CAUSAL REPRESENTATION LEARNING

LECTURER: M.A. MIRZAIE

TOPICS

- Intelligence
- Causality and Correlation
- Levels of Causal Modeling
- Causal Models and Inference
- Structural Causal Models (SCMs)
- Causal Learning and Reasoning
- Causal Representation Learning
- Causal VAE
- Conclusion

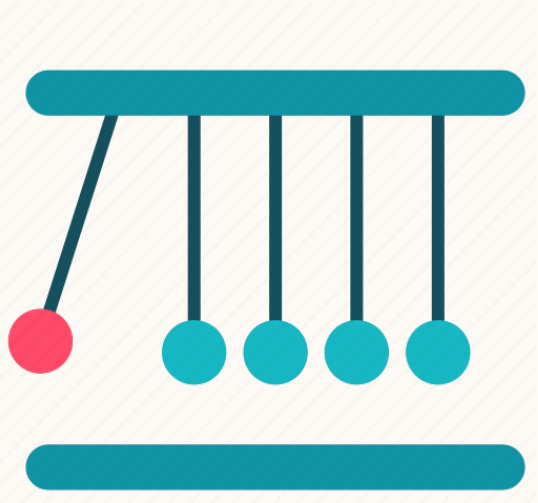
INTELLIGENCE

- Machine Learning vs. Animal's Intelligence
  - Limited at Some Crucial Feats
  - Out of Distribution Generalization (from one problem to another rather than one data point to another)
  - Interventions in the world
  - Domain Shift
  - Temporal Structure
- Large-Scale Pattern Recognition on suitably collected i.i.d data

CAUSATION(1/2)

x causes y if, were we to intervene and *change* the value of x, then the distribution of y would *also* change as a result.

x changing doesn't *always* change y , but just changes the *probability* that y occurs. As we said earlier, it changes the *distribution* of y.



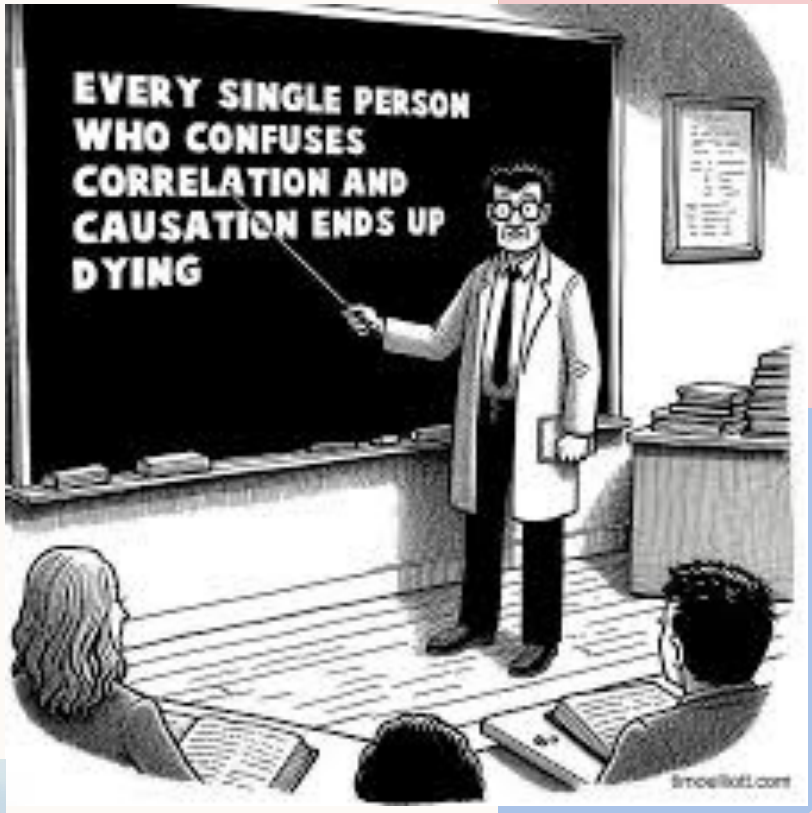
CAUSATION(2/2)

- For many research questions, in order to identify an answer to them we need to have an idea of the data generating process.
- If we can think of some variables as causing others, then the causal relationships between them must be *a part of* that data generating process. If x causes y , then x must be a part of what generates our observations of y.

CAUSALITY AND CORRELATION (1/2)

- **Correlation** describes a statistical association between types of variables
- **Causation** means that changes in one variable brings about changes in the other

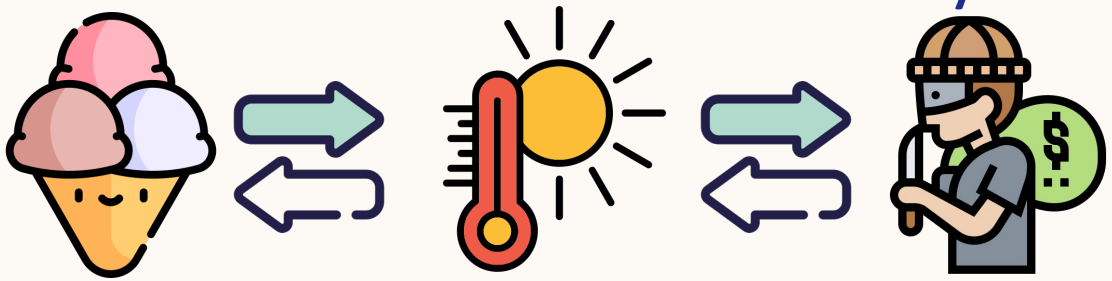
A correlation doesn't imply causation, but causation always implies correlation.



CAUSALITY AND CORRELATION(2/2)

There are two main reasons:

