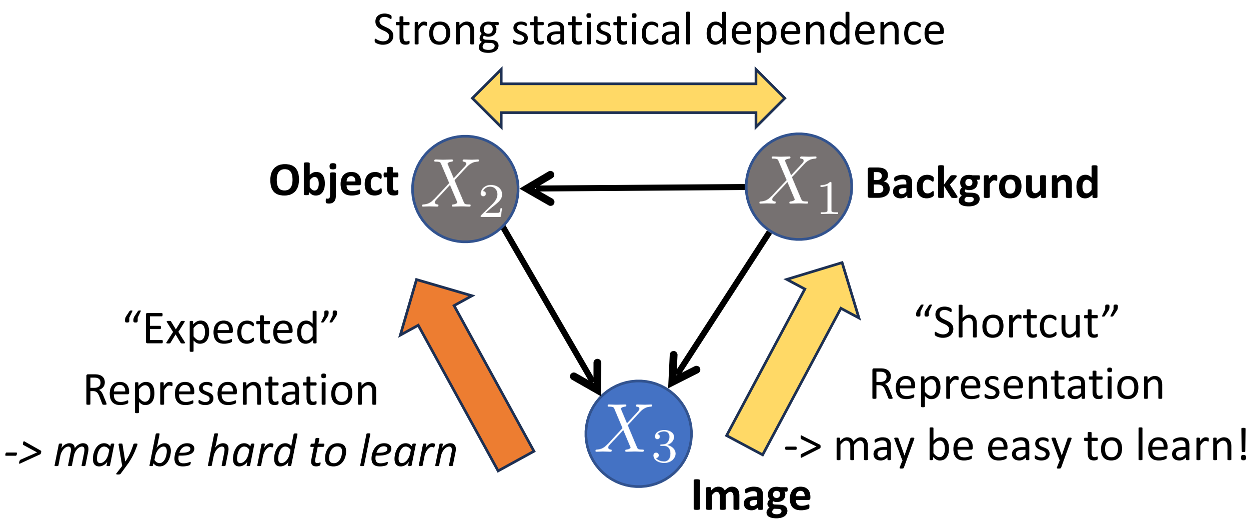
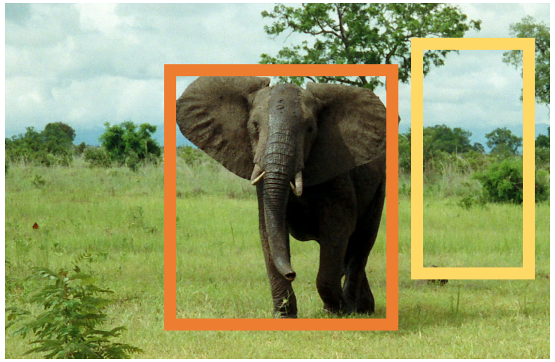


Illustration: shortcut features

Features we are not « interested in » might be used to improve classification...

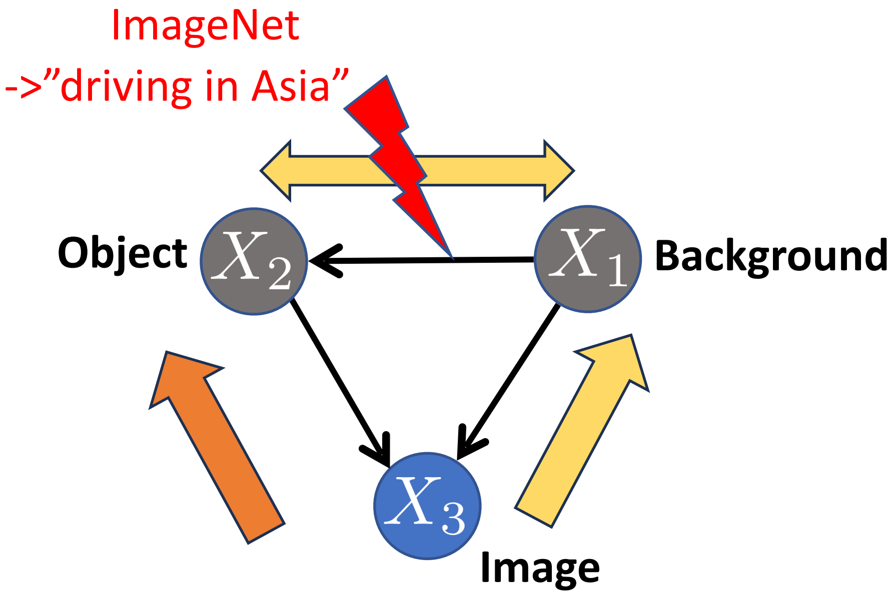




- Shortcut features are found because they are highly predictive on the training data.
- They are **not an issue** if the deployment data has the same distribution as the training data **-> "same environment"**
- Actually, dropping all such features would typically **decrease the performance**!
- But this makes the AI system not robust to changes of the environment…

Illustration: shortcut features

We can model the changes of environment with interventions.



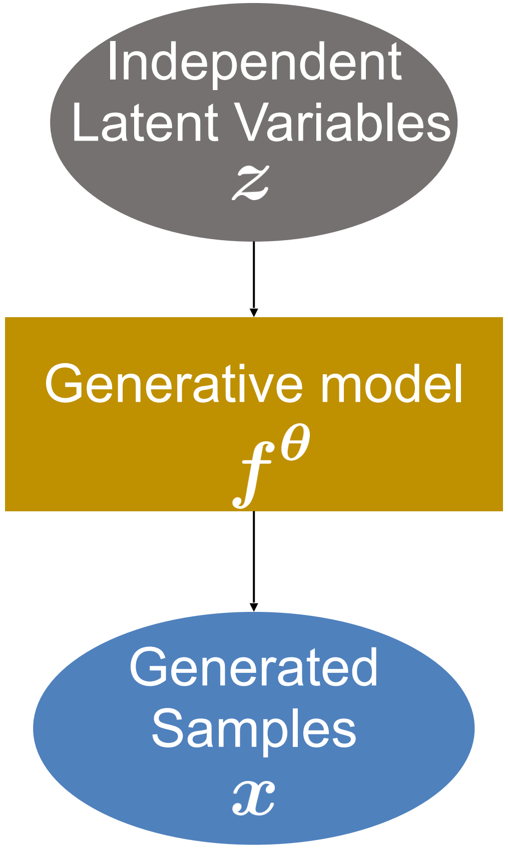
- The intervention affects the shortcut pathway by changing the learnt statistical dependency between object class and some backround features.
- Avoiding such settings requires scrutinizing the data generating process, in particular dataset collection! -> Hard for most complex real-world data.
- Typical dangerous cases: data collected with different measurement systems…
  …e.g.: replacing a particle detector.
- Automatically learning a good approximation of the causal data generating process could help.

Intervening in Deep generative models

[Besserve et al., ICLR 2020]

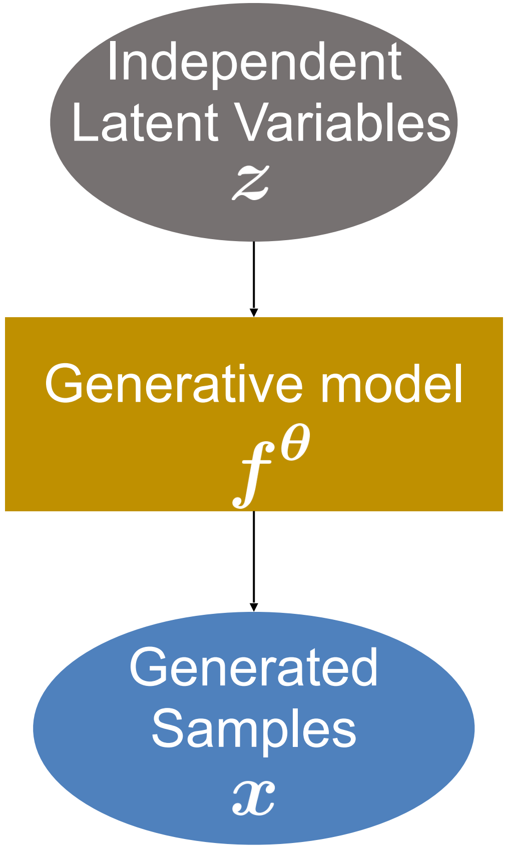
Deep generative models (unsupervised learning)

Architecture



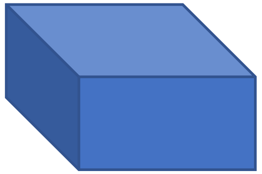
Deep generative models (unsupervised learning)

Architecture



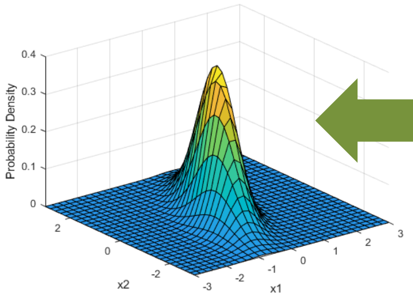
Joint density

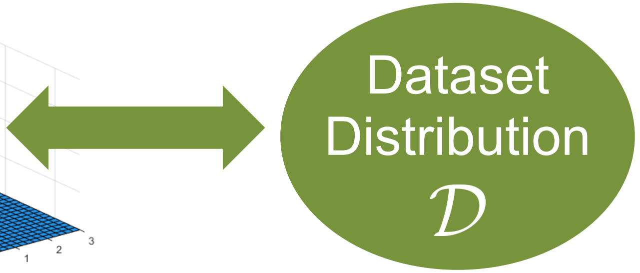
$$p_0(\boldsymbol{z}) = \prod p_{0,k}(z_k)$$



Pushforward density

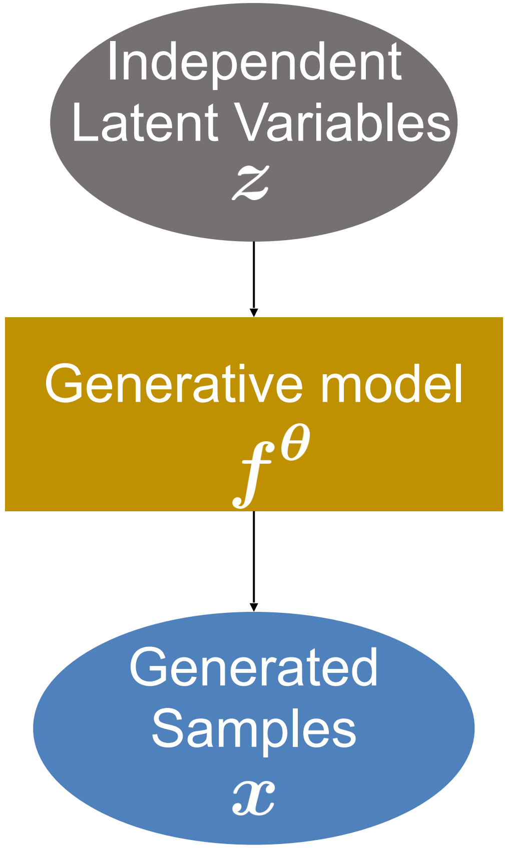
$$p(\boldsymbol{x}) = \overbrace{\boldsymbol{f}_*[p_0(\boldsymbol{z})]} \sim \mathcal{D}$$





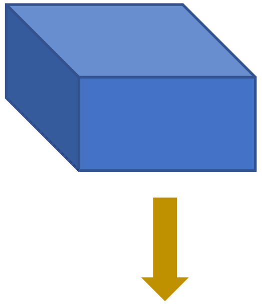
Deep generative models (unsupervised learning)

Architecture



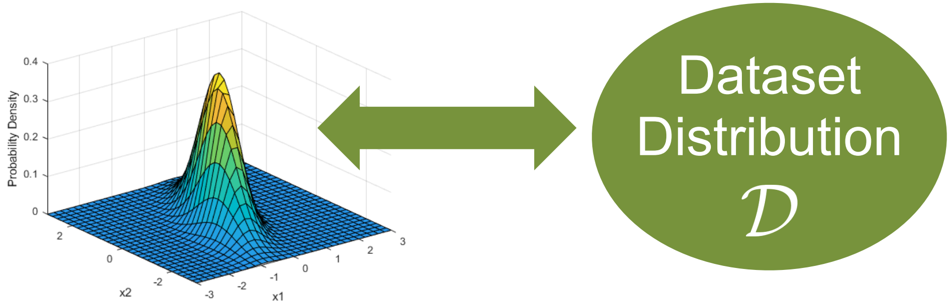
Joint density

$$p_0(\boldsymbol{z}) = \prod p_{0,k}(z_k)$$



Pushforward density

$$p(\boldsymbol{x}) = \overbrace{\boldsymbol{f}_*[p_0(\boldsymbol{z})]} \sim \mathcal{D}$$

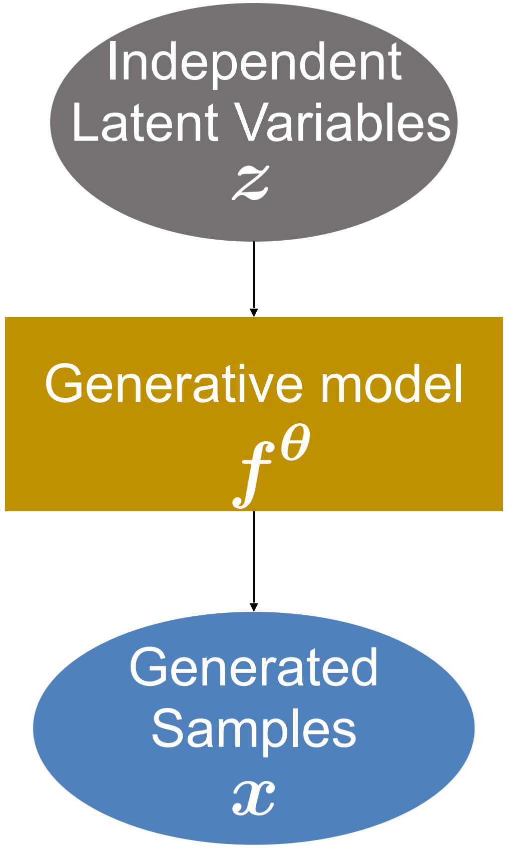


- Many approaches:

[Brock,et al. ICLR 2019]

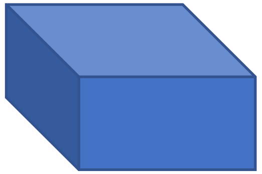
Deep generative models (unsupervised learning)

Architecture



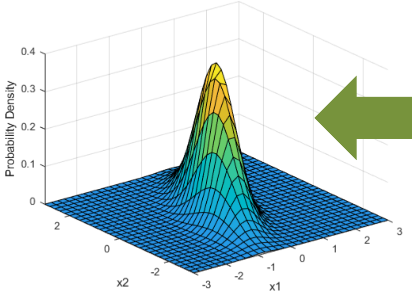
Joint density

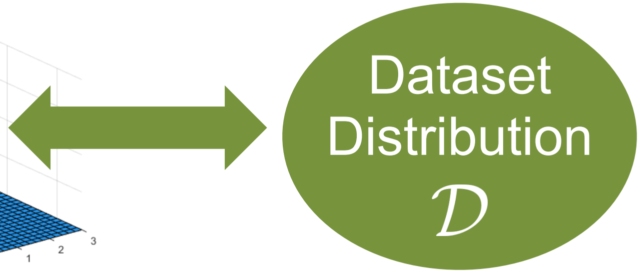
$$p_0(\boldsymbol{z}) = \prod p_{0,k}(z_k)$$



Pushforward density

$$p(\boldsymbol{x}) = \overbrace{\boldsymbol{f}_*[p_0(\boldsymbol{z})]} \sim \mathcal{D}$$





- Many approaches:

VAE, GAN, Stable Diffusion, ….

[Brock,et al. ICLR 2019]

Deep generative models (unsupervised learning)

Architecture

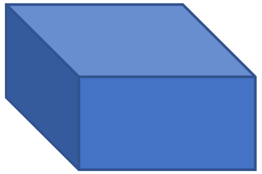
Independent Latent Variables $\boldsymbol{z}$

Generative model $\boldsymbol{f}^{\boldsymbol{\theta}}$

Generated Samples $\boldsymbol{x}$

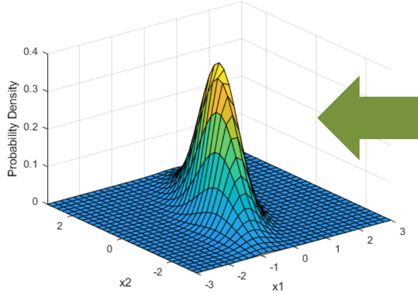
Joint density

$$p_0(\boldsymbol{z}) = \prod p_{0,k}(z_k)$$



Pushforward density

$$p(\boldsymbol{x}) = \overbrace{\boldsymbol{f}_*[p_0(\boldsymbol{z})]}^{} \sim \mathcal{D}$$



Dataset Distribution $\mathcal{D}$

- Many approaches:

VAE, GAN, Stable Diffusion, ….