1. The **third variable problem**
   - Ice Cream Sale and Rubbery

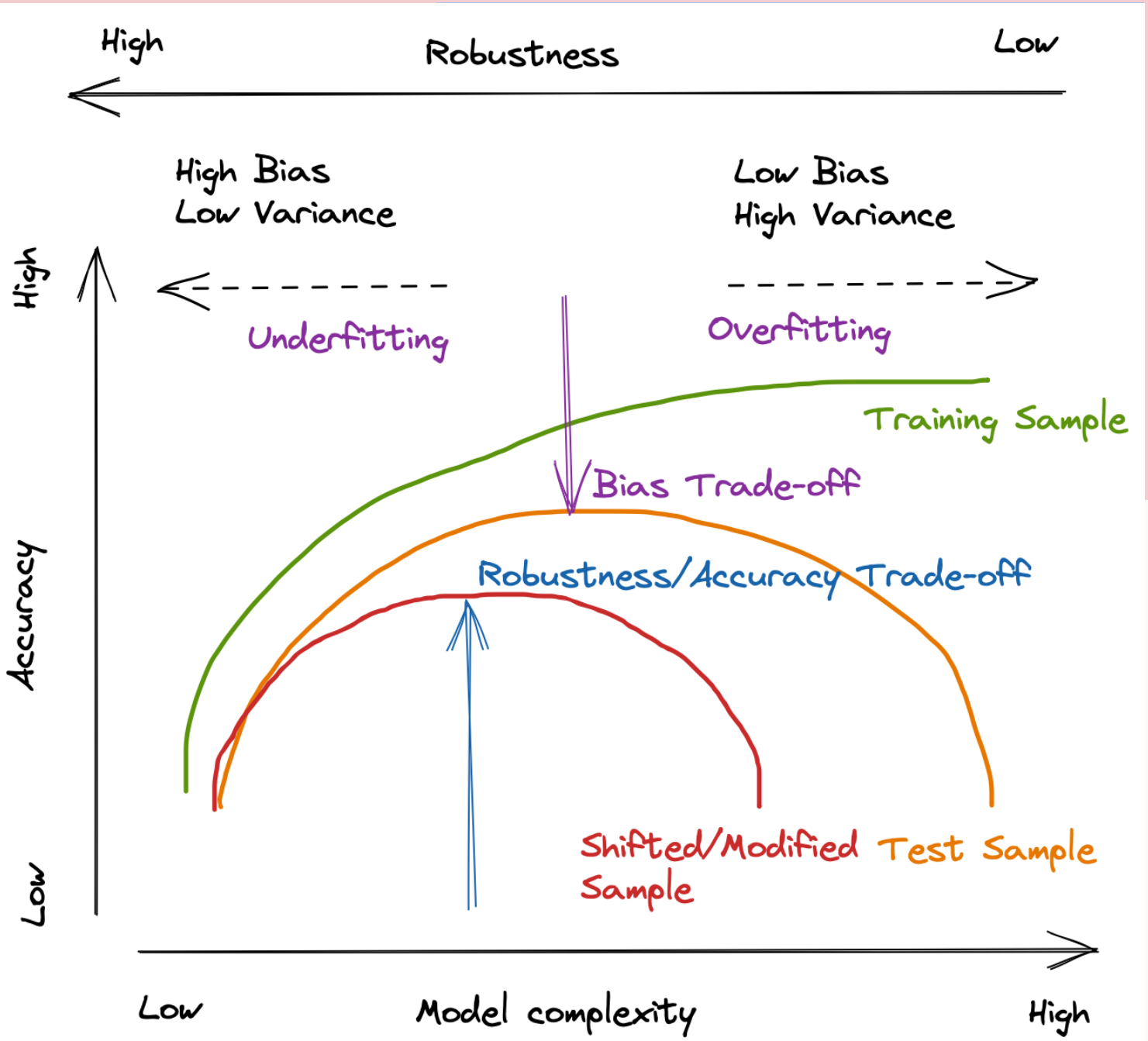


2. The **directionality problem**
   - Vitamin D and Depression

CAUSALITY AND INTELLIGENCE

Issue 1. Robustness

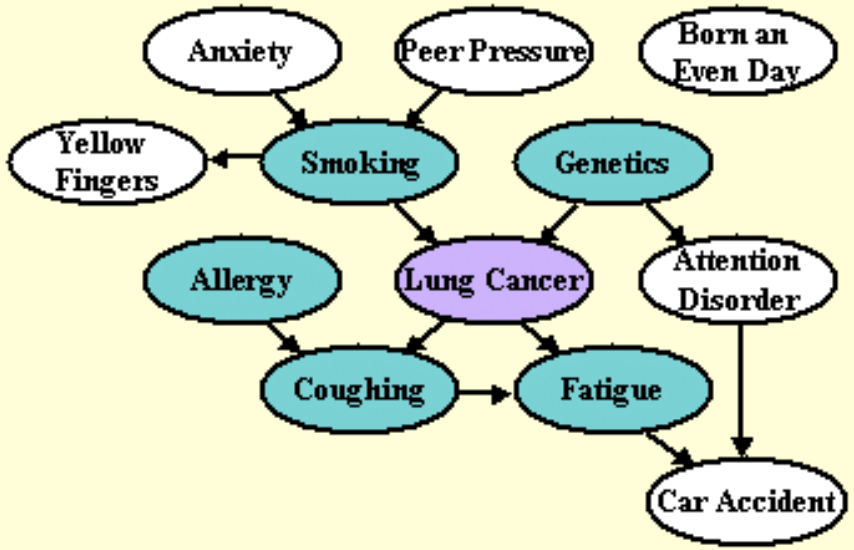
- Robustness of Prediction
- Out of Control Observation in Real World
- Changes in Distribution of Test Data
- Handy Tricks: Data Augmentation, Pre-Training, Self-Supervision, Inductive Biases
- i.i.d Persumation



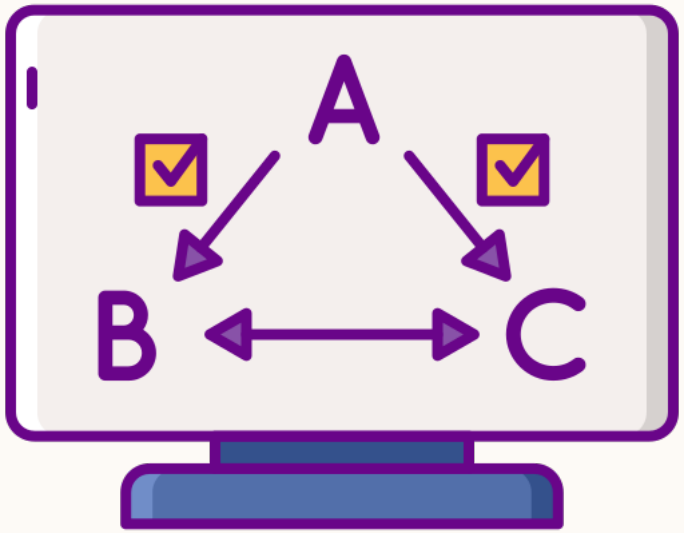
CAUSALITY AND INTELLIGENCE

Issue 2. Learning Reusable Mechanisms

- Infant's Understanding of Physiscs
- Consistency of Rules
- Facing New Environment, Using TRUE Previous Knowledge
- Adapting Parameters, NOT the RULES



CAUSALITY PERSPECTIVE



- Can’t be Fully Described by Boolean Logic or Probabilistic Inference
- Needs Aditional Notion of Intervention
- Conditional probabilities (“seeing people with open umbrellas suggests that it is raining”) cannot reliably predict the outcome of active intervention (“closing umbrellas does not stop the rain”)
- Causal relations can be viewed as components of reasoning chain when prediction based on situations very far from trained distribution
- Discovering causal relations means acquiring robust knowledge that holds beyond the support of observed data distribution and set of training tasks.

SO …

- In the following slides, we'll argue that causality, with its focus on representing STRUCTURAL KNOWLEDGE about data generating process that allows interventions and changes, can contribute toward understanding and resolving some limitations of current ML methods.

LEVELS OF CAUSAL MODELING

- For natural phenomena, set of differential equations modeling mechanisms responsible for time evolution to:
  - ✓ Reason about the effect of interventions
  - ✓ Predict statistical dependencies between variables
  - ✓ Predict future behavior of physical system

$$\frac{dx}{dt} = f(x), \qquad x \in \mathbb{R}^d$$

$$x(t_0) = x_0 \ (Initialization)$$

$$dx = x(t + dt) - x(t), so : x(t + dt) = x(t) + dt.f(x(t))$$

LEVELS OF CAUSAL MODELING

- Although differential equation is a rather comprehensive description of a system, a statistical model can be veiwed as a much more superficial one.
- It often doesn't refer to dynamic processes, but tells us how some of variables allow the prediction of the others as long as experimental conditions do not change. Statistical models learn from observed data and do not have dynamics of the system.

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

PREDICTING IN THE I.I.D SETTING

- Statistical models model the associations of given data and target labels, P(Y|X):It can be proved these questions can be answered by observing a sufficiently large amount of i.i.d data from P(X,Y).
- Despite the impressive advances of machine learning, causality offers an underexplored complement: accurate predictions may not be sufficient to inform decision-making. For example, the frequency of storks is a reasonable predictor for human birth rates in Europe.
- However, as there is no direct causal link between these two variables, a change to the stork population would not affect the birth rates, even though a statistical model may predict so.

PREDICTING UNDER DISTRIBUTION SHIFTS

- Interventions may affect both the value of a subset of causal variables and their relations. For example, “is increasing the number of storks in a country going to boost its human birth rate?” and “would fewer people smoke if cigarettes were more socially stigmatized?”
- As interventions change the joint distribution of the variables of interest, classical statistical learning guarantees **no longer apply.**
- On the other hand, learning about interventions may allow training predictive models that are robust against the changes in distribution that naturally happen in the real world.
- Statistical relations may change due to time or mismatch in train/test. Robustness must be guaranteed in any case.

ANSWERING COUNTERFACTUAL QUESTIONS

Counterfactual Reasoning in 3-year olds (Harris et al 1986)

- Carol **didn't take** her muddy shoes off and walked over the sparkling clean floor.
- The floor is all **dirty**
- If Carol **had taken** her shoes off, **would** the floor be clean or dirty?