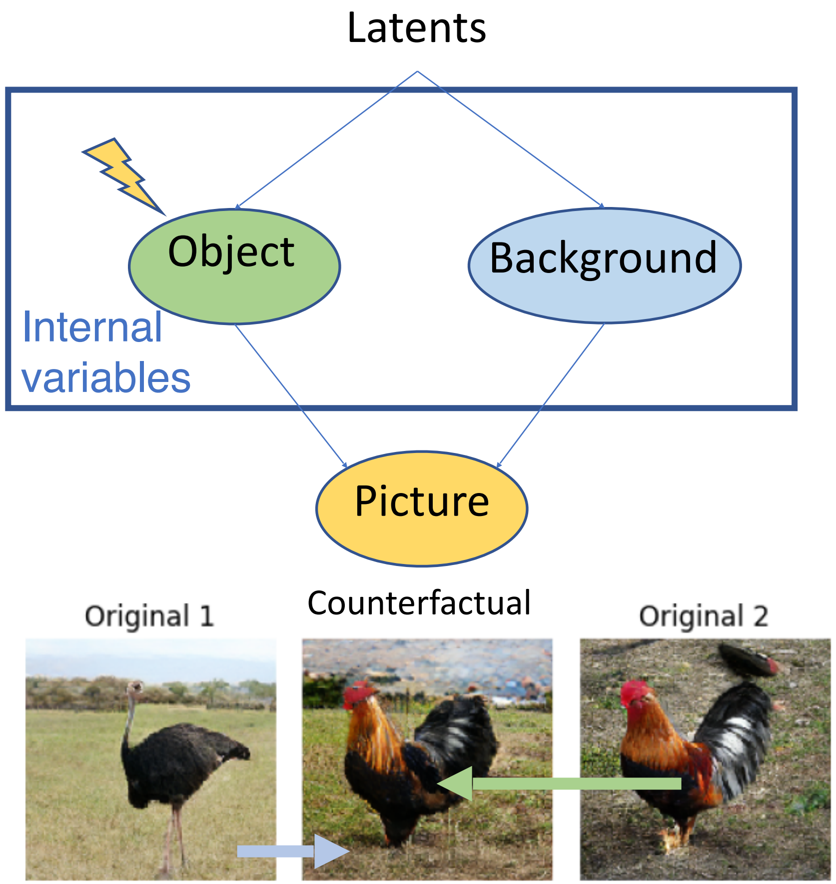
- Can generate highly realistic samples

[Brock,et al. ICLR 2019]

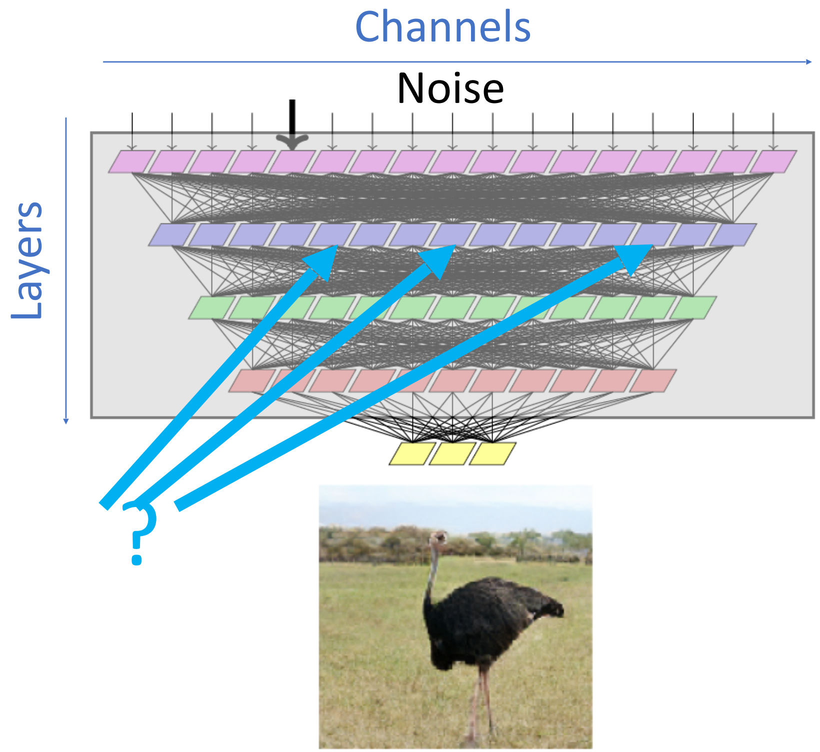
ICM for generating counterfactuals

- Generative models have independent latents but dependent properties
- Look for internal variables with proper causal interpretation!

Disentangled/modular model



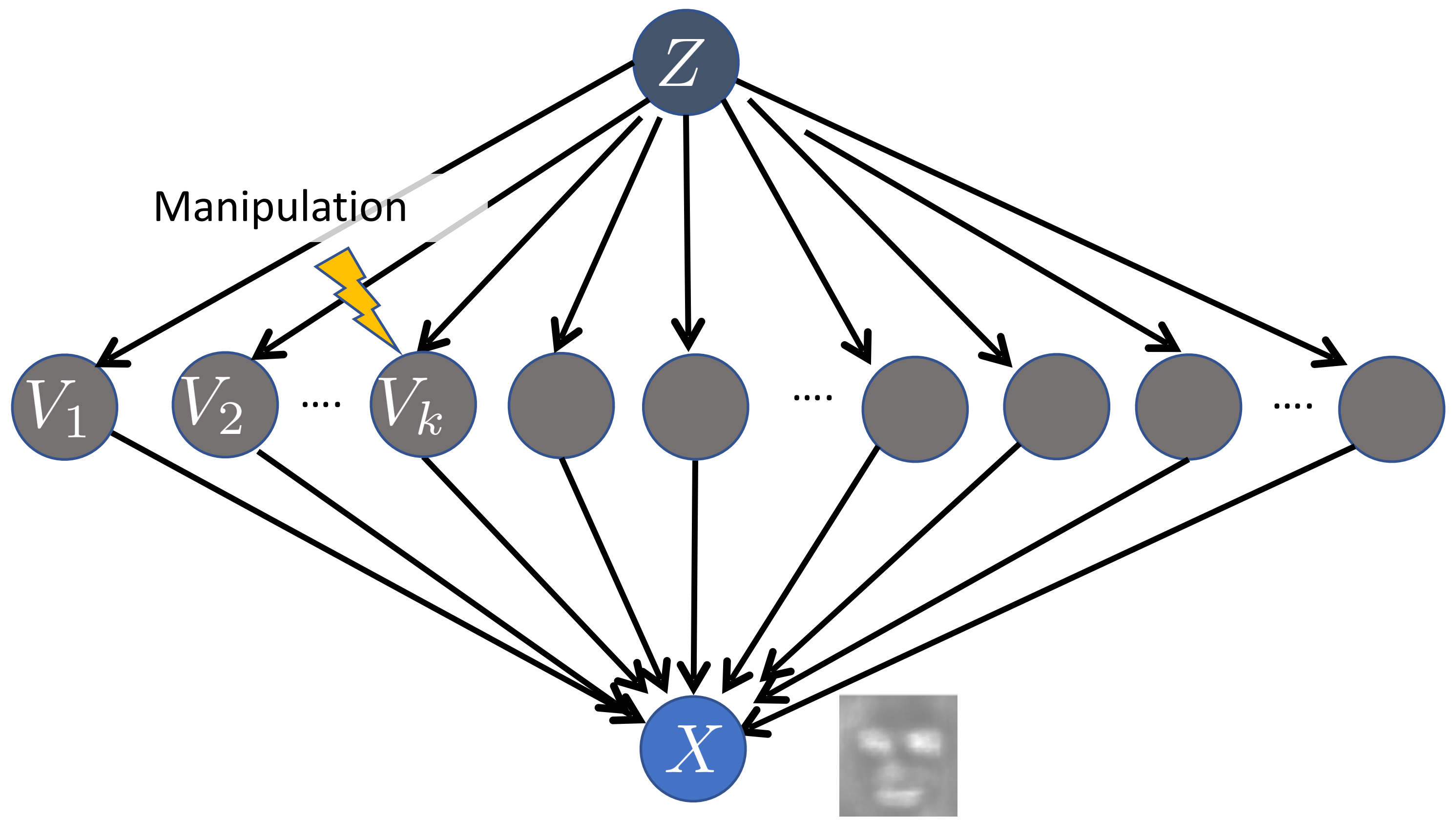
Deep generative model



An approximate causal abstraction

We build a form of causal abstraction [Rubenstein et al, 2017, Beckers et al., 2020], where we group hidden layer neurons into macro-variables on which we want to intervene at once. Groups are meant to encode high level properties of the images.

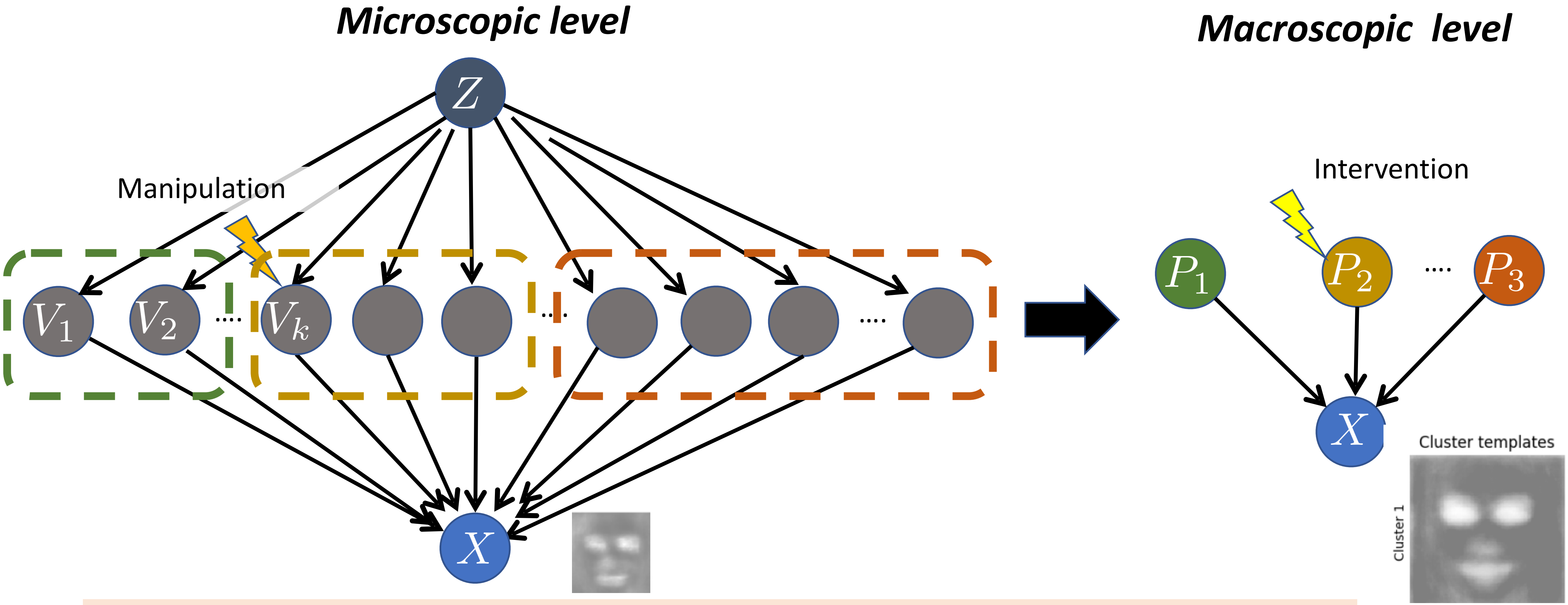
Microscopic level



Idea: group neurons that have similar effect on the generated images.

An approximate causal abstraction

We build a form of causal abstraction [Rubenstein et al, 2017, Beckers et al., 2020], where we group hidden layer neurons into macro-variables on which we want to intervene at once. Groups are meant to encode high level properties of the images.

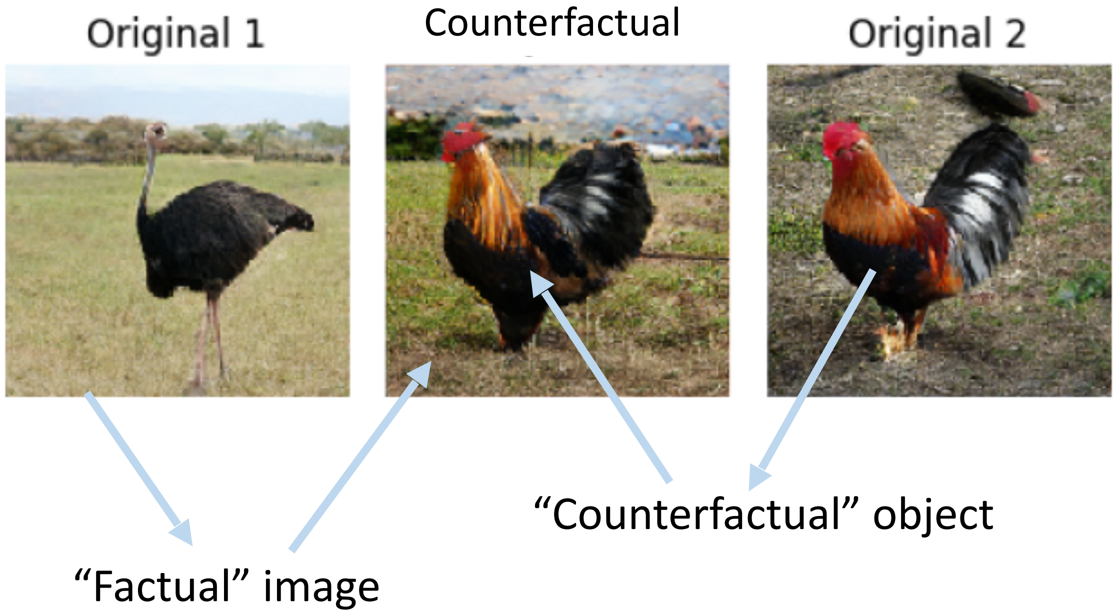


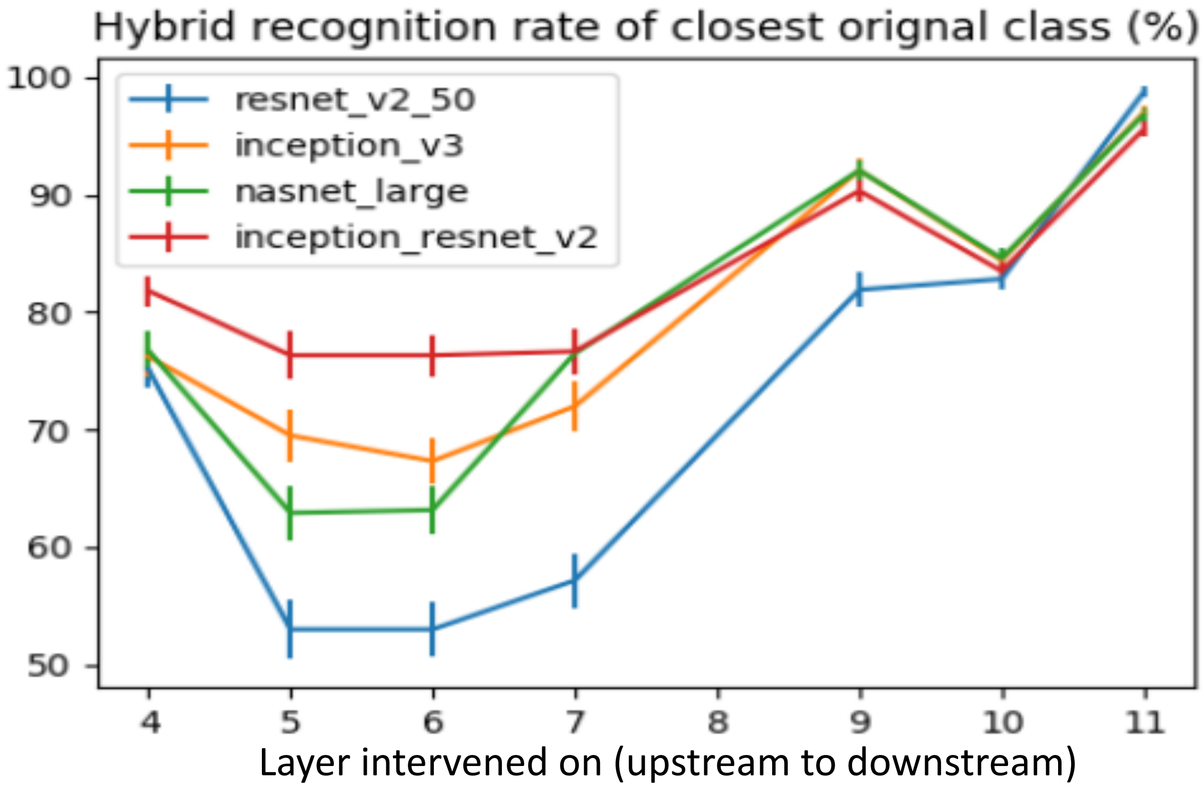
Idea: group neurons that have similar effect on the generated images.

Counterfactual generation

We use counterfactuals to probe state of the art classifiers.

Examples for BigGAN (Brock et al., 2018) on ImageNet



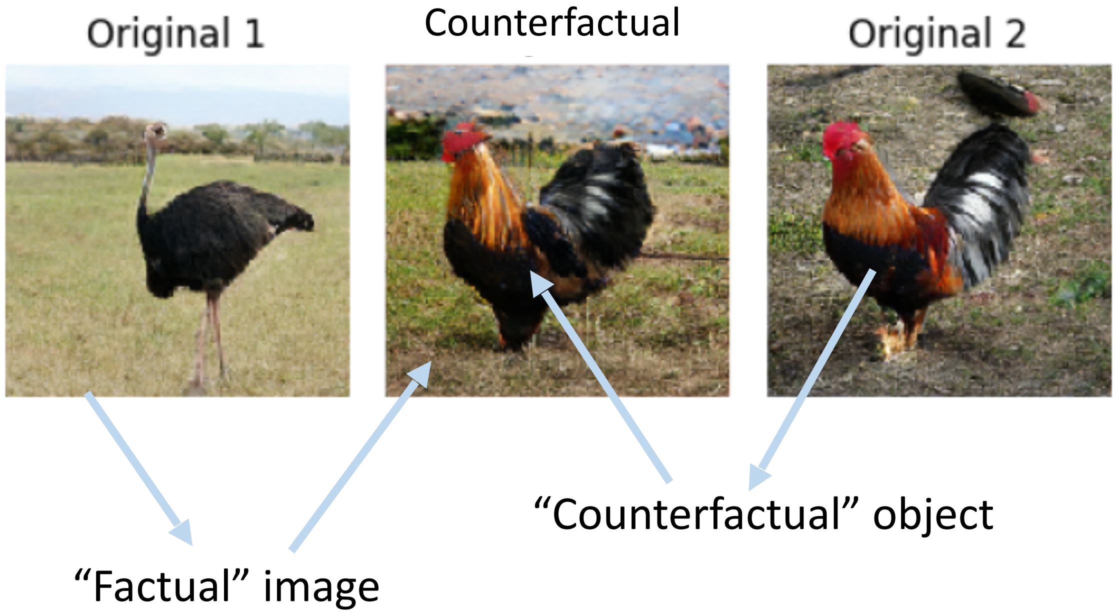


- Several early layers allow object-background separation,
- Other separate shape-texture

[Besserve et al., ICLR 2020]

Counterfactual generation

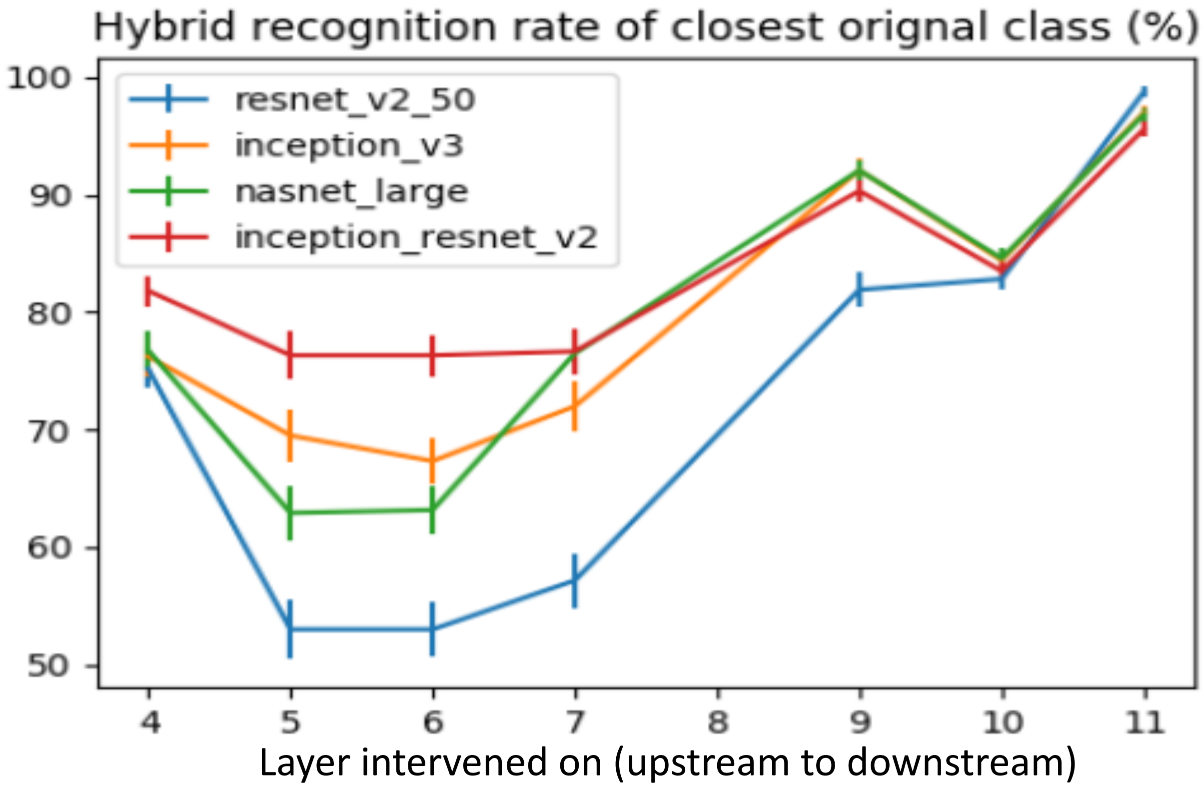
Examples for BigGAN (Brock et al., 2018) on ImageNet

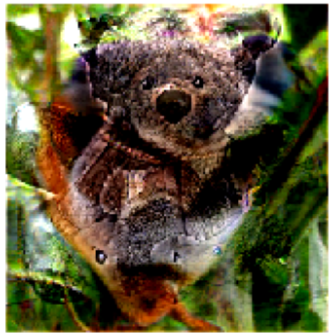


- Several early layers allow object-background separation,
- Other separate shape-texture

[Besserve et al., ICLR 2020]

We use counterfactuals to probe state of the art classifiers.

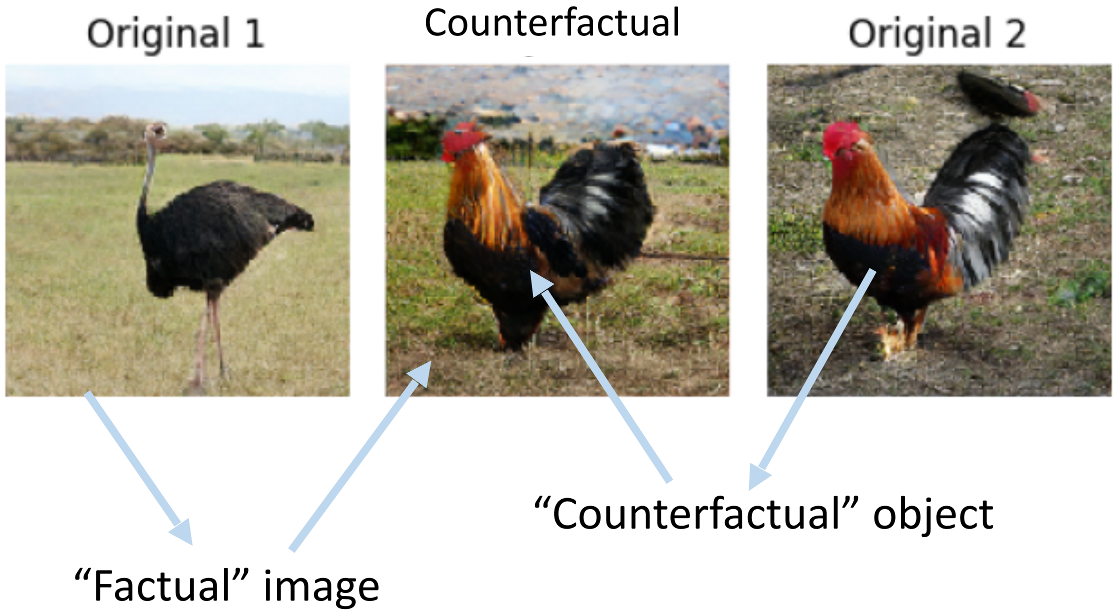




Resnet_v2_50	**Inception_v3**
Koala	Koala
Nasnet_large	**Inception_resnet_v2**
Teddy	Teddy

Counterfactual generation

Examples for BigGAN
(Brock et al., 2018) on ImageNet



- Several early layers allow object-background separation,
- Other separate shape-texture

[Besserve et al., ICLR 2020]

We use counterfactuals to probe state of the art classifiers.

Identifiability of causal representations

[Gresele*, von Kügelgen* et al, NeurIPS 2021]

[Buchholz et al., NeurIPS 2022]

[Reizinger*, Gresele*, Brady* et al. NeurIPS2022]

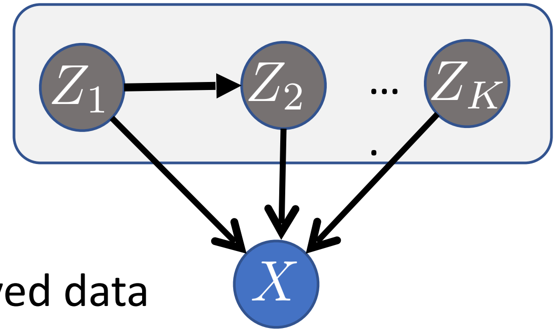
Unsupervised causal representation learning

Assumption: Observed data is parameterized by hidden causal variables

$$X := f(Z_1, Z_2, \ldots, Z_K) \Rightarrow (Z_1, Z_2, \ldots, Z_K) = f^{-1}(X)$$

Ground truth
Causal representation

Hidden



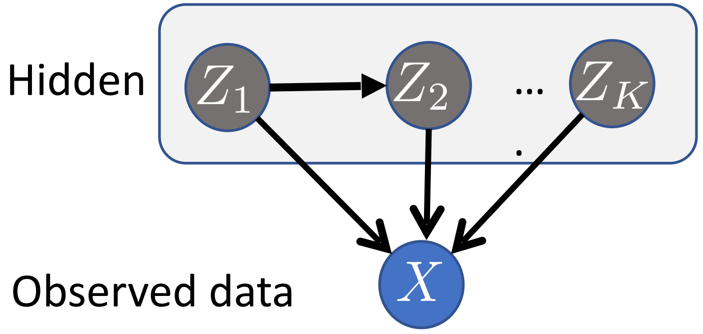
Observed data

Unsupervised causal representation learning

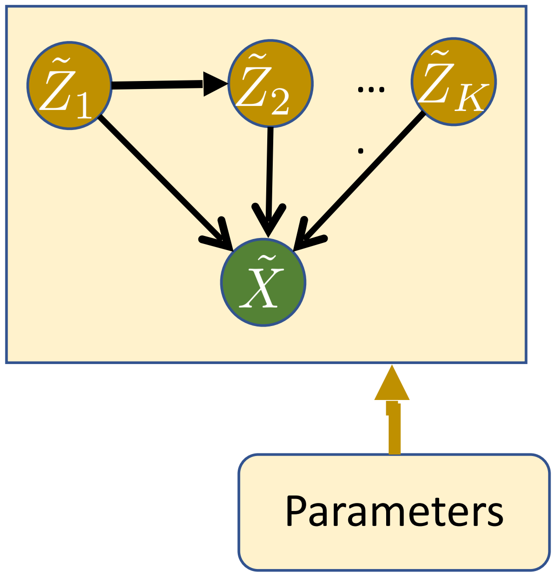
Assumption: Observed data is parameterized by hidden causal variables

$$X := f(Z_1, Z_2, \ldots, Z_K) \Rightarrow (Z_1, Z_2, \ldots, Z_K) = f^{-1}(X)$$

Ground truth
Causal representation



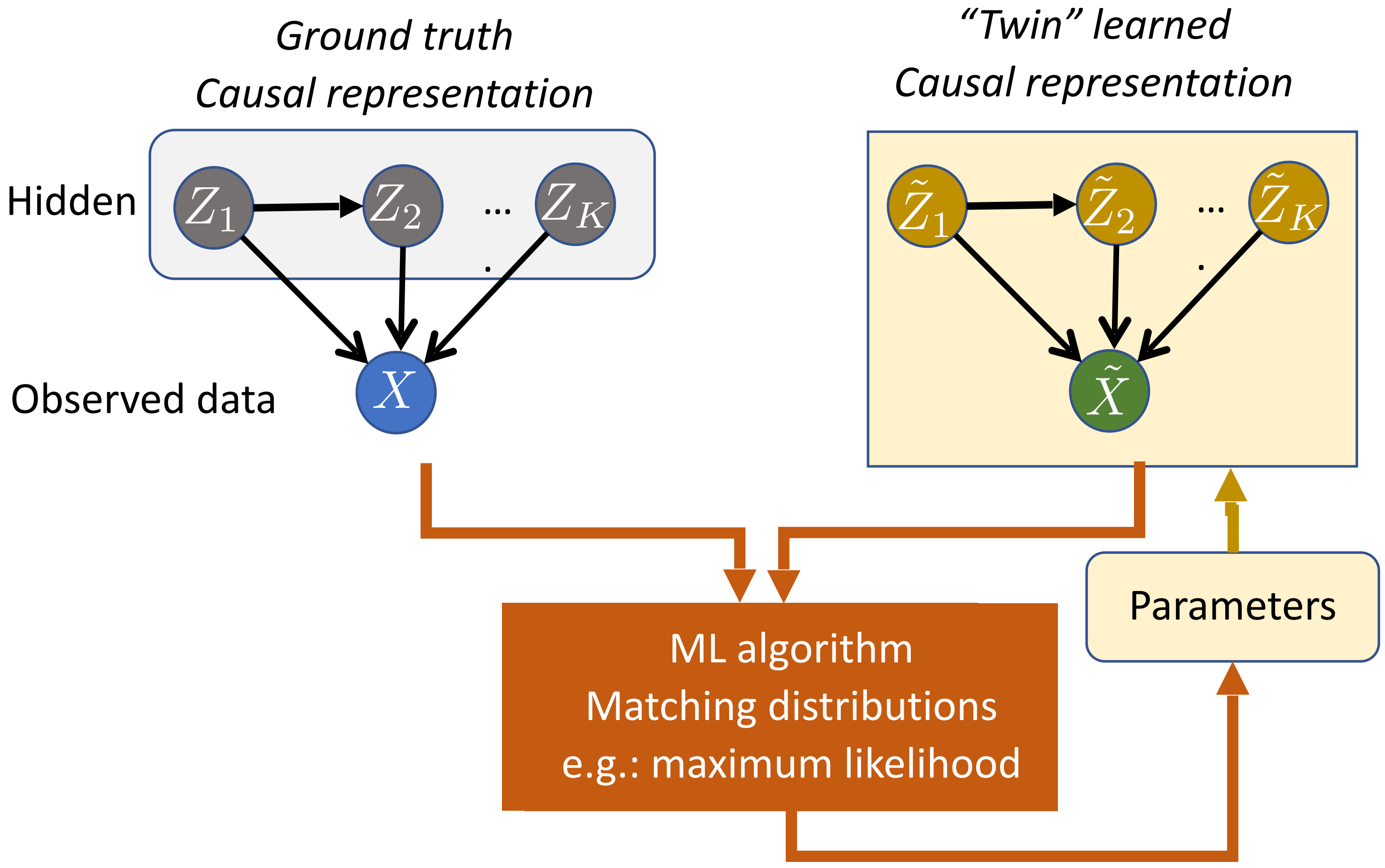
"Twin" learned
Causal representation



Unsupervised causal representation learning

Assumption: Observed data is parameterized by hidden causal variables

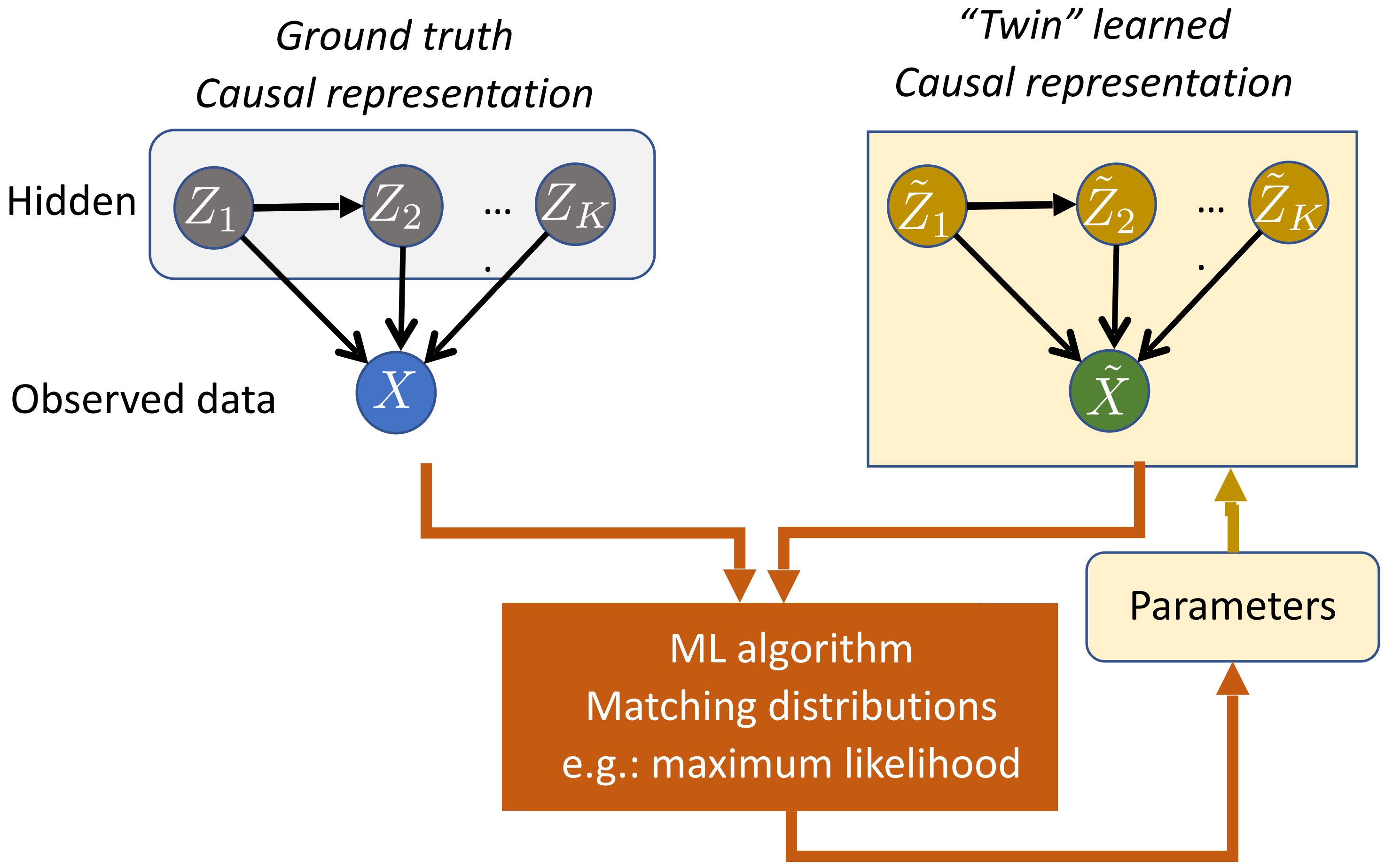
$$X := f(Z_1, Z_2, \ldots, Z_K) \Rightarrow (Z_1, Z_2, \ldots, Z_K) = f^{-1}(X)$$