→[**clean**]

Counterfactual (subjunctive) Question

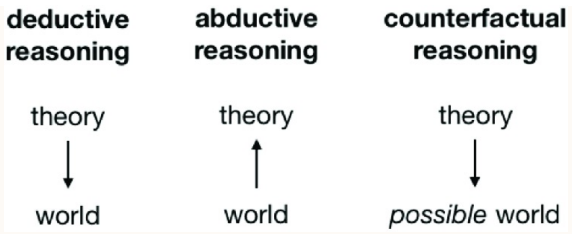
- correct answer

→ they can reason counterfactually (**??**)

Distinction: Reasoning with assumptions **counter-to-fact**

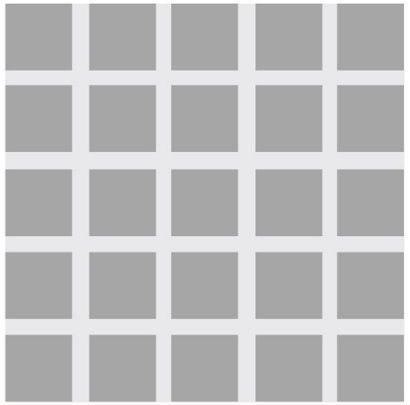
Counterfactual reasoning

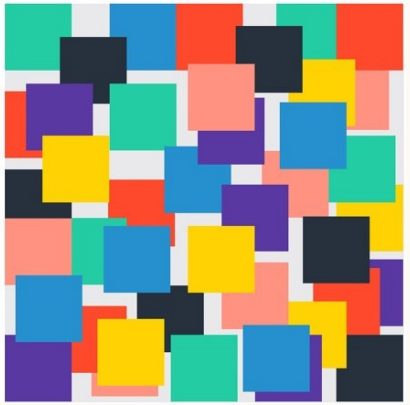
28-11-2010 ESF-LogiCCC



- Harder than Interventions
- This may be a key challenge for AI, as an intelligent agent may benefit from imagining the consequences of its actions and understanding in retrospect what led to certain outcomes, at least to some degree of approximation.
- An interventional question would be "how does the probability of heart failure change if we convince a patient to exercise regularly?" A counterfactual one would be "would a given patient have suffered heart failure if they had started exercising a year earlier?"
- Counterfactuals, or approximations thereof, are especially critical in RL. They can enable agents to reflect on their decisions and formulate hypotheses that can be empirically verified in a process akin to the scientific method.

NATURE OF DATA





The data format plays a substantial role in which type of relation can be inferred. We can distinguish two axes of data modalities:

1. Observational vs. Interventional Data: This is observational in the sense that the data is only observed passively, but it is interventional in the sense that there are interventions/shifts, but unknown to us.
2. Hand-Engineered vs. Raw Data: In classical AI, data are often assumed to be structured into high level and semantically meaningful variables, which may partially correspond to the causal variables of the underlying graph. Raw data, in contrast, are unstructured and do not expose any direct information about causality.

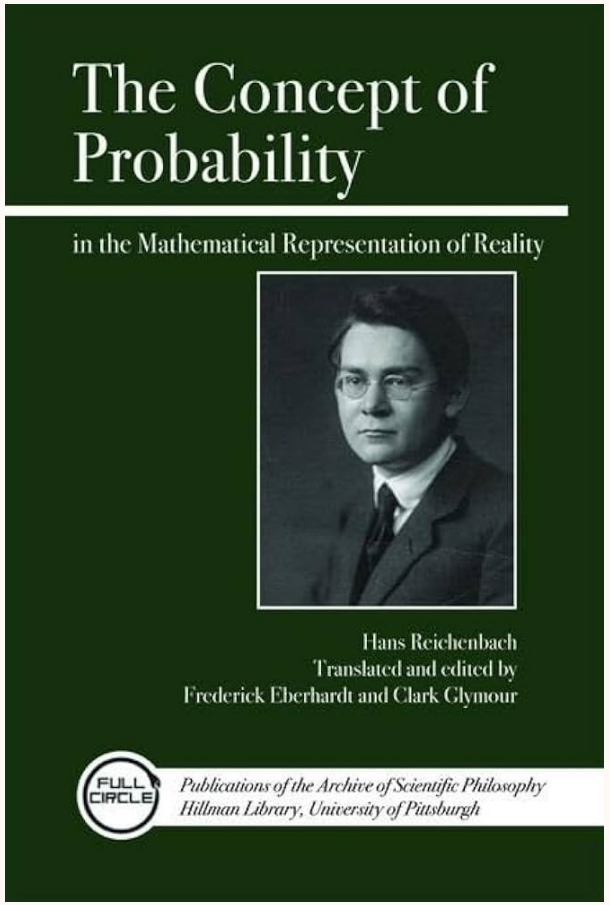
CASUAL MODELS AND INFERENCE

- ✓ Methods Driven by i.i.d. data
- ✓ Reichenbach Principle: From Statistics to Causality
- ✓ Structural Causal Models
- ✓ Differences Between Statistical Models, Causal Graphical Models, and SCMs

METHODS DRIVEN BY I.I.D. DATA

- Strong universal consistency results from statistical learning theory apply, guaranteeing convergence of a learning algorithm to the lowest achievable risks.
- With i.i.d. assumption, the directionality of cause-effect will be lost.
- Recommending is such an intervention, which takes us outside the i.i.d. setting. We no longer work with the observational distribution but a distribution where certain variables or mechanisms have changed.

REICHENBACH PRINCIPLE: FROM STATISTICS TO CAUSALITY



- Common Cause Principle: If two observables X and Y are statistically dependent, then there exists a variable Z that causally influences both and explains all the dependence in the sense of making them independent when conditioned on Z.
- As a special case, this variable can coincide with X or Y . Suppose that X is the frequency of storks and Y the human birth rate. If storks bring the babies, then the correct causal graph is $X \rightarrow Y$. If babies attract storks, it is $X \leftarrow Y$. If there is some other variable that causes both (such as economic development), we have $X \leftarrow Z \rightarrow Y$.

DIRECTED ACYCLIC GRAPH (DAG)

A graphical structure used to represent causal relationships between variables in a system.

Directed

Each edge in the graph has a direction, indicating the direction of causality. For example, if variable A causes variable B, there will be a directed edge from A to B.

Acyclic

There are no cycles in the graph, meaning you can't follow a sequence of edges and return to the same node.

WHY DO WE USE DAG?

- Interventional Studies
- Causal Inference
- Interpretability
- Modeling Causal Mechanisms

STRUCTURAL CAUSAL MODEL(1/3)

The SCM viewpoint considers a set of observables X_1 , . . . , X_n associated with the vertices of a directed acyclic graph (DAG) and assumes that each observable is the result of an assignment :

$$X_i := f_i(PA_i, U_i)\ (i = 1, \ldots, n)$$

f_i is a deterministic function depending on X_i's parents in the graph (denoted by PA_i) and on an unexplained random variable U_i. the set of noises U_1, . . . , U_n is assumed to be jointly independent.

If we specify distributions of U_i , recursive application of the formula allows us to compute the entailed observational joint distribution P(X_1 , . . . , X_n). This distribution has structural properties inherited from the graph and satisfies caudal Markov condition:

Each node X_j conditioned on its parents, is independent of its non descendants.