Unsupervised causal representation learning

Assumption: Observed data is parameterized by hidden causal variables

$$X := f(Z_1, Z_2, \ldots, Z_K) \Rightarrow (Z_1, Z_2, \ldots, Z_K) = f^{-1}(X)$$



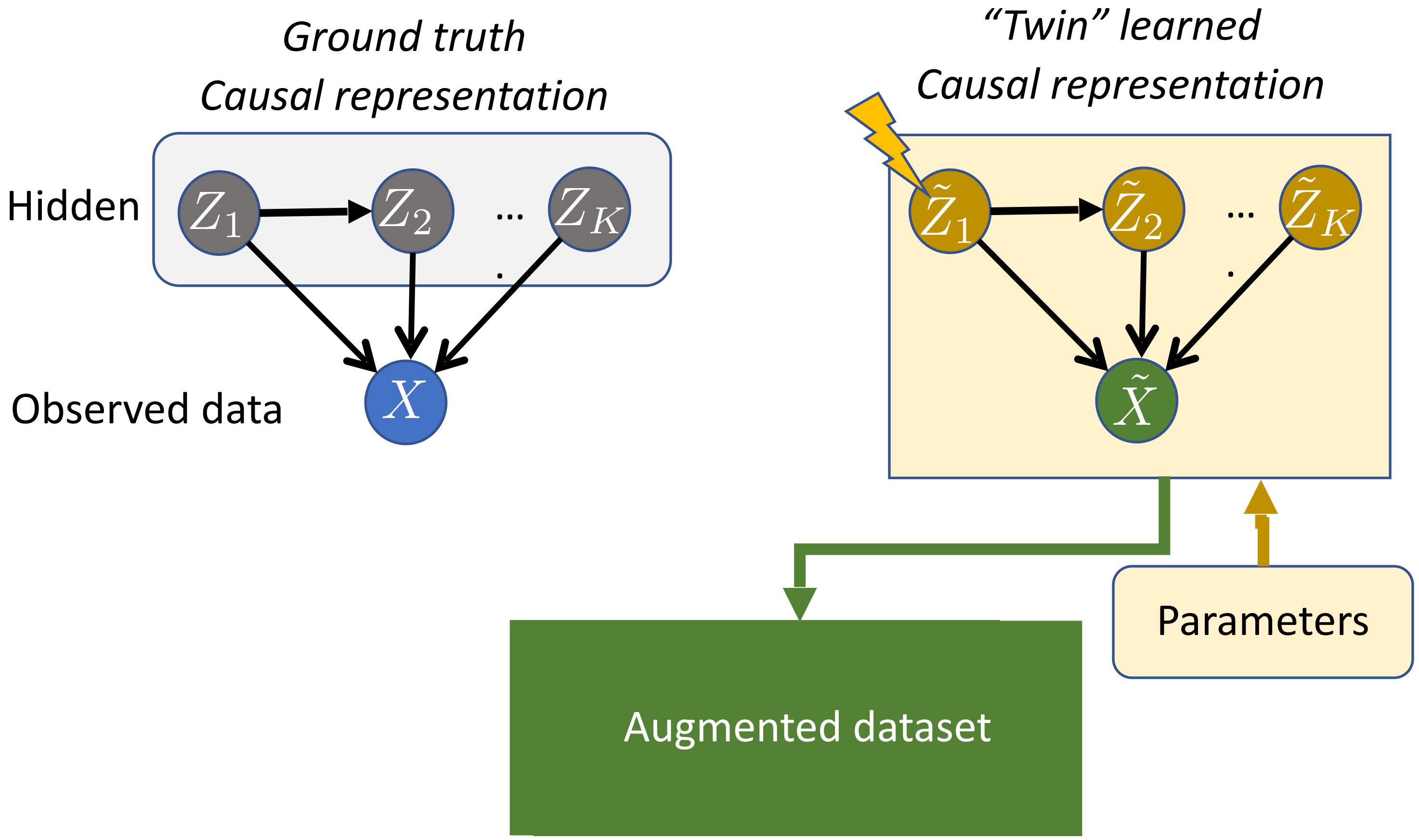
Identifiability:
If the generative model fits the data perfectly, it corresponds to the ground-truth model (up to trivial transformations).

Implications:
We can do "data augmentation" by fitting the model and performing interventions and counterfactuals.

Unsupervised causal representation learning

Assumption: Observed data is parameterized by hidden causal variables

$$X := f(Z_1, Z_2, \ldots, Z_K) \Rightarrow (Z_1, Z_2, \ldots, Z_K) = f^{-1}(X)$$



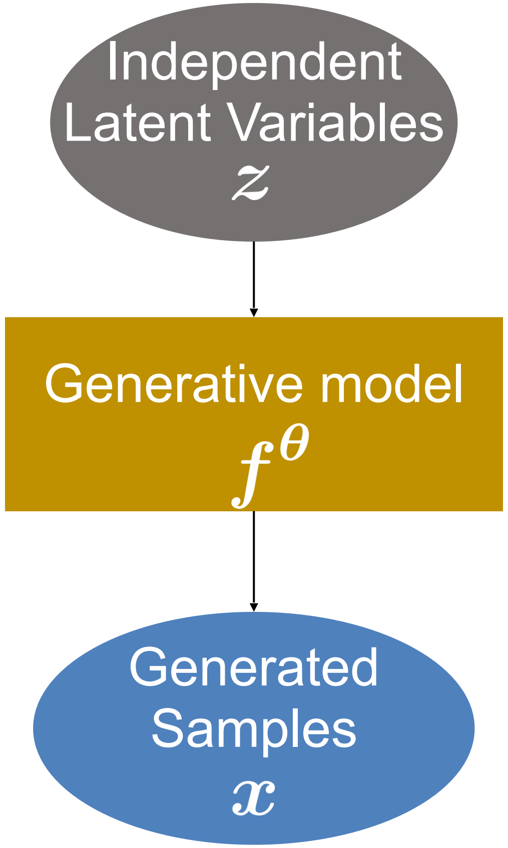
Identifiability:
If the generative model fits the data perfectly, it corresponds to the ground-truth model (up to trivial transformations).

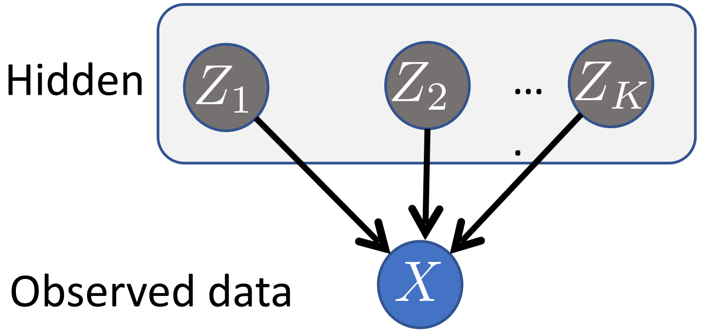
Implications:
We can do "data augmentation" by fitting the model and performing interventions and counterfactuals.

Generic Non-identifiability in Generative Models

[Hyvärinen & Pajunen, Neur. Netw. 1999, Locatello et al, ICML 2019]

Architecture

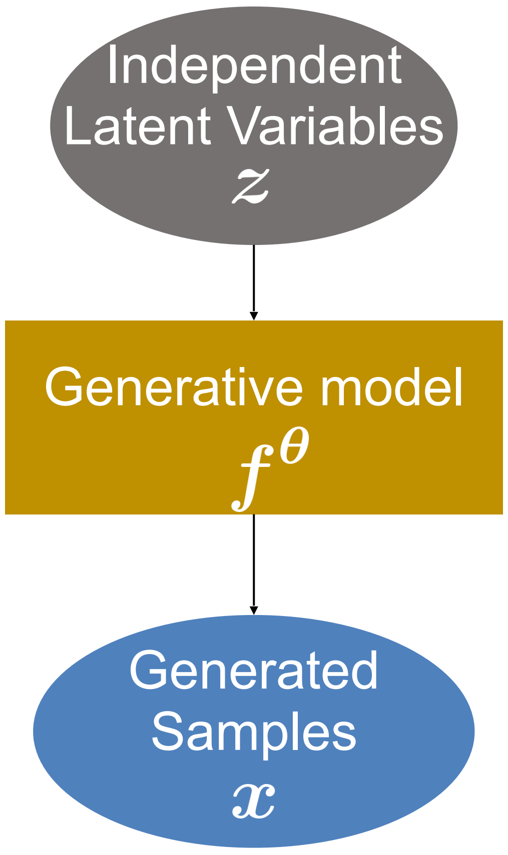




Generic Non-identifiability in Generative Models

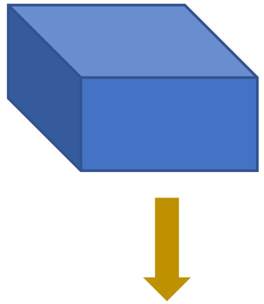
[Hyvärinen & Pajunen, Neur. Netw. 1999, Locatello et al, ICML 2019]

Architecture

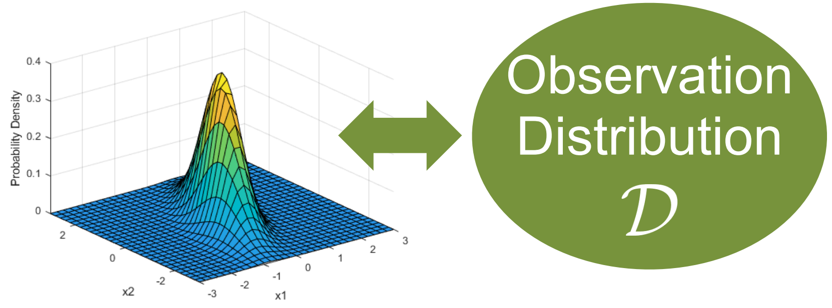


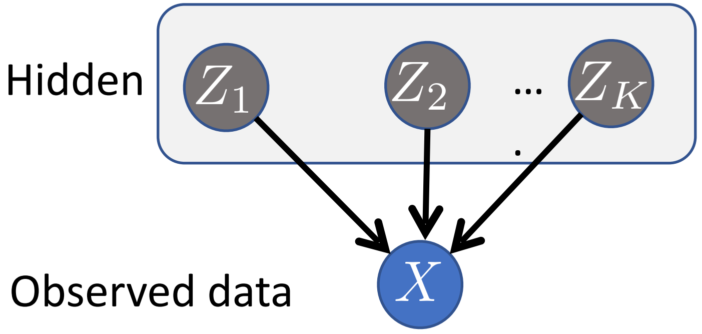
Distributions

$$p_0(\boldsymbol{z}) = \prod p_{0,k}(z_k)$$



$$p(\boldsymbol{x}) = \overbrace{\boldsymbol{f}_*[p_0(\boldsymbol{z})]}^{\textit{Pushforward}} \sim \mathcal{D}$$

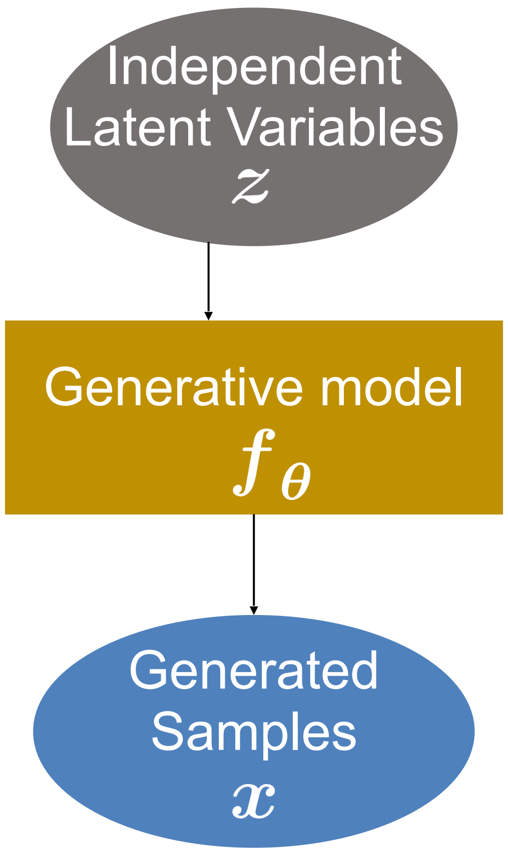




Generic Non-identifiability in Generative Models

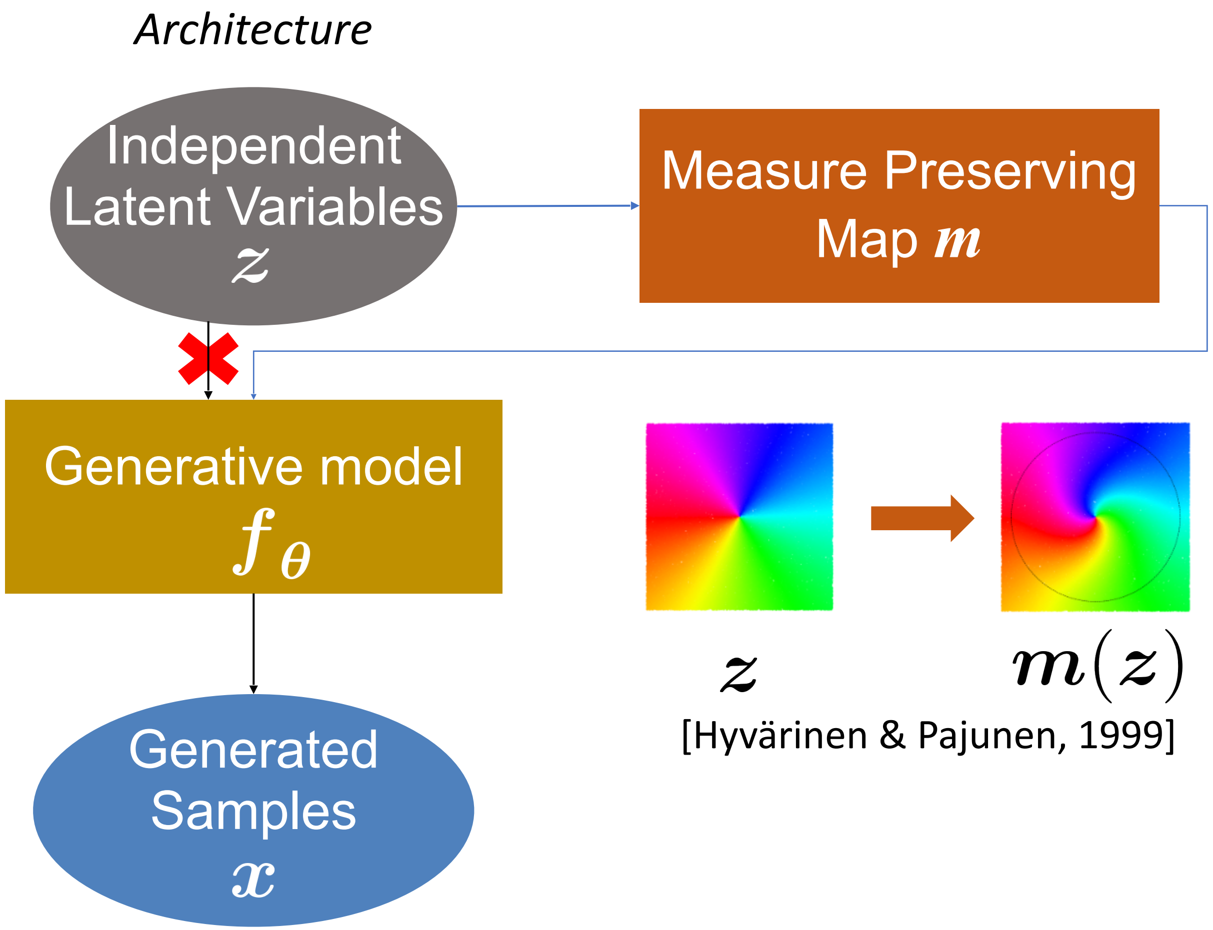
[Hyvärinen & Pajunen, Neur. Netw. 1999, Locatello et al, ICML 2019]

Architecture



Generic Non-identifiability in Generative Models

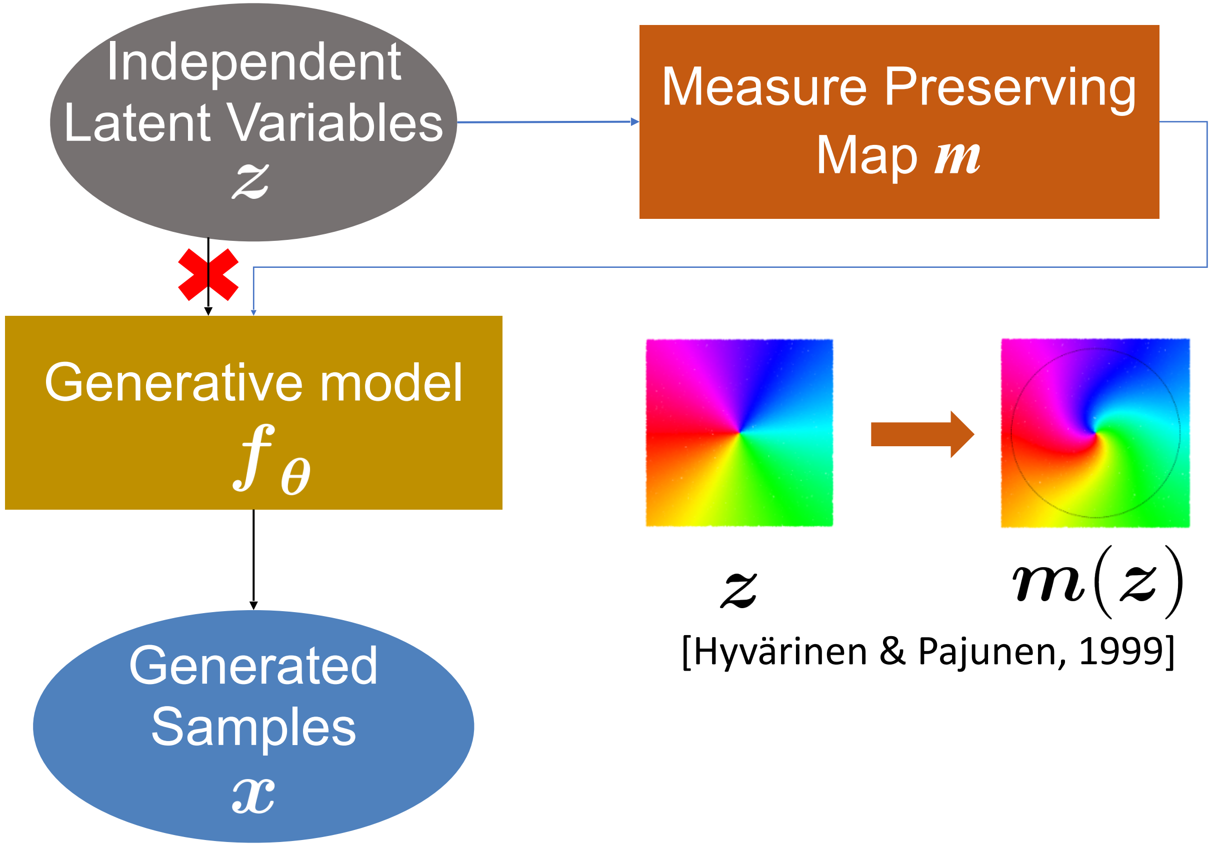
[Hyvärinen & Pajunen, Neur. Netw. 1999, Locatello et al, ICML 2019]



Generic Non-identifiability in Generative Models

[Hyvärinen & Pajunen, Neur. Netw. 1999, Locatello et al, ICML 2019]

Architecture

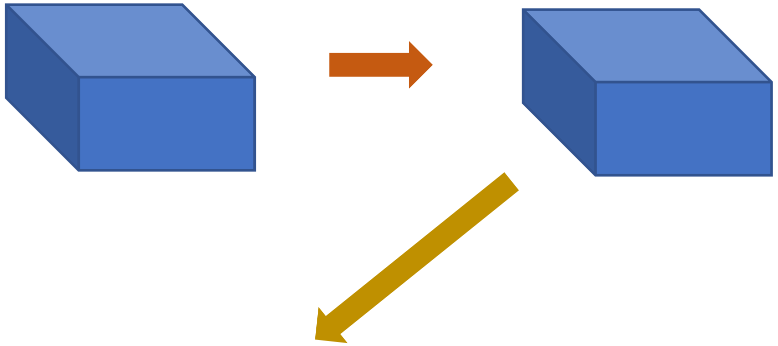


$\boldsymbol{z}$ $\boldsymbol{m}(\boldsymbol{z})$

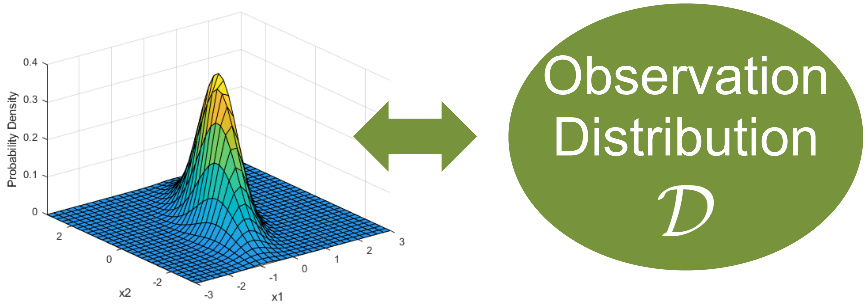
[Hyvärinen & Pajunen, 1999]

Distributions

$p_0(\boldsymbol{z})$ $\boldsymbol{m}[p_0] = p_0(\boldsymbol{z})$

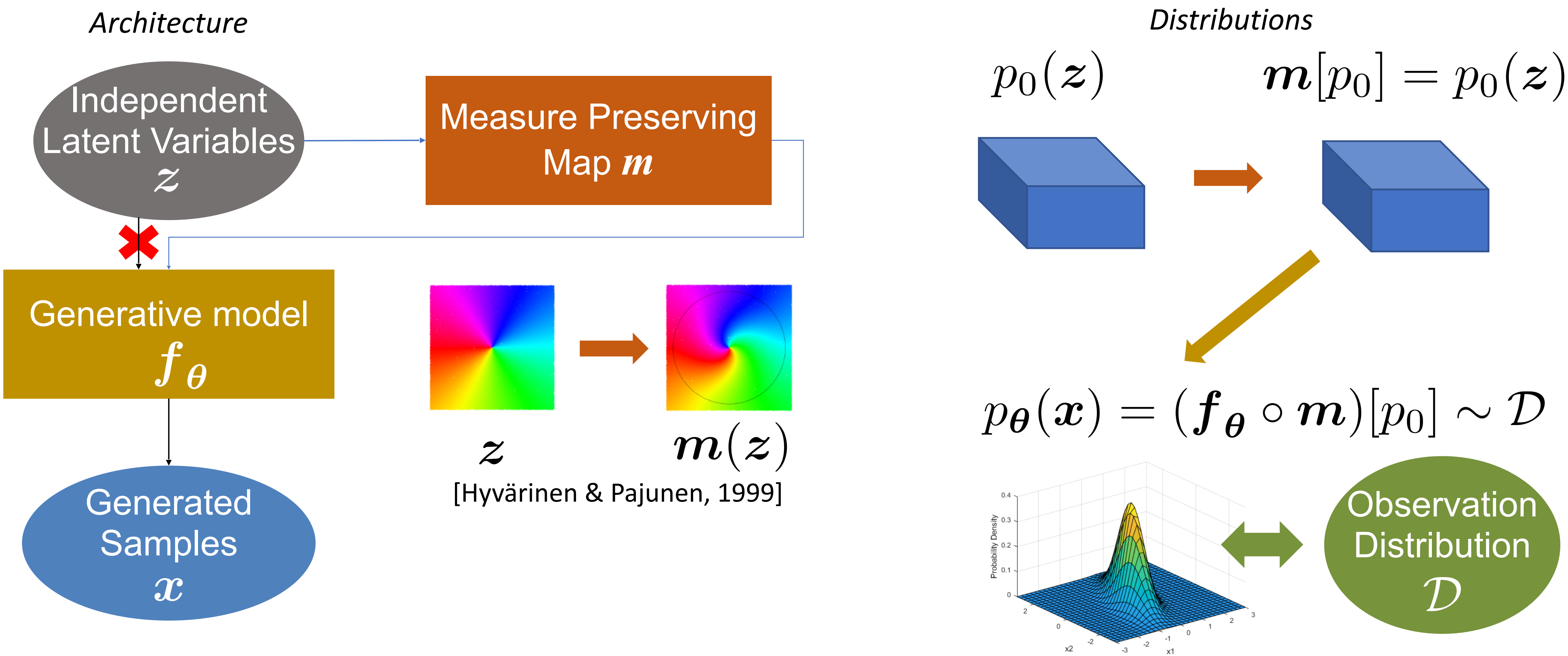


$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = (\boldsymbol{f}_{\boldsymbol{\theta}} \circ \boldsymbol{m})[p_0] \sim \mathcal{D}$$



Generic Non-identifiability in Generative Models

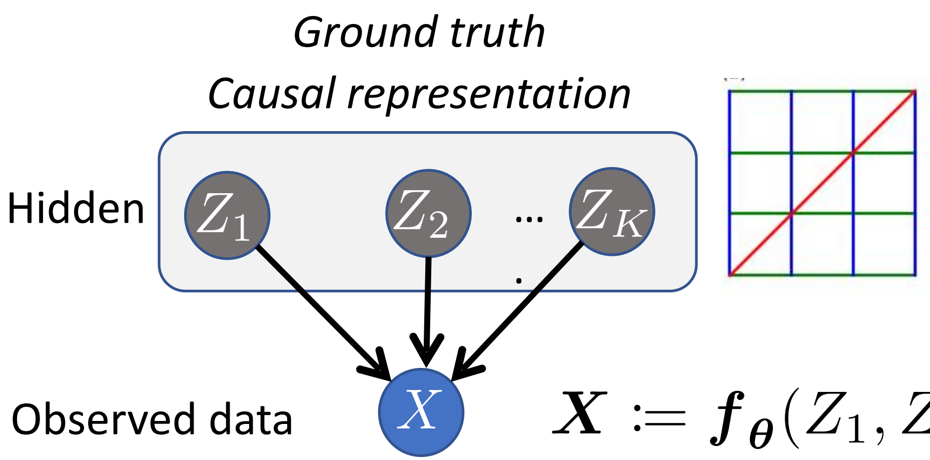
[Hyvärinen & Pajunen, Neur. Netw. 1999, Locatello et al, ICML 2019]



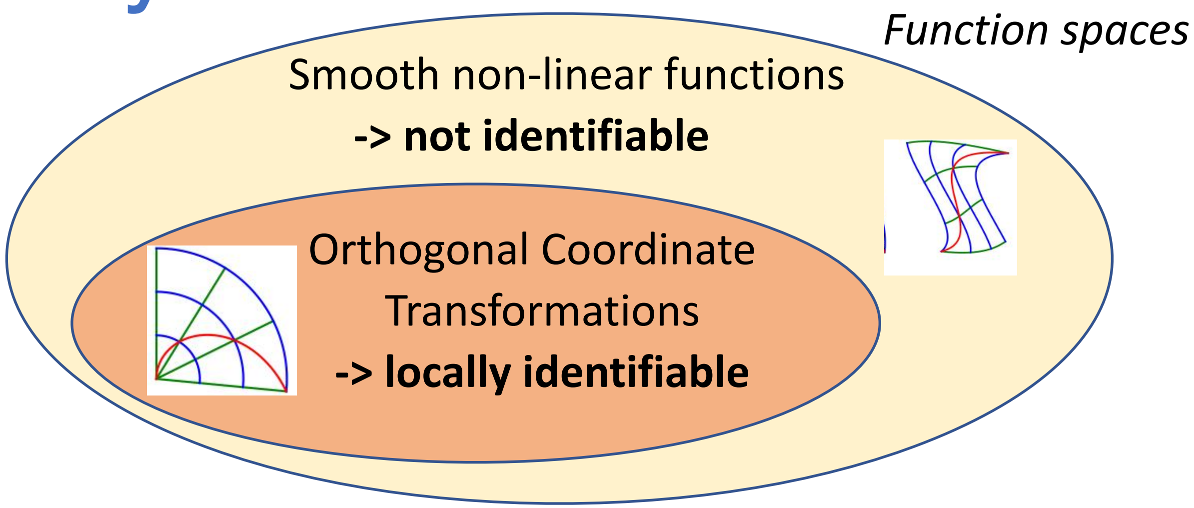
For generic nonlinear ***f***, there exist large families of **spurious solutions** $\mathcal{S}_{\boldsymbol{f}}$ such that

$$\boldsymbol{f}_{\boldsymbol{\theta}}[p_0] = \tilde{\boldsymbol{f}}[p_0], \forall \tilde{\boldsymbol{f}} \in \mathcal{S}_{\boldsymbol{f}}$$

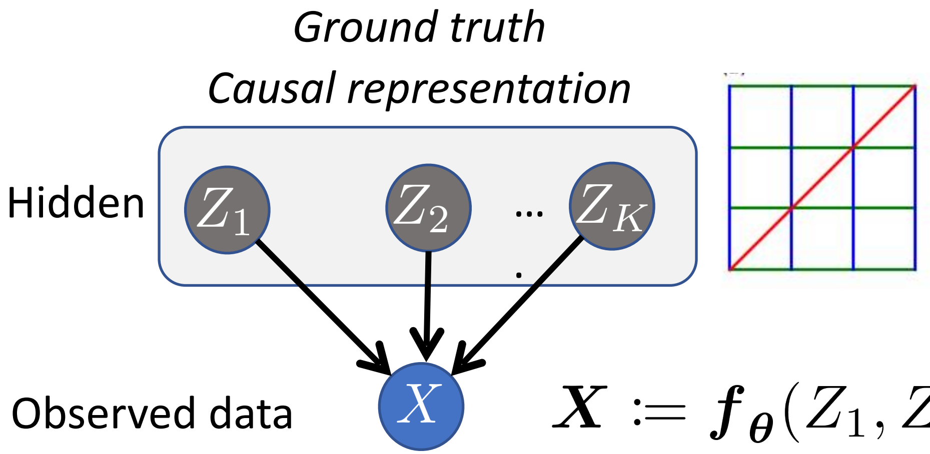
Independent Mechanism Analysis



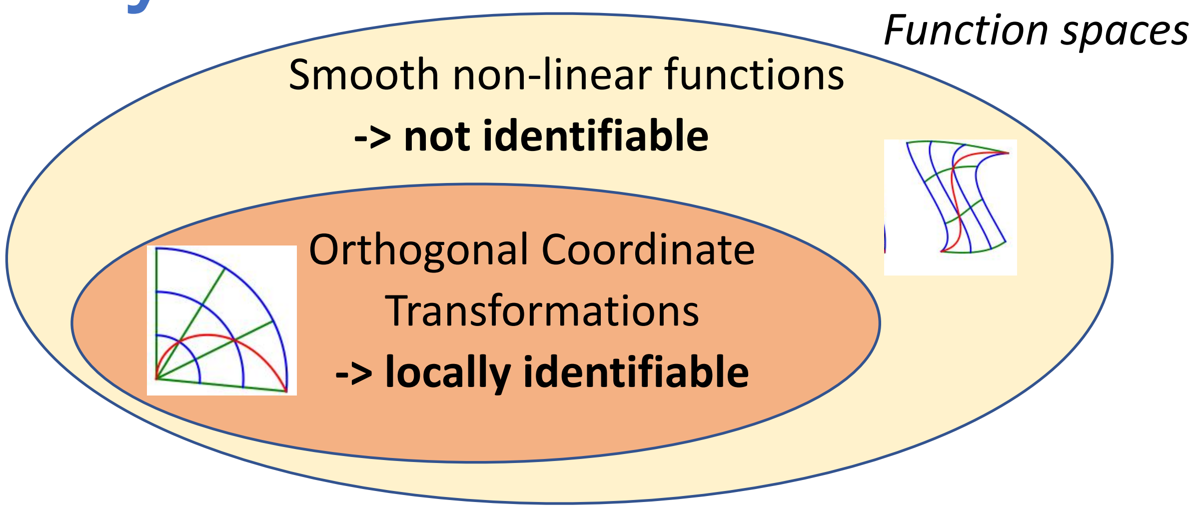
$$\boldsymbol{X} := \boldsymbol{f}_{\boldsymbol{\theta}}(Z_1, Z_2, \ldots, Z_K)$$



Independent Mechanism Analysis



$$\boldsymbol{X} := \boldsymbol{f}_{\boldsymbol{\theta}}(Z_1, Z_2, \dots, Z_K)$$



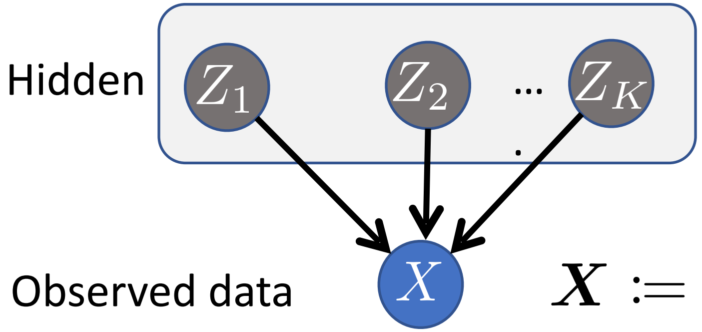
➢ *Constraining* **f** to a smaller (large!) model class, via regularized maximum likelihood, with Normalizing flows, favors identifiability [Gresele*, von Kügelgen* et al, NeurIPS 2021; Buchholz et al., NeurIPS 2022]

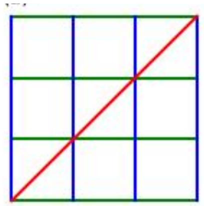
$$\text{maximize}/\boldsymbol{\theta} \quad \mathcal{L}_{IMA}(\boldsymbol{x}; \boldsymbol{f}_{\boldsymbol{\theta}}, \lambda) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \lambda . c_{IMA}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x}))$$