STRUCTURAL CAUSAL MODEL(2/3)

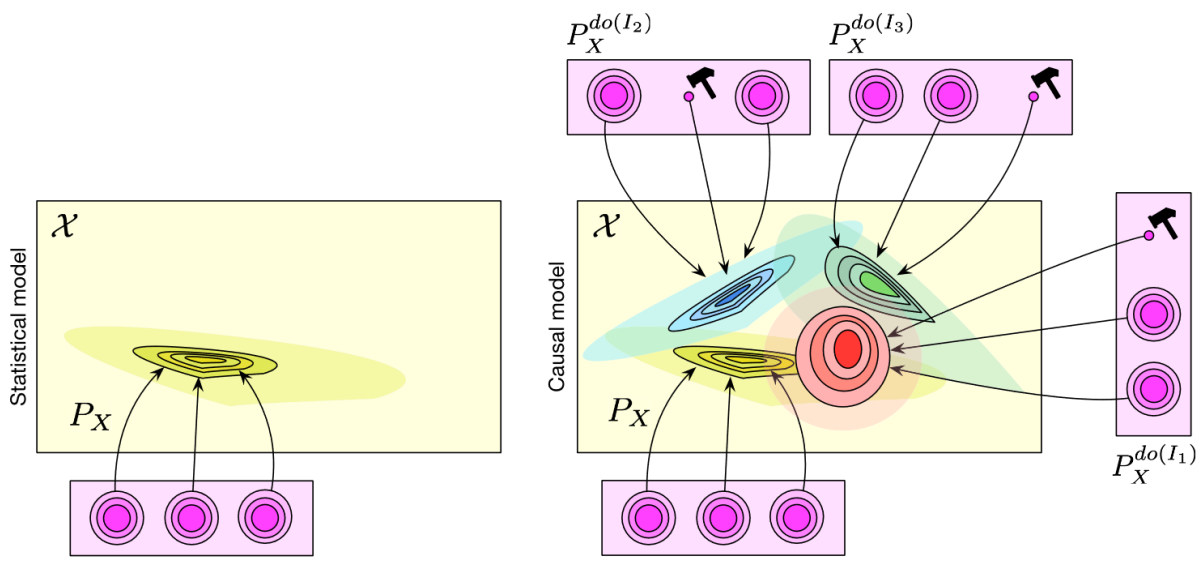
By considering the graph structure and the joint independence of the noises, a canonical factorization of the joint distribution can be defined, which requires causal conditions, which we refer to as causal (or disentangled) factorization:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | PA_i)$$

While many other entangled factorizations are possible, for example:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | X_{i+1}, \dots, X_n)$$

STRUCTURAL CAUSAL MODEL(3/3)



Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention.

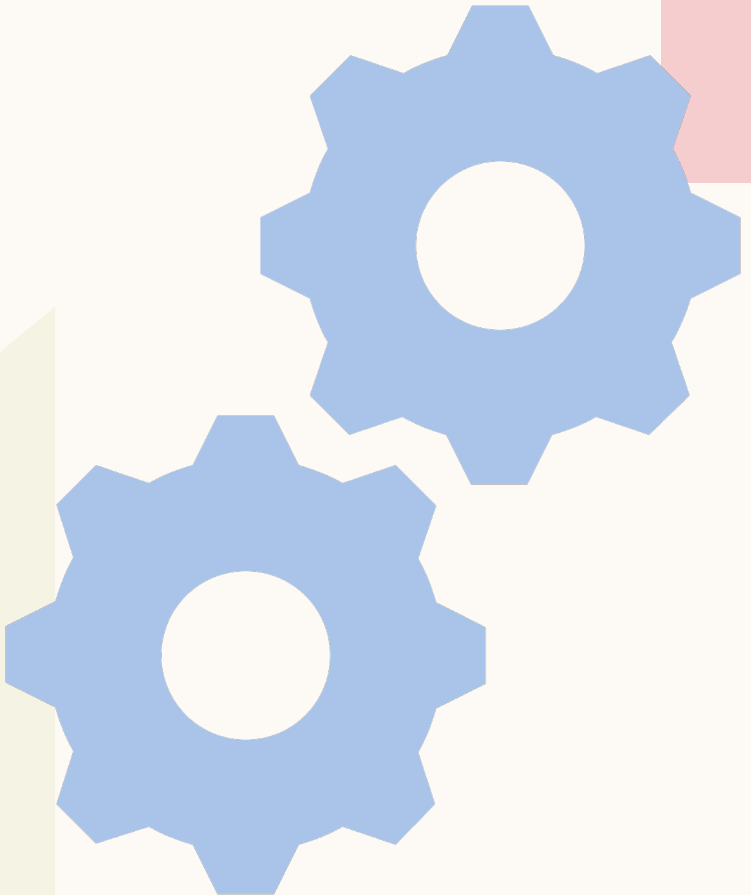
CAUSAL LEARNING AND REASONING

- The conceptual basis of statistical learning is a joint distribution P(X1,...,Xn), and we make assumptions about function classes used to approximate.
- Causal learning considers a richer class of assumptions and seeks to exploit the fact that the joint distribution possesses a causal factorization. It involves the causal conditionals P (Xi|PAi), how these conditionals relate to each other, and interventions or changes that they admit.
- Once a causal model is available, either by external human knowledge or a learning process, causal reasoning allows drawing conclusions on the effect of interventions, counterfactuals, and potential outcomes. In contrast, statistical models only allow reasoning about the outcome of i.i.d. experiments.

WHY CAUSAL REPRESENTATION LEARNING?

	be learned from unstructured data such as images and text	predict reliably under real-world data distribution shifts
Statistical models	✓	✗
Causal models	✗	✓

Causal representation learning aims to incorporate ideas from both representation learning and causal inference in order to learn models from unstructured data which have desirable properties of causal models, such as robustness to data distribution shifts.



INTERVENTION(1/6)

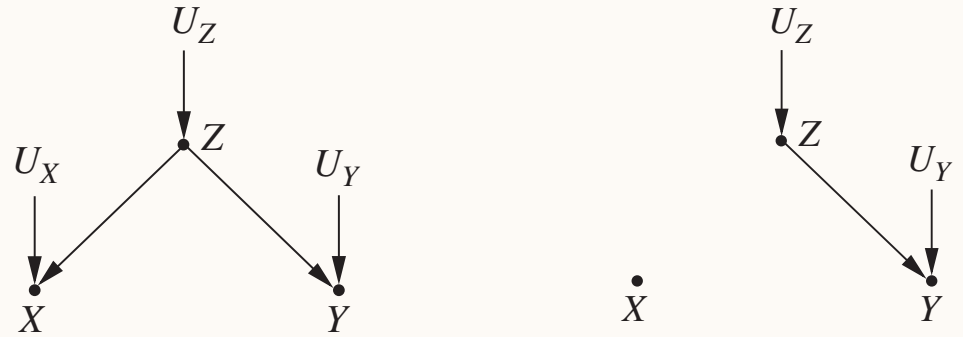
The University of Winnipeg study that showed that heavy text messaging in teens was correlated with “shallowness.” Media outlets jumped on this as proof that texting makes teenagers more shallow. (Or, to use the language of intervention, that intervening to make teens text less would make them less shallow.) The study, however, proved nothing of the sort. It might be the case that shallowness makes teens more drawn to texting. It might be that both shallowness and heavy texting are caused by a common factor—a gene, perhaps—and that intervening on that variable, if possible, would decrease both.

INTERVENTION(2/6)

The difference between intervening on a variable and conditioning on that variable should, hopefully, be obvious. When we intervene on a variable in a model, we fix its value. We change the system, and the values of other variables often change as a result. When we condition on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value we are interested in. What changes, then, is our perception about the world, not the world itself.

INTERVENTION(3/6)

When we intervene to fix the value of a variable, we curtail the natural tendency of that variable to vary in response to other variables in nature. This amounts to performing a kind of surgery on the graphical model, removing all edges directed into that variable.



INTERVENTION(4/6)

The notion of an intervention is a defining characteristic of causal modeling that differentiates it from statistical modeling. Consider X → Y :

- If we intervene on X, then P (Y | do(X = x)) is the population distribution of Y if we fix everyone in the population's X value to x
- The conditional probability P (Y | X = x) is the distribution of Y in the subset of the population where X was x

In general, P (Y | do(X = x)) does not equal P (Y | X = x))

INTERVENTION(5/6)

1) No intervention: Only observational data are obtained from the causal model.

2) Hard/perfect: The function in the structural assignment of a variable (or, analogously, of multiple variables) is set to a constant (implying that the value of the variable is fixed), and then, the entailed distribution for the modified SCM is computed.