Independent Mechanism Analysis

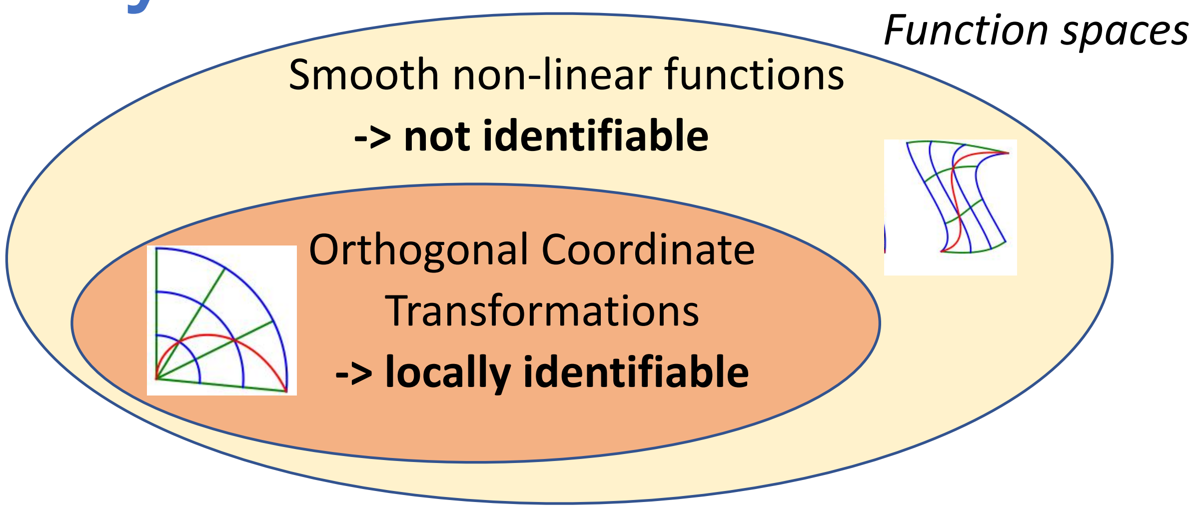
Ground truth

Causal representation





$$\boldsymbol{X} \coloneqq \boldsymbol{f_\theta}(Z_1, Z_2, \ldots, Z_K)$$



➢*Constraining* **f** to a smaller (large!) model class, via regularized maximum likelihood, with Normalizing flows, favors identifiability [Gresele*, von Kügelgen* et al, NeurIPS 2021; Buchholz et al., NeurIPS 2022]

$$\text{maximize}/\boldsymbol{\theta} \quad \mathcal{L}_{IMA}(\boldsymbol{x}; \boldsymbol{f_\theta}, \lambda) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \lambda . c_{IMA}(\boldsymbol{f_\theta}, \boldsymbol{f_\theta}^{-1}(\boldsymbol{x}))$$

➢*Variational Auto-Encoders (VAE) implicitly constrain f in the above model class and inherit identifiability benefits.* [Reizinger*, Gresele*, Brady*,et al. NeurIPS2022]

Evidence Lower BOund :

$$\underbrace{\text{ELBO}(\boldsymbol{x};\, \boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \underbrace{\text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \| p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))}_{\substack{\textit{ELBO gap} \geq 0 \\ \text{minimize}/\boldsymbol{\phi}}}}_{\text{maximize}/\boldsymbol{\theta}}$$

Assumptions for *learning* causal generative AIs

To identify causal representations, we need either:

(1) **inductive biases** (assumptions on the ground truth model)

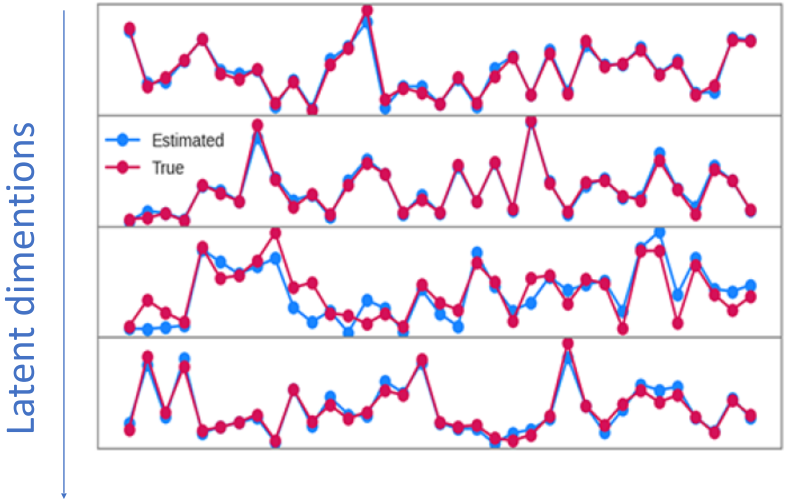
[Gresele*, von Kügelgen* et al, ***NeurIPS*** 2021; Buchholz et al., ***NeurIPS*** 2022;

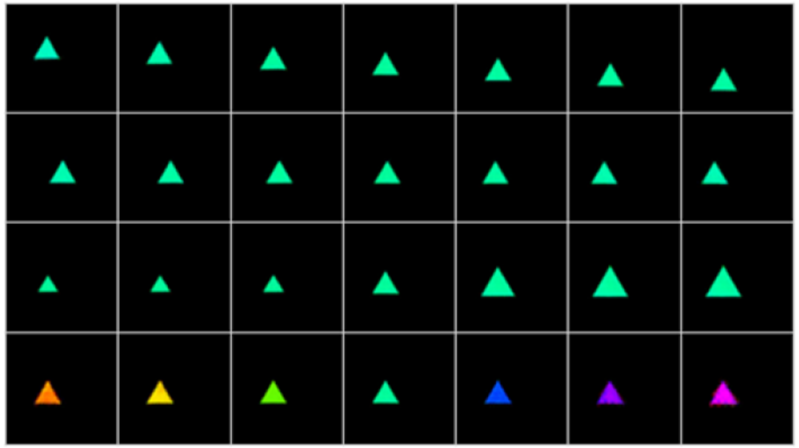
Reizinger*, Gresele*, Brady* et al. ***NeurIPS*** 2022; Leeb et al., ***ICLR*** 2023]

(2) **multi-environment data** = interventions.
[Keurti et al., ***ICML*** 2023; von Kügelgen et al., ***NeurIPS***, 2023; Liang et al., ***NeurIPS***, 2023]

(3) **multi-view data** = counterfactuals.

[Besserve et al. ***AAAI*** 2021; von Kügelgen et al., ***NeurIPS***, 2021]





[Reizinger*, Gresele*, Brady*,et al. NeurIPS2022]

Assumptions for *learning* of causal generative AI

To identify causal representations, we need either:

(1) **inductive biases** (assumptions on the ground truth model)

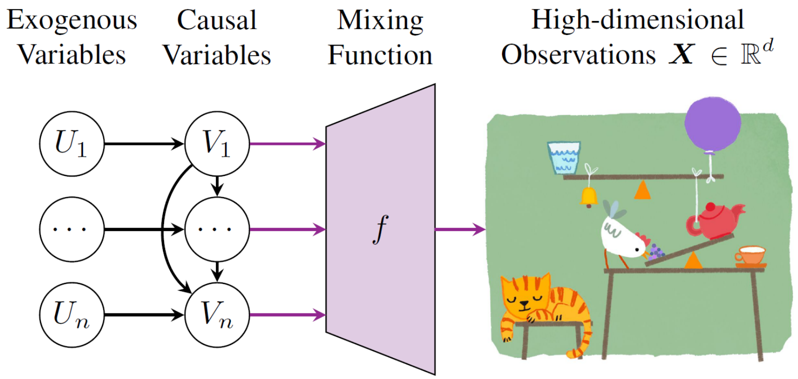
[Gresele*, von Kügelgen* et al, ***NeurIPS*** 2021; Buchholz et al., ***NeurIPS*** 2022; Reizinger*, Gresele*, Brady* et al. ***NeurIPS*** 2022; Leeb et al., ***ICLR*** 2023]

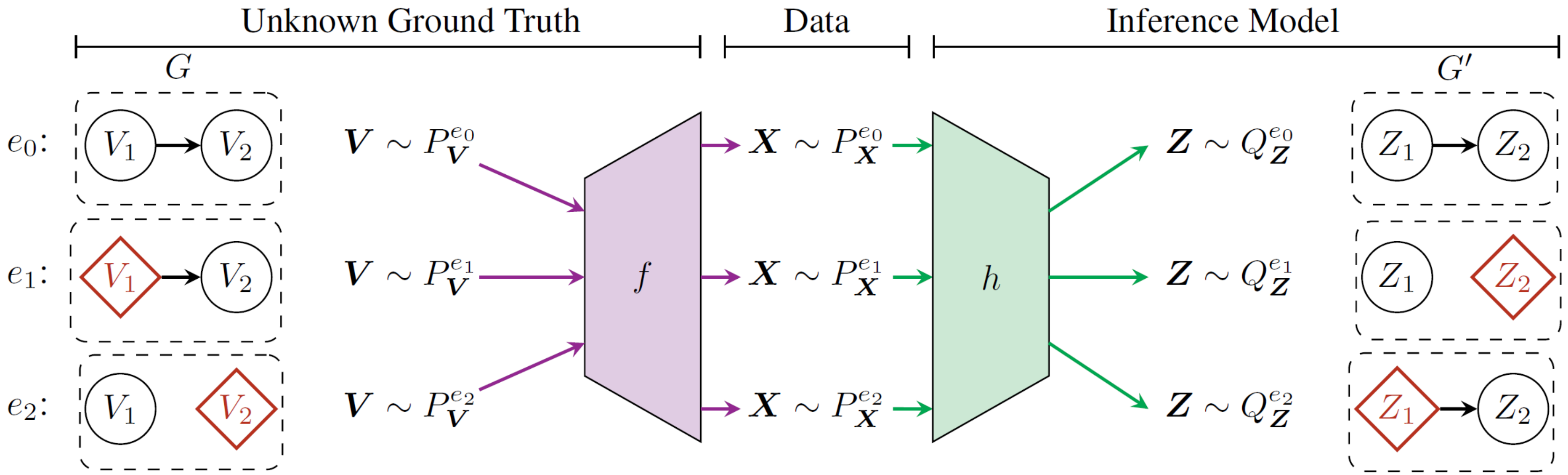
(2) multi-environment data = interventions.

[Keurti et al., ***ICML*** 2023; von Kügelgen et al., ***NeurIPS***, 2023; Liang et al., ***NeurIPS***, 2023]

(3) **multi-view data** = counterfactuals.

[Besserve et al. ***AAAI*** 2021; von Kügelgen et al., ***NeurIPS***, 2021]





[Von Kügelgen et al., NeurIPS, 2023]

Assumptions for *learning* of causal generative AI

To identify causal representations, we need either:

(1) **inductive biases** (assumptions on the ground truth model class)

[Gresele*, von Kügelgen* et al, ***NeurIPS*** 2021; Buchholz et al., ***NeurIPS*** 2022;

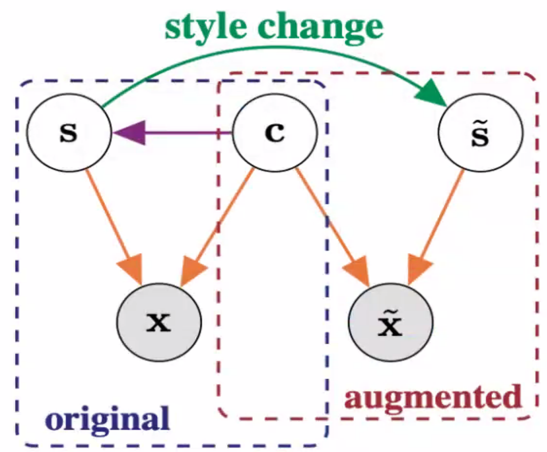
Reizinger*, Gresele*, Brady* et al. ***NeurIPS*** 2022; Leeb et al., ***ICLR*** 2023]

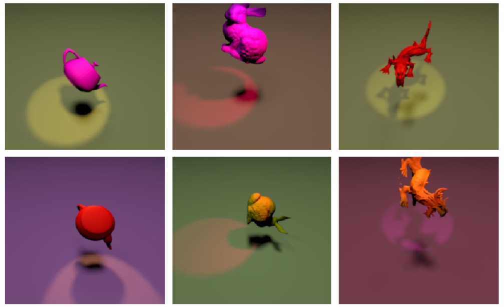
(2) **multi-environment data** = interventions.

[von Kügelgen et al., ***NeurIPS***, 2023; Liang et al., ***NeurIPS***, 2023]

(3) **multi-view data** = counterfactuals, and structured sample-dependency

[Besserve et al. ***AAAI*** 2021: von Kügelgen et al., ***NeurIPS***, 2021, Keurti et al., ***ICML*** 2023]





DA: data augmentation

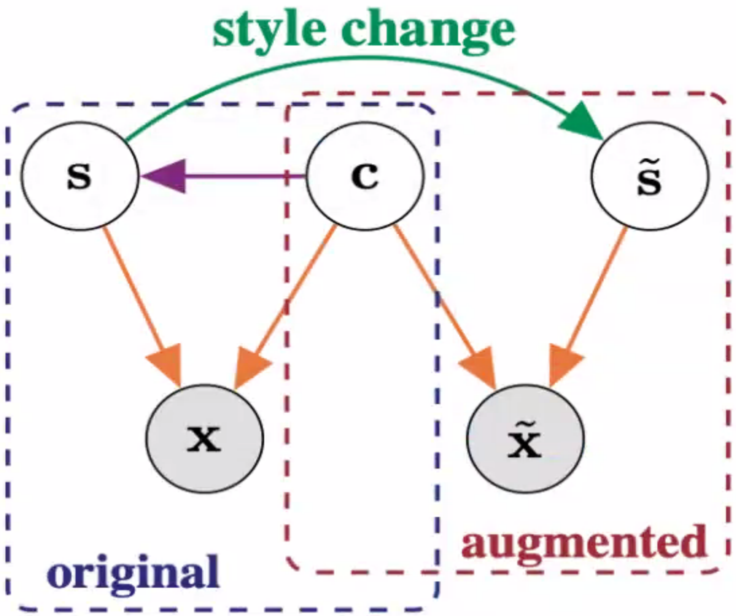
LT: latent transformation

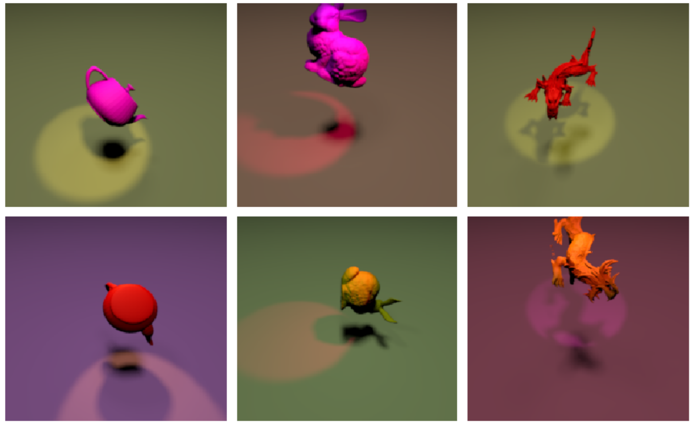
Views generated by	Class
DA: colour distortion	0.42 ± 0.01
LT: change hues	$\mathbf{1.00} \pm 0.00$
DA: crop (large)	0.28 ± 0.04
DA: crop (small)	0.14 ± 0.00
LT: change positions	$\mathbf{1.00} \pm 0.00$
DA: crop (large) + colour distortion	$\mathbf{0.97} \pm 0.00$
DA: crop (small) + colour distortion	$\mathbf{1.00} \pm 0.00$
LT: change positions + hues	$\mathbf{1.00} \pm 0.00$
DA: rotation	0.33 ± 0.06
LT: change rotations	$\mathbf{1.00} \pm 0.00$
DA: rotation + colour distortion	$\mathbf{0.59} \pm 0.01$
LT: change rotations + hues	$\mathbf{1.00} \pm 0.00$

[Von Kügelgen et al., NeurIPS, 2021]

Sample dependency-based identification: Case 1

- **Data augmentation** is a particular case of **self-supervised learning**:
  - Augmentations change "style" of the image but not "content",
  - Data-set contains *(original, augmented)* pairs of samples
  - Result: augmentation-invariant representations identify content.





DA: data augmentation
LT: latent transformation

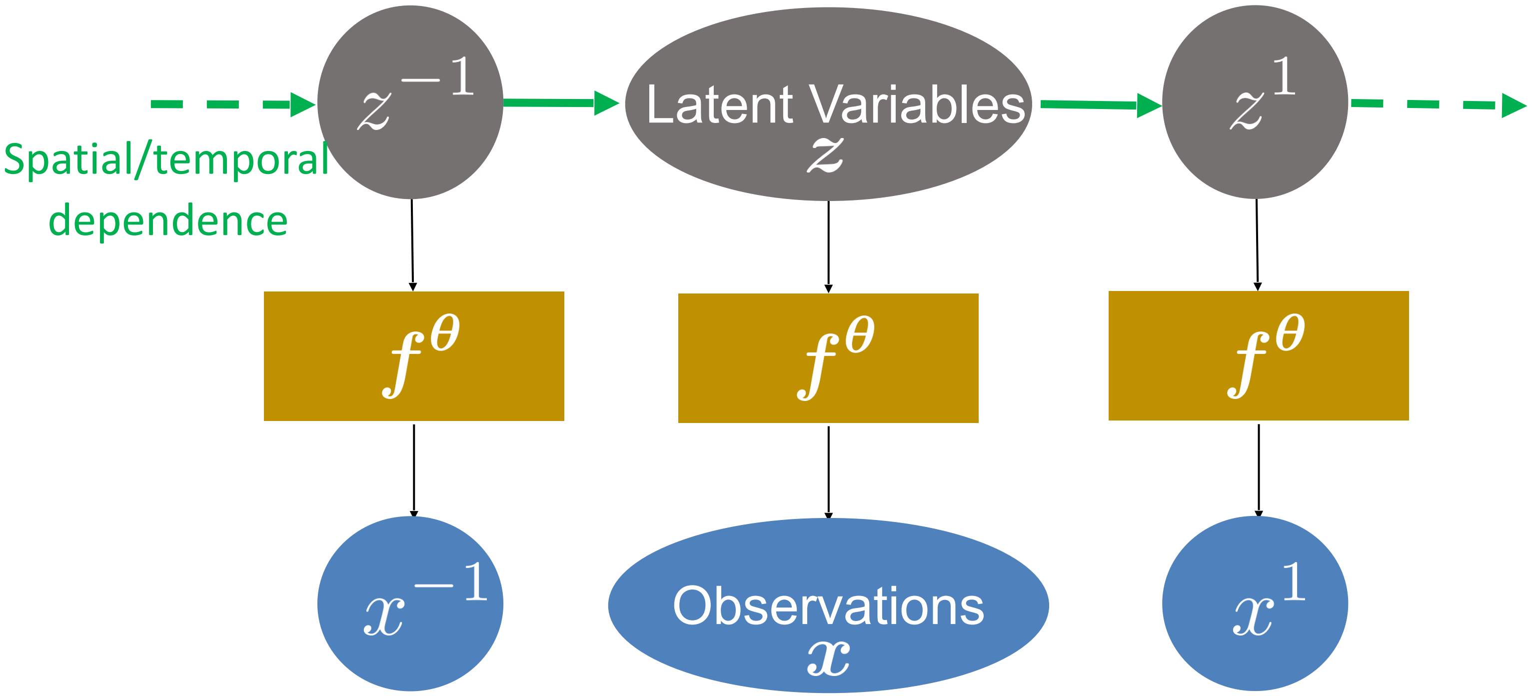
Views generated by	Class
DA: colour distortion	0.42 ± 0.01
LT: change hues	$\mathbf{1.00} \pm 0.00$
DA: crop (large)	0.28 ± 0.04
DA: crop (small)	0.14 ± 0.00
LT: change positions	$\mathbf{1.00} \pm 0.00$
DA: crop (large) + colour distortion	$\mathbf{0.97} \pm 0.00$
DA: crop (small) + colour distortion	$\mathbf{1.00} \pm 0.00$
LT: change positions + hues	$\mathbf{1.00} \pm 0.00$
DA: rotation	0.33 ± 0.06
LT: change rotations	$\mathbf{1.00} \pm 0.00$
DA: rotation + colour distortion	$\mathbf{0.59} \pm 0.01$
LT: change rotations + hues	$\mathbf{1.00} \pm 0.00$

[Von Kügelgen et al., NeurIPS, 2021]

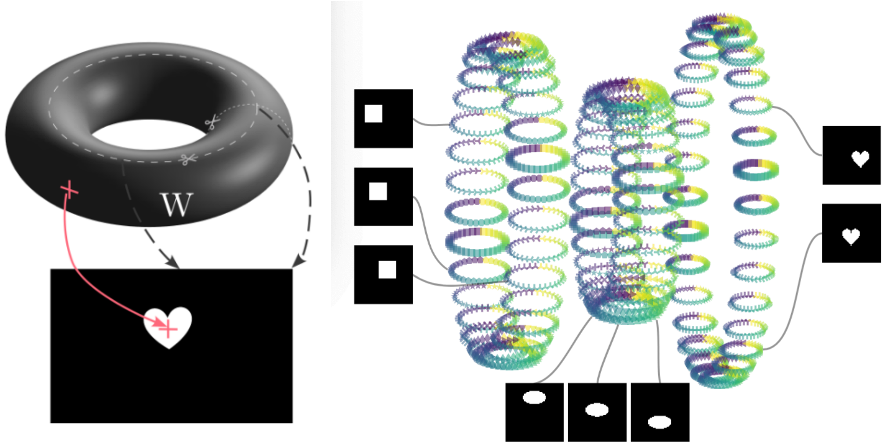
- Challenges: augmentation need to have enough variability, and respect the latent causal structure.

Sample dependency-based identification: Case 2

- Latent time/space structure



Learning latent Lie group structure by manipulation



Keurti et al., ***ICML*** 2023

Several other results on sequential/spatial models: Hälvä & Hyvärinen, 2020, and many others…

Relevance for spatio-temporal emulators and data assimilation (Jordi and Andrew's talks)

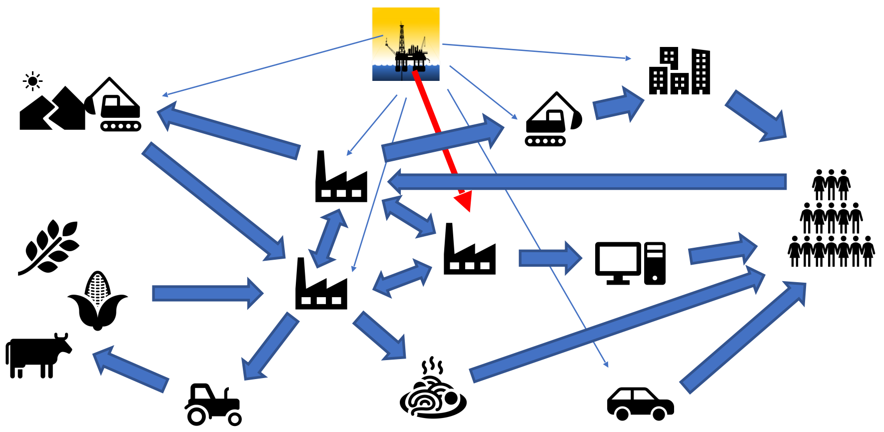
Learning high-level causal explanations

[Kekić et al. and MB, UAI 2024]

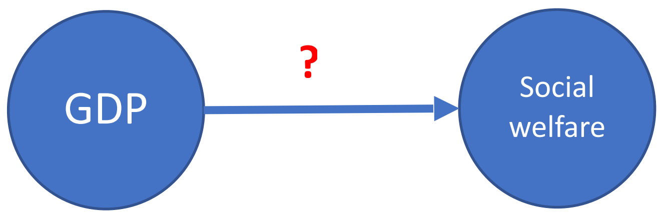
Causal explanations in complex systems

- Experts and decision makers need to get a high-level, simplified representation of complex systems to take decisions.
- This relies on finding “aggregated” indicators to describe the system.
- Those are used to make explicit or implicit causal claims.
- Not obvious these claims are correct!

Low-level representation

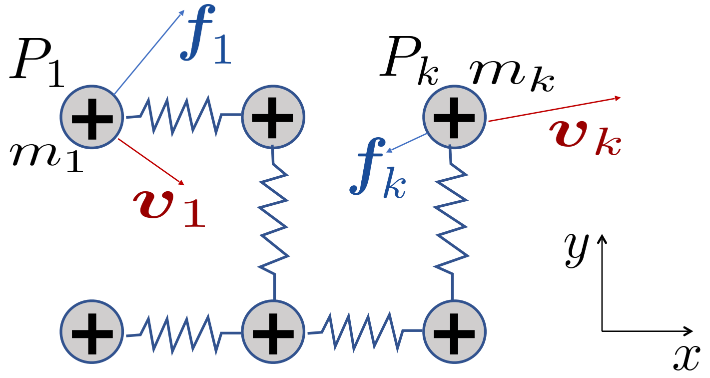


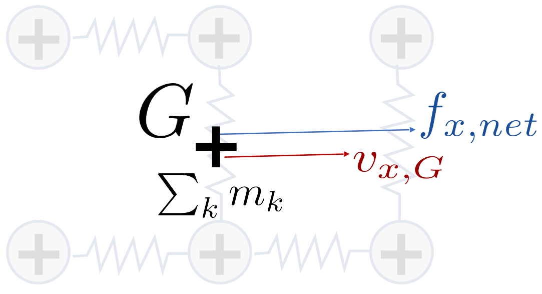
High-level, reduced representation



Analogy to mechanics

	Microscopic system "low-level"		Macroscopic system "high-level"
Interventions:	$(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N)$	$\xrightarrow{\omega}$	$f_{x,net} = \sum f_{x,k}$





System variables:	$(\boldsymbol{v}_1, \ldots \boldsymbol{v}_N)$	$\xrightarrow{\tau}$	$v_{x,G} = \sum_k \frac{m_k}{\sum_l m_l} v_{x,k}$

[Kekić et al., UAI, 2024]

Targeted causal model reduction algorithm

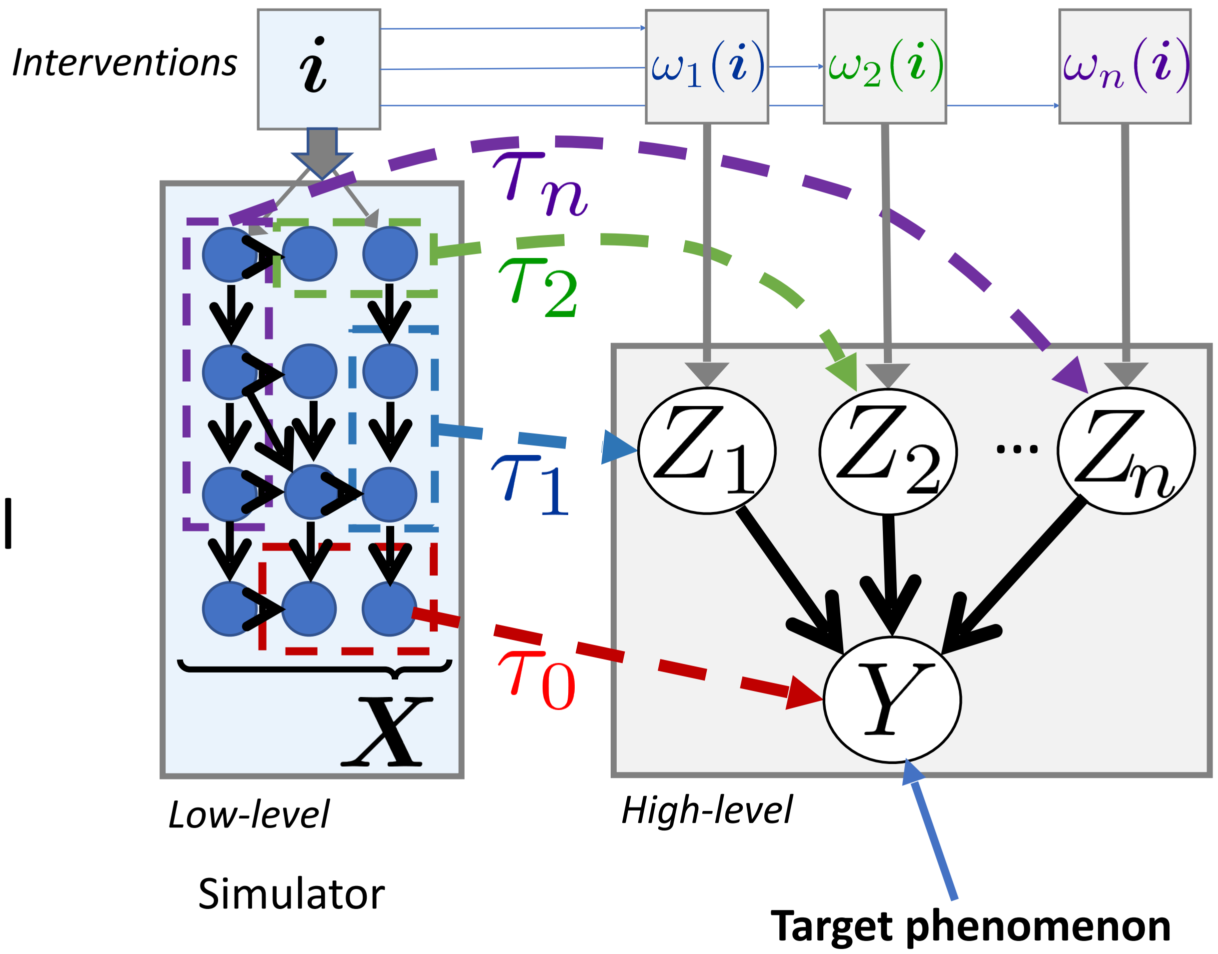
[Kekić et al., UAI 2024]

- Causal consistency principle

Low-level *High-level*

$$\begin{tikzcd} L1 \arrow[r, "\text{Reduction}"] \arrow[d, "\text{intervention}"'] & H1 \arrow[d, "\text{intervention}"] \\ L2 \arrow[r, "\text{Reduction}"'] & H2 \end{tikzcd}$$

- An ML algorithm to learn high-level causes from simulations.

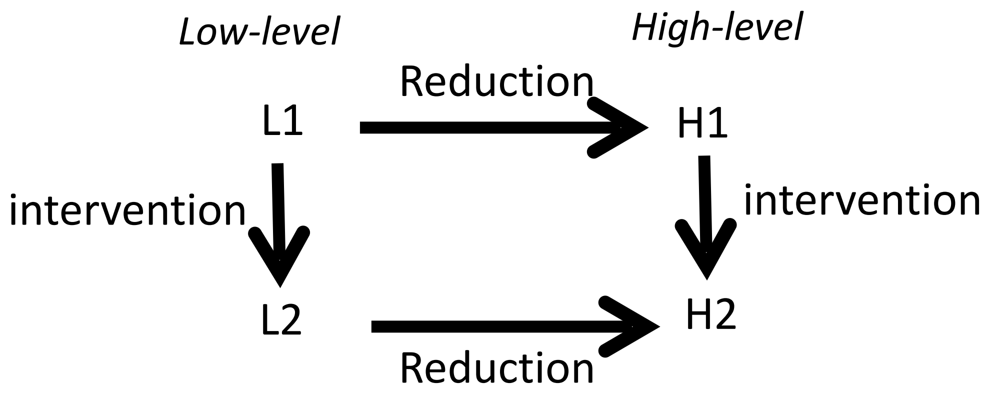


[Kekić et al., UAI 2024]

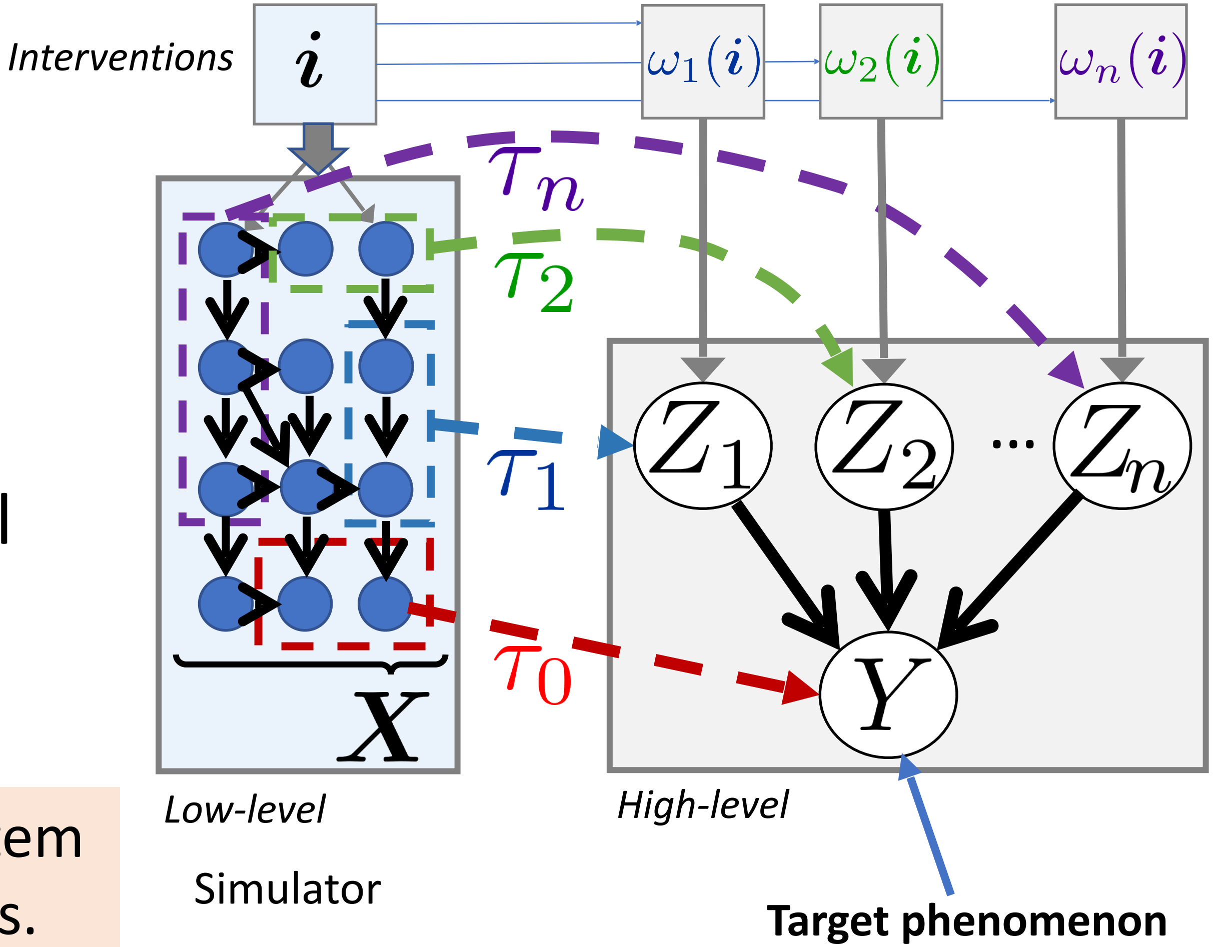
Targeted causal model reduction algorithm

[Kekić et al., UAI 2024]

- Causal consistency principle



- An ML algorithm to learn high-level causes from simulations.