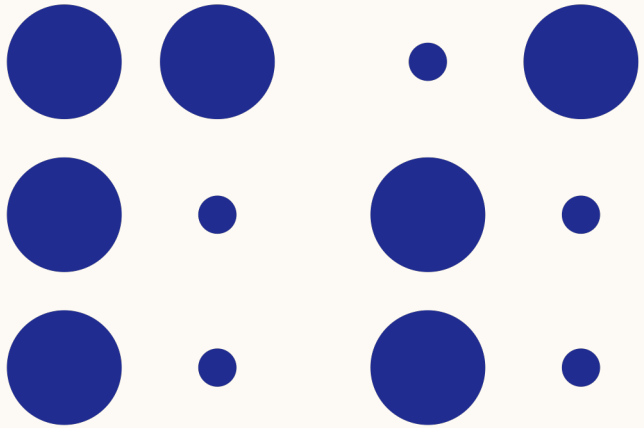
INTERVENTION(6/6)

3) Soft/imperfect: The structural assignment for a variable is modified by changing the function or the noise term (this corresponds to changing the conditional distribution given its parents).

4) Uncertain: The learner is not sure which mechanism/variable is affected by the intervention.

INDEPENDENT CAUSAL MECHANISM PRINCIPLE (1/2)

The causal generative process of a system's variables is composed of **autonomous modules** that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

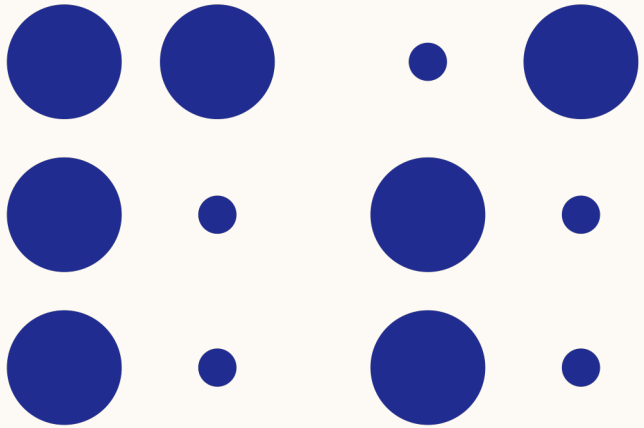


INDEPENDENT CAUSAL MECHANISM PRINCIPLE (2/2)

ICM principle consequences

No flow of influence: intervening upon one mechanism $p(X_i|PA_i)$ does not change the other mechanisms $p(X_j|PA_j), i \neq j$

No flow of information: knowing a mechanism $p(X_i|PA_i)$ does not give us information about another mechanism $p(X_j|PA_j), i \neq j$



REPRESENTATION LEARNING

Representation Learning is a process in machine learning where algorithms extract meaningful patterns from raw data to create representations that are easier to understand and process.

Representation learning can be divided into:

- Supervised representation learning
- Unsupervised representation learning

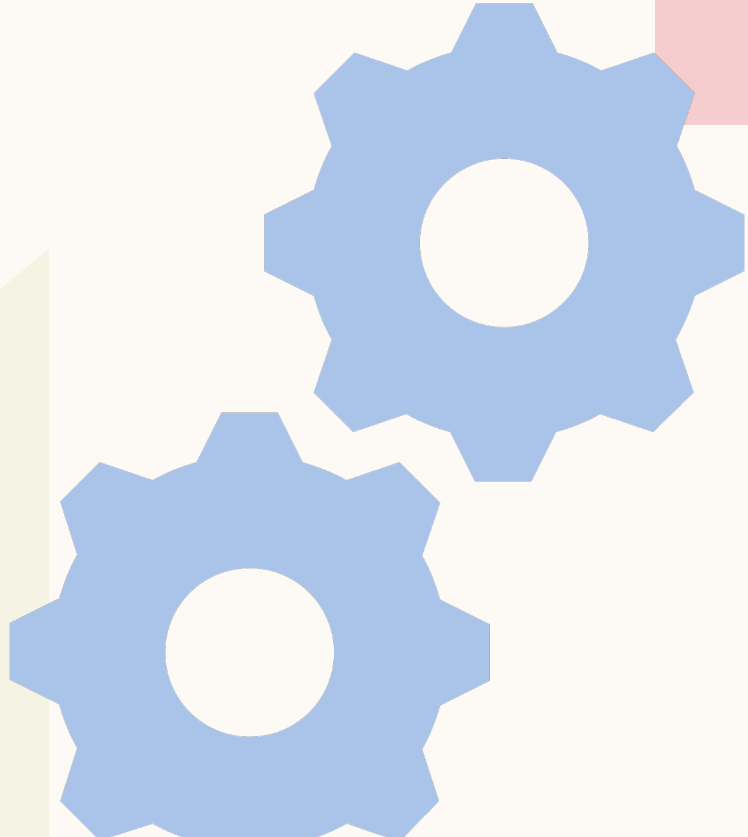
Goals of representation learning are:

- Interpretability
- Reveal hidden features
- Be used for transfer learning

CAUSAL REPRESENTATION LEARNING

CAUSAL REPRESENTATION LEARNING

Causal representation model is a mathematical framework used to understand causal relationships between variables in a system. It aims to uncover how changes in one variable affect other variables over time.



CAUSAL REPRESENTATION LEARNING

Due to SCM, noise terms are independent so the disentangled representation is feasible:

$$P(S_1, \dots, S_n) = \prod_{i=1}^{n} P(S_i | PA_i)$$

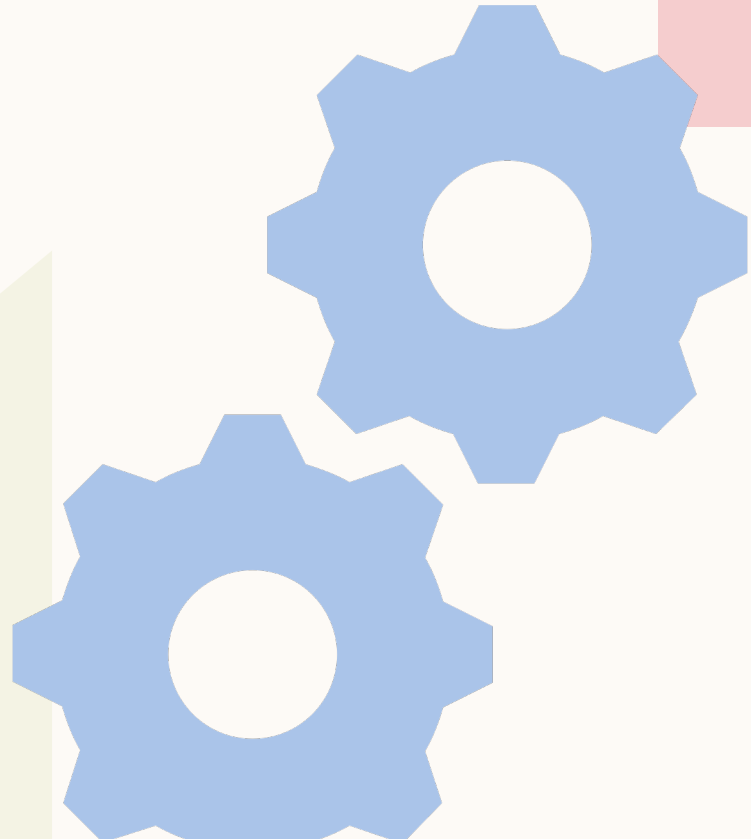
Suppose that we seek to reconstruct such a disentangled representation using independent mechanisms from data, but the causal variables S_i are not provided to us a priori. Rather, we are given (possibly high- dimensional) X = $(X_1, \dots, X_d)$. we should construct causal variables (n $\ll$ d) as well as mechanisms

$$S_i := f_i(PA_i, U_i)$$

CAUSAL REPRESENTATION LEARNING

1. Use an encoder q: $\mathbb{R}^d \to \mathbb{R}^n$ taking X to a latent "bottleneck" representation comprising the unexplained noise variables $U = (U_1, \dots, U_n)$
2. Map f(U) determined by structural assignments $f_1, \dots, f_n$
3. Apply a decoder p: $\mathbb{R}^n \to \mathbb{R}^d$

For suitable n, the system can be trained using reconstruction error to satisfy p ○ f ○ q. If the causal graph is known, the topology of a neural network implementing f can be fixed accordingly; if not, the neural network decoder learns the composition p̃ = p○f. In practice, one may not know f and, thus, only learn an autoencoder p̃ ○ q, where the causal graph effectively becomes an unspecified part of the decoder p̃, possibly aided by a suitable choice of architecture.



CAUSAL VAE

DISENTANGLED REPRESENTATION LEARNING VIA NEURAL STRUCTURAL CAUSAL MODELS

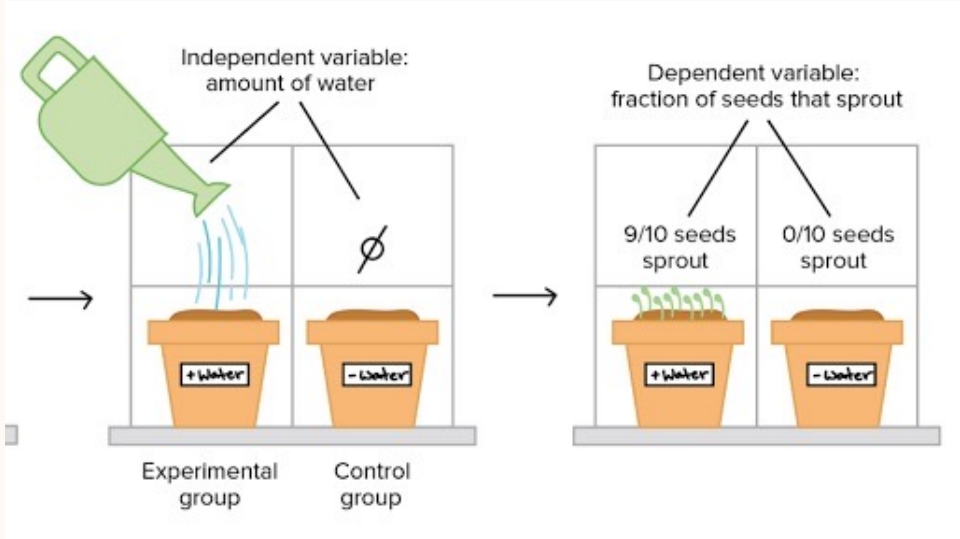
A new Variational Autoencoder (VAE) based framework named CausalVAE, which includes a Causal Layer to transform **independent exogenous factors** into causal **endogenous** ones that correspond to causally related concepts in data.

VAE + SCM ⟶ Causal VAE

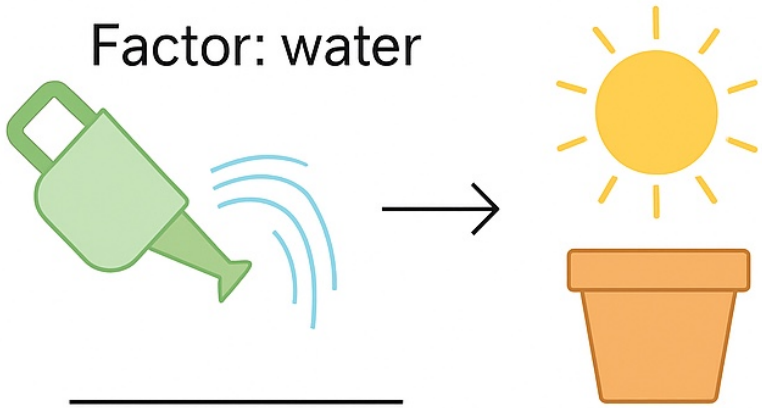
WHY CAUSAL VAE?

Most existing works of disentangled representation learning make a common assumption that **the real world observations are generated by countable independent factors**.

we argue that in many real world applications, latent factors with semantics of interest are causally related and thus we need a new framework that supports causal disentanglement.



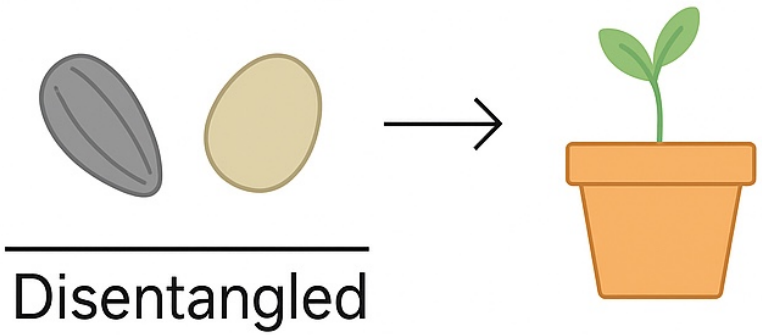
Factor: water



Disentangled

Causal

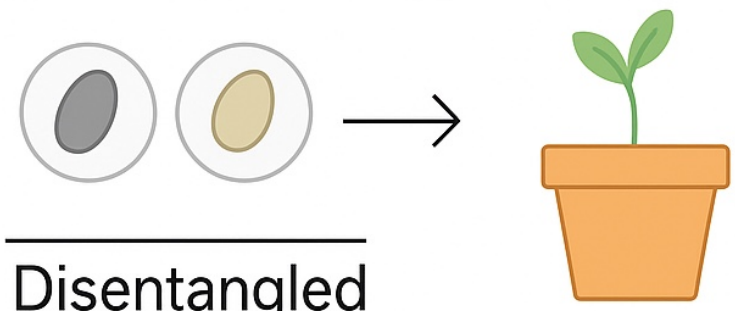
Factor: sunlight



Disentangled

Causal

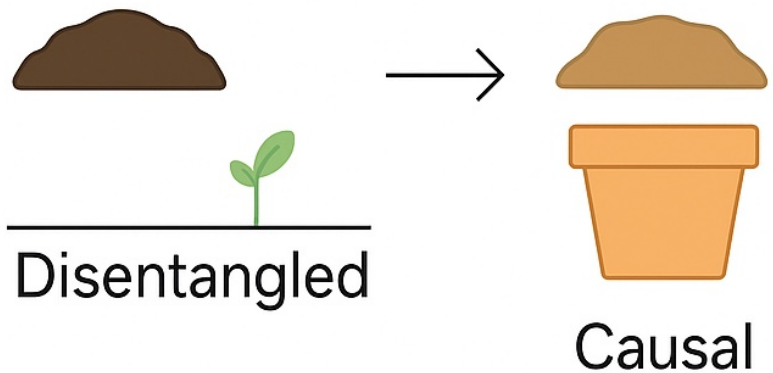
Factor: seed type



Disentangled

Causal

Factor: soil type



Disentangled

Causal

HOW CASUAL VAE WORKS?