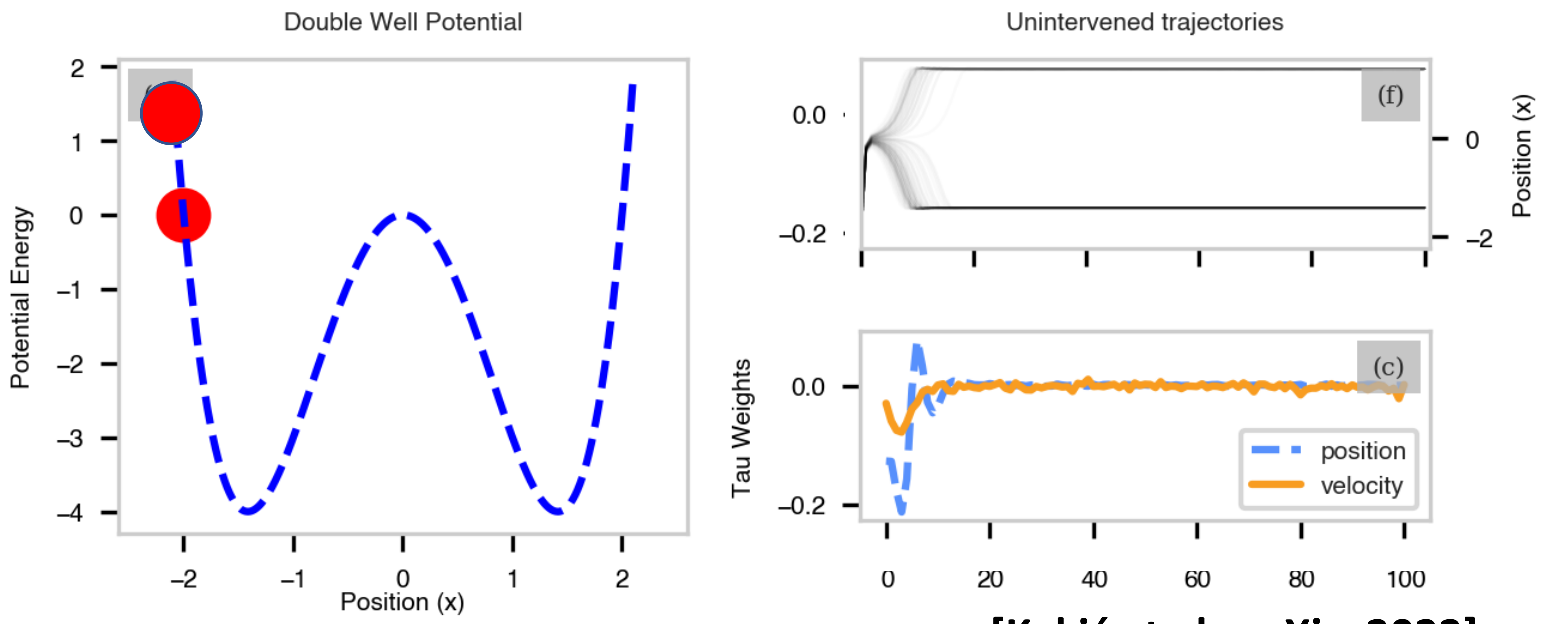


-> A simplified description of a complex system we can interact with through interventions.

[Kekić et al., UAI 2024]

Example: double well simulator (Euler method)

- Interventions via random forces applied along the trajectory.
- Target phenomenon: Final position of the ball



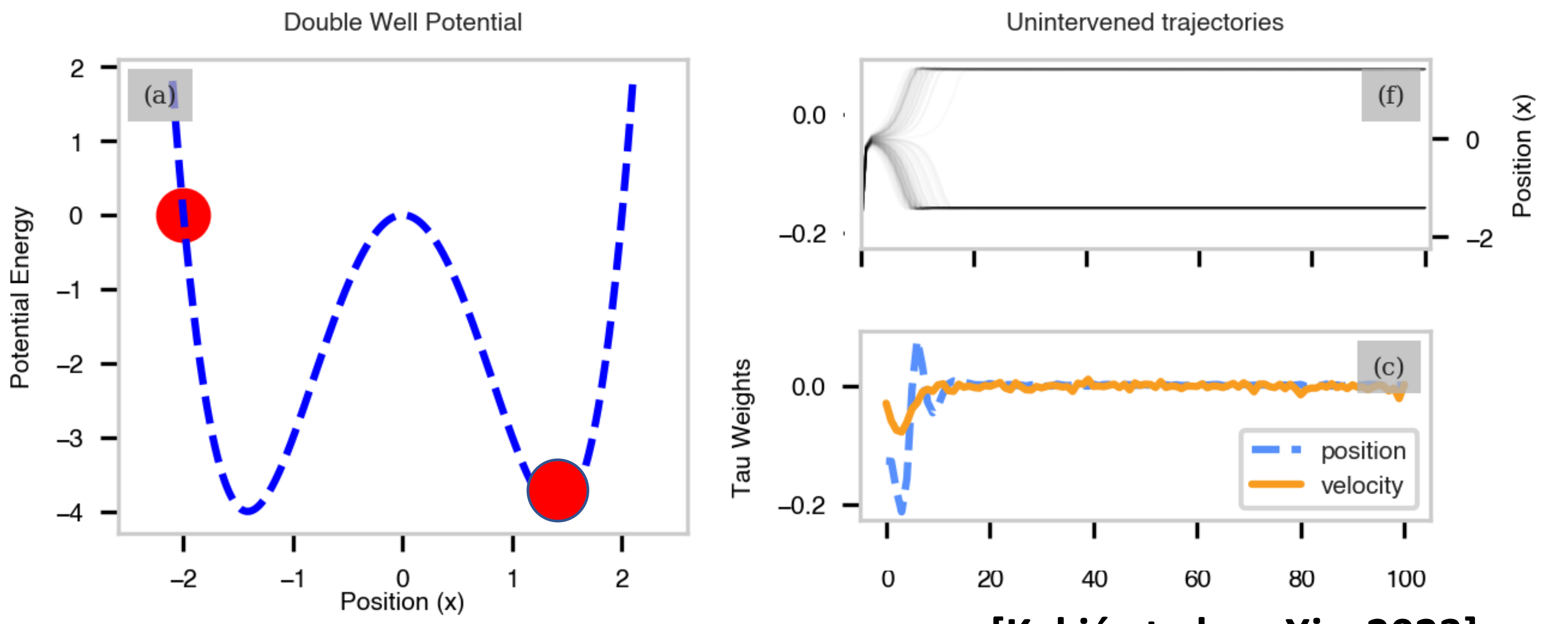
[Kekić et al., arXiv, 2023]

- The algorithm locates the time range where changes are critical for the outcome to happen.
- Ongoing work: enhance interpretability and scalability.

[Kekić et al., UAI 2024]

Example: double well simulator (Euler method)

- Interventions via random forces applied along the trajectory.
- Target phenomenon: Final position of the ball



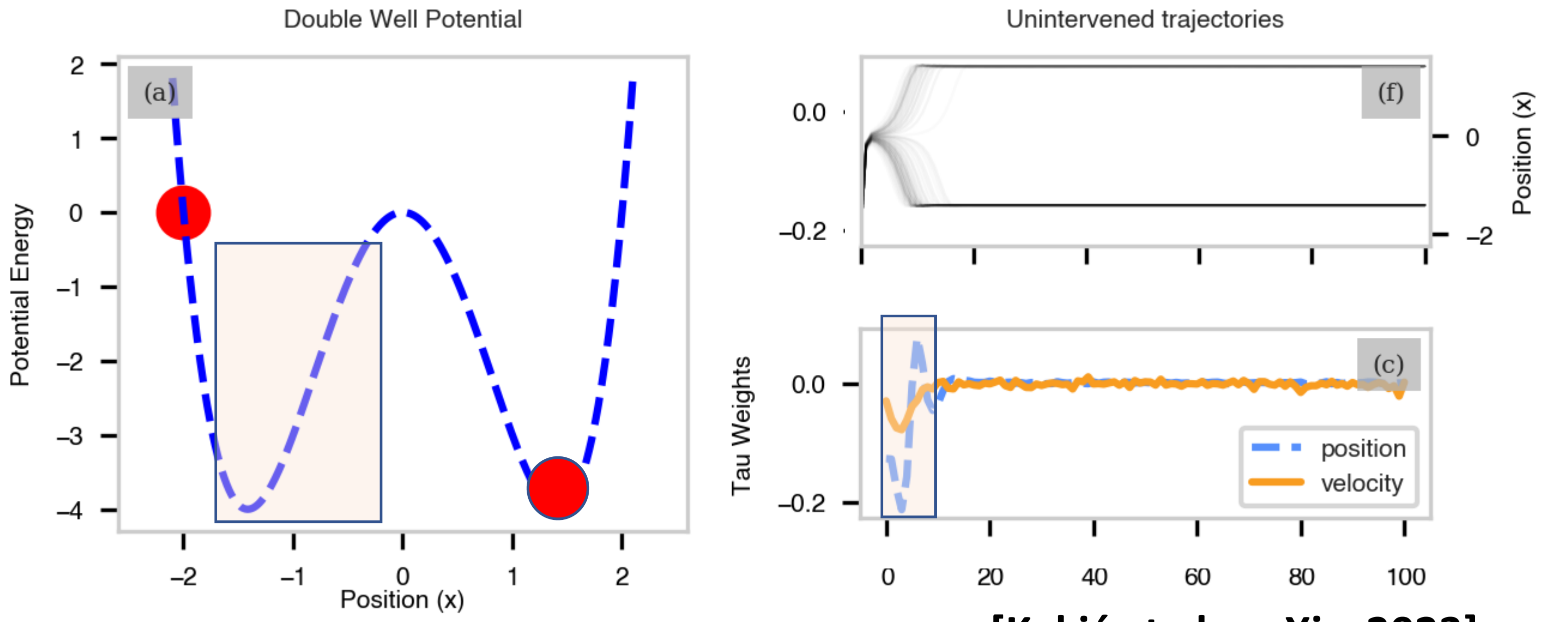
[Kekić et al., arXiv, 2023]

- The algorithm locates the time range where changes are critical for the outcome to happen.
- Ongoing work: enhance interpretability and scalability.

[Kekić et al., UAI 2024]

Example: double well simulator (Euler method)

- Interventions via random forces applied along the trajectory.
- Target phenomenon: Final position of the ball



[Kekić et al., arXiv, 2023]

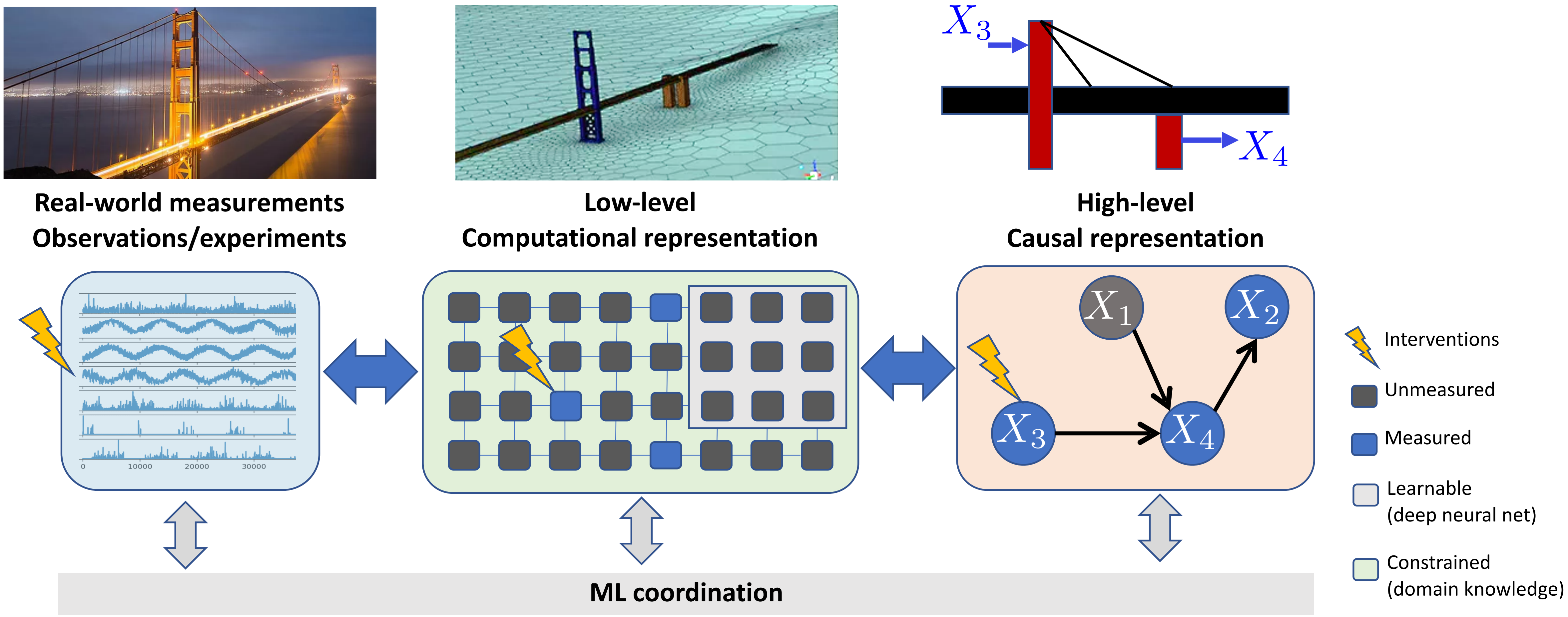
- The algorithm locates the time range where changes are critical for the outcome to happen.
- Ongoing work: enhance interpretability and scalability.

[Kekić et al., UAI 2024]

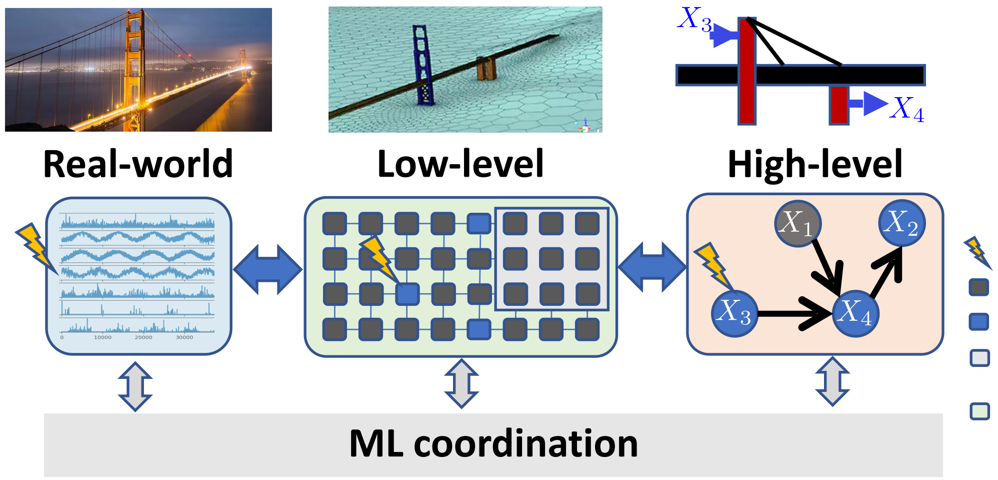
Take-home

- Causality is a promising framework to design interpretable AI robust to changes, but also ***a philosophy for scientific practice: pay attention to data generating mechanisms.***
- There are **guaranties for causal generative AI**, using diverse inductive biases on function spaces, structure, invariances, multiple environments: can guide the choice of AI model and dataset collection.
- Causal reduction/abstraction of complex models into simpler ones allows human-interpretable explanations and control…
  … ***and a way for scientists to prioritize the aim of understanding reality despite the growing complexity of models and data-science pipelines.***

Towards Computational Causal Models (CCM)



Open questions



- How to optimally merge scientific knowledge (e.g. physics equations and principles) with flexible machine learning:
    - ML tools improve predictive power, but may lose physical consistency (Laure’s talk) -> need an identifiability theory for hybrid/”grey box” models (Takeishi et al., 2021)
    - How learning high-level causal models, can help generalize better from models to the real world? (Sim2real gap),
- Training of smaller and modular AI systems whose parts can be reused for other tasks.

Thank you!

Bernhard Schölkopf,
Remy Sun, Arash Mehrjou, Luigi Gresele, Julius von Kügelgen, Patrik Reizinger, Jack Brady, Felix Leeb, Hsiao-Ru Pan, Hamza Keurti, Armin Kekić.

Starting a new lab (TU Braunschweig):
open positions, collaborations, contact me! ☺

TÜBINGEN AI CENTER

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B

References

- Besserve et al., Counterfactuals Uncover the Modular Structure of Deep Generative Models, ICLR 2020
- Besserve et al., Learning Soft Interventions in Complex Equilibrium Systems, UAI 2022
- Buchholz et al., Function Classes for Identifiable Nonlinear ICA, NeurIPS 2022
- Gresele*, von Kügelgen* et al., Independent Mechanism Analysis (IMA), NeurIPS *2021*
- Hälvä & Hyvärinen, Hidden Markov Nonlinear ICA: Unsupervised Learning from nonstationary Time Series, UAI 2020
- Kekić et al., Trageted Reduction of Causal Models, arXiv 2023
- Keurti et al., Homomorphism AutoEncoder — Learning Group Structured Representations from Observed Transitions, ICML 2023
- Reizinger*, Gresele*, Brady* et al., Embrace the Gap, VAEs perform IMA, *NeurIPS 2022*
- Takeishi et al., Physics-Integrated Variational Autoencoders for Robust and Interpretable Generative Modeling, NeurIPS 2021
- von Kügelgen et al., Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style, NeurIPS 2021.