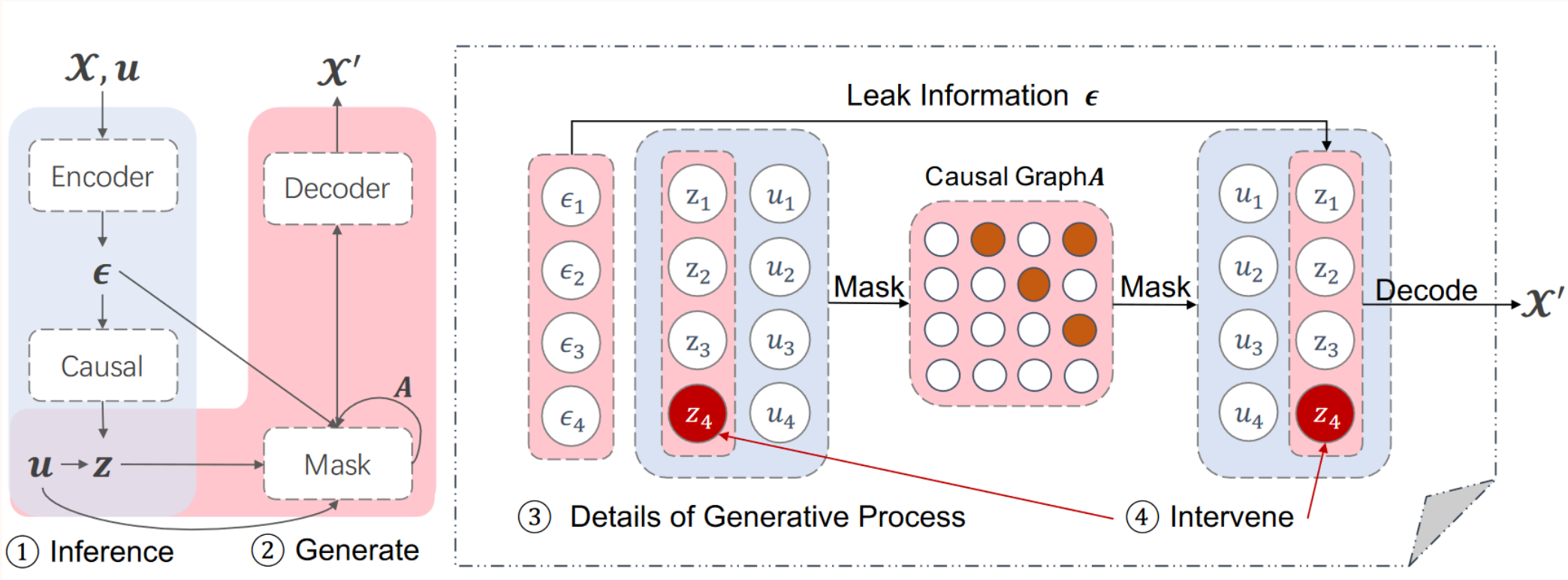
CasualVAE is a VAE-based causal disentangled representation learning framework by introducing a novel **Structural Causal Model layer (Mask Layer)**, which allows it to recover the latent factors with semantics and structure via a causal DAG.

The input signal passes through an encoder to obtain independent exogenous factors and then a Causal Layer to generate causal representation which is taken by the decoder to reconstruct the original input.

additional information is required as weak supervision signals to achieve causal representation learning. By **weak supervision** the causal structure of the latent factors is automatically learned, instead of being given as a prior in.

To train the model, a new loss function used which includes the VAE evidence lower bound (ELBO) loss and an acyclicity constraint imposed on the learned causal graph to guarantee its **DAGness**.

HOW CASUAL VAE WORKS?



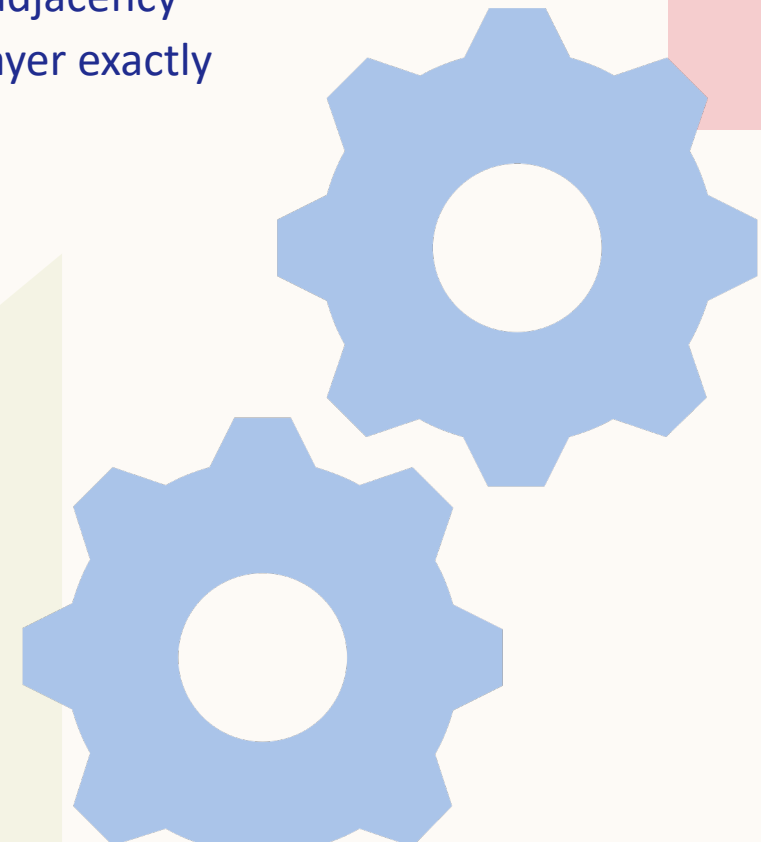
HOW CASUAL VAE WORKS?

FORMALIZED CAUSAL REPRESENTATION

To formalize causal representation, we consider n concepts of interest in data. The concepts in observations are causally structured by a Directed Acyclic Graph (DAG) with an adjacency matrix A. Though a general nonlinear SCM is preferred, for simplicity, the Causal Layer exactly implements a Linear SCM as described in Equation:

$$\mathbf{z} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{z} + \epsilon = \left(\boldsymbol{I} - \boldsymbol{A}^{\boldsymbol{T}}\right)^{-\mathbf{1}}\epsilon,\ \epsilon \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{I})\ \ (1)$$

A is the parameters to be learnt in this layer.ϵ are independent Gaussian exogenous factors and $z \in \mathbb{R}^n$ is structured causal representation of n concepts that is generated by a DAG and thus A can be permuted into a strictly upper triangular matrix.



HOW CASUAL VAE WORKS?

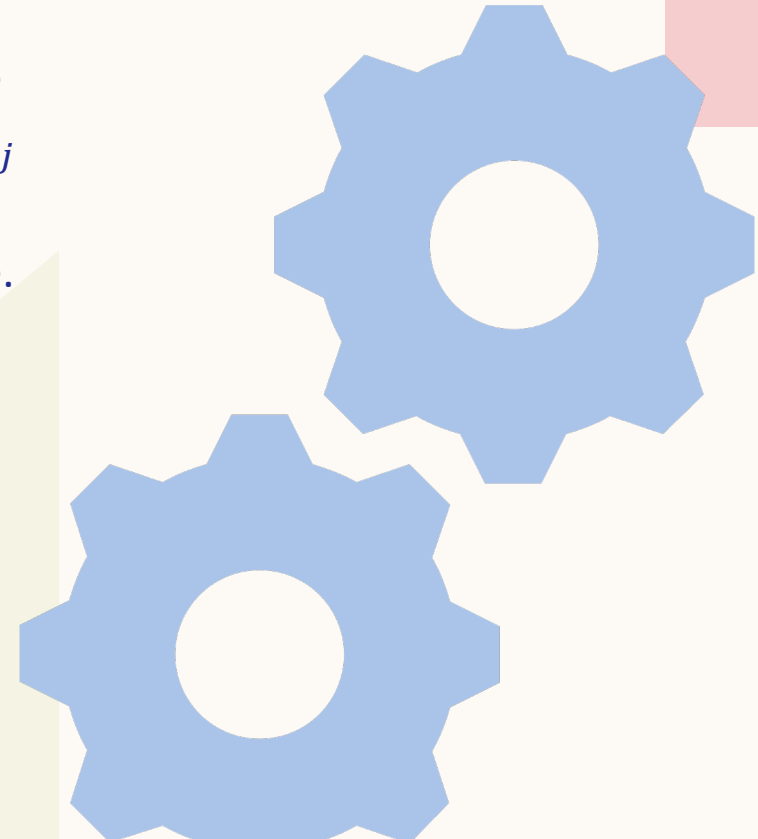
STRUCTURAL CASUAL MODEL LAYER (1/2)

Once the causal representation z is obtained, it passes through a Mask Layer to reconstruct itself.

Let z_i be the i-th variable in the vector z. The adjacency matrix associated with the causal graph $is\ A\ =\ [A_1|\ldots|A_n]$ where $A_i \in \mathbb{R}^n$ is the weight vector such that A_{ij} encodes the causal strength from z_j to z_i. We have a set of mild nonlinear and invertible functions $[g_1, g_2\ldots, g_n]$ that map parental variables to the child variable. Then we write:

$$\boldsymbol{z_i = g_i(A_i \circ z\ ;\ \eta_i) + \epsilon_i}$$

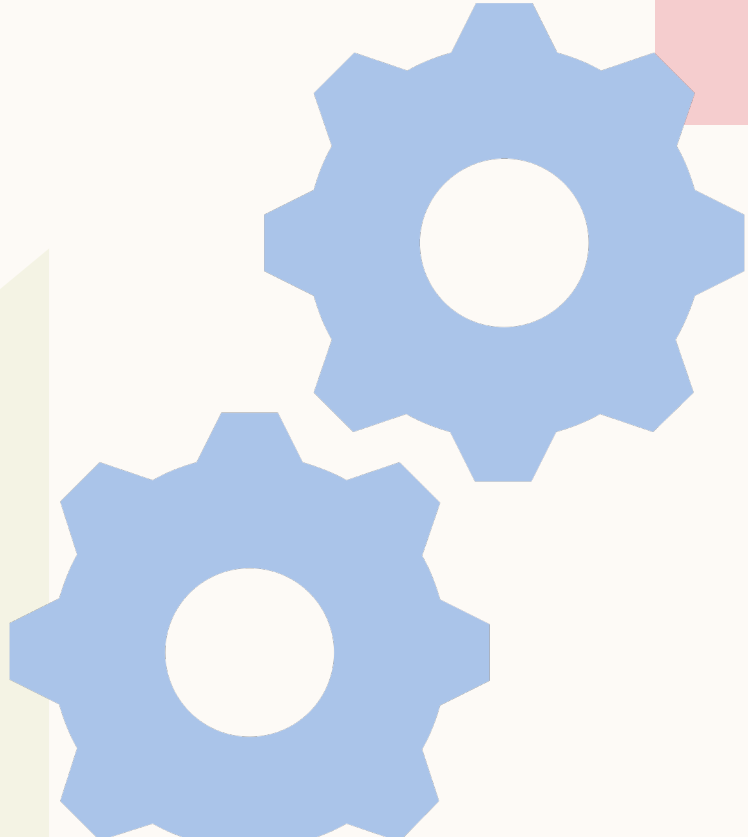
○ is the element-wise multiplication and η_i is the parameter $g_i(.)$



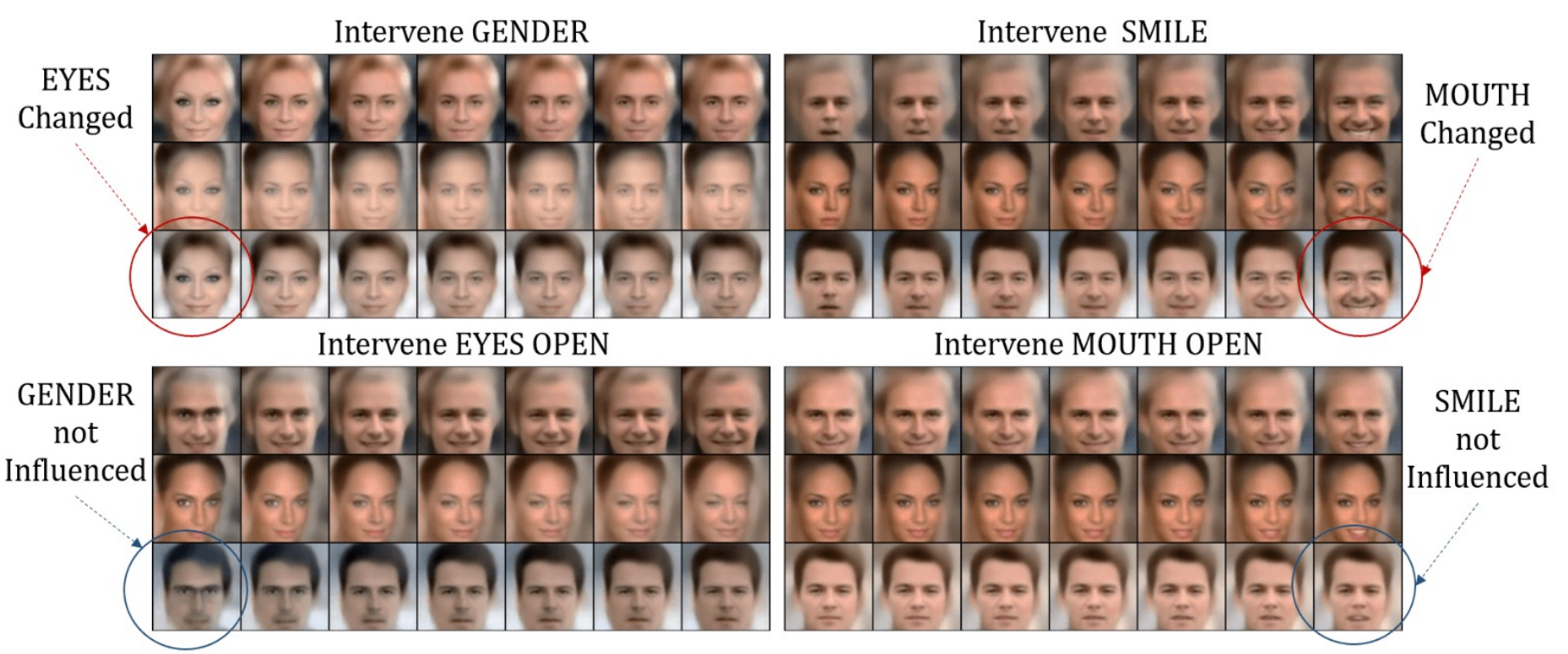
HOW CASUAL VAE WORKS?

STRUCTURAL CASUAL MODEL LAYER (2/2)

we find that adding a mild nonlinear function g_i results in more stable performances. To show how this masking works, consider a variable z_i and $A_i \circ z$ equals a vector that only **contains its parental information** as it masks out all z_i's non-parent variables. By **minimizing the reconstruction error**, the adjacency matrix A and the parameter η_i of the mild nonlinear function g_i are trained.



RESULTS OF CAUSALVAE MODEL ON CELEBA(SMILE).



The controlled factors are GENDER, SMILE, EYES OPEN and MOUTH OPEN respectively.

CONCLUSION

1. Efficiency in Learning Causal Dynamics: Causal representation learning excels in understanding complex systems by directly modeling causal relationships, enabling efficient learning of dynamic systems' behavior.

2. Robust Decision Making: Causal representations provide a more robust foundation for decision-making in uncertain environments by capturing the underlying causal mechanisms driving observed phenomena.

3. Generalization Across Contexts: Unlike disentangled representations, causal representations generalize well across diverse contexts, facilitating transfer learning and adaptation to new environments without extensive retraining.

4. Interpretability and Explainability: Causal representations offer interpretable and explainable models, allowing humans to understand why certain predictions or actions are made, which is crucial in critical applications like healthcare and finance.

5. Counterfactual Reasoning: Causal representations enable sophisticated counterfactual reasoning, allowing systems to understand the consequences of different actions and interventions, essential for planning and policy-making.

6. Discovering Latent Variables: Causal representation learning can automatically discover latent variables and their causal relationships, leading to a more compact and informative representation of complex data.

7. Robustness to Distribution Shifts: Causal representations are more robust to distribution shifts and changes in the data generating process, making them suitable for real-world applications where data distribution may vary over time.

REFRENCES

- Toward Causal Representation Learning

  BERNHARD SCHÖLKOPF , FRANCESCO LOCATELLO , STEFAN BAUER , NAN ROSEMARY KE, NAL KALCHBRENNER, ANIRUDH GOYAL, YOSHUA BENGIO, Proceedings of the IEEE, vol. 109, no. 5, pp. 612-634, May 2021

- Causal Representation Learning

  Sanae Lotfi, Taro Makino, Lily Zhang, Inference and Representation, IEEE, DS-GA 1005, Fall 2021

- CausalVAE: Structured Causal Disentanglement in Variational Autoencoder

  Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, Jun Wang, CVPR2021

- From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling

  Aneesh Komanduri, Xintao Wu, Yongkai Wu, Feng Chen, arxiv2023

THANKS FOR YOUR ATTENTION

Any Questions ?